# Regression Analysis of Facebook Metric Data of Cosmetic company

Vinay Sharma
Sonia Mehra
Rajat Agarwal
Souvik Paul
Ankit Gupta

July 2020

**Abstract**

Here , we are working with facebook metric data published in 2014 on a facebook page of a cosmetic brand. Initially we check missing observation in the data(if any found we delete those) then we split data into test and train for further analysis, the EDA( exploratory data analysis) , changing categorical variable into indicators and then move to linear regression assumptions. Next we detect Leverage/influential points with help of Cook's Distance , DFFITS , DFBETAS , COVRATIO. Then move to deal with different problems like normality( checking using plot and Shapiro-Wilk Test if found transform our response), heteroscedasticity(checking using plots of regressor vs residuals and also by Breusch Pagon test if found transformed our regressors) , multicollinearity(check using Condition Number and Variance Inflation Factor if found we deal with different techniques like Principal Component Regression(PCR),variable elimination,Partial Least Square Regression (PLS) and lastly the LASSO using cross validation). LASSO also gives us a way for variable selection as well as a way to remove multicollinearity . Finding MAPE in each of model allows us to compare each models to get the model with best predicting power. Finally we Checking for Significance of first level of each Categorical Variable and correct all with suitable remedies to get a model with minimum MAPE.

**Key words:** MAPE , LASSO , Breusch Pagon test , Cross Validation , VIF , Multicollinearity , Influential , Cook's Distance , EDA , Regression , PLS , PCR.

# 1 Acknowledgement

We would like to express our heartfelt gratitude to Dr. Minerva Mukhopadhyay for helping us in every difficulty which we face during this project. It has been a great learning experience and has also provided us with a practical insight of the theoretical knowledge gathered during the course Regression Analysis.

We also take this opportunity to thank the authors and publishers of the various books and journals we have consulted. Without those this work would not have been completed.

We would also like to thank our parents for their extensive support throughout the session. Their constant encouragement has enabled us to complete the project within the stipulated time-period.

# Contents

## 2 Introduction

Social media marketing is marketing of any product via social media like Facebook , WhatsApp ,Instagram e.t.c. It allows individuals , businesses and other organizations to interact with one another on a single online platform. A major advantage is there direct interactions which helps them to build relationship and communities online .
Although , business retailers have gained over 133% of there revenues from social media marketing . From the recent data there are more than 3 billion active user on social media including Facebook , Instagram , WhatsApp e.t.c. Any company marketing strategies must include one important part of social media marketing . As it is the easiest and fastest way to interact directly with the consumer of a company . It helps company to increase their sales and helps in their overall development.
Here we are working on Facebook metric data collected from a post published on a Facebook page of renowned cosmetic brand . The company want to increase the number of lifetime engaged user with company . While modeling the data with Linear Regression we have encountered the problem of normality, heteroscedasticity ,multicollinearity and fixed all of them with suitable remedies .

## 3 Objective

Our objective is to increase the number of lifetime engaged user . So it will benefit the company in overall development. Keeping focus on features our aim is to maximize the number of users(lifetime engaged).So we need to know,

Which month the brand post more so as to increase lifetime engaged users?

Which type of post engages mostly users?

Which time of the day engaged users?

Do weekday ,paid or unpaid post also play a role in lifetime engaging of users?

What other variables one focus on to increase lifetime engaged users?

# 4 MAPE

We will use Mean Absolute Percentage Error (MAPE) to check how good our model accuracy is and MAPE also will be used for comparing the models. formula for MAPE is as :-

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\ \frac{A_i - F_i}{A_i}\ \right|$$

Where $A_i$ is the actual value and $F_i$ is the forecast value. Multiplying it by 100 gives percentage.

# 5 Data Description

**Title of the Data Set :-** Facebook Metrics Data Set.

**Source :-** This data Set is created by SÃ©rgio Moro, Paulo Rita and Bernardo Vala (ISCTE-IUL) @ 2016.
**http://archive.ics.uci.edu/ml/datasets/Facebook+metrics**

**Data Set Information :-** The data is related to posts published during the year of 2014 on the Facebook's page of a renowned cosmetics brand. This is a multivariate Data Set which contains **500 Instances** and **19 Attributes**

We found that few column names are too large to handle, so, need to shorten them. Now see the summary of our data at a glance -

**Information About Attributes :-** We see that there are 19 Attributes and from the above picture this implies both categorical and numeric variables are present in this data and clearly variable **"Type,Category,Post.Month,Post.Weekday,Post.Hour,Paid"** is categorical and others are **"Numeric Integer"**
**(1) Page Total Likes** :- Total no of pages likes.
**(2) Type** :- Factor:{Photo,Status,Video,Link}
**(3) Category** :- Factor:{action,product,inspiration}
**(4) Post.month** :- Factor
**(5) Post.weekday** :- Factor
**(6) Post.hour** :- Factor
**(7) Paid** :- Factor{1-yes,0-no}
**(8)Lifetime Post total Reach**
**(9) Lifetime post total impressions** :- No of times a post from page is displayed.
**(10) Lifetime engaged users** :- No of people who clicked anywhere in a post(unique users).
**(11) Lifetime post consumers** :- The number of people who clicked anywhere in a post.
**(12) Lifetime post consumption** :- The number of clicks anywhere in a post.
**(13) Lifetime post impressions by people who have liked your page** :- Total number of impressions just from people who have liked a page
**(14) Lifetime post reach by people who like your page** :- The number of people who saw a page post because they have liked that page.
**(15) Lifetime people who have liked your page and engaged with your post** :- The number of people who have liked a Page and clicked anywhere in a post.
**(16) Comment** :-Number of comments on the publication.
**(17) Like** :- Number of "Likes" on the publication.
**(18) share** :- Number of times the publication was shared.
**(19) total interaction** :-The sum of "likes", "comments", and "shares" of the post.
**"Lifetime engaged users"**,The number of people who clicked anywhere in a post(unique users) is taken as **response variable (Y)**.

# 6 Data Pre-Processing

Let's check out more information and summary statistics of our data including missing values, histograms etc using skim() function.

### 6.1 Missing Value

We found that there are 6 missing value - 1 in 'Paid', 1 in 'like' and 4 missing values in 'share'. We are not bother with the last 5 missing values as later we will exclude the columns 'like', 'comments', 'share' as these column sums is represented by the column 'Total.interaction'. So, now we only exclude the missing values in 'Paid'.

Thus we have no missing value (except in 'like' and 'share' which will be omitted later) now.

### 6.2 Categorical Variable

We found out some categorical variables but few of them were in integer type. Now convert all the categorical variables into factors.

### 6.3 Splitting Data into Train and Test

Now we have 499 observations. For checking the model efficiency we need to divide the data into train data and test data such that we check the performance of our model (built using train data) on test data (prediction on test data). We will split the data into 4:1 ratio that means train data consists of 80% observation and test data will contain 20% observations. We will omit the columns 'like', 'comments' and 'share' from both the data and make 'Lifetime Engaged users' as our response.

Now we will apply every Regression Technique on Train data and check goodness of the model on the prediction of Test data.

## 7 Exploratory Data Analysis

Now we again see data summary on the train data after pre-processing. We store the index of categorical variables in **'cat_col'**.

We find that there is no missing values now. There 6 categorical variables and rest are numerical. Now we run separate analysis for categorical and numerical variables.

## 7.1 Analysing Categorical Variable using Box-Plot And frequency table

### 7.1.1 Type

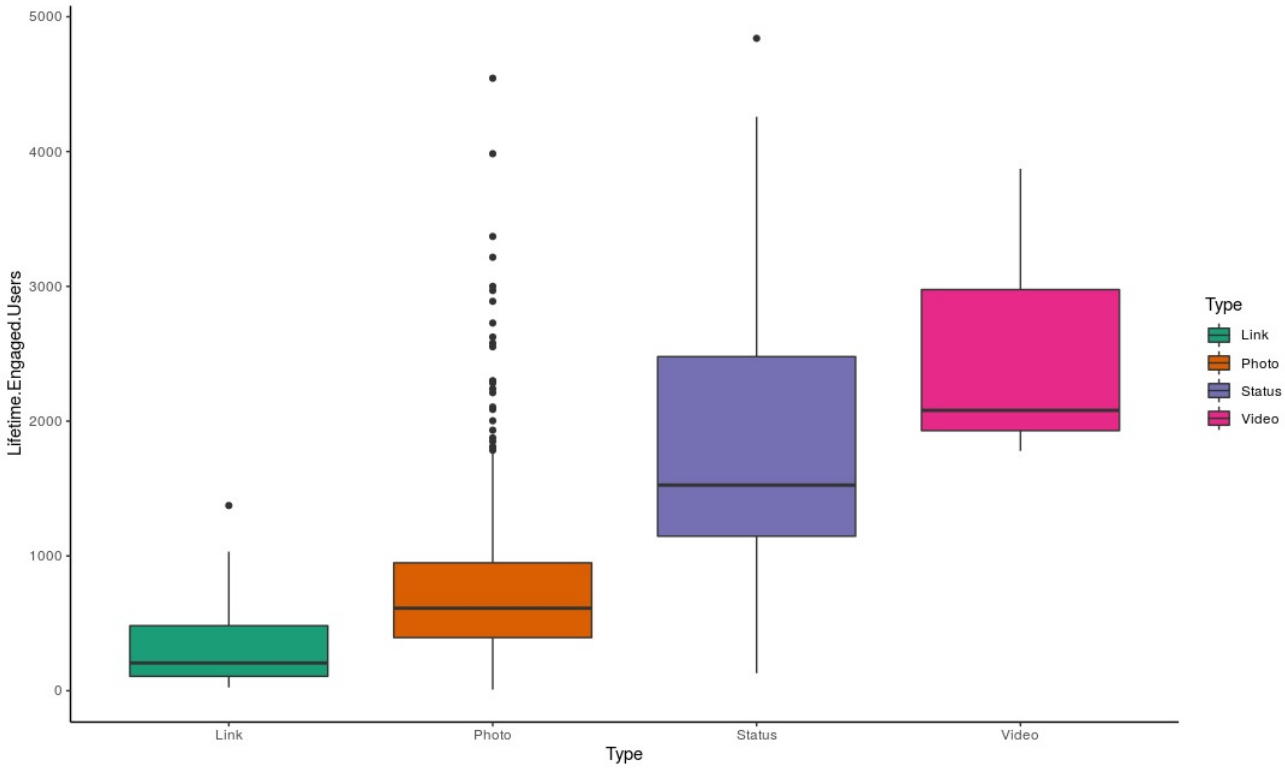


Figure 1: Distribution of Lifetime.Engaged.Users by Type of Post

It has four label namely **'Link','Photo','Status','Video'** Clearly 'Photo' is much more Frequent (345) than other categories. 'Video' has much higher central tendency then comes 'Status' . 'Status' has higher variability and 'Link' has lowest Central tendency and lowest variability with respect to Lifetime.Engaged.Users.All the types here have long tail that implies positive skewness.

### 7.1.2 Category

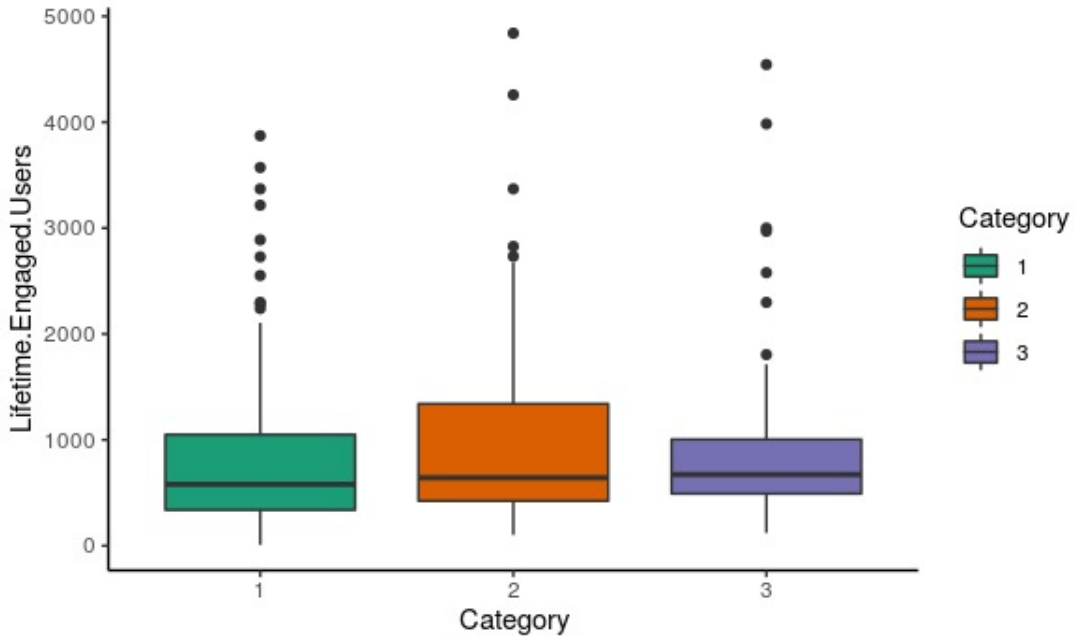


Figure 2: Distribution of Lifetime.Engaged.Users by Category of Post

Variable 'Category' has three label **'1','2','3'**. Label 1 is most frequent(174). In Fig.12 all three label have almost same central tendency and label 2 has slightly more variability. All three Categories are positively skewed.

### 7.1.3 Paid

Variable 'Paid' has two label **'0','1'**. Label '0' is most frequent(301). Both labels have almost same central tendency . Both have too much outliers but their variabilities is are not high. Both labels have positive skewness.

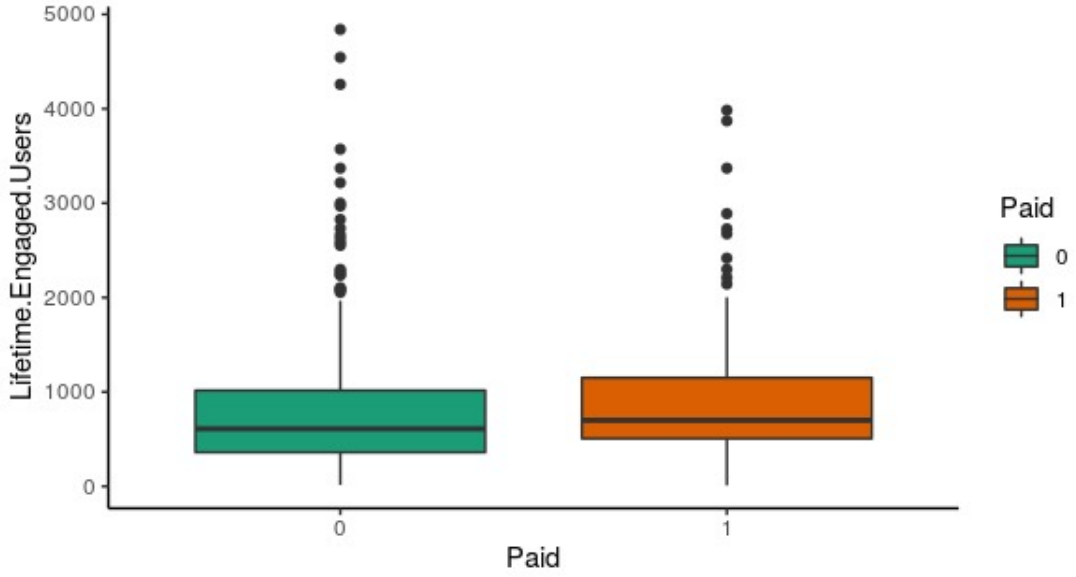


Figure 3: Distribution of Lifetime.Engaged.Users by Paid scenario of Post

### 7.1.4 Post Month

Variable Post.Month has 12 label as **'1','2'...'12'**. Which represent all 12 months of a year. Month October (10th month) is most frequent(50) i.e it has most lifetime engaged users, 'Jun' and 'Dec' are second highest .'Dec' has lowest central tendency where 'Dec' has highest variability and 'May' has lowest variability.

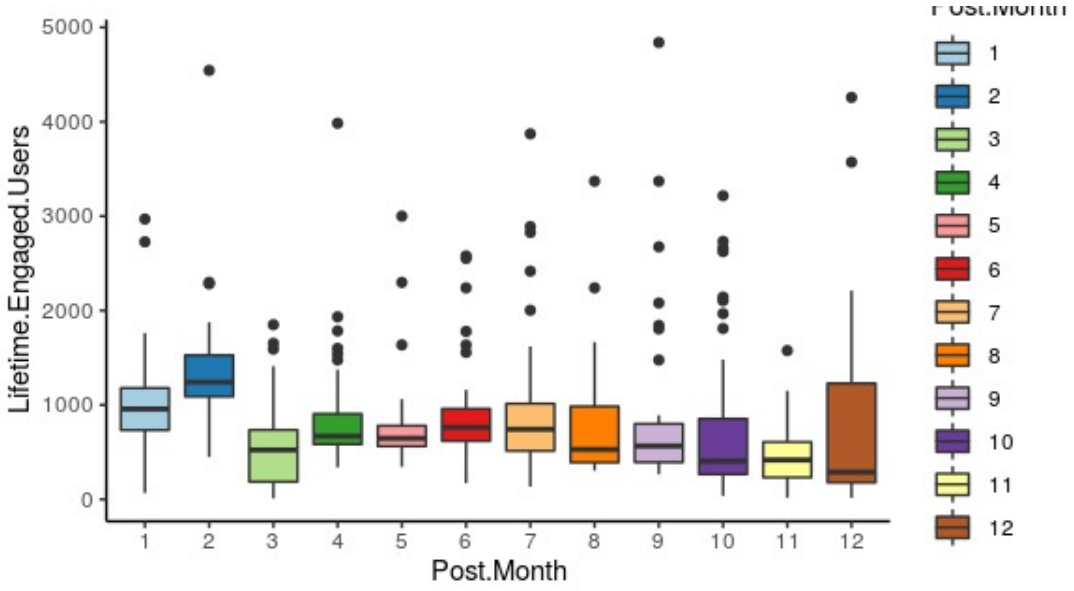


Figure 4: Distribution of Lifetime.Engaged.Users by Post.Month

### 7.1.5 Post Hour

Variable Post.Hour has 23 label as **'1','2'...'23'**. Which represent hours of a day. Third hour is most frequent(85). Hour 6,10,12,14 have more variability in comparison to others.Hours 3,5,6,7,9,10,13,14 are positively skewed and 1,8,17 are negatively skewed

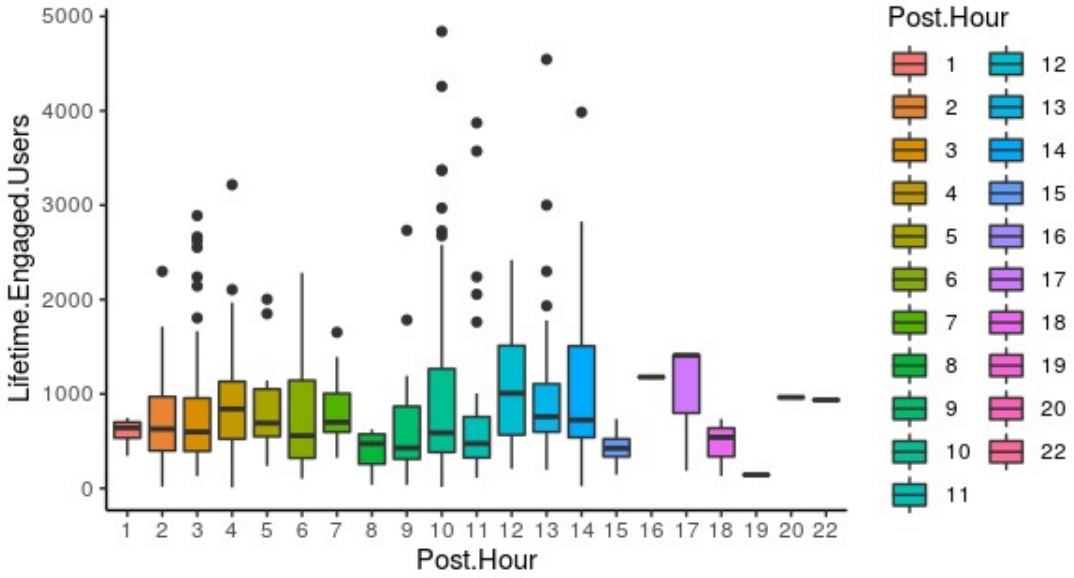


Figure 5: Distribution of Lifetime.Engaged.Users by Post.Hour

### 7.1.6 Post Weekday

Variable Post.Weekday has 7 label as **'1','2'...'7'**. Which represent days of a week. 6th day and 7th day are most frequent(68 and 63 frequency respectively).All days have almost same central tendency .All days have higher number of outliers.

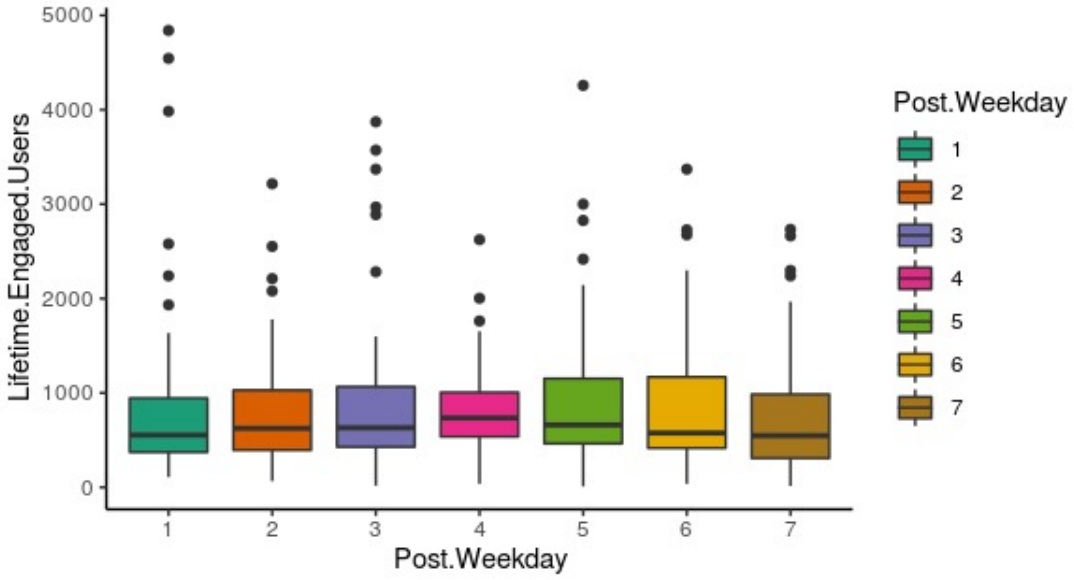


Figure 6: Distribution of Lifetime.Engaged.Users by Post.Weekday

## 7.2 Analysing Numerical Variable using Pair-Plot and correlation matrix

Pairs plot is as follows:

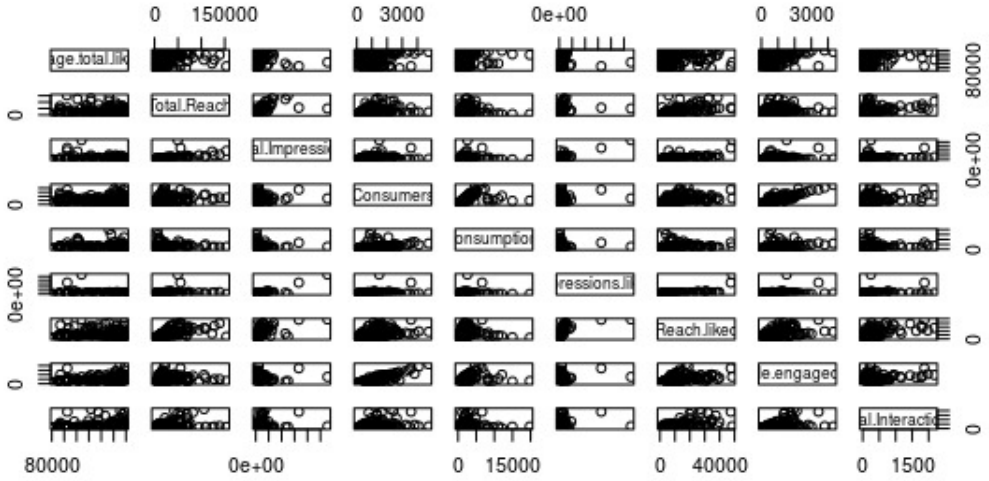


Figure 7: Pair plot among numerical variable

From pairs plot(Fig.21) and correlation matrix. one can infer that there is few strong linear relations among the variables :

1. Lifetime.Post.Total.Reach & Lifetime.Post.Total.Impressions(0.65)

2. Lifetime.Post.Total.Reach & Lifetime.Post.reach.by.people.who.liked.your.page(0.74)
3. Lifetime.Post.Total.Impressions & Lifetime.Post.Impressions.by.people.who.liked.your.page(0.87)
4. Lifetime.Post.reach.by.people.who.liked.your.page & Lifetime.Post.Total.Impressions(0.65),
5. Lifetime.Post.Consumers & Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post(0.91)
6. Lifetime.Post.reach.by.people.who.liked.your.page & Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post(0.66)

Again From Fig.22 one can infer that response is correlated with:

1. Lifetime.Post.Consumers(0.97)
2. Lifetime.Post.reach.by.people.who.liked.your.page(0.69)
3. Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post(0.92).

# 8 Changing categorical variable into Indicator variable and Scaling Numerical Quantity.

It is clear from Fig 23 that there are 6 categorical variable (denoted by length of "cat_col" variable). After changing data into categorical variables we have 53 predictors in which 9 are continuous numerical variables (present in first 9 column of X) and then perform scaling on these variables.
p: no of predictors=53
n:no of observations=399

# 9 Primary Model

Consider our multiple linear regression model as :

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \quad .....(1)$$

Where $\mathbf{X}_{n\times p+1} = (\mathbf{1}, x_1, x_2, \cdots, x_{53})$ is the non-stochastic matrix of predictors (here $\mathbf{1}$ is the column whose all elements are 1) and $\beta = (\beta_0, \beta_1, \cdots, \beta_{53})$ is the vector of regression coefficients and $\epsilon = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)$ is the error vector.

## 9.1 Assumptions

(i) $\epsilon_i \overset{iid}{\sim} \mathbf{N}(\mathbf{0}, \sigma^{\mathbf{2}})\ \forall i$ , where $\sigma^2$ is an unknown constant.
(ii) $\mathbf{X}$ is of full column rank

Let's see our first model and its performance -

So, performance of this model is quite well as MAPE is only 0.046 or 4.6%

# 10 Diagnostics for Leverage/ Influential Points

## 10.1 Leverage Points

Leverage points are those which have unusual x values but does not affect the model estimates much. Leverage points can be find out with the help of Hat matrix(H), where Hat matrix is given as:

$$\mathbf{H} = \mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{X^T}$$

if $\mathbf{h_{ii}}$ is ith diagonal element of Hat matrix, then traditionally ith point is considered as leverage if

$$\mathbf{h}_{ii} > \frac{\mathbf{2 * p}}{\mathbf{n}} \quad .....(a)$$

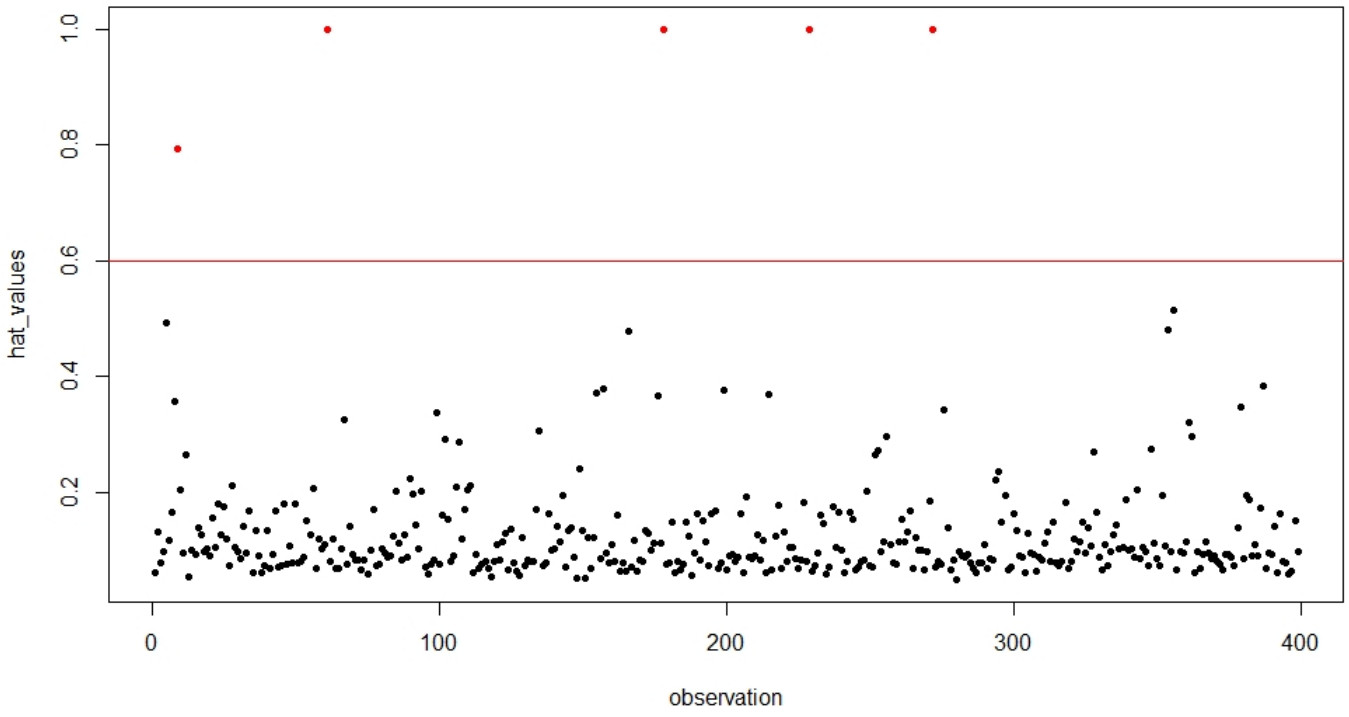


Figure 8: Leverage point detection

If cut-off is considered as given above( which is given in equation (a)) then too many points become leverage,which is clear from fig 22.So to overcome this problem, cut-off is taken to be 0.6, thus yielding 5 leverage points.

Now we delete the leverage points from the data, thus new data is formed. As there are columns that contain only '0' and '1', so, by deleting rows (or observations) it may

happen that few columns is left with '0' only. Then the **X** matrix becomes rank deficient. So, we check those columns and delete them.
Then finally we form a new model without leverage points, then we check its performance.

New model performs well with MAPE 4.7% but slightly worse than previous model.

## 10.2 Diagnostics for Influential points

Influential points are those which have both X ad Y unusual.These points have considerable influence on estimates of coefficients and these pull the direction of regression line towards itself.
Our aim is to compute $Cook'sdistance$ ,$DFFITs_i$, $DFBETAs_{ji}$, $CovRatio$ and identify the influential point .

### 10.2.1 Cook's Distance

Cook's distance is one of the measure used for identification of Influential Points. Cook's distance is calculated as :

$$\mathbf{C}_i = \frac{(\hat{\beta} - \hat{\beta}_{(\mathbf{i})})^T (\mathbf{X^T X})(\hat{\beta} - \hat{\beta}_{(\mathbf{i})})}{\mathbf{pMSres}}$$

$$\mathbf{C}_i = \frac{r_i^2 * h_{ii}}{1 - h_{ii}}$$

where,p is number of regressors, n is total no. of observations . $r_i = \frac{e_i}{\sqrt{MSres(1-h_{ii})}}$is the internal studentized residual and $e_i$ is the ith residuals.
It is clear from formula that it contain both part $r_i$(studentized residual) and $h_{ii}$(ith diagonal element of hat matrix) which measure how far observation is from data and how well the model fits the ith observation $y_i$

Traditionally, the points for which $\mathbf{C_i}$ is greater than 1 is consider influential. But,in this case 0.1 considered as cut-off point and it is clear from Fig.29,30 that 5 points are influential points

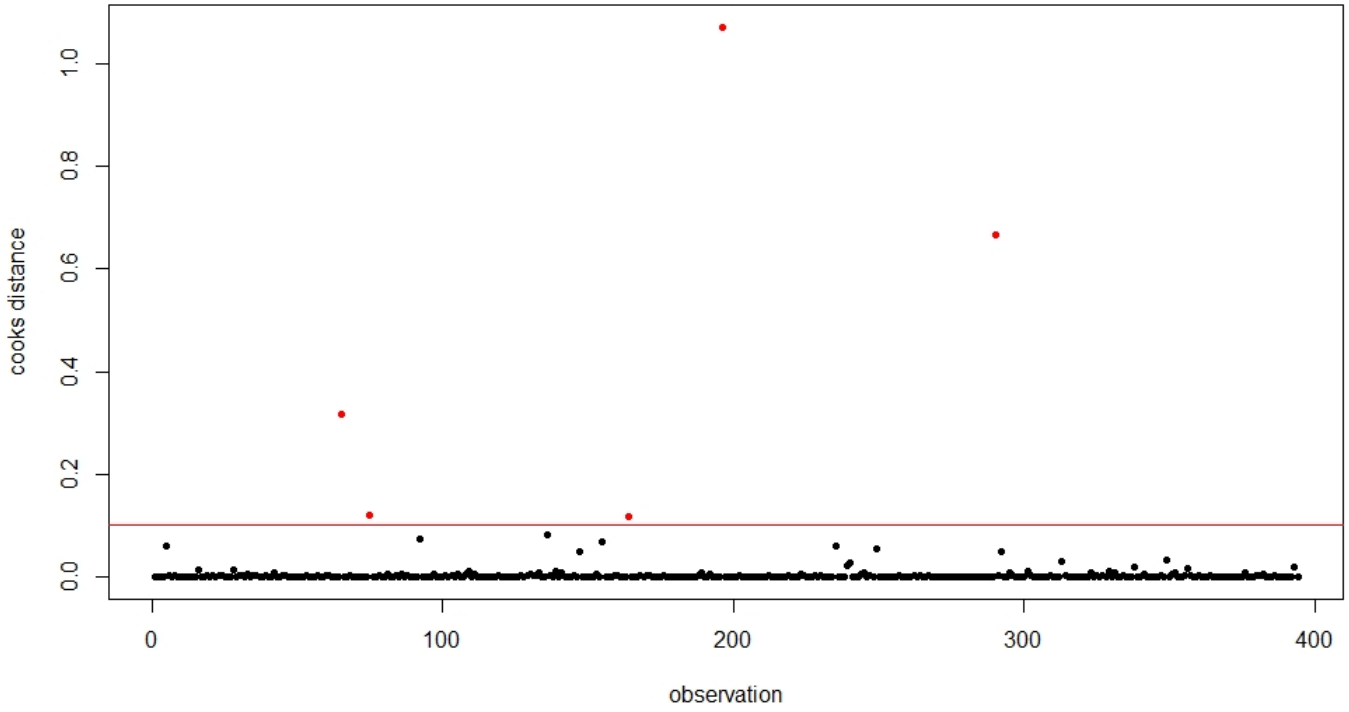


Figure 9: Cook's distance representation when threshold=0.1

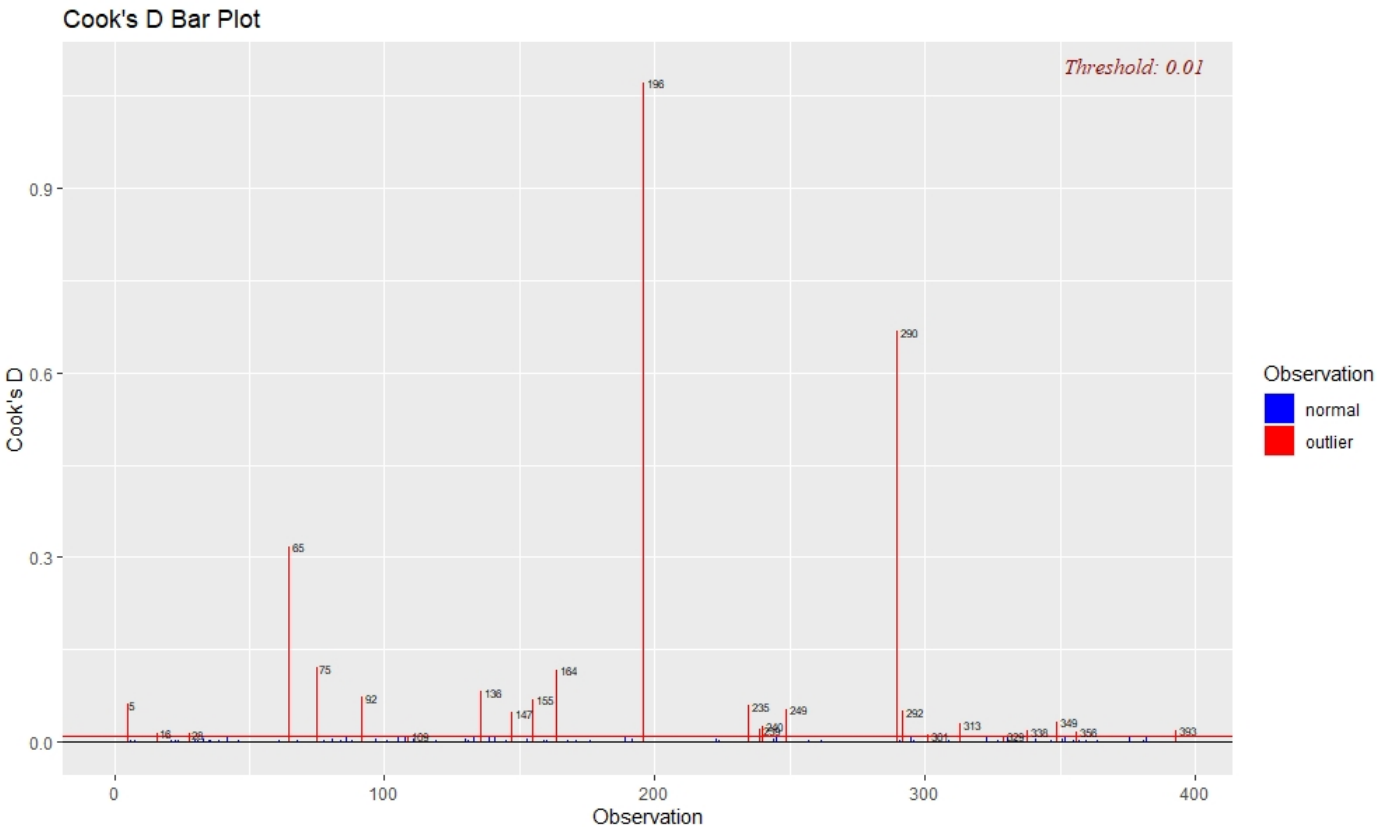


Figure 10: Cook's distance Bar plot when threshold=0.01

### 10.2.2 DFFITS

DFFITS is another method for influential point detection and it also contain both part ($r_i^{*2}$ and $h_{ii}$) which measure how far observation is from data and how well the model fits the ith observation $y_i$ $DFFIts_i$ is calculated as :

$$\mathbf{DFFITs}_i = \sqrt{\frac{r_i^{*2} * h_{ii}}{1 - h_{ii}}}$$

$r_i^* = \frac{e_i}{\sqrt{MSres_i(1-h_{ii})}}$is the external studentized residual. $e_i$ is the ith residuals. $MSres_i$=

$$\frac{\sum_{j\neg=i}^{n} \mathbf{e}_j^2}{(\mathbf{n-1-p})}$$

$\mathbf{h}_{ii}$ is the ith diagonal element of matrix H=$\mathbf{X}(\mathrm{X}^T X)^{-1} X^T$

Traditionally cut-off is given by :

$$|\mathbf{DFFITs}|_i > \frac{\mathbf{2} * \sqrt{p}}{\sqrt{n}}$$

and if $DFFITS_i$ is greater than this cut-off that indicates that ith sample point is influential point. In this case to avoid too much points to be influential, 2.5 is consider as threshold.

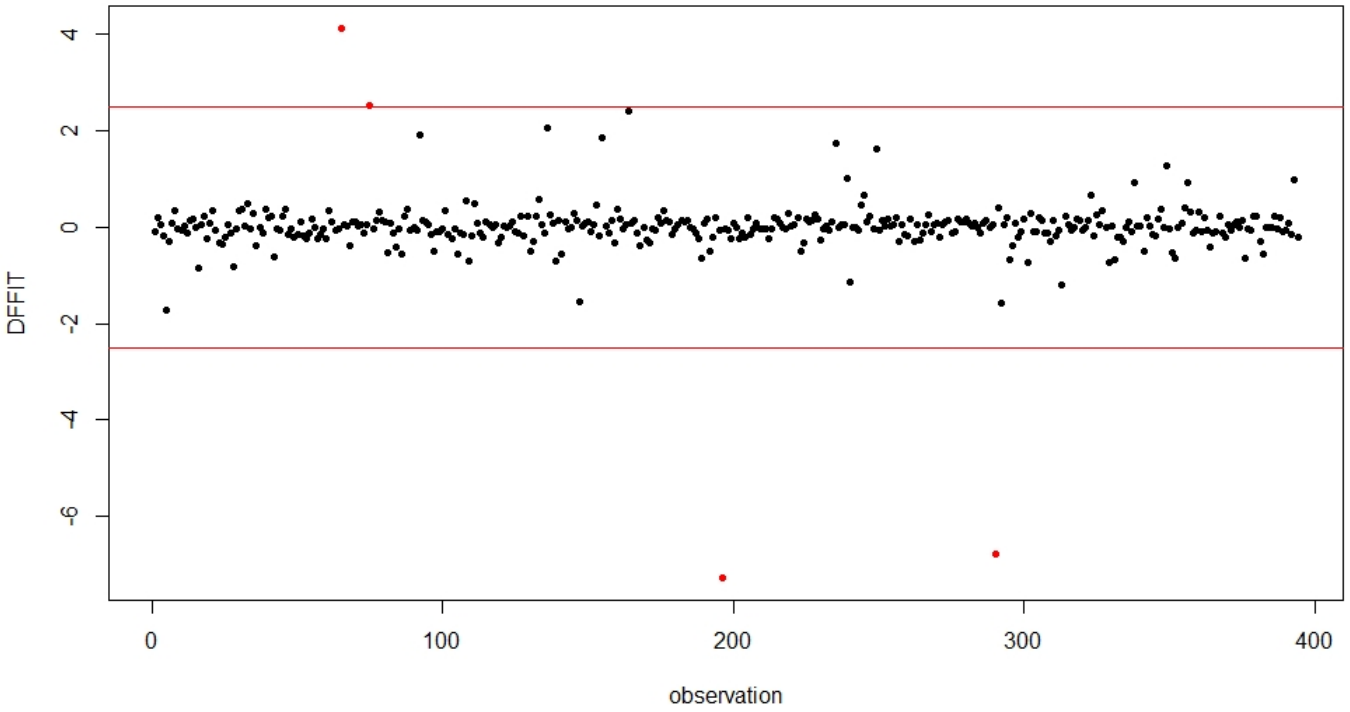


Figure 11: DFFITS representation when threshold=2.5

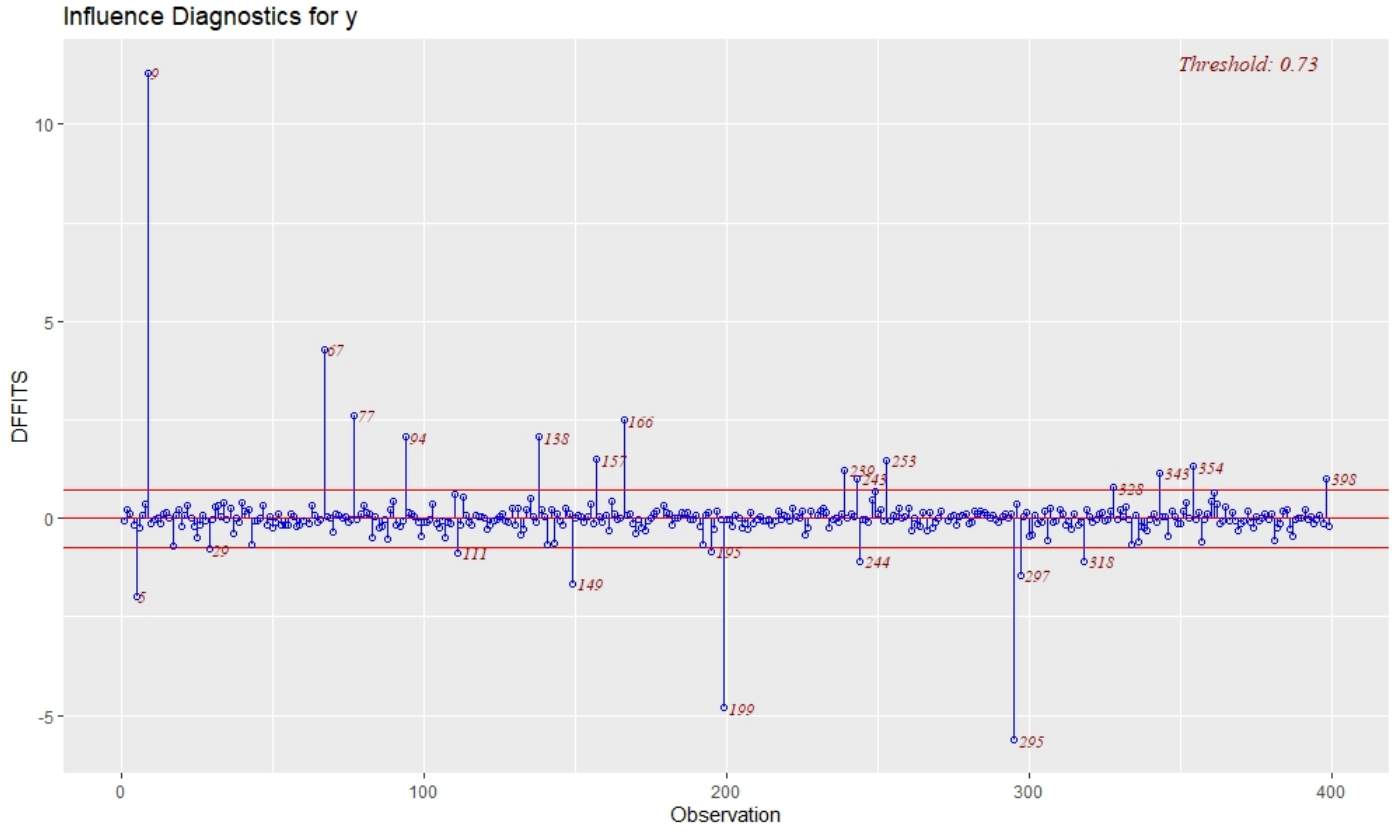


Figure 12: DFFITS representation when threshold=0.73

### 10.2.3 DFBETAS

It is also used for influential point detection and it also contain both part ($r_i^{*2}$ and $h_{ii}$) which measure how far observation is from data and how well the model fits the ith observation $y_i$, which is clear from formula of $DFBETAs_{ji}$ $DFBETAs_{ji}$ is calculated as:

$$\mathbf{DFBETAs}_{ji} = \frac{r_i^* * C_j}{\sqrt{h_{jj}(1-h_{ii})}}$$

$r_i^*$ is the external studentized residual. $\mathbf{h}_{ii}$ , $\mathbf{h}_{jj}$is the ith,jth diagonal element of matrix H=$\mathbf{X}(\mathrm{X}^T X)^{-1}X^T$ respectively and $C_{ji}$ is the is $j^{th}$ row of the matrix C= $((\mathbf{X^T X})^{-1}\mathbf{X^T}$

Generally threshold is given as for this case

$$\mathbf{DFBETAs}_{ji} > \frac{\mathbf{2}}{\sqrt{n}}$$

And if $DFBETAS_{ji}$ exceeds this cut off than we say that ith observation is influential.From Fig.34 it is clear that only those points are considered to be influential points which are considered to be influential by 7 or more regressors out of 9 numerical regressors which are tested for DFBETA

### 10.2.4 COVRATIO

The $Cook'sdistance$,$DFFITS_i$,$DFBETAS_{j,i}$ provides insight about the effect of observation on estimates of coefficients and fitted value y but they do not provide precision of estimation. COVRATIO also used for this purpose.We express the role of ith observation on precision of estimates in terms of $COVRATIO_i$.
We define $COVRATIO_i$ as

$$\mathbf{covratio_i} = \frac{(MSres_i)^p}{(MSres)^p * (1-h_{ii})}$$

$MSres_i$ is the esimate of the variance of response variable when ith data point is removed from the data.$MSres$ is the esimate of the variance of response variable using all the data. $\mathbf{h}_{ii}$ is the ith diagonal element of matrix H=$\mathbf{X}(\mathrm{X}^T X)^{-1}X^T$ .

If $COVRATIO_i > 1$ then ith observation improves the precision of estimates and if $COVRATIO_i$ <1 then ith observation degrades precision. Traditionally, Cut-off for COVRATIOS is defined as

$$|\mathbf{covratio_i - 1}| > \frac{\mathbf{3 * p}}{n}$$

There is one influential point using this approach

Overall there are 5 influential points (selected by at least 2 methods of finding influentials) and after removing influential points from data we check if there is any columns whose sum is zero and found none. After deleting leverage and influential points we have now 389 observations are left.

Now with new data we build a new model and check its performance.

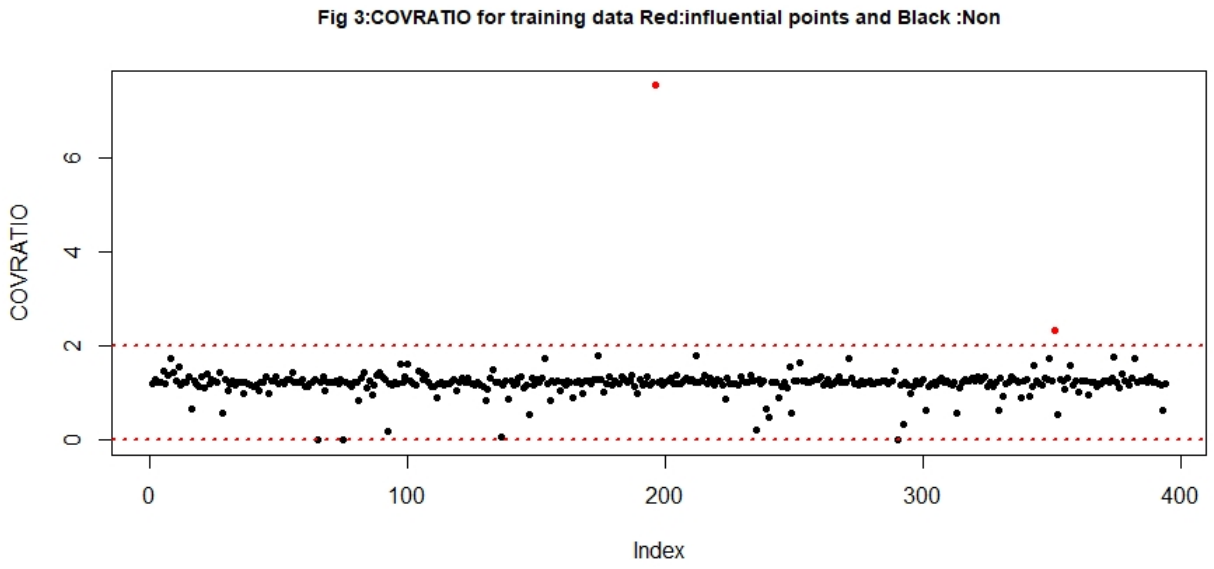


Figure 13: COVRATIO representation

Its MAPE is 5.1% which implies it is very good model but worse than previous ones. That means There is no effect of Outliers in the model w.r.t. to this train and test data.

# 11 Normality Checking

In section 7.1 we assumed $\epsilon_i \overset{iid}{\sim} \mathbf{N}(\mathbf{0}, \sigma^{\mathbf{2}})$ which is equivalent to assume $\mathbf{y} \sim \mathbf{N}_n(X\beta, \sigma^2 I_n)$. Now as y is observed quantity, we will check only its Normality.

## 11.1 Checking Through Plot

We will check Normality by Density curve of $\mathbf{y}$ and by Normal Q-Q Plot. In Density Curve if the curve looks like bell shaped like normal curve , then normal assumption is appropriate. On the other hand in Normal Q-Q Plot order statistics of vector of interest are drawn in the y-axis corresponding to the theoretical order statistics from $\mathbf{N}(0, 1)$ in the x-axis. Normality assumption can be regarded as true if the points almost in a straight line.

## 11.2 Checking Through Test

Shapiro-Wilk Test is used to test the Normality. Suppose we want to check the normality of the vector $y = (y_1, y_2, \cdots, y_n)$. Then the corresponding test statistic is

$$W = \frac{(\sum_{i=1}^{n} a_i y_{(i)})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where $(a_1, a_2, \cdots, a_n) = \frac{m^T V^{-1}}{||V^{-1}m||}$, V is the covariance matrix of the order statistic $y_{(i)}$ and m is the mean vector of $y_{(i)}$.
We check run test using p-value.

## 11.3 Calculations

At first we check that distribution of **y** is not normal by following graphs and test:

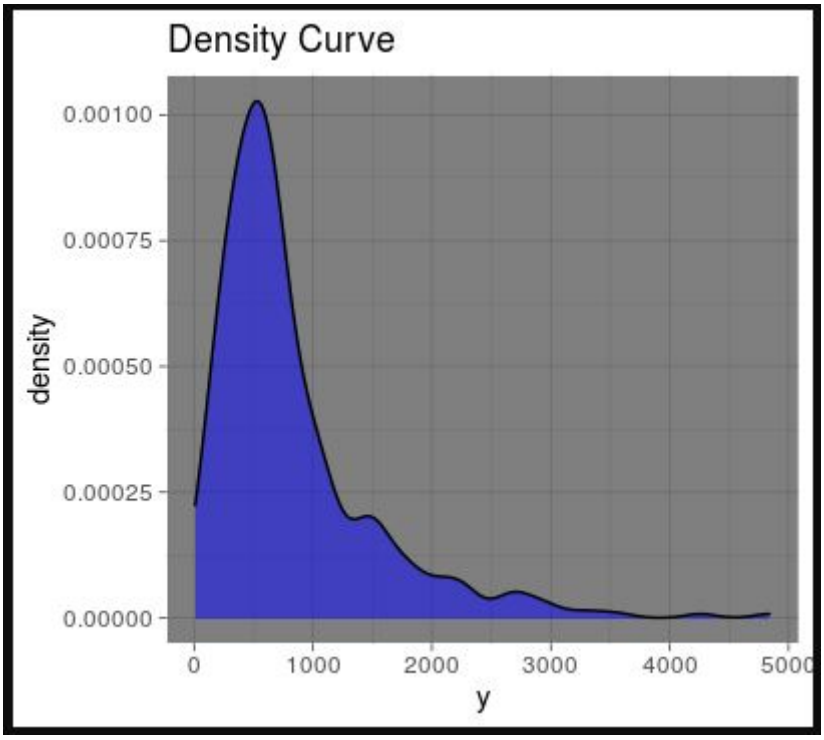


Figure 14: Density Plot

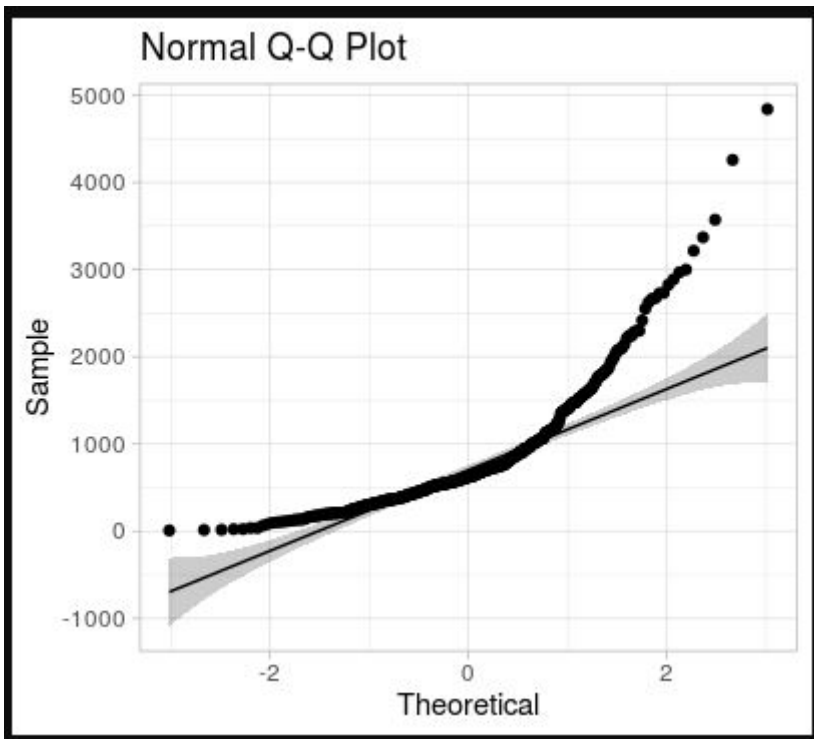


Figure 15: Q-Q Plot for y

From the test p-value came as less than $2.2\times10^{-16}$ (very very low). So,We need to do transformation of y. We see that density plot is positively skewed and also the Q-Q plot is showing deviation from normality behaviour of response .

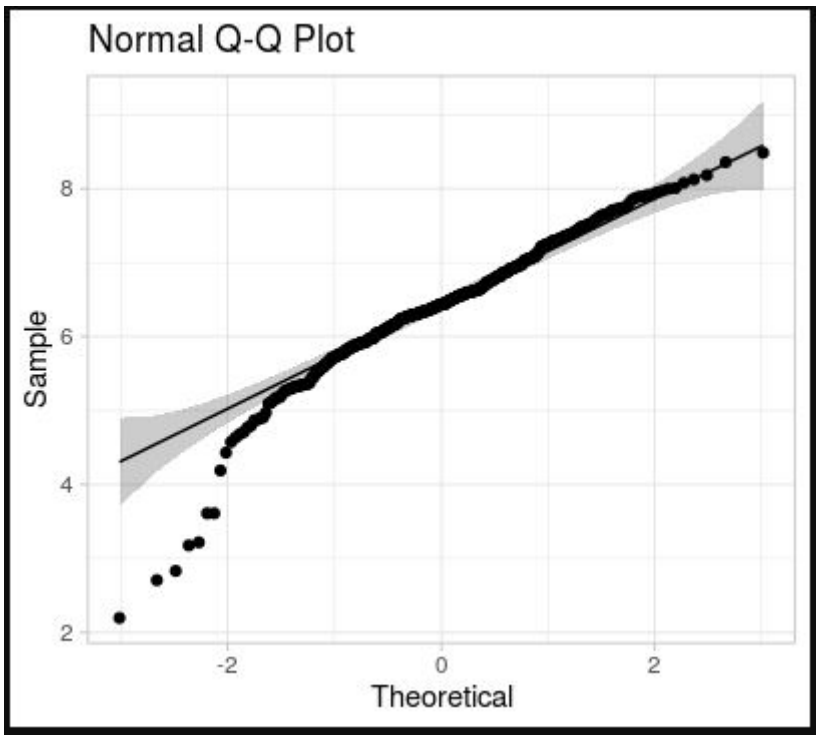


Figure 16: Q-Q Plot for log(y)

Here we see that the Q-Q plot is almost appropriate except for the first few points. But Shapiro test tells that still it is non normal as the p-value=$1.28\times10^{-16}$ is very small. Now we will check excluding some lower order statistic in the begining.

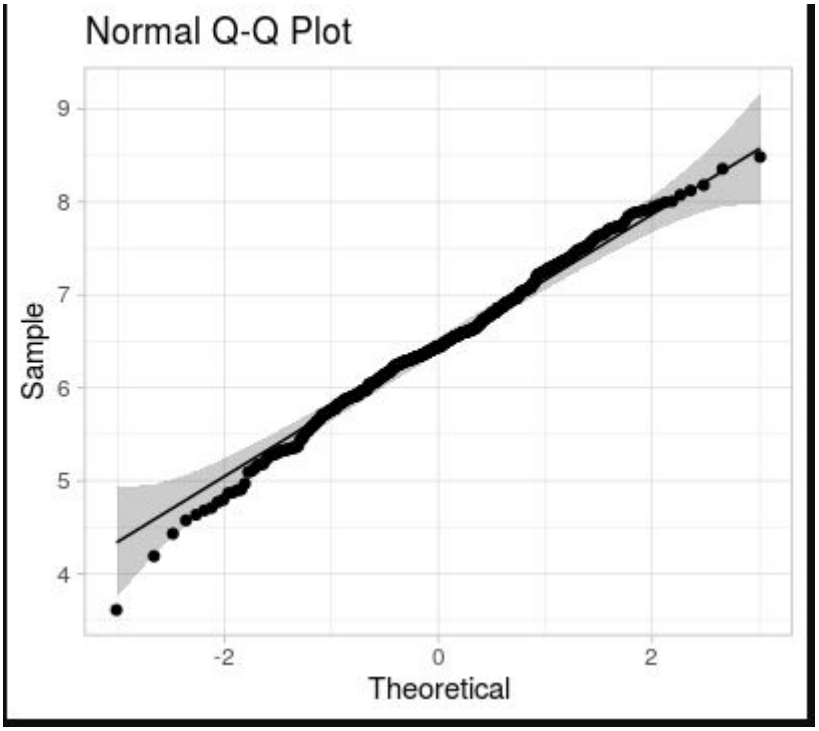


Figure 17: Q-Q Plot for log(y)

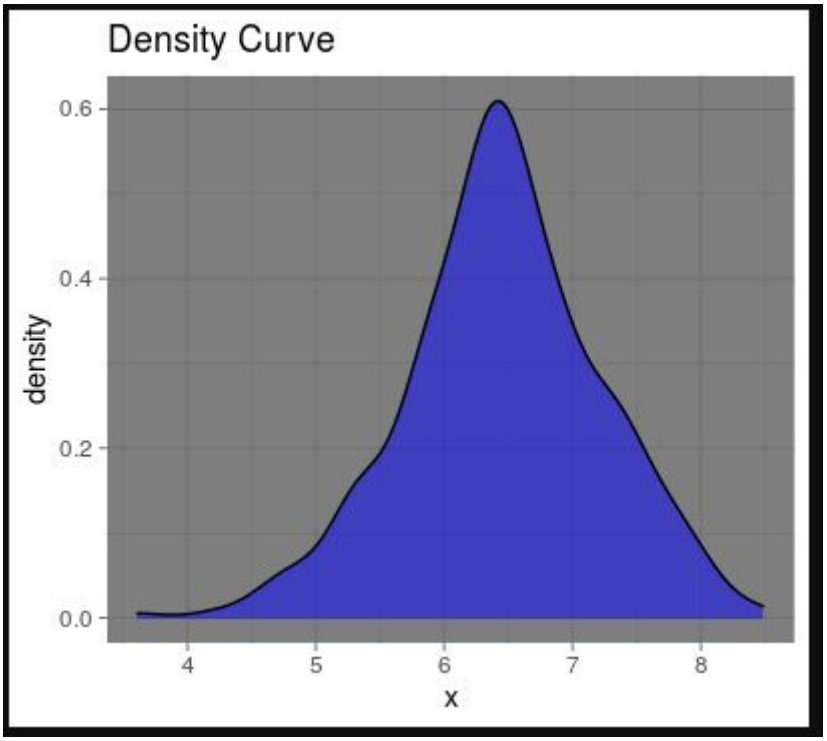


Figure 18: Density Plot for log(y)

Here we see an impressive result through the graphs and Shapiro-Wilk Test. P-value is 0.1075 >0.05. So, we can assume normality of log(y) at 5% level of significance.
Thus our new response variable becomes z=log(y) and number of observations reduce to (389-6)=383. But instead of 383 we have 382 observation as we have removed 7 observation because of a common observtion i.e. 5 distinct and 2 observation with same value that is z[217]=3.610318=z[368].
Now with new data (transformed response), new model is built and 876% MAPE is found which implies this model's prediction power is very bad. We should look for betterment of it.

### 11.4 New Model

Now we assume our new model as:

$$\mathbf{z} = \mathbf{X}\beta + \epsilon.....(2)$$

Where $\mathbf{X}, \beta, \epsilon$ are same as in (1). And assumptions are also same as in 7.1.

## 12 Heteroscedasticity

One of the assumption of our regression model is that our data should be homoscedastic. Consider the regression equation $y_i = \beta x_i + \epsilon_i$. i=1(1)n
In this model, if $\epsilon_i$ has constant variance say $\sigma^2$ than data is said to be homoscedastic and if variance of $\epsilon_i$ depends on i then data is said to be heteroscedastic and it is said that heteroscedasticity is present in data.

### 12.1 Heteroscedasticity Diagnostics

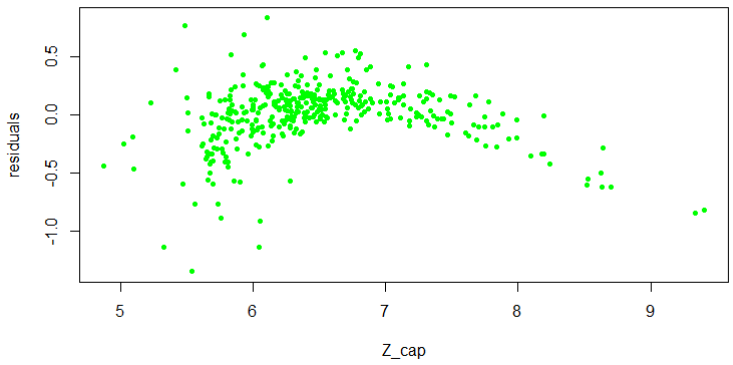


Figure 19: Graph between predicted value and residuals

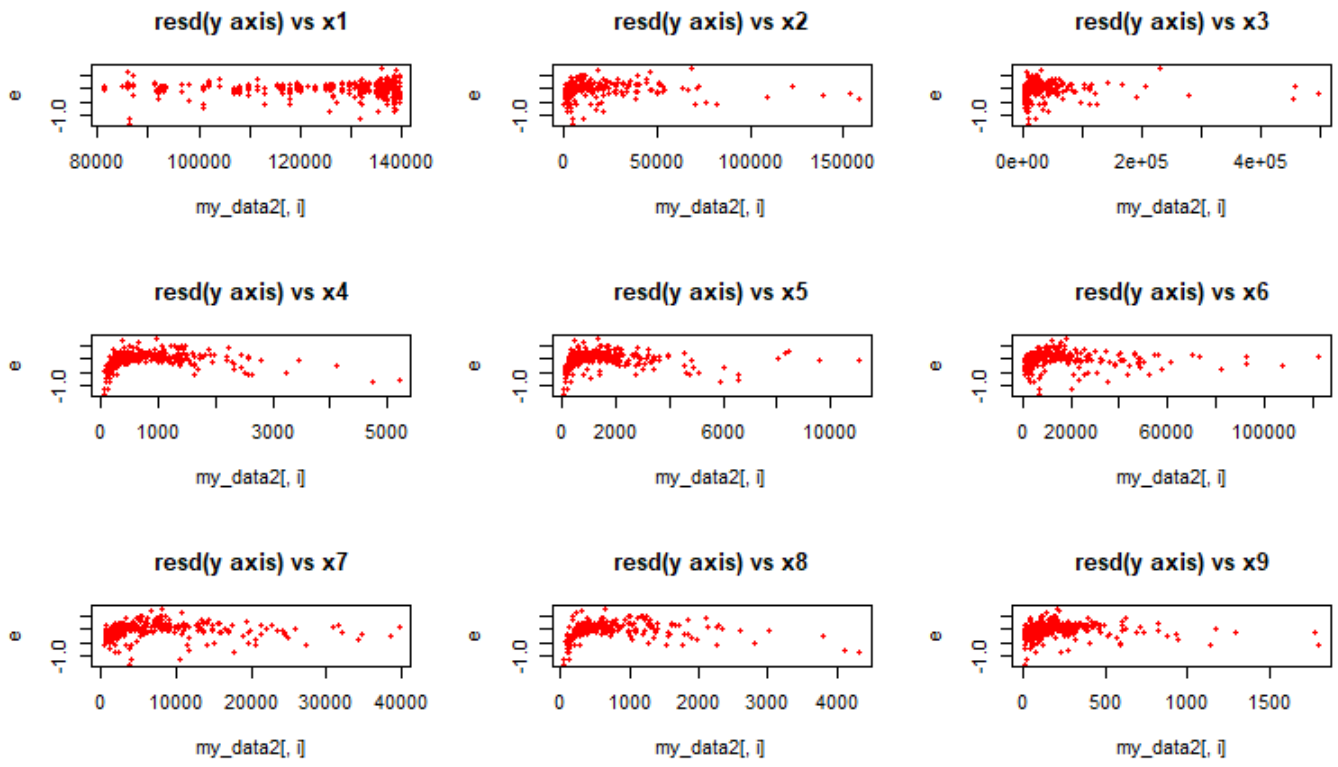


Figure 20: Graph between regressors and residuals

Graph between regressors and residuals have pattern and the graph between residuals and fitted values has also pattern (concave shape) which is indication of heteroscedasticity.
For confirming Heteroscedasticity, we also perform Breusch Pagon test (BP test). This is the test which is used for testing heteroscedasticity. After performing this test,
**p value comes out 3.147e-05 which is very low.** This confirm that heteroscedasticity is present in model.

For removal of Heteroscedasticity, We transformed the data.For doing so, all numerical variable of X transformed by log.(response has already transformed into log).

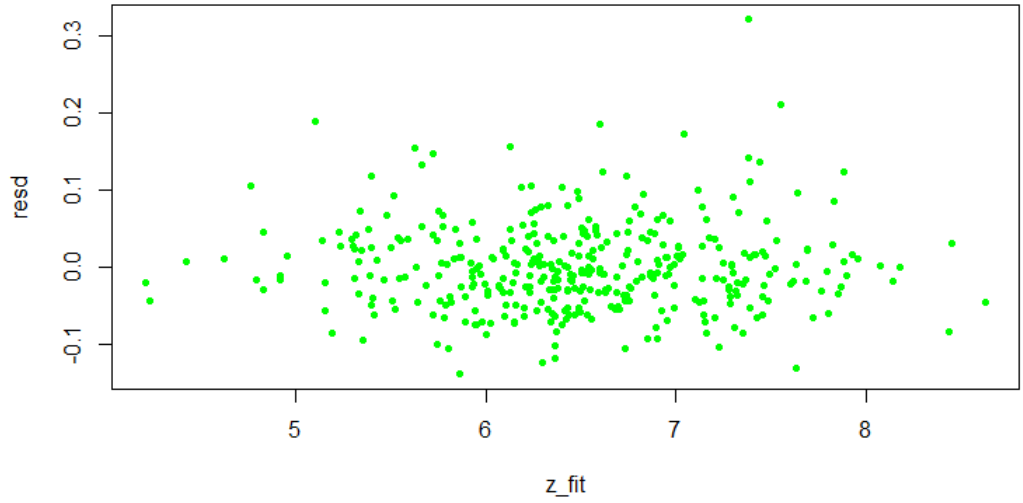


Figure 21: graph between predicted value and residuals

From these graphs it can be seen that there is no pattern in graph between predicted

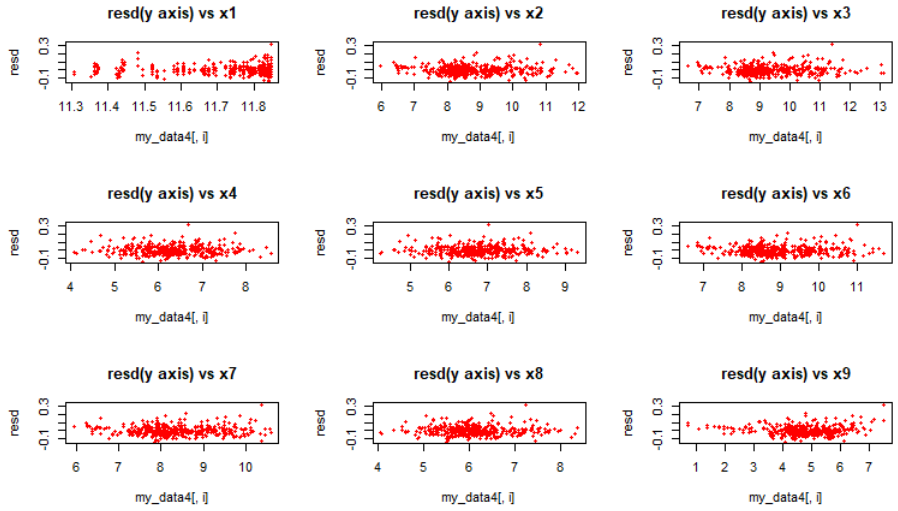


Figure 22: graph between regressors and residuals

value and residuals and graph between regressors and predictors.This indicate that heteroscedasticity is not significant now.

Now we have performed Breusch Pagan test
On applying Breusch Pagan test it can be seen that p value is 0.02235 which is more than 0.02. So heteroscedasticity is not significant at 0.02 level of significant.

So now it is clear that after removing heteroscedasticity, MAPE(mean absolute percentage erroe)= 0.05810667 or 5.8% which is quite better than previous model.

# 13 Multicollinearity

One of the assumptions of linear model is that there should be no multicollinearity in the model i.e there should be no high correlation between two or more predictor variable.

## 13.1 Multicollinearity diagnostics

It is detected with the help of

1. **Condition Number**

$$CN = \frac{\lambda_{max}}{\lambda_{min}}$$

where $\lambda_{max}$ and $\lambda_{min}$ are maximum and minimum values of $X^T X$.Here, X is matrix containing only numerical values of X(numerical data)
Convention: Condition number >100 indicates moderate multicolinearity and > 1000 indicates severe multicollinearity.
In **my_data4**( dataset obtained after removing influential points,normality,heteroscedasticity)Condition number is coming out to be **972.91** which is near to 1000 and it is strong evidence of Multicollinearity.
Note: Only numerical variables are considered in calculation of condition number since dummy variables don't contribute much to multicollinearity

2. **Variance Inflation Factor**

$$VIF_i = \frac{1}{1 - R_i^2}$$

where $R_i^2$ is $R^2$ obtained after regressing $i^{th}$ regressors on remaining other regressors.
Convention: VIF > 10 indicates severe multicollinearity
**my_data4** VIFs are shown in Fig.57 .It is once again evident from Fig.57 that VIFs of all the numerical variables except Total.Interactions is greater than 10 indicating multicollinearity

## 13.2 Dealing with multicollinearity

This work will consider various approaches for removing mulicollinearity as discussed below:

### 13.2.1 Principal Component Regression

In this instead of regressing on X's,regression is done on Principal Components explaining majority of variability. PCR is done for Total.Reach,Total.Impressions,Impressions.liked, Reach.liked and Consumers,Consumptions and People.engaged.liked separately. The choice

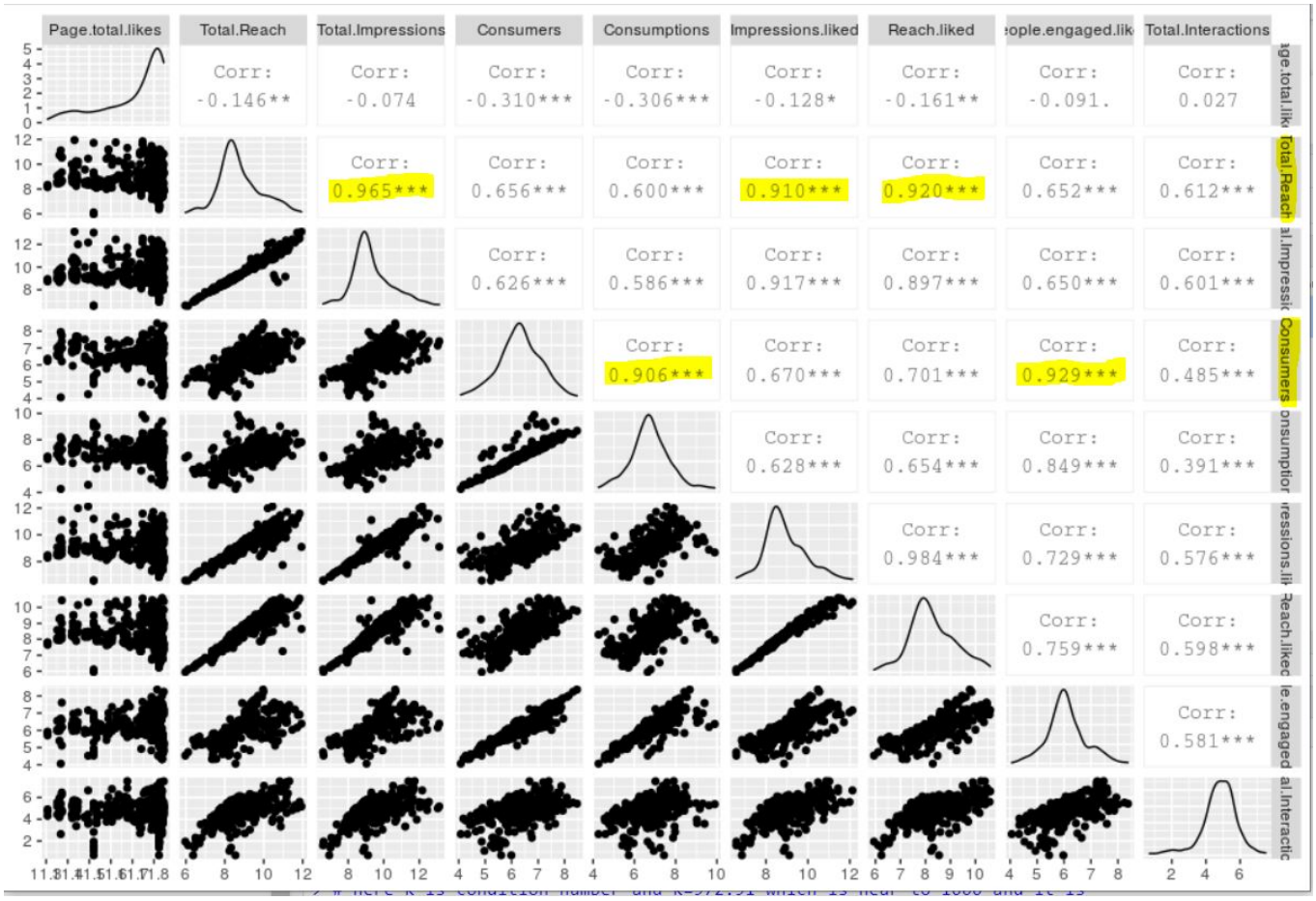


Figure 23: Calculation of VIF

of these combination is made from Fig 58.

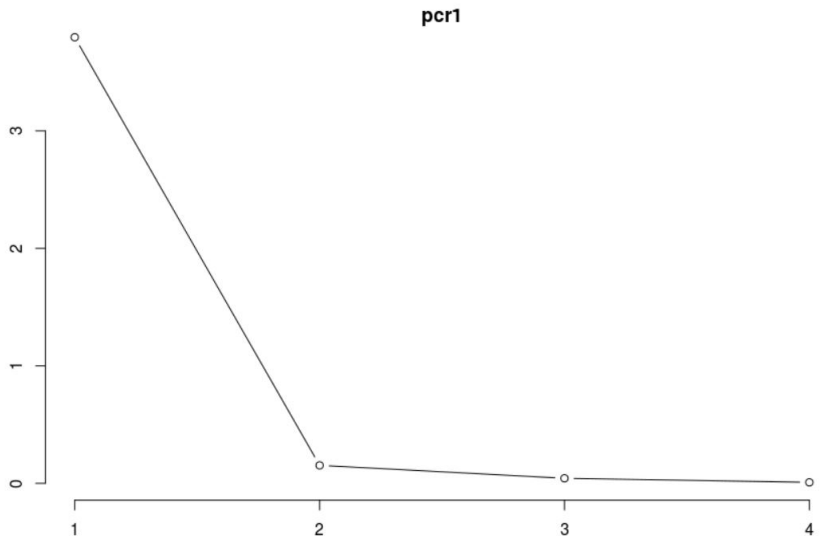


Figure 24: Scree Plot for PCR1

**PCR1**-Principal Component Regression for Total.Reach,Total.Impressions,Impressions.liked, Reach.liked is shown in Fig 59 and 60 indicating majority of variability(98.7%) is explained by first two principal components which will be further used for regression rather than all four Principal Components.

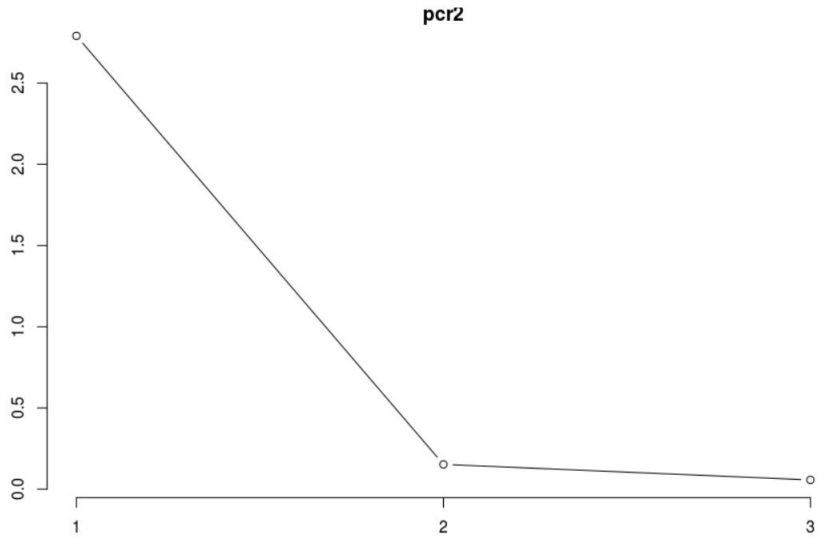


Figure 25: Scree Plot for PCR2

**PCR2**-Principal Component Regression for Consumers,Consumptions and People.engaged.liked is shown in Fig 61 and 62 indicating majority of variability(98.09%) is explained by first two principal components which will be further used for regression rather than all four Principal Components.

### 13.2.2 Variable Elimination

Page.Total.Likes is removed because of low correlation with response(-0.24688) and Large VIF(172.02) (See Fig.57)

Finally model 5 is fitted with following regressors shown in Fig.63 and VIFs are again calculated which are all satisfactorily less than 10 except for 3 dummy variables

When model 5 is applied on test data it shows a MAPE of 103.9799 which clearly show that Principal Component of numerical variables is not a fruitful approach always.Sometimes using it may create a mess as in Figure 64.

Still an attempt is made to improve performance of model 5 by stepwise regression of variables which is a combination of forward selection and backward elimination starting with Backward elimination in this case since full model 5 is being considered.This approach gives model shown in fig 65 with a MAPE of 96.976.So, this model won't be used for analysis since results are not as expected.

### 13.2.3 Principal Component Regression(PCR) on entire dataset

The principal component regression approach involves constructing the first M principal components $Z_1$,$Z_2$,...,$Z_M$ and then using these components as the predictors in a linear regression model that is fit using least squares.The key idea is that often a small number of principal components suffice to explain most of the variability in the data as well as the relationship with the response.In other words, *we assume that the direction in which* $X_1$,$X_2$...$X_p$ *show the most variation are the directions that are associated with* $Y$
PCR is performed on entire dataset and validation="CV" causes pcr() to compute the ten-fold cross-validation error for each possible value of M.Fig 66 shows validation plot for MSE vs no.of components.By looking at Fig.66, 38 principal components are opted which are explaining 98.65% variability of X and 97.78% variability of Z(Use summary(pcr.fit) to get these numbers)

When PCR is tested on test data it provides a MAPE of 0.1178 which is pretty good and this model can be used for analysis. However, as a result of the way PCR is implemented ,the final model is more difficult to interpret because it does not perform any kind of variable selection or even directly produce coefficient estimates.

### 13.2.4 Partial Least Square Regression (PLSR)on entire dataset

PCR suffers from a drawback: there is no guarantee that the direction that best explains the predictors will also be the best directions to use for predicting the response.PLS identifies the new feature in a supervised way-that is it makes use of response Y in order to identify new features that not only approximate old features well,but also that are related to response.PLS places the highest weight on the variable that are most strongly related to the response
PLS is performed on entire dataset and validation="CV" causes pls() to compute the ten-fold cross-validation error for each possible value of M.Fig.68 shows validation plot for MSE vs no.of components.By looking at Fig.68,10 components are opted which are explaining 37.68% variability of X and 99.37% variability of Z(Use summary(pls.fit) to get these numbers)

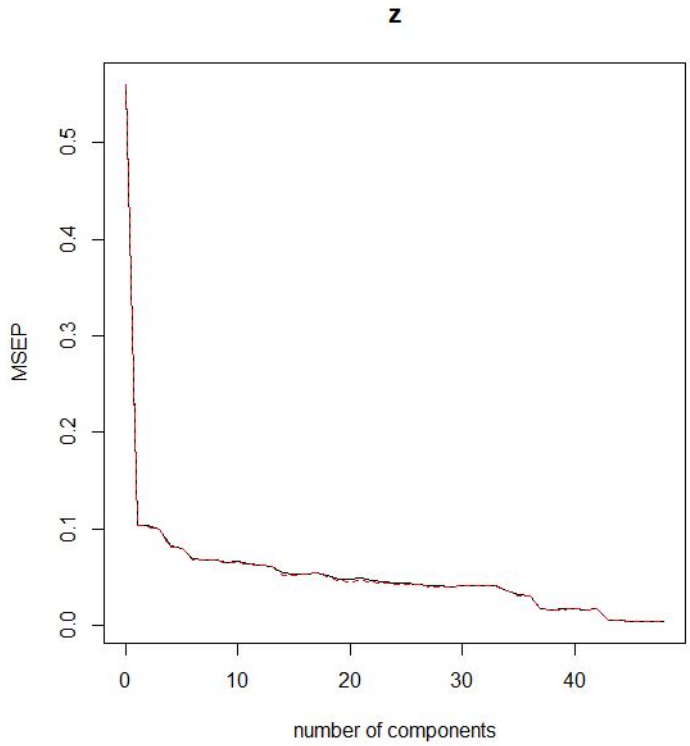


Figure 26: Cross validation MSE v/s number of components using PCR approach

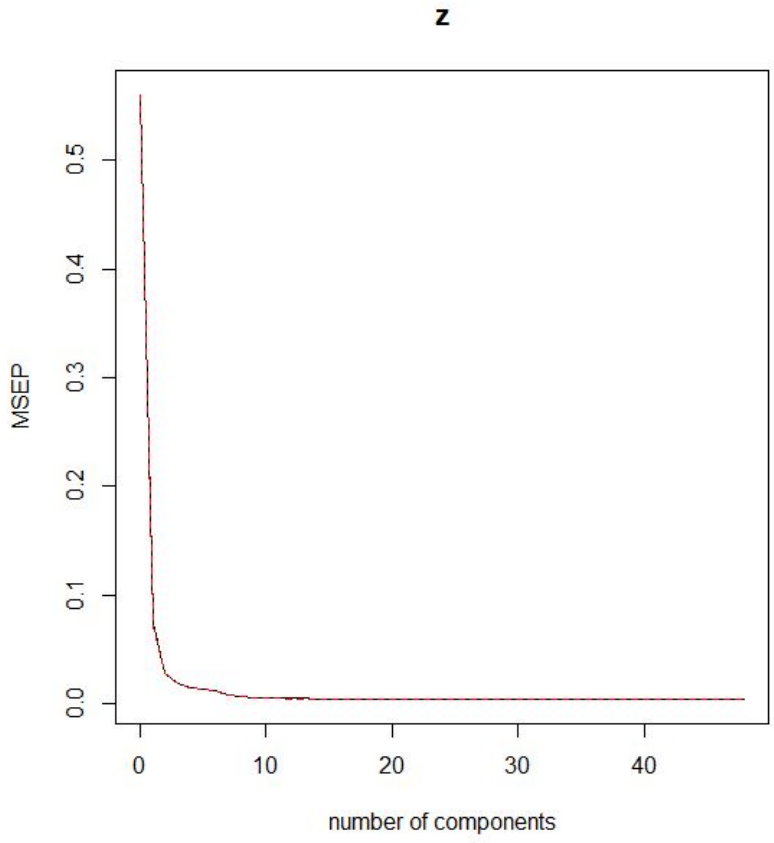


Figure 27: Cross validation MSE v/s number of components using PLS approach

When PLS is tested on test data it provides a MAPE of 0.0617 which is lower than MAPE of PCR so this model can also be used for analysis. However, as a result of the way PLS is implemented ,the final model is more difficult to interpret because it does not perform any kind of variable selection or even directly produce coefficient estimates.

### 13.2.5 The Lasso

Lasso is a techniques in which we fit a model containing all p predictors and use the techniques regularise the some of coefficients of estimates or shrinks the estimates towards zero.Shrinking of the coefficient estimates reduces their variance.
**Consider the equation**
L=$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j|$
The coefficients obtained by minimising this equation is called lasso coefficients is denoted by $\hat{\beta}_{\lambda}^{L}$.In this equation first is RSS and second term is known as **LASSO PENALTY** and $\lambda$ is tuning parameter.
As $\lambda$ increases coefficients of estimates shrinks towards zero and as $\lambda$ become sufficient large some of the coefficients become zero.So we can say Lasso also performs variable selection that is model contain a subset of the variable.

Here LASSO is performed on training data set using CROSS VALIDATION and we have the model corresponding to minimum MSE ( cv.glmnet() and alpha=1 indicate we have performed lasso using cross validation). In this case tuning parameter $\lambda$ is 10 (by default). It is known that as $\lambda$ increases coefficient estimates shrinks towards zero.From the fig70. It is clear that 17 coefficients estimates of predictors are non zero and rest 32 are shrinks to zero.So these predictor having non zero coefficients estimates are significant and rest are insignificant ( when we use $\lambda$=10).Name of significant predictors is given in figure70

Fig 72: is the graph between log($\lambda$) and Mean-Squared Error where x axis represent log($\lambda$) and y axis represent Mean-Squared Error.From this figure 72 and Fig 71: it is clear that when $\lambda$ =0.000429268 or log($\lambda$)= -7.753429 (as log(0.000429268)= -7.753429) than Mean-squared Error is minimum as "cv.out$lambda.min" provide the tuning Parameter $\lambda$ corresponding the model having minimum MSE Using Cross Validation.("cv.out" is defined in fig 70:).

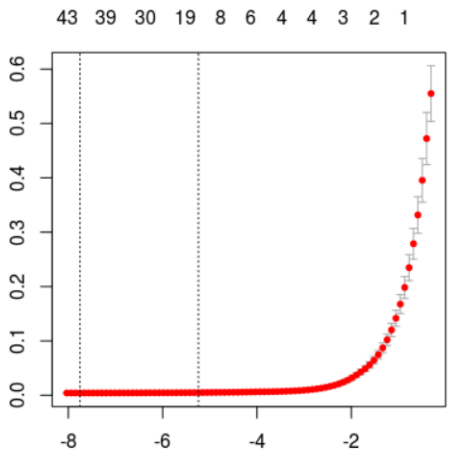


Figure 28: Graph Between Log(lambda) and Mean Squared Error

When the model obtained by LASSO using Cross-Validation is tested on test data it provides a MAPE of 0.0571 which is lower than MAPE of PCR and MAPE os PCS so this model is well enough than other model obtained by PCS and PCR.

# 14 Checking for Significance of first level of each Categorical Variable

In our model, in train data, 'Post.Hour_16', 'Post.Hour_19', 'Post.Hour_20' and 'Post.Hour_22' contain only one observation as '1', but due to exclusion of leverage points these observations are also removed. As a result now these columns contain only elements as '0'. But this will reduce rank of full rank matrix **X** which will create a problem as full rank was our basic assumption. So, to overcome this difficulty we have removed those four columns (this will put no effect on our analysis as we found that in the whole data set there were only one observation of these 'Post.Hour's).
In our initial model we found that there are 52 variables (including dummies) but only few of them have serious effect on the model and others are insignificant in prediction. Checking the p-vlaue and setting our level of significance at 5% we get a sets of significant and insignificant variables which shown in the table below have bold effects on the model. This model consists of 399 observations and 52 variables.
**Highly Significant Variables :** 'Total.Reach', 'Total.Impression', 'Consumer', 'Impression.liked', 'Reach.liked', 'People.engaged.liked' and 'Total.interaction'
**Moderately Significant Variables :** 'Post.week_4', 'Consumption', 'Type_photo', 'Type_video', 'Paid_1
**Insignificant Variables :** Intercept, 'Page.total.like', 'Type_status', All 'Category' (except Category_1), All 'Post.Month' (except Month_1), 'Post.week_5,6,7' , All 'Post.Hour' (except Hour_1) and 'Post.week_2,3'
It is notable that 'Intercept' has no significant effect. From separate Analysis first level of each categorical variable we found that

'Type_link' is significant, 'Category_1' is insignificant, 'Post.Month_1' is significant, 'Post.week_1' is insignificant, 'Post.Hour_1' is insignificant and 'Paid_0' is insignificant . This implies that categorical variables 'Type', 'Post.Month', 'Paid' are significant and 'Category', 'Weekday', 'Hour' are insignificant in predicting.
We found F-statistic=2790 with df 52 and 346 And $F_{0.05;52,346}$= 1.38, So, with a high confidence we can say some of the variables have very significant effects.

After excluding leverage points we built another model and find the following table significant and insignificant varibales. This model is built with 394 observations and 48 variables.
**Highly Significant Variables :** 'Total.Reach', 'Total.Impression', 'Consumer', 'Impression.liked', 'People.engaged.liked' and 'Total.interaction'
**Moderately Significant Variables :** 'Post.week_2,4', 'Consumption', 'Type_photo', 'Type_video', 'Paid_1
**Insignificant Variables :** Intercept, 'Page.total.like','Reach.liked', 'Type_status', All 'Category' (except Category_1), All 'Post.Month' (except Month_1), All 'Post.Hour' (except Hour_1) and 'Post.week_3,5,6,7'
It is notable that 'Intercept' has no significant effect. From separate Analysis first level of each categorical variable we found that

'Type_link' is insignificant, 'Category_1' is insignificant, 'Post.Month_1' is significant, 'Post.week_1' is insignificant, 'Post.Hour_1' is insignificant and 'Paid_0' is insignificant . This implies that categorical variables 'Type', 'Category', 'Post.Month', 'Post.weekday', 'Paid' are significant and 'Post.Hour' is insignificant in prediction.

We found F-statistic=3286 with df 48 and 345 And $F_{0.05;48,345}$= 1.395, So, with a high confidence we can say some of the variables have very significant effects.

After deleting influential points we built another model with 389 observations and 48 variables. Let's see its significant and insignificant variables.
**Highly Significant Variables :** 'Total.Reach', 'Total.Impression', 'People.engaged.liked', 'Consumer', 'Total.interaction' and 'Type_video'
**Moderately Significant Variables :** 'Consumption','Impression.liked', 'Reach.like', 'Type_photo', 'Post.week_4' and 'Category_2,3'
**Insignificant Variables :** Intercept, 'Page.total.like', 'Type_status', All 'Post.Month' (except Month_1), 'Post.week_2,3,5,6,7' , All 'Post.Hour' (except Hour_1) and 'Paid_1'
It is notable that 'Intercept' has no significant effect. From separate Analysis first level of each categorical variable we found that

'Type_link' is insignificant, 'Category_1' is insignificant, 'Post.Month_1' is significant, 'Post.week_1' is insignificant, 'Post.Hour_1' is insignificant and 'Paid_0' is insignificant . This implies that categorical variables 'Type', 'Category', 'Post.Month', 'Post.Weekday' are significant and 'Post.Hour', 'Paid' are insignificant in prediction.
We found F-statistic=4531 with df 48 and 340 And $F_{0.05;48,340}$= 1.396, So, with a high confidence we can say some of the variables have very significant effects.

After Normality correction we made a logarithmic transformation of our original response and built a new model. Let's check its significant and insignificant variables after transformation.
**Highly Significant Variables :** 'Intercept', 'Consumer', 'Total.interaction' and 'Type_photo,status'
**Moderately Significant Variables :** 'Consumption'
**Insignificant Variables :** Intercept, 'Page.total.like', 'People.engaged.liked' , 'Type_video', 'Total.Reach', 'Total.Impression', 'Impression.liked' , 'Total.Reach.like', All 'Category' (except Category_1), All 'Post.Month' (except Month_1), All 'Post.week' (except week_1), All 'Post.Hour' (except Hour_1)and 'Paid_1'
Here 'Intercept' has high significance (this is due to transformation in response). From separate Analysis first level of each categorical variable we found that

'Type_link' is insignificant, 'Category_1' is significant, 'Post.Month_1' is significant, 'Post.week_1' is insignificant, 'Post.Hour_1' is insignificant and 'Paid_0' is insignificant . This implies that categorical variables 'Type', 'Post.Month', 'Post.weekday' are significant and 'Category', 'Post.Hour' , 'Paid' are insignificant.
We found F-statistic=40 (as response is reduced in measure F-statistics fell by a significant amount) with df 48 and 333 And $F_{0.05;48,333}$= 1.396, So, still with a high confidence we can say some of the variables have very significant effects.

To remove heteroscedasticity we made a logarithmic transformation of our numeric continuous variables this time. With transformed response and regressors a new model was built with 382 observations and 48 variables. Now we check its significant and insignificant variables.
**Highly Significant Variables :** 'Total.Reach', 'Total.Impressions', 'Consumer', 'Consumption', 'People.engaged.liked', 'Total.interaction', 'Type_photo'
**Moderately Significant Variables :** 'Category_3' and 'Post.week_5'
**Insignificant Variables :** Intercept, 'Page.total.like', 'Type_video', 'Impression.like', 'Total.Reach.like', 'Category_2', All 'Post.Month' (except Month_1), 'Post.week_2,3,4,6,7'

, All 'Post.Hour' (except Hour_1) and 'Paid_1'
Here again 'Intercept' has no significant effect. From separate Analysis first level of each categorical variable we found that

'Type_link' is insignificant, 'Category_1' is insignificant, 'Post.Month_1' is significant, 'Post.week_1' is insignificant, 'Post.Hour_1' is insignificant and 'Paid_0' is insignificant . This implies that categorical variables 'Type', 'Category', 'Post.Month', 'Post.week' are significant and 'Post.Hour', 'Paid' are insignificant.
We found F-statistic=1273 with df 48 and 333 And $F_{0.05;48,333}$= 1.396, So, with a high confidence we can say some of the variables have very significant effects.

In order to remove multicollinearity from the model two PCA are done for 'Total.Reach', 'Total.impression', 'Impression.liked' , 'Reach.liked' and 'Consumer', 'Consumption' and 'People.engaged.like', then instead of all pcs only first two pcs are taken from both. Variability explained by first two principal components in both cases are 98%. 'Page.total.like' is removed due to low correlation with response and high VIF. Finally the model was built with 382 observations and 44 variables whose significant and insignificant variables are listed below.
**Highly Significant Variables :** Intercept, PC12, PC21, PC22, 'Total.interaction' and 'Post.Month_3,4,...
**Moderately Significant Variables :** 'Type_status', 'Post.Month_2' and 'Post.Hour_2'
**Insignificant Variables :** PC11, 'Type_photo,video', All 'Category' (except Category_1), All 'Post.week' (except week_1), 'Post.Hour_3,4,....,22' and 'Paid_1'
Here 'Intercept' has high significance (this is due to transformation in response). From separate Analysis first level of each categorical variable we found that

'Type_link' is insignificant, 'Category_1' is insignificant, 'Post.Month_1' is significant, 'Post.week_1' is insignificant, 'Post.Hour_1' is insignificant and 'Paid_0' is insignificant . This implies that categorical variables 'Type', 'Post.Month', 'Post.Hour' are significant and 'Category', 'Post.week', 'Paid' are insignificant.
We found F-statistic=672 (as response is reduced in measure F-statistics fell by a significant amount) with df 44 and 337 and $F_{0.05;44,337}$= 1.41, So, still with a high confidence we can say some of the variables have very significant effects.

On the above model we applied Stepwise variable selection method, thus only 25 important variables (these variables resulted minimum AIC) are selected from 44 variables and others totally ignored. So, our new model built with 25 variables and 382 observations. Let's check significant and insignificant variables.
**Highly Significant Variables :**
Intercept, PC12, PC21, PC22, 'Total.interaction' , 'Post.Month_3,4,...,12', 'Type_status' and 'Post.Hour_2'
**Moderately Significant Variables :** 'Post.Month_2' and 'Post.Hour_5,6'
**Insignificant Variables :** 'Post.week_5' and 'Post.Hour_8,9,....,12'
Clearly due searching best model using variable selection method there are very few insignificant variables and relatively larger number of significant variables. From separate Analysis first level of each categorical variable we found that

'Type_link' is insignificant, 'Category_1' is significant, 'Post.Month_1' is significant, 'Post.week_1' is significant, 'Post.Hour_1' is insignificant and 'Paid_0' is insignificant . This implies among the categorical variables only 'Paid' is insignificant.
We found F-statistic=1213 (as response is reduced in measure F-statistics fell by a signif-

icant amount) with df 25 and 356

To remove multicollinearity in another way, again we applied same procedure (PCA) but in different approach. This time PCA technique is applied over all the 48 variables and obviously got a better result than before using first 38 principal components (which explain 97.78% variability).

Another method (PLS) was used to remove multicollinearity and this time using only first 10 component (which explain 99.37% variability) we got very good result in prediction. Here a good number of variables are reduced.

Lastly Lasso was used to remove multicollinearity and to select variables simultaneously (on 48 variables). Lasso signified only 17 main variables into account.
**Significant Variables (taken into the model) :** Intercept, 'Total.Reach', 'Consumer', 'People.engaged.like', 'Total.interaction', 'Type_photo', 'Category_3', 'Post.Month_2,4,5,7,11,12', 'Post.week_5' and 'Post.Hour_2,9,14'.
**Insignificant Variables (not taken into the model) :** 'Page.Total.like', 'Total.impression', 'Impression.like', 'Reach.like', 'Consumption', 'Type_status,video', 'Category_2', 'Post.Month_3,6,8,9, 'Post.week_2,3,4,6,7' and 'Post.Hour_3,4,. . . ,8,10,11,12,15,17,18'
From separate Analysis first level of each categorical variable we found that

'Type_link' is insignificant, 'Category_1' is significant, 'Post.Month_1' is significant, 'Post.week_1' is significant, 'Post.Hour_1' is insignificant and 'Paid_0' is insignificant. This implies among the categorical variables only 'Paid' is insignificant and other categorical variables have significant effect in prediction.

# 15 Conclusion

Finally we got three models **(i)** firstly using principal component of all variables with MAPE 11.8%
**(ii)** secondly using partial least square with MAPE 6.17%
**(iii)** thirdly using Lasso with MAPE 5.71%.
Though the first model's efficiency is not bad, it took all the regressors (48 regressors are hard to interpret at a time ) into account and then used first 38 components (which explain 97.78% of total variability ). This will take too much time and cost for larger number of observations and from this model we can't specify the regressors which have severe effects on the response (Lifetime.engaged.users) . Second model, with better efficiency than first model, also took all the regressors into account (which will again alert a statistician regarding time and cost) but then using only first 10 components (which explain 99.37% of total variability) it predicted very well. Also, here is no scope for choosing best regressors. Thirdly with Lasso we overcame all the difficulties. It identified 17 regressors and with the help of these regressors it predicted very much well than first two models. Clearly, here we can check the effects of these selected 17 regressors and time, cost will be much less here. So, regarding the response Lifetime.engaged.users few factors are there to be noticed. Those are -
**(i)** Posts in April, May, July induced lesser number of users where Feb, Nov, Dec ensured

higher number of users. So, the company should post in year-end or in the beginning of the year.
**(ii)** More users are engaged with post type 'Photo'
**(iii)** Inspirational posts gathered more users
**(iv)** More users can be engaged with the posts in the night and noon
**(v)** Weekdays have not such heavy effects on the users
**(vi)** No effects of paid or unpaid posts
**(vi)** As 'Total.Reach', 'like', 'comment', 'share', 'Consumer', 'People.engaged.like' are directly related to clicking or seeing any post by the users, more amount of these variables certainly will imply more the engaged users.

# 16 Bibliography

1. An Introduction to Statistical Learning,with Applications in R by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani
2. Introduction to Linear Regression Analysis by Douglas C Montgomery, Elizabeth A Peck, G. Geoffrey Vining
3. The Elements of Statistical Learning:Data Mining, Inference, and Prediction by Trevor Hastie,Robert Tibshirani,Jerome Friedman
4. http://dx.doi.org/10.1016/j.jbusres.2016.02.010 ((Moro et al., 2016) Moro, S., Rita, P., Vala, B. (2016))
5. https://stackoverflow.com
6. www.analyticsvidhya.com

# 17 Appendix

The files used in this project can be found at the following link. This include data file and R code.

https://drive.google.com/drive/folders/1_Tw7cTaX78s8HzM6nmt-PVEgtDvDLcTN?usp=sharing