DSC 441: Fundamentals of Data Science
Project Proposal
Group Name: Speed Demons

Members:	Mashall Jahangir	**Email Address:**	mashaljahangir@gmail.com
	Chris Noreikis		chris.nore@gmail.com
	Annie Nguyen		A.Tho.Nguyen@gmail.com
	Hima Barla		spandanabarla26@gmail.com
	Tianhao Tan		tthtantianhao@gmail.com

Motivation:

The goal of this project is to analyze Airline Performance data, collected by the Bureau of Transportation Statistics (BTS) to better understand the causes of delay and build a prediction model to predict departure delay for air travel. So far, we were able to obtain a sample dataset and perform some preliminary analysis and visualization to help us better understand the data itself. At this stage, we are becoming familiar with the data set and are able to determine the dependent variable and independent variables for our prediction model. There are some changes from the proposal that we had decided. For development, we are going to work against a subset of one thousand data rows for each month which means there will be twelve thousand data used in development. The reason behind reducing the data size is we realized that our computers will overload if we are going to download more than a year's worth of data which have approximately 600,00 observations per month.

Exploratory Analysis

For the project, we focus on the Top 10 airlines based on the total number of flights and the cause of the delay which includes Weather, National Aviation System, Late Aircraft Arrival and Security delay. We performed 5 number summary and correlation analysis to understand the relation between variables. In conclusion, some outliers were present in the dataset and between the numerical variables there isn't any strong correlation.

Methodology

Our team applied different data mining techniques to aid us in gathering, cleaning, and enriching our data set. The BTS website only allowed us to download one month of data at a time, and each column within that month had to be manually selected. We then had to manually merge all the files after they were downloaded. This process was tedious and had to be repeated whenever we wanted to add data columns. To solve this problem, our team created a data pipeline to automatically load data from BTS's API and merge all the different files together. When analyzing our data set, our team noticed that when flights were not delayed, a number of other columns related to delay time were marked as NA. As a part of our pipeline, we added a normalization step to default these values to "0".

When attempting to run queries against our data set, we realized that the "Airport" and "Airline" columns were unique identifiers, not human-readable names. These data fields were not available for download from the BTS API. To solve this problem, we added a data enrichment step to our pipeline that merged in the missing data.

Experiment Results:

We started predicting flight departure delay by establishing a baseline model. Most of the independent variables in this model did not play any statistically significant role in the model of predicting departure delay. To measure a baseline model, we computed the mean of **DEP_DELAY_NEW** and used it to predict the same variable of every observation. Given a mean of **13.79** minutes, the baseline model came up with a mean squared error of **2174.09.** With this baseline established, we built a linear regression model. After removing independent variables with high correlation and low P-Value, our linear regression model had a much better MSE of **116.08.**

Next Steps:

Once we started plotting, we realized that we did not clear outliers. Therefore, the range of the y-axes in some scatter plots is very wide. For the final report, our team will determine outliers and prepare a cleaner dataset to work on. Our correlation analysis and linear regression model focused mainly on numeric variables. However, we also learned that there are many categorical variables in our dataset and these variables could be important for our final model, so we will work on converting these variables into numeric and build a model based on that. Additionally, we would like to have better visualization and we will utilize Tableau to build more meaningful visualization.

Project Contribution:

At this stage of the projects, most members of our team have been attending our meetings.