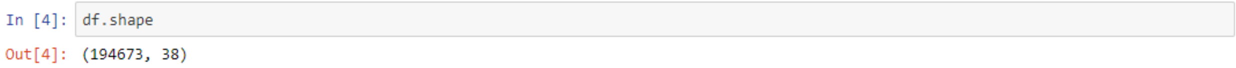
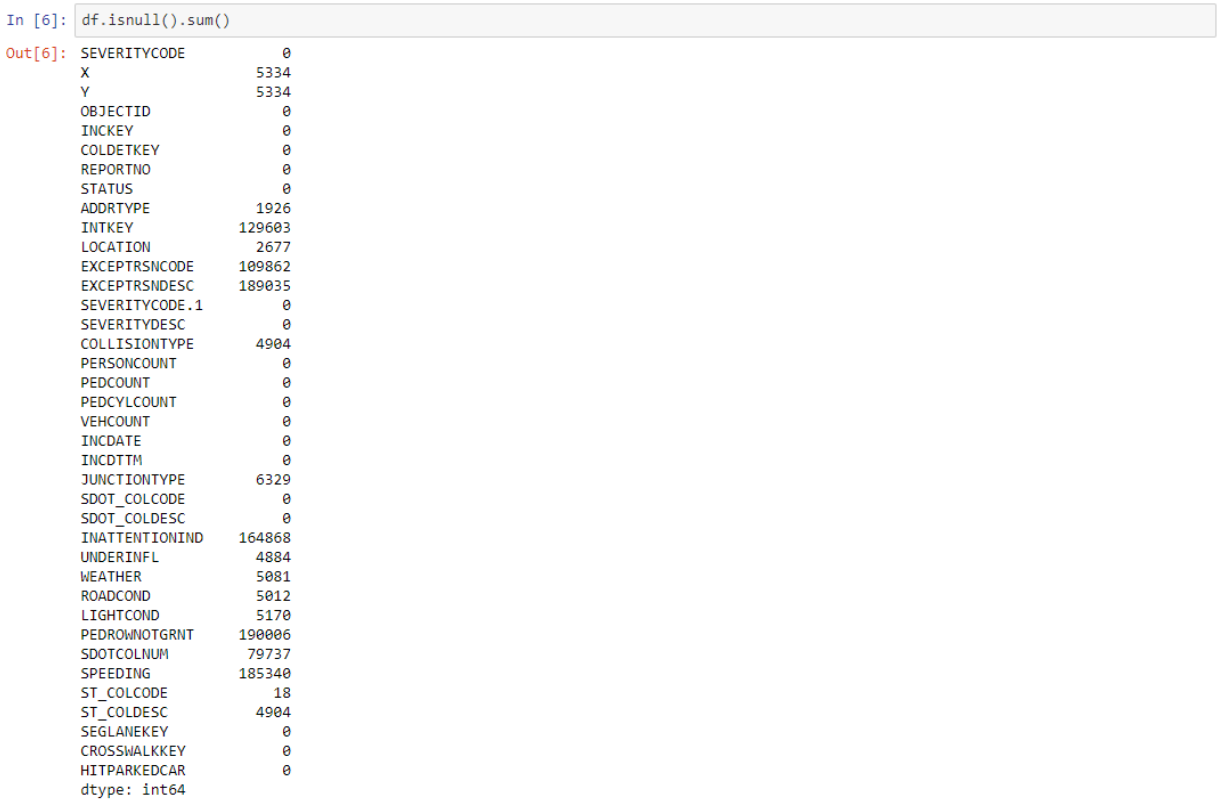
By checking the dataset, the dataset has a total of 38 attributes and 194,673 data.



```
In [4]: df.shape
Out[4]: (194673, 38)
```

At the same time, many attributes have null values.



```
In [6]: df.isnull().sum()
Out[6]: SEVERITYCODE           0
        X                   5334
        Y                   5334
        OBJECTID               0
        INCKEY                 0
        COLDETKEY              0
        REPORTNO               0
        STATUS                 0
        ADDRTYPE            1926
        INTKEY            129603
        LOCATION            2677
        EXCEPTRSNCODE     109862
        EXCEPTRSNDESC     189035
        SEVERITYCODE.1         0
        SEVERITYDESC           0
        COLLISIONTYPE       4904
        PERSONCOUNT            0
        PEDCOUNT               0
        PEDCYLCOUNT            0
        VEHCOUNT               0
        INCDATE                0
        INCDTTM                0
        JUNCTIONTYPE        6329
        SDOT_COLCODE           0
        SDOT_COLDESC           0
        INATTENTIONIND    164868
        UNDERINFL           4884
        WEATHER             5081
        ROADCOND            5012
        LIGHTCOND           5170
        PEDROWNOTGRNT     190006
        SDOTCOLNUM         79737
        SPEEDING          185340
        ST_COLCODE            18
        ST_COLDESC          4904
        SEGLANEKEY             0
        CROSSWALKKEY           0
        HITPARKEDCAR           0
        dtype: int64
```

I will process the data like this:

1. Remove text attributes like description that cannot be converted into numeric values. Then remove the unrelated latitude and longitude. Then fill in the null value (delete or mean or mode) of each attribute according to the attributes of different attributes, and then regularize the data.

2. Divide the data into training set and test set.

3. Establish the Logistics Regression Model and Random Forest model.

4. Optimize the two models and compare their accuracy.

5. Choose the most suitable model.