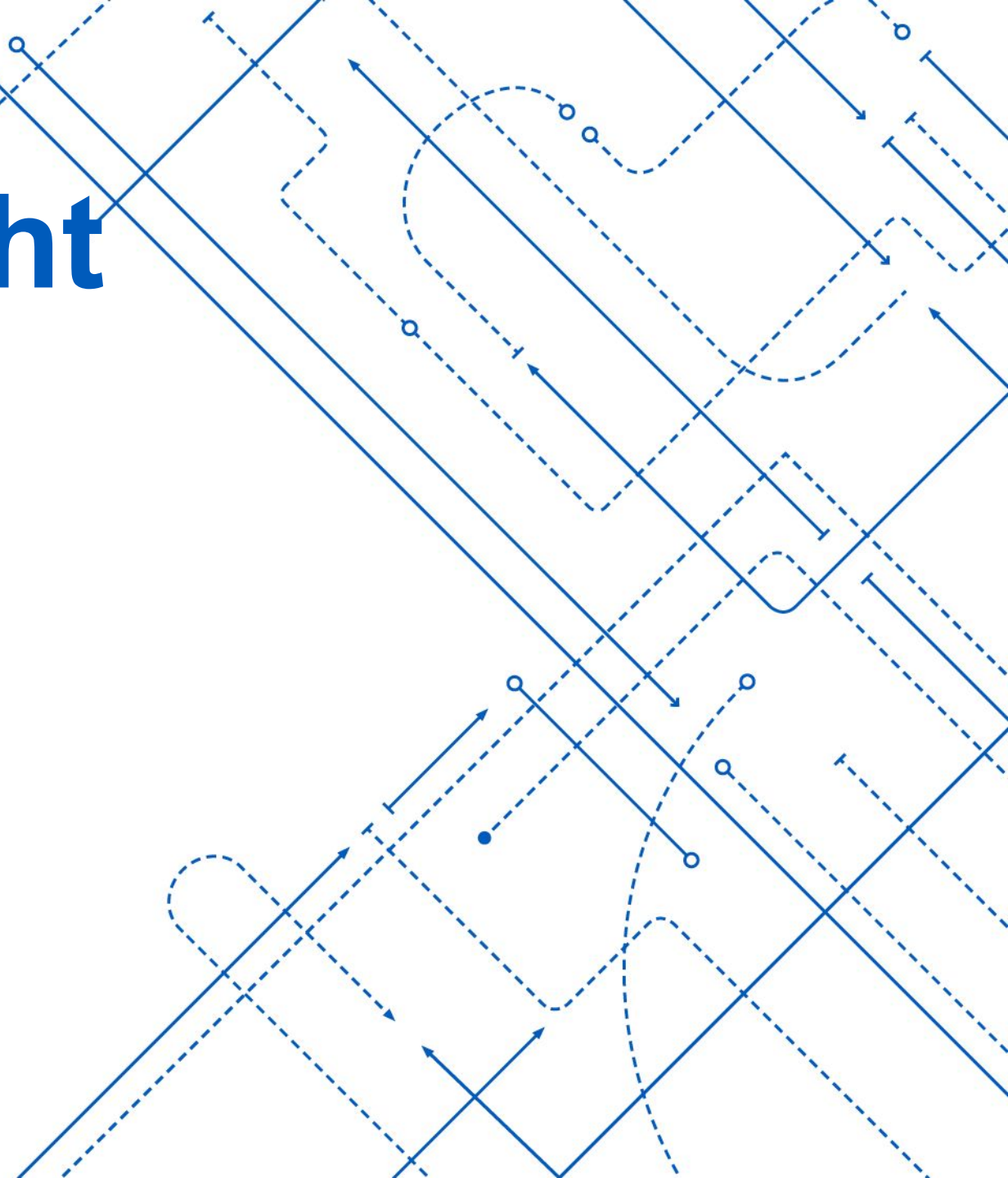
Analysis of Flight Delays

Group 12

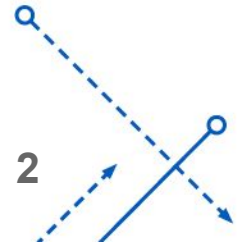
Amrutha Dokkadi
Ashutosh Rastogi
Tejas Prakash Agrawal
Vineeth Reddy Tati



University at Buffalo The State University of New York

Outline

- Motivation
- Data Description
- Methodology
- Results
- Conclusion
- References



Motivation

After a summer of flight delays and cancellations, federal officials issue an ultimatum to the airlines

NATIONAL

More than 4,000 flights were delayed as holiday travel spikes in the U.S.

Updated November 27, 2022 · 5:05 PM ET

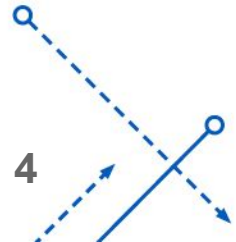
Flight delays cost $32.9 billion, passengers foot half the bill

Data Description

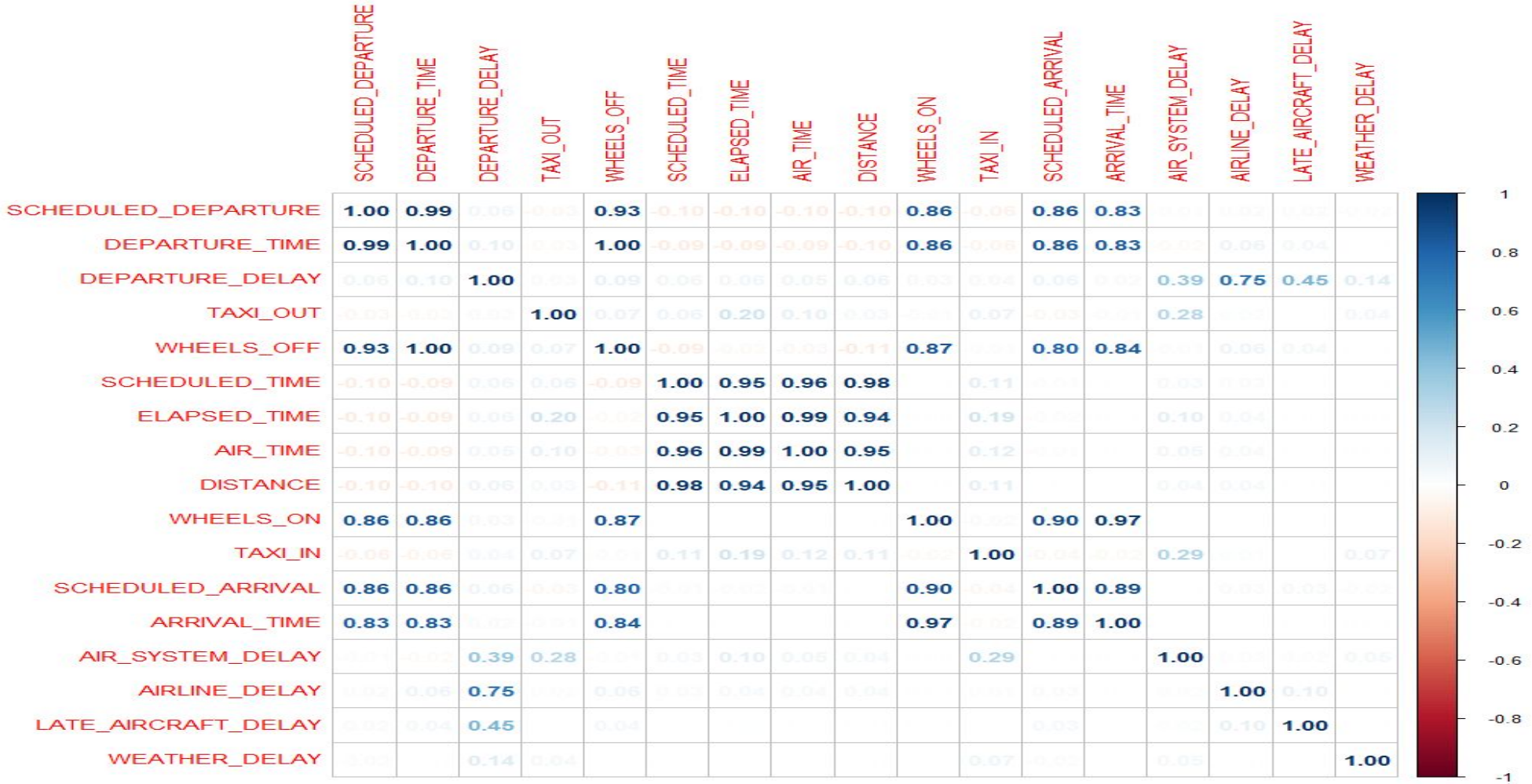
- Source: ***2015 Flight Delays and Cancellations*** from Kaggle
- Response: Arrival Delay
- Predictors: 30. Few of them are:

  Continuous: Departure Delay, Elapsed Time, Distance etc.

  Categorical: Airline Name, Flight Number, Origin Airport etc.
- Due to computation constraints, we used the subset of data points from the first 14 days for the month of February for American Airlines (AA) and Delta Airlines (DL) from their hubs at Dallas (DFW) and Atlanta (ATL) respectively for predictive modeling
- Dataset after preprocessing has the size 14284 observations and 22 predictors

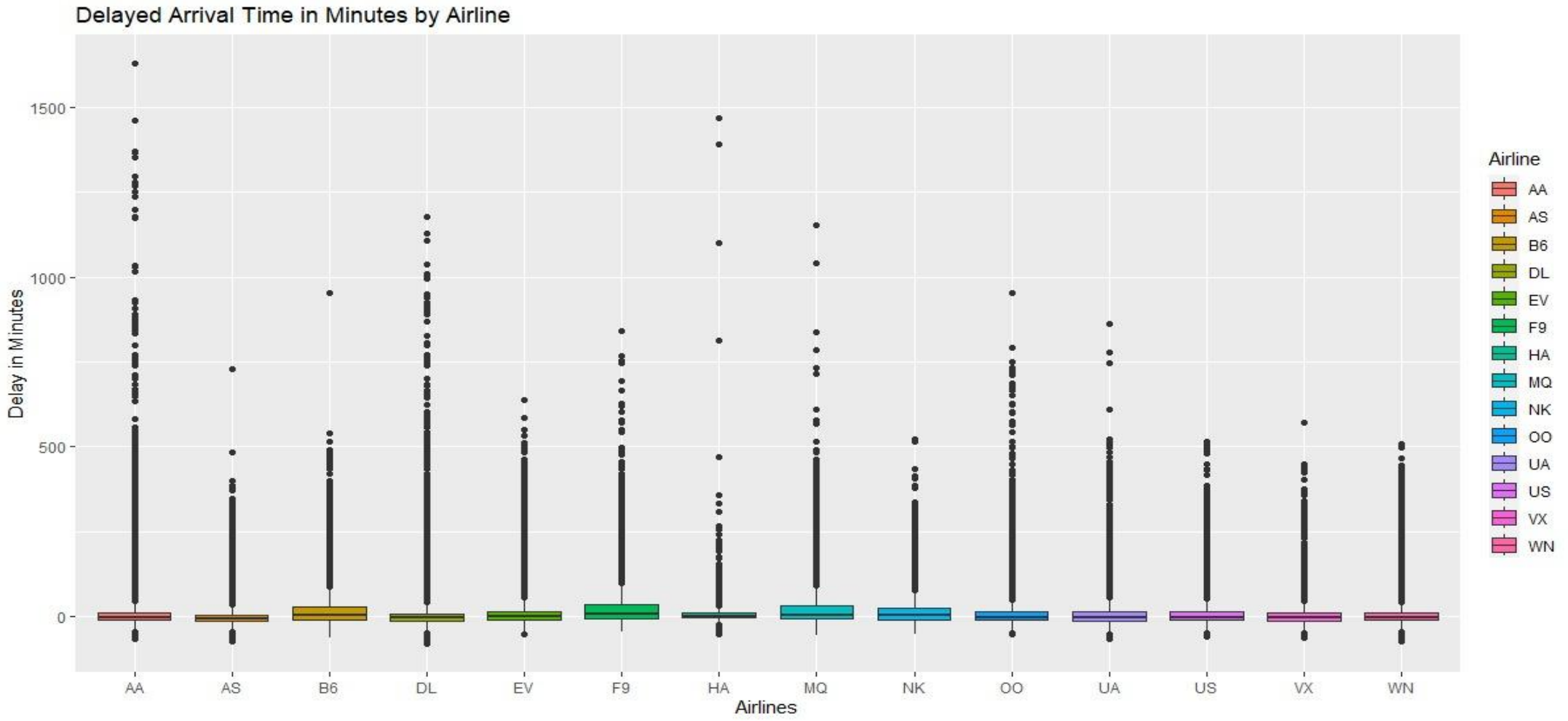


Data Description

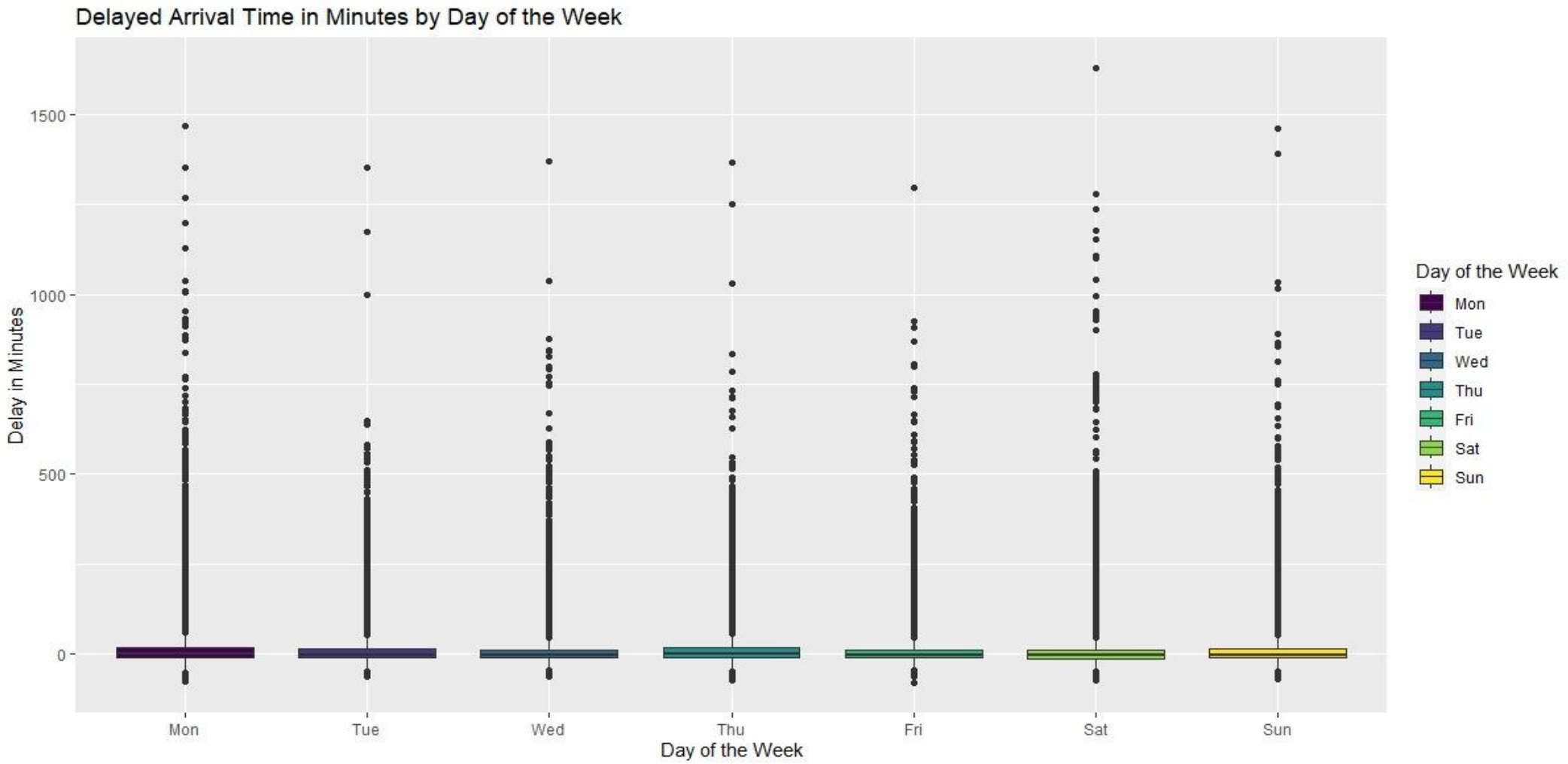


	SCHEDULED_DEPARTURE	DEPARTURE_TIME	DEPARTURE_DELAY	TAXI_OUT	WHEELS_OFF	SCHEDULED_TIME	ELAPSED_TIME	AIR_TIME	DISTANCE	WHEELS_ON	TAXI_IN	SCHEDULED_ARRIVAL	ARRIVAL_TIME	AIR_SYSTEM_DELAY	AIRLINE_DELAY	LATE_AIRCRAFT_DELAY	WEATHER_DELAY
SCHEDULED_DEPARTURE	1.00	0.99	0.06	-0.03	0.93	-0.10	-0.10	-0.10	-0.10	0.86	-0.06	0.86	0.83	0.01	0.02	0.03	-0.02
DEPARTURE_TIME	0.99	1.00	0.10	0.03	1.00	-0.09	-0.09	-0.09	-0.10	0.86	-0.06	0.86	0.83	0.02	0.06	0.04	
DEPARTURE_DELAY	0.06	0.10	1.00	0.03	0.09	0.06	0.06	0.05	0.06	0.03	0.04	0.06	0.05	0.39	0.75	0.45	0.14
TAXI_OUT	-0.03	-0.03	0.03	1.00	0.07	0.06	0.20	0.10	0.03	-0.01	0.07	-0.03	0.00	0.28	0.03	0.01	0.04
WHEELS_OFF	0.93	1.00	0.09	0.07	1.00	-0.09	-0.08	-0.08	-0.11	0.87	-0.05	0.80	0.84	0.01	0.06	0.04	
SCHEDULED_TIME	-0.10	-0.09	0.06	0.06	-0.09	1.00	0.95	0.96	0.98		0.11	-0.03	-0.03	0.03	0.03	0.01	
ELAPSED_TIME	-0.10	-0.09	0.06	0.20	-0.08	0.95	1.00	0.99	0.94		0.19	-0.03	-0.03	0.10	0.04	0.01	
AIR_TIME	-0.10	-0.09	0.05	0.10	-0.08	0.96	0.99	1.00	0.95		0.12	-0.03	-0.03	0.05	0.04	0.01	
DISTANCE	-0.10	-0.10	0.06	0.03	-0.11	0.98	0.94	0.95	1.00		0.11	-0.03	-0.03	0.04	0.04	0.01	
WHEELS_ON	0.86	0.86	0.03	-0.01	0.87					1.00	-0.05	0.90	0.97				
TAXI_IN	-0.06	-0.06	0.04	0.07	-0.05	0.11	0.19	0.12	0.11	-0.05	1.00	-0.04	-0.03	0.29	0.01		0.07
SCHEDULED_ARRIVAL	0.86	0.86	0.06	-0.03	0.80	-0.03	-0.03	-0.03		0.90	-0.04	1.00	0.89		0.03	0.03	0.01
ARRIVAL_TIME	0.83	0.83	0.05	-0.01	0.84					0.97	-0.03	0.89	1.00				
AIR_SYSTEM_DELAY	0.01	0.02	0.39	0.28	0.01	0.03	0.10	0.05	0.04	0.01	0.29	0.01	0.01	1.00	0.03	0.01	0.05
AIRLINE_DELAY	0.02	0.06	0.75	0.03	0.06	0.03	0.04	0.04	0.04	0.01	0.01	0.03	0.02	0.03	1.00	0.10	
LATE_AIRCRAFT_DELAY	0.03	0.04	0.45	0.01	0.04	0.01	0.01	0.01	0.01	0.01	0.01	0.03	0.02	0.01	0.10	1.00	
WEATHER_DELAY	-0.02	0.01	0.14	0.04	0.01	0.01	0.01	0.01	0.01	0.01	0.07	-0.02	0.01	0.05	0.01	0.01	1.00

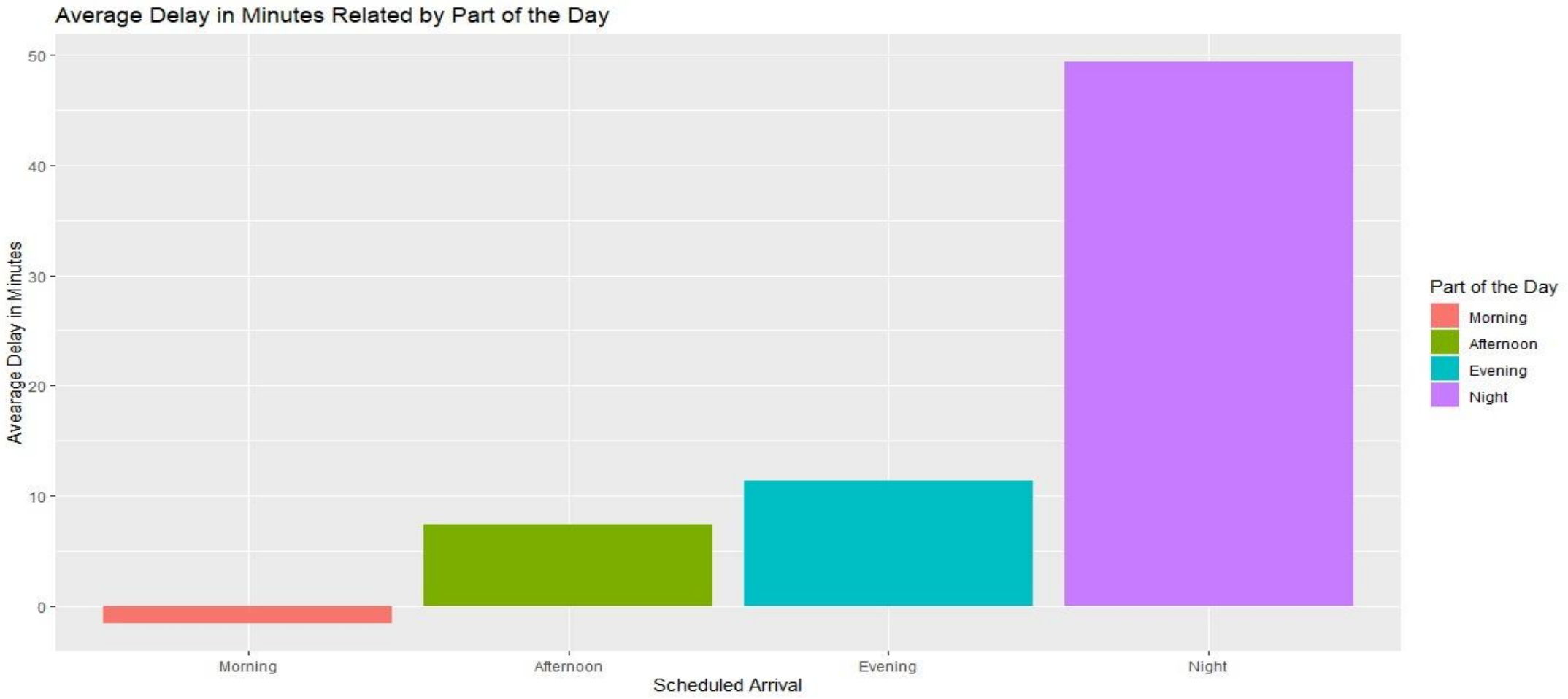
Data Description



Data Description



Data Description



Methodology

```
Call:
lm(formula = ARRIVAL_DELAY ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-44.116  -5.114   0.082   5.177  37.269

Coefficients: (4 not defined because of singularities)
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1.538e+01  6.531e-01 -23.553  < 2e-16 ***
TAXI_OUT             4.450e-01  1.375e-02  32.369  < 2e-16 ***
DISTANCE            -4.005e-03  1.679e-04 -23.860  < 2e-16 ***
WHEELS_ON            1.954e-03  1.689e-04  11.574  < 2e-16 ***
TAXI_IN              3.234e-01  1.601e-02  20.197  < 2e-16 ***
AIR_SYSTEM_DELAY     9.676e-01  7.504e-03 128.953  < 2e-16 ***
AIRLINE_DELAY        1.077e+00  5.779e-03 186.271  < 2e-16 ***
LATE_AIRCRAFT_DELAY  1.076e+00  1.042e-02 103.206  < 2e-16 ***
WEATHER_DELAY        1.002e+00  2.933e-02  34.180  < 2e-16 ***
DAY_OF_WEEKFri       2.052e+00  2.852e-01   7.195 6.61e-13 ***
DAY_OF_WEEKMon       3.408e+00  2.882e-01  11.824  < 2e-16 ***
DAY_OF_WEEKSat       7.720e-01  3.116e-01   2.478   0.0132 *
DAY_OF_WEEKSun       3.846e-01  2.945e-01   1.306   0.1916
DAY_OF_WEEKThu       1.426e+00  2.863e-01   4.979 6.47e-07 ***
DAY_OF_WEEKTue       6.867e-01  2.894e-01   2.373   0.0177 *
DAY_OF_WEEKWed              NA         NA      NA       NA
AIRLINEAA           -7.464e-02  5.056e-01  -0.148   0.8826
AIRLINEDL                   NA         NA      NA       NA
ORIGIN_AIRPORTATL   -2.725e+00  5.052e-01  -5.394 7.03e-08 ***
ORIGIN_AIRPORTDFW           NA         NA      NA       NA
DIVERTED                    NA         NA      NA       NA
SECURITY_DELAY       1.224e+00  2.210e-01   5.539 3.11e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.394 on 11409 degrees of freedom
Multiple R-squared:  0.876,     Adjusted R-squared:  0.8758
F-statistic:  4743 on 17 and 11409 DF,  p-value: < 2.2e-16
```

Linear Regression

Test RMSE is 8.430

Test R-Squared = 0.865

Methodology

```
22 x 1 sparse Matrix of class "dgCMatrix"
                                s1
(Intercept)         -15.339375005
TAXI_OUT              0.445929362
DISTANCE             -0.003654689
WHEELS_ON             0.001842941
TAXI_IN               0.340050276
AIR_SYSTEM_DELAY      0.906309293
AIRLINE_DELAY         1.012544049
LATE_AIRCRAFT_DELAY   1.020180519
WEATHER_DELAY         0.945799489
DAY_OF_WEEKFri        0.714825313
DAY_OF_WEEKMon        2.214833940
DAY_OF_WEEKSat       -0.529469344
DAY_OF_WEEKSun       -0.835526730
DAY_OF_WEEKThu        0.185208421
DAY_OF_WEEKTue       -0.628493727
DAY_OF_WEEKWed       -1.345770695
AIRLINEAA             0.209298120
AIRLINEDL            -0.260039750
ORIGIN_AIRPORTATL    -1.111499099
ORIGIN_AIRPORTDFW     1.059874939
DIVERTED              .
SECURITY_DELAY        1.141173708
```

Ridge Regression

Cross Validation: 10-Fold

Optimal Lambda: 1.5872

Test RMSE: 8.533

Test R-Squared: 0.866

```
22 x 1 sparse Matrix of class "dgCMatrix"
                                s1
(Intercept)         -14.554590431
TAXI_OUT              0.438706144
DISTANCE             -0.003883741
WHEELS_ON             0.001870127
TAXI_IN               0.316748384
AIR_SYSTEM_DELAY      0.966394673
AIRLINE_DELAY         1.074325610
LATE_AIRCRAFT_DELAY   1.071778918
WEATHER_DELAY         0.990916338
DAY_OF_WEEKFri        1.189710627
DAY_OF_WEEKMon        2.565396682
DAY_OF_WEEKSat        .
DAY_OF_WEEKSun       -0.291918993
DAY_OF_WEEKThu        0.576172821
DAY_OF_WEEKTue        .
DAY_OF_WEEKWed       -0.687530018
AIRLINEAA             .
AIRLINEDL             .
ORIGIN_AIRPORTATL    -2.539051077
ORIGIN_AIRPORTDFW     .
DIVERTED              .
SECURITY_DELAY        1.119393780
```

Lasso Regression

Cross Validation: 10-Fold

Optimal Lambda: 0.0375

Test RMSE: 8.491

Test R-Squared: 0.864

Methodology

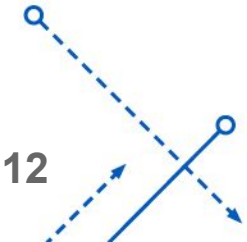
Random Forest Regression

Tress: 1000

Sampled Predictors at Each Split: 5

Test RMSE: 9.298

Test R-Squared: 0.8512



Methodology

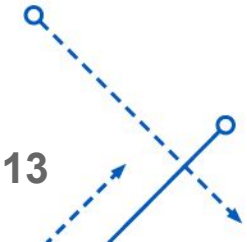
Support Vector Machine Regression

Kernel: Linear

Optimal Cost: 5.573

Test RMSE: 9.573

Test R-Squared: 0.8467

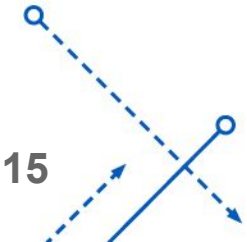


Results

Models	Test RSME	Test R-Squared
Linear Regression	8.430	0.8658
Ridge Regression	8.533	0.8661
Lasso Regression	8.491	0.8643
Random Forest	9.298	0.8512
Support Vector Machines	9.573	0.8467

Conclusion

- Linear Models i.e, Simple, Ridge and Lasso Regression have slightly better prediction performance compared to Random Forest and Support Vector Machine Regression
- Data preprocessing have significant influence on the prediction performance.
- Non-parametric models may have performed better with other pre-processing criteria



References

- Dataset:

  https://www.kaggle.com/datasets/usdot/flight-delays?select=flights.csv

- Analysis:

  https://nycdatascience.com/blog/r/flight-delays-r-shiny/

- Books:
  - An Introduction to Statistical Learning: with Applications in R. Second Edition.
  - The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition.

THANK YOU