**Classification of Road Accident Severity**

**G Chan**

**14 Sep 2020**

|  |  |
| --- | --- |
| **Introduction** | 2 |
| **Data Understanding** |  |
| Data Sources | 2 |
| Target Variable Definition | 3 |
| Data Selection | 3 |
| Data Manipulation | 4 |
| **Methodology** | 5 |
| **Modelling** | 5 |
| **Results** | 6 |
| **Discussion** | 6 |
| **Conclusion** | 6 |
|  |  |
|  |  |

**1. Introduction**

**1.1 Background**

Road accident can occur anytime and anywhere and it may cause heavy casualties.

The cause of traffic accident could be attributed to various factors such as weather, road quality , drivers and pedestrian’s awareness

To minimize the occurrence, we may leverage a machine learning model to classify the severity of the road accident to remind public

**1.2 Problem**

Data that might contribute to determining road safety might cover weather, road quality, drivers

and pedestrian’s awareness. This project is to classify road accident severity.

**1.3 Interest**

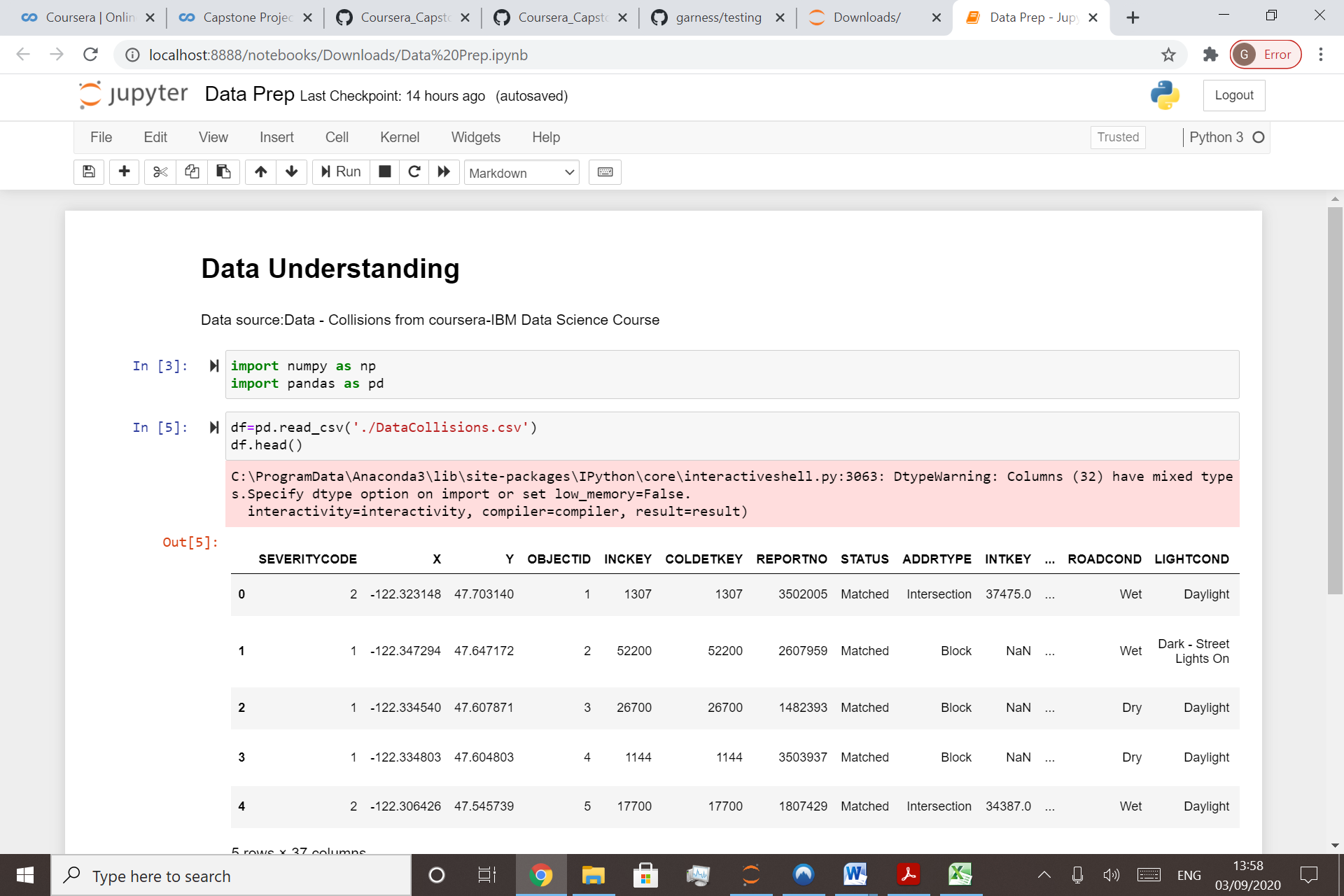
Transportation Bureau , drivers and public would be very interested in indicative information about road accident severity . Maths , Statistic major students and data scientist would also be interested in the development of classification model

**2. Data Understanding**

**2.1 Data Sources**

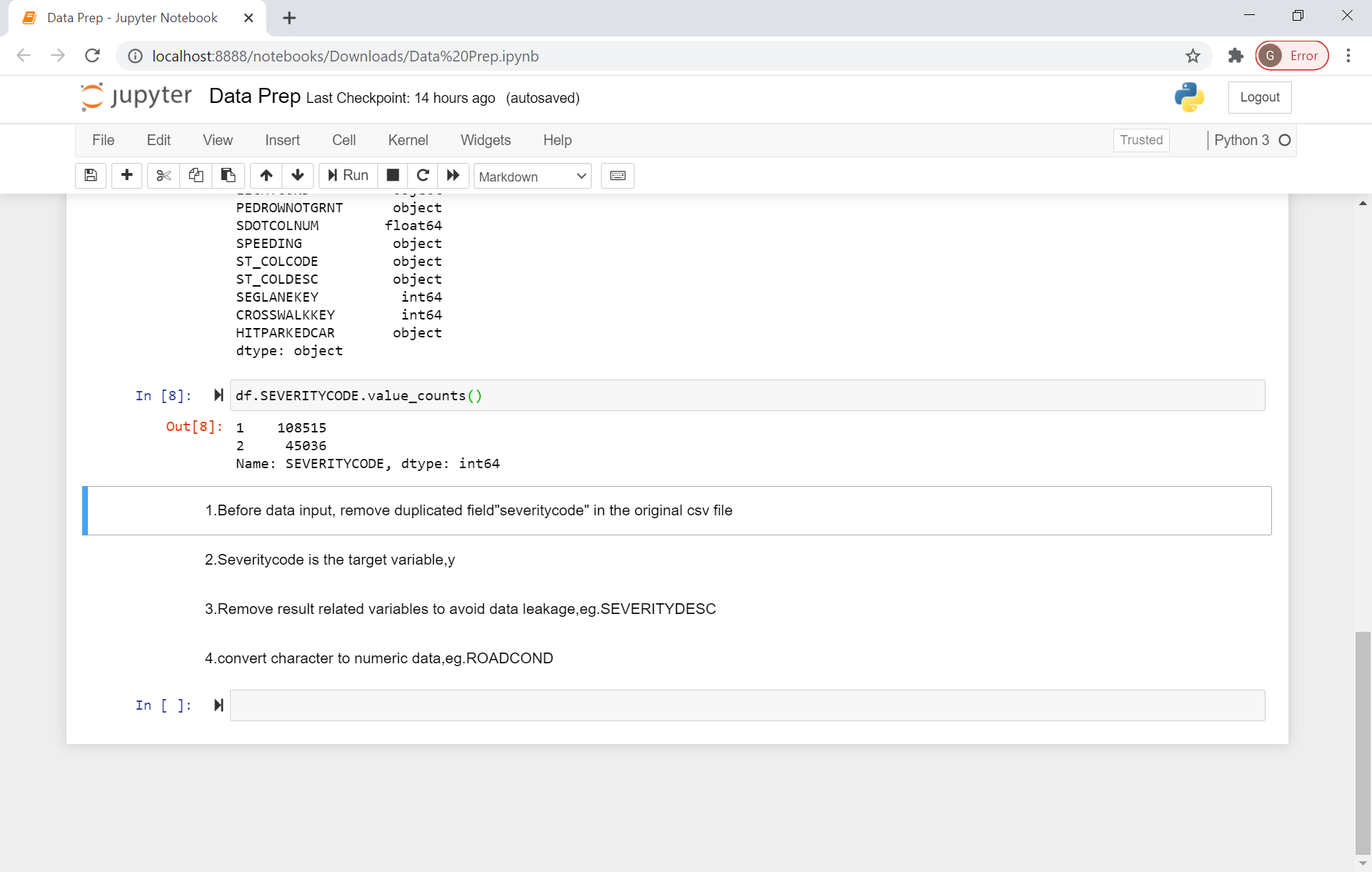
The data is source by public transportation data for Seattle city

Link: https://github.com/garness/testing/blob/master/Data-Collisions.zip



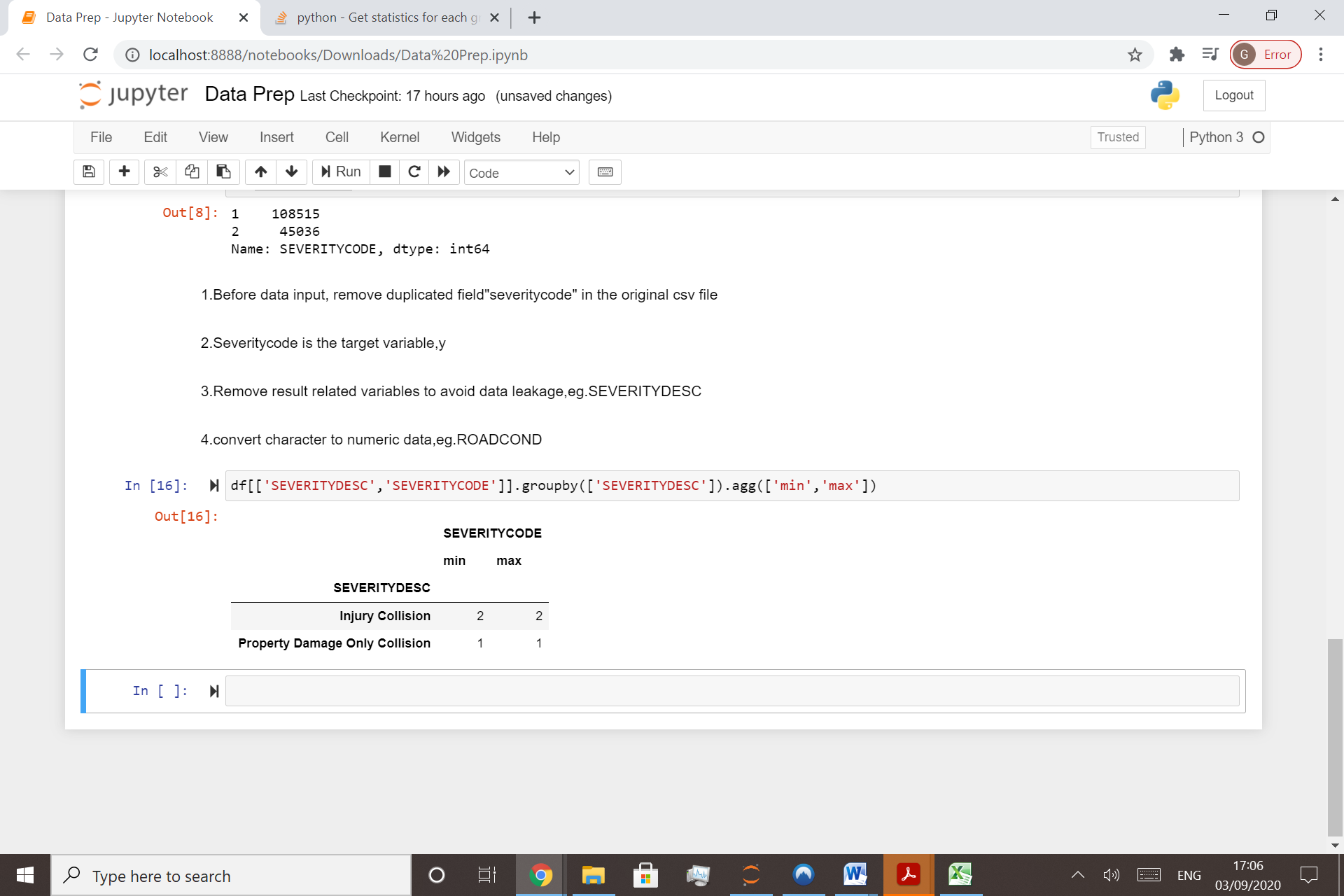
**2.2 Target Variable Definition**

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

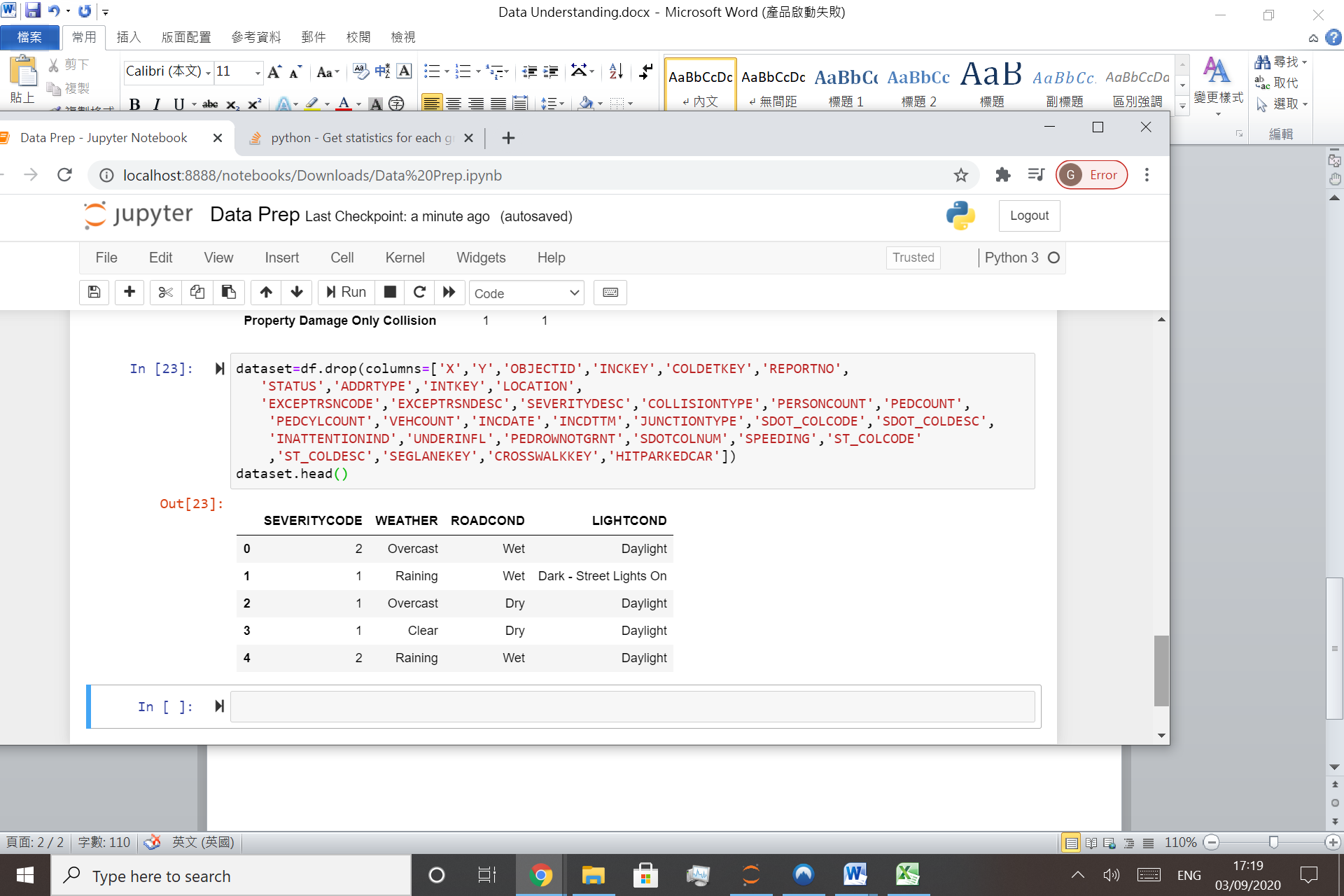


**2.3 Data Selection**

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

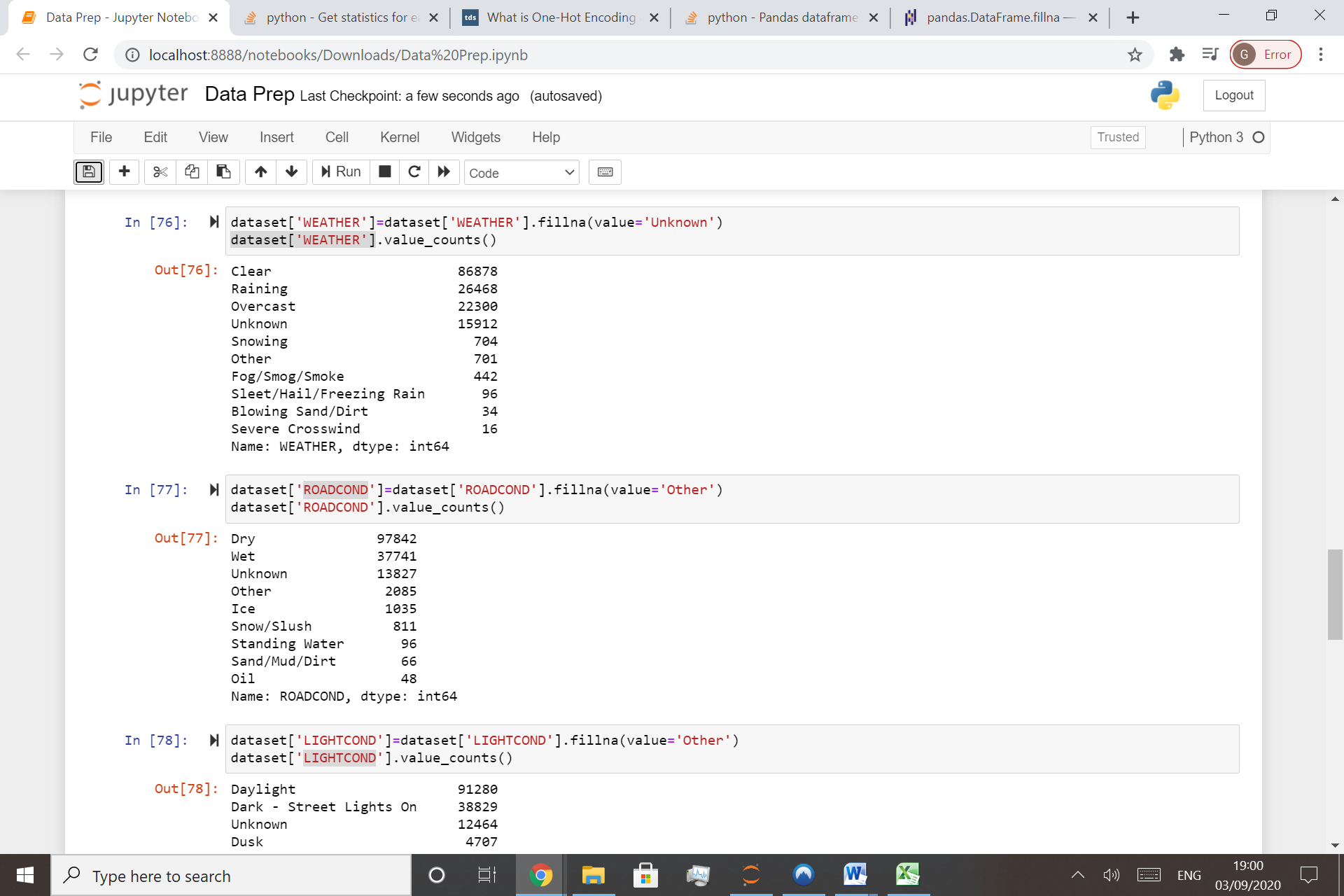


Other meaningless variable are also removed

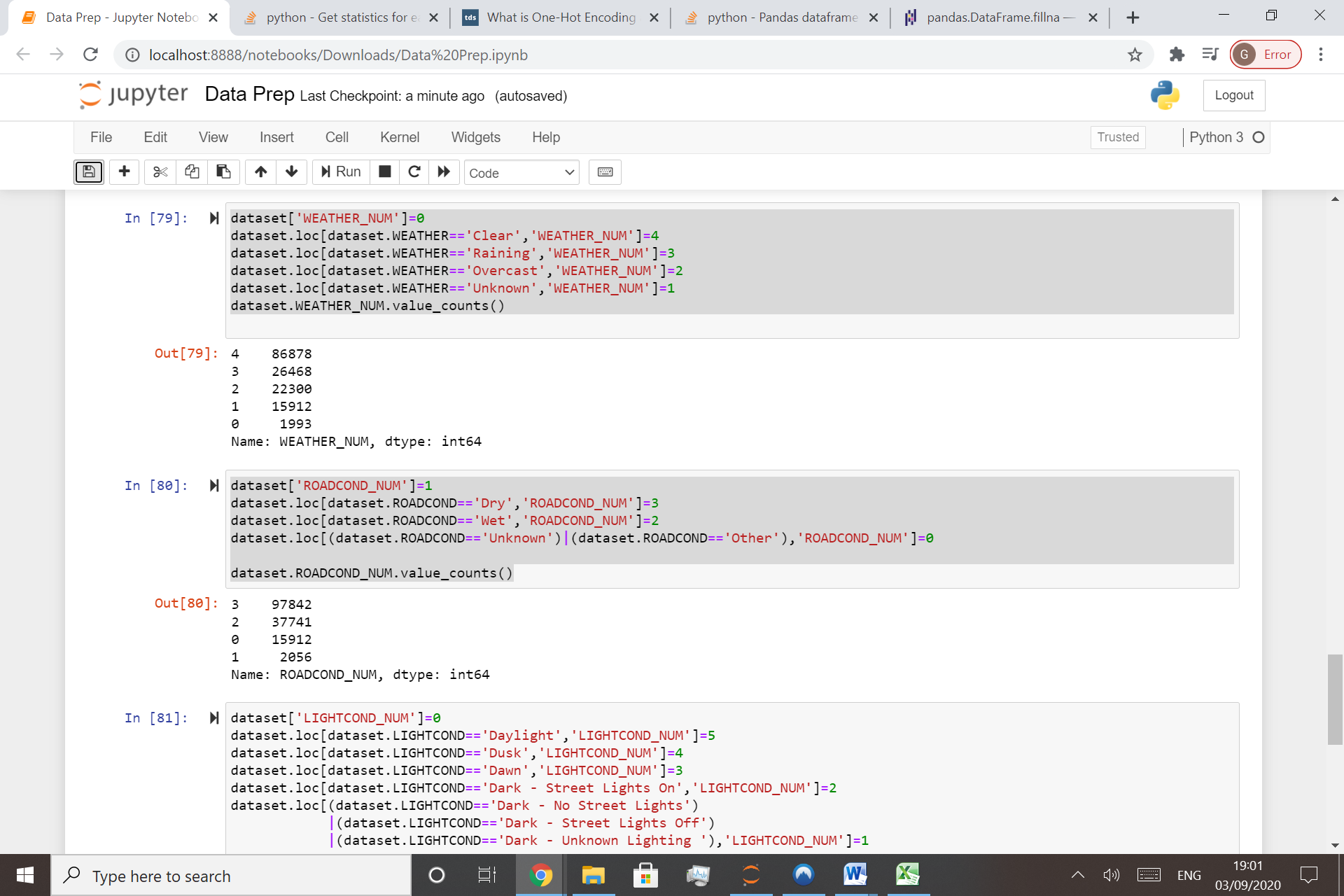


**2.4 Data Manipulation**

Assign missing to the field Weather,Roadcond,lightcond

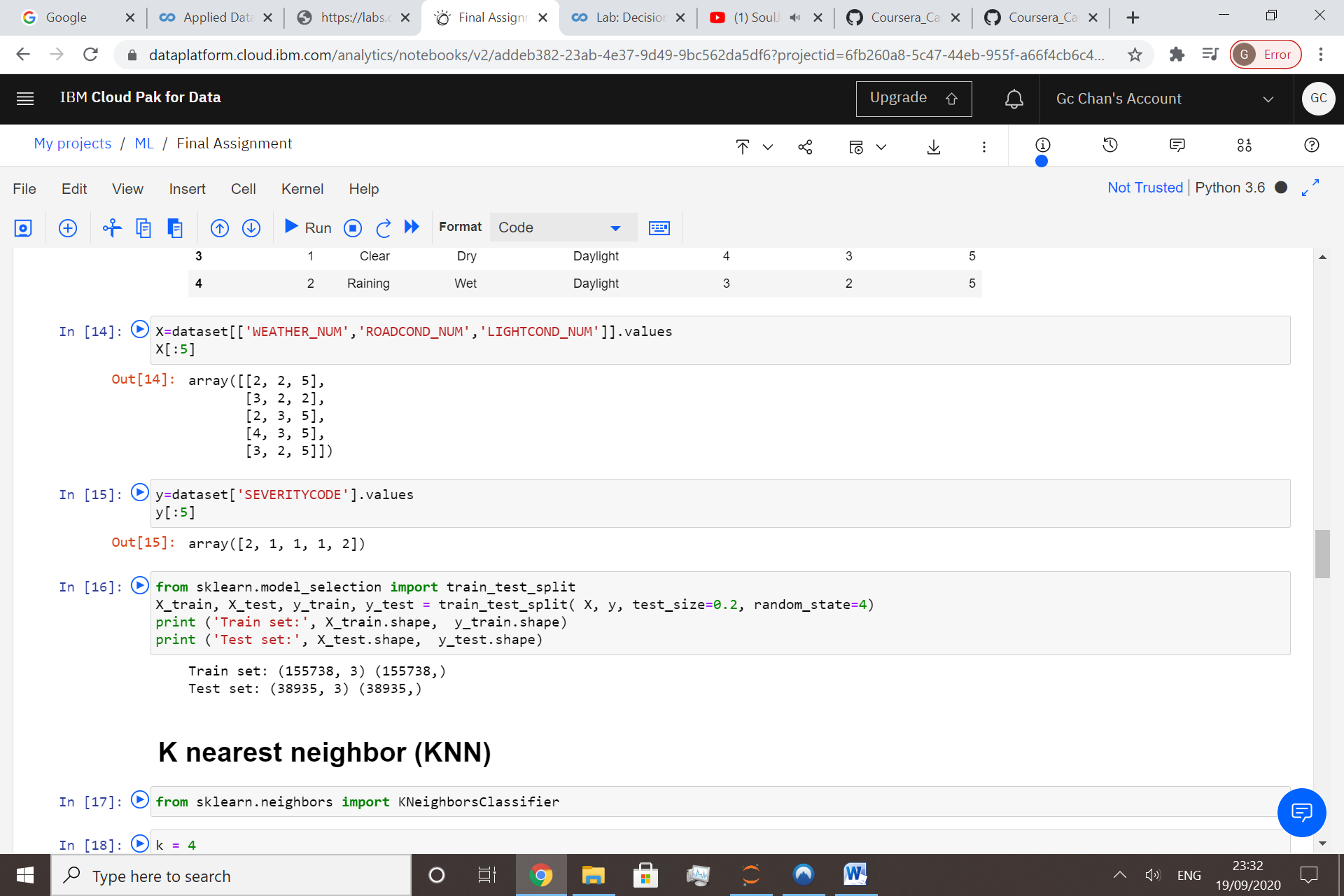


Convert the string to number



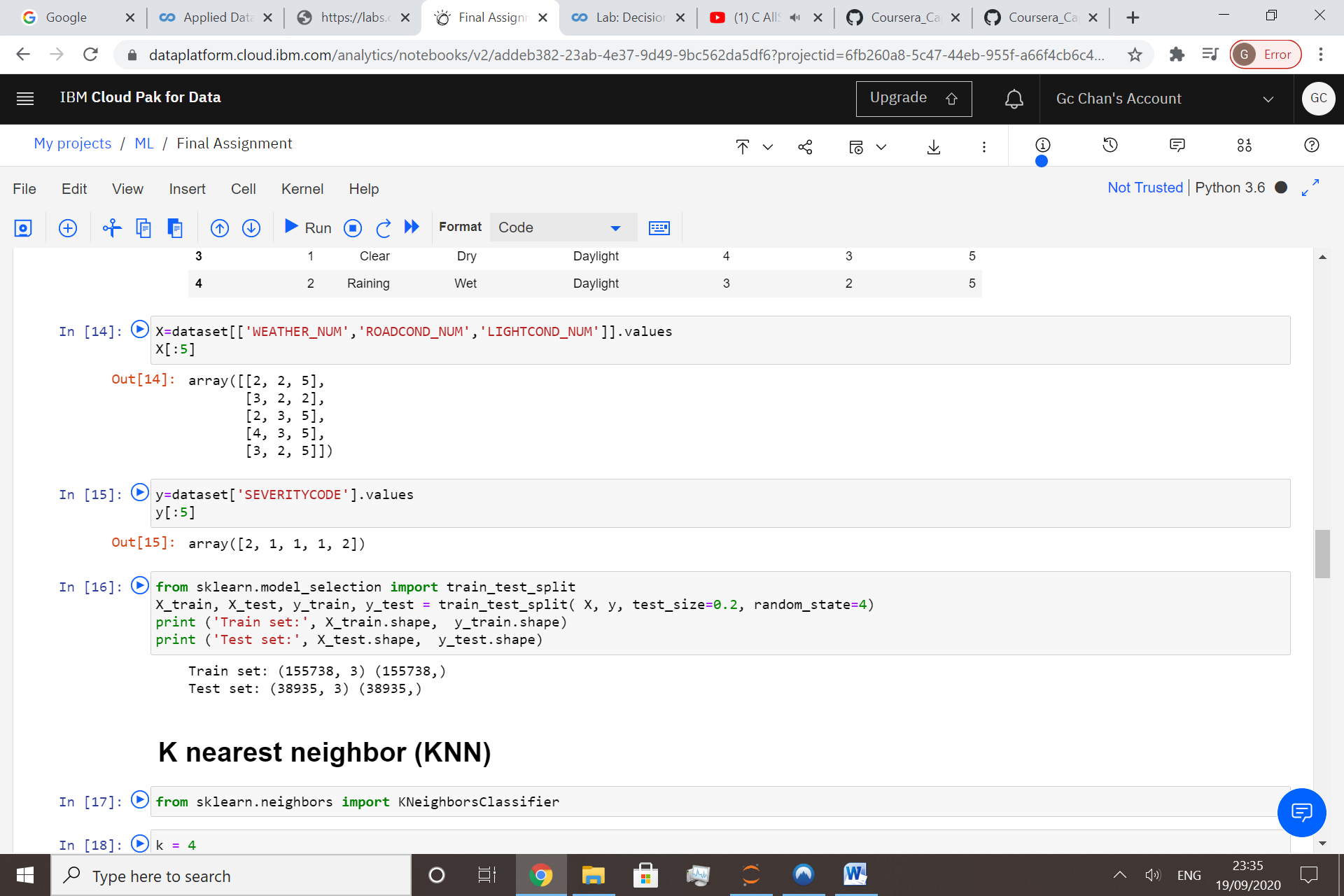
**3. Methodology**

For model development, Variables X and Y are necessary



As no continuous variables are involved, no need to apply normalization

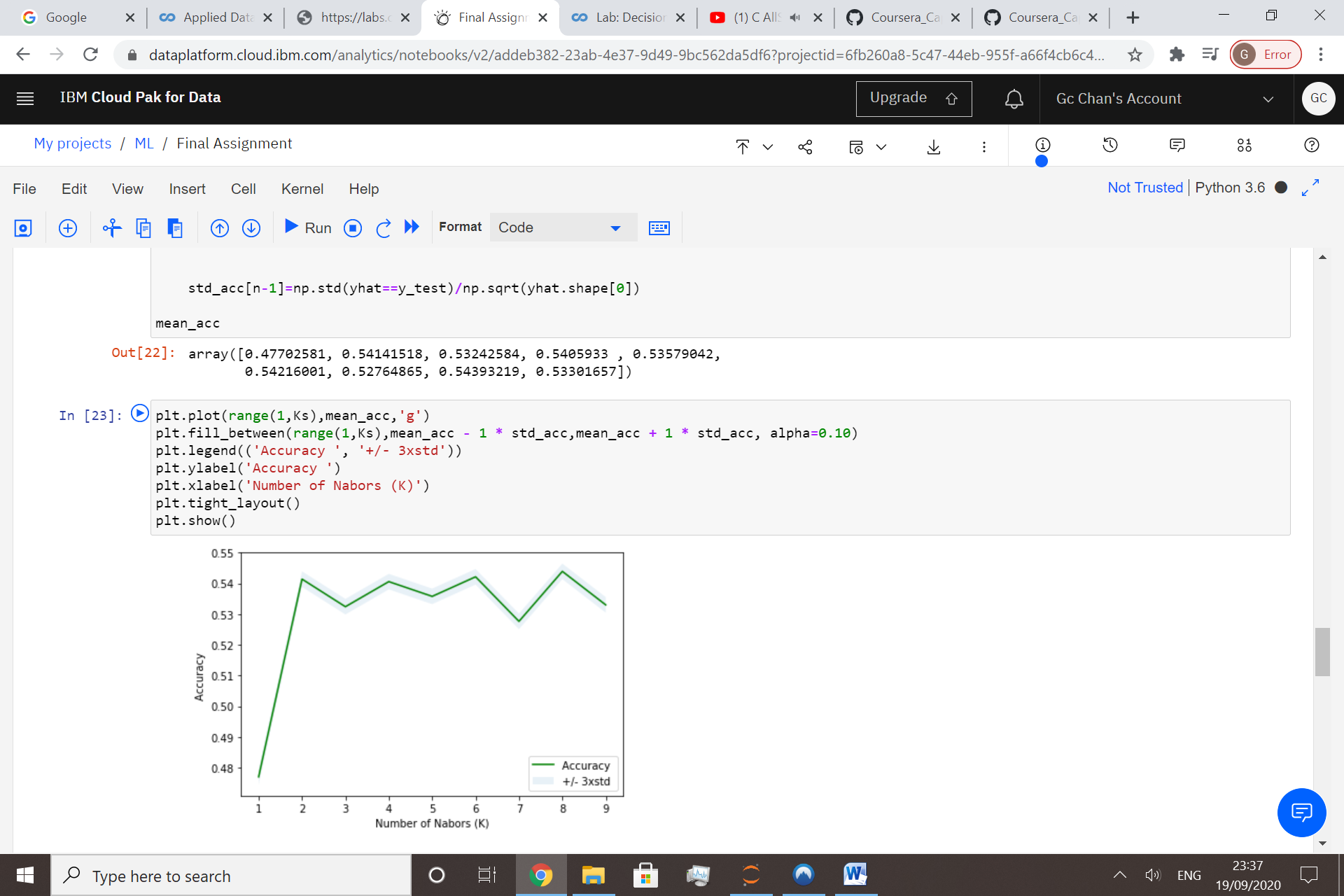
Split the data into training set(70%) and testing set(30%)

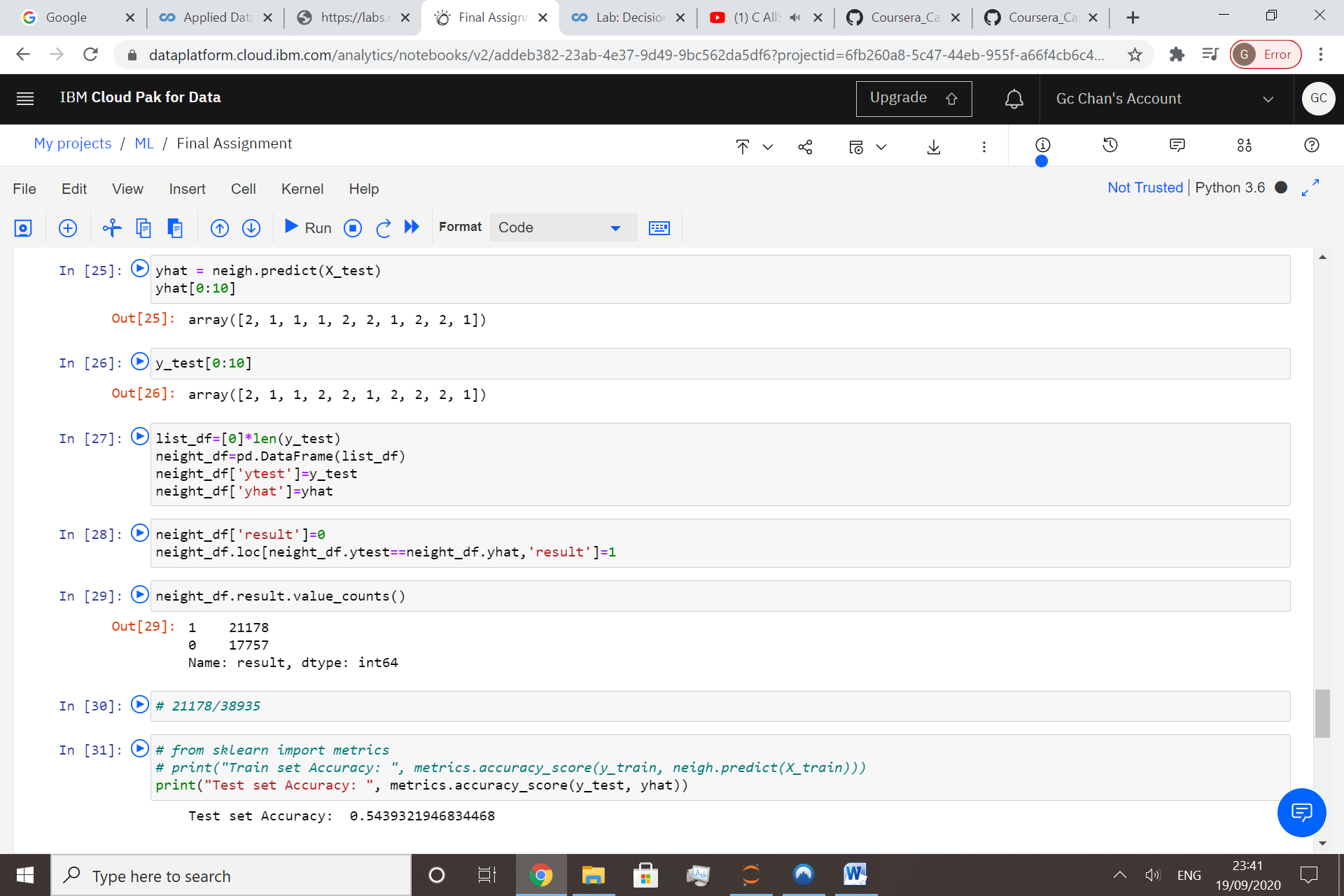


**4. Modelling**

KNN would be used as machine learning model developement

**K-Nearest Neighbors** is an algorithm for supervised learning. Where the data is 'trained' with data points corresponding to their classification. Once a point is to be predicted, it takes into account the 'K' nearest points to it to determine it's classification.



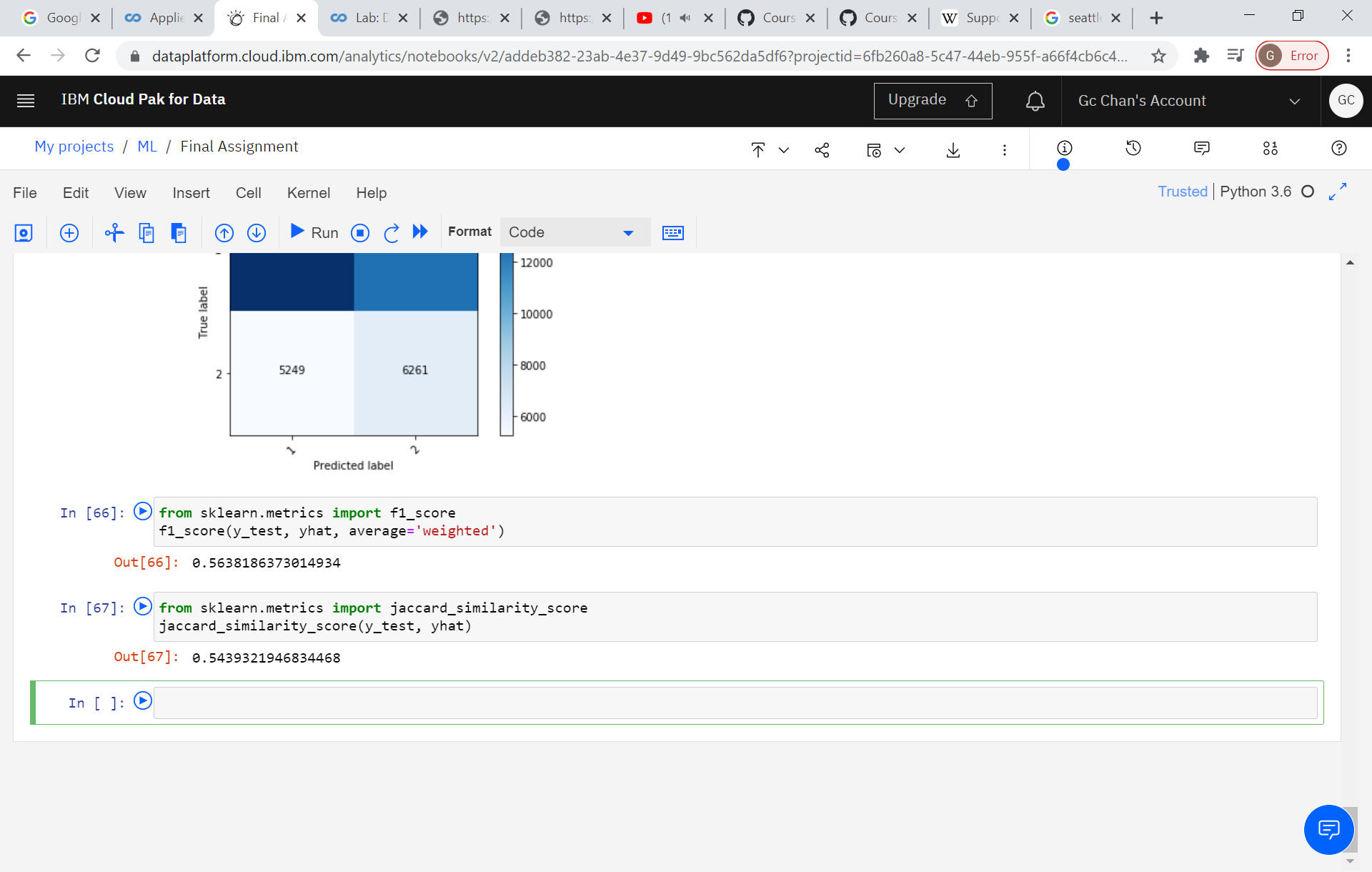


When K=8, the accuracy reached the highest, accuracy rate around 55%

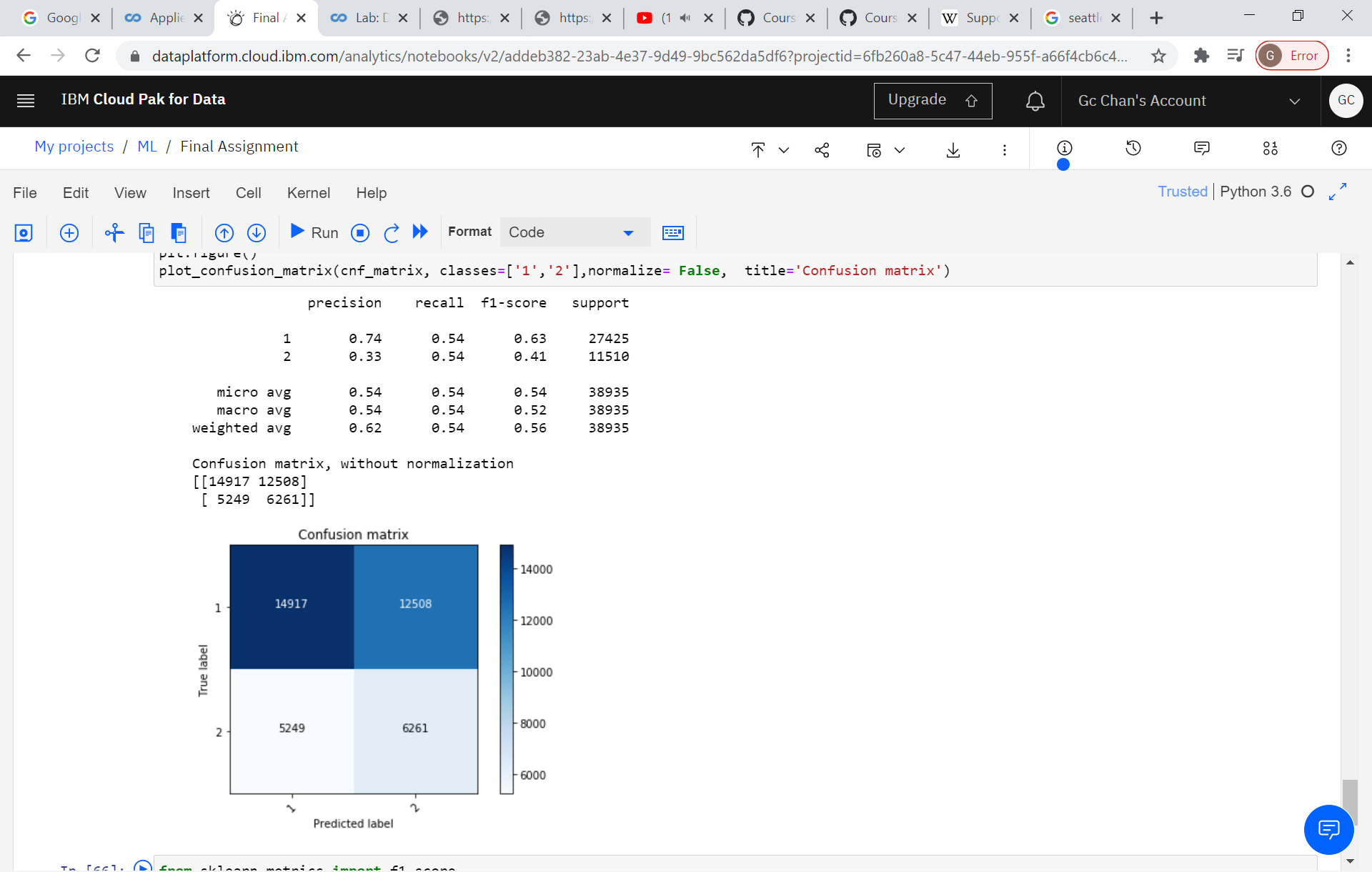
**5. Result**

Apart from accuracy, there are other measurement about model performance.

1. Jaccard index: Similarity for the two sets of data , with a range 0% to 100%
2. F1 score:Model accuracy on a dataset



Confusion matrix:



**6. Discussion**

1. The suitable data is limited. IF more data such as car type, drivers’ condition, car speed are available in model development, there is a change the accuracy can be enhanced
2. Suggest number of KNN be limited within 10 to avoid wide range of group. It may be too difficult to apply in practical implement if number of K is too large

7**. Conclusion**

KNN with K=8 is selected to be a classification model about road severity in Seattle city