Data

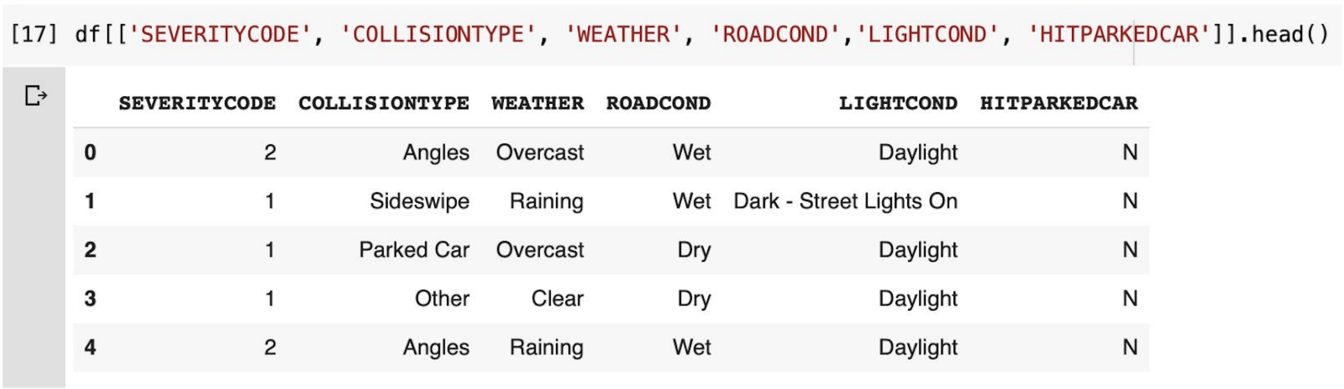
Data Source

We will use the data source provided as part of the course. It can be found here: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

The attributes used to train the machine learning model are:

- **SEVERITYCODE**: The attribute we are trying to predict, the higher, the bigger loss.
- **COLLISIONTYPE**: Gives clues about the situation when the accident happened, i.e: “Parked Car”, “Rear Ender”, etc.
- **WEATHER**: Weather conditions may contribute to the severity of the accident, so we need to consider it.
- **ROADCOND**: This attribute tells if the condition of the road was wet, or dry,, which definitely contributes to collisions.
- **LIGHTCOND**: This attribute defines if the road was dark, if the collision happened during the day, or dusk.
- **HITPARKEDCAR**: Yes/No flag indicating if the collision was against a parked car.

Example of data



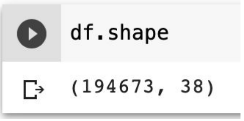
```
[17] df[['SEVERITYCODE', 'COLLISIONTYPE', 'WEATHER', 'ROADCOND','LIGHTCOND', 'HITPARKEDCAR']].head()
```

	SEVERITYCODE	COLLISIONTYPE	WEATHER	ROADCOND	LIGHTCOND	HITPARKEDCAR
0	2	Angles	Overcast	Wet	Daylight	N
1	1	Sideswipe	Raining	Wet	Dark - Street Lights On	N
2	1	Parked Car	Overcast	Dry	Daylight	N
3	1	Other	Clear	Dry	Daylight	N
4	2	Angles	Raining	Wet	Daylight	N

Analysis of Data

Missing values

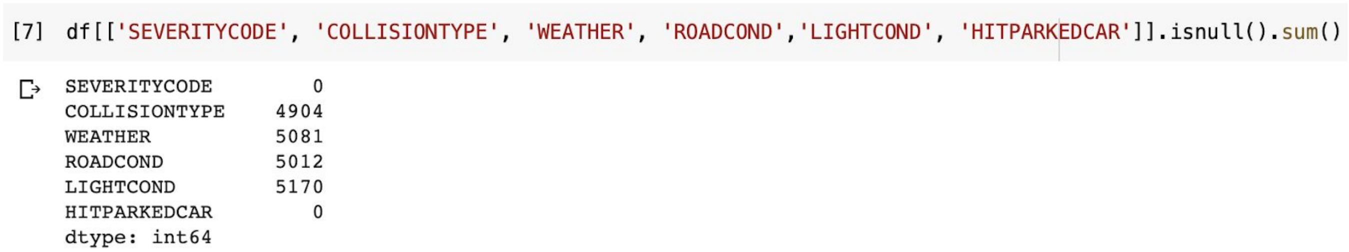
We have a total of 194,673 rows of data.



```
df.shape
```

```
(194673, 38)
```

From this dataset, the number of missing values is:



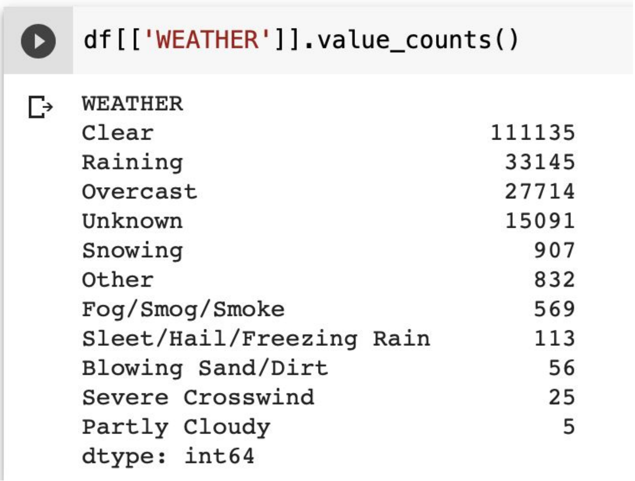
```
[7] df[['SEVERITYCODE', 'COLLISIONTYPE', 'WEATHER', 'ROADCOND','LIGHTCOND', 'HITPARKEDCAR']].isnull().sum()
```

```
SEVERITYCODE        0
COLLISIONTYPE    4904
WEATHER          5081
ROADCOND         5012
LIGHTCOND        5170
HITPARKEDCAR        0
dtype: int64
```

The attribute LIGHTCOND has the most missing values, which represents the 2.6% of the data. Missing values won’t help in predicting severity code, so we should drop them from the dataset.

Balance in data

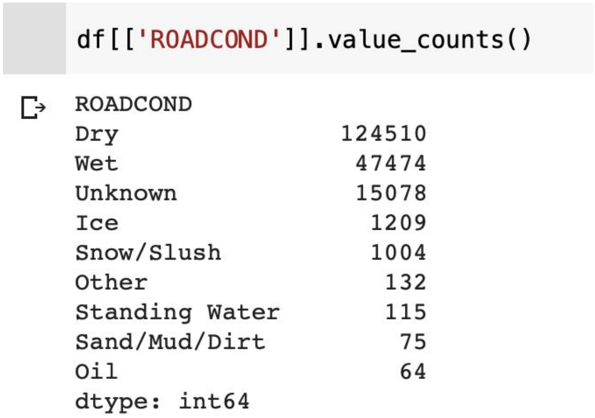
The attribute WEATHER presents imbalance in data, so it would be better to ignore the “Partly Cloudy”, “Severe Crosswind” and “Blowing Sand/Dirt” categories.



```
df[['WEATHER']].value_counts()
```

```
WEATHER
Clear                       111135
Raining                      33145
Overcast                     27714
Unknown                      15091
Snowing                        907
Other                          832
Fog/Smog/Smoke                 569
Sleet/Hail/Freezing Rain       113
Blowing Sand/Dirt               56
Severe Crosswind                25
Partly Cloudy                    5
dtype: int64
```

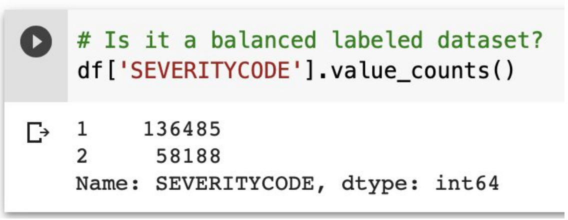
When analyzing the ROADCOND attribute, we find that we should ignore the “Sand/Mud/Dirt” and “Oil” categories.



```
df[['ROADCOND']].value_counts()
```

```
ROADCOND
Dry               124510
Wet                47474
Unknown            15078
Ice                 1209
Snow/Slush          1004
Other                132
Standing Water       115
Sand/Mud/Dirt         75
Oil                   64
dtype: int64
```

The attribute we are trying to predict, SEVERITYCODE, is unbalanced, we would need to balance it before building our model.



```
# Is it a balanced labeled dataset?
df['SEVERITYCODE'].value_counts()
```

```
1    136485
2     58188
Name: SEVERITYCODE, dtype: int64
```