BUSINESS CASE

Presented by: Erick Cuevas

PROCESSING DATA

Cleansing

- Pandas library
- Obtain info of the dataset (size, columns, data type)
- Imputation of null values
- Drop duplicates
- Drop data mismatch values / errors
- Remove special characters

```
#Obtain a first glance / initial info of the dataset
df2.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
```

Transforming

- Pandas/Numpy libraries
- Merging both datasets
- Convert data type of key columns/variables
- Functions: sort_values, groupby, nlargest, nsmallest, reset_index
- Obtain relevant information from the datasets

```
#Group the registers by category and obtain the size of each one
df2_cat_counts = df2_no_dups.groupby('Category').size()
df2_cat_counts.sort_values(ascending=False)
```

```
#Obtain the percentage of each category based on the size
#Create a dataset with that info

df2_cat_perc = df2_cat_counts.reset_index()
total_cat = df2_cat_perc[0].sum()
df2_cat_perc['Percentage'] = (df2_cat_perc[0]/total_cat)*100
df2_cat_perc.sort_values(by=0,ascending=False)
```

Plotting

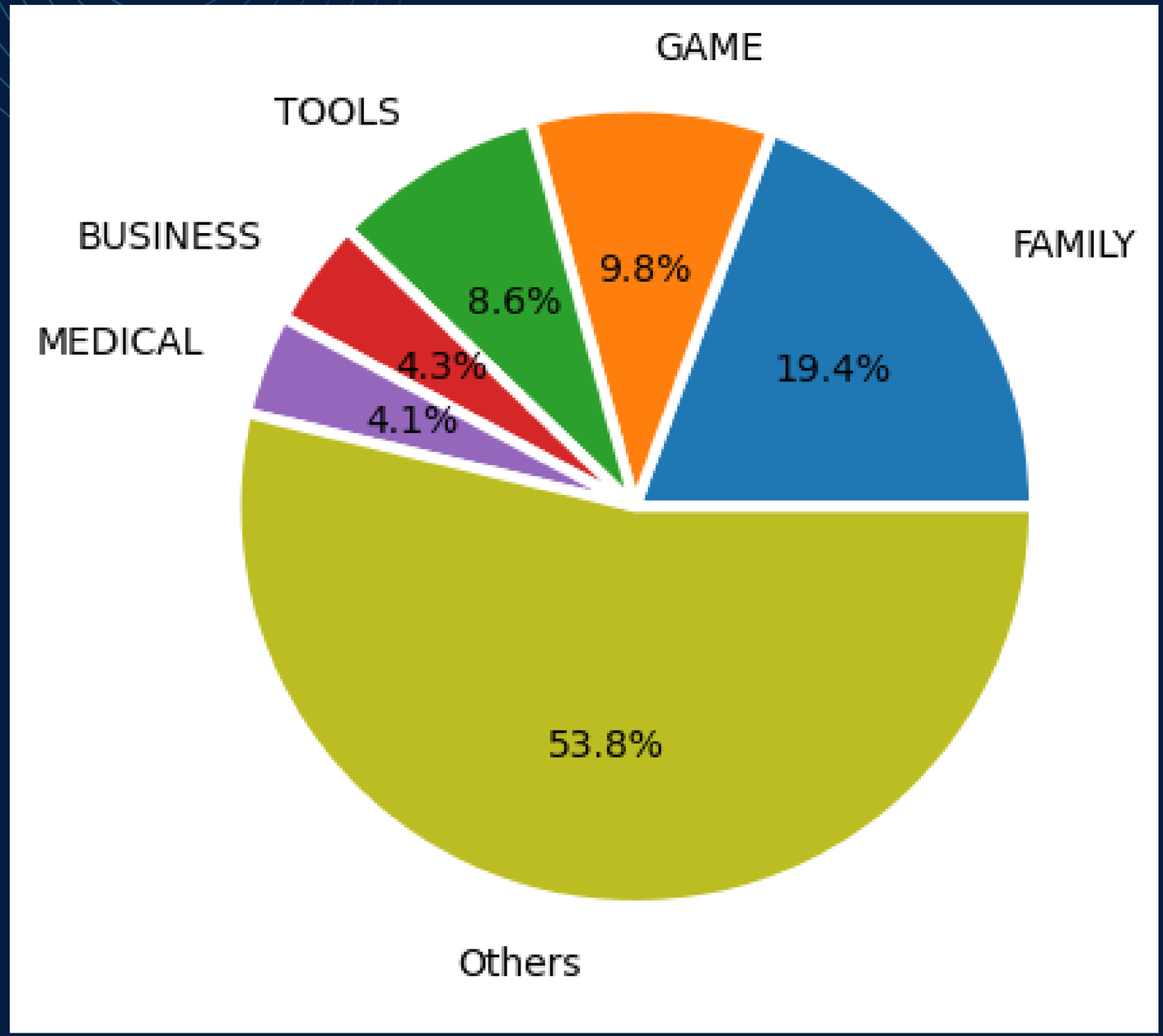
- Matplotlib library
- Creation of visualizations
- Bar Charts & Pie Charts
- Use of for loops due to given approach

```
#Create a loop to obtain a bar chart with the top 5 ratings for filtered datasets with diff number of reviews

filtered_datasets = [df2_no_dups, filtered_df_10M, filtered_df_1M, filtered_df_500k, filtered_df_100k]
colors = ['#1f77b4', '#ff7f0e', '#2ca02c', '#d62728', '#9467bd']

for j, color in zip(filtered_datasets,colors):
  top_5 = j.nlargest(5, 'Rating')
  #top_5 = j.nlargest(5, 'Rating',keep='all')

  plt.barh(top_5['App'], top_5['Rating'], color=color)

  for i, value in enumerate(top_5['Rating']):
    plt.text(value, i, str(value), color= 'black')

  plt.xlabel('Rating')
  plt.ylabel('Apps')
  plt.title('Top 5 Rated Apps')
  plt.show()
```

TOP 5 App Categories

Others

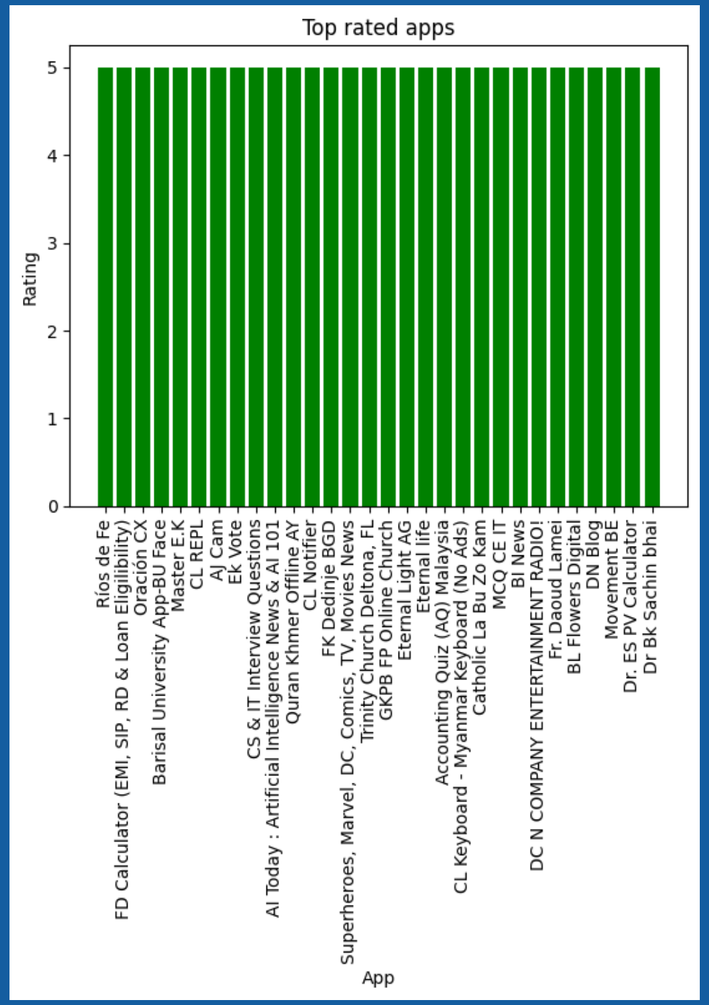
PERSONALIZATION	376	3.892743
PRODUCTIVITY	374	3.872036
LIFESTYLE	369	3.820271
FINANCE	345	3.571798
SPORTS	325	3.364738
COMMUNICATION	315	3.261207
HEALTH_AND_FITNESS	288	2.981675
PHOTOGRAPHY	281	2.909204



TOP 5 RATED APPS

Not possible to obtain the top 5 apps with the best rating

	App	Category	Rating	Reviews
10407	Jigsaw Volvo FH 16 Trucks	FAMILY	5.0	5
7100	CA Speakers	LIFESTYLE	5.0	12
10564	FK Dedinje BGD	SPORTS	5.0	36
8014	Morse Player	FAMILY	5.0	12
8526	DL Image Manager	PRODUCTIVITY	5.0	2
6043	Exam Result BD	FAMILY	5.0	2
8601	DN Calculators	FINANCE	5.0	12
7533	Color CL	LIFESTYLE	5.0	2
8888	Spring flowers theme couleurs d t space	ART_AND_DESIGN	5.0	1
7517	CL Notifier	TOOLS	5.0	36

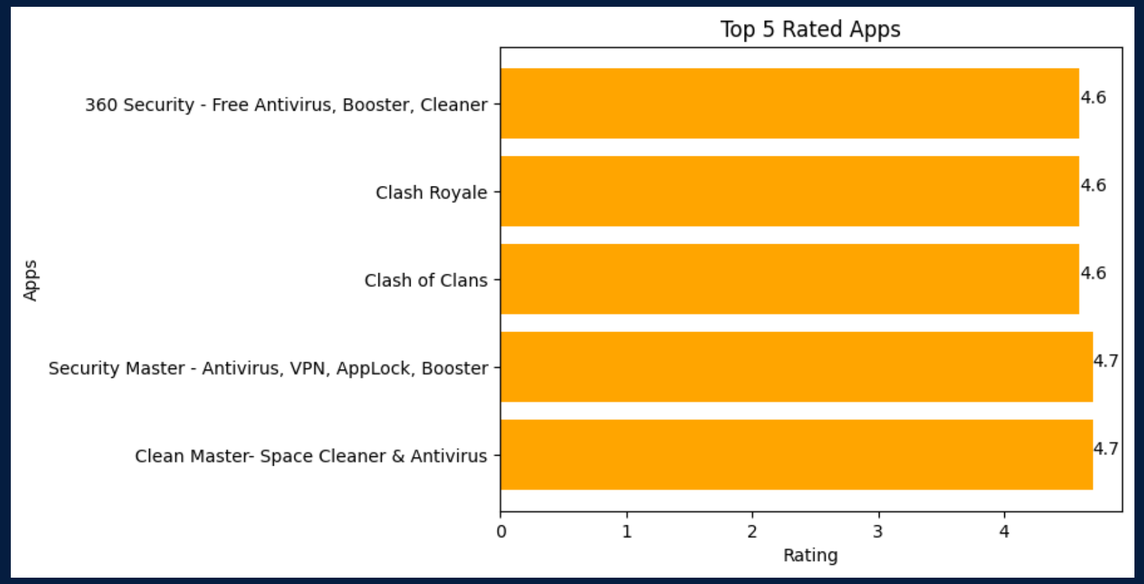


There are *271 registers/apps* with the same "best" rating of 5.0

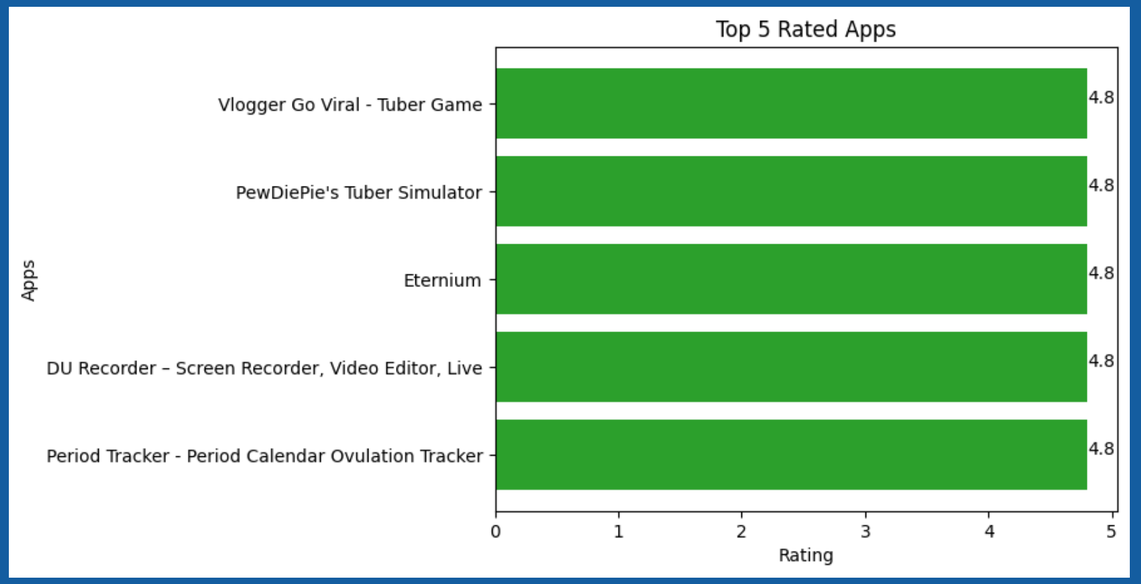
Best approach would be filtering the apps with a larger number of reviews

TOP 5 RATED APPS

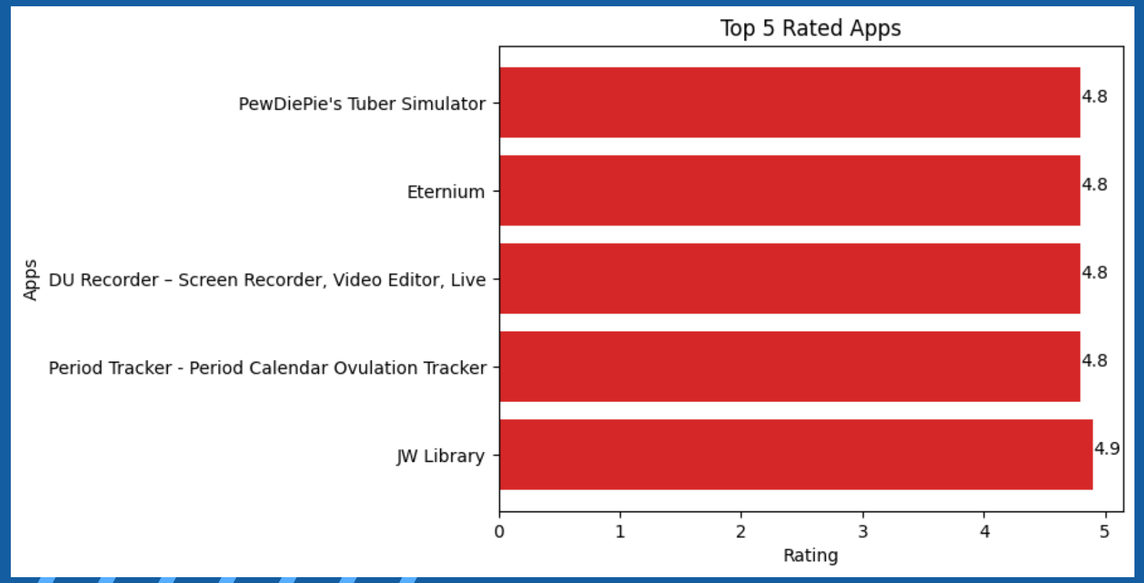
+10M Reviews



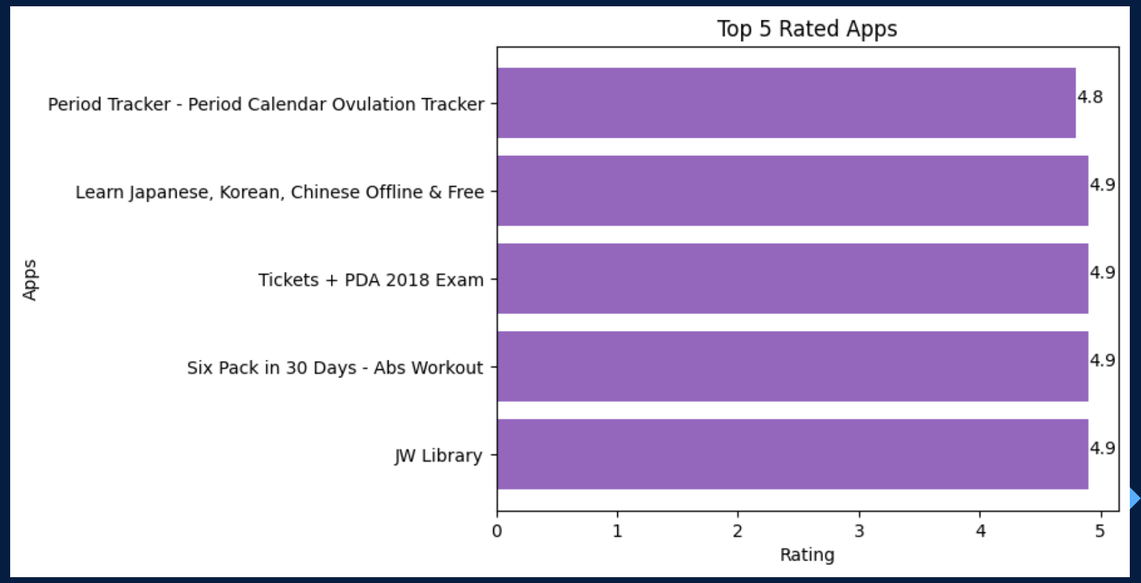
+1M Reviews



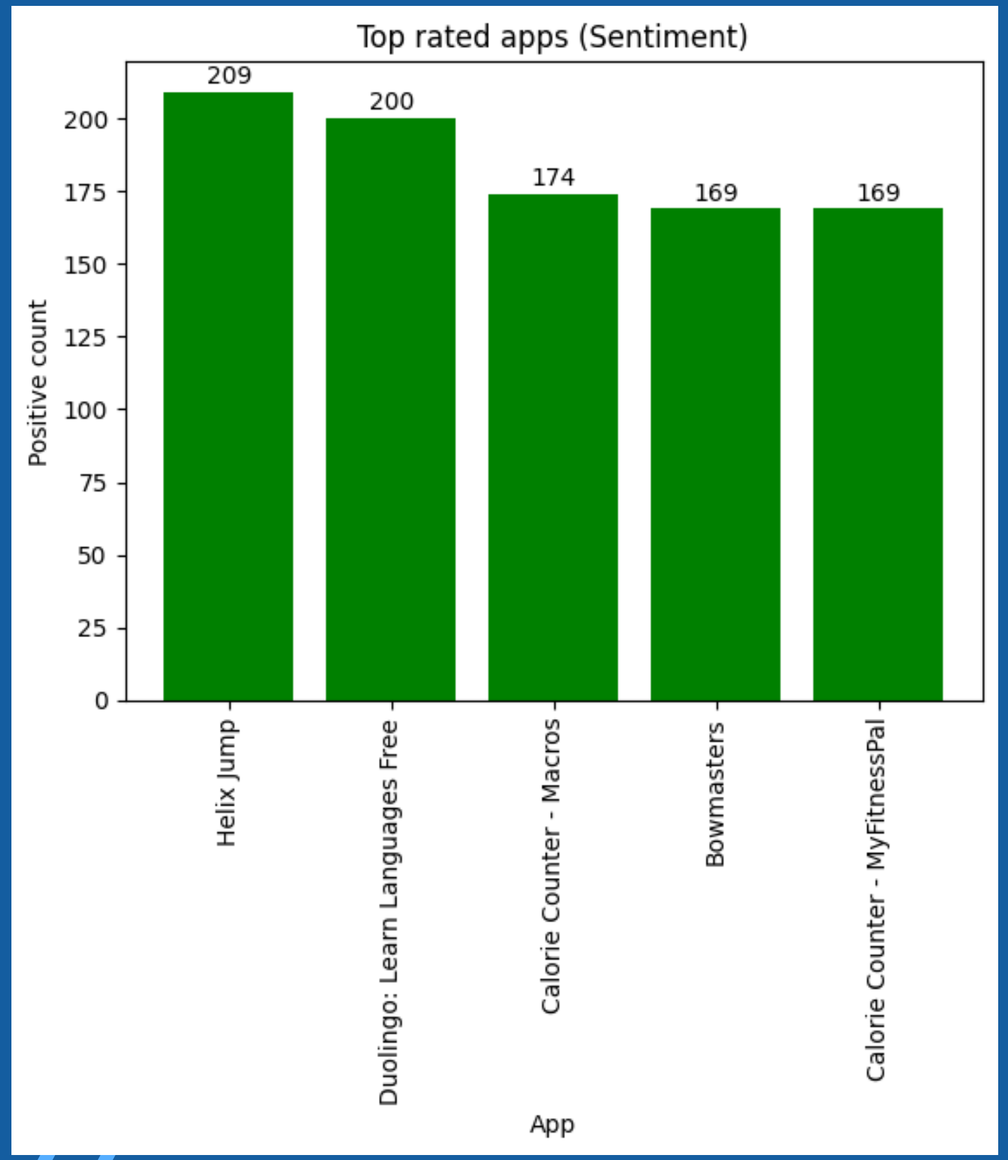
+500k Reviews



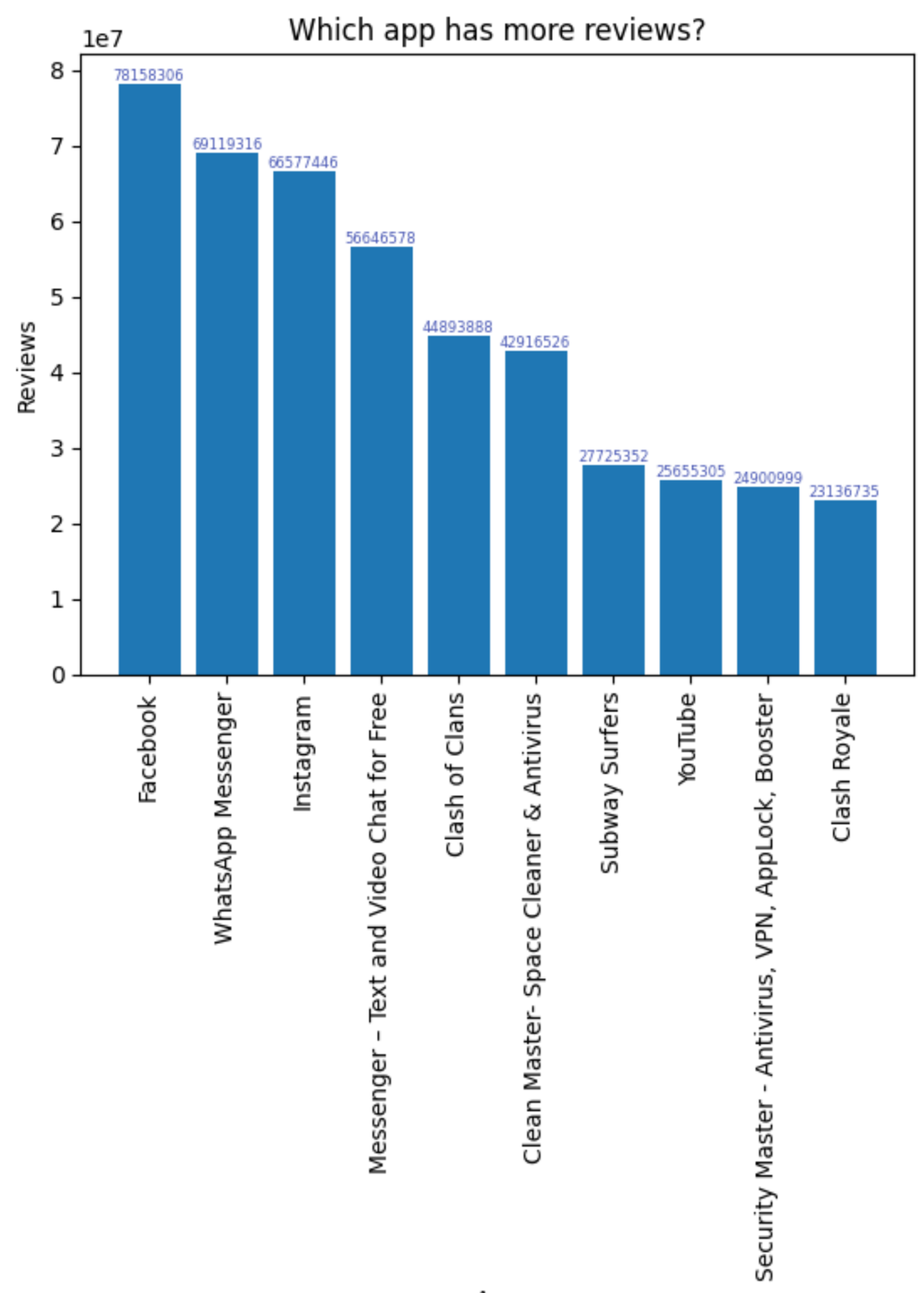
+100k Reviews



TOP 5 RATED APPS (SENTIMENT)



Based on the positive count in the column "Sentiment" from the User Reviews file



WHICH APP HAS MORE REVIEWS?

78,158,306

69,119,316

66,577,446

WHICH APP IS THE LEAST LIKED BY THE USERS?

	App	Category	Rating	Reviews
6329	Basic Strategy Training BJ 21	GAME	0.0	0
6819	BU Questrom Launch	TRAVEL_AND_LOCAL	0.0	3
6808	BT Speed	SPORTS	0.0	3
7702	CP Cloud	BUSINESS	0.0	3
9504	Ek Biladi Jadi Video Song	FAMILY	0.0	3
8575	DM Collection	SHOPPING	0.0	3
5110	Wind & Weather Meter for Ag	WEATHER	0.0	3
5120	West Central Ag	BUSINESS	0.0	3
5119	AG test	TOOLS	0.0	3
6838	BU Alsace	BOOKS_AND_REFERENCE	0.0	3
6843	Bu Hangi Film ?	FAMILY	0.0	3
5116	AG EMS Tour	FAMILY	0.0	3
9874	European Union Flag LWP	PERSONALIZATION	0.0	3
5927	Learn Quran with Elif Ba	FAMILY	0.0	3
5841	Ay Yıldız Duvar Kağıtları	PERSONALIZATION	0.0	3
7897	CT and XR Dose Calculator	MEDICAL	0.0	3
6164	BG Television	FAMILY	0.0	3
10365	Radio FG Paris Underground	FAMILY	0.0	3
5213	Artificial intelligence	FAMILY	0.0	3
7503	CL 2ne1 Wallpaper KPOP HD Best	PERSONALIZATION	0.0	3

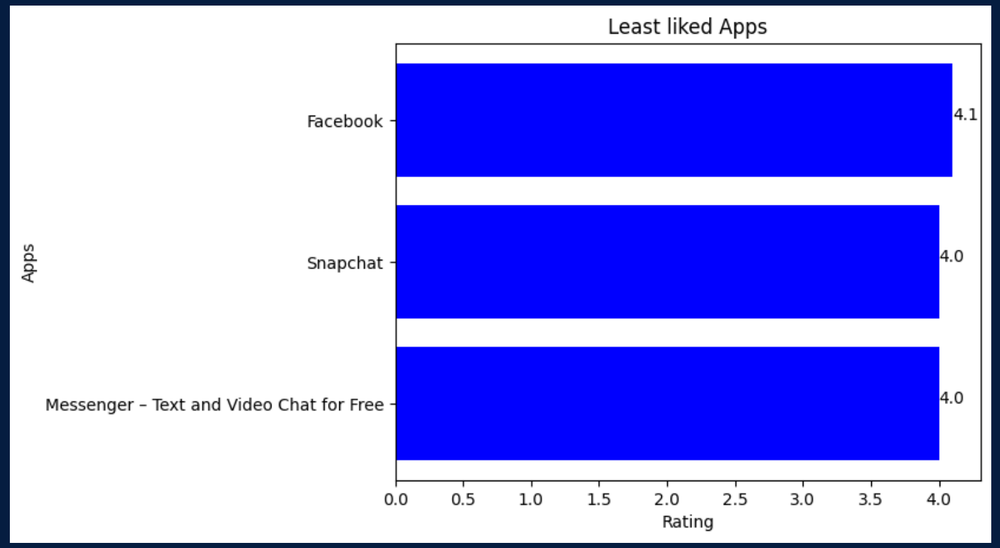
1463 registers/apps with a rating = 0

Following the same approach previously used =

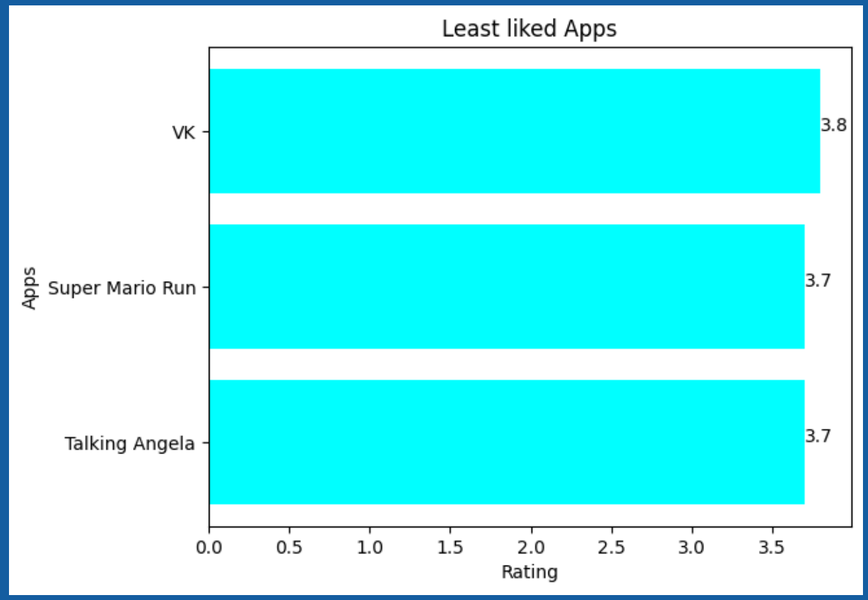
Filter apps with greater number of reviews

WHICH APP IS THE LEAST LIKED BY THE USERS?

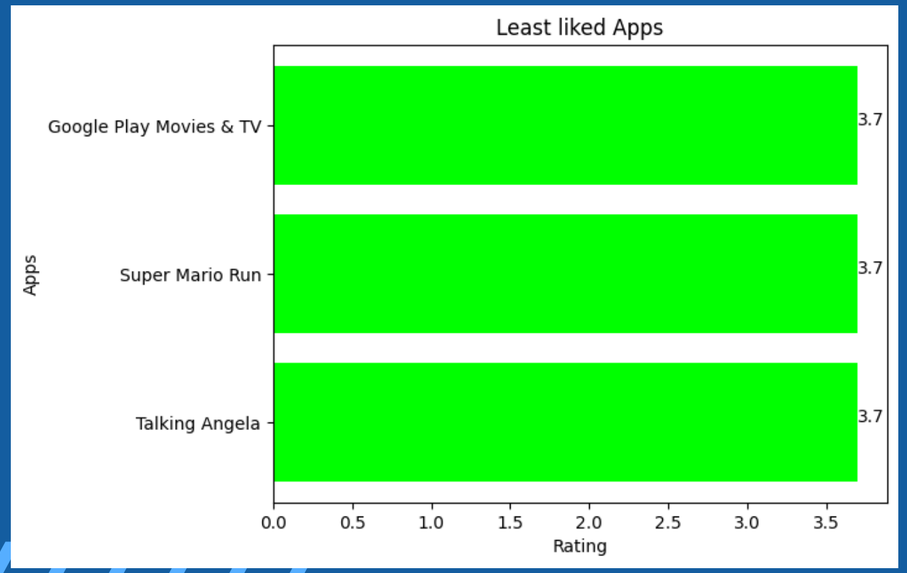
+10M Reviews



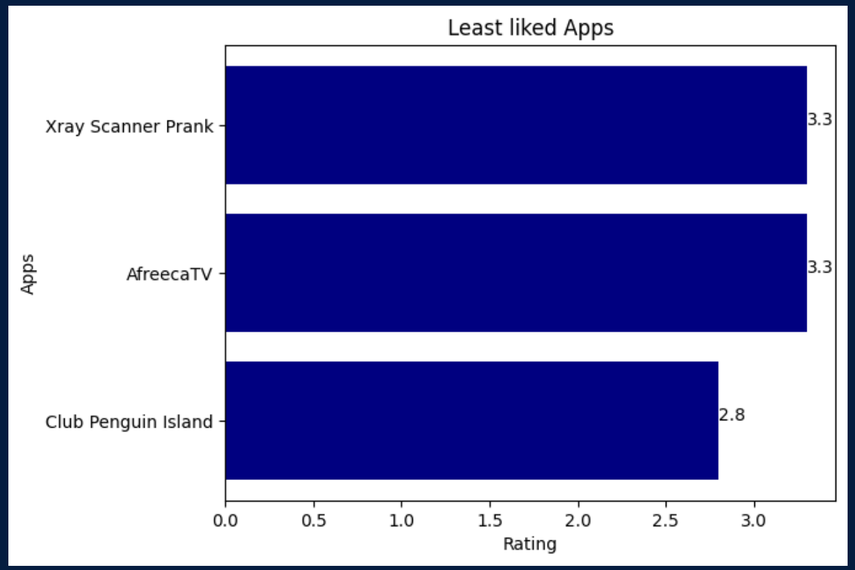
+1M Reviews



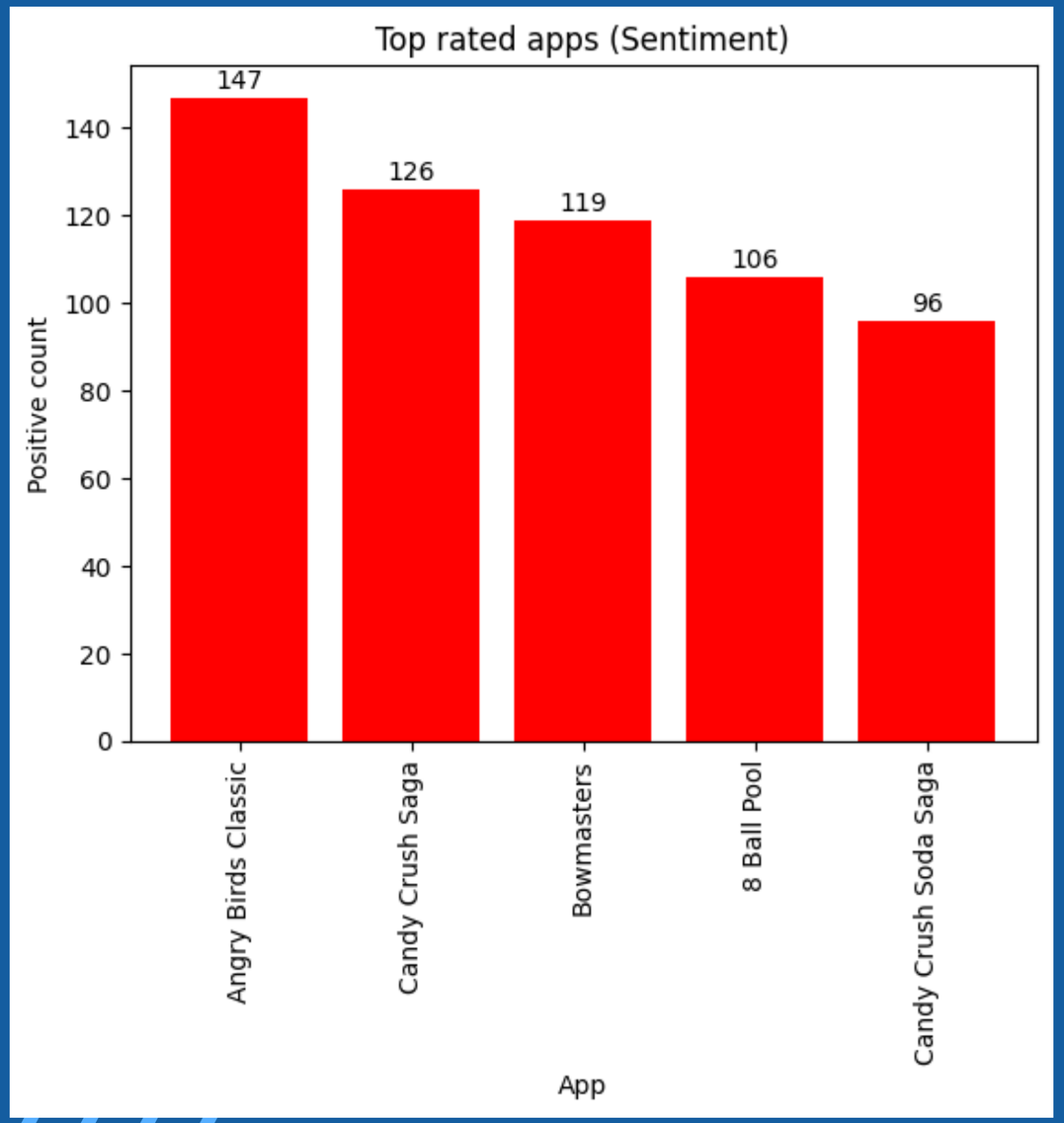
+500k Reviews



+100k Reviews



WHICH APP IS THE LEAST LIKED BY THE USERS?
(SENTIMENT)



Based on the negative count in the column "Sentiment" from the User Reviews file

Had you more data, what would you like to further investigate?

Correlation

Is there any statistical relationship between certain numeric variables? E.G. Does the price affect the rating?

Regression

Is it possible to generate a regression model that can predict the rating given to an app based on other numeric variables? (price, size, etc.)

Clustering

It would be great for google to segment their customers so that they may develop unique marketing strategies for each group for example. This with aid of Clustering algorithms and ML.

What approach would you take with the business to improve the report?

- Identify the most relevant KPIs for the business. E.G. sales, revenue, profit margins.
- Data Analysis using coding tools such as Python, SQL and R.
- Use data visualization tools like Tableau, Power BI, or even Python libraries to create interactive dashboards.
- Implement automated reporting systems that generate and distribute reports on a regular basis.

Strategy to gain the confidence of the business to make changes to future reporting

- Identify the most used tools for reporting & data visualization by benchmarking.
- Get involved in all areas that collects relevant data.
- Get an idea of how data is processed and managed.
- Learn more about the current ETL process.

What advantages could you bring to the business with enhanced reporting?

- An informed and more efficient decision-making process.
- Create strategies based on insights obtained from raw data.
- Constant performance measurement through KPIs in dashboards.
- Creation of interactive and automated reports (dashboards) in Python, Power BI and other tools.

Thank you for your attention