IBM Data Science Capstone Project
Car Accident Severity (Week 1)

Lekha J

september 14 2020

Abstract

It is important to reflect on the stark human and economic cost of road traffic accidents, the first order consequences of which can include property damage, personal injury or death. Key aggravating factors in determining the likelihood and severity of a road traffic accident can include – but are not limited to – current weather, local road/visibility conditions, time of day and individual negligence (e.g. driving under the influence of alcohol/narcotics, driving without due care and attention, driving at excessive speed, etc). In this report I will use Machine Learning techniques to probe publicly-available data for 221,006 road traffic accidents over the past 17 years in the Seattle City Council area in order to determine the relative prevalence and influence of these factors in the severity of road traffic accidents recorded in the City.

1 Business Understanding

Road traffic accidents are a major source of human and economic hardship in most advanced economies, with consequences which can range from minor property/vehicular damage, to major damage, personal injury or death. It is estimated that road traffic accidents cost the United States' economy $\sim$ \$810 billion per year, including costs due to property damage, legal costs and associated medical bills [1]. It is therefore of paramount importance that we understand the factors influencing the likelihood of a road traffic accident occuring at a given location, as well as those which influence the severity of those accidents that do occur.

Intuitively, we might expect that some of the factors which influence the likelihood and severity of a road traffic accident include: the weather, local road contitions (i.e. highways, urban areas or rural roads), time of day (and the presence or absence of street lights), and the number and type of vehicles in the area. Additional factors which may influence the frequency and severity of road traffic accidents include those which can be traced to individual irresponsibility, such as driving under the influence of alcohol/narcotics, driving without due care and attention or driving at excessive speed. While it is intuitive that a combination of these factors may be important, intuition alone cannot determine the relative significance of these factors, which is required in order fully understand the causes of road traffic accidents and devise new strategies to minimise their occurrance and severity.

The target audience for this work will be city planners and emergency service responders: it is hoped that by understanding the factors influencing the frequency and severity of accidents, it will be possible to improve the design of the road network and better prepare for how to deal with accidents when they unfortunatley arise.

2 Data and Proposed Methodology

2.1 Data Understanding

I will use the Cross-Industry Standard Process for Data Mining (CRISP-DM) in order to quantify the impact of these factors on the frequency and severity of car accidents. I will build and test Machine Learning models using data for 221,006 road traffic accidents in the Seattle municipal area between 2004–2020, recorded by Seattle Department of Transport (SDOT) and obtained from the Seattle Open Data Portal (SODP: [2]). **Note that these data were obtained from the Seattle Open Data Portal directly, and that the dataset differs in number of rows and columns from the example dataset provided in the IBM Data Science Capstone introduction. Moreover, the SEVERITYCODE target variable takes on one of five discrete values, rather than the binary options presented in the test dataset**.

The data can be downloaded in Comma Separated Value (CSV) format and read in to a Pandas Dataframe using the Pandas READ_CSV function, and the contents and data types displayed using the HEAD and DTYPES functions. A Jupyter Notebook containing the code for exploring of the dataset, along with data cleaning, model building and model evaluation will be submitted and published on GitHub in next week's submission. Excerpts from the notebook describing the contents of the dataframe are shown in this week's Figs 1 and 2.

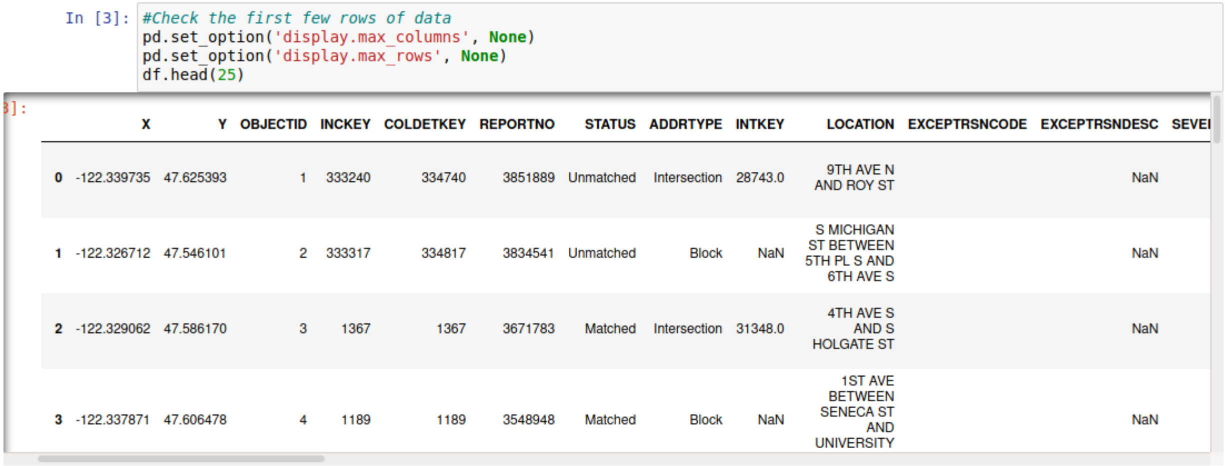
The target/dependent variable is SEVERITYCODE which, in its default form, takes the values 0, 1, 2, 2b or 3. The definitions of these severity codes are provided in the "Attribute Information" metadata which accompany the data release [3] and are given in Table 1.

Severity Code	Meaning
0	Unknown
1	Property Damage
2	Injury (Minor)
2b	Injury (Serious)
3	Fatality

Table 1: SDOT accident severity codes and their definitions

As is clear from Fig. 2 there are 39 candidate predictor variables in this dataset.

In Figs 3 & **??** we begin to explore the properties of the categorical data, in order to get a quick visual overview of which factors commonly coincide with road traffic accidents.



```
In [3]: #Check the first few rows of data
        pd.set_option('display.max_columns', None)
        pd.set_option('display.max_rows', None)
        df.head(25)
```

3]:

	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	LOCATION	EXCEPTRSNCODE	EXCEPTRSNDESC	SEVEI
0	-122.339735	47.625393	1	333240	334740	3851889	Unmatched	Intersection	28743.0	9TH AVE N AND ROY ST		NaN	
1	-122.326712	47.546101	2	333317	334817	3834541	Unmatched	Block	NaN	S MICHIGAN ST BETWEEN 5TH PL S AND 6TH AVE S		NaN	
2	-122.329062	47.586170	3	1367	1367	3671783	Matched	Intersection	31348.0	4TH AVE S AND S HOLGATE ST		NaN	
3	-122.337871	47.606478	4	1189	1189	3548948	Matched	Block	NaN	1ST AVE BETWEEN SENECA ST AND UNIVERSITY		NaN	

Figure 1: Screenshot from Jupyter Notebook showing the output of DF.HEAD(25). Note that only 4 rows and 12 columns are visible on the screenshot; the remaining 21 rows and 28 colums are visible within the Notebook using scroll bars. We see that some columns contain duplicate/redundant data (INCKEY, COLDETKEY), while others contain categorical (ADDRTYPE) or no data (EXCEPTRSNCODE). Cleaning of the data will be essential before meaningful analysis and modelling can be undertaken.

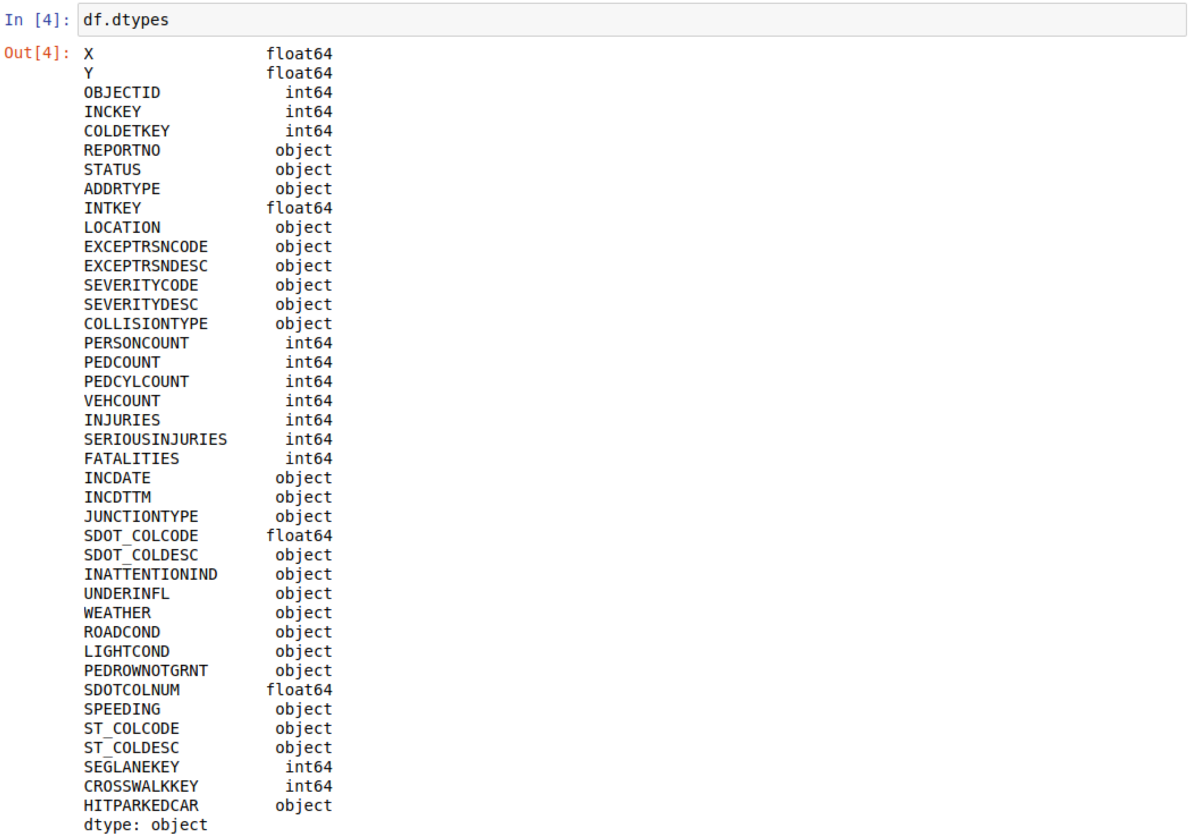
2.2 Data Preparation

In its original form, this dataset is not suitable for quantitative analysis. There are three key reasons for this:

1. The dataset contains columns which are superfluous (i.e. they contain information which is unrelated to the causes or severity of accidents) or are redundant (i.e. they largely replicate information which is already present in other columns). Examples of superfluous columns include OBJECTID, INCKEY and COLDETKEY, which all identify the accident records with respect to other data held by SDOT which are not included in this dataset. Examples of redundant columns include SEVERITYDESC (which provides a textual description of the accompanying SEVERITYCODE) and SDOT_COLCODE/SDOT_COLDESC (which replicate the information that is in the ST_COLCODE column).

2. The dataset contains categorical data, e.g. WEATHER, which takes one of eleven categorical values, or ROADCOND which describes road conditions and takes one of eight categorical values. Machine learning models require numerical data, not categorical data. For this reason it will also be necessary to re-cast the accident severity scale such that it is strictly numerical: 0, 1, 2, 2b, 3 → 0, 1, 2, 3, 4.

3. The dataset contains missing entries, where one or more of the key predictor variables are absent or uninformative (e.g. 6.8% of accidents have “Unknown” listed in the WEATHER column). Including these data entries in the model is likely to increase noise. In some cases, the target variable itself is not in a usable form (4.25% of accidents have SEVERITYCODE “Unknown”).



```
In [4]: df.dtypes

Out[4]: X                    float64
        Y                    float64
        OBJECTID               int64
        INCKEY                 int64
        COLDETKEY              int64
        REPORTNO              object
        STATUS                object
        ADDRTYPE              object
        INTKEY               float64
        LOCATION              object
        EXCEPTRSNCODE         object
        EXCEPTRSNDESC         object
        SEVERITYCODE          object
        SEVERITYDESC          object
        COLLISIONTYPE         object
        PERSONCOUNT            int64
        PEDCOUNT               int64
        PEDCYLCOUNT            int64
        VEHCOUNT               int64
        INJURIES               int64
        SERIOUSINJURIES        int64
        FATALITIES             int64
        INCDATE               object
        INCDTTM               object
        JUNCTIONTYPE          object
        SDOT_COLCODE         float64
        SDOT_COLDESC          object
        INATTENTIONIND        object
        UNDERINFL             object
        WEATHER               object
        ROADCOND              object
        LIGHTCOND             object
        PEDROWNOTGRNT         object
        SDOTCOLNUM           float64
        SPEEDING              object
        ST_COLCODE            object
        ST_COLDESC            object
        SEGLANEKEY             int64
        CROSSWALKKEY           int64
        HITPARKEDCAR          object
        dtype: object
```

Figure 2: Screenshot from Jupyter Notebook showing the output of DF.DTYPES, which lists the data types present in each column of the dataset. We see that some dependent variables are categorical (of type OBJECT), whereas they need to be numerical for most Machine Leaning approaches to work. We will use one-hot encoding to recast each of these categorical variables as a series of numerical variables, with values 0 or 1.

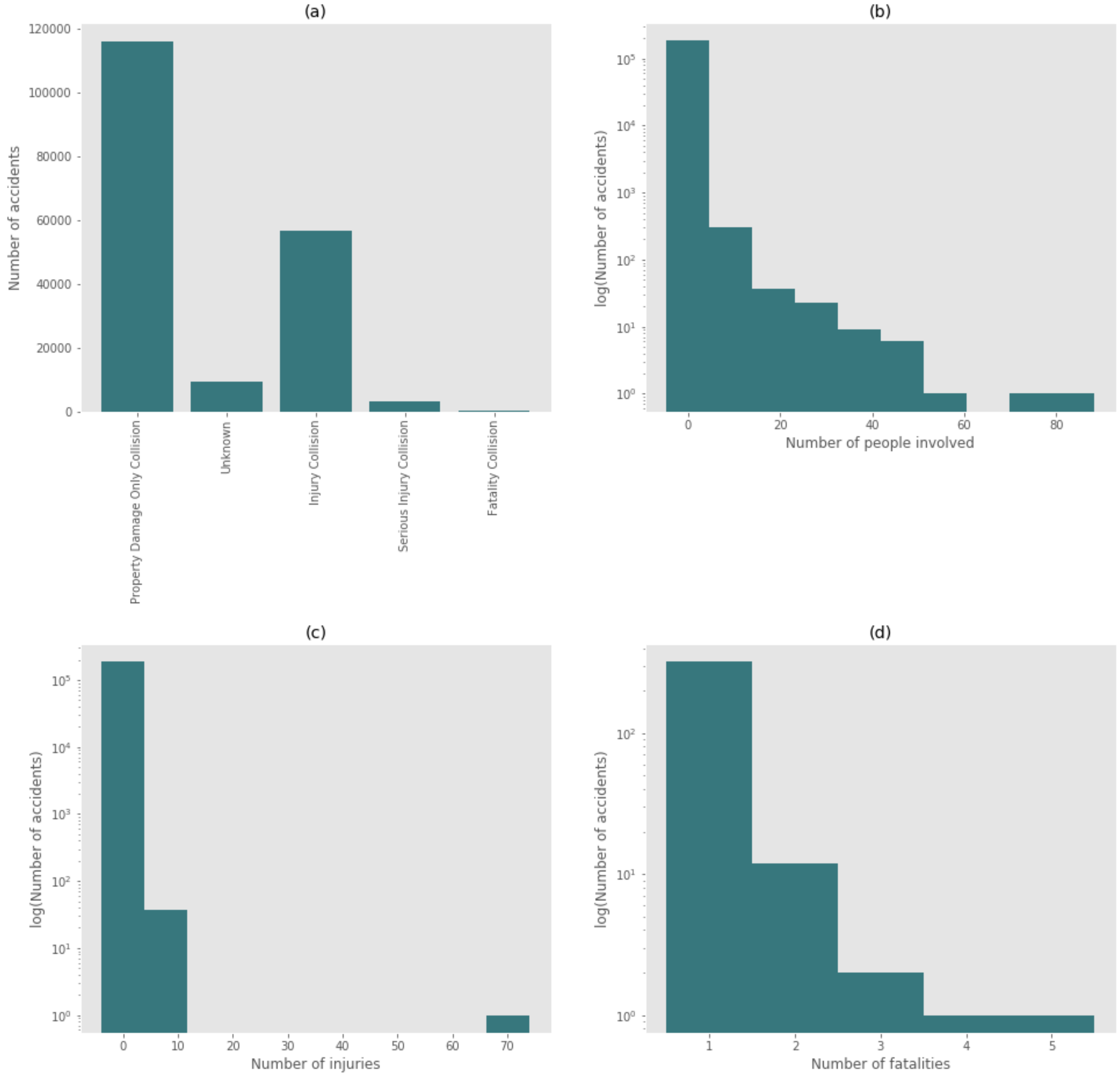


Figure 3: Overview of the severity of accidents in the Seattle municipal area, 2003-2020. *(a):* Of the road traffic accidents in the dataset we see that nearly two-thirds (65.6%) involved only property damage. A significant minority (30.3%) involved minor injuries while 1.6% involved serious injuries. Sadly there have been 335 fatal accidents over this period. 9,396 accidents ($\sim$ 5%) have "Unknown" outcomes: these data are therefore not useful in training or testing the model, as the outcome of the accident is the target variable of this work. *(b):* Number of persons involved per accident. The majority of accidents have few participants. *(c):* Number of persons injured per accident. We see that the majority of accidents involve a small number of injuries, however 16 accidents involved injuries to ≥ 10 people, including one accident in which 78 people were injured. *(d):* Number of fatalities per accident. The vast majority of road traffic accidents (99.8%) in the Seattle area have non-fatal outcomes, however there were 335 fatal accidents in the last 16 years, including one accident with five fatalities.

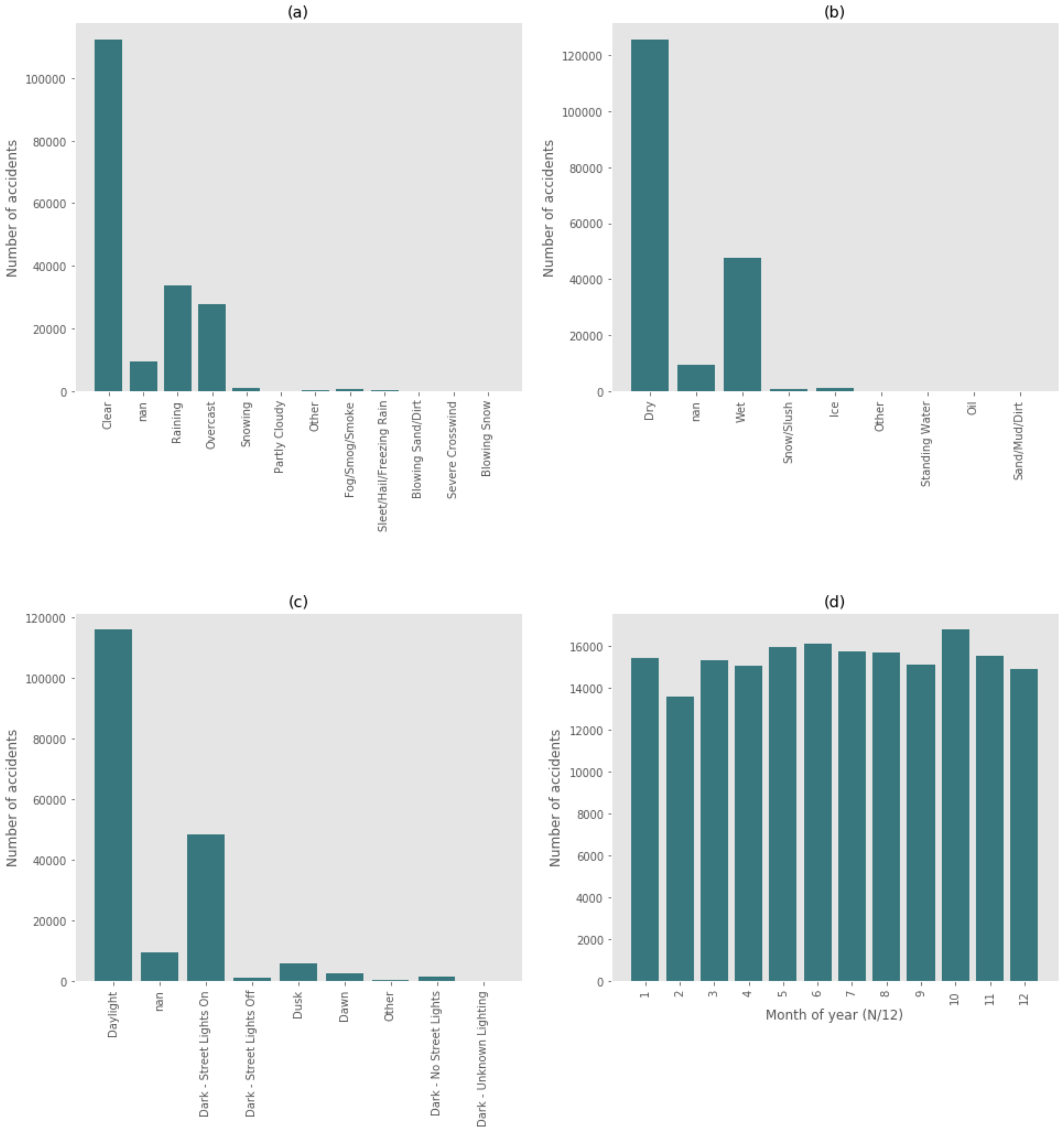


Figure 4: An illustration of the local conditions associated with each accident in the Seattle SDOT accident database, 2004-2020. *(a):* The majority of accidents (75.6%) occurred in clear or overcast (i.e. dry) weather conditions. The remaining 24.4% took place either in severe conditions (such as severe winds) or during periods of precipitation (rain, snow, fog, etc). *(b):* Road conditions at the time of each accident. Clearly the road conditions are related to the prevailing weather at the time (e.g. if there is rain, the roads are likely to be wet), however conditions are not wholly determined by the weather. For instance, 61 accidents occurred on roads where oil was present. *(c):* The light conditions at the time of each accident. 62.6% accidents occured during daylight hours, while 26.2% of accidents occured at night time in areas with streetlights (i.e. urban areas). The remaining 11.2% of accidents include those which happened at dawn/dusk, or on roads with no/faulty streetlights. *(d):* The month of year on which accidents occurred. There is no obvious tendency for accidents to happen at any specific time of the year: the month with the fewest accidents is also the shortest month (February), but otherwise the number of accidents recorded in each month shows no trend throughout the year. The lack of correlation with time in the year is surprising, as one might have expected to see more accidents in the winter months, when the hours of daylight are shortest.

4. The numerical data are imbalanced (there are $\sim 345\times$ as many accidents with SEVERITYCODE=1 as there are accidents with SEVERITYCODE=3) and are not well normalised (e.g. after one-hot encoding many of the categorical variables will be assigned binary values 0/1, whereas the latitude, X and longitude, Y of the accident location are in decimal degrees, and typically cluster around $X = -122.33$, $Y = 47.61$)

In order to use this dataset to build and evaluate a Machine Learning model for predicting accident severity it will be necessary to clean the data using the following standard techniques: (i) discarding rows which are missing crucial data; (ii) discarding columns which contain unnecessary/redundant data; (iii) use of one-hot encoding to create numerical data from categorical variables; (iv) data balancing using downsampling techniques; (v) feature scaling using SCIKITLEARN's STANDARDSCALER function.

2.3 Modelling, Evaluation and Deployment (a forward look)

After cleaning the data I will split the data in to testing (30%) and training (70%) subsamples using TRAIN_TEST_SPLIT, and will then build the following three models for evaluation:

1. **K-Nearest Neighbour (KNN):** this model will attempt to predict the severity of the accident in the test dataset based on the severity of the K accidents whose preceding conditions are most similar in the training dataset.

2. **Decision Tree:** this model will build a decision tree by splitting and branching the data on all the possible values of every attribute in the dataset in order to determine the most predictive features in the dataset. The decision tree will then be used to predict the severity of an accident in the test dataset based on the values of those predictive features.

3. **Support Vector Machine (SVM):** the target variable SEVERITYCODE is not binary in this dataset, and therefore is not suited to logistic regression techniques. Instead, SVM will be used to map the training data to a multi-dimensional space (allowing hyperplanes to be fit which cleanly separate accidents with different SEVERITYCODES), and then these hyperplanes will be used to predict the SEVERITYCODE of accidents in the test dataset, given the values of its independent variables.

Having built these three models I will then evaluate them using the F1 and Jaccard Similarity scores in order to identify the best model.

It is hoped that the best-performing model would then be suitable for deployment and capable of providing useful results to guide the decision-making of town planning authorities and emergency service responders in order to reduce the frequency of accidents and lessen their severity.

References

[1] https://www.pbs.org/newshour/nation/motor-vehicle-crashes-u-s-cost-871-billion-year-federal-study-finds

[2] http://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0

[3] https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf