facebook-posts.R

Tue Sep 12 22:33:22 2017

setwd(getwd())  
library(readr)  
library(ggplot2)  
library(class)  
library(caret)

## Loading required package: lattice

options(scipen = 999) # removes scientfic notation  
dd <- read\_csv("~/Desktop/fb/P1TrainingData.csv",   
 col\_names = FALSE)

## Parsed with column specification:  
## cols(  
## .default = col\_double(),  
## X1 = col\_integer(),  
## X2 = col\_integer(),  
## X3 = col\_integer(),  
## X4 = col\_integer(),  
## X5 = col\_integer(),  
## X31 = col\_integer(),  
## X32 = col\_integer(),  
## X33 = col\_integer(),  
## X34 = col\_integer(),  
## X35 = col\_integer(),  
## X36 = col\_integer(),  
## X37 = col\_integer(),  
## X38 = col\_integer(),  
## X39 = col\_integer(),  
## X40 = col\_integer(),  
## X41 = col\_character(),  
## X42 = col\_character()  
## )

## See spec(...) for full column specifications.

head(dd)

## # A tibble: 6 x 42  
## X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11  
## <int> <int> <int> <int> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0 634995 0 463 1 0 806 11.29104 1 70.49514 0  
## 2 0 634995 0 463 1 0 806 11.29104 1 70.49514 0  
## 3 0 634995 0 463 1 0 806 11.29104 1 70.49514 0  
## 4 0 634995 0 463 1 0 806 11.29104 1 70.49514 0  
## 5 0 634995 0 463 1 0 806 11.29104 1 70.49514 0  
## 6 0 634995 0 463 1 0 806 11.29104 1 70.49514 0  
## # ... with 31 more variables: X12 <dbl>, X13 <dbl>, X14 <dbl>, X15 <dbl>,  
## # X16 <dbl>, X17 <dbl>, X18 <dbl>, X19 <dbl>, X20 <dbl>, X21 <dbl>,  
## # X22 <dbl>, X23 <dbl>, X24 <dbl>, X25 <dbl>, X26 <dbl>, X27 <dbl>,  
## # X28 <dbl>, X29 <dbl>, X30 <dbl>, X31 <int>, X32 <int>, X33 <int>,  
## # X34 <int>, X35 <int>, X36 <int>, X37 <int>, X38 <int>, X39 <int>,  
## # X40 <int>, X41 <chr>, X42 <chr>

tail(dd)

## # A tibble: 6 x 42  
## X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11  
## <int> <int> <int> <int> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 7 7170111 70 497000 9 0 1881 497.2 269 502.3184 0  
## 2 1 7170111 70 497000 9 0 1881 497.2 269 502.3184 0  
## 3 2 7170111 70 497000 9 0 1881 497.2 269 502.3184 0  
## 4 72 7170111 70 497000 9 0 1881 497.2 269 502.3184 0  
## 5 28 7170111 70 497000 9 0 1881 497.2 269 502.3184 0  
## 6 11 7170111 70 497000 9 0 1881 497.2 269 502.3184 0  
## # ... with 31 more variables: X12 <dbl>, X13 <dbl>, X14 <dbl>, X15 <dbl>,  
## # X16 <dbl>, X17 <dbl>, X18 <dbl>, X19 <dbl>, X20 <dbl>, X21 <dbl>,  
## # X22 <dbl>, X23 <dbl>, X24 <dbl>, X25 <dbl>, X26 <dbl>, X27 <dbl>,  
## # X28 <dbl>, X29 <dbl>, X30 <dbl>, X31 <int>, X32 <int>, X33 <int>,  
## # X34 <int>, X35 <int>, X36 <int>, X37 <int>, X38 <int>, X39 <int>,  
## # X40 <int>, X41 <chr>, X42 <chr>

#convert to a data frame  
dd <- as.data.frame(dd)  
  
#look at the variable names  
str(dd)

## 'data.frame': 40949 obs. of 42 variables:  
## $ X1 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ X2 : int 634995 634995 634995 634995 634995 634995 634995 634995 634995 634995 ...  
## $ X3 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ X4 : int 463 463 463 463 463 463 463 463 463 463 ...  
## $ X5 : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ X6 : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ X7 : num 806 806 806 806 806 806 806 806 806 806 ...  
## $ X8 : num 11.3 11.3 11.3 11.3 11.3 ...  
## $ X9 : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ X10: num 70.5 70.5 70.5 70.5 70.5 ...  
## $ X11: num 0 0 0 0 0 0 0 0 0 0 ...  
## $ X12: num 806 806 806 806 806 806 806 806 806 806 ...  
## $ X13: num 7.57 7.57 7.57 7.57 7.57 ...  
## $ X14: num 0 0 0 0 0 0 0 0 0 0 ...  
## $ X15: num 69.4 69.4 69.4 69.4 69.4 ...  
## $ X16: num 0 0 0 0 0 0 0 0 0 0 ...  
## $ X17: num 76 76 76 76 76 76 76 76 76 76 ...  
## $ X18: num 2.6 2.6 2.6 2.6 2.6 ...  
## $ X19: num 0 0 0 0 0 0 0 0 0 0 ...  
## $ X20: num 8.51 8.51 8.51 8.51 8.51 ...  
## $ X21: num 0 0 0 0 0 0 0 0 0 0 ...  
## $ X22: num 806 806 806 806 806 806 806 806 806 806 ...  
## $ X23: num 10.6 10.6 10.6 10.6 10.6 ...  
## $ X24: num 1 1 1 1 1 1 1 1 1 1 ...  
## $ X25: num 70.3 70.3 70.3 70.3 70.3 ...  
## $ X26: num -69 -69 -69 -69 -69 -69 -69 -69 -69 -69 ...  
## $ X27: num 806 806 806 806 806 806 806 806 806 806 ...  
## $ X28: num 4.97 4.97 4.97 4.97 4.97 ...  
## $ X29: num 0 0 0 0 0 0 0 0 0 0 ...  
## $ X30: num 69.9 69.9 69.9 69.9 69.9 ...  
## $ X31: int 0 0 0 7 1 0 0 1 0 0 ...  
## $ X32: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ X33: int 0 0 0 3 0 0 0 1 0 0 ...  
## $ X34: int 0 0 0 7 1 0 0 1 0 0 ...  
## $ X35: int 0 0 0 -3 0 0 0 -1 0 0 ...  
## $ X36: int 65 10 14 62 58 60 68 32 35 48 ...  
## $ X37: int 166 132 133 131 142 166 145 157 177 126 ...  
## $ X38: int 2 1 2 1 5 1 2 2 5 1 ...  
## $ X39: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ X40: int 24 24 24 24 24 24 24 24 24 24 ...  
## $ X41: chr "WED" "THU" "FRI" "FRI" ...  
## $ X42: chr "SAT" "FRI" "SAT" "MON" ...  
## - attr(\*, "spec")=List of 2  
## ..$ cols :List of 42  
## .. ..$ X1 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ X2 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ X3 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ X4 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ X5 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ X6 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X7 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X8 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X9 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X10: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X11: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X12: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X13: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X14: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X15: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X16: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X17: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X18: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X19: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X20: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X21: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X22: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X23: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X24: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X25: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X26: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X27: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X28: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X29: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X30: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X31: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ X32: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ X33: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ X34: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ X35: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ X36: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ X37: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ X38: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ X39: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ X40: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ X41: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ X42: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## ..$ default: list()  
## .. ..- attr(\*, "class")= chr "collector\_guess" "collector"  
## ..- attr(\*, "class")= chr "col\_spec"

#looks like we will need to rename the columns to keep things organized according to the appendix.  
  
dd <- subset(dd, select = -c(X6:X30))  
  
#change column names to match the appendix  
names(dd)[names(dd)=="X1"] <- "TargetVariable"  
names(dd)[names(dd)=="X2"] <- "PagePopularity"  
names(dd)[names(dd)=="X3"] <- "Checkins"  
names(dd)[names(dd)=="X4"] <- "PageTalkingAbout"  
names(dd)[names(dd)=="X5"] <- "PageCatagory"  
names(dd)[names(dd)=="X31"] <- "Comments"  
names(dd)[names(dd)=="X32"] <- "Comments24"  
names(dd)[names(dd)=="X33"] <- "Comments2448"  
names(dd)[names(dd)=="X34"] <- "Comment24Post"  
names(dd)[names(dd)=="X35"] <- "CommentDiff2448"  
names(dd)[names(dd)=="X36"] <- "BaseTime"  
names(dd)[names(dd)=="X37"] <- "PostLength"  
names(dd)[names(dd)=="X38"] <- "PostShareCount"  
names(dd)[names(dd)=="X39"] <- "PostPromotionStatus"  
names(dd)[names(dd)=="X40"] <- "Hours"  
names(dd)[names(dd)=="X41"] <- "PostPublishedWeekday"  
names(dd)[names(dd)=="X42"] <- "BaseDateTimeWeekday"  
  
#check that the column name where changed  
str(dd)

## 'data.frame': 40949 obs. of 17 variables:  
## $ TargetVariable : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ PagePopularity : int 634995 634995 634995 634995 634995 634995 634995 634995 634995 634995 ...  
## $ Checkins : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ PageTalkingAbout : int 463 463 463 463 463 463 463 463 463 463 ...  
## $ PageCatagory : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Comments : int 0 0 0 7 1 0 0 1 0 0 ...  
## $ Comments24 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Comments2448 : int 0 0 0 3 0 0 0 1 0 0 ...  
## $ Comment24Post : int 0 0 0 7 1 0 0 1 0 0 ...  
## $ CommentDiff2448 : int 0 0 0 -3 0 0 0 -1 0 0 ...  
## $ BaseTime : int 65 10 14 62 58 60 68 32 35 48 ...  
## $ PostLength : int 166 132 133 131 142 166 145 157 177 126 ...  
## $ PostShareCount : int 2 1 2 1 5 1 2 2 5 1 ...  
## $ PostPromotionStatus : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Hours : int 24 24 24 24 24 24 24 24 24 24 ...  
## $ PostPublishedWeekday: chr "WED" "THU" "FRI" "FRI" ...  
## $ BaseDateTimeWeekday : chr "SAT" "FRI" "SAT" "MON" ...

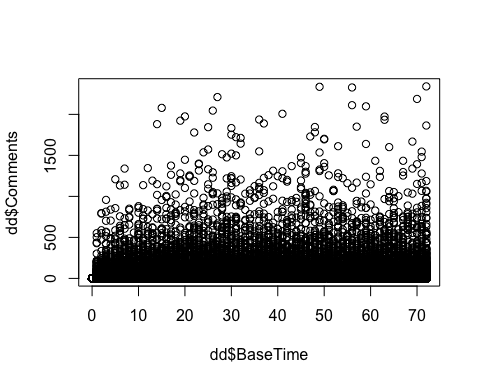
#let get some descriptive stats  
summary(dd$Comments)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 2.00 11.00 55.72 46.00 2341.00

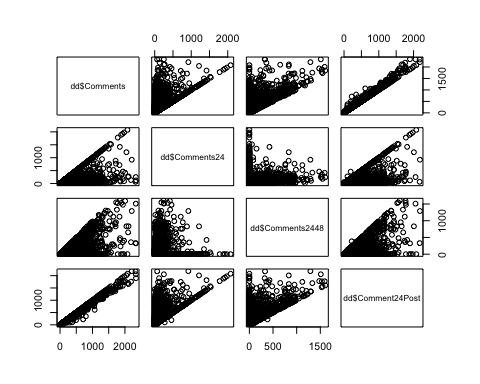
summary(dd$BaseTime)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 17.00 35.00 35.32 53.00 72.00

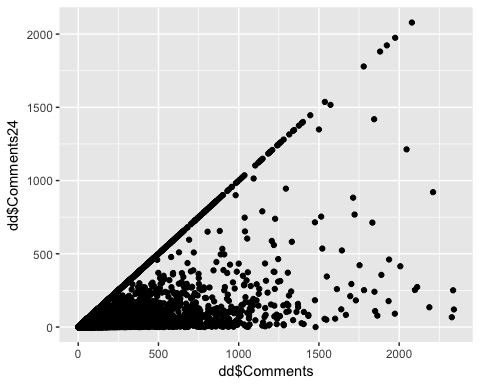
#quick look at the data with a pairs plot and scatter plot to see any patterns in comment volume  
plot(dd$BaseTime, dd$Comments)



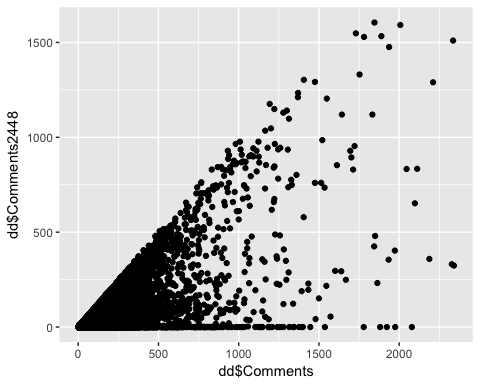
#shows some linear regression  
pairs(~ dd$Comments + dd$Comments24 + dd$Comments2448 + dd$Comment24Post)



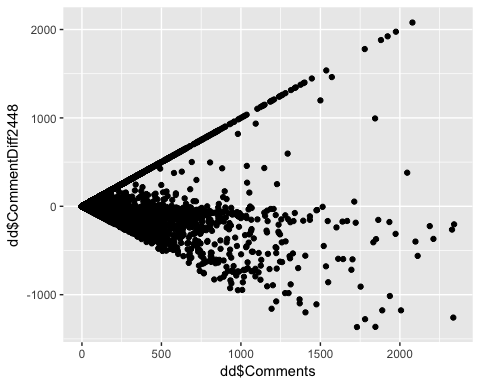
#total comments vs 24hr, 24-48, 24-48 diff  
qplot(dd$Comments, dd$Comments24)



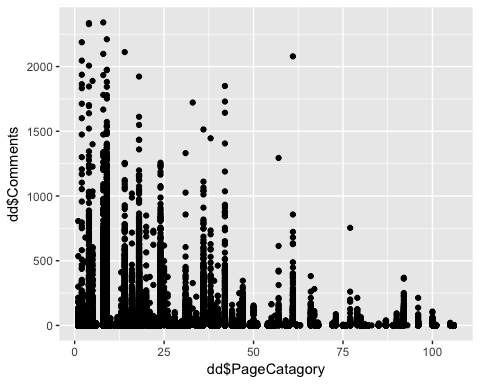
qplot(dd$Comments, dd$Comments2448)



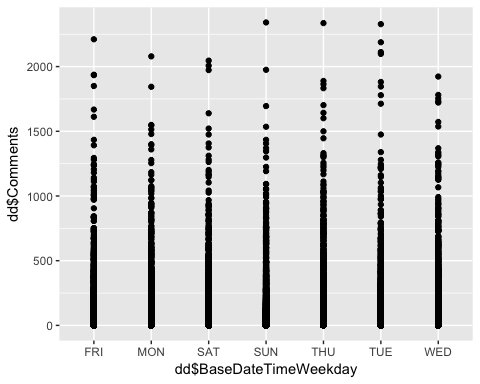
qplot(dd$Comments, dd$CommentDiff2448)



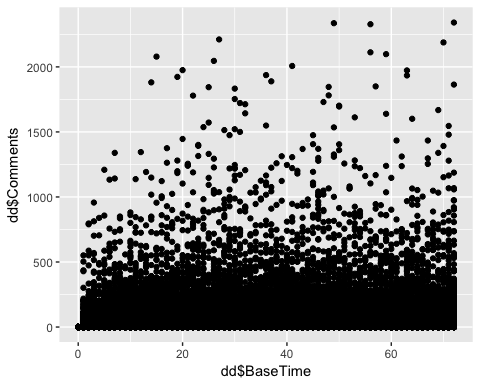
qplot(dd$PageCatagory, dd$Comments)



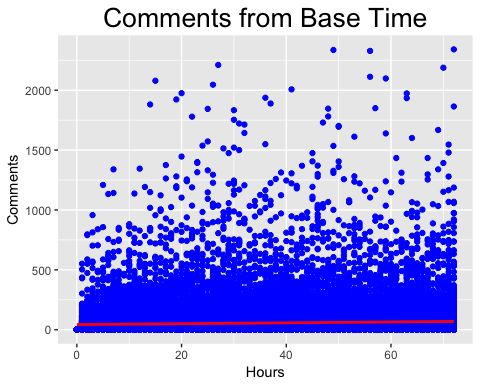
#histogram on Weekday values  
qplot(dd$BaseDateTimeWeekday, dd$Comments)



#looks like the number of comments is clustered under 1000 comments, you would think we would see a spike on the WeekEnds.   
  
#comments vs baseline  
qplot(dd$BaseTime, dd$Comments)



#scatter plot of number of comments vs the basetime  
ggplot(dd, aes(x=dd$BaseTime, y=dd$Comments)) +  
 geom\_point(color="blue") +  
 labs(x="Hours", y="Comments", title="Comments from Base Time") +   
 theme(plot.title = element\_text(hjust = 0.5, size = 20) ) +  
 geom\_smooth(colour="red", method="lm")



#ggplot shows the comment volume of time it linear in fashion  
  
#attach will enable the ability to only the variable not the whole dd$Variable just Variable  
attach(dd)  
  
#need to determine which variables are statistically significant to use in the model.  
#lets examine the data before creating a model  
  
#3.1 ==========================================================================================  
#lets subset the data and look at there corvariance  
#total comments in relation to basetime  
ef <- subset(dd, select = c("BaseTime", "Comments"))  
summary(ef)

## BaseTime Comments   
## Min. : 0.00 Min. : 0.00   
## 1st Qu.:17.00 1st Qu.: 2.00   
## Median :35.00 Median : 11.00   
## Mean :35.32 Mean : 55.72   
## 3rd Qu.:53.00 3rd Qu.: 46.00   
## Max. :72.00 Max. :2341.00

cor(ef)

## BaseTime Comments  
## BaseTime 1.00000000 0.05567851  
## Comments 0.05567851 1.00000000

ef <- subset(dd, select = c("Comments", "PagePopularity")) # strong corelation  
summary(ef)

## Comments PagePopularity   
## Min. : 0.00 Min. : 36   
## 1st Qu.: 2.00 1st Qu.: 36734   
## Median : 11.00 Median : 292911   
## Mean : 55.72 Mean : 1313814   
## 3rd Qu.: 46.00 3rd Qu.: 1204214   
## Max. :2341.00 Max. :486972297

cor(ef)

## Comments PagePopularity  
## Comments 1.0000000 0.1056238  
## PagePopularity 0.1056238 1.0000000

ef <- subset(dd, select = c("Comments", "PageCatagory")) # weak corelation  
summary(ef)

## Comments PageCatagory   
## Min. : 0.00 Min. : 1.00   
## 1st Qu.: 2.00 1st Qu.: 9.00   
## Median : 11.00 Median : 18.00   
## Mean : 55.72 Mean : 24.25   
## 3rd Qu.: 46.00 3rd Qu.: 32.00   
## Max. :2341.00 Max. :106.00

cor(ef)

## Comments PageCatagory  
## Comments 1.0000000 -0.1459321  
## PageCatagory -0.1459321 1.0000000

ef <- subset(dd, select = c("Comments", "PostLength")) # weak corelation  
summary(ef)

## Comments PostLength   
## Min. : 0.00 Min. : 0.0   
## 1st Qu.: 2.00 1st Qu.: 38.0   
## Median : 11.00 Median : 97.0   
## Mean : 55.72 Mean : 163.7   
## 3rd Qu.: 46.00 3rd Qu.: 172.0   
## Max. :2341.00 Max. :21480.0

cor(ef)

## Comments PostLength  
## Comments 1.000000000 -0.005036467  
## PostLength -0.005036467 1.000000000

ef <- subset(dd, select = c("Comments", "PostShareCount")) # strong corelation  
summary(ef)

## Comments PostShareCount   
## Min. : 0.00 Min. : 1.0   
## 1st Qu.: 2.00 1st Qu.: 2.0   
## Median : 11.00 Median : 13.0   
## Mean : 55.72 Mean : 117.2   
## 3rd Qu.: 46.00 3rd Qu.: 61.0   
## Max. :2341.00 Max. :144860.0

cor(ef)

## Comments PostShareCount  
## Comments 1.000000 0.225711  
## PostShareCount 0.225711 1.000000

ef <- subset(dd, select = c("Comments", "Checkins")) # strong corelation  
summary(ef)

## Comments Checkins   
## Min. : 0.00 Min. : 0   
## 1st Qu.: 2.00 1st Qu.: 0   
## Median : 11.00 Median : 0   
## Mean : 55.72 Mean : 4676   
## 3rd Qu.: 46.00 3rd Qu.: 99   
## Max. :2341.00 Max. :186370

cor(ef)

## Comments Checkins  
## Comments 1.00000000 0.09835245  
## Checkins 0.09835245 1.00000000

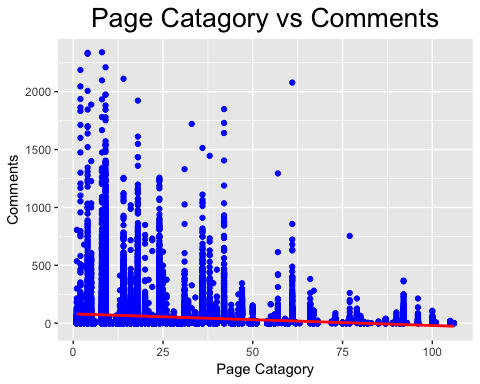
ef <- subset(dd, select = c("Comments", "PageTalkingAbout")) # strong corelation  
summary(ef)

## Comments PageTalkingAbout   
## Min. : 0.00 Min. : 0   
## 1st Qu.: 2.00 1st Qu.: 698   
## Median : 11.00 Median : 7045   
## Mean : 55.72 Mean : 44800   
## 3rd Qu.: 46.00 3rd Qu.: 50264   
## Max. :2341.00 Max. :6089942

cor(ef)

## Comments PageTalkingAbout  
## Comments 1.0000000 0.3291393  
## PageTalkingAbout 0.3291393 1.0000000

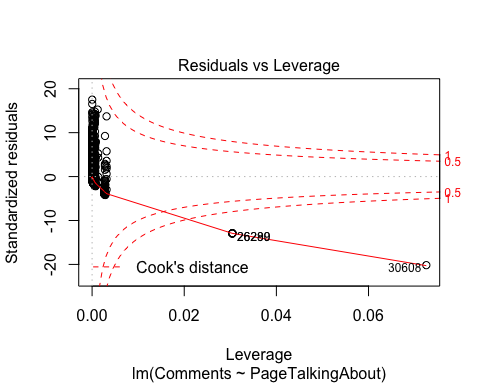
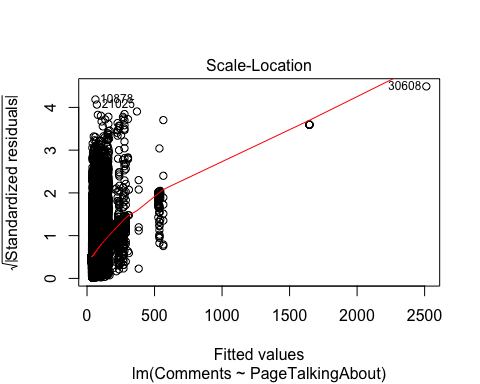
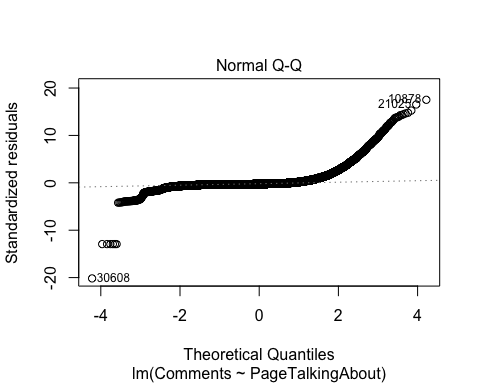
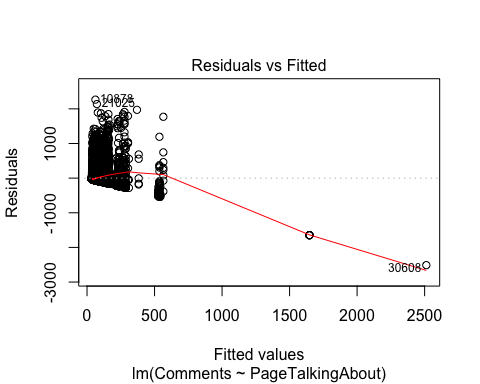
#lets check out if all catagories are in the data set and if there is a trend  
ggplot(dd, aes(x=dd$PageCatagory, dd$Comments)) +  
 geom\_point(color="blue") +  
 labs(x="Page Catagory", y="Comments", title="Page Catagory vs Comments") +   
 theme(plot.title = element\_text(hjust = 0.5, size = 20) ) +  
 geom\_smooth(colour="red", method="lm")



#since this is a simple regression we will only use one variable   
#create a model by page talking about showed the strongest correlation  
model.lm <- lm(Comments ~ PageTalkingAbout, data = dd)  
summary(model.lm)

##   
## Call:  
## lm(formula = Comments ~ PageTalkingAbout, data = dd)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2512.50 -37.60 -33.51 -10.17 2265.83   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 37.513321847 0.689344690 54.42 <0.0000000000000002 \*\*\*  
## PageTalkingAbout 0.000406405 0.000005762 70.53 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 129.3 on 40947 degrees of freedom  
## Multiple R-squared: 0.1083, Adjusted R-squared: 0.1083   
## F-statistic: 4975 on 1 and 40947 DF, p-value: < 0.00000000000000022

plot(model.lm)



#Regression equation  
#CommentsVolume = 37.513321847 + 0.000406405 \* Comments  
  
#predictor pagetalkingabout predict will tell you how many comments in x from baseline   
#BaseTime = x  
newdata = data.frame(PageTalkingAbout = 64641) #67 comments  
predict(model.lm, newdata)

## 1   
## 63.78377

newdata = data.frame(PageTalkingAbout = 57833) #61 comments  
predict(model.lm, newdata)

## 1   
## 61.01696

newdata = data.frame(PageTalkingAbout = 1242488) #542 comments  
predict(model.lm, newdata)

## 1   
## 542.4671

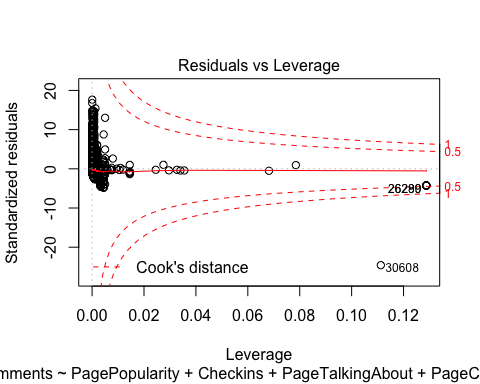
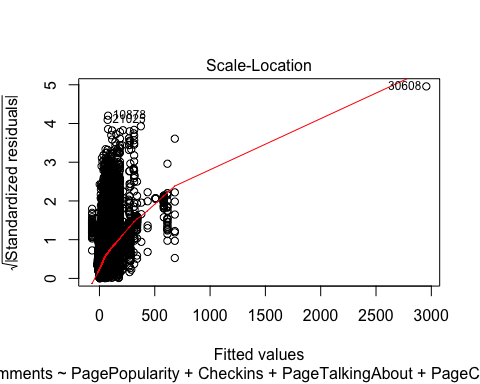
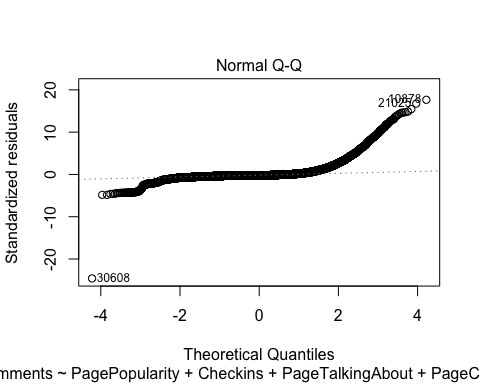
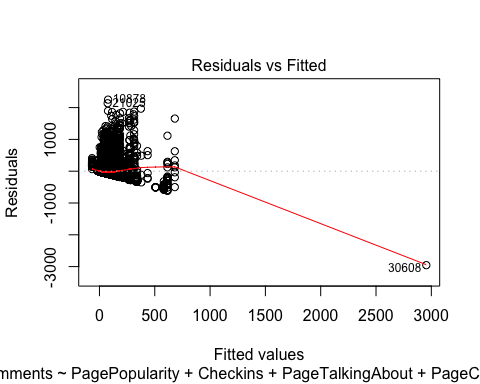
newdata = data.frame(PageTalkingAbout = 1) #37 comments  
predict(model.lm, newdata)

## 1   
## 37.51373

#3.2 ============================================================================================  
# Multiple Regression  
  
model2 <- lm(Comments ~ PagePopularity + Checkins + PageTalkingAbout + PageCatagory + PostLength, data = dd)  
summary(model2)

##   
## Call:  
## lm(formula = Comments ~ PagePopularity + Checkins + PageTalkingAbout +   
## PageCatagory + PostLength, data = dd)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2954.67 -40.37 -28.74 -2.94 2250.15   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 51.2161358839 1.0998896823 46.565 < 0.0000000000000002  
## PagePopularity -0.0000030748 0.0000001193 -25.764 < 0.0000000000000002  
## Checkins 0.0002137216 0.0000311392 6.863 0.00000000000682  
## PageTalkingAbout 0.0005005108 0.0000074654 67.044 < 0.0000000000000002  
## PageCatagory -0.6195993744 0.0320420518 -19.337 < 0.0000000000000002  
## PostLength 0.0009154428 0.0016757852 0.546 0.585  
##   
## (Intercept) \*\*\*  
## PagePopularity \*\*\*  
## Checkins \*\*\*  
## PageTalkingAbout \*\*\*  
## PageCatagory \*\*\*  
## PostLength   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 127.5 on 40943 degrees of freedom  
## Multiple R-squared: 0.1336, Adjusted R-squared: 0.1335   
## F-statistic: 1263 on 5 and 40943 DF, p-value: < 0.00000000000000022

#the summary show that Checkins and PostLength doesn't relate to comment volume.  
plot(model2)



#ouput the coefficients and show fitted values  
model2

##   
## Call:  
## lm(formula = Comments ~ PagePopularity + Checkins + PageTalkingAbout +   
## PageCatagory + PostLength, data = dd)  
##   
## Coefficients:  
## (Intercept) PagePopularity Checkins PageTalkingAbout   
## 51.216135884 -0.000003075 0.000213722 0.000500511   
## PageCatagory PostLength   
## -0.619599374 0.000915443

# y = 51.2161358839 + -0.0000030748 + 0.0002137216 + 0.0005005108 + -0.6195993744 + 0.0009154428  
  
#The output shows that F = 322.5(p < 0.00000000000000022), indicating that we should clearly reject the null hypothesis that the variables  
#PostLength, Checkins collectively have no effect on Comment Volume. The results also show that the variable PagePopularity is significant controlling for  
#the variable PageTalkingAbout(p = 67.044), as is Comment Volume controlling for the variable Comments(p = 46.565).  
#In addition, the output also shows that R2 = 0.1336 and R2adjusted = 0.1335.  
#We could perform a partial F-test by creating a full model and a reduced model then compare them with an ANOVA analysis  
  
#3.3 ==============================================================================================================  
#Free form  
#lets create a new model and remove Checkins and Post Length and PageTalkingAbout  
model3 <- lm(Comments ~ PagePopularity + PageTalkingAbout, data = dd)  
summary(model3)

##   
## Call:  
## lm(formula = Comments ~ PagePopularity + PageTalkingAbout, data = dd)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3144.80 -36.22 -31.59 -9.41 2260.98   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 36.2157592680 0.6846777956 52.90 <0.0000000000000002  
## PagePopularity -0.0000032879 0.0000001194 -27.54 <0.0000000000000002  
## PageTalkingAbout 0.0005317888 0.0000073022 72.83 <0.0000000000000002  
##   
## (Intercept) \*\*\*  
## PagePopularity \*\*\*  
## PageTalkingAbout \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 128.2 on 40946 degrees of freedom  
## Multiple R-squared: 0.1246, Adjusted R-squared: 0.1245   
## F-statistic: 2913 on 2 and 40946 DF, p-value: < 0.00000000000000022

#Prediction Interval 95% CI  
predict(model3, data.frame(PagePopularity = 456, PageTalkingAbout = 1242488), interval = "confidence")

## fit lwr upr  
## 1 696.9555 679.576 714.3349

#A 95 % confidence interval is given by(679.576, 714.3349)  
#the numbers above indicate that with a page popularity of 456 and page talking about of 1224488 will get at least 679-714 comments from baseline  
  
predict(model3, data.frame(PagePopularity = 162624443, PageTalkingAbout = 1242488), interval = "prediction")

## fit lwr upr  
## 1 162.269 -90.7484 415.2865

#A 95 % prediction interval is given by(-90.7484, 415.2865) . Note that this is quite a bit wider than the confidence interval,   
#indicating that the variation about the mean is fairly large.  
  
##Conclusion  
#With the three models complete the best predictors for comments volume from basetime are   
#Page Popularity and Page Talking Abount