Brandon Coleman

Data Analytics Bootcamp

Project 1

Team Yellow

This part of the project will analyze the LA Parking Dataset by focusing on violation. It will

Include visualizations such as Bar charts, and stacked Bar charts. It will run a statistical analysis

on violations by location via the chi-square test.

I began by importing my dependencies, matplotlib, pandas, scipy and numpy.

Find Data:

We searched Kaggle for datasets that would be interesting to work with. Initially that would

have been the fitbit api but the api would only allow one to read in the data for one, personally owned, fitbit and not an entire database. When that fell through we settled on Kaggle’s own

Maintained dataset of LA Parking Citations.

Clean Data:

I was tasked with cleaning the data violations - the laws that were broken. When I began to

clean the violation column, using the nunique function call, I saw that for the entire dataset

There were 830 unique violations. There were ones that were just a sequence of numbers

And letters, and I wasn’t able to find any further information for them online. Because of that and being that they were mostly single entries, I decided to drop them from the dataset.

I then used the replace function to clean up my remaining violations. Ones that meant the same,

But were spelled differently, or were similar enough to group together. THere isn’t one way to clean up a dataset, and particularly descriptive data, but I tried to make it as concise and clear as possible. I whittled it down to around 116 violations. If I needed to do so even further I could have.

Analysis:

From there the most logical question was what are the most commonly ticketed citations?

Another obvious question was which violations resulted in the steepest fines?

Another curious question I had was what was the distribution of fines of a certain car.

Finally, I wondered if there was any correlation between location and violations cited.

For the first two questions my analysis had me perform a groupby of violations on my cleaned database. From there I used the count function in addition with the nlargest function to get the 10 most ticketed violations. For the steepest fines, I also used the nlargest function chained to

a mean function on the ‘Fine amount’ column. Storing both of these results in a series, I then created bar charts for both.

For my third question, of an example of Tesla’s violations I grabbed the make column where ‘Make’ was equal to TESL. A value\_counts call on ‘violation description’ resulted in TESLA’s largest violations, where I kept the top 10. I created a stacked bar chart using a for loop to plot each violation one on top of the other.

Lastly my statistical analysis was a chi-square test of an observed dataframe of the top 10 violations at the top 10 addresses. The null hypothesis was that violations and locations are

Independent of each other. My critical value was calculated with a degree of freedom of 99 because of the 100 rows - 1. Because my resulting chi-square was higher than my critical value

of 123.25 we can reject the null hypothesis and conclude that where you are in LA matters

as to what you might be in violation of.

Things I learned:

* Working in a group on a dataset this large was an experience. A first ever experience.
* I learned to use github and pull and merge data. I learned to coordinate schedules

and communicate with teammates and the benefits of tools like slack and git bash.

* I learned about statistical analysis, maybe not completely but the basics.
* I learned how to create a stacked bar chart. I’ve learned for the future that there’s always

More you can do on a project and it’s rewarding to go ahead and get your hands dirty.

“Move fast and break things”

Things I would do if I had more time:

* Making the database searchable, and filtering across different columns.
  + Keying in a car model and returning the violations most tickets
  + Looking for trends across time : can we deduce what day of the week street cleaning happens on?
  + A heatmap as Kevin suggested, of violations across zipcodes