Descriptive Analytics and Customer Insight

What the Average Is Made Of: Counting, Comparing, and Decomposing on Certified Data

Dr. Jose Mendoza, Academic Director and Clinical Associate Professor
Version 1.0 · July 2026

Chapter Information

ABSTRACT

This chapter develops descriptive analytics as the discipline of summarizing certified data without distorting it. It treats counting and the denominator problem, the choice of a measure of center against distribution shape, and spread as a finding in its own right. Its structural core is the cross-tabulation: three percentage bases, the direction rule, and Simpson's paradox computed rather than merely cited. The chapter then assembles the applied marketing metric families and the decomposition identity, which factors a composite metric such as average order value exactly into basket size and revenue per unit. Cohort retention tables and campaign funnels extend description along time and along the customer journey; concentration analysis, benchmarking, and the insight statement close the sequence. The labs decompose StyleCraft's suburban-versus-urban average-order-value gap and build profiles, cohorts, concentration, and campaign metrics on the certified file.

KEYWORDS

descriptive analytics; frequency tables; cross-tabulation; simpson's paradox; average order value; cohort analysis; conversion funnel; concentration analysis; customer insight; marketing analytics

VERSION AND DATE

Version 1.0 · July 2026 · Language: English (United States)

SUGGESTED CITATION

Mendoza, J. (2026). Descriptive analytics and customer insight. In *Applied business analytics for marketing decision-making: Business analytics and data visualization* (Chapter 5, Version 1.0) [Open educational resource]. CC BY 4.0.

identification only and do not imply endorsement. StyleCraft Collective is a fictional company created for instruction.

COMPANION REPOSITORY
Datasets, notebooks, and figure sources for this chapter: https://github.com/jrmst102/businessanalytics

Chapter Learning Objectives

By the end of this chapter, students should be able to:

1. Produce and interpret frequency tables, proportions, summary statistics, and cross-tabulations from transaction and customer data.
2. State the denominator problem and, for any reported proportion or rate, identify the population it counts over and the populations it silently excludes.
3. Choose the measure of center that fits a distribution's shape, and defend the choice of mean, median, or mode for a skewed marketing quantity.
4. Report spread alongside center — range, interquartile range, and standard deviation — and recognize heterogeneity as a finding rather than noise.
5. Read a cross-tabulation three ways, compute row and column percentages correctly, and detect Simpson's paradox by checking group composition before trusting a comparison of rates.
6. Build customer profiles and group comparisons on distributions, not just on means, from marketing data.
7. Summarize revenue, acquisition, conversion, retention, and engagement metrics at stated grains, and verify every summary against known certified totals.
8. Decompose a composite metric exactly into its factors — including average order value into basket size and revenue per unit — and verify that the factors multiply back.
9. Construct and read a period-active cohort retention table and a descriptive campaign funnel, and explain what cohort alignment does and does not remove from a comparison.
10. Quantify concentration with deciles and Pareto shares, place numbers in context with benchmarks, and turn descriptive findings into insight statements written in managerial language — while keeping descriptive language descriptive.

Chapter 4 ended with two things in hand: a certified table of record, reconciled to finance's known total with every departure from the raw file logged, and a first exploratory look that turned up a right-skewed, arguably two-humped distribution of order values — a shape that hinted at two customer worlds living inside one company. Chapter 4 also made a promise: that the next chapter would finally explain what the suburban stores' celebrated average order value is actually made of. This chapter keeps that promise. It develops descriptive analytics as a precision discipline — counting, summarizing, cross-tabulating, profiling, and decomposing — and shows that the most consequential analytical work at most companies is exactly this work, done carefully and narrated honestly. The destination is customer insight: descriptive findings translated into statements a decision-maker can act on without being misled by them.

CONCEPT

What This Chapter Is Really About

Chapter 4 argued that cleaning is the first analysis. This chapter makes the companion argument for the step that follows: summarizing is the second analysis, and it is nowhere near as innocent as it looks. Every descriptive choice — which denominator, which measure of center, which grouping, which percentage direction, which comparison — changes the story a certified table tells, without changing a single number in the table. The suburban stores' high average order value is a true number; whether it means the expansion is working depends entirely on what that average is made of and who is underneath it. An undisciplined analyst reports the number; a disciplined analyst decomposes it, reconciles every summary to the certified totals, reports the spread with the center, and writes insight statements that say exactly as much as the description supports and no more. In other words, description is not the warm-up before the real analytics. Done precisely and narrated well, description is most of the analytics value a business ever receives.

5.1 Marketing Decision Context: The Meeting About One Number

Chapter 4 closed with a quiet institutional achievement: ten days of disciplined preparation produced a certified table of record, reconciled to finance's quarterly total, with a verification log standing behind every row that was removed or changed. The quarterly expansion review can now argue about the business instead of arguing about the file. This chapter is that argument.

The review convenes in seven days, and the Chapter 2 memo committed the analytics team to two deliverables for it: a same-age cohort comparison of the new suburban stores against the urban stores at the same age, and a store-economics summary. Both are, in the vocabulary of Section 1.4, descriptive work — no model, no forecast, no experiment. Just counts, averages, rates, and comparisons, computed on the certified table. On paper, the easy week.

Then the pre-read goes out, and one number swallows the meeting before it starts. The finance team's summary shows that the average order value at the suburban and resort stores runs far above the average order value online and at the urban stores — not by a rounding margin, but by a multiple. Within a day, the number has acquired a life of its own. The head of retail circulates it with a one-line message: proof the affluent-suburban strategy is working, and grounds to unpause the next wave of store openings. The CFO reads the same number more cautiously and asks the question that gives this chapter its work: catalog prices top out near ninety dollars, the same weekly drops ship to every store, and nothing at Greenwich costs more than it costs in SoHo — so what, exactly, is that average made of? And the CRM manager, whose worry about declining repeat purchase opened this guide back in Chapter 1, adds the uncomfortable companion question: an average is computed over orders — what does it say, if anything, about the customers placing them?

The decision at stake is capital allocation: whether the review recommends resuming the suburban store pipeline, and the pre-read number is on track to carry that decision by itself. The analytics lead's task is therefore not to produce more numbers but to make the numbers already

on the table mean something defensible. That requires answering three questions in descriptive language. First, the composition question: the average-order-value gap must be decomposed — since unit prices cannot explain it, is it larger baskets, a different product mix, or both, and by what measurable factors? Second, the people question: averages summarize orders and customers into single figures — who is underneath the suburban average, how much do they vary, and is the healthy-looking mean hiding an unhealthy distribution? Third, the comparison question: every comparison in the pre-read pools stores, channels, and customer vintages of very different compositions — do the headline comparisons survive when the groups are made comparable, or is the review about to be steered by a mix effect wearing a trend costume?

The deadline is fixed and the audience is not technical. Whatever the analytics lead finds must arrive as clear tables and short sentences: this is what the number is made of, this is who is underneath it, this is what it does and does not support. The chapter builds that toolkit in order. Sections 5.2 through 5.4 establish precise counting and honest summarizing — frequency tables and the denominator problem, the choice of center, and spread as a finding. Sections 5.5 and 5.6 build the comparison machinery: cross-tabulations, the three percentages, Simpson's paradox, and group profiles. Sections 5.7 through 5.9 assemble the applied metric families, the decomposition identity that answers the CFO's question, cohorts and funnels, and the concentration and benchmarking context that turns numbers into claims. Section 5.10 examines how AI assistants help and fail at exactly this work, and the labs in Section 5.11 perform the decomposition on StyleCraft's certified data — where the answer, by design, is checkable.

5.1.1 Opening Case Questions

Keep these questions in mind while reading, and return to them after completing the labs.

1. The suburban average order value is a true number computed from certified data. List three different decisions a stakeholder might take from it, and what additional descriptive fact each decision silently assumes.
2. Catalog prices top out near $90 in every store. If unit prices cannot explain the average-order-value gap, what two quantities could? Write the arithmetic identity that connects them to average order value before reading Section 5.7.
3. The CRM manager objects that an average over orders says nothing about customers. What single change of grain (Section 3.3) turns the order-level average into a customer-level claim, and what new denominator does it require?
4. The pre-read compares all suburban stores against all urban stores. The stores opened in different waves and their customers differ in tenure. Which of this chapter's tools protects that comparison, and from what?

5.2 Counting Done Right: Frequency Tables, Proportions, and the Denominator Problem

The opening case turns on the meaning of one average, but averages are built from counts, so the chapter begins where description begins: counting things and stating shares. This looks beneath a graduate course's dignity until the first dispute, which in practice arrives within minutes of the first table.

A frequency table is the simplest descriptive object there is: the distinct values of one variable, and how many observations take each value. Run against the certified transactions table, a frequency table of channel answers "how many order lines came through Online, App, and Store"; run against customers, a frequency table of loyalty_tier answers "how many customers sit in each tier." The count column is rarely the useful one, however. Decision language runs on shares, so the working form of the frequency table adds a proportion column — each count divided by the total — and the moment a proportion appears, so does the chapter's first discipline: the total it is divided by must be named.

DEFINITION

Frequency Table

A frequency table is a summary that lists each distinct value (or binned range) of a variable together with the count of observations taking that value and, in its working form, each count's proportion of a stated total. A frequency table is complete only when its unit of analysis and its denominator — what one row of the underlying data is, and what the proportions are shares of — are stated alongside it.

Source: Adapted from Freedman et al. (2007) and Agresti (2019).

Table 5.1 shows the working form on an illustrative extract of one thousand order lines, and it already contains a small trap worth pausing on. The proportions in the table are shares of order lines, because the underlying data sits at the order-line grain established in Section 3.3. They are not shares of orders (an order can span several lines), not shares of revenue (lines carry different values), and not shares of customers (a frequent customer contributes many lines). All four statements — share of lines, of orders, of revenue, of customers — are legitimate descriptive claims, they will generally all differ, and each answers a different business question. "Store accounts for 44 percent of order lines" and "Store accounts for 31 percent of customers" can both be true of the same file at the same time.

Table 5.1

A frequency table in working form (illustrative extract, 1,000 order lines)

Channel	Order lines	Share of lines
Online	460	46.0%
Store	440	44.0%
App	100	10.0%
Total	1,000	100.0%

This is the denominator problem, and it deserves its name because it is the single most common way a correct count becomes a wrong claim. Chapter 3 established that a complete ratio metric names its numerator, denominator, window, and filters (Section 3.5.1); the denominator problem is what happens in descriptive work when that discipline lapses — a share is reported, the reader supplies a denominator from imagination, and the imagined denominator is not the one used. An email "open rate of 38 percent" means one thing per email delivered and another per email sent; a "repeat purchase rate" moves by several points depending on which customers the denominator admits, which is exactly the dispute Chapter 3's lab staged (Section 3.10.3); and the pre-read's average order value is revenue per order — a denominator choice with everything riding on it, as Section 5.7 will show. The remedy is not sophistication but habit: every proportion, rate, and average travels with its denominator, in the table note or in the sentence, every time.

CONCEPT

The Denominator Problem

Any count can be turned into several different proportions, each correct over its own denominator and each supporting a different claim. A share is not a fact about the numerator; it is a fact about a numerator-denominator pair. Never report or accept a proportion, rate, or average without naming what it is divided by — and when two stakeholders quote different values for the "same" rate, check the denominators before checking the arithmetic (see Section 3.5.1 for the full metric-definition discipline).

Moreover, counting done right respects one more habit inherited from Chapter 4: the frequency table's total row is a verification instrument, not decoration. The counts in Table 5.1 must sum to the certified extract's row count, and the proportions to 100 percent; a frequency table of the full transactions file whose total disagrees with the verification log's certified row count is announcing that a filter ran silently somewhere. This chapter will make that reconciliation habit systematic in Section 5.7, but it starts here: the first thing to check about any summary is that it still adds up to the file it summarizes. With counting and shares established,

the next section turns to the summary that dominates every marketing conversation — the average — and to the choice hiding inside it.

5.3 The Center That Fits the Shape: Mean, Median, and Mode

Section 5.2 counted values one variable at a time. The natural next step is to collapse a numeric variable into a single representative figure — a measure of central tendency — and here description makes its first genuinely consequential choice, because there are three standard centers and skewed data makes them disagree.

The three candidates are familiar. The mean is the arithmetic average: the sum of the values divided by their count. The median is the middle value: the point with half the observations below and half above. The mode is the most frequent value, or for binned data the tallest bar — the peak of the distribution. On many symmetric, unimodal distributions the three sit close together or coincide, and the choice among them is nearly cosmetic. Marketing data is rarely symmetric. As Section 4.9.1 established, quantities such as order value, customer spend, and engagement counts are chronically right-skewed: most values are modest, and a long tail of large values stretches to the right. On such a shape the three centers pull apart, and on the right-skewed distributions marketing analysts meet most often they fall in a familiar order — the mode lowest at the peak, the median above it, and the mean highest, dragged toward the tail. That ordering is a dependable working heuristic rather than a theorem; the exact arrangement is not guaranteed for every distribution, and what does hold wherever the mean exceeds the median is the consequence that matters: each center is now answering a different question.

DEFINITION

Central Tendency

Measures of central tendency summarize a numeric variable with a single representative value: the mean (the arithmetic average, sensitive to every value and therefore to extreme ones), the median (the middle observation, resistant to extreme values), and the mode (the most frequent value or tallest bin, the only center defined for categorical data). On skewed distributions the three diverge, and the choice among them is an analytic decision, not a formatting preference.

Source: Adapted from Freedman et al. (2007).

The choice discipline is best stated as questions. If the business question concerns totals — revenue, budget, capacity — the mean is the right center, because the mean is the total in disguise: mean order value times order count is total revenue, exactly, which is why finance thinks in means. If the question concerns the typical case — what does a normal order look like, what will the typical customer spend — the median serves better on skewed data, because it refuses to let a handful of large baskets speak for the many small ones. And if the question concerns the most common case — the price point most units actually transact at, the most frequent basket size — the mode answers it, and is additionally the only center available for categorical variables like channel or loyalty tier. Table 5.2 summarizes the mapping. The

professional habit this guide asks for is stricter than choosing well: on skewed data, report mean and median together, because the gap between them is itself descriptive information — it tells the reader, in one comparison, that a tail exists and how hard it pulls.

Table 5.2

Choosing the measure of center

The question being asked	Right center	Why
Totals: revenue, budget, capacity	Mean	The mean times the count reproduces the total exactly
The typical case on skewed data	Median	Resistant to the long tail; speaks for the middle observation
The most common case; categorical data	Mode	The peak of the distribution; the only center for categories
Skewed data, any audience	Mean and median together	Their gap discloses the tail's pull

One computational trap deserves a boxed warning of its own, because it recurs from Chapter 1's exercises to the executive pre-read: the averaged average. The mean of group means is not, in general, the overall mean; the overall mean is the weighted mean, each group's mean weighted by its size. Averaging StyleCraft's per-store average order values gives every store an equal vote regardless of how many orders it processed, and the resulting figure reconciles to no total at all. The pooled figure — total revenue divided by total orders — is the one that adds up. Exercise 1.10, item 2, “The Averaged Average,” previewed this trap on conversion rates (the simple average of four campaign rates versus the pooled rate); it returns here as a standing rule, and it returns once more in this chapter's labs as a verification check, because the equality “weighted recombination of the parts equals the certified whole” is exactly the kind of known-totals check this chapter's verification theme is built on.

In other words, the average is not a neutral summary. It is a choice among centers, made against a shape the analyst is obliged to have looked at (Section 4.9.1), reported with its companion center when the shape is skewed, and recombined by weights, never by simple averaging, when groups are involved. What the average deliberately discards — how much the values vary around it — is the subject of the next section, and discarding it is not always acceptable.

5.4 Spread Is a Finding: Range, IQR, Standard Deviation, and Heterogeneity

Section 5.3 compressed a variable to one number. This section argues that the compression loses something a decision-maker often needs, and introduces the measures that recover it: measures of spread, or dispersion. The section's larger claim is cultural as much as technical — in

marketing data, heterogeneity is not noise around the real answer; heterogeneity frequently is the answer.

Three measures cover the practical ground. The range — maximum minus minimum — is the bluntest: easy to state, ruined by a single extreme value, and mostly useful in inspection (a range check is how Chapter 4 caught the out-of-band prices). The interquartile range (IQR) — the distance between the 25th and 75th percentiles — describes the width of the middle half of the data and inherits the median's resistance to extremes; readers of Chapter 4 have already used it structurally, since the IQR is the box in a box plot and the basis of the outlier fences of Section 4.5. The standard deviation (SD) summarizes the typical distance of observations from the mean; it uses every value, which makes it informative on well-behaved data and tail-sensitive on skewed data, where it should be read alongside the IQR rather than instead of it.

DEFINITION

Dispersion

Measures of dispersion summarize how much a numeric variable varies: the range (maximum minus minimum, maximally sensitive to extremes), the interquartile range (the width of the middle 50 percent of observations, resistant to extremes), and the standard deviation (the typical distance from the mean, computed from every observation). A center reported without a companion measure of spread is an incomplete description.

Source: Adapted from Freedman et al. (2007) and Tukey (1977).

Why insist on spread in a managerial document? Because two groups with identical means can demand opposite marketing treatments. Consider two stores, each with a mean order value of sixty dollars. In the first, orders cluster tightly between fifty and seventy: one customer story, one merchandising story, one number that genuinely represents its orders. In the second, the same mean arises from many thirty-dollar baskets and a steady stream of ninety-dollar occasionwear baskets: two customer stories sharing a checkout line, and a mean that represents neither. Every summary statistic in a pre-read looks like the first store; only the spread — and, for the sharpest cases, the distribution's shape — reveals the second. This is precisely the situation Chapter 4's exploration hinted at for StyleCraft as a whole: a two-humped order-value distribution (Section 4.9.1) is bimodality, and bimodality is the distribution announcing that it describes a mixture of populations. A mean computed across a mixture is an average of nobody.

CONCEPT

Heterogeneity Is a Finding

When a group's members vary widely — in spend, frequency, basket composition, or engagement — that variation is itself a descriptive result, often more decision-relevant than the group's average. Report spread with every center; inspect shape (Section 4.9.1) before trusting any center; and treat wide or two-humped dispersion as a prompt to split the group and describe the parts, not as noise to be averaged away. The question “how much do they vary?” is the question “is this really one group?” wearing statistical clothes.

The practical reporting standard, then, is a small package rather than a single figure: a center chosen for the shape, a second center when skew makes it informative, a spread measure suited to the audience (the IQR travels well because “the middle half of orders fell between X and Y” is plain English), and the group's size — because a mean over forty orders and a mean over forty thousand deserve very different confidence, a point Section 5.9 returns to under benchmarking. This package is what Section 5.6 will assemble into profiles.

Finally, the forward significance of this section should be named, because it shapes the whole book. Heterogeneity as a finding is the descriptive doorway to Part II: if StyleCraft's customers are not one population, the natural next question is which populations they are, and answering it formally is segmentation — Chapter 6's subject. This chapter stays descriptive: it will split groups by known dimensions (metro, channel, tier) and describe the parts. Discovering unknown groups is next month's work; noticing that the average hides them is this month's.

5.5 Cross-Tabulations: Two Variables, Three Percentages, and Simpson's Paradox

The tools so far describe one variable at a time. Marketing questions are almost never about one variable: they are about relationships between categories — channel by metro, tier by age band, repeat purchase by store type. The descriptive instrument for two categorical variables is the cross-tabulation, and it is this chapter's structural core, because both the power and the danger of the expansion pre-read live inside it.

A cross-tabulation (or contingency table) counts observations for every combination of two categorical variables: rows for the values of one, columns for the values of the other, a count in each cell, and totals on the margins. Table 5.3 shows an illustrative extract crossing customer home metro against primary channel. As with the frequency table, the counts are only the beginning; the analytical content arrives when counts become percentages — and a cross-tabulation supports three different percentage bases, which support three different claims. Take the single cell holding the 540 suburban store-primary customers. Cell percentages divide by the grand total: 540 ÷ 2,000, so 27.0 percent of all customers are suburban store-primary. Row percentages divide by the row total: 540 ÷ 800, so 67.5 percent of suburban customers are store-primary. Column percentages divide by the column total: 540 ÷ 840, so 64.3 percent of store-

primary customers are suburban. One cell, three correct percentages, three different sentences — and reading one as another is among the most common table-reading errors in professional life.

Table 5.3

A cross-tabulation read three ways (illustrative extract, 2,000 customers)

Home metro	Online/App primary	Store primary	Row total
NYC Urban	900	300	1,200
NYC Suburban	260	540	800
Column total	1,160	840	2,000

Which percentage should be reported? The classic guidance is directional: percentage in the direction of the explanatory variable — compute percentages within the groups you are comparing, so that the comparison reads across groups on a common base (Zeisel, 1985). To ask "do suburban customers shop differently from urban customers?" is to compare metros, so the percentages run within each metro row: 75.0 percent of urban customers are online/app-primary against 32.5 percent of suburban customers, and the behavioral contrast is stated cleanly. The column percentages answer a different, equally legitimate question — "who is the store channel serving?" — and belong to conversations about channel investment, not customer behavior. The discipline is to decide which question is being asked, percentage in that direction, and label the table so a reader cannot mistake the base: a cross-tabulation without labeled percentage bases is the denominator problem of Section 5.2 in two dimensions.

Now the danger. Because a cross-tabulation compares rates across groups, it inherits the deepest trap in descriptive comparison: the groups being compared may differ in composition, and composition can reverse a comparison outright. Section 3.3.1 established the general principle — aggregation changes the claim, and a relationship at one level of analysis need not hold at another. Simpson's paradox is that principle's table-level case, and this chapter's obligation, having cited the principle, is to compute it (Simpson, 1951; Blyth, 1972).

DEFINITION

Simpson's Paradox

Simpson's paradox is the reversal or disappearance of an association between two variables when the data are disaggregated by a third: a group can show the higher rate in every stratum yet the lower rate in the aggregate, because the groups are unevenly distributed across strata with very different base rates. The paradox is arithmetic, not error — the aggregate and the strata are all computed correctly — which is why it is detected by examining composition, not by rechecking sums.

Source: Adapted from Simpson (1951) and Blyth (1972).

Work the arithmetic on numbers built for clarity. Suppose an analyst compares repeat-purchase rates by primary channel and finds, in aggregate, that store-primary customers repeat at 42.5 percent (357 of 840) while online-primary customers repeat at 52.2 percent (605 of 1,160) — apparently a 9.7-point indictment of the store experience. Table 5.4 disaggregates the same two thousand customers by home metro, and the indictment reverses: within the urban metro, store-primary customers repeat at 65 percent versus 60 percent online; within the suburban metro, 30 percent versus 25 percent. Store-primary customers repeat at a higher rate in every metro and a lower rate overall. The resolution is composition: store-primary customers are concentrated in the suburban metro, where everyone — online or store — repeats far less, so the store column inherits the suburban base rate. The aggregate comparison is not wrong arithmetic; it is a channel comparison contaminated by a metro comparison, and the contaminant is exactly the designed StyleCraft tension this book keeps meeting: the suburban expansion attracts customers who buy richly and return rarely.

Table 5.4

Simpson's paradox in repeat-purchase rates (illustrative extract, 2,000 customers)

Group	Online/App primary	Store primary
NYC Urban: repeaters / customers	540 / 900 (60.0%)	195 / 300 (65.0%)
NYC Suburban: repeaters / customers	65 / 260 (25.0%)	162 / 540 (30.0%)
Aggregate: repeaters / customers	605 / 1,160 (52.2%)	357 / 840 (42.5%)

Note what the aggregate row does and does not say. The 9.7-point aggregate gap favors the online channel; every stratum favors the store channel by five points. Nothing was computed wrongly in either direction, and an analyst who reports only one of the two rows has chosen a conclusion rather than described a table.

Three working rules follow. First, before trusting any comparison of rates across groups, check the composition of the groups on the variables that plausibly drive the rate — a one-line cross-tab of group by stratum is usually enough to see the imbalance. Second, when composition differs, report the stratified comparison and say why: “within each metro, store-primary customers repeat slightly more; the aggregate reverses because store customers skew suburban” is one sentence and it is the honest one. Third, resist the promotion of either table to a causal story: neither the aggregate nor the strata say that the store channel causes anything — confounding and its formal treatment are Chapter 7's subject, and the descriptive analyst's job ends at an accurate, composition-aware account of what the rates are. Where does that leave the pre-read? Every headline comparison in it — suburban versus urban, new stores versus old — pools groups of different composition, and Section 5.8's cohort tables are the standing defense. First, though, the comparison toolkit needs its general form: profiling groups on many descriptives at once.

5.6 Group Comparisons and Customer Profiling

The cross-tabulation compares groups on one categorical outcome. Real stakeholder questions ask for more: not just “do suburban customers repeat less” but “who are these customers?” The descriptive answer is the profile — and this section's discipline is that groups are compared on distributions, not on lone averages.

A group comparison on a numeric variable should, by Sections 5.3 and 5.4, travel as a package: each group's size, a shape-appropriate center (mean and median together when skewed), and a spread measure. Side-by-side five-number summaries — minimum, first quartile, median, third quartile, maximum — do this compactly, and their visual form is the grouped box plot whose reading Chapter 4 established (Section 4.9.2); this chapter adds only the comparative habit of putting the groups on one axis and reading widths and overlaps, not just medians. Grouped box plots disclose differences in medians, in spread, in overlap, and in the tails, and those differences are how an analyst identifies the comparisons worth investigating. What overlap alone does not establish is whether a difference is statistically reliable or operationally important. Two groups whose boxes barely overlap are a strong lead, not a proven distinction; two groups whose means differ while their boxes coincide almost entirely may matter to finance and not at all to a campaign planner deciding whom to treat differently. The picture nominates comparisons; the decision context, and eventually the evidence standards of Chapters 7 and 11, adjudicate them.

A customer profile assembles such comparisons into one table: rows for the groups being compared, columns for a chosen set of descriptives, every cell computed at a stated grain. Table 5.5 sketches the form for the expansion review's central comparison. Note the grain discipline the table forces: order-level facts (average order value, units per order) and customer-level facts (orders per customer, repeat rate, revenue per customer) coexist in one profile, so each column must declare its unit of analysis, and building the customer-level columns requires exactly the deliberate grain change — transactions grouped to customers — that Section 3.3 defined and Chapter 4's lab practiced (Section 4.8).

Table 5.5

A customer profile at two grains (structure; values computed in Lab 5.2)

Descriptive (grain)	NYC Urban	NYC Suburban
Customers (customer)	—	—
Orders per customer, mean / median (customer)	—	—
Repeat-purchase rate, headline definition per Section 3.10.3 (customer)	—	—
Average order value (order)	—	—
Units per order, mean (order)	—	—
Occasionwear share of units (order line)	—	—
Revenue per customer, mean / median (customer)	—	—

CONCEPT

A Profile Is an Argument

A profile is not a data dump; it is a curated set of comparisons chosen to answer the analytic specification (Section 2.5). Every row earns its place by bearing on the decision; every cell declares its grain and denominator; skewed columns carry median beside mean; and group sizes appear first, because every other number in the table is only as steady as the count behind it. Two analysts profiling the same groups for different decisions should produce different tables — and both should reconcile to the same certified totals.

The profile's power is that it converts an adjective into arithmetic. “The suburban shopper is older, more affluent, less frequent, and occasion-driven” is the strategy deck's sentence; the profile is where each clause acquires a number, a denominator, and a spread — or fails to. Its risk is the same as every comparison in this chapter: a profile of pooled groups inherits every composition problem of Section 5.5, so profiles of store groups with different opening dates or customer vintages must either stratify (profile within cohorts, per Section 5.8) or disclose the imbalance in a note. The labs enforce both habits. With the comparison machinery built, the chapter now turns to the specific quantities marketing profiles are made of — the applied metric families — and to the identity that finally answers the CFO's question.

5.7 The Applied Metric Families

The descriptive machinery of Sections 5.2 through 5.6 is general. Marketing runs it over a recurring repertoire of quantities, and fluency in that repertoire — what the standard metrics are, what denominators define them, and how they compose — is part of the job description (Farris et

al., 2010). This section organizes the repertoire into five families and introduces the decomposition identity, the chapter's central analytic move. One boundary first: this section describes the families and their watchpoints; the complete formula catalog lives in Appendix B and only there (Section 3.8.1's rule that metrics belong in the dictionary applies to the guide itself), and every rate below is a ratio metric owing the full four-element definition of Section 3.5.1.

The revenue family summarizes what customers spend. Its workhorses are average order value (AOV) — total revenue divided by order count, an order-grain mean — and revenue per customer, a customer-grain mean over a stated window. The two are routinely conflated and must not be: AOV knows nothing about frequency, so a store can lead on AOV and trail badly on revenue per customer, which is precisely the CRM manager's objection in the opening case. The acquisition family prices the top of the relationship: customer acquisition cost (CAC) — acquisition spend divided by customers acquired, with the window and the definition of "acquired" doing heavy work — and cost per click (CPC), spend divided by clicks. The conversion family describes movement through stages: click-through rate (whose construction Chapter 1 previewed in Section 1.10), and the stage-to-stage rates Section 5.8 will structure. The retention family describes continuation: the repeat-purchase rate — whose three-definitions dispute Chapter 3 staged and settled with a headline definition (Section 3.10.3) — and its complement in spirit, the descriptive churn rate, the share of a stated customer population inactive by a stated rule over a stated window. Churn here is purely descriptive bookkeeping: predicting which individual customers will churn is Chapter 9's subject, and nothing in this chapter models anything. Finally, the engagement family covers the permission channels: open rate and click rate for email and SMS, where the denominator problem is at its most treacherous (sent, delivered, and unique-opened denominators all circulate) and where a boxed reminder from Section 5.2 applies with full force.

Table 5.6

The applied metric families (formulas cataloged in Appendix B)

Family	Representative metrics	Grain of the claim	Denominator watchpoint
Revenue	AOV; revenue per customer	Order; customer	Per order and per customer differ by frequency — never interchange them
Acquisition	CAC; CPC	Cohort of acquired customers; click	What counts as "acquired," and over what window
Conversion	CTR (Section 1.10); stage rates	Impression; stage population	Rate of the step versus rate from the top, and whether the stages share a unit (Section 5.8)
Retention	Repeat-purchase rate (Section 3.10.3); descriptive churn rate	Customer	Which customers the denominator admits, and the window
Engagement	Open rate; click rate	Message or recipient	Sent versus delivered versus unique — name the base

Now the identity. Many headline metrics are composites — products of simpler factors — and a composite can be decomposed exactly: split into factors that multiply back to the original, so that a gap in the headline becomes the product of gaps in its factors. The decomposition is multiplicative, and saying so plainly prevents the most common misuse of it: the identity yields *factors*, not additive percentage shares. A sentence of the form "basket size accounts for sixty percent of the gap" claims an allocation the identity does not supply, and supplying one requires an additional convention (a logarithmic contribution decomposition, for instance) that this chapter does not need. For the review's number, the identity is short. Revenue over orders equals (units over orders) times (revenue over units):

AOV = units per order × revenue per unit.

Average order value is basket size times average realized price per unit — an identity, true by construction, at every grain and in every group. And the second factor decomposes once more, because average price per unit in a catalog with stable list prices is driven largely by mix: revenue per unit equals the sum, across product categories, of each category's share of units times its average realized unit price. Together the two steps answer the CFO's question in the only way arithmetic can: if suburban AOV exceeds urban AOV while list prices are consistent across stores, the difference must arise from basket size, realized revenue per unit, or both — and the decomposition states the factor each contributes, exactly. The realized-price factor may in turn reflect product and category mix (occasionwear at the top of the $24-to-$90 band), discount depth, promotional participation, or, depending on how revenue is defined, returns and adjustments; which of those is doing the work is a second question, and Section 5.11 shows how to ask it. The extension to the CRM manager's question is one more factor: revenue per customer

equals AOV times orders per customer, which is where a weak repeat rate quietly taxes a strong basket.

> **CONCEPT**
>
> **The Decomposition Identity and the Known-Totals Check**
>
> A composite metric is decomposed by writing it as a product of factors that cancel exactly (revenue/orders = units/orders × revenue/units), computing each factor per group, and verifying two equalities before any interpretation: the factors must multiply back to the composite in every group, and the group summaries must recombine — weighted, never simply averaged (Section 5.3) — to the certified totals of Chapter 4's verification log. A decomposition that does not multiply back has a computation error; a summary that does not recombine to the certified total has a silent filter.
>
> Report factors, not shares. The identity produces multiplicative factors whose product is the composite; it does not apportion the gap into percentages, and a percentage claim requires a stated allocation convention the identity alone does not provide. This is the chapter's expression of predict-then-verify (Section 1.7): the answers are checkable, so check them.

The identity does one more quiet service: it disciplines interpretation. “Suburban AOV is higher because suburban customers are wealthier” is a story; “suburban AOV is 2.5 times urban AOV, factoring into a 1.67× basket-size factor and a 1.50× realized-price-per-unit factor, with the price factor tracking category mix rather than list prices” is a description — auditable, decomposable, and silent about causes it cannot see. The labs compute exactly this on StyleCraft's certified data, where the decomposition is a designed finding the dataset ships with, so students can grade themselves. Before the labs, two more descriptive structures complete the toolkit: description along time, and description along the journey.

5.8 Cohorts and Funnels: Description Along Time and Along the Journey

Sections 5.5 and 5.6 warned repeatedly that pooled comparisons inherit composition problems. This section builds the two descriptive structures that manage composition by construction: the cohort, which holds customer vintage fixed, and the funnel, which holds the journey's stages in order.

Begin with the cohort. Comparisons of retention across stores or channels are chronically contaminated by tenure: a store opened eight months ago cannot have two-year-loyal customers, so pooling its customers with a flagship's and comparing repeat rates compares store quality and customer age at once — the Simpson's structure of Section 5.5 with time as the lurking stratum. The fix is to compare like vintages with like.

DEFINITION

Cohort

A cohort is a group of units defined by a shared starting event in a shared time window — for StyleCraft, typically all customers whose first purchase (or signup) fell in a given month or quarter. Cohort analysis follows each cohort forward from its own starting point, so that groups are compared at the same age rather than at the same calendar date.

Aligning on age removes tenure from the comparison and makes vintage differences visible. It does not make the comparison clean. Two cohorts reach a given age in different calendar months, so seasonality, campaigns, pricing changes, and economic conditions can still differ between them, and cohorts acquired through different channels differ in composition from the start. Cohort alignment reduces tenure-driven mix bias; it does not eliminate period or compositional effects.

Source: Adapted from Glenn (2005).

The cohort's standard descriptive artifact is the cohort retention table: one row per cohort, one column per period since the starting event, and in each cell a measure of how much of the cohort is still there. Two structures answer to that description, and a table that does not say which it is cannot be read. In a *period-active* table, a cell holds the share of the cohort that was active — by a stated activity rule — during that individual period; because a customer can be inactive in Month 2 and return in Month 3, a period-active rate can rise as well as fall. In a *survival retention* table, a cell holds the share retained continuously through that period under a stated rule, and survival rates are non-increasing by construction. This chapter, and Lab 5.2, build the period-active table; the activity rule that defines “active in this period” is a measurement decision owing a data dictionary entry per Section 3.8, and naming the structure is part of that entry. Table 5.7 shows the period-active form with illustrative values.

Reading it is a learned skill worth two minutes. Reading along a row shows one cohort aging — period-active shares usually decline with age, though reactivation can lift a later month above an earlier one, which is a finding rather than an error. Reading down a column compares cohorts at the same age: a more comparable, same-age description, subject to the period and compositional differences that alignment does not remove, and exactly the “same-age comparison” the Chapter 2 memo committed to. Reading down the left edge watches cohort sizes, because a triumphant retention story over shrinking cohorts is two findings, not one. The lower-right of the table is empty by construction — young cohorts have not lived long enough to fill it — and the empty triangle is a feature: it displays, rather than hides, which comparisons the data can support.

Table 5.7

A period-active cohort retention table (illustrative values; share of cohort active in month since first purchase)

First-purchase cohort	Size	Month 0	Month 1	Month 2	Month 3
2026 Q1	620	100%	34%	27%	24%
2026 Q2	585	100%	32%	26%	—
2026 Q3	710	100%	30%	—	—
2026 Q4	655	100%	—	—	—

The funnel manages a different composition — the ordered stages of a journey. Descriptively, a funnel is a sequence of stage populations with counts at each stage and rates between them: impressions, clicks, sessions, carts, orders in an acquisition funnel; sent, delivered, opened, clicked in an email funnel. Its intellectual ancestry is the hierarchy-of-effects tradition in advertising — the idea that customers pass through ordered stages from attention to action (Strong, 1925; Lavidge & Steiner, 1961) — but this chapter needs it only as descriptive structure.

DEFINITION

Conversion Funnel

A conversion funnel is a descriptive structure that arranges a customer journey as an ordered sequence of stages, records the count of units reaching each stage, and reports conversion as rates between stages. Each stage's membership rule is an operational definition (Section 3.2) that must be stated, and every funnel rate must declare whether it is step-to-step (this stage over the previous stage) or cumulative (this stage over the top of the funnel).

A funnel's stages are strictly nested — each stage a subset of the one before — only when every stage counts the same unit, such as unique users or sessions. A sequence that mixes units, as impressions, clicks, and orders do, is a stage sequence rather than a set of nested populations, and must be labeled as such.

Source: Adapted from Strong (1925) and Lavidge and Steiner (1961).

Three disciplines make funnels trustworthy. The first is the definition discipline: stage boundaries are measurement decisions — what counts as a session, a cart, an “opened” email — and moving a boundary moves every rate downstream of it, which is why contested funnels are almost always definition disputes rather than behavior changes (a pattern Section 5.13 revisits in industry). The second is the denominator discipline in its funnel form: a “checkout conversion rate of 60 percent” (step) and a “visit-to-order conversion of 3 percent” (cumulative) can

describe the same funnel, and a reader told one while imagining the other is misinformed by true numbers.

The third is the unit discipline, and it is the one most often skipped. Impressions and clicks are event counts, orders are transaction counts, one person generates many of the first two, and an attributed order may be credited on a view rather than a click. “Impressions → clicks → attributed orders” is therefore a campaign-performance sequence, not a nested population, and the arithmetic that nesting licenses — stage counts must be non-increasing, cumulative rates are products of step rates — does not apply to it. Two designs are defensible. Build the funnel on one consistent unit (unique users or sessions, counting click-attributed orders only) and verify that each stage's count does not exceed the previous stage's; or keep the event metrics separate, report click-through rate, click-to-attributed-order rate, and cost per attributed order as what they are, and state the attribution caveat in a sentence. Lab 5.2 takes the second route, because StyleCraft's campaign data carries impressions, clicks, and spend but no session-level events.

Both cohort tables and funnels, finally, submit to the chapter's verification theme: a cohort row's cells are computed over its stated size, and stage counts must reconcile to the certified tables they came from. Where a funnel is genuinely nested, its stage counts must also be non-increasing. Checking that attributed orders do not exceed total orders is a necessary reconciliation, not a proof of nesting. Description along time and along the journey established, one section of context-setting remains before the toolkit goes to work.

5.9 Concentration, Benchmarks, and the Insight Statement

The toolkit can now count, center, spread, cross-tabulate, profile, decompose, and follow cohorts and funnels. This closing conceptual section adds the three finishing disciplines that turn competent description into usable insight: measuring concentration, supplying context, and writing the sentence.

Concentration first. Marketing quantities are not just skewed; they are concentrated — a minority of customers typically accounts for a majority of revenue, a pattern documented in consumer research for six decades as the “heavy half” (Twedt, 1964) and studied formally since (Schmittlein et al., 1993). The descriptive instrument is simple: rank customers by revenue over a stated window, cut the ranking into deciles (ten equal-count groups), and report each decile's share of total revenue. The headline is usually the top decile's share, the popular shorthand is the Pareto ratio (“the top 20 percent account for X percent”), and the working caution comes from the research: the true split is rarely the folkloric 80/20, it varies meaningfully by category and window, and measured concentration is inflated by observation error over short windows — so the window belongs in the sentence (Schmittlein et al., 1993).

DEFINITION

Concentration Analysis

Concentration analysis describes how unevenly a total — revenue, orders, engagement — is distributed across the units that produce it, typically by ranking units, cutting them into deciles or percentiles, and reporting each group's share of the total. Its summary claims (top-decile share; the "heavy half"; Pareto ratios) are window-dependent and must state the ranking variable, the window, and the population ranked.

Source: Adapted from Twedt (1964) and Schmittlein et al. (1993).

Concentration matters to the expansion review for a specific reason: it converts the heterogeneity finding of Section 5.4 into money. If the top decile of customers carries a large share of revenue, then the question "which customers does the suburban expansion acquire?" is not answered by average anything — it is answered by where those customers land in the concentration ranking, and by whether their weak repeat behavior keeps them there. A decile table is one groupby away and reframes the entire meeting.

Context second. A number in isolation is not yet a claim; a number becomes a claim when placed against a benchmark — a reference value that supplies the comparison the reader will otherwise invent. Benchmarks come in three practical grades. Internal-historical (this quarter against last quarter, this cohort against the prior cohort at the same age) is the most trustworthy, because definitions and measurement are held constant. Internal-cross-sectional (this store against sister stores, this campaign against the campaign portfolio) is nearly as strong, subject to the composition cautions of Section 5.5. External (industry reports, published category rates) is the weakest and the most seductive: external benchmarks almost never share your metric's four elements — numerator, denominator, window, filters (Section 3.5.1) — so a gap against them is as often a definition gap as a performance gap, and the honest use is directional context, clearly labeled. Benchmarking is this chapter's descriptive cousin of the baseline discipline Chapter 2 introduced for interpretation (Section 2.7): both exist because a number's meaning is a comparison, and the analyst's job is to make the comparison explicit and fair rather than leave it to imagination.

Third, the sentence. The difference between a competent analyst and a trusted one is frequently one sentence long: the difference between an observation and an insight statement. An observation states what the data shows: "suburban AOV is 2.5 times urban AOV." An insight statement adds the three things a decision-maker needs and description can honestly supply: the decomposition or comparison that explains the number's composition, the magnitude in decision-relevant units, and the implication for the named decision — while stopping, deliberately, short of causal language the description cannot support. "Suburban orders are 2.5 times urban orders in value, factoring into a 1.67× basket-size factor and a 1.50× price-per-unit factor that tracks occasionwear mix rather than list prices — but suburban customers order far less often, so their revenue per customer leads by much less than the AOV headline suggests, and the expansion case should be argued on revenue per customer, not AOV" is an insight statement. It is longer

than the observation, every clause is checkable against a table in this chapter, and nothing in it claims to know why.

> **CONCEPT**
>
> **Insight Statements versus Observations**
>
> An observation reports a number; an insight statement equips a decision. An insight statement (1) names the comparison or decomposition that gives the number meaning, (2) states magnitude in the decision's units, (3) connects to the named decision from the analytic specification (Section 2.5), and (4) stays inside descriptive language — "is associated with," "accounts for," "coincides with" — leaving causal verbs to the evidence that can carry them (Chapters 7 and 11 build that evidence). If a claim's verb asserts more than the tables show, it is not insight; it is speculation with a citation to itself.

With the conceptual sequence complete — precise counting, honest centers and spreads, composition-aware comparison, the metric families and the decomposition identity, cohorts, funnels, concentration, benchmarks, and the insight sentence — the chapter turns to the collaborator who will draft most of this work's code in practice, and to the audit that collaboration requires.

5.10 AI as a Descriptive Analytics Assistant

Descriptive analytics is the domain where AI assistance is at its most fluent and, for exactly that reason, at its most quietly dangerous. Every tool in this chapter — frequency tables, groupbys, cross-tabs, pivot tables, cohort matrices, decile cuts — is one short prompt away from working pandas code, and the code will very likely be correct. The risk has moved: in descriptive work, assistants rarely fail at computation; they fail at the analytic decisions hiding inside the computation, and they fail silently, because the output is a tidy table that looks finished.

Where the assistant genuinely helps, use it. Drafting the mechanical layer — groupby-aggregate scaffolds (whose mechanics Chapter 4's lab established), crosstab and pivot calls, cohort-table reshaping, formatting for presentation — is legitimate delegation that compresses hours into minutes. Asking an assistant to propose which descriptive cuts might illuminate a question is a useful brainstorm, provided proposals are treated as candidates for the specification, not as analysis. And a first-pass narration of a finished, verified table can accelerate writing, provided every sentence survives the insight-statement discipline of Section 5.9.

What must never be delegated are the decisions this chapter spent nine sections on, because they are judgment calls wearing default settings. The characteristic failure modes are worth naming individually. The silent mean: asked to "summarize order value," assistants overwhelmingly reach for the mean — on data Section 4.9.1 already showed is skewed — and report a center the median contradicts. The invented denominator: asked for a "repeat rate" or "open rate," the assistant picks a denominator, does not say which, and the four-element discipline of Section 3.5.1 dies in a code comment. The averaged average: asked to summarize

by group and then overall, assistants routinely average the group means — the exact trap of Section 5.3 — producing an overall figure that reconciles to nothing. The pooled comparison: asked to compare groups, assistants compare aggregates with no composition check, walking directly into Section 5.5's paradox. The unreconciled table: filtered summaries whose totals no longer match the certified file, with the filter mentioned nowhere. And the narrative overreach: asked to "explain" a table, assistants supply causal stories — customers "because," stores "driving" — converting description into speculation at exactly the step stakeholders will quote.

AI IN PRACTICE

The Summary That Must Add Up

A productive pattern for delegated description: paste the relevant data dictionary entries and the certified totals from the Chapter 4 verification log (row count, distinct orders and customers, total revenue — never confidential data; the synthetic StyleCraft files are safe), then prompt — "Write pandas code to produce [one specific table]. Before the code, state: the unit of analysis of each output row; the exact denominator of every rate or proportion; which measure of center you will use and why, given that order value is right-skewed; and the reconciliation I should expect — which totals of your table must equal the certified totals above. After the code, print those reconciliation figures." Then audit in three moves: recompute one cell by hand from the certified table; check the printed totals against the log; and strike every causal verb from any narration before it travels. The predictions come first, per Section 1.7's predict-then-verify, and the whole exchange is documented per Appendix D — the assistant drafted the table, but the analyst of record certified that it adds up.

The deeper point extends this guide's standing division of labor into new territory. In Chapter 4 the assistant drafted cleaning code and the analyst audited the counts; here the assistant drafts summaries and the analyst audits the claims — denominators, centers, compositions, reconciliations, and verbs. Appendix C provides prompt templates for the common descriptive tasks; Appendix D's documentation requirement applies to every exchange; and the labs that follow are built so that the audit has teeth, because on StyleCraft's designed data the right answers are known.

5.11 Hands-On Application in Python and Google Colab

The preceding sections built the toolkit conceptually. This section runs it against the opening case. Lab 5.1 performs the chapter's central move — decomposing the suburban-versus-urban average-order-value gap — twice: first on an eight-order miniature constructed in the notebook, so every figure can be verified by hand, and then on StyleCraft's full certified data, where the decomposition recovers a finding the dataset was designed to contain. Lab 5.2 builds the review's remaining descriptive artifacts: profiles, a cohort retention table, a concentration analysis, and campaign funnel metrics. Throughout, follow the course division of labor: AI assistants may draft the code (Appendix C has prompting templates; Appendix A covers Colab mechanics), but

every summary is predicted before it is computed, reconciled to the certified totals afterward, and every AI exchange is documented per Appendix D.

5.11.1 Lab 5.1, Part A: The Decomposition by Hand

The miniature below contains eight orders — six urban, two suburban — as fourteen order lines, with quantities and prices inside the catalog's $24-to-$90 band and one column, is_occasionwear, carried in from the products dimension. It is built so that the pooled and group summaries are hand-checkable and so that the decomposition identity of Section 5.7 can be verified with a calculator. Before running the first cell, start the predictions: given six urban orders and two suburban ones, will the simple average of the two metro AOVs equal the pooled AOV? (Section 5.3 says it will not, and says which one reconciles to total revenue.)

Code 5.1. Create the hand-checkable miniature

```
import numpy as np
import pandas as pd

lines = pd.DataFrame({
    "order_id":  ["O5001","O5002","O5002","O5003","O5004","O5004",
                  "O5005","O5006","O5006",
                  "O5007","O5007","O5007","O5008","O5008"],
    "metro":     ["NYC Urban"]*9 + ["NYC Suburban"]*5,
    "quantity":  [1]*14,
    "unit_price":[34.0, 36.0, 30.0, 38.0, 28.0, 40.0,
                  80.0, 38.0, 36.0,
                  72.0, 78.0, 36.0, 75.0, 39.0],
    "is_occasionwear": [False, False, False, False, False, False,
                        True,  False, False,
                        True,  True,  False, True,  False]
})
lines["line_revenue"] = lines["quantity"] * lines["unit_price"]

# Part A's certified totals, written as checks rather than as prose.
assert len(lines) == 14
assert lines["order_id"].nunique() == 8
assert np.isclose(lines["line_revenue"].sum(), 660.0)

# metro is carried through the grain change below, so it must be
# constant within every order.
assert (lines.groupby("order_id")["metro"]
             .nunique(dropna=False).le(1).all())

print("lines:", len(lines),
      "| orders:", lines["order_id"].nunique(),
      "| revenue:", lines["line_revenue"].sum())
```

Expected output: 14 lines, 8 orders, revenue 660.0, and four assertions that pass silently.

Input: fourteen literal order lines. Transformation: one derived column, line_revenue. Output: the three certified totals every later summary must reconcile to, in the spirit of Chapter 4's verification log. The assertions are the point of the cell. Chapter 4 argued that a prediction belongs in the log before the step runs; an assertion is that prediction written where the notebook

itself will enforce it, and a printed figure a student forgets to read is not a check. The fourth assertion guards the grain change that follows: carrying metro from lines to orders is honest only if every line in an order shares one metro.

The first computation changes grain deliberately (Section 3.3; mechanics per Section 4.8): order value lives at the order grain, so lines are grouped to orders before any AOV is touched. Predict the output shape first: eight rows, one per order.

Code 5.2. Change grain from lines to orders, then AOV by metro

```
orders = (lines
          .groupby(["order_id", "metro"], as_index=False)
          .agg(order_value=("line_revenue", "sum"),
               units=("quantity", "sum")))

# The grain change must preserve the certified totals exactly.
assert len(orders) == lines["order_id"].nunique()
assert np.isclose(orders["order_value"].sum(),
                  lines["line_revenue"].sum())

aov = (orders
       .groupby("metro")
       .agg(orders=("order_id", "count"),
            revenue=("order_value", "sum"),
            aov=("order_value", "mean"),
            units_per_order=("units", "mean")))

pooled_aov   = orders["order_value"].sum() / len(orders)
weighted_aov = ((aov["aov"] * aov["orders"]).sum()
                / aov["orders"].sum())
simple_aov   = aov["aov"].mean()

# Weighted recombination reconciles to the pooled figure;
# the simple average of group means does not.
assert np.isclose(pooled_aov, weighted_aov)

print(orders)
print(aov)
print("pooled:", round(pooled_aov, 2),
      "| weighted:", round(weighted_aov, 2),
      "| simple average of metro AOVs:", round(simple_aov, 2))
```

Expected output: an eight-row orders frame; urban AOV 60.00 over 6 orders and suburban 150.00 over 2; pooled 82.50, weighted 82.50, and the simple average of the metro AOVs at 105.00.

Input: the fourteen certified lines. Transformation: rows are collected by order (and by the metro that travels with the order), then summarized by metro. Output: the headline gap, plus the averaged-average trap made numerical. The two assertions state the grain-change identities of Section 4.8 — one row per distinct order, and a preserved total — and the third states Section 5.3's standing rule: the weighted recombination equals the pooled figure, while the $105.00 simple average reconciles to no total in the file.

VERIFICATION CHECK

Before you run: write four predictions before running. (1) Reconciliation: the two metros' order counts must sum to 8 and their revenues to 660.0 — the known-totals check of Section 5.7. (2) Urban AOV: urban revenue is 360.0 over 6 orders, so 60.0; suburban is 300.0 over 2 orders, so 150.0. (3) The pooled AOV is 660/8 = 82.50; the simple average of the metro AOVs is (60 + 150)/2 = 105.00 — the averaged-average gap of Section 5.3 — and the weighted recombination (6×60 + 2×150)/8 must return exactly 82.50. (4) Units per order: urban 9/6 = 1.5; suburban 5/2 = 2.5.

After you run: every assertion passes silently, the printed metro table matches predictions (2) and (4), and the three AOV figures read 82.50, 82.50, and 105.00.

Investigate if: any assertion raises. A failed row-count assertion means the grain change dropped or duplicated something; a failed total assertion means a group key contained a missing value; a failed weighted-recombination assertion means the weights are wrong. Find the cause before proceeding.

With the headline gap on screen — suburban AOV of 150 against urban 60, a ratio of 2.5 — the decomposition identity does its work. AOV = units per order × revenue per unit, so the 2.5× gap must factor exactly into a basket factor and a price-per-unit factor.

Code 5.3. Decompose AOV into basket size and revenue per unit

```
decomp = aov.copy()
decomp["units"] = decomp["units_per_order"] * decomp["orders"]
decomp["revenue_per_unit"] = decomp["revenue"] / decomp["units"]
decomp["identity_check"] = (decomp["units_per_order"]
                            * decomp["revenue_per_unit"])

# Verification 1: the factors multiply back in every group.
assert np.allclose(decomp["identity_check"], decomp["aov"])

sub, urb = "NYC Suburban", "NYC Urban"
ratio_aov    = decomp.loc[sub, "aov"] / decomp.loc[urb, "aov"]
ratio_basket = (decomp.loc[sub, "units_per_order"]
                / decomp.loc[urb, "units_per_order"])
ratio_price  = (decomp.loc[sub, "revenue_per_unit"]
                / decomp.loc[urb, "revenue_per_unit"])

# Verification 2: the ratio of composites is the product of
# the ratios of the factors.
assert np.isclose(ratio_aov, ratio_basket * ratio_price)

print(decomp[["aov", "units_per_order",
              "revenue_per_unit", "identity_check"]])
print(round(ratio_aov, 4), "=",
      round(ratio_basket, 4), "x", round(ratio_price, 4))
```

Expected output: revenue per unit of 40.0 urban and 60.0 suburban; identity_check equal to aov in both rows; and the ratio line reading 2.5 = 1.6667 × 1.5.

Input: the metro summary. Transformation: the composite is written as the product of its two factors, group by group. Output: the decomposition and its two verifications. Read the result in

the vocabulary Section 5.7 insisted on: the 2.5× AOV ratio *factors* into a 1.67× basket-size factor and a 1.50× realized-price-per-unit factor. It does not divide the gap into percentage shares, and a sentence claiming that baskets explain "two-thirds of the gap" would be asserting an allocation this identity does not produce.

The last step interrogates the price factor. List prices for the same product do not differ by store, although realized price per unit can still differ because of discounts, product mix, or promotion mix — so a 1.5× gap in revenue per unit is a claim about what was bought, not about what things cost, and the miniature carries the category flag to examine it.

Code 5.4. Examine category mix and realized prices

```
mix = (lines
       .groupby(["metro", "is_occasionwear"])
       .agg(units=("quantity", "sum"),
            revenue=("line_revenue", "sum")))
mix["avg_price"] = mix["revenue"] / mix["units"]

# Occasionwear unit share, computed with explicit columns rather
# than groupby.apply, whose treatment of the grouping columns is
# both version-sensitive and hard for a reader to audit.
share = (lines
         .assign(occasionwear_units=lines["quantity"]
                 .where(lines["is_occasionwear"], 0))
         .groupby("metro")
         .agg(occasionwear_units=("occasionwear_units", "sum"),
              total_units=("quantity", "sum")))
share["occasionwear_unit_share"] = (share["occasionwear_units"]
                                    / share["total_units"])

# Standardized comparison: each metro's own mix, priced at the
# urban metro's within-category average prices.
urban_price = mix.loc["NYC Urban", "avg_price"]
share["revenue_per_unit_at_urban_prices"] = (
    share["occasionwear_unit_share"] * urban_price[True]
    + (1 - share["occasionwear_unit_share"]) * urban_price[False]
)

print(mix.round(2))
print(share[["occasionwear_unit_share",
             "revenue_per_unit_at_urban_prices"]].round(3))
```

Expected output: within-category average prices of $80.00 (occasionwear) and $35.00 (other) in the urban metro against $75.00 and $37.50 in the suburban metro; occasionwear unit shares of 0.111 and 0.600; and standardized revenue per unit of $40.00 urban and $62.00 suburban.

Input: the fourteen lines with their category flag. Transformation: a two-way summary of units, revenue, and realized price; an explicit occasionwear unit share; and one standardization that holds within-category prices constant while letting the mix vary. Output: three views that together locate the price factor.

Read them in order. The metros do not pay materially different prices for the same kind of item — the within-category averages sit a few dollars apart, and in opposite directions — while

the occasionwear share of units differs by more than five times. The standardization makes the comparison exact: buying its own mix at the urban metro's within-category prices, the suburban metro would post revenue per unit of $62.00 against the urban $40.00, a factor of 1.55, slightly *more* than the 1.50 factor actually observed. Mix, in other words, more than accounts for the price factor, and the small within-category price differences work marginally against it.

Be precise about what this establishes. The miniature shows a revenue-per-unit gap strongly associated with category mix, together with small within-category differences in realized price. It does not show that identical products sold at identical realized prices, which would require product identifiers the miniature does not carry — and which Part B, working on the full file with product_id available, can check directly. Section 5.7's identity is exact by construction; this second step is a standardization, and keeping the difference straight is the difference between a decomposition and an explanation.

VERIFICATION CHECK

Before you run: predict that urban units split 1 occasionwear to 8 other and suburban 3 to 2, giving occasionwear unit shares of 0.111 and 0.600. Predict that within-category average prices will be close across metros but not identical. Then predict the recombination: each metro's (occasionwear share × its own occasionwear price + other share × its own other price) must reproduce its revenue per unit — 40.0 and 60.0 — exactly.

After you run: the shares are 0.111 and 0.600; urban within-category prices are $80.00 and $35.00 against suburban $75.00 and $37.50; each metro's recombination returns 40.0 and 60.0; and the standardized figures are $40.00 and $62.00.

Investigate if: a recombination misses its metro's revenue per unit. That means a share or a price was computed over the wrong base: both are unit-weighted, so units — not order lines and not orders — is the denominator for each.

Close Part A the way Chapter 4 closed its miniature: with the certified account. Eight orders and 660.0 in revenue, fully reconciled; a 2.5× AOV gap factored exactly into a 1.67× basket factor and a 1.50× realized-price-per-unit factor; and the price factor shown to track category mix — 60 percent of suburban units are occasionwear against 11 percent of urban, and at common within-category prices that mix alone would produce a 1.55× factor. Every number above is reproducible with a calculator from the fourteen lines, and reproducing them is the point: the full file will not be hand-checkable, so the habits must be certified while it is.

5.11.2 Lab 5.1, Part B: The Designed Finding at Full Scale

Part B runs the same sequence on StyleCraft's certified data — transactions_clean (or your own certified equivalent from Lab 4.1), joined to products for is_occasionwear and to stores for metro, all available from Brightspace and the companion repository. The dataset was designed so that the suburban and resort AOV advantage is produced by basket size and occasionwear mix, not unit price; your decomposition either recovers that designed finding or it does not, and either

outcome is gradable — this is the book's known-truth verification thread at work, the same arrangement that gave Chapter 4's lab an answer key.

Work in the Part A order, with five scale disciplines added. First, open the log: record the certified totals (rows, distinct orders, distinct customers, total revenue) from your Chapter 4 verification log before computing anything, and reconcile every groupby against them. Second, predict directions before magnitudes: before the AOV cell runs, write down which metro group will lead, whether the basket factor or the price factor will contribute more, and roughly where the occasionwear unit shares will land — then compare. Third, use the product identifiers the miniature lacked: for the products that sell in both metros, compare realized price per unit directly, which turns Part A's standardized argument into a direct test of whether the same item transacts at the same price in both places. Fourth, let an AI assistant draft each step per Section 5.10's AI in Practice pattern, requiring its stated denominators and reconciliation figures in advance; at full scale at least one delegated summary will fail a reconciliation or default to the wrong center, and finding it is part of the assignment. Fifth, close with the identity audit: factors multiply back in every group, weighted recombination reproduces the pooled AOV, and within-category average prices sit inside the catalog band in every metro.

VERIFICATION CHECK

Before you run: predict the metro AOV table, the two factors, and the occasionwear unit shares before any cell runs, and write down which of the two factors you expect to be larger.

After you run: the factors multiply back in every metro, the weighted recombination reproduces the pooled AOV, and the same-product realized-price comparison shows no systematic metro difference.

Investigate if: your decomposition table differs from a classmate's in any way not attributable to a stated choice. The designed finding is the answer key, not the assignment: matching it does not excuse a missing reconciliation, and a defensible pipeline that reveals the same structure with slightly different figures (from different but documented join or filter choices) is better work than matching numbers with an unlogged filter. An unexplained difference means a silent step ran somewhere — the Chapter 4 rule, now applied to summaries.

5.11.3 Lab 5.2: Profiles, Cohorts, Concentration, and Campaign Metrics

Lab 5.2 builds the review's remaining artifacts on the certified data, one per part, each opened with a written prediction and closed with a reconciliation. Each part below ships a worked cell that demonstrates the pattern and leaves a stated extension for homework; the cells assume transactions_clean, customers, and campaigns are loaded, that order_date has been parsed to a true date, and that numpy and pandas are imported as in Code 5.1.

Part 1, the profile. Complete Table 5.5: build the customer-grain columns by grouping transactions to customers, the order-grain columns from the Part B orders table, and the line-grain occasionwear share; report mean and median together wherever Section 5.3's skew rule applies, group sizes first per Section 5.6. Code 5.5 builds the customer-grain half of the profile and its reconciliations; the order-grain and line-grain rows are yours to add.

Code 5.5. Build the customer-grain profile

```
# Step 1: transactions to orders, then orders to customers.
orders = (transactions_clean
          .groupby(["order_id", "customer_id"], as_index=False)
          .agg(order_value=("line_revenue", "sum"),
               units=("quantity", "sum"),
               order_date=("order_date", "min")))

assert len(orders) == transactions_clean["order_id"].nunique()
assert np.isclose(orders["order_value"].sum(),
                  transactions_clean["line_revenue"].sum())

cust = (orders
        .groupby("customer_id", as_index=False)
        .agg(orders=("order_id", "count"),
             revenue=("order_value", "sum"),
             units=("units", "sum")))

# The headline repeat-purchase definition of Section 3.10.3:
# two or more orders in the observation window.
cust["is_repeater"] = cust["orders"] >= 2

# Step 2: attach the grouping dimension, cardinality declared.
cust = cust.merge(customers[["customer_id", "home_metro"]],
                  on="customer_id", how="left",
                  validate="many_to_one")
print("customers with no home metro:",
      int(cust["home_metro"].isna().sum()))

# Step 3: the profile. dropna=False keeps any unmatched customers
# visible instead of deleting them from the denominator.
profile = (cust
           .groupby("home_metro", dropna=False)
           .agg(customers=("customer_id", "count"),
                orders_mean=("orders", "mean"),
                orders_median=("orders", "median"),
                repeat_rate=("is_repeater", "mean"),
                revenue_mean=("revenue", "mean"),
                revenue_median=("revenue", "median")))

# Reconciliations: the profile counts every customer and every dollar.
assert profile["customers"].sum() == len(cust)
assert np.isclose((profile["customers"] * profile["revenue_mean"]).sum(),
                  transactions_clean["line_revenue"].sum())

print(profile.round(2))
```

Expected output: a printed count of customers with no home metro (zero on a correctly joined certified file), one profile row per metro, all four assertions passing, and a mean above the median in every revenue column — the skew of Section 4.9.1 arriving at customer grain.

Input: the certified line-grain table and the customers dimension. Transformation: two deliberate grain changes (lines to orders, orders to customers) and one validated join. Output: the customer-grain half of Table 5.5, with group sizes first. The second assertion is the chapter's known-totals check in profile form: the customer counts times the group revenue means

recombine — weighted, never simply averaged — to the certified total. Extension for homework: add the order-grain rows (AOV, units per order) from the Part B orders table and the line-grain occasionwear share, and label each row with its grain as Table 5.5 requires.

Part 2, the cohort table. Define first-purchase monthly cohorts, choose and document an activity rule (the dictionary entry is part of the deliverable, per Section 3.8, and it must name the table's structure), and produce the retention table split urban versus suburban. Read it down the columns: at the same age, how do the metros compare — and does the pooled repeat-rate gap survive the same-age comparison? This part is the same-age cohort analysis the Chapter 2 memo committed to, and its verification is structural: each row's cells are proportions of that row's stated size, and the young-cohort triangle stays empty.

Code 5.6. Construct the period-active cohort retention table

```
# Activity rule (dictionary entry): a customer is ACTIVE in a
# period if the customer placed at least one order in that
# calendar month. This is a PERIOD-ACTIVE table, per Section 5.8,
# so a cell may exceed the cell to its left when customers return.
orders["order_month"] = orders["order_date"].dt.to_period("M")

cohort = (orders.groupby("customer_id")["order_month"]
          .min().rename("cohort"))
orders = orders.join(cohort, on="customer_id")
orders["period"] = [(m - c).n for m, c
                    in zip(orders["order_month"], orders["cohort"])]

size = cohort.value_counts().sort_index().rename("size")

# A complete, contiguous period axis. Without the reindex, a period
# in which no cohort happened to be active would drop out of the
# table altogether and the ages would silently stop being adjacent.
periods = pd.Index(range(int(orders["period"].max()) + 1),
                   name="period")

# Two different absences have to be filled here, and reindex alone
# fills only one of them. unstack leaves NaN for a (cohort, period)
# pair that produced no rows; reindex adds whole periods that no
# cohort reached. fillna after the reindex covers both, so every
# cell is a count before any masking happens.
active = (orders
          .drop_duplicates(["customer_id", "period"])
          .groupby(["cohort", "period"])["customer_id"]
          .count()
          .unstack("period")
          .reindex(index=size.index, columns=periods)
          .fillna(0))

# Zero activity and an unobserved future are different facts and
# must not share a cell value. Elapsed-but-inactive cells are 0.0;
# periods a cohort has not yet lived through are masked to missing.
analysis_month = orders["order_month"].max()

observable = pd.DataFrame(
    [[(cohort_month + p) <= analysis_month for p in periods]
     for cohort_month in size.index],
    index=size.index,
    columns=periods,
)

retention = active.div(size, axis=0).where(observable)

# Structural verifications.
assert (retention[0] == 1.0).all()          # period 0 is the cohort itself
assert size.sum() == orders["customer_id"].nunique()
assert retention.max().max() <= 1.0
assert retention.isna().equals(~observable)  # missing means unobserved

print(size)
print((retention.iloc[:, :7] * 100).round(1))
```

Expected output: a cohort-size column summing to the distinct customer count; a retention table with contiguous period columns whose first column is 100 percent in every row; 0.0 wherever a cohort lived through a period with nobody active; and blanks only in the lower-right triangle no cohort has yet reached.

Input: the order-grain table from Code 5.5. Transformation: each customer is assigned to the month of the first order; each order is placed at its age in months since that cohort month; the count of distinct customers active at each age is divided by the cohort's size; and the result is masked to the periods that have actually elapsed. Output: the period-active table of Table 5.7. The reindex and the mask together carry the section's most easily lost distinction. A blank cell in a cohort table can mean two opposite things — the cohort reached that age and nobody bought, or the cohort has not reached that age at all — and a table that renders both as missing invites the reader to mistake a collapse for an absence of data. Filling elapsed cells with zero and masking only the unobserved future makes the empty triangle mean exactly one thing, which is what licenses reading down a column. The four assertions encode what the structure guarantees: everyone is active at period 0, the cohort sizes partition the customer base, no cell can exceed 100 percent, and a cell is missing if and only if the period has not yet occurred. Note what is deliberately *not* asserted — that each row declines — because in a period-active table a return visit legitimately raises a later month. Extension for homework: split the table by home_metro and read the two versions down the columns.

Part 3, concentration. Rank customers by revenue over the trailing twelve months, cut deciles, and report each decile's share of revenue plus the top-decile and top-two-decile shares, window stated per Section 5.9. Verification: decile revenues must sum to the window's total revenue exactly. Then the review question: cross the top revenue decile against home metro — where do the expansion's customers land?

Code 5.7. Build revenue deciles and concentration shares

```
# The window is part of the claim, so it is declared, not implied.
window_end   = transactions_clean["order_date"].max()
window_start = window_end - pd.DateOffset(months=12)
window = transactions_clean[
    transactions_clean["order_date"] > window_start
]

# Every dollar ranked must belong to a customer, or the deciles
# and the window total will not reconcile.
assert window["customer_id"].notna().all(), \
    "lines without a customer cannot be ranked; resolve them first"

rev = (window
       .groupby("customer_id", as_index=False)
       .agg(revenue=("line_revenue", "sum")))

# Decile 1 is the highest-revenue tenth. rank(method="first")
# breaks ties deterministically so the cut is reproducible.
rev["decile"] = pd.qcut(
    rev["revenue"].rank(method="first", ascending=False),
    10, labels=range(1, 11)
)

conc = (rev.groupby("decile", observed=True)
        .agg(customers=("customer_id", "count"),
             revenue=("revenue", "sum")))
conc["share_of_revenue"] = conc["revenue"] / conc["revenue"].sum()

# The deciles must partition the window exactly.
assert conc["customers"].sum() == len(rev)
assert np.isclose(conc["revenue"].sum(), window["line_revenue"].sum())

print(f"window: after {window_start.date()} "
      f"through {window_end.date()}")
print(conc.round(3))
print("top decile share:",
      round(float(conc["share_of_revenue"].iloc[0]), 3))
print("top two deciles:",
      round(float(conc["share_of_revenue"].iloc[:2].sum()), 3))
```

Expected output: the window printed as the interval it actually is — open at the start, closed at the end — then ten decile rows of roughly equal customer counts with steeply unequal revenue shares, all three assertions passing, and top-decile and top-two-decile shares well above ten and twenty percent respectively.

Input: the certified line-grain table. Transformation: a declared twelve-month window, a customer-grain revenue total inside it — the filter is strictly greater than the start date and inclusive of the end date, so the interval is (window_start, window_end] and the printed label says so — a ranked cut into ten equal-count groups, and each group's share of the window's revenue. Output: the concentration table and the two headline shares. The assertions carry the section's caution: deciles are a partition, so the customer counts and the revenue must both add back to the window, and any figure quoted from this table travels with the window that produced

it (Schmittlein et al., 1993). Extension for homework: cross the top decile against home_metro and against loyalty tier, and repeat the whole cell on a ninety-day window to see how much the headline moves.

Part 4, the campaign metrics. From the campaigns table, compute CTR (per Section 1.10's construction) and CPC per campaign and per channel, labeling every rate step-versus-cumulative and naming denominators; then extend to the campaign-performance sequence of Section 5.8 — impressions, clicks, attributed orders — reporting the click-to-attributed-order rate and cost per attributed order, and noting in one sentence that these are three different units of count rather than a nested funnel. Attribution is itself a definition: campaign_id's nulls carry the two-kinds-of-missing ambiguity documented in Chapter 4.

Code 5.8. Calculate campaign and funnel metrics

```
# The fact table is at line grain, so one order can carry several
# campaign_id values unless the attribution rule forbids it. If it
# can, distinct orders summed across campaigns double-count.
assert (transactions_clean
        .groupby("order_id")["campaign_id"]
        .nunique(dropna=True)
        .le(1)
        .all()), "at least one order carries multiple campaign identifiers"

attributed = (transactions_clean
              .dropna(subset=["campaign_id"])
              .groupby("campaign_id", as_index=False)
              .agg(attributed_orders=("order_id", "nunique"),
                   attributed_revenue=("line_revenue", "sum")))

perf = campaigns.merge(attributed, on="campaign_id",
                       how="left", validate="one_to_one")
perf[["attributed_orders", "attributed_revenue"]] = (
    perf[["attributed_orders", "attributed_revenue"]].fillna(0)
)

def add_rates(frame):
    """Four rates, four denominators, always recomputed from counts.
    A zero denominator becomes missing rather than infinite, so a
    campaign with no clicks reports no rate instead of a record one."""
    imp  = frame["impressions"].replace(0, np.nan)
    clk  = frame["clicks"].replace(0, np.nan)
    ordr = frame["attributed_orders"].replace(0, np.nan)
    frame["ctr"] = frame["clicks"] / imp
    frame["cpc"] = frame["spend"] / clk
    frame["orders_per_click"] = frame["attributed_orders"] / clk
    frame["cost_per_attributed_order"] = frame["spend"] / ordr
    return frame

perf = add_rates(perf)

# Channel roll-up: pool the counts, then run the same function again.
# Rates are recomputed, never averaged across campaigns (Section 5.3).
by_channel = add_rates(
    perf.groupby("channel").agg(
        impressions=("impressions", "sum"),
        clicks=("clicks", "sum"),
        spend=("spend", "sum"),
        attributed_orders=("attributed_orders", "sum")))

# Reconciliations. These check the sequence against its sources;
# they do NOT establish that the stages are nested populations.
assert (perf["attributed_orders"].sum()
        <= transactions_clean["order_id"].nunique())
assert np.isclose(perf["spend"].sum(), campaigns["spend"].sum())
assert (perf["ctr"].dropna() <= 1).all(), \
    "clicks exceed impressions somewhere"
print(perf[["campaign_id", "channel", "impressions", "clicks",
            "spend", "attributed_orders", "ctr", "cpc",
            "orders_per_click",
            "cost_per_attributed_order"]].round(4).head(10))
print(by_channel.round(4))
```

Expected output: a per-campaign table of impressions, clicks, spend, attributed orders and the four rates; a channel roll-up carrying the same four rates recomputed from pooled counts; four assertions passing; and a missing value, rather than an infinity, wherever a campaign or channel produced no clicks or no attributed orders.

Input: the campaigns dimension and the certified transactions. Transformation: attributed orders and revenue are rolled up per campaign, joined one-to-one to the campaign row, and converted into the same four rates at two levels by one shared function, with every zero denominator converted to a missing value first — a rate over nothing is undefined, not infinite, and a table full of inf is a denominator problem wearing a printing error. Output: campaign performance as a sequence of event metrics at two levels. The shared add_rates function is doing more than saving lines: because the channel figures are produced by pooling the counts and running the same function again, the weighted recombination of Section 5.3 is structural rather than advisory — there is no code path in this cell that could average the campaigns' rates, which is exactly the number that would reconcile to nothing. Note the first assertion too: because the fact table is at line grain, an order carrying two campaign identifiers would be counted once under each, and the total-orders reconciliation could still pass while the per-campaign figures double-count. Note finally what the assertions do not claim — attributed orders not exceeding total orders is a reconciliation, not a proof that orders nest inside clicks (Section 5.8). Extension for homework: add a cumulative rate from impressions to attributed orders, state in the notebook why it is not a conversion rate in the nested sense, and report how many attributed orders would be lost under a click-only attribution rule.

Deliverables for both labs: the notebook, the reconciliation log, the completed profile and cohort tables, the data dictionary entries for the activity rule and the attribution rule, and the AI-Use Appendix (Appendix D) documenting every assistant exchange, including at least one delegated summary you corrected and why.

5.12 Marketing Interpretation and Managerial Insight

The labs produced tables; the review needs sentences. This section returns to the meeting about one number and translates the chapter's outputs into decision language — including, per this guide's standing practice, the wrong reading that must be corrected before it circulates.

The headline the analytics lead can now deliver is a decomposition, not a defense or a debunking. The suburban AOV advantage is real, it reconciles to certified totals, and it is made of two measurable things: bigger baskets and a mix that skews to occasionwear at the top of an unchanged price band — with each factor stated exactly, courtesy of Section 5.7's identity. That single sentence converts the pre-read's Rorschach test into an agenda: if the advantage is baskets and mix, then the questions worth the room's time are whether occasion-driven baskets repeat (the cohort table answers at like ages), what the suburban customer is worth per customer rather than per order (the profile's revenue-per-customer row), and where those customers land in the concentration ranking that carries the company's revenue.

Now the wrong managerial reading, because it is already in circulation: “Suburban AOV is 2.5 times urban, so suburban customers are 2.5 times as valuable — resume the openings.” Every clause outruns the tables. AOV is an order-grain mean; customer value is a customer-grain quantity, and the bridge between them is order frequency — revenue per customer equals AOV times orders per customer — which is precisely where the suburban profile is weakest. A group can lead on AOV and trail on revenue per customer, and whether StyleCraft's suburban group does is an empirical cell in Table 5.5, not a matter of emphasis. Nor does even a favorable revenue-per-customer comparison carry the conclusion by itself: the pooled comparison mixes customer vintages (the new stores' customers are structurally young), which is why the memo demanded the same-age cohort comparison, and Section 5.5 showed with two thousand illustrative customers how badly a pooled rate can misrepresent every stratum it contains. The corrected sentence for the room is the insight statement built in Section 5.9 — AOV factored, frequency named, the expansion case re-argued on revenue per customer at like ages — and its final clause matters most: nothing in this chapter says why any of it is so. “The suburban format attracts occasion shoppers” is a hypothesis the description makes plausible; testing whether format causes behavior is Chapter 11's kind of evidence.

One more translation earns its three minutes of meeting time: the definitions underneath the tables. The repeat rate uses the Chapter 3 headline definition; the cohort activity rule, the cohort table's period-active structure, and the attribution rule behind the campaign metrics are documented measurement choices; every rate's denominator is named in its table note. Surfacing this is not pedantry — it is the reason the meeting cannot dissolve into the “one metric, three numbers” dispute that consumed Chapter 3's version of this room, and it is what the analyst of record's signature on a descriptive deliverable means.

5.13 Business Analytics in Practice

This section turns from the fictional case to how the same descriptive disciplines appear in industry — where descriptive analytics is not the remedial tier of the field but its daily, decision-carrying core. The three vignettes below are composite: they synthesize recurring practices in retail, subscription, and digital analytics teams rather than reporting on any one named organization, and they are offered as patterns a new analyst will recognize rather than as prevalence claims about most companies.

5.13.1 The Monday Trading Report

The first vignette is the most ordinary artifact in retail analytics: the Monday trading report. In many retail and e-commerce organizations, the analytical product executives actually read every week is descriptive — a page or two, assembled by an analyst or an automated job overnight into Monday morning, summarizing the trading week. Its conventions are worth studying because they encode this chapter. The reading order is fixed and hierarchical: headline revenue against plan and against last year first (benchmarks before raw numbers, always — a number without its comparison is not yet information); then the decomposition of any gap into traffic, conversion,

and average order value, because executives triage by factor ("we missed on conversion, not on basket"); then exceptions — the stores, categories, or campaigns whose deviation from expectation exceeds a threshold — rather than exhaustive listings. What is deliberately absent is as instructive: no models, no forecasts beyond plan, and very few adjectives.

Behind the page sits a verification routine that looks a great deal like Chapter 4's log. Before the report is released, the week's revenue is reconciled to the finance or order-management system of record; the row counts feeding each summary are compared against the prior week's, because a step change in volume is usually a pipeline failure rather than a demand event; the denominators are re-derived rather than inherited, so that a store closed mid-week does not quietly change what "comparable stores" means; and any definition that moved is annotated on the page itself. Trading analysts describe the job as protecting that page's precision — every rate's denominator stable week to week, every comparison like-for-like (this year's week aligned to the retail calendar, not the civil one), every "up 4 percent" decomposable on demand. The craft looks humble and is not: a trading report whose definitions wobble teaches an executive team to stop trusting numbers on Mondays, and the damage generalizes.

5.13.2 Healthy Averages Over a Deteriorating Base

The second vignette is a pattern subscription and repeat-purchase businesses rediscover regularly: healthy averages over a deteriorating base. A subscription-commerce company — the composite is generic, but the pattern is among the most cited reasons cohort tables exist — tracked overall monthly retention and average revenue per subscriber, and both lines held steady for quarters while the business quietly weakened. The mechanism was composition, exactly Section 5.5's: retention rises with subscriber tenure, so as acquisition slowed, the subscriber base aged, and the growing share of long-tenured, high-retention subscribers propped the overall averages while every new cohort retained worse than the one before it. The pooled metrics were true and reassuring; the cohort retention table — the same rows-by-age artifact as Table 5.7 — showed the down-and-to-the-right deterioration a year before the pooled lines bent.

What such a finding costs the analyst is worth describing, because it is rarely a single chart. Cohort deterioration reaches management as an escalation with a specific shape: the same-age comparison first (this quarter's cohort against the prior four at Month 3), then the reconciliation showing that the pooled metric is unchanged and why the two are not in conflict, then a decomposition of the deterioration into the channels or offers the weak cohorts came from, and only then a recommendation. The order matters because the pooled number is already in the room's memory, and an analyst who leads with "retention is falling" contradicts a dashboard the room trusts. The practice lesson lands on two of this chapter's disciplines at once: averages computed over shifting mixes answer questions nobody asked, and the same-age comparison is not an academic refinement — it is frequently the earliest honest signal a growth business gets.

5.13.3 Who Controls the Funnel's Definitions

The third vignette concerns the funnel — specifically, who controls its definitions. At streaming, e-commerce, and consumer-subscription firms, the conversion funnel is an operational institution: teams are named for stages, dashboards are organized by stage rates, and targets attach to stage boundaries. Which makes the boundaries themselves contested territory. What counts as a "visit" (does a bounce within five seconds count?), an "add to cart," a "trial start" (card required or not?), an "active" account — each is an operational definition in Section 3.2's sense, each moves every rate downstream of it, and each has an owner with a target riding on it.

Mature organizations respond institutionally, and the mechanics are worth knowing before a new analyst is asked to change a definition. Stage definitions live in a governed dictionary with named owners; a proposed change is written as a diff against the current rule, accompanied by a backfill showing the affected rates under both definitions over a trailing window, so the size of the discontinuity is known before it appears; the change is approved by the metric's owner and the teams whose targets move; and on the day it ships, an annotation is written onto every chart that uses the rate, the way a store opening or a site migration would be. Teams that skip the backfill discover the change the way everyone else does — as a performance story that turns out to be an instrumentation story. The lesson reinforces rather than extends the chapter: funnels are powerful precisely because they are descriptive structure imposed on a journey, and the structure is only as trustworthy as the written definitions of its stages — whoever controls those definitions controls the funnel's story, which is Chapter 3's closing thread, operationalized.

5.13.4 In Your First Analyst Job

In your first analyst job, these vignettes compress into an expectation worth internalizing before the job begins. Much of the analytical work organizations actually consume — the trading page, the cohort table, the funnel review, the profile behind a targeting brief — is descriptive, and the analysts who become indispensable are rarely the ones who reached for models fastest. They are the ones whose descriptions could be trusted blind: denominators named, centers matched to shapes, comparisons composition-checked, summaries that reconcile to certified totals, and narration that stops exactly where the evidence stops. Most delivered analytics value is descriptive, done precisely and narrated well — and precision plus narration is a career skill, not a stepping stone.

5.14 Ethics, Averages That Hide People

The Business Analytics in Practice section described description as the analytics organizations trust. This section examines the ethical weight of that trust, extending the guide's running discussions — data use (Section 1.13), problem framing (Section 2.12), measurement design (Section 3.13), and cleaning as editorial power (Section 4.14) — to the summarizing step itself. The chapter's angle is specific: an average is a compression of people into a number, the

compression always loses someone, and the ethical discipline is reporting heterogeneity honestly.

Begin with why this is an ethical matter and not merely a technical one. Every summary statistic in this chapter is a decision about whose experience the reported number represents. A mean order value is pulled upward by high-value baskets and, because it is computed at order grain, gives frequent purchasers more representation than infrequent ones; a median speaks for the middle and silences both tails; a pooled rate lets large groups speak for small ones; a "typical customer" profile manufactures a person who may not exist — the average of a bimodal distribution describes nobody in either hump, which for StyleCraft means a composite of the urban loyalist and the suburban occasion shopper that matches neither. These compressions are unavoidable; description is compression. What is avoidable is presenting the compression as the population — the instinct to mistake a group's summary for its members that Rosling and colleagues catalog as the gap instinct, the human readiness to see two summary numbers as two homogeneous, separated groups when the underlying distributions overlap almost entirely (Rosling et al., 2018). Marketing runs on group summaries — segments, personas, metro comparisons — so marketing analysts hold this particular editorial power in unusual quantity, a concentration of influence continuous with the aggregation warnings of Section 3.3.1.

The risk concentrates where summaries steer treatment of people unevenly. Three of this chapter's own artifacts illustrate. The profile: a suburban-customer profile built on means invites a strategy addressed to the mean — and if suburban order values are wide or two-humped, a meaningful minority of suburban customers (the modest-basket regulars the format also serves) becomes invisible in every deck that quotes it, and decisions about staffing, assortment, and marketing are made as if they were not there. The pooled rate: Section 5.5's paradox is an ethics case as soon as the rate allocates anything — a channel or store group condemned on an aggregate rate that reverses within every stratum is a misallocation manufactured by arithmetic, and the burden lands on whoever staffed, stocked, or was employed by the condemned group. The concentration table: "the top decile carries the revenue" slides quickly into "the bottom deciles do not matter," a sentence that pre-decides service levels for the majority of customers on a window-dependent ranking (Schmittlein et al., 1993) — defensible as strategy only if stated and examined, not smuggled as description.

The discipline this guide asks for is procedural, and it has three parts. First, spread travels with center: no group mean is reported for decision use without its dispersion and group size, and no rate is reported without its numerator, denominator, and underlying group count. Distributions that Section 4.9.1's shape checks show to be bimodal are shown, not averaged — a two-humped histogram in an appendix is more honest than any single row of summary statistics. Second, composition is checked and disclosed: every cross-group comparison of rates carries the one-line mix check of Section 5.5, and where stratification reverses or materially moves a headline, the stratified version is the headline. Third, the silenced are named: every summary that will steer differential treatment states, in a sentence, who it underweights — "this profile is order-weighted and speaks least for infrequent customers" costs a line and converts a hidden editorial choice into

an inspectable one, exactly the conversion Section 4.14 demanded for exclusions. None of this pads the deliverable. All of it keeps faith with the people a certified table compresses — and with the decision-makers entitled to know whom their numbers are quietly not about.

> **CONCEPT**
>
> **Report the People, Not Just the Average**
>
> Every group summary underrepresents someone. Before a descriptive artifact travels: pair every center with its spread and group size; show, rather than average over, distributions the shape checks flag as mixtures; run and disclose the composition check on every cross-group rate comparison; and state in one sentence whom the summary speaks for least. An average reported with these disclosures is description; reported without them, it is a decision about people who were never mentioned.

5.15 Chapter Summary

This chapter took the expansion review from one seductive number to a defensible account of what that number is made of, and built the descriptive disciplines the trip requires. The main point is that summarizing is the second analysis: every descriptive choice — denominator, center, grouping, percentage direction, comparison — changes the story a certified table tells without changing the table, so the choices are made explicitly, checked against composition, and reconciled to known totals rather than left to defaults.

The chapter's instruments arrived in sequence. Frequency tables and proportions established precise counting and the denominator problem — a share is a numerator-denominator pair, and the pair is named every time. The choice of center was matched to distribution shape, mean and median reported together on skewed data, and the averaged average replaced by the weighted recombination that reconciles to totals. Spread was elevated to a finding: heterogeneity, especially the two-humped kind Chapter 4 flagged in StyleCraft's order values, is the question “is this one group?” in statistical clothes. Cross-tabulations brought the three percentage bases and the direction rule, and Simpson's paradox — computed, per this chapter's ownership, as the table-level case of Section 3.3.1's aggregation warning — made composition checks a standing obligation for every comparison of rates. Profiles assembled group comparisons on distributions at declared grains. The applied metric families organized the marketing repertoire — revenue, acquisition, conversion, retention, engagement, formulas cataloged in Appendix B — and the decomposition identity did the chapter's central work: average order value factored exactly into basket size and revenue per unit, revenue per unit examined against category mix, and the whole audited by the twin checks that factors multiply back and summaries recombine to certified totals. Period-active cohort retention tables put comparisons at like ages, while naming what alignment does not remove; the funnel disciplined description along the journey and exposed stage definitions and stage units as measurement decisions; concentration analysis, benchmarks, and the insight statement turned findings into context-bearing sentences that stop where description stops. Throughout, AI assistants drafted the summaries while the analyst audited the

claims — denominators, centers, compositions, reconciliations, and verbs — and the ethics section named the stakes: averages hide people, and honest heterogeneity reporting is what keeps the compression accountable.

Looking ahead, the labs left StyleCraft with a decomposed AOV gap, a profile of two customer worlds, and cohort evidence about how each world repeats — description that keeps insisting the company's customers are not one population. The next chapter takes that insistence formally: it opens Part II by asking which customer groups the data actually contains, building the customer-grain feature table and using recency, frequency, and monetary structure — and then clustering — to discover segments rather than assume them. Description found the humps; segmentation names them.

5.16 Exercises for Practice and Homework

The following exercises practice the chapter's main habits: name the denominator, match the center to the shape, check composition before comparing, decompose before interpreting, and reconcile every summary to known totals. They are organized into two groups. Core chapter practice is the required path and should be completed by every student, and it holds the two homework submissions from which your instructor will assign a subset; Self-Assessment #3 draws on the required-practice sets. In-class activities are prepared for discussion rather than submitted. Each exercise also carries its assignment label — required practice, homework submission, or in-class discussion — so that instructors can assign selectively.

5.16.1 Core Chapter Practice

Exercise 5.1 Concept Check (Required Practice)

1. State the denominator problem in one sentence, and give two denominators that turn the same email-open numerator into materially different rates.
2. On a right-skewed order-value distribution, state the order in which mean, median, and mode commonly fall, explain why that ordering is a working heuristic rather than a theorem, and give one business question each center is the right answer to.
3. Why is the simple average of per-store AOVs generally not the company's AOV, and what recombination is?
4. Two stores share a mean order value of $60. Describe two spread patterns that would demand different marketing responses, and name the statistics that distinguish them.
5. A cross-tabulation supports three percentage bases. Name them, compute all three for the 540-customer cell of Table 5.3, and state Zeisel's direction rule for choosing which to report.
6. Define Simpson's paradox without the word “paradox,” and state the one-line check that detects the conditions for it.
7. Write the decomposition identity for AOV, extend it by one factor to revenue per customer, and state the two verification equalities every decomposition must pass. Then

explain why “basket size accounts for 60 percent of the gap” is not a claim the identity supports.

8. Distinguish a period-active cohort table from a survival retention table, state which one can show a rise from one period to the next and why, and explain what reading down a column compares that the pooled rate cannot — and what it still does not control for.

Exercise 5.2 Pick the Denominator (Required Practice)

For each reported figure, name at least two defensible denominators, state how the choice changes the claim, and pick one with a one-sentence justification tied to a decision.

1. “Our SMS click rate is 9 percent.”
2. “Store accounts for 44 percent of our business.”
3. “Repeat rate at the Greenwich store is 22 percent.”
4. “The Welcome Series converts at 3.1 percent.”
5. “Occasionwear is 60 percent of suburban sales.”
6. “CAC came in at $41 last quarter.”

Exercise 5.3 Choose the Center and Report the Spread (Required Practice)

For each situation, choose the center (or pair) to report, the spread measure to accompany it, and state in one sentence what the choice protects the reader from.

1. Order value across all channels, for the trading report's headline.
2. The “typical basket” for a store-associate training deck.
3. Units per order, which takes only small integer values.
4. Revenue per customer for the expansion review, where Chapter 4's exploration showed a two-humped shape.
5. Time between first and second purchase, where a third of customers have no second purchase yet.

Exercise 5.4 Reconcile the Summary (Required Practice)

A certified file has 80,000 order lines, 52,000 orders, 8,000 customers, and total revenue of $4,160,000. An AI-drafted summary reports: metro AOVs of $71 (urban, 40,000 orders) and $118 (suburban, 11,600 orders); an overall AOV of $94.50 described as “the average of the metro AOVs”; and a repeat rate of 41 percent “of active customers.” Write the reconciliation audit in advance: which totals the metro rows must sum to and whether they do; the correct pooled AOV from the certified totals and the size of the averaged-average error; what is wrong, per Section 3.5.1, with the repeat-rate sentence as reported; and the two questions you would ask before this summary is allowed into a pre-read. (This exercise applies the chapter's verification theme; it is the summaries' version of Exercise 4.4.)

Exercise 5.5 Decompose the Gap End-to-End (Homework Submission)

Complete Lab 5.1 Part B on the certified full file: the metro AOV table with reconciliation to certified totals, the two-step analysis (the exact basket × revenue-per-unit decomposition, then

the standardized comparison of revenue per unit at common within-category prices, plus the same-product realized-price check the miniature could not perform), the identity audit, and a five-sentence insight statement per Section 5.9 addressed to the expansion review. Submit the notebook, the reconciliation log, and the completed AI-Use Appendix (Appendix D) recording every assistant exchange, including at least one delegated summary whose stated denominator, center, or reconciliation you had to correct.

Exercise 5.6 AI-Assisted Summary Audit (Homework Submission)

Give an AI assistant the Lab 5.1 Part A miniature and this deliberately vague prompt, verbatim: “Summarize this sales data and tell me what it means.” Run its code on a copy. Then audit against the chapter: identify every center it chose and whether the shape justified it; every rate or share it reported and whether the denominator was stated; whether any overall figure was an averaged average; whether its totals reconcile to the miniature's certified figures (14 lines, 8 orders, 660.0); and every causal verb in its narration, each rewritten into descriptive language. Conclude with a rewritten prompt per the AI in Practice pattern that would have forced the choices into the open, and submit the audit with the AI-Use Appendix.

5.16.2 In-Class Activities

Exercise 5.7 Find the Flaw in the Description (In-Class Discussion)

Each scenario below describes a summary that is computed correctly. Find the flaw, name the section whose discipline it violates, and propose the repair.

1. A deck reports “our customers' average age band is 25–34” from the modal age_band, and a second slide computes a numeric “mean age” from the band midpoints.
2. The overall conversion rate in a pre-read is the simple average of forty campaigns' conversion rates, and it exceeds every large campaign's rate.
3. A channel comparison shows App with the highest repeat rate; App launched fourteen months ago and its customers skew to the newest signups' most engaged adopters.
4. A funnel review celebrates checkout conversion rising from 58 to 66 percent in the week a new definition stopped counting abandoned guest sessions as checkouts.
5. A campaign dashboard reports a single “funnel conversion rate” computed as attributed orders divided by impressions, and labels the three stages a nested funnel.
6. A profile of “the suburban customer” reports means only; the order-value histogram for the metro shows two humps.
7. The top-decile revenue share is reported from a 30-day window to argue that “a tiny elite carries the company,” and the deciles were cut on customers active in that window only.

Exercise 5.8 Averages That Hide People Mini-Cases (In-Class Discussion)

For each mini-case, identify who the summary hides, what Section 5.14's discipline requires before the artifact travels, and what one disclosure sentence you would attach.

1. A staffing model allocates store hours by mean transaction time per store; two stores serve a bimodal mix of quick pickup orders and long occasionwear fittings.
2. A “customers like these don't respond” conclusion about a metro, drawn from a pooled response rate, when response within every age band in that metro matches the company average.
3. A persona deck describes “the StyleCraft customer” as a 26-year-old app user, built from modal values of every attribute independently.
4. A proposal to cut service investment for bottom-decile customers, ranked on a quarter of revenue, presented without the ranking's window or churn-of-decile-membership.
5. An executive asks for “just the one number” for suburban performance, and the analyst must choose what to send and what to attach.

5.17 Glossary of Terms

This glossary includes only the terms introduced in this chapter. Each definition is tied to the sources used in the chapter rather than added for decoration.

Benchmark. A reference value — internal-historical, internal-cross-sectional, or external — against which a metric is placed so that the number becomes a claim; external benchmarks rarely share a metric's four defining elements and are used as directional context only.

Central tendency. The family of single-value summaries of a numeric variable — mean, median, and mode — whose members diverge on skewed data, making the choice among them an analytic decision (adapted from Freedman et al., 2007).

Cohort. A group of units defined by a shared starting event in a shared time window, followed forward from that event so that groups are compared at the same age rather than the same date; alignment removes tenure from the comparison but not calendar-period or compositional differences (adapted from Glenn, 2005).

Cohort retention table. The descriptive artifact of cohort analysis: rows for cohorts, columns for periods since the starting event, cells holding each cohort's share still present under a stated activity rule; read down a column for same-age comparisons. A period-active table records activity in each individual period and may rise from one period to the next; a survival retention table records continuous retention and is non-increasing by construction.

Concentration analysis. The description of how unevenly a total is distributed across the units producing it, typically via ranked deciles and their shares; its summary claims are window-dependent (adapted from Twedt, 1964; Schmittlein et al., 1993).

Conversion funnel. A descriptive structure arranging a journey as ordered stages with counts at each stage and rates between stages, each stage bounded by a stated operational definition

and each rate labeled step-to-step or cumulative; stages are strictly nested only when every stage counts the same unit (adapted from Strong, 1925; Lavidge & Steiner, 1961).

Cross-tabulation. A table counting observations for every combination of two categorical variables, supporting cell, row, and column percentages; percentages are computed in the direction of the comparison being made (adapted from Zeisel, 1985; Agresti, 2019).

Customer profile. A curated table of descriptive comparisons across groups — sizes first, centers with spreads, each cell at a declared grain — chosen to answer an analytic specification rather than to exhaust the data.

Decomposition identity. An exact factoring of a composite metric into components that multiply back to it — for example, average order value into units per order times revenue per unit — used to express a gap as a product of factors and audited by recombination against known totals. It yields multiplicative factors, not additive percentage shares.

Denominator problem. The failure mode in which a proportion, rate, or average is reported or read without its denominator, allowing several correct numbers to support several different claims about the same numerator.

Dispersion. The family of summaries of a variable's variability — range, interquartile range, and standard deviation — one of which accompanies every reported center (adapted from Freedman et al., 2007; Tukey, 1977).

Frequency table. A summary listing each distinct value of a variable with its count and, in working form, its proportion of a stated total, complete only with its unit of analysis and denominator declared (adapted from Freedman et al., 2007; Agresti, 2019).

Heterogeneity. Wide or structured variation within a group — including bimodality — treated in this guide as a descriptive finding prompting the question of whether the group is one population, rather than as noise around an average.

Insight statement. A descriptive claim equipped for decision use: the comparison or decomposition that gives the number meaning, magnitude in decision units, connection to the named decision, and verbs that assert no more than the description supports.

Simpson's paradox. The reversal or disappearance of an association between two variables upon disaggregation by a third, arising when compared groups are unevenly composed across strata with different base rates; detected by composition checks, not arithmetic rechecks (adapted from Simpson, 1951; Blyth, 1972).

Weighted mean. The recombination of group means into an overall mean using group sizes as weights; the only recombination that reconciles to the pooled total, and the standing correction to the averaged average.

5.18 Further Readings

Students who want additional background may begin with the following readings. General descriptive foundations are listed first, marketing applications second.

- Freedman et al. (2007) for the clearest available treatment of summary statistics, what they assume, and how they mislead — the statistical conscience behind Sections 5.2 through 5.4.
- Zeisel (1985) for the classic, still-unmatched short book on percentages, tables, and the direction rule, and Blyth (1972) for a compact formal treatment of Simpson's paradox; readers wanting the original note may consult Simpson (1951).
- Farris et al. (2010) for the standard practitioner reference on marketing metrics — the extended companion to Section 5.7 and Appendix B.
- Schmittlein et al. (1993) for the definitive skeptical examination of 80/20 claims, including why measured concentration depends on the observation window.
- Glenn (2005) for a rigorous short treatment of cohort analysis and the identification problems that make same-age comparison necessary — and that same-age comparison does not fully solve.
- Rosling et al. (2018) for the gap instinct and the broader case that summary comparisons of groups systematically overstate their separation — the accessible companion to Section 5.14.

5.19 References

Agresti, A. (2019). *An introduction to categorical data analysis* (3rd ed.). Wiley.

Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association, 67*(338), 364–366. https://doi.org/10.1080/01621459.1972.10482387

Farris, P. W., Bendle, N. T., Pfeifer, P. E., & Reibstein, D. J. (2010). *Marketing metrics: The definitive guide to measuring marketing performance* (2nd ed.). Pearson Education.

Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics* (4th ed.). W. W. Norton.

Glenn, N. D. (2005). *Cohort analysis* (2nd ed.). Sage.

Lavidge, R. J., & Steiner, G. A. (1961). A model for predictive measurements of advertising effectiveness. *Journal of Marketing, 25*(6), 59–62. https://doi.org/10.1177/002224296102500611

Rosling, H., Rosling Rönnlund, A., & Rosling, O. (2018). *Factfulness: Ten reasons we're wrong about the world — and why things are better than you think*. Flatiron Books.

Schmittlein, D. C., Cooper, L. G., & Morrison, D. G. (1993). Truth in concentration in the land of (80/20) laws. *Marketing Science, 12*(2), 167–183. https://doi.org/10.1287/mksc.12.2.167

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological), 13*(2), 238–241. https://doi.org/10.1111/j.2517-6161.1951.tb00088.x

Strong, E. K. (1925). *The psychology of selling and advertising*. McGraw-Hill.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Twedt, D. W. (1964). How important to marketing strategy is the "heavy user"? *Journal of Marketing, 28*(1), 71–72. https://doi.org/10.1177/002224296402800117

Zeisel, H. (1985). *Say it with figures* (6th ed.). Harper & Row.