STAT 19000 Optional Projects

Topics: Optional Projects

How to format your solution: Use the files:

`stat19000optionalproject1.R`

`stat19000optionalproject2.R`

`stat19000optionalproject3.R`

`stat19000optionalproject4.R`

from the `/class/datamine/data/examples` folder.

You are welcome to use a combination of R and UNIX to solve the questions.

——— Optional Project 1 ———

Question 1:

Use UNIX to extract the states, airport codes, and departure delays across all years of the flight data.

1a. Which 5 states have the most flight departures? (i.e., which 5 states are the origins of flights most often?) How many flight departures do each of those 5 states have?

1b. Which 5 airport codes have the most flight departures? (i.e., which 5 airport codes are the origins of flights most often?) How many flight departures do each of those 5 airport codes have?

Question 2:

2a. Which 5 states have the longest average departure delays? How long are each of those longest average flight delays, corresponding to those 5 states? You probably will want to extract the relevant data from the flight files, using UNIX, and import this smaller data into R. It will be challenging to import all of the full flight data sets into R.

2b. Treating the airplane tailnum and the origin airport as a pair, find which airplane departed from which airport, the most number of times altogether. How many times has that airplane departed from that airport? (Hint: It will be necessary to remove all of the unknown tailnums. We want an actual airplane and origin airport for the solution.)

Question 3:

3. Make a map of the entire continental United States, using your Google API Key and Google Maps.

Now add some black dots to the map, indicating where the airports in the Great Plains are located. You can get the data for the latitudes and longitudes of the airports from here:

http://stat-computing.org/dataexpo/2009/airports.csv

The black dots that we want to add are *only* for airports in these states: Colorado, Kansas, Montana, Nebraska, New Mexico, North Dakota, Oklahoma, South Dakota, Texas, and Wyoming. Do not add airports for any other states on your map.

——— Optional Project 2 ———

Question 1:

1a. How long is the longest review in the Amazon database? ("Longest" means longest in terms of characters.)

1b. What music product does it review? Find the original music product on Amazon, and give a link to that music product in your solution.

Question 2:

2a. How many album reviews mention both John Coltrane and Miles Davis in the same review?

2b. Choose ten artists who each have one-word names (e.g., Madonna, Adele, Sting, etc.), and make a dotchart that shows the number of reviews for each of these artists. (You can choose which artists.)

2c. How many reviews mention the word jazz? Classical? Rap? Make a dotchart that shows the number of reviews for each of (at least) 10 genres of music. (You can choose which genres.)

Question 3:

3. Make a chart that, for each year from 1959 to 2018, shows the number of reviews that mention that year somewhere in the review. Your chart should have 60 points on it (one per year). The x-axis should show the year and y-axis should show the number of reviews that mention that year.

——— Optional Project 3 ———

Question 1:

1a. In the election data, across all years, how many contributions were $10.00 or less ? (For the sake of this question, do not include negative contributions. Only include amounts in the range of 0 dollars to 10 dollars, inclusive.)

1b. What is the sum of the donation amounts from 1a?

1c. How many donations were $500 or more?

1d. What is the sum of the donation amounts from 1c?

Question 2:

2a. For each election file, from 1980 to 2020, find the percentage of donations in that election cycle that were in the range $0 to $10.00. Make a plot to illustrate this trend.

2b. Find a donor who made a substantial number of donations over the years. Characterize their giving pattern over time (for this one donor of your choice), using a visualization.

Question 3:

3. Make a chart that shows the respective amounts of giving – per election cycle – in each of the 5 Midwest states: IN, IL, MI, OH, WI. How have these states' giving patterns changed over the years? (This question is intentionally a little bit open to your interpretation.)

——— Optional Project 4 ———

Question 1:

1a. Taking into account all AirBnB properties in California, which individual review is the longest (in terms of the number of charaters)? What does that longest review say? (Please display the contents of the longest review among all California properties. It will be necessary to search the files for all 8 cities and counties in the AirBnB California data.)

1b. Within the Los Angeles listings (only), do you think that there is any correlation between the length of the listing (i.e., the review length in characters) and the price?

Question 2:

2a. Consider these states: Massachusetts, New Jersey, New York, and Rhode Island. Compute the average length of an AirBnB review (in characters) for each state. Make a dotchart that shows the average length of a review in each of these states.

2b. Plot all of the AirBnB listings from the four states in question 2a on a Google Map (using your Google API key). Center your map in such a way that it mostly shows the northeast region of the country, where these four states are located.

Question 3:

3. Use rbind to bind together the data frames from all 8 California `listings.csv` files. Then use the tapply function to discover which 5 neighbourhoods have the most expensive listings, on average, across all California properties.

Project Submission:

Submit your solutions for the project(s) at these URLs:

Optional Project 1: https://classroom.github.com/a/deupxJQ2

Optional Project 2: https://classroom.github.com/a/VVWx8SbO

Optional Project 3: https://classroom.github.com/a/T_jWPpsR

Optional Project 4: https://classroom.github.com/a/G1WbAKuI

using the instructions found in the GitHub Classroom instructions folder on Blackboard.