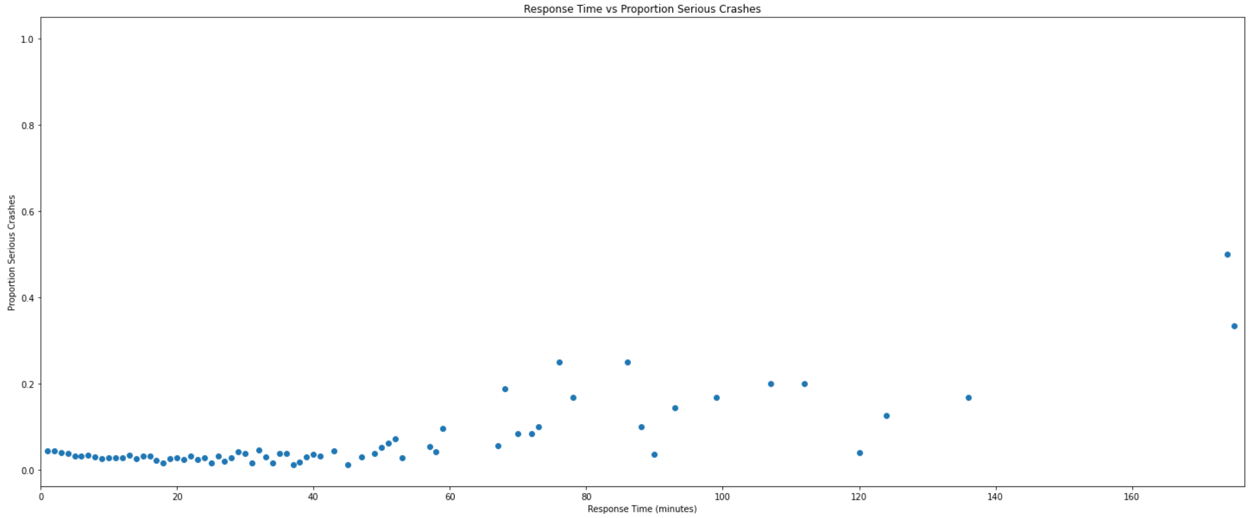
Recommendations for Police Presence to Decrease Traffic Accidents

Team 11: Alex Acra, Will Denton, Bradley Hu, Shannon Xiao

Non-Technical Executive Summary:

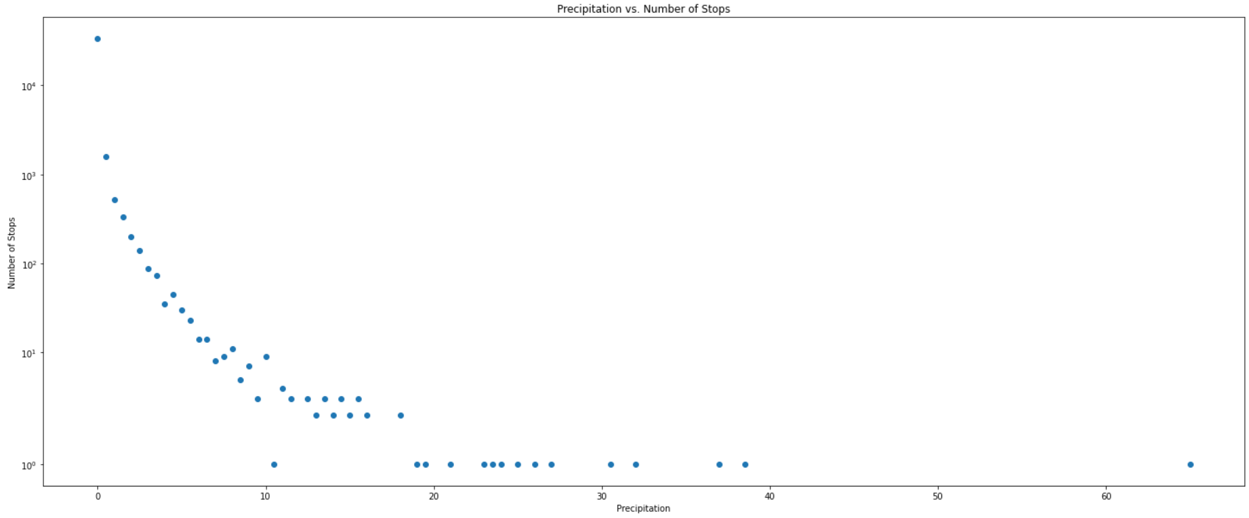
For this project, our team set out to find different ways to help prevent traffic accidents and mitigate the severity of the resulting injuries in Philadelphia. To do so, we explored factors such as traffic stop outcomes, vehicle crash information, weather data, and first responders' response time in order to find underlying patterns influencing critical traffic accidents. Our objective was to investigate how we can improve the police presence around Philadelphia to be more effective in preventing the occurrence and severity of traffic accidents.

An underlying assumption we made in examining this is that increasing police presence, thereby reducing response times, is correlated to the severity of traffic accidents. The first analysis we conducted was to examine the relationship between response time and the proportion of serious traffic accidents – fatal and serious injuries.

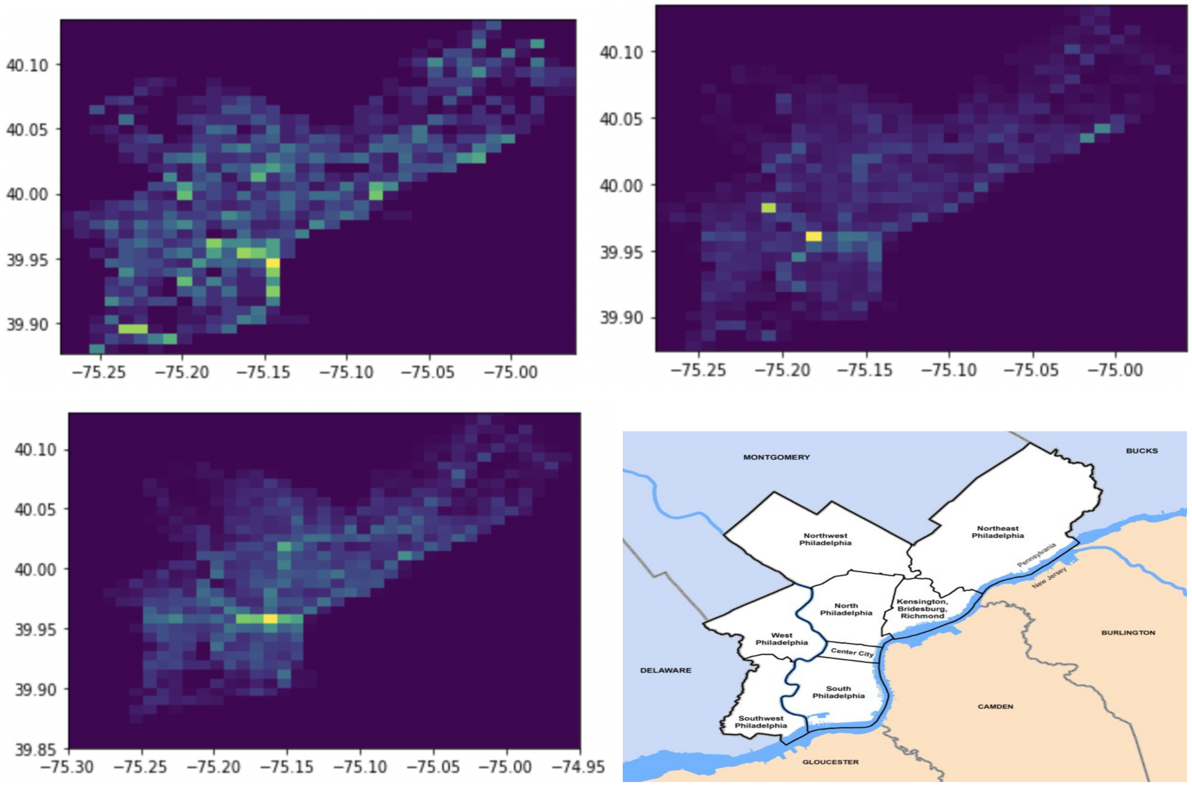


We observe that police response time has a correlation of 0.70 with the proportion of serious traffic accidents, so it makes sense to proceed with our analysis. In examining traffic accident prevention, we then focused on location and weather variables.

We first analyzed the relationship between weather and traffic accidents. In order to include police activity in our analysis, we used traffic stops as an intermediate variable and observed the relationship between traffic stops and weather and the relationship between traffic stops and crash activity. When plotted against each other, we found that traffic stops decreased as precipitation increased, as shown below. This was a surprising result as people tend to drive poorly during poor weather conditions, and we realized that was most likely due to the tendency for people to drive less when there is more precipitation.



Our experiments comparing crashes directly with weather conditions yielded better results. We sought to map the likelihood of a crash in different locations depending on the weather. To this end, we found that in certain weather conditions, areas of Philadelphia were more likely to have accidents. Our key findings can in part be summarized by the following heat maps:

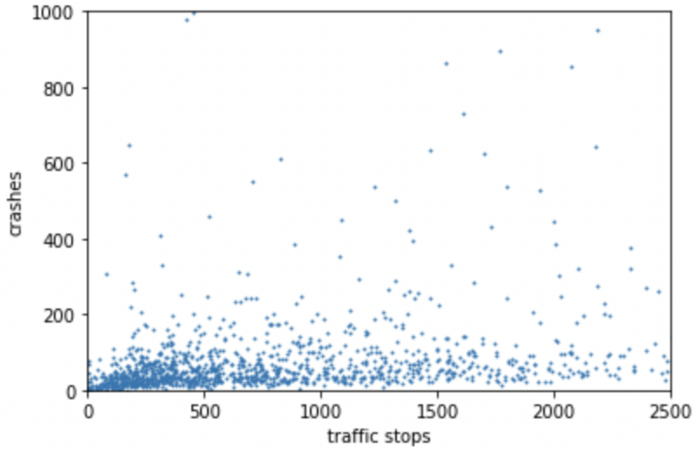


Heat maps (lighter colors represent higher likelihood and darker colors represent lower likelihood) of accidental collision probability under different weather conditions of snow (top left), rain (top right), clear weather (bottom left) and a map of Philadelphia for reference (bottom right).

From these heat maps, we can conclude that under the relevant weather conditions, police and emergency response resources should be allocated accordingly so as to dissuade reckless driving and provide faster response times in the case of accidents.

Another theory we tested was the correlation between traffic stops and the number of crashes in each census-designated area, for the purpose of understanding whether increased police vigilance corresponded to decreased crashes, or rather if increased crimes caught by police could potentially explain increased crashes due to unsafe driving. In fact, we found that there was no significant relationship between the two, with a low correlation coefficient and no apparent visual trend.

```
r = 0.24467628004480982
```



Finally, given these conclusions about where police should be depending on weather conditions, we set out to model how police can prioritize reaching crashes to save the most lives and prevent the most serious injuries. Ideally, this model would be used by first responders to assess the likelihood of a positive outcome and triaging in the case of multiple crashes. We did so using a logistic regression model, which predicts whether a crash will result in serious injury or death based on the location of the crash, the type of collision that occurred, and the initial distance from the police team that responded to it. The model reached 97% accuracy in its predictive ability when training on initially skewed data, and 62.5% accuracy when the data was rebalanced, with a final 72% accuracy rate of identifying deadly/serious crashes.

Technical Exposition:

To examine the relationship between response time and the proportion of serious traffic accidents, we loaded the *crash_info_general.csv* into a pandas dataframe, before directly querying it using the *pandassql* library. We then cleaned the data by removing null values and unifying the “Fatal” and “Suspected Serious Injury” values from the original severity level column under a new ‘flag’ column. Since the officer dispatch and arrival times were valuable to this analysis, we needed to convert the float values into datetime-manipulatable values. Since the response_time data was heavily concentrated toward lower response_times (e.g.. more traffic accidents have lower response times than higher response times), we normalized the distribution by calculating the proportion of serious traffic accidents out of all traffic accidents, rather than examining the cumulation of serious traffic accidents alone. We then removed zero values from the proportion of serious traffic accidents, since we are not concerned with non-serious traffic accidents in this analysis outside of normalization purposes. From here, we found the correlation between response times and proportion of serious traffic accidents to be 0.70.

To plot the weather and traffic stop data, we joined the t*raffic_stops_philadelphia* and *hourly_weather_ philadelphia* datasets together based on the date and time of each. To tune the data, we pruned the samples without any precipitation data and also rounded the time of each traffic stop sample down to the nearest hour and also discretized the precipitation level to the nearest .5 mm. Since there was a large variance in the number of traffic stops for the different precipitation levels, we log-scaled the y-values to observe a clearer correlation. The r-value between the precipitation level in mm and the number of traffic stops for that level was -0.69. To account for the imbalance of total number of days dry-drizzling-pouring days, we also plotted the daily average number of traffic stops for each precipitation level and found that no precipitation still yielded the greatest number of traffic stops, but there was a minimal correlation between the two variables (r-value of -0.22).

In order to explore the relationship between weather and car accidents, we began by simply inquiring how many accidents occurred under different weather conditions. Given a large data imbalance for different types of weather, we decided that the best way to proceed would be by distinctly analyzing different types of weather.

We first explored if there were any insights that could be gained by clustering crashes by location, crash type, and other data, but found little practical insight by clustering these data. We then sought to create heat maps for the distinct types of weather. The only underlying assumption that we made in justification of our conclusions is that frequency of collisions is proportional to the probability of a collision. Given the law of large numbers and the vast size of these datasets, we believe this assumption is justified.

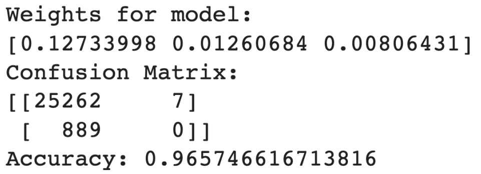
To build these maps, we loaded the *crash_info_general.csv* into a pandas dataframe, before directly querying it using the *pandassql* library. We then cleaned up the data by excluding crash reports which either included null latitudes or null longitudes so that our heatmap wouldn't be unduly influenced by null values.

The heat maps and a relative map of Philadelphia can be found in the above executive summary. Important to note is the fact that the heat maps are relative to the type of weather, rather than absolute to all types of weather. Overlaying the heat maps with the map of Philadelphia yields the likelihood of crashes in different neighborhoods under distinct weather conditions. Thus, civil resources should be apportioned appropriately in order to minimize damage caused by these accidents. The main feature engineering performed in this task was in translating latitude and longitude string values into doubles to be used in our heat maps.

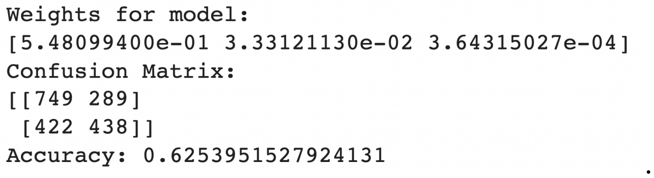
Our examination of traffic stops versus crashes entailed joining the Traffic Stops data set with the Crashes data set on FIPS (census-designated geographic regions), and then plotting/calculating the correlation between the number of traffic stops and the number of crashes for each region. We found an r-value of 0.24, indicating a weak positive relationship between the two variables, and showing that there was no significant relationship to be drawn between the two variables.

We furthermore sought to build a model to predict the likelihood of a fatality or serious injury dependent upon factors such as response time, crash location, and collision type–factors that may be available to a dispatcher upon call and may influence response decisions. To this end, we had to manually compute response times for accidents given times which were in hhmm.0 format (where hh represents 24-hour time, and mm represents time in minutes). Because the time was not standardized in format, i.e. there were times of different lengths, computing response time included excluding null values, standardizing time format, and converting time into minutes since the start of the day, so that response time in minutes could be handled correctly. Furthermore, an edge case in feature engineering arose from situations in which dispatchers arrived the day after they were dispatched (i.e. dispatched at 11:52 pm and arrived 12:15 am). To this end, an alternate formula had to be employed to compute the difference in minutes taking into account next day arrival.

With this data cleaned, we built a logistic regression model using the *sklearn* library, and its StandardScaler feature to make the categorical parameters continuous. In particular, the model predicts whether a crash will result in serious injury or death, based on its type of location, the type of collision that takes place, and the time it takes the nearest police unit to reach it. The model has around 97% accuracy—however, this figure is confounded by the high number of non-fatal or serious crashes in the data (ie data imbalance), and thus skews its predictions. Therefore, we retrained the model by resampling from the larger dataset to create a smaller dataset in order to train on roughly even numbers of fatal/seriously injured and non-fatal/non-seriously injured crashes and its accuracy moved to 62.5%. Based on the confusion matrix of the retrained model, it is more useful in cases of positive identification (i.e. identifying that a crash would be fatal/cause serious injury), with a true positive rate of around 72% (both regression results shown below).



```
Weights for model:
[0.12733998 0.01260684 0.00806431]
Confusion Matrix:
[[25262     7]
 [  889     0]]
Accuracy: 0.965746616713816
```



```
Weights for model:
[5.48099400e-01 3.33121130e-02 3.64315027e-04]
Confusion Matrix:
[[749 289]
 [422 438]]
Accuracy: 0.6253951527924131
```

Future work might involve deploying models to automatically allocate police and emergency response resources in different weather conditions and suggesting courses of action given a limited set of resources and competing emergencies that require assistance.