
Workshop on Data Science

Contents

- Introduction to Data Science
- Exploratory data analysis (EDA)
- Data Visualization
- Hands-on Power BI
- Python for Data Science

Why Data Science?

Data is everywhere

It’s a digital era

What is Data Science?

Data

&

Science

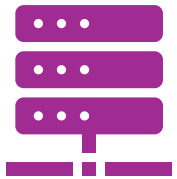
Data science is a set of techniques and tools that help extract knowledge from data to make accurate decisions

Components of Data Science

Probability and statistics

Linear Algebra

Machine learning



Computer Science

Real-world applications

- Personalized experiences
- Healthcare
- Transportation
- Climate change
- Agriculture
- Sports
- Public policy
- Civic engagement
- Social challenges
- Supply chains
- Targeted advertising campaigns
- Fraud detection
- Risk assessment, and
- Algorithmic trading

Foundational Questions for Deriving Insights

What kind of problem needs to be solved? Do the documents need to be classified into predefined categories or clustered?

Can we quantify the caliber of the data being considered for analysis? Is its quality up to the mark?

Is the data ready for analysis? Or are any transformations required?

Is the available data sufficient to solve the given problem?

What are the potential sources of the data?

How many observations are there in the data?

How many attributes/features are there in the data?

Foundational Questions for Deriving Insights

Are there any missing values in the data?

Is there any correlation between the different features of the data?

Can any pattern be identified in the set of features?

Which type of analysis is required: Descriptive / Predictive / Prescriptive?

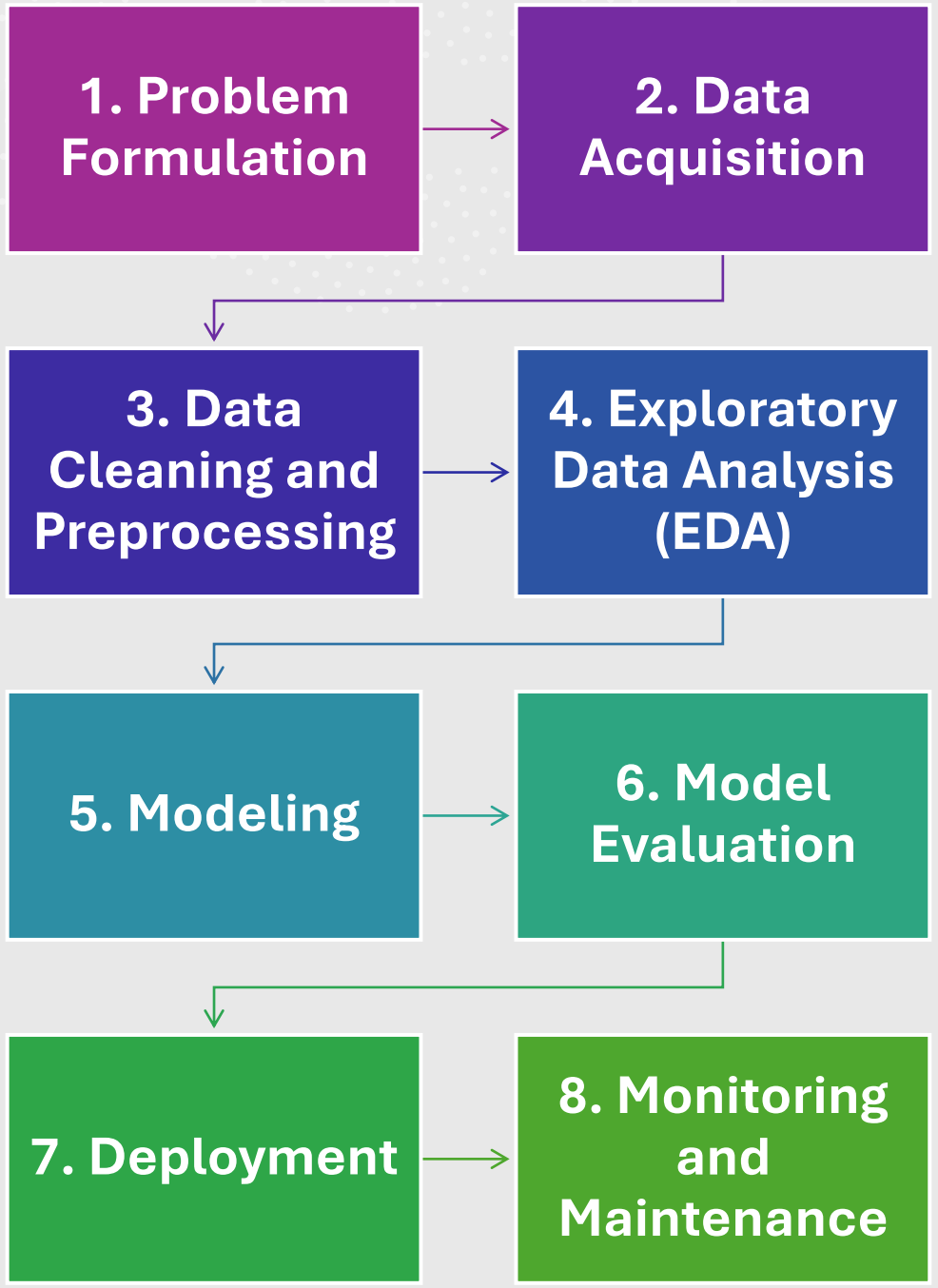
Does the result need optimization?

Which tool may be used during data orientation and alignment?

And there may be many more such questions to ask.

Data Science is about handling all such questions!

Life cycle of Data Science



Exploratory Data Analysis

EDA

A set of procedures for examining data

Summarize the data with the help of descriptive and graphical tools

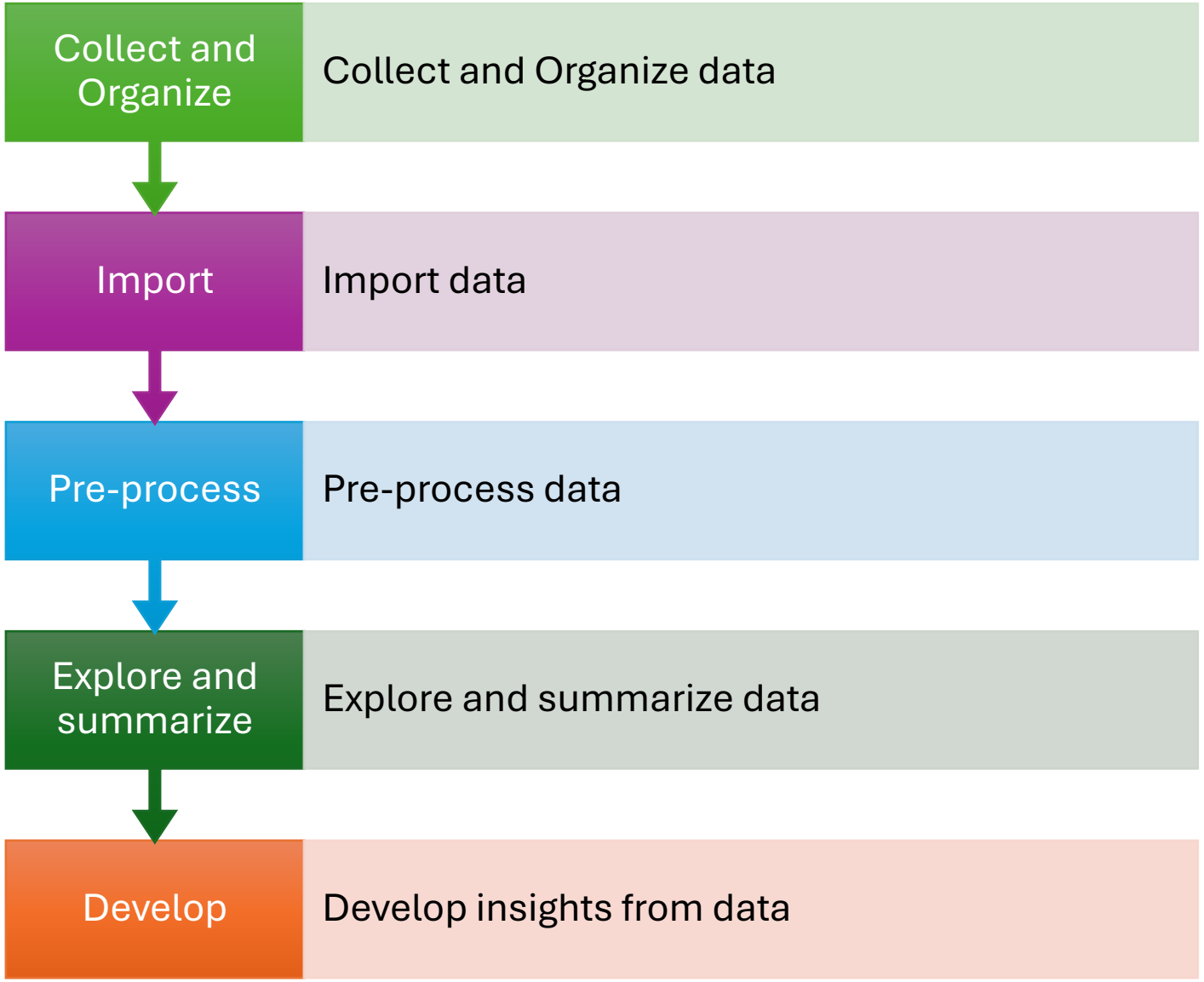
“No assumption” study of the data

Determine the relationship between variables

Handle missing values

To draw a meaning from the data

Steps involved



Steps involved

Collect and Organize data

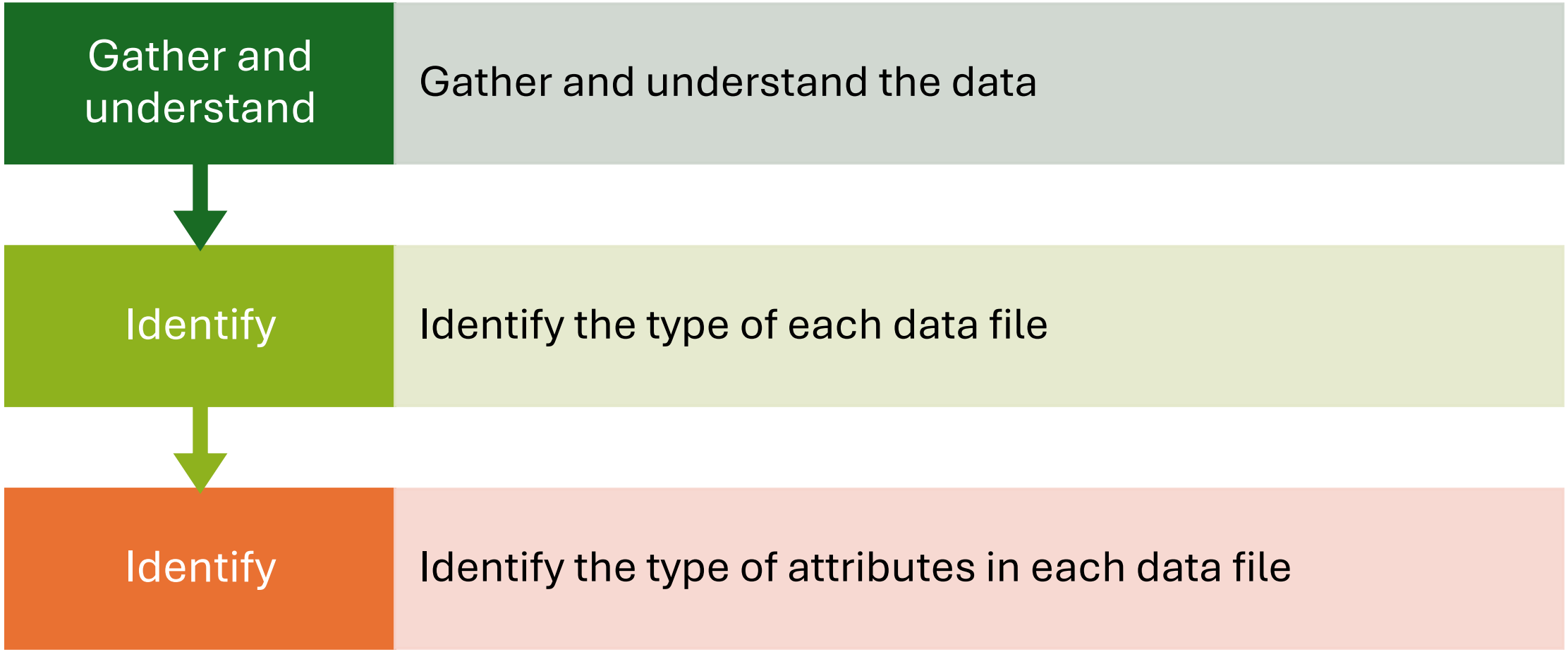
Import data

Pre-process data

Explore and summarize data

Develop insights from data

1. Collect and Organize data



Gather relevant data

Data

For example: Restaurant business
consumer_survey.xls
dining_preferences.xls
Feedback.txt
RD_Insert_Script.txt
restaurant_cuisine.accdb
restaurant_cuisine.xls
restaurant_parking,xls
reviews.xml
user_rating.xls

Source: Infosys Springboard

Classify the data

Structured data

Data in DBs, Spreadsheets.....

Semi-structured data

Emails, xml file, html file.....

Unstructured data

Images, Videos, text files....

eg: Categories Restaurant data

consumer_survey.xls
dining_preferences.xls
RD_Insert_Script.txt
restaurant_cuisine.accdb
restaurant_cuisine.xls
restaurant_parking,xls
user_rating.xls
reviews.xml
Feedback.txt

Structured data

- consumer_survey.xls
- dining_preferences.xls
- RD_Insert_Script.txt
- restaurant_cuisine.accdb
- restaurant_cuisine.xls
- restaurant_parking,xls
- user_rating.xls

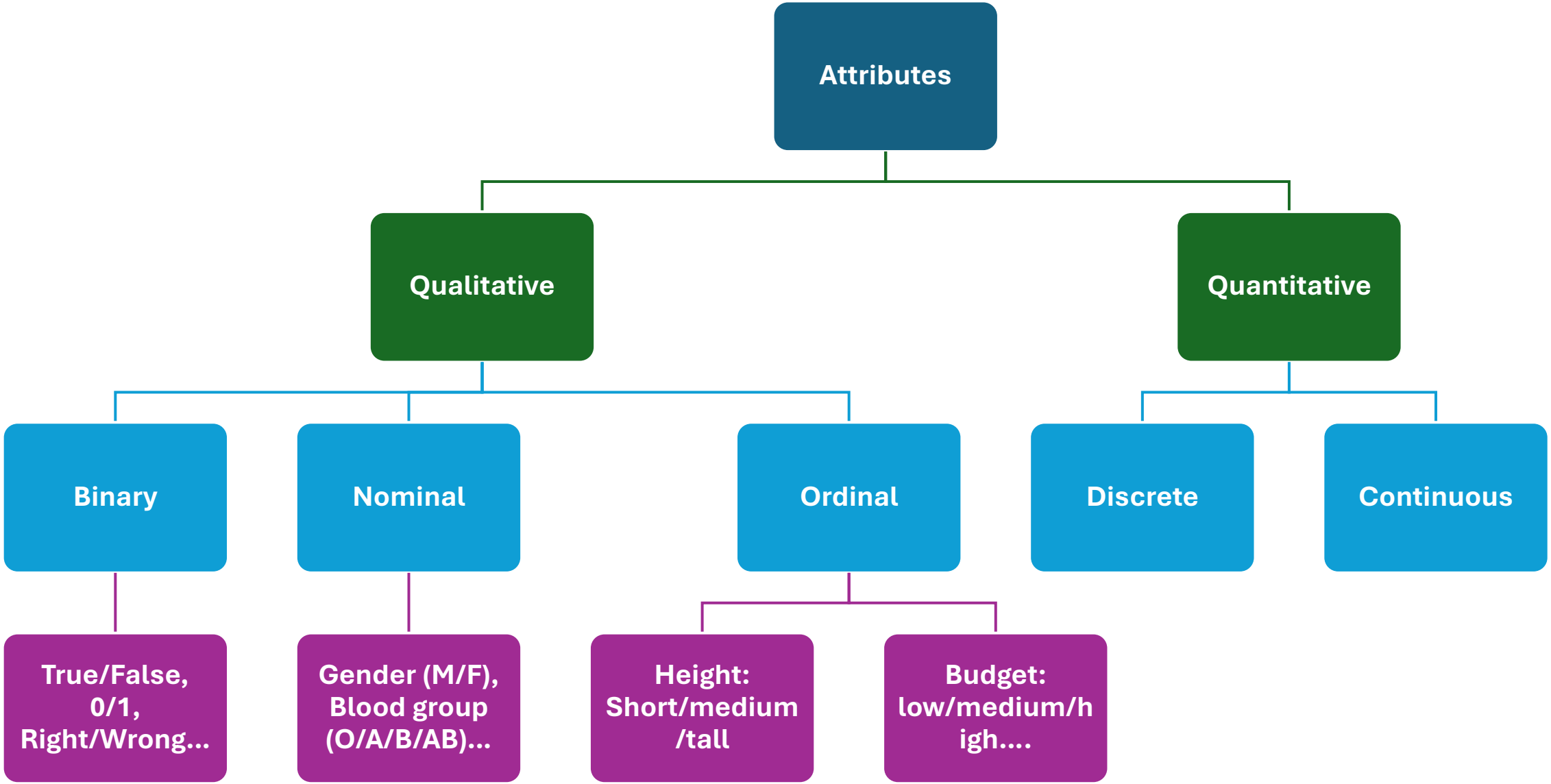
Semi-structured data

- reviews.xml

Unstructured data

- Feedback.txt

Identifying types of attributes in structured data



Classifying attributes of a structured dataset Consumer Survey.xls file

Consumer Survey.xls

• Gender	• Male/Female
	• Qualitative (Binary)
• Smoker	• True/False
	• Qualitative (Binary)
• marital_status	• Single/Married/Widow
	• Qualitative (Nominal)
• Income	• low/medium/high
	• Qualitative (Ordinal)
• FRVPM-Freq of restaurant visits per month	• Quantitative (Discrete)
• AERPM-Avg expense in Restaurants per month	• Quantitative (Continuous)

Steps involved

Collect and Organize data

Import data

Pre-process data

Explore and summarize data

Develop insights from data

Steps involved

Collect and Organize data

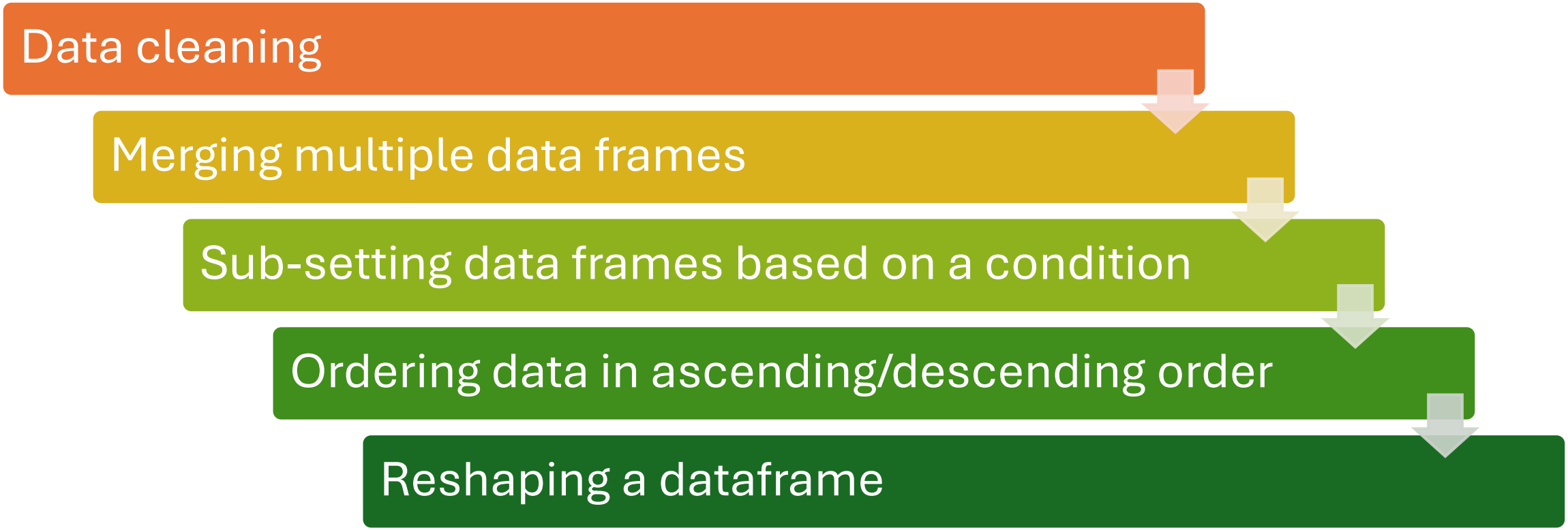
Import data

Pre-process data

Explore and summarize data

Develop insights from data

Pre-process data



Steps involved

Collect and Organize data

Import data

Pre-process data

Explore and summarize data

Develop insights from data

Steps involved

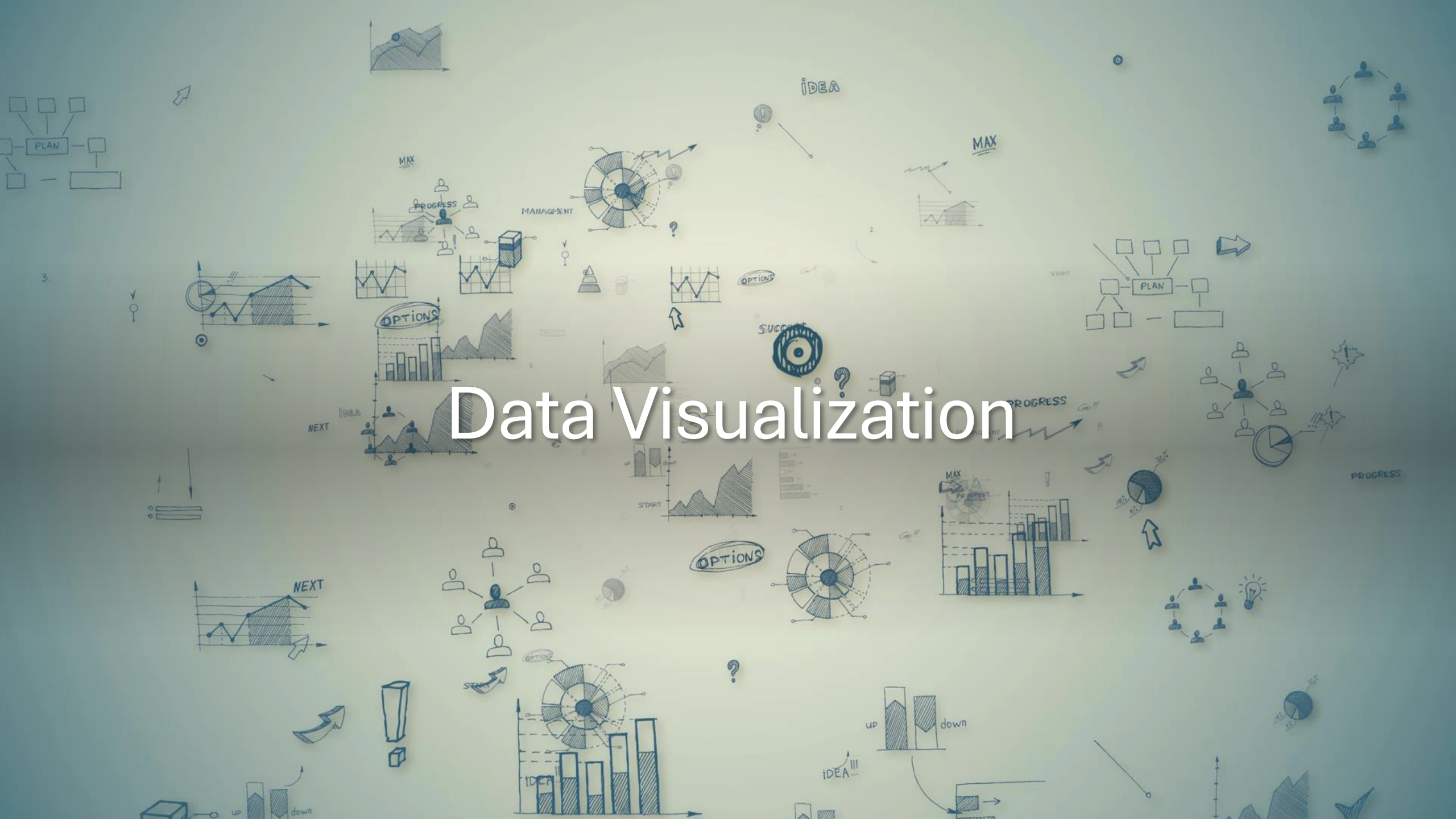
Collect and Organize data

Import data

Pre-process data

Explore and summarize data

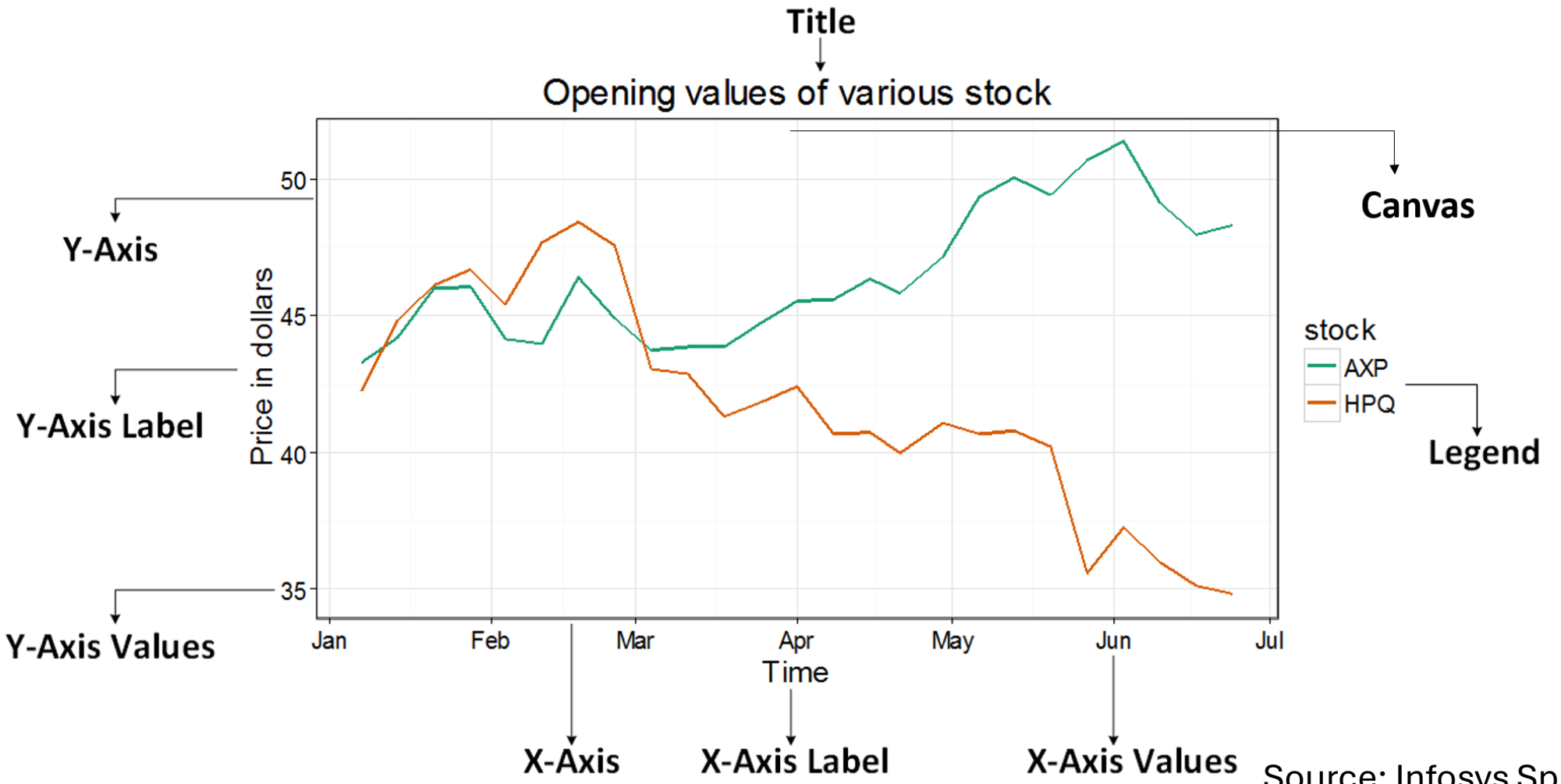
Develop insights from data


Data Visualization

Data Visualization

- Graphical representation of data or information using visual elements like graphs, charts and maps.
- This representation helps us understand patterns, trends, and outliers in the data.
- Data visualization:
  - helps in finding patterns and connections between variables
  - requires less effort from the reader to understand the visuals
  - condenses a large amount of information into a small space for quick analysis
  - provides relevant answers and clarity on certain questions swiftly

Plot

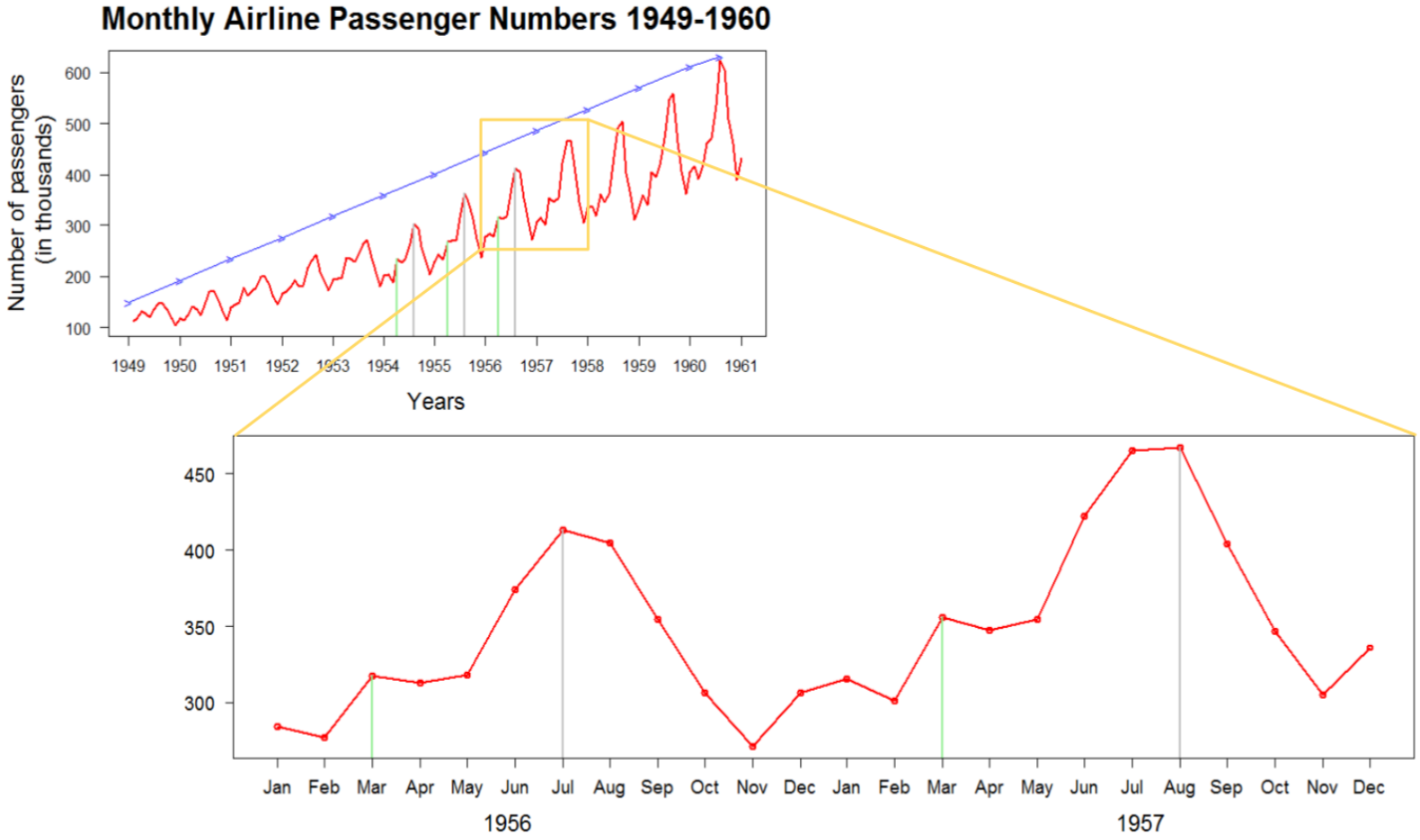


Source: Infosys Springboard

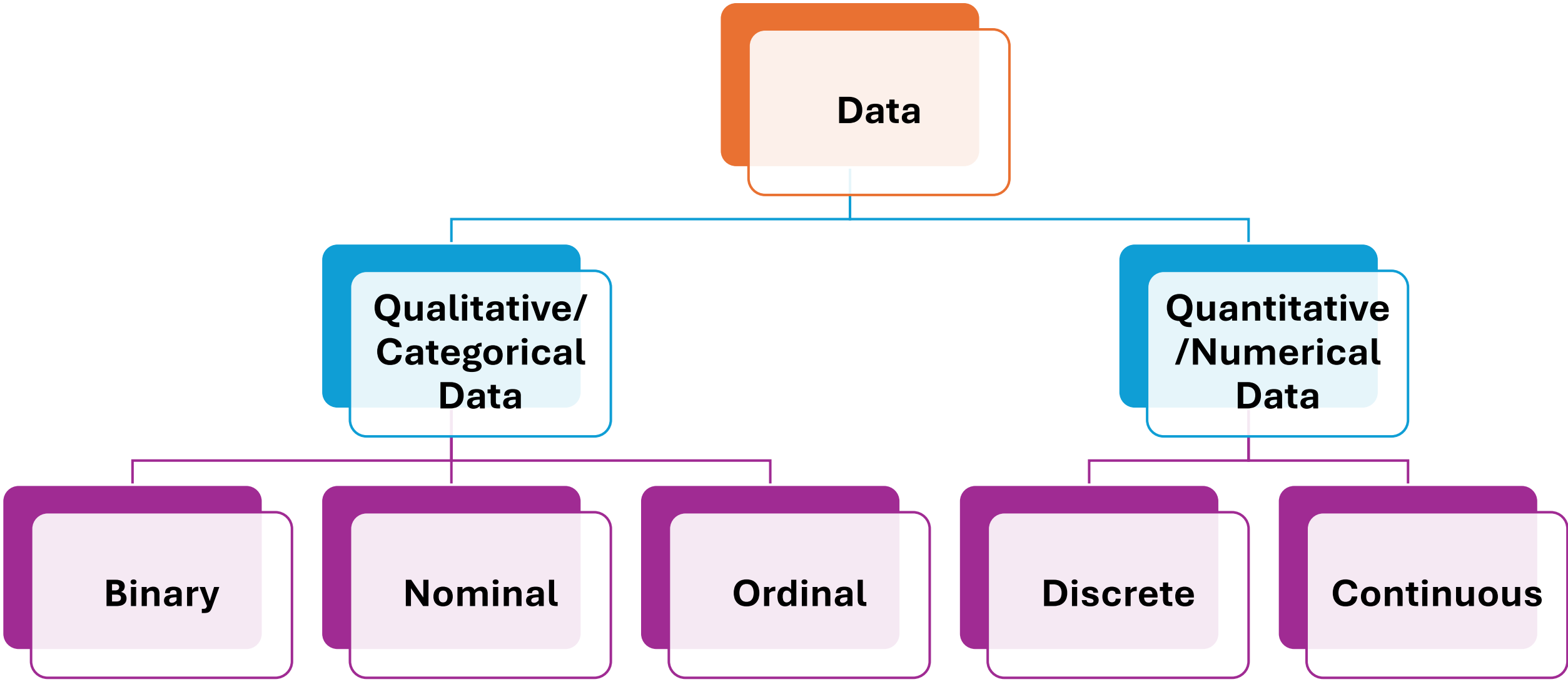
You can now observe with ease that:

Number of passengers has increased over the years (refer the first graph above)

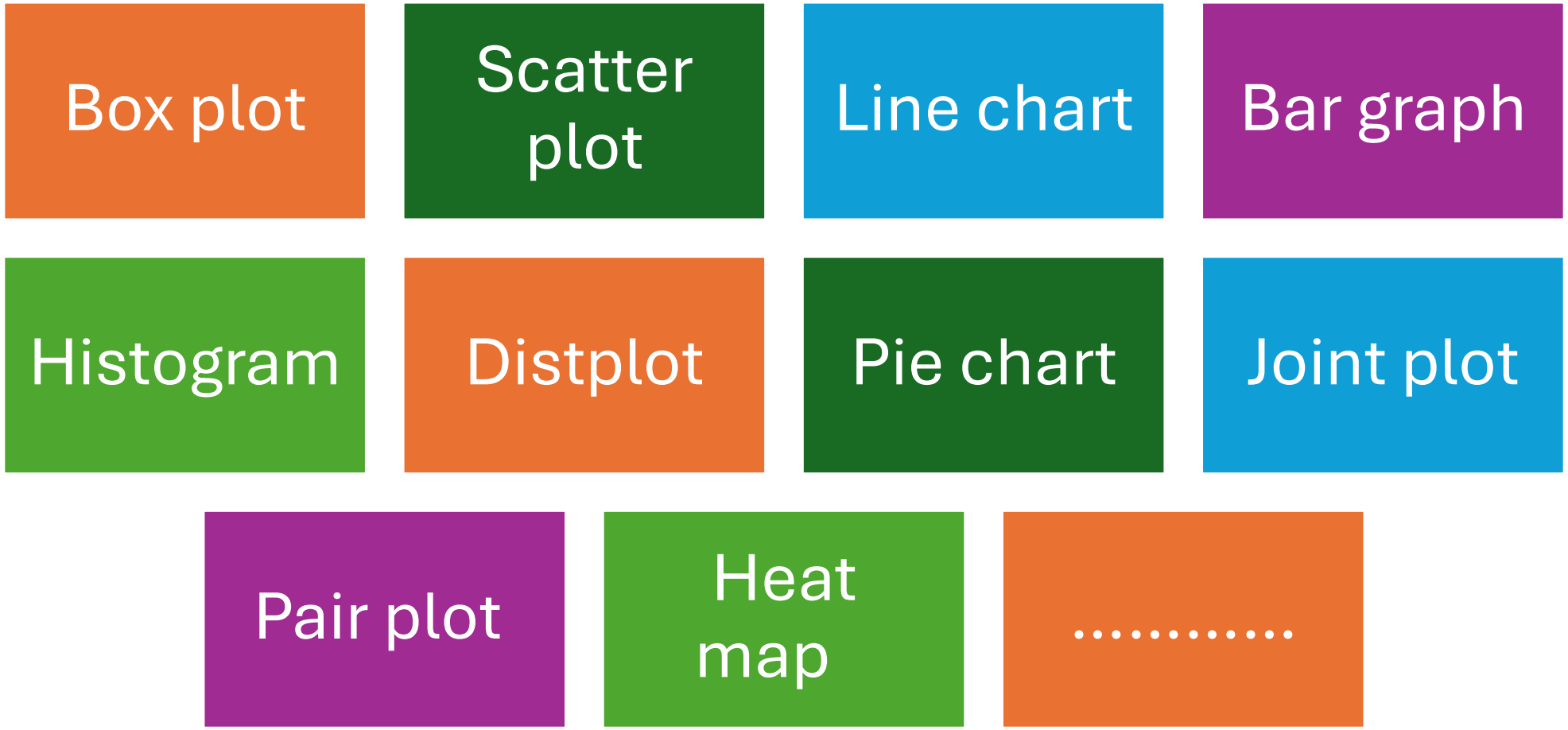
There is a peak in the middle of each year during the months of July and August (refer the second graph above)



Source: Infosys Springboard


Data
Qualitative/ Categorical Data
Quantitative /Numerical Data
Binary
Nominal
Ordinal
Discrete
Continuous

Visualization Constructs

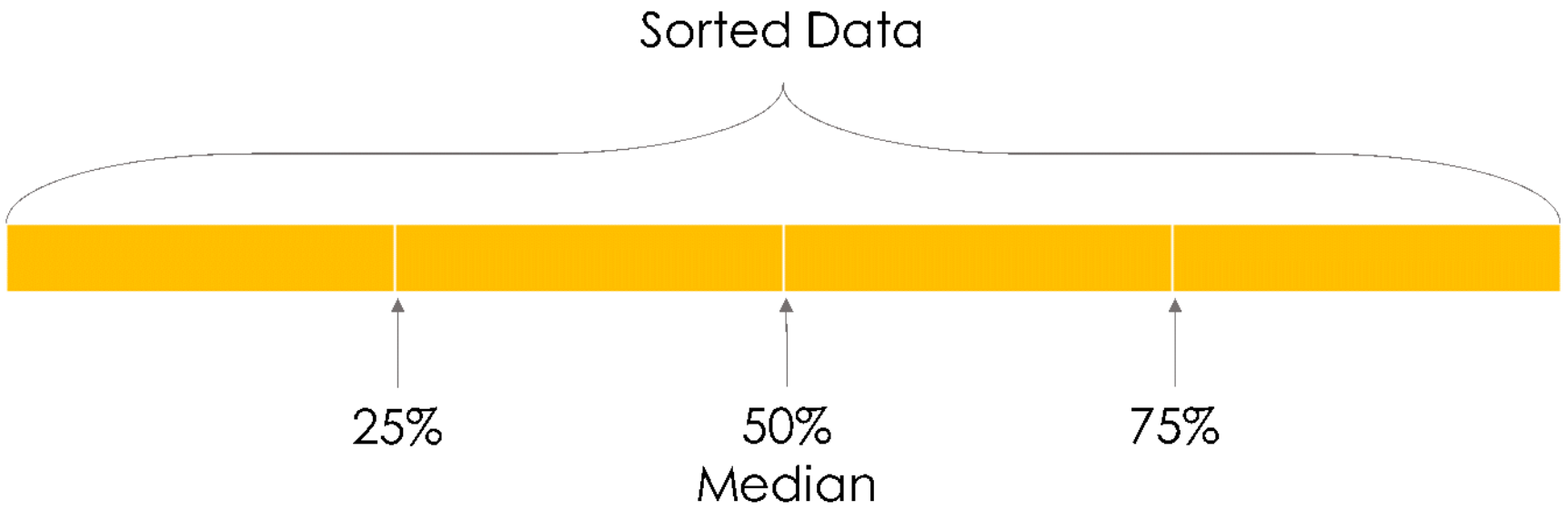


Data Analysis

- Exploratory data analysis to handle
  - missing values,
  - outliers, etc. and
  - analyze the relationship between the variables
- Require the knowledge of statistical concepts to
  - select the right type of graph
  - infer information like outliers, correlation of variables, redundant features etc.

Outliers & Quartiles

- **Outliers**
  - extreme values present in the dataset
- **Quartiles**
  - divide the number of data points into four equal-sized groups, or quarters.

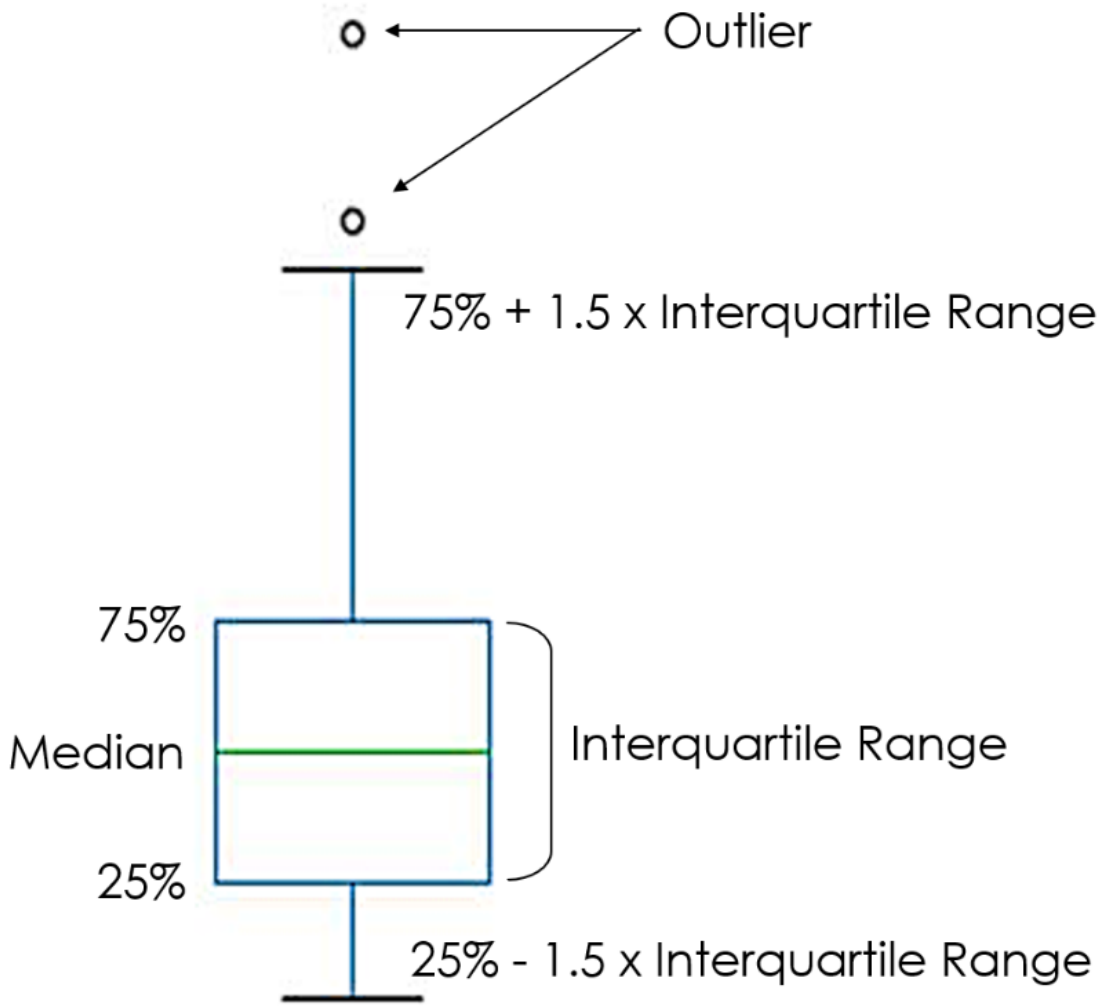


Source: Infosys Springboard

Quartiles

- Following are the steps to find quartiles:
  - sort the dataset in ascending order
  - find median of the sorted dataset (median divides the dataset into two halves - Quartile 2 or Q2)
  - repeat step 2 with the first and second half of the data (this gives Q1 and Q3, dividing the dataset into four equal parts)

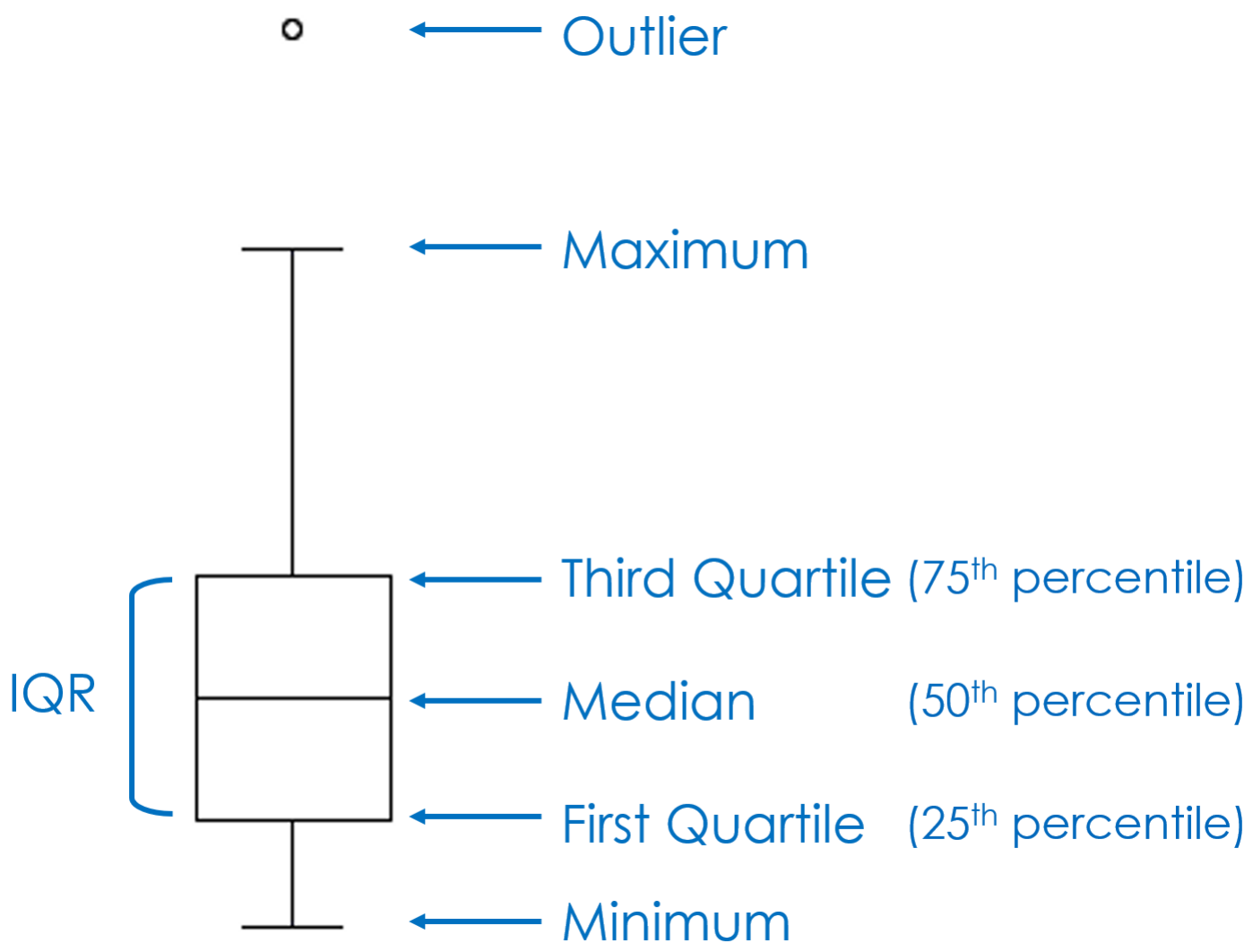
- Inter-Quartile Range (IQR)
  - IQR = Q3 - Q1
- **Upper Limit:** Q3 + (1.5 * IQR)
- **Lower Limit:** Q1 – (1.5 * IQR)



Source: Infosys Springboard

Box plot

- A standardised way of displaying the distribution of data based on a five-number summary (**minimum, first quartile (Q1), median, third quartile (Q3), and maximum**).



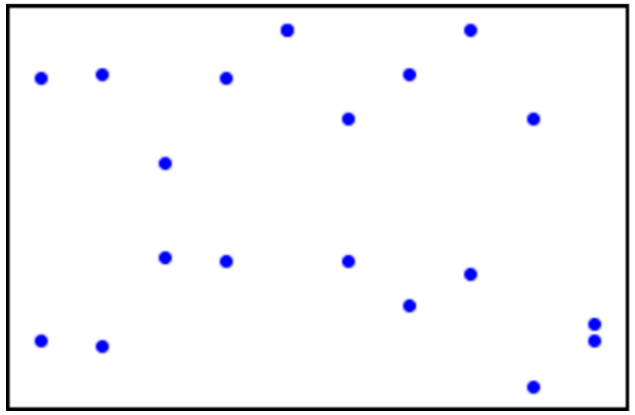
Source: Infosys Springboard

Scatter plot

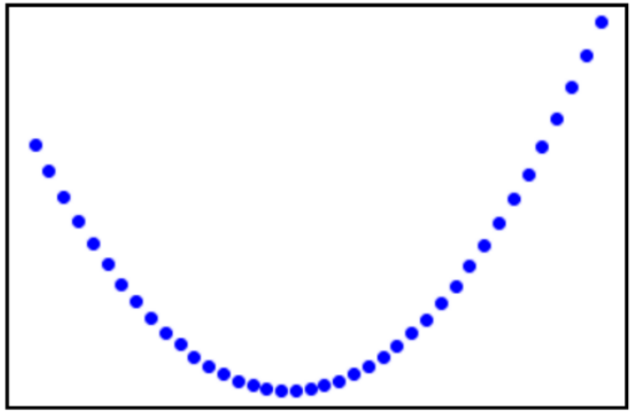
- Uses dots or markers to represent a value in the hyperplane
- Accept both quantitative and qualitative values

Finding Correlation Between Variables

No Correlation

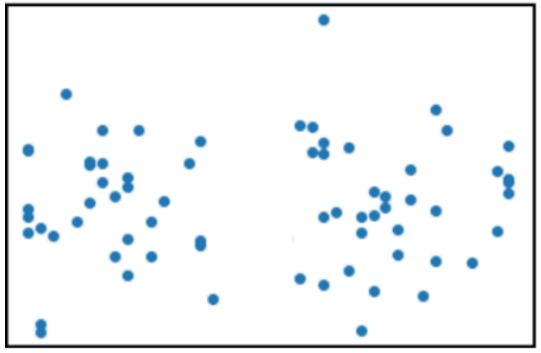


Strong Non-Linear Correlation

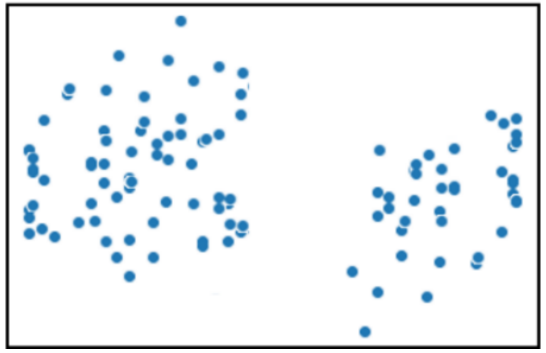


Identifying Patterns in Data

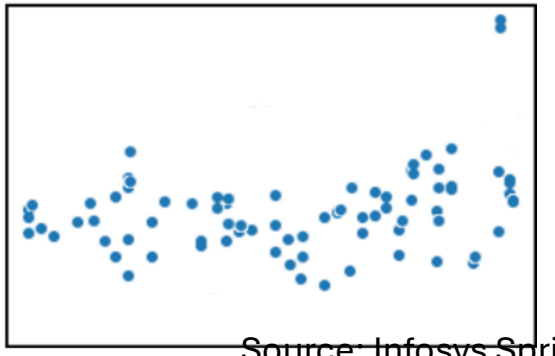
Gaps in Values



Data Clusters



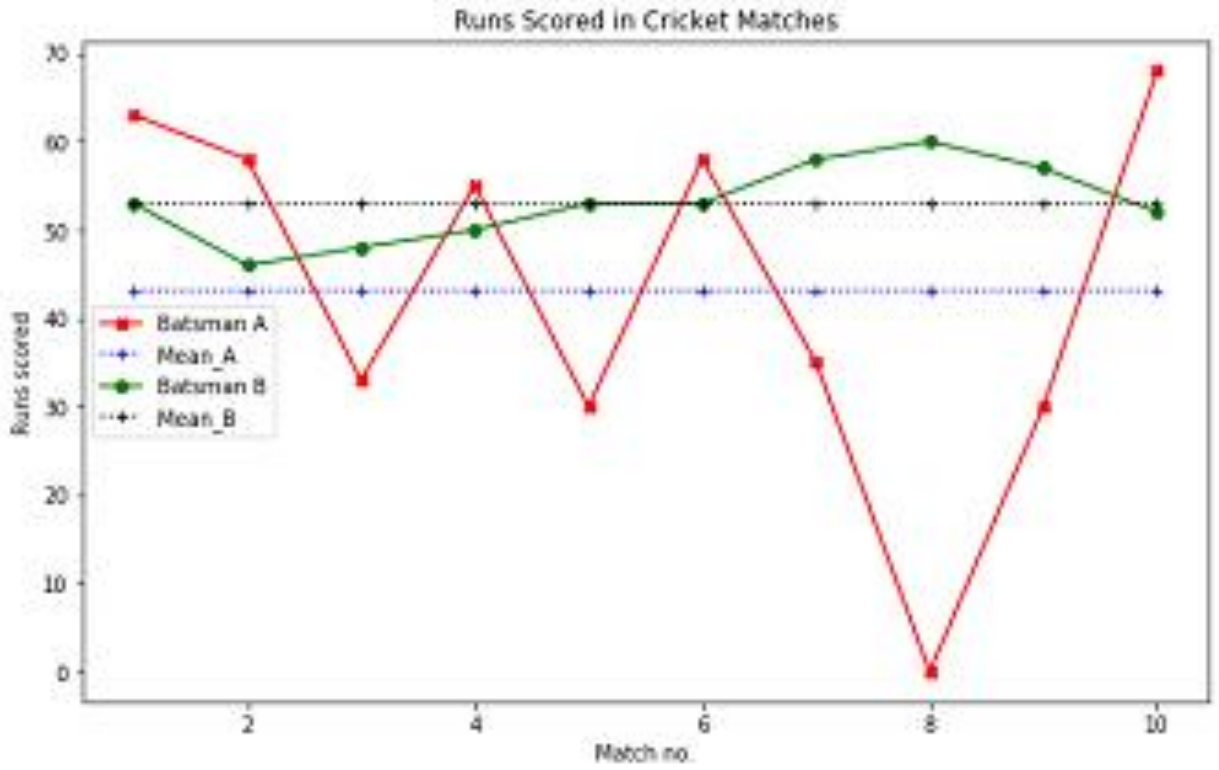
Outliers

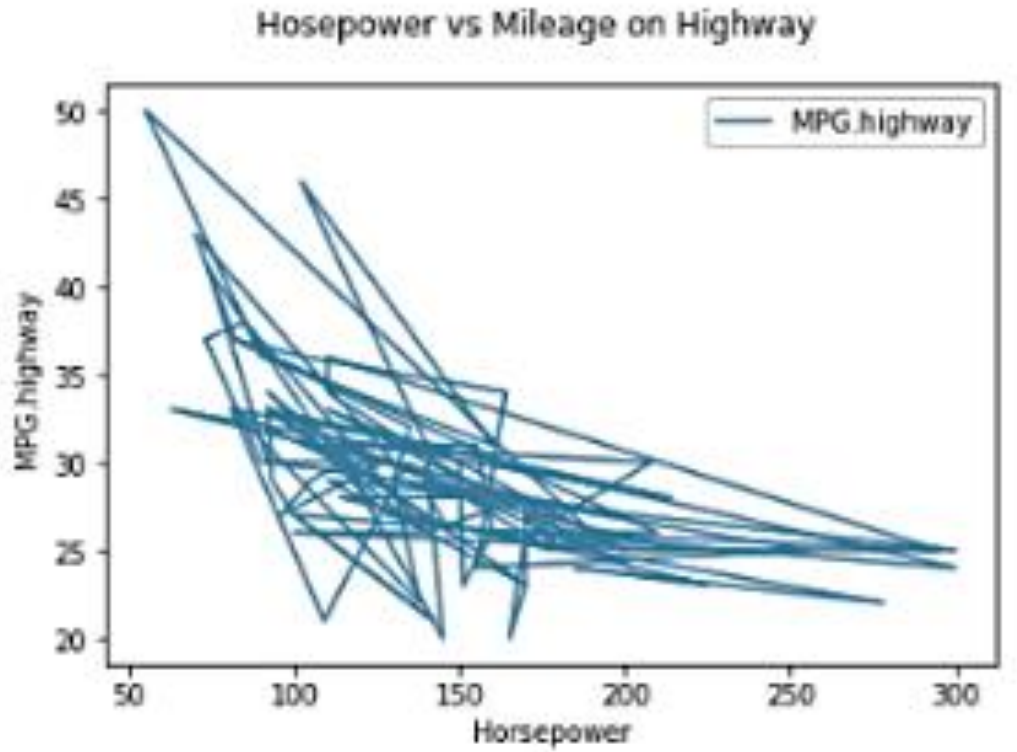


Source: Infosys Springboard

Line chart

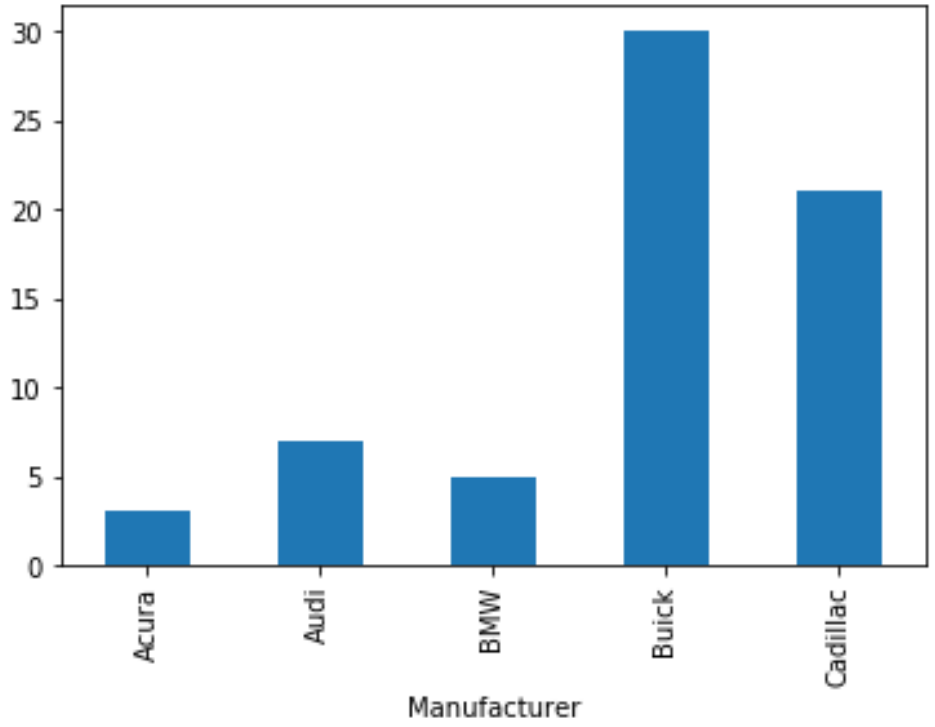
- Interconnecting all data points using straight line segments
- Used to analyse historic variations and trends in data
- For example, times series data.





Source: Infosys Springboard

Bar chart

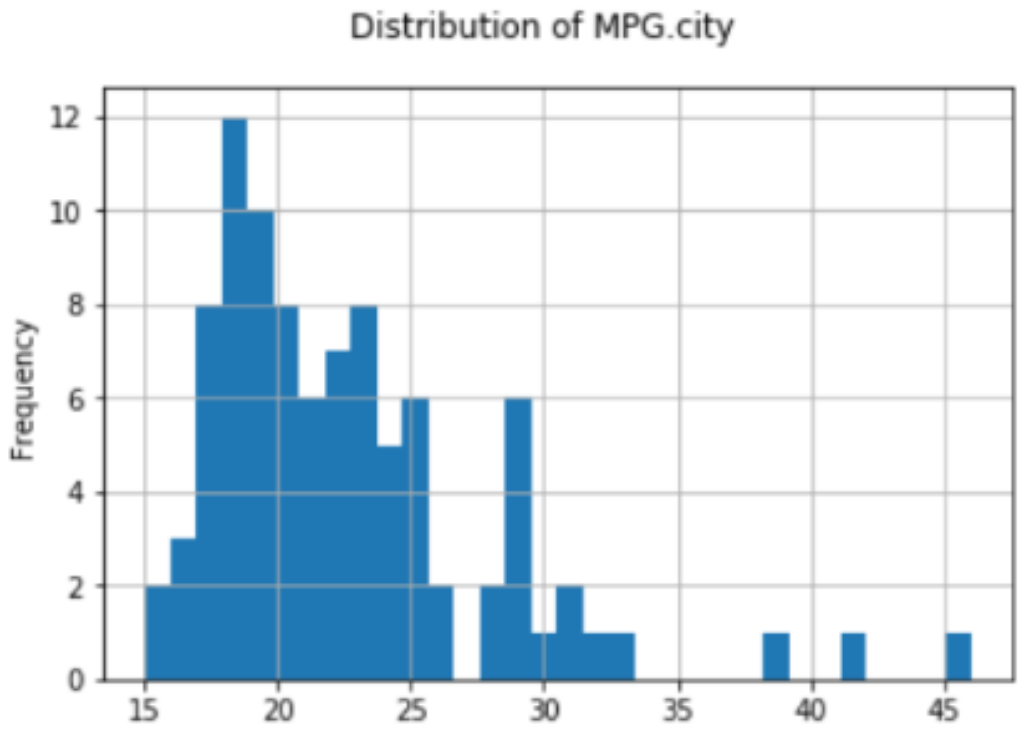


- A graph with rectangular bars that compares different categories
- Each bar represents a particular category and the length indicates the total number of values or items in that category.
- Multiple variations of bar charts including multiple, stacked, error bars, etc.

Source: Infosys Springboard

Histogram

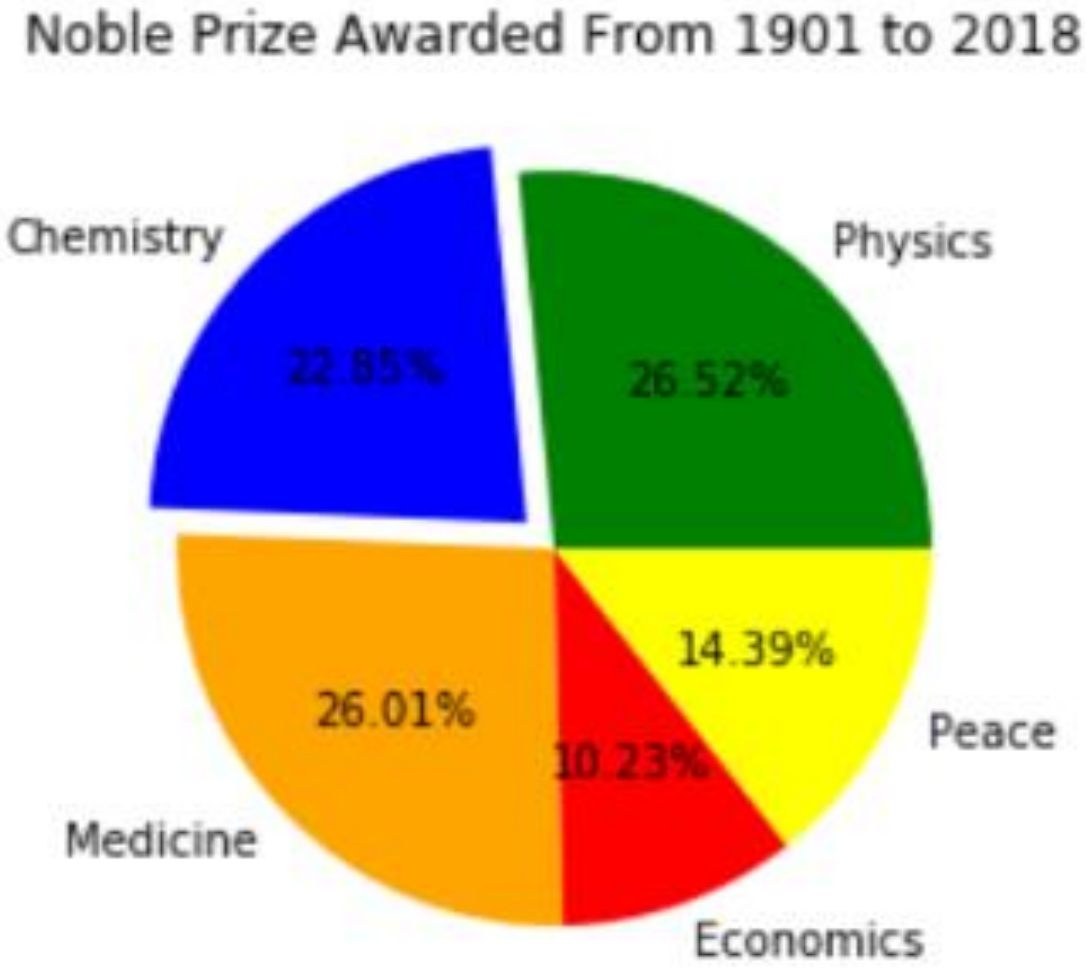
- data as rectangular bars
- it is used for continuous data
- ideally suited to obtain
  the frequency distribution of a given
  data



Source: Infosys Springboard

Pie chart

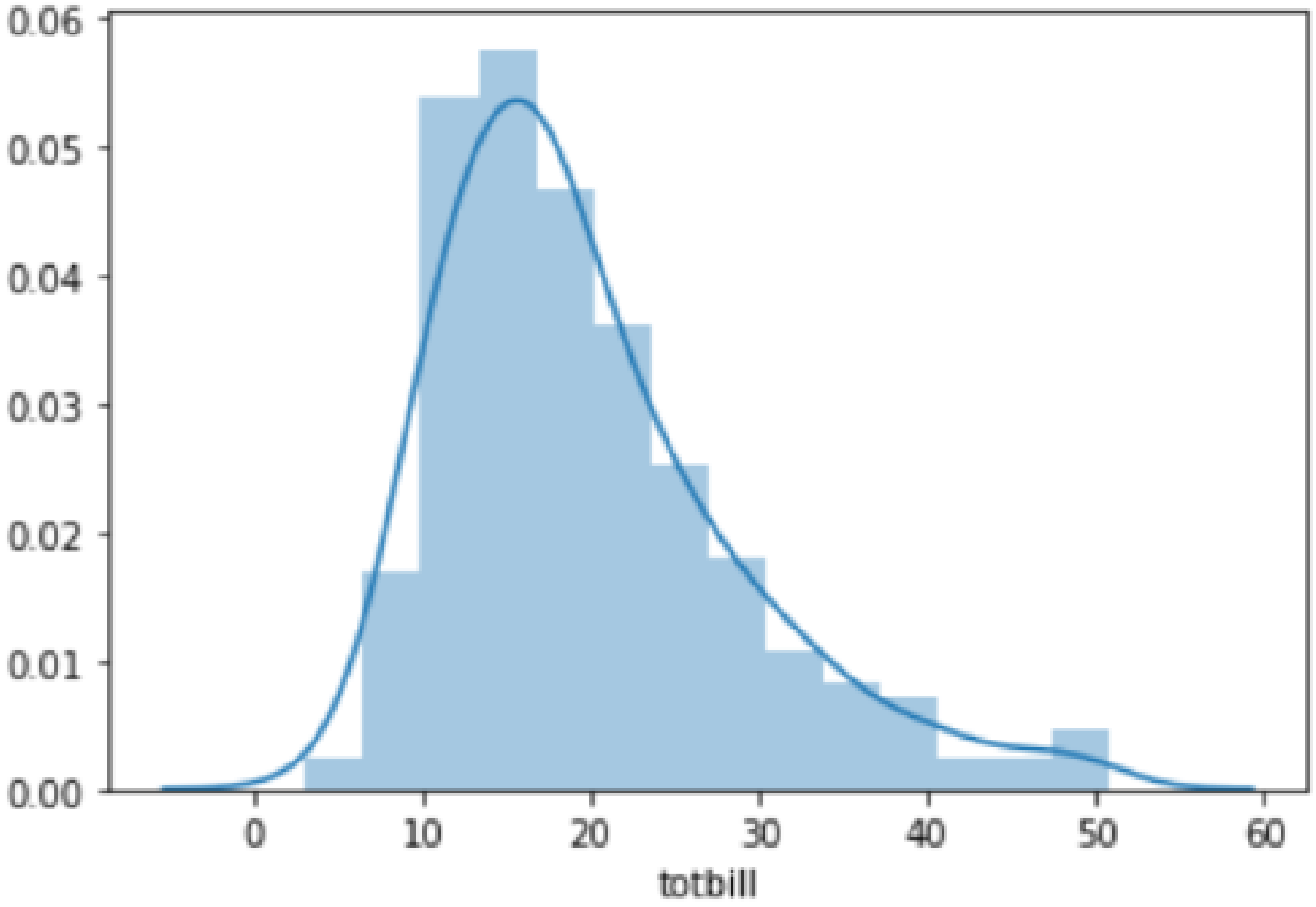
- divides the entire dataset into distinct groups
- The chart consists of a circle split into slices and each slice represents a group



Source: Infosys Springboard

Distribution plot

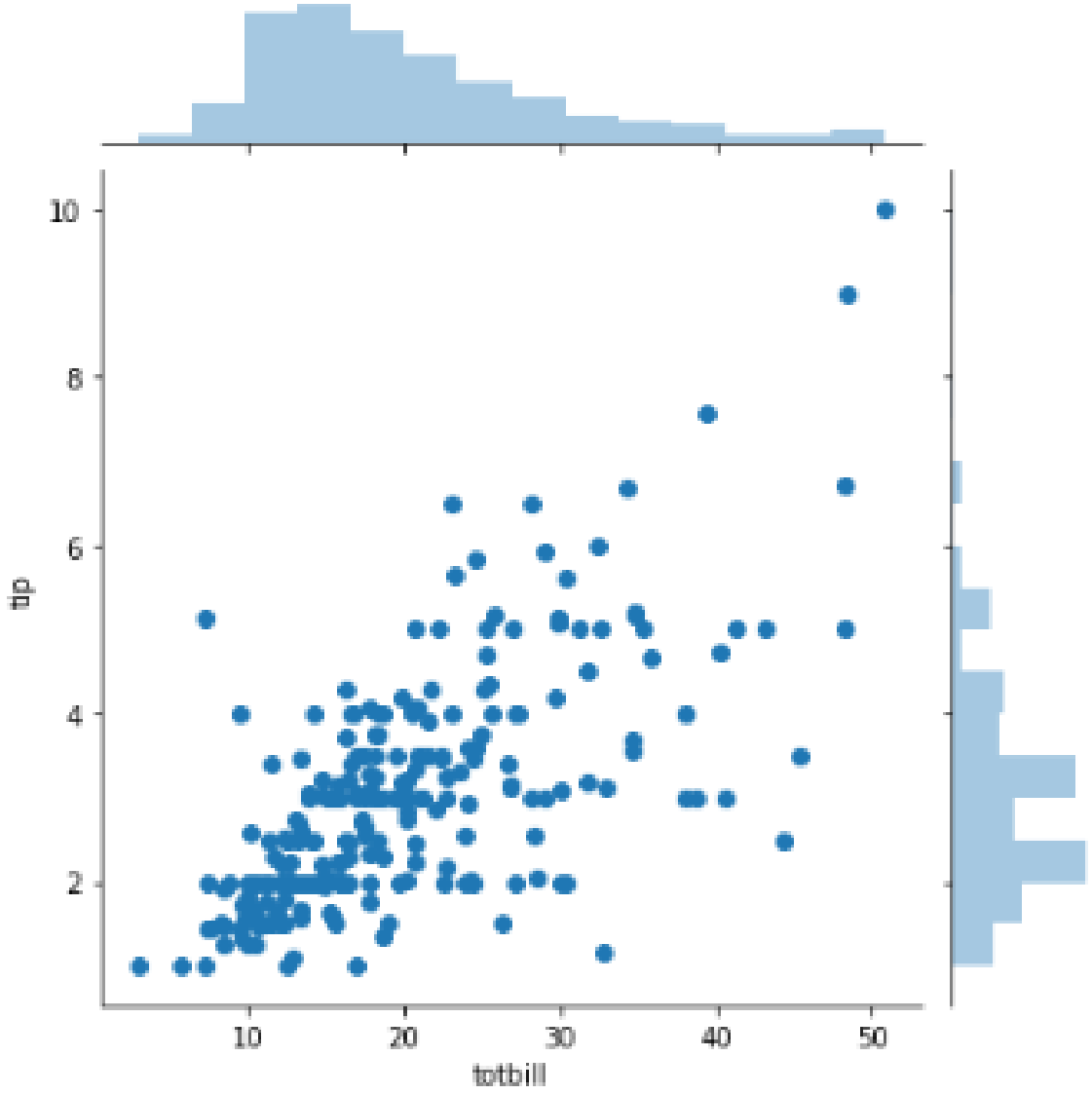
- Depicts the variation in a data distribution
- It represents the overall distribution of continuous data variables.
- Depicts the data by a histogram and a line in combination with it



Source: Infosys Springboard

Joint plot

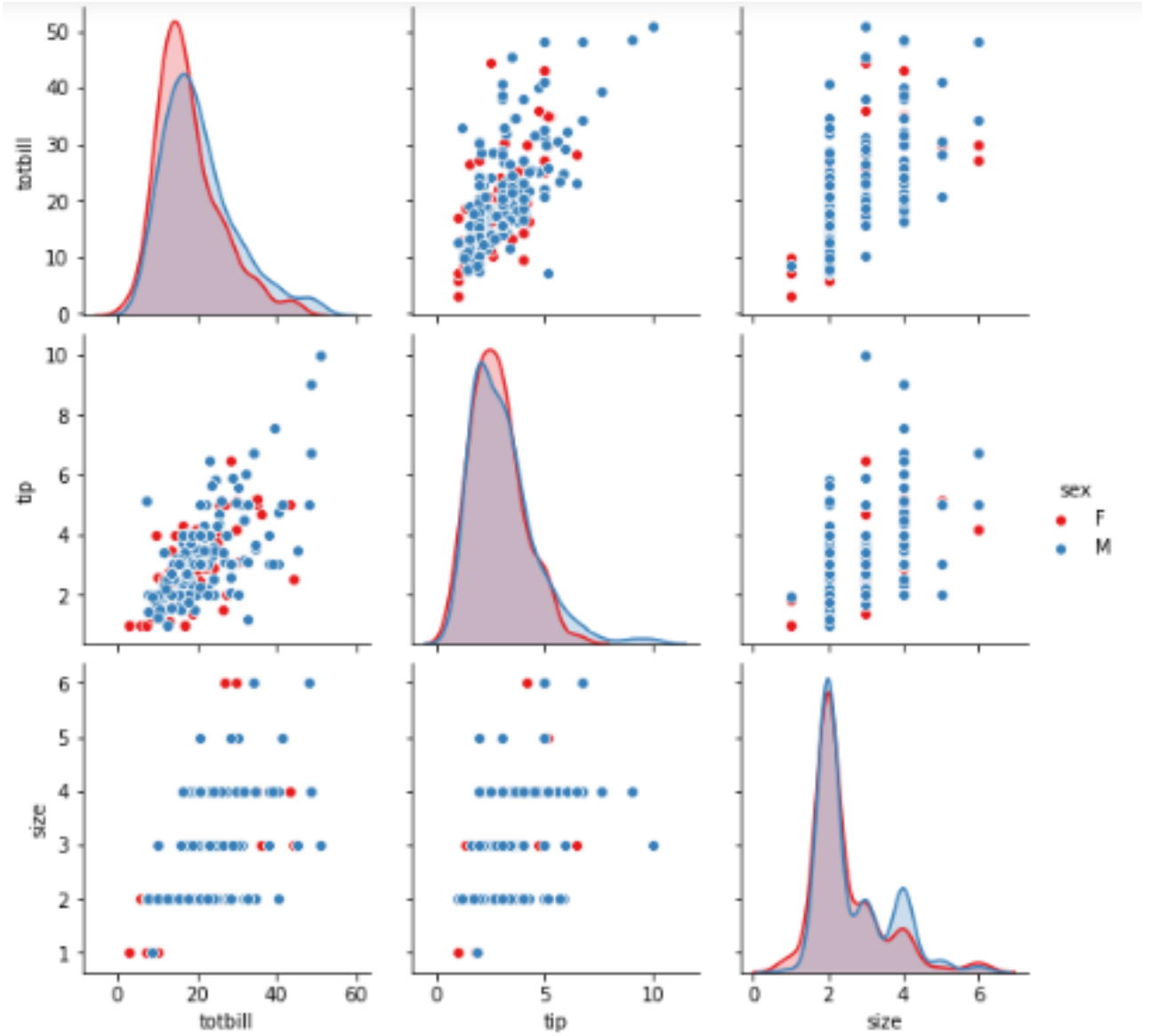
- Combination of two univariate and one bivariate plots
- The bivariate plot (in the center) helps in analysing the relationship between two variables.
- The univariate plot describes the distribution of data in each variable as a marginal plot.



Source: Infosys Springboard

Pair plot

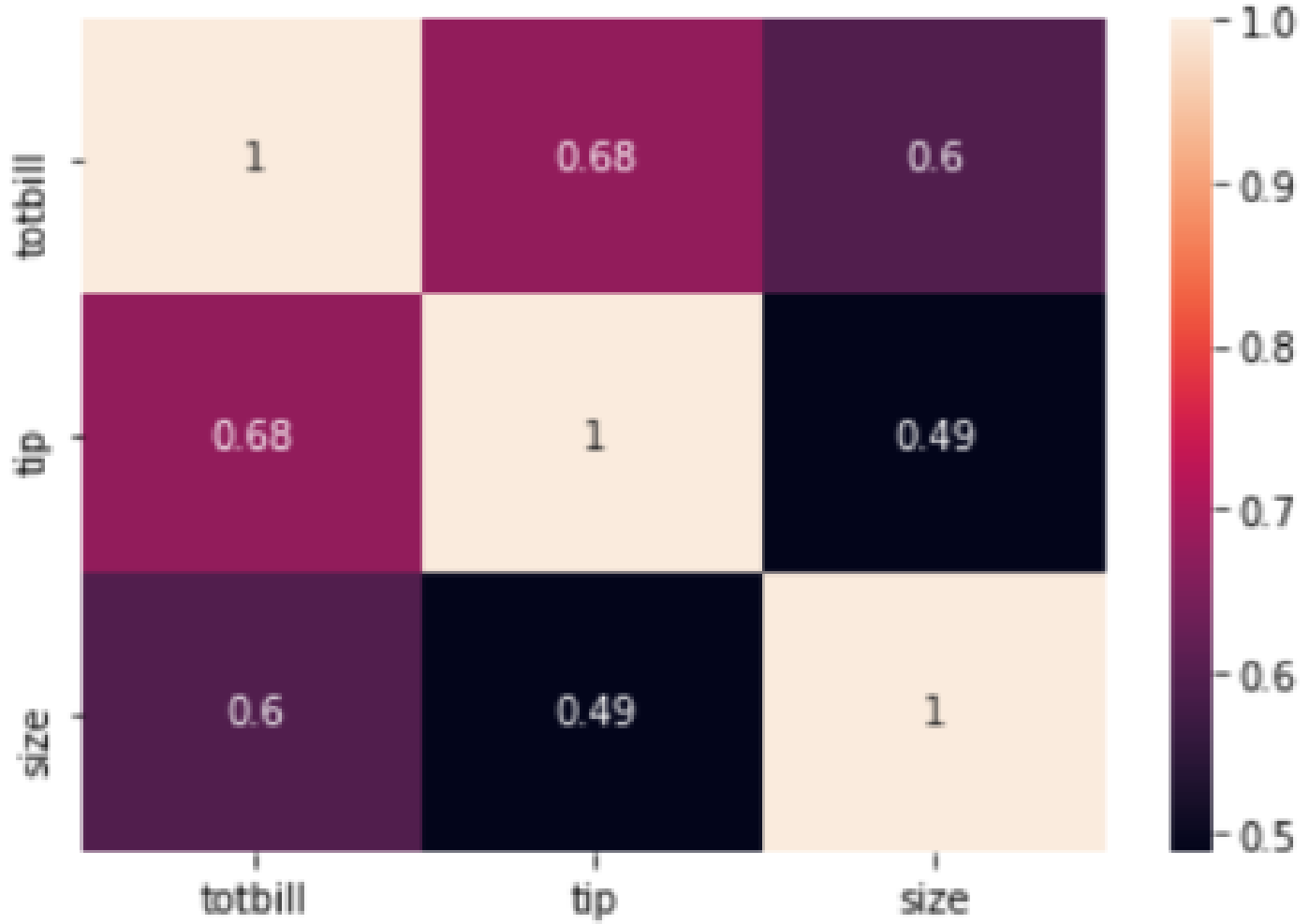
- Pairwise relationships between all the variables in a dataset in a matrix format
- Each row and column in the matrix represents a variable in the dataset.
- The plots present in the diagonal are univariate plots as the variables are compared with themselves and the others are bivariate scatter plots



Source: Infosys Springboard

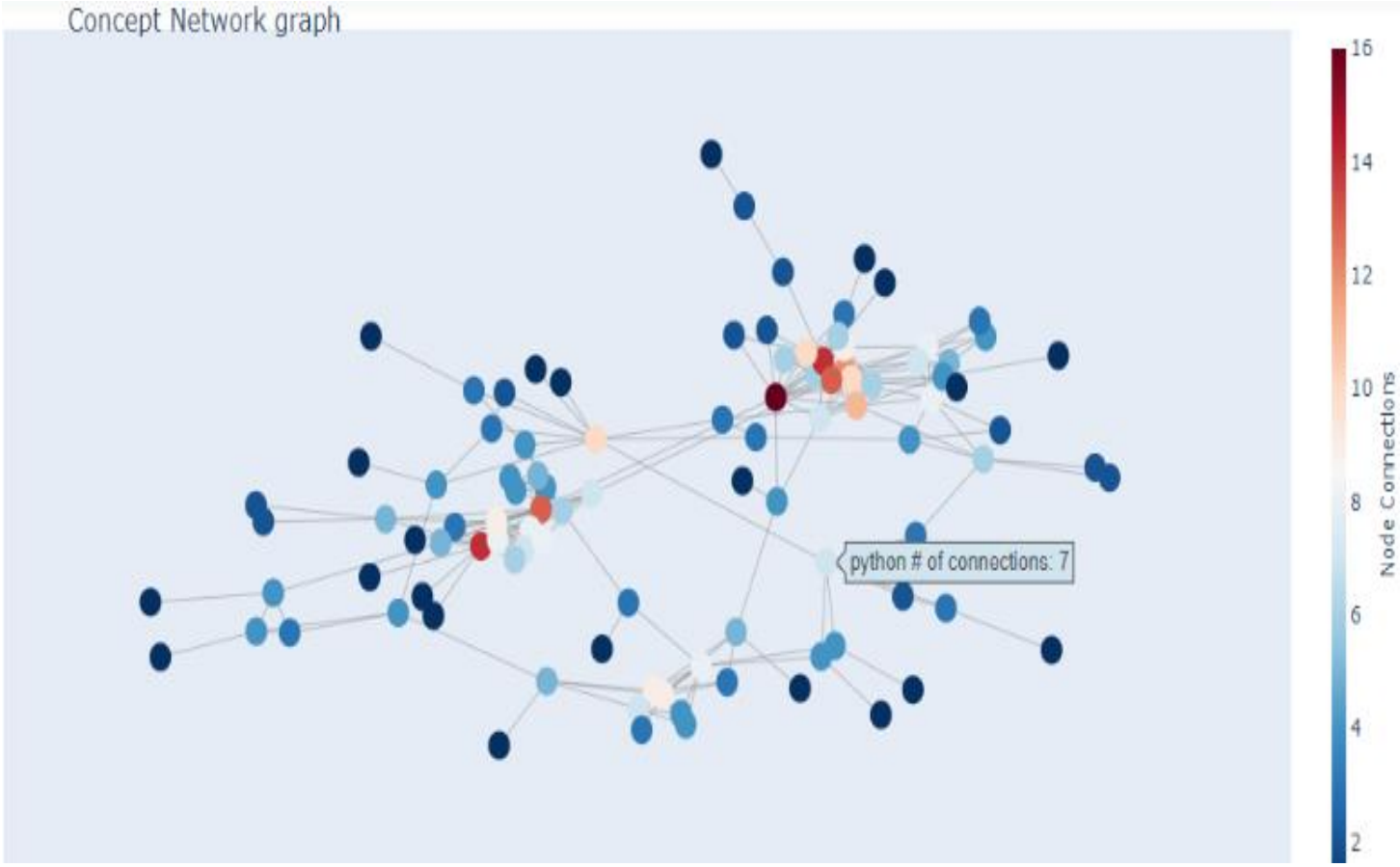
Heat map

- Graphical representation of data where similar values are depicted by the same colours
- The colours vary based on the intensity of the results.



Network graph

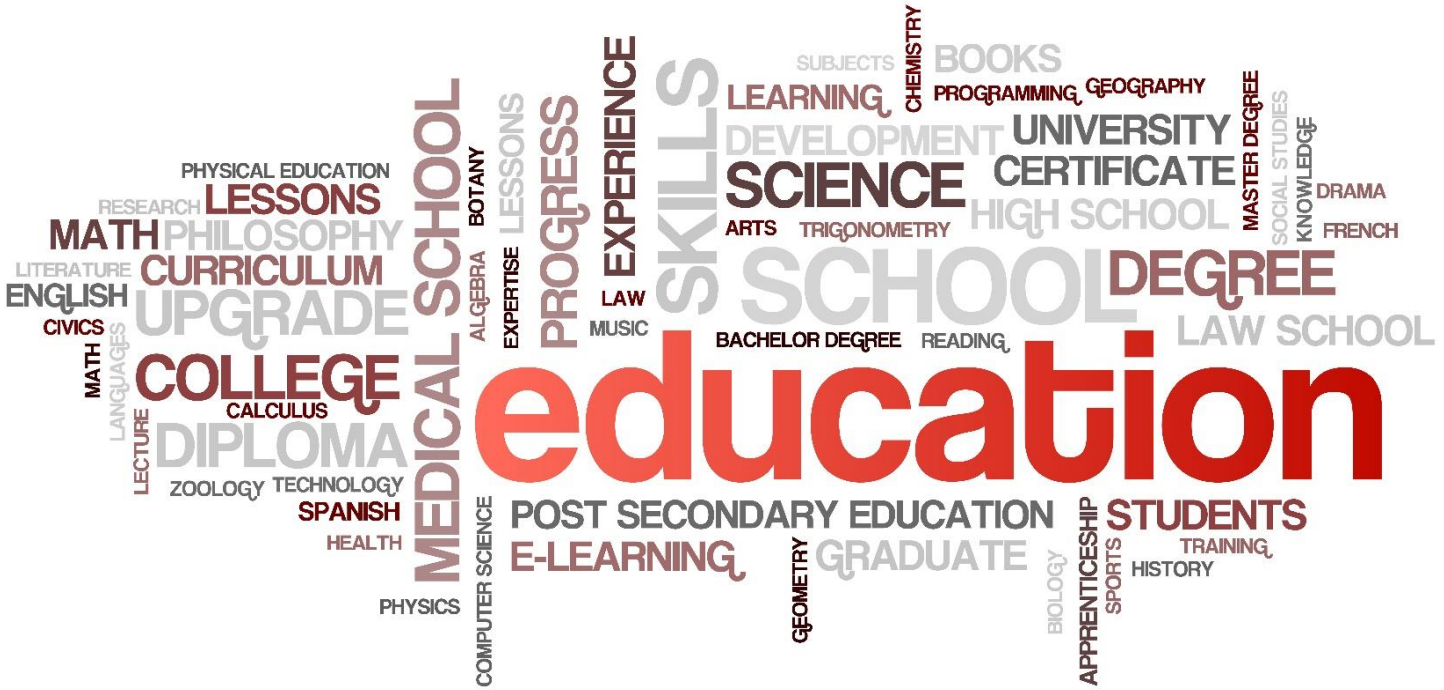
- A network is a set of objects (called nodes or vertices) that are connected.
- The connections between the nodes are called edges or links.
- If the edges in a network are directed, i.e., pointing in only one direction, the network is called a directed network
- If all edges are bidirectional, or undirectional, the network is an undirected network.



Source: Infosys Springboard

Word cloud

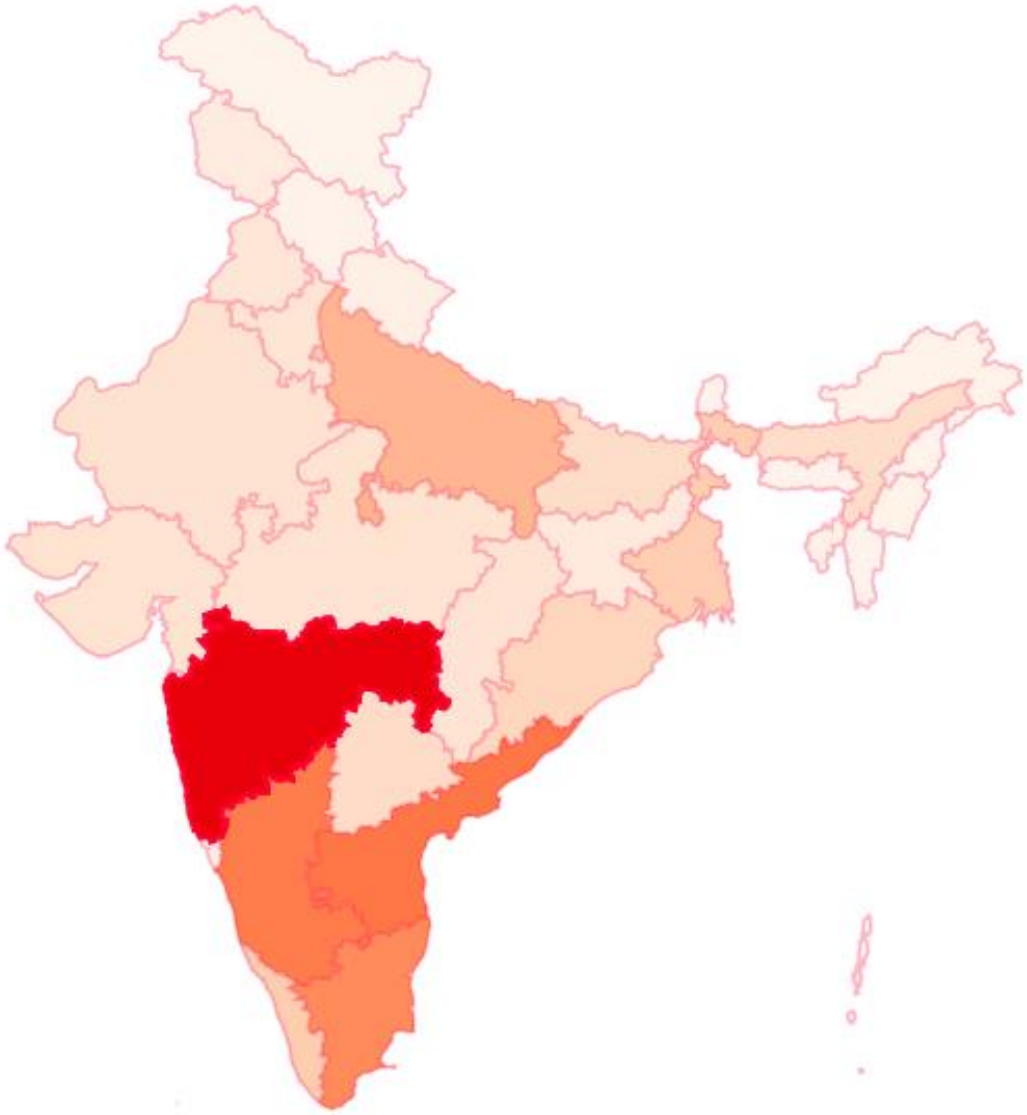
- Visual representation of free form text, which is like a collage
- Depict keyword metadata of websites, articles, reviews, feedbacks etc.
- The frequency and significance of the words are depicted by the font, font size and colour of the text in the cluster.



Source: Infosys Springboard

Choropleth map

- Pictorial representation of data on a geographical map
- The intensity of color in a region on the map corresponds to the respective values



Source: Infosys Springboard

Microsoft Power BI

Data Visualization

- Can help to:
  - Identify key areas and hidden patterns.
  - Get factors that give better customer insights.
  - Analyze and associate data and products properly.
  - Make proper predictions.

Need For Power BI

Spot trends in real-time

Automatically search hidden insights

Custom visualizations

Enterprise-ready

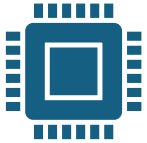
Power BI

The Power BI for data science and analytics explores the AI and analytics features present in Power BI.

It is a business analytics service provided by Microsoft.

It offers data warehouse capabilities, including

- data preparation,
- data discovery, and
- interactive dashboards
- ETL operations and
- deliver reports



There are a variety of range of features

- anomaly detection,
- QnA,
- language detection,
- text analytics,
- forecasting predictions,
- linear detection model, and so on

Power BI's architecture has three phases

1. Data Integration
2. Data Processing
3. Data Presentation

Let’s start Power BI

Install and run

1. Download from the link

- Windows

https://www.microsoft.com/en-us/power-platform/products/power-bi/desktop

- MAC

https://apps.apple.com/us/app/microsoft-power-bi/id929738808

Navigating Power BI Desktop & Power BI service interfaces

Import Data

1. web

https://www.fool.com/research/best-states-to-retire

2. web

https://en.wikipedia.org/wiki/List_of_U.S._state_abbreviations

3. Import the OData feed's order data

https://services.odata.org/V3/Northwind/Northwind.svc/

4. Import excel/csv/txt... from the secondary device

Browse and load

Transform data (**financial sample.xlsx**)

- Cross-check data types (eg: Unit Sold- Whole number)
- Rename columns
  - Month number: month
  - Sale price: unit price
- Convert values to Uppercase
  - Country
- We know the Montana product was discontinued last month,
  - filter this data from our report to avoid confusion.
- Remove columns
  - Discount band & discount

Create a financial dashboard on “financial sample.xlsx” dataset.

The dashboard should have the following:

- A dashboard heading including title and author name
- Total sales
- Units sold
- A table to show category wise total sales of products
- A plot to show Sales by year/month
- A plot to show sales by segments
- A plot to show sales by country
- A plot to show profit by segments
- A plot to show sales by product

Create a new measure

Total Units Sold = SUM(financials[Units Sold])

Finance Dashboard By: Dr. Sachi

Sum of Sales
118,726,350.26

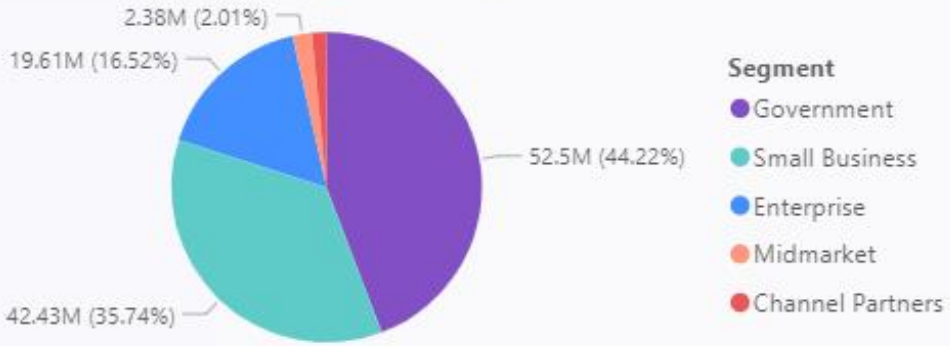
Sum of Units Sold
1125806

Product	Sum of Sales
Paseo	33,011,143.95
VTT	20,511,921.02
Velo	18,250,059.47
Amarilla	17,747,116.06
Montana	15,390,801.88
Carretera	13,815,307.89
Total	**118,726,350.26**

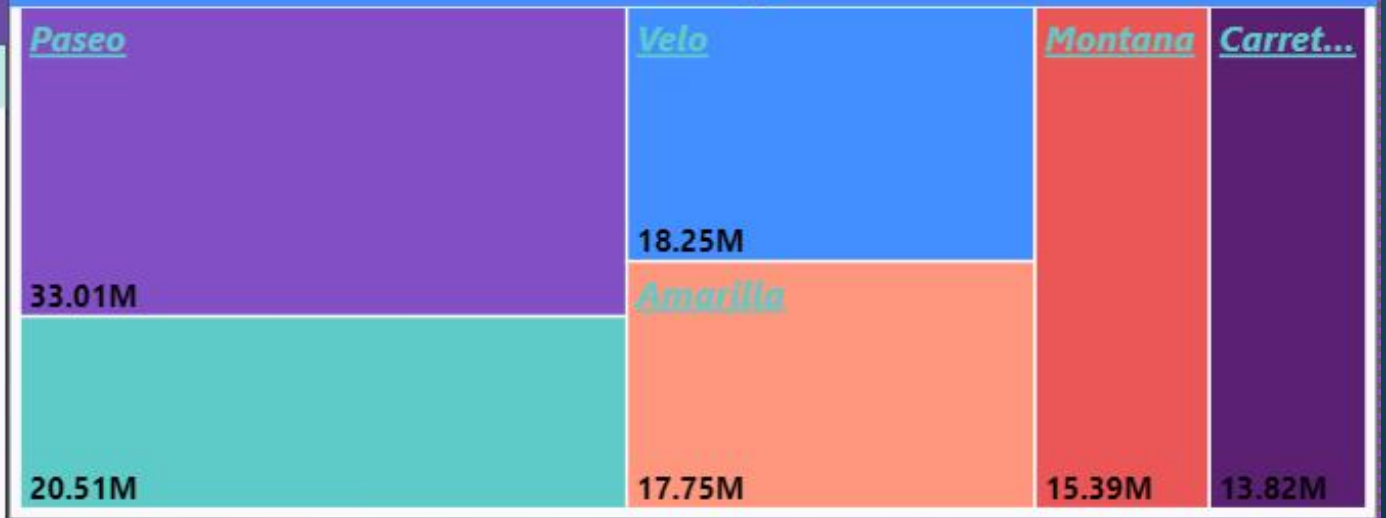
Sum of Sales by Country

Sum of Sales by Month

Sum of Sales by Segment



Sum of Sales by Product



Shape and combine data in Power BI Desktop

Doc

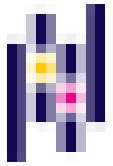
In this tutorial, you'll learn how to:

- Shape data by using Power Query Editor.
- Connect to different data sources.
- Combine those data sources, and create a data model to use in reports.

Python for Data Visualization

Python libraries and frameworks for Data Science

NumPy



pandas

matplotlib

seaborn

dask

bokeh

plotly

SciPy

.....

Task

Let us consider a use case, where a customer wishes to buy a car. Following are some of the questions that the customer might have before the purchase.

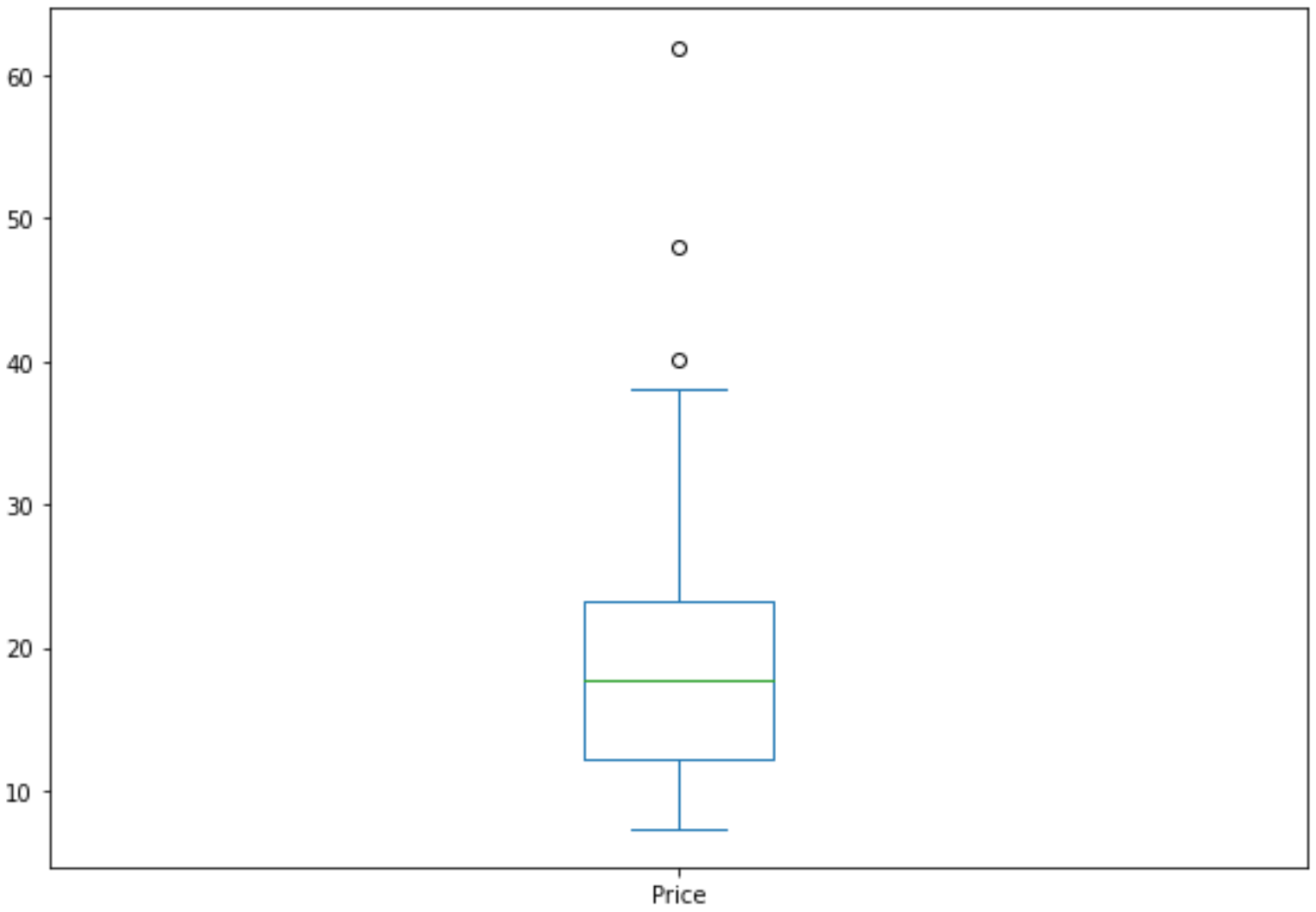
- **What is the price range of different types of cars available in the market?**
- **What is the range of horsepower and mileage of various cars?**
- **Does a car with higher horsepower give lower mileage?**
- **How much leg space does the car have?**
- **How many passengers can the car carry based on its type?**

- **Let us use 'Cars93' dataset to answer the above questions. Click here to downloaded the dataset.**

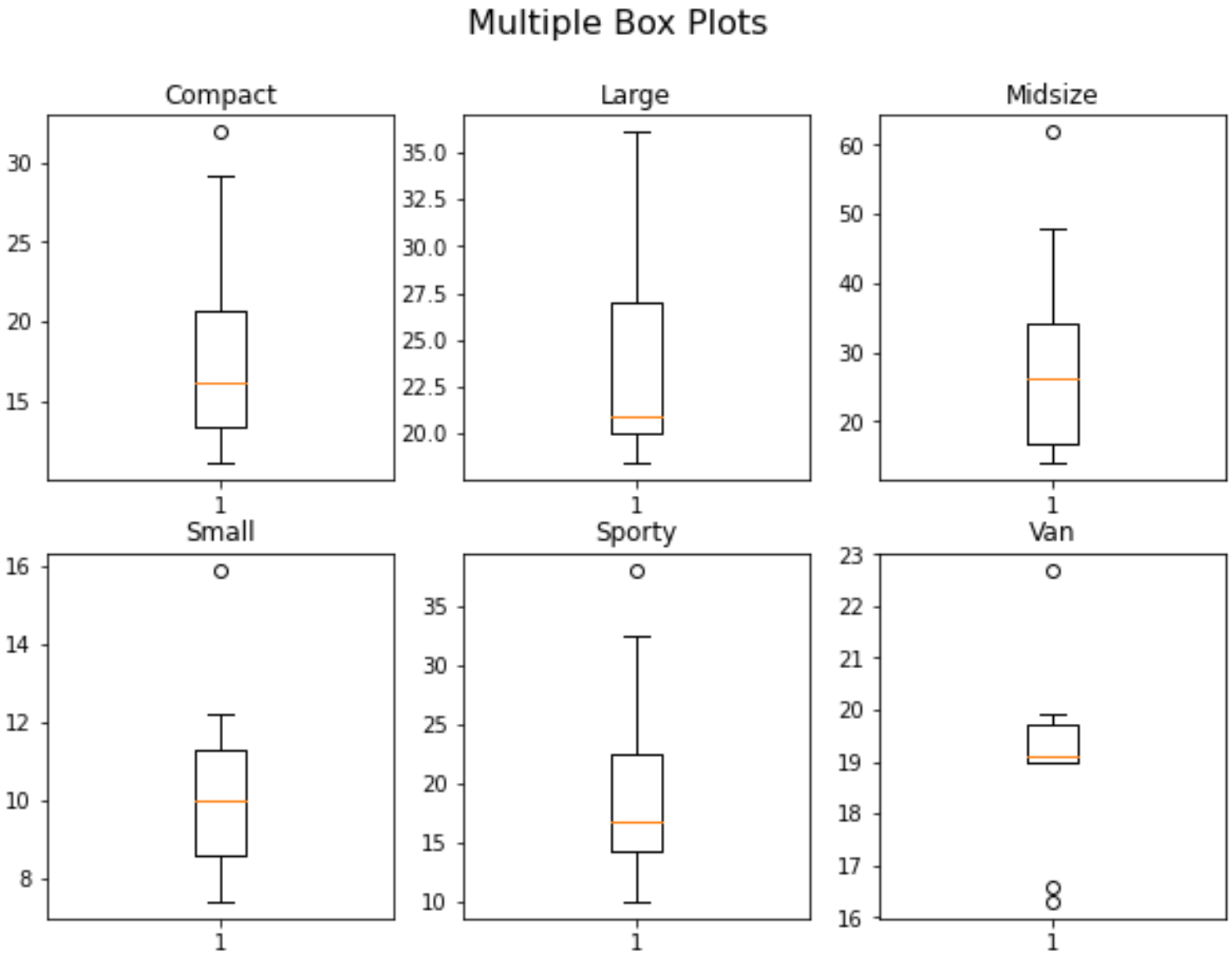
	Manufacturer	Model	Type	Price	MPG.city	MPG.highway	Horsepower	Rear.seat.room	Passengers
0	Acura	Integra	Small	15.9	25	31	140	26.5	5
1	Acura	Legend	Midsize	33.9	18	25	200	30.0	5
2	Audi	90	Compact	29.1	20	26	172	28.0	5
3	Audi	100	Midsize	37.7	19	26	172	31.0	6
4	BMW	535i	Midsize	30.0	22	30	208	27.0	4

Dataset view

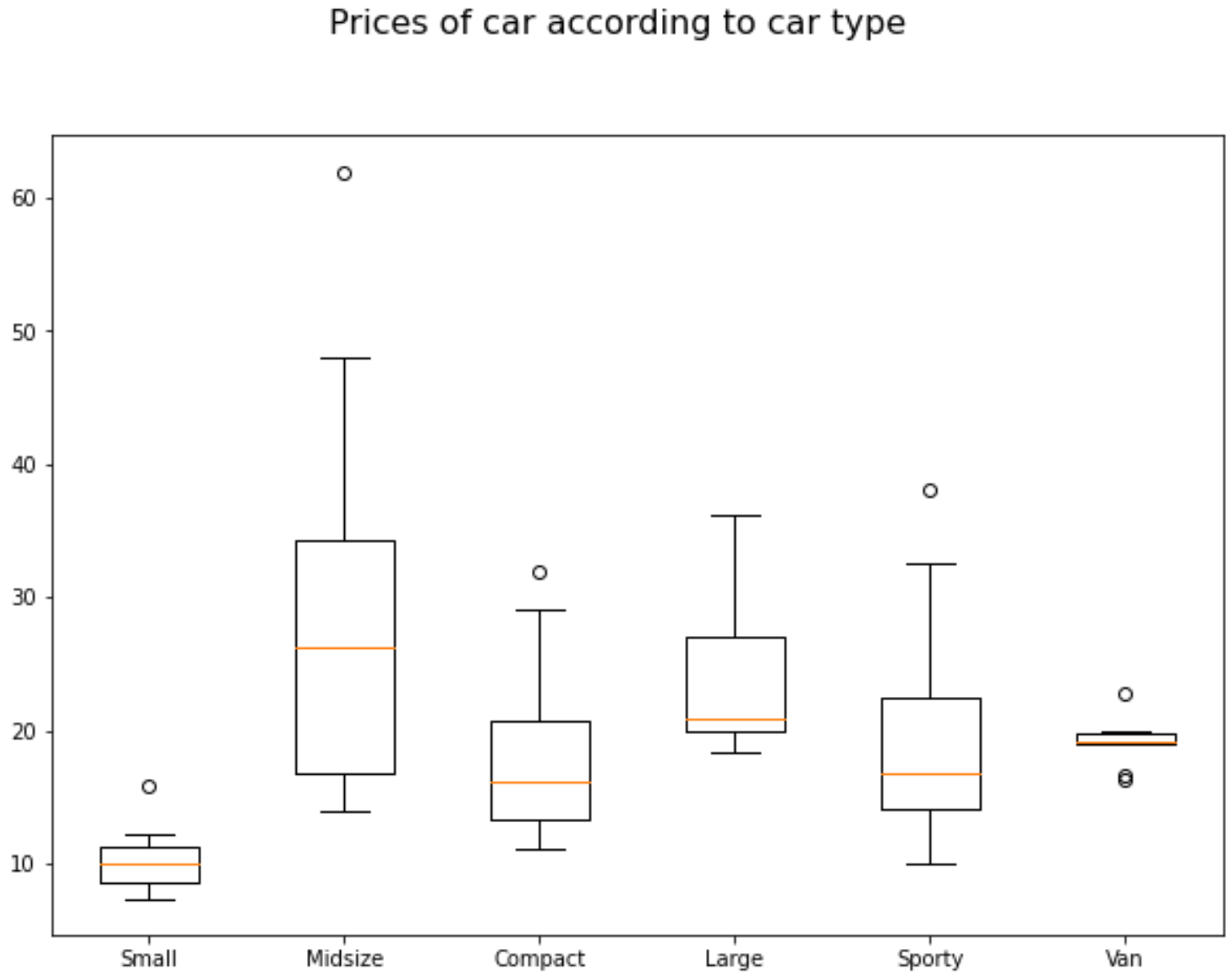
What is the price range of cars available in the market?



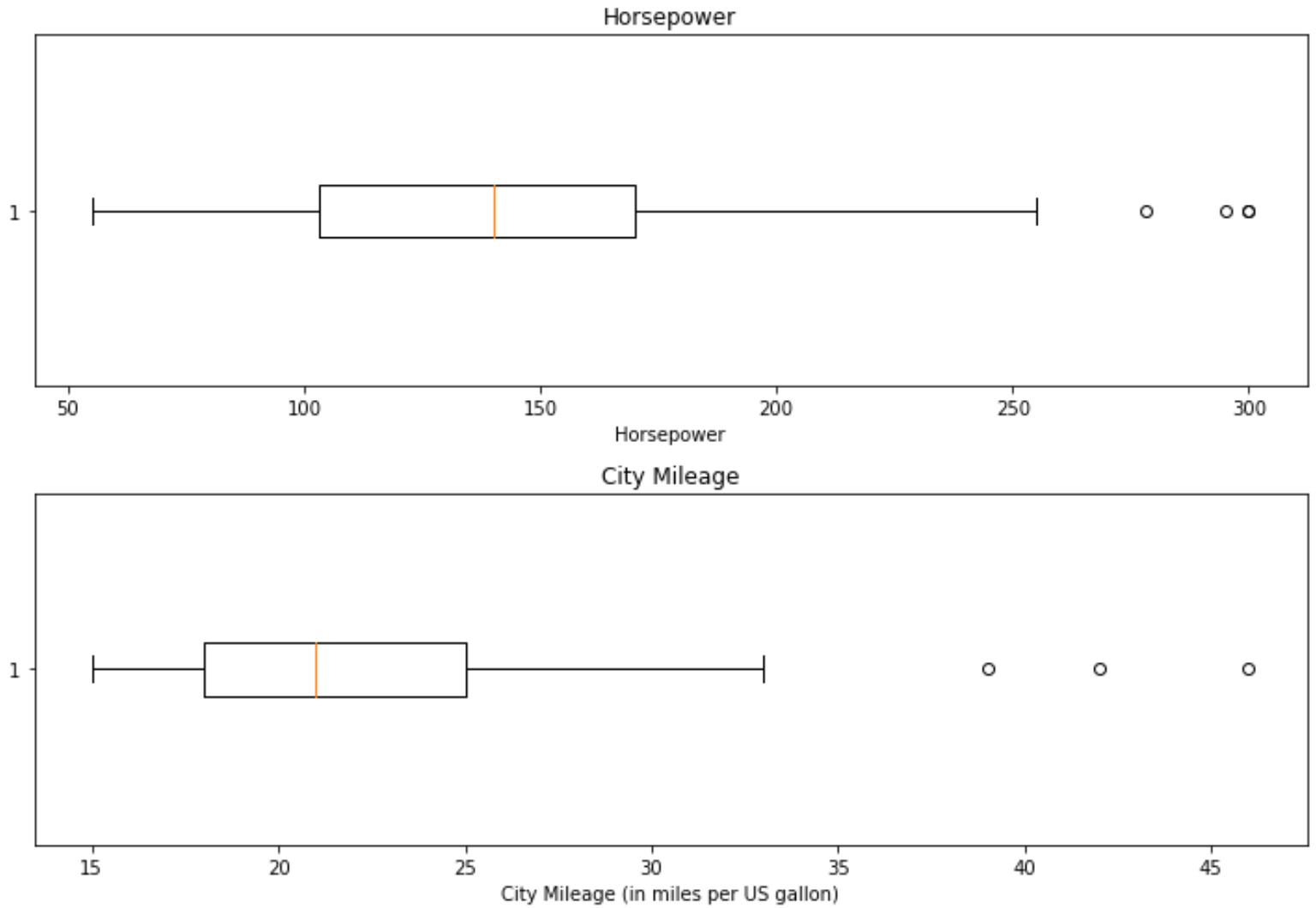
What is the price range of different types of cars available in the market?



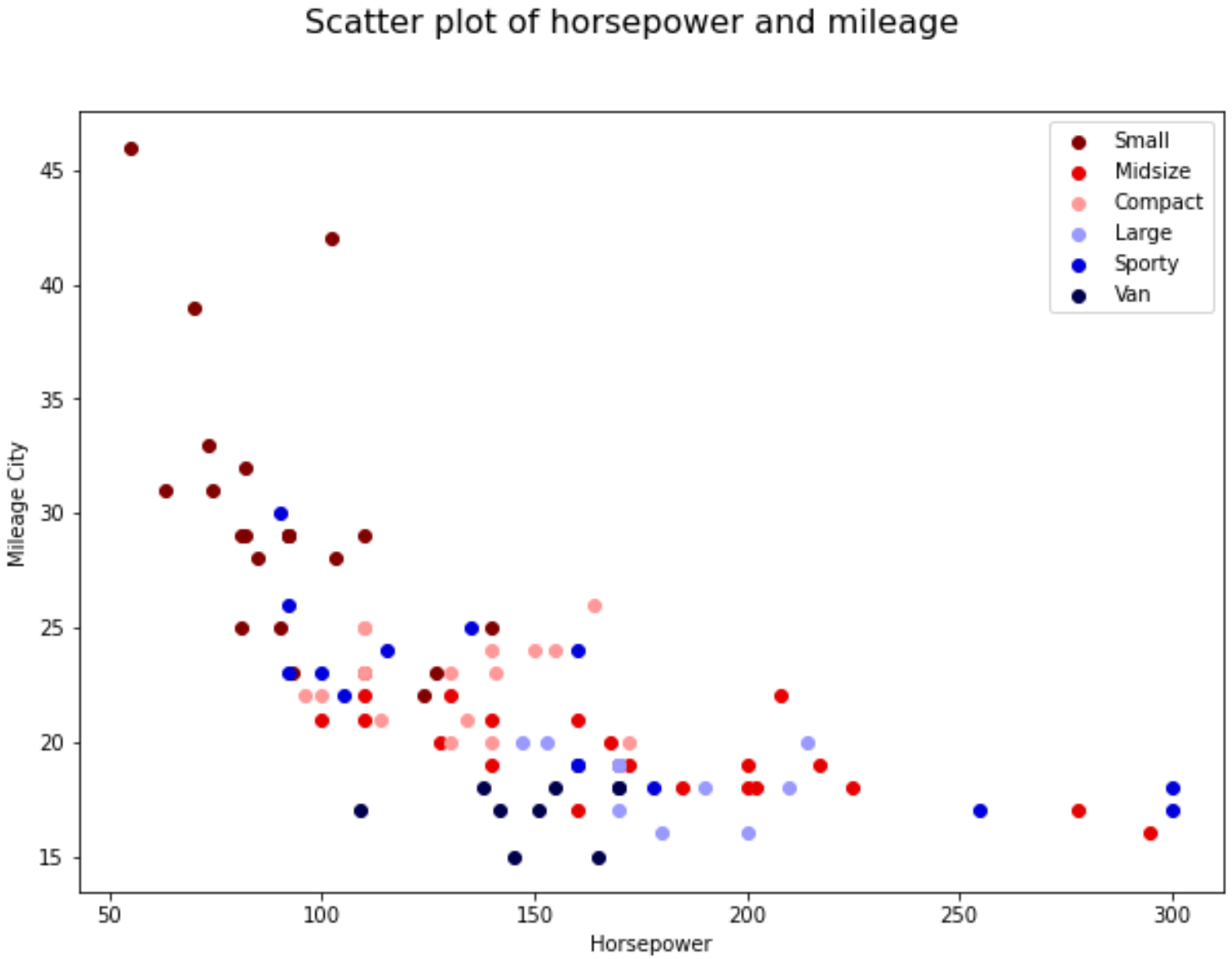
What is the price range of different types of cars available in the market?



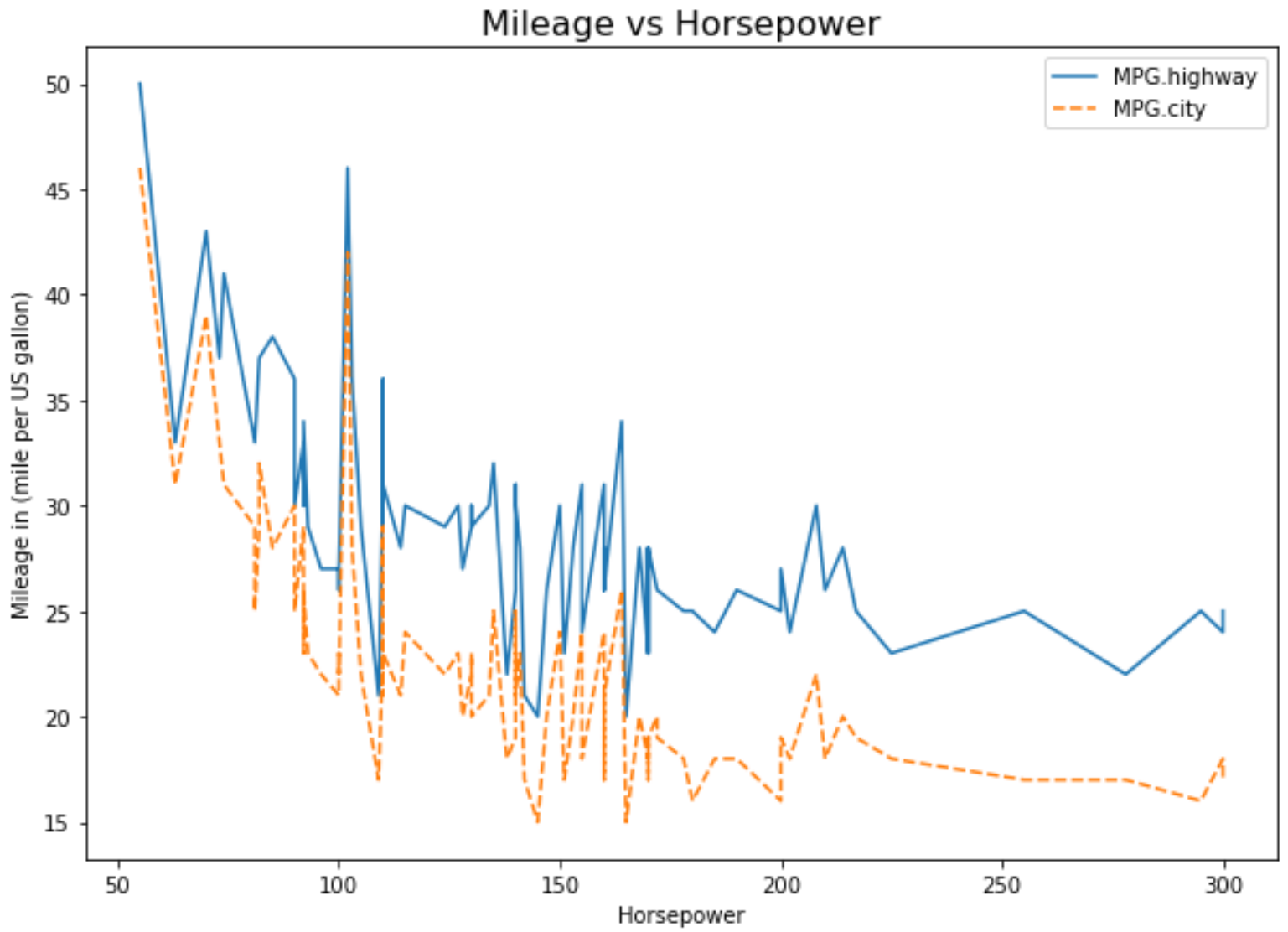
What is the range of horsepower and mileage of various cars?



Does a car with higher horsepower give lower mileage?

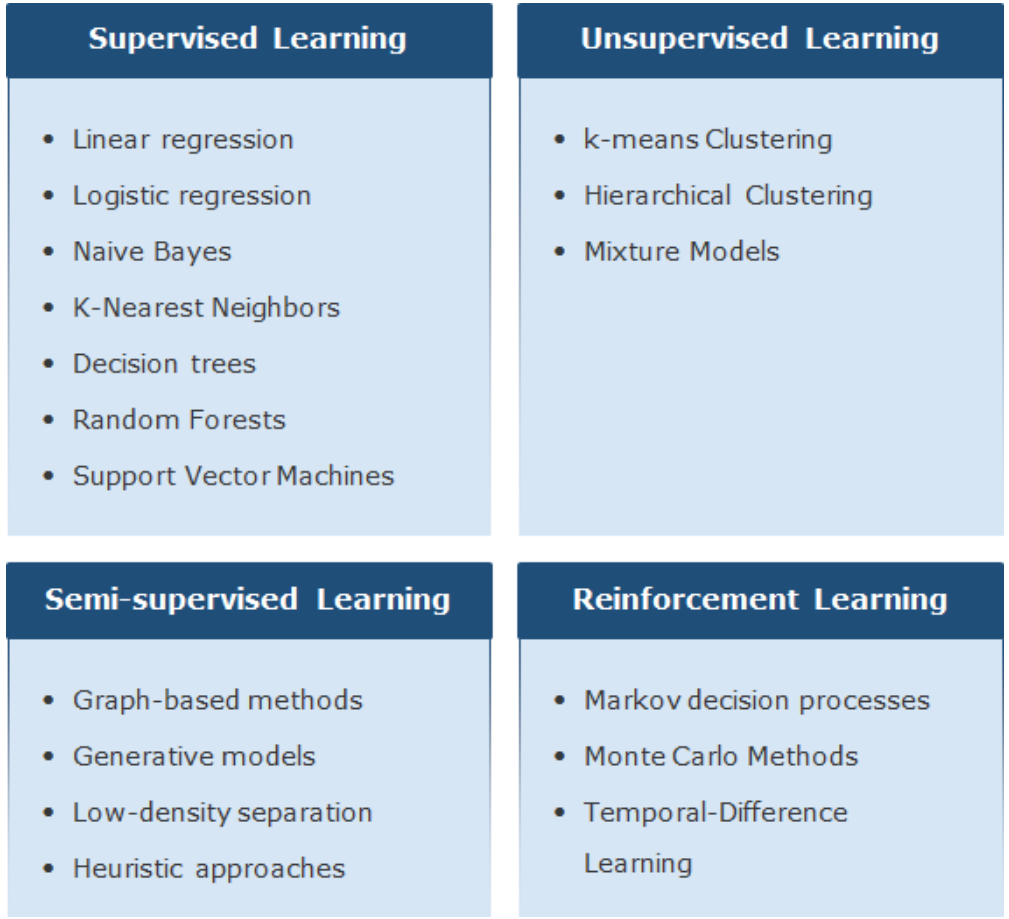


Does a car with higher horsepower give lower mileage?



How many passengers can the car carry based on its type?

famous machine learning algorithms



Tools and packages for Data Science

Programming	Machine Learning Tools
• R • Python • Java • Julia • Fortran • C++ etc.	• R • Python (SciKit-Learn) • Spark • Weka • Julia etc.
Database Management	**Statistical Tools**
• SQL (MySQL, Oracle, DB2, Maria etc.) • NoSQL DBs (Cassandra, MongoDB etc.)	• R • Python (Pandas, Matplotlib, etc.) • Matlab • Julia etc.