Google Play Store Apps

Project in data exploration and visualization

—

Maya Al-hatmi
124062

Contents

Introduction:

Google Play Store is a digital distribution service developed by google to allow users to browse and download applications developed with android software and published through google. I'm interested in studying and analyzing a dataset of google play store apps.

Research Question:

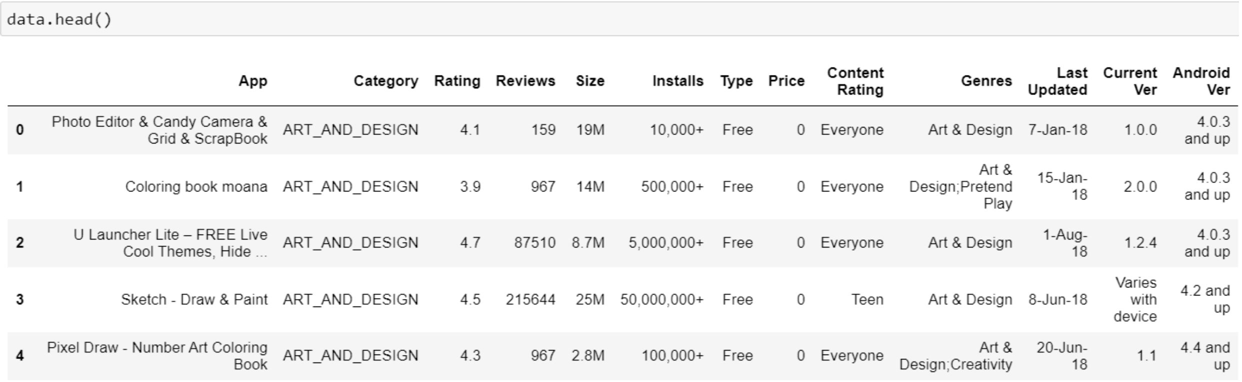
1. What category of apps are most prevalent among teenagers.
2. How business apps compared to lifestyle apps according to the size.
3. What is the relationship between size of app (as independent variable) and rating of app as outcome variable.
4. Measure the performance of the users reviews sentiment analysis.

Dataset (Description and summary):

Google play store apps dataset from kaggle website source : https://www.kaggle.com/lava18/google-play-store-apps/discussion/103760

Is a dataset collected by Lavanya Gubta . The dataset organized on CSV file contains 17 features . This multivariate structured dataset contains records of 10841 apps in google play store app ,each record has a unique name.

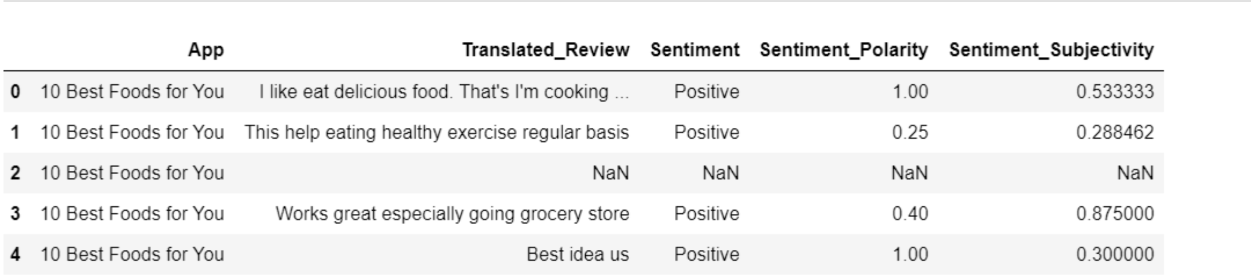
- The first csv file contains 13 features:



```
data.head()
```

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	7-Jan-18	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	15-Jan-18	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	1-Aug-18	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	8-Jun-18	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	20-Jun-18	1.1	4.4 and up

- The second csv contains 4 more features:



	App	Translated_Review	Sentiment	Sentiment_Polarity	Sentiment_Subjectivity
0	10 Best Foods for You	I like eat delicious food. That's I'm cooking ...	Positive	1.00	0.533333
1	10 Best Foods for You	This help eating healthy exercise regular basis	Positive	0.25	0.288462
2	10 Best Foods for You	NaN	NaN	NaN	NaN
3	10 Best Foods for You	Works great especially going grocery store	Positive	0.40	0.875000
4	10 Best Foods for You	Best idea us	Positive	1.00	0.300000

Data preprocessing

I. Replace value:

A. Variable size: there are values like '100,00+' ,'100M', and '1000k' , therefore ,I replace the character of ',' '+' 'M' 'K' with ' ' so it is better when converting from string to number.

II. Remove nan values:

A. Replace all nan values in the dataset by 0.

III. Remove duplicated rows.

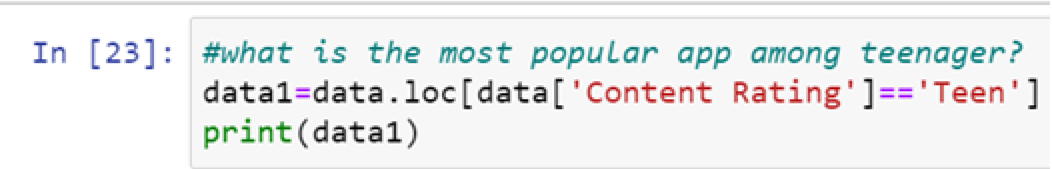
IV. Remove the raw number 10472.

A. Since it causes errors in modelling.

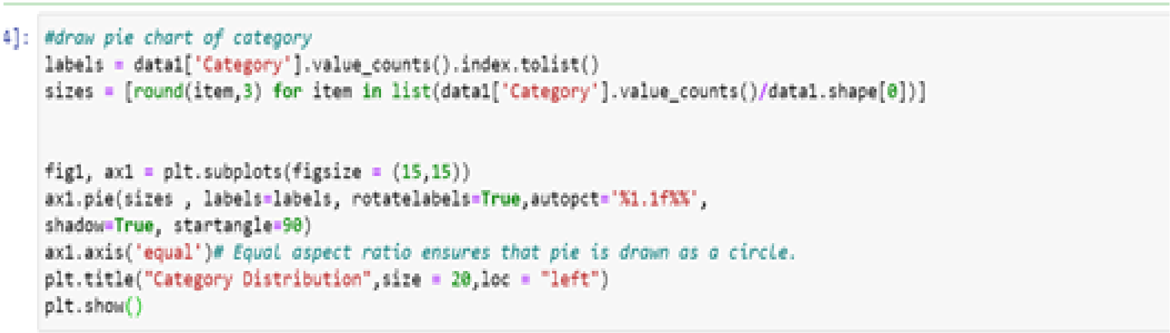
Methods:

1. What category of apps are most prevalent among teenagers:

- visualize the category using pie charts . The visualization contains the percentage of each category according to where the content is for teens.



```
In [23]: #what is the most popular app among teenager?
         data1=data.loc[data['Content Rating']=='Teen']
         print(data1)
```

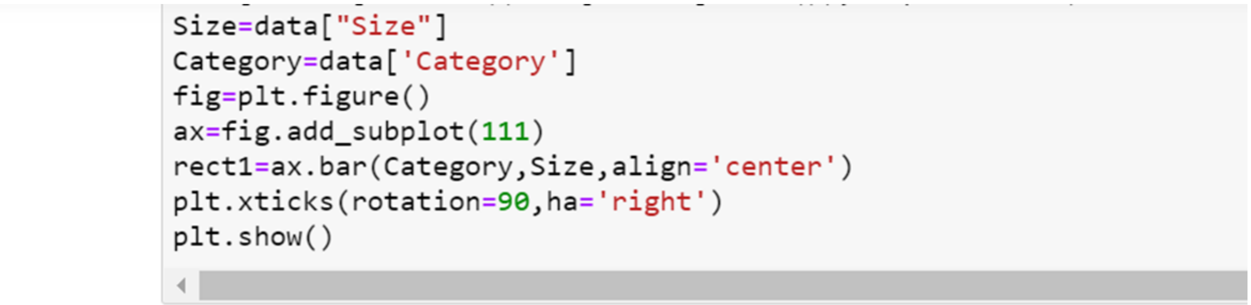


```
4]: #draw pie chart of category
    labels = data1['Category'].value_counts().index.tolist()
    sizes = [round(item,3) for item in list(data1['Category'].value_counts()/data1.shape[0])]


    fig1, ax1 = plt.subplots(figsize = (15,15))
    ax1.pie(sizes , labels=labels, rotatelabels=True,autopct='%1.1f%%',
    shadow=True, startangle=90)
    ax1.axis('equal')# Equal aspect ratio ensures that pie is drawn as a circle.
    plt.title("Category Distribution",size = 20,loc = "left")
    plt.show()
```

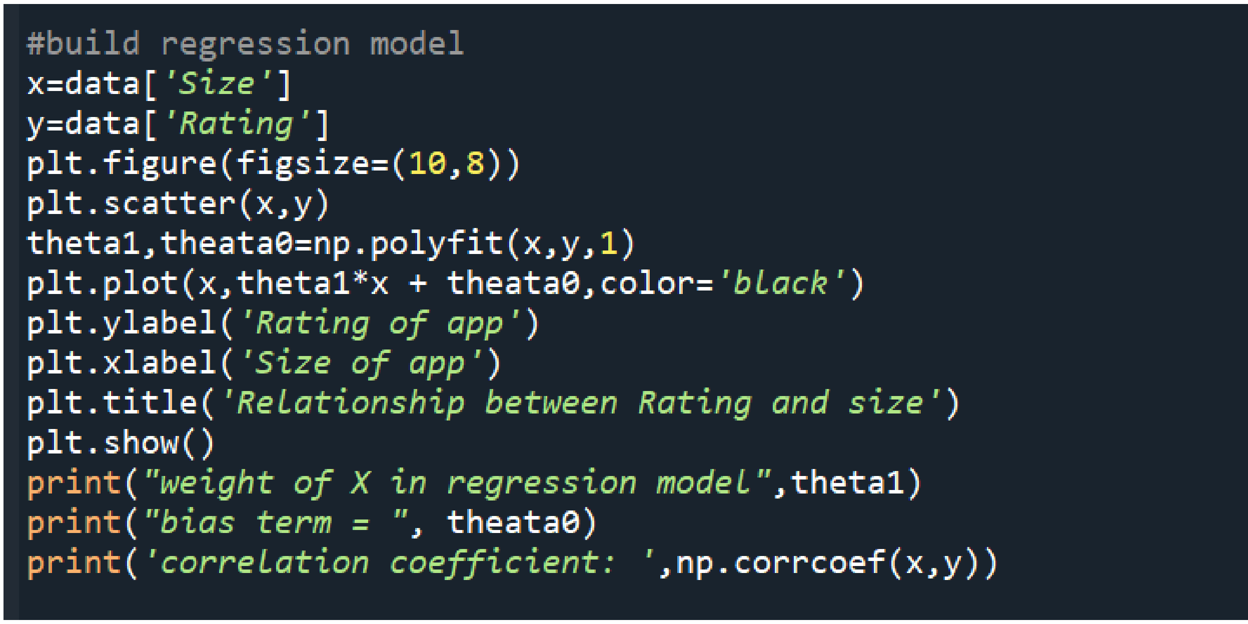
2. How business apps compared to lifestyle apps according to the size.

- visualize the size and category using a bar chart.



```
Size=data["Size"]
Category=data['Category']
fig=plt.figure()
ax=fig.add_subplot(111)
rect1=ax.bar(Category,Size,align='center')
plt.xticks(rotation=90,ha='right')
plt.show()
```

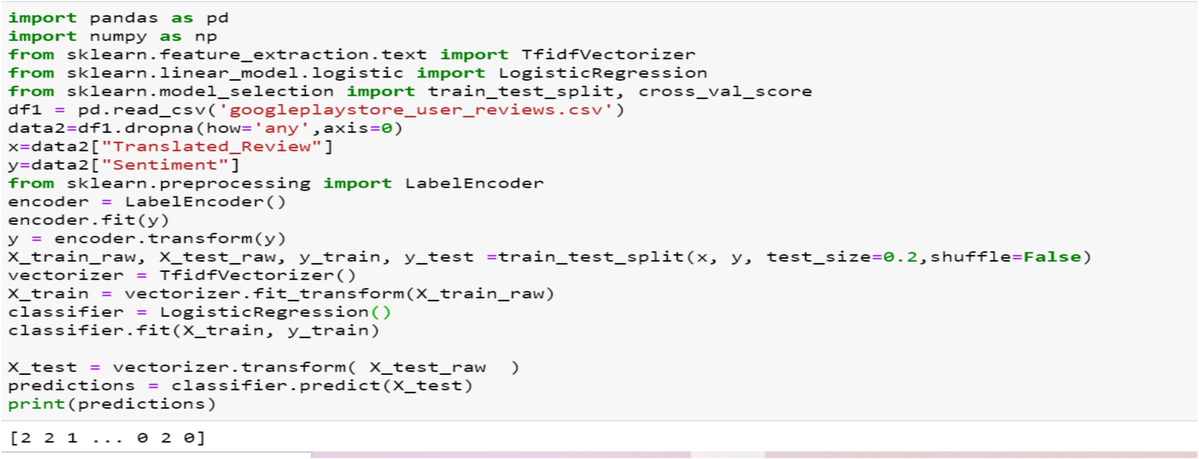
3. What is the relationship between size of app (as independent variable) and rating of app as outcome variable.
   - build a linear regression model to study the relation between size and rating , and find correlation coefficient.



```
#build regression model
x=data['Size']
y=data['Rating']
plt.figure(figsize=(10,8))
plt.scatter(x,y)
theta1,theata0=np.polyfit(x,y,1)
plt.plot(x,theta1*x + theata0,color='black')
plt.ylabel('Rating of app')
plt.xlabel('Size of app')
plt.title('Relationship between Rating and size')
plt.show()
print("weight of X in regression model",theta1)
print("bias term = ", theata0)
print('correlation coefficient: ',np.corrcoef(x,y))
```

4. Measure the performance of the users reviews sentiment analysis.

- Build a logistic regression model to compute the accuracy of the Sentiment analysis.

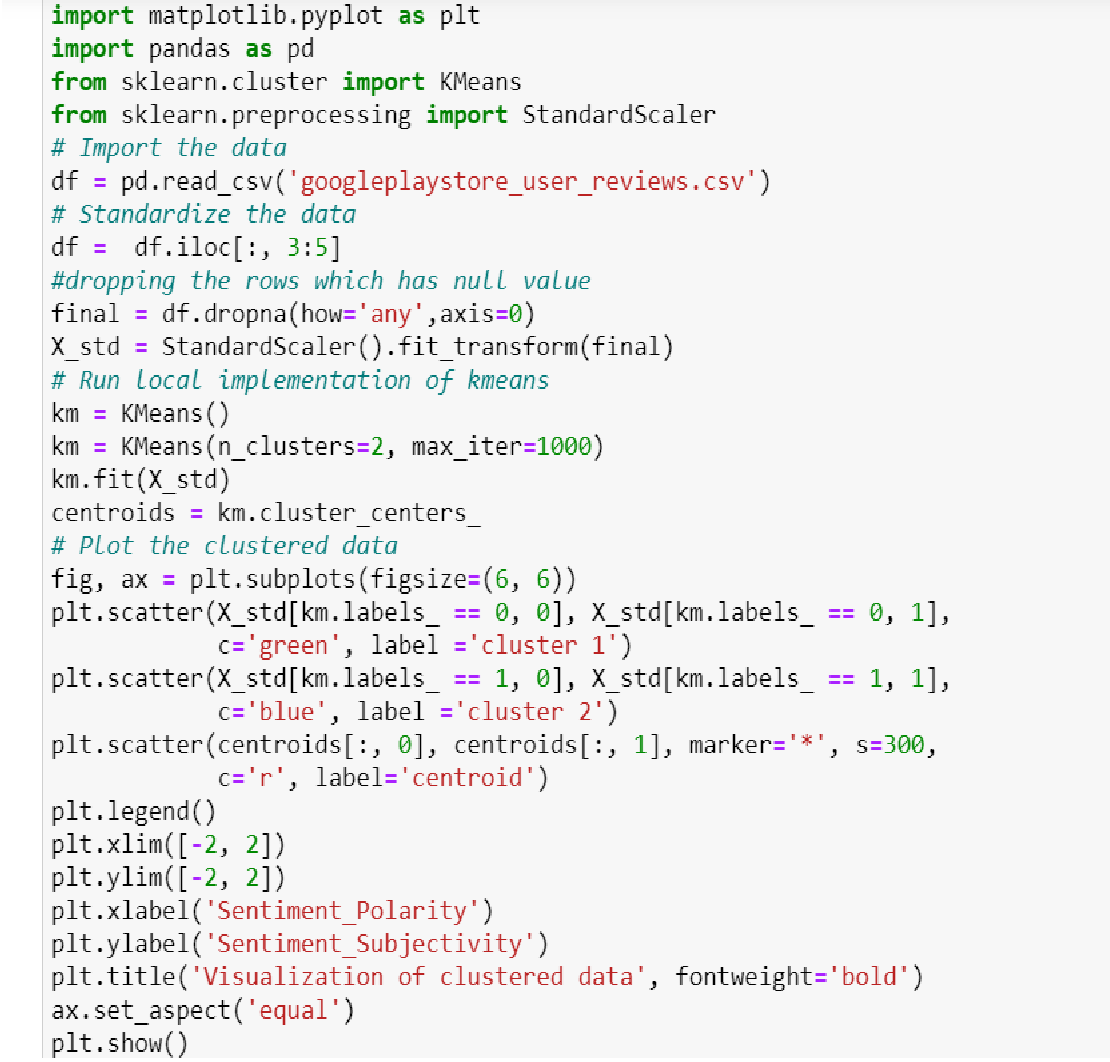


```
import pandas as pd
import numpy as np
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model.logistic import LogisticRegression
from sklearn.model_selection import train_test_split, cross_val_score
df1 = pd.read_csv('googleplaystore_user_reviews.csv')
data2=df1.dropna(how='any',axis=0)
x=data2["Translated_Review"]
y=data2["Sentiment"]
from sklearn.preprocessing import LabelEncoder
encoder = LabelEncoder()
encoder.fit(y)
y = encoder.transform(y)
X_train_raw, X_test_raw, y_train, y_test =train_test_split(x, y, test_size=0.2,shuffle=False)
vectorizer = TfidfVectorizer()
X_train = vectorizer.fit_transform(X_train_raw)
classifier = LogisticRegression()
classifier.fit(X_train, y_train)

X_test = vectorizer.transform( X_test_raw  )
predictions = classifier.predict(X_test)
print(predictions)
```

```
[2 2 1 ... 0 2 0]
```

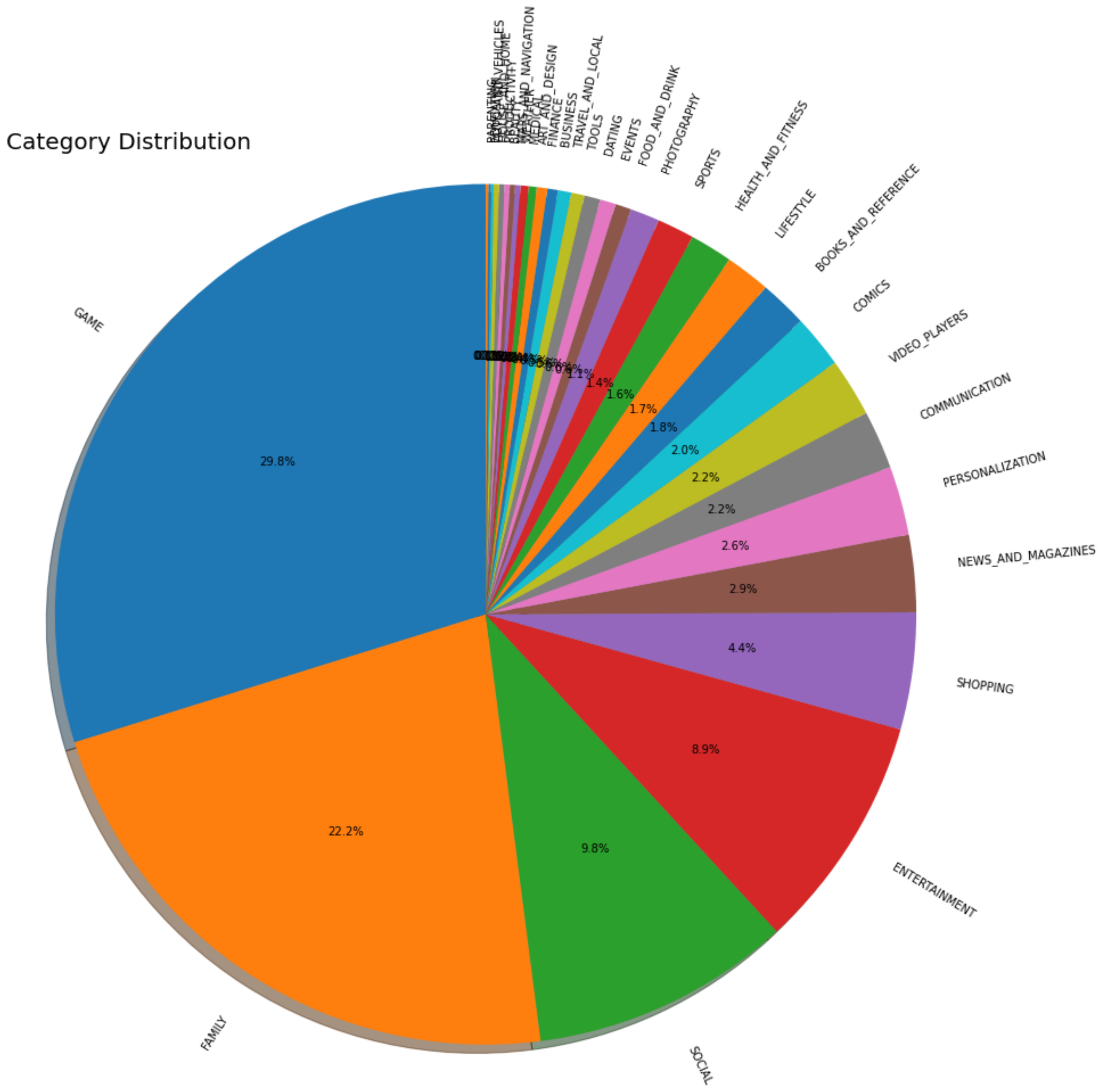
- Clustering: cluster the data across Sentiment_Polarity and Sentiment_Subjectivity.



```
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
# Import the data
df = pd.read_csv('googleplaystore_user_reviews.csv')
# Standardize the data
df =  df.iloc[:, 3:5]
#dropping the rows which has null value
final = df.dropna(how='any',axis=0)
X_std = StandardScaler().fit_transform(final)
# Run local implementation of kmeans
km = KMeans()
km = KMeans(n_clusters=2, max_iter=1000)
km.fit(X_std)
centroids = km.cluster_centers_
# Plot the clustered data
fig, ax = plt.subplots(figsize=(6, 6))
plt.scatter(X_std[km.labels_ == 0, 0], X_std[km.labels_ == 0, 1],
            c='green', label ='cluster 1')
plt.scatter(X_std[km.labels_ == 1, 0], X_std[km.labels_ == 1, 1],
            c='blue', label ='cluster 2')
plt.scatter(centroids[:, 0], centroids[:, 1], marker='*', s=300,
            c='r', label='centroid')
plt.legend()
plt.xlim([-2, 2])
plt.ylim([-2, 2])
plt.xlabel('Sentiment_Polarity')
plt.ylabel('Sentiment_Subjectivity')
plt.title('Visualization of clustered data', fontweight='bold')
ax.set_aspect('equal')
plt.show()
```

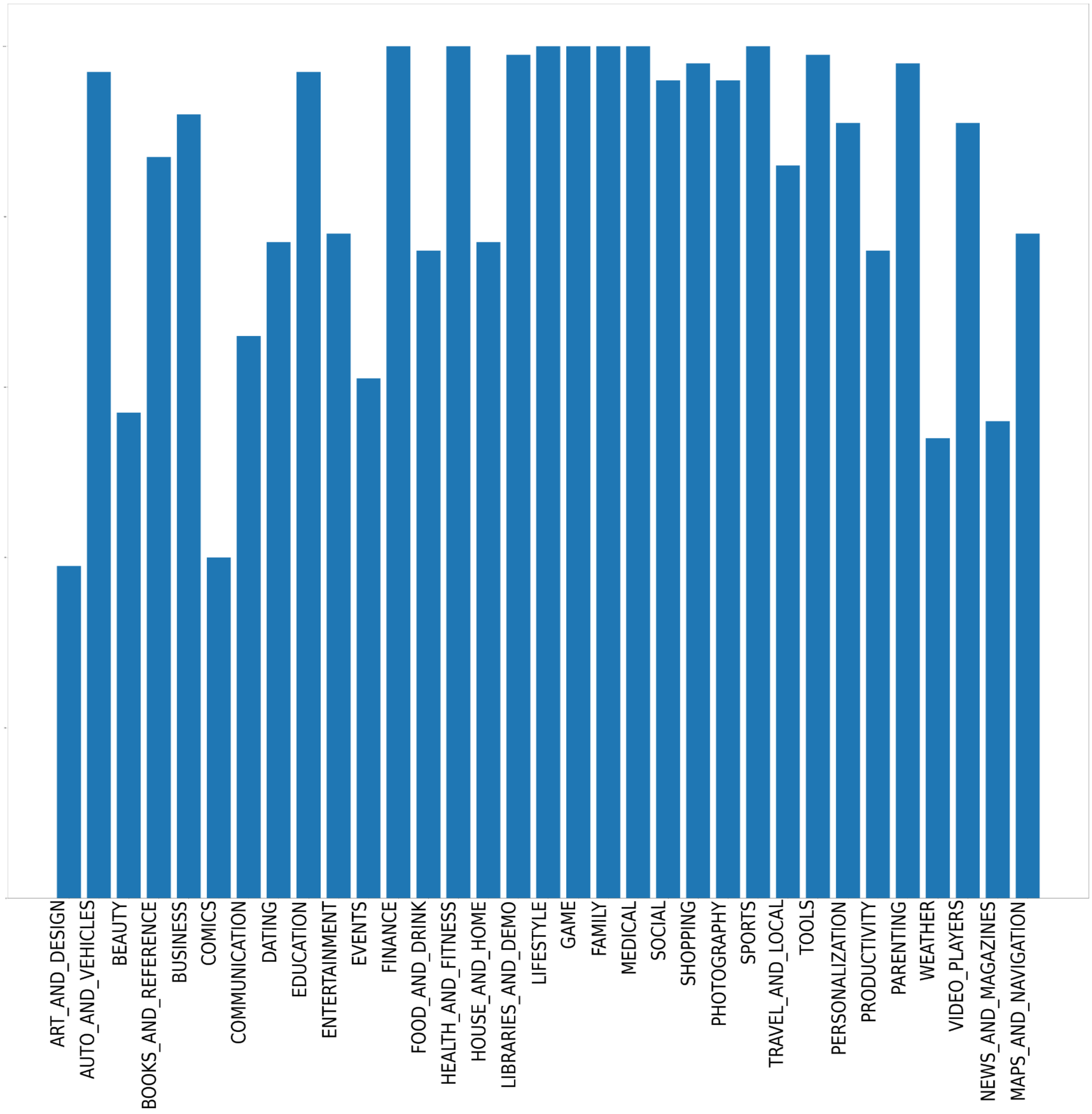
Results:

a. visualize the category using pie charts. The visualization contains the percentage of each category according to the content for teens.



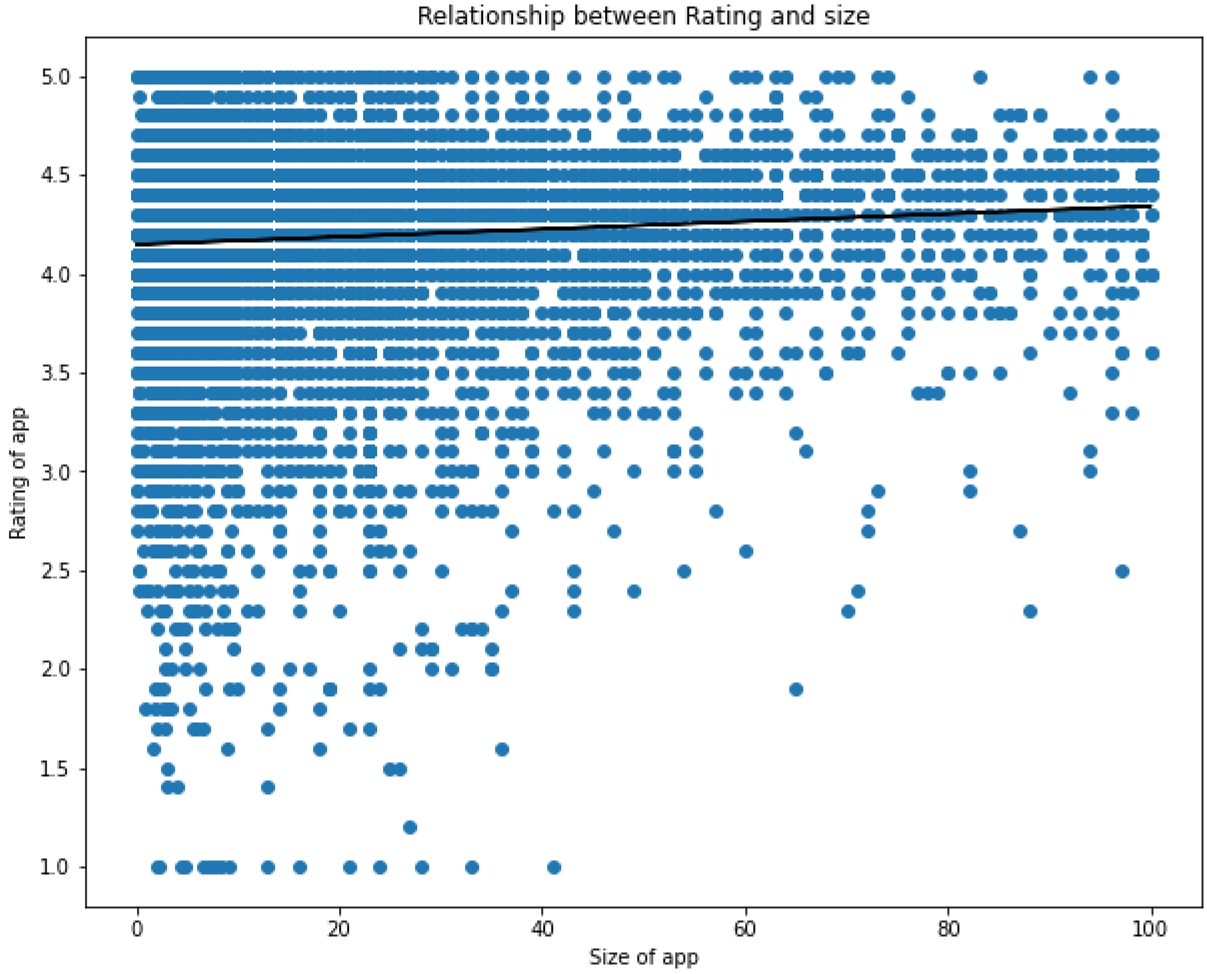
- The result shows that the game is the most prevalent category of apps among teens by 29.8%.

b. visualize the size and category using a bar chart.



- The bar chart shows that lifestyle apps are larger in size than business apps.

c. build a linear regression model to study the relation between size and rating ,and calculate correlation coefficient.



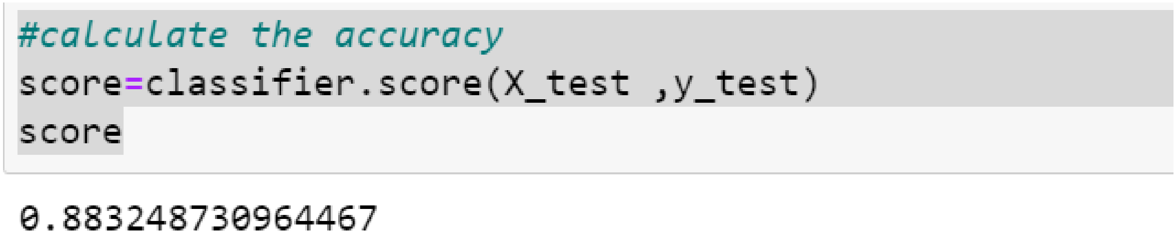
weight of X in regression model 0.0019426500553753837

bias term = 4.147214702424731

correlation coefficient: [[1.   0.0803078][0.0803078 1.   ]]

- Positive correlation shows that as the number of size increases , the mean of rating also tends to increase.

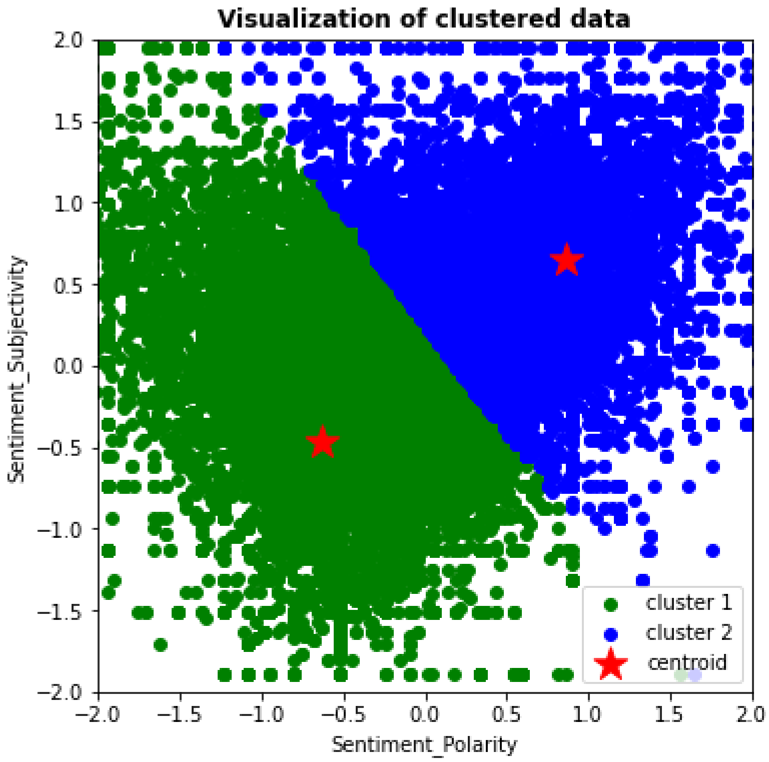
d. Build a logistic regression model to compute the accuracy of the Sentiment analysis and cluster the data across Sentiment_Polarity and Sentiment_Subjectivity.



```
#calculate the accuracy
score=classifier.score(X_test ,y_test)
score
```

```
0.883248730964467
```

- The accuracy of logistic regression is 88% , meaning that the model predicts the Sentiment correctly by 88%.
- The sentiment function of test returns two properties , polarity and subjectivity the clustering of this properties:



Conclusion:

In conclusion, there is an accelerated increase in the number of apps in google play store.

Depending on the results obtained, I expect to see more gaming apps that target teens in particular. I hope to do future analysis on the dataset and perform better results .

References :

https://www.kaggle.com/lava18/google-play-store-apps/tasks?taskId=276