Assignment 5: Data Science with Airlines

This assignment continues our dive into the world of real data sets and the interesting questions that can be asked and answered with data science.

In order to do this assignment, here are the concepts that you will need to know:

- *File handling - the csv file has to be read*
- *HashMaps and ArrayLists - once the file is read, we have to store the data in a data structure that will make data retrieval easy*

*The activities in **Weeks 8 and 9 Recitation** will help you practice for this assignment, so it is highly recommended that you attend the live sessions or review the recordings and be sure to attempt the recitation activities.*

Introduction

If you have ever taken a flight, you are likely to have been concerned about whether or not your flight is going to be delayed. The Bureau of Transportation Statistics provides publicly available data sets on every aircraft that has taken off in the United States. The csv that we provided has data from 2016.

We would like you to answer questions using this data. Are you frustrated with the airport that is closest to you? Do you think others couldn’t possibly have it worse than you? What is the likelihood that the flight you have to take to visit your family during Thanksgiving will get there on time? Here’s the assignment that will help answer these questions using your newly acquired programming skills.

Data Science Questions

Using the flights.csv file that we have supplied, we want you to write Java code to read the file and then answer the following questions. *We have also supplied another small csv file that has the details on the cancellation codes.*

Please do not create one giant main method to answer these. Remember to make your code DRY.

There is a specific manner in which we want you to answer these questions. We need you to use the FormattedOutput.java class and use that to write your answers to a file. The file needs to be called **answers.txt** and that needs to be submitted along with the actual Java code.

Cancelled flights - These flights do not count in the source/sink calculations. They should also not count in Q7 and Q8. It is probably best to put them in a separate data structure.

You do need them when you calculate the percent of cancelled flights.

Diverted flights - do not count them while solving Q7 and Q8. For the other questions also we would like you to ignore them but it turns out that it does not seem to matter for the answers.

Erroneous data - Some flights are flagged as cancelled, yet have an arrival time and/or departure time. Likewise, some flights have a departure time, but no arrival time. In other cases, flights may have a cancellation code but are not flagged as being cancelled. For flights with incomplete information, you may ignore them entirely. For flights that are flagged as being cancelled, you should treat them as any other cancellations regardless of any other data associated with them.

Blank cells in the .csv file - There will always be missing data in files--data is messy! Just discard the blank cells if the information isn't affecting your code. Do NOT delete the entire row if there is one blank cell (for example if there is a blank cell in 'cancellation reason' then don't delete the whole row-- maybe the flight wasn't cancelled).

1. Which carrier has the highest percentage of cancelled flights? Output the 2-letter Carrier ID and the chance of a cancelled flight, as a percentage (Example: AA,1.22%). This percentage is defined as the number of canceled flights over the total number of flights.
   When computing this percentage please do not do any rounding. Report it to the default level of accuracy (do not use Math.round etc).

2. What's the most common cause of cancellations? Output the one-letter code.

3. Which plane (tail number) flew the furthest (most miles)? Output the complete tailnumber (Example: N775AJ).

4. Which airport is the busiest by total number of flights in and out? Use the number OriginAirportID (Example: 12478).

5. You need planes to put people on! Which airport is the biggest "source" of airplanes? Use the difference between arrivals and departures to compute this value. Output the OriginAirportID (Example: 12478) Biggest source means find the greatest difference between the number of departing flights and number of arriving flights. (Do not consider cancelled flights in this calculation.)

6. Which airport is the biggest "sink" of airplanes? Again, use the difference between arrivals and departures, outputting the OriginAirportID (Example: 12478). Biggest sink means find the greatest difference between the number of arriving flights and number of departing flights. (Do not consider cancelled flights in this calculation.)

7. How many American Airlines (Unique Carrier ID ‘AA’) flights were delayed by 60 minutes or more? If a flight was delayed departing and arriving, only count that as 1. Output an integer.

8. What was the largest delay that was made up (arrived early/on time)? Output the Day of Month (the number), departure delay (as a number), and the tail-number. Example: (10,30,N947JB).

9. Come up with a question of your own and answer it! (You can put the question and answer right in here)

Checking your Solution

We have provided a smaller data set flights_small.csv that is used for our validation tests. Please use this data set and a corresponding unit test that we have written to validate that your code is working as we expect it to.

To validate your solution:

1. Use your code to read flights_small.csv.
2. Write your answers to answers.txt. Please use our FormattedOutput.java file for this.
3. Run our ValidationTest.java file as a unit test and ensure it passes.

What to Submit

Submit all of your code (all your .java files).
Do not submit the flights.csv and the ValidationTest.java file--we already have these two files on our end. If you submit ValidationTest.java the code will crash on Codio.

Make sure that you have run your code on the full input file called “flights.csv”. This should generate an output file answers.txt that has the answers to all nine questions. Remember to submit answers.txt in addition to the code.