University of BRISTOL

Exploratory data analysis (EDA)

EMATM0048- Teaching session 9

Dr. Zahraa Abdallah

Last Session!

- What are the different types of data.
  - Structured vs unstructured
  - Quantitative vs qualitative
  - Discrete vs continuous

- Different sources of data
  - APIs
  - Web scraping

Today's session

- First step in analysis: Exploratory data analysis (EDA)
- Graphical and non-graphical EDA
- Relationships among variables

Readings

- Experimental Design and Analysis, Howard J. Seltman
  - Chapter 4 http://www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf

Exploratory data analysis

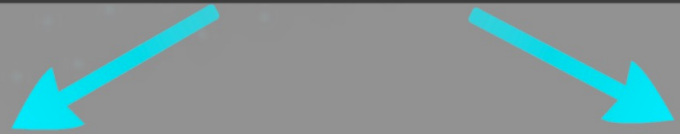
- Exploratory data analysis or “EDA” is a critical first step in analysing the data.
- Here are the main reasons we use EDA:
  - Catching mistakes and anomalies
  - Gaining new insights into data
  - Detecting outliers in data
  - Testing assumptions
  - Identifying important factors in the data
  - Understanding relationships

Data Types (Recap)

- It helps to know the different types of data we are working with. Different data types often require different techniques.
  - Nominal data: Data with categories that don't have any natural order attached to them. For example –gender, hair colour, eye colour
  - Scale data: Strictly numerical data that is ordered against a constant and consistent scale. For example –height, time, specific age.

Exploratory Data Analysis

Exploratory data analysis (EDA) is an approach to process data to summarise their main characteristics, often with visual and non-visual methods. Exploratory data analysis was promoted by John Tukey.



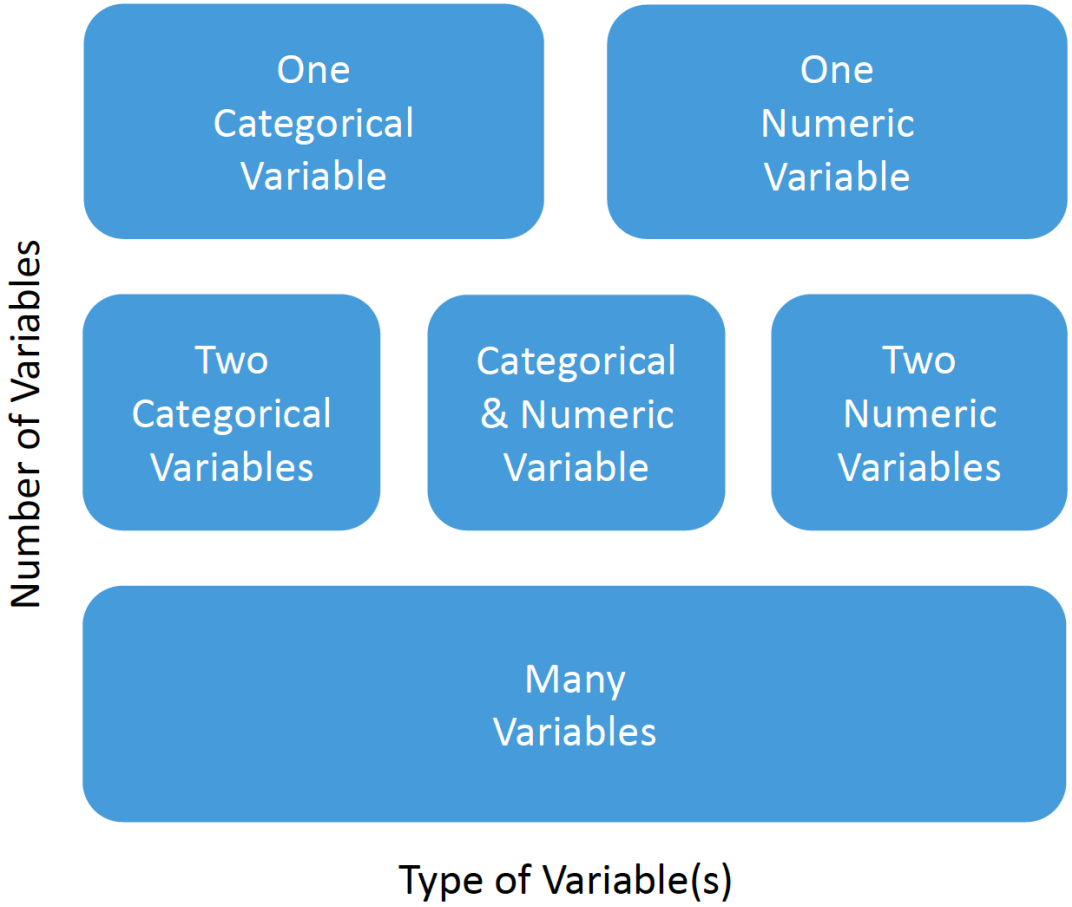
Non-Graphical

- Central Location
- Variability
- Correlation
- .....

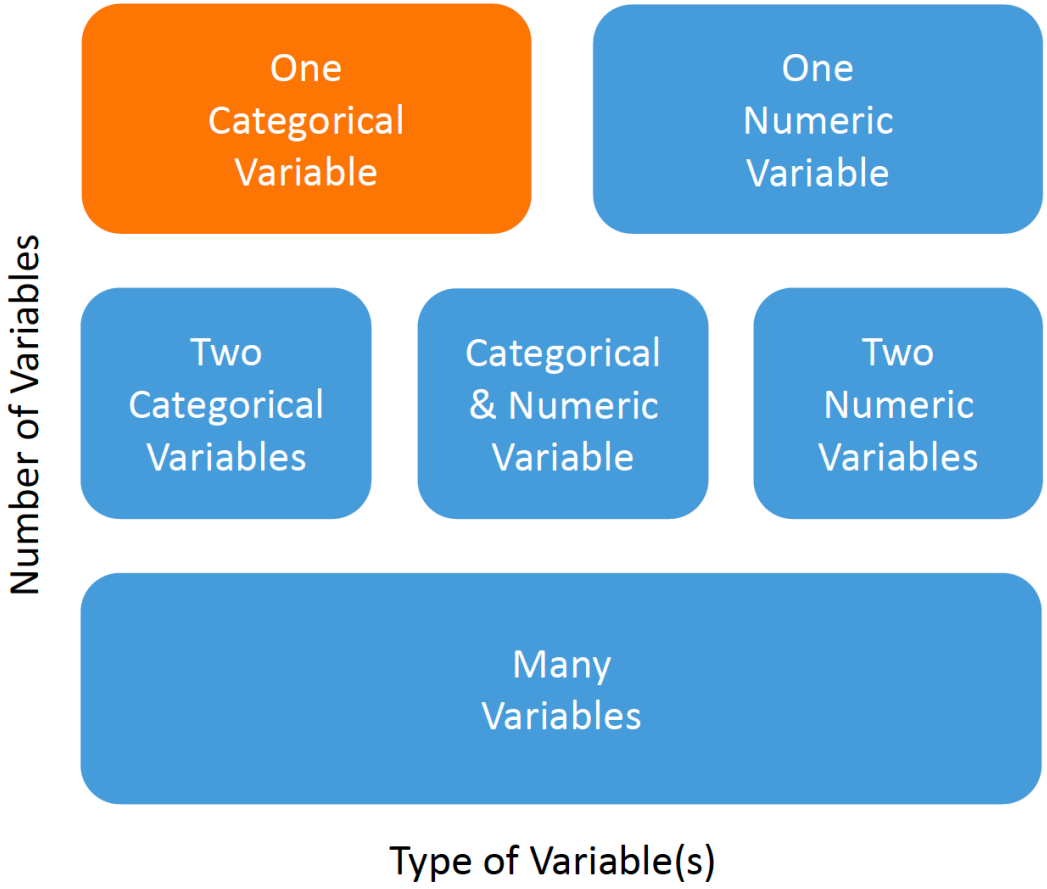
Graphical

- Scatter plots
- Box plots
- Histograms
- ...

Exploratory data analysis



Exploratory data analysis



One Categorical Variable

When you are dealing with nominal data, you collect information through:

- **Frequencies**: The Frequency is the rate at which something occurs over a period of time or within a dataset.
- **Proportion**: You can easily calculate the proportion by dividing the frequency by the total number of events. (e.g how often something happened divided by how often it could happen)
- **Percentage.**

One Categorical Variable

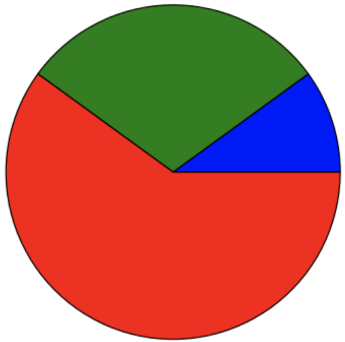
- A simple tabulation of the frequency of each category is the best univariate non-graphical EDA for categorical data.

Statistic/College	H&SS	MCS	SCS	other	Total
Count	5	6	4	5	20
Proportion	0.25	0.30	0.20	0.25	1.00
Percent	25%	30%	20%	25%	100%

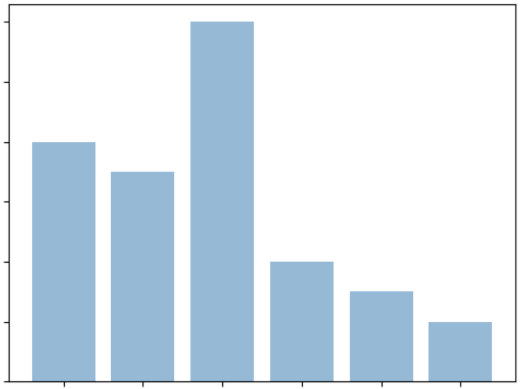
Graphical or non Graphical

One Categorical Variable

Pie Chart

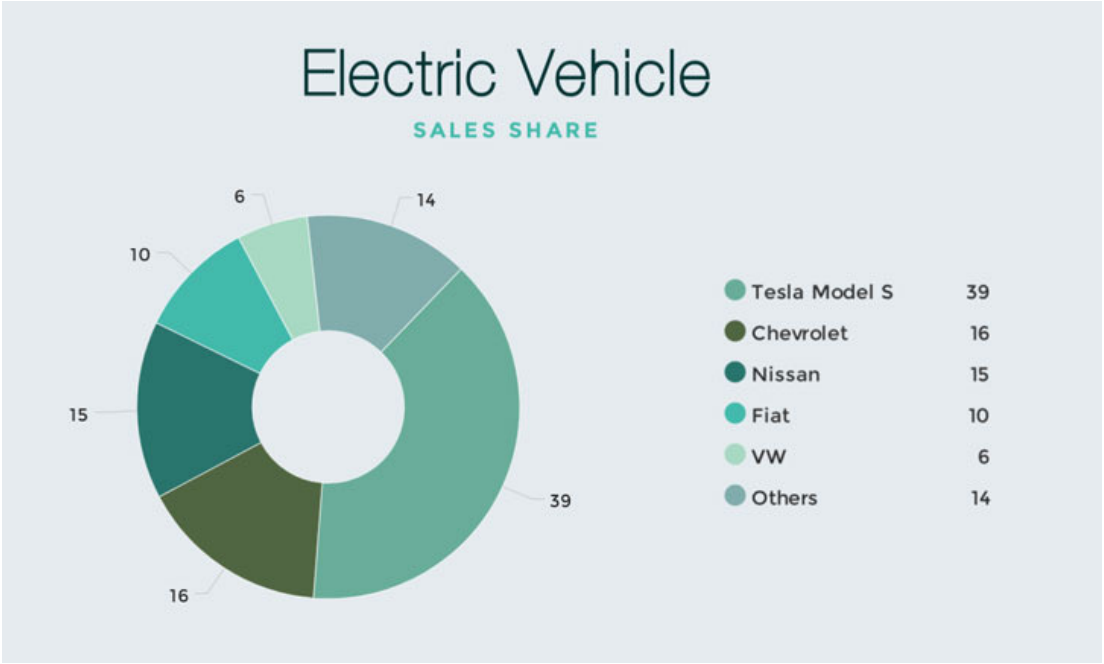


Bar Chart



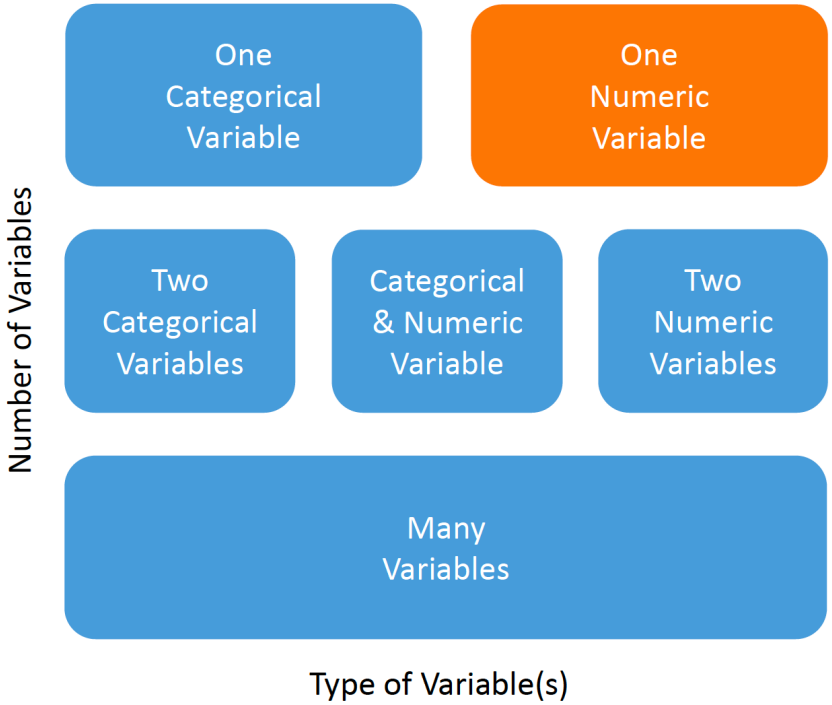
Composition with pie charts

- Pie charts show categories as a proportion or a percentage of the whole.
- Use pie charts to show the composition of categorical data with each segment proportional to the quantity it represents.



(source: https://www.visme.co/pie-chart-maker/)

Exploratory data analysis

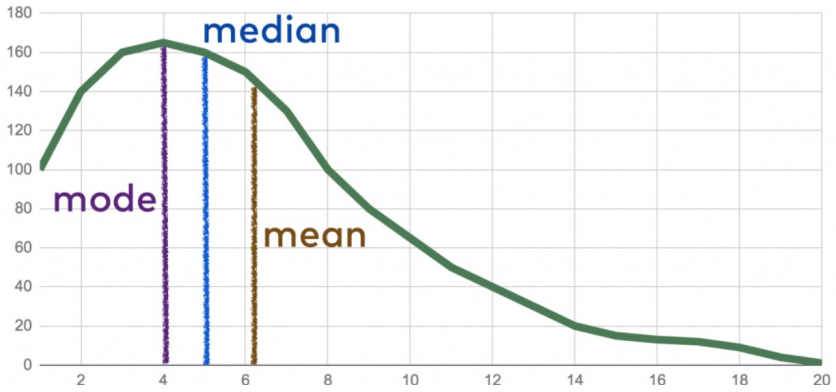


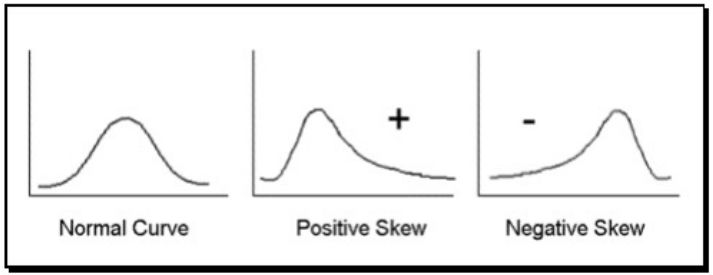
One Numerical Variable (Non-Graphical):

Preliminary assessments of the population distribution of the variable using the data of the observed sample.

– Centrality:

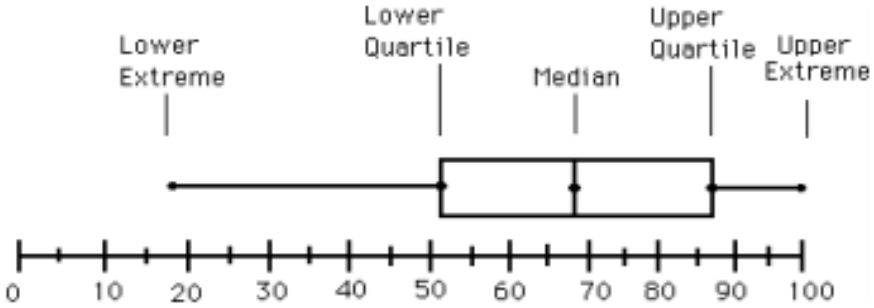
- The most common measure of central tendency is the mean.
- For skewed distribution or when there is concern about outliers, the median may be preferred.
- Mode can also be used.





One Numerical Variable (Non-Graphical):

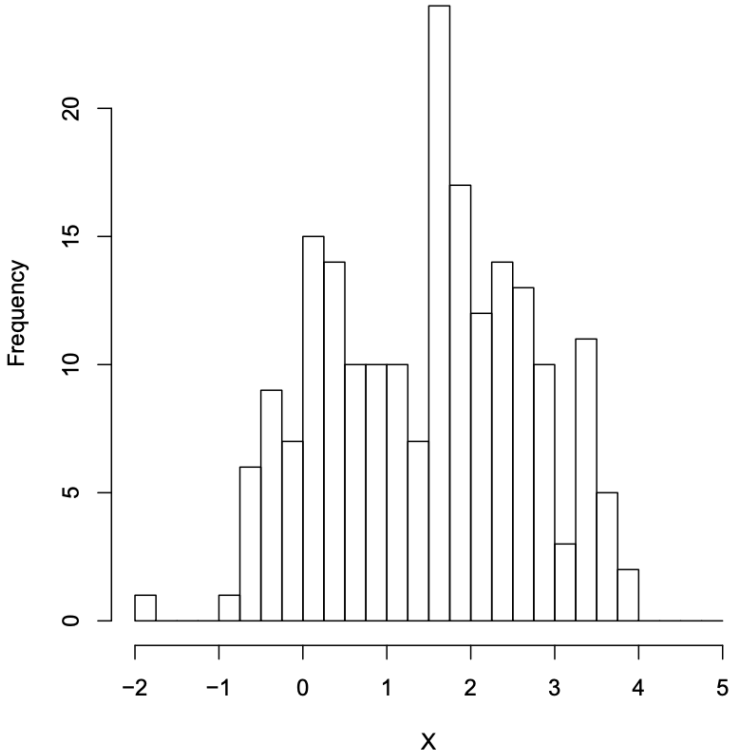
- Spread:
  - Range: Difference between Max and Min
  - The variance and standard deviation are two useful measures of spread
  - The interquartile range (IQR) is a robust measure of spread: Difference between the values of 25% and 75% of the distribution.



One Numerical Variable (Graphical):

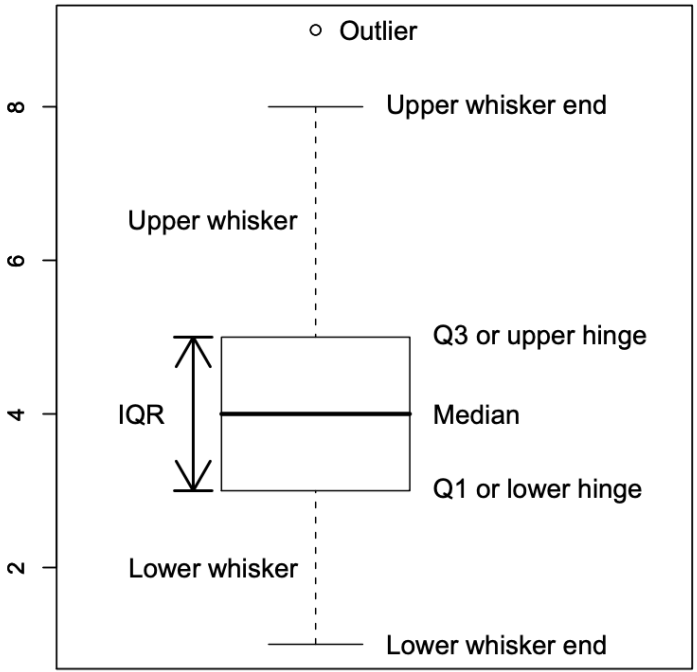
- Histogram:

With practice, histograms are one of the best ways to quickly learn a lot about your data, including central tendency, spread, modality, shape and outliers.

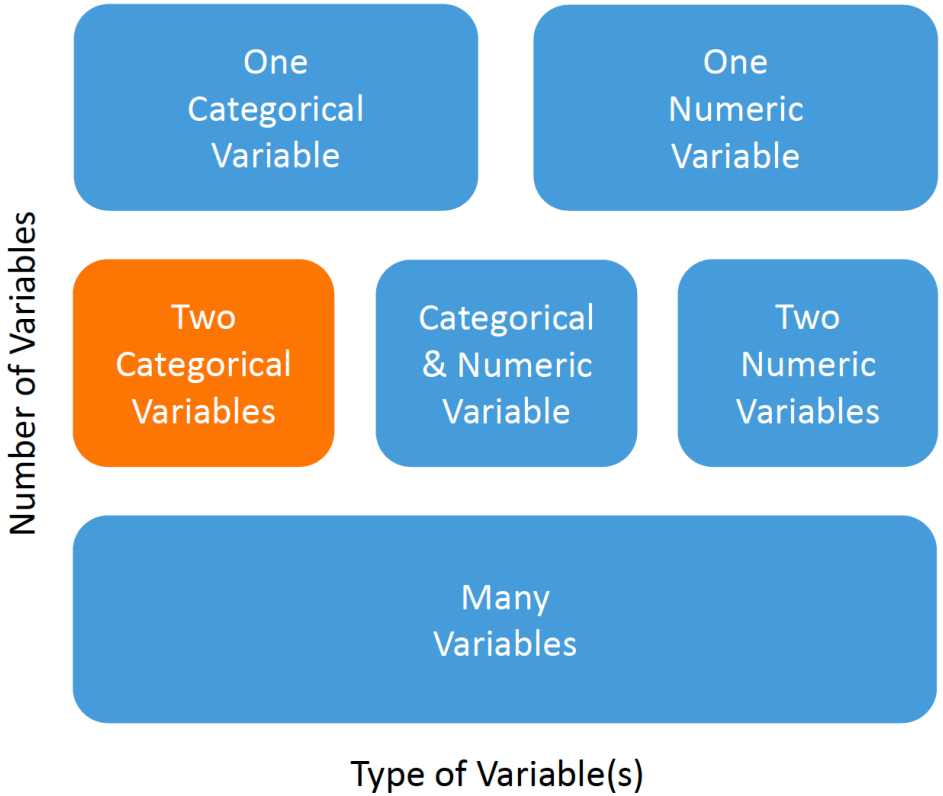


One Numerical Variable (Graphical):

- Boxplots show robust measures of location and spread as well as providing information about symmetry and outliers.



Exploratory data analysis

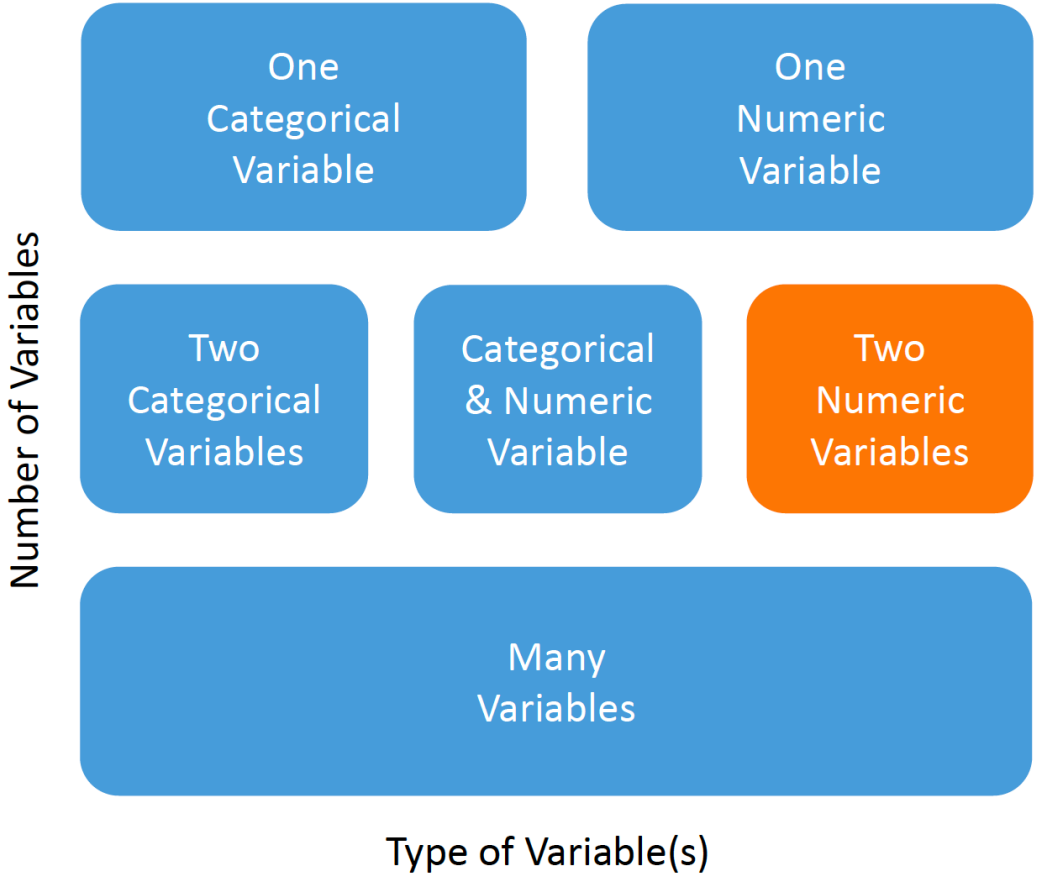


Two categorical variables

Cross -tabulation

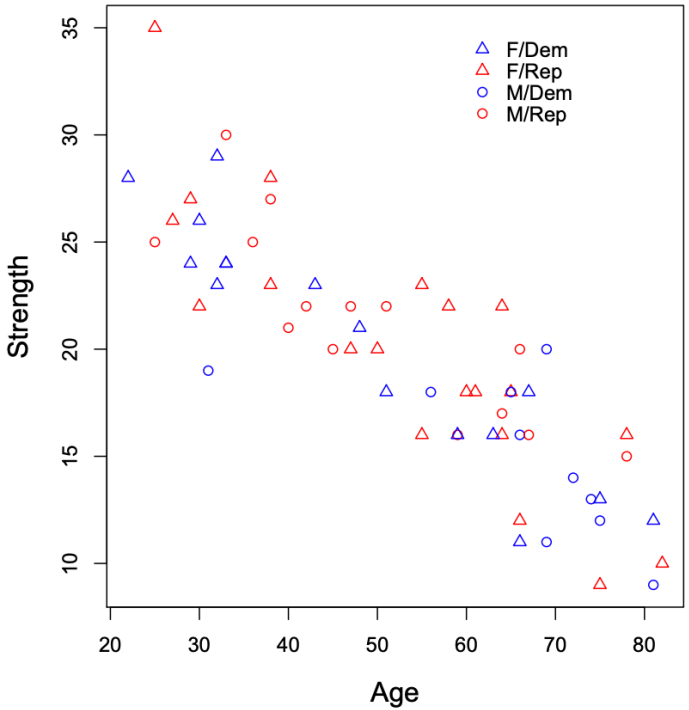
Movies by Genre and Rating					
Genre	G	PG	PG-13	R	Total
Action	2	70	311	229	612
Adventure	44	179	209	64	496
Animation	43	111	8	6	168
Comedy	45	258	472	506	1218
Drama	12	136	586	836	1570
Family	38	181	10	1	230
…	…	…	…	…	…
Total	230	1207	2686	3058	7181

Exploratory data analysis



Two Numerical Variables (Graphical)

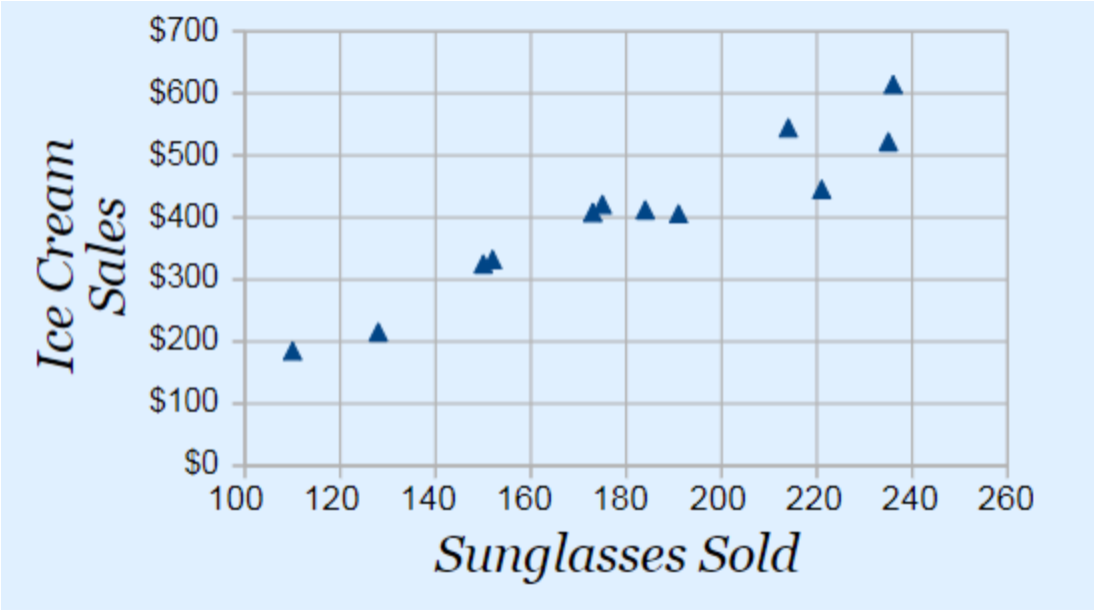
- For two quantitative variables, the basic graphical EDA technique is the scatterplot which allows us to see relationships between two scale variables.
- If one variable is explanatory and the other is outcome, it is a very strong convention to put the outcome on the y (vertical) axis.



Correlation

- **Correlation** is a statistical technique which tells us how strongly the pair of variables are linearly related and change together.

- **Example:** Correlation between Ice cream sales and sunglasses sold.



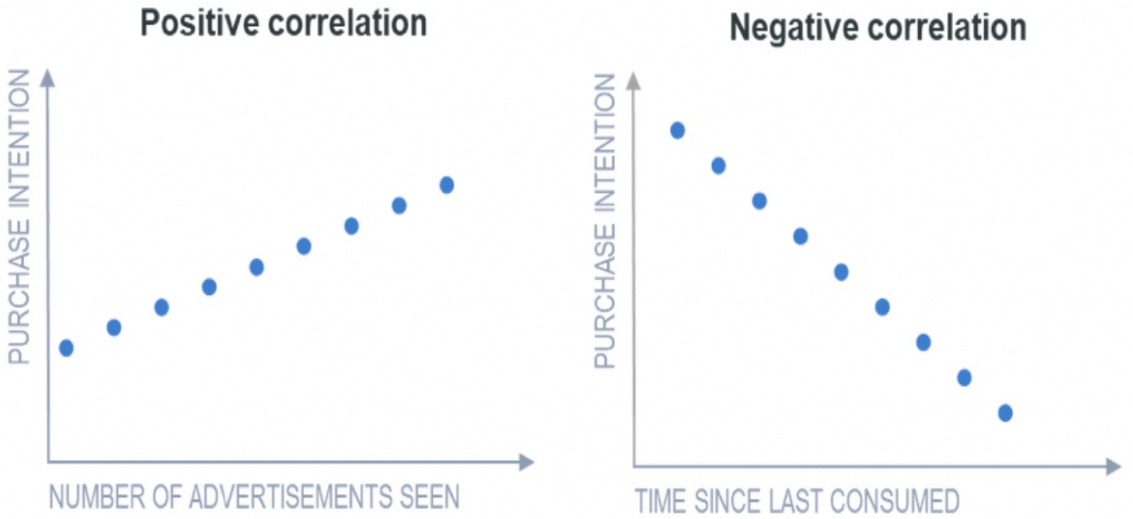
Correlation Coefficient

- The most well-known correlation coefficient is the Pearson's product moment coefficient.
- The correlation between two random variables is a number that runs from -1 through 0 to +1 and indicates a strong inverse relationship, no relationship, and a strong direct relationship, respectively.

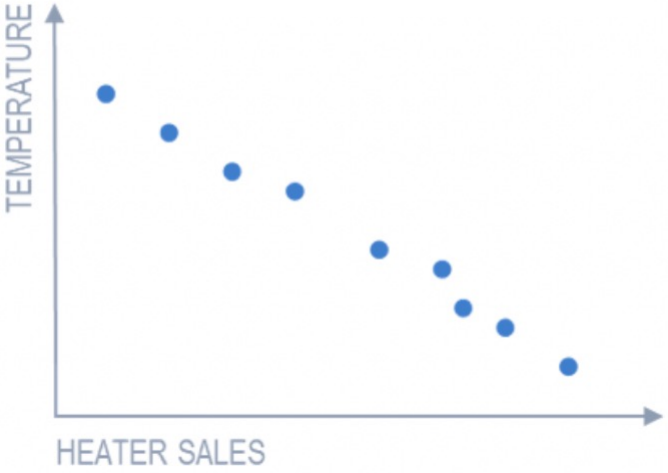
$$r_{A,B} = \frac{\sum_{i=1}^{m}(a_i - \bar{A})(b_i - \bar{B})}{m\sigma_A\sigma_B}$$

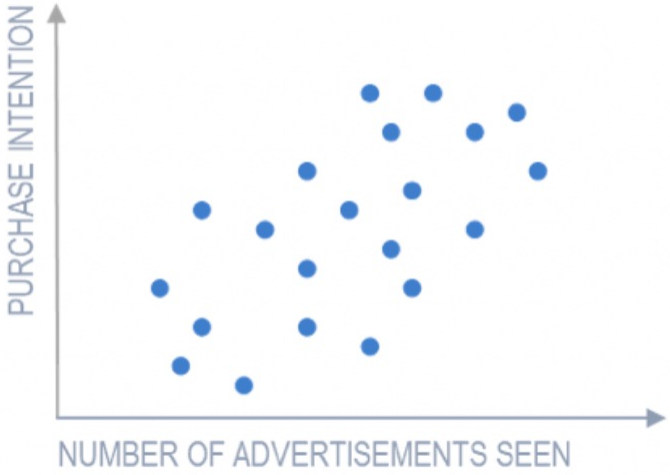
- If is **positive** that means both attributes are positively correlated. This indicates that if **one attribute increased the other will increase too.**
- **Zero** value means **the two attributes are independent.**
- Positive or negative coefficient does not say that any of them can possibly replace the other.

Correlation Coefficient



What type of correlation? Pos/Neg … Strong/weak





Correlation and causation

- **Causation** takes a step further than correlation. It says any change in the value of one variable will **cause** a change in the value of another variable, which means one variable makes other to happen. It is also referred as cause and effect.

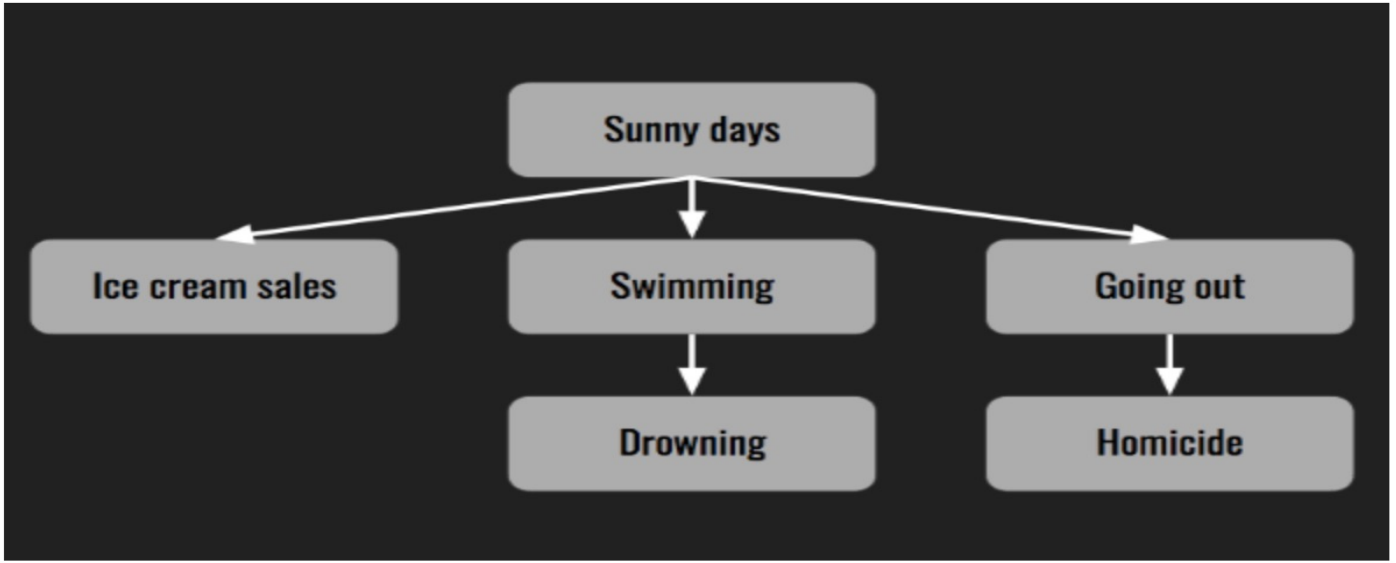
- When two attributes are found correlated that doesn't mean one attribute is causing the occurrence of the other.

Example

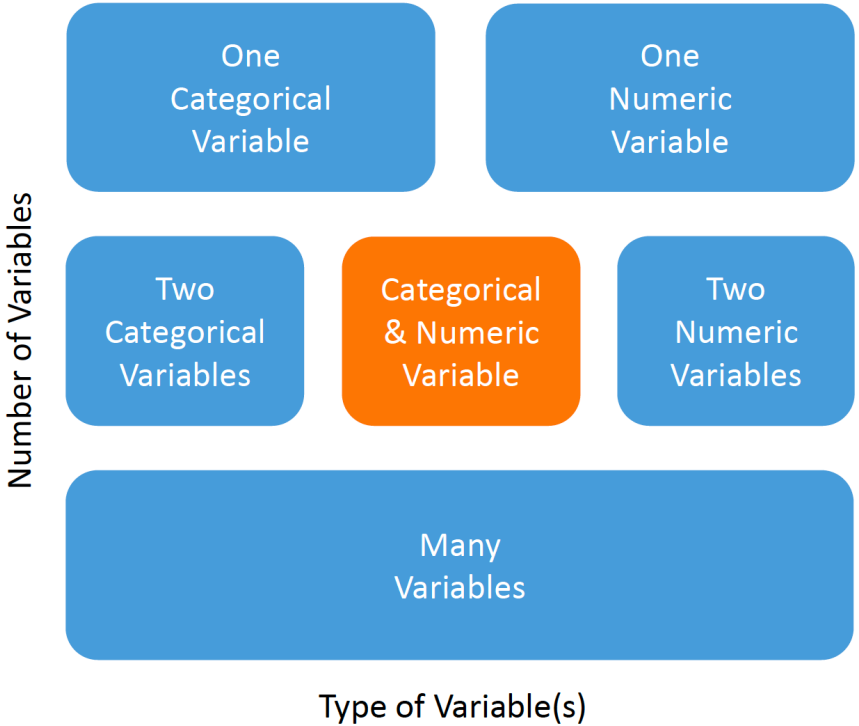
- ***Ice cream sales is correlated with homicides in New York (Study)***

Example

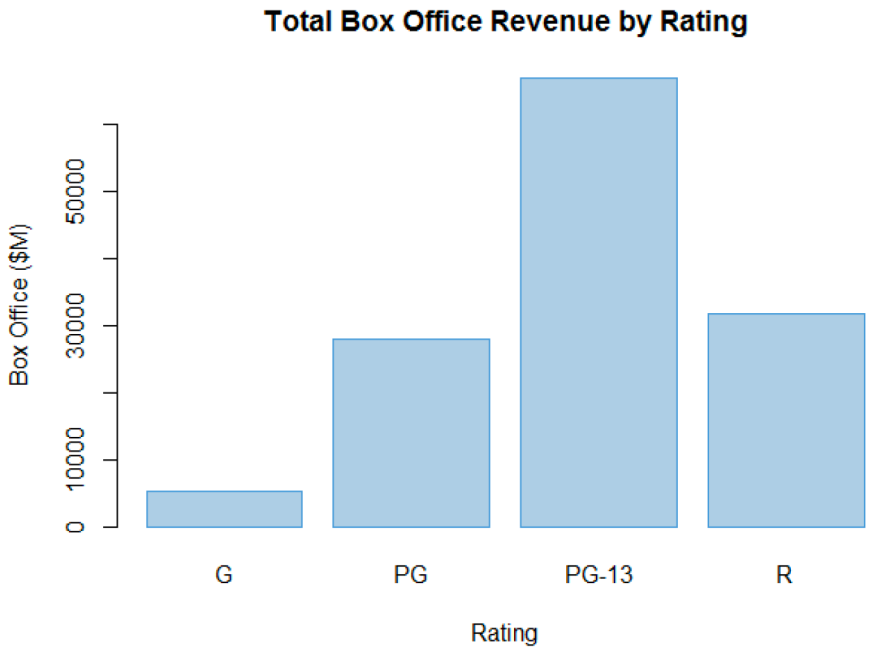
- ***Ice cream sales is correlated with homicides in New York (Stu***



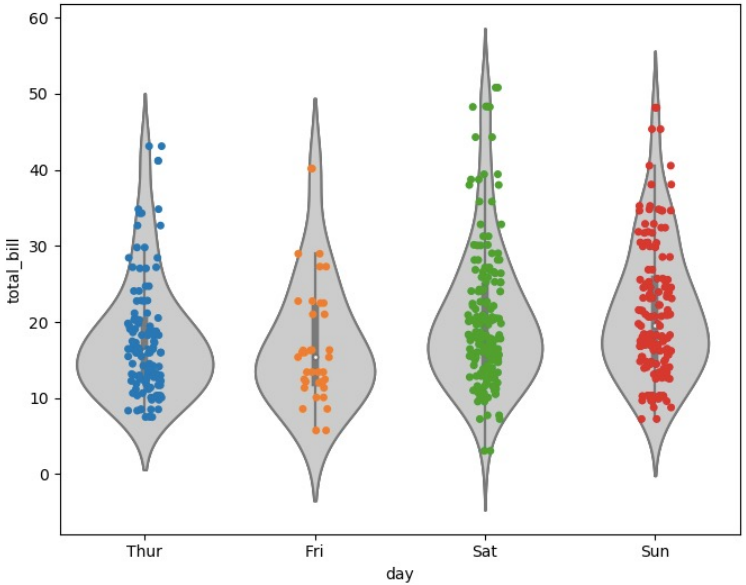
Exploratory data analysis



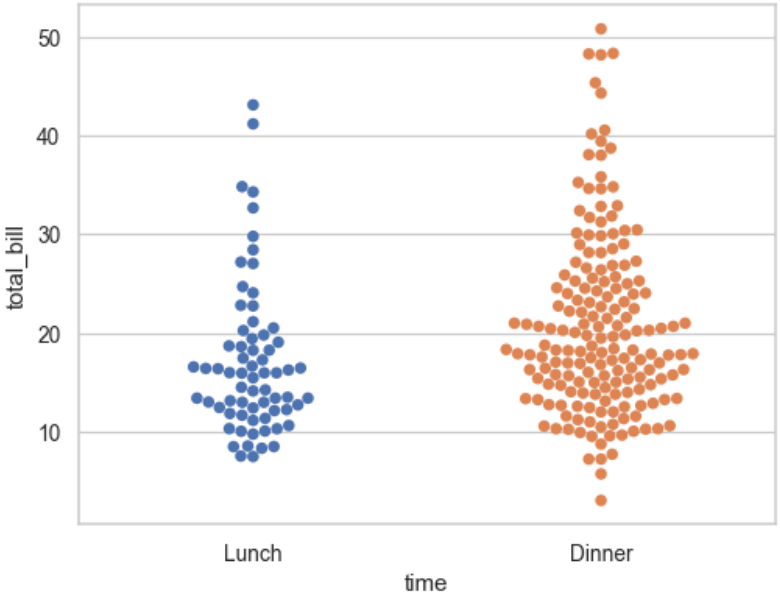
One numerical and one categorical



One numerical and one categorical



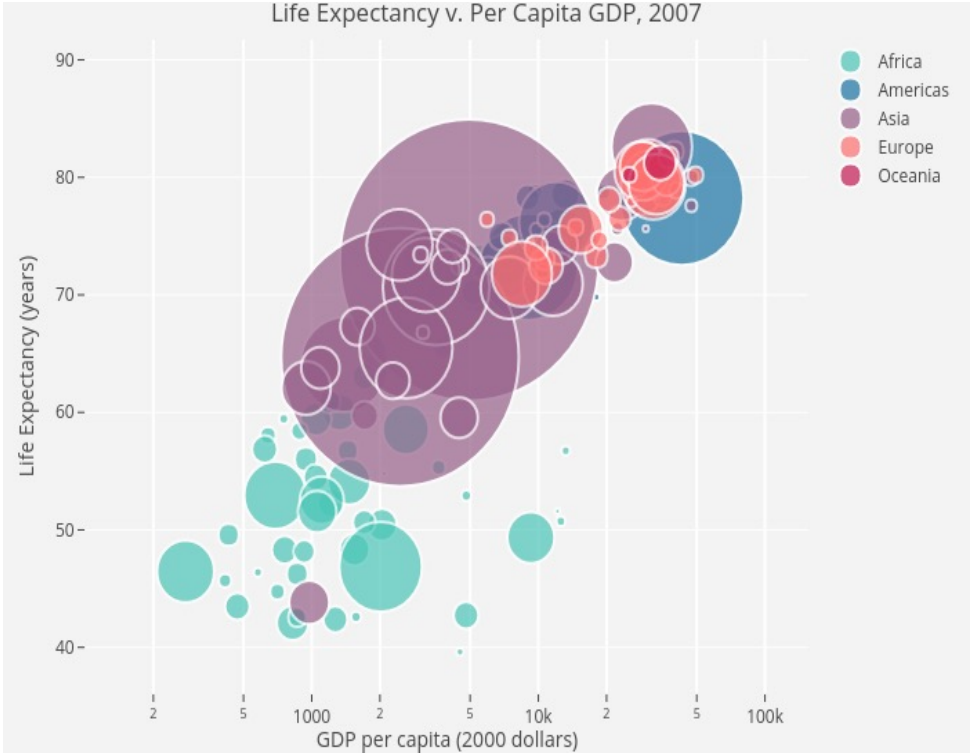
Violin plot



Swarm plot

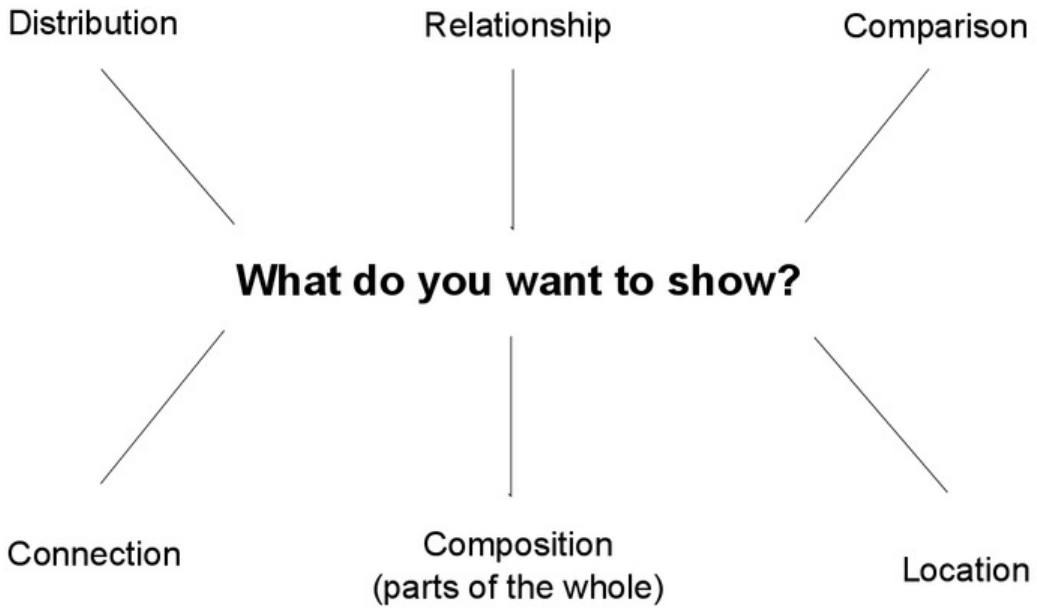
More than two variables

- Third and fourth dimensions can be added.



(source: https://plot.ly/)

Which chart type should I use?

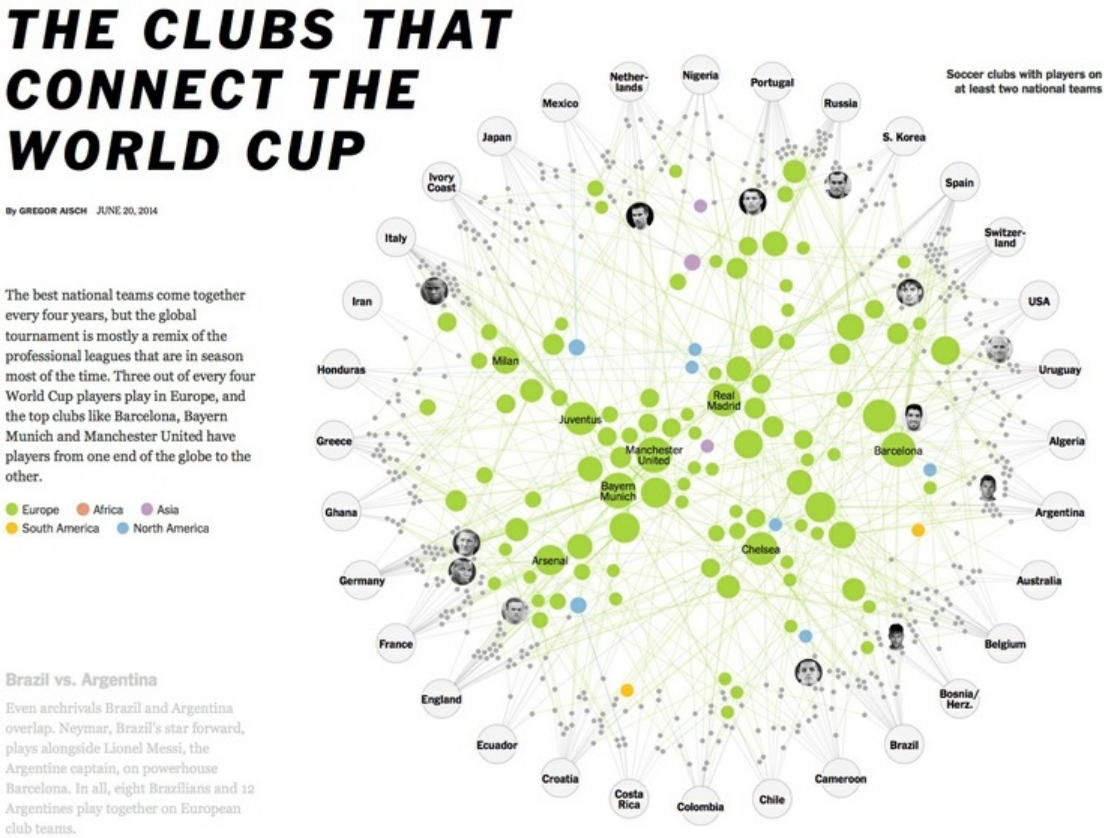


University of BRISTOL

More graphs

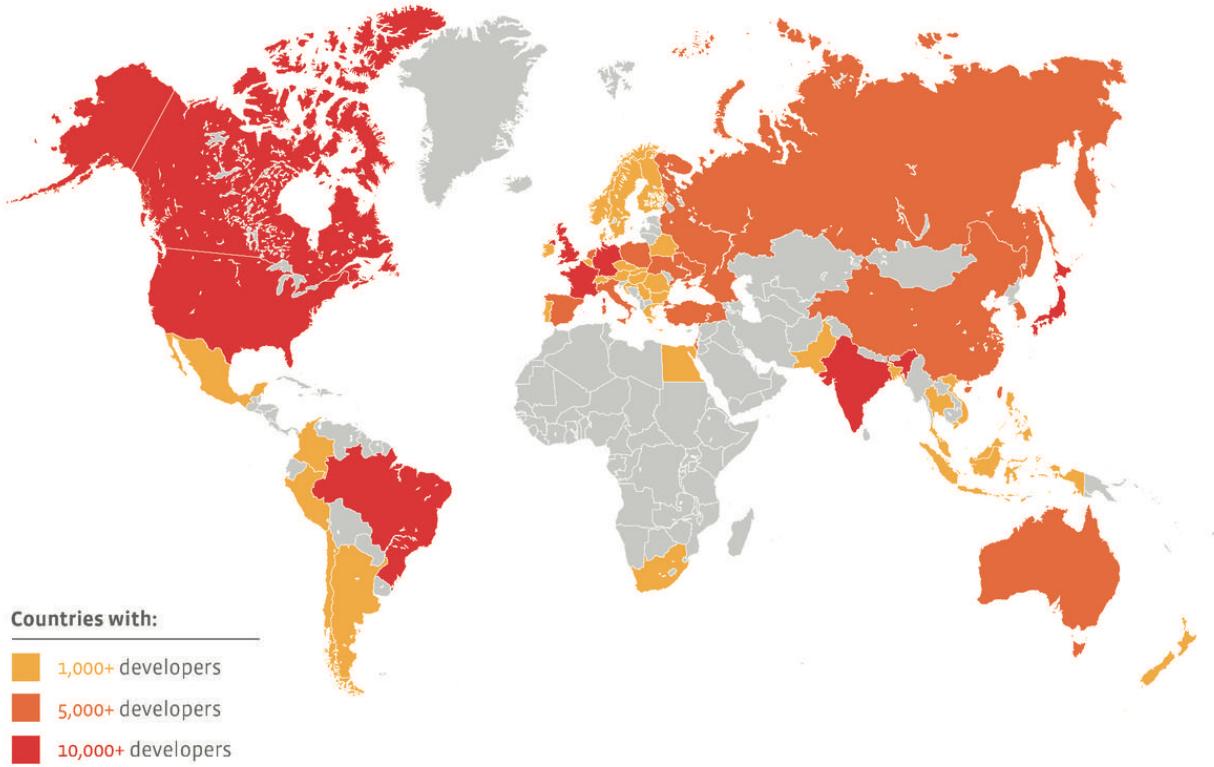
bristol.ac.uk

Connection



(Source: *The New York Times*)

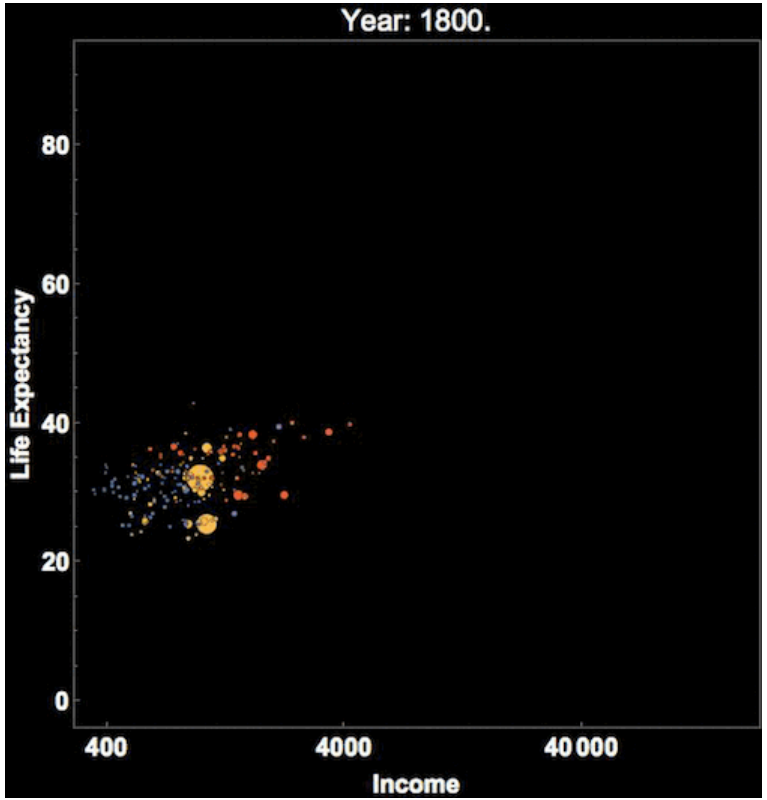
Location



(Source: https://www.pinterest.co.uk/)

We can also go higher

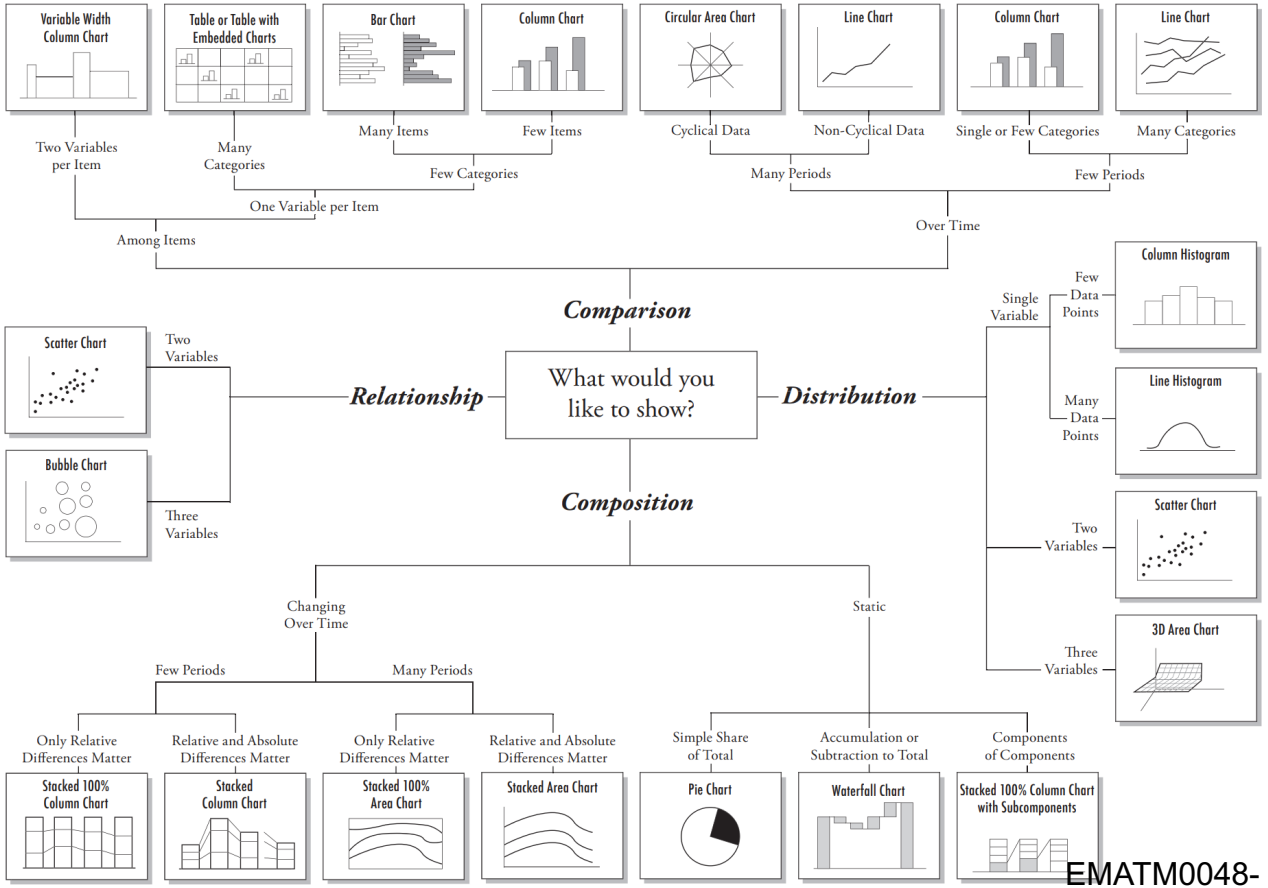
The pressing question is, can we go higher than six dimensions?



Hans Rosling's famous visualization of global population, health and economic indicators

Which chart type should I use?

(Source: A. Abela, Extreme Presentation Method)



University of BRISTOL

Outliers

bristol.ac.uk

Outliers

"An outlier is an entry in a dataset that is anomalous with respect to the behavior seen in the majority of the other entries in the dataset."

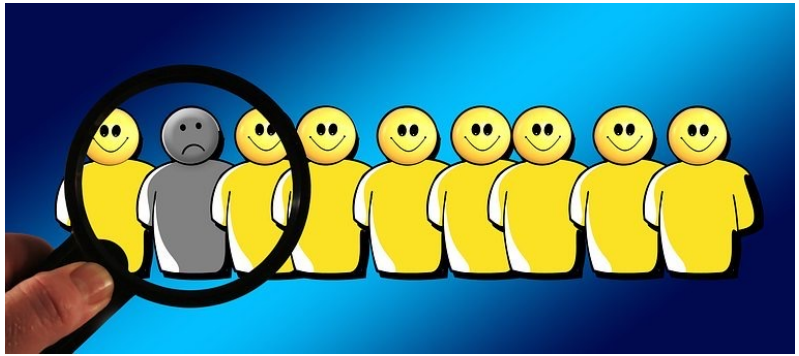
- Outliers can be seen from two different perspectives:
  - They might be seen as glitches in the data.
  - They might be also seen as an interesting element as they could potentially represent the consequential elements in the data.

Examples of outliers

- Measurement
- Execution errors
- Inherent data variability
- Fraud detection
- Customised marketing
- Medical Analysis
- Terrorist attacks

Why to handle outlier?

- Outliers increase the error variance and reduces the power of statistical tests
- If the outliers are non-randomly distributed, they can decrease normality,
- They can bias or influence estimates that may be of substantive interest.
- They can also impact the basic assumption of Regression, ANOVA and other statistical model assumptions.



Types of outliers

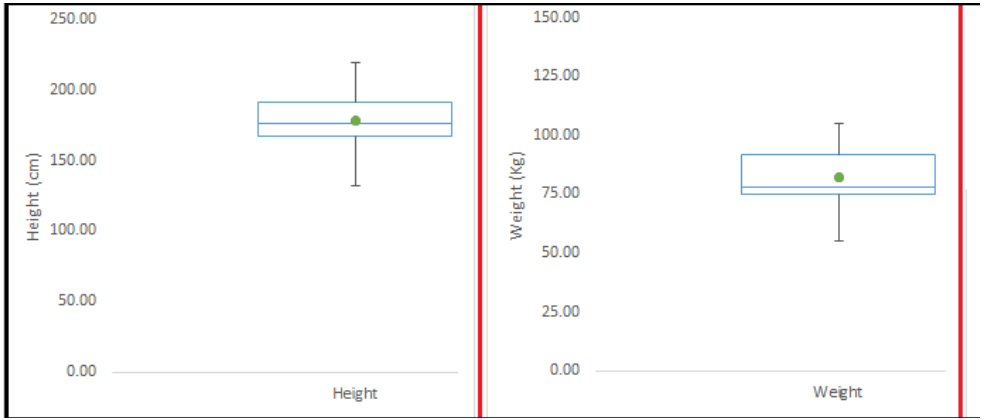
Outlier can be of two types: Univariate and Multivariate.

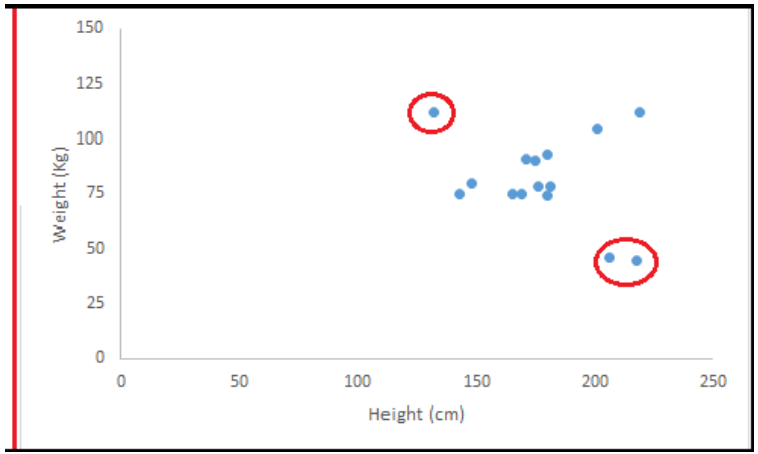
- **Univariate outlier**: concerns the distribution of a single variable.

- **Multi-variate outliers:** are outliers in an n-dimensional space. In order to find them, you have to look at distributions in multi-dimensions.

Types of outliers

Univariate outlier





Multi- variate outlier

(image source https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#three)

Handling outliers

- Some basic methods to handle outliers:
  - 3σ rule
  - The Hample identifier
  - Boxplots

- These methods are for detecting outliers. Upon detecting outliers, you can either:
  - filter them out
  - leave them or
  - impute them with other proper values.

3σ Rule

For example, consider the two datasets:

- 27 23 25 22 23 20 20 25 29 29
- 12 31 31 16 28 47 9 5 40 47

Any outliers?

What is the mean for each?

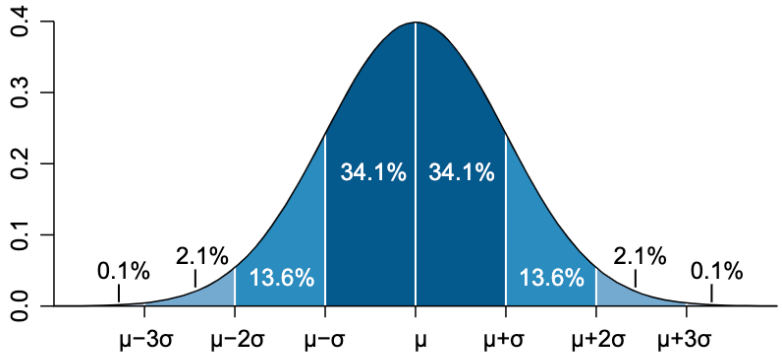
- Both have the same mean 25.
- However, the first dataset has values closer to the mean and the second dataset has values more spread out.

What is the SD for each?

- The first dataset is 3.13 and for the second set is 14.67.

3σ Rule

- About 68% of values drawn from a normal distribution are within one standard deviation σ away from the mean.
- About 95% of the values lie within two standard deviations.
- About 99.7% are within three standard deviations.



This fact is known as the 68-95-99.7 (empirical) rule, or the 3-sigma rule.

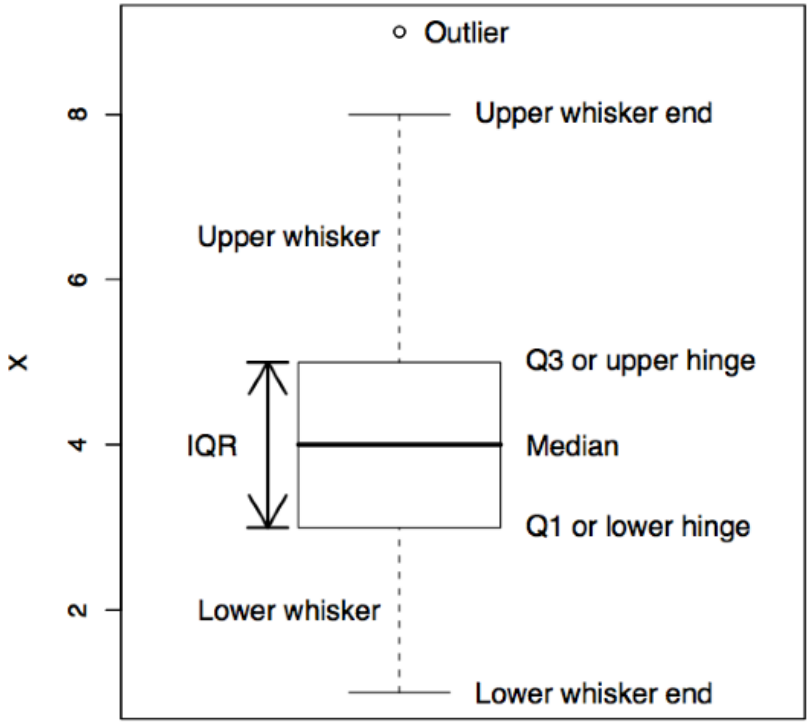
Boxplots:

A box and whisker plot—also called a box plot—displays the five-number summary of a set of data.

The five-number summary is

- Minimum
- First quartile,
- Median
- Third quartile
- Maximum.

Boxplots


Outlier
Upper whisker end
Upper whisker
Q3 or upper hinge
IQR
Median
Q1 or lower hinge
Lower whisker
Lower whisker end
x
8
6
4
2

Takeaways

- Exploratory data analysis is the first step to process your data.
- EDA can be graphical or non-graphical, could be for one variable or across several.
- The type of EDA to apply relies on the type of data and what it means.
- Outliers need to be handled in the exploration phase.