



Week 12:

Data Analysis Pipeline

Data Science Pipeline



Finding Data

Access and evaluate a relevant dataset



Research Questions

Identify questions about this dataset



Data Refining

Filter, clean and trim the data



Data Visualization

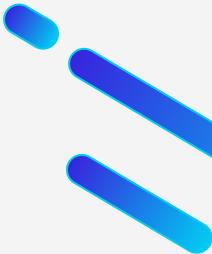
Create data visualizations and interpret



Communicating Results

Share your conclusions and insights

Today's focus



Finding Data

Access and evaluate a relevant dataset



Research Questions

Identify questions about this dataset



Data Refining

Filter, clean and trim the data



Data Visualization

Create data visualizations and interpret



Communicating Results

Share your conclusions and insights

Sources of Data

Primary data is collected directly from the source (target population or system under study).

Primary Data

- Surveys
- Interviews
- Observations
- Experiments
- Internet of Things (devices with sensors that exchange data)



Secondary Data

- Internet
- Social media
- Financial reports
- Journals & papers
- Databases
- Government publications
- Open Data



Secondary data are sources that have already been collected, processed, and made available for use by researchers, analysts, or the public.

Secondary Source - Example

A financial report, presented by the company itself or a reporter presenting their findings, is an example of a **secondary source**.

- How do we know this is a **secondary source**?

FACEBOOK

Meta Reports Fourth Quarter and Full Year 2022 Results

MENLO PARK, Calif., Feb. 1, 2023 /PRNewswire/ -- Meta Platforms, Inc. (Nasdaq: META) today reported financial results for the quarter and full year ended December 31, 2022.

"Our community continues to grow and I'm pleased with the strong engagement across our apps. Facebook just reached the milestone of 2 billion daily actives," said Mark Zuckerberg, Meta founder and CEO. "The progress we're making on our AI discovery engine and Reels are major drivers of this. Beyond this, our management theme for 2023 is the 'Year of Efficiency' and we're focused on becoming a stronger and more nimble organization."

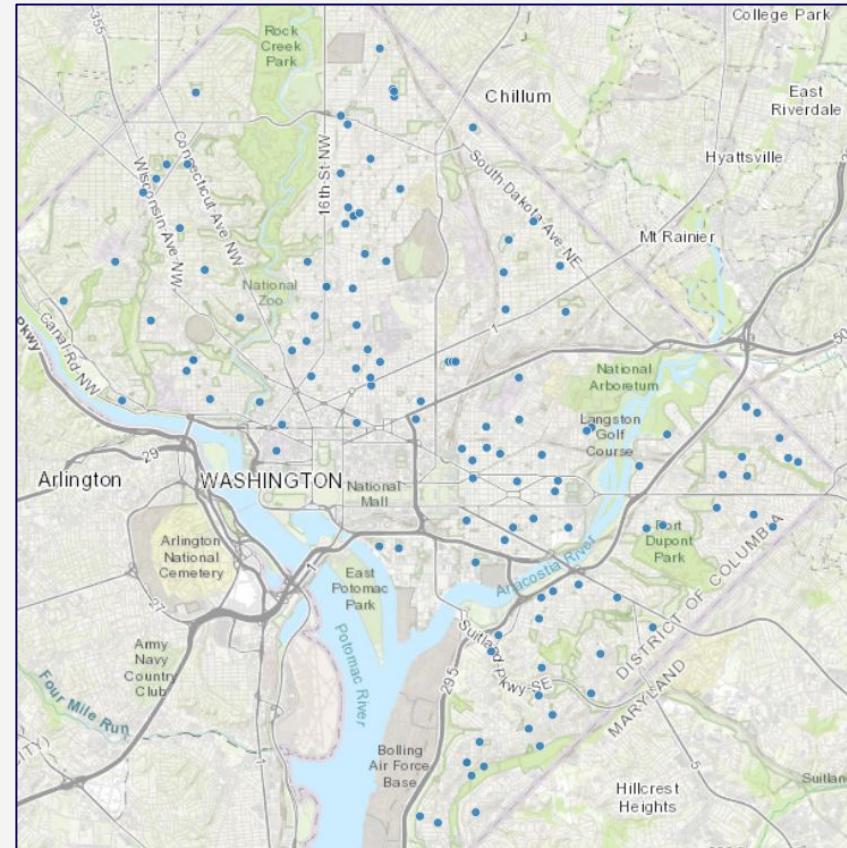
Fourth Quarter and Full Year 2022 Financial Highlights

In millions, except percentages and per share amounts	Three Months Ended December 31,		Year Ended December 31,		% Change	
	2022	2021	2022	2021		
Revenue	\$ 32,165	\$ 33,671	(4)%	\$ 116,609	\$ 117,929	(1)%
Costs and expenses	25,766	21,086	22%	87,665	71,176	23%
Income from operations	\$ 6,399	\$ 12,585	(49)%	\$ 28,944	\$ 46,753	(38)%
Operating margin	20%	37%		25%	40%	
Provision for income taxes	\$ 1,497	\$ 2,417	(38)%	\$ 5,619	\$ 7,914	(29)%
Effective tax rate	24%	19%		19%	17%	
Net income	\$ 4,652	\$ 10,285	(55)%	\$ 23,200	\$ 39,370	(41)%
Diluted earnings per share (EPS)	\$ 1.76	\$ 3.67	(52)%	\$ 8.59	\$ 13.77	(38)%

Secondary Source - Example

Datasets hosted on government database websites like OpenDataDC, like this one on the location of DCPS schools, is another example of a **secondary source**.

- How can you distinguish this from a **primary source**?



Secondary Source - Example

API hubs such as RapidAPI, which we'll use later in this course, are another example of a **secondary** source that provide information directly from an app or website (can be real-time!)

The screenshot shows the RapidAPI platform interface. A specific endpoint, "Get Daily Scoreboard - Live - Real Time", is selected. The interface includes a search bar at the top, a list of endpoints on the left, and detailed description and configuration options on the right.

Selected Endpoint: GET Get Daily S...

Description: This call will pull game scores and no other stats. /getNFLScoresOnly

Details: Call it with no parameters, it will return the current date's game list with their scores.

Request URL: rapidapi.com

Header Parameters:

Code Snippets:

```
(Python) Requests Copy Code
import requests

url = "https://tank01-nfl-live-in-game-real-time-statistics-nfl.p.rapidapi.com/getNFLScoresOnly"

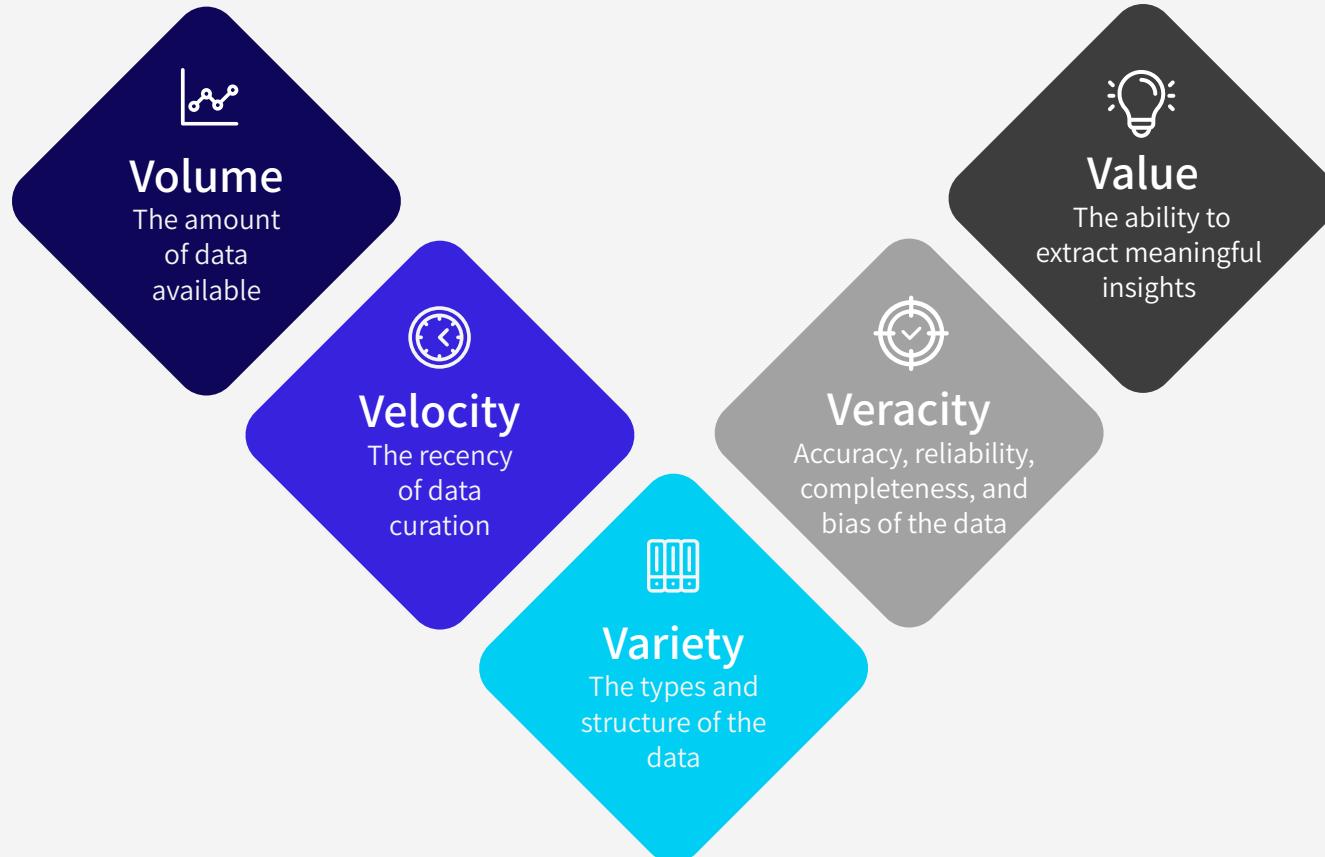
querystring = {"gameDate": "20221211", "topPerformers": "true"}

headers = {
    "X-RapidAPI-Key": "SIGN-UP-FOR-KEY",
    "X-RapidAPI-Host": "tank01-nfl-live-in-game-real-time-statistics-nfl.p.rapidapi.com"
}

response = requests.get(url, headers=headers, params=querystring)

print(response.json())
```

Rethinking the 5Vs of Big Data



Evaluating Data: Volume

The amount of data available and whether it is sufficient to support the investigation at hand.

- Datasets that are too small may result in incomplete or misleading conclusions, while very large datasets may be difficult to manage or interpret.
- Think of insights or the different questions you could draw/pose from datasets, looking at different timespans (e.g., daily weather vs. past 10 days vs. past 10 years)



Volume
The amount
of data
available

Evaluating Data: Volume

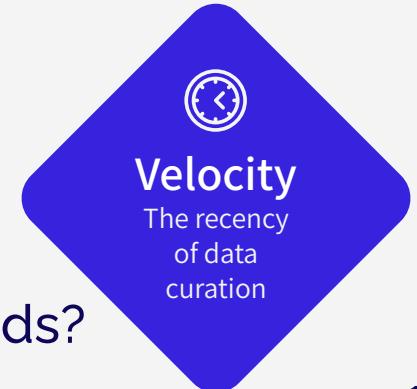
- How much data is included in the dataset? Or, How many data points/observations/rows are included?
- Is there enough data to answer the research questions?
- Is there enough data to draw meaningful insights or conclusions?
- Is there too much data or too many observations/rows? Or is some of it not relevant or not related to the main question?



Evaluating Data: Velocity

How current or recent the data is and whether it reflects a static snapshot or changes over time.

- Some questions require up-to-date data (e.g., social media trends), whereas others can be addressed using historical data (e.g., rates of population change).
- Think of a regional COVID-19 dataset that is updated **weekly** vs. a regional dataset that is updated **daily**. Which one is more appropriate for identifying sudden spikes or long-term trends?



Evaluating Data: Velocity

- When was the data curated or last updated?
- Does it includes real-time or recent data?
- Is the dataset relevant to the investigation period?
- How recent does the data need to be to be useful for the current investigation?



Evaluating Data: Variety

The types of data included (e.g., numerical, categorical, etc.) and how the data are structured and organized (e.g., tabular format, JSON, map)

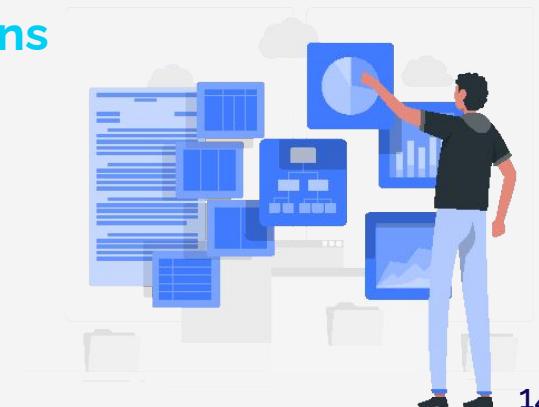
- Various types of data can support richer analysis, but can also be more complex to interpret.
- Think of a mammal's dataset that includes their diet (categorical), speed (numerical), life span (numerical), habitat description (text), and latitude and longitude values (spatial). Which variables are the most useful for investigation?



Variety
The types and structure of the data

Evaluating Data: Variety

- What data types are included in the dataset (e.g., numerical, categorical, text, images, etc.)?
- Is the dataset structured and organized (e.g., tabular format, JSON, map)?
- Is the dataset well-documented? Are there details about the variables included?
- Is the data format consistent across observations and easy to interpret with the available tools?



Evaluating Data: Veracity

The accuracy, reliability, completeness, and potential biases in the dataset.

- Inaccurate, incomplete or biased data can lead to false conclusions.
- Let's break it down and further understand each dimension separately.



Evaluating Data: Accuracy

- What do we know about where the data came from? For example, how was the data collected? By whom? For what purpose(s)?
- Is the dataset accurate compared to **known fact** or **external source**?



Accurate, Reliable



Accurate, Unreliable



Inaccurate, Reliable



Inaccurate, Unreliable

Evaluating Data: Reliability

- Is the data source **reliable**? Is the source consistently measured?
- Some measuring instruments and designs may reliably measure data; others may introduce measurement variability (such as a malfunctioning thermometer or poorly worded question on a survey)



Accurate, Reliable



Accurate, Unreliable



Inaccurate, Reliable



Inaccurate, Unreliable

Evaluating Data: Completeness

- Is there any **missing data**?
- Check to make sure the dataset does not have a lot of missing data points – particularly if they are enough to make the sample small, or if they are systematically missing in certain populations
- **Benchmark:** Does your dataset represent all sub-populations you're interested in?



Evaluating Data: Data Biases

- Is the data potentially **biased**?
- Are there potential biases in data sampling, reporting, or measurement? Are there any **under- or over-represented** populations?
Are responses influenced by social context?
Are tools used systematically mismeasuring a phenomenon?
- If there are biases, how might these biases **potentially influence the results**?



Evaluating Data: Value

The relevance and usefulness of the data in answering a given question or generating meaningful insights.

- Even well-structured, accurate datasets may not be useful for answering a specific question and drawing meaningful insights.
- A dataset is valuable if it helps generate evidence in support of (or against) an explanation.



Evaluating Data: Value

- Does the dataset contain the **necessary information to address the scientific question** or the phenomenon under study?
- Can it be analyzed to **derive meaningful insights** or evidence-based claims?
- **Which variables are most helpful or insightful?**
- What **additional data** could make the dataset more useful?



Quiz

Match these 5 scenarios to the V (from the 5Vs) that they do the best job of representing:

- A researcher collects 10,000 responses to a survey on favorite color.
- A student uses an API to collect live-updated data from Spotify on top hits.
- A professor uses a verified government database to find population statistics for an upcoming presentation.
- A company uses not only survey data, but also video-recorded interviews and field-testing data to draw conclusions about a new product.
- The San Francisco Zoo decides to use a new research paper specifically about gorillas, instead of one written about chimpanzees, to determine what food to give to their gorillas.



Data Provenance

- A historical record of where the data came from, how it was collected, and how it was handled.
- Based on these records you should be able to answer the following questions:
 - Who collected it?
 - How was the data collected?
 - Were there any modifications along the way?



Data Provenance can help identify errors and biases

- Who collected it?
 - Some orgs are biased and may cherry pick data.
- How was the data collected?
 - Some methods are flawed and biased.
- Were there any modifications along the way?
 - Helps explain the types of errors that might have been introduced
 - Computers make formatting errors
 - Humans make spelling mistakes



Case Study: the Literacy Dataset

- Who collected it?
- How was the data collected?
- Were there any modifications along the way?
- What kinds of data quality issues could be introduced?



Today's Focus



Finding Data

Access and evaluate a relevant dataset



Research Questions

Identify questions about this dataset



Data Refining

Filter, clean and trim the data



Data Visualization

Create data visualizations and interpret



Communicating Results

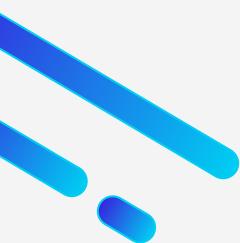
Share your conclusions and insights

Today, we'll focus on identifying questions about our datasets.

Review of Final Project

Exemplars

- Take a look at some example final projects from previous years:
 - Pokemon GO Moves
 - Soccer Players
 - DC Affordable Houses
- Look at the **structure** of these project presentations and **what kinds of questions** they ask!
 - Look at the **section headers**. What headers are included?
 - Look at the **content** within each section. What would go in each section in your project?



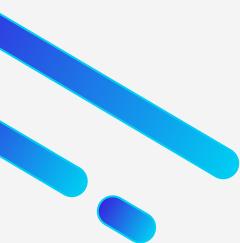
Types of Data Science Questions

- **Descriptive** questions ask about features and statistics of a particular sample or population
 - **Example:** How many movies are comedies?
 - **Example:** What are the win rates across NFL teams?
- **Comparative** questions compare two samples or populations
 - **Example:** Does the most popular dog differ by country?
 - **Example:** Do pop songs get more streams than rock songs?



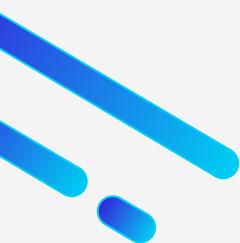
Types of Data Science Questions

- **Evaluative** questions often look for the “best” or most extreme cases in a sample or population.
 - **Example:** Who is the best soccer player in the USA?
 - **Example:** Who is the most popular pop music artist?
- **Predictive** questions ask what factors can be used to predict outcomes
 - **Example:** Do heavier dogs live longer?
 - **Example:** Is there a relationship between roller coaster height and roller coaster speed?



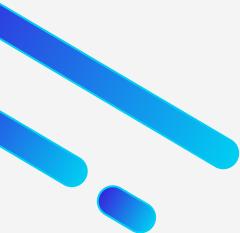
First Draft of Questions

- Write **4-5 data science questions** that you could use your dataset to answer
- These questions might be **descriptive, comparative, evaluative, or predictive**
 - Try to think of a few different types of questions that you could ask about your dataset!



Considering Additional Data

- After this first draft, you might feel like you need an additional data source to answer some of your questions.
- Consider searching other data sources, or consider data augmentation with LLM API.



Today's Focus



Finding Data

Access and evaluate a relevant dataset



Research Questions

Identify questions about this dataset



Data Refining

Filter, clean and trim the data



Data Visualization

Create data visualizations and interpret



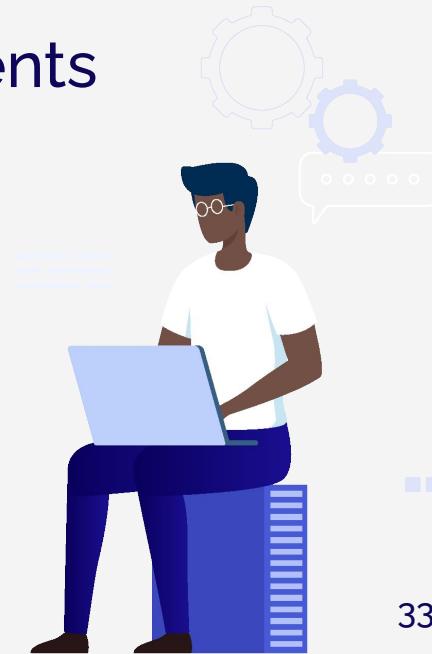
Communicating Results

Share your conclusions and insights

Today's Focus: Data Refining

There are four main goals for this stage:

- Filter your data with conditional statements
- Filter the variables you want to see
- Identify and quantify any missing data
- Identify and fix any duplicate responses



Today's Focus



Finding Data

Access and evaluate a relevant dataset



Research Questions

Identify questions about this dataset



Data Refining

Filter, clean and trim the data



Data Visualization

Create data visualizations and interpret



Communicating Results

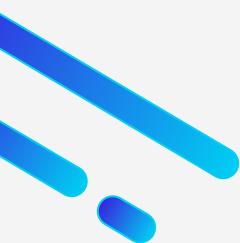
Share your conclusions and insights

Data Visualization

- Often interwoven with data cleaning and exploratory analysis. Helps you “orient” yourself to the data.
- Need more ideas? <https://datavizcatalogue.com/>
- Consider:
 - What one-variable graphs could you create to focus on the distribution of a particular variable?
 - What two-variable graphs could you create to examine relationships between two or more variables?

Interpreting Data Visualizations

- For each data visualization,
 - identify **what kind of graph it is**, and
 - write an **interpretation of a feature of the graph**, or the graph as a whole, that **helps you answer one of your questions**



Today's Focus



Finding Data

Access and evaluate a relevant dataset



Research Questions

Identify questions about this dataset



Data Refining

Filter, clean and trim the data



Data Visualization

Create data visualizations and interpret



Communicating Results

Share your conclusions and insights

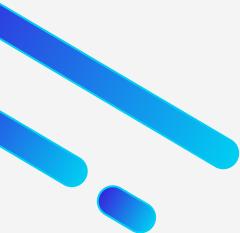
Final Project Structure

For the final project, make sure you answer the following questions:

- Dataset used
- Question posed
- Techniques used in program
- 3-5 graphs you created
- Description of what each graph is showing

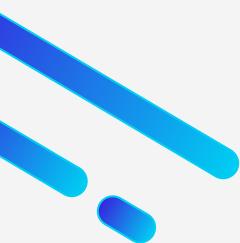
Dataset

- Name the dataset you chose, and explain the Volume, Variety, Velocity, and Value of the data.
 - Do you have “a lot” of data to answer your question?
 - Is the data from the right time period for the question you are asking?
 - What data types are present in the dataset, and what do they represent?
 - Why is this data valuable/meaningful, and applicable to your question?



Questions for Investigation

- Explain the question(s) that drove your investigation of this dataset.
 - What subquestions did you have to answer?
 - How did these smaller questions contribute to finding a more complete answer for your driving question?



Data Visualization & Interpretations

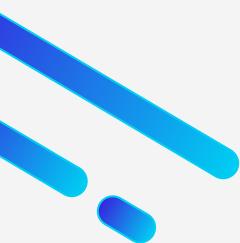
When creating a presentation (3-5) you include in your presentation, explain the following:

- What type of graph is this, and what are some key features to look at?
- How would you interpret what the graph is showing?
- How does this help you answer your questions for investigation?



Conclusions

- Address each of the following:
 - What answer did you reach for your questions?
 - How did your questions relate to one another? Why are they important to answer?
 - How confident are you in your answers? (did you have enough data & the right kind of data to fully answer your questions?)



What is Data Storytelling?

Data storytelling is the practice of using **data**, **visualizations**, and **narrative** to communicate insights in a way that informs, engages, and inspires an audience to act.



Data Storytelling



Data: Analyzed information collected through observation or measurement (e.g., Survey results, test scores, app visits)

Visualizations: Visual representation of the data, such as charts, graphs, images, and maps

Narrative: A structured story that defines the context, problem, and solution. “Why” the data matters and “What” it means

Building a Narrative

- **Settings:** Establish the context, background, subject, and timeframe of the data.
- **Conflict:** Introduce the main question/ issue.
- **Insights:** Present the data and the analysis performed, uncover patterns/trends.
- **The “Aha!” Moment:** Present the most significant finding. Answer the main question.
- **Resolution:** Summarize the key takeaways and provide recommendations.



Next week - Happy Thanksgiving!

- There will still be lab sessions, but
 - Attendance is optional
 - No lab notebook submission
- Make sure to submit your midterm!
- We will have one more programming assignment, after the thanksgiving break

