Identifying Key Predictors of Crash Outcome Severity using PennDOT Crash Data

Phillip Chau, Kevin Lu, Brian Wong | *Dec 15.*

Introduction

In this study we analyze Pennsylvania Department of Transportation Crash Data over the course of a decade. Our goal is to predict the instances where crashes (identified uniquely by crash report number) had fatal or serious injuries as a result, based on different factors such as road conditions, driver behavior, and location type, to name a few. By creating a model to predict this, we wish to then look back at which features contributed the most to predicting outcomes with fatalities or serious injuries. After doing so, we take a deeper dive into the main predictors and do some more research as well to provide context on the problem and to give recommendations on how to improve it.

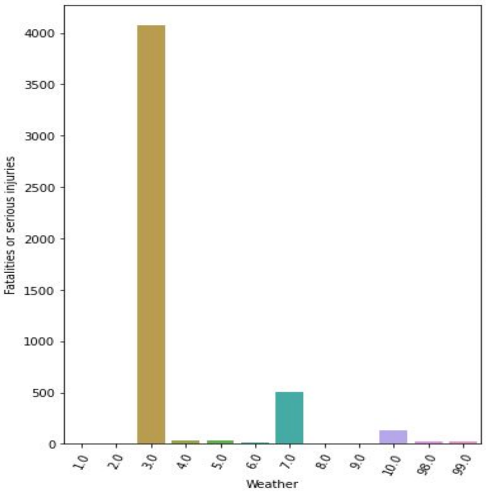
Section 1: Exploratory Analysis

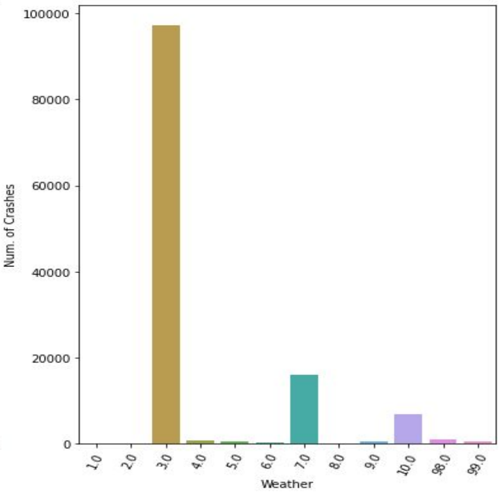
1.1 Analyzing Impact of Time and Weather Hazards

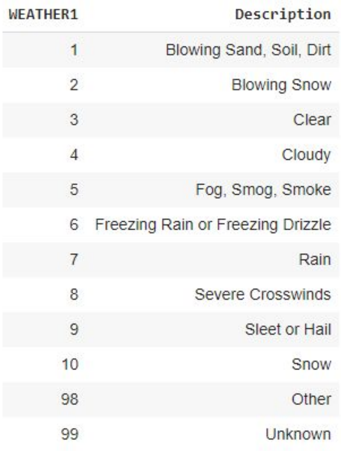
In this section, we look into the impact of the time of day, day of week, weather and speed limit on the number and fatality of crashes. We hope that a preliminary understanding of when and under what circumstances crashes are fatal would allow us to obtain insight into how to attribute the fatality of the crash to respective drivers and the infrastructure of the road. By looking into when fatal crashes are most often, we hope to isolate factors that are intrinsic to the timing and situation of crashes caused fatalities.

Fatal Accidents by Weather

We first look into the distribution of the weather during which the crash occurred, and the fatality percentage under each weather type. By preliminarily looking into the crash count of the dataset, we observe that there are most crashes on clear days, followed by rainy and snowy days. However, this is not reflective of the fatality rate, as the larger number of crashes on such days is due to the higher number of clear days in a year. The number of fatal crashes follows a similar pattern, and is also heavily influenced by the disproportionate nature of the dataset.





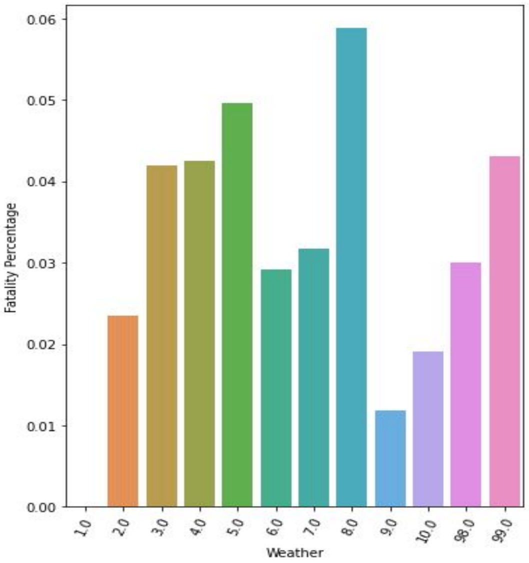


WEATHER1	Description
1	Blowing Sand, Soil, Dirt
2	Blowing Snow
3	Clear
4	Cloudy
5	Fog, Smog, Smoke
6	Freezing Rain or Freezing Drizzle
7	Rain
8	Severe Crosswinds
9	Sleet or Hail
10	Snow
98	Other
99	Unknown

We want to look into the ***fatality percentage*** for each type of weather. This refers to the number of fatal crashes divided by the number of crashes in each group. We believe that this will be a better indicator to see how weather impacts a crash’s fatality.

After scaling for the number of crashes, severe crosswinds pose the highest fatality rate, followed closely by freezing rain. Snow has a relatively low fatality.

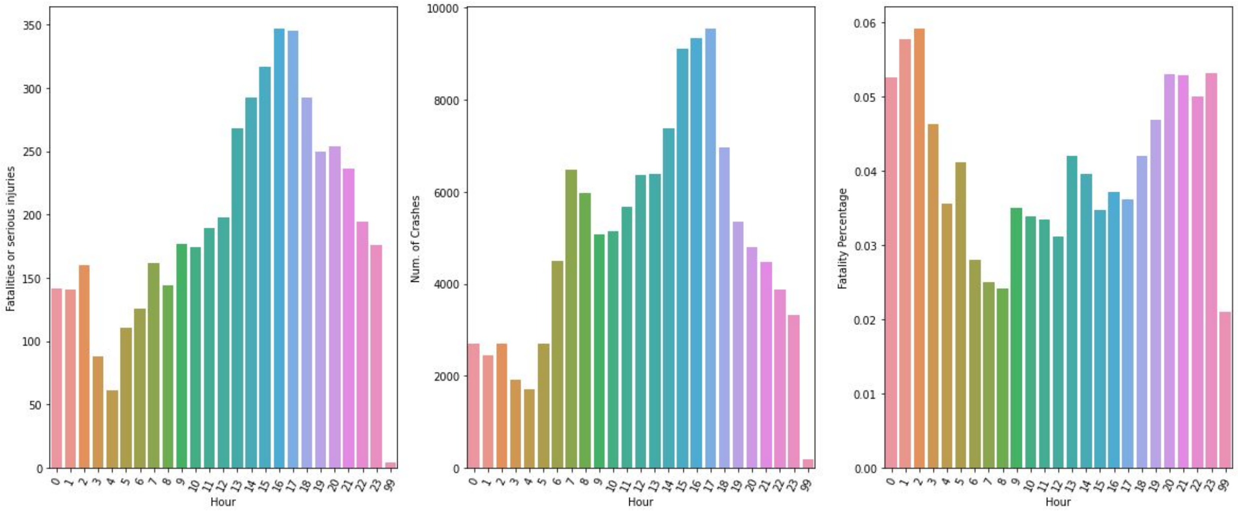
Looking at the datasets for 2015-2019, we have chosen clear, fog, rain, snow as variables for our model. This takes into account their respective frequencies and their impacts on fatality. While some weathers may have higher fatality percentages than others, their absence from earlier datasets make it a bad indicator for our model.



Fatal Accidents by Time

We then look into which hours of day fatal crashes are most likely to occur.

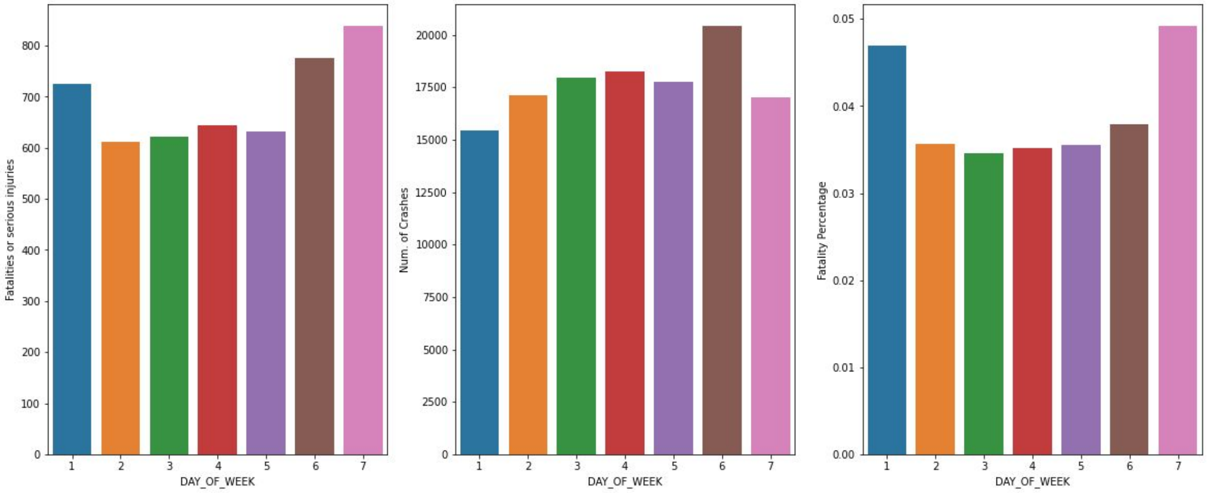
We observe that crashes and fatal crashes most likely happen from 3-6 pm. On the contrary, crashes are much fewer in earlier times of day. Again, this might simply be due to traffic situations and cannot be reflective of what we want to analyze. We look into distributions for *fatality percentage* for further analysis.



The fatality percentage chart shows a different picture. The highest fatality crashes happen during midnight and night hours, with fewer in the morning.

Fatal Accidents by Day of Week

Next, we group by day of week. Note 1 = Sunday, 7 = Saturday.

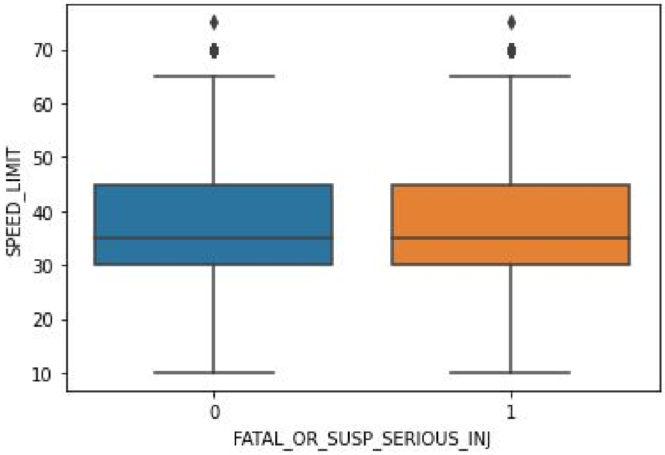


From the graphs above, the counts of crashes seem fairly average over the week, with a higher number of fatal crashes on weekends (Sunday and Saturday). This is clearly reflected in the *fatality percentage* graph, a phenomenon that makes sense as drivers on weekends may not be as accustomed to driving as drivers on weekdays.

Fatal Accidents by Speed Limit

Finally, we look into the speed limit at which the crash happened. We want to see if there is an obvious difference in distribution of speed limits for fatal and nonfatal crashes. As such, we plotted a boxplot as follows.

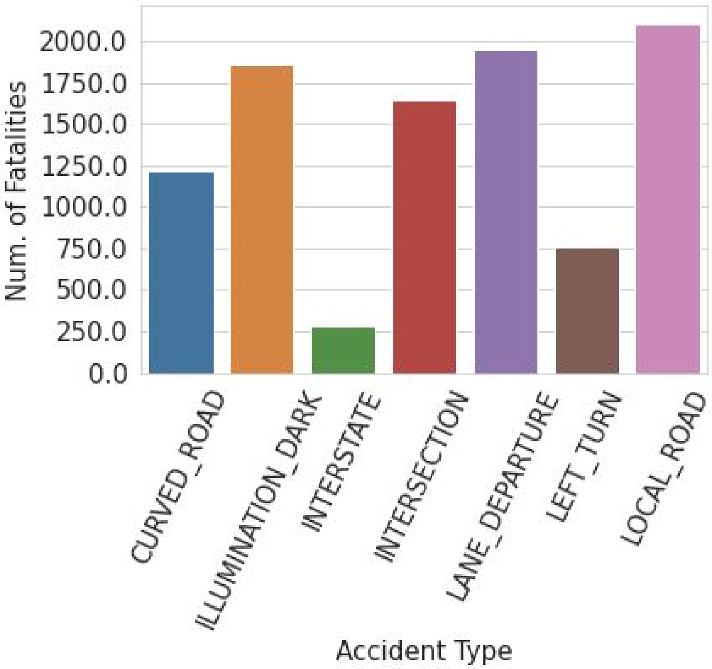
Looking at the plots for the distributions of speed limits, there does not seem to be a big difference in distribution between the crashes that are fatal and those that are not. However, we do think it should have some

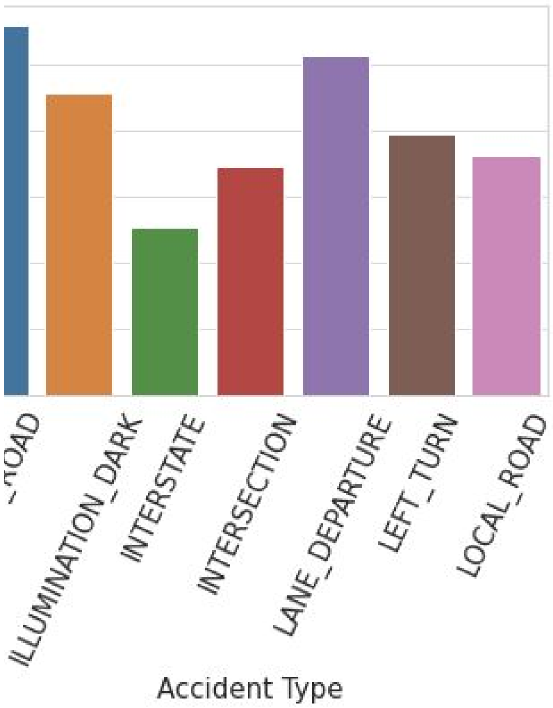


impact on fatality hence we decided to still include it in the model.

1.2 Analyzing Safety of Road Infrastructure

Sometimes, accidents could occur due to issues that are not completely within the driver’s control. Naturally, some aspects of the road in which the drivers are on are more difficult to navigate on than others, which leads us to believe that certain road conditions and designs could lead to more fatal accidents. Thus, we decided to look at 7 different types of collisions due to the road infrastructure: accidents on curved roads, accidents with dark illumination, accidents at interstates, intersections, from lane departure, left turns and on local roads. As seen below, accidents on local roads and from lane departure tended to have the most fatal accidents. However, we had to keep in mind that the number of crashes aren't evenly distributed amongst these categories. For example, local road crashes tended to occur a lot more than interstate crashes.

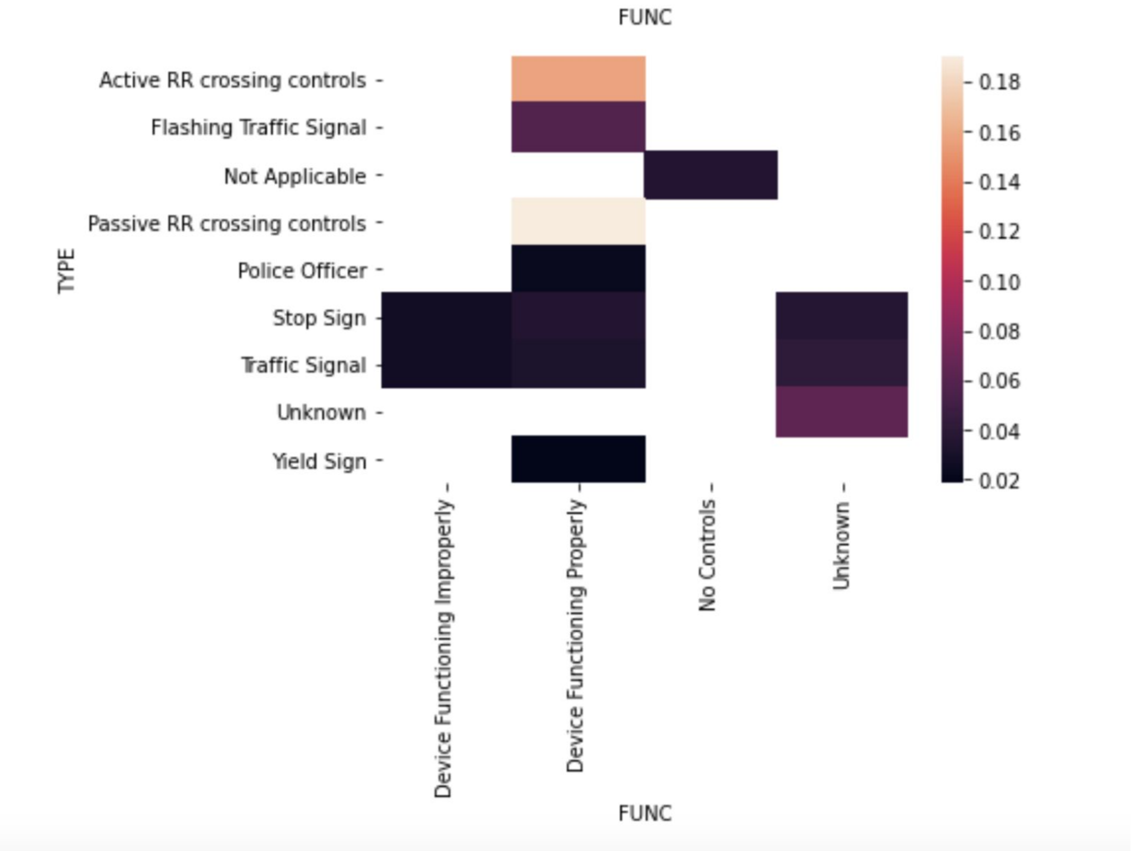




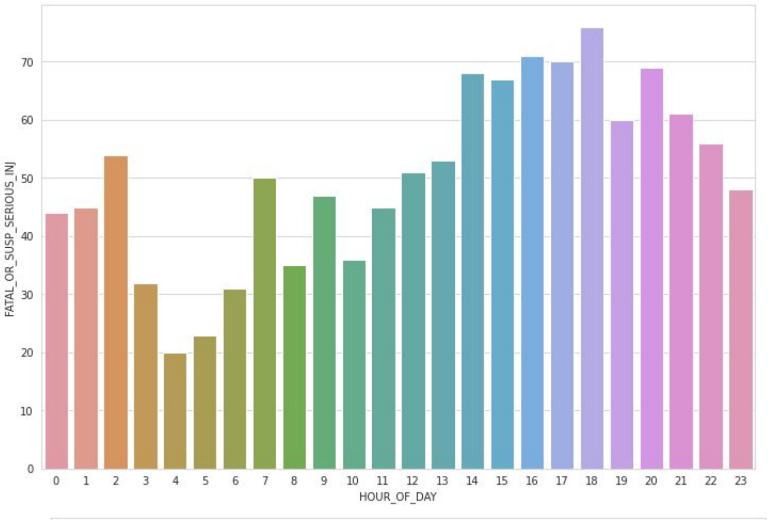
Thus, analyzing the percentage of crashes that are fatal, we saw a slightly different situation where curved roads had the highest percentage. In either case, it appears lane departures still were significant in terms of fatal accidents.

Fatal Accidents at Intersections

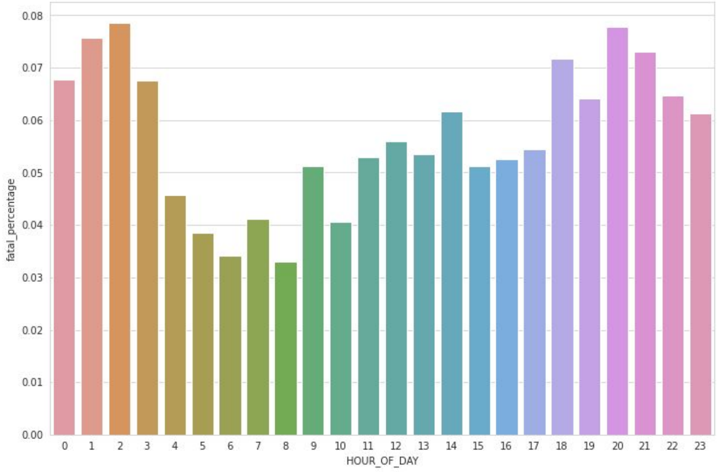
We first analyzed crashes at intersections. The provided dataset also included flags indicating what kind of traffic controls were present at the intersection. Thus, we found it interesting to see if the functionality of these traffic controls had any impact on the fatality of the crashes. Grouping the intersection crashes based on these traffic control factors and plotting them on a heatmap, we saw that the highest number of fatal crashes occurred at intersections with traffic signals or stop signs. However, note that there tend to be more intersections with traffic signals or stop signs than crossing controls for example. Thus, we also analyzed these factors against the percentage of fatal crashes over total crashes. From this heatmap, we saw that the highest percentage of fatal crashes occurred at crossing roads, which would intuitively make sense given that a full on collision with a train sounds quite fatal.



Fatal Accidents at Curved Roads



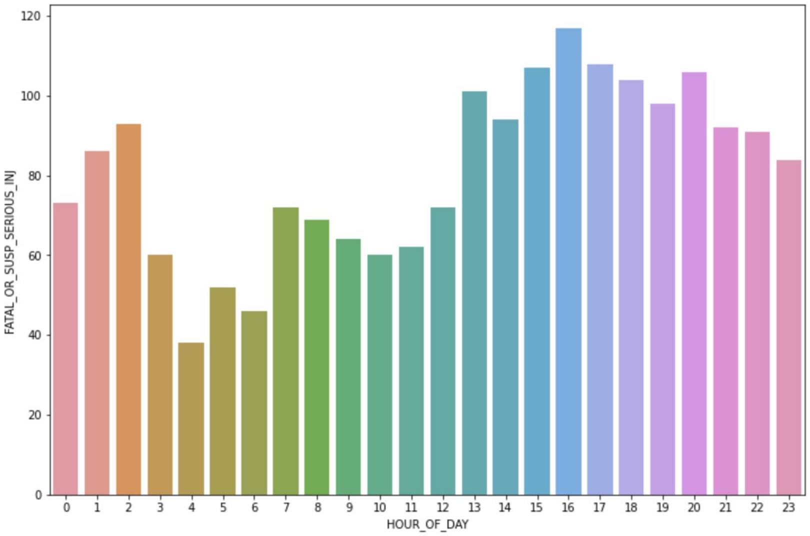
Now we'll take a look at factors that relate to fatal curved road accidents, as these types of crashes have one of the highest fatality percentages. First, plotting the distribution of fatal crashes at curved roads throughout the day, we see that a significant amount of fatal crashes occurred towards the evening, most notably at around 6 pm est. Keep in mind that this is the **total count,** and this may be the case due to the fact it gets dark around the time, hindering visibility of the road. Thus, that leaves greater opportunities for potential crashes.



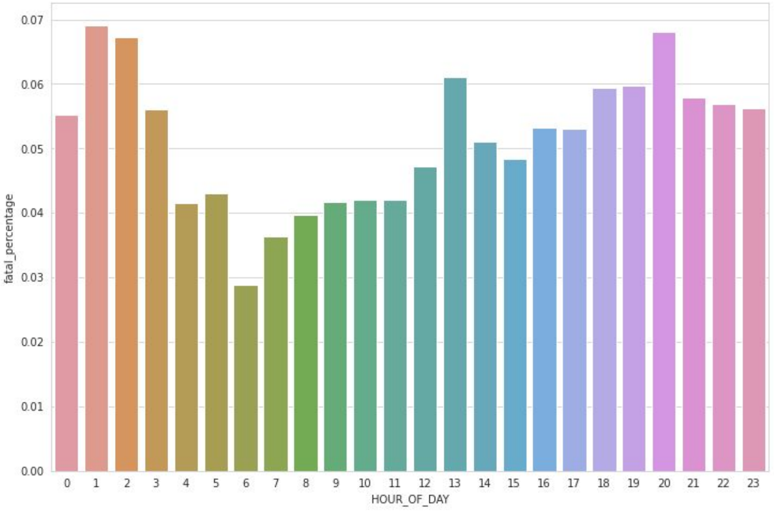
At the same time, the total percentage, however, spiked at around 1 to 2 am, which can possibly account for the fact that drivers are either fatigued or have drank prior to driving and became reckless as a result.

Fatal Accidents from Lane Departures

Finally, we observe some factors involved with fatal lane departure crashes, which are some of the most fatal accidents according to our analysis above. Analyzing how fatal lane departure crashes vary throughout the day, we observe a very similar pattern as curved road fatal accidents. We see that a significant amount of fatal crashes occurred towards the evening, most notably at around 4 pm est. Keep in mind that this is the **total count,** and this may be the case due to the fact that 4 pm is the intersection of many events, including people coming home from work and people going out before it gets dark. Thus, that leaves greater opportunities for potential crashes and more congested lanes.



However, observing the fatal percentage, we see that the most fatal hours are actually 1 - 2 am, which could be attributed to the fact that people could be coming home from drinking or are fatigued while driving.



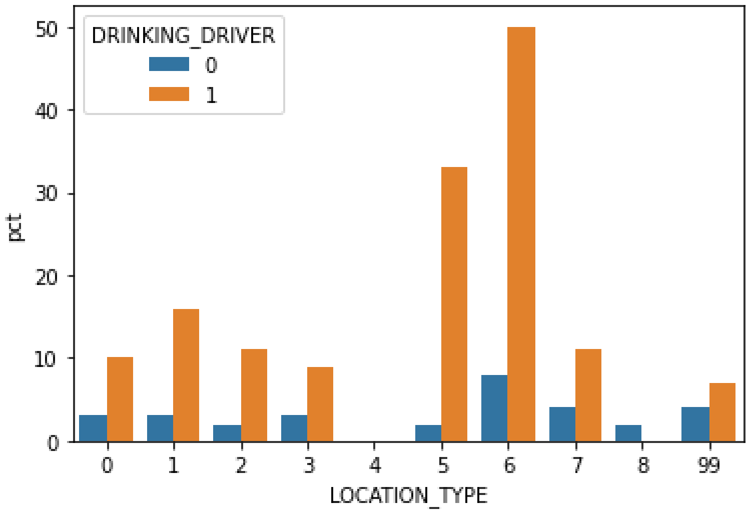
1.3 Analyzing Percentage of Fatal or Suspected Serious Given Driver Related Fault By Location Type

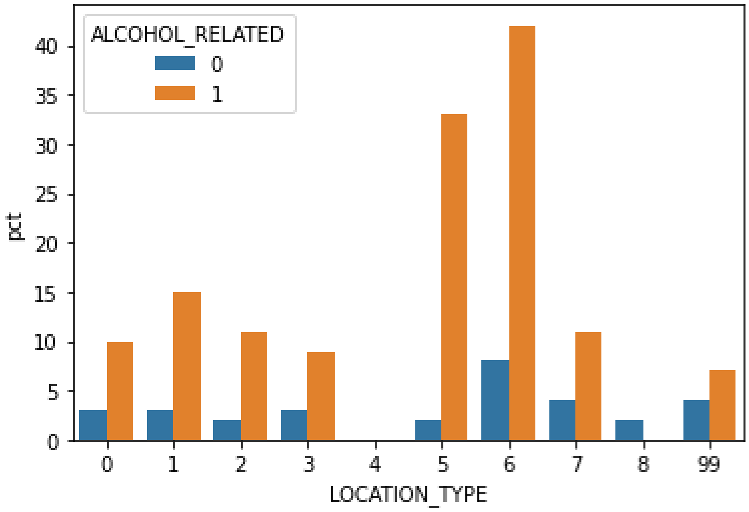
In this section we analyze the various driver faults that may influence the likelihood and severity of a crash occurring. Many of these are quite intuitive, so we were interested in knowing whether the percentage of severe crashes was significantly higher for any of these behaviors, grouped by location type so that we might also get an idea of any locations that see higher than normal severity rates.

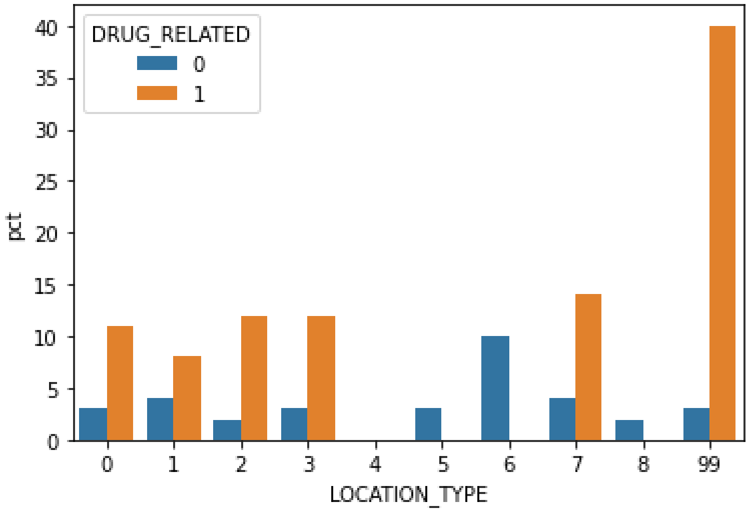
This section will use the following legend:
LOCATION_TYPE
0 – Not applicable
1 – Underpass
2 – Ramp
12 - In open area (back of pickup etc.)
13 - Trailing unit
14 - Riding on vehicle exterior
15 - Bus passenger
98 - Other
99 - Unknown

Fatal and Serious Suspected Injury Percentage By Location Type Due to Drinking Driver, Related to Alcohol, and Drug Use

Here we find that Drinking Driver and Alcohol Related are, as one might expect, almost exactly the same. And of course, alcohol is a significant factor when looking at crashes as the number of alcohol related vehicular deaths annually would suggest. Drug use, which may include alcohol, is also quite serious, particularly in a few location types, indicating that drug usage may be a localized affair.

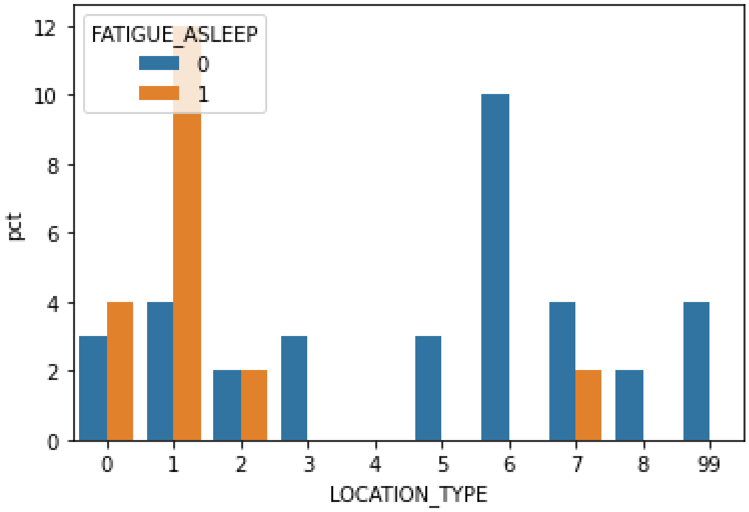

DRINKING_DRIVER
0
1
pct
50
40
30
20
10
0
0
1
2
3
4
5
6
7
8
99
LOCATION_TYPE


ALCOHOL_RELATED
0
1
pct
40
35
30
25
20
15
10
5
0
0
1
2
3
4
5
6
7
8
99
LOCATION_TYPE


DRUG_RELATED
0
1
pct
40
35
30
25
20
15
10
5
0
0
1
2
3
4
5
6
7
8
99
LOCATION_TYPE

Fatal and Serious Suspected Injury Percentage By Location Type Due to Driver Fatigue and Falling Asleep

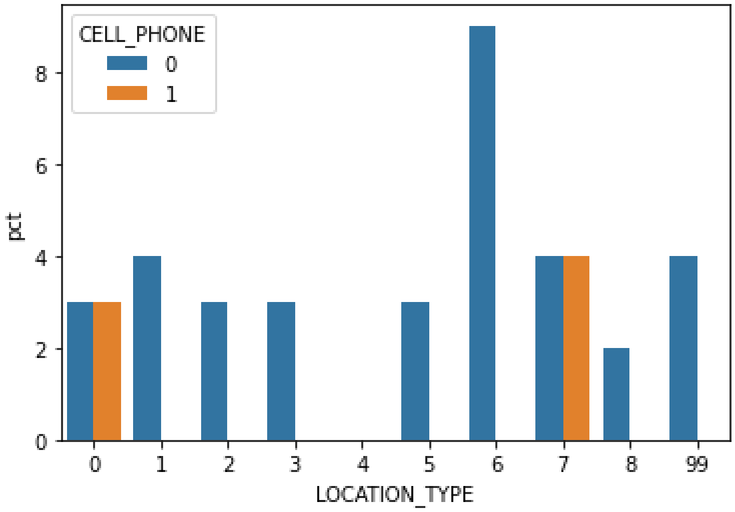
Here we look at what happens when drivers fall asleep at the wheel, which apparently causes serious accidents only in some places, especially location type 1. Overall, it does not seem to be a big problem as even when it does happen to be severe, it does not appear to be significantly more so than when it isn't severe.

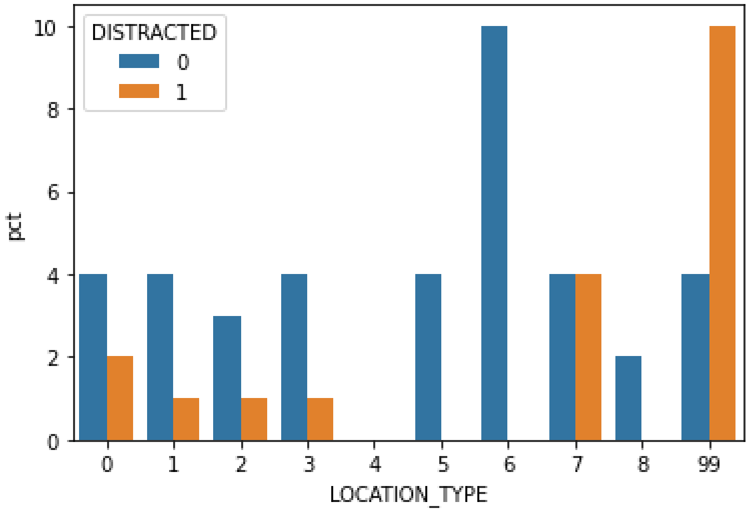


Fatal and Serious Suspected Injury Percentage By Location Type Due to Cellphone Use and Being Distracted

Here we take a look at drivers using their cellphones while driving as well as the more general distractions. It appears that cellphone use only happens to result in serious crashes in two location types, and that neither has a higher rate of severe incidence than non-severe. This indicates it is probably not of too much importance, although some investigation into why location types 0 and 7 are particularly susceptible to cellphone use while driving in the first place may add value. Distracted driving in general appears to result in crashes more often in different location types, but generally at a very low frequency of serious crashes. Location types 6 and

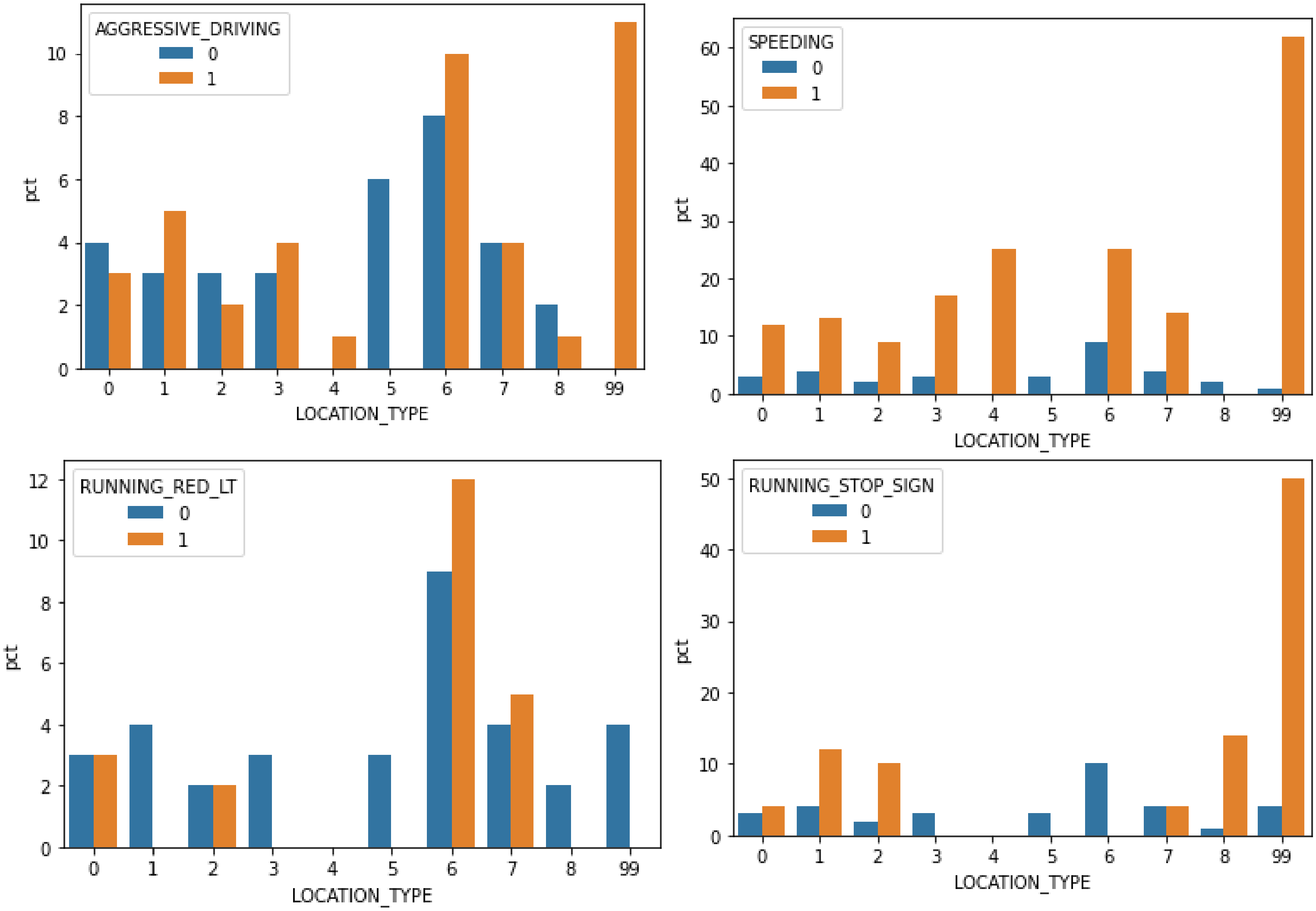
99 (unknown) are a bit intriguing given that they have an incredibly high incidence of crashes, despite 6 being all minor in nature and 99 having a very high incidence of serious crashes worthy of investigation and rectification. On the whole, however, it doesn't look as if the distractions are a particularly big deal.





Fatal and Serious Suspected Injury Percentage By Location Type Due to Aggressive Driving, Speeding, Running Red Lights and Stop Signs

In this section we look at 4 factors that are aggressive driving and behaviors derivative of it. We see a much higher incidence of severe cases on the whole, with only running red lights a bit less dangerous than the other three. In addition, we see that location type 99 (unknown locations) is once again the center of danger as it has incredibly high levels of serious or fatal injuries. However, one thing to keep in mind is that for 99 there are very few crashes there in total, so it may be that it simply appears abnormally high due to small sample size.



Section 2: Data Pre-Processing

2.1 Splitting the Data for Training and Testing

We split the dataset into a training set and a test set in 3 different ways - once training 2015-2018 and testing in 2019, the second time trained on 2015-2017 and tested on 2018-2019, and the third time on an 80-20 train-test random split. Data is available for 2020, but it is not complete and is currently being updated as this project is done in the December of 2020).

Despite having data 2 decades old, we believe that enough significant construction and changes have been made in the past 2 decades to render such data obsolete. Within 20 years, any amount of construction and policy changes may have occurred, which would drastically impact our model's ability to generalize based on past data. In terms of performance, the previous two managed to underfit significantly such that training accuracy was ~0.6 while testing accuracy was ~0.7, across several different models. Our random split iteration ended up resolving this issue.

While it is true that anything that happened in say 2017 would still cause our model harm in terms of predictive power, the odds are lower and given more time we would explore a mixed approach with both random split and dividing by year. Even more interesting would be to research what construction went on during the last 5 years and if there were any significant policy changes that would influence the nature of the accidents that occurred.

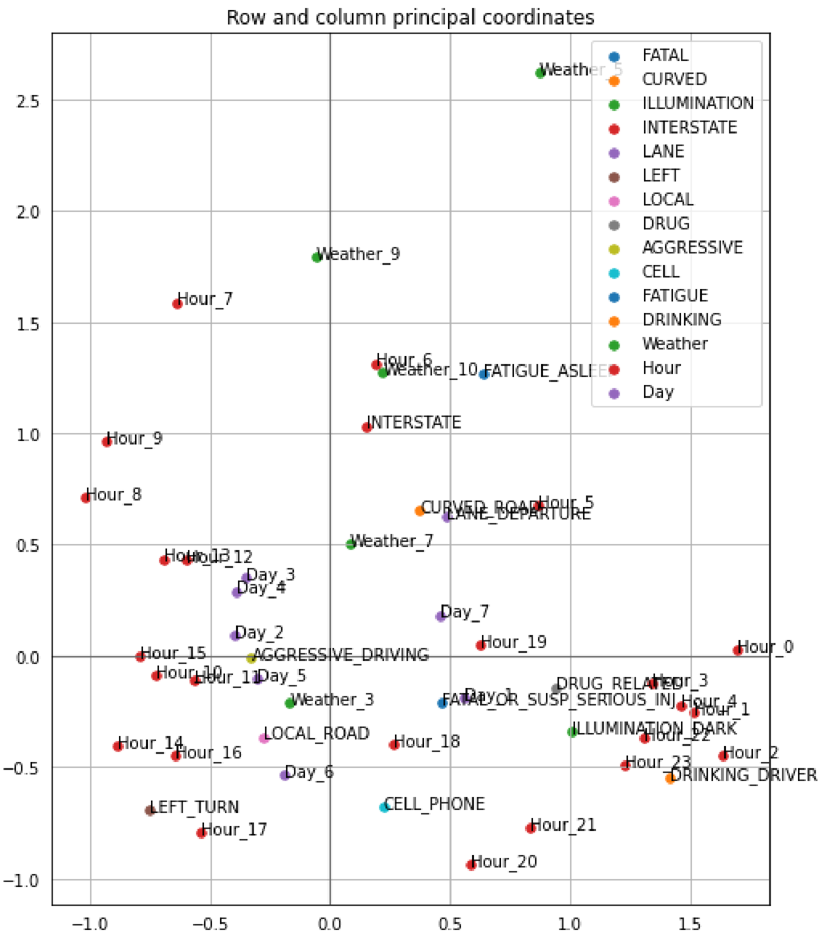
Oversampling

One point to note is that after splitting our data, we investigated the proportion of fatal and serious injury crashes in the training set and found that it was ~5%, which is too small to effectively train on. Thus, we oversampled to sklearn's default 50%, which started returning us meaningful results. Had we not done this, we would simply predict not serious crashes each time, as it would have a 95% accuracy simply saying no every time, which would outperform whatever our models gave us.

2.2 Feature Extraction

MCA

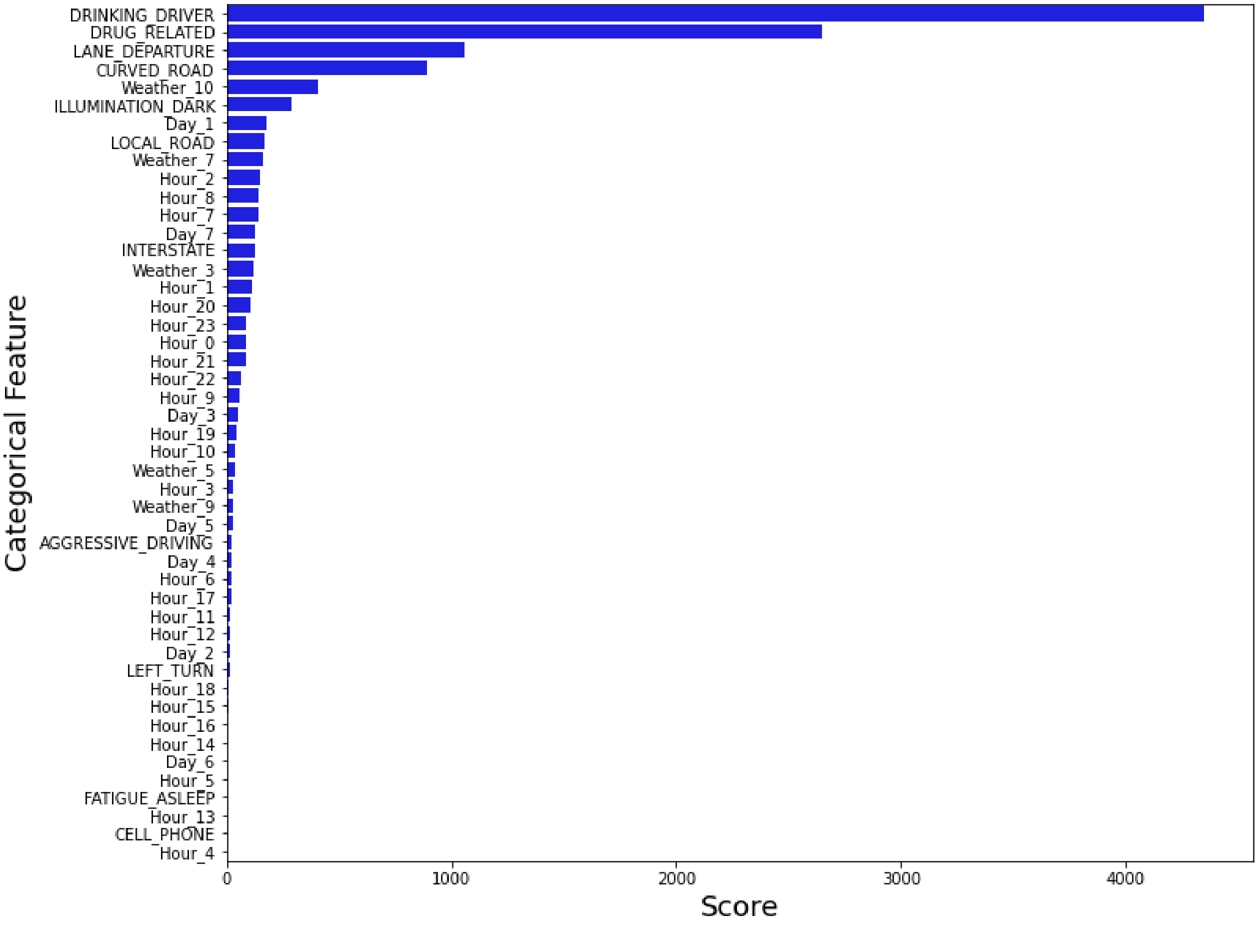
As seen by the exploratory analysis, some of the variables appear to be potentially correlated. Thus, in the interest of filtering out redundant and irrelevant features, we hoped to identify which features are most correlated and simply take the most relevant features in order to have our feature set convey as much information as possible. Traditionally, we thought of Principal Component Analysis in order to do dimensionality reduction. However, given that all of our features were binary, this simply wasn't possible. Thus, we resorted to using MCA (Multiple Correspondence Analysis), which is a correspondence analysis technique used to discover and reduce interrelationships between categorical variables.



The plot above essentially shows the relationship of the features to each other. We noticed that factors such as drug related crashes, crashes that occured in late hours and crashes from drinking all are grouped close to each other and close to the class variable. This is a sign that these variables could potentially be relevant. But at the same time, the plot also shows correlation between some of these variables. It makes sense that crashes that occur from 11 pm to 3 am and crashes from drinking are correlated for example, which is a sign that they can possibly convey similar information.

Chi-Squared Statistic

To reaffirm our understanding of the MCA plot, we used the selectKBest function in sklearn in order to do feature selection. In feature selection, we often attempt to choose features that are highly dependent on the class/response. Since chi-square tests are often used to test the independence of two events, we usd selectKBest to rank the most relevant features based on those with the highest chi-square value (the most dependent to the class). The rankings are seen below.



Through MCA we noted that crashes from drinking, crashes in dark illumination and drug related crashes appeared correlated to fatal crashes. This conclusion continues to be reflected in our feature extraction, helping us narrow down our features based on those with the greatest relevance while filtering out those that may convey repetitive information.

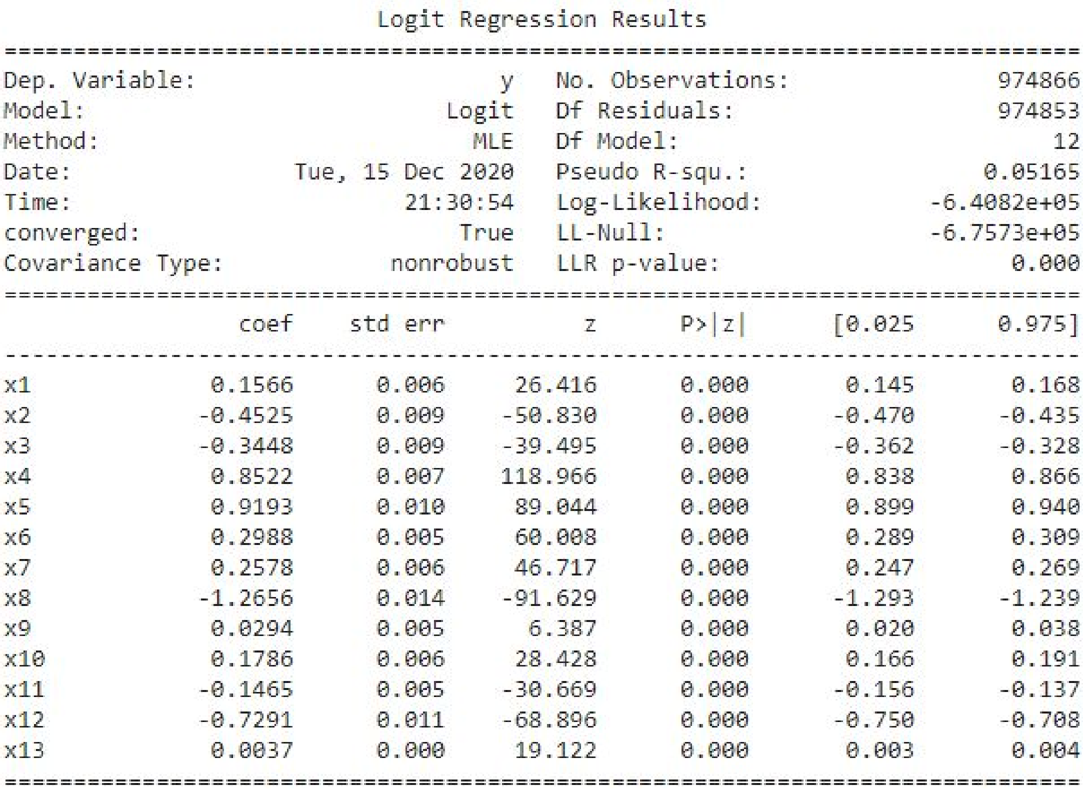
Section 3: Modelling

Logistic Regression

We chose to first attempt a logistic regression because we believed that it would do better representing the binary variables than a linear regression. As we are attempting to classify the crashes as fatal vs non-fatal, we can use a logistic regression model to predict the probability of the fatality of a crash. For our analysis, we used *p = 0.5* as the hard limit threshold to classify our results.

We decided to use the MCA-processed train and test datasets for model fitting, as an assumption of logistic regression is the absence of multicollinearity, which we hope to have eliminated through MCA. In terms of accuracy, it also performs better than the non-MCA processed dataset.

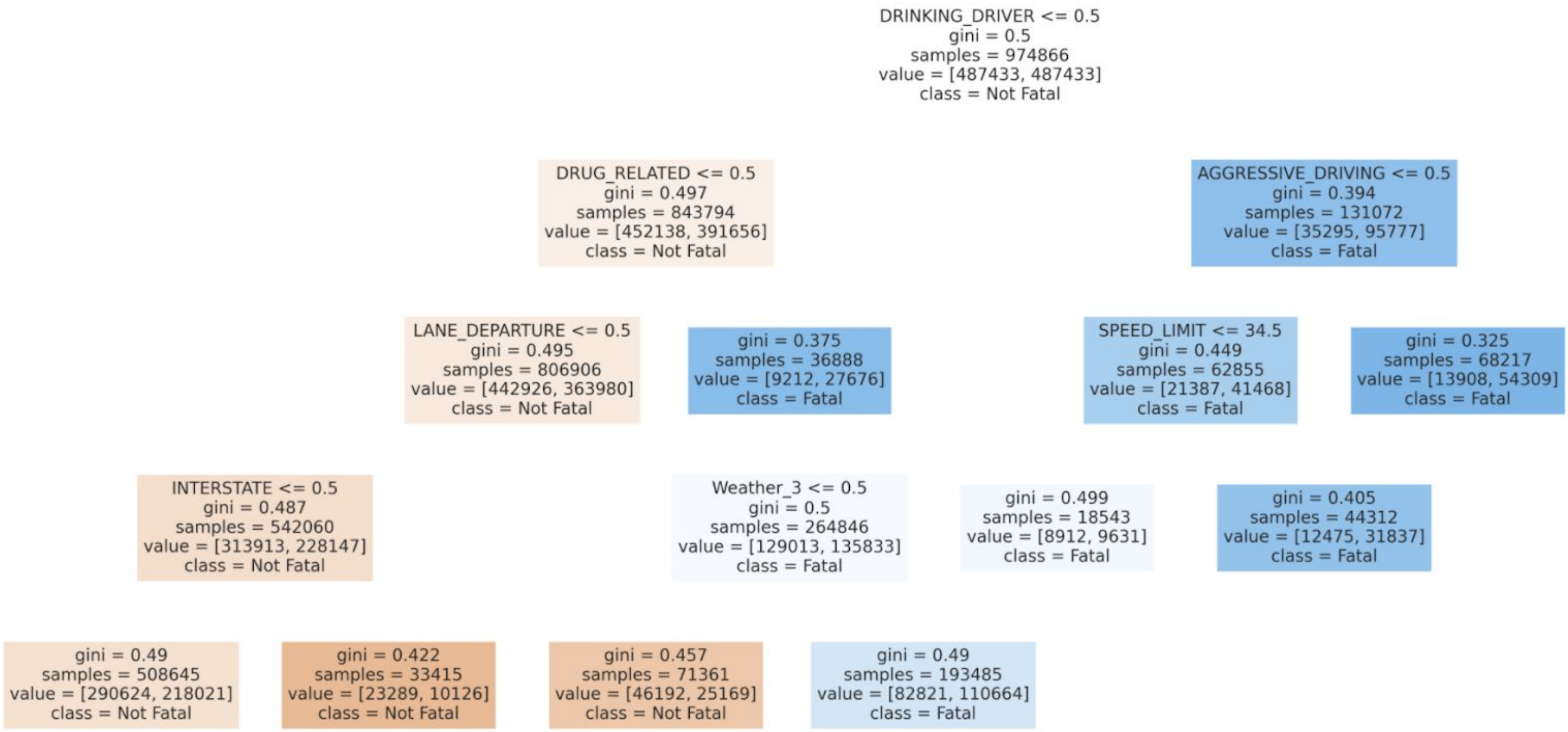
The results for the logistic regression model is shown below.



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:               974866
Model:                          Logit   Df Residuals:                   974853
Method:                           MLE   Df Model:                           12
Date:                Tue, 15 Dec 2020   Pseudo R-squ.:                 0.05165
Time:                        21:30:54   Log-Likelihood:            -6.4082e+05
converged:                       True   LL-Null:                   -6.7573e+05
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.1566      0.006     26.416      0.000       0.145       0.168
x2            -0.4525      0.009    -50.830      0.000      -0.470      -0.435
x3            -0.3448      0.009    -39.495      0.000      -0.362      -0.328
x4             0.8522      0.007    118.966      0.000       0.838       0.866
x5             0.9193      0.010     89.044      0.000       0.899       0.940
x6             0.2988      0.005     60.008      0.000       0.289       0.309
x7             0.2578      0.006     46.717      0.000       0.247       0.269
x8            -1.2656      0.014    -91.629      0.000      -1.293      -1.239
x9             0.0294      0.005      6.387      0.000       0.020       0.038
x10            0.1786      0.006     28.428      0.000       0.166       0.191
x11           -0.1465      0.005    -30.669      0.000      -0.156      -0.137
x12           -0.7291      0.011    -68.896      0.000      -0.750      -0.708
x13            0.0037      0.000     19.122      0.000       0.003       0.004
==============================================================================
```

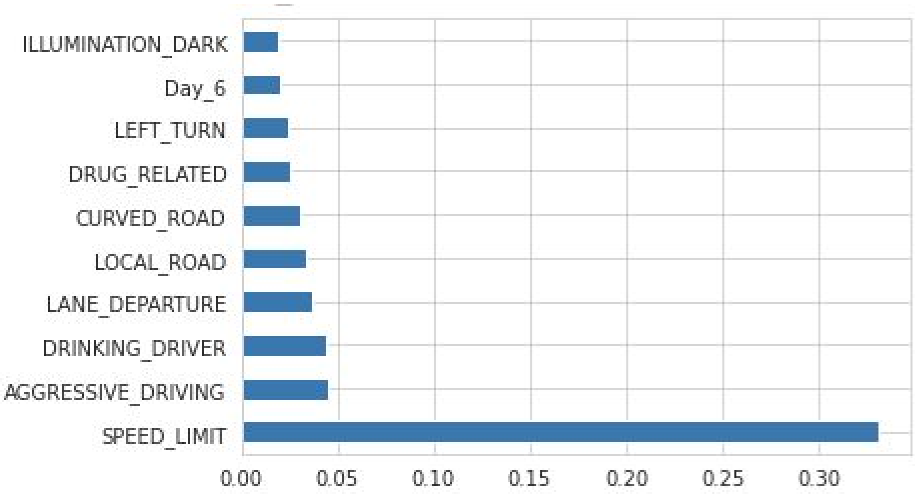
Decision Trees

Switching from regression techniques, we wanted to try a single decision tree first to see if this would, as we would expect, perform better on predicting a binary variable on binary features. This model was more or less a proof of concept before running the random forest, so that we could observe if it was indeed able to make predictions based on the features given. As seen below, the decision tree performed just like how we thought it would. Crashes due to drinking, drugs, aggressive driving and lane departures seemed to be good predictors for fatal accidents, just as we saw from the best feature extraction.



Random Forest

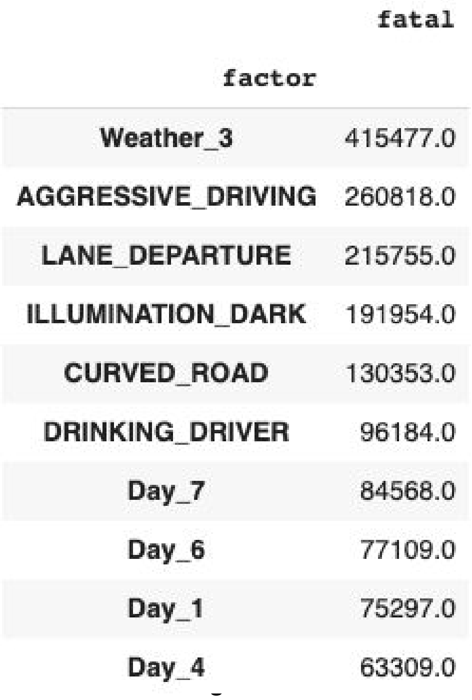
Expanding on the idea of decision trees, we decided to use the random forest model, as it is a particularly common and robust model able to cover many cases. The performance of the random forest was indeed the best overall, and using the built in feature importance ranking we were able to identify several of the most valuable predictors for the severity of a given crash. As seen below, speed-limit appears to be a very important factor in determining fatal accidents. Unlike most of our other models, random forest is able to capture this single continuous variable along with the binary features, which helped us see the importance of all our features rather than filtering out the non-binary ones.



Naive Bayes Classification

Naive Bayes Classifiers are simple yet effective probabilistic machine learning models for classification. For our purpose, since all of our features are categorical, we used a Bernoulli Naive Bayes Classifier, which is designed specifically for binary variables. However, since these models are naive, it assumes that all the features are independent. In our original dataset, it seems a bit obvious that not all the features are completely independent. For example, just based on real world observation, crashes from drinking and crashes that occur at late hours like 1 am are probably related. Fortunately, before running our models, we did dimensionality reduction with MCA on our data in order to account for these related variables. This allowed us to proceed with creating our naive bayes classifier.

Our Bernoulli Naive Bayes Classifier helped us see which features had the greatest predictive power in determining whether a crash was fatal or not by telling us how many instances the model predicted that a given feature led to a fatal accident. As seen below, some of these features were exactly predicted by our previous models as well, such as aggressive driving, lane departures and drinking and driving.



	fatal
factor	
Weather_3	415477.0
AGGRESSIVE_DRIVING	260818.0
LANE_DEPARTURE	215755.0
ILLUMINATION_DARK	191954.0
CURVED_ROAD	130353.0
DRINKING_DRIVER	96184.0
Day_7	84568.0
Day_6	77109.0
Day_1	75297.0
Day_4	63309.0

Results

	Accuracy	***MSE***	***AUC Score***
Logistic Regression	70.4%	0.544	0.602
Decision Trees	72.7%	0.522	0.603
Random Forest	78.4%	0.465	0.524
Naive Bayes	68.9%	0.558	0.598

As shown above, in terms of accuracy, it appears that random forest performed the best out of all our models. However, the AUC score of decision trees appears to be a lot higher than that of random forest. While this means that technically decision trees did a better job at distinguishing between fatal and non-fatal crashes, we must keep in mind that decision trees are prone to overfitting. Models such as logistic regression and naive bayes performed pretty decently; however, the fact that we may not have filtered out all dependent features may have weighed down Naive Bayes' performance. In the end, random forest was the most preferred model for our dataset since it allowed us to capture both continuous features such as speed as well as binary features when making a classifier for fatal crashes.

Further Exploration

On the technical side, we would have liked to spend some more time on looking into neural networks, support vector machines, and other extensions of what we currently have. In addition, we would take some more time to work on hyperparameter tuning as our accuracy results were not particularly impressive.

After finding the features with the most predictive power, we could actually look into what Pennsylvania has done in regards to these issues as well as look into any other datasets related to them for further research. After knowing what causes fatalities, we could actively research where crashes related to that feature happened and thus create a scatterplot to show the density of crashes in any given area. This way, we will be able to identify specific roads and locations that are particularly problematic in regards to safety, at least statistically.

Conclusion

Our project created a model capable of predicting the severity of a crash to an acceptable degree of accuracy. With this model, we are able to identify the most important features contributing the most to one's likelihood of being involved in a serious crash given the event of a crash. We tested a variety of different data science methods here, such as oversampling, MCA, and using Chi-Squared Statistics. Oversampling as mentioned earlier was instrumental to training our data. MCA and Chi-Squared statistics were used to remove variables with high correlation and low predictive power. However, perhaps due to the nature of the data, we found that including these two feature reduction techniques actually decreased our accuracy slightly from just evaluating the model on all the variables. In the end, iterating over different splits of the data for training and testing as well as over different models, we ended up with a random forest model that performed reasonably well. From this model we were able to isolate a few features believed to be particularly important to gauging the severity of a crash, and we believe they are worthy of further exploration such that

Pennsylvania’s Department of Transportation can investigate and improve our infrastructure in accordance with these findings.

Link to Notebook:
https://colab.research.google.com/drive/1sqB6r4EEKIM14ks-npEw0UP8gZhubQrF?usp=sharing