DSCI550: Data Science at Scale

Homework 1, Fall 2024

Due: 11:59 PM, 9/20/2024

This assignment must be done independently.

The purpose of this assignment is to provide you with some experience exploring and analyzing data using Python.

1. See the "flights.csv" file about flight delay data. The dataset should be pretty self-explanatory. You should explore and analyze this data using Python programs. Your goal here is to perform a simple analysis of the dataset and answer the following questions. Submit your Python code with the filename "YourLastname-YourFirstname.Python-file-extension" using the class web page.

   Note: When you find missing values, remove the flight record (i.e., entire row which has missing value(s)) in the analysis. When you see a negative delay number, it can be interpreted as an early departure and arrival, so it will not be considered as a delay (i.e., treat them as zeros).

   1) (20 pts) How many flights were recorded for each month in the dataset? And what is the average departure delay time for each month?
   2) (20 pts) Which were the top 5 most common destinations from the 'JFK' airport, based on the frequency?
   3) (20 pts) Calculate the total distance traveled by each carrier. Which carrier traveled the most distance overall?
   4) (20 pts) How much departure delay do I need to expect when I use AA in December? (based on average)
   5) (20 pts) Determine the top 5 busiest days (based on the number of flights). Report the year, month, day, and the number of flights.