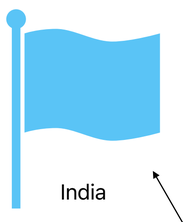


10 Billion - Population

India

Generalize the sample data anaysis to the population

Select the individual city of Bangalore - Sample

$2000

Statistical Analysis

Collect, Analyse and discover the patterns

Insights

1. Descriptive Analysis - summarisation - charts, table
2. Inferential Analysis -draw conclusion
3. Predictive Analysis - analyze the past trend and predict future events.
4. Prescriptive Analysis
5. EDA
6. Causal Analysis

Common Steps for Statistical Analysis:

1. Collect the data
2. Understand relationship between data
3. Create Model
4. Validate the model
5. Generate predictions

Common Methods for Statistical Analysis:

- Mean
- Standard Deviation
- Regression
- Hypothesis Testing
- Sampling

Descriptive Statistics

Create Summary of the data

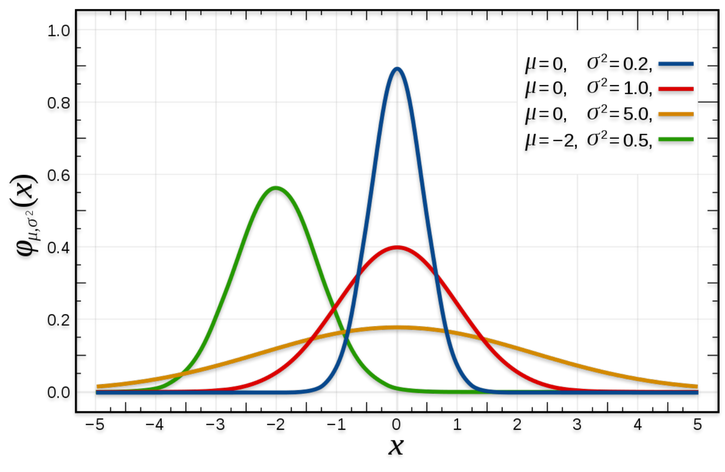
8.5 GPA

Dataset : 10 , 5 ,6,12 and 5

Sum = 38

Average/mean = 38/5 = 7.6

1. **Distribution (Frequency Distribution) :**



- **Measure of Central Tendency :**

Mean
Median
Mode

- **Measure of Variability :**

Range
Standard Deviation
Variance

Univariate , Bivariate and Multi Variate

Outliers

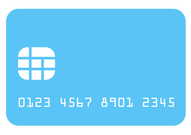
Case Study : Educational Achievement in a Sample

Most of the data of the subjects in range of 12 to 16 years
One of subject = 0
Another Subject = 30 Years

Ways to Find Outliers

- Observe the values with Box Plot (less than q1-1.5*iqr/ more than q3+1.5*iqr)
- Graphical methods

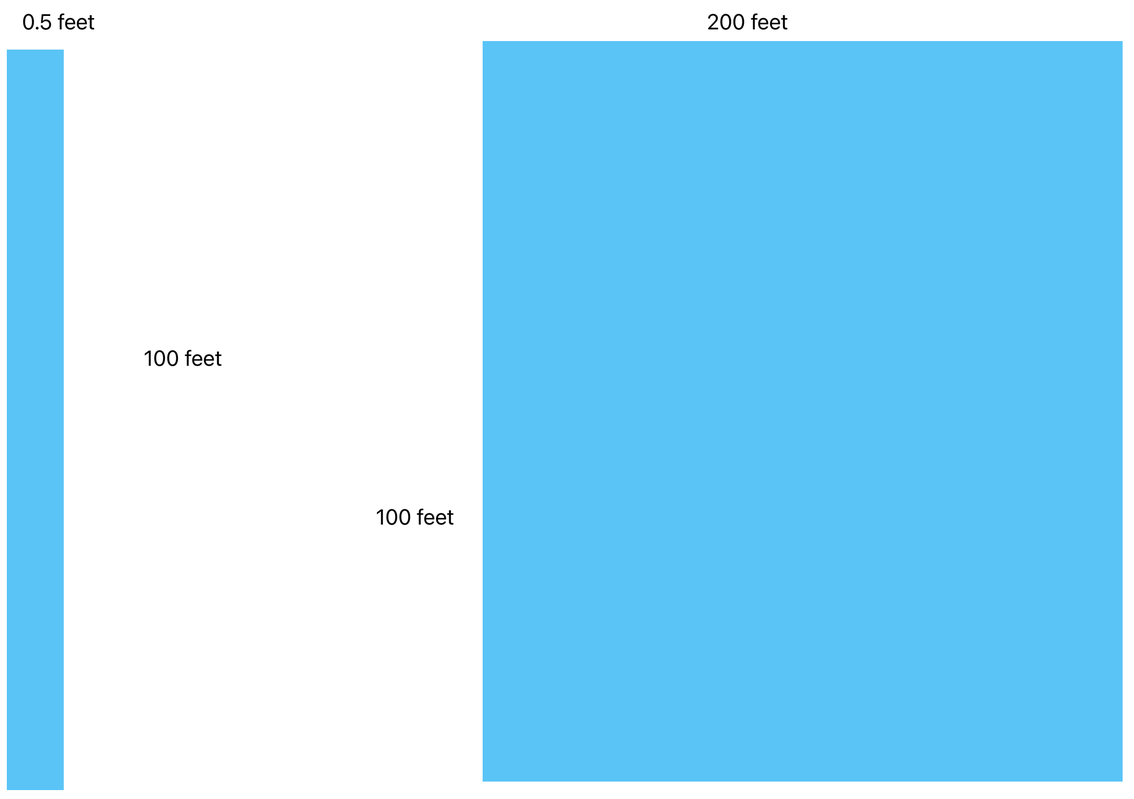
Imbalanced Dataset



Fraudulent - 100 Yes
99000 - No

Under sampling - remove random records from majority class
Over Sampling - duplicate the random records from minority class

Dimensionality Reduction - PCA



Dimensionality Reduction is a technique to reduce the number of features in the dataset.

- Remove unnecessary columns
- taking care of missing values
- remove the duplicate columns

Feature Extraction and Dimensionality Reduction