library(readxl)

Homezilla\_Dataset <- read\_excel("C:/Users/Sowmya/Desktop/fall2018/IDS515/Homezilla/Homezilla Dataset.xlsx",

sheet = "62 Listing Properties")

View(Homezilla\_Dataset)

listing\_properties<-Homezilla\_Dataset

> str(listing\_properties)

Classes ‘tbl\_df’, ‘tbl’ and 'data.frame': 62 obs. of 10 variables:

$ Web ID : chr "F1410261" "V1089633" "V1052961" "V1071997" ...

$ type : chr "house" "house" "house" "house" ...

$ subtype : chr "Single Family Detached" "Condo Apartment" "Single Family Detached" "Single Family Detached" ...

$ sqfoot : num 4115 990 4359 2769 1975 ...

$ bedrooms : num 3 2 3 5 3 6 2 2 1 2 ...

$ bathrooms : num 4 2 4 4 3 4 2 1 1 1 ...

$ half baths : num 0 0 0 0 0 0 0 0 0 0 ...

$ price : num 1050000 199900 1399000 2798000 598800 ...

$ status : chr "STACT" "STACT" "STINA" "STINA" ...

$ last update: chr "2014-07-09 155732" "2014-10-10 152509" "2014-09-15 133614" "2014-08-19 154230" ...

is.na(listing\_properties)

colnames(listing\_properties) <- c("Web\_ID", "type", "subtype", "sqfoot", "bedrooms", "bathrooms",

"half\_baths", "price", "status", "last\_update")

library(readxl)

Homezilla\_Dataset <- read\_excel("C:/Users/Sowmya/Desktop/fall2018/IDS515/Homezilla/Homezilla Dataset.xlsx",

sheet = "Browsing Data")

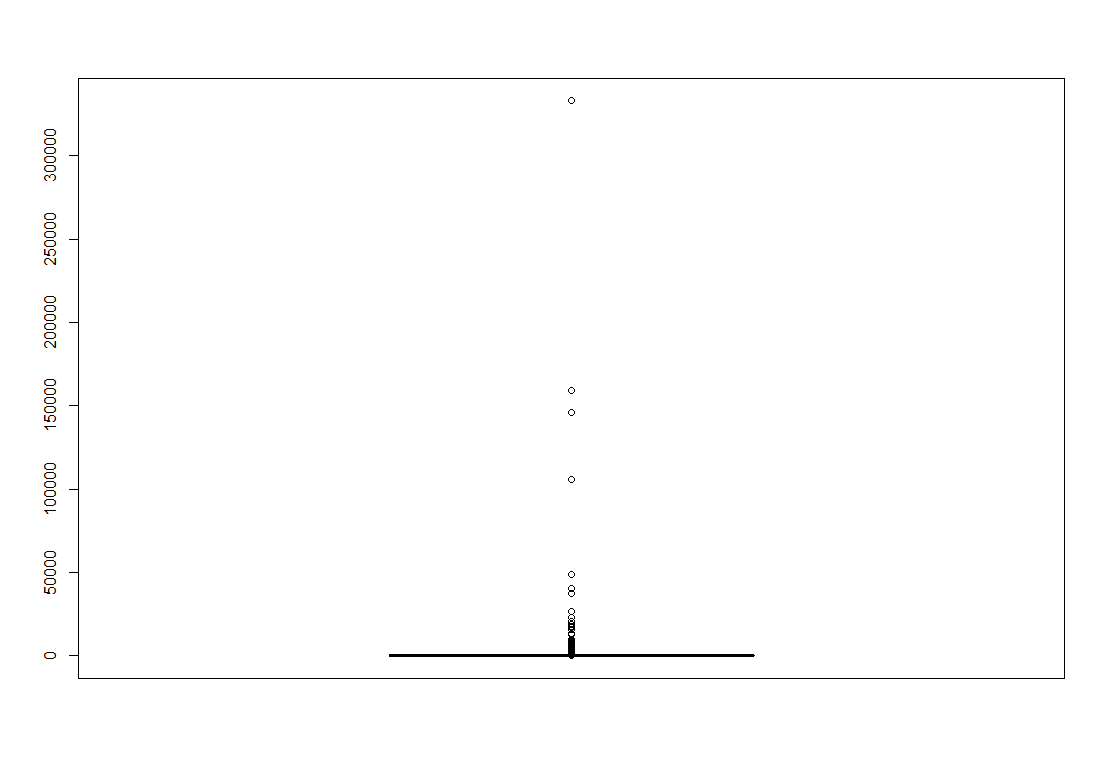
browsing\_data<-Homezilla\_Dataset

**Data Cleaning:**

colnames(browsing\_data) <- c("Web\_ID", "Time\_Viewed", "Timestamp", "Direction", "Photo\_ID", "PhotoTag1", "PhotoTag2", "PhotoTag3", "PhotoTag4", "PhotoTag5", "PhotoTag6", "PhotoTag7", "PhotoTag8", "UserAgent","CustomerID")

#boxplot(browsing\_data$Time\_Viewed)

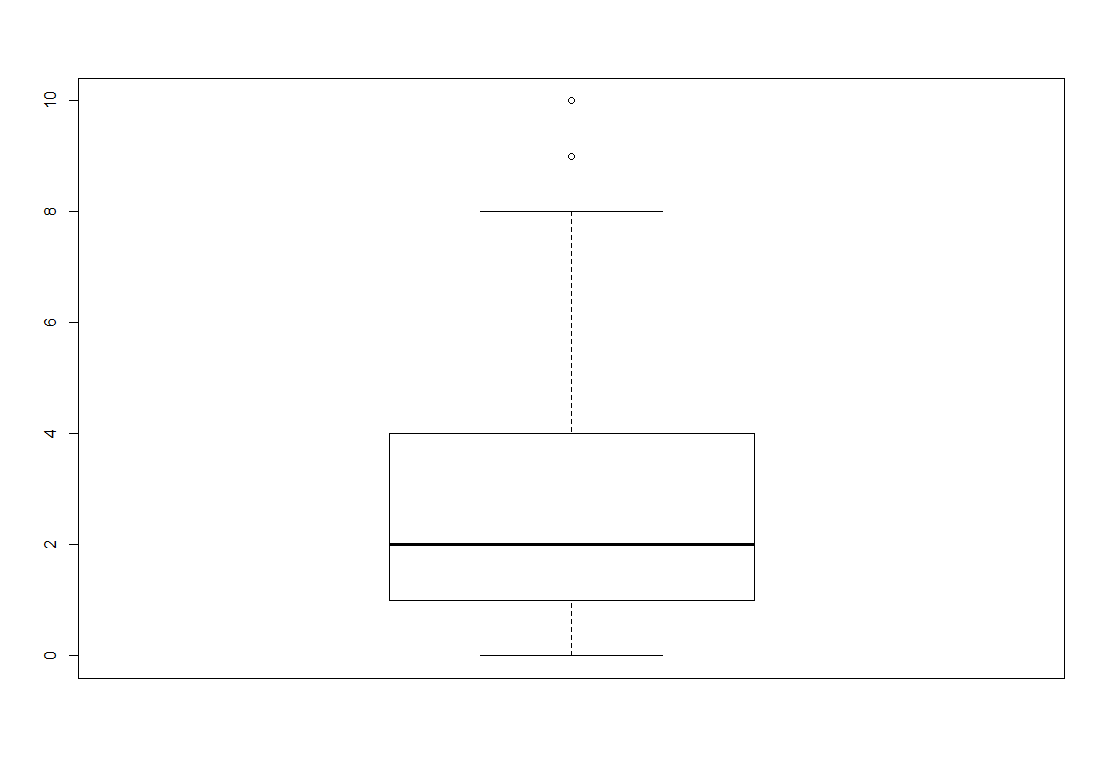
boxplot(browsing\_data$Time\_Viewed)



outliersBrowse = boxplot.stats(browsing\_data$Time\_Viewed, coef = 2)$out

browsing\_data$Time\_Viewed = ifelse(browsing\_data$Time\_Viewed %in% outliersBrowse, NA, browsing\_data$Time\_Viewed)

boxplot(browsing\_data$Time\_Viewed)



browsing\_data$Direction <- as.factor(browsing\_data$Direction)

browsing\_data$Direction[is.na(browsing\_data$Direction)] <- "First"

levels(browsing\_data$Direction)

> levels(browsing\_data$Direction)

[1] "First" "left" "right"

browsing\_data$Time\_Viewed[is.na(browsing\_data$Time\_Viewed)] = mean(browsing\_data$Time\_Viewed, na.rm=TRUE)

browsing\_data$Time\_Viewed <- as.factor(ifelse(browsing\_data$Time\_Viewed >=3, 1,0))

levels(browsing\_data$Time\_Viewed)

> levels(browsing\_data$Time\_Viewed)

[1] "0" "1"

browsing\_data$PhotoTag3[is.na(browsing\_data$PhotoTag3)] <- "Null"

browsing\_data$PhotoTag3<- as.factor(browsing\_data$PhotoTag3)

browsing\_data$PhotoTag4[is.na(browsing\_data$PhotoTag4)] <- "Null"

browsing\_data$PhotoTag4<- as.factor(browsing\_data$PhotoTag4)

browsing\_data$PhotoTag5[is.na(browsing\_data$PhotoTag5)] <- "Null"

browsing\_data$PhotoTag5<- as.factor(browsing\_data$PhotoTag5)

browsing\_data$PhotoTag6[is.na(browsing\_data$PhotoTag6)] <- "Null"

browsing\_data$PhotoTag6<- as.factor(browsing\_data$PhotoTag6)

browsing\_data$PhotoTag7[is.na(browsing\_data$PhotoTag7)] <- "Null"

browsing\_data$PhotoTag7<- as.factor(browsing\_data$PhotoTag7)

browsing\_data$PhotoTag8[is.na(browsing\_data$PhotoTag8)] <- "Null"

browsing\_data$PhotoTag8<- as.factor(browsing\_data$PhotoTag8)

**Filter and Join dataset:**

#filter data frame based on subtype

condo\_webid<-as.data.frame(listing\_properties[ listing\_properties$subtype == "Condo Apartment", 1:3])

is.na(condo\_webid)

condo\_webid<-na.omit(condo\_webid)

head(condo\_webid)

townhouse<-as.data.frame(listing\_properties[ listing\_properties$subtype == "Townhouse", 1:3])

is.na(townhouse)

townhouse<-na.omit(townhouse)

head(townhouse)

single\_family <- as.data.frame(listing\_properties[

listing\_properties$subtype == "Single Family Detached", 1:3 ])

is.na(single\_family)

single\_family<-na.omit(single\_family)

head(single\_family)

#join with browsing data set

head(browsing\_data)

is.na(browsing\_data$Web\_ID) #no rows without webid

single\_web<-merge(single\_family, browsing\_data, by="Web\_ID")

head(single\_web)

nrow(single\_web)

condo\_web<-merge(condo\_webid, browsing\_data, by="Web\_ID")

head(condo\_web)

nrow(condo\_web)

townHouse\_web<-merge(townhouse, browsing\_data, by="Web\_ID")

head(townHouse\_web)

nrow(townHouse\_web)

**Logistic regression**

#Model for condo apartment

myFormula<- Time\_Viewed ~ Direction + PhotoTag2 + PhotoTag3 + PhotoTag4 + PhotoTag5 + PhotoTag6 + PhotoTag7

condo\_regression<-glm(myFormula, data = condo\_web, family = binomial(logit))

summary(condo\_regression)

#modify based on NA

levels(condo\_web$PhotoTag4)[levels(condo\_web$PhotoTag4)=="machines-visible"] <- "Null"

levels(condo\_web$PhotoTag6)[levels(condo\_web$PhotoTag6)=="living-window"] <- "Null"

levels(condo\_web$PhotoTag3)[levels(condo\_web$PhotoTag3)=="tableview"] <- "Null"

levels(condo\_web$PhotoTag5)[levels(condo\_web$PhotoTag5)=="living-window"] <- "Null"

levels(condo\_web$PhotoTag5)[levels(condo\_web$PhotoTag5)=="machines-not-visible"] <- "Null"

levels(condo\_web$PhotoTag6)[levels(condo\_web$PhotoTag6)=="bath-sink"] <- "Null"

levels(condo\_web$PhotoTag3)[levels(condo\_web$PhotoTag3)=="bath-tub"] <- "Null"

levels(condo\_web$PhotoTag3)[levels(condo\_web$PhotoTag3)=="machines-visible"] <- "Null"

myFormula<- Time\_Viewed ~ Direction + PhotoTag2 + PhotoTag2 + PhotoTag3

condo\_regression<-glm(myFormula, data = condo\_web, family = binomial(logit))

summary(condo\_regression)

#Predict the factors affecting single\_family house types

myFormula<- Time\_Viewed ~ Direction + PhotoTag2 + PhotoTag3 + PhotoTag4 + PhotoTag5

single\_regression<-glm(myFormula, data = single\_web, family =binomial(logit))

summary(single\_regression)

#Model for TownHouse

myFormula<- Time\_Viewed ~ Direction + PhotoTag2 + PhotoTag3 + PhotoTag4

town\_regression<-glm(myFormula, data = townHouse\_web, family = binomial(logit))

summary(town\_regression)

**Variable selection:**

library(Rcmdr)

finalCondo<-stepwise(condo\_regression, direction = "forward", criterion = "BIC")

summary(finalCondo)

#Smaller the AIC better the model