Statistical Analysis Plan

Martin Olarte

2023-02-16

Administrative Information

Authorship Information

This statistical analysis will be primarily conducted by Martin Olarte and reviewed by peers as well as an instructional team composed of Dr. Sam Berchuck and Youran Wu.

Ethical Assurances

All statistical analyses included in an abstract or manuscript will reflect the SAP and no changes will be made to the SAP without discussing with the SAP author.

Timeframe

- Submission I: Aims, Introduction, and EDA, is due at 10am EDT on *February 3* .
- Submission II: Project Methods, Analysis Plan, & Preliminary Results is due at 10am EDT on *March 3*.
- Submission III: Final Report is due at 10am EDT on *April 6*.
- Presentation: The project will be presented in class in *April 11* to get feedback and suggestions from classmates.
- Final revisions of the individual project (a final report and comprehensive response to reviews) are due at 10am EDT on *April 25*.

Study Overview

Background/Introduction

The aviation industry is one of the most critical transportation sectors, providing fast and convenient means of travel for people and goods. On-time performance of flights is a critical aspect of the aviation industry that affects the satisfaction of customers and the reputation of the airlines. The aim of this project is to estimate the delays of flights between RDU (Raleigh-Durham International Airport) and MIA (Miami International Airport) using carrier on-time performance data from the Bureau of Transportation Statistics and weather data from NCDC NOAA datasets.

The rationale behind this project is to provide insights into the factors that contribute to flight delays and to develop a model that can predict the delay of flights. This information can be useful for both airlines to plan their operations and for customers to make informed decisions about their travel plans.

Study Aims

The main aims of the project are:

Aim 1

1) To develop a model that predicts the delay of flights between RDU and MIA based on historical carrier performance, scheduled departure time, and weather conditions.

Aim 2

2) To identify the factors that contribute to flight cancellations, such as weather conditions or specific seasonal trends.

Primary Hypotheses

The primary hypothesis is that harsh weather conditions typically result in longer delay times and cancellations for flights between RDU and MIA.

Secondary Hypotheses

The secondary hypotheses are that the relationship between flight delay time / cancellation and weather condition may vary depending on:

- Departure time of day (e.g. early morning vs. night time)
- Flight date (hinting towards seasonality effects)
- Airline-specific factors

Primary Outcome

Table 1: Primary Outcomes

Outcome	**Description**	**Variable Name and Source**	**Specifications**
Departure Delay	Departure delay (minutes), defined as actual departure time - CRS (Computer Reservation System) departure time	DEP_DELAY (carrier.csv)	Minutes (continuous and can be negative)
Cancellation Code	Cancellation codes used by the Bureau of Transportation Statistics (BTS)	CANCELLATION_CODE (carrier.csv)	A=Carrier Caused, B=Weather, C=National Aviation System, D=Security

Additional Variables of Interest

Table 2: Additional Variables of Interest

Variable	Description	Variable Name and Source	Specifications
Unique Carrier Code	Unique carrier code used to identify airlines	OP_UNIQUE_CARRIER (carrier.csv)	Character
Scheduled Departure Time	CRS (Computer Reservation System) departure time	CRS_DEP_TIME (carrier.csv)	Military time (integer from 0 (midnight) to 2359 (11:59pm))
Average Wind Speed	Average daily wind speed (miles per hour)	AWND.<origin or dest> * (weather.csv)	Miles per hour (double)
Precipitation	Precipitation (inches)	PRCP.<origin or dest> * (weather.csv)	Inches (double)
Average Temperature	Average temperature (degrees Fahrenheit)	TAVG.<origin or dest> * (weather.csv)	Degrees Fahrenheit (integer)
Maximum Temperature	Maximum temperature (degrees Fahrenheit)	TMAX.<origin or dest> * (weather.csv)	Degrees Fahrenheit (integer)
Minimum Temperature	Minimum temperature (degrees Fahrenheit)	TMIN.<origin or dest> * (weather.csv)	Degrees Fahrenheit (integer)
Direction of fastest 2-minute wind	Direction of fastest 2-minute wind (degrees)	WDF2.<origin or dest> * (weather.csv)	Degrees (integer from 10 to 360 in intervals of 10)
Direction of fastest 5-minute wind	Direction of fastest 5-minute wind (degrees)	WDF5.<origin or dest> * (weather.csv)	Degrees (integer from 10 to 360 in intervals of 10)
Fastest 2-minute wind speed	Fastest 2-minute wind speed (miles per hour)	WSF2.<origin or dest> * (weather.csv)	Miles per hour (double)
Fastest 5-minute wind speed	Fastest 5-minute wind speed (miles per hour)	WSF5.<origin or dest> * (weather.csv)	Miles per hour (double)

*Data is available for both origin and destination locations on the same day (e.g. TMAX.origin and TMAX.dest)

Table 3: Carrier Code to Airline Translation

Unique Carrier Code	Airline
9E	Endeavor Air Inc.
AA	American Airlines Inc.
DL	Delta Air Lines Inc.
F9	Frontier Airlines Inc.
MQ	Envoy Air
NK	Spirit Air Lines
YX	Republic Airways

The complete variable encoding documentation for **weather** data can be found here. Similarly, the complete variable encoding documentation for **carrier** data can be found here.

Study Population

Inclusion Criteria

Flight data was downloaded on a monthly basis from January 2020 to November 2022 for the state of North Carolina only (includes arrival and departure data).

Exclusion Criteria

The time frame was chosen to gather enough data for statistical significance, but at the same time limiting confounding variables around pre-covid commercial flying patterns. Thus, earlier year data (1987-2019) is still available, but not considered for this analysis.

Study Design

This is an observational study of commercial flights. Each data point represents a flight at a specified date and time, and attached are core weather statistics for the flight day at both origin and destination airports. All available flights between RDU and MIA were selected, but weather information is less reliable due to potential instrumental/human errors, as well as accessibility to certain data points.

Data Acquisition

The carrier data was gathered from the BTS website directly, on a monthly basis from January 2020 to December 2022 for the state of North Carolina only (includes arrivals and departures). After binding all monthly data together, and filtering for only flights between RDU and MIA, the weather data (collected from the NCDC NOAA) was joined using the corresponding station IDs for both airports. Weather data was joined twice, once for the origin and once for the destination, and appropriate variable renaming was performed to avoid confusion. Next, irrelevant predictors that would not be available at the time of estimation (other response variables like whether the flight was diverted or arrival information) were immediately discarded for the purpose of this analysis.

After data wrangling, variable selection began by determining NA percentages within each column (see the histogram below). The table below shows only numerical columns with more than 70% of observations with their respective summary statistics. Some of the weather data was extremely rare, making it difficult to implement in model fitting, and other data was insignificant (e.g. snowing patterns since it has barely ever snowed at the airport locations).

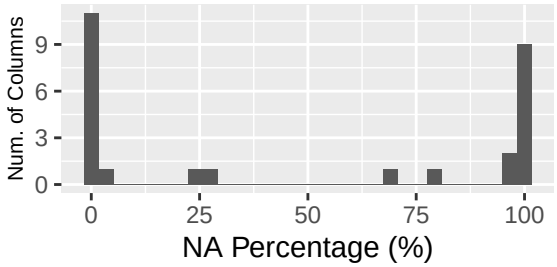


Figure 1: Columns NA Percentage

Table 4: Column Summary Statistics

	Minimum	1Q	Median	Mean	3Q	Maximum	Percent NA (%)
CRS_DEP_TIME	530.00	900.00	1408.00	1384.159	1830.00	2230.00	0.000
DEP_DELAY	-19.00	-5.00	-2.00	12.932	7.00	2619.00	2.163
DIVERTED	0.00	0.00	0.00	0.003	0.00	1.00	0.000
AWND.origin	0.45	4.47	6.49	6.846	8.72	20.13	0.000
PRCP.origin	0.00	0.00	0.00	0.155	0.06	5.26	0.000
SNOW.origin	0.00	0.00	0.00	0.004	0.00	1.60	24.202
SNWD.origin	0.00	0.00	0.00	0.008	0.00	2.00	26.161
TAVG.origin	25.00	63.00	75.00	70.297	80.00	88.00	0.000
TMAX.origin	30.00	74.00	83.00	79.656	89.00	102.00	0.000
TMIN.origin	15.00	52.00	68.00	61.882	74.00	84.00	0.000
WDF2.origin	10.00	90.00	130.00	154.247	230.00	360.00	0.000
WDF5.origin	10.00	90.00	140.00	159.598	230.00	360.00	0.522
WSF2.origin	6.90	14.10	16.10	17.039	19.90	47.00	0.000
WSF5.origin	10.10	19.00	21.90	23.309	27.10	64.00	0.522

- Contact information for data collection/acquisition: Martin Olarte (mo144@duke.edu)
- Carrier data was downloaded from the BTS website directly on February 10, 2023.
- The raw carrier data is stored in `RawData` as monthly csv files and the combined and clean carrier data is available in `Data\carrier.csv`.
- Weather data was ordered from the NOAA website on February 9, 2023 and downloaded after the order was processed.
- The raw weather data is stored in `RawData\weather.csv` and the clean weather data is available in `Data\weather.csv`.

Analysis Plans

Analysis Plan for Aim 1

Aim 1 will be addressed in the framework of a T-student robust linear model. First, instead of the traditional Ordinary Least Squares (OLS) method for reducing the residuals, robust regression uses different estimators (M-estimators, MM-estimators, etc.) that essentially adjust the weight of each point in an iterative process. Thus, such estimation is much less sensitive to outliers than OLS. This is particularly useful in the context of this aim, since the response variable of interest (Departure Delay) has a great number of outlying values. Furthermore, the T-student distribution has a heavy tail to accommodate outlying errors. This model will be fit in R software using the `rlm()` (robust linear model) function from the MASS package on only the data from January 2020 to December 2021. As noted in an article provided by Johns Hopkins Bloomberg School of Public Health, "the R^2 and F-statistics are not given because they cannot be calculated (at least not in the same way)", so "the bootstrap is a general purpose inferential method which is useful in these situations". Similarly, Wilcox et.al also suggest that "percentile bootstrap methods can be used to test hypotheses, which control Type I error probabilities relatively well even when there is heteroscedasticity" (R. Wilcox, A. Granger, and Clark 2013). In order to evaluate the hypotheses, we will examine the bootsrap distribution of coefficient estimates, with 95% confidence intervals not containing 0 considered evidence of a significant effect. Moreover, interaction terms will be explored using residuals to evaluate model fit too, along with external validation using data from 2022. We will conduct a sensitivity analysis on airport-segregated data

by fitting two individual models for RDU and MIA to determine if there are any airport-specific variations. A second sensitivity analysis will focus exclusively on estimating flight delays after the peak of the COVID-19 pandemic by fitting a model solely on data after April 18, 2022 when a federal judge in Florida struck down the U.S. federal transportation mask mandate and all of the major U.S. airlines lifted their pandemic-era mask requirements for domestic flights. The motivation for this being that the pandemic clearly had an effect on the transportation sector, which could be confounding some of the results of delays during periods where passengers were required to wear a face mask and airlines had to go through an additional step of checking this requirement. Finally, time permitting, we will address issues of generalizability of the data via exploratory analysis incorporating available data from other airports.

Analysis Plan for Aim 2

Aim 2 will rely on a logistic regression model to identify the factors that contribute to flight cancellations due to weather. The response variable Cancellation Code is categorical, but we are only interested in weather-related cancellations (Code C), since the other types of cancellations are much more rare and have a sparse nature that makes it difficult to model (e.g. a national security alert). Thus, we can remove such records and the response can be converted to binary, where 1 represents flights cancelled due to weather and 0 represents any other flight that was not cancelled. The logistic regression model will be fit in R software, using the `glm()` function with a binomial family. To identify the most important predictors in the model, we will conduct a variable selection procedure such as stepwise selection or LASSO regularization using AIC and BIC as information criteria to decide whether to include a predictor or not. If AIC and BIC do not agree, further investigation would be required to determine if that predictor should be included, always erring on the side of interpretability. During model fitting, assumptions must be checked. Absence of multicollinearity will be assessed with the `vif()` function from the `car` package, residual plots will be used to check for linearity and homoscedasticity assumptions, and the normality of residuals will be checked via a normal probability plot using the `qqnorm()` and `qqline()` functions. To evaluate the fit of the model, we will use standard goodness-of-fit measures, such as the Hosmer-Lemeshow test to compare the observed and expected frequencies in groups defined by the predicted probabilities, and the receiver operating characteristic (ROC) curve. We will calculate the area under the ROC curve (AUC) to quantify the discrimination ability of the model. To measure the predictive power of the model, we will also use 10-fold cross-validation due to the size of the dataset. Coefficient interpretation will be crucial to identify the factors that contribute most strongly to flight cancellations. We will calculate odds ratios and their confidence intervals to quantify the effect size of each predictor, summarizing the results in a table with visualizations for where the odd ratio of each coefficient stands to determine statistical and practical significance. All 95% confidence intervals not containing 0 considered evidence of a significant effect. Sensitivity analyses will test the robustness of the model, including the segregation of airport data like in Aim 1, but also segregation of airline data. Also, the use of alternative model specifications, such as different forms of the predictors (linear vs. quadratic) and their effect on the model can be explored. The generalizability of the findings for this aim could be hindered considering the potential impact of external factors (such as natural disasters) on flight cancellations. Moreover, the number of cancelled flights in the current dataset could be too small for a powerful analysis, and could call for older data which would not be confounded by the Covid-19 pandemic as much since the aim is limited to weather-related cancellations.

Exploratory Analyses

The plots shown below from Figure 3 and 4 are limited to data with a departure delay less than 3 hours or 180 minutes. Only 127 out of the total 7,859 flight observations have departure delays above 3 hours, which is only 1.62% of the data. Thus, it seemed reasonable to fit a model with a heavy-tailed distribution for the error terms. All carriers seem to have a similar highly skewed distribution of delay times, with center around 0 and a long tail towards longer delay times. However, there are differences in variance for outlying values particularly because 4,093 out of the 7,859 flights are from American Airlines. This further motivates the sensitivity analysis with respect to airline segregation for delays and cancellations.

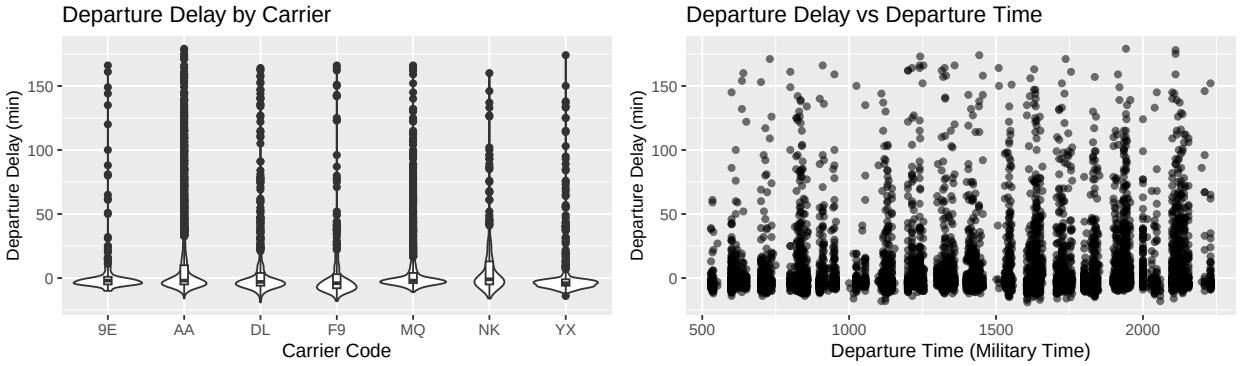


Figure 2: Response Variable EDA

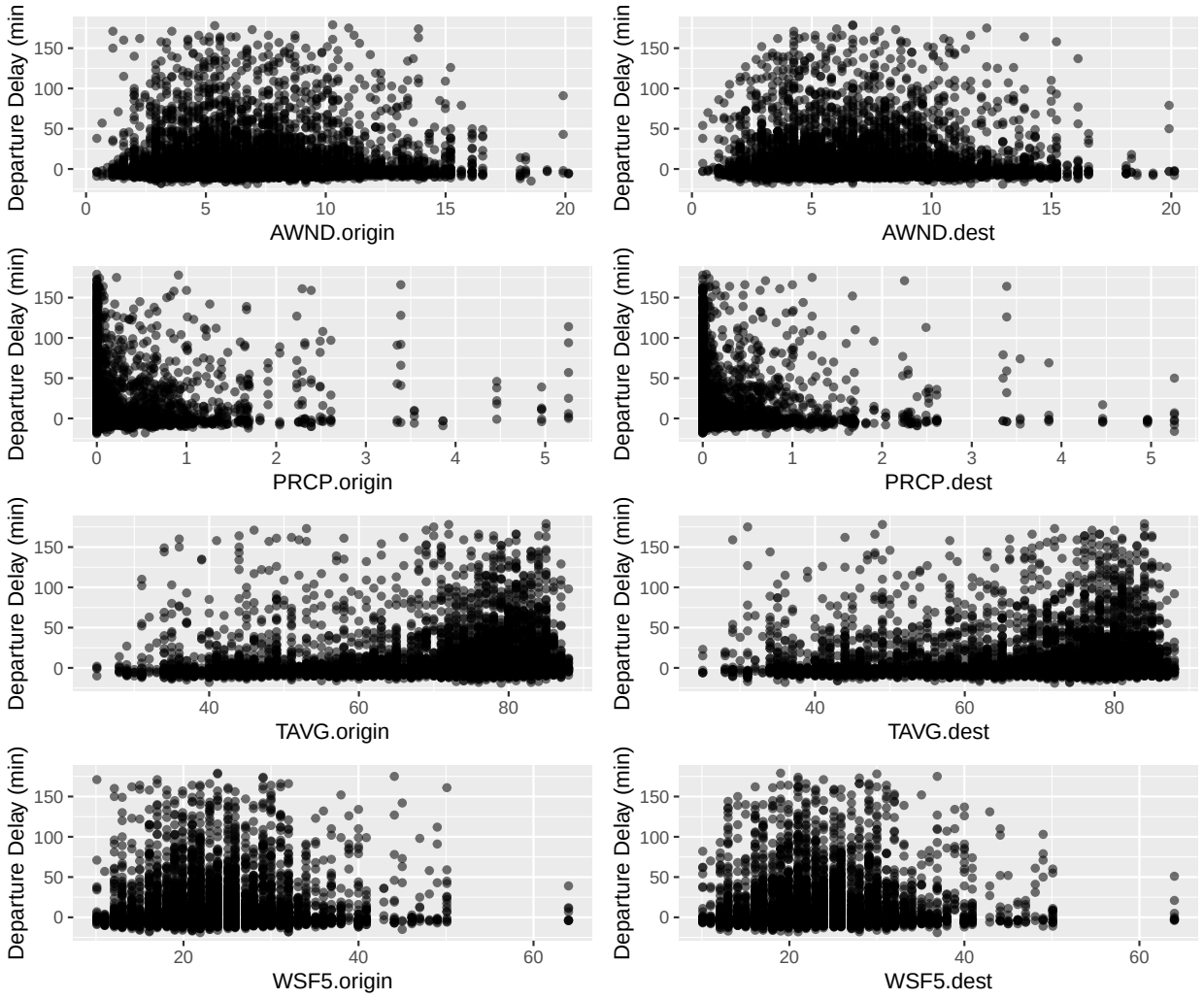


Figure 3: Predictors vs Response Variable (See Table 2 for variable names)

From the EDA, it is evident that pairs of weather covariates from the destination and origin will be highly correlated due to the fact that measurements are daily summaries, which could be an issue looking forward but will be decided when addressing multicollinearity.

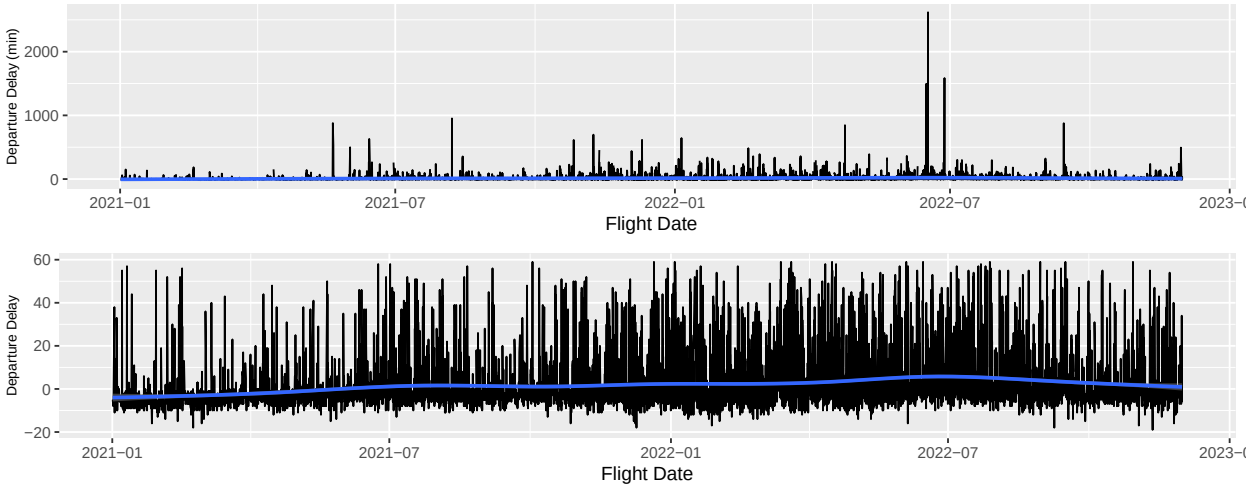
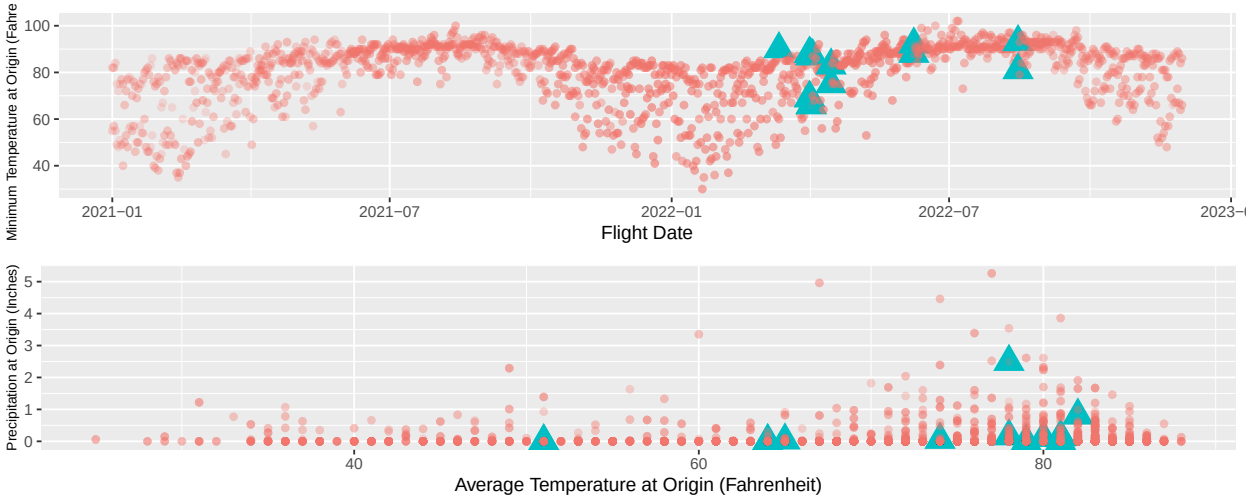


Figure 4: Time Series Plots

From the time series data in Figure 5, which could be interesting to explore further to identify seasonal or weekly changes, we can see that the outlying values are dominating the dataset, and even when filtering extremely for delays of less than 1 hour, fitting the default cubic spline GAM does not indicate any significant trend initially.



Finally, regarding flight cancellation trends, there were a total of 172 cancelled flights in the dataset, and only 17 of those were weather related (Code C). From Figure 6 it is evident that the cancelled flights due to weather (represented by blue triangles) were only present after 2022 in the data, and there seems to be a correlation with temperature. However, precipitation is not seen as a particularly strong predictor as expected.

Software

- R version 4.2.2 (2022-10-31) using the tidyverse, ggplot2, gridExtra, MASS, and car packages, along with their dependencies.

References

R. Wilcox, Rand, Douglas A. Granger, and Florence Clark. 2013. "Modern Robust Statistical Methods: Basics with Illustrations Using Psychobiological Data." *Universal Journal of Psychology* 1 (2): 21–31. https://doi.org/10.13189/ujp.2013.010201.