**Predicting TTC Bus Delays**

**Initial Code and Results Summary**

**Mary-Kate Fraser**

CKME136 – Data Analytics Capstone Course - Winter 2019

**Data Cleaning**

Bus\_Measure <- read\_csv("~/Desktop/Big Data/CKME 136/Bus\_Measure.csv")

Bus3 <- Bus\_Measure

Bus3$Date <- paste('20',Bus3$Date,sep = '')

Bus3$Date\_Time <- paste(Bus3$Date,Bus3$Time,sep = ' ')

Bus3$Date\_Time <- strptime(Bus3$Date\_Time,"%Y-%m-%d %H:%M:%S", tz = "America/New\_York")

Bus3$Date <- as.Date(Bus3$Date)

Bus3$mday <- as.factor(format(Bus3$Date,'%d'))

table(Bus3$mday, useNA = 'ifany')

Bus3$Month <- as.factor(format(Bus3$Date,'%m'))

table(Bus3$Month, useNA = 'ifany')

Bus3$Year <- as.factor(format(Bus3$Date,'%Y'))

table(Bus3$Year, useNA = 'ifany')

Bus3$Hour <- as.factor(format(Bus3$Date\_Time,'%H'))

table(Bus3$Hour,useNA = 'ifany')

Bus3$Minute <- as.factor(format(Bus3$Date\_Time,'%M'))

table(Bus3$Minute,useNA = 'ifany')

Bus3$Seconds <- as.factor(format(Bus3$Date\_Time,'%S'))

table(Bus3$Seconds,useNA = 'ifany')

Bus3$Route <- as.factor(toupper(Bus3$Route))

table(Bus3$Route,useNA = 'ifany')

Bus3$Location <- as.factor(toupper(Bus3$Location))

Bus3$Measure <- as.factor(Bus2$Measure=='High')

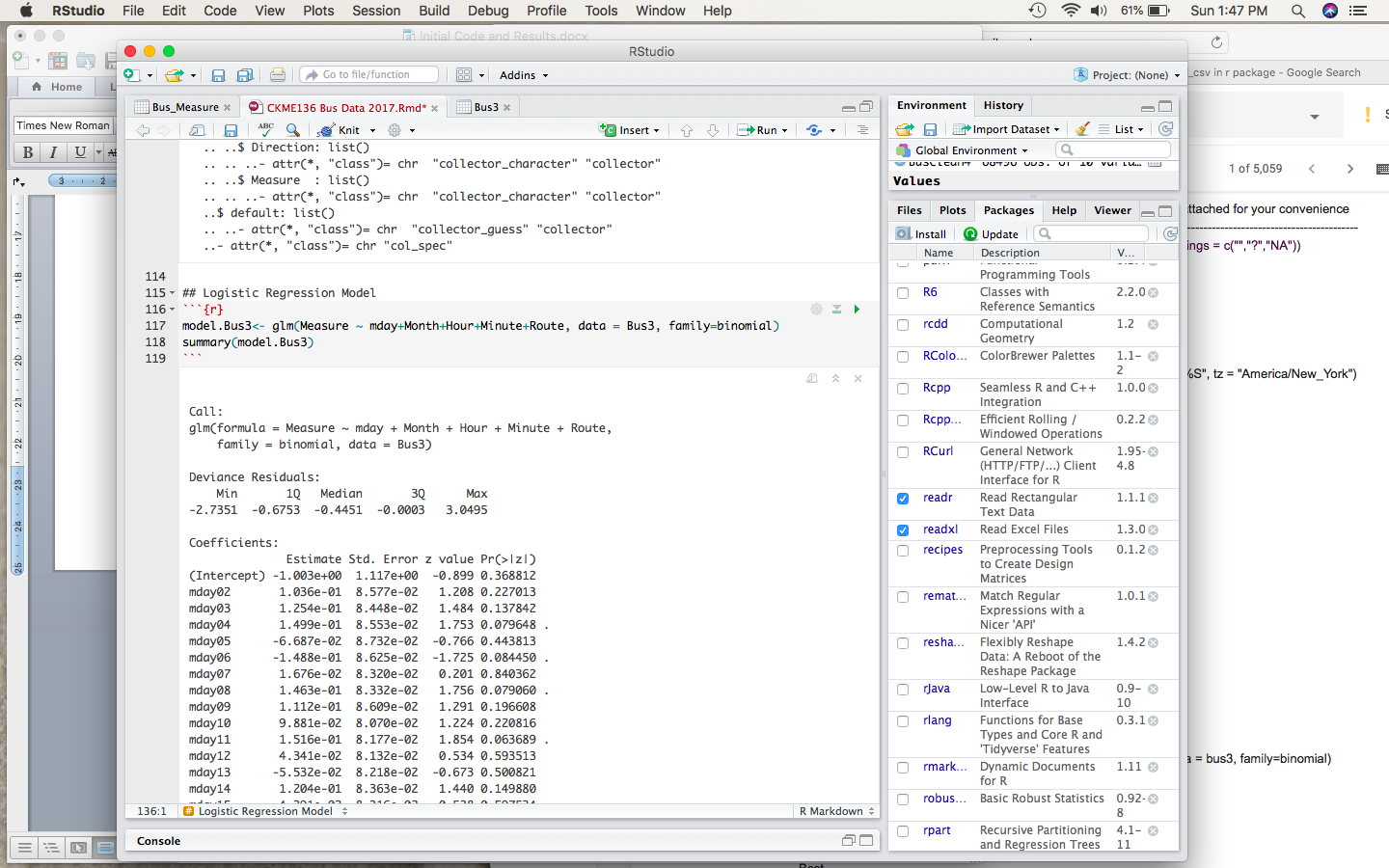
table(Bus3$Measure, useNA = 'ifany')

str(Bus3)

View(Bus3)

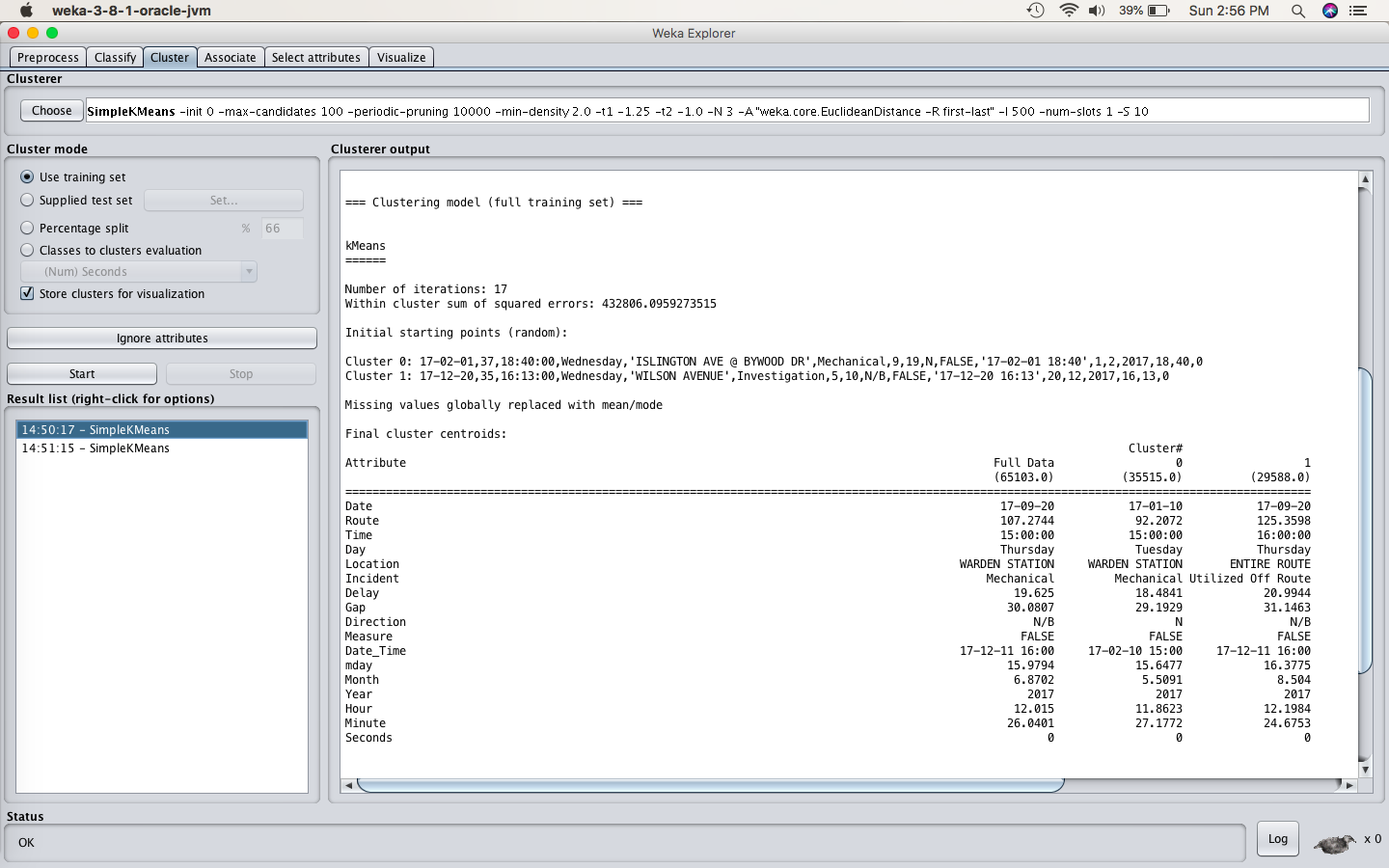
**Logistic Regression**

model.Bus3<- glm(Measure ~ mday+Month+Hour+Minute+Route, data = Bus3, family=binomial)

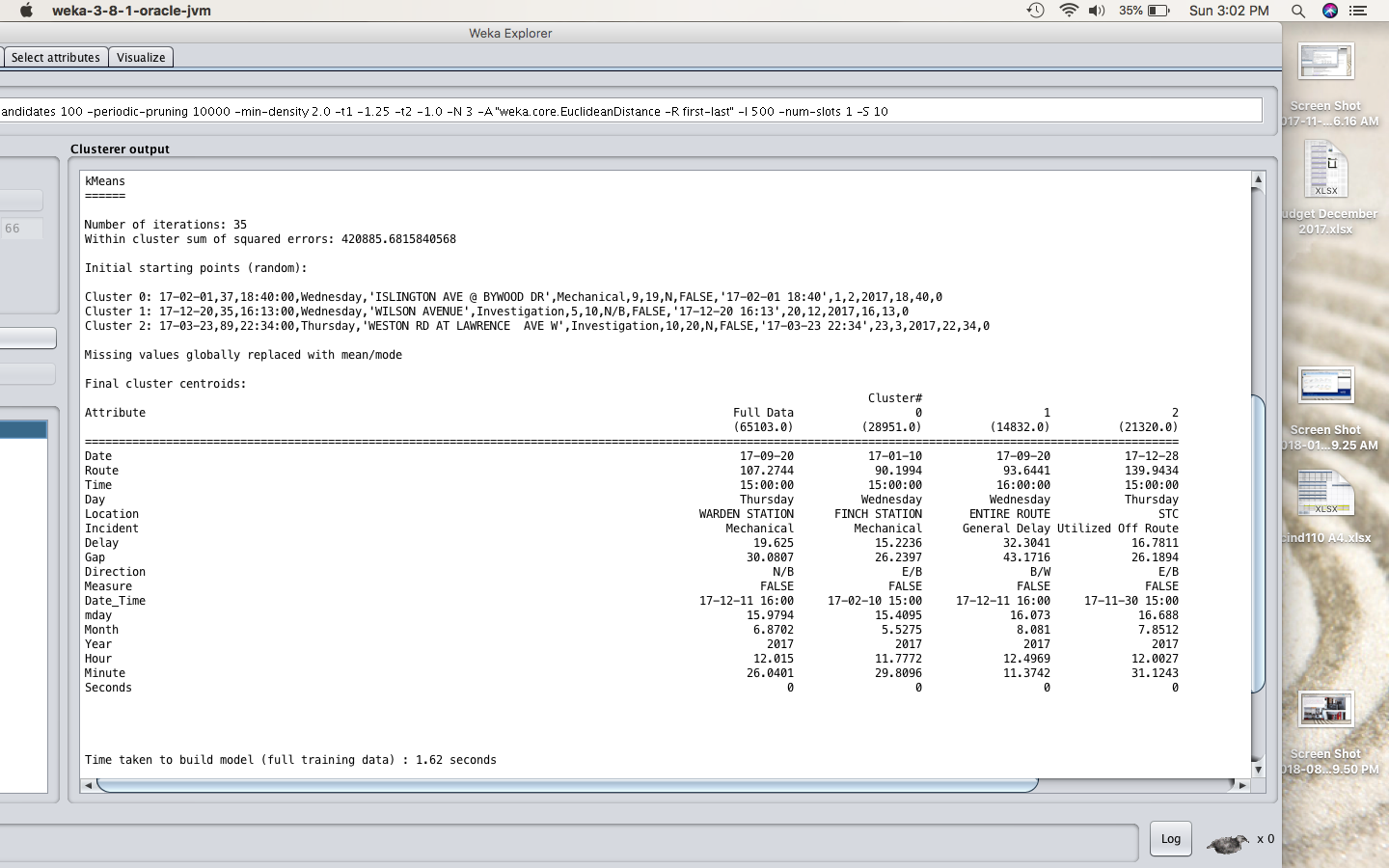


**Clustering**

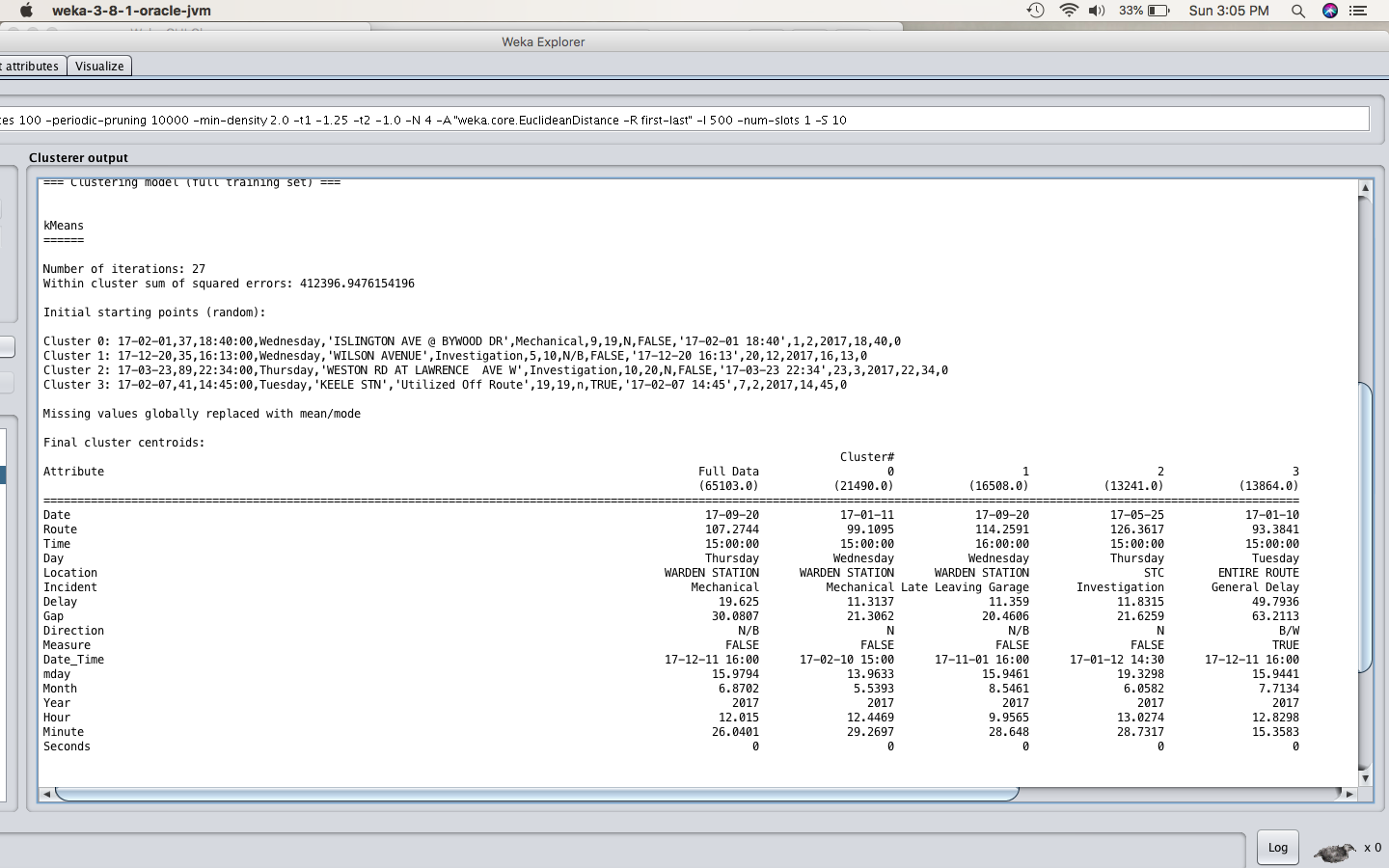
**2 Clusters:**



**3 Clusters:**



**4 Clusters**



**Cluster Summary**

|  |  |
| --- | --- |
| **# of Clusters** | **Clustered Instances** |
| 2 Clusters | 0 35515 ( 55%)  1 29588 ( 45%) |
| 3 Clusters | 0 28951 ( 44%)  1 14832 ( 23%)  2 21320 ( 33%) |
| 4 Clusters | 0 21490 ( 33%)  1 16508 ( 25%)  2 13241 ( 20%)  3 13864 ( 21%) |

**Decision Trees**

## Set up the test set for the classification tree.

set.seed(101)

alpha <- 0.7

inTrain <- sample(1:nrow(Bus3), alpha\*nrow(Bus3))

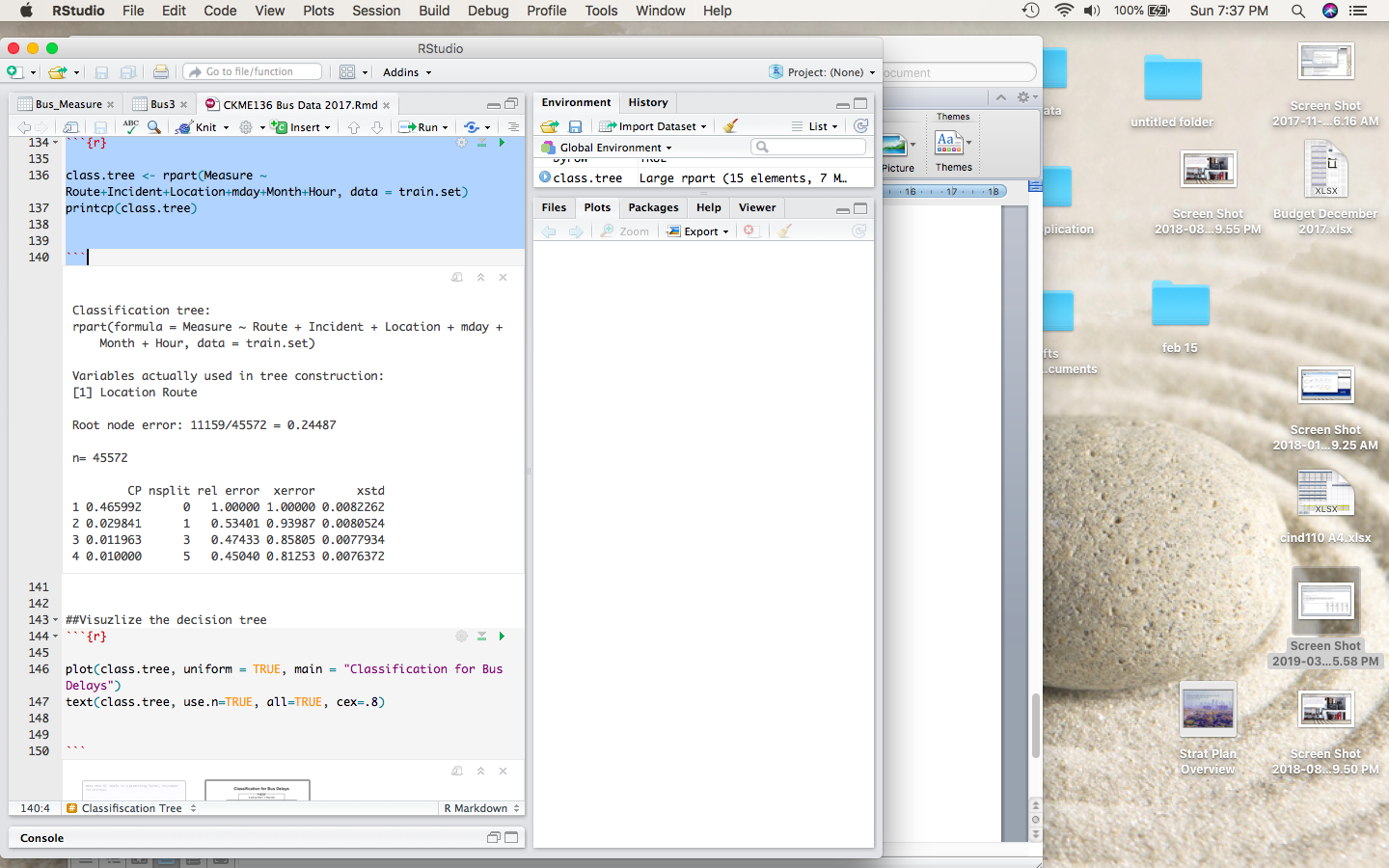
train.set <- Bus3[inTrain,]

test.set <- Bus3[-inTrain,]

## Classifiscation Tree

class.tree <- rpart(Measure ~ Route+Incident+Location+mday+Month+Hour, data = train.set)

printcp(class.tree)



##Visuzlize the decision tree

plot(class.tree, uniform = TRUE, main = "Classification for Bus Delays")

text(class.tree, use.n=TRUE, all=TRUE, cex=.8)

