CSX415_Project_flight_delay_cancellation_analysis

Ajey Patil

5/7/2018

R Markdown

Data Cleaning done in load.project() data munging

```
#currdir <- getwd()
#install.packages('ProjectTemplate')
library(ProjectTemplate)
load.project()
```

{r setup, include=FALSE, echo=FALSE} #knitr::opts_chunk$set(echo=TRUE)

Install Packages

```
#Example package for RMSE calculation in Regression Analysis
#devtools::install_github("ajeypatil/rmse")
#library(rmse)
```

Perform Exploratory Data Analysis

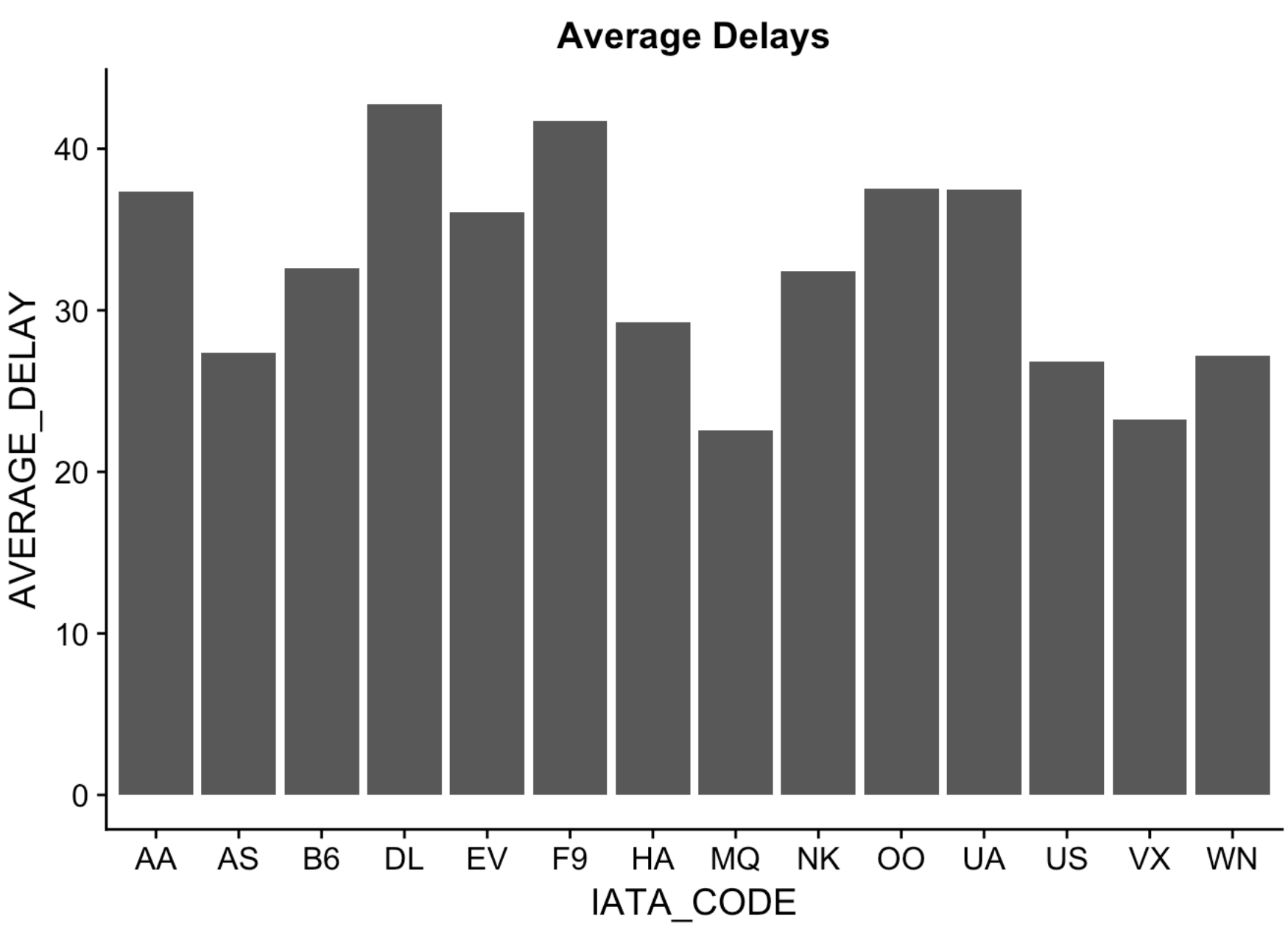
graphs stored in graph directory

```
source('src/01-EDA/CSX415_Project_ExploratoryDataAnalysis.R')
```

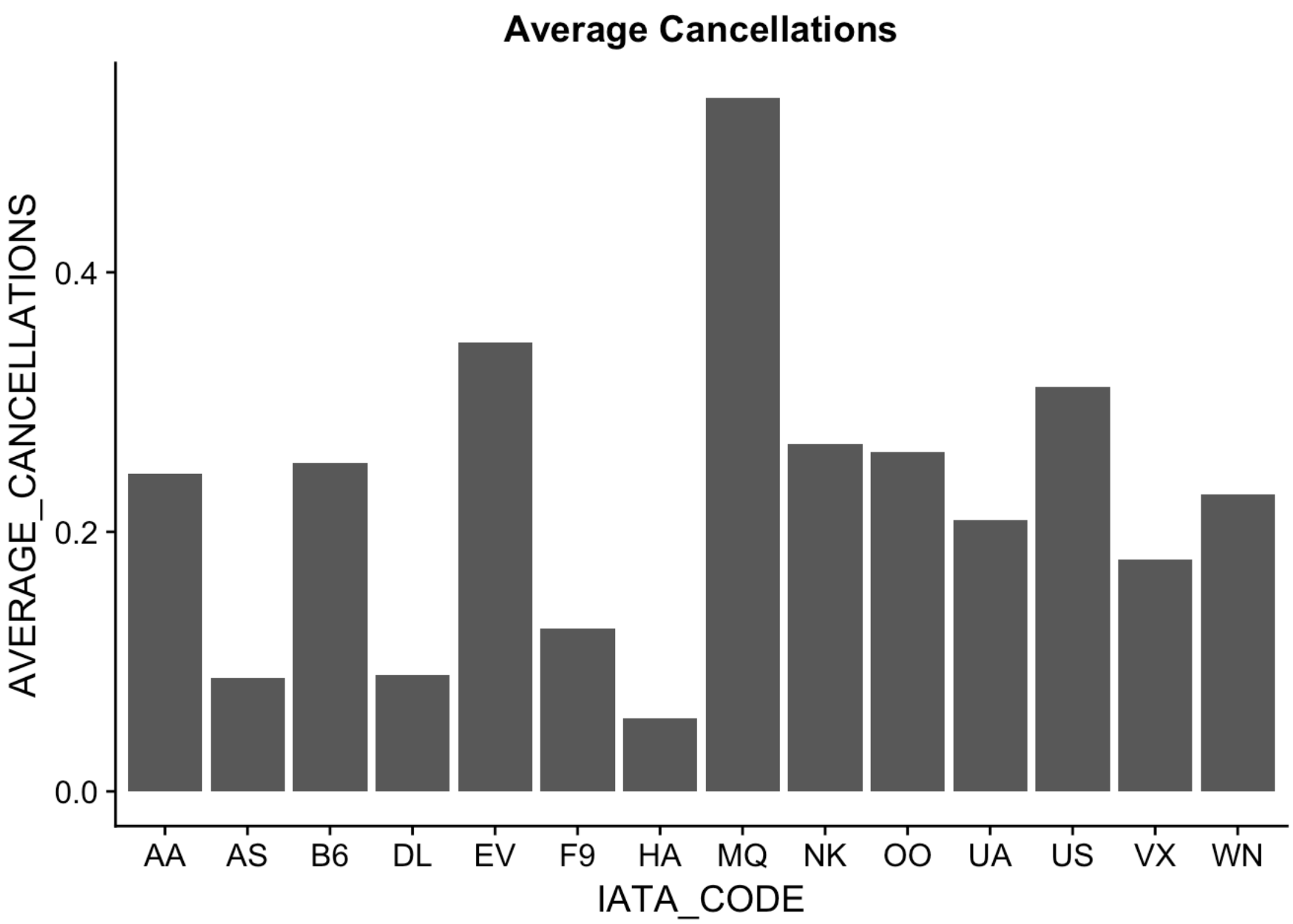
```
## Saving 7 x 5 in image
## Saving 7 x 5 in image
## Saving 7 x 5 in image
## Saving 7 x 5 in image
## Saving 7 x 5 in image
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

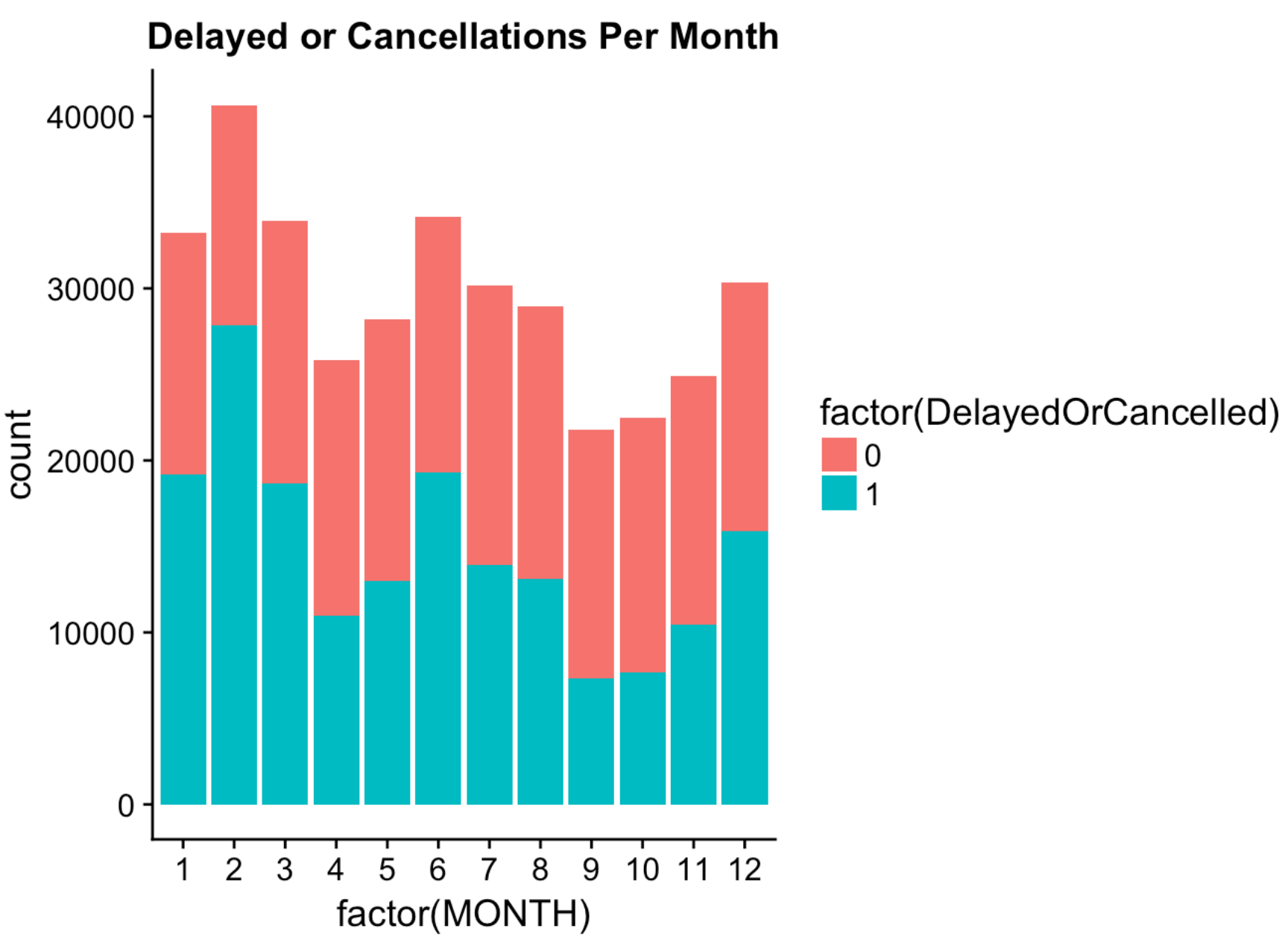
```
ggplot(alldelays, aes(x=IATA_CODE, y=AVERAGE_DELAY)) + geom_bar(stat='identity') + ggtitle("Average Delays")
```



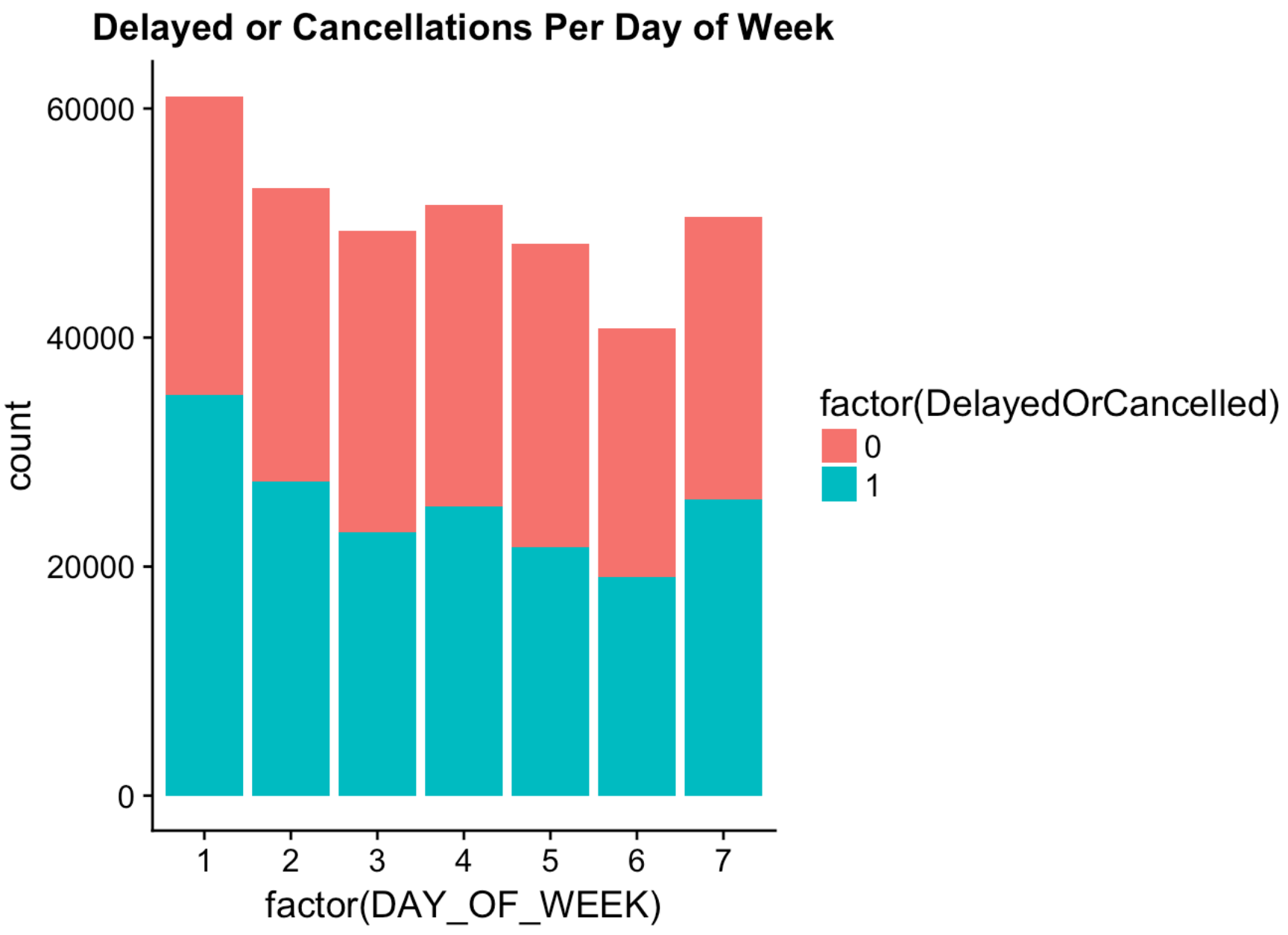

```
ggplot(allcancellations, aes(x=IATA_CODE, y=AVERAGE_CANCELLATIONS)) +  geom_bar(stat='identity') +  ggtitle("A
verage Cancellations")
```




```
ggplot(DelaysAndCancellations,aes(factor(MONTH), group=DelayedOrCancelled, fill=factor(DelayedOrCancelled))) +
geom_bar() +  ggtitle("Delayed or Cancellations Per Month")
```

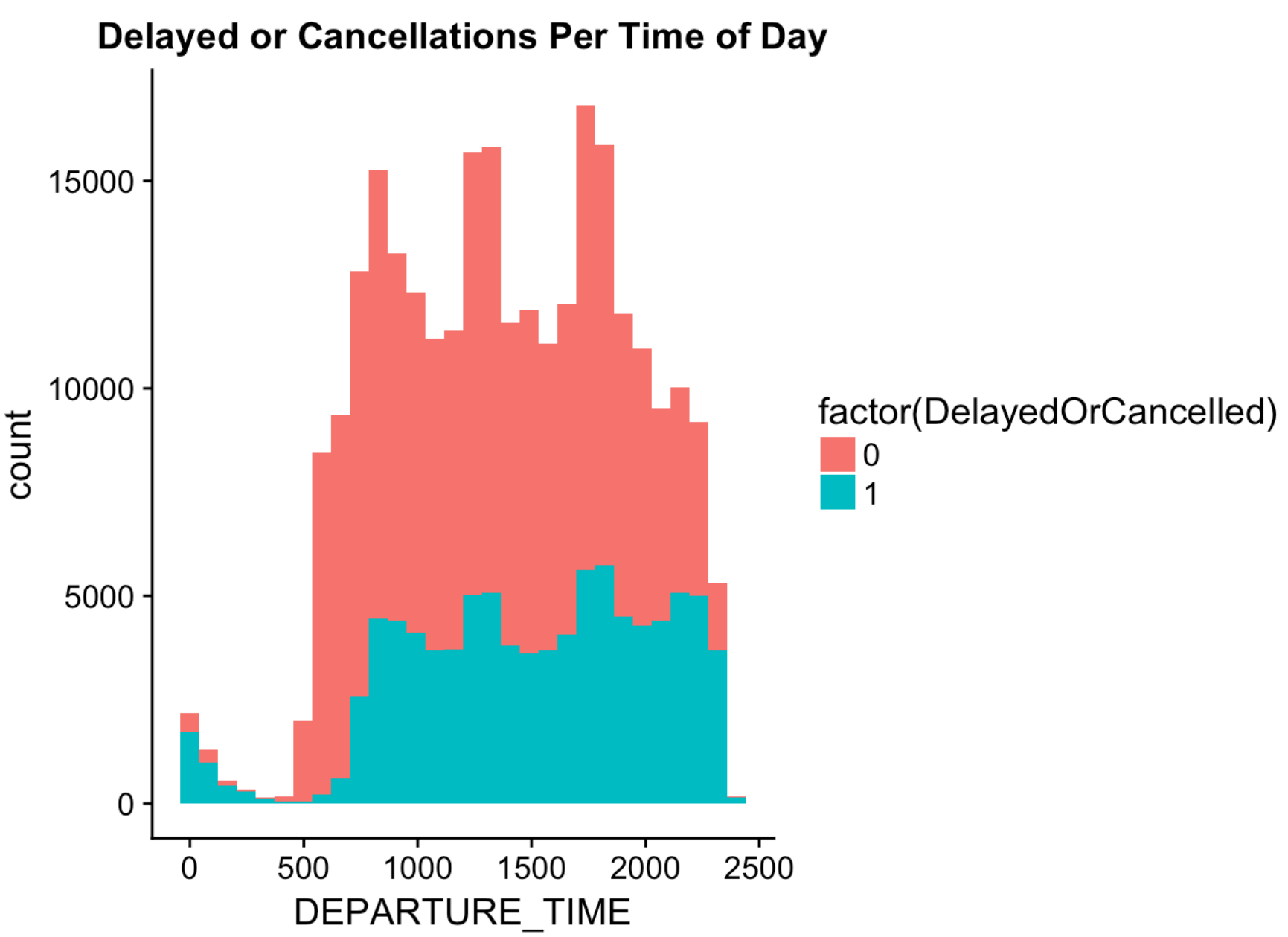



```
ggplot(DelaysAndCancellations,aes(factor(DAY_OF_WEEK), group=DelayedOrCancelled, fill=factor(DelayedOrCancelle
d))) + geom_bar() + ggtitle("Delayed or Cancellations Per Day of Week")
```




```
ggplot(DlyCnclDb,aes(DEPARTURE_TIME, group=DelayedOrCancelled, fill=factor(DelayedOrCancelled))) +  geom_histo
gram() + ggtitle("Delayed or Cancellations Per Time of Day")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



PreProcess Data

Remove zero variance columns

Test-Train split

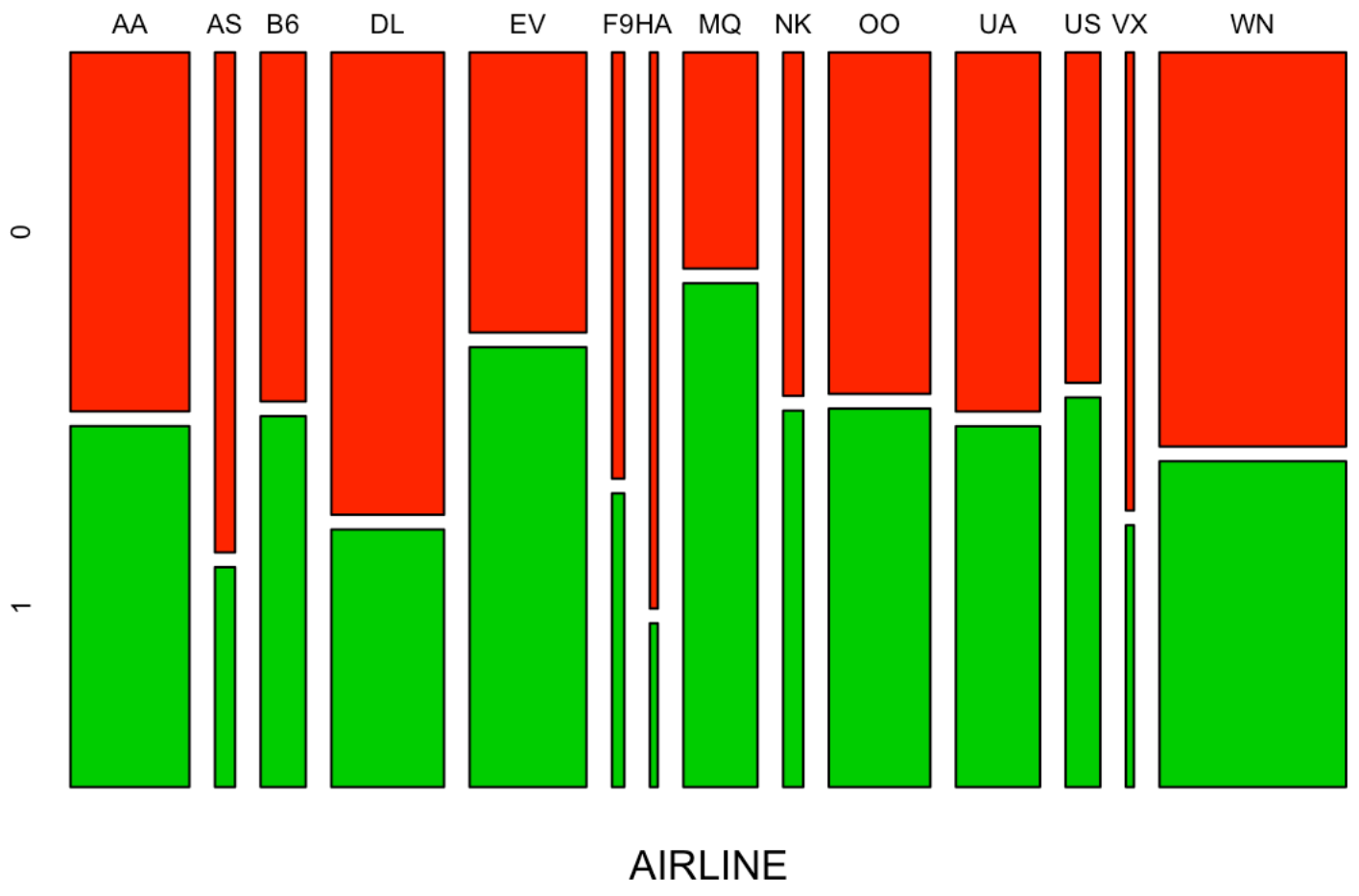
```
source('src/02-PREPROCESS/CSX415_Project_process.R')
```

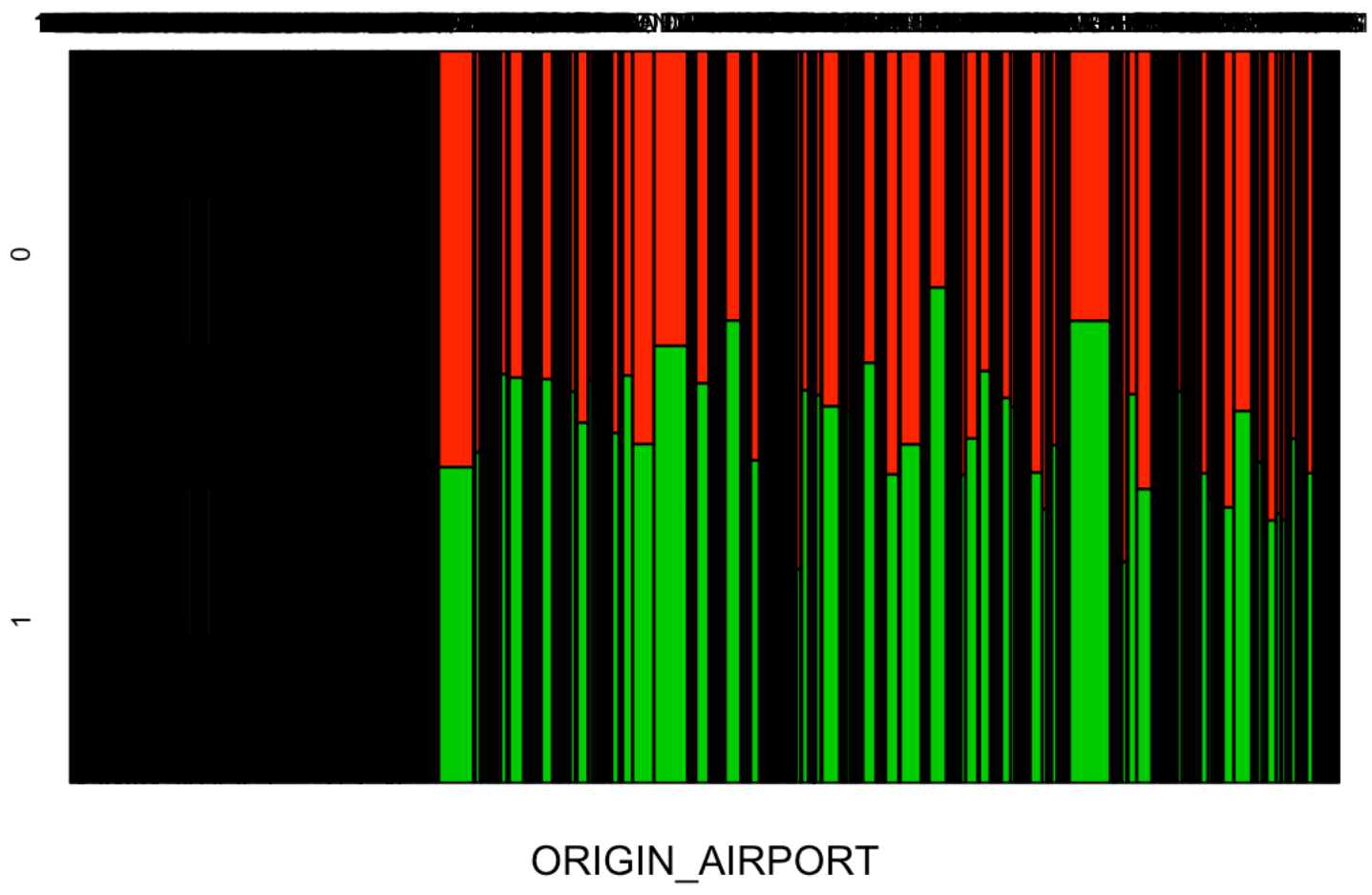
Modelling

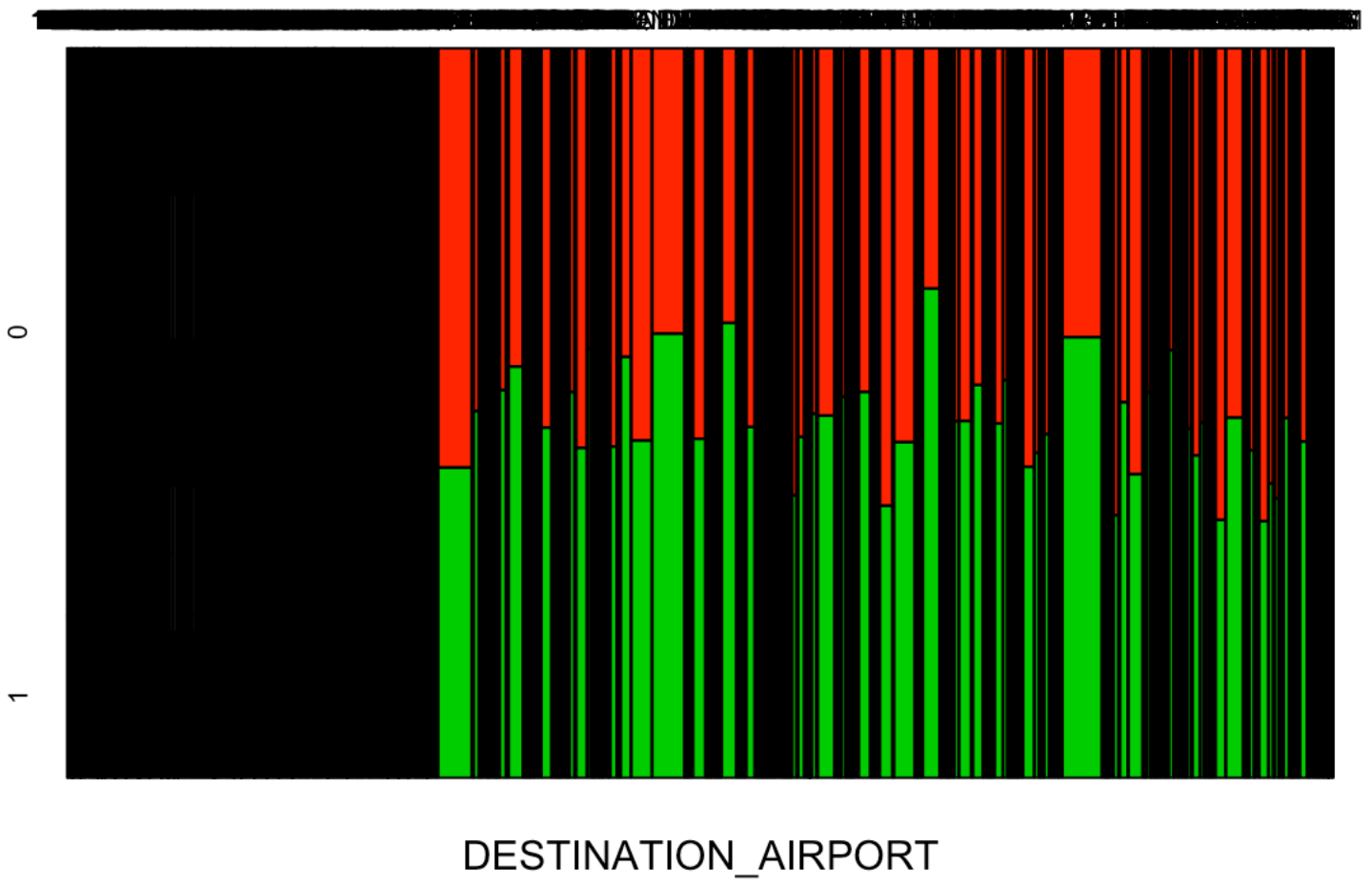
Apply Model

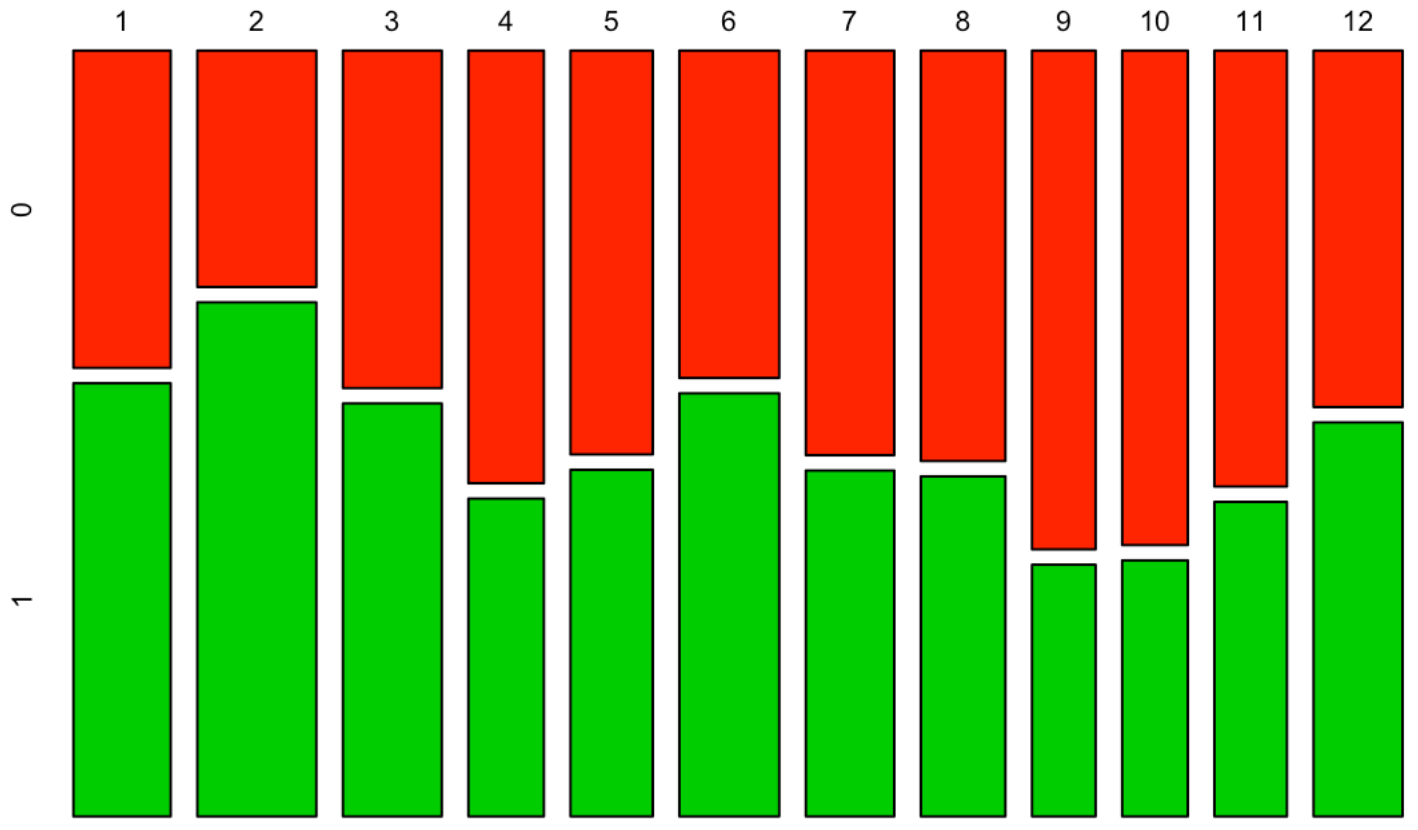
Naive Bayes

```
source('src/03-MODELS/CSX415_Project_Data_Model_Naive.R')
#nb.model
#summary(nb.model)
plot(nb.model)
```

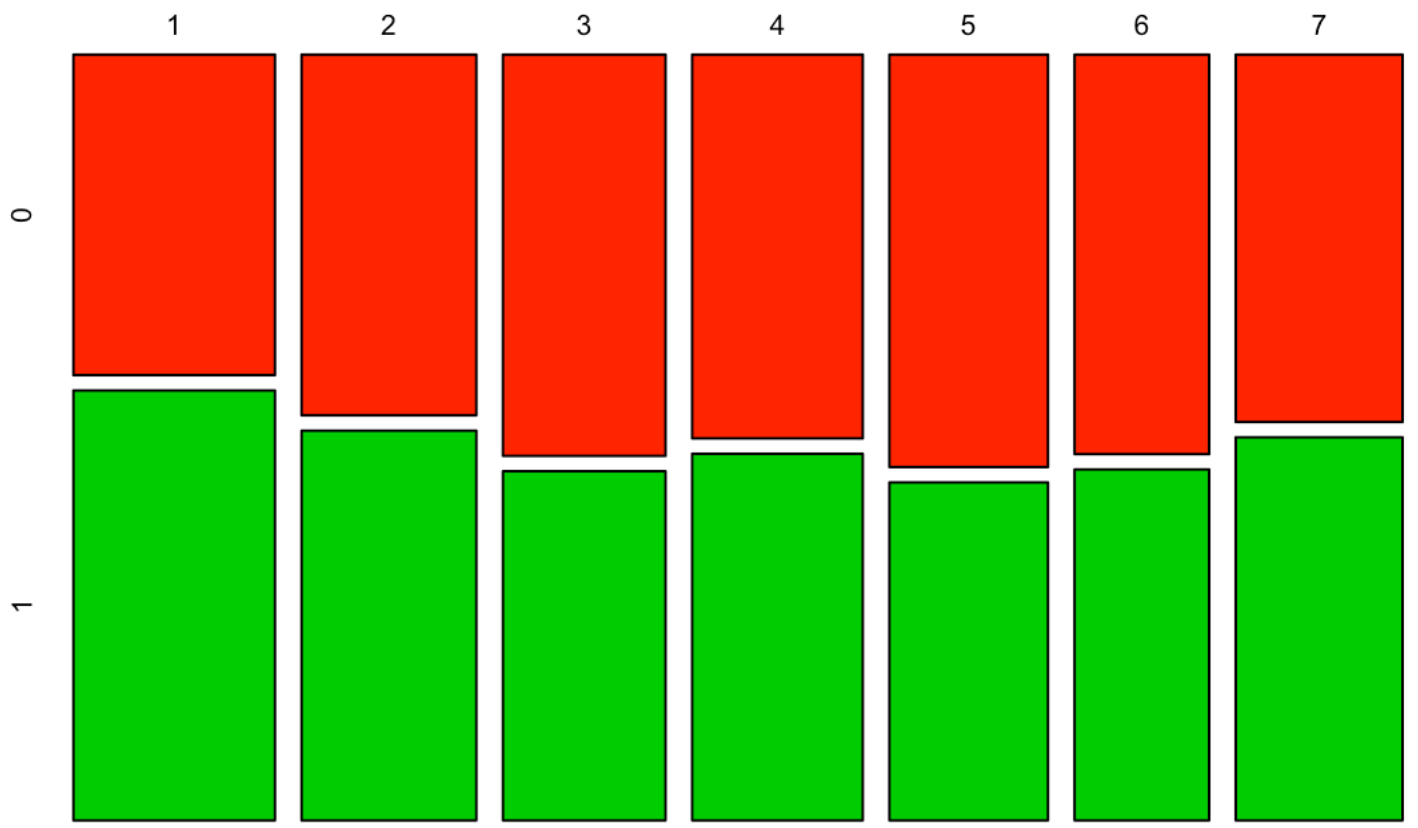

AA
AS
B6
DL
EV
F9
HA
MQ
NK
OO
UA
US
VX
WN
0
1
AIRLINE


0
1
ORIGIN_AIRPORT

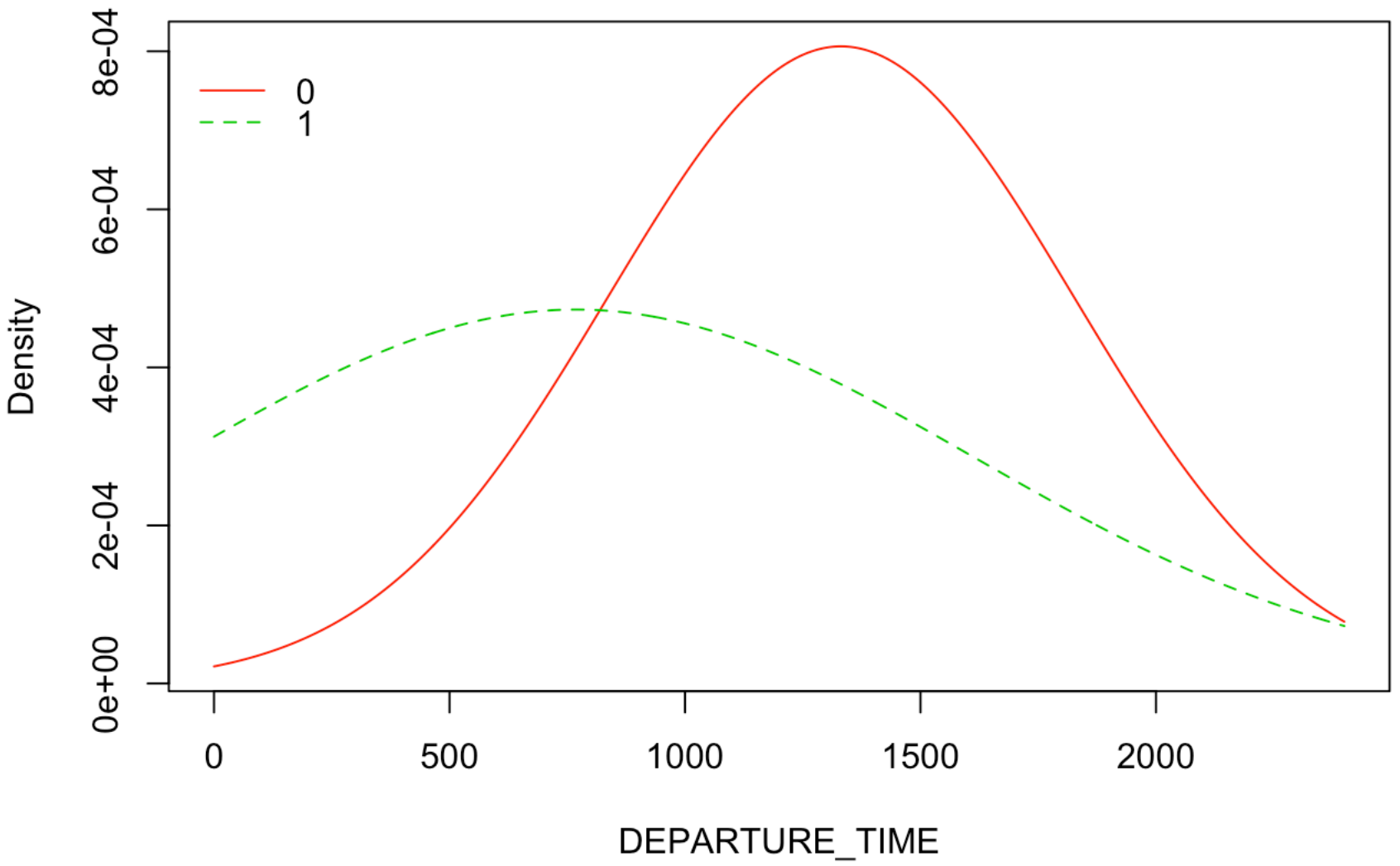

0
1
DESTINATION_AIRPORT



MONTH

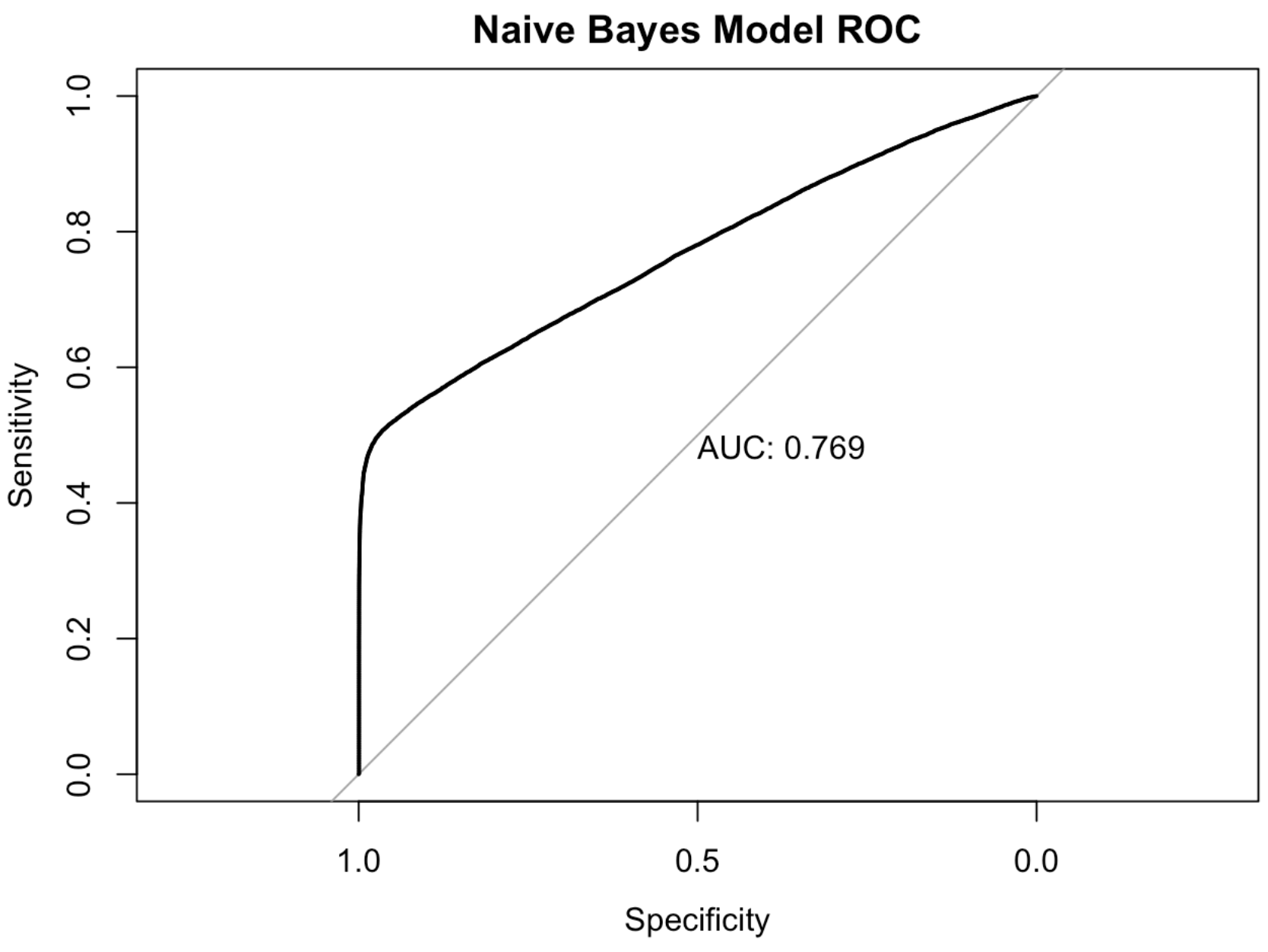


DAY_OF_WEEK



Naive Bayes Model ROC plot

```
plot.roc(TestData$DelayedOrCancelled,nb_pred_prob[,2],print.auc=TRUE,main="Naive Bayes Model ROC")
```

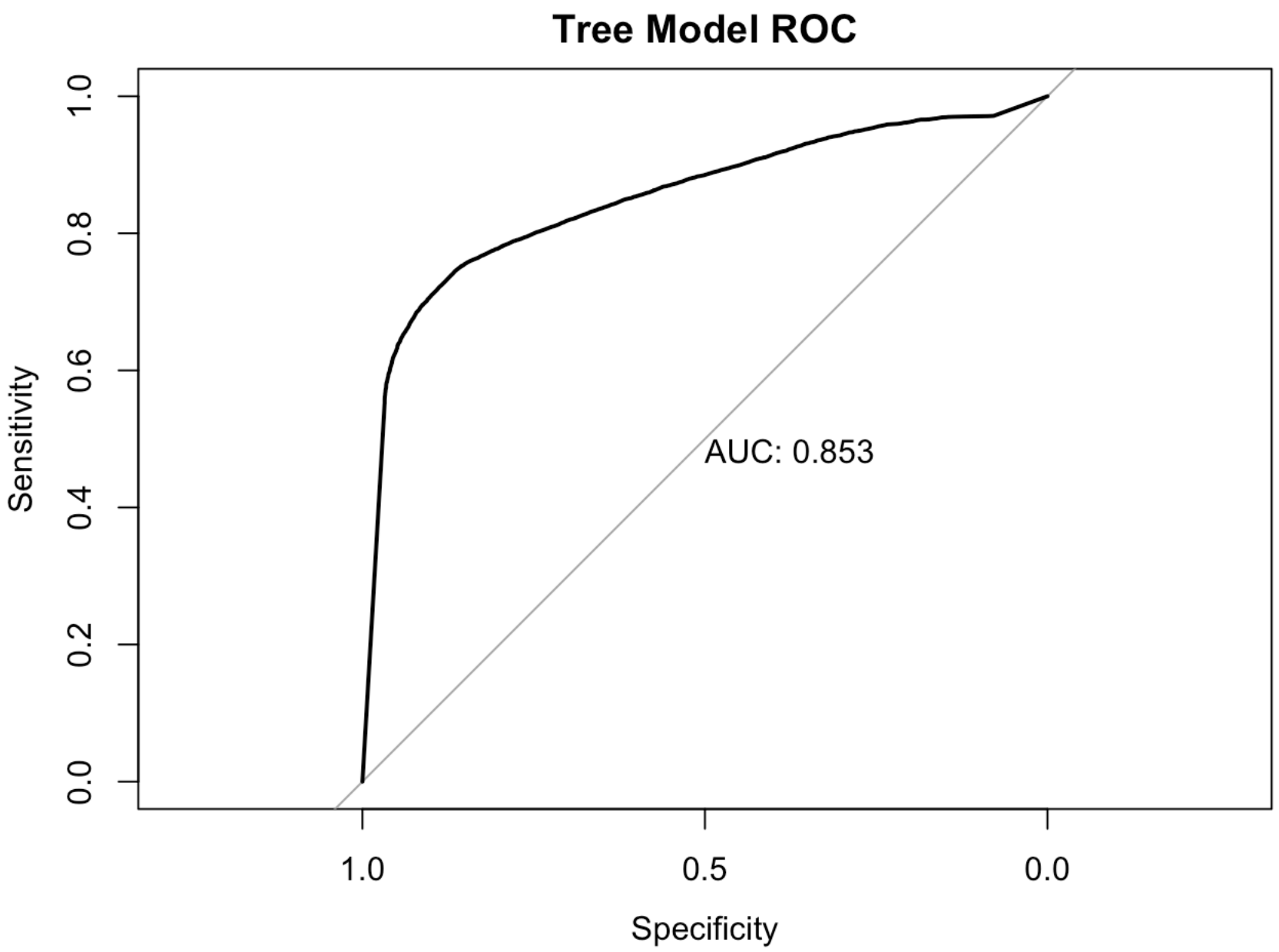


Logistic Regression model

```
# Logistic Regression model for flights data takes more than 10 hours
# to train, the trained model is saved but is 3.9 Gb in size (.rds)
# Also the ROC calculated is only 0.592, hence not using this model
#source('src/03-MODELS/CSX415_Project_Data_Model_LogisticRegression.R')
#glm.model
#summary(glm.model)
#plot.roc(TestData$DelayedOrCancelled,glm_predictions,print.auc=TRUE,main="GLM Model ROC")
```

Tree

```
source('src/03-MODELS/CSX415_Project_Data_Model_Tree.R')
#tree.model
#summary(tree.model)
#plot(tree.model)
plot.roc(TestData$DelayedOrCancelled,tr_predictions[,2],print.auc=TRUE,main="Tree Model ROC")
```



Model Evaluation

Naive Bayes Model Evaluation

```
confusionMatrix(TestData$DelayedOrCancelled,nb_predictions)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 26559  8898
##          1 12612 22845
##
##                Accuracy : 0.6967
##                  95% CI : (0.6933, 0.7001)
##     No Information Rate : 0.5524
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3933
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.6780
##             Specificity : 0.7197
##          Pos Pred Value : 0.7490
##          Neg Pred Value : 0.6443
##              Prevalence : 0.5524
##          Detection Rate : 0.3745
##    Detection Prevalence : 0.5000
##       Balanced Accuracy : 0.6989
##
##        'Positive' Class : 0
##
```

Tree Model Evaluation

```
tr_pred <- ifelse((tr_predictions[,2]>0.8), 1,0)
confusionMatrix(TestData$DelayedOrCancelled,tr_pred)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 31833  3624
##          1 10237 25220
##
##                Accuracy : 0.8045
##                  95% CI : (0.8016, 0.8075)
##     No Information Rate : 0.5933
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6091
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.7567
##             Specificity : 0.8744
##          Pos Pred Value : 0.8978
##          Neg Pred Value : 0.7113
##              Prevalence : 0.5933
##          Detection Rate : 0.4489
##    Detection Prevalence : 0.5000
##       Balanced Accuracy : 0.8155
##
##        'Positive' Class : 0
##
```

Model Selection

Tree Model is more accurate and ROC is greater than 0.75 as required compared to Naive Bayes model.

.

Logistic Regression model takes long time to train and ROC is less than 0.65 hence not selecting Logistic Regression model also because saved model .rds is 3.9 Gb in size hence not suitable for deloyment

.

Comparing the metrics, the accuracy and Kappa values of Tree Model are greater than Naive Bayes Model

.

Conclusion: Tree Model satisfies the requirements criteria of accuracy greater than 70% and ROC(AUC) greater than 0.65 and hence used for deployment