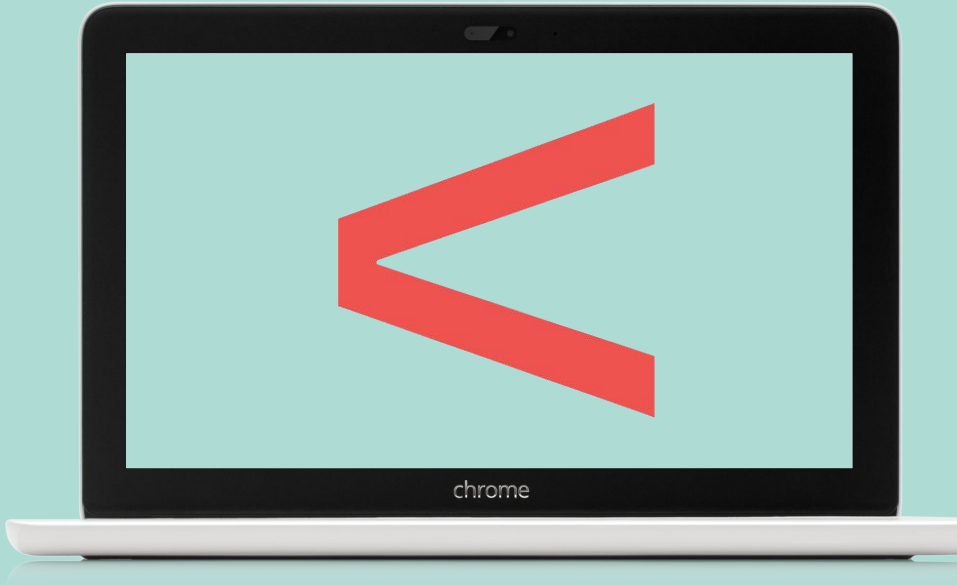
Basic statistics for Data Analysis



Importance of statistics for Data Science

- Identify the importance of features by using various statistical tests.
- Finding the relationship between features to eliminate the possibility of duplicate features.
- Converting the features into the required format.
- Normalizing and scaling the data. This step also involves the identification of the distribution of data and the nature of data.
- Taking the data for further processing by using required adjustments in the data.
- After processing the data identify the right mathematical approach/model.
- Once the results are obtained the results are verified on the different accuracy measurement scales.

Source: Analytics Vidhya

What are statistics?

Statistics involve collecting, organizing, analyzing and interpreting data to make decisions.

They help us **make sense and get information from data**, data being raw numbers that don’t say anything in themselves.

4 types of Data Analytics in Business

Descriptive	Diagnostic / Inferential
What happened? ● Provides essential insight into past performance	**Why did it happen?** ● Identifies anomalies in the data ● Uses statistical techniques to find relationships between variables and make hypotheses
Predictive	**Prescriptive**
What may happen in the future? ● Uses historical data to identify trends and determine if they are likely to recur ● Uses a variety of statistical and Machine Learning techniques, such as: classification, regression and neural networks	**How can we make it happen?** ● Allows data-driven decisions ● Recommends actions you can take to reach different outcomes

Types of descriptive statistics

- Frequency: frequency of each value.
- Central tendency: averages of the values.
- Spread (also, dispersion or variability): how spread out the values are.

You can apply these to assess only one variable at a time (univariate analysis), or to compare two or more (bivariate and multivariate analysis).

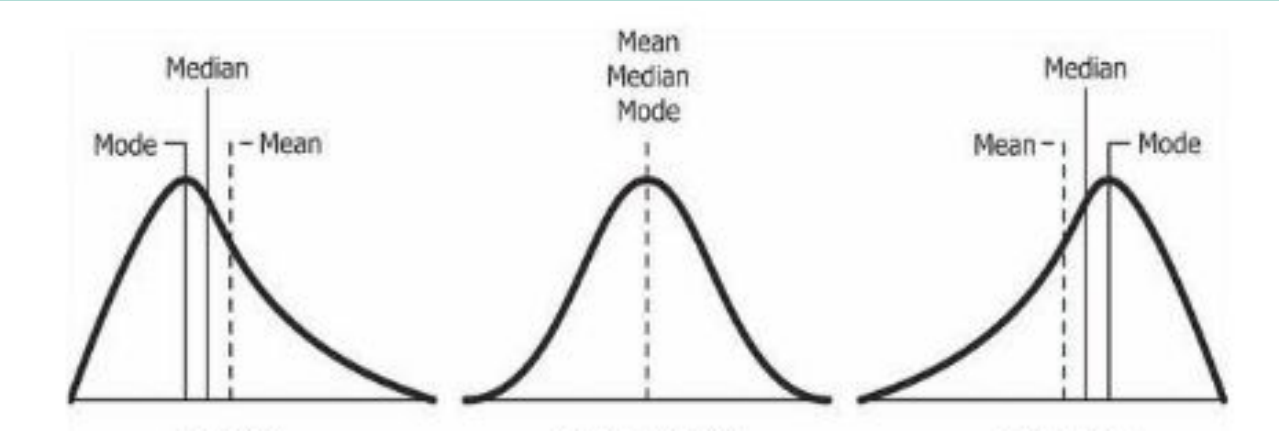
Frequency

- Values can be given in numbers or percentages

	Number	Percentage
Uses glasses	4	57.14
Doesn’t use glasses	3	42.86

- Frequencies can be simple (integers) or grouped (0-4, 5-8)

Central tendency

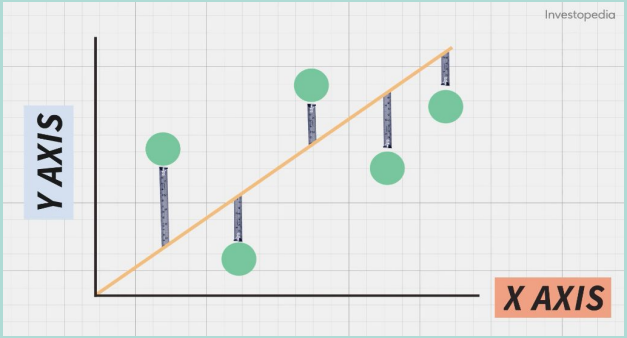


- Mean: The average of the dataset
- Median: The middle value of an ordered dataset
- Mode: The most frequent value in the dataset

Spread: range and variance

- **Range:** the difference between the highest and lowest value in the dataset
- **Variance:** it is the average of the squared differences from the mean

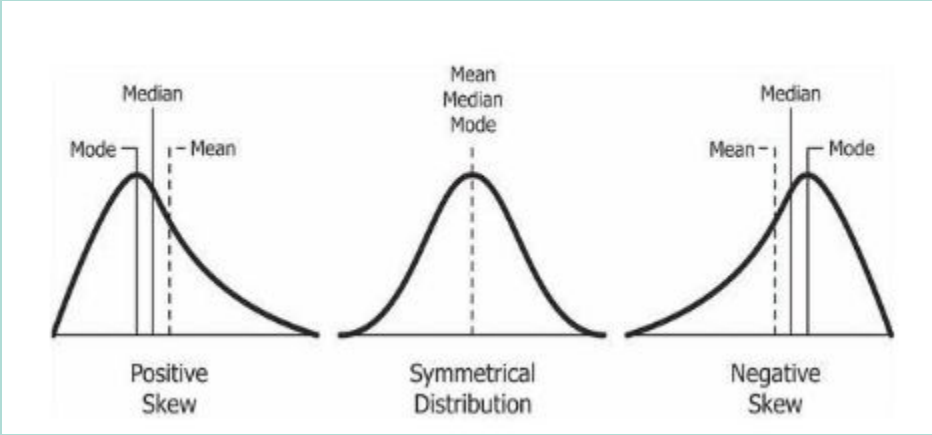
Variance



Outliers

- Outliers are data points that are far from other data points
- They can be caused by different reasons, such as data entry and measurement errors, sampling problems or natural variation
- Few examples are:
  - Wrong measure in height of group of people
  - Including a sample point that doesn't fit in the target population
  - Unexpected variation in nature / processes
- There is not a general rule whether we should remove them or not, it depends on the case
- Skewness and Kurtosis can be used to detect outliers
- The Interquartile Range (IQR) can also be used to find outliers
- Also, graphing the data

Distribution for detection of outliers: skewness

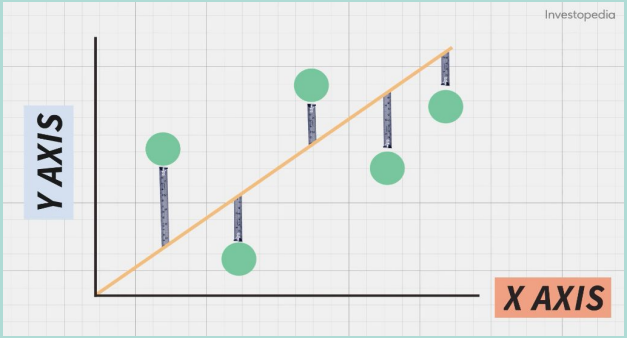


- Skewness: A measure of symmetry or balance. Can be positive or negative
  - Positive skewness is observed when there are outliers greater than the mean
  - Negative skewness is observed in distributions with outliers less than the mean

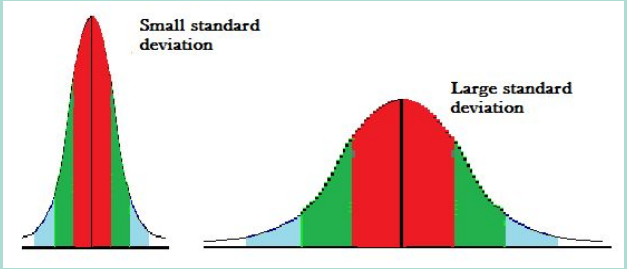
Spread: standard deviation

- **Standard deviation:** is a measure of inconsistency showing the spread of a data distribution.
  - It's the square root of the variance. The more spread out a data distribution is, the greater its standard deviation. Outliers can affect the standard deviation.
  - It is used when you need to determine the dispersion of data points (whether or not they're clustered).

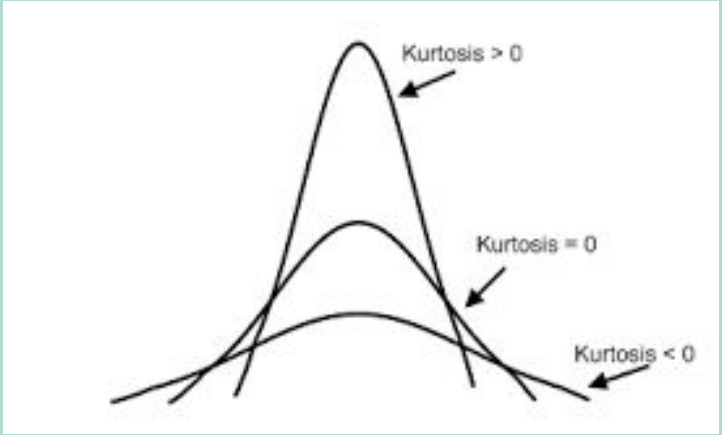
Variance



Standard deviation



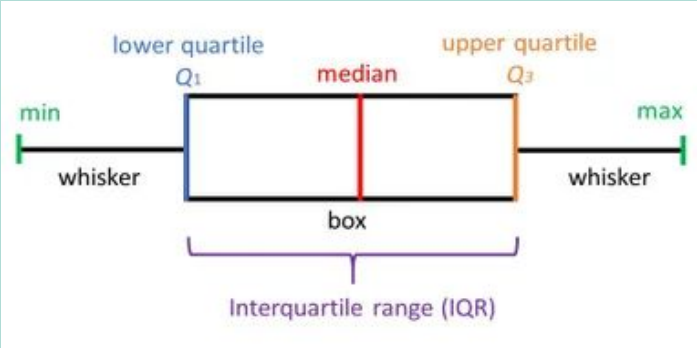
Distribution for detection of outliers: kurtosis

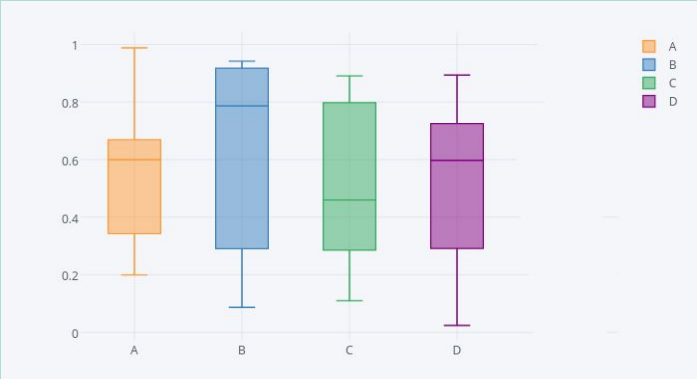


- Kurtosis: A measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution
- The larger the kurtosis, the more extreme are the outlier values
- In finance, kurtosis is used as a measure of financial risk. A large kurtosis is associated with a high risk for an investment because it indicates high probabilities of extremely large and extremely small returns.

Rank-ordered statistics

- **Percentiles:** helps to compare the given value with the rest of the data (i. e., if you have the 7th best score in a test from a group of 100 people, that means that you have a better score than 92 people, so you are at the 93rd percentile)
- **Quartiles:** divides the number of data points into four more or less equal parts, or quarters (25%, 50%, 75% and 100%)
- **Interquartile Range** (**IQR**) is basically Q3 – Q1



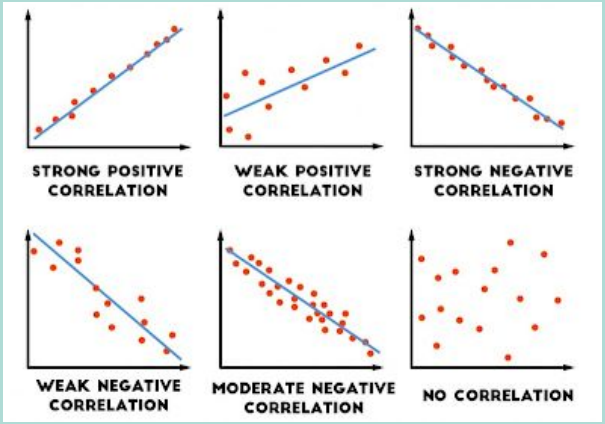


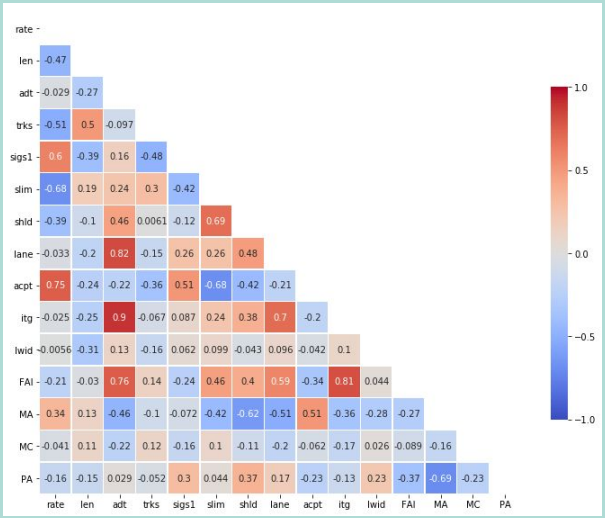
Distribution may vary

Relationship between variables

- Causation: indicates that **one event is the result of the occurrence of the other event** (cause and effect). For example, standing under the rain makes you wet
- Covariance: measures **the extent to which two random variables change in tandem** (i. e., when you bring your umbrella because it's going to rain). The value of covariance lies between -∞ and +∞
- Correlation: measures **how strongly two variables are related** (i. e. , weight and height). The value of correlation takes place between -1 and +1

"Correlation does not imply causation!"





Descriptive statistics with .describe( )

The .describe( ) method allows you to have a fast overview of basic descriptive statistics for the numeric values in your dataset.

	rating	food_rating	service_rating
count	1161.000000	1161.000000	1161.000000
mean	1.199828	1.215332	1.090439
std	0.773282	0.792294	0.790844
min	0.000000	0.000000	0.000000
25%	1.000000	1.000000	0.000000
50%	1.000000	1.000000	1.000000
75%	2.000000	2.000000	2.000000
max	2.000000	2.000000	2.000000

For more information, visit mathisfun.com