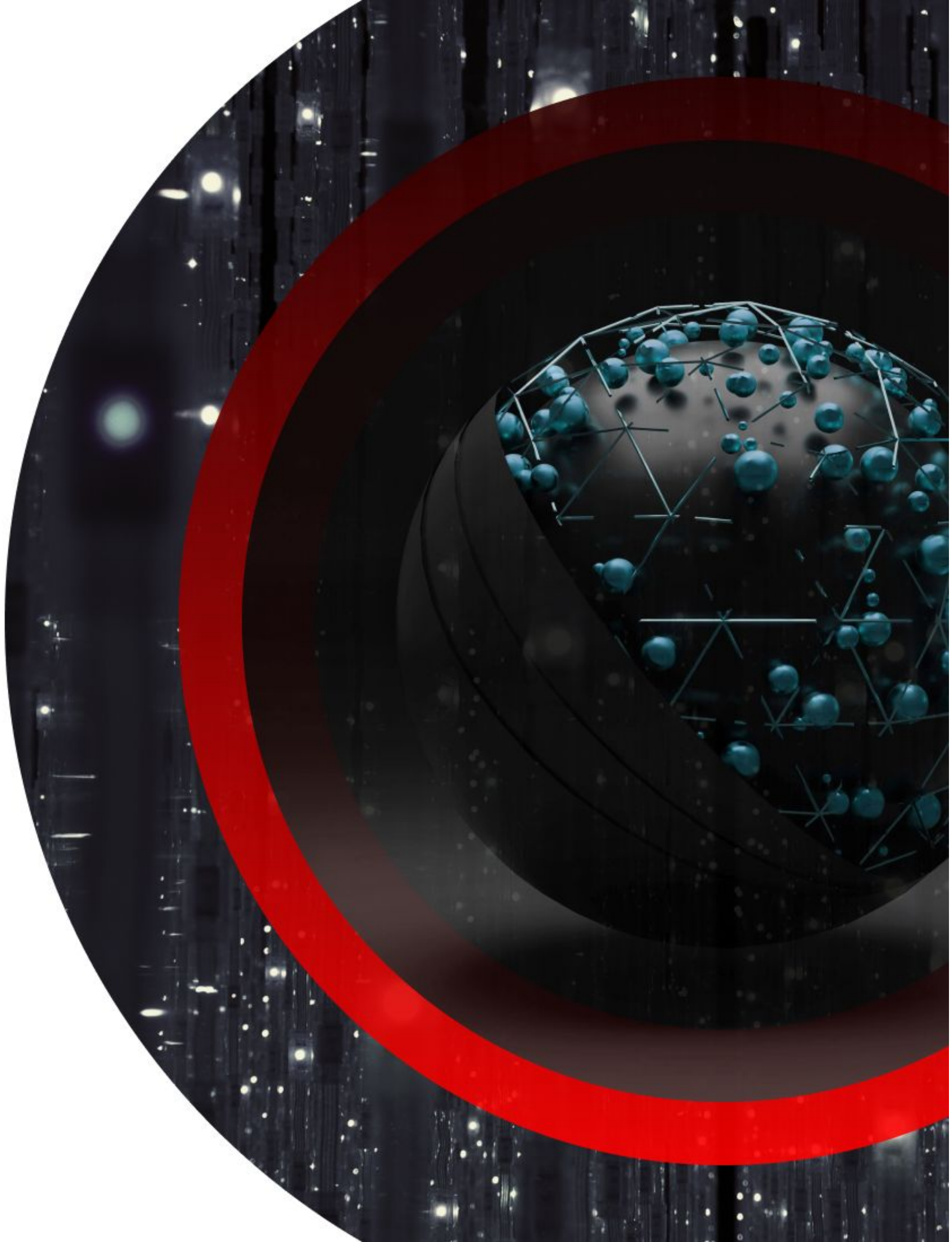
DATA SCIENCE AND ANALYTICS

Introductory Course



CLASS NORMS

COURTESY IN CLASS

Remaining on mute unless called on, exercising courtesy during breakout rooms, using the chat box for questions only

ATTENDANCE

100% attendance is expected and contributes to success in passing the course and the program

PARTICIPATION

Keeping an open mind in discussions and sharing experiences, making contributions during team assignments, submitting assignments in Canvas, and participating in discussion boards

USE OF CLASS RESOURCES

Follow along during the lecture with the lesson companion and download any in-class documents prior to class.

PROGRAM PATH

1 Introductory Course

2 SQL and Databases

3 Statistics and Probability

4 Data Storytelling

Milestone 1: Building and Presenting Data Stories

5 Python Programming

6 Data Wrangling

7 Visual Communications

8 Advanced SQL Programming

Milestone 2: Data Integration, Preparation, Reporting, and Presentation

9 Business Intelligence

10 Big Data

11 Machine Learning

12 Applied AI

Milestone 3: Capstone Project: Delivering Insights and Presentations

INTRODUCTORY COURSE PATH

1
INTRODUCTION TO DATA SCIENCE AND ANALYTICS

2
COMPUTING PRIMER

3
PROGRAMMING CONCEPTS

4
DISCOVERING AND CURATING DATA

5
STRUCTURING AND ANALYZING DATA

6
CLEANING AND ENRICHING DATA

7
VALIDATING AND PRESENTING DATA

8
INTRODUCTION TO DATA SCIENCE PROJECTS

9
ASSESSMENT NIGHT

LESSON 1: DATA CLEANING: IDENTIFYING USER ERRORS

LESSON OUTLINE

- Clean data for more accurate analysis.
- Identify errors to prevent redundancy.
- Verify/create consistent formatting.

DISCUSSION

In a tweet about the Paycheck Protection Program (PPP) loan data set, author Steven Rich notes that **Philadelphia is misspelled in 57 different ways**.

Without cleaning data (the topic we're about to explore), what problems might arise from using a data set like this?

LESSON GOALS

WHAT ARE THE GOALS?

You will demonstrate an understanding of techniques to clean data.

WHY ARE THEY IMPORTANT?

It's important to recognize issues in data sets and correct them, as well as to avoid common pitfalls.

YOUR TAKE

What are your goals for the lesson?

Share three things you know about the lesson topic and two things you want to know about the lesson topic.

REVIEW

Previous Lesson: The Analytic Process

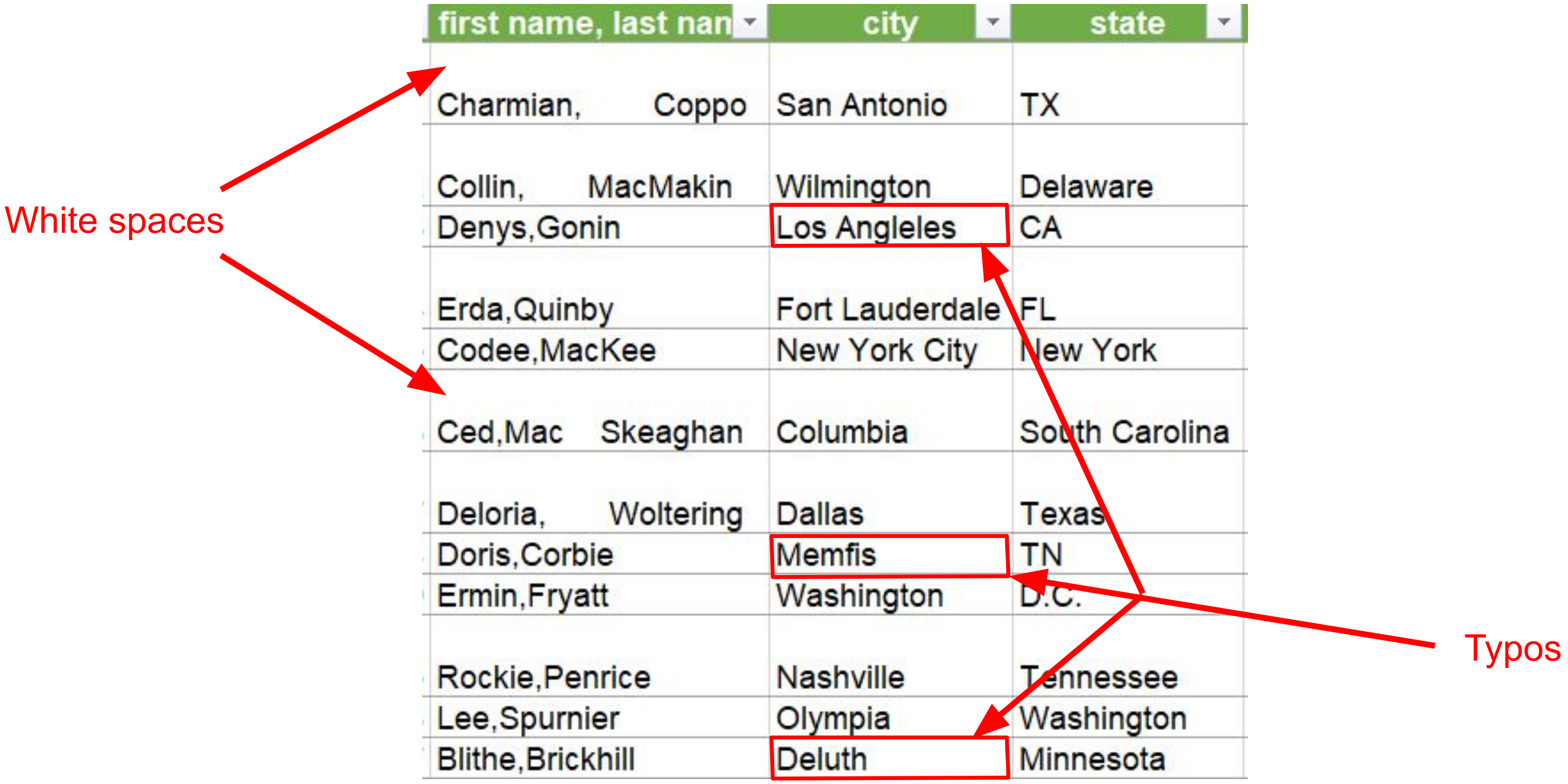
- We defined data analysis.
- We described the steps in a data analysis.
- We collected, cleaned, and processed data.

TYPES OF ERRORS

White space: The s p a c e between characters

Typo: When there is a spelling or typiing mistke

TYPES OF ERRORS



first name, last nan	city	state
Charmian, Coppo	San Antonio	TX
Collin, MacMakin	Wilmington	Delaware
Denys,Gonin	Los Angleles	CA
Erda,Quinby	Fort Lauderdale	FL
Codee,MacKee	New York City	New York
Ced,Mac Skeaghan	Columbia	South Carolina
Deloria, Woltering	Dallas	Texas
Doris,Corbie	Memfis	TN
Ermin,Fryatt	Washington	D.C.
Rockie,Penrice	Nashville	Tennessee
Lee,Spurnier	Olympia	Washington
Blithe,Brickhill	Deluth	Minnesota

White spaces

Typos

DISCUSSION

Think about what kind of data is collected about you.

What would happen if that data was input incorrectly?

What small mistake in entering data would have big consequences?

1.6.1 GUIDED ACTIVITY

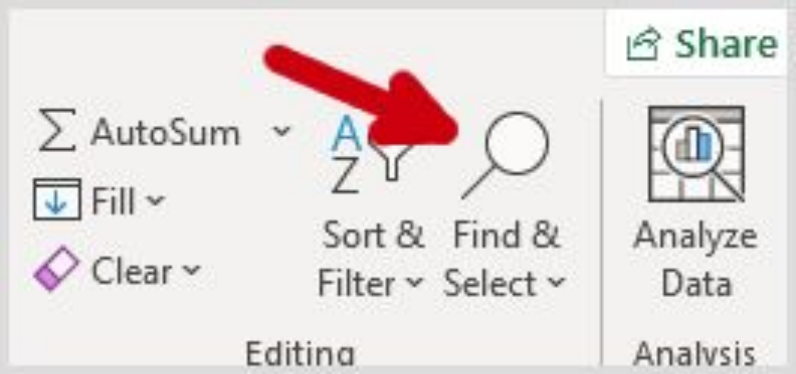
ın this activity, you’ll locate each error in the PPP loan application file **(1.6.1 Activity.xlsx).**

- Which columns have errors?
- What types of errors or inconsistencies are present?
- Approximately how many of each error are there?

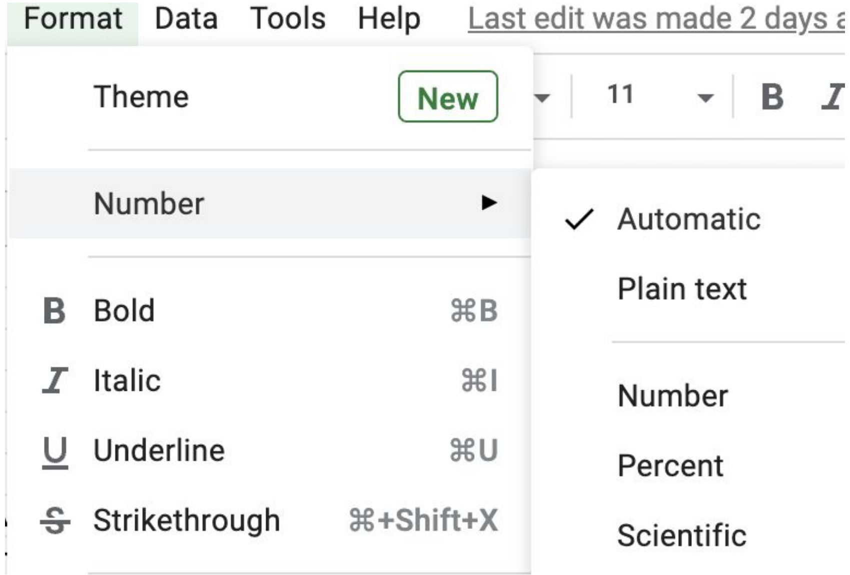
Hint: You may need to use Google to determine if cities and states are spelled correctly.

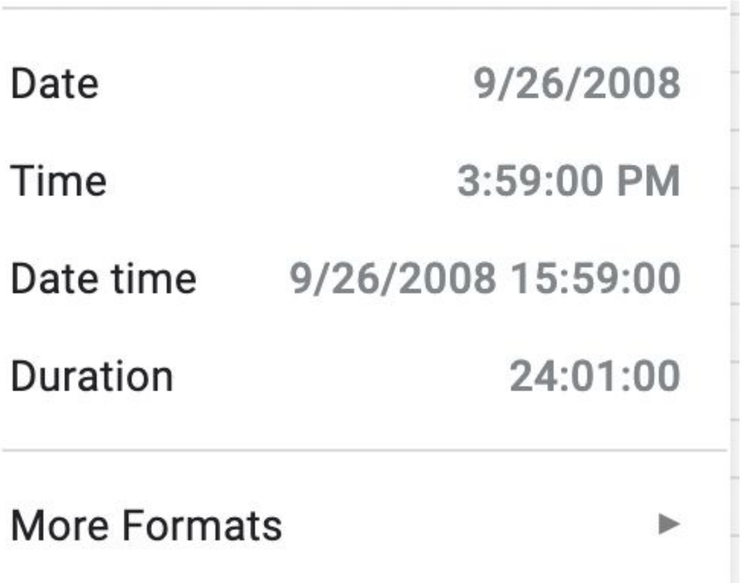
FIND & SELECT

The **Find & Select** tool can make massive changes across a data set to fix data errors.



DATE AND TIME





DATE AND TIME

You can use formatting to make your data more uniform.

It’s important the date and time are accurate so contextual information (such as analysis over time) can be interpreted properly.

CURRENCY

Consider the following when choosing how to format currency data:

- Decimal places
- Symbol
- Negative numbers

1.6.1 INDEPENDENT ACTIVITY

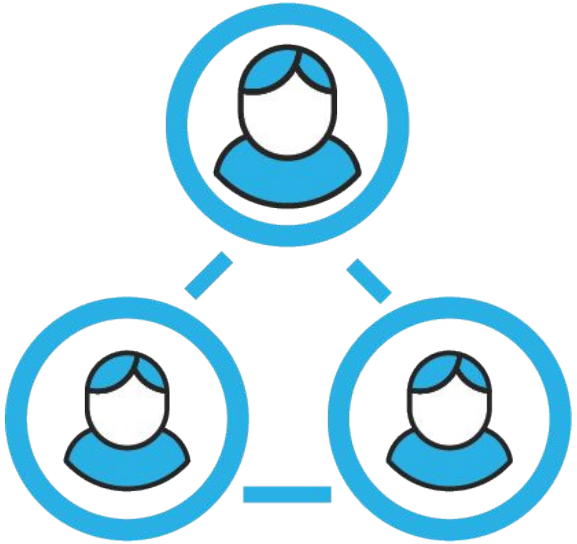
Now that you know what errors are in the PPP loan file, set out to correct them! You will need to use a combination of Google, spell check, Find & Select, and manual correction to assist you in your data cleaning task.

In case you missed any errors in your search, make sure you fix any:

- Typos
- White space
- Abbreviations
- Date and time formatting
- Currency formatting

INTERVIEW TIME

Why is it important to properly format the date, time, and currency?



REVIEW AND WRAP-UP

Today we learned to:

- Clean data for more accurate analysis.
- Identify errors to prevent redundancy.
- Verify/create consistent formatting.

QUESTIONS?

BREAK TIME

LESSON 2:
NULL VALUES AND OUTLIERS

LESSON OUTLINE

- Define null values and outliers.
- Determine whether to omit or alter null values and outliers.
- Examine null value and outlier solutions.

LESSON GOALS

WHAT ARE THE GOALS?

Learn how to detect, analyze, and manage null values and outliers.

WHY ARE THEY IMPORTANT?

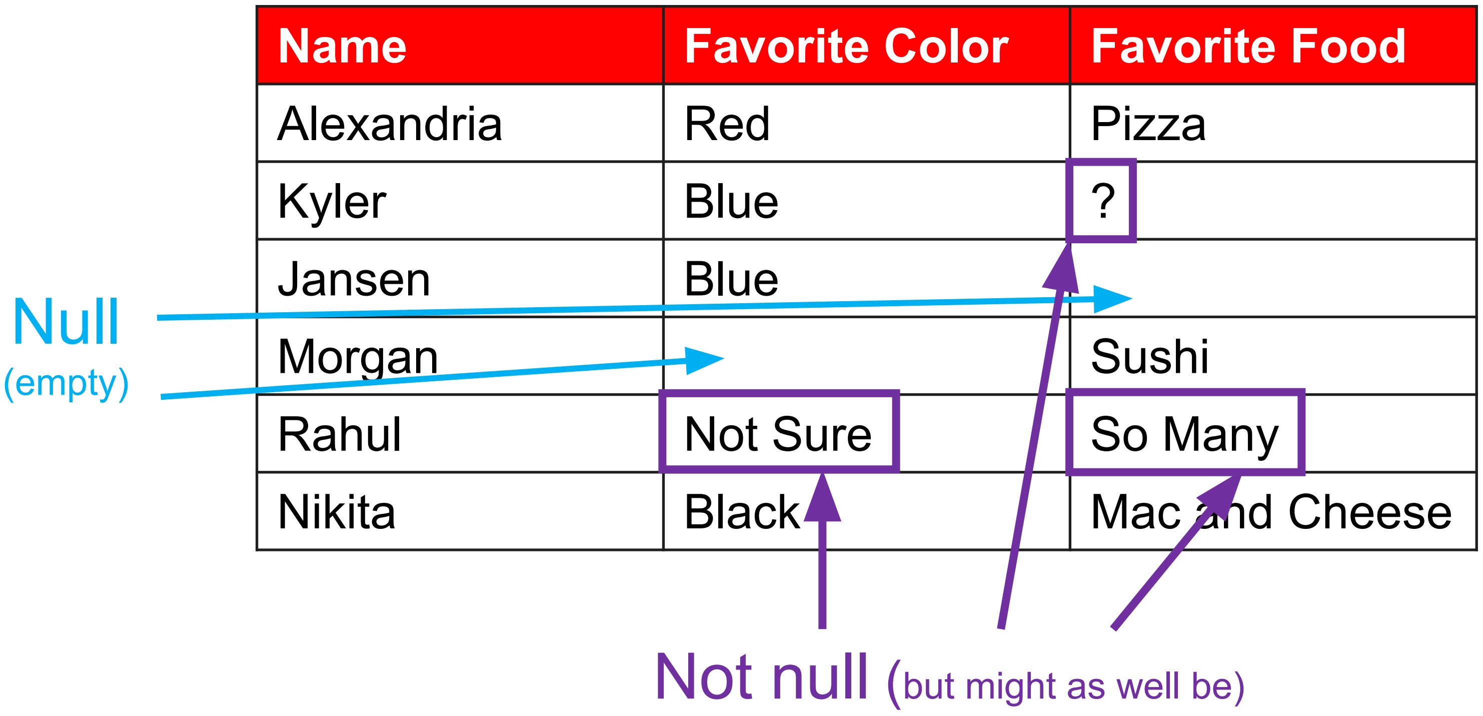
Null values and outliers can be symptoms/indicators of a significant event, can interfere with data analysis, and can cause confusion. We need to know how to recognize them and what to do when we spot them.

NULL VALUES

"Null, in a database context, is the total absence of a value in a certain field and means that the field value is unknown. Null is not the same as a zero value for a numerical field, text field or space value."
—Techopedia

A null value may be caused by user error or if no value was collected. For example, if you skipped a question on a test, that would be a null value, even though you handed in the test and have values for other answers.

NULL VALUES



Name	Favorite Color	Favorite Food
Alexandria	Red	Pizza
Kyler	Blue	?
Jansen	Blue	
Morgan		Sushi
Rahul	Not Sure	So Many
Nikita	Black	Mac and Cheese

Null
(empty)

Not null (but might as well be)

NULL VALUES

We should keep null values (keep rows) if:

- The values in other features can still contribute information.
- There is a small number of null values that will not affect the overall outcome.

We should drop null values (delete rows) if:

- There are multiple null values in multiple features (blank/empty data).
- There are only a few values being dropped.

We should drop null values (delete column) if:

- The data is unrelated to the research question.

We should change null values if:

- Information from the null values is required to get a complete picture.
- Changing a null to a field like “Other” or “No Response” can add information.

1.6.2 GUIDED ACTIVITY

- Examine the scenarios described in the 1.6.2 Guided Activity file.

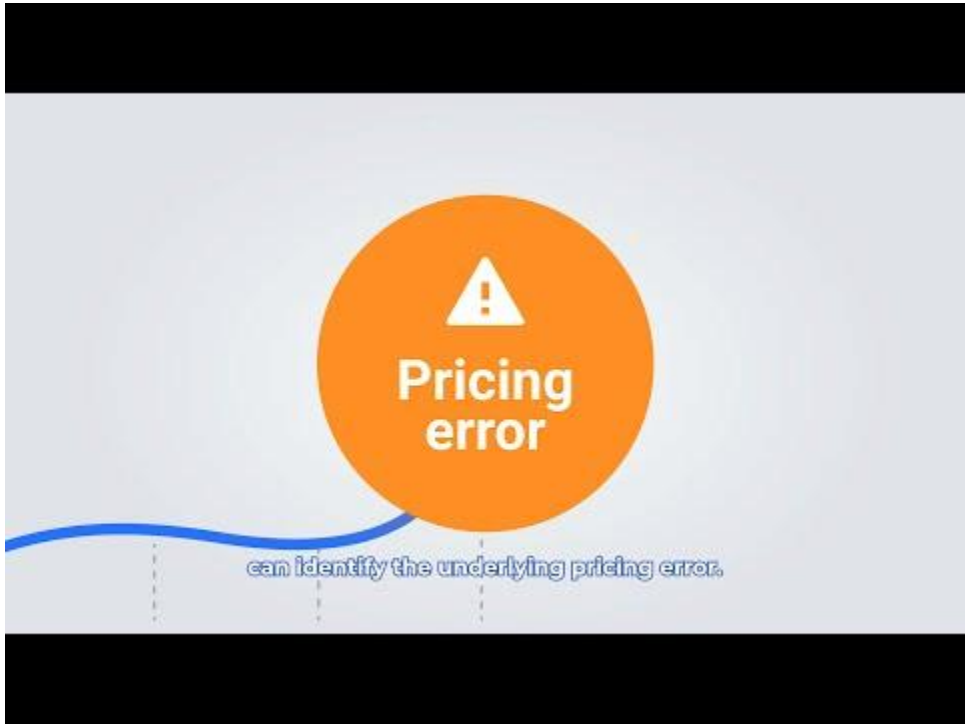
- Follow the prompts to determine whether to keep the values null, change the values to something else, or remove the data completely.

OUTLIERS

"An outlier, in mathematics, statistics and information technology, is a specific data point that falls outside the range of probability for a data set. In other words, the outlier is distinct from other surrounding data points in a particular way."
—Techopedia

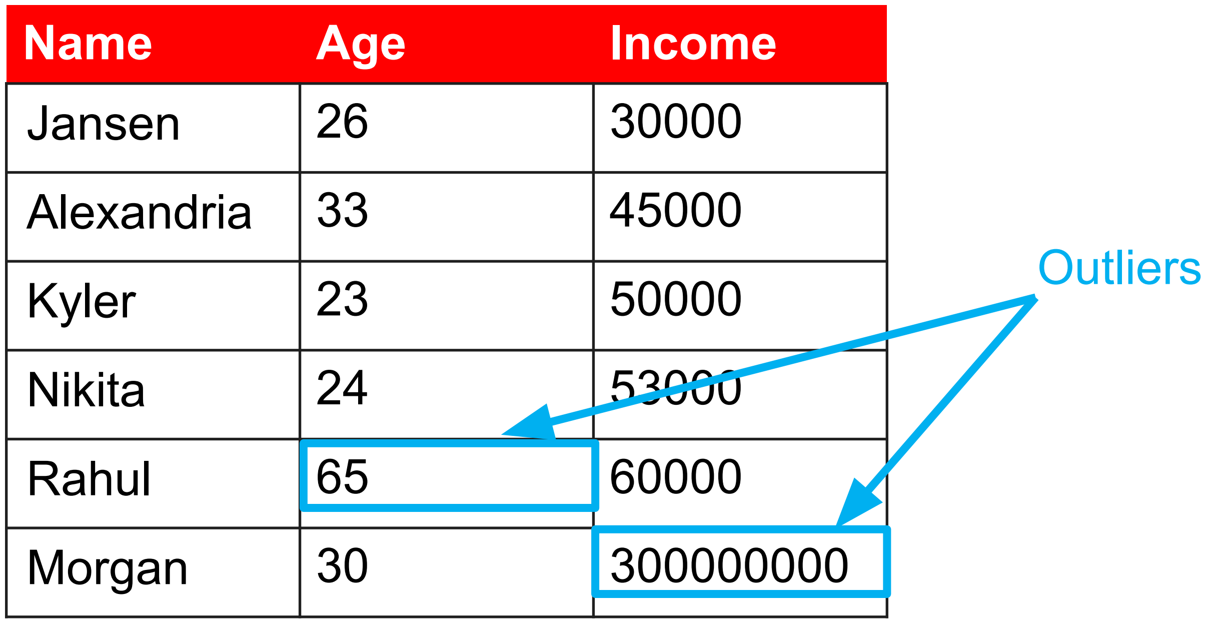
An outlier could be arrival time. You get to work between 8 and 8:15 a.m. every day, but one day you don't get in until noon. That is not the norm but an outlier.

OUTLIERS



From: Anodot on YouTube

OUTLIERS



Name	Age	Income
Jansen	26	30000
Alexandria	33	45000
Kyler	23	50000
Nikita	24	53000
Rahul	65	60000
Morgan	30	300000000

Outliers

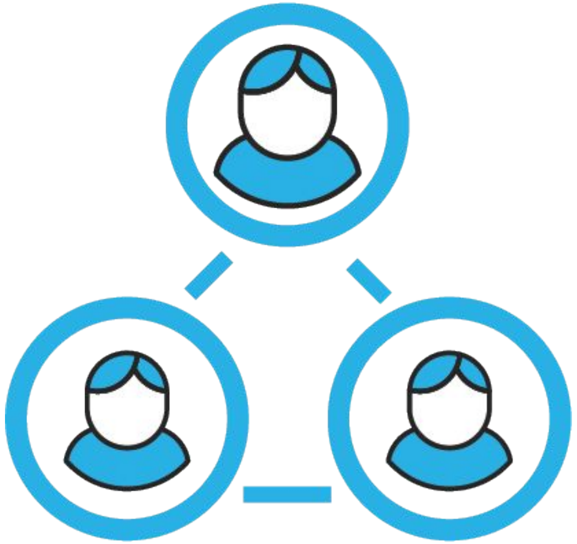
1.6.2 INDEPENDENT ACTIVITY

Examine the scenarios described in the 1.6.2 Independent Activity file.

Follow the prompts to determine how to handle the outliers described in the data.

INTERVIEW TIME

How do you handle null values and outliers?



REVIEW AND WRAP-UP

Today we learned:

- To define null values and outliers
- To determine whether to omit or alter null values and outliers
- To examine null value and outlier solutions

QUESTIONS?

BREAK TIME

LESSON 3: DATA ENRICHMENT

LESSON OUTLINE

- Define feature engineering and use it to make new data.
- Use grouping as a feature to organize categorical variables.
- Use binning as a feature to categorize numeric variables.

DATA ENRICHMENT

From the mid-nineteenth century, stock prices would run through ticker tape. It’s amazing how one piece of tape contained all the stock information of the day.

Today, we receive stock information similarly, from streaming stock data. However, whenever we see stock summaries, we usually see a graph with the stock history, which can be shown in months and years. This provides a picture that is much stronger than the tape.

How did we get here? Stock data on ticker tape can only hold a limited number of features. However, using knowledge of stocks and the stock market (**domain knowledge**), we can create new features to help organize this ticker tape data into a format that we can put into charts and graphs, which will make these numbers easier to understand.

DATA ENRICHMENT

“Data enrichment is a general term that refers to processes used to enhance, refine, or otherwise improve raw data…A common data enrichment process could, for example, correct likely misspellings or typographical errors in a database.”

—Stitch

LESSON GOALS

WHAT ARE THE GOALS?

You will learn how to apply feature selection techniques.

WHY ARE THEY IMPORTANT?

Making data sets easier to understand and organize is important for data analysis.

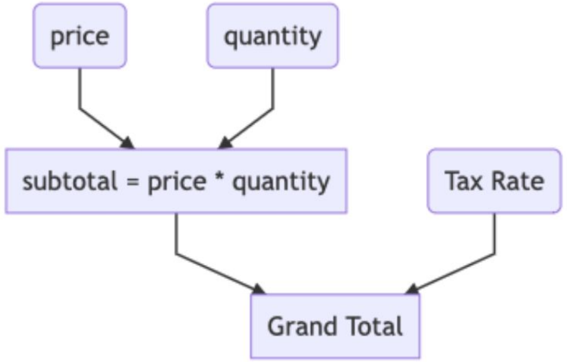
FEATURE ENGINEERING

“Feature engineering is focused on using the variables you already have to create additional features that are (hopefully) better at representing the underlying structure of your data.”
—Max Steele

FEATURE ENGINEERING

Feature engineering is when we create features from one or more existing features.

Features are variables in data sets that are important for predicting a target or answering a question the data analyst might have.



Price and quantity from the café example

1.6.3 GUIDED ACTIVITY

In this activity, you'll calculate daily pay for an individual working at a company. Follow the instructions in the 1.6.3 Activity file and complete the spreadsheet labeled **1.6.3 Activity.xlsx**.

GROUPING

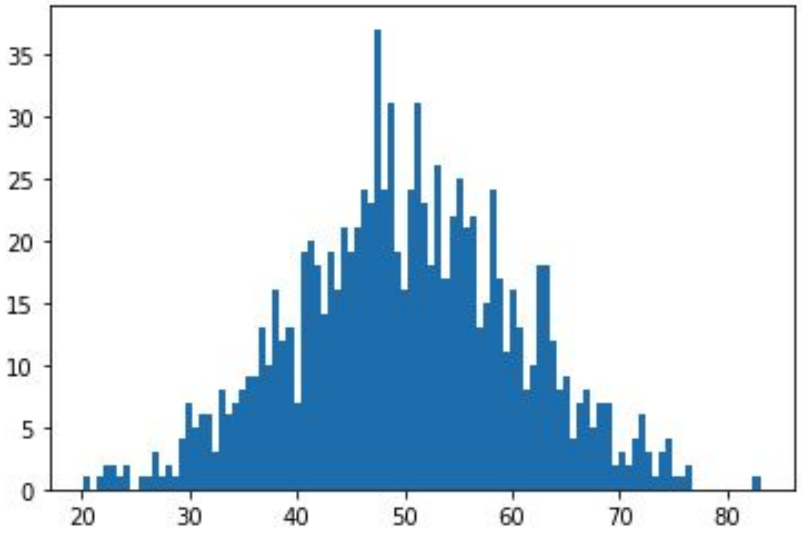
“Grouped data is data that has been bundled together in categories. Histograms and frequency tables can be used to show this type of data:”
—Statistics How To

BINNING

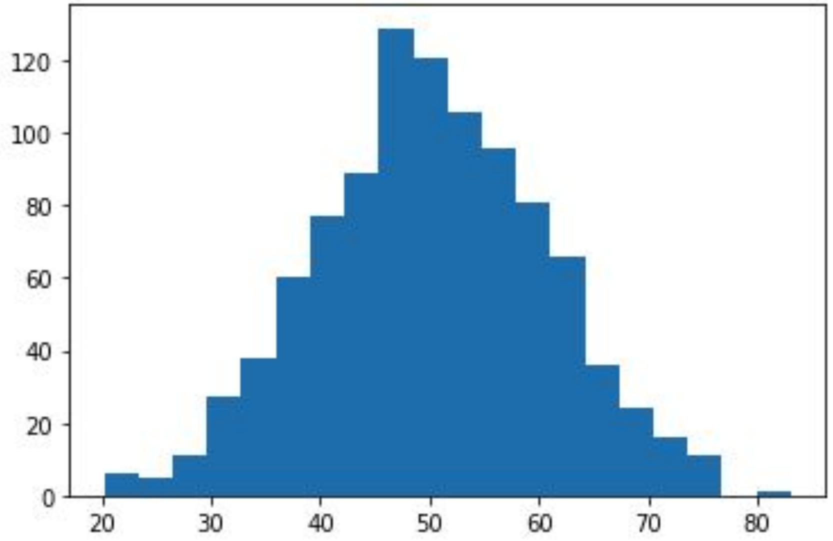
“Binning is a way to group a number of more or less continuous values into a smaller number of ‘bins.’ For example, if you have data about a group of people, you might want to arrange their ages into a smaller number of age intervals.”
—TIBCO

Binning

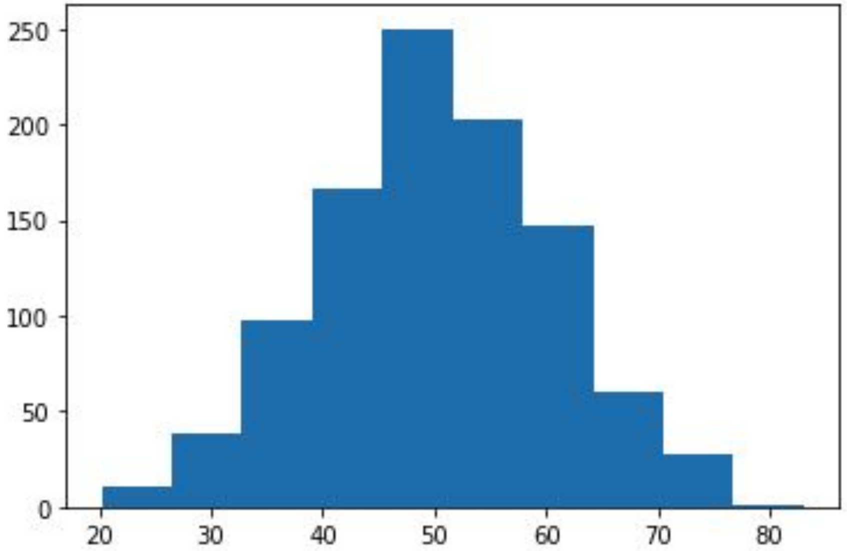
No bins



Bins = 20



Bins = 10



DISCUSSION

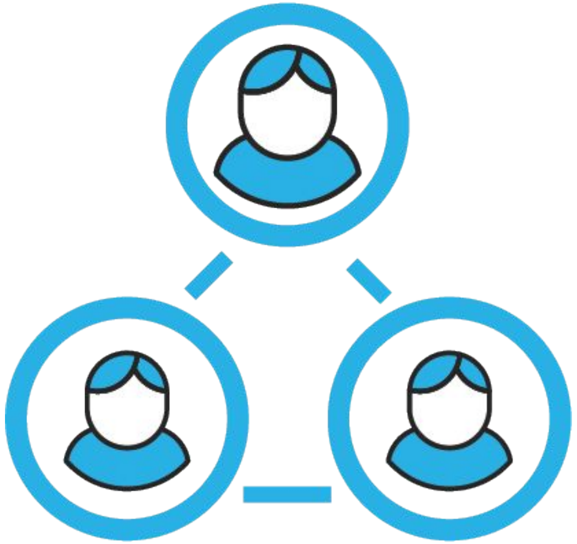
Are there any questions on the binning versus grouping concepts?

How does binning work?

Does it come as naturally as grouping data?

INTERVIEW TIME

What types of features can be engineered with clickstream data?



REVIEW AND WRAP-UP

Today we learned:

- What feature engineering is
- About grouping and binning data

QUESTIONS?

NEXT STEPS

☑ Assigned Activities

☑ Reminders