GOOGLE PLAY STORE APPS

RATING PREDICTION

HRISAV BHOWMICK

A20011
MSE PROJECT
PRAXIS BUSINESS SCHOOL

CONTENTS

INTRODUCTION

Google Play is a digital distribution service operated and developed by Google Inc. It serves as the official app store for the Android operating system, allowing users to browse and download. Applications are available through Google Play either free of charge or at a cost. They can be downloaded directly on an Android device through the mobile app or by deploying the application to a device from the Google Play website.

The aim of the analysis is to provide insights about android applications. Deeper dive in data will help to find out the factors of influences on an application. An analysis on category, content ratings, type, price, ratings and installs would be carried out for this purpose. It would help to find out how they are inter related and will help android developers to analyze and understand the factors behind the ratings of the apps. The ultimate objective is to predict the rating of the apps based on the information provided.

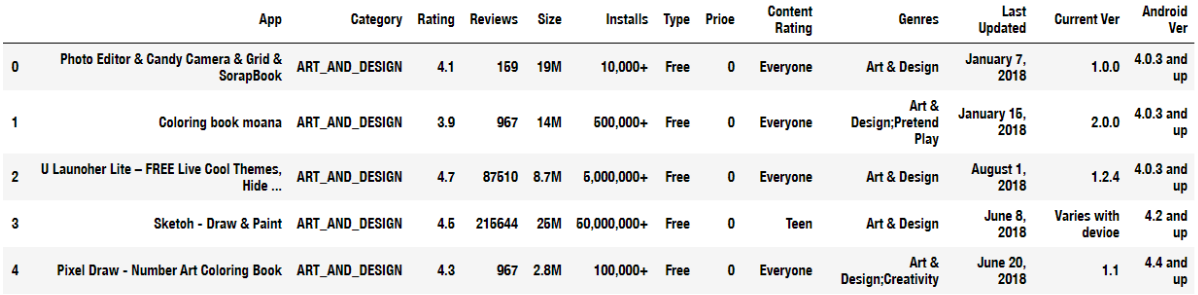
DATASET

The dataset is extracted from Kaggle. It contains details of Android applications from Google Play. There are 13 includes that depict each application and an aggregate of 10,841 applications. Below table consists of the column headers of the dataset and its brief description.

COLUMN HEADERS	DESCRIPTION
App	Name of the App
Category	Category of the app
Rating	Over all user rating of the app out of 5
Installs	Number of user downloads for the app
Price	Cost of the App
Content Rating	Age group the app is targeted at
Reviews	Number of reviews provided

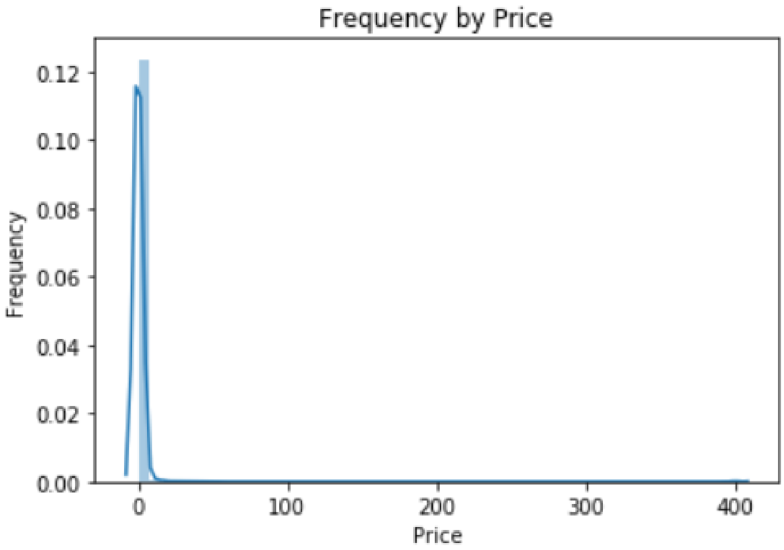
Size	Size of the app
Type	Free or Paid app
Genres	Genre of the app
Last Updated	Last updated date of the app
Current Ver	Current version of the app
Android Ver	Android version of the app

The first 5 rows of the dataset is provided below.

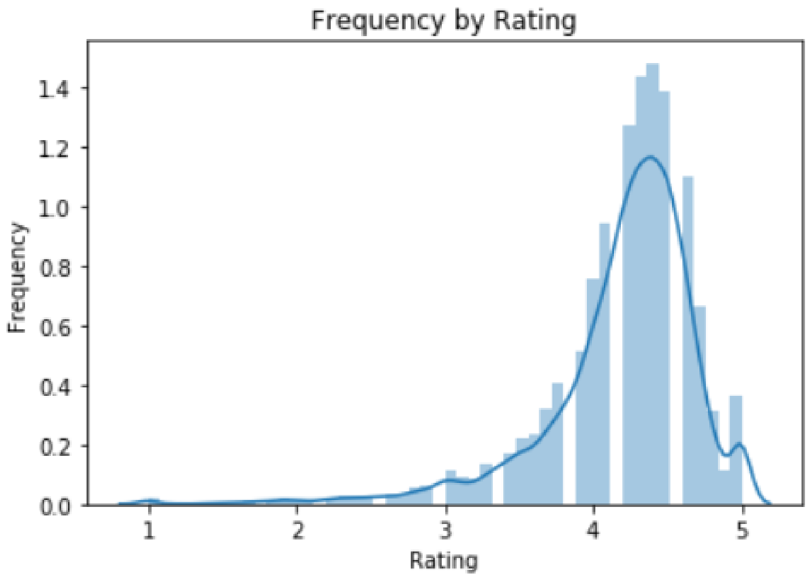


	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	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

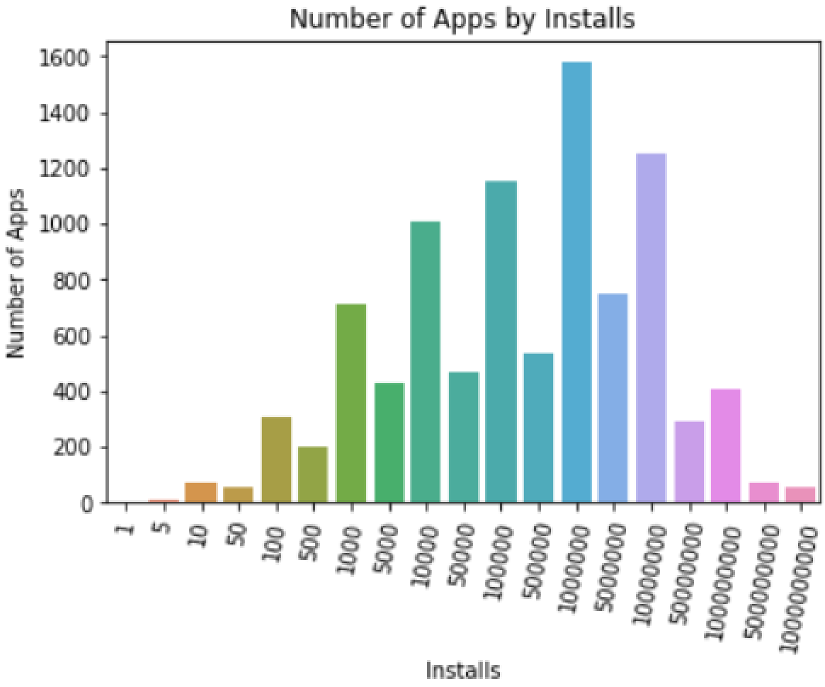
EXPLORATORY ANALYSIS



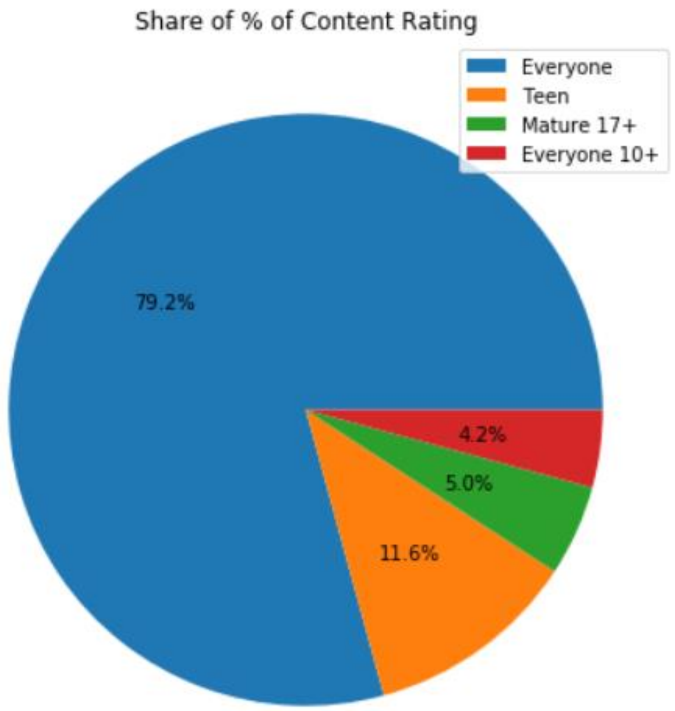
Price is positively skewed with mean value of $0.96.



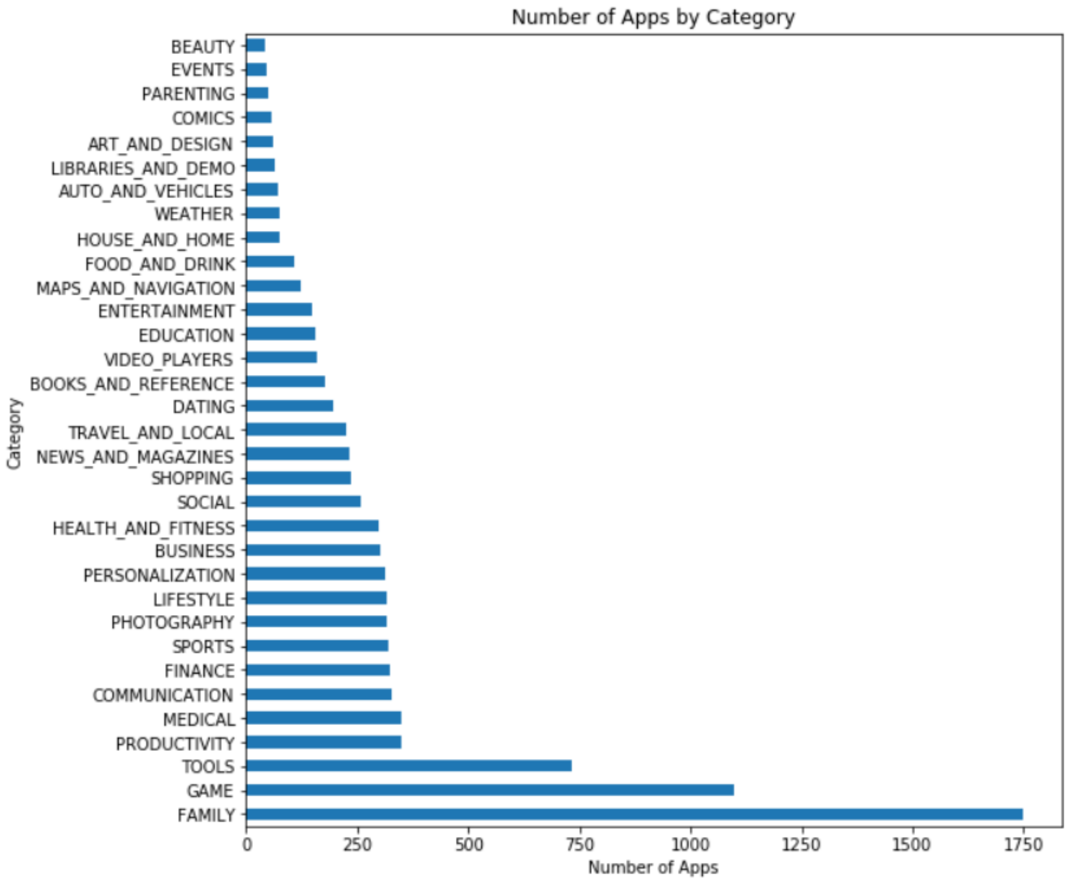
Rating is negatively skewed with mean value of 4.19.



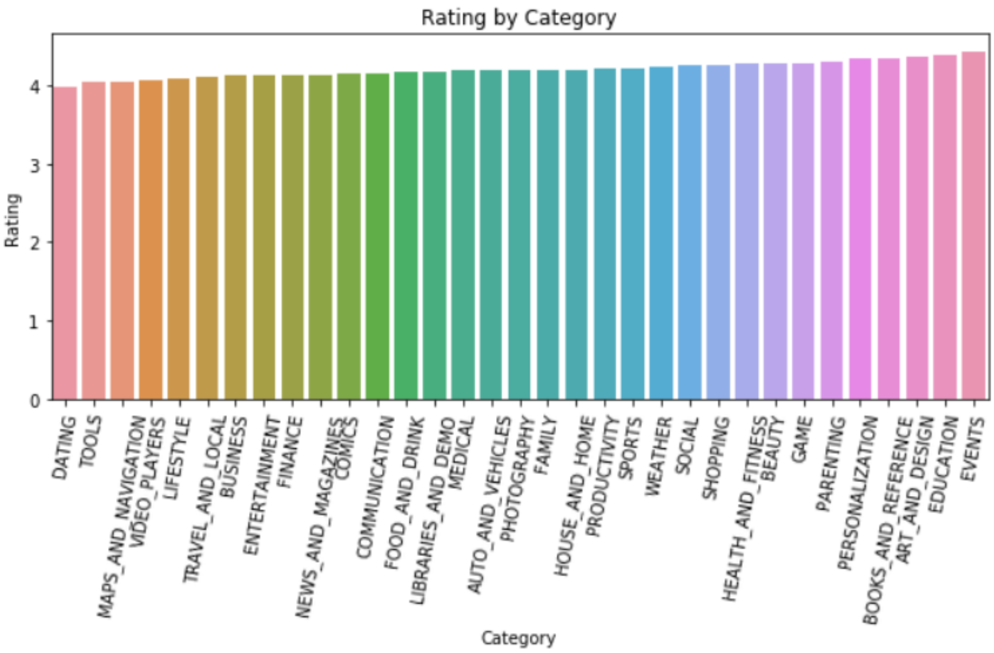
The Installs category of 1 million + has the highest number of apps.



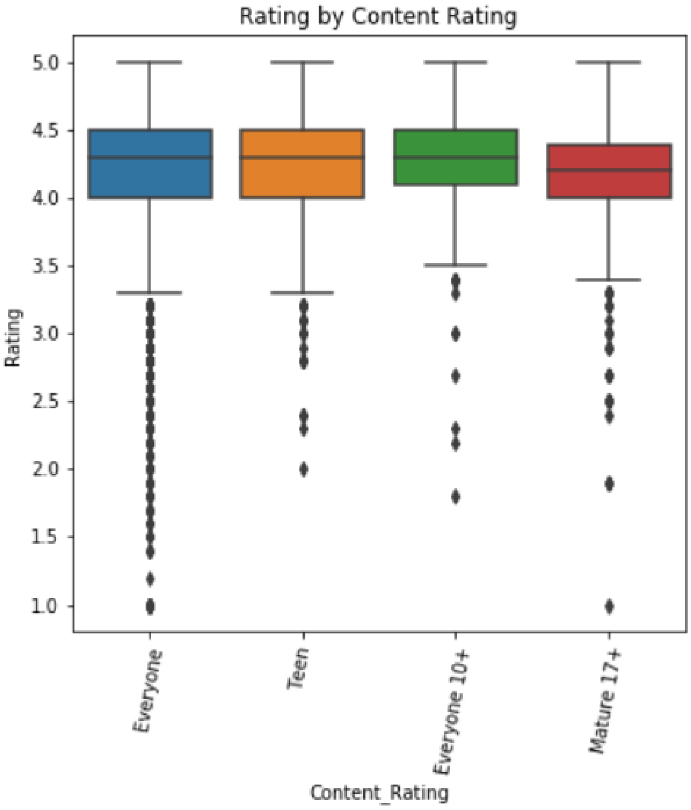
Content Rating category 'EVERYONE' has the maximum share percentage of around 79%, whereas, Content Rating category 'EVERYONE 10+' has the least share percentage of 4.2%.



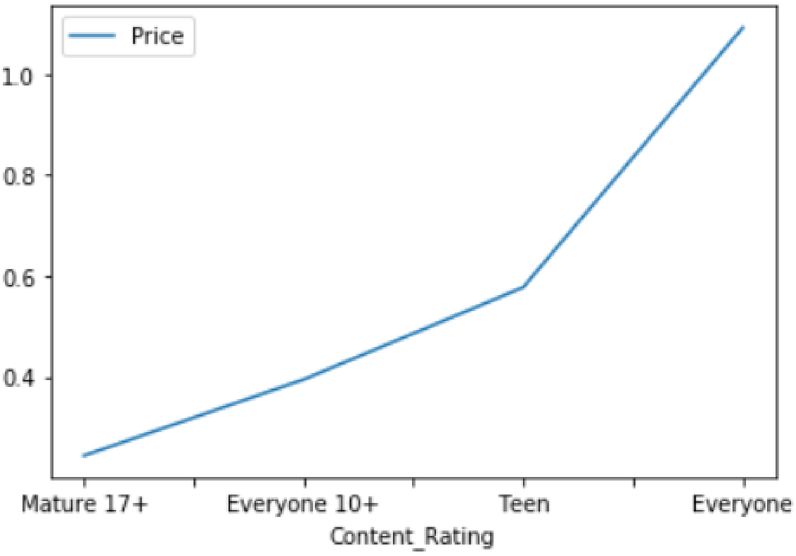
The horizontal bar plot shows that the maximum number of apps belong to the Family Category, followed by Game category. The least number of apps belong to the Beauty category.



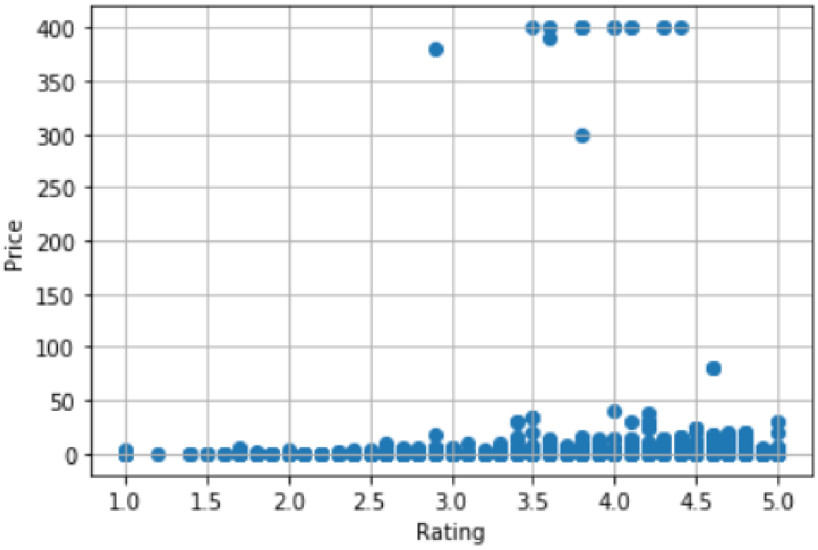
'EVENTS' category has the highest mean rating followed by 'EDUCATION'. Meanwhile, 'DATING' category has the lowest mean rating.



Apps with 'Everyone 10+' category has the highest mean rating and 'Mature 17+' has the lowest mean rating. Inter quartile range for Everyone and Teen categories is greater than Mature 17+ and Everyone 10+ categories. It means Everyone and Teen has more dispersed data.



Apps having Content Rating 'Mature 17+' has lowest average Price, where as, 'Everyone' category has highest average price.



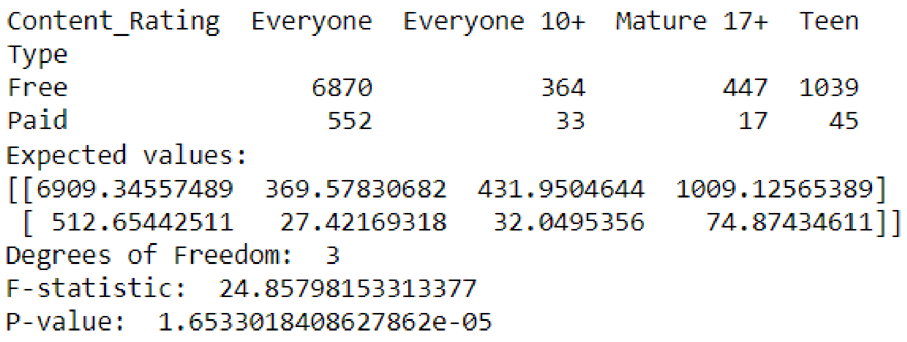
Majority of the apps priced above $50 has a rating between 3.5 to 4.5.

STATISTICAL TESTS

Chi Square Test

H0 - Type and Content Rating are dependent.

H1 - Type and Content Rating are independent.



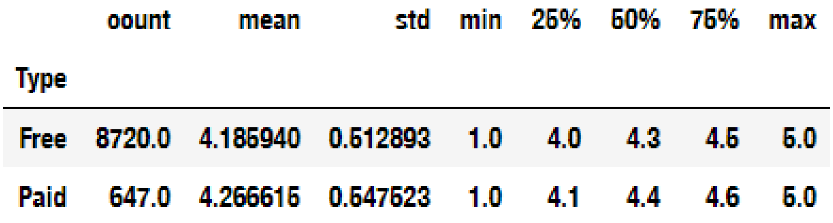
```
Content_Rating  Everyone  Everyone 10+  Mature 17+  Teen
Type
Free                6870           364         447  1039
Paid                 552            33          17    45
Expected values:
[[6909.34557489  369.57830682  431.9504644  1009.12565389]
 [ 512.65442511   27.42169318   32.0495356    74.87434611]]
Degrees of Freedom:  3
F-statistic:  24.85798153313377
P-value:  1.6533018408627862e-05
```

As f-stat > f-critical, we can say that Type and Content Rating are independent.

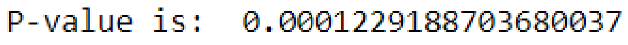
t- Test

H0 - Rating of free and paid apps is same.

H1 - Rating of free and paid apps is different.



	oount	mean	std	min	26%	60%	76%	max
Type								
Free	8720.0	4.185940	0.612893	1.0	4.0	4.3	4.6	6.0
Paid	647.0	4.266616	0.647623	1.0	4.1	4.4	4.6	6.0



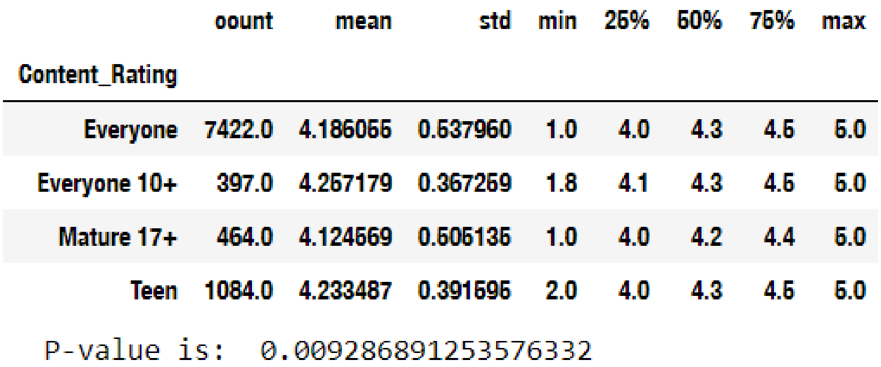
```
P-value is:  0.0001229188703680037
```

As P-value < 0.05, we can say that rating of free and paid apps is different.

ANOVA for F-Test

H0 - Mean Rating for all Content Ratings is same.

H1 - Mean Rating for one or more Content Ratings is different.

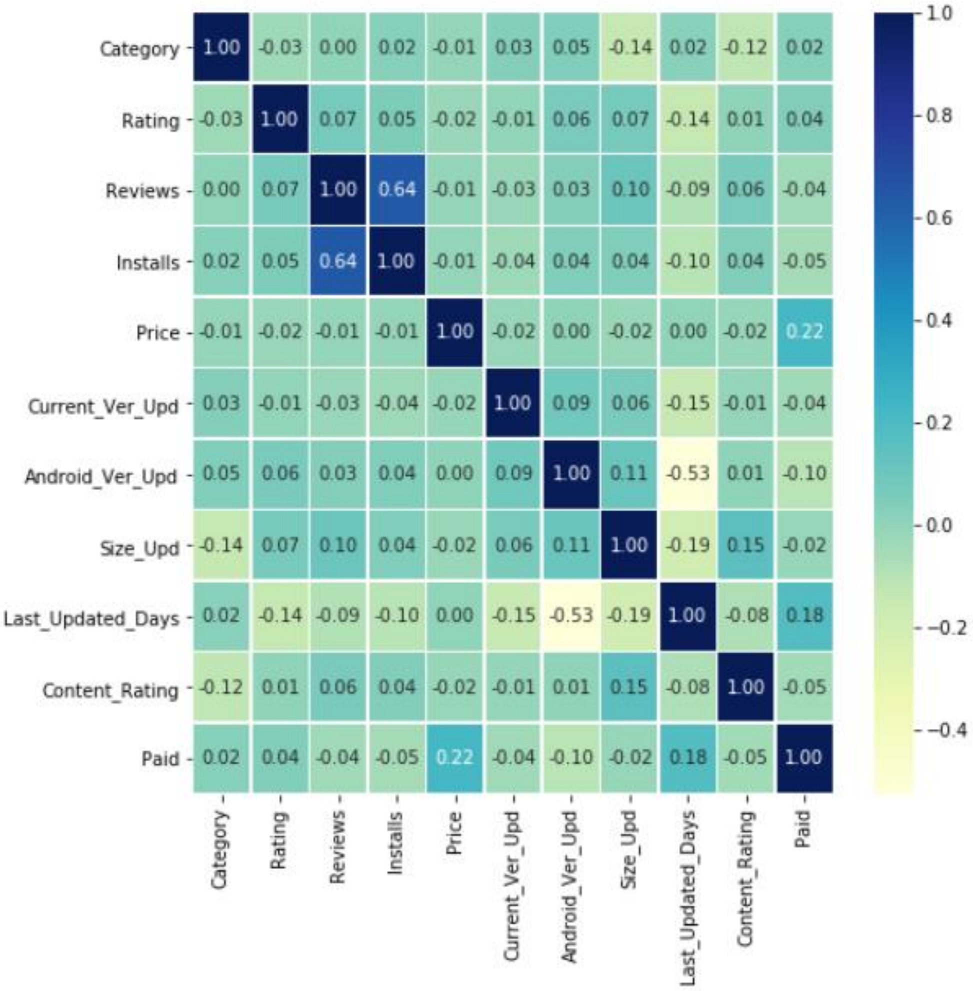


	count	mean	std	min	25%	50%	75%	max
Content_Rating								
Everyone	7422.0	4.186055	0.537960	1.0	4.0	4.3	4.5	5.0
Everyone 10+	397.0	4.257179	0.367259	1.8	4.1	4.3	4.5	5.0
Mature 17+	464.0	4.124569	0.505135	1.0	4.0	4.2	4.4	5.0
Teen	1084.0	4.233487	0.391595	2.0	4.0	4.3	4.5	5.0

```
P-value is:  0.009286891253576332
```

As P-value < 0.05, we can say that the mean rating for one or more Content Ratings is different.

CORRELATION

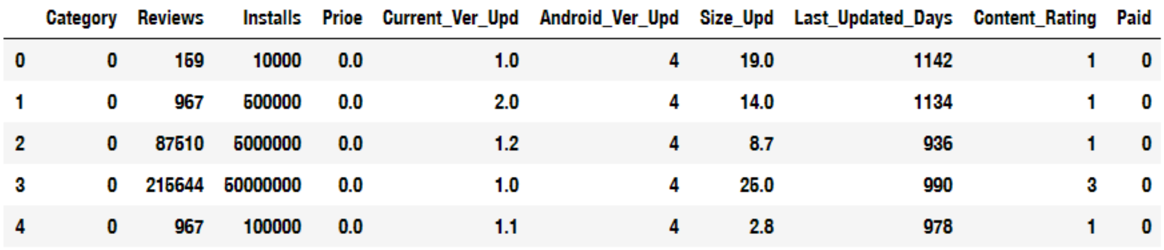


	Category	Rating	Reviews	Installs	Price	Current_Ver_Upd	Android_Ver_Upd	Size_Upd	Last_Updated_Days	Content_Rating	Paid
Category	1.00	-0.03	0.00	0.02	-0.01	0.03	0.05	-0.14	0.02	-0.12	0.02
Rating	-0.03	1.00	0.07	0.05	-0.02	-0.01	0.06	0.07	-0.14	0.01	0.04
Reviews	0.00	0.07	1.00	0.64	-0.01	-0.03	0.03	0.10	-0.09	0.06	-0.04
Installs	0.02	0.05	0.64	1.00	-0.01	-0.04	0.04	0.04	-0.10	0.04	-0.05
Price	-0.01	-0.02	-0.01	-0.01	1.00	-0.02	0.00	-0.02	0.00	-0.02	0.22
Current_Ver_Upd	0.03	-0.01	-0.03	-0.04	-0.02	1.00	0.09	0.06	-0.15	-0.01	-0.04
Android_Ver_Upd	0.05	0.06	0.03	0.04	0.00	0.09	1.00	0.11	-0.53	0.01	-0.10
Size_Upd	-0.14	0.07	0.10	0.04	-0.02	0.06	0.11	1.00	-0.19	0.15	-0.02
Last_Updated_Days	0.02	-0.14	-0.09	-0.10	0.00	-0.15	-0.53	-0.19	1.00	-0.08	0.18
Content_Rating	-0.12	0.01	0.06	0.04	-0.02	-0.01	0.01	0.15	-0.08	1.00	-0.05
Paid	0.02	0.04	-0.04	-0.05	0.22	-0.04	-0.10	-0.02	0.18	-0.05	1.00

Installs and Reviews are highly correlated with correlation of 0.64. Android_Ver_Upd and Last_Updated_Days are highly negatively correlated with correlation of -0.53. Reviews and Category has no correlation. Also Android_Ver_Upd and Price has no correlation.

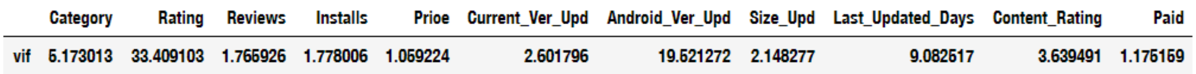
OLS REGRESSION

Here is the final output after encoding the categorical variables.



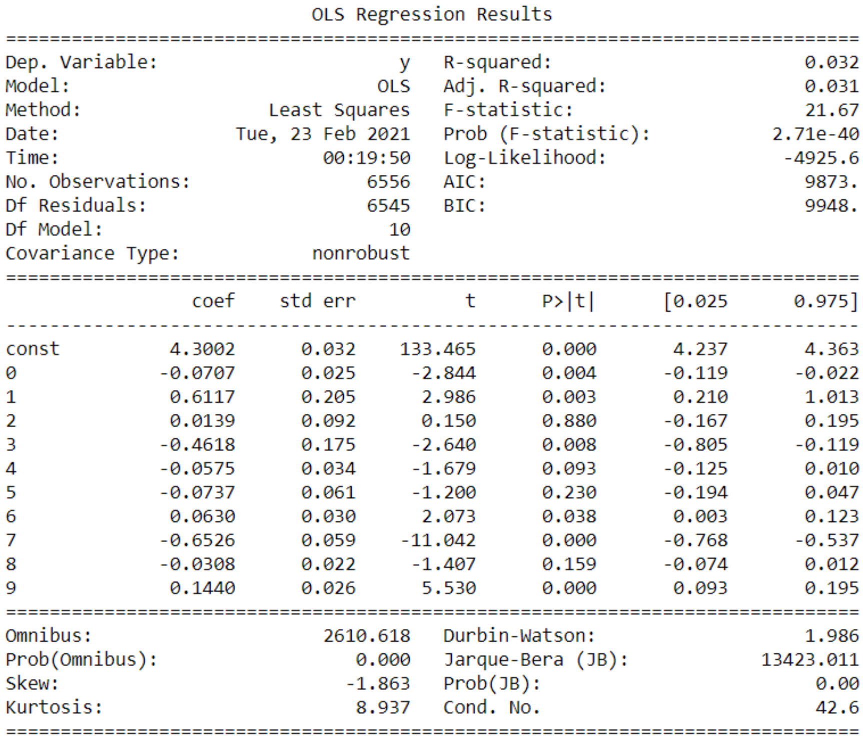
	Category	Reviews	Installs	Price	Current_Ver_Upd	Android_Ver_Upd	Size_Upd	Last_Updated_Days	Content_Rating	Paid
0	0	169	10000	0.0	1.0	4	19.0	1142	1	0
1	0	967	500000	0.0	2.0	4	14.0	1134	1	0
2	0	87510	5000000	0.0	1.2	4	8.7	936	1	0
3	0	215644	50000000	0.0	1.0	4	25.0	990	3	0
4	0	967	100000	0.0	1.1	4	2.8	978	1	0

We check the VIF scores, to understand if any variable is showing multicollinearity. Multicollinearity occurs when two or more independent variables are highly correlated with one another in a regression model.



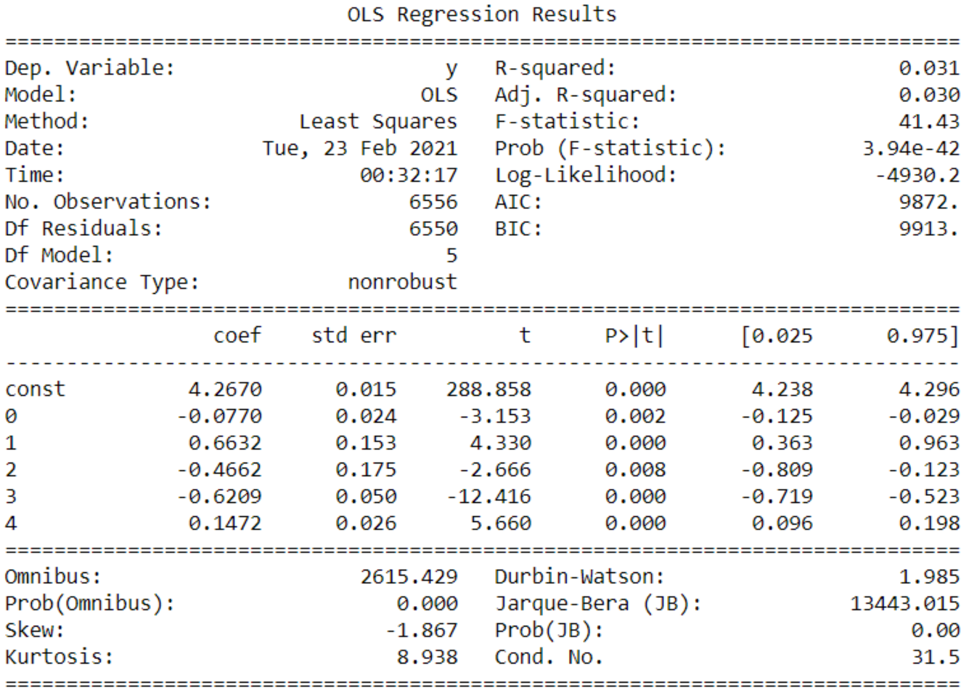
	Category	Rating	Reviews	Installs	Price	Current_Ver_Upd	Android_Ver_Upd	Size_Upd	Last_Updated_Days	Content_Rating	Paid
vif	5.173013	33.409103	1.765926	1.778006	1.059224	2.601796	19.521272	2.148277	9.082517	3.639491	1.175169

We can see a lot of variables are having VIF score more than 5, which means multicollinearity exists. As of now we will go ahead with OLS Regression and later on handle this issue.



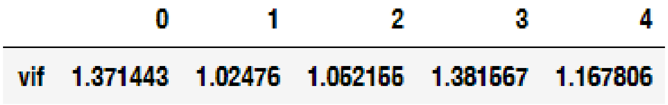
```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                      y   R-squared:                       0.032
Model:                            OLS   Adj. R-squared:                  0.031
Method:                 Least Squares   F-statistic:                     21.67
Date:                Tue, 23 Feb 2021   Prob (F-statistic):           2.71e-40
Time:                        00:19:50   Log-Likelihood:                -4925.6
No. Observations:                6556   AIC:                             9873.
Df Residuals:                    6545   BIC:                             9948.
Df Model:                          10                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.3002      0.032    133.465      0.000       4.237       4.363
0             -0.0707      0.025     -2.844      0.004      -0.119      -0.022
1              0.6117      0.205      2.986      0.003       0.210       1.013
2              0.0139      0.092      0.150      0.880      -0.167       0.195
3             -0.4618      0.175     -2.640      0.008      -0.805      -0.119
4             -0.0575      0.034     -1.679      0.093      -0.125       0.010
5             -0.0737      0.061     -1.200      0.230      -0.194       0.047
6              0.0630      0.030      2.073      0.038       0.003       0.123
7             -0.6526      0.059    -11.042      0.000      -0.768      -0.537
8             -0.0308      0.022     -1.407      0.159      -0.074       0.012
9              0.1440      0.026      5.530      0.000       0.093       0.195
==============================================================================
Omnibus:                     2610.618   Durbin-Watson:                   1.986
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            13423.011
Skew:                          -1.863   Prob(JB):                         0.00
Kurtosis:                       8.937   Cond. No.                         42.6
==============================================================================
```

P-value of F-statistic is significant, as it is less than 0.05. But individual P-values for some of the variables is insignificant as it is greater than 0.05. So we run this OLS multiple times so as to get rid of the insignificant variables.



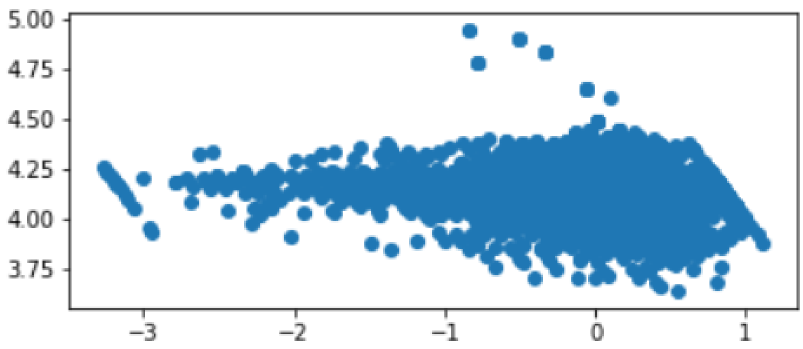
```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                      y   R-squared:                       0.031
Model:                            OLS   Adj. R-squared:                  0.030
Method:                 Least Squares   F-statistic:                     41.43
Date:                Tue, 23 Feb 2021   Prob (F-statistic):           3.94e-42
Time:                        00:32:17   Log-Likelihood:                -4930.2
No. Observations:                6556   AIC:                             9872.
Df Residuals:                    6550   BIC:                             9913.
Df Model:                           5                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.2670      0.015    288.858      0.000       4.238       4.296
0             -0.0770      0.024     -3.153      0.002      -0.125      -0.029
1              0.6632      0.153      4.330      0.000       0.363       0.963
2             -0.4662      0.175     -2.666      0.008      -0.809      -0.123
3             -0.6209      0.050    -12.416      0.000      -0.719      -0.523
4              0.1472      0.026      5.660      0.000       0.096       0.198
==============================================================================
Omnibus:                     2615.429   Durbin-Watson:                   1.985
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            13443.015
Skew:                          -1.867   Prob(JB):                         0.00
Kurtosis:                       8.938   Cond. No.                         31.5
==============================================================================
```

Finally we can see only 'Category', 'Reviews', 'Price', 'Last_Updated_Days', 'Paid' are significant for our OLS model. Also when we checked VIF score again, it shows now the multicollinearity problem has been tackled.



	0	1	2	3	4
vif	1.371443	1.02476	1.062166	1.381667	1.167806

REGRESSION ASSUMPTIONS

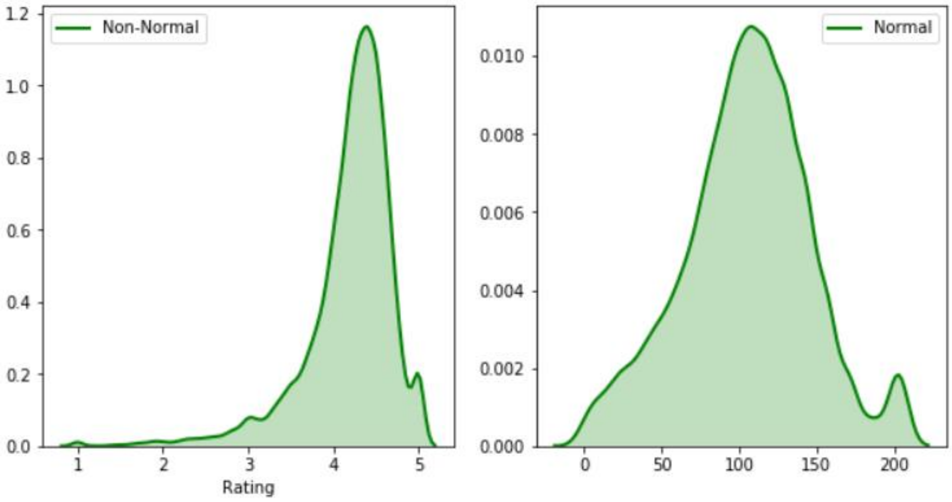


Residual plot shows a cone shaped pattern, which clearly tells us that it is heteroscedastic data. Data is non-linearly associated and some outliers is also present.

Heteroscedasticity is a systematic change in the spread of the residuals over the range of measured values. To check heteroscedasticity, we do **Breusch Pagan Test**. The null hypothesis is errors are homoscedastic. But our test shows P-value less than 0.05. Hence, we reject the null hypothesis, and our data is heteroscedastic.

Autocorrelation refers to the degree of correlation between the values of the same variables across different observations in the data. To check autocorrelation, we do **Durbin Watson Test**. As our DW value is 1.79 and it is within 0 to 2, so positive autocorrelation exists.

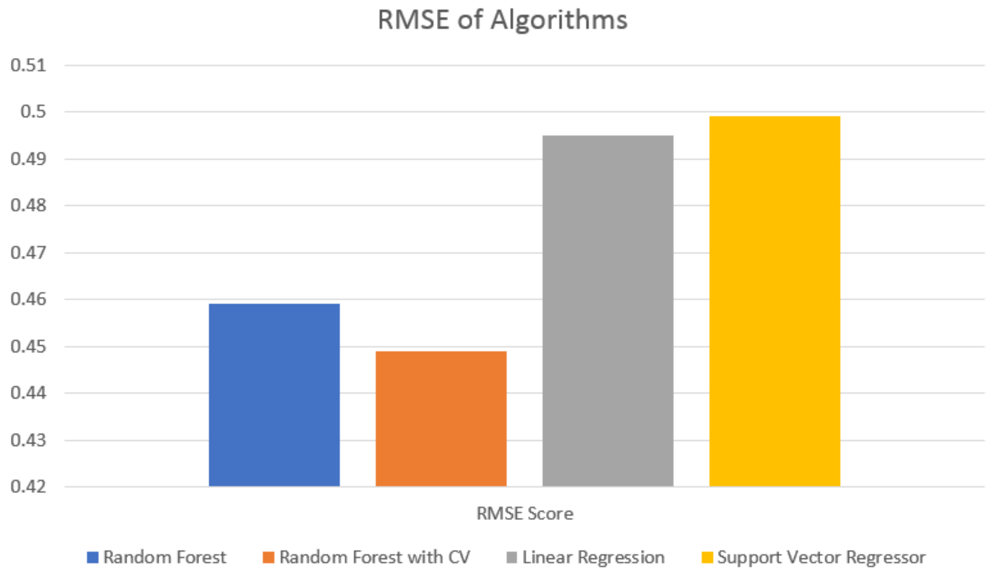
So to tackle these problems, we do a Box-Cox transformation, and make the distribution of dependent variable normal.



ALGORITHMS

Next we try to perform model building using Random Forest, Support Vector Regressor. Also we tried to find out the best result by applying Randomized Search cross validation.

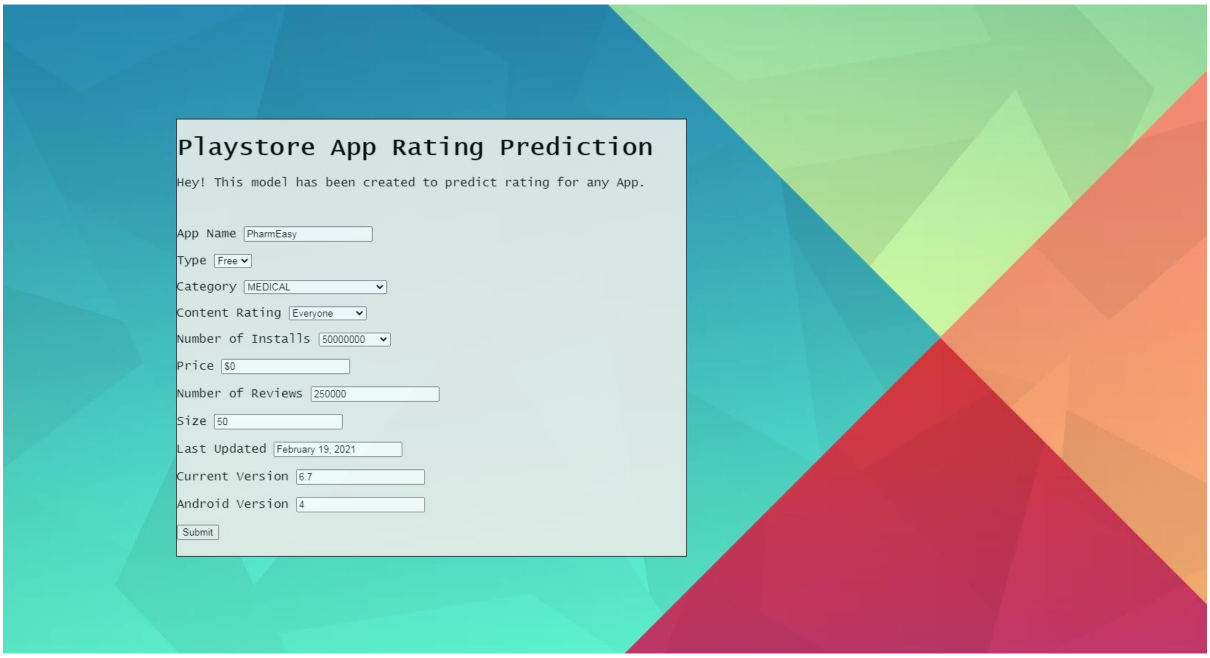
Residuals are a measure of how far from the regression line data points are. RMSE is a measure of how spread out these residuals are. In other words, it tells us how concentrated the data is around the line of best fit. Lesser the RMSE score, better is the model.



We are getting the best RMSE score when we are applying Random Forest with cross validation. So we go ahead with this algorithm for predicting the rating for new android apps.

UI BUILDING

So finally a UI was built regarding the same. We will have to put in details for any app whose rating we would like to predict. Here we have filled in with details of PharmEasy.



Playstore App Rating Prediction

Hey! This model has been created to predict rating for any App.

App Name PharmEasy

Type Free

Category MEDICAL

Content Rating Everyone

Number of Installs 50000000

Price $0

Number of Reviews 250000

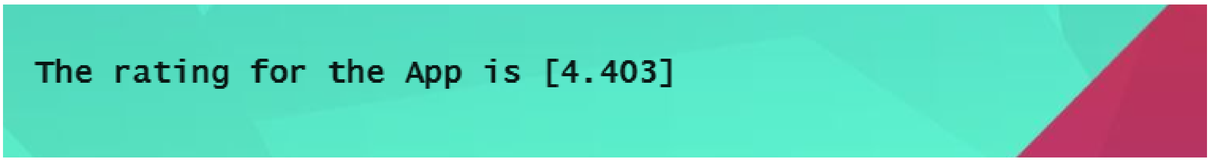
Size 50

Last Updated February 19, 2021

Current Version 6.7

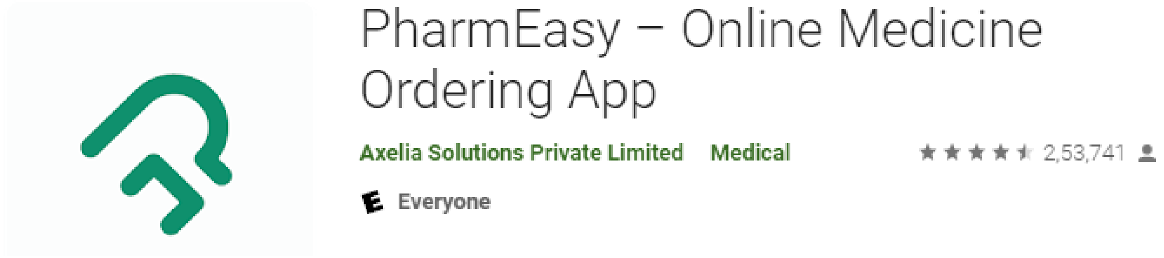
Android Version 4

Our output shows the rating as 4.403.



```
The rating for the App is [4.403]
```

When we checked the rating for PharmEasy on Google Play Store it came out to be very close.



PharmEasy – Online Medicine Ordering App

Axelia Solutions Private Limited Medical ★★★★☆ 2,53,741

Everyone

CONCLUSION

The project was started from scratch where the dataset which was taken was totally raw. A lot of cleaning was done to put it out in a presentable form. Missing values were also imputed during this process.

The motive during the project was to analyze the data and provide insights regarding them, which would in turn help the android developers to understand the factors behind the ratings of any app. The ultimate objective was to predict the ratings of the new android apps, and it was successfully completed.

During the entire process, we got a deep understanding of Statistical Tests like t-Test, ANOVA and Chi Square Test. Also we have built an understanding of OLS Regression and analyzing its outputs. Finally we tried out some other algorithms, and Random Forest with Cross Validation gave us the best RMSE score.