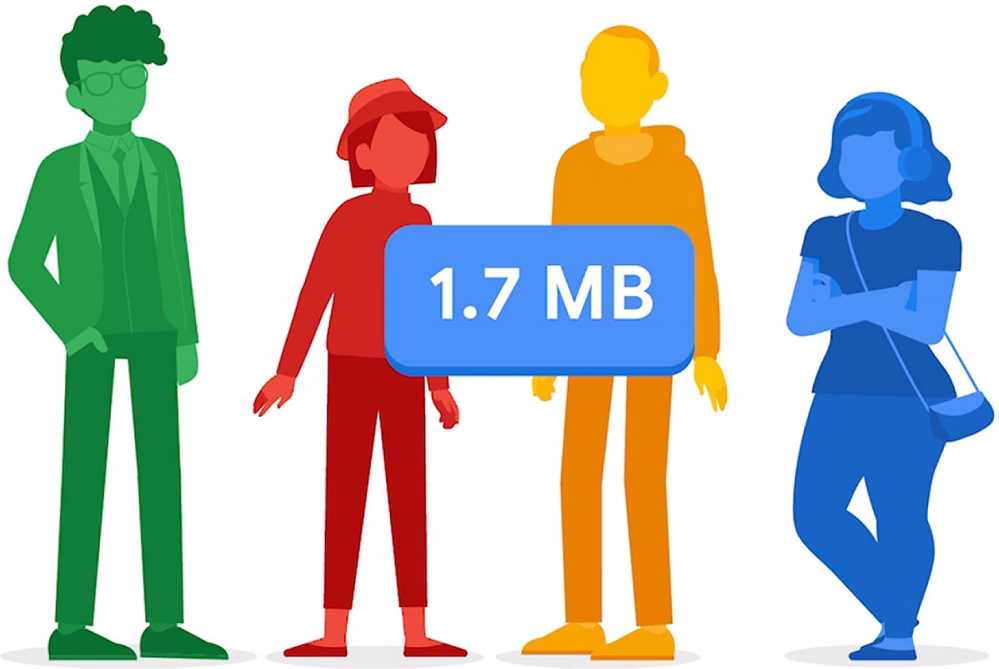
Exploratory Data Analysis

Parham Zilouchian Moghaddam

May 2025

Every person, every second, produces 1.7 MB

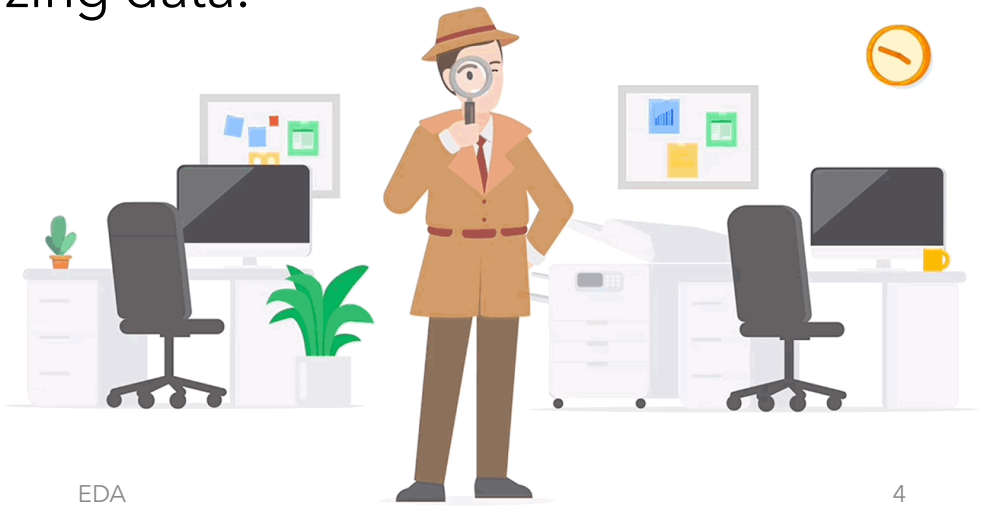


Over 2.5 quintillion bytes of data every single day worldwide.

- There is a **huge demand**, now and for the foreseeable future, for people who can organize data and interpret the stories locked within.

- Data professional: A term used to describe any individual who works with data and/or has data skills.

- At a minimum, a data professional is capable of exploring, cleaning, selecting, analyzing, and visualizing data.

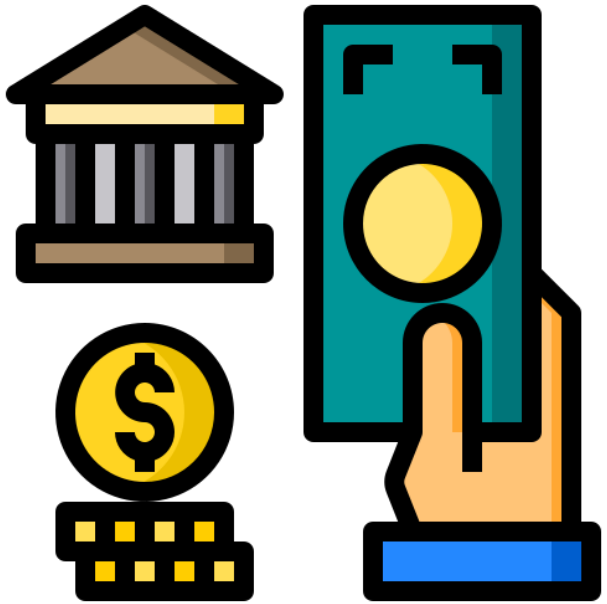


Technical data professionals

- Expertise in mathematics, statistics, and computing.
- Build models and make predictions.
  - (Use tools such as R and Python to extract value from business datasets.)
- The result is a solution that has a direct positive impact.
- Explore datasets
- Interpret information for an organization's operations, finance, research, and development.
- Work aligns with business strategy.

Finance

- Assess risks
- Monitor market trends
- Detect anomalies to reduce fraud
- Create a more stable financial system

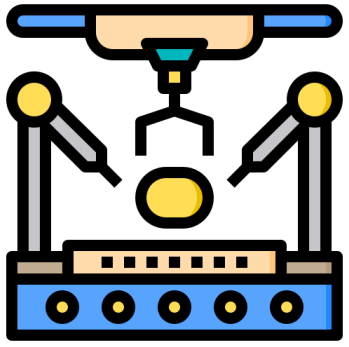


Healthcare

- Process clinical data
- Support early detection
- More precise diagnoses

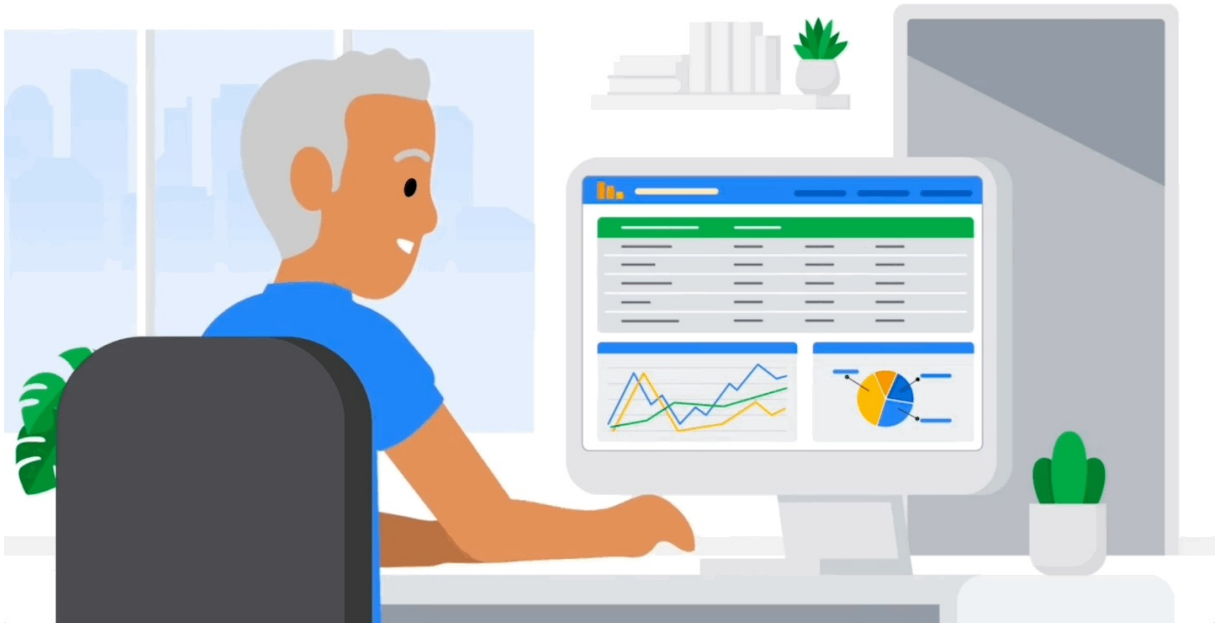
Manufacturing

- Predict when to perform preventative maintenance to avoid production line breakdowns.
- Use data to maximize quality assurance and defect tracking.
- AI models help respond to logistical issues and reduce delivery truck miles on the road. (Advancing key sustainability goals.)
- When supply chains reach every corner of the world, data enables clear and near-real-time communication with suppliers, retailers, and other network partners.



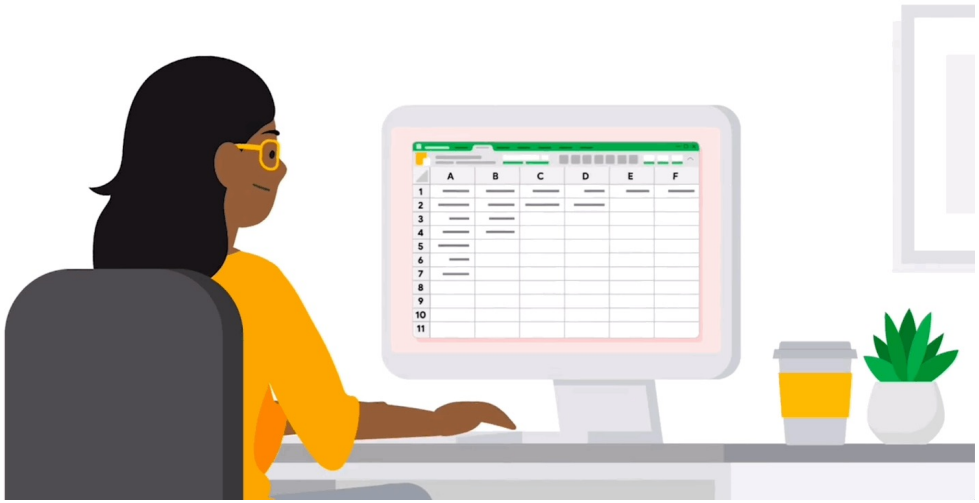
Open data

- Data that is available to the public and free to use, with guidance on how to navigate the datasets and acknowledge the source.

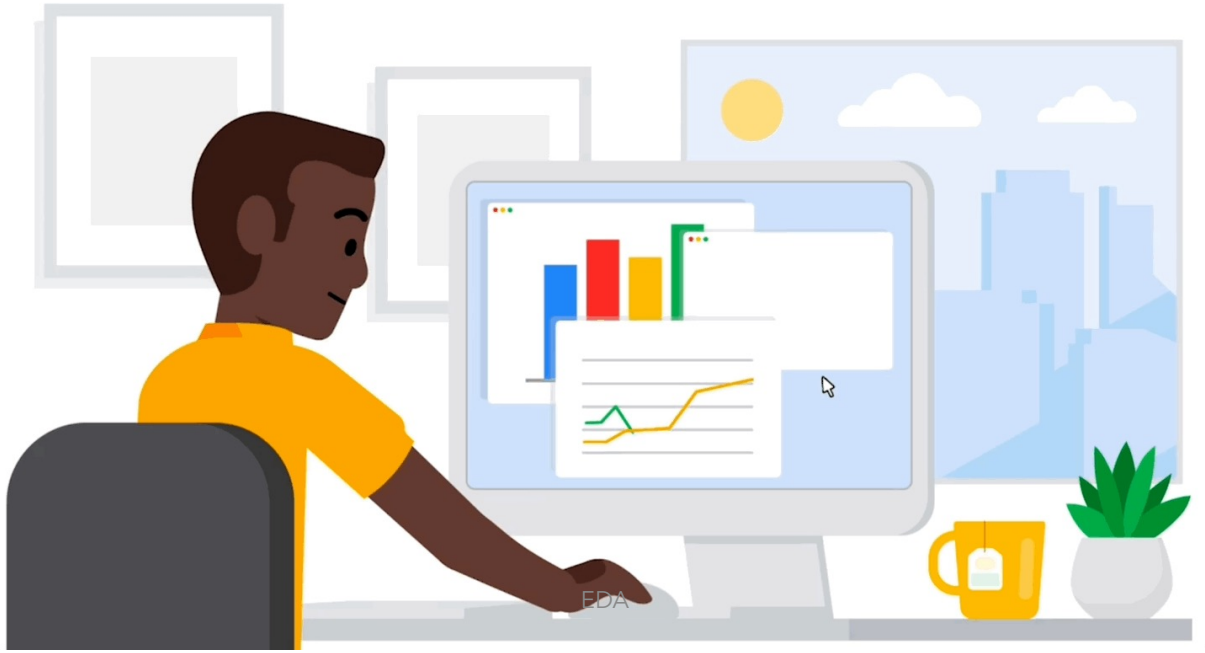


Data Cleaning

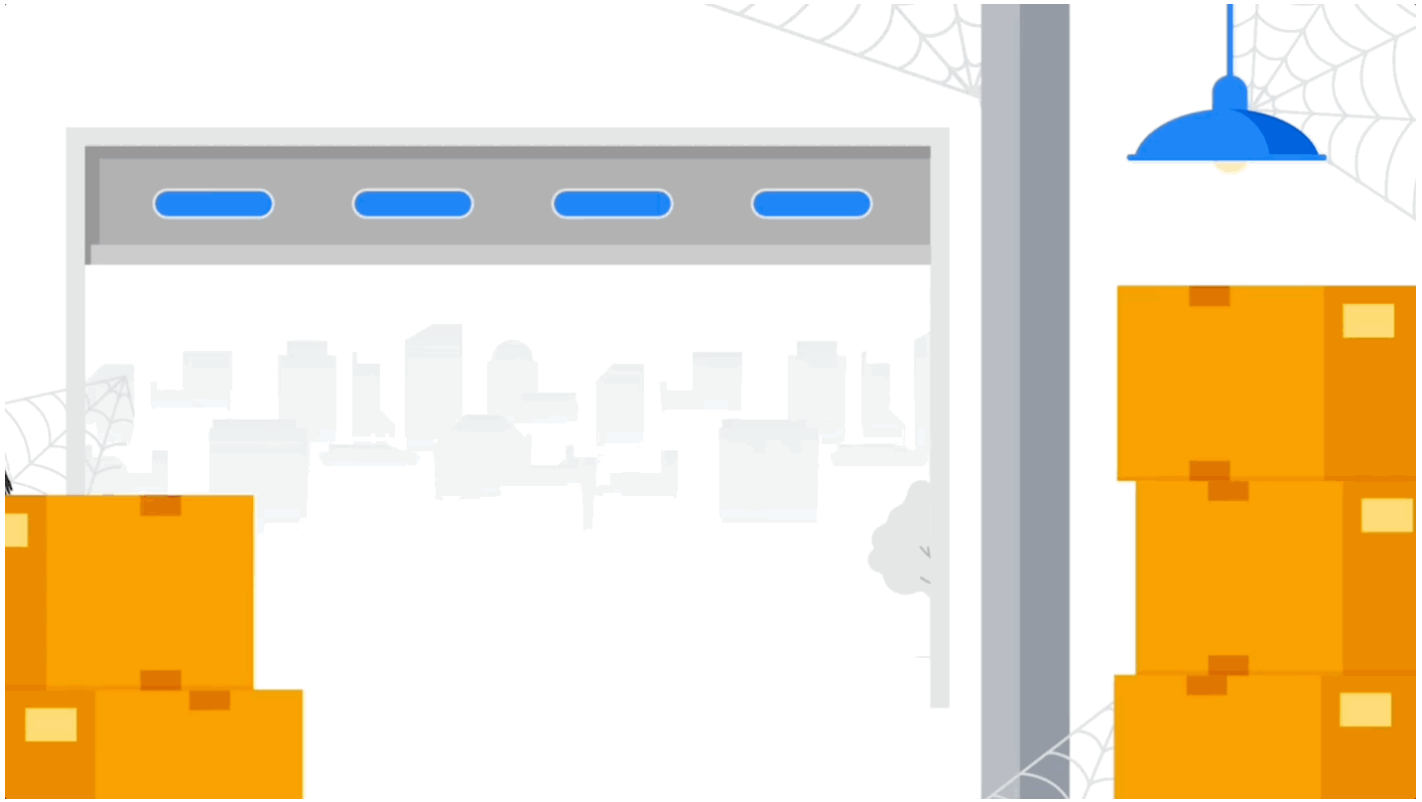
- The process of formatting data and removing unwanted material.
- The goal is to remove anything that could create an error during analysis.
- This process includes tagging and consolidating duplicates, irrelevant entries, structural errors, and empty space.



- Once you have everything in the proper format, you can then filter out unwanted material.
- Often, it is helpful to render the data visually to reveal additional insights through charts, dashboards, and reports.
- Graphic tools are very useful in identifying patterns as well as in sharing information with others.



Example of an old warehouse



Exploratory data analysis

- The process of investigating, organizing, and analyzing datasets and summarizing their main characteristics often employs data wrangling and visualization methods.

The 6 Practices of EDA

Discovering

Structuring

Cleaning

Joining

Validating

Presenting

- These practices do not necessarily need to be followed in this order, and depending on the data team's needs and the type of data being studied, they may perform EDA in different ways.

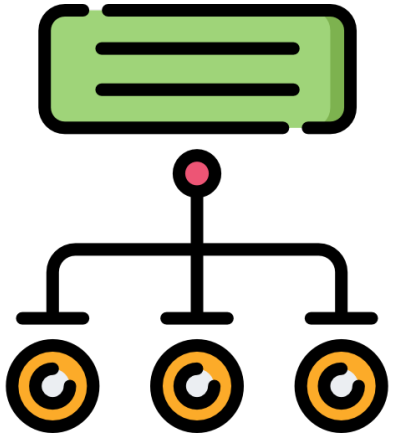
Discovering

- It is like walking into the room and removing coverings to get the number and types of items.
- During this practice, data professionals familiarize themselves with the data so they can start conceptualizing how to use it.

- They review the data and ask questions about it!
- What are the column headers? What do they mean?
- How many total datapoints are there?

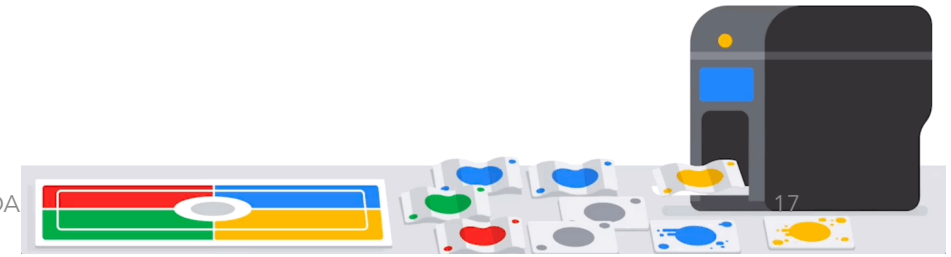
Structuring

- Now, in the old warehouse, you have to start organizing. (Into metal and non-metal categories)
- The process of taking raw data and organizing or transforming it to be more easily visualized, explained, or modeled.
- Structuring refers to categorizing and organizing data columns based on what is already in the dataset.
- Example: In terms of the calendar data, it might look like categorizing data into months or quarters rather than years.



Cleaning

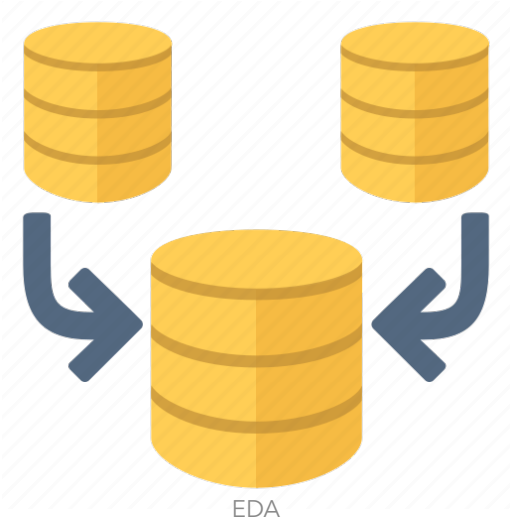
- Cleaning is the process of removing errors that may distort your data or make it less useful.
- Missing values, misspellings, duplicate entries, extreme outliers
- In the warehouse example, we might decide to put broken or unusable items in a separate box away from the other items.



Joining

- The process of augmenting or adjusting data by adding values from other datasets.

- In other words, we might add more value or context to the data by adding more information from other data sources.



Validating

- The process of verifying that the data is consistent and of high quality.

- Validating is the process of checking for misspellings and inconsistent numbers or date formats. And checking that the data cleaning process didn't introduce more errors.

Presenting

- Making your cleaned dataset or data visualizations available to others for analysis or further modeling.

- In other words, presenting practice is sharing what you've learned through EDA.

Data Visualization

- A graph, chart, diagram, or dashboard that is created as a representation of information.
- Presenting can come at any point in the EDA.
- Data visualizations aren't exclusive to the presenting practice. They should be used throughout the EDA.
- They help you understand data and point out trends and insights to others.

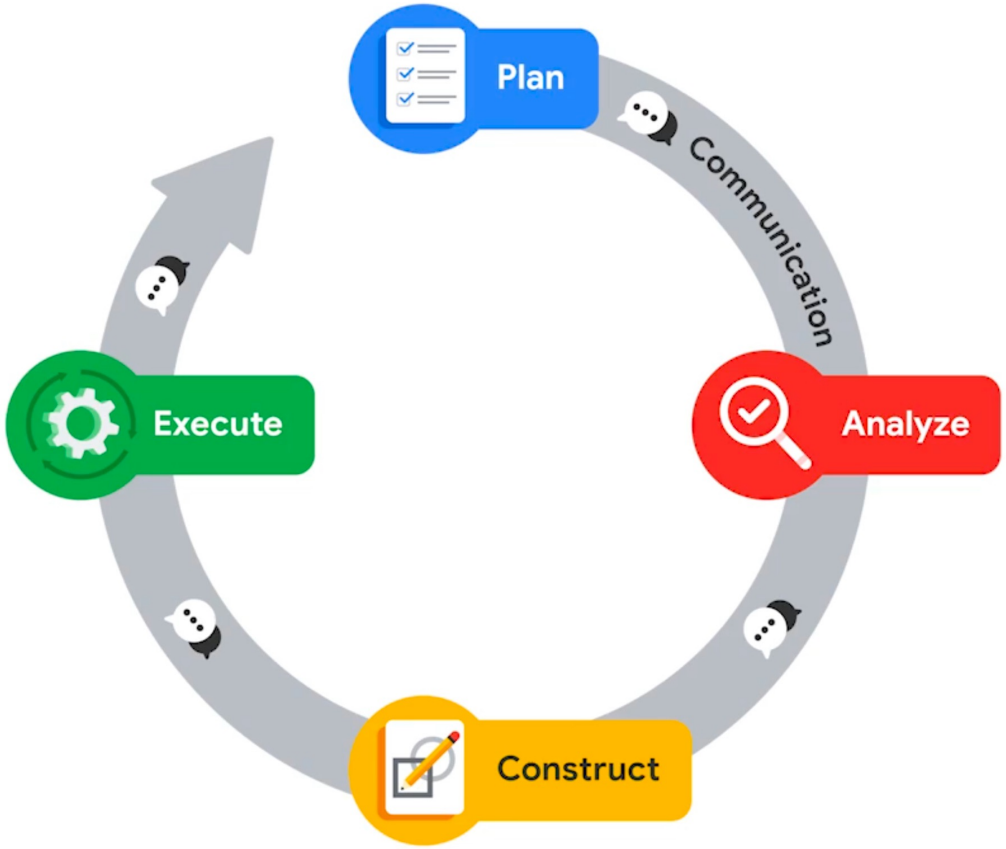
- The story you uncover should come from the data, not from your mind or biases in the data.

- It is your duty to convey your data in both an ethical and accessible way.

- In the workplace, communicating data **ethically** would be presenting sales numbers in context, year over year.
- So that **the rises and the falls** don't appear exaggerated in the data visualizations.

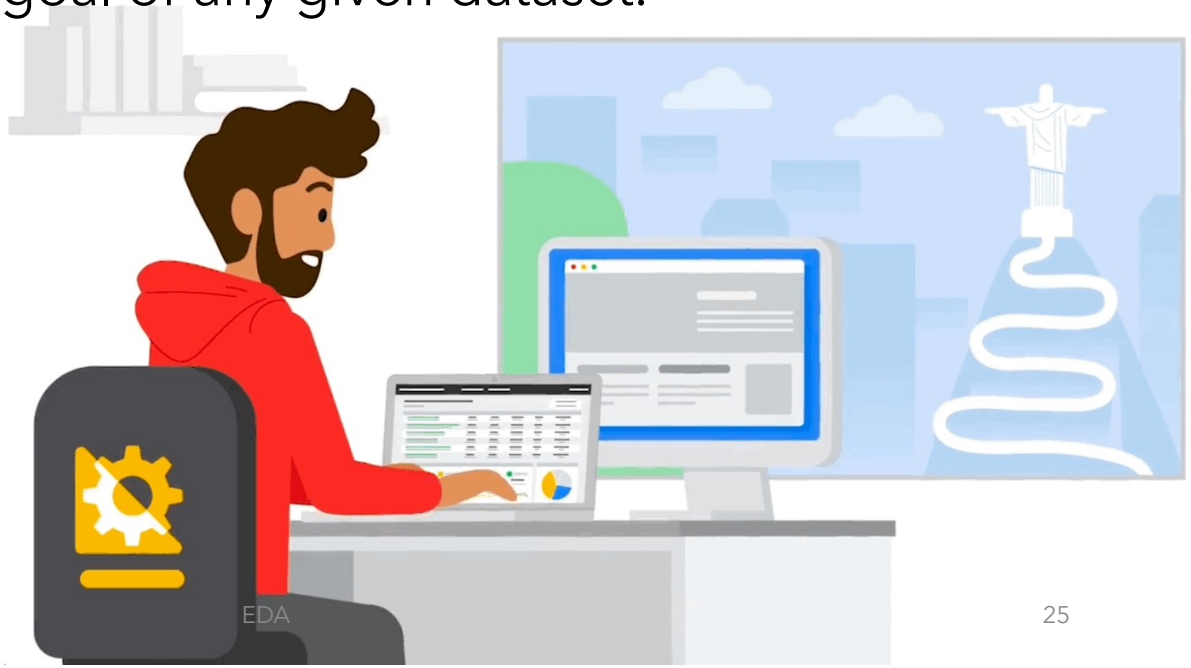
- Forgetting about the real purpose of what you were doing is natural!
- As data professionals, our curiosity and excitement for finding stories in data might cause us to forget the original purpose of our data exploration.
- We want to focus that curiosity on what questions need to be answered or what problems need to be solved.
- We seek a balanced mindset, one of targeted curiosity.
- This balance can be achieved by using PACE.

PACE Workflow:



PACE Workflow:

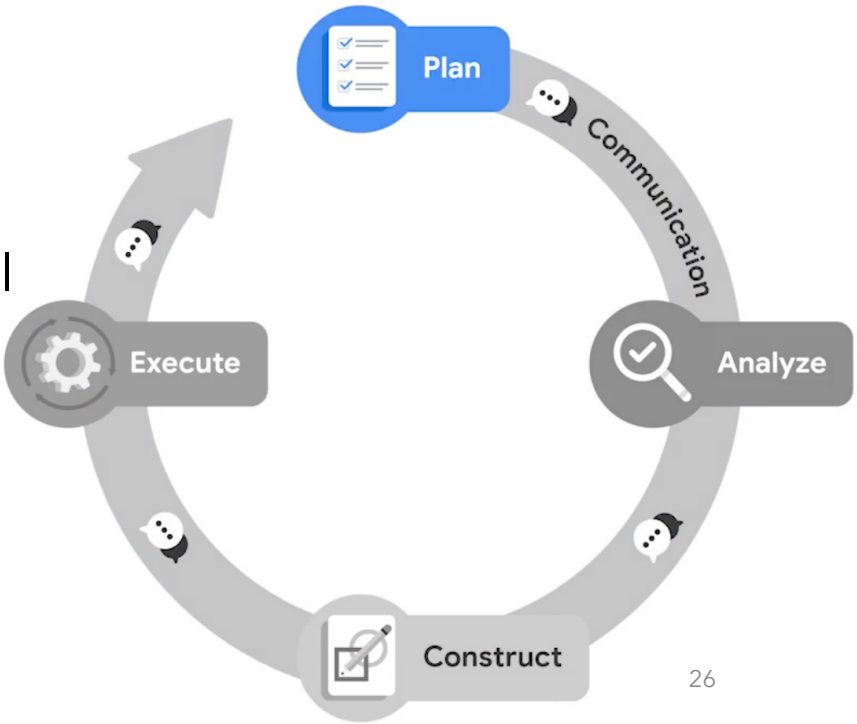
- PACE is the workflow some data professionals use to remain focused on the end goal of any given dataset.



Plan Stage

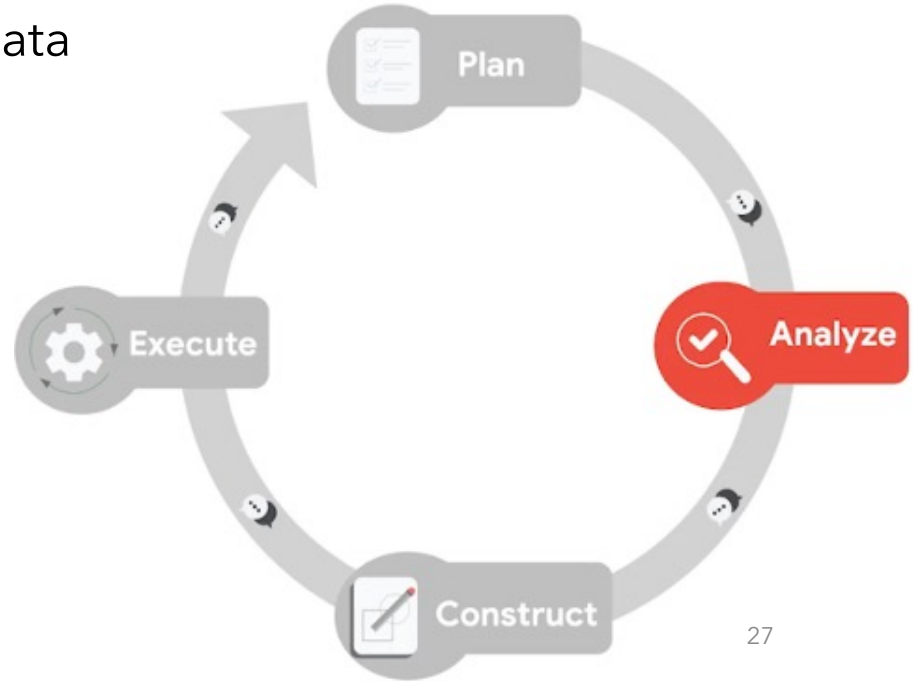
- We define the scope of our project.

- What are the goals of the project?
- What strategies will be needed?
- What will be the business or operational impacts of this plan?



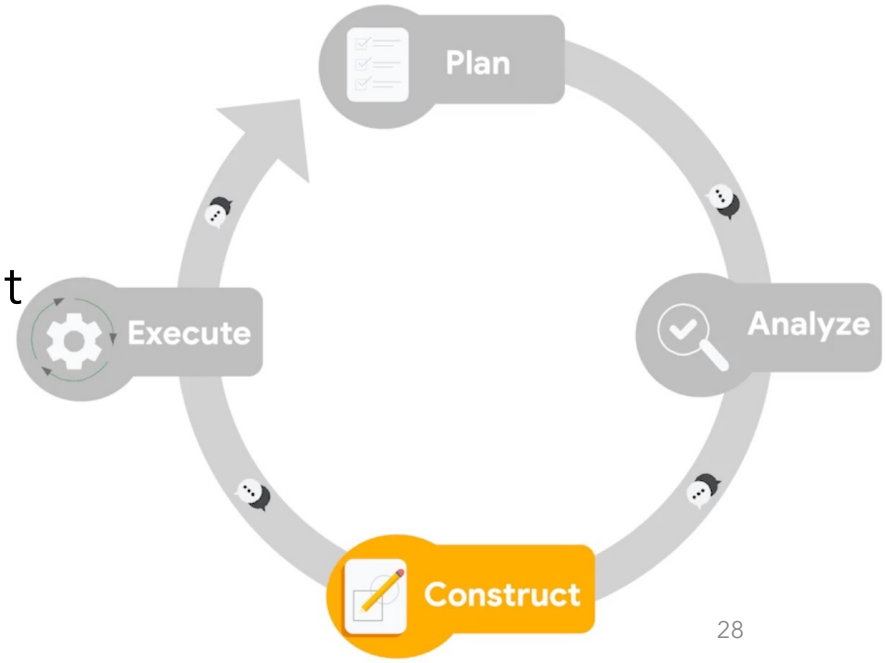
Analyze Stage

- We will engage with the data.
- Acquire data from primary and secondary sources.
- Clean, reorganize, and transform data for analysis.
- Engage in EDA
- Work with stakeholders.



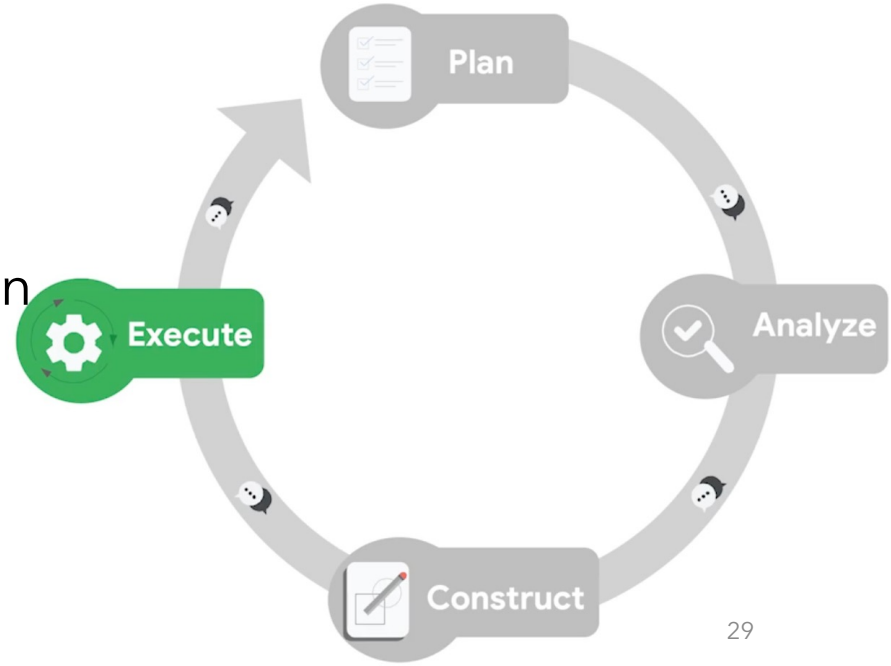
Construct Stage

- We will engage with the data.
- Build and revise machine learning models.
- Uncover relationships in the data.
- Applying statistical inference about data relationships.



Execute Stage

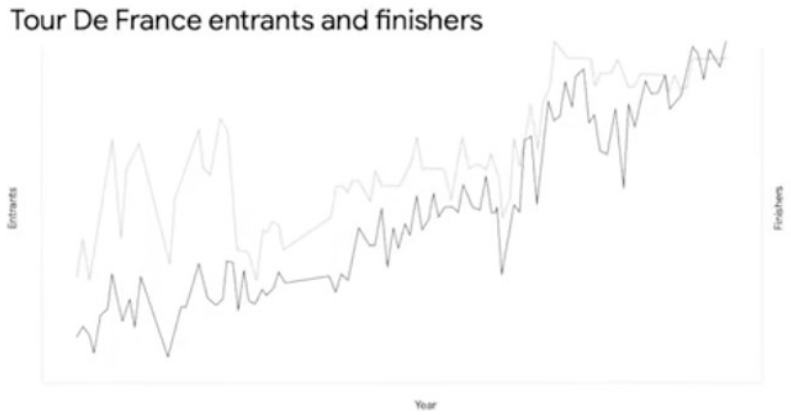
- Present findings to the internal and external stakeholders.
- Answer questions.
- Consider different viewpoints.
- Present recommendations based on the new findings in the data.

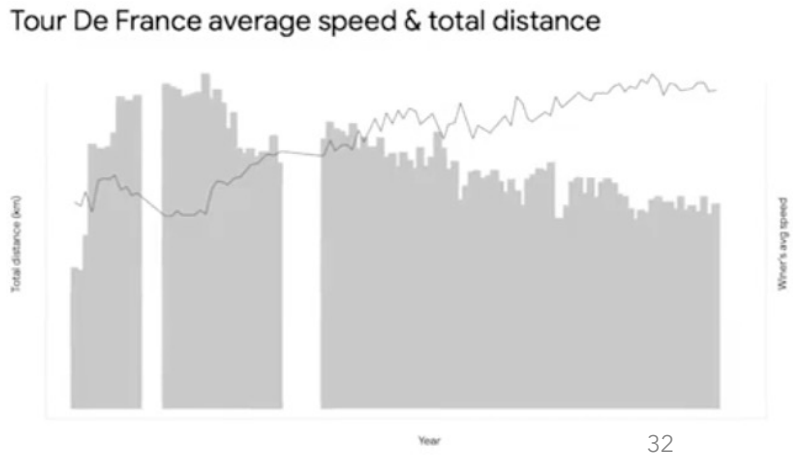


- You have to understand who the data is from.
- You have to understand why it was created.
- Understand what important caveats are sitting on top of it.
- Storytelling is the way that your insights make in to other people and really make change.

Data Visualizations vs. Data Tables

Year	Winner's avg speed	Total distance (km)	Number of stages	Finishers	Entrants
1903	25.6800	2,428,000	6	21	60
1904	25.2700	2,420,000	6	27	88
1905	27.1100	2,994,000	11	24	60
1906	24.4600	4,545,000	13	14	82
1907	28.4700	4,488,000	14	33	93
1908	28.7400	4,488,000	14	36	112
1909	28.6600	4,497,000	14	55	150
1910	29.1000	4,737,000	15	41	110
1911	27.3200	5,344,000	15	28	84
1912	27.7600	5,289,000	15	41	131
1913	26.7200	5,287,000	15	25	140
1914	26.8400	5,380,000	15	54	145
1919	24.0600	5,560,000	15	10	69
1920	24.0700	5,503,000	15	22	113
1921	24.7200	5,485,000	15	38	123
1922	24.2000	5,375,000	15	38	121



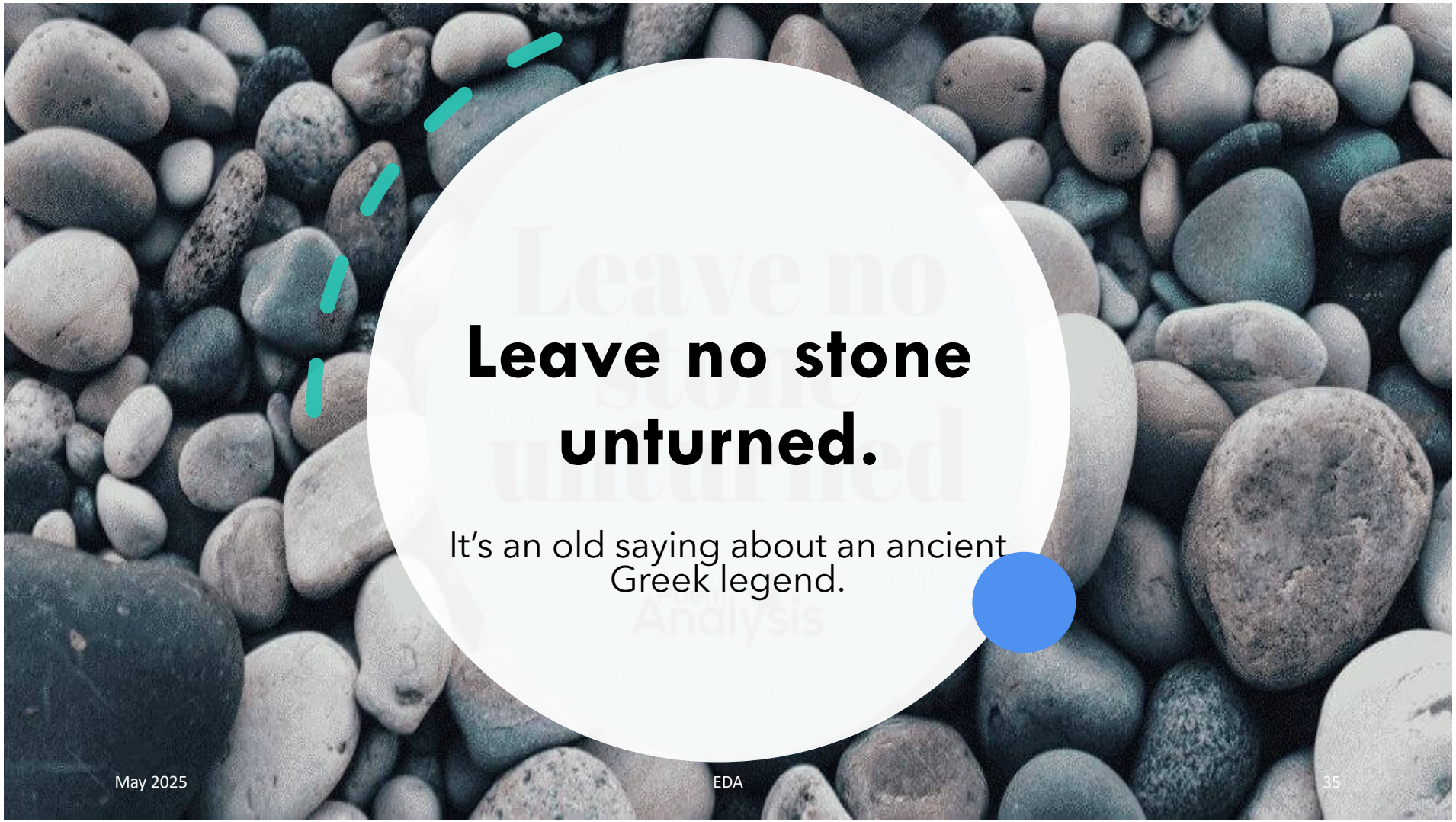


ID	Item number	Cost	Sale	Distributor
1	476905	23.98	34.13	Calco
2	238859	54.76	65.13	Calco
3	852858	1.22	3.14	Reineer
4	769940	345.76	412.24	Tycom
5	868959	34.98	41.14	Reineer
6	859948	45.67	52.54	Grainner
7	002848	NaN	245.65	Tycom
8	212737	65.78	72,13	Calco
9	883994	34.01	41.10	Burries
10	885040	12.66	19.12	Grainner

Column Headers

ID	Item number	Cost	Sale	Distributor
1	476905	23.98	34.13	Calco
2	238859	54.76	65.13	Calco
3	852858	1.22	3.14	Reineer
4	769940	345.76	412.24	Tycom
5	868959	34.98	41.14	Reineer
6	859948	45.67	52.54	Grainner
7	002848	NaN	245.65	Tycom
8	212737	65.78	72.13	Calco
9	883994	34.01	41.10	Burries
10	885040	12.66	19.12	Grainner

Missing data


Leave no stone unturned.
It's an old saying about an ancient Greek legend.

Data Formats

- Tabular files
- XML files
- CSV files
- Excel files
- DB files
- JSON files (JavaScript Object Notation)

Excel

Work item	Vendor	Labor	Equipment	Materials	Subcontr.	Subtotal	Markup%	Markup	Total
Permits/Fees	City of LA				$1,500.00	$1,500.00		$0.00	$1,500.00
Evacuation		$6,000.00	$8,000.00	$500.00		$14,500.00	15.00%	$2,175.00	$16,675.00
Utilities		$3,500.00	$2,500.00	$2,750.00	$1,000.00	$9,750.00	15.00%	$1,462.50	$11,212.50
Water Well						$0.00		$0.00	$0.00
Septic Tank						$0.00		$0.00	$0.00
Foundation	Connie's Concrete				$3,500.00	$3,500.00	5.00%	$175.00	$3,675.00
Concrete floor	Connie's Concrete				$1,900.00	$1,900.00	5.00%	$95.00	$1,995.00
Framing		$3,500.00	$1,500.00	$9,000.00		$14,00.00	15.00%	$2,100.00	$16,100.00
Roofing	Robert's Roofing				$3,500.00	$3,500.00	5.00%	$175.00	$3,675.00
Windows/doors	Wally's Windows				$8,000.00	$8,000.00	5.00%	$400.00	$8,400.00
Siding						$0.00		$0.00	$0.00
Electric	Ernie's Electric				$18,500.00	$18,500.00	5.00%	$925.00	$19,425.00
Plumbing	Mac's Mechanical				$16,500.00	$16,500.00	5.00%	$825.00	$17,325.00
Insulation		$3,500.00		$1,000.00		$4,500.00	5.00%	$0.00	$4,500.00
Masonry	Mason's Masonry				$14,500.00	$14,500.00	5.00%	$725.00	$15,225.00
Drywall	Doug's Drywall				$12,500.00	$12,500.00	5.00%	$625.00	$13,125.00
Interior Trim	Doug's Drywall				$9,000.00	$9,000.00	5.00%	$450.00	$9,450.00
Painting	Paul's Painting				$13,500.00	$13,500.00	5.00%	$575.00	$14,175.00
TOTALS						**$145,650**	**6.8%**	**$6,800**	**$156,370**

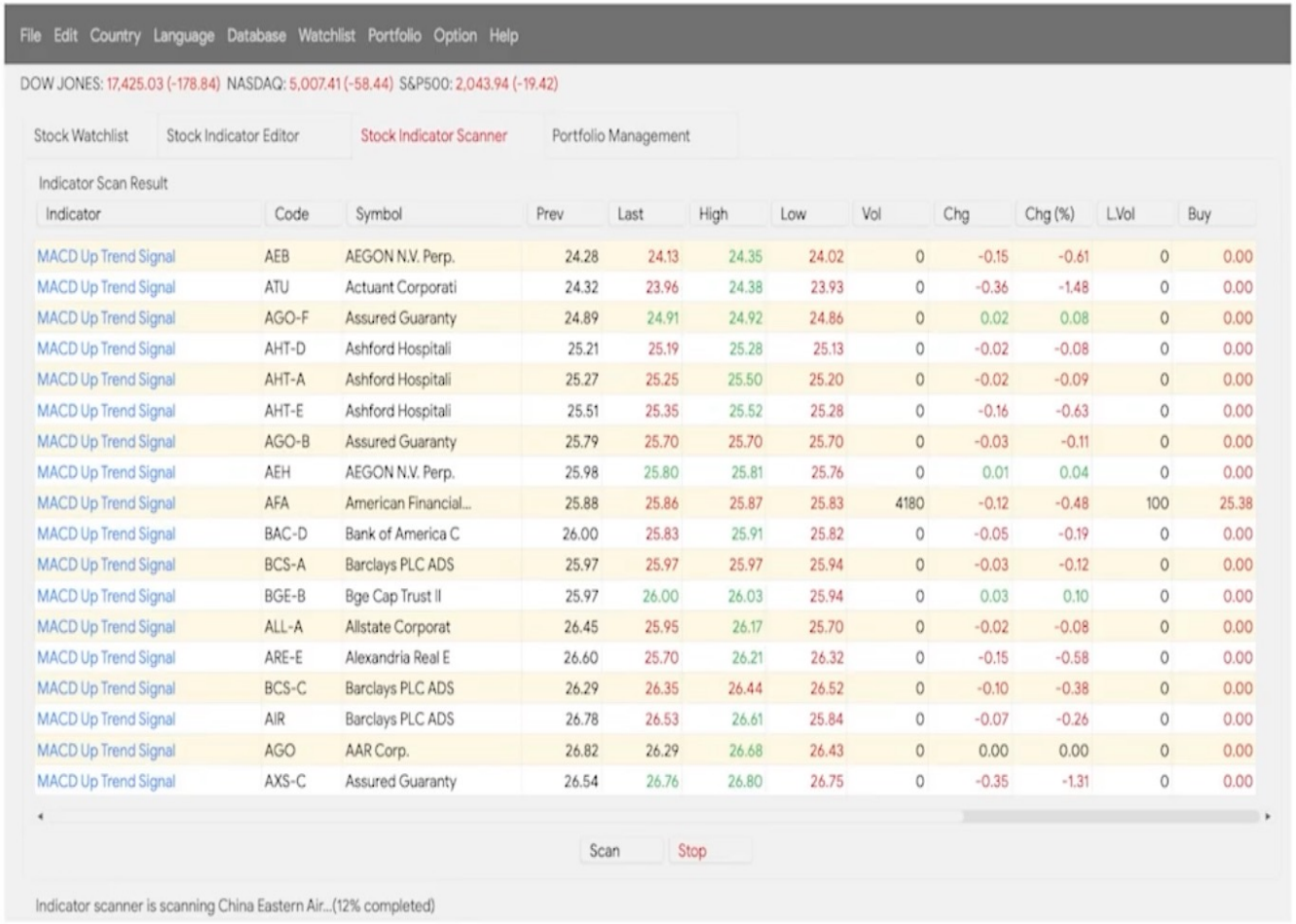
CSV

- Easy to
  - Read
  - Store
  - Create
  - Manipulate

1	Mouse biometric data, created Tue Dec 11 18:20:25 EST 2012
2	18
3	Pedro Xavier/?/?, ?, ?, PedroXavier_Brower_001.xml 0, 208, 40, 32, 1, 1,
4	Pedro Xavier/?/?, ?, ?, PedroXavier_Brower_002.xml 1, 208, 34, 30, 1, 3,
5	Pedro Xavier/?/?, ?, ?, PedroXavier_Brower_003.xml 2, 208, 31, 30, 0, 1,
6	Pedro Xavier/?/?, ?, ?, PedroXavier_Brower_004.xml 3, 208, 37, 35, 0, 2,
7	Pedro Xavier/?/?, ?, ?, PedroXavier_Brower_005.xml 4, 208, 48, 48, 0, 0,
8	Pedro Xavier/?/?, ?, ?, PedroXavier_Brower_006.xml 5, 208, 38, 35, 1, 1,
9	Robert Xavier/?/?, ?, ?, RobertXavier_Other_001.xml 6, 208, 1, 1, 0, 0,
10	Robert Xavier/?/?, ?, ?, RobertXavier_Other_002.xml 7, 208, 27, 26, 0,
11	Robert Xavier/?/?, ?, ?, RobertXavier_Other_003.xml 8, 208, 27, 27, 0,
12	Robert Xavier/?/?, ?, ?, RobertXavier_Other_004.xml 9, 208, 50, 44, 1,
13	Robert Xavier/?/?, ?, ?, RobertXavier_Other_005.xml 10, 208, 33, 32, 0,
14	Robert Xavier/?/?, ?, ?, RobertXavier_Other_006.xml 11, 208, 54, 54, 0,
15	Robert Xavier/?/?, ?, ?, RobertXavier_Other_007.xml 12, 208, 22, 22, 0,
16	Robert Xavier/?/?, ?, ?, RobertXavier_Other_008.xml 13, 208, 42, 38, 1,
17	Robert Xavier/?/?, ?, ?, RobertXavier_Other_009.xml 14, 208, 24, 24, 0,

Database

- Great for
  - Searching
  - Storage
- Use SQL



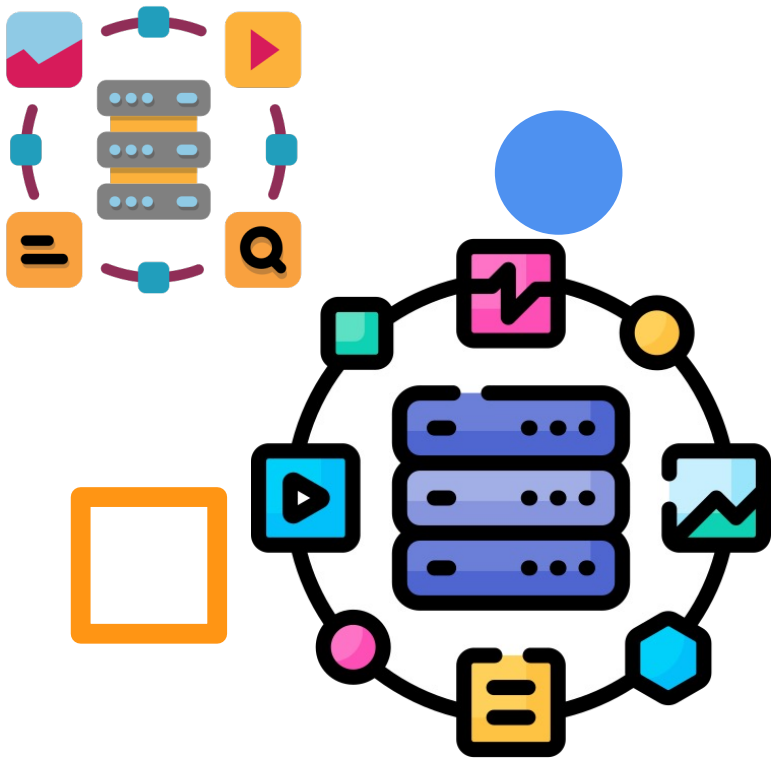
JSON

- Saved in JavaScript format
- It more closely resembles Python code, albeit with a distinct syntax, set of functions, and formatting.
- They have a small message size.
- Readable by almost any programming language.
- Programmers can easily distinguish between strings and numbers.


```
"users": [
  {
    "userId": 1,
    "firstName": "Jon",
    "lastName": "Smith",
    "phoneNumber": "111-111-1111",
    "emailAddress": "jsmith@info.com"
  },
  {
    "userId": 2,
          "firstName": "Bethany",
    "lastName": "Brown",
    "phoneNumber": "111-111-1112",
    "emailAddress": "bbrown@info.com"
  },
  {
    "userId": 3,
          "firstName": "Kate",
    "lastName": "Johnson",
    "phoneNumber": "111-111-1113",
    "emailAddress": "kjohnson@info.com"
  },
  {
    "userId": 4,
          "firstName": "David",
    "lastName": "Wilson",
    "phoneNumber": "111-111-1114",
    "emailAddress": "dwilson@info.com"
  },
  {
    "userId": 5,
          "firstName": "Sarah",
    "lastName": "Williams",
    "phoneNumber": "111-111-1115",
    "emailAddress": "swilliams@info.com"
  }
```

Types of Data

- First-party data
- Second-party data
- Third-party data



Missing data is often encoded as:

- N/A
- NaN ('Not a Number')
- [blank]

- Missing data (or null values): A value that is not stored for a variable in a set of data.

Let’s imagine a scenario about sleep habits

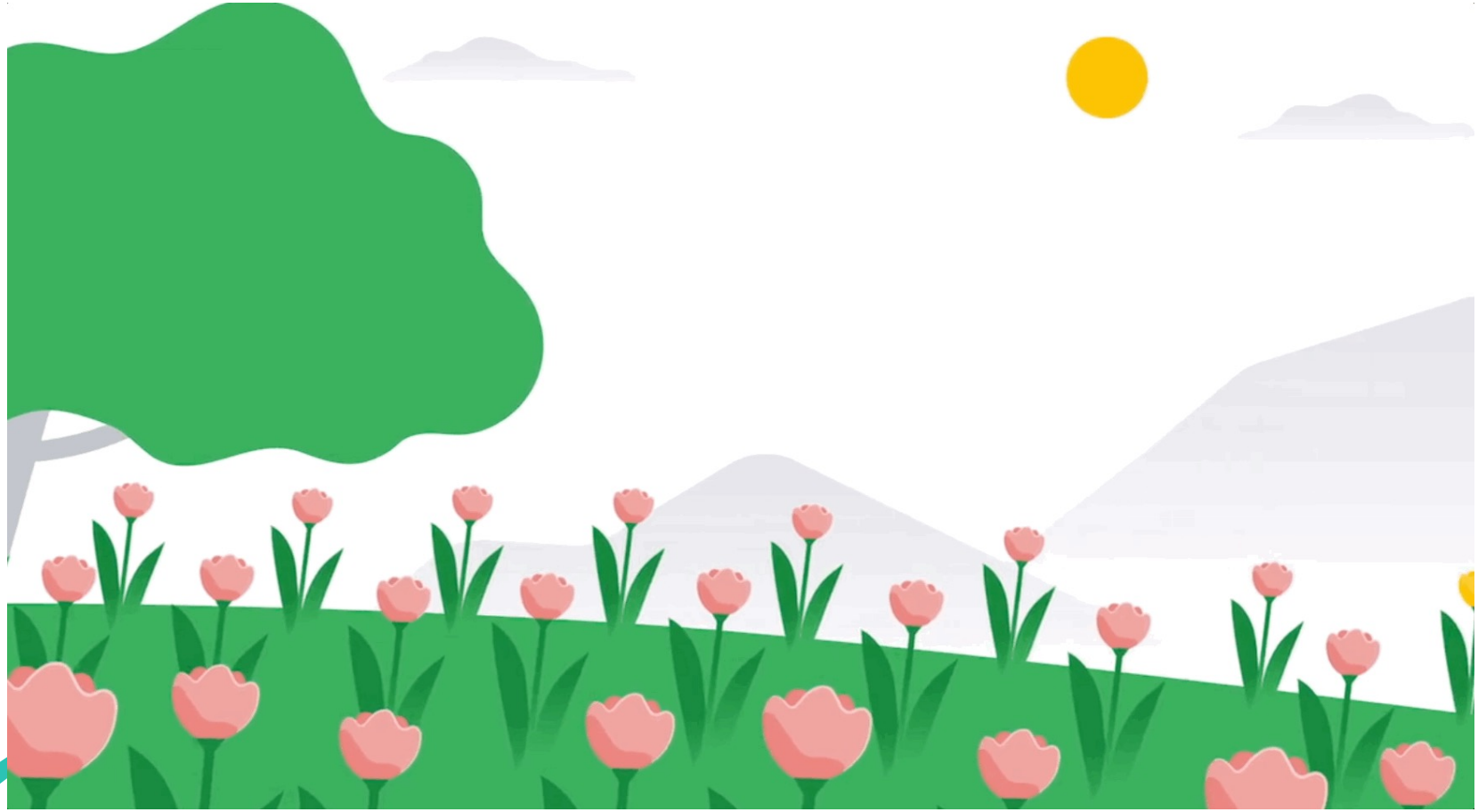
What to do with missing data

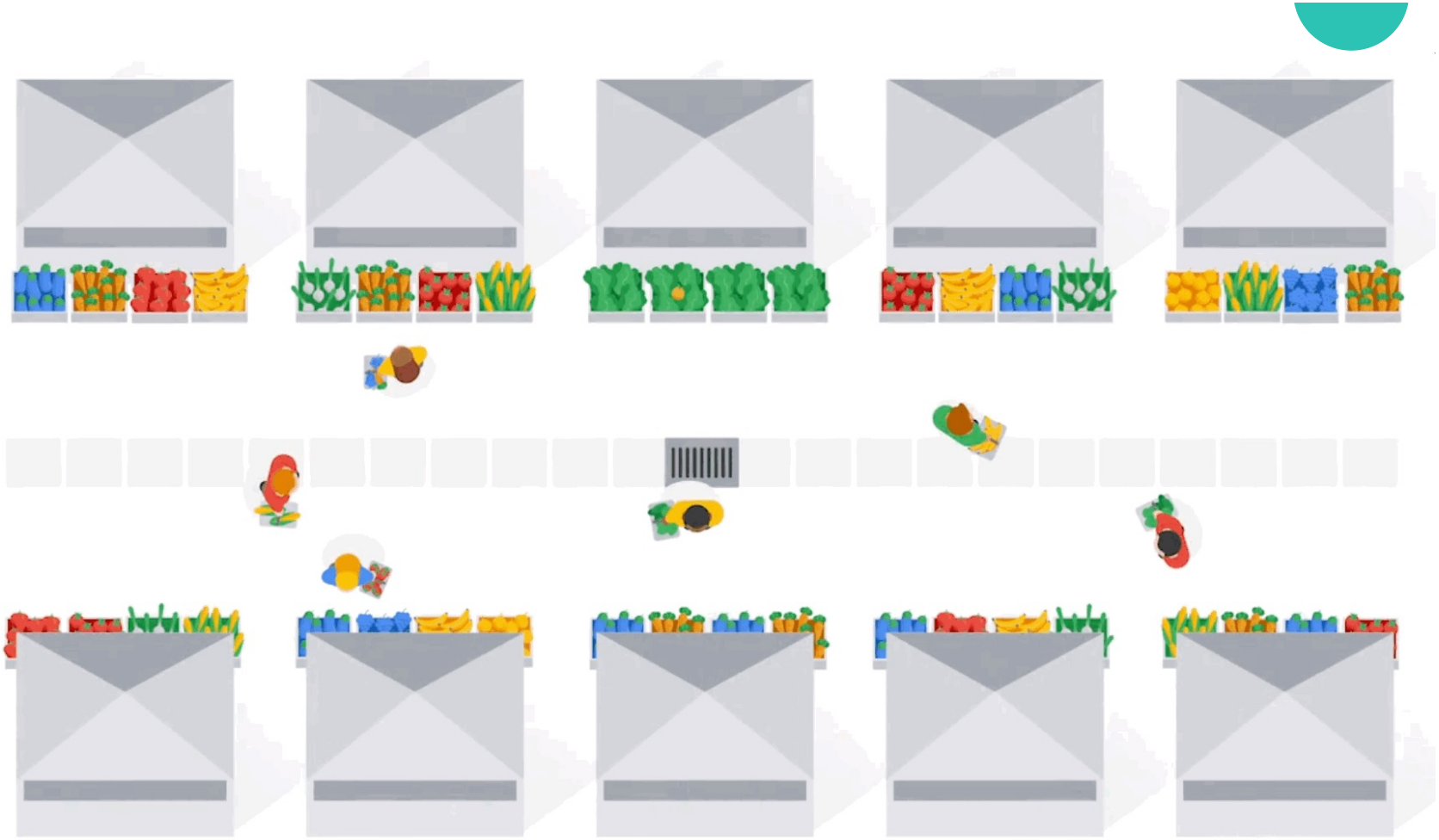
- Request for missing values to be filled in by the owner of the data.
- Delete the missing column(s), row(s), or value (s).
- Create a NaN category
- Drive new representative values (Filling in the missing data.)

Drive new representative value(s) strategy

- Forward filling
- Backward filling (backfilling)
- Deriving mean values


Outliers
What are they?
How to deal with them?

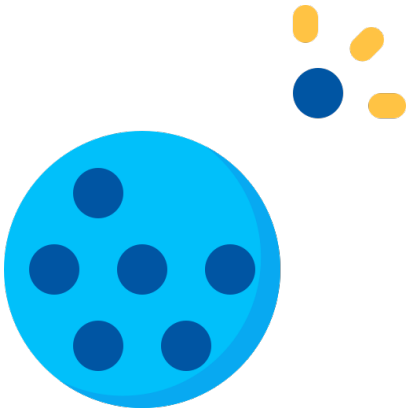




Outliers

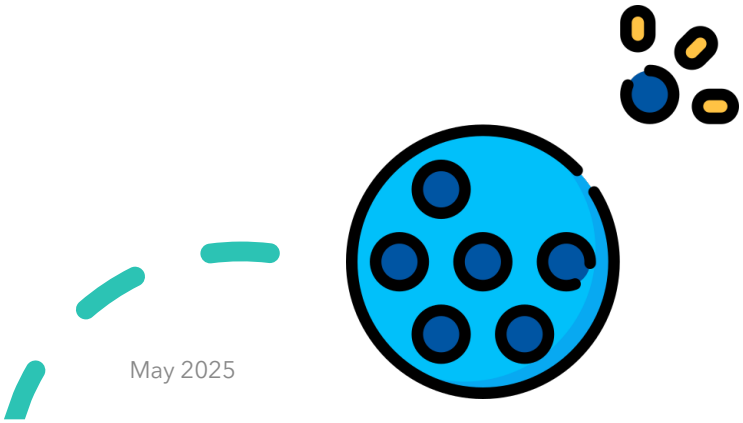
- These are observations that are an abnormal distance from other values or an overall pattern in a data population.

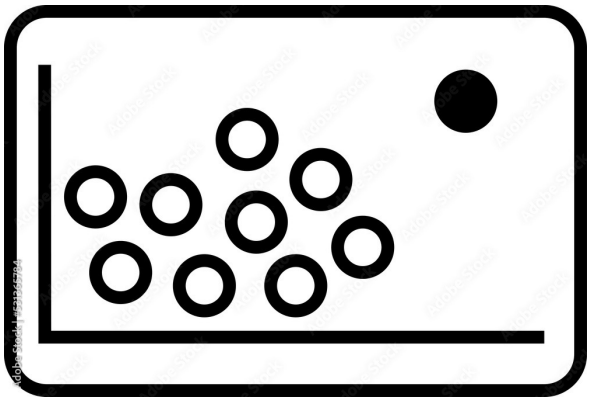
- As a data professional, you should be fully aware of the beginnings and ends, highs and lows, and extreme points of your data across every variable in the dataset. (These values can often be the outliers.)



3 types of outliers

- Global outliers
- Contextual outliers
- Collective outliers





Global outliers

- Global outliers are values that are entirely different from the overall data group and have no association with any other outliers.

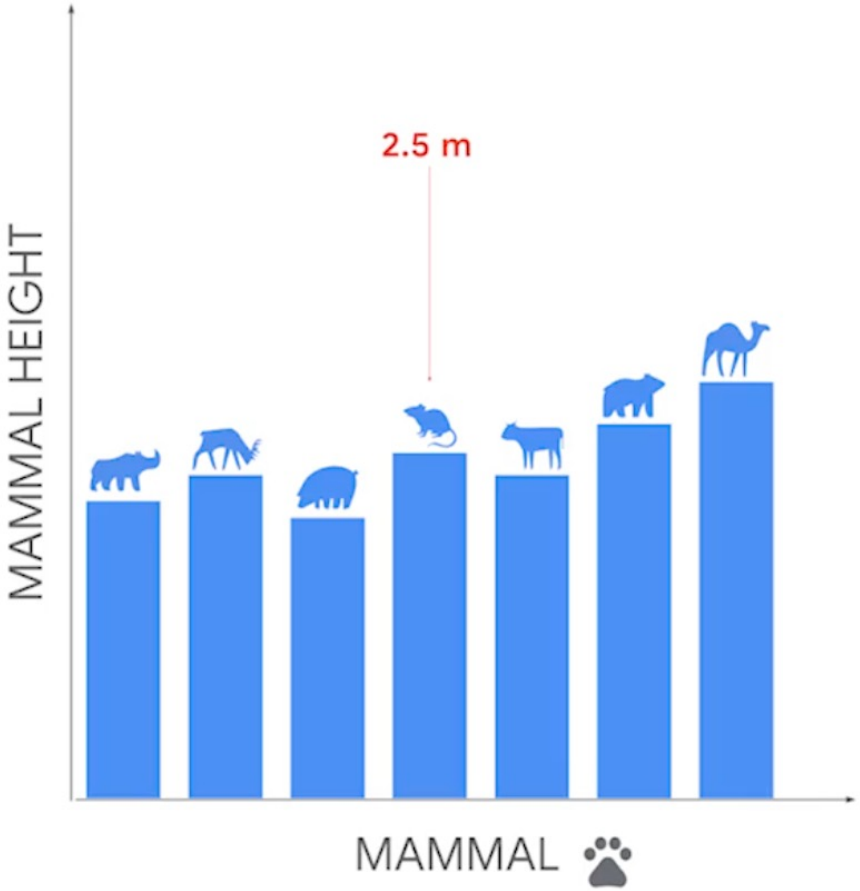
- They may be inaccuracies, typographical errors, or just extreme values you typically don't see in a dataset.
- Global outliers should be thrown out to create a predictive model.

Contextual outliers

- Normal data points under certain conditions but become anomalies under most other conditions.

- These outliers are more common in time series data.

Outlier in only one category



Collective outliers

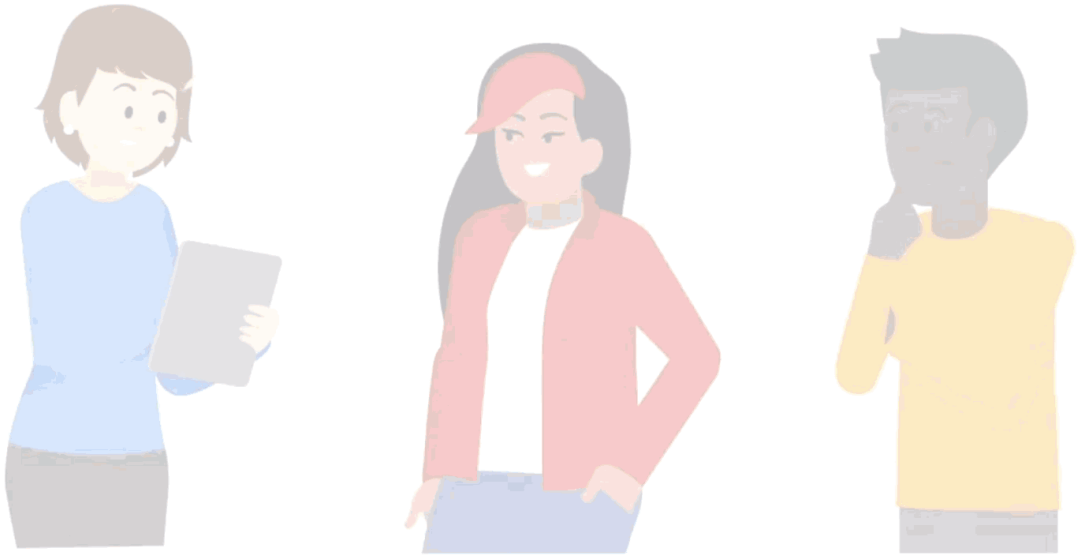
- A group of abnormal points that follow similar patterns and are isolated from the rest of the population.

- Example: Store parking lot after working hours!

Strategy about outliers

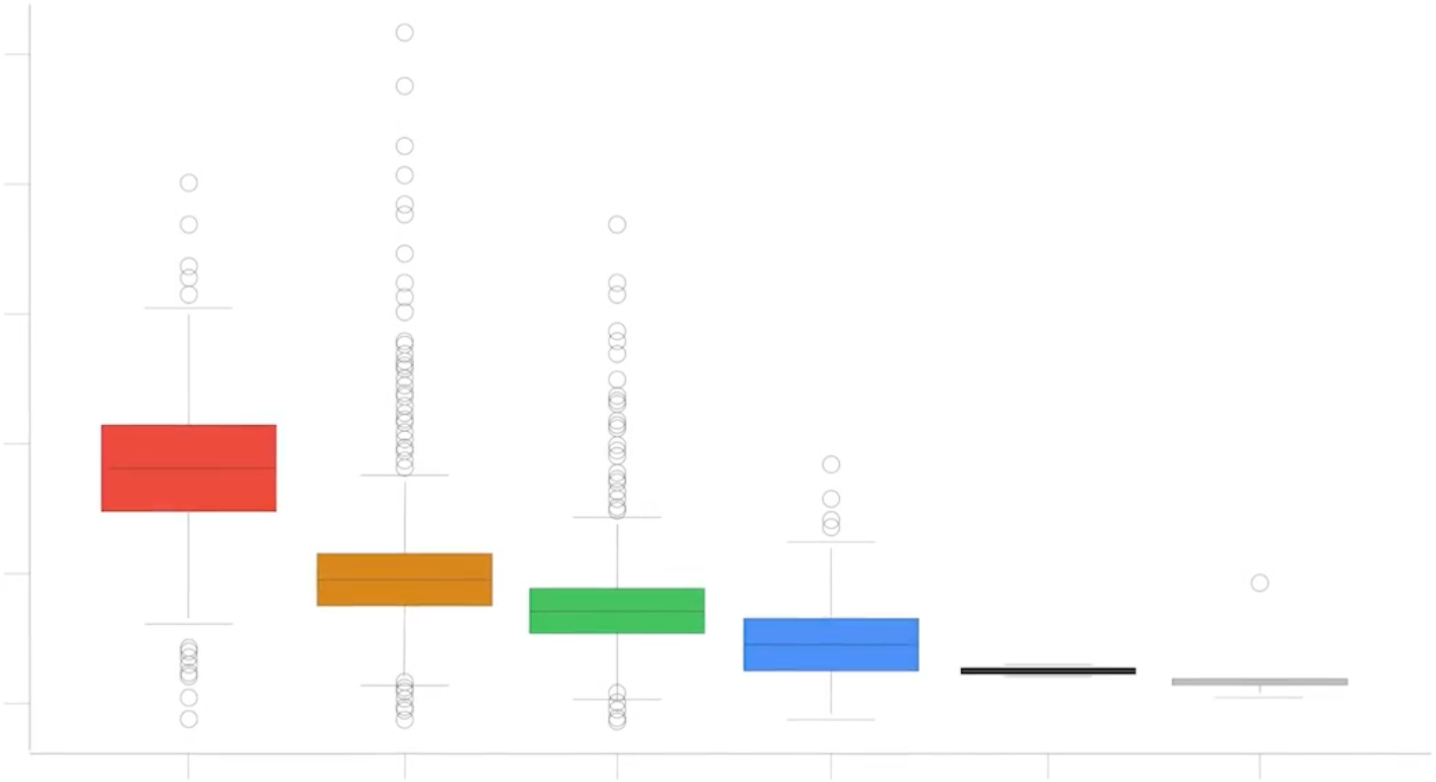
- It is essential that our EDA includes a strategy for dealing with outliers.
- It is up to us to decide how to represent them or remove them completely.
- The decision on what to do must always be made in the context of the dataset and the business plan for the data.
- Always consider the ethical implications of any decision that we make about outliers.

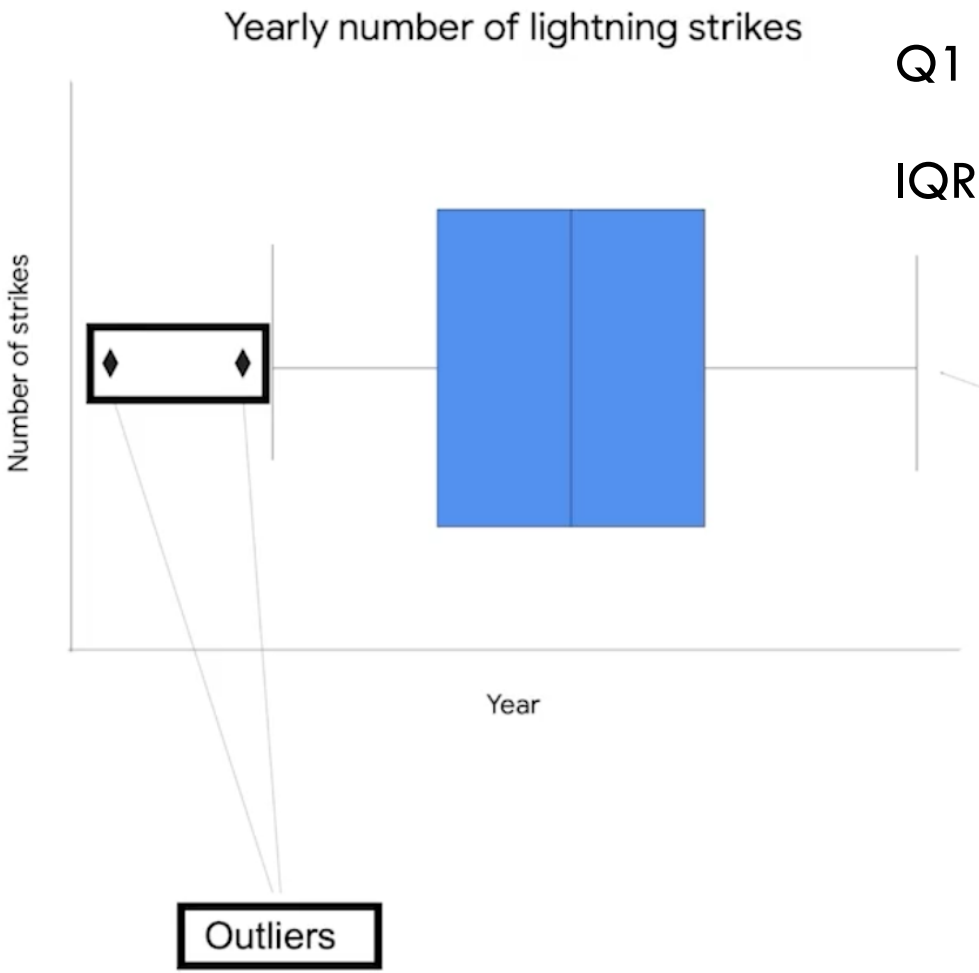
A new AD strategy and the absence of several marketing people



The business team will use that information

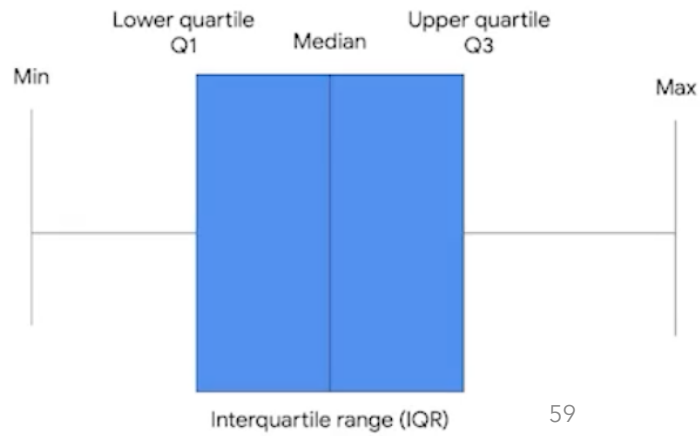
Box plots





$$Q1 - (1.5 \times IQR) , Q3 + (1.5 \times IQR)$$

$$IQR = Q3 - Q1$$



Parham Zilouchian Moghaddam

Email: parham_zm@hotmail.com

Telegram: @parham_zm