the United States National Civilian Aviation Network Statistical Research

Advanced Data Science Final Project Report

Class: CSCI 4022
Professor: Zach Mullen
Date:12/16/2017

Group Members:
Keyu Wu, Han Yin, Ziyue Guo

Nowadays, taking the plane to travel and to work has become increasingly common in people's lives, but people always meet troubles during the flight such as flight delay, terrible service, and unpleasant flight experience. Such phenomenons are largely attributed to the airport and the air company. Therefore, ranking the airport and the airlines around the USA can help people from all over the world to get the best flight experience. In the meantime, the ranking report of the airport and air companies also contributes to the enhancement of air companies. In this project, we are going to do data mining about the United States' flight information. By exploring the flight information we can solve many doubts about American airports. For instance, what are the most important airport in America? How to rank American airports? What is the best route between the two locations? What is the best US Air company? and what are the important factors that affect customers' flight experience? Besides, we expect to visualize the United States airline network in various ways as well. So the user can have more clear and intuitive feedback about our report. Furthermore, since the dataset is about the flight delay, it is meaningful to explore the factors that influence the flight delay, there are many potential answers such as the flight distance, the influx, and the outflux of the airport. Being familiar with the reasons that cause flight delay helps customers to avoid such an awkward situation.

Our datasets comes from the U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics that tracks the on-time performance of domestic flights operated by large air carriers, Summary information on the number of on-time, delayed, canceled, and diverted flights is published in DOT's monthly Air Travel Consumer Report and in this dataset of 2015 flight delays and cancellations. If possible, our report readers can go to https://www.kaggle.com/usdot/flight-delays/ to download the dataset. When we deal with the data, we only use the flight.csv, but the users who download the dataset can have a better understanding of our report by reading the other two CSV files. One thing to focus on this dataset is the airport name when our team explores the data, we found that some airports are named by the number, these airports may belong to the military and it is useless for our project, and there are some NAN data inside the data frame. So we clean the dataset and delete the rows that contain numerical airport.

In the real world, there are plenty of media trying to rank the airports but with different methods, I want to talk about two representative company that tries to solve the problem of airport ranking, the first company is CNBC, they got the same result as us that Hartsfield-Jackson Atlanta International Airport (ATL) is the most efficient airport in the US. Another company is the AirHelp, they collected the data from the customer feedback and their final score is the average score based on On-Time performance, Service quality, Food, and Shops. And their best US airport in the US is ATL as well. In the academic area, some scholars do research on the same topic as we do. For example, Volte-Dorta (2017) analyze the European Airport Network, by doing so to illustrate the vulnerability of European Airports. Compared with our team, he used the completed different method, according to his paper, he illustrates that
"First, we generate a baseline scenario by assigning a flight combination with minimum travel time for each observed passenger itinerary. Second, we simulate the closure of one major airport and the affected passengers are sequentially relocated in

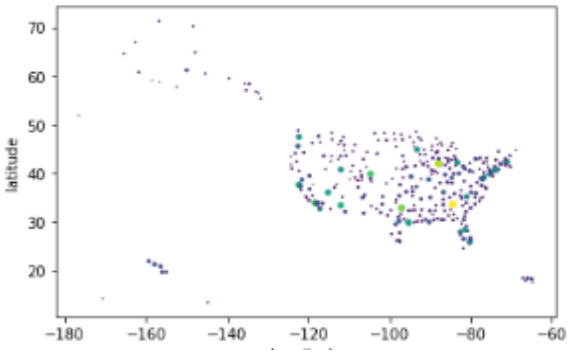
delay-minimizing alternative itineraries within a predefined network, taking into account capacity constraints and allowing for ground transfers between airports. Third, the aggregate delay is used as a measure of network damage generated by the particular airport closure."(Volte-Dorta, 2017)
Even though our method and Volte-Dorta's method are different, Volte-Dorta's idea reminds us that this topic not only shows the importance of the airports but also demonstrate the vulnerabilities of the airport network. That makes us more confident to explore on this topic.

We use PageRank is the first method we come up in mind. During the analyzing process, we used 322 airports listed in airports.csv as nodes and 5,819,079 fights from January 1st to December 31th as links to perform Page Rank to get the vector of importance for each airport. This data frame we are dealing with is extremely large and it takes a long time to run our code. Luckily we get valid some outputs.

Here are the results. As you can see, ATL, ORD, DFW, DEN and LAX become the top 5 airports in the US, And on the graph. It is convenient to see the airport density no matter it is regarded as the original airport or the destination. Data in the graph are presented in descending order.

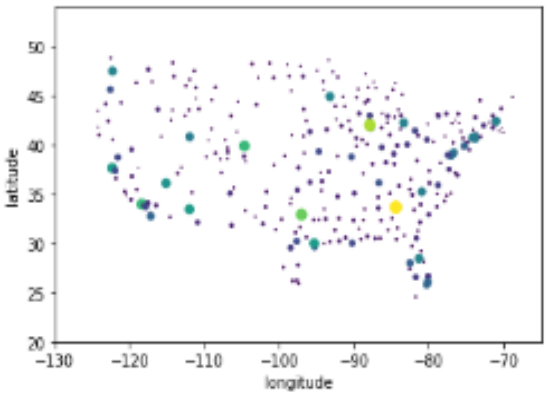
	Ori_Density	Des_Density	PageRank
ATL	0.065036864	0.065049615	0.065053897
ORD	0.053607465	0.053611590	0.053612750
DFW	0.044919344	0.044925157	0.044923674
DEN	0.036763203	0.036754765	0.036753724
LAX	0.036504058	0.036508370	0.036506243
SFO	0.027753682	0.027745807	0.027744547
PHX	0.027529977	0.027529415	0.027528862
IAH	0.027493787	0.027505225	0.027505282
LAS	0.024973401	0.024976589	0.024975433
MSP	0.021023590	0.021025653	0.021025704
MCO	0.020810761	0.020810386	0.020810701
SEA	0.020795198	0.020796698	0.020796021

After getting the rank of the airport, our team utilizes the longitude and latitude information for each airport in the data frame to generate a 2-dimensional map of airports. Then we include a mathematical algorithm to find distances between airports. By combining these data, we geometrically visualize the importance of the airport and its location on the 2- dimensional map.



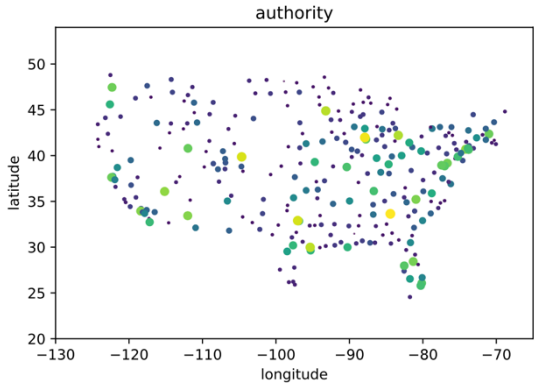
P1: Plot of airport importance

The importance of the airport is revealed by the size of the point, here is the graph. In order to make this graph more intuitive and comprehensive, our team deals with the outlier by restricting the scope on the x-axis and y-axis between N20 ~ 54 latitude and W65 ~ 130 longitude (e.g. ignore the airports in Alaska due to the distortion of projection at high latitude). The final graph looks like the graph below.



P2: Plot of airport importance (only mainland US)

Furthermore, we also implemented Hubs and Authority algorithm. This is an algorithm very similar to PageRank but in more detail since it divided information into 2 parts, and what we care more is the authority core. Here is the authority output:



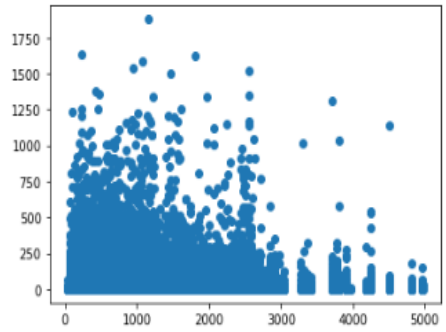
P3: Plot of airport authority score (only mainland US)

In case one is curious about what hubs scores looks like, the hubs score plot look almost the same as the authority score plot since for almost every single air route connects 2 airports, the flights are commutative.

By far, there are a lot of columns of data that are not used yet, for example, the delay of each individual fights, times each fight take, fight companies, delay or cancelation reason, etc. We would further explore these data and apply more data science methods. To evaluate the airport, the flight delay time and delay frequency of the airports are essential factors. Therefore, our team count the delay ratio of each airport and here are the results of several airports' total delay time as either the original airport or the destination.

```
ORD 35778 285884
ATL 26397 346836
DFW 24684 239551
DEN 22720 196055
LAX 20904 194673
IAH 16416 146622
BOS 14209 107847
LGA 13882 99605
PHX 13626 146815
SFO 13551 148008
LAS 13150 133181
```

At the beginning of the analysis, we wonder what factors influence the delay time. So we analyze the relationship between the flight distance and the delay time, we build a graph with flight distance as x and delay time as y, the overall graph looks like below. If we only look at the graph of the point distribution, we may think the smaller the flight distance is there will be more frequent flight delays. However, according to the Pearson test, the Pearson correlation is only 0.012, and the slope is extremely low, only 0.0007752 indicating that it is unlikely the flight distance will influence the delay time.

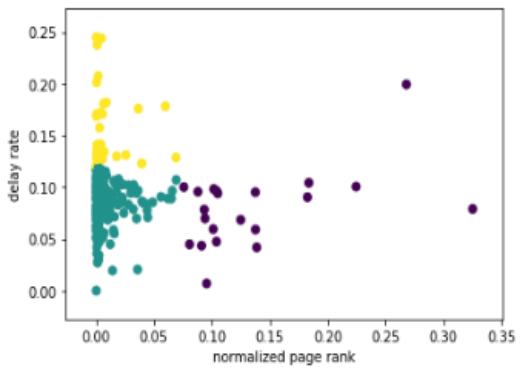


P4: Plot of flight distance (mile) vs. delay time (min)

After finding there is almost no relationship between the delay time and the flight distance, we calculate the flight rate based on the number of flight delays and the number of total flights. But the result of the relationship between flight delay time with flight distance and relationship between

delay rate and flight distance is almost the same, we draw the conclusion that there is almost no relationship between flight distance and flight delay. But our exploration will not end.

Since our original graph that shows the importance of the airport around the US does not consider the delay time, we come up with new ideas to combine the PageRank that we compute with the delay time of each airport. In this part, our primary method is K-means clustering. The reason we choose K-means clustering is that clustering the airport can more clearly indicate the features of the airport. For example, one airport may be very efficient but have relatively more flight delays, another airport may be inefficient but they hardly have any flight delays. To make the result become generally comprehensive and fit in the corresponding data, we do the normalization of the delay ratio by multiplying 5 in advance. In our case k = 3, the three clusters represent high page rank, lower page rank with high delay rate, lower page rank with low delay rate respectively.

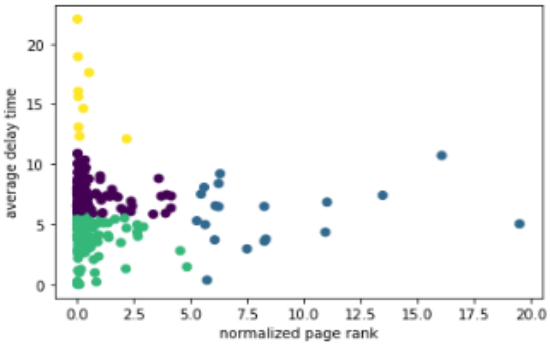


P5: 2 factor K means (k = 3) clustering (delay rate and normalized page rank)

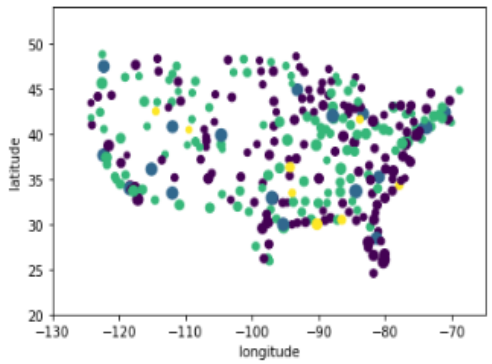
Now the data of our Airport can have the numerical label based on its features. This kind of label can also be represented on the 2-dimension graph we created before by assigning the airports that have the same label with the same color.

P6: color labeled mainland US airports using data from P5

Now anyone who lives in a relatively smaller country can also choose a better airport when they go out. Then we do the K-means clustering again for the average delay time and the normalized PageRank, but this time we use k=4 because of there exit low page rank airport with extremely high average delay time. And we created another US map to illustrate the result.



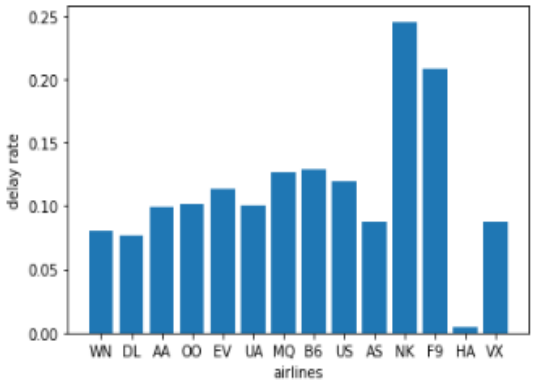
P7: 2 factor K means (k=4) clustering (average delay time and normalized page rank)



P8: color labeled mainland US airports using data from P7

The new US map is more suitable for people who cared more about the arrival delay time than the delay rate. Compared with average delay time vs. normalized page rank graph and delay rate vs. normalized PageRank graph, their shape is almost the same, showing that the delay time and the delay rate of each airport are closely associated.

So far, the rank of the airports around the US has been pretty obvious, a new question comes out, what is the US best Airline company? For everyone, they may have different answers based on their flight experience, air companies' food quality and ticket price. But for us, the determinant of the air company is the delay time and the flight details. And of course, we will rank the airline company in technical ways. Here is a bar graph of the US air companies' delay rate.

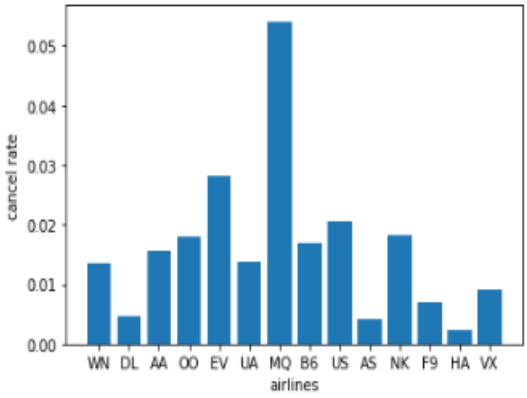


P9: delay rate of various air companies

Here is the reference graph of the integral flight company name of the abbreviation air company name.

	IATA_CODE	AIRLINE
0	UA	United Air Lines Inc.
1	AA	American Airlines Inc.
2	US	US Airways Inc.
3	F9	Frontier Airlines Inc.
4	B6	JetBlue Airways
5	OO	Skywest Airlines Inc.
6	AS	Alaska Airlines Inc.
7	NK	Spirit Air Lines
8	WN	Southwest Airlines Co.
9	DL	Delta Air Lines Inc.
10	EV	Atlantic Southeast Airlines
11	HA	Hawaiian Airlines Inc.
12	MQ	American Eagle Airlines Inc.
13	VX	Virgin America

According to the graph, Hawaiian Airlines is the air company with the lowest delay rate, on the contrary, Spirit Airlines is the air company with the highest delay rate, so people who are serious with the timeliness should try to avoid taking Spirit Airlines when they buy the tickets. Moreover, there is another way to rank the air company by utilizing the cancellation rate.



P10: cancelation rate of various air companies

According to the graph, American Eagle Airlines has the highest cancellation rate, and the record of MQ is much higher than other airlines, MQ should figure out how to solve this serious problem. Hawaiian Airlines has the lowest cancellation rate. However, we cannot say the Hawaiian airport is the best airline because HA has a much lower number of flights than other airlines. One thing we are sure of is that American Eagle Airline and the Spirit Airlines need to improve urgently.

In conclusion, throughout the report, we have successfully answered most of the problems that we initially wonder. For instance, ATL is the most efficient airport in the USA, the worst air company based on delay time and cancellation rate are Spirit Airline or American Eagle Airline. The importance of each airport is shown on the graph. The flight distance doesn't affect the flight delay and the flight delay time is associated with the flight delay rate. The whole American airports are clustering into 3-group and 4-group by instinct features. But we still have plenty of work to do due to the time constraint and technical restrictions. We still didn't answer what is the best route to transfer for the customers, even though we can use dynamic programming to find the shortest path and find the best route, but the combination of the route of all airports is too large to deal with, the data in the original CSV file is millions level, the complexity of the algorithm need to be largely reduced even with dynamic programming. During the process of doing the project, we gradually realize, while we are working on real-life problems, the knowledge on only one field is not enough, we have to combine all methods from many classes such as Advanced Data Science, Machine learning, Algorithms to overcome the obstacles. Besides, we rank the American airport based on the network and flight information, but this is not enough, the airport quality is determined by many other factors such as the traffic, the service, the accident rate, and the crews' abilities. The single standard that we evaluated needs to be strengthened and improved.

References:

Voltes-Dorta, Augusto, et al. "Vulnerability of the European Air Transport Network to Major Airport Closures from the Perspective of Passenger Delays: Ranking the most Critical Airports." Transportation Research Part A, vol. 96, 2017, pp. 119-145.

AirHelp https://www.airhelp.com/en/airhelp-score/airport-ranking/

Kaggle https://www.kaggle.com/usdot/flight-delays/

Wall Street Journal
https://www.wsj.com/articles/the-best-and-worst-u-s-airports-of-2019-11573658675