30538 Problem Set 1: Parking Tickets

Peter Ganong, Maggie Shi, and Ozzy Houck

2024-09-30

1. **PS1:** Due Sat Oct 5 at 5:00PM Central. Worth 50 points. Initiate your **repo**

We use `(*)` to indicate a problem that we think might be time consuming.

Steps to submit (5 points on PS1 and 10 points on PS2)

1. "This submission is my work alone and complies with the 30538 integrity policy." Add your initials to indicate your agreement: **___**
2. "I have uploaded the names of anyone I worked with on the problem set **here**" **___** (1 point)
3. Late coins used this pset: **___** Late coins left after submission: **___**
4. Knit your `ps1.qmd` to make `ps1.pdf`.
    - The PDF should not be more than 25 pages. Use `head()` and re-size figures when appropriate.
5. Push `ps1.qmd` and `ps1.pdf` to your github repo. It is fine to use Github Desktop.
6. Submit `ps1.pdf` via Gradescope (4 points)
7. Tag your submission in Gradescope

Background

Read **this** article and **this** shorter article. If you are curious to learn more, **this** page has all of the articles that ProPublica has done on this topic.

PS1

Read in one percent sample (15 Points)

1. To help you get started, we pushed a file to the course repo called `parking_tickets_one_percent.csv` which gives you a one percent sample of tickets. We constructed the sample by selecting

ticket numbers that end in `01`. How long does it take to read in this file? (Find a function to measure how long it takes the command to run. Note that everytime you run, there will be some difference in how long the code takes to run). Add an `assert` statement which verifies that there are 287458 rows.

2. Using a function in the `os` library calculate how many megabytes is the CSV file? Using math, how large would you predict the full data set is?

3. The rows on the dataset are ordered or sorted by a certain column by default. Which column? Then, subset the dataset to the first 500 rows and write a function that tests if the column is ordered.

Cleaning the data and benchmarking (15 Points)

1. How many tickets were issued in the data in 2017? How many tickets does that imply were issued in the full data in 2017? How many tickets are issued each year according to the ProPublica article? Do you think that there is a meaningful difference?

2. Pooling the data across all years what are the top 20 most frequent violation types? Make a bar graph to show the frequency of these ticket types. Format the graph such that the violation descriptions are legible and no words are cut off.

Visual Encoding (15 Points)

1. In lecture 2, we discussed how Altair thinks about categorizing data series into four different types. Which data type or types would you associate with each column in the data frame? Your response should take the form of a markdown table where each row corresponds to one of the variables in the parking tickets dataset, the first column is the variable name and the second column is the variable type or types. If you argue that a column might be associated with than one type, explain why in writing below the table.

2. Compute the fraction of time that tickets issued to each vehicle make are marked as paid. Show the results as a bar graph. Why do you think that some vehicle makes are more or less likely to have paid tickets?

3. Make a plot for the number of tickets issued over time by adapting the Filled Step Chart example online. Go back to Bertin's taxonomy of visual encoding, which we discussed in lecture. What visual encoding channel or channels does this use?

4. Make a plot for the number of tickets issued by month and day by adapting the Annual Weather Heatmap example online. What visual encoding channel or channels does this use?

5. Subset to the five most common types of violations. Make a plot for the number of tickets issued over time by adapting the Lasagna Plot example online. What visual encoding channel or channels does this use?

6. Compare and contrast the plots you made for the prior three questions. What are the pros and cons of each plot?

7. Suppose that the lesson you want a reader to take away is that the enforcement of violations is not evenly distributed over time? Which plot is best and why?