Michael Hayden
November 2018 Cohort
Springboard

Data Rubric Story

The data narrative for the parking project began early on in the data harvesting process. My goal in completing this assignment was to cover the programs I used and explain how the python libraries were utilized to find the necessary information. Upon completion of tasks for the project thus far, I feel that I entered the course with an incorrect or shallow understanding of what data science involves. In hindsight, I am pleased to have entered into the program with an incomplete understanding because my newfound skillset and more holistic view of data science has been grown through hard work and knowledge acquisition efforts.

After getting a ticket for being parked on a street that needed street cleaning, I decided to learn more about the parking data in the Chicagoland region. My desire to learn more about the world around me using data was bolstered by my intent to figure out the probability of getting a ticket if I parked my car illegally for a set number of hours. I wanted to base the conclusions from this study on the historical parking ticket data in the area. Upon starting this data science course, I quickly realized that I need a firmer foundation in data analytics before I can dive deeper into my capstone project and into understanding how to use data to solve problems I encounter daily.

The first hurdle that I had to overcome was getting the data set to upload to Github. After many days with little success, my mentor and I decided to break down the data set into smaller quantities. After breaking it down, we were able to produce a sample of the data set that was smaller than 2GB. At this point, the dataset was loaded to Github using Gitlfs.

While looking at the data early on, we figured out that the zip code of the data was not showing up in the proper format. To fix the zip code issues, we decided to use Geocoder. While using geocoder, we set up a function that would take the ticket issue address then add Chicago, IL to the end of the address geocoder which would append the zip code to a new list. It took a while to get the data in the correct format. Some of the initial problems we faced included time-out errors and struggling with the function crashing randomly. After a few weeks of

fidgeting with the data and reducing the data down to a single day, we were able to produce the correct zip codes.

The only problem with using the zip codes in the data was that we were planning on using Folium to create a heat map of the ticket issue locations. This presented an issue because Folium works with latitudes and longitudes rather than zip codes. We then had to re-run the code with Geolocator to produce the latitude and longitude coordinates. Finally, using Folium we were able to produce a heat map of the parking tickets in the Chicago area.

After seeing the completed map, the visual prompted a new question: are there more tickets issued near the police stations in comparison to areas without increased police presence? To answer this, I found a public data set of police station locations in Chicago given in latitude and longitude coordinates. I imported the dataset, then used the children method of Folium to place the fifteen or so police stations on the map. The process of importing the second data set was a lot quicker for me because I was able to practice the same technique as before. I found that there was a loose correlation between more tickets being given around police stations. I found that the driving factor behind there being more tickets in some areas over other areas had to do with the type of parking available. The areas with fewer tickets issued, were typically safer areas with lower crime rates and areas with free street parking.

I did not expect to find the number one ticket area to be at O'Hare Airport. I originally thought that the area with the most tickets would be downtown. Through my research, I did find that there are special tickets that can be given downtown for parking during rush hour. Finally, I broke down the number of tickets per hour. After studying the trend, it appears that most ticket issuers start their shift at 6 AM and finish at 3 PM.

While my mentor and I were brainstorming about how to create a machine learning aspect to the data we decided upon running some basic naive based statistics on the provided info. I wanted to learn more about my daily illegal parking habits and the probability of acquiring a ticket. To figure out the probability, I recreated June 9, 2017, and then broke down tickets by zip code. Since I typically park in 60657 zip code and I can't be parked on that street from 6 pm to 6 am without risking getting a parking ticket I investigated this further.

My first findings from this data showed that 8,052 tickets were issued that day where 449 of that total were issued within the 60657 zip code. At this point, I was becoming a little concerned because this averaged out to over 15 tickets per block in the neighborhood. While this prompted concern due to the high ticket total per block radius, I was reminded that tickets could only be issued from 6 pm to 6 am. I broke this data set down further to see how many of the 449 tickets were issued between 6 pm and 12 am of that day. I found that 21 tickets were issued within that six hour time frame. Overall, 4.6% of the tickets issued that day were given the time that I spent illegally parked. After analyzing this data, I feel like I can confidently take on those odds.