2. Data Wrangling / Data Pre-processing:

2.1 Data Sources:

I am using the dataset provided by the Seattle Police Department (SPD) and recorded by the Seattle Department of Transportation (SDOT)(made available to me by Coursera). It contains all the collision records and was updated weekly. This dataset contains a total of 1,94,673 records and 37 attributes. To get detailed the detailed information and the metadata of this dataset, you can contact SDOT Traffic Management Division, Traffic Records Group.

2.2 Data Cleaning & Features Selection:

There were a lot of problems with the existing dataset. To start with, there were a lot of missing values and they were represented in 2 forms (i.e. NaN values, string values with 'Unknown' Label).
Example: ROADCOND attribute has 5012 missing values (NaN Type), but it also has 15078 values with 'Unknown' label.

```
In [9]: df.ROADCOND.value_counts().to_frame()
Out[9]:
```

	ROADCOND
Dry	124510
Wet	47474
Unknown	15078
Ice	1209
Snow/Slush	1004
Other	132
Standing Water	115
Sand/Mud/Dirt	75
Oil	64

```
In [10]: df.ROADCOND.isnull().sum()
Out[10]: 5012
```

Same was the case for all other attributes. So, the first step I took was to convert all the missing values into NaN type so that it will be easy for me while modelling the dataset. Also, I renamed 'X' and 'Y' attribute to 'LONGITUDE' and 'LATITUDE' respectively to make it more meaningful.

After fixing these problems, now it was time for Dimensionality Reduction. There were a lot of attributes which needs to be removed from data frame like,

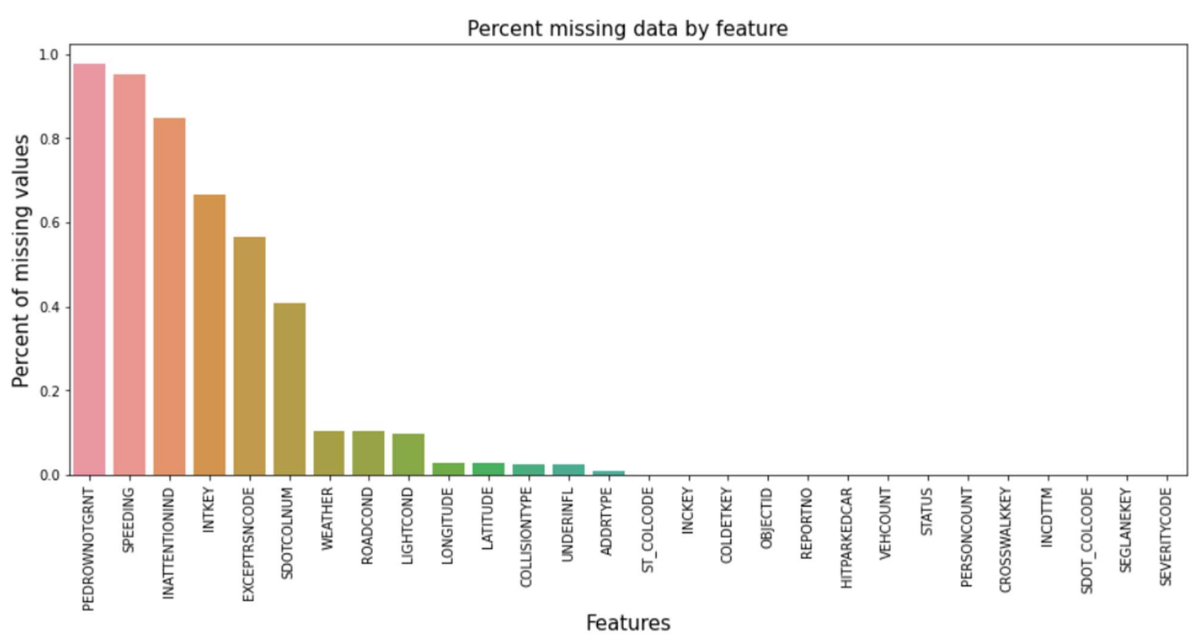
- Handling attributes which were giving the same information repeatedly.
- Handling attributes which have a significant amount of missing data.

Attribute like ‘LOCATION’ is redundant because this information is already available to us in the form of latitude and longitude. So ‘LOCATION’ needs to be dropped. Same was the case for the following attributes: SEVERITYCODE.1, SEVERITYDESC, PEDCOUNT, PEDCYLCOUNT, INCDATE, JUNCTIONTYPE, EXCEPTRSNDESC, SDOT_COLDESC. Hence all these attributes have been dropped.

If an attribute has a lot of missing data, it can create a bias in the model. In this project, I have neglected all those attributes where missing value is more than 40%. Following are the attributes that I dropped:

- PEDROWNOTGRNT, missing value: 97.6%
- SPEEDING, missing value: 95.2%
- INATTENTIONIND, missing value: 84.7%
- INTKEY, missing value: 66.5%
- EXCEPTRSNCODE, missing value: 56.43%
- SDOTCOLNUM, missing value: 43.7%

PEDROWNOTGRNT	190006	0.976026
SPEEDING	185340	0.952058
INATTENTIONIND	164868	0.846897
INTKEY	129603	0.665747
EXCEPTRSNCODE	109862	0.564341



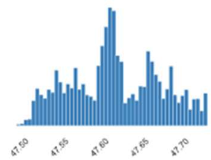
Now LATITUDE & LONGITUDE has 2.7 % missing value. Although I can use imputer to predict the missing values, it is very complicated. These missing values can't be imputed by simple SK Learn Imputer. This is a geographical data and needs a geography-specific library to predict these missing values. But, the missing values % is very small, I decided to drop those missing values.

LATITUDE

Real number ($\mathbb{R}_{\geq 0}$)

MISSING

Distinct	23839	Mean	47.61954252
Distinct (%)	12.6%	Minimum	47.49557292
Missing	5334	Maximum	47.73414158
Missing (%)	2.7%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Memory size	1.5 MiB

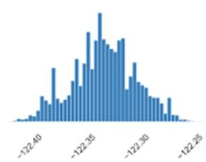


LONGITUDE

Real number ($\mathbb{R}$)

MISSING

Distinct	23563	Mean	-122.3305184
Distinct (%)	12.4%	Minimum	-122.4190911
Missing	5334	Maximum	-122.2389494
Missing (%)	2.7%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Memory size	1.5 MiB



Now there are some attributes which have NO impact on our target variable (i.e. SEVERITYCODE). They are simply post-crash details recorded by the department. These attributes are not responsible to cause any crash and hence are not appropriate in predicting the severity of crashes. Such attributes are:

- OBJECT_ID
- INCKEY
- COLDETKEY
- REPORTNO
- STATUS
- SDOT_COLCODE
- ST_COLCODE
- SEGLANEKEY
- CROSSWALKKEY

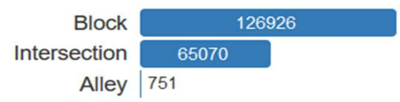
I decided to drop all these attributes.

Handling Missing Values in remaining attributes

a. ADDRTYPE

ADDRTYPE
Categorical

Distinct	3
Distinct (%)	< 0.1%
Missing	1926
Missing (%)	1.0%
Memory size	1.5 MiB



Toggle details

Frequencies | Length

Common Values | Overview | Chart

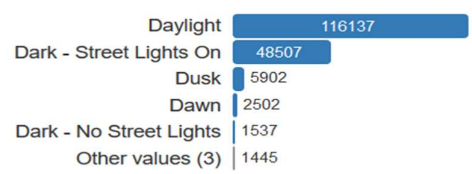
Value	Count	Frequency (%)
Block	126926	65.2%
Intersection	65070	33.4%
Alley	751	0.4%
(Missing)	1926	1.0%

As 65% of crashes happened in Block, there is a high chance that the remaining 1% (missing values) also happened at Block. Therefore, I replaced NaN values by a label value 'Block'.

b. LIGHTCOND

LIGHTCOND
Categorical
MISSING

Distinct	8
Distinct (%)	< 0.1%
Missing	18643
Missing (%)	9.6%
Memory size	1.5 MiB



Toggle details

Frequencies | Length

Common Values | Overview | Chart

Value	Count	Frequency (%)
Daylight	116137	59.7%
Dark - Street Lights On	48507	24.9%
Dusk	5902	3.0%
Dawn	2502	1.3%

Around 60% of accidents happened during daylight, the probability of those missing values being ‘Daylight’ is high. Hence, I replaced NaN values with the label “Daylight”.

c. ROADCOND

Common Values | Overview | Chart

Value	Count	Frequency (%)
Dry	124510	64.0%
Wet	47474	24.4%
Ice	1209	0.6%
Snow/Slush	1004	0.5%
Other	132	0.1%
Standing Water	115	0.1%
Sand/Mud/Dirt	75	< 0.1%
Oil	64	< 0.1%
(Missing)	20090	10.3%

In 64% of accidents, the condition of the road was ‘Dry’. Hence I decided to replace the NaN values with label “Dry”.

d. WEATHER

Common Values | Overview | Chart

Value	Count	Frequency (%)
Clear	111135	57.1%
Raining	33145	17.0%
Overcast	27714	14.2%
Snowing	907	0.5%
Other	832	0.4%
Fog/Smog/Smoke	569	0.3%
Sleet/Hail/Freezing Rain	113	0.1%
Blowing Sand/Dirt	56	< 0.1%
Severe Crosswind	25	< 0.1%
Partly Cloudy	5	< 0.1%
(Missing)	20172	10.4%

Around 57% of the cases happened during a clear day. Hence, I replaced NaN values with label ‘Clear’.