########################################################################

install.packages("xlsx")

install.packages("openxlsx")

library(xlsx)

library(openxlsx)

########################################################################

install.packages(aod)

install.packages(aod3)

install.packages("Hmisc")

install.packages("corrplot")

install.packages("corrgram")

library(aod)

library(corrplot)

library("Hmisc")

library("ggpubr")

library(ggplot2)

library(data.table)

library(corrplot)

par(mar=c(1,1,1,1))

dff <- read.csv("F:/project/delays/On\_Time\_Reporting\_Carrier\_On\_Time\_Performance\_manipulated.csv")[ ,c(16,25,4,34,45,52,55,31,32,42,43,54,57,57,60)]

dff = as.data.frame(dff)

names(dff)

dff$ArrivalDelayGroups[is.na(dff$ArrivalDelayGroups)] <- 13

dff$DepartureDelayGroups[is.na(dff$DepartureDelayGroups)] <- 13

dff<-dff[(dff$ArrivalDelayGroups < 13 ),]

groupadd=min(dff$ArrivalDelayGroups)

dff[,5]=dff[,5]-groupadd

dff[,4]=dff[,4]-groupadd

dff[is.na(dff)] <- 0

head(dff)

summary(dff)

str(dff)

unique(dff$DepartureDelayGroups)

unique(dff$DistanceGroup)

unique((dff$ArrivalDelayGroups))

unique(dff$DayOfWeek)

unique(dff$OriginState)

unique(dff$DestState)

x <- dff[3:15]

y <- dff[3:15]

dff.cor=cor(x, y)

palette = colorRampPalette(c("green", "white", "red","brown","blue","darkred",

"yellow","gray","black","pink","orange","cyan")) (12)

heatmap(x = dff.cor, col = palette, symm = TRUE)

corrplot(dff.cor)

pl1=ggplot(dff, aes(x = dff$ArrDelayMinutes)) +

geom\_histogram(colour="darkblue", size=1, fill="red")

pl2=ggplot(dff, aes(x = dff$DepDelayMinutes)) +

geom\_histogram(colour="darkblue", size=1, fill="red")

pl3=ggplot(dff, aes(x = dff$DistanceGroup)) +

geom\_histogram(colour="darkblue", size=1, fill="blue")

pl4=ggplot(dff, aes(x = dff$Distance)) +

geom\_histogram(colour="darkblue", size=1, fill="green")

pl5=ggplot(dff, aes(x = dff$ArrivalDelayGroups)) +

geom\_histogram(colour="darkblue", size=1, fill="blue") + scale\_x\_log10()

pl6=ggplot(dff, aes(x = dff$DepartureDelayGroups)) +

geom\_histogram(colour="darkblue", size=1, fill="blue") + scale\_x\_log10()

ggarrange(pl1, pl2, pl3,pl4,pl5,pl6 + rremove("x.text"),

labels = c("A", "B", "C","D", "E", "F","G","H"),

ncol = 2, nrow = 3)

bp1=ggplot(dff, aes(x=factor(dff$ArrivalDelayGroups),

y=dff$ArrDelayMinutes,color=dff$ArrivalDelayGroups)) +

geom\_boxplot(outlier.colour="red", outlier.shape=2,outlier.size=1)

bp2=ggplot(dff, aes(x=factor(dff$DepartureDelayGroups),

y=dff$ArrDelayMinutes,color=dff$DepartureDelayGroups)) +

geom\_boxplot(outlier.shape=NA)

bp3=ggplot(dff, aes(x=factor(dff$DayOfWeek),

y=dff$ArrDelayMinutes,color=dff$DayOfWeek)) +

geom\_boxplot(outlier.shape=NA)

bp4=ggplot(dff, aes(x=factor(dff$DistanceGroup),

y=dff$ArrDelayMinutes,color=dff$DistanceGroup)) +

geom\_boxplot(outlier.shape=NA)

bp5=ggplot(dff, aes(x=dff$OriginState,

y=dff$ArrDelayMinutes,color=dff$OriginState)) +

geom\_boxplot(outlier.shape=NA) + ylim(0, 50)

bp6=ggplot(dff, aes(x=factor(dff$DestState),

y=dff$ArrDelayMinutes,color=dff$DestState)) +

geom\_boxplot(outlier.shape=NA) + ylim(0, 200)

# Remove outliers

#dff$Distance[which(dff$Distance > 1400)]=1400

#dff$ArrDelayMinutes[which(dff$ArrDelayMinutes > 180)]=180

#dff$DepDelayMinutes[which(dff$DepDelayMinutes > 180)]=180

#ggarrange(bp1,bp2,bp3,bp4 + rremove("x.text"),

# labels = c("A", "B","C","D","E","F"),

# ncol = 1, nrow = 6)

ps1=ggscatter(dff, x = "DepDelayMinutes",

y = "ArrDelayMinutes",add = "reg.line") +

stat\_cor(method = "pearson",label.x = 3, label.y = 32) +

stat\_regline\_equation(label.x = 3, label.y = 32)

ps2=ggscatter(dff, x = "ArrivalDelayGroups",

y = "ArrDelayMinutes",add = "reg.line") +

stat\_cor(method = "pearson",label.x = 3, label.y = 32) +

stat\_regline\_equation(label.x = 3, label.y = 32)

ps3=ggscatter(dff, x = "DepartureDelayGroups",

y = "ArrDelayMinutes",add = "reg.line") +

stat\_cor(method = "pearson",label.x = 3, label.y = 32) +

stat\_regline\_equation(label.x = 3, label.y = 32)

ps4=ggscatter(dff, x = "DistanceGroup",

y = "ArrDelayMinutes",add = "reg.line") +

stat\_cor(method = "pearson",label.x = 3, label.y = 32) +

stat\_regline\_equation(label.x = 3, label.y = 32)

#ggarrange(ps1, ps2,ps3 + rremove("x.text"),

#labels = c("A", "B","C","D"),

#ncol = 1, nrow = 3)

dff$ArrivalDelayGroups <- factor(dff$ArrivalDelayGroups)

train\_n=as.integer((2/3)\*length(dff$ArrDelayMinutes))

train\_n

test\_n=length(dff$ArrDelayMinutes)-train\_n

test\_n=train\_n+1

dff\_train=dff[1:train\_n,]

dff\_test=dff[test\_n:length(dff$ArrDelayMinutes),]

length(dff\_train$ArrDelayMinutes)

length(dff\_test$ArrDelayMinutes)

fit1 <- lm(dff\_train$ArrDelayMinutes ~ dff\_train$DepDelayMinutes ,

data=dff\_train)

summary(fit1)

###################################################################

mylogit1 <- glm(ArrDelayMinutes ~ DepDelayMinutes +

ArrivalDelayGroups,data=dff\_train )

summary(mylogit1)

exp(coef(mylogit1))

anova(mylogit1, test = "Chisq")

dff\_test$rankP <-

predict(mylogit1, newdata=dff\_test,type = "response")

dff\_test

histogram(dff\_test$rankP)

newdata2 <- with(dff\_test, data.frame(DepDelayMinutes =

rep(seq(from = 0, to = 1500, length.out = 5),20),

ArrivalDelayGroups = factor(rep(1:14, each = 50))))

newdata3 <- cbind(newdata2, predict(mylogit1, newdata = newdata2,

type = "link",se = TRUE))

newdata3 <- within(newdata3, {

PredictedProb <- plogis(fit)

LL <- plogis(fit - (1.96 \* se.fit))

UL <- plogis(fit + (1.96 \* se.fit))

})

histogram(newdata3$PredictedProb)

newdata3

ggplot(newdata3, aes(x = DepDelayMinutes, y = PredictedProb)) +

geom\_ribbon(aes(ymin = LL,ymax = UL,

fill = ArrivalDelayGroups), alpha = 0.2) +

geom\_line(aes(colour = ArrivalDelayGroups), size = 1)