Play Store App Review Analysis

Nikhil Roeewal, Zayeem Ahmed,
Ajit Padole, Manoj Sridala, Sumanta Banerjee

ABSTRACT

The Google play store is one of the largest and most popular Android app stores. It has an enormous amount of data that can be used to make an optimal model. We have used Play Store data from the team capstone project dashboard. This data set contains 13 different features that can be used for predicting key factors responsible for app engagement & success stories.

INTRODUCTION

Mobile applications are absolutely vital because specific software is required for almost every purpose be it personal, social, business & for any such functions. It is one of the fastest-growing segments of downloadable software application markets It has become more important as the android market has gone to a real amelioration among mankind over the last decade. One of the main reasons for this popularity is the fact that about 81% of the apps are free of cost. The market has increased to over 3.5 million Apps and around 3000+ apps are being added per day as per a Google survey report. Thus, the market, in turn, led to around 5 billion users downloading all over the world. Developers and users play key roles in determining the impact that market interactions have on future technology. However, the lack of a clear understanding of the inner working and dynamics of popular app markets impacts both the developers and users. This journal talks about the dynamics of the Play Store app & analyzes the actionable insights for the developers to work on and capture the Android market and also analyzes factors for app engagement and success with classifier models used for finding the user engagement, success parameters, and the complete Data visualization.

INTEGRAL METHODOLOGY

The entire Analysis is divided into the following phases: Dataset Description, Breakdown of Datasets, Examining the null values & missing values, Data Cleaning, followed by Exploratory Data Analysis by and applying different models. First, we collect the data from Alma's better dashboard. Thereafter we did basic data cleaning and data visualization. After visualizing the data set, we removed some unnecessary features and made it ready for analyzing the data set using different plots. Next, we conduct data modelling by using Bar plot graphs, violin plots, density plots, etc. Finally, we narrate the analysis results to provide a clear vision of the relationship among the areas of interest

DATASET DESCRIPTION

Let's take a look at the data, which consists of tw o files:
playstore.csv: contains all the details of the appli cations on Google Play. There are 13 features th at describe a given app.
user_reviews.csv: contains 100 reviews for each app, most helpful first. The text in each review h as been preprocessed and attributed with three n ew features: Sentiment (Positive, Negative or Neutral), Sentiment Polarity and Sentiment Subjectivity,

About Dataset Most regularly a dataset relates to the matter of the single database table, or the single factual information framework, where each segment of the table speaks to a specific variable, and each column compares to a given individual from the informational collection being referred to. This dataset has 13 columns of varied categories of the appliance. In this project, I have analyzed all these various columns of the dataset.

App	Name of the App
Category	Category under which it falls
Rating	Application’s rating on playstore
Reviews	No. of reviews of the app
Size	Size of the app
Installs	No. of Installs of the app
Type	If the App is free/ paid
Price	Price of the App (o if it is free)
Content Rating	Appropriate target audience of the App
Genres	Genres under which the App falls
Last Updated	Date when the App was last updated
Current Version	Current version of the App
Android Version	Minimum Android version required to run the App.

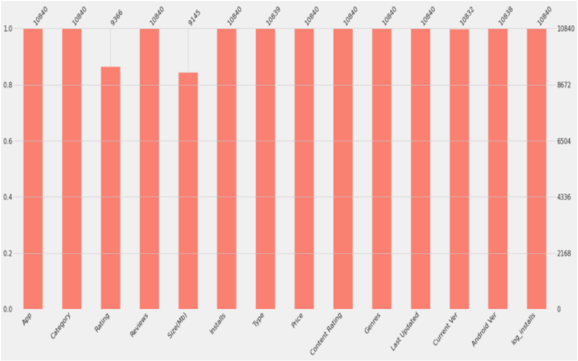
BREAKDOWN OF DATASETS

In order to go ahead for data visualization upon key factors we need to go for certain extra steps before proceeding to the main segment. In this part we are going with the following steps:

1. Importing Analytical necessary library classes for future analysis.
2. Reading the csv data file from Google drive.
3. Setting figure size for future visualization.
4. Removing future warnings in seaborne plots.
5. Visualizing all the columns of the respective Data frame.
6. Viewing all data information
7. Checking the Unique values in the column (if any)
8. Converting the data types to similar objects as the Analysis Demands.
9. Formatting the “size” column into a single column in the dataset.
10. Eradicating special characters from the dataset columns.

EXAMINING NULL VALUES

The most critical thing from which we can draw some observations is Dataset, however data comes with unexpected values too i.e. sometimes it may be Null or missing in other words the space might be blank. Thus, at the time of analysing the first thing which we will do is to examine the null or missing values on the Dataset. It is the first step that will make the results “more” accurate &should be handled before it affects the performance of the models that predict the outcome.Byplotting a graph it can be seen that missing values are more in Size & Rating columns. Hence, several methods to eradicate those null values.



DATA CLEANING

Data cleaning is one of the most essential subtask of any data science project. Although it can be a very tedious process, it's worth should never be undermined.

By looking at a random sample of the dataset rows (from the above task), we observe that some entries in the columns like Installs, Price and Size have a few special characters (+ , $,M , k) .This prevents the columns from being purely numeric, making it difficult to use them in subsequent future mathematical calculations. Ideally, as their names suggest, we would want these columns to contain only digits from [0-9].

Hence, we now proceed to clean our data. Specifically, the special characters "," and "+" present in Installs column and "$" present in Price column need to be removed.

By finding all unique values of each row the inappropriate values can be identified. Different methods can then be used for removing them or to change those values accordingly to use them to make predictions better.

As the proverb goes by saying “More Data beats clever algorithm, but better data beats more Data” – Peter Norvig.So going with the method firstly we have found the categorical null values and replacing them by a textual string, secondly finding out the numerical ‘Nan’ values & replacing them with the median of that respective column .After we had check the entire datasets for any null values (if, any exists after eradicating).

Now, cleaning all the null values we would drop certain labels/columns which is unnecessary for actionable insights. Therefore, we are can proceed for the Exploratory Data Analysis and observations regarding the datasets.

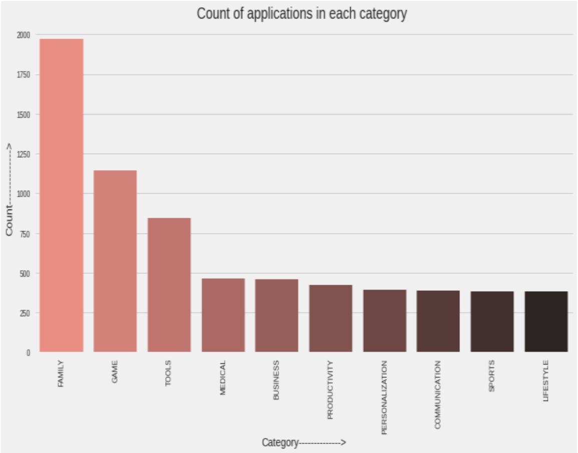
DATA VISUALIZATIONS

With more than 1 billion active users in 190 countries around the world, Google Play continues to be an important distribution platform to build a global audience. For businesses to get their apps in front of users, it's important to make them more quickly and easily discoverable on Google Play. To improve the overall search experience, Google has introduced the concept of grouping apps into categories

In this step, we will perform some initial analysis and visualizations. In order to understand which category has the most number of application installations from the dataset, we have made a bar plot to visualize it.

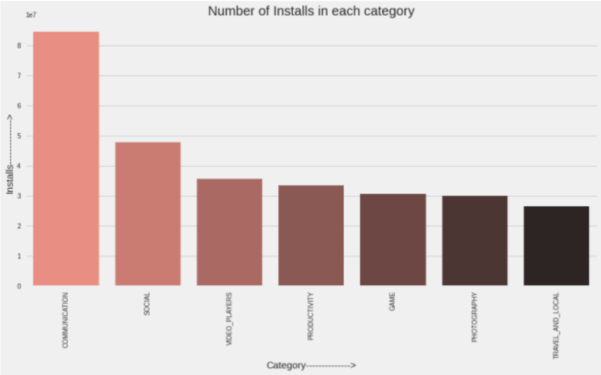
Observation-1

so we take the Category column and check the maximum count of apps so that we can our top 10 apps.



Now we know that the 'Family' and 'Game' category rules the play store market, followed by Tools, Medical, and Business.

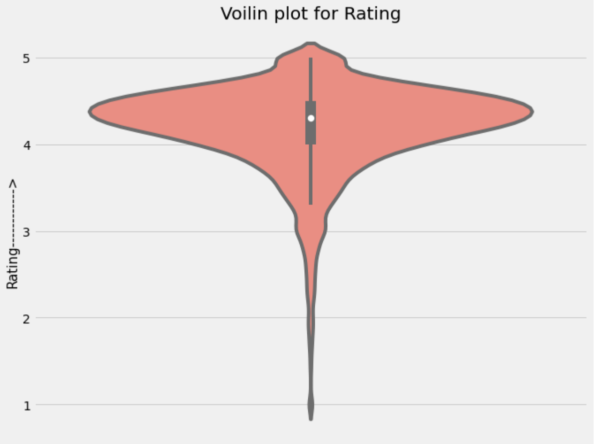
let's compare "installs" with "category" which is one of the other characters



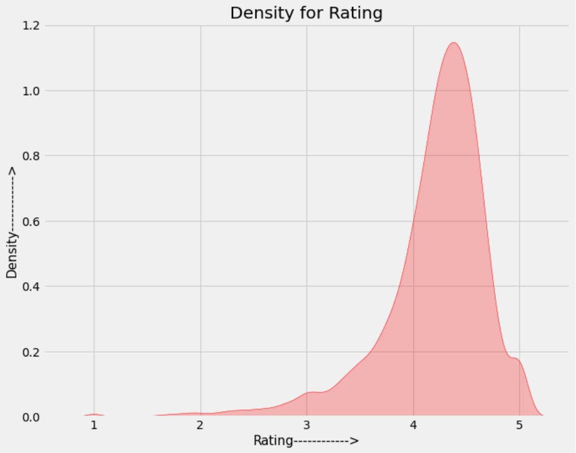
Family betrayed our 'Installs'.As we have seen so far, the list of Top 7 Characteristics of 'Category' (acc. to the number of apps developed) and Top 6 Characteristics (acc. to number of Installs) differ a lot.
As data analyst, we can use this information to decide our future insights.

Observation-2

let's see some other character "ratings" for which we are using violin plot and density plot that makes our story content more interesting in visual aspects.



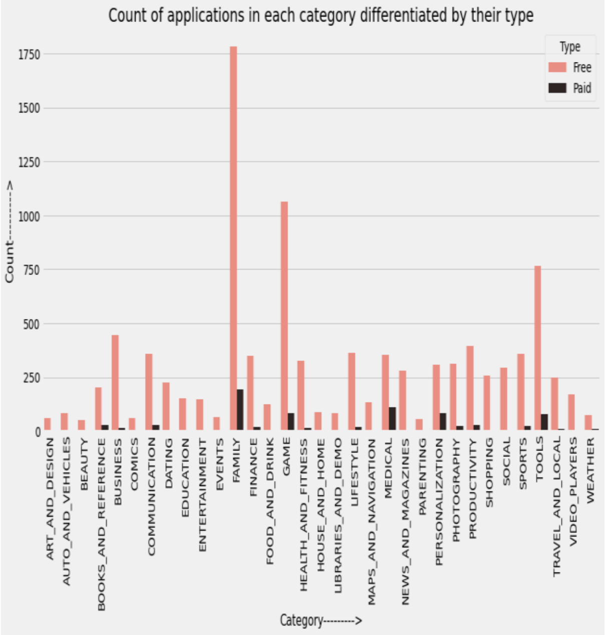
From our research, we found that the average volume of ratings across all app categories is between 4-4.5.



The histogram plot is skewed to the left indicating that the majority of the apps are highly rated with only a few exceptions in the low-rated apps.

Observation-3

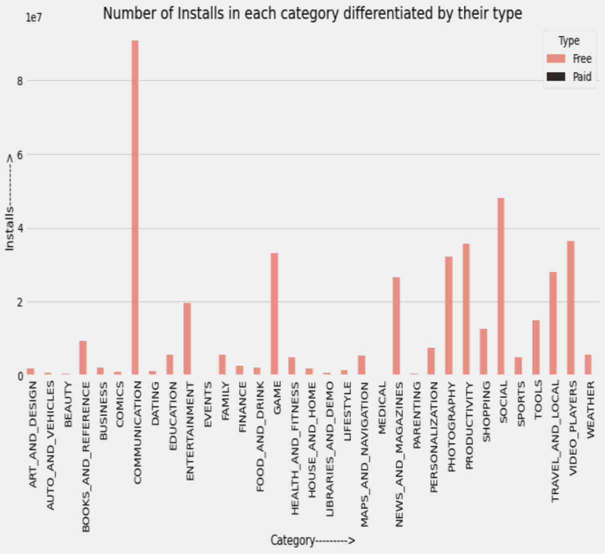
Here we are using count plot of Catergory column with hue as Type .



It looks like certain app categories have more free apps available for download than others. In our story, the majority of apps in the Family, Food & Drink, and Tools, as well as Social categories were free to install.

At the same time Family, Sports, Tools, and Medical categories had the biggest number of Paid apps available for download.

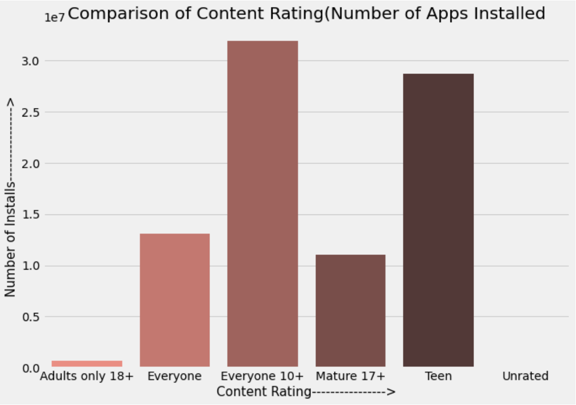
let's see how "Installs" affects the results of the above chart.



It can be concluded that the number of free applications installed by the user is high when compared with the paid ones

Observation-4

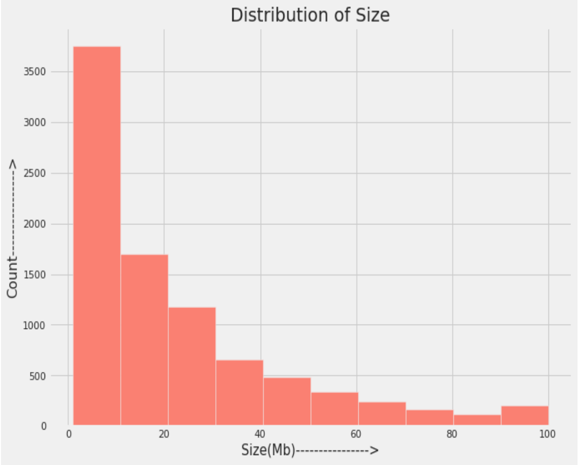
Impact of the Content Rating on its number of Installations of apps.



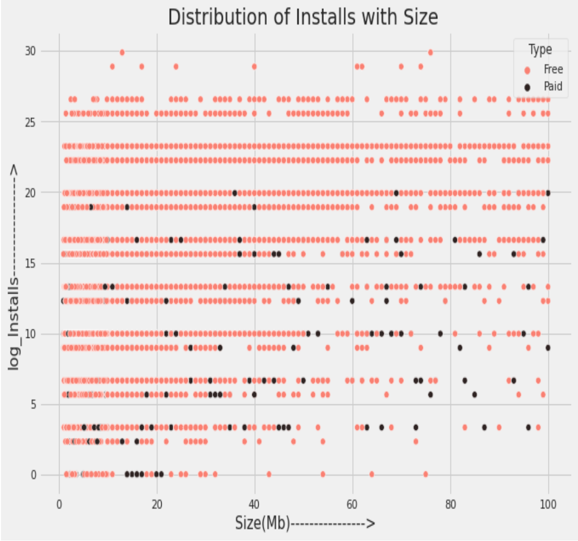
Number of ’Teen’ Apps is few as compared to ' Everyone' but when we check its 'Number of Installations', it seems like a good second best Choice. Few apps but Considerable Installations 'Everyone' is an easy option but 'Teen' and '10+' are the most rewarding.

Observation-5

Impact of size on installs



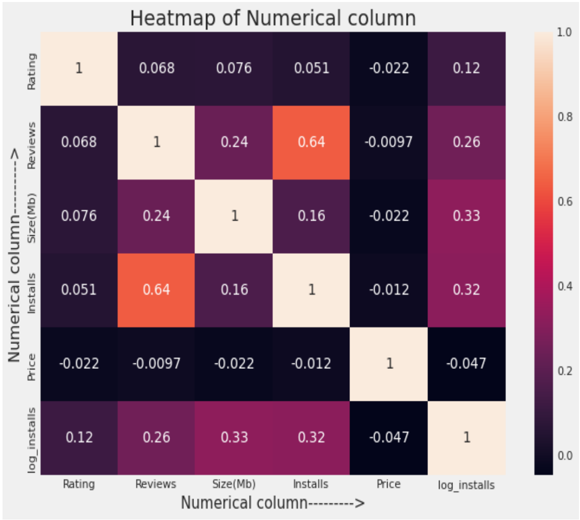
From the above histogram, it can be concluded that maximum number of applications present in the dataset are of small size



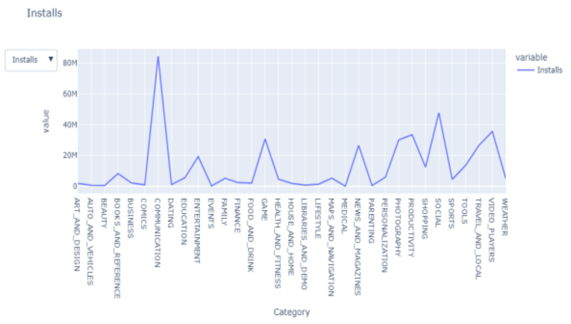
It is clear from the above mentioned plot that size may impact the number of installations. Bulky applications are less installed by the users

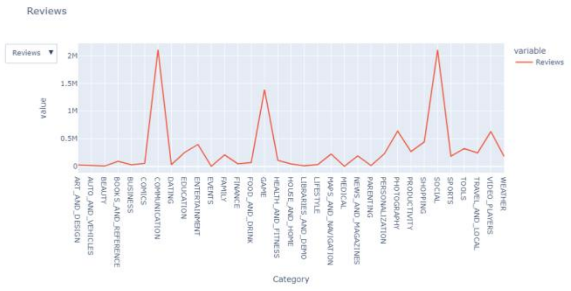
Correlation check

Next we can easily visualize the correlations between ‘Installs’ with other columns.



Certain visualization depicts that it is absolutely uncorrelated except ‘Reviews’.
Reiews have correlation of 0.64 with installs lets look into this

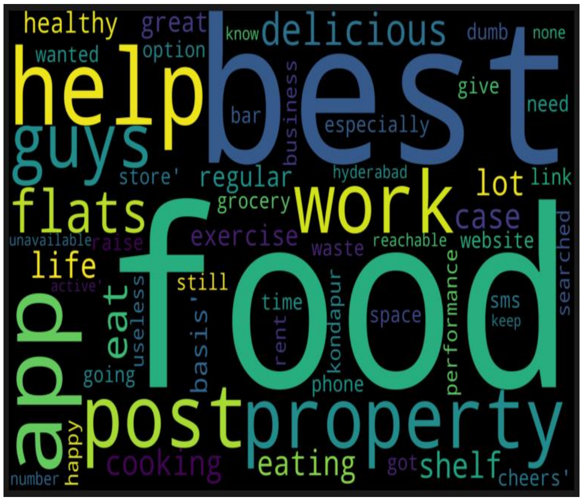




"Installs" vs "Reviews" have a correlation value of 0.64, shows perfect and strong correlationfrom above line chart as both "installs" and "Reviews" follow the same pattern approximately.

Observation-6

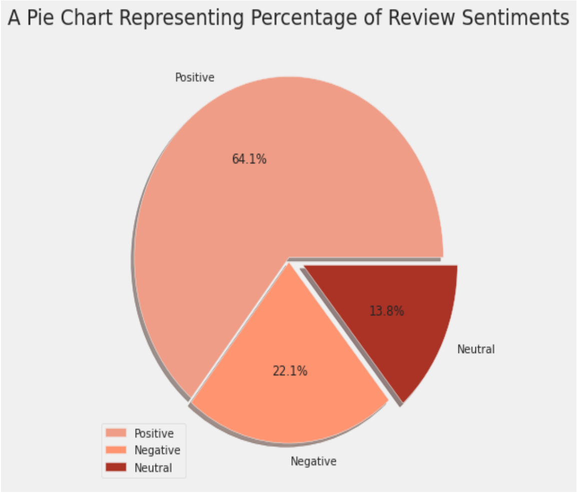
let's describe a poster/hashtag for our storyline/data, which helps the audience to see some repetitive words of our story.



Observation-7

sentiments give us an idea about the emotions of

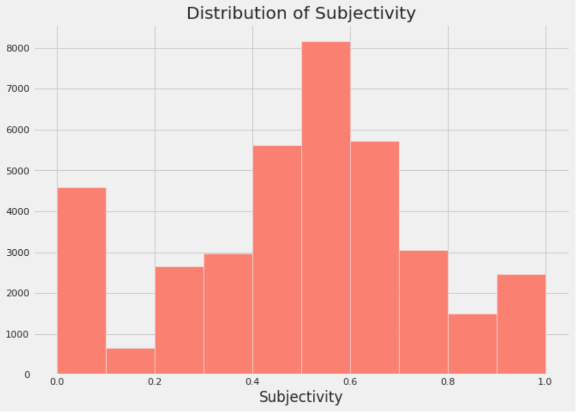
the story whether it's positive,negative or neutral



As is clear from the pie charts there are 64.1 % of Positive sentiments, 22.1% of Negative sentiments, and 13.8% neutral sentiments.

Observation-8

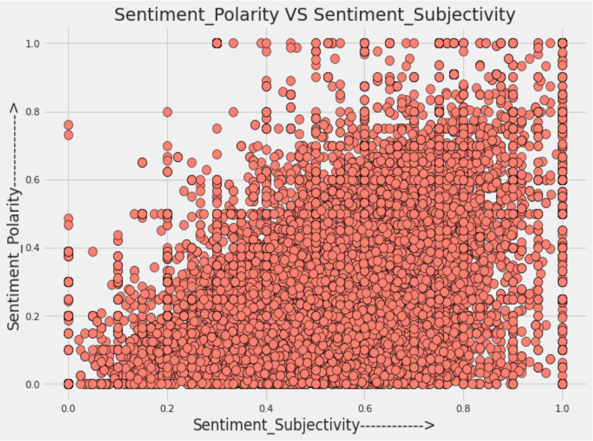
let's dive deep down into sentiments to see a more depth understanding of sentiment Polarity and Subjectivity.



It can be seen that the maximum number of sentiment subjectivity lies between 0.4 to 0.7. From this, we can conclude that the maximum number of the audience give reviews to the applications, according to their experience.

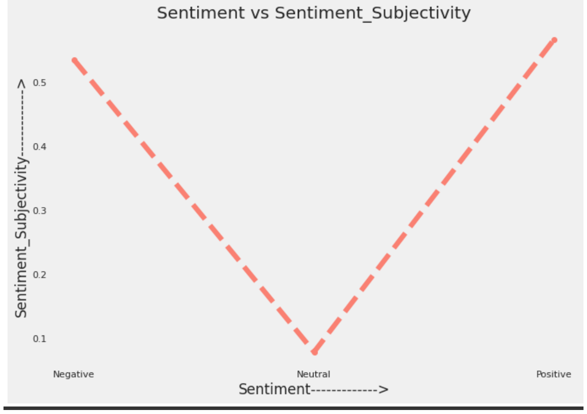
Observation-9

Does sentiments Polarity is proportional to sentiments subjectivity in our story?



From the above scatter plot it can be concluded that sentiment subjectivity is not always proportional to sentiment polarity but in maximum number of case, shows a proportional behavior, when variance is too high or low.

Observation-10



The above line has a Sentiment_Subjectivity of 0.55 for Negative sentiments and Positive sentiments of 0.60.

* Sentiment_Subjectivity > 0.5(refers to that mostly it is public opinion and not a piece of factual information)

CONCLUSION AND FUTURE WORK

Thus the app development companies could decide what application should be developed and they can also see the prediction of their developed application. In this they also get to see the categorized reviews of all the application in one interface which will help them decide which app is liked by the users and which apps need to be developed more.

The dataset contains immense possibilities to improve business values and have a positive impact. It is not limited to the problem taken into consideration for this project.

We could add a system that would create application on its own by using the data set and creating the best user interface by the highly rated apps.

ACKNOWLEDGEMENT

This project was completed by Nikhil Roeewal, Zayeem Ahmed,Ajit Padole, Manoj Sridala, Sumanta Banerjee. We are extremely grateful to the celebrated authors whose precious works have been consulted and referred in our project work. We also wish to convey our appreciation to our peers who provided encouragement and timely support in the hour of need.