Indian School of Business
Certification in Business Analytics

Statistical Analysis & Modelling
Project Report

Linear Regression on
Facebook Page - Performance Metrics

Submitted By:
Rajiv Vanipenta
(7th April; 2017)

Table of Contents

Executive Summary:

Social Media advertising is the new boom. This research project provides an approach through linear modelling for predicting the performance of the brands Facebook page and hence the type of posts it publishes. The model was developed through the KPI's provided by the Facebook Insights for the brands page. The final equation predicts the "consumption rate" of a given post with 87.2% accuracy and with an error rate of 0.354. We found that the consumption of a post depends on type of post, category of post, total impressions, reach, and engagement of the post. We found that the if a post is paid or not it is not going to impact our consumption of post.

Introduction:

Regression Analysis is a simple statistical method for investigating functional relationship among variables. The relationship is expressed in terms of an equation or a model connecting the response or dependant variable and one or more explanatory or predictor variables. (Chatterjee, 2014) . A simple regression equation can be shown as under.

$$Y = \beta 0 + \beta 1.X1 + \beta 2.X2 + \cdots + \beta n.Xn + \varepsilon \quad \text{Eqn.(1)}$$

The regression process, starts by defining a problem statement or a simple business objective in hand. Once a problem is identified as regression problem, we decide on our variables of interest. Post that, we need to collect the relevant data that helps in meeting our goals defined above. We need to ensure the data is clean and reliable; else we need to take necessary steps in order to bring the data into a good shape. Next, is to analyse the behaviour of different variables and check if there are any abnormalities. Further, we go on building the basic model and validate our assumptions that are necessary to carry on a linear regression. The assumptions and limitations are the key factors that are needed to be kept in mind while building a model. We need to improve the model which we have built, such that it meets all the assumptions we considered. After fitting different model's that meet our assumptions we need to evaluate the models to choose the best possible model that fits the data well. Finally, the model will be ready to use and can be used for intended purpose.

In regression, we predict an output based on relationship between different variables. The output may be a quantitative variable i.e. a numeric output or a categorical variable a binary output. The type of output we are going to predict will determine the type of regression to be used. If our output is a quantitative variable we use a Simple / Multiple Linear regression and if output is a qualitative variable we use logistic regression.

Throughout the scope of this document, we use different terminologies with the same meaning. Few of them are listed here: Dependant variable, output variable, Determining variable, response variable, Y variable are all the same. Independent variables, Input variables, regressors, explanatory variables, predictor variables, X's means the same. Terms like residuals, errors, unexplained part, Noise, $\acute{\varepsilon}$ will are of same meaning. Forecasted values, predicted values are the values that we get out of our regression output after feeding in the input variables. The regression coefficients, β values estimates, regression parameters are the unknown constants that need to be determined.

Problem Statement:

A renowned cosmetics brand is spending a good amount of time and efforts in various marketing platforms in order to improve its brand value which in turn hopefully be converted to sales.

The Management especially is interested in knowing about the social media promotions. The company has been posting various content on its Facebook (FB) page and wants to know how people are reacting to the marketing campaign that are taking place on that page. Company wants to analyse the efficiency of their Facebook page by using the performance metrics provided by Facebook Insights (Joss, 2012).

This should help the company to determine the factors affecting the reachability and improve the content of the page, so that the page is more effective in reaching to the customers. Thereby improving the brands visibility, this might in turn increase the profits in the due course.

Understanding the Domain:

Facebook page is a place where a particular community / network / brand can interact with customers / members. The main objective of opening the page will be to connect to the members of page effectively and make them engaged with the posts by knowing their interests.

FB Insights (Joss, 2012) provides various parameters to the owner of the page to calculate the performance of the page and there by actions that can improve the page efficiency. On a higher level few of the page metrics provided by FB (an excel sheet of metrics as highlighted in Figure 1) to the owner of the page are: Impressions, Reach, Engaged Users, Consumers, Consumptions, Interactions.

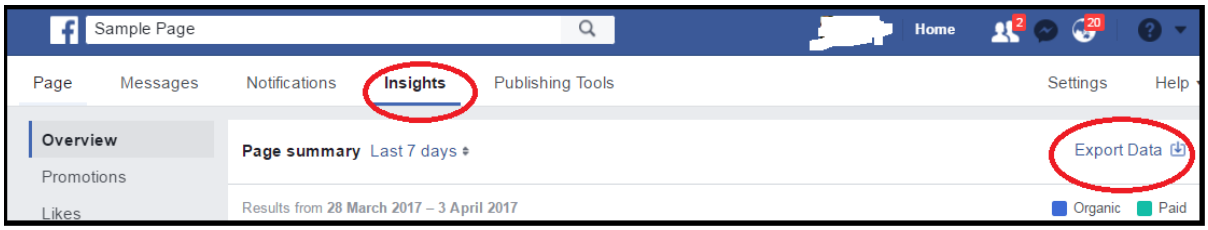


Figure 1: A Facebook page showing Insights tab and Export Data Option to the owner of page

Below is the small explanation of what the various FB metric's mean.

Impressions: After a post is posted on our page, it is displayed on the time line of most of the members of our page. This again depends on the previous level of engagement of the user on the page. If a user who is a member of the page, but not active on the page, might have the post displayed at the bottom of his/her time line. The count of total number of displays, of the post on all the users' timelines is called impressions.

Reach: It is not necessary that if a particular post is displayed on a user's time line, the user scrolls down in the time line and sees the post. (Our assumption here is that: FB might have an algorithm that will decide on, in which order the posts are displayed on a user time line is based of personal interests). Suppose,, the user has viewed the post on his time line. This means that the post has reached the user. Else it would just remain as "Impression" and is not counted as reach to the user.

There is a serious confusion between the metric "Engaged users" and the metric "Consumers".

Before going into these, we need to define what a "story" is, what an "engagement" of a post and what "consumption" is.

Story: A story is something that is displayed on users' personal time line. That is, if a user liked, commented, shared any post, friends of that user will be able to see that this user has liked or commented or shared this post. This is termed as a story.

Engagement: When any user clicks anywhere on your post that creates a story. This is also counted for De-storying (means un-liking, un-commenting actions).

Consumptions: Whenever a user performs any clicks on your post, which creates a story along with some other clicks like, playing a video, viewing a photograph, clicking on the link provided, this comes under consumptions.

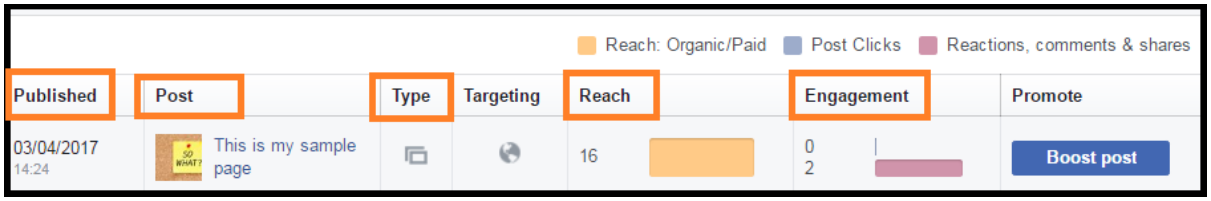


Figure 2: Facebook page metric dashboard showing different KPI's of the page

So, above two are explained here with an example. Say, I have posted picture as a post, and a user named Bob like's my post. Here Bob has created a story. All friends of the bob's can see that Bob has liked my post. Hence, Bob has engaged with my post. Another user say, Jack has just viewed the image that I posted by clicking on the image, here Jack did not create any story. None of Jack's friends will know that he has clicked the picture on my post. Hence jack has just consumed my post and this action is counted under consumption.

Consumers: These are the people who have consumed my post (DonKor, 2013).

Engaged Users: The users who have been engaged with my post (DonKor, 2013)

Interactions: This is the sum of total Likes, Shares, Comments received for the post. (DonKor, 2013)

Hope all the confusing terms / metrics which are provided by in the FB insights excel file, are clear after reading this.

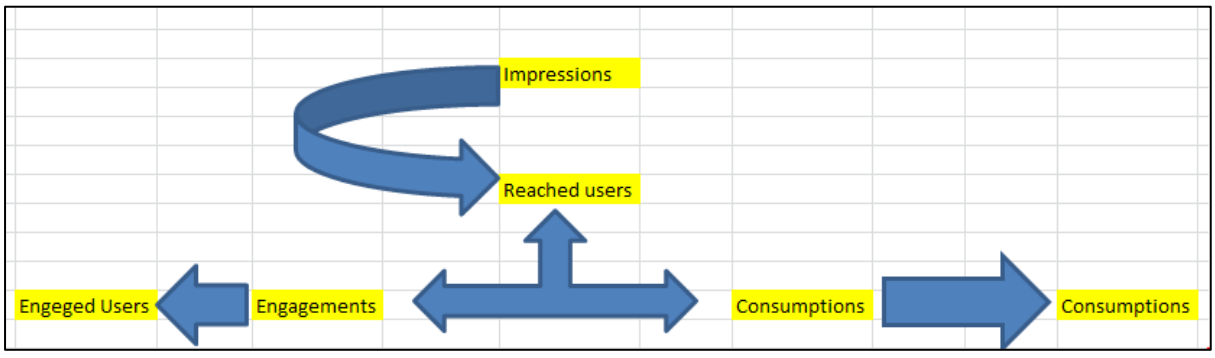


Figure 3: The Simple process flow diagram of the KPI's

If a particular type of post say, posting an image has generated a lot of engagement or consumed more by your users in the past, next time when such type of post from your page is posted, that will have much higher visibility (as per the FB post rank algorithm) than other posts posted by your competitor pages on a members timeline.

Data Collection and Preparation:

We are given a dataset that contains various metrics of the cosmetic brand's Facebook page. The dataset is the composition of all the posts that are posted on the brand's Facebook page for the year of 2014. The given Dataset consists of 500 rows (posts) and 19 columns (metrics). In these 19 different parameters for all the related posts, there is a composition of both categorical and numerical data. A brief description of the 19 parameters and the type of data they contain can be found below:

1. *Page Total Likes:* Total Number of people liked your page (Numeric Data).
2. *Type:* Defines if the type of post is a (Web) Link, Photo, Status or a Video (Categorical Data).
3. *Category:* Defines to which category a post belongs to: Action, Product, and Inspiration (Categorical Data).
4. *Paid:* Defines if the post is a Paid post (Categorical Data).
5. *Post Hour:* Which hour of the day was the post published (Categorical Data).
6. *Post Weekday:* Which Day of the week was the post published (Categorical Data)
7. *Post Month:* Which Month was the post published (Categorical Data)
8. *Lifetime post total impressions:* Impressions are the number of times a post from your Page is displayed (Numeric Data).
9. *Lifetime post impressions by people who have liked your page:* Impressions are the number of times a post from your Page is displayed on the people who likes your page (Numeric Data).
10. *Lifetime Post total Reach:* The total number of people who saw the post in its lifetime (Numerical Data).
11. *Lifetime post reaches by people who like your page:* The total number of people who saw the post in its lifetime and also like your page (Numerical Data).
12. *Lifetime engaged users:* Engaged Users are people who clicked anywhere in your posts that generates a story (Numeric Data).
13. *Lifetime people who have liked your page and engaged with your post:* The people who engaged with the post which generates a story and also like the page (Numeric Data).
14. *Lifetime post consumptions:* Consumptions measure any click on Post content, whether it generates a Story or not (Numeric Data).
15. *Lifetime post consumers:* Consumers are the total number of people who have consumed your post(Numeric Data).
16. *Comments:* Total Comments for the post (Numeric Data).
17. *Likes:* Total Number of Likes for the Post (Numeric Data).
18. *Shares:* Total Number of people who shared the Post (Numeric Data).
19. *Total Interactions:* Sum of Total Likes, Comments, Shares (Numeric Data).

This is the raw data which are available to us. We convert them to the metrics in a meaningful manner that will be aligning with our business KPI's.

The most important KPI's are the, reach rate given by:

$$\text{Reach Rate} = \frac{\text{Total reach}}{\text{Total Impressions}} * 100$$

$$\text{Reach rate by liked users} = \frac{\text{Reach to users who liked our page}}{\text{Impressions to users who liked our page}} * 100$$

The engagement rate KPI's are given by:

$$\text{Engagement rate} = \frac{\text{Engaged Users}}{\text{Total Reach}} * 100$$

$$\text{Engagement rate By Liked Users} = \frac{\text{Engaged Users who liked our page}}{\text{Reach to the people who liked our page}} * 100$$

Consumer rate is given by:

$$\text{Consumer Rate} = \frac{\text{Total Consumer}}{\text{Total Reach}} * 100$$

We can calculate the Average consumptions as:

$$\text{Consumption rate} = \frac{\text{Total Consumptions}}{\text{Total Reach}} * 100$$

$$\text{Avg. Consumptions} = \frac{\text{Total Consumptions}}{\text{Total Consumers}}$$

Total impression's which are converted to 100's scale to make more sense.

$$\text{Impressions in 100} = \text{Total Impression's} / 100$$

$$\text{Impression rate by liked} = \frac{\text{Impressions for liked users}}{\text{Total Page likes}} * 100$$

Using these parameters, we need to decide on which KPI we need to model in terms of others. As per our business goal we need to determine the factors which give us more consumption rate for our posts. More consumption's of a post means the more the connection to customers. So, here **'Consumption Rate'** will be our Determining variable which we need to predict in terms of others.
We want to analyse the effect of 'consumption rate' of a post based on: Type, Category, Paid, Impressions, Impressions rate by liked users, reach of the post, reach to the people who liked our page, engaged users rate, engaged users who liked our page, consumer rate, interactions.

Loading the dataset in R:

The Data file (facebook.csv) which has the raw data is loaded in to the Git repository and is read into our R program. The URL for the file: https://raw.githubusercontent.com/Rajiv2806/SA-2-Mini-Project-1-Facebook-Page-Sales/master/Facebook.csv

```
library(RCurl)
facebook_page <- read.table(text = getURL("https://raw.githubusercontent.com/Rajiv2806/SA-2-Mini-Project-1-Facebook-Page-Sales/master/Facebook.csv"),header = T,sep = ",")
```

After loading the raw data and **Renaming columns:** and including the calculated fields, our final data set will look as below.

Renaming columns:

As we can see that few columns are having very long names by the default way from which they are read in. We shorten the field names by renaming them for coding comfort. The new column names will be used in the place of actual column names for referencing. But the Interpretation will remain the same. Below are the renamed Columns.

Actual Column Name	Renamed Column Name
Page.total.likes	Page_Likes
Post.Month	Month
Post.Weekday	Weekday
Post.Hour	Hour
Lifetime.Post.Total.Reach	Reach
Lifetime.Post.reach.by.people.who.like.your.Page	Reach_ByLiked
Lifetime.Post.Total.Impressions	Impressions
Lifetime.Post.Impressions.by.people.who.have.liked.your.Page	Impressions_ByLiked
Lifetime.Engaged.Users	Engaged
Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post	Engaged_ByLiked
Lifetime.Post.Consumers	Consumers
Lifetime.Post.Consumptions	Consumptions
Total_Interactions	Interactions

Table 1: Original Columns and Renamed Columns

Code for renaming columns and re-arranging the order of columns:

```
names(facebook_page) <- c("Page_Likes","Type","Category","Month","Weekday","Hour","Paid"
                          ,"Reach","Impressions","Engaged","Consumers","Consumptions"
                          ,"Impressions_ByLiked","Reach_ByLiked","Engaged_ByLiked"
                          ,"Comment","Like","Share","Interactions")
facebook_page <- facebook_page[,c(19,2,3,7,4,5,6,8,14,9,13,10,15,11,12,16,17,18,1)]
```

Adding calculated fields to data set:

Below is the code converting the raw fields into our KPI's. We have removed all the other variables from our original data set and are retaining only our KPI fields which will help in our modelling. Our data will be reduced to 500*12 dimensions now.

```
facebook_page$Impressions_in_100 <- facebook_page$Impressions/100
facebook_page$ImpressionsRate_Liked <- (facebook_page$Impressions_Liked/facebook_page$Page_Likes)*100
facebook_page$ReachRate <- (facebook_page$Reach/facebook_page$Impressions)*100
facebook_page$ReachRate_Liked <- (facebook_page$Reach_Liked/facebook_page$Impressions_Liked)*100
facebook_page$EngagedRate <- (facebook_page$Engaged/facebook_page$Reach)*100
facebook_page$EngagedRate_Liked <- (facebook_page$Engaged_Liked/facebook_page$Reach_Liked)*100
facebook_page$ConsumerRate <- (facebook_page$Consumers/facebook_page$Engaged)*100
facebook_page$consumptionRate <- (facebook_page$Consumptions/facebook_page$Reach) * 100
facebook_page <- facebook_page[,c(27,2,3,4,20,21,22,23,24,25,26,1)]
```

Exploratory Data Analysis:

Exploratory data analysis (EDA) is the first step in regression analysis, before we proceed on to modelling. EDA is an important approach to analyse the data by summarising them and looking into the main characteristics. EDA tells us the structure of the dataset, data types of variables, Missing values, Outlier detection. We can take a call at this stage on how to handle any kind of abnormalities present in our data. The data structure we have at this stage will look as below:

```
sapply(facebook_page,class)
```

```
##       consumptionRate                  Type              Category
##             "numeric"              "factor"              "factor"
##                  Paid    Impressions_in_100 ImpressionsRate_Liked
##              "factor"             "numeric"             "numeric"
##             ReachRate       ReachRate_Liked           EngagedRate
##             "numeric"             "numeric"             "numeric"
##     EngagedRate_Liked          ConsumerRate          Interactions
##             "numeric"             "numeric"             "integer"
```

Univariate characteristics can be seen through summarising the data, looking into the measures of central tendency, by plotting the histograms, boxplots. Bivariate characteristics can be known through the correlation coefficients, scatter plot diagrams.

Checking the datatypes:

After looking into each variable and its data type we can observe interesting things here. The variables like 'Category', 'Paid' which we considered as categorical variables are taken as numeric / integer datatypes. We need to handle this situation by converting them as factors variables.

Before converting these variables into factors, we can take a call to convert 'Paid' posts to Yes or No.

```
facebook_page$Paid[facebook_page$Paid == 1] = "Yes"
facebook_page$Paid[facebook_page$Paid == 0] = "No"
facebook_page$Paid <- as.factor(facebook_page$Paid)

facebook_page$Category[facebook_page$Category == 1] = "Action"
facebook_page$Category[facebook_page$Category == 2] = "Product"
facebook_page$Category[facebook_page$Category == 3] = "Inspiration"
facebook_page$Category <- as.factor(facebook_page$Category)
```

The levels of each of the factor variables assigned can be seen in the below tables:

Converting 'Paid' Post to factor:

Values	Levels
1	Yes
0	No

Converting the 'Category' of post to factor:

Values	Levels
1	Action
2	Product
3	Inspiration

Finally, if we recheck our data set, we can find that all the columns are assigned correct data types.

Null Values:

Null values / Missing values, do not contain any data. Handling Null values is important, because they may lead to computational errors and hence might lead to wrong interpretation of results.
In our given dataset, there is one Missing value in 'Paid' post. We impute it using the most frequent occurring (mode) type of paid post.

This is the first assumptions we made. We don't have any missing values in our dataset now.

```
facebook_page$Paid[is.na(facebook_page$Paid)] = "No"
sum(is.na(facebook_page))
```

```
## [1] 0
```

Summary statistics:

Measures of central tendency like the mean, median, Quartiles, Minimum, maximum can be determined for each parameter. They are shown below.

```
summary(facebook_page)
```

```
##  consumptionRate        Type            Category    Paid
##  Min.   :  0.4878   Link  : 22   Action     :215   No :361
##  1st Qu.:  6.7254   Photo :426   Inspiration:155   Yes:139
##  Median : 16.1678   Status: 45   Product    :130
##  Mean   : 20.9524   Video :  7
##  3rd Qu.: 21.6428
##  Max.   :350.4202
##  Impressions_in_100 ImpressionsRate_Liked   ReachRate
##  Min.   :    5.70   Min.   :   0.5553     Min.   :  4.47
##  1st Qu.:   56.95   1st Qu.:   3.1164     1st Qu.: 53.32
##  Median :   90.51   Median :   5.6189     Median : 56.78
##  Mean   :  295.86   Mean   :  14.5853     Mean   : 60.46
##  3rd Qu.:  220.85   3rd Qu.:  12.0230     3rd Qu.: 60.13
##  Max.   :11102.82   Max.   :1064.5075     Max.   :790.63
##  ReachRate_Liked   EngagedRate      EngagedRate_Liked  ConsumerRate
##  Min.   : 4.366   Min.   : 0.4878   Min.   : 0.8929   Min.   : 35.73
##  1st Qu.:51.452   1st Qu.: 6.4101   1st Qu.: 7.6525   1st Qu.: 81.46
##  Median :55.097   Median :12.1717   Median :12.3771   Median : 90.48
##  Mean   :53.984   Mean   :12.4339   Mean   :12.7234   Mean   : 86.46
##  3rd Qu.:58.012   3rd Qu.:16.2629   3rd Qu.:16.1054   3rd Qu.: 95.33
##  Max.   :73.220   Max.   :60.0840   Max.   :49.8580   Max.   :100.00
##   Interactions
##  Min.   :   0.0
##  1st Qu.:  71.0
##  Median : 123.5
##  Mean   : 212.1
##  3rd Qu.: 228.5
##  Max.   :6334.0
```

One of the interesting observations here is, the fields which are in terms of rate or percentages are having values of more than 100. Since we know this is impossible in reality, we need to take corrective action's here to verify the data set and remove or take fill in correct values for those observations.

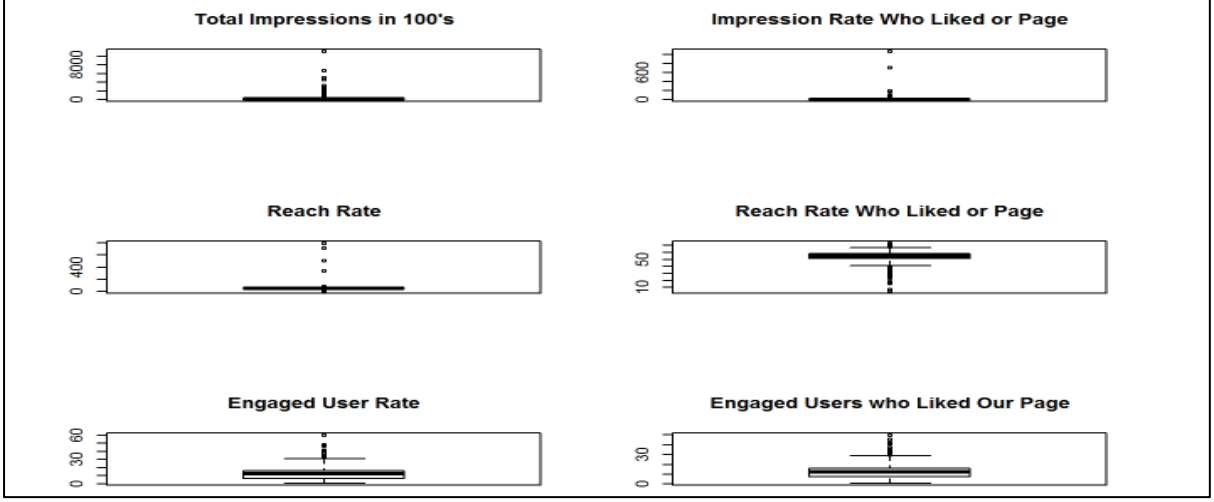
But, since the data is coming from the FB insights automatically, we do not know what the appropriate mechanism to handle this situation. So, we continue with the corrupted values and build our model.

Univariate Analysis:

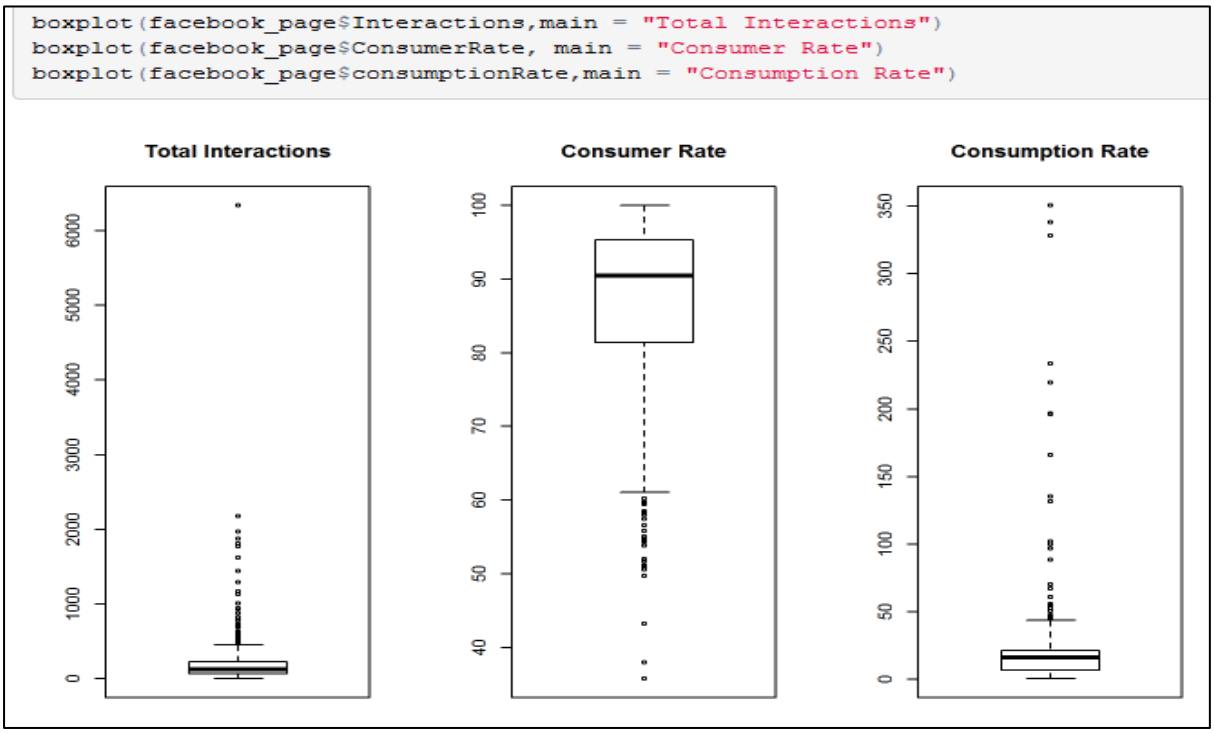
We need to check the behaviour of the individual variables. By checking the box plots we can see how our variables are behaving. We will be able to say if they are normally distributed or they are skewed

to one side. We can detect any possible outliers that might be present in each variable and this identification can help us to take any corrective measures or apply appropriate transformations on these observations, so that our final regression output is not influenced by these observations.

```
par(mfrow=c(3,2))
boxplot(facebook_page$Impressions_in_100,main = "Total Impressions in 100's")
boxplot(facebook_page$ImpressionsRate_Liked,main = "Impression Rate Who Liked or Page")
boxplot(facebook_page$ReachRate,main = "Reach Rate")
boxplot(facebook_page$ReachRate_Liked,main = "Reach Rate Who Liked or Page")
boxplot(facebook_page$EngagedRate,main = "Engaged User Rate")
boxplot(facebook_page$EngagedRate_Liked,main = "Engaged Users who Liked Our Page")
```




```
boxplot(facebook_page$Interactions,main = "Total Interactions")
boxplot(facebook_page$ConsumerRate, main = "Consumer Rate")
boxplot(facebook_page$consumptionRate,main = "Consumption Rate")
```

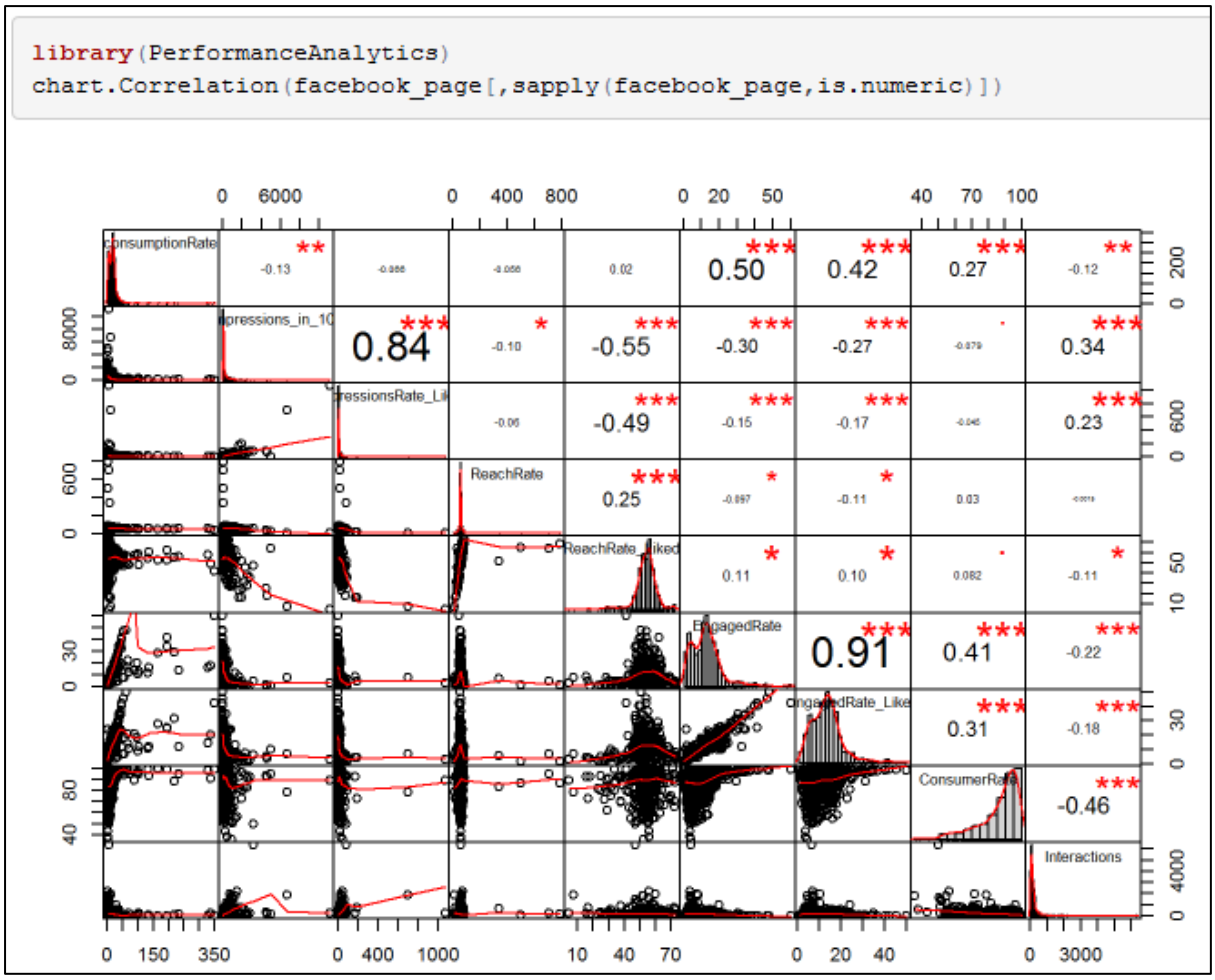


From the above figures, we can see the visually the problem we discussed in the previous section, the outlier variables.

Bivariate Analysis:

In order to perform the regression, we start with the assumption that our independent variables are independent of each other and are not correlated to one another. The correlation coefficients and the Scatter plots between a pair of variables will say us how much they are related with each other.

```
library(PerformanceAnalytics)
chart.Correlation(facebook_page[,sapply(facebook_page,is.numeric)])
```



From the above figure looking at the correlation coefficients on the above half of the table, we can see the a correlation between '*Engaged User rate*' and '*Engaged user rate of liked users*' with a degree of 0.91. We can also observe significant level of correlation for 'Total Impressions' and 'Impression Rate *by liked users*' with a degree od 0.84. Both these correlation pairs make sense as they are related as per the definitions.

The diagonal elements of the above plot say that few of the variables are skewed to the right. This is because of the reason that few of the posts received high attention of the users.

The scatter plots in the lower half of the above figure again say the same story. Almost all our observations are grouped together only a single or couple of data points are pulling the skewness on to one side.

This exercise of the bivariate data, helped us see, how a single or couple of observations are making an impact on the whole. These observations can be neutralized by treating with appropriate methods which might enable us to see dramatic changes while designing our model and also applying

transformations on these variables might help in some cases.

Assumptions:

Before performing the Linear Regression on the given dataset we assume the following:

- There is a linear relationship between our response variable and our predictor variables. This is also called the Linearity assumption.
- The residuals generated are normally distributed $\epsilon \sim N(0, \sigma^2)$. This is Normality Assumption.
- The residuals generated are not influenced by any of our independent variables.
- The predictors should not have and significant correlation between them. This is also called the Multi Collinearity assumption.
- The residuals generated are independent of each other and should not have any pattern or auto-correlation between them. This is called homoscedasticity.

Model Building:

Before we proceed on to build our first model, we will explore our findings in the EDA in relation with our Model Assumptions discussed above. We can see that our data has a problem of multi collinearity among few pairs of our regressors. We can also see the effect of few of observations which are exercising more influence on our model.

Partitioning data:

The dataset will be divided into training and validations sets. We build our model on the training data set and test our best model fit on the validation data set. After choosing the best model, we apply that model on our complete data to use it for intended purpose.

```
rownumbers <- sample(1:nrow(facebook_page),size = 0.8*nrow(facebook_page))
facebook_train <- facebook_page[rownumbers,]
facebook_Validation <- facebook_page[-rownumbers,]
```

Our training data set has 400 rows and validation set has 100 rows.

Model 0:

We build our first model by ignoring all the assumptions and building a simple model with all regressors we need to check the influence of all regressors on our output variable.

The summary of the Model-0 fit is shown below:

```
Model0 <- lm(consumptionRate~
                +Type + Category + Paid
                +Impressions_in_100 + ImpressionsRate_Liked
                +ReachRate + ReachRate_Liked
                +EngagedRate + EngagedRate_Liked
                +ConsumerRate + Interactions
                ,data = facebook_train)
```

From the Model 0 summary statistics, we can see that that our F test for the model is significant with the F Statistic value of 13.81 and the corresponding p-value is 2.2e-16. This lies far in the right hand side of the rejection region of our null hypothesis. So, we conclude that there exists some strong statistical evidence that there is a linear relation between our predictor variable and one of our regressors.

Further, we check the significance of the individual regressors in explaining our predictor variable. What we observe is that, the type of post, total number of impressions, paid or unpaid post etc.., not significant and do not have any explanatory power in defining our outcome at this stage. We make transformations on our regressors and then try to make them as significant variables.

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.853  -7.923  -1.910   1.938 299.503
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.526840  17.775756   0.086 0.931595
## TypePhoto                2.918967   7.077105   0.412 0.680238
## TypeStatus               5.285383   9.326490   0.567 0.571244
## TypeVideo               -2.151852  12.729549  -0.169 0.865851
## CategoryInspiration    -15.558255   3.726260  -4.175 3.68e-05 ***
## CategoryProduct        -15.359109   4.124330  -3.724 0.000225 ***
## PaidYes                 -0.582266   3.180610  -0.183 0.854842
## Impressions_in_100       0.001778   0.004058   0.438 0.661623
## ImpressionsRate_Liked   -0.022420   0.049143  -0.456 0.648488
## ReachRate               -0.005274   0.026738  -0.197 0.843748
## ReachRate_Liked          0.059355   0.224688   0.264 0.791793
## EngagedRate              2.871838   0.467965   6.137 2.09e-09 ***
## EngagedRate_Liked       -0.732503   0.495903  -1.477 0.140464
## ConsumerRate            -0.058133   0.163772  -0.355 0.722812
## Interactions             0.002823   0.004461   0.633 0.527245
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.9 on 385 degrees of freedom
## Multiple R-squared:  0.3343, Adjusted R-squared:  0.3101
## F-statistic: 13.81 on 14 and 385 DF,  p-value: < 2.2e-16
```

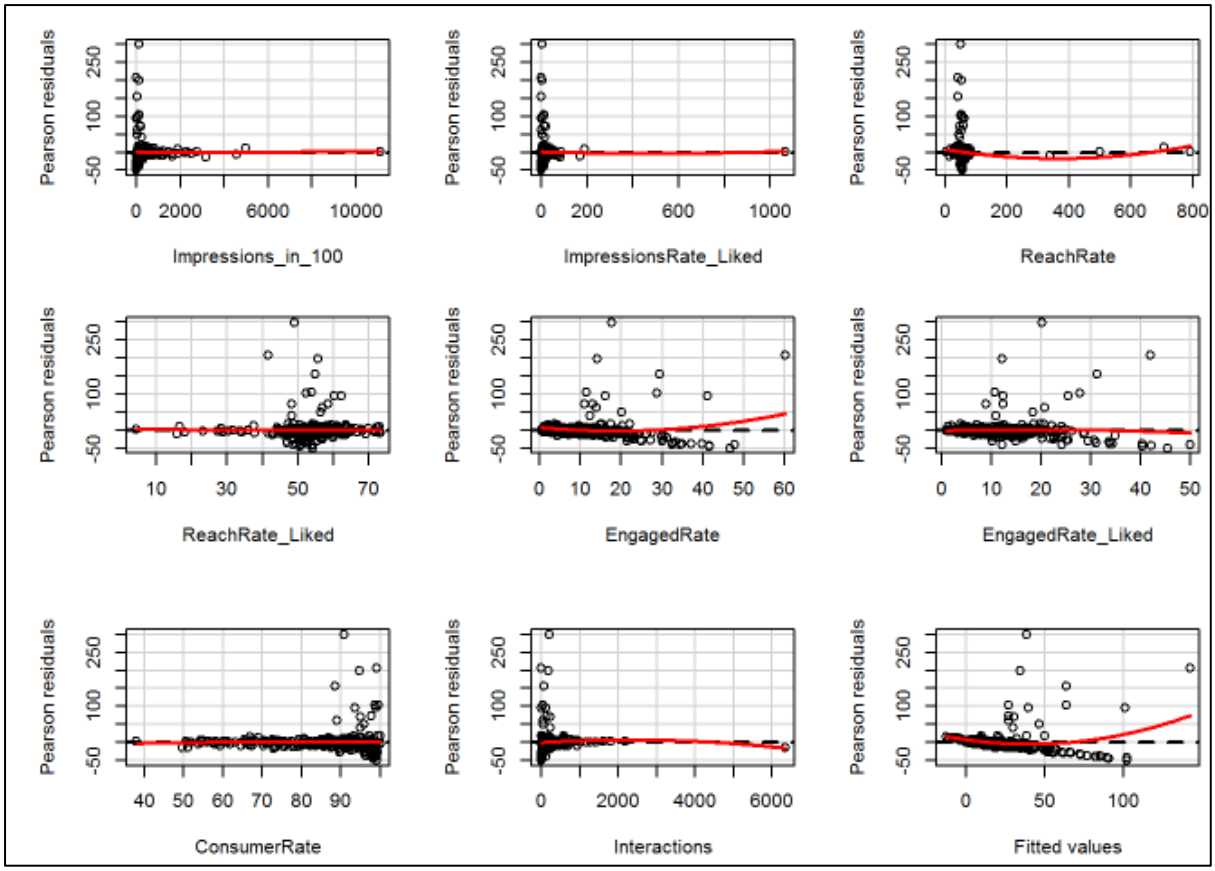
The Residual Standard error: 27.62. Says the interval in which our predicted value might fall at a given confidence interval.
The parameter R Squared (R Sq.) is 0.3343. This means that, 33.43% of the total variation in the data is captured by our model. In other words our model will give us 33.43% accurate results. The Adjusted R Squared (Adj R Sq.) is 0.3101 gives us indication that the effect if any new parameters are added or removed to our model.

From the above discussion on Model 0, we can be happy with our results and say that our model is accurate with 33% predictive capability. This is just the result of simple model we have built. We can still make improvements to this model to get better R Sq. value and reduce the standard error.

Checking Validity of Model 0:

Checking the residual plots will give us an indication of how the regressors are preforming in accordance with the residuals.

From the below plots the engaged user rate is having an exponential effect on our output variable. So, we need to apply an appropriate transformation on this in order to make the regressor to behave linearly. The transformations that are supposed to be made are also shown

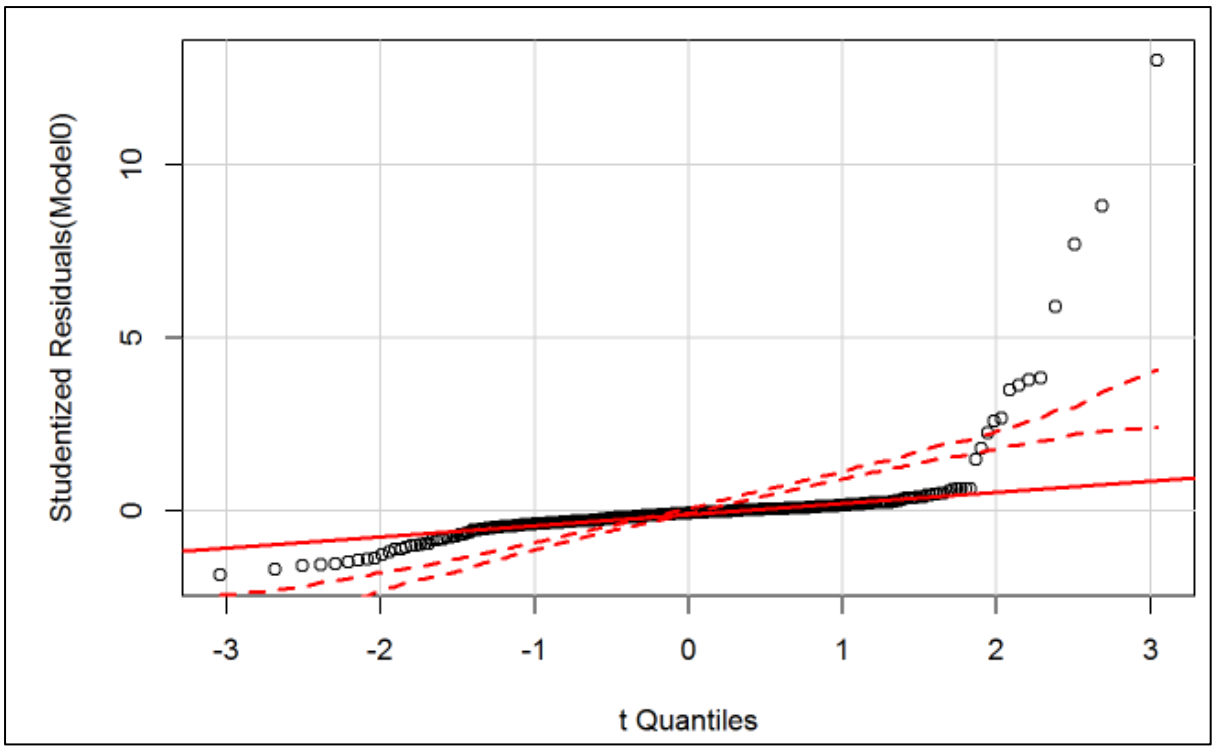



```
consumptionRate_Log <- log(facebook_train$consumptionRate)
Impressions_in_100_Log <- log(facebook_train$Impressions_in_100)
ImpressionsRate_Liked_Log <- log(facebook_train$ImpressionsRate_Liked)
ReachRate_Log <- log(facebook_train$ReachRate)
ReachRate_Liked_Log <- log(max(facebook_train$ReachRate_Liked)+1-facebook_train$ReachRate_Liked)
EngagedRate_Log <- log(facebook_train$EngagedRate)
EngagedRate_Liked_Log <- log(facebook_train$EngagedRate_Liked)
ConsumerRate_Log <- log(max(facebook_train$ConsumerRate)+1- facebook_train$ConsumerRate)
facebook_train <- cbind(facebook_train,ConsumerRate_Log,Impressions_in_100_Log,ImpressionsRate_Liked_Log,ReachR
ate_Log,ReachRate_Liked_Log,EngagedRate_Log,EngagedRate_Liked_Log,ConsumerRate_Log)
```

we applied log transformations and the reflected log transformations on the independent variables. So, each regressor is behaving either exponentially with residuals. We convert them so as they behave linearly with our outcome '*consumption rate*'

Residual plot of Model 0:

We know from our assumptions that that for our model the residuals must be normally distributed. From the qq-polt we see that our residuals are not following the normality assumption.



Model 1:

We build our Model 1 based on the findings from the Model 0. The regression equation for Model 1 will be:

```
Model1 <- lm(consumptionRate_Log~
                +Type+Category+Paid
                +Impressions_in_100_Log + ImpressionsRate_Liked_Log
                +ReachRate_Log + ReachRate_Liked_Log
                +EngagedRate_Log + EngagedRate_Liked_Log
                +ConsumerRate_Log
                +Interactions
                ,data = facebook_train)
```

The summary statistics of Model 1 are shown below: If we look into the summary of model 1, the F Statistic, value is 198. We can see that the F-statistic value has increased from Model 0, which strengthens our linearity assumptions.

The consumer rate has become additionally significant in addition to other significant variables in which were significant in model 0.

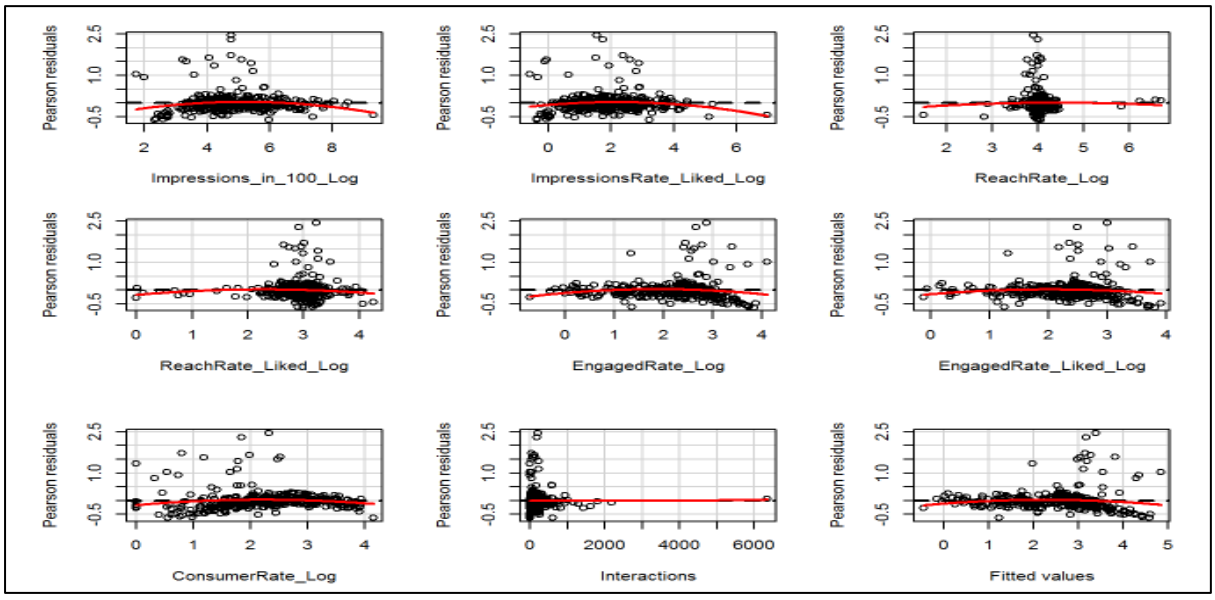
On applying the transformations in order to form a linear model the R Sq. Value which is 87% is dramatically improved. By applying the transformations in order to form a linear model has increased our R Sq. value by 52%. So our model 1 has more predictive power than the Model 0. We have certainly improved the performance.

The Residual standard error value is now 0.3497 when compared to 27 in last model. We have improved on our prediction accuracy too.

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62226 -0.14575 -0.03534  0.06221  2.44781
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   7.429e-01  5.853e-01   1.269 0.205074
## TypePhoto                     7.446e-02  9.413e-02   0.791 0.429434
## TypeStatus                   -6.724e-02  1.262e-01  -0.533 0.594451
## TypeVideo                    -4.522e-02  1.633e-01  -0.277 0.781987
## CategoryInspiration          -2.383e-01  4.790e-02  -4.976 9.82e-07 ***
## CategoryProduct              -2.019e-01  5.438e-02  -3.713 0.000235 ***
## PaidYes                      -4.691e-03  4.011e-02  -0.117 0.906956
## Impressions_in_100_Log       -1.677e-02  8.483e-02  -0.198 0.843373
## ImpressionsRate_Liked_Log     1.044e-01  8.378e-02   1.245 0.213711
## ReachRate_Log                -1.044e-01  8.072e-02  -1.293 0.196686
## ReachRate_Liked_Log          -3.788e-02  4.574e-02  -0.828 0.408029
## EngagedRate_Log               9.883e-01  1.103e-01   8.957  < 2e-16 ***
## EngagedRate_Liked_Log         1.742e-01  1.226e-01   1.421 0.156170
## ConsumerRate_Log             -1.479e-01  2.534e-02  -5.839 1.12e-08 ***
## Interactions                 -7.007e-05  5.501e-05  -1.274 0.203533
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3497 on 385 degrees of freedom
## Multiple R-squared:  0.8781, Adjusted R-squared:  0.8736
## F-statistic:   198 on 14 and 385 DF,  p-value: < 2.2e-16
```

Checking validity of Model 1:

Though we performed well on our R Sq. and Standard error values. We need to check the other assumptions too. Below is the residual plot's vs regressors plots for model 1.



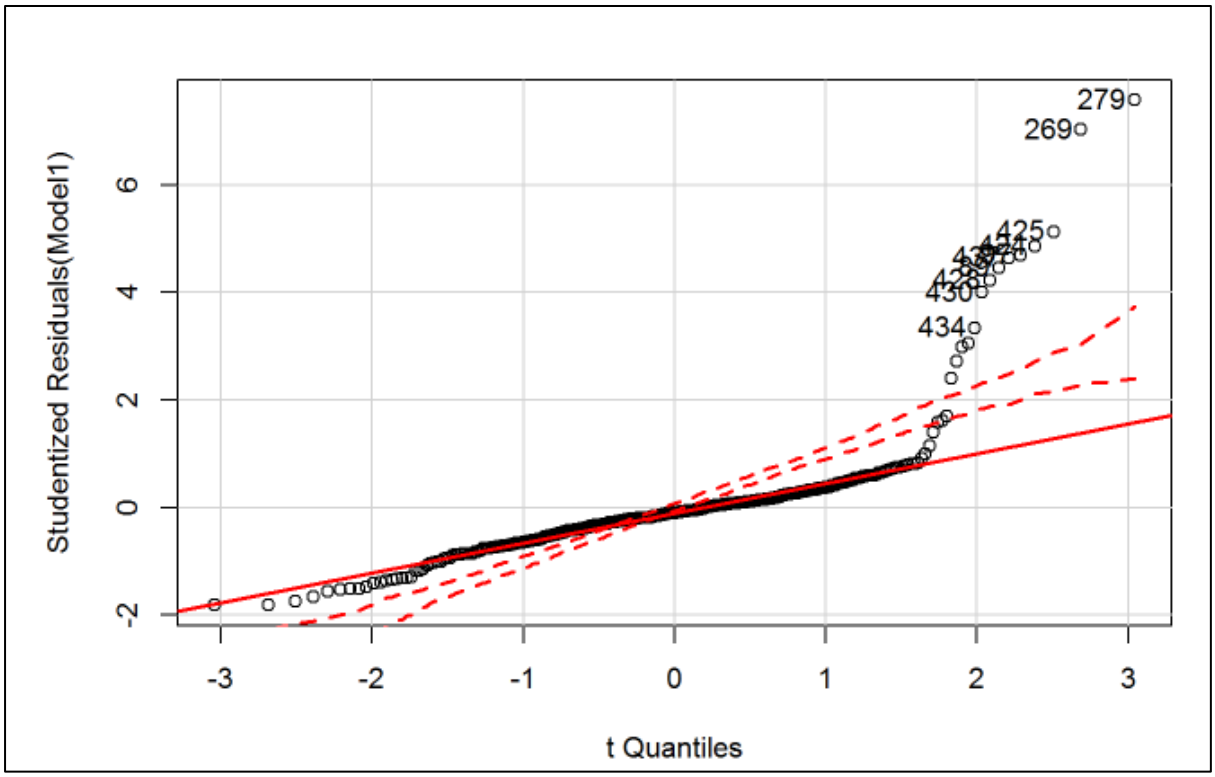
From the residual plots we see that the regressors are not behaving linearly and are showing a quadratic pattern with the residuals. So, we need to apply the squared transformations to the regressors which are showing the characteristics. Below are the transformations made.

```
Impressions_in_100_Log_Sq <- facebook_train$Impressions_in_100_Log^2
ImpressionsRate_Liked_Log_Sq <-  facebook_train$ImpressionsRate_Liked_Log^2
ReachRate_Log_Sq <- facebook_train$ReachRate_Log^2
ReachRate_Liked_Log_Sq <- facebook_train$ReachRate_Liked_Log^2
EngagedRate_Log_Sq <- facebook_train$EngagedRate_Log^2
EngagedRate_Liked_Log_Sq <- facebook_train$EngagedRate_Liked_Log^2
ConsumerRate_Log_Sq <- facebook_train$ConsumerRate_Log^2
```

These transformations will be applied in our next model where we will improve on our current one.

Residual Plot of model 1:

Still we do not see the normality assumption of the residuals is not properly followed. But we can see on the left lower edge few data points have come closer to the red line we have.



Concluding on Model 1, we see that we have improved from model 0 on our R Sq. and Standard error values form Model 0. But we still were not able to meet the assumptions we talked about, in this model. So we will try and improve the areas so as we get more reliable model.

Model 2:

The regression equation for model 2 is as shown below:

```
Model2 <- lm(consumptionRate_Log~
                  +Type+Category+Paid
                  +Impressions_in_100_Log + Impressions_in_100_Log_Sq
                  +ImpressionsRate_Liked_Log + ImpressionsRate_Liked_Log_Sq
                  +ReachRate_Log + ReachRate_Log_Sq
                  +ReachRate_Liked_Log + ReachRate_Liked_Log_Sq
                  +EngagedRate_Log + EngagedRate_Log_Sq
                  +EngagedRate_Liked_Log + EngagedRate_Liked_Log_Sq
                  +ConsumerRate_Log + ConsumerRate_Log_Sq
                  +Interactions
                  ,data = facebook_train)
```

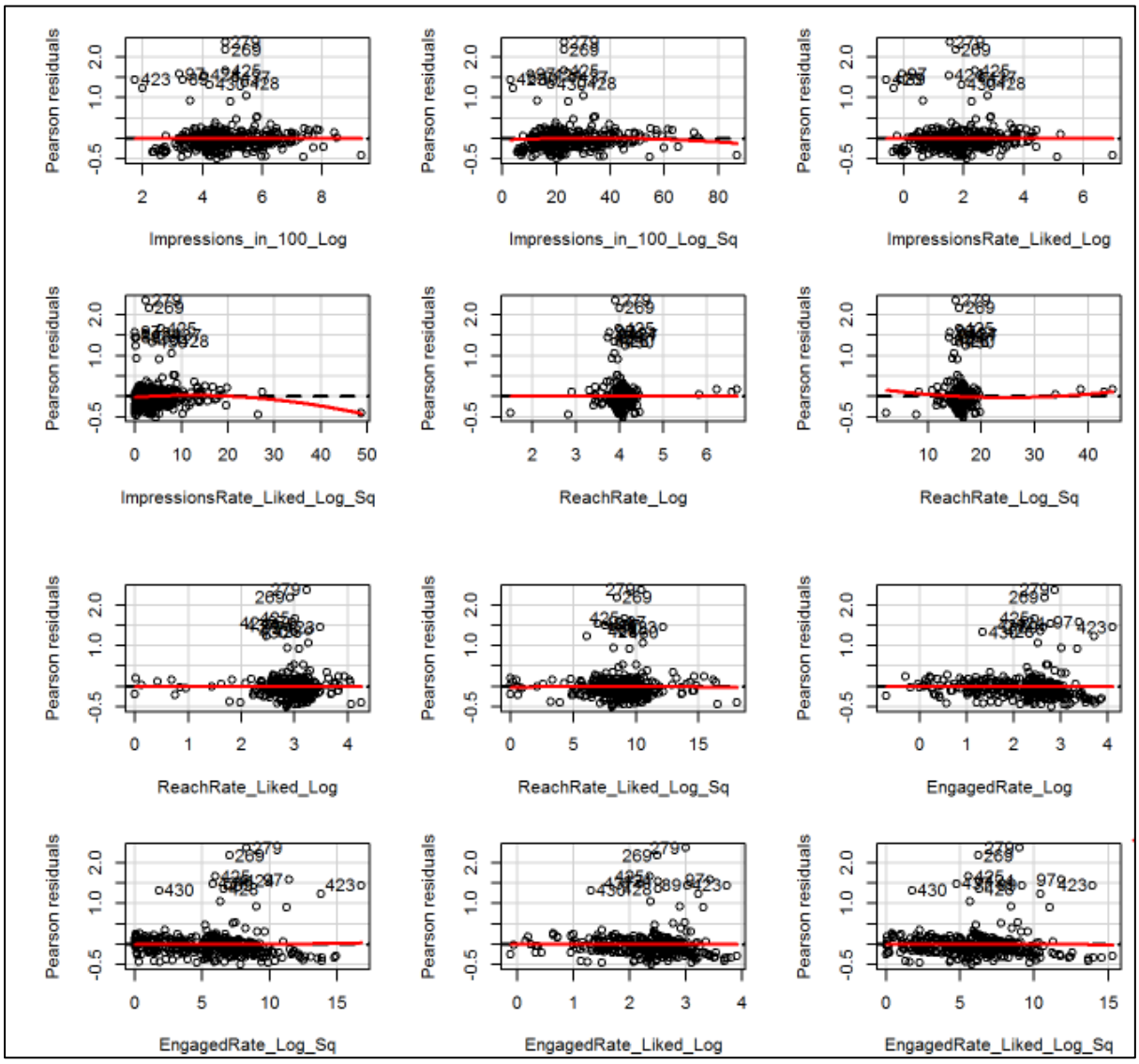
The summary statistics of the model are shown below:

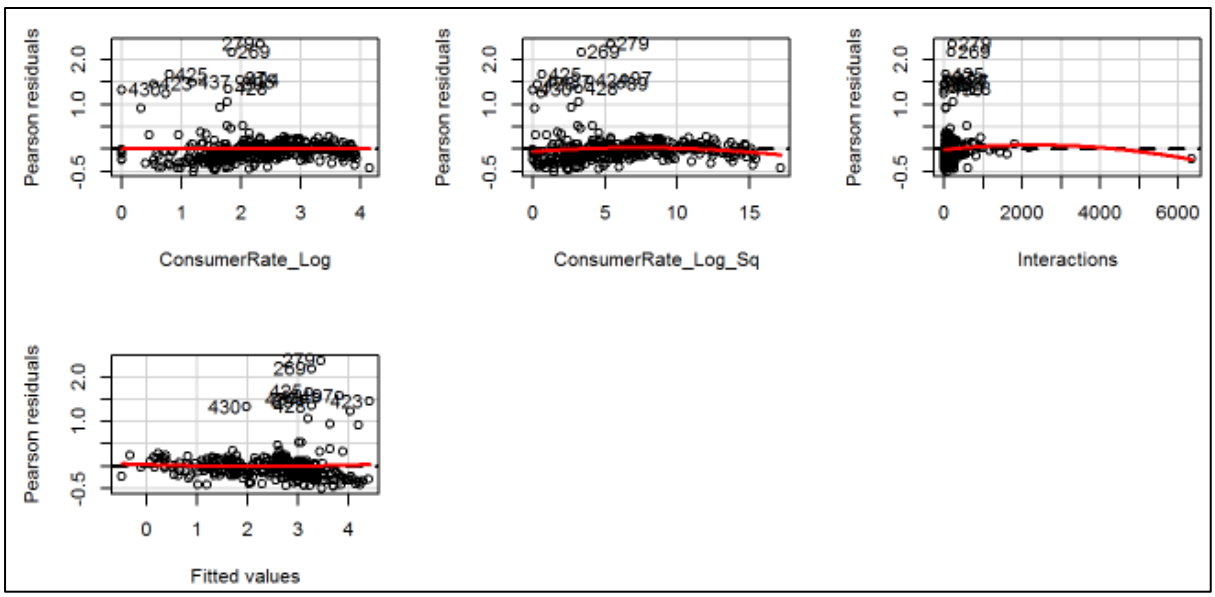
```
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.903e+00  1.287e+00   1.479 0.140099
## TypePhoto                     1.504e-01  9.927e-02   1.515 0.130482
## TypeStatus                   -1.618e-02  1.334e-01  -0.121 0.903480
## TypeVideo                     6.646e-02  1.677e-01   0.396 0.692156
## CategoryInspiration          -2.754e-01  4.995e-02  -5.514  6.5e-08 ***
## CategoryProduct              -2.379e-01  5.535e-02  -4.297  2.2e-05 ***
## PaidYes                      -7.555e-03  3.968e-02  -0.190 0.849106
## Impressions_in_100_Log        2.581e-01  2.819e-01   0.916 0.360441
## Impressions_in_100_Log_Sq    -3.164e-02  2.431e-02  -1.301 0.193900
## ImpressionsRate_Liked_Log     1.322e-01  1.665e-01   0.794 0.427727
## ImpressionsRate_Liked_Log_Sq -8.962e-03  2.616e-02  -0.343 0.732103
## ReachRate_Log                -8.585e-01  4.520e-01  -1.899 0.058296 .
## ReachRate_Log_Sq              6.317e-02  4.647e-02   1.359 0.174825
## ReachRate_Liked_Log           1.662e-01  1.484e-01   1.120 0.263232
## ReachRate_Liked_Log_Sq       -4.947e-02  3.563e-02  -1.388 0.165848
## EngagedRate_Log               9.262e-01  2.559e-01   3.619 0.000336 ***
## EngagedRate_Log_Sq           -1.642e-02  5.228e-02  -0.314 0.753656
## EngagedRate_Liked_Log         2.639e-01  3.201e-01   0.824 0.410202
## EngagedRate_Liked_Log_Sq     -5.731e-03  6.414e-02  -0.089 0.928853
## ConsumerRate_Log              3.411e-02  9.342e-02   0.365 0.715191
## ConsumerRate_Log_Sq          -5.193e-02  2.082e-02  -2.494 0.013040 *
## Interactions                  7.611e-05  6.664e-05   1.142 0.254155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3426 on 378 degrees of freedom
## Multiple R-squared:  0.8851, Adjusted R-squared:  0.8787
## F-statistic: 138.6 on 21 and 378 DF,  p-value: < 2.2e-16
```

Our R Sq. Value has improved by 1% to 88.51% in this model. The Standard error value has also been reduced by a small fraction to 0.3426. The user reach rate parameter of a post has become significant additionally to the previous model. So, yes making changes to our previous model has worked, but lets look if the regression assumptions are met.

Checking the Validity of Model 2:

The regressors Vs the residual plots for model 2 are shown below:

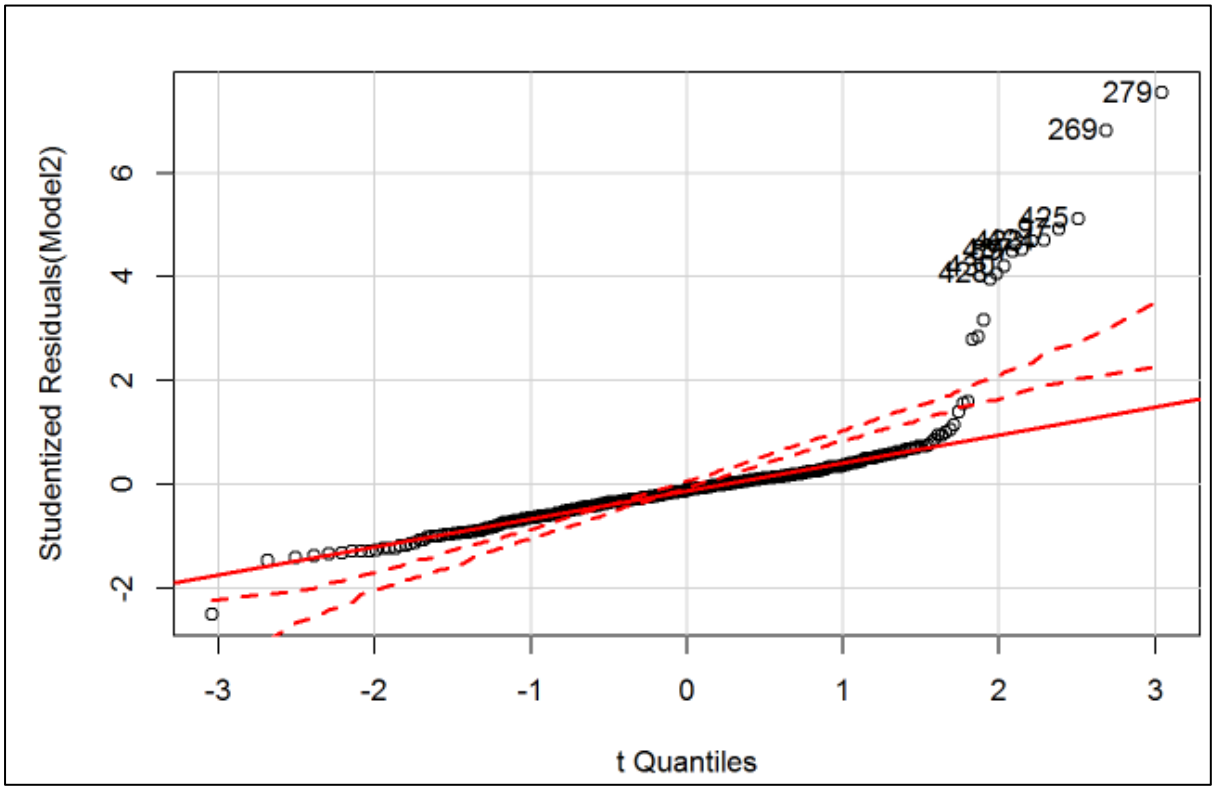




From this we can see that almost all the regressors are not having any influence on our residuals. There is an exception in 'Impressions rate by logged square' where there is influence of an outlier. The 'reach rate logged square' and 'consumer rate logged square' are also having influence on the residuals which are minimal and these will be taken care at the next step.

Residual plot of model 2:

From the normality plot of the residuals, there are still some data points which affect our residuals. But few data points have come closer to the normality line. From the below plot we can see that observations 279, 269, 425 are the most influential observations in our training set. We can also check the overall outliers through cooks distance method.

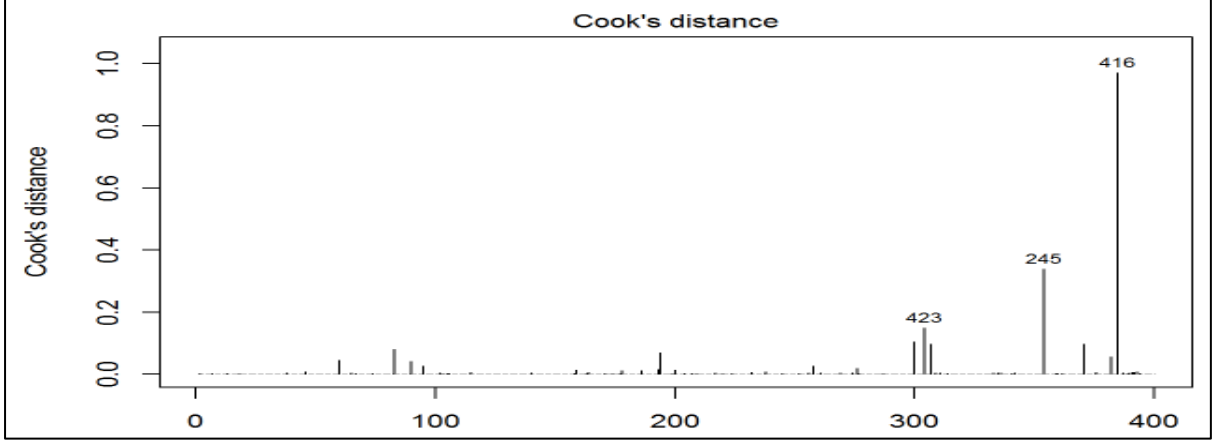


Outlier detection through cooks distance:

This will give us the observations which are very distinct from other observations in our data set. From the plot we can see observations: 416, 245, 423 are the outliers. And from the above qq-plot we can see that there are not many common observations. So, our outliers are not influencing our model. Our model is influenced due to other parameters that are present in our model.

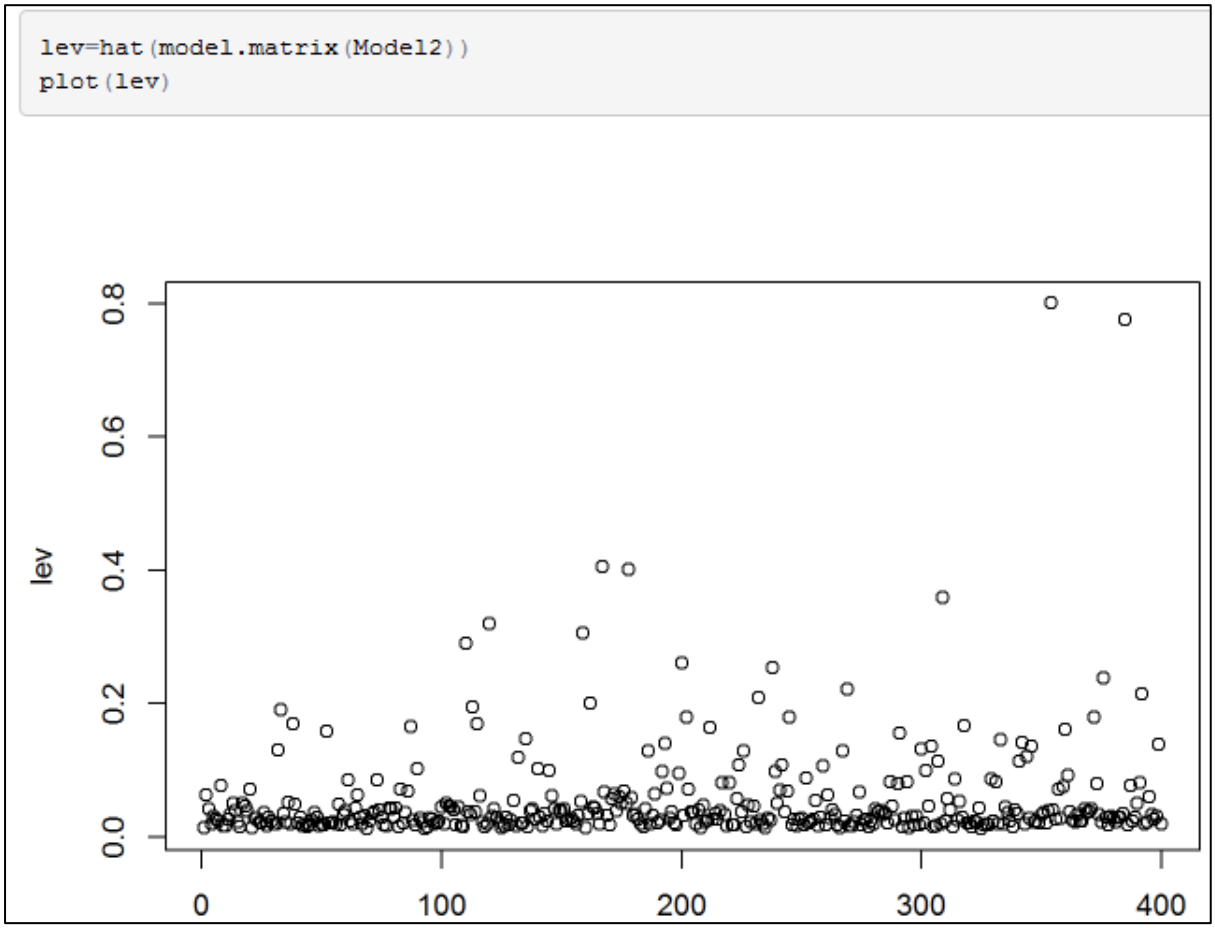
Note: These observations will be changing for each run, as the method of partitioning done before the regression is performed is a random selection.

```
cutoff2 <- 4/((nrow(facebook_train)-length(Model2$coefficients)-2))
plot(Model2, which=4, cook.levels=cutoff2)
```



Looking at the leverage plots will give us an graphical representation to know how much a data point is influential in the x direction that influence our regression line. They have the ability to have high influence on our model. Observations on the right side top corner which are very distant from others, and they might be good or bad leverage points. The observations which are having relatively high influence (greater than 0.2) in this case are: 447, 477, 442,403, 464, 493, 305, 141, 309, 478, 483, 245, 369, 416, 278.

```
lev=hat(model.matrix(Model2))
plot(lev)
```



With these many influential observations and leverage points, we need to recheck our model with and without these observations and how the estimates are influenced. So, here we go to our next model with the possible improvements from current model, Model 2.

Model 3:

The regression equation for the Model 3 will be as shown below. We can see all the regression parameters are the same and we are only removing the influential observations form modelling. (These observations are not exactly the ones shown above as the whole program is run in one shot and the values change for each run..

```
Model3 <- lm(consumptionRate_Log~
                  +Type+Category+Paid
                  +Impressions_in_100_Log + Impressions_in_100_Log_Sq
                  +ImpressionsRate_Liked_Log + ImpressionsRate_Liked_Log_Sq
                  +ReachRate_Log + ReachRate_Log_Sq
                  +ReachRate_Liked_Log + ReachRate_Liked_Log_Sq
                  +EngagedRate_Log + EngagedRate_Log_Sq
                  +EngagedRate_Liked_Log + EngagedRate_Liked_Log_Sq
                  +ConsumerRate_Log + ConsumerRate_Log_Sq
                  +Interactions
                  ,data = facebook_train[-c(97,245,279,413)])
```

The summary stats of the model from the below snapshot are almost identical to Model 2. None of our estimates or their significance, the R Sq value, the residual standard errors are effected by the removing these observations. So, we can say that model 3 is almost same as model 2, and these influential observations or leverage points do not have hardly any influence on our regression line.

```
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.903e+00  1.287e+00   1.479 0.140099
## TypePhoto                     1.504e-01  9.927e-02   1.515 0.130482
## TypeStatus                   -1.618e-02  1.334e-01  -0.121 0.903480
## TypeVideo                     6.646e-02  1.677e-01   0.396 0.692156
## CategoryInspiration          -2.754e-01  4.995e-02  -5.514  6.5e-08 ***
## CategoryProduct              -2.379e-01  5.535e-02  -4.297  2.2e-05 ***
## PaidYes                      -7.555e-03  3.968e-02  -0.190 0.849106
## Impressions_in_100_Log        2.581e-01  2.819e-01   0.916 0.360441
## Impressions_in_100_Log_Sq    -3.164e-02  2.431e-02  -1.301 0.193900
## ImpressionsRate_Liked_Log     1.322e-01  1.665e-01   0.794 0.427727
## ImpressionsRate_Liked_Log_Sq -8.962e-03  2.616e-02  -0.343 0.732103
## ReachRate_Log                -8.585e-01  4.520e-01  -1.899 0.058296 .
## ReachRate_Log_Sq              6.317e-02  4.647e-02   1.359 0.174825
## ReachRate_Liked_Log           1.662e-01  1.484e-01   1.120 0.263232
## ReachRate_Liked_Log_Sq       -4.947e-02  3.563e-02  -1.388 0.165848
## EngagedRate_Log               9.262e-01  2.559e-01   3.619 0.000336 ***
## EngagedRate_Log_Sq           -1.642e-02  5.228e-02  -0.314 0.753656
## EngagedRate_Liked_Log         2.639e-01  3.201e-01   0.824 0.410202
## EngagedRate_Liked_Log_Sq     -5.731e-03  6.414e-02  -0.089 0.928853
## ConsumerRate_Log              3.411e-02  9.342e-02   0.365 0.715191
## ConsumerRate_Log_Sq          -5.193e-02  2.082e-02  -2.494 0.013040 *
## Interactions                  7.611e-05  6.664e-05   1.142 0.254155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3426 on 378 degrees of freedom
## Multiple R-squared:  0.8851, Adjusted R-squared:  0.8787
## F-statistic: 138.6 on 21 and 378 DF,  p-value: < 2.2e-16
```

The residual plots Vs regressors, the normal probability plot are the same for Model 2 and Model 3.

Checking the Multicollinearity:

The assumption that all the regressors should be independent of each other in our model is also important. If collinearity is present within the regressors of our model our regression estimate coefficients or the Beta values are highly inflated.
The problem of multicollinearity can be checked using methods: VIF and Colldiag.

The VIF gives is the variance inflation factor: The GVIF column in the below snapshot tells us how much each regression estimate is influenced because of collinearity. For eg. The estimate of '*Impressions Logged*' is 354 times higher (or lower) than it should have been in the normal conditions. This is a very bad sign. Influence on the estimates more than 20 is a serious problem. And this says that our estimates are not very reliable.
But this can be neutralized by the fact, what we have seen during our exploratory data analysis stage. We have seen there is collinearity present between the total impressions – impressions by liked users and engagement and engagement by liked users. And also the several of the squared components or logged components of the same variable which are present explain the problem of why we are seeing the inflated estimates.

```
library(perturb)
library(MASS)
#colldiag(fit1,add.intercept=FALSE,center=TRUE)
vif(Model2)
```

```
##                                    GVIF Df GVIF^(1/(2*Df))
## Type                           3.666372  3        1.241765
## Category                       2.116392  2        1.206144
## Paid                           1.081702  1        1.040049
## Impressions_in_100_Log       354.865523  1       18.837875
## Impressions_in_100_Log_Sq    290.507302  1       17.044275
## ImpressionsRate_Liked_Log    103.381232  1       10.167656
## ImpressionsRate_Liked_Log_Sq  55.448134  1        7.446350
## ReachRate_Log                 67.773993  1        8.232496
## ReachRate_Log_Sq              58.288713  1        7.634705
## ReachRate_Liked_Log           17.567221  1        4.191327
## ReachRate_Liked_Log_Sq        22.969469  1        4.792647
## EngagedRate_Log              152.136688  1       12.334370
## EngagedRate_Log_Sq            95.618184  1        9.778455
## EngagedRate_Liked_Log        156.266438  1       12.500657
## EngagedRate_Liked_Log_Sq     109.723837  1       10.474915
## ConsumerRate_Log              22.563106  1        4.750064
## ConsumerRate_Log_Sq           23.026891  1        4.798634
## Interactions                   2.280953  1        1.510282
```

Next is the collinearity Diagnostic matrix, which gives us the variance decomposition matrix. Upon complete inspection of the conditional indices given in the second column with the variance > 30 we

can see all the transformed variable pairs are contributing to collinearity in this case. So we can say there are not any big findings other than that.

```
colldiag(facebook_train[,c(14:27)])
```

```
Condition
Index    Variance Decomposition Proportions
           intercept Impressions_in_100_Log ImpressionsRate_Liked_Log ReachRate_Log ReachRate_Liked_Log
1      1.000 0.000     0.000                0.000                     0.000         0.000
2      3.383 0.000     0.000                0.000                     0.000         0.000
3      5.264 0.000     0.000                0.000                     0.000         0.000
4     11.205 0.000     0.000                0.000                     0.000         0.000
5     12.574 0.000     0.000                0.001                     0.000         0.004
6     17.194 0.000     0.000                0.001                     0.000         0.000
7     24.863 0.000     0.000                0.040                     0.000         0.000
8     39.097 0.001     0.000                0.045                     0.000         0.000
9     49.706 0.001     0.000                0.034                     0.000         0.001
10    57.327 0.003     0.000                0.010                     0.000         0.001
11    74.841 0.010     0.002                0.049                     0.000         0.001
12   113.077 0.001     0.001                0.000                     0.000         0.910
13   237.085 0.010     0.061                0.153                     0.000         0.001
14   362.517 0.078     0.935                0.666                     0.013         0.003
15   679.062 0.896     0.000                0.001                     0.986         0.077
```

Best Subset Selection from model 2(3):

Finally we have checked all our assumptions, of the best model we have built that is model 2. Now we perform a method of best subset selection which is an automated process, which selects the best regressors and gives us the variables that are necessary to be in our final regression equation.

```
###Best subset regression
step <- stepAIC(Model2, direction="both")
```

```
## Step:  AIC=-803.09
## consumptionRate_Log ~ Type + Category + Impressions_in_100_Log_Sq +
##     ImpressionsRate_Liked_Log + ReachRate_Log + ReachRate_Liked_Log +
##     ReachRate_Liked_Log_Sq + EngagedRate_Log + EngagedRate_Log_Sq +
##     EngagedRate_Liked_Log + ConsumerRate_Log_Sq
##
##                                 Df Sum of Sq    RSS     AIC
## <none>                                       49.837 -803.09
## - ReachRate_Liked_Log            1    0.2940 50.131 -802.73
## - ReachRate_Liked_Log_Sq         1    0.3266 50.163 -802.47
## + ConsumerRate_Log               1    0.1661 49.671 -802.42
## + Interactions                   1    0.1494 49.687 -802.29
## + ReachRate_Log_Sq               1    0.1138 49.723 -802.00
## + Paid                           1    0.0461 49.791 -801.46
## + EngagedRate_Liked_Log_Sq       1    0.0077 49.829 -801.15
## + ImpressionsRate_Liked_Log_Sq   1    0.0004 49.836 -801.09
## + Impressions_in_100_Log         1    0.0000 49.837 -801.09
## - Impressions_in_100_Log_Sq      1    0.8075 50.644 -798.66
## - ImpressionsRate_Liked_Log      1    0.9231 50.760 -797.74
## - EngagedRate_Liked_Log          1    1.0971 50.934 -796.38
## - Type                           3    1.7254 51.562 -795.47
## - ReachRate_Log                  1    1.5712 51.408 -792.67
## - EngagedRate_Log_Sq             1    1.8292 51.666 -790.67
## - Category                       2    2.8876 52.724 -784.56
## - EngagedRate_Log                1    5.3050 55.142 -764.62
## - ConsumerRate_Log_Sq            1   10.9488 60.785 -725.65
```

The Step AIC method gave us the best possible subset of variables that can be included in our model. We had total of 11 variables are selected from 18 variables we had in model 2.

So, our best regression equation can be written as:

```
(consumptionRate_Log~
                Type+Category
                +Impressions_in_100_Log_Sq+ImpressionsRate_Liked_Log
                +ReachRate_Log+ReachRate_Liked_Log
                +EngagedRate_Log+EngagedRate_Log_Sq+EngagedRate_Liked_Log
                +ConsumerRate_Log_Sq
```

Model4:

After getting the best subset of variables, we check the equation by modelling it on our training data set: The regression equation will be:

```
Model4 <- lm(consumptionRate_Log~
                            Type+Category
                            +Impressions_in_100_Log_Sq+ImpressionsRate_Liked_Log
                            +ReachRate_Log+ReachRate_Liked_Log
                            +EngagedRate_Log+EngagedRate_Log_Sq+EngagedRate_Liked_Log
                            +ConsumerRate_Log_Sq
             ,data = facebook_train
             )
```

The summary stats for the model:

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65879 -0.16917 -0.04272  0.07282  2.42335
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.399717   0.481243   2.909 0.003841 **
## TypePhoto                     0.183003   0.105476   1.735 0.083537 .
## TypeStatus                   -0.041438   0.132740  -0.312 0.755077
## TypeVideo                    -0.007320   0.183960  -0.040 0.968281
## CategoryInspiration          -0.231196   0.048949  -4.723 3.26e-06 ***
## CategoryProduct              -0.195186   0.054509  -3.581 0.000386 ***
## Impressions_in_100_Log_Sq    -0.017170   0.007372  -2.329 0.020371 *
## ImpressionsRate_Liked_Log     0.211140   0.079152   2.668 0.007963 **
## ReachRate_Log                -0.268511   0.088566  -3.032 0.002595 **
## ReachRate_Liked_Log          -0.014939   0.046814  -0.319 0.749822
## EngagedRate_Log               1.011564   0.163936   6.170 1.72e-09 ***
## EngagedRate_Log_Sq           -0.075924   0.027310  -2.780 0.005700 **
## EngagedRate_Liked_Log         0.333462   0.122333   2.726 0.006706 **
## ConsumerRate_Log_Sq          -0.056907   0.006365  -8.940  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3581 on 386 degrees of freedom
## Multiple R-squared:  0.8736, Adjusted R-squared:  0.8694
## F-statistic: 205.2 on 13 and 386 DF,  p-value: < 2.2e-16
```

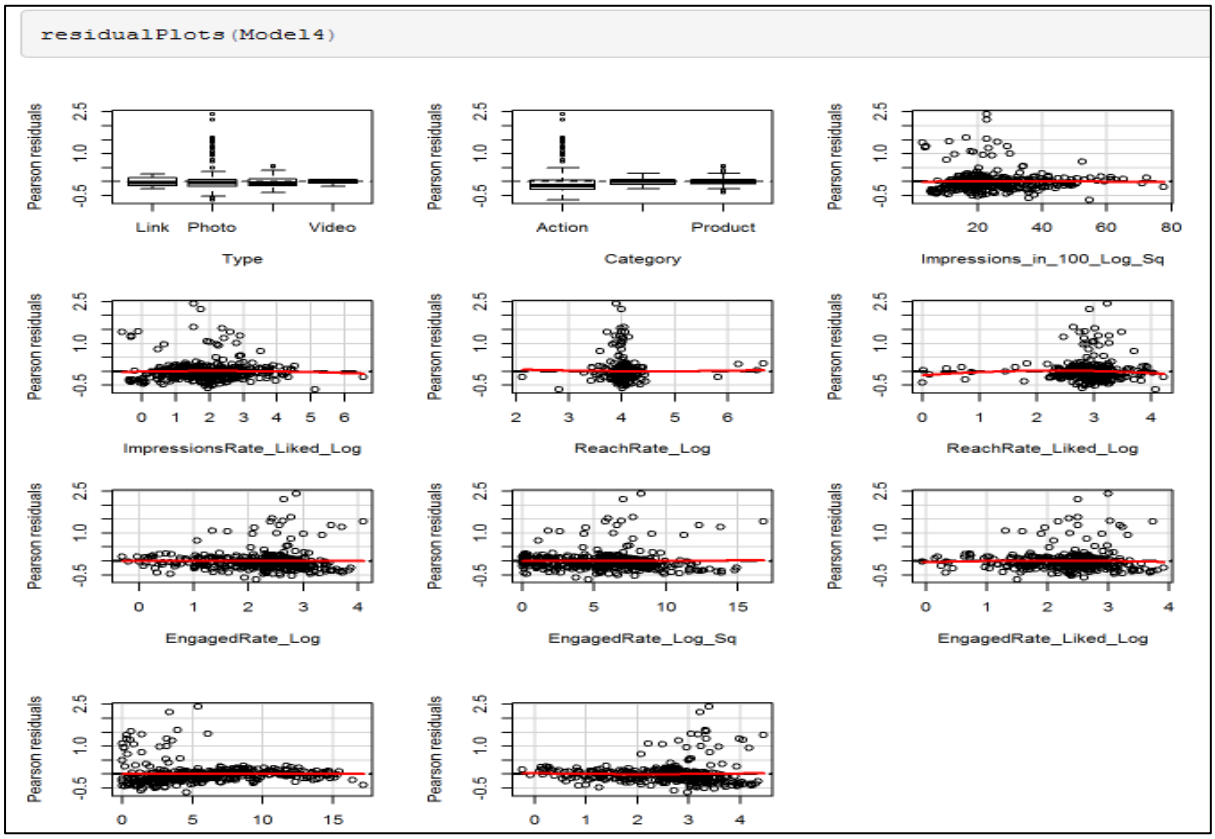
From the summary stats we can see that F-statistic is significant and R Sq. Value is 87.36% and residuals standard error is 0.35.

Along with that we can see many parameters which have become significant than they were in model 2. This say's the importance of subset selection.

Checking the validity of Model 4:

The residuals Vs the repressor's plots say that that the residuals are not affected by any way due to any of the regressors used the equation.

```
residualPlots(Model4)
```



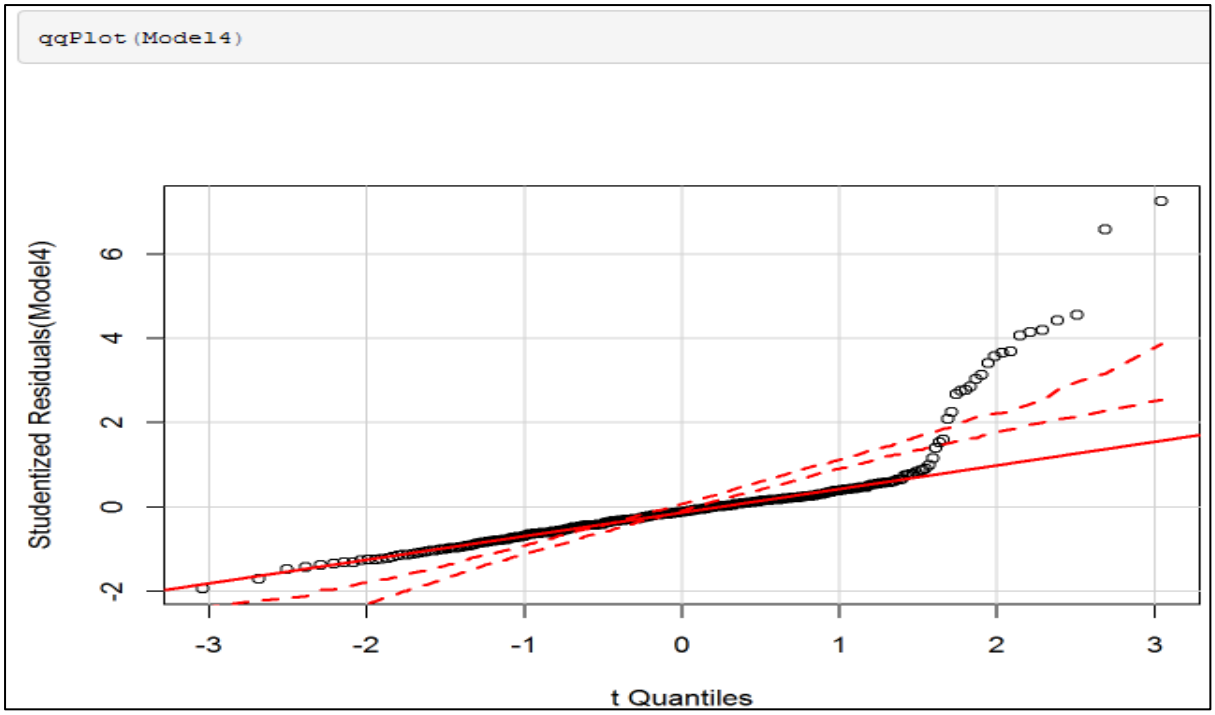
	Test stat	Pr(>|t|)
Type	NA	NA
Category	NA	NA
Impressions_in_100_Log_Sq	-0.181	0.856
ImpressionsRate_Liked_Log	-0.551	0.582
ReachRate_Log	0.304	0.761
ReachRate_Liked_Log	-1.536	0.125
EngagedRate_Log	-1.272	0.204
EngagedRate_Log_Sq	0.286	0.775
EngagedRate_Liked_Log	-1.103	0.271
ConsumerRate_Log_Sq	-0.045	0.964
Tukey test	1.269	0.204

The graphical representation and the corresponding p values shown in the table says that there is no significant effect on residuals by the regressors. This shows our Model 4 has overcome the challenges faced in Model 2. This is also due to the effect of best subset selection.

Residual plot of Model 4:

The qq plot of the residuals says that the residuals are not normally distributed, and that is because of some influential observations and leverage observation. But we have improved a lot from our Model 0 through Model 4 in making the residuals look normal.

```
qqPlot(Model4)
```



Validation of the model:

Now that we have decided on the best model we need to check it's effectiveness by testing it on our validation data. Before that we need to prepare our data in the validation set into the transformations used on the regressors in the Model 4. We keep only the regressors that are necessary by eliminating other variables and keeping only the necessary ones in the test data set. We validate our model on this dataset and check the accuracy of the model.

From the below code output, our model has performed 84% more accurately on the testing data. This is usually a very good precision accuracy. So, we can conclude that the model we have built is a suitable one and we now have to fit this model back to our original data set.

```
y_hat<-predict.lm (Model4,newdata= facebook_test,se.fit=TRUE)$fit
y_hat<-as.vector (y_hat)
dev<- consumptionRate_Log_Val-(y_hat)
num<-sum(dev^2)
dev1<-consumptionRate_Log_Val-mean(consumptionRate_Log_Val)
den<-sum(dev1^2)
Predicted.Rsq<-1-(num/den)
Predicted.Rsq
```

```
## [1] 0.843533
```

Final Model:

Same as done in the above validation dataset, we transform the regressors in the original data set and fit our data on to the model.

```
Final_Model <- lm(consumptionRate_Actual~., data = facebook_final)
summary(Final_Model)
```

Summary statistics of our final model are:

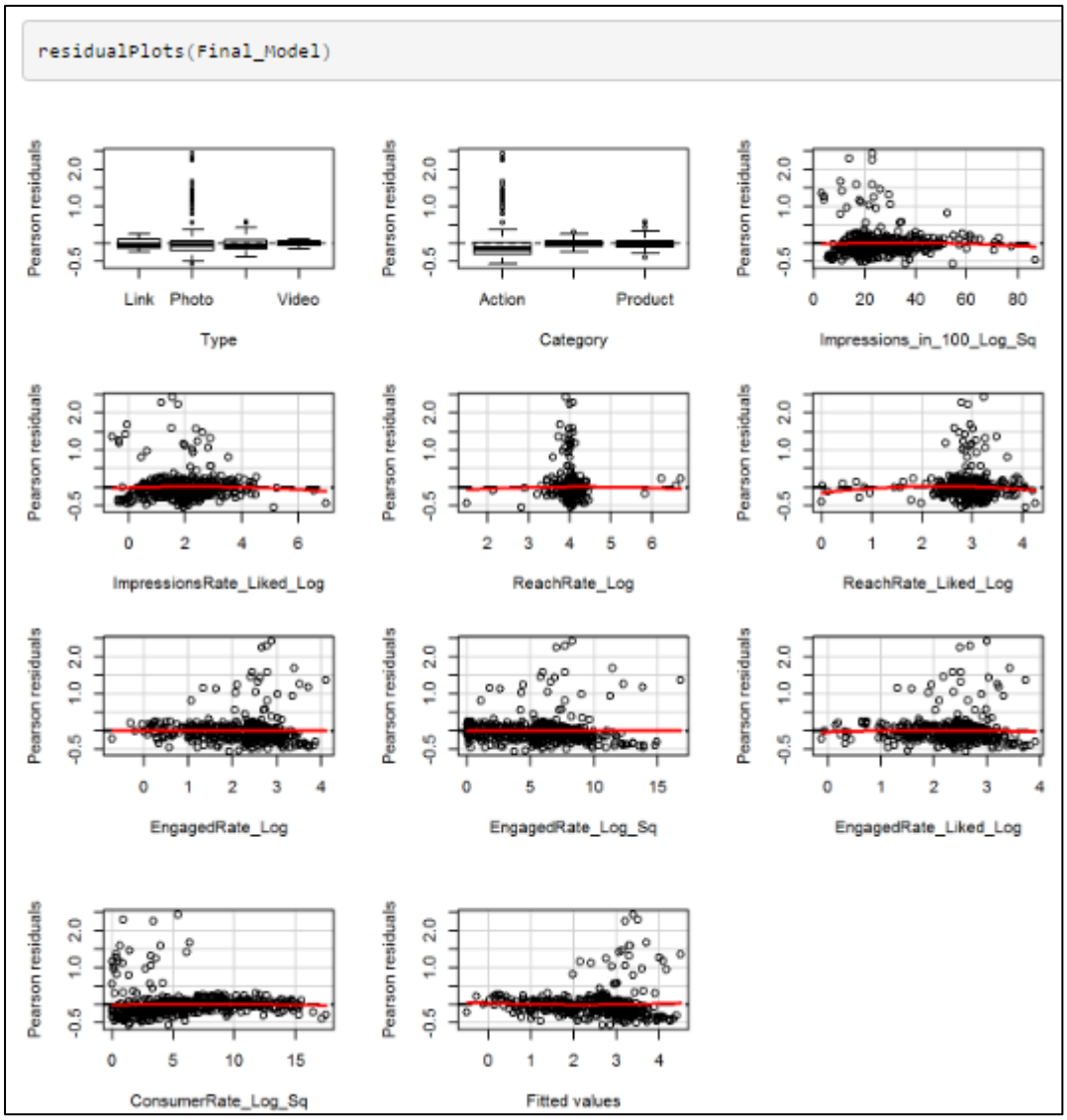
```
##
## Call:
## lm(formula = consumptionRate_Actual ~ ., data = facebook_final)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55431 -0.16758 -0.03544  0.06822  2.42801
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.318364   0.430055   3.066  0.00229 **
## TypePhoto                   0.156413   0.089924   1.739  0.08260 .
## TypeStatus                 -0.061118   0.115563  -0.529  0.59714
## TypeVideo                   0.011553   0.162050   0.071  0.94319
## CategoryInspiration        -0.231600   0.043137  -5.369 1.23e-07 ***
## CategoryProduct            -0.192807   0.048654  -3.963 8.52e-05 ***
## Impressions_in_100_Log_Sq -0.016981   0.006653  -2.552  0.01101 *
## ImpressionsRate_Liked_Log  0.201324   0.071244   2.826  0.00491 **
## ReachRate_Log              -0.247120   0.078402  -3.152  0.00172 **
## ReachRate_Liked_Log        -0.027519   0.043329  -0.635  0.52566
## EngagedRate_Log             0.982067   0.145539   6.748 4.28e-11 ***
## EngagedRate_Log_Sq         -0.065394   0.024091  -2.714  0.00688 **
## EngagedRate_Liked_Log       0.348820   0.110056   3.169  0.00162 **
## ConsumerRate_Log_Sq        -0.052692   0.005561  -9.476  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3542 on 486 degrees of freedom
## Multiple R-squared:  0.872,  Adjusted R-squared:  0.8686
## F-statistic: 254.8 on 13 and 486 DF,  p-value: < 2.2e-16
```

In the final model we were able to capture 87% of the total variation in the data from our model and with standard errors of residuals is only 0.35. The F-stat value is very significant and most are our regressors are highly significant. Though few regressors are not significant, they are in overall important parameters in deciding the final predicted value of our model.

Checking the validity assumptions on Final Model:

The residual Vs the regressor plots say us the residual values are not influenced by our independent variables and the final residual plot says that the all the residual values are independent of each other and we cannot witness any pattern in the residual values.

```
residualPlots(Final_Model)
```




```
##                                Test stat Pr(>|t|)
## Type                                  NA       NA
## Category                              NA       NA
## Impressions_in_100_Log_Sq         -0.821    0.412
## ImpressionsRate_Liked_Log         -0.884    0.377
## ReachRate_Log                     -0.438    0.661
## ReachRate_Liked_Log               -1.716    0.087
## EngagedRate_Log                   -0.653    0.514
## EngagedRate_Log_Sq                 0.084    0.933
## EngagedRate_Liked_Log             -0.940    0.348
## ConsumerRate_Log_Sq               -0.449    0.654
## Tukey test                         1.739    0.082
```

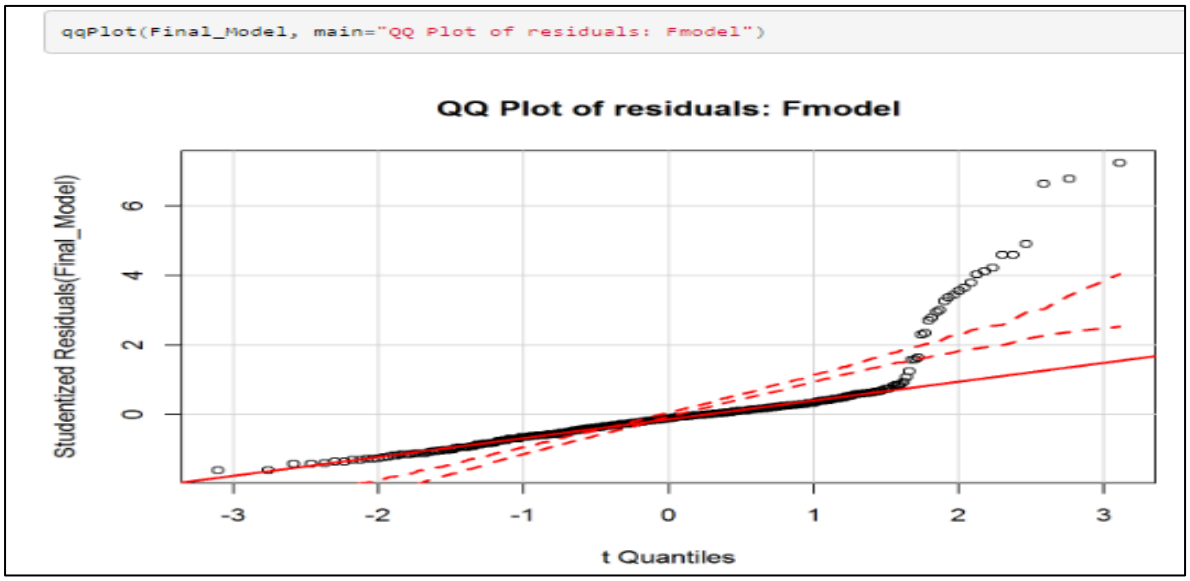
The table says the statistical significance values of the above plots. We can see that none of them are significant at usual confidence levels.

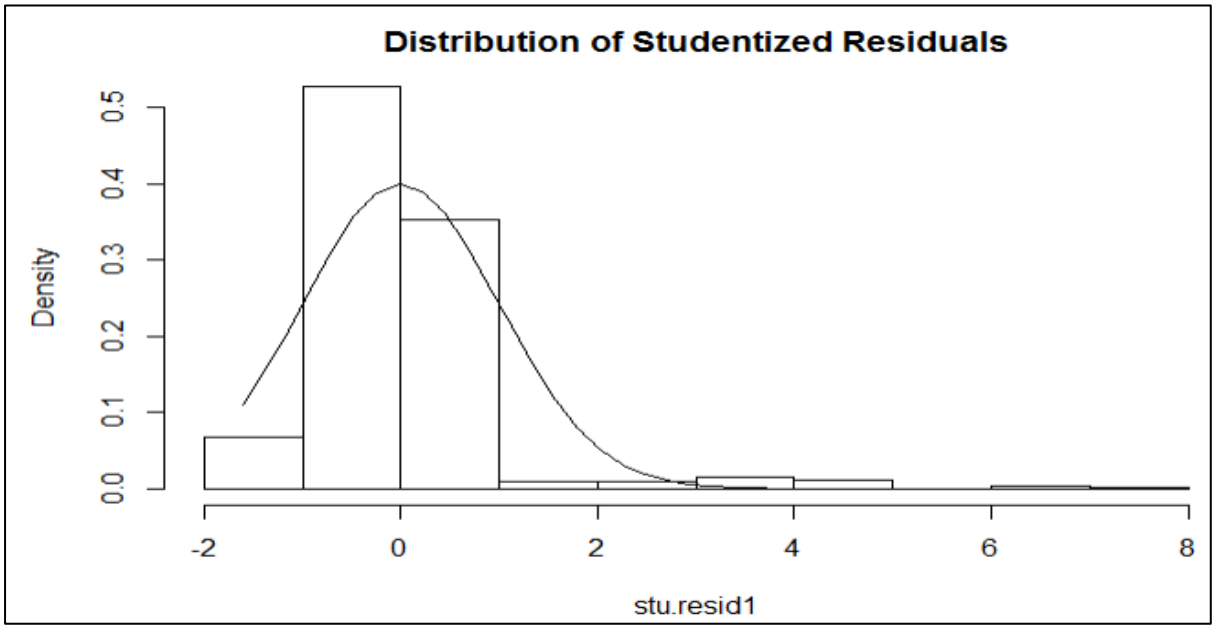
So, we have now just seen that our final model will align with our assumptions that residuals are not dependant / influenced by any of our regressors and the residuals are independent of each other.

Residual plot of the Final Model:

The next assumption we need to check is if the residuals generated are normally distributed or not. We check the normal qq-plot and the histogram of distribution of the residuals.

```
qqPlot(Final_Model, main="QQ Plot of residuals: Fmodel")
```





From this plot it is evident that our regressors are not normally distributed. This problem is occurring because of the fact that we did not treat the corrupted data present in our data set. If we could have had much precise data we could have achieved better results.

Multicollinearity check:

The next assumption is that there is no correlation present between the regressors. The Variance inflation factor values of the model independent variables are described below.

```
vif(Final_Model)
```

	GVIF	Df	GVIF^(1/(2*Df))
Type	2.488291	3	1.164082
Category	1.789611	2	1.156617
Impressions_in_100_Log_Sq	27.282371	1	5.223253
ImpressionsRate_Liked_Log	22.899746	1	4.785368
ReachRate_Log	2.217130	1	1.489003
ReachRate_Liked_Log	1.615168	1	1.270893
EngagedRate_Log	56.301009	1	7.503400
EngagedRate_Log_Sq	23.177996	1	4.814353
EngagedRate_Liked_Log	20.539815	1	4.532087
ConsumerRate_Log_Sq	1.980869	1	1.407434

we can see the pairs like total impressions and the impression rare by the liked users, the engaged user rate parameters are having VIF's values greater than 20. this is because of the transformations in the same variables.

The collinearity diagnostic matrix is also saying the same story that we described above.

```
colldiag(facebook_final[,c(3:10)])
```

```
Condition
Index    Variance Decomposition Proportions
           intercept Impressions_in_100_Log_Sq ImpressionsRate_Liked_Log ReachRate_Log ReachRate_Liked_Log
1      1.000 0.000     0.000                  0.000                     0.000         0.000
2      3.353 0.000     0.003                  0.004                     0.000         0.000
3      5.241 0.000     0.003                  0.008                     0.000         0.000
4     11.708 0.004     0.003                  0.026                     0.012         0.040
5     16.661 0.001     0.118                  0.134                     0.011         0.039
6     18.957 0.001     0.028                  0.015                     0.017         0.461
7     32.914 0.004     0.064                  0.041                     0.011         0.031
8     61.539 0.140     0.191                  0.256                     0.113         0.116
9    104.539 0.850     0.591                  0.515                     0.836         0.312
```

Conclusion of model Building:

We have started from building a simple basic null model from the KPI's we decided that will influence our output total consumptions. We have partitioned the data, and built the model, we started with an R Sq. value of 33% and residual standard error value of 27 and we performed transformation on our input variables to meet the modelling assumptions. We used the method of best subset selection to decide on the final regressors and we have built model that explains almost 88% of our training data and the small residuals standard error 0.38 on our training data.

Next we fit the model on the validation data to check the accuracy of the model and we achieved nearly 84% accuracy level. This we considered as a good approximation.

Finally, we built our model on the complete data set which can be used for prediction purposes.

We can give our final regression equation to predict the consumption rate as:

Log (Consumption rate of a post) = 0.15 * Image - 0.02 * status – 0.01 * Video -0.23 * Inspiration -0.19 * Product – 0.01 * (log(total Impression's in 100's))^2 + 0.17 * log(percentage of users who liked our page and had an impression) - 0.20 * log(percentage of reach) – 0.03 * log(percent of users who liked our page and the post has reached to them) + 1.06 * log (total users who engaged with post) – 0.06 * (log(total users who engaged with the post))^2 + 0.27 * log(users who liked our page and engaged with post) – 0.05 * (log(Total number of consumers of our post))^2.

Business Interpretation of the results:

We now have successfully built a regression model based on the key indicators that we obtain from the Facebook metrics to determine which parameter will generate more user interests on our posts. So, building the model is one side of the coin and interpreting the model and taking business actions based on the model is very important part of the task.

Our model says that to increase the users interests in our posts (i.e to have high consumptions rate) the factors like: type of the post, category of the post are important. The total impressions of the post and total impression of the post on the users who liked our page are also determining the post consumption rate. Reach of the post to all the audience and reach to the people who like our page and the total engaged users and consumers of the post are all having a final effect on the audience driving the interest towards social media campaigns. Our model says that, weather the type of post is either a paid or non-paid or the total interactions (the likes, shares, comments of a post) is not going to affect the user interest into the post.

It is also important to check which factors are contributing more positively and which ones in a negative way. When compared to the post with the web link, the image posts are 0.15 times more effective and the status posts are -0.02 times less effective than web link post and video posts are -0.01 times less effective. From this we can say that the Image posts grab the attention of the users followed by the web link, video and finally status post's.
The Category of the post which is bringing more user attention is the Action type, followed by product and lastly inspirational posts.

The impression of the post on the users who liked our post is important to the company as the total impressions rate increases by 1% our post consumption increases by 0.17 times normal. If the total engagement actions which will enable to create a story on the users time line is the most effective medium for pulling the customers for each % increase in the engaged users we have 1.06 more consumptions of our post than usual.

On a concluding note, if a social media manager of the company want to take a call on which type of post he / she needs to post, in order that gets more visibility (consumptions) to the cosmetic company's brand, the understanding of the parameters from the above model will be greatly helpful.

Limitations:

Before applying the model, the consumer of this model must keep in mind that the data corruption issue we found in the data is not taken into account. We also ignored the time parameters which were available to us thinking that, the time when a post was posted will not affect the total lifetime consumption of that post. The point estimates provided by the model might not be accurate, but can fall within a given confidence interval.

Works Cited

Chatterjee, H. (2014). *Regression Analysis by Example; Fifth Edition.* Wiley.

code, E. (n.d.). *Project on Facebook Page Performance of a Cosmetic Brand*. Retrieved from RPubs: http://rpubs.com/rajiv2806/RegressionModel_FacebookMetrics_CosmeticCompany

DonKor, B. (2013, Sept 10). *Engagement-vs-consumption-the-facebook-dilemma*. Retrieved 04 01, 2017, from http://brnrd.me/engagement-vs-consumption-the-facebook-dilemma/

Joss, E. (2012). *A Beginner's Guide to Facebook Insights*. Retrieved 04 03, 2017, from blog.kissmetrics.com: https://blog.kissmetrics.com/guide-to-facebook-insights/

Execution Code Can be found here:

http://rpubs.com/rajiv2806/RegressionModel_FacebookMetrics_CosmeticCompany

https://github.com/Rajiv2806/Application_of_Linear_Regression_on_Facebook_Page_Metrics