2022-2023

VISVESVARAYA TECHNOLOGICAL UNIVERSITY

“Jnana Sangama”, BELAGAVI – 590018,

Karnataka,India

INTERNSHIP REPORT
ON
“DATA ANALYSIS WITH PYTHON”

Submitted in partial fulfilment of the requirements for the award of degree

BACHELOR OF ENGINEERING

in

ELECTRONICS AND COMMUNICATION ENGINEERING

Submitted by:

NAME : TEJAS M S

USN : 1BC21EC008

Conducted at
Cranes Varsity

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

BANGALORE COLLEGE OF ENGINEERING & TECHNOLOGY

**Opp, Heelalige Railway Station,
Chandapura Hosur Main Road,
Bengaluru-560099**

BANGALORE COLLEGE OF ENGINEERING & TECHNOLOGY

Chandapura, Bengaluru – 560099

Department of Electronics & Communication Engineering

CERTIFICATE

This is to certify that the internship titled "**DATA ANALYSIS WITH PYTHON** " , carried out by, **TEJAS.M.S**, bearing USN : 1BC21EC008 a bonifide student of Bangalore College of Engineering and Technology, in partial fulfillment for the award of Bachelor of Engineering in Electronics and Communication Engineering of the Visvesvaraya Technological University, Belagavi during the year 2022–2023. It is certified that all the corrections/suggestions indicated for Internal Assessment have been incorporated in the Report deposited in the departmental library. The project report has been approved as it satisfies the academic requirements.

Signature of the guide

Mrs. Karthika V
Department of ECE
BCET, Bangalore

Signature of the HOD

Dr. John Clement Sunder
Professor and Head
BCET, Bangalore

Signature of the Principal

Dr.Channankaiah
Principal
BCET, Bangalore

BANGALORE COLLEGE OF ENGINEERING & TECHNOLOGY

Chandapura, Bengaluru - 560099

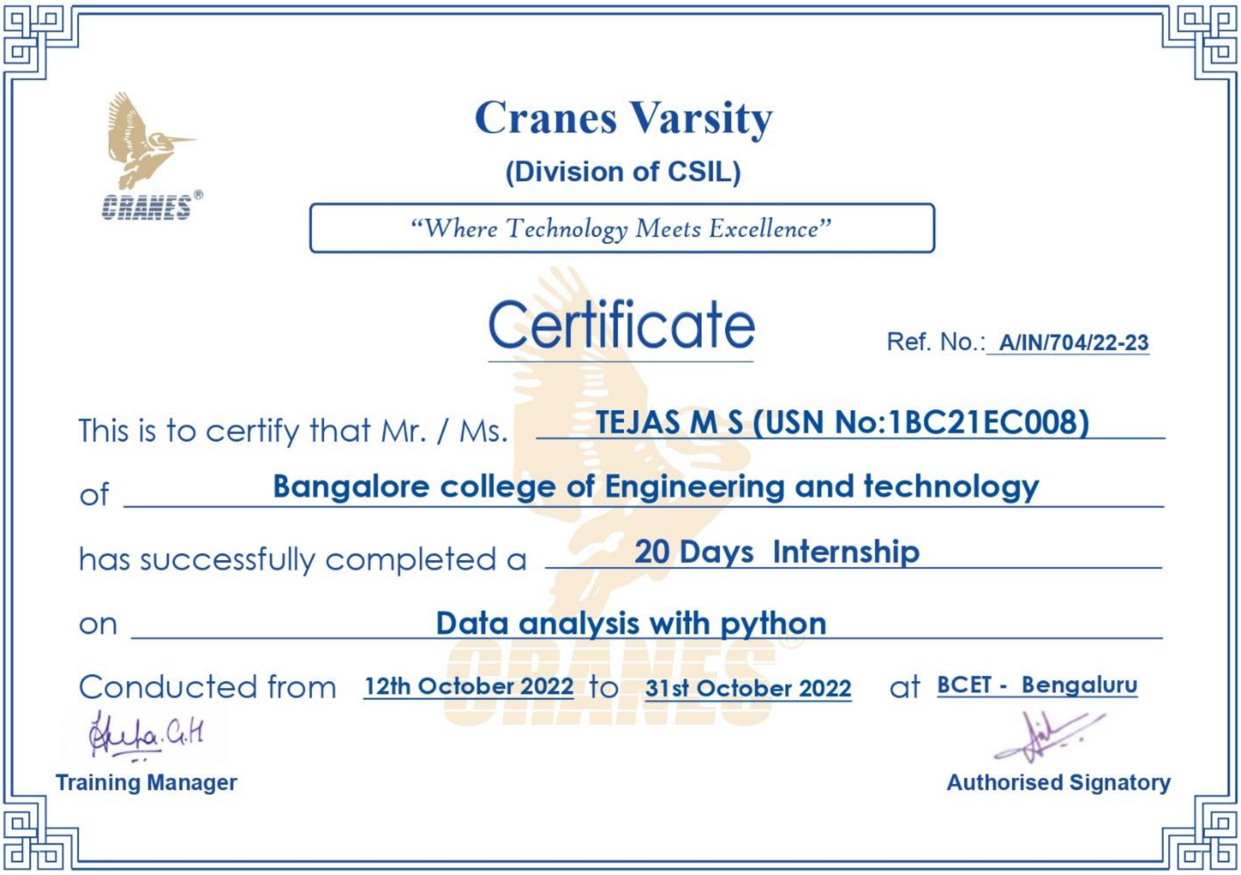
Department of Electronics & Communication Engineering

DECLARATION

I, **TEJAS M S** , first year student of ECE, Bangalore College of engineering and technology- 560 099, hereby declare that, the Internship has been successfully completed, in **Cranes Varsity**. This report is submitted in partial fulfillment of the requirements for award of Bachelor Degree in ECE, during the academic year 2022-2023.

Place : Bangalore
Date : 17/11/2022
USN : 1BC21EC008
Name of the student : TEJAS M S

CERTIFICATE PROVIDED BY THE COMPANY



Cranes Varsity

(Division of CSIL)

CRANES®

"Where Technology Meets Excellence"

Certificate

Ref. No.: A/IN/704/22-23

This is to certify that Mr. / Ms. TEJAS M S (USN No:1BC21EC008)

of Bangalore college of Engineering and technology

has successfully completed a 20 Days Internship

on Data analysis with python

Conducted from 12th October 2022 to 31st October 2022 at BCET - Bengaluru

Shila.G.H

Training Manager

Authorised Signatory

ACKNOWLEDGEMENT

This Internship is a result of accumulated guidance, direction and support of several important persons. We take this opportunity to express our gratitude to all who have helped us to complete the Internship.

We express our sincere thanks to our Principal, for providing us adequate facilities to undertake this Internship.

We would like to thank our Head of Dept – branch code, for providing us an opportunity to carry out Internship and for his valuable guidance and support.

We would like to thank our (Lab assistant name) Software Services for guiding us during the period of internship.
We express our deep and profound gratitude to our guide, Guide name, Assistant/ Associate Prof, for her keen interest and encouragement at every step in completing the Internship.

We would like to thank all the faculty members of our department for the support extended during the course of Internship.

We would like to thank the non-teaching members of our dept, for helping us during the Internship.

Last but not the least, we would like to thank our parents and friends without whose constant help, the completion of Internship would have not been possible.

NAME : TEJAS M S
USN : 1BC21EC008

TABLE OF CONTENTS:

1. INTRODUCTION OF COMPANY:

CRANES VARSITY: A Division of Cranes Software International Ltd.

Head Quartered: Bangalore, India.

Establishment: In 1998

COMPANY PROFILE

Cranes Varsity has over decade long relationship in Indian education industry as a pioneer in introducing embedded systems and digital signal processing education services to the

Indian market through the late 90’s much before the world recognized the Indian education market size to be one of $40 Billions.

Driven by a passion of embedded systems and the digital signal processing and shortage in skilled man power, in 1988 cranes varsity introduced hands on technical training to the graduates and the working professionals in the fields of mathematical modeling & simulation.

Cranes varsity is best known and credited in the Indian education sector seeding and nurturing of the legendary technical computing tools like MatlabTM and Texas Instruments, DSP’s for the Indian engineering sector through the 90’s.

“Bridging the chasm between the Engineer and the Industry”

Cranes is Authorized Partner to: Texas Instruments, National Instruments, IBM etc.

1.1 ACHIEVEMENTS :

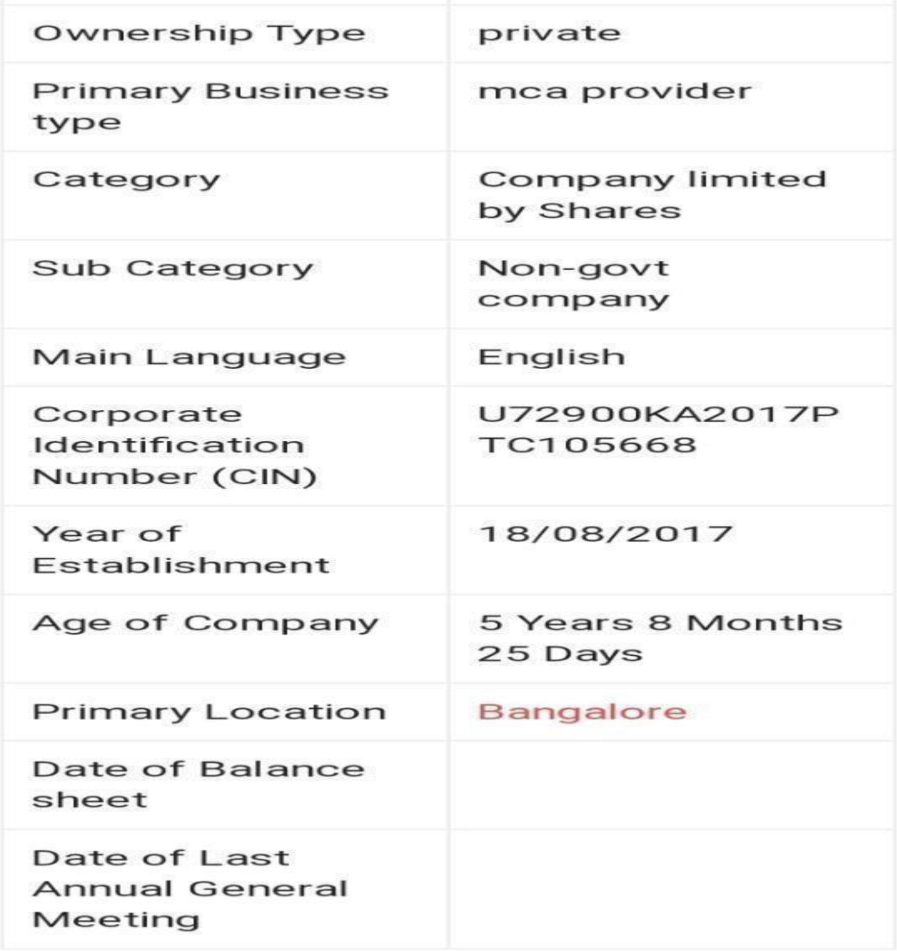
- Cranes Varsity has trained over 50,000 graduates directly through its education model. Over 4.5 million engineering graduates indirectly trained through its university.
- Employment in various Indian institute and Multinational Corporation is an achievement that is deeply valued and cherished.

2. ABOUT THE COMPANY:

Cranes Varsity is a pioneer Technical Training institute turned EdTech Platform offering Technology educational services for over 24 years. A division of Cranes Software International Ltd, Cranes Varsity was established with an ambitious vision of bridging the gap between the technology academia and the industry. The team continuously strives to be an organization that brings together technology and education, empowering aspiring professionals to seek assured placements and a lucrative career path. Cranes Varsity offers high-impact hands-on technology training that catapults engineering students, graduates, and working professionals to be quickly employable in Niche high-end engineering fields. The inhouse placement team further ensures that these students get placed in leading corporate firms – with whom Cranes Varsity has decades-old relationship. Cranes Varsity carries a legacy of being the Authorized-training partner for Texas Instruments, Math Works, Wind River & ARM. Cranes Varsity also has the honor of being a trusted partner of over 5000 reputed Academia, Corporate & Defense Organizations. Cranes Varsity has training leadership in EMBEDDED, MATLAB & DSP, extending training domains to emerging industry trends like Automotive, IoT, VLSI, Java full-stack, Data Science, Business Analytics and Software Programming.

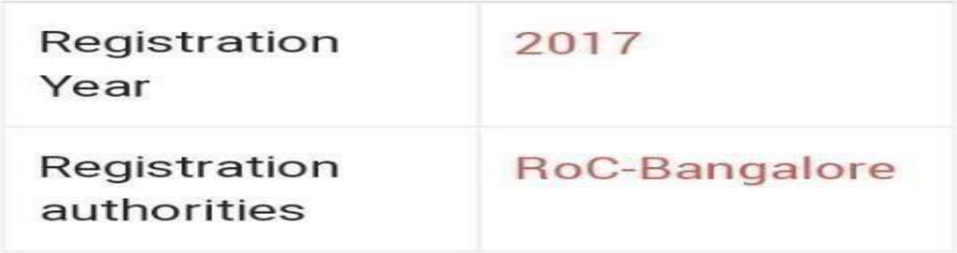
3. Cranes Varsity Private Limited Overview

General Details:

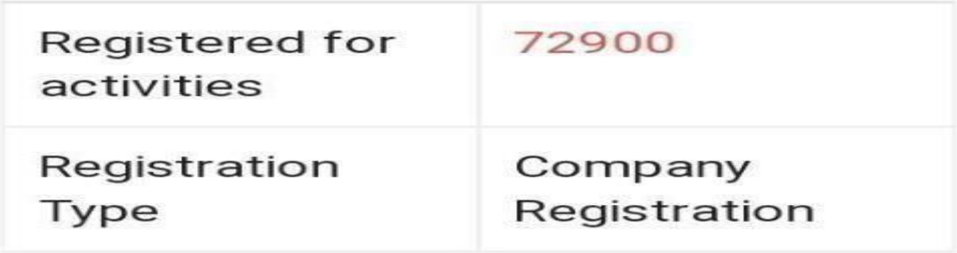


Ownership Type	private
Primary Business type	mca provider
Category	Company limited by Shares
Sub Category	Non-govt company
Main Language	English
Corporate Identification Number (CIN)	U72900KA2017PTC105668
Year of Establishment	18/08/2017
Age of Company	5 Years 8 Months 25 Days
Primary Location	Bangalore
Date of Balance sheet	
Date of Last Annual General Meeting	

4. Registration Details:



Registration Year	2017
Registration authorities	RoC-Bangalore



Registered for activities	72900
Registration Type	Company Registration

Services provided by Cranes Varsity:

Placement Oriented Training Program:

The objective of the Placement Oriented Training Program is to enhance the student's skills in the core areas, make them industry-ready, and get placement opportunities in their core domain. Our training expertise is in the core domains like Embedded, Automotive, IOT, VLSI, Artificial Intelligence, Machine Learning, Data Science, and Machine Learning.

The purpose of POP is to guide students to choose the right career and enhance core subject knowledge & Technical skills to meet the employment opportunities in the current Industry.

5. INTRODUCTION:

Data are those raw facts and figures with no proper information hence need to be processed to get the desired information. While information is those results which we get after processing the raw data in different levels or extracted conclusions from a given dataset through a process called data analysis.

Data Analysis is simply the analysis of various data means cleaning the data, transforming it into understandable form, and then modeling data to extract some useful information forb use or an organizational use. It is mainly used in taking business decisions. Many libraries are available for doing the analysis. For example, NumPy, Pandas, Seaborn, Matplotlib, Sklearn, etc.

- NumPy: NumPy is a library written in Python, used For numerical analysis in Python. It stores the data in The form of nd-arrays (n-dimensional arrays).

- Pandas: Pandas is mainly used for converting data into Tabular form and hence, makes the data more Structured and easily to read.

- Matplotlib: Matplotlib is a data visualisation and Graphical plotting package for Python and its Numerical extension NumPy that runs on all platforms.

- Seaborn: Seaborn is a Python data visualisation Package based on matplotlib that is tightly connected With pandas data structures. The core component of Seaborn is visualisation, which aids in data exploration and comprehension.

- Sklearn: Scikit-learn is the most useful library for machine learning in Python. It includes numerous useful tools for classification, regression, clustering, and dimensionality reduction.

Data visualization will help the data analysis to make it more Understandable and interactive by plotting or displaying the Data in pictorial form. Pandas, a Python open-source package that deals with three different data structures: series, data Frames, and panels, solves that need of analyzing and Visualization of data.

Data analysis using Python makes task easier since Python Programming language has many advantages over any other Programming language. It has prominent features like being a High-level programming language (the codes are in human Readable form) it is easy to understand and use by any Programmer or user. Many libraries and functions for statistical, Numerical analysis are available in Python. Moreover, the Source code is freely available to anyone (free and open source).

There are primarily five steps involved in the data analytics process, which include:

Data Collection: The first step in data analytics is to collect or gather relevant data from multiple sources. Data can come from different databases, web servers, log files, social media, excel and CSV files, etc.

Data Preparation: The next step in the process is to prepare the data. It involves cleaning the data to remove unwanted and redundant values, converting it into the right format, and making it ready for analysis. It also requires data wrangling.

Data Exploration: After the data is ready, data exploration is done using various data visualization techniques to find unseen trends from the data.

Data Modeling: The next step is to build your predictive models using machine learning algorithms to make future predictions.

Result interpretation: The final step in any data analytics process is to derive meaningful results and check if the output is in line with your expected results.

Why Data Analytics Using Python?

There are many programming languages available, but Python is popularly used by statisticians, engineers, and scientists to perform data analytics.

Here are some of the reasons why Data Analytics using Python has become popular:

- Python is easy to learn and understand and has a simple syntax.
- The programming language is scalable and flexible.
- It has a vast collection of libraries for numerical computation and data manipulation.
- Python provides libraries for graphics and data visualization to build plots.
- It has broad community support to help solve many kinds of queries.

5.1 OBJECTIVES OF ANALYSIS :

Objectives of Google-Play Store Analysis:

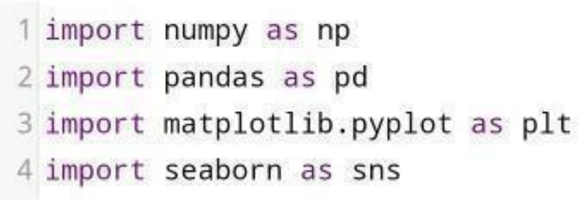
This project focuses on the analysis of the Play Store data set in Kaggle.

The aim of this project is:

1. Using the data to analyze consumer trends and determine which type of apps are the most popular and profitable.
2. Classifying applications based on their categories.
3. Presenting the growth of applications from 2016 to 2018.
4. Comparing different categories of applications based on the Android version.
5. Comparing the rates in different kinds of applications.
6. Assessing supported Android version with numbers of reviews based on different categories.

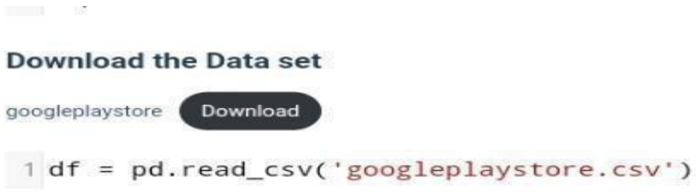
5.2 IMPLEMENTATION OF DATA ANALYSIS:

Lets start with importing the libraries



```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

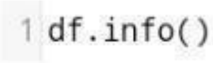
Linking the file path to the code



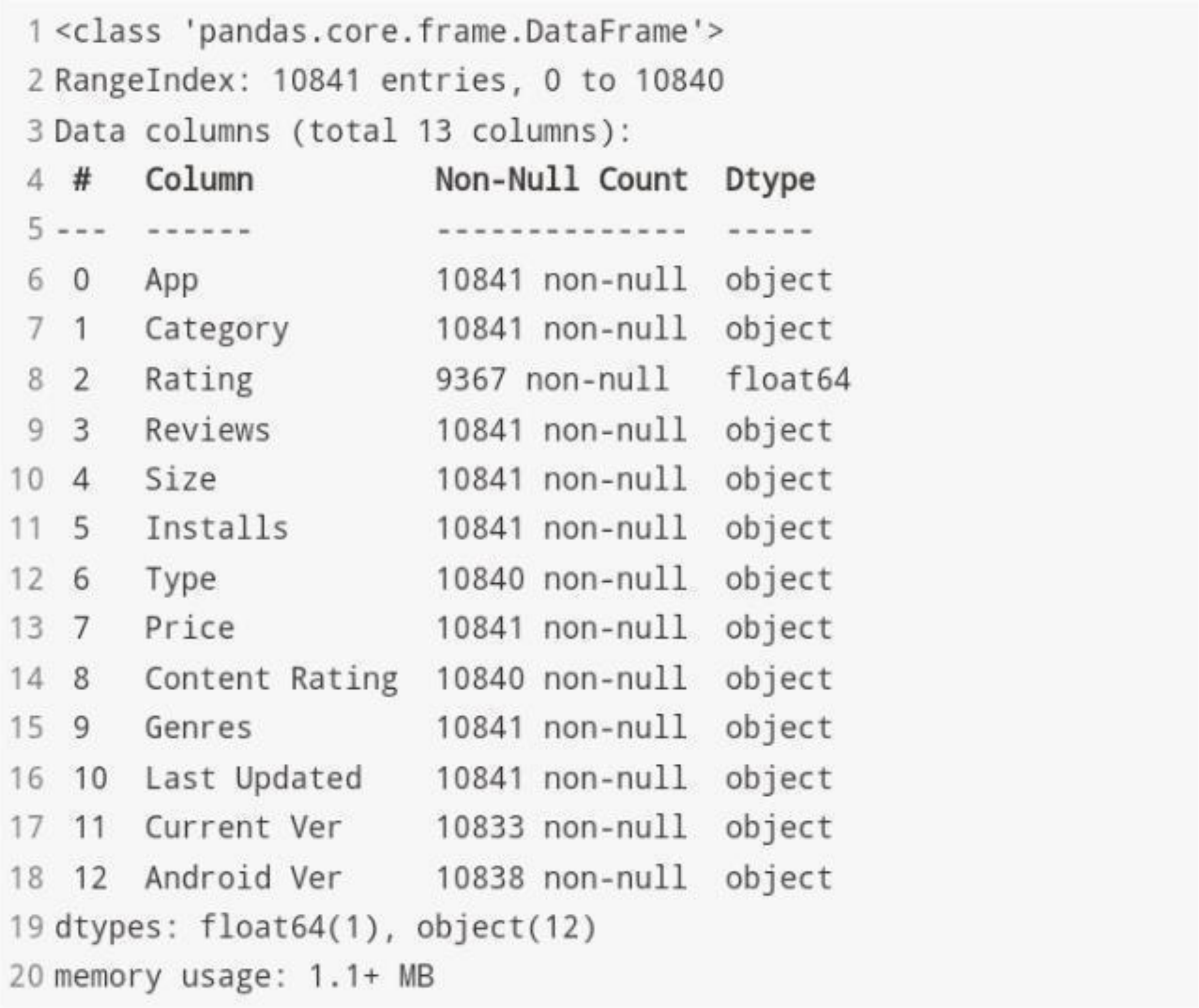
Download the Data set

```
df = pd.read_csv('googleplaystore.csv')
```

Lets see at some insights of the data



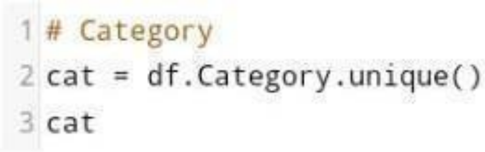
```
df.info()
```



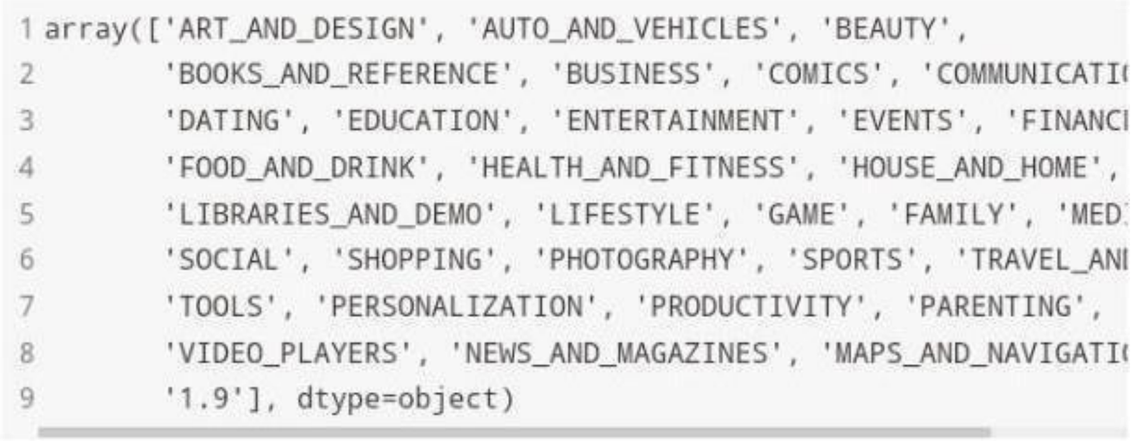
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10841 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   App             10841 non-null  object
 1   Category        10841 non-null  object
 2   Rating          9367 non-null   float64
 3   Reviews         10841 non-null  object
 4   Size            10841 non-null  object
 5   Installs        10841 non-null  object
 6   Type            10840 non-null  object
 7   Price           10841 non-null  object
 8   Content Rating  10840 non-null  object
 9   Genres          10841 non-null  object
 10  Last Updated    10841 non-null  object
 11  Current Ver     10833 non-null  object
 12  Android Ver     10838 non-null  object
dtypes: float64(1), object(12)
memory usage: 1.1+ MB
```

5.3 EXPLORATORY DATA ANALYSIS ON GOOGLE PLAY STORE :

Let’s take a look on all the category

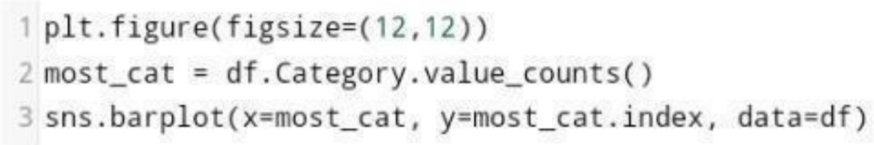


```
# Category
cat = df.Category.unique()
cat
```



```
array(['ART_AND_DESIGN', 'AUTO_AND_VEHICLES', 'BEAUTY',
       'BOOKS_AND_REFERENCE', 'BUSINESS', 'COMICS', 'COMMUNICATI
       'DATING', 'EDUCATION', 'ENTERTAINMENT', 'EVENTS', 'FINANC
       'FOOD_AND_DRINK', 'HEALTH_AND_FITNESS', 'HOUSE_AND_HOME',
       'LIBRARIES_AND_DEMO', 'LIFESTYLE', 'GAME', 'FAMILY', 'MED
       'SOCIAL', 'SHOPPING', 'PHOTOGRAPHY', 'SPORTS', 'TRAVEL_AN
       'TOOLS', 'PERSONALIZATION', 'PRODUCTIVITY', 'PARENTING',
       'VIDEO_PLAYERS', 'NEWS_AND_MAGAZINES', 'MAPS_AND_NAVIGATI
       '1.9'], dtype=object)
```

So we got 34 category on this data set, let’s see which one is the famous category



```
plt.figure(figsize=(12,12))
most_cat = df.Category.value_counts()
sns.barplot(x=most_cat, y=most_cat.index, data=df)
```

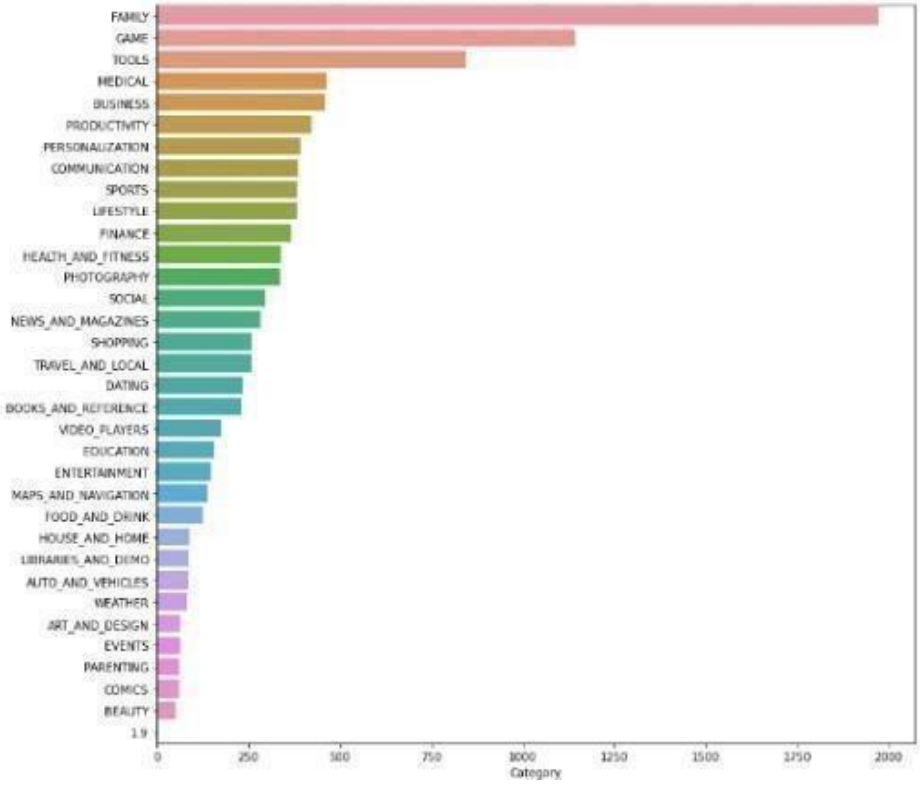
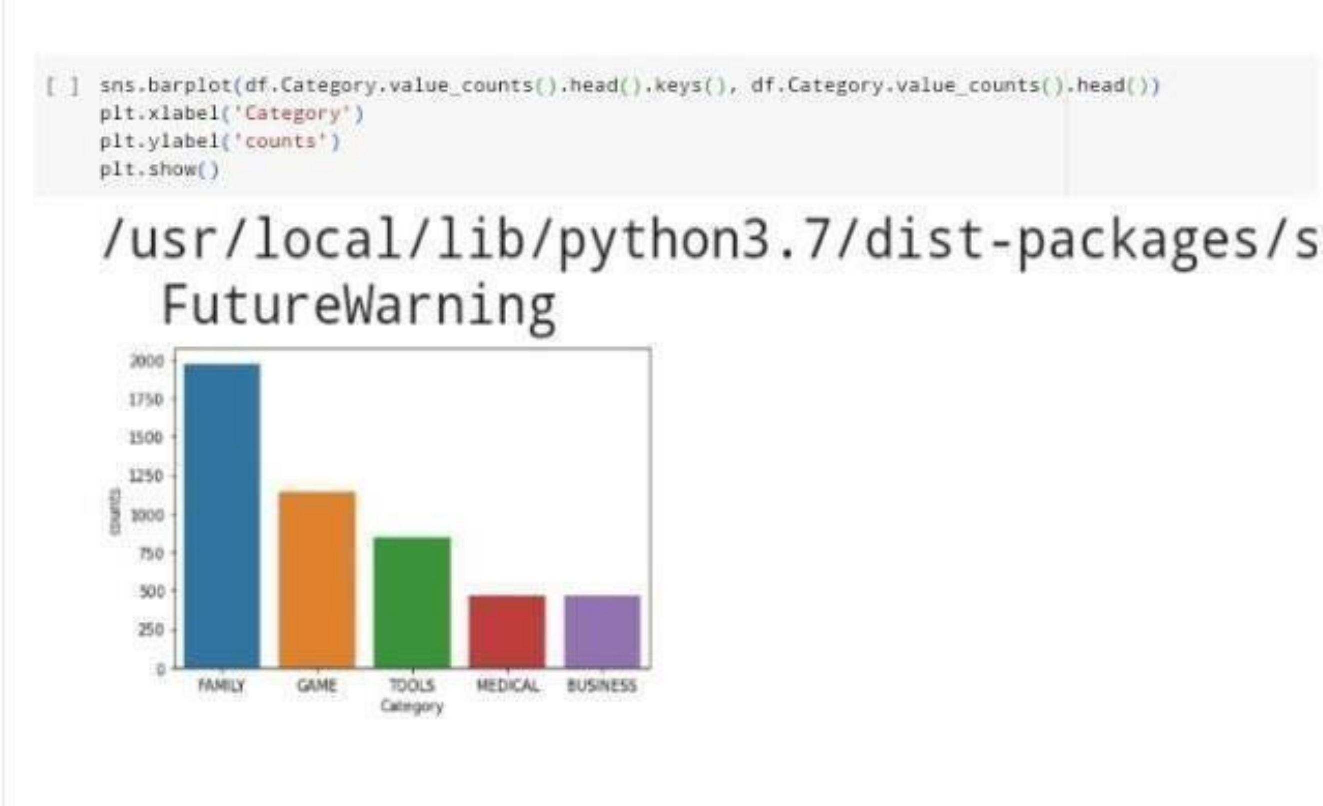


Fig 1.

So, there is around 2000 app with family category, followed by game category with 1200 app. And this ‘1.9’ Category, I don’t know what it is, but it only had 1 app so far, so its not visible on the graph.

Top 5 category of apps released in Playstore.

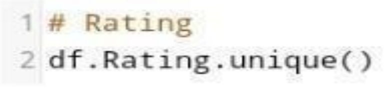


```
[ ] sns.barplot(df.Category.value_counts().head().keys(), df.Category.value_counts().head())
    plt.xlabel('Category')
    plt.ylabel('counts')
    plt.show()
```

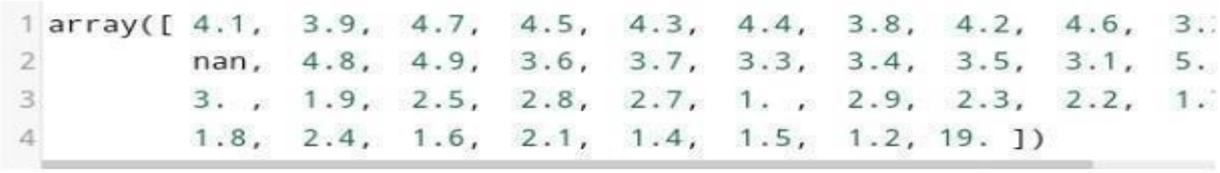
```
/usr/local/lib/python3.7/dist-packages/s
  FutureWarning
```

Fig 2.

Let’s look at the rating, and what kind of correlation share between category and rating.

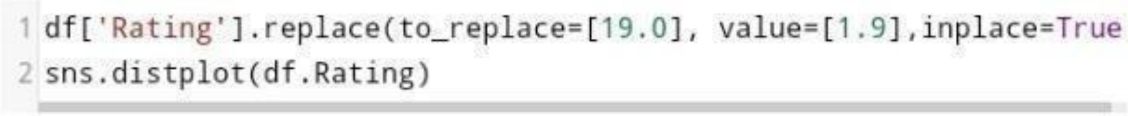


```
# Rating
df.Rating.unique()
```



```
array([ 4.1,  3.9,  4.7,  4.5,  4.3,  4.4,  3.8,  4.2,  4.6,  3.
        nan,  4.8,  4.9,  3.6,  3.7,  3.3,  3.4,  3.5,  3.1,  5.
        3. ,  1.9,  2.5,  2.8,  2.7,  1. ,  2.9,  2.3,  2.2,  1.
        1.8,  2.4,  1.6,  2.1,  1.4,  1.5,  1.2, 19. ])
```

There we had a null values, I am going to leave it as it is. And a 19 for rating is not possible, so I assume it’s a ‘1.9’. So let’s change it and see the distribution value on rating column.



```
df['Rating'].replace(to_replace=[19.0], value=[1.9],inplace=True
sns.distplot(df.Rating)
```

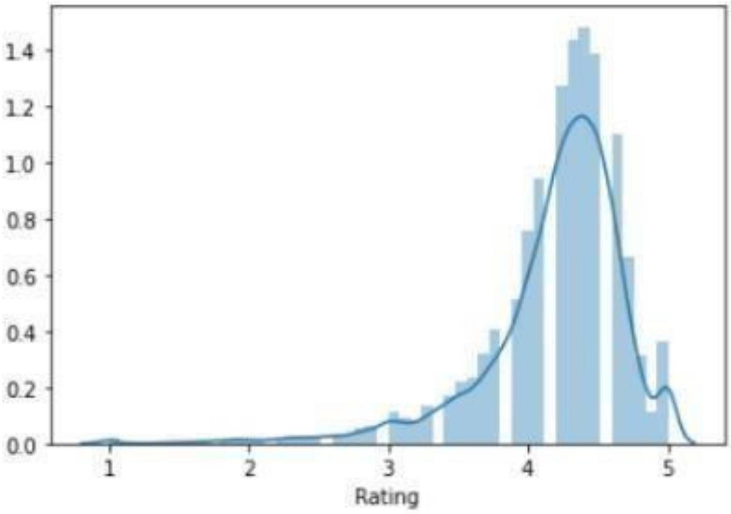
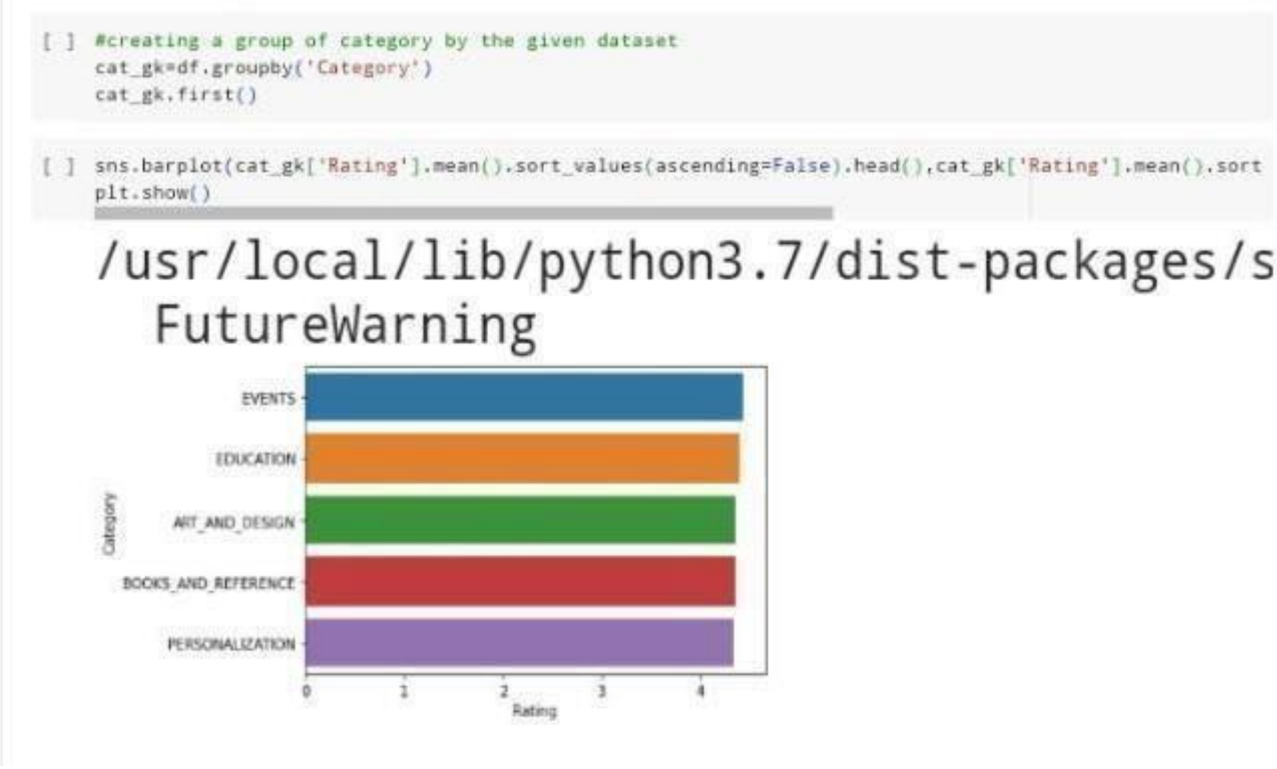


Fig 3.

Most of the rating is around 4. Let’s see how rating is distributed by category column



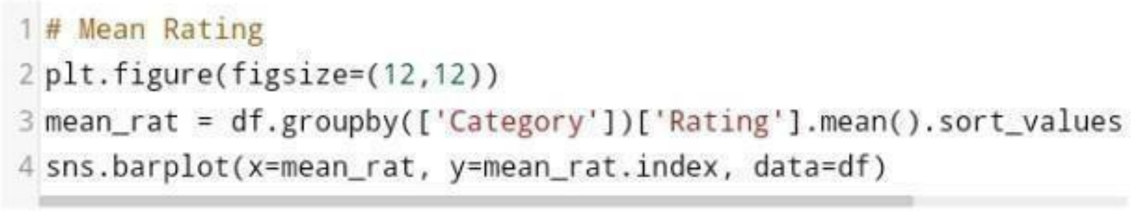
```
[ ] #creating a group of category by the given dataset
    cat_gk=df.groupby('Category')
    cat_gk.first()
```

```
[ ] sns.barplot(cat_gk['Rating'].mean().sort_values(ascending=False).head(),cat_gk['Rating'].mean().sort
    plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/s
  FutureWarning
```

Fig 4.

By the horizontal is the rating value, and vertically is quantity of the rating.



```
# Mean Rating
plt.figure(figsize=(12,12))
mean_rat = df.groupby(['Category'])['Rating'].mean().sort_values
sns.barplot(x=mean_rat, y=mean_rat.index, data=df)
```

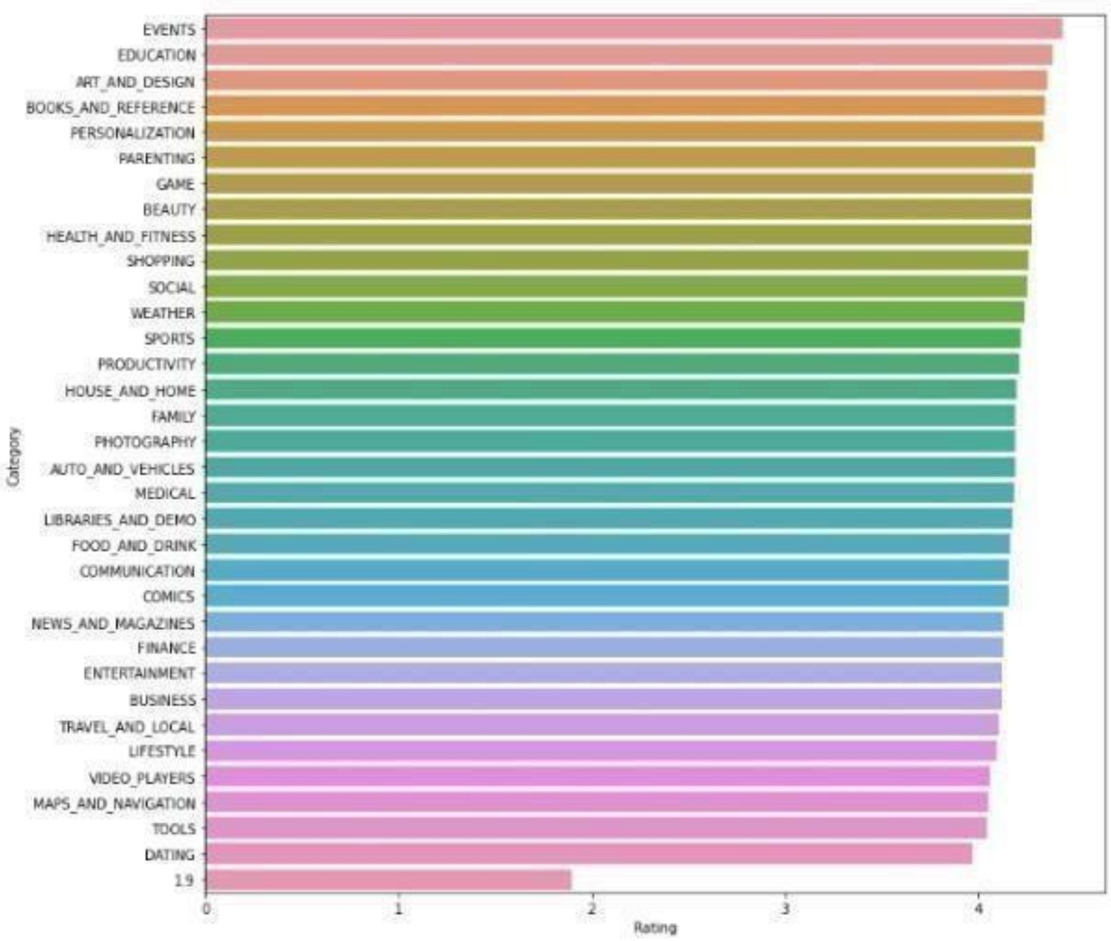
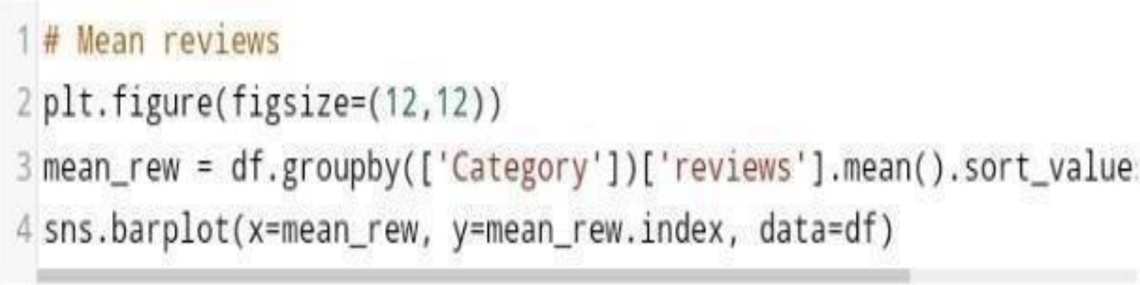


Fig 5.

And this is the average of rating by category, family and game has a lot of quantity causing the low on average rating, on the other side event has the highest average rating by category.

Showing the amount of total reviews.



```
# Mean reviews
plt.figure(figsize=(12,12))
mean_rew = df.groupby(['Category'])['reviews'].mean().sort_value
sns.barplot(x=mean_rew, y=mean_rew.index, data=df)
```

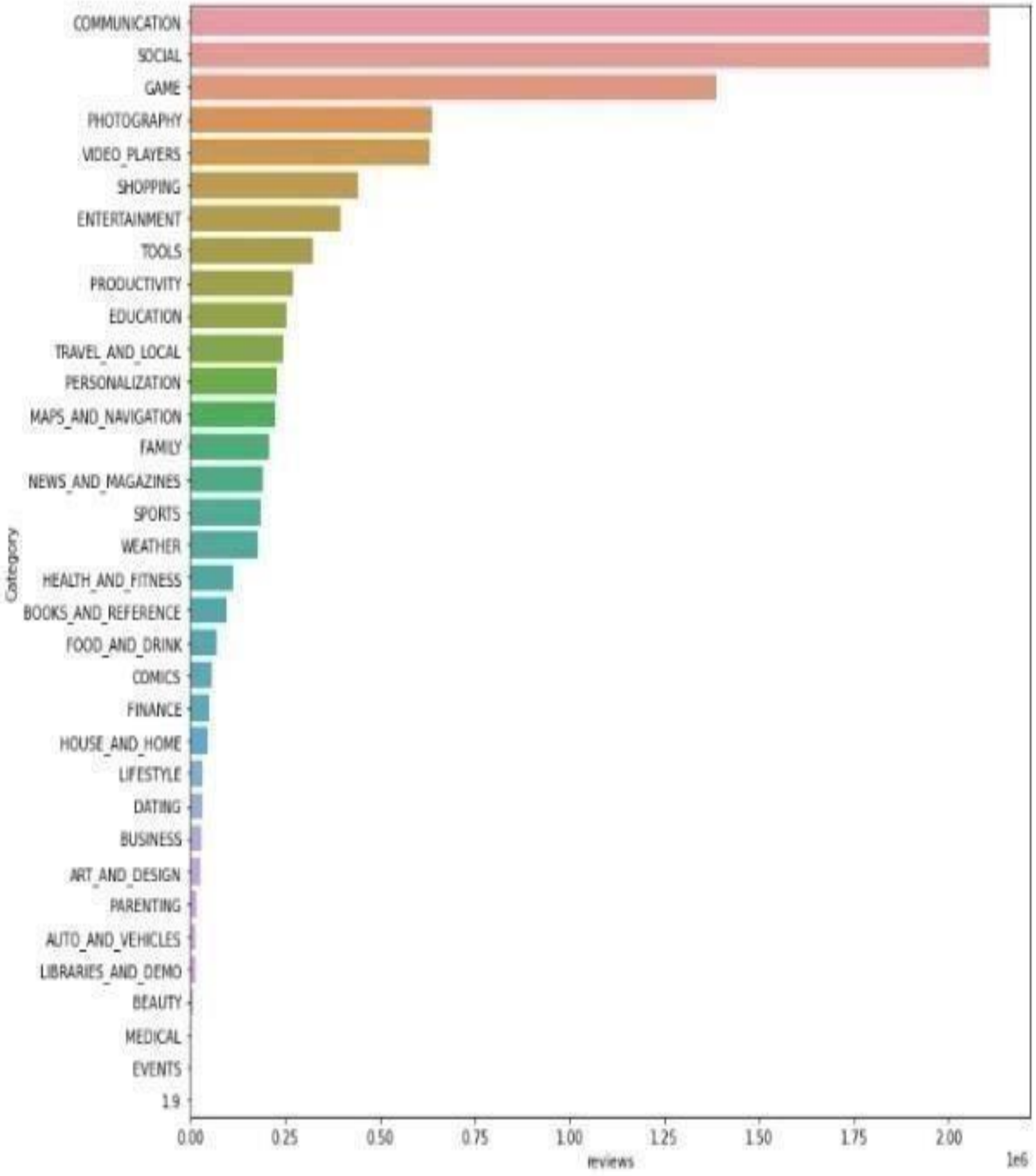
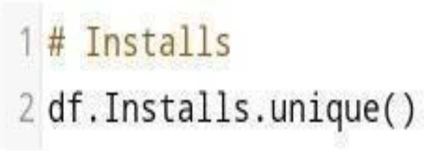
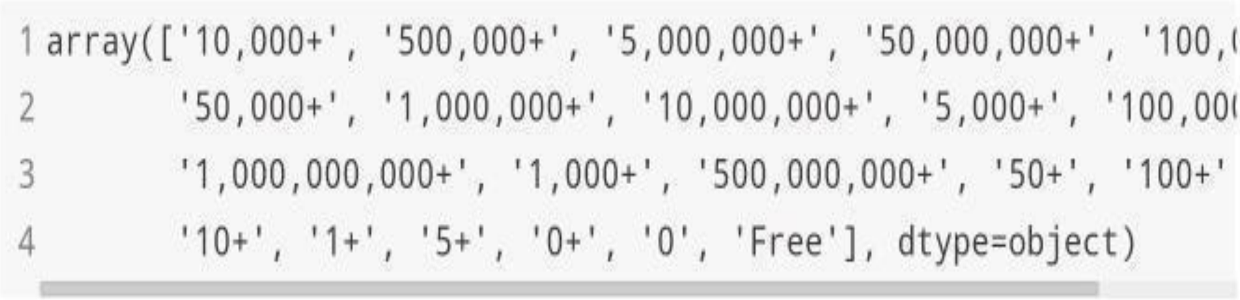


Fig 6.

This is the average of reviews on each category. Let’s move on to next column, installs.

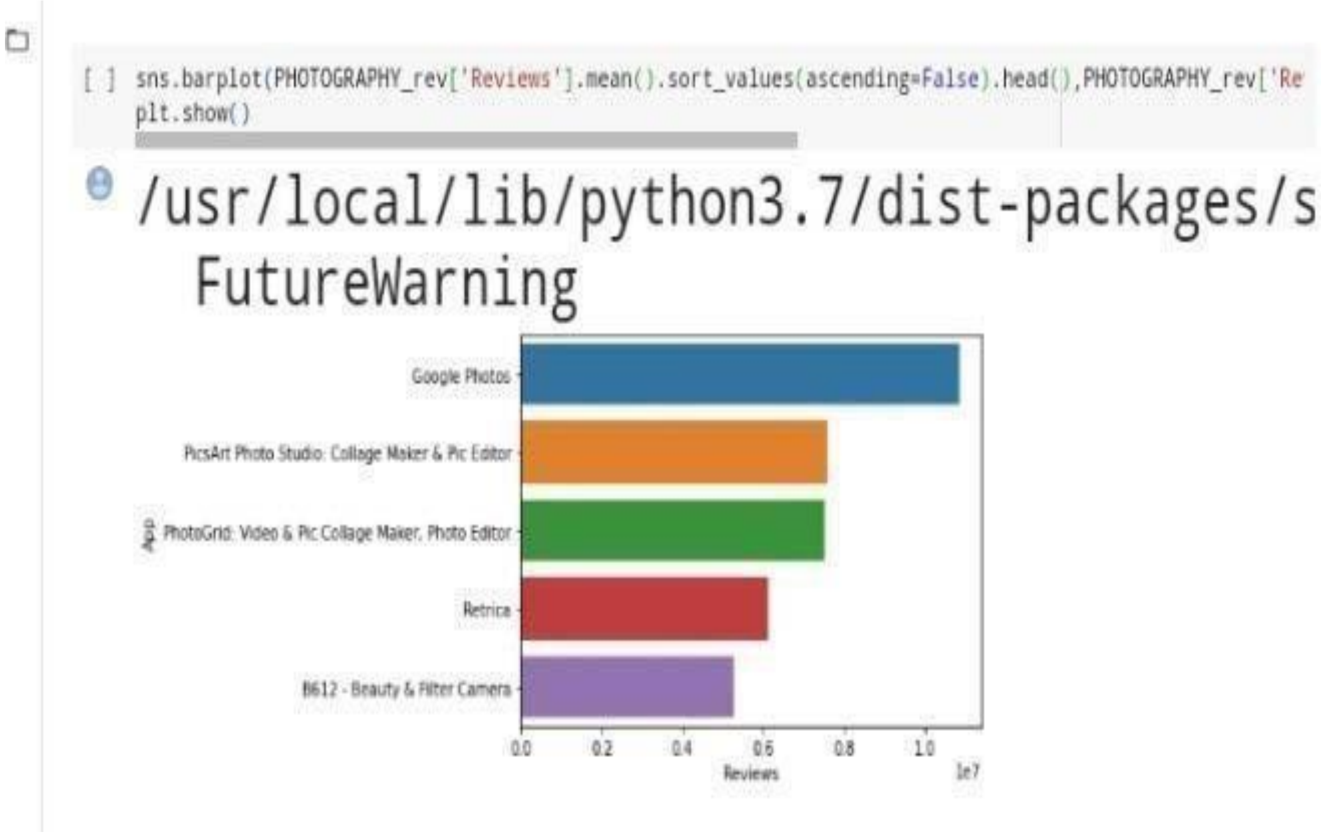


```
# Installs
df.Installs.unique()
```



```
array(['10,000+', '500,000+', '5,000,000+', '50,000,000+', '100,(
       '50,000+', '1,000,000+', '10,000,000+', '5,000+', '100,00(
       '1,000,000,000+', '1,000+', '500,000,000+', '50+', '100+'
       '10+', '1+', '5+', '0+', '0', 'Free'], dtype=object)
```

This are the top 5 reviewed PHOTOGRAPHY apps in playstore



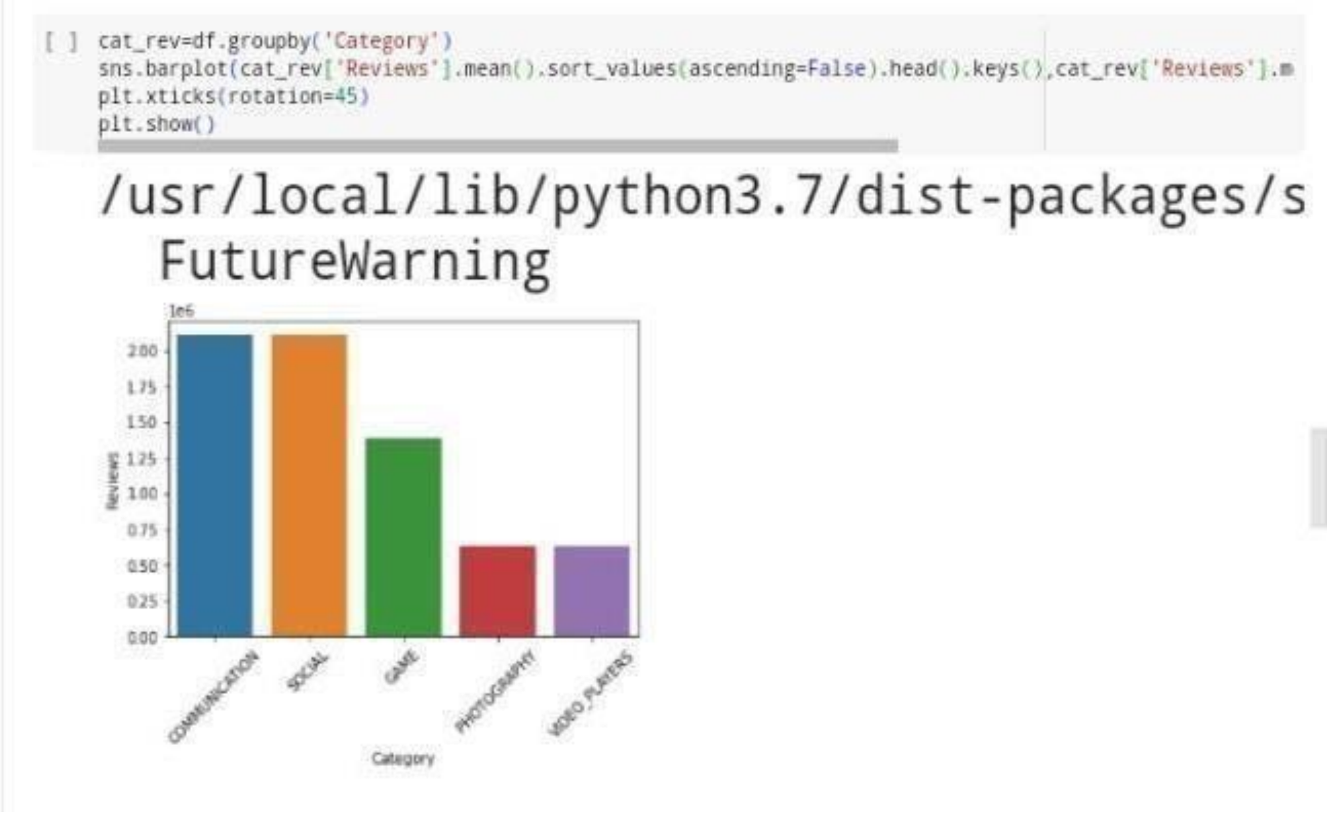
```
[ ] sns.barplot(PHOTOGRAPHY_rev['Reviews'].mean().sort_values(ascending=False).head(),PHOTOGRAPHY_rev['Re
    plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/s
  FutureWarning
```

Google Photos
PicsArt Photo Studio: Collage Maker & Pic Editor
PhotoGrid: Video & Pic Collage Maker, Photo Editor
Retrica
B612 - Beauty & Filter Camera
App
0.0
0.2
0.4
0.6
0.8
1.0
Reviews
1e7

Fig 7.

top 5 category of apps having maximum mean of reviews

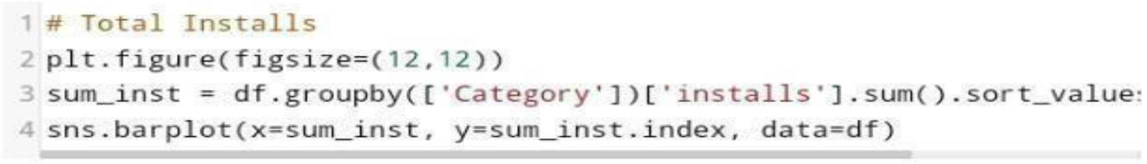


```
cat_rev=df.groupby('Category')
sns.barplot(cat_rev['Reviews'].mean().sort_values(ascending=False).head().keys(),cat_rev['Reviews'].m
plt.xticks(rotation=45)
plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/s
  FutureWarning
```

Fig 8.

- In above bargraph visualisation 1 unit is equals to 1 million reviews



```
# Total Installs
plt.figure(figsize=(12,12))
sum_inst = df.groupby(['Category'])['installs'].sum().sort_value:
sns.barplot(x=sum_inst, y=sum_inst.index, data=df)
```

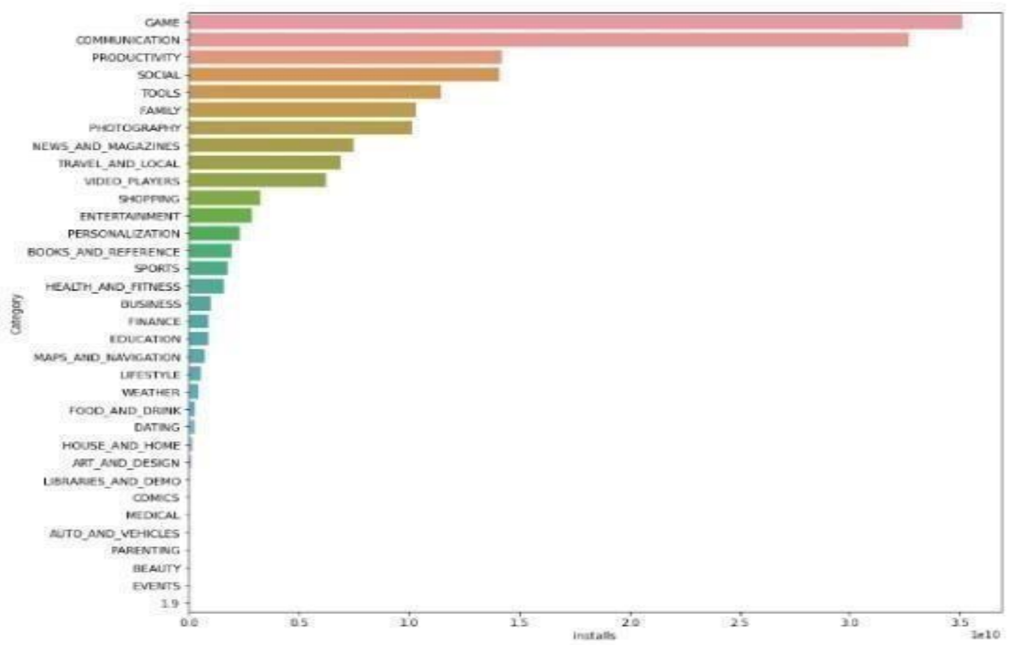
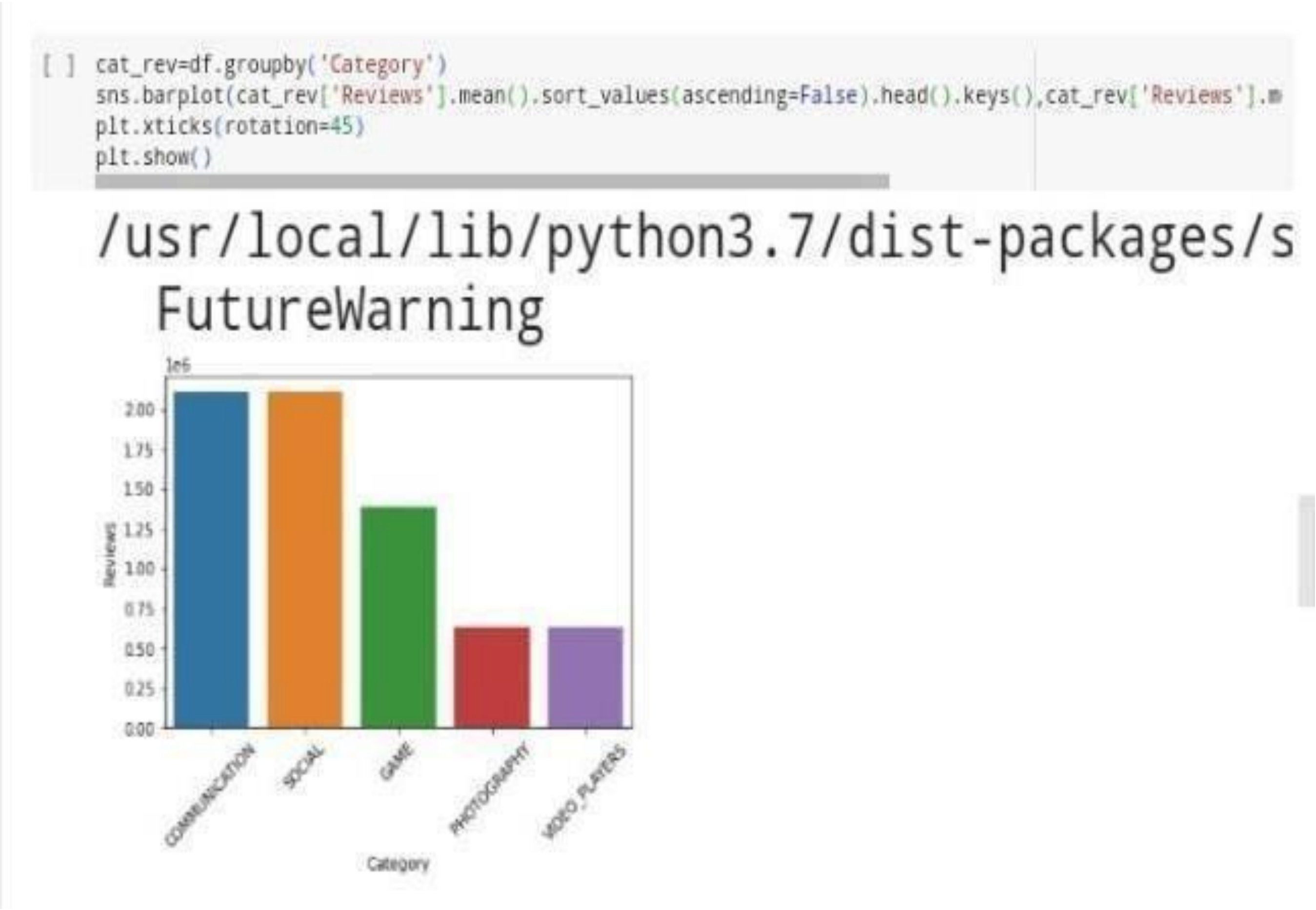


Fig 9.

This are the top 5 reviewed SOCIAL MEDIA apps in playstore.



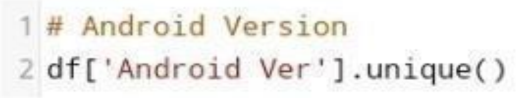
```
[ ] cat_rev=df.groupby('Category')
    sns.barplot(cat_rev['Reviews'].mean().sort_values(ascending=False).head().keys(),cat_rev['Reviews'].m
    plt.xticks(rotation=45)
    plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/s
  FutureWarning
```

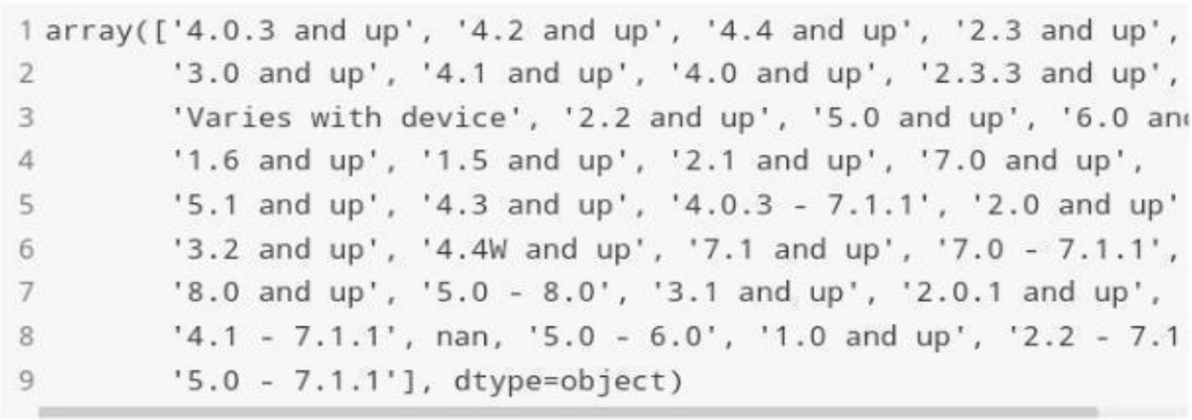
Fig 10.

The medical has a high amount of paid app considering quantity of medical app is not much.

Last is the version of android you should have before accessing the app.



```
# Android Version
df['Android Ver'].unique()
```



```
array(['4.0.3 and up', '4.2 and up', '4.4 and up', '2.3 and up',
       '3.0 and up', '4.1 and up', '4.0 and up', '2.3.3 and up',
       'Varies with device', '2.2 and up', '5.0 and up', '6.0 an
       '1.6 and up', '1.5 and up', '2.1 and up', '7.0 and up',
       '5.1 and up', '4.3 and up', '4.0.3 - 7.1.1', '2.0 and up'
       '3.2 and up', '4.4W and up', '7.1 and up', '7.0 - 7.1.1',
       '8.0 and up', '5.0 - 8.0', '3.1 and up', '2.0.1 and up',
       '4.1 - 7.1.1', nan, '5.0 - 6.0', '1.0 and up', '2.2 - 7.1
       '5.0 - 7.1.1'], dtype=object)
```

Below bar graph showing the content rating and it’s count

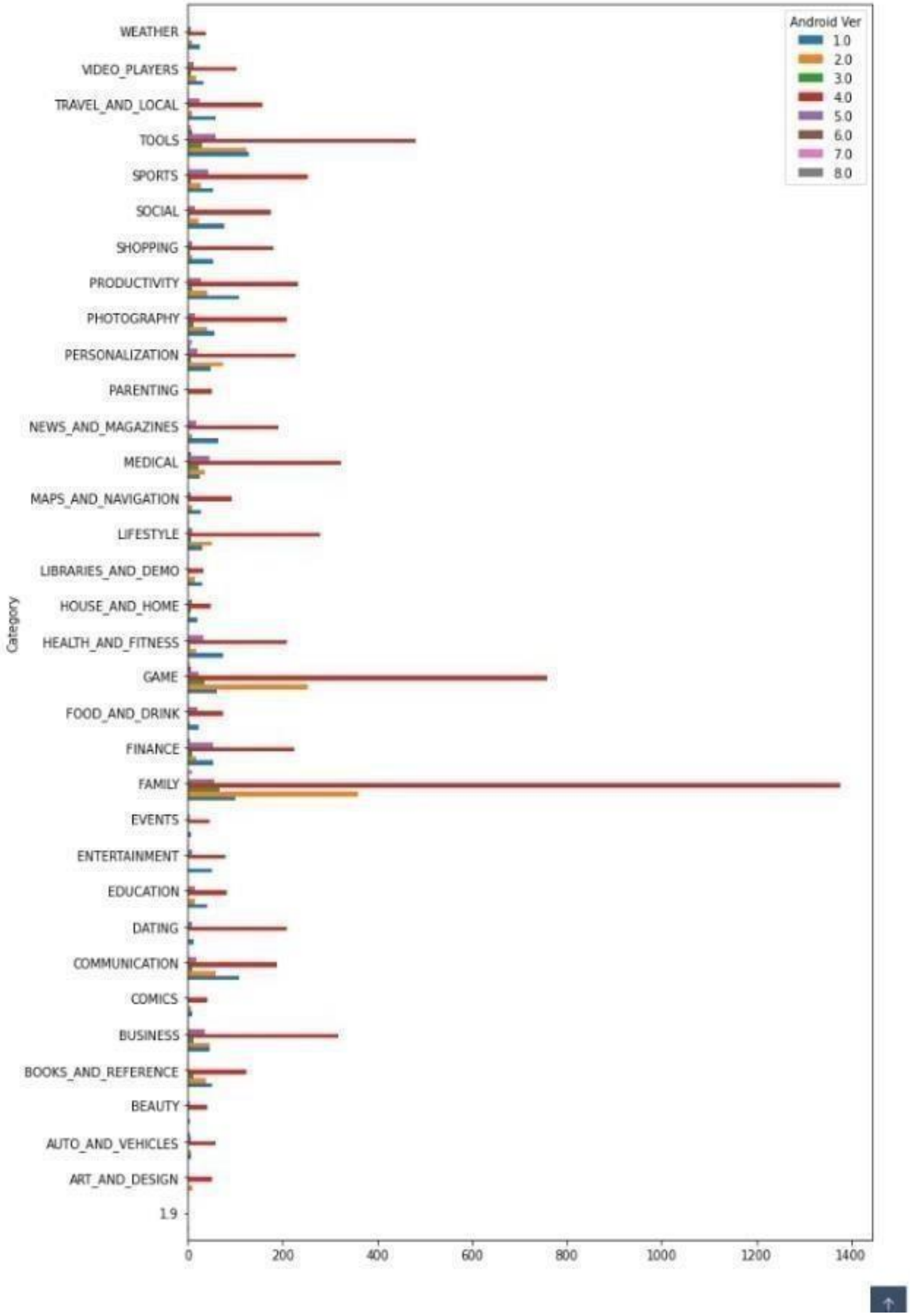
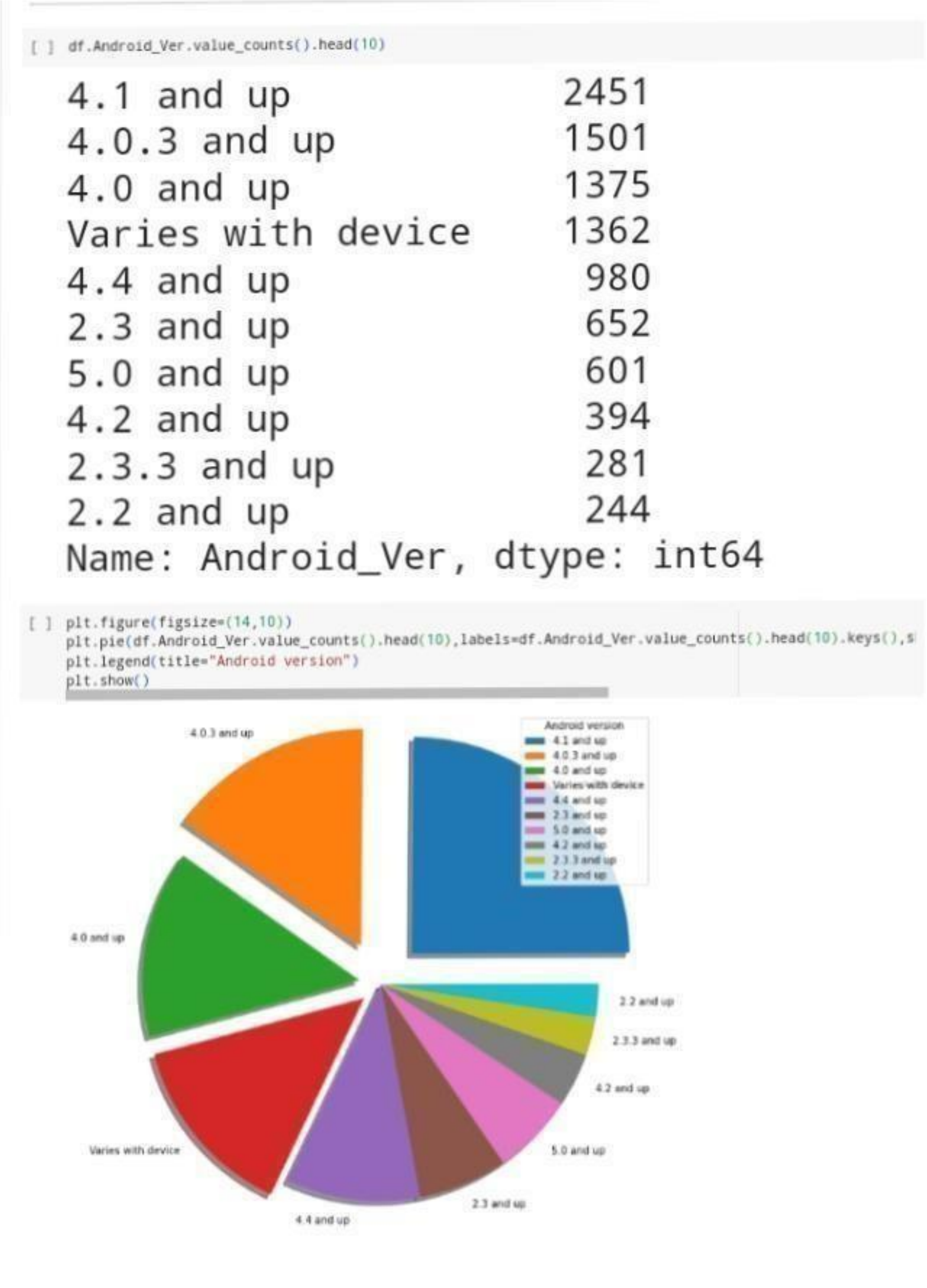


Fig 11.

Analysis of Android version:



```
[ ]  df.Android_Ver.value_counts().head(10)
```

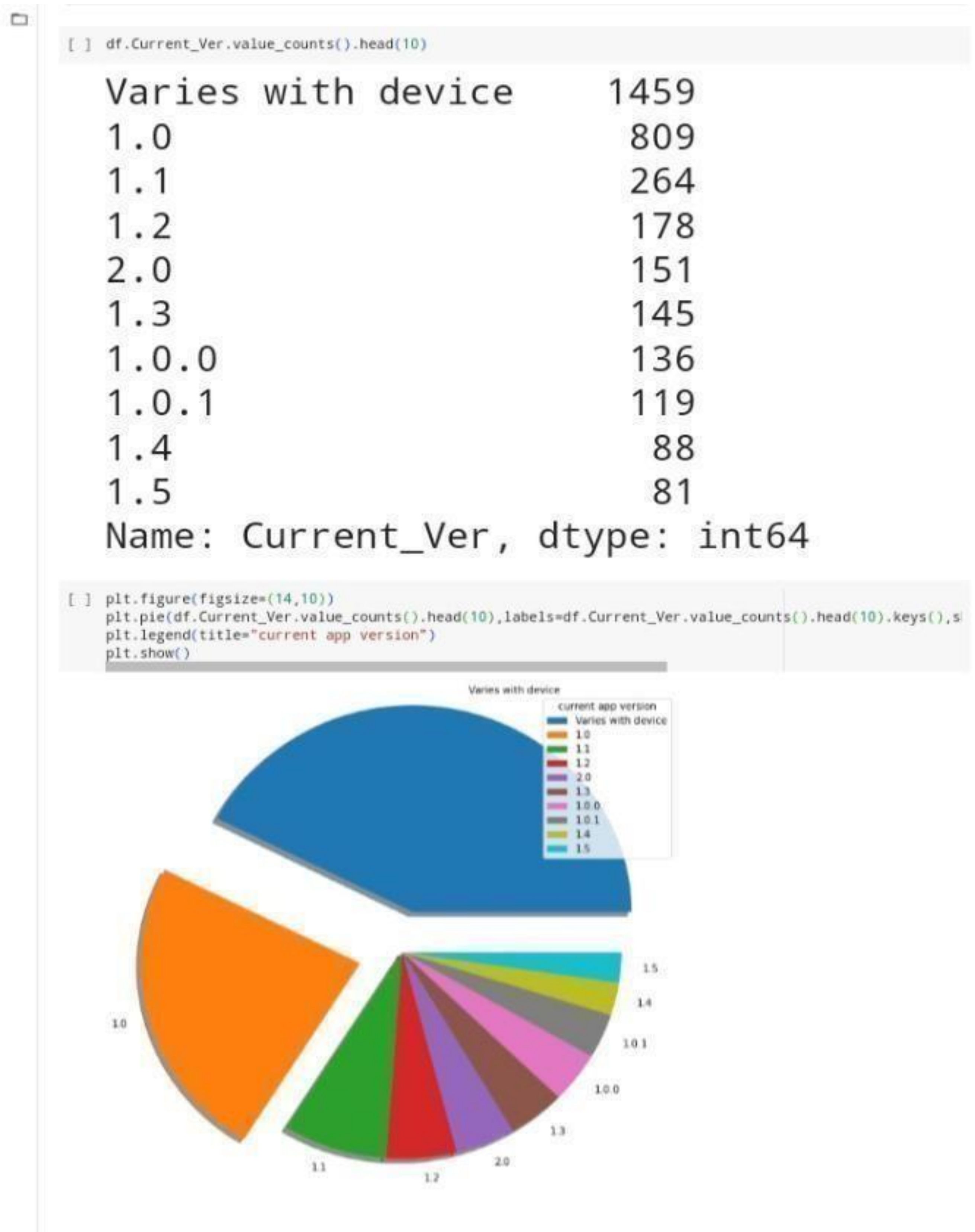
```
4.1 and up              2451
4.0.3 and up            1501
4.0 and up              1375
Varies with device      1362
4.4 and up               980
2.3 and up               652
5.0 and up               601
4.2 and up               394
2.3.3 and up             281
2.2 and up               244
Name: Android_Ver, dtype: int64
```

```
[ ]  plt.figure(figsize=(14,10))
     plt.pie(df.Android_Ver.value_counts().head(10),labels=df.Android_Ver.value_counts().head(10).keys(),s
     plt.legend(title="Android version")
     plt.show()
```

Fig 12.

Required Android version for most of the apps in given dataset should be 4.1 or Above.

Analysis of Current version of App.



```
[ ] df.Current_Ver.value_counts().head(10)
```

```
Varies with device    1459
1.0                    809
1.1                    264
1.2                    178
2.0                    151
1.3                    145
1.0.0                  136
1.0.1                  119
1.4                     88
1.5                     81
Name: Current_Ver, dtype: int64
```

```
[ ] plt.figure(figsize=(14,10))
    plt.pie(df.Current_Ver.value_counts().head(10),labels=df.Current_Ver.value_counts().head(10).keys(),s
    plt.legend(title="current app version")
    plt.show()
```

Fig 13.

Most of the app versions varies with Devices.

6. CONCLUSION :

Data analysis is an important process of research or simply discovering information related to any work. Data derived from the observation, experiment, and other primary and secondary data collection methods is large and cannot be taken as it is. Not all data is relevant, neither can it directly signify any trends, relations, facts, and associations within the data. To find out those required trends and relations, the data needs to be reconstructed in the relevant form and modified. This process is called data analysis. Data analysis and conclusion take forward the research.

The data need Is to be clearly defined before the collection of data itself and the process is as follows:

Data Collection: It is gathering information based on research objectives and variables identified previously. The data gathered should be accurate and related to the research question. Data is collected from various sources using secondary collection techniques from organizational databases, previous surveys, and documents. Data is primarily collected through personal interactions and surveys. Then data is arranged and cleaned, removing insignificant information.

Data Processing: After arranging, data needs to be organized in tabular form with suitable analysis tools. Data needs to be arranged in spreadsheets and other statistical tools, then data modeling has to be created.

Data Analysis: Data needs to be cleaned of errors before analysis. Statistical tools provide analysis like regression, correlation, averages, and others. After tools are applied, data needs to be understood and its findings are interpreted according to the research question.

Communicate results: Visualized data needs to be written in clearly and results should be shown in a classified and organized way. The data findings with diagrams and graphs make the process of data interpretation and presentation complete.

Finally, the conclusion is the essential step in completing the data analysis process. Data analysis gives out certain results, but in big research studies, it is difficult to understand the essence or crux of the findings, relevant to the topic under study. The conclusion gives important inferences derived from the study and bind them together as a final summary of findings.

Cause and effect: The conclusion should be derived based on cause and effect relations. The cause and effect among the data variables, classes, samples, and groups provide a final conclusion.

Generalizations: Though generalization should be avoided; certain large samples can be generalized to derive conclusions. The populations with simple structures, small populations that can find certain general characteristics among themselves can be generalized.

Data Reporting: All the organized data, along with findings, and results in the visualized form, should be reported on the paper in the form of a document and following a certain format that is called data report or research paper/thesis. Final reporting of data in the prescribed format, along with research question, methodology, and literature review, must be put together as a report in the final step of data analysis and conclusion.

Clearly defining limitations is also important within concussion. If the current study has found some inferences similar to what earlier done, those can be underlined. Data should be error-free and statistically significant. The patterns of significance are also an important part of the conclusion.

7. Reference:

- https://thecleverprogrammer.com/2020/06/22/google-play-store-data-analysis-with-python/
- https://www.kaggle.com/lava18/google-play-store-apps
- https://www.kaggle.com/code/ecemboluk/google-play-store-analysis
- https://github.com/souravskr/Google-Play-Store-Analysis