Detecting leading causes of fatalities in NY vehicle accidents

Zachary Labkovski, Jonathan Mager, Nikunj Patel

I. Problem Statement

For a city as densely populated as New York, safety both for motorists and pedestrians is at the forefront. However, over the course of the first three months of 2022, traffic-related deaths rose 44 percent compared to the same time period in 2021. This trend mirrors nationwide motor vehicle statistics, with motor vehicle fatalities increasing 12 percent from 2020 to 2021, the highest percentage increase year over year since data was first collected in 1975 (Axios). Our goal is to identify any trends or features that are especially important in what makes car crashes in New York fatal as well as the probability of a crash being fatal given certain conditions. By doing so, we hope to identify ways in which the city can look to reduce fatal accidents, including potentially identifying times of day, days of the year, and locations for targeted enforcement, and changing traffic patterns for different vehicle types.

II. Data Source

We pulled our data, a csv file containing motor vehicle collision data for New York City, from the City of New York's public data from their Open Data initiative.

Each row of the data table represents a single collision incident. The file contains the information from each crash contained in police reports–that is, form MV104-AN, required to be completed by the responding officer in cases of collisions resulting in injury or death, or damages totaling at least $1,000. It is important to note that, because of this criteria, the dataset only includes accidents that would be considered major.

The information for each collision includes address information (borough, ZIP code, primary street, cross streets, etc.), along with exact location (longitude, latitude), time of crash to the minute, injuries and fatalities by person type (motorist/pedestrian/cyclist), types of vehicles involved (sedan/pickup/SUV/etc.), and contributing factors.

One thing to consider when we do our analysis is that the causes of crashes can be subjective. Things like distraction, unsafe speed, and driver inexperience could be what the police considered the main cause of the crash. This could also have been a "he said/she said" scenario where the actual reason for the crash could be more complicated. While we cannot control subjectivity, we should be reasonably careful with any conclusions we draw from this feature.

III. Methodology

A. Geographic Analysis of Accidents

Our initial aim was to find areas that proved to have higher rates of accidents than others. Using this information could be vital for further city planning, or understanding where emergency responders need to be located to best reach regions that have tend to see more accidents.

We chose to use the data that contained latitude and longitude coordinates and eliminated the data points that did not include values in these columns. There is an assumption to be made on that the data eliminated does not particularly skew the overall geographical projection of where these accidents occurred. With those points eliminated, there were still 1.73 million accidents over the course of a decade to be examined.

We created a new dataframe using these accidents and restricted the latitude and longitude further depending on the bounds of a borough or neighborhood that needed further investigation. There was also an added column that combined the total number of people injured or killed which would be passed in as a third argument as a weighted measurement for the severity of each accident. The street names would all have to be treated as categorical variables making each accident extremely unique. There is also a lack of consistency in the way the street names were recorded depending on whether the accident occurred in between intersections, at an intersection, or along a major highway. Thus, solely relying on the coordinates will produce a better result.

The initial attempt is to create a K-means cluster diagram to identify where the most dangerous intersections may be. If this does not deliver easily interpretable results, then a density map could help visualize where these areas might be and we could use the severity factor as an added feature to construct it. Constructing the heat map with a geographical map of NYC projected onto it will also further prove that the coordinates are enough to construct a density map and there is no need for addresses with street names for the accidents.

B. Density Estimation of Time Series Data

After answering the question of where to focus our attention for catastrophic vehicle accidents, we then sought to identify when accidents occurred the most, and whether there are trends in the data to help NYC municipal decision makers adjust their resources to better respond.

When doing a basic analysis of the datasource, we found that most datapoints missing coordinate data (latitude and longitude) came from incidents reported on major bridges and highways, where traffic patterns (e.g. speed limits) and individual access (bicyclists, pedestrians) differ greatly from the rest of the data. As a result, we decided to omit those datapoints for our time series analysis. Additionally, as part of the data preparation, date and time data needed to be converted to better suit our analysis. From our original data, we created a new data frame containing the crash month, day of month, day of week, day of year, and the time of crash presented as the minute of the day in which the crash occurred (ranging from 1 to 1440).

In addition to removing datapoints missing coordinate information, we also chose to filter data for different levels of analysis; that is, we looked at all reported accidents, only at those involving deaths, and at those involving death and/or injury. By doing so, we would be able to better visualize differences in trends that would impact resource needs. At each level, we assembled histograms and Gaussian Kernel Density Estimation plots for day of week, day of month, day of year, and time of

day. Once we identified in which analyses trends were most prevalent, we could then combine scopes in a two-dimensional plot to isolate specific combinations of time-based factors that may not be easily identifiable in one-dimensional plots.

C. Logistic Regression to Identify Contributing Factors of Fatality

To determine the contributing factors for an accident to be fatal, we decided a logistic regression model would be best for two reasons:

1. We could look at the factors with the greatest positive and negative magnitudes to see which had the greatest effect on the overall probability
2. If first responders could quickly identify the main contributing factors to the accident and run the model, then the probability of fatality could be a good indicator of how many units need to be sent.

Starting from the initial dataset, this process needed a few specific steps of data cleaning in order for it to be ready for the model to be built. First, the only factors used will be Contributing Factor Vehicle 1, Contributing Factor Vehicle 2, Vehicle Type Code 1, Vehicle Type Code 2, and Borough. We decided to use these factors since first responders would only have time to transmit back a few key details, and not enough time to build a full report.

Next, we decided to impute all missing data fields with “Unspecified” since this distinction was already being used in the police reports as an indicator that the value of the field was unknown. Then, we needed a predicted response variable. This was obtained by taking the Number of Persons Killed variable, since this includes all people killed (pedestrians, motorists, cyclists, etc) and made a new field where it was 1 if Number of Persons Killed was greater than or equal to 1 and 0 otherwise.

The most important step followed with making binary dummy variables for each of the independent variable fields because they are all categorical. This meant creating an indicator field for all 5 boroughs, every single model of car, etc. This added up to 3,189 fields.

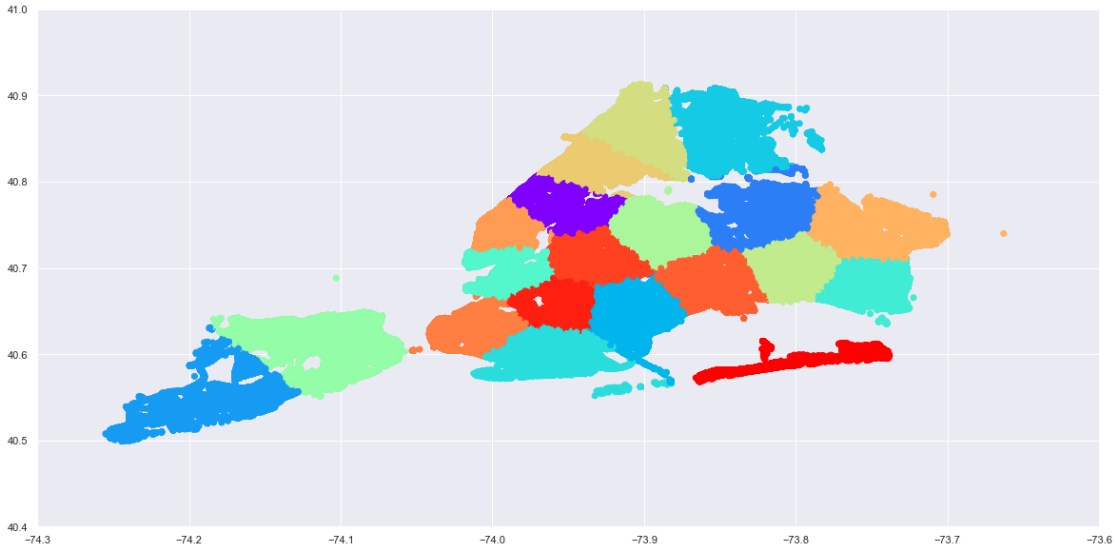
Finally, there was a train-test split where 67% was used for training and 33% for testing. However, because our dataset was so large with data going back 10+ years, we could only use part of the training set so that it would run on a computer. Since the training set was randomized, this was not an issue.

After all of the preprocessing was done, we used scikit-learn’s logistic regression package. At first, no parameters were used, and then a second logistic regression model with a parameter was added.

IV. Evaluation and Final Results

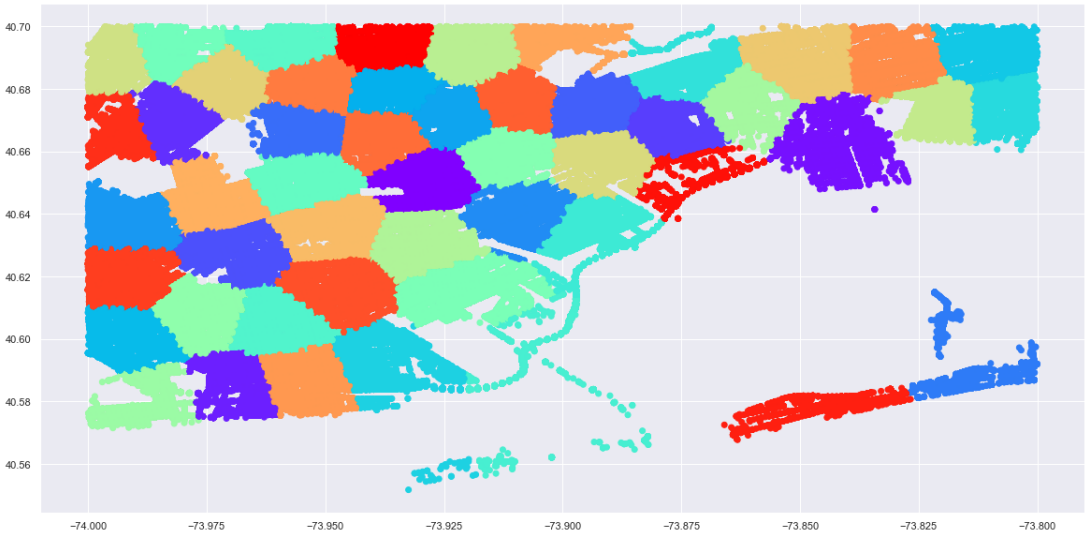
A. Takeaways from Geographic Analysis

Our initial attempt included using K-means clustering to dissect intersections with high frequencies of accidents. Due to the sheer number of data points that contained latitude and longitude coordinates (~1.73 million), using 25 clusters only separated the points by the neighborhoods which they occurred in. The map below shows every data point color matched to their respective cluster for the entire NYC metroplex.

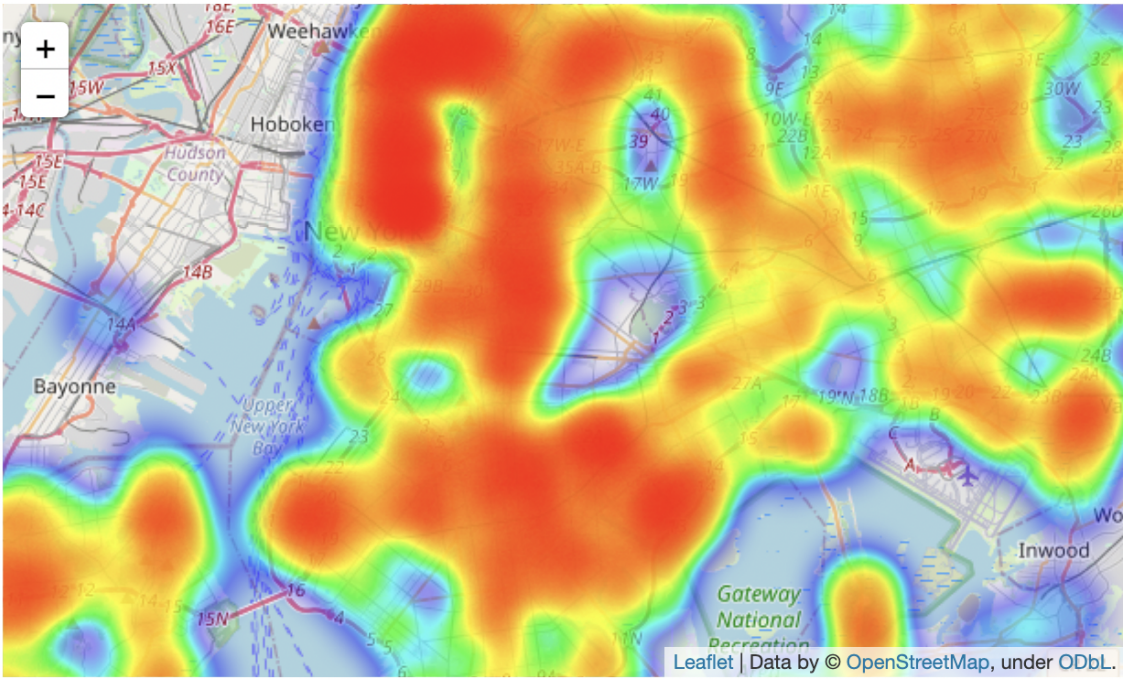


We can distinguish the geographical regions with the very left being New Jersey. Going eastward we see Brooklyn at the middle bottom, followed by Queens, and Manhattan island is sitting on top. However, this does not aid us in finding vital intersections as the centroids just follow the population density.

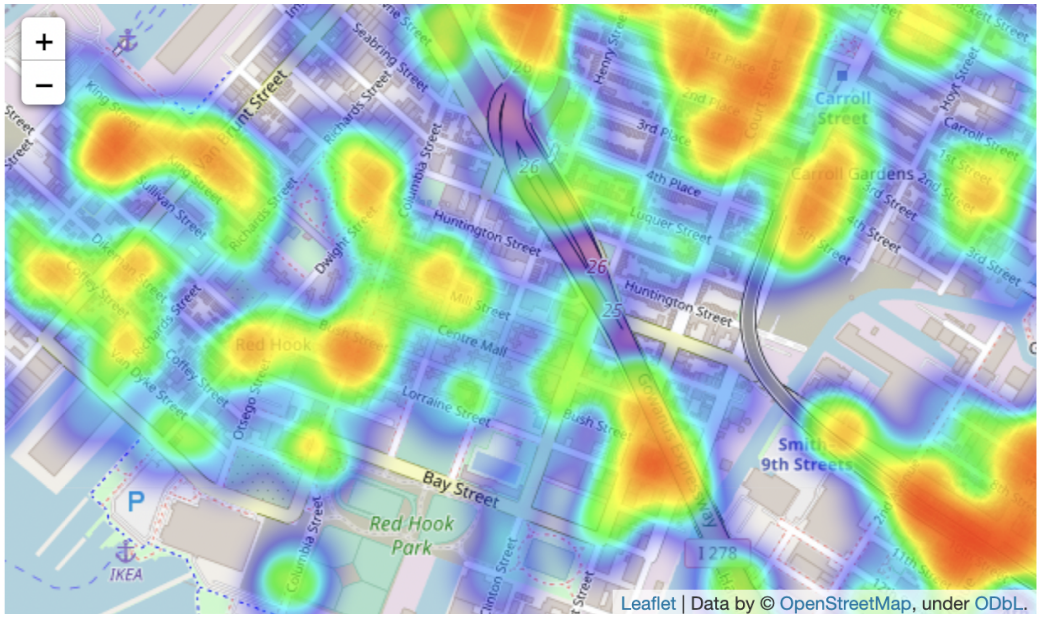
If we attempt to limit the coordinates to the geographical bounds of a particular neighborhood and also increase the number of clusters (to 50), the outputted map still results in a slate with many points all attributed to the highest concentrated centers of the population. Furthermore, most centroids were along main streets with higher traffic (i.e. Broadway, Pennsylvania Ave, etc) which does not aid in showing any particular correlation with dangerous areas. The map below shows every data point and their respective cluster for just the Brooklyn borough.



By utilizing the gmaps plugin in Python, we then constructed heatmaps which did a better job at showing high frequency areas. The map below shows a general heatmap based on the frequency of all accidents in the NYC area.



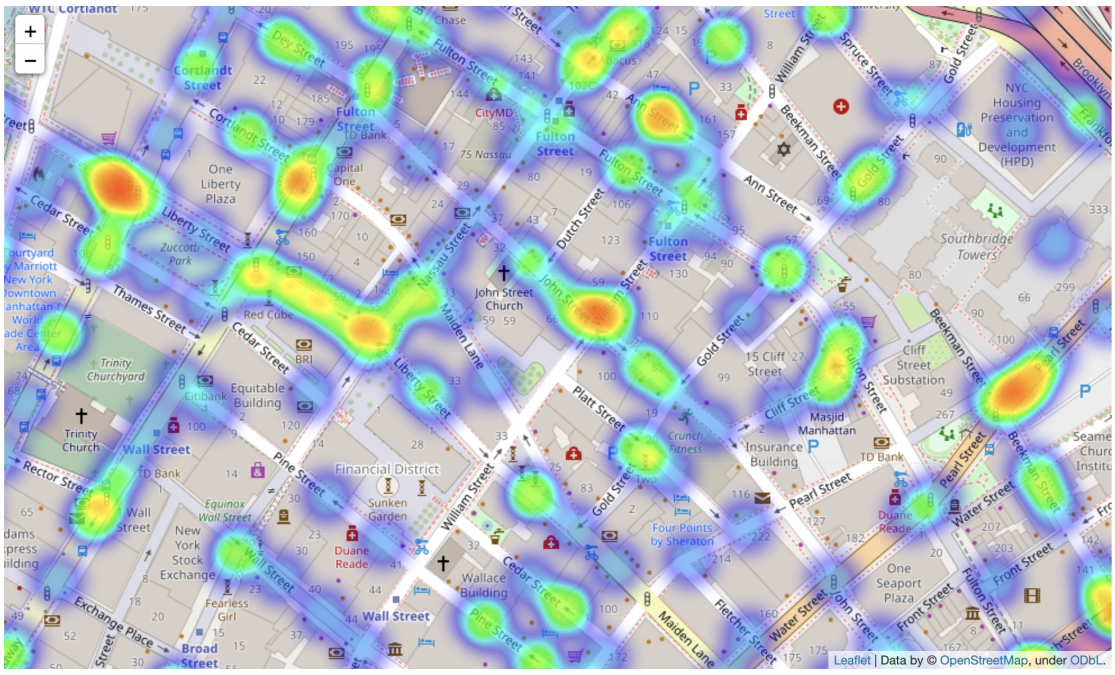
The map below shows the restricted zone of the Carroll Gardens neighborhood in Brooklyn to give a more accurate depiction of where improvements need to be made. The other benefit of the gmaps plugin is it that the building features and shapes easily identify whether these districts are primarily residential, industrial, or commercial areas. We can use this information to better contextualize the traffic that occurs in or around these zones.



We took this one step further by scaling the number of people injured and killed into a weight that the heatmap takes in as a third argument. This helped better identify key intersections where urban planning may need to be rethought. These could be areas where distinct bike and bus lanes need to be constructed or stop signs need to be replaced with traffic lights. The map below shows the same area of Carroll Gardens with improved interpretability due to the weight of severity added in. It becomes clear that there the on and off ramps before the Hamilton Avenue Bridge need to be rethought and the streets surrounding the Red Hook Houses can be further looked into.

If we zoom in further to the expressway access roads, we are able to notice that the Bush Street and Court Street feeders are the major cause.

In another example, the map below shows us a constricted zone of the SoHo neighborhood in Manhattan and highlights three major areas of improvement at Liberty Street near One Liberty Plaza, John Street, and Pearl Street.



We can further investigate into the area surrounding One Liberty Plaza and notice that there is a looming danger on Southbound Broadway as well as pose the question of how Liberty Street can further be improved. The area is pictured below.

The map displays that bike lanes and traffic lights already thus it became a question of whether the region is accessible to those on foot. This is an area with a high level of pedestrian activity as shown with Zucotti Park, the 9/11 Memorial just west and Wall Street just south of the parameters. This could be marked as an area of utter importance to be improved to prevent pedestrian injuries with the high volume of vehicles that must traverse the area.

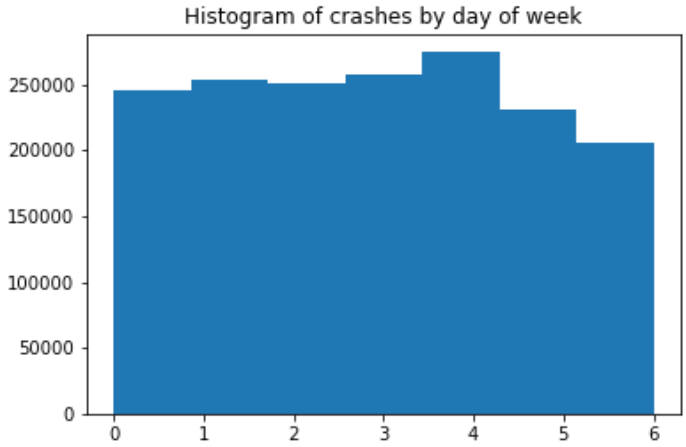
Ultimately, the density map proved to be more useful as a way of modeling areas that required further investigation. It allows us to use additional clues such as the current state of urban infrastructure, zone type, and severity of accidents to visualize detrimental traffic features and intersections.

B. Dates and Times to Focus On

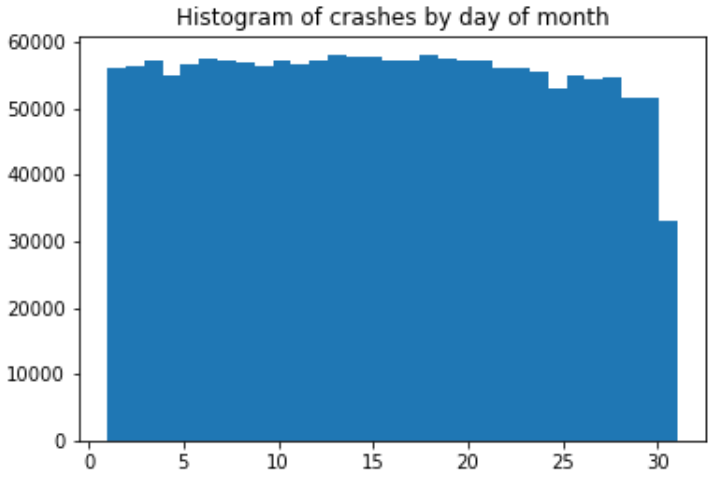
For all time series analysis, we felt it was best for smaller increments of time (e.g. day of the week, day of the month) to visualize through histograms, whereas for larger increments of time (e.g. day of the year, minute of the day) we chose to focus on Gaussian KDE plots to minimize information loss from binning and improve ease of trend analysis with smoothing. Additionally, once identifying trends in one-dimensional time series data, for viewing data in two dimensions, we used two dimensional histograms, since it is more easily interpretable than 2 dimensional KDE plots.

Prevalence of All Accidents

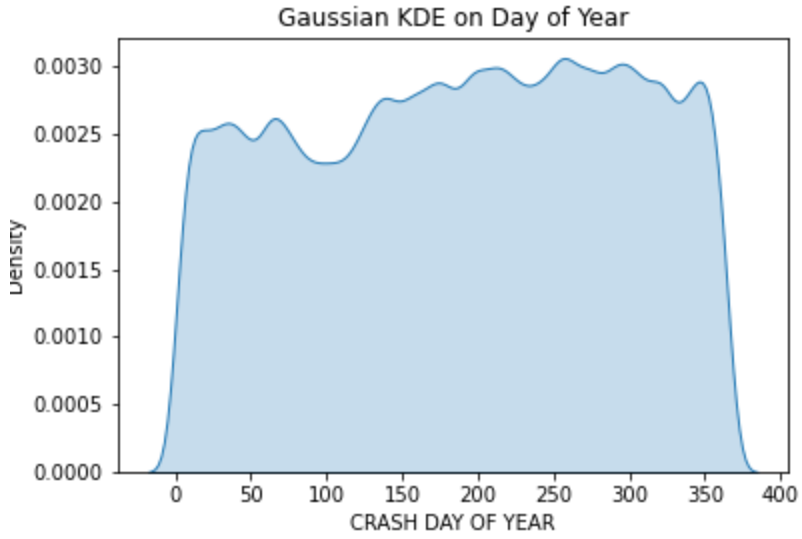
When considering all accidents in the database, we produced the following visualizations:



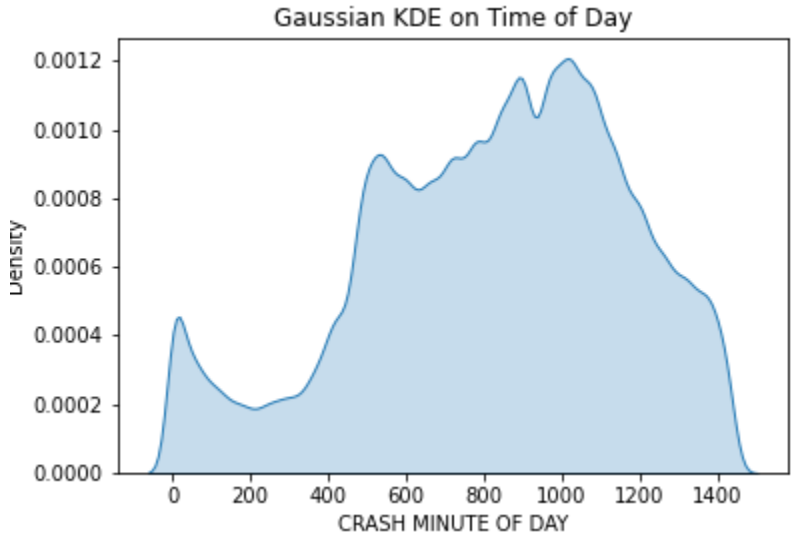
In this visualization, the numbers represent days of the week, ranging from 0 (Monday) to 6 (Sunday). We can see that total number of accidents tends to be higher on weekdays compared to weekends, likely due to commuter traffic. Additionally, there is definitely a material spike in accidents on Fridays, which for a city like New York with a high volume of commuters, tourists, and day visitors from the surrounding region, is logically sound given that Friday would likely be the day in which those three groups of motorists converge the most.



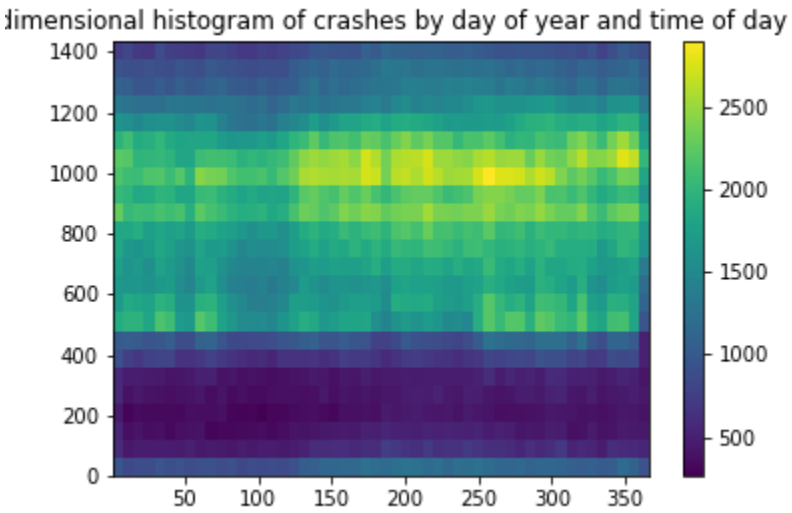
In analyzing the histogram of accidents by day of the month, the only significant differences seem to occur at the end of the month, which can be explained simply by the variation in number of days for each month of the year. Otherwise, there does not appear to be any material difference in accident frequency by day of the month.



Considering accident frequency by day of the year can help in both identifying trends throughout the course of the year and identifying key dates where there may be a spike in vehicle collisions compared to surrounding dates. In the KDE plot above, we can see that, on the whole, accident frequency tends to increase in early summer and stay at that higher level until the end of the year, while staying lower in the winter. One plausible reason is the prevalence of "snowbirds" in New York–that is, (mostly) retired individuals who live in the northern half of the United States most of the year but live in Florida for the winter. Because accident rates and accident fatality rates are higher for those 65 and over than for the population on the whole, this may be a major contributing factor to the variation in cyclical trends. However, we were unable to find data to conduct analysis on any correlation with "snowbirds", so that possibility is merely a theory at this stage. Additionally, we see accident rates initially plunge slightly around early March and stay at a lower level until around the start of summer, which is most easily explained by variations in weather.



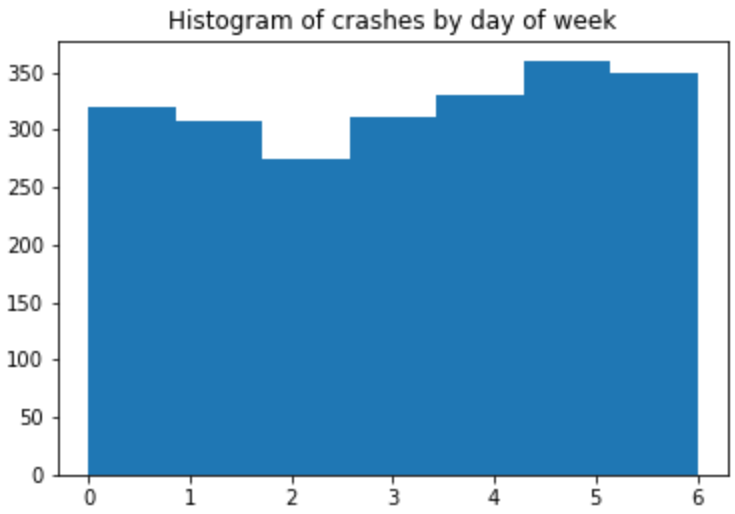
Throughout the course of the day, we can see accident rates drop off initially after 1 am (after the bar/club crowd typically goes home), before seeing a gradual increase between 7 and 9 am, which is typically referred to as the morning "rush hour". After that time, accident rates stay fairly level until around 3 pm, when accidents spike and stay high until about 6 pm, where they gradually decrease until the end of the night. This information can be quite useful in preparing for emergency response throughout the course of the day, both in determining staffing needs and medical/vehicle resource needs.



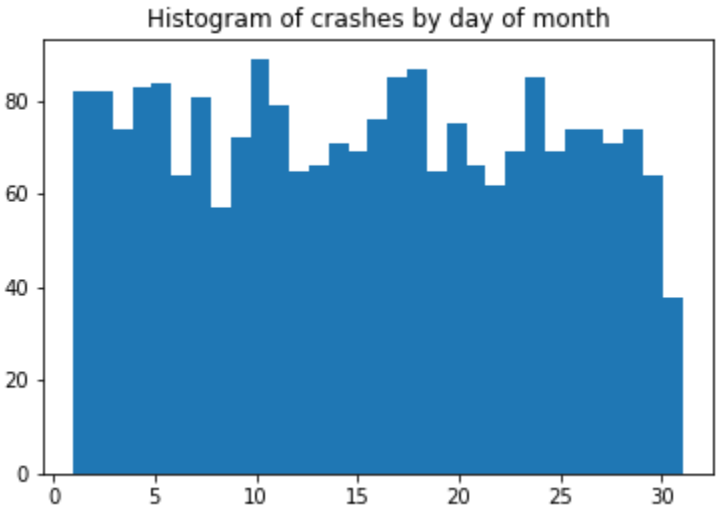
Since trends appeared mostly in analyzing time throughout the course of the day and in cyclical patterns in days of the year, we decided to combine the two into a two-dimensional histogram in hopes of identifying any major combinations of date and time that see particularly high accident levels. However, there do not appear to be any significant differences in trends compared to looking at each time period individually.

Prevalence of Fatal Accidents

After looking at time series data for all accidents, we decided to separate accidents involving fatalities to see whether there are any differences in trends that may impact logistics for New York City decision-makers. It is important to note, however, that compared to over 1.7 million data points looked at overall, there were only 2,252 fatal accidents in our dataset. As a result, the sample size may not be large enough to be truly representative of real trends in accident fatality.

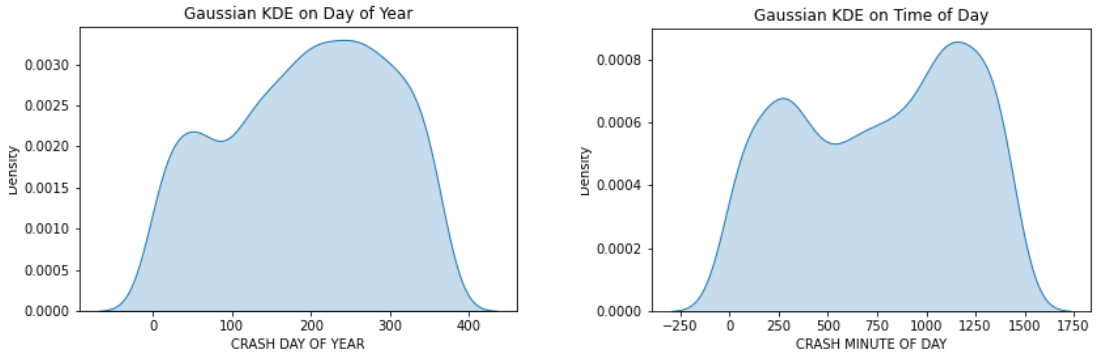


For day of week frequency, weekends tend to see more fatal accidents within the scope of the distribution compared to all accidents. One possible explanation is increased use of drugs and alcohol on weekends, which–as we will discuss later–are a leading cause of accident fatality. However, as aforementioned, it is important to note that fatal accidents comprise a much smaller sample size than all accidents.

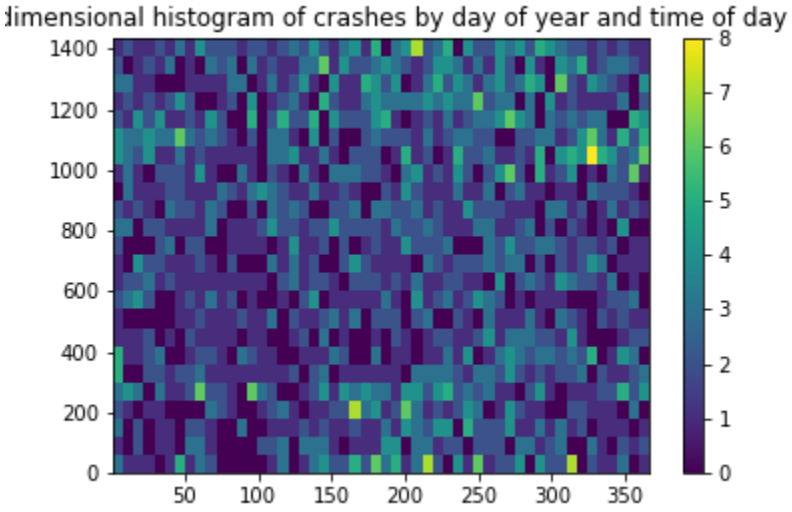


There seems to be much more variation in accident frequency by day of the month for fatal accidents compared to accidents overall. One thing to note however, in addition to the aforementioned small sample size, is in looking at the data, spikes and drops in fatal accidents seem to be fairly cyclical, with peaks and valleys occurring approximately every seven days. In constructing a frequency matrix with day of the week and day of the month as the axes, we actually found that there were major discrepancies throughout the course of our data (ranging from July 1, 2012 to November 5, 2022) in terms of weekends falling more frequently on certain dates in the month. As a result, although there

seems to be a trend in fatal accident frequency on certain days of the month, it is a continuation of the previously seen cyclical trend in fatal accidents occurring more often on weekends.



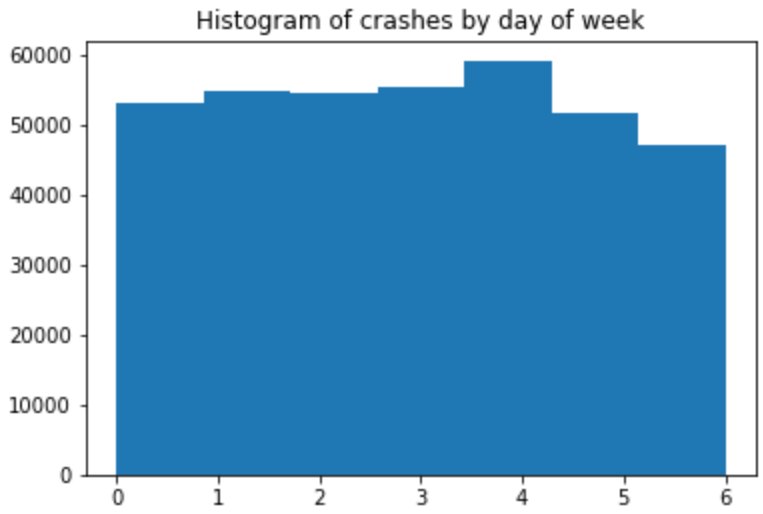
The distribution of both day of year and time of day frequency do not differ much for fatal accidents compared with all accidents; however, they do appear somewhat different because of differences in smoothing due, once again, to the smaller sample size.



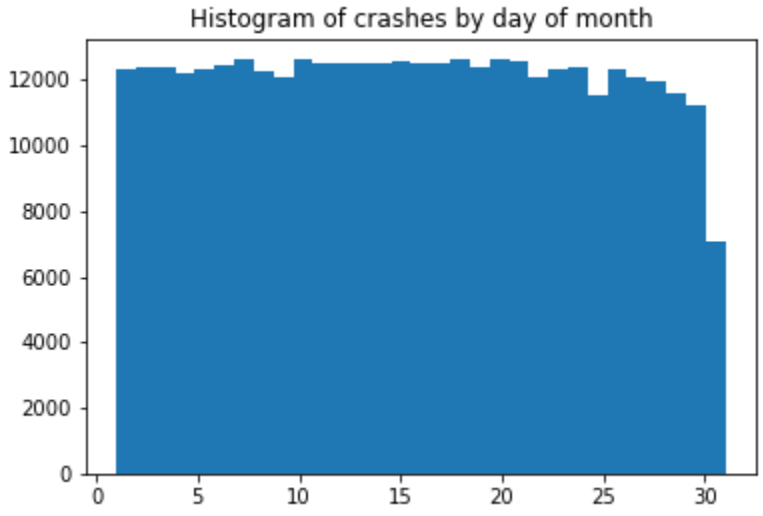
Unlike for the other plots, where the sample size is merely a disclosure but not a disqualification, for the two-dimensional histogram with day of year and time of day data, I am willing to fully dismiss the results because of the small sample size. As shown above, the scale for the histogram ranges from 0 to 8 data points, meaning there is no single combination of day of the year and minute of the day where there were more than eight fatal accidents in our data source.

Prevalence of Harm-Inducing Accidents

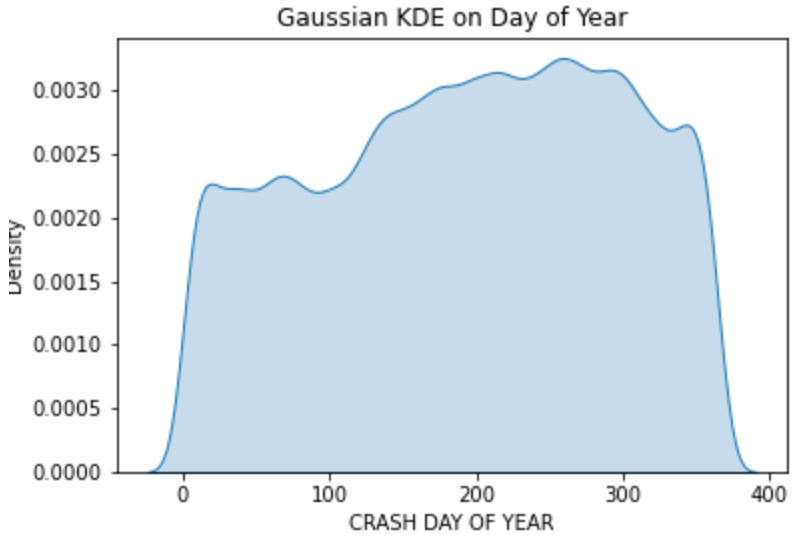
Because of the small sample size concern previously mentioned for data points involving fatalities, we decided to look separately at data points involving fatalities and/or injuries, which provided a much larger sample size of over 375,000. The results are shown below:



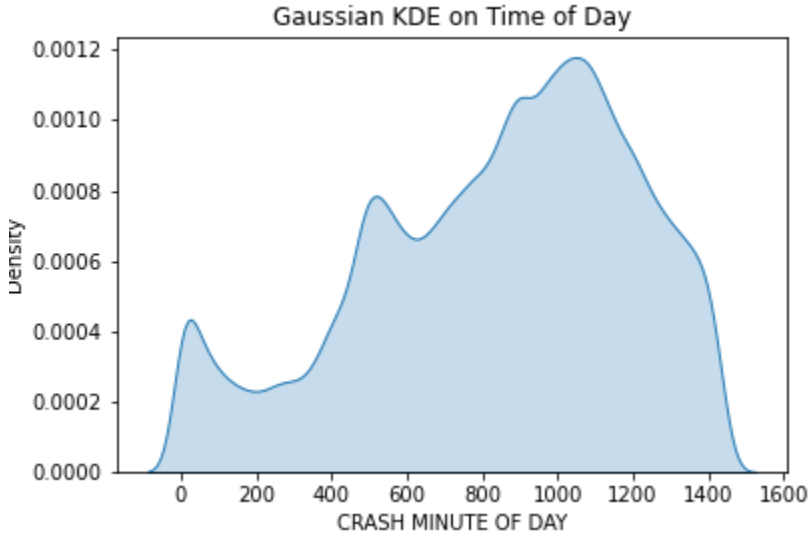
Similar to the trends seen in all data points, for data involving any sort of harm, weekdays tend to produce more accidents than weekends, and accident prevalence is at its highest on Fridays.



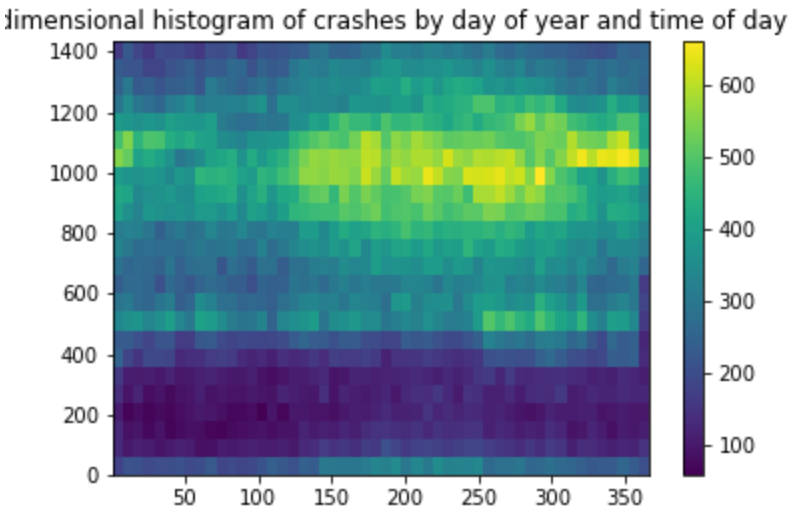
As with all data points, there are no apparent trends in day of the month, aside from lower totals for the 31st of the month, which does not exist for all months.



One key difference compared to all data points for day of the year is a greater difference between winter months and the rest of the year. The increase from winter months to the start of the summer in accident frequency is much more stark for accidents involving injury or death. If the aforementioned theory about "snowbirds" is true, this would make sense, since the increase in accident fatality rate for retired age drivers compared with the rest of the population is greater than the increase in general vehicle accident rate.



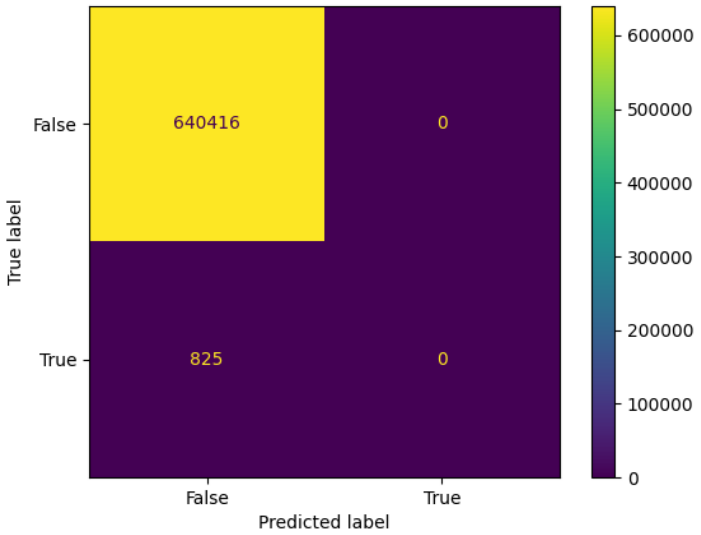
The Kernel Density Estimation plot for fatality- and injury-involved accidents by time of day compared with all accidents is nearly identical. However, similar to the plot for fatal accidents, there are differences in smoothing due to the sample size.



The two-dimensional histogram for day of year and time of day for harm-inducing accidents, as with that for all accidents, does not indicate any trends not seen in looking at either time increment individually.

C. Leading Contributing Factors on Accident Fatality

We started with a logistic regression model with no parameters and fed in the data that was described in the methodology. This resulted in 99% accuracy on the test set. However, on further inspection, it only predicted 0's or "non-fatalities" since less than 1% of accidents in this dataset were fatal as seen below.



Since this was not useful from a prediction standpoint, we made another model, but there were still some useful insights.

The top 10 most contributing factors to fatalities were:

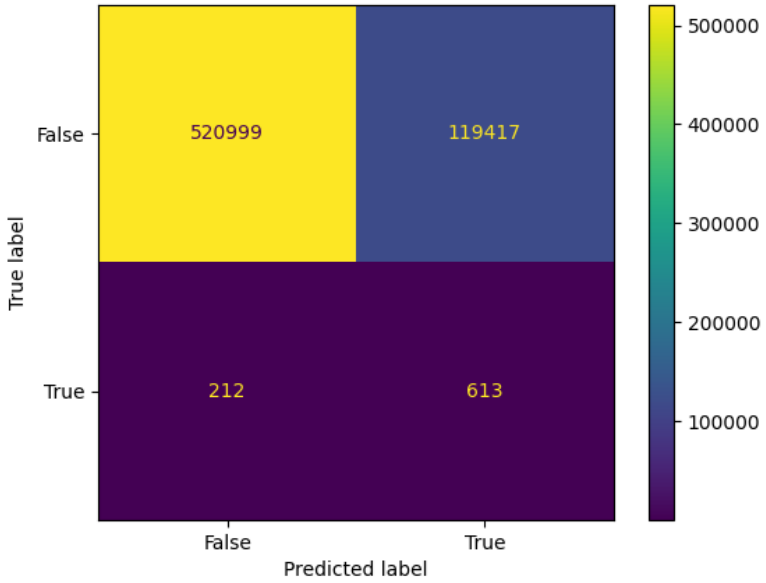
1. Motorcycle
2. Unsafe Speed
3. Illegal Drugs
4. Dump Truck
5. Illness
6. Motorbike
7. Lost Consciousness
8. Pedestrian/Bicycle/Other Pedestrian Error/Confusion*
9. E-Bike
10. Traffic Control Disregarded

*This was a contributing factor from the police report that seems to imply a non-motor vehicle error.

The top 10 least contributing factors to fatalities (i.e. the 'safest' accidents):

1. Sedan
2. Passing Too Closely
3. Bike
4. Following Too Closely
5. Unsafe Lane Changing
6. Passing or Lane Usage Improper
7. Outside Car Distraction
8. Reaction to Uninvolved Vehicle
9. Driver Inexperience
10. Taxi

While these leading contributing factors were useful, we were looking for a model that has better prediction capabilities. We used the same scikit-learn logistic regression model, but with the parameter class_weight = 'balanced'. This parameter accounts for the fact that the prediction classes are not even and that there are many more non-fatalities than fatalities. The resulting accuracy of this model was 81.3% on the test set.



Although this model predicted a lot more accidents to be fatal that were not, it does a much better job at trying to predict than the last model.

The top 10 most contributing factors to fatalities were:

11. Obstruction/Debris
12. USPS
13. Snow Plow
14. Illness
15. Pavement Defective
16. E-Bike
17. Drugs (illegal)
18. Motorcycle
19. Failure to Keep Right
20. Unsafe Speed

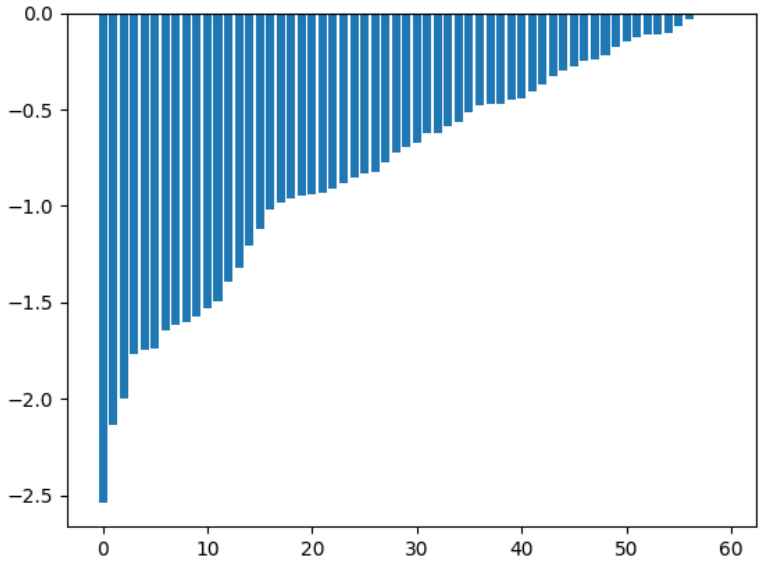
The top 10 least contributing factors to fatalities (i.e. the 'safest' accidents):

21. Passing Too Closely
22. Outside Car Distraction
23. SMALL COM VEH(4 TIRES) *
24. Pick-up Truck
25. Brakes Defective
26. Failure to Yield Right-of-Way
27. Delivery Vehicle
28. Glare
29. PK*
30. Lost Consciousness

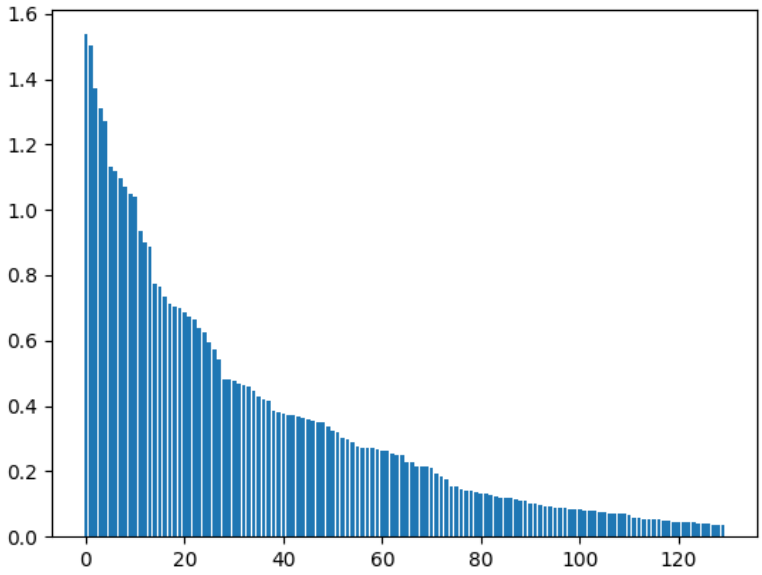
* These are types of vehicles

One thing that was apparent about these lists of fatal and non-fatal factors is they were much more specific than the set of factors for the previous model. It seemed to weigh factors that did not occur much at all very high. This is due to some factors like USPS only being involved in a few accidents where a few also happened to be fatal so the percentage of fatal accidents was high.

For these reasons, we only analyzed the accidents listed in the first models leading factors. The following graph shows the magnitude of the 60 most leading factors for non-fatalities.

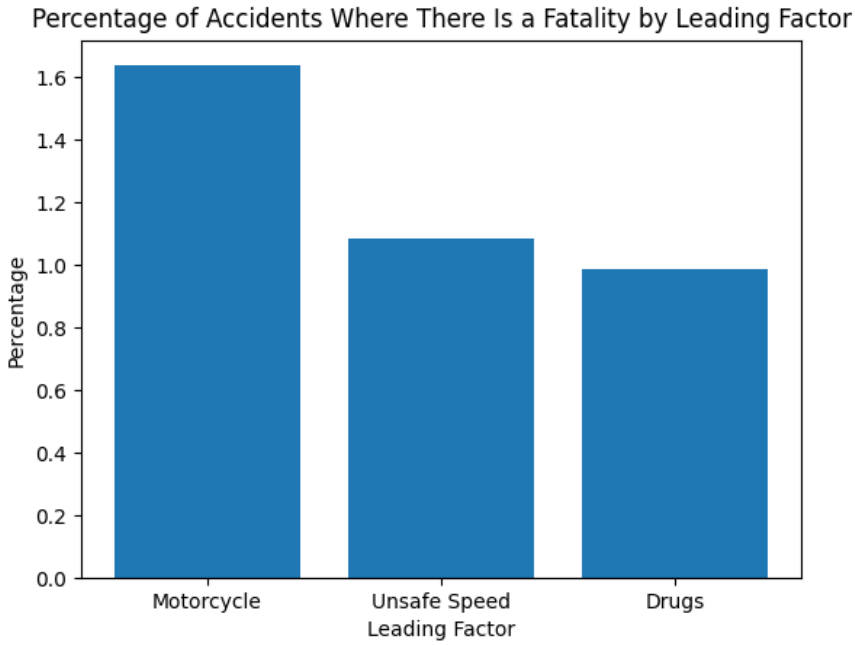


The following graph is the magnitude of the 120 leading factors for fatal crashes.



As shown in the previous 2 graphs, there are only 60 factors that contribute to the model negatively and 120 factors that contribute to the model positively. There were over 3000 factors that went into the model, so most of them had a negligible effect on the outcome of the model.

We looked at the top 3 leading factors for fatality and compared the percentage they were listed in fatal accidents vs. non-fatal accidents.



The graph shows percentages that are very low which is to be expected since fatalities are rare. However it is important to remember that it is the combination of factors that the model uses to predict.

Individual Contributions:

Geographic Analysis of Accidents – Nikunj Patel
Density Estimation of Time Series Data – Zachary Labkovski
Logistic Regression to Identify Leading Factors of Fatality – Jonathan Mager