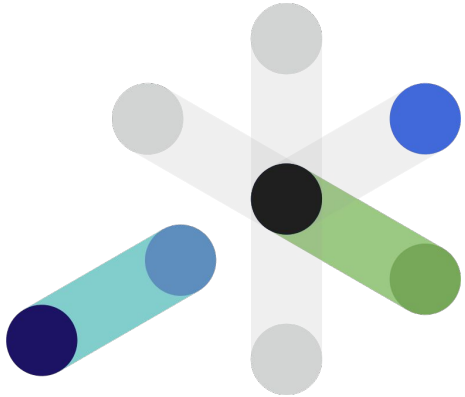
srijan:

ITM UNIVERSITY
GWALIOR • MP • INDIA

Exploratory Data Analysis (EDA)



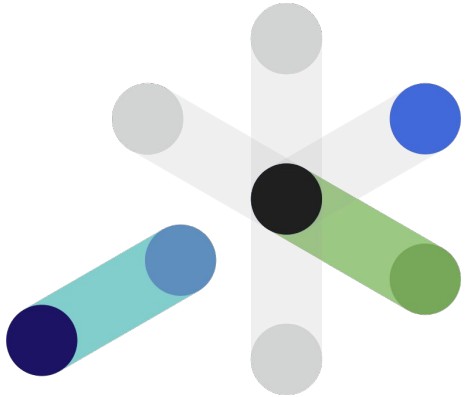
By: Mayank Kumar

About Me

Mayank Kumar

Data Scientist - II @ Srijan Technologies

Experience Across
Machine Learning, Deep Learning,
MLOps, Cloud, Algorithms, Optimization



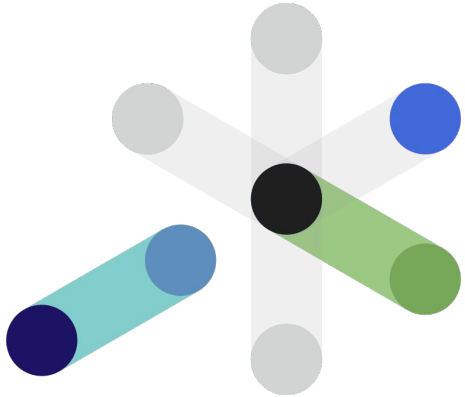
srijan:

LinkedIn : https://www.linkedin.com/in/mayank-k-jha/

Things to cover

srijan:

1. Introduction to EDA
2. Some important terminologies
3. Descriptive Statistics
4. Univariate Analysis
5. Bivariate Analysis
6. Dimensionality Reduction
7. Process of Bootstrapping a Machine Learning Project
8. Conclusion

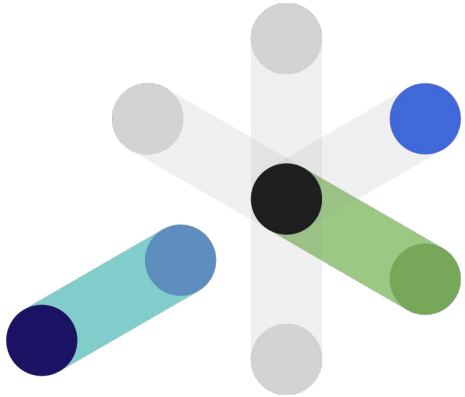


Introduction to EDA

srijan:

What is Exploratory Data Analysis (EDA) ?

a. How to ensure we are ready to use machine learning algorithms in a project?
b. How to choose the most suitable algorithms for our data set?
c. Which features has more impact on business ?

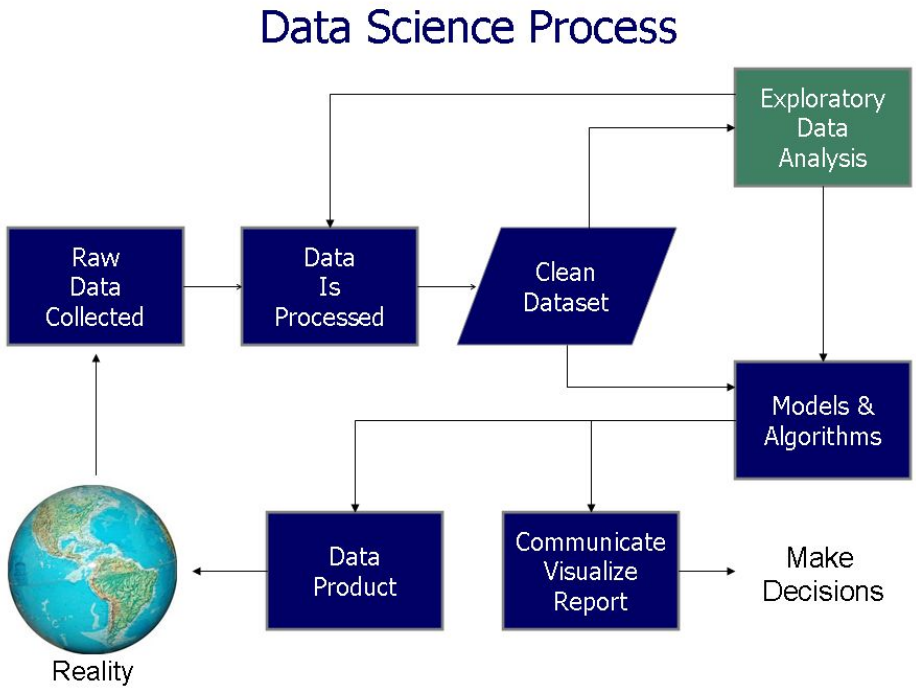


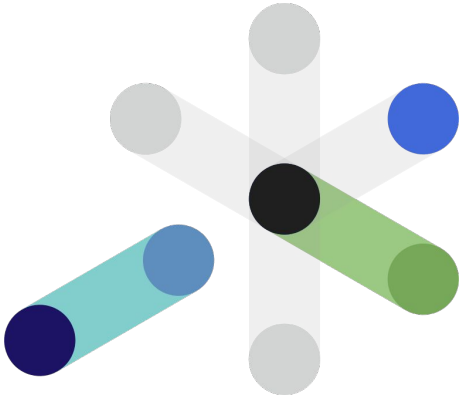
- An approach for summarizing, visualizing, and becoming familiar with the important characteristics of a data set.
- EDA is one big detour. There's no real structured way to do it. It's an iterative process.

Introduction to EDA

srijan:

Where does Exploratory Data Analysis fits in ?



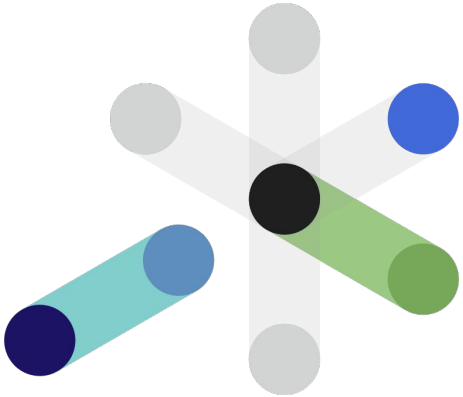


Introduction to EDA

srijan:

Exploratory Data Analysis is majorly performed using the following methods:

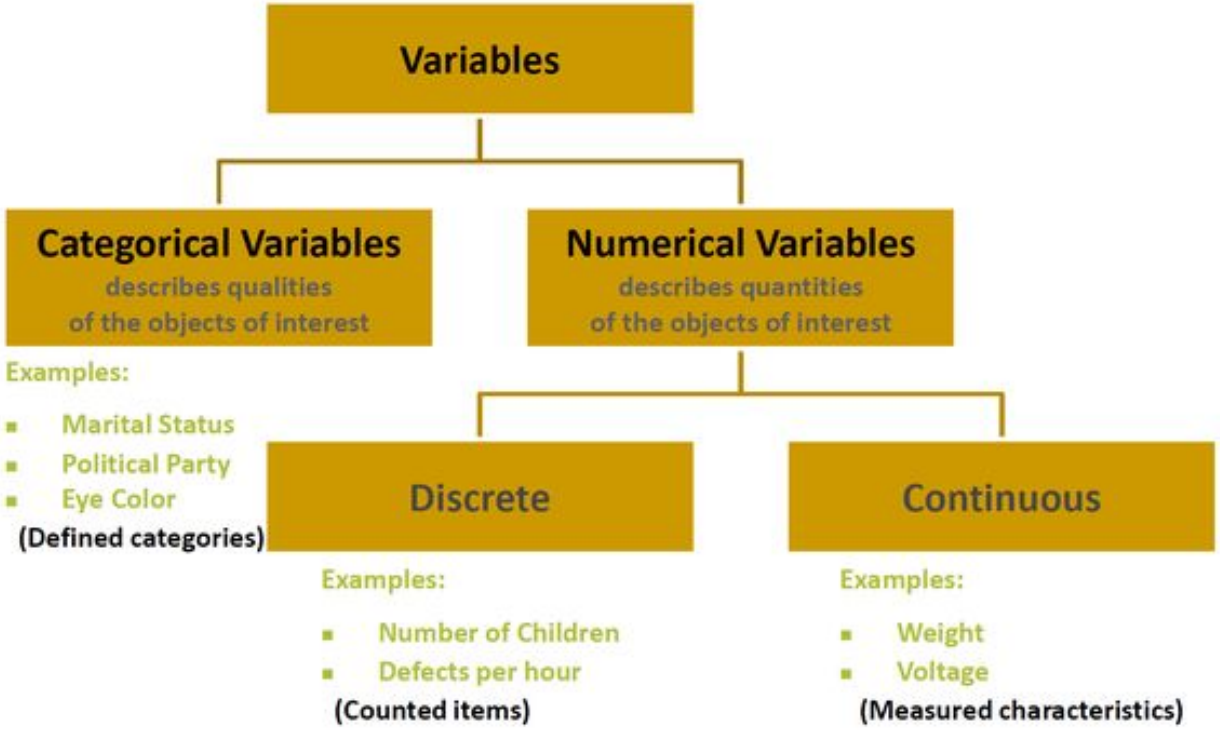
- Descriptive Statistics
- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis
- Dimensionality Reduction

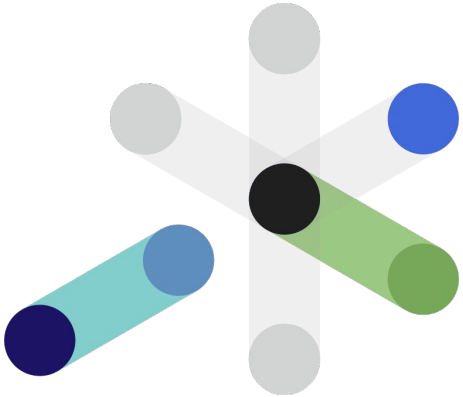


Some important terminologies

srijan:

- Variables / Columns / Features types

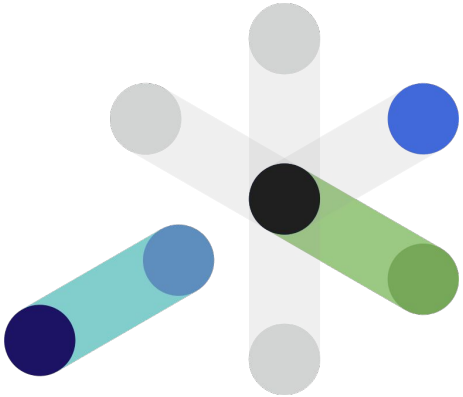




Some important terminologies

srijan:

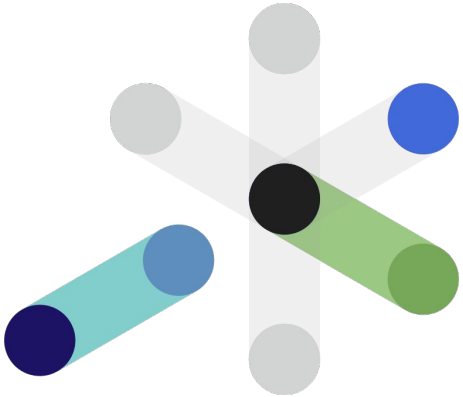
- Missing / Null values
  - In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation.
  - Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.
  - Missing values are generally represented by a null NaN values or some flags like -9999 , etc
  - How to handle?
    - Understanding the reasons why data are missing is important for handling the remaining data correctly.
    - Imputation can be done through domain knowledge, interpolation, machine learning models, descriptive statistics, etc
  - Example: Imputing missing values for gender using names title



Some important terminologies

srijan:

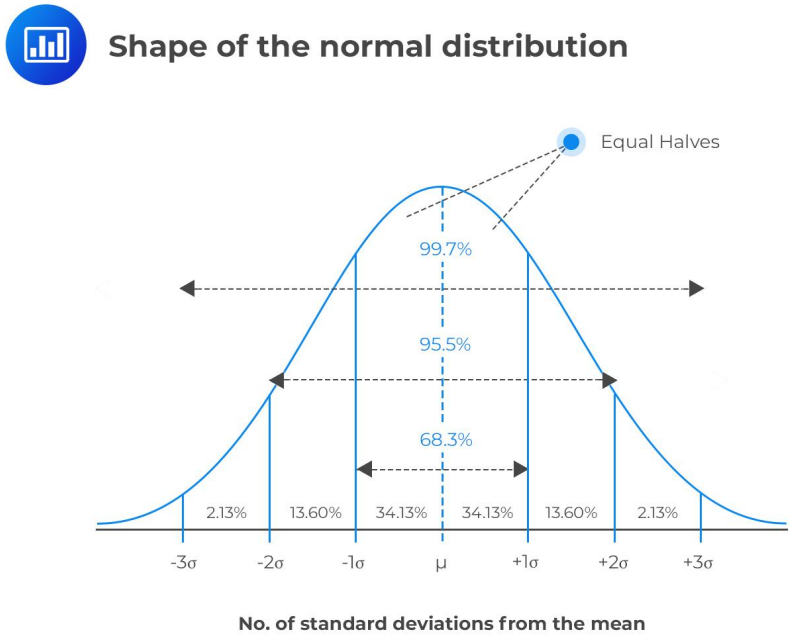
- Outliers
  - In statistics, an outlier is a data point that differs significantly from other observations.
  - An outlier can cause serious problems in statistical analyses.
  - Outliers can occur by chance in any distribution, but they often indicate either measurement error or that the population has a heavy-tailed distribution.
  - Best way to treat outliers is either to drop them or replace them using either
    - domain knowledge
    - descriptive statistics
    - clipping, etc

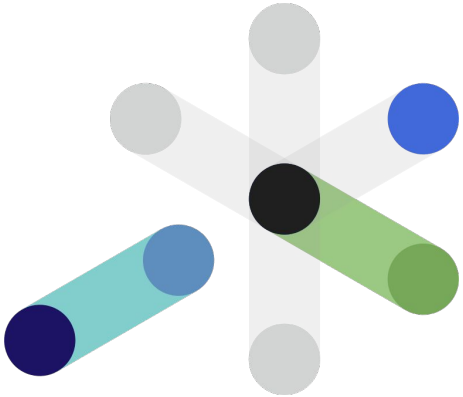


Some important terminologies

srijan:

- Normal distribution
  - Have overlapping mean, median and mode
  - Symmetric around mean
  - Satisfies 68–95–99.7 rule

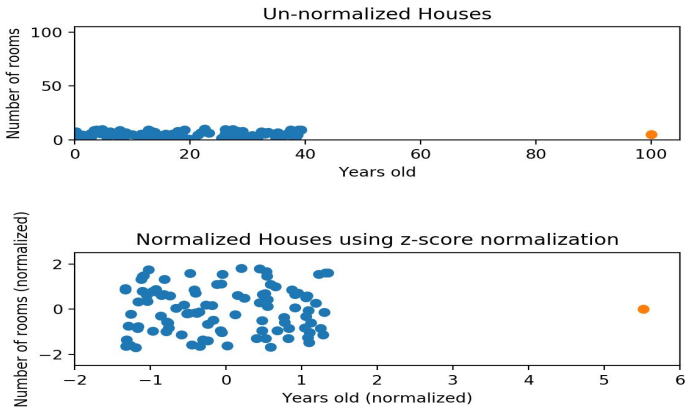


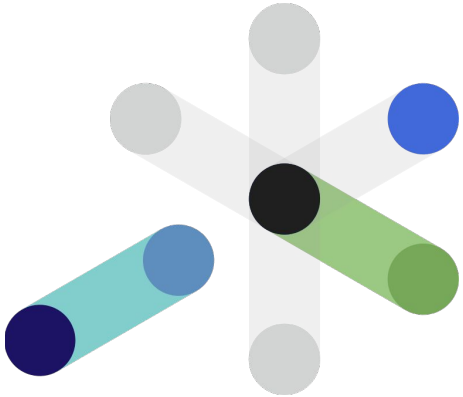


Some important terminologies

srijan:

- Normalization
  - May refer to more sophisticated adjustments where the intention is to bring the entire probability distributions of adjusted values into alignment
  - In more simpler terms, it means to align distributions to a normal distribution.
  - Methods includes:
    - Z-Score
    - Power transform
    - Log Transform
    - Min-Max Scaling

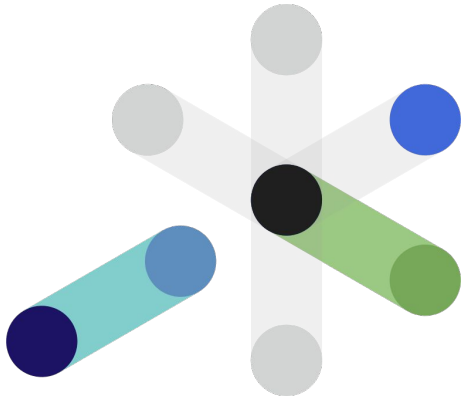




Descriptive Statistics

srijan:

- A summary statistic that quantitatively describes or summarizes features from a collection of information.
- Most commonly used measures are:
  - Mean
  - Mode
  - Median
  - Standard deviation
  - Variance
  - Skewness
  - Kurtosis



Descriptive Statistics

srijan:

3, 7, 10, 8, 31, 10, 2

Mean (avg) $= \frac{3+7+10+8+31+10+2}{7} = \frac{71}{7}$

↓

10.14

7 numbers

Median = 2, 3, 7, 8, 10, 10, 31

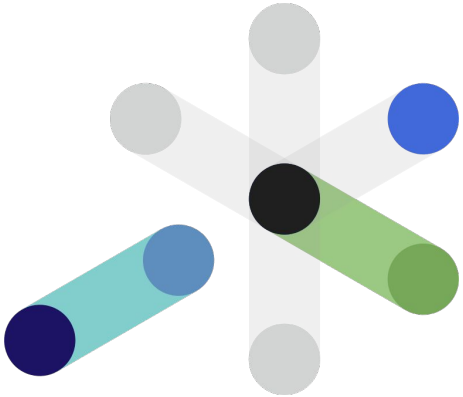
↓ ↑ middle

8

Mode 3, 7, (10) 8, 31, (10) 2

↓

10



Mean Vs Median Vs Mode

Descriptive Statistics

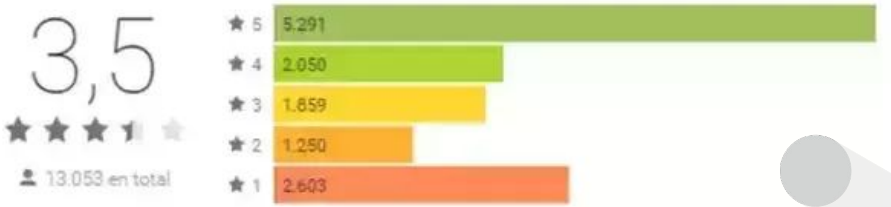
srijan:

Some Real life examples of these stats:

Mean

- Example could be
  - Ratings in Play Store

(5*5291+4*2050+3*1659+2*1250+2603)/13053
3.42718149



Median

- Example could be
  - Calculating average income for a country, median is used

Mode

- Example could be
  - Most viewed videos on YouTube
  - Most popular hotels recommendation in a city based on views/bookings

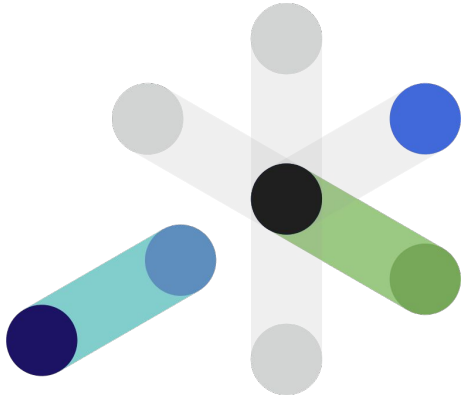
Descriptive Statistics

srijan:

Variance

- The average of the squared differences from the Mean.
- To calculate the variance follow these steps:
  - Calculate the Mean (the simple average of the numbers)
  - Then for each number: subtract the Mean and square the result (the squared difference).
  - Then calculate the average of those squared differences.

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



Standard Deviation

- The Standard Deviation is a measure of how spread out numbers are.
- Its symbol is σ (the greek letter sigma)
- The formula is easy: it is the square root of the Variance.

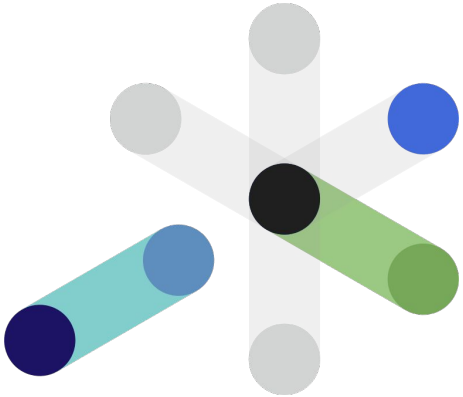
$$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}}$$

Descriptive Statistics

srijan:

Some Real life examples of Variance and Standard Deviation

- Example could be
  - Customer survey analysis
    - A market researcher is analyzing the results of a recent customer survey.
    - He wants to have some measure of the reliability of the answers received in the survey in order to predict how a larger group of people might answer the same questions.
    - A low standard deviation shows that the answers are very projectable to a larger group of people.
    - A high standard deviation shows that the answers are not projectable to a larger group of people.
  - Outlier detection and removal
  - Data normalization using z-score

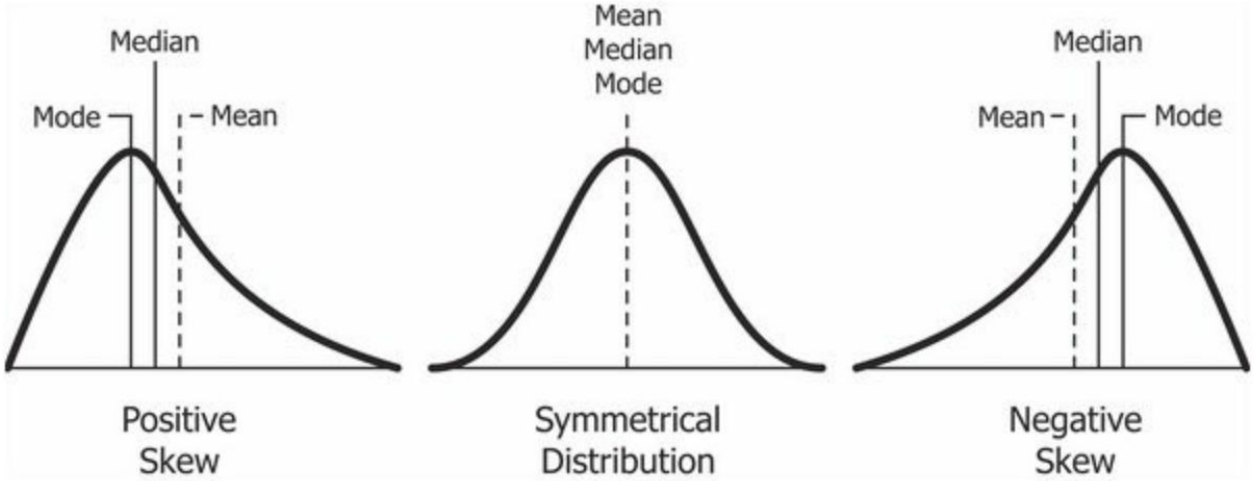


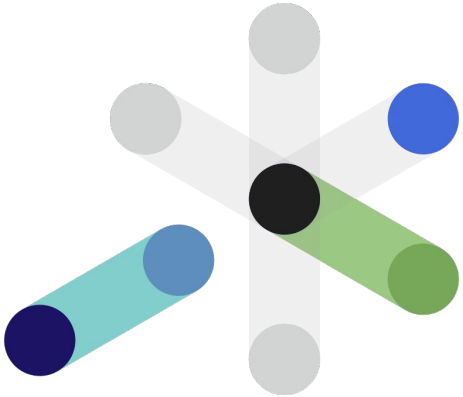
Descriptive Statistics

srijan:

Skewness

- Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean
- A perfectly symmetrical data set will have a skewness of 0.
  Example: The normal distribution has a skewness of 0
- The skewness value can be positive, zero, negative, or undefined



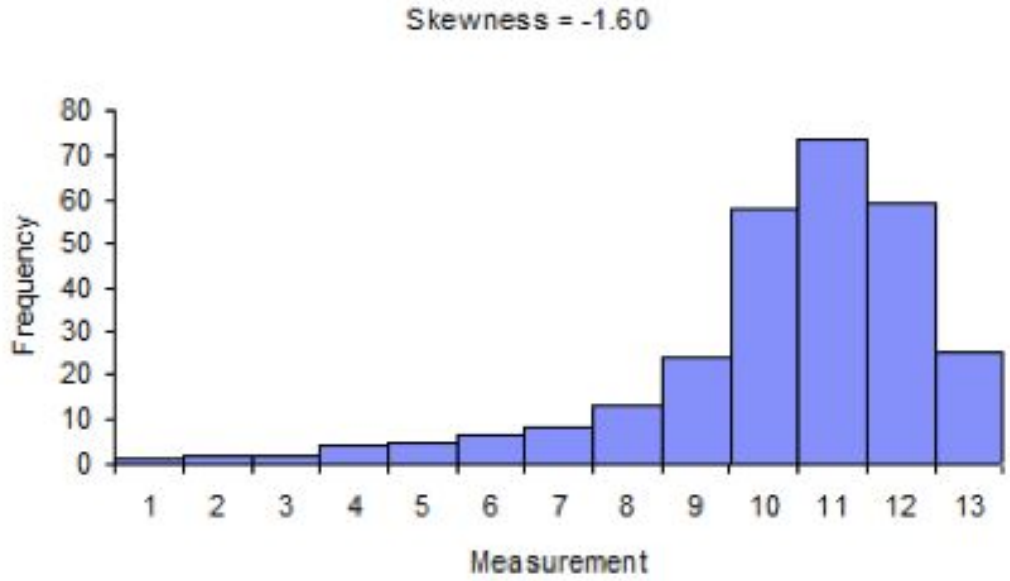


$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right)^{3/2}}$$

Descriptive Statistics

srijan:

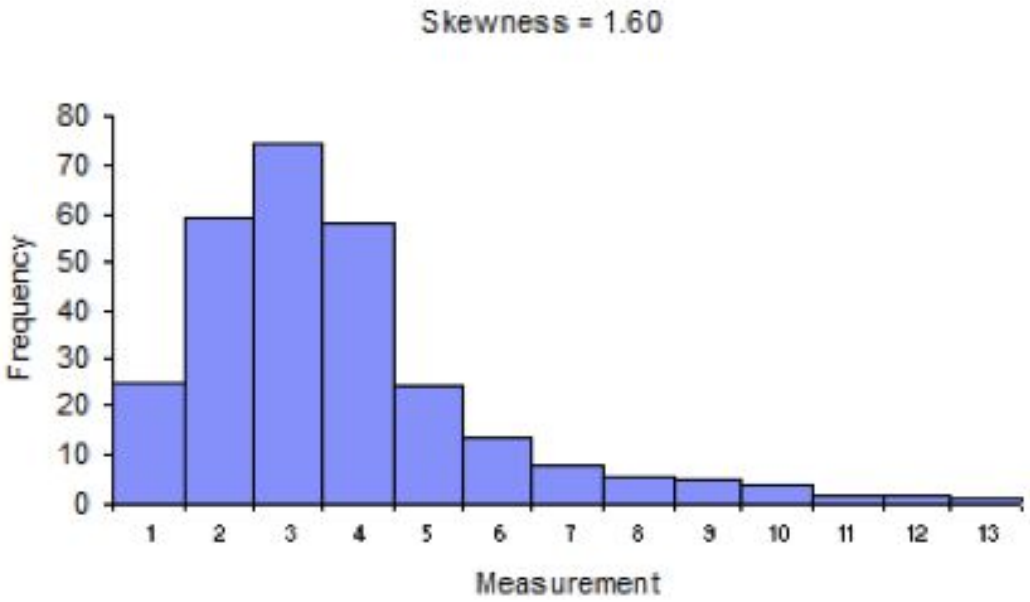
Skewness



Descriptive Statistics

srijan:

Skewness

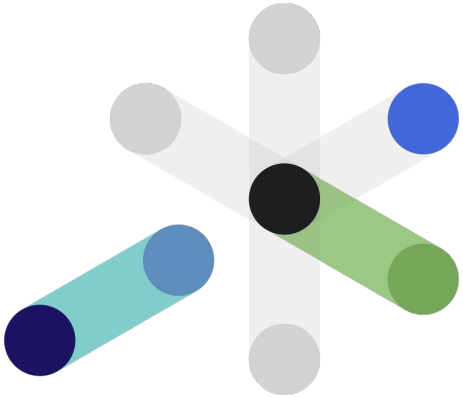


Descriptive Statistics

srijan:

Some Real life examples of Skewness

- Example for
  - Skewness
    - Common examples for positive skewness include people's incomes; mileage on used cars for sale; house prices; number of accident claims by an insurance customer; number of children in a family.

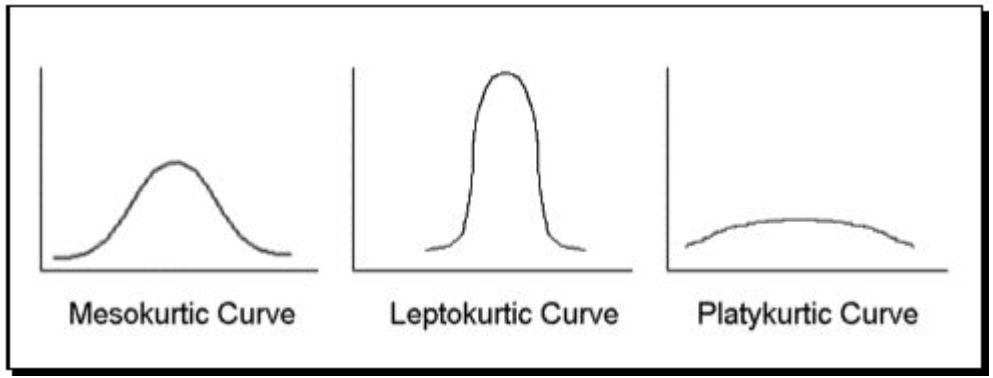


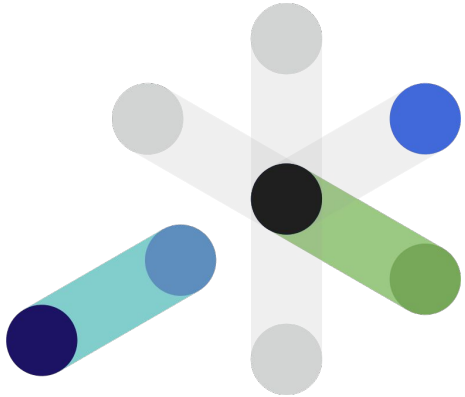
Descriptive Statistics

srijan:

Kurtosis

- Kurtosis tells us about the height and sharpness of the central peak, relative to that of a standard bell curve / normal distribution
- There are three types of kurtosis:
  - Mesokurtic : Distributions that are moderate in breadth and curves with a medium peaked height.
  - Leptokurtic : More values in the distribution tails and more values close to the mean
  - Platykurtic : Fewer values in the tails and fewer values close to the mean



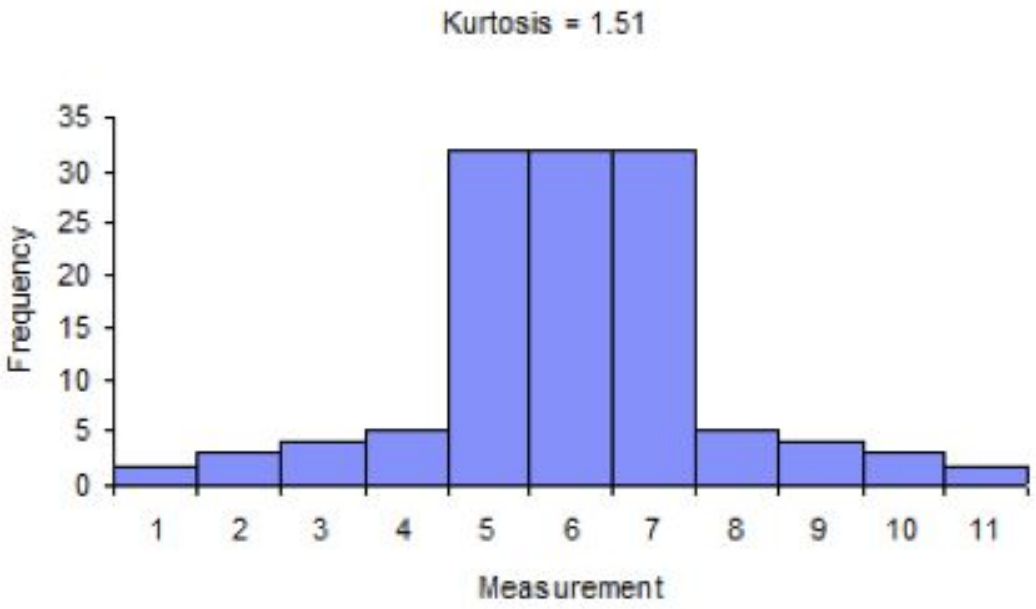


$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right)^2}-3$$

Descriptive Statistics

srijan:

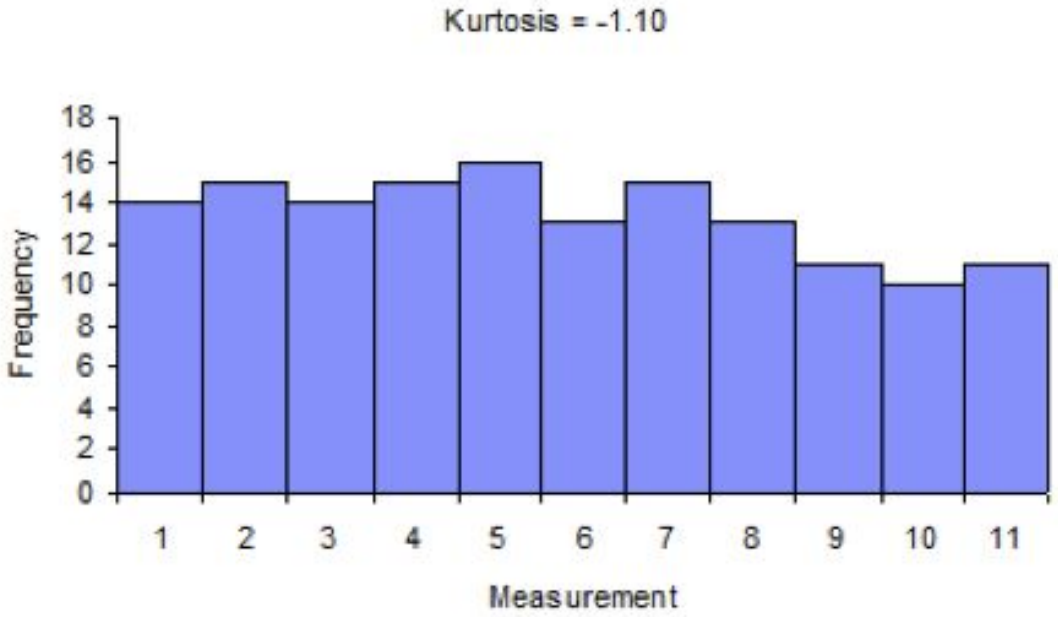
Kurtosis



Descriptive Statistics

srijan:

Kurtosis

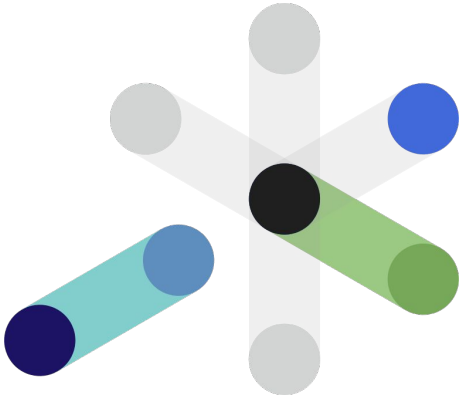


Descriptive Statistics

srijan:

Some Real life examples of Kurtosis

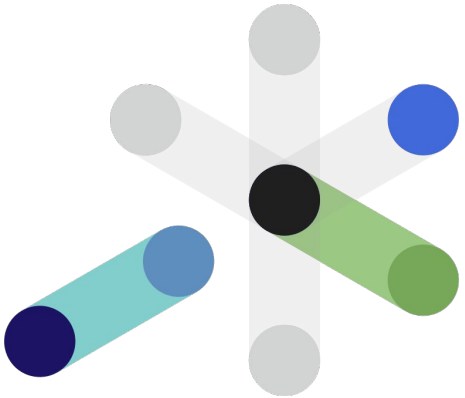
- Example for
  - Kurtosis
    - Outlier detection
    - Deciding over proper sample size for Stock Returns



Univariate Analysis

srijan:

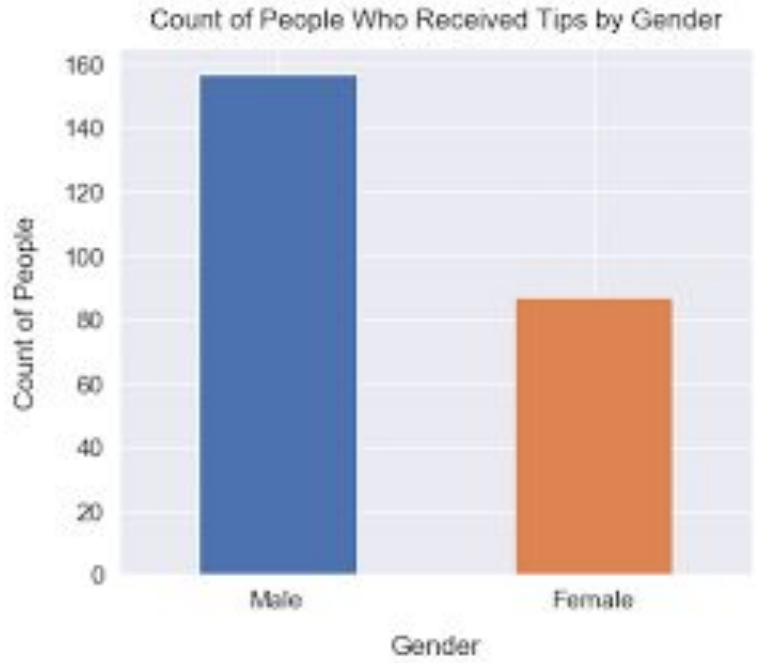
- “Uni” means one and “Variate” means variable
- It is the simplest form of analysis where single variable is analyzed
- Variable could be categorical or continuous
- Most commonly used methods includes:
  - Categorical variables
    - Count plot
  - Continuous variables
    - Histogram plot
    - Box plot

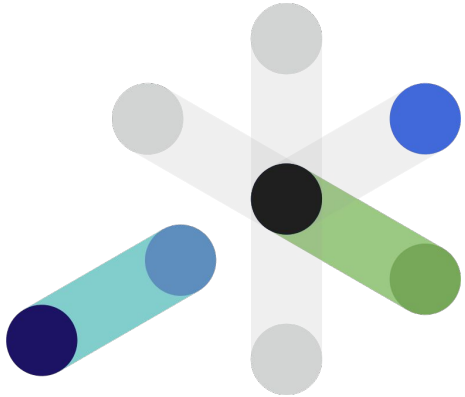


Univariate Analysis

srijan:

- Categorical variables
  - Count plot

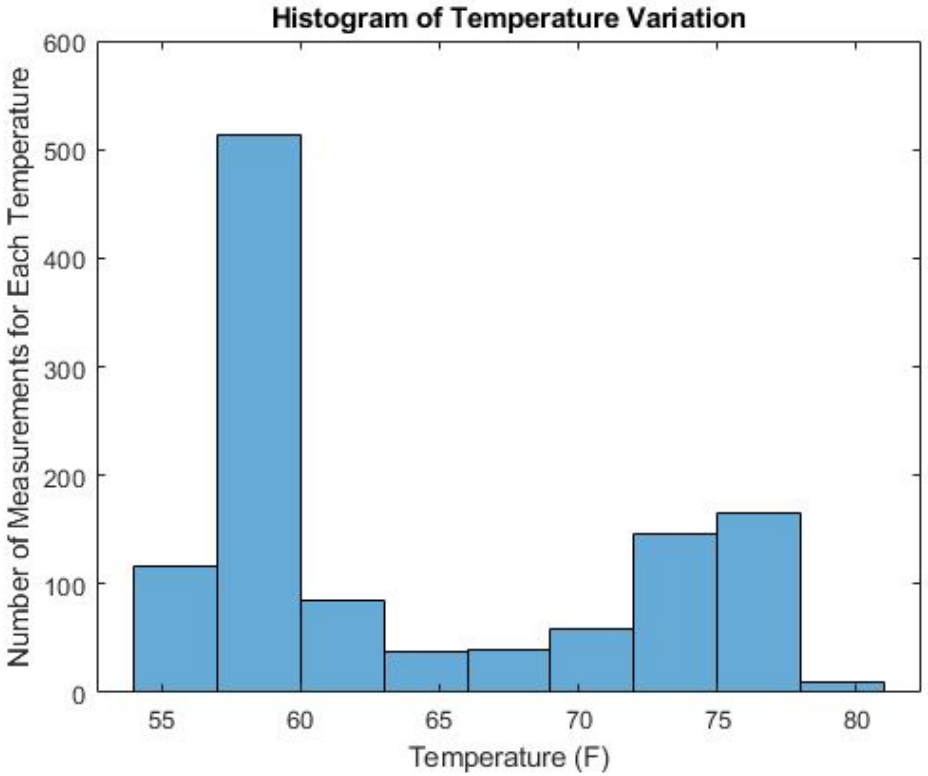


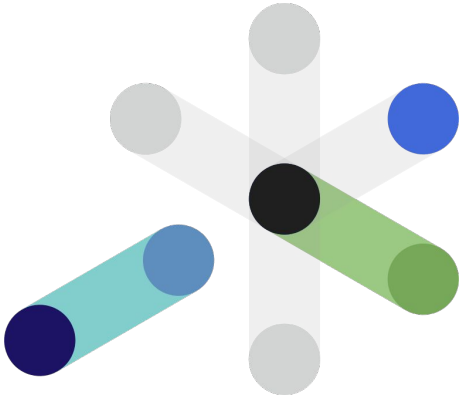


Univariate Analysis

srijan:

- Continuous variables
  - Histogram plot

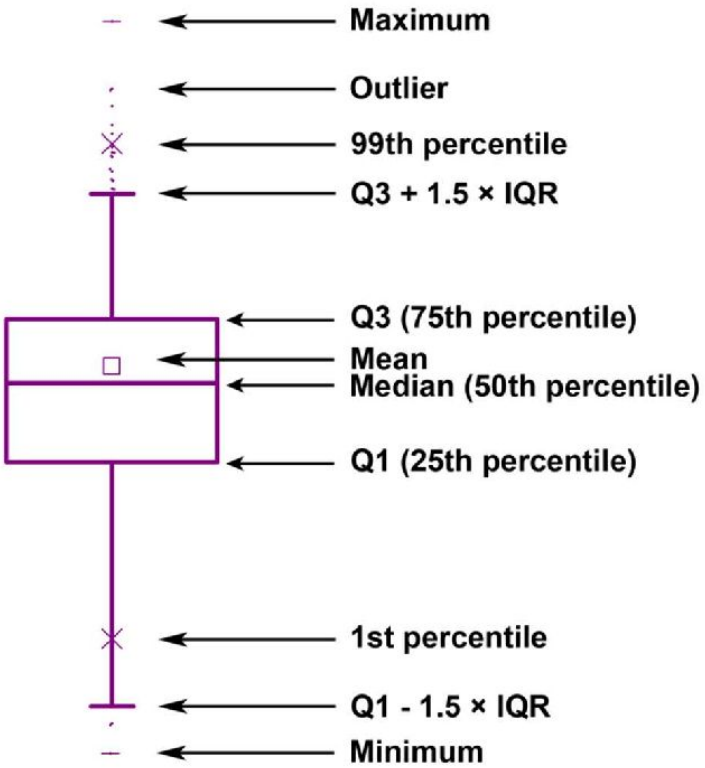




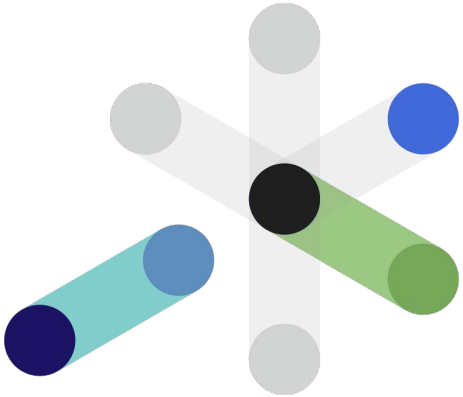
Univariate Analysis

srijan:

- Continuous variables
  - Box plot



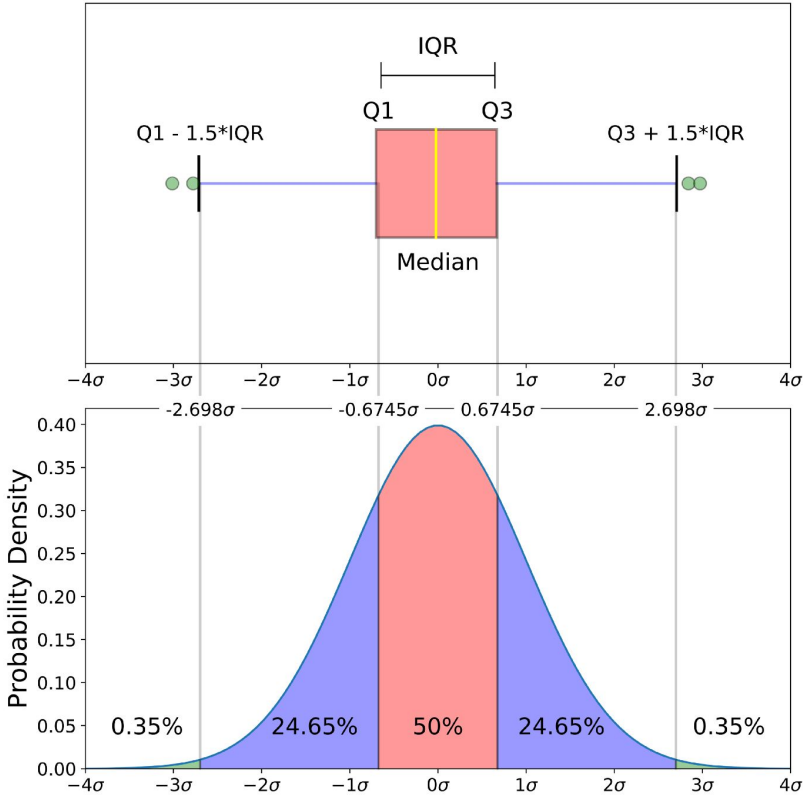
Note: IQR = Q3 - Q1

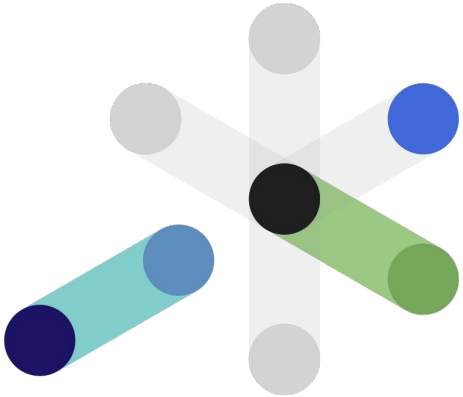


Univariate Analysis

srijan:

- Continuous variables
  - Box plot

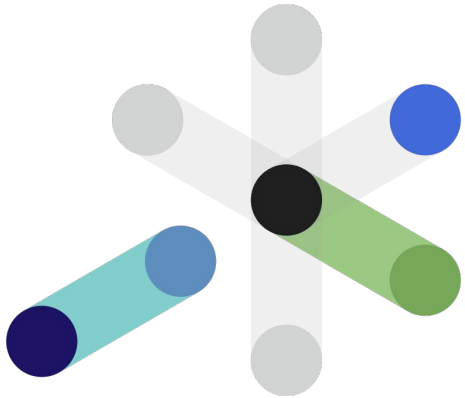




Bivariate Analysis

srijan:

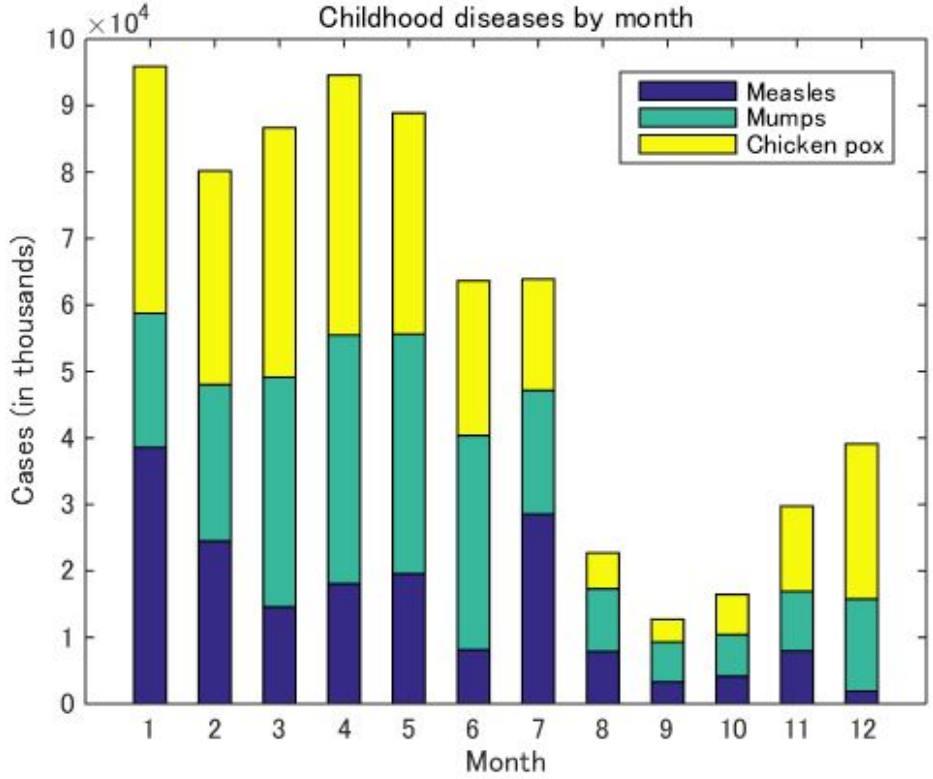
- “Bi” means two and “Variate” means variable
- Form of analysis where two variables are analyzed to determine relationship between each other
- Variable could be categorical or continuous
- Most commonly used methods includes:
  - Categorical variables
    - Stacked plot
    - Cross tables
  - Continuous variables
    - Pair plot

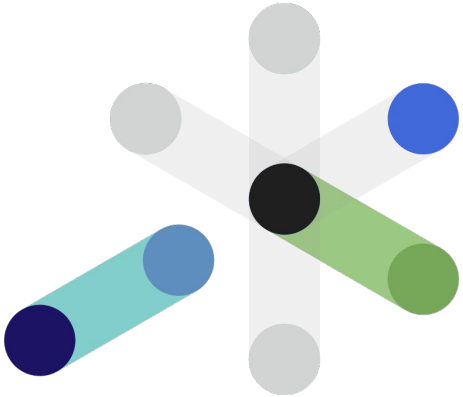


Bivariate Analysis

srijan:

- Categorical variables
  - Stacked plot



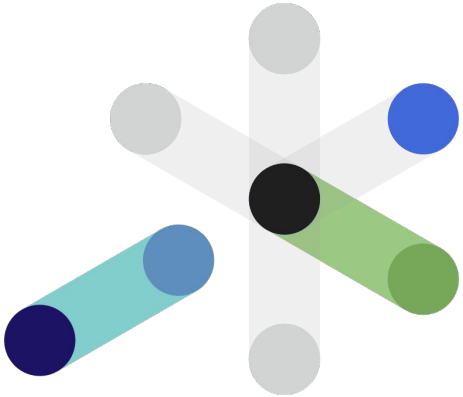


Bivariate Analysis

srijan:

- Categorical variables
  - Cross Tables

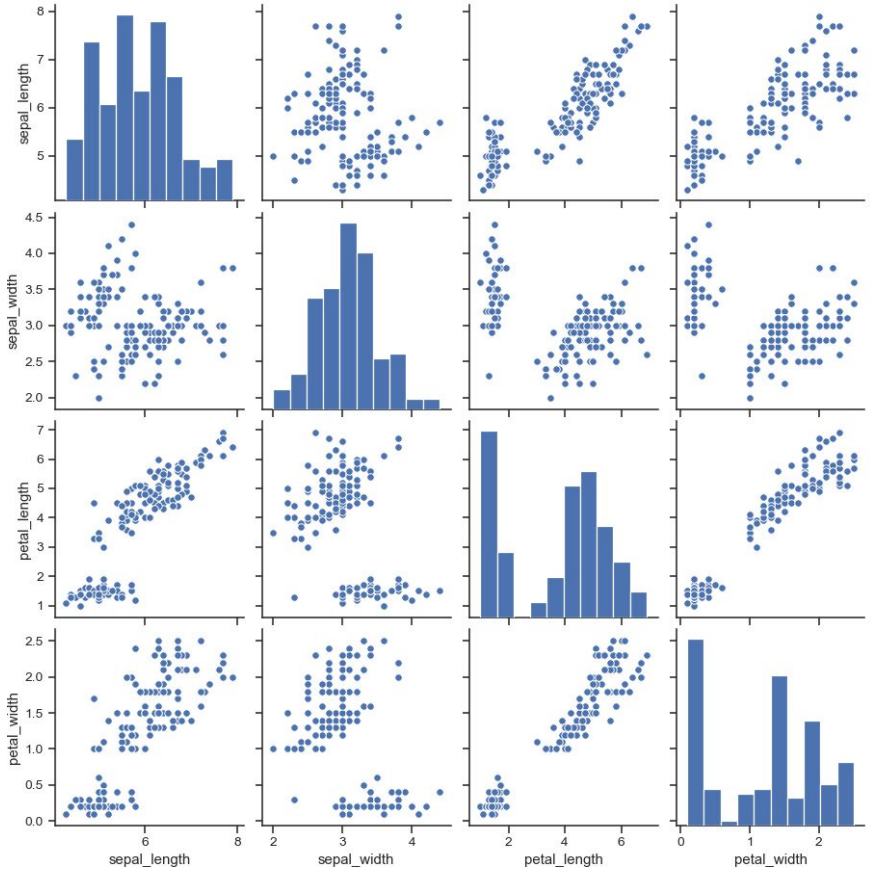
Color Car	black	blue	brown	gray	green	red	white
BMW	1	0	0	0	0	0	0
Ford	3	0	0	1	0	0	1
Honda	0	1	0	0	0	2	0
Toyota	0	0	0	0	1	1	0
Volvo	0	2	1	0	0	0	0

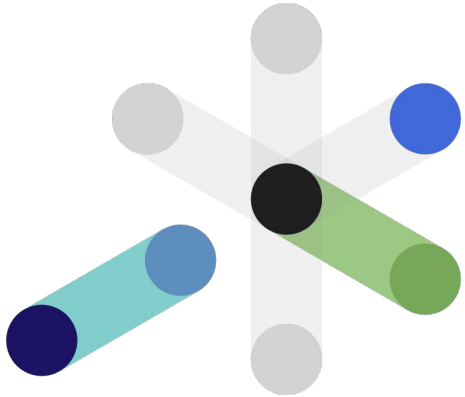


Bivariate Analysis

srijan:

- Continuous variables
  - Pair plot

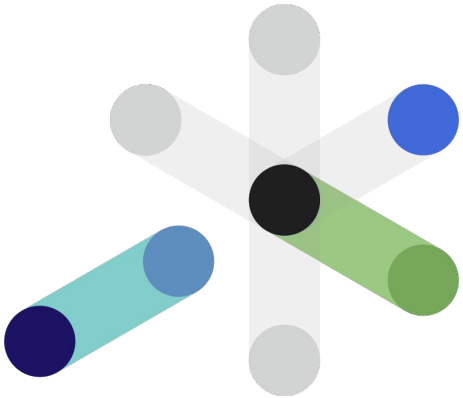




Dimensionality Reduction

srijan:

- Transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its intrinsic dimension.
- It is very difficult to visualize data in higher dimensions so reducing our space to 2D or 3D may allow us to plot and observe patterns more clearly
- Most commonly used techniques:
    - Low Variance Filter
    - Backward Feature Elimination



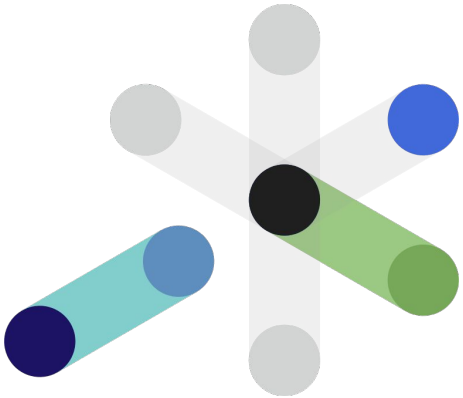
Dimensionality Reduction

srijan:

- Most commonly used techniques:
  - Low Variance Filter : Filtering out variables with low variance
    - Calculate the variance of each variable we are given.
    - Then drop the variables having low variance as compared to other variables in our dataset. Since variables with a low variance will not affect the target variable.

```
train.var()
```

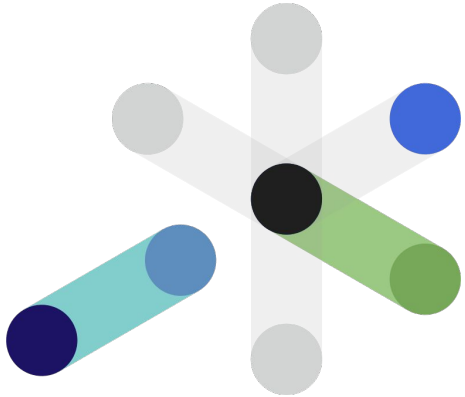
```
Item_Weight                  1.786956e+01
Item_Visibility              2.662335e-03
Item_MRP                     3.878184e+03
Outlet_Establishment_Year    7.008637e+01
Item_Outlet_Sales            2.912141e+06
dtype: float64
```

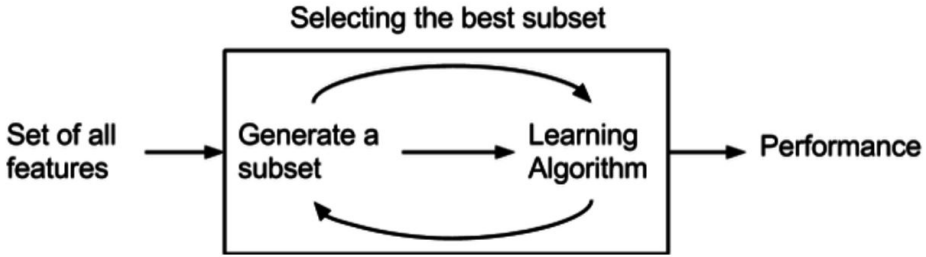


Dimensionality Reduction

srijan:

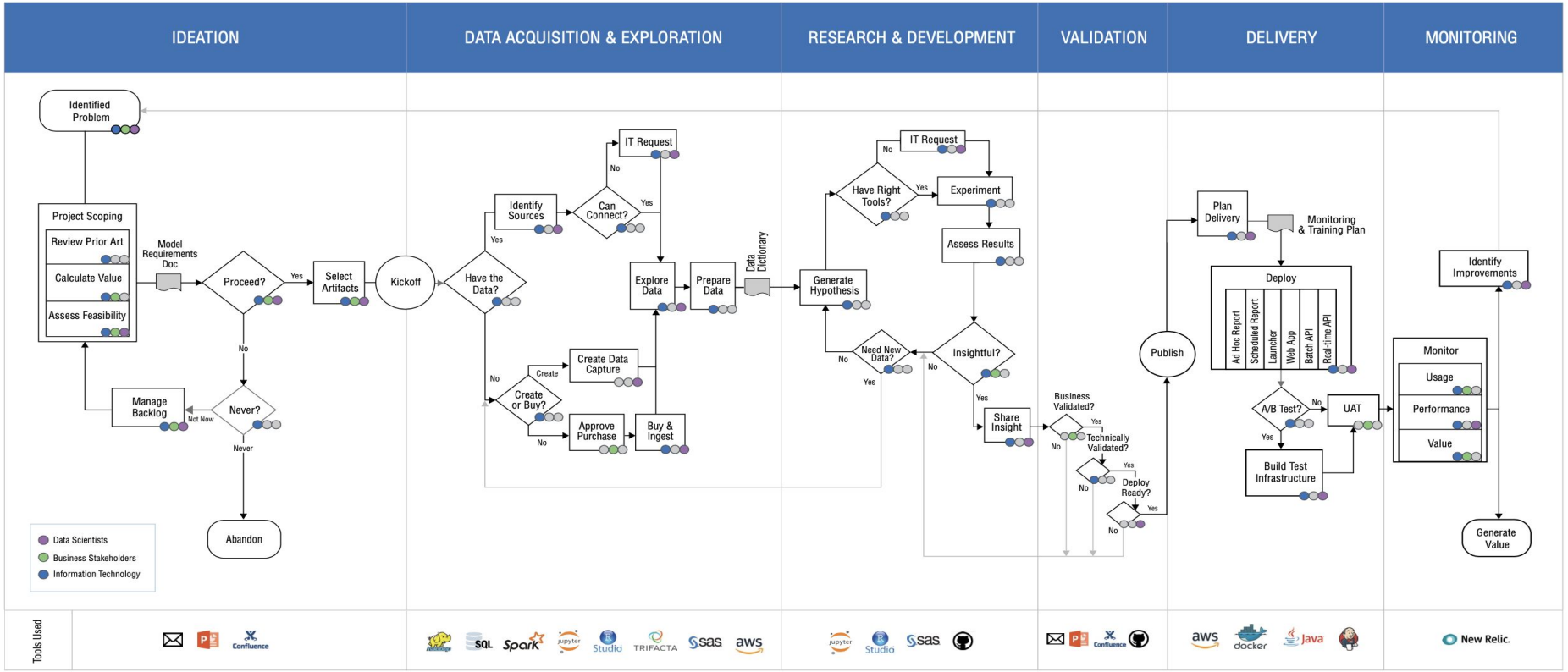
- Most commonly used techniques:
  - Backward Feature Elimination :
    - We first take all the n variables present in our dataset and train the model using them
    - We then calculate the performance of the model and save it.
    - Now, we compute the performance of the model after eliminating each variable (n times), i.e., we drop one variable every time and train the model on the remaining n-1 variables
    - We identify the variable whose removal has produced the smallest (or no) change in the performance of the model, and then drop that variable
    - Repeat this process until no variable can be dropped





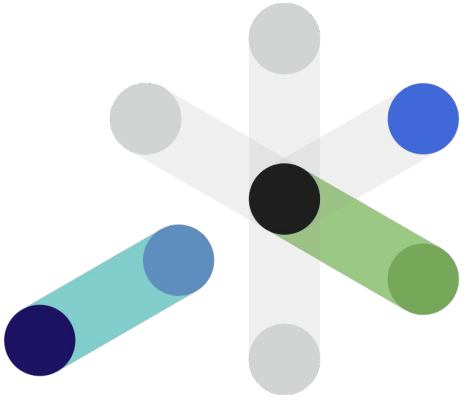
Process of Bootstrapping a Machine Learning Project

srijan:



Source: Domino Data Lab

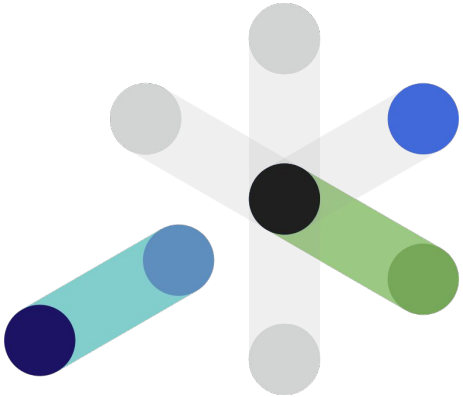
Any Questions ?



References/Credits

srijan:

- https://towardsdatascience.com/a-gentle-introduction-to-exploratory-data-analysis-f11d843b8184
- https://www.analyticsvidhya.com/blog/2018/08/dimensionality-reduction-techniques-python/
- https://www.statisticshowto.com/clustering/
- https://www.kaggle.com/residentmario/univariate-plotting-with-pandas
- https://medium.com/@purnasaigudikandula/exploratory-data-analysis-beginner-univariate-bivariate-and-multivariate-habberman-dataset-2365264b751
- https://adataanalyst.com/data-analysis-resources/visualise-categorical-variables-in-python/
- https://www.kaggle.com/residentmario/bivariate-plotting-with-pandas
- https://www.kaggle.com/etakla/exploring-the-dataset-bivariate-analysis
- https://dfrieds.com/data-visualizations/bar-plot-python-pandas.html
- https://www.spcforexcel.com/knowledge/basic-statistics/are-skewness-and-kurtosis-useful-statistics#skewness
- https://help.gooddata.com/doc/en/reporting-and-dashboards/maql-analytical-query-language/maql-expression-reference/aggregation-functions/statistical-functions/predictive-statistical-use-cases/normality-testing-skewness-and-kurtosis



Thank You

srijan: