Flight Delay Prediction

Outline

- Introduction
- Data Pre-processing and Analysis
- Modelling Techniques
- Results
- Conclusion

Introduction

- Travel by air has become natural choice for business or personal
- Punctuality has become an important factor of travel
- Concern for improvement in delays
- Better quality of on-time airline performance

Data Set Info

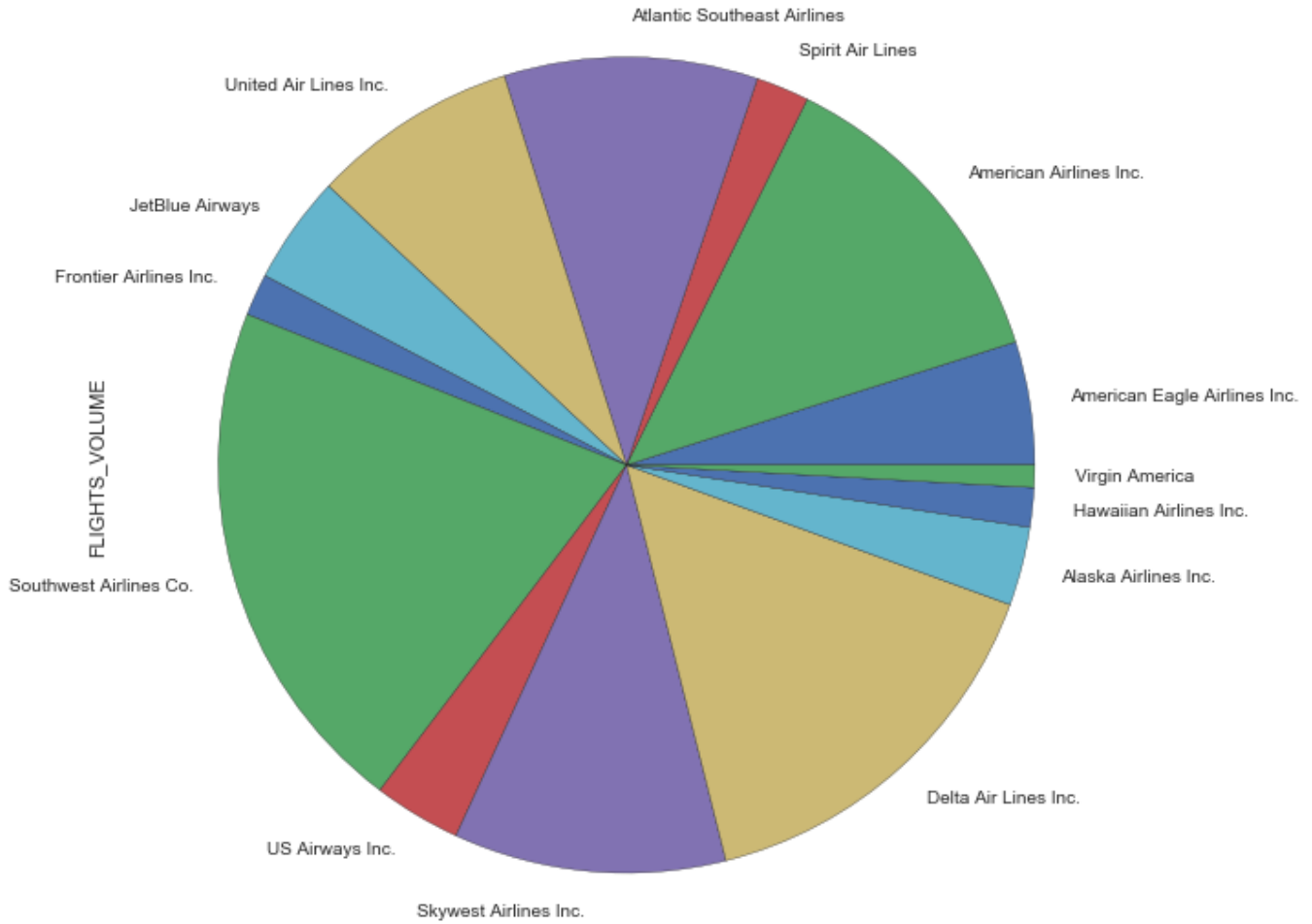
- Number of records: Approx 1 Million
- For analysis purpose, we used 5000 sampled data
- Number of variables : 31
- Building Model (Train :80% / Test : 20%)

Data Pre-processing

- Firstly, we analysed the number of missing values and their importance
- Removed NA values for building well-structured data
- Feature selection: Most significant attributes chosen (departure time and arrival time, the origin and destination airports, distance, taxi-in time, taxi-out time, elapsed time in air, delay time, if cancelled, distribution of delay time among multiple reasons)

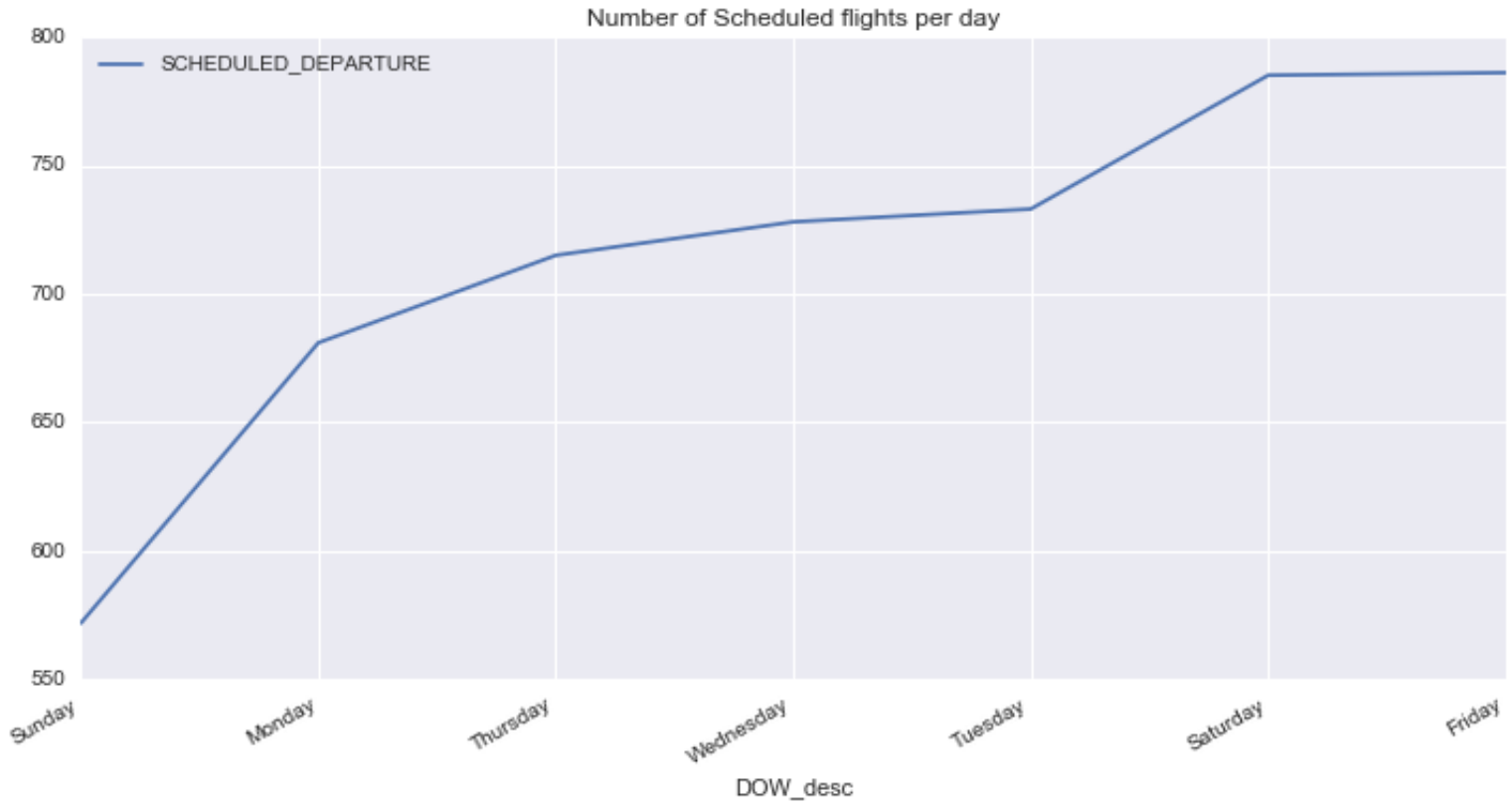
Airline Flight Volume

Shows that Southwest Airlines runs most of the airplanes in US in 2015

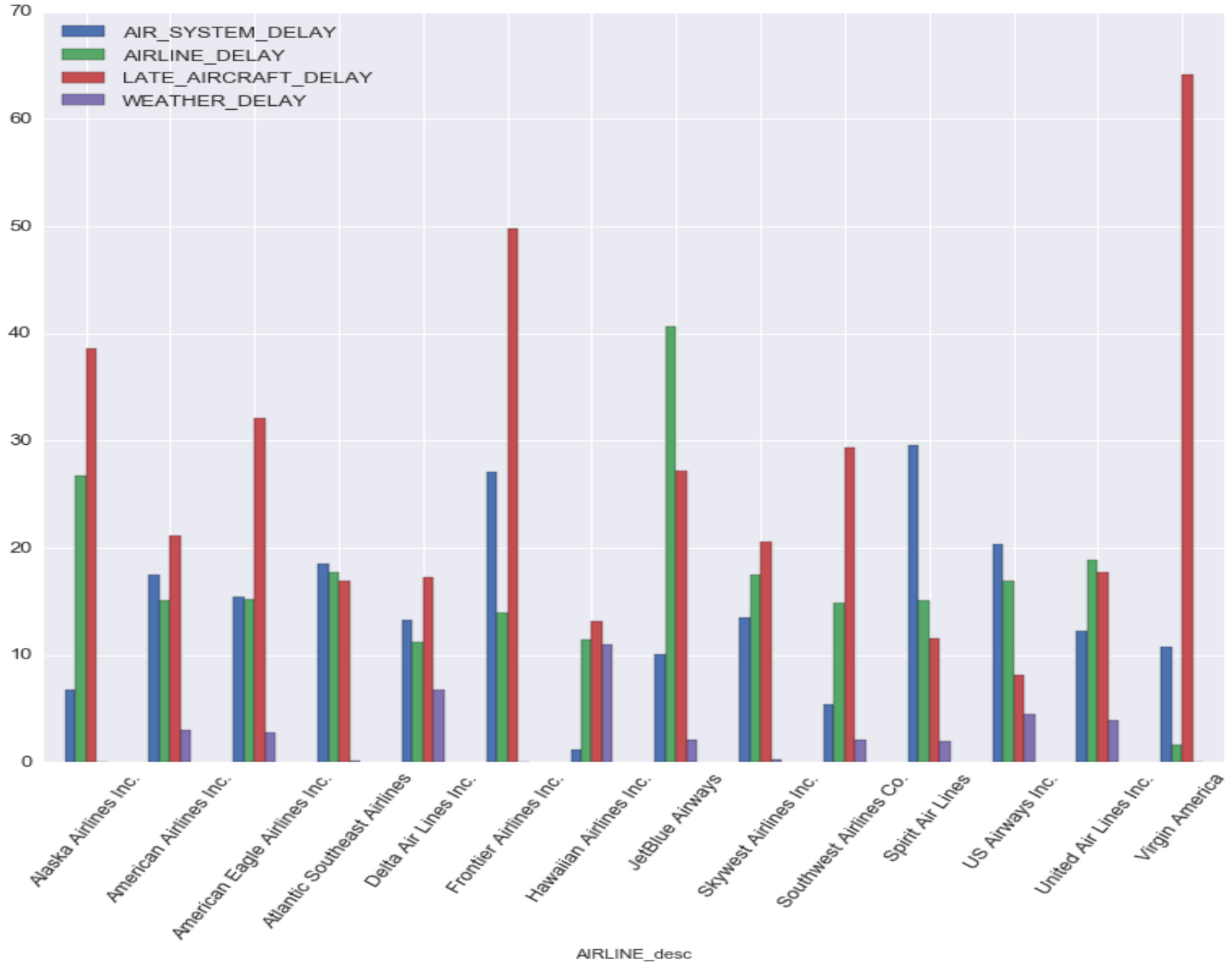


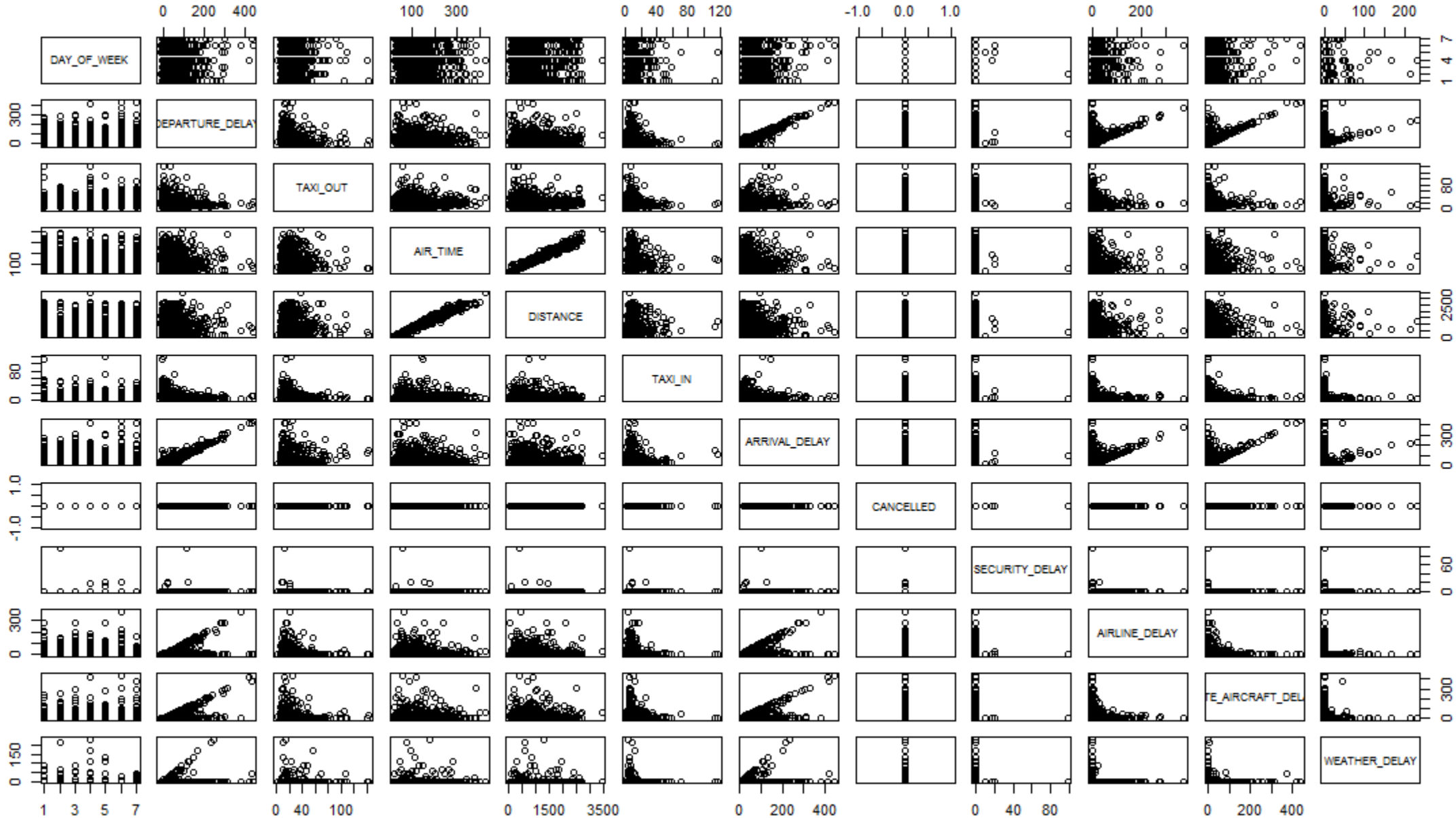
Shows Tuesday, Saturday and Friday as the busiest days

Scheduled flights in a week



Reasons why a flight could get delayed




DAY_OF_WEEK
DEPARTURE_DELAY
TAXI_OUT
AIR_TIME
DISTANCE
TAXI_IN
ARRIVAL_DELAY
CANCELLED
SECURITY_DELAY
AIRLINE_DELAY
TE_AIRCRAFT_DEL
WEATHER_DELAY

Attributes Details

```
 DAY_OF_WEEK    DEPARTURE_DELAY      TAXI_OUT         AIR_TIME        DISTANCE
Min.   :1.000   Min.   :-13.00   Min.   :  1.00   Min.   : 18.0   Min.   :  67.0
1st Qu.:2.000   1st Qu.: 17.00   1st Qu.: 12.00   1st Qu.: 63.0   1st Qu.: 392.5
Median :4.000   Median : 37.00   Median : 16.00   Median :101.5   Median : 690.0
Mean   :3.969   Mean   : 51.81   Mean   : 20.98   Mean   :119.5   Mean   : 845.4
3rd Qu.:5.000   3rd Qu.: 66.25   3rd Qu.: 24.00   3rd Qu.:149.0   3rd Qu.:1076.0
Max.   :7.000   Max.   :436.00   Max.   :143.00   Max.   :425.0   Max.   :3417.0
   TAXI_IN         ARRIVAL_DELAY      CANCELLED SECURITY_DELAY    AIRLINE_DELAY
Min.   :  1.000   Min.   : 15.00   Min.   :0   Min.   : 0.0000   Min.   :  0.00
1st Qu.:  4.000   1st Qu.: 23.00   1st Qu.:0   1st Qu.: 0.0000   1st Qu.:  0.00
Median :  6.000   Median : 36.00   Median :0   Median : 0.0000   Median :  2.00
Mean   :  8.824   Mean   : 56.29   Mean   :0   Mean   : 0.1856   Mean   : 16.85
3rd Qu.:  9.000   3rd Qu.: 65.25   3rd Qu.:0   3rd Qu.: 0.0000   3rd Qu.: 18.00
Max.   :118.000   Max.   :445.00   Max.   :0   Max.   :98.0000   Max.   :377.00
LATE_AIRCRAFT_DELAY WEATHER_DELAY
Min.   :  0.00      Min.   :  0.000
1st Qu.:  0.00      1st Qu.:  0.000
Median :  5.00      Median :  0.000
Mean   : 23.17      Mean   :  2.613
3rd Qu.: 29.00      3rd Qu.:  0.000
Max.   :436.00      Max.   :231.000
```

Predictive Task

- Feature addition: We classified the values in ARRIVAL_DELAY greater than 20 mins as delayed flights and ARRIVAL_DELAY less than 20 min as non-delayed flights. Filled the column "Delayed" with Yes or No.
- Ran a cross-validation of count of 12 for each model

```
control <- trainControl(method="cv", number=12)
metric <- "Accuracy"
```

Predictive Models used

- Naïve bayes
  - Used this as a baseline for our analysis
- SVM
- Random Forest
- KNN
- Linear Regression

Naïve Bayes

```
model.naiveBayes <- naiveBayes(Delayed~., data=train, metric=metric, trControl=control)
prediction.NB <- predict(model.naiveBayes, test)
nba.acc=confusionMatrix(prediction.NB, test$Delayed)$overall[1]
confusionMatrix(prediction.NB, test$Delayed)
```

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No   28  14
       Yes   1 136

               Accuracy : 0.9162
                 95% CI : (0.8656, 0.9523)
    No Information Rate : 0.838
    P-Value [Acc > NIR] : 0.001677

                  Kappa : 0.7386
 Mcnemar's Test P-Value : 0.001946

            Sensitivity : 0.9655
            Specificity : 0.9067
         Pos Pred Value : 0.6667
         Neg Pred Value : 0.9927
             Prevalence : 0.1620
         Detection Rate : 0.1564
   Detection Prevalence : 0.2346
      Balanced Accuracy : 0.9361

       'Positive' Class : No
```

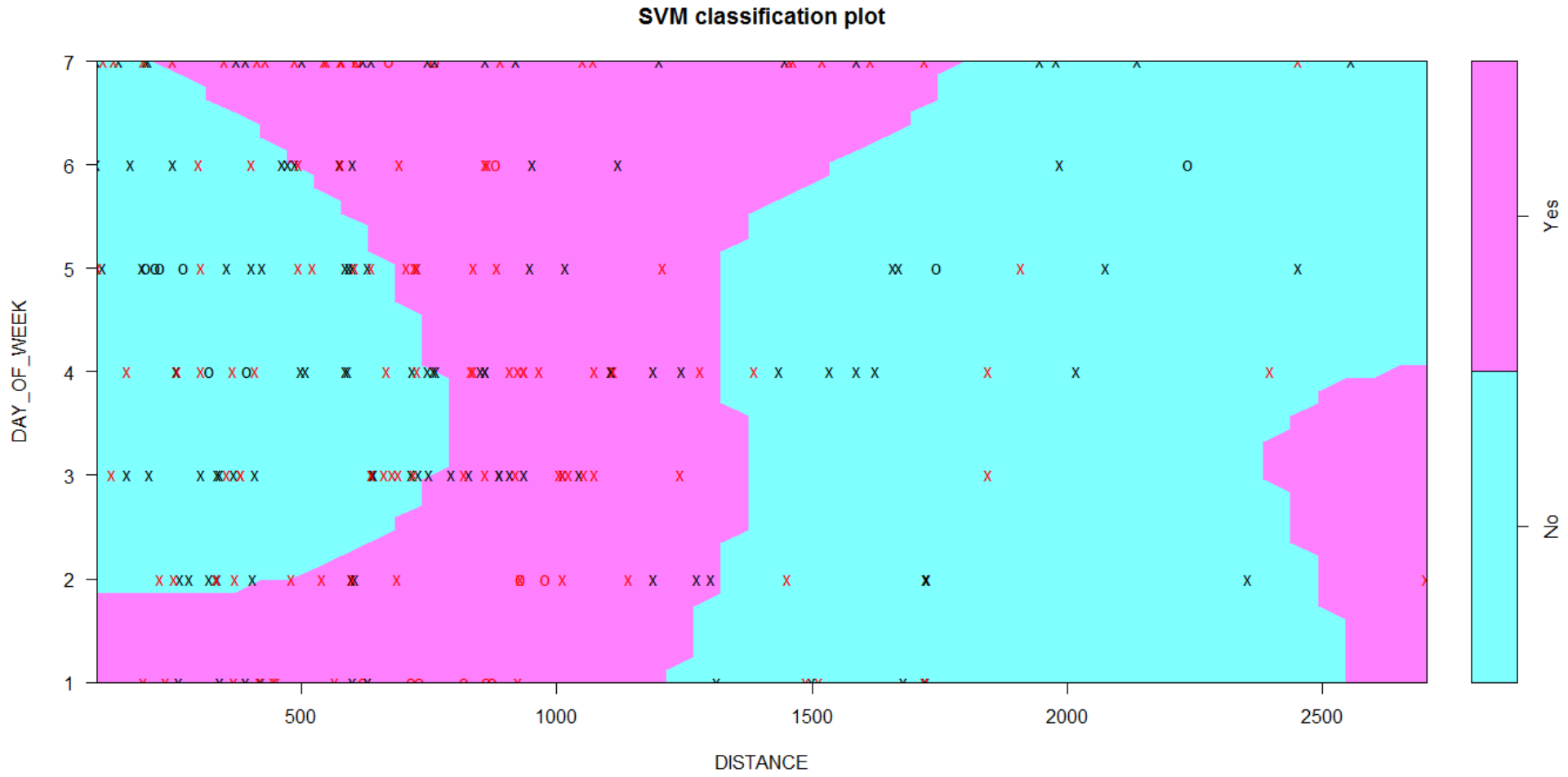
SVM

```
#SVM

model.svm <- train(Delayed~., data=train, method="svmRadial", metric=metric, trControl=control)
prediction.svm <- predict(model.svm, test)
confusionMatrix(prediction.svm, test$Delayed)

tuneResult <- tune(svm, Delayed ~. , data=train,
                   ranges = list(epsilon = seq(0,0.1,0.01), cost = 2^(2:9)))
tunedModel <- tuneResult$best.model
tunedModelPrediction <- predict(tunedModel, test)
confusionMatrix(tunedModelPrediction, test$Delayed)
```

```
> confusionMatrix(prediction.svm, test$Delayed)
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No    0   0
       Yes  29 150

               Accuracy : 0.838
                 95% CI : (0.7757, 0.8887)
    No Information Rate : 0.838
    P-Value [Acc > NIR] : 0.5494

                  Kappa : 0
 Mcnemar's Test P-Value : 1.999e-07

            Sensitivity : 0.000
            Specificity : 1.000
         Pos Pred Value :   NaN
         Neg Pred Value : 0.838
             Prevalence : 0.162
         Detection Rate : 0.000
   Detection Prevalence : 0.000
      Balanced Accuracy : 0.500

       'Positive' Class : No
```

```
> confusionMatrix(tunedModelPrediction, test$Delayed)
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No    0   0
       Yes  29 150

               Accuracy : 0.838
                 95% CI : (0.7757, 0.8887)
    No Information Rate : 0.838
    P-Value [Acc > NIR] : 0.5494

                  Kappa : 0
 Mcnemar's Test P-Value : 1.999e-07

            Sensitivity : 0.000
            Specificity : 1.000
         Pos Pred Value :   NaN
         Neg Pred Value : 0.838
             Prevalence : 0.162
         Detection Rate : 0.000
   Detection Prevalence : 0.000
      Balanced Accuracy : 0.500

       'Positive' Class : No
```

With this classification plot, we can understand the relation between Day of week and Distance. With this plot, we can see that Day of week 1, and distance from (0-1250) chances of the flight being delayed is more than others.
For the range distance of 750-1250, there is more chance of flight being delayed for the entire week.



Random Forest

```r
#RF
model.rf <- train(Delayed~., data=train, method="rf", metric=metric, trControl=control)
prediction.rf <- predict(model.rf, test)
confusionMatrix(prediction.rf, test$Delayed)
```

```
> confusionMatrix(prediction.rf, test$Delayed)
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No   29   0
       Yes   0 150

               Accuracy : 1
                 95% CI : (0.9796, 1)
    No Information Rate : 0.838
    P-Value [Acc > NIR] : 1.818e-14

                  Kappa : 1
 Mcnemar's Test P-Value : NA

            Sensitivity : 1.000
            Specificity : 1.000
         Pos Pred Value : 1.000
         Neg Pred Value : 1.000
             Prevalence : 0.162
         Detection Rate : 0.162
   Detection Prevalence : 0.162
      Balanced Accuracy : 1.000

       'Positive' Class : No
```

KNN

```
#KNN

model.knn <- train(Delayed~., data=train, method="knn", metric=metric, trControl=control)
prediction.knn <- predict(model.knn, test)
confusionMatrix(prediction.knn, test$Delayed)
```

```
> confusionMatrix(prediction.knn, test$Delayed)
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No    8   7
       Yes  21 143

               Accuracy : 0.8436
                 95% CI : (0.7819, 0.8935)
    No Information Rate : 0.838
    P-Value [Acc > NIR] : 0.46868

                  Kappa : 0.2846
 Mcnemar's Test P-Value : 0.01402

            Sensitivity : 0.27586
            Specificity : 0.95333
         Pos Pred Value : 0.53333
         Neg Pred Value : 0.87195
             Prevalence : 0.16201
         Detection Rate : 0.04469
   Detection Prevalence : 0.08380
      Balanced Accuracy : 0.61460

       'Positive' Class : No
```

Linear Regression

```
#Building various models for comparison
l_model1 = lm(ARRIVAL_DELAY ~. , data=flights_v01_train)
l_model2 = lm(ARRIVAL_DELAY ~DEPARTURE_DELAY+TAXI_IN , data=flights_v01_train)
l_model3 = lm(ARRIVAL_DELAY ~DEPARTURE_DELAY+TAXI_IN+SECURITY_DELAY , data=flights_v01_train)
l_model4 = lm(ARRIVAL_DELAY ~DEPARTURE_DELAY+TAXI_IN+SECURITY_DELAY+WEATHER_DELAY , data=flights_v01_train)
l_model5 = lm(ARRIVAL_DELAY ~DEPARTURE_DELAY+TAXI_IN+DISTANCE , data=flights_v01_train)

print(anova(l_model1, l_model2, l_model3, l_model4,l_model5))
```

```
> print(anova(l_model1, l_model2, l_model3, l_model4,l_model5))
Analysis of Variance Table

Model 1: ARRIVAL_DELAY ~ DAY_OF_WEEK + DEPARTURE_DELAY + TAXI_OUT + AIR_TIME +
    DISTANCE + TAXI_IN + CANCELLED + SECURITY_DELAY + AIRLINE_DELAY +
    LATE_AIRCRAFT_DELAY + WEATHER_DELAY
Model 2: ARRIVAL_DELAY ~ DEPARTURE_DELAY + TAXI_IN
Model 3: ARRIVAL_DELAY ~ DEPARTURE_DELAY + TAXI_IN + SECURITY_DELAY
Model 4: ARRIVAL_DELAY ~ DEPARTURE_DELAY + TAXI_IN + SECURITY_DELAY +
    WEATHER_DELAY
Model 5: ARRIVAL_DELAY ~ DEPARTURE_DELAY + TAXI_IN + DISTANCE
  Res.Df    RSS Df Sum of Sq        F    Pr(>F)
1    709  64973
2    717 239122 -8   -174149 237.5440 < 2.2e-16 ***
3    716 239045  1        77   0.8348 0.3611909
4    715 236826  1      2220  24.2211 1.069e-06 ***
5    716 238226 -1     -1400  15.2760 0.0001018 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  17  15
         1  11 137

               Accuracy : 0.8556
                 95% CI : (0.7956, 0.9034)
    No Information Rate : 0.8444
    P-Value [Acc > NIR] : 0.3872

                  Kappa : 0.4805
 Mcnemar's Test P-Value : 0.5563

            Sensitivity : 0.60714
            Specificity : 0.90132
         Pos Pred Value : 0.53125
         Neg Pred Value : 0.92568
             Prevalence : 0.15556
         Detection Rate : 0.09444
   Detection Prevalence : 0.17778
      Balanced Accuracy : 0.75423

       'Positive' Class : 0
```

Performance Comparison

Model	Accuracy	Precision	Recall	F1- score
Naïve Bayes	0.92	0.68	0.97	0.80
Random Forest	1.00	1.00	1.00	1.00
Support Vector Machine	0.84	NaN	0.0	NaN
k-NN	0.82	0.533	0.28	0.37
Linear Regression	0.86	0.53	0.61	0.58

Conclusion

- With regression model, we found several attributes to be statistically significant like Departure Delay, Taxi In, Taxi Out,etc.
- We found out relationships between variables which helped us conclude that Wednesday flights are more likely to be on time.
- With SVM we pattern recognition where we found that distance from 750-1250 means more chance of delay
- With our analysis, we can say that if we travel on weekdays, though air traffic is less we will reach on time

References

- DOT's Bureau of Transportation Statistics Dataset from kaggle.com
- Website: http://usatoday30.usatoday.com/travel/flights/2007-12-20-flight-delays_N.htm
- Michael Ball, Cynthia Barnhart, Martin Dresner, Mark Hansen, Kevin Neels, Amedeo Odoni, Everett Peterson, Lance Sherry, Antonio A Trani, and Bo Zou. 2010. Total delay impact study: a comprehensive assessment of the costs andimpacts of flight delay in the United States. (2010).
- Juan Jose Rebollo and Hamsa Balakrishnan. 2014. Characterization and prediction of air traffic delays. Transportation research part C: Emerging technologies 44(2014), 231–241.

Thank You!