APP DATA

A project report submitted in the partial fulfillment of the requirements for the

Award of the degree of

BACHELOR OF TECHNOLOGY

In

Computer Science and Engineering

Submitted By

Gogineni Hemanth Sai

Reg.no: 121910320015

Under the Guidance of Mr. P. Vishak Bharadwaj

24th June 2022



DECLARATION BY THE CANDIDATE

I the undersigned solemnly declare that the project report APP DATA is based on my own work carried out during the course of our study under the supervision of Mr. P. Vishak Bharadwaj. I assert the statements made and conclusions drawn are an outcome of my research work. I further certify that

1. The work contained in the report is original and has been done by me under the general supervision of my supervisor.
2. The work has not been submitted to any other institution for any other degree/diploma/certificate in this university or any other University of India or abroad.
3. We have followed the guidelines provided by the university in writing the report.
4. Whenever we have used materials (data, theoretical analysis, and text) from other sources, we have given due credit to them in the text of the report and giving their details in the references.

Gogineni Hemanth Sai

(121910320015)

(PG-2223-ETSD-470)

CERTIFICATE

This is to certify that this project work entitled

“APP DATA”

is the Bona fide work carried out by Gogineni Hemanth Sai, Reg. No: 121910320015 submitted in partial fulfilment of the requirement for the Degree of Bachelor of Technology in May-June 2022.

The results submitted in this project have been verified and are found to be satisfactory. The results embodied in this thesis have not been submitted to any other university for the award of any other degree/diploma.

Signature of project supervisor

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crowned the efforts with success. It is a pleasant aspect that I have now the opportunity to express my gratitude for all of them.

The first person I would like to thank is my project guide Mr. P. Vishak Bharadwaj, who had given continuous critical suggestions and extension of proper working atmosphere, abiding interest has finally evolved into this research work.

It is indeed with a great sense of pleasure and immense sense of guidance that I acknowledge the help, and I am highly indebted to Prof. Dr. Dharma Raj Cheruku, Principal, and School of Technology, for his support during the tenure of the internship.

I would like to express my sincere thanks to Prof. R. Sireesha, Head of the Department of Computer Science and Engineering for providing the opportunity to undertake this internship and for encouragement in the completion of the project.

I am also thankful to all the staff members of the Computer Science and Engineering Department for their valuable suggestions. Furthermore, I would like to thank my teammates and parents who extended their help, encouragement and moral support either directly or indirectly in this project.

Gogineni Hemanth Sai

(121910320015)

CONTENTS

|  |  |  |
| --- | --- | --- |
| Chapter | Title of the Chapter / Chapter Name | Page No’s |
| A | ABSTRACT | 8 |
| I | INTRODUCTION | 9 |
| II | EDA | 10-11 |
| III | STEPS PERFORMED IN EDA | 12-13 |
| IV | DATA CLEANING AND PREPARATION | 14-20 |
| V | DATA VISUALIZATION | 21-25 |
| B | References and Appendix | 26 |

ABOUT

Phoenix Global is a skill-development company that helps students acquire and master professional and soft skills as per the requirements of the industry benchmarked to world’s top firms, trained by top class industry professionals.

Phoenix Global is a platform having Industry professionals with esteemed alma mater including the IITs and IIMs to mentor and train students on cutting-edge skills, critical to the emerging industries while also giving them an opportunity to intern on a project under the mentorship of industry professionals from the IITs /IIMs.

Our vision is to be a national leader in skill development and industry readiness training by providing differentiated training from top-class industry experts. The mission is to be a go-to skill development platform for students, imparting skills benchmarked at global standards that help them realize their dream careers profitably

Our core values, the 4Ps – Professionalism, Punctuality, Passion, Perseverance stand for who and what we are as an organization.

SCHEDULE OF INTERNSHIP

|  |  |
| --- | --- |
| **Day** | **Activity Plan** |
| 1 | Induction Program |
| 2 | Pre-Readings/Material Distribution |
| 3 | Training Session - 1 |
| 4 | Training Session - 2 |
| 5 | Training Session - 3 |
| 6 | Training Session - 4 |
| 7 | Training Session - 5 |
| 8 | Teams formation for Project |
| 9 | Weekend Off |
| 10 | Training Session - 6 |
| 11 | Training Session - 7 |
| 12 | Training Session - 8 |
| 13 | Training Session - 9 |
| 14 | Training Session - 10 |
| 15 | Project Title Allocation |
| 16 | Weekend Off |
| 17 | Project Session - 1 |
| 18 | Project Session - 2 |
| 19 | Project Session - 3 |
| 20 | Project Session - 4 |
| 21 | Project Session - 5 |
| 22 | Project Mid Review |
| 23 | Weekend Off |
| 24 | Project Session - 6 |
| 25 | Project Session - 7 |
| 26 | Project Session - 8 |
| 27 | Project Session - 9 |
| 28 | Project Session - 10 |
| 29-44 | Project Working Sessions |
| 45 | Project Final Presentation and Thesis Defense |

ABSTRACT

The Play Stores in Android Market is one of the largest and most popular Android app stores. It has an enormous amount of data that can be used to make an optimal model. We have used a raw data set of Google Play Store from the Kaggle website.

This data set contains 13 different features that can be used for predicting whether an app will be successful or not using different features. This data set is scraped from the Google Play Store. This journal talks about different classiﬁer models that we used for prediction purposes and ﬁnding which one gives the highest accuracy. This journal also gives detailed information on feature extraction and the complete Data visualization done on this data set. Our project code can be found at [GitHub-AppData](https://github.com/SunithaMeher/App-Data).

CHAPTER 1: INTRODUCTION

Mobile apps are altering their business landscape by reinvigorating companies, disrupting industries, and creating opportunities previously non-existent. Many companies are scrambling to satisfy customers from the point of view, resulting in a reactive approach to app investment.

There has been an exponential rise in smartphones in the last decade. The growth is due to the lowering down of data prices in developing countries such as India coupled with plummeted prices of these devices. With such a humungous rise, there has been a considerable increase in application development for mobile devices. The apps are available in diverse genres with an intent to simplify people's life.

In the present VUCA world, there are disruptions ongoing in every sphere.

Though it is hard to predict what will happen and very little room to avoid unforeseen circumstances, few measures can come to our rescue. Analytics is one such weapon where we can get the best out of hand.

Here’s where analytics comes into play. It captures data from mobile app, website, and web app visitors to identify unique users, track their journeys, record their behaviour, and report on the app’s performance.

Instead of relying on intuition or domain expertise, analytics assists companies with unparalleled insights into the otherwise hidden lives of app users.

Furthermore, this project is all about exploring, visualizing the ongoing exponential rise in creating mobile applications and developing insights about their growth in current android market.

CHAPTER 2: EDA

We have performed an EDA(Exploratory Data Analysis ) on a dataset collected from Kaggle, a platform which allows users to find and publish datasets.

What’s an EDA anyway?

EDA -it is a process to understand the data in depth and learn the different data characteristics, often with visual means .It is an integral part of working with data finding out recurring patterns and significant correlations that might be present in the data.



It is crucial to understand it in depth before you perform data analysis and run your data through an algorithm. You need to know the patterns in your data and determine which variables are important and which do not play a significant role in the output. Further, some variables may have correlations with other variables. You also need to recognize errors in your data.

All of this can be done with Exploratory Data Analysis. It helps you gather insights and make better sense of the data, and removes irregularities and unnecessary values from data.

* Helps you prepare your dataset for analysis.
* Allows a [machine learning model](https://www.simplilearn.com/machine-learning-models-article) to predict our dataset better.
* Gives you more accurate results.
* It also helps us to choose a better machine learning model.

EDA is very essential because it is a good practice to first understand the problem statement and the various relationships between the data features before getting your hands dirty.

 Technically, the primary motive of EDA is to

* Examine the data distribution
* Handling missing values of the dataset(a most common issue with every dataset)
* Handling the outliers
* Removing duplicate data
* Encoding the categorical variables
* Normalizing and Scaling

To understand the steps involved in EDA, we have used Python as the programming language and Jupyter Notebook for this project as it is an open-source platform, an excellent IDE and preferably good for visualization and presentation.

CHAPTER 3: Steps performed in EDA

This process can be summarized into five steps: capture, maintain, process, analyze, and communicate. First, we gather the data that has meaningful variables leading to appropriate classes. Then clean the data so that it is easy for a computer to read and process modeling. Next, we apply algorithms to train the model and test it using data acquired from the Kaggle dataset and analyze the model's performance. We then view the results and attempt to extract any relevant learning or information.

Steps Involved in Exploratory Data Analysis:

1.Data Collection

2.Data Cleaning

3.Data Preparation

4.Handling Outliers

5.Visualizations.

* + - 1. Data Collection and importing:

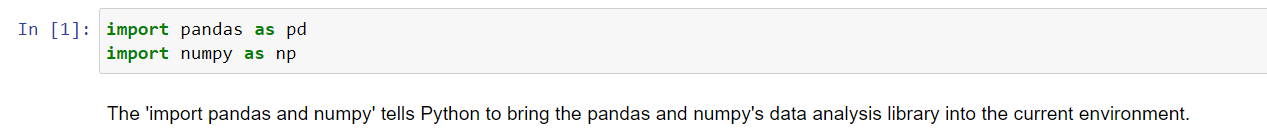
[Data collection](https://www.simplilearn.com/data-collection-methods-article) is an essential part of exploratory data analysis. It refers to the process of finding and loading data into our system. Good, reliable data can be found on various public sites or bought from private organizations.

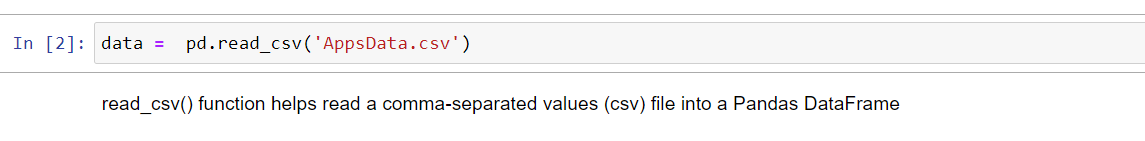
Some reliable sites for data collection are Kaggle, GitHub, [Machine Learning](https://www.simplilearn.com/tutorials/machine-learning-tutorial/what-is-machine-learning) Repository, etc.

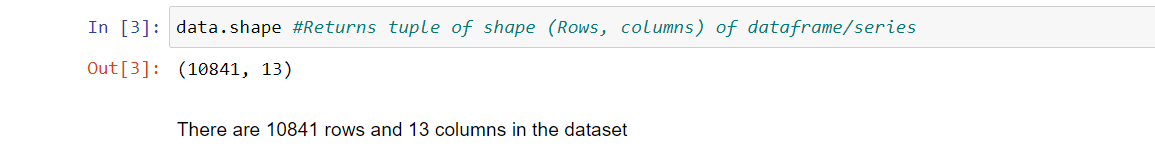
The dataset we have utilized for our project-AppData is from Kaggle.

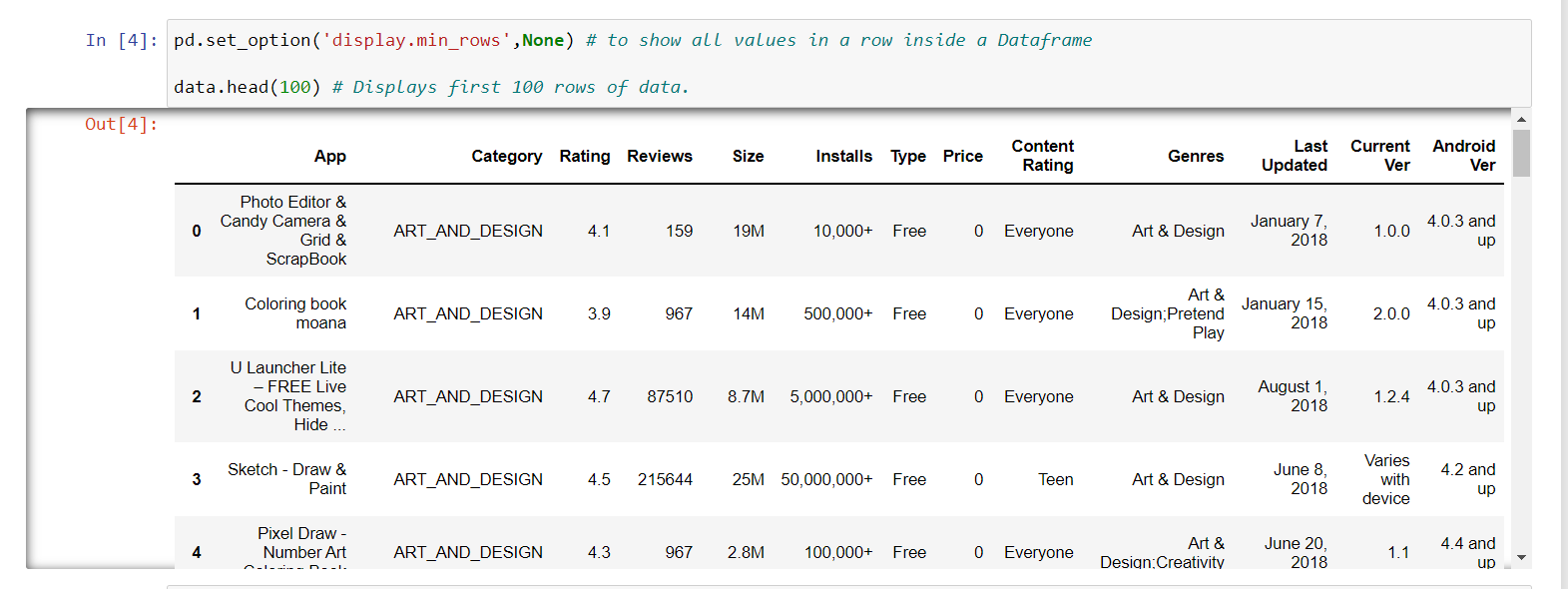
It depicted analyzations of the Apps found in the market with certain fields to gain an insight into the present android market.

Furthermore, we import the data as per the requirements and easy access to python libraries, followed by reading the csv into the kernel and exploring other data characteristics using functions and methods.









CHAPTER 4

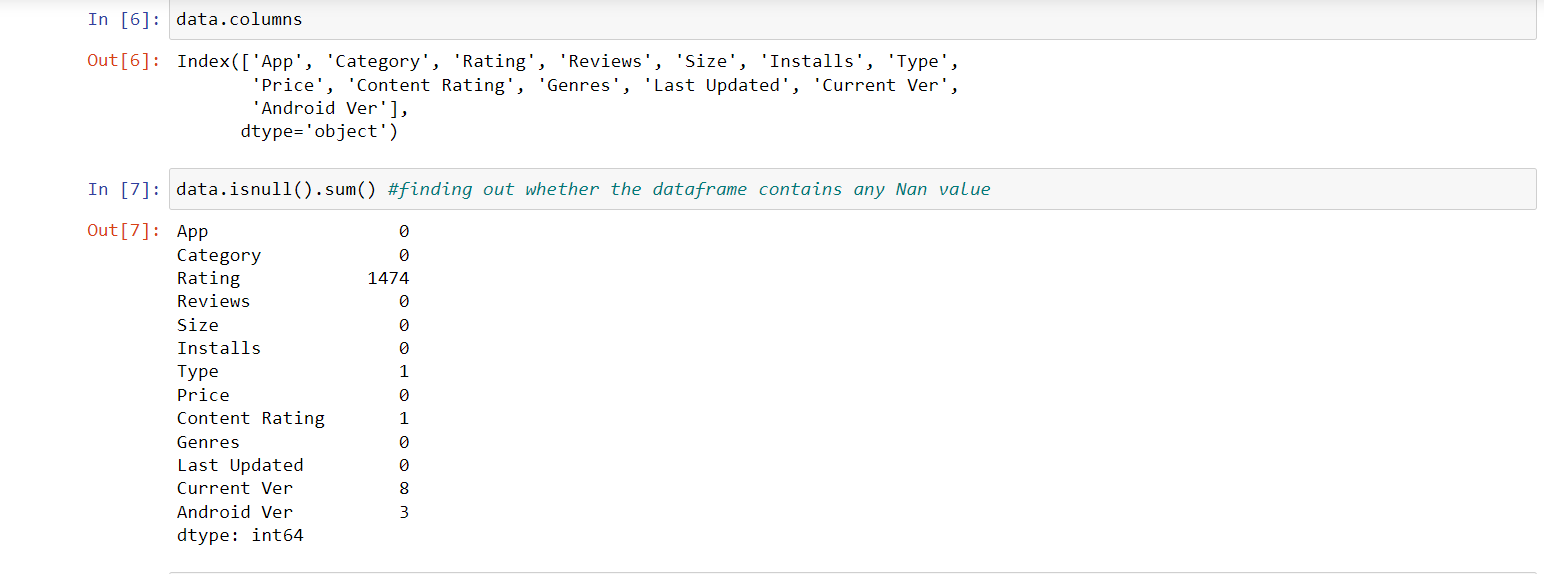
Data Cleaning and Data Preparation

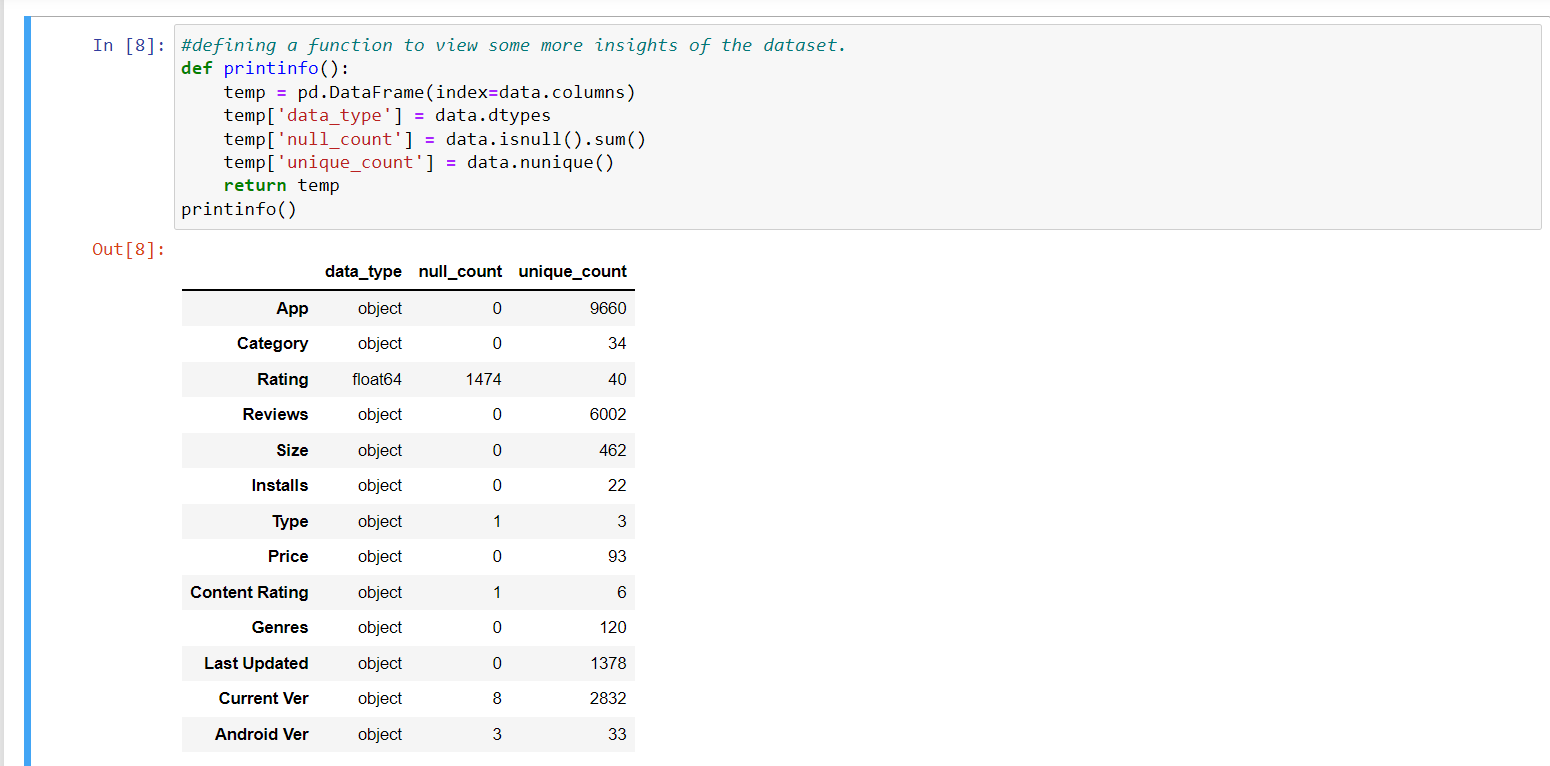
* + - 1. Data Cleaning:

[Data cleaning](https://www.simplilearn.com/data-cleaning-why-and-how-to-get-started-article) refers to the process of removing unwanted variables and values from your dataset and getting rid of any irregularities in it. Such anomalies can disproportionately skew the data and hence adversely affect the results. Some steps that can be done to clean data are:

* Removing missing values, outliers, and unnecessary rows/ columns.
* Re-indexing and reformatting our data.

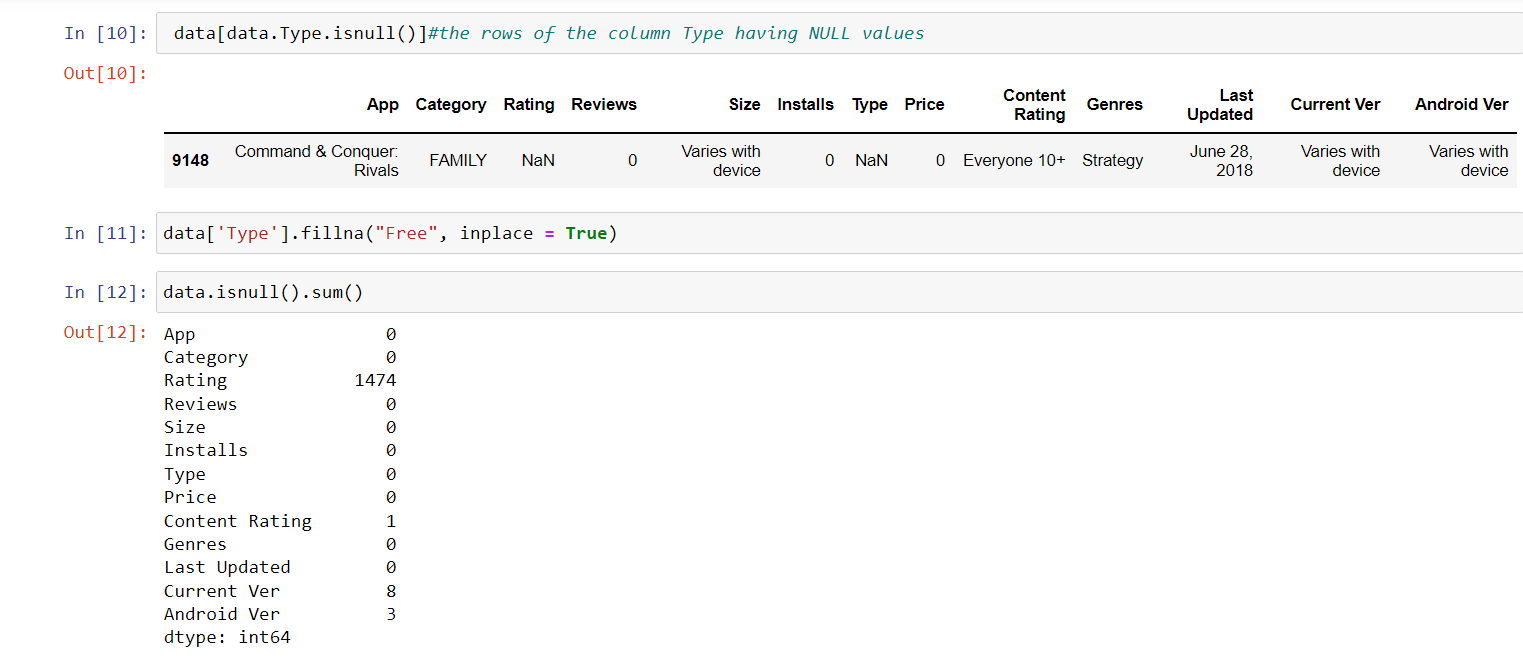
We first need to check to see the number of missing values in each column and the percentage of missing values they contribute to and here are the functions and methods used to perform the respective actions:

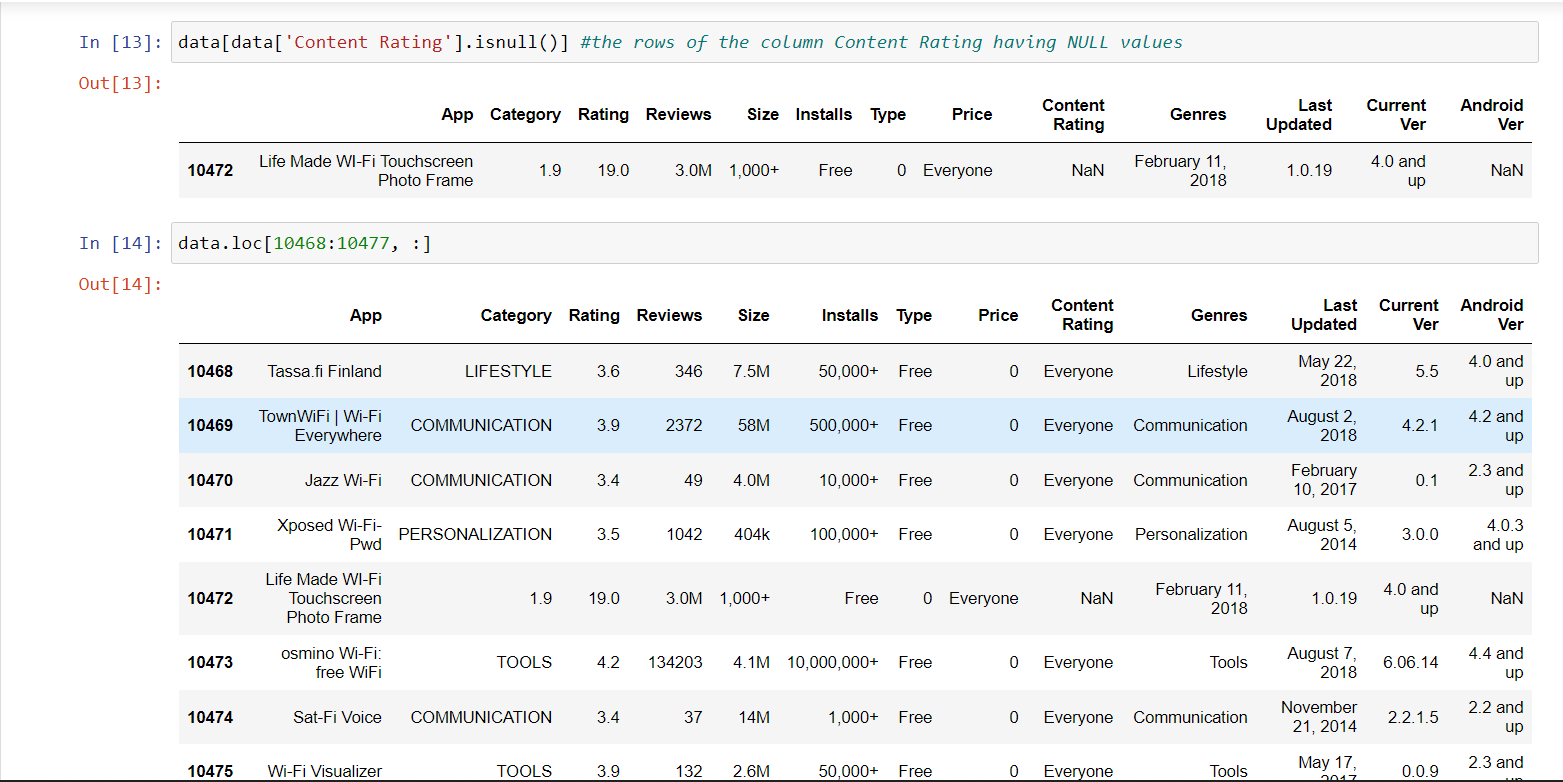


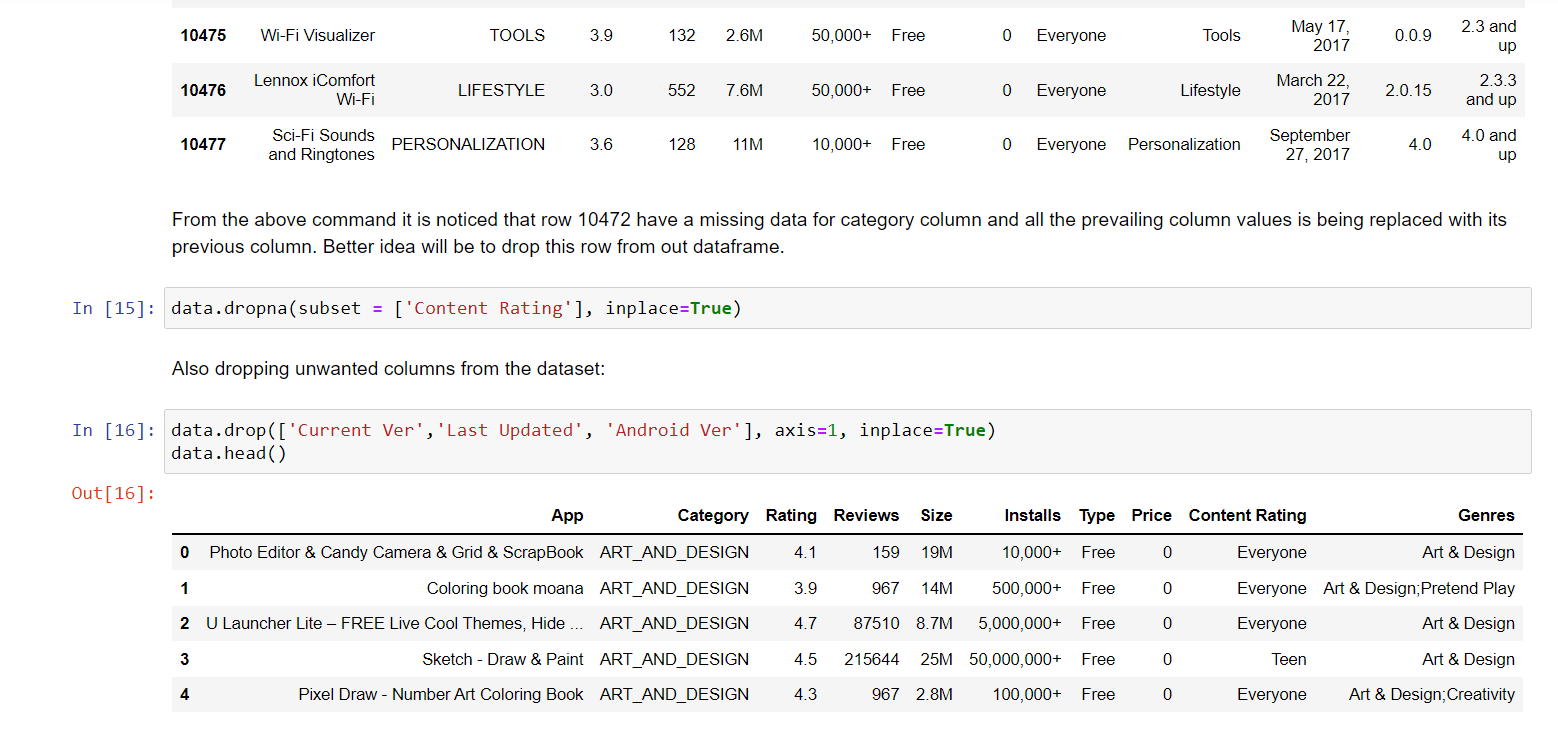


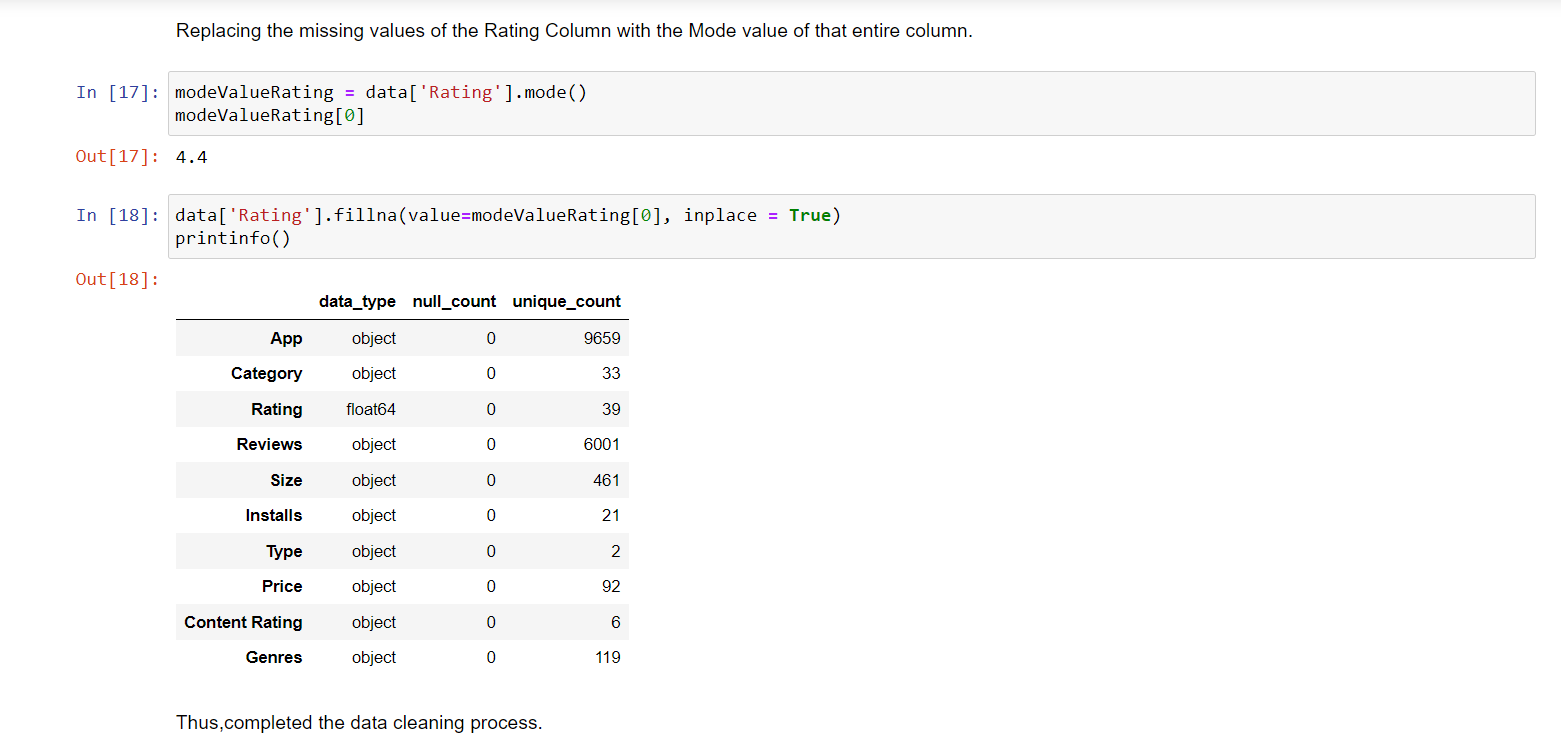
Cleaning the data having null values:







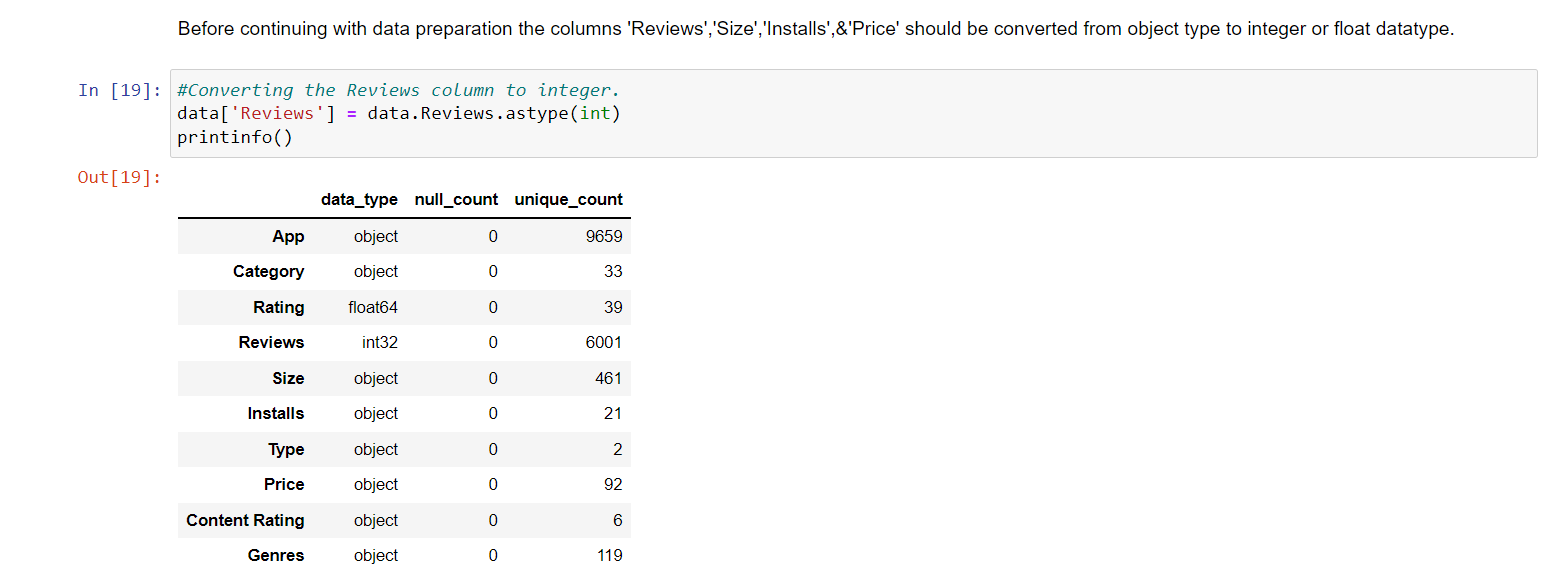


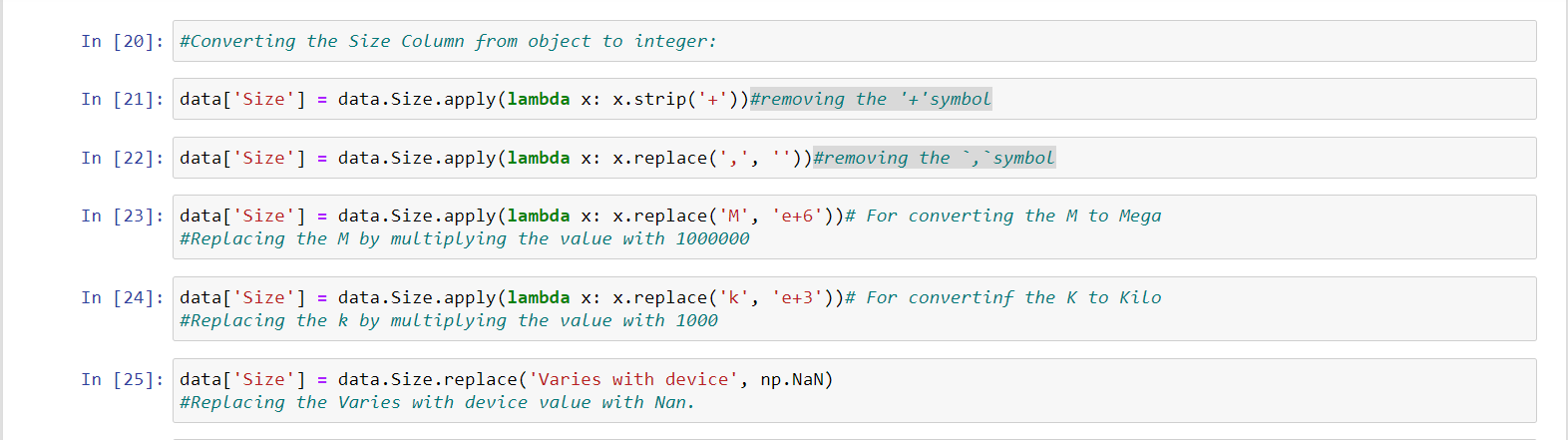


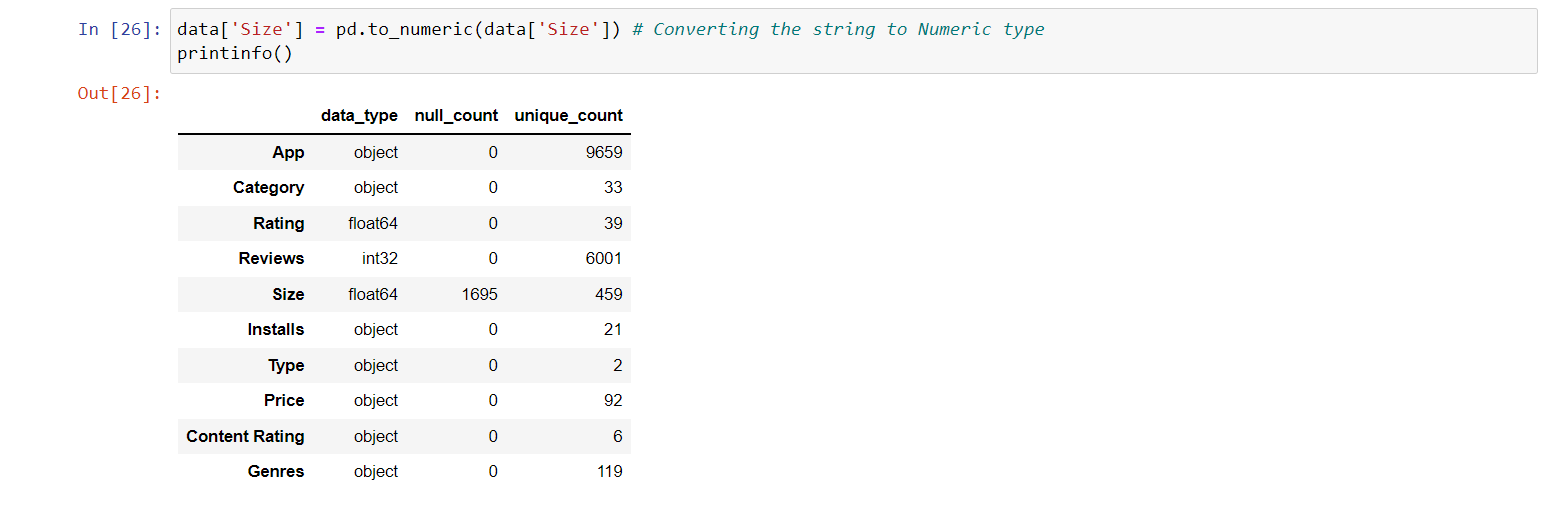
* + - 1. Data Preparation:

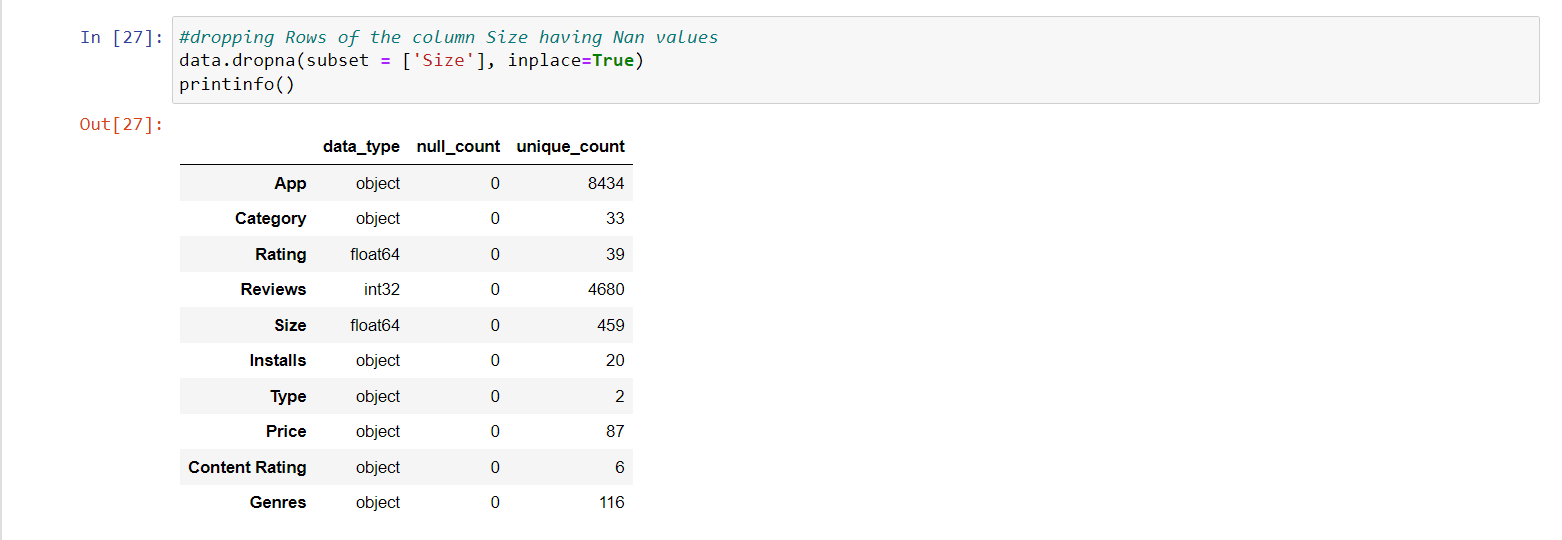
Data preparation is the process of gathering, combining, structuring and organizing data so it can be used in business intelligence (BI), analytics and data visualization applications.

For better overall operational efficiency, data preparation step is continued.

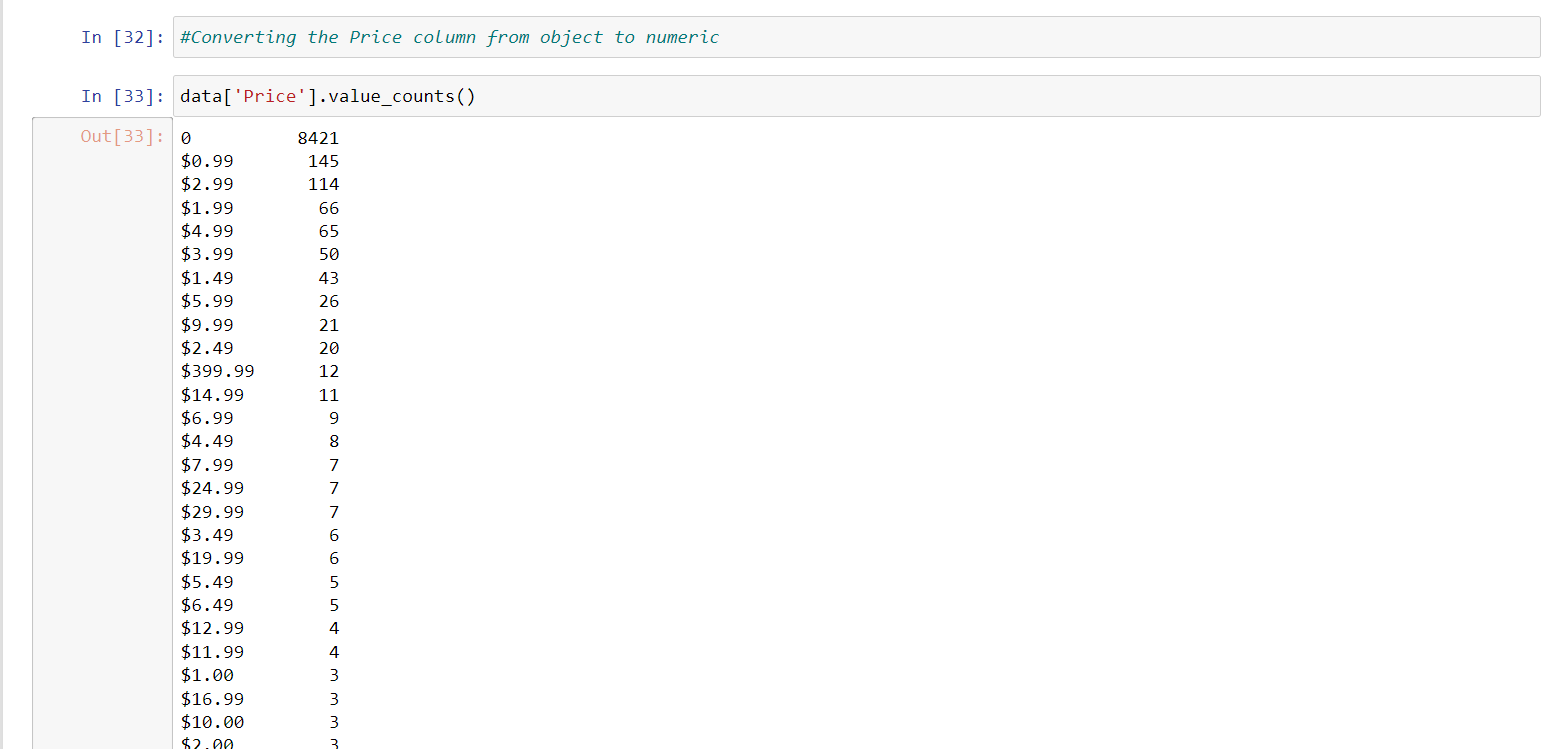


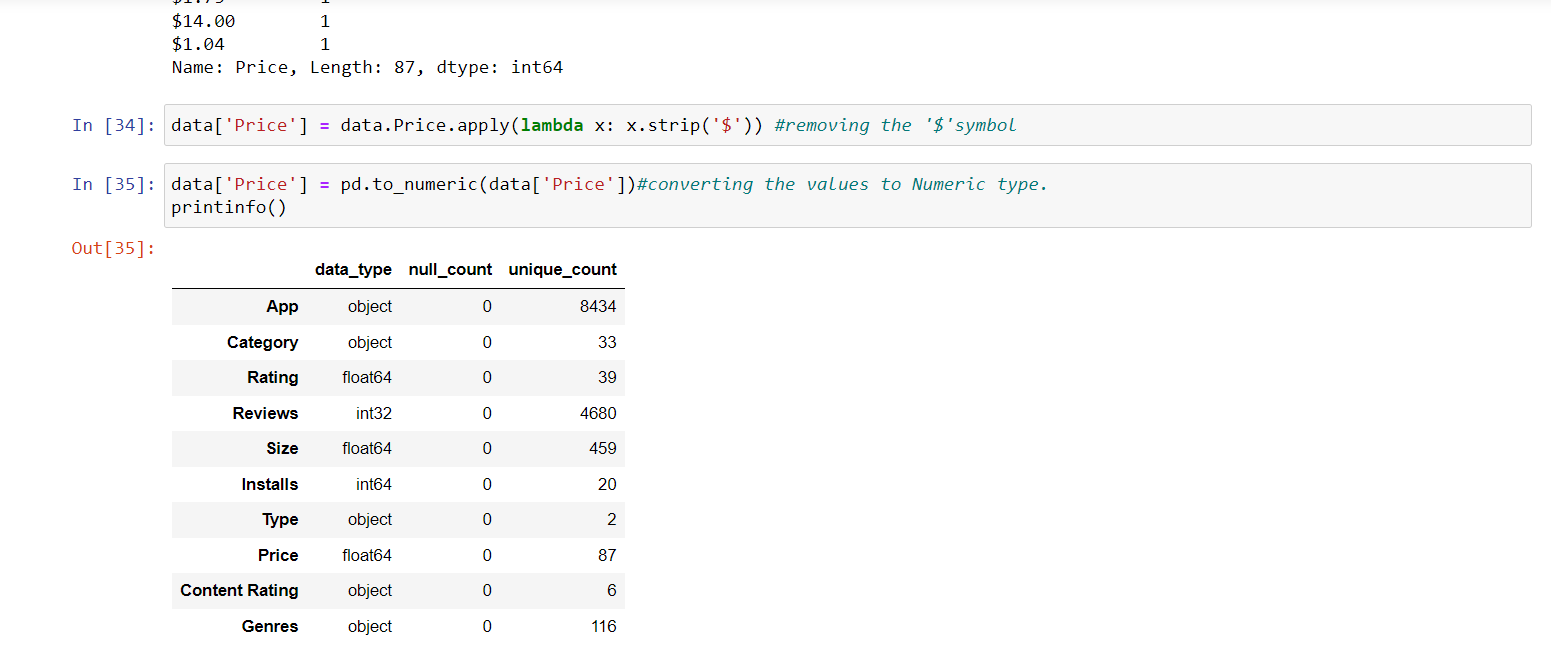


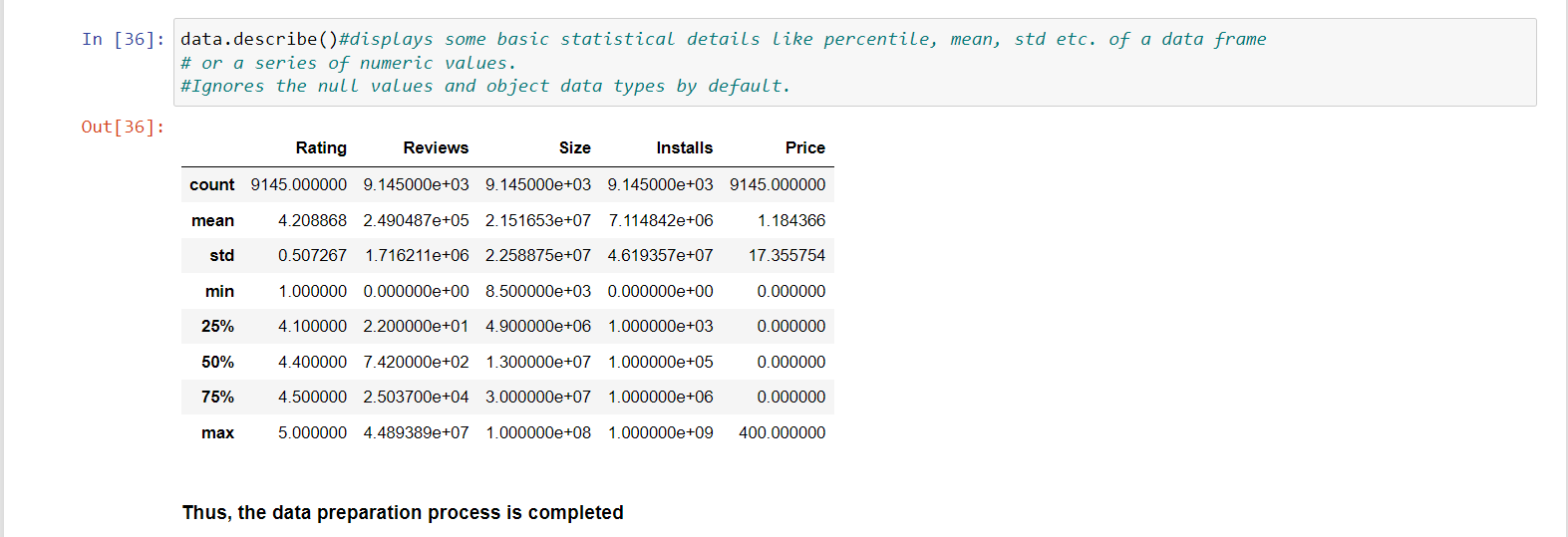










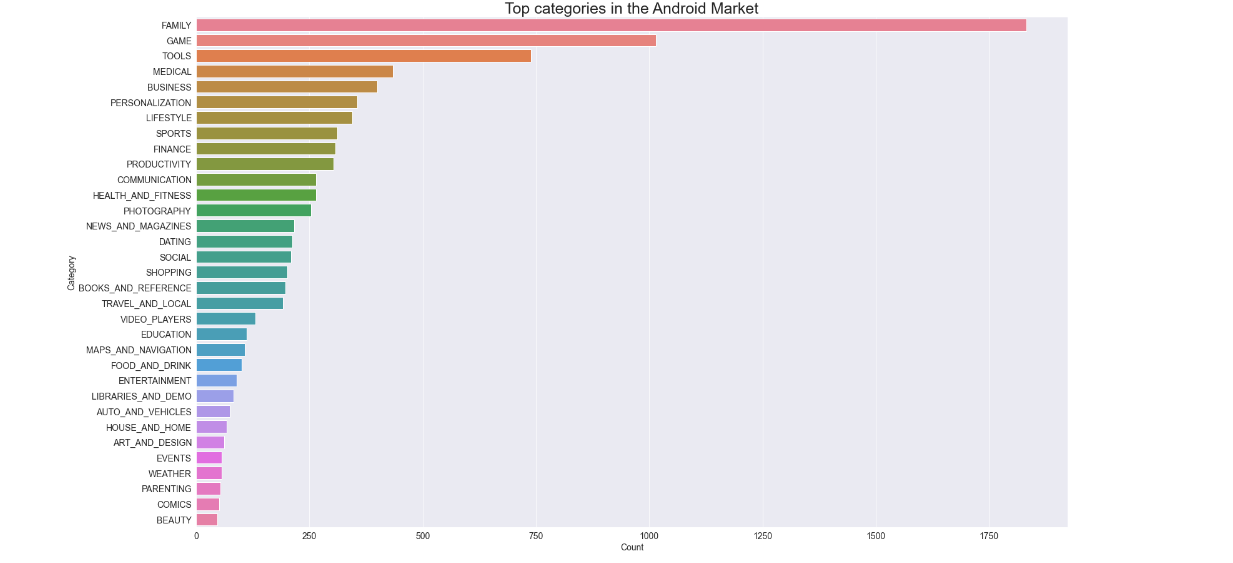


CHAPTER 5: Data Visualization

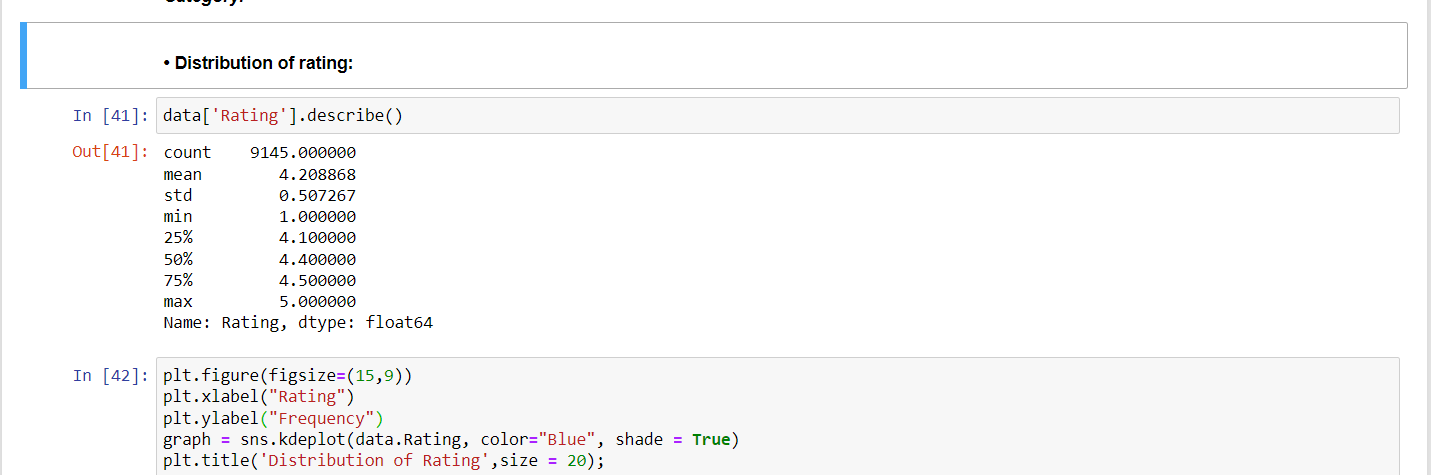
Data Visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

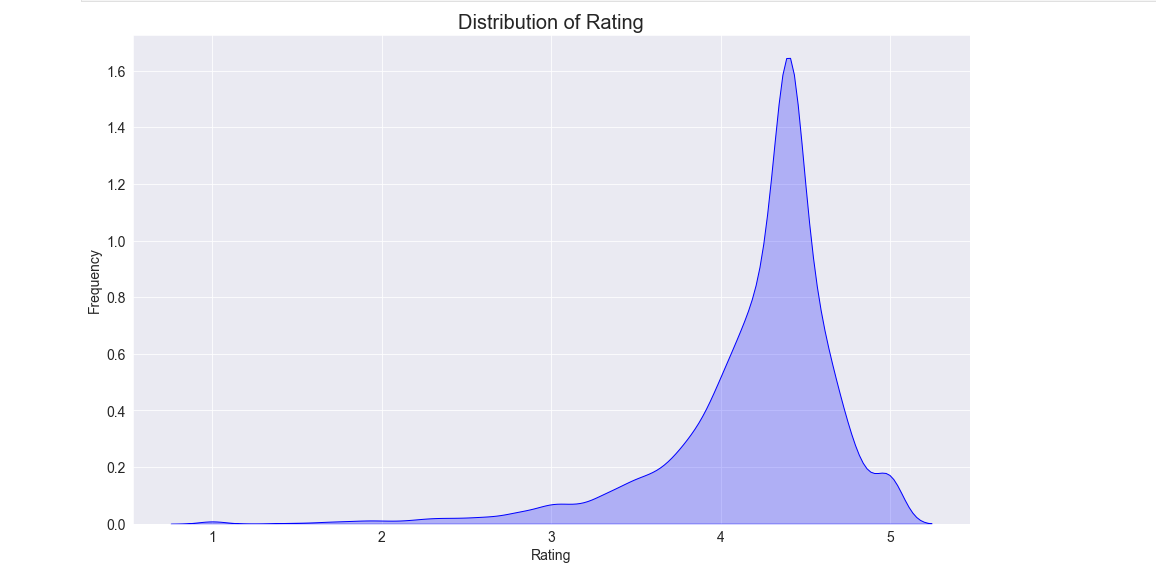
data visualization sits right in the middle of analysis and visual storytelling and helps the target audience arrive at the intended insight.



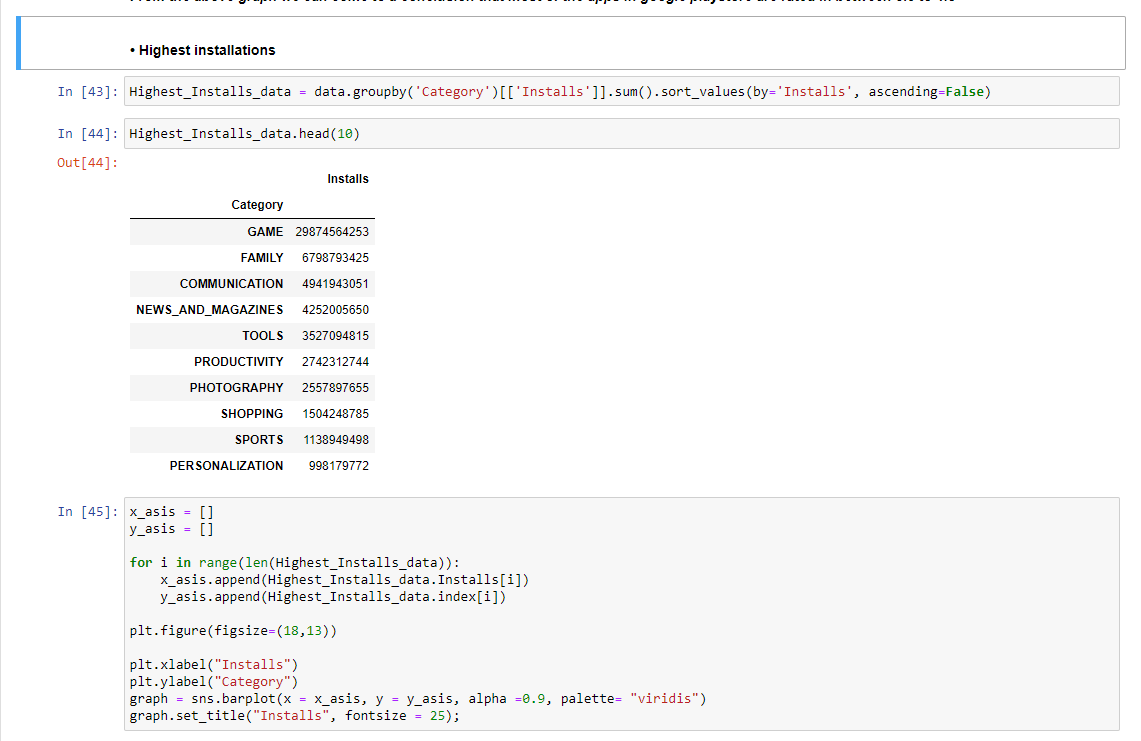


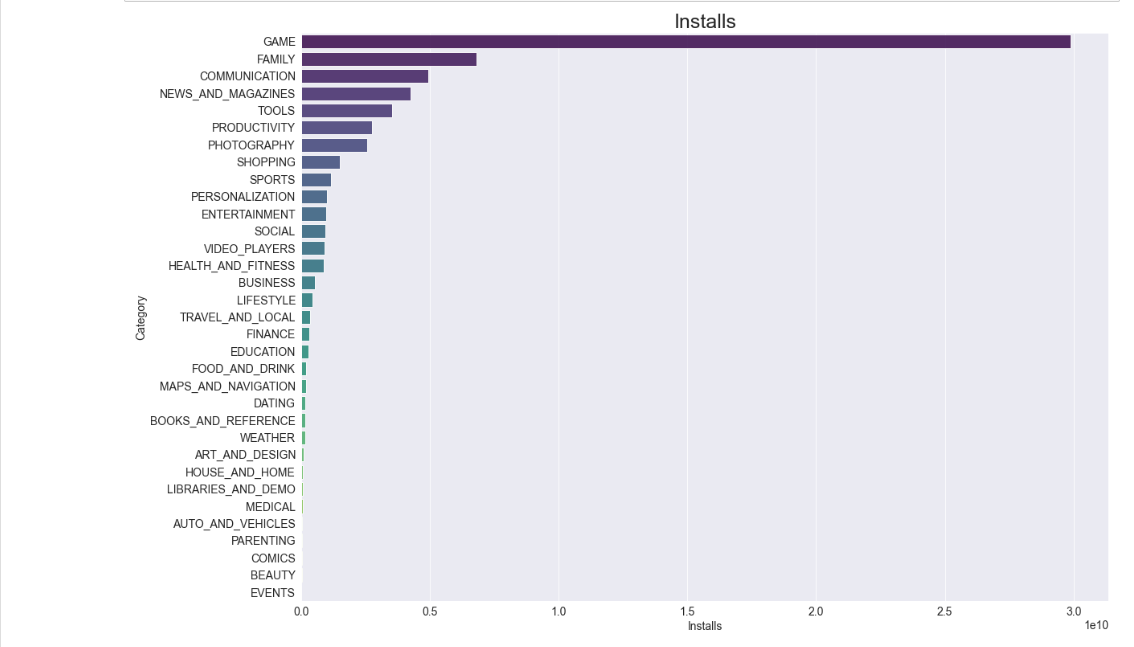
From the above output we can conclude that in Play Store most of the apps are under Family & Game category and least are of Beauty & Comics Category.



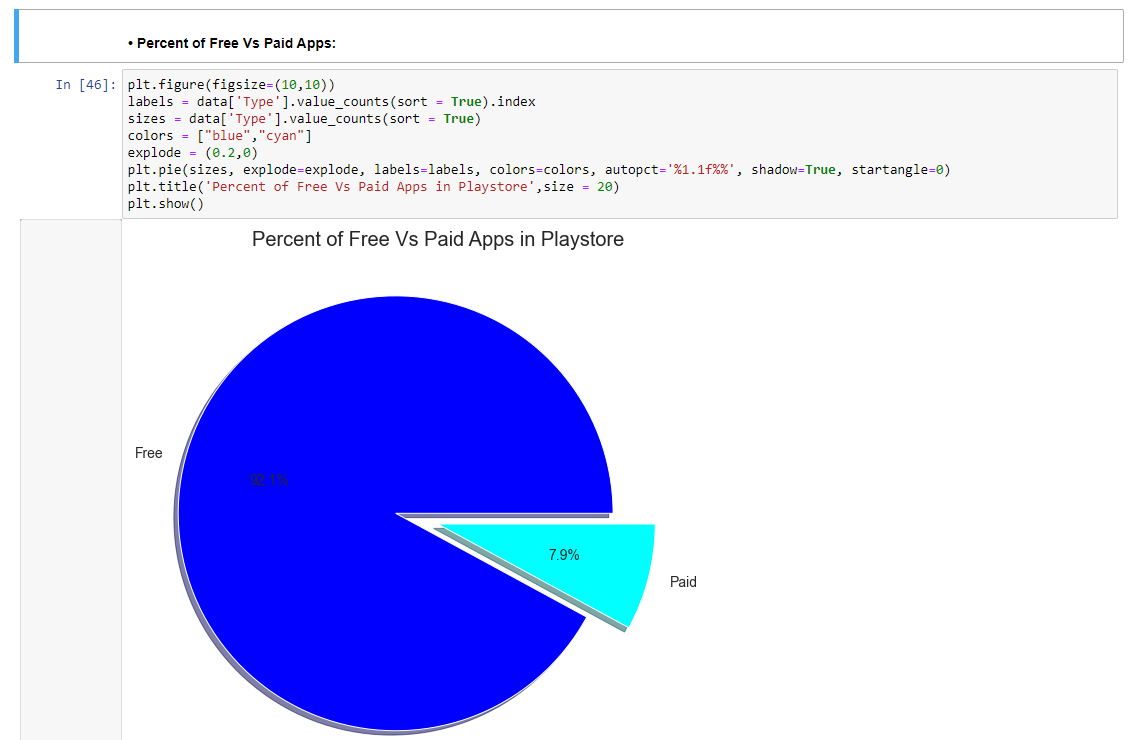


From the above graph we can come to a conclusion that most of the apps in google Play Store are rated in between 3.5 to 4.8.

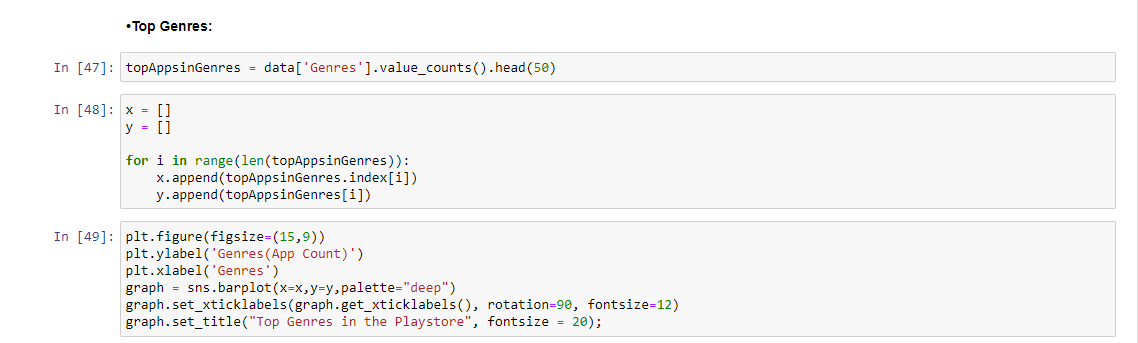


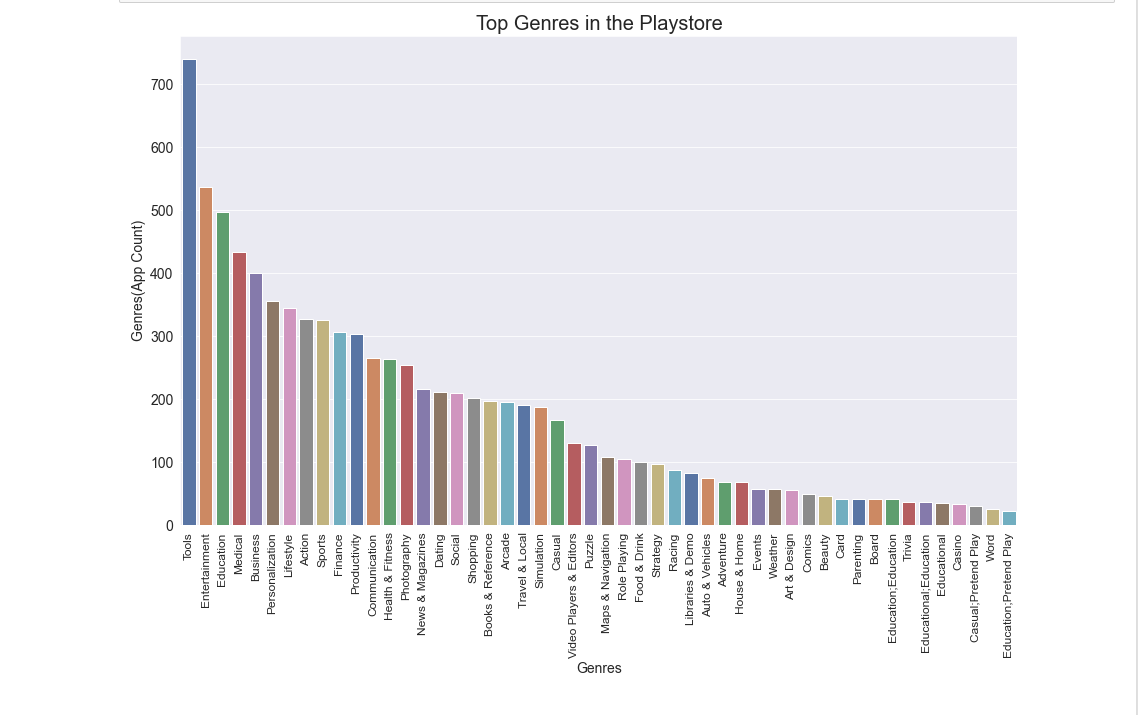


From the above visualization, it can be interpreted that the top categories with highest installs are Game, Family, Communication, News & Magazines, & Tools.



From the above graph we can see that 92%(Approx.) of apps in google play store are free and 8%(Approx.) are paid.





From the above visualization we can see that the Highest Number of Apps found in the Tools and Entertainment genres followed by Education, Medical and many more.

REFERENCES

1. Jiawei Han, Micheline Kamber, Jian Pei , Data Mining: Concepts and Techniques, 3/e, Morgan Kaufmann publishers, 2011.
2. <https://www.kaggle.com/datasets/lava18/google-play-store-apps>
3. <https://www.irjet.net/archives/V7/i12/IRJET-V7I1248.pdf>
4. <https://www.geeksforgeeks.org/graph-plotting-in-python-set-1/>