**Air Traffic Data Analysis using Elasticsearch (U.S. International & Domestic Air Traffic Data)**

Roxana Cuevas

Jenny Iglesias

Dante Villalobos

Abrar Hasan

Computer Information Systems 3200

California State University, Los Angeles

**Abstract:**

This poject utilized a dataset that captured information regarding U.S. International Air Traffic for the years 1990-2020. Among the information captured was the date of occurrence, U.S. airport name, foreign airport name, carrier name, airport ID, area code, airline ID and total number of passengers in each flight. With this information we will be able to analyze the data to determine the busiests months for international travel throughout the year and thus recommend which months are the busiest and least busy for international travel. This can be done by analyzing the total number of passengers and number of departures leaving each U.S. airport on any given day for our analysis period. The same information will also provide a snapshot of the busiest U.S. airports for international travel and concurrently a snapshot of the busiest foreign airports receiving these same U.S. departing travelers. This dataset will also allow us to see the busiest and least busy airline carriers. Both pieces of information can be a crucial planning factor in determining when to travel and where. Finally, we will also be able to visualize the historical changes in travel trends for the past 3 decades and notice some interesting changes and we will build a predictive model that will take all of this data into consideration and predict the number to traveling passengers in the future.

**1. Introduction**

We chose this topic because we know there is a lot of information out there regarding flight data and statistics, but we wanted to put together some more useful information That would present facts many of us already suspect or have a clue about, but we wanted to present these in an easy-to-understand visual format. If you google for example the “busiest airports in the U.S.” you will most likely see a list that shows you the largest airports instead of the busiest in terms of passengers. Second on the list of searches will most likely be an article based on misleading facts from the first article, but this will only show you the busiest airports within the past year. We felt that viewing only the past year of travel data would not really do justice to the actual numbers you hear about in the news when airports are their busiest.

We wanted to really take advantage of our newly found knowledge of how to manipulate large data and use that to our advantage. By analyzing a large dataset, we were able to capture passenger numbers for every U.S. airport for the past 3 decades and see where they went and when they went. We were even able to take advantage of the large scale of the data narrow down our findings to see what the busiest airlines were.

We felt this topic was important because travel is something everyone has done or will do at some point in their life, and it is both cost effective and wise to have advanced knowledge of what to expect when traveling especially if traveling internationally. With recent events and COVID restrictions soon being lifted we foresee travel numbers going back to normal very soon. With everyone being so eager to travel it will be beneficial to plan and save yourself the money and/or the headaches. The data we present can help in doing just that. Plan.

The direction of our project was based solely on the numbers and making our visuals as interesting as possible, that anyone who has traveled recently would be able to nod in agreement. Our dataset contained many columns, but we highlighted those that we felt would be most interesting and useful to visualize and use in our predictive model which will be discussed in more detail later. We chose columns such as *date*, *U.S. airport name*, *foreign airport name*, *carrier name*, and *total* number of passengers in each flight as those that would drive most of our analysis. Then we thought about all possible ways to represent this data and which of those we would find interesting if we were outsiders viewing this report. We then proceeded to breakdown our totals by years to see an overview of the trend over the past 30 years and how this has changed and finally we put together detaliled representations of the more recent numbers our dataset provided.

2. Related Work

Some related work that we found was from ACI. ACI is Airport Council International. They had done research on arrival and departure of flights around the world in 2019. According to their list the busiest of airports were Chicago O’Hare International Airport (ORD), Hartsfield–Jackson Atlanta International Airport (ATL), Dallas/Fort Worth International Airport (DFW), Los Angeles International Airport (LAX), Denver International Airport (DEN), Beijing Capital International Airport (PEK). That is data collected only in 2019. In contrast, in our data set we were able to conduct research on 1990 - 2020. From 2016 to 2019 the busiest airports according to our dataset were LHR-Heathrow, United Kingdom, YYZ-Mississauga, Canada, CUN-Cancun, Mexico, MEX-Mexico City, Mexico, CDG-Paris, France, NRT-Narita, Japan. That was an interesting find. Another related work we found was from the Airport Trade Organization. They had listed the busiest airports from 2020. According to Tulsa world the top busiest airports in 2020 were: Guangzhou (CAN), Atlanta (ATL), Chengdu (CTU), Dallas/Fort Worth (DFW), and Shenzhen (SZX). Compared to our research, we had busiest international airports from 2010 - 2020. According to our dataset the busiest among that time were: YYZ-Mississauga, Canada, LHR-London Heathrow, United Kingdom, MEX-Mexico City, Mexico. Even though the time periods were close the results were not remarkably close. The results completely differ. We had a lot of historical data, which helped us prepare a list from multiple years.

According to an article published by Skyscanner they gathered information for the year 2020 on when is the best time to book a domestic and international flight. Their research relates to our topic because they are as well focusing on helping the viewers with both type of traveling options. They concluded that the best time to book a domestic flight would be January and the worst time of the year to book a domestic flight would be May. The best time of the year to book an international flight from the US would as well be January and the worst time of the year would be May. Not only did they leave off there, but they were also able to gather results on what day of the week would be best to book a flight and that day would be a Saturday. Another research that relates to our paper would be an article that was published by The Wall Street Journal. They wrote an article that helps travelers know which airline the best is to travel with domestically. The gathered results were that Southwest out beat data by having a punctual arrival time of 84% compared to 79%. The worst airline to travel with is Allegiant because they do not worry a lot about customer satisfaction. These two articles relate to our project subject because they focus on ensuring that everyone can get a good deal out of a travel experience.

Although the two articles do relate to our project they differ from our topic because we were able to provide a more reliable answer. We had more data field to work with rather than just a year or two. We had more factors included in our research such as months, dates, airlines and passengers. While our related work articles only based their results solely on their specific criteria meaning only on the airlines or the prices. The numbers provided to us by our data file gave us the chance to expand our project past domestic results and we were able to get a more coherent look of airline travel internationally as well. When putting the numbers in GitHub we were able to give a clearer image to certain aspects of our project. For instance, when we were trying to show the results for only airlines, we were able to do so. This was a big benefit to us because it helped us give results to what we were looking for.

3. General Instructions

We first wanted to understand all of the data within our data set. The best way to do so was to begin visualizing the numbers we wanted to see for all years contained within our data set. This would highlight interesting patterns and how these have changed over the past 3 decades. Figure 1. below shows in green the total number of passengers who traveled between 1990-200, in lavender the number of passengers that traveled between 2000-2010, and finally in darker purple he number of passengers that traveled between 2010-2020. Then in order of most heavily traveled from to least we see that JFK New York and LAX in Los Angeles were the busiest airports regardless of the time period analyzed.

Chart

Description automatically generated

Figure 1. Top 10 Busiest U.S. Airports 1990-2020 (by passengers)

Then we looked at the busiest world airports utilizing the same metrics as previous figure. Three separate buckets, one for travels occurring from 1990-200, 2000-2010, and 2010-2020. Visualizing our data in this format allowed us to see that the busiest foreign airports have remained the same for the past 30 years which are Heathrow in UK and Mississauga in Canada and seen in Figure 2 below.

Chart, bar chart

Description automatically generated

Figure 2. Top 10 busiest world airports 1990-2020 (by passengers)

To truly be able to make a recommendation on which time is best to travel we decided to analyze the past 4 years' worth of data. Not to include 2020 since this was not a normal traveling year. We narrowed down our analysis to 2016-2019 to capture the most recent international traveling trends that were closest to what we can expect today. Figure 3. Pie chart below shows the top busiest U.S. airports in terms of passengers flying out on a given day JFK in New York and LAX in Los Angeles being the top two. Additionally, the Figure 3 below shows that the top two airports remain unchanged since the past 3 decades.

Chart, sunburst chart

Description automatically generated

Figure 3. Top 10 Busiest U.S. Airports 2016-2019

Next, we sought to drilldown our data to find out who were the busiest and least busiest airlines.

Figure 4. below shows the output of our Dashboard detail containing the top busiest airlines from 2016-2019 concluding that American Airlines and United Airlines were the busiest airlines for the past four-year period of our analysis. Each airline transporting between 108,000,000-115,000,000 passengers for the duration of the four-year analysis period. We can also conclude from the below dashboard that for the most past American Airlines and United Airlines have dominated in number of passengers during most of the four-year period. But from the pie chart in our dashboard, we can see that as of 2016 a third airline started gaining numbers and became more of a competition for these top two airlines. This third airline is Delta Airlines. Who as highlighted in blue below had low numbers initially but then for the following three years came close to matching the numbers of United Airlines.

Graphical user interface, chart, application

Description automatically generated

Figure 4. Busiest Airline Detail Dashboard 2016-2019

We then wanted to create a visualization that showed in a map where the highest concentration of passengers was. This step required creating a small sub data set from our already existing large data set. In Kibana we created a simple report that would extract in the form of a table the top 25 busiest airports for U.S. and the top 25 busiest world airports from 2016-2019. The output of this report gave us fifty-line items and two columns. One column recording the name of the airport another recorded the sum of passengers for each airport during our date range. In order, to create a world map we needed to obtain coordinates (latitude and longitude) for each location. We exported the report as a csv file and looked up the coordinates for each airport ourselves. This data can now be located within our GitHub small data set [3]. Once we had a csv file containing four columns airport name, sum of passengers, latitude, and longitude we uploaded this data set as an index in Kibana to create our aero spatial heat map visualization shown in Figure 5 below. Which further confirms what we have already stated about the top busiest airports.

Map

Description automatically generated

Figure 5. Top 50 Busiest U.S. and world airports

One of the main questions we wanted to answer at the start of our project was which months were the busiest to travel. By putting together, a simple line graph such as Figure 6 below that displayed the total number of passengers each month, we were able to easily answer this question. We ran this same line graph for only 2016, then for 2017, 2018 and finally for 2019. Each year displayed nearly similar results but constantly showed the busiest international travel month was always July and the least busing being February. Understandably July is the beginning of the summer, and everyone wants to travel and alternatively February is shortly after the holidays so most people are home saving for that July vacation. We also feel weather may be a contributing factor but for that we will need to analyze weather data. It is, however, common knowledge that beginning March and into late summer is the ideal weather for both London and Canada, which is where our analysis found most tourists were headed.

Chart, line chart

Description automatically generated

Figure 6. Travel totals by month for 2016-2019

Finally, we utilized Azure Machine learning to build and test a model that could predict the total number of passengers using the same large data set we previously analyzed in depth. We are very confident that this model could predict our total number of passengers accurately because as shown in Figure 7 below our coefficient of determination was 0.80 and the closer, we are to 1 the better results our model will produce. Our Relative Squared Error was 0.19 and the closer this number is to 0 the better results we will obtain.

Graphical user interface, application

Description automatically generated

Figure 7. Azure Predictive Model

**4. Conclusion**

In closing, we feel that the work we have put together is both interesting and important because we were able to analyze a large data set to see travel logs and identify the busiests airports in the U.S. as well as world airports. In doing so, we looked at what these trends have been for the past 30 years and seen that although travel has significanly increased over the past years, the most popular airports pretty much remain the same which are the major metropolitan areas of both the east coast and west coast of the United States. We were then able to narrow down our reports to only show data that has taken places over more recent years in order to more accurately attempt to pedict which is a better time to fly. This work is importat because we used data that is publicly available and truned this into useful information that we can now utiliza to show others and assist them in travel planning to avoid crowds at airports or perhaps help in choosing a destination that isnt as popular for a nice relaxing vacation.

In completing this project we learned that July is always the time of the year where most people travel and February is when the least number of passengers are leaving for international flights. We also learned that LAX our airport remains and has been one of the busiests airports for the past 30 years and these numbers do not look like they are about to chahnge.

### References

[1] Data Source URL: [https://www.kaggle.com/parulpandey/us-international-air-traffic-data](https://www.kaggle.com/parulpandey/us-international-air-traffic-data%E2%80%8B)

[2] GitHub URL: [https://github.com/rcueva12/CIS3200-Data\_Processing\_Analytics/tree/main](https://github.com/rcueva12/CIS3200-Data_Processing_Analytics/tree/main%E2%80%8B)

[3] Small dataset for Maps URL: <https://github.com/rcueva12/CIS3200-Data_Processing_Analytics/blob/main/airports4.csv>

[4] Azure URL: <https://studio.azureml.net/Home/ViewWorkspaceCached/17993356d8fa48389e9a5796ceb6afcb#Workspaces/Experiments/Experiment/17993356d8fa48389e9a5796ceb6afcb.f-id.1bc20085248a4dba86f25479eb229049/ViewExperiment>

[4] Kibana URL: <https://e4370728e6b04da7a2f0204627978298.us-central1.gcp.cloud.es.io:9243/app/home#/>

[5] Article: <https://www.internationalairportreview.com/article/110871/top-20-busiest-airports-world-aircraft-movements/>

[6]Article: <https://tulsaworld.com/travel/worlds-busiest-airport-in-2020-was-in-china-not-atlanta/article_055e116d-c0d6-5042-929f-0b624b33e111.html>

[7]Article: <https://www.skyscanner.com/tips-and-inspiration/best-time-to-book-usa>

[8]Article: <https://www.wsj.com/articles/the-best-and-worst-u-s-airlines-of-2020-11611756016>