1.1 Data Understanding

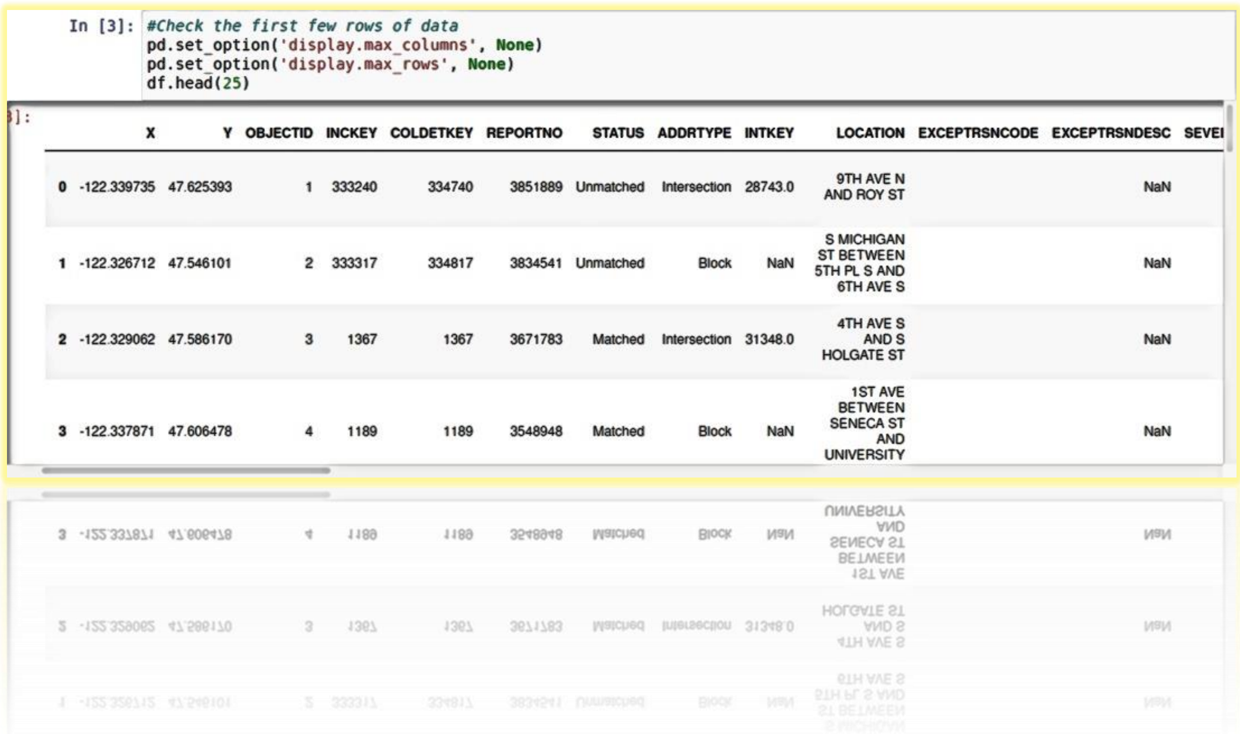
Data Mining process shall be used to determine the output that determines the road accidents and their severity. A Machine learning model shall be used to evaluate 221,006 data sets in Seattle area for past 15 years that is available at Seattle Open Data Portal.

It is noteworthy, the data was obtained from Seattle Open Data Portal directly. The data can be taken in the CSV format and can be read by pandas read_csv format and the content can be printed on Python Jupyter format. The Jupyter notebook will then be used to display the datasets and Heads of the data. Th

The target/dependent variable is SEVERITY 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 and are given in Table 1.

Severity	Impact
0	Not known
1	Property Damage
2	Minor Injury
2b	Serious Injury
3	Death

Table 1: SDOT accident severity codes and their definitions



```
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)
```

	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	

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.

1.2 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 iden- tify the accident records with respect to other data held by SDOT which are not in- cluded in this dataset. Examples of redundant columns include severitydesc (which provides a textual description of the accompanying severitycode) and sdot_col- code/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 cat- egorical 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.

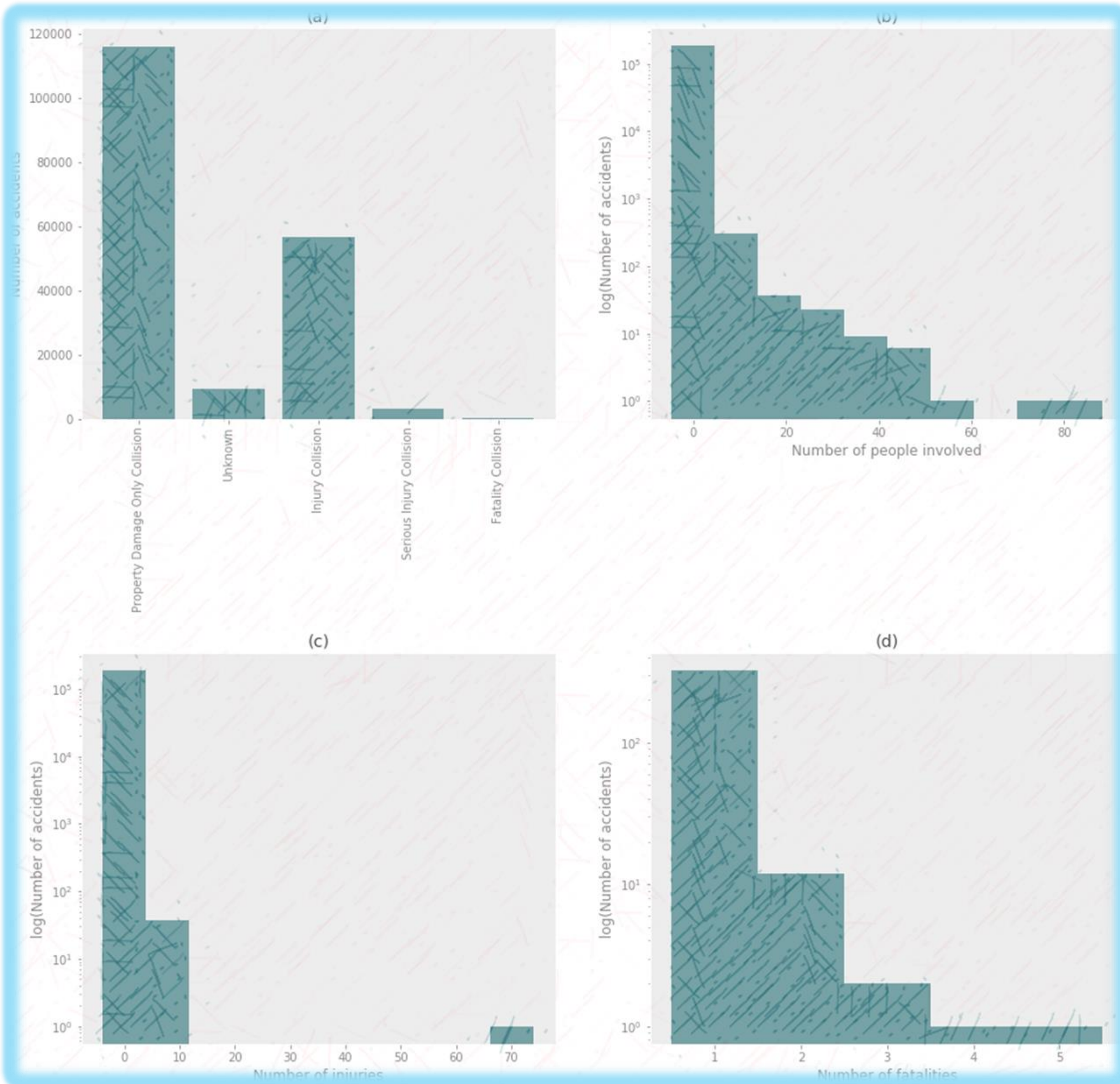
3. The dataset contains missing entries, where one or more of the key predictor variables are absent or uninformative (e.g. 6.8% of accidents have "Unknown" listed in the weather

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 Severity "Unknown")

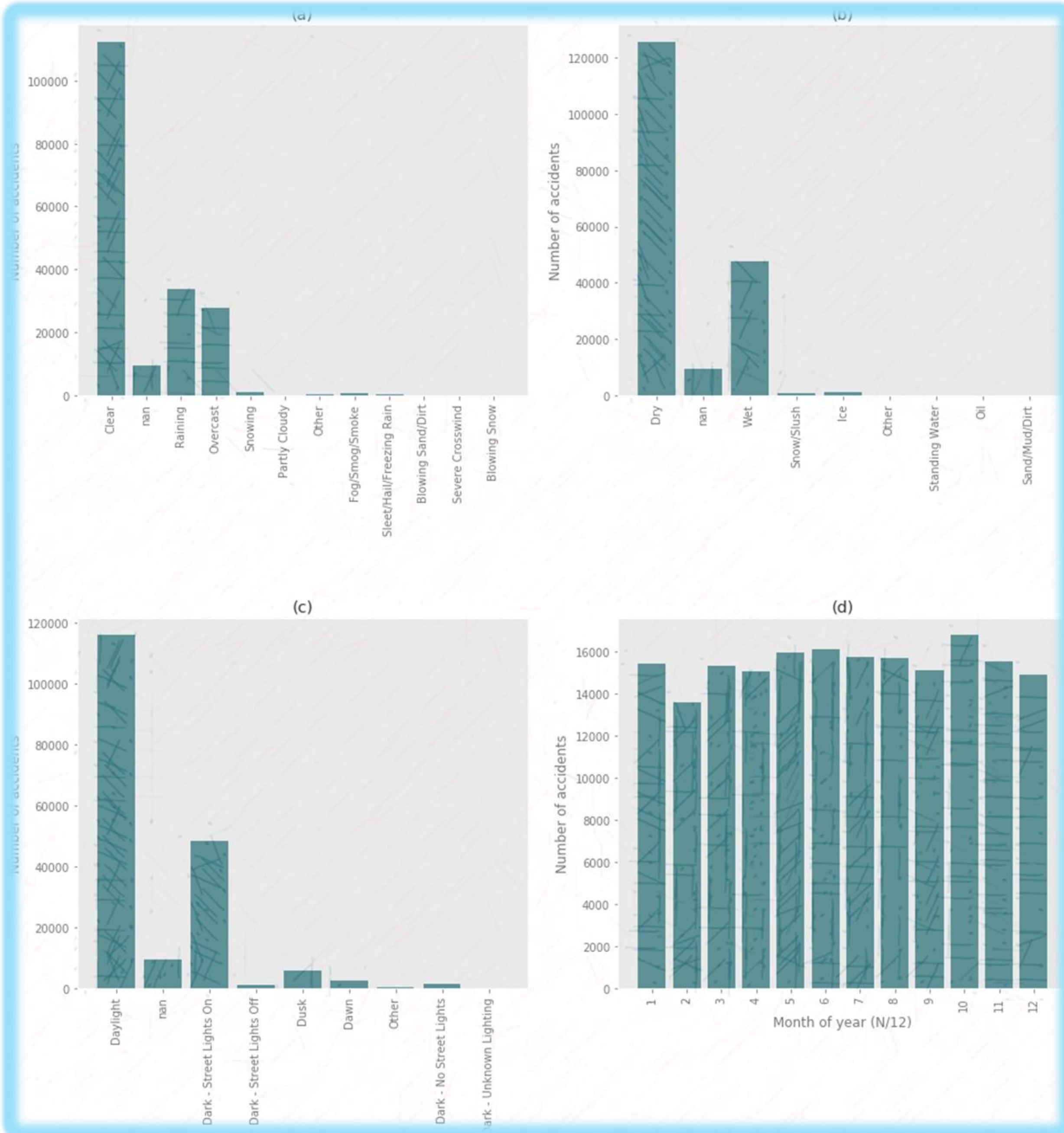
```
In [4]: df.dtypes

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
```

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.



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 (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.



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 345 as many accidents with severi- tycode=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 pre- dicting 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.