R Code:

## Data input of csv file

a <- read.csv("C:\\Users\\Guest\\Documents\\R training\\Casper\\data.csv")

a$dateordered <- as.Date(a$dateordered,format = "%m/%d/%Y")

a$datereturned <- as.Date(a$datereturned,format = "%m/%d/%Y")

## Subset to get Month returned orders

b <- subset(a,a$orderstatus == "returned")

library(plyr)

mnthreturn <- plyr::count(b,vars = "monthordered")

mnthreturn <- data.frame(mnthreturn)

mnthreturn <- mnthreturn[c(1,5,4,2,3),]

mnthreturn$monthordered<- ordered(mnthreturn$monthordered,levels=c("August","September","October","November","December"))

## Plot of returns

library(ggplot2)

gg <- ggplot(mnthreturn,aes(mnthreturn$monthordered,mnthreturn$freq,group=1))

gg + geom\_point() + geom\_smooth() +

xlab("Month") + theme\_bw() +

geom\_text(aes(label=mnthreturn$freq),hjust=1, vjust=-0.5) +

ylab("Frequency") + ggtitle("Monthly Frequency of Returned Orders")

## Subset for complete orders to compare this

d <- subset(a,a$orderstatus == "complete")

dd <- aggregate(x = d$orders, by = list(d$monthordered),FUN = sum)

dd <- data.frame(dd)

colnames(dd) <- paste(c("month","sum"))

dd$month <- ordered(dd$month,levels = c("August","September","October","November","December","January"))

## Plot of sum of sales

dg <- ggplot(dd,aes(dd$month,dd$sum,group=1))

library(ggthemes)

dg + geom\_point() + geom\_smooth(color = "lightgreen") +

geom\_text(aes(label=dd$sum),hjust=1, vjust=-0.5) +

theme\_calc()+ scale\_colour\_calc() + xlab("Month") + ylab("Sum of sales") + ggtitle("Sum of Sales per Month")

## Correlation

cor(e$numberreturn,e$sum)

## Table comparison

z <- plyr::count(b,vars = "monthordered")

z <- data.frame(z)

z$monthordered <- ordered(z$monthordered,levels=c("August","September","October","November","December"))

colnames(z)[1] <- paste("month\_orig\_return")

e <- cbind(z[c(1,5,4,3,2),],dd[c(1,5,4,3,2),2])

colnames(e) <- paste(c("month","numberreturn","sum"))

e <- cbind(e,e$numberreturn/e$sum)

colnames(e)[4] <- paste("percent")

## Elbow method chart

u <- e[,2:ncol(e)]

set.seed(123)

k.max <- 4

wss <- sapply(1:k.max,function(k){kmeans(u,k,nstart=20)$tot.withinss})

plot(1:k.max,wss,type="b",pch=19,frame=FALSE,xlab="Number of clusters K",ylab="Total within-clusters sum of squares",main="Elbow Method")

## K means using 2 clusters

km <- kmeans(e[,2:ncol(e)],2)

km$cluster

## Hierarchical Clustering

t <- e[,2:3]

rownames(t) <- e$month

dis <- dist(t)

hc <- hclust(dis)

op = par(bg = "#DDE3CA")

plot(hc, col = "#487AA1", col.main = "#45ADA8", col.lab = "#7C8071",

col.axis = "#F38630",xlab = "Month", lwd = 3, lty = 3, sub = "", hang = -1, axes = FALSE)

axis(side = 2, at = seq(0, 400, 100), col = "#F38630", labels = FALSE,

lwd = 2)

mtext(seq(0, 400, 100), side = 2, at = seq(0, 400, 100), line = 1,

col = "#A38630", las = 2)

## SQL Statement

library(sqldf)

a <- read.csv("C:\\Users\\Guest\\Documents\\R training\\Casper\\data.csv")

sqla <- sqldf(

'SELECT MONTHORDERED AS MONTH\_ORDER,SUM(ORDERS) AS SUM\_ORDERS

FROM a

WHERE ORDERSTATUS = "complete"

GROUP BY MONTHORDERED

ORDER BY MONTHORDERED ASC'

)