Predicting Train Arrival Status - On Time or Late

Introduction

Adanna Alutu

June 6, 2017

At the beginning of the project, it was hard to come up with a good data to analyze and predict the outcome. Initially I wanted to work on data from my job but after we coudn't see much dependence among the fields that made sense, my mentor Dr. Shmuel Naaman advised me to scout for data from other internet sites he recommended.

Since I take the train most of the time and experienced delay issues many times that has ranged from 10 mins to 2 hours, I became interested in working on transportation data for trains. This is because I want to experience the process of predicting outcomes which is made possible through Data Science. I want to focus on the steps that will make it possible for me and my mentor Dr Shmuel Naaman to predict the arrival times of the train. The possibility of cutting down the delays experienced in waiting for the train no longer seems to be far fetched. My mentor agreed with me and the Septa Train data from Kaggle website was a good option to work on. There were 3 different datasets available to work on but I chose the "on time performance" which I felt has more relevant features, variables and observations and also has sufficient data for the analysis, tests involved.

The variables in the dataset include: 1. train_id 2. status 3. origin 4. direction 5. next_station 6. timeStamp 7. date

subtitle:

Data Exploration

Several steps were taken to ensure elaborate data analysis and wrangling. Every bit of the data was maximized. We went beyond using the provided variables by creating new ones, removing unnecessary data and testing with reliable tools to get quality, reliable results that can be tested with any dataset.

It was necessary to take the following steps to ensure that all the combinations, dicing and testing would yield a meaningful interpretation and prediction that will help tell us with high confidence when the train will be late:

I. We first tried to plot charts with the entire data but the plots were too crowded and blurry to make any sense. The scales were distorted with big units affected by the outliers.

II. GGPlot bar charts were used to plot and observe the trends and statistics summary but the dataset was too huge for the charts.

III. My mentor suggested shuffling the data and taking the first 20percent as sample to work on. Using the formula below, the row-wise shuffling was done first before the column was then shuffled:

IV. We used the data to fit in several models which include:

- GGPlot with different combination of the variables.
- Linear regression model which was used different ways to get the best stastical summary. Including using some of the observations as variables.
- CART model with focus on the classification method because most of the variables in the data are categorical and the prediction is binary with 0 as "on Time" and 1 as "Late"
- Random Forest which created it's own model that highlighted the top more meaningful variables that contributed majorly in predicting the outcome.

Each of these models were implemented because the train dataset contains a mixture of numerical and categorical variables. Converting their types to either numeric or factors wasn't sufficient. To get the benefit of all the variables, it was essential to test these models.

subtitle:

Data Wrangling

Some data manipulations were done which include: + splitting some of the original variables nto separate variables. For example, time stamp variable was split into six variables. year, month day, hour, min, seconds. + Irrelevant variables were removed or set to null so they would not appear in the dataframe used for the predictions. + Some of observations from the wkday and day of month variables were converted to variables and they significantly improved the statistics of the models. The additions however increased the number of variables from 11 to 58. + Units attached to the dependent variable observations were removed to enable convesions to different types and allow plotting with only the observations of the same type. + The dependent variable "status" observations of "on Time" were replaced with "0" using gsub so that all the observations for the variable will match and easier to manipulate."On time" meant the train arrived as scheduled so it made sense to use "0" to represent no delay.

subtitle:

A Peek into some new variables

This section shows the summary of the SEPTA train data and the first few records using the head().

```
##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date

## Warning: Too many values at 150009 locations: 1, 2, 3, 4, 5, 6, 7, 8, 9,
## 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...

## 'data.frame':    150009 obs. of  45 variables:
##  $ status      : num  1.61 1.79 0 0 0 ...
##  $ origin      : Factor w/ 177 levels "16th St Jct",..: 54 141 115 65 32 151 49 49 151 101 ...
##  $ hour        : chr  "16" "08" "05" "09" ...
##  $ minute      : chr  "55" "54" "04" "32" ...
##  $ month       : num  4 9 6 3 6 10 9 6 10 6 ...
##  $ wkday1      : num  1 0 0 0 1 0 0 0 0 0 ...
##  $ wkday2      : num  0 0 1 0 0 0 0 1 0 0 ...
##  $ wkday3      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ wkday4      : num  0 0 0 1 0 0 0 0 0 0 ...
##  $ wkday5      : num  0 1 0 0 0 0 1 0 0 1 ...
##  $ wkday6      : num  0 0 0 0 0 1 0 0 1 0 ...
##  $ wkday7      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthday1   : num  1 0 0 0 0 0 0 0 0 0 ...
##  $ monthday2   : num  0 0 0 0 0 0 0 0 0 1 ...
##  $ monthday3   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthday4   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthday5   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthday6   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthday7   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthday8   : num  0 1 0 0 0 0 0 0 0 0 ...
```

```
##  $ monthday9   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthday10  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthday11  : num  0 0 0 0 0 1 0 0 1 0 ...
##  $ monthday12  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthday13  : num  0 0 1 0 0 0 0 1 0 0 ...
##  $ monthday14  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthday15  : num  0 0 0 0 0 0 1 0 0 0 ...
##  $ monthday16  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthday17  : num  0 0 0 0 1 0 0 0 0 0 ...
##  $ monthday18  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthday19  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthday20  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthday21  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthday22  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthday23  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthday24  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthday25  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthday26  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthday27  : num  0 0 0 1 0 0 0 0 0 0 ...
##  $ monthday28  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthday29  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthday30  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthday31  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ next_station: Factor w/ 155 levels "30th Street Station",..: 87 125 96 39 120 43 99 91 19 47 ...
##  $ direction   : Factor w/ 2 levels "N","S": 2 2 2 1 2 1 2 2 1 1 ...
```

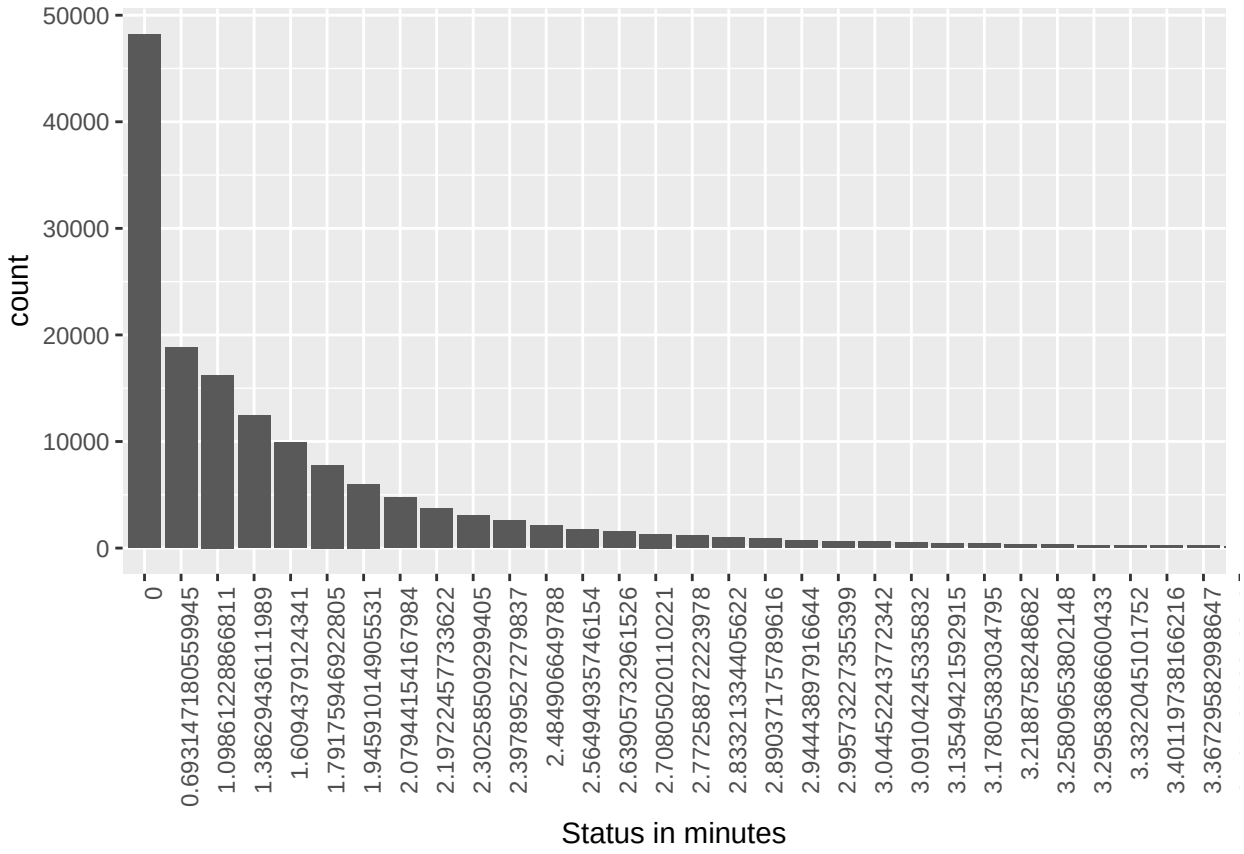
subtitle:

Some Initial plots

GGplot graphs used initally to see trends and relationships within the datasets. ####Status variable chart Status is the name of the dependent variable being predicted in this project. The bar chart shows the frequency of the delays experienced by passengers at the train station when the train is late.

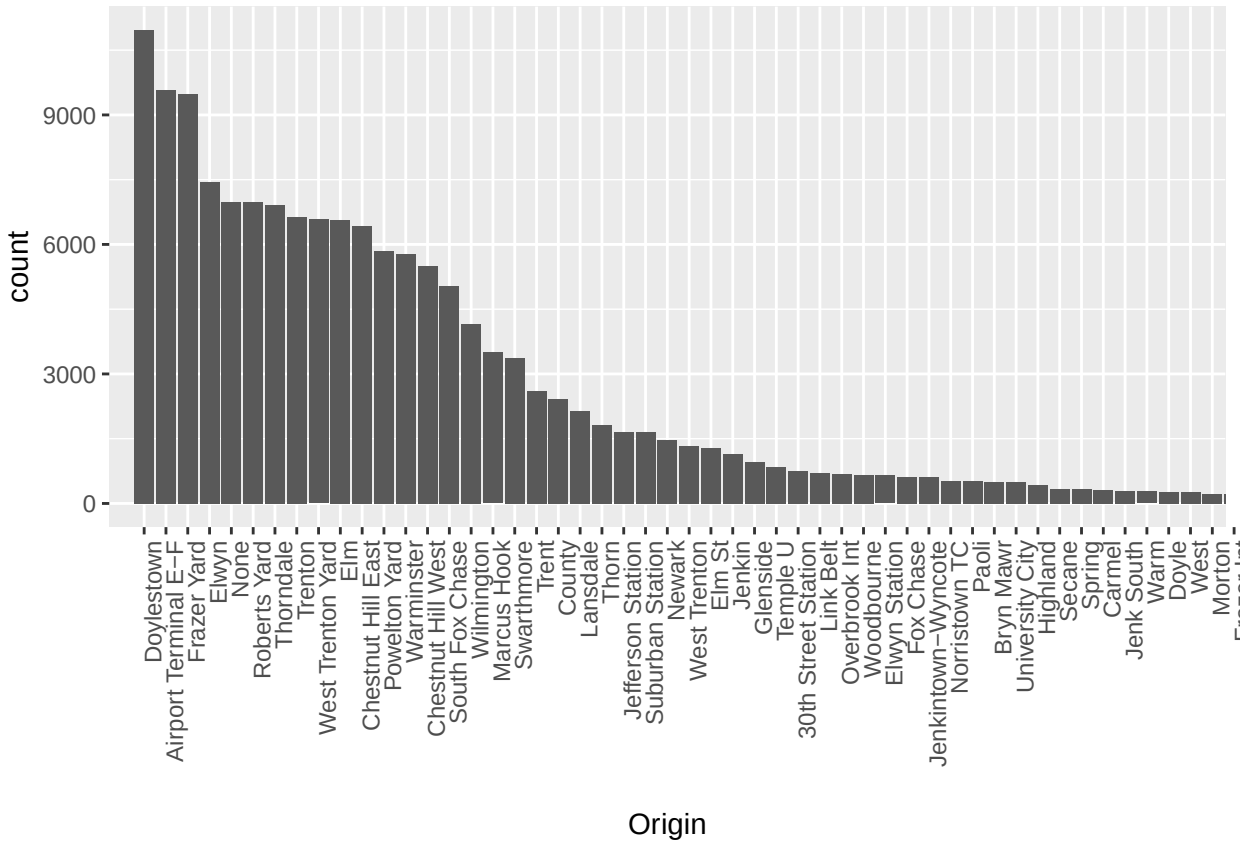
From the chart, we can tell that the trains are on time ~50% of the time and late 50% of the time. In this project, we want to predict when to expect the train to be late and when it will be early to avoid waste of time when possible.

```
library("ggplot2")
#set bar levels in descending order

train_var <- train_data$status
train_data2 <- within(train_data,
                      train_var <- factor(train_var,
                                          levels = names(sort(table(train_var),
                                                              decreasing = TRUE))))
trainstat_graph <- ggplot(train_data2, aes(x = train_var)) +
  geom_bar() +
   theme(axis.text.x = element_text(angle=90, hjust=1)) + coord_cartesian(xlim = c(1, 30)) + scale_x_dis

trainstat_graph
```



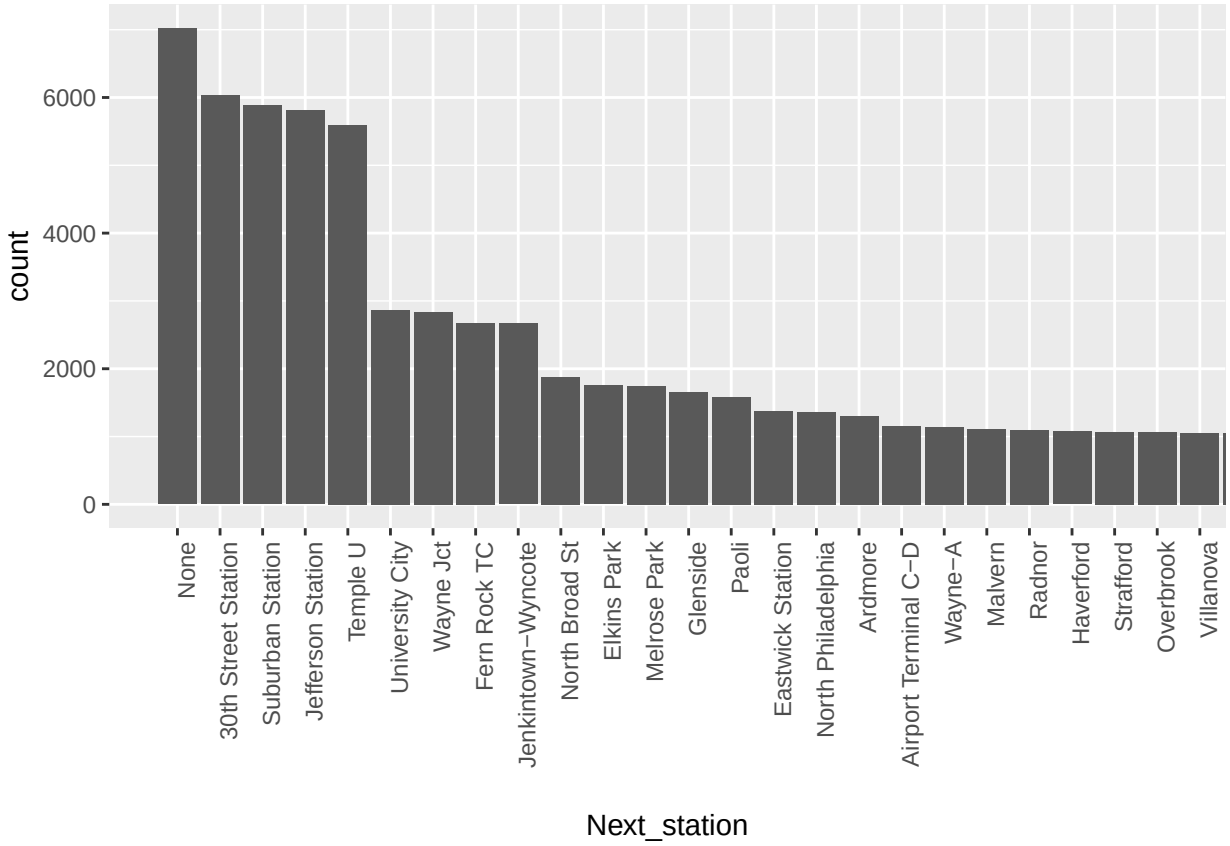
Origin variable bar chart This chart shows the origin which is also the station where each trip begins.

```
train_var <- train_data$origin

train_data2 <- within(train_data,
                      train_var <- factor(train_var,
                                          levels = names(sort(table(train_var),
                                                              decreasing = TRUE)), ordered = TRUE))

trainorig_graph <- ggplot(train_data2, aes(x =train_var)) +
  geom_bar() +
   theme(axis.text.x = element_text(angle=90, hjust=1)) + coord_cartesian(xlim = c(0, 50)) + scale_x_dis


trainorig_graph
```



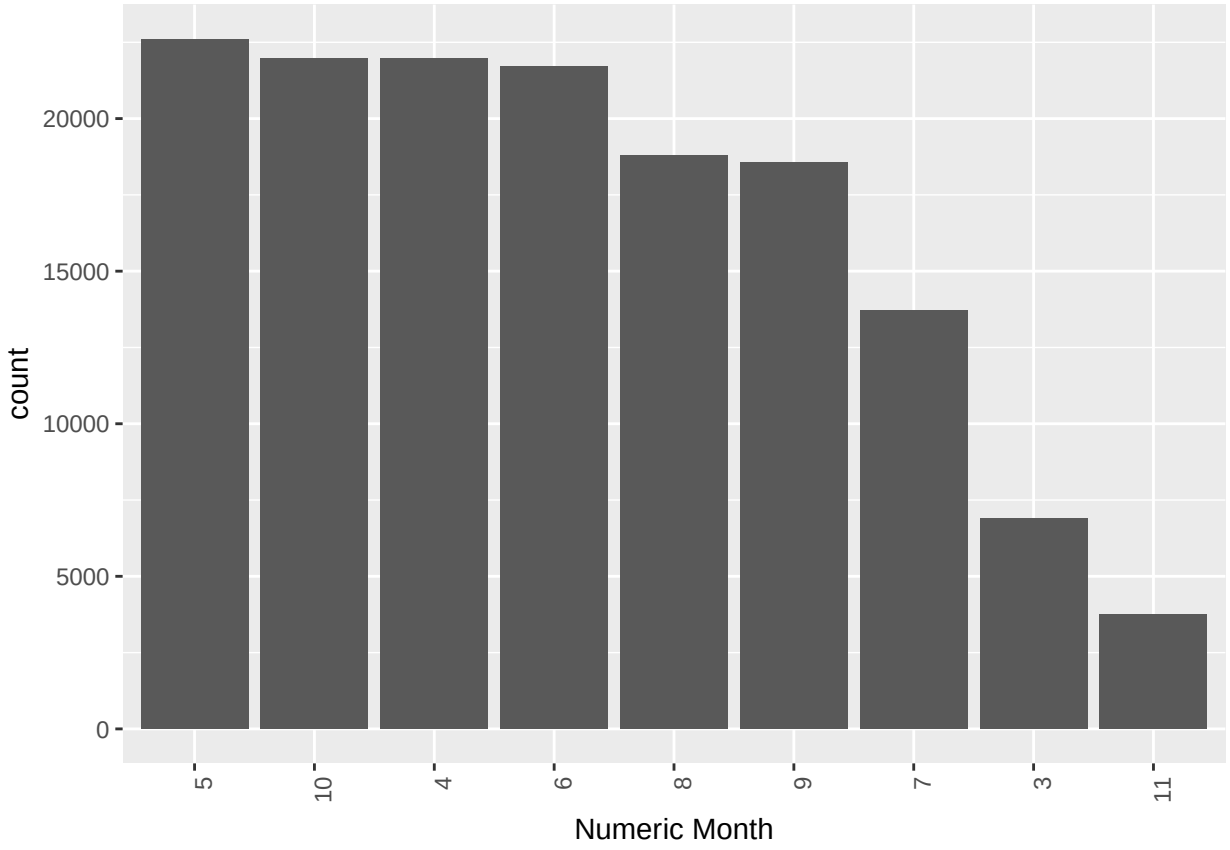
Next Station variable bar chart NextStation variable represents destination for each train ride.

```
train_var <- train_data$next_station


train_data2 <- within(train_data,
                      train_var <- factor(train_var,
                                          levels = names(sort(table(train_var),
                                                              decreasing = TRUE)), ordered = TRUE))

trainnext_graph <- ggplot(train_data2, aes(x = train_var)) +
  geom_bar() +
   theme(axis.text.x = element_text(angle=90, hjust=1)) + coord_cartesian(xlim = c(0, 25))+ scale_x_dis

trainnext_graph
```



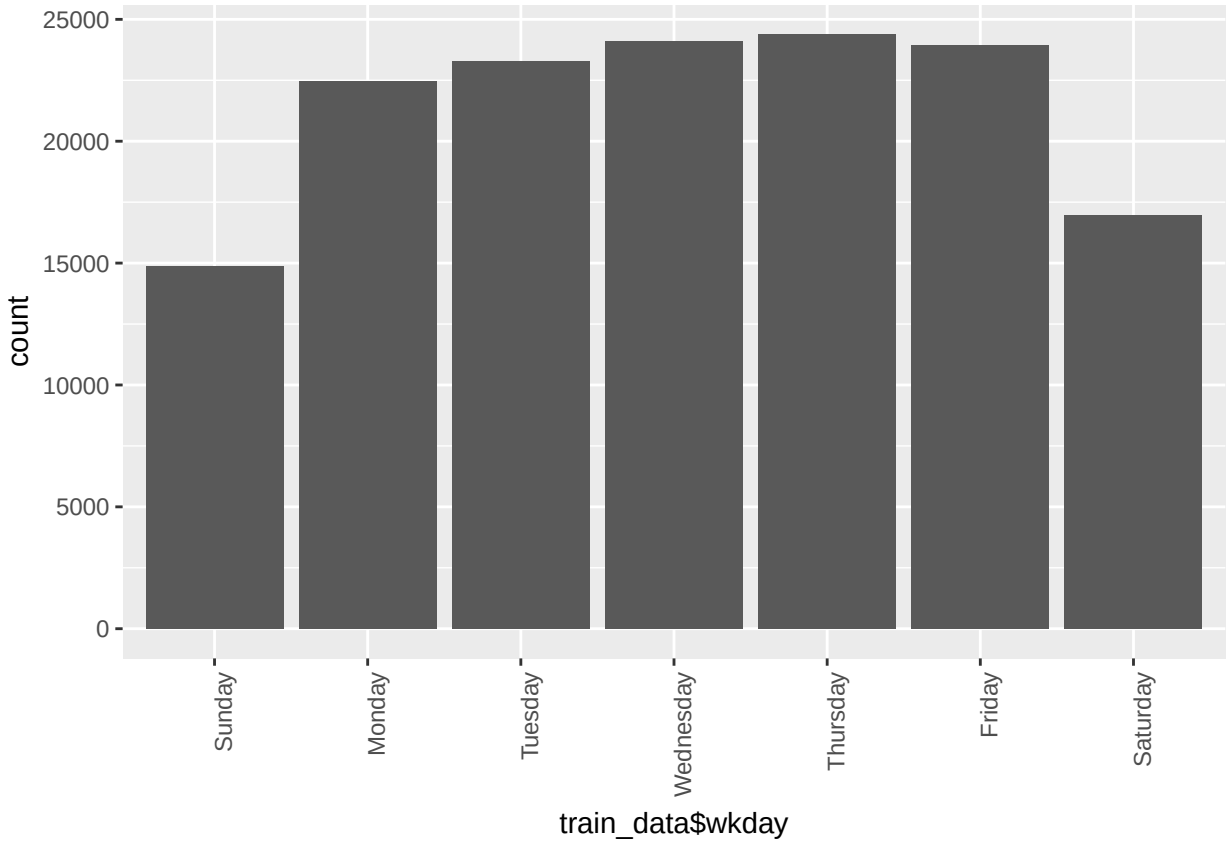
Month bar chart This is one of the new variables improvised by splitting up the timestamp variable.

```
train_var <- train_data$month


train_data2 <- within(train_data,
                      train_var <- factor(train_var,
                                          levels = names(sort(table(train_var),
                                                              decreasing = TRUE)), ordered = TRUE))
trainmonth_graph <- ggplot(train_data2, aes(x = train_var)) +
  geom_bar() +
   theme(axis.text.x = element_text(angle=90, hjust=1))+ scale_x_discrete(name = "Numeric Month")

trainmonth_graph
```



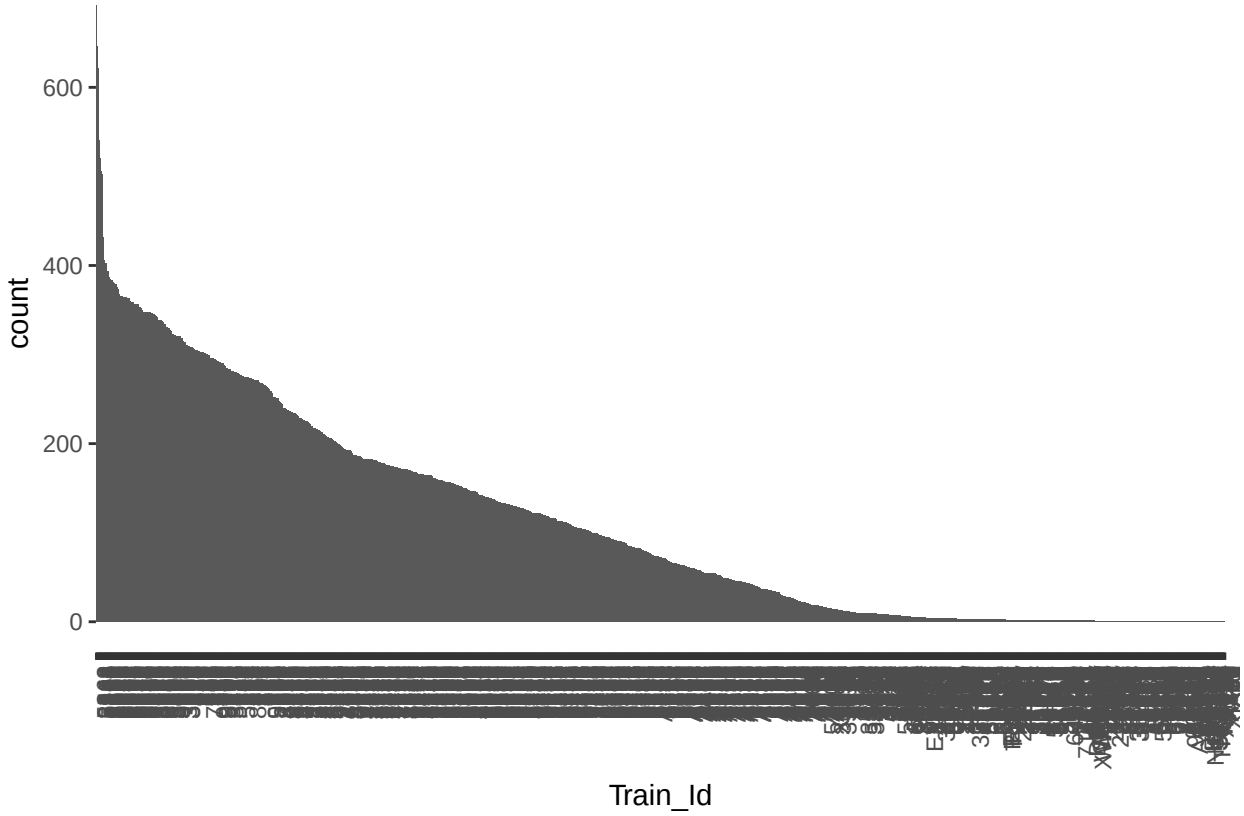
The Weekday chart The weekday chart shows the number of trains that run different days of the week. More trains run during the week and fewer trains on the weekends. The busiest day is Thursday.

```
train_data2 <- within(train_data,
                      train_data$wkday <- factor(train_data$wkday,
                                                 levels = c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursda


trainwkday_graph <- ggplot(train_data2, aes(x = train_data$wkday)) +
  geom_bar() +
   theme(axis.text.x = element_text(angle=90, hjust=1))

trainwkday_graph
```



Train id This is a chart of all the train ids in descending order. These are all the trains that transported passengers during the period of one year in our dataset.

```
train_var <- train_data$train_id


train_data2 <- within(train_data,
                      train_var <- factor(train_var,
                                          levels = names(sort(table(train_var),
                                                              decreasing = TRUE)), ordered = TRUE))

trainid_graph <- ggplot(train_data2, aes(x = train_var)) +
  geom_bar() +
   theme(axis.text.x = element_text(angle=90, hjust=1))+ scale_x_discrete(name = "Train_Id")

trainid_graph
```



subtitle:

More conversion of observations into variables

Convert the weekday and monthday observations to columns. The purpose is to increase the number of variables that contribute to the status (delay and on time arrivals) of the train.

To achieve this, a matrix was used. In this case, the matrix translated the values of the new columns to "0" and "1"."1" was printed when the train travelled in the specified day or month. "0" was used to fill the rest of the observations that the train did not ride during the week day or day of the month. for example, the new column "wkday1" represents "Monday". The value "1" in that column represents the train rides that happened on Mondays.

```
#Added contrasts to print all travel days and all days of the months otherwise some are skipped.
```

subtitle:

Linear Regression model test

Linear model statistical summary for Status based on the independent variable x = origin. The linear regression model is one of the models implemented in this project in efforts to predict the delays of the train.

```
train_data$monthday <- as.numeric(train_data$monthday)

library(broom)

glance(summary(lm(log(as.numeric( train_data$status)+1) ~ origin, data = train_data)))
```

```
##   r.squared adj.r.squared     sigma statistic p.value  df
## 1 0.1600799     0.1590933 0.4543964  162.2523       0 177
```

Added more variables to see the effect on the status of the train, which is the variable we are trying to predict.

```
logvar <- log(as.numeric( train_data$status)+1)
glance(summary(lm(logvar ~ origin+ hour + month + monthday + wkday1 + wkday2 + wkday3 + wkday4 + wkday5
```

```
##   r.squared adj.r.squared     sigma statistic p.value  df
## 1 0.2162341     0.2151563 0.4389879  200.6264       0 207
```

This is the same regression model with all the 45 variables.

```
glance(summary(lm(status ~ ., data = train_data)))
```

```
##   r.squared adj.r.squared     sigma statistic p.value   df
## 1 0.3721483     0.3645101 0.7935369  48.72211       0 1804
```

Next is the linear model chat which was used to get better statistics. To achieve a much better R-value > 24%, all significant independent variables were added including the new ones created by the matrix that were converted from observations to variables.

The original variable count was 11, the addition of the new variables increased the variable count to 58.The improvement of the variables count definitely contributed to a better statistics which increased from .06% to 24%.

```
#Use glance() to print only the statistics and not both statistics and train_data summary
#the . represents all variables in dataframe e
glance(summary(lm(status ~ ., data = train_data)))
```

```
##   r.squared adj.r.squared     sigma statistic p.value   df
## 1 0.3721483     0.3645101 0.7935369  48.72211       0 1804
```

subtitle:

CART Model /Decision Tree

:

In this section, the CART model is implemented. The two options considered are Classification and Regression CART models/trees but the reression model is preferred so that the results can be compared with the linear regression model used above. It's like comparing apples to apples or oranges to oranges.

I found this site very helpful because they explained in detail the conditions for the variables before a succesful model can be achieved - https://rstudio-pubs-static.s3.amazonaws.com/27179_e64f0de316fc4f169d6ca300f18ee2aa.html.

Prior to finding this site, only the root or just one circle with a number (4.6) was drawn.

```
library(caTools)
set.seed(3000)

smp_size <- floor (0.8 *nrow(Data))
```

```
train_ind <- sample (seq_len(nrow(Data)), size=smp_size)

train <- Data[train_ind, ]
test  <- Data[-train_ind,]


#build CART model

library(rpart)
library(rpart.plot)

#now create the CART model. Use rpart to build a linear regression tree since the status variable being

#use rpart formula to fit the data.

#logvar <- log(as.numeric( train_data$status)+1)
TraindataTree2 = rpart(status ~. , data = train)



actual <- test$status
predicted <- predict(TraindataTree2, newdata = test )

R2_test <- 1 - (sum((actual-predicted )^2)/sum((actual-mean(actual))^2))



actual <- train$status
predicted <- predict(TraindataTree2, newdata = train )

R2_train <- 1 - (sum((actual-predicted )^2)/sum((actual-mean(actual))^2))
R2_test
```

```
## [1] 0.207283
```

```
R2_train
```

```
## [1] 0.2086067
```

```
#print R2
```

The r-squared value is .318 ~ 32%. So far, the CART linear regression model performance is slightly better because the r-squared value is .335 ~ 34%.

subtitle:

Check Performance of CART Regression model

Further test to check the performance of the Training and Test variables. The Trainset data will be trained and performance measured at 5%, 10%, 30%, 50%, 70%, 80% of the observations. A for loop is used to accomplish this in the code.

The result from the plot shows that the Training set performed better than the Testset. The Training set started leveling around 15% of the data used. So no need to use the rest of the data since they don't add more value or improve the performance. The test data on the other hand started leveling at around 10% and the numbers are on the negative side. So only the first 5% contributes to the performance of the Test set.

```
#set.seed(2)

#actualtst <- log(Testset$status + 1)
rm (R80trn)
```

First is a check on the Test and Training data.

```
## Warning in rm(R80trn): object 'R80trn' not found
```

```
 rm (R80tst)
```

```
## Warning in rm(R80tst): object 'R80tst' not found
```

```
 rm(jj)
```

```
## Warning in rm(jj): object 'jj' not found
```

```
jj<-0
j <- .01 * 150000
k <- 0

R80trn <- 0
R80tst <- 0
for(i in 1:j){
    if(j >=150000){
      #get out of loop once all data has been processed
      break
    }
  k = k + 1
  train_d= train[1:j,]
 jj[k] <- j
  TraindataTreePerf = rpart(status ~ ., train_d)


  predstrn <- predict(TraindataTreePerf, newdata = train)

predstest <- predict(TraindataTreePerf, newdata = test)


R80trn[k] =   1-sum((train$status-predstrn)^2)/sum((train$status-mean(train$status))^2)


R80tst[k] =   1-sum((test$status-predstest)^2)/sum((test$status-mean(test$status))^2)


jj
print(j)

j <- j+j

}
```

```
## [1] 1500
## [1] 3000
## [1] 6000
## [1] 12000
## [1] 24000
## [1] 48000
## [1] 96000
```

```
#r-squared for CART Training set
R80trn
```

```
## [1] -0.12627769  0.08038021  0.17325549  0.20986960  0.20760042  0.20648902
## [7]  0.20828764
```

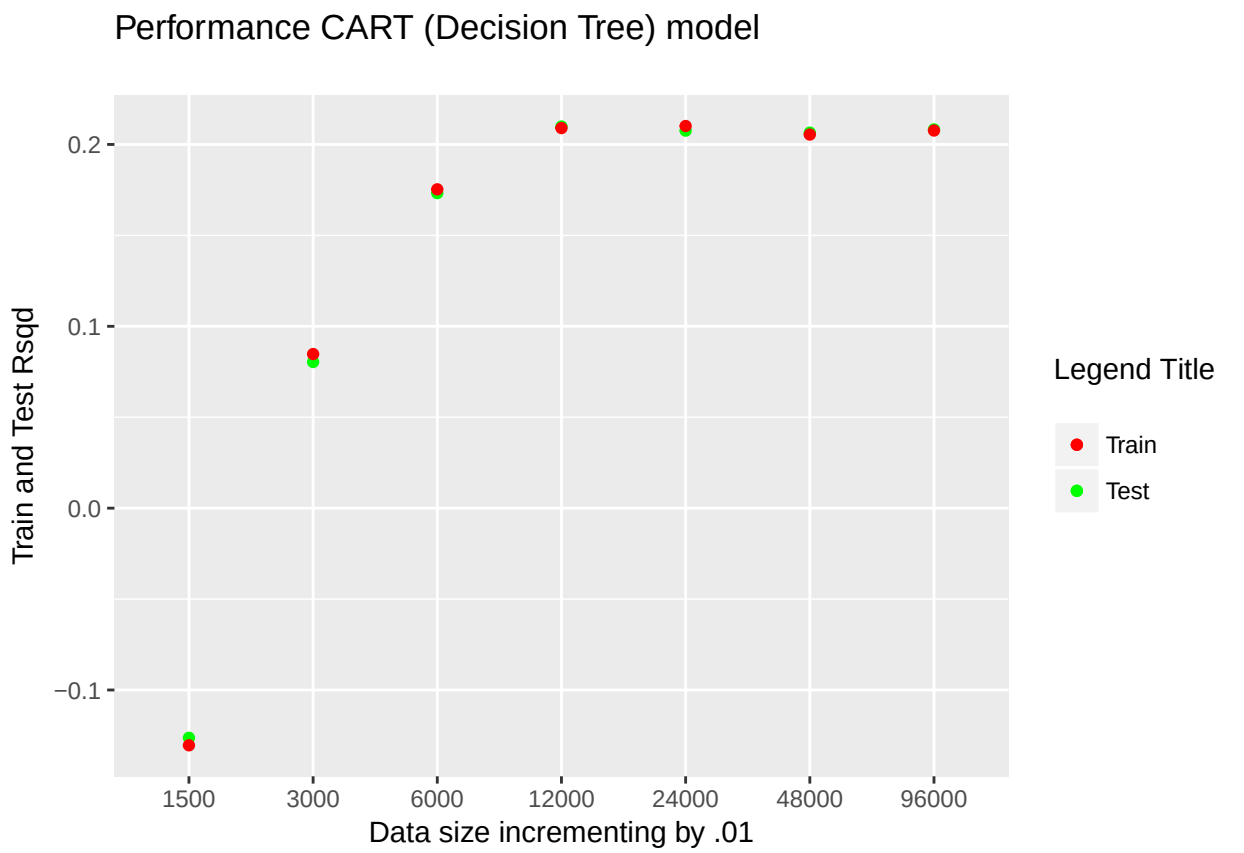
```
#r-squared for CART Testing set
R80tst
```

```
## [1] -0.13042049  0.08474824  0.17532162  0.20901182  0.21014193  0.20543047
## [7]  0.20764966
```

```
b <- data.frame(jj, R80trn, R80tst)
print(b)
```

```
##      jj      R80trn      R80tst
## 1  1500 -0.12627769 -0.13042049
## 2  3000  0.08038021  0.08474824
## 3  6000  0.17325549  0.17532162
## 4 12000  0.20986960  0.20901182
## 5 24000  0.20760042  0.21014193
## 6 48000  0.20648902  0.20543047
## 7 96000  0.20828764  0.20764966
```

```
  ggplot(b) +
    geom_point(aes(x = factor(jj), y = R80trn, color = "red")) +
  geom_point(aes(x = factor(jj), y = R80tst, color = "green"))+
      labs(title = "Performance CART (Decision Tree) model\n", x = "Data size incrementing by .01", y =
  scale_color_manual(labels = c("Train", "Test"), values = c("red", "green"))
```




```
  #xlab(
  #xlab("Train -Rsqd")+
   # ylab("Test -Rsqd")
```

subtitle:

Checking max depth on the Test set for the CART models

The depths start leveling around the sixth level. There is no need to increase the depth more than the sixth level. Testing max depth on training set. The result from the plot shows that the best performance is reached at max_depth = 2. There is no need to grow a tree with any depth greater than 5.

```
set.seed(2)

#R2ts <- numeric()
  rm(R2ts)
```

```
## Warning in rm(R2ts): object 'R2ts' not found
```

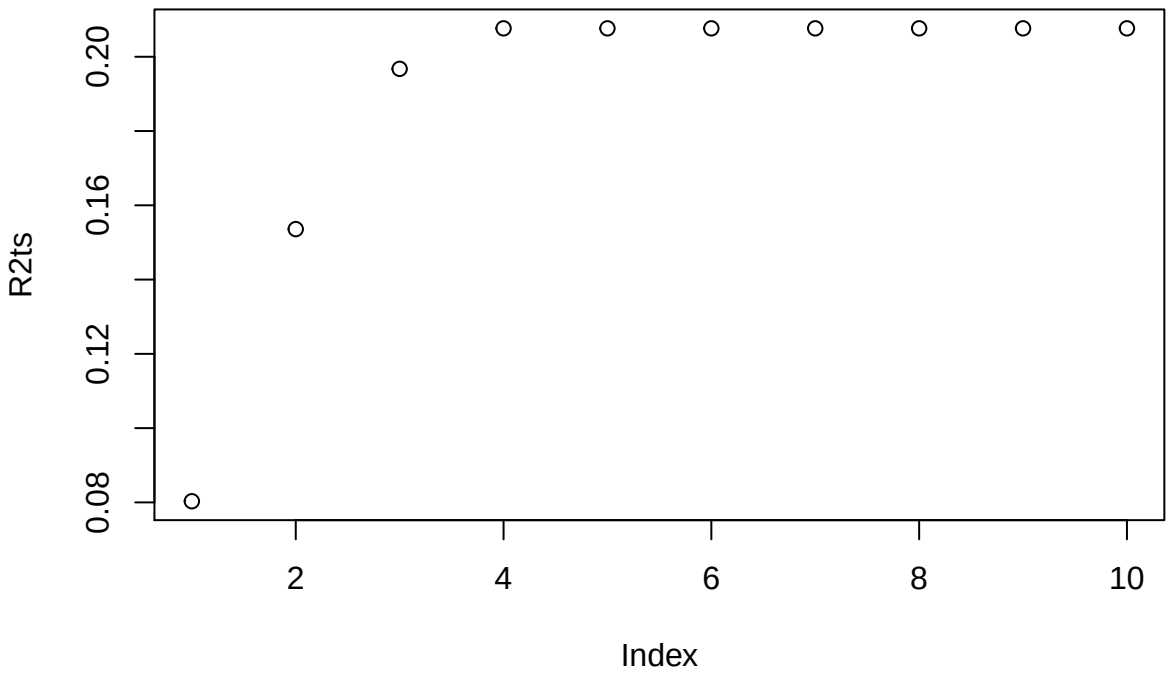
```
  rm(R2tn)
```

```
## Warning in rm(R2tn): object 'R2tn' not found
```

```
  rm(jj)
R2ts <- 0
R2tn <- 0
k <-0
jj <-0

for (iv in 1:10){

  #TraindataTreePerf = rpart(status ~ ., train_d)
TestsetTree100p = rpart(status ~ ., control = list(maxdepth = iv), train_d)

k <- iv
jj[k] <- k
preds <- predict(TestsetTree100p, newdata = test)
#k=k+1

R2ts[k] <- (1 - (sum((test$status-preds )^2)/sum((test$status-mean(test$status))^2)))

predstn <- predict(TestsetTree100p, newdata = train)
#k=k+1

R2tn[k] <- (1 - (sum((train$status-predstn )^2)/sum((train$status-mean(train$status))^2)))


}
#list the r-squared values at different depths
R2ts
```

```
##  [1] 0.08031145 0.15358830 0.19674656 0.20764966 0.20764966 0.20764966
##  [7] 0.20764966 0.20764966 0.20764966 0.20764966
```

```
plot(R2ts)
```




```
R2tn
```

```
##  [1] 0.07738756 0.15218426 0.19646121 0.20828764 0.20828764 0.20828764
##  [7] 0.20828764 0.20828764 0.20828764 0.20828764
```

```
md <- data.frame(jj, R2ts, R2tn)
print(md)
```

```
##    jj       R2ts       R2tn
## 1   1 0.08031145 0.07738756
## 2   2 0.15358830 0.15218426
## 3   3 0.19674656 0.19646121
## 4   4 0.20764966 0.20828764
## 5   5 0.20764966 0.20828764
## 6   6 0.20764966 0.20828764
## 7   7 0.20764966 0.20828764
## 8   8 0.20764966 0.20828764
## 9   9 0.20764966 0.20828764
## 10 10 0.20764966 0.20828764
```

```
  ggplot(md) +
    geom_point(aes(x = factor(jj), y = R2tn, color = "red")) +
  geom_point(aes(x = factor(jj), y = R2ts, color = "green"))+
      labs(title = " Maximum Depth for Training and Testing data - Decision Tree \n", x = "Maximum depth
  scale_color_manual(labels = c("Train", "Test"), values = c("red", "green"))
```


Maximum Depth for Training and Testing data − Decision Tree
Train and Test
0.20
0.16
0.12
0.08
1
2
3
4
5
6
7
8
9
10
Maximum depth inc by 1
Legend Title
Train
Test