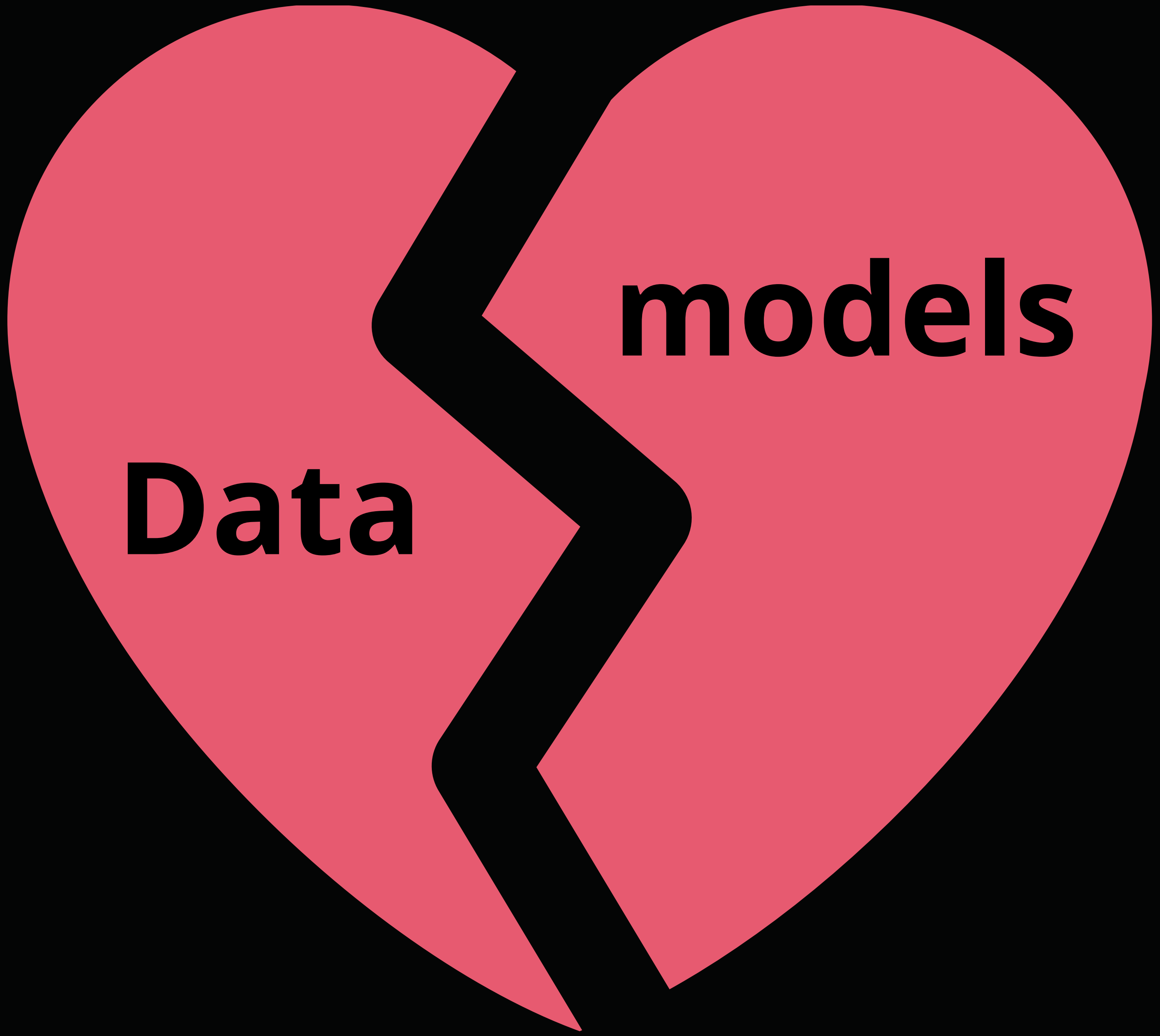
Exploratory data analysis

Stella Dong

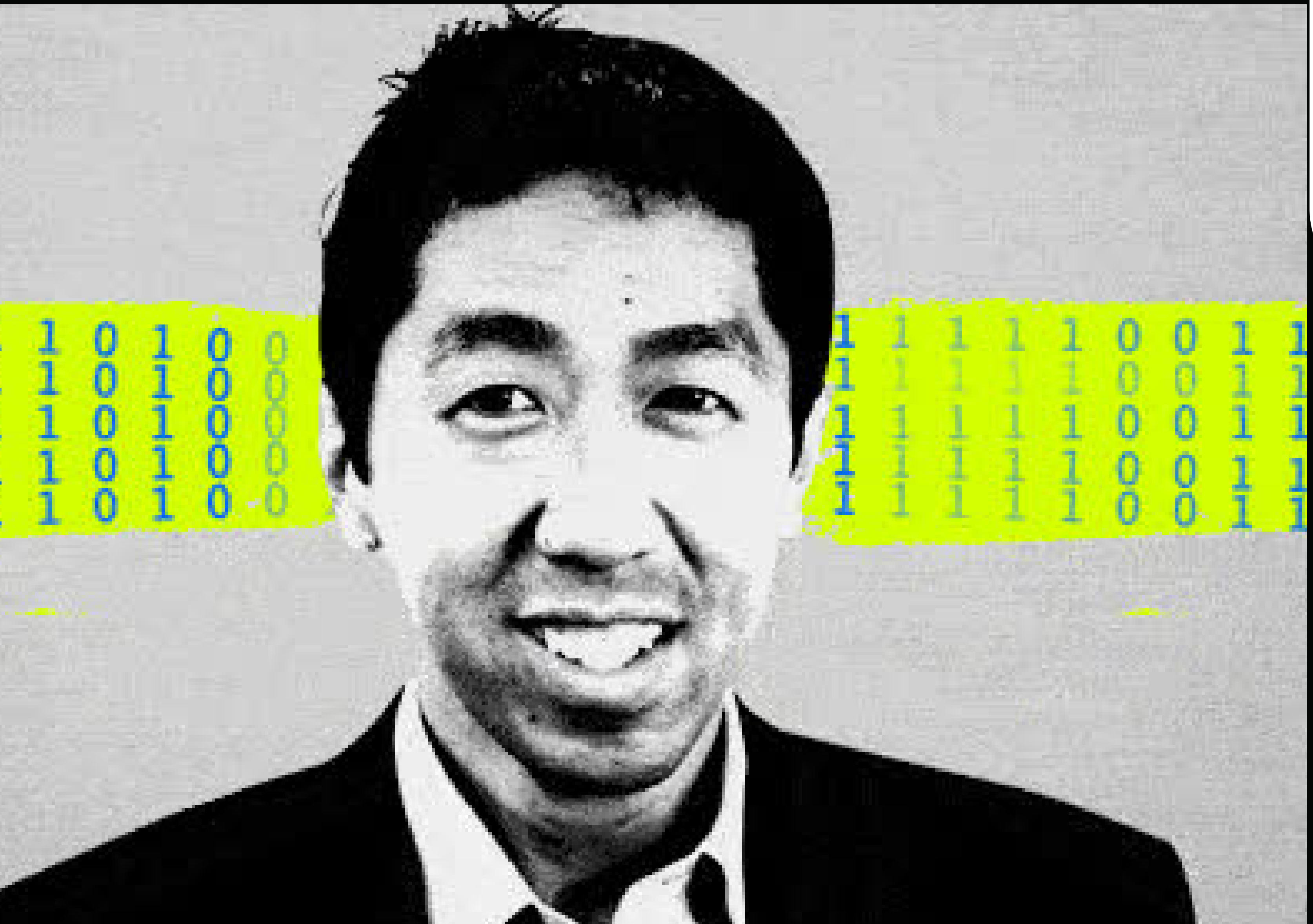
any AI project =



Data

models

Andrew Ng is Founder of DeepLearning.AI, General Partner at AI Fund, Chairman and Co-Founder of Coursera, and an Adjunct Professor at Stanford University.



“If 80 percent of our work is data preparation, then ensuring data quality is the important work of a machine learning team.”

What's EDA?

Exploratory Data Analysis (EDA) refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

Steps of Data Exploration and Preparation

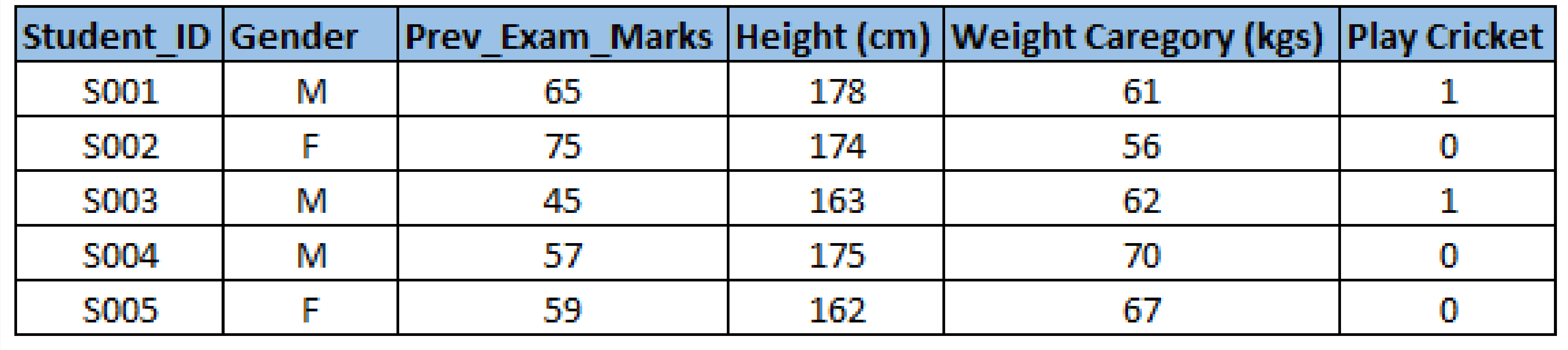
1. Variable Identification
2. Univariate Analysis
3. Bi-variate Analysis
4. Missing values treatment
5. Outlier treatment
6. Variable transformation
7. Variable creation

we will need to iterate over steps 4 – 7 multiple times before we come up with our refined model.

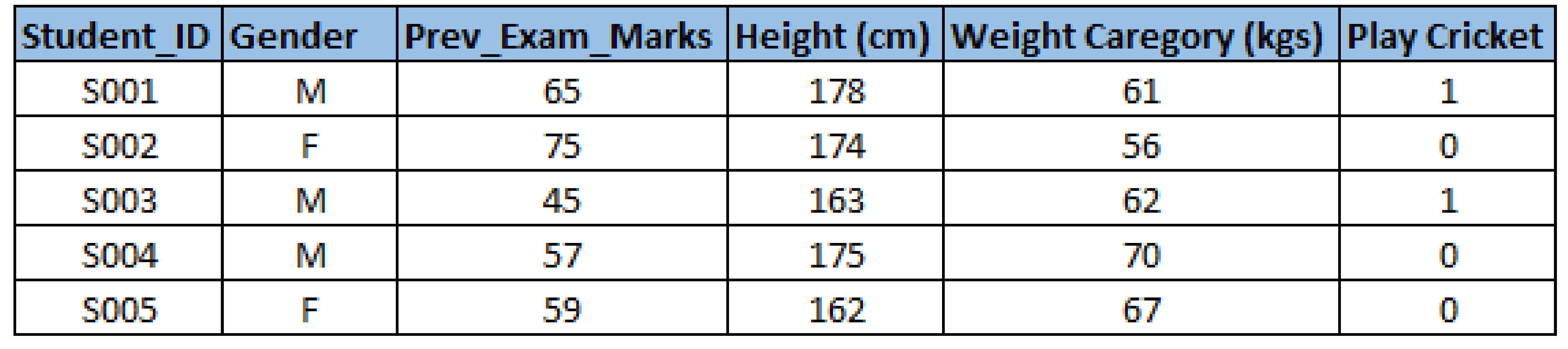
1. Variable Identification

Variable Identification

Suppose, we want to predict, whether the students will play cricket or not. You need to identify **predictor variables**, **target variable, data type of variables and category of variables.**



Student_ID	Gender	Prev_Exam_Marks	Height (cm)	Weight Caregory (kgs)	Play Cricket
S001	M	65	178	61	1
S002	F	75	174	56	0
S003	M	45	163	62	1
S004	M	57	175	70	0
S005	F	59	162	67	0



Student_ID	Gender	Prev_Exam_Marks	Height (cm)	Weight Caregory (kgs)	Play Cricket
S001	M	65	178	61	1
S002	F	75	174	56	0
S003	M	45	163	62	1
S004	M	57	175	70	0
S005	F	59	162	67	0

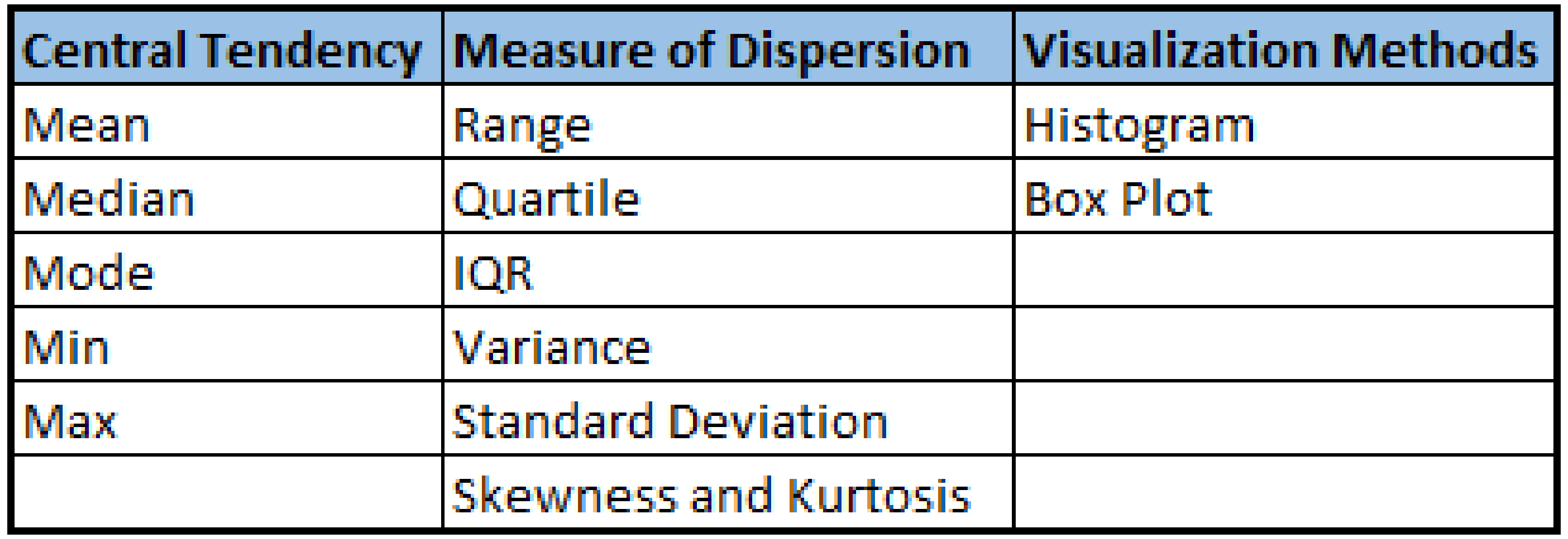
The variables have been defined in different category:

- **predictor variables:** Gender, Prev_Exam_Marks, Height, Weight
- **target variable:** Play Cricket
- **numeric:** Prev_Exam_Marks, Height, Weight
- **categorical:** Gender, Play Cricket

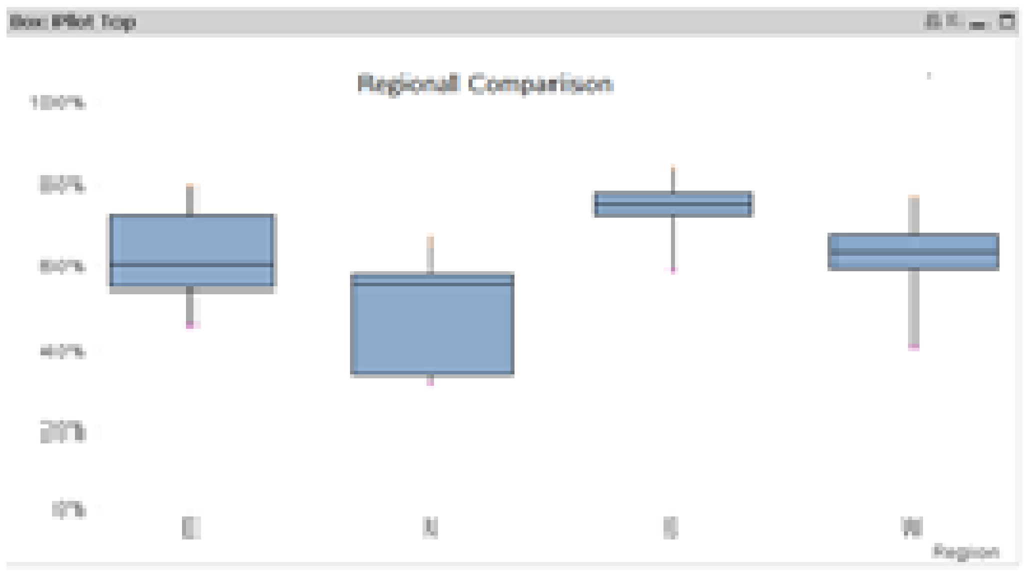
2. Univariate Analysis

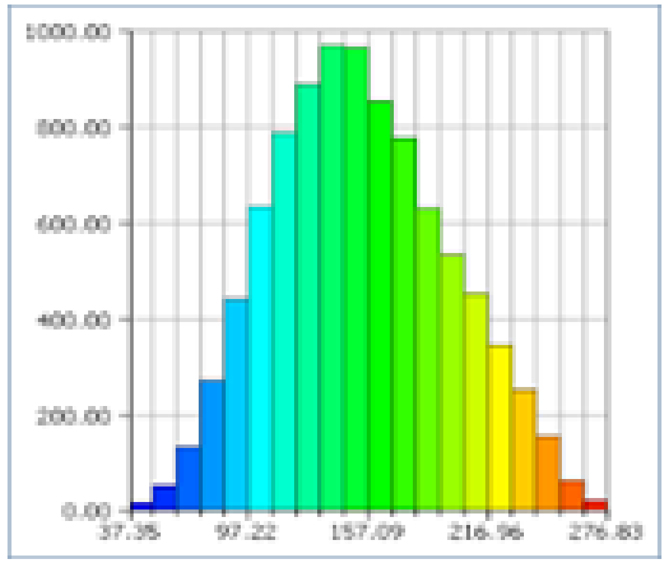
Univariate Analysis

Numeric Variables: we need to understand the **central tendency and spread of the variable**. These are measured using various statistical metrics visualization methods as shown below:



Central Tendency	Measure of Dispersion	Visualization Methods
Mean	Range	Histogram
Median	Quartile	Box Plot
Mode	IQR	
Min	Variance	
Max	Standard Deviation	
	Skewness and Kurtosis	





Categorical Variables: we use **frequency table** to understand distribution of each category. We can also read as percentage of values under each category. It can be be measured using two metrics, **Count and Count%** against each category. **Bar chart** can be used as visualization.

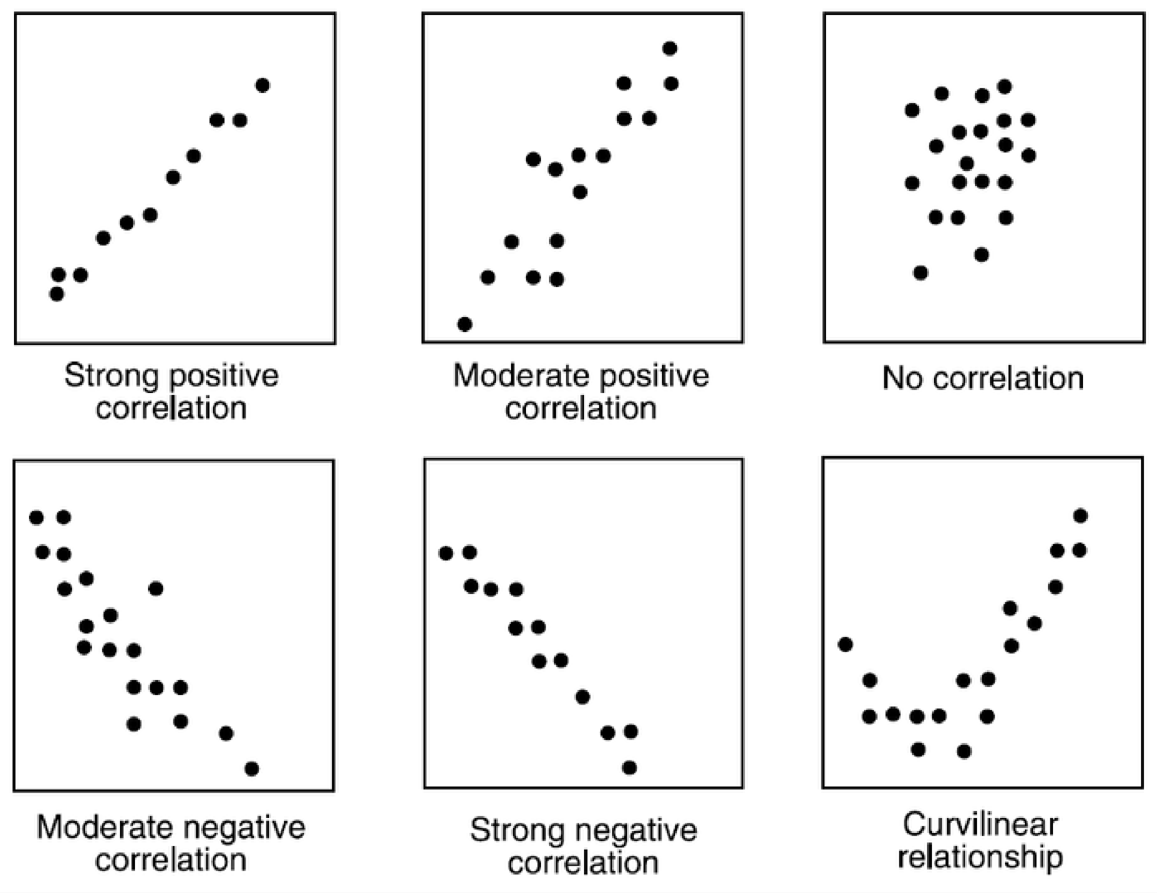
3. Bi-variate Analysis

Bi-variate Analysis

Bi-variate Analysis finds out the relationship between two variables. We can perform bi-variate analysis for any combination of categorical and numeric variables:

1. Numeric vs. Numeric
2. Categorical vs. Categorical
3. Categorical vs. Numeric

numeric & numeric: we use a scatter plot to find out the relationship between two variables. The pattern of scatter plot indicates the relationship between variables. The relationship can be linear or non-linear.

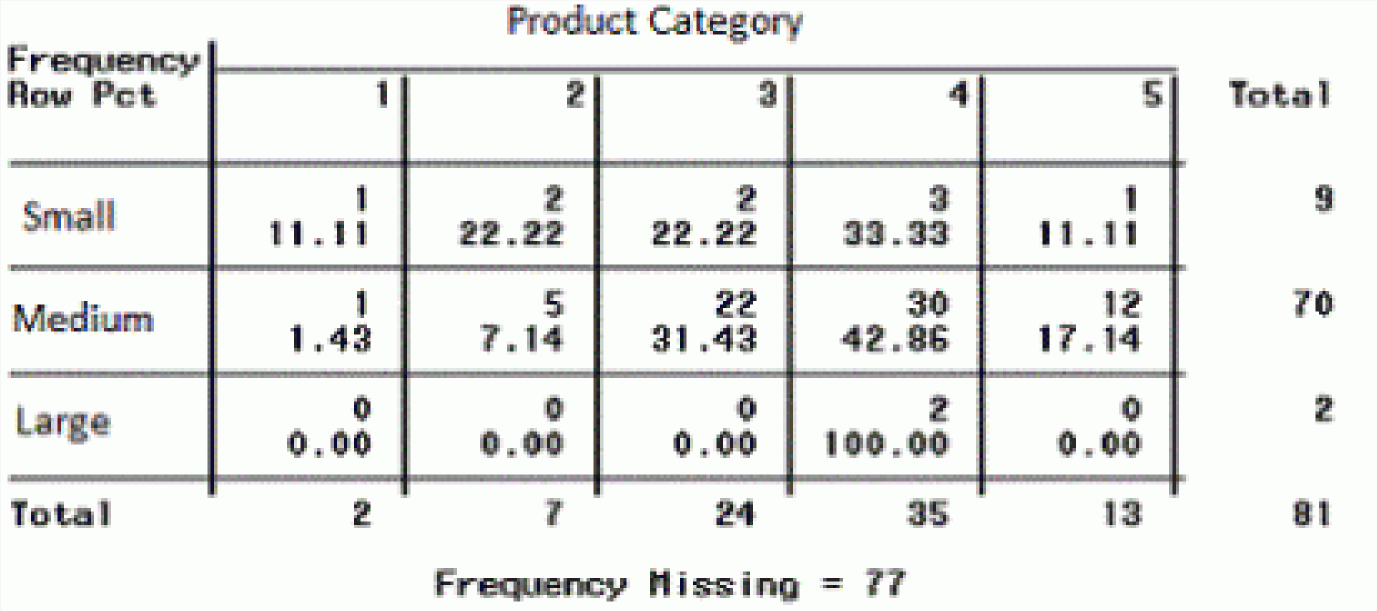


To find the strength of the relationship, we use Correlation. Correlation varies between -1 and +1.

- -1: perfect negative linear correlation
- +1:perfect positive linear correlation and
- 0: No correlation

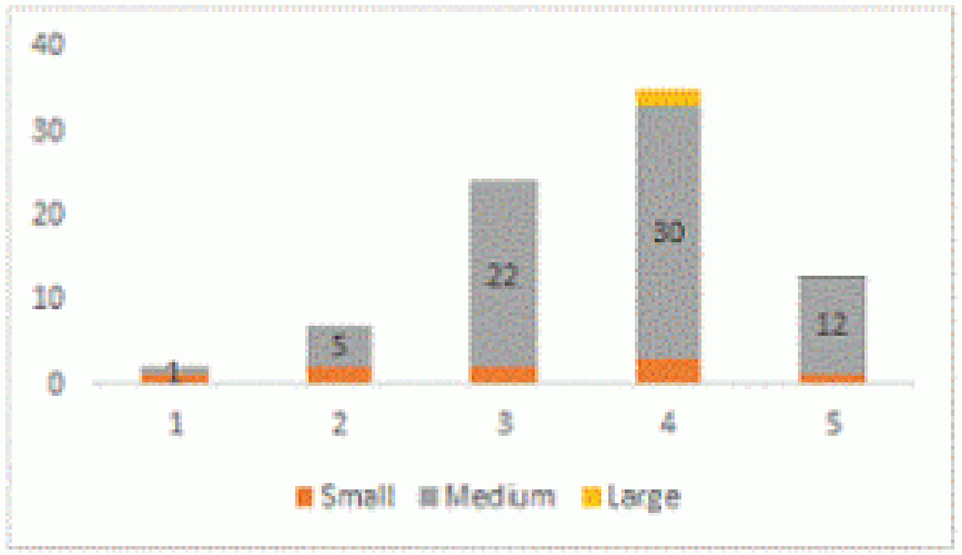
Categorical & Categorical: To find the relationship between two categorical variables, we can use following methods:

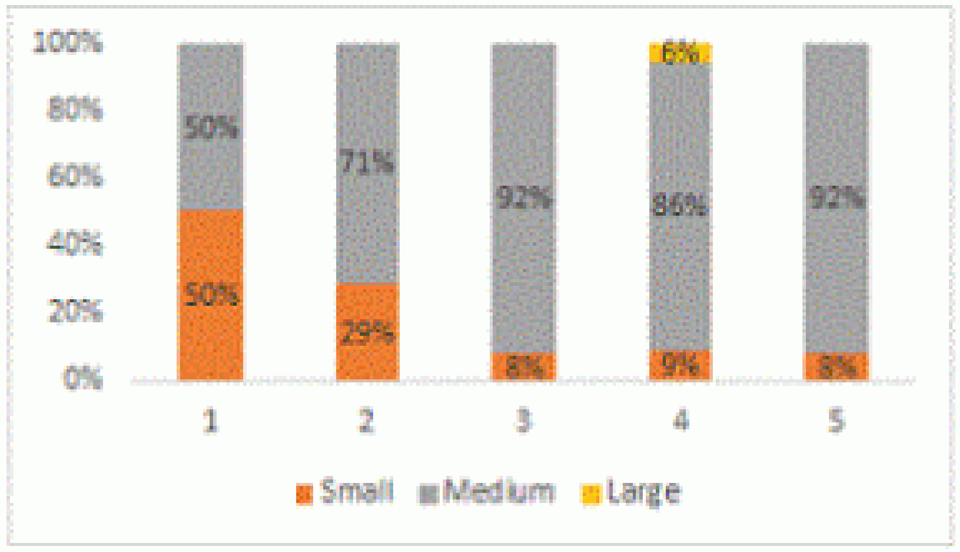
1. **Two-way table:** The rows represents the category of one variable and the columns represent the categories of the other variable. We show count or count% of observations available in each combination of row and column categories.
2. **Stacked Column Chart:** This method is more of a visual form of Two-way table.



	Product Category					
Frequency Row Pct	1	2	3	4	5	Total
Small	1 11.11	2 22.22	2 22.22	3 33.33	1 11.11	9
Medium	1 1.43	5 7.14	22 31.43	30 42.86	12 17.14	70
Large	0 0.00	0 0.00	0 0.00	2 100.00	0 0.00	2
Total	2	7	24	35	13	81

Frequency Missing = 77

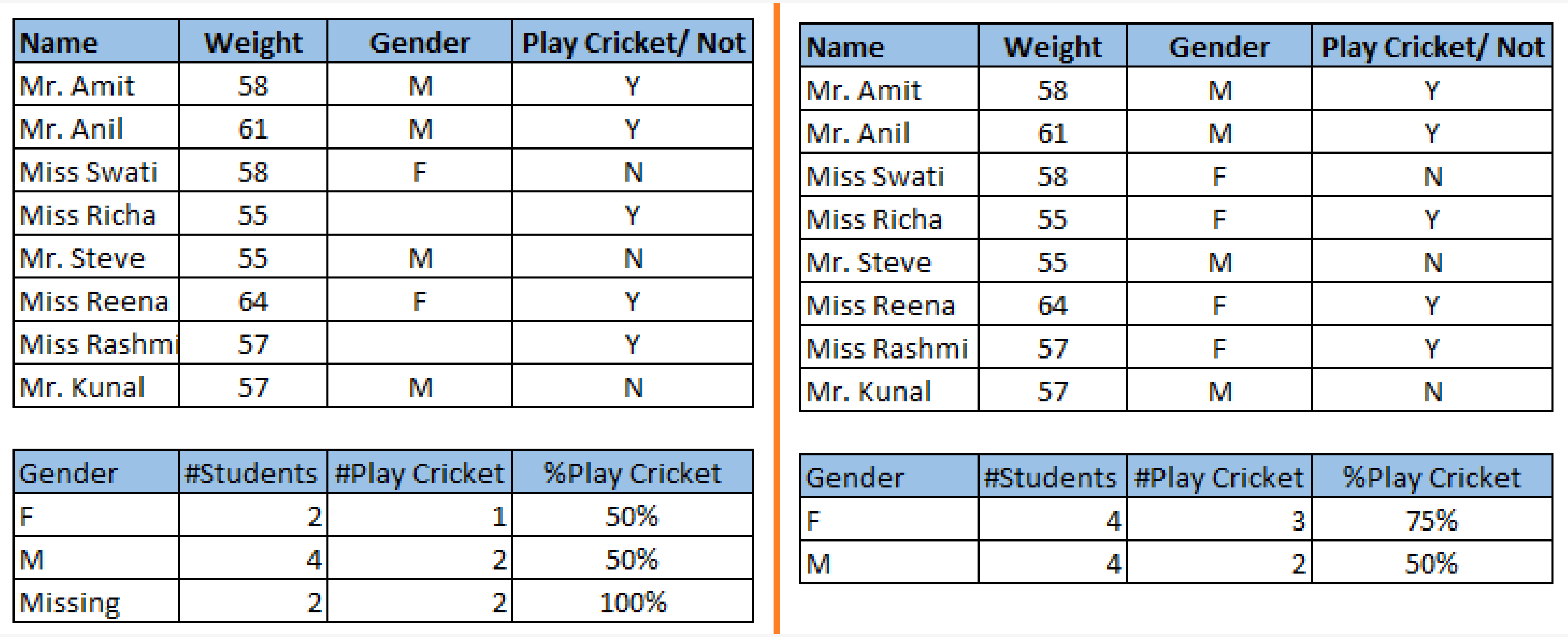




Categorical & Numeric: While exploring relation between categorical and Numeric variables, we can draw **box plots** for each level of categorical variables. If levels are small in number, it will not show the statistical significance. To look at the statistical significance we can perform Z-test, T-test or ANOVA.

4. Missing Value Treatment

Missing Value Treatment



Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55		Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57		Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	2	1	50%
M	4	2	50%
Missing	2	2	100%

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55	F	Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57	F	Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	4	3	75%
M	4	2	50%

In the left figure, we have not treated missing values. The inference from this data set is that the chances of playing cricket by males is higher than females. On the other hand, the right figure shows data after treatment of missing values (based on gender), we can see that females have higher chances of playing cricket compared to males.

Why my data has missing values?

Missing completely at random: This is a case when the probability of missing variable is same for all observations. For example: respondents of data collection process decide that they will declare their earning after tossing a fair coin.

Missing at random: This is a case when variable is missing at random and missing ratio varies for different values / level of other input variables. For example: We are collecting data for age and female has higher missing value compare to male.

Missing that depends on unobserved predictors: This is a case when the missing values are not random and are related to the unobserved input variable. For example: In a medical study, if a particular diagnostic causes discomfort, then there is higher chance of drop out from the study.

Missing that depends on the missing value itself: This is a case when the probability of missing value is directly correlated with missing value itself. For example: People with higher or lower income are likely to provide non-response to their earning.

Which are the methods to treat missing values ?

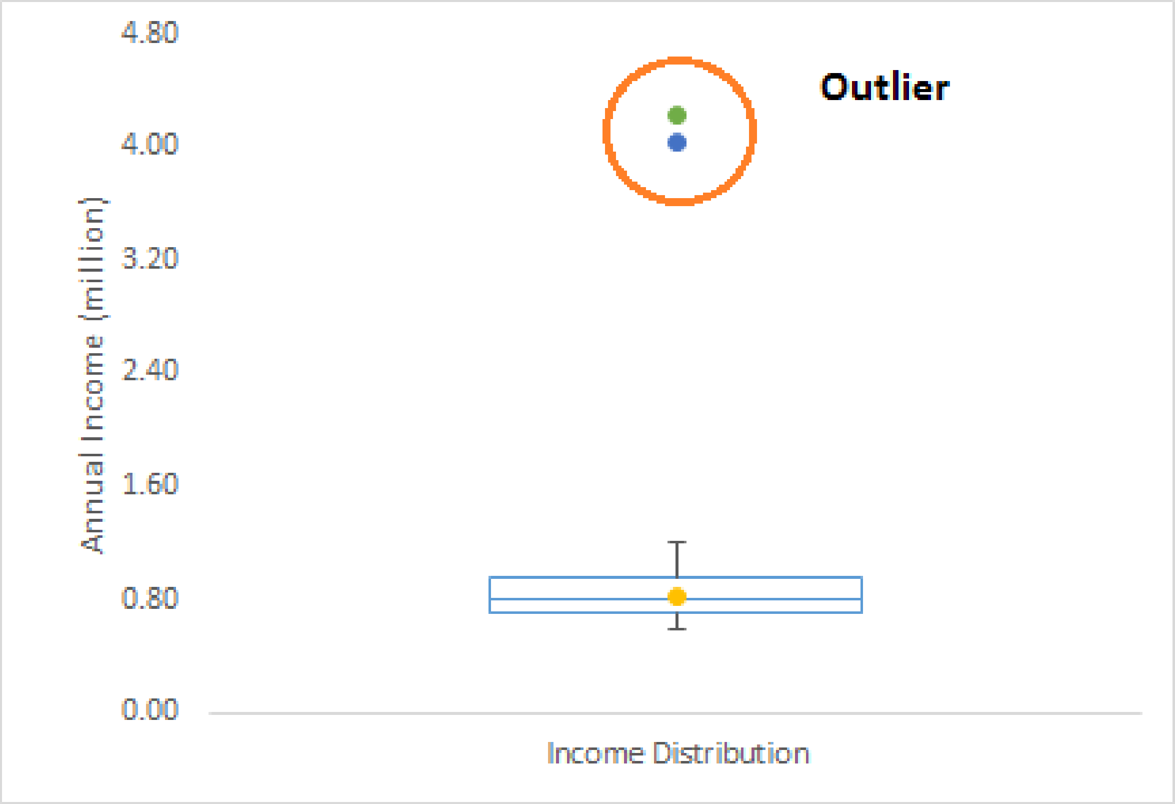
1. **Deletion of the missing rows if not many missing values**
2. **Mean/ Mode/ Median Imputation**
3. **Prediction Model:** In this case, we divide our data set into two sets: one set with no missing values for the variable and another one with missing values. We create a model (regression, ANOVA, Logistic regression, etc.) to predict target variable based on other attributes of the training data set and populate missing values of test data set. (Warnings: be aware that the estimated values are usually more well-behaved than the true values And, if there are no relationships with attributes in the data set and the attribute with missing values, then the model will not be precise for estimating missing values.)
4. **KNN (k-nearest neighbor) Imputation:** The missing values of an attribute are imputed using the given number of attributes that are most similar to the attribute whose values are missing. The similarity of two attributes is determined using a distance function. (Warning: KNN algorithm is very time-consuming in analyzing large database. It searches through all the dataset looking for the most similar instances. Choice of k-value is very critical. Higher value of k would include attributes which are significantly different from what we need whereas lower value of k implies missing out of significant attributes.)

5. Techniques of Outlier Detection and Treatment

What is an Outlier?

Outlier is an observation that appears far away and diverges from an overall pattern in a sample.

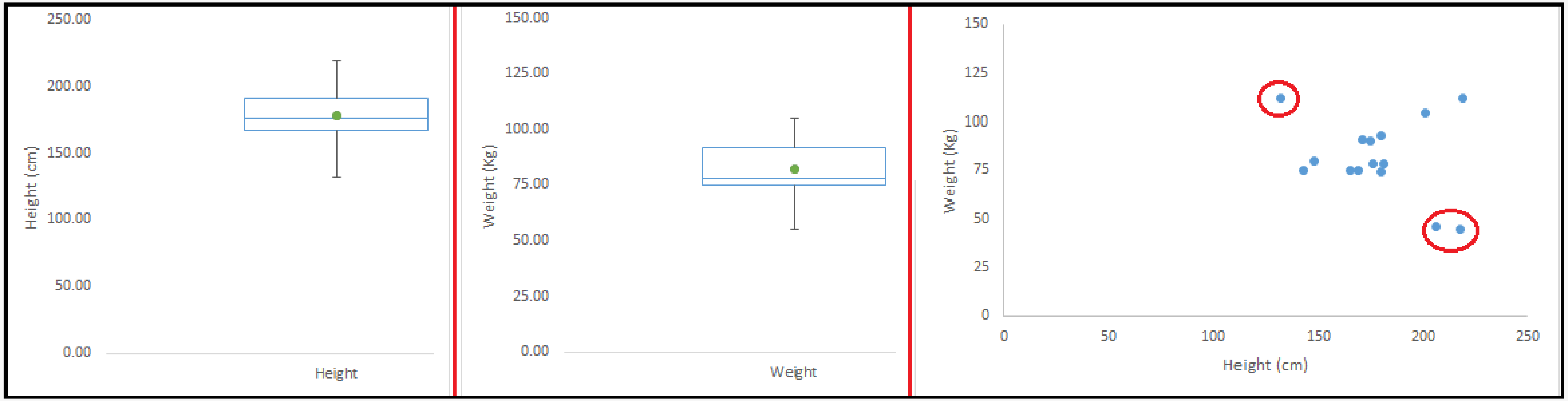
For example, the average annual income of customers is $0.8 million. But, there are two customers having annual income of $4 and $4.2 million. These two customers annual income is much higher than rest of the population. These two observations will be seen as Outliers.



What are the types of Outliers?

Outlier can be of two types: **Univariate and Multivariate**.

For example, the first two figures below, show the univariate distribution for Height, Weight, and they don't have outliers. However, the scatter plot on the right showing bivariate distribution for Height, Weight, which has outliers.



What causes Outliers?

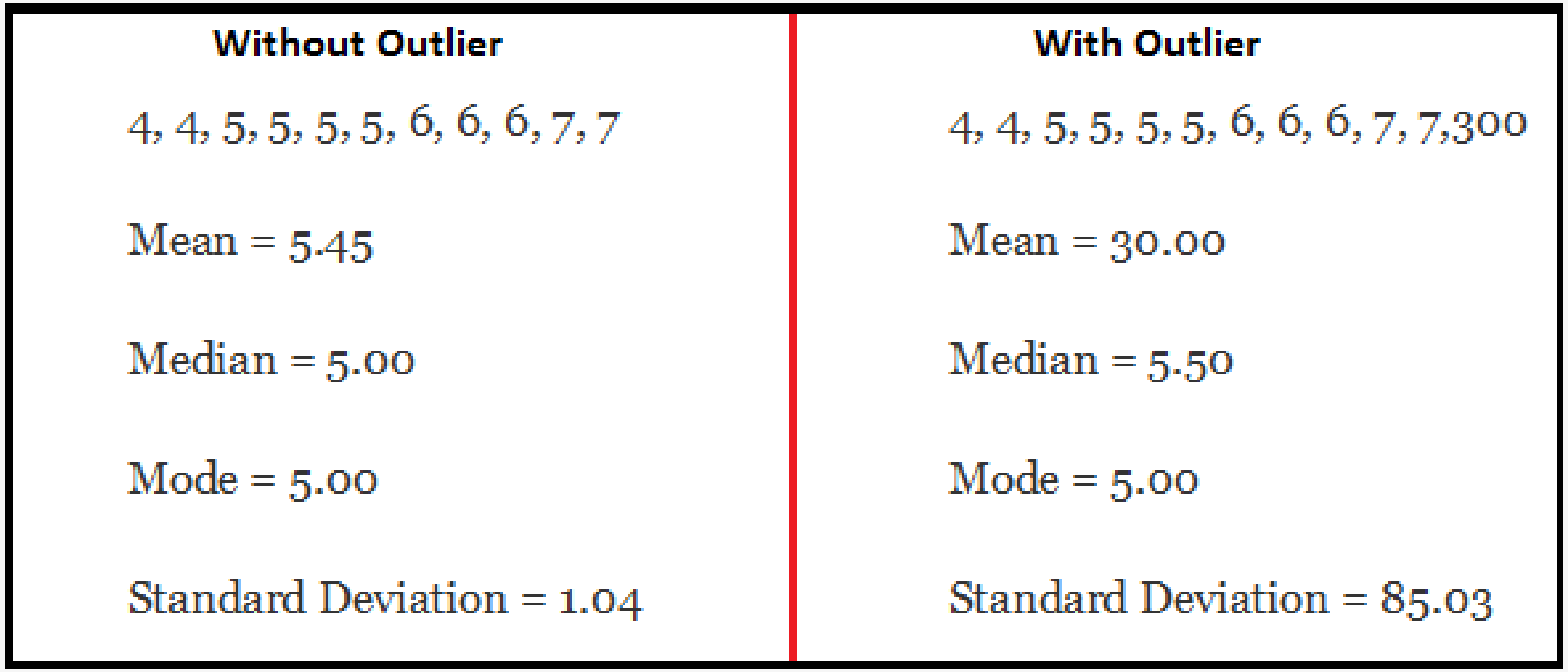
1. **Data Entry Errors:** Human errors such as errors caused during data collection, recording, or entry can cause outliers in data.
2. **Measurement Error:** It is the most common source of outliers. This is caused when the measurement instrument used turns out to be faulty.
3. **Experimental Error:** Another cause of outliers is experimental error.
4. **Intentional Outlier:** This is commonly found in self-reported measures that involves sensitive data.
5. **Data Processing Error:** Whenever we perform data mining, we extract data from multiple sources. It is possible that some manipulation or extraction errors may lead to outliers in the dataset.
6. **Sampling error:** For instance, we have to measure the height of athletes. By mistake, we include a few basketball players in the sample.
7. **Natural Outlier:** When an outlier is not artificial (due to error), it is a natural outlier.

What is the impact of Outliers on a dataset?

Outliers can change the results of the data analysis and statistical modeling. There are numerous unfavourable impacts of outliers in the data set:

- It **increases 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.**

Example of how outliers impact the dataset?



Without Outlier	With Outlier
4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7	4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7,300
Mean = 5.45	Mean = 30.00
Median = 5.00	Median = 5.50
Mode = 5.00	Mode = 5.00
Standard Deviation = 1.04	Standard Deviation = 85.03

As you can see, data set with outliers has significantly different mean and standard deviation. In the first scenario, we will say that average is 5.45. But with the outlier, average soars to 30. This would change the estimate completely.

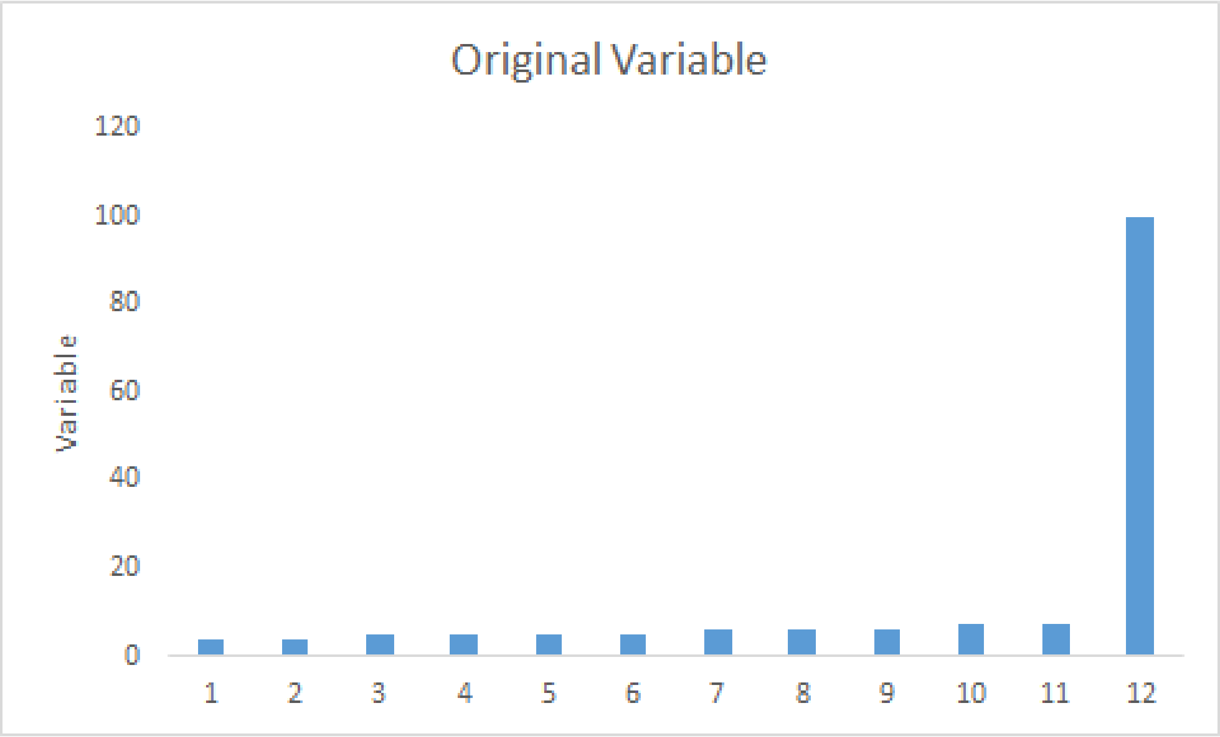
How to detect Outliers?

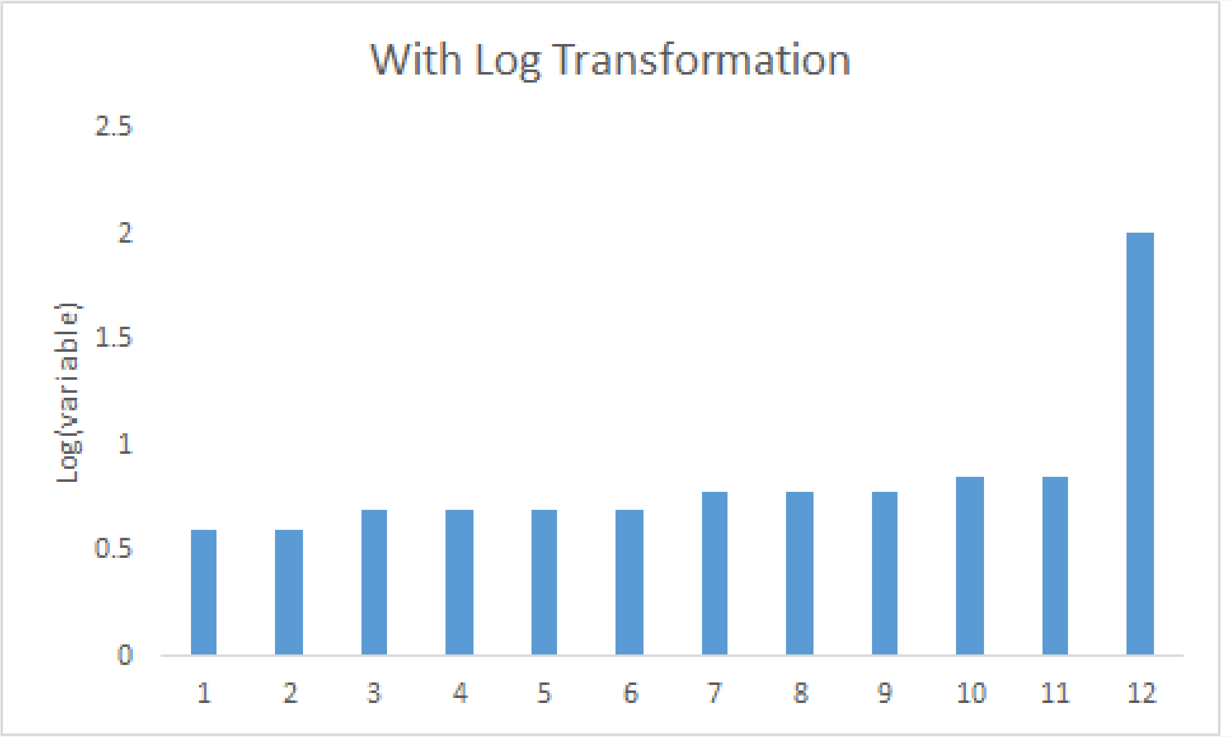
Most commonly used method to detect outliers is **visualization**. We use various visualization methods, like **Box-plot, Histogram, Scatter Plot.**

Some analysts also think that any value, which is **beyond the range of -1.5 x IQR to 1.5 x IQR**

How to remove Outliers?

1. **Deleting observations:** we delete outlier values if it is due to data entry error, data processing error or outlier observations are very small in numbers. We can also use trimming at both ends to remove outliers.
2. **Transforming and binning values:** Transforming variables can also eliminate outliers. Natural log of a value reduces the variation caused by extreme values. Binning is also a form of variable transformation.





The Art of Feature Engineering

What is Feature Engineering?

Feature engineering is the science (and art) of extracting more information from existing data. You are not adding any new data here, but you are actually making the data you already have more useful.

What is the process of Feature Engineering ?

You perform feature engineering once you have completed the first 5 steps in data exploration – Variable Identification, Univariate, Bivariate Analysis, Missing Values Imputation and Outliers Treatment.

Feature engineering itself can be divided in 2 steps:

- Variable transformation.
- Variable / Feature creation.

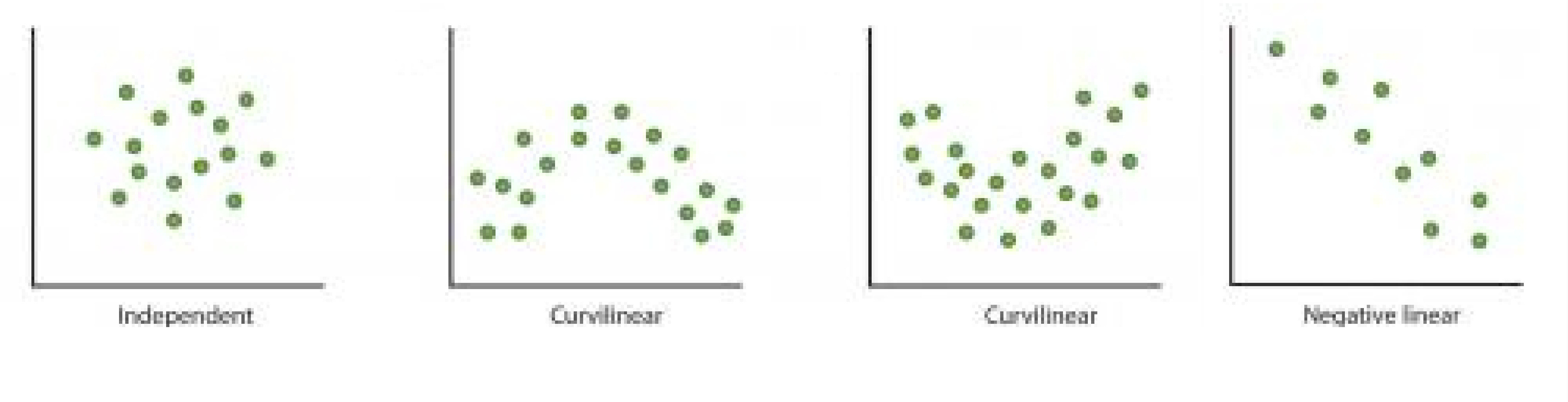
6. Variable Transformation

What is Variable Transformation?

In data modelling, transformation refers to the replacement of a variable by a function. For instance, replacing a variable x by the square / cube root or logarithm x is a transformation. In other words, transformation is a process that changes the distribution or relationship of a variable with others.

When should we use Variable Transformation?

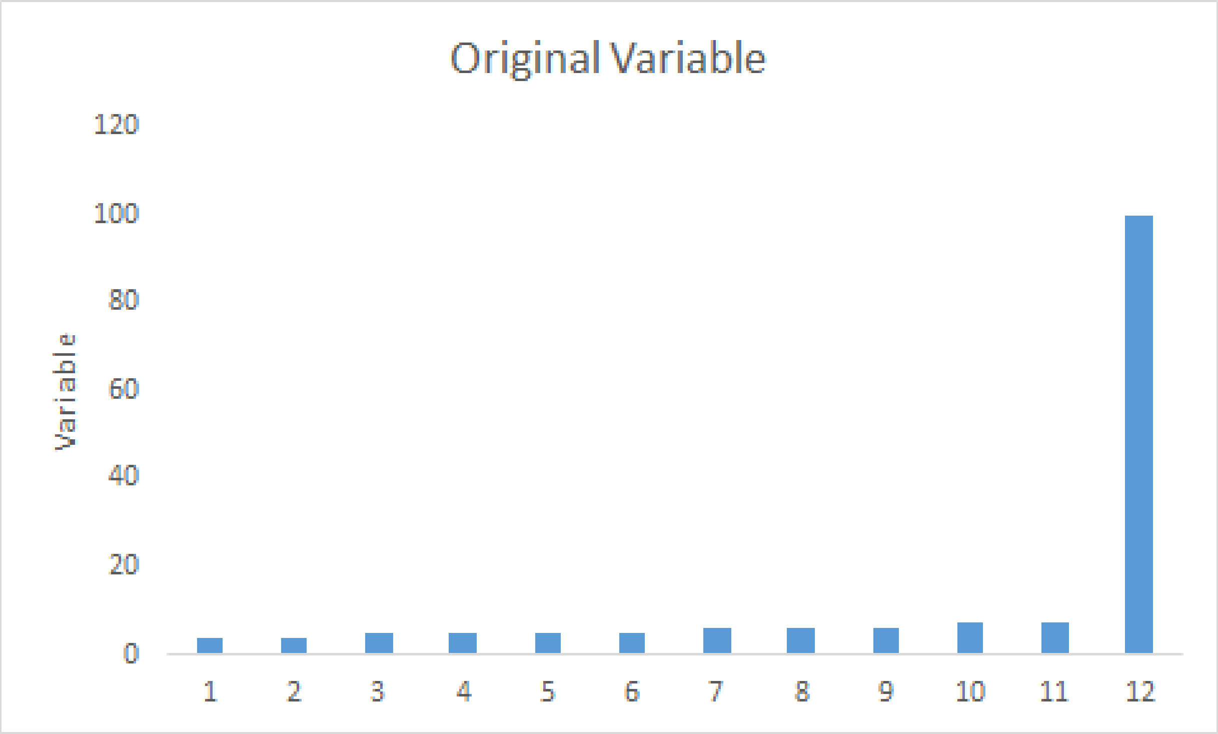
- When we want to change the scale of a variable or standardize the values of a variable for better understanding.
- When we can transform complex non-linear relationships into linear relationships. Scatter plot can be used to find the relationship between two numeric variables. These transformations also improve the prediction. Log transformation is one of the commonly used transformation technique used in these situations.

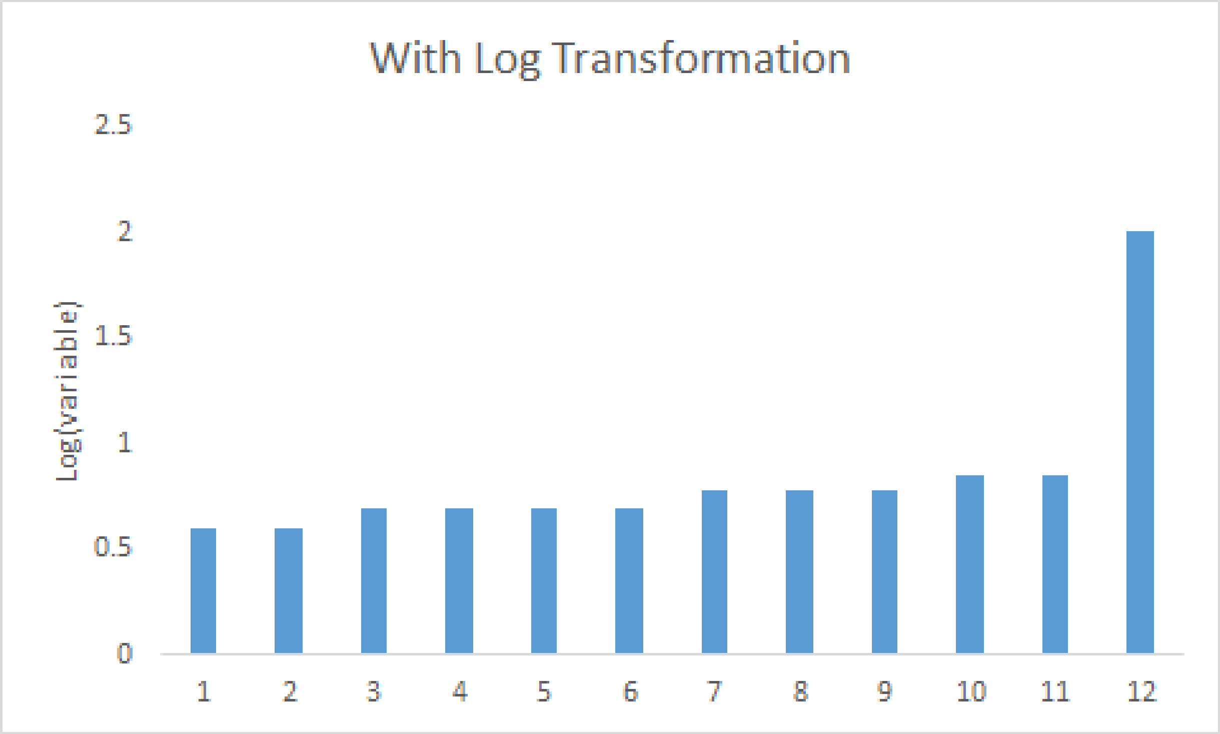


When should we use Variable Transformation?

Symmetric distribution is preferred over skewed distribution as it is easier to interpret and generate inferences. Some modeling techniques requires normal distribution of variables. So, whenever we have a skewed distribution, we can use transformations which reduce skewness.

For right skewed distribution, we take square / cube root or logarithm of variable and for left skewed, we take square / cube or exponential of variables.





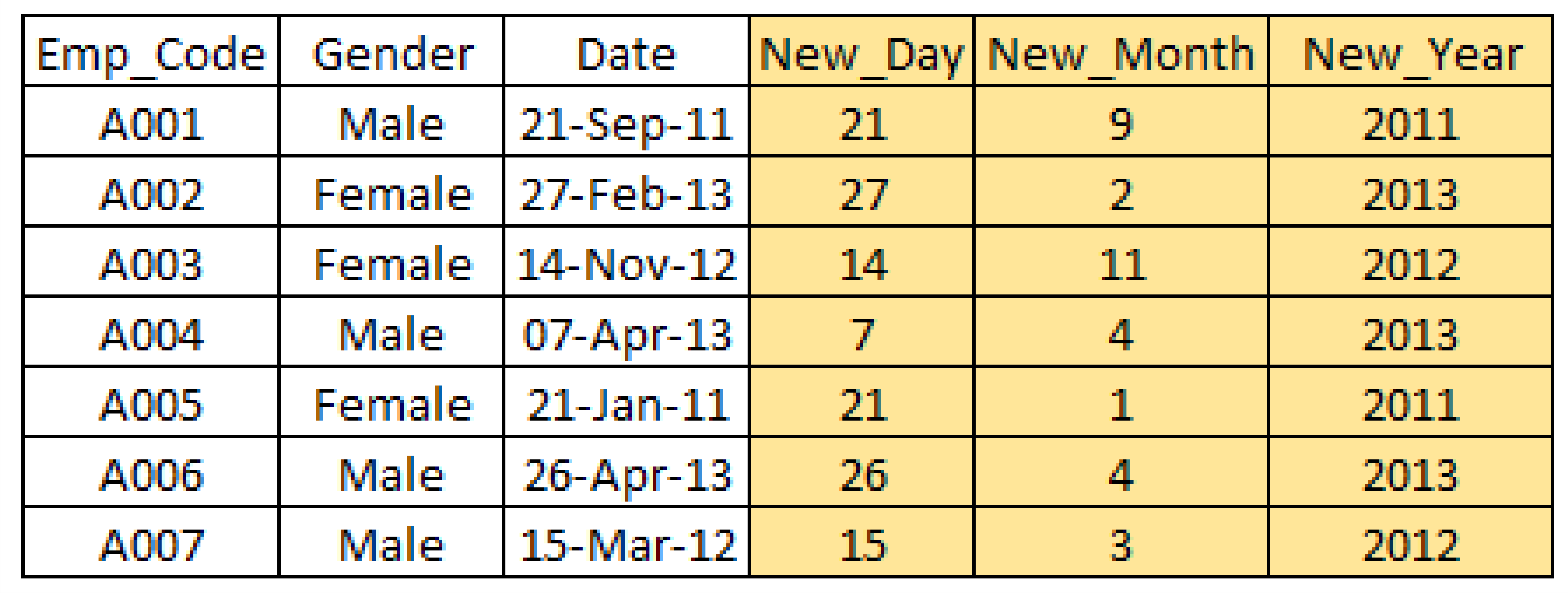
What are the common methods of Variable Transformation?

- Logarithm: Log of a variable is used to change the shape of distribution of the variable on a distribution plot. It is generally used for reducing right skewness of variables. Though, It can't be applied to zero or negative values as well.
- Square / Cube root: The square and cube root of a variable is not as significant as logarithmic transformation. Cube root has its own advantage. It can be applied to negative values including zero. Square root can be applied to positive values including zero.
- Binning: It is used to categorize variables. It is performed on original values, percentile or frequency.

7. What is Feature / Variable Creation & its Benefits?

What is Feature / Variable Creation & its Benefits?

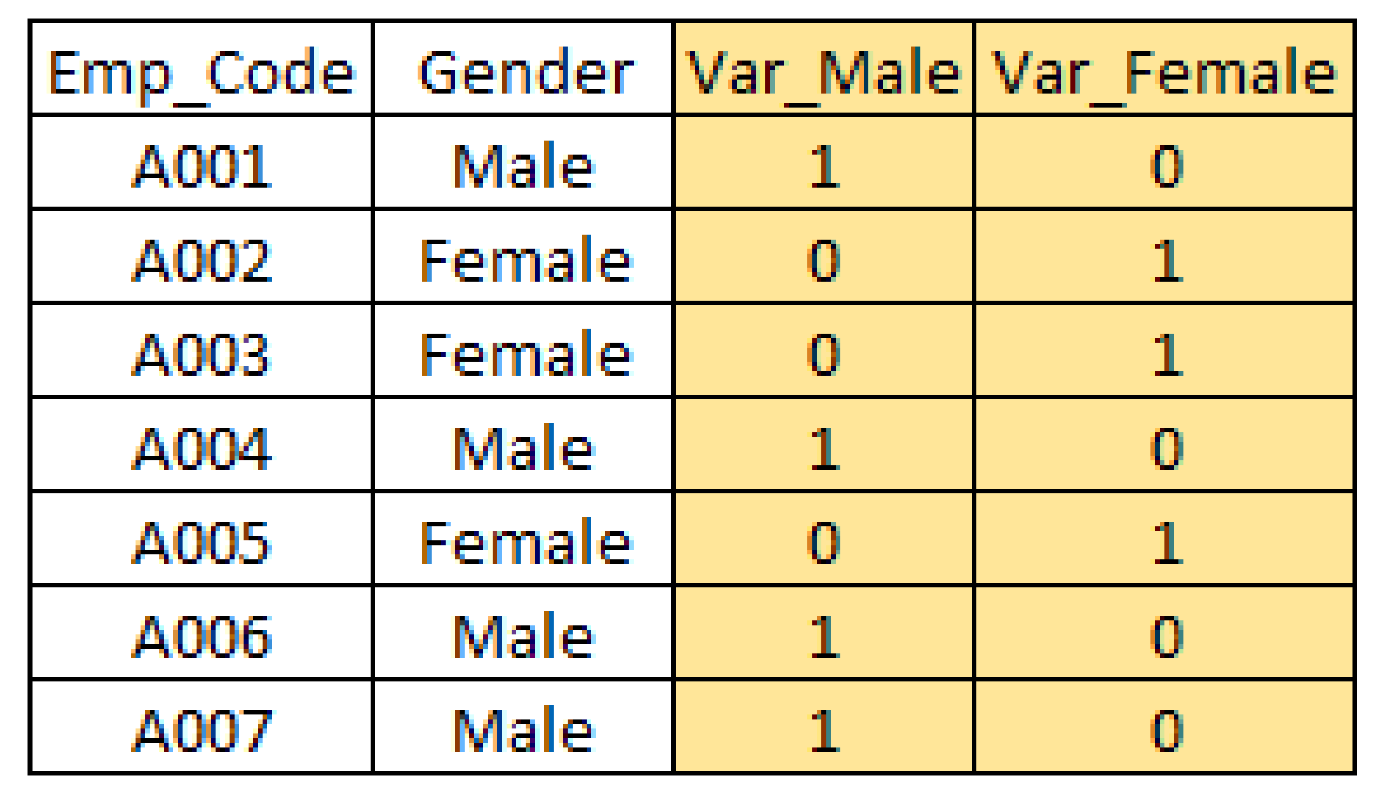
Feature / Variable creation is a process to generate a new variables / features based on existing variable(s). For example, say, we have date(dd-mm-yy) as an input variable in a data set. We can generate new variables like day, month, year, week, weekday that may have better relationship with target variable.



Emp_Code	Gender	Date	New_Day	New_Month	New_Year
A001	Male	21-Sep-11	21	9	2011
A002	Female	27-Feb-13	27	2	2013
A003	Female	14-Nov-12	14	11	2012
A004	Male	07-Apr-13	7	4	2013
A005	Female	21-Jan-11	21	1	2011
A006	Male	26-Apr-13	26	4	2013
A007	Male	15-Mar-12	15	3	2012

Common techniques to create new features

- **Creating derived variables:** This refers to creating new variables from existing variable(s) using set of functions or different methods. For example, YY/MM/DD => Year, Month, Day.
- **Creating dummy variables:** convert categorical variable into numerical variables.



Emp_Code	Gender	Var_Male	Var_Female
A001	Male	1	0
A002	Female	0	1
A003	Female	0	1
A004	Male	1	0
A005	Female	0	1
A006	Male	1	0
A007	Male	1	0

Visualization tools

Box plots are used where there is a need to summarize data on an interval scale like the ones on the stock market, highlighting the lowest, highest, median and outliers.

Heatmaps are most often used for the representation of the correlation between variables. Here is an example of a heatmap.

Histogram is the graphical representation of numerical data that splits the data into ranges. The taller the bar, the greater the number of data points falling in that range.