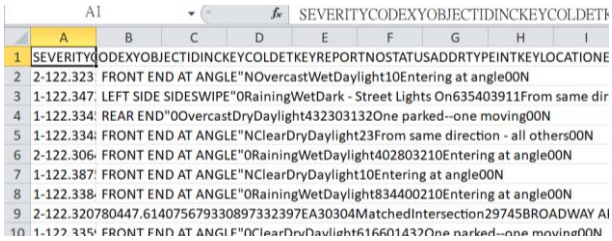
Data Understanding –

Classification of Road Accident Severity

To build a classification model about road accident severity, I decided to leveraged 'Data Collision' file provided by coursea – IBM Data Science Course

To begin with, csv file has to be read in python. However, the format of the csv is unstructured



	A	B	C	D	E	F	G	H	I
1	SEVERITYC	ODEXY	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	LOCATIONE
2	2-122.323	FRONT END AT ANGLE"	NOvercast	Wet	Daylight	10	Entering at angle	00	N	
3	1-122.347	LEFT SIDE SIDESWIPE"	0	Raining	Wet	Dark - Street Lights On	635403911	From same dir		
4	1-122.334	REAR END"	0	Overcast	Dry	Daylight	432303132	One parked--one moving	00	N
5	1-122.334	FRONT END AT ANGLE"	N	Clear	Dry	Daylight	23	From same direction - all others	00	N
6	2-122.306	FRONT END AT ANGLE"	0	Raining	Wet	Daylight	402803210	Entering at angle	00	N
7	1-122.387	FRONT END AT ANGLE"	N	Clear	Dry	Daylight	10	Entering at angle	00	N
8	1-122.338	FRONT END AT ANGLE"	0	Raining	Wet	Daylight	834400210	Entering at angle	00	N
9	2-122.320	780447.6	140756793308973323	97	EA30304	Matched	Intersection	29745	BROADWAY A	
10	1-122.335	FRONT END AT ANGLE"	0	Clear	Dry	Daylight	616601432	One parked--one moving	00	N

I need to do the reformatting work

SEVERITYC	X	Y	OBJECTID	INCKEY	COLDETKE	REPORTN	STATUS	ADDRTYPE	INTKEY	LOCATION	EXCEPTRSI	EXCEPTRSI	SEVERITYC
2	-122.323	47.70314	1	1307	1307	3502005	Matched	Intersectic	37475	5TH AVE N			Injury Coll
1	-122.347	47.64717	2	52200	52200	2607959	Matched	Block		AURORA BR BETWEEN	RAYE ST		Property D
1	-122.335	47.60787	3	26700	26700	1482393	Matched	Block		4TH AVE BETWEEN	SENECA ST	/	Property D
1	-122.335	47.6048	4	1144	1144	3503937	Matched	Block		2ND AVE E			Property D
2	-122.306	47.54574	5	17700	17700	1807429	Matched	Intersectic	34387	SWIFT AVE S AND	SWIFT AV	OFI	Injury Coll
1	-122.388	47.69058	6	320840	322340	E919477	Matched	Intersectic	36974	24TH AVE			Property D
1	-122.338	47.61853	7	83300	83300	3282542	Matched	Intersectic	29510	DENNY WAY AND	WESTLAKE A	V	Property D
2	-122.321	47.61408	9	330897	332397	EA30304	Matched	Intersectic	29745	BROADWA			Injury Coll
1	-122.336	47.6119	10	63400	63400	2071243	Matched	Block		PINE ST BETWEEN	5TH AVE AN	D	Property D

After a quick scan, found that there is duplicated data field'SEVERITYCODE' so I removed it

Then read the csv and read the first 5 row of the data and the datatype

Data Understanding

Data source:Data - Collisions from coursera-IBM Data Science Course

In [3]:
```
import numpy as np
import pandas as pd
```

In [5]:
```
df=pd.read_csv('./DataCollisions.csv')
df.head()
```

```
C:\ProgramData\Anaconda3\lib\site-packages\IPython\core\interactiveshell.py:3063: DtypeWarning: Columns (32) have mixed type
s.Specify dtype option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

Out[5]:

	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

```
Out[7]: 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
        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
```

SEVERITYCODE is a target variable ‘’y’’ for model development

```
In [8]: df.SEVERITYCODE.value_counts()

Out[8]: 1    108515
        2     45036
        Name: SEVERITYCODE, dtype: int64
```

1.Before data input, remove duplicated field"severitycode" in the original csv file

2.Severitycode is the target variable,y

3.Remove result related variables to avoid data leakage,eg.SEVERITYDESC

4.convert character to numeric data,eg.ROADCOND

Remove result related variables to avoid data leakage,eg.SEVERITYDESC

```
In [16]: df[['SEVERITYDESC','SEVERITYCODE']].groupby(['SEVERITYDESC']).agg(['min','max'])
```

Out[16]:

	SEVERITYCODE	
	min	max
SEVERITYDESC		
Injury Collision	2	2
Property Damage Only Collision	1	1

Other meaningless variable are also removed

```
In [23]: dataset=df.drop(columns=['X','Y','OBJECTID','INCKEY','COLDETKEY','REPORTNO',
             'STATUS','ADDRTYPE','INTKEY','LOCATION',
             'EXCEPTRSNCODE','EXCEPTRSNDESC','SEVERITYDESC','COLLISIONTYPE','PERSONCOUNT','PEDCOUNT',
              'PEDCYLCOUNT','VEHCOUNT','INCDATE','INCDTTM','JUNCTIONTYPE','SDOT_COLCODE','SDOT_COLDESC',
              'INATTENTIONIND','UNDERINFL','PEDROWNOTGRNT','SDOTCOLNUM','SPEEDING','ST_COLCODE'
             ,'ST_COLDESC','SEGLANEKEY','CROSSWALKKEY','HITPARKEDCAR'])
         dataset.head()
```

Out[23]:

	SEVERITYCODE	WEATHER	ROADCOND	LIGHTCOND
0	2	Overcast	Wet	Daylight
1	1	Raining	Wet	Dark - Street Lights On
2	1	Overcast	Dry	Daylight
3	1	Clear	Dry	Daylight
4	2	Raining	Wet	Daylight

Assign missing to the field Weather,Roadcond,lightcond

```
In [76]: dataset['WEATHER']=dataset['WEATHER'].fillna(value='Unknown')
         dataset['WEATHER'].value_counts()

Out[76]: Clear                    86878
         Raining                  26468
         Overcast                 22300
         Unknown                  15912
         Snowing                    704
         Other                      701
         Fog/Smog/Smoke             442
         Sleet/Hail/Freezing Rain    96
         Blowing Sand/Dirt           34
         Severe Crosswind            16
         Name: WEATHER, dtype: int64
```

```
In [77]: dataset['ROADCOND']=dataset['ROADCOND'].fillna(value='Other')
         dataset['ROADCOND'].value_counts()

Out[77]: Dry              97842
         Wet              37741
         Unknown          13827
         Other             2085
         Ice               1035
         Snow/Slush         811
         Standing Water      96
         Sand/Mud/Dirt       66
         Oil                 48
         Name: ROADCOND, dtype: int64
```

```
In [78]: dataset['LIGHTCOND']=dataset['LIGHTCOND'].fillna(value='Other')
         dataset['LIGHTCOND'].value_counts()

Out[78]: Daylight                   91280
         Dark - Street Lights On    38829
         Unknown                    12464
         Dusk                        4707
```

Convert the string to number

```
In [79]: dataset['WEATHER_NUM']=0
         dataset.loc[dataset.WEATHER=='Clear','WEATHER_NUM']=4
         dataset.loc[dataset.WEATHER=='Raining','WEATHER_NUM']=3
         dataset.loc[dataset.WEATHER=='Overcast','WEATHER_NUM']=2
         dataset.loc[dataset.WEATHER=='Unknown','WEATHER_NUM']=1
         dataset.WEATHER_NUM.value_counts()

Out[79]: 4    86878
         3    26468
         2    22300
         1    15912
         0     1993
         Name: WEATHER_NUM, dtype: int64
```

```
In [80]: dataset['ROADCOND_NUM']=1
         dataset.loc[dataset.ROADCOND=='Dry','ROADCOND_NUM']=3
         dataset.loc[dataset.ROADCOND=='Wet','ROADCOND_NUM']=2
         dataset.loc[(dataset.ROADCOND=='Unknown')|(dataset.ROADCOND=='Other'),'ROADCOND_NUM']=0

         dataset.ROADCOND_NUM.value_counts()

Out[80]: 3    97842
         2    37741
         0    15912
         1     2056
         Name: ROADCOND_NUM, dtype: int64
```

```
In [81]: dataset['LIGHTCOND_NUM']=0
         dataset.loc[dataset.LIGHTCOND=='Daylight','LIGHTCOND_NUM']=5
         dataset.loc[dataset.LIGHTCOND=='Dusk','LIGHTCOND_NUM']=4
         dataset.loc[dataset.LIGHTCOND=='Dawn','LIGHTCOND_NUM']=3
         dataset.loc[dataset.LIGHTCOND=='Dark - Street Lights On','LIGHTCOND_NUM']=2
         dataset.loc[(dataset.LIGHTCOND=='Dark - No Street Lights')
                     |(dataset.LIGHTCOND=='Dark - Street Lights Off')
                     |(dataset.LIGHTCOND=='Dark - Unknown Lighting '),'LIGHTCOND_NUM']=1
```

Next step is to consider a suitable model for classification:

KNN,Decision Tree & SVM