Hongyao Zhang

Prof. Ernst

CSE 140

3/15/2013

Report

0. Title and author

   Factors for a flight to be late from 2009

1. Overview of research questions and results

   a) What are the top 10 tail numbers that are more likely to be late?

      In this question, I will rank the tail numbers by how many times each tail number is late. The first 10 tail numbers are the top 10 that are more likely to be late.

      Answer:

      ('N27318', 288)

      ('N77302', 276)

      ('N87353', 241)

      ('N37342', 183)

      ('N77278', 177)

      ('N506AU', 165)

      ('N471CA', 142)

      ('N506CA', 129)

('N303SW', 128)

('N323SW', 125)

b) What are the top 10 airline companies that are more likely to be late?

In this question, I will rank the airline companies by how many times each airline companies' flights are late. The first 10 airline companies are the top 10 that are more likely to be late.

Answer:

('19393', 47356)

('19805', 23018)

('20304', 19546)

('19790', 18911)

('20355', 17519)

('20398', 17319)

('19977', 14219)

('19704', 13447)

('20366', 12786)

('19386', 11900)

c) What are the top 10 destinations that are more likely to be late?

In this question, I will rank the destination cities by how many times each flight to that city is late. The first 10 destination cities are the top 10 that are more likely to be late.

Answer:

('Atlanta, GA', 18509)

('Chicago, IL', 16134)

('Denver, CO', 9569)

('Dallas/Fort Worth, TX', 9122)

('New York, NY', 8976)

('Houston, TX', 8254)

('Phoenix, AZ', 7333)

('Los Angeles, CA', 7315)

('Las Vegas, NV', 6635)

('Newark, NJ', 5993)

d) What are the top 10 origins that are more likely to be late?

In this question, I will rank the origin cities by how many times each flight from that city is late. The first 10 origin cities are the top 10 that are more likely to be late

Answer:

('Atlanta, GA', 18370)

('Chicago, IL', 17954)

('Dallas/Fort Worth, TX', 11189)

('Denver, CO', 10159)

('Houston, TX', 9582)

('New York, NY', 8535)

('Phoenix, AZ', 7570)

('Detroit, MI', 6887)

('Los Angeles, CA', 6768)

('Newark, NJ', 6304)

e) What is the most major factor for a flight to be late?

In this question, I will compare the results from question a to e and get a conclusion. I will compute probabilities of being late with factor of each category. The category with largest probability is the one that matters the most for flight lateness.

Answer:

In all the factors, the one matter the most for flight lateness is tail number, because it has the largest probability of being late. The probability is 0.524900075939

f) What is the best flight recommendation from XXX to XXX based on the performance since 2009?

In this question, I will take two names of cities from user's input and traverse the dataset to find the name. Then I will find the flight that matches these two names with shortest average late times.

No specific answer.

2. Motivation and background

My motivation for this research is to make convenience for people who often travel by air. We might notice that flights of some certain routes or some certain airlines are really more likely to be late than others. My research will find out what they are and what the most major reason of being late is so that people could avoid them on the next time they choose a flight. Also my program could take user's input of origin's city name and destination's city name and recommend the best flight on that route for the user.

3. Dataset

My dataset is about flights performance from January 2009. I found it from data.gov website. Since my dataset was gotten from the online data extraction tool, I will upload it to Dropbox so that the staff could get the exact same data as I use for this project. There are 11 columns in my CSV file. They are airline id, carrier, tail number, flight number, origin and destination city, origin and destination state, departure and arrival delay and distance.

4. Methodology

First, I will extract data from CSV file and clean the data so that I can use it. Then in order to find the top 10 things that are more likely to be late, I will separate the data by categories and rank each category's data by the number of lateness flights (0 in departure and arrival delay means the flight is on time) under that category. The first 10 things in the rank are the things I am trying to get. After that I will add up the ratio of the late times of each element under each category and the occurrence of that element. And I will divide the sum by the total number of

elements under that category to get the overall probability. The one with the largest probability is the most major factor for a flight to be late. To achieve the recommendation function, I will search through the dataset for the city names (origin and destination) that the user inputs. I will compute the average delay time of each flight. To get the average delay time, I will add up the late time in minutes of each flights that matches the city names and divide the sum by the number of flights that goes from the origin to the destination. The one with shortest average delay time is recommended flight.

5. Results

   a) I am trying to find some airplanes that is much more likely to be late than others. What I find interesting is, airplanes with tail number start with "N" and follow with digits only is significantly more likely to be late than other airplanes.

   b) I am trying to find some airline companies that always have significantly more late flights so people could avoid taking their flights. Then I find an airline with id "19393" that has even twice as many late flights as the second airline in the rank.

   c) The reason to find top 10 destinations is to see if there are any airports always cause planes to be late coming in. People could know their flights may very possibly to be late if they are going there before hand. The result is not surprising because large cities and large airports are usually quite busy and cause late coming in. The cities I got in my rank are all huge cities.

That’s why the answer doesn’t surprise me.

d) Pretty much the same reason with question e, I am trying to find some cities that always cause plane to be late going out. The result surprises me a little because the rank list is almost exactly the same with destination rank list. This tells me that huge cities not only could cause planes to come in late but could also cause planes to go out late.

e) The tail number seems to be the most major factor for a plane of being late because it has the largest probability. We can see that mechanical problem of a plane is a considerable factor of being late.

f) I think this is a quite useful function. Most people hate flight delay and this function may help them efficiently plan their trip and avoid delay.

6. Reproduce

Download the data file from Dropbox and save it in the same folder with my source code. Open cmd and enter “python final_proj.py the data file name”. Type tail number, yes, destination, yes, origin, yes, airline, no to get my result from question a to e. Then the rest is up to you. The search function is more like an application and very flexible. No specific answer is provided for that.

7. Collaboration

There is no one helped me on this project. I learned how to make an assignment by myself and to think in the way that the one who creates the assignment does. I also learned how to use the technique we learned in class for something in our real life.