Capstone Project - Car accident severity (Week 2)

In this week, you will continue working on your capstone project. Please remember by the end of this week, you will need to submit the following:

1. A full report consisting of all of the following components (15 marks):
- Introduction where you discuss the business problem and who would be interested in this project.
- Data where you describe the data that will be used to solve the problem and the source of the data.
- Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.
- Results section where you discuss the results.
- Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.
- Conclusion section where you conclude the report.

1. A link to your Notebook on your Github repository pushed showing your code. (15 marks)
2. Your choice of a presentation or blogpost. (10 marks)

Table of contents

1. Introduction and Business Problem
2. Data
3. Methodology
4. Conclusion

1. Introduction and Business Problem

The initial phase is to understand the project's objective from the business or application perspective. Accidents are nowadays really common.There are many types of collisions, including transportation, traffic, car accidents and so on. Collisions will display at the intersection or mid-block of a segment. This project will focus on predicting the severity of an accident, using the shared data for Seattle city as an example of how to deal with the accidents data. According to 37 different attributes, a supervised machine learning will be used to predict accidents to improve the predictability of the model. Using different sectors like weather, vehcount etc. is really important to avoid accidents. Once the problem confirmed, data understanding, preparation, modeling, evaluation and deployment will be used to anaylse this situation.

2. Data

In the phase of data understanding, dataset should be collected or extracted from various sources such as csv file or SQL database like excel document and background introduction. Then, the attributes

(columns) that will be used to train the machine learning model should be determined. Also, we will assess the condition of chosen attributes by looking for trends, certain patterns, skewed information, correlations, and so on. Using the dataset given from Seattle Police Department, attributes can be used to weigh the severity of an accident are severitycode, weather, vehcount, personcount, lightcond and so on. Data Preparation: The data preparation includes all the required activities to construct the final dataset which will be fed into the modeling tools. Data preparation can be performed multiple times and it includes balancing the labeled data, transformation, filling missing data, and cleaning the dataset.

3. Methodology

In modeling phase, various algorithms and methods can be selected and applied to build the model including supervised machine learning techniques.We can select a single or multiple machine learning models for the same data mining problem. At this phase, stepping back to the data preparation phase is often required. In evaluation phase, before proceeding to the deployment stage, the model needs to be evaluated thoroughly to ensure that the business or the applications' objectives are achieved. Certain metrics can be used for the model evaluation.

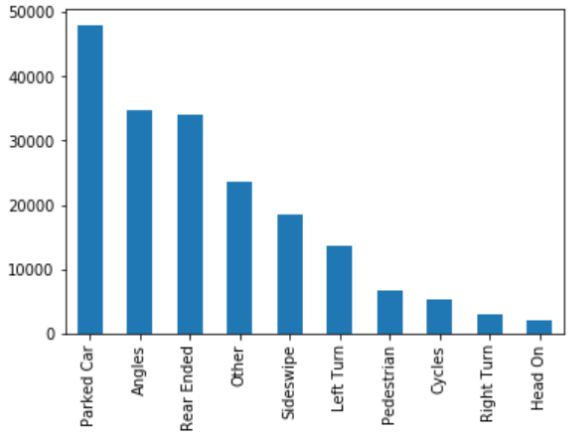
Firstly, I create a formular to download all data that I use. And then, I get the amount of severitycode.

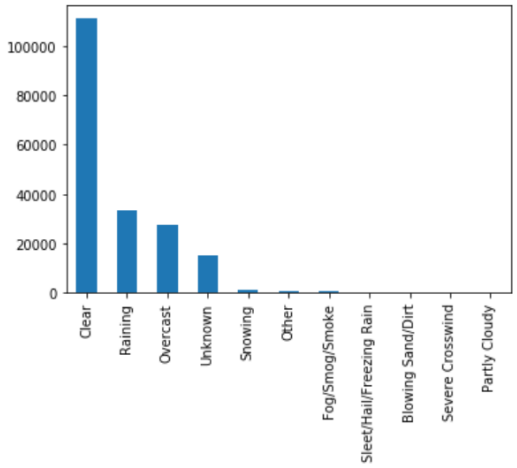
	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND	PI
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Daylight	N
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark - Street Lights On	N
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry	Daylight	N
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry	Daylight	N
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0	...	Wet	Daylight	N

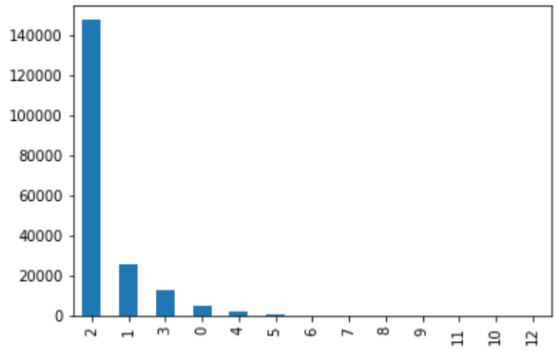
5 rows × 38 columns

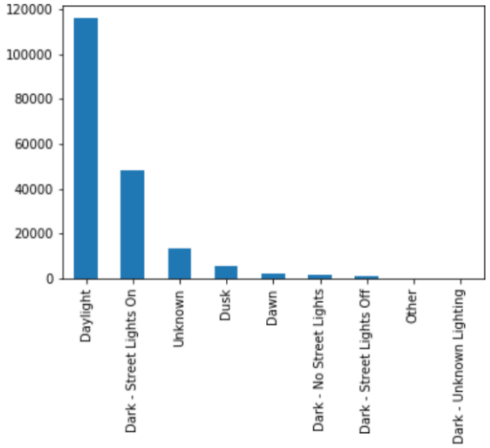
	SEVERITYCODE
1	136485
2	58188

Let's take collisontype, weather, vehcount, lightcond, roadcount for a look.

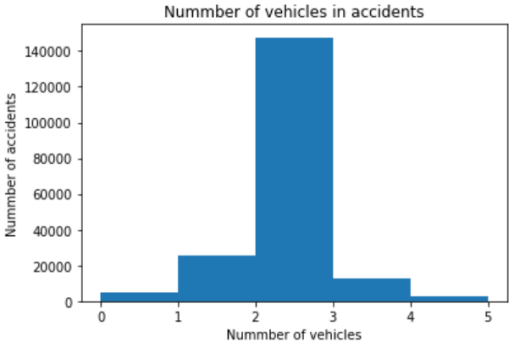


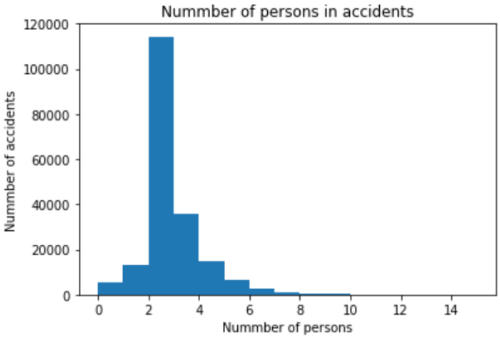






The analysis of the relationship between number of vehicles and people in accidents shows us the most common accidents occur in two to three vehicles between two to four people.

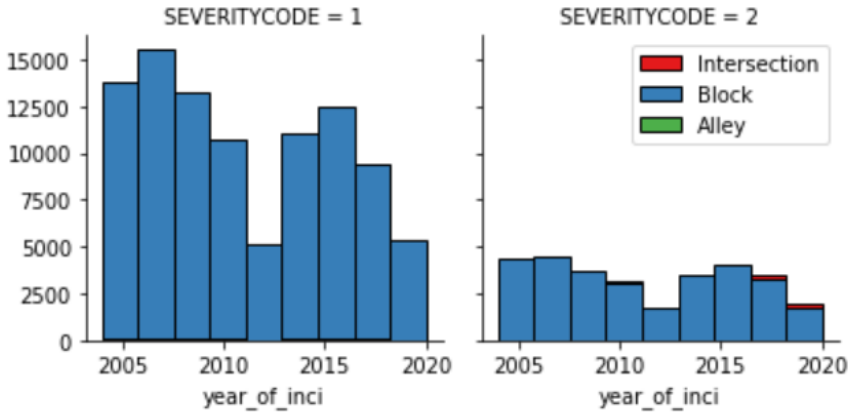


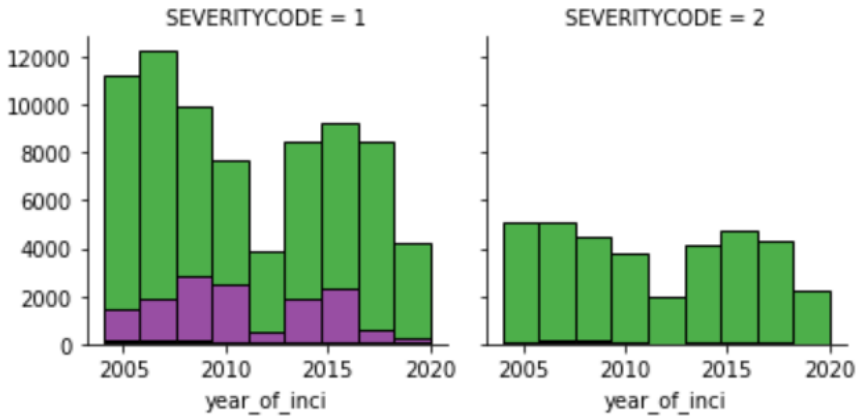


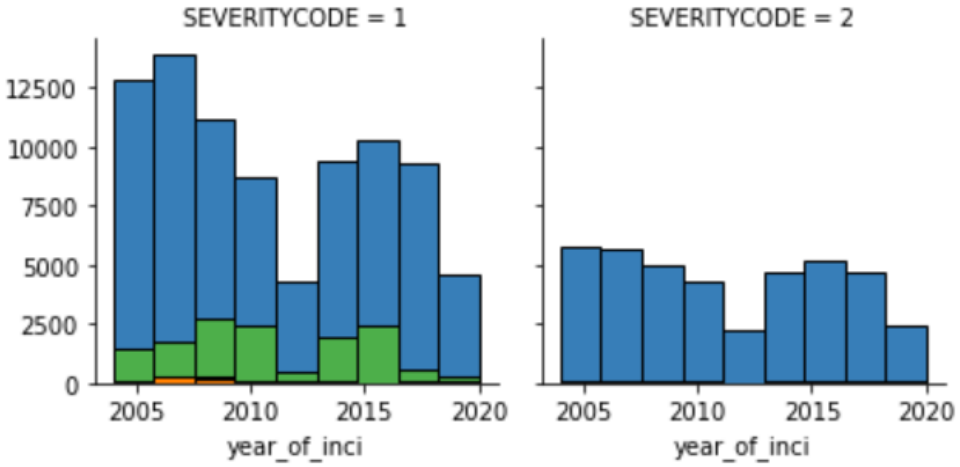
From this form, we can see that the most important attributes and what happen between them.

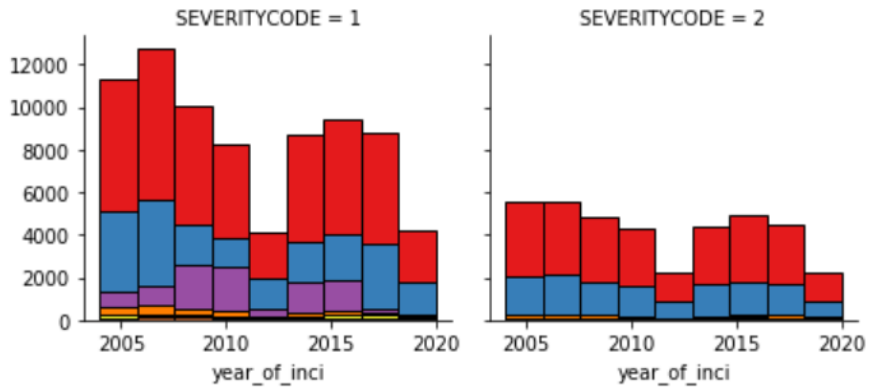
	ADDRTYPE	COLLISIONTYPE	LIGHTCOND	ROADCOND	WEATHER	SEVERITYCODE
0	Intersection	Angles	Daylight	Wet	Overcast	2
1	Block	Sideswipe	Dark - Street Lights On	Wet	Raining	1
2	Block	Parked Car	Daylight	Dry	Overcast	1
3	Block	Other	Daylight	Dry	Clear	1
4	Intersection	Angles	Daylight	Wet	Raining	2

Analysing the severity based on weather, road condition, light condition against each year, when the severitycode is one, there are more accidents in blocks.









4. Conclusion

As the result, the number of people involved in these accidents at any given time. Most accidents included two to four people between two to three cars. It is also important to find out where most accidents taken place, intersections are the most common accident zones.

Using the important attributes like weather, vehcount, lightcond, roadcond, we can find that the amount of car accidents are related to these factors and are influenced much by them. But there are still many missing factors and we need to drop from the data to get good results, like the attribute "SPEED", "EXCEPTRSNDESC" and "PEDROWNOTGRNT" . However, They are also important factors that should be considered especially the speed, despite the data is not entire.

Most accidents occur in clear, dry and bright environments and include two to three vehicle and two to four persons in blocks. The most common collisiontype is parked car and the next is angles, which means that these kind of collison can be avoided by improving carefulness. The severity of these accidents are not serious like crashes. The data results indicate to official department that they should ask drivers to be more careful or take some measures like installing the monitoring in the places of high pedestrian and vehicular traffic or using specific technology in our cars to notice people.This could be helpful to prevent future accidents.