# Flight Delay Analysis (Unsupervised Leaning)

Joseph O'Malley

https://github.com/ultrajoseph

# Defining the Problem

- Focusing on Kansas City (MCI)
- Targeting management at Southwest Airlines
- Conditions that lead to severe delays
- Most/least reliable groups of flights offered
- Benchmarking performance

# Data Description

- U.S. Department of Transportation (retrieved from Kaggle.com)
- 2015 flight data from “large US carriers”
- Collected by U.S. Department of Transportation
- 5.8 million records (subset to 77,320 – MCI Flights)
- 33 total feature columns: schedule departure time (of day), taxi time, month, etc.
- Appended Latitude/Longitude Data

# Model Choice

- Looking to explore interactions that lead to longer delays
- Expect time of year & time of day to have strong effects
- information can help them to schedule flights more strategically to avoid delays
- inform certain departure/arrival cities to target/avoid, flight time of day, flight time of year, regions to avoid, and an combination of these (and other) feature columns
- show competitor airlines that are performing better/worse in certain conditions

# Analysis Methods

# Analysis: EDA

- 53.6 percent of MCI flights on Southwest
- Highest travel month in July (10.3 percent)
- Saturday is lowest travel day (11 percent)
- Flight distance range of 152-1499 miles - median 643 miles

Figure 2.1 – Flights by Day of Week

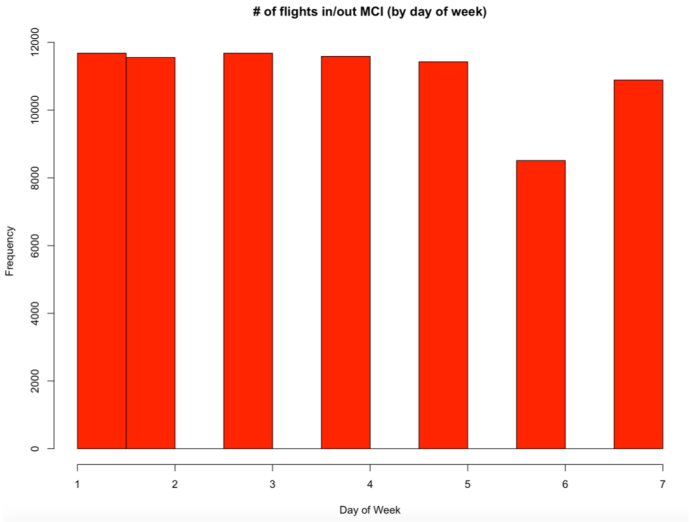
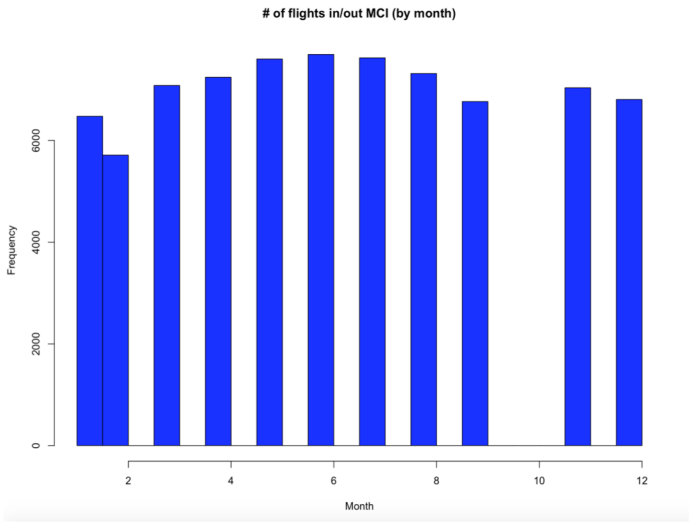



Figure 2.2 – Flights by Month



# Analysis: Correspondence Analysis

- Binned continuous variables
- Ran Chi-squared significance test

Figure 3.1 – Delays by Airline

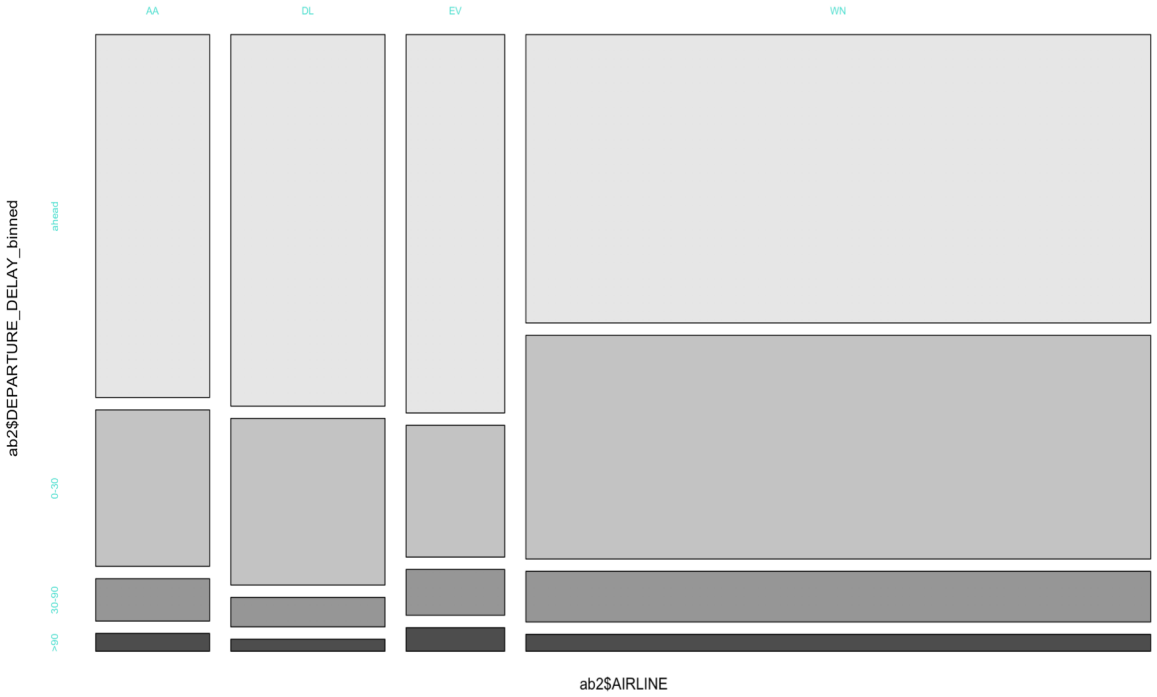
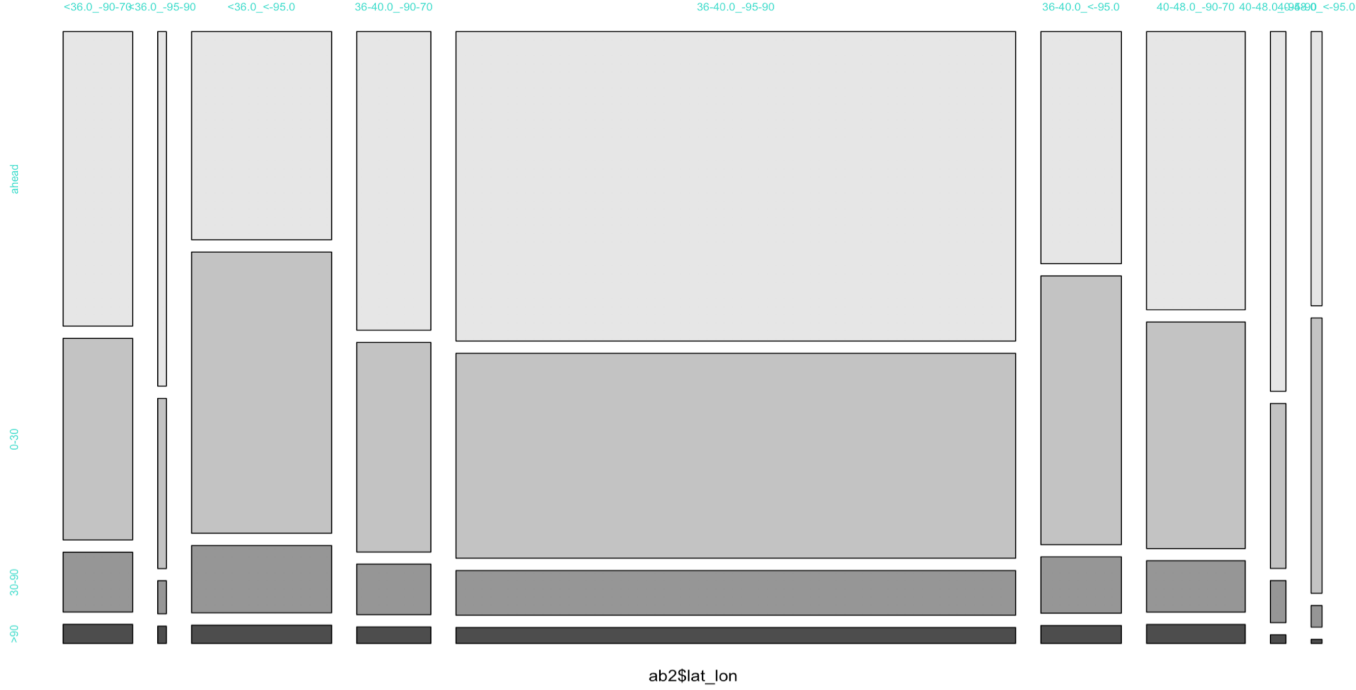



Figure 3.2 – Delays by Latitude/Longitude



# Findings: Correspondence Analysis

- top 4 airlines account for 84.8% of all flights
- EV (skywest) - serves as a regional connector for American, Delta, United, and Alaskan airline
- Southwest Airlines have more intermediate delays in the southwest United States (Lat/Lon)
- 2.9 percent of Southwest flights had “severe” delays…90minuted> (Delta had just 2.1 percent)

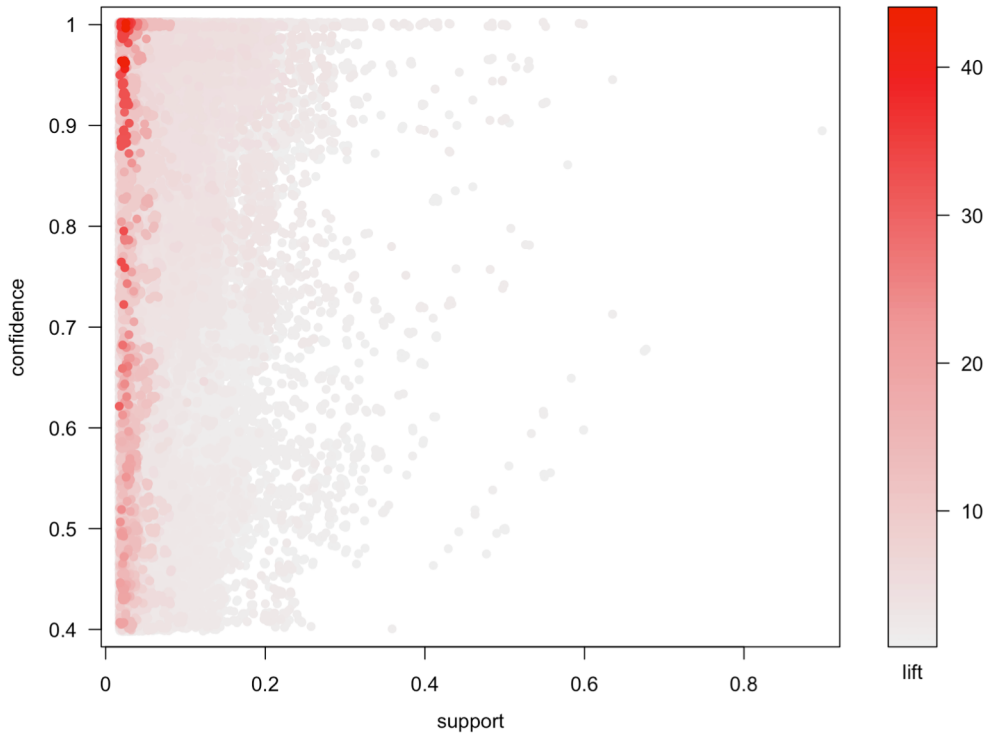
# Analysis: Association Rules

- Transformed Dataset
- Initial set contained 241,489 rules (figure 4.1)
- tested several values for lift, support, and confidence. There was a tradeoff for finding rules that had a large enough sample size to be generalizable and rules that gave us high lift (odds ratios)
- Set optimal parameters: apriori(seg.trans, parameter=list(support=0.02, conf=0.4,target="rules",maxlen=4))
- Looked at effects Lat/Lon on weather delays, day of the week, and causes of long delays (10+ minutes – example below)

```
weather_rules <- apriori(sw_flights_trans, parameter = list(support =
                                0.0002, confidence = 0.20, maxlen=4), appearance=list(lhs=c("WEATHER_DELAY_binned=10+")))


weather_rules_partial <- subset(weather_rules, items %pin% "LON" |items %pin% "LAT")
inspect(sort(weather_rules_partial[1:15], by = c("lift","count", "confidence")))
```

Figure 4.1 Scatter plot for 241489 rules



# Findings: Association Rules - Findings

- 13 percent more likely to be delayed 30-90 minutes on Thursday
- longer flights (1.16 lift) 2-3 hours on Saturdays
- 5.7x more likely to be on a delayed flight between 12p-9a
- South and West US more likely to get delayed for weather

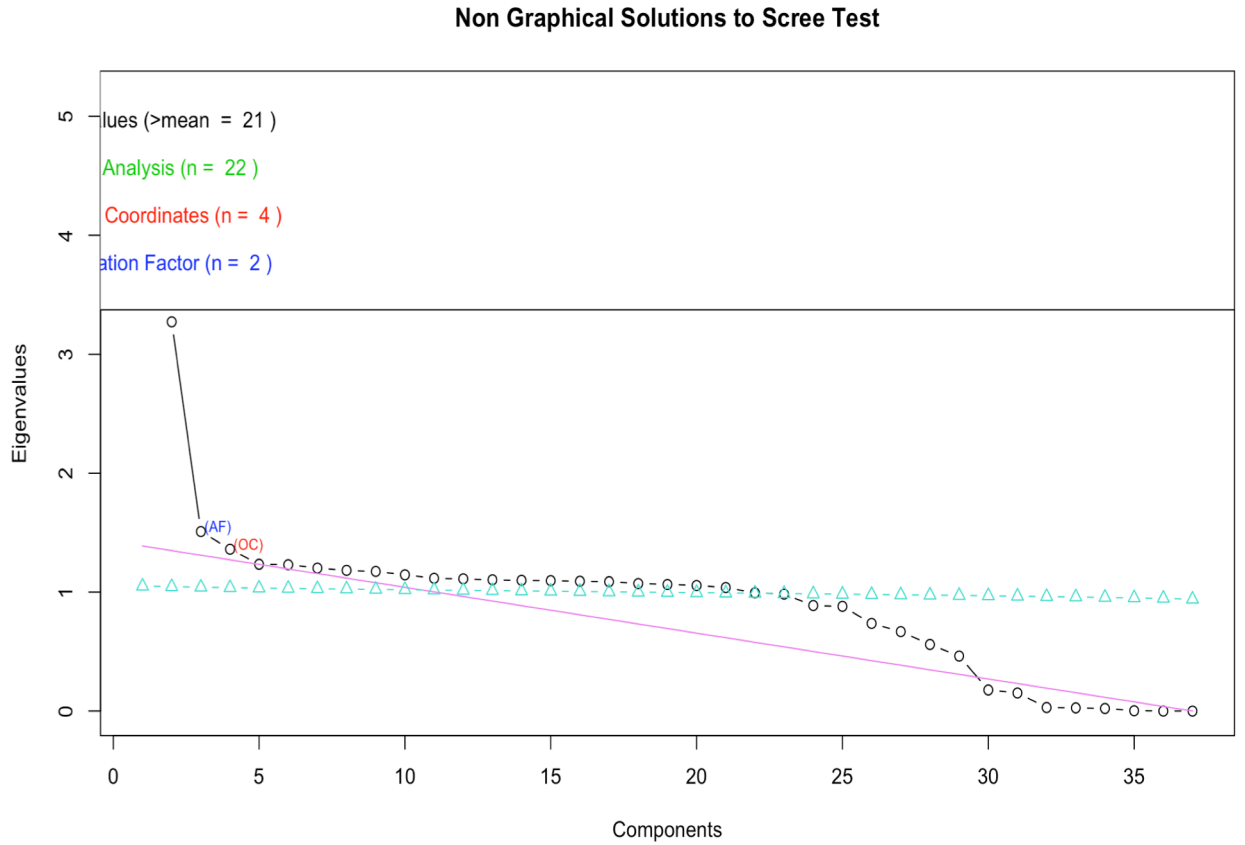
# Analysis: Factor Analysis

- One-hot encoded & scaled (z-score) data
- Top 3 eigenvalues – 5.2, 3.3, 1.5

Factor Loadings

| | MR1 | MR2 | MR3 | MR4 |
|---|---|---|---|---|
| SS loadings | 5.08 | 3.17 | 0.79 | 0.74 |
| Proportion Var | 0.14 | 0.09 | 0.02 | 0.02 |
| Cumulative Var | 0.14 | 0.22 | 0.24 | 0.26 |
| Proportion Explained | 0.52 | 0.32 | 0.08 | 0.08 |
| Cumulative Proportion | 0.52 | 0.84 | 0.92 | 1.00 |

- Chose 4 factors:
  - MR1 - "timing_vs_scheduled"
  - MR2 -"time_in_air_lat/lon"
  - MR3 - "delays_taxi_time"
  - MR4 - "factor_4"



# Analysis: Regression using 4 ”Factors”

- Dependent – ’departure_delay’ (continuous)
- ‘scheduled_vs_actual_timing’ had the highest positive weight (increases the delay the most)
- All factors ‘statistically significant’
- $R^2 = 0.17$

```
> summary(southwest_flights_fa.final$southwest_flights_fa_dependent)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-15.00   -3.00    0.00   10.15   10.00  495.00
```

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               10.16218    0.13871  73.263  < 2e-16 ***
scheduled_v_actual_timing  7.07116    0.13860  51.020  < 2e-16 ***
time_in_air_and_latlon     0.75872    0.13801   5.498 3.87e-08 ***
delays_and_taxi_time      -0.42638    0.02155 -19.783  < 2e-16 ***
MR4                      -11.11660    0.14575 -76.272  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.12 on 41106 degrees of freedom
Multiple R-squared:  0.1716,    Adjusted R-squared:  0.1715
F-statistic:  2129 on 4 and 41106 DF,  p-value: < 2.2e-16
```

# Conclusion: Key Findings

- Delta has less “severe” delays (0.8% less)
- South & West US more prone to delays
- More likely to have long delayed (30-90 minutes) on Thursday
- 5.7x more likely to be on a delayed flight between 12a-9a

# Recommendations

- Management should analyze Delta's staffing/policies
- Increase staff scheduled on Thursdays
- Look to areas other than the Southwest United States for expansion

# Limitations & Future State

- Granularity of cause of missed flight
- Pricing Data
- Additional Cities
- Interconnection of flights w/Network Analysis – (i.e. weather in a previous city)
- New Data (2018+)