

**VISVESVARAYA TECHNOLOGICAL UNIVERSITY,
BELAGAVI 590018**



Report on

“Google Play Store App Data Analytics ”

By

Riyanchhi Agrawal (1BM17CS076)
Shehyaaz Khan Nayazi (1BM17CS094)

Under the Guidance of

Mrs. PALLAVI G B
(Assistant Professor)

Department of CSE
BMS College of Engineering

Work carried out at



Department of Computer Science and Engineering
BMS College of Engineering
(Autonomous college under VTU)
P.O. Box No.: 1908, Bull Temple Road, Bangalore-560 019
2017-2018

BMS COLLEGE OF ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND
ENGINEERING



CERTIFICATE

This is to certify that the BIG DATA and ANALYTICS mini Project titled “**Google Play Store App Data Analytics**” has been carried out by Riyanhhi Agrawal (1BM17CS076), Shehyaaz Khan Nayazi (1BM17CS094) during the academic year 2020-2021.

Signature of the guide

Mrs. PALLAVI G B

Assistant Professor,

Department of Computer Science and Engineering

BMS College of Engineering, Bangalore

BMS COLLEGE OF ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND
ENGINEERING



DECLARATION

We, Riyanchhi Agrawal (1BM17CS076) , Shehyaaz Khan Nayazi (1BM17CS094), students of 7th Semester, B.E, Department of Computer Science and Engineering, BMS College of Engineering, Bangalore, hereby declare that, this assignment work entitled "**Google Play Store App Data Analytics**" has been carried out by us under the guidance of **Mrs. PALLAVI G B (Assistant Professor)**, Department of CSE, BMS College of Engineering, Bangalore during the academic semester Aug- Dec 2020. We also declare that to the best of our knowledge and belief, the assignment reported here is not from part of any other report by any other students.

Signature of the Candidates

RIYANCHHI AGRAWAL (1BM17CS076)

SHEHYAAZ KHAN NAYAZI (1BM17CS094)

INDEX

| Sl. No. | TOPIC | PAGE No. |
|----------------|---------------------------------------|-----------------|
| 1. | OBJECTIVE | |
| 2. | INTRODUCTION | |
| 3. | DESIGN MODULES | |
| 4. | DETAILED DESCRIPTION OF MODULES | |
| 5. | SCREENSHOTS | |
| 6. | NEW LEARNINGS AND FUTURE ENHANCEMENTS | |

1. OBJECTIVE

Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes.

Big data analytics applications allow data analysts, data scientists, predictive modelers, statisticians and other analytics professionals to analyze growing volumes of structured transaction data, plus other forms of data that are often left untapped by conventional BI and analytics programs.

This includes a mix of semi-structured and unstructured data. For example, internet clickstream data, web server logs, social media content, text from customer emails and survey responses, mobile phone records, and machine data captured by sensors connected to the internet of things (IoT).

Kaggle, a subsidiary of Google , is an online community of data scientists and machine learning practitioners.

Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

Apart from a huge number of public datasets (on Kaggle) , there are not many counterpart datasets available for Google Play Store apps anywhere on the web. Google Play Store uses sophisticated modern-day techniques (like dynamic page load) using JQuery making scraping more challenging.

The Play Store apps data has enormous potential to drive app-making businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market.

The dataset is the web scraped data of Play Store apps for analyzing the Android market. It consists of a total of 4470 records with 13 columns.

2. INTRODUCTION

Analysis of data examines large and different types of data and study their growth and development.

It involves three major stages-

- Pre-processing the data
- Performing Analytics over data
- Visualizing the data.

Data Pre-processing - It involves methods and techniques used to discover knowledge from the data. As data is most likely to be imperfect, inconsistent and sometimes redundant it cannot be directly used in the next step i.e. Data Mining process. Data preprocessing stage enables us to process data by adapting it to the requirements presented by each data mining algorithm.

General steps in pre-processing-

- **Noise Identification:** The veracity of data depends on identifying and removing additional meaningless information present in data. The noise can also be random fluctuations in data values that harm analytic predictions.
- **Data Cleaning:** Detecting and correcting or removing corrupt or inaccurate records in raw data by removing typographical errors and validating values against a known list of entities. Noisy data is also smoothened out or removed in this stage.
- **Data Normalization:** Where skew in data is removed by transforming all variables in data to specific range for better analytics processing.
- **Data Transformation:** Converting data from one format or structure to another format suitable for integration, warehousing and wrangling.
- **Data Integration:** Combining data from multiple sources into an integrated view. As volume of data increases and more machine learning algorithms are applied, data integration becomes more critical.

The **Data Analysis** stage is dedicated to carrying out the actual analysis task, which typically involves one or more types of analytics. This stage can be iterative in nature, especially if the data analysis is exploratory, in which case analysis is repeated until the appropriate pattern or correlation is uncovered. The exploratory analysis approach will be explained shortly, along with confirmatory analysis.

Depending on the type of analytic result required, this stage can be as simple as querying a dataset to compute an aggregation for comparison. On the other hand, it can be as challenging as combining data mining and complex statistical analysis techniques to discover patterns and anomalies or to generate a statistical or mathematical model to depict relationships between variables.

Data analysis can be classified as

- Confirmatory analysis
- Exploratory analysis.

Confirmatory data analysis is a deductive approach where the cause of the phenomenon being investigated is proposed beforehand. The proposed cause or assumption is called a hypothesis. The data is then analyzed to prove or disprove the hypothesis and provide definitive answers to specific questions. Data sampling techniques are typically used. Unexpected findings or anomalies are usually ignored since a predetermined cause was assumed.

Exploratory data analysis is an inductive approach that is closely associated with data mining. No hypothesis or predetermined assumptions are generated. Instead, the data is explored through analysis to develop an understanding of the cause of the phenomenon. Although it may not provide definitive answers, this method provides a general direction that can facilitate the discovery of patterns or anomalies.

Data Visualization- The ability to analyze massive amounts of data and find useful insights carries little value if the only ones that can interpret the results are the analysts. The Data Visualization stage is dedicated to using data visualization techniques and tools to graphically communicate the analysis results for effective interpretation by business users.

Business users need to be able to understand the results in order to obtain value from the analysis and subsequently have the ability to provide feedback. The results of completing the Data Visualization stage provide users with the ability to perform visual analysis, allowing for the discovery of answers to questions that users have not yet even formulated. Visual analysis techniques are covered later in this book. The same results may be presented in a number of different ways, which can influence the interpretation of the results. Consequently, it is important to use the most suitable visualization technique by keeping the business domain in context.

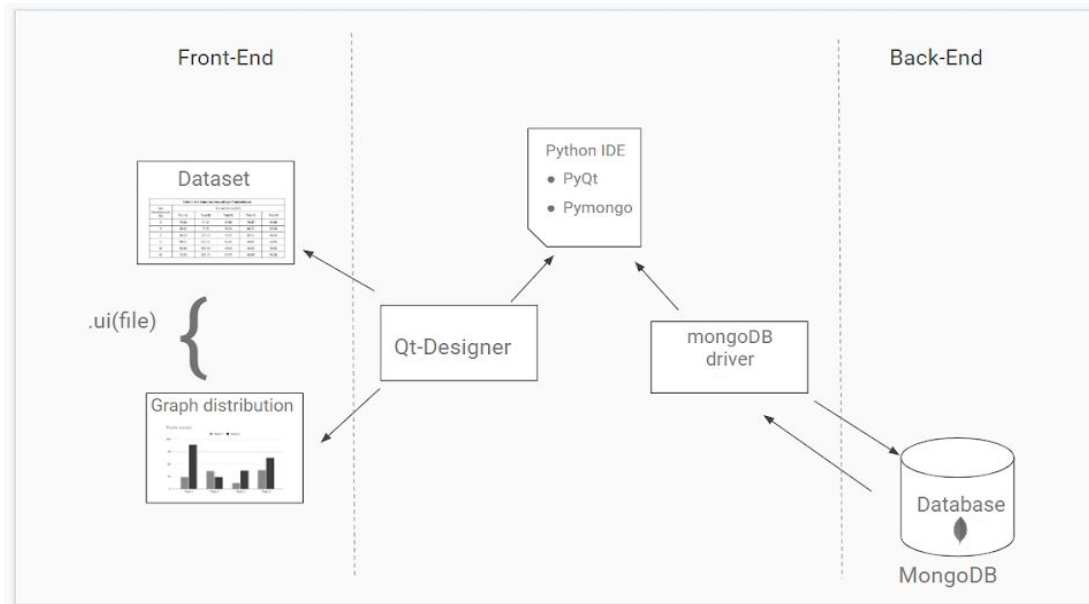
Another aspect to keep in mind is that providing a method of drilling down to comparatively simple statistics is crucial, in order for users to understand how the rolled up or aggregated results were generated.

Here we analyze the data on apps present in Google playstore and based on certain logic we process the data from above mentioned stages. The Play Store apps data has enormous potential to drive app-making businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market. The dataset is the web scraped data of Play Store apps for analyzing the Android market. It consists of a total of 4470 rows with 13 columns.

The dataset contains following columns -

1. App - Application name on Google Playstore
2. Category - General category to which the app belongs.
3. Rating - App rating ranges between 0 to 5.
4. Reviews - No. of user reviews.
5. Size - Size of the app in MB.
6. Installs - No.of Installs.
7. Type - App is capitalized or not.
8. Price - Based on Type charging amount of the particular app. (in \$)
9. Content Rating - Based on user age period access given by developer.
10. Genres - A more detailed category to which the application belongs.
11. Last Updated - Date on which the app was last updated.
12. Current Ver - Current version of the app.
13. Android Ver - Android version supported by the app.

3. DESIGN MODULES



Front-End- It is all about the visual aspects of the application that a user can see and experience. Here User can view data set and graphical representation of data set. The files known UI files are named with .ui extensions.

Python driver- Connects front-end and back-end.

Back-End-It is a 'database-side' programming, happens on the database. It's the machinery that works behind the scenes to power those features which users interact with on the client-side.

4. REQUIREMENTS AND FUNCTIONALITY

➤ Functional Requirements-

○ System Specifications-

- Windows(version 7 or above)/Linux OS
- 500Gb-1Tb HDD/SSD
- 4GB RAM or above

○ Tools Required-

- VS Code/Pycharm (or any required library supporting IDE)
- MongoDB
- Qt Designer

Google Play Store App Data Analytics provides the following productive accessibility on data manipulation -

- **Import / Export Dataset-** MongoDB supports import and export for both JSON and CSV files by exporting MongoDB information we can acquire a human readable text file with your data. By default, information is exported in json format but you can also export to csv (comma separated value).
- **Insert Document-** Any new details of an app can be inserted in the dataset with the help of inserting a document in the dataset. Here with the help of application form, users can perform the functionality.
- **Delete Document(s)-** Similar to insertion any new details of an app can be deleted in the dataset with the help of deleting a document in the dataset. Here with the help of application form, users can perform the deletion with the help of App name.
- **Update Document(s)-** Any details of an app can be updated in the dataset with the help of updating a document in the dataset.
- **Delete Collection-** Whole collection can be dropped at once by accessing the button.
- **Perform Analysis (using graphs)-** With the help of the dropdown menu any graph distribution can be loaded and analyzed . Following seven graphs can be analyzed and represents the dataset in a best way-
 1. Bar graph of distribution of apps across Category.
 2. Pie chart of free and paid apps.
 3. Pie chart of Content rating in each category.
 4. Bar graph of android version.
 5. Histogram of Rating.
 6. Boxplot of Installs.
 7. Ratings vs Size.
- **Recommendations (using Demographic Filtering)-**Demographic filtering is used to generate top 20 recommended Apps. Based on Rating and Reviews, a score is generated and using the score we can rank the records.

To perform the above mentioned functionalities following python libraries are used-

- **Pymongo** (MongoDB driver for Python)- PyMongo is a Python distribution containing tools for working with MongoDB, and is the recommended way to work with MongoDB from Python. This documentation attempts to explain everything you need to know to use PyMongo.
- **PyQt5**- PyQt is a Python binding of the cross-platform GUI toolkit Qt, implemented as a Python plug-in.
- **Pandas**- Pandas is a software library written for the Python programming language for data manipulation and analysis.
- **Numpy**-NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.
- **Matplotlib**- Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits.
- **MongoDB** (Backend NoSQL Database)-MongoDB is a cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with optional schemas.

5. SCREENSHOTS

Database-

MongoDB Compass - localhost:27017/mydb.google_playstore_apps

Connect View Collection Help

Local

8 DBS 20 COLLECTIONS

☆ FAVORITE

HOST: localhost:27017

CLUSTER: Standalone

EDITION: MongoDB 4.0.20-36-g8006ee Community

Q google|

mydb

google_playstore_apps

mydb.google_playstore_apps

DOCUMENTS 4.5k TOTAL SIZE 1.4MB AVG. SIZE 319B INDEXES 1 TOTAL SIZE 48.0KB AVG. SIZE 48.0KB

Documents Aggregations Schema Explain Plan Indexes Validation

FILTER

ADD DATA VIEW

Displaying documents 1 - 20 of 4470

Options FIND RESET

Refresh

```
{
  "_id": ObjectId("5f0d4dd8be9292e6b1a73471"),
  "App": "Photo Editor & Candy Camera & Grid & ScrapBook",
  "Category": "ART_AND_DESIGN",
  "Rating": 4.1,
  "Reviews": 159,
  "Size": "19M",
  "Installs": "10,000+",
  "Type": "Free",
  "Price": 0,
  "Content Rating": "Everyone",
  "Genres": "Art & Design",
  "Last Updated": "January 7, 2018",
  "Current Ver": "1.0.0",
  "Android Ver": "4.0.3 and up"
}
```

Dataset-

| googleplaystore | | | | | | | | | | | | | | Share | |
|--|---|---------------------|--------|---------|--------------------|-------------|------|--------|----------------|-------------------|--------------------|--------------------|--------------------|-------|--|
| File Edit View Insert Format Data Tools Add-ons Help Last edit was seconds ago | | | | | | | | | | | | | | | |
| Fossil Q | | | | | | | | | | | | | | | |
| | A | B | C | D | E | F | G | H | I | J | K | L | M | | |
| 1 | App | Category | Rating | Reviews | Size | Installs | Type | Price | Content Rating | Genres | Last Updated | Current Ver | Android Ver | | |
| 2 | Photo Editor & Candy Camera & Grid & ScrapBook | ART_AND_DESIGN | 4.1 | 159 | 19M | 10,000+ | Free | | 0 Everyone | Art & Design | January 7, 2018 | 1.0.0 | 4.0.3 and up | | |
| 3 | U Launcher Lite - FREE Live Cool Themes & Icons | ART_AND_DESIGN | 4.7 | 87510 | 8.7M | 5,000,000+ | Free | | 0 Everyone | Art & Design | August 1, 2018 | 1.2.4 | 4.0.3 and up | | |
| 4 | Sketch - Draw & Paint | ART_AND_DESIGN | 4.5 | 215044 | 25M | 50,000,000+ | Free | | 0 Teen | Art & Design | June 8, 2018 | Varies with device | 4.2 and up | | |
| 5 | Paper flowers instructions | ART_AND_DESIGN | 4.4 | 167 | 5.6M | 50,000+ | Free | | 0 Everyone | Art & Design | March 26, 2017 | | 1.2.3 and up | | |
| 6 | Smoke Effect Photo Maker - Smoke Editor | ART_AND_DESIGN | 3.8 | 178 | 19M | 50,000+ | Free | | 0 Everyone | Art & Design | April 26, 2018 | | 1.1 4.0.3 and up | | |
| 7 | Infinite Painter | ART_AND_DESIGN | 4.1 | 36815 | 29M | 1,000,000+ | Free | | 0 Everyone | Art & Design | June 14, 2018 | 6.1.61.1 | 3.0 and up | | |
| 8 | Garden Coloring Book | ART_AND_DESIGN | 4.4 | 13791 | 33M | 1,000,000+ | Free | | 0 Everyone | Art & Design | September 20, 2018 | 2.2.9.2 | 4.2 and up | | |
| 9 | Kids Paint Free - Drawing Fun | ART_AND_DESIGN | 4.7 | 121 | 3.1M | 10,000+ | Free | | 0 Everyone | Art & Design | July 3, 2018 | | 2.8 4.0.3 and up | | |
| 10 | Text on Photo - Fontee | ART_AND_DESIGN | 4.4 | 13880 | 28M | 1,000,000+ | Free | | 0 Everyone | Art & Design | October 27, 2017 | 1.0.4 | 4.1 and up | | |
| 11 | Name Art Photo Editor - Focus n Filters | ART_AND_DESIGN | 4.4 | 8788 | 12M | 1,000,000+ | Free | | 0 Everyone | Art & Design | July 31, 2018 | 1.0.15 | 4.0 and up | | |
| 12 | Tattoo Name On My Photo Editor | ART_AND_DESIGN | 4.2 | 44829 | 20M | 10,000,000+ | Free | | 0 Teen | Art & Design | April 2, 2018 | | 3.8 4.1 and up | | |
| 13 | Mandala Coloring Book | ART_AND_DESIGN | 4.6 | 4326 | 21M | 100,000+ | Free | | 0 Everyone | Art & Design | June 26, 2018 | 1.0.4 | 4.4 and up | | |
| 14 | 3D Color Pixel by Number - Sandbox Art & Design | ART_AND_DESIGN | 4.4 | 1518 | 37M | 100,000+ | Free | | 0 Everyone | Art & Design | August 3, 2018 | 1.2.3 | 2.3 and up | | |
| 15 | Learn To Draw Kawaii Characters | ART_AND_DESIGN | 3.2 | 55 | 2.7M | 5,000+ | Free | | 0 Everyone | Art & Design | June 8, 2018 | NaN | 4.2 and up | | |
| 16 | Photo Designer - Write your name with shapes | ART_AND_DESIGN | 4.7 | 3632 | 5.5M | 500,000+ | Free | | 0 Everyone | Art & Design | July 31, 2018 | | 3.1 4.1 and up | | |
| 17 | 350 Diy Room Decor Ideas | ART_AND_DESIGN | 4.5 | 27 | 17M | 10,000+ | Free | | 0 Everyone | Art & Design | November 7, 2017 | | 1.2.3 and up | | |
| 18 | FipaClip - Cartoon animation | ART_AND_DESIGN | 4.3 | 194216 | 39M | 5,000,000+ | Free | | 0 Everyone | Art & Design | August 3, 2018 | 2.2.5 | 4.0.3 and up | | |
| 19 | ibis Paint X | ART_AND_DESIGN | 4.6 | 224399 | 31M | 10,000,000+ | Free | | 0 Everyone | Art & Design | July 30, 2018 | 5.5.4 | 4.1 and up | | |
| 20 | Logo Maker - Small Business | ART_AND_DESIGN | 4 | 450 | 14M | 100,000+ | Free | | 0 Everyone | Art & Design | April 20, 2018 | | 4.1 and up | | |
| 21 | Boys Photo Editor - Six Pack & Men's Suit | ART_AND_DESIGN | 4.1 | 654 | 12M | 100,000+ | Free | | 0 Everyone | Art & Design | March 20, 2018 | | 1.1 4.0.3 and up | | |
| 22 | Superheroes Wallpapers 4K Backgrounds | ART_AND_DESIGN | 4.7 | 7699 | 4.2M | 500,000+ | Free | | 0 Everyone 10+ | Art & Design | July 12, 2018 | 2.2.6.2 | 4.0.3 and up | | |
| 23 | McQueen Coloring pages | ART_AND_DESIGN | NaN | 61 | 7.0M | 100,000+ | Free | | 0 Everyone | Art & Design | March 7, 2018 | 1.0.0 | 4.1 and up | | |
| 24 | HD Mickey Minnie Wallpapers | ART_AND_DESIGN | 4.7 | 118 | 23M | 50,000+ | Free | | 0 Everyone | Art & Design | July 7, 2018 | 1.1.3 | 4.1 and up | | |
| 25 | Harley Quinn wallpapers HD | ART_AND_DESIGN | 4.8 | 192 | 6.0M | 10,000+ | Free | | 0 Everyone | Art & Design | April 25, 2018 | | 1.5 3.0 and up | | |
| 26 | ColorIt - Drawing & Coloring | ART_AND_DESIGN | 4.7 | 20260 | 25M | 500,000+ | Free | | 0 Everyone | Art & Design | October 11, 2017 | 1.0.8 | 4.0.3 and up | | |
| 27 | Animated Photo Editor | ART_AND_DESIGN | 4.1 | 203 | 6.1M | 100,000+ | Free | | 0 Everyone | Art & Design | March 21, 2018 | | 1.03 4.0.3 and up | | |
| 4459 | myQ | TOOLS | 3.6 | 3642 | 29M | 100,000+ | Free | | 0 Everyone | Tools | July 23, 2018 | 3.2.0 | 4.0.3 and up | | |
| 4460 | QR Code Reader | TOOLS | 4 | 495971 | 2.7M | 50,000,000+ | Free | | 0 Everyone | Tools | March 16, 2016 | | 2.3 2.3.3 and up | | |
| 4461 | ICQ - Video Calls & Chat Messenger | COMMUNICATION | 4.4 | 697639 | Varies with device | 10,000,000+ | Free | | 0 Everyone | Communication | July 18, 2018 | 7.3(823181) | 4.4 and up | | |
| 4462 | Q Operator | BUSINESS | 4.6 | 60 | Varies with device | 1,000+ | Free | | 0 Everyone | Business | November 8, 2017 | 1.0.4 | 4.3 and up | | |
| 4463 | Q Call | TOOLS | NaN | 20 | 16M | 1,000+ | Free | | 0 Everyone | Tools | July 16, 2018 | | 2.1 4.0 and up | | |
| 4464 | Q-Slope | TOOLS | 4.5 | 2 | 3.9M | 50+ | Paid | \$1.79 | 0 Everyone | Tools | December 3, 2017 | | 5.4.3 and up | | |
| 4465 | Q-Tech Companion App | FOOD_AND_DRINK | 1.8 | 17 | 3.1M | 10,000+ | Free | | 0 Everyone | Food & Drink | March 8, 2017 | 1.6.3 | 4.4 and up | | |
| 4466 | Q Mobile Banking | FINANCE | 4.7 | 460 | 12M | 10,000+ | Free | | 0 Everyone | Finance | June 26, 2018 | 5.9.1.0 | 5.0 and up | | |
| 4467 | Q Quiz SK | FAMILY | 3.9 | 159 | 36M | 5,000+ | Free | | 0 Everyone | Education | December 8, 2017 | | 2.3 2.3 and up | | |
| 4468 | Bq Q Car Service | MAPS_AND_NAVIGATION | 4.4 | 75 | 25M | 5,000+ | Free | | 0 Everyone | Maps & Navigation | May 15, 2017 | | 3.002 4.0.3 and up | | |
| 4469 | Q-See QT View | TOOLS | 3.6 | 7357 | 19M | 100,000+ | Free | | 0 Everyone | Tools | July 12, 2018 | 4.4.1 | 5.0 and up | | |
| 4470 | DataCamp - Learn R, Python & SQL | FAMILY | 4.4 | 944 | 12M | 100,000+ | Free | | 0 Everyone | Education | August 1, 2018 | 3.1.0 | 4.1 and up | | |
| 4471 | R Language Reference Guide | BOOKS_AND_REFERENCE | 3.9 | 54 | 2.8M | 10,000+ | Free | | 0 Everyone | Books & Reference | December 24, 2017 | | 1.11 4.0 and up | | |

Code-

mongo_db.py

```
import pymongo
import pandas as pd

class Mongo_db:
    def __init__(self):
        # connecting to mongodb and getting the database and collection
        self.mongoClient = pymongo.MongoClient("mongodb://localhost:27017/")
        self.mydb = self.mongoClient["mydb"] # database name
        self.mycol = self.mydb["google_playstore_apps"] # collection name

    def getCollection(self):
        return self.mycol

    def insertDocument(self, doc):
        try:
            res = self.mycol.insert_one(doc)
            return res.inserted_id
        except Exception:
            return None

    def updateDocument(self, query, newVal, multiple=False):
        try:
            newValues = {"$set": newVal}
            res = self.mycol.update_one(
                query, newValues) if not multiple else self.mycol.update_many(query, newValues)
            return res.modified_count
        except Exception:
            return None

    def deleteDocument(self, query, multiple=False):
        try:
            res = self.mycol.delete_one(
                query) if not multiple else self.mycol.delete_many(query)

            return res.deleted_count
        except Exception:
            return None

    def searchData(self, query, projection, sortField="", sortOrder=0):
        # returns pandas dataframe of the result
        try:
            if sortField == "" or sortOrder == 0:
                res = self.mycol.find(query, projection)
                return pd.DataFrame(list(res))
            else:
                res = self.mycol.find(query, projection).sort(
                    sortField, sortOrder)
                return pd.DataFrame(list(res))
        except Exception:
            return None

    def deleteCollection(self):
        try:
            self.mycol.drop()
            return True
        except Exception:
            return False
```

graph.py

```
import pymongo
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from PyQt5 import QtCore, QtWidgets
```

```
def ratingHistogram(col, fig):
    ''' Histogram of App Ratings '''
    plt.clf()
    # mongodb query
    data = pd.DataFrame(
        list(col.find({}, {"_id": 0, "Rating": 1})))
    ax = fig.add_subplot(111)

    data.hist(bins=50, ax=ax, facecolor='purple', alpha=0.8) # histogram of ratings
```

```
def barGraphNumApps(col):
    ''' Bar graph of number of apps in each Category '''
    plt.clf() # clear the current figure
    # mongodb query
    x = list(col.aggregate(
        [{"$group": {"_id": "$Category", "number": {"$sum": 1}}}]))
    yval = []
    xval = []
    for doc in x:
        xval.append(doc['_id'])
        yval.append(doc['number'])

    # horizontal bar graph
    plt.barh(xval, yval, color=('deebskyblue',
                                'palegreen', 'orangered', 'gold'))
    plt.xlabel('No. of Apps')
    plt.ylabel("Categories")
    for i, v in enumerate(yval):
        plt.text(v + 3, i-0.25, str(v), color='black', fontsize=5)
    plt.yticks(rotation=0, fontsize=5)
    plt.title('Distribution of Apps across Categories')
```

```
def barGraphAndroidVer(col):
    ''' Bar graph of Android Version '''
    plt.clf() # clear the current figure
    # mongodb query
    x = list(col.aggregate(
        [{"$group": {"_id": "$Android Ver", "number": {"$sum": 1}}}]))
    yval = []
    xval = []
    for doc in x:
        if str(doc['_id']) != "nan":
            xval.append(str(doc['_id']))
            yval.append(doc['number'])

    # bargraph
    plt.barh(xval, yval, color=('slategrey', 'lightcoral', 'orangered', 'gold'))
    plt.xlabel('No. of Apps')
    plt.ylabel("Android Version")
    for i, v in enumerate(yval):
        plt.text(v + 3, i-0.25, str(v), color='black', fontsize=5)
    plt.yticks(rotation=0, fontsize=5)
    plt.title('Distribution of Android Versions')
```

```
def pieChartContentRating(col):
    """ Pie chart of Content Rating of apps as per Age """
    plt.clf() # clear the current figure
    x = list(col.aggregate(
        [{"group": {"_id": "$Content Rating", "number": {"$sum": 1}}}]))
    yval = []
    xval = []
    for doc in x:
        xval.append(doc['_id'])
        yval.append(doc['number'])

    colors = ['gold', 'palegreen', 'tomato', 'lightcoral', 'deepskyblue']

    plt.pie(yval, labels=xval, explode=[0.03]*len(yval), colors=colors, radius=0.6, startangle=290, shadow=False,
            autopct='%1.2f%%', rotatelabels=0)
    plt.title('Content Rating on Playstore')
    plt.axis('equal')
```

```
def pieChartFreePaidApps(col):
    """ Pie chart of free and paid apps """
    plt.clf() # clear the current figure
    x = list(col.aggregate(
        [{"group": {"_id": "$Type", "number": {"$sum": 1}}}]))
    yval = []
    xval = []
    for doc in x:
        xval.append(doc['_id'])
        yval.append(doc['number'])

    colors = ['lightgrey', 'lightcoral']

    plt.pie(yval, labels=xval, explode=[0.03]*len(yval), colors=colors, radius=0.6, startangle=20, shadow=False,
            autopct='%1.2f%%', rotatelabels=0)
    plt.title('Distribution of Free and Paid apps')
    plt.axis('equal')
```

```
def scatterPlotRatingSize(col):
    """ scatter plot of install vs size """
    plt.clf() # clear the current figure
    data = pd.DataFrame(
        list(col.find({}, {"_id": 0, "Rating": 1, "Size": 1})))
    # data.head()
    data["Rating"].dropna(inplace=True)
    # The size column has M which denotes Mb so we need to make it in numeric values too
    data["Size"].dropna(how='any', inplace=True)
    arr = data['Size'].values
    temp = []
    for i in range(len(arr)):
        if arr[i] == 'Varies with device':
            temp.append(21.51) # approximate value
        elif 'k' in arr[i]:
            arr[i] = arr[i].replace('k', '')
            arr[i] = float(arr[i])/1000
            temp.append(arr[i])
        elif 'G' in arr[i]:
            arr[i] = arr[i].replace('G', '')
            arr[i] = float(arr[i])*1000
            temp.append(arr[i])
        elif 'M' in arr[i]:
            arr[i] = arr[i].replace('M', '')
            arr[i] = float(arr[i])
            temp.append(arr[i])
    data['Size'] = temp
    # Now plot with matplotlib
    plt.title("Rating Vs. App Size")
    plt.xlabel("App Size")
    plt.ylabel("Rating")
    plt.scatter(data.Size, data.Rating, s=10, c=data.Rating, marker='*', cmap='plasma')
```

```
def boxPlotInstall(col, fig):
    ''' Box plot of installs, grouped according to type '''
    plt.clf() # clear the current figure
    data = pd.DataFrame(
        list(col.find({}, {"_id": 0, "Installs": 1, "Type": 1})))
    data.dropna(inplace=True)
    data["Installs"] = data["Installs"].str.replace(",", "")
    data["Installs"] = data["Installs"].str.replace("+", "")
    data["Installs"] = pd.to_numeric(data["Installs"])
    data = data[data!=0].dropna()
    data["Installs"] = np.log10(data["Installs"]) # converting to log 10 scale

    ax = fig.add_subplot(111)
    data.boxplot(by="Type", column=[
        "Installs"], ax=ax)
    plt.title("Boxplot of Installs(log-scale) grouped by Type")
    plt.suptitle("")
```

Output-

Google Play Store App Analytics

File Edit

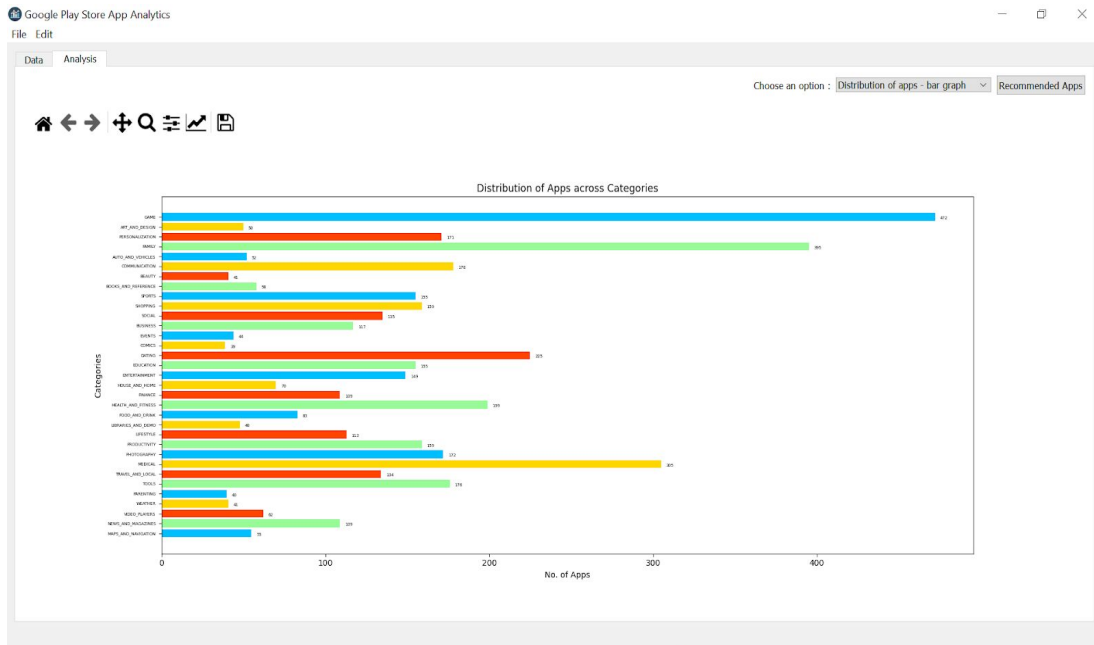
Data Analysis

Load Data Search App Name Search Advanced Search

| App | Category | Rating | Reviews | Size | Installs | Type | Price | Content Rating | Genres | Last Updated | Current Ver | |
|----------------------|-----------------|--------|---------|------|-------------|------|-------|----------------|--------------|-------------------|--------------------|-----|
| Photo Editor & ... | ART_AND_DESL... | 4.1 | 159 | 19M | 10,000+ | Free | 0 | Everyone | Art & Design | January 7, 2018 | 1.0.0 | 4.0 |
| U Launcher Lite ... | ART_AND_DESL... | 4.7 | 87510 | 8.7M | 5,000,000+ | Free | 0 | Everyone | Art & Design | August 1, 2018 | 1.2.4 | 4.0 |
| Sketch - Draw &... | ART_AND_DESL... | 4.5 | 215644 | 25M | 50,000,000+ | Free | 0 | Teen | Art & Design | June 8, 2018 | Varies with device | 4.2 |
| Paper flowers ... | ART_AND_DESL... | 4.4 | 167 | 5.6M | 50,000+ | Free | 0 | Everyone | Art & Design | March 26, 2017 | 1 | 2.3 |
| Smoke Effect ... | ART_AND_DESL... | 3.8 | 178 | 19M | 50,000+ | Free | 0 | Everyone | Art & Design | April 26, 2018 | 1.1 | 4.0 |
| Infinite Painter | ART_AND_DESL... | 4.1 | 36815 | 29M | 1,000,000+ | Free | 0 | Everyone | Art & Design | June 14, 2018 | 6.1.61.1 | 4.2 |
| Garden Colorin... | ART_AND_DESL... | 4.4 | 13791 | 33M | 1,000,000+ | Free | 0 | Everyone | Art & Design | September 20, ... | 2.9.2 | 3.0 |
| Kids Paint Free ~... | ART_AND_DESL... | 4.7 | 121 | 3.1M | 10,000+ | Free | 0 | Everyone | Art & Design | July 3, 2018 | 2.8 | 4.0 |
| Name Art Photo... | ART_AND_DESL... | 4.4 | 8788 | 12M | 1,000,000+ | Free | 0 | Everyone | Art & Design | July 31, 2018 | 1.0.15 | 4.0 |
| Tattoo Name O... | ART_AND_DESL... | 4.2 | 44829 | 20M | 10,000,000+ | Free | 0 | Teen | Art & Design | April 2, 2018 | 3.8 | 4.1 |
| Text on Photo ~ ... | ART_AND_DESL... | 4.4 | 13880 | 28M | 1,000,000+ | Free | 0 | Everyone | Art & Design | October 27, 2017 | 1.0.4 | 4.1 |
| Mandala Colori... | ART_AND_DESL... | 4.6 | 4326 | 21M | 100,000+ | Free | 0 | Everyone | Art & Design | June 26, 2018 | 1.0.4 | 4.4 |
| 3D Color Pixel b... | ART_AND_DESL... | 4.4 | 1518 | 37M | 100,000+ | Free | 0 | Everyone | Art & Design | August 3, 2018 | 1.2.3 | 2.3 |
| Learn To Draw ... | ART_AND_DESL... | 3.2 | 55 | 2.7M | 5,000+ | Free | 0 | Everyone | Art & Design | June 6, 2018 | nan | 4.2 |
| Photo Designer ... | ART_AND_DESL... | 4.7 | 3632 | 5.5M | 500,000+ | Free | 0 | Everyone | Art & Design | July 31, 2018 | 3.1 | 4.1 |
| 350 Diy Room ... | ART_AND_DESL... | 4.5 | 27 | 17M | 10,000+ | Free | 0 | Everyone | Art & Design | November 7, ... | 1 | 2.3 |
| FlipaClip ~ ... | ART_AND_DESL... | 4.3 | 194216 | 39M | 5,000,000+ | Free | 0 | Everyone | Art & Design | August 3, 2018 | 2.2.5 | 4.0 |

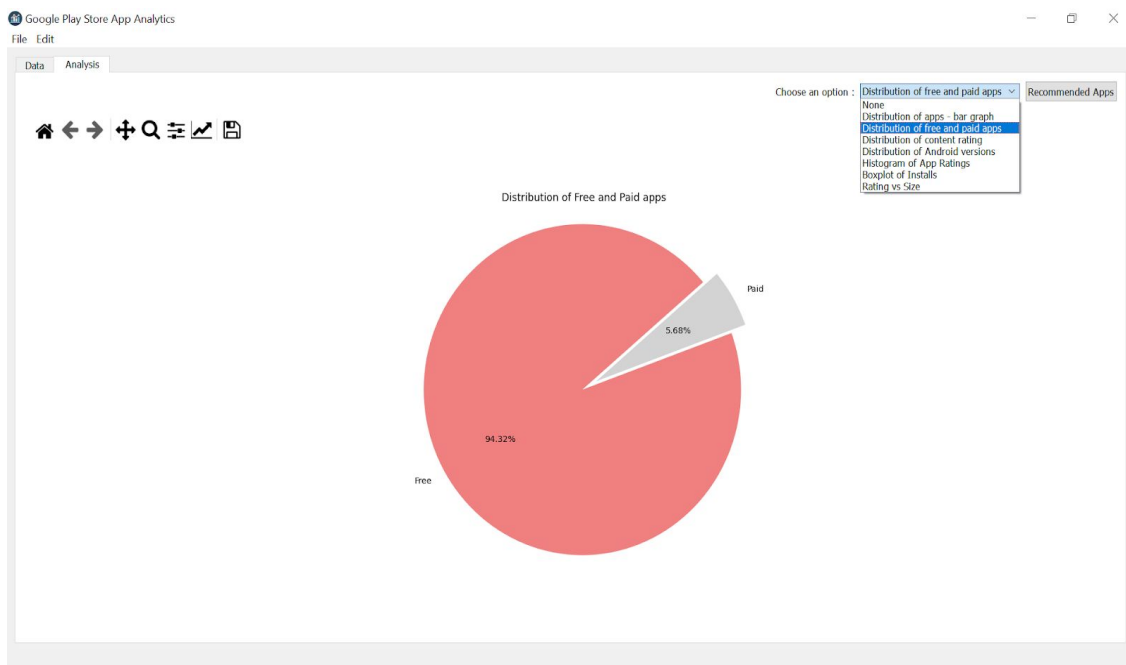
Graph Analysis -

a. Distribution of Apps across Categories



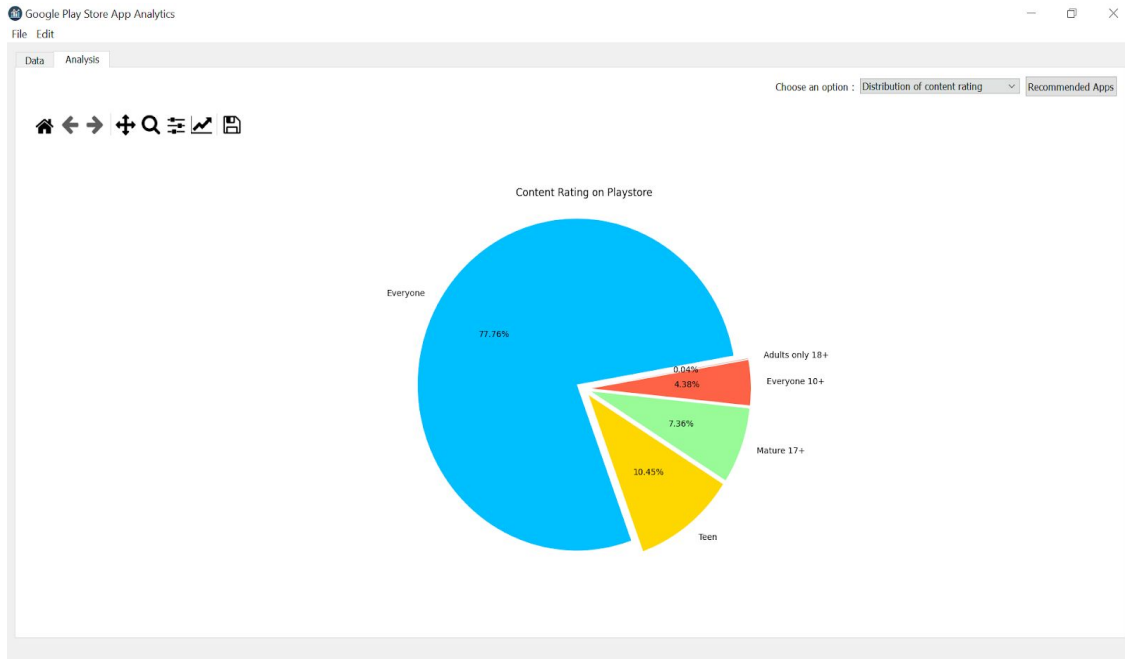
This graph shows that the categories with the maximum number of apps are Gaming, Family and Medical. This indicates that Gaming, Family and Medical are popular categories.

b. Distribution of Free and Paid apps



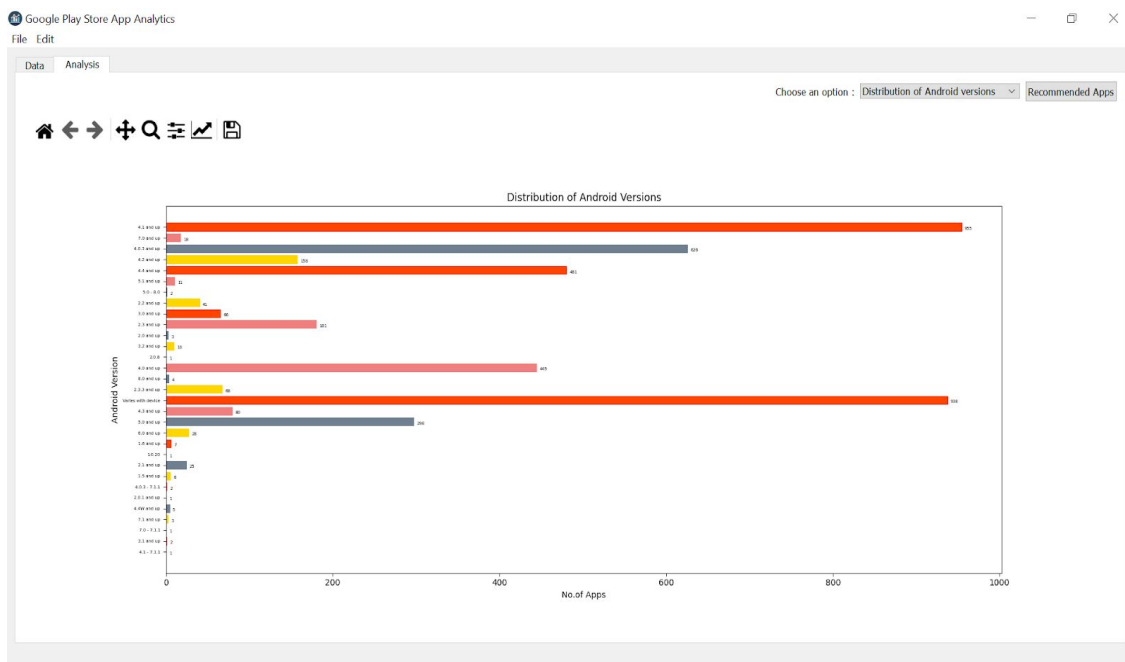
This graph shows that the distribution of Free apps is much more than the distribution of Paid apps, which is an indicator of the higher popularity of free apps compared to paid apps.

c. Distribution of Content Rating



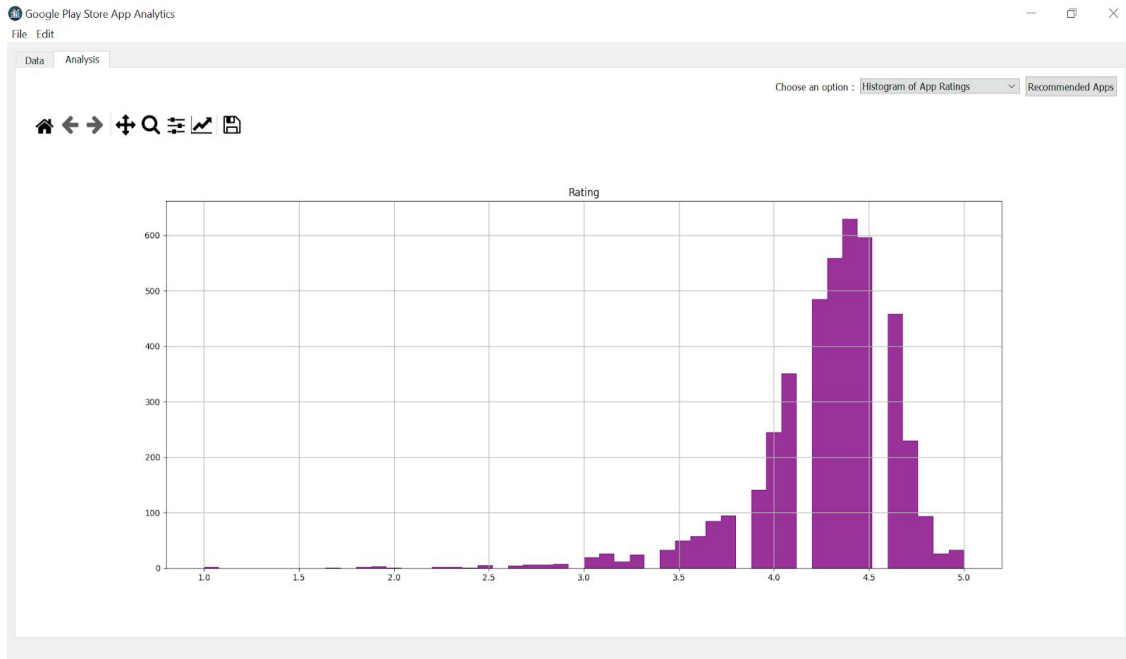
This pie chart shows that the majority of the apps provide content that is suitable for everyone, appealing to the majority of the audience. Also, very few apps provide content that is suitable only for adults.

d. Distribution of Android Versions



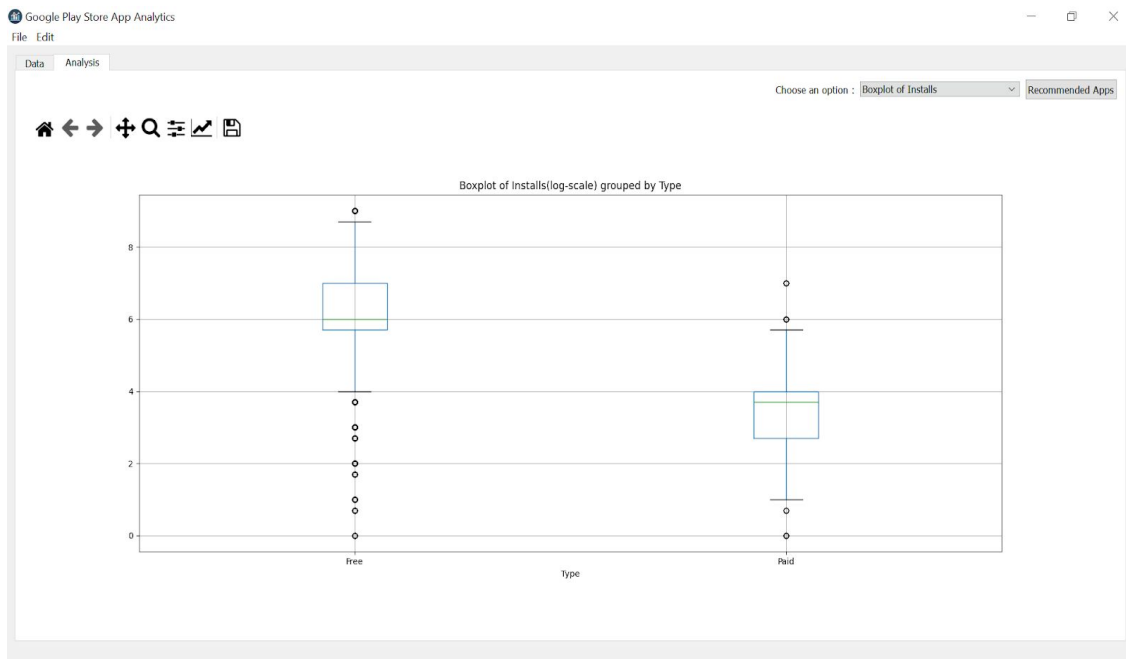
This bar graph shows that the majority of the apps support Android 4.1 and above. Also, the Android version varies for a significant number of apps.

e. Histogram of App Ratings



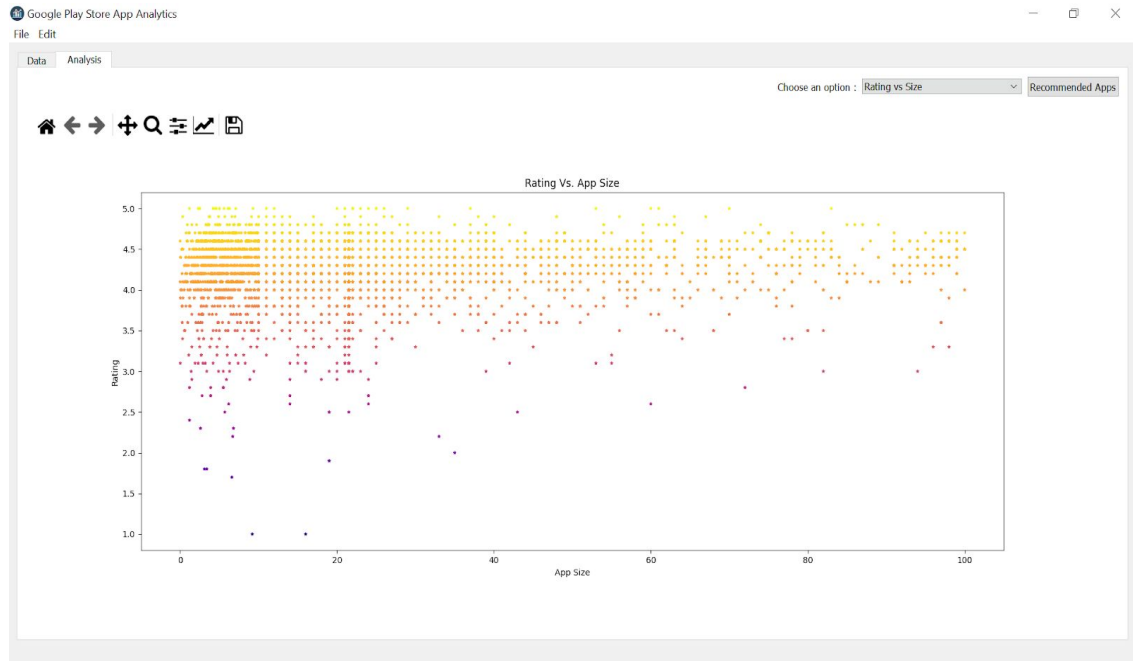
This histogram indicates that the majority of the apps have a rating above 4.0, showing that most of the apps provide quality content to the users.

f. Boxplots of Installs



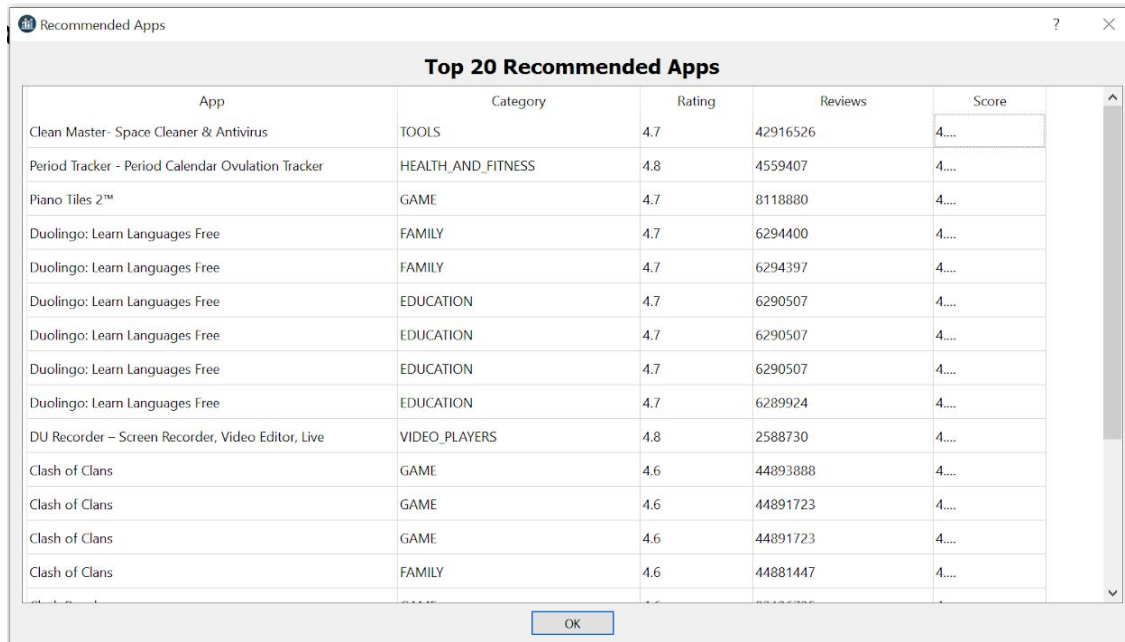
The boxplot of App installs for free and paid apps shows that the number of installs of Free apps is higher than that of Paid apps. This means that users tend to download free apps more than paid apps.

g. Scatterplot of Rating vs Size



The scatterplot of App Rating vs App size shows that most of the high rated apps do not consume a lot of space, i.e., are light-weighted apps. Though bulky apps are also highly rated, the frequency is very less compared to light-weighted apps. Apps that have a size of less than 40 MB are considered light-weighted.

App Recommendations -



| App | Category | Rating | Reviews | Score |
|--|--------------------|--------|----------|-------|
| Clean Master- Space Cleaner & Antivirus | TOOLS | 4.7 | 42916526 | 4.... |
| Period Tracker - Period Calendar Ovulation Tracker | HEALTH_AND_FITNESS | 4.8 | 4559407 | 4.... |
| Piano Tiles 2™ | GAME | 4.7 | 8118880 | 4.... |
| Duolingo: Learn Languages Free | FAMILY | 4.7 | 6294400 | 4.... |
| Duolingo: Learn Languages Free | FAMILY | 4.7 | 6294397 | 4.... |
| Duolingo: Learn Languages Free | EDUCATION | 4.7 | 6290507 | 4.... |
| Duolingo: Learn Languages Free | EDUCATION | 4.7 | 6290507 | 4.... |
| Duolingo: Learn Languages Free | EDUCATION | 4.7 | 6290507 | 4.... |
| Duolingo: Learn Languages Free | EDUCATION | 4.7 | 6289924 | 4.... |
| DU Recorder – Screen Recorder, Video Editor, Live | VIDEO_PLAYERS | 4.8 | 2588730 | 4.... |
| Clash of Clans | GAME | 4.6 | 44893888 | 4.... |
| Clash of Clans | GAME | 4.6 | 44891723 | 4.... |
| Clash of Clans | GAME | 4.6 | 44891723 | 4.... |
| Clash of Clans | FAMILY | 4.6 | 44881447 | 4.... |

Using Demographic Filtering, the top 20 best apps are recommended. The recommendation is based on a score calculated as follows :

$$Score = \left(\frac{v}{v+m} \right) R + \left(\frac{m}{v+m} \right) C$$

Here,

- v is the number of votes for the app;
- m is the minimum votes required to be listed in the chart;
- R is the average rating of the app; And
- C is the mean vote across the whole report

6. NEW LEARNINGS FROM THE PROGRAMMING ASSIGNMENT

Data extraction and manipulation on huge amount of data has added to the experience of the learning journey. The application of big data analysis, treats the ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.

Analysis of big data allows analysts, researchers and business users to make better and faster decisions using data that was previously inaccessible or unusable. Businesses can use advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, statistics and natural language processing to gain new insights from previously untapped data sources independently or together with existing enterprise data.

7. FUTURE ENHANCEMENTS

- a. If volume of data increases then migrating the data set to the cloud can bring an improvement in Quantity analysis.
- b. More prediction increases more competition in best methods and can result in better forecasting.
- c. Introducing the method to more new technologies and better functional dependent libraries can be an advancement.