#session13_assignment_13.1

#1. Use the below given data set

DataSet

#Problem- prediction of the number of comments in the upcoming 24 hours on those blogs,

#The train data was generated from different base times that may temporally overlap.

#Therefore, if you simply split the train into disjoint partitions, the underlying time intervals may overlap.

#Therefore, the you should use the provided, temporally disjoint train and test splits to ensure that the evaluation is fair.

```
library(readr)

library(data.table)

library(foreach)

getwd()

path="C:/Users/Swapna/Documents"

setwd(path)


train<-fread("C:/Users/Swapna/Documents/R files test/BlogFeedback/blogData_train.csv")

View(train)


test1<-fread("C:/Users/Swapna/Documents/R files test/BlogFeedback/blogData_test-2012.02.01.00_00.csv")

test2<-fread("C:/Users/Swapna/Documents/R files test/BlogFeedback/blogData_test-2012.02.06.00_00.csv")

test3<-fread("C:/Users/Swapna/Documents/R files test/BlogFeedback/blogData_test-2012.02.12.00_00.csv")

test4<-fread("C:/Users/Swapna/Documents/R files test/BlogFeedback/blogData_test-2012.02.18.00_00.csv")

test5<-fread("C:/Users/Swapna/Documents/R files test/BlogFeedback/blogData_test-2012.02.24.00_00.csv")

test6<-fread("C:/Users/Swapna/Documents/R files test/BlogFeedback/blogData_test-
2012.02.29.00_00.csv")

test7<-fread("C:/Users/Swapna/Documents/R files test/BlogFeedback/blogData_test-2012.03.01.00_00.csv")

test8<-fread("C:/Users/Swapna/Documents/R files test/BlogFeedback/blogData_test-2012.03.10.00_00.csv")

test9<-fread("C:/Users/Swapna/Documents/R files test/BlogFeedback/blogData_test-2012.03.20.00_00.csv")

test10<-fread("C:/Users/Swapna/Documents/R files test/BlogFeedback/blogData_test-2012.03.31.01_00.csv")

test<-rbind(test1,test2,test3,test4,test5,test6,test7,test8,test9,test10)

View(test)

# log-transform

train[, V281 := log(1 + V281)]

test[, V281 := log(1 + V281)]

# drop continous variables without variation

drop = c(8, 13, 28, 33, 38, 40, 43, 50, 278)

train[, (drop) := NULL]

test[, (drop) := NULL]

# write to files

write.csv(train, "BlogFeedback-Train.csv", row.names = F)

write.csv(test, "BlogFeedback-Test.csv", row.names = F)
```

#a. Read the dataset and identify the right features

```
# log-transform

train[, V281 := log(1 + V281)]

test[, V281 := log(1 + V281)]
```

#b. Clean dataset, impute missing values and perform exploratory data analysis.

```
# drop continous variables without variation

drop = c(8, 13, 28, 33, 38, 40, 43, 50, 278)
```

```
train[, (drop) := NULL]
test[, (drop) := NULL]

str(train)
table(train)

# write to files
write.csv(train, "BlogFeedback-Train.csv", row.names = F)
write.csv(test, "BlogFeedback-Test.csv", row.names = F)

# missing values
sum(is.na(train))
sum(is.na(test))

is.na(train)
#c. Visualize the dataset and make inferences from that
library(ggplot2)

gg <- ggplot(train, aes(x=V16, y=V281)) +
geom_point() +
geom_smooth(method="loess", se=F) +
labs(subtitle="Visualization of blog train",
y="V281",
x="V16",
title="Scatterplot")

plot(gg) # show data set is right sweked with ouliers


hist(train$V4) # column V4 is right distributed , right skewed

barplot(train$V237)

#d. Perform any 3 hypothesis tests using columns of your choice, make conclusions


wilcox.test(test$V21, data = test)
```

```
Wilcoxon signed rank test with continuity
correction

data: test$V281
V = 517640, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 0

# T test

t.test(test$V281)

One Sample t-test

data: test$V281
t = 21.33, df = 1327, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
0.2533693 0.3046945
sample estimates:
mean of x
0.2790319

t.test(test$V100)

One Sample t-test

data: test$V100
t = NaN, df = 1327, p-value = NA
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
NaN NaN
sample estimates:
mean of x
```

```
0

# correlation test at .95 significance level between two independent variables

cor.test(train$V4,train$V214)

#Pearson's product-moment correlation

data: train$V4 and train$V214
t = 2.5913, df = 52395, p-value = 0.009565
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.002757668 0.019880320
sample estimates:
cor
0.01131982

#e. Create a linear regression model to predict the number of comments in the next 24 hours
#(relative to basetime)

library(tree)
library(C50)

model<-tree(train$V281~.,data = train) # tree based model for non linear complex data
model
summary(model)
model1<-lm(train$V281~., data = train)
model1
summary(model1)
```