TTC Bus Delay Analysis using Linear Predictive Model of Delay-Causing Incidents

Load readxl & stringr libraries

library(readxl)  
library(stringr)  
library(ggplot2)  
require(coefplot)

## Loading required package: coefplot

library(tree)  
library(rpart)  
library(caret)

## Loading required package: lattice

# Create one dataframe for the entire worksheet

df <- data.frame()  
  
for (b in c(14,15,16,17,18)){  
   
  
 x <- sprintf("D:/DataAnalytics/CKME\_136/TTC\_Bus\_Delay/Capstone/Bus\_20%s.xlsx",b)  
  
   
 y <- excel\_sheets(x)  
   
   
 # Create one dataframe for the entire worksheet  
   
 for (a in y){  
   
 df1 <- read\_excel(x,sheet = a)  
   
 df <- rbind(df,df1)  
   
 }  
   
}

# Store Combined file in one CSV file

write.csv(df,"D:/DataAnalytics/CKME\_136/TTC\_Bus\_Delay/Capstone/combined.csv")

# Load the combined file

delay <- read.csv("D:/DataAnalytics/CKME\_136/TTC\_Bus\_Delay/Capstone/combined.csv", header = TRUE, stringsAsFactors = FALSE, sep = ",")  
  
head(delay)

## X Report.Date Route Time Day Location  
## 1 1 2014-01-01 95 1899-12-31 00:23:00 Wednesday York Mills station  
## 2 2 2014-01-01 102 1899-12-31 00:55:00 Wednesday Entire run for route  
## 3 3 2014-01-01 54 1899-12-31 01:28:00 Wednesday lawrence and Warden  
## 4 4 2014-01-01 112 1899-12-31 01:30:00 Wednesday Kipling Station  
## 5 5 2014-01-01 24 1899-12-31 01:37:00 Wednesday VP and Ellesmere  
## 6 6 2014-01-01 129 1899-12-31 01:50:00 Wednesday Scarborough Town  
## Incident Min.Delay Min.Gap Direction Vehicle  
## 1 Mechanical 10 20 E 1734  
## 2 General Delay 33 66 b/w 8110  
## 3 Mechanical 10 20 WB 7478  
## 4 Emergency Services 18 36 N 8084  
## 5 Investigation 10 20 n 7843  
## 6 Mechanical 10 20 N 1755

delay$Incident <- as.factor(delay$Incident)  
  
delay$Day <- as.factor(delay$Day)  
  
class(delay$Incident)

## [1] "factor"

class(delay$Day)

## [1] "factor"

summary(delay$Incident)

## Diversion Emergency Services General Delay   
## 20648 6934 49782   
## Investigation Late Leaving Garage Mechanical   
## 32132 57234 141163   
## Utilized Off Route   
## 60867

# Data Cleaning

# Clean direction feature

direct\_fun <- function(x){  
   
 switch(x,  
 "W/N"="WNB","w/b"="WB","b/w"="BW","w.b"="WB","w"="WB","UP"="NB","u"="NB",  
 "SS"="SB","s/d"="SB","s/b"="SB","s"="SB","Ou"="NB","W"="WB","N"="NB",  
 "S"="SB","E"="EB","n"="NB","s"="SB","e"="EB","WB"="WB","NB"="NB","SB"="SB",  
 "EB"="EB","BW"="BW","E/W"="BW","N/E"="NE","BW's"="BW","BWS"="BW","B/W's"="BW",  
 "b/ws"="BW","b/b"="BW","'N"="NB", "N & S"="BW","1w/b"="WB"  
  
 )  
}  
  
delay$Direction <- sapply(delay$Direction,direct\_fun)

# Renaming the columns

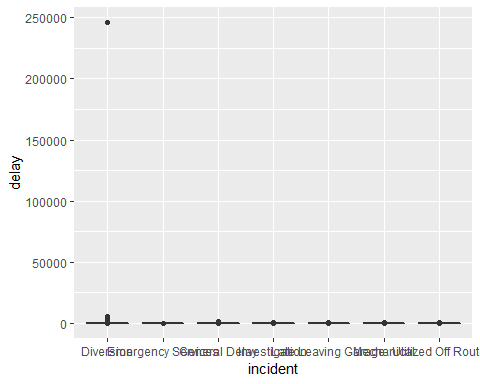
names(delay) <- c("SrNo","date","route","time","day","location","incident","delay","gap","direction","vehicle")  
  
delay$direction <- as.character(delay$direction)  
delay$direction <- as.factor(delay$direction)

# Removing “NULLs”

delay <- delay[delay$direction != "NULL",]

# removing “NAs”

delay <- delay[which(!is.na(delay$delay)),]  
  
  
ggplot(delay,aes(y=delay,x=incident))+geom\_boxplot()



# removing outliers

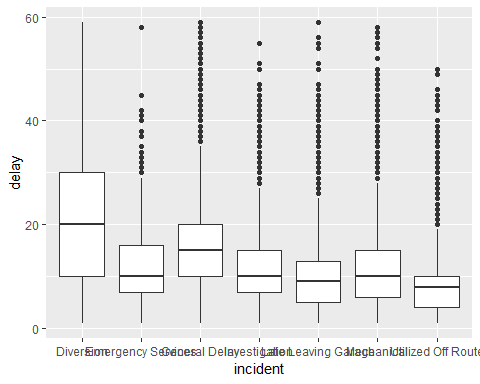
delay <- delay[which(delay$delay < 60 & delay$delay > 0 ),]  
  
head(delay)

## SrNo date route time day location  
## 1 1 2014-01-01 95 1899-12-31 00:23:00 Wednesday York Mills station  
## 2 2 2014-01-01 102 1899-12-31 00:55:00 Wednesday Entire run for route  
## 3 3 2014-01-01 54 1899-12-31 01:28:00 Wednesday lawrence and Warden  
## 4 4 2014-01-01 112 1899-12-31 01:30:00 Wednesday Kipling Station  
## 5 5 2014-01-01 24 1899-12-31 01:37:00 Wednesday VP and Ellesmere  
## 6 6 2014-01-01 129 1899-12-31 01:50:00 Wednesday Scarborough Town  
## incident delay gap direction vehicle  
## 1 Mechanical 10 20 EB 1734  
## 2 General Delay 33 66 BW 8110  
## 3 Mechanical 10 20 WB 7478  
## 4 Emergency Services 18 36 NB 8084  
## 5 Investigation 10 20 NB 7843  
## 6 Mechanical 10 20 NB 1755

class(delay$date)

## [1] "character"

ggplot(delay,aes(y=delay,x=incident))+geom\_boxplot()



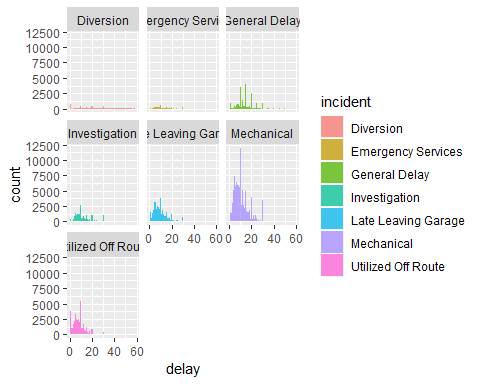
aggregate(delay ~ incident, data = delay, mean)

## incident delay  
## 1 Diversion 21.423803  
## 2 Emergency Services 12.475490  
## 3 General Delay 15.171343  
## 4 Investigation 11.833964  
## 5 Late Leaving Garage 10.164624  
## 6 Mechanical 11.180707  
## 7 Utilized Off Route 8.261773

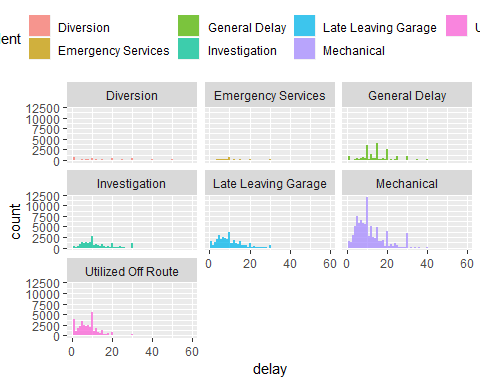
aggregate(delay ~ incident, data = delay, var)

## incident delay  
## 1 Diversion 245.32399  
## 2 Emergency Services 58.78262  
## 3 General Delay 64.11348  
## 4 Investigation 54.29743  
## 5 Late Leaving Garage 43.08443  
## 6 Mechanical 49.07003  
## 7 Utilized Off Route 30.76704

ggplot(delay, aes(x= delay,fill=incident))+geom\_histogram(binwidth = 1,alpha=3/4)+facet\_wrap(~incident)



ggplot(delay, aes(x= delay,fill=incident))+geom\_histogram(binwidth = 1,alpha=3/4)+facet\_wrap(~incident)+theme(legend.position = "top")



# Analysis of Variance “ANOVA” using linear regression coefficient

delay.lm <- lm(delay ~ incident, data = delay)  
summary(delay.lm)

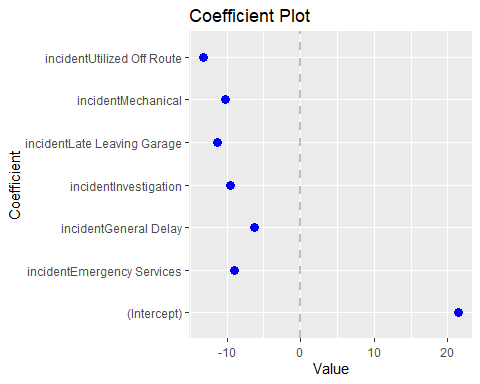
##   
## Call:  
## lm(formula = delay ~ incident, data = delay)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.424 -5.165 -1.181 3.166 48.835   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 21.42380 0.09430 227.18 <2e-16 \*\*\*  
## incidentEmergency Services -8.94831 0.15001 -59.65 <2e-16 \*\*\*  
## incidentGeneral Delay -6.25246 0.10730 -58.27 <2e-16 \*\*\*  
## incidentInvestigation -9.58984 0.10819 -88.64 <2e-16 \*\*\*  
## incidentLate Leaving Garage -11.25918 0.10236 -109.99 <2e-16 \*\*\*  
## incidentMechanical -10.24310 0.09738 -105.19 <2e-16 \*\*\*  
## incidentUtilized Off Route -13.16203 0.10197 -129.08 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.263 on 206624 degrees of freedom  
## Multiple R-squared: 0.1057, Adjusted R-squared: 0.1057   
## F-statistic: 4070 on 6 and 206624 DF, p-value: < 2.2e-16

delay.lm2 <- lm(delay ~ incident-1, data = delay)  
summary(delay.lm2)

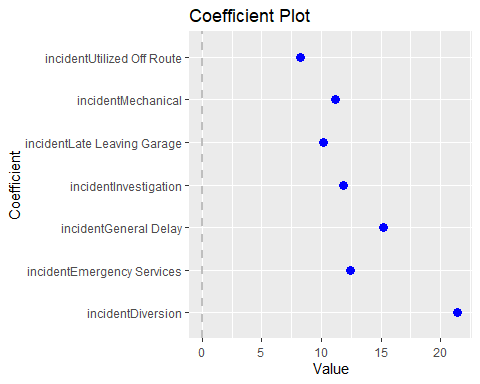
##   
## Call:  
## lm(formula = delay ~ incident - 1, data = delay)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.424 -5.165 -1.181 3.166 48.835   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## incidentDiversion 21.42380 0.09430 227.2 <2e-16 \*\*\*  
## incidentEmergency Services 12.47549 0.11666 106.9 <2e-16 \*\*\*  
## incidentGeneral Delay 15.17134 0.05119 296.4 <2e-16 \*\*\*  
## incidentInvestigation 11.83396 0.05303 223.2 <2e-16 \*\*\*  
## incidentLate Leaving Garage 10.16462 0.03981 255.3 <2e-16 \*\*\*  
## incidentMechanical 11.18071 0.02427 460.7 <2e-16 \*\*\*  
## incidentUtilized Off Route 8.26177 0.03878 213.0 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.263 on 206624 degrees of freedom  
## Multiple R-squared: 0.717, Adjusted R-squared: 0.717   
## F-statistic: 7.479e+04 on 7 and 206624 DF, p-value: < 2.2e-16

# Plot linear regression coefficient

coefplot(delay.lm)



coefplot(delay.lm2)

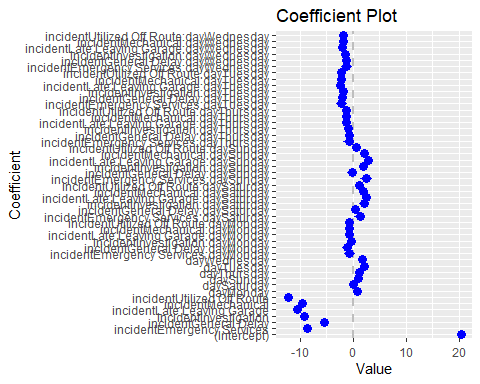


# Consider the interaction with the day of incident

delay.lm3 <- lm(delay ~ incident \* day, data = delay)  
summary(delay.lm3)

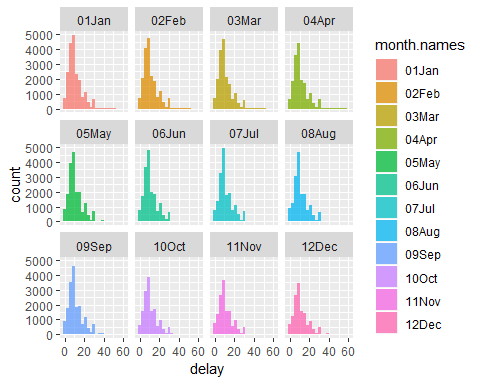
##   
## Call:  
## lm(formula = delay ~ incident \* day, data = delay)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.508 -4.825 -1.140 3.199 47.400   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 20.4116 0.2338 87.316  
## incidentEmergency Services -8.5990 0.3640 -23.625  
## incidentGeneral Delay -5.4490 0.2632 -20.706  
## incidentInvestigation -9.2719 0.2691 -34.454  
## incidentLate Leaving Garage -10.4880 0.2524 -41.552  
## incidentMechanical -9.6105 0.2413 -39.832  
## incidentUtilized Off Route -12.3447 0.2522 -48.953  
## dayMonday 0.8136 0.3376 2.410  
## daySaturday 0.0582 0.3549 0.164  
## daySunday 1.0343 0.3722 2.779  
## dayThursday 1.1969 0.3336 3.588  
## dayTuesday 2.0966 0.3341 6.275  
## dayWednesday 1.7193 0.3322 5.176  
## incidentEmergency Services:dayMonday -0.6819 0.5313 -1.284  
## incidentGeneral Delay:dayMonday -1.1951 0.3881 -3.079  
## incidentInvestigation:dayMonday -0.2950 0.3886 -0.759  
## incidentLate Leaving Garage:dayMonday -0.8058 0.3630 -2.220  
## incidentMechanical:dayMonday -0.6825 0.3480 -1.961  
## incidentUtilized Off Route:dayMonday -0.7836 0.3663 -2.139  
## incidentEmergency Services:daySaturday 1.2522 0.5539 2.261  
## incidentGeneral Delay:daySaturday 0.4111 0.4172 0.985  
## incidentInvestigation:daySaturday 2.0269 0.4090 4.955  
## incidentLate Leaving Garage:daySaturday 2.4355 0.3963 6.146  
## incidentMechanical:daySaturday 1.9660 0.3682 5.339  
## incidentUtilized Off Route:daySaturday 1.1676 0.3880 3.009  
## incidentEmergency Services:daySunday 2.4534 0.5932 4.136  
## incidentGeneral Delay:daySunday -0.2448 0.4559 -0.537  
## incidentInvestigation:daySunday 1.9408 0.4340 4.472  
## incidentLate Leaving Garage:daySunday 2.8705 0.4166 6.890  
## incidentMechanical:daySunday 2.0664 0.3865 5.346  
## incidentUtilized Off Route:daySunday 0.6799 0.4060 1.675  
## incidentEmergency Services:dayThursday -0.7443 0.5240 -1.420  
## incidentGeneral Delay:dayThursday -0.6701 0.3720 -1.802  
## incidentInvestigation:dayThursday -0.9745 0.3817 -2.553  
## incidentLate Leaving Garage:dayThursday -1.3740 0.3597 -3.819  
## incidentMechanical:dayThursday -1.2665 0.3441 -3.681  
## incidentUtilized Off Route:dayThursday -1.3290 0.3594 -3.698  
## incidentEmergency Services:dayTuesday -2.1633 0.5240 -4.129  
## incidentGeneral Delay:dayTuesday -2.1301 0.3744 -5.689  
## incidentInvestigation:dayTuesday -1.8562 0.3826 -4.851  
## incidentLate Leaving Garage:dayTuesday -2.5282 0.3596 -7.031  
## incidentMechanical:dayTuesday -2.3081 0.3444 -6.702  
## incidentUtilized Off Route:dayTuesday -2.2678 0.3588 -6.320  
## incidentEmergency Services:dayWednesday -1.2231 0.5347 -2.287  
## incidentGeneral Delay:dayWednesday -1.3133 0.3721 -3.530  
## incidentInvestigation:dayWednesday -1.4689 0.3802 -3.863  
## incidentLate Leaving Garage:dayWednesday -2.0992 0.3590 -5.848  
## incidentMechanical:dayWednesday -1.9201 0.3426 -5.604  
## incidentUtilized Off Route:dayWednesday -1.8203 0.3573 -5.095  
## Pr(>|t|)   
## (Intercept) < 2e-16 \*\*\*  
## incidentEmergency Services < 2e-16 \*\*\*  
## incidentGeneral Delay < 2e-16 \*\*\*  
## incidentInvestigation < 2e-16 \*\*\*  
## incidentLate Leaving Garage < 2e-16 \*\*\*  
## incidentMechanical < 2e-16 \*\*\*  
## incidentUtilized Off Route < 2e-16 \*\*\*  
## dayMonday 0.015962 \*   
## daySaturday 0.869739   
## daySunday 0.005452 \*\*   
## dayThursday 0.000333 \*\*\*  
## dayTuesday 3.50e-10 \*\*\*  
## dayWednesday 2.27e-07 \*\*\*  
## incidentEmergency Services:dayMonday 0.199311   
## incidentGeneral Delay:dayMonday 0.002075 \*\*   
## incidentInvestigation:dayMonday 0.447802   
## incidentLate Leaving Garage:dayMonday 0.026411 \*   
## incidentMechanical:dayMonday 0.049875 \*   
## incidentUtilized Off Route:dayMonday 0.032406 \*   
## incidentEmergency Services:daySaturday 0.023775 \*   
## incidentGeneral Delay:daySaturday 0.324467   
## incidentInvestigation:daySaturday 7.22e-07 \*\*\*  
## incidentLate Leaving Garage:daySaturday 7.99e-10 \*\*\*  
## incidentMechanical:daySaturday 9.34e-08 \*\*\*  
## incidentUtilized Off Route:daySaturday 0.002622 \*\*   
## incidentEmergency Services:daySunday 3.54e-05 \*\*\*  
## incidentGeneral Delay:daySunday 0.591360   
## incidentInvestigation:daySunday 7.75e-06 \*\*\*  
## incidentLate Leaving Garage:daySunday 5.59e-12 \*\*\*  
## incidentMechanical:daySunday 9.01e-08 \*\*\*  
## incidentUtilized Off Route:daySunday 0.094017 .   
## incidentEmergency Services:dayThursday 0.155532   
## incidentGeneral Delay:dayThursday 0.071625 .   
## incidentInvestigation:dayThursday 0.010675 \*   
## incidentLate Leaving Garage:dayThursday 0.000134 \*\*\*  
## incidentMechanical:dayThursday 0.000232 \*\*\*  
## incidentUtilized Off Route:dayThursday 0.000218 \*\*\*  
## incidentEmergency Services:dayTuesday 3.65e-05 \*\*\*  
## incidentGeneral Delay:dayTuesday 1.28e-08 \*\*\*  
## incidentInvestigation:dayTuesday 1.23e-06 \*\*\*  
## incidentLate Leaving Garage:dayTuesday 2.06e-12 \*\*\*  
## incidentMechanical:dayTuesday 2.06e-11 \*\*\*  
## incidentUtilized Off Route:dayTuesday 2.62e-10 \*\*\*  
## incidentEmergency Services:dayWednesday 0.022186 \*   
## incidentGeneral Delay:dayWednesday 0.000416 \*\*\*  
## incidentInvestigation:dayWednesday 0.000112 \*\*\*  
## incidentLate Leaving Garage:dayWednesday 5.00e-09 \*\*\*  
## incidentMechanical:dayWednesday 2.10e-08 \*\*\*  
## incidentUtilized Off Route:dayWednesday 3.49e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.205 on 206582 degrees of freedom  
## Multiple R-squared: 0.1201, Adjusted R-squared: 0.1199   
## F-statistic: 587.5 on 48 and 206582 DF, p-value: < 2.2e-16

coefplot(delay.lm3)

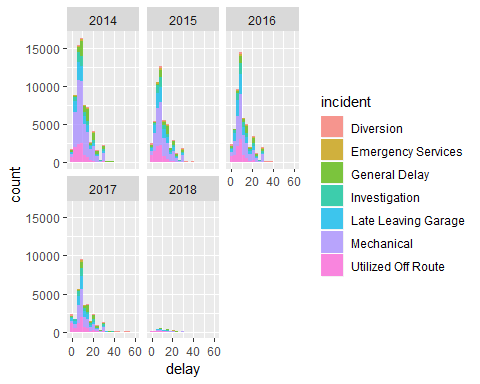


# Studying monthly distribution of the delay

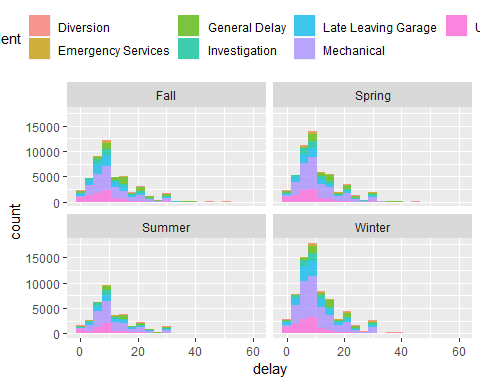
month.col <- str\_sub(delay$date,start = 6, end = 7)  
  
month.name <- function(x){  
   
 switch(x,  
 "01"="01Jan",  
 "02"="02Feb",  
 "03"="03Mar",  
 "04"="04Apr",  
 "05"="05May",  
 "06"="06Jun",  
 "07"="07Jul",  
 "08"="08Aug",  
 "09"="09Sep",  
 "10"="10Oct",  
 "11"="11Nov",  
 "12"="12Dec"  
 )  
   
}  
  
month.names <- sapply(month.col, month.name)  
  
delay <- data.frame(delay,month.names)  
  
ggplot(delay, aes(x= delay,fill=month.names))+geom\_histogram(binwidth =3,alpha=3/4)+facet\_wrap(~month.names)



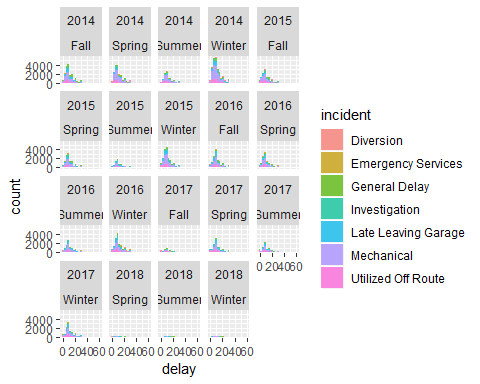
yearly <- str\_sub(delay$date,start = 1, end = 4)  
  
delay <- data.frame(delay,yearly)  
  
ggplot(delay, aes(x= delay,fill=incident))+geom\_histogram(binwidth = 3,alpha=3/4)+facet\_wrap(~yearly)



season <- function(x){  
   
 switch(x,  
 "01"="Winter",  
 "02"="Winter",  
 "03"="Winter",  
 "04"="Spring",  
 "05"="Spring",  
 "06"="Spring",  
 "07"="Summer",  
 "08"="Summer",  
 "09"="Fall",  
 "10"="Fall",  
 "11"="Fall",  
 "12"="Winter"  
 )  
   
}  
season.names <- sapply(month.col, season)  
  
delay <- data.frame(delay,season.names)  
  
ggplot(delay, aes(x= delay,fill=incident))+geom\_histogram(binwidth = 3,alpha=3/4)+facet\_wrap(~season.names)+theme(legend.position = "top")



ggplot(delay, aes(x= delay,fill=incident))+geom\_histogram(binwidth = 3,alpha=3/4)+facet\_wrap(~yearly+season.names)



# Create a prediction model for the delay through linear regression

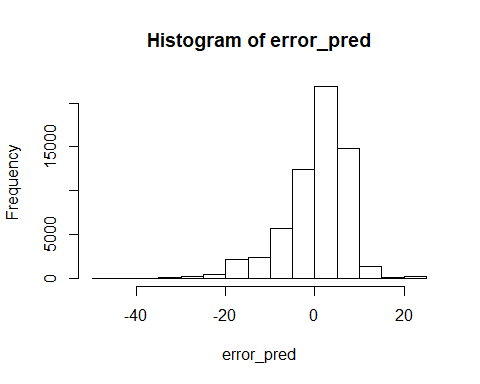
index <- sample(length(delay$route), .7 \*length(delay$route))  
  
train <- delay[index,]  
test <- delay[-index,]  
  
  
delay\_single <- lm(delay ~ incident-1, data = train)  
  
summary(delay\_single)

##   
## Call:  
## lm(formula = delay ~ incident - 1, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.700 -5.118 -1.179 3.197 46.821   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## incidentDiversion 21.70041 0.11253 192.83 <2e-16 \*\*\*  
## incidentEmergency Services 12.53541 0.13941 89.92 <2e-16 \*\*\*  
## incidentGeneral Delay 15.14111 0.06098 248.29 <2e-16 \*\*\*  
## incidentInvestigation 11.80319 0.06313 186.95 <2e-16 \*\*\*  
## incidentLate Leaving Garage 10.11770 0.04743 213.30 <2e-16 \*\*\*  
## incidentMechanical 11.17904 0.02893 386.39 <2e-16 \*\*\*  
## incidentUtilized Off Route 8.29739 0.04612 179.91 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.24 on 144634 degrees of freedom  
## Multiple R-squared: 0.7184, Adjusted R-squared: 0.7184   
## F-statistic: 5.272e+04 on 7 and 144634 DF, p-value: < 2.2e-16

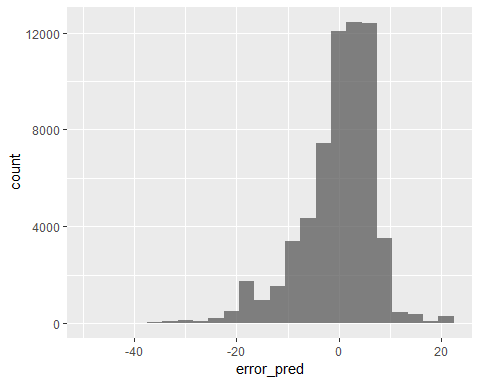
delay\_pred <- predict(delay\_single, newdata = test, interval = "prediction")  
  
pred <- data.frame(delay\_pred)  
  
  
error\_pred <- pred$fit - test$delay  
  
error <- data.frame(error\_pred)  
head(error)

## error\_pred  
## 1 -4.8209649  
## 2 4.1790351  
## 3 -13.1968061  
## 4 0.1176976  
## 5 -0.7026051  
## 6 -17.8209649

hist(error\_pred)



ggplot(error,aes(x=error\_pred))+geom\_histogram(binwidth = 3,alpha=3/4)



summary(error\_pred)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -48.88230 -3.19681 1.17904 0.00363 5.14111 20.70041

mse\_single <- sqrt(sum((error\_pred)^2)/nrow(test))  
  
mse\_single

## [1] 7.318057

relative\_error <- 1- ((test$delay - abs(error\_pred))/test$delay)  
#relative\_error  
relative\_error\_less25 <- table(relative\_error < .25)["TRUE"]/nrow(test)  
relative\_error\_less25\_pct <- paste("relative\_error\_less 25% is ",100\*relative\_error\_less25)  
relative\_error\_less25\_pct

## [1] "relative\_error\_less 25% is 33.4941119535409"

# Hypothesis: delays due to “Mechanical” incidents has no significant difference from delays caused by “All incidents”

Mechanical\_mean <- mean(delay$delay[delay$incident == "Mechanical"])  
  
Mechanical\_mean

## [1] 11.18071

Mechanical\_sd <- sd(delay$delay[delay$incident == "Mechanical"])  
  
Mechanical\_var <- var(delay$delay[delay$incident == "Mechanical"])  
  
  
  
Population\_mean <- mean(delay$delay)  
  
Population\_mean

## [1] 11.28802

Population\_sd <- sd(delay$delay)  
  
Population\_sd

## [1] 7.680392

Population\_var <- var(delay$delay)  
  
Population\_var

## [1] 58.98842

n <- length(delay$delay[delay$incident == "Mechanical"])  
  
n

## [1] 89565

z <- (Mechanical\_mean - Population\_mean)/(Population\_sd/sqrt(n))  
  
z

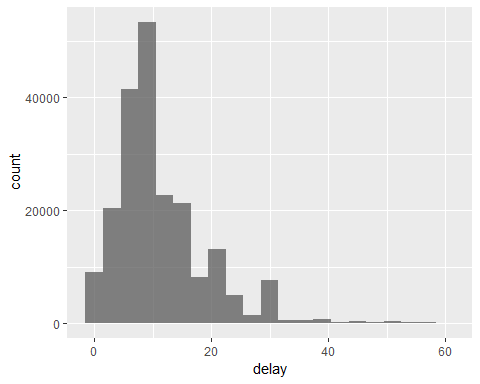
## [1] -4.181594

z1 <- (Mechanical\_mean - Population\_mean)/(Mechanical\_sd/sqrt(n))  
  
p\_value <- pnorm(z)  
  
p\_value

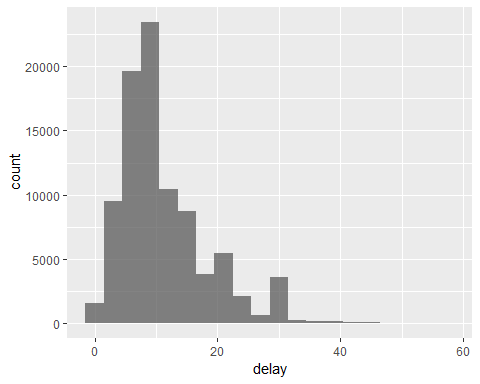
## [1] 1.447362e-05

# side by side histograms of whole delays and delays due to Mechanical reasons

ggplot(delay, aes(x=delay))+geom\_histogram(binwidth = 3,alpha=3/4)



delay\_mechanical <- delay[delay$incident == "Mechanical",]  
  
ggplot(delay\_mechanical, aes(x=delay))+geom\_histogram(binwidth = 3,alpha=3/4)



# study Multi regression cosidering “gap” attribute

# Pearson correlation test between Delay and Gap

delay.gap <- delay[!is.na(delay$gap), ]  
  
cor.test(delay.gap$delay,delay.gap$gap, method = "pearson")

##   
## Pearson's product-moment correlation  
##   
## data: delay.gap$delay and delay.gap$gap  
## t = 395.66, df = 206610, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.6541040 0.6590104  
## sample estimates:  
## cor   
## 0.6565642

index.gap <- sample(length(delay.gap$route), .7 \*length(delay.gap$route))  
  
train.gap <- delay.gap[index.gap,]  
test.gap <- delay.gap[-index.gap,]  
  
delay\_Multi.gap <- lm(delay ~ incident-1+gap, data = train.gap)  
  
summary(delay\_Multi.gap)

##   
## Call:  
## lm(formula = delay ~ incident - 1 + gap, data = train.gap)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -295.616 -2.507 -0.635 1.705 50.713   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## incidentDiversion 9.9636278 0.0925765 107.63 <2e-16 \*\*\*  
## incidentEmergency Services 5.8761211 0.1068527 54.99 <2e-16 \*\*\*  
## incidentGeneral Delay 9.2972463 0.0495153 187.76 <2e-16 \*\*\*  
## incidentInvestigation 5.5721506 0.0516861 107.81 <2e-16 \*\*\*  
## incidentLate Leaving Garage 5.0212847 0.0392677 127.87 <2e-16 \*\*\*  
## incidentMechanical 5.3145439 0.0283142 187.70 <2e-16 \*\*\*  
## incidentUtilized Off Route 3.9051182 0.0375619 103.97 <2e-16 \*\*\*  
## gap 0.2660248 0.0008119 327.64 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.5 on 144621 degrees of freedom  
## Multiple R-squared: 0.8373, Adjusted R-squared: 0.8373   
## F-statistic: 9.303e+04 on 8 and 144621 DF, p-value: < 2.2e-16

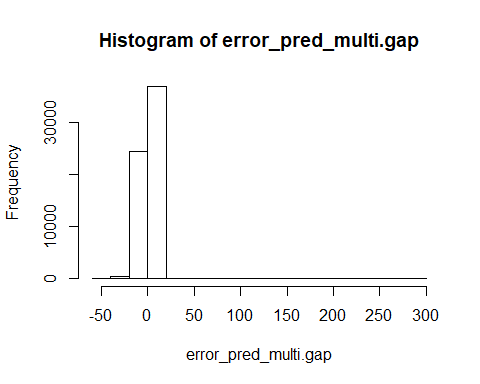
delay\_pred.gap <- predict(delay\_Multi.gap, newdata = test.gap, interval = "prediction")  
  
pred\_Multi.gap <- data.frame(delay\_pred.gap)  
  
error\_pred\_multi.gap <- pred\_Multi.gap$fit - test.gap$delay  
  
error\_multi.gap <- data.frame(error\_pred\_multi.gap)  
head(error\_multi.gap)

## error\_pred\_multi.gap  
## 1 -6.1451201  
## 2 -4.0444660  
## 3 -7.9258967  
## 4 0.3417797  
## 5 4.5805687  
## 6 -13.7784412

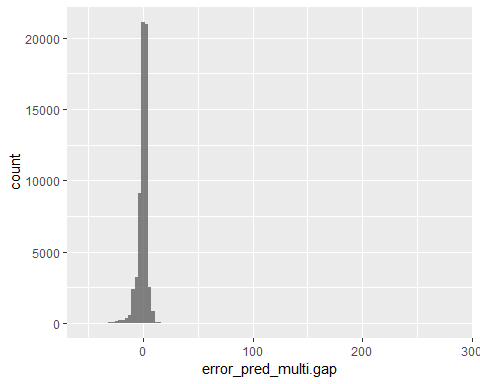
sum(is.na(train.gap$gap))

## [1] 0

hist(error\_pred\_multi.gap)



ggplot(error\_multi.gap,aes(x=error\_pred\_multi.gap))+geom\_histogram(binwidth = 3,alpha=3/4)



summary(error\_pred\_multi.gap)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -52.1239 -1.7047 0.6350 -0.0179 2.5068 283.3986

mse\_multi.gap <- sqrt(sum((error\_pred\_multi.gap)^2)/nrow(test.gap))  
  
mse\_multi.gap

## [1] 5.768736

# study Multi regression cosidering “direction” attribute

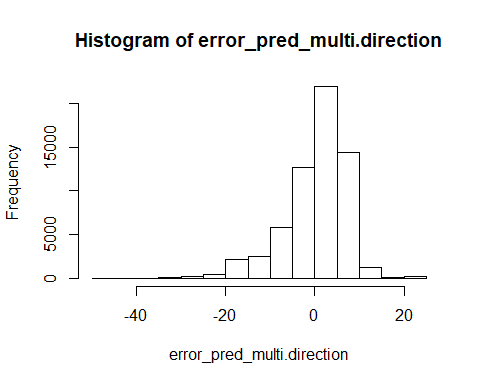
delay\_Multi.direction <- lm(delay ~ incident-1+direction, data = train)  
  
summary(delay\_Multi.direction)

##   
## Call:  
## lm(formula = delay ~ incident - 1 + direction, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.945 -5.047 -1.233 3.136 47.018   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## incidentDiversion 21.68945 0.15277 141.971 <2e-16 \*\*\*  
## incidentEmergency Services 12.51527 0.18533 67.529 <2e-16 \*\*\*  
## incidentGeneral Delay 15.13675 0.06448 234.733 <2e-16 \*\*\*  
## incidentInvestigation 11.79146 0.13792 85.497 <2e-16 \*\*\*  
## incidentLate Leaving Garage 10.11262 0.13139 76.968 <2e-16 \*\*\*  
## incidentMechanical 11.16065 0.12595 88.610 <2e-16 \*\*\*  
## incidentUtilized Off Route 8.24789 0.12983 63.528 <2e-16 \*\*\*  
## directionEB 0.07284 0.12852 0.567 0.5709   
## directionNB 0.25517 0.12754 2.001 0.0454 \*   
## directionNE -1.27636 1.70975 -0.747 0.4554   
## directionSB -0.17837 0.12777 -1.396 0.1627   
## directionWB -0.06569 0.12816 -0.513 0.6083   
## directionWNB 4.83935 7.23944 0.668 0.5038   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.238 on 144628 degrees of freedom  
## Multiple R-squared: 0.7186, Adjusted R-squared: 0.7185   
## F-statistic: 2.84e+04 on 13 and 144628 DF, p-value: < 2.2e-16

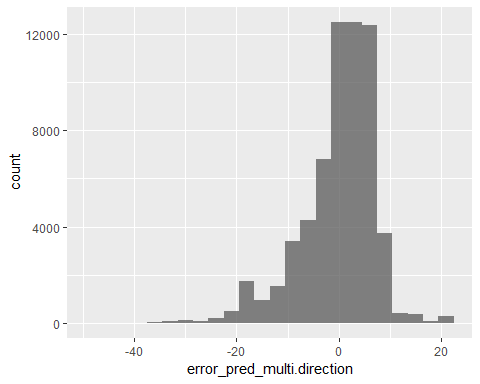
delay\_pred.direction <- predict(delay\_Multi.direction, newdata = test, interval = "prediction")  
  
pred\_Multi.direction <- data.frame(delay\_pred.direction)  
  
error\_pred\_multi.direction <- pred\_Multi.direction$fit - test$delay  
  
error\_multi.direction <- data.frame(error\_pred\_multi.direction)  
head(error\_multi.direction)

## error\_pred\_multi.direction  
## 1 -4.90503528  
## 2 3.98227980  
## 3 -12.95336987  
## 4 -0.06575049  
## 5 -0.93047800  
## 6 -17.76650918

hist(error\_pred\_multi.direction)



ggplot(error\_multi.direction,aes(x=error\_pred\_multi.direction))+geom\_histogram(binwidth = 3,alpha=3/4)



mse\_multi.direction <- sqrt(sum((error\_multi.direction)^2)/nrow(test))  
  
mse\_multi.direction

## [1] 7.317375

# Consider “gap” & “direction” features

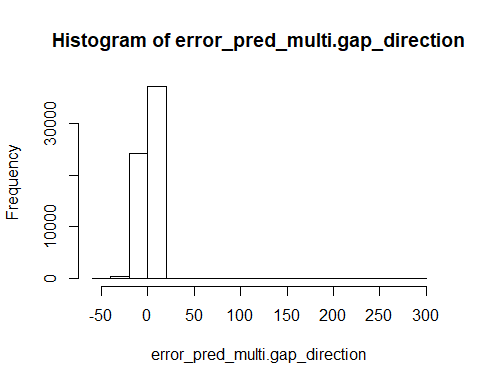
delay\_Multi.gap\_direction <- lm(delay ~ incident-1+gap+direction, data = train.gap)  
  
summary(delay\_Multi.gap\_direction)

##   
## Call:  
## lm(formula = delay ~ incident - 1 + gap + direction, data = train.gap)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -295.655 -2.594 -0.755 1.605 50.791   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## incidentDiversion 10.042242 0.120773 83.150 <2e-16 \*\*\*  
## incidentEmergency Services 5.964337 0.141404 42.180 <2e-16 \*\*\*  
## incidentGeneral Delay 9.313124 0.051989 179.136 <2e-16 \*\*\*  
## incidentInvestigation 5.663648 0.106337 53.261 <2e-16 \*\*\*  
## incidentLate Leaving Garage 5.119229 0.100846 50.763 <2e-16 \*\*\*  
## incidentMechanical 5.402635 0.097066 55.659 <2e-16 \*\*\*  
## incidentUtilized Off Route 3.976237 0.099338 40.028 <2e-16 \*\*\*  
## gap 0.265960 0.000812 327.542 <2e-16 \*\*\*  
## directionEB -0.034024 0.097476 -0.349 0.7271   
## directionNB 0.035479 0.096814 0.366 0.7140   
## directionNE -1.290611 1.175694 -1.098 0.2723   
## directionSB -0.169138 0.096853 -1.746 0.0808 .   
## directionWB -0.176399 0.097265 -1.814 0.0697 .   
## directionWNB 2.086645 5.500353 0.379 0.7044   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.5 on 144615 degrees of freedom  
## Multiple R-squared: 0.8373, Adjusted R-squared: 0.8373   
## F-statistic: 5.317e+04 on 14 and 144615 DF, p-value: < 2.2e-16

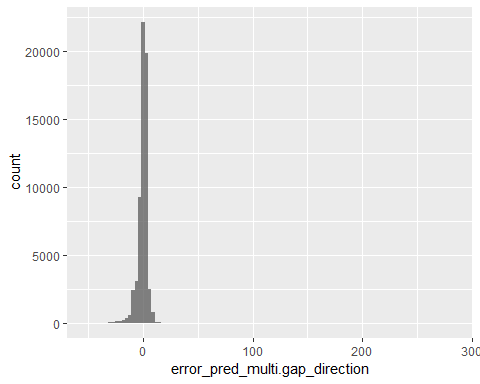
delay\_pred.gap\_direction <- predict(delay\_Multi.gap\_direction, newdata = test.gap, interval = "prediction")  
  
pred\_Multi.gap\_direction <- data.frame(delay\_pred.gap\_direction)  
  
error\_pred\_multi.gap\_direction <- pred\_Multi.gap\_direction$fit - test.gap$delay  
  
error\_multi.gap\_direction <- data.frame(error\_pred\_multi.gap\_direction)  
head(error\_multi.gap\_direction)

## error\_pred\_multi.gap\_direction  
## 1 -6.1335157  
## 2 -4.1353641  
## 3 -7.8064062  
## 4 0.2692915  
## 5 4.4921955  
## 6 -13.7270292

hist(error\_pred\_multi.gap\_direction)



ggplot(error\_multi.gap\_direction,aes(x=error\_pred\_multi.gap\_direction))+geom\_histogram(binwidth = 3,alpha=3/4)



summary(error\_pred\_multi.gap\_direction)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -52.00018 -1.64424 0.73737 -0.01786 2.56013 283.44221

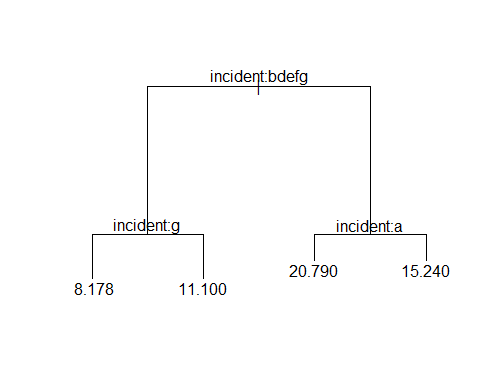
mse\_multi.gap\_direction <- sqrt(sum((error\_pred\_multi.gap\_direction)^2)/nrow(test.gap))  
  
mse\_multi.gap\_direction

## [1] 5.769012

# Create a prediction model for the delay through regression tree

# consider “incident” feature

tree\_model <- tree(delay ~ incident, data = test)  
plot(tree\_model)  
text(tree\_model)



class(delay$direction)

## [1] "factor"

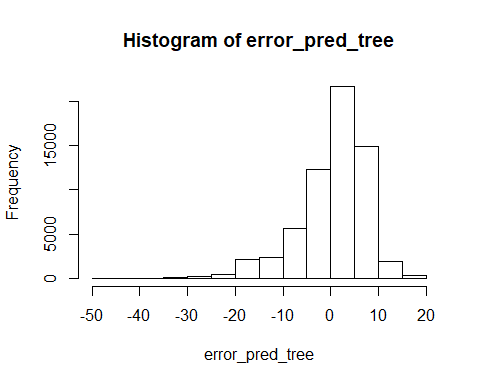
levels(delay$direction)

## [1] "BW" "EB" "NB" "NE" "NULL" "SB" "WB" "WNB"

delay\_pred\_tree <- predict(tree\_model, newdata = test, interval = "prediction")  
  
pred\_tree <- data.frame(delay\_pred\_tree)  
  
error\_pred\_tree <- pred\_tree$delay\_pred\_tree - test$delay  
  
error\_tree <- data.frame(error\_pred\_tree)  
head(error\_tree)

## error\_pred\_tree  
## 1 -4.8996363  
## 2 4.1003637  
## 3 -13.8996363  
## 4 1.1003637  
## 5 -0.8223457  
## 6 -17.8996363

hist(error\_pred\_tree)



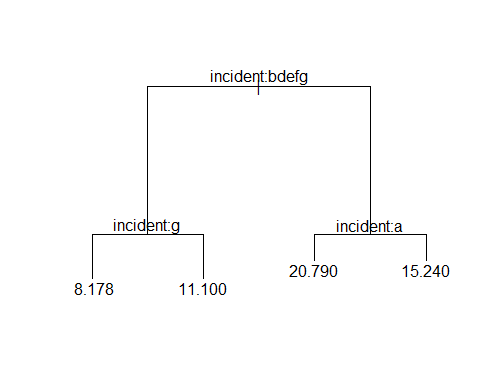
mse\_tree <- sqrt(sum((error\_tree)^2)/nrow(test))  
  
mse\_tree

## [1] 7.329504

# Create a prediction model for the delay through regression tree

# consider “incident” & “direction” features

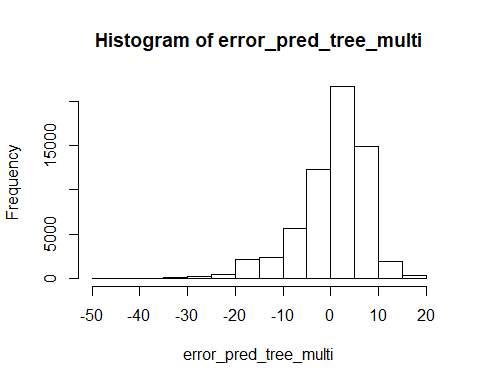
tree\_model\_multi <- tree(delay ~ incident+direction, data = test)  
plot(tree\_model\_multi)  
text(tree\_model\_multi)



delay\_pred\_tree\_multi <- predict(tree\_model\_multi, newdata = test, interval = "prediction")  
  
pred\_tree\_multi <- data.frame(delay\_pred\_tree\_multi)  
  
error\_pred\_tree\_multi <- pred\_tree\_multi$delay\_pred\_tree\_multi - test$delay  
  
error\_tree\_multi <- data.frame(error\_pred\_tree\_multi)  
head(error\_tree\_multi)

## error\_pred\_tree\_multi  
## 1 -4.8996363  
## 2 4.1003637  
## 3 -13.8996363  
## 4 1.1003637  
## 5 -0.8223457  
## 6 -17.8996363

hist(error\_pred\_tree\_multi)



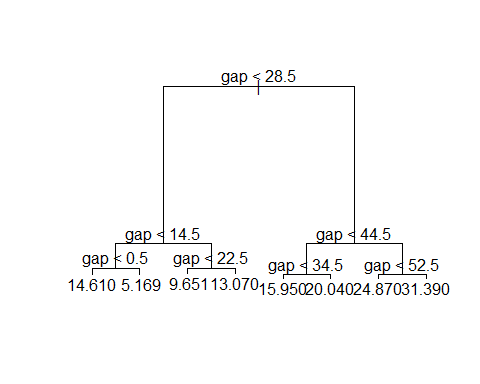
mse\_tree\_multi <- sqrt(sum((error\_tree\_multi)^2)/nrow(test))  
  
mse\_tree\_multi

## [1] 7.329504

# Create a prediction model for the delay through regression tree

# consider “incident”, “gap” features

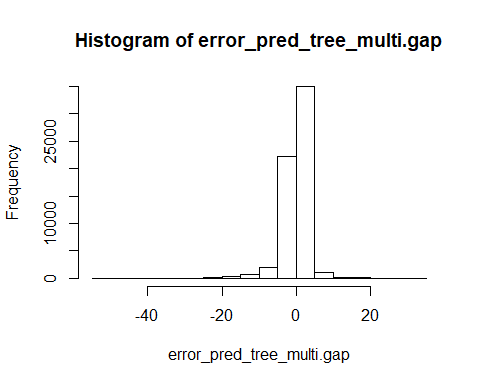
tree\_model\_multi.gap <- tree(delay ~ incident+gap, data = test.gap)  
plot(tree\_model\_multi.gap)  
text(tree\_model\_multi.gap)



delay\_pred\_tree\_multi.gap <- predict(tree\_model\_multi.gap, newdata = test.gap, interval = "prediction")  
  
pred\_tree\_multi.gap <- data.frame(delay\_pred\_tree\_multi.gap)  
  
error\_pred\_tree\_multi.gap <- pred\_tree\_multi.gap$delay\_pred\_tree\_multi.gap - test.gap$delay  
  
error\_tree\_multi.gap <- data.frame(error\_pred\_tree\_multi.gap)  
head(error\_tree\_multi.gap)

## error\_pred\_tree\_multi.gap  
## 1 -1.61048416  
## 2 0.04326255  
## 3 1.38951584  
## 4 -0.34941600  
## 5 4.16898017  
## 6 -9.95673745

hist(error\_pred\_tree\_multi.gap)



mse\_tree\_multi.gap <- sqrt(sum((error\_tree\_multi.gap)^2)/nrow(test.gap))  
  
mse\_tree\_multi.gap

## [1] 3.766878

# Cross Validation 10 folds

# using incidents and directions

# Define training control  
set.seed(123)   
train.control <- trainControl(method = "cv", number = 10)  
# Train the model  
model <- train(delay ~ incident+direction, data = delay, method = "lm", trControl = train.control)

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading  
  
## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient  
## fit may be misleading

# Summarize the results  
print(model)

## Linear Regression   
##   
## 206631 samples  
## 2 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 185968, 185969, 185968, 185968, 185967, 185967, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 7.262021 0.1060437 5.370674  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

# Cross Validation 10 folds

# using incidents

# Define training control  
set.seed(123)   
train.control <- trainControl(method = "repeatedcv", number = 10,repeats = 3)  
# Train the model  
model <- train(delay ~ incident, data = delay, method = "lm", trControl = train.control)  
# Summarize the results  
print(model)

## Linear Regression   
##   
## 206631 samples  
## 1 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 3 times)   
## Summary of sample sizes: 185968, 185969, 185968, 185968, 185967, 185967, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 7.263374 0.1057237 5.372744  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

# Cross Validation 10 folds

# using incidents and gap

# Define training control  
set.seed(123)   
train.control <- trainControl(method = "repeatedcv", number = 10,repeats = 3)  
# Train the model  
model <- train(delay ~ incident+gap, data = delay.gap, method = "lm", trControl = train.control)  
# Summarize the results  
print(model)

## Linear Regression   
##   
## 206613 samples  
## 2 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 3 times)   
## Summary of sample sizes: 185954, 185952, 185952, 185952, 185952, 185951, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 5.582148 0.4811505 3.201688  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE