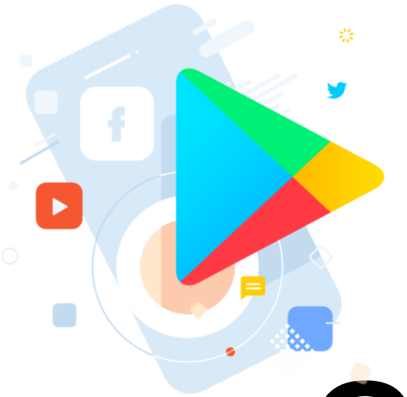


Google Play Store

Group 9: Kevin Lee, Yatian Lu, Yuhan Wang, Gloria Yu, Lanxue Zhang

Agenda

1. Introduction
2. Data Inspecting & Cleaning
3. Analysis (Text Mining, Sentiment Analysis, Pivot Table)
4. Visualization
5. Predictive Model
6. Conclusion

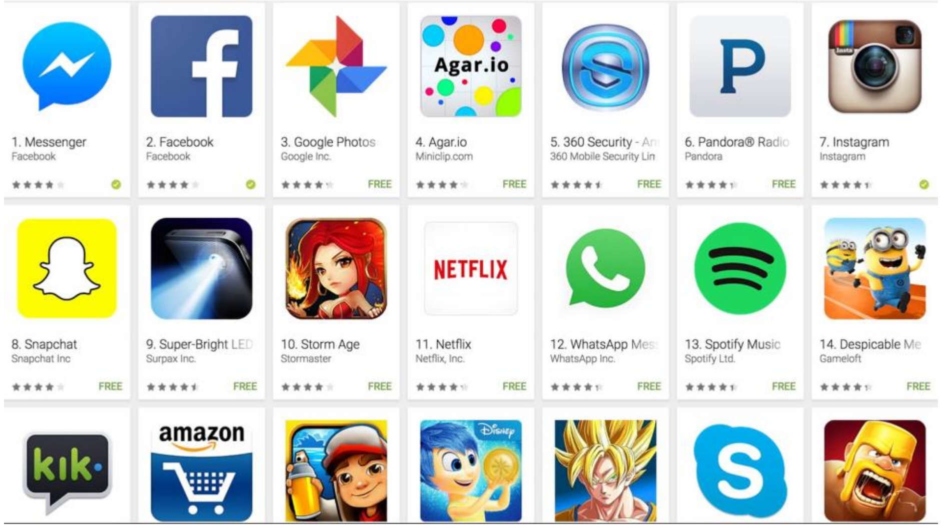
1.1 Introduction: Background

Background of Google Play Store

- Over 3.5 million apps
- For each app, Google Play Store will show its name, category, rating, reviews, size, updated time, price, genre, current version and minimum Android version required.

Background of Dataset

- The data was scraped from the Google Play Store
- The dataset includes 10.8k rows and 13 features.



1.2 Introduction: Define Problem

Current Situation

- Many public datasets for Apple Store Apps
- Not many datasets available for Google Play Store

ANDROID APP ON Google play

Available on the App Store

Our Objective

- Analyze and predict what factors affect the rating of the app
- Generating insights to help with decision-making, driving app-making business to success

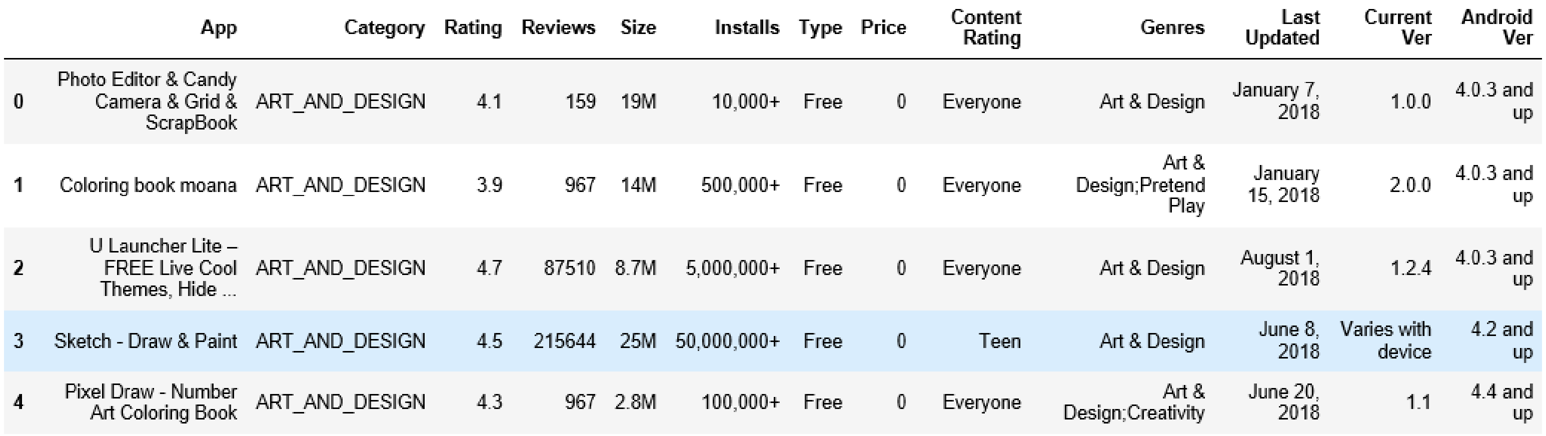
1.3 Introduction: Variable Definition

Variable	Type	Definition
App	string	Application name
Category	string	Category the app belongs to, such as family, game, tools and medical
Rating	Decimal	Overall user rating of the app
Reviews	Integer	Number of user reviews for the app
Size	String	Size of the app
Installs	String	Number of user downloads/installs for the app
Type	String	Paid or Free
Price	String	Price of the app
Content Rating	String	Age group the app is targeted at - Children / Mature 21+ / Adult
Genres	String	An app can belong to multiple genres (apart from its main category)
Last Updated	Date	Date when the app was last updated on Play Store
Current Ver	String	Current version of the app available on Play Store
Android Ver	String	Min required Android version

Note: all the values of each variable are as when scraped.

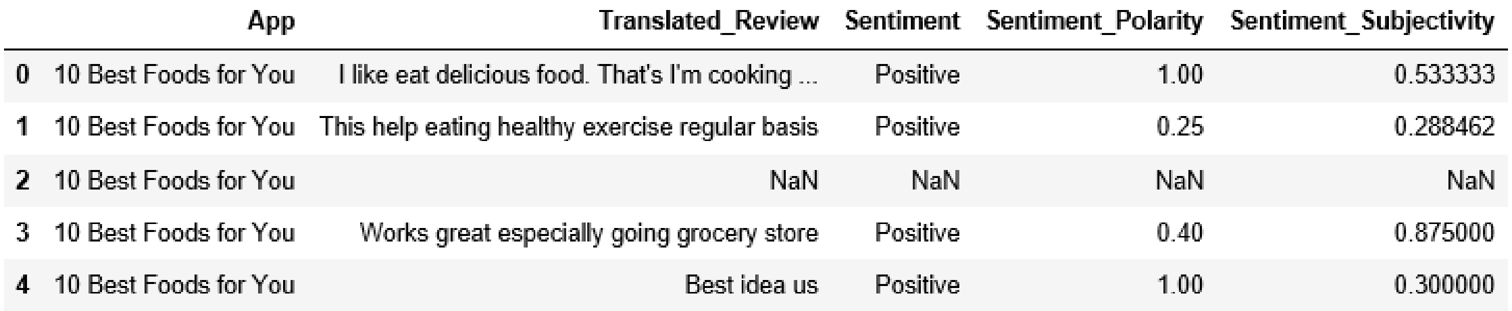
2.1 Data Inspecting & Cleaning: Inspecting

app.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	January 7, 2018	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	January 15, 2018	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	August 1, 2018	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	June 8, 2018	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	June 20, 2018	1.1	4.4 and up

review.head()



	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

2.2 Data Inspecting & Cleaning: Cleaning

App data:

app.shape[0:2]: (10841,13)

app.info(): Only Rating is float64 type, others are all object.

app.isnull().any(): Rating, Type, Content Rating, Current Ver, Android Ver

app.Rating.describe(): Rating values range from 0 to 5, with an outlier of 19.0, removed.

app.drop_duplicates()

app.dropna()

Cleaned Shape: (8886, 13)

Review data:

review.shape[0:2]: (64295, 5)

review.info(): Reviews and Sentiment are object. Sentiment Polarity and Subjectivity are float64

review.isnull().any(): Translated_Review, Sentiment, Sentiment_Polarity, Sentiment_Subjectivity

review.drop_duplicates()

app.dropna()

Cleaned Shape: (29692, 5)

2.3 Data Inspecting & Cleaning: Exploring

Variable exploration:

App data:

Category: 33 unique values. Event category has the highest avg rating at 4.435556

Genres: 115 unique values. Adventure;Brain Games genre has the highest avg rating at 4.6

Content Ratings: 6 unique values. Adults only 18+ has the highest avg rating at 4.3

Type: Free and Paid. Paid type gets higher ratings.

Review data:

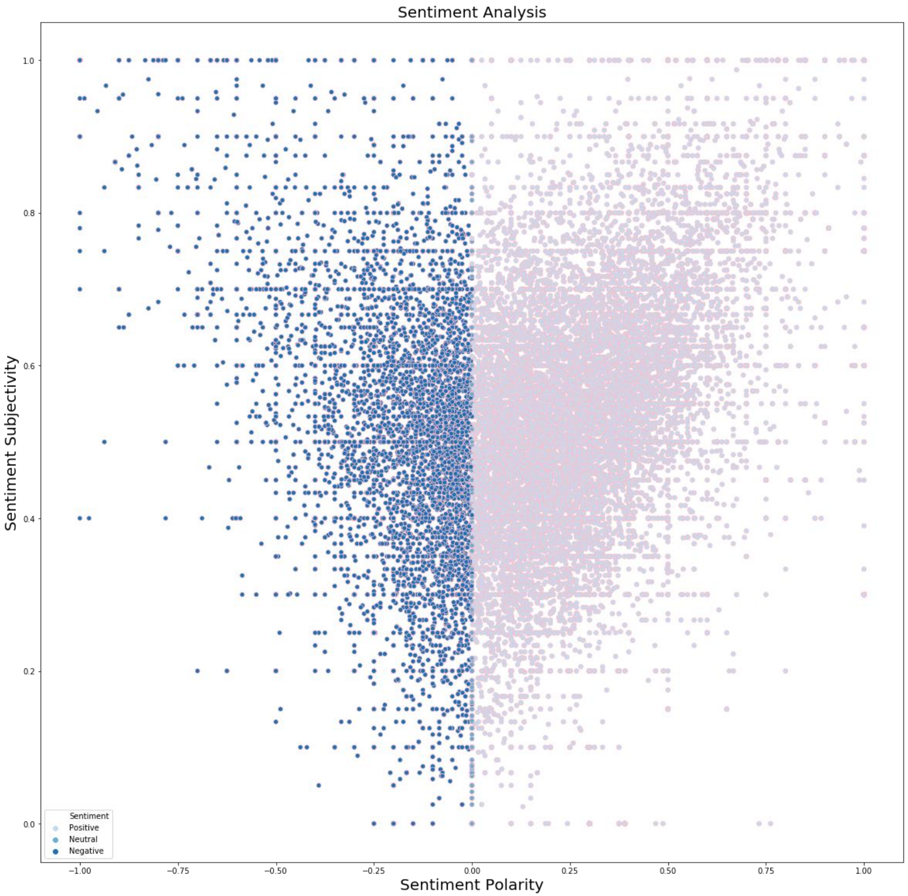
Sentiment: Positive, Negative and Neutral. Over 60% reviews are positive.

Sentiment Polarity and Subjectivity are consistent with sentiment categories.

Key code:

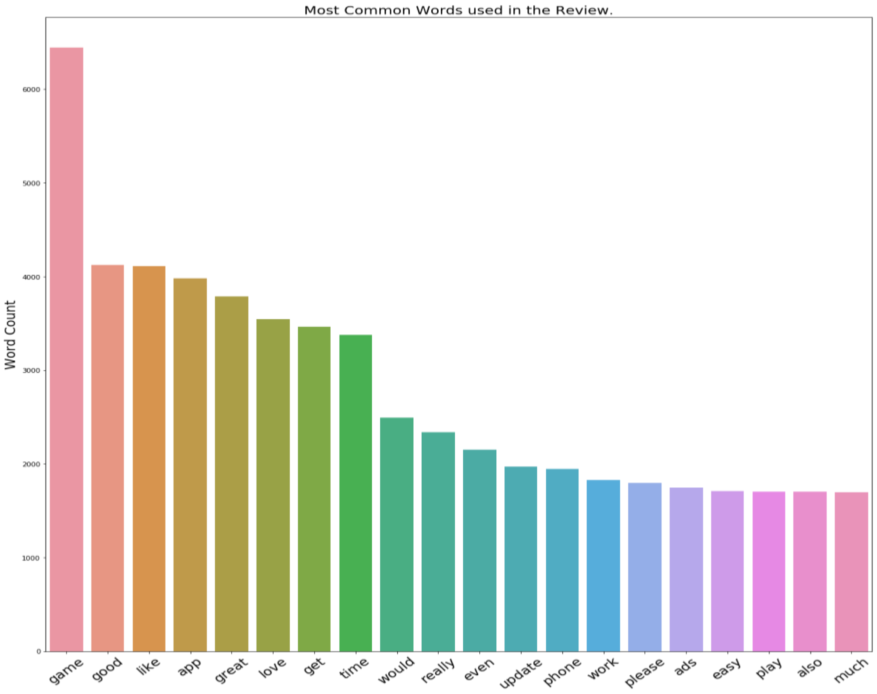
.describe()
.value_counts()
.groupby()
.mean()

3.1 Analysis: Text Mining & Sentiment Analysis I



Sentiment Analysis

- More polarity, more subjectivity
- “Game” is the most common word in the reviews



Most Common Words used in the Review.

3.2 Analysis: Text Mining & Sentiment Analysis II

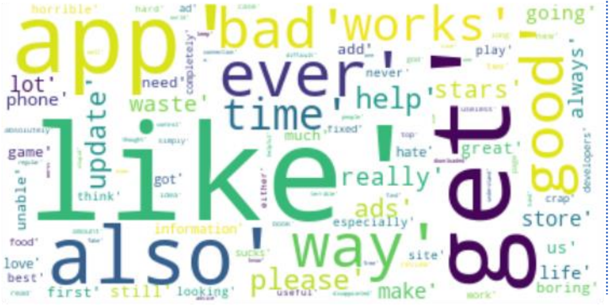
Methods

- WordClouds are Generated by Naive Bayesian Classifier--a type of Machine Learning Approach
- 75% train data & 25% test data
- Accuracy: 76.06

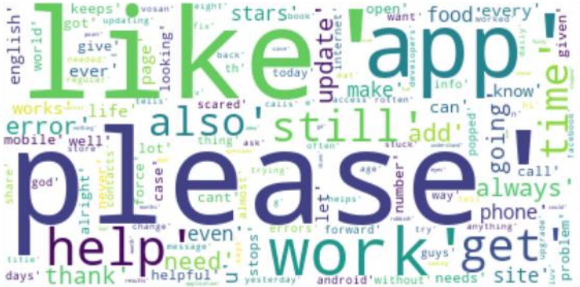
Analysis

- “Like”,”app” should be neutral words since people always leave reviews such as “I like app” or “I don’t like app”
- The outcome is acceptable since all positive words are classified in the positive wordcloud and all negative words are classified in the negative wordcloud, excluding the word “good”.

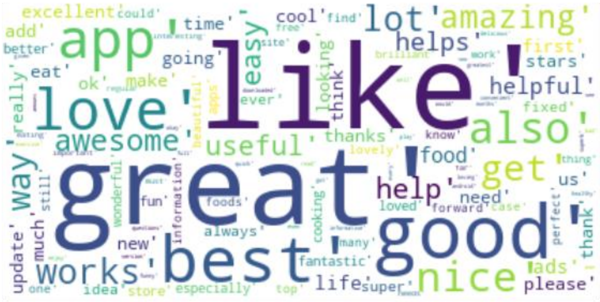
Negative WordCloud



Neutral WordCloud

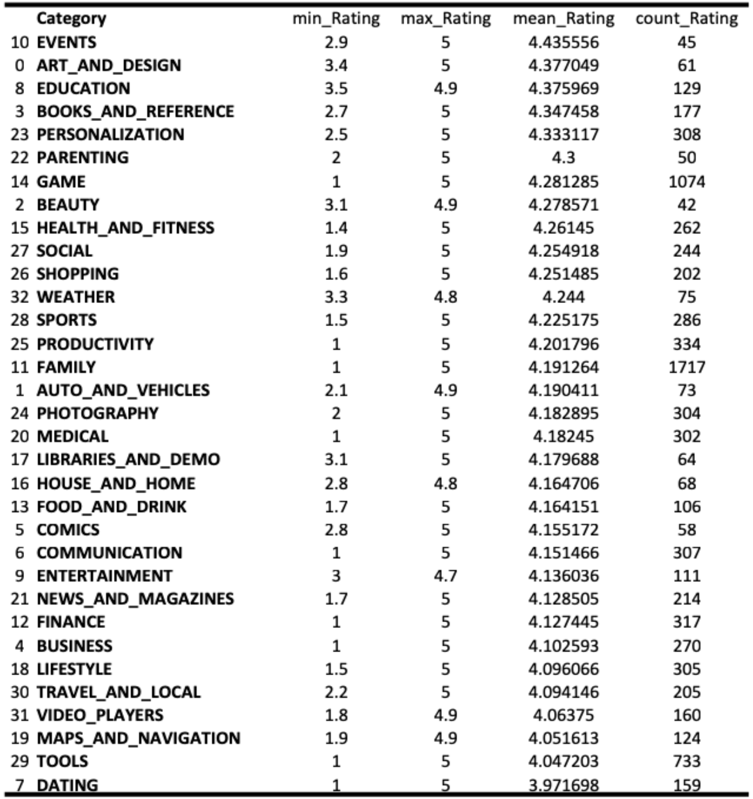


Positive WordCloud



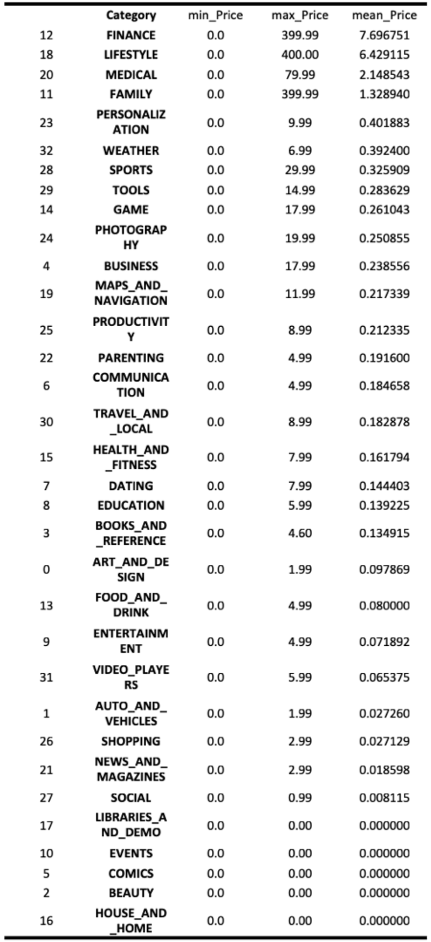
3.3 Analysis: Pivot Tables

Different Types of Apps Ranked by Mean_Rating



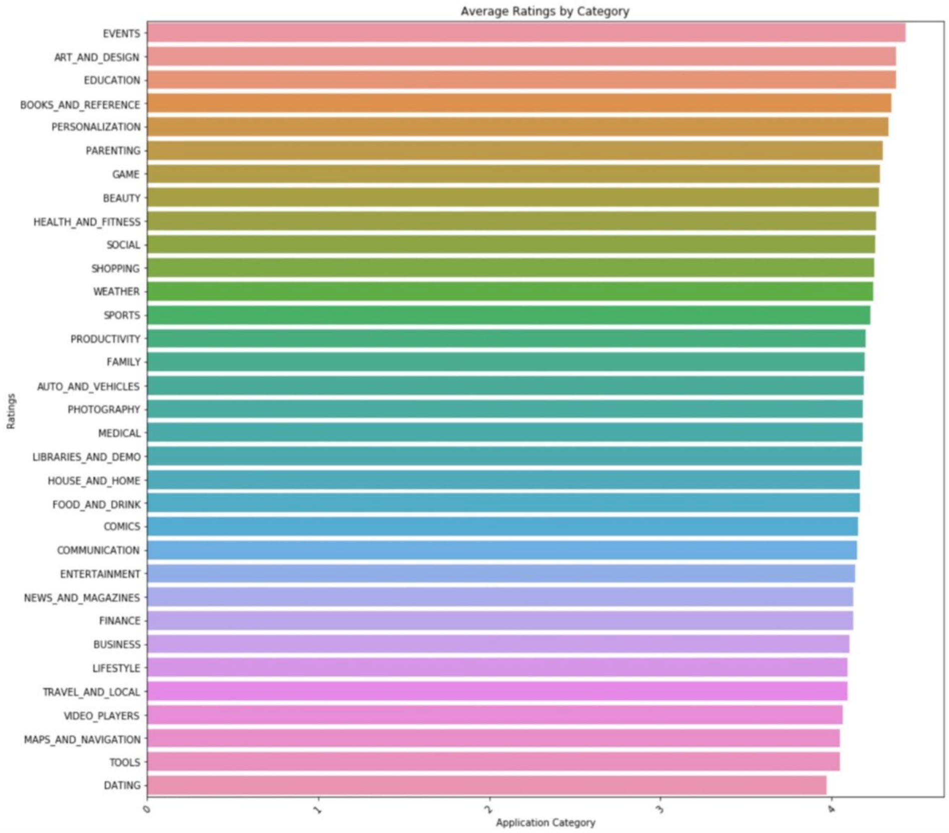
	Category	min_Rating	max_Rating	mean_Rating	count_Rating
10	EVENTS	2.9	5	4.435556	45
0	ART_AND_DESIGN	3.4	5	4.377049	61
8	EDUCATION	3.5	4.9	4.375969	129
3	BOOKS_AND_REFERENCE	2.7	5	4.347458	177
23	PERSONALIZATION	2.5	5	4.333117	308
22	PARENTING	2	5	4.3	50
14	GAME	1	5	4.281285	1074
2	BEAUTY	3.1	4.9	4.278571	42
15	HEALTH_AND_FITNESS	1.4	5	4.26145	262
27	SOCIAL	1.9	5	4.254918	244
26	SHOPPING	1.6	5	4.251485	202
32	WEATHER	3.3	4.8	4.244	75
28	SPORTS	1.5	5	4.225175	286
25	PRODUCTIVITY	1	5	4.201796	334
11	FAMILY	1	5	4.191264	1717
1	AUTO_AND_VEHICLES	2.1	4.9	4.190411	73
24	PHOTOGRAPHY	2	5	4.182895	304
20	MEDICAL	1	5	4.18245	302
17	LIBRARIES_AND_DEMO	3.1	5	4.179688	64
16	HOUSE_AND_HOME	2.8	4.8	4.164706	68
13	FOOD_AND_DRINK	1.7	5	4.164151	106
5	COMICS	2.8	5	4.155172	58
6	COMMUNICATION	1	5	4.151466	307
9	ENTERTAINMENT	3	4.7	4.136036	111
21	NEWS_AND_MAGAZINES	1.7	5	4.128505	214
12	FINANCE	1	5	4.127445	317
4	BUSINESS	1	5	4.102593	270
18	LIFESTYLE	1.5	5	4.096066	305
30	TRAVEL_AND_LOCAL	2.2	5	4.094146	205
31	VIDEO_PLAYERS	1.8	4.9	4.06375	160
19	MAPS_AND_NAVIGATION	1.9	4.9	4.051613	124
29	TOOLS	1	5	4.047203	733
7	DATING	1	5	3.971698	159

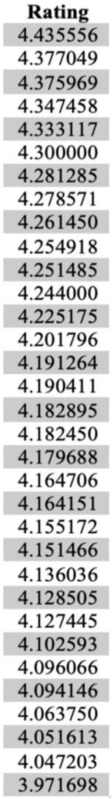
Different Types of Apps Ranked by Mean_Price



	Category	min_Price	max_Price	mean_Price
12	FINANCE	0.0	399.99	7.696751
18	LIFESTYLE	0.0	400.00	6.429115
20	MEDICAL	0.0	79.99	2.148543
11	FAMILY	0.0	399.99	1.328940
23	PERSONALIZATION	0.0	9.99	0.401883
32	WEATHER	0.0	6.99	0.392400
28	SPORTS	0.0	29.99	0.325909
29	TOOLS	0.0	14.99	0.283629
14	GAME	0.0	17.99	0.261043
24	PHOTOGRAPHY	0.0	19.99	0.250855
4	BUSINESS	0.0	17.99	0.238556
19	MAPS_AND_NAVIGATION	0.0	11.99	0.217339
25	PRODUCTIVITY	0.0	8.99	0.212335
22	PARENTING	0.0	4.99	0.191600
6	COMMUNICATION	0.0	4.99	0.184658
30	TRAVEL_AND_LOCAL	0.0	8.99	0.182878
15	HEALTH_AND_FITNESS	0.0	7.99	0.161794
7	DATING	0.0	7.99	0.144403
8	EDUCATION	0.0	5.99	0.139225
3	BOOKS_AND_REFERENCE	0.0	4.60	0.134915
0	ART_AND_DESIGN	0.0	1.99	0.097869
13	FOOD_AND_DRINK	0.0	4.99	0.080000
9	ENTERTAINMENT	0.0	4.99	0.071892
31	VIDEO_PLAYERS	0.0	5.99	0.065375
1	AUTO_AND_VEHICLES	0.0	1.99	0.027260
26	SHOPPING	0.0	2.99	0.027129
21	NEWS_AND_MAGAZINES	0.0	2.99	0.018598
27	SOCIAL	0.0	0.99	0.008115
17	LIBRARIES_AND_DEMO	0.0	0.00	0.000000
10	EVENTS	0.0	0.00	0.000000
5	COMICS	0.0	0.00	0.000000
2	BEAUTY	0.0	0.00	0.000000
16	HOUSE_AND_HOME	0.0	0.00	0.000000

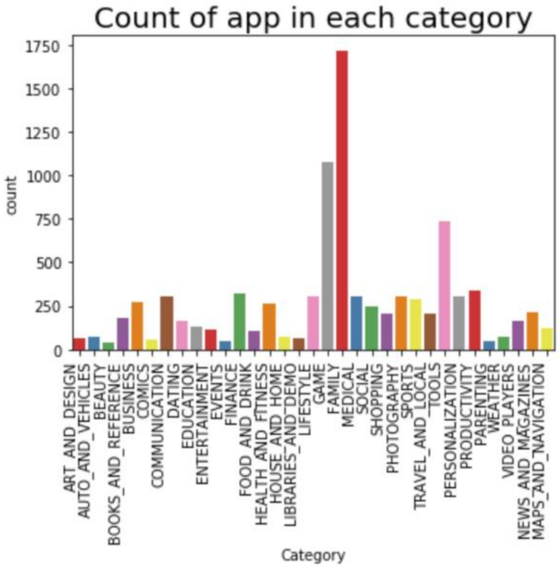
4.1 Visualization - Different Types of Apps



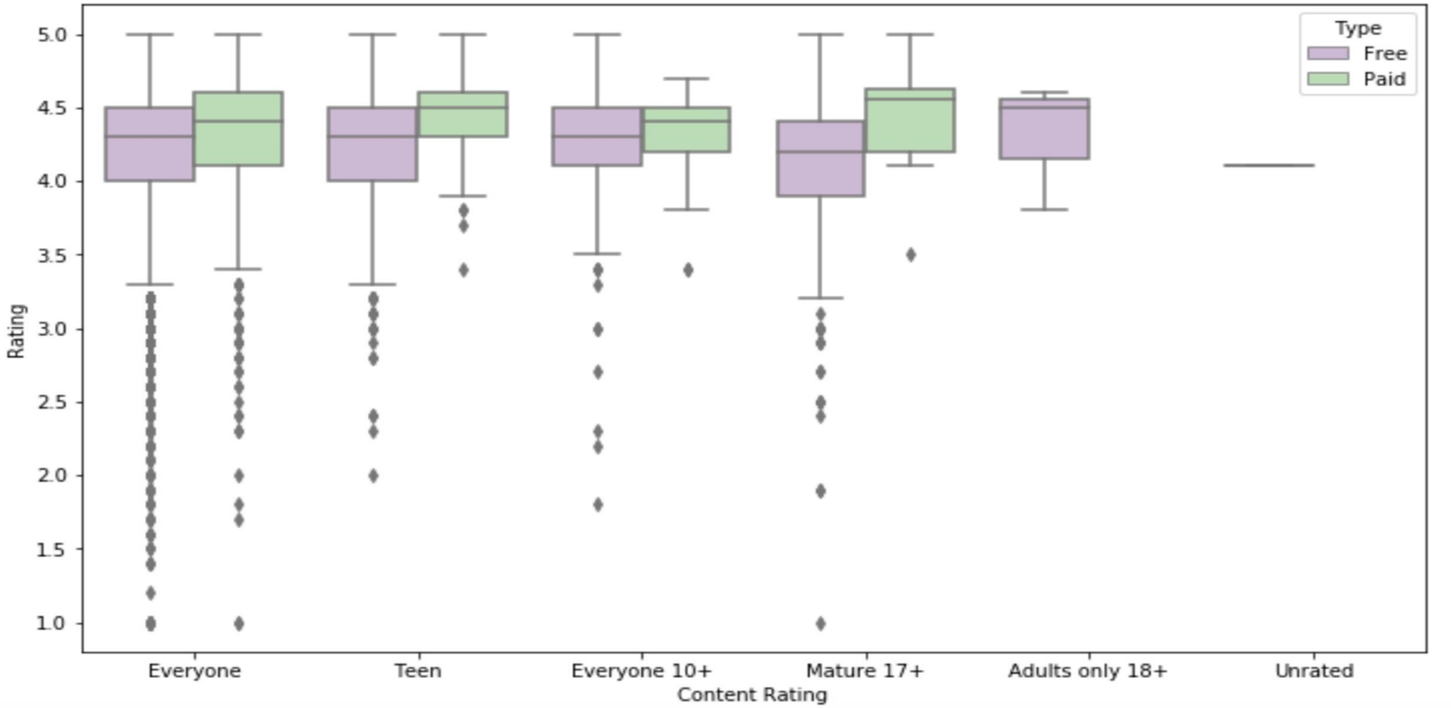


Rating
4.435556
4.377049
4.375969
4.347458
4.333117
4.300000
4.281285
4.278571
4.261450
4.254918
4.251485
4.244000
4.225175
4.201796
4.191264
4.190411
4.182895
4.182450
4.179688
4.164706
4.164151
4.155172
4.151466
4.136036
4.128505
4.127445
4.102593
4.096066
4.094146
4.063750
4.051613
4.047203
3.971698

- Different types of apps have similar ratings
- Events, Art & Design, Education apps have the highest rating, while Dating, Tools, Maps & Navigation apps have the lowest ratings.
- The top three amount of apps in google play store are Family,Game and personalization apps. The number of Beauty apps is the smallest in the google play store.

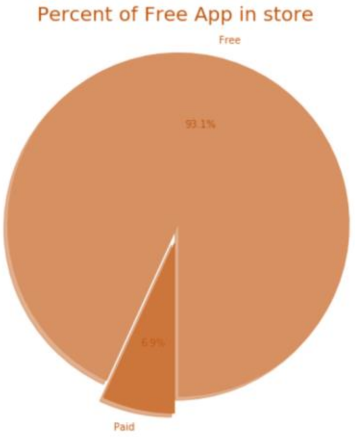


4.2 Visualization: Paid Apps Have Higher Ratings

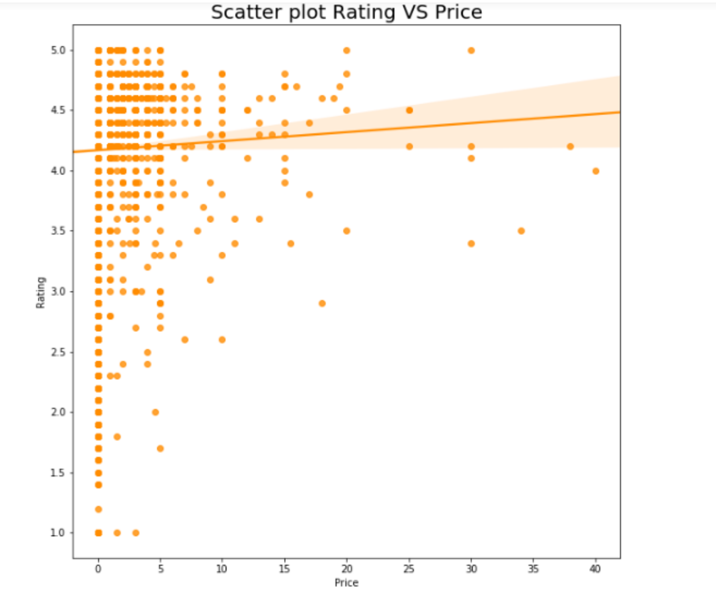


- Paid apps in general have higher rating than free apps.
- Apps that are avaliable to different age groups have similar ratings.
- Most apps in the Google Play Store are for all age groups, there are no paid apps that are only for Adults 18+ users

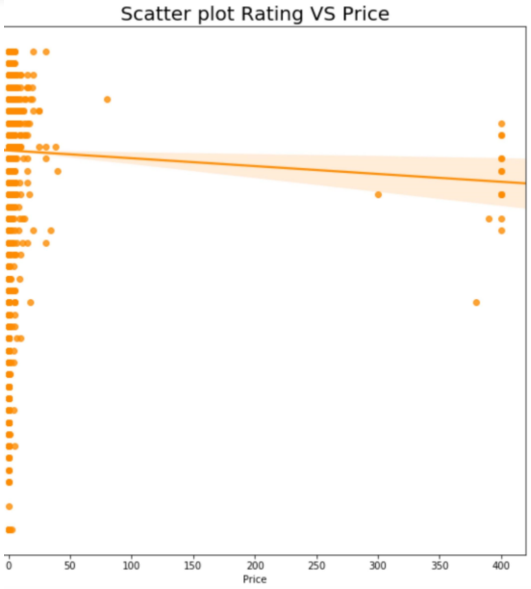
4.3 Visualization: Price and Ratings



Percent of Free App in store



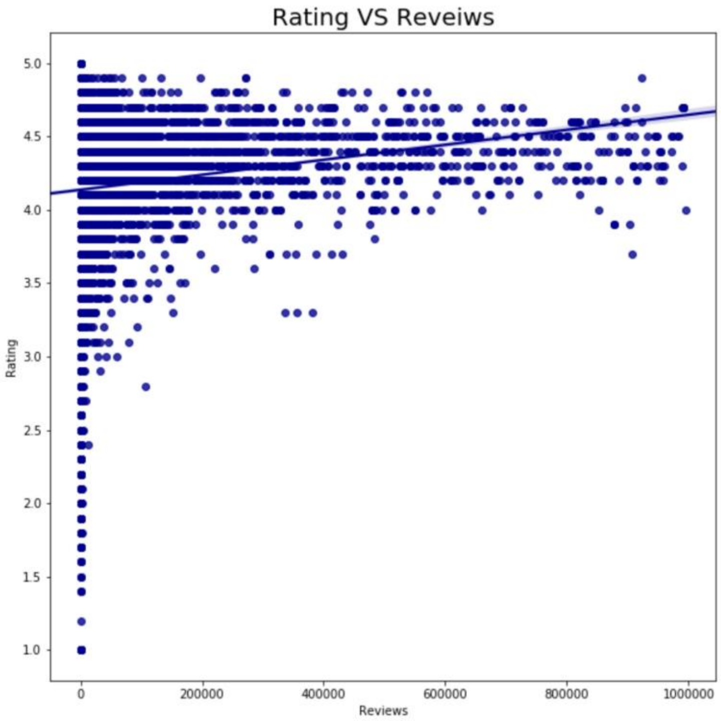
Scatter plot Rating VS Price



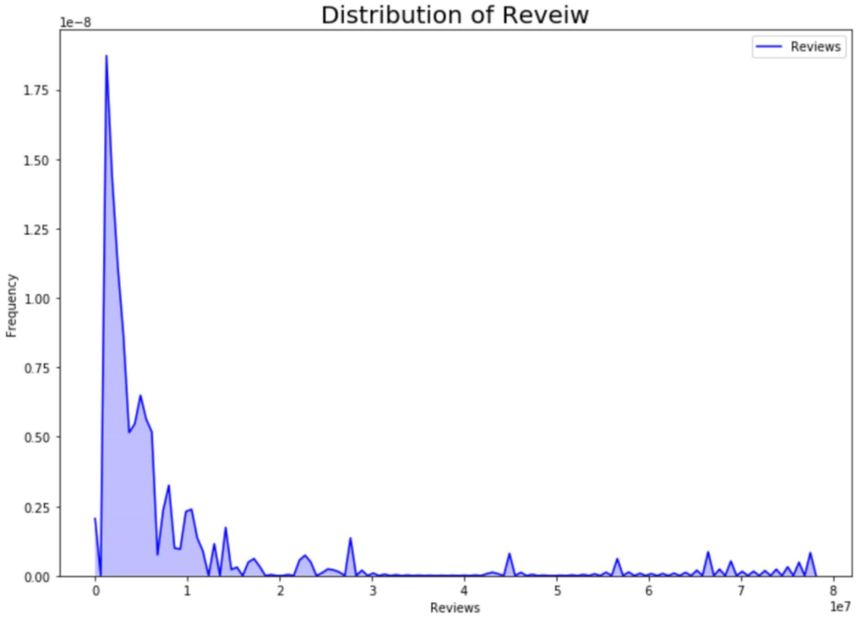
Scatter plot Rating VS Price

- Only 7% Apps are paid apps.
- For Apps that are less than $50, as price increase, the rating decreases.
- For Apps charges higher than $50, high price tend to come with low ratings

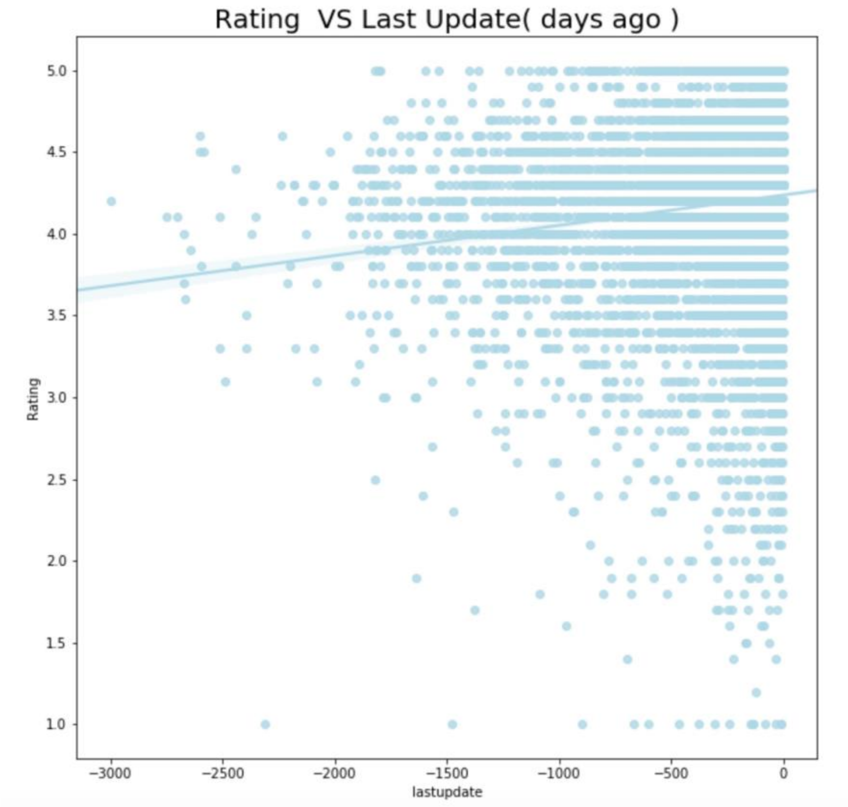
4.4 Visualization: More reviews, Higher ratings



- Most of application in this store have less than 1M in reviews.
- Popular Apps with more reviews tend to have a good review

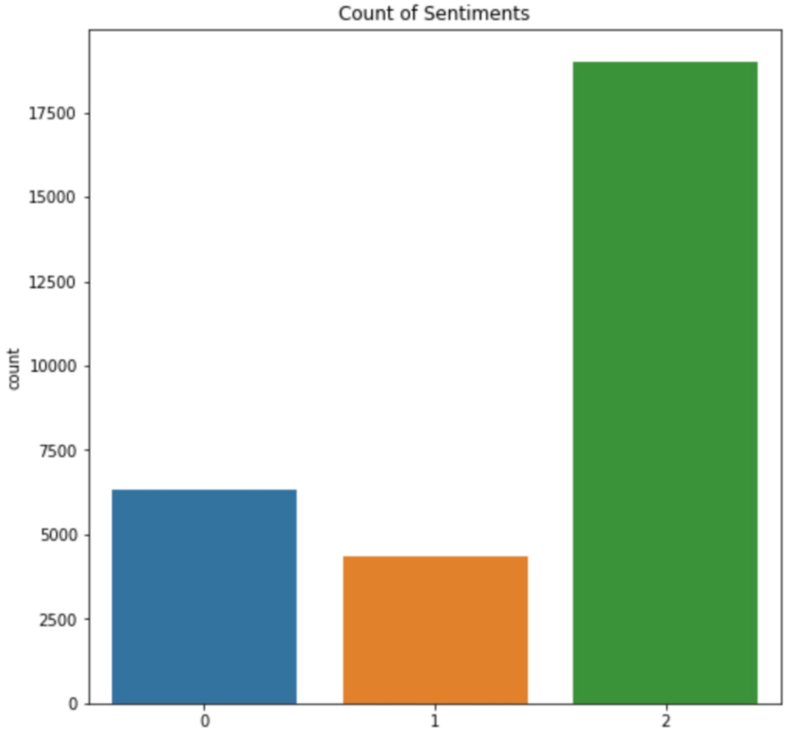


4.5 Visualization: Newly Updated Apps Have Higher ratings



- As the difference in days between today and last update date become smaller, the rating gets higher.

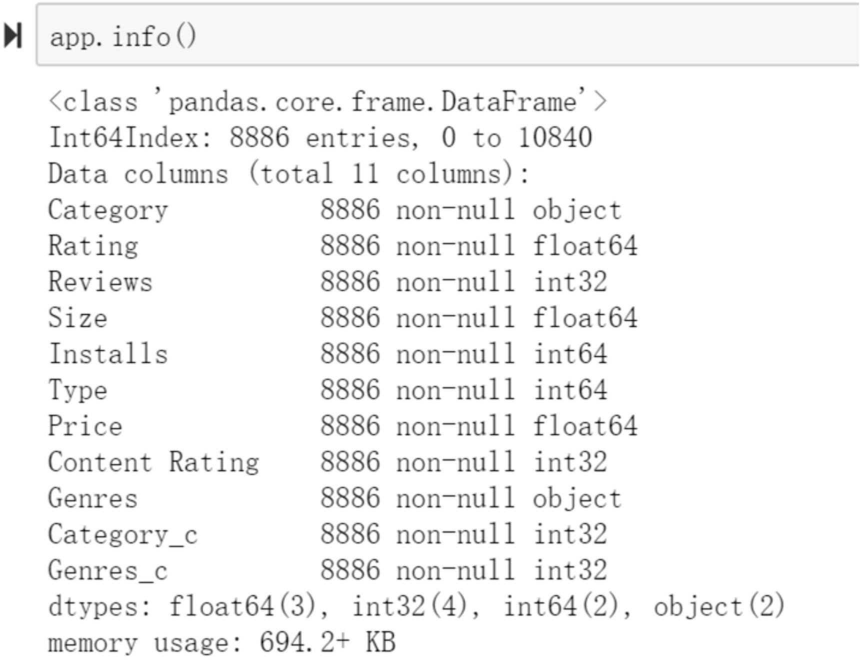
4.6 Visualization: Sentiments of Reviews



- Most reviews are positive, with total counts of 19015.
- Less reviews are neutral (4356) and negative (6321).

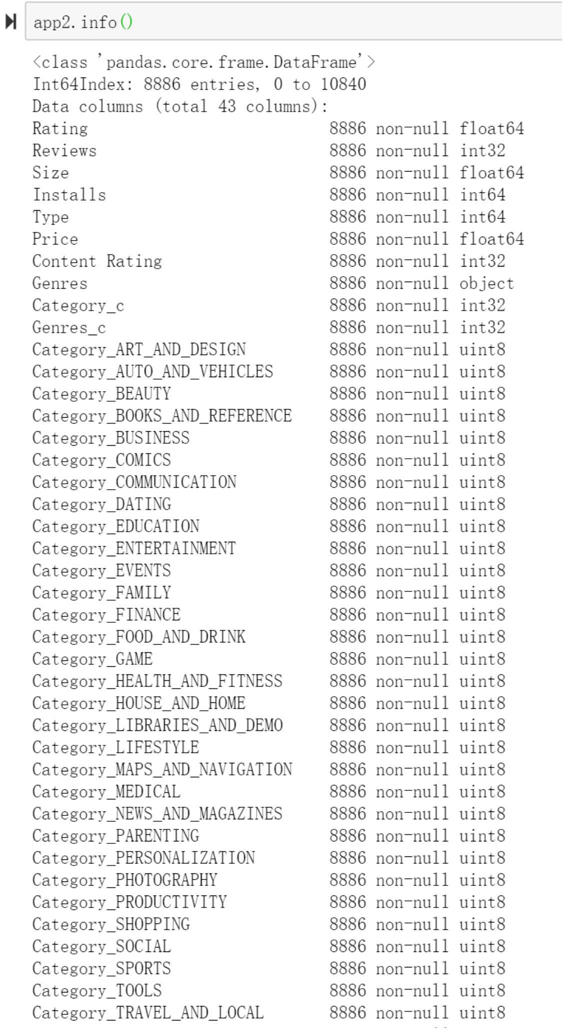
5.1 Predictive Model: Data Processing

Transfer all variables into numeric and scalling variables



```
app.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8886 entries, 0 to 10840
Data columns (total 11 columns):
Category          8886 non-null object
Rating            8886 non-null float64
Reviews           8886 non-null int32
Size              8886 non-null float64
Installs          8886 non-null int64
Type              8886 non-null int64
Price             8886 non-null float64
Content Rating    8886 non-null int32
Genres            8886 non-null object
Category_c        8886 non-null int32
Genres_c          8886 non-null int32
dtypes: float64(3), int32(4), int64(2), object(2)
memory usage: 694.2+ KB
```

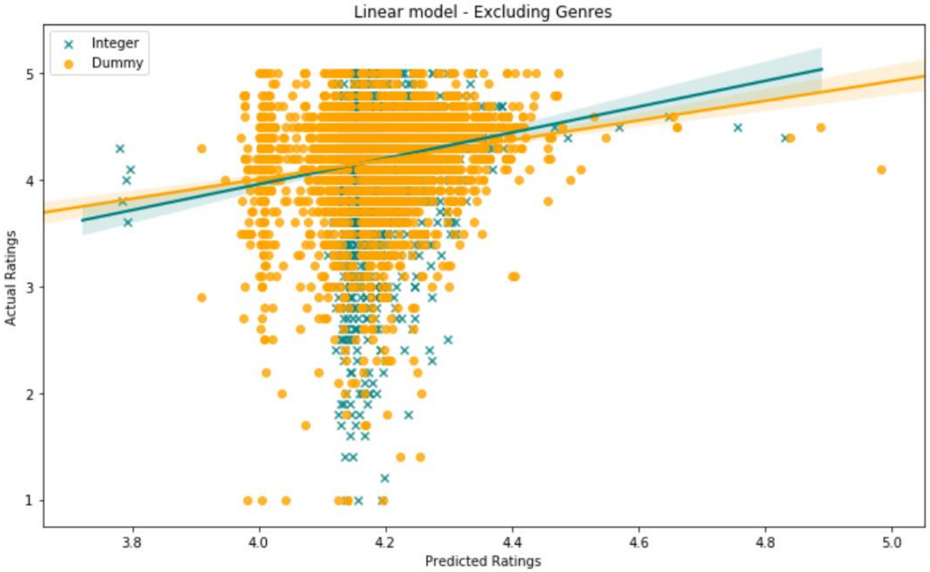


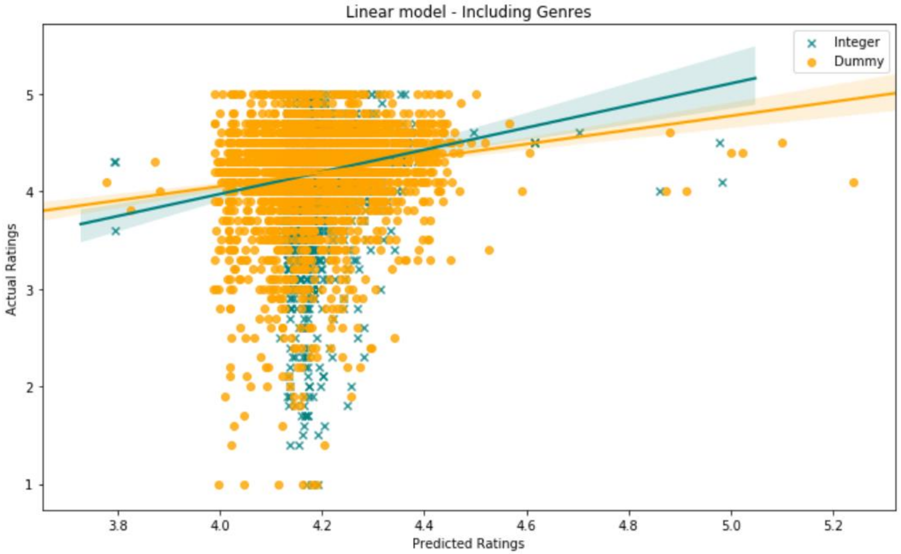
```
app2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8886 entries, 0 to 10840
Data columns (total 43 columns):
Rating                          8886 non-null float64
Reviews                         8886 non-null int32
Size                            8886 non-null float64
Installs                        8886 non-null int64
Type                            8886 non-null int64
Price                           8886 non-null float64
Content Rating                  8886 non-null int32
Genres                          8886 non-null object
Category_c                      8886 non-null int32
Genres_c                        8886 non-null int32
Category_ART_AND_DESIGN         8886 non-null uint8
Category_AUTO_AND_VEHICLES      8886 non-null uint8
Category_BEAUTY                 8886 non-null uint8
Category_BOOKS_AND_REFERENCE    8886 non-null uint8
Category_BUSINESS               8886 non-null uint8
Category_COMICS                 8886 non-null uint8
Category_COMMUNICATION          8886 non-null uint8
Category_DATING                 8886 non-null uint8
Category_EDUCATION              8886 non-null uint8
Category_ENTERTAINMENT          8886 non-null uint8
Category_EVENTS                 8886 non-null uint8
Category_FAMILY                 8886 non-null uint8
Category_FINANCE                8886 non-null uint8
Category_FOOD_AND_DRINK         8886 non-null uint8
Category_GAME                   8886 non-null uint8
Category_HEALTH_AND_FITNESS     8886 non-null uint8
Category_HOUSE_AND_HOME         8886 non-null uint8
Category_LIBRARIES_AND_DEMO     8886 non-null uint8
Category_LIFESTYLE              8886 non-null uint8
Category_MAPS_AND_NAVIGATION    8886 non-null uint8
Category_MEDICAL                8886 non-null uint8
Category_NEWS_AND_MAGAZINES     8886 non-null uint8
Category_PARENTING              8886 non-null uint8
Category_PERSONALIZATION        8886 non-null uint8
Category_PHOTOGRAPHY            8886 non-null uint8
Category_PRODUCTIVITY           8886 non-null uint8
Category_SHOPPING               8886 non-null uint8
Category_SOCIAL                 8886 non-null uint8
Category_SPORTS                 8886 non-null uint8
Category_TOOLS                  8886 non-null uint8
Category_TRAVEL_AND_LOCAL       8886 non-null uint8
```

5.2 Predictive Model: Prediction I - Linear Regression

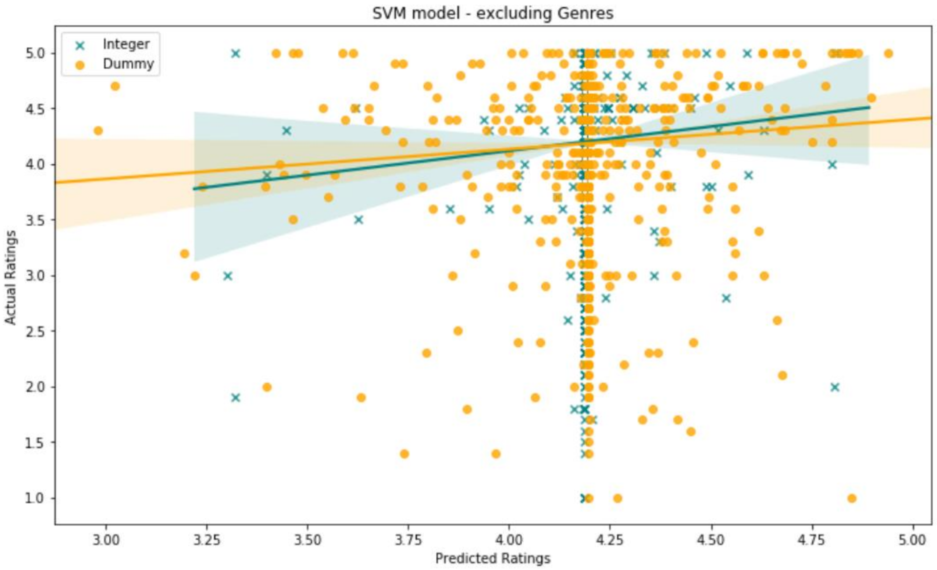
Excluding Genres (categorical variable) vs Including Genres (categorical variable)

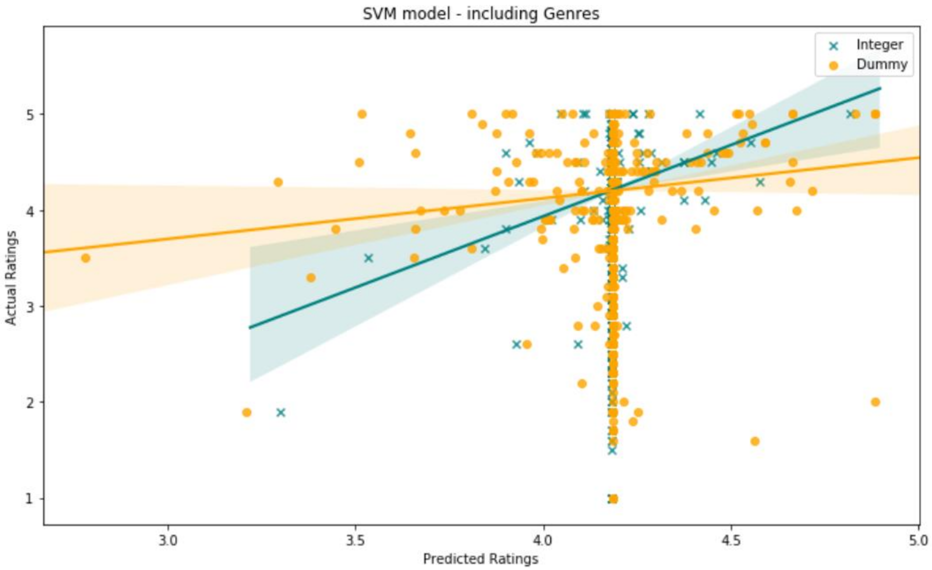




5.3 Predictive Model: Prediction II - SVR model

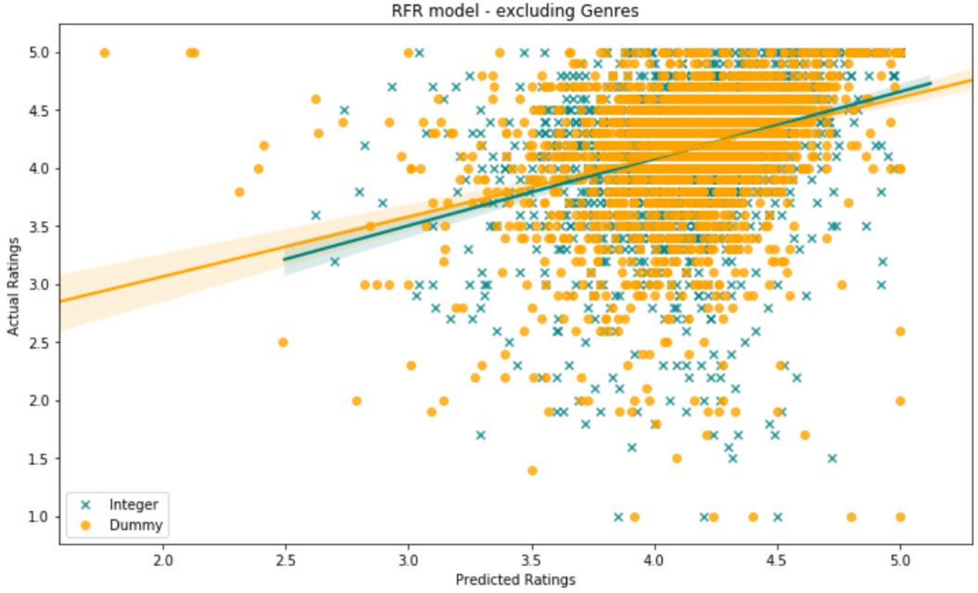
Excluding Genres (categorical variable) vs Including Genres (categorical variable)

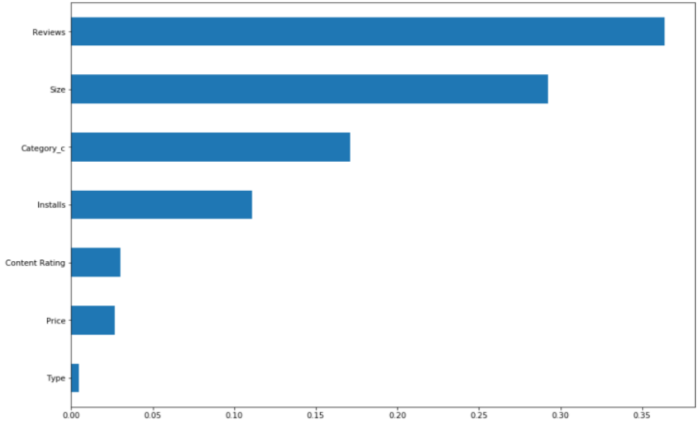


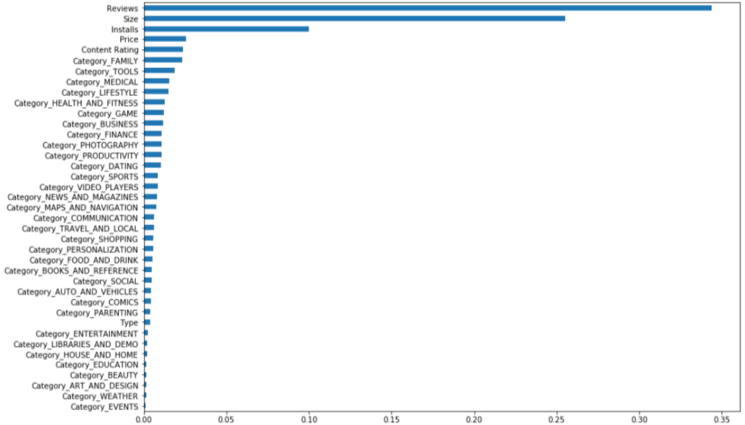


5.4 Predictive Model: Prediction III - Random Forest model

Excluding Genres (categorical variable) and Feature Importance (for integer and dummy)

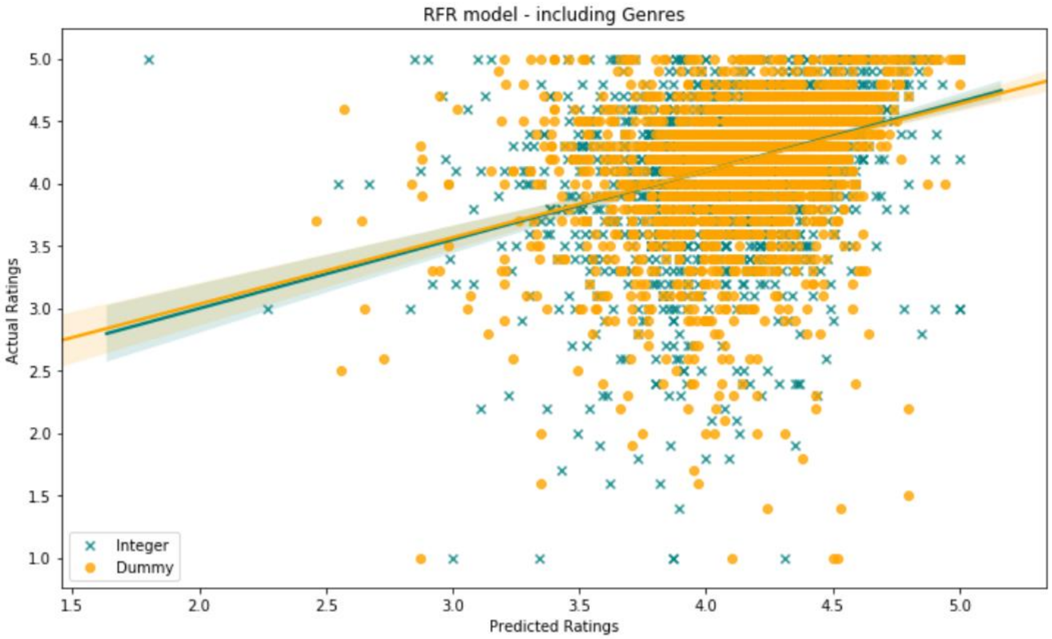


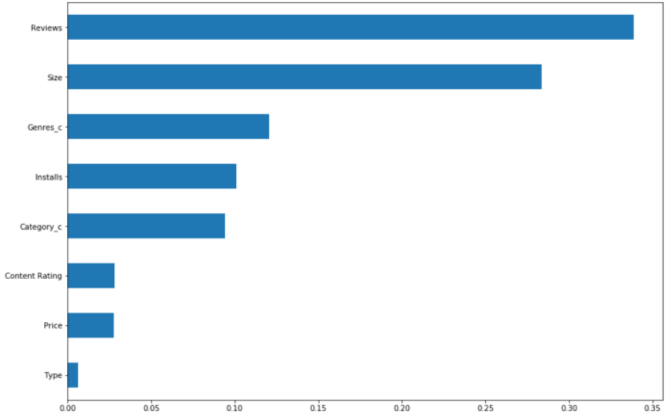


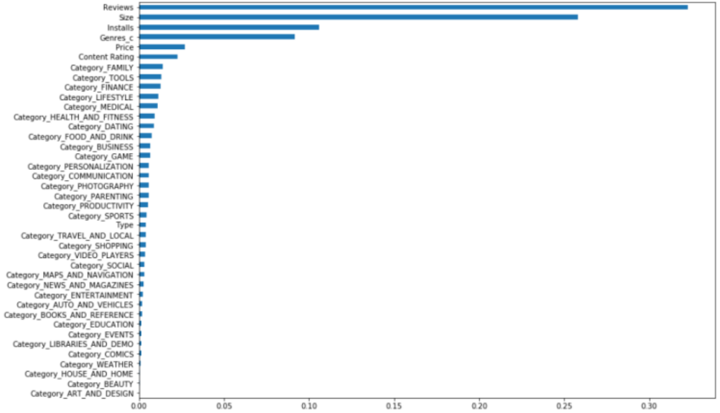


5.4 Predictive Model: Prediction III - Random Forest model

Including Genres (categorical variable) and Feature Importance (for integer and dummy)

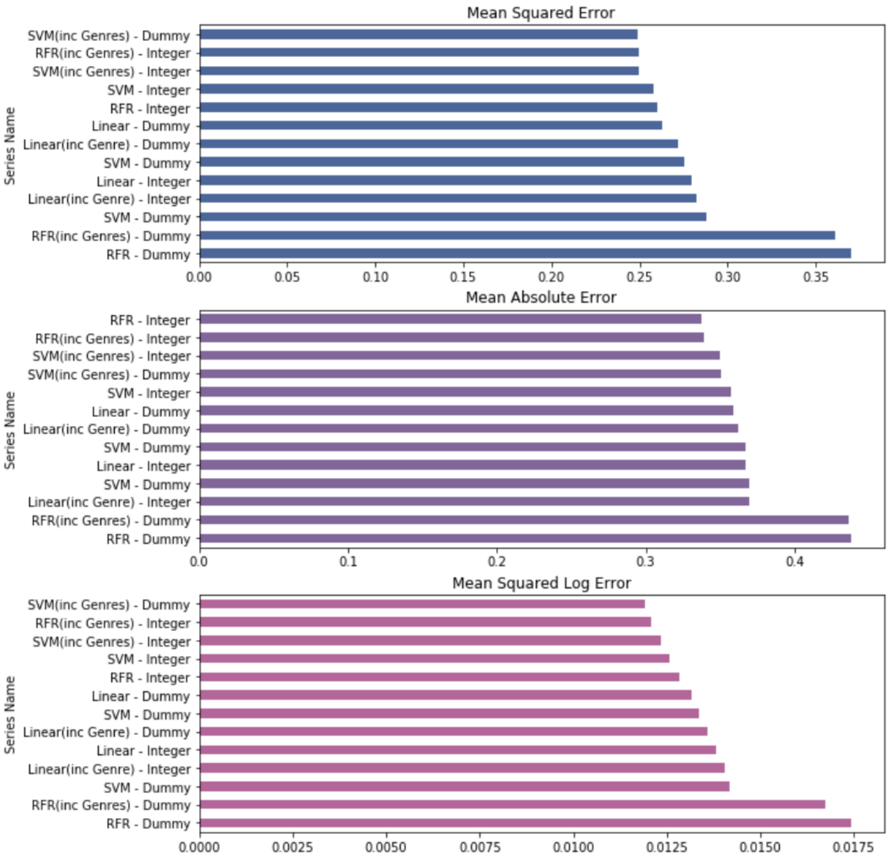






5.5 Predictive Model: Prediction III - Results

Including Genres (categorical variable) and Feature Importance (for integer and dummy)



Linear Regression Model:

```
Out[122]: 'Accuracy: 3.17%'
```

SVM Model:

```
Out[144]: 'Accuracy: 0.64%'
```

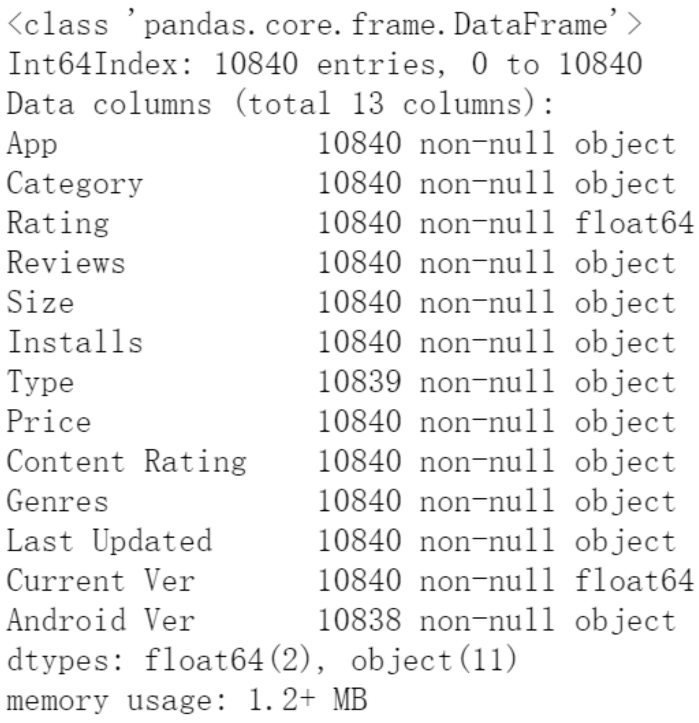
Random Forest Model:

```
Out[152]: 'Accuracy: 4.99%'
```

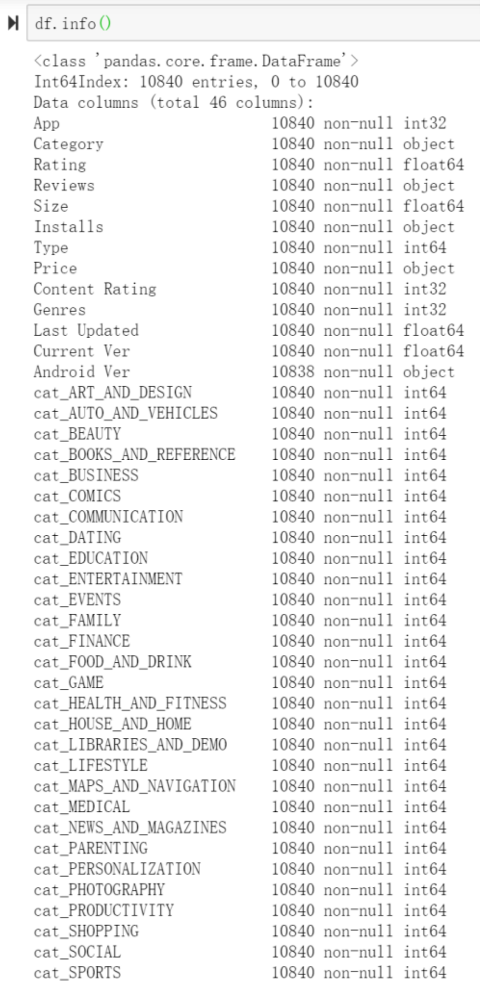
5.6 Predictive Model: Prediction IV - Random Forest Model

Re-processing Data | Using Median to Replace Na Values

Encoding Categorical Variables



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10840 entries, 0 to 10840
Data columns (total 13 columns):
App               10840 non-null object
Category          10840 non-null object
Rating            10840 non-null float64
Reviews           10840 non-null object
Size              10840 non-null object
Installs          10840 non-null object
Type              10839 non-null object
Price             10840 non-null object
Content Rating    10840 non-null object
Genres            10840 non-null object
Last Updated      10840 non-null object
Current Ver       10840 non-null float64
Android Ver       10838 non-null object
dtypes: float64(2), object(11)
memory usage: 1.2+ MB
```

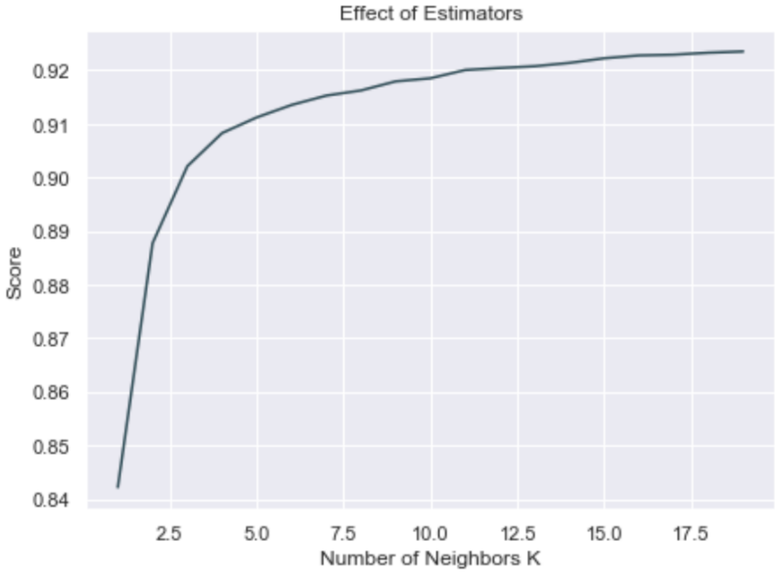


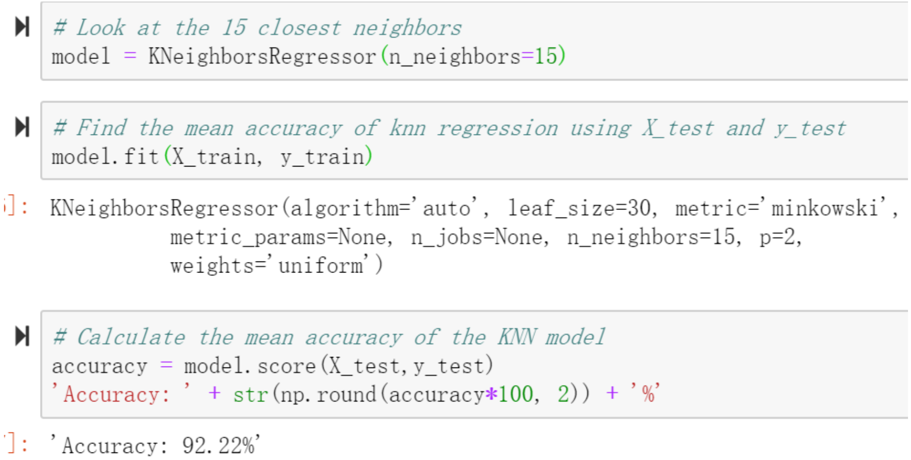
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10840 entries, 0 to 10840
Data columns (total 46 columns):
App                        10840 non-null int32
Category                   10840 non-null object
Rating                     10840 non-null float64
Reviews                    10840 non-null object
Size                       10840 non-null float64
Installs                   10840 non-null object
Type                       10840 non-null int64
Price                      10840 non-null object
Content Rating             10840 non-null int32
Genres                     10840 non-null int32
Last Updated               10840 non-null float64
Current Ver                10840 non-null float64
Android Ver                10838 non-null object
cat_ART_AND_DESIGN         10840 non-null int64
cat_AUTO_AND_VEHICLES      10840 non-null int64
cat_BEAUTY                 10840 non-null int64
cat_BOOKS_AND_REFERENCE    10840 non-null int64
cat_BUSINESS               10840 non-null int64
cat_COMICS                 10840 non-null int64
cat_COMMUNICATION          10840 non-null int64
cat_DATING                 10840 non-null int64
cat_EDUCATION              10840 non-null int64
cat_ENTERTAINMENT          10840 non-null int64
cat_EVENTS                 10840 non-null int64
cat_FAMILY                 10840 non-null int64
cat_FINANCE                10840 non-null int64
cat_FOOD_AND_DRINK         10840 non-null int64
cat_GAME                   10840 non-null int64
cat_HEALTH_AND_FITNESS     10840 non-null int64
cat_HOUSE_AND_HOME         10840 non-null int64
cat_LIBRARIES_AND_DEMO     10840 non-null int64
cat_LIFESTYLE              10840 non-null int64
cat_MAPS_AND_NAVIGATION    10840 non-null int64
cat_MEDICAL                10840 non-null int64
cat_NEWS_AND_MAGAZINES     10840 non-null int64
cat_PARENTING              10840 non-null int64
cat_PERSONALIZATION        10840 non-null int64
cat_PHOTOGRAPHY            10840 non-null int64
cat_PRODUCTIVITY           10840 non-null int64
cat_SHOPPING               10840 non-null int64
cat_SOCIAL                 10840 non-null int64
cat_SPORTS                 10840 non-null int64
```

5.7 Predictive Model: Prediction IV - Random Forest Model

K-Nearest Neighbors Model





```
# Look at the 15 closest neighbors
model = KNeighborsRegressor(n_neighbors=15)
```

```
# Find the mean accuracy of knn regression using X_test and y_test
model.fit(X_train, y_train)
```

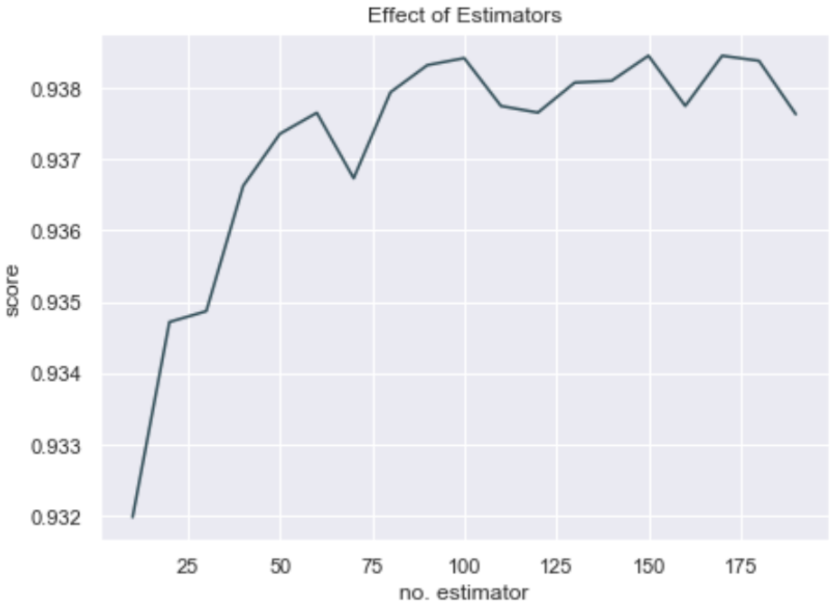
```
]: KNeighborsRegressor(algorithm='auto', leaf_size=30, metric='minkowski',
                       metric_params=None, n_jobs=None, n_neighbors=15, p=2,
                       weights='uniform')
```

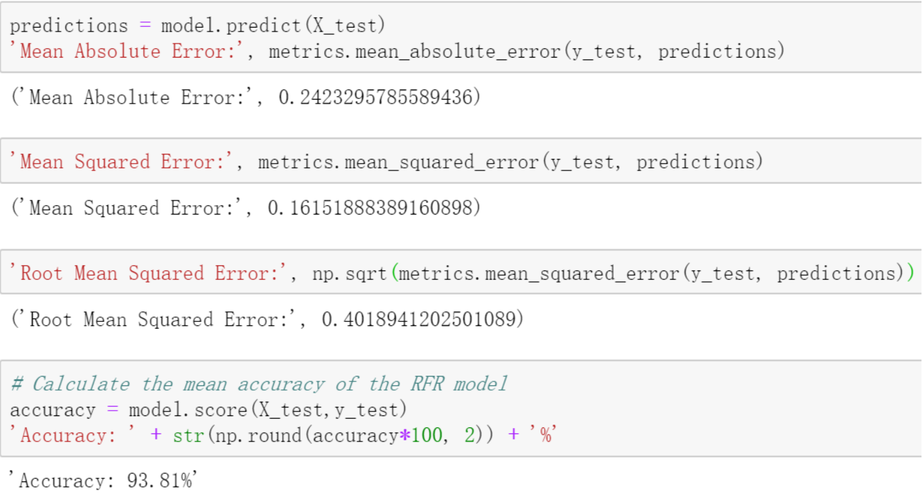
```
# Calculate the mean accuracy of the KNN model
accuracy = model.score(X_test,y_test)
'Accuracy: ' + str(np.round(accuracy*100, 2)) + '%'
```

```
]: 'Accuracy: 92.22%'
```

5.8 Predictive Model: Prediction IV - Random Forest Model

Random Forest Model





```
predictions = model.predict(X_test)
'Mean Absolute Error:', metrics.mean_absolute_error(y_test, predictions)
```

```
('Mean Absolute Error:', 0.2423295785589436)
```

```
'Mean Squared Error:', metrics.mean_squared_error(y_test, predictions)
```

```
('Mean Squared Error:', 0.16151888389160898)
```

```
'Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, predictions))
```

```
('Root Mean Squared Error:', 0.4018941202501089)
```

```
# Calculate the mean accuracy of the RFR model
accuracy = model.score(X_test,y_test)
'Accuracy: ' + str(np.round(accuracy*100, 2)) + '%'
```

```
'Accuracy: 93.81%'
```

6. Conclusion

Text Mining Analysis Conclusion:

- Attitude towards apps could be analyzed by the subjectivity on the reviews.
- Machine Learning could be employed into word sentiment analysis

Visualization Conclusion:

- Type of Apps, Price, Number of Reviews, and Update Date all affect the ratings of the Apps.

Prediction Conclusion:

- From the prediction models we come to the conclusion that, random forest tree model is the best prediction model for ratings. The accuracy is 93.81%, and the RMSE is 0.4034.

Thanks