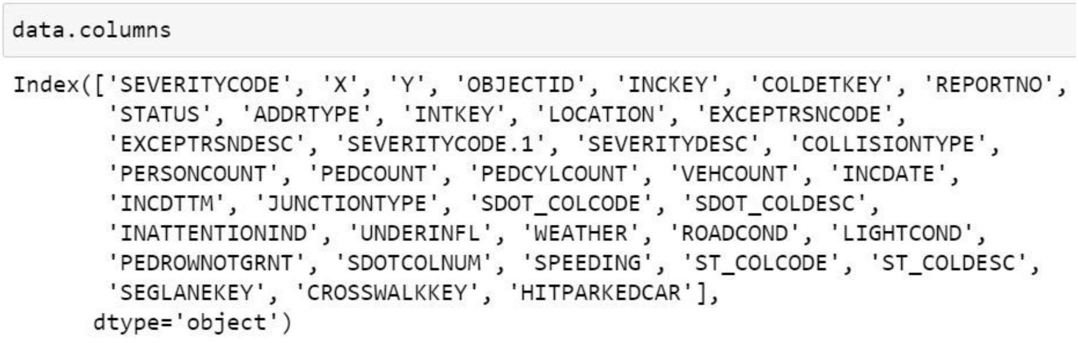
Data Collection and Understanding

The data being used for this ML based prediction problem is the Seattle Accident Dataset, which is available in a .csv format. There are 194673 accident/collision instances have 38 features/columns each, containing recordings of possible factors such as road condition, if the vehicle was speeding, weather, location, junction type etc. One of the columns contains the severity class of the accident – ‘1’ signifying property damage while ‘2’ signifies human injury. This column would serve as the labels in the machine learning binary (class ‘1’ or ‘2’) classification problem.



```
data.columns
```

```
Index(['SEVERITYCODE', 'X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO',
       'STATUS', 'ADDRTYPE', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE',
       'EXCEPTRSNDESC', 'SEVERITYCODE.1', 'SEVERITYDESC', 'COLLISIONTYPE',
       'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INCDATE',
       'INCDTTM', 'JUNCTIONTYPE', 'SDOT_COLCODE', 'SDOT_COLDESC',
       'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND',
       'PEDROWNOTGRNT', 'SDOTCOLNUM', 'SPEEDING', 'ST_COLCODE', 'ST_COLDESC',
       'SEGLANEKEY', 'CROSSWALKKEY', 'HITPARKEDCAR'],
      dtype='object')
```

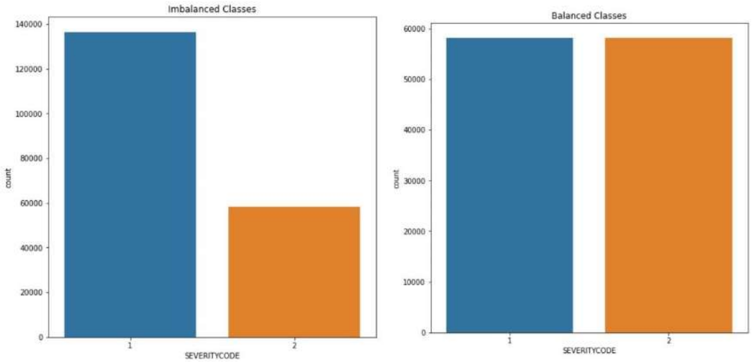
The different columns for each collision instance

Data Preprocessing – Cleaning and Preparation

Not all of the features mentioned above are useful in the classification problem. The data available in each of the column may also not be present in desired numerical format which can be fed in as input in any classification algorithm. Hence, there is a need for preprocessing of data and feature selection based.

1. **Balancing the Imbalanced Dataset**

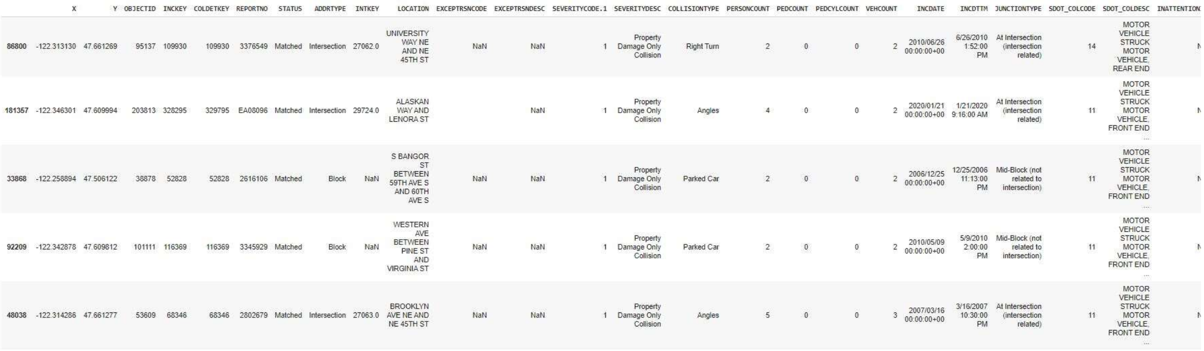
   As the two target classes in the dataset are imbalanced



Dataset before and after down-sampling

2. Data Cleaning

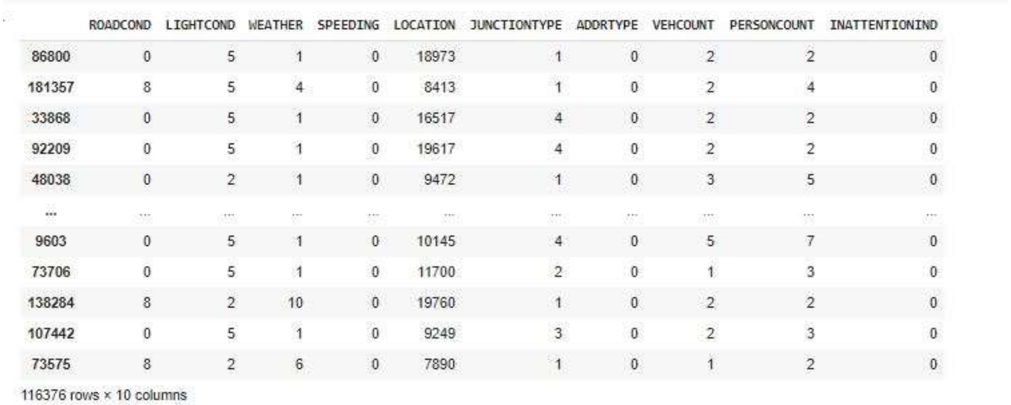
It can be observed that the data has a lot of text and categorical data. These are not suitable to be fed as inputs to any ML classification algorithm.



	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	LOCATION	EXCEPTRSNCODE	EXCEPTRSNDESC	SEVERITYCODE.1	SEVERITYDESC	COLLISIONTYPE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	INCDATE	INCDTTM	JUNCTIONTYPE	SDOT_COLCODE	SDOT_COLDESC	INATTENTION
86800	-122.313130	47.661269	95137	109930	109930	3376549	Matched	Intersection	27062.0	UNIVERSITY WAY NE AND NE 45TH ST	NaN	NaN	1	Property Damage Only Collision	Right Turn	2	0	0	2	2010/06/26 00:00:00+00	6/26/2010 1:52:00 PM	At Intersection (intersection related)	14	MOTOR VEHICLE STRUCK MOTOR VEHICLE, REAR END	N
181357	-122.346301	47.609994	203813	328295	329795	EA08096	Matched	Intersection	29724.0	ALASKAN WAY AND LENORA ST		NaN	1	Property Damage Only Collision	Angles	4	0	0	2	2020/01/21 00:00:00+00	1/21/2020 9:16:00 AM	At Intersection (intersection related)	11	MOTOR VEHICLE STRUCK MOTOR VEHICLE, FRONT END ...	N
33868	-122.258894	47.506122	38878	52828	52828	2616106	Matched	Block	NaN	S BANGOR ST BETWEEN 59TH AVE S AND 60TH AVE S	NaN	NaN	1	Property Damage Only Collision	Parked Car	2	0	0	2	2006/12/25 00:00:00+00	12/25/2006 11:13:00 PM	Mid-Block (not related to intersection)	11	MOTOR VEHICLE STRUCK MOTOR VEHICLE, FRONT END ...	N
92209	-122.342878	47.609812	101111	116369	116369	3345929	Matched	Block	NaN	WESTERN AVE BETWEEN PINE ST AND VIRGINIA ST	NaN	NaN	1	Property Damage Only Collision	Parked Car	2	0	0	2	2010/05/09 00:00:00+00	5/9/2010 2:00:00 PM	Mid-Block (not related to intersection)	11	MOTOR VEHICLE STRUCK MOTOR VEHICLE, FRONT END ...	N
48038	-122.314286	47.661277	53609	68346	68346	2802679	Matched	Intersection	27063.0	BROOKLYN AVE NE AND NE 45TH ST	NaN	NaN	1	Property Damage Only Collision	Angles	5	0	0	3	2007/03/16 00:00:00+00	3/16/2007 10:30:00 PM	At Intersection (intersection related)	11	MOTOR VEHICLE STRUCK MOTOR VEHICLE, FRONT END ...	N

i) There are also a lot of 'NaN' values under most columns. Replacing them with suitable values.
ii) The categorical values are label encoded.
iii) Picking the relevant columns for the classification problem as many of the columns may not be very useful.

3. Final Preprocessed Data



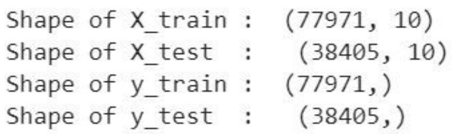
	ROADCOND	LIGHTCOND	WEATHER	SPEEDING	LOCATION	JUNCTIONTYPE	ADDRTYPE	VEHCOUNT	PERSONCOUNT	INATTENTIONIND
86800	0	5	1	0	18973	1	0	2	2	0
181357	8	5	4	0	8413	1	0	2	4	0
33868	0	5	1	0	16517	4	0	2	2	0
92209	0	5	1	0	19617	4	0	2	2	0
48038	0	2	1	0	9472	1	0	3	5	0
...	...	...	...	...	...	...	...	...	...	...
9603	0	5	1	0	10145	4	0	5	7	0
73706	0	5	1	0	11700	2	0	1	3	0
138284	8	2	10	0	19760	1	0	2	2	0
107442	0	5	1	0	9249	3	0	2	3	0
73575	8	2	6	0	7890	1	0	1	2	0

116376 rows × 10 columns

The data is now cleaned and all the values are numerical and hence suitable to be fed as input to ML algorithms.

4. Train-test split of data

Splitting the data into train and test data. 33% of the data is taken to be test data while 66% is used for training the algorithm.



```
Shape of X_train :  (77971, 10)
Shape of X_test  :   (38405, 10)
Shape of y_train :  (77971,)
Shape of y_test  :   (38405,)
```

This data can now be used to input to various ML algorithms and check for model performance based on various metrics. The model with best metrics would be chosen for the classification and prediction.