Play Store Apps Rating Prediction

Submitted by:

HRISAV BHOWMICK
A20011

MSE Project
PRAXIS BUSINESS SCHOOL
KOLKATA, INDIA

FEBRUARY 23, 2021

Agenda

- About the Data
- Exploratory Analysis
- Statistical Tests
- OLS Regression
- Regression Assumptions
- Model Comparison
- Conclusion

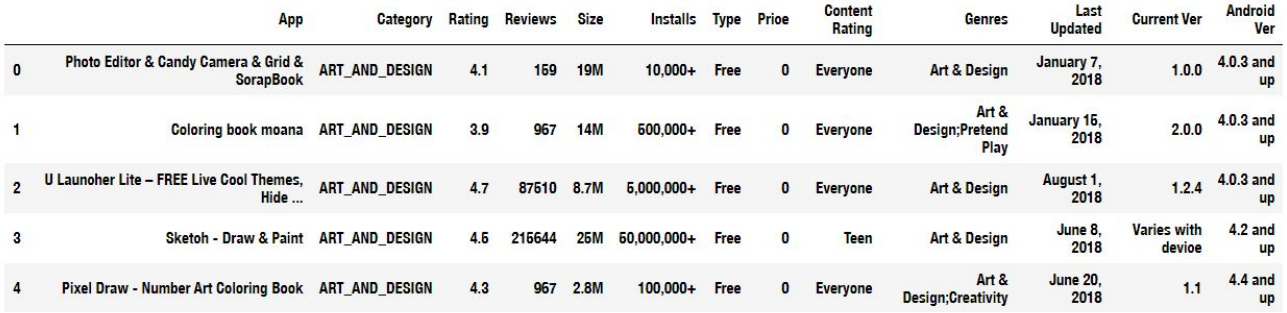
About the Data

- Dataset is about **Android Apps** from Google Play Store.

- It has **10,841 applications** and **13 features**. We are dropping those apps which has missing Rating values. So finally, we are working on **9,367 applications**.

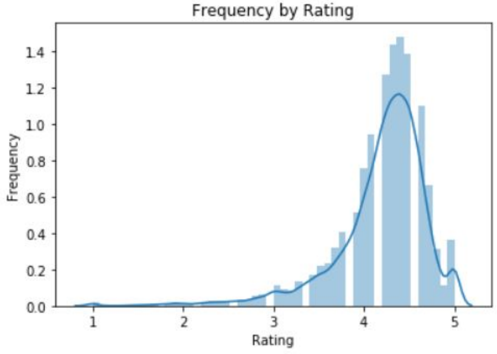
- The objective is to provide insights about the apps and to **understand the factor of influences** on these apps and ultimately, we will build a model which will **predict the ratings** of the new apps.

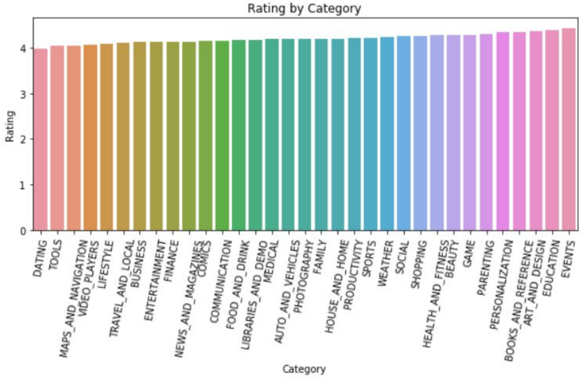
- The first 5 rows of the dataset is shown below.

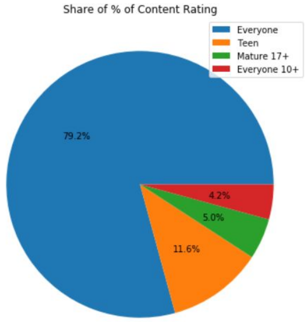


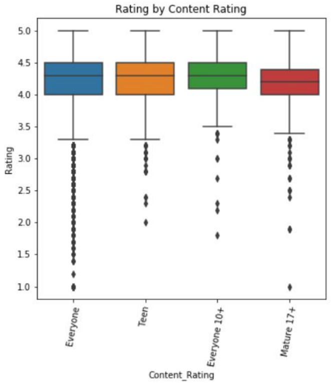
	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9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



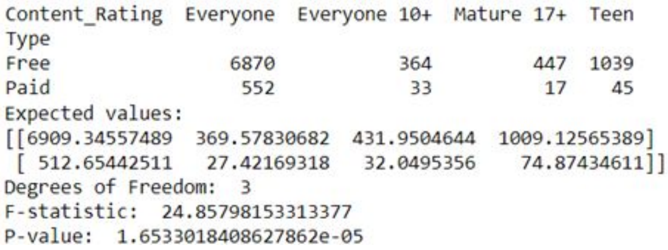






- Rating is **negatively skewed** with mean value of 4.19.
- EVENTS category has the **highest mean rating** followed by EDUCATION. Meanwhile, DATING category has the **lowest mean rating**.
- Content Rating category EVERYONE has the **maximum share percentage** of around 79%.
- Apps with EVERYONE 10+ category has the highest mean rating and MATURE 17+ has the lowest mean rating.

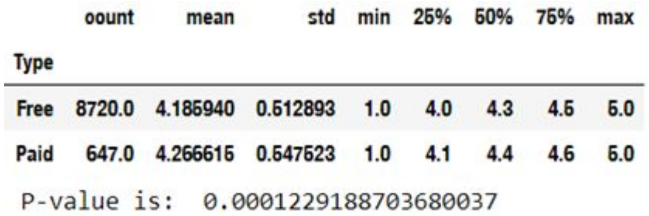
Statistical Tests



```
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
```

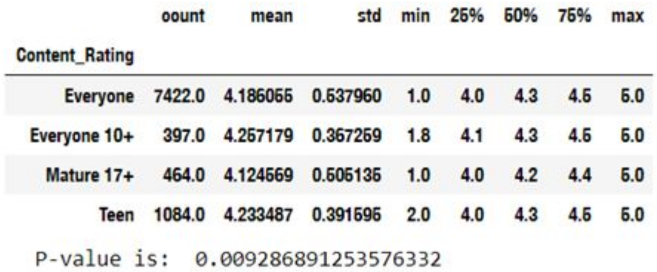
From **Chi-Square Test**, as f-stat > f-critical, we can say that Type and Content Rating are independent.

From **T-Test**, as P-value < 0.05, we can say that rating of free and paid apps is different.



	count	mean	std	min	25%	50%	75%	max
Type								
Free	8720.0	4.185940	0.512893	1.0	4.0	4.3	4.6	5.0
Paid	647.0	4.266615	0.547623	1.0	4.1	4.4	4.6	5.0

```
P-value is:  0.00012291887036800037
```

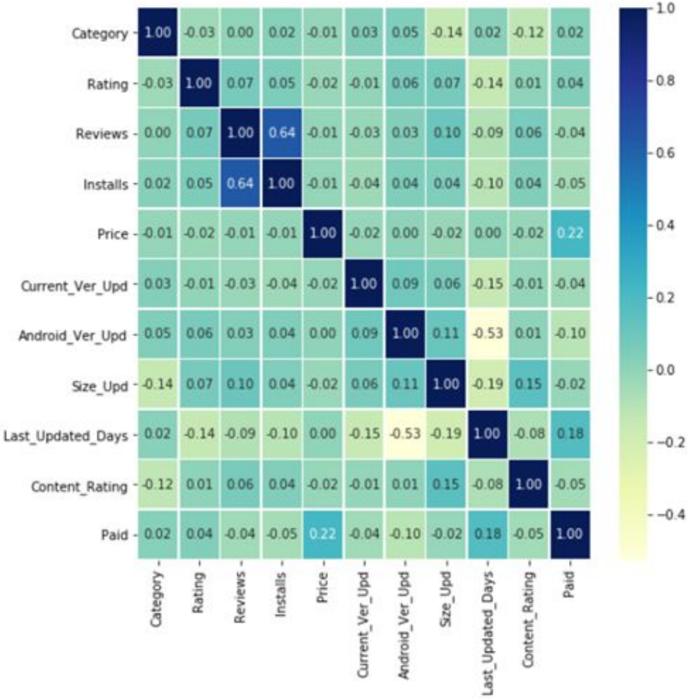


	count	mean	std	min	25%	50%	75%	max
Content_Rating								
Everyone	7422.0	4.186065	0.537960	1.0	4.0	4.3	4.5	5.0
Everyone 10+	397.0	4.257179	0.367259	1.8	4.1	4.3	4.5	5.0
Mature 17+	464.0	4.124569	0.505135	1.0	4.0	4.2	4.4	5.0
Teen	1084.0	4.233487	0.391596	2.0	4.0	4.3	4.5	5.0

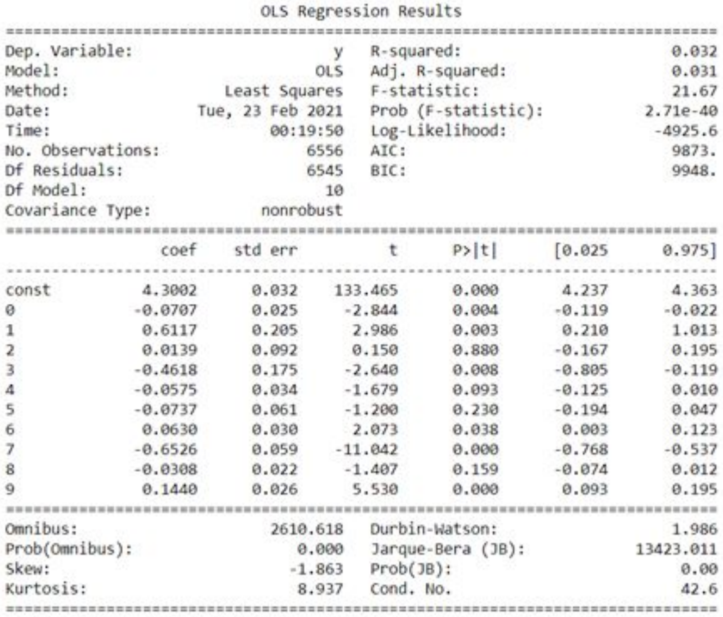
```
P-value is:  0.009286891253576332
```

From **ANOVA**, we can say that the mean rating for one or more Content Ratings is different.

OLS Regression



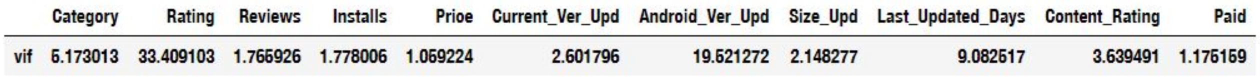
	Category	Rating	Reviews	Installs	Price	Current_Ver_Upd	Android_Ver_Upd	Size_Upd	Last_Updated_Days	Content_Rating	Paid
Category	1.00	-0.03	0.00	0.02	-0.01	0.03	0.05	-0.14	0.02	-0.12	0.02
Rating	-0.03	1.00	0.07	0.05	-0.02	-0.01	0.06	0.07	-0.14	0.01	0.04
Reviews	0.00	0.07	1.00	0.64	-0.01	-0.03	0.03	0.10	-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



```
                            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
==============================================================================
```

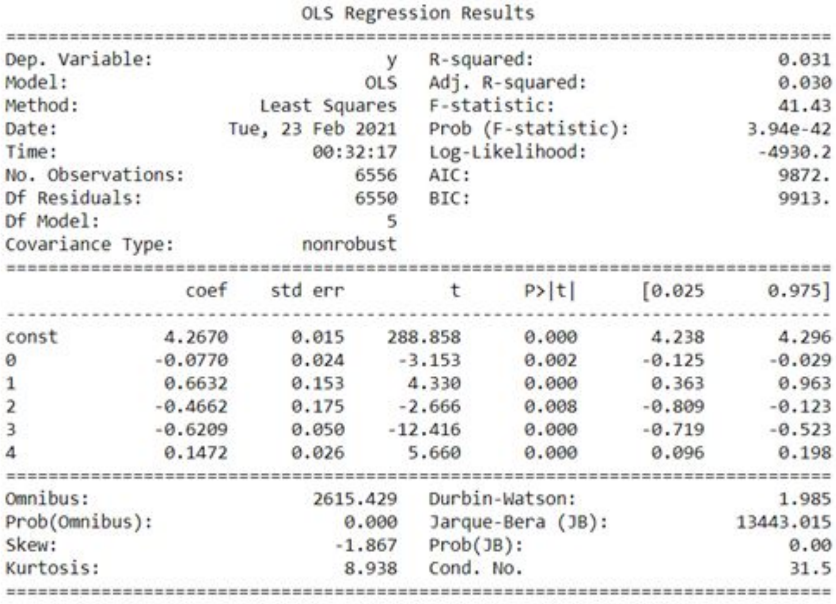
Lot of variables are having **VIF score** more than 5, which means multicollinearity exists.

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.

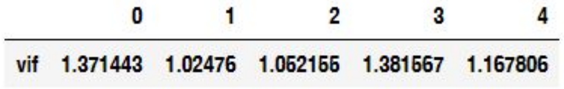


	Category	Rating	Reviews	Installs	Prioe	Current_Ver_Upd	Android_Ver_Upd	Size_Upd	Last_Updated_Days	Content_Rating	Paid
vif	6.173013	33.409103	1.766926	1.778006	1.069224	2.601796	19.621272	2.148277	9.082617	3.639491	1.176169

OLS Regression



```
                            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
==============================================================================
```



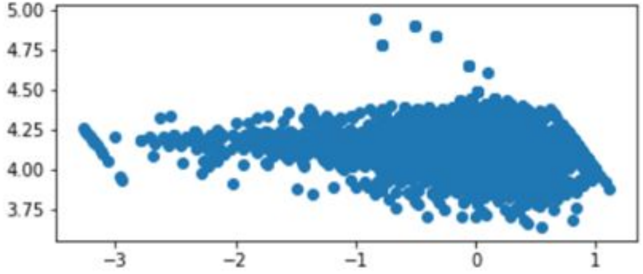
	0	1	2	3	4
vif	1.371443	1.02476	1.062166	1.381567	1.167806

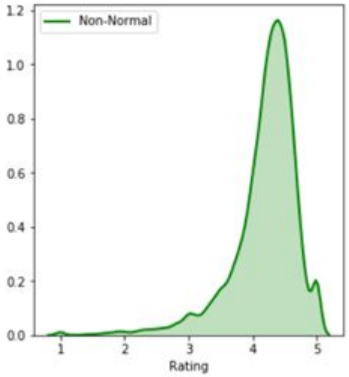
We run this OLS multiple times so as to get rid of the insignificant variables.

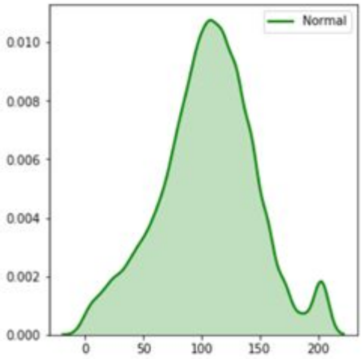
Finally we can see only 'Category', 'Reviews', 'Price', 'Last_Updated_Days', 'Paid' are **significant for our OLS model**.

VIF score shows now the **multicollinearity problem has been tackled**.

Regression Assumptions





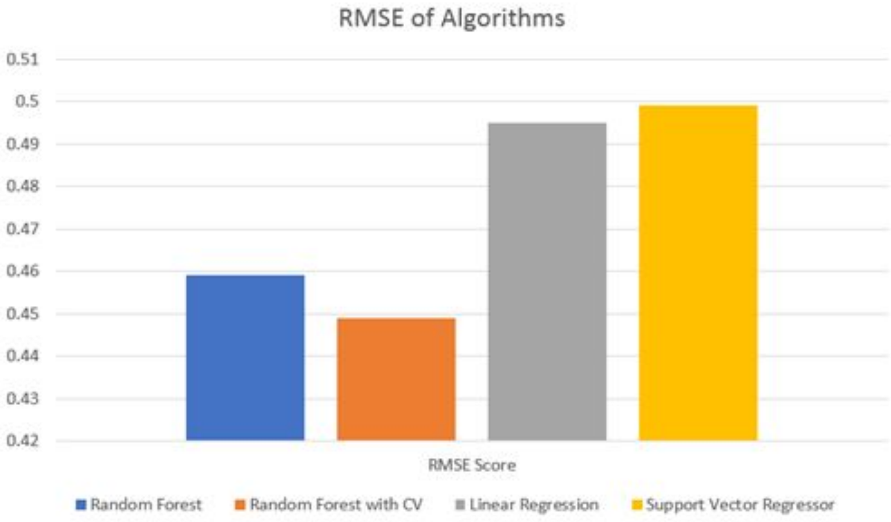


- Residual plot shows a **cone shaped pattern**, which clearly tells us that it is heteroscedastic data. Data is **non-linearly associated** and some outliers is also present.
- To check heteroscedasticity, we do **Breusch Pagan Test**. Our test shows P-value less than 0.05. Hence, we reject the null hypothesis, and our data is heteroscedastic.
- 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.

Model Comparison

- We tried 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**.

- 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. RMSE is a measure of how spread out these residuals are.



Conclusion

- We analysed the data and tried to provide insights, which would help the client to understand the factors behind the rating of any app.

- We applied OLS regression and tried to check autocorrelation and homoscedasticity of the residuals. We applied Box-Cox transformation to obtain a normal distribution of the transformed data and a constant variance. But even after doing so, it is not affecting our model.

- So we applied other algorithms, and found out Random Forest with cross validation gives the best RMSE score. So we went ahead with this algorithm for predicting the rating for new android apps.

Thank You