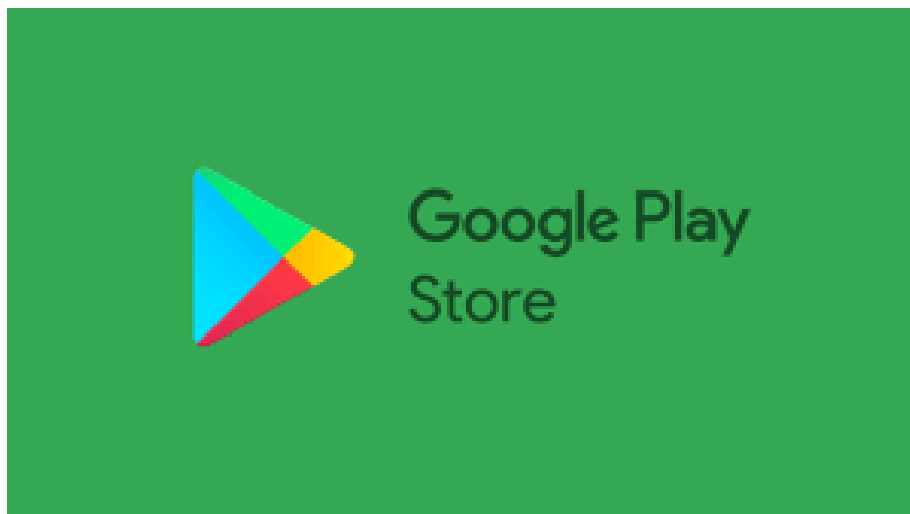# DATA SCIENCE MINOR PROJECT REPORT

(Project Semester August-December 2018)

Submitted by

Aashutosh Singh

Registration No. **11717459**

Programme and section: **B.Tech(RKM061A34)**

Course Code. **INT217**

Under the Guidance of

Hargobind Singh

Discipline of CSE/IT

Lovely School of Computer Science and Engineering

Lovely Professional University, Phagwara

<u>CERTIFICATE</u>

This is to certify that Aashutosh Singh bearing Registration no. 11717459 has completed INT217 project titled, **"GOOGLE PLAY STORE Analysis and Visualization"** under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.

Signature and Name of the Supervisor

Designation of the Supervisor

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab.

Date: /11/2019

<u>DECLARATION</u>

I, **Aashutosh Singh** student of **B.Tech (CSE)** under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: /11/2019   Signature

Registration No. 11717459   Aashutosh Singh

<u>ACKNOWLEDGEMENT</u>

I would like to express my thanks to the people who have helped me most throughout my project. I am grateful to my teacher Mr. Hargobind Singh for nonstop support for the project.

A special thank of mine goes to my colleague who helped me out in completing the project, where they all exchanged their own interesting ideas, thoughts and made this possible to complete my project with all accurate information. I wish to thank my parents for their personal support or attention who inspired me to go my own way.

At last but not the least I want to thank my friends who treasured me for my hard work and encouraged me and finally to God who made all the things possible for me till the end.

TABLE OF CONTENT

INTRODUCTION

Google Play (previously Android Market) is a digital distribution service operated and developed by Google. It serves as the official app store for the Android operating system, allowing users to browse and download applications developed with the Android software development kit (SDK) and published through Google. Google Play also serves as a digital media store, offering music, magazines, books, movies, and television programs. It previously offered Google hardware devices for purchase until the introduction of a separate online hardware retailer, Google Store, on March 11, 2015.

Applications are available through Google Play either free of charge or at a cost. They can be downloaded directly on an Android device through the Play Store mobile app or by deploying the application to a device from the Google Play website. Applications exploiting hardware capabilities of a device can be targeted to users of devices with specific hardware components, such as a motion sensor (for motion-dependent games) or a front-facing camera (for online video calling). The Google Play store had over 82 billion app downloads in 2016 and has reached over 3.5 million apps published in 2017.It has been the subject of multiple issues concerning security, in which malicious software has been approved and uploaded to the store and downloaded by users, with varying degrees of severity

Play Store is Google's official pre-installed app store on Android-certified devices. It provides access to content on the Google Play Store, including apps, books, magazines, music, movies, and television programs.

Play Store filters the list of apps to those compatible with the user's device. Developers can target specific hardware components (such as compass), software components (such as widget), and Android versions (such as 7.0 Nougat). Carriers can also ban certain apps from being installed on users' devices, for example tethering applications.

There is no requirement that Android applications must be acquired using the Play Store. Users may download Android applications from a developer's website or through a third-party app store alternative. Play Store applications are self-contained Android Package files (APK), similar to .exe files to install programs on Microsoft Windows computers. On Android devices, an "Unknown sources" feature in Settings allows users to bypass the Play Store and install APKs from other sources. Depending on developer preferences, some apps can be installed to a phone's external storage card.

Android users have complained that the Google Play store access cannot be blocked and there is constant data exchange with the google cloud. Also, valuable CPU resources are used, slowing down the Android system.

SCOPE OF ANALYSIS

The system provides different types of services based on the different types of application in the play store. The analysis provides we count the downloaded application in each category wise number of software that costumer had downloaded and which software has most downloaded.
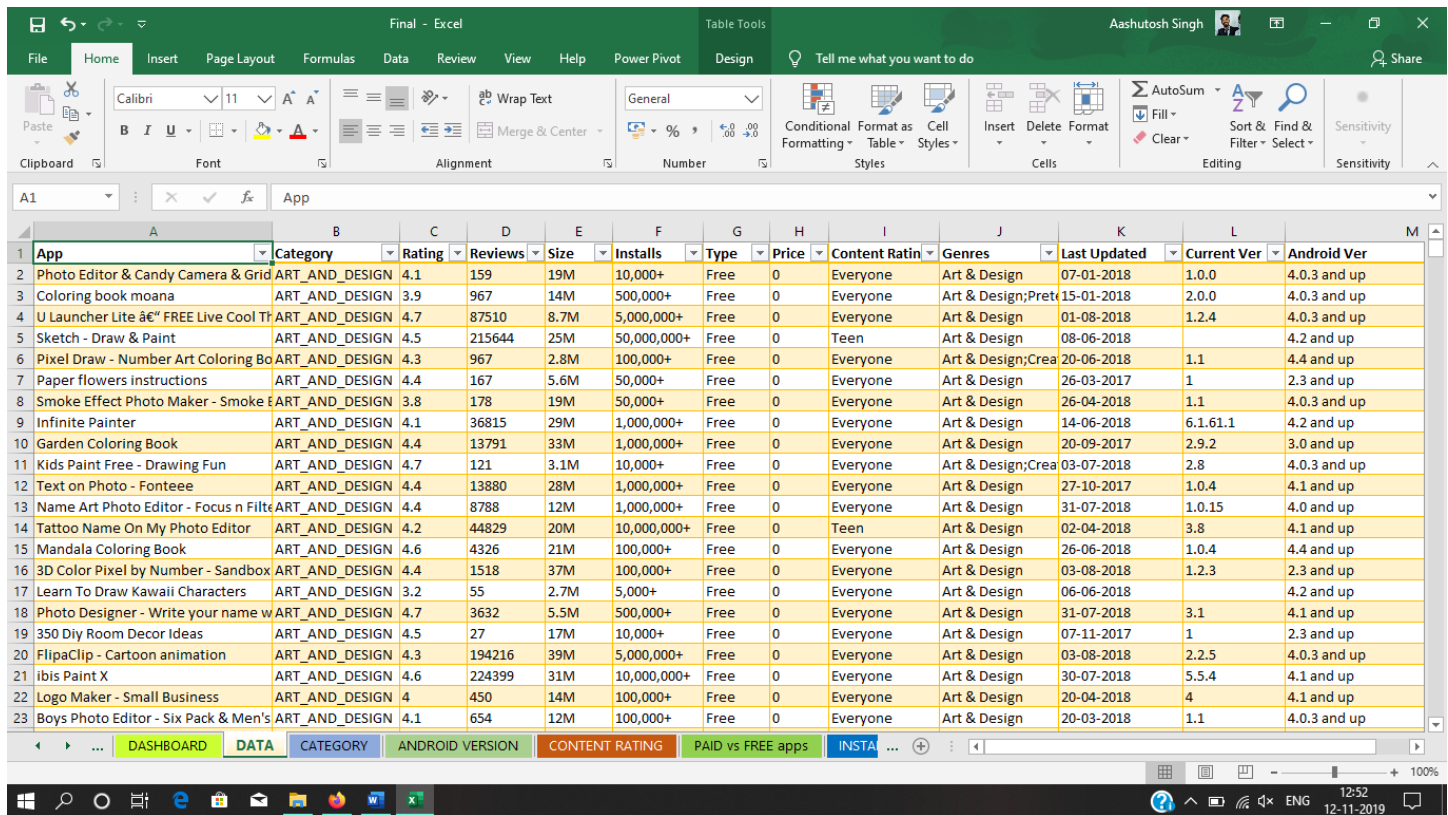
The objective in play store data management are:

1.   Total number of software category wise

2.   Android version

3.   Total number of software in each content rating

4.   Total number of free software and paid

5.   Number of install software

6.   Updation

7.   Reviews

8.   Sentiments

By this analysis costumer know that Health & Fitness, Travel & Local and Education were the most common categories, accounting for about 15%, 14% and 13%, respectively, of the total number of apps in our dataset. Food & Drink was the least prevalent category with only 87 observations (less than 1% of the total number of apps in our dataset). Four categories of applications, including Health & Fitness, Travel & Local, Education and Finance, accounted for more than 50% of the apps in our dataset.

My System

The data looks like this:

In the data the columns like **apps, category, rating, reviews, size, installs, type, price, content rating, genres**

Last update, current version, android version

apps: In this column the data is filled according to total number of install.

category: In this column the data filled category wise.

rating: In this column the data filed with increasing to decreasing wise.

reviews: In the given data set filled with reviews.

size: In this column the number of software ordered by each size.

install: In this column total number of software downloaded by each category.

types: In this column date filled with given software free or paid.

price: In this column different types of price are given.

Content rating: In this column the given data set told about who can access the data like everyone10+ or teen.

genres: In this column the told about given dataset is video player or photographics .

last update: In this column last update of software.

Current version: In this column tell about current version of software.

Android version: In this column we know which type of android version required for given software.

Sales: In this we use formula for counting sales by multiplying selling price by total number of books.

In existing system, we can find out

- **Total number of software category wise**
- **Android version**
- **Total number of software in each content rating**
- **Total number of software free or paid**
- **Total number of installed applications**
- **Updation**

App: Name of app

Sentiment: Positive/Negative/Neutral (Preprocessed)

Sentiment_Polarity: Sentiment polarity score

Sentiment_Subjectivity: Sentiment subjectivity score

In existing system, we can find out

- **Reviews each application.**
- **Sentiments for particular application.**

Drawbacks or limitations of existing system:

We can't enter large amount of data in excel sheet manually so we need some software for managing the data sheet. In given data analysis can be done only once after that if data set has changed then for updated data analysis is required again. For managing larger data we use power pivot and one of the limitation of power pivot is it does not provide functionality as compared to excel sheet.

Source of dataset

The source of data is online site   https://www.kaggle.com/datasets

ETL PROCESS

In computing, extract, transform, load (ETL) is a process in database usage to prepare data for analysis, especially in data warehousing. The ETL process became a popular concept in the 1970s. Data extraction involves extracting data from homogeneous or heterogeneous sources, while data transformation processes data by transforming them into a proper storage format/structure for the purposes of querying and analysis; finally, data loading describes the insertion of data into the final target database such as an operational data store, a data mart, or a data warehouse. A properly designed ETL system extracts data from the source systems, enforces data quality and consistency standards, conforms data so that separate sources can be used together, and finally delivers data in a presentation-ready format so that application developers can build applications and end users can make decisions.

Since the data extraction takes time, it is common to execute the three phases in parallel. While the data is being extracted, another transformation process executes while processing the data already received and prepares it for loading while the data loading begins without waiting for the completion of the previous phases.

ETL systems commonly integrate data from multiple applications (systems), typically developed and supported by different vendors or hosted on separate computer hardware. The separate systems containing the original data are frequently managed and operated by different employees. For example, a cost accounting system may combine data from payroll, sales, and purchasing.

Extract

The first part of an ETL process involves extracting the data from the source system(s). In many cases, this represents the most important aspect of ETL, since extracting data correctly sets the stage for the success of subsequent processes. Most data-warehousing projects combine data from different source systems. Each separate system may also use a different data organization and/or format. Common data-source formats include relational databases, XML, JSON and flat files, but may also include non-relational database structures such as Information Management System (IMS) or other data structures such as Virtual Storage Access Method (VSAM) or Indexed Sequential Access Method (ISAM), or even formats fetched from outside sources by means such as web spidering or screen-scraping. The streaming of the extracted data source and loading on-the-fly to the destination database is another way of performing ETL when no intermediate data storage is required. In general, the extraction phase aims to convert the data into a single format appropriate for transformation processing.

An intrinsic part of the extraction involves data validation to confirm whether the data pulled from the sources has the correct/expected values in a given domain (such as a pattern/default or list of values). If the data fails the validation rules it is rejected entirely or in part. The rejected data is ideally reported back to the source system for further analysis to identify and to rectify the incorrect records.

Transform

In the data transformation stage, a series of rules or functions are applied to the extracted data in order to prepare it for loading into the end target. Some data does not require any transformation at all; such data is known as "direct move" or "pass through" data.

An important function of transformation is the cleaning of data, which aims to pass only "proper" data to the target. The challenge when different systems interact is in the relevant systems' interfacing and communicating. Character sets that may be available in one system may not be so in others.

In other cases, one or more of the following transformation types may be required to meet the business and technical needs of the server or data warehouse:

Selecting only certain columns to load: (or selecting null columns not to load). For example, if the source data has three columns (aka "attributes"), roll_no, age, and salary, then the selection may take only roll_no and salary. Or, the selection mechanism may ignore all those records where salary is not present (salary = null).

Translating coded values: (e.g., if the source system codes male as "1" and female as "2", but the warehouse codes male as "M" and female as "F")

Encoding free-form values: (e.g., mapping "Male" to "M")

Deriving a new calculated value: (e.g., sale_amount = qty * unit_price)

Sorting or ordering the data based on a list of columns to improve search performance

Joining data from multiple sources (e.g., lookup, merge) and deduplicating the data

Aggregating (for example, rollup — summarizing multiple rows of data — total sales for each store, and for each region, etc.)

Generating surrogate-key values

Transposing or pivoting (turning multiple columns into multiple rows or vice versa)

Splitting a column into multiple columns (e.g., converting a comma-separated list, specified as a string in one column, into individual values in different columns)

Disaggregating repeating columns

Looking up and validating the relevant data from tables or referential files

Applying any form of data validation; failed validation may result in a full rejection of the data, partial rejection, or no rejection at all, and thus none, some, or all of the data is handed over to the next step depending on the rule design and exception handling; many of the above transformations may result in exceptions, e.g., when a code translation parses an unknown code in the extracted data

LOAD

The load phase loads the data into the end target, which may be a simple delimited flat file or a data warehouse. Depending on the requirements of the organization, this process varies widely. Some data warehouses may overwrite existing information with cumulative information; updating extracted data is frequently done on a daily, weekly, or monthly basis. Other data warehouses (or even other parts of the same data warehouse) may add new data in a historical form at regular intervals—for example, hourly. To understand this, consider a data warehouse that is required to maintain sales records of the last year. This data warehouse overwrites any data older than a year with newer data. However, the entry of data for any one year window is made in a historical manner. The timing and scope to replace or append are strategic design choices dependent on the time available and the business needs. More complex systems can maintain a history and audit trail of all changes to the data loaded in the data warehouse.

As the load phase interacts with a database, the constraints defined in the database schema — as well as in triggers activated upon data load — apply (for example, uniqueness, referential integrity, mandatory fields), which also contribute to the overall data quality performance of the ETL process.

For example, a financial institution might have information on a customer in several departments and each department might have that customer's information listed in a different way. The membership department might list the customer by name, whereas the accounting department might list the customer by number. ETL can bundle all of these data elements and consolidate them into a uniform presentation, such as for storing in a database or data warehouse.

Another way that companies use ETL is to move information to another application permanently. For instance, the new application might use another database vendor and most likely a very different database schema. ETL can be used to transform the data into a format suitable for the new application to use.

An example would be an Expense and Cost Recovery System (ECRS) such as used by accountancies, consultancies, and legal firms. The data usually ends up in the time and billing system, although some businesses may also utilize the raw data for employee productivity reports to Human Resources (personnel dept.) or equipment usage reports to Facilities Management.
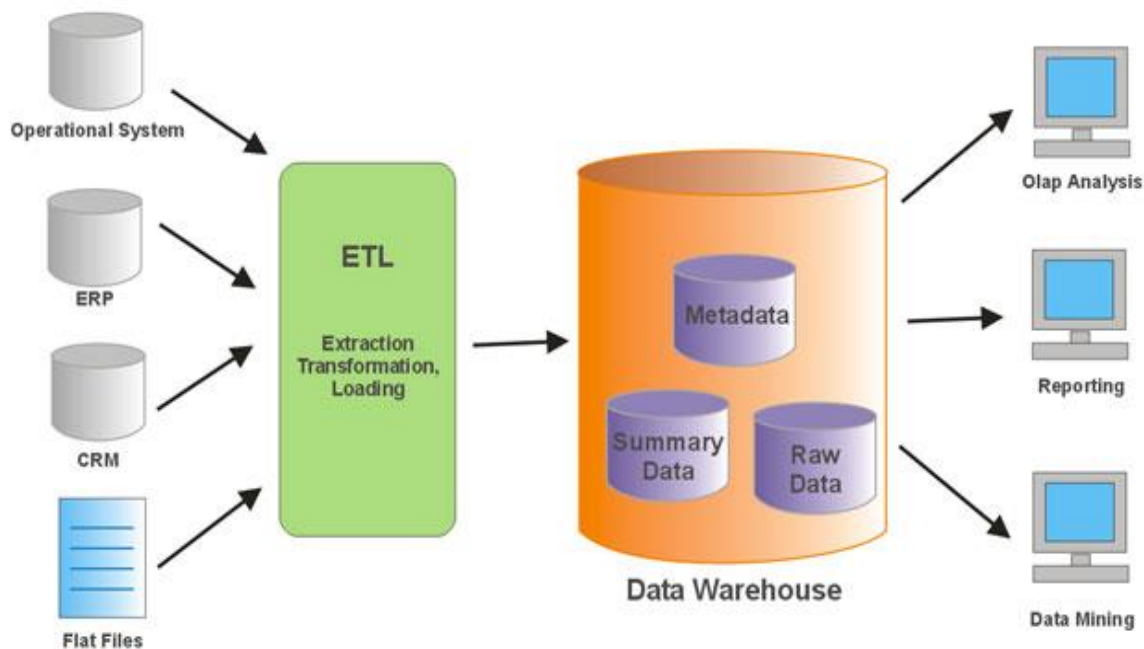
In our scenario, dataset is Mobileappstore.csv, so during the ETL process the data is extracted from this dataset, transformed to eliminate irrelevant data mentioned in the scope of analysis section and loaded into the excel where the required data resides. From this analysis reporting can be done.

Analysis and Visualization on dataset (for each analysis)

Introduction:

Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, while being used in different business, science, and social science domains.

Data visualization is viewed by many disciplines as a modern equivalent of visual communication. It involves the creation and study of the visual representation of data.

To communicate information clearly and efficiently, data visualization uses statistical graphics, plots, information graphics and other tools. Numerical data may be encoded using dots, lines, or bars, to visually communicate a quantitative message. Effective visualization helps users analyze and reason about data and evidence. It makes complex data more accessible, understandable and usable. Users may have particular analytical tasks, such as making comparisons or understanding causality, and the design principle of the graphic (i.e., showing comparisons or showing causality) follows the task. Tables are generally used where users will look up a specific measurement, while charts of various types are used to show patterns or relationships in the data for one or more variables.

The process of data analysis:

- Data requirements
- Data collection
- Data processing
- Data cleaning
- Modeling and algorithms

Specific Requirements, functions and formulas:

We use different formulas for data cleaning, data processing and modelling and algorithms.

Main function of excel that I have used are

- Power pivot
- Pivot table

- Different arithmetic formulas
- Conditional formatting
- Create graph using power pivot
- Create relation between different table
- Function analyser (Slicer)
- Different types of graph
- Different functionality of graph
- Comparison between different quantities
- Absolute cell reference
- Cell formatting

Analysis results:

Analysis result are mention below:

1. **Total number of software category wise**

   - **Description:** In this analysis, we can see how many applications are their installed category-wise.

   - **Analysis Result:**

Row Labels	Count of Installs
ART_AND_DESIGN	62
AUTO_AND_VEHICLES	73
BUSINESS	303
COMMUNICATION	328
DATING	195
MAPS_AND_NAVIGATION	124
PERSONALIZATION	312
SPORTS	319
VIDEO_PLAYERS	160
Grand Total	**1876**

   - **Visualisation:**
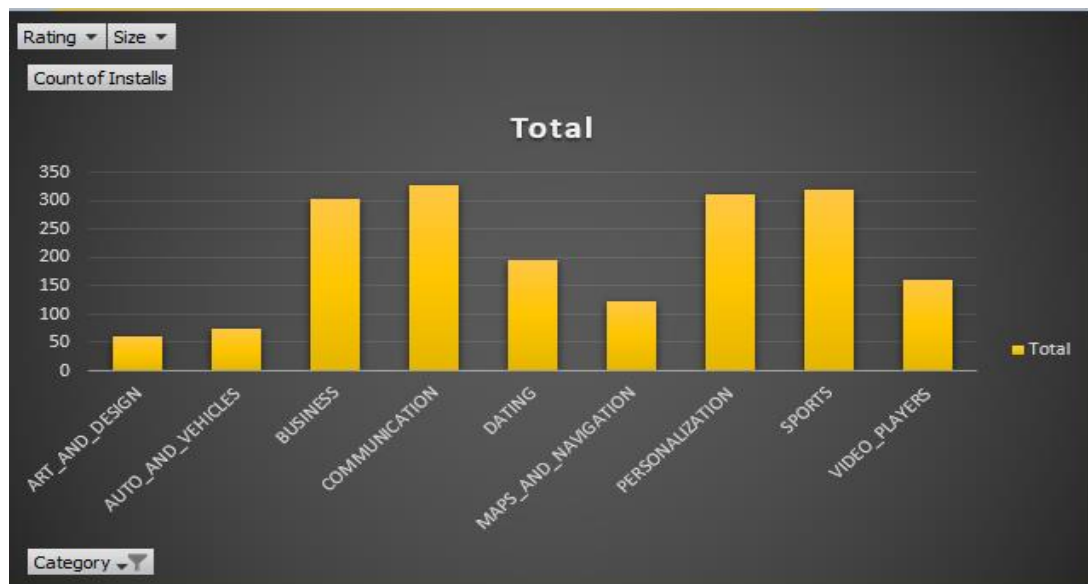
2. Android version:

- **Description:** Here we can depict that how many application are paid and free according to their version.

- **Analysis Result:**

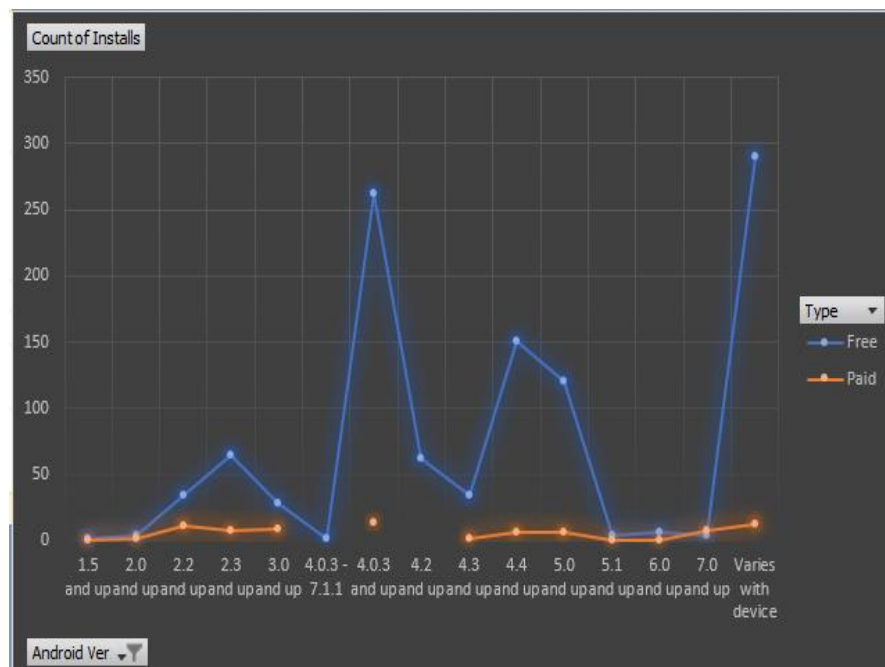| Count of Installs	Column Labels		
Row Labels	Free	Paid	Grand Total
1.5 and up	2	1	3
2.0 and up	4	2	6
2.2 and up	34	11	45
2.3 and up	65	8	73
3.0 and up	28	9	37
4.0.3 - 7.1.1	2		2
4.0.3 and up	262	14	276
4.2 and up	62		62
4.3 and up	35	2	37
4.4 and up	151	6	157
5.0 and up	120	7	127
5.1 and up	4	1	5
6.0 and up	6	1	7
7.0 and up	4	8	12
Varies with device	290	12	302
Grand Total	1069	82	1151

- **Visualization:**

3. **Total number of software in each content rating:**

- **Description:** In this analysis we will use type of application section for representing the number of applications installed as per content rating they have.

- **Analysis Result:**

Count of Installs	Column Labels		
Row Labels	Free	Paid	Grand Total
Adults only 18+	1		1
Everyone	1397	135	1532
Everyone 10+	30		30
Mature 17+	203	3	206
Teen	107		107
Grand Total	1738	138	1876

- **Visualization:**
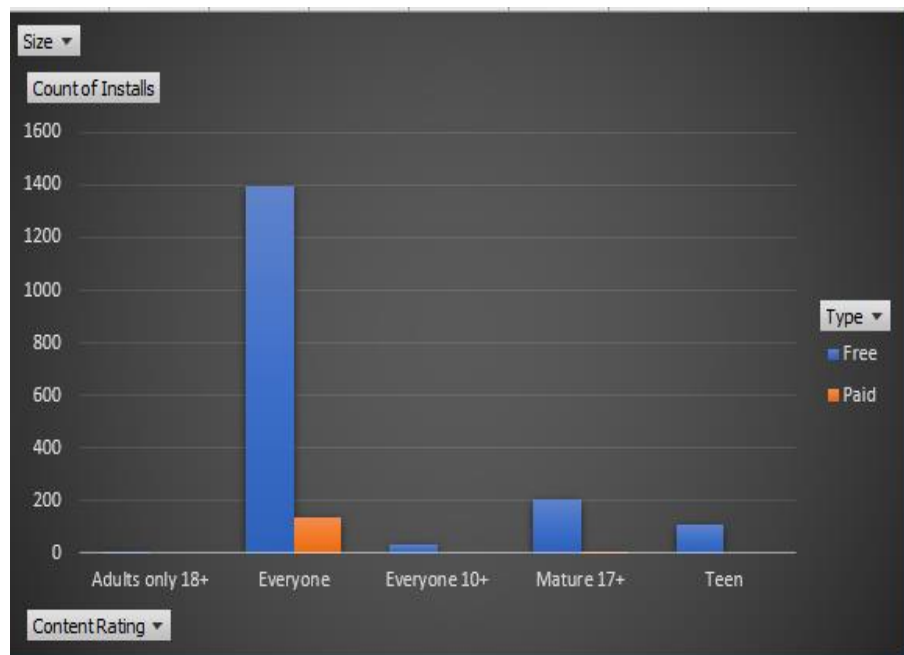
**4. Total number of free software and paid:**

- **Description:** Here, we will analyze the apps which are paid and free on the basis of their category they have.

- **Analysis Result:**

| Count of Installs	Column Labels		
Row Labels	Free	Paid	Grand Total
ART_AND_DESIGN	59	3	62
AUTO_AND_VEHICLES	72	1	73
BUSINESS	292	11	303
COMMUNICATION	306	22	328
DATING	191	4	195
MAPS_AND_NAVIGATION	119	5	124
PERSONALIZATION	246	66	312
SPORTS	297	22	319
VIDEO_PLAYERS	156	4	160
Grand Total	1738	138	1876

- **Visualization:**
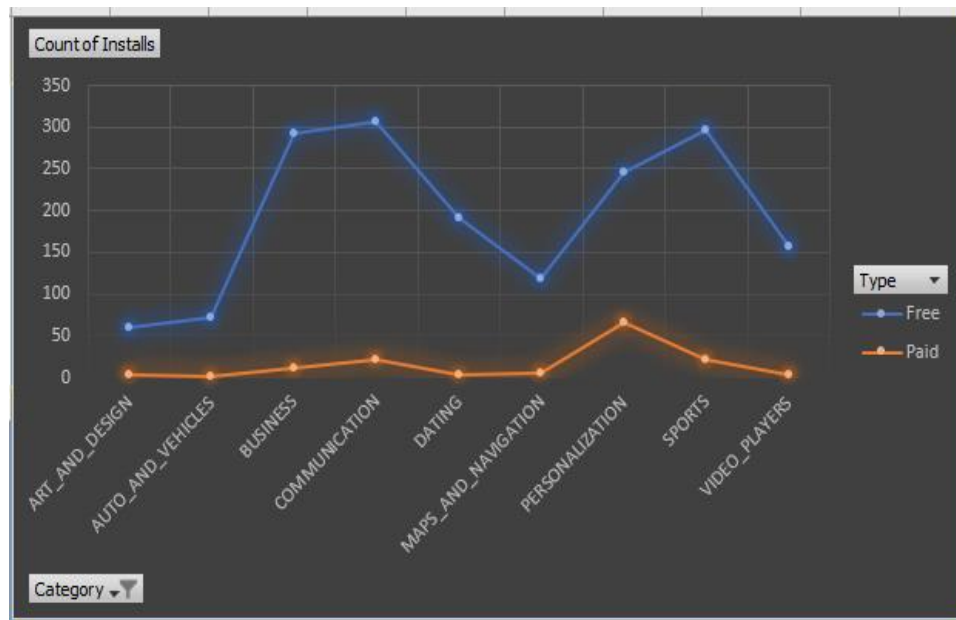
5. **Number of install software:**

    - **Description:** We can depict from the analysis about the count of installation taken place according to the android version.

    - **Analysis Result:**

Row Labels	Count of Installs
2.5	1
2.7	1
2.9	1
3.1	2
3.4	4
3.5	3
3.6	1
3.7	3
3.8	3
3.9	2
4	3
4.1	4
4.2	6
4.3	3
4.4	7
4.5	2
4.6	3
Grand Total	49

    - **Visualization:**
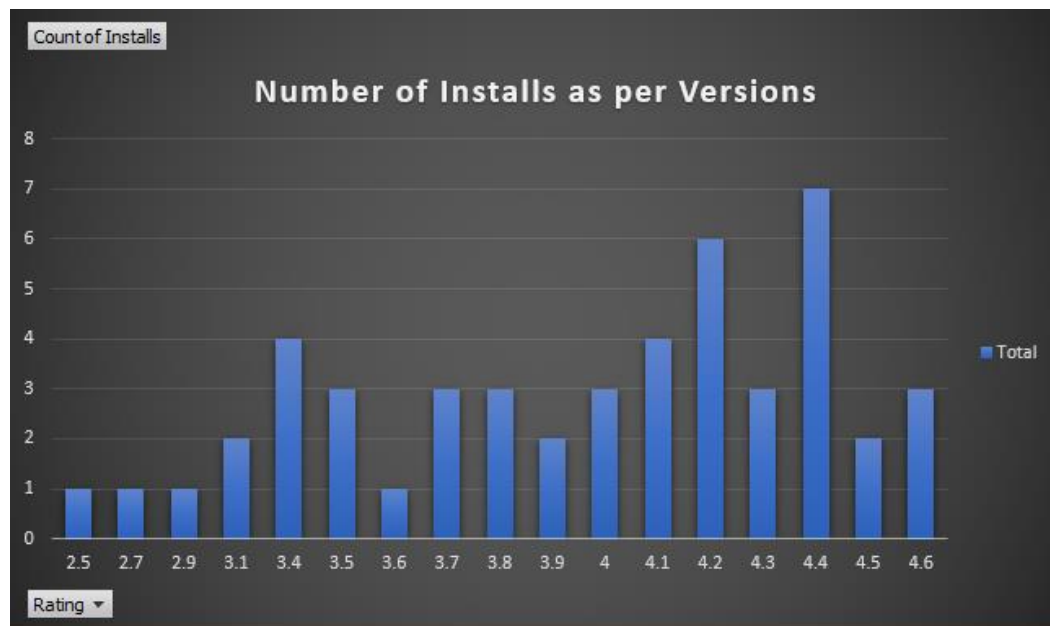
6. **UPDATION:**

- **Description:** This analysis shows the count of application updated on the basis of the months and years.

- **Analysis Result:**

| Row Labels	Count of Current Ver
⊟ Jan	76
2012	1
2014	3
2015	6
2016	5
2017	18
2018	43
⊞ Feb	71
⊞ Mar	97
⊞ Apr	108
⊞ May	162
⊞ Jun	194
⊞ Jul	612
⊞ Aug	330
⊞ Sep	49
⊞ Oct	58
⊞ Nov	55
⊞ Dec	61
Grand Total	**1873**

- **Visualization:**

7. REVIEWS:

- **Description:** This analysis helps to depict the different application having positive, negative and neutral reviews.

- **Analysis Result:**

| Count of App	Column Labels			
Row Labels	Negative	Neutral	Positive	Grand Total
10 Best Foods for You	10	22	162	194
Airbnb	19	7	36	62
Asphalt 8: Airborne	31	3	39	73
Chrome Beta	4	12	22	38
Google Earth	10	5	23	38
Grand Total	74	49	282	405

- **Visualization:**
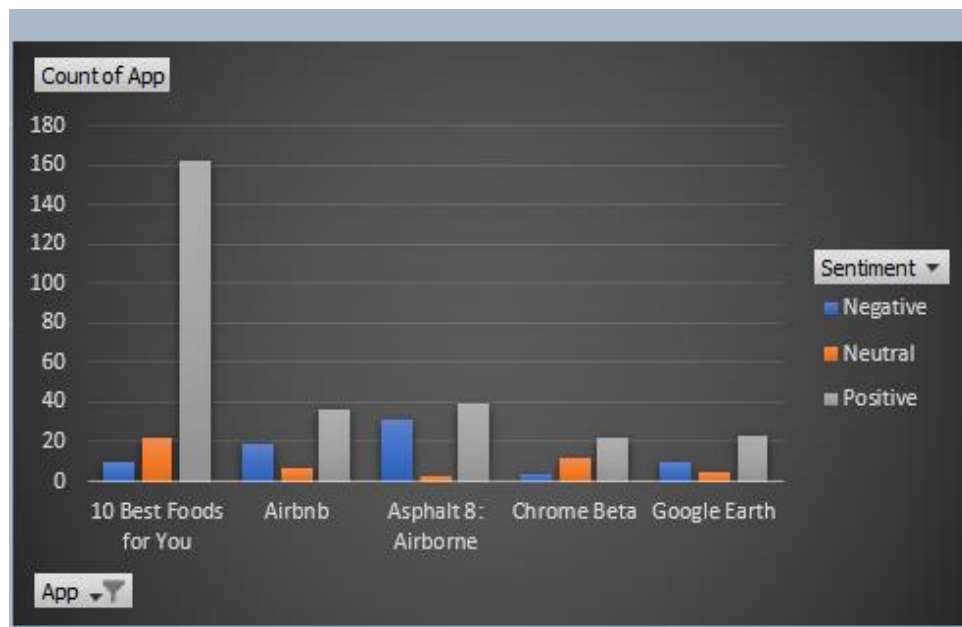
8. **Sentiments:**

- **Description:** Through we can see for each application having how much positive and negative feedback.

- **Analysis Result:**

Row Labels .T	Count of Sentiment
⊟ Facebook	113
Negative	60
Positive	53
Grand Total	**113**

- **Visualization:**
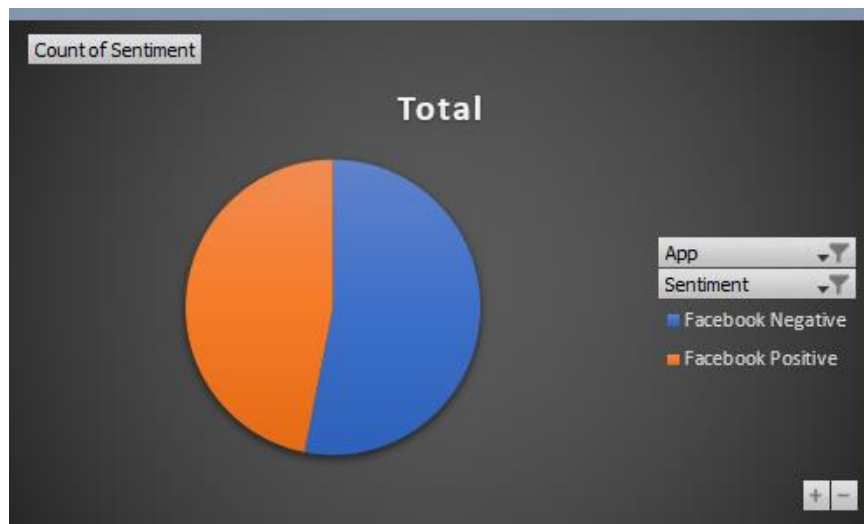
Future Scope

Data analytics is a process through which data is cleaned, analyzed and modelled using tools. This data is then used to derive insights. The insights are then used for business related decision-making purposes. There are many techniques that data analysts use in different fields of work. In the world of business, Data analytics is used for making strategies to get the desired business results. Today, data analytics has become a big career option in India. As a result, big data analytics courses are in huge demand.

Businesses have realized the importance of utilizing big data analytics to maximize their profits. They know that it is vital for their growth and for the future health of their business. Today, major business decisions are taken by utilizing the insights derived from data related to the organization or industry related data. As competition increases and customers are flooded with choices, it has become important to move faster in the market and that too with accuracy.

Bibliography

- https://nycdatascience.com/blog/student-works/web-scraping/analysis-of-apps-in-the-google-play-store/
- https://www.kaggle.com/datasets
- https://www.youtube.com/results?search_query=dashboard+on+google+play+store+in+excel