```
#Admin
setwd("/Users/bennetthellman/Desktop/OneDrive - Massachusetts Institute of
Technology/AE")
df <- read.csv("WranglerElantra2018.csv", stringsAsFactors = FALSE, header
= TRUE)
library(tidyverse)
library(miscTools)
library(Metrics)


#Data Wrangling
df<-df%>%mutate(Month = as.factor(Month))

#a.i
#set.seed(15072)
#row.number <- sample(1:nrow(df), 0.7*nrow(df))
#train = df[row.number,]
#test = df[-row.number,]

train = df%>%filter(Year != 2018)
test = df%>%filter(Year == 2018)


#a.i
lmai_train =
lm(Wrangler.Sales~Year+Unemployment.Rate+Wrangler.Queries+CPI.Energy+CPI.All,
data = train) #Create the linear regression
summary(lmai_train)
predai_train <- predict(lmai_train, newdata = train)
c(RMSE = rmse(train$Wrangler.Sales, predai_train), MAE =
mae(train$Wrangler.Sales, predai_train), R2=rSquared(train$Wrangler.Sales,
resid = train$Wrangler.Sales-predai_train))
predai <- predict(lmai_train, newdata = test)
train.mean <- mean(train$Wrangler.Sales)
SSE <- sum((predai - test$Wrangler.Sales)^2)
SST <- sum((train.mean - test$Wrangler.Sales)^2)
OSR2 <- 1 - SSE/SST
c(RMSE = rmse(test$Wrangler.Sales, predai), MAE = mae(test$Wrangler.Sales,
predai), OSR2 = OSR2)


#b.i
cov(df[,c('Unemployment.Rate', 'Wrangler.Queries', 'CPI.Energy',
'CPI.All')])
summary(lmai_train)
#b.ii
lmbii_train =
lm(Wrangler.Sales~Year+Unemployment.Rate+Wrangler.Queries+CPI.Energy, data
= train) #Create the linear regression
summary(lmbii_train)
predbi_train <- predict(lmbii_train, newdata = train)
```

```
c(RMSE = rmse(train$Wrangler.Sales, predbi_train), MAE =
mae(train$Wrangler.Sales, predbi_train), R2=rSquared(train$Wrangler.Sales,
resid = train$Wrangler.Sales-predbi_train))
predbii <- predict(lmbii_train, newdata = test)
train.mean <- mean(train$Wrangler.Sales)
SSE <- sum((predbii - test$Wrangler.Sales)^2)
SST <- sum((train.mean - test$Wrangler.Sales)^2)
OSR2 <- 1 - SSE/SST
c(RMSE = rmse(test$Wrangler.Sales, predbii), MAE = mae(test$Wrangler.Sales,
predbii), OSR2 = OSR2)
#b.iii


#c.i
lmci_train =
lm(Wrangler.Sales~Year+Unemployment.Rate+Wrangler.Queries+CPI.Energy+CPI.All+Month,
data = train) #Create the linear regression
summary(lmci_train)
predci_train <- predict(lmci_train, newdata = train)
c(RMSE = rmse(train$Wrangler.Sales, predci_train), MAE =
mae(train$Wrangler.Sales, predci_train), R2=rSquared(train$Wrangler.Sales,
resid = train$Wrangler.Sales-predci_train))
predci <- predict(lmci_train, newdata = test)
train.mean <- mean(train$Wrangler.Sales)
SSE <- sum((predci - test$Wrangler.Sales)^2)
SST <- sum((train.mean - test$Wrangler.Sales)^2)
OSR2 <- 1 - SSE/SST
c(RMSE = rmse(test$Wrangler.Sales, predci), MAE = mae(test$Wrangler.Sales,
predci), OSR2 = OSR2)
#c.ii

#d.i
lmdi_train = lm(Elantra.Sales~ Year + Unemployment.Rate + Elantra.Queries +
CPI.Energy + CPI.All + Month, data = train) #Create the linear regression
summary(lmdi_train)
preddi_train <- predict(lmdi_train, newdata = train)
c(RMSE = rmse(train$Elantra.Sales, preddi_train), MAE =
mae(train$Elantra.Sales, preddi_train), R2=rSquared(train$Elantra.Sales,
resid = train$Elantra.Sales-preddi_train))
preddi <- predict(lmdi_train, newdata = test)
train.mean <- mean(train$Elantra.Sales)
SSE <- sum((preddi - test$Elantra.Sales)^2)
SST <- sum((train.mean - test$Elantra.Sales)^2)
OSR2 <- 1 - SSE/SST
c(RMSE = rmse(test$Elantra.Sales, preddi), MAE = mae(test$Elantra.Sales,
preddi), OSR2 = OSR2)

#e.i

#e.ii

library(corrplot)
library(RColorBrewer)
```

```
df_wrang<-
df%>%select(Wrangler.Sales,Year,Unemployment.Rate,Wrangler.Queries,CPI.Energy,CPI.All)
M <-cor(df_wrang)
corrplot(M, type="upper", order="hclust",
         col=brewer.pal(n=8, name="RdYlBu"))

df_elantra<-df%>%select(Elantra.Sales,
Year,Unemployment.Rate,Elantra.Queries,CPI.Energy)
M <-cor(df_elantra)
corrplot(M, type="upper", order="hclust",
         col=brewer.pal(n=8, name="RdYlBu"))


#http://www.sthda.com/english/wiki/correlation-analyses-in-r
```