INTRODUCTION

This data analysis investigation has been conducted in order to assist the Seattle Department of Transport (SDOT) to obtain a better understanding of the collisions data they have in their records with the aim of using the resulting key insights to better predict and ultimately adopt preventive measures to prevent (where possible) accidents in the future. This has come about to solve one of the key issues the SDOT have been attempting to comprehend for decades – to what degree does certain conditions influence the severity of a collision.

The key goal has been to build a model that is able to predict (with high certainty), accident "severity" based on various conditions such as, location, time of day, weather, road condition etc. Such a model's potential benefit to the overall Seattle community would be priceless and may assist to save lives, prevent unnecessary stress, save money and time. No doubt such potential could as a whole make Seattle a better place to live. If the model were to be successful in Seattle, the basic concepts and theories behind the model, could even be adopted in other parts of America as well as in other countries, given the negative impacts of collisions and accidents would be fairly similar around the world, and thus so would be the potential benefits of being able to predict and prevent such accidents, where possible.

DATA

The data that has been analysed in this project has been obtained from a very reliable source, the Seattle Department of Transport (SDOT). The data relates to all collisions from 2004 onwards, that has been provided by the Seattle Police Department (SPD) and recorded by Traffic Records. Therefore nearly 200,000 such collision records over a period of more than 15 years have been incorporated in the thorough analysis performed. Such a rich and abundant data set has allowed for a very deep analysis, and thus increasing the confidence in the resulting insights and recommendations provided with this report.

The target of this data analysis has been to predict with high certainty the severity of a potential collision under certain conditions. For simplicity yet practicality, we have separated each recorded collision into one of two outcomes – injury (high severity) or property damage (low severity). The impact of various conditions including, location, time of day, weather, road condition etc. on the severity of a collision has been well analysed. For instance, the data has assisted us to make evidence based inferences to questions such as;

- Does wet road conditions increase the severity of a collision compared to dry conditions?
- Does driving in the dark increase the severity of a collision compared to the daylight?
- Do certain locations increase the severity of a collision?

METHODOLOGY

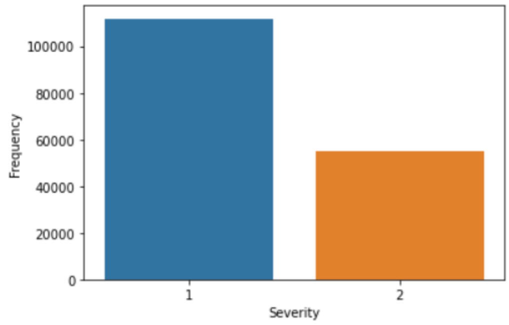
Data Wrangling

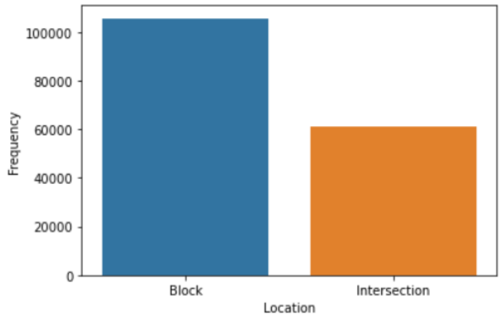
All the data was initially cleansed in order to obtain the most effective data set to commence the analysis on. Firstly missing data inputs that related to the key independent variables/attributes of either the location, road condition, whether or light condition, were deleted. This data was chosen to be deleted because replacing the missing values with any other value (i.e. most frequent categorical value) would make the data invalid and redundant for data analysis purposes. Furthermore once missing data was deleted, more than 85% of the original dataset still remained, thus further providing justification for the approach taken in handling the missing data. Furthermore attributes/independent variables not relevant to our analysis were also removed from our investigation. For instance various SDOT codes and incident keys/IDs were removed. Finally the resulting data set for analysis only included the key attributes; severity, location, weather, road condition, light condition and under influence.

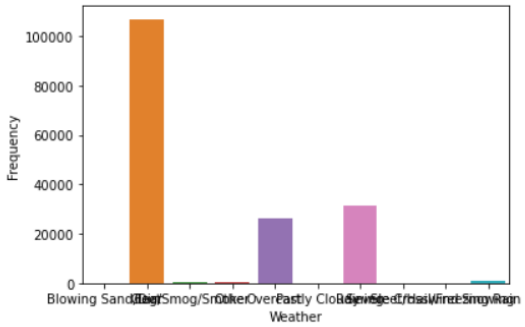
Exploratory Data Analysis

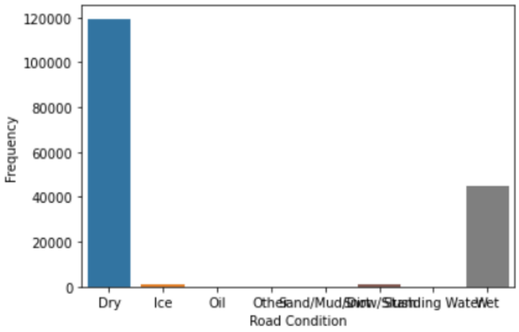
Initial visualisation of the results were as follows:

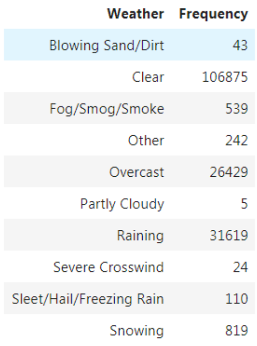
Refer to the Appendix Section of this report for a link to the coding notebook.



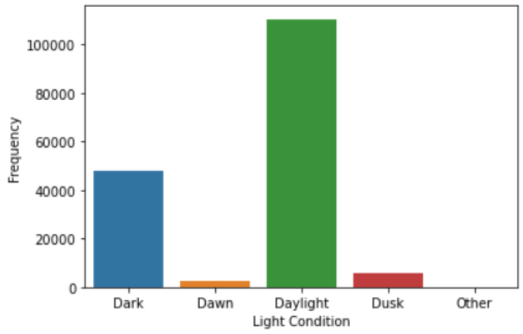


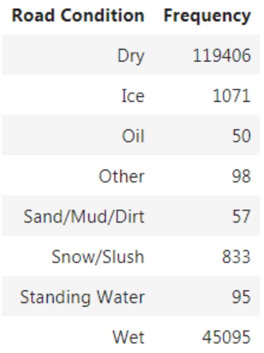






Weather	Frequency
Blowing Sand/Dirt	43
Clear	106875
Fog/Smog/Smoke	539
Other	242
Overcast	26429
Partly Cloudy	5
Raining	31619
Severe Crosswind	24
Sleet/Hail/Freezing Rain	110
Snowing	819





Road Condition	Frequency
Dry	119406
Ice	1071
Oil	50
Other	98
Sand/Mud/Dirt	57
Snow/Slush	833
Standing Water	95
Wet	45095

As can be inferred from the above, some highlights of the initial analysis of the last 15 years of Seattle's collision records were as follows:

- Thankfully only one in every three accidents have resulted in injury (highest severity) - as opposed to property damage (which is considered less severe)
- Approximately one in every three accidents occurred at intersections
- Surprisingly nearly two thirds of all accidents occurred during clear weather conditions and good light conditions (i.e. day time)
- Yet again surprisingly 70% of all accidents occurred when the road conditions were perfect and dry

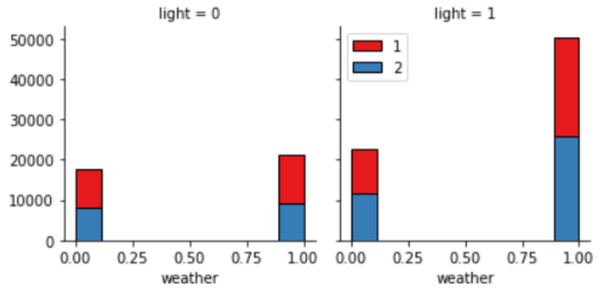
Following the initial analysis of the overall trends of collisions as per above, the data was then further categorised, normalised and transformed to uncover deeper trends/correlations. This was done with the ultimate aim of transforming the data to assist in developing a model that would predict the severity of a future collision, which was the ultimate goal. Please refer to the *Appendix Section* of this report for a table of original verses transformed data variables. For instance the 'weather' attribute data was transformed to either 0 (clear) or 1 (not clear).

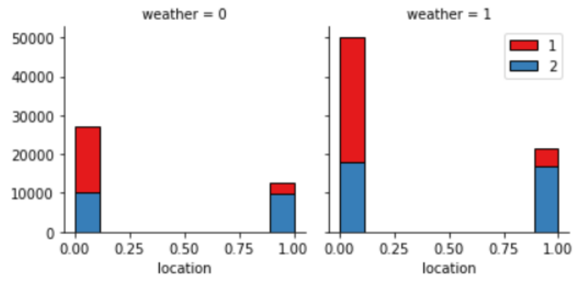
The correlations identified (which are further discussed in the *Results Section* of this report below), assisted to develop the most suitable model. After a thorough analysis of the various potential models via incorporating a training dataset, the Decision Tree Classifier Model, was chosen as the best model to predict the severity of a collision for the independent variables analysed. Given a certain combination of independent variables (i.e. bad weather, poor road conditions, day time, not intoxicated), it was able to predict with high certainty on the testing data set, the resulting severity of a collision – i.e. personal injury (high severity) or only property damage (low severity). This model's testing evaluation results were the best, with a high accuracy score of nearly 70%. Please see the coding notebook for detailed model training, testing and evaluation procedures and results.

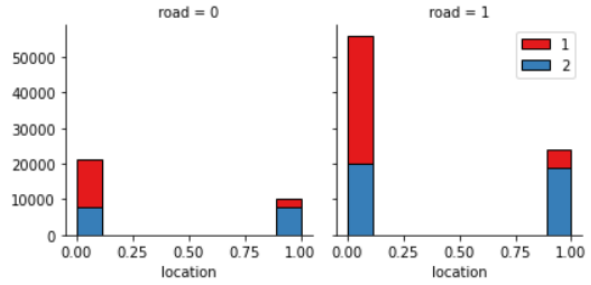
RESULTS

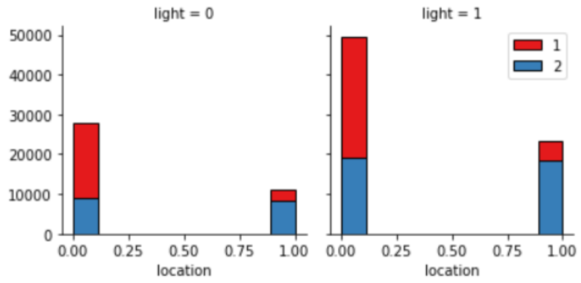
As per the graphical representations below, some of the key findings were:

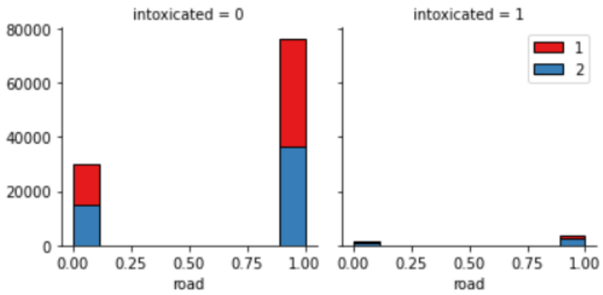
- Regardless of weather, road or light conditions (good or bad), accidents at intersections led to more severe injuries
- On the contrary, again regardless of weather, road or light conditions, accidents at any other locations (other than intersections) resulted in more property damage instead (i.e. less severe)
- Driving intoxicated even on good road conditions led to more severe collisions
- In good or bad weather or road conditions, the time of day (day time or night time) did not adversely affect the severity of collisions

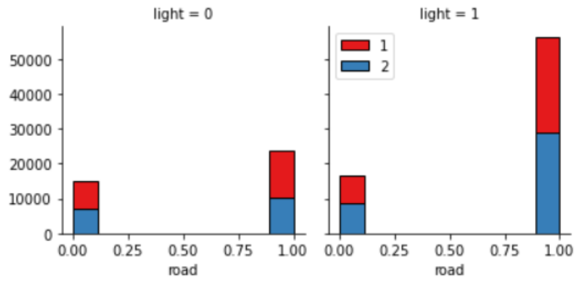












DISCUSSION

Further to the key observations/findings presented in the *Results Section* above, the following recommendations were provided to SDOT, with the aim of minimising accidents in Seattle in the future:

1) Severity of accidents are highest at intersections (regardless of weather, road or light conditions) – therefore it is recommended that SDOT review safety procedures at major intersections in Seattle going forward and ensure
2) Severity of accidents are also highest when intoxicated (regardless of weather, road or light conditions) – therefore it is recommended that SDOT continue to enforce stricter policies and checks to minimise such occurrences in the future
3) It is recommended that SDOT allow us to deploy the current model developed which would allow SDOT to provide safety warnings or via other methods to inform all road users of when certain conditions cause the model to output warnings of potential higher severity of collisions

CONCLUSION

In conclusion, this report has presented some extraordinary findings from more than 15 years of collision data from Seattle. The key findings discussed in the report along with the key recommendations made, if followed by SDOT would benefit the community of Seattle as a whole.

The data scientists behind this project are now actively awaiting the deployment of this model and seeing how it performs on a live basis and if it can be successful in providing Seattle pre-warnings of conditions to be more cautious or avoid if possible.

Following deployment of the model for the next 12-18 months, data will be reviewed again and the model calibrated. The ultimate goal of the data scientists behind this model is to further expand their model to other cities/geographical areas. Consequently it is hoped that one day, this can lead to reduction in the severity of collisions worldwide and lead to the betterment of all human beings.

APPENDIX

1) Coding notebook can be accessed via:

https://github.com/mayura9/testrepo/blob/master/Final%20Capstone%20Project%20(2).ipynb

2) Transformed data as follows:

Original data variables		Transformed data variables	
Severity	Property damage = 1	No change	
	Injury = 2	No change	
Weather	Clear	Clear	= 1
	Raining	Not clear	= 0
	Blowing sand/dirt		
	Fog/smog/smoke		
	Overcast		
	Partly cloudy		
	Severe crosswind		
	Snowing		
	Sleet/hail/freezing rain		
	Other		
Road Conditions	Dry	Dry	= 1
	Wet	Not dry	= 0
	Ice		
	Oil		
	Standing water		
	Snow/ slush		
	Sand/mud/dirt		
	Other		
Light Conditions	Daylight	Daylight	= 1
	Dark	Not daylight	= 0
	Dawn		
	Dusk		
	Other		
Location	Intersection	Intersection	= 1
	All other locations	Not intersection	= 0
Intoxicated	True = 1	No change	
	False = 0	No change	