1 Data collection and understanding

1.1 Data source

We used shared data for Seattle city as basis to deal with the accidents data. At first glance at the CSV file, we could see what type of data we have with us. The target label for the data set is Severity, which describes fatality of an accident. The remaining columns have different types of attributes. Also noticed that the data had some unbalanced attributes which need to be attended in next steps.

We also used the collisions meta data available for the years to understand the nature of all attributes. Having about 1.95L data observations, we could plan for split of the observations that could be used to train and test the prospective model.

1.2 Data understanding

The dataset basics are provided as follows;

Title: Collisions—All Years
Abstract: All collisions provided by Traffic Records.
Description: This includes all types of collisions. Collisions are displayed at the intersection or mid-block of a segment in the Annexure.
Timeframe: 2004 to Present.
Keyword(s): SDOT, Seattle, Transportation, Accidents, Bicycle, Car, Collisions, Pedestrian, Traffic, Vehicle
Types: The data is a mix of numerical and categorical types.

The data set provides labelled data for severity of accident. It shows a dual class categorical type of variable. The attributes (36 columns) convey information mainly about;

- the incident: such as identification no., location coordinates, date, time etc.
- the collision: such as code, type, description, injuries, fatalities etc.
- the impact: such as count of pedestrians, cyclists, vehicles involved etc.
- preconditions: such as inattention, influence of drugs, road condition, weather, speeding etc.

Attributes are almost complete with the information such as name, data type, length and description as shown in next section. State Collision Code Dictionary comprising about 85 codes with descriptions is also provided in supplement.

With the given dataset, severity code is identified as the target variable (labelled or dependent) while rest of the fields are noted as independent variables or the attributes. The case objective along with given data does qualify it as a classification problem of the supervised machine learning. All columns that could influence the cause and impact of an accident need to be selected for training and testing the model.

2 Data preparation

2.1 Basic insight of dataset

After reading data into Pandas data frame, it became a good start to explore the dataset. Following ways are followed to obtain essential insights of the data to help better understand the dataset;

Columns:
It provides list of columns that exist in the dataset.

Data types:
This step is to know the variety of types viz. object, float, int, bool and datetime64. In order to better learn about each attribute, it is necessary to know the data type of each column, which were identified using Python as in screen shot below;

```
[4]: df.dtypes

[4]: SEVERITYCODE       int64
     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.1      int64
     SEVERITYDESC       object
     COLLISIONTYPE      object
     PERSONCOUNT         int64
     PEDCOUNT            int64
     PEDCYLCOUNT         int64
     VEHCOUNT            int64
     INCDATE            object
     INCDTTM            object
     JUNCTIONTYPE       object
     SDOT_COLCODE        int64
     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
```

Data description:

We could get statistical summary, such as count, unique value, column mean value, column standard deviation, etc of each column. It provides various summary statistics, excluding NaN (Not a Number) values.

```
[6]: df.describe(include='all')
```

[6]:

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND	PEDROWNO
count	194673.000000	189339.000000	189339.000000	194673.000000	194673.000000	194673.000000	194673	194673	192747	65070.000000	...	189661	189503	
unique	NaN	NaN	NaN	NaN	NaN	NaN	194670	2	3	NaN	...	9	9	
top	NaN	NaN	NaN	NaN	NaN	NaN	1782439	Matched	Block	NaN	...	Dry	Daylight	
freq	NaN	NaN	NaN	NaN	NaN	NaN	2	189786	126926	NaN	...	124510	116137	
mean	1.298901	-122.330518	47.619543	108479.364930	141091.456350	141298.811381	NaN	NaN	NaN	37558.450576	...	NaN	NaN	
std	0.457778	0.029976	0.056157	62649.722558	86634.402737	86986.542110	NaN	NaN	NaN	51745.990273	...	NaN	NaN	
min	1.000000	-122.419091	47.495573	1.000000	1001.000000	1001.000000	NaN	NaN	NaN	23807.000000	...	NaN	NaN	
25%	1.000000	-122.348673	47.575956	54267.000000	70383.000000	70383.000000	NaN	NaN	NaN	28667.000000	...	NaN	NaN	
50%	1.000000	-122.330224	47.615369	106912.000000	123363.000000	123363.000000	NaN	NaN	NaN	29973.000000	...	NaN	NaN	
75%	2.000000	-122.311937	47.663664	162272.000000	203319.000000	203459.000000	NaN	NaN	NaN	33973.000000	...	NaN	NaN	
max	2.000000	-122.238949	47.734142	219547.000000	331454.000000	332954.000000	NaN	NaN	NaN	757580.000000	...	NaN	NaN	

11 rows × 38 columns

Info:

Similar to above, but it provides a concise summary of the DataFrame as shown below;

```
RangeIndex: 194673 entries, 0 to 194672
Data columns (total 38 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   SEVERITYCODE    194673 non-null  int64
 1   X               189339 non-null  float64
 2   Y               189339 non-null  float64
 3   OBJECTID        194673 non-null  int64
 4   INCKEY          194673 non-null  int64
 5   COLDETKEY       194673 non-null  int64
 6   REPORTNO        194673 non-null  object
 7   STATUS          194673 non-null  object
 8   ADDRTYPE        192747 non-null  object
 9   INTKEY          65070 non-null   float64
 10  LOCATION        191996 non-null  object
 11  EXCEPTRSNCODE   84811 non-null   object
 12  EXCEPTRSNDESC   5638 non-null    object
 13  SEVERITYCODE.1  194673 non-null  int64
 14  SEVERITYDESC    194673 non-null  object
 15  COLLISIONTYPE   189769 non-null  object
 16  PERSONCOUNT     194673 non-null  int64
 17  PEDCOUNT        194673 non-null  int64
 18  PEDCYLCOUNT     194673 non-null  int64
 19  VEHCOUNT        194673 non-null  int64
 20  INCDATE         194673 non-null  object
 21  INCDTTM         194673 non-null  object
 22  JUNCTIONTYPE    188344 non-null  object
 23  SDOT_COLCODE    194673 non-null  int64
 24  SDOT_COLDESC    194673 non-null  object
 25  INATTENTIONIND  29805 non-null   object
 26  UNDERINFL       189789 non-null  object
 27  WEATHER         189592 non-null  object
 28  ROADCOND        189661 non-null  object
 29  LIGHTCOND       189503 non-null  object
 30  PEDROWNOTGRNT   4667 non-null    object
 31  SDOTCOLNUM      114936 non-null  float64
 32  SPEEDING        9333 non-null    object
 33  ST_COLCODE      194655 non-null  object
 34  ST_COLDESC      189769 non-null  object
 35  SEGLANEKEY      194673 non-null  int64
 36  CROSSWALKKEY    194673 non-null  int64
 37  HITPARKEDCAR    194673 non-null  object
dtypes: float64(4), int64(12), object(22)
memory usage: 56.4+ MB
```

2.2Feature selection

In the first screening, it was noticed that some of the attributes were not significant to the cause of or to assess the impact of the severity. So these could be dropped for removing bias during design of the model. Rest of the columns were retained for further analysis.

2.3Cleansing data

In the second step, columns were analysed for missing values. Columns were treated for substituting missing values as shown in the table below;

Sr. No.	Attribute	Data type, length	Description	Wrangling Method	Rationale
1	OBJECTID	OBJECTID	ESRI unique identifier	Dropped	Insignificance
2	X	Longitude	ESRI geometry field	Dropped	Insignificance
3	Y	Latitude	ESRI geometry field	Dropped	Insignificance
4	ADDRTYPE	Text, 12	Collision address type: • Alley • Block • Intersection	Retained	1% missing data, replaced by max. frequency
5	INTKEY	Double	Key that corresponds to the intersection associated with a collision	Dropped	Insignificance
6	LOCATION	Text, 255	Description of the general location of the collision	Dropped	Insignificance
7	EXCEPTRSNCODE	Text, 10		Dropped	Insignificance
8	EXCEPTRSNDESC	Text, 300		Dropped	Insignificance
9	SEVERITYCODE	Text, 100	A code that corresponds to the severity of the collision: • 3—fatality • 2b—serious injury • 2—injury • 1—prop damage • 0—unknown	Retained	Target variable
10	SEVERITYDESC	Text	A detailed description of the severity of the collision	Retained	Target variable
11	COLLISIONTYPE	Text, 300	Collision type	Retained	2.5% missingdata, replaced by max. frequency
12	PERSONCOUNT	Double	The total number of people involved in the collision	Retained	
13	PEDCOUNT	Double	The number of pedestrians involved in the collision. This is entered by the state.	Retained	
14	PEDCYLCOUNT	Double	The number of bicycles involved in the collision. This is entered by the state.	Retained	
15	VEHCOUNT	Double	The number of vehicles involved in the collision. This is entered by the state.	Retained	

Sr. No.	Attribute	Data type, length	Description	Wrangling Method	Rationale
16	INJURIES	Double	The number of total injuries involved in the collision. This is entered by the state.	Retained	
17	SERIOUSINJURIES	Double	The number of serious injuries involved in the collision. This is entered by the state.	Retained	
18	FATALITIES	Double	The number of fatalities involved in the collision. This is entered by the state.	Retained	
19	INCDATE	Date	The date of the incident.	Dropped	Insignificance
20	INCDTTM	Text, 30	The date and time of the incident.	Dropped	Insignificance
21	JUNCTIONTYPE	Text, 300	Category of junction at which collision took place	Retained	3.3% missingdata, replaced by max. frequency
22	SDOT_COLCODE	Text, 10	A code given to the collision by SDOT.	Dropped	Insignificance
23	SDOT_COLDESC	Text, 300	A description of the collision corresponding to the collision code.		
24	INATTENTIONIND	Text, 1	Whether or not collision was due to inattention (Y/N).	Dropped	85% data is missing
25	UNDERINFL	Text, 10	Whether or not a driver involved was under the influence of drugs or alcohol.	Dropped	only 3% observations are influencing
26	WEATHER	Text, 300	A description of the weather conditions during the time of the collision.	Retained	2.6% missingdata, replaced by max. frequency
27	ROADCOND	Text, 300	The condition of the road during the collision.	Retained	2.6% missingdata, replaced by max. frequency
28	LIGHTCOND	Text, 300	The light conditions during the collision.	Retained	2.7% missingdata, replaced by max. frequency
29	PEDROWNOTGRNT	Text, 1	Whether or not the pedestrian right of way was not granted. (Y/N)	Dropped	97.6% data missing
30	SDOTCOLNUM	Text, 10	A number given to the collision by SDOT.	Dropped	Insignificance

Sr. No.	Attribute	Data type, length	Description	Wrangling Method	Rationale
31	SPEEDING	Text, 1	Whether or not speeding was a factor in the collision. (Y/N)	Dropped	only 3% observations are influencing, rest data unavailable
32	ST_COLCODE	Text, 10	A code provided by the state that describes the collision. For more information about these codes, please see the State Collision Code Dictionary.	Dropped	Insignificance
33	ST_COLDESC	Text, 300	A description that corresponds to the state's coding designation.	Retained	2.5% missingdata, replaced by max. frequency
34	SEGLANEKEY	Long	A key for the lane segment in which the collision occurred.	Dropped	Insignificance
35	CROSSWALKKEY	Long	A key for the crosswalk at which the collision occurred.	Dropped	Insignificance
36	HITPARKEDCAR	Text, 1	Whether or not the collision involved hitting a parked car. (Y/N)	Dropped	only 3.7% observations are influencing
37	STATUS	Text, 10	Matched, Unmatched	Dropped	Insignificance
38	REPORTNO	Long	Sr. No. of report for internal purposes	Dropped	Insignificance

2.4Transforming data

The last step in data cleansing would be to check and make sure that all data is in the correct format (int, float, text or other). To use categorical variables for regression analysis, indicator variables (or dummy variable) were used for transforming categorical variables into binary values (0s and 1s). This would make the data ready for next tests of correlation and determining significance. The results are as shown in screen shots below;

```
[15]: df_clean.info(max_cols=157)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 194673 entries, 0 to 194672
Data columns (total 156 columns):
 #   Column                                                                  Non-Null Count   Dtype
---  ------                                                                  --------------   -----
 0   SEVERITYCODE                                                            194673 non-null  int64
 1   SEVERITYDESC                                                            194673 non-null  object
 2   PERSONCOUNT                                                             194673 non-null  int64
 3   PEDCOUNT                                                                194673 non-null  int64
 4   PEDCYLCOUNT                                                             194673 non-null  int64
 5   VEHCOUNT                                                                194673 non-null  int64
 6   ADDRTYPE_Alley                                                          194673 non-null  uint8
 7   ADDRTYPE_Block                                                          194673 non-null  uint8
 8   ADDRTYPE_Intersection                                                   194673 non-null  uint8
 9   COLLISIONTYPE_Angles                                                    194673 non-null  uint8
 10  COLLISIONTYPE_Cycles                                                    194673 non-null  uint8
 11  COLLISIONTYPE_Head On                                                   194673 non-null  uint8
 12  COLLISIONTYPE_Left Turn                                                 194673 non-null  uint8
 13  COLLISIONTYPE_Other                                                     194673 non-null  uint8
 14  COLLISIONTYPE_Parked Car                                                194673 non-null  uint8
 15  COLLISIONTYPE_Pedestrian                                                194673 non-null  uint8
 16  COLLISIONTYPE_Rear Ended                                                194673 non-null  uint8
 17  COLLISIONTYPE_Right Turn                                                194673 non-null  uint8
 18  COLLISIONTYPE_Sideswipe                                                 194673 non-null  uint8
 19  JUNCTIONTYPE_At Intersection (but not related to intersection)          194673 non-null  uint8
 20  JUNCTIONTYPE_At Intersection (intersection related)                     194673 non-null  uint8
 21  JUNCTIONTYPE_Driveway Junction                                          194673 non-null  uint8
 22  JUNCTIONTYPE_Mid-Block (but intersection related)                       194673 non-null  uint8
 23  JUNCTIONTYPE_Mid-Block (not related to intersection)                    194673 non-null  uint8
 24  JUNCTIONTYPE_Ramp Junction                                              194673 non-null  uint8
 25  JUNCTIONTYPE_Unknown                                                    194673 non-null  uint8
 26  SDOT_COLDESC_DRIVERLESS VEHICLE RAN OFF ROAD - HIT FIXED OBJECT         194673 non-null  uint8
 27  SDOT_COLDESC_DRIVERLESS VEHICLE RAN OFF ROAD - NO COLLISION             194673 non-null  uint8
 28  SDOT_COLDESC_DRIVERLESS VEHICLE STRUCK MOTOR VEHICLE FRONT END AT ANGLE 194673 non-null  uint8
 29  SDOT_COLDESC_DRIVERLESS VEHICLE STRUCK MOTOR VEHICLE LEFT SIDE AT ANGLE 194673 non-null  uint8
 30  SDOT_COLDESC_DRIVERLESS VEHICLE STRUCK MOTOR VEHICLE LEFT SIDE SIDESWIPE 194673 non-null  uint8
 31  SDOT_COLDESC_DRIVERLESS VEHICLE STRUCK MOTOR VEHICLE REAR END           194673 non-null  uint8
 32  SDOT_COLDESC_DRIVERLESS VEHICLE STRUCK MOTOR VEHICLE RIGHT SIDE AT ANGLE 194673 non-null  uint8
 33  SDOT_COLDESC_DRIVERLESS VEHICLE STRUCK MOTOR VEHICLE RIGHT SIDE SIDESWIPE 194673 non-null  uint8
 34  SDOT_COLDESC_DRIVERLESS VEHICLE STRUCK OBJECT IN ROADWAY                194673 non-null  uint8
 35  SDOT_COLDESC_DRIVERLESS VEHICLE STRUCK PEDESTRIAN                       194673 non-null  uint8
 36  SDOT_COLDESC_MOTOR VEHCILE STRUCK PEDESTRIAN                            194673 non-null  uint8
 37  SDOT_COLDESC_MOTOR VEHICLE OVERTURNED IN ROAD                           194673 non-null  uint8
 38  SDOT_COLDESC_MOTOR VEHICLE RAN OFF ROAD - HIT FIXED OBJECT              194673 non-null  uint8
 39  SDOT_COLDESC_MOTOR VEHICLE RAN OFF ROAD - NO COLLISION                  194673 non-null  uint8
 40  SDOT_COLDESC_MOTOR VEHICLE STRUCK MOTOR VEHICLE, FRONT END AT ANGLE     194673 non-null  uint8
 41  SDOT_COLDESC_MOTOR VEHICLE STRUCK MOTOR VEHICLE, LEFT SIDE AT ANGLE     194673 non-null  uint8
 42  SDOT_COLDESC_MOTOR VEHICLE STRUCK MOTOR VEHICLE, LEFT SIDE SIDESWIPE    194673 non-null  uint8
 43  SDOT_COLDESC_MOTOR VEHICLE STRUCK MOTOR VEHICLE, REAR END               194673 non-null  uint8
 44  SDOT_COLDESC_MOTOR VEHICLE STRUCK MOTOR VEHICLE, RIGHT SIDE AT ANGLE    194673 non-null  uint8
 45  SDOT_COLDESC_MOTOR VEHICLE STRUCK MOTOR VEHICLE, RIGHT SIDE SIDESWIPE   194673 non-null  uint8
 46  SDOT_COLDESC_MOTOR VEHICLE STRUCK OBJECT IN ROAD                        194673 non-null  uint8
 47  SDOT_COLDESC_MOTOR VEHICLE STRUCK PEDALCYCLIST, FRONT END AT ANGLE      194673 non-null  uint8
 48  SDOT_COLDESC_MOTOR VEHICLE STRUCK PEDALCYCLIST, LEFT SIDE SIDESWIPE     194673 non-null  uint8
 49  SDOT_COLDESC_MOTOR VEHICLE STRUCK PEDALCYCLIST, REAR END                194673 non-null  uint8
 50  SDOT_COLDESC_MOTOR VEHICLE STRUCK PEDALCYCLIST, RIGHT SIDE SIDESWIPE    194673 non-null  uint8
 51  SDOT_COLDESC_MOTOR VEHICLE STRUCK TRAIN                                 194673 non-null  uint8
 52  SDOT_COLDESC_NOT ENOUGH INFORMATION / NOT APPLICABLE                    194673 non-null  uint8
 53  SDOT_COLDESC_PEDALCYCLIST OVERTURNED IN ROAD                            194673 non-null  uint8
 54  SDOT_COLDESC_PEDALCYCLIST RAN OFF ROAD - HIT FIXED OBJECT               194673 non-null  uint8
 55  SDOT_COLDESC_PEDALCYCLIST STRUCK MOTOR VEHICLE FRONT END AT ANGLE       194673 non-null  uint8
 56  SDOT_COLDESC_PEDALCYCLIST STRUCK MOTOR VEHICLE LEFT SIDE AT ANGLE       194673 non-null  uint8
 57  SDOT_COLDESC_PEDALCYCLIST STRUCK MOTOR VEHICLE LEFT SIDE SIDESWIPE      194673 non-null  uint8
 58  SDOT_COLDESC_PEDALCYCLIST STRUCK MOTOR VEHICLE REAR END                 194673 non-null  uint8
 59  SDOT_COLDESC_PEDALCYCLIST STRUCK MOTOR VEHICLE RIGHT SIDE AT ANGLE      194673 non-null  uint8
 60  SDOT_COLDESC_PEDALCYCLIST STRUCK MOTOR VEHICLE RIGHT SIDE SIDESWIPE     194673 non-null  uint8
 61  SDOT_COLDESC_PEDALCYCLIST STRUCK OBJECT IN ROAD                         194673 non-null  uint8
 62  SDOT_COLDESC_PEDALCYCLIST STRUCK PEDALCYCLIST FRONT END AT ANGLE        194673 non-null  uint8
 63  SDOT_COLDESC_PEDALCYCLIST STRUCK PEDALCYCLIST REAR END                  194673 non-null  uint8
 64  SDOT_COLDESC_PEDALCYCLIST STRUCK PEDESTRIAN                             194673 non-null  uint8
 65  WEATHER_Blowing Sand/Dirt                                               194673 non-null  uint8
 66  WEATHER_Clear                                                           194673 non-null  uint8
 67  WEATHER_Fog/Smog/Smoke                                                  194673 non-null  uint8
 68  WEATHER_Other                                                           194673 non-null  uint8
 69  WEATHER_Overcast                                                        194673 non-null  uint8
 70  WEATHER_Partly Cloudy                                                   194673 non-null  uint8
 71  WEATHER_Raining                                                         194673 non-null  uint8
 72  WEATHER_Severe Crosswind                                                194673 non-null  uint8
 73  WEATHER_Sleet/Hail/Freezing Rain                                        194673 non-null  uint8
 74  WEATHER_Snowing                                                         194673 non-null  uint8
 75  WEATHER_Unknown                                                         194673 non-null  uint8
 76  ROADCOND_Dry                                                            194673 non-null  uint8
 77  ROADCOND_Ice                                                            194673 non-null  uint8
 78  ROADCOND_Oil                                                            194673 non-null  uint8
 79  ROADCOND_Other                                                          194673 non-null  uint8
```

```
 80  ROADCOND_Sand/Mud/Dirt                                                                  194673 non-null  uint8
 81  ROADCOND_Snow/Slush                                                                     194673 non-null  uint8
 82  ROADCOND_Standing Water                                                                 194673 non-null  uint8
 83  ROADCOND_Unknown                                                                        194673 non-null  uint8
 84  ROADCOND_Wet                                                                            194673 non-null  uint8
 85  LIGHTCOND_Dark - No Street Lights                                                       194673 non-null  uint8
 86  LIGHTCOND_Dark - Street Lights Off                                                      194673 non-null  uint8
 87  LIGHTCOND_Dark - Street Lights On                                                       194673 non-null  uint8
 88  LIGHTCOND_Dark - Unknown Lighting                                                       194673 non-null  uint8
 89  LIGHTCOND_Dawn                                                                          194673 non-null  uint8
 90  LIGHTCOND_Daylight                                                                      194673 non-null  uint8
 91  LIGHTCOND_Dusk                                                                          194673 non-null  uint8
 92  LIGHTCOND_Other                                                                         194673 non-null  uint8
 93  LIGHTCOND_Unknown                                                                       194673 non-null  uint8
 94  ST_COLDESC_All Other Multi Vehicle                                                      194673 non-null  uint8
 95  ST_COLDESC_All other non-collision                                                      194673 non-null  uint8
 96  ST_COLDESC_Breakage of any part of the vehicle resulting in injury or in further property damage  194673 non-null  uint8
 97  ST_COLDESC_Domestic animal other (cat, dog, etc)                                        194673 non-null  uint8
 98  ST_COLDESC_Entering at angle                                                            194673 non-null  uint8
 99  ST_COLDESC_Fire started in vehicle                                                      194673 non-null  uint8
100 ST_COLDESC_Fixed object                                                                  194673 non-null  uint8
101 ST_COLDESC_From opposite direction - all others                                          194673 non-null  uint8
102 ST_COLDESC_From opposite direction - both going straight - one stopped - sideswipe       194673 non-null  uint8
103 ST_COLDESC_From opposite direction - both going straight - sideswipe                     194673 non-null  uint8
104 ST_COLDESC_From opposite direction - both moving - head-on                               194673 non-null  uint8
105 ST_COLDESC_From opposite direction - one left turn - one right turn                      194673 non-null  uint8
106 ST_COLDESC_From opposite direction - one left turn - one straight                        194673 non-null  uint8
107 ST_COLDESC_From opposite direction - one stopped - head-on                               194673 non-null  uint8
108 ST_COLDESC_From same direction - all others                                              194673 non-null  uint8
109 ST_COLDESC_From same direction - both going straight - both moving - rear-end           194673 non-null  uint8
110 ST_COLDESC_From same direction - both going straight - both moving - sideswipe          194673 non-null  uint8
111 ST_COLDESC_From same direction - both going straight - one stopped - rear-end           194673 non-null  uint8
112 ST_COLDESC_From same direction - both going straight - one stopped - sideswipe          194673 non-null  uint8
113 ST_COLDESC_From same direction - one left turn - one straight                           194673 non-null  uint8
114 ST_COLDESC_From same direction - one right turn - one straight                          194673 non-null  uint8
115 ST_COLDESC_Non-domestic animal (deer, bear, elk, etc)                                   194673 non-null  uint8
116 ST_COLDESC_Not stated                                                                   194673 non-null  uint8
117 ST_COLDESC_One car entering driveway access                                             194673 non-null  uint8
118 ST_COLDESC_One car entering parked position                                             194673 non-null  uint8
119 ST_COLDESC_One car leaving driveway access                                              194673 non-null  uint8
120 ST_COLDESC_One car leaving parked position                                              194673 non-null  uint8
121 ST_COLDESC_One parked--one moving                                                       194673 non-null  uint8
122 ST_COLDESC_Other object                                                                194673 non-null  uint8
123 ST_COLDESC_Pedalcyclist All Other Involvements ONE UNIT - PEDALCYCLIST ONLY or PEDALCYCLIST STR  194673 non-null  uint8
124 ST_COLDESC_Pedalcyclist Strikes Moving Vehicle                                          194673 non-null  uint8
```

```
114 ST_COLDESC_From same direction - one right turn - one straight                          194673 non-null  uint8
115 ST_COLDESC_Non-domestic animal (deer, bear, elk, etc)                                   194673 non-null  uint8
116 ST_COLDESC_Not stated                                                                   194673 non-null  uint8
117 ST_COLDESC_One car entering driveway access                                             194673 non-null  uint8
118 ST_COLDESC_One car entering parked position                                             194673 non-null  uint8
119 ST_COLDESC_One car leaving driveway access                                              194673 non-null  uint8
120 ST_COLDESC_One car leaving parked position                                              194673 non-null  uint8
121 ST_COLDESC_One parked--one moving                                                       194673 non-null  uint8
122 ST_COLDESC_Other object                                                                194673 non-null  uint8
123 ST_COLDESC_Pedalcyclist All Other Involvements ONE UNIT - PEDALCYCLIST ONLY or PEDALCYCLIST STR  194673 non-null  uint8
124 ST_COLDESC_Pedalcyclist Strikes Moving Vehicle                                          194673 non-null  uint8
125 ST_COLDESC_Pedalcyclist Strikes Pedalcyclist or Pedestrian                              194673 non-null  uint8
126 ST_COLDESC_Person fell, jumped or was pushed from vehicle                               194673 non-null  uint8
127 ST_COLDESC_Railway Vehicle Strikes Pedalcyclist                                         194673 non-null  uint8
128 ST_COLDESC_Railway Vehicle Strikes Pedestrian                                           194673 non-null  uint8
129 ST_COLDESC_Railway Vehicle Strikes Vehicle                                              194673 non-null  uint8
130 ST_COLDESC_Same direction -- both turning left -- both moving -- rear end               194673 non-null  uint8
131 ST_COLDESC_Same direction -- both turning left -- both moving -- sideswipe              194673 non-null  uint8
132 ST_COLDESC_Same direction -- both turning left -- one stopped -- rear end               194673 non-null  uint8
133 ST_COLDESC_Same direction -- both turning left -- one stopped -- sideswipe              194673 non-null  uint8
134 ST_COLDESC_Same direction -- both turning right -- both moving -- rear end              194673 non-null  uint8
135 ST_COLDESC_Same direction -- both turning right -- both moving -- sideswipe             194673 non-null  uint8
136 ST_COLDESC_Same direction -- both turning right -- one stopped -- rear end              194673 non-null  uint8
137 ST_COLDESC_Same direction -- both turning right -- one stopped -- sideswipe             194673 non-null  uint8
138 ST_COLDESC_Strikes or Was Struck by Object from the Load of Another Vehicle            194673 non-null  uint8
139 ST_COLDESC_Strikes or Was Struck by a Part of Another Vehicle (Not from Load)          194673 non-null  uint8
140 ST_COLDESC_Vehicle - Pedalcyclist                                                      194673 non-null  uint8
141 ST_COLDESC_Vehicle Hits City Road or Construction Machinery                            194673 non-null  uint8
142 ST_COLDESC_Vehicle Hits Other Road or Construction Machinery                           194673 non-null  uint8
143 ST_COLDESC_Vehicle Hits State Road or Construction Machinery                           194673 non-null  uint8
144 ST_COLDESC_Vehicle Strikes All Other Non-Domestic Animal                               194673 non-null  uint8
145 ST_COLDESC_Vehicle Strikes Deer                                                        194673 non-null  uint8
146 ST_COLDESC_Vehicle Strikes Pedalcyclist                                                194673 non-null  uint8
147 ST_COLDESC_Vehicle Strikes Railway Vehicle                                             194673 non-null  uint8
148 ST_COLDESC_Vehicle Struck by City Road or Construction Machinery                       194673 non-null  uint8
149 ST_COLDESC_Vehicle Struck by Other Road or Construction Machinery                      194673 non-null  uint8
150 ST_COLDESC_Vehicle backing hits pedestrian                                             194673 non-null  uint8
151 ST_COLDESC_Vehicle going straight hits pedestrian                                      194673 non-null  uint8
152 ST_COLDESC_Vehicle hits Pedestrian - All Other Actions                                 194673 non-null  uint8
153 ST_COLDESC_Vehicle overturned                                                          194673 non-null  uint8
154 ST_COLDESC_Vehicle turning left hits pedestrian                                        194673 non-null  uint8
155 ST_COLDESC_Vehicle turning right hits pedestrian                                       194673 non-null  uint8
dtypes: int64(5), object(1), uint8(150)
memory usage: 36.8+ MB
```

2.5Test of correlation and significance

To get a better measure of the important characteristics, we would look at the correlation of the variables vis-a-vis target variable i.e. Accident Severity. 2 measures are followed for the analysis.

A. Pearson Correlation:

The Pearson Correlation measures the linear dependence between two variables X and Y of 'int64' or 'float64' types.

The resulting coefficient is a value between -1 and 1 inclusive, where:

1: Total positive linear correlation.
0: No linear correlation, the two variables most likely do not affect each other.
-1: Total negative linear correlation.
The closeness to terminal values (-1 and 1) would decide strength of the correlation.

B. P-value:

The P-value is the probability value that the correlation between these two variables is statistically significant. Normally, we choose a significance level of 0.05, which means that we are 95% confident that the correlation between the variables is significant. We would use "stats" module in the "Scipy" library to get the P-value.

By convention, when the

p-value is < 0.001: we say there is strong evidence that the correlation is significant.
the p-value is < 0.05: there is moderate evidence that the correlation is significant.
the p-value is < 0.1: there is weak evidence that the correlation is significant.
the p-value is > 0.1: there is no evidence that the correlation is significant.

[19]:
```
from scipy import stats

dict={"Column Name":["Pearson Correlation Coefficient","P-value"]}
for column in df_corr.columns.values.tolist():
    pearson_coef, p_value = stats.pearsonr(df_corr[column], df_corr['SEVERITYCODE'])
    dict[column]=[pearson_coef,p_value]
df_dict = pd.DataFrame(dict)
df_dict_clean=df_dict.transpose(copy=True)
df_dict_clean.rename(columns={0: 'Pearson Correlation Coefficient',1:'P-value'}, inplace=True)
df_dict_clean.reset_index(inplace=True)
df_dict_clean.drop([0,1], axis=0, inplace=True)
df_dict_clean
```

•••

[21]:
```
df_dict_clean[df_dict_clean['Pearson Correlation Coefficient'] > 0.2]
```

[21]:

	index	Pearson Correlation Coefficient	P-value
3	PEDCOUNT	0.246338	0
4	PEDCYLCOUNT	0.214218	0
10	COLLISIONTYPE_Cycles	0.213271	0
15	COLLISIONTYPE_Pedestrian	0.245441	0
20	JUNCTIONTYPE_At Intersection (intersection rel...	0.201628	0
36	SDOT_COLDESC_MOTOR VEHCILE STRUCK PEDESTRIAN	0.243571	0

2.6Conclusion: Important Variables

By now we have a better idea of what our data looks like and which variables are important to take into account when predicting the ‘Severity’ class.

As we move into building machine learning modelling steps to automate our analysis, feeding the model with variables that meaningfully affect our target variable would help improve the model's prediction performance.