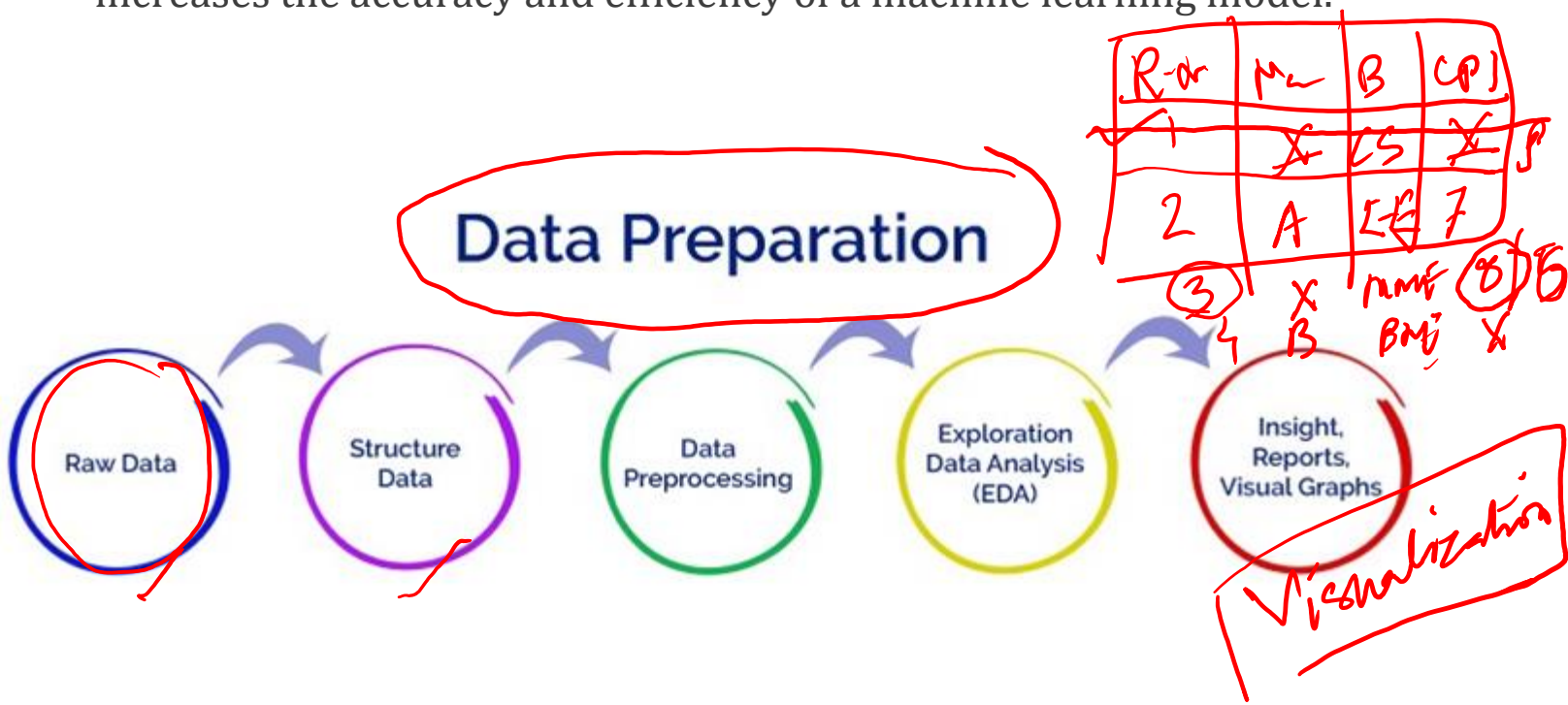
DATA PRE-PROCESSING AND VISUALIZATION

Dr. Aakanksha Sharaff
Department of Computer Science and Engineering
National Institute of Technology Raipur C.G. India

DATA PRE-PROCESSING

- Data preprocessing is a process of preparing the raw data and making it suitable for any learning model. It is the first and crucial step while creating a model.
- A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine/deep learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.



DATA PRE-PROCESSING STEPS

- **Data Cleaning**
- **Data Integration**
- **Data Transformation**
- **Data Reduction**
- **Data Sampling**

DATA CLEANING

Missing Data:
This situation arises when some data is missing in the data. It can be handled in various ways.

- **Ignore the tuples:**
  This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
- **Fill the Missing values:**
  There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

DATA CLEANING

Noisy Data:

Noisy data is a meaningless data that can't be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :

- **Binning Method:**
  - This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.
  - Binning is a way to group a number of more or less continuous values into a smaller number of "bins". For example, if you have data about a group of people, you might want to arrange their ages into a smaller number of age intervals.
- **Regression:**
  Here data can be made smooth by fitting it to a regression function. The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).
- **Clustering:**
  This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

DATA TRANSFORMATION

This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

1. **Normalization:**
   It is done in order to scale the data values in a specified range (0.0 to 1.0). e.g. Min-max, z-score

2. **Attribute Selection:**
   In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

3. **Discretization:**
   This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

4. **Concept Hierarchy Generation:**
   Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

DATA REDUCTION

Data reduction is preprocessing technique that helps in obtaining reduced representation of dataset from the available dataset that is much smaller in volume.

- Integrity of the original data should be maintained even after reduction in data volume.
- It should produce same analytics result as on original data.

Need of Data Reduction

- Visualization
- Increase efficiency of data science/mining algortihms
- Need less memory space

DATA REDUCTION

The various steps to data reduction are:

1. **Data Cube Aggregation:**

Year 2019

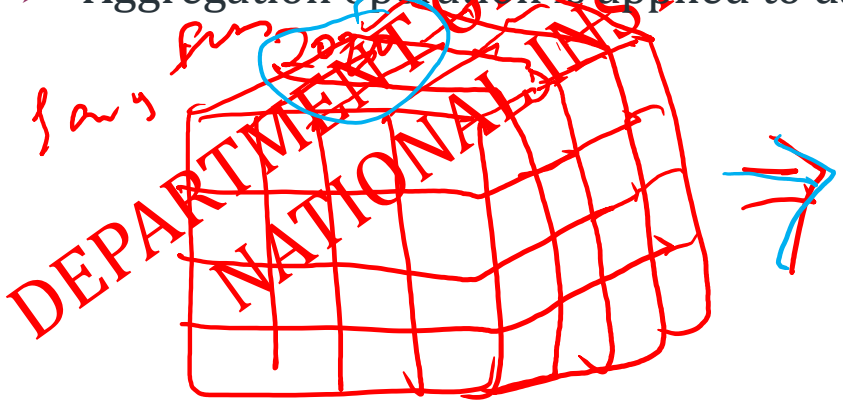
Quarter	Sales
Q1	500
Q2	300

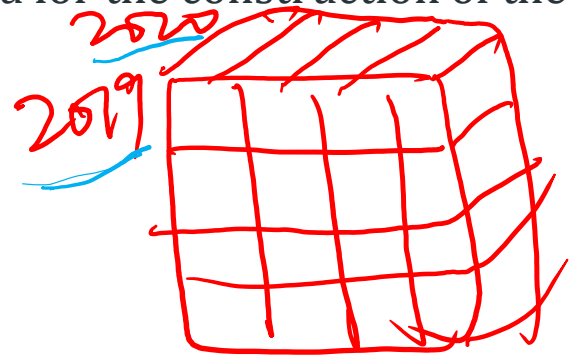
Year 2020

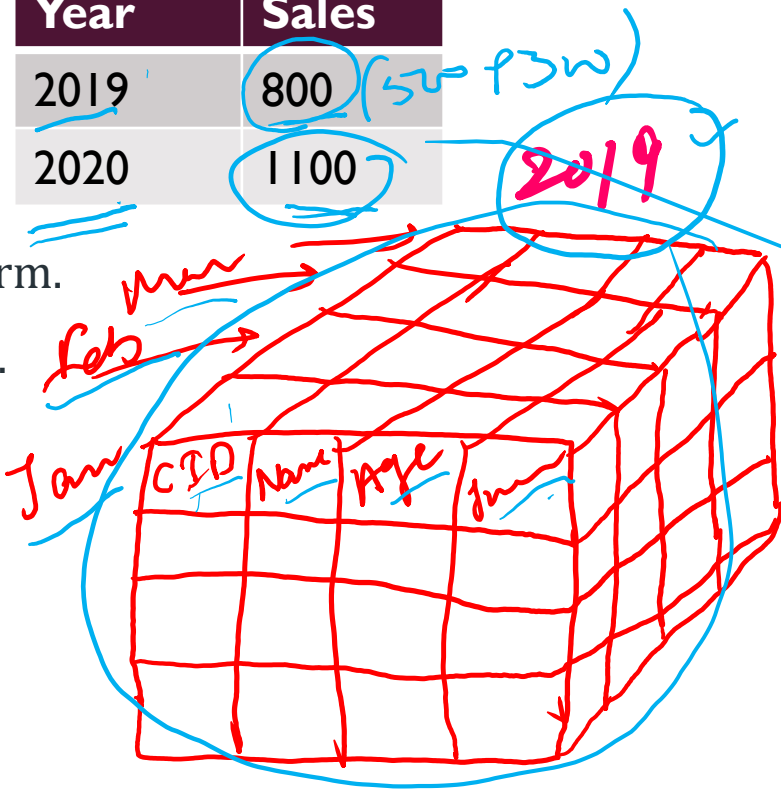
Quarter	Sales
Q1	600
Q2	500

Year	Sales
2019	800
2020	1100

- It is a process in which information is gathered and expressed in summary form.
- Aggregation operation is applied to data for the construction of the data cube.



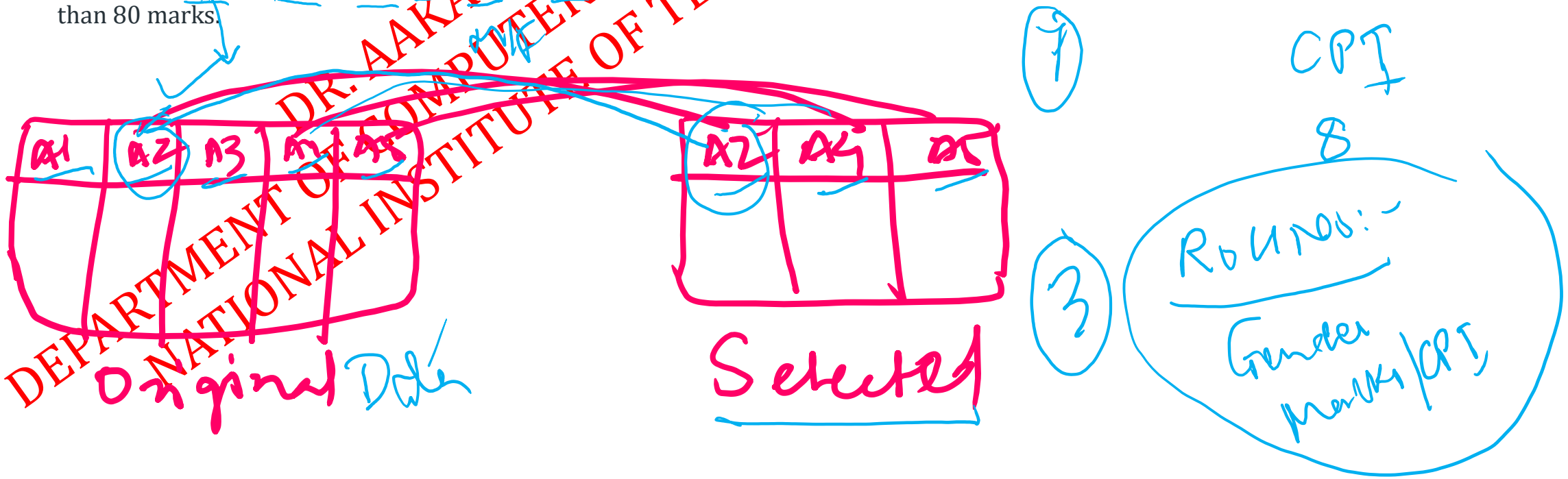




DATA REDUCTION

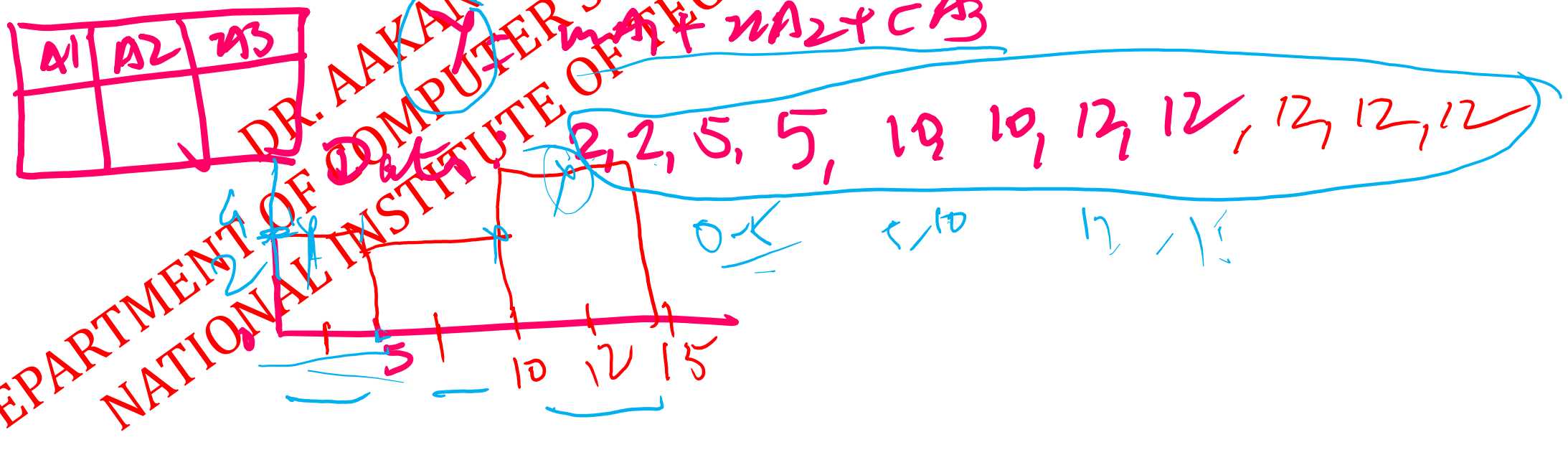
2. Attribute Subset Selection:

- Attribute subset selection reduces the dataset size by removing irrelevant or redundant features/attributes/dimensions.
- The highly relevant attributes should be used, rest all can be discarded.
- Keeping irrelevant attributes may be detrimental (non-beneficial), causing confusion for mining algorithm employed. This can result in discovering pattern of poor quality.
- Moreover, the added volume of irrelevant or redundant attribute can slow down the mining process.
- E.g. S. No., Roll No., Name, Age, DoB, Gender, Marks are attributes. The query is to find the number of males and females who secured more than 80 marks.



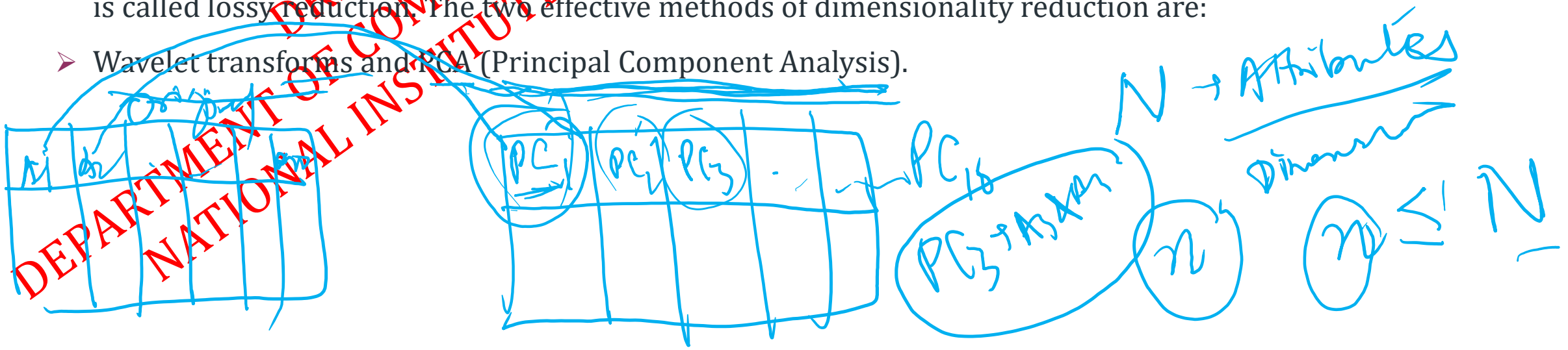
DATA REDUCTION

3. **Numerosity Reduction:**
   Data is replaced by estimated/alternative value. This enable to store the model of data instead of whole data, for example: Regression Models.



DATA REDUCTION

4. **Dimensionality Reduction:**
   - Remove redundant attributes.
   - This reduce the size of data by encoding mechanisms. It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are:
   - Wavelet transforms and PCA (Principal Component Analysis).



DATA PRE-PROCESSING

- Data Preprocessing contains the following necessary steps:
  - Getting the dataset
  - Importing libraries
  - Importing datasets
  - Finding Missing Data
  - Encoding Categorical Data
  - Feature scaling

DATA PRE-PROCESSING (CONTD.)

Get the Dataset

- To create a machine learning model, the first thing we required is a dataset as a machine learning model completely works on data. The collected data for a particular problem in a proper format is known as the dataset. To use the dataset in our code, we usually put it into a CSV file. However, sometimes, we may also need to use an HTML or xlsx file. CSV stands for "Comma-Separated Values" files. It is a file format which allows us to save the tabular data, such as spreadsheets. It is useful for huge datasets and can use these datasets in programs.

Importing Libraries

- In order to perform data preprocessing using Python, we need to import some predefined Python libraries. These libraries are used to perform some specific jobs. There are three specific libraries that we will use for data preprocessing, which are:
  - Numpy
  - Pandas
  - Matplotlib

Importing the Datasets

- Now we need to import the datasets which we have collected for our machine learning project. Now to import the dataset, we will use read_csv() function of pandas library, which is used to read a csv file and performs various operations on it. Using this function, we can read a csv file locally as well as through an URL.

DATA PRE-PROCESSING (CONTD.)

Handling Missing data

- If our dataset contains some missing data, then it may create a huge problem for our machine learning model. Hence it is necessary to handle missing values present in the dataset.
- There are mainly two ways to handle missing data, which are:
  - **By deleting the particular row:** The first way is used to commonly deal with null values. In this way, we just delete the specific row or column which consists of null values. But this way is not so efficient and removing data may lead to loss of information which will not give the accurate output.
  - **By calculating the mean:** In this way, we will calculate the mean of that column or row which contains any missing value and will put it on the place of missing value. This strategy is useful for the features which have numeric data such as age, salary, year, etc. Here, we will use this approach.

Encoding Categorical data

- Since machine learning model completely works on mathematics and numbers, but if our dataset would have a categorical variable, then it may create trouble while building the model. So it is necessary to encode these categorical variables into numbers.
- We have two methods to do so:
  - **LabelEncoder** :When only two class variables are present.
  - **OneHotEncoder** :When more than two class variables are present.

DATA PRE-PROCESSING (CONTD.)

Feature Scaling

- Feature scaling is the final step of data preprocessing in machine learning. It is a technique to standardize the independent variables of the dataset in a specific range. In feature scaling, we put our variables in the same range and in the same scale so that not a single variable dominate the other variable.
- For feature scaling, we will import ***StandardScaler*** class of ***sklearn.preprocessing.***
- By Applying above scalar, we will get all the values range between -1 to 1.

EXPLORATORY DATA ANALYSIS

- Exploratory Data Analysis 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.
- Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.
- EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today.
- Exploratory data analysis is a simple classification technique usually done by visual methods. It is an approach to analyzing data sets to summarize their main characteristics. When you are trying to build a machine learning model you need to be pretty sure whether your data is making sense or not.
- Exploratory data analysis (EDA) is a task of analyzing data using simple tools from statistics, simple plotting tools.

IMPORTANCE OF EXPLORATORY DATA ANALYSIS

- The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.
- Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals. EDA also helps stakeholders by confirming they are asking the right questions. EDA can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modeling, including machine learning.

NEED OF EXPLORATORY DATA ANALYSIS

- Every machine learning problem solving starts with EDA. It is probably one of the most important part of a machine learning project. With the growing market, the size of data is also growing. It becomes harder for companies to make decisions without analyzing it properly.
- With the use of charts and certain graphs, one can make sense out of the data and check whether there is any relationship or not.
- Various plots are used to determine any conclusions. This helps the company to make a firm and profitable decisions. Once Exploratory Data Analysis is complete and insights are drawn, its feature can be used for supervised and unsupervised machine learning modelling.

EXPLORATORY DATA ANALYSIS TOOLS

Some of the most common data science tools used to create an EDA include:

- **Python:** An interpreted, object-oriented programming language with dynamic semantics. Its high-level, built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for rapid application development, as well as for use as a scripting or glue language to connect existing components together. Python and EDA can be used together to identify missing values in a data set, which is important so you can decide how to handle missing values for machine learning.
- **R:** An open-source programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians in data science in developing statistical observations and data analysis.

EDA LEVEL OF ANALYSIS (STATISTICS: QUANTITATIVE DATA ANALYSIS)

EDA level of analysis depends on various quantitative data analysis and the analysis of number of variables/features considered for a particular case study is one of them. There are three different analysis mentioned below:

- Univariate analysis
- Bivariate analysis
- Multivariate analysis

UNIVARIATE ANALYSIS

- Univariate analysis is the most basic form of statistical data analysis technique. When the data contains only one variable and doesn't deal with a causes or effect relationships then a Univariate analysis is used.
- The key objective of univariate analysis is to simply describe the data to find the patterns within the data. The relationship or pattern within data can be found by looking into the mean, median, mode, dispersion, variance, range, standard deviation etc.

e.g. In a survey of classroom, the researcher may be looking to count the number of boys and girls. In this instance, the data would simply reflect the number, i.e. a single variable and the quantity of boys and girls.

Another example could be, if we have given height, and weight as input and output should be predicted as type (Obesity, Slim, Fit). So, by considering only weight can we predict type. The answer is Yes but it might be possible that the data may get overlapped between Obese and Slim/Slim and Fit etc. So this will work only when we have continuous data not categorical data.

STATISTICAL TECHNIQUES TO CONDUCT UNIVARIATE ANALYSIS

- Frequency Distribution Tables
- Histograms
- Frequency Polygons
- Pie Charts
- Bar Charts

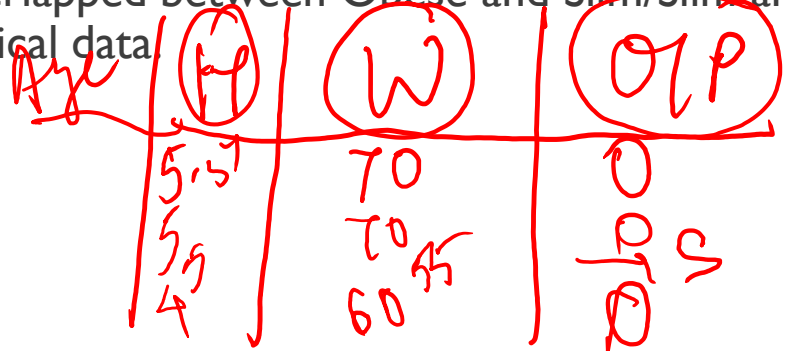
BIVARIATE ANALYSIS

- Bivariate analysis is slightly more analytical to univariate analysis.
- When the dataset contains two variables and researchers aim to undertake comparisons between the two dataset then bivariate analysis is the right type of analysis technique.
- Bivariate analysis will measure the correlations between the two variables.

e.g. In a survey of data science class, the researcher want to analyse the ratio of students who scored above 85% corresponding to their genders.

In this case there are two variables- gender X (independent variable) and result Y (dependent variable)

Another example could be, if we have given height, and weight as input and output should be predicted as type (Obesity, Slim, Fit). So, by considering both variables can we predict type more accurately. The answer is Yes but it might be possible that the data may get overlapped between Obese and Slim/Slim and Fit etc. So this will work only when we have continuous data not categorical data.



STATISTICAL TECHNIQUES TO CONDUCT BIVARIATE ANALYSIS

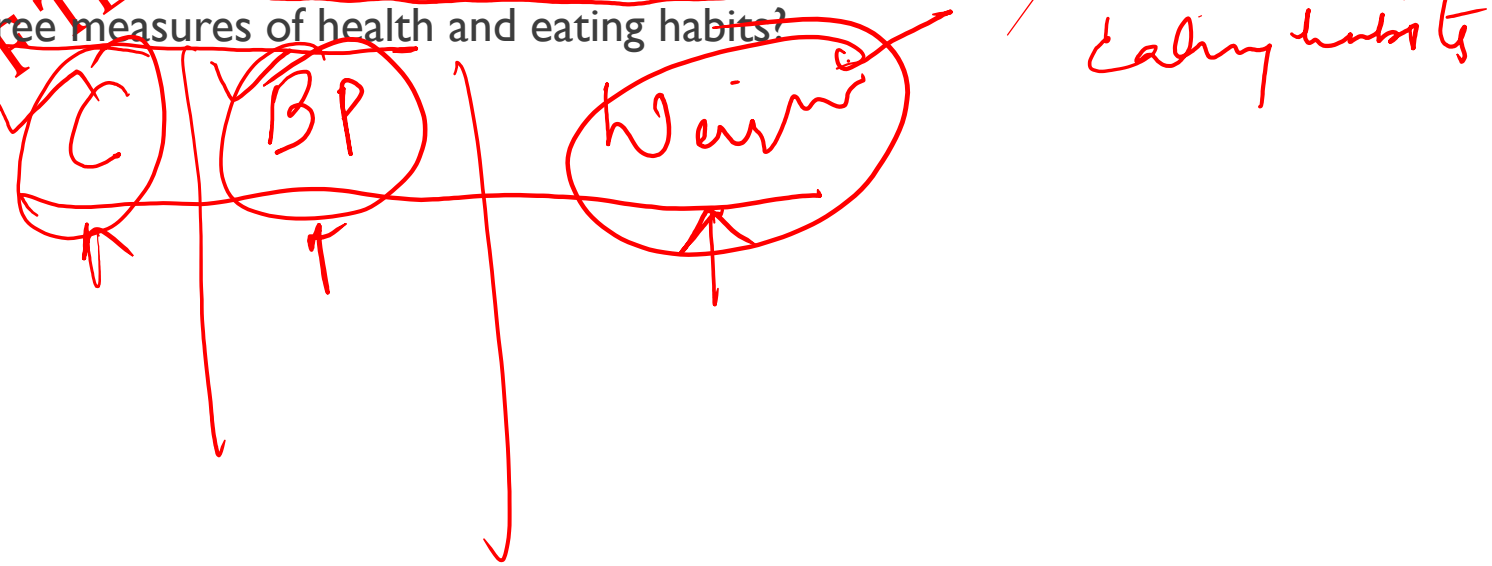
Bivariate analysis is conducted using

- Correlation coefficients
- Regression analysis (Linear regression, Logistics regression)

MULTIVARIATE ANALYSIS

- Multivariate analysis is a more complex form of statistical analysis technique and used when there are more than two variables in the dataset and tries to understand the relationship of each variable with each other.

e.g. A doctor has collected data on cholesterol, blood pressure, and weight. She also collected data on eating habits of the subjects (e.g. how many ounces of red meat, fish, dairy products, and chocolate consumed per week). She wants to investigate the relationship between the three measures of health and eating habits?



STATISTICAL TECHNIQUES TO CONDUCT MULTIVARIATE ANALYSIS

Multivariate analysis is conducted using commonly known techniques-

- Factor Analysis
- Cluster Analysis
- Variance Analysis
- Discriminant Analysis
- Multidimensional Scaling
- Principal Component Analysis
- Redundancy Analysis

EXPLORATORY DATA ANALYSIS TOOLS

Specific statistical functions and techniques you can perform with EDA tools include:

- Clustering and dimension reduction techniques, which help create graphical displays of high-dimensional data containing many variables.
- Univariate visualization of each field in the raw dataset, with summary statistics.
- Bivariate visualizations and summary statistics that allow you to assess the relationship between each variable in the dataset and the target variable you're looking at.
- Multivariate visualizations, for mapping and understanding interactions between different fields in the data.
- K-means Clustering is a clustering method in unsupervised learning where data points are assigned into K groups, i.e. the number of clusters, based on the distance from each group's centroid. The data points closest to a particular centroid will be clustered under the same category. K-means Clustering is commonly used in market segmentation, pattern recognition, and image compression.
- Predictive models, such as linear regression, use statistics and data to predict outcomes.

TYPES OF EXPLORATORY DATA ANALYSIS

There are four primary types of EDA:

- **Univariate non-graphical.** This is simplest form of data analysis, where the data being analyzed consists of just one variable. Since it's a single variable, it doesn't deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.
- **Univariate graphical.** Non-graphical methods don't provide a full picture of the data. Graphical methods are therefore required. Common types of univariate graphics include:
  - Stem-and-leaf plots, which show all data values and the shape of the distribution.
  - Histograms, a bar plot in which each bar represents the frequency (count) or proportion (count/total count) of cases for a range of values.
  - Box plots, which graphically depict the five-number summary of minimum, first quartile, median, third quartile, and maximum.
- **Multivariate nongraphical:** Multivariate data arises from more than one variable. Multivariate non-graphical EDA techniques generally show the relationship between two or more variables of the data through cross-tabulation or statistics.
- **Multivariate graphical:** Multivariate data uses graphics to display relationships between two or more sets of data. The most used graphic is a grouped bar plot or bar chart with each group representing one level of one of the variables and each bar within a group representing the levels of the other variable.

TYPES OF EXPLORATORY DATA ANALYSIS

Other common types of multivariate graphics include:

- Scatter plot, which is used to plot data points on a horizontal and a vertical axis to show how much one variable is affected by another.
- Multivariate chart, which is a graphical representation of the relationships between factors and a response.
- Run chart, which is a line graph of data plotted over time.
- Bubble chart, which is a data visualization that displays multiple circles (bubbles) in a two-dimensional plot.
- Heat map, which is a graphical representation of data where values are depicted by color.

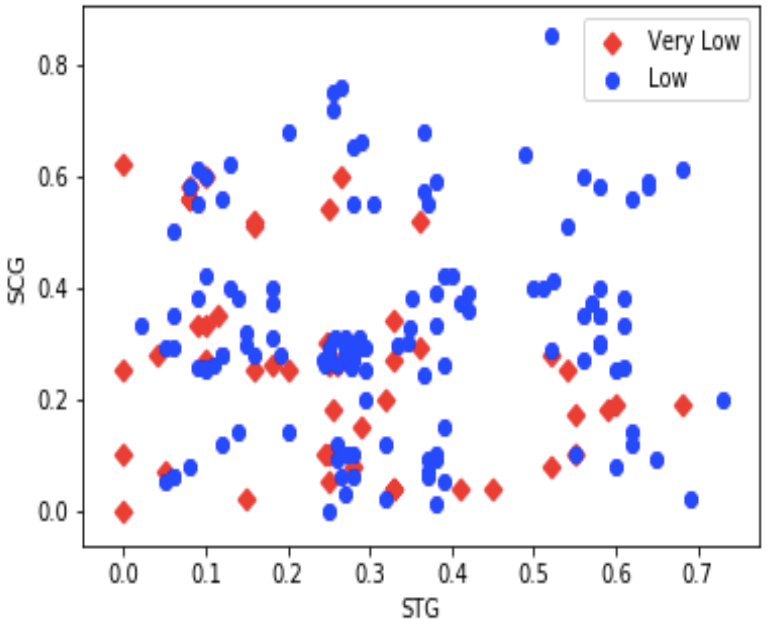
EXPLORATORY DATA ANALYSIS AND DATA VISUALIZATION (CONTD.)

How can we perform EDA

- Talking about Python, we use certain libraries like NumPy, Pandas, Matplotlib and seaborn for EDA.
- The more creative we become with data more insights we can visualize. So, while performing EDA, always ask the right question, be more creative towards data and understand the pattern thoroughly.
- **Some methods and plots are distinguished as: -**
  - **Univariate analysis**: This is the analysis of one ("uni") variable.
  - **Bivariate analysis**: This is the analysis of exactly two variables.
  - **Multivariate analysis**: This is the analysis of more than two variables.
- **Here are the common graphs used while performing EDA:-**
  - Scatter Plot
  - Pair plots
  - Histogram
  - Box plots
  - Violin Plots

EXPLORATORY DATA ANALYSIS AND DATA VISUALIZATION (CONTD.)

- **Scatter plot**: — It is a type of plot which will be in a scatter format. It is typically between 2 features. This is used to check if there is any linearity between these two features.
- Here two colours (Orange & Blue) represents two features.

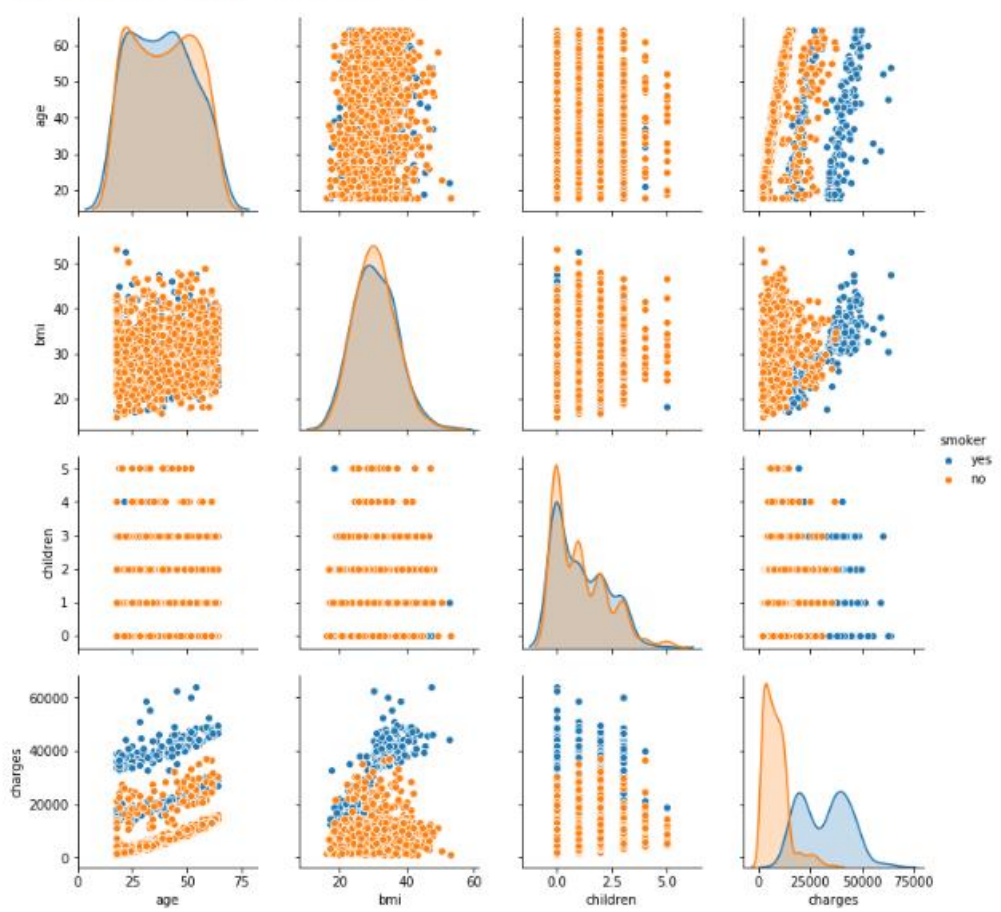


EXPLORATORY DATA ANALYSIS AND DATA VISUALIZATION (CONTD.)

- **Pair plots:-** This is used to see the behavior of all the features present in the dataset. Also we get to see the PDF representation.

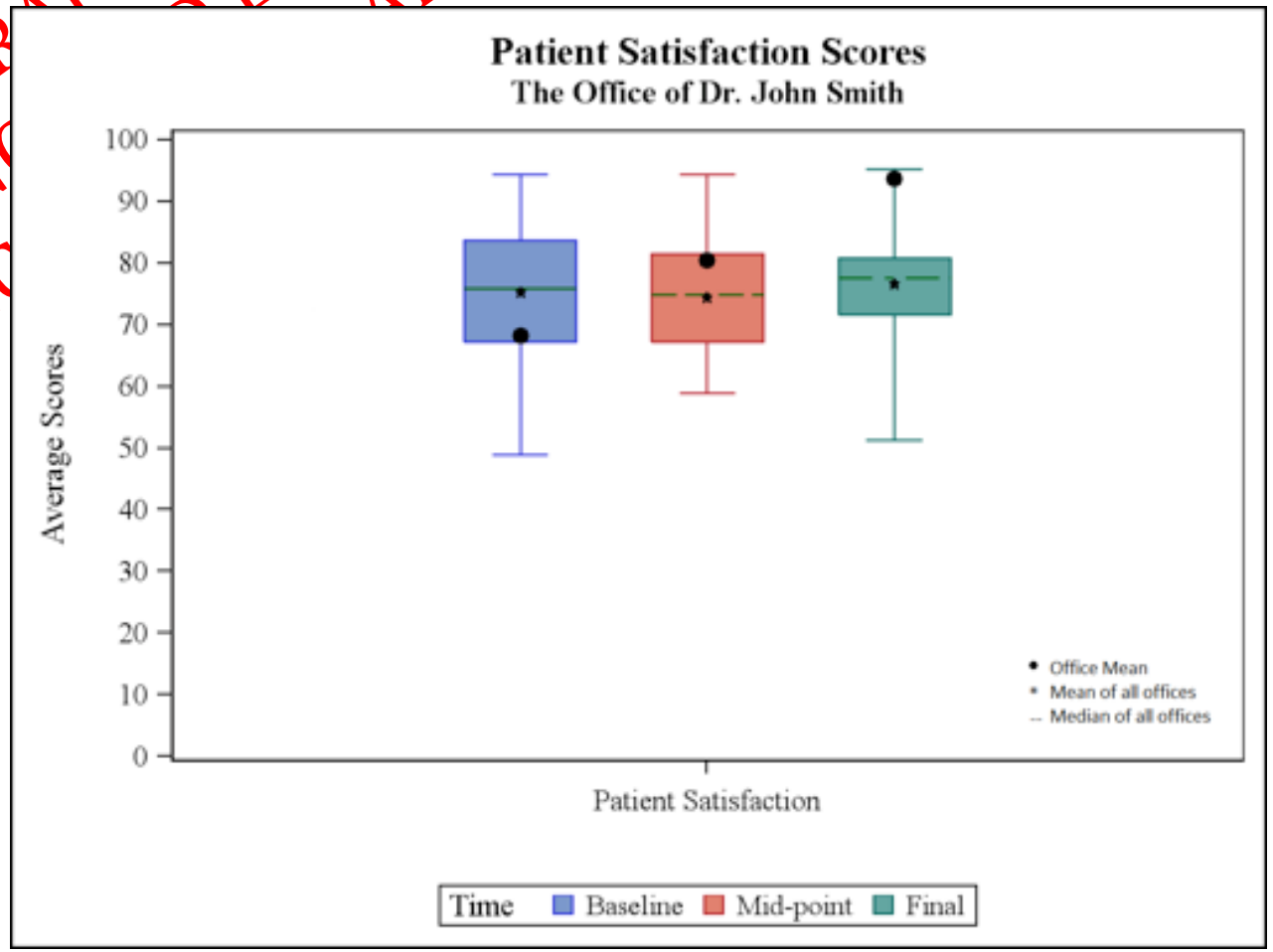

```
sns.pairplot(df, hue="smoker")
```

```
<seaborn.axisgrid.PairGrid at 0x1afb4b02128>
```



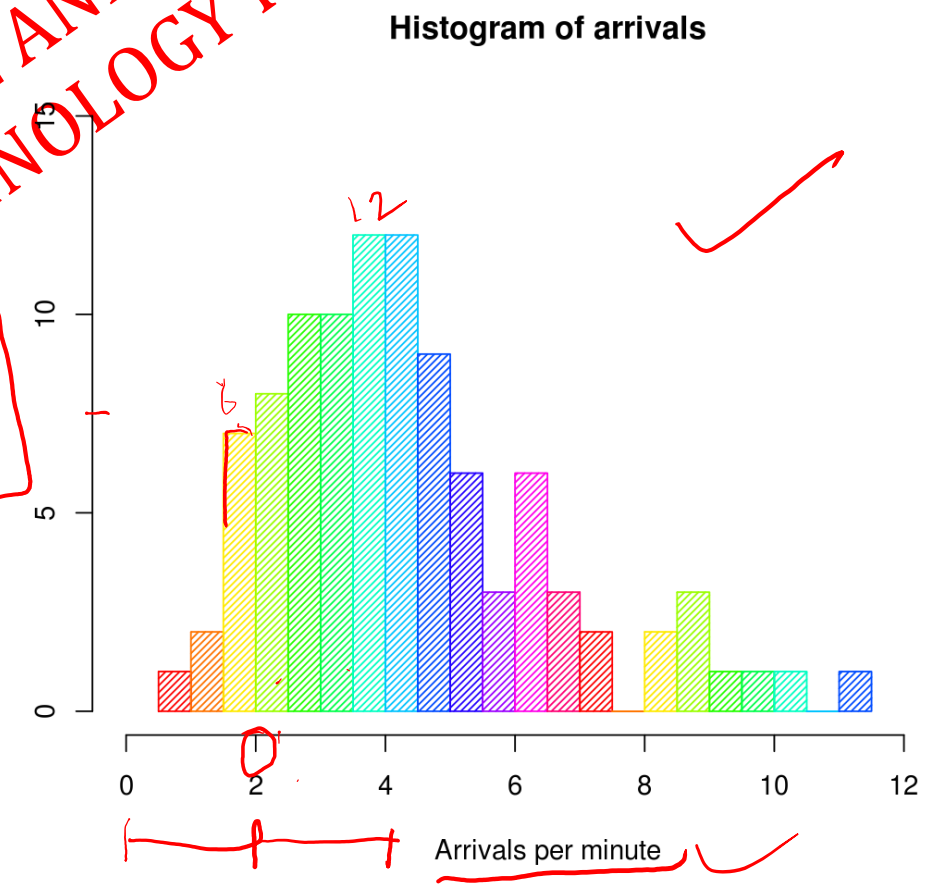
EXPLORATORY DATA ANALYSIS AND DATA VISUALIZATION (CONTD.)

- **Box-Plots:** Box plots tell us the percentile plotting which other plots cant tell easily. It also helps in detection of outliers.
- Here we can know about the quartile range and the outliers situation.



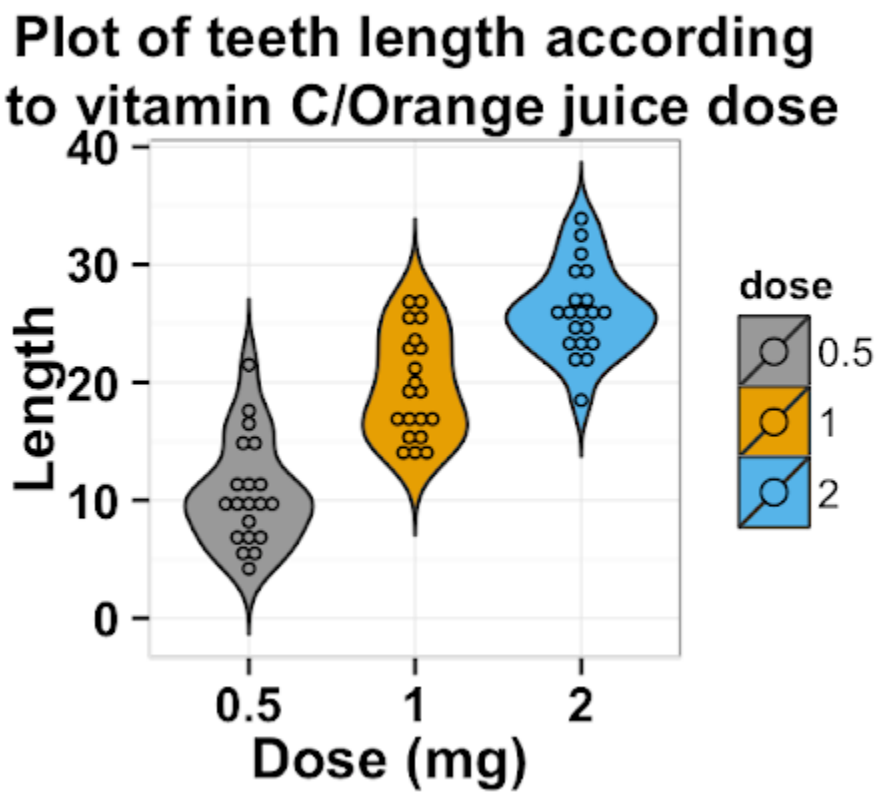
EXPLORATORY DATA ANALYSIS AND DATA VISUALIZATION (CONTD.)

- **Histogram:** Histogram plots are used to depict the distribution of any continuous variable. These types of plots are very popular in statistical analysis.



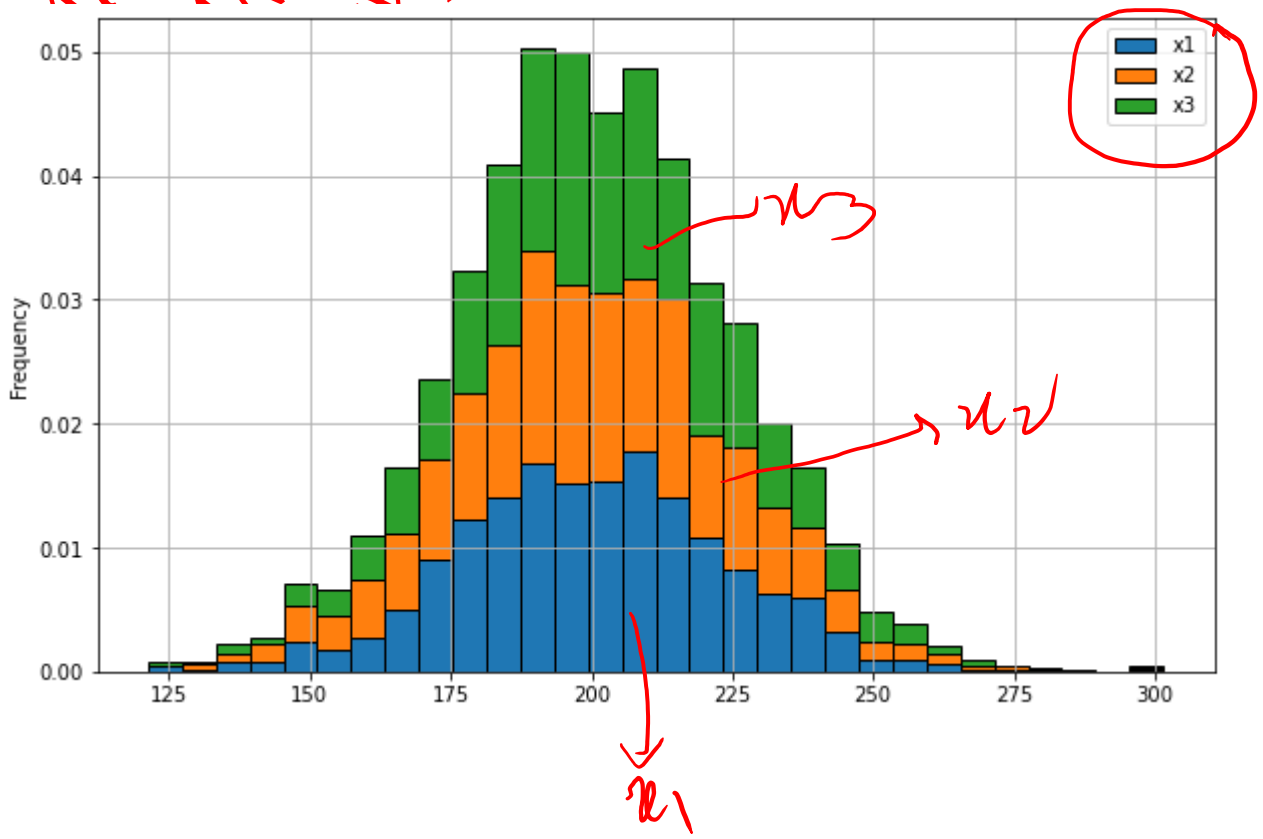
EXPLORATORY DATA ANALYSIS AND DATA VISUALIZATION (CONTD.)

- **Violin plots-** It is a extension of box plots in this the kernel density plot is also plotted with box plots.



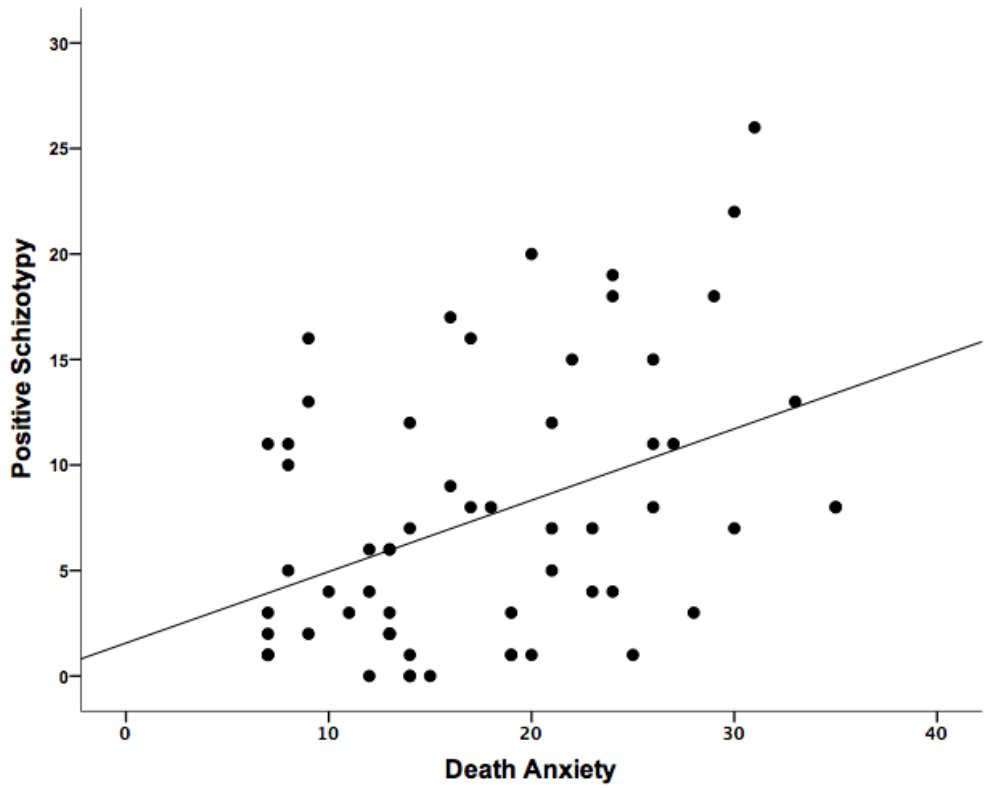
STACKED HISTOGRAMS FOR MULTIVARIATE DATA

- Stacked Histogram is a type of graph or graphical representation of different items on a single bar with different colour where each colour represents different item.
- This histogram is normally used when we need to represent multiple variables. It will be easy for multiple categorical labels.



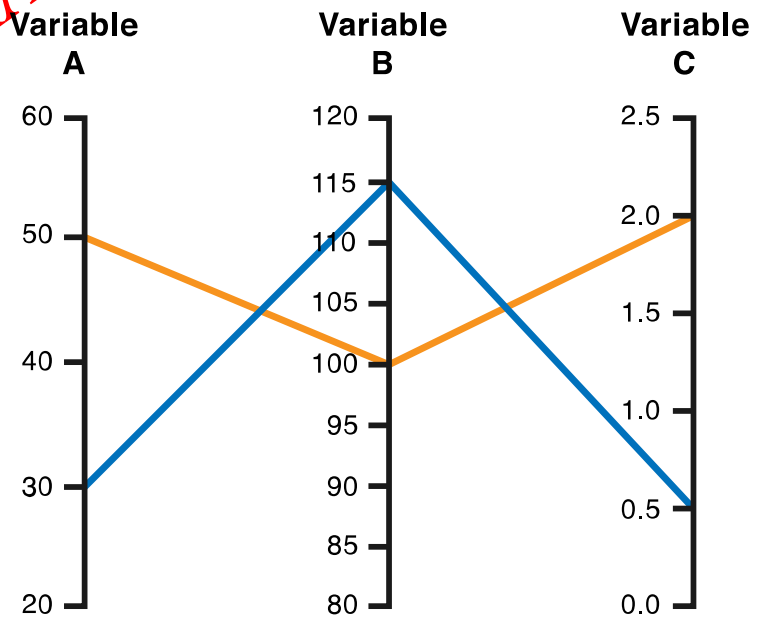
BIVARIATE SCATTER PLOT

- A bivariate plot graphs the relationship between two variables that have been measured on a single sample of subjects. Such a plot permits you to see at a glance the degree and pattern of relation between the two variables.
- On a bivariate plot, the abscissa (X-axis) represents the potential scores of the predictor variable and the ordinate (Y-axis) represents the potential scores of the predicted or outcome variable.
- This is what we mean by "bivariate" plot. Each point represents two variables.



PARALLEL COORDINATE PLOT

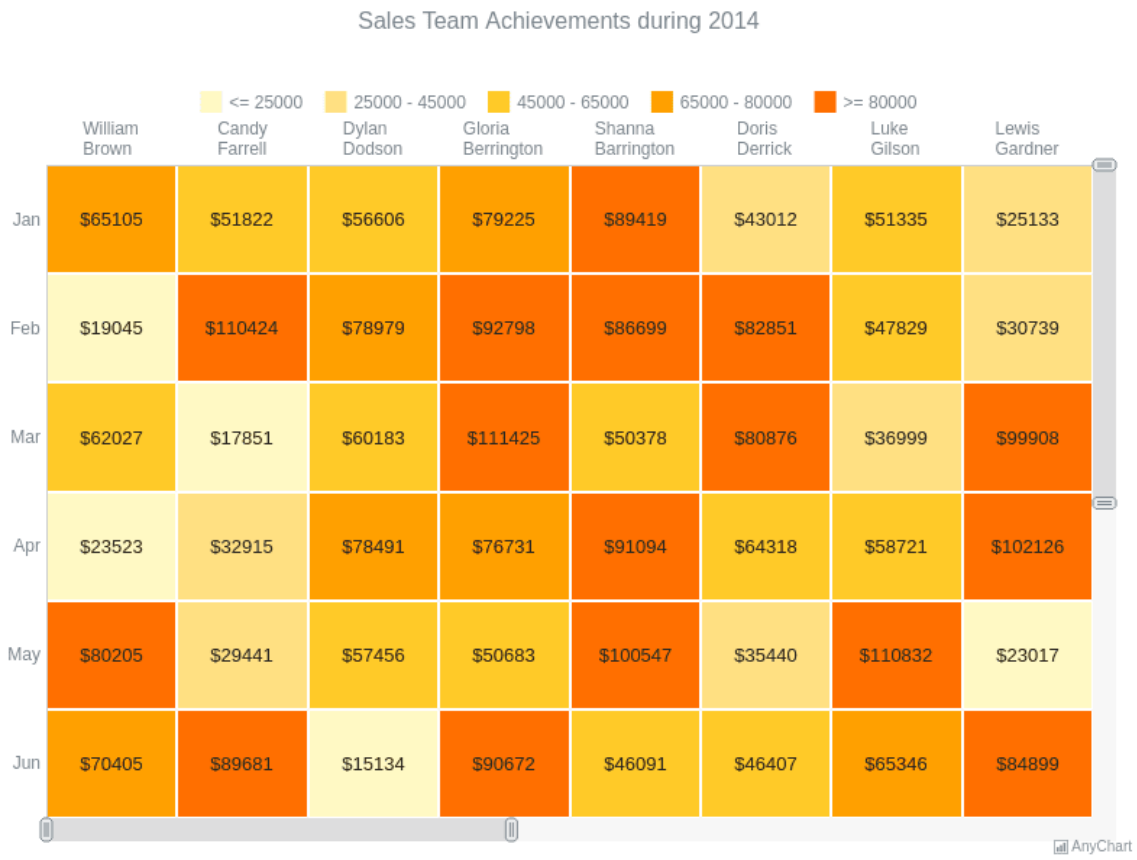
- This type of visualisation is used for plotting multivariate, numerical data.
- Parallel Coordinates Plots are ideal for comparing many variables together and seeing the relationships between them.
- In a Parallel Coordinates Plot, each variable is given its own axis and all the axes are placed in parallel to each other. Each axis can have a different scale, as each variable works off a different unit of measurement, or all the axes can be normalized to keep all the scales uniform. Values are plotted as a series of lines that connected across all the axes. This means that each line is a collection of points placed on each axis, that have all been connected together.
- EX- Comparing computer or cars specs across different models.



Data			
	Variable A	Variable B	Variable C
Item 1	50	100	2.0
Item 2	30	115	0.5

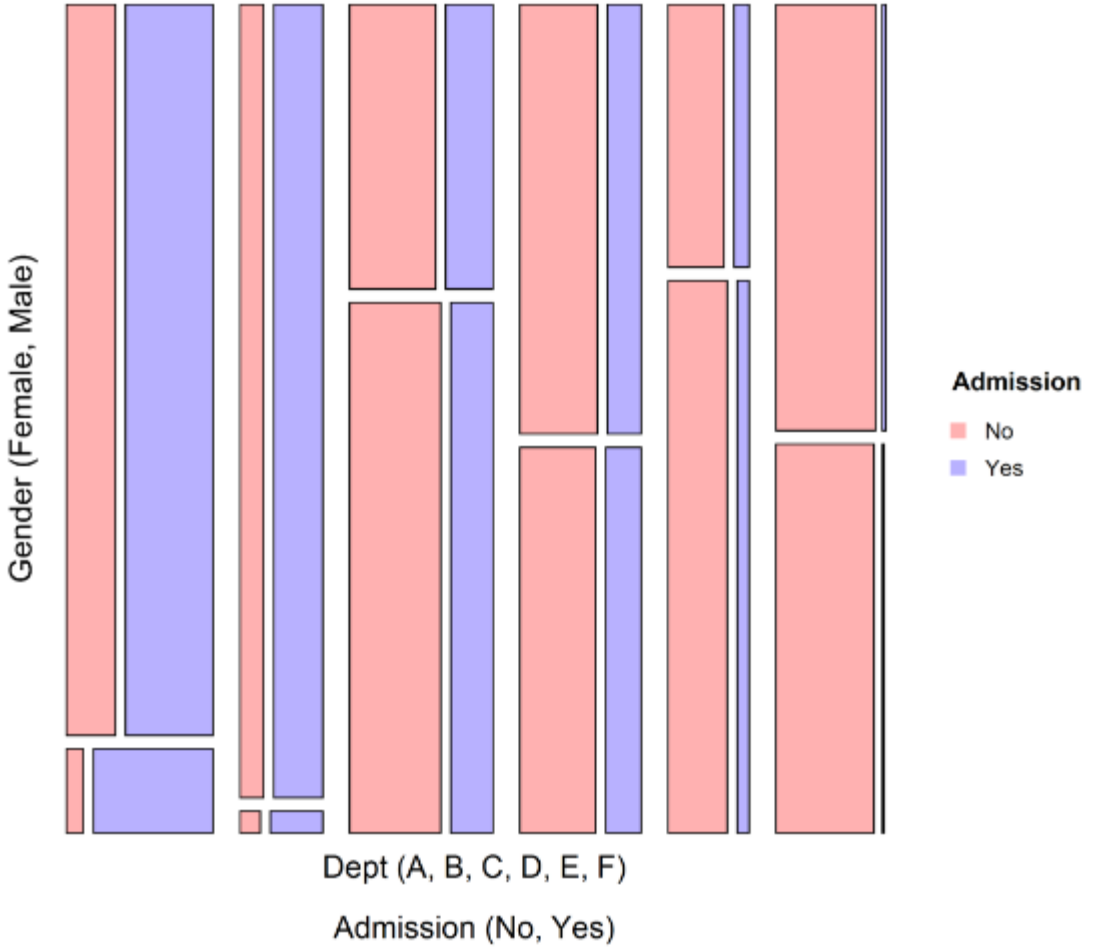
HEAT MAP (TABLE PLOT)

- Heat maps are a great tool for visualizing complex statistical data.
- Doctors, engineers, marketers, sociologists, and researchers of every kind use heat maps to make data sets comprehensible and actionable.
- A heat map is data analysis software that uses color the way a bar graph uses height and width as a data visualization tool.
- A heat map uses a warm-to-cool colour spectrum to show you which parts of a page receive the most attention.
- This also help you answer a crucial question: "Where should the most important content be on this dataset?"



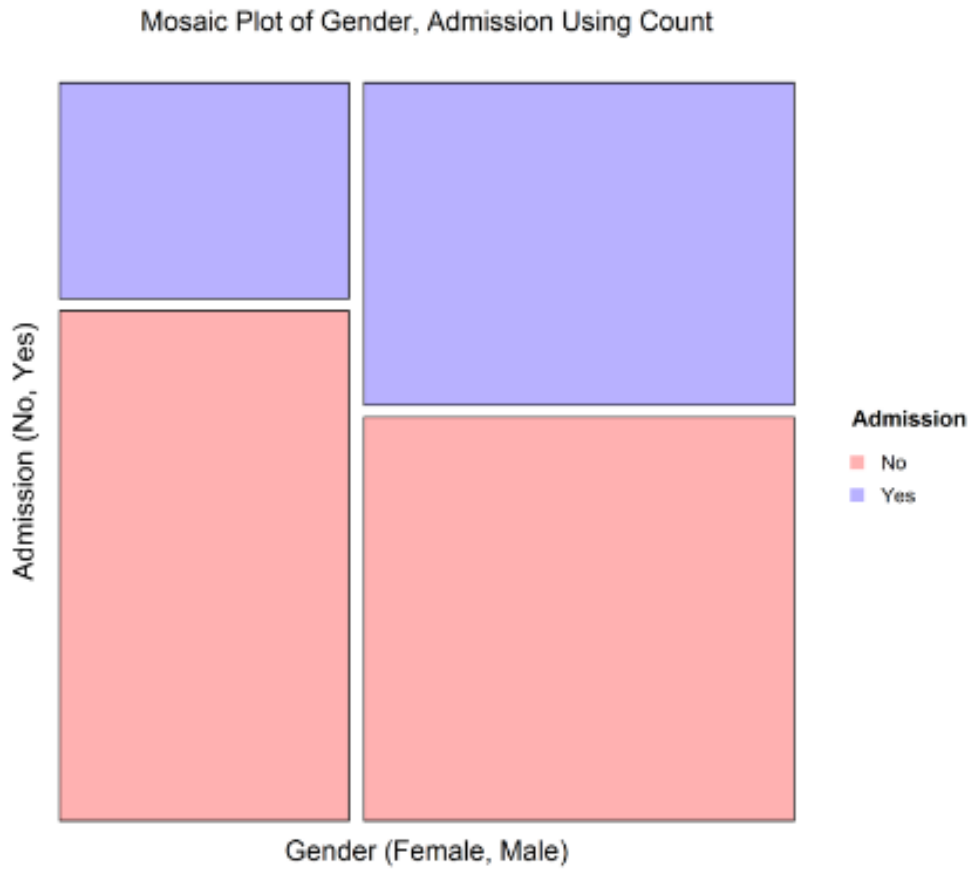
MOSAIC PLOTS

- A mosaic plot is a graphical display of the cell frequencies of a contingency table in which the area of boxes of the plot are proportional to the cell frequencies of the contingency table. This procedure can construct mosaic plots for up to four-way contingency tables.
- Mosaic plot is based on conditional probabilities.



MOSAIC PLOTS (CONTD.)

- For example the admission details of female and male candidates has taken in to consideration.
- The widths of the boxes are proportional to the percentage of females and males, respectively. In fact, 41% of applicants are female and 59% are male.
- The heights of the boxes are proportional to percent admitted. In fact, 45% of the male applicants are admitted, while only 30% of the female applicants are admitted. This seems to show a large gender-bias in admission.
- To make the plot easier to interpret, the boxes for admitted females and males are colored blue while the not-admitted females and males are colored pink.
- It is easy to see that females' blue box on the left is much shorter than the males' blue box on the right.



REFERENCES

- https://www.javatpoint.com/data-preprocessing-machine-learning
- https://medium.com/analytics-vidhya/data-visualization-and-exploratory-data-analysis-eda-in-data-science-984e84942fda
- https://datavizcatalogue.com/methods/parallel_coordinates.html
- https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Mosaic_Plots.pdf