CS 422: Data Mining

Department of Computer Science
Illinois Institute of Technology
Vijay K. Gurbani, Ph.D.

Fall 2024: Homework 2 (Exploratory Data Analysis) (10 points)

Due date: Friday, September 13 2024, 11:59:59 PM Chicago time.

1 Exercises (3 points)

1.1 [1 pt.] Read the Exploratory Data Analysis (EDA) section in Wikipedia (https://en.wikipedia.org/wiki/Exploratory_data_analysis). Based on your reading, answer the following questions:

(a) What is EDA?
(b) Why is it important in data science and machine learning? (List 2-3 advantages of EDA.)

1.2 [1 pt.] Tan, Chapter 1, questions 1, 3.

1.3 [1 pt.] Tan, Besides the lecture, make sure you read Chapter 2, sections 2.1 – 2.3. After doing so, answer the following questions at the end of the chapter: 2, 3, 7, 12.

2 Practicum problems (7 points)

EDA on the DelayedFlights dataset.

The DelayedFlights dataset consists of 1.9 million entries pertaining delayed flights. A smaller version of this dataset (387,351 rows) has been made available to you on Blackboard for EDA. The dataset contains the following dimensions (or variables):

1. **Year** 2008
2. **Month** 1-12
3. **DayofMonth** 1-31
4. **DayOfWeek** 1 (Monday) - 7 (Sunday)
5. **DepTime** actual departure time (local, hhmm)
6. **CRSDepTime** scheduled departure time (local, hhmm)
7. **ArrTime** actual arrival time (local, hhmm)
8. **CRSArrTime** scheduled arrival time (local, hhmm)
9. **UniqueCarrier** unique carrier code
10. **FlightNum** flight number
11. **TailNum** plane tail number: aircraft registration, unique aircraft identifier
12. **ActualElapsedTime** in minutes
13. **CRSElapsedTime** in minutes
14. **AirTime** in minutes
15. **ArrDelay** arrival delay, in minutes: A flight is counted as "on time" if it operated less than 15 minutes later the scheduled time shown in the carriers' Computerized Reservations Systems (CRS).
16. **DepDelay** departure delay, in minutes
17. **Origin** origin IATA airport code
18. **Dest** destination IATA airport code

19. **Distance** in miles
20. **TaxiIn** taxi in time, in minutes
21. **TaxiOut** taxi out time in minutes
22. **Cancelled** was the flight cancelled?
23. **CancellationCode** reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
24. **Diverted** 1 = yes, 0 = no
25. **CarrierDelay** in minutes: Carrier delay is within the control of the air carrier. Examples of occurrences that may determine carrier delay are: aircraft cleaning, aircraft damage, awaiting the arrival of connecting passengers or crew, baggage, bird strike, cargo loading, catering, computer, outage-carrier equipment, crew legality (pilot or attendant rest), damage by hazardous goods, engineering inspection, fueling, handling disabled passengers, late crew, lavatory servicing, maintenance, oversales, potable water servicing, removal of unruly passenger, slow boarding or seating, stowing carry-on baggage, weight and balance delays.
26. **WeatherDelay** in minutes: Weather delay is caused by extreme or hazardous weather conditions that are forecasted or manifest themselves on point of departure, enroute, or on point of arrival.
27. **NASDelay** in minutes: Delay that is within the control of the National Airspace System (NAS) may include: non-extreme weather conditions, airport operations, heavy traffic volume, air traffic control, etc.
28. **SecurityDelay** in minutes: Security delay is caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas.
29. **LateAircraftDelay** in minutes: Arrival delay at an airport due to the late arrival of the same aircraft at a previous airport. The ripple effect of an earlier delay at downstream airports is referred to as delay propagation.

Note that the dataset has many missing values as indicated by the presence of "NA" in the data frame. You will leave these missing values in the dataset instead of removing them or imputing them. You will work around these missing values. When using correlation or covariance functions, study the "use" parameter and decide what value to provide this parameter such that the "NA"'s are omitted from consideration. Also

2(a) **[1 pt.]** How many columns have "NA"'s in them? Write R code to display a table of all columns and the frequency of "NA"'s in that column. (Hint: You can use the colSums() function in conjunction with the is.na() function for this.). Return a scalar that represents number of columns that have NAs in them.

2(b) **[2 pts.]** Focus on the cause of delays: CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, and LateAircraftDelay.

(i) Which of these delays do you *think* leads to **most** travel problems and which one leads to the **least** travel? Your function should return a string with two concatenated delays separated by a comma (e.g., "most=carrierdelay,least=securitydelay").

Don't do any modeling for 2(b)(i), just think about the problem and determine which of the delays leads to most travel problems and which to least.

(ii) Create a new data frame that has three fields called "cause" (string), "mean" (numeric) and "median" (numeric). Find the mean and median of each of the five delays and put the cause (one of "carrier", "weather", "nas", "security", and "late_aircraft") along with the mean and median as rows in the data frame. **As you insert the mean and median into the data frame, round them to 2 decimal places.**

Sort the data frame such that it is sorted by the mean column, highest mean at the top and lowest at the bottom. Return the data frame.

(iii) Does the data in 2(b)(ii) match your expectations in 2(b)(i)? That is did you guess accurately which of the delays will lead to **most** travel problems and which to the **least** problems? Your function should return one of the following strings: "yes" (guess accurately for which delay leads to most and which delay leads to least problems), "no" (did not guess either correctly), or "half" (guessed one of them correctly, does not matter which).

2(c) **[2 pts.]** Find all the observations where the departure delay was greater than 30 minutes and save that into a new data frame. Use that dataframe to answer the following questions:
(i) How many observations are there where the departure delay is > 30 minutes? Your function should return a scalar indicating the number of observations.

(ii) Of all of the observations where the delay is > 30 minutes, find which airline had the **least** delays and which had the **most** delays. (Hint: Look at table() API.) Return these two airlines and their total delays as a string in the following format: "xx=n,yy=m". Here, "xx" and "yy" are the 2 letter airline code (UniqueCarrier[1]) corresponding to the airlines with the least and most delays, respectively; and "n" and "m" are the least and most delays by these airlines, respectively.

(iii) Use barplot() to plot a histogram of the delay frequencies ordered from lowest to highest. The x-label should be "carrier code", y-label should be "delay (minutes)", and the title of the plot should be "delayed flights".

2(d) **[2 pts.]** Focus on the delays: Carrier, Weather, NAS, Security, and LateAircraft and the variable DepDelay (the departure delay). You are trying to determine which of these delays are correlated with late departure. Save these six variables to a new data frame and using the dataframe, answer the following questions:
(i) Produce a correlation matrix from the new dataframe using the cor() API. Note that that there are NA's in the Carrier, Weather, NAS, Security, and LateAircraft delays, so take a look at the cor() API help page to see how you can work around them. Return the correlation matrix rounded to two decimal points.

(ii) Using corrplot() (from package "corrplot") , plot the correlations between DepDelay and each of the five delays.

(iii) The corrplot() API from 2(d)(ii) returns a list containing three named elements: "corr", "corrPos", and "args". Focus on the "corrPos" element (which is a data frame), print it out and see what it contains. Extract the rows that contain the correlation of "DepDelay" with all of the other elements and save them to a data frame; your data frame will contain 6 rows and 5 columns. Sort this data frame using the column "corr" with the smallest correlation value at the top and the largest at the bottom. Your function should return this data frame. (Do not perform any rounding.)

(iv) Examine the data frame in 2(d)(iii) and find the delays that are correlated positively and negatively with "DepDelay". Return a string in the format: "strongest positive=X,strongest negative=Y", where X is delay that is positively correlated, and Y is the delay that is negatively correlated with "DepDelay". Note: Make sure X and Y are lowercase, i.e., "SecurityDelay" should be specified as "securitydelay".

---

1 If you are curious to see which airlines these are, go to the following web page to translate the airline code (UniqueCarrier) to the airline name: https://www.iata.org/en/publications/directories/code-search/