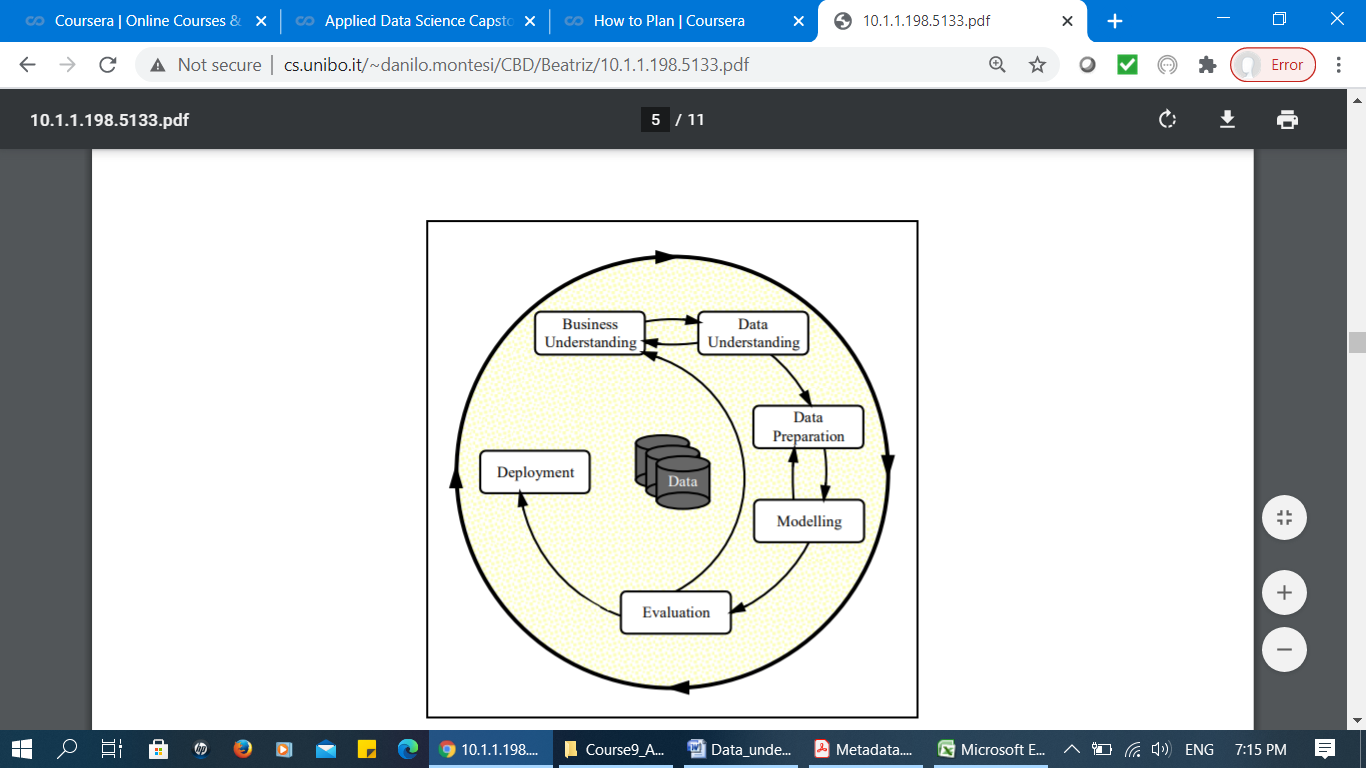
**CHAPTER 2: DATA UNDERSTANDING AND ANALYSIS**

**INTRODUCTION**

In this study, we follow CRISP-Data Mining (CRoss-Industry Standard Process for Data Mining) methodology [2]. It provides a guideline of a data mining\data science project life cycle (refer to Figure 1). We have implemented these phases and the detailed process is explained in the following sections.

****

**Figure 1: Phases of the CRISP-DM Process Model**

**1. Business Understanding**

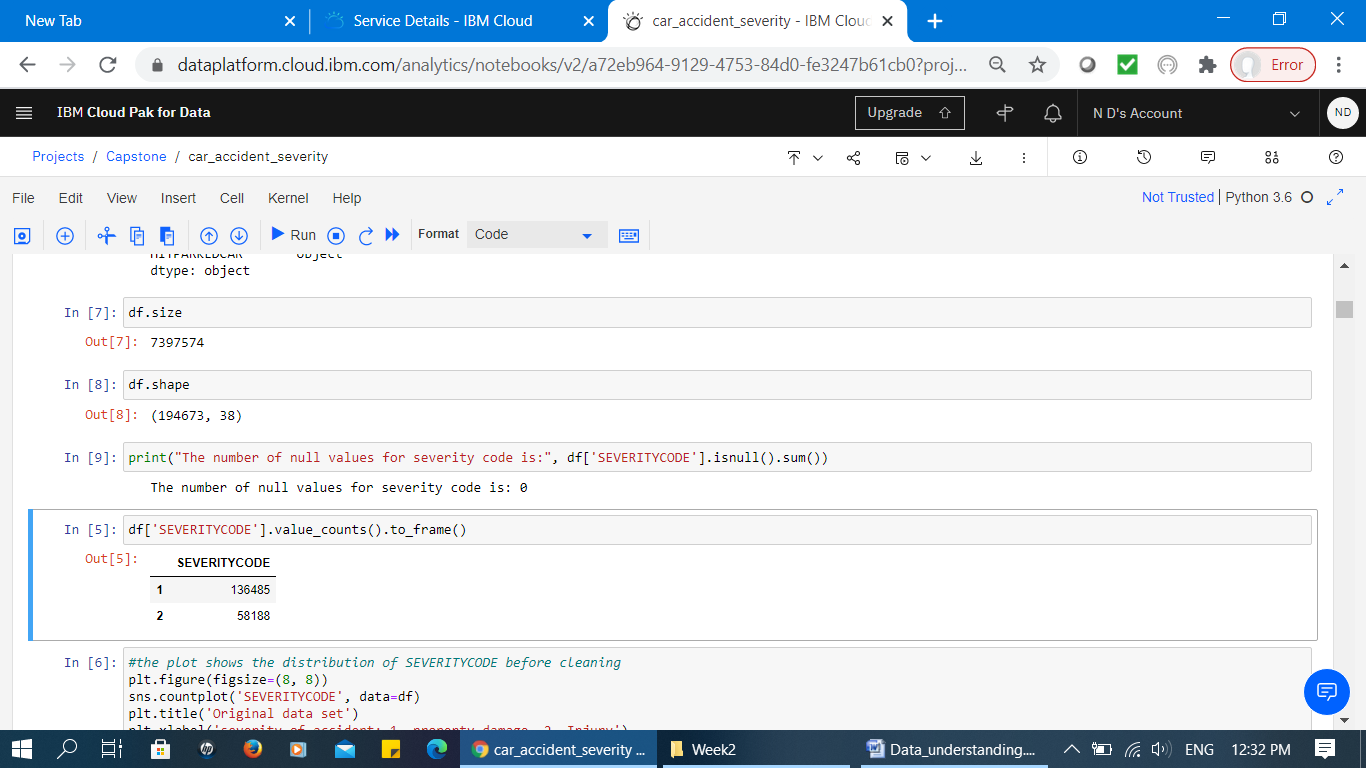
The study focuses on car accident severity in Seattle city. We studied the project and identified the problems and requirements. The study aims to develop a model that could predict the severity of car accident given by the factors affecting the collision. Basically, the beneficiaries in this study is not limited to the emergency department units that could potentially present more advanced help carrier to the community but also everyone who tend to travel by any motor vehicle. This is briefly explained in “Chapter 1: Introduction” section of this study.

**2. Data Understanding**

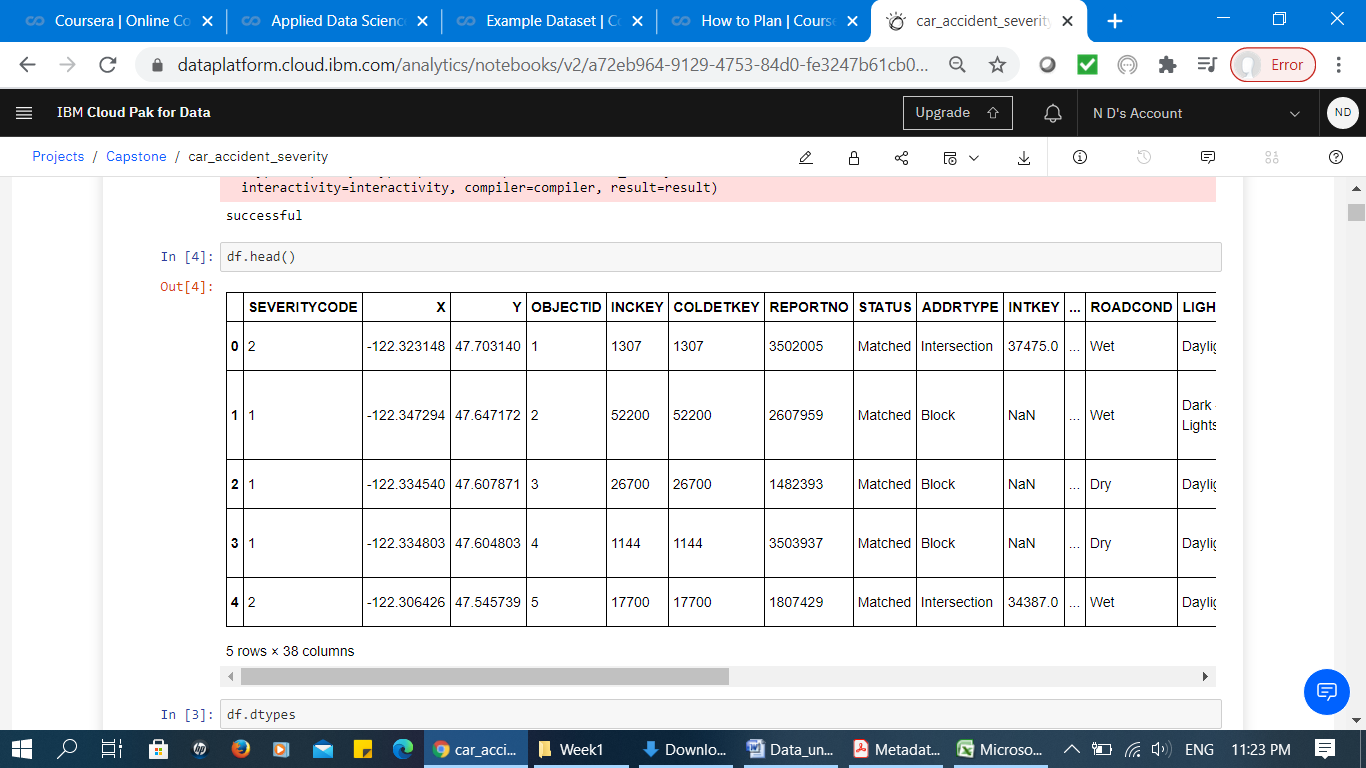
1. *Introduction*

This section initiates with data collection process and it is comprised of the steps involved in getting familiar with dataset and to identify the quality of collected data. Basically, the data describe all types of collisions displayed at the intersection or mid-block of a segment from January 2004 to May 2020. The collisions are provided by Seattle Police Department (SPD) and recorded by Traffic Records group. They are accessed and collected from the following link (<https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv>).

The original dataset is comprised of 194,673 rows and 38 columns (Figure 2). The objective is to identify the impact of traffic using the severity of accident. Therefore, the “SEVERITYCODE” attribute, which describes the fatality of an accident, will be used as the dependant (target) variable. This code corresponds to the severity of the collision through 5 values (3: fatality, 2b: serious injury, 2: injury, 1: property damage, 0: unknown). In the current dataset, the total number of cases relevant to code 1 and 2 (injury and property damage respectively) is available. We identified there are no missing values in this attribute (Figure 2). Moreover, the remaining 37 columns are described as independent variables and their corresponding values (Figure 3 and Figure 4).

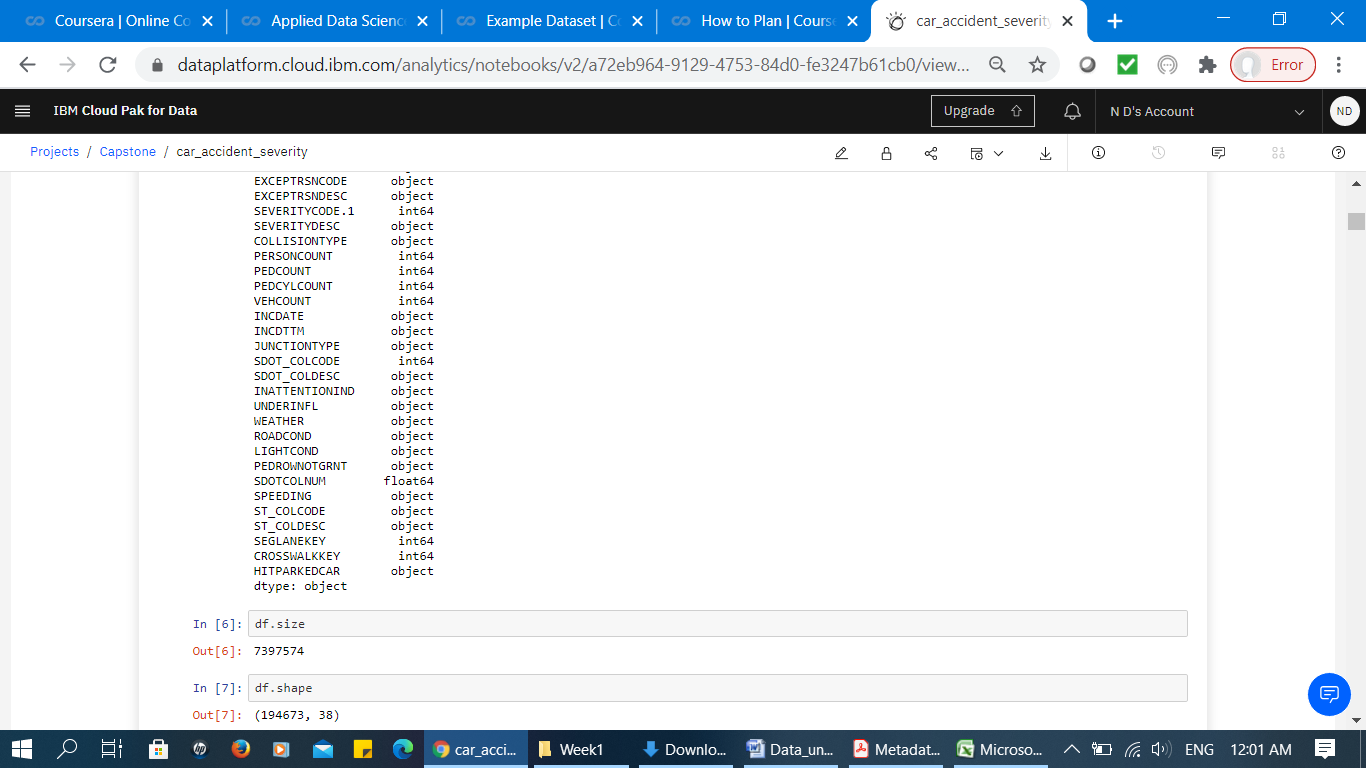
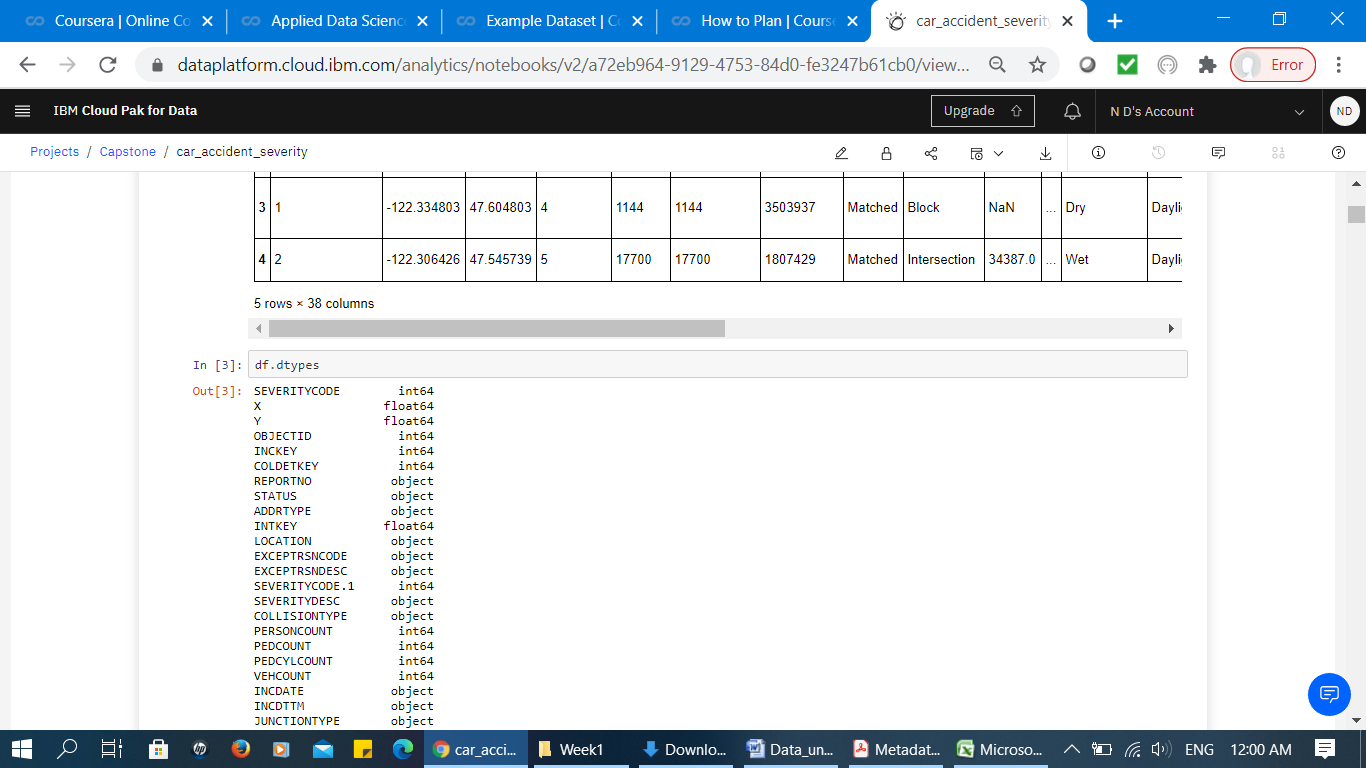


**Figure 2: Dataset size**



**Figure 3: Sample of original dataset (dependant and independent variables)**

The data types of these attributes are presented in Figure 4. There are total numbers of 22 attributes that are presented as object data type and the remaining 16 are presented as integer or float data types. It provides a clear perspective about which attributes needs to be normalized before applying into the models.



**Figure 4: Attribute data types**

1. *Explanation of attribute values*

This section presents the attributes codes and their description collected for the study. This phase is essential since it assists to identify and use the attributes in the analysis precisely. It is also critical to the success of pre-processing phase. Table 1 presents these attributes and their description.

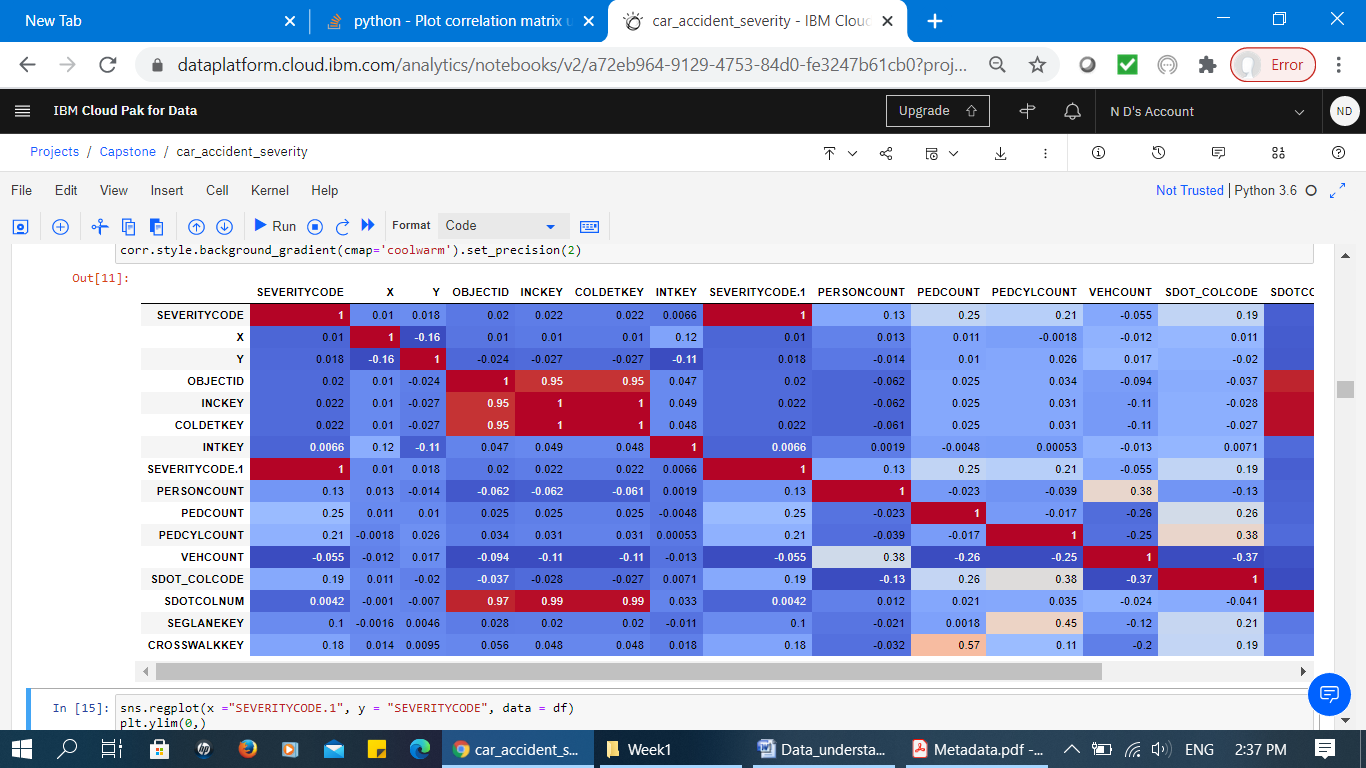
**Table 1: Attributes, vales and descriptions**

| ***Attribute Code*** | ***Description*** |
| --- | --- |
| OBJECTID | ESRI unique identifier |
| SHAPE(X, Y) | ESRI geometry field |
| INCKEY | A unique key for the incident |
| COLDETKEY | Secondary key for the incident |
| ADDRTYPE | Collision address type including; alley, block, intersection |
| INTKEY | Key that corresponds to the intersection associated with a collision |
| LOCATION | Description of the general location of the collision |
| EXCEPTRSNCODE | A code—not known |
| EXCEPTRSNDESC | A code description —not known |
| SEVERITYCODE | A code that corresponds to the severity of the collision: 3—fatality, 2b—serious injury, 2—injury, 1—prop damage, 0—unknown |
| SEVERITYDESC | A detailed description of the severity of the collision |
| COLLISIONTYPE | Collision type |
| PERSONCOUNT | The total number of people involved in the collision |
| PEDCOUNT | The number of pedestrians involved in the collision |
| PEDCYLCOUNT | The number of bicycles involved in the collision. |
| VEHCOUNT | The number of vehicles involved in the collision |
| INJURIES | The number of total injuries in the collision |
| SERIOUSINJURIES | The number of serious injuries in the collision |
| FATALITIES | The number of fatalities in the collision |
| INCDATE | The date of the incident |
| INCDTTM | The date and time of the incident |
| JUNCTIONTYPE | Category of junction at which collision took place |
| SDOT\_COLCODE | A code given to the collision by SDOT |
| SDOT\_COLDESC | A description of the collision corresponding to the collision code |
| INATTENTIONIND | Whether or not collision was due to inattention. (Y/N) |
| UNDERINFL | Whether or not a driver involved was under the influence of drugs or alcohol |
| WEATHER | A description of the weather conditions during the time of the collision |
| ROADCOND | The condition of the road during the collision |
| LIGHTCOND | 300 The light conditions during the collision |
| PEDROWNOTGRNT | Whether or not the pedestrian right of way was not granted. (Y/N) |
| SDOTCOLNUM | A number given to the collision by SDOT |
| SPEEDING | Whether or not speeding was a factor in the collision. (Y/N) |
| ST\_COLCODE | A code provided by the state that describes the collision. |
| ST\_COLDESC | A description that corresponds to the state’s coding designation |
| SEGLANEKEY | A key for the lane segment in which the collision occurred |
| CROSSWALKKEY | A key for the crosswalk at which the collision occurred |
| HITPARKEDCAR | Whether or not the collision involved hitting a parked car. (Y/N) |

1. *Data Redundancy analysis*

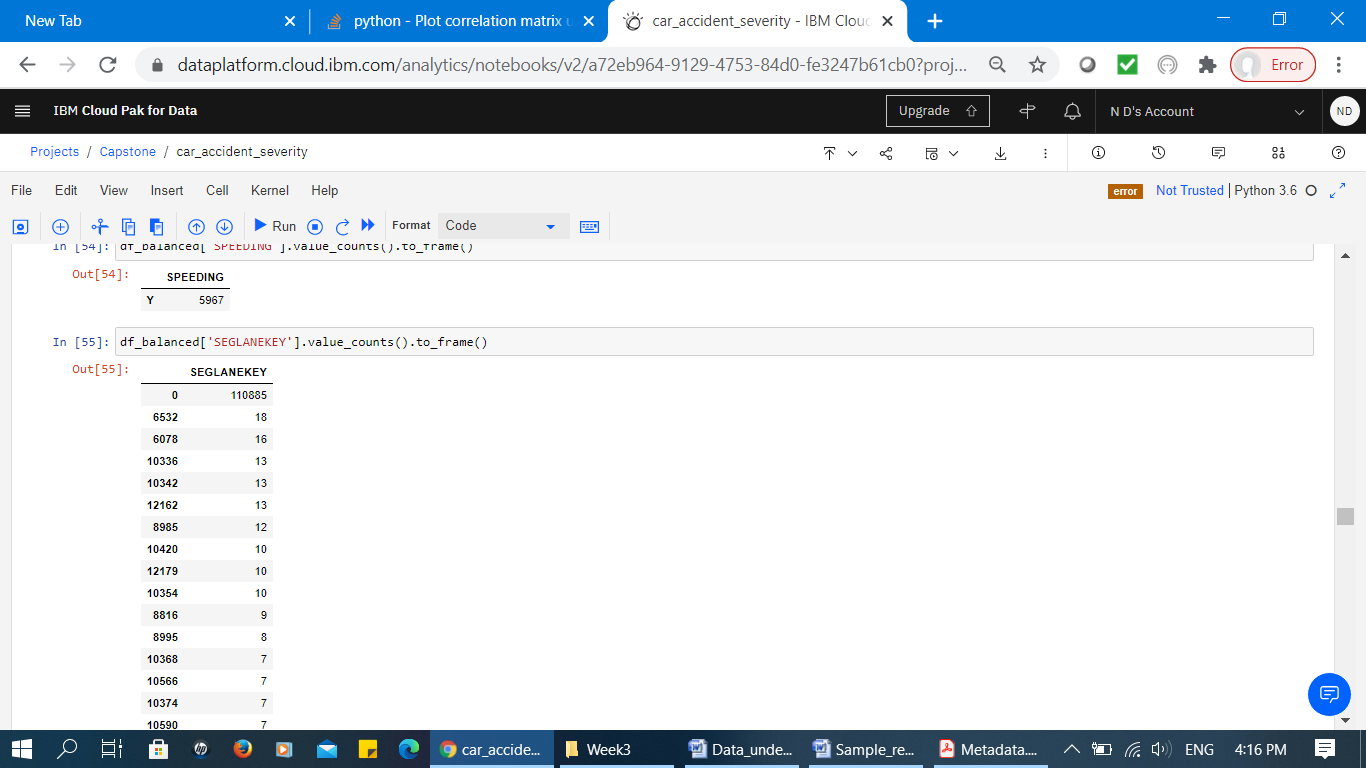
This section presents the data redundancy/ data incompleteness analysis based on the result of correlation analysis. Basically, correlation analysis provides a brief overview of the original data. Figure 5 visualizes a portion of this analysis. The red sections present the existence of highly correlated attributes. These attributes are “OBJECT ID”, “INCKEY” and “COLDETKEY”. According to the description of these attributes presented in Table 1, these are the unique keys which eventually has no effect in prediction of car severity accident. Therefore, they will be excluded from the analysis.

Moreover, “SEVERITYCODE.1” which is highly correlated to “SEVERITYCODE” found to be a duplicate attribute. We believe “SEVERITYCODE.1” is redundant and it needs to be excluded from the analysis.

**

**Figure 5: Visualizing Data Correlation Analysis**

SDOTCOLNUM (number given to the collision by SDOT) is also highly correlated to the mentioned key numbers. However, the coding of this attribute is not provided by experts. Moreover, the number by itself does not provide any meaningful information. Therefore, it is also excluded from the analysis (Figure 6).



**Figure 6: Portion of “SDOTCOLNUM” Attribute values and its counts**

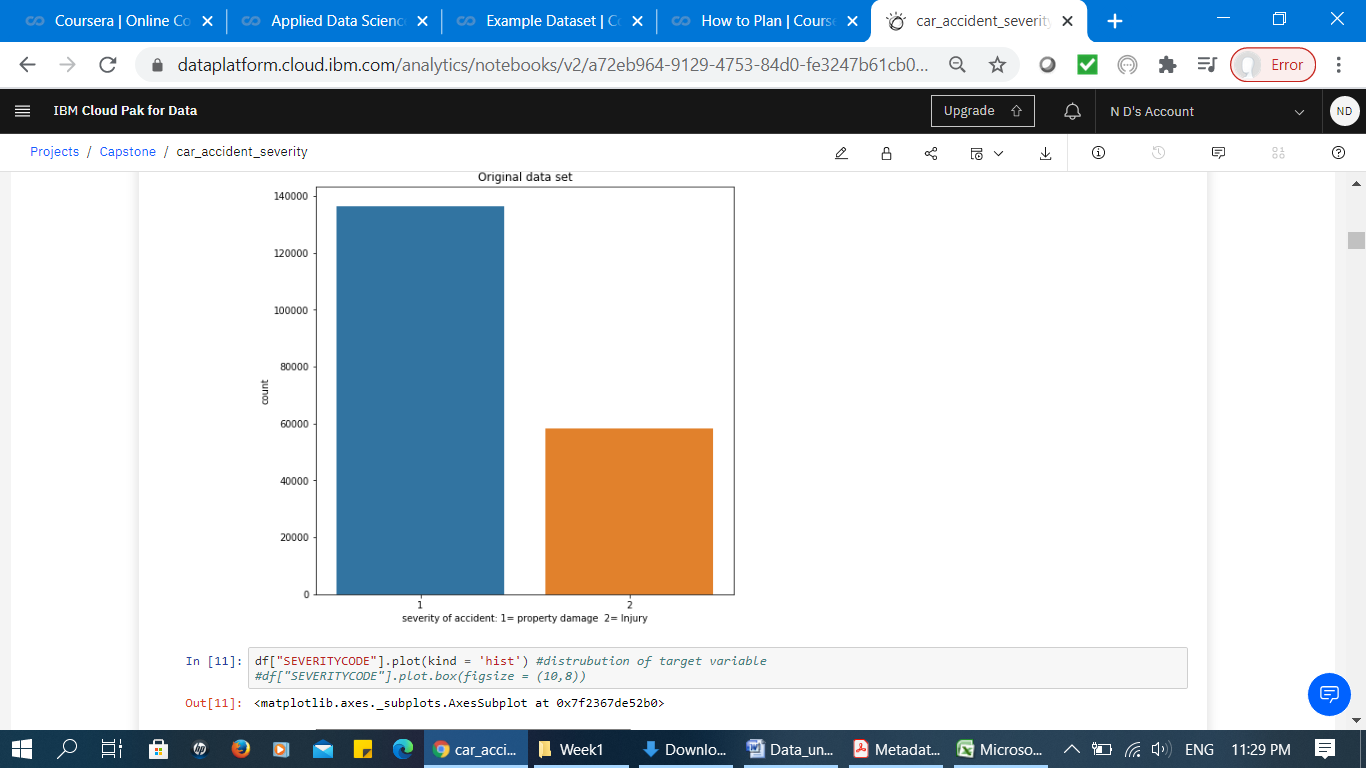
Since the data is not pre-processed yet, the correlation analysis does not present very strong correlations. This indicates that data need to be cleaned first. We will pre-process the data and present the results in the subsequent sections.

**3. Data Preparation**

The data preparation is a critical task of data pre-processing because it mainly includes all the required activities to construct the final dataset which will be fed into the modelling tools. The data preparation section aims to develop a clean dataset. In this section, the process of balancing the labelled data, handling missing data, data standardization and all the required feature engineering tasks for some attributes are explained.

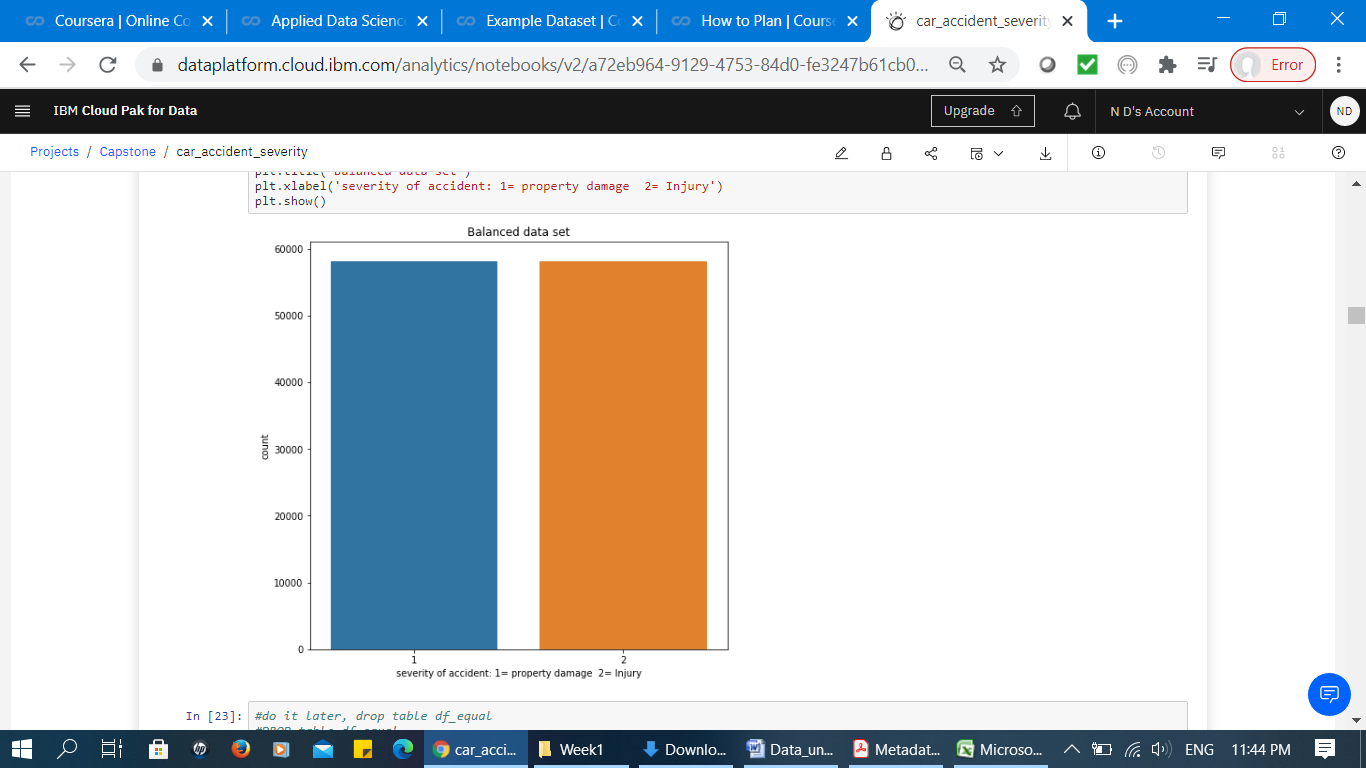
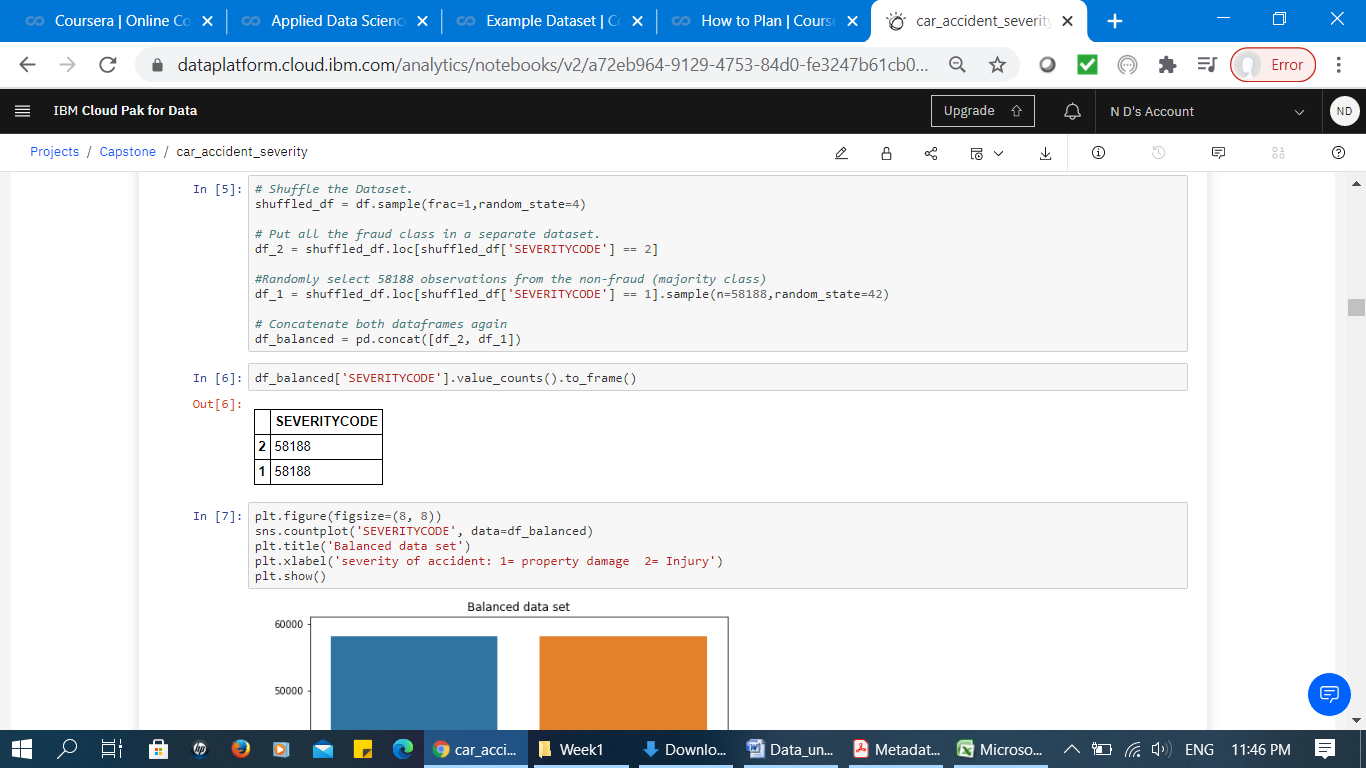
1. ***Balancing the Labelled Data***

To identify whether the data is in a good state used for the modelling algorithm, we have analyzed the data, attributes and their values. Figure 7 illustrates the distribution of dependant variable in respects to its values; property damage and injury (code 1 and 2 respectively). It presents that the data has unbalanced labels. The unbalanced labels potentially create a biased machine learning model. Therefore, the first task is to balance the data according to the labels.



**Figure 7: Dependant variable values before balancing**

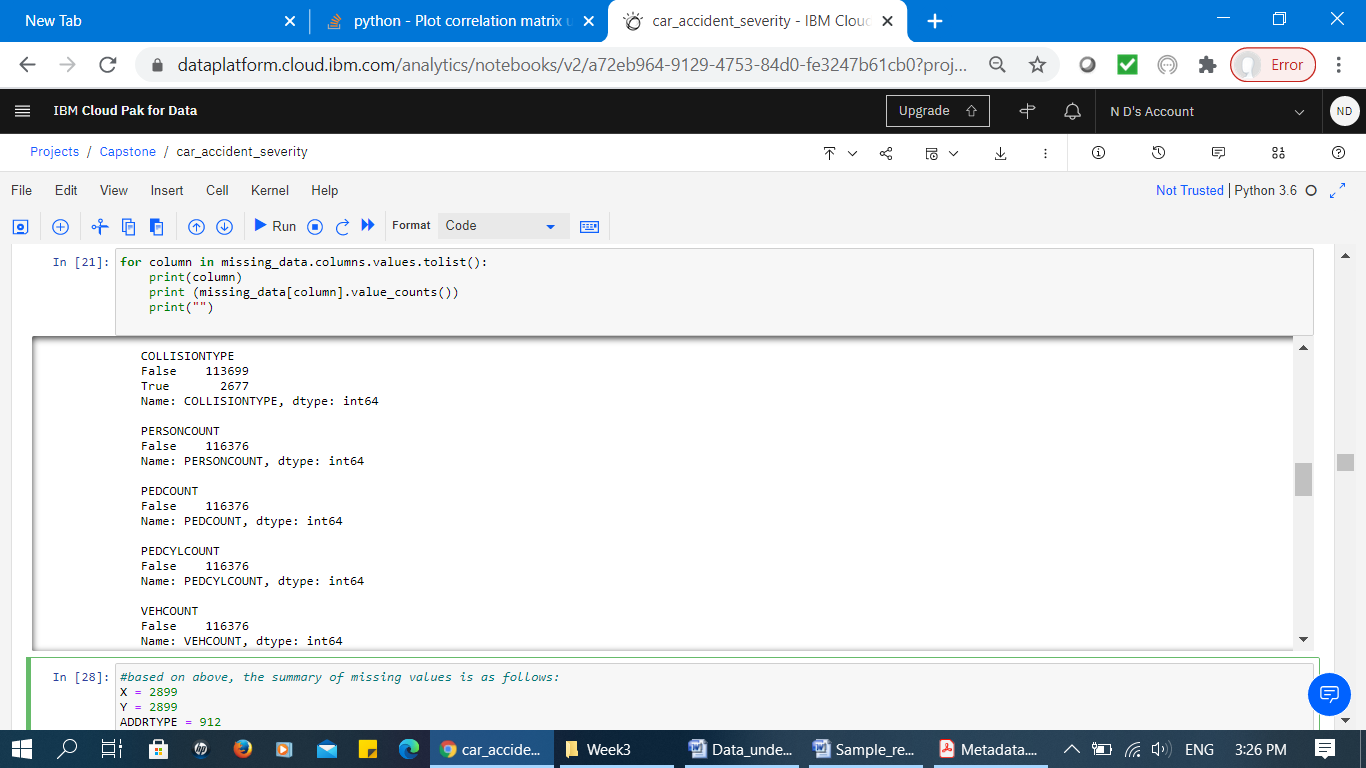
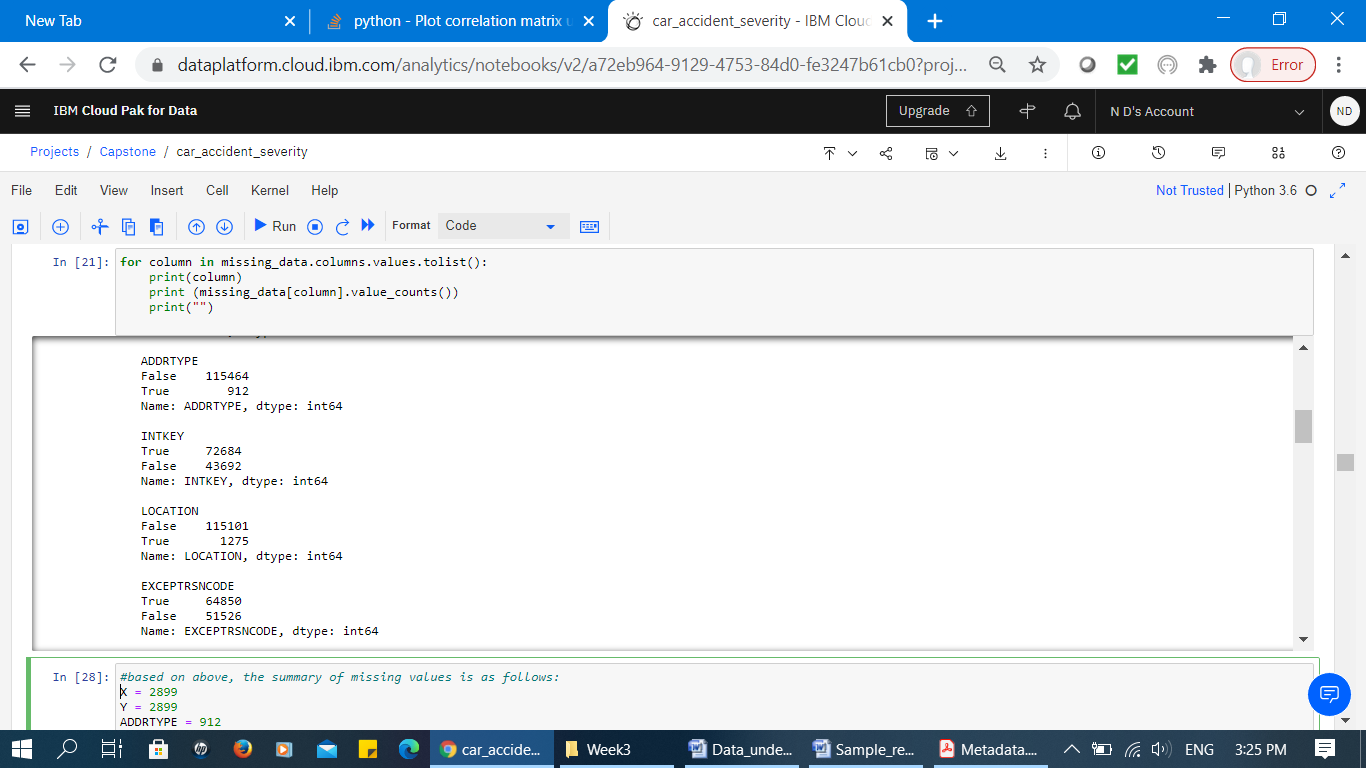
The following figure presents how the data is balanced corresponds to the two code values 1 and 2. As presented next, there are exactly 58188 numbers of labels for each category in the dataset.



**Figure 8: Balanced dataset**

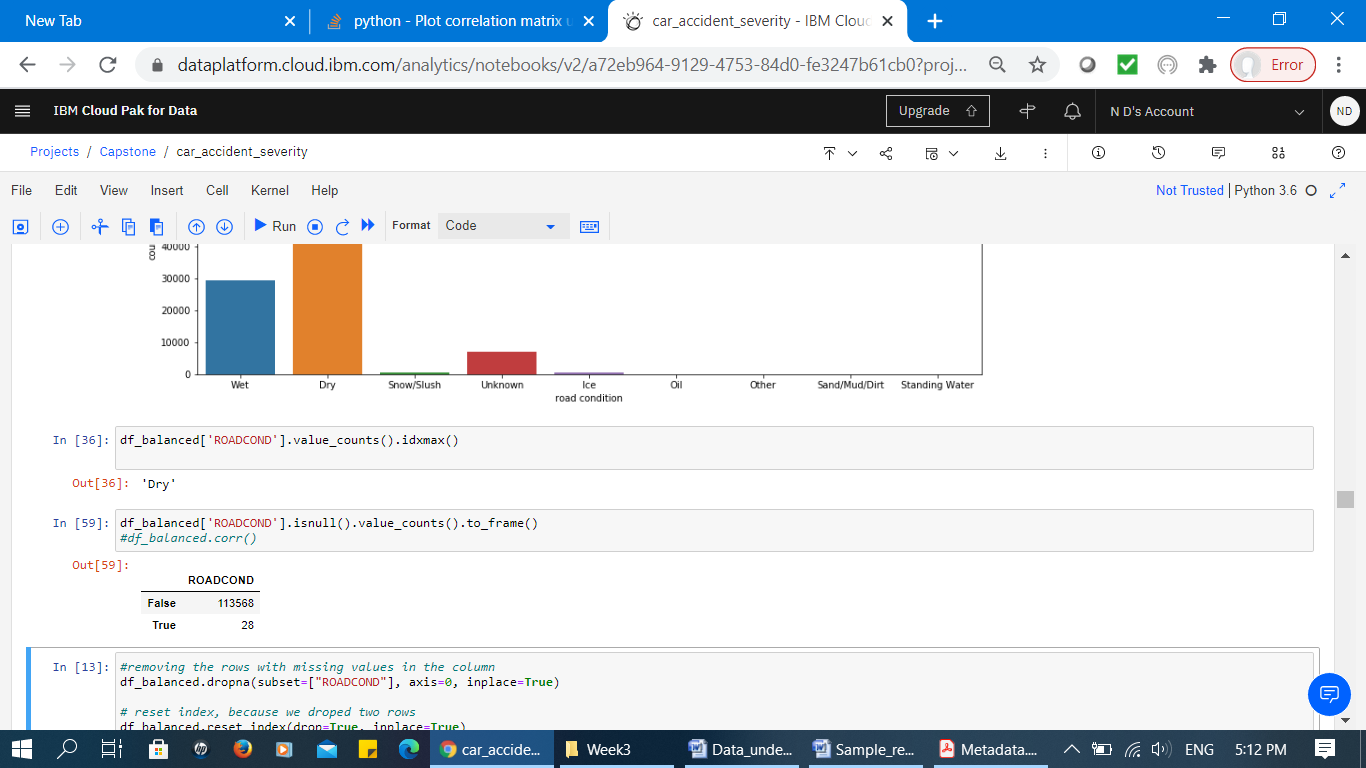
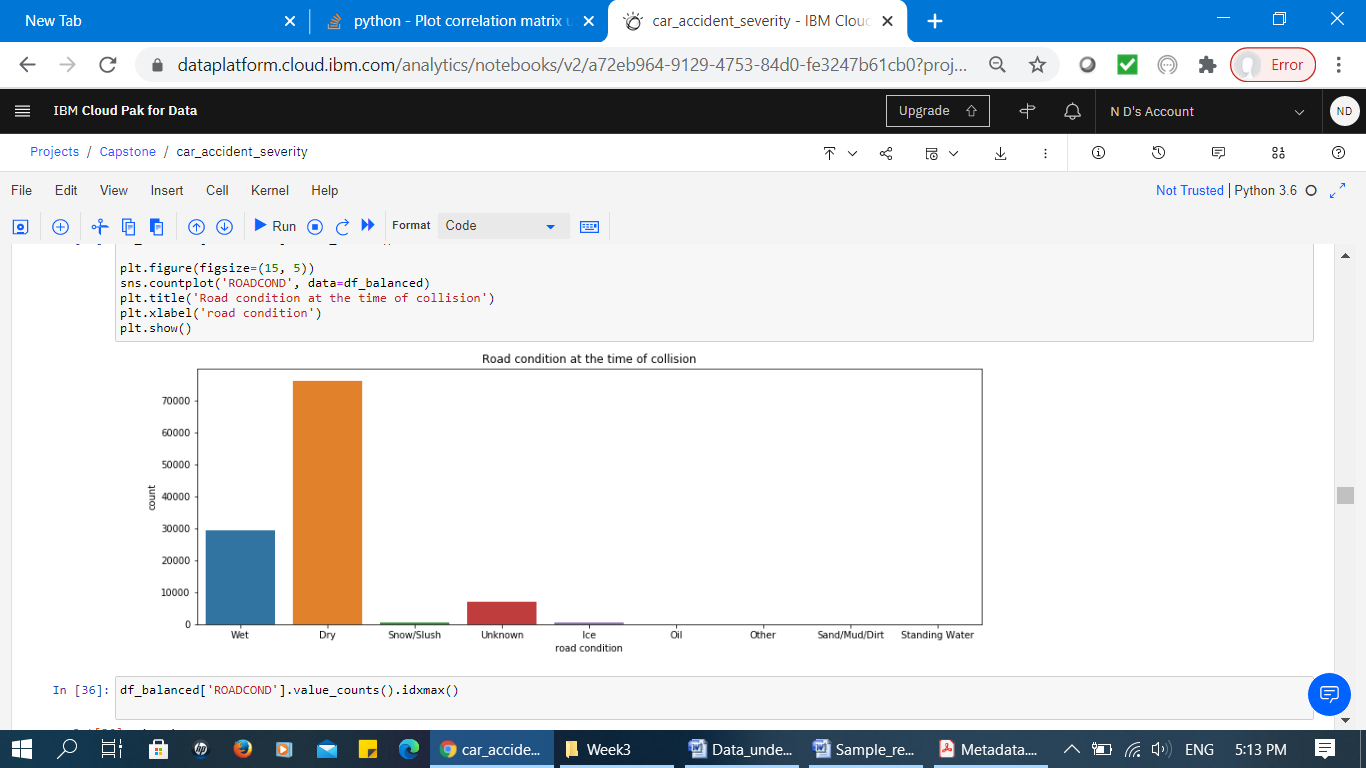
1. ***Handling Missing Data***

In this section, the attributes and their values are analysed to find the availability of missing values. Figure 9 presents a sample of codes illustrating the number of missing values for the attributes (True presents the number of missing values).



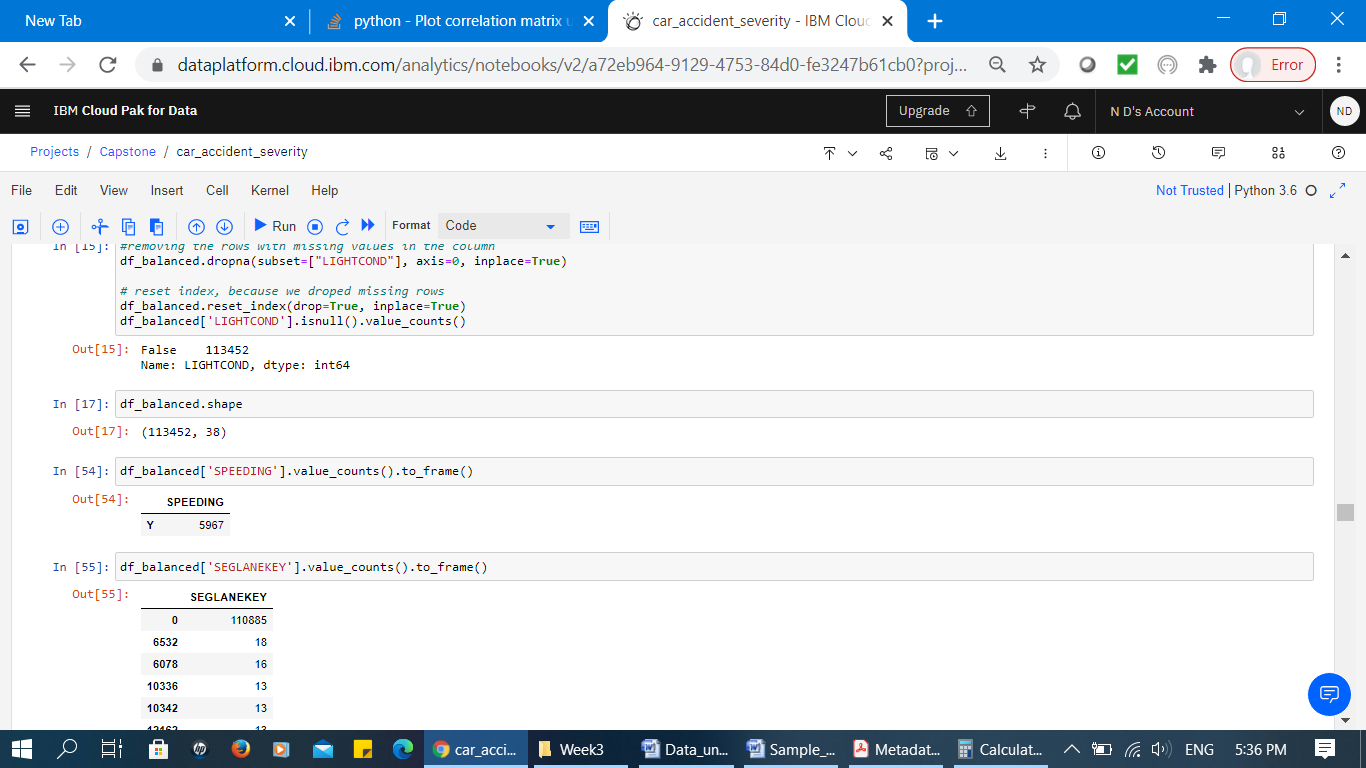
**Figure 9: Number of Missing Values (Corresponds to True value)**

Next, all the attributes are evaluated individually to identify the quality of the attribute values. The result identifies that the missing values for the possible predictable attributes are removed. For example, the attribute “ROADCOND” which presents the condition of road at the time collision has less than 0.1% missing values, so the rows related to the missing values are excluded from dataset (Figure 10). Similarly the attribute “LIGHTCOND” which includes 0.1% missing values, the related missing rows are excluded.



**Figure 10: Left diagram presents attribute values “ROADCOND” and right diagram presents the number of missing values (28 missing value)**

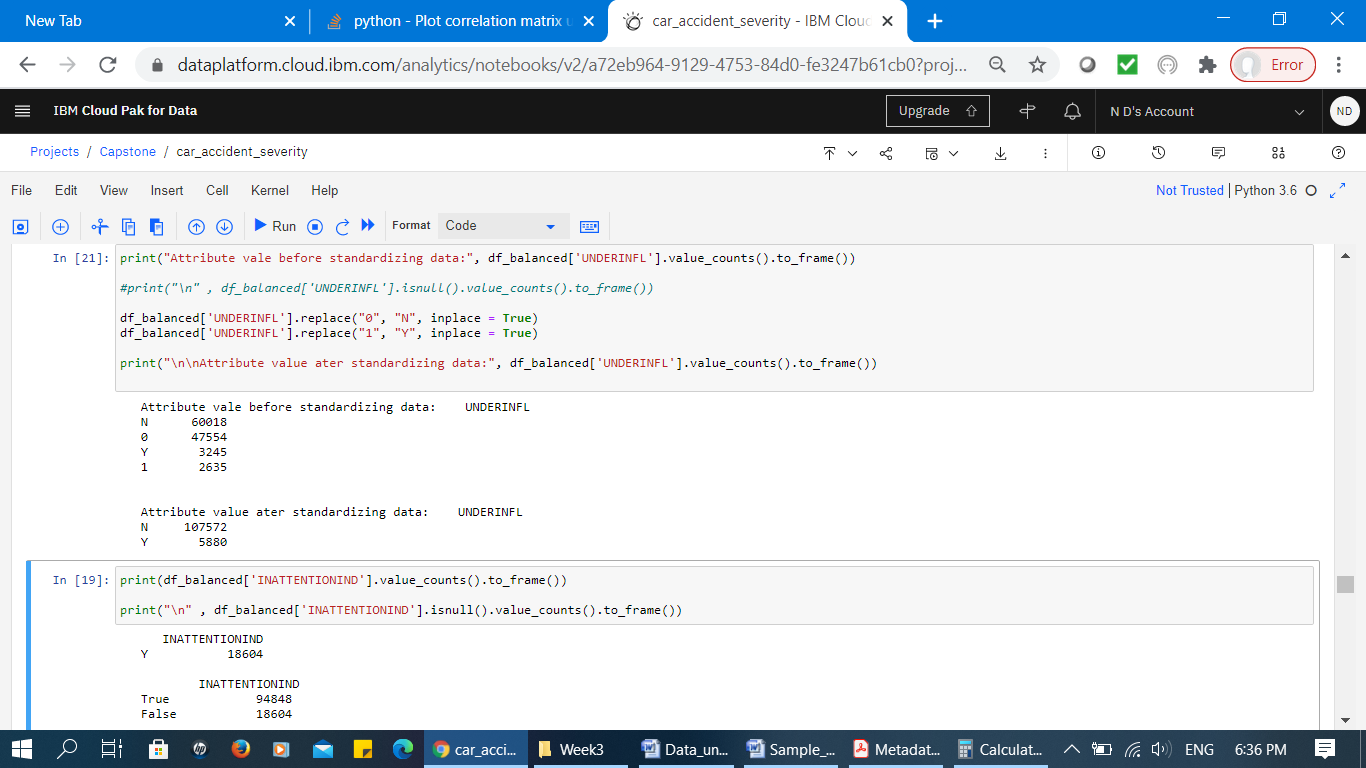
Attribute “SPEEDING” which presents whether or not speeding was a factor in the collision presented as one values “Y” (Figure 11). The corresponding missing values don’t confirm whether they are missed or speed is not a factor of collision. Therefore due to lack of expert knowledge, we have to exclude it from analysis.

**

**Figure 11: Attribute “SPEEDING” value**

1. ***Data Standardization and Encoding***

The attribute “UNDERINFL” which presents whether or not a driver involved was under influence of drugs or alcohol is presented as “N”, “Y”, “0” and “1” (Figure 12). We assume “0” means “N” and “1” means “Y”. Therefore the values are standardized to form a consistent data attribute.

******

**Figure 12: Attribute “UNDERINFL” - Before and After Data Standardization**

Similarly, attribute “INATTENTIONIND” which presents whether or not collision was due to inattention is presented as “Y”. We assume the missing values are not related to being inattention. Therefore, we encoded the missing values to “N”.

**REFERENCES**

[2] The CRISP-DM process model (1999), http://www.crisp-dm.org/