**Google Play Store Exploration**

**Table of contents:**

**List of Tables & Figures.**

|  |  |  |
| --- | --- | --- |
| **No. Of Contents** | **Table of Contents** | **Page No.** |
| 1 | Abstract | 6 |
| 2 | Introduction   * 1. Motivation   2.2 Project scope   * 1. Project Goal   2. Literature/Market survey   2.5 Organisation of the report | 7-11  7  8  9  10  11 |
| 3 | Project Description   * 1. Business/Domain Understanding   2. Datasets understanding   3. Data Limitations   4. Benefits of project | 12-15  12  13  14  15 |
| 4 | Exploratory Data Analysis  4.1 Data collection  4.2 Data exploration  4.3 Complexity of Data  4.4 Data cleaning  4.5 Data Transformation | 16-20  16  17  18  19  20 |
| 5 | Design   * 1. Analytical methods and Technology used.   2. Descriptive Statistical Analysis   3. Data Visualization   4. Feature Engineering   5. Short data snapshots   6. Short code snippets | 21-32  21  22  23-29  30  31  32 |
| 6 | Modelling (add extra 4 page)   * 1. Selection of model/technique   2. Challenges faced.   3. Evaluation and Cross Validation   4. Model Interpretation   5. What worked/What didn’t work   6. Short data output/snapshots   7. Short code snippets | 33-39  33  34  35  36  37  38  39 |
| 7 | Key Results   * 1. Output of intermediate steps   2. Outcome/Sample outputs   7.3 Analysis of the results | 40-42  40  41  42 |
| 8 | Conclusion   * 1. Summary of the project outcome   2. Future work | 43-44  43  44 |
| 9 | References | 45 |

1. **Abstract**

In today’s era electronic devices such as phone, tablets, computers, iPads, and lot other devices has been a most popular and trending. These sectors are rapidly growing on daily basis and this is the future. There are lot more to come soon. In those devices there are numerous applications are used by the user throughout the world to get various services and fulfil their needs.

Google Play is one among the top used and vast varieties/categories of application that is available in today’s market. Billions of people around the world use ‘Google Play Store’ to download and get their preferred services.

Similarly, there are developers who makes the applications available in play store for the consumers. The developers’ market in Google play is very vast and here developers can earn money by providing their service to the end consumer.

From this project, we hope to suggest the areas where developers will be able to generate more income.

1. **Introduction**
   1. **Motivation**

Google Play, formerly Android Market, is a [digital distribution](https://en.wikipedia.org/wiki/Digital_distribution) service operated and developed by [Google](https://en.wikipedia.org/wiki/Google). It serves as the official app store for certified devices running on the [Android operating system](https://en.wikipedia.org/wiki/Android_(operating_system)), allowing users to browse and download applications developed with the [Android software development kit](https://en.wikipedia.org/wiki/Android_software_development) (SDK) and published through Google. Google Play also serves as a [digital media](https://en.wikipedia.org/wiki/Digital_media) store, offering music, books, movies, and television programs.

It previously offered Google hardware devices for purchase until the introduction of a separate online hardware retailer, [Google Store](https://en.wikipedia.org/wiki/Google_Store). 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 [proprietary](https://en.wikipedia.org/wiki/Proprietary_software) Play Store [mobile app](https://en.wikipedia.org/wiki/Mobile_app) or by [deploying](https://en.wikipedia.org/wiki/Software_deployment) the application to a device from the Google Play website.

* 1. **Project Scope**

In this project based on the details of various application given in the dataset our scope is to find out the streams where developers can generate income through their application.

Few of the streams that developers used to generate income is through advertisement, in-purchases, buy and sell their products, subscriptions to use their services.

Another thing to consider is that paid apps can be difficult to promote in Google Play Store. Not saying that it is impossible. Certainly not. However, with free or premium apps, developers are much more likely to get users to download the app.

**2.3 Project Goal**

Our goal here is to predict the income generated through the category of apps. Constraint could be some apps might not generate income consistently, so there is a chance that prediction could go wrong in some specific cases.

To analyze the way where a developer can more likely to generate income through the app. As mentioned, there are certain ways a developer can generate income. However, we will be able to study the area where the developers are lacking in generating their income.

We will be able to provide few recommendations for the developers for their application in the ‘Google Play App Store.’ Thus, we can increase the productivity of the application and even generate income.

**2.4 Literature/Market survey**

This dataset is taken from Kaggle, this data is regarding the mobile/tablet application used widely among the users. This dataset contains 1118136 rows and 23 columns. Each attribute contains the details such as ratings, downloads, installs, price, and in-app purchases of numerous applications.

Identifying the areas where developers are lacking in terms of income generations from their application.

Building a predictive model to interpret the results.

* 1. **Organisation of the report**

Understanding the business.

Basic Data understanding.

Exploratory Data Analysis.

Feature Engineering.

Data Visualization.

Statistical Analysis.

Model Interpretation.

Model Validation.

Performance Tuning.

Interpreting Results.

Drawing Conclusion.

End recommendations for developers.

1. **Project Description** 
   1. **Business/Domain Understanding**

Google Play Store, formerly Android Market, is a digital distribution service operated and developed by Google. It serves as the official app store for certified devices running on the Android operating system, allowing users to browse and download applications developed with the Android software development kit (SDK) and published through Google.

. Google Play also serves as a digital media store, offering music, books, movies, and television programs.

The number of apps has risen back to over 3 million Android applications. As of 2017, developers in more than 150 locations could distribute apps on Google Play, though not every location supports merchant registration. Developers receive 70% of the application price, while the remaining 30% goes to the distribution partner and operating fees. Developers can set up sales, with the original price struck out and a banner underneath informing users when the sale ends.

Google Play allows developers to release early versions of apps to a select group of users, as alpha or beta tests.[8] Users can pre-order select apps (as well as movies, music, books, and games) to have the items delivered as soon as they are available.

* 1. **Dataset understanding**

This dataset is taken from Kaggle, this data is regarding the mobile/tablet application used widely among the users. The excel dataset is named as Google-Playstore.csv. This dataset contains 1118136 rows and 23 columns. Each attribute contains the details such as name, ratings, downloads, installs, price, and in-app purchases of numerous applications.

The dataset contains missing values, unwanted characters and these must be treated accordingly. Using various statistical methods, we can interpret the relationship between different attributes.

Data transformation must be done for categorical columns for modeling purposes.

* 1. **Data Limitations**

The dataset of Google Play App Stores contains various missing values. Conversion of the data, removal of unwanted characters in the dataset.

Feature Engineering to reduce the unnecessary columns in the dataset and finding the relationship between the independent features of the application. Data Transformation of the data and outliers (the extreme values must be treated).

This will improve the accuracy of the model and interpret better results

* 1. **Benefits of project**

Developers in the market will be able to explore more options to generate income for themselves. Also, can interpret the areas where developers are lacking, and issues can be minimised, and income can be generated.

Based on the category of the application, can conclude that which is the best stream to generate income for developers. We will be able to conclude the top categories of the application generating income and vice-versa.

1. **Exploratory Data Analysis**
   1. **Data collection**

This dataset is taken from Kaggle, this data is regarding the mobile/tablet application used widely among the users. This dataset contains 1118136 rows and 23 columns. Each attribute contains the details such as ratings, downloads, and installs, price and in-app purchases of numerous applications.

The excel dataset is named as Google-Playstore.csv. This data set was obtained from Kaggle and contains 25717128 records used for exploration of mobile/android applications. There are various types of application under different markets and hence are widely used for multipurpose.

* 1. **Data exploration**
* The dataset contains 23 columns and 1118136 rows with int, float and object data types.
* Unnecessary columns such as App Id, Developer Website, App Id, Minimum Android, Developer Id, Developer Website, Developer Email, and Privacy Policy should be dropped.
* Date time conversion to be done in order to find out the maintenance and tenurity of the applications.
* Removal of unwanted characters in columns and conversion of the data types for better analysis.
* Detection of outliers and treatment of missing values (3.37 % in Size, 0.69 in Released, 0.61 in Rating Count).
* Handling duplicated values in App name column.

**4.3 Complexity of Data**

* The data is not normally distributed and there are missing values that must be treated in order to make a better model for analysis.
* The values in columns differ from each other and the range is also different hence data transformation can rectify this error.
* The dataset also contains lots of outliers which effects the performance of the model, hence this should be detected and treated.
* When it comes to categorical columns, it should be encoded for machine learning and modelling purposes.

**4.4 Data cleaning**

* Dropped unnecessary columns such as App Id, Developer Website, App Id, Minimum Android, Developer Id, Developer Website, Developer Email, and Privacy Policy.
* Converted Last Updated and released column using date time package and found out tenurity and maintenance in terms of days for each application.
* Removed alphabetic and special characters in size column, minimum android column and converted the columns from object to float data type.
* Removed special characters from installs columns and converted the data type to float.
* 106647 records were deleted as it was duplicate records based on the name of the application.

**4.5 Data Transformation**

1. In order to implement machine learning to the dataset the data has to be transformed in such a way that the algorithm can study/train the data in efficient manner.
2. In this project the dataset there are different type of attributes that is categorical, numerical and Boolean values. As our target variable is Boolean we use classification algorithm.
3. However, here the categorical data is transformed to numerical.
4. We use Sklearn Pre-processing package and the encoder for the data transformation that is applied is Label Encoder.
5. As a result, we can observe that the categorical data is now converted to numerical data and then made available for implementing machine learning model for prediction.
6. Category and Currency columns are now converted to numerical column.
7. **Design** 
   1. **Analytical methods and Technology used.**
8. Google Colab.
9. Python Libraries :
   1. Data Visualisation – Seaborn and MatplotLib.
   2. Scipy for statistical analysis.
   3. Sklearn for model selection, accuracy score, confusion matric, machine learning, and f1 score etc.
   4. Date Time Package for conversion.
   5. Pandas and numpy libraries for computations and analysis.
   6. Regular Expression and string operations to remove unwanted characters’ in the dataset.

**5.2 Descriptive Statistical Analysis**

* Understood the mean, median and maximum values of each attributes.
* Understood the correlation between the columns and observed that installs and minimum installs are highly correlated.
* Figured Interquartile Range of each numerical columns which also helped to detect the outlier.
* Found out the skewness of the data. Some attributes are positively skewed.

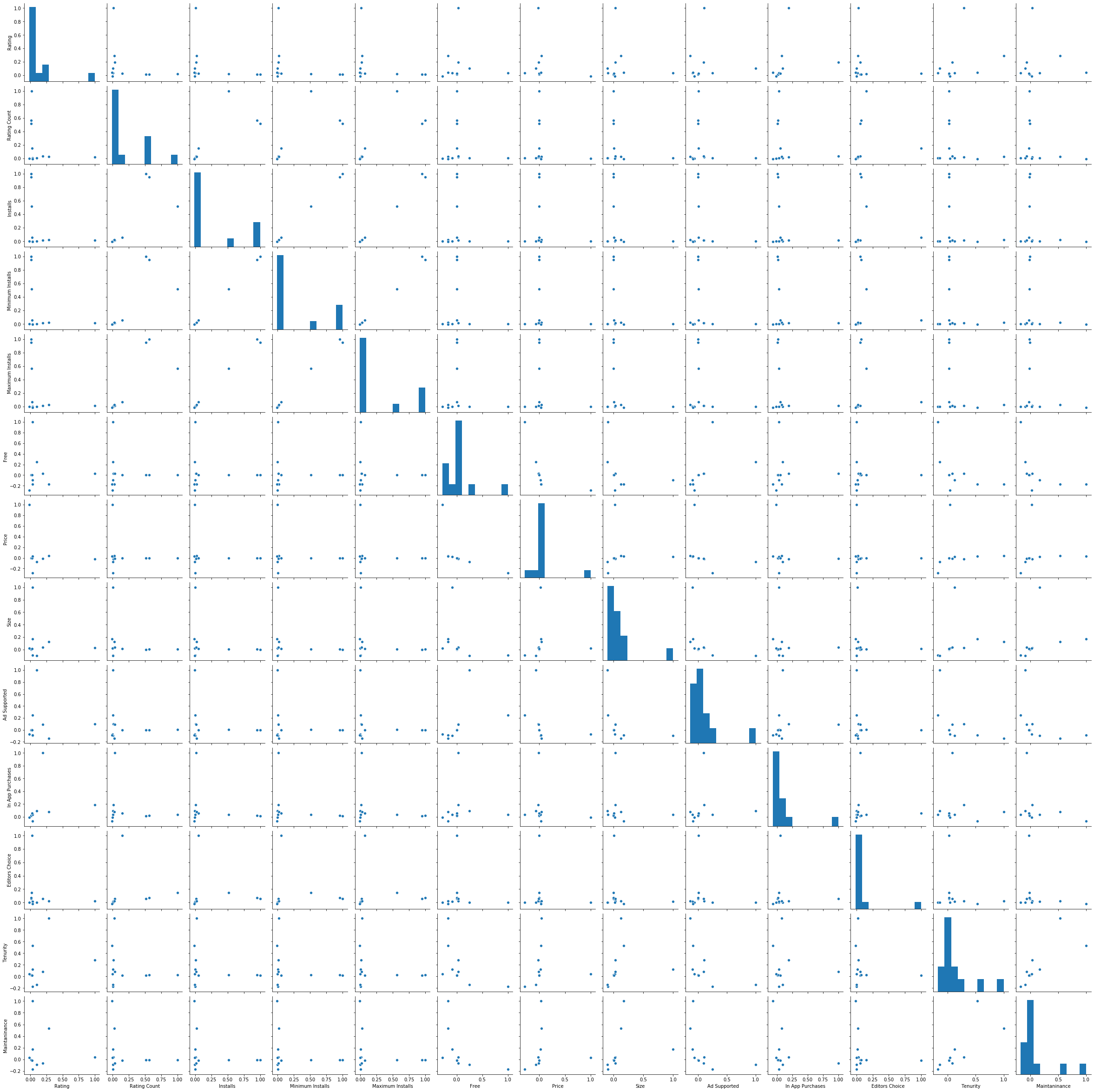
Here are the few attributes with their skewness, here we can see that currency is negatively skewed data will be piled up to the right.

|  |  |
| --- | --- |
| Free | -4.575370 |
| Ad Supported | -0.624968 |
| Rating | -0.293478 |
| Category | 0.021406 |
| Tenurity | 0.940520 |
| Maintenance | 1.160412 |

**5**.**3 Data Visualization**

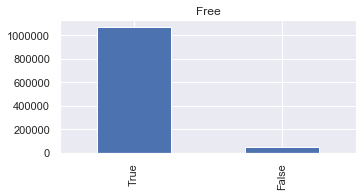
Pair Plot:

This represents the data is not normally distributed and has outliers present in the data.



**Count Plot:**

This graph represents the application those are free in the app store.

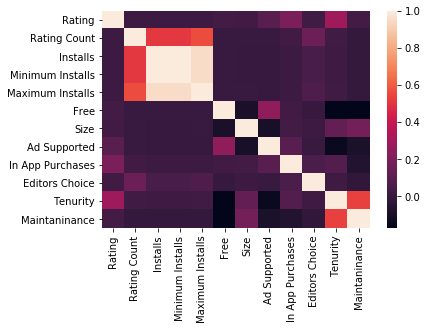


Here 722729 applications in the app store are available for free and also support advertisement.

We can conclude that developers can provide the application for free, however can generate income through free apps by advertisement.

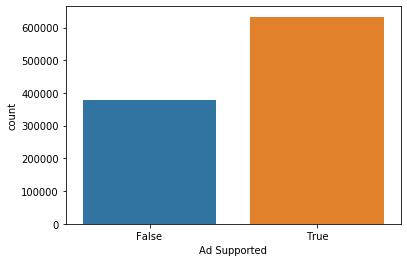
Heat Map: This represents the relationship between each attributes.

Here as we see that that installs and Minimum Installs are highly correlated with each other.



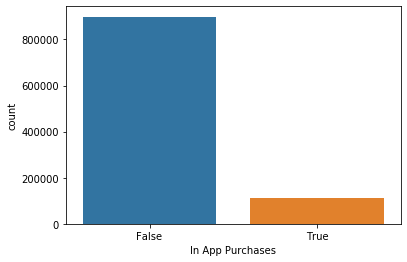
Count Plot:

This represents that more than 600,000 applications in the dataset supports advertisement for generating income. 64.9130% of the applications supports advertisement and 35.0869% of the applications doesn’t support.



Similarly the below graph represents that less than 200,000 application supports in app purchases.

Thus we can conclude that most of the developers use advertisement to generate income.



Pair Plot:

Below chart represents the data distribution after outlier treatment using Interquartile range.

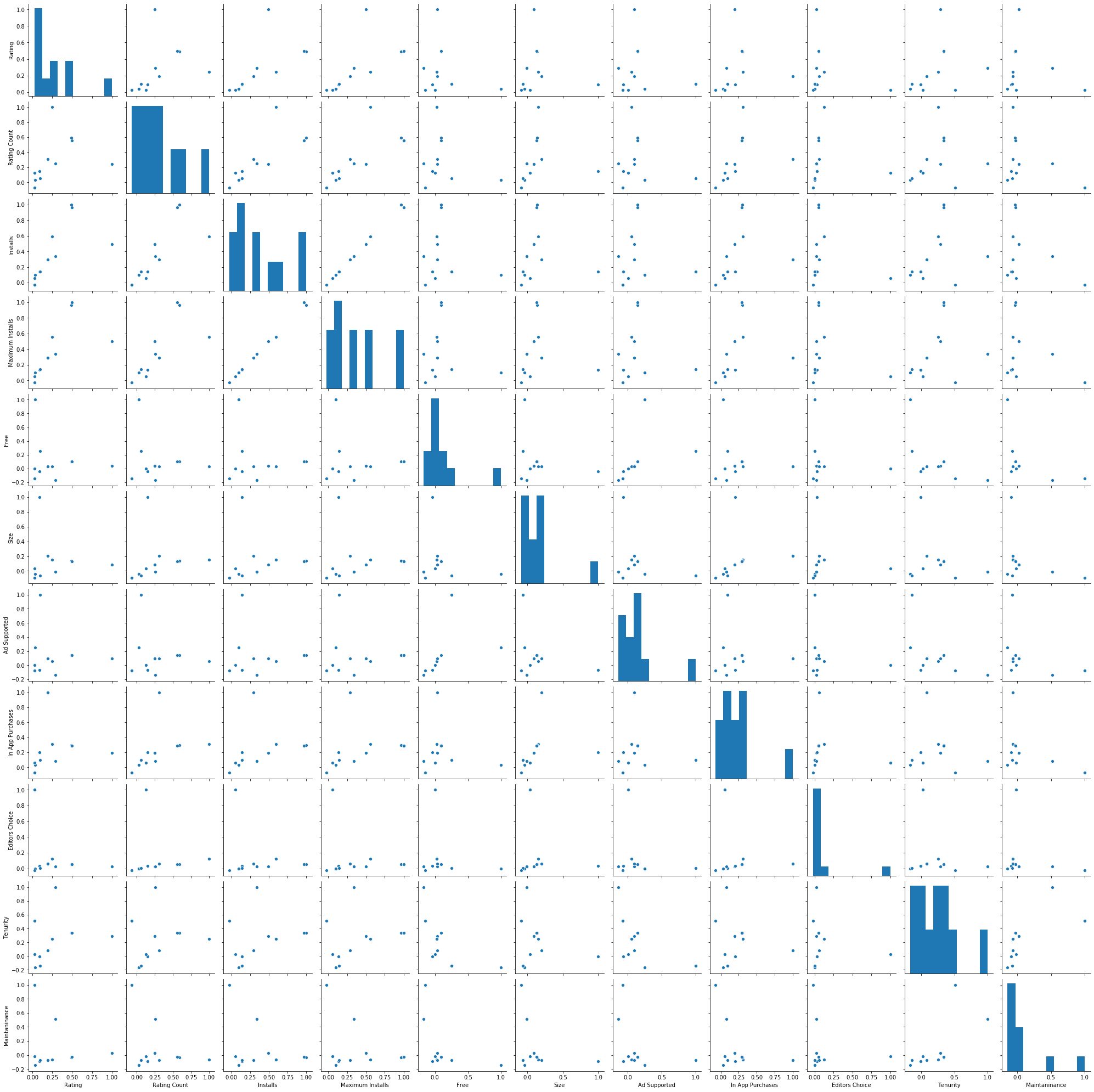


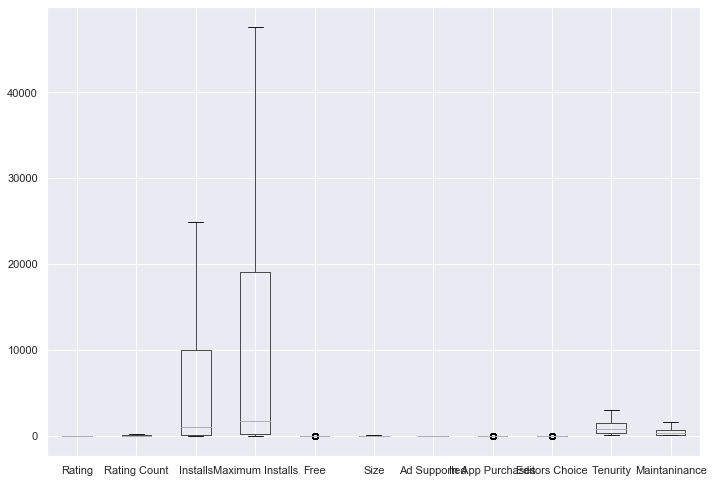
Table that represents the various categories of applications that supports advertisements for generating income.

|  |  |  |
| --- | --- | --- |
| **Categories of application** | **Does Not Support Advertisements** | **Supports Advertisements** |
| Action | 1543 | 11085 |
| Adventure | 1838 | 8286 |
| Arcade | 2732 | 14983 |
| Art & Design | 1566 | 10756 |
| Auto & Vehicles | 3783 | 3089 |
| Beauty | 1739 | 4497 |
| Board | 1125 | 4136 |
| Books & Reference | 15864 | 63105 |
| Business | 36648 | 5562 |
| Card | 868 | 3806 |
| Casino | 474 | 2174 |
| Casual | 2532 | 17977 |
| Comics | 352 | 1785 |
| Communication | 9790 | 8601 |

Box Plot:

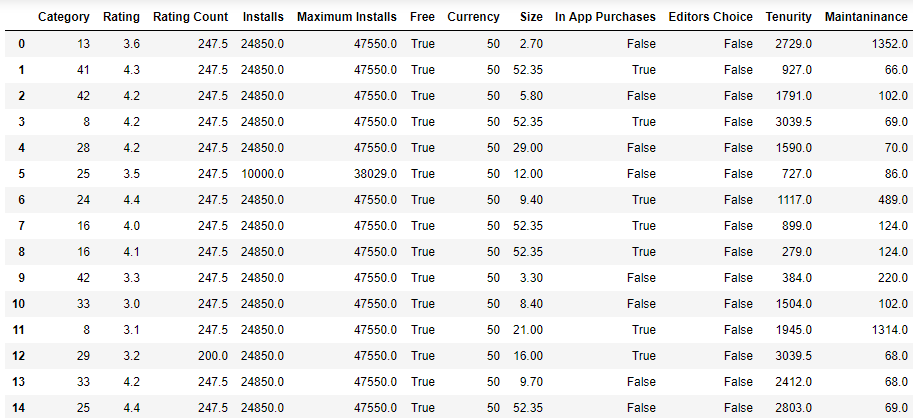
The below chart represent the data of each columns after removing the outliers.

Now the data is ready for further analysis.

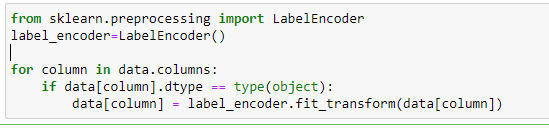
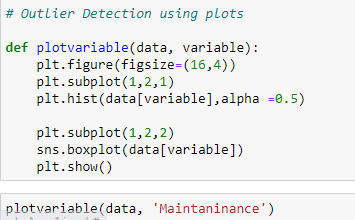


**5.4 Feature Engineering**

* The accuracy of the model is dependent on how clean and transformed the data is, thus feature engineering comes in to picture.
* In this project here we dropped the unnecessary columns that are not useful for machine learning algorithm.
* As per the earlier analysis we found that the columns : Installs and Minimum Installs were highly co-related against each other hence we had to keep Installs and delete Minimum Installs.
* This process is also called as dimensionality reduction or feature selection.
  1. **Short data snapshots**



**5.6 Short code snippets**



**6. Modelling**

**6.1 Selection of model/technique**

* Selection of the right model is most important step in machine learning.
* According to the analysis we found that most of the developers’ generate income through advertisements.
* As the target variable is of Boolean type we will use classification models.
* We will be using hit and trial.
* The models that we will be implementing in this project are :
  1. Logistic Regression.
  2. Random Forest Classifier.

**6.2 Challenges faced.**

* As the dataset is too large the model consumes lot more time to implement.
* Logistic Regression might not be the best algorithm for the dataset. Hence we can try out random forest.
* Evaluation and cross validation has to be done for improving the accuracy of the model.

**6.3 Evaluation and Cross Validation**

* As we are implementing Random Forest algorithm we will be applying some evaluation and cross validation techniques.
* There are few hyper parameter tuning in random forest classification algorithm.
* However, random forest algorithm is based on decision tree.
* In this project we use n\_jobs, **criterion as "entropy" and n\_estimators** as hyper parameter tuning.
* n\_jobs: This is actually the number of processor.

**6.4 Model Interpretation**

* Logistic Regression :
* The accuracy score resulted is 67.57% and f1\_score 78.97.
* [[ 14944, 63101],

[9402, 136181]]

* The above figure shows the confusion matrix, that is the true positive, false positive, false positive and true negative.
* Random Forest Classifier :
* The accuracy score resulted is 83.89% and f1\_score 88.14%.
* [[ 53675, 11647],

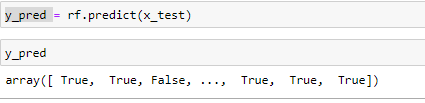
[24370, 133936]]

* The above figure shows the confusion matrix, that is the true positive, false positive, false positive and true negative.

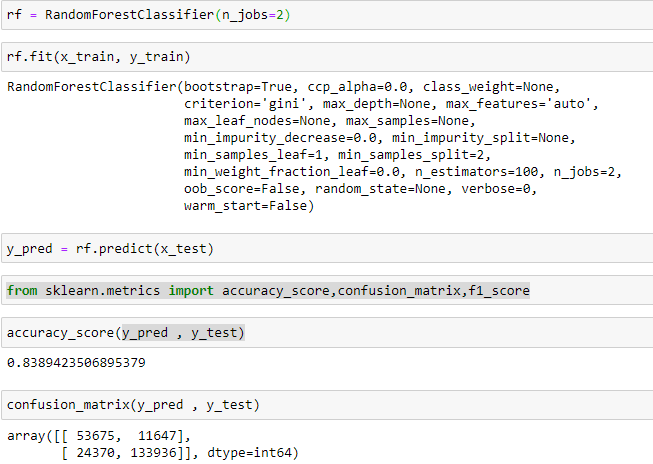
**6.5 What worked/What didn’t work?**

* When we compare both Logistic Regression and Random Forest Classifier for this dataset. We can conclude that Random Forest classifier algorithm is the best fit for the dataset.
* **Random Forest Classifier gave the best accuracy of 83.89% whereas Logistic Regression resulted in** 67.57% accuracy.
* **Hyper parameter tuning with n\_jobs, also resulted in 0.01% increase in the accuracy of the model hence it is not much effective.**
* **Also used criterion as "entropy" and n\_estimators as this slightly decreased the accuracy of the model hence removed.**

**6.6 Short data output/snapshots**



* 1. **Short code snippets**



-

1. **Key Result**

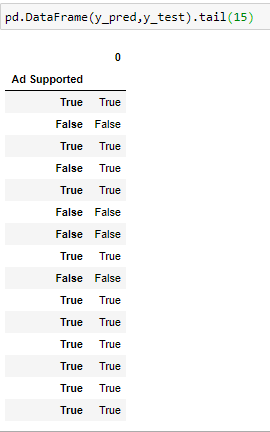
**7.1 Output of intermediate steps**

* Confusion Matrix : [[ 53675, 11647],

[24370, 133936]]

* The above figure shows the confusion matrix, that is the true positive, false positive, false positive and true negative.
* Accuracy Score 83.89%
* F1\_score 88.14%.

**7.2 Outcome/Sample outputs**

Here Ad Supported Column is the actual value that is test set values and columns ‘0’ is the predicted values.

* 1. **Analysis of the results**

We can see that the Random Forest Classifier is providing the accuracy of 83.89%. This tells us that the model predicted results are 83.89% are accurate information.

When we dug deeper into the confusion matrix we see that:

True Positive: 53675 – No. data points whose actual outcomes were positive and the algorithm correctly identified it as positive.

False Positive: 11647 - No. data points whose actual outcomes were negative but the algorithm incorrectly identified it as positive.

False Negative: 24370 - No. data points whose actual outcomes were positive but the algorithm incorrectly identified it as negative.

True Negative:: 133936 – No. data points whose actual outcomes were negative and the algorithm correctly identified it as negative.

**F1 score** calculated from the [precision](https://en.wikipedia.org/wiki/Precision_(information_retrieval)) and [recall](https://en.wikipedia.org/wiki/Recall_(information_retrieval)) of the test resulted 88.14%

**8 Conclusion**

**8.1 Summary of the project outcome**

* Developers’ in the market can use this model to predict the in which all categories of the application will support advertisement to generate better income

.

* The random forest model predicts the target variable 83.89% accurately when compared to the actual output given in the dataset.
* We also concluded that the developers’ preferred option to generate the income is through advertisement.
* As there is millions of records considering neural networks for the further analysis.
  1. **Future work**
* As the dataset is very huge it’s better to choose neural networks.
* In case of new data adding to the same model will work efficiently. However, some changes has to be done.

1. **References** 
   1. <https://towardsdatascience.com/>
   2. <http://www.analyticsvidhya.com/>
   3. <https://www.geeksforgeeks.org/>

* 1. Wikipedia.