ETC5512: Instructions for Assignment 4

Di Cook

Due date: Jun 19 2020 12.00pm (noon)

Overview of problem

This assignment encourages you to explore the US airline traffic data base in more detail. This builds on the tutorial exercises started in weeks 10 and 11.

Turn in

There are multiple parts with specific exercises to complete. There isn't a report format this time. You need to answer the questions using an Rmarkdown document that compiles to an html output, similar to the weekly tutorials. In addition, you will need to provide a set of data files, that you construct.

A zipped archive of a template folder is provided. The folder you need to turn in will have this structure:

```
|--- analysis
|   |--- LastName-FirstName.html
|   |--- LastName-FirstName.Rmd
|--- data
|   |--- airlines.rda
|   |--- airports.tda
|   |--- flights.rda
|   |--- planes.rda
|   |--- planes.rda
|   |--- weather.rda
|   |--- city.rda
|   |--- myairport_mymonth_mycarrier.rda # Change the name to reflect your subset  choice
|--- raw_data # These files will be large - DON"T INCLUDE IN YOUR ZIP
|   |--- On_Time_Reporting_Carrier_On_Time_Performance_(1987_present)_2020_2.csv # Example
|   |--- 402312038_T_MASTER_CORD.csv # Airports data file
|   |--- other raw files
|--- Assignment4_instructions.Rmd # Provided to you
|--- Assignment4_instructions.pdf # Provided to you
|--- README.md # Summary of files in the folder
|--- etc5512-assignment4.Rproj
```

Notes

- In the Rmd file, include your code to complete the analysis, but it should be hidden. The code processing the raw data should be in a chunk with `eval=FALSE` but it should be clear how you constructed your subsets of data, to a level that someone else can easily re-do your work.
- Plots should be nicely themed, and easy to read. Where necessary include titles indicating what's plotted. Code should be documented with short comments in the script. Short sentences or paragraphs should overview your steps for each part or each exercise.
- The data files need to be provided as `.rda` files. When you create the data in R, use the function `save(mydata, file="data/mydata.rda")` to create these files.
- The `README.md` provides an overview of your data files.
- *IMPORTANT: Your raw_data folder should be empty for turning in! Use this folder to store the downloaded files, to keep your work organised.*

Marks

The total assignment is worth 20 points.

- Reproducibility and readability (2 points)
- Exercise 1 (4 points)
- Exercise 2 (6 points, 2 points for part e)
- Exercise 3 (6 points, 0.5 point for each of air traffic tables, 1 point for the weather data, 0.5 point for city data, 1 point for scraping code for city information, 1 point for completing README.md file with overview of files, 0.5 point for correct variables and types.)
- Exercise 4 (2 points)

Exercises

1.

a. Choose a month of flight data from the ontime database, between March 2019 through March 2020 that corresponds to your **birthday** month. Subset the data to ONE of the major carriers, eg, DL, UA, WN, AA, OO. Choose ONE MAJOR airport, eg ATL, ORD, DFW, DEN, CLT, LAX, PHX, IAH, LAS, SFO, DTW, MCO, BOS, MSP, SEA, DCA, IAD, MIA, SLC, and subset flights into and out of the airport. *You need to report the name of the flights data file, and your choices of month, carrier and airport.*

b. Download the table on airport information. Join the flights data to the airports information. Save this data to your data folder.

c. Plot all the flight paths on a map. Write a sentence or two summarising the extent of this carrier's service to and from the airport.

2.

The overall goal for this exercise is to make a movie of air traffic into and out of an airport during the course of one day. Completing the steps building up to this will help you work sequentially to this goal.

a. Convert all times into a standard time unit. Compute a new variable, which is the number of minutes from midnight (NYC time).
b. Select a day, and one plane, that has flown into and out of your chosen airport in the day. Show the flight path of this plane.
c. Break the paths down into steps of 15 mins, plot these as dots, coloured by time.
d. Animate the flight path of this plane, showing its position every 15 mins.
e. Extend your animation for all flights for the one carrier for the day. The display should also have (1) a representation of the time, (2) delay indicated.

3.

Motivated by the `nycflights13` data, your task here is to construct a new data set that has your choice of airport, one month of data, and ALL CARRIERS, with separate tables made available as `.rda` files. The same variables as the `nycflights13` should be provided.

- data: contains the compiled data, including these files
  - flights: Use `?flights` to see variables to be included
  - airlines: carrier, name
  - airports: Use `?airports` to see variables to be included
  - planes: Use `?planes` to see variables to be included
  - weather: Look at the help for this table `?weather` to see how the weather data was compiled, and the variables ideally included

  - city: This table provides information about the city where the airport is located, as recorded from the Wikipedia page (1 point reserved for providing the code to scrape this information) and will have these variables:
    * faa: same as the airport table
    * pop: 2019 metro population, maybe estimated
    * mayor: name of current mayor
    * area: Metro in sq miles
    * zip: zip codes, may be several, provided as a text string
    * fip: FIPS code

4.

Write a paragraph (less than 100 words) explaining how your data files from the previous question are provided in a format that lends itself to analysis. (See the care and feeding of wild data week of material.)