30538 Problem Set2: Parking Tickets

Peter Ganong, Maggie Shi, and Ozzy Houck

2024-09-30

1. **PS2:** Due Sat Oct 19 at 5:00PM Central. Worth 100 points.

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

Steps to submit (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**" **___** (2 point)
3. Late coins used this pset: **___** Late coins left after submission: **___**
4. Knit your `ps2.qmd` to a pdf named `ps2.pdf`.
   - The PDF should not be more than 25 pages. Use `head()` and re-size figures when appropriate.
5. Push `ps2.qmd` and `ps2.pdf` to your github repo. It is fine to use Github Desktop.
6. Submit `ps2.pdf` via Gradescope (8 points)
7. Tag your submission in Gradescope

Background Recap

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. This problem set is a continuation of PS1 using the same data. Please start by loading the data in the same way as PS1.

Data cleaning continued (15 points)

1. 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.

2. Three variables are missing much more frequently than the others. Why? (Hint: look at some rows and read the data dictionary written by ProPublica)
3. Some of the other articles on the propublica website discuss an increase in the dollar amount of the ticket for not having a city sticker. What was the old violation code and what is the new violation code?
4. How much was the cost of an initial offense under each code? (You can ignore the ticket for a missing city sticker on vehicles over 16,000 pounds.)

Revenue increase from "missing city sticker" tickets (35 Points)

1. Using pandas, create a new value for violation codes which combines the two codes that you found in the previous question. Again using pandas, collapse the data to capture the number of missing city sticker tickets by month. Then, using Altair, plot the number of tickets over time.
2. Suppose that your reader wants to be able to use the plot to deduce when the price increase occurred. Add frequent or custom date labels on the x-axis of your plot such that the date of the price increase is readily apparent. We haven't covered Altair's date labeling features in class so you'll first need to find the relevant help page in the documentation. Which help page did you use?
3. The City Clerk said the price increase would raise revenue by $16 million per year. For now, ignore the fact that many tickets are not paid and assume that the number of tickets issued is the same before and after the policy change. Using only the data available in the calendar year prior to the increase, how much of a revenue increase should they have projected? Remember that you are working with a one percent sample of the data.

   Assume that the number of tickets of this type issued afterward would be constant and you can assume that there are no late fees or collection fees, so a ticket is either paid at its face value or is never paid.
4. What happened to repayment rates (percentage of tickets issued that had payments made) on this type of ticket in the calendar year after the price increase went into effect? Suppose for a moment that the number of tickets issued was unchanged after the price increase. Using the new repayment rates in the year after the price increase occurred, what would the change in revenue have been?
5. Make a plot with the repayment rates on "missing city sticker" tickets and a vertical line at when the new policy was introduced. Interpret.

6. Suppose that the City Clerk were committed to getting more revenue from tickets. What three violation types would you as an analyst have recommended they increase the price of? Consider both the number of tickets issued for each violation type and the repayment rate for each violation type. You may assume there is no behavioral response to price changes (ie. people continue to commit violations at the same rate and repay at the same rate). Make a plot to support your argument and explain in writing why it supports your argument.

Headlines and sub-messages (20 points)

1. The City Clerk has now begun to wonder... maybe raising ticket prices will lead to a decline in repayment rates after all. Make a data frame where each row is a violation description, the fraction of time that the ticket is paid, and the average level 1 fine. Sort this dataframe based on how many total tickets of each type have been issued. Print the rows for the 5 most common violation descriptions.
2. Make a scatter plot which shows the relationship between fine amount and the fraction of tickets that are paid. Focus only on violations that appear at least 100 times. There will be one outlier with a high fine and you can exclude that ticket type from the plot. Then make two other plots which show the same relationship in different ways. For all three plots, write out what are the headlines and what are sub-messages.
3. The City Clerk doesn't understand regressions and only has time to look at one plot. Which plot are you going to bring to them and why?

Understanding the structure of the data and summarizing it (Lecture 5, 20 Points)

1. Most violation types double in price if unpaid.
   - Does this hold for all violations?
   - If not, find all violations with at least 100 citations that do not double. How much does each ticket increase if unpaid?
2. Many datasets implicitly contain information about how a case can progress. Draw a diagram explaining the process of moving between the different values of `notice_level` (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 tree drawings above.

3. Go back to your scatter plot from the previous section. We want to add labels to each dot (which conveniently you constructed in the previous step). Implement this in two ways: (a) label every dot with adjacent text or (b) put the text in a legend. Either way, you will find the same problem – there are too many labels and the plot is illegible. Revise the plots. First, do this the easy way, which is to pick the ten most commonly used violation descriptions and mark all the other dots as "Other". Second, for (b), try to construct meaningful categories by marking violation descriptions which sound similar with a common label and a common color.

Extra Credit (max 5 points)

1. Which violation codes, if any, are associated with multiple violation descriptions? In these cases, using pandas, create a new column which records the most common violation description associated with this code. If there are any with multiple descriptions print the three codes with the most observations.

2. Above you made a diagram on paper to show how a case can progress. Although Vega-Lite cannot support tree layout plots like this, Vega can!. Make the digital version of your diagram using Vega. You can use the (online Vega console](https://vega.github.io/editor/#/custom/vega). Submit your `JSON` as a separate file and submit your plot alongside your pset as a `.png`.