30538 Problem Set 2: Parking Tickets

Peter Ganong, Maggie Shi, and Richard Chen

2026-01-07

Due Sat Jan 17 at 5:00PM Central.

"This submission is my work alone and complies with the 30538 integrity policy." Add your initials to indicate your agreement: **___**

Github Classroom Assignment Setup and Submission Instructions

1. **Accepting and Setting up the PS2 Assignment Repository**
   - Each student must individually accept the repository for the problem set from Github Classroom ("ps_2") – https://classroom.github.com/a/edXgVAAq
     - You will be prompted to select your cnetid from the list in order to link your Github account to your cnetid.
     - If you can't find your cnetid in the link above, click "continue to next step" and accept the assignment, then add your name, cnetid, and Github account to this Google Sheet and we will manually link it: https://rb.gy/9u7fb6
   - If you authenticated and linked your Github account to your device, you should be able to clone your PS2 assignment repository locally.
   - Contents of PS2 assignment repository:
     - `ps2_template.qmd`: this is the Quarto file with the template for the problem set. You will write your answers to the problem set here.
2. **Submission Process (5%)**:
   - Knit your completed solution `ps2.qmd` as a pdf `ps2.pdf`.
     - Your submission does not need runnable code. Instead, you will tell us either what code you ran or what output you got.
   - To submit, push `ps2.qmd` and `ps2.pdf` to your PS2 assignment repository. Confirm on Github.com that your work was successfully pushed.

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. See the documentation from Propublica's Github on these data **here** and **here**

Visualization guidelines (5%)

To receive full credit on visualizations throughout the problem set, your visualizations should be coded using `altair`, and should:

- Explicitly declare encodings and data types within `altair`
- Format the graph such that:
  - All axes and units are properly labeled and legible
  - No words or data points are cut off in your compiled PDF
  - Variables should be encoded in a sensible/intuitive way

Read in and characterize data (30%)

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 in 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.
4. For each column, how many rows are `NA`? Write a function which returns a two column data frame where each row is a variable, the first column of the data frame is the name of each variable, and the second column of the data frame is the number of times that the column is NA. Test your function. Then, report the results applied to the parking tickets data frame. There are several ways to do this, but we haven't covered them yet in class, so you will need to work independently to set this up.
5. Three variables are missing much more frequently than the others. Why? (Hint: look at some rows and read the data dictionary written by ProPublica)

6. 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?

Visual encodings (10%)

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.

Understanding the structure of the data and summarizing it (20%)

1. Many datasets implicitly contain information about how the data are generated. In this setting, we can use the data to learn how a case progresses.

- Draw a diagram explaining your understanding of the process of moving between the different values of `notice_level`, and how you used the data or other sources to arrive at this conclusion. If you draw it on paper, take a picture and include the image in your write up.
- Draw a second diagram explaining the different values of `ticket_queue`. If someone contests their ticket and is found not liable, what happens to `notice_level` and to `ticket_queue`? Include this in your diagram above.

Aggregating, transforming, visualizing the data (30%)

1. 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.
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. List two observations or takeaways that jump out from this plot. Go back to the taxonomy of visual encodings 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. List two observations or takeaways that jump out from this plot. 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. Explore what the use of `color_condition = alt.condition(...)` does in the plot, and describe below. List two observations or takeaways that jump out from this plot. 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? What kinds of questions is each plot best-suited for answering?

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?