30535 Applied Problem Set 3

Peter Ganong

03/23/2020

June 3, 2020 at 4:59PM.

Name your files `applied_ps_3.Rmd` and `applied_ps_3.html`. (5 pts.)

Follow the style guide (10 pts.)

This submission is our work alone and complies with the 30535 integrity policy.

Add your initials to indicate your agreement: **___**

Add names of anyone you discussed this problem set with: **___**

Submit by pushing your code to your repo on Github Classroom: https://classroom.github.com/g/5vDiXToZ.

waze data

- You can find the waze data dictionary **here**.
- At the start of the course that you agreed to follow **these** data usage terms. Here are the most important parts:
  - you may download the data onto your computer
  - you will delete the data at the end of the quarter
  - you agree not to use the data to create a competitor to Waze
  - you agree not to share the data or your analysis[1]

Prelim questions

1. Have you deleted any Waze data that you downloaded onto your computer (answer this at the end of the problem set in the affirmative)?

1 Waze data start-up (5 points)

Working with data on a server adds a challenge as you have to make calls to the database which take time to process. A call to the database can be slow for several reasons.

1) the data you are trying to pull is very large.
2) many people are making requests at the same time.
3) something else is going on.

We can adjust for 1 and 2 by testing our code on small subsections of the data.

Next week we will provide an opt out where you can use `csv` we provide. Using this option will result in a 10 percent discount on your problem set final grade. For example, if you earn 90 pts based on your solutions, your final grade will be $90 \cdot .9 = 81$.

1. Which of the following methods will cause problems as you develop your solutions?

[1] Unless you opt out, I am planning to submit your work for this problem set to Waze for disclosure review so that you can include your work in your portfolio going forward.

a. Use `filter()` to reduce the amount of data you pull while exploring data. For example, you can filter by time and location to only get data for a small part of the city and/or over a short time period.
b. `collect()` a small sample data set so that the you have data in memory on your computer.
c. `collect()` the entire data set each time you want to work with it.

2. As is the case with any data set, Waze has to make decisions about what data to store and how to measure it. Review the data documentation and the rest of the problem set. Propose a variable that Waze could feasibly track that is not available now or feasible and better way a to measure a variable currently in the dataset. Support your proposal with reasoning.

3. As is the case with most consumer data, Waze users are self-selecting. Write a few detailed sentences about how you think self-selection influences what data is present.

2 Waze vision zero (15 points)

Read up on the `ggmap` package, which will be useful for doing these problems. Particularly, get to know the `get_stamenmap()` function. If you find yourself downloading 1000s of tiles, check your settings. You are welcome to try using google basemaps as well; while free for new users, this will require a credit card. The version of `ggmap` on CRAN is out of date, instead find and install it from github.

1. Look at Vision Zero Neighborhood High Crash Corridor #7. Plot the accidents in this corridor on a map.

2. Around what intersection are accidents most common? Use Google Street View to look at this intersection. Do you see any problems?

3 Transit Oriented Development (15 points)

1. On October 21, the City of Chicago declared the 79 and 66 bus routes as areas of focus for transit oriented development. The City says the plan addresses bus "slow zones". Note: Watch out for "179th St".

   a. For each corridor, plot traffic alerts by time of day.
   b. Using a reasoned approach, choose two additional corridors for comparison.
      i. What corridors did you choose and why?
      ii. Make comparison plots.
   c. Looking beyond traffic, what other alerts are very common in this area? Do you think these alerts would slow down the 66 / 79? If so, what steps could the City take to address the issues?

4 Waze single event (20 point)

1. Revisit the event which caused c5a73cc6-5242-3172-be5a-cf8990d70cb2.

   a. Define a bounding box around the cause of the event.
   b. What causes all these jams? Some googling might help.
   c. Plot the number of jams 6AM-6PM CST. Why are there two humps?
   d. Place one vertical line at each hump.
   e. Next, propose a quantitative measure of traffic jam severity that combines the number of traffic `JAM` alerts with information in the `subtype` variable.
   f. Plot this measure from 6AM-6PM CST. Is there any information that is conveyed by your severity measure that was not captured by plotting the number of jams? If so, what is it?

5 Waze aggregate over multiple events (30 points)

1. Pick one major accident. What is the uuid? Sample alerts from the two hours before the accident first appeared in the data and two hours after the accident for a geographic box of 0.1 miles around the accident. Make a plot where the y-axis is the number of traffic jam alerts and the x-axis is the five-minute interval from two hours before the accident to two hours after the accident. Warning: This question is harder than it first appears. You might want to review R4DS chapter 12.5 (lecture note 5) on missing values and chapter 16.4 (lecture note 9).

2. Building on your work for the prior question, write a function that takes as its arguments `uuid`, a `date-time`, a latitude and a longitude and returns a data frame with the number of alerts in each five-minute interval from two hours before to two hours after.

3. Make a data frame with every major accident on Nov 20, 2017. Feed each row of this data frame to your function. Collapse the output into the mean number of traffic jam alerts in each five-minute interval in the two hours before the accident and two hours after the accident for a geographic box of 0.1 miles. Tip: This may take upwards of 20 minutes to run on all major accidents. Use your function on a small sample of accidents first to make sure your code is working as expected before trying to run on all accidents.

4. Plot the mean number of jam alerts around major accident. To be clear, the correct answer here is a single plot that summarizes jams across major accidents, not one plot for each accident. Congratulations! This is your first event study.