Hugh Esterson (hhe4), Maddie Demming (med100), Michaela Kotarba (mk401), Jason Zhang (jzz4) & Nick Reddy (npr12)
Prof. Stephens-Martinez
COMPSCI 216
April 6, 2022

COMPSCI 216: Project Prototype

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, commercial aviation is a logistical 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.

As Duke students and residents of the greater Research Triangle of North Carolina, we're interested in looking at travel out of Raleigh-Durham International Airport (RDU). The Triangle has boomed and continues to grow in recent years, lending itself as an interesting city that we could analyze. With population growth comes the need for transportation and, thus, more traffic through adjacent airports. Thus, for our COMPSCI 216 final project, our team is using data-based approaches to better understand the aviation industry with a focus on RDU. Our main research question is: "How did commercial aviation routes from Raleigh-Durham International Airport change from 1990 to 2009, specifically in terms of the number of flights, the breadth of destinations, and the number of carriers?" Our main points include a geospatial analysis of the flights originating from RDU, as well as quantitative analyses regarding the volume of travelers flying out of RDU. To aid in our analysis, we have split our data temporally (i.e., by smaller date ranges) to illustrate the change in flights out of RDU.

We believe that our main research question is substantial, feasible, and relevant for the scope of this project. Firstly, this research question is a *substantial* one, as it can be analyzed in a variety of different ways. We have taken on a substantial undertaking to effectively aggregate various large-scale data sources together, filter data, and calculate new data points. These calculations and processes are more complex than running simple summary statistics, as it involves heavy data cleaning and creating our own statistics of interest. 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.

So far, this project has proven to be *feasible* as well. We have successfully incorporated a geospatial package into our Jupyter Notebook environments and produced several visualizations with geospatial components. Additionally, we have chosen to analyze only the flights originating from RDU, which has ensured 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 prototype report, the aviation industry and any associated consultants would be interested in this analysis and its application to other quickly-growing or soon-to-grow metropolitan areas. The results of this project can be used to more specifically understand how the airline industry may have adapted to meet increased demand by focusing on RDU as a case example. 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.

Since our proposal, we have made a number of revisions regarding details discussed in this section. While all relevant changes have been implemented into the writing above, this paragraph serves to explicitly acknowledge the changes. Most centrally, we have shifted the focus of our research question to focus on a different time period than we had listed in our proposal. While we were initially interested in reviewing changes in aviation during and after the COVID-19 pandemic, the most comprehensive data source listed flights with a timeframe from 1990 to 2009. Thus, we have decided to focus on this range instead, which coincides well with the population growth seen in the region at the time. Next, we have re-emphasized the project's focus on a geospatial analysis to analyze how destinations out of RDU have changed. This has been paired with other more quantitative analyses that regard the number of passengers and flights, as well as airlines, that depart from RDU. Both the geospatial and quantitative analyses will also now incorporate a temporal component. Lastly, we have subset the data into four 5-year periods, which allows us to better illustrate the changing nature of aviation from RDU and more directly answer the posed research question.

Data Sources

Following from our project proposal, our group identified four data sources 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 the "USA Airport Dataset" and provides us with information on some 3.6M+ flights within the US from 1990 through 2009. This dataset was filtered to focus only on flights originating from RDU, but additional airports may be used once further analysis is complete. Furthermore, and as mentioned in the previous section, we have also subset this dataset into four smaller sections, with four 5-year periods. Crucially, this dataset provides latitude and longitude coordinates for the airports, which is the basis for our geospatial analysis. It also provides the counts of passengers and number of seats on the flight, which are analyzed in our quantitative analysis.

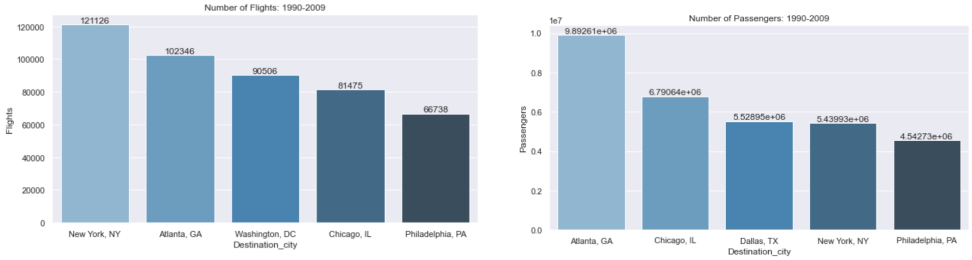
Next, the "Flight Route Database" contains information about each commercial flight route in the United States. Again, we have filtered to focus on only the flights originating from RDU. This dataset provides information on the number of destinations from RDU, which can be linked to the main dataset for its geospatial component. Finally, this dataset provides information about the airlines that fly out of RDU. Thirdly, the "Airline Fleets" dataset provides an array of information on the fleets of the major airlines in the world. And finally, a 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. Each dataset provides a different perspective on the question at hand, from the airports to the airplanes to the routes. Altogether, the data is exhaustive and encompasses the focus of our research project, further justifying its usage.

An additional type of data that we have implemented since the writing of our proposal are shape files sourced from the U.S. Census Bureau. These files allow us to map the geospatial components of our data, providing boundaries and polygon objects that aid in visualization. Correspondingly, we have set a popular coordinate reference system (EPSG 4326), which is used for most GPS satellite navigation purposes. This part of the data is crucial in allowing us to complete our geospatial analyses and originates from an authoritative source in the U.S. government.

Preliminary Results and Methods

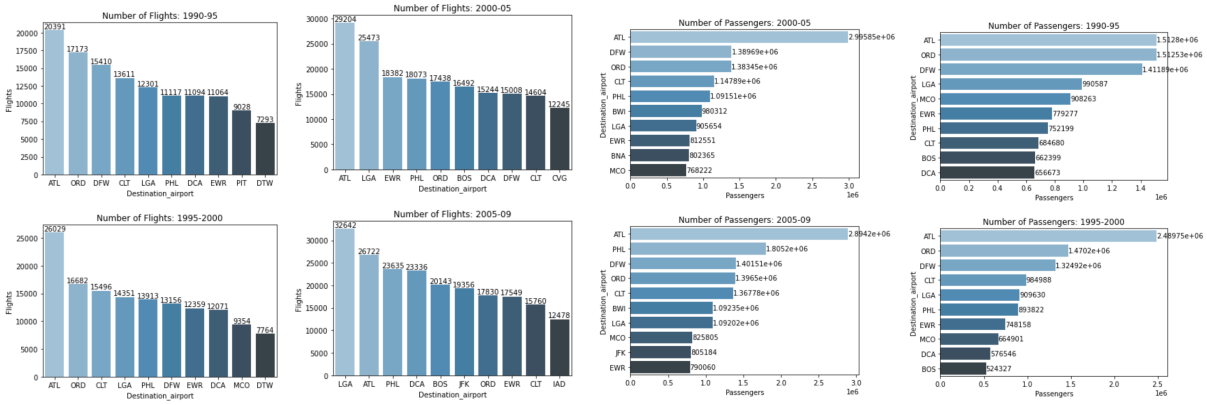
Before dealing with any piece of data, we underwent several data cleaning steps to arrive at a clean and work-able dataframe. First, we created several "composite" statistics such as "seats per flight" by using Pandas operations to create new columns. We also parsed the timestamp into a year variable for our temporal analysis. Similarly, we stripped address strings to isolate the postal state code, which then could link to our geospatial shape file. Next, we applied several masks for filtering, including focusing only on the origin of RDU and flights that had more than 30 passengers to remove any private or charter flights in the dataset. We then subset our data into the four 5-year ranges with another filter and completed all subsequent group-bys and visualizations from these four subsets.

To begin to answer our research question, we began by running several exploratory data analyses to become accustomed to our data and arrive at a few overarching findings. We first pulled in our CSV-formatted data on airports, routes, and fleets as Pandas DataFrames and started looking at what destinations are most popular, by both the number of flights and the number of passengers. First, we filtered for RDU specific flights and found that, between 1990 and 2009, the most frequent flights out of RDU are to New York, NY (121,126 flights), Atlanta, GA (102,346), Washington, DC (90,506), Chicago, IL (81,475), and Philadelphia, PA (66,738). By number of passengers, however, Atlanta is, by a wide margin, the most popular destination, followed by Chicago, Dallas, New York, and Philadelphia.



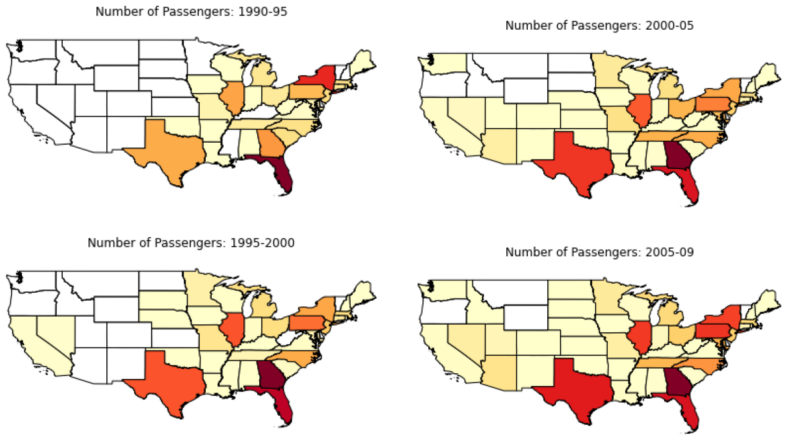
As an additional aspect of our strictly quantitative approach, we melted our sorted dataset to see the number of flights and number of passengers serviced out of RDU over each year included in the dataset. Here, we note a sharp increase in the popularity of flights leaving RDU in the mid-to-late-1990s, peaking in 2005 before declining shortly after.

After this very preliminary exploration of the data, we focused on the temporal subsetting of the data to identify how the number of flights and number of passengers to different destinations changed from 1990 through 2009. The following 8 graphs show this change, with the left two columns showing the number of flights to the 10 most popular destinations across each time period and the right columns showing the same for the number of passengers. As some very rough conclusions, we can see that the ranking of most popular destinations differs a bit when sorting by the number of flights and the number of passengers. We also note the predominance of Atlanta as the most popular destination (n.b., largest Delta hub in the country) in seven of the eight panels.



We then turned to our geospatial analysis by linking the destination airports to their states and the corresponding polygons given for each state in the shape file. With this methodology, we grouped by the number of flights to each state and created a choropleth map for each of the four time periods to show

where flights originating from RDU arrived. In these maps, we can see that flights from RDU have gradually expanded from a strict focus on the East Coast to a more diverse selection of destinations throughout the West. We see the strong dependence on the Georgia (ATL), New York, and Florida markets as well. Lastly, we can see the growing popularity of Texas over time, as the darker choropleth filling indicates a higher concentration of passengers to that state.



To access our code and data, visit our GitHub folder here, where you will find all the materials we used to complete this analysis.

Reflection and Next Steps

So far in our project, we have successfully implemented a holistic analysis on the quantitative, temporal, and geospatial fronts. We have created basic visualizations that directly address our research question, which is a very important component of our progress. Many of these visualizations are essentially complete and can be used in our final report. Our teamwork has also been very amicable and team members have been helpful in assisting one another.

The most challenging part of the project in this prototype step has been merging the various datasets together. Several of the datasets do not overlap in terms of their timing, which necessitates very careful consideration to avoid any data inconsistencies that could emerge when merging datasets from different time periods. Additionally, the implementation of the geospatial packages into Python was quite challenging, but this hurdle has been cleared.

As next steps, our team hopes to clean our code and ensure that it is efficient and iterable. We also hope to finalize our visualizations to make sure they are aesthetically pleasing and also ensure that they tell an informative "story" in regard to our research question. Thus, each team member will be assigned a portion of the visualizations to refine, before combining our individual work together. Another to-do will be to create print-able, read-able tables that can be presented in the final report to complement the provided visualizations. We will also create an additional time (Monday 6:00pm) to meet each of the remaining weeks to complement our existing 7:30pm Wednesday Zoom calls. This will help us address all the tasks in the final stages of our project.

GitHub link: https://github.com/michaelakotarba/216project