BAN 602   Quantitative Fundamentals   Fall 2019
Homework Assignment 1
RegionEx case

You can find the case here. https://doi.org/10.1287/ited.1110.0066cs) (Click on "View pdf" in the link) The key objective behind your report is to analyze whether the flight delay rankings tell the whole story. In order to answer this question, you should start by answering the following questions.

Note: When you load the data, some columns may not be loaded in the correct format. Use the class function to check if the variables are in the correct format. If you want to convert a variable into numeric, for example, if you would like to change column "Tonumber" in numeric format, use the function as.numeric(as.character(Tonumber)). You may use the function head to ensure correct conversions. There are also some blank rows in the dataset. Feel free to remove them if it is troublesome (see hint 1 below).

1. Compute the mean, median, 90th percentile, and standard deviation of the arrival delay minutes of RegionEx's flights. Do the same for MDA's flights. How do the two airlines compare? What's your conclusion from this quick analysis? (Hint 1: Some of the MDA flights were cancelled. You may want to remove these rows before you proceed with numeric calculations. To identify the row numbers that were cancelled, you can use the function which() in R. For example, to identify the row numbers that had a value of "remove" in column "thisVar", use the function which(filename$thisVar=="remove"). Then you may remove these rows from your dataset by using the – sign. For example to remove row 3 you can say newdf = df[-3,]. To remove a set of rows, that is, say you want to remove rows 5,6,8. Then denote "a" as a <- c(5,6,8). Then newdf = df[-a,]. Hint 2. Similar to hint 1, you may want to just split the database into two. One that has RegionEx and the other that has MDA. This may make the analysis simpler.)

2. Inspect the distribution of RegionEx's arrival delays by constructing a histogram (or a frequency distribution) of the arrival delay minutes of RegionEx's flights. Do the same for MDA's flights. How do these two distributions compare? See if your conclusions from comparison of descriptive statistics (in Question 1) is consistent with these histograms. What, if any, additional information do the histograms provide?

3. Using the FAA definition of a "late" flight, what percentage of RegionEx's September flights were "late"? What percentage of MDA's September flights were "late"? What percentage were "on-time" for each airline, according to the FAA definition? (Note: The data already incorporates the FAA definition and calls it "Delay indicator").

4. Compare the performance of the two airlines on each flight leg by calculating the descriptive statistics (mean, median, 90th percentile, standard deviation) of delay minutes for each of the four routes. Calculate the % delay flights for each of the route. Do any of the comparisons change? If so, why?

5. Consider only the RegionEx flights. Prepare a scatter plot of arrival delay minutes versus number of passengers. Your scatter plot should consist of 240 data points, one for each flight in the data set where the vertical coordinate is arrival delay minutes of that flight and the horizontal

coordinate is the number of passengers. What is the correlation coefficient between arrival delay minutes and number of passengers for RegionEx's flights? Interpret your results.

6. Compare the scheduled flight durations for the two airlines on each of their four routes. Compare the actual flight durations. What do you notice? If the two airlines had the same scheduled duration, what impact would this have on their delay records? (Hint: Again, the numbers may not be loaded in the right format. Use the function as.POSIXct(TimeColumn, format="%H:%M") to ensure you can perform mathematical operations).

7. What other factors should Marion Volero take into consideration regarding the data analysis? Feel free to use crosstabs (or any other method) which can make answering questions really easy. Please also submit the R file (just one R file) in which you performed the analysis along with your answers in pdf.

Good luck!