Capstone Project - Car accident severity (Week 1)

|  |  |
| --- | --- |
| **Title Collisions** | All Years |
| **Abstract** | All collisions provided by SPD and recorded by Traffic Records. |
| **Description** | This includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. Timeframe: 2004 to Present |

**Introduction | Business Understanding**

**Background**

Road traffic are becoming one of the leading causes of death across all age groups globally. Analyzing a significant range of factors, including weather conditions, special events, roadworks, traffic jams among others, an accurate prediction of the severity of the accidents can be performed.

These insights will allow law enforcement bodies to allocate their resources more effectively in advance of potential accidents, preventing when and where a severe accidents can occur as well as saving both, time and money. In addition, this knowledge of a severe accident situation can be warned to drivers so that they would drive more carefully or even change their route if it is possible or to hospital which could have set everything ready for a severe intervention in advance.

**Problem Description**

In an effort to reduce the frequency of car collisions in a community, an algorithm must be developed to predict the severity of an accident given the current weather, road and visibility conditions. When conditions are bad, this model will alert drivers to remind them to be more careful.

**Target Audience**

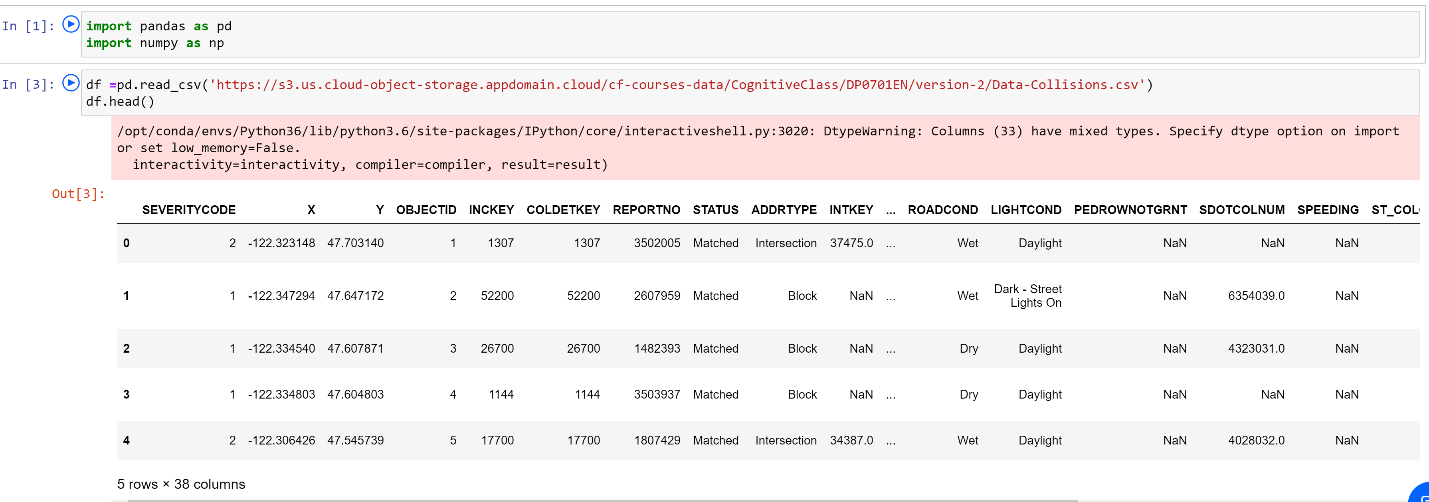
Governments should be highly interested in accurate predictions of the severity of an accident, in order to reduce the time of arrival and thus save a significant amount of people each year.

The public especially the travelers to a city would also be interested in accurate predictions of the severity of an accident under current conditions.

**Data**

**Data Understanding**

For this project I’ll use data collected by “SDOT Traffic Management Division, Traffic Records Group”, in “Data\_Collision.csv”. The data contains all types of collisions provided by SPD and recorded by Traffic Records from Timeframe: 2004 to Present.



This dataset contains 194674 rows and 38 columns with most of the information that will be needed for the project such as current weather, road and visibility conditions.



Our predictor or target variable will be 'SEVERITYCODE' because it is used measure the severity of an accident in 5 levels [0,1,2,2b,3] within the dataset.

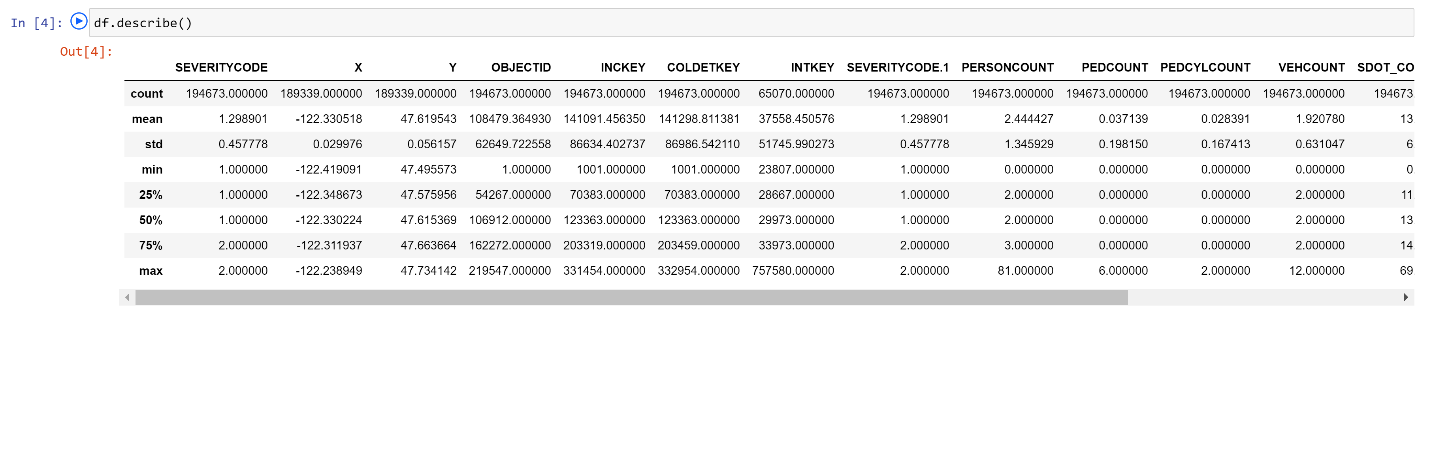
Severity codes are as follows:

|  |  |
| --- | --- |
| **Severity Code** | **Severity Description** |
| 3 | fatality |
| 2b | serious injury |
| 2 | injury |
| 1 | prop damage |
| 0 | unknown |

Attributes used to weigh the severity of an accident are ‘'WEATHER', 'ROADCOND' and 'LIGHTCOND'.

**Data Preparation and Cleaning**

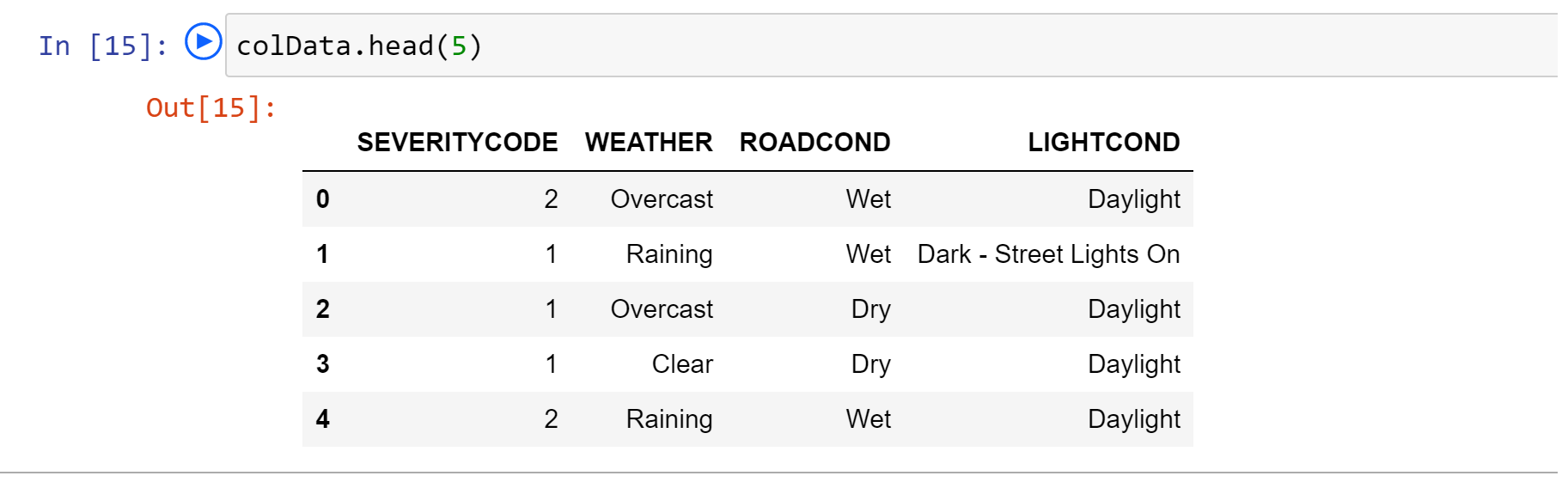
However, this dataset contains mixed datatypes and needs extensive cleaning before it can be used for the project.



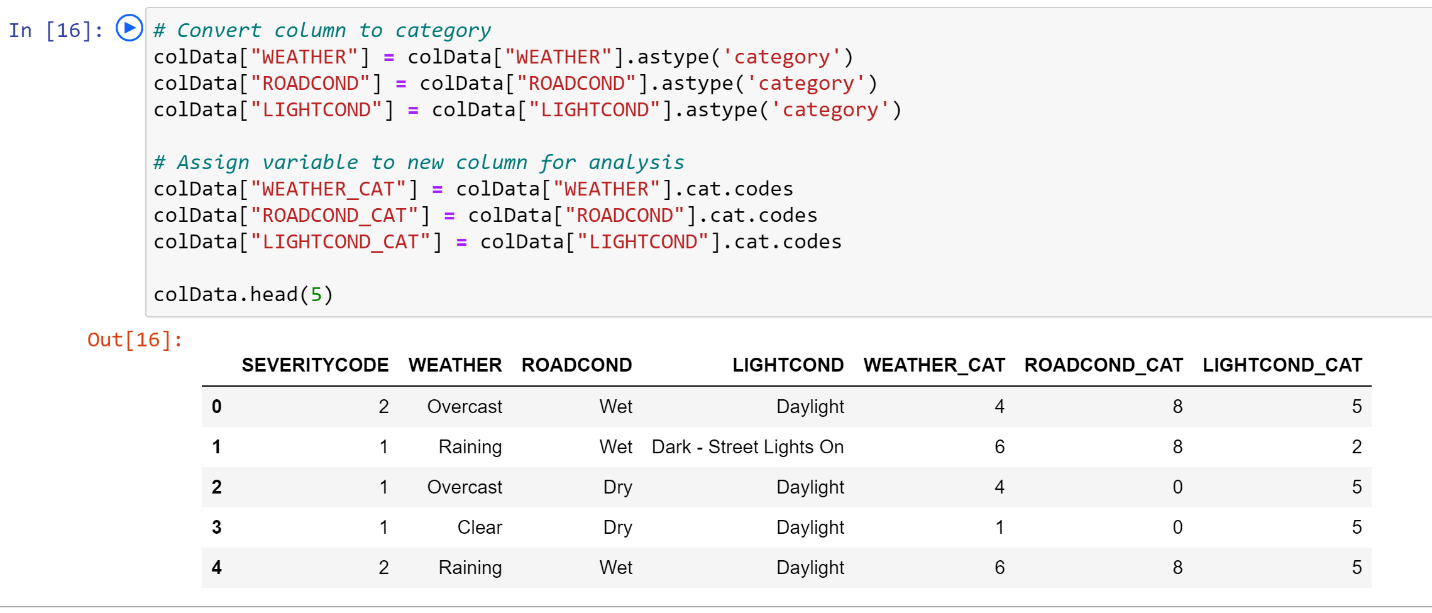
* Select data that is related to our problem
* Remove unnecessary data [Drop these columns]
* Remove the rows with null values.
* Do data profiling to validate correctness of data.
* Convert columns with text field to numerical data like categorizing the weather conditions
* Eliminated rows that greatly contributed to the imbalance of the data and were not significant in volume
* Replaced null values with fillers
* Remove unnecessary data [Drop these columns]



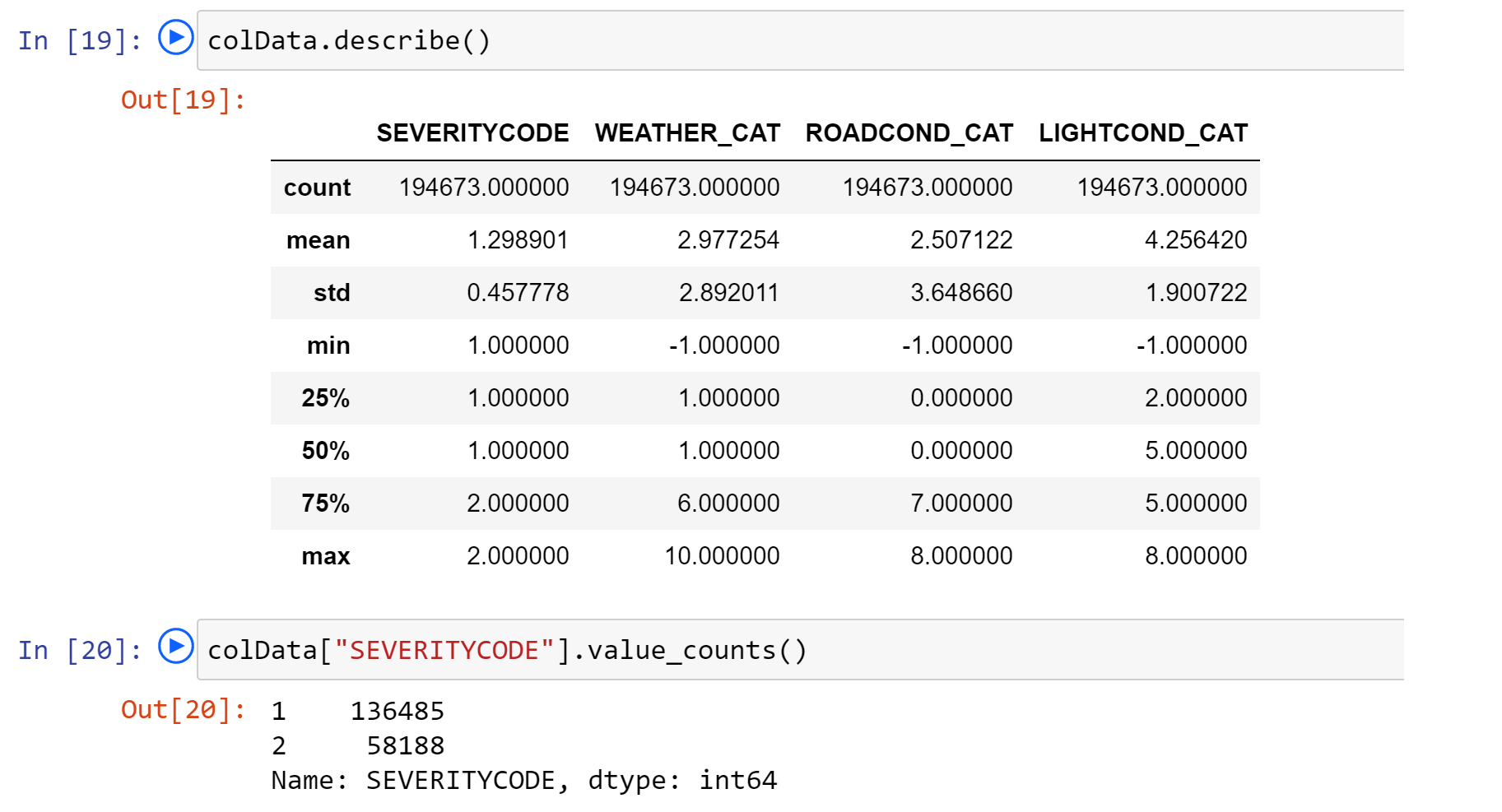
* Select data that is related to our problem

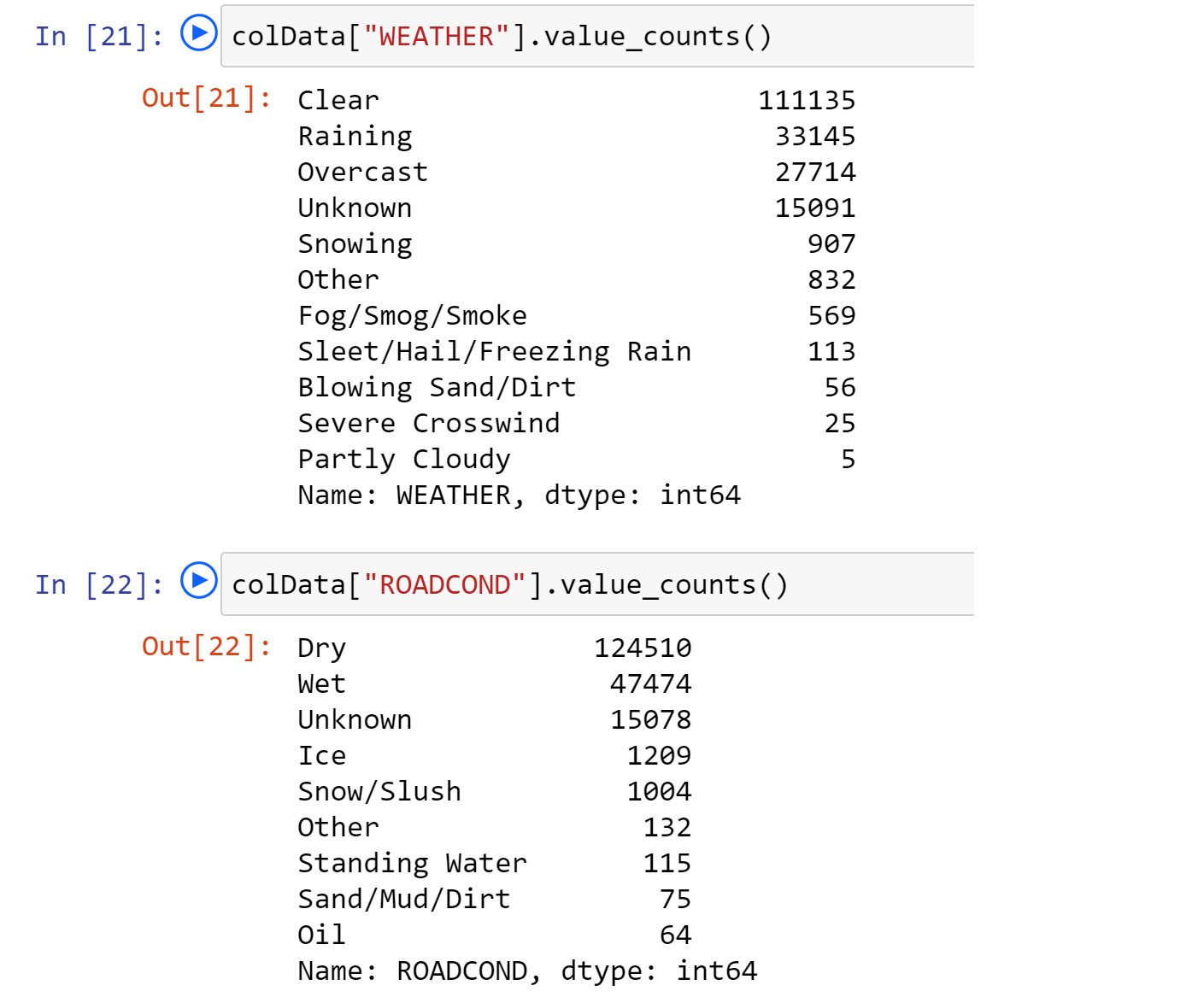


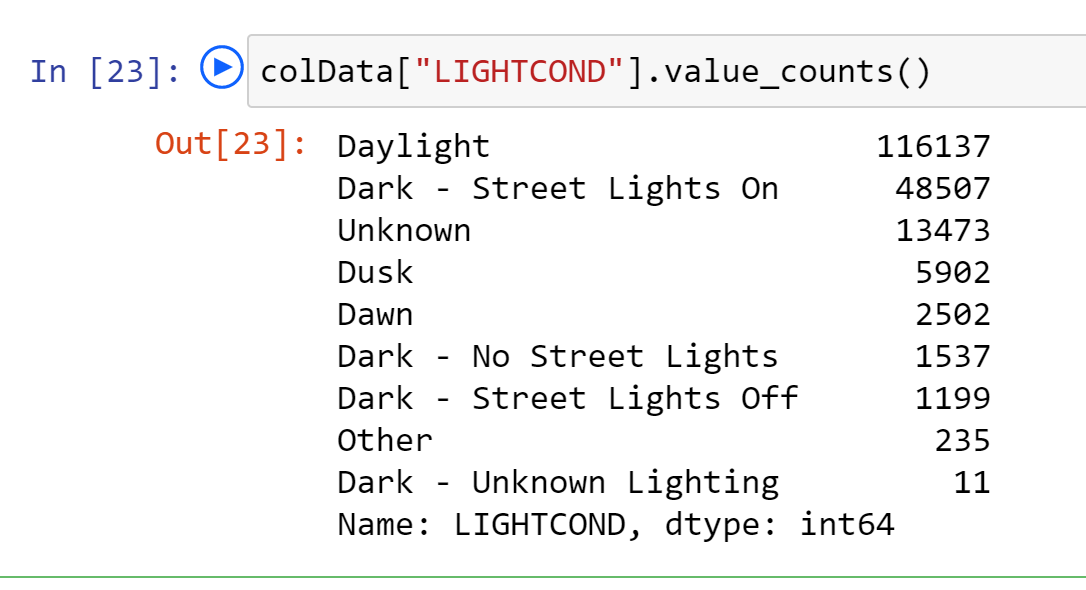
* Convert columns with text field to numerical data like categorizing the weather conditions



* Eliminated rows that greatly contributed to the imbalance of the data and were not significant in volume







Our target variable SEVERITYCODE is not balanced. SEVERITYCODE in class 1 is nearly three times the size of class 2.

We can fix this by **downsampling** the majority class.

Down-sampling involves randomly removing observations from the majority class to prevent its signal from dominating the learning algorithm. Here are the steps:

* First, we'll separate observations from each class into different DataFrames.
* Next, we'll resample the majority class without replacement, setting the number of samples to match that of the minority class.
* Finally, we'll combine the down-sampled majority class DataFrame with the original minority class DataFrame.

