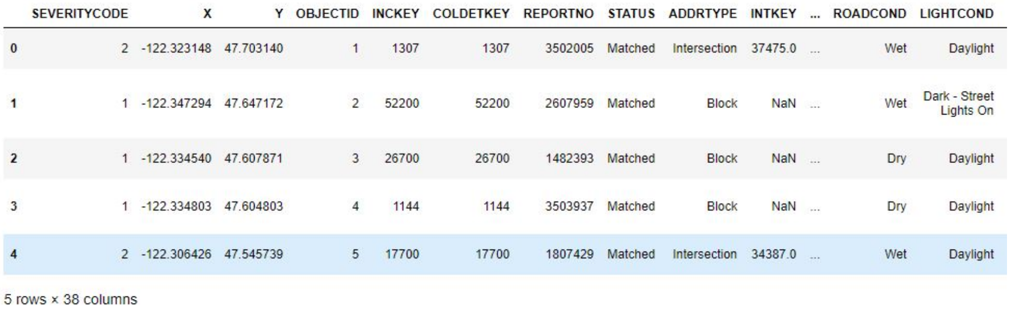
Predicting Accident Severity in Seattle using Weather, Road conditions and Light conditions.

Elijah Chipato

September 6, 2020

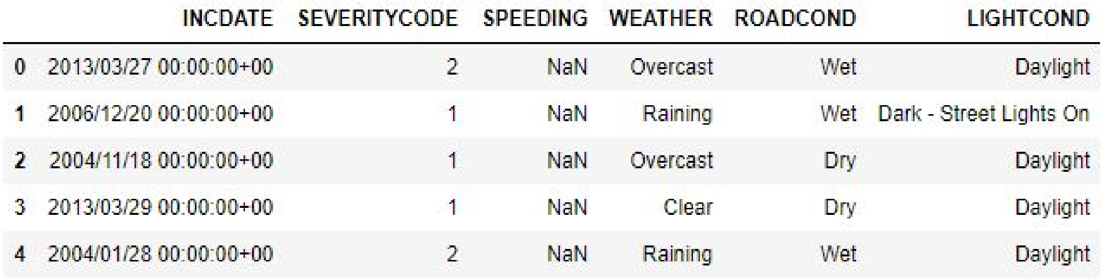
1 Data Section



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

5 rows × 38 columns

Figure 1: First 5 rows of the dataset and a couple of columns.



	INCDATE	SEVERITYCODE	SPEEDING	WEATHER	ROADCOND	LIGHTCOND
0	2013/03/27 00:00:00+00	2	NaN	Overcast	Wet	Daylight
1	2006/12/20 00:00:00+00	1	NaN	Raining	Wet	Dark - Street Lights On
2	2004/11/18 00:00:00+00	1	NaN	Overcast	Dry	Daylight
3	2013/03/29 00:00:00+00	1	NaN	Clear	Dry	Daylight
4	2004/01/28 00:00:00+00	2	NaN	Raining	Wet	Daylight

Figure 2: First 5 rows of the selected feature to be used in this study.

The example dataset for Seattle city is used in this study. Figure 1 shows the first 5 rows of the original dataset which has about 38 features/columns. However, in this study only a couple of features will be used for the modelling. These features are shown on Figure 2 below. The severity code feature will be used as the label of the dataset to train the models. In Figure 2 there are columns of incident date (INCDATE) and speeding (SPEEDING), these will be used during feature engineering to assess whether the day of the week is important in predicting accident severity. If observed that it has no impact these two will be dropped. To assess the quality of the data a heat-map is used to reveal the location or concentration of null values in our dataset. Figure 3 shows a heatmap which reveals that the speeding column has the most null values in the selected features for modelling. WEATHER, ROADCOND and LIGHTCOND also have similar number of null values. For the speeding column the attributes are Y for yes and NaN for No meaning the vehicle was not speeding. The NaN values are replaced by N to represent No. The null values for the rest of the features namely WEATHER, ROADCOND and LIGHTCOND are dropped. Figure 4 shows the new heatmap after dropping all null values and it shows that the datset is now clean and free of null values.

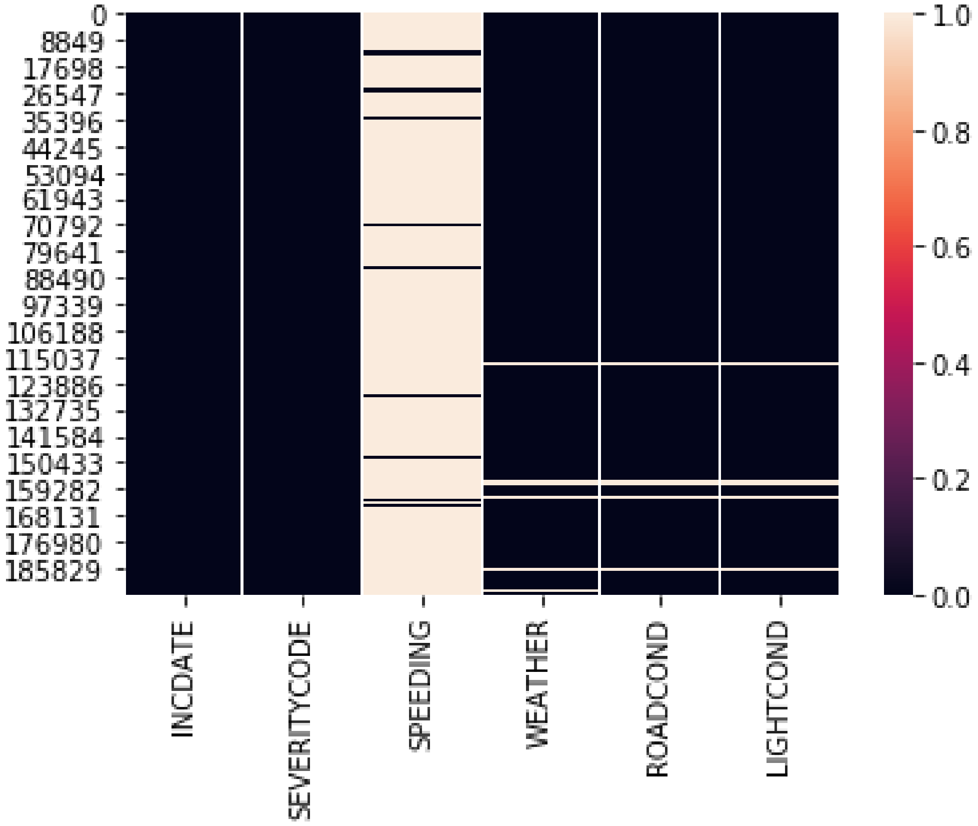


Figure 3: Heatmap showing null values in the dataset.

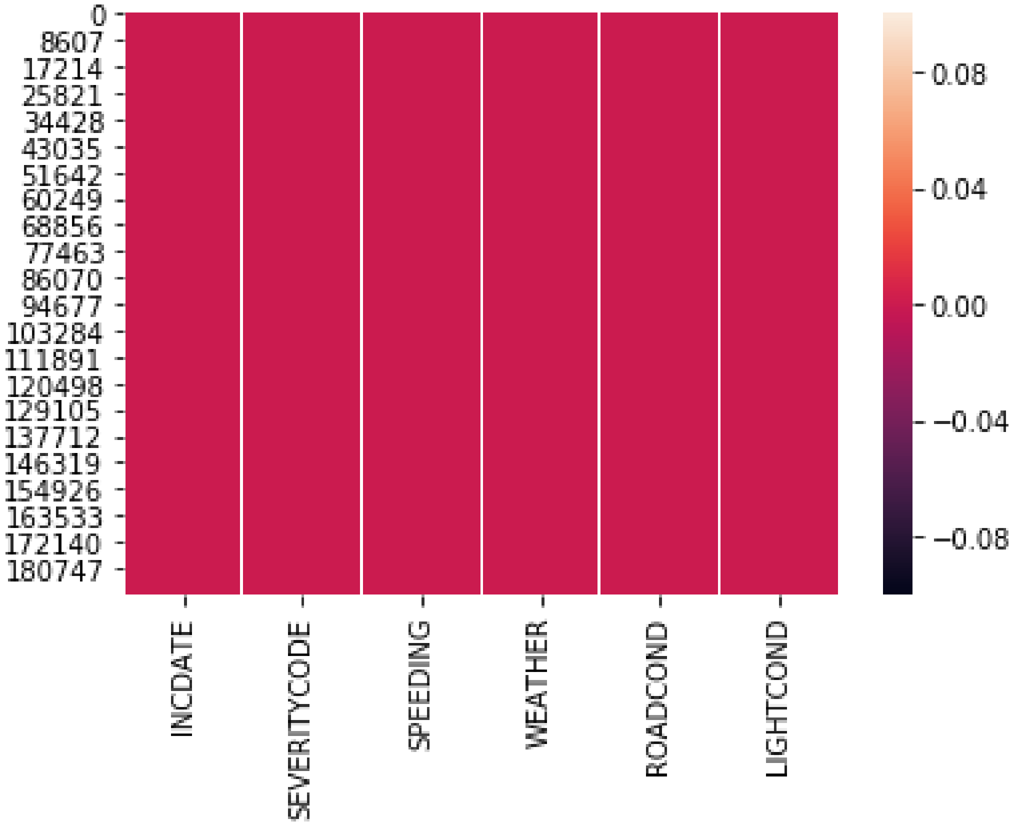


Figure 4: Heatmap showing absence of null values in the dataset.