Introduction & Business Problem

Traffic accidents have been, and will continue to be, a problem plaguing society for the foreseeable future. With the world population expected to keep growing, we would expect there to be an increase in cars on the road, which will not be balanced equally by an increase in roadways. To follow that, we expect the numbers of accidents to also increase. However, with enough data, it may be possible to identify factors that increase the likelihood of accidents, and measures can be taken to prevent that. In some cases, measures may be infrastructure upgrades (signage, lightning, street bumps, speed limits, etc.), or they may be campaigns to spread awareness and change human behavior. With self-driving cars starting to become a reality in the upcoming decades, action must be taken now to create safe and reliable roadways. Government transportation departments need actionable data-driven recommendations to improve roadway conditions and reduce accidents and fatalities. Accident prevention will result in many positives for communities, such as less traffic, fewer emissions, improved public transportation reliability, and most importantly, safety. Information and conclusions drawn from this data will help with future urban planning and design, as well as traffic control, and can even be of benefit to car insurance related issues.

Data

In order to solve the problem of traffic-related accidents, a publicly accessible data set from Seattle, Washington will be used to draw recommendations. The labeled dataset, accessible here, includes roughly 190,000 collisions from January 2004 to May 2020, courtesy of SPD and recorded by Traffic Records. Each accident is assigned a unique identifier, with as many attributes to describe the accident as could be collected. Accidents here are defined as including, but not limited to, personal vehicles, bicycles, and pedestrians. Other data attributes present include timestamps, location data, accident severity, injuries, collision description, weather conditions, and driver characterization, among others. There is a total of 37 potential data attributes which may be used for analysis and recommendations, although not all of them may be used. Other data will be brought in, such as Seattle road maps and geography, which will be cross referenced with location identifiers to view accident density across the city. Clustering accident information with location will be key to this project, as it will serve to guide recommendations on what action is needed to be taken in which neighborhoods. Regression, correlation, and other methods will be used in order to determine data attributes which may effect severity, and machine learning models will be used in order to predict the severity of an accident based on a combination of multiple data attributes.