City of Raileigh Analysis

Name Sriranganath Shankam Manjunatha Prasanna

```r
library(dplyr)
library(caTools)
library(tidyverse)
library(caret)
```

```r
#loading the Dataset
raileigh = read.csv("Building_Permits.csv")
```

```r
#No of rows and coumns respectively
nrow(raileigh)
```

```
[1] 141953
```

```r
ncol(raileigh)
```

```
[1] 87
```

```r
#Total Different Types of Construction
unique(raileigh$const_type, incomparables = FALSE)
```

```
 [1] V  B      II B II A IIIB I   A I   B IV    V   A IV U VI U V   U VI P V   P IIIA I     III   II    IV P
Levels:  I I   A I   B II II A II B III IIIA IIIB IV IV P IV U V   A V   B V   P V   U VI P VI U
```

```r
#Mean and Median Stories
no_na_stories = na.omit(raileigh$numberstories) #removing blanks
summary(raileigh$numberstories)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
     0.0      1.0      2.0      9.7      2.0 342365.0    19511 
```

Intial Linear Regression for the Data columns Estimated cost and Issue Date year for raw data

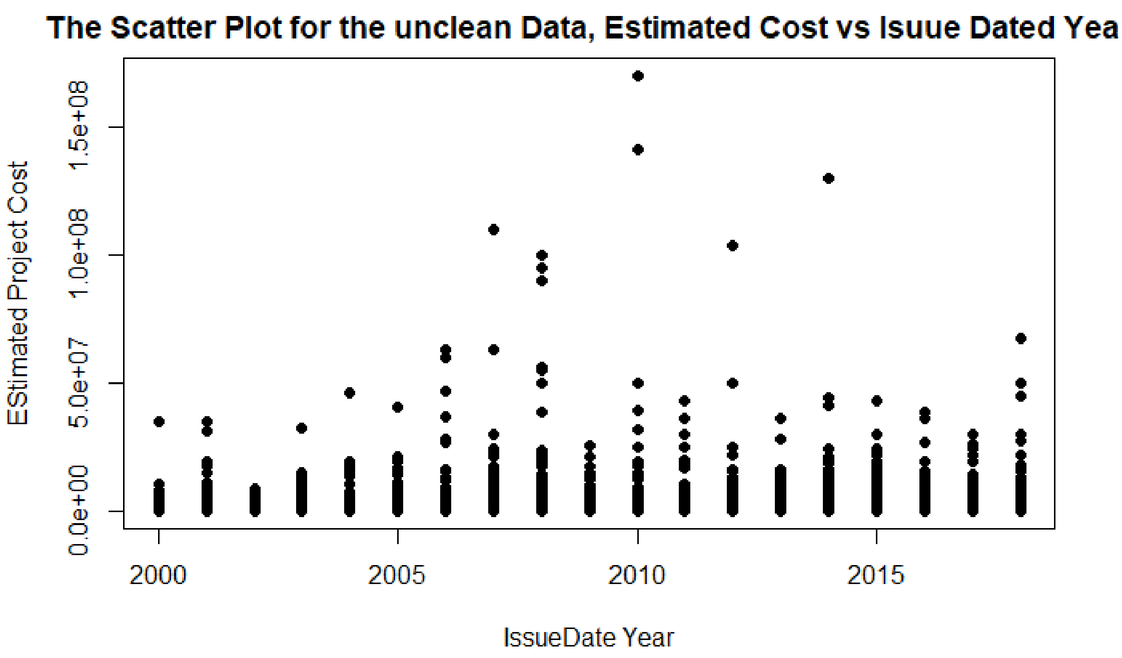
```r
initial_pred=lm(estprojectcost~issueddate_yr,data=raileigh)
summary(initial_pred)
```

```
Call:
lm(formula = estprojectcost ~ issueddate_yr, data = raileigh)

Residuals:
      Min        1Q    Median        3Q       Max
  -210069   -176009   -128066    -41855 169811934

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -5340098.2  1399266.5  -3.816 0.000135 ***
issueddate_yr     2750.3      696.6   3.948 7.87e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1422000 on 138517 degrees of freedom
  (3434 observations deleted due to missingness)
Multiple R-squared:  0.0001125, Adjusted R-squared:  0.0001053
F-statistic: 15.59 on 1 and 138517 DF,  p-value: 7.871e-05
```

The P-values are very high as well as the Rsquare value is too low and its not significant al all.

Hide

```
#Scatter Plot for Estimated Cost
plot(raileigh$issueddate_yr,raileigh$estprojectcost,pch=19,main="The Scatter Plot for the unclean Data, Esti
mated Cost vs Isuue Dated Year",xlab = "IssueDate Year", ylab = "EStimated Project Cost")
```



Standard Deviation of X and Y of permits

Hide

```
sd(raileigh$i..X, na.rm = TRUE)
```

```
[1] 0.06880534
```

Hide

```
sd(raileigh$Y, na.rm = TRUE)
```

```
[1] 0.05839208
```

Estimate cost

Hide

```
range(raileigh$estprojectcost)
```

```
[1] 0.0e+00 1.7e+08
```

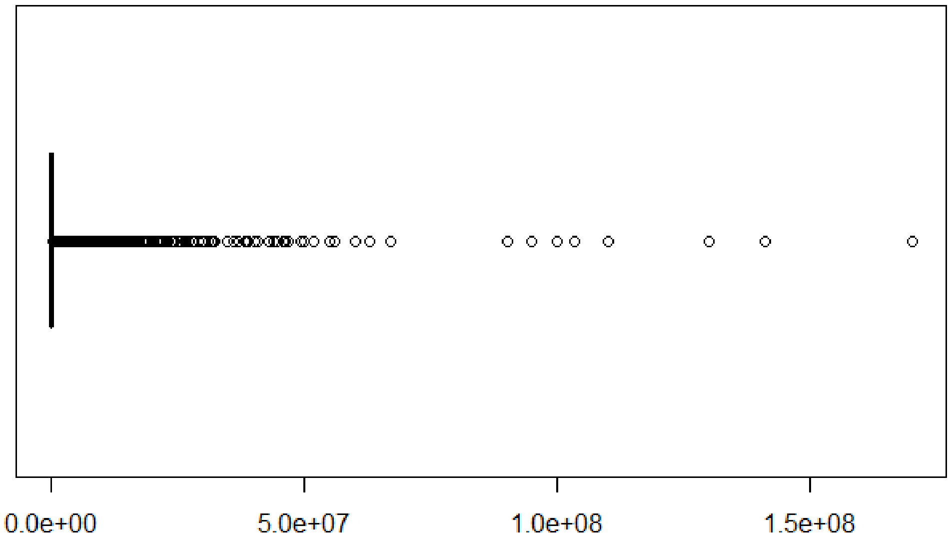
Hide

```
summary(raileigh$estprojectcost) #Checking the range and summaryfor knowing the end values of Estimated cost
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
0.00e+00 1.00e+04 5.44e+04 1.92e+05 1.35e+05 1.70e+08 
```

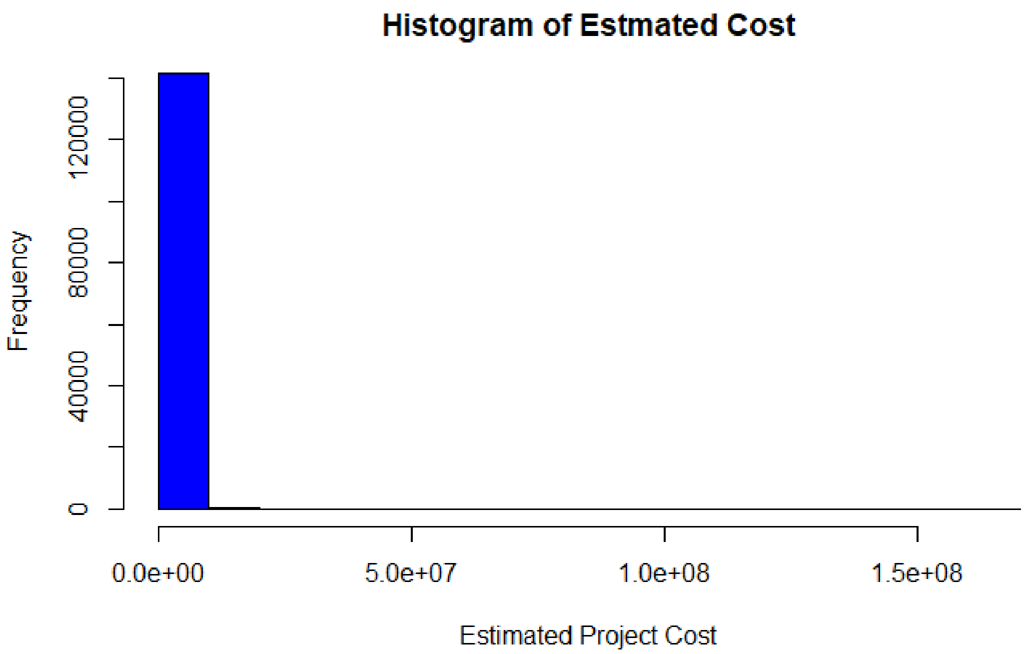
Hide

```
boxplot(raileigh$estprojectcost,horizontal=TRUE)  # BoXplot to see how there are so many outliers
```



Hide

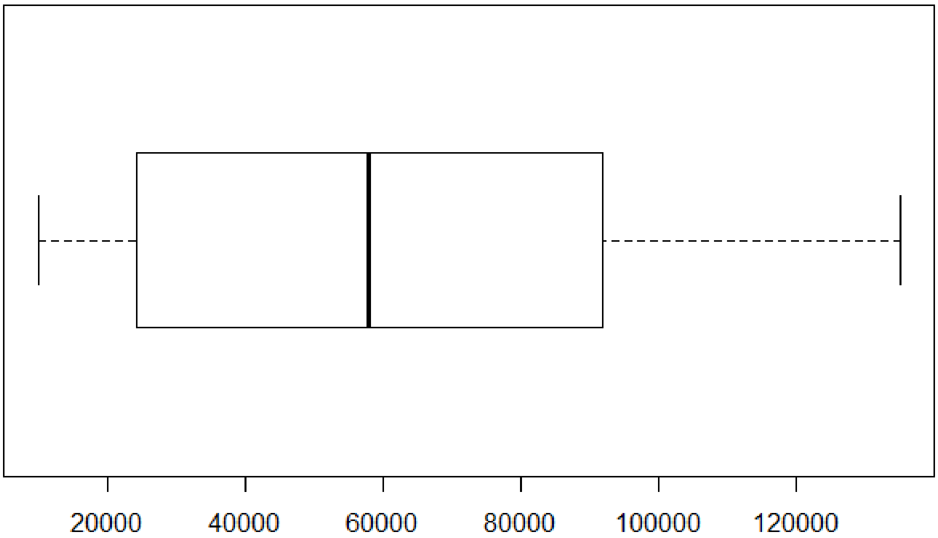
```
hist(raileigh$estprojectcost, main = "Histogram of Estmated Cost",xlab = "Estimated Project Cost", col = "Bl
ue") # Histogram of Estimated Cost
```



This shows the range of the Estimated Project Cost, as the data has very high outliers were getting one big bar on the histogram, so we can remove the outliers and check for the distribution of this field. Box plot also shows the same as well. How the data is positively skewed.

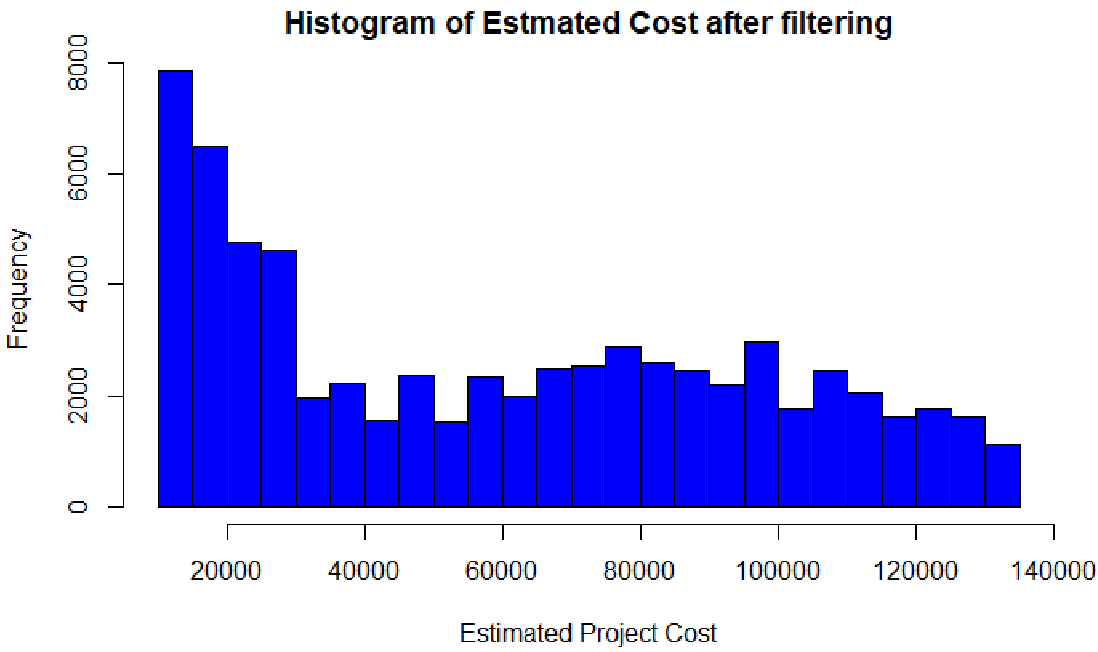
Hide

```
#Filtering the Estimate cost by Inter quartile range to show the Distribution of Estimated Cost
filtered_estimate<-raileigh %>%
  dplyr::filter(estprojectcost > quantile(estprojectcost, 0.25),
                estprojectcost < quantile(estprojectcost, 0.75))
boxplot(filtered_estimate$estprojectcost,horizontal=T)
```



Hide

```
hist(filtered_estimate$estprojectcost, main = "Histogram of Estmated Cost after filtering",xlab = "Estimated Project Cost", col = "Blue")
```



Now we can see that the Estimated cost is positively skewed , with tail towards the right. We can see many pojects in cost range of 8000 to 22000.

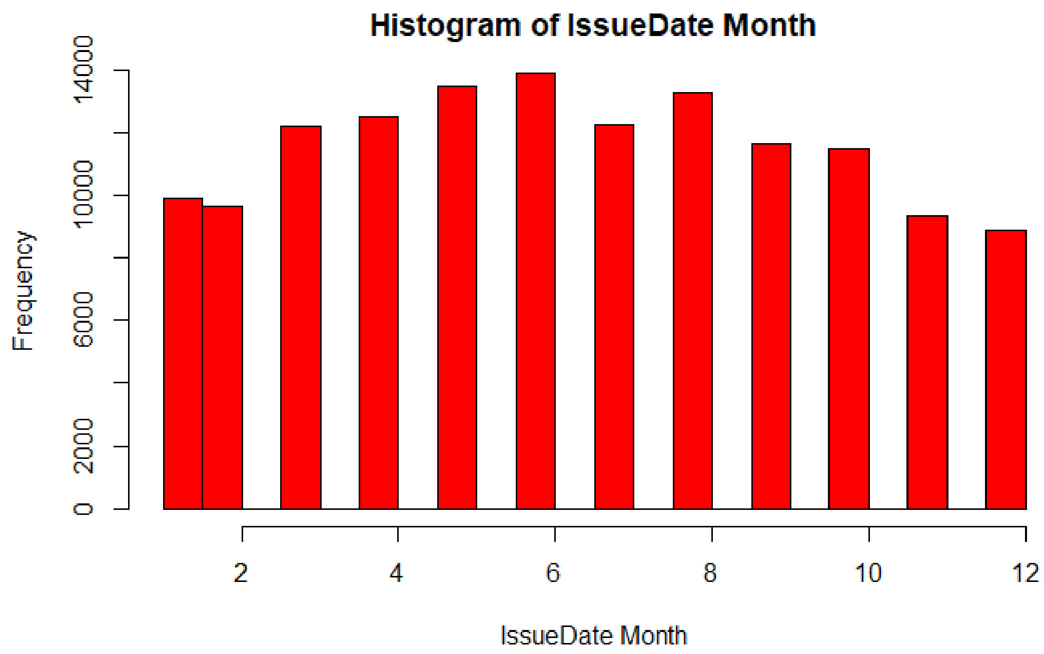
Hide

```
#Issue Date Month
no_na_issuedatemth = na.omit(raileigh$issueddate_mth) #Removing NA values in Issue date Month
summary(no_na_issuedatemth)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   4.000   6.000   6.407   9.000  12.000 
```

Hide

```
hist(no_na_issuedatemth,main = "Histogram of IssueDate Month",xlab = "IssueDate Month", col = "Red")
```



I have handled the null values for the Issue month and have plotted the histogram. we can see thatmore projects have been issue in the months of 4-8 in any given year.

Hide

```
#Filtering the data for Workclass = "New", Construction Type ="V  B" and Number of Stories is less than 3
filtered_raileigh = filter(raileigh, workclassmapped=="New" & const_type=="V  B" & numberstories < 3)
range(filtered_raileigh$issueddate_yr) #checking the range of Issue date year
```

```
[1] NA NA
```

Were filetering the data according to the conditions Workclass = "New", Construction Type ="V B" and Number of Stories is less than 3 Also were checking the issuedate year range, as we can see its shows NA, because it has a lot of NA values. No lwts handle it

Hide

```
#Finding NA values in both our values for Regression, and replacing the NA values by mean of the other value
s.
filtered_raileigh$estprojectcost = ifelse(is.na(filtered_raileigh$estprojectcost),ave(filtered_raileigh$estp
rojectcost, FUN = function(x) mean(x,na.rm=TRUE)),filtered_raileigh$estprojectcost)
filtered_raileigh$issueddate_yr = ifelse(is.na(filtered_raileigh$issueddate_yr),ave(filtered_raileigh$issued
date_yr, FUN = function(x) mean(x,na.rm=TRUE)),filtered_raileigh$issueddate_yr)
range(filtered_raileigh$issueddate_yr)  #checking after replacing NA values
```

```
[1] 2002 2018
```

So now its shows the highes year and the lowest year in the coulumn.

Hide

```
range(filtered_raileigh$estprojectcost) #range of projectcost
```

```
[1] 0e+00 1e+08
```

Sampling the whole filetered data set and splitting into 75% train and 25% into test data

Hide

```
smp_size <- floor(0.75 * nrow(filtered_raileigh))
set.seed(2196) # Set the seed to make your partition reproducible
train_ind <- sample(seq_len(nrow(filtered_raileigh)), size = smp_size)
train_raileigh <- filtered_raileigh[train_ind, ]
test_raileigh <- filtered_raileigh[-train_ind, ]
```

Linear Regression with Train dataset without cleaning

Hide

```
raileigh_pred=lm(estprojectcost~issueddate_yr,data=train_raileigh)
summary(raileigh_pred)
```

```
Call:
lm(formula = estprojectcost ~ issueddate_yr, data = train_raileigh)

Residuals:
     Min       1Q   Median       3Q      Max 
 -332336   -87718   -43393    27559 99789843 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -25034959    2818097  -8.884   <2e-16 ***
issueddate_yr     12572       1403   8.960   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 982600 on 23281 degrees of freedom
Multiple R-squared:  0.003436,	Adjusted R-squared:  0.003394 
F-statistic: 80.28 on 1 and 23281 DF,  p-value: < 2.2e-16
```

Hide

```
plot(train_raileigh$issueddate_yr,train_raileigh$estprojectcost,pch=19,main="Linear regression on Train data
set without cleaning",xlab = "IssueDate Year", ylab = "EStimated Project Cost")
```

Because of the high range of data and the outlier, which can be seen of the scatter plot we can see much of relation between our variable chosen for Regressio. The P-values for the intercept shows that, IssueDat year (inedependent variable) has some kinda relation with our

Dependent Variable (Estimated Cost). But due to the high error of an 982600 and the R square value of 0.003 we can say that this model is not a right fit for these selected variables.

Filtering Train Data, by multiplying the interquartile range by the number 1.5 and adding it to the third quartile and subrtracting it from the first quartile. Any number less than this is a suspected outlier.

Hide

```
filtered_train<-train_raileigh %>%
  dplyr::filter(estprojectcost > quantile(train_raileigh$estprojectcost, .25) - 1.5*IQR(train_raileigh$estpr
ojectcost),
                train_raileigh$estprojectcost < quantile(train_raileigh$estprojectcost, .75) + 1.5*IQR(train
_raileigh$estprojectcost))
```

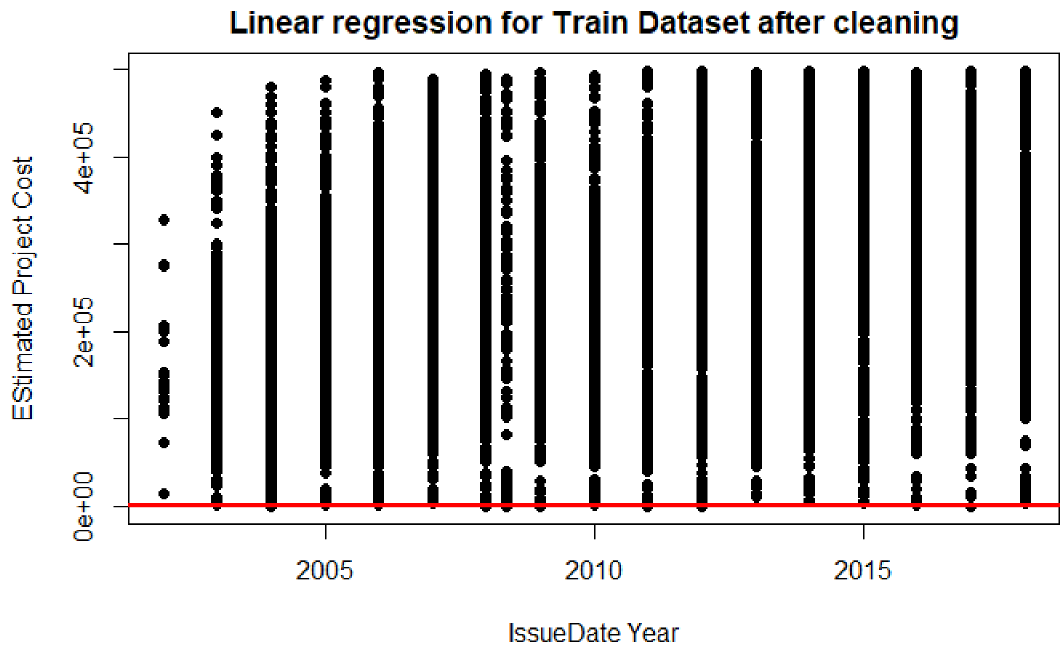
Hide

```
raileigh_pred2=lm(estprojectcost~issueddate_yr,data=filtered_train)
summary(raileigh_pred2)
```

```

Call:
lm(formula = estprojectcost ~ issueddate_yr, data = filtered_train)

Residuals:
    Min      1Q  Median      3Q     Max 
-268005  -60033  -16548   49421  343667 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -1.922e+07  2.828e+05  -67.96   <2e-16 ***
issueddate_yr  9.658e+03  1.408e+02   68.59   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 95490 on 22205 degrees of freedom
Multiple R-squared:  0.1748,	Adjusted R-squared:  0.1748 
F-statistic:  4704 on 1 and 22205 DF,  p-value: < 2.2e-16
```

Hide

```
plot(filtered_train$issueddate_yr,filtered_train$estprojectcost,pch=19,main="Linear regression for Train Dat
aset after cleaning",xlab = "IssueDate Year", ylab = "EStimated Project Cost")
abline(0,1,col="red",lwd=3)
```



After removing the outlier and limiting our Datasets, from the linear regression results we can see that the R squared value increased to 17% which is a significant difference from the previous value.But still it doesnt prove any significant relation between our dependent and independent variables.Providing us with no meaningful insights.

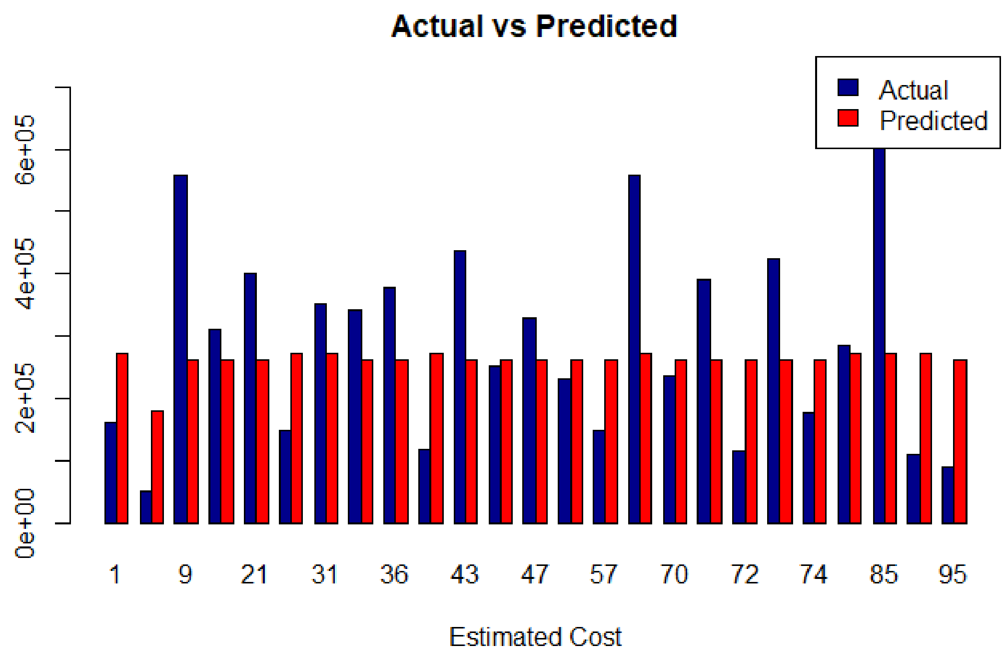
Hide

```
#Predicting values on Test Dataset
predictions <- predict(raileigh_pred2,test_raileigh)
output <- cbind(test_raileigh, predictions)
actuals_preds <- data.frame(cbind(actuals=test_raileigh$estprojectcost, predicteds=predictions))
```

Plotting Actual vs Predicted with test data

Hide

```
mx <- t(as.matrix(ab[-3]))
ab = head(actuals_preds,n=25)
barplot(mx,main="Actual vs Predicted", xlab='Estimated Cost',beside = TRUE,
        col=c("darkblue","red"))
legend('topright',fill=c("darkblue","red"),legend=c('Actual','Predicted')) # add a legend
```



As we can see the Data is predicting value is similar range, and when compared to the actual data, it is nowhere close.

Conclusion : I feel the linear regression for the variables of Estimated cost has no significant relation to IssueDate year, Probable solution to try would be to see if any other variable in the dataset Raileigh has any relation with it.There is no linear relation whatsoever Also, we can try more advanced Regression models - one model which might work is Robust Regression, which applies re-weighing to remove outlier influence with these kinda data set with heavy outlier influence.

Thank you for your time.