Bike and Car Accident Prediction Milestone Report

By Linnea Hartsuyker

Problem Statement

Many drivers hate sharing the road with bicyclists, but perhaps they could become allies, if road improvements for bikes made the road safer for cars as well and vice versa. In this project, I am examining what road attributes predict a greater density of car accidents and bike accidents, in hopes of finding road qualities that could be changed to produce a lower accident density for both cars and bikes.

If no commonalities can be found, I will still determine what road attributes contribute to either type of accident, independently, information that road maintenance and design could take into account.

This project focuses on bike and car accidents in the greater Boston area.

Datasets

The data sets for this project come from:

- A database of Boston bike accidents from 2009 to 2012, constructed from police reports during this time: https://dataverse.harvard.edu/dataverse/BARI
- A database of car accident/crashes for the entire US in 2017
- The Massachusetts Department of Transportation (Mass DOT) road inventory https://geo-massdot.opendata.arcgis.com/datasets/road-inventory-2018
- The Mass DOT road inventory has a data dictionary

From these datasets, I needed to construct a single file 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 density score, a car accident density score, and road attributes

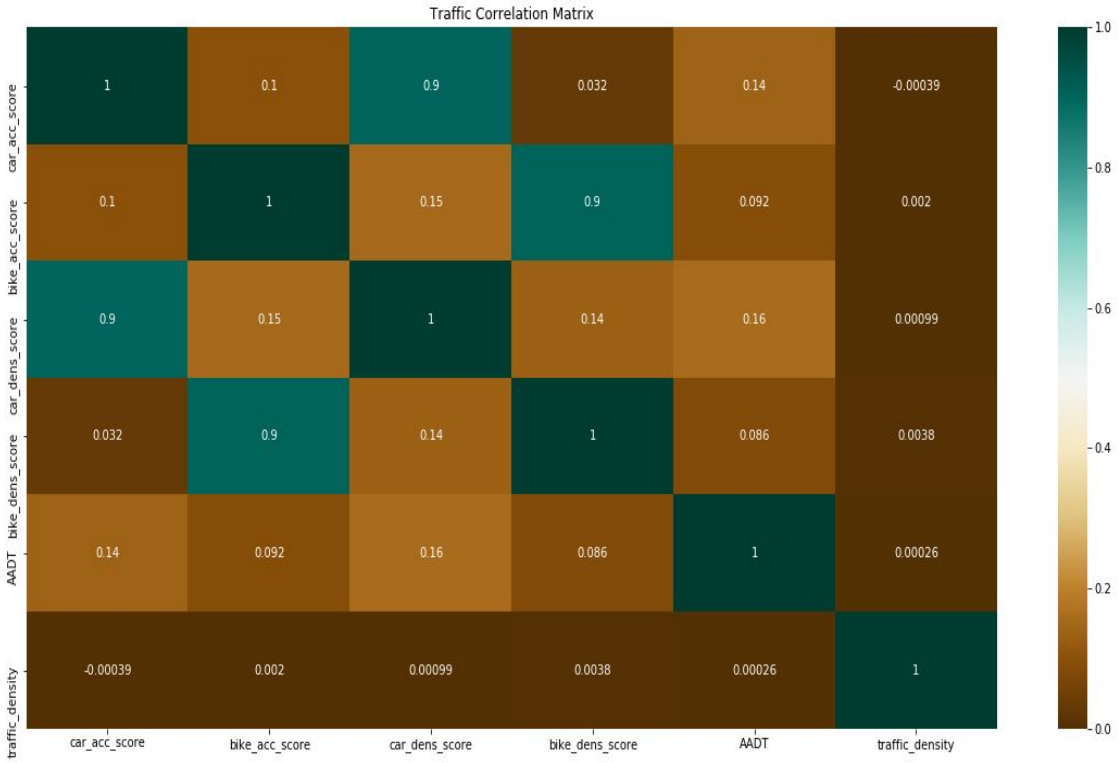
For more detail on creating this file, see the Bike and Car Accident Prediction Data Wrangling report.

Initial Findings

For code and more visualizations, see Exploratory Data Analysis and Bike and Car Accident Statistics notebooks.

Traffic

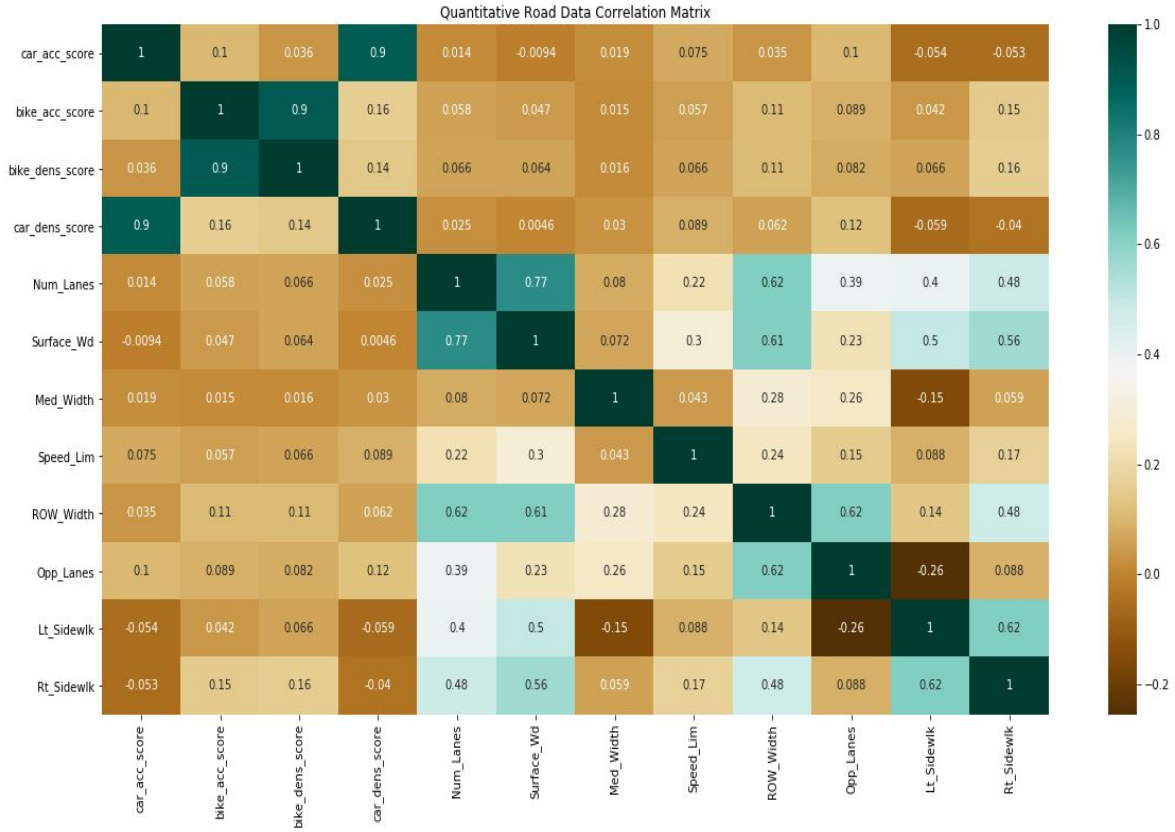
My first question was whether bike and car accidents were correlated on their own. And I was also interested in whether traffic volume was correlated with accidents. I created a data set that contained only accident and traffic data and plotted a correlation heatmap.



A correlation coefficient near 1 would mean a strong correlation, and none of the correlation coefficients are close to one, except for properties derived from one another, like accident score and accident density values.

Other Quantitative Road Attributes

Next I wanted to see if any of the numeric values had obvious correlations, with another heatmap. These values included: number of lanes, surface width, median width, speed limit, right of way width, opposite lanes, left sidewalk width, right sidewalk width, and right of way width.



This also showed no strong, direct correlations between accident scores and any quantitative values.

Digging into Qualitative Road Attributes

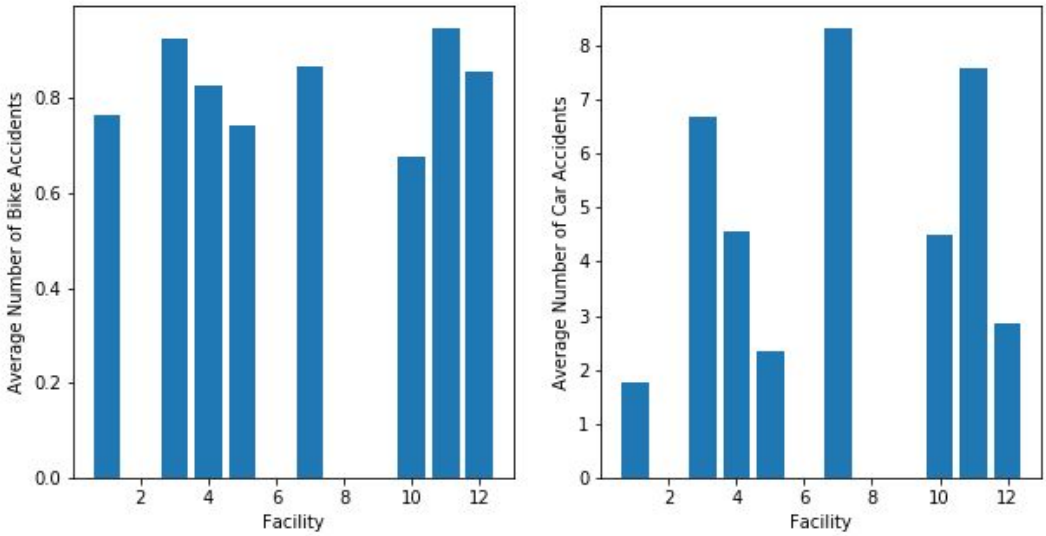
Bar charts of qualitative data against accident scores showed that several attributes appeared to be associated with high accident densities.

For reference, the average number of car accidents per area was 2.11 and the average number of bike accidents 0.75. Looking at bar charts, I compared with these values to determine “larger” average number of accidents.

For all hypothesis tests I examined the accident density scores, which are accidents per square meter, rather than average number of accidents, since this seemed like a more consistent metric.

Facility Types

Roads are designated with facility types, and the graphs showed a much larger average number of car accidents for facility types Tunnel (3), Simple ramp (7), and Simple ramp - tunnel (11), and a moderately larger average for some others. It showed a much larger average number of bike accidents for facility types Doubledeck (4), Rotary (5), Collector - Distributor (10), and Bicycle (12).

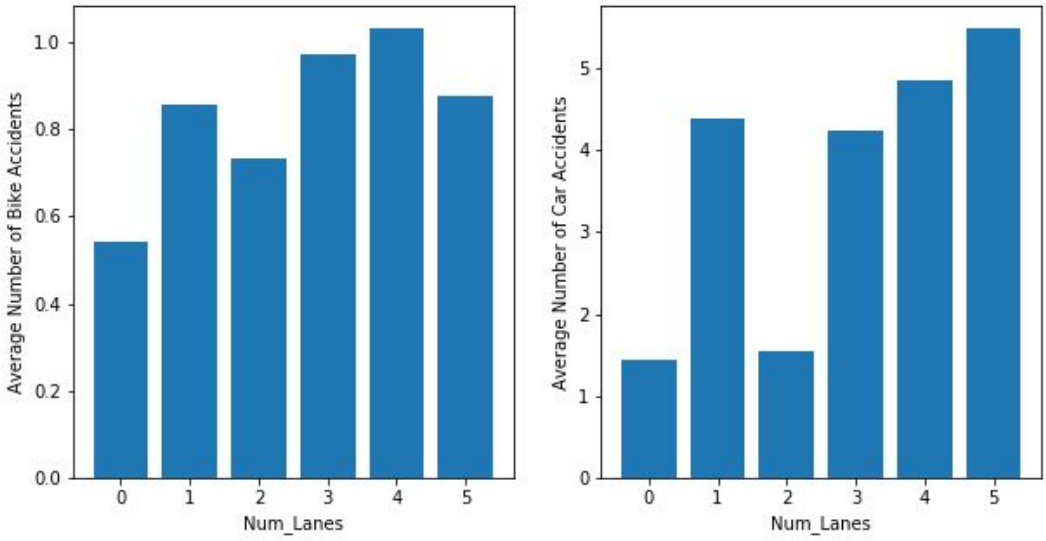


Hypothesis: Facility types 3, 4, 5, 7, 10, 11, and 12 have higher mean car accident density scores than the average across all facility types

I performed a p-test and found statistical significance to *reject the null hypothesis* that there is no difference in the mean car accident density for all of these facility types. I could only *reject the null hypothesis* bike accidents occurring in facility types 11 and 12.

Number of Lanes

Number of lanes doesn't have a high correlation coefficient, but roads with 1, 3, 4, and 5 lanes appear to have more car accidents than with 2 lanes.



Since number of lanes and speed limit are not totally independent, I decided to look at mean accident densities for roads with a 30mph speed limit.

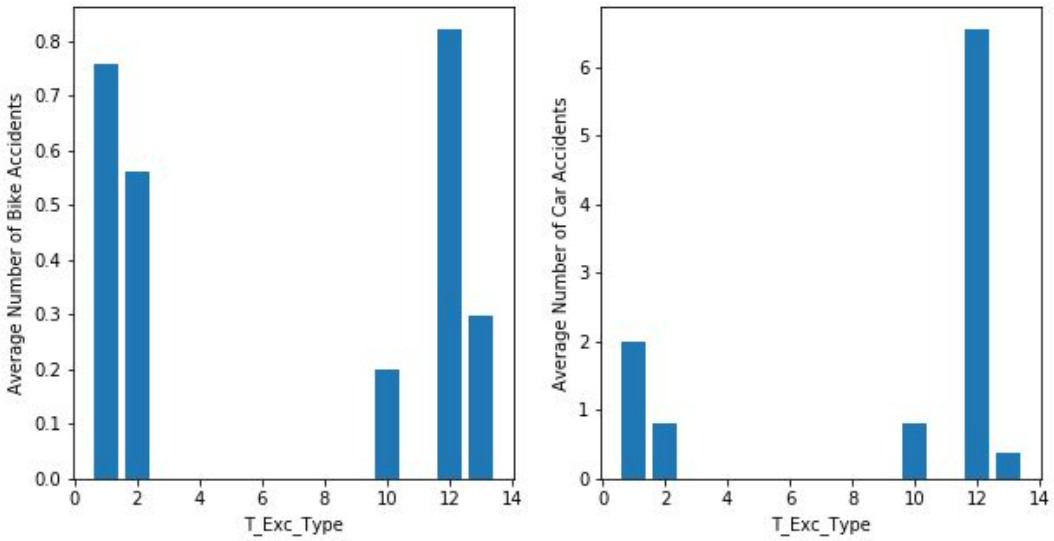
Hypothesis: roads with 1, 3, 4 and 5 lanes have a higher mean car accident density than the average car accident density

I found a p-value for rejecting the null hypothesis that there was the same car accident density for roads with 1 and 3 lanes, i.e. they had a statistically significant greater average car accident density.

Using a similar test I found that roads with 4 lanes had a higher mean bike accident density than average.

Truck Access

Higher levels of truck access does seem to have an effect on accidents, but this still may be because of what data is collected. Routes that allow hazardous trucks (Exclusion Type 12) have much higher than average number of car accidents.



The vast majority of roads have no truck exclusions, but around 5000 have some kind of exclusion.

Hypothesis: roads with any kind of truck exclusion have a lower mean accident density than roads with no truck limits

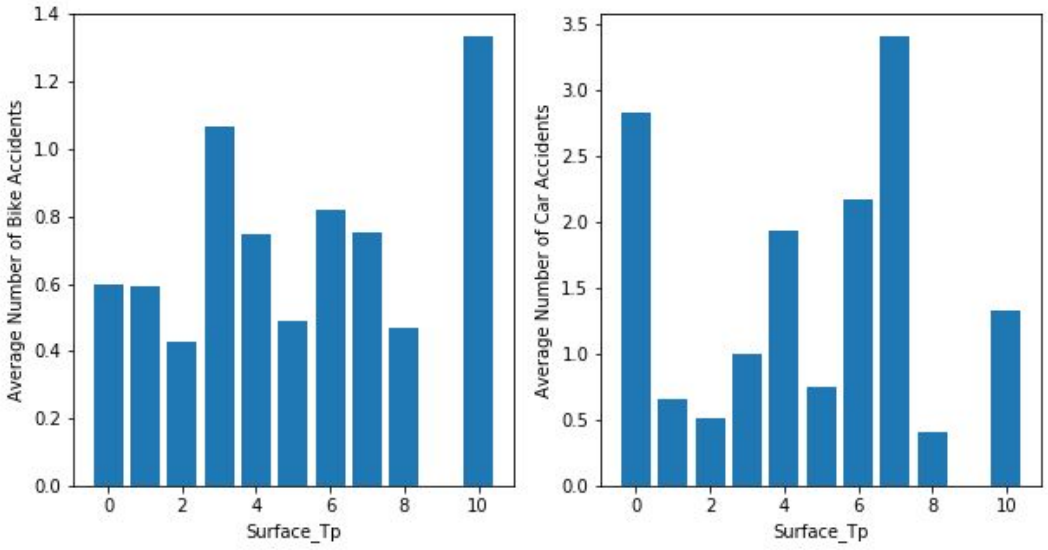
I found extremely small p-values for rejecting the null hypothesis, so truck exclusion does have a strong effect on accident density for both bikes and cars.

Hypothesis: allowing hazardous trucks leads to a higher mean accident density than average

I found a very small p-value for rejecting the null hypothesis with respect to cars, indicating that the presence of hazardous material trucks leads to a higher number of car accidents, but not enough evidence to say the same for bikes.

Surface Type

The data for surface types is not very good, but we do have some.



These graphs can be a bit misleading because only types 2, 4, 5, 6, and 7 have much data at all.

Surface Type	**Number of Data Points**
-1 = Unknown	2051
0 = Unknown	8743
1 = unimproved, graded earth or soil surface road	49
2 = gravel or stone road	960
3 = brick road	73
4 = block road	392
5 = surface-treated road	7895
6 = bituminous concrete road	53584

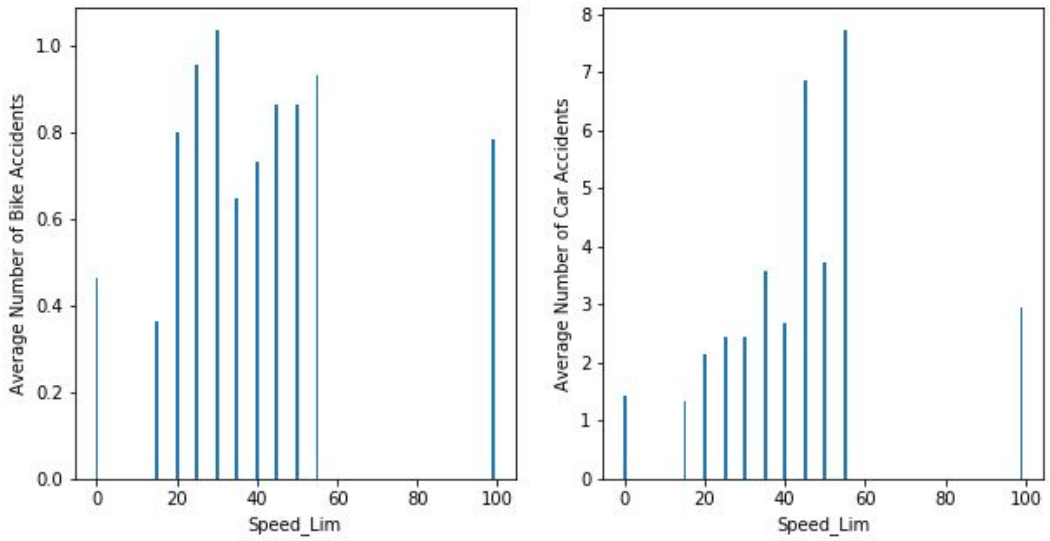
7 = portland cement concrete road	337
8 = Composite road, flexible over rigid	34
9 = Composite road, rigid over flexible	1
10 = stone dust	3

Hypothesis: Road types 2, 4, 5, and 7 have higher mean accident densities than type 6

The p-values support this hypothesis for higher bike and car accident density for types 2, 5, and 7 but not 4.

A Deeper Dive Into Speed Limit

Speed limit doesn't have a high correlation coefficient, but looking at the bar chart shows the number of bike accidents may be growing from 15 to 30, and then again from 35 to 55. Car accidents mostly increase from 20 to 55.



Hypothesis: the mean car accident density increases from 20 to 30 mph, as does mean bike accident density. (I'm starting with 20 mph because there weren't many data at 15 mph)

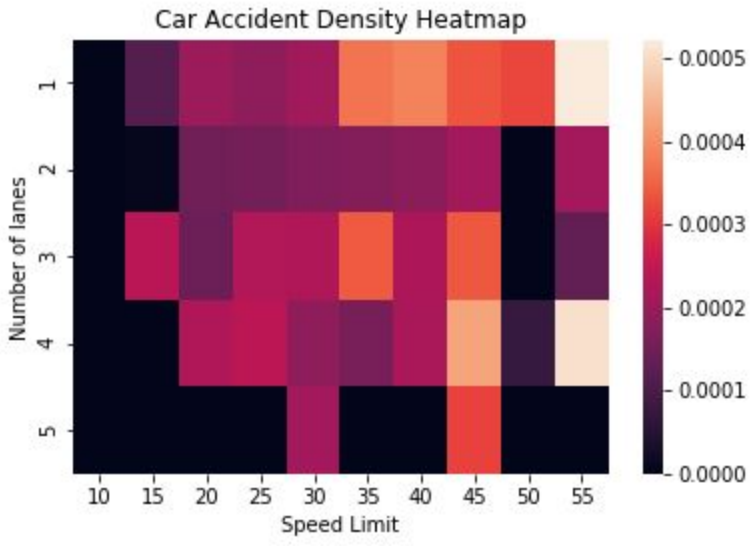
I found the p-value only showed statistical significance for bike accidents. The p-value for rejecting the null hypothesis that there is no difference between bike accidents in:

- 20 mph zones and 25 mph zones is 0.0009764585459075452
- 25 mph zones and 30 mph zones is 0.0006733426655266123
- 20 mph zones and 30 mph zones is 9.245160606207875e-11

Conclusions and Next Steps

I have identified some road attributes that are correlated with higher accident density for both bikes and cars.

The next step is to use machine learning to find out if the data can be used to predict accident density, especially where there might be multiple variable interactions. For instance, during exploratory data analysis I found evidence that increasing the speed limit on a one lane road had a correlation with increased car accidents.



There may be many other interactions between variables that could be interesting but could only be found with more advanced machine learning.