# PROJECT CHARTER PURPOSE

**Description:**

An Out-patient department is a part and parcel of any hospital. People with health problems who visit the hospital for diagnosis and treatment but do not stay overnight are called Out-patients.

Health care providers are under a great deal of pressure to reduce costs and improve quality of service provided. Outpatient services are gradually becoming an essential component in health care. A good OPD service can reduce the load on In-patient services. An Out-patient clinic is a place for rendering preventive and promotive health activities. An Out-patient department performs the function of diagnosis, curative, preventive, rehabilitative, care. Besides, an OPD must screen the patients for admission to hospital, and follow-up care after discharge. It is a place for training of medical and nursing students. An OPD clinic must maintain UpToDate records for future treatments, medical education, epidemiological and social research. Many Out-patient service centres specialize in specific area of medicine such as Orthopaedics, Cardiac Sciences, Neurosciences, Endocrinology, Lifestyle diseases etc. which are managed by medical personnel, with specialized knowledge and expertise. With the advent of advancement of technology, the average length of stay in hospitals has reduced. These results in the hospitals are more prone to admit acute patients for In-patient care. As a result, the pressure on the Out-patient department has increased.

# 1.Purpose:

The project charter defines the scope, objectives, and overall approach for the work to be completed. It is a critical element for initiating, planning, executing, controlling, and assessing the project. It should be the single point of reference on the project for project goals and objectives, scope, organization, estimates, work plan, and budget. In addition, it serves as a contract between the Project Team and the Project Sponsors, stating what will be delivered according to the budget, time constraints, risks, resources, and standards agreed upon for the project.

**2.PROJECT EXECUTIVE SUMMARY**

● Project goals:

* perform descriptive analysis on the hospital outpatient count and analyze the data pertaining to Out-patients visiting different hospitals with respect to year, age, specialty and time-bound and perform EDA of every category individually.

● Objectives:

* Collecting the data.
* Exploratory data Analysis.
* Visualizations
* Data Cleansing and Data preprocessing.
* Feature engineering

● Approach

1. Data Sourcing
2. Data Cleaning
3. EDA

**3.PROJECT OVERVIEW**

**Use case:**

To analyse the Out-patient count with respect to age, adult and children, adult and children per hospital, time-bands, time-bands per hospital, hospital-wise, specialty-wise, specialty-time band, specialty count by descriptive and predictive analysis Using Python and R.

**1. Data Sourcing**

Data Sourcing is the process of finding and loading the data into our system. Broadly there are two ways in which we can find data.

**2. Data Cleaning**

After completing the Data Sourcing, the next step in the process of EDA is Data Cleaning. It is especially important to get rid of the irregularities and clean the data after sourcing it into our system.

**3. Exploratory Data Analysis (EDA):**

What is Exploratory Data Analysis (EDA)?

If we want to explain EDA in simple terms, it means trying to understand the given data much better, so that we can make some sense out of it.

In statistics, exploratory data analysis is an approach to analysing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task.

EDA in Python uses data visualization to draw meaningful patterns and insights. It also involves the preparation of data sets for analysis by removing irregularities in the data.

**Steps in Exploratory Data Analysis**



**Variable Identification**:  In this step, we identify every variable by discovering its type. According to our needs, we can change the datatype of any variable.

**Univariate Analysis**: In Univariate Analysis, we study individual characteristics of every feature/variable available in the dataset.

**Bivariate Analysis**: In Bivariate Analysis, we study the relationship between any two variables which can be categorical-continuous, categorical-categorical, or continuous-continuous.

**Multivariate Analysis**: In MultivariateAnalysis, we study the relationship between multiple variables.

**Missing Values:**

If there are missing values in the Dataset before doing any statistical analysis, we need to handle those missing values.

There are mainly three types of missing values.

MCAR (Missing completely at random): These values do not depend on any other features.

MAR (Missing at random): These values may be dependent on some other features.

MNAR (Missing not at random): These missing values have some reason for why they are missing.

**Outliers:**

Outliers are the values that are far beyond the next nearest data points.

There are two types of outliers:

1. Univariate outliers: Univariate outliers are the data points whose values lie beyond the range of expected values based on one variable.

2. Multivariate outliers: While plotting data, some values of one variable may not lie beyond the expected range, but when you plot the data with some other variable, these values may lie far from the expected value.

Three quantitative methods commonly used in statistics for the detection of univariate outliers:

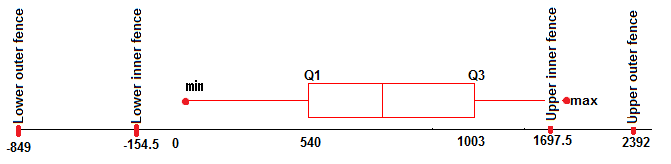
# Tukey’s box plot method

# Internally studentized residuals (AKA z-score method)

# Median Absolute Deviation method

# ****Tukey’s box plot method****

Next to its visual benefits, the box plot provides useful statistics to identify individual observations as outliers. Tukey distinguishes between **possible** and **probable outliers**. A possible outlier is located between the **inner** and the **outer fence**, whereas a probable outlier is located outside the outer fence.



Example of a box plot including the inner and outer fences and minimum and maximum observations (known as whiskers).

While the inner (often confused with the whiskers) and outer fence are usually not shown on the actual box plot, they can be calculated using the **interquartile range**(IQR) like this:

IQR =Q3 - Q1, whereas q3 := 75th quartile and q1 := 25th quartile

Inner fence = [Q1-1.5\*IQR, Q3+1.5\*IQR]

Outer fence = [Q1–3\*IQR, Q3+3\*IQR]

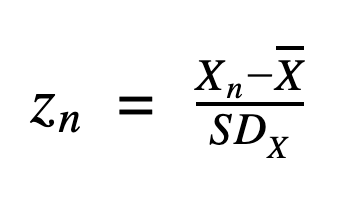
The distribution’s inner fence is defined as 1.5 x IQR below Q1, and 1.5 x IQR above Q3. The outer fence is defined as 3 x IQR below Q1, and 3 x IQR above Q3. Following Tukey, only the probable outliers are treated, which lie outside the outer fence. For the underlying example, this means:

You can easily find the outliers of all other variables in the data set by calling the function **tukeys\_method** for each variable.

The great advantage of Tukey’s box plot method is that the statistics (e.g. IQR, inner and outer fence) are robust to outliers, meaning to find one outlier is independent of all other outliers. Also, the statistics are easy to calculate. Furthermore, this method does not require a normal distribution of the data, which is often not guaranteed in real-life settings. If a distribution is highly skewed (usually found in real-life data), the Tukey method can be extended to the **log-IQ method**. Here, each value is transformed to its logarithm before calculating the inner and outer fences.

## ****Internally studentized residuals AKA z-score method****

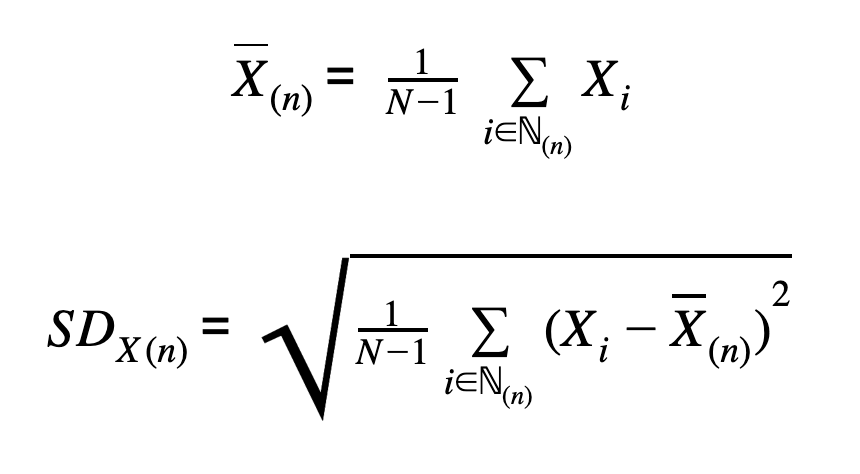
Another commonly used method to detect univariate outliers is the **internally standardized residuals**, aka the **z-score method**. For each observation (Xn), it is measured how many standard deviations the data point is away from its mean (X̄).



Following a common rule of thumb, if z > C, where C is usually set to 3, the observation is marked as an outlier. This rule stems from the fact that if a variable is normally distributed, 99.7% of all data points are located 3 standard deviations around the mean.

When using the z-score method, 8 observations are marked as outliers. However, this method is highly limited as the distributions mean and standard deviation are sensitive to outliers. This means that finding one outlier is dependent on other outliers as every observation directly affects the mean.

Moreover, the z-score method assumes the variable of interest to be normally distributed. A more robust method that can be used instead is the externally studentized residuals. Here, the influence of the examined data point is removed from the calculation of the mean and standard deviation, like so:



Nevertheless, the externally studentized residuals have limitations as the mean and standard deviations are still sensitive to other outliers and still expect the variable of interest X to be normally distributed.

## ****Median Absolute Deviation method****

The **median absolute deviation method** (MAD) replaces the mean and standard deviation with more robust statistics, like the median and median absolute deviation. The median absolute deviation is defined as:



The test statistic is calculated like the z-score using robust statistics. Also, to identify outlying observations, the same cut-off point of 3 is used. If the test statistic lies above 3, it is marked as an outlier. Compared to the internally (z-score) and externally studentized residuals, this method is more robust to outliers and does assume X to be parametrically distributed (Examples of discrete and continuous parametric distributions).