CSE 140 Final project

Airline delay analyze

I. Question and result:

In this project, I will analyze the airline delay information. May times when we go to the airport and try to catch our plane, there may be a coincident that your plane may get delay, so what is the possibility that your plane may get delay? And what is the factor that the plane may delay? This project will analyze this information by different key factor and get a result.

Questions:

1. Is the delay are appear more common in some specific days in the week?
2. Is some airline just may cause delay much more than others?
3. Is the travel distance a factor that causes the delay more often?
4. Is that some city or airport have more delay that the other?

Answers:

1. Not really, as my results turned out, the days in the week have a delay that have the similar delay percentage that is below the 5% different, which in case it common to happen
2. Yes, the airline performance has a huge impact and difference. Some airlines have a delay percentage up about 58% and some are low as 20%.
3. Yes. It is kind of opposite of what we think, the shorter the distance travel, the easier it get delay. My results show that if the travel distance is less than 500 miles the delay rate is much higher than others. Specific detail wills shows later.
4. Yes. Different airports perform huge different. The worst are Los Angeles LAX and San Francisco SFO.

II. Motivation and Background:

As an international student, flying schedule is so important to us, the delay may cause us to do a schedule change or even worse, miss a transaction flight. So it would be good if we could find a way to determine what are the key factors that cause the delay so we can try to avoid it. For say, if is just some airline that is often delay, we could try to find a alternate airline, or if just the airport is busy, alternate airport my also be an good option. In my personal experience, there is once I had to sleep on the floor in the LAX airport because the flight has late for 16 hours and I also missed my transaction flight. So after this analyze, guess those things will never happen again.

III. Dataset

The data that I am using is from http://www.transtats.bts.gov/ the exact URL is:
http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time

And for this analyze, I had choose the 2012 first quarter information. And I had also included he entire 2012 information for a longer and bigger performance. The result will be more accurate, but the assumption will be the same.

IV. Methodology

The program first by read the entire pre-downloaded csv file in the directory or user passed in directory. And store them in a list. By saving runtime, I had choose to use the build in csv reader and save it to a set of tuple form, this saves a lot of time since you don't have to run through all the data and then sort them.

(1) Check by airline:

The function "all_airline" take all the stored information and return all the airlines in a set. And then the "check_all_airline" function takes all the stored information and all the airlines and returns a list of airline names as a string and pair with their delay percentage. The delay percentage is calculated by the function called "Air_line_percentage" it take an airline's name as a string and take all stored information and return the delay percentage of the air company as a float. The airline_mostdelay function get a set of all the airlines and all the info and a top number and return the delay rank of those top delay airline

(2) Check by city.

The depart_desitnation function gets all stored information and returns a set of departure and destination city pair.

(3) Check by distance

The air_line_distance take the all stored information and return a pair of the name of the airline, delay time and the travel distance. The check by distance function gets all the distance information and sorts them by scale.

(4) print_airline_info function and the plot_city function will print the calculated Information

V. Result

Printed airline information:

Airline: OO
Top delay of this Airline: [(('SAN', 'LAX'), 811), (('LAX', 'SAN'), 787), (('SLC', 'DEN'), 677), (('LAS', 'LAX'), 666), (('SBA', 'SFO'), 626)]
Delay Ratio: 36.9313297889

Airline: AA
Top delay of this Airline: [(('DFW', 'LAX'), 958), (('DFW', 'ORD'), 879), (('LGA', 'DFW'), 863), (('ORD', 'DFW'), 842), (('LAX', 'DFW'), 841)]
Delay Ratio: 38.1862943811

Airline: DL
Top delay of this Airline: [(('LGA', 'ATL'), 813), (('ATL', 'LGA'), 802), (('MCO', 'ATL'), 750), (('ATL', 'DCA'), 730), (('DCA', 'ATL'), 694)]
Delay Ratio: 32.7498498202

Airline: HA
Top delay of this Airline: [(('OGG', 'HNL'), 694), (('HNL', 'OGG'), 600), (('LIH', 'HNL'), 565), (('KOA', 'HNL'), 542), (('HNL', 'KOA'), 532)]
Delay Ratio: 46.11593827

Airline: WN
Top delay of this Airline: [(('HOU', 'DAL'), 837), (('DAL', 'HOU'), 814), (('LAX', 'SFO'), 671), (('LAS', 'PHX'), 669), (('SFO', 'LAX'), 616)]
Delay Ratio: 32.5532784601

Airline: AS
Top delay of this Airline: [(('SEA', 'ANC'), 731), (('ANC', 'SEA'), 665), (('SEA', 'LAX'), 598), (('LAX', 'SEA'), 588), (('SEA', 'SFO'), 494)]
Delay Ratio: 33.7868558832

Airline: US
Top delay of this Airline: [(('BOS', 'PHL'), 728), (('DCA', 'LGA'), 720), (('LGA', 'DCA'), 682), (('PHL', 'BOS'), 666), (('BOS', 'LGA'), 664)]
Delay Ratio: 34.118175803

Airline: B6
Top delay of this Airline: [(('JFK', 'FLL'), 655), (('FLL', 'JFK'), 630), (('JFK', 'MCO'), 620), (('MCO', 'JFK'), 553), (('JFK', 'PBI'), 496)]
Delay Ratio: 37.5165569876

Airline: MQ
Top delay of this Airline: [(('CMH', 'ORD'), 557), (('EWR', 'ORD'), 552), (('ORD', 'EWR'), 551), (('SAN', 'LAX'), 522), (('RDU', 'LGA'), 519)]

Delay Ratio: 34.5578157881

Airline: FL
Top delay of this Airline: [(('LGA', 'ATL'), 616), (('ATL', 'LGA'), 585), (('MCO', 'ATL'), 521), (('ATL', 'MCO'), 491), (('FLL', 'ATL'), 473)]
Delay Ratio: 28.4522667556

Airline: F9
Top delay of this Airline: [(('DEN', 'LAS'), 405), (('LAS', 'DEN'), 395), (('DEN', 'SFO'), 342), (('SFO', 'DEN'), 333), (('DEN', 'DFW'), 327)]
Delay Ratio: 51.2119799557

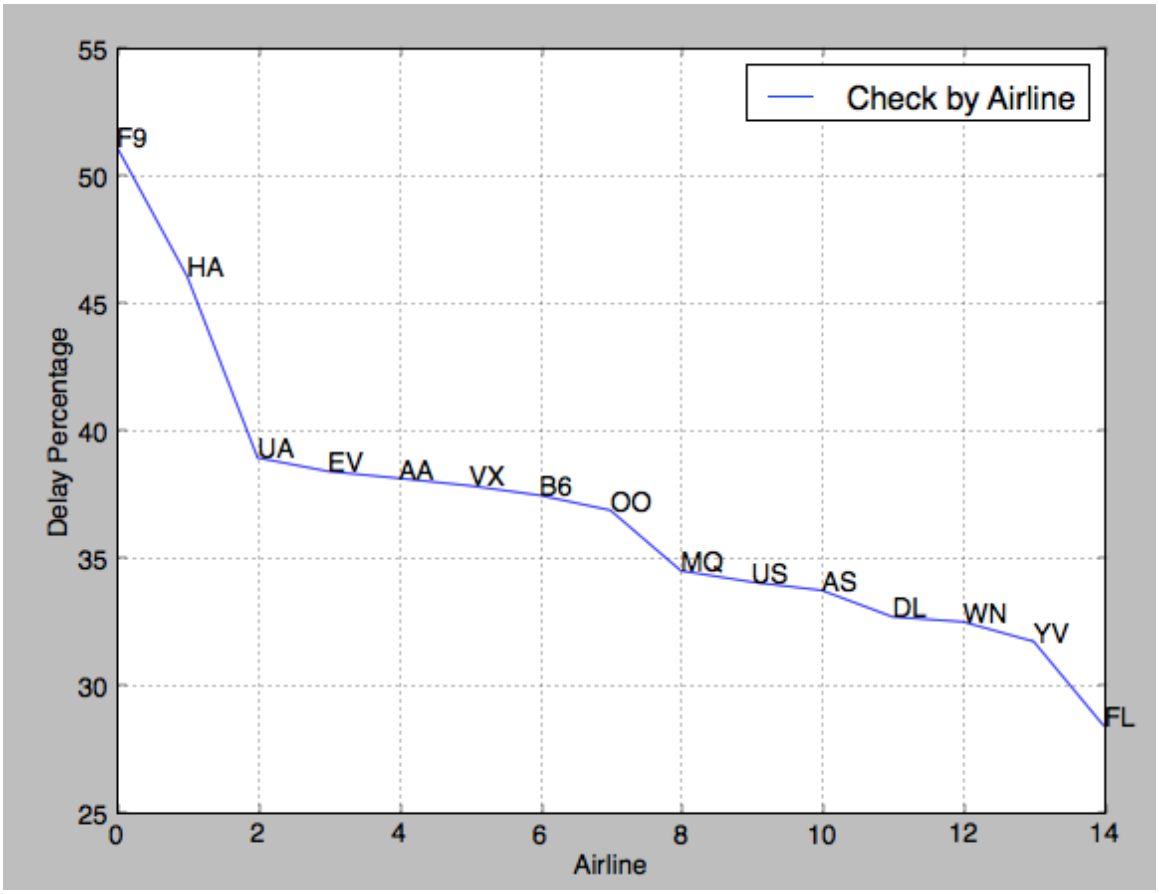
Airline: VX
Top delay of this Airline: [(('LAX', 'SFO'), 510), (('LAS', 'SFO'), 499), (('SFO', 'LAX'), 478), (('JFK', 'LAX'), 446), (('SFO', 'LAS'), 436)]
Delay Ratio: 37.895841282

Airline: EV
Top delay of this Airline: [(('CHA', 'ATL'), 591), (('ORD', 'CLE'), 589), (('ATL', 'CHA'), 575), (('IAH', 'CRP'), 575), (('CLE', 'ORD'), 562)]
Delay Ratio: 38.4573558603

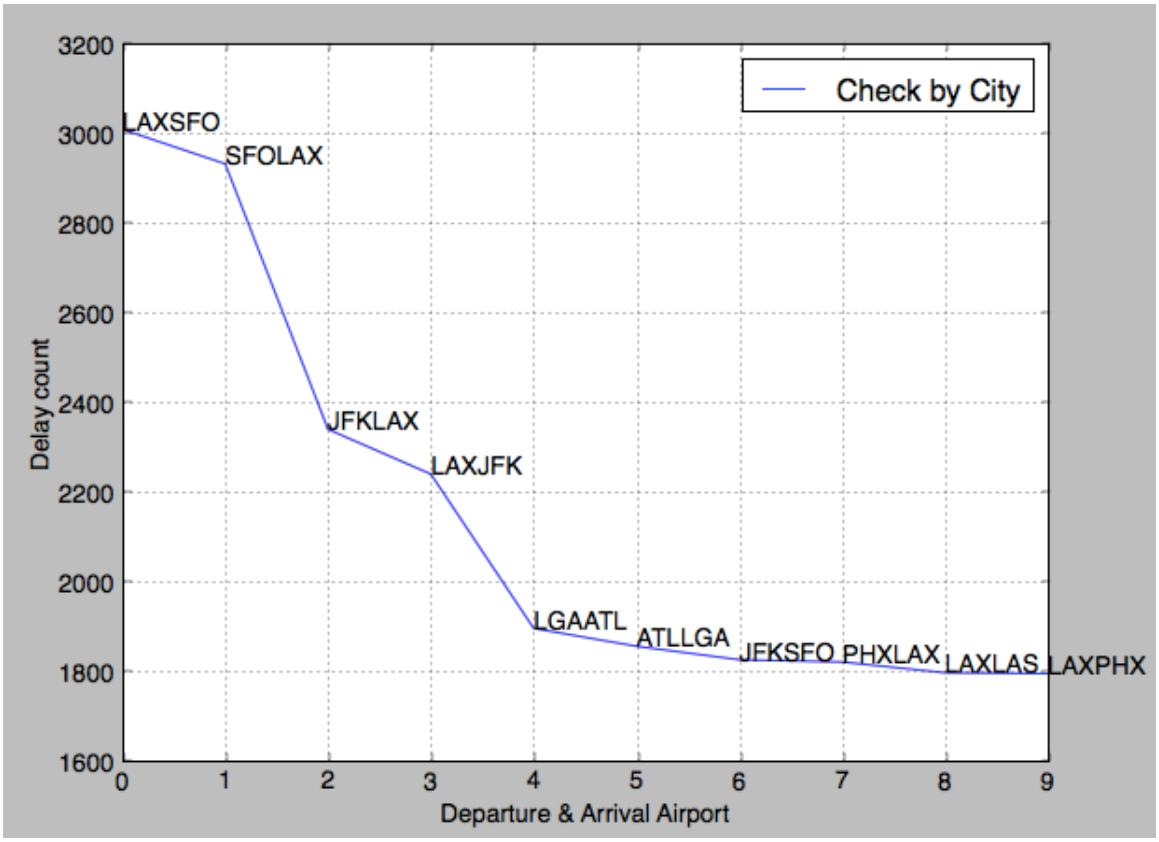
Airline: UA
Top delay of this Airline: [(('ORD', 'SFO'), 923), (('SFO', 'ORD'), 902), (('SFO', 'LAX'), 861), (('LAX', 'SFO'), 846), (('LGA', 'ORD'), 769)]
Delay Ratio: 38.9905743604

Airline: YV
Top delay of this Airline: [(('CLT', 'IAD'), 520), (('HNL', 'OGG'), 502), (('OGG', 'HNL'), 490), (('TUS', 'PHX'), 473), (('IAD', 'CLT'), 473)]
Delay Ratio: 31.782297451

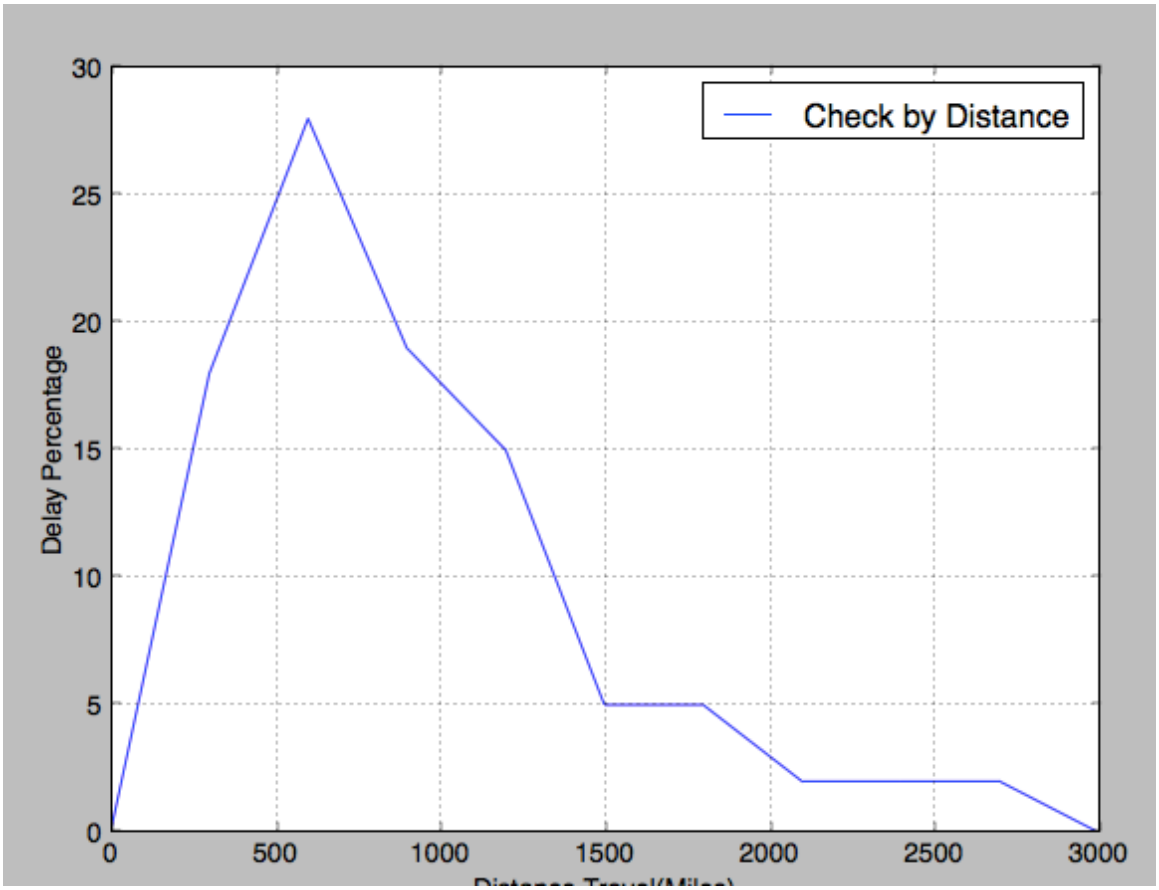
The delay percentage of each individual airline:



The delay count by the departure and arrival airport.



The delay percentage by the travel distance:



And days of weeks and it's delay percentage:

Days of Weeks	Delay percentage
1	14.75%
2	13.906%
3	14.11%
4	14.82%
5	14.96%
6	12.40%
7	13.64%

VI. Reproducing the result.
User could either type a specific directory where all the allowed files are store all just run the file with python Module. The user could put all the file in the "On Time On Time Performance 2012" folder and run the program

VII.
I did this assignment myself since it is not that much work required, and from this program we have learn that Python is powerful tool to take almost any kind of file and run the model you build. It is nice to get to do this assignment and get some practice on it.