AVIATION DELAY TIMES

STA 325 Final Project:

Calleigh Smith, Hannah Bogomilsky, Hugh Esterson, Maria Henriquez & Mariana Izon

01

EXECUTIVE SUMMARY

PROJECT OBJECTIVES

- **Motivations:**
  - US aviation as growing transportation method
    - 2019: 925.5M passengers (4.1% increase)
  - Number one complaint: delayed flights
- **Project Goals:**
  - Understand the market for US airline industry
  - Be able to improve upon airline arrival times to improve customer satisfaction
  - Allow airports to better plan for unexpected delays
- **Approach:**
  - Analyze flight data from January 2020
    - JFK to California airports (team members' hometown airports)
  - Determine which variables are significant in predicting arrival delays
  - Predict arrival delays with high accuracy

EMPHASIS ON PREDICTION

- **Industry Factors**
  - Customer satisfaction relies on an airline's ability to get clients where they need to be on time
    - Airline industry is competitive
  - Issues arise when flights are delayed
    - Missed connections → greater internal pressure for airlines/risk of losing revenue
    - Logistics as optimization
- **Incentives**
  - Being able to predict delays can lead to more accurate arrival times
    - Trickle-down effects for all parties involved
  - *Still want to understand how variables affect delays*

PROJECT SUMMARY

- **Data Cleaning**
  - Huge dataset of all US domestic flights from January 2020
  - Focus on a specific route(s)
  - Select relevant variables
- **Modeling**
  - General → Complex
    - Linear regression → GAM → Trees
  - Emphasis on prediction but inference is important as well
- **Prediction**
  - Test error metrics (80-20 division of training-test sets)
  - Cross-validation throughout to corroborate decisions

MODELING STRATEGIES

- **Multivariate Linear Model:**
  - Determine interaction effects through ANOVA
  - Correct degree of variables through CV
  - Check diagnostics
- **Generalized Additive Model:**
  - Understand *effects* of significant individual variables
- **Regression Tree:**
  - Interpretable and generally good for prediction
  - Test out random forest and boosting
    - Choose the best method and tuning parameters via cross-validation

MODEL PROGRESSION

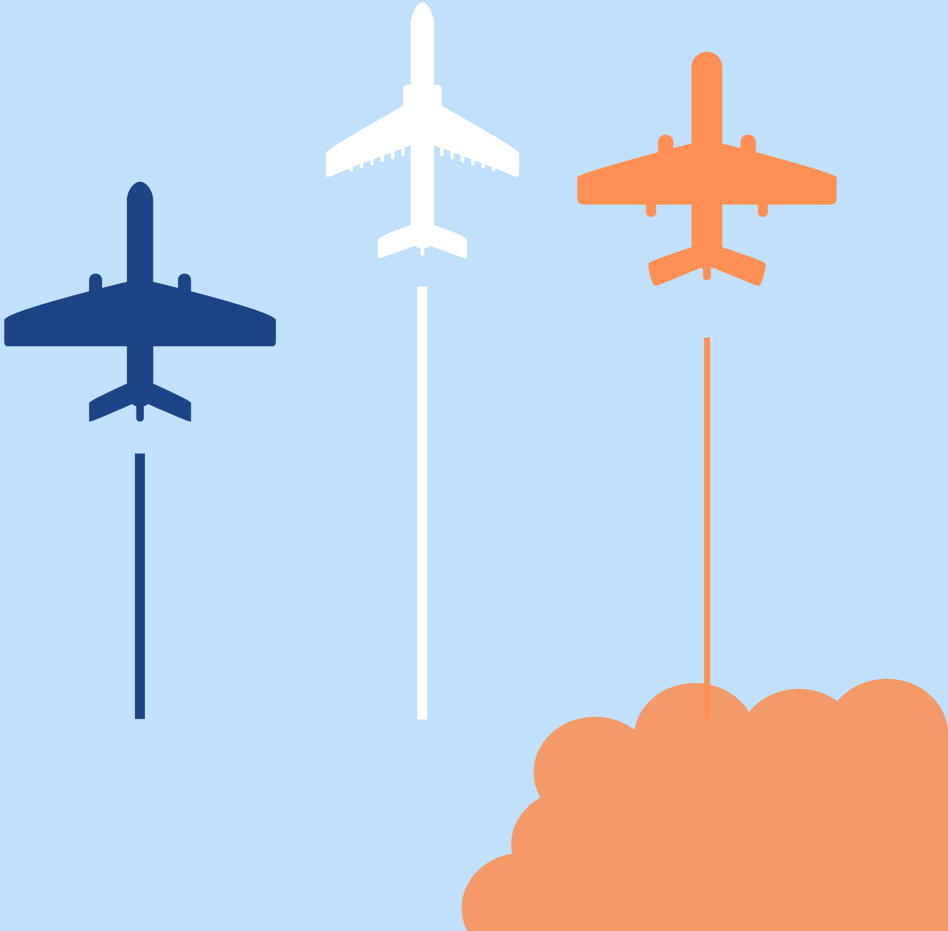
START SIMPLE
Check for underlying relationships
MLR

GAM
BUILD
Understand variables and their effects

TOWARDS CUTTING-EDGE
Improve prediction and inference
TREES

02

DATA DESCRIPTION



DATA BREAKDOWN

- **Source**
  - US Department of Transportation
    - Bureau of Transportation Statistics
  - *Reporting Carrier On-Time Performance*
    - Data bank of flight statistics, per month, since 1987
- **Variables of interest**
  - Time-based: *DayOfWeek*, *DayofMonth*
  - Route-based: *Origin*, *Dest*
  - Flight-based: *Reporting_Airline*, *TaxiOut*, *TaxiIn*, *DepDelay*
  - Delay indicators: *Carrier*, *Weather*, *NAS*, *Security*, *LateAircraft*
- **Data prep**
  - Download as CSV
  - Cleaned externally in Excel / Numbers
  - Filtered in R with ***dplyr***
  - Created indicator variables to replace time-additive variables

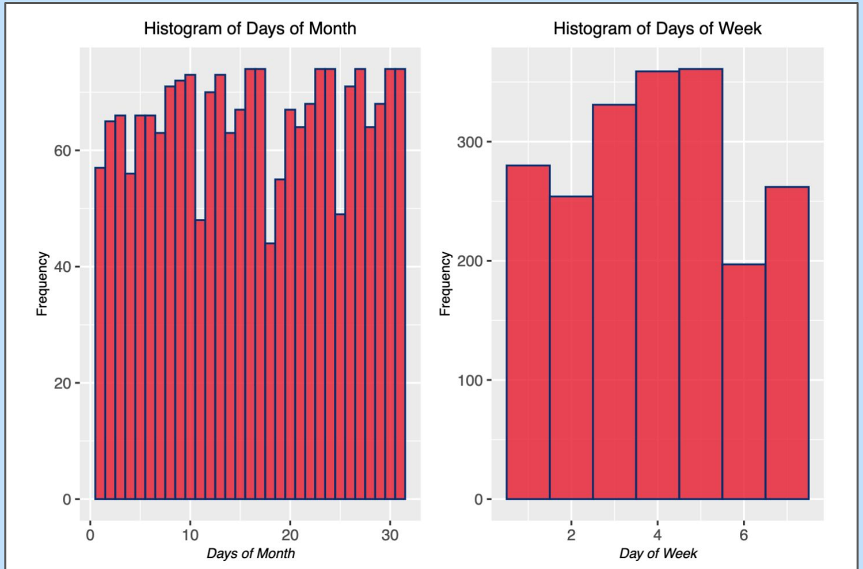
DATA DICTIONARY

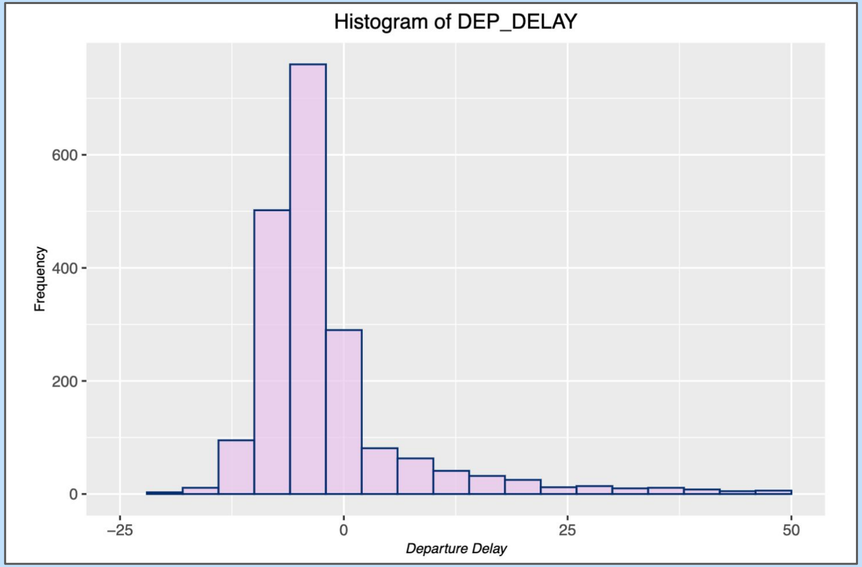
Table 1: Data Dictionary

Variables	Type	Description
General Flight Variables		
DAY_OF_MONTH	numeric	flight's day of week; Monday (1), Tuesday (2), ..., Sunday (7)
DAY_OF_WEEK	numeric	flight's day of month
OP_CARRIER	factor	airline providing flight; American (AA), Delta (DL), Alaska Airlines (AS), JetBlue (B6)
TYPE_DELAY	factor	classifaction type of delay; weather, National Air System, security, late aircraft
Departure-Based Variables		
ORIGIN	factor	flight's origin airport code; all JFK
CRS_DEP_TIME	numeric	Computerized Reservation System/scheduled time of departure; reported in military time, e.g. 7:30pm as 1930
DEP_TIME	numeric	flight's actual time of departure
DEP_DELAY	numeric	difference in flight's scheduled and actual time of departure; negative values indicate an early departure
TAXI_OUT	numeric	time duration from gate pushback to takeoff upon departure
Arrival-Based Variables		
DEST	factor	flight's destination airport code; SFO or LAX
CRS_ARR_TIME	numeric	Computerized Reservation System/scheduled time of arrival
ARR_TIME	numeric	flight's actual time of arrival
ARR_DELAY	numeric	difference in flight's scheduled and actual time of arrival; negative values indicate an early departure
TAXI_IN	numeric	time duration from landing to gate parking upon arrival

EXPLORATORY DATA ANALYSIS: INITIAL

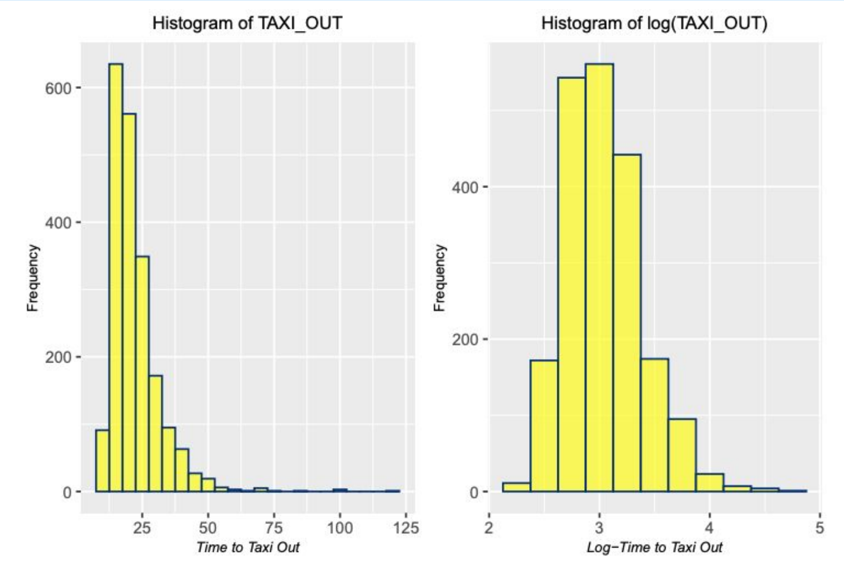
- Quick stats:
  - **2044 flights** with originally **34 variables**
  - **Carriers**: 4 included - American, Delta, JetBlue, Alaska
  - **Destinations**: 10 California destinations from JFK origin
  - **Delay Cause**: relatively few delays, with **NAS delays** as most common
    - National Air System delay: weather, airport operations, ATC

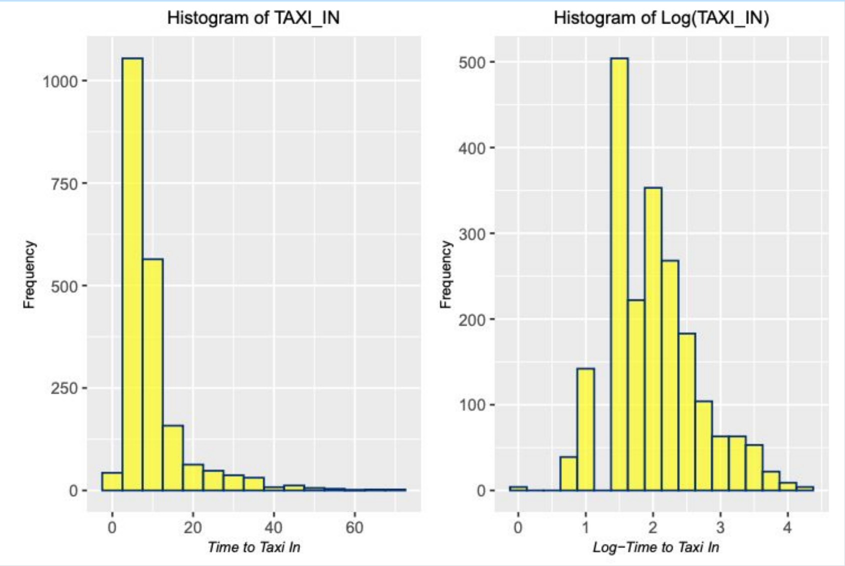




EXPLORATORY DATA ANALYSIS: FINAL

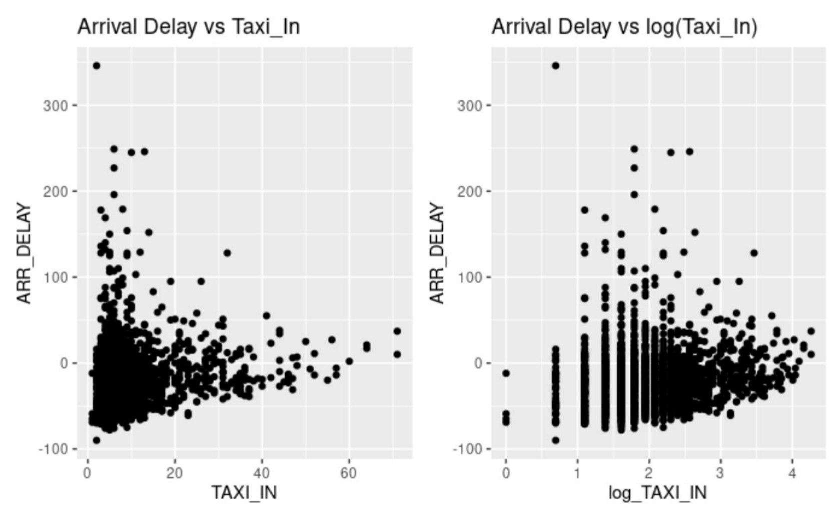
- Final data cleaning:
  - **80-20** training-test set split
  - **Carriers**: 4 included - American, Delta, JetBlue, Alaska
  - **Destinations**: 2 California destinations (SFO and LAX) from JFK origin
  - **Some transformations:** log-transformations on predictors; Box-Cox on response



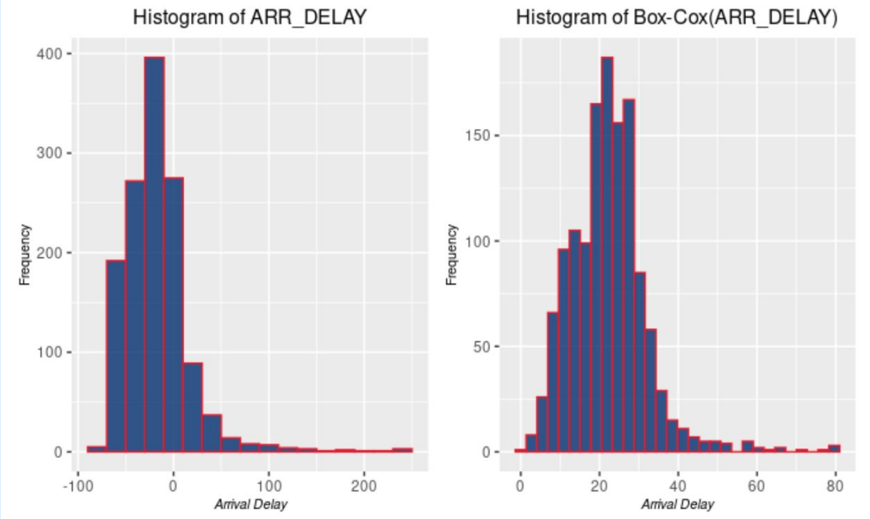


EXPLORATORY DATA ANALYSIS: CLEANED

Non-Linearity in Response vs. Predictors



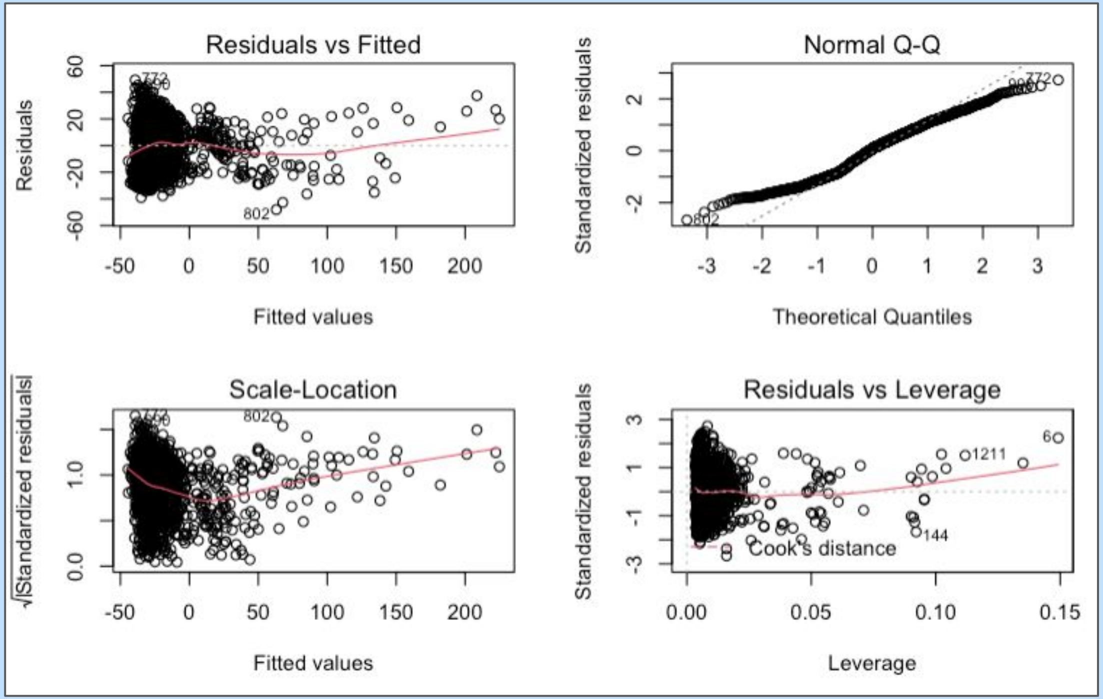
Box-Cox Transformation on Response



03

MULTIPLE LINEAR REGRESSION

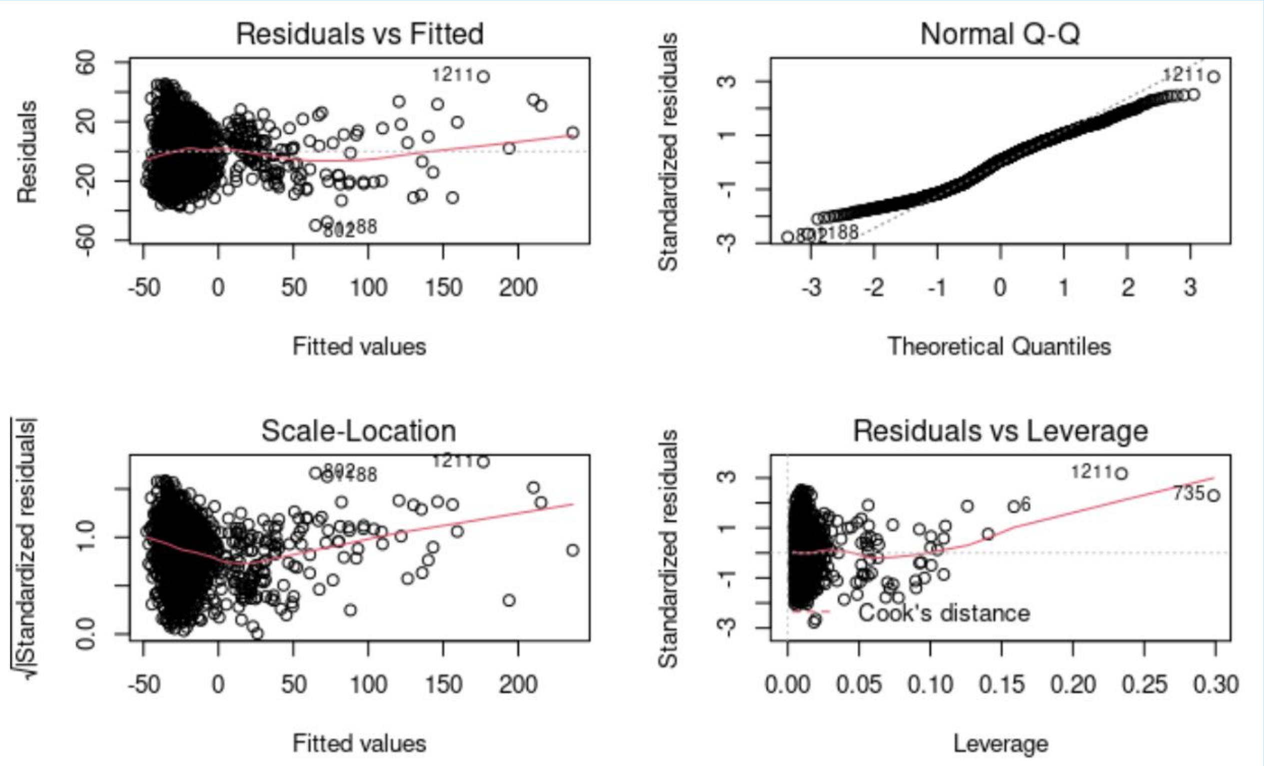
MLR: ATTEMPT 1



Baseline Model

- Performed model selection using AIC to get rid of insignificant predictors
- No interactions or transformations to variables

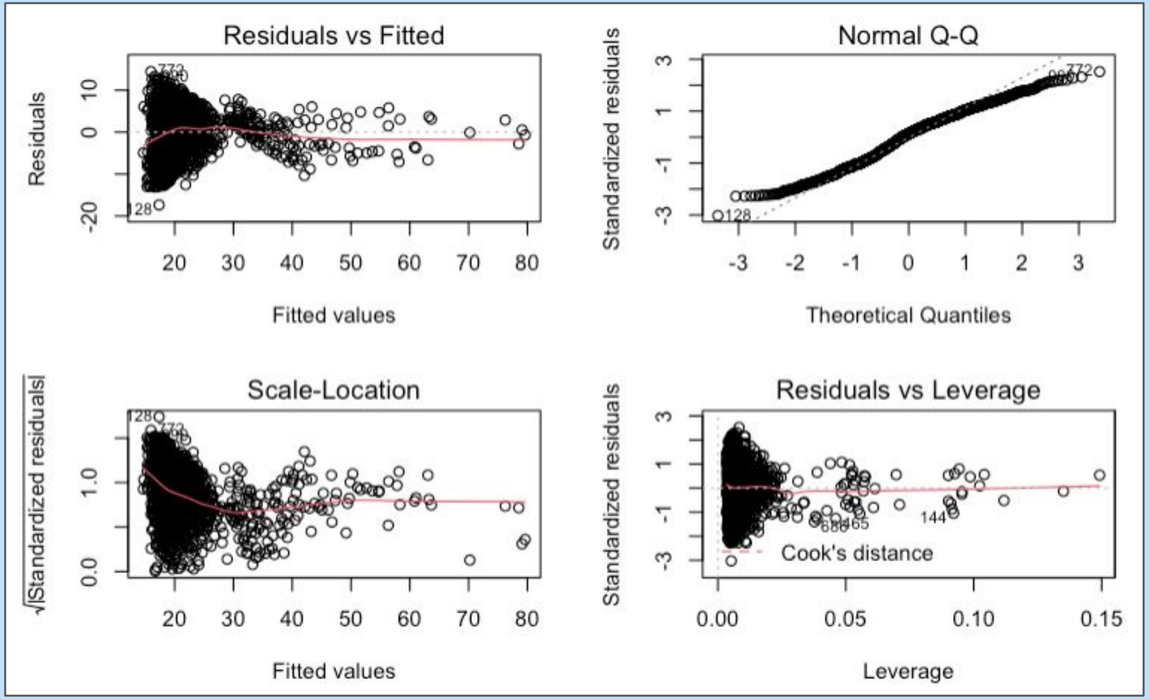
MLR: ATTEMPT 2



Log-transformed predictors model

MLR: ATTEMPT 3

Box-Cox transformed response model



MLR: PROS AND CONS

- **Pros**
  - Yields relatively interpretable models
  - Computationally inexpensive to implement
  - No hyperparameter tuning
- **Cons**
  - High test error
  - Evidence of non-linearity in data

MLR: ERROR TABLE

Model Name	Model MSE
Baseline Linear	322.46
Selected Linear w/ Log-Transformed Predictors	333.90
Selected Linear w/ Box-Cox	334.92

$$\begin{aligned}\widehat{ARR\_DELAY} = {} & -24.10 + 0.87(DEP\_DELAY) - 1.57(OP\_CARRIER(AS)) + 1.92(OP\_CARRIER(B6)) \\ & - 2.30(OP\_CARRIER(DL)) - 1.83(DEST(SFO)) - 0.004(CRS\_DEP\_TIME) - 0.002(CRS\_ARR\_TIME) \\ & + 0.87(TAXI\_OUT) + 0.47(TAXI\_IN) - 2.22(TYPE\_DELAY(LATE\_AIRCRAFT)) \\ & + 25.09(TYPE\_DELAY(NAS)) - 13.60(TYPE\_DELAY(No\ Delay))\end{aligned}$$

04

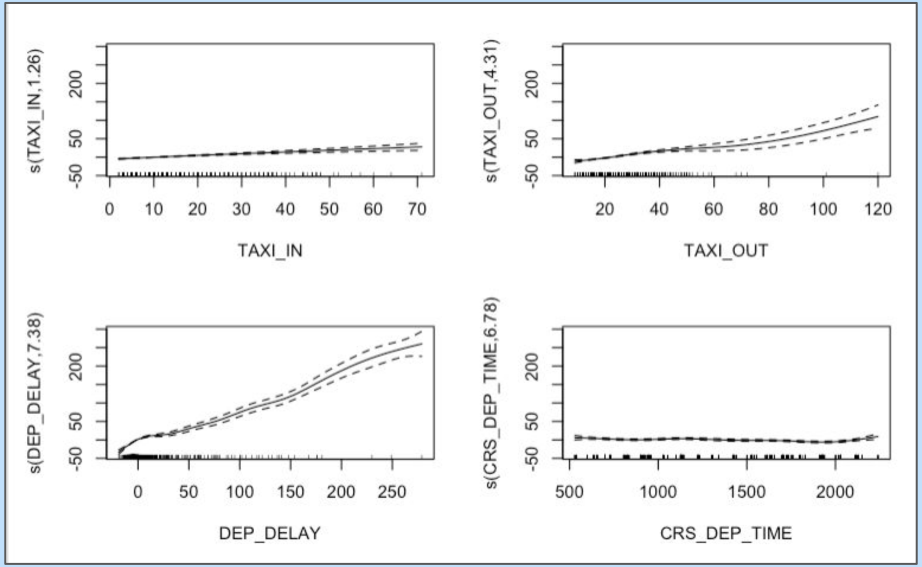
GENERAL ADDITIVE MODELING

GAM: ATTEMPT 1

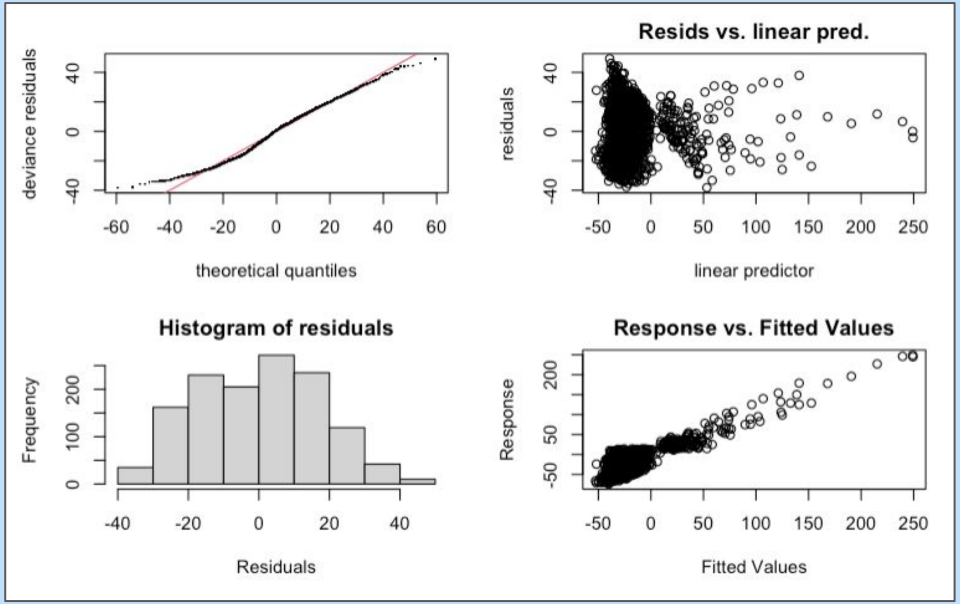
Original Response Model

- ANOVA to check linearity on TAXI_IN
  - Smoothing spline performs better
- ANOVA to remove insignificant predictors

Numerical variables with cubic smoothing splines



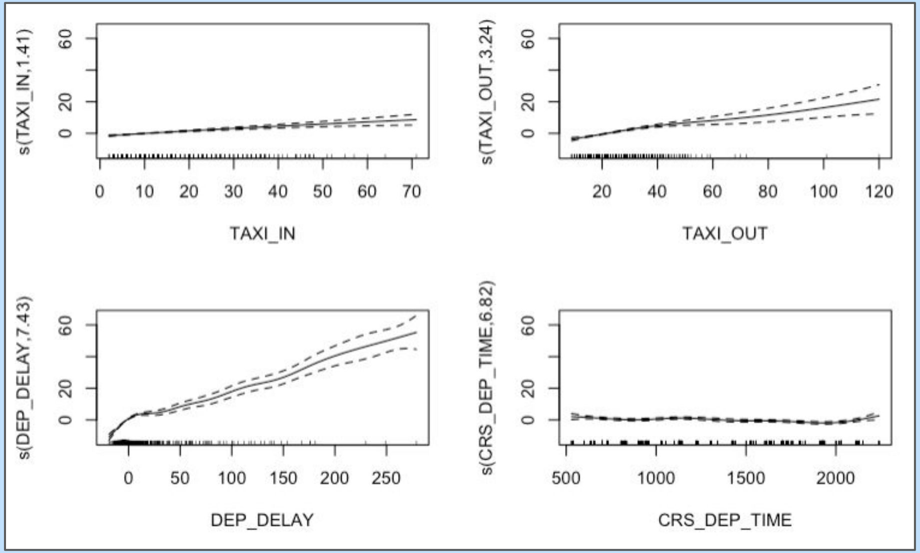
Model Diagnostics



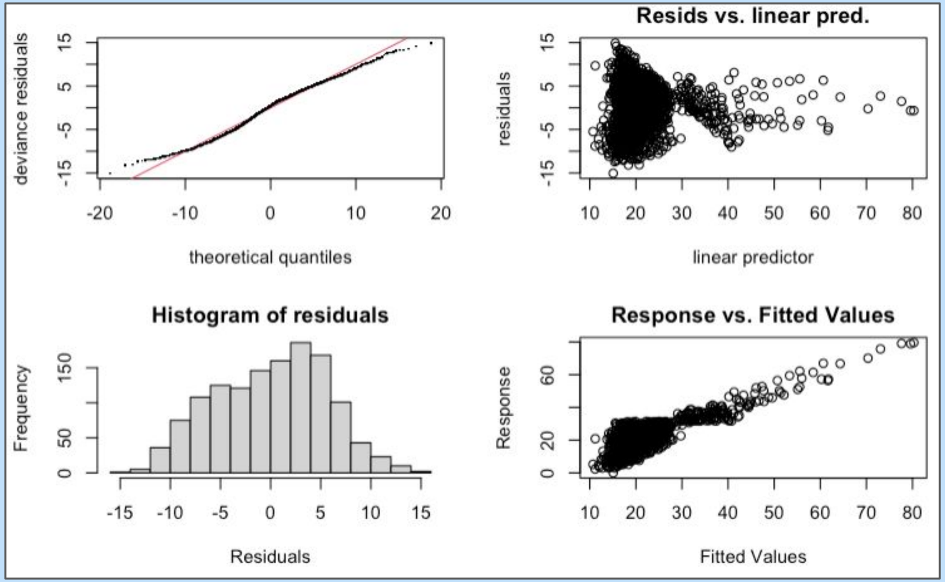
GAM: ATTEMPT 2

Box-Cox Transformation Model

Numerical variables with cubic smoothing splines



Model Diagnostics



GAM: PROS AND CONS

- **Pros**
  - Relatively computationally inexpensive
  - Good for inference
  - Has the ability to model highly complex nonlinear relationships
- **Cons**
  - Somewhat high test error
  - Could be potentially overfitting

GAM: ERROR TABLE

Model Name	Model MSE
GAM	**312.30**
GAM w/ Box-Cox	317.45

$$\widehat{ARR\_DELAY} = 1.83 - 1.68(OP\_CARRIER(AS)) + 2.49(OP\_CARRIER(B6)) - 3.14(OP\_CARRIER(DL)) - 3.20(TYPE\_DELAY(LATE\_AIRCRAFT)) + 18.80(TYPE\_DELAY(NAS)) - 22.41(TYPE\_DELAY(No\ Delay)) + s(TAXI\_IN) + s(TAXI\_OUT) + s(DEP\_DELAY) + s(CRS\_DEP\_TIME)$$

05

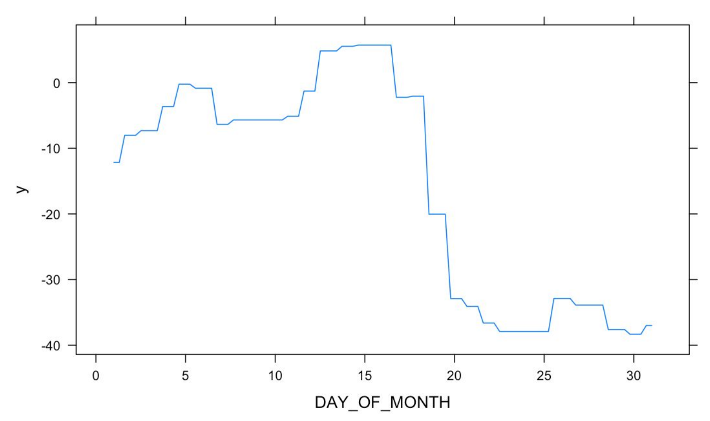
TREE-BASED REGRESSION

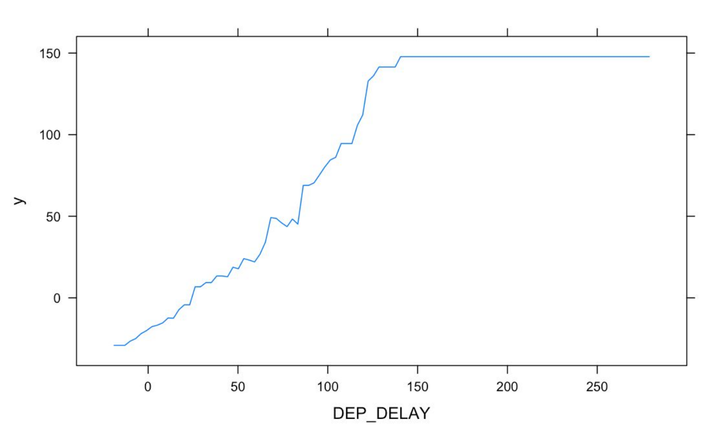
TREES: RANDOM FOREST

- Performs well when underlying relationships are **non-linear**
- Parameters to tune through **cross-validation**
  - Number of predictors sampled to build each tree → 2
  - Number of trees → 10,000

TREES: BOOSTING

- Parameters to tune through cross-validation
  - Shrinkage/learning rate → 0.1
  - Number of trees → 150
  - Interaction depth → 3
- Importance Metrics
  - DEP_DELAY most significant (understandably)
  - **DAY_OF_MONTH**
  - NAS_DELAY
  - TAXI variables





TREES: PROS AND CONS

- **Pros**
  - Performs well when true relationship is non-linear
  - Model gives some insight into which variables are most important in predicting the response
  - Increased predictive performance compared to our other models
- **Cons**
  - Possibility of overfitting
  - Random Forests and boosting can sometimes yield results that are difficult to interpret
  - Not as “plug-and-chug” compared to other models

TREES: ERROR TABLE

Model Name	Model MSE
Random Forest	155.01
Boosting	**129.80**

06

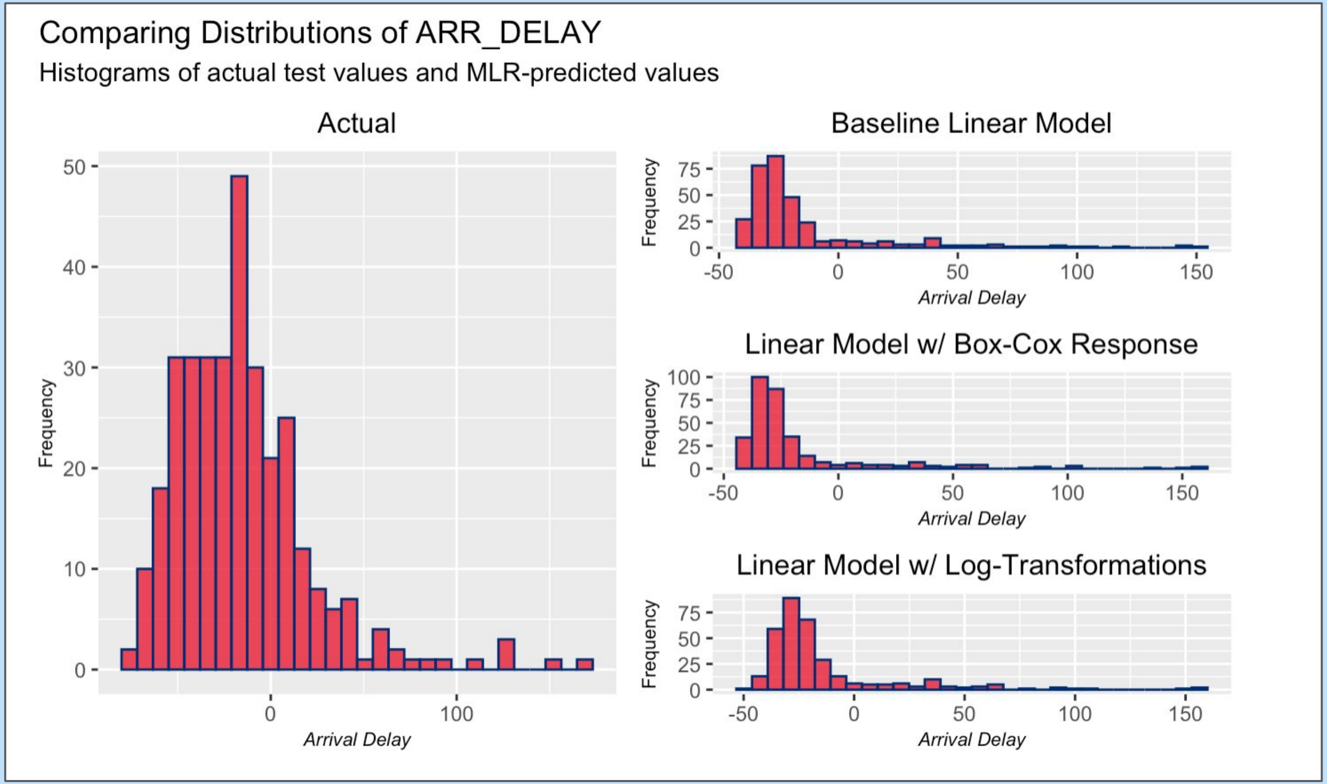
RESULTS

OVERALL ERROR TABLES

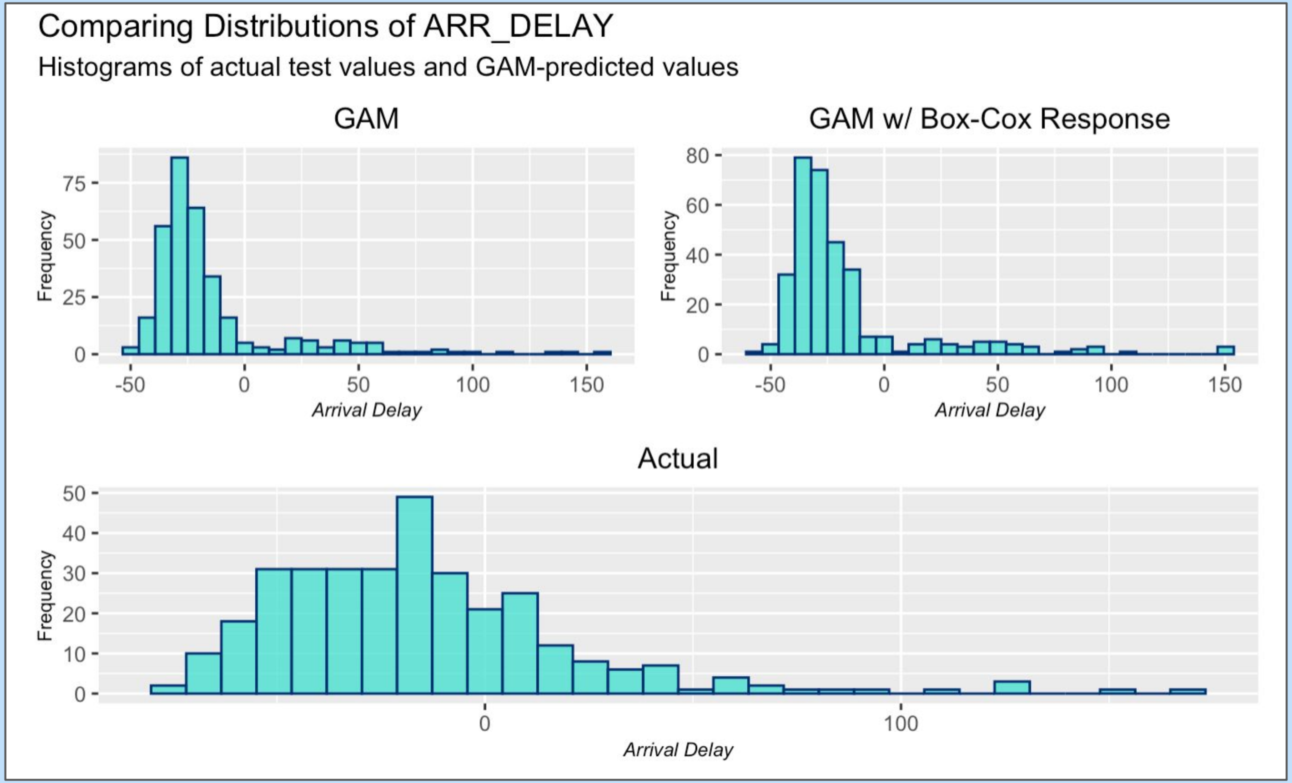
- Boosting provided the best performing model with a large percent increase and the **lowest MSE**

Model Name	Model Type	Model MSE	Model Percent Improvement
Baseline Linear	Multiple Linear Regression	322.46	↓ ---
Selected Linear w/ Log-Transformed Predictors	Multiple Linear Regression	333.90	↓ -3.5469
Selected Linear w/ Box-Cox	Multiple Linear Regression	334.92	↓ -3.865
GAM	Generalized Additive Model	312.30	↑ 3.1519
GAM w/ Box-Cox	Generalized Additive Model	317.45	↑ 1.5523
Random Forest	Tree-Based Regression	155.01	↑ 51.9272
Boosting	**Tree-Based Regression**	**129.80**	↑ 59.7479

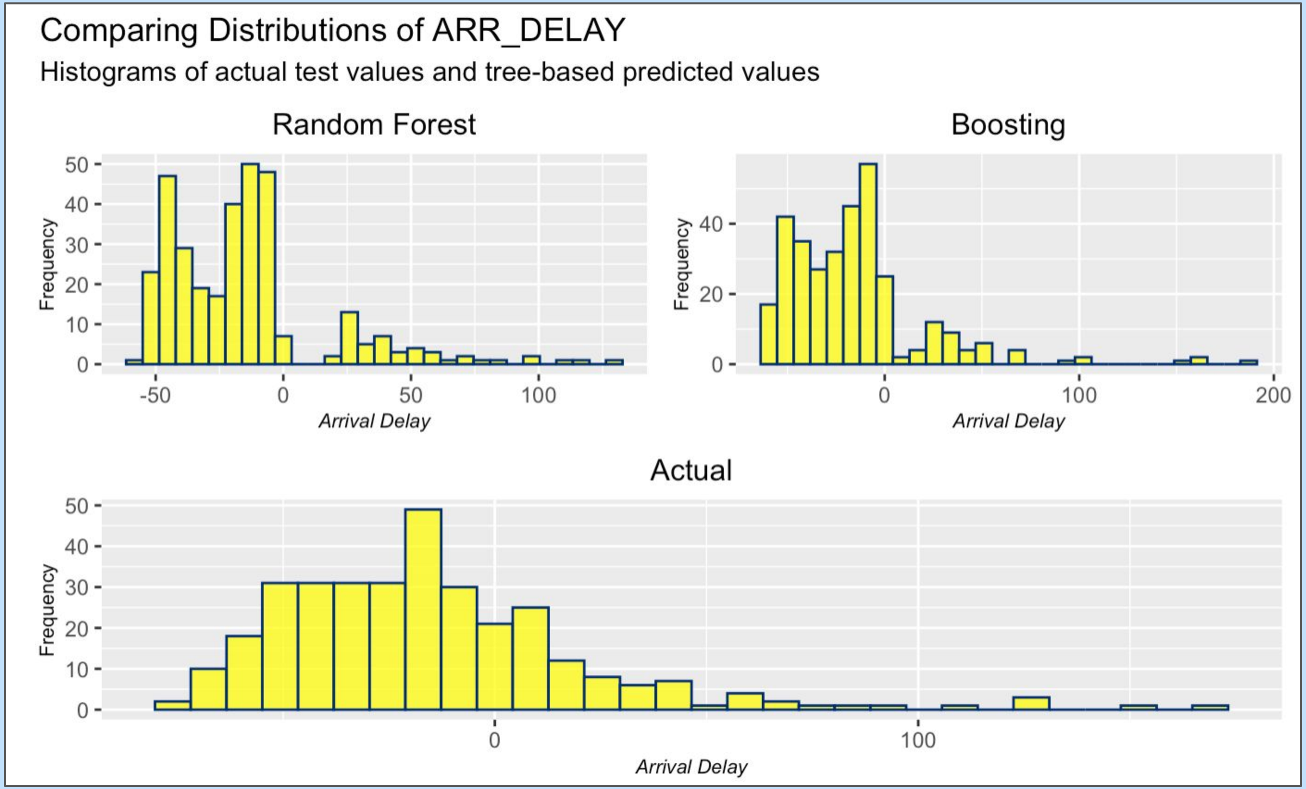
HISTOGRAMS: MLR



HISTOGRAMS: GAM

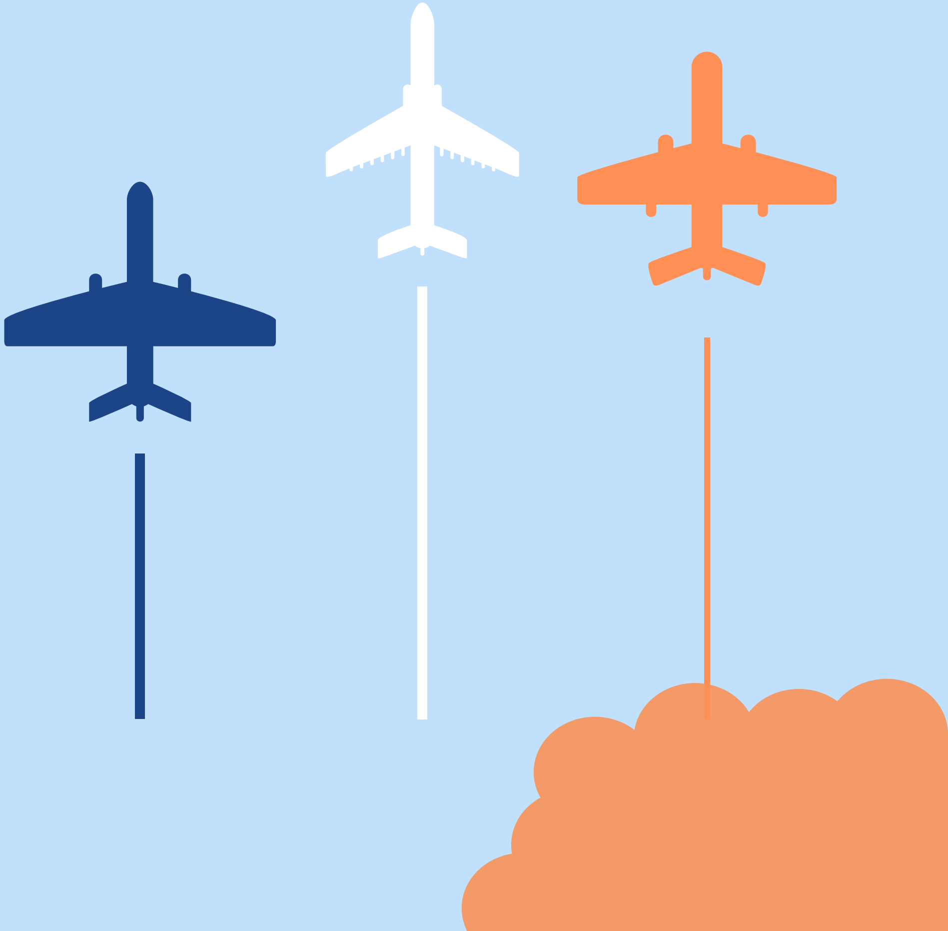


HISTOGRAMS: TREES



07

FINAL CONCLUSIONS



TAKEAWAYS

- **Optimal model** for predicting arrival delay times for flights from JFK to SFO and LAX is a **tree-based regression with boosting**
    - Large increase in predictive performance
- Using several types of models with several iterations helped identify weaknesses and increase comprehension of the dataset as a whole
    - Ultimately allowed for the best model selection
- Robust model would help all parties involved
    - Passengers, airlines, and airports
    - Each with their own priorities, but all helped by less delays and/or more efficient recovery from delays

FUTURE DIRECTIONS

- Focused on one month in one year, January of 2020
  - Expand to a longer period of time
- Only one originating airport, JFK, to two destinations, SFO and LAX
  - Explore the opposite, originating from CA and landing in NY
  - Add more origins and/or destinations
- Interesting to analyze effect of COVID-19
  - Decreased air traffic → less delays?
- Strong modeling procedure, but more holistic application to larger-scale data would uncover most crucial effect

THANKS!

For questions, please email any of our team members:

-Calleigh Smith (cas175@duke.edu)

-Hugh Esterson (hugh.esterson@duke.edu)

-Maria Henriquez (meh83@duke.edu)

-Hannah Bogomilsky (hlb25@duke.edu)

-Mariana Izon (mariana.izon@duke.edu)