

# Comprehensive Predictions on Google Play Store Apps

## Table of Contents

<b>1. Introduction</b>	2
<b>2. Data</b>	2
<b>3. Problems to be Solved</b>	3
<b>4. Solutions</b>	3
<b>5. Experiments and Results</b>	3
5.1. Methods and Process	3
5.2. Evaluations and Results	9
5.3. Findings	22
<b>6. Conclusions and Future Work</b>	23
6.1. Conclusions	23
6.2. Limitations	23
6.3. Potential Improvements or Future Work	23

## 1. Introduction

Everyday many applications are developed based on different categories that can be business, games, lifestyle, health and fitness, etc. And at the same time there are many apps already present on the biggest platform for all user's i.e. Google Play Store. For the developers to find and predict whether the application they are developing in any category is appropriate enough to compete with other applications of similar type. For ex. Music Players, there are n number of music player applications already present on the platform and if there's a new music player app launched on google play store how it will perform where the competition is so high.

To help this issue we are proposing a solution in this project by predicting the Rating of the app before it is launched by keeping various other attributes as independent and Ratings variable as the only dependent variable. The developers can predict the Rating of the app according to the category, price and size. **This project will predict the Rating of the app without taking reviews into consideration as we are trying to predict the Rating before the launch of the application. Once the application has been launched and the reviews start to build-in then it will change the overall rating of the application (which has not been included in this project).**

## 2. Data

Our dataset consists of over 10,800 rows of entries of application with 13 columns of various attributes. The dataset has been taken from Kaggle (<https://www.kaggle.com/rodolfooluna/google-play-store-apps>). **In our dataset Rating is the dependent variable and other attributes are independent variables.**

The 13 attributes used in this project are:

- App - Name of the Application
- Category- The belonging of the app
- Rating- Overall User rating of the app
- Reviews- Number of reviews by the users for that particular app.
- Size- Application Size
- Type- Whether it is Free or Paid
- Price- Application Price
- Content Rating- Target age group
- Genres- Contains Multiple Categories
- Last Updated- Date when the App was Recently updated
- Current Ver- Latest Version Available for the App
- Android Ver- Minimum android specifications requirements

Attribute	Type
App	Factor
Category	Factor
Rating	Number
Reviews	Integer
Size	Factor
Type	Factor
Price	Factor
Content Rating	Factor
Genres	Factor
Last Updated	Factor
Current Version	Factor
Android version	Factor

## 3. Problems to be Solved

### 3.1 Exploring the Data

This will include some key observations, how the performance of the application can be optimized from the reviews obtained and finding various ways to improve the business as well. Exploring the correlation between the price and size of the app, version and many more based on the number of installations.

Data Preprocessing: Transforming our data language into machine language which can be used for further encoding or decoding of the data required for the process.

- It requires data quality assessment which includes checking for missing values, inconsistent values and duplicate values.
- Dimensional reduction: In data analytics algorithm works better when the dimensions are lower and irrelevant features and noise could be eliminated.

### 3.2 Predicting the Ratings

Again, with the usage of Multiple Regression Model using various attributes, the rating of the application could be solved.

Prediction Analysis: The process of using data analysis to make predictions on data. It uses data along with analysis, statistics, and machine learning techniques to create a predictive model for forecasting future events.

- Using ANOVA to find out about the hypothesis used for linear or multiple regression. Here, we have considered that the average mean value of the ratings with respect to categories are same.
- Here, we will be predicting the rating of the app before its launch on the google play store platform using N- fold Cross Validation.

## 4. Solutions

### 4.1 Exploring the data:

For this problem we will be performing the preprocessing and data cleaning that can be finally be without inconsistent, duplicate and missing values. For example, removing "\$", ",", " from the column Price, "M", "K" from the size column. For dimensional cleaning we will be removing few columns which are not giving helpful information required in the final prediction of the model i.e. Current Version, last updated and Android version.

### 4.2 Predicting the Analysis:

For final prediction of app, we will first use linear model, step wise and finally using N- fold cross validation for the final prediction. We will also build ANOVA model and find out about the hypothesis. **For the prediction models we have used Rating as the dependent variable. Category, Price, Type, Size, Reviews, Installs Content.Rating, Genres, Last.Updated, Current.ver, Android.ver.**

## 5. Experiments and Results

### 5.1. Methods and Process

## 5.1.1 Exploring the data

**Data Preprocessing:** Here we took the dataset from Kaggle for the applications present in google play store, we had around 10841 different rows of application with 13 different columns defining the various details about a particular app. In this there were many entries which had null values, duplicate entries and some inconsistent entries as well.

### Quality Assessment

- First, we found the columns having null values. In the screenshot below we can see that only Ratings column has missing values. For that we have used (summary) to find the missing values in our dataset.

```
> #Gives the count of number of rows and columns present in the dataset.
> paste("No of Observation Is",nrow(app))
[1] "No of Observation Is 10841"
> paste("No of Variable Is",ncol(app))
[1] "No of Variable Is 13"
> #provides the statistics of the dataset
> summary(app)
```

App	Category	Rating	Reviews	Size
ROBLOX	9 FAMILY	1972 Min. :1.000	Min. : 0	25M :1839
CBS Sports App - Scores, News, Stats & Watch Live	8 GAME	1144 1st Qu.:4.000	1st Qu.: 38	11M : 198
8 Ball Pool	7 TOOLS	843 Median :4.300	Median : 2094	12M : 196
Candy Crush Saga	7 MEDICAL	463 Mean :4.191	Mean : 444112	14M : 194
Duolingo: Learn Languages Free	7 BUSINESS	460 3rd Qu.:4.500	3rd Qu.: 54768	13M : 191
ESPN	7 PRODUCTIVITY	424 Max. :5.000	Max. : 78158306	15M : 184
(Other)	10796 (Other)	5535 NA's :1474	(Other):8039	(Other):8039

Installs	Type	Price	Content.Rating	Genres	Last.Updated	Current.Ver
1,000,000+ :1579	Free:10040	0 :10041	Adults only 18+ : 3	Tools : 842	3-Aug-18 : 326	Varies with device:1459
10,000,000+ :1252	NaN : 1	\$0.99 : 148	Everyone :8715	Entertainment: 623	2-Aug-18 : 304	1 : 842
100,000+ :1169	Paid: 800	\$2.99 : 129	Everyone 10+ : 414	Education : 549	31-Jul-18: 294	1.1 : 276
10,000+ :1054		\$1.99 : 73	Mature 17+ : 499	Medical : 463	1-Aug-18 : 285	1.2 : 185
1,000+ : 908		\$4.99 : 72	Teen :1208	Business : 460	30-Jul-18: 211	2 : 165
5,000,000+ : 752		\$3.99 : 63	Unrated : 2	Productivity : 424	25-Jul-18: 164	1.3 : 145
(Other) :4127		(Other): 315		(Other) :7480	(Other) :9257	(Other) :7769

Android.Ver
4.1 and up :2451
4.0.3 and up :1501
4.0 and up :1376
Varies with device:1362
4.4 and up : 980
2.3 and up : 652
(Other) :2519

To get rid of those we have used the mean values of the rating column and have replaced all the NA's with those mean values. By using (is.na) we have replaced the values with mean values and now we can see there are no missing values in the dataset.

```
> # Data Preprocessing For Rating column.
> app$Rating<-ifelse(is.na(app$Rating),mean(app$Rating,na.rm=TRUE),app$Rating)
> app$Rating= round(app$Rating, digits=1)
> summary(app$Rating)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.000	4.100	4.200	4.193	4.500	5.000

```
> summary (app)
```

App	Category	Rating	Reviews	Size
ROBLOX	9 FAMILY	1972 Min. :1.000	Min. : 0	25M :1839
CBS Sports App - Scores, News, Stats & Watch Live	8 GAME	1144 1st Qu.:4.100	1st Qu.: 38	11M : 198
8 Ball Pool	7 TOOLS	843 Median :4.200	Median : 2094	12M : 196
Candy Crush Saga	7 MEDICAL	463 Mean :4.193	Mean : 444112	14M : 194
Duolingo: Learn Languages Free	7 BUSINESS	460 3rd Qu.:4.500	3rd Qu.: 54768	13M : 191
ESPN	7 PRODUCTIVITY	424 Max. :5.000	Max. : 78158306	15M : 184
(Other)	10796 (Other)	5535	(Other):8039	(Other):8039

Installs	Type	Price	Content.Rating	Genres	Last.Updated	Current.Ver
1,000,000+ :1579	Free:10040	0 :10041	Adults only 18+ : 3	Tools : 842	3-Aug-18 : 326	Varies with device:1459
10,000,000+ :1252	NaN : 1	\$0.99 : 148	Everyone :8715	Entertainment: 623	2-Aug-18 : 304	1 : 842
100,000+ :1169	Paid: 800	\$2.99 : 129	Everyone 10+ : 414	Education : 549	31-Jul-18: 294	1.1 : 276
10,000+ :1054		\$1.99 : 73	Mature 17+ : 499	Medical : 463	1-Aug-18 : 285	1.2 : 185
1,000+ : 908		\$4.99 : 72	Teen :1208	Business : 460	30-Jul-18: 211	2 : 165
5,000,000+ : 752		\$3.99 : 63	Unrated : 2	Productivity : 424	25-Jul-18: 164	1.3 : 145
(Other) :4127		(Other): 315		(Other) :7480	(Other) :9257	(Other) :7769

Android.Ver
4.1 and up :2451
4.0.3 and up :1501
4.0 and up :1376
Varies with device:1362
4.4 and up : 980
2.3 and up : 652
(Other) :2519

- We can notice that there is one more missing value in the column type. Here as the Type column is categorical, we cannot determine the mean or median for this hence I have replaced the NaN value with "Free" as it has larger number of entries.

```
> # Data Preprocessing for type column.
> summary(app$type)
Free   NaN   Paid
10040    1    800
> app$type<- str_replace(app$type, "NaN", "Free")
> app$type <- as.factor(app$type)
> summary(app)
```

App	Category	Rating	Reviews	Size
ROBLOX	9 FAMILY	1972	Min. :1.000	Min. : 0 25M :1839
CBS Sports App - Scores, News, Stats & Watch Live	8 GAME	1144	1st Qu.:4.100	1st Qu.: 38 11M : 198
8 Ball Pool	7 TOOLS	843	Median :4.200	Median : 2094 12M : 196
Candy Crush Saga	7 MEDICAL	463	Mean :4.193	Mean : 444112 14M : 194
Duolingo: Learn Languages Free	7 BUSINESS	460	3rd Qu.:4.500	3rd Qu.: 54768 13M : 191
ESPN	7 PRODUCTIVITY	424	Max. :5.000	Max. : 78158306 15M : 184
(Other)	(Other)	5535		(Other):8039

```

Installs      Type      Price      Content.Rating      Genres      Last.Updated      Current.Ver
1,000,000+ :1579 Free:10041 0 :10041 Adults only 18+ : 3 Tools : 842 3-Aug-18 : 326 Varies with device:1459
10,000,000+:1252 Paid: 800 $0.99 : 148 Everyone :8715 Entertainment: 623 2-Aug-18 : 304 1 : 842
100,000+ :1169 $2.99 : 129 Everyone 10+ : 414 Education : 549 31-Jul-18: 294 1.1 : 276
10,000+ :1054 $1.99 : 73 Mature 17+ : 499 Medical : 463 1-Aug-18 : 285 1.2 : 185
1,000+ : 908 $4.99 : 72 Teen :1208 Business : 460 30-Jul-18: 211 2 : 165
5,000,000+ : 752 $3.99 : 63 Unrated : 2 Productivity : 424 25-Jul-18: 164 1.3 : 145
(Other) :4127 (Other): 315 (Other) :17480 (Other) :9257 (Other) :7769

Android.ver
4.1 and up :2451
4.0.3 and up :1501
4.0 and up :1376
Varies with device:1362
4.4 and up : 980
2.3 and up : 652
(Other) :2519
```

## Dimension Reduction

- In Data Analytics it is important that our data should be as simplified as it can so that the analysis we want to perform are easily conducted and for that we can remove some irrelevant features and dimensions from our columns. We can notice from above screenshot that in the column for size there are the dimensions like M and K representing the size in bytes. Hence, we have removed them. Similarly, for the price column we have removed the [\$] dimension. Last, we have removed the [+], [,] dimensions from Installs column.

```
> # Data Cleaning for Size column.
> app$Size<-gsub("M","",app$Size)
> app$Size<-gsub("K","",app$Size)
> app$Size<-as.numeric(app$Size)
> head(app$Size)
[1] 19.0 14.0 8.7 25.0 2.8 5.6
> # Data Cleaning for Price column.
> app$Price<- str_replace(app$Price,"[$]","")
> app$Price = as.numeric(app$Price)
> summary(app$Price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.000  0.000  0.000  1.027  0.000 400.000
> # Data Cleaning for Installs column.
> app$Installs<-gsub("[,]", "", app$Installs)
> app$Installs<-gsub("[+]", "", app$Installs)
> app$Installs <- as.factor(app$Installs)
> head(app$Installs)
[1] 10000 500000 5000000 50000000 100000 50000
Levels: 0 1 10 100 1000 10000 100000 1000000 10000000 ;
```

- It becomes easy to convert or replace some values or entries if they are numeric to remove noise, but some character entries are difficult to replace. Hence, in the Content.Rating where were two entries which couldn't be replaced and at last, we have removed them as they were irrelevant entries too.



This is done using the function [indices], we found the two rows which were defined as Unrated and hence removed them. This ended up with no [Unrated] entries and displayed the unique entries in the Content.Rating column.

```
> # Data Cleaning for Content.Rating column.
> summary(app$Content.Rating)
Adults only 18+      Everyone      Everyone 10+      Mature 17+      Teen      Unrated
           3           8715           414           499           1208           2
> indices= which(app$Content.Rating == "Unrated")
> indices
[1] 7313 8267
> app= app[c(-7313, -8267),]
> app$Content.Rating <- as.factor(app$Content.Rating)
> summary(app$Content.Rating)
Adults only 18+      Everyone      Everyone 10+      Mature 17+      Teen      Unrated
           3           8715           414           499           1208           0
> unique(app$Content.Rating)
[1] Everyone      Teen      Everyone 10+      Mature 17+      Adults only 18+
Levels: Adults only 18+ Everyone Everyone 10+ Mature 17+ Teen Unrated
```

- Again, there were few columns which were not providing enough information to us for the final output we were looking for i.e. prediction of the Rating of the App and decided to remove them. These columns were Last.Updated, Current.ver, Android.Ver. Moreover, Genres was same as Category i.e. giving the same information. Changed the reviews datatype to numeric to make the dataset easy for processing and the final attributes we will be using for the model to predict using [str] function.

```
> # Removing the columns.
> # They have been removed as they are not giving me enough information which is required
> # in my prediction analysis
> # Removing Columns like: Android.ver, Current.Ver, Last.Updated. Also
> # As category is similar to Genres. hence removed.
> app= app[,c(-10,-11,-12,-13)]
> # Changing the attribute
> app$Reviews<-as.numeric(app$Reviews)
> # Defines the various attributes of the columns.
> str(app)
'data.frame':   10839 obs. of  9 variables:
 $ App      : Factor w/ 9660 levels "\i DT\ FÅ°tbo1. Todos Somos TÃ©cnicos.",...: 7229 2563 8998
 $ Category : Factor w/ 33 levels "ART_AND_DESIGN",...: 1 1 1 1 1 1 1 1 1 ...
 $ Rating   : num  4.1 3.9 4.7 4.5 4.3 4.4 3.8 4.1 4.4 4.7 ...
 $ Reviews  : num  159 967 87510 215644 967 ...
 $ Size     : num  19 14 8.7 25 2.8 5.6 19 29 33 3.1 ...
 $ Installs : Factor w/ 20 levels "0","1","10","100",...: 6 17 18 19 7 16 16 8 8 6 ...
 $ Type     : Factor w/ 2 levels "Free","Paid": 1 1 1 1 1 1 1 1 1 1 ...
 $ Price    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Content.Rating: Factor w/ 6 levels "Adults only 18+",...: 2 2 2 5 2 2 2 2 2 2 ...
```

- Finally, the final data has been renamed and viewed. Also, the final .CSV file has been generated after preprocessing and cleaning and is now ready for prediction analysis.

```

> #Changing the name
> dataset<-app
> #Viewing of the dataset after the cleaning and preprocessing.
> View(dataset)
> #Importing the preprocessed and cleaned CSV file.
> write.csv(dataset, "C:\\Users\\13128\\Desktop\\google-play-store-apps\\final output.csv")

```

	App	Category	Rating	Reviews	Size	Installs	Type	Price	ContentRating
1	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19.0	10000	Free	0	Everyone
2	Coloring book moana	ART_AND_DESIGN	3.9	967	14.0	500000	Free	0	Everyone
3	U Launcher Lite &™ FREE Live Cool Themes, Hide Apps	ART_AND_DESIGN	4.7	87510	8.7	5000000	Free	0	Everyone
4	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25.0	50000000	Free	0	Teen
5	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8	100000	Free	0	Everyone
6	Paper flowers instructions	ART_AND_DESIGN	4.4	167	5.6	50000	Free	0	Everyone
7	Smoke Effect Photo Maker - Smoke Editor	ART_AND_DESIGN	3.8	178	19.0	50000	Free	0	Everyone
8	Infinite Painter	ART_AND_DESIGN	4.1	36815	29.0	1000000	Free	0	Everyone
9	Garden Coloring Book	ART_AND_DESIGN	4.4	13791	33.0	1000000	Free	0	Everyone
10	Kids Paint Free - Drawing Fun	ART_AND_DESIGN	4.7	121	3.1	10000	Free	0	Everyone

### 5.1.2 Predicting the Analysis:

The detailed analysis will be in section (5.2). Here we will work on the ANOVA model.

ANOVA model for hypothesis is used to differentiate between more than 2 box plots to determine the statistics value of the model.

For the prediction we decide to make 2 hypotheses based on our project:

- 1) Null Hypothesis ( $\mu_0$ ) = Average mean of the ratings is same for all the categories.  
 $\mu_{\text{CategoryGames}} = \mu_{\text{CategoryLifestyles}} = \mu_{\text{CategoryComics}} = \mu_{\text{CategoryShopping}} \dots\dots = \mu_0$
- 2) Alternate Hypothesis ( $\mu_a$ ) = Average mean of ratings is not same for all the categories.  
 $\mu_{\text{CategoryGames}} \neq \mu_{\text{CategoryLifestyles}} \neq \mu_{\text{CategoryComics}} \neq \mu_{\text{CategoryShopping}} \dots\dots = \mu_a$

```

> #Building Anova model
> anova=lm(dataset$Rating~dataset$Category)
> summary(anova)

Call:
lm(formula = dataset$Rating ~ dataset$Category)

Residuals:
    Min       1Q   Median       3Q      Max
-3.2828 -0.1262  0.0558  0.2795  0.9910

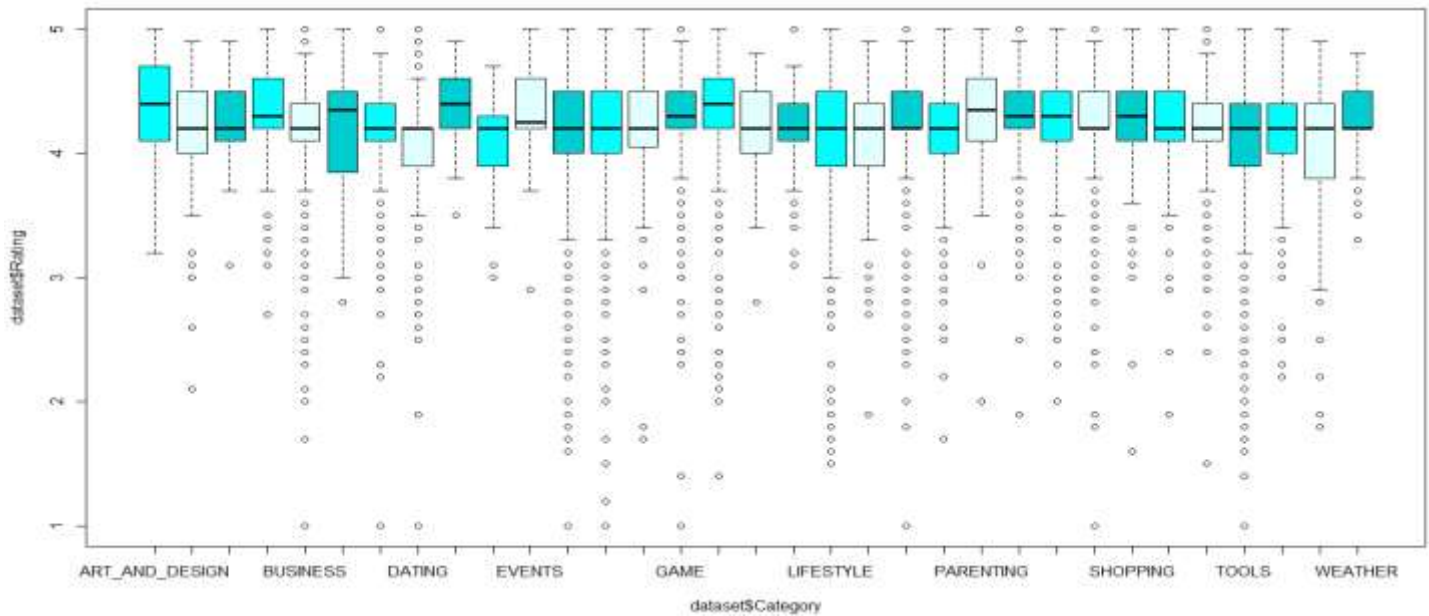
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.35077    0.05877   74.029 < 2e-16 ***
dataset$CategoryAUTO_AND_VEHICLES -0.15900    0.07807   -2.037  0.041712 *
dataset$CategoryBEAUTY -0.08851    0.08769   -1.009  0.312872
dataset$CategoryBOOKS_AND_REFERENCE -0.03822    0.06653   -0.574  0.565692
dataset$CategoryBUSINESS -0.20251    0.06279   -3.225  0.001262 **
dataset$CategoryCOMICS -0.19410    0.08483   -2.288  0.022146 *
dataset$CategoryCOMMUNICATION -0.18591    0.06351   -2.927  0.003429 **
dataset$CategoryDATING -0.34179    0.06643   -5.145  2.72e-07 ***
dataset$CategoryEDUCATION  0.03705    0.06995    0.530  0.596350
dataset$CategoryENTERTAINMENT -0.22459    0.07043   -3.189  0.001433 **
dataset$CategoryEVENTS  0.01486    0.08344    0.178  0.858691
dataset$CategoryFAMILY -0.15762    0.05973   -2.639  0.008333 **
dataset$CategoryFINANCE -0.21088    0.06378   -3.307  0.000948 ***
dataset$CategoryFOOD_AND_DRINK -0.17912    0.07226   -2.479  0.013202 *
dataset$CategoryGAME -0.06799    0.06042   -1.125  0.260474
dataset$CategoryHEALTH_AND_FITNESS -0.08361    0.06413   -1.304  0.192310
dataset$CategoryHOUSE_AND_HOME -0.15304    0.07749   -1.975  0.048305 *
dataset$CategoryLIBRARIES_AND_DEMO -0.16724    0.07807   -2.142  0.032207 *
dataset$CategoryLIFESTYLE -0.23716    0.06357   -3.730  0.000192 ***
dataset$CategoryMAPS_AND_NAVIGATION -0.28508    0.07136   -3.995  6.52e-05 ***
dataset$CategoryMEDICAL -0.15898    0.06276   -2.533  0.011321 *
dataset$CategoryNEWS_AND_MAGAZINES -0.20660    0.06517   -3.170  0.001528 **
dataset$CategoryPARENTING -0.06744    0.08483   -0.795  0.426649
dataset$CategoryPERSONALIZATION -0.04797    0.06345   -0.756  0.449611
dataset$CategoryPHOTOGRAPHY -0.15823    0.06422   -2.464  0.013759 *
dataset$CategoryPRODUCTIVITY -0.14134    0.06312   -2.239  0.025155 *
dataset$CategorySHOPPING -0.09615    0.06571   -1.463  0.143399
dataset$CategorySOCIAL -0.10196    0.06492   -1.570  0.116352
dataset$CategorySPORTS -0.13124    0.06355   -2.065  0.038938 *
dataset$CategoryTOOLS -0.28367    0.06100   -4.651  3.35e-06 ***
dataset$CategoryTRAVEL_AND_LOCAL -0.23023    0.06576   -3.501  0.000465 ***
dataset$CategoryVIDEO_PLAYERS -0.27534    0.06883   -4.001  6.36e-05 ***
dataset$CategoryWEATHER -0.11053    0.07869   -1.405  0.160173
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4738 on 10806 degrees of freedom
Multiple R-squared:  0.02618, Adjusted R-squared:  0.02329
F-statistic: 9.078 on 32 and 10806 DF, p-value: < 2.2e-16

```

From the above output of ANOVA model, we could find that many categories p value is less than 0.05 when taking confidence level of 95%, which falls in the rejection region and hence we can conclude that with 95% confidence level we have the evidence to reject the null hypothesis and come to conclusion that our alternative hypothesis is true that different categories have different mean of ratings or in other terms there will be at-least one category which will have different average mean value of rating from other categories mean.





To prove it further we will find which category have different or same average mean value from others, and for that we must perform individual parameter test.

To perform individual parameter testing we will convert the ANOVA model into linear regression model and clarify about the Beta Values.

For example:  $\beta_{\text{VideoPlayer}} = \mu_{\text{VideoPlayer}} - \mu_{\text{CategoryDating}}$

If we find the individual parameter test, we can notice there are many categories with different p values some are greater than 0.05 (95% Confidence level) and some are less than 0.05. If we watch closely the lowest value is of Video player category and lesser than that is Dating category. Here we can conclude that we can reject the null hypothesis and accept the alternate hypothesis.

Our Null Hypothesis, in this case, is,  $\beta$  coefficient corresponding to  $\text{CategoryVideoPlayer}$  and  $\text{CategoryDating} = 0$ , whereas the Alternate Hypothesis  $\neq 0$ .

And  $\beta$  value here is the difference between the average mean value of Rating of  $\text{CategoryVideoPlayer}/\text{CategoryDating}$  and  $\text{CategoryLifeStyle}$  ( $\text{CategoryLifeStyle}$  is considered as the base value here)

We can further do the analysis; which category has largest mean value of Rating and which has the lowest. Since t-value in case of  $\text{CategoryVideoPlayer}$  is positive, hence  $\beta_{\text{VideoPlayer}} = \mu_{\text{VideoPlayer}} - \mu_{\text{CategoryLifeStyle}} > 0$  which means  $\mu_{\text{VideoPlayer}}$  is greater than  $\text{CategoryLifeStyle}$

Also, since t-value in case of  $\text{CategoryDating}$  is negative, hence  $\beta_{\text{CategoryDating}} = \mu_{\text{CategoryDating}} - \mu_{\text{CategoryLifeStyle}} < 0$  which means  $\mu_{\text{CategoryDating}} < \mu_{\text{CategoryLifeStyle}}$

We can come up with the complete conclusion now that  $\mu_{\text{CategoryDating}}$  will have the lowest average mean value of Rating whereas  $\mu_{\text{VideoPlayer}}$  will have the highest value.

## 5.2. Evaluations and Results

Predicting the Analysis: Firstly, we have used multiple linear regression to do the feature selection and finally used N-Fold cross validation to predict the model.

We have used stepwise analysis for all the directions and then selected the features that can be used for the final output used in N-Fold Cross validation.

- Building the linear model and rejecting the variables which have the p value greater than 0.05 as they are non-significant.

```
> #Building linear regression model.
> # our multiple linear model
> multiple<- lm(Rating~ Category+Type+Size+Price+Reviews+Installs+Content.Rating, data= dataset)
> summary(multiple)

Call:
lm(formula = Rating ~ Category + Type + Size + Price + Reviews +
    Installs + Content.Rating, data = dataset)

Residuals:
    Min       1Q   Median       3Q      Max
-3.4003 -0.1489  0.0426  0.2575  1.0816

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.474e+00  3.012e-01  14.852  < 2e-16 ***
CategoryAUTO_AND_VEHICLES -1.838e-01  7.624e-02  -2.411  0.015907 *
CategoryBEAUTY -8.520e-02  8.560e-02  -0.995  0.319587
CategoryBOOKS_AND_REFERENCE -8.458e-02  6.508e-02  -1.300  0.193737
CategoryBUSINESS -2.591e-01  6.155e-02  -4.210  2.58e-05 ***
CategoryCOMICS -1.915e-01  8.353e-02  -2.293  0.021876 *
CategoryCOMMUNICATION -2.992e-01  6.235e-02  -4.798  1.63e-06 ***
CategoryDATING -3.507e-01  6.898e-02  -5.084  3.75e-07 ***
CategoryEDUCATION -4.993e-02  6.846e-02  -0.729  0.465808
CategoryENTERTAINMENT -3.584e-01  6.973e-02  -5.140  2.80e-07 ***
CategoryEVENTS -7.161e-03  8.159e-02  -0.088  0.930066
CategoryFAMILY -2.029e-01  5.842e-02  -3.472  0.000518 ***
CategoryFINANCE -2.281e-01  6.236e-02  -3.658  0.000256 ***
CategoryFOOD_AND_DRINK -2.435e-01  7.062e-02  -3.447  0.000568 ***
CategoryGAME -1.879e-01  5.945e-02  -3.161  0.001578 **
CategoryHEALTH_AND_FITNESS -1.570e-01  6.273e-02  -2.503  0.012334 *
CategoryHOUSE_AND_HOME -2.077e-01  7.571e-02  -2.744  0.006078 **
CategoryLIBRARIES_AND_DEMO -1.341e-01  7.662e-02  -1.750  0.080166 .
CategoryLIFESTYLE -2.643e-01  6.218e-02  -4.251  2.15e-05 ***
CategoryMAPS_AND_NAVIGATION -3.336e-01  6.974e-02  -4.784  1.74e-06 ***
CategoryMEDICAL -1.899e-01  6.151e-02  -3.087  0.002026 **
CategoryNEWS_AND_MAGAZINES -2.518e-01  6.400e-02  -3.935  8.38e-05 ***
CategoryPARENTING -5.914e-02  8.278e-02  -0.714  0.475022
CategoryPERSONALIZATION -1.277e-01  6.216e-02  -2.055  0.039901 *
CategoryPHOTOGRAPHY -2.858e-01  6.296e-02  -4.539  5.72e-06 ***
CategoryPRODUCTIVITY -2.378e-01  6.182e-02  -3.847  0.000120 ***
CategorySHOPPING -2.046e-01  6.437e-02  -3.179  0.001480 **
CategorySOCIAL -1.905e-01  6.420e-02  -2.967  0.003013 **
CategorySPORTS -2.049e-01  6.218e-02  -3.295  0.000987 ***
CategoryTOOLS -3.370e-01  5.965e-02  -5.650  1.65e-08 ***
CategoryTRAVEL_AND_LOCAL -3.016e-01  6.432e-02  -4.688  2.79e-06 ***
CategoryVIDEO_PLAYERS -3.524e-01  6.732e-02  -5.235  1.68e-07 ***
CategoryWEATHER -1.916e-01  7.690e-02  -2.491  0.012737 *
TypePaid  1.279e-01  1.842e-02  6.943  4.05e-12 ***
Size -1.137e-04  4.963e-05  -2.292  0.021952 *
Price -7.168e-04  2.922e-04  -2.453  0.014188 *
Reviews  4.705e-09  2.047e-09  2.299  0.021533 *

Installs1  4.114e-02  1.340e-01  0.307  0.758849
Installs10 9.989e-02  1.239e-01  0.806  0.420157
Installs100 1.073e-01  1.229e-01  0.873  0.382495
Installs1000 -7.785e-02  1.225e-01  -0.635  0.525154
Installs10000 -1.190e-01  1.224e-01  -0.972  0.331252
Installs100000 -5.482e-02  1.225e-01  -0.448  0.654456
Installs1000000 6.995e-02  1.224e-01  0.571  0.567714
Installs10000000 1.660e-01  1.226e-01  1.354  0.175762
Installs100000000 2.482e-01  1.245e-01  1.993  0.046235 *
Installs1000000000 4.670e-02  1.431e-01  0.326  0.744230
Installs5  7.283e-02  1.320e-01  0.552  0.581159
Installs50  9.086e-02  1.259e-01  0.722  0.470496
Installs500  2.683e-02  1.244e-01  0.216  0.829224
Installs5000 -1.182e-01  1.234e-01  -0.958  0.338209
Installs50000 -1.130e-01  1.235e-01  -0.915  0.360076
Installs500000 2.050e-03  1.235e-01  0.017  0.986749
Installs5000000 9.138e-02  1.230e-01  0.743  0.457628
Installs50000000 2.024e-01  1.250e-01  1.619  0.105393
Installs500000000 1.815e-01  1.352e-01  1.343  0.179392
Content.RatingEveryone -9.450e-02  2.702e-01  -0.350  0.726583
Content.RatingEveryone 10+ -9.899e-02  2.711e-01  -0.365  0.715017
Content.RatingMature 17+ -1.264e-01  2.712e-01  -0.466  0.641206
Content.RatingTeen -9.397e-02  2.704e-01  -0.347  0.728225
--
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.462 on 10779 degrees of freedom
Multiple R-squared:  0.0766,    Adjusted R-squared:  0.07155
F-statistic: 15.16 on 59 and 10779 DF,  p-value: < 2.2e-16
```

- From the above two screenshots we can find that all the values of install and content rating are greater than 0.05. hence, we can remove them for the final prediction and take the variables which are significant. The variables which are significant and can be used in the process of predictions are: Category, Type, Size, Price, Reviews.
- Now we will build another model which will have just the above variables.**

```
> multiple2<- lm(Rating~ Category+Type+Size+Price+Reviews, data= dataset)
> summary(multiple2)
```

Call:

```
lm(formula = Rating ~ Category + Type + Size + Price + Reviews,
    data = dataset)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-3.2679 -0.1423  0.0568  0.2569  1.0002
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	4.349e+00	5.861e-02	74.197	< 2e-16	***
CategoryAUTO_AND_VEHICLES	-1.554e-01	7.785e-02	-1.996	0.045960	*
CategoryBEAUTY	-8.461e-02	8.744e-02	-0.968	0.333260	
CategoryBOOKS_AND_REFERENCE	-4.114e-02	6.636e-02	-0.620	0.535320	
CategoryBUSINESS	-1.994e-01	6.261e-02	-3.185	0.001452	**
CategoryCOMICS	-1.884e-01	8.459e-02	-2.228	0.025923	*
CategoryCOMMUNICATION	-2.051e-01	6.344e-02	-3.232	0.001231	**
CategoryDATING	-3.400e-01	6.624e-02	-5.133	2.91e-07	***
CategoryEDUCATION	3.830e-02	6.976e-02	0.549	0.582987	
CategoryENTERTAINMENT	-2.247e-01	7.024e-02	-3.199	0.001382	**
CategoryEVENTS	1.963e-02	8.320e-02	0.236	0.813529	
CategoryFAMILY	-1.594e-01	5.958e-02	-2.676	0.007459	**
CategoryFINANCE	-2.032e-01	6.364e-02	-3.193	0.001414	**
CategoryFOOD_AND_DRINK	-1.760e-01	7.206e-02	-2.443	0.014591	*
CategoryGAME	-8.016e-02	6.030e-02	-1.329	0.183795	
CategoryHEALTH_AND_FITNESS	-8.165e-02	6.395e-02	-1.277	0.201749	
CategoryHOUSE_AND_HOME	-1.482e-01	7.728e-02	-1.918	0.055177	.
CategoryLIBRARIES_AND_DEMO	-1.439e-01	7.829e-02	-1.838	0.066093	.
CategoryLIFESTYLE	-2.306e-01	6.342e-02	-3.636	0.000278	***
CategoryMAPS_AND_NAVIGATION	-2.841e-01	7.116e-02	-3.993	6.57e-05	***
CategoryMEDICAL	-1.670e-01	6.268e-02	-2.664	0.007722	**
CategoryNEWS_AND_MAGAZINES	-2.049e-01	6.499e-02	-3.153	0.001621	**
CategoryPARENTING	-6.505e-02	8.458e-02	-0.769	0.441878	
CategoryPERSONALIZATION	-5.888e-02	6.335e-02	-0.930	0.352648	
CategoryPHOTOGRAPHY	-1.647e-01	6.404e-02	-2.571	0.010146	*
CategoryPRODUCTIVITY	-1.425e-01	6.294e-02	-2.264	0.023595	*
CategorySHOPPING	-9.540e-02	6.553e-02	-1.456	0.145496	
CategorySOCIAL	-1.192e-01	6.482e-02	-1.839	0.065877	.
CategorySPORTS	-1.317e-01	6.337e-02	-2.078	0.037719	*
CategoryTOOLS	-2.861e-01	6.085e-02	-4.701	2.62e-06	***
CategoryTRAVEL_AND_LOCAL	-2.308e-01	6.557e-02	-3.520	0.000434	***
CategoryVIDEO_PLAYERS	-2.786e-01	6.864e-02	-4.060	4.95e-05	***
CategoryWEATHER	-1.144e-01	7.847e-02	-1.458	0.144849	
TypePaid	7.634e-02	1.827e-02	4.179	2.95e-05	***
Size	-1.271e-04	5.069e-05	-2.507	0.012179	*
Price	-7.664e-04	2.943e-04	-2.604	0.009221	**
Reviews	1.041e-08	1.583e-09	6.577	5.03e-11	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4725 on 10802 degrees of freedom

Multiple R-squared: 0.03216, Adjusted R-squared: 0.02894

F-statistic: 9.972 on 36 and 10802 DF, p-value: < 2.2e-16

- Now we will perform forward selection model for the new model that we have used.

# there are many screenshots of each step but here included only the final output. Rest can be found in the Output pdf.

```
> #Performing stepwise forward regression
> ols_step_forward_p(multiple2, details= TRUE)
Forward Selection Method
```

Candidate Terms:

1. Category
2. Type
3. Size
4. Price
5. Reviews

We are selecting variables based on p value...

Forward Selection: Step 1

+ Category

Model Summary					
R	0.162	RMSE	0.474		
R-Squared	0.026	Coef. var	11.301		
Adj. R-Squared	0.023	MSE	0.225		
Pred R-Squared	0.021	MAE	0.316		
RMSE: Root Mean Square Error					
MSE: Mean Square Error					
MAE: Mean Absolute Error					
ANOVA					
	Sum of Squares	DF	Mean Square	F	Sig.
Regression	65.218	32	2.038	9.078	0.0000
Residual	2426.061	10806	0.225		
Total	2491.278	10838			

Variables Entered:

- + Category
- + Reviews
- + Type
- + Price
- + Size

Final Model output

Model Summary					
R	0.179	RMSE	0.472		
R-Squared	0.032	Coef. var	11.269		
Adj. R-Squared	0.029	MSE	0.223		
Pred R-Squared	0.026	MAE	0.314		
RMSE: Root Mean Square Error					
MSE: Mean Square Error					
MAE: Mean Absolute Error					
ANOVA					
	Sum of Squares	DF	Mean Square	F	Sig.
Regression	80.129	36	2.226	9.972	0.0000
Residual	2411.149	10802	0.223		
Total	2491.278	10838			



Parameter Estimates							
model	Beta	Std. Error	Std. Beta	t	Sig.	lower	upper
(Intercept)	4.349	0.059		74.197	0.000	4.234	4.464
CategoryAUTO_AND_VEHICLES	-0.155	0.078	-0.029	-1.996	0.046	-0.308	-0.003
CategoryBEAUTY	-0.085	0.087	-0.012	-0.968	0.333	-0.256	0.087
CategoryBOOKS_AND_REFERENCE	-0.041	0.066	-0.012	-0.620	0.535	-0.171	0.089
CategoryBUSINESS	-0.199	0.063	-0.084	-3.185	0.001	-0.322	-0.077
CategoryCOMICS	-0.188	0.085	-0.029	-2.228	0.026	-0.354	-0.023
CategoryCOMMUNICATION	-0.205	0.063	-0.079	-3.232	0.001	-0.329	-0.081
CategoryDATING	-0.340	0.066	-0.103	-5.133	0.000	-0.470	-0.210
CategoryEDUCATION	0.038	0.070	0.010	0.549	0.583	-0.098	0.175
CategoryENTERTAINMENT	-0.225	0.070	-0.055	-3.199	0.001	-0.362	-0.087
CategoryEVENTS	0.020	0.083	0.003	0.236	0.814	-0.143	0.183
CategoryFAMILY	-0.159	0.060	-0.128	-2.676	0.007	-0.276	-0.043
CategoryFINANCE	-0.203	0.064	-0.077	-3.193	0.001	-0.328	-0.078
CategoryFOOD_AND_DRINK	-0.176	0.072	-0.040	-2.443	0.015	-0.317	-0.035
CategoryGAME	-0.080	0.060	-0.051	-1.329	0.184	-0.198	0.038
CategoryHEALTH_AND_FITNESS	-0.082	0.064	-0.030	-1.277	0.202	-0.207	0.044
CategoryHOUSE_AND_HOME	-0.148	0.077	-0.028	-1.918	0.055	-0.300	0.003
CategoryLIBRARIES_AND_DEMO	-0.144	0.078	-0.026	-1.838	0.066	-0.297	0.010
CategoryLIFESTYLE	-0.231	0.063	-0.089	-3.636	0.000	-0.355	-0.106
CategoryMAPS_AND_NAVIGATION	-0.284	0.071	-0.066	-3.993	0.000	-0.424	-0.145
CategoryMEDICAL	-0.167	0.063	-0.070	-2.664	0.008	-0.290	-0.044
CategoryNEWS_AND_MAGAZINES	-0.205	0.065	-0.068	-3.153	0.002	-0.332	-0.078
CategoryPARENTING	-0.065	0.085	-0.010	-0.769	0.442	-0.231	0.101
CategoryPERSONALIZATION	-0.059	0.063	-0.023	-0.930	0.353	-0.183	0.065
CategoryPHOTOGRAPHY	-0.165	0.064	-0.059	-2.571	0.010	-0.290	-0.039
CategoryPRODUCTIVITY	-0.143	0.063	-0.058	-2.264	0.024	-0.266	-0.019
CategorySHOPPING	-0.095	0.066	-0.030	-1.456	0.145	-0.224	0.033
CategorySOCIAL	-0.119	0.065	-0.040	-1.839	0.066	-0.246	0.008
CategorySPORTS	-0.132	0.063	-0.051	-2.078	0.038	-0.256	-0.007
CategoryTOOLS	-0.286	0.061	-0.160	-4.701	0.000	-0.405	-0.167
CategoryTRAVEL_AND_LOCAL	-0.231	0.066	-0.073	-3.520	0.000	-0.359	-0.102
CategoryVIDEO_PLAYERS	-0.279	0.069	-0.073	-4.060	0.000	-0.413	-0.144
CategoryWEATHER	-0.114	0.078	-0.021	-1.458	0.145	-0.268	0.039
Reviews	0.000	0.000	0.064	6.577	0.000	0.000	0.000
TypePaid	0.076	0.018	0.042	4.179	0.000	0.041	0.112
Price	-0.001	0.000	-0.025	-2.604	0.009	-0.001	0.000
Size	0.000	0.000	-0.024	-2.507	0.012	0.000	0.000

Selection Summary						
Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Category	0.0262	0.0233	33.8041	14603.0460	0.4738
2	Reviews	0.0299	0.0269	-5.3479	14563.9289	0.4729
3	Type	0.0310	0.0279	-15.8804	14553.3760	0.4727
4	Price	0.0316	0.0285	-20.7133	14548.5259	0.4726
5	Size	0.0322	0.0289	-25.0000	14544.2195	0.4725

- In the above screenshots we can find what all variables were used for forward stepwise regression. At last the final model we can find the variables with different values of prediction errors:

1) RMSE: 0.472    2) R-Squared: 0.032



- Now we will find the stepwise regression model for backward direction.

```
> #Performing stepwise backward regression
> ols_step_backward_p(multiple2, details= TRUE)
Backward Elimination Method
-----
```

Candidate Terms:

```
1 . Category
2 . Type
3 . Size
4 . Price
5 . Reviews
```

we are eliminating variables based on p value...

No more variables satisfy the condition of p value = 0.3

Variables Removed:

Final Model Output

Model Summary				
R	0.179	RMSE		0.472
R-squared	0.032	Coef. Var		11.269
Adj. R-squared	0.029	MSE		0.223
Pred R-squared	0.026	MAE		0.314

RMSE: Root Mean Square Error  
MSE: Mean Square Error  
MAE: Mean Absolute Error

ANOVA					
	Sum of Squares	DF	Mean Square	F	sig.
Regression	80.129	36	2.226	9.972	0.0000
Residual	2411.149	10802	0.223		
Total	2491.278	10838			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	sig	lower	upper
(Intercept)	4.349	0.059		74.197	0.000	4.234	4.464
CategoryAUTO_AND_VEHICLES	-0.155	0.078	-0.029	-1.996	0.046	-0.308	-0.003
CategoryBEAUTY	-0.085	0.087	-0.012	-0.968	0.333	-0.256	0.087
CategoryBOOKS_AND_REFERENCE	-0.041	0.066	-0.012	-0.620	0.535	-0.171	0.089
CategoryBUSINESS	-0.199	0.063	-0.084	-3.185	0.001	-0.322	-0.077
CategoryCOMICS	-0.188	0.085	-0.029	-2.228	0.026	-0.354	-0.023
CategoryCOMMUNICATION	-0.205	0.063	-0.079	-3.232	0.001	-0.329	-0.081
CategoryDATING	-0.340	0.066	-0.103	-5.133	0.000	-0.470	-0.210
CategoryEDUCATION	0.038	0.070	0.010	0.549	0.583	-0.098	0.175
CategoryENTERTAINMENT	-0.225	0.070	-0.055	-3.199	0.001	-0.362	-0.087
CategoryEVENTS	0.020	0.083	0.003	0.236	0.814	-0.143	0.183
CategoryFAMILY	-0.159	0.060	-0.128	-2.676	0.007	-0.276	-0.043
CategoryFINANCE	-0.203	0.064	-0.077	-3.193	0.001	-0.328	-0.078
CategoryFOOD_AND_DRINK	-0.176	0.072	-0.040	-2.443	0.015	-0.317	-0.035
CategoryGAME	-0.080	0.060	-0.051	-1.329	0.184	-0.198	0.038
CategoryHEALTH_AND_FITNESS	-0.082	0.064	-0.030	-1.277	0.202	-0.207	0.044
CategoryHOUSE_AND_HOME	-0.148	0.077	-0.028	-1.918	0.055	-0.300	0.003
CategoryLIBRARIES_AND_DEMO	-0.144	0.078	-0.026	-1.838	0.066	-0.297	0.010
CategoryLIFESTYLE	-0.231	0.063	-0.089	-3.636	0.000	-0.355	-0.106
CategoryMAPS_AND_NAVIGATION	-0.284	0.071	-0.066	-3.993	0.000	-0.424	-0.145
CategoryMEDICAL	-0.167	0.063	-0.070	-2.664	0.008	-0.290	-0.044
CategoryNEWS_AND_MAGAZINES	-0.205	0.065	-0.068	-3.153	0.002	-0.332	-0.078
CategoryPARENTING	-0.065	0.085	-0.010	-0.769	0.442	-0.231	0.101
CategoryPERSONALIZATION	-0.059	0.063	-0.023	-0.930	0.353	-0.183	0.065
CategoryPHOTOGRAPHY	-0.165	0.064	-0.059	-2.571	0.010	-0.290	-0.039
CategoryPRODUCTIVITY	-0.143	0.063	-0.058	-2.264	0.024	-0.266	-0.019
CategorySHOPPING	-0.095	0.066	-0.030	-1.456	0.145	-0.224	0.033
CategorySOCIAL	-0.119	0.065	-0.040	-1.839	0.066	-0.246	0.008
CategorySPORTS	-0.132	0.063	-0.051	-2.078	0.038	-0.256	-0.007
CategoryTOOLS	-0.286	0.061	-0.160	-4.701	0.000	-0.405	-0.167
CategoryTRAVEL_AND_LOCAL	-0.231	0.066	-0.073	-3.520	0.000	-0.359	-0.102
CategoryVIDEO_PLAYERS	-0.279	0.069	-0.073	-4.060	0.000	-0.413	-0.144
CategoryWEATHER	-0.114	0.078	-0.021	-1.458	0.145	-0.268	0.039
TypePaid	0.076	0.018	0.042	4.179	0.000	0.041	0.112
Size	0.000	0.000	-0.024	-2.507	0.012	0.000	0.000
Price	-0.001	0.000	-0.025	-2.604	0.009	-0.001	0.000
Reviews	0.000	0.000	0.064	6.577	0.000	0.000	0.000

[1] "No variables have been removed from the model."

- We can notice in backward stepwise regression no variable was eliminated in the final model with prediction metrics:

1) RMSE: 0.472    2) R-Squared: 0.032

- Now we will perform stepwise regression for both the directions.**

```
> ##Performing stepwise both direction regression
> ols_step_both_p(multiple2, details= TRUE)
Stepwise Selection Method
```

Candidate Terms:

1. Category
2. Type
3. Size
4. Price
5. Reviews

we are selecting variables based on p value...

Stepwise Selection: Step 1

+ Category

Model Summary			
R	0.162	RMSE	0.474
R-Squared	0.026	Coef. Var	11.301
Adj. R-Squared	0.023	MSE	0.225
Pred R-Squared	0.021	MAE	0.316

RMSE: Root Mean Square Error  
MSE: Mean Square Error  
MAE: Mean Absolute Error

ANOVA					
	Sum of Squares	DF	Mean Square	F	Sig.
Regression	65.218	32	2.038	9.078	0.0000
Residual	2426.061	10806	0.225		
Total	2491.278	10838			

Parameter Estimates							
model	Beta	Std. Error	Std. Beta	t	Sig.	lower	upper
(Intercept)	4.351	0.059		74.029	0.000	4.236	4.466
CategoryAUTO_AND_VEHICLES	-0.159	0.078	-0.029	-2.037	0.042	-0.312	-0.006
CategoryBEAUTY	-0.089	0.088	-0.013	-1.009	0.313	-0.260	0.083
CategoryBOOKS_AND_REFERENCE	-0.038	0.067	-0.012	-0.574	0.566	-0.169	0.092
CategoryBUSINESS	-0.203	0.063	-0.085	-3.225	0.001	-0.326	-0.079
CategoryCOMICS	-0.194	0.085	-0.030	-2.288	0.022	-0.360	-0.028
CategoryCOMMUNICATION	-0.186	0.064	-0.072	-2.927	0.003	-0.310	-0.061
CategoryDATING	-0.342	0.066	-0.104	-5.145	0.000	-0.472	-0.212
CategoryEDUCATION	0.037	0.070	0.009	0.530	0.596	-0.100	0.174
CategoryENTERTAINMENT	-0.225	0.070	-0.055	-3.189	0.001	-0.363	-0.087
CategoryEVENTS	0.015	0.083	0.002	0.178	0.859	-0.149	0.178
CategoryFAMILY	-0.158	0.060	-0.127	-2.639	0.008	-0.275	-0.041
CategoryFINANCE	-0.211	0.064	-0.079	-3.307	0.001	-0.336	-0.086
CategoryFOOD_AND_DRINK	-0.179	0.072	-0.040	-2.479	0.013	-0.321	-0.037
CategoryGAME	-0.068	0.060	-0.044	-1.125	0.260	-0.186	0.050
CategoryHEALTH_AND_FITNESS	-0.084	0.064	-0.030	-1.304	0.192	-0.209	0.042
CategoryHOUSE_AND_HOME	-0.153	0.077	-0.029	-1.975	0.048	-0.305	-0.001
CategoryLIBRARIES_AND_DEMO	-0.167	0.078	-0.031	-2.142	0.032	-0.320	-0.014
CategoryLIFESTYLE	-0.237	0.064	-0.091	-3.730	0.000	-0.362	-0.113
CategoryMAPS_AND_NAVIGATION	-0.285	0.071	-0.066	-3.995	0.000	-0.425	-0.145
CategoryMEDICAL	-0.159	0.063	-0.067	-2.533	0.011	-0.282	-0.036
CategoryNEWS_AND_MAGAZINES	-0.207	0.065	-0.069	-3.170	0.002	-0.334	-0.079
CategoryPARENTING	-0.067	0.085	-0.010	-0.795	0.427	-0.234	0.099
CategoryPERSONALIZATION	-0.048	0.063	-0.019	-0.756	0.450	-0.172	0.076
CategoryPHOTOGRAPHY	-0.158	0.064	-0.057	-2.464	0.014	-0.284	-0.032
CategoryPRODUCTIVITY	-0.141	0.063	-0.057	-2.239	0.025	-0.265	-0.018
CategorySHOPPING	-0.096	0.066	-0.031	-1.463	0.143	-0.225	0.033
CategorySOCIAL	-0.102	0.065	-0.035	-1.570	0.116	-0.229	0.025
CategorySPORTS	-0.131	0.064	-0.051	-2.065	0.039	-0.256	-0.007
CategoryTOOLS	-0.284	0.061	-0.158	-4.651	0.000	-0.403	-0.164
CategoryTRAVEL_AND_LOCAL	-0.230	0.066	-0.073	-3.501	0.000	-0.359	-0.101
CategoryVIDEO_PLAYERS	-0.275	0.069	-0.072	-4.001	0.000	-0.410	-0.140
CategoryWEATHER	-0.111	0.079	-0.020	-1.405	0.160	-0.265	0.044

# Final Model Output

## Model Summary

R	0.179	RMSE	0.472
R-Squared	0.032	Coef. var	11.269
Adj. R-Squared	0.029	MSE	0.223
Pred R-Squared	0.026	MAE	0.314

RMSE: Root Mean Square Error  
MSE: Mean Square Error  
MAE: Mean Absolute Error

## ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	80.129	36	2.226	9.972	0.0000
Residual	2411.149	10802	0.223		
Total	2491.278	10838			

## Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig.	lower	upper
(Intercept)	4.349	0.059		74.197	0.000	4.234	4.464
CategoryAUTO_AND_VEHICLES	-0.155	0.078	-0.029	-1.996	0.046	-0.308	-0.003
CategoryBEAUTY	-0.085	0.087	-0.012	-0.968	0.333	-0.256	0.087
CategoryBOOKS_AND_REFERENCE	-0.041	0.066	-0.012	-0.620	0.535	-0.171	0.089
CategoryBUSINESS	-0.199	0.063	-0.084	-3.185	0.001	-0.322	-0.077
CategoryCOMICS	-0.188	0.085	-0.029	-2.228	0.026	-0.354	-0.023
CategoryCOMMUNICATION	-0.205	0.063	-0.079	-3.232	0.001	-0.329	-0.081
CategoryDATING	-0.340	0.066	-0.103	-5.133	0.000	-0.470	-0.210
CategoryEDUCATION	0.038	0.070	0.010	0.549	0.583	-0.098	0.175
CategoryENTERTAINMENT	-0.225	0.070	-0.055	-3.199	0.001	-0.362	-0.087
CategoryEVENTS	0.020	0.083	0.003	0.236	0.814	-0.143	0.183
CategoryFAMILY	-0.159	0.060	-0.128	-2.676	0.007	-0.276	-0.043
CategoryFINANCE	-0.203	0.064	-0.077	-3.193	0.001	-0.328	-0.078
CategoryFOOD_AND_DRINK	-0.176	0.072	-0.040	-2.443	0.015	-0.317	-0.035
CategoryGAME	-0.080	0.060	-0.051	-1.329	0.184	-0.198	0.038
CategoryHEALTH_AND_FITNESS	-0.082	0.064	-0.030	-1.277	0.202	-0.207	0.044
CategoryHOUSE_AND_HOME	-0.148	0.077	-0.028	-1.918	0.055	-0.300	0.003
CategoryLIBRARIES_AND_DEMO	-0.144	0.078	-0.026	-1.838	0.066	-0.297	0.010
CategoryLIFESTYLE	-0.231	0.063	-0.089	-3.636	0.000	-0.355	-0.106
CategoryMAPS_AND_NAVIGATION	-0.284	0.071	-0.066	-3.993	0.000	-0.424	-0.145
CategoryMEDICAL	-0.167	0.063	-0.070	-2.664	0.008	-0.290	-0.044
CategoryNEWS_AND_MAGAZINES	-0.205	0.065	-0.068	-3.153	0.002	-0.332	-0.078
CategoryPARENTING	-0.065	0.085	-0.010	-0.769	0.442	-0.231	0.101
CategoryPERSONALIZATION	-0.059	0.063	-0.023	-0.930	0.353	-0.183	0.065
CategoryPHOTOGRAPHY	-0.165	0.064	-0.059	-2.571	0.010	-0.290	-0.039
CategoryPRODUCTIVITY	-0.143	0.063	-0.058	-2.264	0.024	-0.266	-0.019
CategorySHOPPING	-0.095	0.066	-0.030	-1.456	0.145	-0.224	0.033
CategorySOCIAL	-0.119	0.065	-0.040	-1.839	0.066	-0.246	0.008
CategorySPORTS	-0.132	0.063	-0.051	-2.078	0.038	-0.256	-0.007
CategoryTOOLS	-0.286	0.061	-0.160	-4.701	0.000	-0.405	-0.167
CategoryTRAVEL_AND_LOCAL	-0.231	0.066	-0.073	-3.520	0.000	-0.359	-0.102
CategoryVIDEO_PLAYERS	-0.279	0.069	-0.073	-4.060	0.000	-0.413	-0.144
CategoryWEATHER	-0.114	0.078	-0.021	-1.458	0.145	-0.268	0.039
Reviews	0.000	0.000	0.064	6.577	0.000	0.000	0.000
TypePaid	0.076	0.018	0.042	4.179	0.000	0.041	0.112
Price	-0.001	0.000	-0.025	-2.604	0.009	-0.001	0.000
Size	0.000	0.000	-0.024	-2.507	0.012	0.000	0.000

## Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Category	addition	0.026	0.023	33.8040	14603.0460	0.4738
2	Reviews	addition	0.030	0.027	-5.3480	14563.9289	0.4729
3	Type	addition	0.031	0.028	-15.8800	14553.3760	0.4727
4	Price	addition	0.032	0.028	-20.7130	14548.5259	0.4726
5	Size	addition	0.032	0.029	-25.0000	14544.2195	0.4725

- After executing we have found all the prediction errors which were in both direction model.

1) RMSE: 0.472    2) R-Squared: 0.032

- Now we will try to execute and do feature selection by stepwise forward AIC model.

```
> ols_step_forward_aic(multiple2, details= TRUE)
Forward Selection Method
```

Candidate Terms:

```
1 . Category
2 . Type
3 . Size
4 . Price
5 . Reviews
```

```
Step 0: AIC = 14826.57
Rating ~ 1
```

Variable	DF	AIC	Sum Sq	RSS	R-Sq	Adj. R-Sq
Category	1	14603.046	65.218	2426.061	0.026	0.023
Reviews	1	14778.848	11.403	2479.875	0.005	0.004
Type	1	14814.490	3.235	2488.043	0.001	0.001
Price	1	14824.204	1.004	2490.274	0.000	0.000
Size	1	14824.602	0.913	2490.366	0.000	0.000

+ Category

```
Step 1 : AIC = 14603.05
Rating ~ Category
```

Variable	DF	AIC	Sum Sq	RSS	R-Sq	Adj. R-Sq
Reviews	1	14563.929	9.186	2416.875	0.030	0.027
Type	1	14594.192	2.428	2423.632	0.027	0.024
Size	1	14599.888	1.154	2424.906	0.027	0.024
Price	1	14602.004	0.681	2425.380	0.026	0.023

+ Reviews

```
Step 2 : AIC = 14563.93
Rating ~ Category + Reviews
```

Variable	DF	AIC	Sum Sq	RSS	R-Sq	Adj. R-Sq
Type	1	14553.376	2.797	2414.077	0.031	0.028
Size	1	14560.575	1.193	2415.681	0.030	0.027
Price	1	14562.950	0.664	2416.211	0.030	0.027

Final Model output

#### Model Summary

R	0.179	RMSE	0.472
R-Squared	0.032	Coef. Var	11.269
Adj. R-Squared	0.029	MSE	0.223
Pred R-Squared	0.026	MAE	0.314

RMSE: Root Mean Square Error

MSE: Mean Square Error

MAE: Mean Absolute Error

ANOVA					
	Sum of Squares	DF	Mean Square	F	Sig.
Regression	80.129	36	2.226	9.972	0.0000
Residual	2411.149	10802	0.223		
Total	2491.278	10838			

Parameter Estimates							
model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	4.349	0.059		74.197	0.000	4.234	4.464
CategoryAUTO_AND_VEHICLES	-0.155	0.078	-0.029	-1.996	0.046	-0.308	-0.003
CategoryBEAUTY	-0.085	0.087	-0.012	-0.968	0.333	-0.256	0.087
CategoryBOOKS_AND_REFERENCE	-0.041	0.066	-0.012	-0.620	0.535	-0.171	0.089
CategoryBUSINESS	-0.199	0.063	-0.084	-3.185	0.001	-0.322	-0.077
CategoryCOMICS	-0.188	0.085	-0.029	-2.228	0.026	-0.354	-0.023
CategoryCOMMUNICATION	-0.205	0.063	-0.079	-3.232	0.001	-0.329	-0.081
CategoryDATING	-0.340	0.066	-0.103	-5.133	0.000	-0.470	-0.210
CategoryEDUCATION	0.038	0.070	0.010	0.549	0.583	-0.098	0.175
CategoryENTERTAINMENT	-0.225	0.070	-0.055	-3.199	0.001	-0.362	-0.087
CategoryEVENTS	0.020	0.083	0.003	0.236	0.814	-0.143	0.183
CategoryFAMILY	-0.159	0.060	-0.128	-2.676	0.007	-0.276	-0.043
CategoryFINANCE	-0.203	0.064	-0.077	-3.193	0.001	-0.328	-0.078
CategoryFOOD_AND_DRINK	-0.176	0.072	-0.040	-2.443	0.015	-0.317	-0.035
CategoryGAME	-0.080	0.060	-0.051	-1.329	0.184	-0.198	0.038
CategoryHEALTH_AND_FITNESS	-0.082	0.064	-0.030	-1.277	0.202	-0.207	0.044
CategoryHOUSE_AND_HOME	-0.148	0.077	-0.028	-1.918	0.055	-0.300	0.003
CategoryLIBRARIES_AND_DEMO	-0.144	0.078	-0.026	-1.838	0.066	-0.297	0.010
CategoryLIFESTYLE	-0.231	0.063	-0.089	-3.636	0.000	-0.355	-0.106
CategoryMAPS_AND_NAVIGATION	-0.284	0.071	-0.066	-3.993	0.000	-0.424	-0.145
CategoryMEDICAL	-0.167	0.063	-0.070	-2.664	0.008	-0.290	-0.044
CategoryNEWS_AND_MAGAZINES	-0.205	0.065	-0.068	-3.153	0.002	-0.332	-0.078
CategoryPARENTING	-0.065	0.085	-0.010	-0.769	0.442	-0.231	0.101
CategoryPERSONALIZATION	-0.059	0.063	-0.023	-0.930	0.353	-0.183	0.065
CategoryPHOTOGRAPHY	-0.165	0.064	-0.059	-2.571	0.010	-0.290	-0.039
CategoryPRODUCTIVITY	-0.143	0.063	-0.058	-2.264	0.024	-0.266	-0.019
CategorySHOPPING	-0.095	0.066	-0.030	-1.456	0.145	-0.224	0.033
CategorySOCIAL	-0.119	0.065	-0.040	-1.839	0.066	-0.246	0.008
CategorySPORTS	-0.132	0.063	-0.051	-2.078	0.038	-0.256	-0.007
CategoryTOOLS	-0.286	0.061	-0.160	-4.701	0.000	-0.405	-0.167
CategoryTRAVEL_AND_LOCAL	-0.231	0.066	-0.073	-3.520	0.000	-0.359	-0.102
CategoryVIDEO_PLAYERS	-0.279	0.069	-0.073	-4.060	0.000	-0.413	-0.144
CategoryWEATHER	-0.114	0.078	-0.021	-1.458	0.145	-0.268	0.039
Reviews	0.000	0.000	0.064	6.577	0.000	0.000	0.000
TypePaid	0.076	0.018	0.042	4.179	0.000	0.041	0.112
Price	-0.001	0.000	-0.025	-2.604	0.009	-0.001	0.000
Size	0.000	0.000	-0.024	-2.507	0.012	0.000	0.000

Selection Summary					
variable	AIC	Sum Sq	RSS	R-Sq	Adj. R-Sq
Category	14603.046	65.218	2426.061	0.02618	0.02329
Reviews	14563.929	74.403	2416.875	0.02987	0.02690
Type	14553.376	77.201	2414.077	0.03099	0.02794
Price	14548.526	78.726	2412.552	0.03160	0.02846
Size	14544.219	80.129	2411.149	0.03216	0.02894

- We have seen that in forward AIC all the variables have been selected and AIC is been reduced and prediction metrics:

1) RMSE: 0.472    2) R-Squared: 0.032



- *We will now perform backward AIC model for the linear model we have used.*

```
> ols_step_backward_aic(multiple2, details= TRUE)
Backward Elimination Method
```

Candidate Terms:

- 1 . Category
- 2 . Type
- 3 . Size
- 4 . Price
- 5 . Reviews

Step 0: AIC = 14544.22  
Rating ~ Category + Type + Size + Price + Reviews

Variable	DF	AIC	Sum Sq	RSS	R-Sq	Adj. R-Sq
Size	1	14548.526	1.403	2412.552	0.032	0.028
Price	1	14549.023	1.514	2412.663	0.032	0.028
Type	1	14559.728	3.898	2415.047	0.031	0.027
Reviews	1	14585.535	9.655	2420.804	0.028	0.025
Category	1	14753.723	61.615	2472.764	0.007	0.007

Variables Removed:

No more variables to be removed.

Variables Removed:

Final Model output

Model Summary			
R	0.179	RMSE	0.472
R-Squared	0.032	Coef. Var	11.269
Adj. R-Squared	0.029	MSE	0.223
Pred R-Squared	0.026	MAE	0.314

RMSE: Root Mean Square Error  
MSE: Mean Square Error  
MAE: Mean Absolute Error

ANOVA					
	Sum of Squares	DF	Mean Square	F	Sig.
Regression	80.129	36	2.226	9.972	0.0000
Residual	2411.149	10802	0.223		
Total	2491.278	10838			

Parameter Estimates								
model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper	
(Intercept)	4.349	0.059		74.197	0.000	4.234	4.464	
CategoryAUTO_AND_VEHICLES	-0.155	0.078	-0.029	-1.996	0.046	-0.308	-0.003	
CategoryBEAUTY	-0.085	0.087	-0.012	-0.968	0.333	-0.256	0.087	
CategoryBOOKS_AND_REFERENCE	-0.041	0.066	-0.012	-0.620	0.535	-0.171	0.089	
CategoryBUSINESS	-0.199	0.063	-0.084	-3.185	0.001	-0.322	-0.077	
CategoryCOMICS	-0.188	0.085	-0.029	-2.228	0.026	-0.354	-0.023	
CategoryCOMMUNICATION	-0.205	0.063	-0.079	-3.232	0.001	-0.329	-0.081	
CategoryDATING	-0.340	0.066	-0.103	-5.133	0.000	-0.470	-0.210	
CategoryEDUCATION	0.038	0.070	0.010	0.549	0.583	-0.098	0.175	
CategoryENTERTAINMENT	-0.225	0.070	-0.055	-3.199	0.001	-0.362	-0.087	
CategoryEVENTS	0.020	0.083	0.003	0.236	0.814	-0.143	0.183	
CategoryFAMILY	-0.159	0.060	-0.128	-2.676	0.007	-0.276	-0.043	
CategoryFINANCE	-0.203	0.064	-0.077	-3.193	0.001	-0.328	-0.078	
CategoryFOOD_AND_DRINK	-0.176	0.072	-0.040	-2.443	0.015	-0.317	-0.035	
CategoryGAME	-0.080	0.060	-0.051	-1.329	0.184	-0.198	0.038	
CategoryHEALTH_AND_FITNESS	-0.082	0.064	-0.030	-1.277	0.202	-0.207	0.044	
CategoryHOUSE_AND_HOME	-0.148	0.077	-0.028	-1.918	0.055	-0.300	0.003	
CategoryLIBRARIES_AND_DEMO	-0.144	0.078	-0.026	-1.838	0.066	-0.297	0.010	
CategoryLIFESTYLE	-0.231	0.063	-0.089	-3.636	0.000	-0.355	-0.106	
CategoryMAPS_AND_NAVIGATION	-0.284	0.071	-0.066	-3.993	0.000	-0.424	-0.145	
CategoryMEDICAL	-0.167	0.063	-0.070	-2.664	0.008	-0.290	-0.044	
CategoryNEWS_AND_MAGAZINES	-0.205	0.065	-0.068	-3.153	0.002	-0.332	-0.078	
CategoryPARENTING	-0.065	0.085	-0.010	-0.769	0.442	-0.231	0.101	
CategoryPERSONALIZATION	-0.059	0.063	-0.023	-0.930	0.353	-0.183	0.065	
CategoryPHOTOGRAPHY	-0.165	0.064	-0.059	-2.571	0.010	-0.290	-0.039	
CategoryPRODUCTIVITY	-0.143	0.063	-0.058	-2.264	0.024	-0.266	-0.019	
CategorySHOPPING	-0.095	0.066	-0.030	-1.456	0.145	-0.224	0.033	
CategorySOCIAL	-0.119	0.065	-0.040	-1.839	0.066	-0.246	0.008	
CategorySPORTS	-0.132	0.063	-0.051	-2.078	0.038	-0.256	-0.007	
CategoryTOOLS	-0.286	0.061	-0.160	-4.701	0.000	-0.405	-0.167	
CategoryTRAVEL_AND_LOCAL	-0.231	0.066	-0.073	-3.520	0.000	-0.359	-0.102	
CategoryVIDEO_PLAYERS	-0.279	0.069	-0.073	-4.060	0.000	-0.413	-0.144	
CategoryWEATHER	-0.114	0.078	-0.021	-1.458	0.145	-0.268	0.039	
TypePaid	0.076	0.018	0.042	4.179	0.000	0.041	0.112	
Size	0.000	0.000	-0.024	-2.507	0.012	0.000	0.000	
Price	-0.001	0.000	-0.025	-2.604	0.009	-0.001	0.000	
Reviews	0.000	0.000	0.064	6.577	0.000	0.000	0.000	

[1] "No variables have been removed from the model."

- Again, we have performed backward AIC model with final prediction error's:

1) RMSE: 0.472    2) R-Squared: 0.029

- **Finally, we are performing N-Fold Cross validation for the final prediction of the rating of application.**

```
> model <- train(Rating~Category+Price+Size+Type+Reviews,data=dataset, trControl=train_control, method="lm",na.action=na.pass)
> summary(model)
```

```
Call:
lm(formula = .outcome ~ ., data = dat)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-3.2679 -0.1423  0.0568  0.2569  1.0002
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.349e+00  5.861e-02  74.197 < 2e-16 ***
CategoryAUTO_AND_VEHICLES -1.554e-01  7.785e-02  -1.996  0.045960 *
CategoryBEAUTY -8.461e-02  8.744e-02  -0.968  0.333260
CategoryBOOKS_AND_REFERENCE -4.114e-02  6.636e-02  -0.620  0.535320
CategoryBUSINESS -1.994e-01  6.261e-02  -3.185  0.001452 **
CategoryCOMICS -1.884e-01  8.459e-02  -2.228  0.025923 *
CategoryCOMMUNICATION -2.051e-01  6.344e-02  -3.232  0.001231 **
CategoryDATING -3.400e-01  6.624e-02  -5.133  2.91e-07 ***
CategoryEDUCATION  3.830e-02  6.976e-02   0.549  0.582987
CategoryENTERTAINMENT -2.247e-01  7.024e-02  -3.199  0.001382 **
CategoryEVENTS  1.963e-02  8.320e-02   0.236  0.813529
CategoryFAMILY -1.594e-01  5.958e-02  -2.676  0.007459 **
CategoryFINANCE -2.032e-01  6.364e-02  -3.193  0.001414 **
CategoryFOOD_AND_DRINK -1.760e-01  7.206e-02  -2.443  0.014591 *
CategoryGAME -8.016e-02  6.030e-02  -1.329  0.183795
CategoryHEALTH_AND_FITNESS -8.165e-02  6.395e-02  -1.277  0.201749
CategoryHOUSE_AND_HOME -1.482e-01  7.728e-02  -1.918  0.055177 .
CategoryLIBRARIES_AND_DEMO -1.439e-01  7.829e-02  -1.838  0.066093 .
CategoryLIFESTYLE -2.306e-01  6.342e-02  -3.636  0.000278 ***
CategoryMAPS_AND_NAVIGATION -2.841e-01  7.116e-02  -3.993  6.57e-05 ***
CategoryMEDICAL -1.670e-01  6.268e-02  -2.664  0.007722 **
CategoryNEWS_AND_MAGAZINES -2.049e-01  6.499e-02  -3.153  0.001621 **
CategoryPARENTING -6.505e-02  8.458e-02  -0.769  0.441878
CategoryPERSONALIZATION -5.888e-02  6.335e-02  -0.930  0.352648
CategoryPHOTOGRAPHY -1.647e-01  6.404e-02  -2.571  0.010146 *
CategoryPRODUCTIVITY -1.425e-01  6.294e-02  -2.264  0.023595 *
CategorySHOPPING -9.540e-02  6.553e-02  -1.456  0.145496
CategorySOCIAL -1.192e-01  6.482e-02  -1.839  0.065877 .
CategorySPORTS -1.317e-01  6.337e-02  -2.078  0.037719 *
CategoryTOOLS -2.861e-01  6.085e-02  -4.701  2.62e-06 ***
CategoryTRAVEL_AND_LOCAL -2.308e-01  6.557e-02  -3.520  0.000434 ***
CategoryVIDEO_PLAYERS -2.786e-01  6.864e-02  -4.060  4.95e-05 ***
CategoryWEATHER -1.144e-01  7.847e-02  -1.458  0.144849
Price -7.664e-04  2.943e-04  -2.604  0.009221 **
Size -1.271e-04  5.069e-05  -2.507  0.012179 *
TypePaid  7.634e-02  1.827e-02  4.179  2.95e-05 ***
Reviews  1.041e-08  1.583e-09  6.577  5.03e-11 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.4725 on 10802 degrees of freedom
Multiple R-squared:  0.03216, Adjusted R-squared:  0.02894
F-statistic: 9.972 on 36 and 10802 DF, p-value: < 2.2e-16
```

```
> model
```

```
Linear Regression
```

```
10839 samples
  5 predictor
```

```
No pre-processing
```

```
Resampling: Cross-validated (10 fold)
```

```
Summary of sample sizes: 9755, 9756, 9754, 9754, 9755, 9754, ...
```

```
Resampling results:
```

```
RMSE      Rsquared    MAE
0.4724772  0.02748946  0.3155124
```

```
Tuning parameter 'intercept' was held constant at a value of TRUE
```

- We can notice that the above output is ready for the prediction model with prediction matrices:  
1) RMSE: 0.4724    2) R-Squared: 0.0274

### 5.3. Findings

Finally, for the prediction of the Rating of the app we have used N- Fold cross validation using linear model with the selected variables using the feature selection of the variables.

```
> #Final N-Fold Cross validation model
> set.seed(100)
> model_Final <- train(Rating~Category+Price+Size,data=data2, trControl=train_control, method="lm",na.action=na.pass)
> summary(model_Final)

Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
    Min       1Q   Median       3Q      Max
-3.2198 -0.1070  0.0270  0.2955  0.9009

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.230e+00  1.095e-02  386.416 < 2e-16 ***
Category     -1.866e-03  5.510e-04  -3.386 0.000711 ***
Price        -6.064e-04  2.886e-04  -2.101 0.035624 *
Size         -9.915e-05  5.062e-05  -1.959 0.050168 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4791 on 10835 degrees of freedom
Multiple R-squared:  0.001814, Adjusted R-squared:  0.001538
F-statistic: 6.564 on 3 and 10835 DF, p-value: 0.0001983

> model_Final
Linear Regression

10839 samples
   3 predictor

No pre-processing
Resampling: Cross-validated (10 fold)
Summary of sample sizes: 9754, 9756, 9753, 9756, 9754, 9757, ...
Resampling results:

    RMSE      Rsquared    MAE
0.4788429  0.002691789  0.3144538

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Although we have all the independent variables present but we have finally worked on only three independent variables for the project that are Category, Price & Size. We have not used the variables such as Type and Reviews as their value in the above output as they are very small.

Finally, our prediction model is ready with prediction metrics to be: 1). RMSE = 0.4788    2) R-Squared = 0.0026

### Comparing all the models

Regression Model	RMSE	R-Squared
Forward Stepwise	0.472	0.032
Backward Stepwise	0.472	0.032
Both Stepwise	0.472	0.032
Forward AIC	0.472	0.032
Backward AIC	0.472	0.029
N-Fold Model	0.472	0.027
N-Fold Final	0.478	0.002

## 6. Conclusions and Future Work

### 6.1. Conclusions

For an application developer, it is beneficial to know about how well the app is going to perform when it is launched on the application like google play store taking into consideration various variables related to the application. We have presented a small project based on the prediction of the app before it's launches, we have used rating as the dependent variable and price & category as the independent variable for the prediction of the rating.

Firstly, we have preprocessed & cleaned the dataset taken from Kaggle having more than 10,000 rows and 13 columns. Then performed the feature selection process using Multiple Linear Regression, ANOVA Hypothesis and at last used N-Fold Cross Validation to predict the rating.

We have been able to successfully predict the rating of the applications. Moreover, using ANOVA we determined that the mean of all the categories are based on the ratings are not same, mean value of videoplayer category is the highest and mean of the Dating category is the lowest. With N-Fold cross validation prediction model we have used three variables for final output with prediction matrices of

1) RMSE = 0.478    2) R-Squared = 0.002

We have also designed the UI for user friendly approach of how our final model is predicting the rating without running the code again and again.

### 6.2. Limitations

Although we have been successful to complete our project, but we were limited to some features which could have been more helpful to get more accurate results. Limitations we faced are:

- 1) The dataset was small, if the dataset could have been with more entries the results would have been more précised.
- 2) We could only use three independent variables to predict the model, but we could have taken more variables to predict but due to lesser data we were restricted to three only.
- 3) For betterment of our project we could have grouped the Size column to various subparts as  
Low: 1Kb – 10MB, Medium: 11MB – 200 MB, High: 200MB – 1GB, Very High: > 1GB.  
Doing this we would have reduced the different entries in Size column and have been able to predict in a better way.

### 6.3. Potential Improvements or Future Work

Future improvement could be executed by adding more entries in dataset, used all the independent variables.

Using Classification models, logistic regression and multiple linear regression to determine which is the best model for the prediction of the ratings.