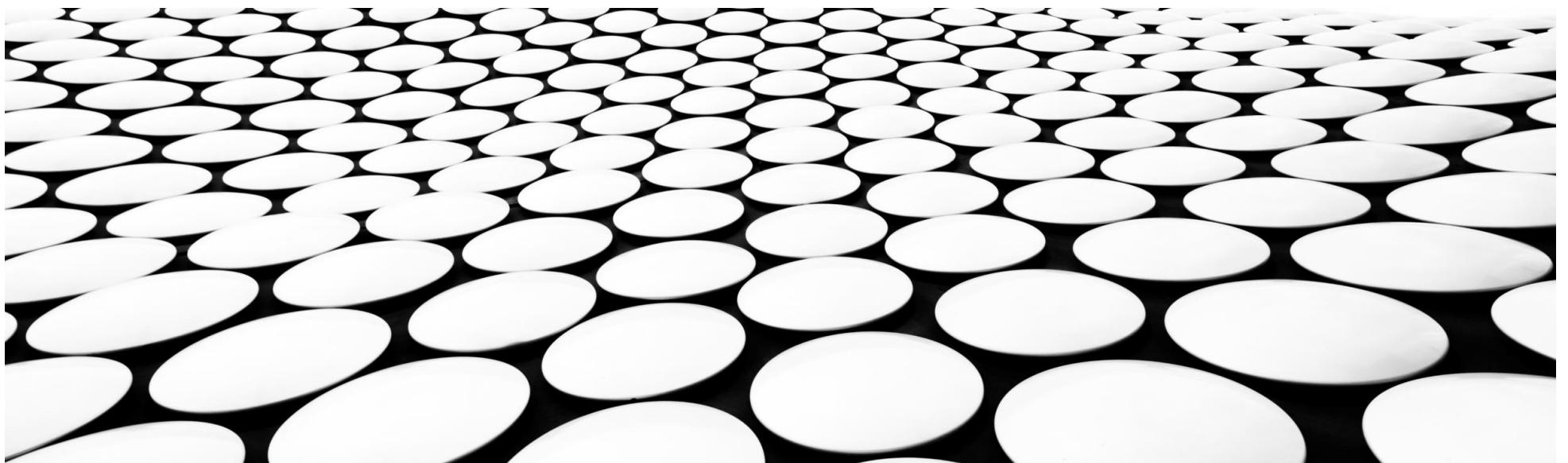


## Contents

Week 1: Insights from data	2
Week 2: Exploratory Data Analysis Part 1	58
Week 2: Exploratory Data Analysis Part 2	113

# DERIVING INSIGHTS FROM DATA

*Input/Output in Data Science*





## DERIVING INSIGHTS FROM DATA

- Types of insights
- Types of data

# Back to Data Science Definition(s)



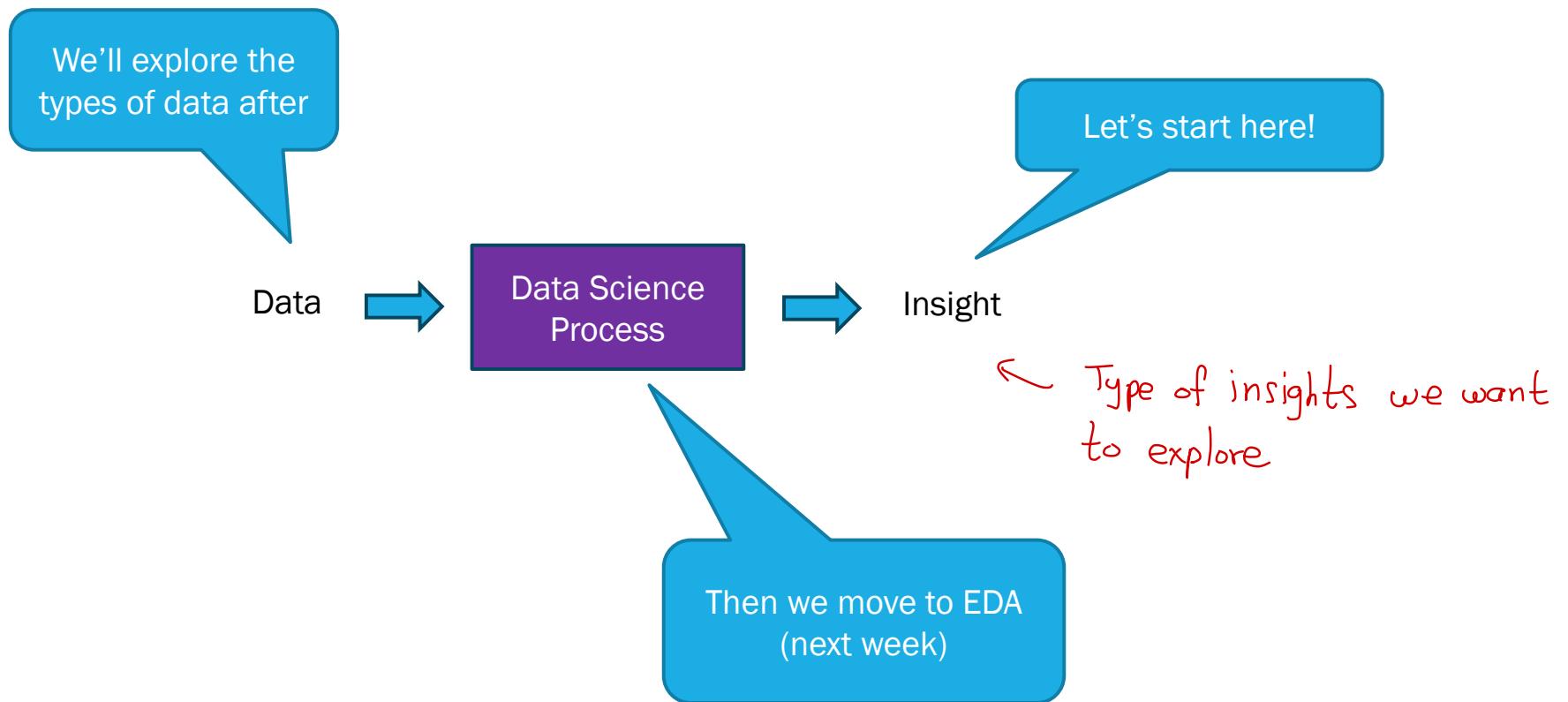
Data science is an interdisciplinary academic field that uses statistics, scientific computing, scientific methods, processing, scientific visualization, algorithms and systems to extract or extrapolate knowledge and **insights** from potentially noisy, structured, or unstructured **data**.

Data science combines math and statistics, specialized programming, advanced analytics, artificial intelligence (AI) and machine learning with specific subject matter expertise to uncover actionable **insights** hidden in an organization's **data**. These insights can be used to guide decision making and strategic planning.



Data science is the study of **data** to extract meaningful **insights** for business. It is a multidisciplinary approach that combines principles and practices from the fields of mathematics, statistics, artificial intelligence, and computer engineering to analyze large amounts of data. This analysis helps data scientists to ask and answer questions like what happened, why it happened, what will happen, and what can be done with the results.

Common: **data as input, insights as output**





Insights

## 8 Types of Data Analytics to Improve Decision-Making



Data analytics helps businesses learn from the past, optimize existing resources, and plan for the future. Find 8 ways to leverage data analysis here.

Contents

Jan 4, 2024 · 8 min read

# The Four Main Types of Data Analytics For Decision-Making

Descriptive

→ What happened? *Storytelling*

Diagnostic

→ Why did something happen?

Predictive

→ Based on what we know now, what will happen in the future?

Prescriptive

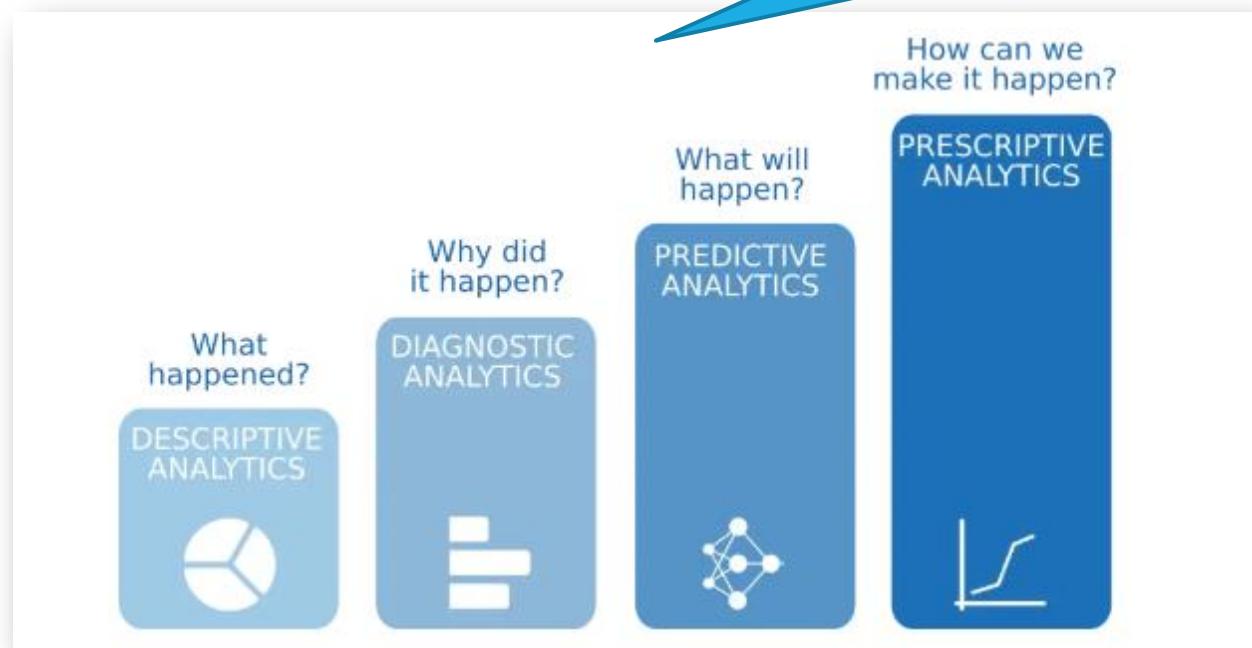
→ Should we implement some rules or regulation or provide advice based on what we know?

*Modify future*

## What is advanced analytics?

Tutorials  
© Pablo Martín August 31, 2023

Many sources (as this one) show these 4 types of analytics





**Descriptive analytics** serves as the foundational layer of data analysis. This type of analysis involves examining historical data to gain an understanding of past events. This type of analysis answers the question, “What happened?” in order to plan for the future. Descriptive analytics helps to summarize and visualize data trends, providing the context needed to assess the current state of affairs and identify potential areas of concern or opportunity.

The table below illustrates these parameters for the total amount of money spent by a customer in an online store.

	Minimum	Maximum	Mean	Standard deviation
Total amount (USD)	12	500	51	56

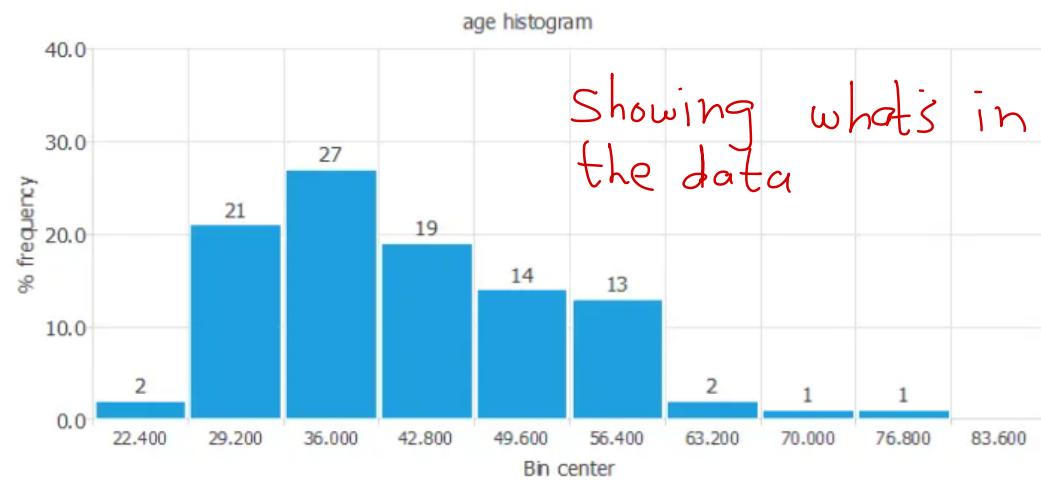
Descriptive  
Statistics



Example in sales

Histogram  
Visualization

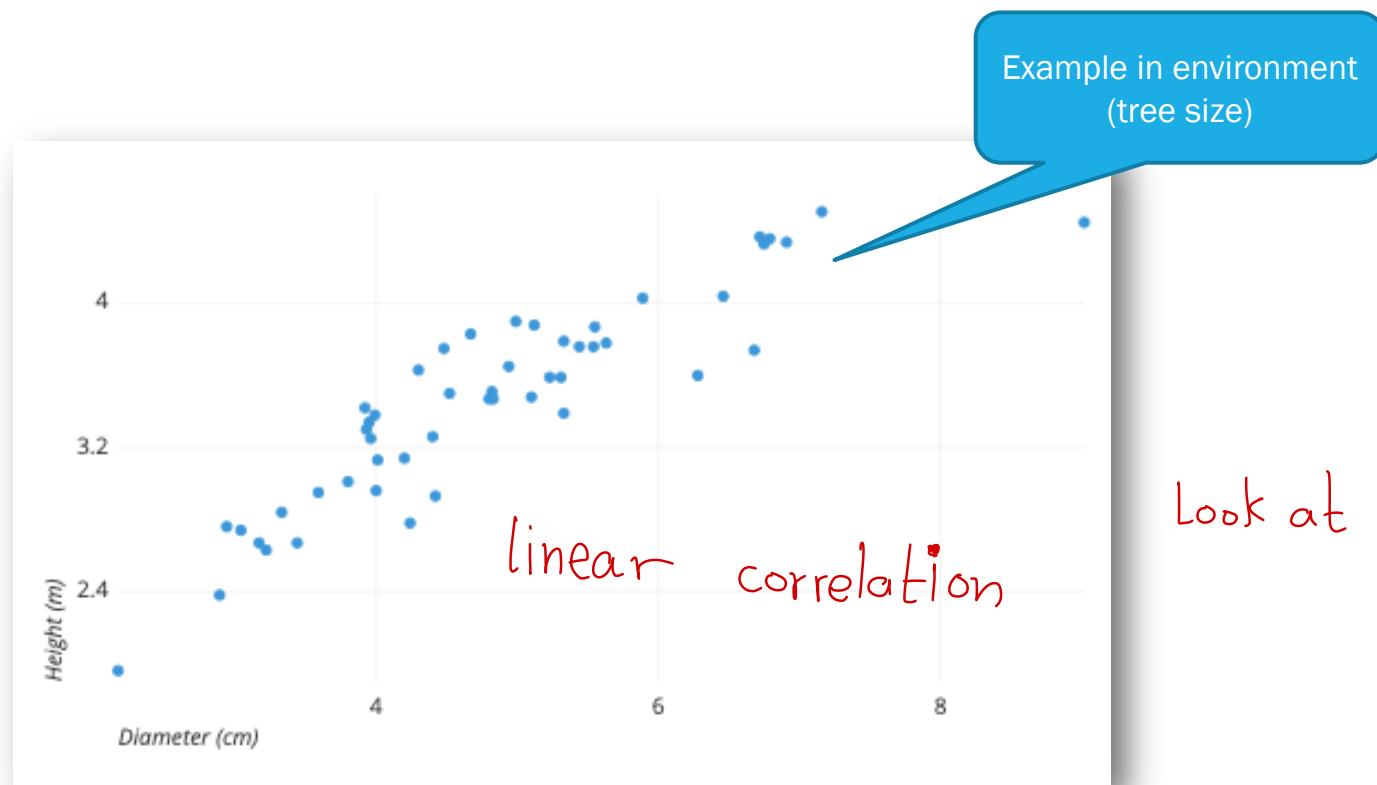
For example, the following figure depicts a histogram for the age of our customers.



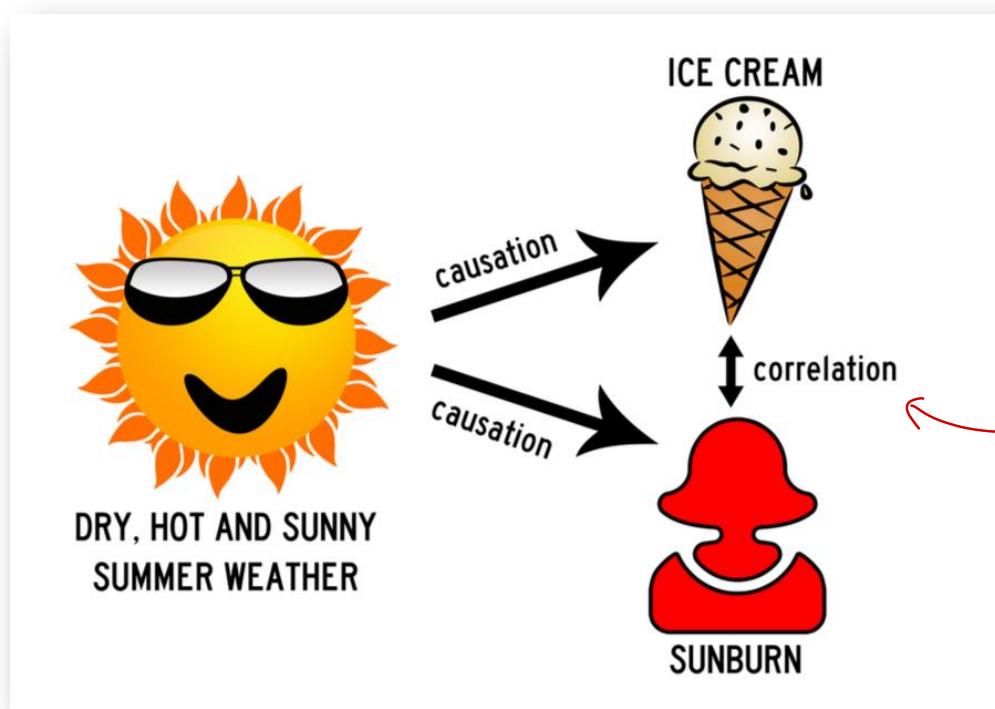


While descriptive analytics looks at the past, diagnostic analytics digs deeper into why specific events or trends occurred, seeking to uncover the root causes of particular outcomes or issues.

Tricky sometimes



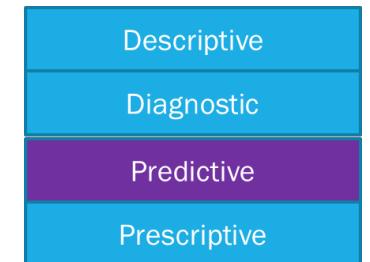
Look at correlation



Don't cause each other



Correlation  $\neq$  Causality

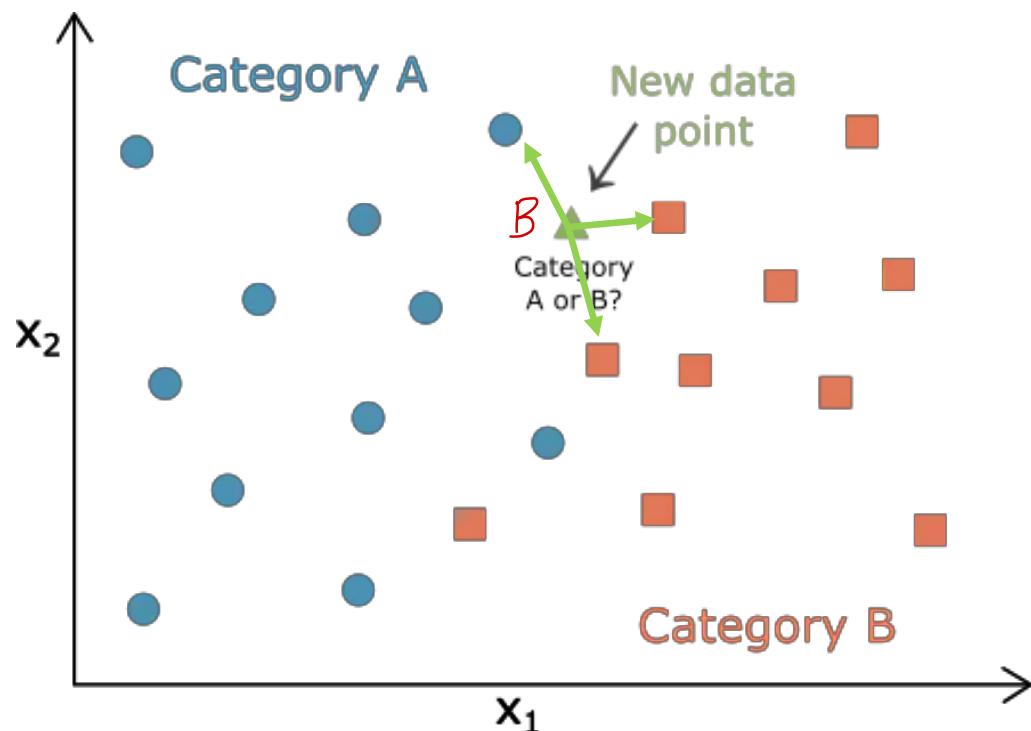


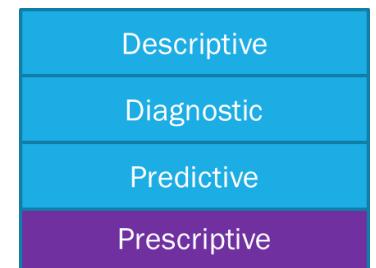
Predictive analytics takes data analysis a step further by forecasting future events or trends, leveraging historical data to build models that can make predictions about what is likely to happen. This type of analysis is an essential tool for proactive decision-making — while no one can see into the future, predictive analytics comes close.

## Logistic Regression



## K – Nearest neighbors





Prescriptive analytics goes beyond predicting future events and offers recommendations for actions that will achieve specific outcomes. This type of analytics is highly valuable in improving operational efficiency.



## 6 EXAMPLES OF PRESCRIPTIVE ANALYTICS IN ACTION

How can we  
make it happen?

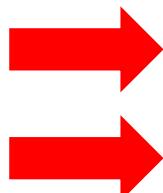
PREScriptive  
ANALYTICS



### 4. Banking: Fraud Detection

Another algorithmic use of prescriptive analytics is the detection and flagging of bank fraud. With the sheer volume of data stored in a bank's system, it would be nearly impossible for a person to manually detect any suspicious activity in a single account. An algorithm—trained using customers' historical transaction data—analyzes and scans new transactional data for anomalies. For instance, perhaps you typically spend \$3,000 per month, but this month, there's a \$30,000 charge on your credit card.

The algorithm analyzes patterns in your transactional data, alerts the bank, and provides a recommended course of action. In this example, the course of action may be to cancel the credit card, as it could have been stolen.



To help you learn... ask for examples.



ChatGPT

To help you learn the terms... explicitly type the different insights to remember them.

Knowing that data science insights are either descriptive,  
diagnostic, predictive or prescriptive, please provide  
some examples of those types of insights.



It's not an oracle.. It can  
make mistakes.

## 1. Descriptive Insights

**Definition:** These insights summarize past data to understand what happened.

- **Example (Environmental):** "The average air quality index (AQI) in the city has improved by 15% over the last decade due to increased green spaces."
- **Example (Medical):** "The hospital's patient admission rate increased by 20% during flu season compared to the annual average."
- **Example (Business):** "Sales increased by 10% during the holiday season compared to the previous quarter."

## 2. Diagnostic Insights

**Definition:** These insights analyze historical data to determine why something happened.

- **Example (Environmental):** "The spike in carbon dioxide levels last quarter was due to a 25% increase in industrial activity."
- **Example (Medical):** "The high readmission rate for patients was linked to incomplete post-discharge instructions provided by nursing staff."
- **Example (Business):** "Customer churn increased because of a delay in shipping during the promotional period."

### 3. Predictive Insights

**Definition:** These insights use historical data and machine learning models to predict what might happen in the future.

- **Example (Environmental):** "The model predicts a 30% probability of severe drought next summer based on current rainfall patterns."
- **Example (Medical):** "Patients with high cholesterol and low activity levels have a 70% increased risk of heart disease within the next five years."
- **Example (Business):** "Based on sales trends, revenue is expected to increase by 15% in the next quarter."

## 4. Prescriptive Insights

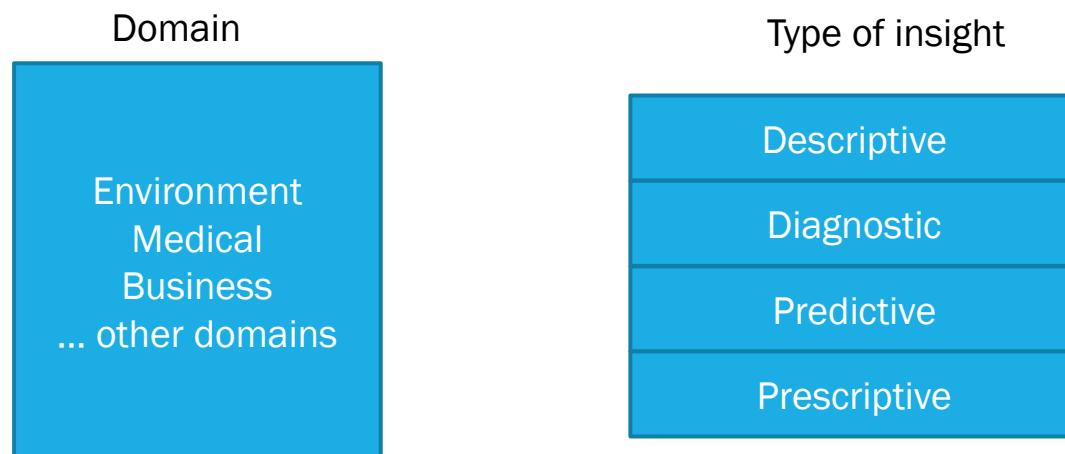
**Definition:** These insights recommend actions based on predictive and diagnostic analyses to achieve desired outcomes.

- **Example (Environmental):** "To reduce air pollution by 20%, prioritize policies that limit industrial emissions and increase public transportation usage."
- **Example (Medical):** "Patients with diabetes should have quarterly check-ups and follow a specific diet plan to reduce the likelihood of complications."
- **Example (Business):** "To increase customer retention, offer a loyalty program and improve response times for customer support inquiries."

Beyond our role  
as data scientists



Looking at many examples can be overwhelming... let's fix one variable.



Can you give one example each of descriptive, diagnostic, predictive and prescriptive for the health domain.



### What happened?

- "Over the past year, 25% of patients visiting the clinic reported symptoms of anxiety or depression, with the highest prevalence among individuals aged 18–35."

### Why did it happen?

- "The increase in anxiety and depression cases among young adults correlates with higher screen time and reduced physical activity during the pandemic."

### What might happen?

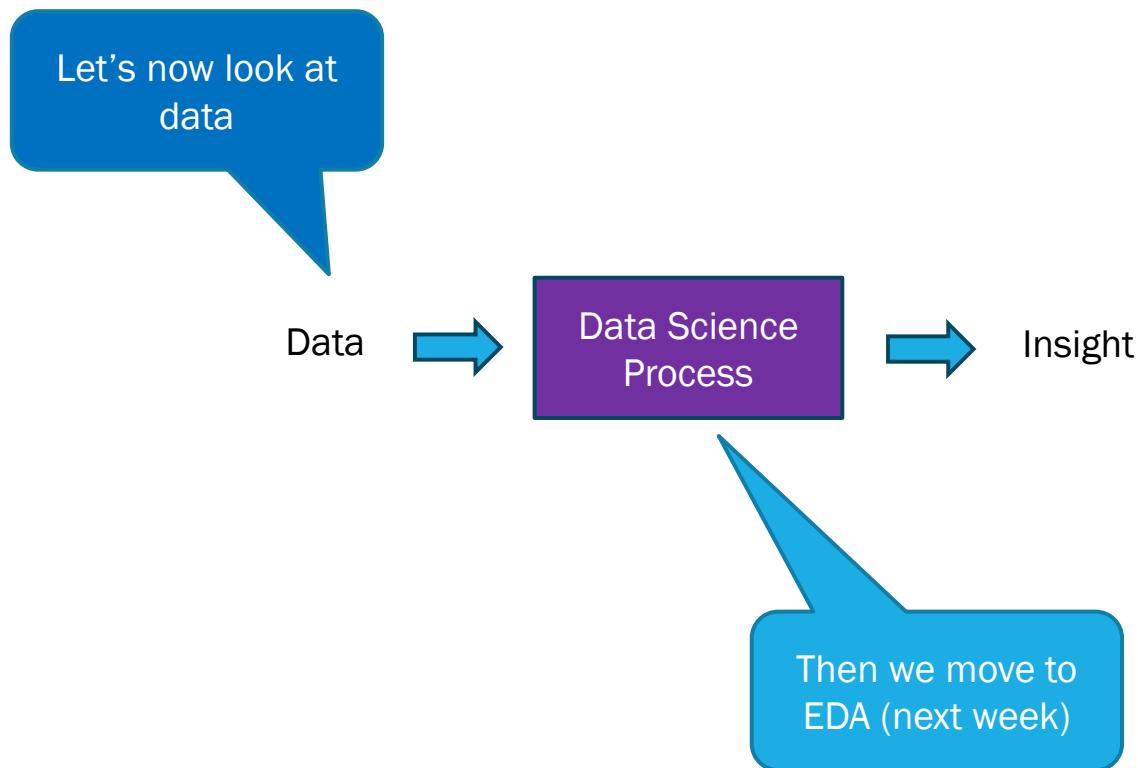
- "Patients with elevated blood pressure and a BMI over 30 have a 40% higher likelihood of developing Type 2 diabetes within the next five years."

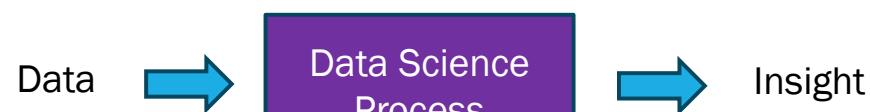
### What should be done?

- "To reduce the risk of diabetes in at-risk patients, implement a lifestyle intervention program that includes weekly nutrition counseling and exercise sessions."



Data



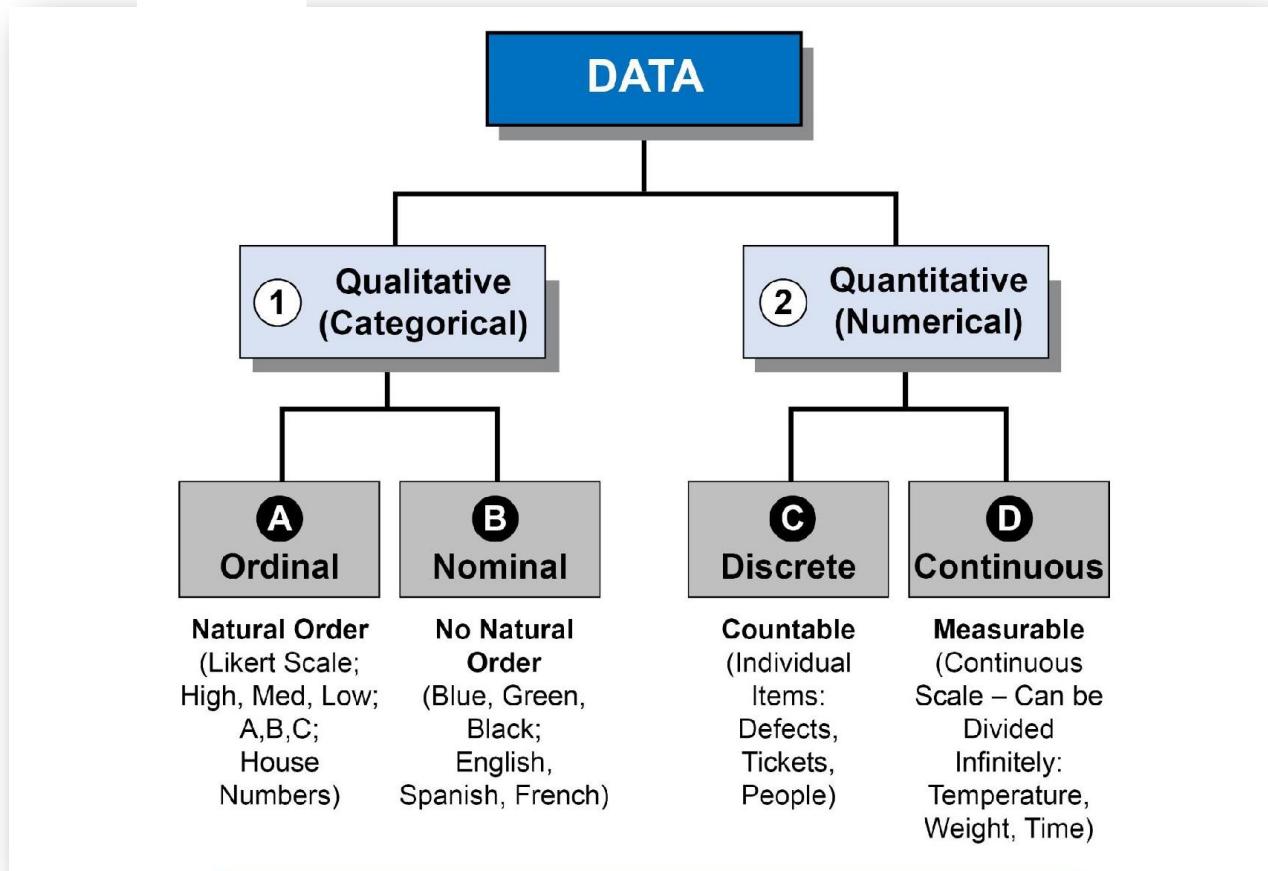


### Dataset schema

Column	Unit/Description
Patient ID	Unique identifier (alphanumeric)
Age	Years
Gender	Male/Female
Length of Stay	Days
Number of Diagnoses	Count
Primary Diagnosis	ICD Code
Comorbidities	Count of co-existing conditions
Treatment Type	Categorical (e.g., surgery, medication)
Readmission Within 30d	Yes/No

How can we predict the likelihood of hospital readmissions within 30 days based on patient demographics, medical history, and treatment details?

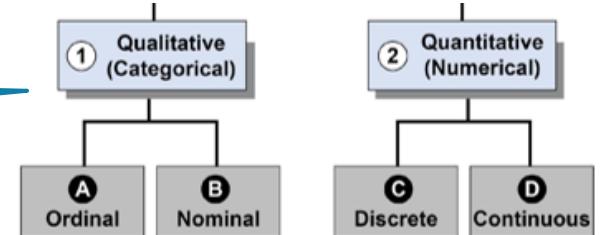
Can we characterize the data? We will not be able to perform the same analysis on the different kinds of data...



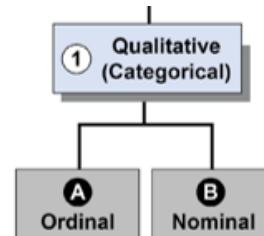
For each row, decide on type:

### Dataset schema

Column	Unit/Description
Patient ID	Unique identifier (alphanumeric)
Age	Years
Gender	Male/Female
Length of Stay	Days
Number of Diagnoses	Count
Primary Diagnosis	ICD Code
Comorbidities	Count of co-existing conditions
Treatment Type	Categorical (e.g., surgery, medication)
Readmission Within 30d	Yes/No



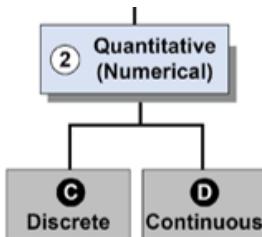
This is a good place to ask ChatGPT for additional examples!

**A**  
Ordinal

- **Definition:** Categorizes data into distinct groups where the order or ranking matters, but the intervals between categories are not meaningful.
- **Examples:**
  - Pain levels: Mild, Moderate, Severe
  - Education levels: High school, Bachelor's, Master's, Ph.D.
  - Star ratings: 1-star, 2-stars, 3-stars, 4-stars, 5-stars
  - Socioeconomic status: Low, Middle, High

**B**  
Nominal

- **Definition:** Categorizes data into distinct, non-ordered groups or labels. No inherent ranking exists.
- **Examples:**
  - Blood types: A, B, AB, O
  - Eye color: Blue, Green, Brown
  - Job roles: Teacher, Engineer, Doctor
  - Types of fruits: Apple, Banana, Orange



**C**  
Discrete

- **Definition:** Data that can take only specific, distinct values, often counted. Values are usually whole numbers.
- **Examples:**
  - **Health:** Number of patients in a clinic, number of medications a patient takes
  - **Environment:** Number of trees in a park, number of rainy days in a month
  - **Business:** Number of items sold, number of employees in a company

**D**  
Continuous

- **Definition:** Data that can take any value within a given range and is often measured. Values can include decimals and fractions.
- **Examples:**
  - **Health:** Body temperature (e.g., 98.6°F), weight (e.g., 68.4 kg), blood pressure (e.g., 120.5 mmHg)
  - **Environment:** Air quality index (e.g., 56.7), rainfall (e.g., 12.8 cm), temperature (e.g., 21.3°C)
  - **Business:** Revenue (e.g., \$10,254.75), product dimensions (e.g., 12.45 cm)

## Different Types of Data



Khushee Kapoor · Follow  
5 min read · Aug 28, 2021

There are many other characterizations

### Measurement

Quantitative (Numerical)  
Qualitative (Categorical)

### Format

Structured  
Unstructured  
Semi-Structured

### Volume

Big Data  
Small Data

### Collection

Primary Data  
Secondary Data

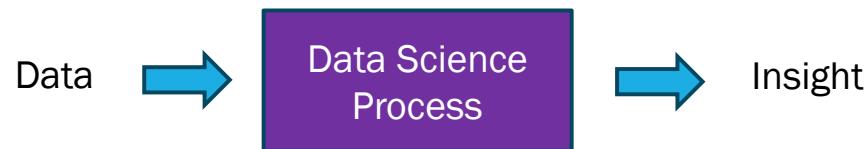
### Source

Internal Data  
External Data

### Time

Historical Data  
Real-Time Data

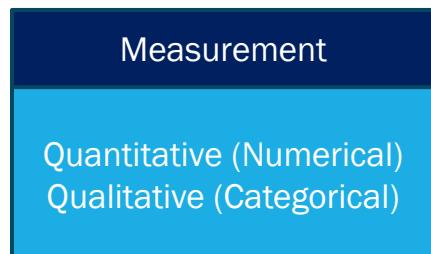
Measurement	Format	Volume
Quantitative (Numerical) Qualitative (Categorical)	Structured Unstructured Semi-Structured	Big Data Small Data
Collection	Source	Time
Primary Data Secondary Data	Internal Data External Data	Historical Data Real-Time Data



Insight: Predicting the Impact of Air Quality on Health Outcomes in Urban Areas

Predictive analysis

Let's explore these dimensions further + ask ChatGPT for examples in support to the provided insight

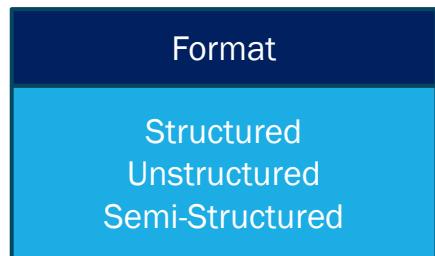


Numerical

Data Type	Example
Air Quality	PM <sub>2.5</sub> = 45 µg/m <sup>3</sup> , NO <sub>2</sub> = 25 ppb, O <sub>3</sub> = 50 µg/m <sup>3</sup>
Health Data	Asthma attacks = 15/month, Admissions = 100/week, Mortality rate = 120/1,000,000 people annually
Weather Data	Temperature = 25°C, Wind Speed = 15 km/h, Humidity = 75%

Categorical

Data Type	Example
Pollution Source	Traffic-related: "Highways", "Residential Areas", "Industrial Zones"
Health Conditions	Asthma, Cardiovascular Disease, COPD
Location	Neighborhood: "Downtown", "Suburban", "Industrial Park"

**Unstructured Data***Text, Images*

The university has 5600 students. Shaun (ID Number: 160801), 18 years old Communication study. Linh with ID number 160802, majoring in Accounting and is 20 years old. Ahmed from Psychology study program, 19 years old, ID number 160803.

GLEEMATIC A.I.

**Semi-Structured Data***JSON, APIs*

```
<University>
  <ID Number="160801">
    <Name="Shaun">
    <Age="18">
    <Program="Communication">
  <ID Number="160802">
    <Name="Linh">
    <Age="20">
    <Program="Accounting">
.....</University>
```

**Structured Data***Tabular*

ID	Name	Age	Program
160801	Shaun	18	Communication
160802	Linh	20	Accounting
160803	Ahmed	19	Psychology

GLEEMATIC A.I.

Source

Format
Structured
Unstructured
Semi-Structured

Date	PM2.5 ( $\mu\text{g}/\text{m}^3$ )	NO2 (ppb)	O <sub>3</sub> ( $\mu\text{g}/\text{m}^3$ )	Location	Asthma Admissions (per week)	Mortality Rate (per 1,000,000)
2025-01-01	40	30	55	Downtown	100	120
2025-01-02	25	22	40	Suburban	80	110
2025-01-03	50	35	60	Industrial	150	140
2025-01-04	45	30	55	Downtown	130	130

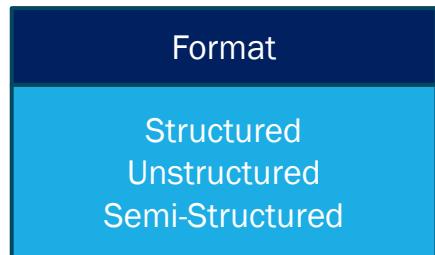
Format	
Structured	
Unstructured	
Semi-Structured	Unstructured

**Social Media/News Posts (Text Data):**

- "The air quality in Downtown today is terrible! PM2.5 levels above 50  $\mu\text{g}/\text{m}^3$ . Breathing difficulties reported."
- "Suburban areas report lower PM2.5 levels today, but still above safe levels."
- "Traffic pollution in industrial zones contributes to worsening respiratory health."
- "More cases of asthma reported in areas with high Ozone levels above 60  $\mu\text{g}/\text{m}^3$ ."

**Medical Notes (Text Data):**

- "Patient presents with shortness of breath, worsening asthma symptoms. High PM2.5 levels in the area."
- "COPD patient diagnosed with severe air pollution exposure, advised to stay indoors."
- "Asthma exacerbation due to high NO<sub>2</sub> levels and traffic-related pollution exposure."



XML (Weather Data):

```
xml

<weather>
  <entry>
    <timestamp>2025-01-01T12:00:00</timestamp>
    <temperature unit="Celsius">25</temperature>
    <humidity unit="percent">75</humidity>
    <windspeed unit="km/h">15</windspeed>
  </entry>
  <entry>
    <timestamp>2025-01-02T12:00:00</timestamp>
    <temperature unit="Celsius">22</temperature>
    <humidity unit="percent">70</humidity>
    <windspeed unit="km/h">20</windspeed>
  </entry>
</weather>
```

JSON (Air Quality Sensor Data):

Semi-structured

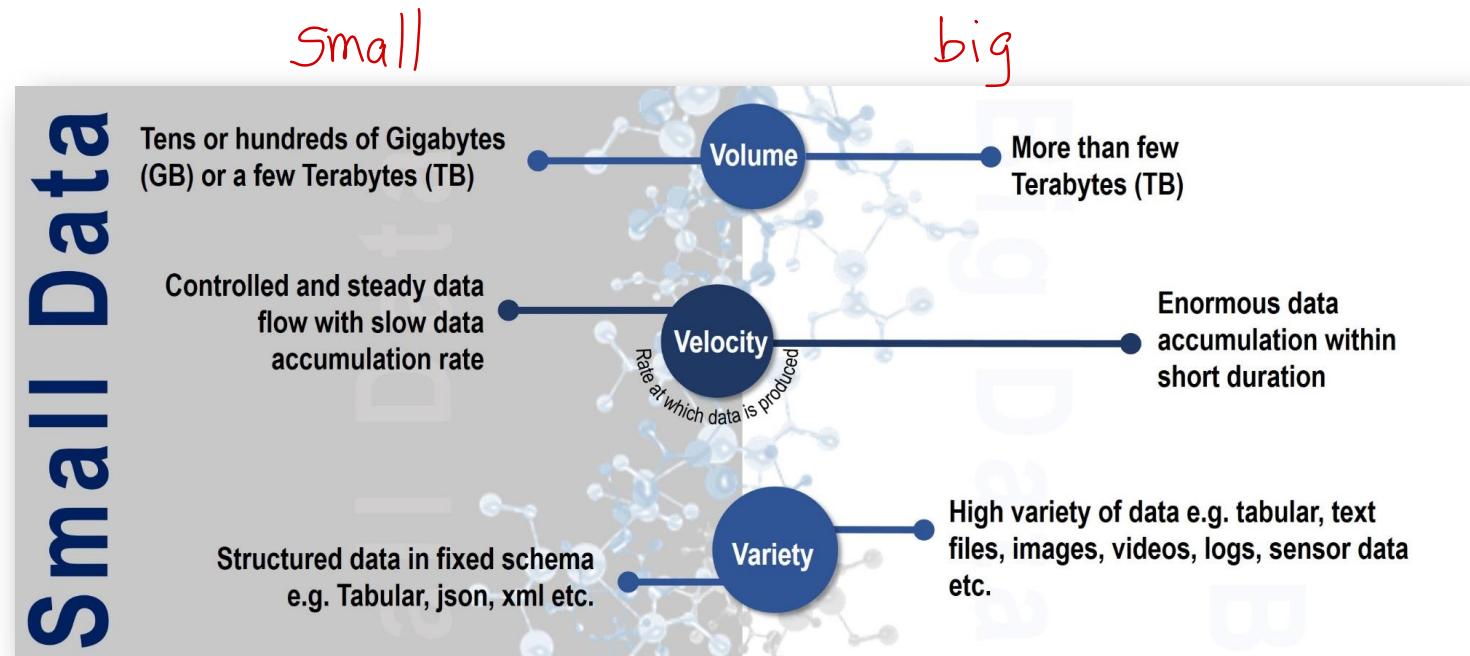
```
json

[
  {
    "timestamp": "2025-01-01T12:00:00",
    "PM2.5": 45,
    "NO2": 30,
    "O33

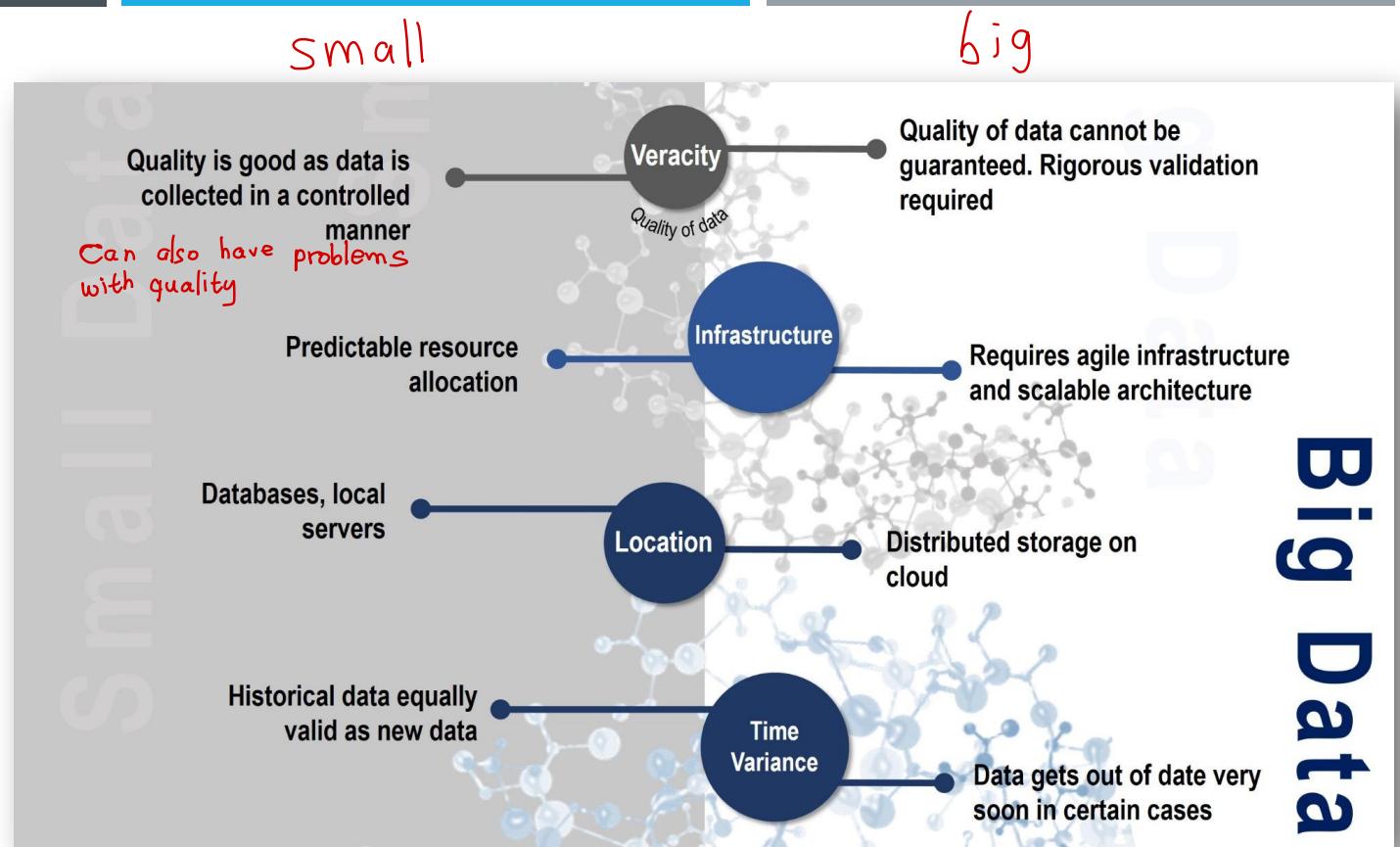
```



# Small Data



[Source](#)



# Big Data

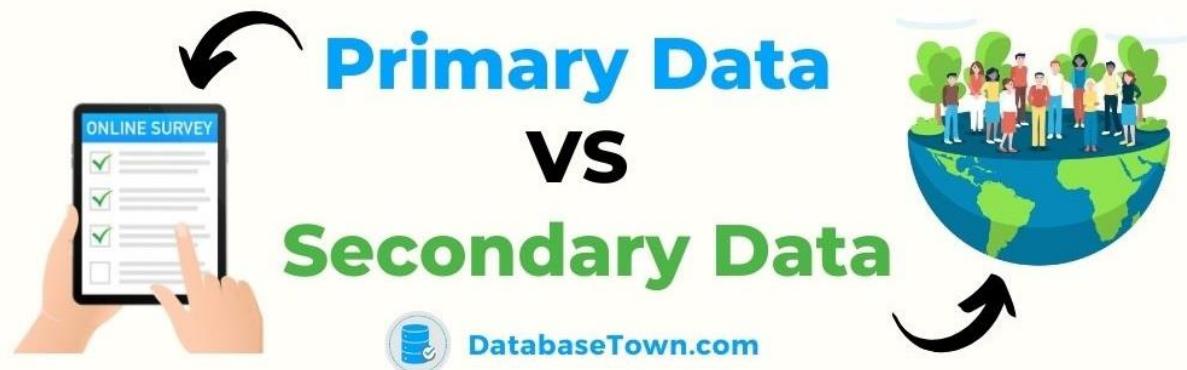
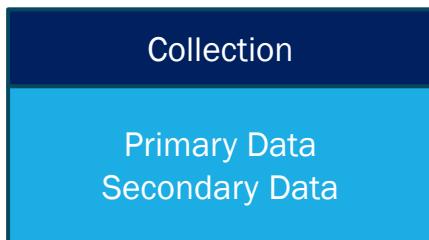


[Source](#)



Data Type	Example
Real-time Air Quality	Continuous data streams from 500 sensors across a city, reporting every minute (e.g., 500 sensors × 1440 minutes per day × 365 days/year)
Health Data	Health records from 1 million patients with diagnoses, prescriptions, and hospital visits across 10 years

Data Type	Example
Local Health Study	Data from 200 patients detailing asthma symptoms, medications, and air quality exposure in a specific neighborhood
Pollution Monitoring	Data from a single sensor over the course of 6 months: { Date: 2025-01-01, PM2.5: 40 µg/m³, NO2: 25 ppb }



	Primary Data	Secondary Data
Source	Collected firsthand by researcher	Collected by others
Purpose	Specific to current research	May have different original purpose
Time	Takes longer to collect	Quickly available

collected within  
company

collected elsewhere



### These are the characteristics of primary data:

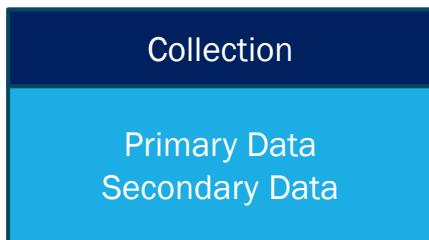
- Primary data is collected directly by the researcher,
- It is specific to the research question,
- Up-to-date and current,
- It can be time-consuming and expensive to collect, and
- It gives the researcher full control over the data collection process.

### Methods of Collecting Primary Data:

1. Surveys and questionnaires
2. Interviews (face-to-face, phone, or online)
3. Focus groups
4. Observations
5. Experiments
6. Field research



[Source](#)



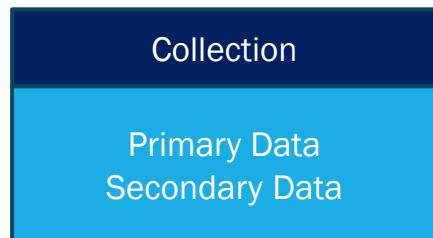
### **Key Characteristics of Secondary Data:**

- Collected by others for different purposes
- Already existing and readily available
- Often less expensive and faster to obtain
- May not perfectly fit the current research needs
- Researcher has no control over data collection methods

### **Sources of Secondary Data:**

1. Government publications
2. Academic journals and books
3. Census data
4. Company records
5. Industry reports
6. Previous research studies
7. Online databases





Primary data comes directly from observations or experiments:

Date	PM2.5 ( $\mu\text{g}/\text{m}^3$ )	NO2 (ppb)	O <sub>3</sub> ( $\mu\text{g}/\text{m}^3$ )	Location	Symptoms
2025-01-01	40	30	55	Downtown	Wheezing
2025-01-02	35	25	50	Suburban	Shortness of Breath
2025-01-03	60	40	70	Industrial	Chest Pain

### Secondary

Secondary data comes from external sources or previous research:

Study Title	Source	PM2.5 Threshold ( $\mu\text{g}/\text{m}^3$ )	Health Outcome
"Air Pollution and Respiratory Health"	WHO	35	Increased asthma attacks
"Traffic Pollution and Cardiovascular Disease"	CDC	40	Heart disease correlation



## Internal data



### User's data

logs, messages, mails



### Internal documents

invoices, contracts, notes



### IoT Devices

cameras, sensors



### Logs

website / platform logs

## External data



### Web

e-commerce, real estate



### Geo

maps, localization, GPS



### Files

invoices, documents, sheets



### 3rd parties

weather, credit card, telco

## Internal and External Data: What's the Difference and Why It Matters



Sebastian Berg

CEO, Co-founder

Product March 15, 2023

Source

Source

Internal Data  
External Data

# The Importance of Internal Data

Internal data is essential because it provides companies with insights into their operations. This data can help companies understand their performance and make informed decisions based on the analysis of their data. It also allows companies to identify areas for improvement and make necessary changes to their operations.

Internal data is also valuable because it is unique to the company. Since internal data is generated within the company, it is specific to the company's operations and can provide a competitive advantage. For example, a company can use its internal sales data to analyze trends and forecast future sales.

## Internal and External Data: What's the Difference and Why It Matters



Sebastian Berg

CEO, Co-founder

Product March 15, 2023

[Source](#)

Source

Internal Data  
External Data

# The Value of External Data

External data is important because it provides companies with a broader perspective on the market and industry. It can help companies understand industry trends, benchmark their performance against competitors, and identify new opportunities.

By using external data, companies can gain insights into the behavior and preferences of their target audience. For example, a company can use social media data to analyze how consumers are talking about its brand and products.

External data can also help companies mitigate risks. By analyzing external data such as economic indicators, companies can identify potential risks to their operations, such as changes in the market or regulatory environment.

## Internal and External Data: What's the Difference and Why It Matters



Sebastian Berg

CEO, Co-founder

Product March 15, 2023

[Source](#)

Source		Internal																											
	Internal Data	<p>Internal data refers to data generated within an organization:</p> <table border="1"> <thead> <tr> <th>Hospital ID</th> <th>Patient ID</th> <th>Disease</th> <th>Admission Type</th> <th>Air Quality Exposure</th> <th>Symptoms</th> </tr> </thead> <tbody> <tr> <td>HOSP001</td> <td>123</td> <td>Asthma</td> <td>Emergency</td> <td>High</td> <td>Wheezing</td> </tr> <tr> <td>HOSP002</td> <td>456</td> <td>COPD</td> <td>Scheduled</td> <td>Low</td> <td>Shortness of Breath</td> </tr> <tr> <td>HOSP003</td> <td>789</td> <td>Asthma</td> <td>Emergency</td> <td>High</td> <td>Coughing</td> </tr> </tbody> </table>				Hospital ID	Patient ID	Disease	Admission Type	Air Quality Exposure	Symptoms	HOSP001	123	Asthma	Emergency	High	Wheezing	HOSP002	456	COPD	Scheduled	Low	Shortness of Breath	HOSP003	789	Asthma	Emergency	High	Coughing
Hospital ID	Patient ID	Disease	Admission Type	Air Quality Exposure	Symptoms																								
HOSP001	123	Asthma	Emergency	High	Wheezing																								
HOSP002	456	COPD	Scheduled	Low	Shortness of Breath																								
HOSP003	789	Asthma	Emergency	High	Coughing																								
	External Data	<p>External data comes from outside sources:</p> <table border="1"> <thead> <tr> <th>Data Source</th> <th>PM2.5 (<math>\mu\text{g}/\text{m}^3</math>)</th> <th>NO2 (ppb)</th> <th><math>\text{O}_3</math> (<math>\mu\text{g}/\text{m}^3</math>)</th> <th>Location</th> <th>Date</th> </tr> </thead> <tbody> <tr> <td>EPA</td> <td>35</td> <td>30</td> <td>60</td> <td>Downtown</td> <td>2025-01-01</td> </tr> <tr> <td>WHO</td> <td>40</td> <td>32</td> <td>58</td> <td>Suburban</td> <td>2025-01-01</td> </tr> </tbody> </table>				Data Source	PM2.5 ( $\mu\text{g}/\text{m}^3$ )	NO2 (ppb)	$\text{O}_3$ ( $\mu\text{g}/\text{m}^3$ )	Location	Date	EPA	35	30	60	Downtown	2025-01-01	WHO	40	32	58	Suburban	2025-01-01						
Data Source	PM2.5 ( $\mu\text{g}/\text{m}^3$ )	NO2 (ppb)	$\text{O}_3$ ( $\mu\text{g}/\text{m}^3$ )	Location	Date																								
EPA	35	30	60	Downtown	2025-01-01																								
WHO	40	32	58	Suburban	2025-01-01																								

Time

Historical Data  
Real-Time Data

## The Essential Fusion: Real-Time and Historical Data

Historical data provides a rich backdrop of information, highlighting long-term trends, patterns and outcomes. Real-time, domain-specific data feeds ensure that ML algorithms are working with the most current information. Integrating real-time data into ML models helps make the predictive insights contextualized and hyper-personalized to the end user in the moment, providing valuable information to inform strategic decisions.

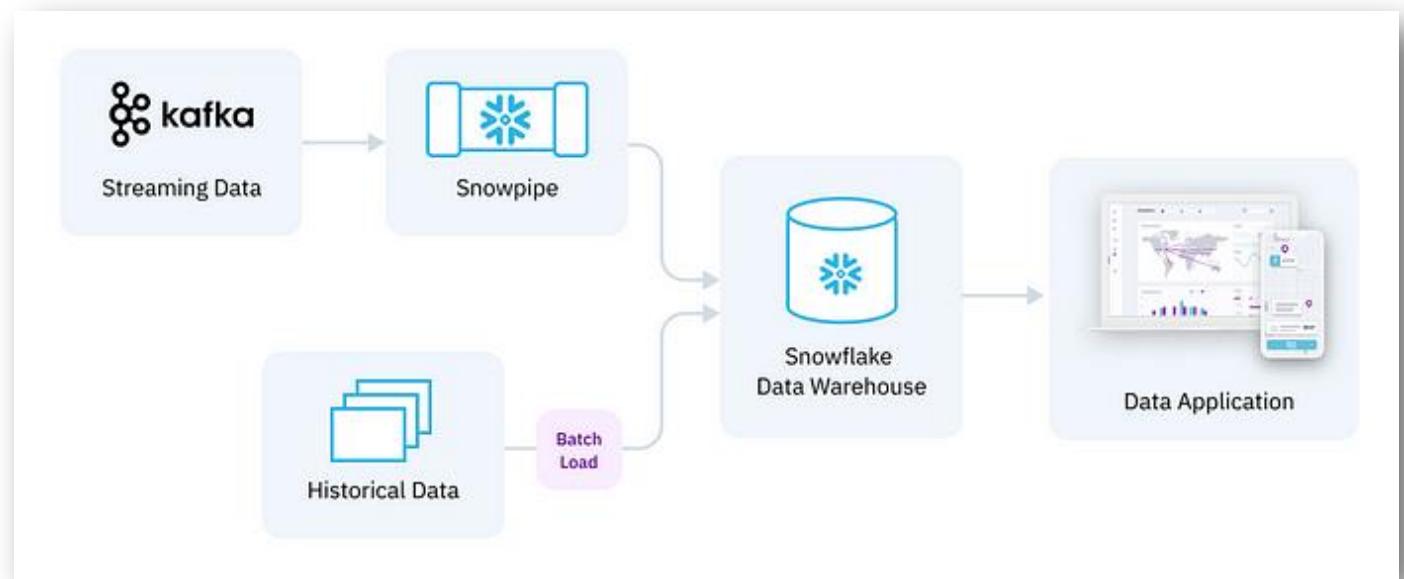
AI / DATA

### Integrating Real-Time and Historical Data Enhances Decision-Making

The accuracy and relevance of machine learning-driven predictions depend significantly on the quality and timeliness of the data fed into the models.

Apr 18th, 2024 6:34am by [Rahul Pradhan](#)

[Source](#)



**Joining Streaming and Historical Data for Real-Time Analytics: Your Options With Snowflake, Snowpipe and Rockset**

[Source](#)

Time

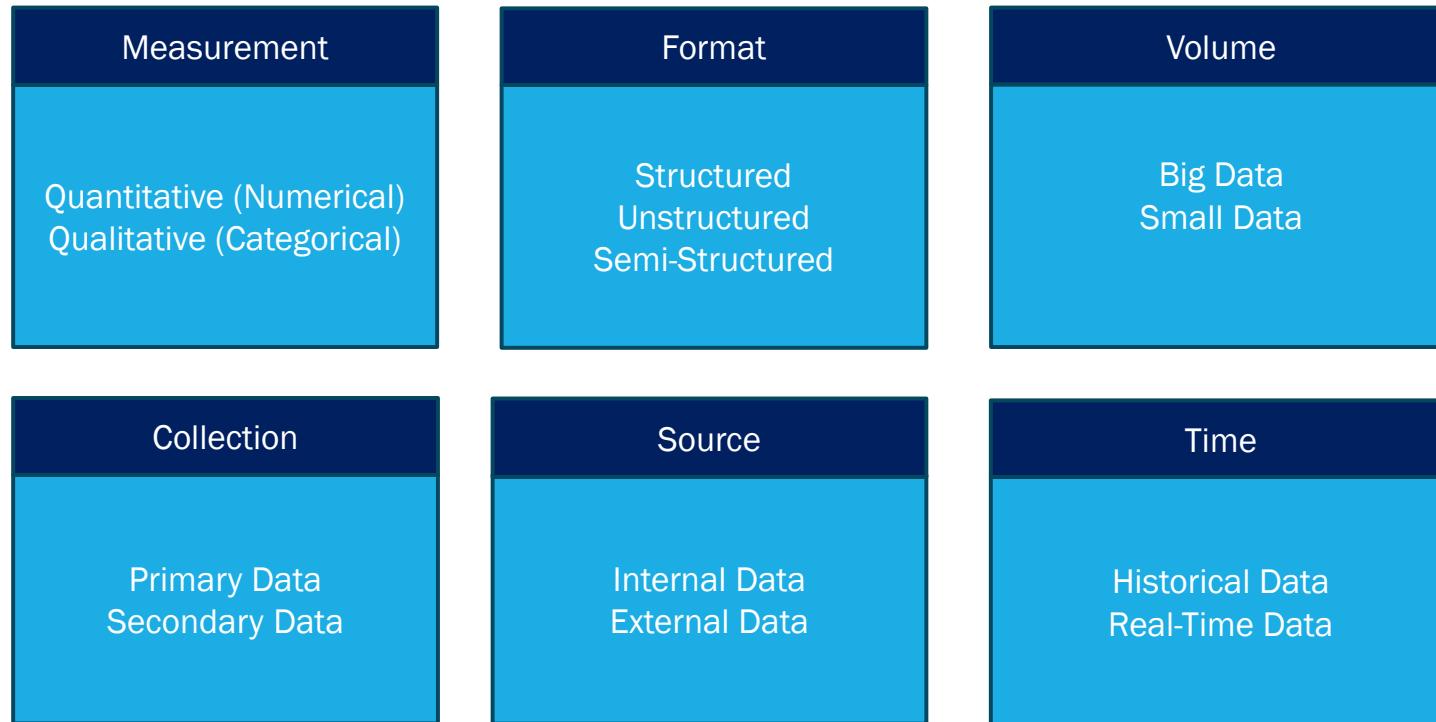
Historical Data  
Real-Time Data

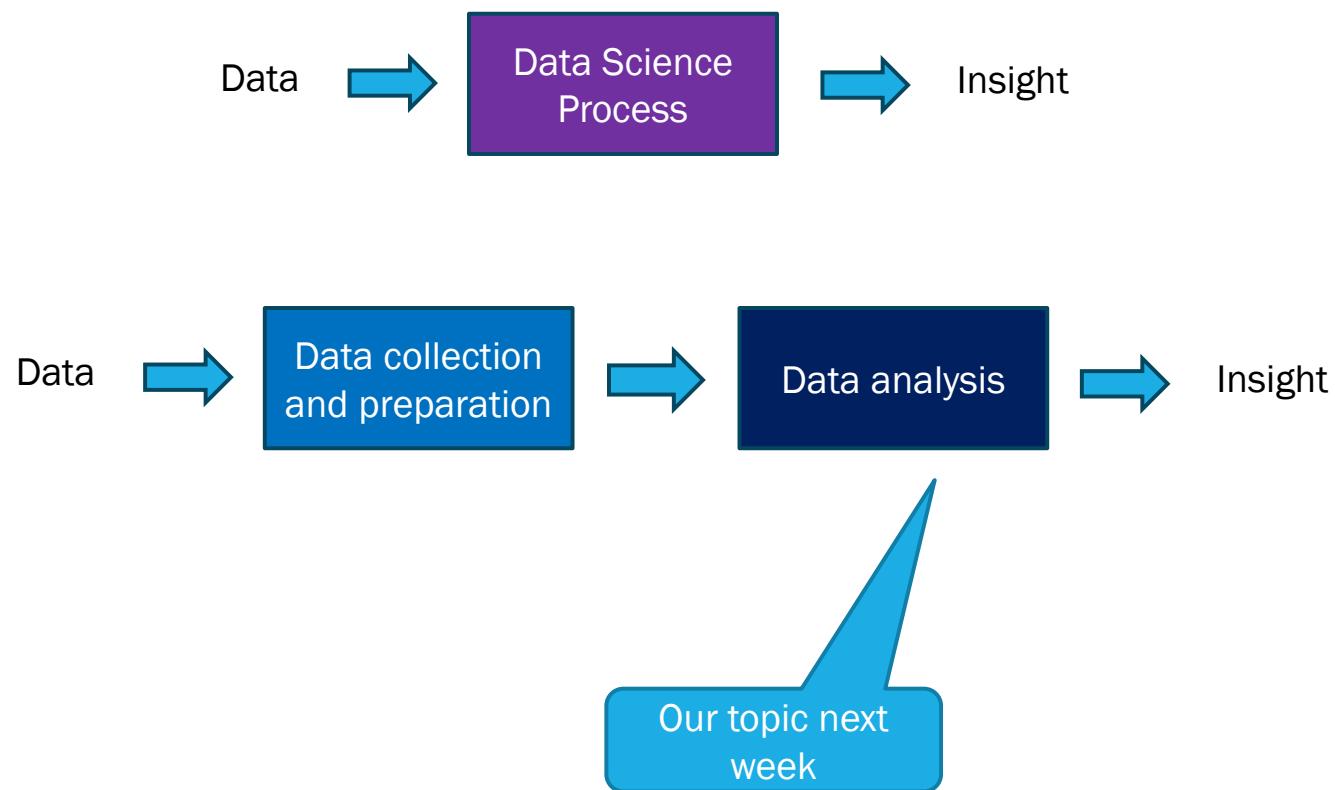
Historical

Data Type	Example
Long-term Air Quality	"PM2.5 levels for the past 10 years in Downtown: 2015 = 40 µg/m³, 2016 = 45 µg/m³, ..., 2025 = 30 µg/m³"
Health Statistics	"Asthma incidence in the East Coast from 2010-2020: 2010 = 0.05%, 2015 = 0.07%, 2020 = 0.10%"

Real-time

Data Type	Example
Pollution Sensors	Real-time data streamed from multiple sensors across a city: { "timestamp": "2025-01-01T12:00:00", "PM2.5": 42 µg/m³, "NO2": 30 ppb }
Wearable Health Data	Data from a fitness tracker: "Patient ID: 123, Heart Rate: 75 bpm, Symptoms: Shortness of breath, PM2.5 exposure: High"





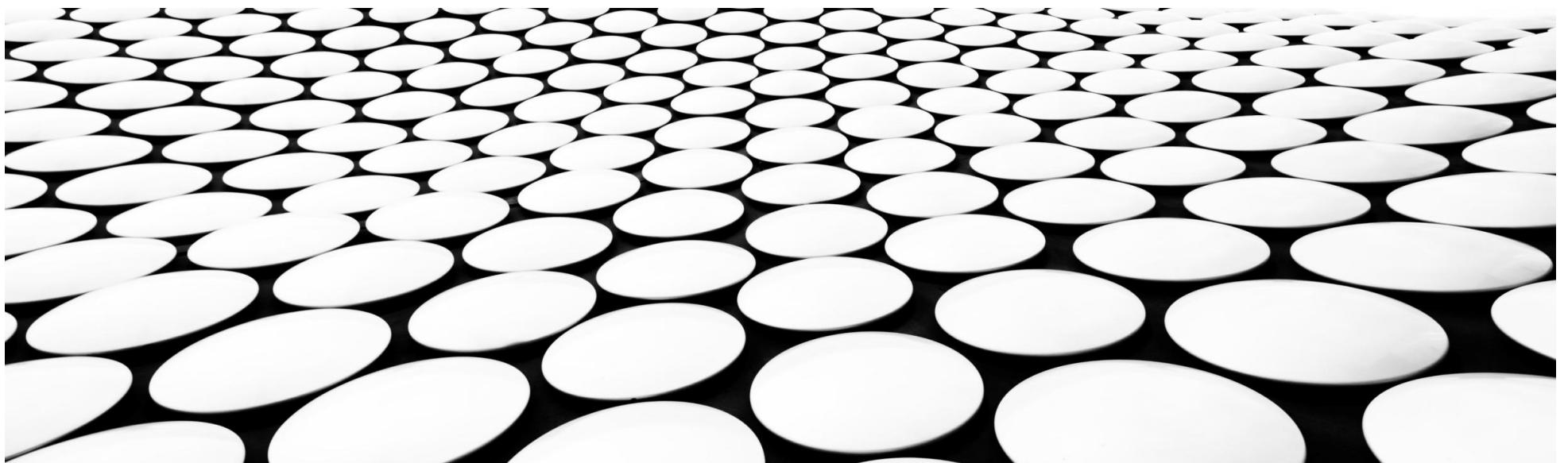


## DERIVING INSIGHTS FROM DATA

- Types of insights
- Types of data

# EXPLORATORY DATA ANALYSIS

*Story behind the data: Titanic dataset*

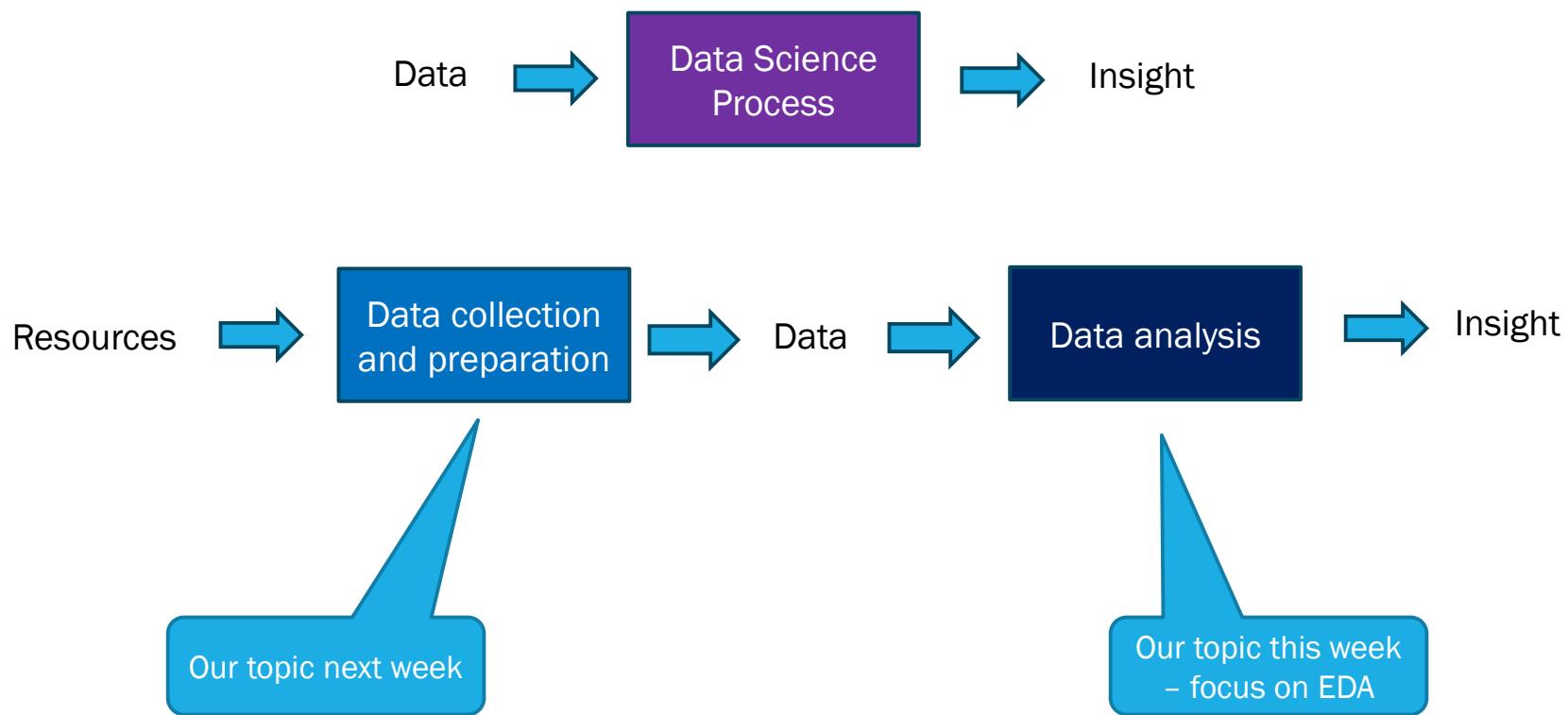




## EXPLORATORY DATA ANALYSIS

- Descriptive statistics
- Diagnostic statistics
- Case Study 1 – Titanic

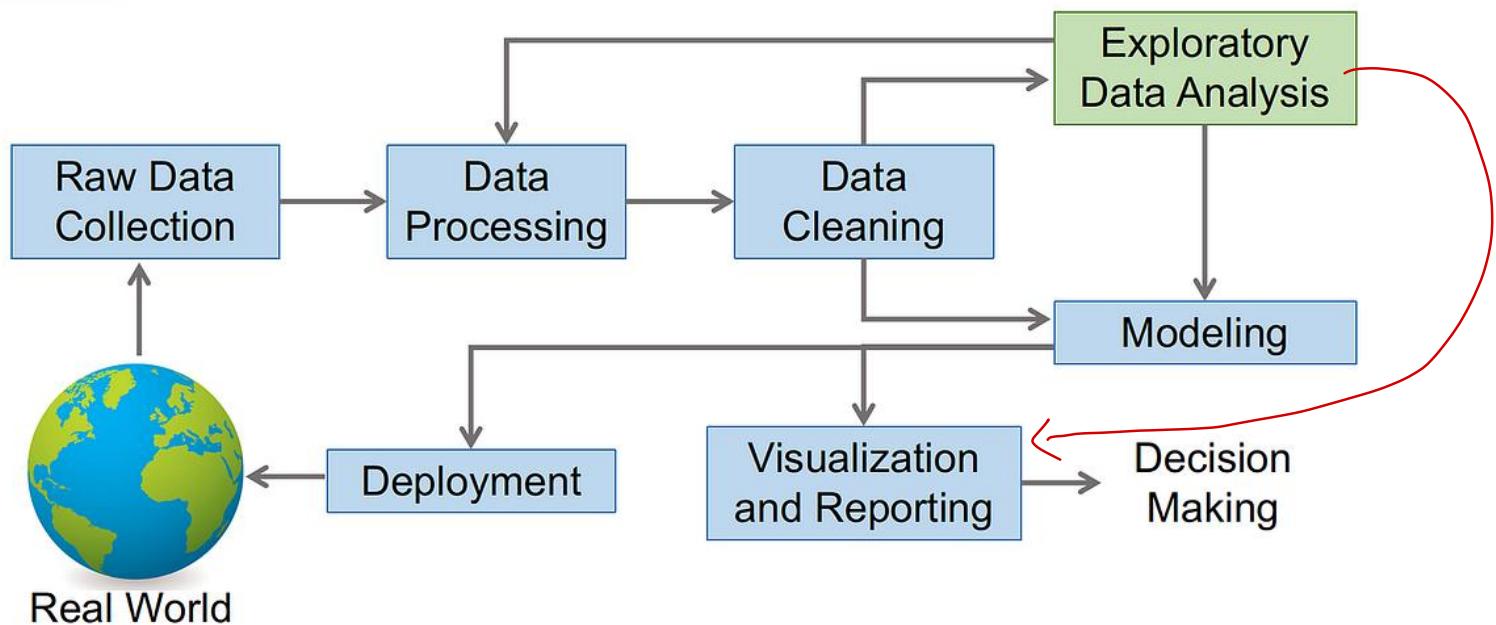
Friday Case Study 2 +  
Examples for Assignment 1



## Data Science Process: A Comprehensive Guide

Abhiljit · Follow  
6 min read · Jan 15, 2024

# Data Science Process



## Introduction to Exploratory Data Analysis

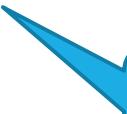


Kaushik Mani · Follow

Published in DataDrivenInvestor · 9 min read · Jan 29, 2019

... one of the most important components to any data science experiment that doesn't get as much importance as it should is Exploratory Data Analysis (EDA).

In short, EDA is “**A first look at the data**”. It is a critical step in analyzing the data from an experiment. It is used to **understand and summarize the content of the dataset** to ensure that the features which we feed to our machine learning algorithms are refined and we get valid, correctly interpreted results.

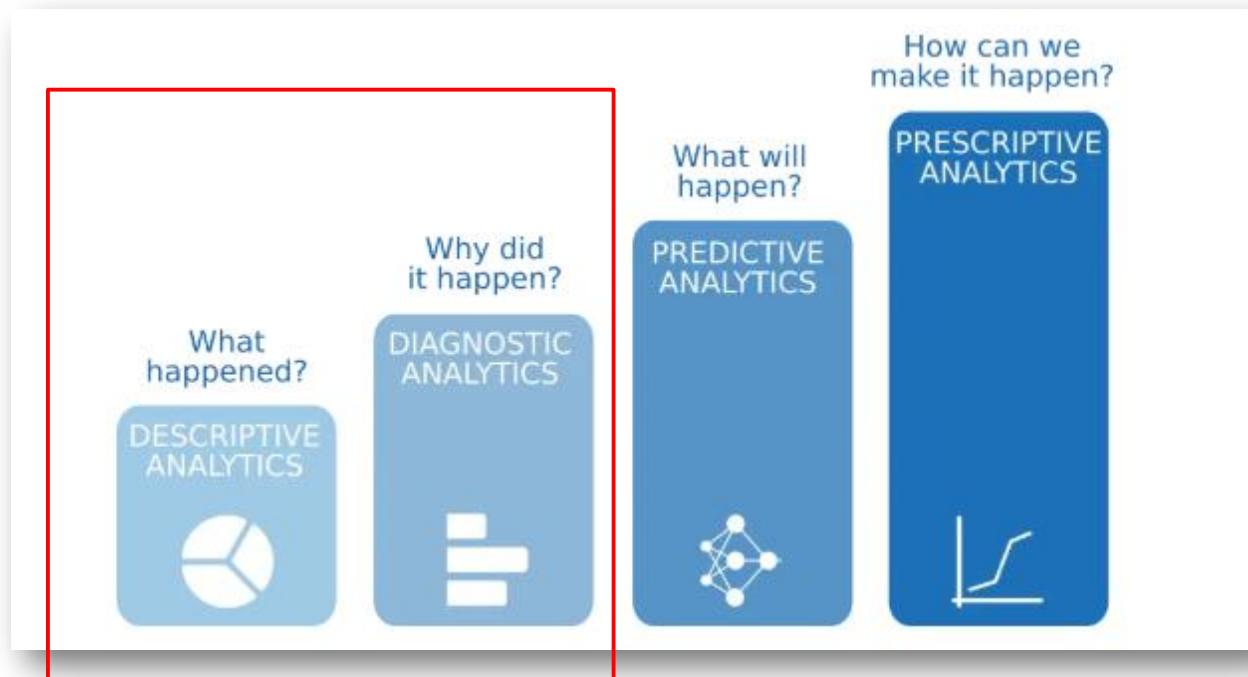


Or that we present valid descriptive insights in reports

# What is advanced analytics?

Tutorials

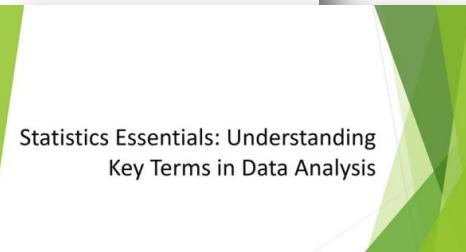
© Pablo Martín    August 31, 2023



# Descriptive Statistics

## Statistics Handbook for Data Analysts

Anita Gupta · Follow  
17 min read · Sep 14, 2024



### Measures of central tendency

1. Mean
2. Median
3. Mode

### Measures of dispersion

1. Range
2. Variance
3. Standard deviation

**Mean:** The average value.

$$\text{Mean}(\mu) = \frac{1}{n} \sum_{i=1}^n X_i$$

### Example

The average salary across United States is \$60,000.

**Median:** The middle value in a sorted dataset. If the dataset contains an even number of values, the median is the average of the two middle values.

### Example

The median divides the data points: 50% are above it, and 50% are below it. So next time someone says the median salary is \$50,000, it means 50% of people earn below that and 50% earn above it.

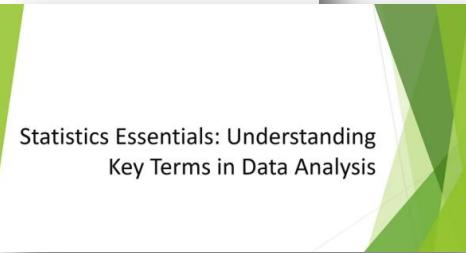
**Mode:** The most frequently occurring value.

Example

If you do a poll and ask what food students want to order for lunch, and the majority vote for pizza, then pizza is the mode — the highest occurring value.

## Statistics Handbook for Data Analysts

Anita Gupta · Follow  
17 min read · Sep 14, 2024



### Measures of central tendency

1. Mean
2. Median
3. Mode

### Measures of dispersion

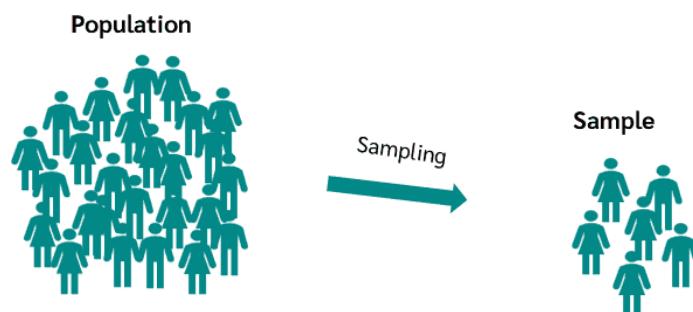
1. Range
2. Variance
3. Standard deviation

**Range** is from the minimum to the maximum value.

### Example

If you're looking to buy a house, you might be considering houses in the price range of \$500,000 to \$800,000. We use ranges all the time in day-to-day conversations.

Variance is a measure of dispersion, meaning it is a measure of how far a set of numbers is spread out from their average value.



$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

$S^2$  = Variance

n = The Number of data Point

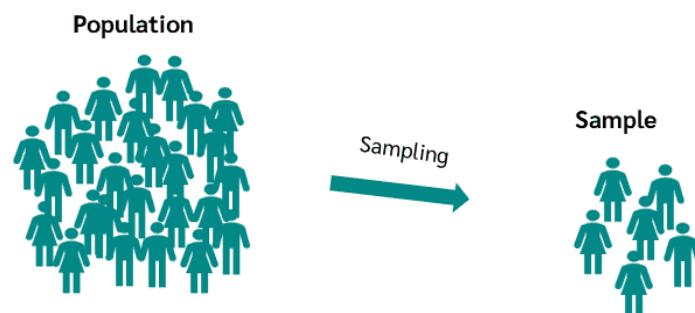
$X_i$  = Each of the values of the data

$\bar{X}$  = The Mean of  $X_i$

Equation for a  
sample

**Standard deviation** is the square root of the variance.

This allows the dispersion metric to be in the same unit as the observed data. For example, we now have dollars instead of dollars squared, making interpretation more useful.



$$S_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

$n$  = Number of Observations  
 $x_i$  = Value of the one Observation  
 $\bar{x}$  = Mean Value of Observations

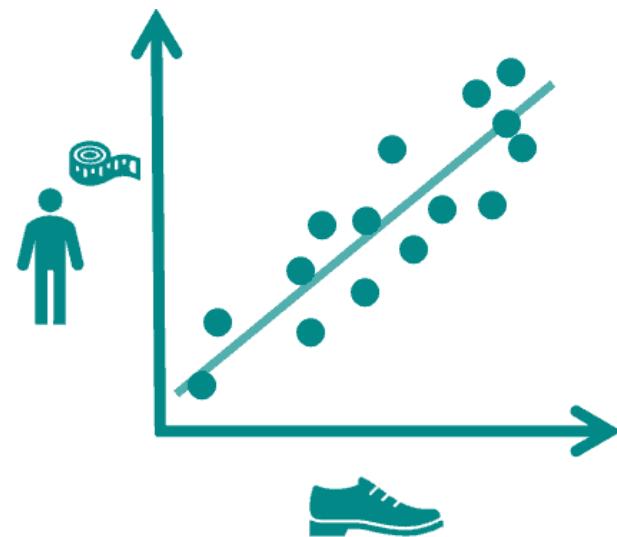
Equation for a sample

# Diagnostic statistics



# Correlation analysis

Correlation analysis is a statistical method used to evaluate the relationship between two variables, such as the association between body size and shoe size.





## Pearson correlation analysis

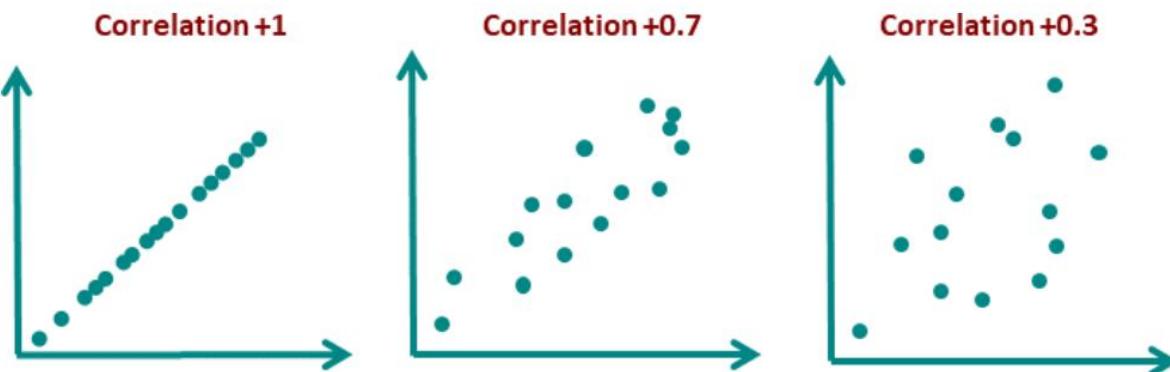
With the [Pearson correlation analysis](#) you get a statement about the linear correlation between metric scaled variables. The respective **covariance** is used for the calculation.

With the help of correlation analysis two statements can be made:

- one about the direction
- and one about the strength

$$Cov(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}}$$

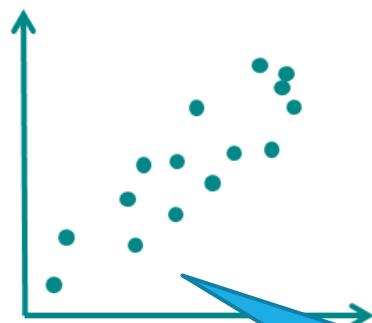


$ r $	Strength of correlation
$0.0 < 0.1$	no correlation
$0.1 < 0.3$	little correlation
$0.3 < 0.5$	medium correlation
$0.5 < 0.7$	high correlation
$0.7 < 1$	very high correlation

## Requirements for causality

1

There is a significant correlation



2

Timely

Variable A surveyed



Variable B surveyed

Theory



Profound theory

If we don't even have correlation...  
for sure we can affirm there is NO  
causality

The screenshot shows a YouTube playlist titled "Correlation Analysis" from the channel "DATAtab". The playlist has 5 videos and 6,937 views. The first video is "Pearson Correlation [Simply explained]" (7:50), followed by "Spearman Rank Correlation [Simply explained]" (5:41), "Point-biserial Correlation [Simply explained]" (6:30), "Kendall's Tau [Easily explained]" (6:32), and finally "Causality, Correlation and Regression" (7:35). Each video thumbnail includes a small icon related to the correlation type.

Choice varies based on conditions (parametric: metric data normally distributed)

# *Case Study 1*

## Titanic Dataset

```
# Import libraries for visualization
import seaborn as sns
import matplotlib.pyplot as plt

# Loading the dataset
titanic = sns.load_dataset('titanic')

# Number of samples, Number of attributes
titanic.shape

(891, 15)
```

Dataset available  
within seaborn

Lots of other titanic  
datasets!!



2,288 Datasets

	<b>Titanic</b> Khashayar Baghizadeh · Updated 8 years ago Usability 7.1 · 1 File (CSV) · 11 kB
	<b>Titanic dataset</b> Brenda N · Updated 3 years ago Usability 10.0 · 1 File (CSV) · 12 kB
	<b>Titanic</b> Azeem Bootwala · Updated 8 years ago Usability 8.2 · 2 Files (CSV) · 12 kB
	<b>Titanic Dataset</b> M Yasser H · Updated 3 years ago Usability 10.0 · 1 File (CSV) · 23 kB
	<b>Titanic</b> Rahul · Updated 5 years ago Usability 6.8 · 3 Files (CSV) · 35 kB
	<b>Titanic Dataset</b> Shubham_Gupta012 · Updated 2 years ago Usability 10.0 · 1 File (CSV) · 7 kB

```
titanic.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 15 columns):
 #   Column      Non-Null Count  Dtype  
 --- 
 0   survived    891 non-null    int64  
 1   pclass       891 non-null    int64  
 2   sex          891 non-null    object  
 3   age          714 non-null    float64 
 4   sibsp        891 non-null    int64  
 5   parch        891 non-null    int64  
 6   fare          891 non-null    float64 
 7   embarked     889 non-null    object  
 8   class         891 non-null    category
 9   who           891 non-null    object  
 10  adult_male   891 non-null    bool   
 11  deck          203 non-null    category
 12  embark_town  889 non-null    object  
 13  alive         891 non-null    object  
 14  alone         891 non-null    bool
```

A first view of attributes and  
data types

#### Siblings (sibsp)

- Brother, sister
- Step-brother, step-sister
- Husband, wife

#### Parents/Children (parch)

- Mother, father
- Son, daughter
- Step-children

▶ titanic.head(10)

...	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True
5	0	3	male	NaN	0	0	8.4583	Q	Third	man	True	NaN	Queenstown	no	True
6	0	1	male	54.0	0	0	51.8625	S	First	man	True	E	Southampton	no	True
7	0	3	male	2.0	3	1	21.0750	S	Third	child	False	NaN	Southampton	no	False
8	1	3	female	27.0	0	2	11.1333	S	Third	woman	False	NaN	Southampton	yes	False
9	1	2	female	14.0	1	0	30.0708	C	Second	child	False	NaN	Cherbourg	yes	False

Measurement

Quantitative (Numerical)  
Qualitative (Categorical)

Format

Structured  
Unstructured  
Semi-Structured

Volume

Big Data  
Small Data

Collection

Primary Data  
Secondary Data

Source

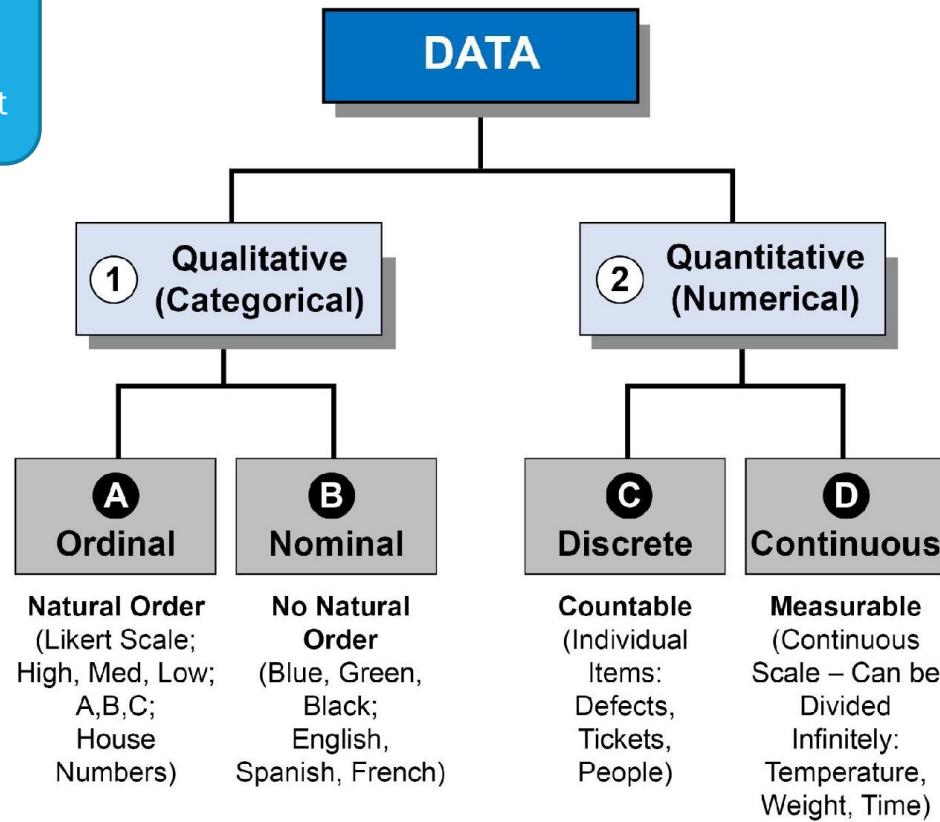
Internal Data  
External Data

Time

Historical Data  
Real-Time Data

How would we characterize our current dataset?

We saw last week that this dimension for data characterization was important



▶ titanic.head(10)

...	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True
5	0	3	male	NaN	0	0	8.4583	Q	Third	man	True	NaN	Queenstown	no	True
6	0	1	male	54.0	0	0	51.8625	S	First	man	True	E	Southampton	no	True
7	0	3	male	2.0	3	1	21.0750	S	Third	child	False	NaN	Southampton	no	False
8	1	3	female	27.0	0	2	11.1333	S	Third	woman	False	NaN	Southampton	yes	False
9	1	2	female	14.0	1	0	30.0708	C	Second	child	False	NaN	Cherbourg	yes	False

## Feature analysis

```
titanic.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 15 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   survived    891 non-null    int64  
 1   pclass       891 non-null    int64  
 2   sex          891 non-null    object  
 3   age          714 non-null    float64 
 4   sibsp        891 non-null    int64  
 5   parch        891 non-null    int64  
 6   fare          891 non-null    float64 
 7   embarked     889 non-null    object  
 8   class         891 non-null    category
 9   who           891 non-null    object  
 10  adult_male   891 non-null    bool   
 11  deck          203 non-null    category
 12  embark_town  889 non-null    object  
 13  alive         891 non-null    object  
 14  alone         891 non-null    bool
```

A view at missing data also...  
we'll do data cleaning later in  
the semester

A view at perhaps redundant  
attributes?

Importance of self-  
explanatory labels when  
documentation is lost...

## Descriptive statistics

Central tendency

Dispersion

`titanic.describe()`

	survived	pclass	age	sibsp	parch	fare
count	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

```
▶ titanic.describe(include=["object", "category"])

...
   sex  embarked  class  who  deck  embark_town  alive
count    891        889    891    891     203           889    891
unique     2          3      3      3      7             3      2
top  male         S  Third  man       C  Southampton  no
freq    577        644    491    537      59           644    549
```

Mode is on  
categorical data

Dispersion on  
categorical data

```
▶ titanic.value_counts('embark_town')
```

```
...          count
```

```
embark_town
```

Southampton	644
Cherbourg	168
Queenstown	77

```
▶ titanic.value_counts('class')
```

```
...          count
```

```
class
```

Third	491
First	216
Second	184

```
▶ titanic.value_counts('who')
```

```
...          count
```

```
who
```

man	537
woman	271
child	83

▶ titanic.value\_counts('class')

	count
class	
Third	491
First	216
Second	184

▶ titanic.value\_counts('pclass')

	count
pclass	
3	491
1	216
2	184

Redundancy

Sometimes a numerical  
feature is really a  
categorical feature

	age
count	714.000000
mean	29.699118
std	14.526497
min	0.420000
25%	20.125000
50%	28.000000
75%	38.000000
max	80.000000

	count
age	
24.0	30
22.0	27
18.0	26
19.0	25
30.0	25
...	...
53.0	1
66.0	1
70.5	1
74.0	1
80.0	1

In general value\_counts  
on a numerical data  
does not make sense

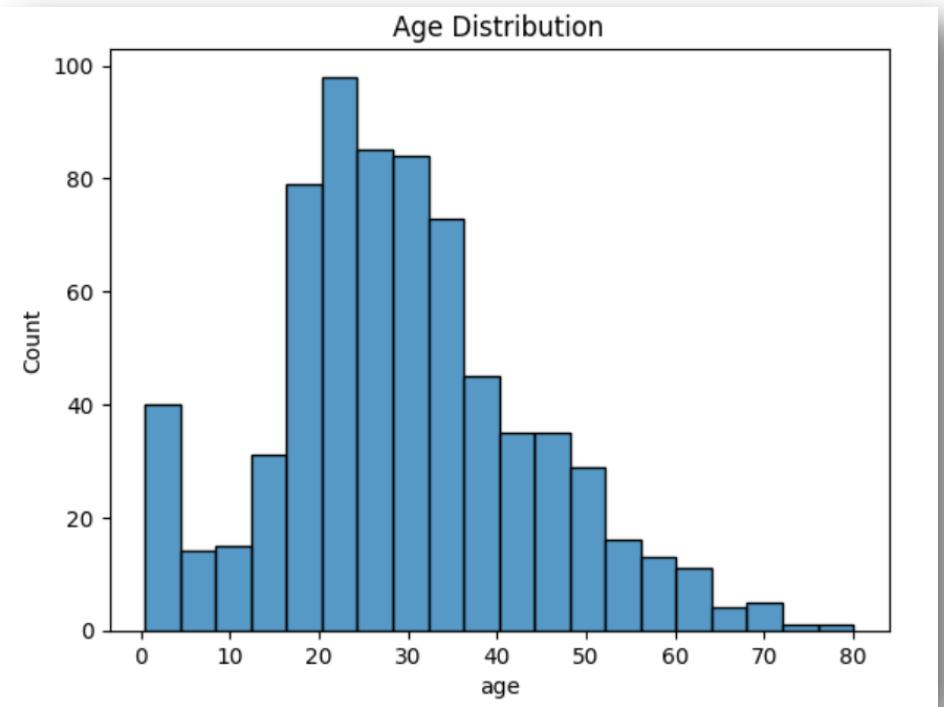
Better with graphs for  
visualisation of  
dispersion

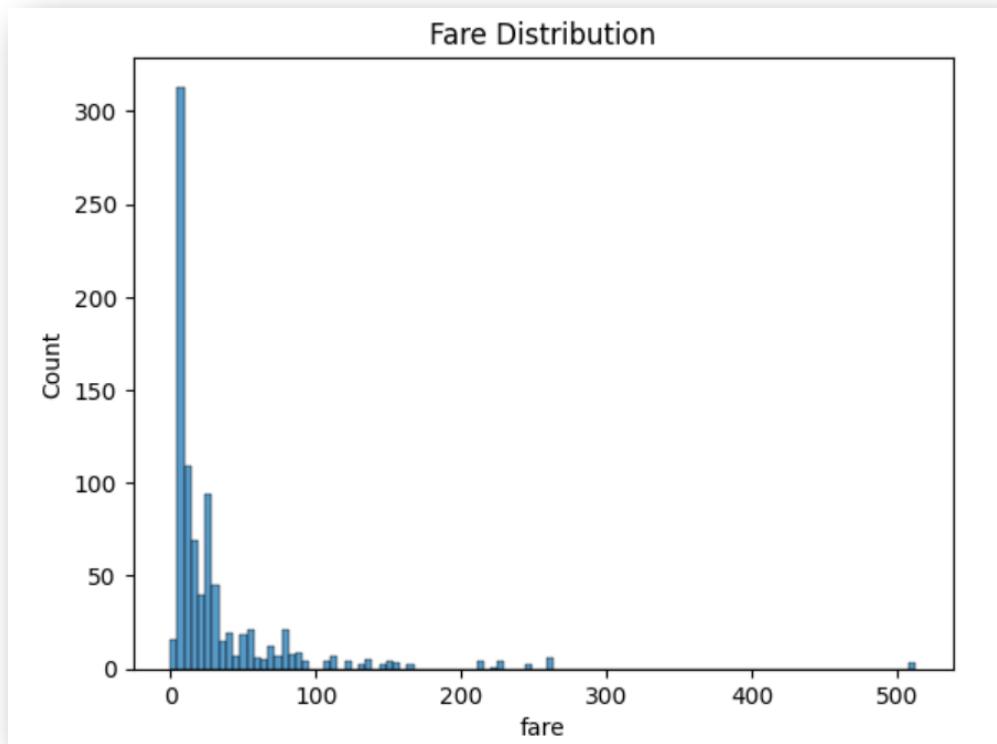
## Univariate analysis Numerical Data

### Histograms

A histogram is a bar chart that displays the frequency of a numerical variable's values. It is created by dividing the data into intervals, called bins, and counting the number of observations that fall within each bin.

```
▶ sns.histplot(data=titanic, x="age")
  plt.title("Age Distribution")
  plt.show()
```





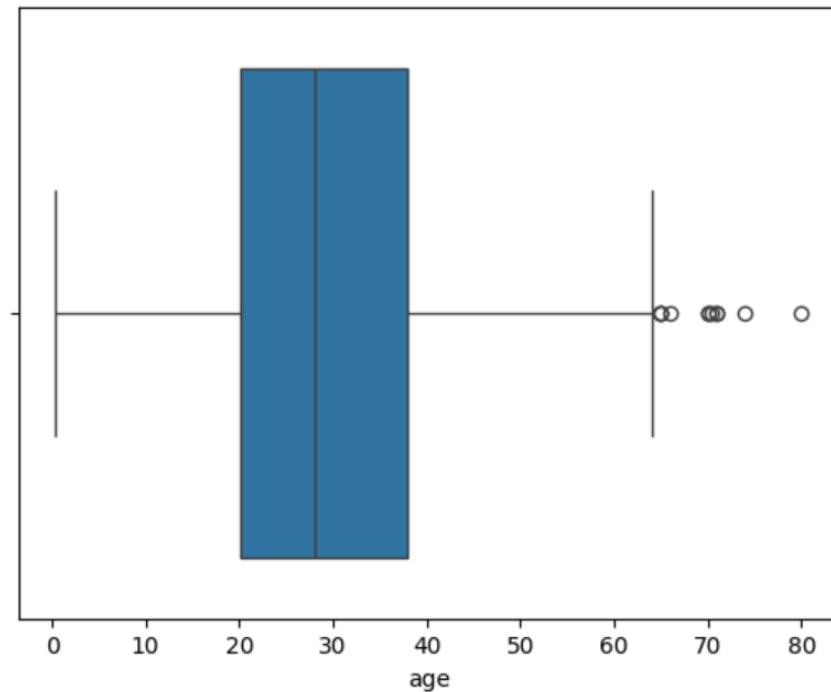
Further study of  
distributions...

Perhaps an outlier?

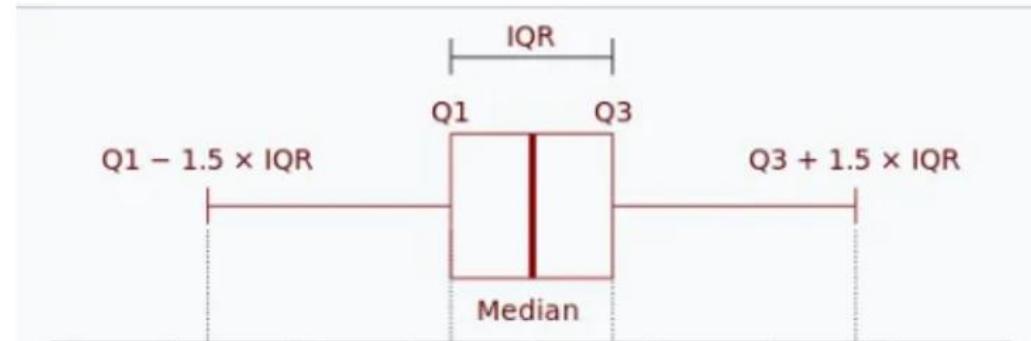
## Box Plots

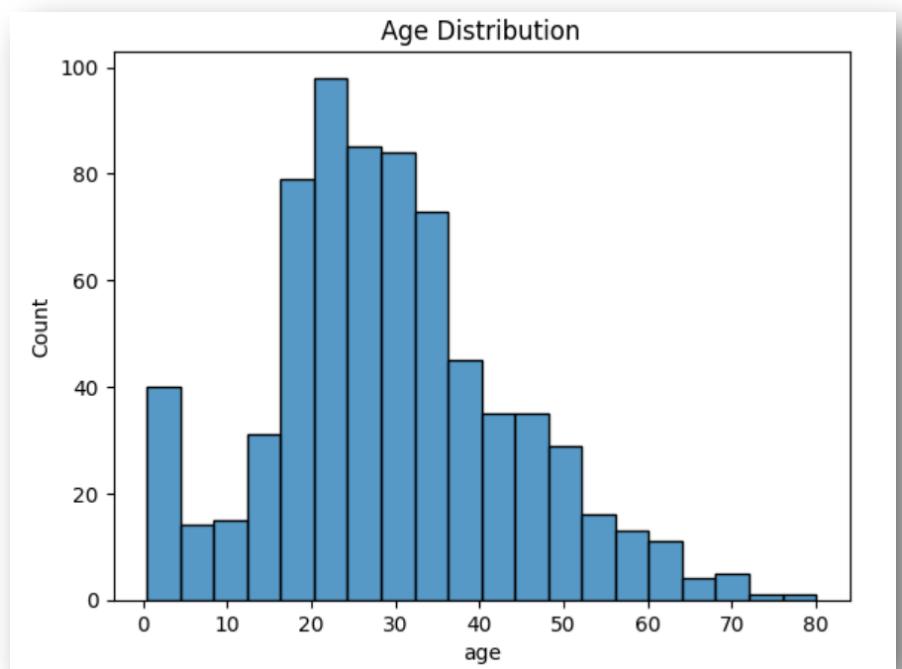
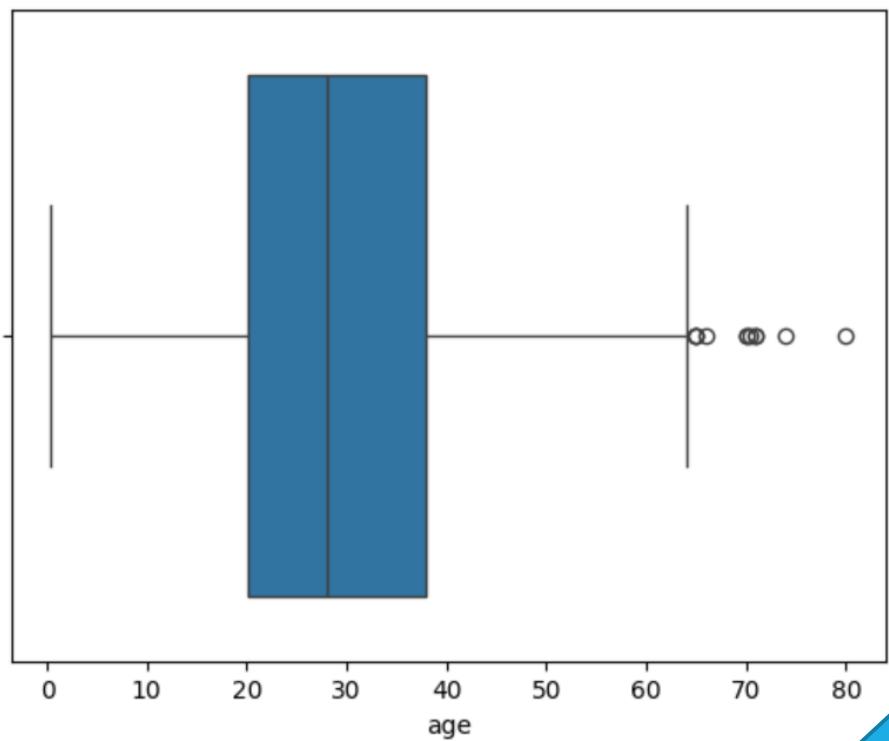
A box plot is a graphical summary of a numerical distribution showing its median, quartiles, spread, and potential outliers.

```
▶ sns.boxplot(x="age", data=titanic)
```



- The **middle line** represents the **median**, which means that 50% of the data points lie below this line, and 50% lie above.
- Then we have the **first quartile (Q1)** and the **third quartile (Q3)**. The median is the second quartile.
- The range between Q1 and Q3 is called the **interquartile range (IQR)**, which represents 50% of the data.
- The **whiskers** extend from Q1 and Q3 to the **minimum** and **maximum** values within a defined range.

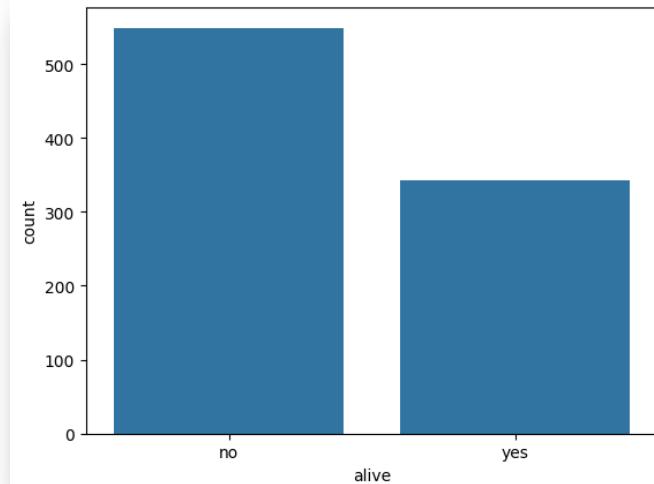




Often used side by side

## Univariate analysis Categorical Data

```
sns.countplot(x="alive", data=titanic)  
plt.show()
```



## Count Plots

A count plot is a type of bar plot that displays the count of observations in each category.

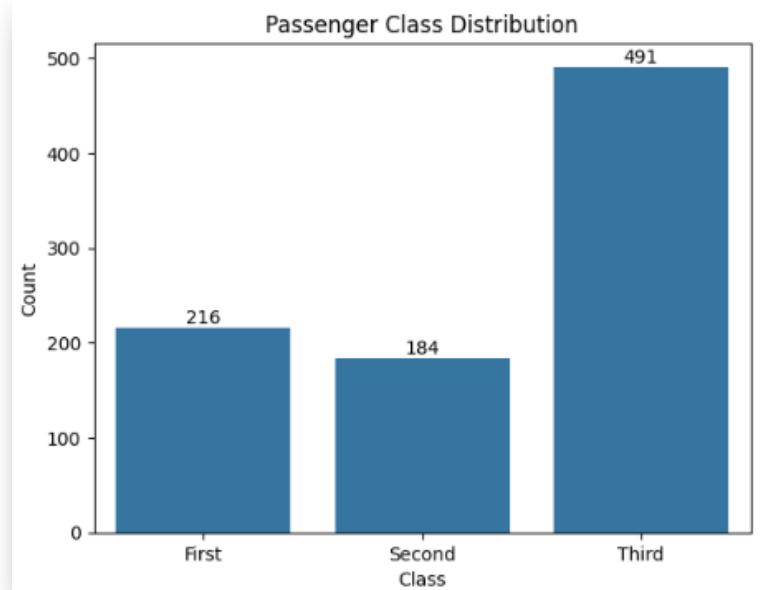
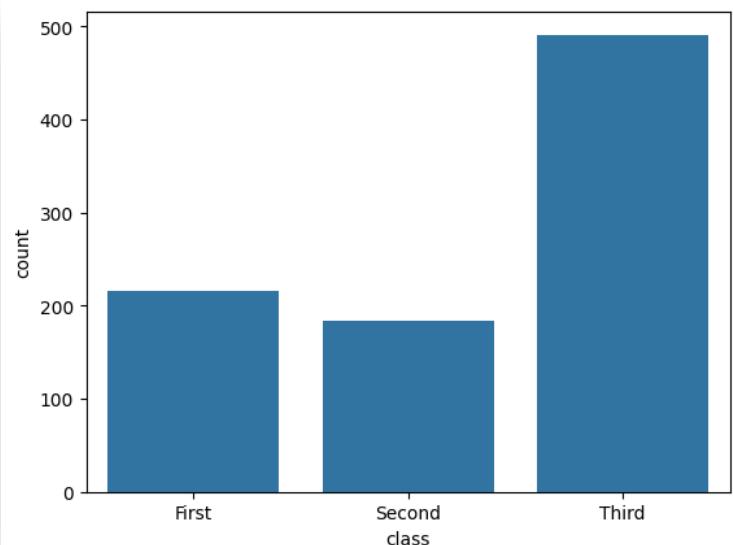
Data obtained about 891 passengers on the Titanic shows that most passengers died in the incident.

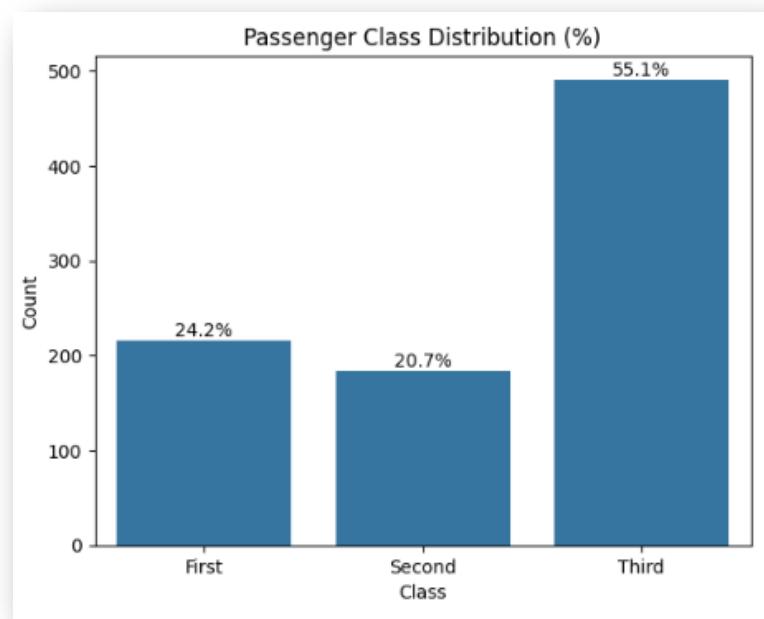
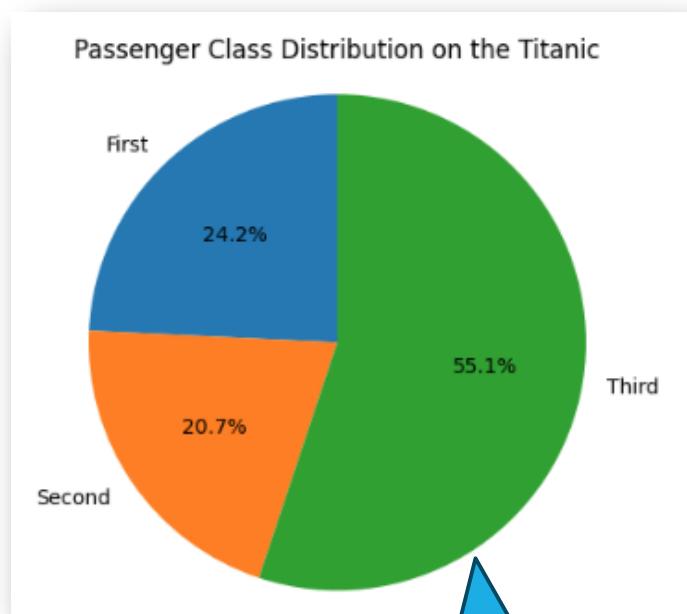


## Communication Skills



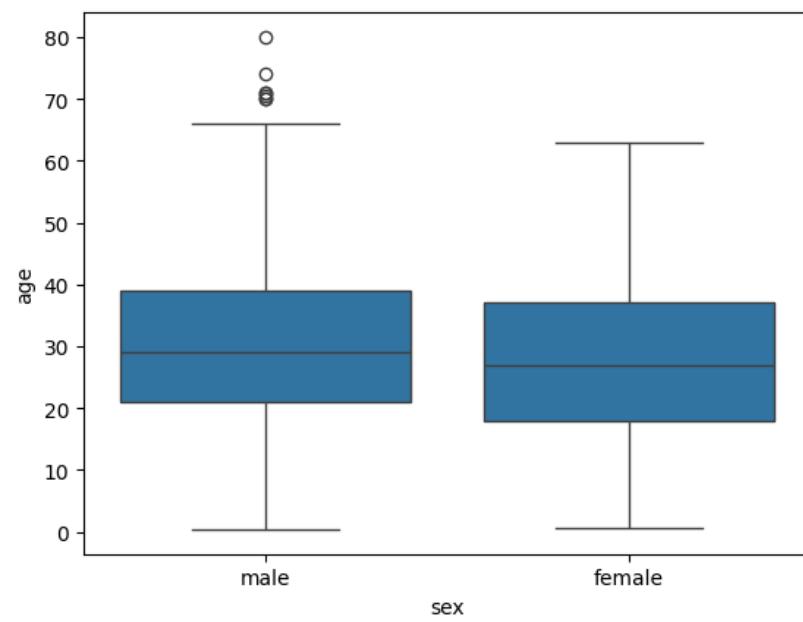
Good practice to write a sentence (insight) from what you observe in the graph.





What about Pie Chart?

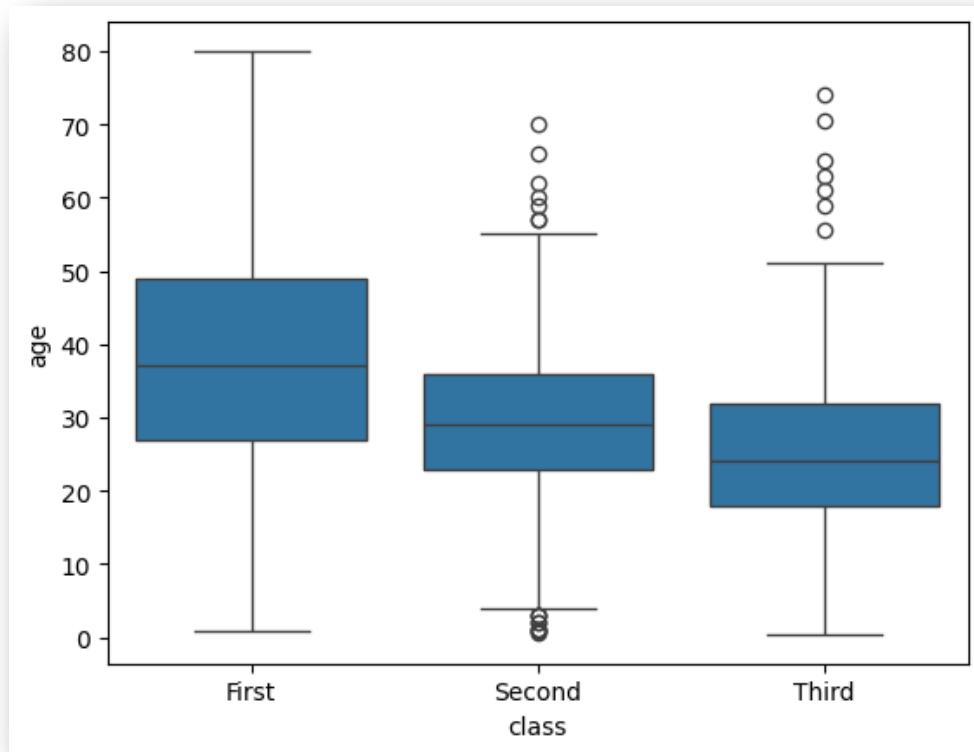
Bivariate analysis  
Numerical/Categorical



Shows central tendency statistics of one variable according to values of a second variable

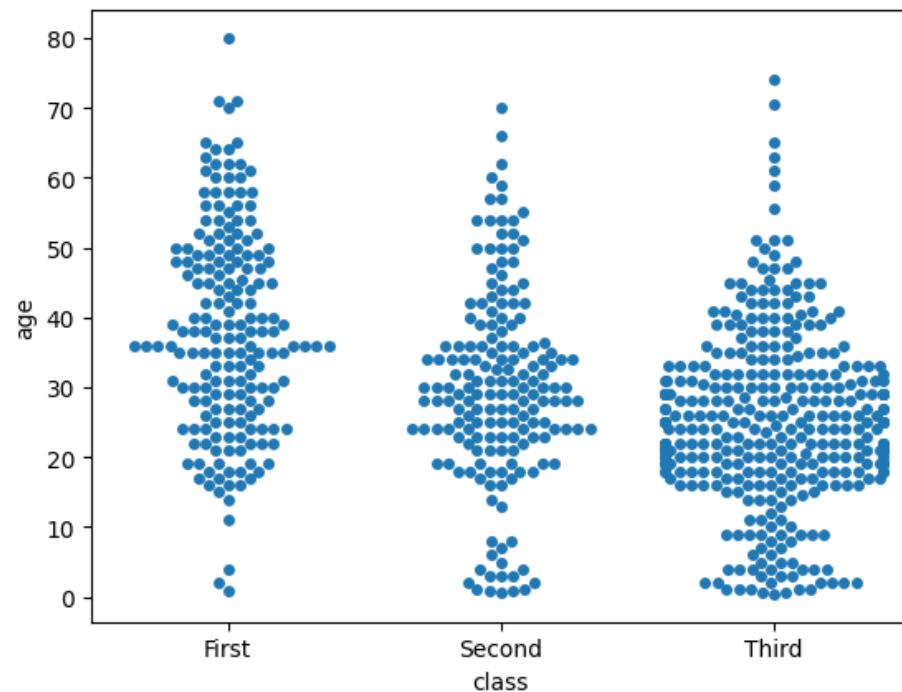
```
sns.boxplot(x="sex", y="age", data=titanic)
```

```
sns.boxplot(x="class", y="age", data=titanic)
```

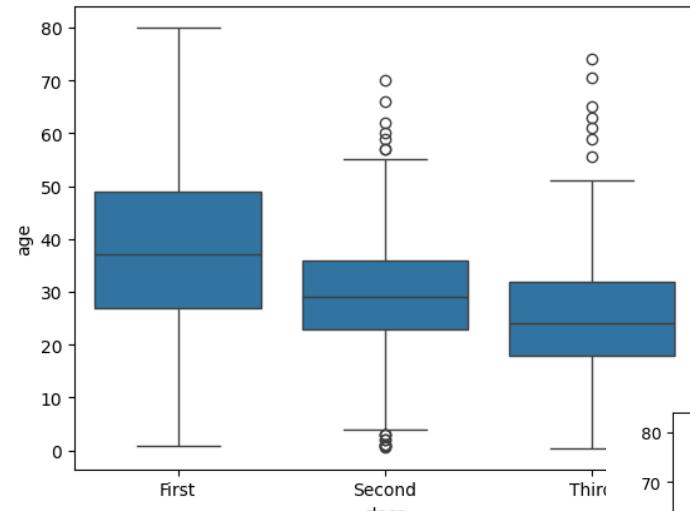


Insight?

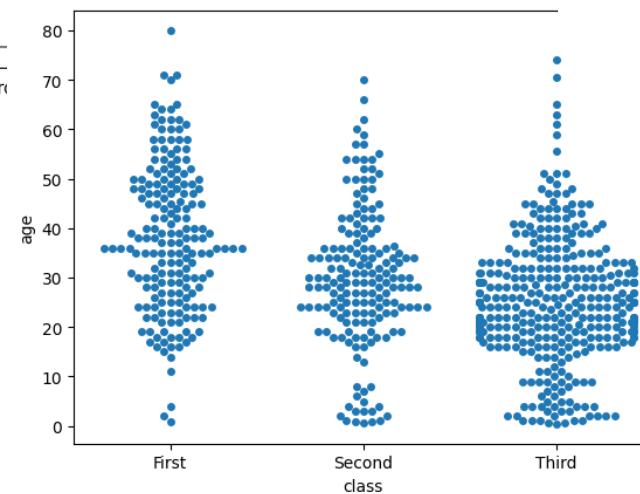
```
sns.swarmplot(x="class", y="age", data=titanic)
```



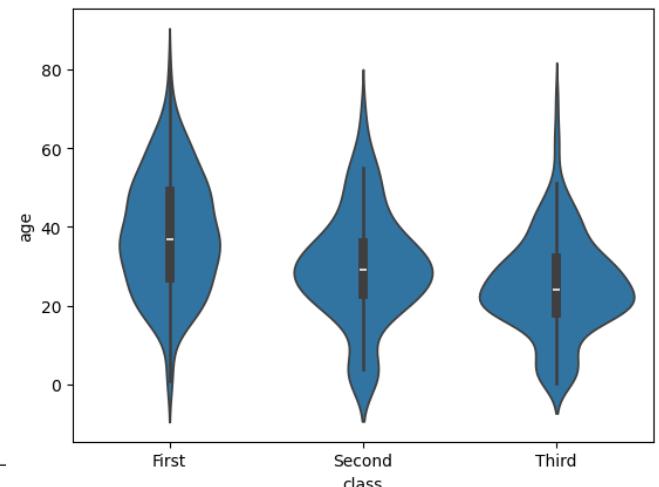
What do you think?



Which would you put in the report? Only a visual (preference) difference?



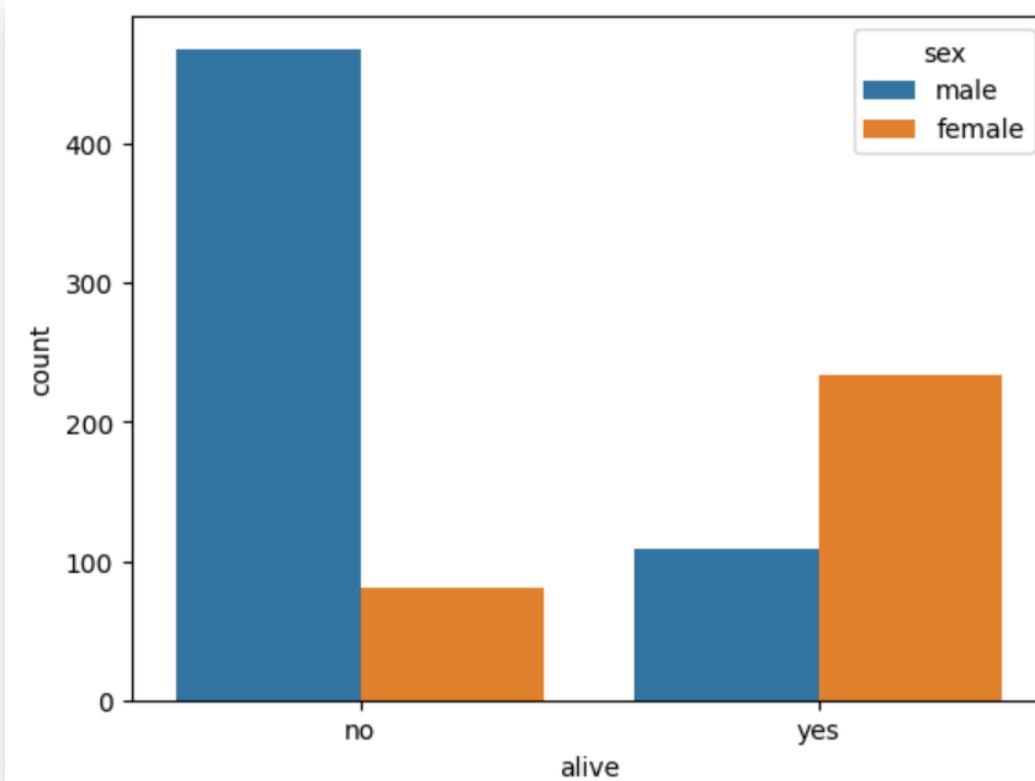
```
sns.swarmplot(x="class", y="age", data=titanic)
```



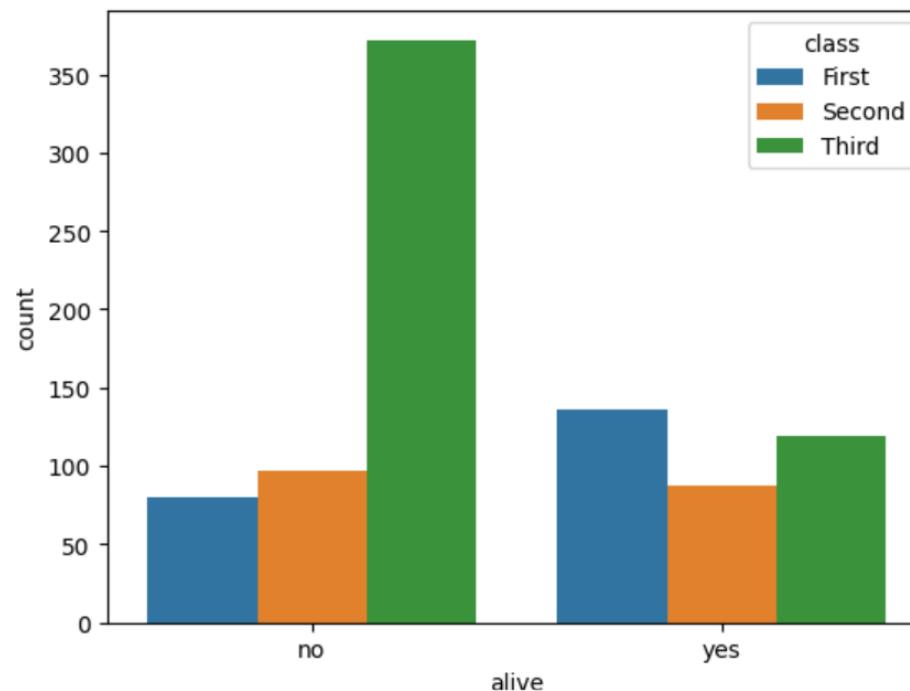
```
sns.violinplot(x="class", y="age", data=titanic)
```

## Bivariate analysis Categorical/Categorical

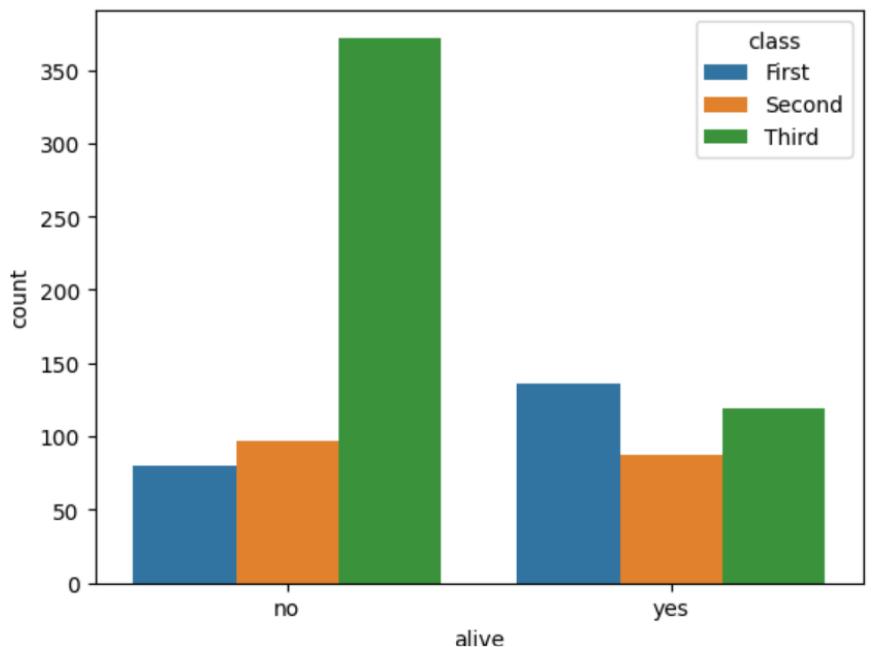
```
sns.countplot(x="alive", hue="sex", data=titanic)  
plt.show()
```



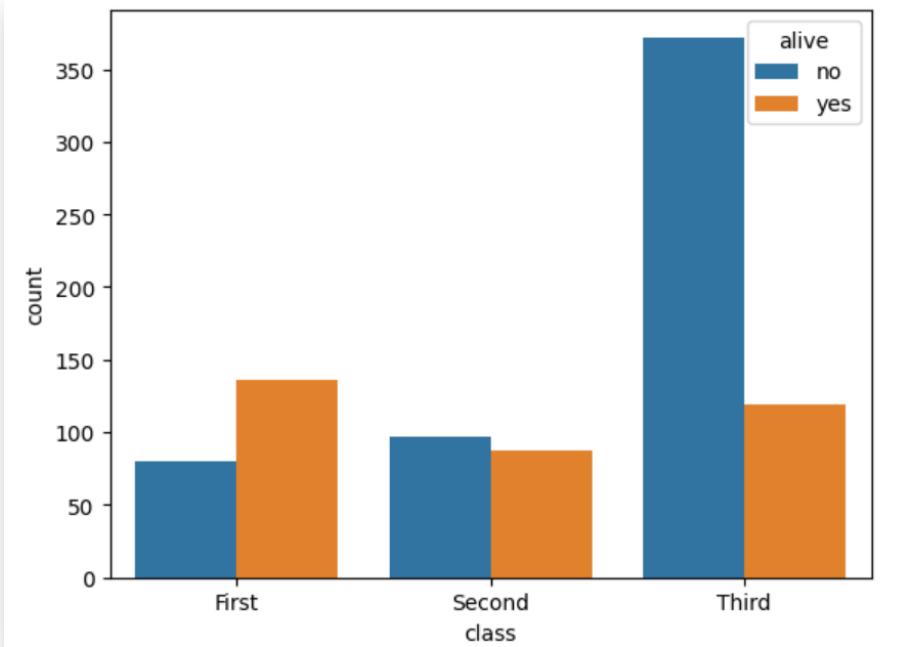
```
sns.countplot(x="alive", hue="class", data=titanic)  
plt.show()
```



```
sns.countplot(x="alive", hue="class", data=titanic)  
plt.show()
```

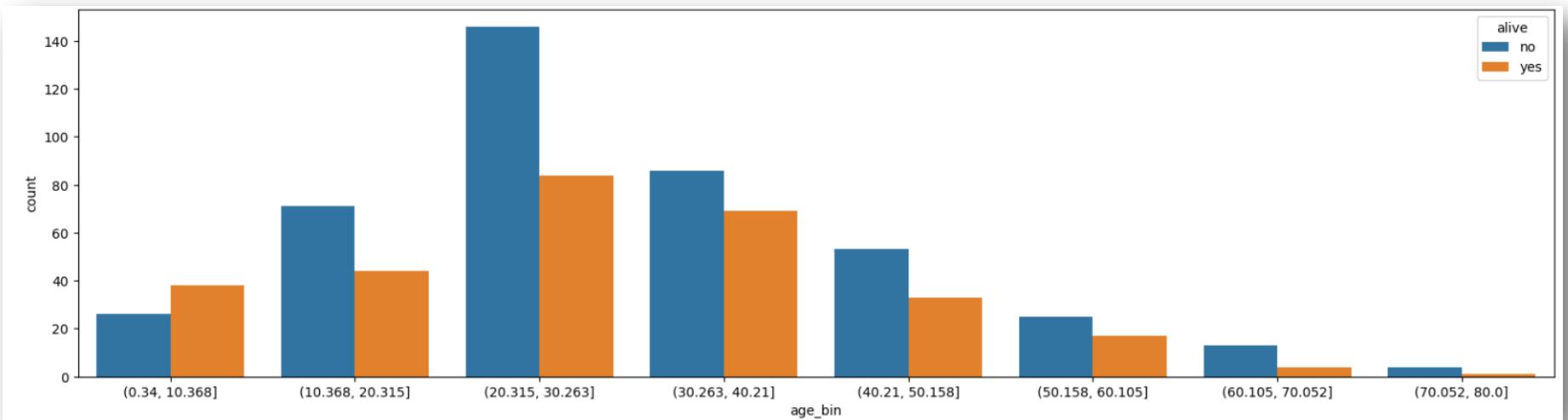


```
sns.countplot(x="class", hue="alive", data=titanic)  
plt.show()
```



## Bivariate analysis Categorical/Numerical

```
titanic['age_bin'] = pd.cut(titanic['age'], 8)
fig = plt.figure(figsize=(20,5))
sns.countplot(x="age_bin", hue="alive", data=titanic)
```



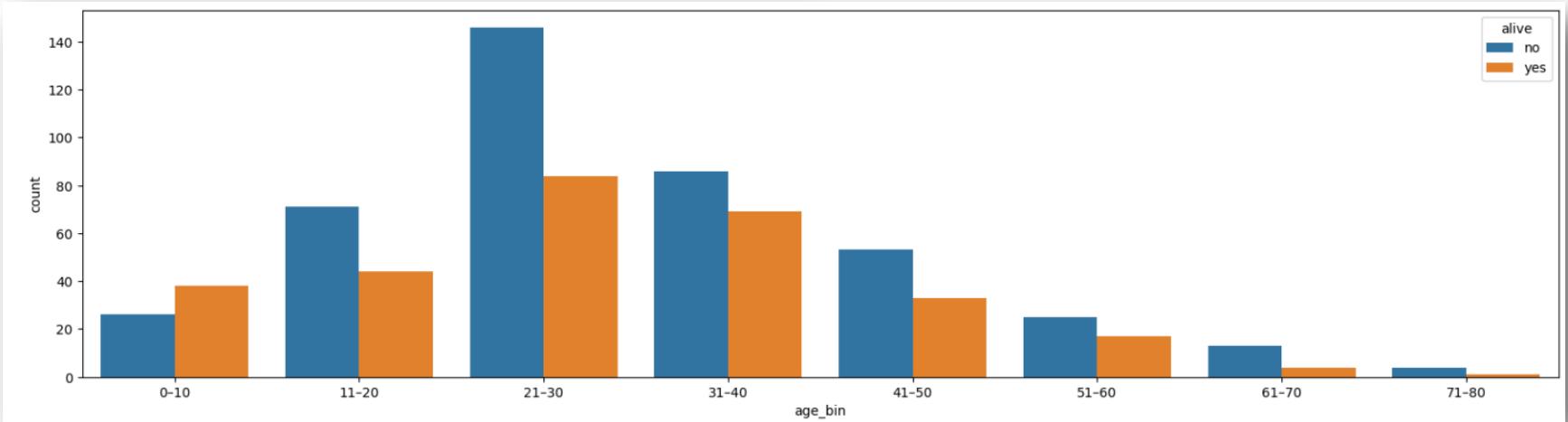
```
bins = [0, 10, 20, 30, 40, 50, 60, 70, 80]
labels = ["0-10", "11-20", "21-30", "31-40", "41-50", "51-60", "61-70", "71-80"]

titanic["age_bin"] = pd.cut(
    titanic["age"],
    bins=bins,
    labels=labels,
    right=True,      # intervals are (a, b]
    include_lowest=True
)

fig = plt.figure(figsize=(20,5))
sns.countplot(x="age_bin", hue="alive", data=titanic)
```



Instead of using the bins like here titanic['age\_bin'] = pd.cut(titanic['age'], 8), how can I make my own bins of 0-10, 11-20, etc

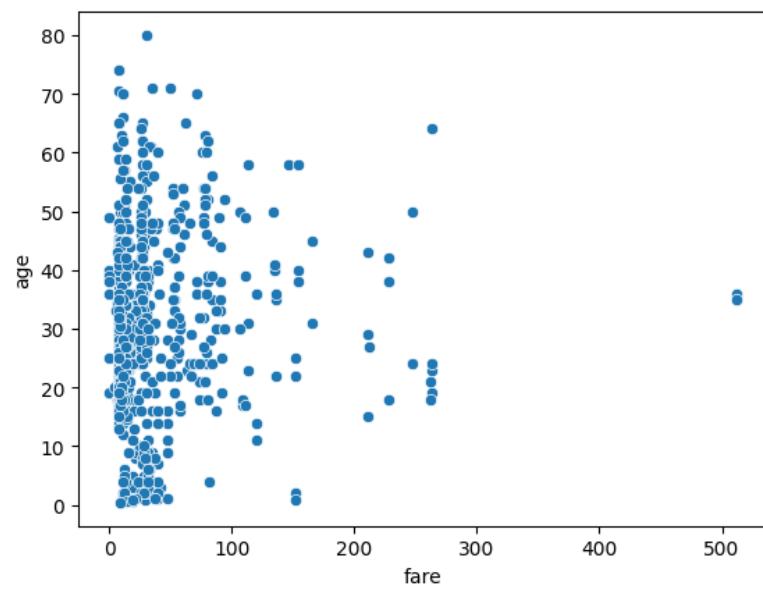


Bivariate analysis  
Numerical/Numerical

## Scatter Plots

A **scatter plot** shows the relationship between two continuous variables, making it great for visualizing correlation.

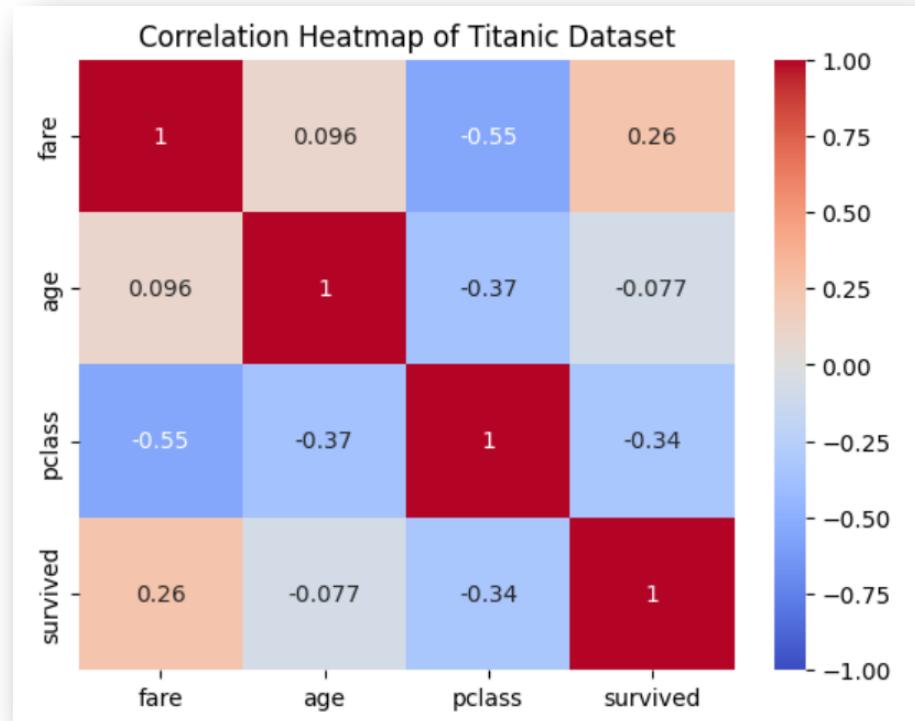
```
sns.scatterplot(x="fare", y="age", data=titanic)
```



```
cols = ["fare", "age", "pclass", "survived"]
corr = titanic[cols].corr()

sns.heatmap(
    corr,
    vmin=-1,
    vmax=1,
    center=0,
    cmap="coolwarm",
    annot=True
)

plt.title("Correlation Heatmap of Titanic Dataset")
plt.show()
```



**Univariate Analysis**

- Numerical feature
- Categorical feature

**Bivariate Analysis**

- Numerical/Categorical
- Categorical/Categorical
- Categorical/Numerical
- Categorical/Categorical



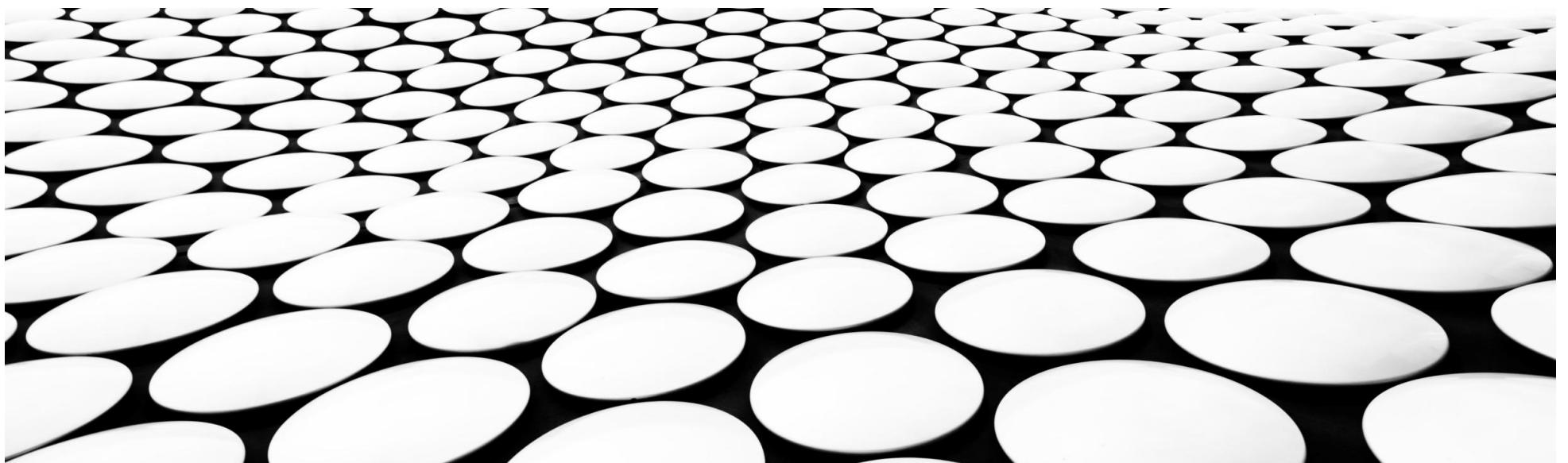
## EXPLORATORY DATA ANALYSIS

- Descriptive statistics
- Diagnostic statistics
- Case Study 1 – Titanic

Friday Case Study 2 +  
Examples for Assignment 1

# EXPLORATORY DATA ANALYSIS

*Story behind the data: Census dataset*





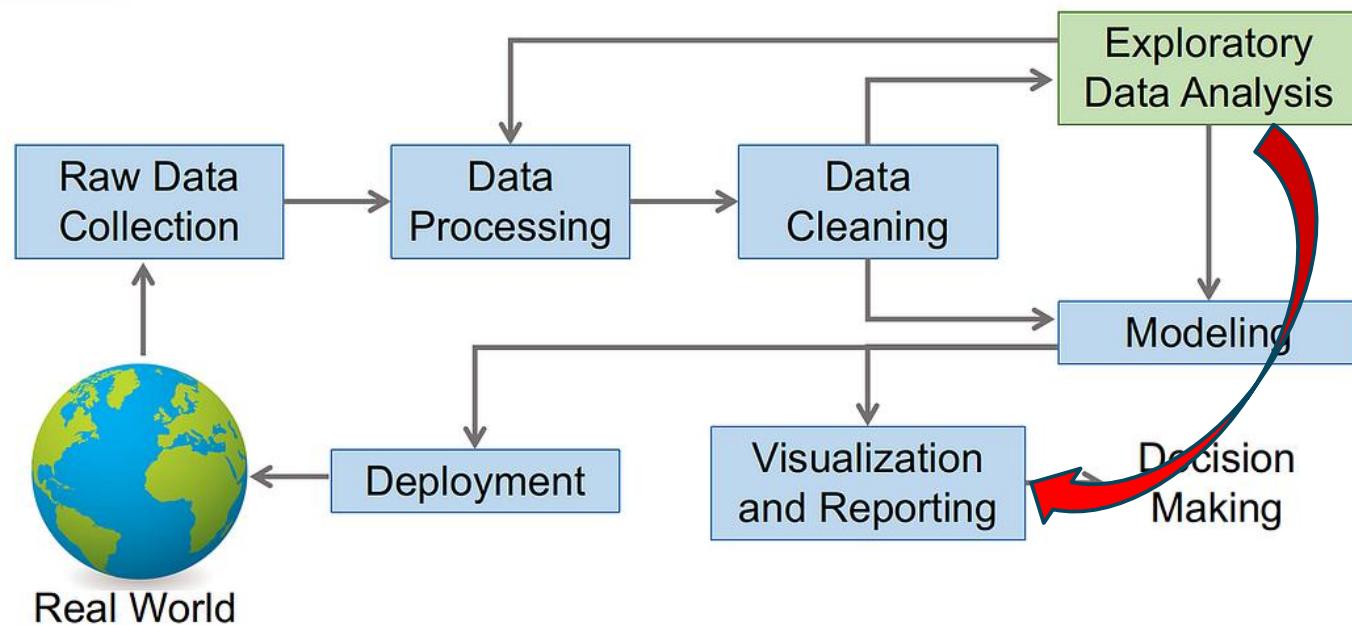
## EXPLORATORY DATA ANALYSIS

- Case Study 2 – Census data
- Assignment 1 Examples

## Data Science Process: A Comprehensive Guide

Abhiljit · Follow  
6 min read · Jan 15, 2024

# Data Science Process

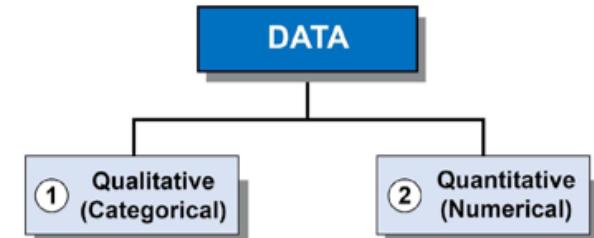


Univariate Analysis

- Numerical feature
- Categorical feature

Bivariate Analysis

- Numerical/Categorical
- Categorical/Categorical
- Categorical/Numerical
- Numerical/Numerical



We didn't have time to explore those...

# *Case Study 2*

## Census Dataset



UC Irvine  
Machine Learning  
Repository

## Adult

Donated on 4/30/1996

Predict whether annual income of an individual exceeds \$50K/yr based on census data. Also known as "Census Income" dataset.

<b>Dataset Characteristics</b>	<b>Subject Area</b>	<b>Associated Tasks</b>
Multivariate	Social Science	Classification
<b>Feature Type</b>	<b># Instances</b>	<b># Features</b>
Categorical, Integer	48842	14

<https://archive.ics.uci.edu/dataset/2/adult>

Explore the dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32561 entries, 0 to 32560
Data columns (total 15 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   age              32561 non-null   int64  
 1   work-class       30725 non-null   object  
 2   fnlwgt           32561 non-null   int64  
 3   education        32561 non-null   object  
 4   education-num    32561 non-null   int64  
 5   marital-status   32561 non-null   object  
 6   occupation       30718 non-null   object  
 7   relationship     32561 non-null   object  
 8   race              32561 non-null   object  
 9   sex               32561 non-null   object  
 10  capital-gain    32561 non-null   int64  
 11  capital-loss    32561 non-null   int64  
 12  hours-per-week  32561 non-null   int64  
 13  native-country   31978 non-null   object  
 14  income            32561 non-null   object
```

Look at a bit of data...

age	work-class	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States <=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States <=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States <=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States <=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba <=50K

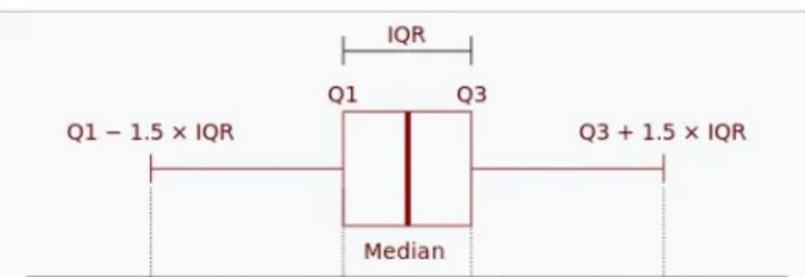
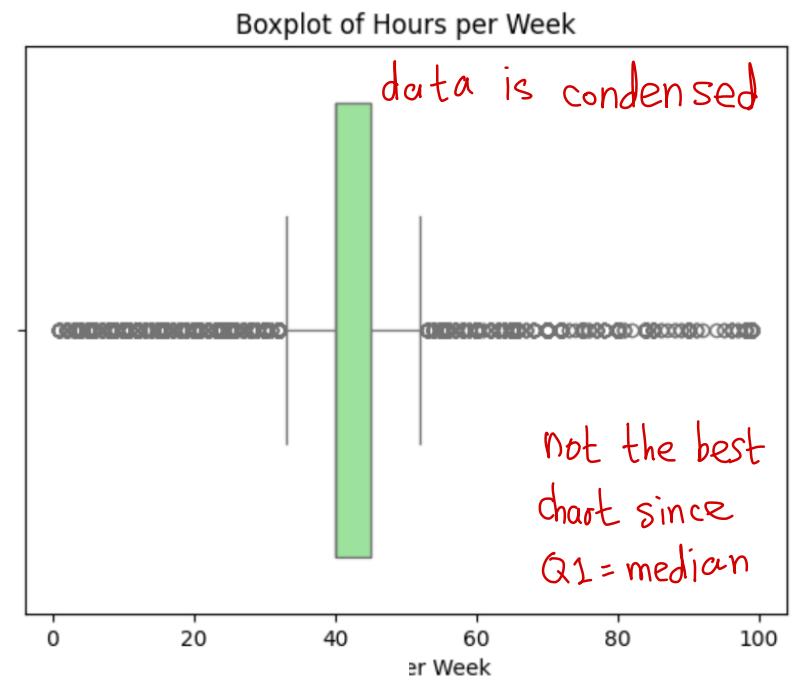
Surveyed

what is fnlwgt in the census dataset from uci 

`fnlwgt` is a sampling weight indicating how many people in the U.S. population a given individual represents.

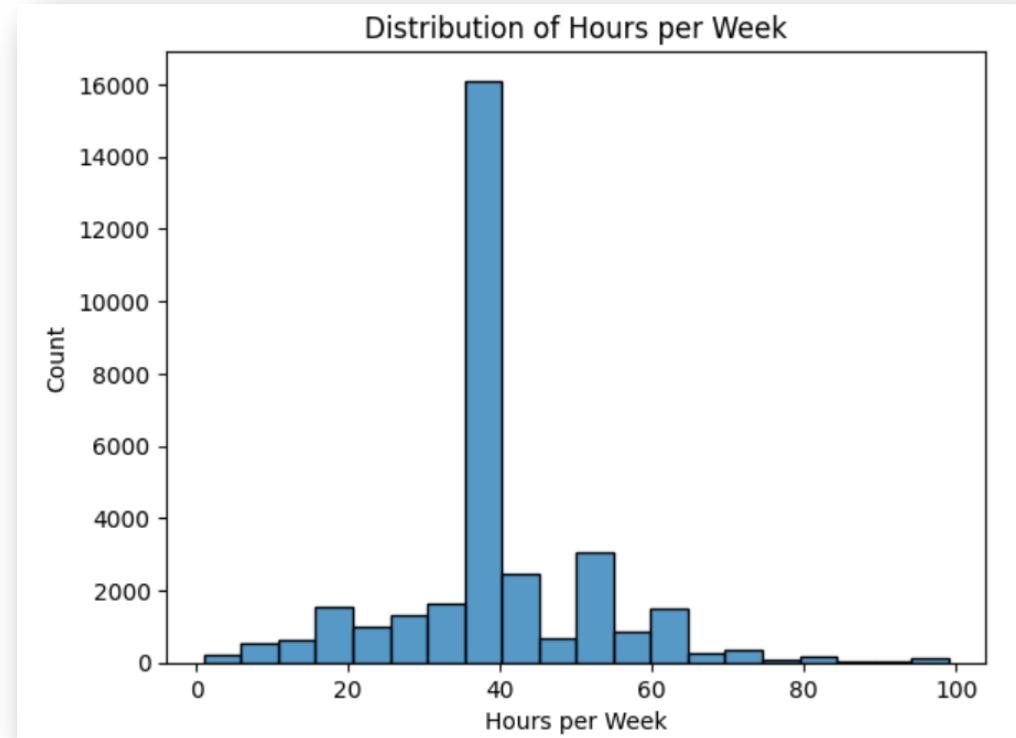
## Univariate analysis Numerical Data

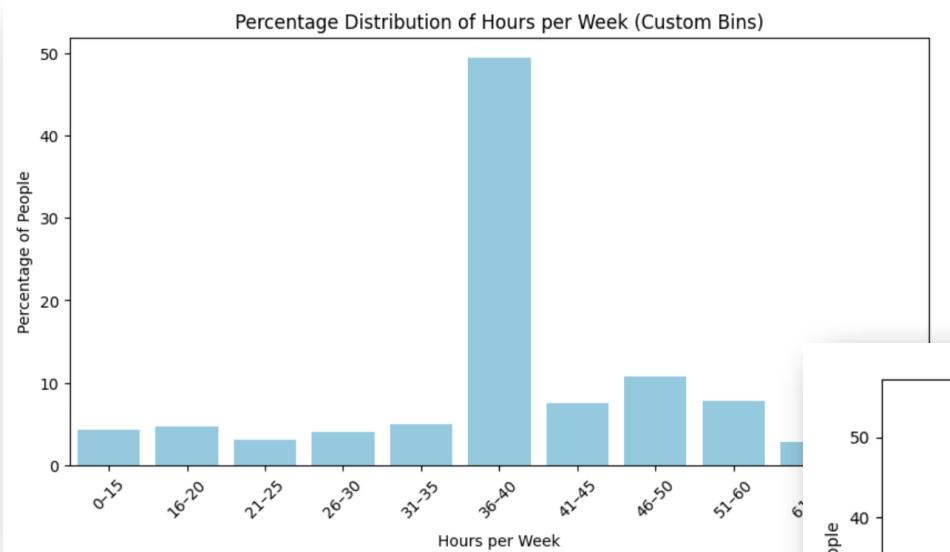
hours-per-week	
<b>count</b>	32561.000000
<b>mean</b>	40.437456
<b>std</b>	12.347429
<b>min</b>	1.000000
<b>25%</b>	40.000000
<b>50%</b>	40.000000
<b>75%</b>	45.000000
<b>max</b>	99.000000



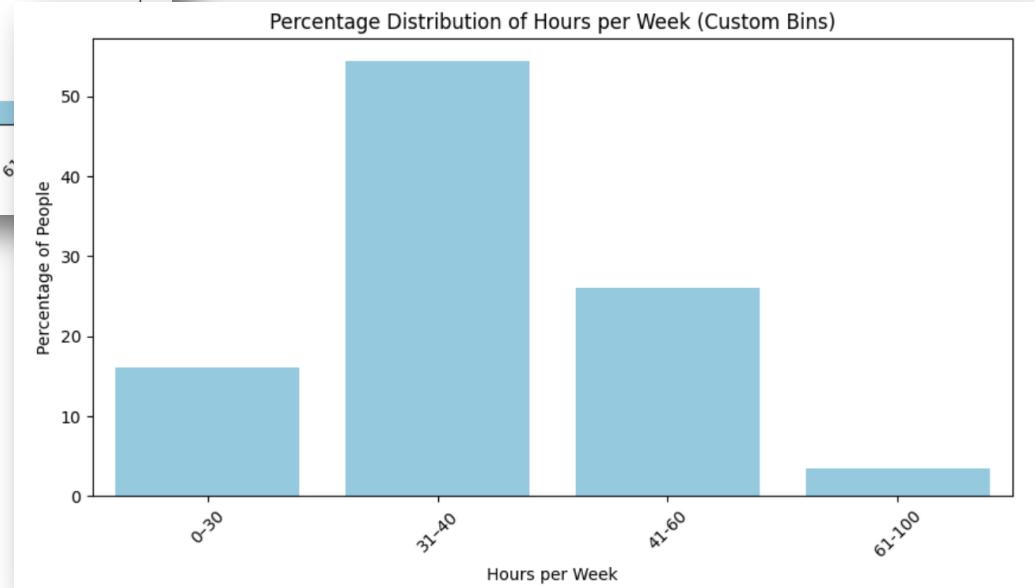
```
census['hours-per-week'].describe()
```

hours-per-week	
count	32561.000000
mean	40.437456
std	12.347429
min	1.000000
25%	40.000000
50%	40.000000
75%	45.000000
max	99.000000

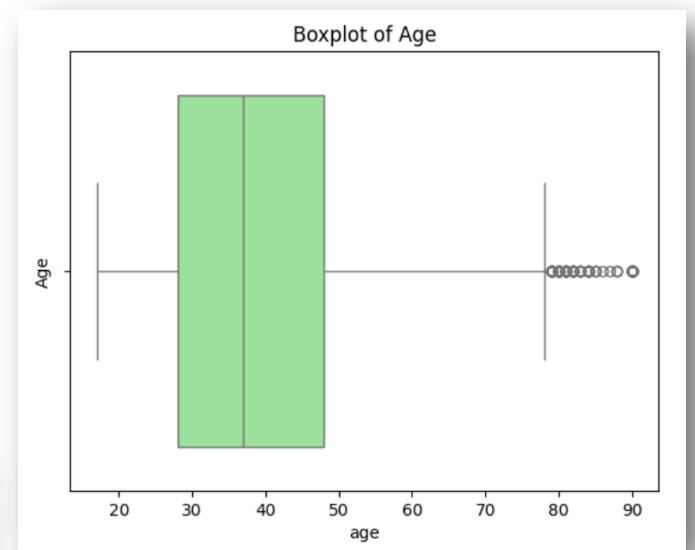
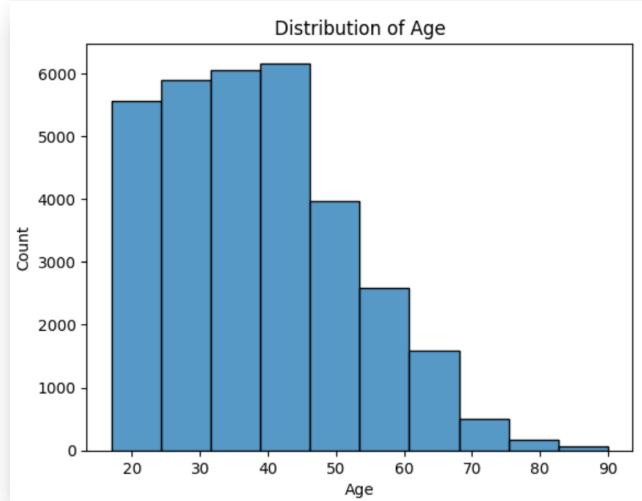




Binning introduces a type of bias...

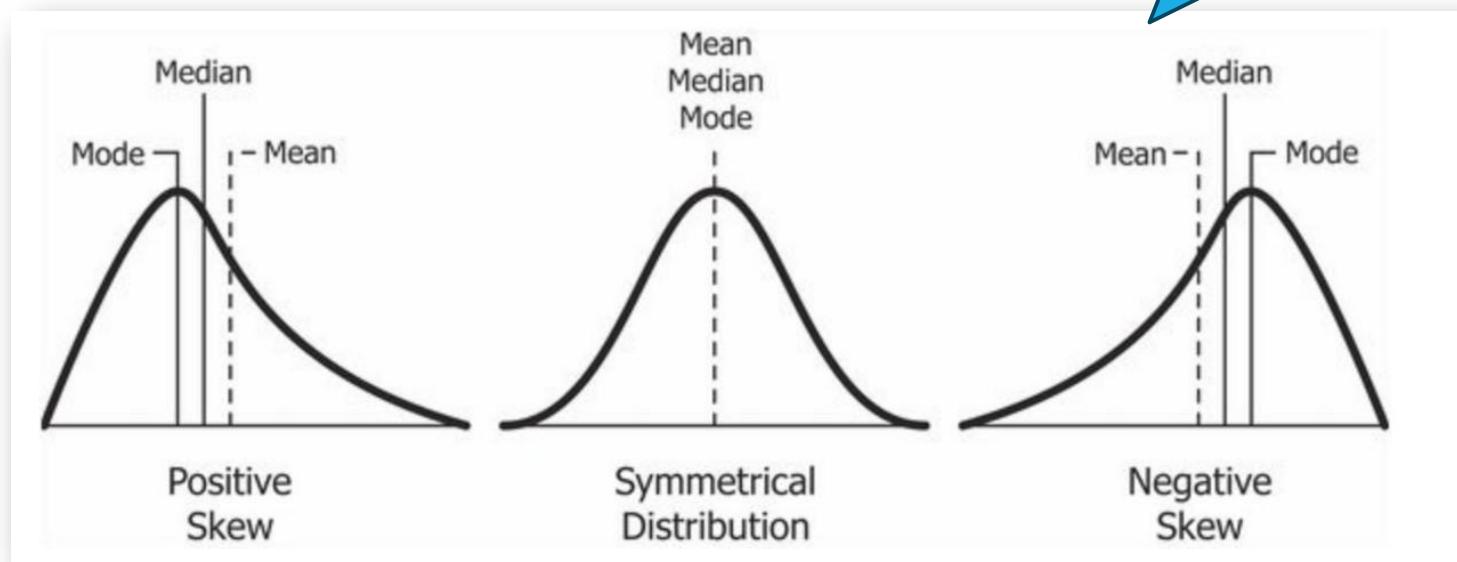


age
count 32561.000000
mean 38.581647
std 13.640433
min 17.000000
25% 28.000000
50% 37.000000
75% 48.000000
max 90.000000





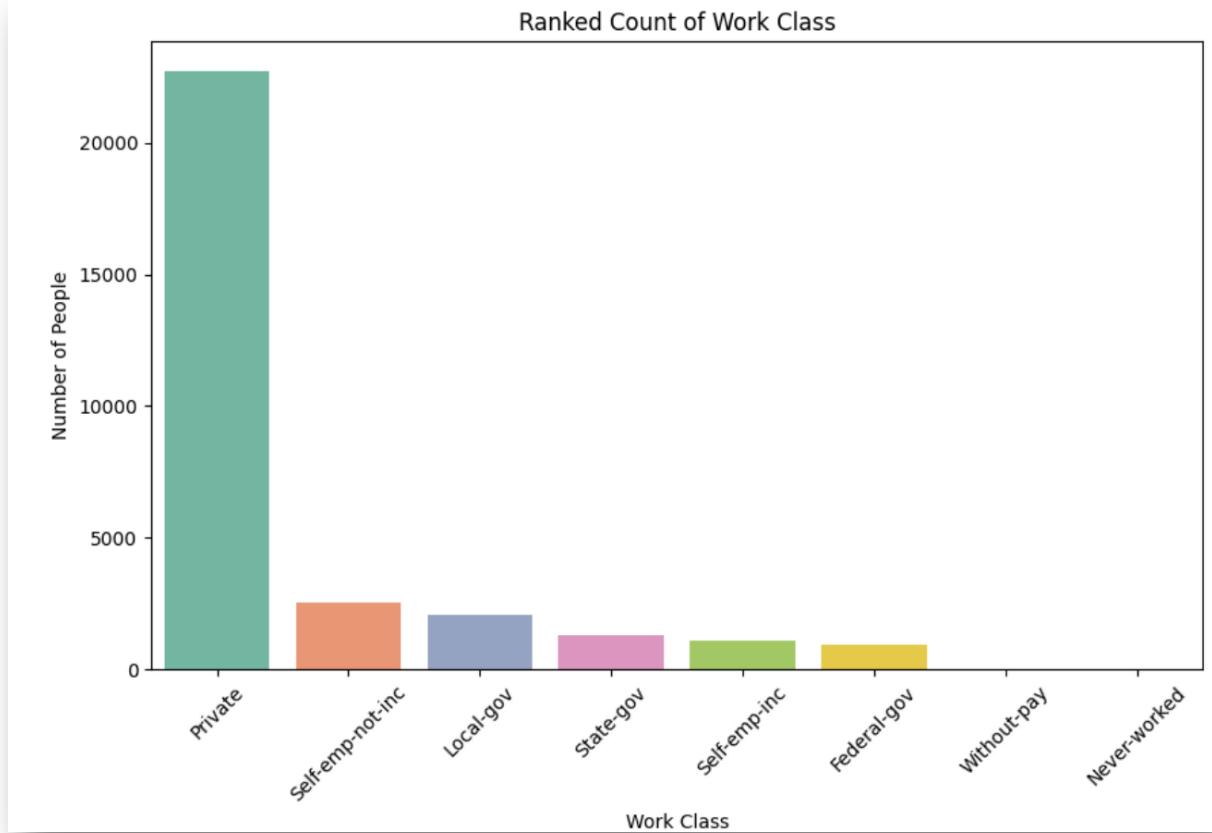
Name refers to direction of the tail



## Univariate analysis Categorical Data

```
census.value_counts("work-class")
```

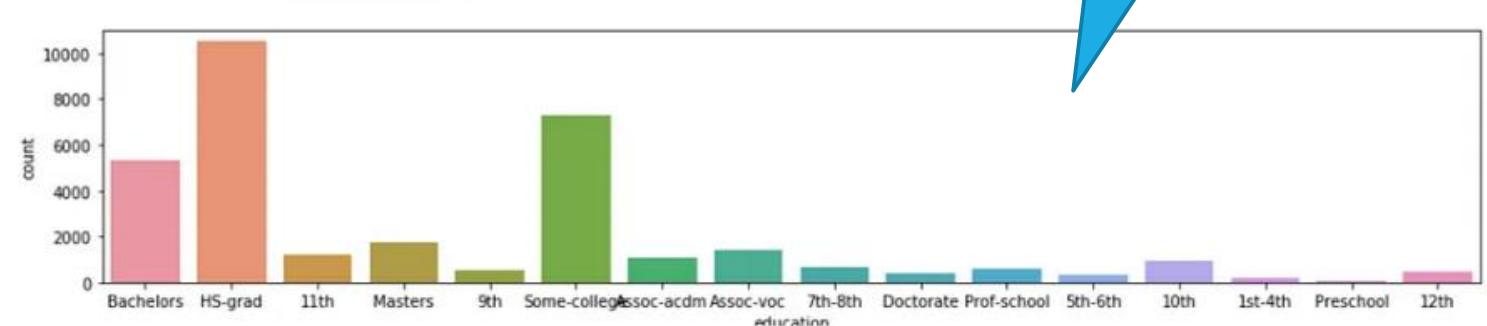
	count
work-class	
Private	22696
Self-emp-not-inc	2541
Local-gov	2093
State-gov	1298
Self-emp-inc	1116
Federal-gov	960
Without-pay	14
Never-worked	7



We don't know meaning of the variable names  
Add a legend

```
census.value_counts("education")
```

	count
education	
HS-grad	10501
Some-college	7291
Bachelors	5355
Masters	1723
Assoc-voc	1382
11th	1175
Assoc-acdm	1067
10th	933
7th-8th	646
Prof-school	576
9th	514
12th	433
Doctorate	413
5th-6th	333
1st-4th	168
Preschool	51

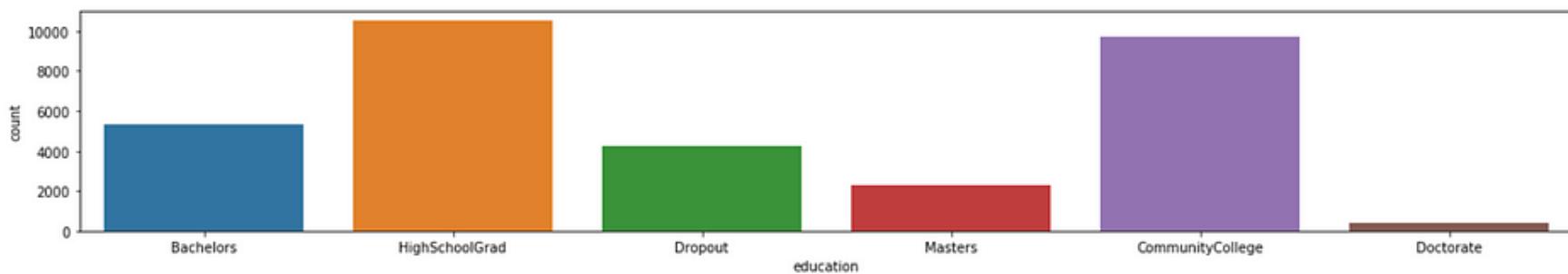


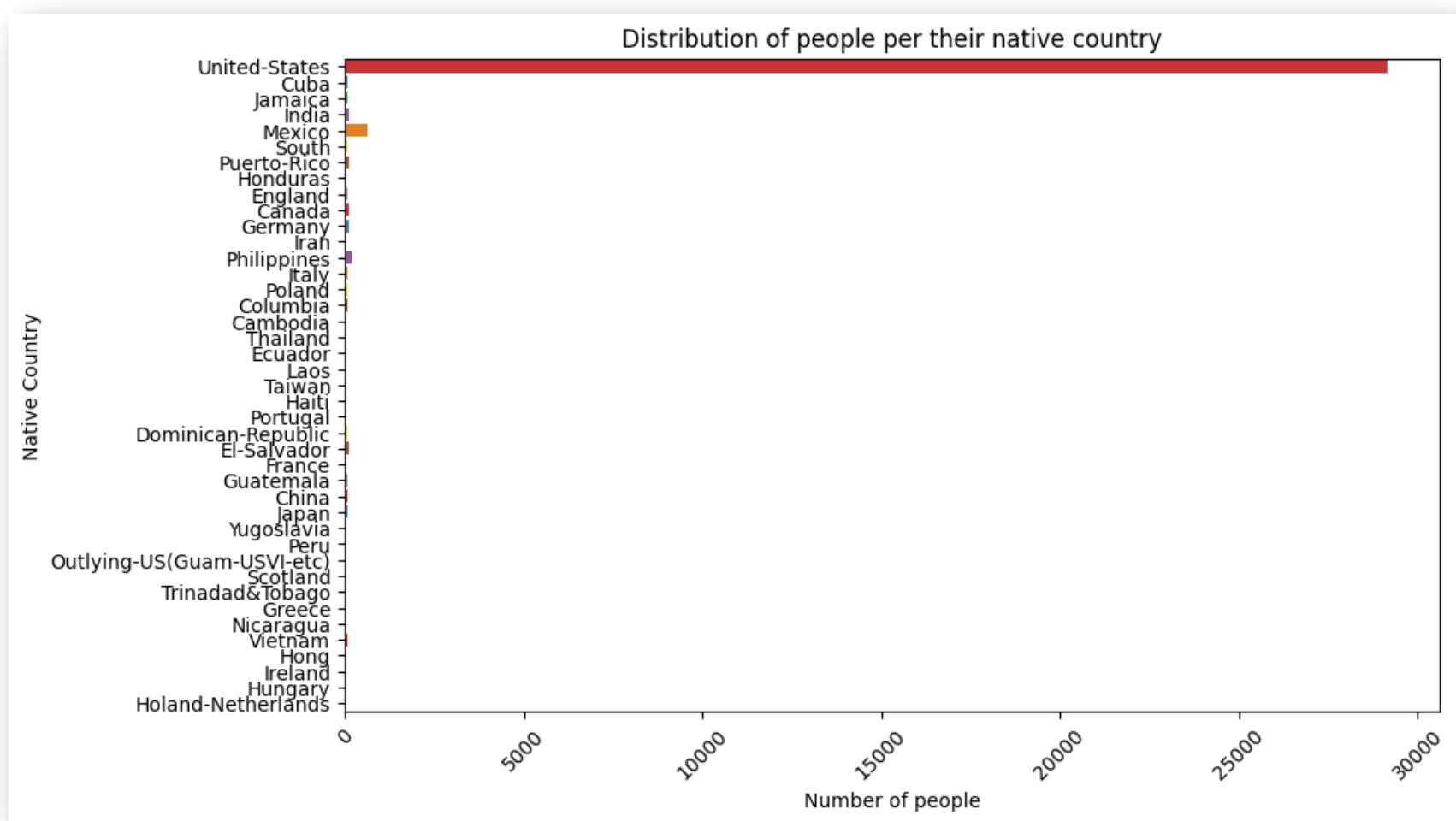
Too many possible values,  
hard to interpret

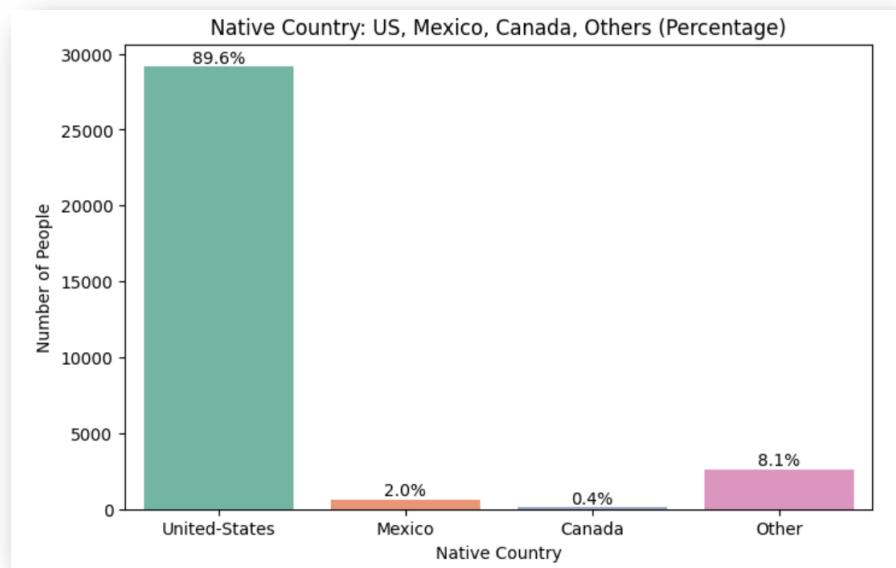
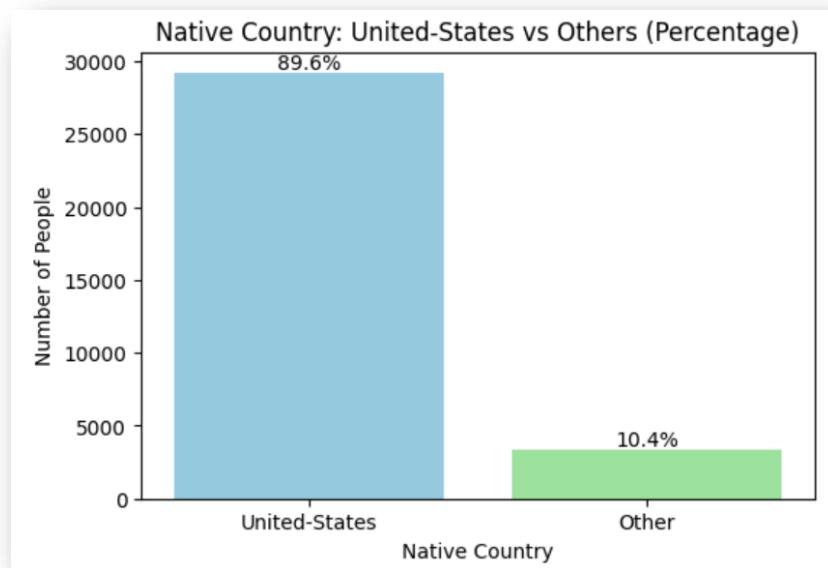
```
In [28]: #Grouping education
data['education'].replace('Preschool', 'Dropout', inplace=True)
data['education'].replace('10th', 'Dropout', inplace=True)
data['education'].replace('11th', 'Dropout', inplace=True)
data['education'].replace('12th', 'Dropout', inplace=True)
data['education'].replace('1st-4th', 'Dropout', inplace=True)
data['education'].replace('5th-6th', 'Dropout', inplace=True)
data['education'].replace('7th-8th', 'Dropout', inplace=True)
data['education'].replace('9th', 'Dropout', inplace=True)
data['education'].replace('HS-Grad', 'HighSchoolGrad', inplace=True)
data['education'].replace('HS-grad', 'HighSchoolGrad', inplace=True)
data['education'].replace('Some-college', 'CommunityCollege', inplace=True)
data['education'].replace('Assoc-acdm', 'CommunityCollege', inplace=True)
data['education'].replace('Assoc-voc', 'CommunityCollege', inplace=True)
data['education'].replace('Prof-school', 'Masters', inplace=True)

fig = plt.figure(figsize=(20,3))
sns.countplot(x="education", data=data)
```

Domain knowledge necessary to group categories, do you agree with their groups?



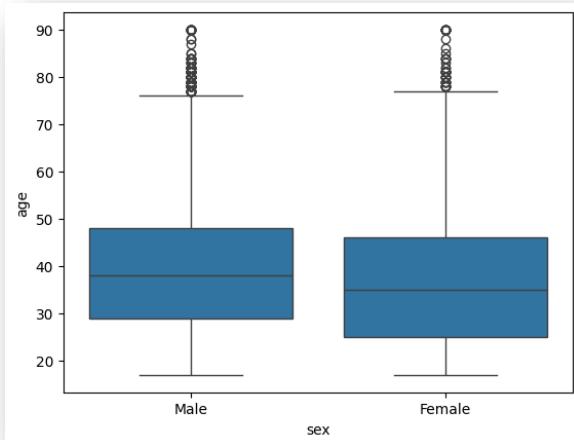




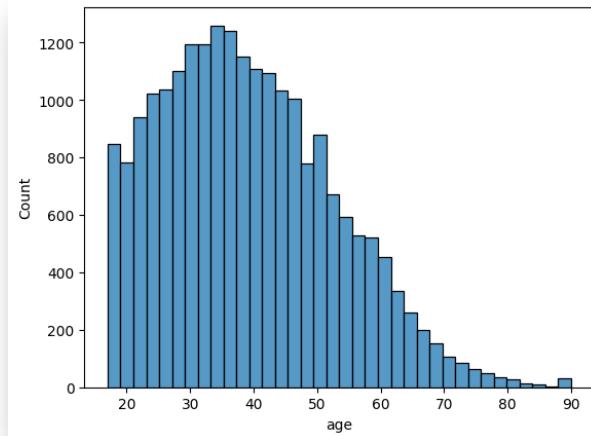
Grouping introduces  
a particular view

Can introduce bias

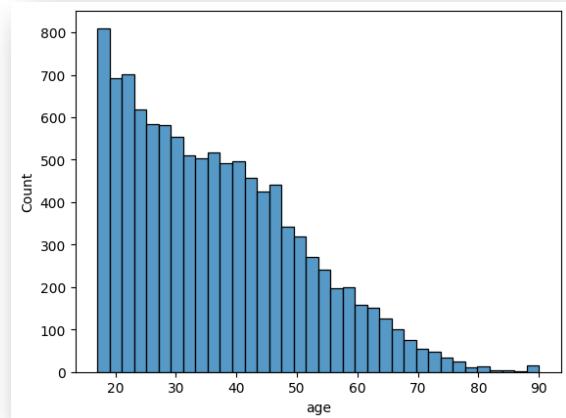
## Bivariate analysis Numerical/Categorical



Age distribution of Male



Age distribution of Female



## Bivariate analysis Categorical/Categorical

```
In [31]: # Plotting cross tabulation values for native-country and sex  
pd.crosstab(data['native-country'], data['sex'], margins=True)
```

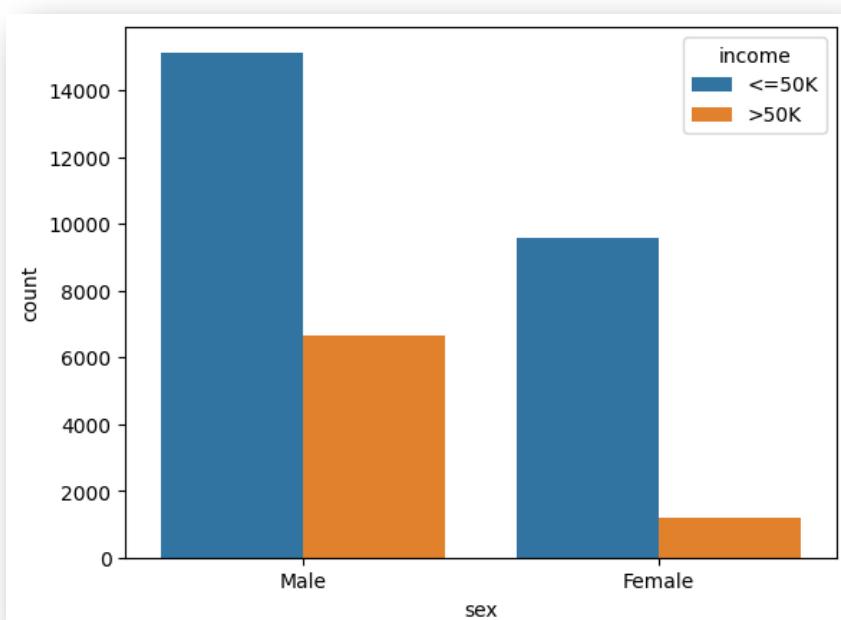
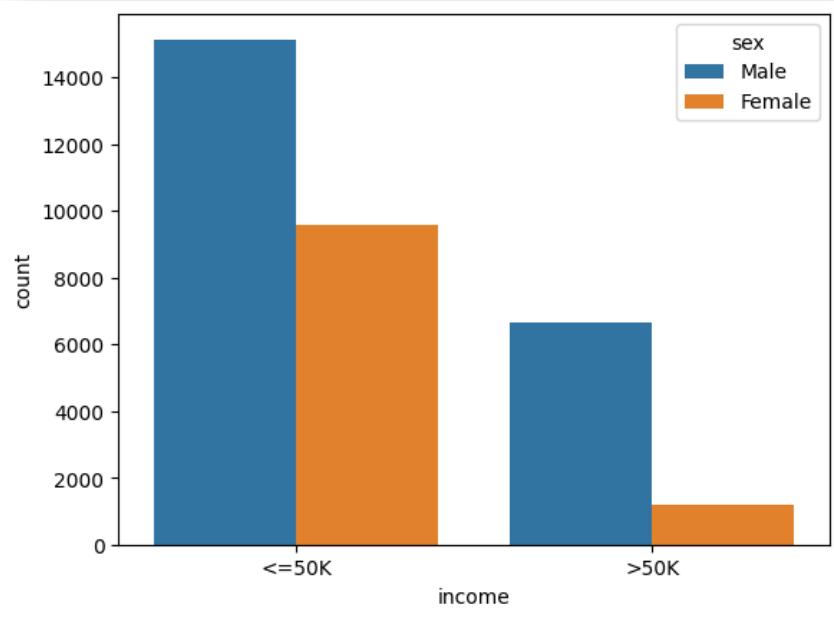
Out[31]:

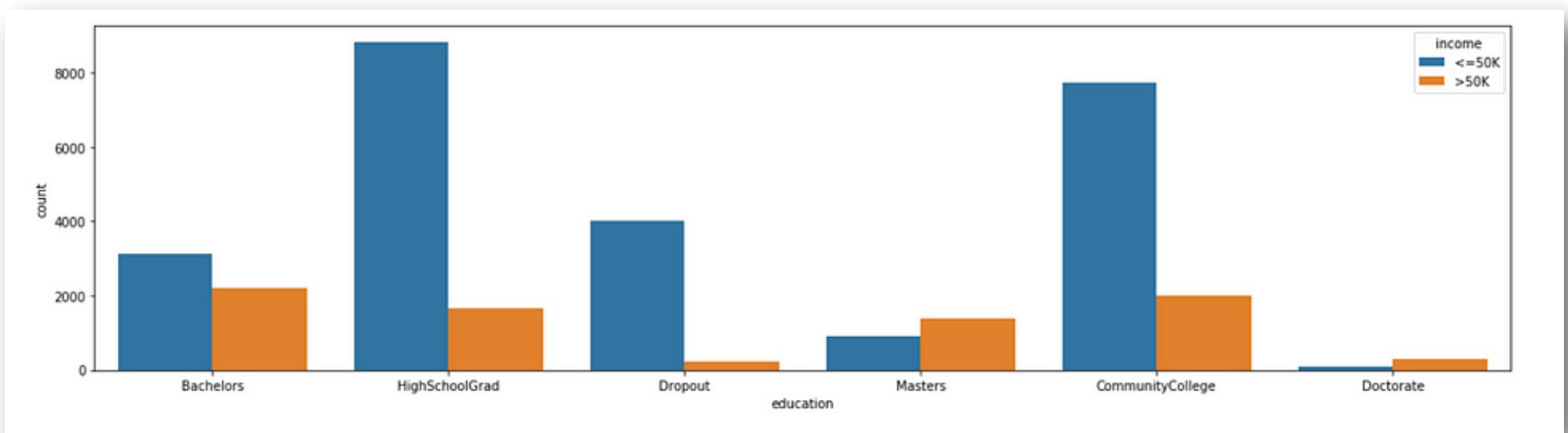
sex	Female	Male	All
native-country			
Others	1089	2302	3391
United-States	9682	19488	29170
All	10771	21790	32561



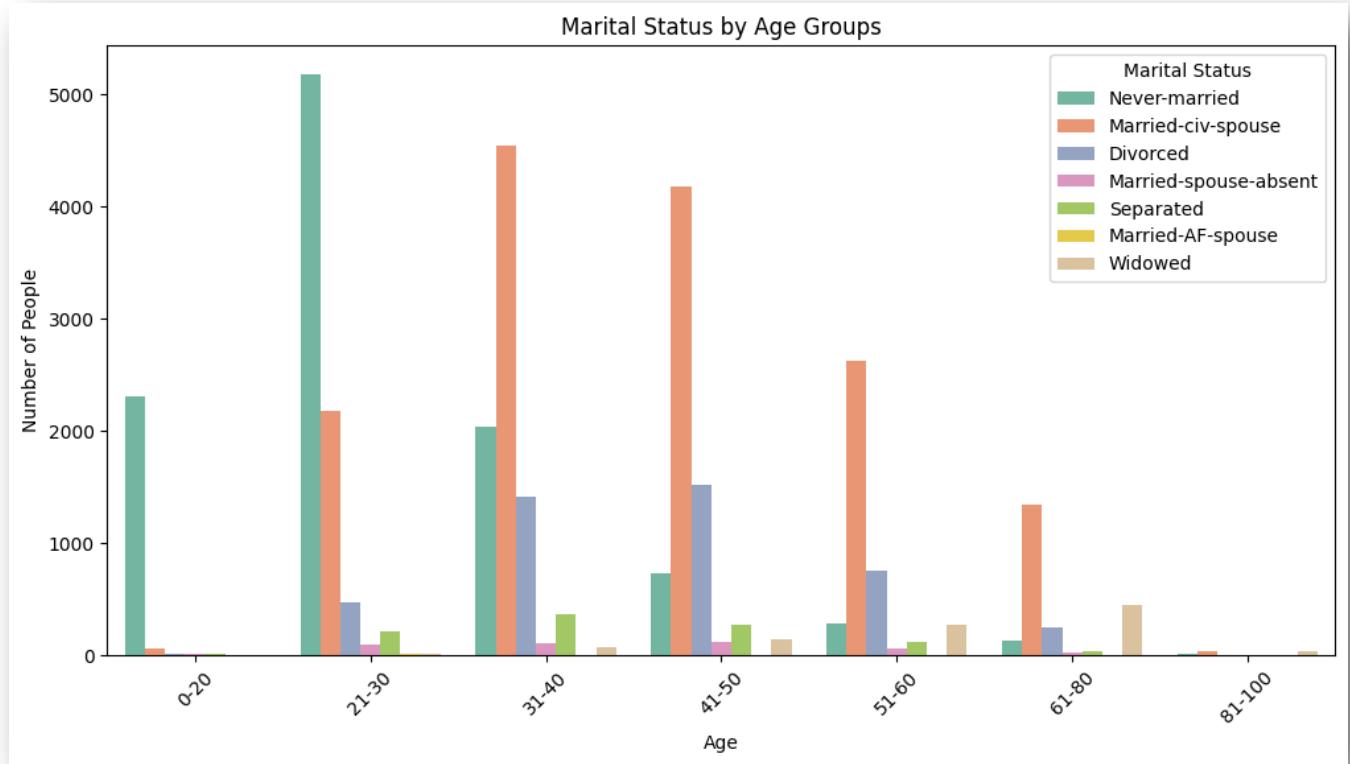
We look at two categorical attributes

Same information, but would it matter  
what to put in a report?





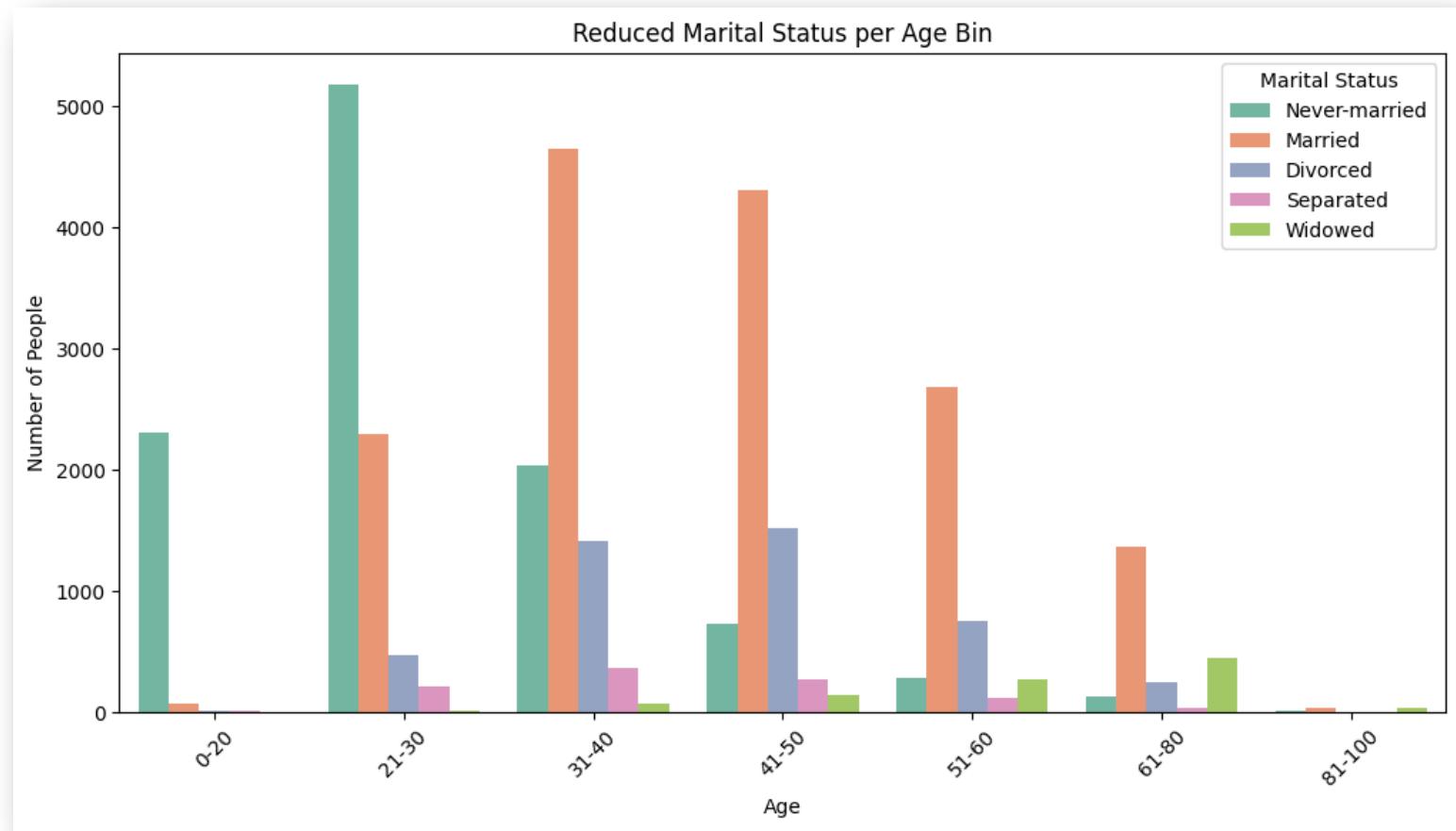
## Bivariate analysis Categorical/Numerical



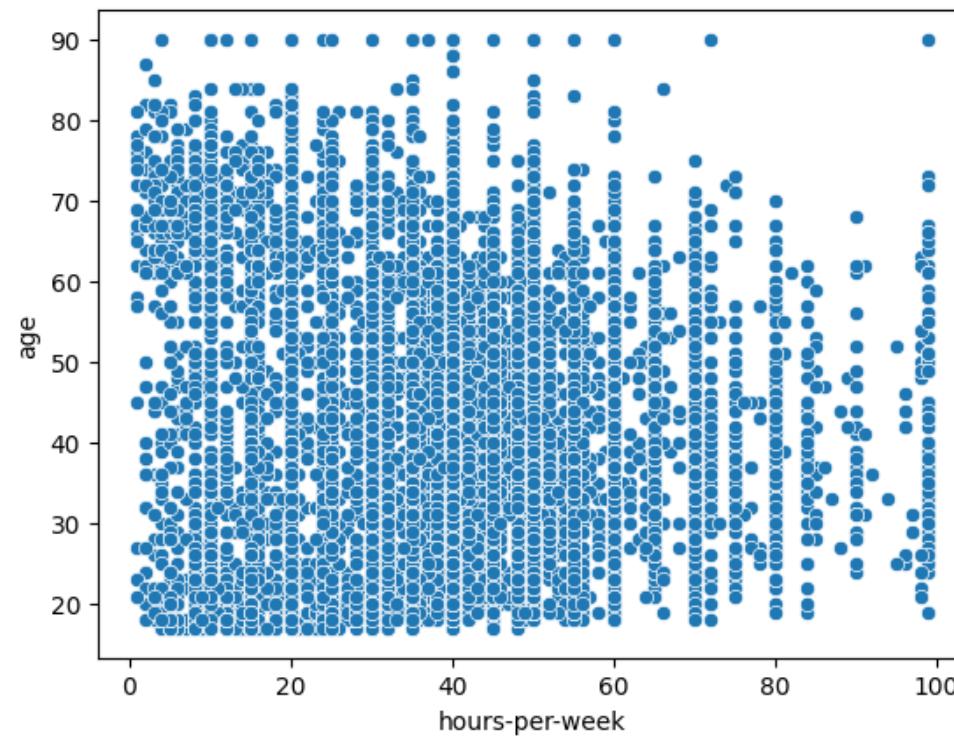
Category	Meaning / Explanation
Never-married	The person has <b>never legally married</b> .
Married-civ-spouse	Married <b>with a civil/legal spouse</b> (living together).
Married-AF-spouse	Married <b>with an Armed Forces spouse</b> (military).
Married-spouse-absent	Legally married but the <b>spouse is not present</b> (e.g., separated, working elsewhere).
Separated	Legally married but <b>currently separated</b> from the spouse.
Divorced	Legally <b>divorced</b> from a spouse.
Widowed	Spouse has <b>died</b> .

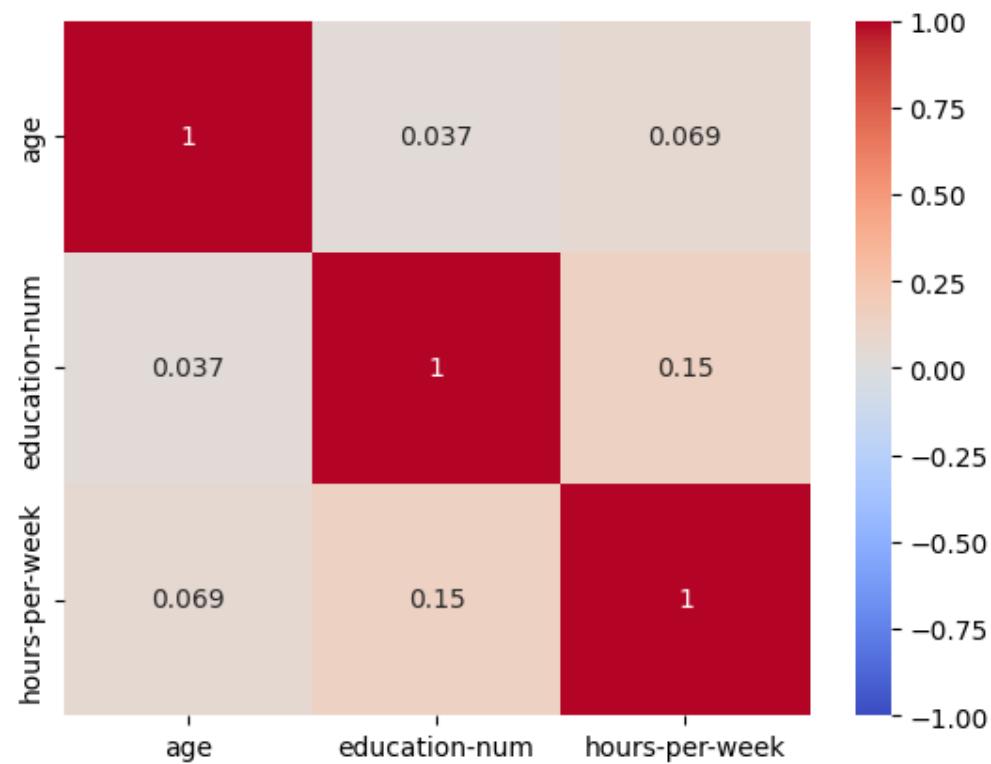


Search for additional domain information to attempt grouping



Bivariate analysis  
Numerical/Numerical





Nothing seems correlated

# Assignment 1

## *Deliverables*

1. Choose 2 datasets

The Kaggle site is a very interesting site to explore as it contains datasets for many tasks in data science and AI. You must select, from Kaggle, 2 datasets to explore.

For **Dataset 1**, it MUST be one among the 3 suggested below. The 3 suggestions are in 3 different domains: finance, healthcare, and mobile usage.

1. German Credit Dataset
  - o Size: 10 columns, 1000 rows
  - o Description: Financial data for credit scoring
  - o [Link](#)
2. Heart Failure Prediction Dataset
  - o Size: 12 columns, 918 rows
  - o Description: Health-related data for heart failure prediction
  - o [Link](#)
3. Mobile Device Usage Dataset
  - o Size: 11 columns, 700 rows
  - o Description: Technology/mobile behavior data
  - o [Link](#)

For **Dataset 2**, you must select a dataset that is NOT within the 3 suggested sets above. Furthermore, your second dataset MUST be in a domain different from the first dataset. I want you to explore 2 different domains. Another constraint is to find a dataset with a minimum of 10 columns so you have various features to explore.

2. Report the story of each dataset

The purpose of the report (written within a Jupyter Notebook) is to illustrate 10 insights that you found through your analysis for each dataset. In the notebook, you will be able to alternate between text and code, both required.

Your Jupyter Notebook should include:

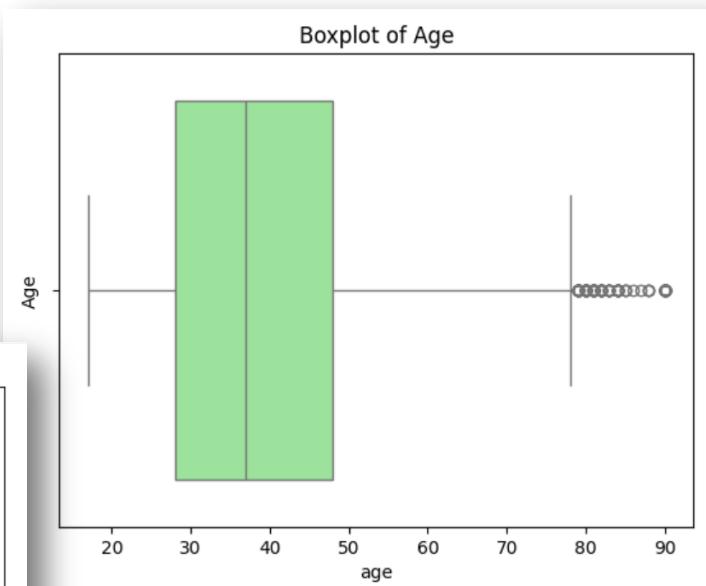
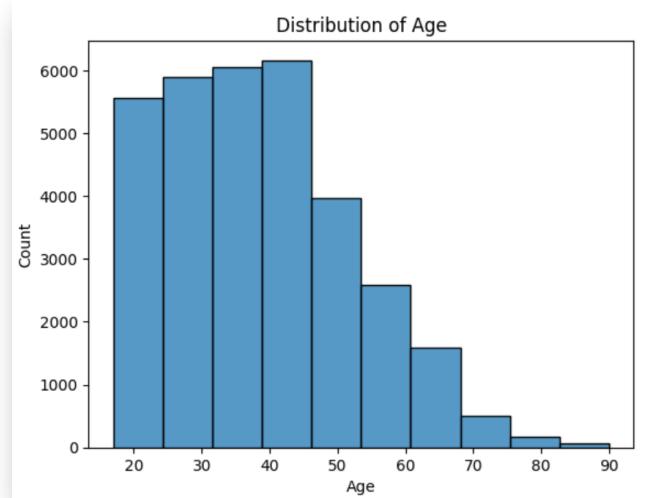
1. Group number, names and student numbers of group members
2. Introduction to provide the goal of the analysis/report and mention who the audience would be (imagine an audience who would read your report).
3. A description of the two datasets used (*see Dataset description requirement section*)
4. **A set of 10 insights for each dataset.** For EACH ONE:
  - a. State the insight in a single sentence.
  - b. Show supporting evidence from the data making sure that evidence is as self-explanatory as possible (graph with title, axis descriptions, etc)
  - c. Mention what type of analysis was done to arrive at such insight (*Analysis description requirement section*).
  - d. Show how this evidence was obtained (and how to reproduce it – provide code)
5. Conclusion
6. References

#### 4. Analysis description requirements

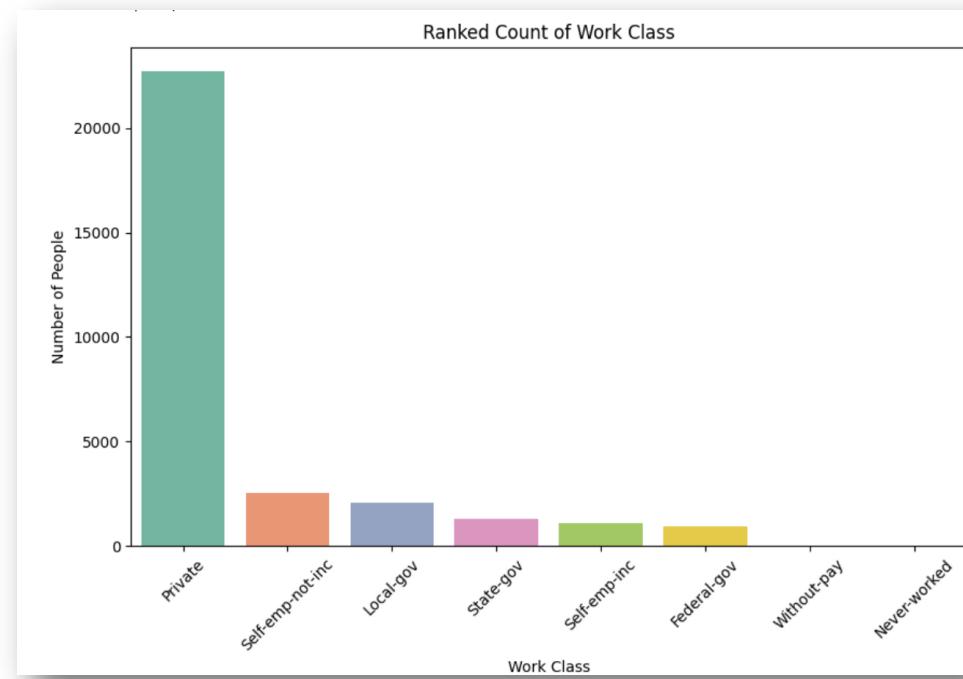
Given that we are in an **academic context**, I have additional requirements to make sure that you explore various types of analysis and visualization tools. Therefore, **among your 20 insights (10 for each dataset)**, you **MUST** have a diversity of supporting evidence including at least one of each of the analysis (written as r1 to r8) below for each dataset:

- a) Univariate analysis
  - a. Numerical data:
    - i. (r1) Simple histogram or boxplot for visualization of dispersion of a numerical variable
  - b. Categorical data:
    - i. (r2) Countplot for a category with multiple values to see the distribution among values.
    - ii. (r3) Grouped-Data countplot in which you group some values (and explain how you did the grouping).
- b) Bivariate analysis
  - a. Numerical/Categorical
    - i. (r4) Explore a numerical variable's distribution according to specific values of a categorical variable.
  - b. Categorical/Categorical
    - i. (r5) Comparing categories with 2 values
    - ii. (r6) Comparing categories with more than 2 values for which you set the order (e.g. increasing counts, or alphabetical order)
  - c. Categorical/Numerical
    - i. (r7) Looking at a categorical variable's distribution among bins of a numerical feature. You should create the bins with specific values.
  - d. Numerical/Numerical
    - i. (r8) Use the scatterplot to highlight correlation.

R1 – Univariate Numerical



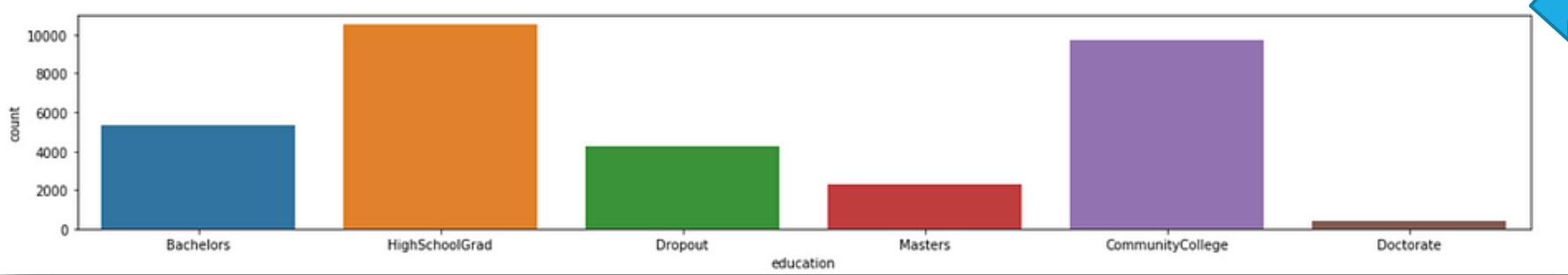
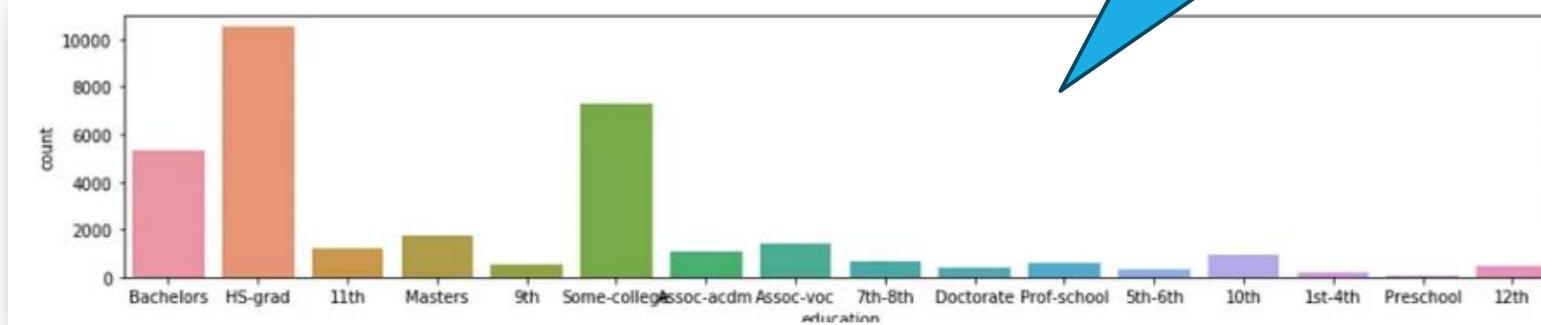
R2 – Univariate Categorical



### R3 – Univariate Categorical - Grouped



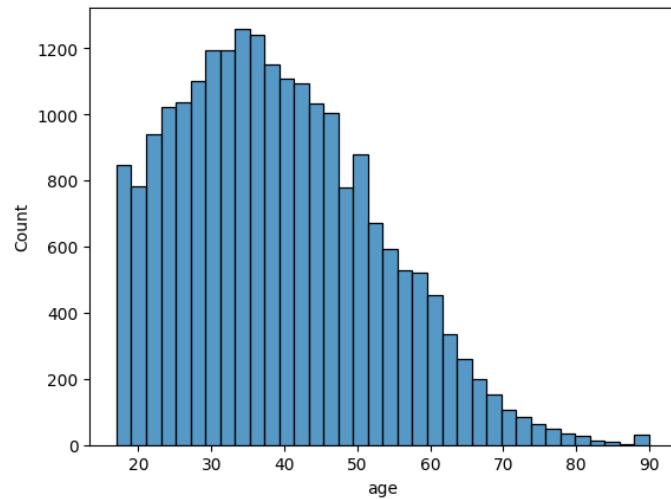
Provide domain justification  
for the grouping



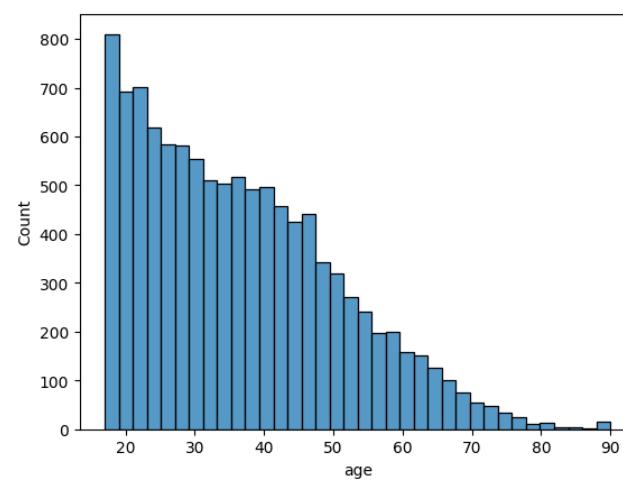
R4 – Bivariate  
Numerical/Categorical



Age distribution of Male



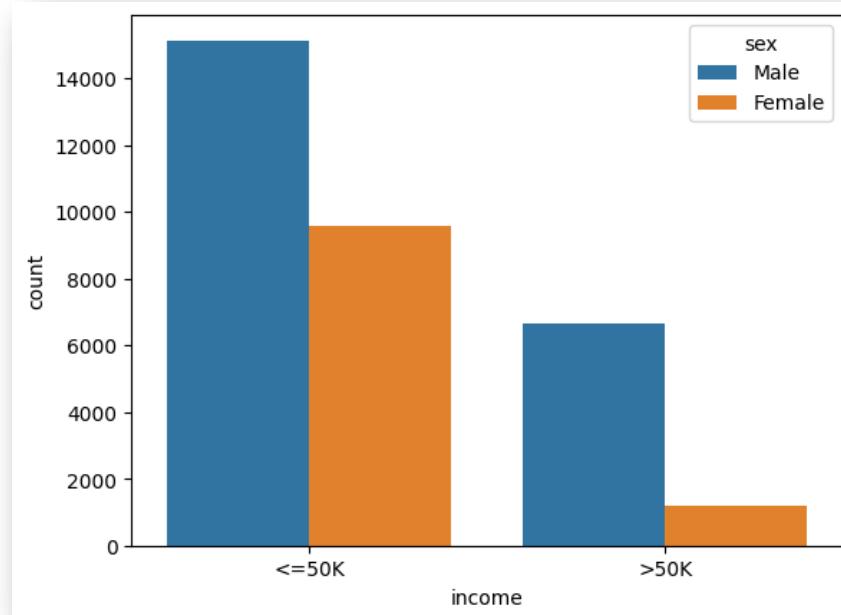
Age distribution of Female



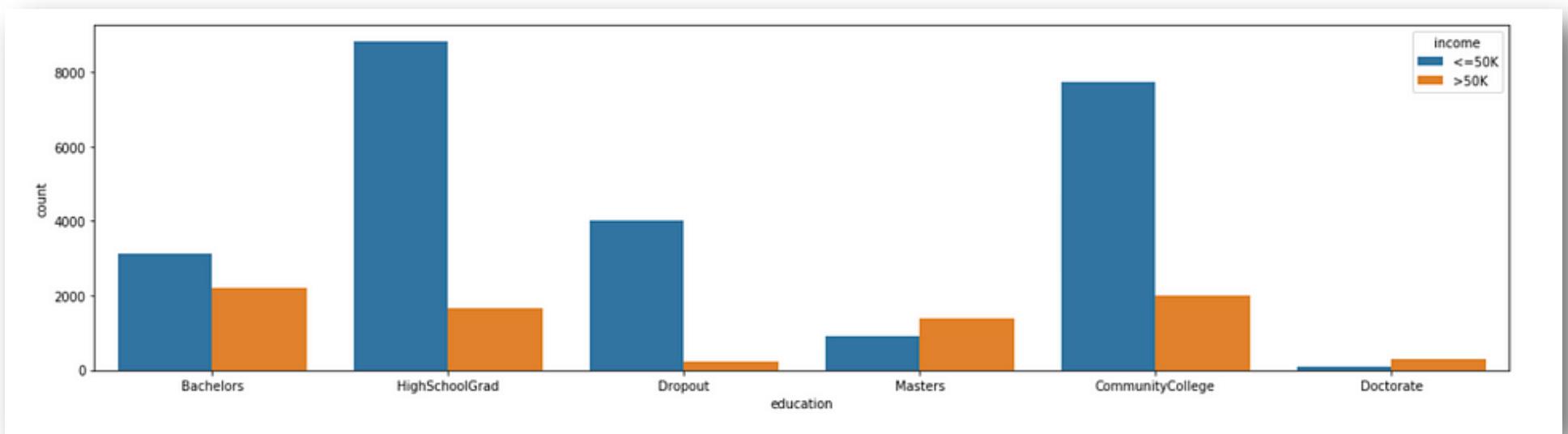
R5 – Bivariate  
Categorical/Categorical  
2 values per category



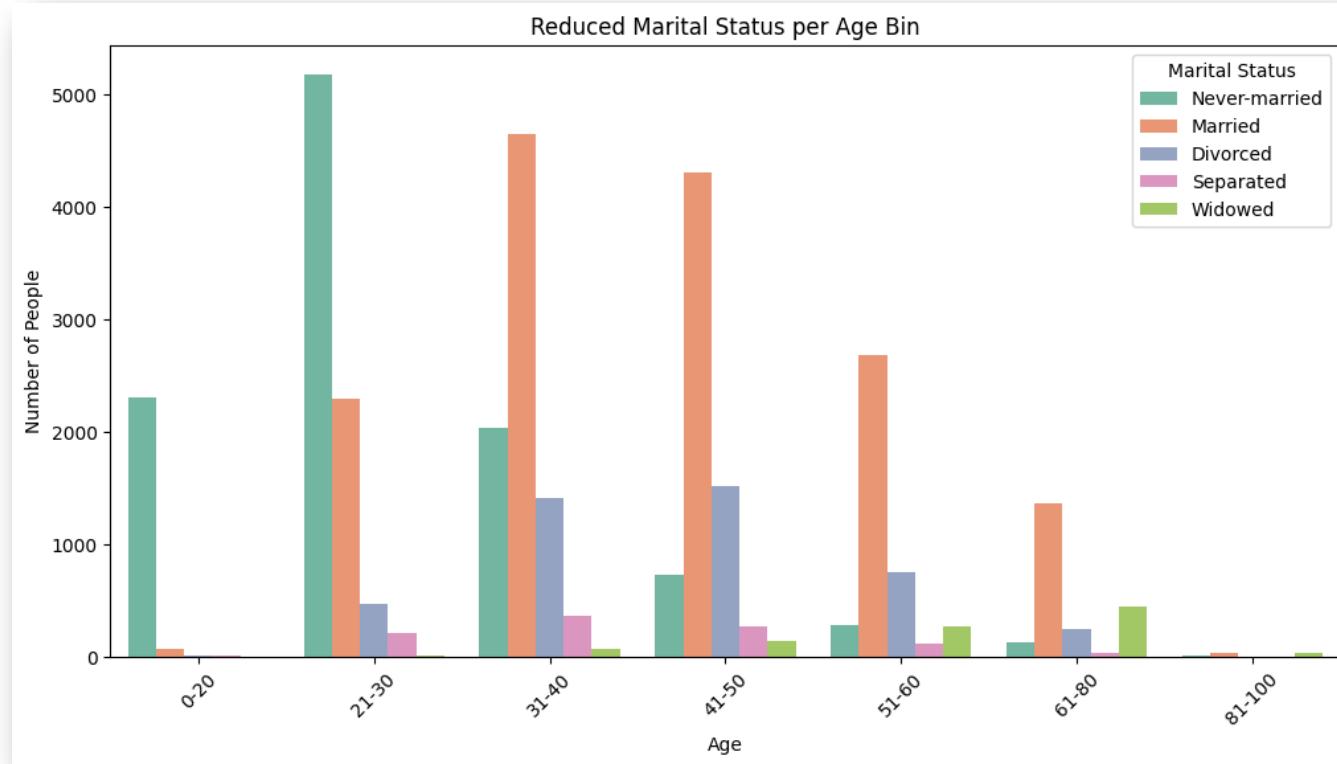
Explain why this view  
(and not flipped)



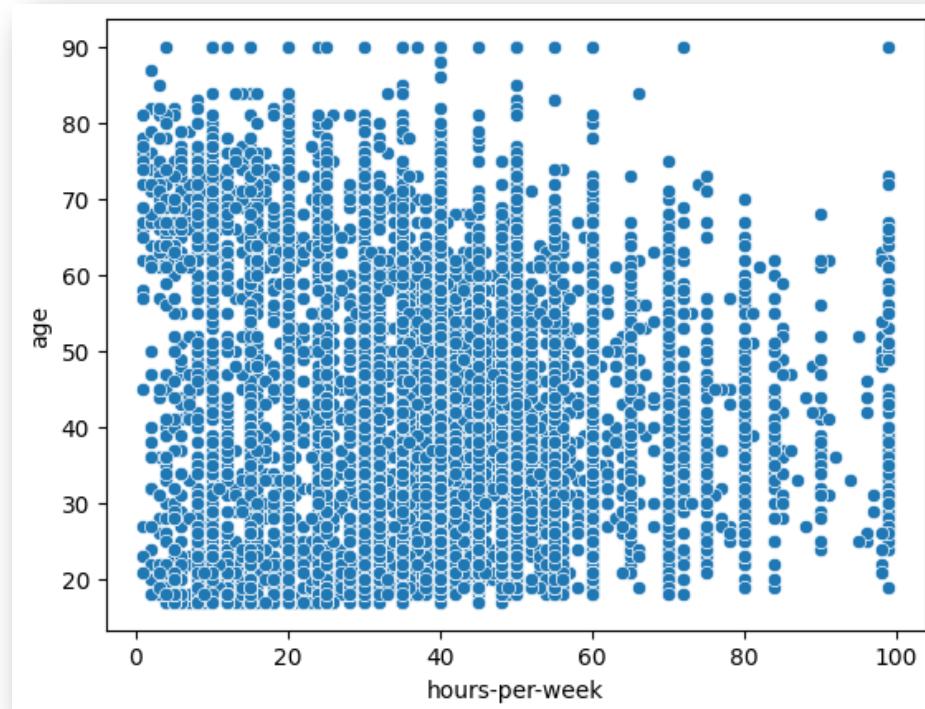
R6 – Bivariate  
Categorical/Categorical  
More than 2 values for at least  
one category



R7 – Bivariate  
Categorical/Numerical



R8 – Bivariate  
Numerical/Numerical





Example of insights

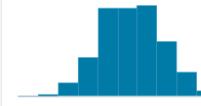
Example 1

## Billionaires Statistics Dataset (2023)

Exploring the Global Landscape of Success



<https://www.kaggle.com/datasets/nelgiriyewithana/billionaires-statistics-dataset/data>

# rank	# finalWorth	▲ category	▲ personName	# age
The ranking of the billionaire in terms of wealth.	The final net worth of the billionaire in U.S. dollars.	The category or industry in which the billionaire's business operates.	The full name of the billionaire.	The age of the billionaire.
		Finance & Invest... 14% Manufacturing 12% Other (1944) 74%	<b>2638</b> unique values	
1	211000	Fashion & Retail	Bernard Arnault & family	74
2	180000	Automotive	Elon Musk	51
3	114000	Technology	Jeff Bezos	59
4	107000	Technology	Larry Ellison	78
5	106000	Finance & Investments	Warren Buffett	92
6	104000	Technology	Bill Gates	67
7	94500	Media & Entertainment	Michael Bloomberg	81
8	93000	Telecom	Carlos Slim Helu & family	83
9	83400	Diversified	Mukesh Ambani	65
10	80700	Technology	Steve Ballmer	67
11	80500	Fashion & Retail	Francoise Bettencourt Meyers & family	69
12	79200	Technology	Larry Page	50

**Key Features**

- **rank:** The ranking of the billionaire in terms of wealth.
- **finalWorth:** The final net worth of the billionaire in U.S. dollars.
- **category:** The category or industry in which the billionaire's business operates.
- **personName:** The full name of the billionaire.
- **age:** The age of the billionaire.
- **country:** The country in which the billionaire resides.
- **city:** The city in which the billionaire resides.
- **source:** The source of the billionaire's wealth.
- **industries:** The industries associated with the billionaire's business interests.
- **countryOfCitizenship:** The country of citizenship of the billionaire.
- **organization:** The name of the organization or company associated with the billionaire.
- **selfMade:** Indicates whether the billionaire is self-made (True/False).
- **status:** "D" represents self-made billionaires (Founders/Entrepreneurs) and "U" indicates inherited or unearned wealth.
- **gender:** The gender of the billionaire.
- **birthDate:** The birthdate of the billionaire.
- **lastName:** The last name of the billionaire.
- **firstName:** The first name of the billionaire.
- **title:** The title or honorific of the billionaire.
- **date:** The date of data collection.
- **state:** The state in which the billionaire resides.
- **residenceStateRegion:** The region or state of residence of the billionaire.
- **birthYear:** The birth year of the billionaire.
- **birthMonth:** The birth month of the billionaire.
- **birthDay:** The birth day of the billionaire.
- **cpi\_country:** Consumer Price Index (CPI) for the billionaire's country.
- **cpi\_change\_country:** CPI change for the billionaire's country.

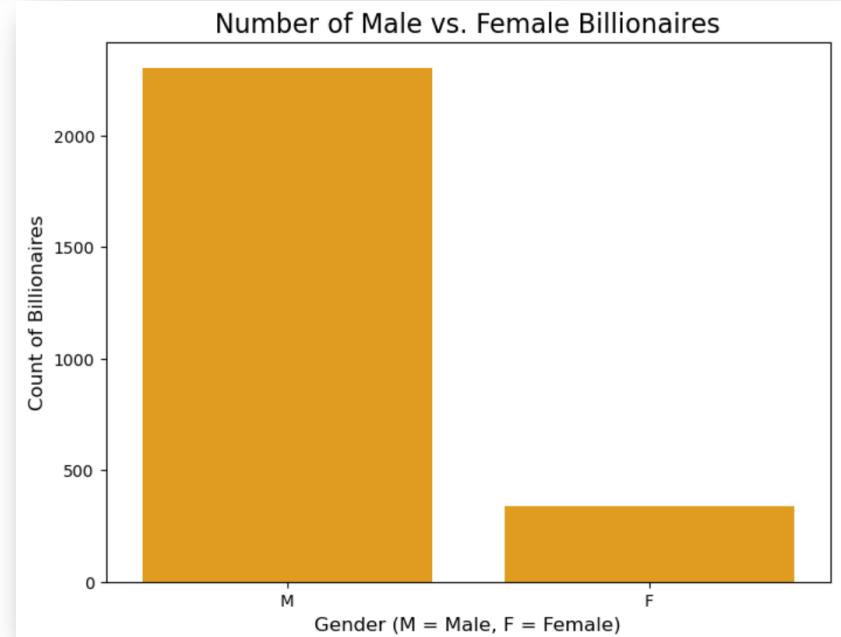
No need to use everything...

## Insight 1

**Insight:** The large majority of billionaires are male, with women making up a small portion of the total number of billionaires.

**Analysis Type:** Countplot for a category with multiple values

**How Evidence Was Obtained:** Using the countplot function on the gender column to count the number of occurrences of male and female billionaires in the dataset. With this we can visualize the disparity between the genders.

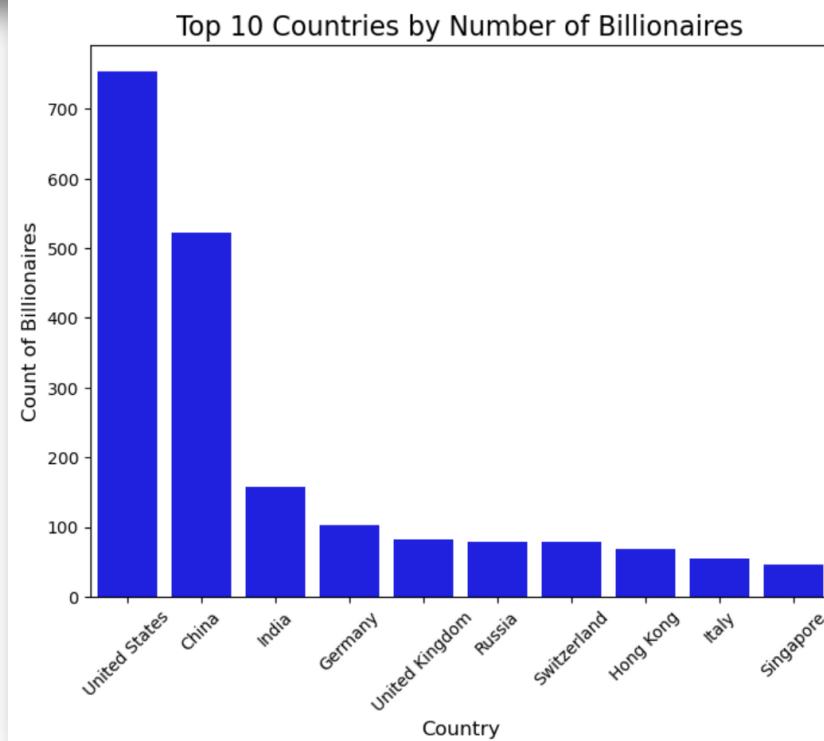


### Insight 3

**Insight:** The U.S. and China dominate the billionaire list, with vastly more billionaires than any other country.

**Analysis Type:** Countplot for a category with multiple values

**How Evidence Was Obtained:** We first get the count of the number of billionaires in each country with the value\_counts() function. Then, we get the top 10 using the head(10) function. We can then visualize in descending order the number of billionaires by country.



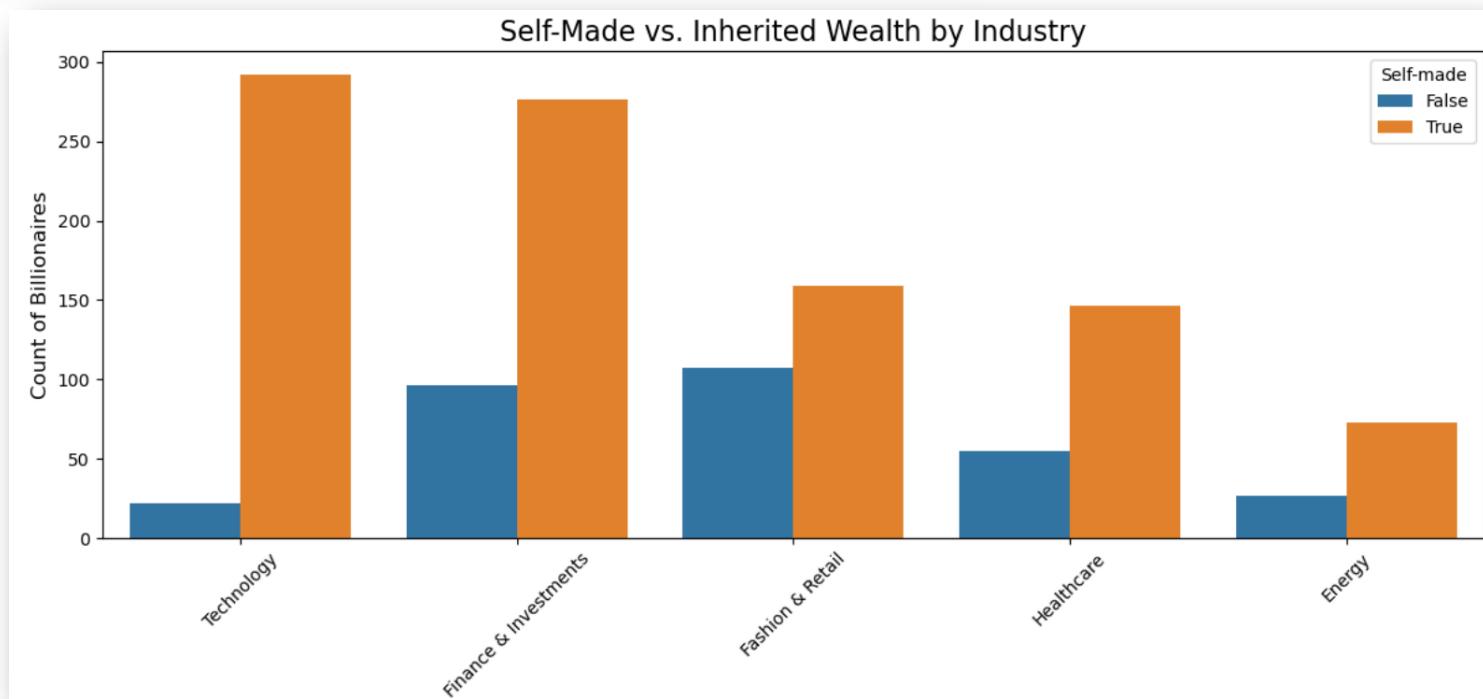
## Insight 10

**Insight:** The technology industry creates the most self-made billionaires because it has a low barrier of entry and it is easy to grow big quickly.

**Analysis Type:** Compare category with 2 values against category with more than 2 values

**How Evidence Was Obtained:** This countplot compares the self-made versus inherited status for billionaires across different industries, focusing on technology and other major sectors. It shows the dominance of self-made billionaires in the tech industry.

That's all in the data?



Example 2

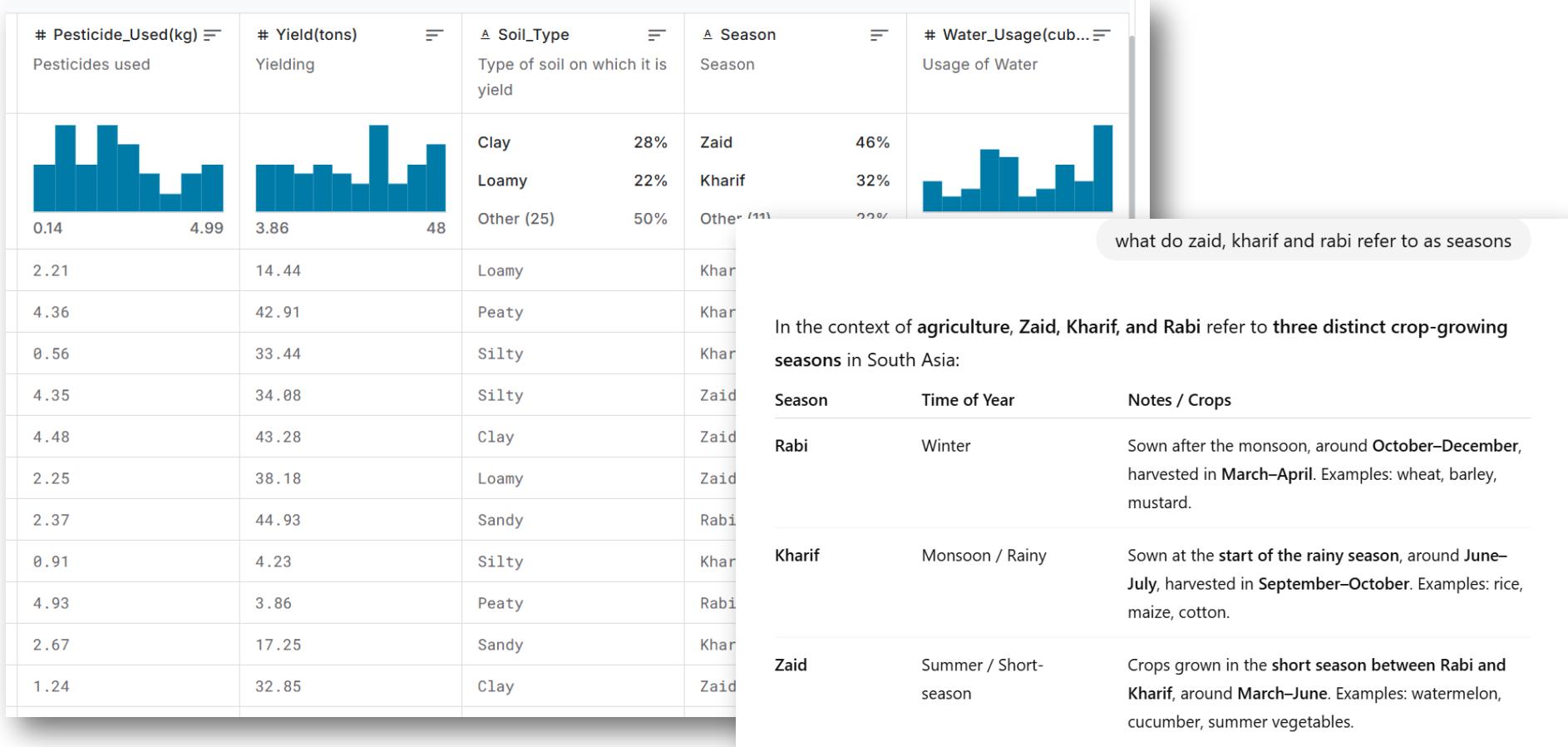
## Agriculture and Farming Dataset

Insights and Trends in Agriculture and Farming: A Comprehensive Data Analysis



<https://www.kaggle.com/datasets/bhadramohit/agriculture-and-farming-dataset>

<b>▲ Farm_ID</b>	<b>▲ Crop_Type</b>	<b># Farm_Area(acres)</b>	<b>▲ Irrigation_Type</b>	<b># Fertilizer_Used(to...)</b>	<b># Pesticide_</b>
FarmID	Type of Crop used	Area of Farm land	Type of Irrigation applied	Fertilizers used	Pesticides us
<b>50</b> unique values	Cotton	14%	Drip	30%	0.14 - 0.62 Count: 5
	Barley	14%	Flood	26%	0.14
	Other (36)	72%	Other (22)	44%	0.5 - 9.96
F001	Cotton	329.4	Sprinkler	8.14	2.21
F002	Carrot	18.67	Manual	4.77	4.36
F003	Sugarcane	306.03	Flood	2.91	0.56
F004	Tomato	380.21	Rain-fed	3.32	4.35
F005	Tomato	135.56	Sprinkler	8.33	4.48
F006	Sugarcane	12.5	Sprinkler	6.42	2.25
F007	Soybean	360.06	Drip	1.83	2.37
F008	Rice	464.6	Drip	5.18	0.91



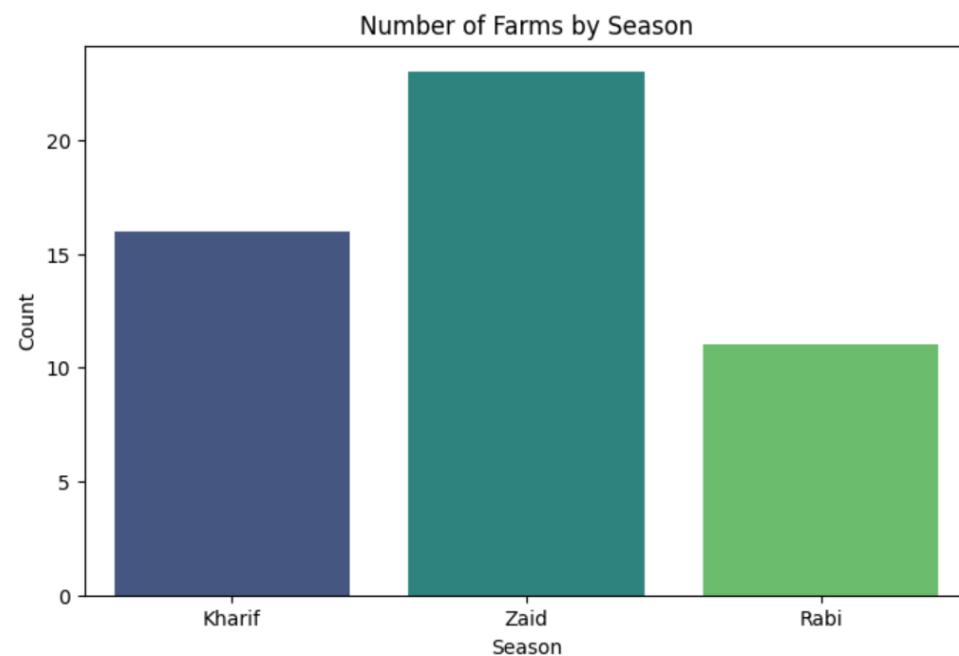
**Insight 2:** Zaid is the most common farming season.

**Supporting Evidence:** A countplot of different seasons shows that Kharif dominates.

**Analysis Type:** Univariate Analysis (Categorical) → Countplot.

Be careful with your explanation... there is a contradiction here.

Be more specific in your title (active farms?)



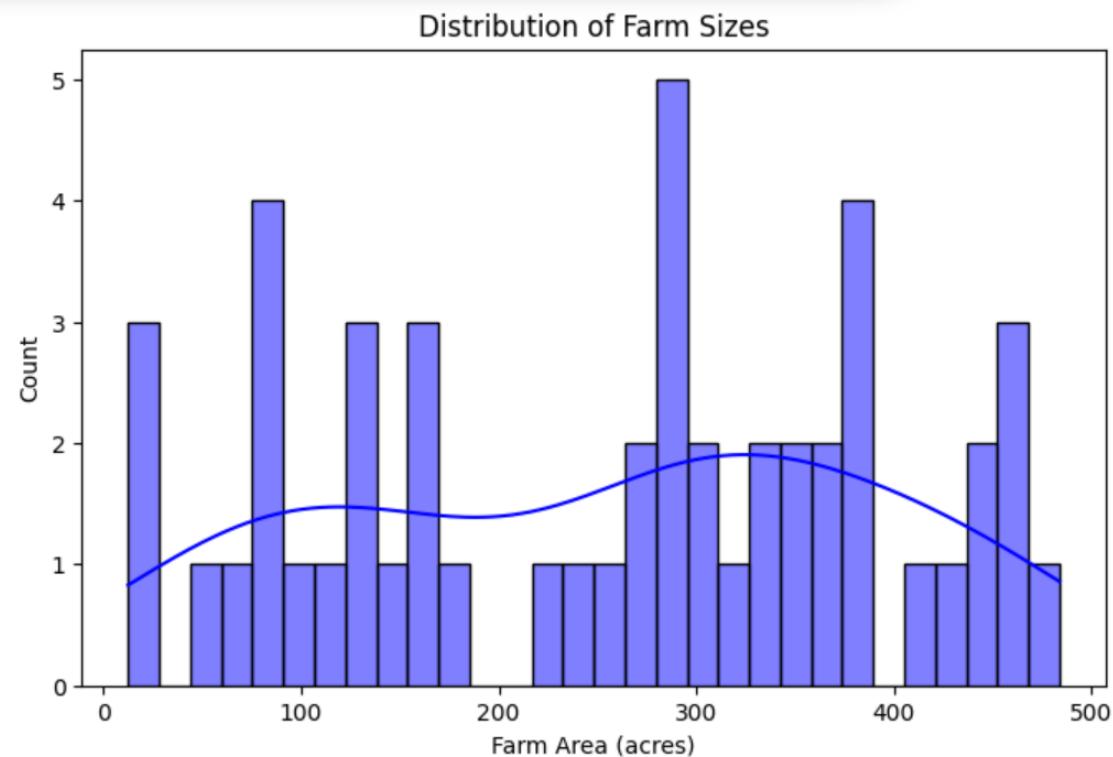
**Insight 1:** The distribution of farm sizes is right-skewed, with many small farms and fewer large farms.

**Supporting Evidence:** A histogram of farm sizes shows most farms are smaller, with a few much larger ones.

**Analysis Type:** Univariate Analysis (Numerical) → Histogram

It's actually left-skewed  
(long tail on the left)

Not that obvious...  
Perhaps say something  
else?

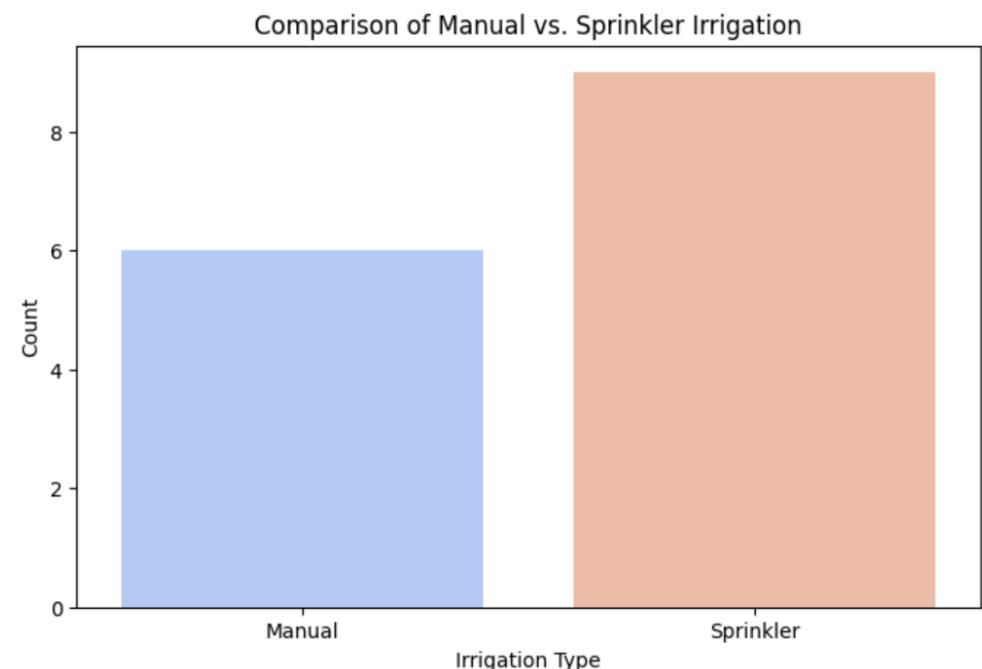


**Insight 4:** Sprinkler irrigation is more common than manual irrigation.

**Supporting Evidence:** A comparison of two irrigation types.

**Analysis Type:** Bivariate Analysis (Categorical/Categorical) → Compare Two Categories

This is a UNIVARIATE analysis... two values of the same variable

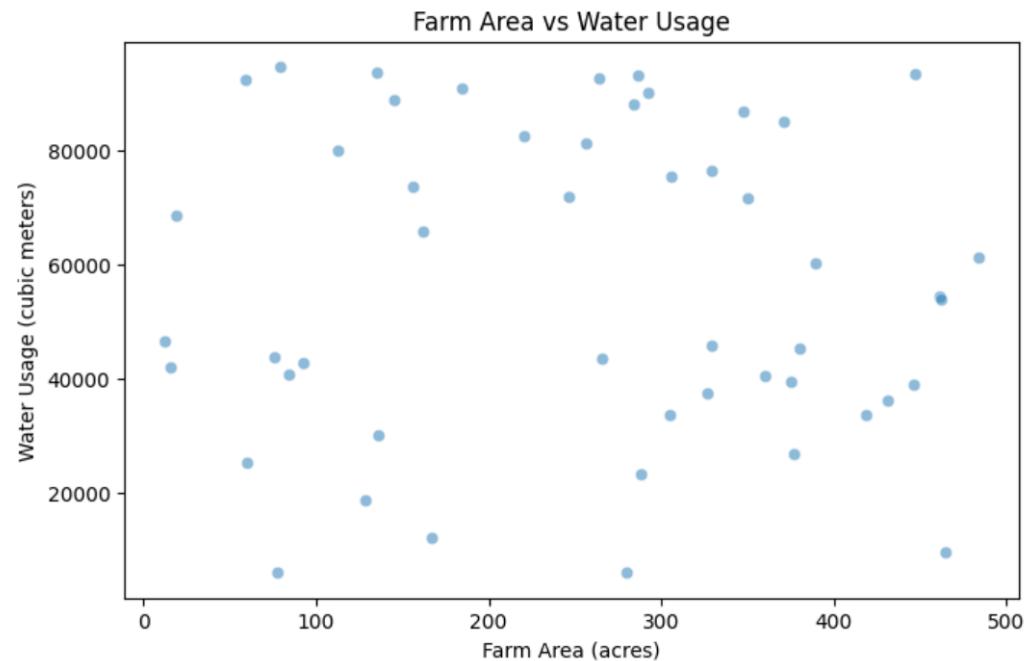


**Insight 6:** There is a moderate correlation between farm size and water usage.

**Supporting Evidence:** A scatter plot of farm area vs. water usage shows an increasing trend.

**Analysis Type:** Bivariate Analysis (Numerical/Numerical) → Correlation Scatterplot

Really?

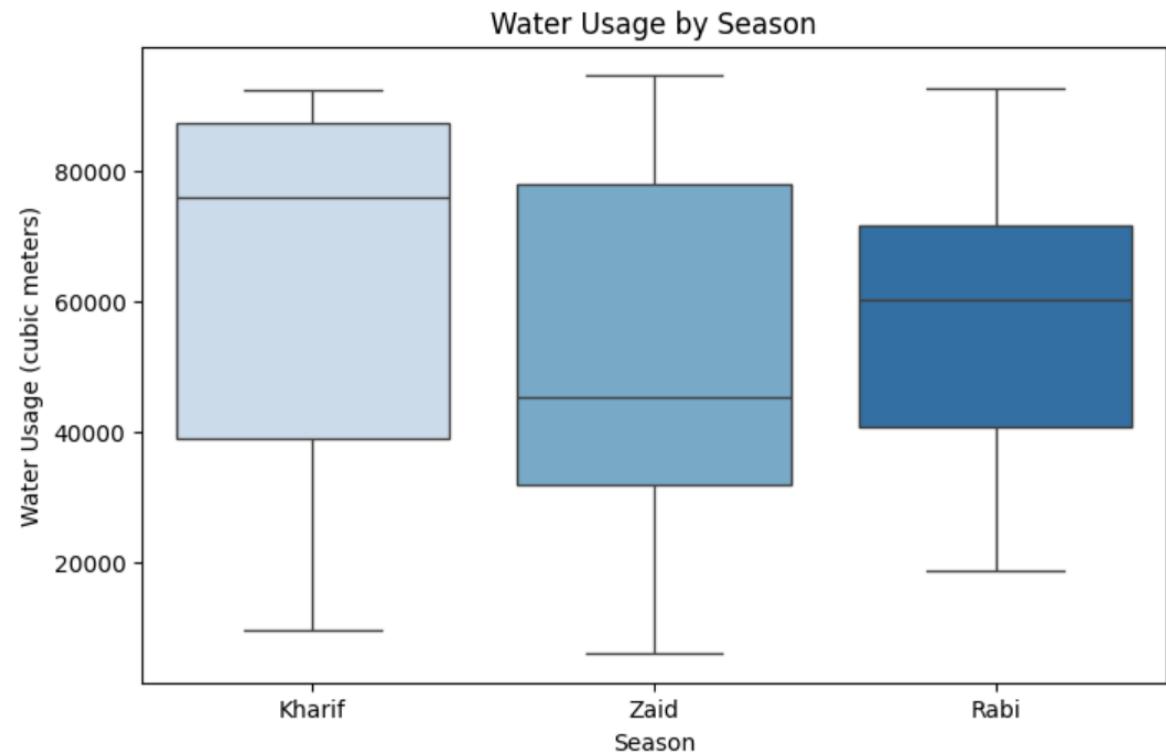


**Insight 9:** Farms in Kharif season use significantly more water than those in other seasons.

**Supporting Evidence:** A box plot of water usage grouped by season.

**Analysis Type:** Bivariate Analysis (Numerical/Categorical) → Boxplots per value

OK...



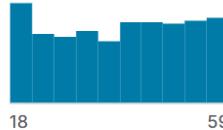
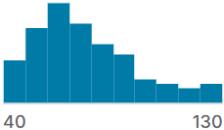
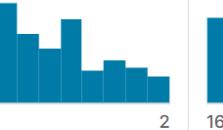
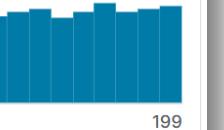
### Example 3

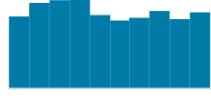
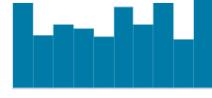
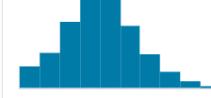
## Gym Members Exercise Dataset

Analyzing Fitness Patterns and Performance Across Diverse Gym Experience Levels



<https://www.kaggle.com/datasets/valakhorasani/gym-members-exercise-dataset>

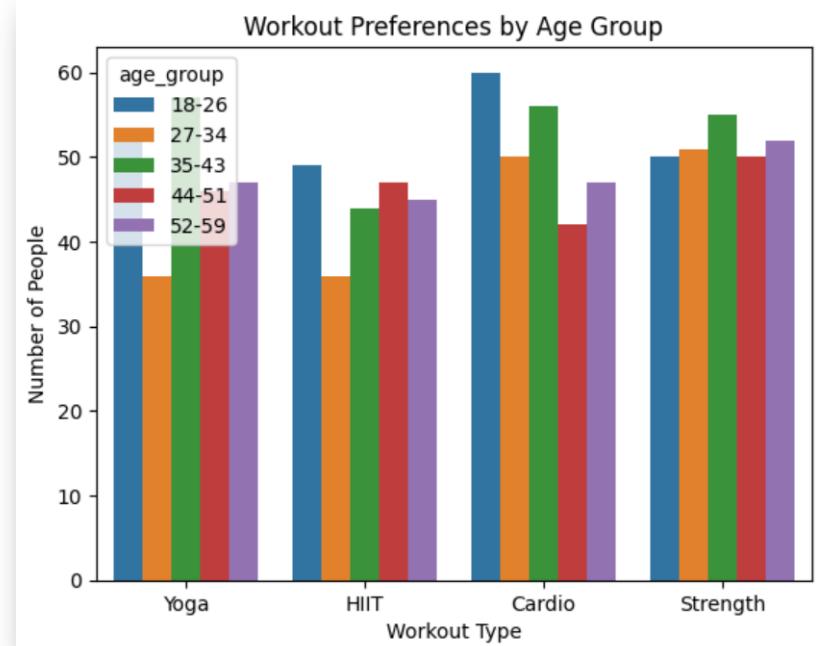
# Age	Gender	# Weight (kg)	# Height (m)	# Max_BPM
The age of the gym member (in years)	The gender of the gym member (Male or Female)	Member's body weight measured in kilograms	Member's height measured in meters	Maximum beats per minute (BPM) recorded during a workout session
	Male 53% Female 47%			
18	Male	88.3	1.71	180
56	Female	74.9	1.53	179
46	Female	68.1	1.66	167
32	Male	53.2	1.7	190
25	Male	46.1	1.79	188
38	Female	58.0	1.68	168
36	Male	70.3	1.72	174
40	Female	69.7	1.51	189
28	Male	121.7	1.94	185
28	Male	101.8	1.84	169
41	Male	120.8	1.67	188
53	Male	51.7	1.7	175

# Avg_BPM	# Resting_BPM	# Session_Duration...	# Calories_Burned	▲ Workout_Type
Average beats per minute (BPM) during the workout session	Member's heart rate (BPM) before starting the workout, at rest	The total time spent during a workout session, measured in hours	The number of calories burned during the workout session	The type of workout performed (e.g., Cardio, Strength, Yoga, HIIT)
 120	 50	 0.5	 303	Strength 27% Cardio 26% Other (460) 47%
157	60	1.69	1313.0	Yoga
151	66	1.3	883.0	HIIT
122	54	1.11	677.0	Cardio
164	56	0.59	532.0	Strength
158	68	0.64	556.0	Strength
156	74	1.59	1116.0	HIIT
169	73	1.49	1385.0	Cardio
141	64	1.27	895.0	Cardio
127	52	1.03	719.0	Strength
136	64	1.08	808.0	Cardio
146	54	0.82	593.0	HIIT
152	72	1.15	865.0	HIIT

Insight:

All age groups like doing Strength exercises equally, while Yoga and HIIT remains less popular, especially for people in age 27-34; and Cardio is more popular among younger people (18-43)

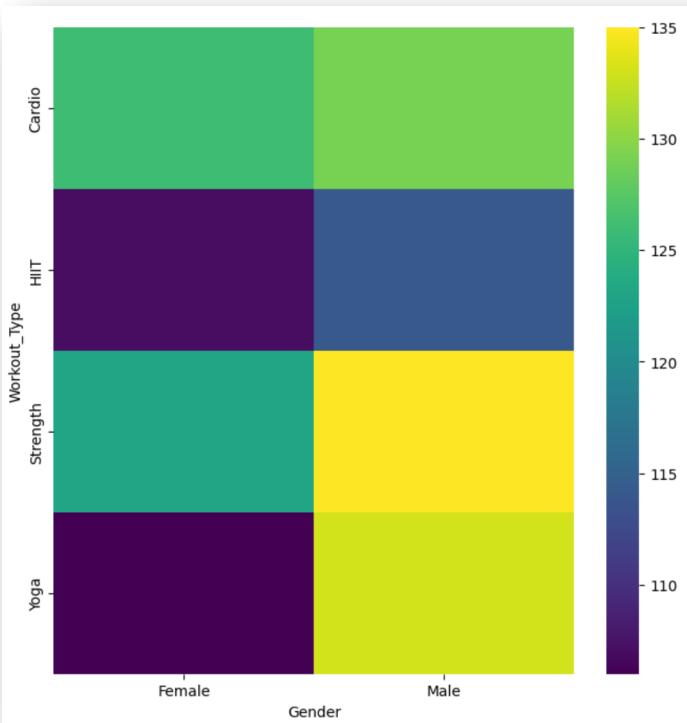
Please be careful  
with hiding the graph



Insight:

Yoga and Strength is most popular for Men, while Woman prefer doing Cardio

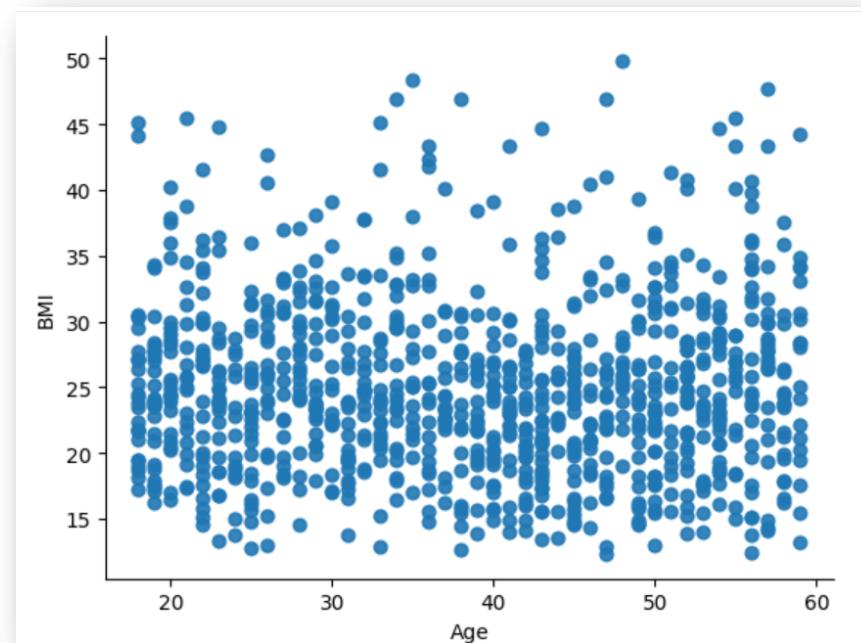
OK if you want to do something different, but still, the insight should be interpreted easily from the evidence provided.



Insight:

BMI level, an indication for obesity and healthiness, remains quite evenly distributed among all age groups, with people have high BMI slightly tends to be in 18-24 and 55-60.

If including additional info...  
make sure there are  
references for it





## EXPLORATORY DATA ANALYSIS

- Case Study 2 – Census data
- Assignment 1 Examples