Homework #2
MEM 410, Managerial Analytics, Winter 2018
Due: Start of Class 2/1/18
Electronic Submission Only

Download the "Chicago – NYC – Boston Flights" dataset from Canvas. Leverage MS Office or any tool that you are comfortable with to address the following questions:

1. [3 points] This data file is supposed to contain flight information among three cities (Chicago, New York, and Boston). Assume that the objective is to understand flight delays for traveling among the three cities. Upon closer review of this data, what limitations do you observe?

Answer:

I. To understand flight delays for traveling among the three cities, there should be to and fro information for all the three cities. However, it is observed that there is limited data for origin and destination. For instance, there is no flight information for: -
   a. Chicago to Boston
   b. New York to Chicago or
   c. flights that originate from Boston.

II. Delay could be owing to multiple reasons namely Carrier, Weather, NASdelay, security, late Aircraft. However, data in these columns is missing for records with arrival delay of less than 15minutes. In view of this missing information, it will be challenging to draw accurate conclusions for flight delays.

2. [3 points] What airlines are included in this file and which airlines may have questionable data?

Answer :

A. Information for following airlines have been included Airlines are included in this file.

Airlines
Altantic Southeast Airline
American Airlines
American Eagle
Delta Airlines
JetBlue
SkyWest
Southwest
Spirit
United
US Airways

B. i) Atlantic Southeast Airline has questionable data since there is only one flight that goes from ORD to EWR.

	EWR	JFK	LGA	MDW		ORD			Grand Total
Row Labels	BOS	BOS	BOS	EWR	LGA	EWR	JFK	LGA	
Altantic Southeast Airline	180					1			181
American Airlines		1456	1498			327	680	5290	9251
American Eagle						286			286
Delta Airlines		1504	3707				33		5244
JetBlue	1765	2611					1040		5416
SkyWest								80	80
Southwest				1978	2391				4369
Spirit								608	608
United	3019					3578		4471	11068
US Airways			2719						2719
Grand Total	**4964**	**5571**	**7924**	**1978**	**2391**	**4192**	**1753**	**10449**	**39222**

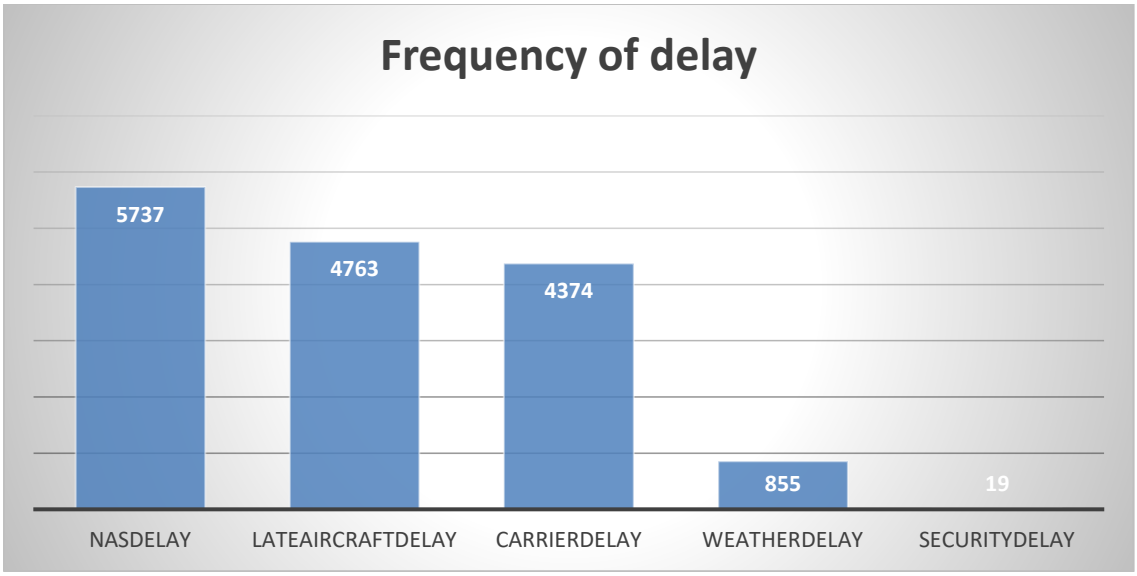
ii) There is significant time gap in flight information over the one year. For eg. For American Eagle, there is missing information from Nov'14 to May'15 and US Airways is missing flight information from Jul'15-Oct'15.

	2014	2014	2015	2015	2015	2015	2015	2015	2015	2015	2015	2015	Grand Total
Row Labels	11	12	1	2	3	4	5	6	7	8	9	10	
Altantic Southeast Airline	20	12	14	17	18	15	17	13	9	15	12	19	181
American Airlines	637	639	606	531	597	688	669	654	1106	1074	1002	1048	9251
American Eagle								50	30	40	56	110	286
Delta Airlines	504	379	480	478	524	511	510	526	332	344	316	340	5244
JetBlue	401	396	402	356	429	425	510	506	516	511	478	486	5416
SkyWest		2	5	7	6	11	9	8		13	15	4	80
Southwest	352	374	358	320	362	363	372	365	379	376	369	379	4369
Spirit			62	56	62	60	62	60	62	62	60	62	608
United	910	853	822	812	983	923	942	988	1000	962	883	990	11068
US Airways	299	254	323	346	385	378	354	380					2719
Grand Total	**3123**	**2909**	**3072**	**2923**	**3366**	**3374**	**3445**	**3550**	**3434**	**3397**	**3191**	**3438**	**39222**

iii) All airlines have some flights cancelled but there is a Departure delay value associated with it which seems odd since they never took off.

3. [3 points] Assume that these five fields ("CarrierDelay", "WeatherDelay", "NASDelay", "SecurityDelay" and "LateAircraftDelay") represent the reasons for flight delays. What's the most common reason for flight delays based on the data? Do you observe any data risks in drawing such a conclusion?

All 5 types of delays have data for equal number of records. Hence, if we select only non-zero values to calculate the frequency of each kind of delay, NASDelay has maximum frequency and is the most common reason for flight delays.



The risk involved in drawing conclusions from this is that there is missing information in these 5 delays for majority of records. It is uncertain how those missing records might play into frequency and hence this approach does pose a risk.

4. [3 points] Base on this data, if I want to choose an airline with the least amount of expected delays when traveling from Chicago to New York City, which airline should I choose, assuming that I am open to using any of the airports in either cities and I might fly any day of the week?

Assumption : EWR has also been included for comparison purposes assuming any of the airports in either cities could be chosen.
Clearly Atlantic Southeast has the minimum Average of Arrival Delay Minutes(0). However, there is only one flight to provide such insight. We would rather consider Skywest that has more data to support that the Average delay is 10.6 minutes. Hence, **SkyWest airline** should be chosen.

Airline	Origin	Dest	Average of Arrival Delay Minutes	Count of Flights
Altantic Southeast Airline	**ORD**	EWR	0.00	1
SkyWest	**ORD**	LGA	10.61	80
American Eagle	**ORD**	EWR	11.63	286
American Airlines	**ORD**	LGA	13.58	5290
American Airlines	**ORD**	JFK	20.75	680
American Airlines	**ORD**	EWR	35.24	327
Delta Airlines	**ORD**	JFK	16.58	33
Southwest	**MDW**	EWR	17.96	1978
Southwest	**MDW**	LGA	18.10	2391
Spirit	**ORD**	LGA	19.06	608
JetBlue	**ORD**	JFK	20.66	1040
United	**ORD**	EWR	20.93	3578
United	**ORD**	LGA	21.49	4471

5. [3 points] Does the above answer change for me if I want to make a weekend trip, i.e., going to New York on Saturday morning (i.e., arriving before 12pm)? Please show the comparison across the different airlines to justify your answer. Are there any risks to your "optimal" selection? What other analyses can you demonstrate to further justify your recommendation?
Yes , the answer does change. In this case, Spirit Airline should be chosen. As shown below, Spirit has an average delay of 2.33 minutes and Southwest comes close to 2.81. The difference is not too much but the number of flights associated can make a difference. We can conclude that greater the number of flights, stronger is the hypothesis of arrival delay minutes.
Assumption : Saturday is counted as 6th day of the week

Airline	Origin	Dest	Average of ArrDelayMinutes	Count of FlightNum
Altantic Southeast Airline	**ORD**	EWR	0.00	1
Spirit	**ORD**	LGA	2.33	43
SkyWest	**ORD**	LGA	5.79	19
Southwest	**MDW**	EWR	17.69	51
Southwest	**MDW**	LGA	2.81	110
American Airlines	**ORD**	EWR	18.36	14
American Airlines	**ORD**	JFK	81.50	2
American Airlines	**ORD**	LGA	5.22	98
United	**ORD**	EWR	10.57	89
United	**ORD**	LGA	7.76	37
JetBlue	**ORD**	JFK	27.81	48
Grand Total			9.62	512

Another alternate method to further justify Spirit being the optimal solution is that if we compute the maximum arrival delay for each airline, Spirit still fairs to be causing the least delay(after Atlantic Southeast's one objectionable record). This further strengthens our hypothesis.

6. [3 points] Assume now that I am planning a trip from Chicago to Boston via New York for 12/24/17. I plan to leave Chicago no earlier than 7:00am and I need to get to Boston no later than 3:00pm. Assume that it will take me 30 minutes to get off one plane and catch the other. I still want to minimize the total amount of time I spend in between. Which airline should I fly?
Assumption: Flights for 12/24/2017(Sunday-Day 7) is assumed to have the same flight schedules as 12/24/2014.

It takes 30mins to de-board from 1st flight and board the next one, Southwest Airlines should be chosen flying from MDW to EWR(arrival time 1122) which gives it ample time to board the next JetBlue flight at 1159 minimizing the wait time to 7 minutes.

Airline	FlightNumber	Departure Time	Arrival Time	Origin From	Origin To
Southwest	1623	826	1122	MDW	EWR
JetBlue	2380	1159	1300	EWR	BOS

7. [7 points] Suppose that you are working for a travel site company such as Expedia. You are managing a project that will rank available options based on the least delays per customer's desired travel date & destination. What are the different considerations for such a capability? What limitation do you see if you can get a "full version" of the dataset above? What other data would you propose to have? What validations would you propose for the additional datasets?
   A) *Considerations* - In order to rank airlines based on least delays per customer's desired travel date & destination, we need to consider
      a. availability of delay data to support such hypothesis
      b. Flight information for at least for a couple of years to draw concrete inferences.
      c. Consider season of travelling or estimate forecasted weather that could potentially affect delay.
      d. Socio-economic conditions of the destination could also play into the incoming and outgoing flights from a location.
      e. Accounting for unexpected delays at the airport (potentially for a connecting flight).
   B) *Limitation* - If we get a full version of dataset available to us, we would have information for many more destinations and they might have other delaying factors, normalizing that data could be a challenge. Dealing with outliers that skew the data could be a possible problem that needs to be accounted for.
   C) *Additional Data* - To manage ranking of an airline, we would need
      a. past data for at least 5 years
      b. customer experiences with regards to how delayed flights were handled by an airline.
      c. Role of delay in determining ticket prices. A consumer might buy a cheaper ticket for an airline that has higher average delays or willing to pay more to travel with an airline having better track records.
   D) Validations –
   i) The additional datasets should be in accordance with the existing dataset.
   ii) There should be consistency among the fields captured on the complete dataset.
   iii) There can be validation checks build around the data to ensure its correctness.
   iv) Account for the fact that different areas can have different reasons for delay and impact differently.