Seattle Paramedic Deployment Optimization
Enabling Quicker Decisions

Section 1 - Introduction:
At a public town hall, several citizens of Seattle raised concerns over response times for Emergency Medical Services (EMS) without sharing key data to support their claims. Having heard their specific accounts of severe accidents and their level of anxiety, the Mayor and head of EMS would like to understand the circumstances and patterns surrounding severe accidents. Findings could lead to quick-win changes in training guides to better prepare their 911 operators for timely and appropriate EMS responses.

As confident as the internal EMS team was in their current procedures, they were open to the review, because real life cases have proven to deviate from even the best training guides. In their experiences, long conversions with someone who is hysterical or not capable of doing an on-scene assessment delays the appropriate course of action or leads to misguided decisions. Sending services for every accident is cost prohibitive. And, there's not enough time to wait for the police to arrive, assess and then phone in a request.

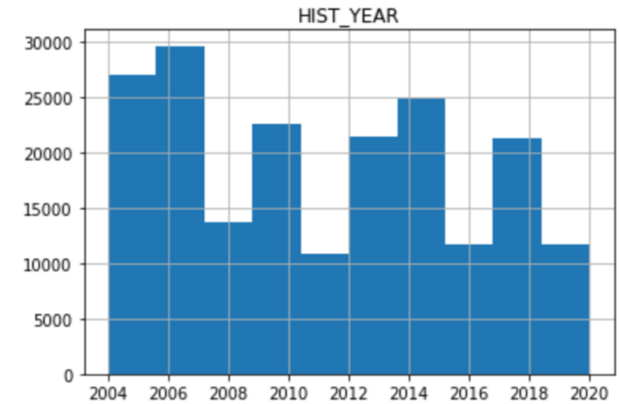
Identifying the key differentiators, in the context of the description of the accident, can aide in guiding the 911 operators' decisions leading to faster deployment, and saving of many lives, without an undue burden of cost. The findings and changes to procedures can then be communicated back to the public in order to restore their full confidence in the EMS department.

Section 2 - Data:
In order to identify those key differentiators of severe accidents, a robust dataset of incidents is required. As it turns out, the city of Seattle has all the data, with companion metadata, available electronically. A dataset for the study was found here: [Seattle Data Collisions].

A candidate dataset should, at foremost, identify if there were injuries involved or somehow identify severity. This dataset had a categorical variable, SEVERITYDESC, the clearly identified injury collisions versus less severe property damage only collisions and could be used as the target variable. However, analysis of the values showed a heavy bias towards property damage incidents, and that would have to be accounted for when splitting the training data.

```
Property Damage Only Collision    136485
Injury Collision                   58188
Name: SEVERITYDESC, dtype: int64
```

The data set should be robust in time with many years of data. However, it does not necessarily need to enable time series evaluation for this study. This dataset happens to have over 150K collisions dating back to 2004.



This dataset was robust in features, such as: road conditions, weather, street and types of vehicles involved. The analysis considered several feature variables, including: address type (alley, block, intersection), weather, road conditions, light conditions, collision type, junction type, under influence flag and collision description.

The variables did have NULL values that were replaced with their most common value. For example, 'Parked Car' replaced empty values in the collision type field. As seen below, 'Parked Car' originally had 47,987 values and 4904 NULLS were replaced with 'Parked Car'.

```
Parked Car     52891
Angles         34674
Rear Ended     34090
Other          23703
Sideswipe      18609
Left Turn      13703
Pedestrian      6608
Cycles          5415
Right Turn      2956
Head On         2024
Name: COLLISIONTYPE, dtype: int64
```

Although the collision description variable contained a standardized list of values, there were 39 values and caused concern for a sparse model. Top ten values shown below:

```
MOTOR VEHICLE STRUCK MOTOR VEHICLE, FRONT END AT ANGLE     85209
MOTOR VEHICLE STRUCK MOTOR VEHICLE, REAR END               54299
MOTOR VEHICLE STRUCK MOTOR VEHICLE, LEFT SIDE SIDESWIPE     9928
NOT ENOUGH INFORMATION / NOT APPLICABLE                     9787
MOTOR VEHICLE RAN OFF ROAD - HIT FIXED OBJECT               8856
Name: SDOT_COLDESC, dtype: int64
```

Although time of day was available and could be grouped, the light conditions variable had adequate groupings and was used instead.

```
Daylight                    121307
Dark - Street Lights On      48507
Unknown                      13473
Dusk                          5902
Dawn                          2502
Dark - No Street Lights       1537
Dark - Street Lights Off      1199
Other                          235
Dark - Unknown Lighting         11
Name: LIGHTCOND, dtype: int64
```

Address type and junction type were found to have significant correlation, and junction type was dropped from the feature set. Same was true for road conditions, and weather was dropped from the feature set.

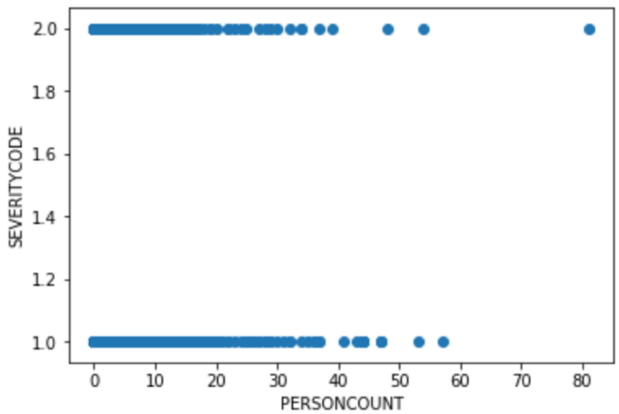
	ADDRTYPE	WEATHER	ROADCOND	LIGHTCOND	COLLISIONTYPE	JUNCTIONTYPE	UNDERINFL
ADDRTYPE	1.000000	-0.090608	-0.030661	-0.053434	-0.466902	-0.915249	-0.041532
WEATHER	-0.090608	1.000000	0.752051	0.208585	0.026320	0.111295	-0.038970
ROADCOND	-0.030661	0.752051	1.000000	0.022630	-0.003217	0.042160	-0.008955
LIGHTCOND	-0.053434	0.208585	0.022630	1.000000	0.034426	0.057661	-0.218037
COLLISIONTYPE	-0.466902	0.026320	-0.003217	0.034426	1.000000	0.470961	0.002491
JUNCTIONTYPE	-0.915249	0.111295	0.042160	0.057661	0.470961	1.000000	0.048592
UNDERINFL	-0.041532	-0.038970	-0.008955	-0.218037	0.002491	0.048592	1.000000

Ordinal features were also considered, such as: number of people, vehicles, pedestrians and bicyclists involved. First step is to check for NULLs and replace with the variable's average value. However, none of the numeric variables had NaN values.

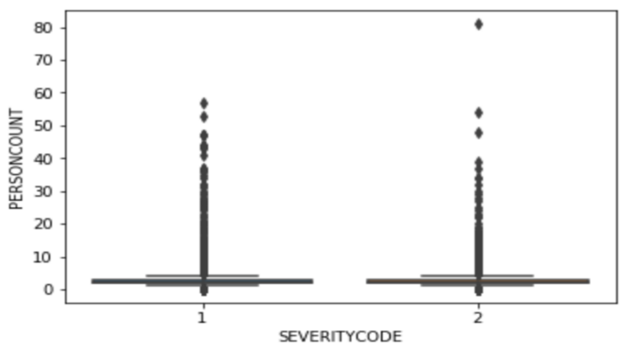
There was a possibility that severity increased with the number of people involved. However, correlation analysis, scatter and boxplots dispelled that theory.

	SEVERITYCODE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT
SEVERITYCODE	1.000000	0.130949	0.246338	0.214218	-0.054686
PERSONCOUNT	0.130949	1.000000	-0.023464	-0.038809	0.380523
PEDCOUNT	0.246338	-0.023464	1.000000	-0.016920	-0.261285
PEDCYLCOUNT	0.214218	-0.038809	-0.016920	1.000000	-0.253773
VEHCOUNT	-0.054686	0.380523	-0.261285	-0.253773	1.000000

Scatter plot



Box plot



Person Count – Top 10

```
2     114231
3      35553
4      14660
1      13154
5       6584
0       5544
6       2702
7       1131
8        533
9        216
Name: PERSONCOUNT, dtype: int64
```

Thus, none of the ordinal variables were carried forward as features.

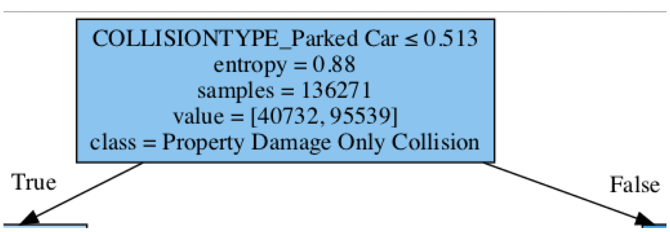
There wasn't be a need for hospital outcomes data or insurance adjuster reports to further assess the severity, since it was provided in the core data. There's no need for specific geographic location, since the location is already confined to the city of Seattle and would otherwise lead to a sparse model. The analysis also does not need to deep-dive into the accounts reported at the town hall nor personal information of the people involved. Thus, no additional datasets were required to perform the analysis.

Section 3 – Methodology:

The data collection and feature engineering should allow for a sufficiently accurate and explainable model such that reasonable confidence can be achieved in its outcomes. The outcome of the model is to provide insights on when to send paramedics. The model does not need to predict number of injured people, estimated medical bills or predicted damage costs. Once the desired model is achieved, those key differentiators can be identified for reasonability checks and inclusion in the updated training guides.

With that being said, we want to help the 911 operators make better decisions, so we just need a decision tree model and we're done. Right? Well, it wasn't that simple.

First, we didn't want to take for granted that the decision tree model could provide accurate results compared to other models. Three different model types were evaluated: decision tree, K-Nearest Neighbor and Logistic Regression. Fortunately, decision tree accuracy on train versus test and against other models was comparable.



'Parked Car'. Yes, parked cars. The top split on the decision tree was based on the Collision Type's Parked Car value. Should the Mayor make parked cars illegal? Should they warn the public about the atrocities caused by vicious parked cars? Let's look for a better answer, like severe accidents happen at night, or on icy roads, or in intersections.

Collision Type had more that 4900 NULL values, and since 'Parked Car' was the most frequent value, we classified even more to the 47K+ existing accidents. Classifying the NULLs into the 'Other' bucket didn't change the outcome.

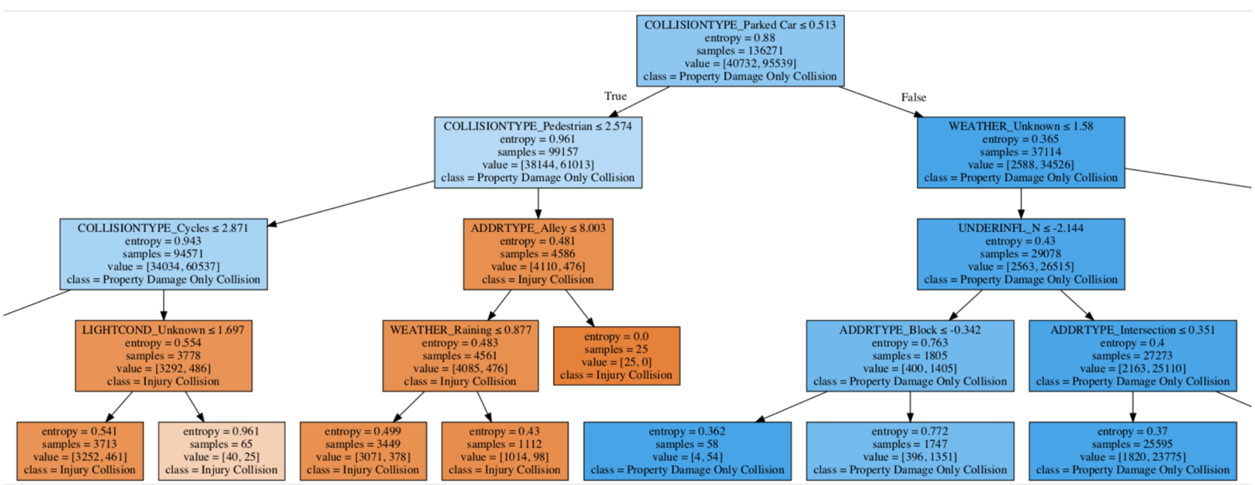
Perhaps, the Collision Type feature itself had too much influence on the model. With the NULLs restored back to 'Parked Car', some of the dropped features were added back in. And, try dropping some that were originally selected. With each change in feature selection, and checking against the KNN and LR models for deviations, there was no change. No luck.

OK. Restore all the features back to the original set and try shaking the tree another way. Change up the decision tree's criterion and max_depth. Yet again, the outcome was dominated by the 'Parked Car' value.

Decision trees can work on numeric and categorical variables. Skip the features converted via get_dummies and use the original values. Still, the outcome was dominated by the 'Parked Car' value.

Was it time to outsource the analysis, so that *we* didn't have to tell the Mayor of the looming Parked Car threat to the citizens?

Section 4 – Results:



Well, there you have it. Parked Cars are chasing down Pedestrians and Cyclists and causing them severe injuries. It clearly says 'True' right there at the top. Except, when the get_dummies is used to convert the categorical feature to numeric, the meaning of the values got switched around. Although the feature says Parked Car, the 'True' path is actually the opposite. Said another way, an accident between a moving vehicle and a pedestrian is more likely to be severe. And, the same can be said for moving vehicles and cyclists. Good news, no one has to tell the Mayor that Parked Cars are out to destroy mankind.

Section 5 – Observations:

We have many years of data, but it is a reasonable size to open in Excel. Using the original data without any feature engineering, we can see that Pedestrian accidents are likely to be severe 89% of the time, and Cyclist accidents are severe 86% of the time. This begins to instill confidence in the outcomes.

While the model's accuracy is only 75%, it is stable. No matter what we throw at it, by changing parameters and/or features, it produces the same outcome. It also compares well to the other KNN and LR models. We could spend hours, days, weeks and so on trying to improve accuracy of the model, but we're limited by the time the Mayor has to respond to the public, and of course, the data itself.

The decision tree gives a clear recommendation. When accidents occur between moving vehicles and pedestrians or cyclists, paramedics should be called. The recommendation itself seems reasonable, compared to a recommendation of calling the paramedics every time there's an accident in the rain in Seattle.

Going back to the training guides for the 911 operators, these updates can be easily applied, so that the operators do not have to continue the conversation to assess bleeding, concussion, broken bones, etc. and further establish probable need. There is some inherent risk of calling for EMS services too soon, but calculated steps need to be taken to deal with the public outcry.

Section 6 – Conclusion:

It's time to go back to the Mayor and Head of EMS. We have a high-level view of the problem, analysis and recommendation for the public, easy to explain changes for training and procedures documents for the Head of EMS, and a more detailed view of the analysis for the Mayor to be able to respond to questions. Based on these deliverables, the Mayor needs to assess the cost of potentially excess EMS calls. Or, choose to spin it another way that they're visibly doing everything they reasonably can to address the publics' health and medical response concerns.