Fourth Session

Alireza Moradi

✉ | Linkedin | | k

Exploratory Data Analysis (EDA)

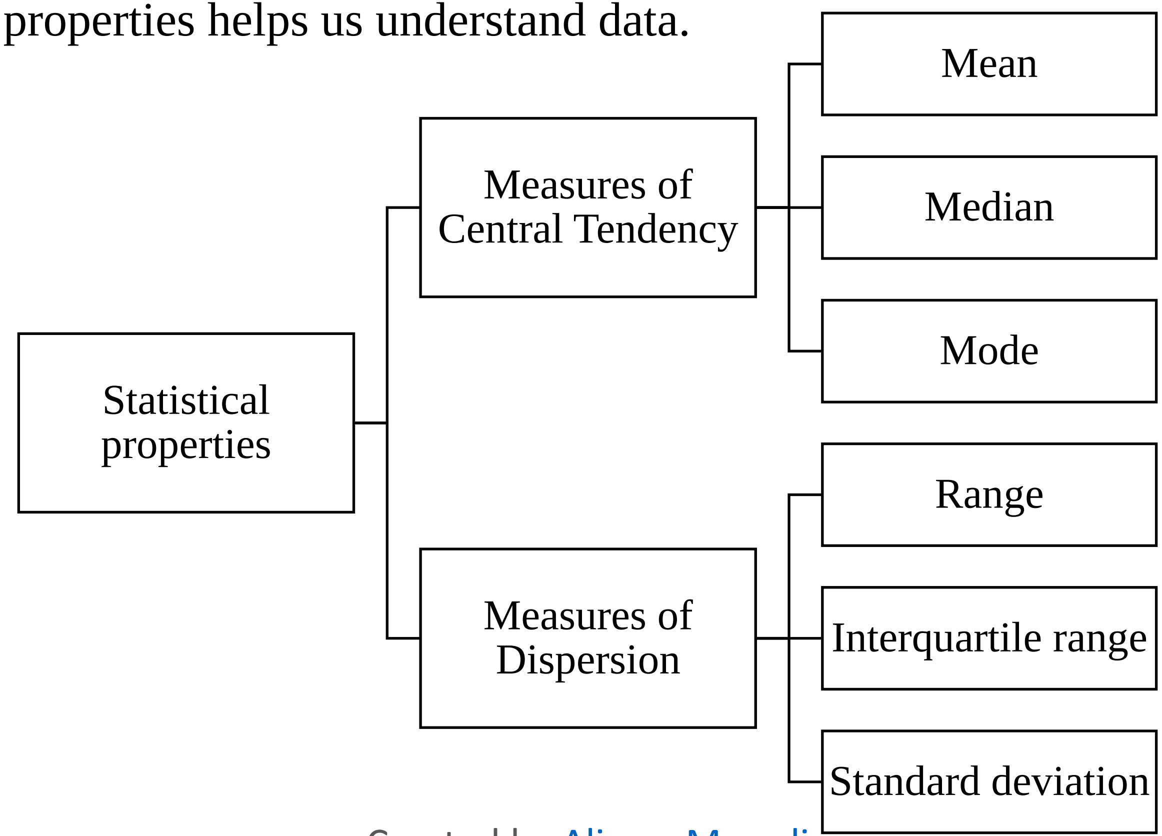
- Exploratory Data Analysis (EDA) is the process of investigating the dataset to discover patterns, anomalies (outliers), and **form hypotheses based on our understanding of the dataset**.

- The main purpose of EDA is to help look at data before making any assumptions.

- EDA involves generating **summary statistics** for numerical data in the dataset and creating **various graphical representations** to understand the data better.

Statistical properties

- Statistical properties helps us understand data.



Statistical properties

Measures of Central Tendency

Mean

Median

Mode

Measures of Dispersion

Range

Interquartile range

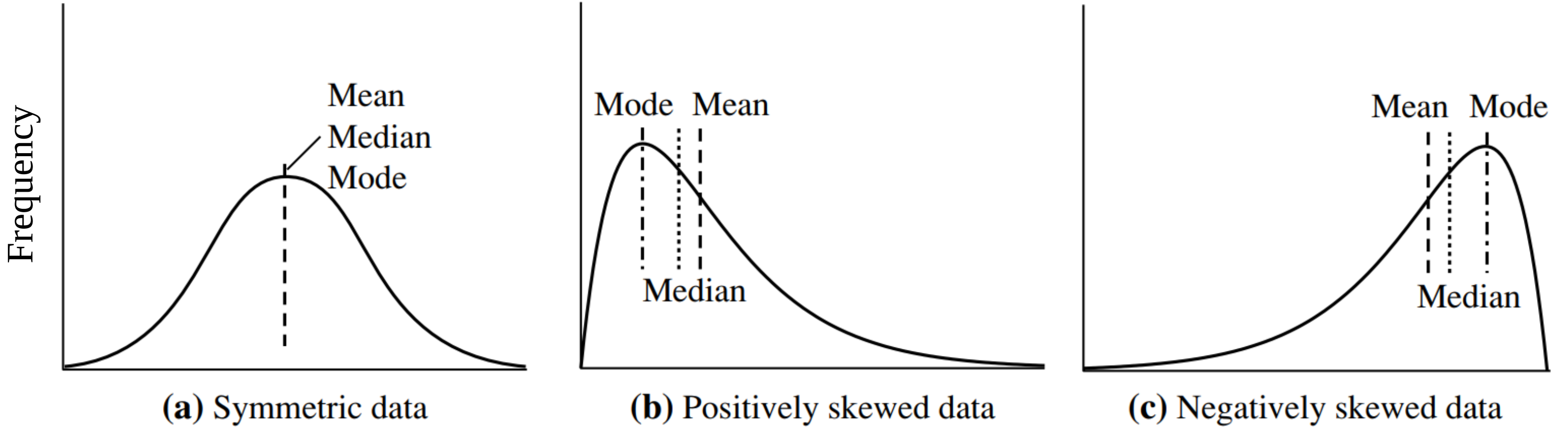
Standard deviation

Created by Alireza Moradi

Measures of Central Tendency

- The 3 most common measures of central tendency are the mean, median and mode.

- The **mode** is the most frequent value. The **median** is the middle number in an ordered data set. The **mean** is the sum of all values divided by the total number of values.



Created by Alireza Moradi

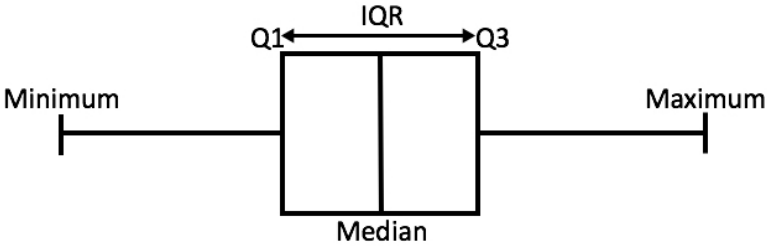
Measures of Dispersion

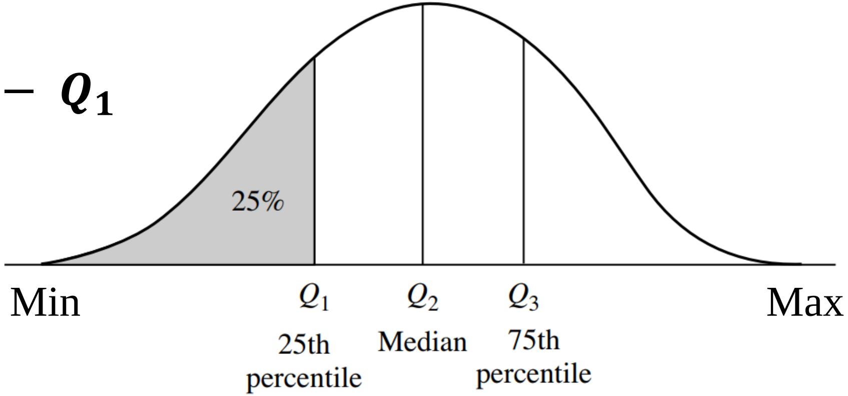
- Measures of dispersion are non-negative real numbers that help to gauge the spread of data about a central value.

- The **Range** is the difference between the lowest and highest values.

- **Quartiles** are three values that split sorted data into four parts, each with an equal number of observations

- interquartile range (IQR): $\boldsymbol{IQR = Q_3 - Q_1}$





Created by Alireza Moradi

Measures of Dispersion

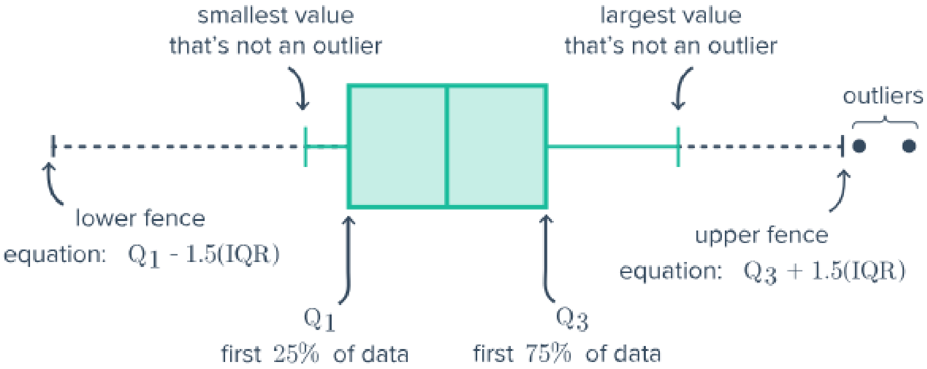
- Five number summary = $(min, Q_1, median, Q_3, max)$

- The upper and lower fences represent the cut-off values for upper and lower outliers in a dataset.

- They are calculated as:

✓ Lower fence $= \; Q_1 - (1.5 \times IQR)$
✓ Upper fence $= \; Q_3 + (1.5 \times IQR)$

- This can form a **box plot**:



Created by Alireza Moradi

Measures of Dispersion

Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Sample Variance

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

Population Standard deviation

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

Sample Standard deviation

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

Data visualization

- Data visualization is the representation of data through use of common graphics, such as charts, plots, and even animations.

- These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.

- We introduce some of them here:
  - ✓ Histogram
  - ✓ Scatter plot
  - ✓ Pie chart
  - ✓ Bar plot
  - ✓ Box Plot
  - ✓ Pair Plot

Data visualization; tools

- Data visualization tools are software applications that transform complex data sets into easily understandable visual representations.

- Some examples are **Power BI**, **Excel**, **tableau**, and **python libraries**.

- Python libraries:
  - ✓ Matplotlib
  - ✓ Seaborn
  - ✓ Plotly ; interactive

Data visualization; Example

- The dataset is available on Kaggle here.

index	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.9
1	18	male	33.770	1	no	southeast	1725.6
2	28	male	33.000	3	no	southeast	4449.4
3	33	male	22.705	0	no	northwest	21984.5
4	32	male	28.880	0	no	northwest	3866.9
5	31	female	25.740	0	no	southeast	3756.6
6	46	female	33.440	1	no	southeast	8240.6
7	37	female	27.740	3	no	northwest	7281.5
8	37	male	29.830	2	no	northeast	6406.4
9	60	female	25.840	0	no	northwest	28923.1

Data visualization; Example

- **age**: age of primary beneficiary (years)
- **sex**: insurance contractor gender, female, male
- **bmi**: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, ideally 18.5 to 24.9 (kg / m ^ 2)
- **children**: Number of children covered by health insurance / Number of dependents
- **smoker**: Smoking
- **region**: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- **charges**: Individual medical costs billed by health insurance ($)

Data visualization; Example

- Let’s first review last session:

Attribute	Type
age	Ratio
sex	Binary
bmi	Ratio
children	Discrete
smoker	Binary
region	Nominal
charges	Ratio

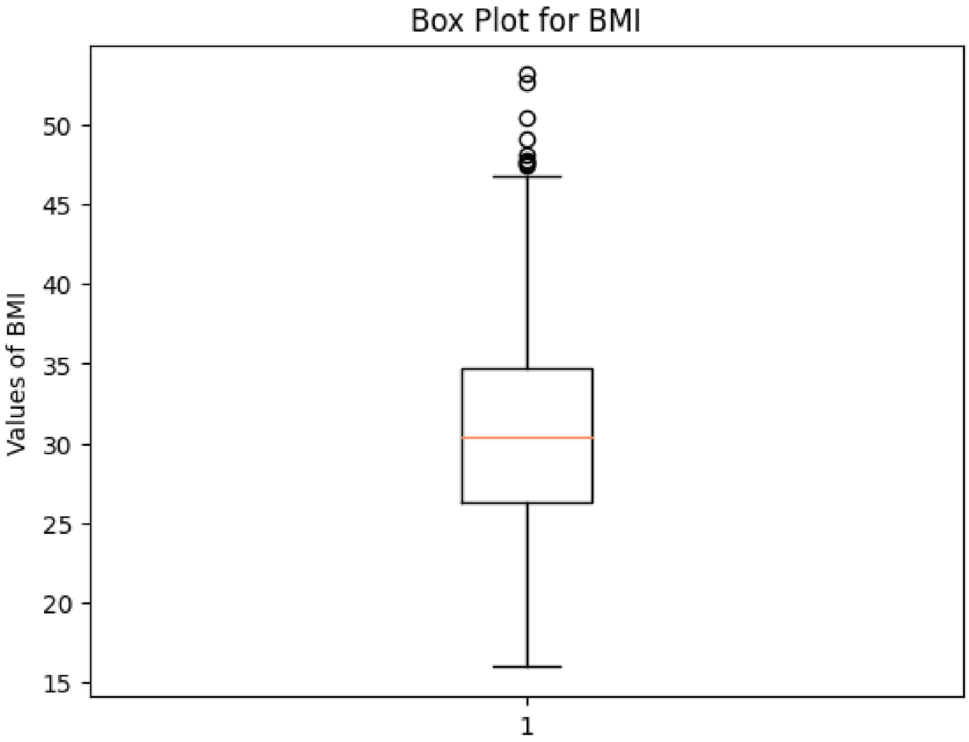
DataFrame.describe()

Statistic	age	bmi	children	charges
count	1337.0	1337.0	1337.0	1337
mean	39.2	30.7	1.1	13279.1
std	14.0	6.1	1.2	12110.4
min	18.0	16.0	0.0	1121.9
25%	27.0	26.3	0.0	4746.3
50%	39.0	30.4	1.0	9386.2
75%	51.0	34.7	2.0	16657.7
max	64.0	53.1	5.0	63770.4

Data visualization; Box plot

- **Box plot** is a graph summarizing a set of data.

- The shape of the boxplot shows how the data is distributed and it also shows any outliers.

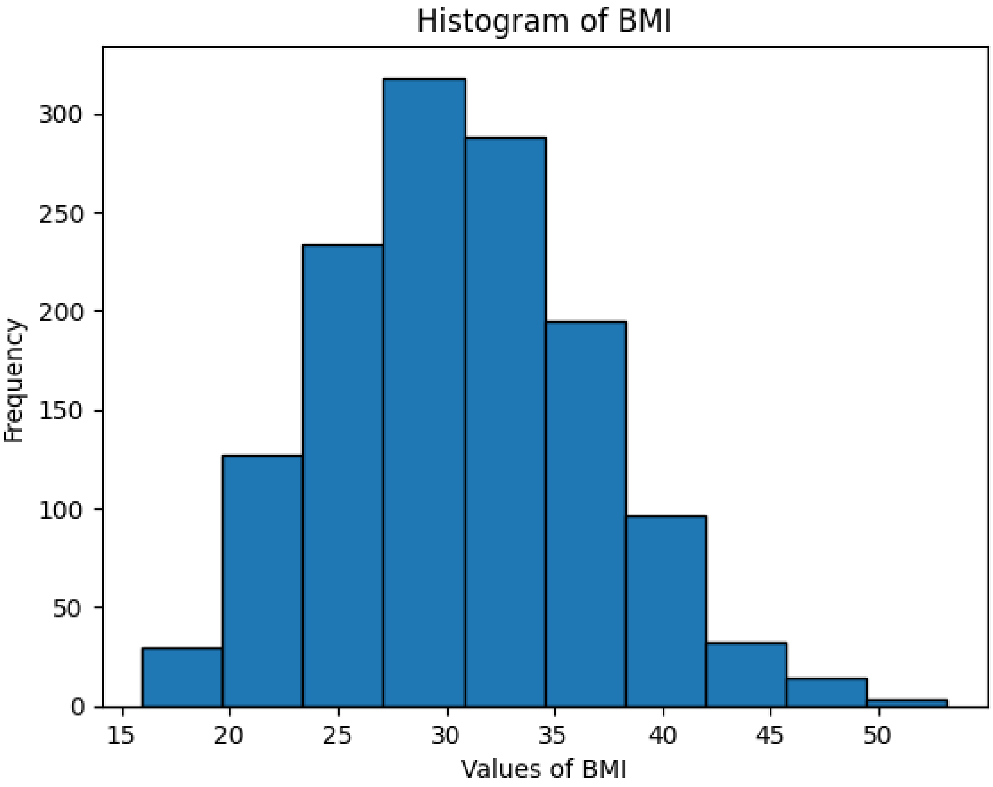


Created by Alireza Moradi

Data visualization; Histogram

- A **histogram** is a graph that shows the frequency of numerical data using rectangles.

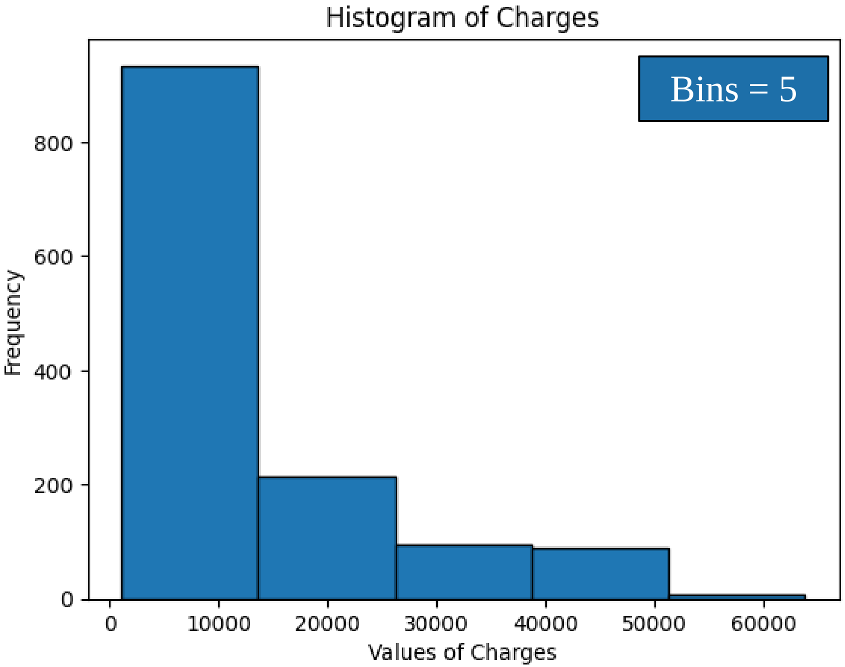
- The height of a rectangle (the vertical axis) represents the distribution frequency of a variable (the amount, or how often that variable appears).

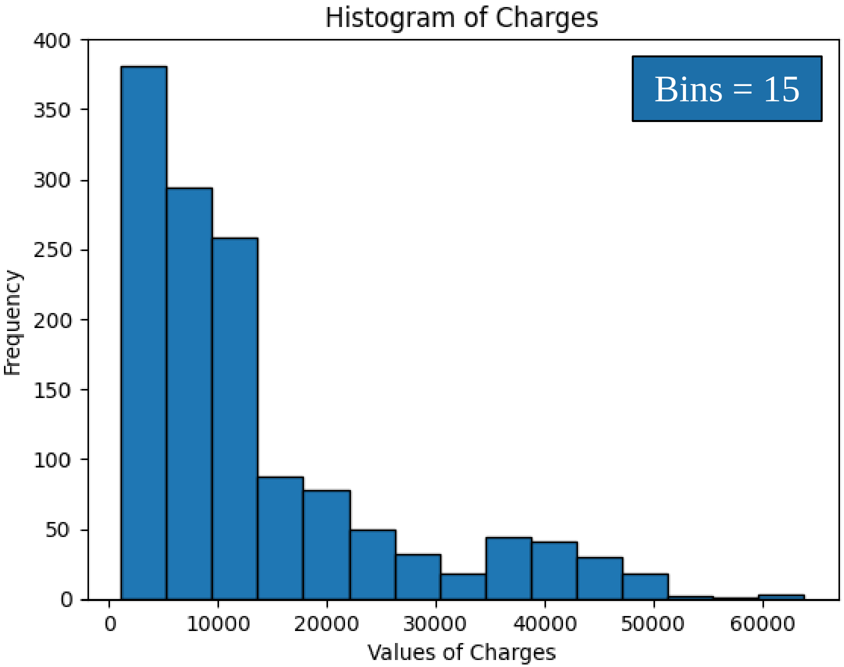


Created by Alireza Moradi

Data visualization; Histogram

- Explore the data in greater detail by increasing the number of **bins**.



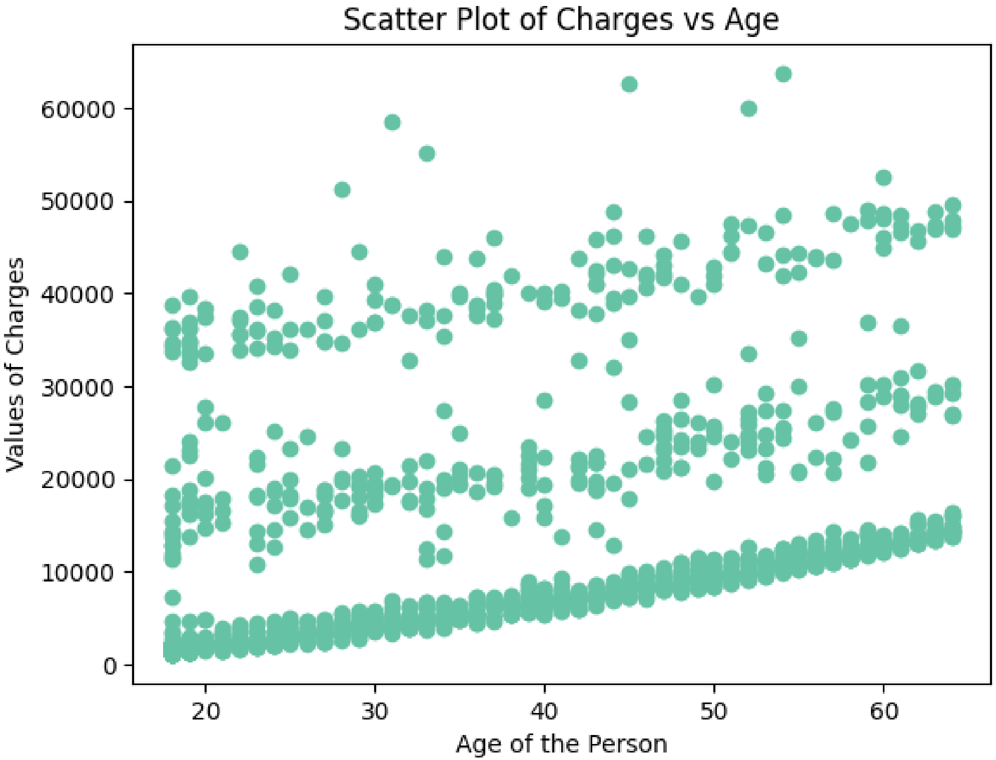


Created by Alireza Moradi

Data visualization; Scatter plot

- **Scatter plots** are the graphs that present the relationship between two variables in a data-set.

- It represents data points on a two-dimensional plane or on a Cartesian system.

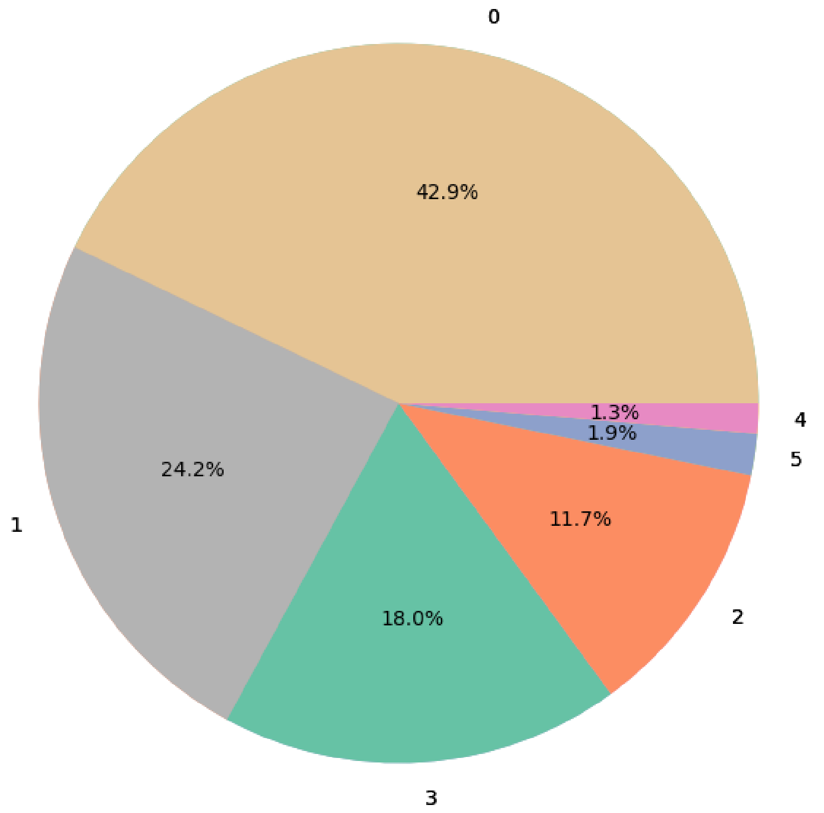


Created by Alireza Moradi

Data visualization; Pie chart

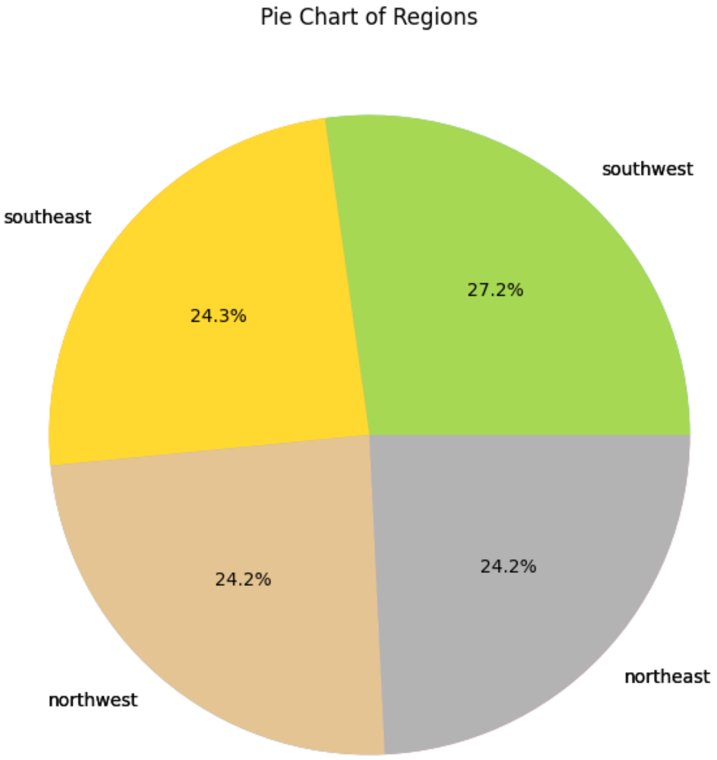
- A **pie chart** (or a **circle chart**) is a circular statistical graphic which is divided into slices to illustrate numerical proportion.

- This shows pie chart for number of children.

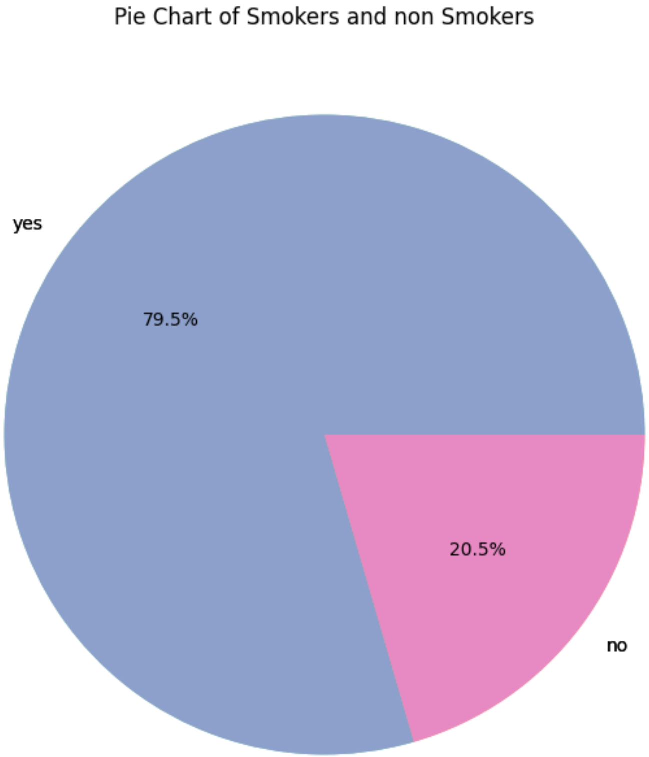


Created by Alireza Moradi

Data visualization; Pie chart



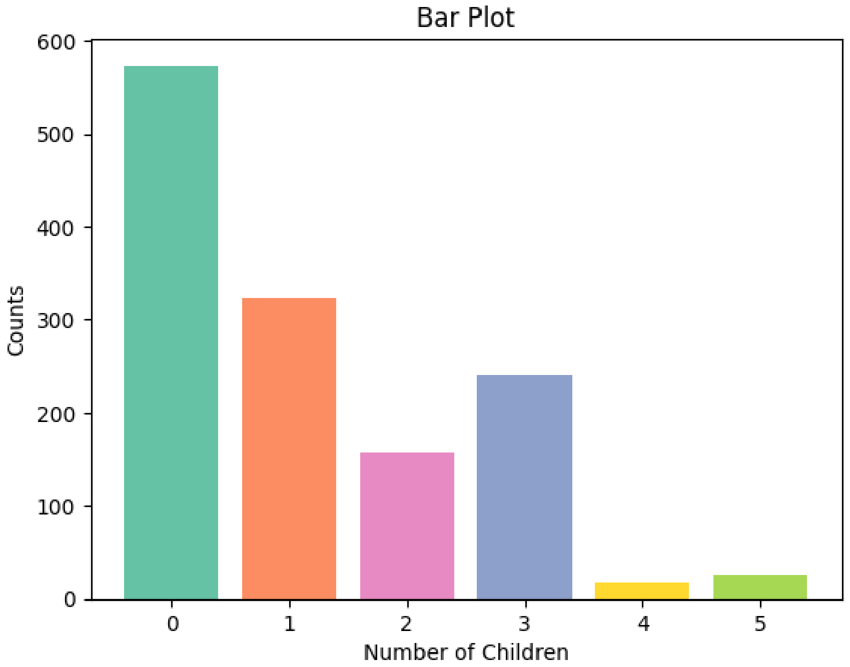
Pie Chart of Regions



Pie Chart of Smokers and non Smokers

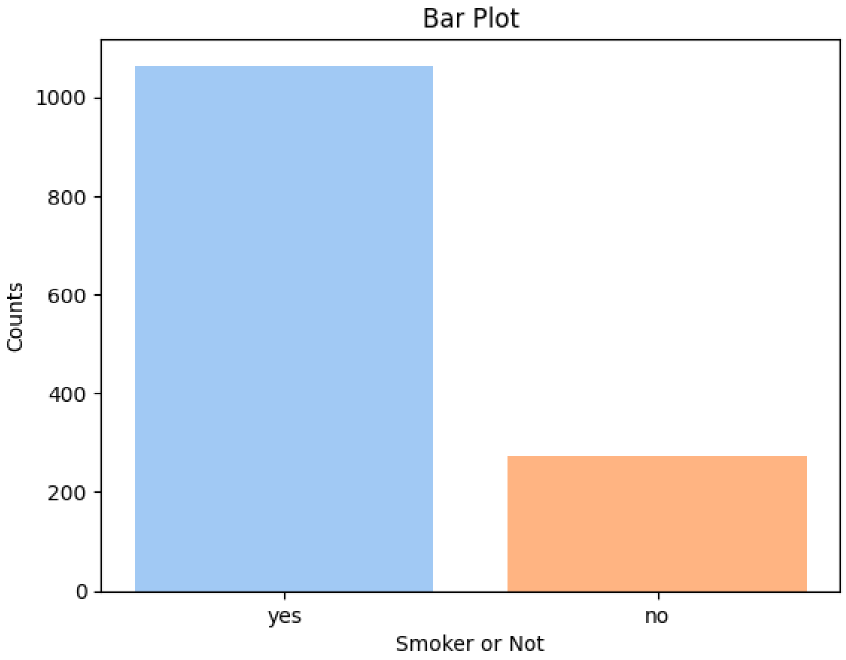
Data visualization; Bar plot

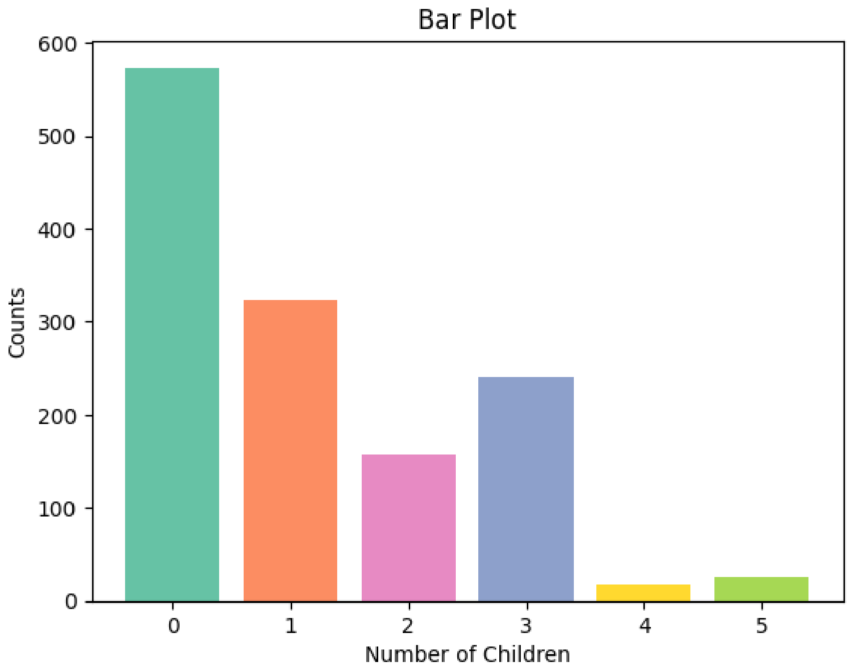
- A **bar plot** (or **bar chart**) is one of the most common types of graphic.
- It shows the relationship between a numeric and a categoric variable.
- Each entity of the categoric variable is represented as a bar. The size of the bar represents its numeric value.



Data visualization; Bar plot

- Two bar plots for *smoker* and *children*.



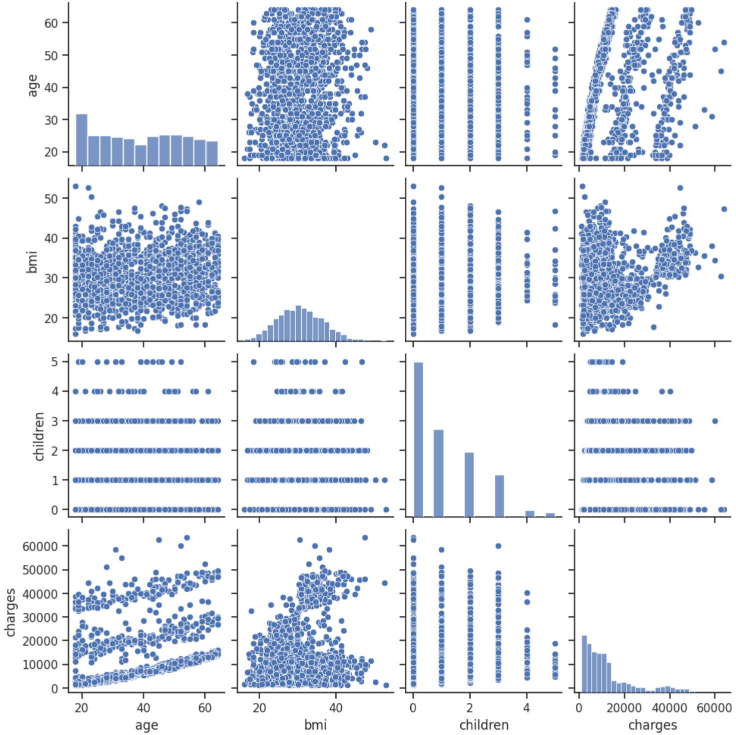


Created by Alireza Moradi

Data visualization; Pair plot

- A pairs plot allows us to see both distribution of single variables and relationships between two variables .

Histogram Or bar plot	Scatter plot	Scatter plot	Scatter plot
Scatter plot	Histogram Or bar plot	Scatter plot	Scatter plot
Scatter plot	Scatter plot	Histogram Or bar plot	Scatter plot
Scatter plot	Scatter plot	Scatter plot	Histogram Or bar plot



Business intelligence (BI)

- BI refers to the strategies and technologies used by organizations to analyze and manage business data.

- It helps them gain insights into their operations, customers, and market trends, enabling them to make better decisions.

- **key performance indicator** (KPI) is a quantifiable measure of performance over time for a specific objective. For example, *number of new customers in a month*.

- BI reports on different KPIs through:
  1. Ad-hoc reports
  2. Dashboards

BI; Ad-hoc reports

- Ad-hoc reporting is a process that involves manually processing data to produce custom-made reports.

- The focus of an ad-hoc report is usually on a specific area of the business, such as its marketing or supply chain management.

- **Storytelling** and **data visualization** are essential elements of effective BI reports.

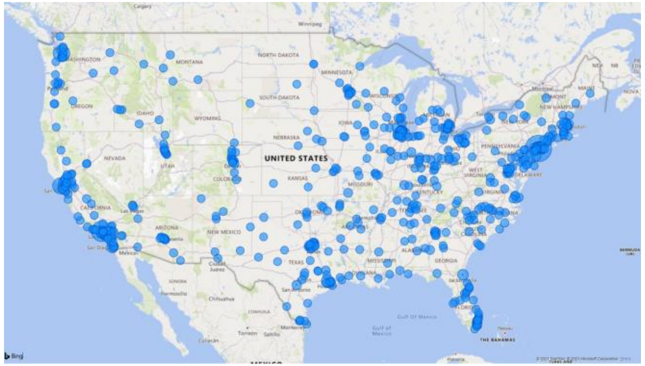
BI; Dashboard

- A dashboard is a way of displaying various types of visual data in one place. Usually, a dashboard is intended to convey different, but related information in an easy-to-digest form.

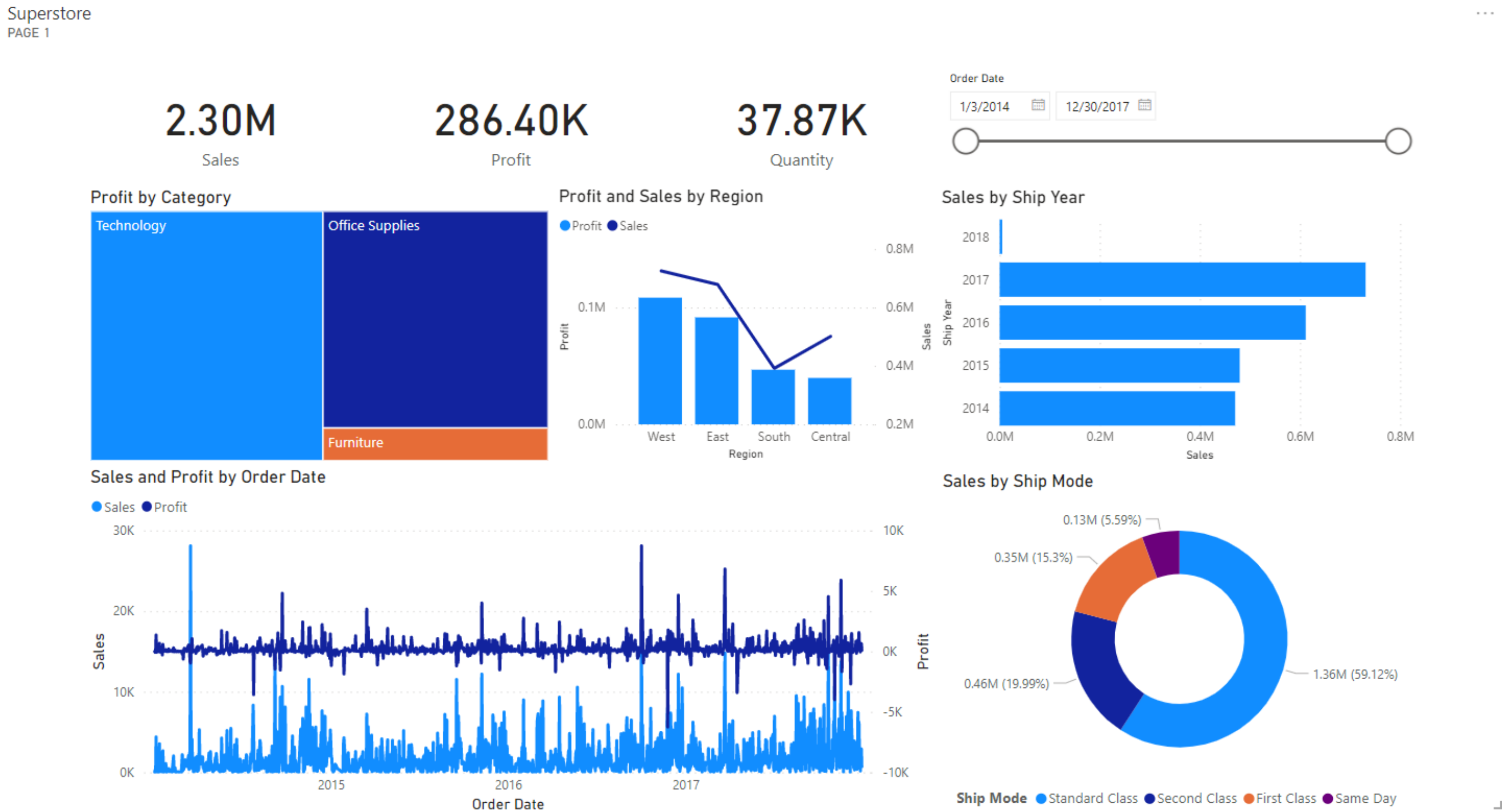
- Dashboards provide a holistic view of key business areas.

- The presentation of data on dashboards is graphical in nature, using bar charts, pie charts, maps and gauges.

gauge



BI; Dashboard


Superstore
PAGE 1
...
2.30M
Sales
286.40K
Profit
37.87K
Quantity
Order Date
1/3/2014
12/30/2017
Profit by Category
Technology
Office Supplies
Furniture
Profit and Sales by Region
● Profit ● Sales
0.1M
0.0M
Profit
West
East
South
Central
Region
0.8M
0.6M
0.4M
0.2M
Sales
Sales by Ship Year
2018
2017
2016
2015
2014
Ship Year
0.0M
0.2M
0.4M
0.6M
0.8M
Sales
Sales and Profit by Order Date
● Sales ● Profit
30K
20K
10K
0K
Sales
2015
2016
2017
Order Date
10K
5K
0K
-5K
-10K
Profit
Sales by Ship Mode
0.13M (5.59%)
0.35M (15.3%)
0.46M (19.99%)
1.36M (59.12%)
Ship Mode ● Standard Class ● Second Class ● First Class ● Same Day