How to Plan Your Trip?

An Analysis of Flights Data from 1989 to 2008

Final Report for CMSC 12300 – Group: Hide on Bush

Group Members: Yaling Xu; Siyuan Peng; Zeyu Xu; Dongcheng Yang

June 2019

1 Introduction

Flight delay has been one of the important problems that deserves attention for not only carriers but also passengers. In the U.S., more than 3 million flights are delayed or cancelled due to several reasons such as weather, National Airspace System (NAS) control and security.

However, air traffic become over concentrated at a relatively small number of airports, and particularly at certain times of the day. Of 450 airports handling commercial flights in the USA, the top 60 have 94 percent of the traffic and the top 20 control over half the traffic. For example, inclement weather in the summers of 1999 and 2000 caused delays not only at airports experiencing the inclement weather, but at airports with flights connecting from the airports experiencing inclement weather.

Numerous delays result in high economic cost. A significant number of passengers have suffered in choosing probable flights and airports. According to a report released by the US Federal Aviation Administration, the economic cost of flight delays on passengers, airlines and other parts of economy is over 31 million in 2015. Most of the cost imposes on passengers who not only waste time waiting flights, but only lose time and money on food and hotels until next available flights.

Previous researchers have analyzed the delays and cancellations from different perspectives. The FAA found that current system for collecting causal data doesn't provide strong conclusions on improving the recommendation system. Barnhart and Bratu (2001) studied the percentage of passengers waiting in the connecting hubs and pointed out general delay statistics didn't solve the issues of passengers missing connecting flights within the context of delays and cancellations.

Existing studies mostly analyze flight delays causes and delay propagation. Few studies focus on the connecting flights and possible reasons considering flights overlapping areas. In order to fill this gap, this paper contributes to the analysis of delay from three perspectives:

First, we explore the flight delay patterns in U.S.. Specifically, we use data visualization to demonstrate delay condition in major airports and carriers. Then we examine more in details on the delay causes. Most importantly, maps are produced in order to generate a geographical overview of the flight condition in the past 20 years. Second, we use historical data to generate optimal flight recommendation for travelers. If two places have direct nonstop flights, flights that have lowest average delay time will be provided. If connecting flights are the only choices, a UI interface is designed to provide optimal flights given original airport, destination airport, minimum and maximum connecting time. Finally, we explore delay cause from a brand new perspective. We assume that bad weather in flights overlapping area may affect all flights that go across this area. Thus we pair all flights and endow weights for each overlapping area. The overlapping areas on the subsequent flights are compared and possible scenarios are designed to examine the hypothesis.

2 Preamble: Overview of the Dataset

Data Source:

https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time

This dataset comes from the Flight Delays and Cancellation Dataset from Bureau of Transportation Statistics, containing approximately 726,919,200 lines of records, 28 variables in total. The data set focuses on 18 carriers and over 200 airports in U.S. domestic market from 1989 to 2008.

This dataset is composed by the following variables:¶

1. Year: 1989-2008
2. Month: 1-12
3. DayofMonth: 1-31
4. DayOfWeek: 1 (Monday) - 7 (Sunday)
5. DepTime: actual departure time (local, hhmm)
6. CRSDepTime: scheduled departure time (local, hhmm)
7. ArrTime: actual arrival time (local, hhmm)
8. CRSArrTime: scheduled arrival time (local, hhmm)
9. UniqueCarrier: unique carrier code
10. FlightNum: flight number
11. TailNum: plane tail number: aircraft registration, unique aircraft identifier
12. ActualElapsedTime: in minutes
13. CRSElapsedTime: in minutes
14. AirTime: in minutes
15. ArrDelay: arrival delay, in minutes: A flight is counted as ”on time” if it operated less than 15 minutes later the scheduled time shown in the carriers’ Computerized Reservations Systems (CRS).

16. DepDelay: departure delay, in minutes

17. Origin: origin IATA airport code

18. Dest: destination IATA airport code

19. Distance: in miles

20. TaxiIn: taxi in time, in minutes

21. TaxiOut: taxi out time in minutes

22. Cancelled: was the flight cancelled

23. CancellationCode: reason for cancellation (A = carrier, B = weather, C = NAS, D = security)

24. Diverted: 1 = yes, 0 = no

25. CarrierDelay: in minutes: Carrier delay is within the control of the air carrier. Examples of occurrences that may determine carrier delay are: aircraft cleaning, aircraft damage, awaiting the arrival of connecting passengers or crew, baggage, bird strike, cargo loading, catering, computer, outage-carrier equipment, crew legality (pilot or attendant rest), damage by hazardous goods, engineering inspection, fueling, handling disabled passengers, late crew, lavatory servicing, maintenance, oversales, potable water servicing, removal of unruly passenger, slow boarding or seating, stowing carry-on baggage, weight and balance delays.

26. WeatherDelay in minutes: Weather delay is caused by extreme or hazardous weather conditions that are forecasted or manifest themselves on point of departure, enroute, or on point of arrival.

27. NASDelay in minutes: Delay that is within the control of the National Airspace System (NAS) may include: non-extreme weather conditions, airport operations, heavy traffic volume, air traffic control, etc.

28. SecurityDelay: in minutes: Security delay is caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas.

29. LateAircraftDelay: in minutes: Arrival delay at an airport due to the late arrival of the same aircraft at a previous airport. The ripple effect of an earlier delay at downstream airports is referred to as delay propagation.

2.1 Data Cleaning and Preprocess

Since time-related variables are inconsistent in the original dataset, we compiled and converted them to standard data frame using the datetime that python provides internally. Basically, we combined dates and times to datetime.time format. We wrote few functions to run the transformation.

In addition, we noticed that DEPARTURE_TIME and ARRIVAL_TIME variables are misleading since they didn't provide the date information. Thus we used DEPARTURE_DELAY and ARRIVAL_DELAY variables instead since detailed delay minutes are provided.

Finally, we cleaned the data frame and reorganized the variables for readability.

2.2 Descriptive Analysis

2.2.1 Understanding the Data

We import the data and show the correlation matrix in order to present the multicollinearity among 29 variables (Figure 1). Few interesting observations are found.

Only when delay time is longer than 15 minutes, there is data to describe the delay reasons. That is due to the fact that a flight is counted as "on time" if it operated less than 15 minutes later the scheduled time shown in the carriers' Computerized Reservations Systems (CRS). Arrival performance is based on arrival at the gate. Departure performance is based on departure from the gate. Arrival Delay is the sum of CarrierDelay, WeatherDelay, NASDelay and LateAircraftDelay. In cases of cancelation or diversion there is no data related to delay causes.

CRS Elapsed Time is usually higher than the Actual Elapsed Time (actual time spent in the Taxi In + Taxi out + Airtime operations) since airport and carriers will allocate a longer CRS Elapsed Time in order to absorb late flights.

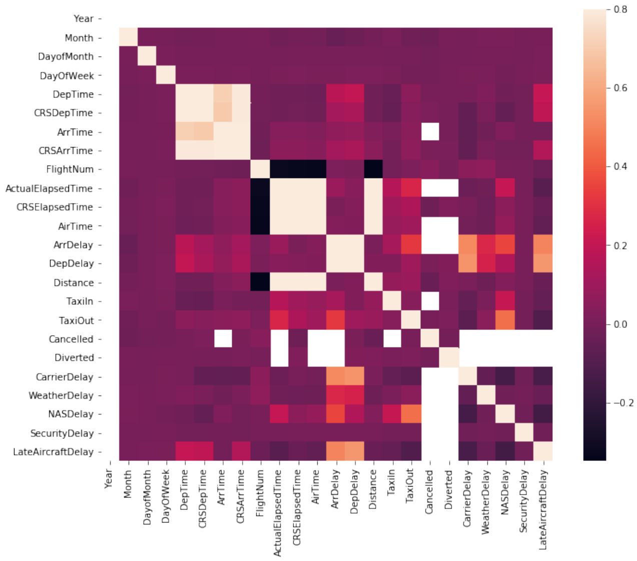


Figure 1: Correlation Matrix

2.2.2 Delay status

We divide delay status into five conditions: flight was on time or delayed less than 15 minutes (0), slightly delayed for delayed more than 15 minutes but less than 1 hour (1), highly delayed for delayed more than one hour but less than 2 hours (2), diverted (3), or cancelled (4).

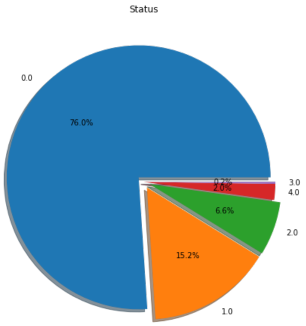


Figure 2: Delay Status

From the bar and histogram plotted, we could see that 76% flights were on on time or delayed less than 15 minutes while more than 20% delayed more than one hour. And only 2.3% flights canceled or diverted.

2.2.3 Cancellation Analysis

We analyze the cancelled flights (0 = carrier, 1 = weather, 2 = NAS, 3 = security) and find that more than 40% cancellation are due to the weather. In other words, it is caused by extreme or hazardous weather conditions that are forecasted or manifest themselves on point of departure, enroute, or on point of arrival.

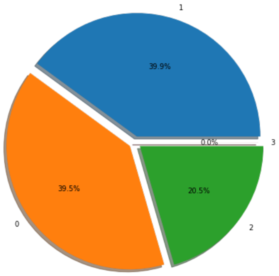


Figure 3: Cancellation Status

2.2.4 Delay Analysis

Then we analyze the delay flights by month. We plot the average delay and the avearge number of minutes delayed by month. It is found that October has the lowest delay possibility and delay minutes. The majority of delays gather in February, June and December.

2.2.5 Overview of the Geographical Area

We plot the airport location and the average number of flights per year in the U.S. map in order to have a geographical overview of the flight condition in the past 20 years.

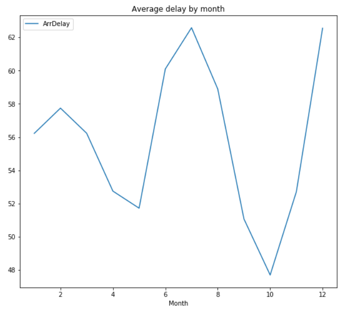


Figure 4: Delay Analysis

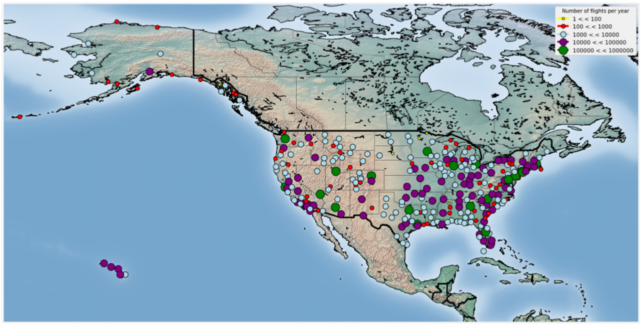


Figure 5: Number of Flights

We find that the busiest airports serving over 100 thousand flights per year mainly concentrate in the east and west coasts in the United States. Then we explore the delay condition of each airport in detail. Basically, we care more about the departure delay for the following two reasons. First, costumers usually arrive in the departure airport before the estimated departure time. Thus they have to wait in the airport if the flight is delay in departure. Second, if carriers are allowed higher speed on planes that departed late, the arrival delay could be decrease significantly. Moreover, we only consider flights that delays more than 15 minutes since short delay minutes are acceptable. In other words, we would use departure delay time that over 15 minutes to calculate the delay percentage for each carrier and airport. We plot the airport location and the average delay percentage for departed planes in this airport in the U.S. map.

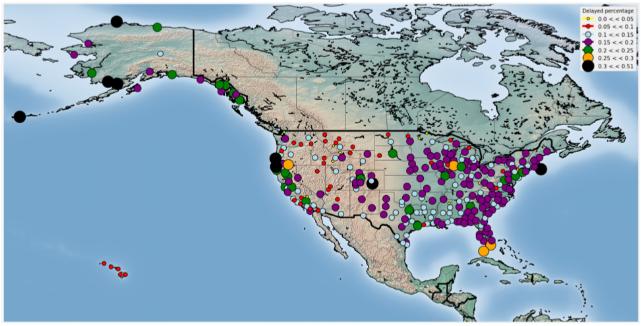


Figure 6: Delay Percentage

2.2.6 Carrier Delay: by carrier

Carrier delay is within the control of the air carrier, which reflects the operation ability and service capability. The growth and quality of carrier control is the valuable source for airlines. To improve the management and scheduling abilities of irregular flight has become one of most important work emphases of airlines' operations control management. Thus we want to examine more in details on the delay condition for each carriers. We plot the average delay and delay distribution by carrier.

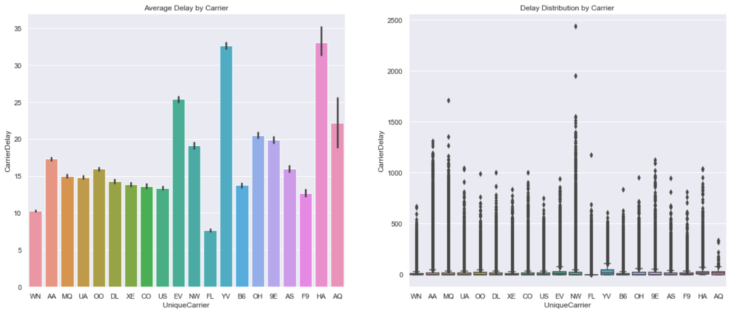


Figure 7: Average Delay and Delay Distribution by Carrier

AirTran Airways create the lowest carrier delay (7.8 minutes) per flight, and Southwest Airlines with 11 minutes per flight, is the second lowest among 18 carriers. ExpressJet and Continental Airlines also have performance that better than the average in the carrier delay. On the other hand, Hawaiian Airlines and Mesa Airlines generate the highest average carrier dalay that over 20 minutes per flight.

As we probe into the delay distribution boxplot, interestingly, carriers with smaller number of flights tend to have a higher carrier delay. A typical example is the Aloha Airlines. This finding implies that the volume of flights is an important factor affecting average carrier delay.

On one hand, large average carrier delay indicates high possibility of waiting. On the other hand, larger number of flights indicate larger group of customers facing extreme waiting. Considering the number of flights and average carrier delay together Southwest Airlines and AirTran Airways have outstanding performance regarding carrier delay.

2.2.7 NAS Delay: by airport

Delay that is within the control of the National Airspace System (NAS) that highly represents the airport operation ability. New York (JFK) and Boston (BOS) generate the highest NAS average delay that both over 25 minutes per flight. Whereas San Francisco (SFO) and Las Vegas (LAS) perform relatively better. However, there is no clear correlation between the flights volume and NAS Delay.

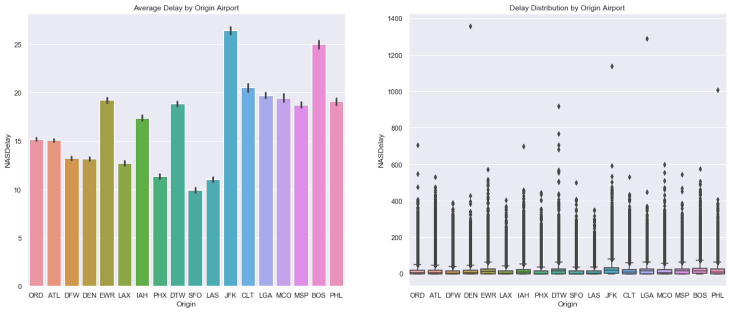


Figure 8: NAS Delay

3 Utilizing Crisscross points to Predict Weather Condition

Intuitively, extreme weather plays a significant role in flight delay and cancellation. As what we discussed above, over 40 percent flight delay is caused by extreme or hazardous weather conditions. However, most lines only cover major cities around the U.S., thus merely analyzing weather conditions in original and destination airports is not enough to understand causes of delay. Here, we assume that weather conditions along the airline routes will affect flights across this area.

Before we move into our analysis, we have few assumptions:

1. If the data can cover a significant amount of time, crisscross points of all flight pairs will almost cover the whole U.S.

2. We assume that extreme weather that affects flights would last a whole day, which means that all flights across this area in the same day would be influenced.

To begin with, we believe that our flight data from 1989 to 2008 is enough to cover the whole U.S and thus the first assumption is satisfied. By the second assumption, we are able to measure the effect of weather in crisscrossing area on flights that across it by day. For example, considering two flights on January 1, 2008. One flies from New York to Salt Lake City and the other one flies from Seattle to Miami. Both of the flights fly across Chicago and got delayed. We assume that the extreme weather in Chicago on that day leads to delay of both flights.

3.1 Key variables definition

1. Delayed flights: flights that are delayed over 30 minutes are defined as delayed flights.

2. Flights crisscross points: we simplify airline to straight line that from departure airport to arrival airport. Then we calculate the intersection of two paired airlines as the crisscross point.

3. Geographical position: Latitude-longitude identification method is used to tag geographical position. Thus, we are able to get geographic position for all airports and crisscross points.

3.2 Algorithm

At first, we pair all flights in one day and calculate the crisscross points. Following the same way, we calculate every day's the crisscross points in the span of 20 years. In this way, we can create a dictionary, where the key is the latitude-longitude coordinate, value is the frequency of this point.

Secondly, in order to map the frequency clearly in the map, we need some modifications. If the latitude-longitude coordinates are decimal, we round them up to the nearest integers. In this way, we could update the dictionary where all keys are integer coordinates and values are sum of frequencies in this area. For example, coordinate (111, 90) covers crisscross points whose latitude is between 110.5 and 111.4, longitude is between 89.5 and 90.4.

Furthermore, we divide the values in dictionary by 20, and thus we could get the average occurrences of each overlapping area per year.

Finally, we plot number of delayed flights in each overlapping area in the map.

3.3 Method

We use "Dask", which is the primary parallel processing library in python to increase running speed. The Dask delayed decorator stages operations that are to be parallelized. Any function Dask delayed touches become lazy and run later. Any data touched by Dask delayed will make instruction that calls the data to be delayed or lazy and run later. The Dask compute code runs those delayed objects parallelly.

3.4 Results

The final heatmap of our results is presented in Figure 9. Each grid in the heatmap has its corresponding measurement of extreme weather condition. Darker color in the heatmap indicates higher weights, or more extreme weather. In Figure 10, we visualized the average Regional Climate Extremes Index from 1989-2008. The data source is National Climate Data Center. Comparing the heatmap and the Climate Extremes Index, we could see that they are quite similar to each other. Especially, we have great estimation in the North West, Northern Rockies and Plains, Upper Midwest, Ohio Valley and North East areas of the country. The weather condition of West area is significantly different from North West part, which is also shown in Figure 10. It is therefore verified that we could use the crisscross point idea to infer extreme weather condition.

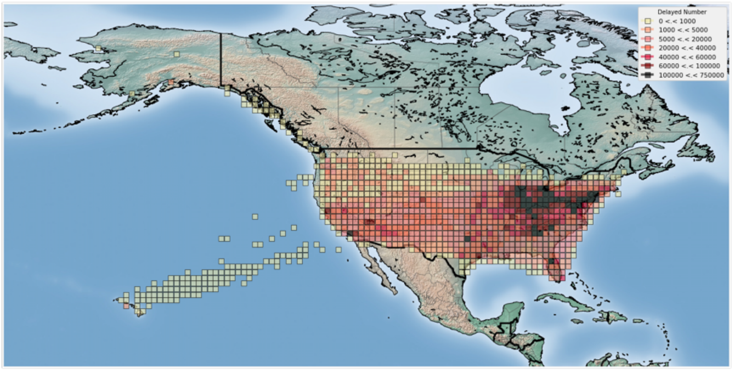


Figure 9: Heatmap

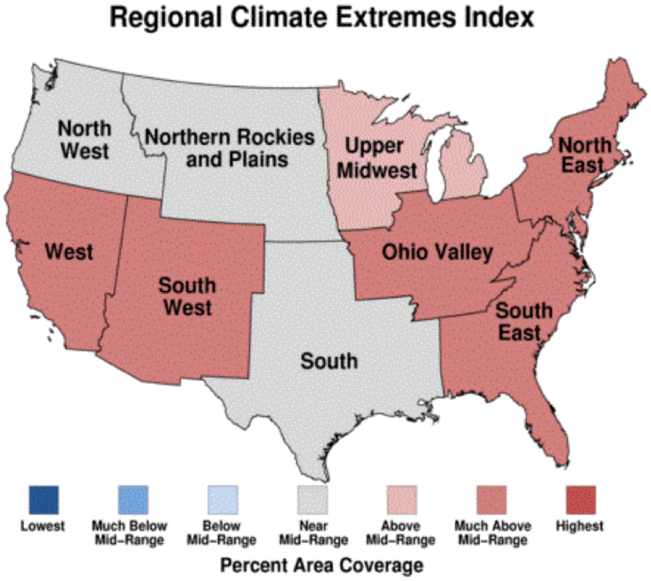


Figure 10: Climate Extreme Index

4 Recommendation System

The U.S. attaches great importance to aviation industry and it grows very fast in the past few decades. However, direct lines only cover major cities around the U.S., and connecting flights usually face flight delay risks. Consumers have to spend time and money on waiting for next available flights, and it costs more in a sprawling airport. In this part, we design a recommendation system for travelers using historical data from 1989 to 2008, and we present results in a search engine.

4.1 Algorithm

Due the lack of flights price data, we design the recommendation system mainly based on following variables: CRSElapsedTime (elapsed time shown in computerized reservations systems), DepDelay (departure delay, in minutes) and ArrDelay (arrival delay, in minutes).

4.1.1 Assumptions

1. We assume that consumers have no specific preference for carriers, departure and arrival time.
2. In terms of direct flights, we assume that consumers care more about the departure delays since they don't want to waste time waiting in airports. We rank direct flights only by departure delays.
3. As far as connecting flights, we assume that consumers consider the whole travel time. So we rank flights by the sum of departure delays, arrival delays and elapsed time.
4. We only consider one stop connecting flights.

4.1.2 Data Clear Algorithm

1. Get all airports from the data.

2. Select two airports as origin and destination, and then collect information on all possible flights between origin and destination to build Dictionary 1.

   Dictionary 1: Flight route data

   Key: string: Origin to Destination

   Value: Information of all flights from origin to destination (a tuple: (string: short name of carrier + flight number, planned departure time, planned arrival time))

3. Collect and take average of delay time for each flight. For records that share the same origin, destination, carrier and flight number, we regard them as the same flight. We build Dictionary 2 to store flight data.

   Dictionary 2: Flight data

   Key: string: {Origin} to {Destination} {Unique code of the carrier} {Flight number}

   Value: Average delay time (a tuple: (average departure delay, average arrival delay))

4. Based on the Dictionary 1, we build 2 lists. List 1 includes all direct flights between origin and destination. List 2 includes paired connecting flights between origin and destination. For each pair in the List 2, we search all airports for possible transit airports and build Dictionary 3 by checking whether there exist direct flights both from origin to transit, and from transit to destination.

   Dictionary 3: Possible transit airports

   Key: tuple: (Origin, Destination)

   Value: Transition airports (If there are flights from origin to transition airport, and flights from transition airport to destination)

4.2 Method

To get the information for each flight and divide them into different groups, we used MapReduce method with three mappers. The first mapper collected IATA codes for all airports in the database and generate possible (origin, destination) set. The second mapper collected the planned departure and arrival flights for each flight and divide them into different groups by their origins and destinations. The third mapper calculated the average departure and arrival delay time for each specific flight. Then, We built data-frame to store the information and wrote the algorithm.

4.3 Results

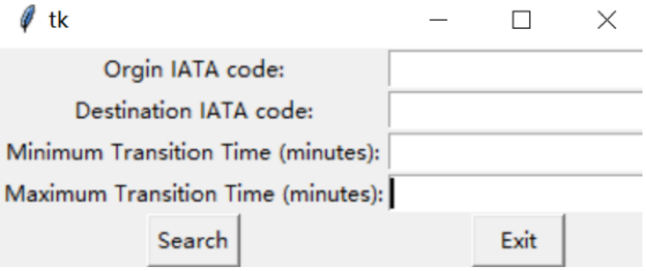
We design a simple and understandable executable for customers. This executable is simple and easy to use, allowing not only customized inputs, but also monitoring possible optimal flights. As shown in Figure 11, the user interface is clear and readable. We have four input boxes, which prompt the user to enter the enter IATA codes of origin and destination airport, minimum transit time and maximum transit time.

If the user doesn't finish all boxes or doesn't enter integer in time boxes, a warning message will be reported (Figure 12).

If the user enters two airports that have a direct flight, the executables will provide at most 3 optimal flights based on customized flight number, planned departure time and arrival time, as shown in Figure 13.

If the user enters two airports that don't have direct flight but exist connecting flights, the executables will provide at most three best optimal flights based on customized flight number, planned departure time and arrival time, as shown in Figure 14.

If the user enters airports that we can't find any route that satisfies both the transit time requirement and one stop requirement, a message will be reported as shown in Figure 15.



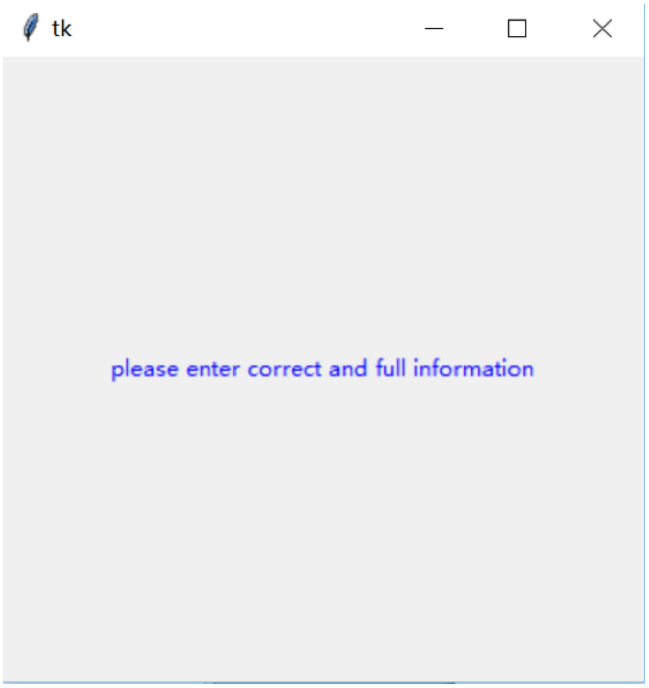
Orgin IATA code:

Destination IATA code:

Minimum Transition Time (minutes):

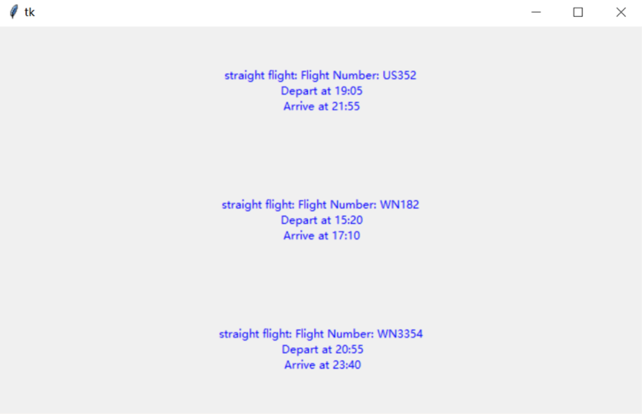
Maximum Transition Time (minutes):

Figure 11: User Interface



please enter correct and full information

Figure 12: Warning

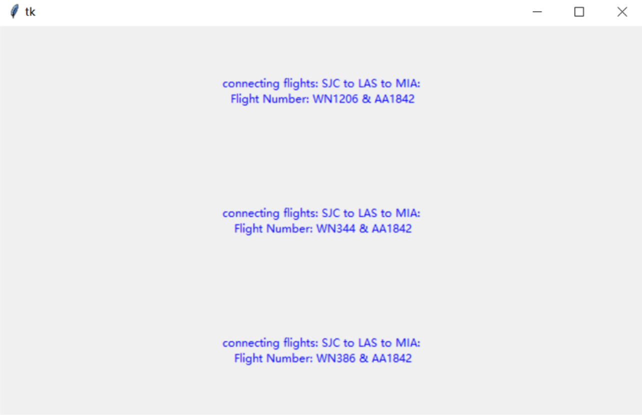


```
straight flight: Flight Number: US352
Depart at 19:05
Arrive at 21:55

straight flight: Flight Number: WN182
Depart at 15:20
Arrive at 17:10

straight flight: Flight Number: WN3354
Depart at 20:55
Arrive at 23:40
```

Figure 13: Direct Flight

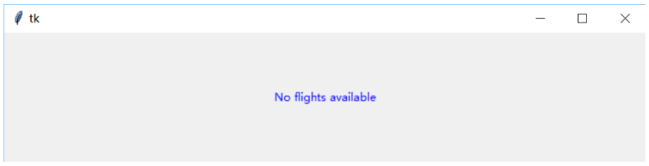


```
connecting flights: SJC to LAS to MIA:
Flight Number: WN1206 & AA1842

connecting flights: SJC to LAS to MIA:
Flight Number: WN344 & AA1842

connecting flights: SJC to LAS to MIA:
Flight Number: WN386 & AA1842
```

Figure 14: Connecting Flights



```
No flights available
```

Figure 15: No Data

5 Limitations

5.1 Weather Prediction

In the weather part, we assume that the airline routes are straight and calculate latitude-longitude coordinates on the plane for simplicity. However, these assumptions may bias our results to some extent.

5.2 Recommendation System

The system can still be further optimized.

Cancellation, time change, and flight number change were not taken into account by the system as they are not included in the data set. If we have data for these variables, we can add features on our system to check for flight cancellation and other changes.

We also can add ranking feature for users to rank our recommendations. Then, we will be able know how they valued our recommended flights by price, departure or arrival delay, historic average departure or arrival time, specific carrier or route. We will use these rankings to improve our recommendation algorithm.

With data for flights' actual arrival or departure time (in recent years), we can calculate the chance of a flight's delay or early arrival. The time factors included in our recommendation metrics will definitely help users to make better decision.

6 Conclusion

Our assumption and processing of the data are rather convincing because the crisscross-points heat-map we generated matched with the extreme climate map well.

For each origin and destination pair, our flight search engine gives the best three flight routes under a customer's request. We recommend flights that have the lowest expectation delay time based on the historical data.

The timeliness of our engine can be improved if we have more recent data from 2008 to 2018. Additionally, if there is price information for each flight, we would also be able to have the system to measure the cost-performance ratio (delay time v.s. financial cost for delay time per minute) for each flight while doing recommendation.

References:

”Federal Aviation Administration Bureau of Transportation Statistics Air Carrier Flight Delays and Cancellations”, Report no. CR-2000-112, 2000.

C. Barnhart, S. Bratu, ”National Trends in Airline Flight Delays and Cancellations and the Impact on Passengers”, Workshop on Airline and National Strategies for Dealing with Airport and Airspace Congestion, 2001.