A5: Extension Plan

Patrick Peng (ID: 2029888)
DATA 512, Autumn 2021

Motivation/Problem Statement

When COVID-19 prompted lockdowns and stay-at-home orders across the country last year, one of the most noticeable effects was the near-total evaporation of traffic congestion in major cities. While this period of low traffic volumes lasted, I began seeing a lot of anecdotal reports of drivers speeding, driving recklessly, and getting into accidents now that roadways were largely free of traffic. Counterintuitively, it seemed like the streets were becoming less safe even though fewer people were driving. Were our efforts to curb one public health issue (COVID-19) unintentionally exacerbating another (motor vehicle crashes)?

I hope to study the relationship between COVID-19 infection rates and the number of reported car accidents in Milwaukee County, Wisconsin. I think this is a good human-centered data science problem because motor vehicle crashes take a profound human toll. They are the leading cause of death among children and young adults and a leading cause of accidental injury death in almost all age groups. Furthermore, studies have shown that this toll falls most heavily on the poor. In a post-industrial city like Milwaukee where 25% of residents are in poverty, particularly concentrated in its segregated minority neighborhoods, this research question is highly relevant. If reactions to rising COVID-19 infections are influencing the number of accidents, that would be an interesting public health dilemma.

Research Question

Is there a relationship between COVID-19 infection rates and the number of reported car accidents in the City of Milwaukee?

The lower traffic volumes that occurred as a result of public health measures to combat COVID-19 could have affected traffic accident rates in two ways: Either accident rates should decrease due to fewer drivers on the road, or they might increase due to expanded opportunities for reckless driving. While the COVID-19 data used in A4 does not include public health measures other than mask mandates, it's possible that the infection rates alone may influence traffic accidents, as residents nervous by media coverage of rising infections may choose to forgo non-essential trips.

My hypothesis is that increases in the COVID-19 infection rate (or reported infections) over some time period correlate to a decrease in the number of motor vehicle accidents, either in the same time period or with some delay due to reporting and media coverage.

Data

In addition to the COVID-19 data used in the A4 common analysis, I am also using the Traffic Accident Data set from the City of Milwaukee's online data portal, which is available under the CC-BY-4.0 license. This dataset lists the incident ID number, date, and approximate location of every traffic accident report taken by city police dating back to 2006, although data is limited for the oldest years. Accidents are only included in this dataset if they caused injury/death or caused more than $200 of damage to government property or more than $1000 of damage to private property.

This dataset does have a few shortcomings. For one, it covers only the City of Milwaukee, which comprises slightly more than 60% of the County's population. The County government, Sheriff's office, and the suburban cities surrounding Milwaukee do not maintain online open data portals, so the City dataset is the only one freely available without filing a public records request. Second, it does not specify any details about each accident other than its approximate location. The data is not fully anonymized - someone could in theory use the incident ID to look up and purchase the detailed crash report and its personally identifiable information, although the fact that the reports cost money represents a moderate frictional barrier against misuse. As a result, I am only able to study trends in the number of reported accidents, rather than in injuries or deaths. Finally, without a corresponding data source for changes in vehicle miles traveled (VMT), I am forced to look only at the absolute number of accidents, rather than as a function of VMT. As a result, I will not be able to tell whether a rise in accidents is due to reckless behavior or simply more drivers on the road.

Using this dataset, I will be able to establish the trend in motor vehicle accidents over the course of the pandemic. I will likely need to aggregate this data into time periods of a week or more to make the trend observable, as day-to-day fluctuations are too noisy. I can then compare the COVID-19 infection rate over the same time periods to see if there is a correlation.

Unknowns and Dependencies

There are many factors in addition to COVID-19 infection rates that may influence the number of motor vehicle accidents in Milwaukee, such as seasonal trends and public health mandates other than masking. Because this analysis is examining *only* the correlation with COVID-19 infection rates, it will be difficult to separate out such effects, although comparing trends to pre-COVID years may offer some insights. Any correlation thus discovered may be entirely spurious.

Methodology

After downloading the dataset from the city's open data portal, I need to bucket the crash data into discrete time periods in order to smooth out the day-to-day fluctuations and make the trend observable. Preliminary exploratory data analysis suggests that time periods of one week may be suitable. I will need to do the same for the COVID-19 infection rate - instead of calculating it on a daily basis, I will do so on a weekly basis.

To account for time delays between rising/falling infection rates and (potential) changes in driver behavior, I may need to offset the time periods slightly. I will investigate multiple offset sizes to see if any discernible trends emerge.

To establish the degree of correlation between COVID-19 infection rate and accident counts, I will perform a linear regression. This is a good choice for this analysis because it is easily interpretable and lends itself easily to graphics, especially because I am only performing regression against one variable.

Timeline to Completion

Time period	**Milestones**
Week of 11/14	Download and clean dataset Exploratory data analysis
Week of 11/21	Prepare data for analysis Perform regression and check for significance Repeat with offsets
Week of 11/28	Prepare presentation deck
Week of 12/5	Complete presentation deck Prepare final report
Week of 12/12	Complete final report