Estimating Flight Delay Time between RDU and MIA Using Historical Carrier Data and Weather Conditions

Martin Olarte

2023-02-10

Aims

The main aims of the project are:

1) To estimate the average delay of flights between RDU and MIA given information that is readily available like the weather.
2) To identify the factors that contribute to flight delays, such as weather conditions or specific seasonal trends.
3) To develop a model that predicts the delay of flights based on historical carrier performance and weather conditions.

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.

The response variable will be the delay of flights, which will be measured in minutes. The datasets provide a rich set of available predictor variables to choose from including carrier performance metrics, such as expected departure time, actual departure time, expected arrival time, actual arrival time, cancellations, diversions, and the different type of delays, as well as daily weather conditions, such as minimum, average, and maximum temperature (in degrees Farenheit), precipitation and snowfall (in inches), average wind speed (in miles per hour), direction of fastest 2-minute and 5-minute wind (in degrees), fastest 2-minute and 5-minute wind speed (in miles per hour), and distinct one-hot-encoded weather types (e.g. fog, thunder, hail, etc.).

Before variable selection, the data cleaning and wrangling steps will involve first defining the response variable, next removing any missing or irrelevant data, and then ensuring that the data is consistent and in a format that can be used for model fitting This will involve transforming the data into a format that is suitable for regression analysis and dealing with any outliers or anomalies in the data.

Exploratory Data Analysis

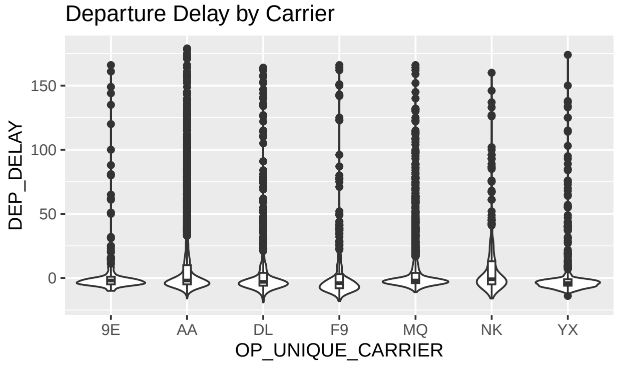
The carrier data was gathered from the BTS website directly, on a monthly basis from January 2020 to December 2022 for the state of North Carolina only (includes arrivals and departures). After binding all monthly data together, and filtering for only flights between RDU and MIA, the weather data (collected 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 variables that would not be available at the time of estimation (other response variables like whether the flight was cancelled or arrival information) were immediately discarded for the purpose of this analysis.

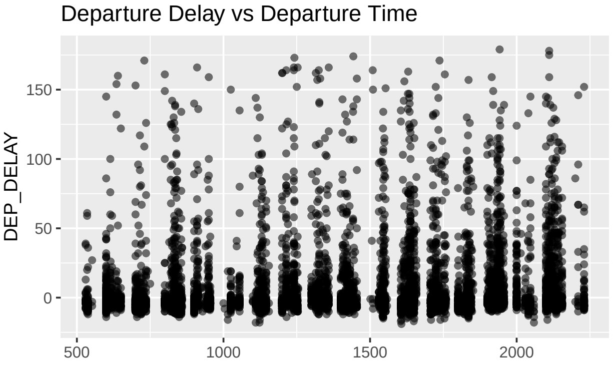
Key Statistics

After data wrangling, variable selection began by determining NA percentages within each column. Some of the weather data was extremely rare, making it difficult to implement in model fitting. The table below shows only numerical columns with more than 70% of observations with their respective summary statistics. The variable encoding documentation can be found *here*.

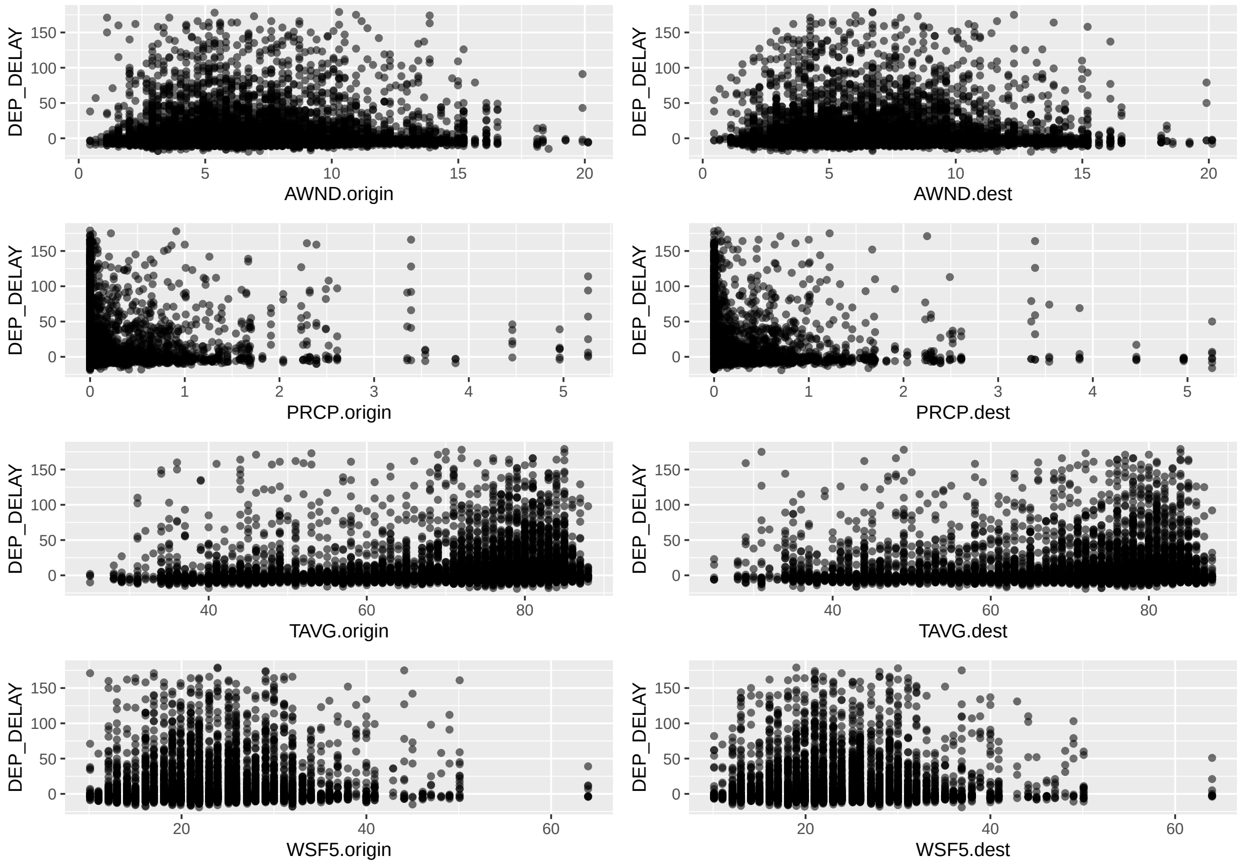
	Minimum	**1Q**	**Median**	**Mean**	**3Q**	**Maximum**	**Percent NA (%)**
CRS_DEP_TIME	530.00	900.00	1408.00	1384.159	1830.00	2230.00	0.000
DEP_DELAY	-19.00	-5.00	-2.00	12.932	7.00	2619.00	2.163
DIVERTED	0.00	0.00	0.00	0.003	0.00	1.00	0.000
AWND.origin	0.45	4.47	6.49	6.846	8.72	20.13	0.000
PRCP.origin	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

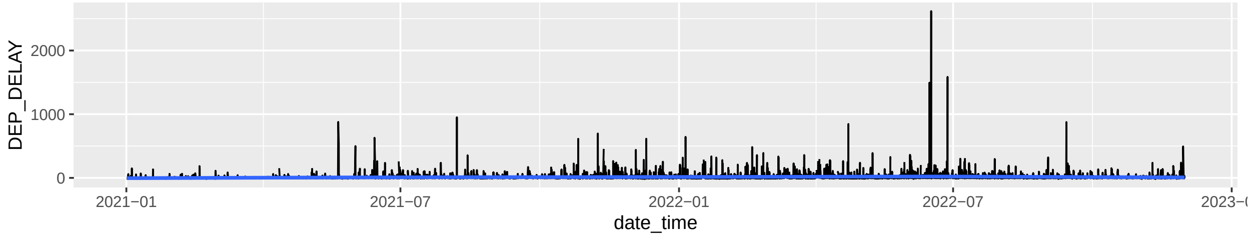
Plots

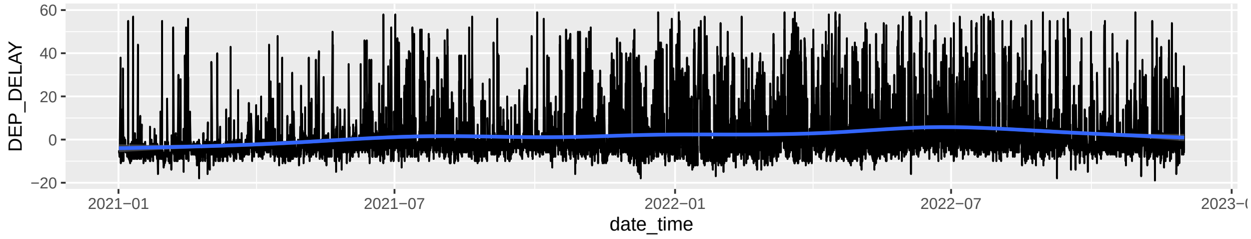




All carriers seem to have a similar highly skewed distribution of delay times, with center around 0 and a long tail towards longer delay times. However, there are differences in variance for outlying values particularly because 4,093 out of the 7,859 flights are from American Airlines.







The plots shown above 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 (at least for EDA) to ignore these values. I am still unsure if these values will be included in the model or not, but that would be decided by assessing the model accuracy and when addressing model diagnostics. 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. From the time series data, 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.