Bike and Car Accident Prediction Data Wrangling

Data Sources

I gathered data from a variety of sources.

- Boston bike accidents 2009-2012 https://dataverse.harvard.edu/dataverse/BARI
- Accident/crash data: US accident data 2017 https://www.kaggle.com/sobhanmoosavi/us-accidents
- Mass DOT (Massachusetts Department of Transportation) road inventory: https://geo-massdot.opendata.arcgis.com/datasets/road-inventory-2018

Wrangling Goals

The goal of these data wrangling steps is to create a single files that contains:

- Roughly similarly sized regions where bike accidents have occurred, car accidents have occured, and where no accidents have occured
- Each region should have a bike accident score, a car accident score, and road attributes

The major challenges were as follows:

- The bike and car accidents list out individual accidents and need to be clustered geographically
- The Mass DOT road data contains a lot of unidentified IDs and codes that need to be translated into meaningful information
- All the data uses a mix of x/y coordinates in an unidentified coordinate system and latitude/longitude to identify locations

Cleaning Steps

See this notebook for more detail on all of the steps below.

Bike Accidents

After reading in the bike data to a pandas dataframe, I found that some of the bike accidents lacked coordinates. Since this project is based on road attributes at the location where accidents took place, I dropped rows without coordinates.

The area the bike accident database covers will determine the car accidents and road data areas as well, so I found the extent of the area by finding the minimum and maximum of the X and Y coordinate columns. Using pyproj, I converted these from the Mass DOT coordinate system into latitude and longitude.

Car Accidents

The database from kaggle was nearly 3 million rows and covered the entirety of the United States. The only tasks here were to limit the dataset to the area covered by the bike accident database, and then convert it to the Mass DOT coordinate system which is more useful for accident clustering and linking road attributes to locations.

Mass DOT Road Data

The Mass DOT road data contains a lot of unidentified IDs and codes that need to be translated into meaningful information, so I found a dictionary that translates most of the codes.

I used geopandas to limit the Mass DOT data to the area containing bike accidents.

Geographic Wrangling Steps

Since this project will examine what attributes predict high volumes of bike and car accidents, I wanted to cluster data around where accidents occured, and create a database of regions with associated car and bike accident scores as well as regions without accidents.

Creating Bike Accident Clusters

More detail in this notebook, Geographical Manipulations Part I.

I looped through the accident dataframe and used shapely to create circles of radius 50m centered on each bike accident. If another bike accident occurred within that radius, I created a new shape that was a 50m radius around both bike accidents. The final shapes were circles and blobs, as follows:

Adding Car Accident to Clusters

More detail can be found in this notebook: Geographic Manipulations Part II.

I used a similar process to add car accidents to this list of clusters. If a car accident was within an existing cluster, I recalculated the shape and added the car accident id to a list of car accident ids associated with the. If the car accident was not in an existing cluster, I created a new cluster.

Joining Road Data

More detail can be found in this notebook: Geographic Manipulations Part III.

At the end of this, I was curious if my accident clusters covered all Boston roads, or if I would also need to add areas without accidents, so I created a multipolygon that encompassed all of my clusters.

Here is the image created, showing accident clusters occurring on lines that look like roads, but with gaps between them that need coverage by creating regions without accidents.

Geopandas provides ways of joining geo-spatial databases based on geometry, so I used their overlay method to intersect the accident cluster dataframe with the road dataframe, which associated road data with all of the accident clusters.

Then I did the same thing, but looking at the difference between the road dataframe and the accident dataframe to create a dataframe of roads without accidents.

Finally, I appended them together to create a database of all regions, and made the bike and car accident scores count the number of accidents in each region, which are zero for the around 50% of road areas that contained neither a car nor bike accident.

Final Cleaning Steps

More detail on these steps can be found here in Final Data Wrangling

This data set still had many unnecessary columns and columns with lots of null values. I dropped columns with the following qualities:

- Too many null values
- Containing geographic data important for creating the set but not important for analysis
- Other descriptive data that wouldn't be important for analysis, e.g. Street names, Maintenance stations

Many columns still had null values for 10 to 20% of the values.

- For categorical data encoded as integers, I assigned null values to be -1
- For categorical data encoded as strings, I assigned null values to be 'NA'

- For measurement data with a fairly small number of options, like road width or number of lanes, I also assigned null values to be -1, but this will need to be monitored. I may try treating these measurements as categorical values
- For traffic volume, which can vary wildly, I created a separate dataframe that only contains accident scores and traffic volume, with all rows with null traffic volume dropped.

Final Thoughts and Next Steps

From this data, I should be able to find correlations, if they exist, between bike accident scores, car accident scores, and all of the other data associated with roads. The dependent variables will be bike accident score and car accident score. The independent variables will be road attributes like size, speed limit, surface type, truck exclusions, and traffic volume.

Because these regions vary in size, I may want to normalize the bike and car accident scores by the area covered.

I have left a large number of columns in the final database for now, since I don't know which of them will predict bike and car accidents. If none of the road attributes predict these, or just if I'm curious, I may add in pavement condition information from this file: https://geo-massdot.opendata.arcgis.com/datasets/pavement-condition