**GOOGLE PLAY STORE APP ANALYSIS**

Samyuktha Nayakanti Lakshmi Sowmya Chennareddy Dedeepya Chintaparthi Sarala Nakka

Student id:801058723 Student id: 801074902 Student id : 801056005 Student id:801074190

[snayaka1@uncc.edu](mailto:snayaka1@uncc.edu) [lchennar@uncc.edu](mailto:lchennar@uncc.edu)  [dchintap@uncc.edu](mailto:dchintap@uncc.edu) [snakka@uncc.edu](mailto:snakka@uncc.edu)

Pragna Reddy Anthireddy Rakesh Gunti Naga Sandhyadevi Yalla Kavya Potluri

Student id:801079345 Student id:801100506 Student id:801097837 Student id:801073398

[panthire@uncc.edu](mailto:panthire@uncc.edu) [rgunti1@uncc.edu](mailto:rgunti1@uncc.edu) [nyalla@uncc.edu](mailto:nyalla@uncc.edu) [kpotluri@uncc.edu](mailto:kpotluri@uncc.edu)

Naga Vineel Golla Likitha Chowdary Punati Gowtham Sabbineni

Student id:801114632 Student id: 801083443 Student id: 801074060

[ngolla@uncc.edu](mailto:ngolla@uncc.edu) [lpunati@uncc.edu](mailto:lpunati@uncc.edu) [gsabbine@uncc.edu](mailto:gsabbine@uncc.edu)

**ABSTRACT:**

The advent of broad range of mobile and its applications is so much that our entire day revolves around a mobile, which is no longer just a medium of communication, but beyond that is an established and proved notion to lay grounds for the fact that these mobile applications are the wind under the booming mobile market. Upon research, and analysis, we see that the android applications occupy the lion's share summing up to 80% of the market.

In this paper, we tried mining the relations and the behavior of each feature that impacts the dataset. We have visualized the dataset components and checked the validity of these visualizations using various data mining algorithms. These various data mining algorithms are implemented using XLMiner, on the hypothesis we built for mining the relationship and drawing inferences.

**Keywords:**

Data Visualization, XLMiner, Classification, Prediction, Clustering

**1.INTRODUCTION**

Google play store was first launched in October 2008 and it is referred as “Android market” by Google. This Android market houses many apps related to various categories and caters to the needs of every age group.

Play store comes as a pre-installed application on all the devices with android OS. Google play store APK runs on every Android device whether it is a smartphone, tablet or PC. We can see the top apps and games. Play store provides parent control feature and you also have an option to choose either Wi-Fi or mobile data to auto-update the installed apps. This app is full of features where you must use the same google id in your new device and you can get a list of all installed apps on your previous device. Apps are neatly categorized so that you don’t have to mess with different genres. You can also get the reviews and ratings (out of 5) of any app before installing it which lets you take decision about the app.

Users can also upload their apps on the play store apart from utilizing the various apps that are already available. The platform can be utilized by creating a google developer account and making an initial payment of $25 in order to create developer console which shows the progress of your app and revenue generated through your app.

Though there are around 2 billion of android device in the world there are not many datasets on android devices as Google Play Store uses sophisticated techniques using jQuery making web scraping challenging. However, this data set has huge potential to drive the android app-making business to a great extent. Developers and business can utilize the insights drawn from the dataset to amplify the market for the apps developed.

The second section contains the business use case on which the entire analysis is based upon. Section 3 encloses the Hypothesis questions based on our use case. Section 4 highlights the dataset and its features. This sections also covers the various pre-processing techniques involved in cleaning the data. Fifth section gives a clear view of the analysis models applied on hypothesis. We have applied the below mentioned models on our hypothesis:

I. K-Means Clustering

II. KNN Algorithm (For Classification and Prediction)

III. Random Trees

IV. Regression Tree

V. Multiple Linear Regression.

VI. Classification Tree.

We have come to conclusions based on the metrics, Lift charts, Metrics etc. These models paved ways to solutions that our hypothesis is aimed at. We identified the features that has highest impact while installing the apps, we found out the best apps All these are discussed in detail in section 6. We used Tableau to perform different Visualizations and they are attached in the Appendix section.

**2. BUSINESS USE-CASE**

The data we’ve chosen has been collected from the Kaggle. In this project, we will be analyzing the dataset and build a model to predict the features of the app that influence the user to install the app.

As many apps in Android are free of charge, keeping in mind the features that are influencing the users to install and use the app helps both developers and businesses to reap profits. Secondly, we would like to classify the app into least popular and most popular app which gives users an idea about the app and its popularity among the community.

**3. HYPOTHESIS QUESTIONS**

1. Identify the most popular and least popular apps by considering the most important app features.
2. To identify the top most factors which influence users to install the apps.
3. To predict the rating of the app based on the app features.
4. To classify the content rating of the app based on other features.

**4. DATASET**

The dataset used for this project is Google play store apps dataset, which is taken from Kaggle​​. This dataset folder contains two sheets, one is the google\_play\_store apps and the other is google\_playstore\_user\_reviews.

**google\_play\_store\_apps:**

This data set contains details about the app like app name, category, rating of the app, number of reviews etc. This dataset has 10,847 rows and 13 columns. The columns are as follows:

**App**: This column contains the Name of the App

**Category**: This column contains the category to which the app belongs to. There are 34 app categories in this column.

**Rating**: This column depicts the rating of the app.

**Reviews**: This column holds the number of reviews received for the app.

**Size:** Size of the particular app is specified in this column.

**Installs:** Thiscolumn tells us about the count of installations per an app.

**Type:** This column specifies whether the app is a free app or a paid app.

**Price:** This column has information about the price of the apps. Free app has the price as zero and the paid app has its respective value.

**Content Rating:** This column throws light on the user group for each app. It has 6 user groups like Everyone, Adults Only 18+, Everyone 10 etc..

**Genres:** This column is like extended version of the Category column. Example: Art and Design category has a genre of Art and Design, Creativity etc.

**Last Updated:** This column has details about when the app is last updated.

**Current Ver:** This column specifies the current version of the particular app.

**Android Ver:** This column specifies the Android version of the particular app.

**Google play store user reviews:**

This data set contains the user reviews for the apps, sentiment of the reviews, sentiment polarity, sentiment subjectivity.

Our hypothesis is related only to the google play store apps dataset and we did not consider the user reviews sheet as the analysis of the sheet is more related to text mining, which would be our future focus.

**4.1 Pre-Processing of Data:**

To apply models to the dataset, the data must be in a format that is acceptable by the model. Our dataset has some missing values, data columns with different data formats, some “Nan” values, some special characters etc.

The following are the pre-processing steps performed on our data set.

**a. Data-type conversions:**

* Converted ‘Price’ column from string datatype to float
* Converted ‘Installs’ column from string datatype to long int
* Converted ‘Last Updated’ column datatype to ‘short date’.
* Converted ‘Current Version’ column’s datatype to string format.
* Converted ‘Size’ column to numeric by replacing ‘K’ with ‘e+3’ and M with ‘e+6’.

**b. Handling Missing Values and deleting unnecessary fields:**

* We have only one record with type as ‘Nan’, It has no reviews, no installs and no size specified. Hence this record is deleted.
* Size column has blank values; hence we handled those values by using the mean of the size column.
* ‘Nan’ values in ‘Rating’ column are replaced with ‘0’. We assume those apps with rating 0 are not yet rated.
* Two rows with Type column having ‘Nan’ and ‘blank’ values are deleted, as they do not have much impact on the dataset.

**c. Removing Duplicates:**

* Duplicate data in the dataset is deleted.

**d. Handling Categorical Data:**

* There are more than 30 categories under the ‘Category’ column, XLMiner supports only until 30 categories. Hence, we have clubbed some of the similar categories. After the following clubbing we have 26 categories under the Category column.
* Clubbed the Comics category into Books and References Category
* Clubbed the Game category into Entertainment category
* Clubbed the Parenting category into Family Category
* Clubbed the Health and Fitness category into Medical Category
* Clubbed the Personalization, Video Players category into Tools category.
* Clubbed the communicated category into Social category.
* Created dummies for the ‘Type’, ’Content rating’ and the ‘Category’ Column.

After the preprocessing is done, the final data set has 9775 rows and 13 columns.

**5. DEMONSTRATING PROCEDURES**

Various models can be applied to the pre-processed data to dataset for predicting and breaking down the information. All the three hypothesis questions are addressed using these demonstrating procedures by choosing the best relevant model for a specific theory. Below mentioned are the general modeling techniques:

**5.1 Regression:**

Regression is a statistical approach for modelling relationship between a target variable with a given set of predictor variables. Regression helps in identifying the behavior of a variable when other variable(s) are changed in the process. Most commonly used model for prediction is Regression analysis.

**5.2 Classification:**

Classification is a technique where it is identified to which category or class a new observation belongs to depending on the dataset that is already trained. It is considered as an instance of supervised learning. i.e. learning where training dataset is well labelled.

**5.3 Clustering:**

Clustering, one of the techniques used for unsupervised learning is the process of dividing the given dataset for values into required number of groups such that similar data is grouped into one group and dissimilar to the data in other groups. It is basically a collection of objects based on similarity and dissimilarity between the data.

**5.4 Prediction:**

Prediction is similar to classification techniques. The only difference is that in prediction we try to predict the value of quantitative variable and in classification , we try to predict the class and this may vary with the project.

**6. BUILDING MODELS**

In this section, various models are built to analyze the hypothesis questions.

**K-Means clustering:**

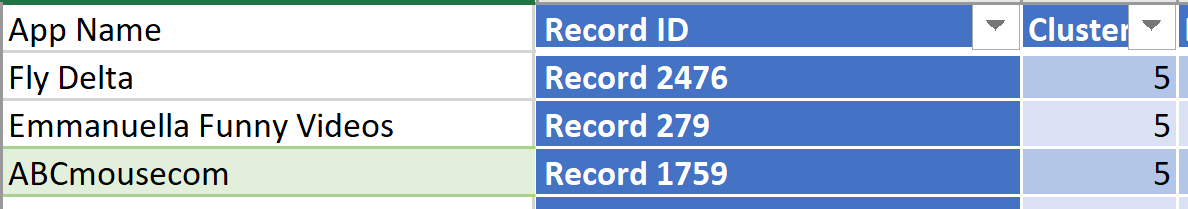
In K-Means clustering, entire set of observations will be partitioned into k clusters, and each observation will be assigned to a cluster with nearest distance.

**HYPOTHESIS 1:**

To find the most and least popular apps using the Clustering Analysis.

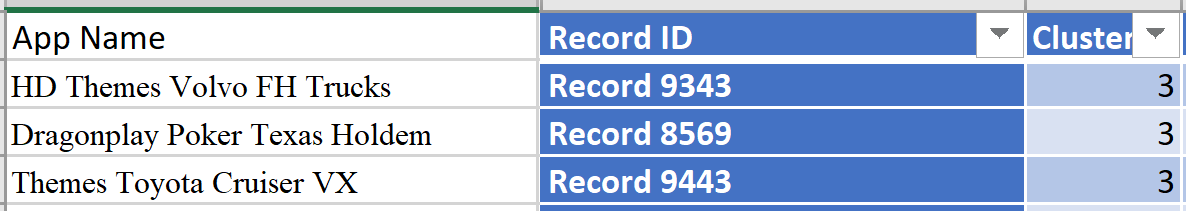
To analyze this, we have used the K-Means Clustering. We have selected the input variables as Reviews, Installs, Rating, Category with dummy variables, Content rating with dummy variables, type with dummy variables, price and treated\_size. We have selected the number of clusters as 10, and the number of iterations as 50. Results are as follows:

Most Popular Apps:



Based on the distance, we see that the ‘Fly Delta’, ‘Emmanuella Funny Videos’ and ‘ABCmousecom’ are the most popular apps.

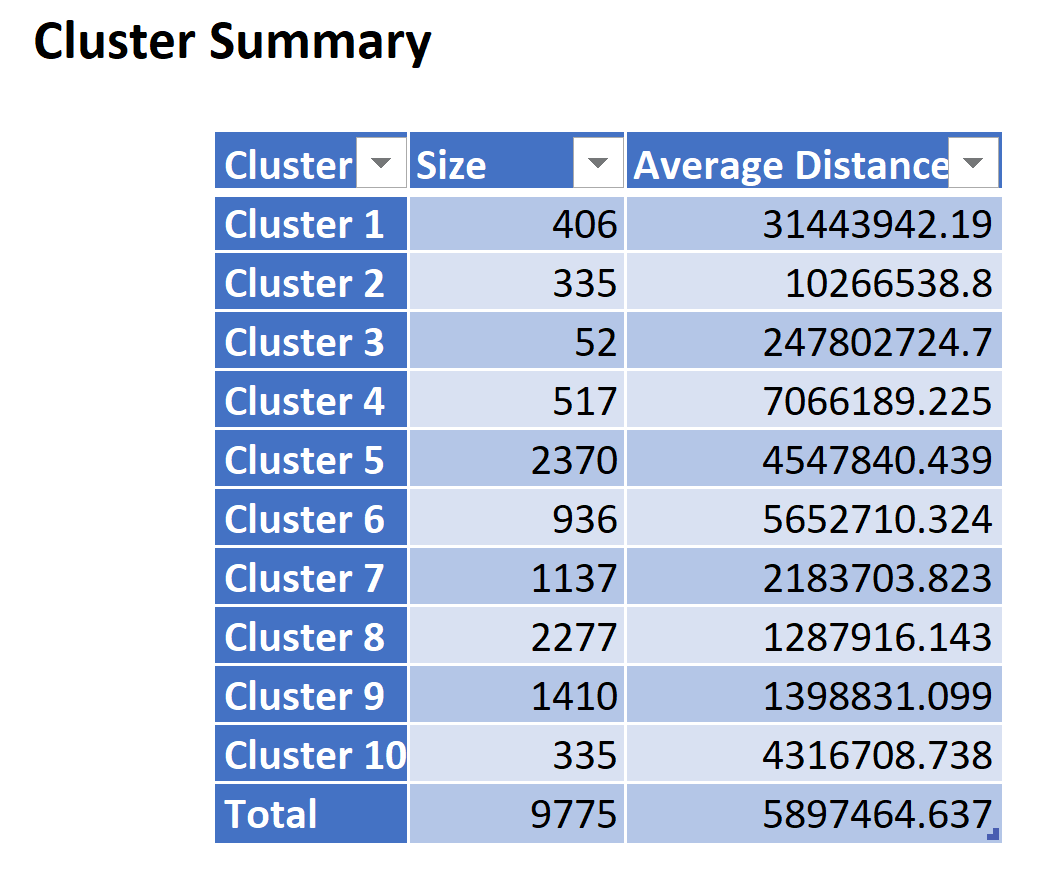
Least Popular Apps:



Based on the distance, we see that the

|  |
| --- |
| ‘HD Themes Volvo FH Trucks’, ‘Dragonplay Poker Texas Holdem’, ’HD Themes Toyota Cruiser VX’ are the least popular apps. |

**Cluster summary is as follows:**

****

Each record is assigned to the cluster to which has the minimum distance. Based on this, the most popular apps are assigned to cluster 5 and the least popular apps are assigned to cluster 3.

**Limitations of K-Means Clustering:**

* Prediction of K Value is not so easy.
* There may be different final clusters based on initial partitions
* Results are highly impacted by initial seeds
* The results tend to change with normalizing and standardizing the data.

**HYPOTHESIS 2:**

To identify the top most factors which influence the users to install the apps.

We have used Regression Tree, Linear Regression and KNN models to find the top predictors which influence the users to install the apps.

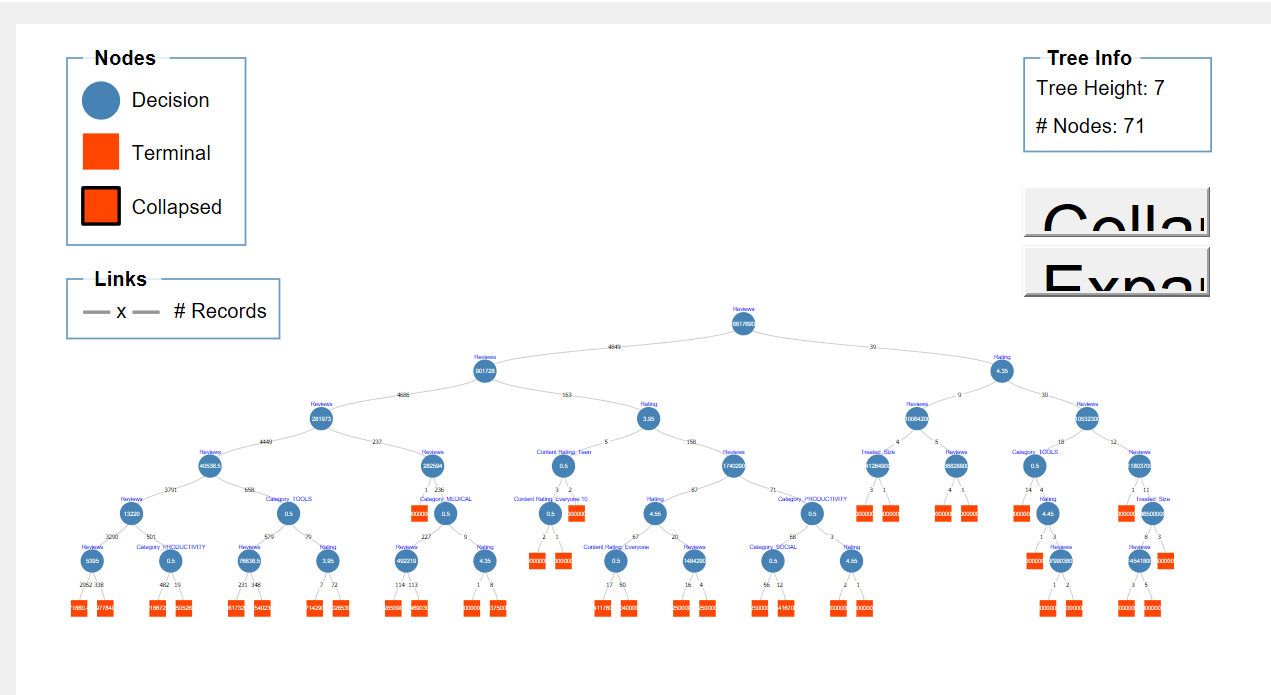
After the data is pre-processed and the dummies are created for the categorical variables (type, category and content rating), data is partitioned into training set (50%), validation set (30%) and test set (20%) and then the above three models are run on the partitioned data.

**Model 1: Regression Tree:**

Regression tree model is used when the output or response variable is a continuous variable. In this hypothesis, the number of installs is the target variable and it is a continuous variable. Hence, Regression tree model is used to predict the top factors that influence the users to install the apps. Also, in Regression tree, we can visualize each step which could help users in making decisions.

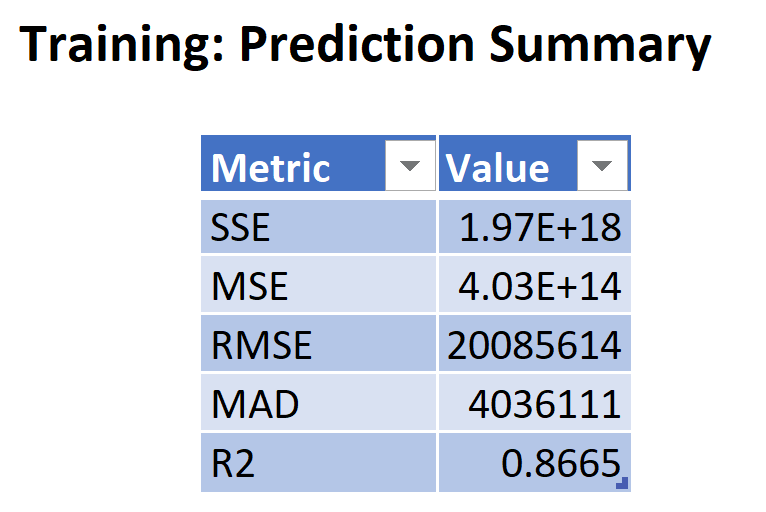
Regression tree model is run on the partitioned data set by selecting the input variables as Reviews, Rating, Category with dummy variables, Content rating with dummy variables, type with dummy variables, price, treated\_size and the output variable as Installs.

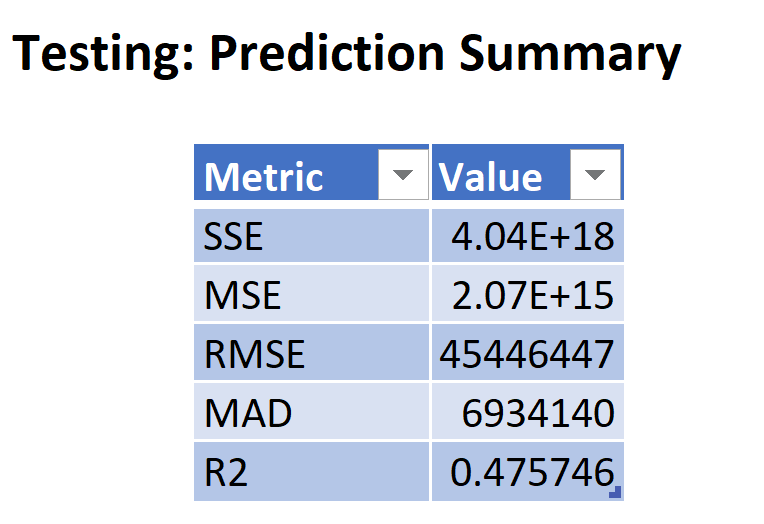
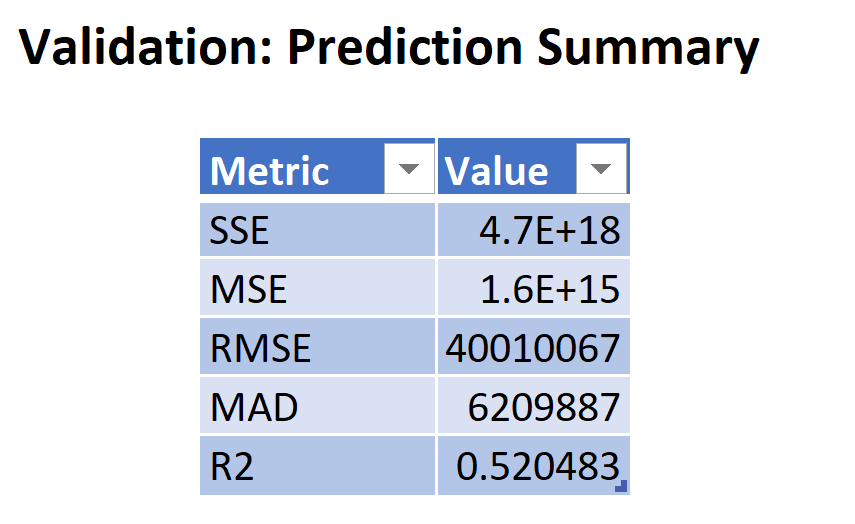
Tree for scoring is selected as Best Pruned and the tree to display is selected as Fully grown while running the Regression tree. The best pruned tree generated is as follows:



From the above best-pruned tree, the top factors that influence the users to install the apps are Reviews, rating and size.

**Metrics for Regression Tree:**





**Limitations of Regression Tree:**

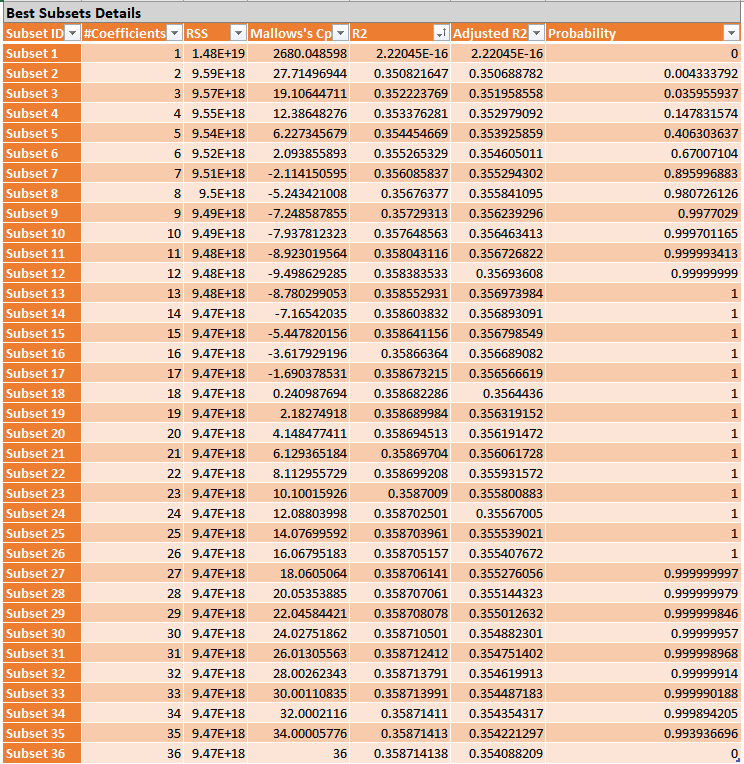
* Regression trees require many parameters(splits) to capture simple models such as linear relationships.
* A small change in the data can have a greater impact on the splits and thus leading to instability.

**Model 2 - Multiple Linear Regression:**

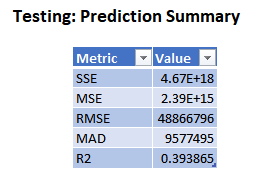
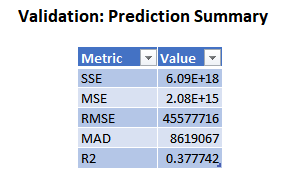
Multiple Linear Regression is used when we have two or more independent variables which determines one output dependent variable.

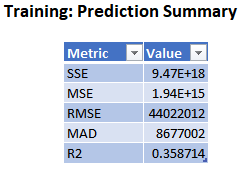
Multiple Linear regression model is run on the partitioned data set by selecting the input variables as Reviews, Rating, Category with dummy variables, Content rating with dummy variables, type with dummy variables, price, treated\_size and the output variable as Installs. We have selected the Best subsets option in the feature selection while running the model.

Below are the best subsets generated:



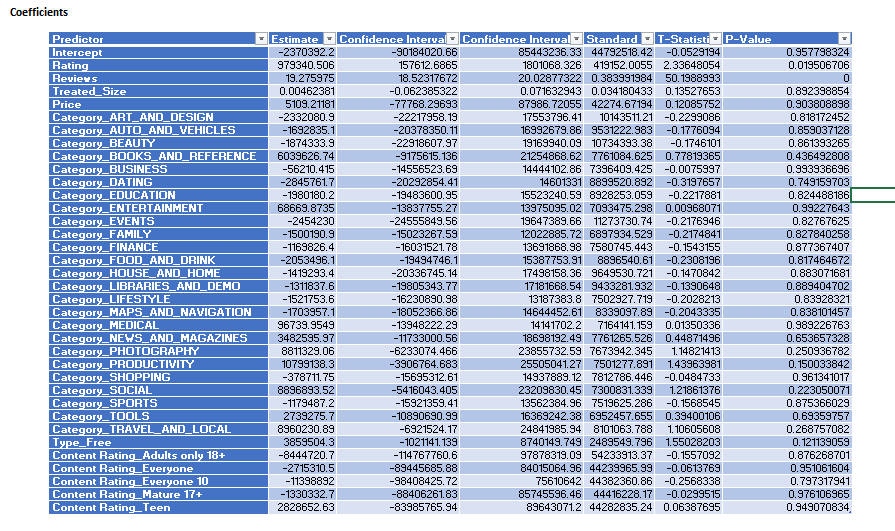
From the above, subset 36 is chosen as the best subset as it has highest R2 value and its c\_ p value is near to p+1, which indicates that it is a better fit. Now, the Multiple Linear regression model is run again on the subset 36 and the results are as follows:





The training, validation and testing metrics are attached for the Multiple Linear Regression.R2 value for the training data is 0.358, for validation data is 0.377 and for testing data it is 0.393.

**MLR\_Output:**

****

From the coefficients table, based on the P-value, the top predictors that influence users to install the app are Reviews, rating and type(free/paid) of the app.

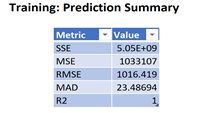
**Limitations of Multiple Linear Regression:**

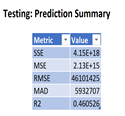
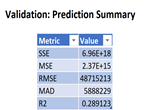
1. Multiple linear regression is applicable only when the output variable is linearly dependent with the independent input variables. Not having linear relationship between input and output variables and outliers in the data will impact the models to a great extent.
2. The model considers noise for modelling rather than the relationship between the variables when number of parameters exceed the number of samples in the data.

**Model 3 - KNN - Prediction:**

KNN Algorithm looks for the k nearest points, determines to which class maximum of the point belongs to and then classifies the new point to the same class. Here K is a number which specifies how many nearest points should we look before classifying the new point.

KNN model is run on the partitioned data set with the same input and output variables us**e**d in the above models. K is chosen as 1. Results are as follows:





**Limitations: KNN Model**

a. Need to determine the value of parameter K (nearest neighbor).

b. Distance-based learning is not clear, and attribute-based selection is not clear.

c. As we need to calculate the distance for each observation, the cost of computation would be very high.

**Hypothesis 2 - Best Model:**

We have run regression tree, Multiple linear regression and KNN for this hypothesis. Among the three models, we see that error rate is lower for Regression tree and also the R2 value for the validation and training data is higher for regression tree. Hence, we chose Regression Tree as the best model for this hypothesis and the predictors that influence the users to install the apps according to the best model are reviews, rating and size of the app.

**HYPOTHESIS 3:**

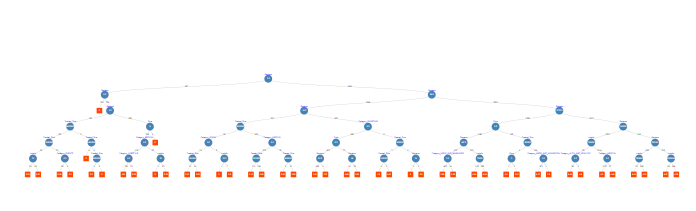
To predict the rating of the app based on other app features.

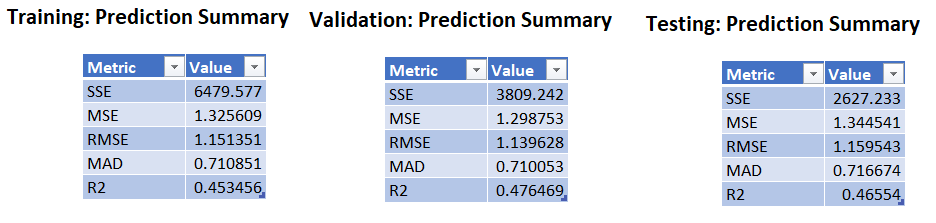
After the data is pre-processed ,dummies are created for the categorical variables (category, type and content\_rating).Then the data is partitioned into training(50%),validation(30%) and testing(20%) and the three models Regression tree, KNN and Multiple Linear Regression are run by using the input variables as reviews, installs, treated\_size, price, type, category, content\_rating and the output variable as Rating.

**Model 1 - Regression Tree**

After the Regression tree is run on the partitioned data set, the best pruned tree is as follows:

From the below best pruned tree, we can say that the top most features for predicting the app rating are reviews, installs and size.

**Metrics for Regression Tree:**

****

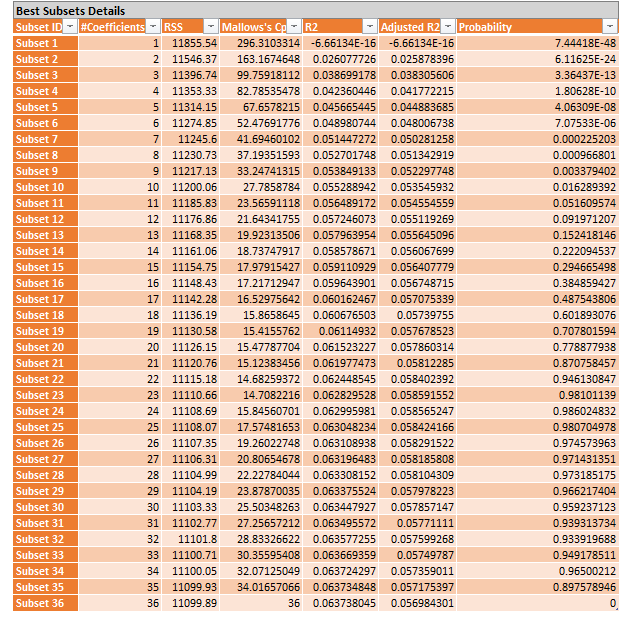
**Model 2 - Multiple Linear Regression:**

The model for MLR, given n observations, is

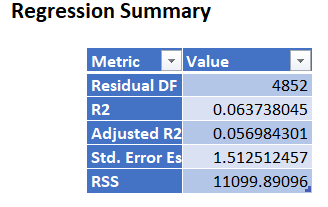
y= β\_0+ β\_1 x\_1+β\_2 x\_2+ …+β\_p x\_1+ε, where i = 1, 2,.. n

To predict the rating of the app, Multiple Linear regression on the partitioned data set, and the best subsets option is selected in the feature selection. The attached picture shows the best subsets generated after running the Multiple Linear Regression.

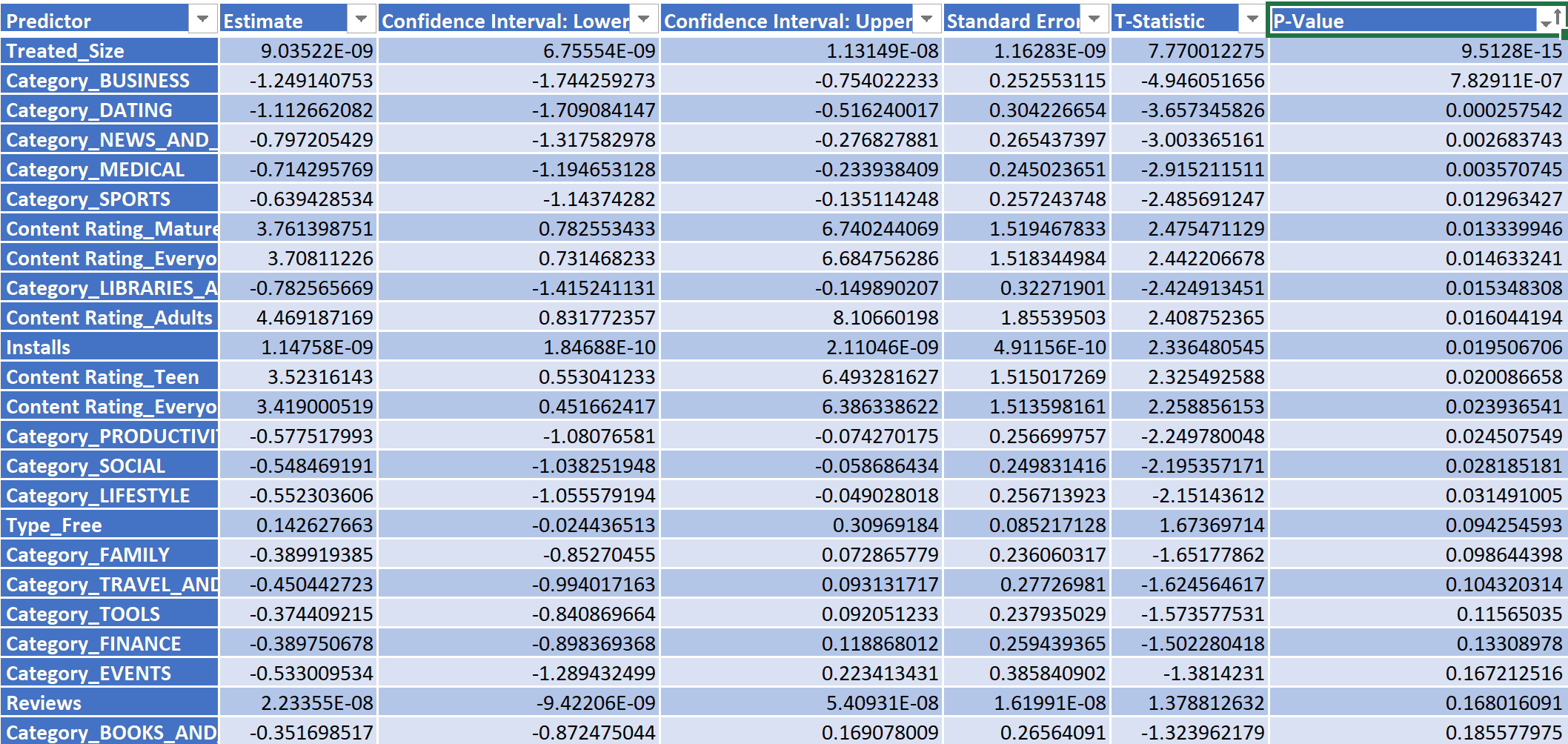
From the subsets generated, we chose subset 36 as the best subset as it has the highest R2 value and also it has the c\_p value is near p+1 indicating that it is a better fit.



Then, the Multiple Linear Regression is run on the subset 36 The coefficients and the final Metrics are as follows:



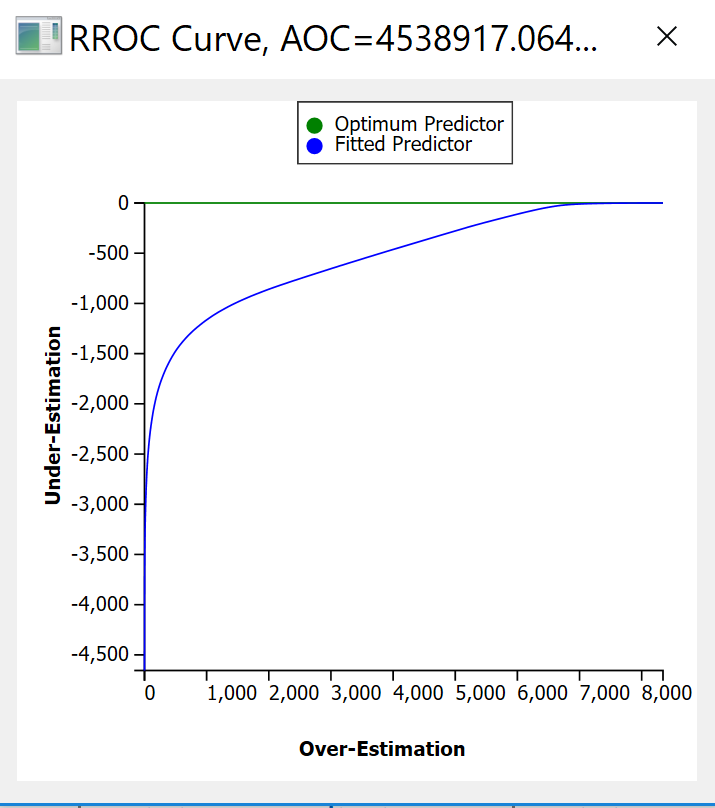
**Coefficients:**



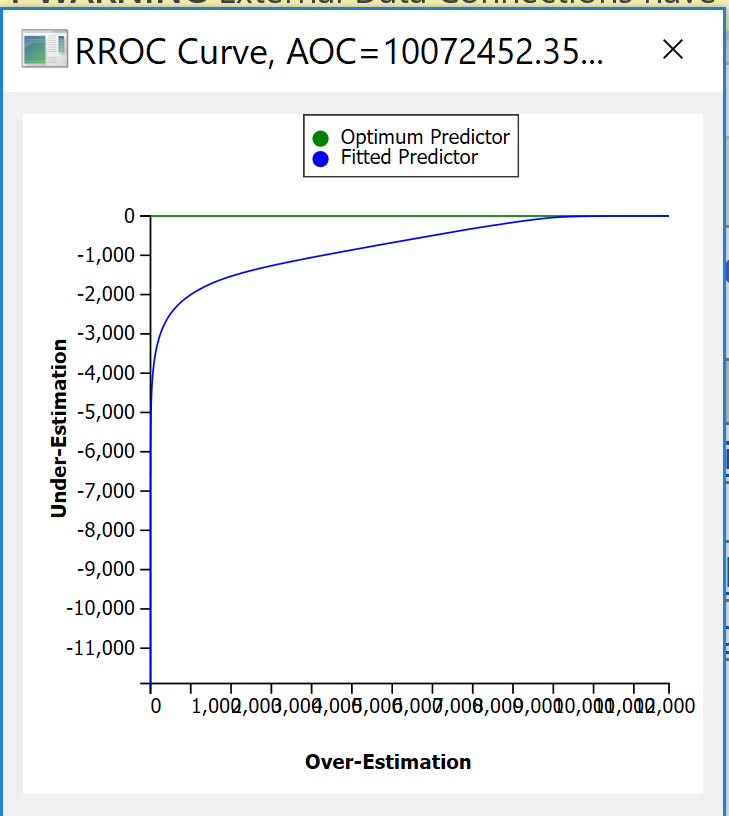
**ROC Curves:**

We have also used ROC curves to evaluate our models and the following are the ROC curves for testing and validation data:

**ROC Curve - Testing**



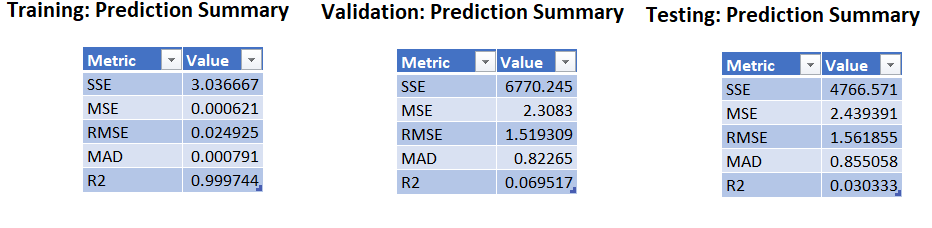
**ROC Curve - Validation:**

****

**Model 3 -KNN**

To predict the app rating based on other app features, KNN model is run on the partitioned data set by choosing K as 1 and the results are attached.

We can see that R2 value for the training data is 0.997 and this indicated that training data might be subjected to over-fitting here as the value of R2 is very close to 1. Also, we can say that the R2 value for validation data is 0.06 and for testing data is 0.03.



**Best-Model - Hypothesis 3:**

Three predictions models i.e. Regression tree, Multiple Linear Regression and KNN are run on the partitioned data set to predict the app rating using other input features. Out of these three models Regression tree is the best model as it has least error rate and higher R2 value.

**HYPOTHESIS 4:**

To classify content-rating of the app based on other app features, that is if the app can be used by everyone or if there are any age restrictions for the app to be used.

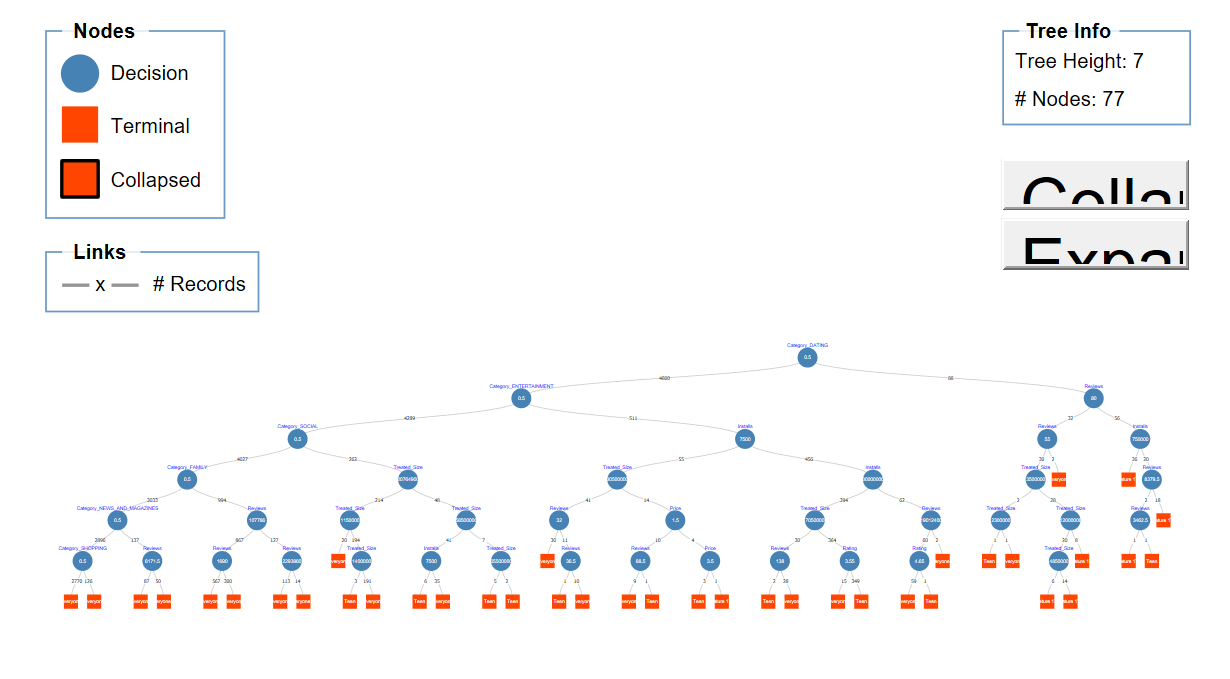
After the data is pre-processed and the dummies are created for the categorical variables (type and category), data is partitioned into training set (50%), validation set (30%) and test set (20%) and then the above three models are run on the partitioned data, with the output variable as Content rating.

**Model 1 - Classification Tree:**

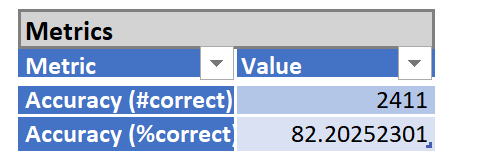
Classification trees are used when the output variable is categorical. As the ‘content rating’ for an app is a categorical variable, we use the classification tree in order to classify the content rating of an app based on the other app features. By looking at the classification tree, we can analyze which features of the app plays an important role in decision making.

To classify the content-rating of the app, classification tree model is run on the partitioned data set (dummies created only for type and Category variables) by giving the input variables as Reviews, Installs, Rating, Category with dummy variables, type with dummy variables, price, treated\_size and the output variable as content rating.

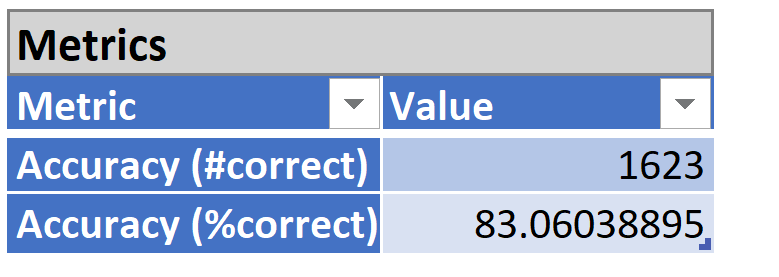
The Best pruned tree generated is as follows:



**Validation set accuracy:**

****

**Testing set accuracy:**

****

**Limitations of Classification Tree:**

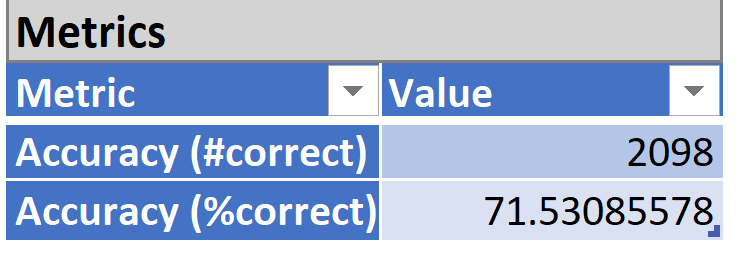
* High Probability of overfitting
* When there are many class variables the calculations are very complex
* This sometimes gives a biased response

**Model 2 - KNN**

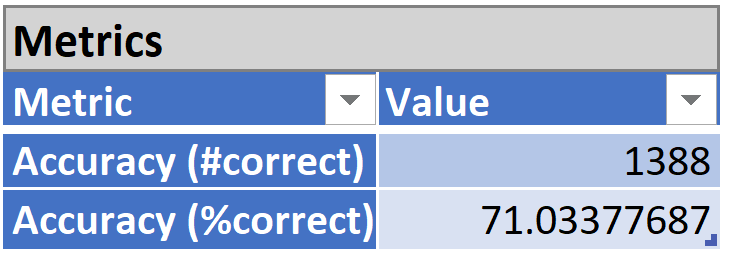
KNN model is run on the partitioned data setby giving the input variables as Reviews, Installs, Rating, Category with dummy variables, type with dummy variables, price, treated\_size and the output variable as content rating.

The results are as follows:

**Validation set accuracy:**

****

**Testing set accuracy:**

****

**Limitations of KNN for Classification:**

* KNN will not be able to handle if there is a minimal distance from two different boundaries.
* It is not advisable to use KNN when we have Categorical data.

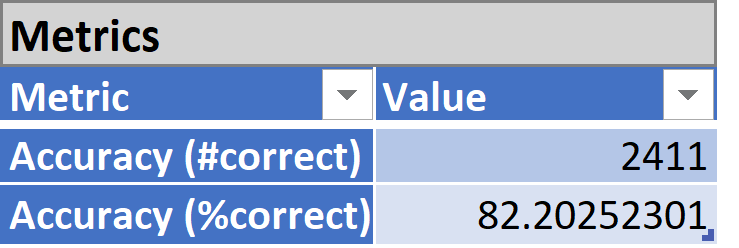
**Model 3 -Ensemble - Random Trees:**

In Random Forest not all the features are taken into consideration, only a random subset of features are taken and while splitting the node ,the model does not check for the most important feature ,but it checks for the best feature among the random subset of features and thus resulting in a better model.

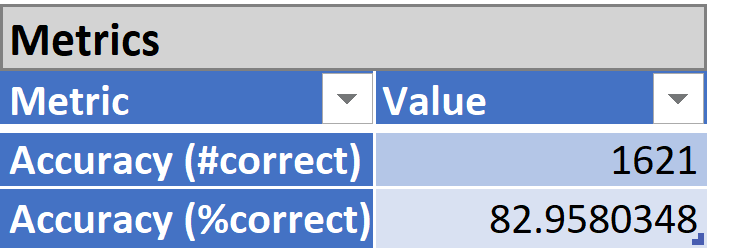
Random trees model from Ensemble is applied on the partitioned data set by giving the input features as reviews, rating, installs, treated\_size, category with dummies, type with dummies, price and the output variable as content rating.

The following are the results

**Validation set accuracy:**



**Test set accuracy:**

****

**Limitations of Random Trees:**

* The large number of trees generated in the random forest algorithm makes the algorithm slow down and ineffective for prediction evaluation
* Predictions slow down once the algorithm is trained thought the training happens quickly. To get an accurate prediction, large number of trees are required and this in turn slows down the performance.

**Best Model -Hypothesis 4:**

To classify the content rating of the app, three models i.e. Classification Tree, KNN and Random Trees are run on the data. Out of these three models Classification Tree and the Ensemble Random tree models have almost same accuracy and the error rates with Classification tree having .1% higher than the ensemble random tree. Also, from Classification tree we can see that the App Category plays an important role in deciding the Content rating of the app.

**7. STRATEGIC RECOMMENDATION**

Analysis of a dataset of this sort requires proper planning and clear understanding of hypothesis. We need to check the data of apps with respect to size, category and various other features. There are both categorical and numerical variables in the data set , but for the category variable named ‘category’ there are more than 30 classes in it and having many number of classes would lead to lower accuracy of the models.Results would be more accurate when the categorical variable has less number of classes/categories and when these are not closely related to each other.

If the dataset would have had the details on whether an app is pre-installed or not, would have the business use case and hypothesis related to it.

Another key strategic recommendation lies in preprocessing the data. For example, removing the special characters and replacing it with space reduces the noise. This could help in improving the results of the model.

**8. CONCLUSION**

The hypothesis is analyzed and studied based on different models built in XLMiner and on visualizations built in Tableau. Finally, the following conclusions are drawn about the hypothesis:

1. Taking into consideration the important app features, the most popular apps are ‘**Fly Delta**’, ‘**Emmanuella Funny Videos**’ and ‘**ABCmousecom**’ and the least popular apps are ‘**HD Themes Volvo FH Trucks**’, ‘**Dragonplay Poker Texas Holdem**’, ’**HD Themes Toyota Cruiser VX’**.
2. Reviews, Rating and Type of the app(free/paid) are the top predictors that influence the users to install the app.
3. While trying to predict the app rating, we see that Multiple Linear Regression is the best model and the app rating can be predicted based on the input features we use.
4. While classifying the content rating of the apps Classification tree is the best model when compared to other models and from the tree, we can infer that the DATING category of apps have the content rating as Mature 17+.

**9. REFERENCES**

[1] <https://www.kaggle.com/lava18/google-play-store-apps>

[2]<https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-regression-analysis-results-p-values-and-coefficients>

[3] Data Mining for business Analysis- Galit Shmueli , Peter C. Bruce, Mia L Stephens, Nitin R Patel

[4] <https://www.solver.com/cluster-analysis>

[5]<https://www.solver.com/k-nearest-neighbors-k-nn-classification-intro>

[6]<https://www.solver.com/using-multiple-linear-regression>

**[7]**[https://www.theanalysisfactor.com/what-is-an-roc-curve](https://www.theanalysisfactor.com/what-is-an-roc-curve/)

[8]<https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd>

[9]<https://www.statista.com/statistics/266210/number-of-available-applications-in-the-google-play-store/>

[10]<https://chrome.google.com/webstore/detail/play-store/mpjpdgddkidcceneniokmmncldibnafj?hl=en>

[11]<https://nycdatascience.com/blog/student-works/web-scraping/analysis-of-apps-in-the-google-play-store/>

[12]<https://www.machinelearningapplications.com/classification/>

[13]<https://magoosh.com/data-science/what-is-a-regression-tree/>

[14]<https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/>

[15]<https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd>

[16]<https://www.saedsayad.com/k_nearest_neighbors.htm>

[17]<https://www.statista.com/statistics/266210/number-of-available-applications-in-the-google-play-store/>

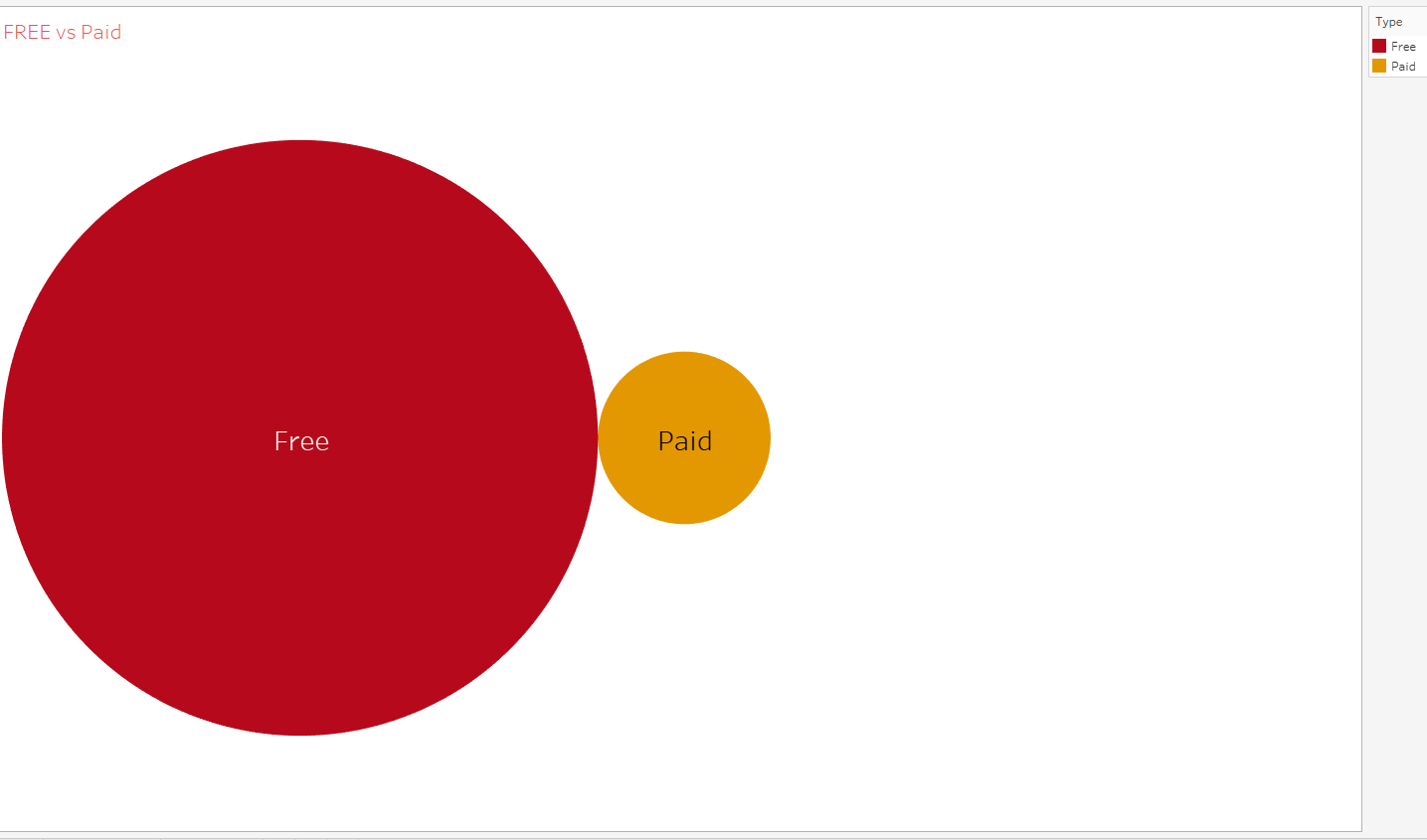
[18]<https://www.android.gs/official-google-play-store-account-is-live-on-twitter>

[19]<https://www.powershow.com/view/ad90a-ODMwZ/Multiple_linear_regression_powerpoint_ppt_presentation>

**10. APPENDIX**

**10.1 Visualizations:**

**i. Free Vs Paid apps:**

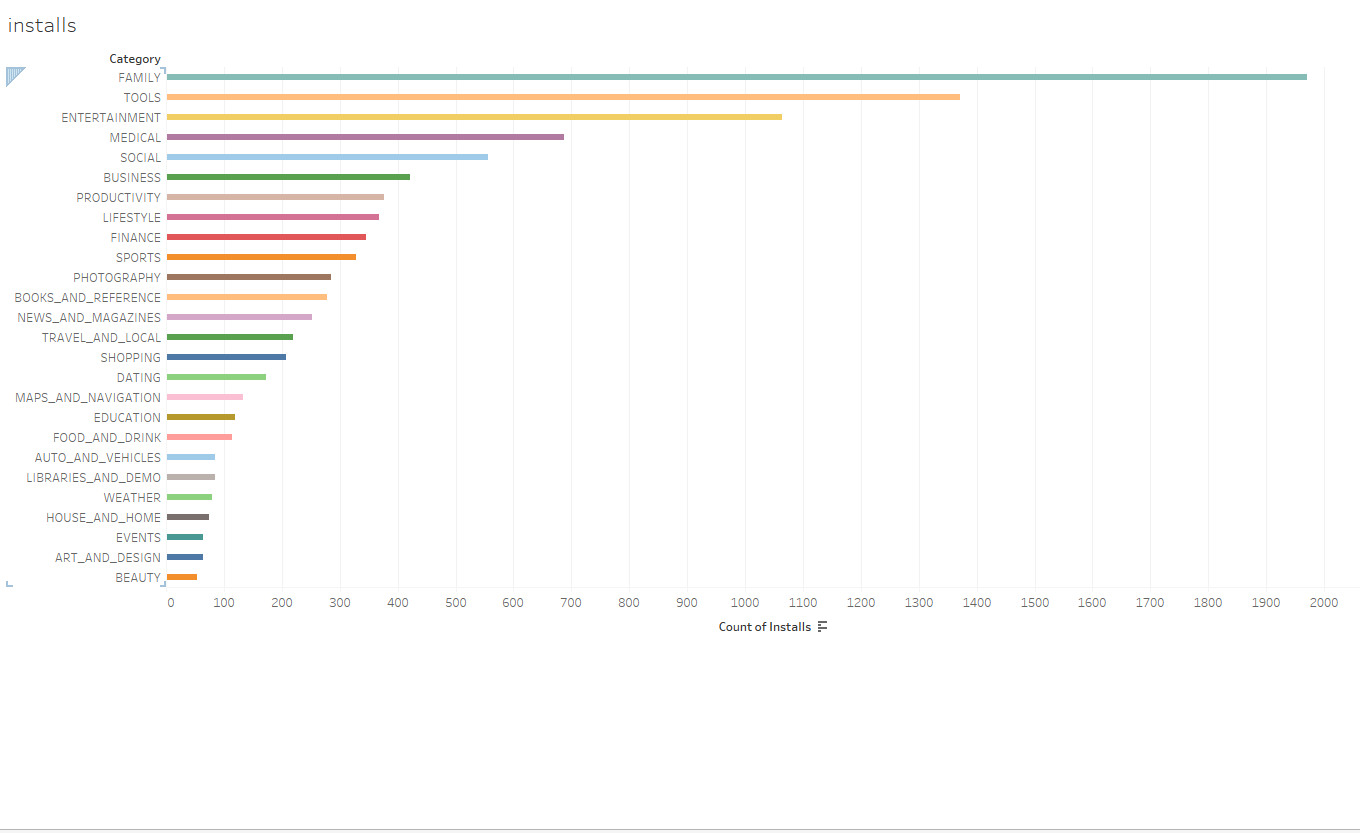
****

The above visualization gives insights about the number of free vs paid apps in our dataset.

From the above visualization, we can conclude that the number of free apps in the data set are more when compared to the paid apps.

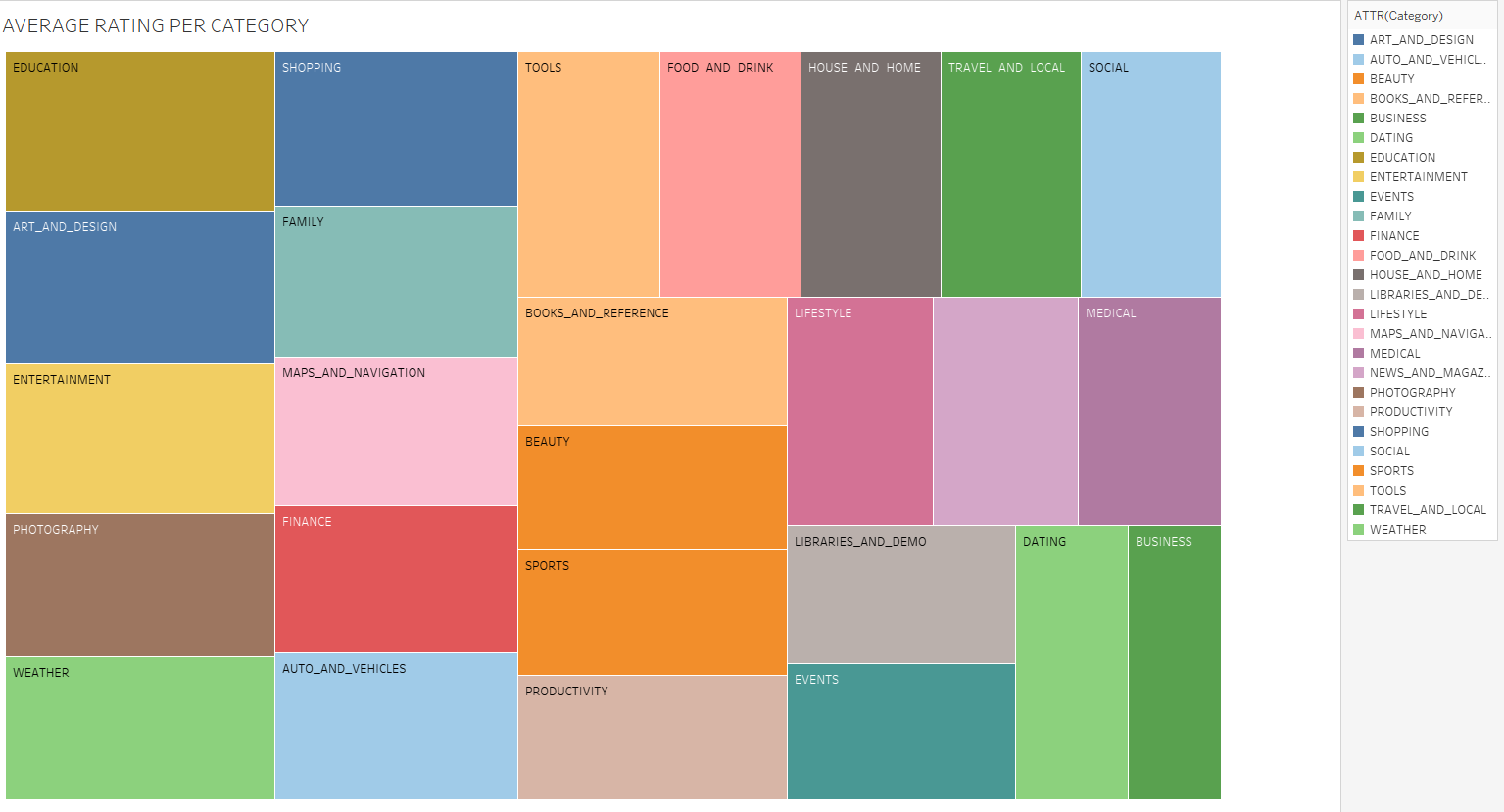
**ii. Count of Installs per category:**

The below visualization shows the number of installs per each category.

****

From the above visualization we can say that the number of installs under Family category are more, the second highest number of installs are in Tools category and least number of installs are in the Beauty Category.

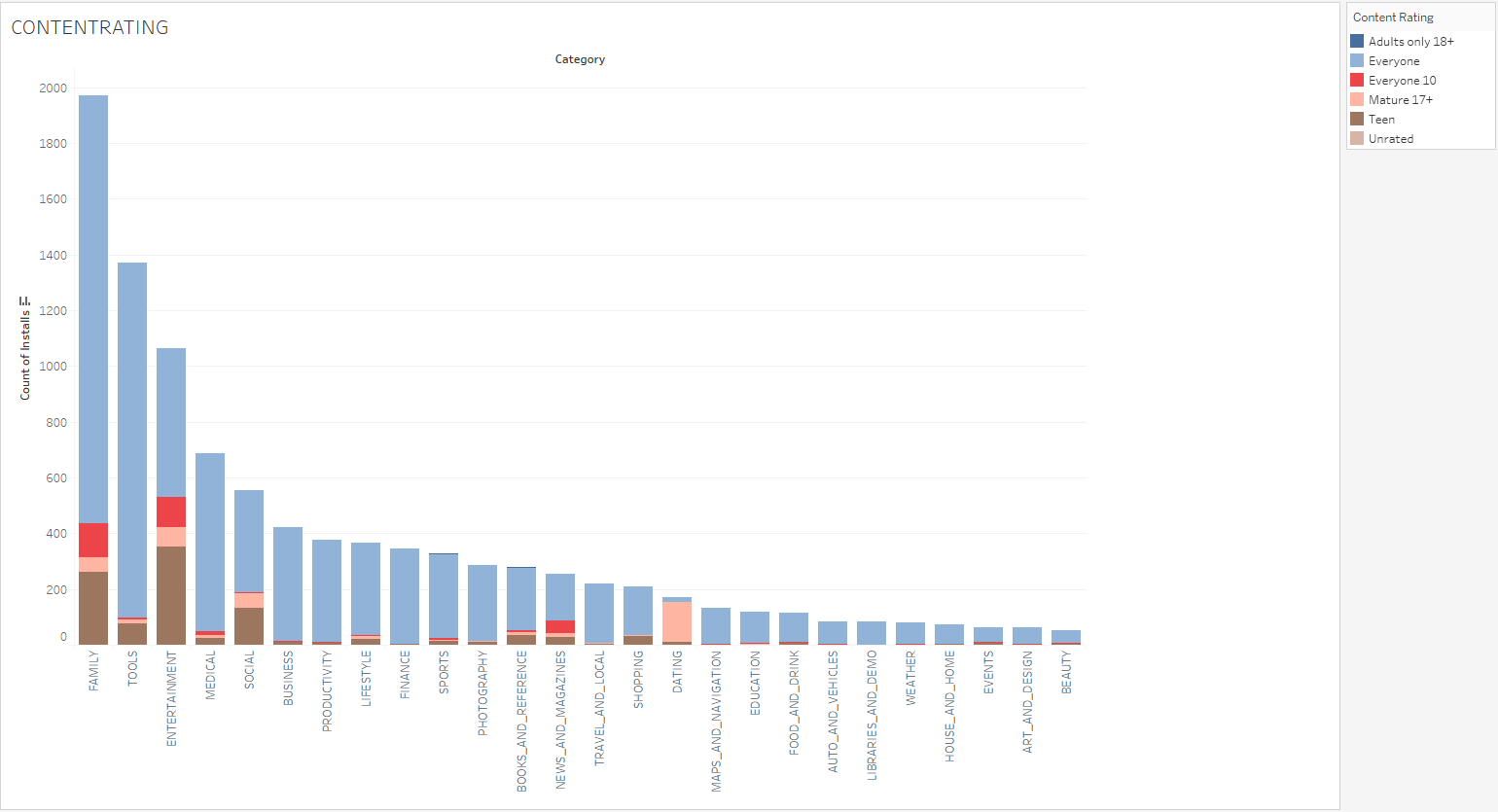
**iii. Average Rating per category:**

****

This visualization gives information about the average rating for the apps present in the category.

Average rating is highest for the Education category (4.328) and the average rating is the least for Business Category (2.570).

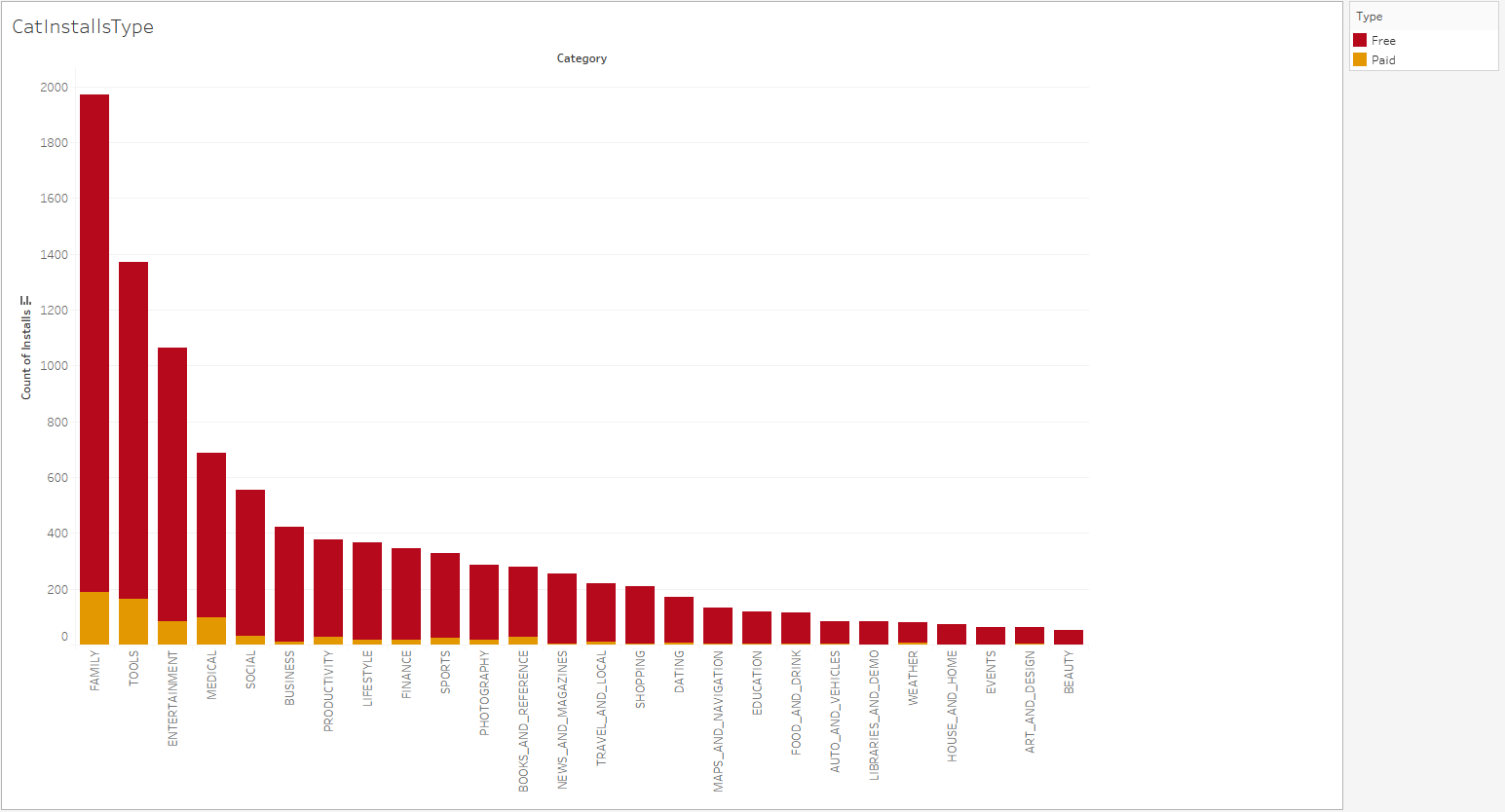
**iv. Content Rating per category:**

****

The above visualization shows the kind of apps(content) installed per each category in the dataset.

From the visualization we can conclude that the number of records with content rating as ‘Everyone’ group are more.

**v. Category Vs Type**

****

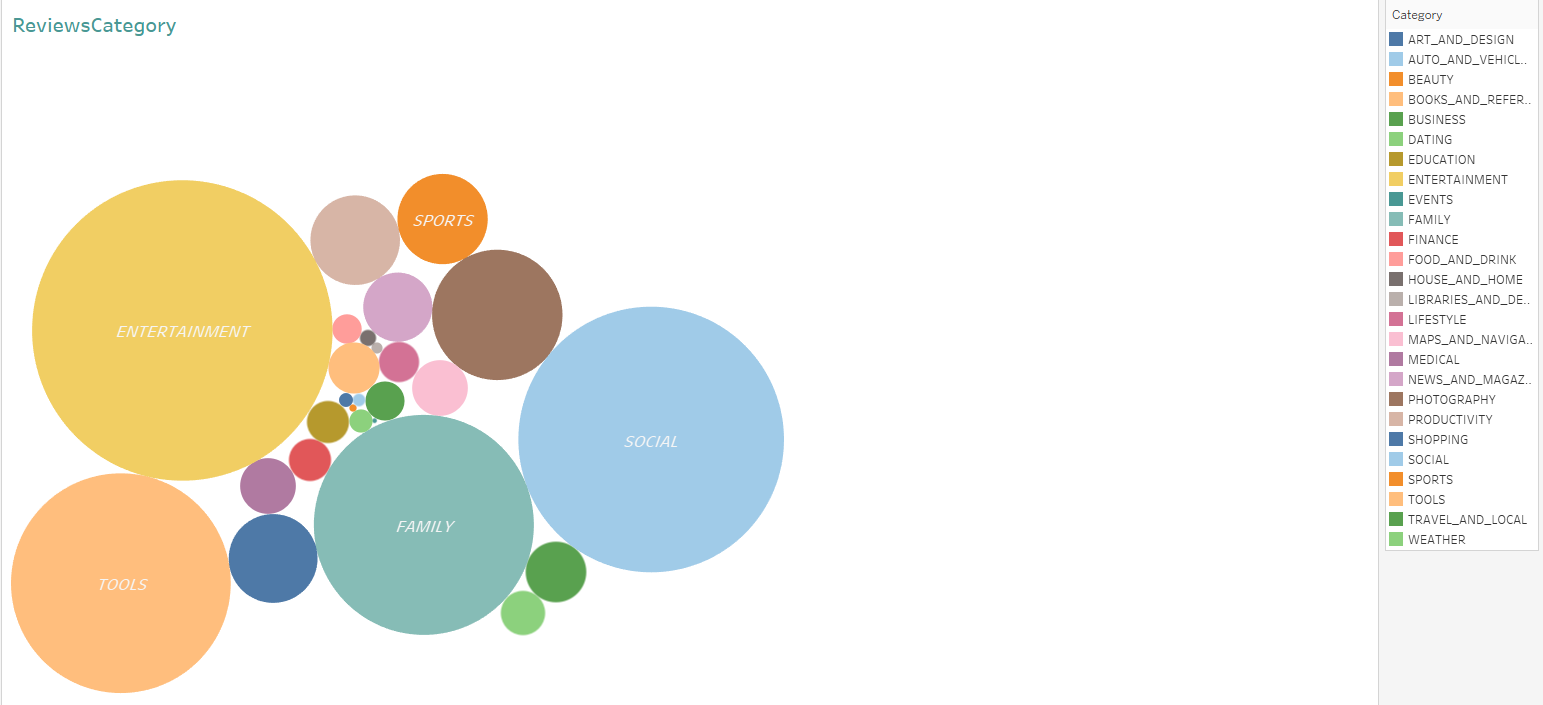
This visualization shows the count of type of apps (Free/paid) installed per each app based on category. From the Visualization we see that most of the categories include free apps and these are the most commonly installed apps.

**vi. Primary Genre:**

****

This visualization throws light on the primary Genre or the most common Genre to which most of the apps in the dataset belongs to. From this visualization we can conclude that, Entertainment is the common Genre in the dataset. Genres like “Role Playing: Education”, “Board: Pretend Play” are the least common genres in the dataset.

**vii. Review Count:**



This visualization aims finding that category that has got the most reviews in the data set. From this visualization, we see that Entertainment followed by Social and Family are the three top most categories in terms of review count. Event followed by Beauty are the two least categories in terms of the review count.