Understanding Flight Delays
CSE 519: Data Science Fundamentals
Project Proposal

Stony Brook University — October 21, 2021

1 Introduction and Objectives

Airline flight delay is one of the most severe problems across the globe in the 21st century. Between 2012 and 2019, the average delay rate for operating carrier flights is about 19.23%, based on the Airline On-Time Statistics and Delay Causes data provided by the U.S. Department of Transportation (DOT) Bureau of Transportation Statistics (BTS) [2]. As the number of flight operations was generally increasing, the Federal Aviation Administration (FAA) handled more than 10 million passenger flights in 2019 with 2.9 million passengers flying every day in and out of United States airports [4]. On the other hand, flight delay has been an issue for many years that can potentially become worse in the future if there is not a clear understanding of the causes and proper solution.

Flight delay releases pressure on everything related to aviation such as the National Airspace System (NAS), air traffic controllers, passengers, airlines, airports and the United States economy [3, 5, 7]. The NAS is a complicated network that controls air navigation facilities, airports, landing areas, aeronautical information, and so on [6]. In particular, flight delay is one of the most pressing problems for the NAS operations and air traffic controllers because of the flexibility to adjust flight and airport schedules. Furthermore, from an economics perspective, flight delays result in the reduction of business productivity and active leisure for passengers to compensate for the additional air travel time [7]. What is more, passengers' preferences on airline carriers and airports can change due to flight delays, which can result in the jeopardization of airline companies' marketing strategies. Last but not least, the total estimated cost of flight delays in the United States has increased every year. FAA [5] reported the total cost of delays separated into categories of airlines, passengers, lost demand and indirect costs, as shown in Table 1.

	2012	2013	2014	2015	2016	2017	2018	2019
Airlines	5.7	6.0	5.8	5.8	5.6	6.4	7.7	8.3
Passengers	9.7	11.0	10.5	13.3	13.3	14.8	16.4	18.1
Lost Demand	1.3	1.4	1.4	1.8	1.8	2.0	2.2	2.4
Indirect	2.5	2.7	2.6	3.1	3.0	3.4	3.9	4.2
Total	**19.2**	**21.1**	**20.3**	**24.0**	**23.7**	**26.6**	**30.2**	**33.0**

Table 1: Total Cost of Flight Delay in the U.S. (in billion dollars)

The total cost of flight delay in the United States increased by 71.88% between 2012 and 2019, with an average total cost increase of 8.30% per year. Additionally, an interesting observation is that the largest increase in the total cost was the increase in the cost of delay to passengers. In a study measuring the econometric estimation of the welfare impact of flight delays, Britto et al. [1] suggests that each passenger from the United States can gain about 1.50 to 2.50 dollars from a 10% reduction in delays. Hence, the influence of flight delays is reinforced again.

The objectives of this project are to examine the impact of flight delays on airports and airlines and to model when flight delays are likely. In this project, we shall follow the general definition of ”flight delay” and ”on-time” according to the FAA, BTS, and NAS. That is, we consider a flight is delayed if the flight arrived at the airport gate 15 minutes or more after the scheduled arrival time as reflected in the Computerized Reservation System (CRS). Similarly, we consider a flight to be on-time if the flight arrived at the gate less than 15 minutes than the scheduled time appeared in the CRS. Besides this overview, Section 2 of the proposal will discuss the data collection process and how we will perform the appropriate data analysis procedures to better understand flight delays. In Section 3, a more detailed plan will be provided on the data analysis and predictive modeling process so that we are ready to answer the proposed research questions. Finally, we will give some basic data descriptions with visualizations in Section 4.

2 Data and Methodology

The dataset used in this project comes from the BTS database called Airline On-Time Arrival Performance Data. BTS database is reliable because the data is open-sourced under the U.S. government. The data consists of essential variables, such as airline, airport, year and month, number of flights delayed due to weather, delayed arrivals in minutes, etc, to record on-time data for the commercial carrier network. We selected the period from January 2012 to December 2019. Besides the primary dataset, we also incorporated some secondary datasets with only one property (e.g. which airline carrier group) from the BTS database to further analyze questions of interest. Note that the secondary datasets need to be cleaned so that the corresponding subset of information can be merged into the primary dataset. For example, records before 2012 and after 2019 in the auxiliary data are deleted. After a series of data cleaning, the datasets will be merged based on the primary key(s) in the main dataset and proceed to further enrich the available data.

We will be using Python to perform all sorts of data manipulation, analysis, and modeling procedures. Here, we list some potentially useful packages. For data manipulation and scientific computing, we will use `pandas` and `NumPy`. Libraries such as `Matplotlib` and `seaborn` are excellent data visualization tools. Furthermore, `scikit-learn` and `SciPy` provide ample powerful methods to perform predictive data analysis and relevant statistical procedures.

3 Data Analysis and Modeling Procedures

3.1 Delays in Busiest Airports and the others

The first research question considers the delay pattern difference between the busiest airports and less busy airports. To properly define which airports are busy, we follow the classification of major airports defined in the BTS Airline On-Time Statistics and Delay Causes dataset. There are 30 major airports, which are all the core international airports around the United States, out of 379 airports in our primary dataset. Therefore, we consider the major airports as the busiest airports among all the airports. We will perform a series of data analysis procedures to compare the delay percentage by year, month, airport, busiest vs. not the busiest, and so on both quantitatively and visually to answer the question.

3.2 Delays Versus Airline Market Size

We are also interested in how airlines with a different proportion of market revenues behave in front of delay rates. We will add two extra secondary datasets to solve the puzzle. One of the datasets divides the 23 airline companies in the primary dataset into different carrier groups, decoded by the BTS Aviation Support Tables database. In all the 23 airline carriers investigated, the carriers are either in group National Carriers (carriers with annual revenue over \$100 million to \$1 billion) or group Major Carriers (carriers with annual revenue over \$1 billion) after an appropriate cleaning to the duplicated values in the secondary dataset.

Another secondary dataset separates the airline carriers into marketing carriers and regular carriers. The marketing carrier group is a special subset of all airline carriers in that marketing carriers have market flights for themselves and also regional codeshare partners in some cases, as illustrated in the BTS Airline On-Time Performance Data database. We will apply similar analysis procedures to figure out the relationship between carrier group and flight delay.

3.3 Delays versus Airline Establishment Time

Some of the airline carriers established since 1960, such as the American Airline Inc. Some others are relatively new, such as JetBlue Airways started in 2000. Does the establishment time of airline companies affect the delay rate? We have a dataset from the BTS Air Carrier Statistics dataset on the starting and ending year of operation (empty of ending year means still in operation). To properly start the research, we propose a categorization of two groups, a relatively newer airline group and a relatively older group of airline companies based on a suitable standard. Furthermore, considering the establishment time as a continuous variable is also plausible for analyzing their difference in delay ratios.

3.4 Modeling and Prediction

Finally, to complete a data analytics project, we will continue to obtain additional insights on the datasets and build a model for predicting future delay trends. Feature selection will be used to both reduce the modeling computational complexity and perhaps to improve the model performance. Several machine learning algorithms will be imposed with hyperparameter tuning to maximize the model's predictive accuracy using an appropriate evaluation metric or a loss function. After a system trial and error of modeling and validation, we will come up with a productive tool to predict when an airline delay is likely.

4 Preliminary Data Exploration

In this section, we present several basic data exploration results to better understand the overall structure of the primary as well as the secondary datasets. Figure 1 is a summary statistics bar plot for flight status from January 2012 to December 2019[1].

Based on the bar plot, we observe that about 80% of the flights during the eight years are on time. About 1.58% of the flights are canceled, and about 0.24% of them are diverted to another destination. The rest of the categories are different reasons for the delay.

[1] Note that it is possible to have multiple causes assigned to one delayed flight. In this scenario, every cause is prorated based on the delayed minutes it is responsible for. Thus, the displayed numbers and percentages are rounded and may not add up to the total.

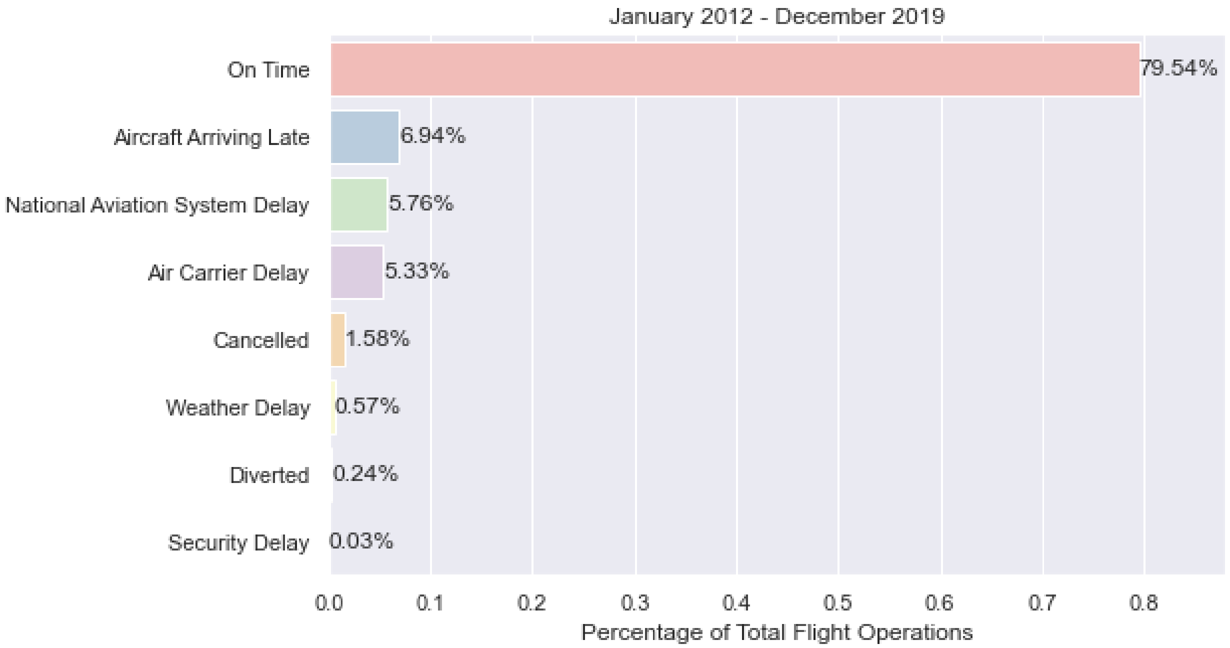

Figure 1: On-Time Arrival Performance By Flight Status

Furthermore, a lot of people are wondering about the delay rate for each year and see if improvements have been made. The bar chart in Figure 2 shows that between 2012 and 2019, most delay rates are between 17% and 22% with the highest delay rate in 2014 and the lowest in 2016. We believe the difference of about 5% in delay rates is significant and deserves a more thorough investigation. In general, the delay rate has not been going down.

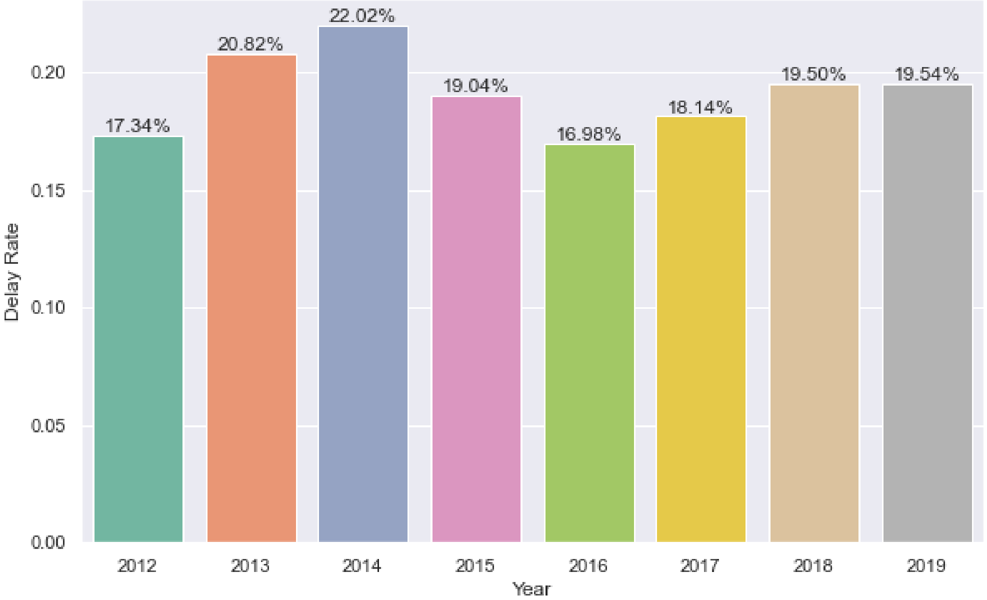

Figure 2: On-Time Arrival Performance By Year

Table 2 shows a numerical comparison between the less busy airports and the busiest airports on delay rates. Except for 2014, delay rates are higher for the busiest airports. The difference between the delay rates for each year is about 2 to 3 percent.

	2012	2013	2014	2015	2016	2017	2018	2019
Less Busy Airports	16.97%	20.66%	22.11%	18.57%	16.19%	17.40%	18.94%	19.02%
Busiest Airports	18.39%	21.30%	21.75%	20.40%	19.16%	20.22%	20.93%	21.23%

Table 2: Delay Rate for Busiest and Less Busy Airports

Last but not least, we can do a similar exploratory data analysis to compare the delay rates for national carriers and major carriers, as shown in Table 3. The statistics appear that national carriers have a higher delay rate than major carriers (about 2% to 3%) for all years between 2012 and 2019. This may suggest that smaller airline companies have a higher probability of flight delay compared to larger airline carriers.

	2012	2013	2014	2015	2016	2017	2018	2019
National Carriers	18.23%	21.95%	23.60%	21.25%	19.37%	19.91%	20.66%	20.36%
Major Carriers	16.97%	20.20%	21.51%	18.36%	16.12%	17.54%	18.70%	18.99%

Table 3: Delay Rate for National Carriers and Major Carriers

References

[1] R. Britto, M. Dresner, and A. Voltes. The impact of flight delays on passenger demand and societal welfare. *Transportation Research Part E: Logistics and Transportation Review*, 48(2):460–469, March 2012.

[2] Bureau of Transportation Statitics. On-time performance - reporting operating carrier flight delays at a glance. `https://www.transtats.bts.gov/HomeDrillChart.asp`. Accessed: October 15, 2021.

[3] L. Carvalho, A. Sternberg, L. M. Gonçalves, A. B. Cruz, J. A. Soares, D. Brandão, D. Carvalho, and E. Ogasawara. On the relevance of data science for flight delay research: a systematic review. *Transport Reviews*, 41(4):499–528, December 2020.

[4] Federal Aviation Adminstration. Air traffic by the numbers. `https://www.faa.gov/air_traffic/by_the_numbers/`. Accessed: October 16, 2021.

[5] Federal Aviation Adminstration. Air traffic by the numbers 2020. `https://www.faa.gov/air_traffic/by_the_numbers/media/Air_Traffic_by_the_Numbers_2020.pdf`. Accessed: October 16, 2021.

[6] Federal Aviation Adminstration. National airspace system. `https://www.faa.gov/air_traffic/nas/`. Accessed: October 16, 2021.

[7] E. B. Peterson, K. Neels, N. Barczi, and T. Graham. The economic cost of airline flight delay. *Journal of Transport Economics and Policy*, 47(1):107–121, January 2013.