TE360
Ziheng (Jack) Chen
11/27/2022

Final Project Report

Hypothesis

The traffic in Chicago is terrible, at least that's what I was told. Like many other metropolitans, there are countless cars on the already ancient and narrow roads, and pedestrians are everywhere, making the driving condition even worse. As a result, aggressive drivers are commonplace. However, For the weeks that I worked in the city, I could not find any evidence to support the previous statement. Specifically, when I went out to a restaurant for lunch on a workday, I did not notice the overflowing amount of cars on the road, and I did not feel the need to stay vigilant of the passing cars for aggressive behavior. Could it be that the density of businesses has little effect on the quality of traffic?

Along with examining the location distribution of companies, the goal of my project is to find out whether the density of businesses negatively affects traffic conditions. My opinion was certainly biased since I did not experience Chicago's traffic prior to COVID, and the city currently is not operating at its maximum. Therefore, for this project, I used the dataset provided by the city of Chicago to determine how bad the traffic actually is over the last three years. More specifically, I determined the quality along two criteria: congestion and the number of crashes. In addition, using the same congestion data, I will examine the congestion throughout the day across the regions that lie between my home and the city.

The result of this project, besides satisfying personal interests, can also help other people plan their routes. Furthermore, it can be informational for newcomers to choose an office in the city that can be reached with relatively low traffic.

In this project, I will be using the following dataset from the Chicago Data Portal that contains the listed fields respectively:

1. Chicago Traffic Tracker - Congestion Estimates by Regions
    a. Congestion information in units of miles per hour of the latest measurement
    Note: "For congestion advisory and traffic maps, this value is compared to a 0-9, 10-20, and 21 & over scale to display heavy, medium, and free flow conditions for the traffic segment. Although expressed in miles per hour, this value is more a reflection of the congestion level in the region than it is indicative of the average raw speed vehicles are traveling within the region.
    b. Region information in terms of latitude and longitude."
2. Chicago Traffic Tracker - Historical Congestion Estimates by Region - 2018-Current
    a. Congestion information in units of miles per hour since 2018
    b. Region information in terms of latitude and longitude
3. Traffic Crashes - Crashes
    a. Crash location in terms of latitude and longitude

4. Business Licenses
    a. Business location in terms of latitude and longitude

Model

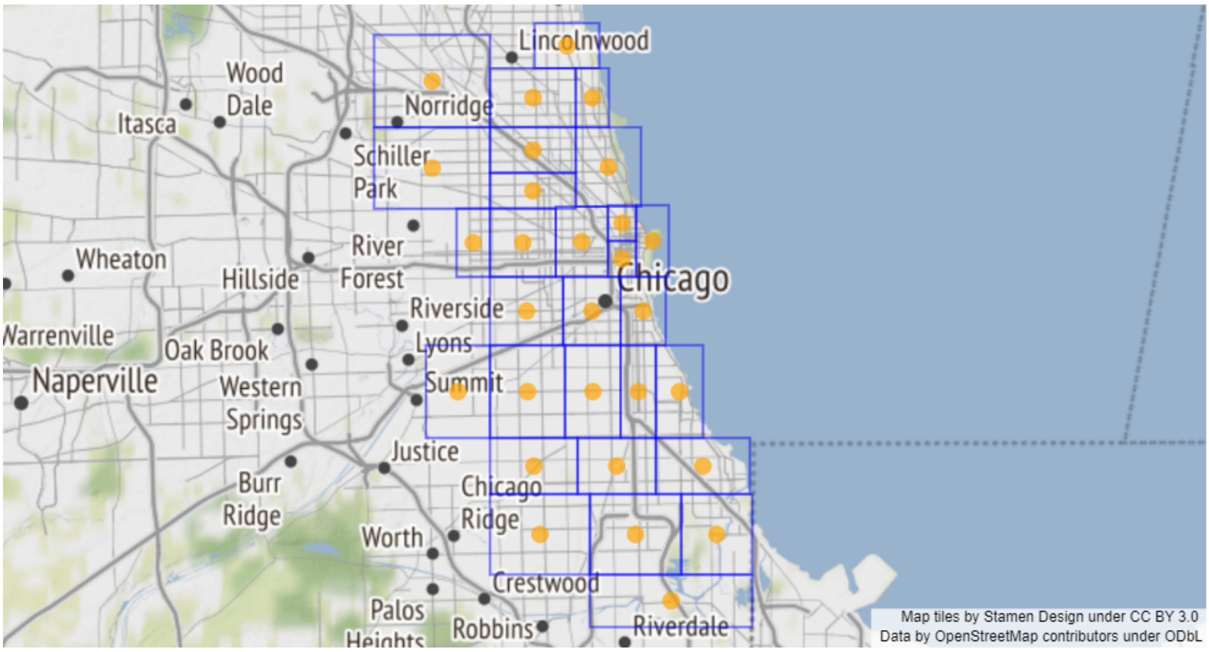
Originally, my plan was to gather congestion data by segment, which provided much more detail than with regions. However, the approach quickly ran into a problem. There are more than 1000 segments in the city, and each segment has a 3-year worth of data, so any manipulation of the data took a long time. Furthermore, having such fine-grained data could also bring too much detail for me to overlook the general trend. Therefore, I switched to using regions instead of segments. There are only about 30 regions dividing up the city, and the geographical locations are plotted later.

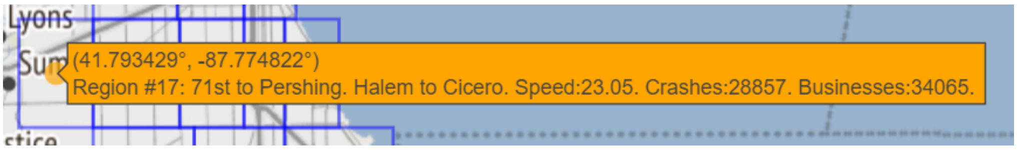
Although using regions provides benefits, it also closes off some opportunities. Since each region covers a large area, it is impossible to tell which road segments are consistently congested and should be avoided. Therefore, I could not achieve my goal of finding out the route that avoids the most congestion as I mentioned in my proposal. To make up for the loss, I decided to append the best timing for the lease congestion using a regression method.

With respect to the above considerations, my approach was rather simple. In order to examine the relationship between the number of businesses and traffic, after importing and cleaning the datasets, I first need to translate the location information of each crash and business license to a corresponding region. Besides geographical modifications, I also need to modify the timing information. Since traffic can be quite different from weekdays to weekends, I filtered out the congestion estimated on weekends. To get the average congestion of a region, I took the mean of all the available estimates. Lastly, the relationship can be found by calculating the correlation matrix. On the other hand, to determine the best timing to avoid congestion, I first found the 4 regions that will be crossed. I then took the average of the historical congestion dataset to find the average congestion in a single day. I then ran a polynomial regression.

Despite carefully executing code line by line to avoid potential errors, my approach has many challenges. For example, besides the lack of detail from the region's dataset, coming up with an efficient data structure took a lot of thinking with trial and error. Also on the coding side, since the dataset was curated ahead of time, I was also unsure of how to convert it to the correct data type to work with. The solution to that was to dig deep into the packages to find useful functions. Lastly, I was not sure how to plot with geographical content. Luckily I found some tutorials on constructing a scattered plot on a map.

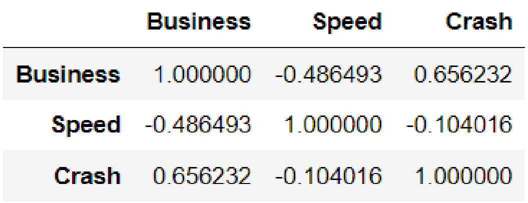
Outcome





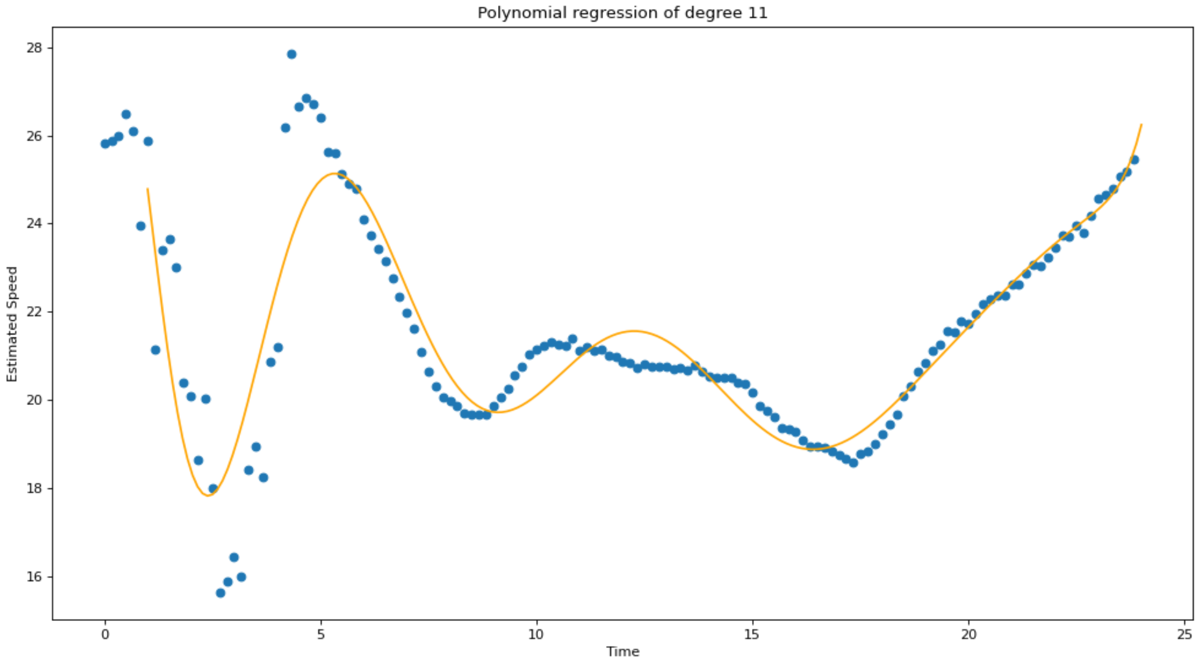
(41.793429°, -87.774822°)
Region #17: 71st to Pershing. Halem to Cicero. Speed:23.05. Crashes:28857. Businesses:34065.

The above plot was an interactive and comprehensive plot that shows all the curated information of a region. When hovering over a yellow dot, the plot shows the region ID, description, average estimated speed, number of businesses, and number of crashes.



	Business	Speed	Crash
Business	1.000000	-0.486493	0.656232
Speed	-0.486493	1.000000	-0.104016
Crash	0.656232	-0.104016	1.000000

The above table further shows the connection between the number of businesses, the number of crashes, and congestion. The number of businesses indeed has a moderate positive correlation with the number of car crashes and a negative correlation with estimated speed. In other words, with more businesses, there are expected to be more crashes, and the traffic will be more congested.



The above plot shows the average estimated speed throughout a day between regions 9, 10, 11, and 13. The traffic becomes the most congested around 8 am and around 5 pm, and it rebounds quickly after 8 pm. The points regress to the yellow polynomial, which further predicts that the difference in estimated speed can vary up to 30% throughout the day.

In conclusion, the outcome very much follows my hypothesis and my intuition. The map plot supported my personal experience as the estimated speed is indeed higher in region 12 than in surrounding regions, which interestingly opposes the general trend as there are also more businesses in region 12 as well. It also makes sense that the congestions are heavier during peak hours, which was solidified by the dataset.

Sources

1. https://data.cityofchicago.org/Transportation/Chicago-Traffic-Tracker-Congestion-Estimates-by-Re/t2qc-9pjd
2. https://data.cityofchicago.org/Transportation/Chicago-Traffic-Tracker-Historical-Congestion-Esti/kf7e-cur8
3. https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if
4. https://data.cityofchicago.org/Community-Economic-Development/Business-Licenses/r5kz-chrr
5. https://plotly.com/python/scattermapbox/
6. https://www.w3schools.com/python/python_ml_polynomial_regression.asp