### Report

For this project, I chose to focus on transportation, more specifically with flight data. When coming up with ideas, I wanted to find the busiest airport in the United States. To my surprise, Atlanta’s Hartsfield-Jackson International Airport is the busiest airport in the *world*. After looking into some information about the airport, its history, and location in the U.S., it makes sense as to why it is so busy. Regardless, TranStats has data available for flights in the U.S. up to April of this year, but I chose January and February of this year for training and predicting, respectively. Additionally, I took the data for those two months and narrowed the csv file to strictly show all non-canceled flights to and from Atlanta International. Even though TranStats has an immense amount of useful information on flights, it does not have any flight rules for various dates. For this, I used Iowa State’s ASOS Network to obtain ceiling and visibility information around Atlanta International to create a flight rules column associated with that data. I also grabbed data from January and February for this.

The goal for this project is to design and deploy a predictive model that can accurately identify flights that were likely to be delayed in the month of February going to and from Atlanta International Airport. Using the month of January as the training set and putting it into Azure ML Studio with the Actual Elapsed Time as the target variable, I will deploy a regression model. With that model and an online endpoint, I will connect the results to Power BI via a python script. In Power BI, I will create a report comparing the real results of the flight times against the regression model’s prediction results. In the end, the accuracy of the model can truly be determined along with its usefulness.

The sources of data for this project were found on TranStats for Atlanta International’s flight data and Iowa State University’s ASOS Network for creating the flight rules. Both flight rules datasets have around 700 rows of data whereas each of the Atlanta International flight datasets have about 50,000 rows.

### Storyboard

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Setup | Actions | Outcomes | Results |
| Current Implementation | * Flights to and from Atlanta International are often delayed. | * The airport attempts to find the source of the delays. | * Some flights that were delayed will end up flying eventually. * Some flights may be canceled outright or due to many delays. | * Those who fly to and from Atlanta International may want to avoid the airport and/or avoid flying with Delta Airlines (Delta uses Atlanta International as their main hub) or whatever airline they used. |
| Future Implementation | * Flights to and from Atlanta may be delayed, though additional information is provided to prevent them. | * Atlanta International will use the pattern from the data to determine the risk of delays and choose to not delay certain flights, cancel others, and still delay some. | * Ideally, a reduced number of flights are delayed and/or canceled. | * Staff can determine delays in flights ahead of time to prepare and possibly prevent some of them. * Customers will have information to know the most efficient flights for them to take to and from Atlanta International. |

### 4V Model

Volume:

The flight data for January has 51,865 records of data across 54 attributes. February’s data has 49,504 records with the same number of attributes as January’s dataset. I also removed all the canceled flights for both months. As for the flight rules datasets, January’s has 760 records and February’s has 653. I only selected 4 attributes from Iowa State’s ASOS data, but I added in a fifth attribute to show the flight rule that would have been assigned to each time in the dataset. With all, I would give volume a 5/5.

Variety:

The TranStats datasets have squeezed just about every bit of information one would need about flights. The attributes for these are all thoroughly descriptive and useful in at least some way. I did not need to include every attribute that you can select from the flight data in this project, but that still adds to its data variety. For Iowa State’s data, I would say the same thing. I only needed 4 attributes (5 with the flight rules) for these datasets, though their data has a multitude of other informative attributes and different locations to filter if I needed them for anything. I would give data variety a 5/5.

Velocity:

Even though the data I am using is historical, the velocity is high quality due to both website’s frequent updates. TranStats flight data is available up until April of this year, and Iowa State’s ASOS data is updated close to real-time. To be more specific, data is synced from the real-time ingest every 10 minutes. As of writing this (July 20th), they have data up to this very day, just a few hours ago. I would give data velocity a 5/5.

Veracity:

The accuracy and reputability of the TranStats data is practically perfect because the website is from the Bureau of Transportation, it is an official website of the U.S. Government. The same goes for Iowa State’s data, the website is entirely educational, and all its data is trustworthy. As for data labeling, some of the attribute names and data values for the flight data and the flight rules data are unclear at first, but both websites provide good descriptions for them. All four datasets that I used in this project do have missing data. For the flights data, there are missing values due to canceled flights. For the flight rules data, I am assuming the reason these have missing values is simply because recording is not available at certain times. Considering everything, I would give data veracity a 5/5.

### Data Source Links

TranStats Flight Data

<https://transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGJ&QO_fu146_anzr=>

Iowa State University ASOS Network Data

<https://mesonet.agron.iastate.edu/request/download.phtml?network=GA_ASOS>