Emily Riederer

Data Disasters

To all the mistakes I've made (data, and otherwise) and those who tolerated my making them.

Contents

List of Tables

List of Figures

Preface

Training in data analysis often begins with Statistics 101 course. Students learn the "happy path" of answer data that adheres to specific assumptions (such as "independent and identically distributed with a Normal density") and answers pre-specified questions (most notably, the infamous null hypothesis significance test). Then, they venture out into the world of real-world data analysis where non-experimental data is rarely so well behaved and the questions asked of it are far more nuanced.

No one course can aim to teach students everything they should know about statistics. In fact, one of the greatest privileges of a career in statistics is the responsibility and privilege of life-long learning. However, **the flaw of introductory statistics is not that it's incomplete, but that it's not obvious how it is *not* complete**. Statistics is a bad salesman. There's no season finale, no cliff hanger, no teasing and hinting and promising more and better to come. Student may leave thinking that answering more complex data analysis questions is trivially easy (by relying on the one-size-fits-all "panacea" that they learned) or intractably difficult (when assumptions are not met.)

This book attempts to add more color to all the dimensions of data analysis while showcasing the nuances throughout the true *life cycle* of data analysis using two strategies.

First, it attempts to showcase common pitfalls in all the parts of data analysis: from data management and computation to visualization, interpretation, and modeling and even to communication and collaboration. Data analysis is fundamentally a *creative* task, so there are rarely canonical one-size-fits-all solutions. Curiously, however, there are plenty of canonical *issues* even if they require different solutions in different settings. Thus, the goal of this book is to highlight common *data disasters* and, in doing so, help students cultivate an intuition for how to detect common problems before they occur in an important analysis.

Second, while exploring these *data disasters*, we humbly put forth a (woefully incomplete!) literature review of more advanced methods from statistics and other quantitative disciplines (e.g. economics, epidemiology), to help learners build a "mental index" of terms to search and techniques to study should they encounter a relevant problem.

0.1 Main Topics

In particular, we will aim to help you avoid ten types of data disasters:

- **Data Dalliances**: Misinterpreting or misuing data based on how it was collected or what it represents
- **Computational Quandaries**: Letting computers do what you said and not what you meant
- **Egregious Aggregations**: Losing critical information when information is condensed
- **Vexing Visualization**: Confusing ourselves or others with plotting choices
- **Incredible Inferences**: Drawing incorrect conclusions for analytical results
- **Cavalier Causality**: Falling prey to spurious correlations masquerading as causality
- **Mindless Modeling**: Failing to get the most value out of models by not tailoring the features, targets, and performance metrics
- **Alternative Algorithms**: Lacking an understanding of alternative methods which may be better suited for the problem at hand
- **Complexifying Code**: Making projects unwieldy or more difficult to understand than necessary
- **Rejecting Reproducibility**: Working inefficiently instead of an efficient, reproducible, and sharable workflow

0.2 Common Themes

In each chapter, we will see numerous examples of each disaster and consider strategies to help us mitigate. Along the way, we'll emphasize:

- The importance of **domain knowledge** and the **data-generating process** to decide what it is you want to do
- The utility of **simulation** as a tool to explore if, in fact, you are doing it
- The exploration of **counterexamples** to build **intuition for common patterns** of problems even where common solutions don't exist

As we go, we will notice how three common themes that challenge the focus of introductory statistics:

- Summary statistics masks interesting stories that we see when focusing on the **variation**
- Similarly, observations and variables are rarely independent; the story is in the **covariance**
- Assumptions of Normality, or more broadly symmetry, are often in appropriate in wonky, **highly skewed** world

About the Author

Emily Riederer is...

1

Introduction

Statistics is not synonymous with data analysis; rigor vs practicality

"Evaluating the Success of a Data Analysis" (Hicks and Peng, 2019)

"Data Alone is not Ground Truth" (Bassa, 2017)

1.1 What is data?

Data is...

1.2 What is analysis?

Analysis is the process of turning information into insight.

1.3 What is data analysis?

Data analysis altogether is...

1.4 A case study

Now unplug your Internet cable, and start doing some serious work.

We have a nice figure in Figure 1.1, and also a table in Table 1.1.

```
par(mar = c(4, 4, 1, .1))
plot(cars, pch = 19)
```

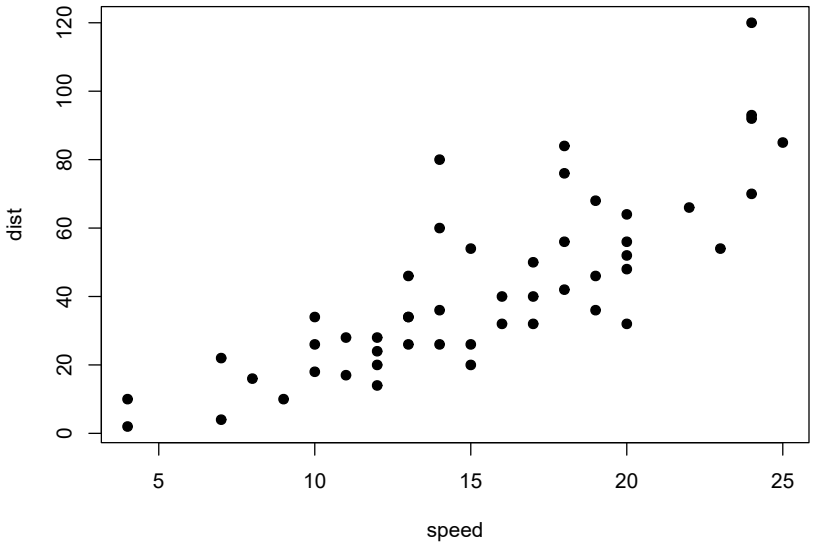


FIGURE 1.1: Hello World!

```
knitr::kable(
  head(iris), caption = 'The boring iris data.',
  booktabs = TRUE
)
```

TABLE 1.1: The boring iris data.

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

Data

2

Data Dalliances

The first step to data analysis is, in fact, data. While this may seem obvious, statistics textbooks often dodge this detail. Discussions of regression analysis often begin with a statement like:

"Let X be the nxp design matrix of independent variables…"

---

but in practice this statement is as absurd as writing a book about how to win a basketball game, assuming your team already has a 20 point lead with 1 minute left to play.

It's very convenient but typically incorrect to assume that the data we happen to have is the ideal (or, more humbly, sufficient) data for the questions we wish to analyze. The specific vagaries of data vary greatly by domain, but a commonality across many fields (such as political science, economics, epidemiology, and market research) is that we are often called to work with *found data* (or, more formally, "observational data") from administrative sources or production systems. In contrast to artisanally crafted data experimental data (like the carefully controlled agricultural experiments which motivated many early methods developments in statistics), this data was generated neither by us not for us. To quote Angela Bassa, the head of data science at an e-commerce company: "Data isn't ground truth. Data are artifacts of systems" (Bassa, 2017).

The analytical implications of observational versus experimental data are well explored in the field of causal inference (which we will discuss some in Chapters 6 and 7). However, this distinction has implications far earlier in the data analysis process, as well. To name a few:

- Records and fields many not represent the entities or measures most conducive to analysis
- Data collection methods may capture a different subset of events or do so at a different frequency than we expected, leading to systemic biases

- Data movement between systems can insert errors (or, at minimum, challenges to our intuition)
- Data transformations may be fragile or transient, reflecting the primary purpose of the system not our unrelated analytical use

In this chapter, we will explore data structures and the full data generating process to better understand how different types of data challenges emerge. In doing so, we will hone sharper intuition for how our data can deceive us and what to watch out for when beginning an analysis.

2.1 Preliminaries

Before we begin our exploration of data dalliances, we must first establish a baseline understanding of data structure, data production, and data quality.

2.1.1 Data Structure Basics

Understanding the content and structure of the data you are using is a critical prerequisite to analysis. In this book, we focus on tabular, structured data like one might find in an Excel spreadsheet or relational database.[1]

In particular, many tools work best with what R developer Hadley Wickham describes as "tidy data" (Wickham, 2014). Namely:

1. Each variable forms a column
2. Each observation forms a row
3. Each type of observational unit forms a table

This is analogous to how one generally finds data arranged in a database and how statisticians are used to conceptualizing it. For example, the design matrix of a linear model consists of one column of data for each independent variable to be included in the model and one row for each observation.[2] As Wickham points out, this is also similar to what is called "3rd normal form" in the world of relational database management systems.

[1] Other types of data that one might encounter in the wild include free text, JSON, spatial data, and more. However, many of these require either more advanced analysis techniques or structuring that converts them into tabular data, so they are out of the scope of this discussion.

[2] When data is arrange this way in a matrix X, linear regression coefficients can be computed as $\beta = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$

Using this data structure is valuable not only because it is similar to what many modern data tools expect, but also because it provides us a framework to think critically about what defined each observation and each variable in our dataset.

2.1.2 Data Production Processes

In statistical modeling we discuss the **data generating process**: we can build models that describe the mechanisms that create our observations. We can broaden this notion to think about the generating process of each of these steps of data production.

Regardless of the type of data (experimental, observational, survey, etc.), there are generally four main steps to production: collection, extraction, loading, and transformation.[3]

- **Collect**: The way in which signals from the real world are captured as data. This could include logging (e.g. for web traffic or system monitoring), sensors (e.g. temperature collection), surveys, and more
- **Extract**: The process of removing data from the place in which it was originally captured in preparation of moving it somewhere in which analysis can be done
- **Load**: The process of loading the extracted data to its final destination
- **Transform**: The process of modeling and transforming data so that its structure is useful for analysis and its variables are interpretable

To better theorize about data quality issues, it's useful to think of four DGPs: the real-world DGP, the data collection/extraction DGP[4], the data loading DGP, and the data transformation DGP.

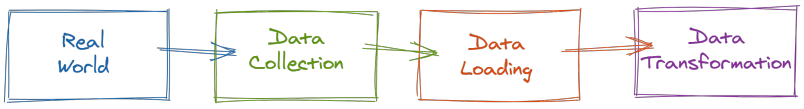


FIGURE 2.1: A schematic of the data production process

For example, consider the role of each of these four DGPs for e-commerce data:

[3] You may hear the last three referred to as ELT or ETL

[4] I don't mean to imply statisticians do not regularly think about the data collection DGP! The rich literature on missing data imputation, censored data in survival analysis, and non-response bias is survey data collection are just a few examples of how carefully statisticians think about how data collection impacts analysis. I chose to break it out here to discuss the more technical aspects of collection

- **Real-world DGP**: Supply, demand, marketing, and a range of factors motivate a consumer to visit a website and make a purchase
- **Data collection DGP**: Parts of the website are instrumented to log certain customer actions. This log is then extracted from the different operational system (login platforms, payment platforms, account records) to be used for analysis
- **Data loading DGP**: Data recorded by different systems is moved to a data warehouse for further processing through some sort of manual, scheduled, or orchestrated job. These different systems may make data available at different frequencies.
- **Data transformation DGP**: To arrive at that final data presentation requires creating a data model[5] to describe domain-specific attributes with key variables crafted with data transformations

Or, consider the role of each of these four DGPs for subway ridership data[6]:

- **Real-world DGP**: Riders are motivated to use public transportation to commute, run errands, or visit friends. Different motivating factors may cause different weekly and annual seasonality
- **Data collection DGP**: To ride the subway, riders go to a station and enter and exit through turnstiles. The mechanical rotation of the turnstile caused by a rider passing through is recorded
- **Data loading DGP**: Data recorded at each turnstile is collected through a centralized computer system at the station. Once a week, each station uploads a flat file of this data to a data lake owned by the city's Department of Transportation
- **Data transformation DGP**: Turnstiles from different companies may have different data formats. Transformation may include harmonizing disparate sources, coding system-generated codes (e.g. Station XYZ) to semantically meaningful names (e.g. Main Street Station), and publishing a final unified representation across stations and across time

Throughout this chapter, we'll explore how understanding key concepts about each of these DGPs can help guide our intuition on where to look for problems.

2.1.3 Data Quality Dimension

To guide our discussion of how data production can affect aspects of data quality, we need a guiding definition of data quality. This is challenging because data quality is *subjective* and *task-specific*. It matters much more if data is

[5] https://en.wikipedia.org/wiki/Data_model

[6] Like NYC's infamously messy turnstile data[7]. I don't claim to know precisely how this dataset is created, but many of the specific challenges it contains are highly relevant.

"fit for purpose" and operates in a way that is *transparent* to its users moreso than meeting some preordained quality standard.

Regardless, it's useful for our discussion to think about general dimensions of data quality. Here, we will rely on six dimensions of data quality outlined by Data Management Association. STheir official definitions are:

1. **Completeness**: The proportion of stored data against the potential of "100% complete"
2. **Uniqueness**: Nothing will be recorded more than once based upon how that thing is identified. It is the inverse of an assessment of the level of duplication
3. **Timeliness**: The degree to which data represent reality from the required point in time
4. **Validity**: Data are valid if it conforms to the syntax (format, type, range) of its definition
5. **Accuracy**: The degree to which data correctly describes the "real world" object or event being described.
6. **Consistency**: The absence of difference, when comparing two or more representations of a thing against a definition

2.2 Data Collection

One of the tricky nuances of data collection is understanding what precisely is getting captured and logged in the first place. No matter how robust the sensors, loggers, or other mechanisms are that record our dataset, that data is still unfit for its purpose so long as the analyst does not fully understand what it represents. In the next section, we will see how what data gets collected (and our understanding of it) can alter our notions of data completion and how we must handle it in our computations.

2.2.1 What Makes a Record (Row)

The first priority when starting to work with a dataset is understanding what a single record (row) represents and what causes it to be generated.

Consider something as simple as a login system where users must enter their credentials, endure a Captcha-like verification process to prove that they are not a robot, and enter a multi-factor authentication code. Figure 2.2 depicts such a process.

Which of these events gets collected and recorded has a significant impact

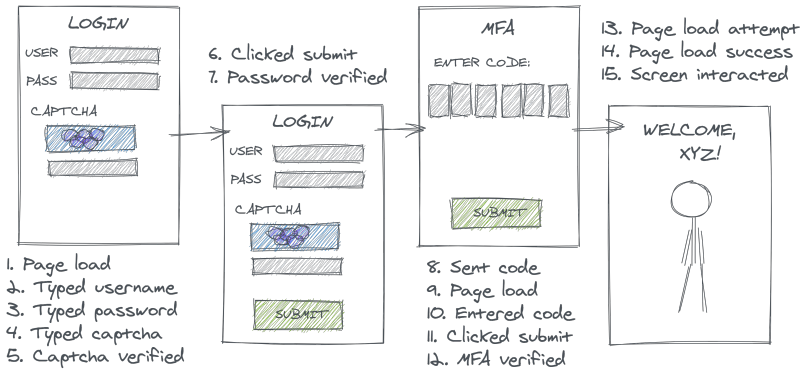


FIGURE 2.2: A diagram illustrating a multi-step process for a user to login to a website or app

on subsequent data processing. In a technical sense, no inclusion/exclusion decision here is *incorrect*, per say, but if the producers' choices don't match the consumers' understandings, it can lead to misleading results.

For example, an analyst might seek out a `logins` table in order to calculate the rate of successful website logins. Reasonably enough, they might compute this rate as the sum of successful events over the total. Now, suppose two users attempt to login to their account, and ultimately, one succeeds in accessing their private information and the other doesn't. The analyst would probably hope to compute and report a 50% login success rate. However, depending on how the data is represented, they could quite easily compute nearly any value from 0% to 100%.

Figure 2.3 depicts a few different realistic cases:

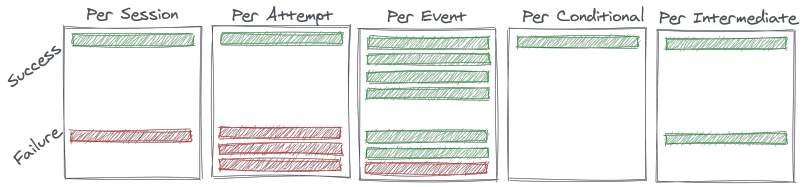


FIGURE 2.3: Login events recorded under different data collection paradigms

- **Per Attempt**: If data is logged once per overall login attempt, successful attempts only trigger one event, but a user who forgot their password may try (and fail) to login multiple times. In the case illustrated above, that deflates the successful login rate to **25%**.

- **Per Event**: If the logins table contains a row for every login-related event, each 'success' will trigger a large number of positive events and each 'failure' will trigger a negative event preceded by zero or more positive events. In the case illustrated below, this inflates our successful login rate to **86%**.
- **Per Conditional**: If the collector decided to only look at downstream events, perhaps to circumvent record duplication, they might decide to create a record only to denote the success or failure of the final step in the login process (MFA). However, login attempts that failed an upstream step would not generate any record for this stage because they've already fallen out of the funnel. In this case, the computed rate could reach **100%**
- **Per Intermediate**: Similarly, if the login was defined specifically as successful password verification, the computed rate could his **100%** even if some users subsequently fail MFA

	Session	Attempt	Attempt	Outcome	Intermediate
Success	1	1	6	1	2
Total	2	4	7	1	2
Rate	50%	**25%**	**86%**	**100%**	**100%**

While humans have a shared intuition of what concepts like a user, session, or login are, the act of collecting data forces us to map that intuition onto an atomic event . Any misunderstanding in precisely what that definition is can have massive impact on the perceived data quality; "per event" data will appear heavily duplicated if it is assumed to be "per session" data.

In some cases, this could be obvious to detect. If the system outputs fields that are incredibly specific (e.g. with some hyperbole, imagine a `step_in_the_login_process` field with values taking any of the human-readable descriptions of the fifteen processes listed in the image above), but depending how this source is organized (e.g. in contrast to above, if we only have fields like `sourceid` and `processid` with unintuitive alphanumeric encoded values) and defined, it could be nearly impossible to understand the nuances without uncovering quality metadata or talking to a data producer.

2.2.2 What Doesn't Make a Record (Row)

Along with thinking about what *does* count (or gets logged), it's equally important to understand what systemically does not generate a record. Consider users who have the intent or desire to login (motivated by a real-world DGP) but cannot find the login page, or users who load the login page but never click a button because they know that they've forgotten their password and see no way to request it. Often, some of these corner cases may be some of

the most critical and informative (e.g. here, demonstrating some major flaws in our UI). It's hard to *computationally* validate what data doesn't exist, so *conceptual* data validation is critical.

2.2.3 Records versus Keys

The preceding discussion on what types of real-world observations will or will not generate records in our resulting dataset is related to but distinct from another important concept from the world of relational databases: **primary keys**.

Primary keys are a minimal subset of variables in a dataset than define a unique record. For example, in the previous discussion of customer logins this might consist of **natural keys**[8] such as the combination of a `session_id` and a `timestamp` or **surrogate keys**[9] such as a global `event_id` that is generated every time the system logs any event.

Understanding a table's primary keys can be useful for many reasons. To name a few reasons, these fields are often useful for linking data from one table to another and for identifying data errors (if the uniqueness of these fields are not upheld). They also can be suggestive of the true granularity of the table.

However, simply knowing a table's primary keys does *not* resolve the issues we discussed in the prior two sections. Any of the many different data collection strategies we considered are *unique* by session and timestamp; however, as we've seen, that is no guarantee that they *must* contain every session and timestamp in the universe of events.

2.2.4 What Defines a Variable (Column)

Just as critical as understanding what constitutes a record (row) in a dataset is understanding the precise definition of each variable (column). Superficially, this task seems easier: after all, each variable has a name which hopefully includes some semantic information. However, quite often this information can provide a false sense of security. Just because you identify a variable with a promising sounding name, that does not mean that it is the most relevant data for your analysis.

For example, consider wanting to analyze patterns in customer spend amounts across orders on an e-commerce website. You might find a table of orders with a field called `amt_spend`. But what might this mean?

- If the dataset is sourced from a payment processor, it likely includes the

[8] Keys with semantic meanings that are naturally part of the dataset
[9] Keys without semantic meaning that exist primarily for the purpose of being keys

total amount billed to a customers' credit card: including item prices less any discounts, shipping costs, taxes, etc. Alternatively, if this order was split across a gift card and a credit card, this field might only reflect the amount charged to the credit card

- If the dataset is created for Finance, it might perhaps include only the total of item prices less discounts if this best corresponded to the data the Finance team needs for revenue reporting
- Someone, somewhere, at some point might have assigned `amt_spend` to the name of the variable containing gross spend (before accounting for any discounts) and there might be a different variable `amt_spend_net` which accounts for discounts applied

It's critical to understand what each variable *actually* means. The upside of this is that it forces analysts to more crisply think about their research questions and what the *ideal* variables for their analysis would be. As we've seen, concepts like "spend" may seem deceptively simple, but are not unambiguous.

2.2.5 The Many Meanings of Null

Related to the presence and absence of full *records* is the presence or absence of individual *fields*. If records contain some but not all relevant information, they may be published with explicitly missing fields or the full record may not be published at all. The difference between implicit and explicit missingness on the resulting data is illustrated in Figure 2.4.

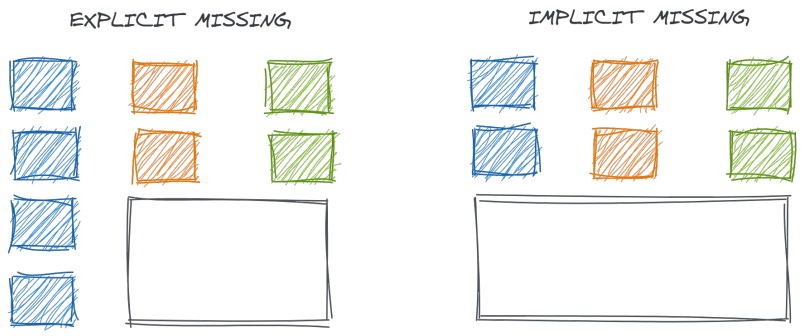


FIGURE 2.4: A comparison of explicit versus implicit missingness

Understanding what the system implies by each *explicitly* missing data field is also critical for validation and analysis. Checks for data completeness usually include counting null values, but null data isn't always incorrect. In fact, null data can be highly informative if we know what it means. Some meanings of null data might include:

- **Field is not relevant**: Perhaps our `logins` table reports the mobile phone operating system (iOS or Android) that was used to access the login page to track platform-specific issues. However, there is no valid value for this
- **Relevant value is not known**: Our `logins` table might also have an `account_id` field which attempts to match login attempts to known accounts/customers using different metadata like cookies or IP addresses. In theory, almost everyone trying to log in should have an account identifier, but our methods may not be good enough to identify them in all cases
- **Relevant value is null**: Of course, sometimes someone without an account at all might try to log in for some reason. In this case, the correct value for an `account_id` field truly *is* null
- **Relevant value was recorded incorrectly**: Sometimes systems have glitches. Without a doubt, every single login attempt *should* have a timestamp, but such a field could be null if this data was somehow lost or corrupted at the source

Similarly, different systems might or might not report out these nulls in different ways such as:

- **True nulls**: Literally the entry in the resulting dataset is null
- **Null-like non-nulls**: Blank values like an empty string (`''`) that contain a null amount of information but won't be detected when counting null values
- **Placeholder values**: Meaningless values like an `account_id` of `00000000` for all unidentified accounts which preserve data *validity* (the expected structure) but have no intrinsic meaning
- **Sentinel/shadow values**: Abnormal values which attempt to indicate the reasons for null-ness such as an `account_id` of `-1` when no browser cookies were found or `-2` when cookies were found but did not help link to any specific customer record

Each of these encoding choices changes the definitions of appropriate completeness and validity for each field and, even more critically, impacts the expectations and assertions we should form for data accuracy. We can't expect 100% completeness if nulls are a relevant value; we can't check validity of ranges as easily if sentinel values are used with values that are outside the normal range (hopefully, or we have much bigger problems!) So, understanding how upstream systems *should* work is essential for assessing if they *do* work.

Similarly, understanding how our null data is collected has significant implications for how we subsequently process it. We will discuss this more in Chapter 3 (Computational Quandaries).

2.3 Data Extraction & Loading

Checking that data contains expected and *only* expected records (that is, completeness, uniqueness, and timeliness) is one of the most common first steps in data validation. However, the superficially simple act of loading data into a data warehouse or updating data between tables can introduce a variety of risks to data completeness which require different strategies to detect. Data loading errors can result in data that is stale, missing, duplicate, inconsistently up-to-date across sources, or complete but for only a subset of the range you think.

While the data quality principles of **completeness**, **uniqueness**, and **timeliness** would suggest that records should exist once and only once, the reality of many haphazard data loading process means data may appear sometime between zero and a handful of times. Data loads can occur in many different ways. For example, they might be:

- manually executed
- scheduled (like a cron[10] job)
- orchestrated (with a tool like Airflow[11] or Prefect[12])

No approach is free from challenges. For example, scheduled jobs risk executing before an upstream process has completed (resulting in stale or missing data); poorly orchestrated jobs may be prevented from working due to one missing dependency or might allow multiple stream to get out of sync (resulting in multisource missing data). Regardless of the method, all approaches must be carefully configured to handle failures gracefully to avoid creating duplicates, and the frequency at which they are executed may cause partial loading issues if it is incompatible with the granularity of the source data.

2.3.0.1 Data Load Failure Modes

To develop our understanding of the true data generating process and to formulate theories on how our data could be broken (and what we should validate), it is useful to understand the different ways data extraction and loading can fail.

Figure 2.5 illustrates a number of examples. Suppose that each row of boxes in the diagram represents one day of records in a table.

Our dataset might be susceptible to:

[10] https://en.wikipedia.org/wiki/Cron
[11] https://airflow.apache.org/
[12] https://www.prefect.io/

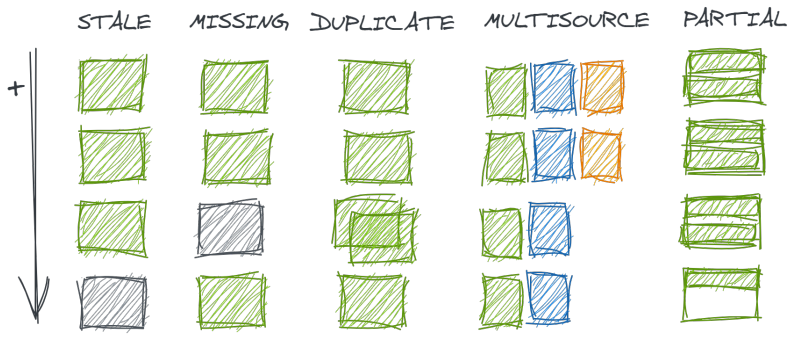


FIGURE 2.5: Different modes of data loading failure

- **Stale data** occurs when the data is not as up-to-date as would be expected from is regular refresh cadence. This could happen if a manual step was skipped, a scheduled job was executed before the upstream source was available, or orchestrated data checks found errors and quarantined new records
- **Missing data** occurs when one data load fails but subsequent loads have succeeded
- **Duplicate data** occurs when one data load is executed multiple times
- **Multisource missing data** occurs when a table is loaded from multiple sources, and some have continued to update as expected while others have not
- **Partial data** occurs when a table is loaded correctly as intended by the producer but contains less data than expected by the consumer (e.g. a table loads ever 12 hours but because there is some data for a given date, the user assumes that all relevant records for that date have been loaded)

The differences in these failure modes become important when an analyst attempts to assess data completeness. One of the first approaches an analyst might consider is simply to check the `min()` and `max()` event dates in their table. However, this can only help detect stale data. To catch missing data, an analyst might instead attempt to count the number of `distinct` days represented in the data; to detect duplicate data, that analyst might need to count records by day and examine the pattern.

Metric	Stale	Missing	Duplicate	Multi	Partial
`min(date)`	**13**	14	14	14	14
`max(date)`					
`count(distinct date)`	**3**	**3**	4	4	4

Metric	Stale	Missing	Duplicate	Multi	Partial
`count(1) by date`	**1001001000100100010010010020010010**			01006666	10010010050
`count(1) count(distinct PKs)`	300300	300300	**400300**	332332	350350

In a case like the toy example above where the correct number of rows per date is highly predictable and the number of dates is small, such eyeballing is feasible; however when the expected number of records varies day-to-day or time series are long, this approach becomes subjective, error-prone, and intractable. Additionally, it still might be hard to catch errors in mutli-source data or partial loads if the lower number of records was still within the bounds of reasonable deviation for a series. These last two types deserve further exploration.

2.3.0.2 Multi-Source

A more effective strategy for assessing data completeness requires a better understanding of how data is being collected and loaded. In the case of multi-source data, one single source stopping loading may not be a big enough change to disrupt aggregate counts but could still jeopardize meaningful analysis. It would be more useful to conduct completeness checks by *subgroup* to identify these discrepancies.

But not any subgroup will do; the subgroup must correspond to the various data sources. For example, suppose we run an e-commerce store and wish to look at sales from the past month by category. Naturally, we might think to check the completeness of the data by category. But what if sales data is sourced from three separate locations: our Shopify site (80%), our Amazon Storefront (15%), and phone sales (5%). Unless we explicitly check completeness by channel (a dimension we don't particularly care about for our analysis), it would be easy to miss if our data source for phone sales has stopped working or loads at a different frequency.

Another interesting aspect of multi-source data, is multiple sources can contribute either to different *rows/records* or different *columns/variables*. Table-level frequency counts won't help us in the latter case since other sources might create the right total number of records but result in some specific fields in those records being missing or inaccurate.

2.3.0.3 Partial Loads

Partial loads really are not data errors at all, but are still important to detect since they can jeopardize an analysis. A common scenario might occur if a job loads new data every 12 hours (say, data from the morning and afternoon of day n-1 loads on day n at 12AM and 12PM, respectively). An analyst retrieving data at 11AM may be concerned to see an approximate ~50% drop in sales in the past day, despite confirming that their data looks to be "complete" since the maximum record date is, in fact, day n-1. Of course, this concern could be somewhat easily allayed if they then checked a timestamp field, but such a field might not exists or might not have been used for validation since its harder to anticipate the appropriate maximum timestamp than it is the maximum date.

2.3.0.4 Delayed or Transient Records

The interaction between choices made in the data collection and data loading phases can introduce their own sets of problems.

Consider an `orders` table for an e-commerce company that analysts may use to track customer orders. It might contain one record per `order_id` x `event` (placement, processing, shipment), one record per order placed, one record per order shipping, or one record per order with a `status` field that changes over time to denote the order's current stage of life. Some of these options are illustrated in Figure 2.6.

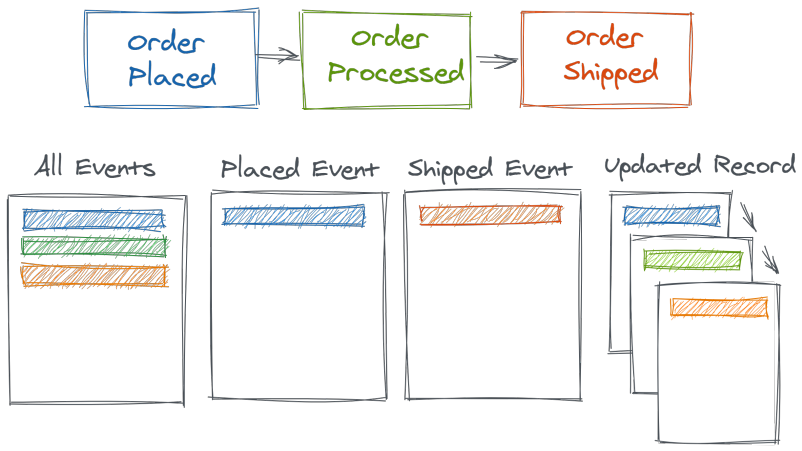


FIGURE 2.6: Illustration of alternative data collection and extraction strategies for order data

Any of these modeling choices seem reasonable and the difference between

them might appear immaterial. But consider the *collection* choice to record and report *shipped* events. Perhaps this might be operationally easier if shipment come from one source system whereas orders could come from many. However, an interesting thing about shipments is that they are often lagged in a variable way from the order date.

Suppose the e-commerce company in question offers three shipping speeds at checkout. Figure 2.7 shows the range of possible shipment dates based on the order dates for the three different speeds (shown in different bars/colors).

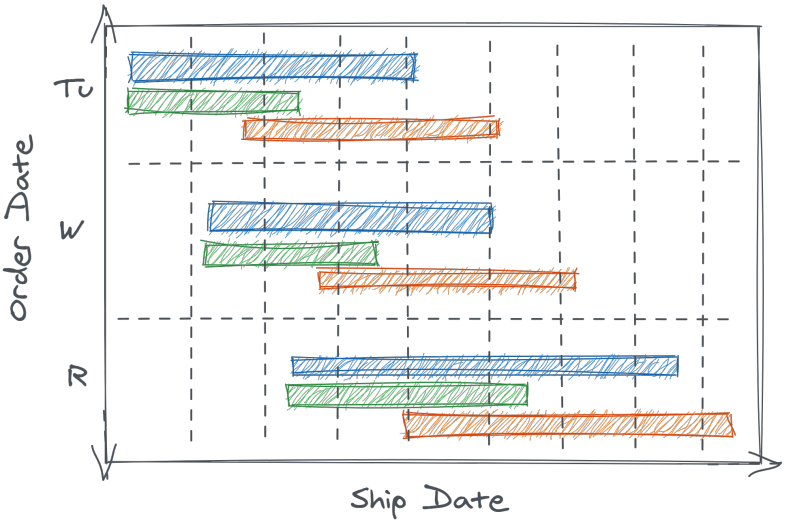


FIGURE 2.7: A conceptual chart of when different classes of real-world events might materialize as records in our dataset

How might this effect our perceived data quality?

- Order data could appear **stale** or not timely since orders with a given `order_date` would only load days later once shipped
- Similar to **missing** or **multisource** data, the data *range* in the table could lead to deceptive and incomplete data validation because some orders from a later order date might ship (and thus be logged) before all orders from a previous order date
- Put another way, we could have multiple order dates demonstrating **partial** data loads
- These features of the data might behave inconsistently across time due to seasonality (e.g. no shipping on weekends or federal holidays), so heuristics developed to clean the data based on a small number of observations could fail

- From an analytical perspective, orders with faster shipping would be disproportionately overrepresented in the "tail" (most recent) data. If shipping category correlated with other characteristics like total order spend, this could create an artificial trend in the data

Once again, understanding that data is *collected* at point of shipment and reasoning how shipment timing varies and impacts *loading* is necessary for successful validation.

If this thought experiment seems to vague, we can make it more concrete by mocking up a dataset with which to experiment.

In the simplest version, we will simply suppose one order is submited on each of 10 days with dates (represented for convenience as integers and not calendar dates) given by the `dt_subm` vector. Suppose shipping always takes three days, so we can easily calculate the shipment date (`dt_ship`) based on the submission date. The shipment date is the same as the date the data will be logged and loaded (`dt_load`).

```
# data simulation: single orders + deterministic ship dates ----
dt_subm <- 1:10
days_to_ship <- 3
dt_ship <- dt_subm + days_to_ship
dt_load <- dt_ship
df <- data.frame(dt_subm, dt_ship, dt_load)
head(df)
```

```
##   dt_subm dt_ship dt_load
## 1       1       4       4
## 2       2       5       5
## 3       3       6       6
## 4       4       7       7
## 5       5       8       8
## 6       6       9       9
```

Suppose we are an analyst living in day 5 and wonder how many orders were submitted on day 3. We can observe all shipments loaded before day 5 so we filter our data accordingly. However, when we count how many records exist for day 3 we find none. Instead, when we move ahead to an analysis date of day 7, we are able to observe the orders submitted on day 3.

```
library(dplyr)

# how many day-3 orders do we observe as of day-5? ----
df %>%
  filter(dt_load <= 5) %>%
  filter(dt_subm == 3) %>%
  nrow()
```

```
## [1] 0
```

```
# how many day-3 orders do we observe as of day-7? ----
df %>%
  filter(dt_load <= 7) %>%
  filter(dt_subm == 3) %>%
  nrow()
```

```
## [1] 1
```

(Note that these conditions could be checked much more succinctly with a base R expression such as `sum(df$dt_load < 7 & df$dt_subm == 3)`. However, there is sometimes virtue in option for more readable code even if it is less compact. Here, we prefer the more verbose option for the claritfy of our exposition. Such trade-offs, and general thoughts on coding style, are explored further in Chapter 10.)

Now, this may seem to trivial. Clearly, if there were *zero* records for a day, we would catch this in data validation, right? We can make our synthetic data slightly more realistic to better illustrate the problem. Let's not imagine that there are 10 orders each day, and each order is shipped sometime between 2 and 4 days after the order with equal probability.

```
# data simulation: multiple orders + random ship dates ----
dt_subm <- rep(x = 1:10, each = 10)
days_to_ship <- sample(x = 2:4, size = length(dt_subm), replace = TRUE)
dt_ship <- dt_subm + days_to_ship
dt_load <- dt_ship
df <- data.frame(dt_subm, dt_ship, dt_load)
head(df)
```

```
##   dt_subm dt_ship dt_load
## 1       1       4       4
## 2       1       3       3
## 3       1       4       4
## 4       1       5       5
## 5       1       3       3
## 6       1       4       4
```

When we repeat the prior analysis, we now see that we have *some* records for orders submitted on day 3 by the time we begin analysis on day 5. In this case, we might be more easily tricked to believe this is *all* orders. However, when we repeat the analysis on day 7, we see the the number of orders on day 3 has increased.

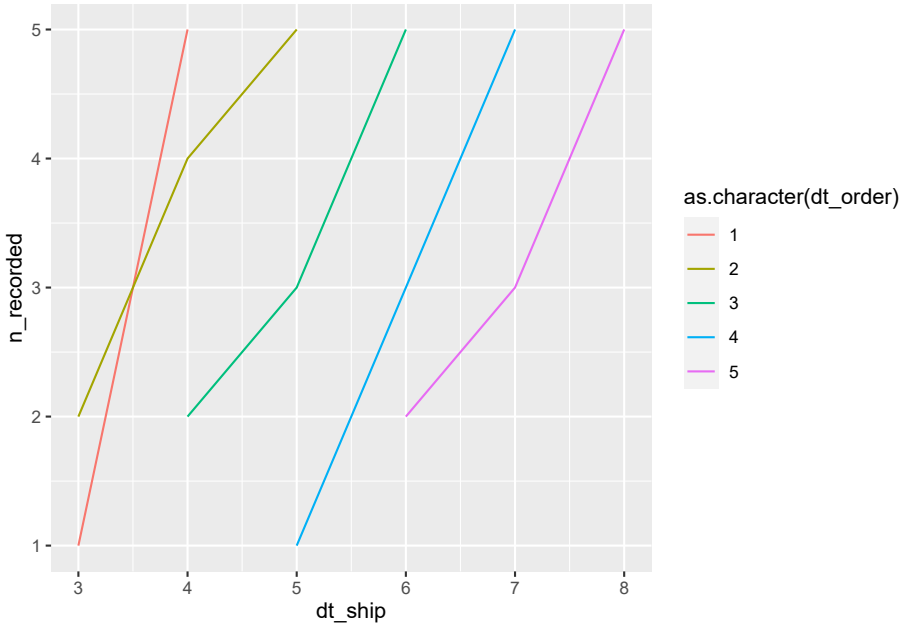
Of course, you can imagine the real world is yet much more complicated than this example. In reality, we would have a random number of orders each day. Additionally, we might have a *mixture* of different *types* of orders. There might be high-priced orders where customers tended to be willing to pay for faster shipping, and low-priced orders where customers tend to chose slower shipping. In a case like this, not only might naive validation miss the lack of data completeness, but the *sample* of shipments we begin to see on day 5 could be unrepresentative of the population of orders placed on day 3. This is a type of **selection bias** that we will examine further in Chapter 6 (Incredible Inferences).

2.4 Data Encoding, Modeling, & Transformation (TODO)

2.5 Strategies (TODO)

```
# but in reality this is just a data loading/recording issue ----
df %>%
  group_by(dt_order, dt_ship) %>%
  count() %>%
  group_by(dt_order) %>%
  arrange(dt_ship) %>%
```

```
  mutate(n_recorded = cumsum(n)) %>%
  ggplot(aes(x = dt_ship, y = n_recorded, col = as.character(dt_order))) + geom_line()
```



2.5.1 Data Transformation

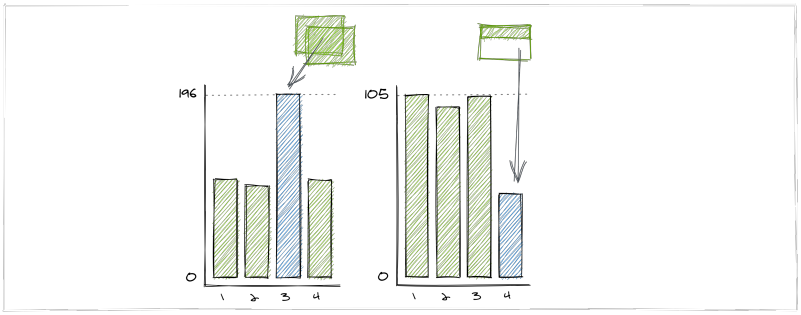
Finally, once the data is roughly where we want it, it likely undergoes many transformations to translate all of the system-generated fields we discussed in data collection into semantically-relevant dimensions for analytical consumers. Of course, the types of transformations that could be done are innumerable with far more variation than data loading. So, we'll just look at a few examples of common failure patterns.

2.5.1.1 Pre-Aggregation

Data transformations may include aggregating data up to higher levels of granularity for easier analysis. For example, a transformation might add up item-level purchase data to make it easier for an analyst to look at spend per *order* of a specific user.

Data transformations not only transform our data, but they also transform how the dimensions of data quality manifest. If data with some of the

completeness or **uniqueness** issues we discussed with data loading is pre-aggregated, these problems can turn into problems of **accuracy**. For example, the duplicate or partial data loads that we discussed when aggregated could suggest inaccurately high or low quantities respectively.



2.5.1.2 Field Encoding

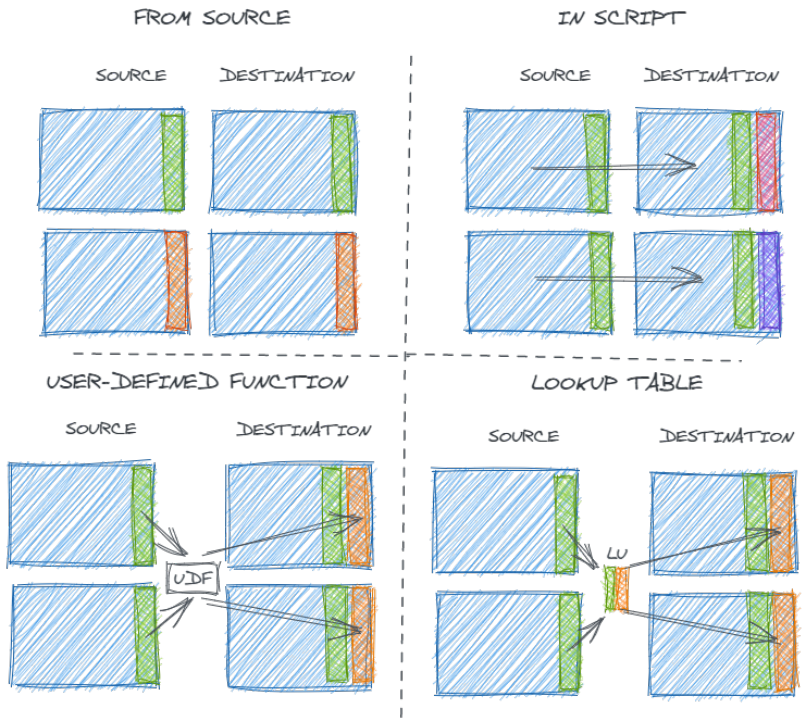
When we assess data consistency across tables,

Categorical fields in a data set might be created in any number of ways including:

- Directly taken from the source
- Coded in a transformation script
- Transformed with logic in a shared user-defined function (UDFs[13]) or macro[14]
- Joined from a shared look-up table

Each approach has different implications on data consistency and usability.

[13] https://docs.snowflake.com/en/sql-reference/user-defined-functions.html
[14] https://docs.getdbt.com/docs/building-a-dbt-project/jinja-macros/#macros



Using fields from the source simply is what it is – there's no subjectivity or room for manual human error. If multiple tables come from the same source, it's likely but not guaranteed that they will be encoded in the same way.

Coding transformations in the ELT process is easy for data producers. There's no need to coordinate across multiple processes or use cases, and the transformation can be immediately modified when needed. However, that same lack of coordination can lead to different results for fields that should be the same.

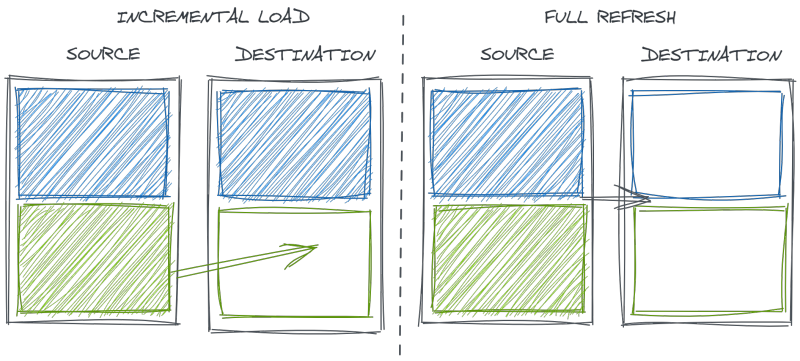
Alternatively, macros, UDFs, and look-up tables provided centralized ways to map source data inputs to desired analytical data outputs in a systemic and consistent way. Of course, centralization has its own challenges. If something in the source data changes, the process of updating a centralized UDF or look-up table may be slowed down by the need to seek consensus and collaborate. So, data is more *consistent* but potentially less *accurate*.

Regardless, such engineered values require scrutiny – paticularly if they are being used as a key to join multiple tables – and the distinct values in them should be carefully examined.

2.5.1.3 Updating Transformations

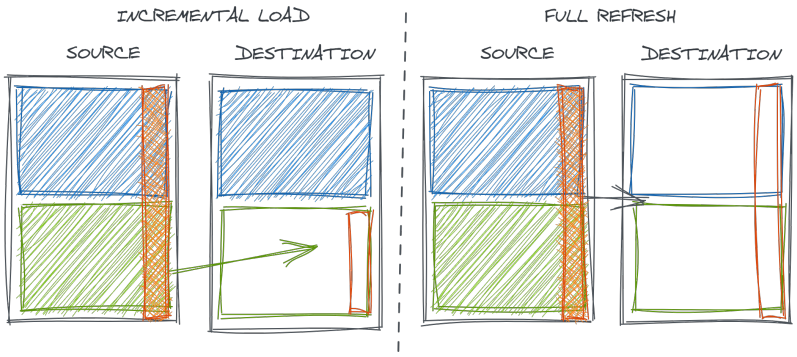
Of course, data consistency is not only a problem across different data sources but within one data source. Regardless of the method of field encoding used in the previous step, the intersection of data loading and data transformation strategies can introduce data consistency errors over time.

Often, for computation efficiency, analytical tables are loaded using an *incremental* loading strategy. This means that only new records (determined by time period, a set of unique keys, or other criteria) from the upstream source are loaded to the downstream table. This is in contrast to a *full refresh* where the entire downstream table is recreated on each update.

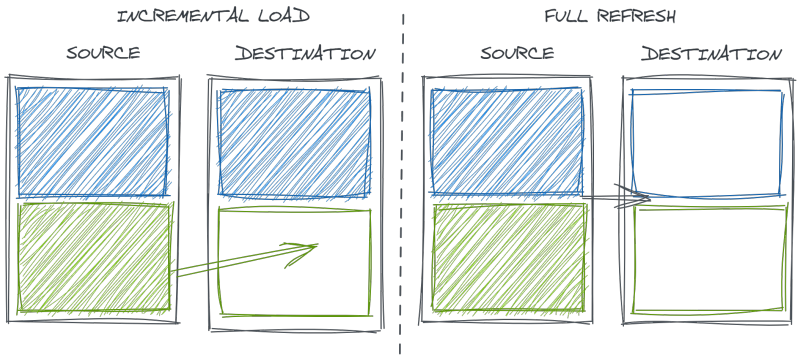


Incremental loads have many advantages. Rebuilding tables in entirety can be very time consuming and computationally expensive. In particular, in non-cloud data warehouses that are not able to scale computing power on demand, this sort of heavy duty processing job can noticeably drain resources from other queries that are trying to run in the database. Additionally, if the upstream staging data is ephemeral, fully rebuilding the table could mean failing to retain history.

However, in the case that our data transformations change, incremental loads may introduce inconsistency in our data overtime as only new records are created and inserted with the new logic.



This is also a problem more broadly if some short-term error is discovered either with data loading or transformation in historical data. Incremental strategies may not always update to include the corrected version of the data.



Regardless, this underscores the need to validate entire datasets and to re-validate when repulling data.

2.6 Human-Generated Data

Ahhhh

2.7 Other Encoding Issues

- for indicators which is 1?
- field values changing over time

2.8 Strategies

–>

3

Computational Quandaries

After gaining confidence in one's data (or, at least, making peace with it), the next step in a data analysis is often to start cleaning and exploring that data with summary statistics, plots, and models. Generally, this requires a computational tool like SQL, R, or python.

The process of computation itself can be fraught with challenges. Computational tools are extremely literal; they are excellent at doing *precisely what they were told to do* but not often what analysts might have *meant* or *wished* that they would do. Additionally, the moment an analyst begins to use a tool, the conversation is no longer between them and the data; suddenly, the mental model of how every single tool developer thought you might want to do analysis affects the tools' behaviors and the analysts' results.

In this chapter, we will explore common ways that tools may do something technically correct, reasonable, and as-intended but very much not what analysts may expect. Along the way, we will see how computational methods interact with the data encoding choices we discussed in Chapter 2 (Data Dalliances).

3.1 Preliminaries - Data Computation

Before we think about specific tools or failure modes, we can first consider the common types of operations that the analytical tools allow us to do with our data.

3.1.1 Single Table Operations

Given a single data table, we may wish to do operations (illustrated in Figure 3.1) such as:

- **Filtering**: Extracting a subset of a dataset for analysis based on certain inclusion criteria for each record

- **Aggregation**: Grouping our data table by one or more variables and condensing information across records with *aggregate functions* like counts, sums, and averages
- **Transformation**: Create new columns or modifying existing columns to represent more complex or domain-specific context

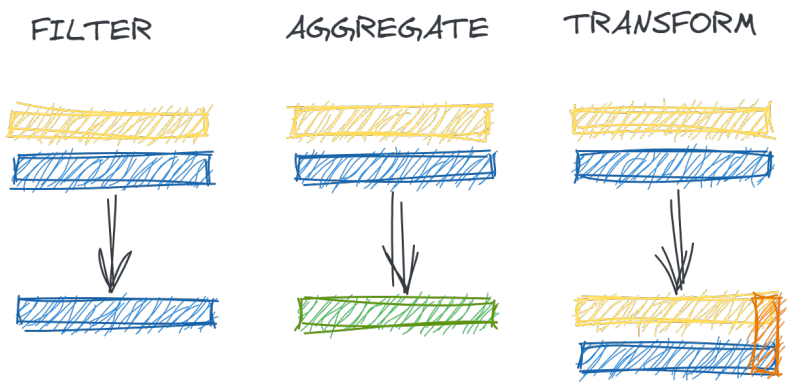


FIGURE 3.1: Illustration of basic single-table data wrangling operations

3.1.2 Multiple Table Operations

Often, we can get additional value in an analysis by combining multiple types of information from difference tables. When working with multiple tables, we may be interested in:

- **Combining Row-wise**: Taking multiple tables with the same schemas (column names and data types) and creating a single table which contains the union (all records), intersection (only matching), or difference (only in one) of the records in the two tables
- **Combining Column-wise**: Appending additional fields to existing records through joining (also known as merging) multiple tables

3.1.3 Mechanics

All of these operations rely on a few core computational tasks:

- **Arithmetic**: Basic addition, subtraction, multiplication, and division to aggregate and transform data

- **Equality**: Comparing whether or not two values are equal is critical for data filtering, column-wise combination, and certain types of data transformation
- **Casting**: Converting data types of different elements into a comparable format is necessary for row-wise combination and often a prerequisite to certain equality and arithmetic tasks

While these operations may seem simple, their behavior within certain tools and when employed for certain data types may sometimes lead to unintuitive or misleading results.

3.2 Null Values

In Chapter 2 (Data Dalliances), we discuss how null values may represent many different concepts and be encoded in multiple different ways. In addition to those semantic challenges, various representations of null values may cause different computational problems.[1] In this section, we will explore these potential failure modes.

3.2.1 Types of Null Values

Not only can null values represent many different things (as explored in Chapter 2), they also may be represented in many different ways. Understanding how nulls are encoded in one's dataset is a critical prerequisite to attempting any of the computations described in the subsequent sections.

3.2.1.1 Language representations

Different programming languages each offer their own versions of null values – and sometimes more than one. For example, the R language includes `NA`, typed `NA`s (e.g. `NA_integer`, `NA_character`), `NaN`, and `NULL`; meanwhile, core python has `None` and the `numpy` module provides a `nan`.

These different values carry different semantic and functional meanings. For example R's `NA` generally means "the presence of an absence" whereas `NULL` is "the absence of a presence". This is articulated more clearly if we examine

[1] This problem is not isolated to data analysis tools. For an entertaining example, see the 2019 WIRED article "How a 'NULL' License Plate Landed One Hacker in Ticket Hell" (Barrett, 2019) which a real-world software system producing unintended and undesirable behavior when asked to deal with a word `'NULL'`.

the lengths of these objects and observe that `NA` has a length 1 whereas `NULL` has a length 0.

```
c(length(NA), length(NULL))
```

```
## [1] 1 0
```

As further proof that these are not interchangeable, we may use the helper functions `is.na()` and `is.null()`. It's false that `NA` is `NULL` and essentially unevaluable if `NULL` is `NA` because `NULL`s are truly nothing.

```
c(
  is.na(NA),
  is.null(NULL),
  is.na(NULL),
  is.null(NA)
)
```

```
## [1]  TRUE  TRUE FALSE
```

To further complicate matters, we have `NaN` ("not a number"), along with `-Inf` and `Inf`, which generally arise when we attempt to abuse R's calculator. Somewhat charmingly, `Inf` and `-Inf` may be used in some rudimentary calculations where the limit is returned.[2]

```
c(
  1/0,   # returns Inf
  0/0,   # returns NaN
  1/Inf  # returns 0
)
```

```
## [1] Inf NaN   0
```

3.2.1.2 Data encoding choices (TODO)

Beyond these null types offered natively by different programming languages, there are also many different data management *conventions* for null values.

[2] From calculus, we know 1/Inf approaches 0, but Inf/Inf is undefined.

Because null values can have many meanings, sometimes missing fields are encoded with "out of range" values which intend to suggest a type of missingness.

For example, the US Census Bureau's Medical Expenditure Panel Survey uses the following reserved codes to denote different types of missingness: (TODO: cite p10 https://www.meps.ahrq.gov/data_stats/download_data/pufs/h206a/h206adoc.pdf)

```
- -1 INAPPLICABLE Question was not asked due to skip pattern
- -7 REFUSED Question was asked and respondent refused to answer question
- -8 DK Question was asked and respondent did not know answer
- -14 NOT YET TAKEN/USED Respondent answered that the medicine has not yet been used
- -15 CANNOT BE COMPUTED Value cannot be derived from data
```

This approach preserves a lot of relevant information while, at the same time, being readily apparent that these values are not valid when the data is manually inspect. Unfortunately, manually inspecting every data field is rarely possible, and such sentinel values may go undetected when looking at higher-level summaries.

Consider a survey of a population of retired adults where age is coded as `999` if not provided. Below, we simulate 100,000 such observations that are uniformly distributed between the age of 65 and 95 (hence, have an expected value of 80). Next, we replace merely *half of a percent* with our "null" values of `999`. Taking the mean with these false values results in a mean of about 85. This number alone might not raise the alarm; after all, we know the dataset's population is older adults. However, accidentally treating these as valid values biases our results by a somewhat remarkable five years.

```
set.seed(123)

n <- 100000
p <- 0.005
ages <- runif(n, 65, 95)

ages_nulls <- ages
ages_nulls[1:(n*p)] <- 999

c(mean(ages), mean(ages_nulls))
```

```
## [1] 79.98 84.57
```

So, the first order of business with null values is understanding how they are

encoded and translation them to the most computationally appropriate form. However, that is only the beginning of the story.

3.2.2 Aggregation

How null values are handled in the simple aggregation of data varies both across different languages and across different functions within a language. To better understand the problems this might cause, we will look at examples in R and SQL.

To explore aggregation, let's build a simple dataset. We will suppose that we are working with a subscription-based e-commerce service and that we are looking at a `monthly_spend` dataset with one record per customer and information about the amount they spent and returned in a given month:

```
spend <-
  data.frame(
    AMT_SPEND = c(10, 20, NA),
    AMT_RETURN = rep(NA, 3)
  )
```

To compute the average amount spent (`AMT_SPEND`) with the `dplyr` package, an analyst might first reasonably write the following `summarize()` statement. However, as we can see, due to the presence of null values within the `AMT_SPEND` column, the result of this aggregation is for the whole quantity of `AVG_SPEND` to be set to the value `NA`.

A glance at the documentation for the `mean()` function reveals that it has an `na.rm` parameter which, when set to true, removes null values from our dataset. Adding this argument to the previous statement allows us to reach a numerical answer.

```
summarize(spend,
          AVG_SPEND = mean(AMT_SPEND),
          AVG_SPEND_NARM = mean(AMT_SPEND, na.rm = TRUE))
```

```
##   AVG_SPEND AVG_SPEND_NARM
## 1        NA             15
```

However, is this the *right* numerical answer? What `na.rm = TRUE` does is *drop* the null values from the set of numbers being averaged. However, suppose the

null values represent that no purchases were made. That is, zero dollars were spent. In effect, we have removed all non-purchasers from the data being averaged.

More precisely, we have switched from taking the average

$$\frac{\sum_1^n Spend}{\sum_1^n 1}$$

over all n customers

to taking the average

$$\frac{\sum_{Spend>0} Spend}{\sum_{Spend>0} 1}$$

over only those customers with spend

At face value, we could say that the code above is giving the incorrect answer; by dropping some low (zero) purchase amounts, the average amount spend per customer is inflated. A second perspective, which is someone more philosophically troubling, is that this tiny change to the code which fixed the *obvious* problem (returning a null value) has introduced a *non-obvious* problem by fundamentally changing the question that we are asking. By dropping all accounts from our table who made no purchases, we are no longer answering "What is the average amount spent by a new registrant?" but rather "What is the average amount spent by an actively engaged customer?" This technical quirk has significant analytical impact.

To answer the real question at hand, we would instead have a couple of options. We could manually `sum()` the amount spent with the option to drop nulls but then divide by the correct denominator (all observations – not just those with spend) or we could explicitly recode null values in `AMT_SPEND` to zero before taking the average.[3] Either of these options lead to the correct conclusion of a lower average spend amount.

```
summarize(
    spend,
    AVG_SPEND_MANUAL = sum(AMT_SPEND, na.rm = TRUE) / n(),
    AVG_SPEND_RECODE = mean(coalesce(AMT_SPEND, 0))
  )
```

[3] Recoding can be done with a number of different general purpose functions like `ifelse` or `dplyr::case_when` in R. Different SQL varaints often offer different options for this purpose with functions such as `nvl()` or `zeroifnull()`. A common version across many platforms is `coalesce()` which takes the first non-null argument listed.

```
##   AVG_SPEND_MANUAL AVG_SPEND_RECODE
## 1               10               10
```

This is all well and good if we could just accept that the behaviors above are simply how nulls work, but further complexity comes as we see that there is no industry standard across tools. For example, as the SQL code below shows, SQL's `avg()` function behaves more like R's `mean()` *with* the `na.rm = TRUE` option set. That is, the default behavior of SQL is to only operate on the valid and available values.

```
SELECT avg(amt_spend) as AVG_SPEND
FROM spend
```

```
##   AVG_SPEND
## 1        15
```

However, this is not to suggest that null values cannot also be destructive in SQL. While aggregation functions (which compute over the *rows/records*) like `sum()` and `avg()` drop nulls, operators like `+` and `-` (which compute *across columns/variables* in the *same row/record*) do not exhibit the same behavior. Consider, for example, if we wish to calculate the average net purchase amount (purchases minus returns) instead of the gross (total) purchase amount.

```
SELECT avg(amt_spend-amt_return) as AVG_SPEND_NET
FROM spend
```

```
##   AVG_SPEND_NET
## 1            NA
```

Despite what we learned above about SQL's `avg()` function, the query above returns only a null value. What has happened? In our `registration` data set, the `amt_return` column is completely null (representing no returns). Because the subtraction occurs before the average is taken, subtracting real numbers in `amt_spend` with null values in `amt_return` creates a column of all null values which are then fed into the `avg()` function. This process is shown step-by-step below.

```
SELECT
  amt_spend,
  amt_return,
  amt_spend-amt_return
FROM spend
```

```
##   AMT_SPEND AMT_RETURN amt_spend-amt_return
## 1        10         NA                    NA
## 2        20         NA                    NA
## 3        NA         NA                    NA
```

3.2.3 Comparison

Null values don't just introduce complexity when doing arithmetic. Difficulties also arise any time multiple variables are assessed for equality or inequality. Since a null value is unknown, most programming languages generally will *not* consider nulls to be comparable with other nulls.

We can simple examples of this in both R and SQL.

```
c(
  NA == 3,
  NA > 10,
  NA == NA
  )
```

```
## [1] NA NA NA
```

```
SELECT
  (NULL = 3) as NULL_EQ_NUM,
  (NULL > 10) as NULL_GT_NUM,
  (NULL = NULL) as NULL_EQ_NULL
```

```
##   NULL_EQ_NUM NULL_GT_NUM NULL_EQ_NULL
## 1          NA          NA           NA
```

In these toy examples, such outcomes may seem perfectly logical. However, this same reasoning can arise in sneakier ways and lead to uninteded results when equality evaluations are *implicit* in the task at hand instead of the singular focus. We'll now see examples from data filtering, joining, and transformation.

3.2.3.1 Filtering

Suppose we want to split our dataset into two datasets based on high or low values of spend. We might assume the following two lines of code will create a clear partition[4] of results.

```
spend_lt20 <- filter(spend, AMT_SPEND < 20)
spend_gte20 <- filter(spend, AMT_SPEND >= 20)
```

However, examining the resulting datasets, we see than *neither* contains the null records.

```
spend_lt20
```

```
##   AMT_SPEND AMT_RETURN
## 1        10         NA
```

```
spend_gte20
```

```
##   AMT_SPEND AMT_RETURN
## 1        20         NA
```

The same situation results in SQL.

```
SELECT *
FROM spend
WHERE AMT_SPEND < 20
```

```
##   AMT_SPEND AMT_RETURN
## 1        10         NA
```

[4] A **paritition** of our data would imply that every record is contained in precisely one group

```
SELECT *
FROM spend
WHERE AMT_SPEND >= 20
```

```
##   AMT_SPEND AMT_RETURN
## 1        20         NA
```

Thus, whenever our data has null values, the very common act of data filtering risks excluding important information.

3.2.3.2 Joining

The same phenomenon as described above also happens when joining multiple datasets.

Suppose we have multiple datasets we wish to merge based on columns denoting a record's name and date of birthday. For ease of exploration, we will make the simplest possible such dataset and simply try to merge it to itself. (This may seem silly, but often when trying to understand *computationally* complex things, it is a good idea to make the scenario as simple as possible. In fact, this idea is core to the concept of computational unit tests which we will discuss at the end of this chapter.)

```
bday <- data.frame(NAME = c('Anne', 'Bob'), BIRTHDAY = c('2000-01-01', NA))
bday
```

```
##   NAME   BIRTHDAY
## 1 Anne 2000-01-01
## 2  Bob       <NA>
```

In SQL, if we try to join this table, the records in row 1 will match because `'Anne' == 'Anne'` and `'2000-01-01' == '2000-01-01'`. However, poor Bob's record is eliminated because his birthdate is logged as null, and `NA == NA` is false.

```
SELECT a.*
FROM
  bday as a
  INNER JOIN
```

```
  bday as b
  ON
  a.NAME = b.NAME and
  a.BIRTHDAY = b.BIRTHDAY
```

```
##   NAME   BIRTHDAY
## 1 Anne 2000-01-01
```

In contrast, R's `dplyr::inner_join()` function will not do this by default. This function lets us specifically control how nulls are matches with the `na_matches` argument, with a default option to match on `NA` values. (You may read more about the argument by typing `?dplyr::inner_join` in the R console to pull up the documentation.)

```
inner_join(bday, bday, by = c('NAME', 'BIRTHDAY'))
```

```
##   NAME   BIRTHDAY
## 1 Anne 2000-01-01
## 2  Bob       <NA>
```

This example then is not only a cautionary tale for how null values may unintentionally corrupt our data transformations but also how "brittle" our knowledge and intuition may be when moving between tools. Neither of these default behaviors is strictly better or worse, but they are definitely different and have real implications on our analysis.

3.2.3.3 Transformation

A common task in data analysis is to aggregate results by subgroup. For example, we might want to summarize how many customers (rows/records) spent more or less than $10. To discern this, we might create a categorical variable for high versus low purchase amounts, group by this variable and count.

The psuedocode would read something like this:

```
data %>%
  mutate(HIGH_LOW = << transform AMT_SPEND >>) %>%
  group_by(HIGH_LOW) %>%
  count()
```

To define the `HIGH_LOW` variable, we might use a function like `ifelse()`, `dplyr::if_else()`, or `dplyr::case_when()`. However, once again, we have the issue of how values are *partitioned* when nulls are included. If we recode any records with `AMT_SPEND` of less than or equal to 10 to "Low" and default the rest to "High", we will accidentally count all null values in the "High" group.

```
spend %>%
  mutate(HIGH_LOW = case_when(
    AMT_SPEND <= 10 ~ "Low",
    TRUE ~ "High")
    ) %>%
  group_by(HIGH_LOW) %>%
  count()
```

```
## # A tibble: 2 x 2
## # Groups:   HIGH_LOW [2]
##   HIGH_LOW     n
##   <chr>    <int>
## 1 High         2
## 2 Low          1
```

Instead, it is more accurate and transparent (unless we know specifically what null values mean and what group they should be part of) to not let one of our "core" categories by the "default" case in our logic. We can explicitly encode any residual values as something like "OTHER" or "ERROR" to help us see that there is a problem requiring extra attention.

```
spend %>%
  mutate(HIGH_LOW = case_when(
    AMT_SPEND <= 10 ~ "Low",
    AMT_SPEND > 10 ~ "High",
    TRUE ~ "OTHER")
    ) %>%
  group_by(HIGH_LOW) %>%
  count()
```

```
## # A tibble: 3 x 2
## # Groups:   HIGH_LOW [3]
##   HIGH_LOW     n
##   <chr>    <int>
```

```
## 1 High         1
## 2 Low          1
## 3 OTHER        1
```

3.3 Dates and Times (TODO)

3.4 Strings (TODO)

3.5 Encoding Choices (TODO)

3.6 Order of Operations (TODO)

3.7 Object References (TODO)

3.8 Trusting Tools

A theme throughout this book is the fundamentally *social* nature of data analysis. Data analysis is fraught without understanding the countless decisions made along the way by those who generated it (whose data is reflected), those who collected it, those who migrated it, and those who have posed questions of it. On one hand, this is a beautiful aspect of analysis; on the other hand, it means that analysts and their analyses are subject to all of the cognitive and social psychological biases of everyday humans.

One such bias is "social proof": assuming that if a tool behaves a certain way, it must be because it is correct.

Assuming that our tools know best is admittedly an attractive proposition. It appeals to a desire to think that someone, somewhere is "in charge" and, perhaps more critically, helps us avoid a domino effect of distrust (If we *don't* trust our tools how can we trust our results? And if we can't trust our results, how can we trust anything at all?) Unfortunately, there are many reasons are tools might not know best. For example, the tool's developer might have:

- Made a mistake
- Had a different analysis problem in mind with a different optimal approach
- Been optimizing for a different constraint (e.g. explainability vs. accuracy, speed vs. theoretical properties)
- Come from a community with different norms
- Been affording users the flexibility to do things many ways even if they don't agree
- Built a certain feature for a different purpose than how you are using it
- Not thought about it at all

As a few concrete examples from popular open source tools. We'll look briefly at the prominent python library `scikitlearn` for machine learning and Apache Spark, an engine for large-scale distributed data processing.

3.8.1 Defaults in `scikitlearn`

`scikitlearn`'s default behavior for logistic regression modeling[5] automatically applies L2 regularization. You might or might not know what this means, and you might or might not want to apply it to your problem. That's fine. The important thing is that it *will* change your estimates and predictions, and it is *not* a part of the classical definition of that algorithm (for modelers coming from a statistical background.)

Of course, there's nothing inherently wrong about this choice; the library authors just had different goals than a typical statistical. `scikitlearn` developer Olivier Grisel explains on Twitter[6] that this choice (and others in the library) is explained because "Scikit-learn was always designed to make it easy to get good predictive accuracy (eg as measured by CV) rather than as statistical inference library." Additionally, this choice is documented in bold in the function documentation[7].

However, an analyst could easily miss this nuance if they do not *read* the documentation. Or, if they *misinterpret* this choice as social proof that regularization is always the right approach, they might not make the best choice for their own analysis.

[5] A classic modeling technique for predicting binary (yes/no) outcomes

[6] https://twitter.com/ogrisel/status/1167438229655773186?s=20

[7] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

3.8.2 Algorithms in `Spark`

For example, according to a 2015 Jira ticket[8], developers of Spark considered multiple methodologies they could use when adding the functionality to compute feature importance for a random forest. Ultimately, a core contributor advised against permutation importance due to its computational cost.

> It's high time we add this to MLlib, so I'm adding this to the 1.5 roadmap. Peter Prettenhofer If you are still interested in this, please feel free to take it. Or if others are interested, please comment on this JIRA.
>
> The initial API should be quite simple; I'm imagining a single method returning importance for each feature, modeled after what R or other libraries return.
>
> I think we should calculate importance based on the learned model. The permutation test would be nice in the future but would be much more expensive (shuffling data).

Clearly, no one wants a workflow that is too costly or timely to run. So, once again, there is no right or wrong. However, since every approach to feature importance has its own biases, pitfalls, and challenges in interpretation, it's a mistake for an end-user to not carefully understand which algorithm is used and why.

3.9 Strategies (TODO)

[8] https://issues.apache.org/jira/browse/SPARK-5133

Analysis

4

Egregious Aggregations

Once armed with an understanding of the data and tools available for analysis, a common start to analysis is exploring data with *aggregation*. At its heart, any sort of data analysis is the process of condensing raw data into something more manageable and useful while giving up as little of the information as possible.

Many elementary tools for this task are much better at the comprehension task than the preservation one. We learn rigorous assumptions to consider and validate when studying linear regression, but basic arithmetic aggregation presents itself as agnostic and welcome to any type of data. However, the underlying distributions of our variables and the relationships between them have a significant impact on the how informative and interpretable various summarizations are.

In this chapter, we will explore different ways that univariate and multivariate aggregations can be naive or uninformative.

4.1 Averages

4.1.1 Averaging skewed data

Arithmetic average versus colloquial meaning of average as "typical"

Skewed data

Multimodal data / mixture models

4.1.2 No "average" observation

In the previous section, the average represented a point in the relevant data *range* even if it was not perhaps the one most representative of a "typical" observation. We discussed how in some situations this quantity may be a

reasonable answer to certain types of questions and an aid for certain types of decisions.

However, when we seek an average *profile* over multiple variables, the problems of averages are further compounded. We may end up with a set of "average" summary statistics that are not representative of any part of our population.

To see this, let's assume we are working with data for a company with a subscription business model. We might be interested in profiling the age of each account (how long they have been a subscriber) and their activity (measured by amount spent on an e-commerce platform, files downloaded on a streaming service, etc.)

The following code simulates a set of observations: 80% of accounts are between 0 to 3 years in age and have an average activity level of 100 while 20% of accounts are older than 3 years in age and have an average activity level of 500. (Don't over-think the specific probability distributions lived here. We are concerned with interrogating the properties of the average and not with simulating a realistic data generating process. Giving ourselves permission to be wrong or "lazy" about unimportant things gives us more energy to focus on what matters.)

```
set.seed(123)

# define simulation parameters ----
## n: total observations
## p: proportion of observations in group 1
n <- 5000
p <- 0.8
n1 <- n*p
n2 <- n*(1-p)

# generate fake dataset with two groups ----
df <-
  data.frame(
    age = c(runif(n1,   0,  3), runif(n2,   3, 10)),
    act = c(rnorm(n1, 100, 10), rnorm(n2, 500, 10))
  )
```

Figure 4.1 shows a scatterplot of the relationship between account age (x-axis) and activity level (y-axis). Meanwhile, the marginal rug plots shows the univariate distribution of each variable. The sole red dot denotes the coordinates of the average age and average activity. Notably, this dot exists in a region of "zero density"; that is, it is not representative of *any* customer. Strategic decisions made with this sort of observation in mind as the "typical" might not be destined for success.

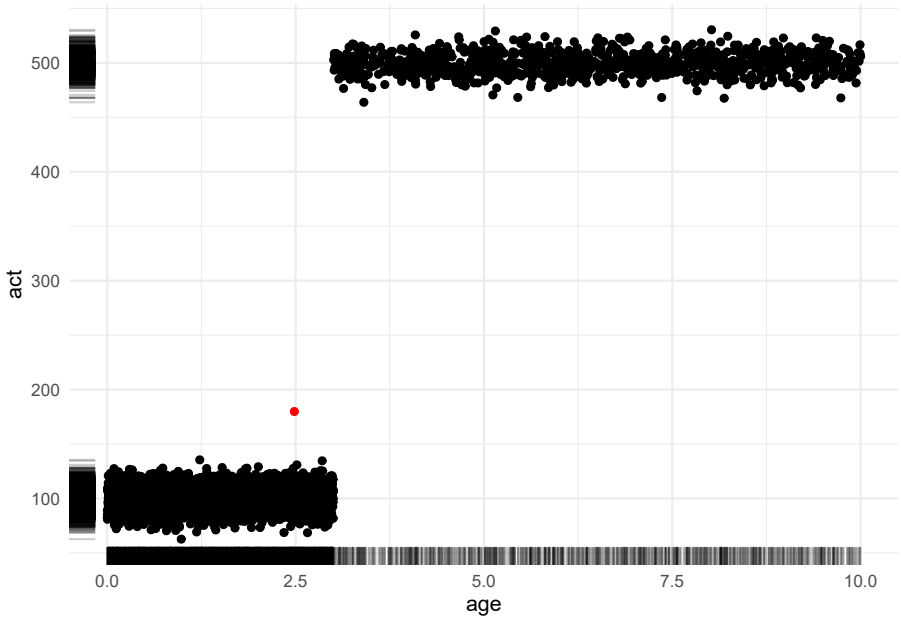


FIGURE 4.1: A scatterplot of two variables and their averages

4.1.3 The product of averages

As the above example shows, averages of multivariate data can produce poor summaries – particularly when these variables are interrelated[1].

A second implication of this observation is that deriving additional computations based on pre-averaged numbers is likely to obtain inaccurate results.

For example, consider that we wish to estimate the average dollar amount of returns per any e-commerce order. Orders may generally be a mixture of low-price orders (around $50 on average) and high-price orders (around $250 on average). Low-price orders may have a 10% probability of being returned while high price orders have a 20% probability. (Again, are these numbers, distributions, or relationships hyper-realistic? Not at all. However, once again we are telling ourselves a story just to reason about numerical properties, so we have to give ourselves permission to not focus on irrelevant details.)

```
set.seed(123)
```

[1] We intentionally avoid the word *correlated* here to emphasize the fact that *correlation* refers more strictly to linear relationships

```
# define simulation parameters ----
## n: observations per group
## pr[1|2]: mean price per group
n <- 100
pr1 <- 50
pr2 <- 250
pr_sd <- 5
re1 <- 0.1
re2 <- 0.2

# simulate spend amounts and return indicatiors ----
amt_spend  <- c(rnorm(n, pr1, pr_sd), rnorm(n, pr2, pr_sd))
ind_return <- c(rbinom(n, 1, re1),    rbinom(n, 1, re2))

# compute summary statistics ----
average_of_product <- mean(amt_spend * ind_return)
product_of_average <- mean(amt_spend) * mean(ind_return)
```

The *true* average amount returned across all of our orders is 36.0438 (from the `average_of_product` variable). However, if instead we already knew an average spend amount and an average return proportion, we might be inclined to compute the `product_of_average` method which returns a value of 26.9923. (This is a difference of 9.05 relative to an average purchase amount of 150.)

At first, this may seem unintuitive until we write out the formulas and realize that these metrics are, in fact, two very different quantities:

$$\frac{\sum_1^n Spend}{\sum_1^n 1} * \frac{\sum_1^n I(Return)}{\sum_1^n 1}$$

over all n orders

versus

$$\frac{\sum_1^n Spend * I(Return)}{\sum_1^n 1}$$

If this still feels counterintuitive, we can see how much of the difference is accounted for by the interrelation between our two variables. In the following code, we break the relationship between the variables by randomly reordering the `ind_return` variable so it is no longer has any true relationship to the corresponding `amt_spend` variable.

```
# randomly reorder one of two variables to break relationships ----
ind_return <- sample(ind_return, size = 200)

# recompute variables ----
average_of_product <- mean(amt_spend * ind_return)
product_of_average <- mean(amt_spend) * mean(ind_return)
```

After redoing the calculations, we find that th two values are much closer. `average_of_product` is now 24.1041 and `product_of_average` is now 26.9923. These are notably still not the same number so that does not mean that these two equations are equivalent if variables are unrelated; however, this second result once again illustrates the extent to which interrelations can defy our naive intuitions.

4.2 Ratios

4.2.1 Picking the right denominator

4.2.2 Sample size effects

4.3 Trends

4.3.1 "If trends continue..."

4.3.2 Seasonality

5

Vexing Visualization

6

Incredible Inferences

Previously, we have seen how different inputs like data, tools, and methods can add risks to our data analysis. However, the battle is not won simply when we get our first set of *outputs*. In this chapter, we will explore common errors in interpreting the results of our analysis by exploring aspects of bias, missingness, and confounding.

7

Cavalier Causality

In Chapter 6 (Incredible Inferences), we began to see that we can be tricked by biases when we lack *causal* thinking and an underlying theory for the data generating process. In this chapter, we will revisit some of these same disasters and introduce some specific frameworks to help us more rigorously explore our analysis for errors and biases and, even better, strategize the best ways to fix them.

8

Mindless Modeling

9

Alternative Algorithms

As the consummate showman, P.T. Barnum is often quoted as saying "Leave them wanting more". Unfortunately, statistics professors have less of a flare for drama. Introductory statistics courses will typically introduce a few types of models (for example, linear and *perhaps* logistic regression), and that's a wrap. It's often until students start *taking* the subsequent courses that they are exposed to the true limitations of previous techniques and taught to demand more.

This chapter attempts to flip that paradigm by briefly surveying a broad number of modeling techniques. The goal is not to go into all of the rigorous deals that one should understand to use these models. Instead, we hope to build a "mental toolbox" of techniques so that you know where to focus your study when you encounter a problem in the real world.

9.1 Modeling Binary Outcomes

9.2 Modeling Counts

9.3 Modeling Time Until an Event

9.4 Modeling Repeated Measures on a Population

9.5 Modeling Observations in a Nested Hierarchy

9.6 Modeling Time & Space Data

Workflow

10

Complexifying Code

11

Rejecting Reproducibility

A

Useful Data Generation Functions (TODO)

B

Common Probability Distributions (TODO)

Bibliography

Barrett, B. (2019). How a 'null' license plate landed one hacker in ticket hell. *WIRED*.

Bassa, A. (2017). Data alone isn't ground truth.

Hicks, S. C. and Peng, R. D. (2019). Evaluating the success of a data analysis.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software, Articles*, 59(10):1–23.