Hugh Esterson, Maddie Demming, Michaela Kotarba, Jason Zhang & Nick Reddy
Prof. Stephens-Martinez
COMPSCI 216
March 6, 2022

COMPSCI 216: Project Proposal

Introduction and Research Questions

The world of aviation is a fast-paced, high-volume industry, shuttling millions of passengers across the globe each and every day. Flying at 30,000 feet above the ground and at speeds upwards of 500 miles per hour, the feat of flying is fascinating. However, the behind-the-scenes world of commercial aviation is just as meticulous and astounding as the flying itself. Complex networks of conveyor belts bring checked luggage through the subterranean maze of international airports while ground agents rush to fuel, cater, and clean the aircraft, but perhaps most importantly, airlines are a logistics masterpiece. Using a variety of data, airlines constantly look to optimize their operations, from the dynamic pricing models to ensuring the correct aircraft match the given routes, all while keeping in mind the safety and comfort of customers.

Thus, the data collected and analyzed by airlines is crucial in ensuring that the logistical ensemble goes without a hitch. According to the United States Bureau of Transportation Statistics, a whopping 68.1 million passengers traveled by air in the month of November 2021 within the United States alone. Crucially, however, the aviation industry is of particular interest since the onset of the COVID-19 pandemic, which brought worldwide, months-long cancellations. While the aviation sector has rebounded since the proliferation of vaccines and the return of leisure travel, the data revolving around the airline industry from 2020 through the present day is of specific interest for further research.

For our COMPSCI 216 final project, our team is interested in using data-based approaches to better understand the aviation industry. In our preliminary project planning, we have decided to build a relational database with several tables, formed by individual and pre-existing datasets, to analyze trends across the industry. This would be done in Python with the SQL knowledge we have accrued in the course so far. A variety of research questions have been brainstormed and discussed by the group so far, all of which seem to be directly answerable given the datasets that we have aggregated. Our main research question as of current is: “How did commercial aviation routes from Raleigh-Durham International Airport (RDU) change during the COVID-19 pandemic, specifically in terms of the number of flights, the breadth of destinations, and the number of carriers? Using these three aspects, how did the total number of potential passengers on flights originating from RDU change from pre-COVID years?” Other potential research topics/questions include a geospatial analysis of the flights to and from RDU, as well as questions regarding the types of aircraft used to service RDU’s passenger base. The aircraft-centric question could concern topics, such as the calculating the number of passengers (by aircraft capacity) per route or classifying the routes by aircraft and its flight range.

We believe that our main research question is substantial, feasible, and relevant for the scope of this project. Firstly, this research question and its three aspects is substantial, as it can be analyzed in a variety of different ways. Creating the database, with several tables, will be a substantial undertaking to effectively aggregate various large-scale data sources together. With the created database, queries will be used to filter data and calculate a total number of potential fliers. This calculation will be more complex than a simple summary statistic, as it involves creating our own statistic. We also plan to implement hypothesis testing during the project to confirm or disconfirm statistical significance. Additionally, this research question will be substantial since we plan to implement some features of a geospatial analysis, which goes beyond the course content at this point of the semester.

While the research question is one that is substantial, we also believe that it is feasible for this time constraint and the skills of our group members. Each member of our group is well-experienced in Python and data science as a whole. While we may not have encountered this exact analysis in Python, many group members are experienced in this sort of analysis and are eager to learn how to implement new skills (e.g., geospatial analysis) in Python. As a quick point of reference, members in this group have had

previous experience in working with: (1) relational databases, (2) geospatial analysis, (3) Python data science projects, and (4) aviation-specific data. Thus, we think that the research question will be feasible, given the experience of the members involved. Additionally, in terms of data, we plan to make this question all the more feasible by beginning by focusing only on the flights originating from RDU, which will ensure that the analysis and subsequent discussion is focused. Another benefit of this proposed research project is that it is easily scalable to different amounts of data. That is, if the analysis seems to be too simplistic, different filters and subsets can be used to analyze the findings across additional airports, comparing them to RDU.

Lastly, this research question is relevant to a greater and more local audience. Focusing on RDU, this research will answer and engage with data that is local to the Duke and wider Durham communities. Additionally, as motivated by the introductory paragraphs of this proposal, the aviation industry and any associated consultants would be interested in this analysis due to the volatile nature of air travel since the spread of COVID-19. The results of this project can be used to more specifically understand *how* the airline industry may have changed during the pandemic, by focusing on RDU as a specific case. Thus, the research question and the resulting project will be relevant to the aviation community as well as a broader populace of Durham residents and travelers.

Data Sources

In our preliminary research, our group has identified four sources of data that will be used in our project. Each of the datasets are publicly available, with three being stored on Kaggle and the fourth being accessible on the wider internet. Visualizing each dataset as its own table within our created database, the "main" dataset is called "Flight Route Database" and contains information about each commercial flight route in the United States. At the beginning of the project, this dataset will be filtered to focus only on flights originating from RDU, while additional airports may be used once further analysis is complete. The "USA Airport Dataset" on Kaggle provides us with information on all airports in the US. Crucially, this dataset provides latitude and longitude coordinates for the airports, which will be the basis for any subsequent geospatial analysis. The "Airline Fleets" dataset provides an array of information on the fleets of the major airlines in the world. Lastly, this publicly-accessible file, posted by a French university contains information on several types of aircraft. This information includes the capacity of each plane, which will be crucial in calculating the total number of potential passengers.

At this stage of the project, we have carried out a very preliminary analysis of the datasets and have found that they will be very useful in answering our overall research objectives. Each dataset provides a different perspective on the question at hand, from the airports to the airplanes to the routes. Each also have columns that can act as linking keys to form the relational database, which will aid in addressing our research question in a database format of querying. Lastly, the data is exhaustive and encompasses the focus of our research project, further justifying its usage.

Collaboration Plan

This project group contains five students who are all friends prior to COMPSCI216, and we are all excited to work together in an academic setting. We plan to divide responsibilities equally to begin but are encouraging each other to take leadership on stages of the project that revolve around topics or analyses that a specific group member may have had previous experience with. We also plan on working together, in-person, as much as possible, but will decide specific responsibilities on a weekly basis. On average, we expect each group member to contribute at least 2 hours per week on the project, but we have agreed that this number may increase as the project comes to a close. As of now, we have found that weekly meetings over Zoom at 7:30pm will be used to keep team members up-to-date and on the same page. Since we all have preexisting relationships with one another prior to this class, we will continue to communicate via a group chat over text. Finally, we will store data and project materials on GitHub in a shared repository, which is accessible here.

Links (Gradescope does not seem to work with the embedded hyperlinks):

- USBTS: https://www.bts.gov/newsroom/november-2021-us-airline-traffic-data
- Flight Route Database: https://www.kaggle.com/open-flights/flight-route-database
- USA Airport Database: https://www.kaggle.com/flashgordon/usa-airport-dataset
- Airline Fleets dataset: https://www.kaggle.com/traceyvanp/airlinefleet
- Aircraft info dataset: http://www.lsv.fr/~sirangel/teaching/dataset/aircrafts.txt
- GitHub repository: https://github.com/michaelakotarba/216project