25 March Assignment

June 14, 2023

```
[ ]: Q1. Load the flight price dataset and examine its dimensions. How many rows and␣
      ↪columns does the
     dataset have?
```

ANS -

```
[23]: import pandas as pd
      import seaborn as sns
      import numpy as np
      import matplotlib.pyplot as plt
      %matplotlib inline
```

```
[24]: df = pd.read_excel('flight_price.xlsx')
```

```
[25]: df.head()
```

```
[25]:        Airline Date_of_Journey    Source Destination                  Route  \
      0       IndiGo      24/03/2019  Banglore   New Delhi              BLR → DEL
      1    Air India       1/05/2019   Kolkata    Banglore  CCU → IXR → BBI → BLR
      2  Jet Airways       9/06/2019     Delhi      Cochin  DEL → LKO → BOM → COK
      3       IndiGo      12/05/2019   Kolkata    Banglore        CCU → NAG → BLR
      4       IndiGo      01/03/2019  Banglore   New Delhi        BLR → NAG → DEL

        Dep_Time  Arrival_Time Duration Total_Stops Additional_Info  Price
      0    22:20  01:10 22 Mar   2h 50m    non-stop         No info   3897
      1    05:50         13:15   7h 25m     2 stops         No info   7662
      2    09:25  04:25 10 Jun      19h     2 stops         No info  13882
      3    18:05         23:30   5h 25m      1 stop         No info   6218
      4    16:50         21:35   4h 45m      1 stop         No info  13302
```

```
[16]: df.shape
```

```
[16]: (10683, 11)
```

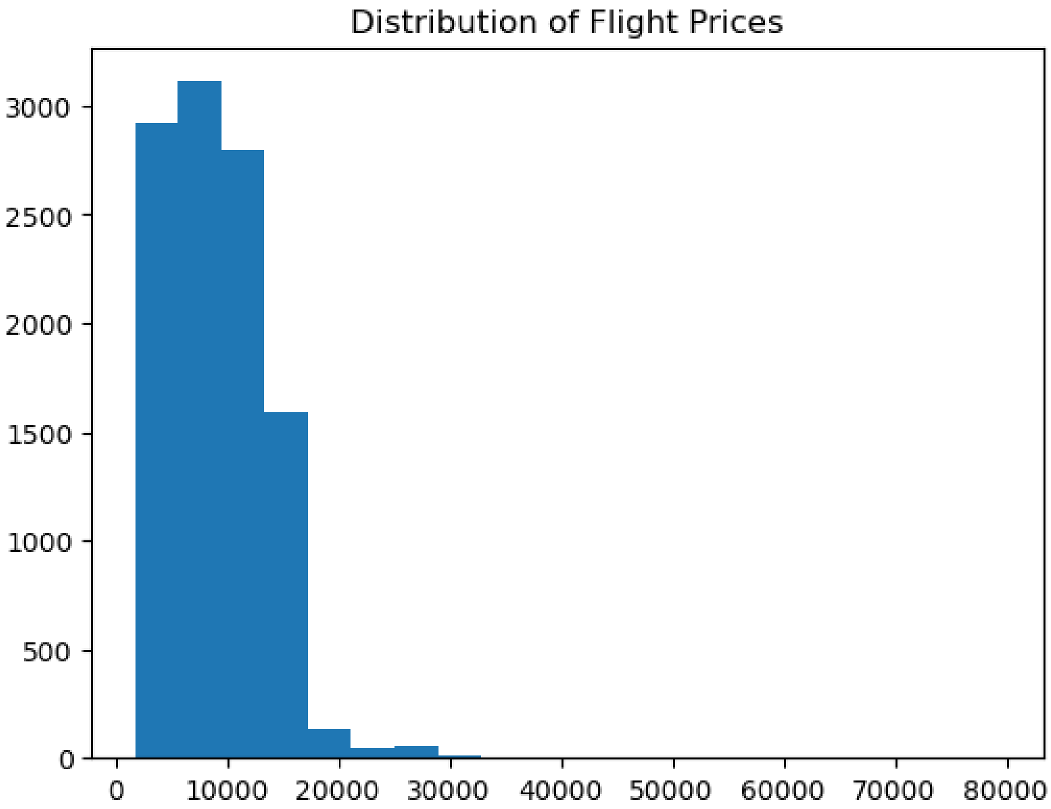
```
[ ]:
```

```
[ ]: Q2. What is the distribution of flight prices in the dataset? Create a␣
      ↪histogram to visualize the
```

```
distribution.
```

ANS -

```
[19]: df = pd.read_excel('flight_price.xlsx')
      plt.hist(df['Price'], bins=20)
      plt.title('Distribution of Flight Prices')
      plt.show()
```



```
[ ]:
```

```
[ ]: Q3. What is the range of prices in the dataset? What is the minimum and maximum␣
       ↪price?
```

ANS -

```
[21]: df.head()
```

```
[21]:        Airline Date_of_Journey    Source Destination                  Route  \
      0       IndiGo      24/03/2019  Banglore   New Delhi              BLR → DEL
      1    Air India       1/05/2019   Kolkata    Banglore  CCU → IXR → BBI → BLR
      2  Jet Airways       9/06/2019     Delhi      Cochin  DEL → LKO → BOM → COK
```

```
3       IndiGo      12/05/2019   Kolkata    Banglore        CCU → NAG → BLR
4       IndiGo      01/03/2019  Banglore   New Delhi        BLR → NAG → DEL

  Dep_Time  Arrival_Time Duration Total_Stops Additional_Info  Price
0    22:20  01:10 22 Mar   2h 50m    non-stop         No info   3897
1    05:50         13:15   7h 25m     2 stops         No info   7662
2    09:25  04:25 10 Jun      19h     2 stops         No info  13882
3    18:05         23:30   5h 25m      1 stop         No info   6218
4    16:50         21:35   4h 45m      1 stop         No info  13302
```

```
[26]: df.describe()
```

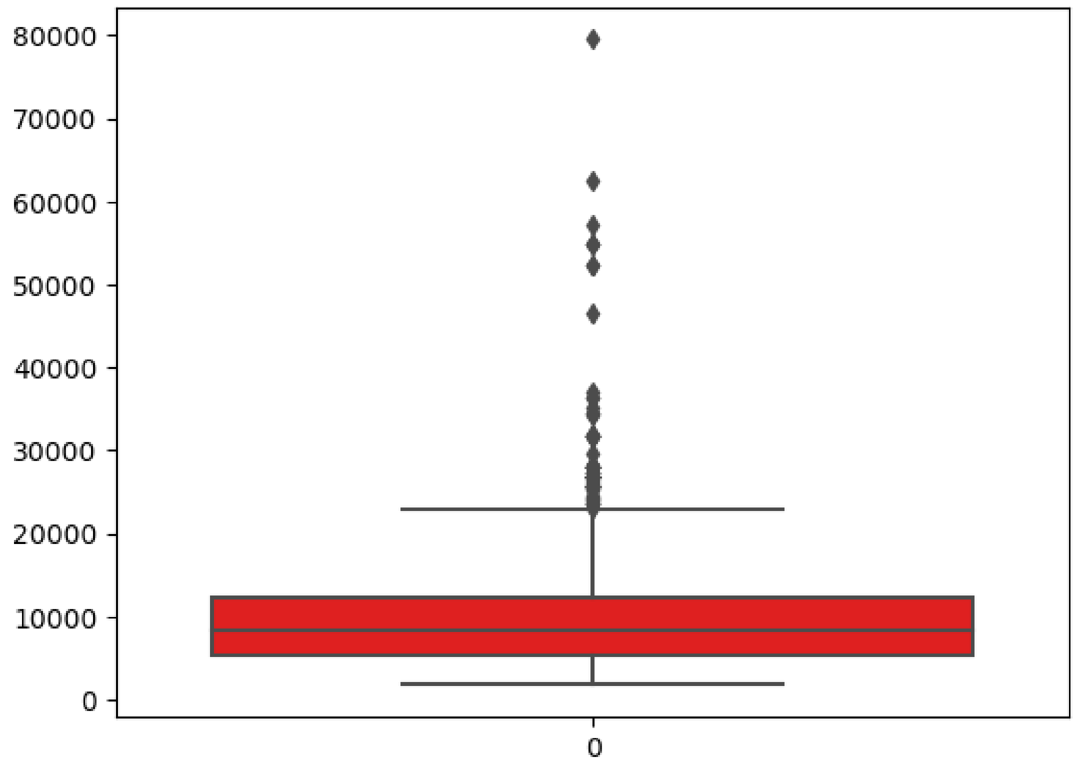
```
[26]:               Price
      count  10683.000000
      mean    9087.064121
      std     4611.359167
      min     1759.000000
      25%     5277.000000
      50%     8372.000000
      75%    12373.000000
      max    79512.000000
```

```
[ ]:
```

```
[ ]: Q4. How does the price of flights vary by airline? Create a boxplot to compare␣
      ↪the prices of different
     airlines.
```

ANS -

```
[26]:
```



[ ]:

[ ]:

```
Q5. Are there any outliers in the dataset? Identify any potential outliers␣
 ↪using a boxplot and describe how
they may impact your analysis.
```

ANS -

[ ]:

[ ]:

[ ]:

```
Q6. You are working for a travel agency, and your boss has asked you to analyze␣
 ↪the Flight Price dataset
to identify the peak travel season. What features would you analyze to identify␣
 ↪the peak season, and how
would you present your findings to your boss?
```

ANS -

[ ]:

```
To identify the peak travel season from the Flight Price dataset, you can␣
 ↪analyze features such as historical ticket price data, ticket
```

```
purchase date and departure date, season, holidays, supply (number of available␣
  ↪airlines and flights), fare class, availability of seats,
recent market demand and flight distance.

You can present your findings in a report format with graphs and charts that␣
  ↪show the trends in ticket prices over time. You can also include
a summary of your findings and recommendations for your boss based on the data.
```

[ ]:

```
```

[ ]:

```
Q7. You are a data analyst for a flight booking website, and you have been␣
  ↪asked to analyze the Flight
Price dataset to identify any trends in flight prices. What features would you␣
  ↪analyze to identify these
trends, and what visualizations would you use to present your findings to your␣
  ↪team?
```

ANS -

[ ]:

```
To identify trends in flight prices, I would analyze the following features:.

1.Date: I would analyze the flight prices over time to identify any seasonal␣
  ↪trends.

2.Departure and Arrival Cities: I would analyze the flight prices for different␣
  ↪departure and arrival cities to identify any regional trends.

3.Airline: I would analyze the flight prices for different airlines to identify␣
  ↪any pricing trends.

4.Flight Duration: I would analyze the flight prices for different flight␣
  ↪durations to identify any pricing trends.

To present my findings to my team, I would use the following visualizations:.

Line Chart: To visualize the trend in flight prices over time.
Bar Chart: To visualize the average flight prices for different departure and␣
  ↪arrival cities, airlines, and flight durations.
```

[ ]:

```
```

[ ]:

```
Q8. You are a data scientist working for an airline company, and you have been␣
  ↪asked to analyze the
Flight Price dataset to identify the factors that affect flight prices. What␣
  ↪features would you analyze to
identify these factors, and how would you present your findings to the␣
  ↪management team?
```

ANS -

[ ]:
```
To identify the factors that affect flight prices, I would analyze the␣
  ↪following features:

1.Departure and Arrival Cities: The prices of flights can vary depending on the␣
  ↪cities of departure and arrival. For example, flights to
popular tourist destinations may be more expensive than flights to less popular␣
  ↪destinations.

2.Departure and Arrival Dates: The prices of flights can also vary depending on␣
  ↪the dates of departure and arrival. For example, flights
during peak travel seasons may be more expensive than flights during off-peak␣
  ↪seasons.

3.Airline: The prices of flights can also vary depending on the airline. Some␣
  ↪airlines may offer cheaper flights than others.

4.Flight Duration: The prices of flights can also vary depending on the␣
  ↪duration of the flight. Longer flights may be more expensive than
shorter flights.

5.Number of Stops: The prices of flights can also vary depending on the number␣
  ↪of stops. Flights with more stops may be cheaper than non-stop
flights.

To present my findings to the management team, I would create a report that␣
  ↪includes visualizations such as graphs and charts to help them
understand the data better. I would also provide a summary of my findings and␣
  ↪recommendations based on my analysis.
```

[ ]:
```

```

##Google Playstore:

[ ]:
```
Q9. Load the Google Playstore dataset and examine its dimensions. How many rows␣
  ↪and columns does
the dataset have?
```

ANS -

[40]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

[41]:

```
df= pd.read_csv('https://raw.githubusercontent.com/krishnaik06/
↪playstore-Dataset/main/googleplaystore.csv')
df.head()
```

[41]:

```
                                                 App        Category  Rating  \
0     Photo Editor & Candy Camera & Grid & ScrapBook  ART_AND_DESIGN     4.1
1                                Coloring book moana  ART_AND_DESIGN     3.9
2  U Launcher Lite - FREE Live Cool Themes, Hide …  ART_AND_DESIGN     4.7
3                              Sketch - Draw & Paint  ART_AND_DESIGN     4.5
4              Pixel Draw - Number Art Coloring Book  ART_AND_DESIGN     4.3

  Reviews  Size     Installs  Type Price Content Rating  \
0     159   19M      10,000+  Free     0       Everyone
1     967   14M     500,000+  Free     0       Everyone
2   87510  8.7M   5,000,000+  Free     0       Everyone
3  215644   25M  50,000,000+  Free     0           Teen
4     967  2.8M     100,000+  Free     0       Everyone

                     Genres      Last Updated         Current Ver  \
0              Art & Design   January 7, 2018               1.0.0
1  Art & Design;Pretend Play  January 15, 2018               2.0.0
2              Art & Design    August 1, 2018               1.2.4
3              Art & Design      June 8, 2018  Varies with device
4    Art & Design;Creativity     June 20, 2018                 1.1

    Android Ver
0  4.0.3 and up
1  4.0.3 and up
2  4.0.3 and up
3    4.2 and up
4    4.4 and up
```

[5]:

```
df.shape
```

[5]:

```
(10841, 13)
```

[ ]:

```

```

[6]:

```
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
```

[ ]:

```

```

[ ]:

```
Q10. How does the rating of apps vary by category? Create a boxplot to compare␣
 ↪the ratings of different
app categories.
```

ANS -

[ ]:

```

```

[ ]:

```
Q11. Are there any missing values in the dataset? Identify any missing values␣
 ↪and describe how they may
impact your analysis.
```

ANS-

[20]:

```
df.isnull().sum()
```

[20]:

```
App                  0
Category             0
Rating            1474
Reviews              0
Size                 0
Installs             0
Type                 1
Price                0
Content Rating       1
Genres               0
Last Updated         0
Current Ver          8
Android Ver          3
dtype: int64
```

[ ]:

```
    Missing values in a dataset can impact the performance of the model by␣
 ↪creating a bias in the dataset. This bias can create a lack of
relatability and trustworthiness in the dataset. The loss in values might␣
 ↪contain crucial insights or information for model development.
Missing values in datasets can cause complications in data handling and␣
 ↪analysis, loss of information and efficiency, and can produce biased
results. You can drop the data with missing values or impute them with mean,␣
 ↪median, or most frequently occurring values or by other
statistical methods.
```

[ ]:

```

```

[ ]:

```
Q12. What is the relationship between the size of an app and its rating? Create␣
 ↪a scatter plot to visualize
the relationship.
```

ANS -

[ ]:

```

```

[ ]:

```

```

[ ]:

```
Q13. How does the type of app affect its price? Create a bar chart to compare␣
 ↪average prices by app type.
```

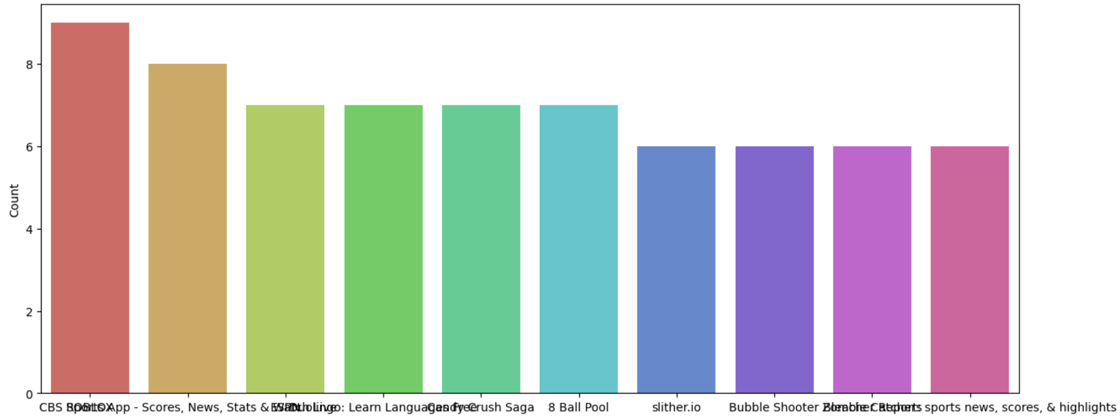
ANS -

[69]:

```
plt.figure(figsize=(15,6))
sns.barplot(x=App.index[:10], y ='Count',data = App[:10],palette='hls')
```

[69]:

```
<AxesSubplot: ylabel='Count'>
```



[ ]:

```

```

```
[ ]: Q14. What are the top 10 most popular apps in the dataset? Create a frequency␣
      ↪table to identify the apps
     with the highest number of installs.
```

ANS-

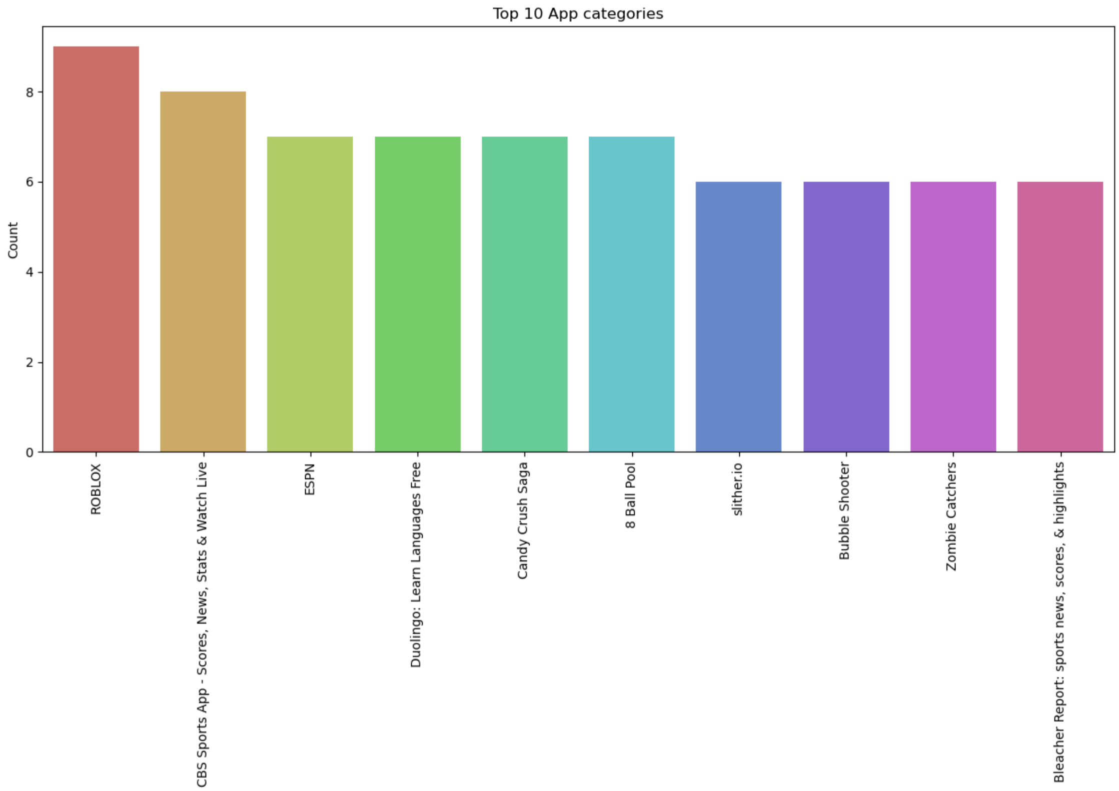
```
[60]: App = pd.DataFrame(df['App'].value_counts())        #Dataframe of apps on the␣
       ↪basis of categ
      App.rename(columns = {'App':'Count'},inplace=True)
```

```
[61]: App
```

```
[61]:                                                      Count
      ROBLOX                                                   9
      CBS Sports App - Scores, News, Stats & Watch Live        8
      ESPN                                                     7
      Duolingo: Learn Languages Free                           7
      Candy Crush Saga                                         7
      …                                                    …
      Meet U - Get Friends for Snapchat, Kik & Instagram       1
      U-Report                                                 1
      U of I Community Credit Union                            1
      Waiting For U Launcher Theme                             1
      iHoroscope - 2018 Daily Horoscope & Astrology            1

      [9660 rows x 1 columns]
```

```
[62]: plt.figure(figsize=(15,6))
      sns.barplot(x=App.index[:10], y ='Count',data = App[:10],palette='hls')
      plt.title('Top 10 App categories')
      plt.xticks(rotation=90)
      plt.show()
```



[ ]:
```

```

[ ]:
```
Q15. A company wants to launch a new app on the Google Playstore and has asked␣
  ↪you to analyze the
Google Playstore dataset to identify the most popular app categories. How would␣
  ↪you approach this
task, and what features would you analyze to make recommendations to the␣
  ↪company?
```

ANS-

[ ]:
```
To analyze the Google Playstore dataset to identify the most popular app␣
  ↪categories, you can start by preprocessing the data to clean raw data
into assorted data that is ready for use. After that, you can group the data by␣
  ↪category and count the number of apps in each category. This
will give you an idea of which categories are most popular. For instance, in␣
  ↪one analysis, Family was found to be the category with the
highest number of applications1. You can also use sentiment analysis to see how␣
  ↪the sentiments comport as you go down through the popularity
rankings. Another way is to plot some graphs against different specifications␣
  ↪of an application. For example, you can divide the apps into
```

```
categories and then plot the number of apps in each category to explore the␣
  ↪most popular category among apps on the play store.
```

[ ]:

```

```

[ ]:

```
Q16. A mobile app development company wants to analyze the Google Playstore␣
  ↪dataset to identify the
most successful app developers. What features would you analyze to make␣
  ↪recommendations to the
company, and what data visualizations would you use to present your findings?
```

ANS-

[ ]:

```
To identify the most successful app developers in the Google Playstore dataset,␣
  ↪you can analyze the following features:

1.Category: The category of the app can help identify which categories are most␣
  ↪popular and which ones have the highest number of downloads.
2.Rating: The rating of an app can help identify which apps are most popular␣
  ↪among users.
3.Reviews: The number of reviews an app has received can help identify how␣
  ↪popular an app is among users.
4.Size: The size of an app can help identify how much storage space an app␣
  ↪takes up on a user’s device.
5.Installs: The number of installs an app has received can help identify how␣
  ↪popular an app is among users.
6.Price: The price of an app can help identify which apps are most popular␣
  ↪among users.

To present your findings, you can use data visualizations such as:

1.Bar charts: Bar charts can be used to show the number of downloads for each␣
  ↪category.
2.Scatter plots: Scatter plots can be used to show the relationship between the␣
  ↪rating and the number of reviews for each app.
3.Pie charts: Pie charts can be used to show the percentage of apps in each␣
  ↪category.
```

[ ]:

```

```

[ ]:

```
Q17. A marketing research firm wants to analyze the Google Playstore dataset to␣
  ↪identify the best time to
launch a new app. What features would you analyze to make recommendations to␣
  ↪the company, and
what data visualizations would you use to present your findings?
```

ANS-

[ ]:

```
To make recommendations to the company, you can analyze the following features of the Google Playstore dataset:

1.Category: The category of the app can help identify the most popular app categories and the ones that have the most competition.
2.Rating: The rating of an app can help identify how well it is received by users.
3.Reviews: The number of reviews can help identify how popular an app is.
4.Price: The price of an app can help identify how much users are willing to pay for an app.
5.Size: The size of an app can help identify how much storage space users are willing to allocate for an app.
6.Content Rating: The content rating of an app can help identify which age groups are most interested in the app.

To present your findings, you can use data visualizations such as:

1.Bar charts: To compare different categories and their popularity.
2.Scatter plots: To show the relationship between two variables such as price and rating.
3.Heat maps: To show the distribution of apps across different categories and their ratings.
```