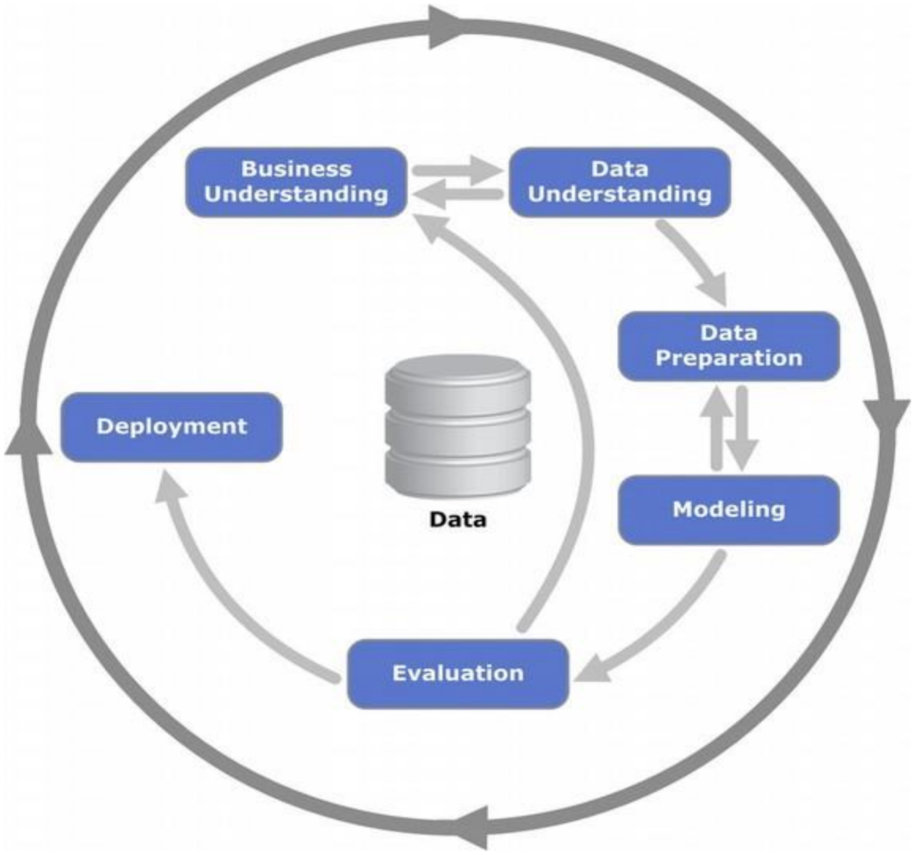
Exploratory DATA Analysis

EDA

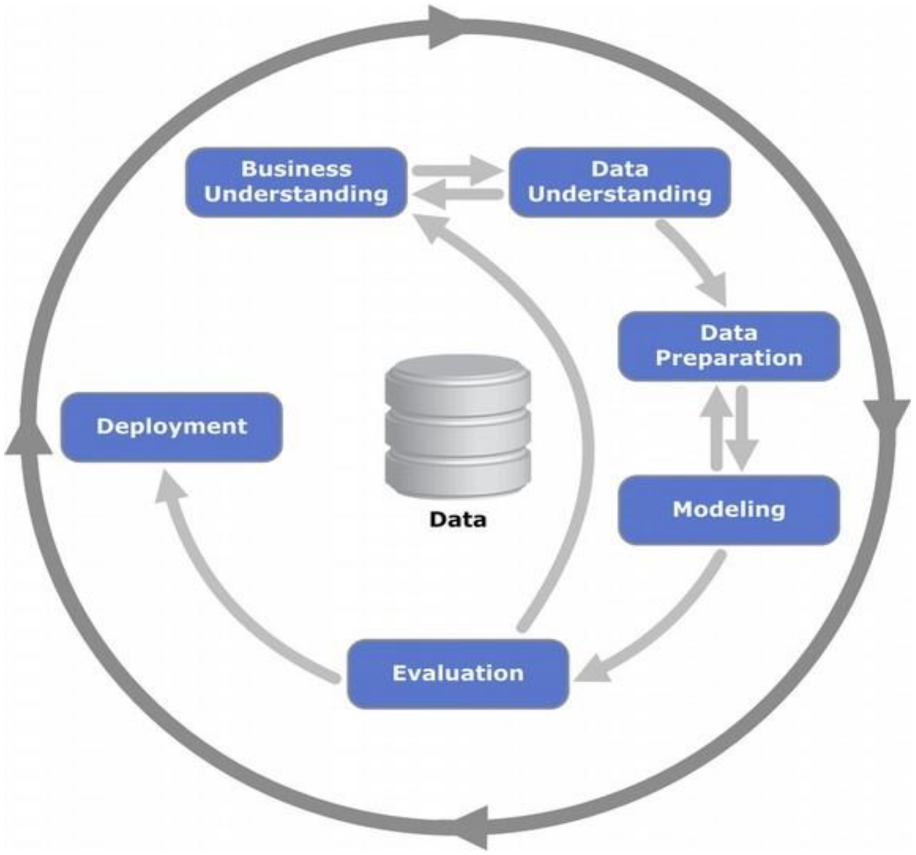
CRISP-DM process

- Cross-industry standard process for data mining, known as CRISP-DM, is an open standard process model that describes common approaches used by data mining experts. It is the most widely-used analytics model



CRISP-DM process

- **Business understanding**
- **Data understanding**
- **Data Preparing**
  - Data Cleaning: Missing value, Noise data,
  - Data Integration: Different sources, Redundancy
  - Data Transformation
  - Data Reduction
- **Data understanding**
- **Data Mining**
- **Evaluation**
- **Knowledge presentation**



So for understanding, explore or cleaning?

- When you think about your data,
  - it might seem logical to explore your data,
  - find out what’s there and which parts need attention before commencing with the cleaning.
  - After all, how can you clean up if you don’t know what the mess is?
- it’s an iterative process.
  - You might not be able to explore and find out what you need about the data with it in the wrong format.
  - Do some back and forth between exploration and cleaning.

Exploratory DATA Analysis (EDA)

- **Data exploration**: an approach similar to initial data analysis, whereby a data analyst uses visual exploration to understand what is in a dataset and the characteristics of the data. These characteristics can include:
  - size or amount of data,
  - completeness of the data,
  - correctness of the data,
  - possible relationships amongst data elements

Primarily EDA is for seeing what the **data** can tell us beyond the formal modeling or hypothesis testing task

Data Types

- Data Types are an important concept of statistics, which needs to be understood, to correctly apply statistical measurements to your data and therefore to correctly conclude certain assumptions about it.

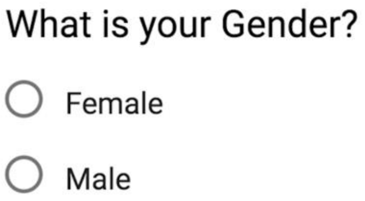
- Type 1: Categorical data. Describes categories or groups.
  - Car brands like Mercedes, BMW and Audi – different *categories*.
  - Numerical values (1 for female and 0 for male).
  - Note that those numbers don't have mathematical meaning.
- Type 2: Numerical data. Represents numbers. and it is further divided into two subsets:
  - discrete
  - continuous.

Categorical Data

- Nominal Data: represent discrete units and are used to label variables, that have no quantitative value.
  - Nominal data that has no order (labels)

- Ordinal Data: represent discrete and ordered units.
  - Difference between Elementary and High School is different than the difference between High School and College
  - Ordinal scales are usually used to measure non-numeric features like happiness, customer satisfaction and so on.

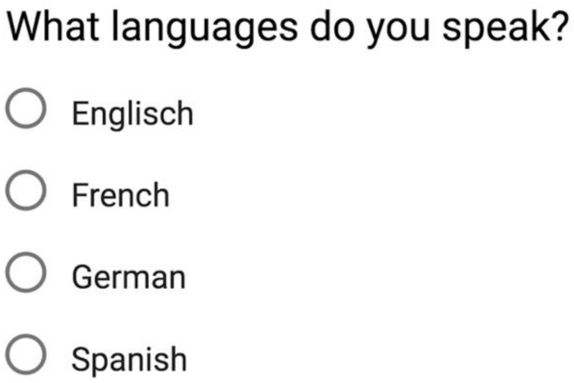
Categorical Data



What is your Gender?

○ Female

○ Male



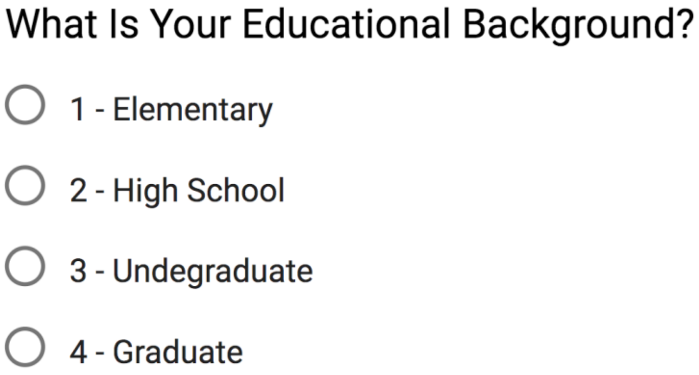
What languages do you speak?

○ Englisch

○ French

○ German

○ Spanish



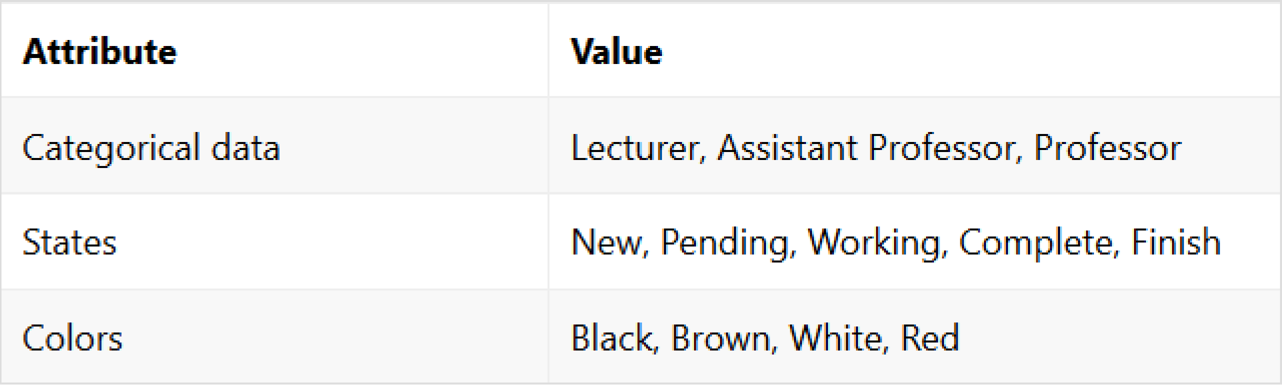
What Is Your Educational Background?

○ 1 - Elementary

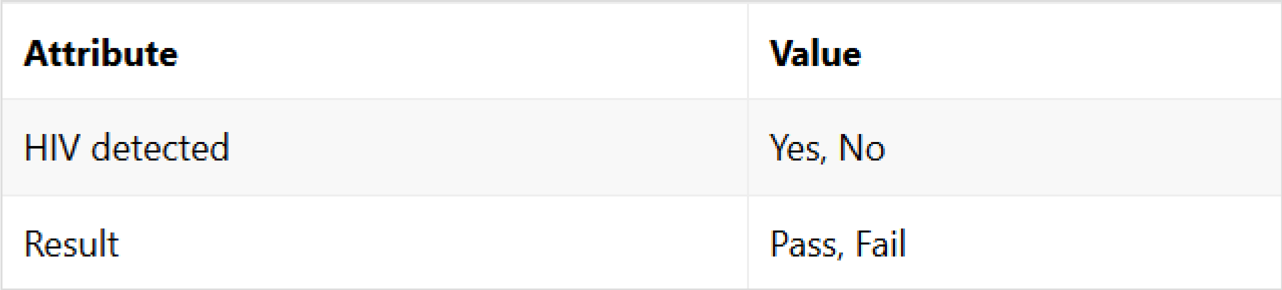
○ 2 - High School

○ 3 - Undegraduate

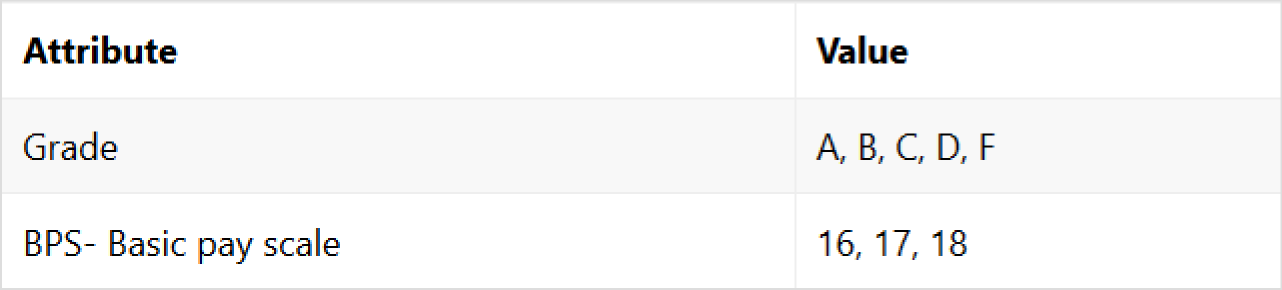
○ 4 - Graduate



Attribute	Value
Categorical data	Lecturer, Assistant Professor, Professor
States	New, Pending, Working, Complete, Finish
Colors	Black, Brown, White, Red



Attribute	Value
HIV detected	Yes, No
Result	Pass, Fail



Attribute	Value
Grade	A, B, C, D, F
BPS- Basic pay scale	16, 17, 18

Numerical Data

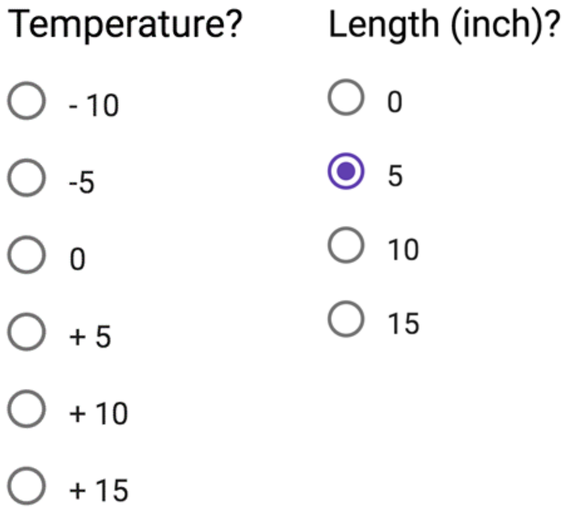
- Discrete data: values are distinct and separate.
  - Data can only take on certain values.
  - Data can't be measured but it can be counted.
  - It basically represents information that can be categorized into a classification.
  - "Can you count it and can it be divided up into smaller and smaller parts?"

- Continuous data: represents measurements.
  - their values **can't be counted but they can be measured**.
  - Continuous data technically have an infinite number of steps (in a float type there are many number between 0 and 1)

Numerical Data

- The number of heads in 100 coin flips.
- Grades at university are discrete – A, B, C, D, E, F, or 0 to 100 percent.
- The number of objects in general. No matter if bottles, glasses, tables, or cars. They can only take integer values
- Money can be considered both, but physical money like banknotes and coins are definitely discrete. You can't pay $1.243. You can only pay $1.24. That's because the difference between two sums of money can be 1 cent at most.

- Apart from weight, other measurements that are also continuous are:
  - Height
  - Area
  - Distance
  - Time



Temperature?

○ - 10

○ -5

○ 0

○ + 5

○ + 10

○ + 15

Length (inch)?

○ 0

● 5

○ 10

○ 15

Some type of continuous data: Interval, Ratio-scaled

Data type

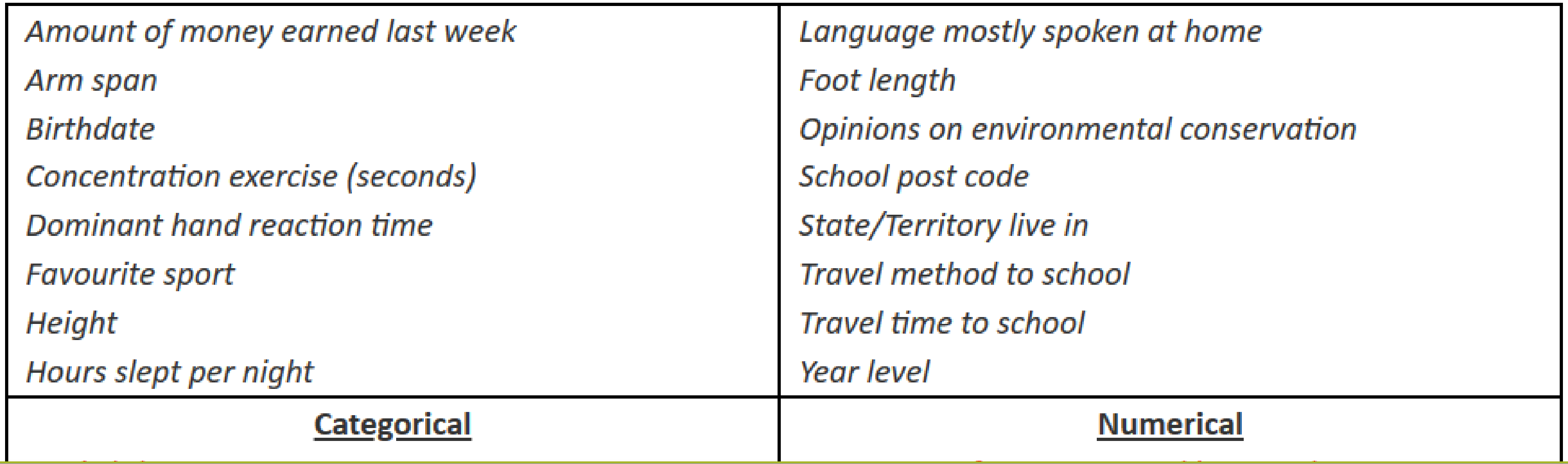
Sometimes we said that:

- Categorical data are qualitative
- Numerical data are quantitative

- **Quantitative**: The information is recorded as numbers and represents an objective measurement or a count. Temperature, weight, and a count of transactions are all quantitative data. When you have a number, doesn't necessarily make it quantitative.
  - zip codes, phone numbers and bank-accounts are numeric, but it doesn't make much sense to find the average phone number or median zip-code.

- **Qualitative**: The information represents characteristics that you do not measure with numbers. Instead, the observations fall within a countable number of groups. In fact, this type of variable can capture information that isn't easily measured and can be subjective.
  - Taste, eye color, architectural style, and marital status are all types of qualitative variables.

Data type



Amount of money earned last week *Arm span* *Birthdate* *Concentration exercise (seconds)* *Dominant hand reaction time* *Favourite sport* *Height* *Hours slept per night*	*Language mostly spoken at home* *Foot length* *Opinions on environmental conservation* *School post code* *State/Territory live in* *Travel method to school* *Travel time to school* *Year level*
Categorical	**Numerical**

[redacted]

Data type

Amount of money earned last week	*Language mostly spoken at home*
Arm span	*Foot length*
Birthdate	*Opinions on environmental conservation*
Concentration exercise (seconds)	*School post code*
Dominant hand reaction time	*State/Territory live in*
Favourite sport	*Travel method to school*
Height	*Travel time to school*
Hours slept per night	*Year level*
Categorical	**Numerical**
1. *Birthdate*	1. *Amount of money earned last week*
2. *Favourite sport*	2. *Arm span*
3. *Language mostly spoken at home*	3. *Concentration exercise (seconds)*
4. *Opinions on environmental conservation*	4. *Dominant hand reaction time*
5. *School post code*	5. *Height*
6. *State/Territory live in*	6. *Hours slept per night*
7. *Travel method to school*	7. *Foot length*
8. *Year level*	8. *Travel time to school*

How to describe and display data?

- **Categorical data: tables**

- **Numerical data: descriptive statistics**

Tables

- **Frequency Tables :** records the counts of totals and category names

- **Relative frequency tables:** displays the percentages, the counts, of the variables in each category

- **Distribution-tables or graphical displays**: show how variables are distributed across categories.

Contingency Tables

- **Contingency tables (*two-way*):** presents categorical data by counting the number of observations that fall into each group for two variables, one divided into rows and the other divided into columns.

Class

Survival

	1st	2nd	3rd	Crew	total
Alive	203	118	178	212	711
Dead	122	167	528	673	1490
Total	325	285	706	885	2201

The totals for each category, also known as ***marginal distributions***, provide the number of individuals in each row or column without accounting for the effect of the other variable

Contingency Tables

1.) Compare the survival rates for each passenger classes.

- 1st class and survived =
- 1st class and died =

2.) What percent of the survivors were in second class?

3.) What percent were second-class who survived?

4.) What percent of the second-class passengers survived?

Class

		1st	2nd	3rd	Crew	total
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201

Contingency Tables

- **Marginal distribution:** frequency distribution of one of the variables.
- **Conditional distribution:** shows one variable for just the individuals who satisfy some condition on another variable.

Class

		1st	2nd	3rd	Crew	total
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201

Contingency Tables

Did the chance of surviving the Titanic sinking depend on ticket class? Look at the conditional distributions to answer this.

Class

		1st	2nd	3rd	Crew	total
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201

Alive

	1st	2nd	3rd	Crew	Total

Dead

	1st	2nd	3rd	Crew	Total

Descriptive statistics

- **Statistics is all about drawing conclusions from data, which is a necessary initial step.**
- **Descriptive Statistics:** describing, presenting, summarizing and organizing your data (population), either through numerical calculations or graphs or tables.
- **Descriptive Statistical Analysis helps you to understand your data and is a very important part of Machine Learning.**
- **Two situations:**
  - Univariate
  - Multivariate

Descriptive statistics (One Variable)

- **Frequencies, quantiles and percentiles**
- **Measures of central tendency**
  - Mean
  - Median
  - Mode
- **Measures of spread**
  - Range
  - Interquartile range (IQR)
  - Variance
  - Mean absolute deviation
  - Median absolute deviation

Descriptive statistics (One Variable)

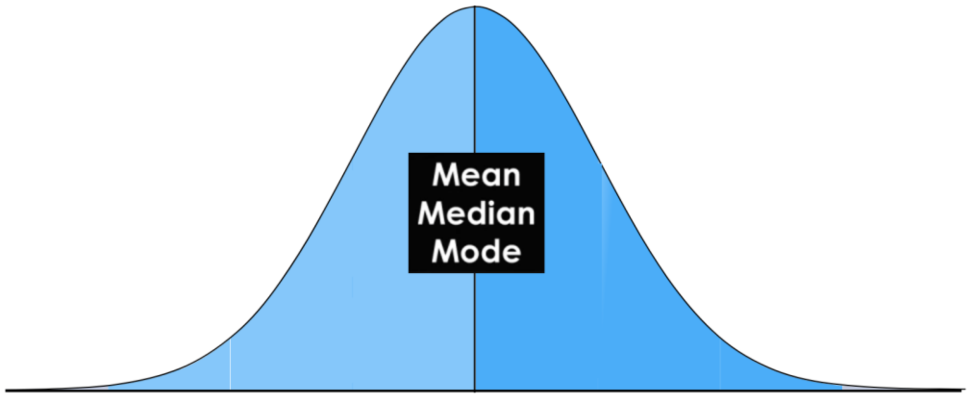
Quantiles are cut points dividing the range of a probability distribution into continuous intervals with equal probabilities: q-quantiles

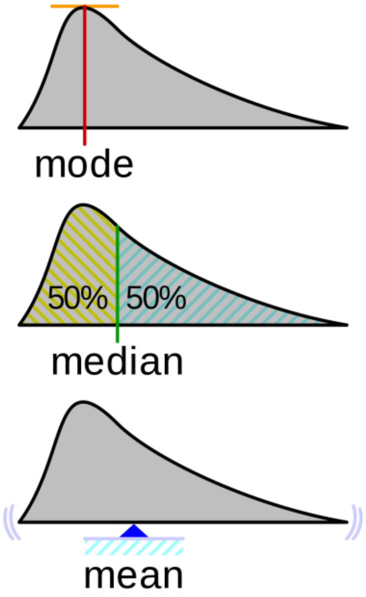
Some *q*-quantiles have special names:

- The only 2-quantile is called the median
- The 4-quantiles are called quartiles → Q
- The difference between upper and lower quartiles is also called the interquartile range, → IQR = Q3 – Q1
- The 100-quantiles are called percentiles → P

Descriptive statistics (One Variable)

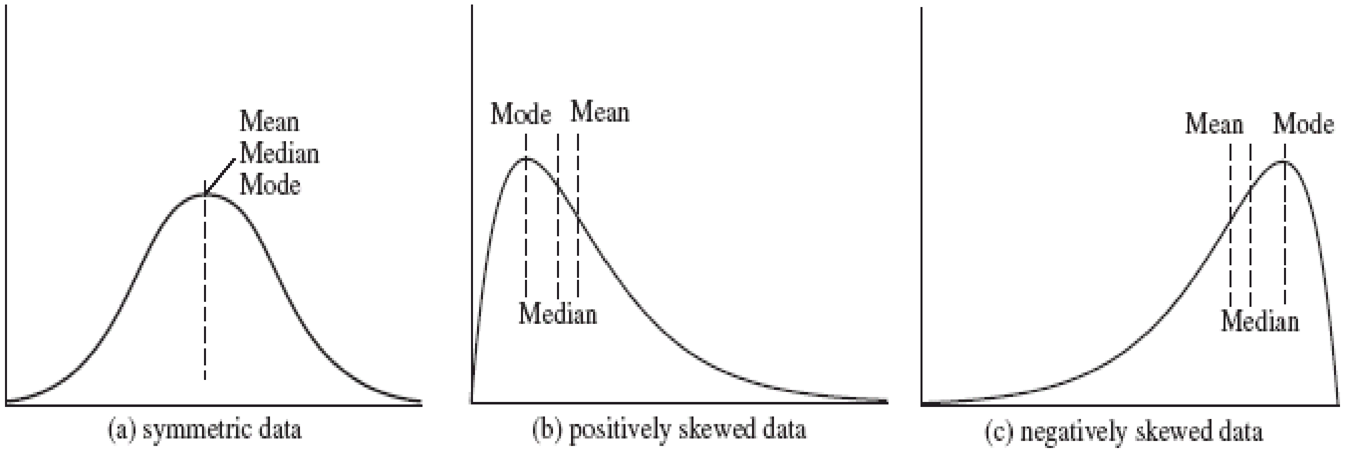
- **Central tendency: Mean, Median, Mode**





Descriptive statistics (One Variable)

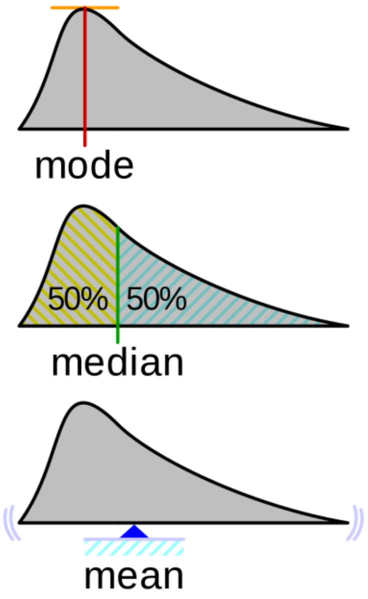
- **Central tendency: symmetric vs skewed data**



(a) symmetric data

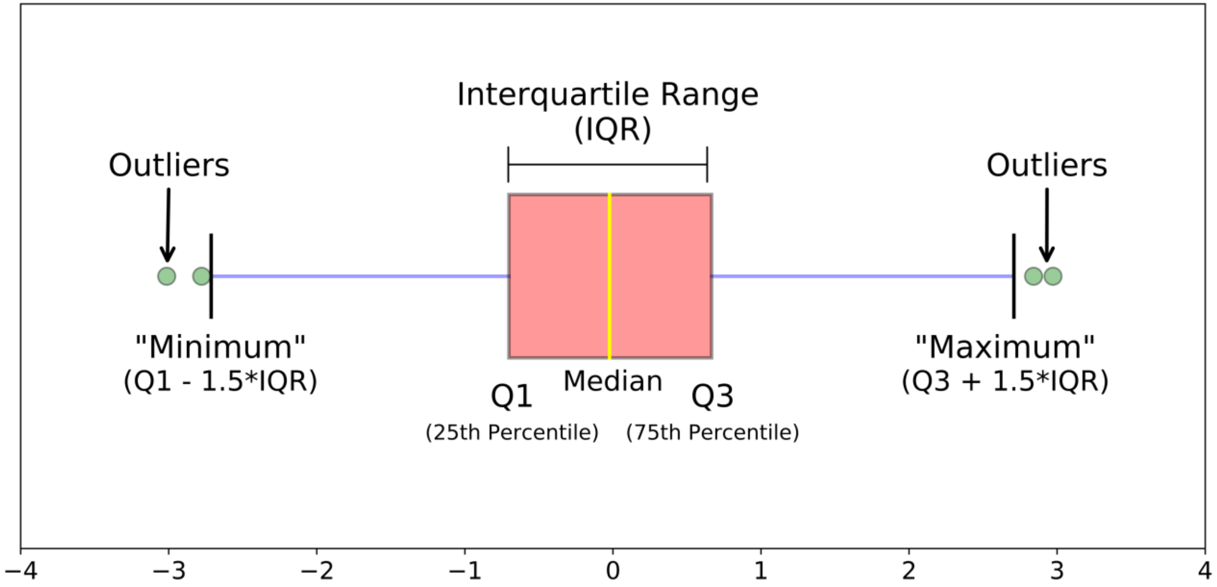
(b) positively skewed data

(c) negatively skewed data



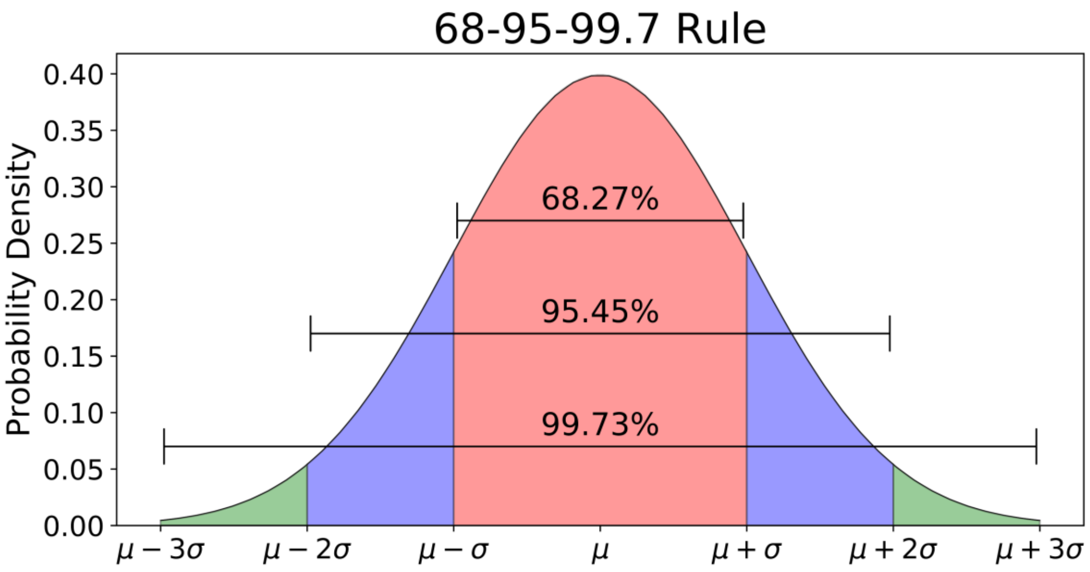
Descriptive statistics (One Variable)

- **Measures of spread:** the most popular variability measures are:
  - Range: difference between the largest and the smallest points in your data.
  - InterQuartile Range (IQR): measure of statistical dispersion between upper (75th) and lower (25th) quartiles



Descriptive statistics (One Variable)

- **Measures of spread:** the most popular variability measures are:
  - Variance : difference between every data point and the mean, squaring them, summing them up and then taking the average of those numbers.
  - Standard deviation: square root of the variance



Descriptive statistics (One Variable)

- **Measures of spread:** the most popular variability measures are:
  - Mean absolute deviation: It is calculated similarly to standard deviation, but it uses absolute values instead of squares to circumvent the issue of negative differences between the data points and their means.
  - Median absolute deviation: a robust measure of the variability of a univariate sample of quantitative data. Defined as the median of the absolute deviations from the data's median.

Descriptive statistics (Multivariate)

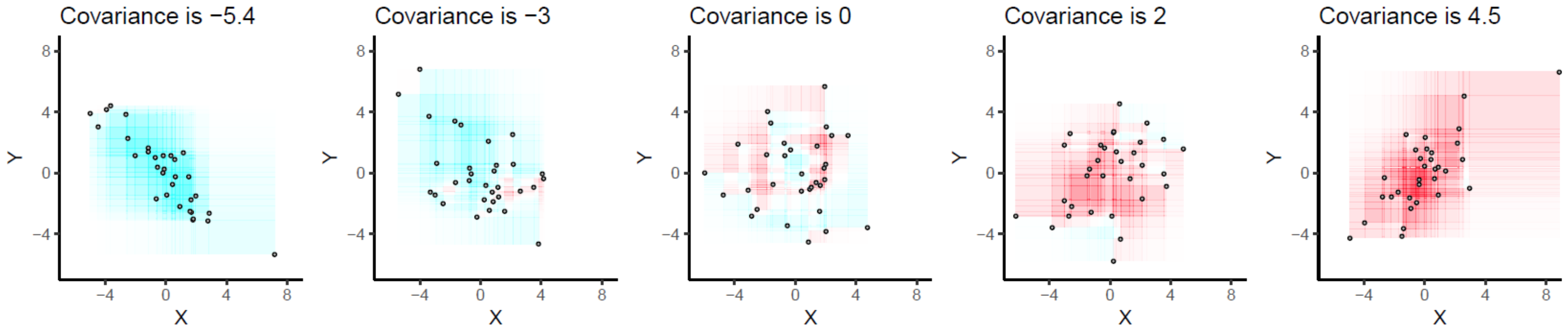
- **Mean**
- **Covariance matrix**
- **Correlation matrix**

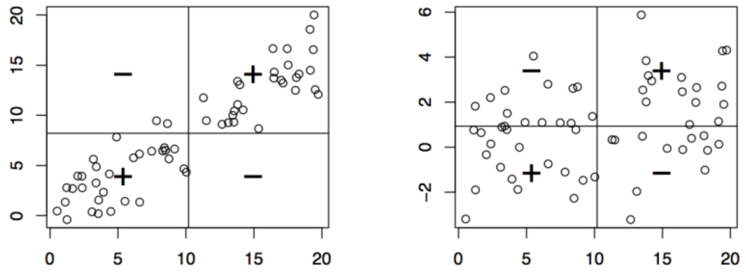
Descriptive statistics (Multivariate)

- Covariance matrix : measures **how much two random variables vary together in a population**.
- When the population contains higher dimensions or more random variables, a matrix is used to describe the **relationship between different dimensions**.
- In a more easy-to-understand way, covariance matrix is to **define the relationship in the entire dimensions as the relationships between every two random variables**

Descriptive statistics (Multivariate)

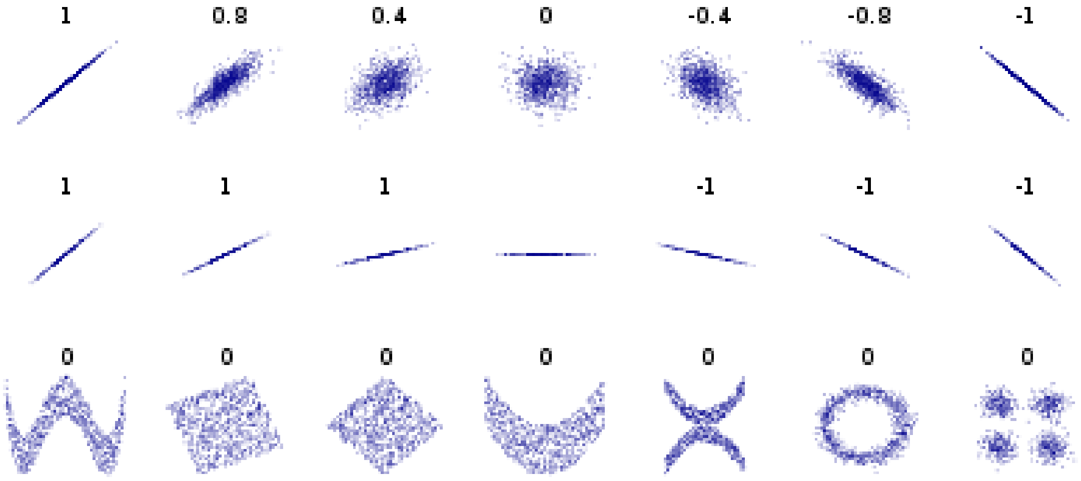
- Covariance matrix :





Descriptive statistics (Multivariate)

- Correlation matrix: the degree to which a pair of variables are linearly related.



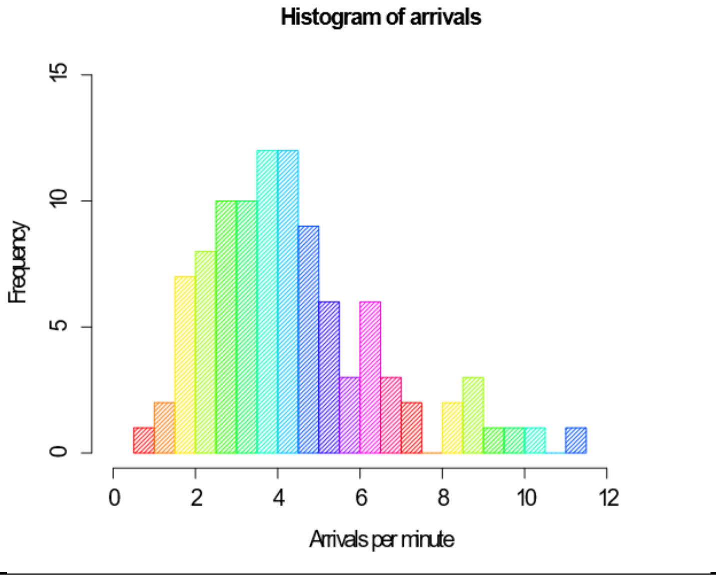
DATA EXPLORATION: Visualization

- **Arrangement**

- **Selection**

Visualization (Univariate)

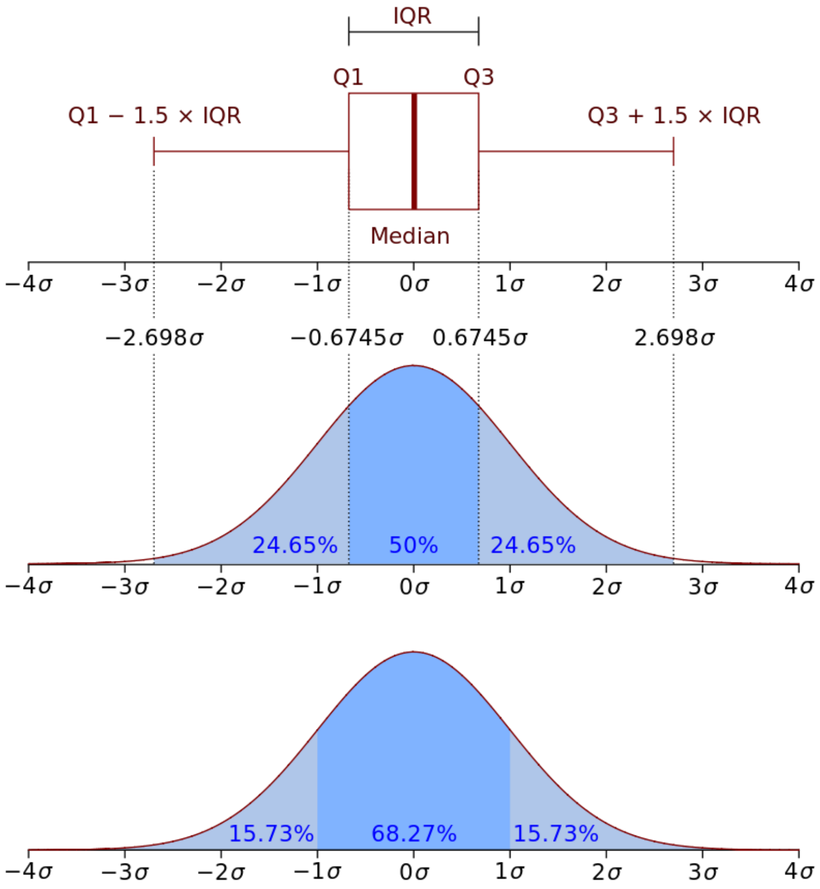
- Numerical variables:
  - histograms
  - boxplots
  - These figures would give us an understanding about the variables' central tendencies and spread.



- Numerical variables:
  - histograms : an approximate representation of the distribution of numerical or categorical data
  - the first step is to "bin" the range of values—that is, divide the entire range of values into a series of intervals—and then count how many values fall into each interval.
  - The vertical axis is the frequency density—the number of cases per unit of the variable on the horizontal axis.

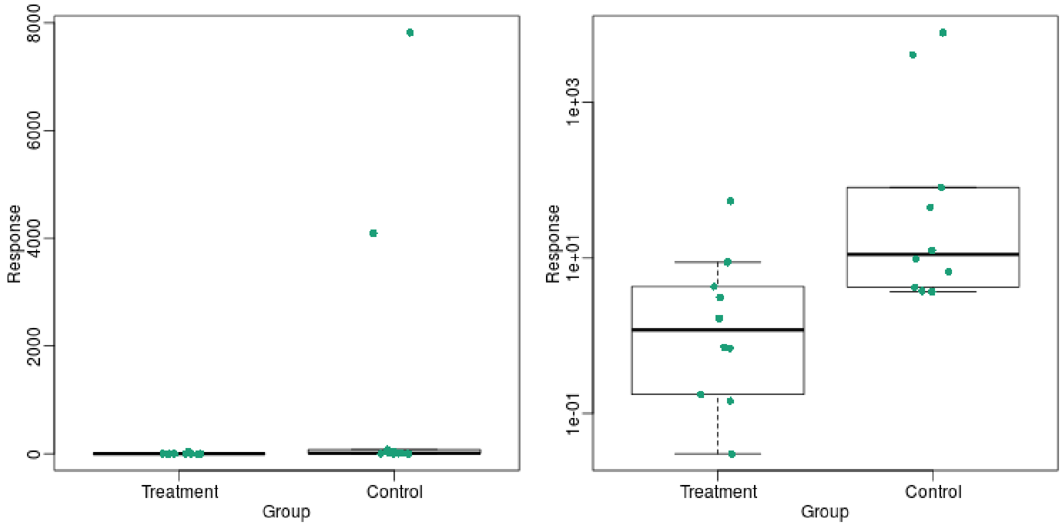
Visualization (Univariate)

- Numerical variables:
  - Box plot: a method for graphically depicting groups of numerical data through their quartiles.
  - Box plots may also have lines extending from the boxes (whiskers) indicating variability outside the upper and lower quartiles, hence the terms box-and-whisker plot and box-and-whisker diagram.
  - Outliers may be plotted as individual points



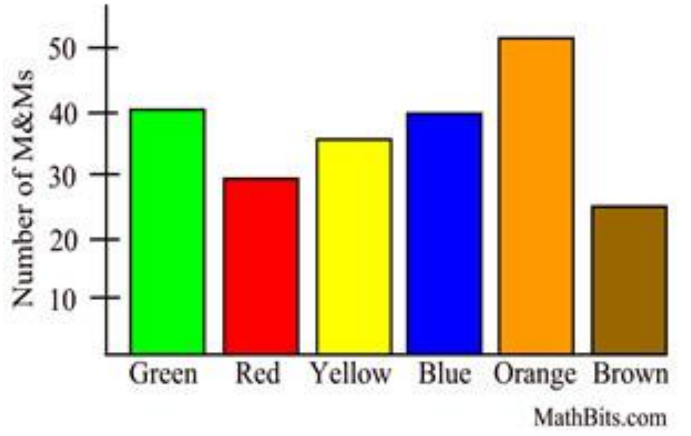
Visualization (Univariate)

- Numerical variables:
  - Box plot: a method for graphically depicting groups of numerical data through their quartiles.
  - Log scale



Visualization (Univariate)

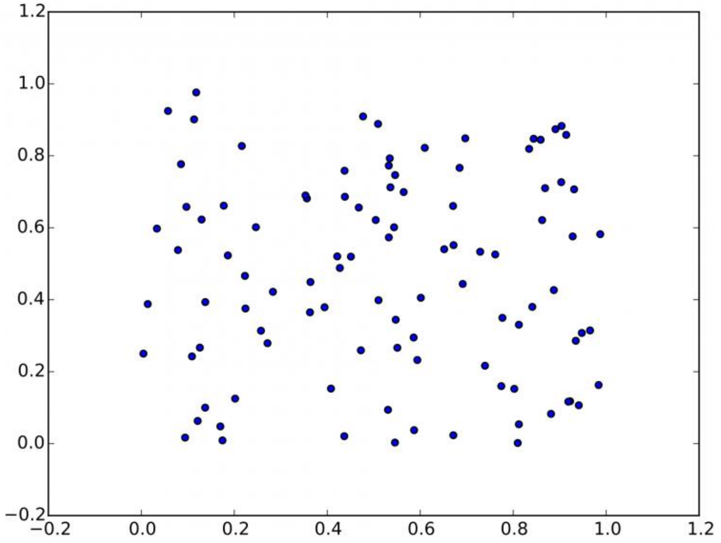
- Categorical variables:
  - build a bar chart visualization that shows the frequencies in each category.



Number of Colors in Bag of M&M Candies

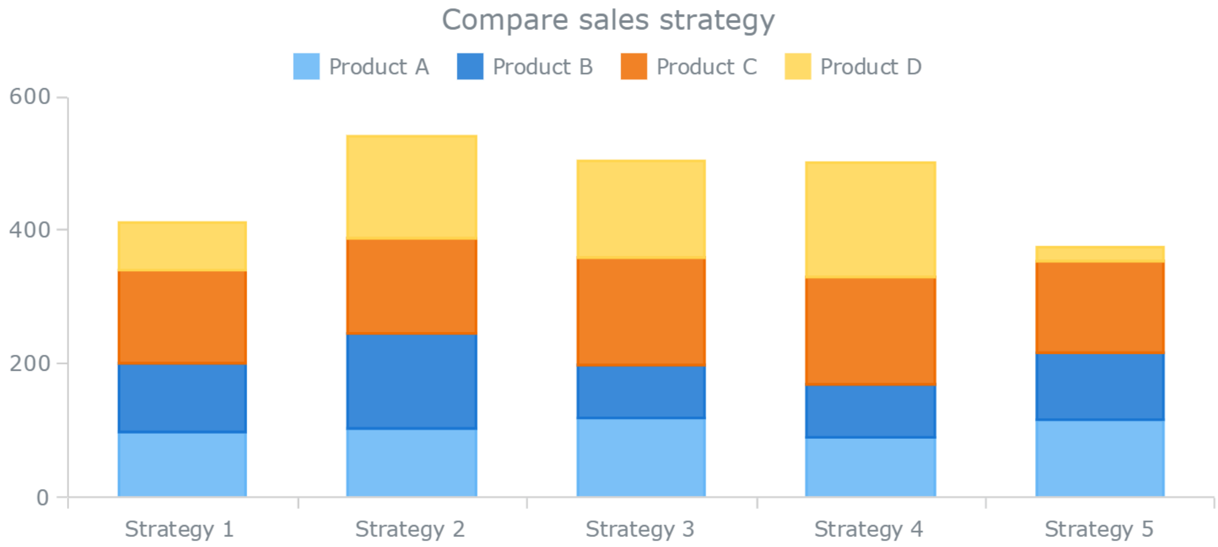
Visualization (Bi-variate)

- Continuous & Continuous (Matrix):
  - scatter plots in order to see how two continuous variables interact between each other.



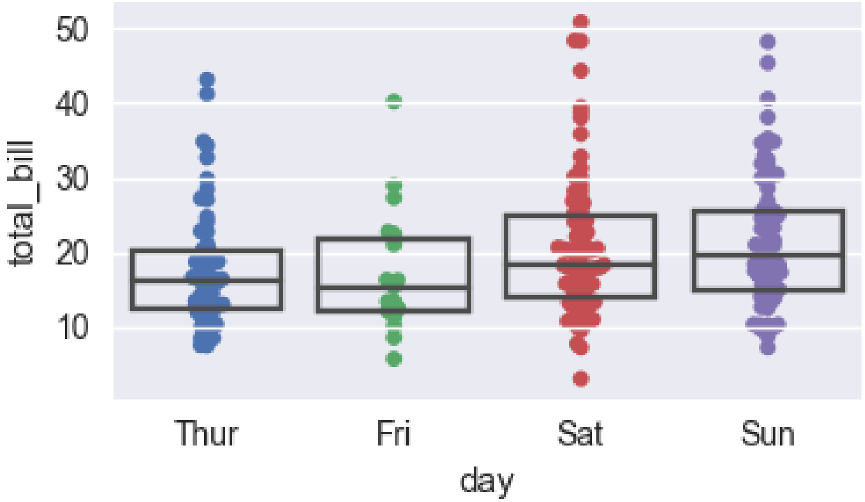
Visualization (Bi-variate)

- Categorical & Categorical:
  - A Stacked Column Chart shows how the frequencies are spread between the two categorical variables.



Visualization (Bi-variate)

- Categorical & Continuous:
    - boxplots combined with swarmplots.

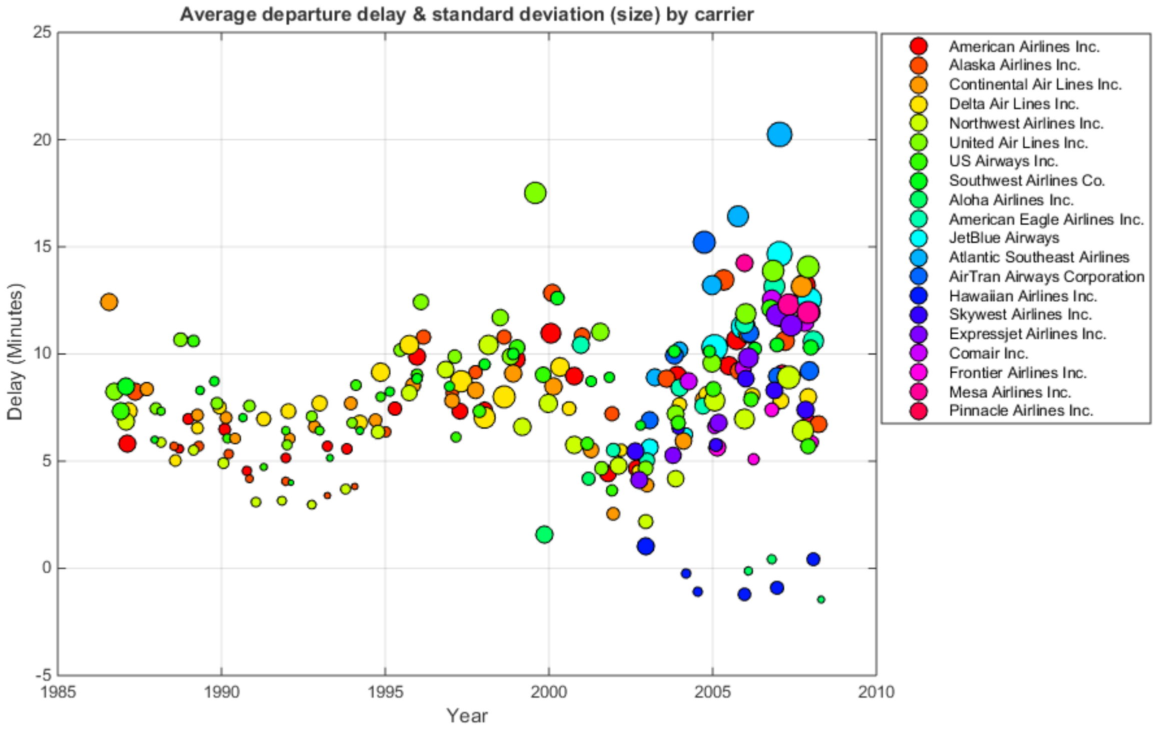


Visualization (Multivariate)

- Displaying additional dimensions
  - Size
  - Color
  - Shape
- Slicing data
- Animation

Visualization (Multivariate)

- Displaying additional dimensions



Visualization

- Higher-dimsional data
  - Matrices
  - Parallel coordinates
  - Star coordinates
  - Multidimensional arrays
  - Data cubes
- Dimensionality reduction
- Pivoting
- Slicing/dicing
- Roll up/drill down