**Capstone Project: Car accident severity**

**Introduction**

Car accident are one of the leading causes of death in US, the consequences may range from minor injury/vehicle damage to major personal injury or even death. It is therefore vital for us to understand which are the factors influencing the likelihood of a car accident occurring at certain location, as well as the factors which influence the severity of these car accidents.

Factors likely to influence the likelihood and severity of car accident may include: weather, road condition, timing in the day, and traffic congestion. Additional factors may include driver’s personal condition, e.g. consumption of alcohol, tiredness etc.

The report aims to provide understanding of the factors influencing the frequency and severity of car accidents. These insights may provide reference for city planners and emergency service providers when allocating road design/medical response resources.

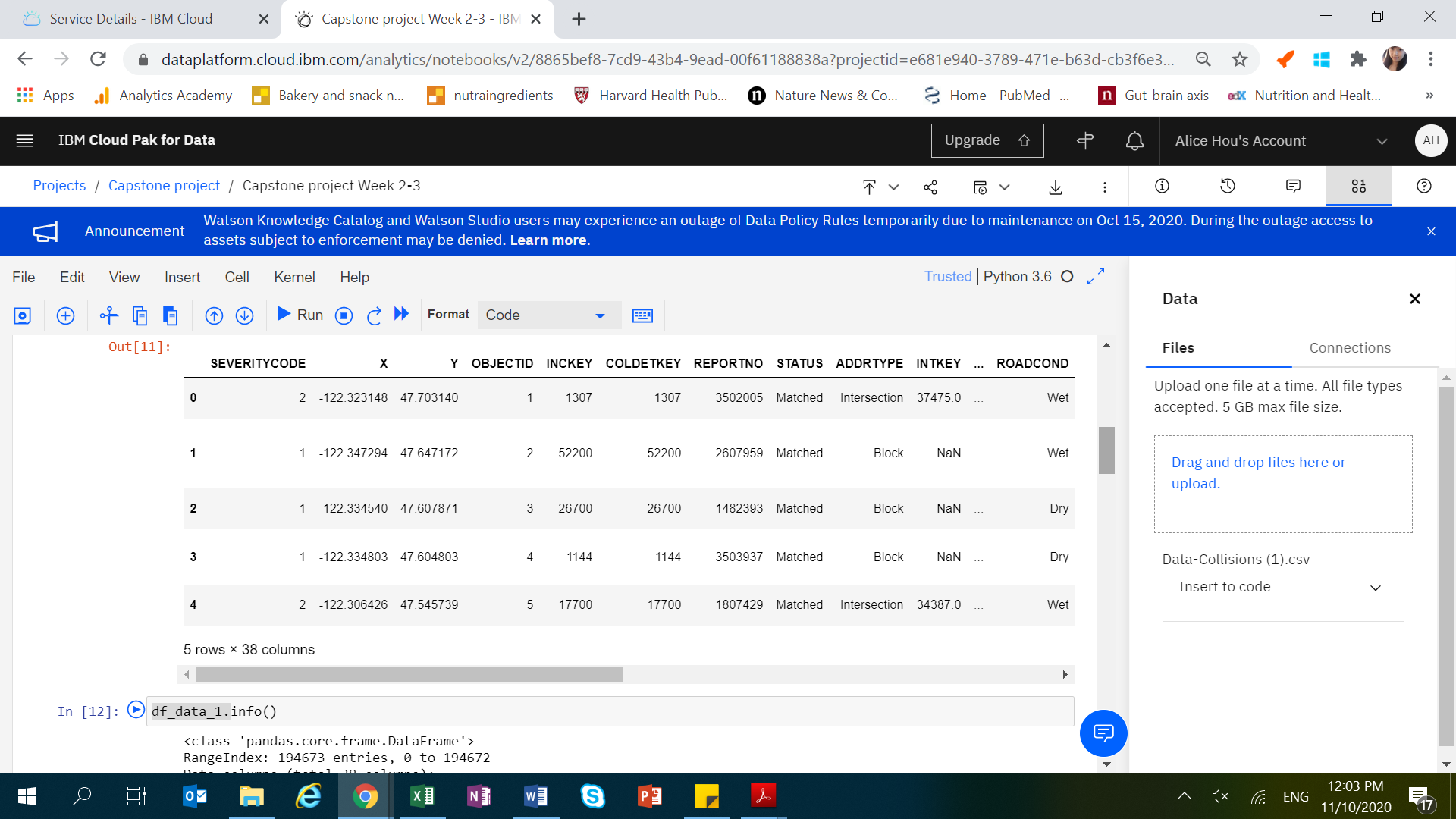
**Data description**

The data is provided by SDOT Traffic management division which include all car collisions from 2004 to present in Seattle and has been updated weekly. The data has been downloaded and read into Pandas Dataframe, and the content and data types displayed using the HEAD and DTYPES functions.

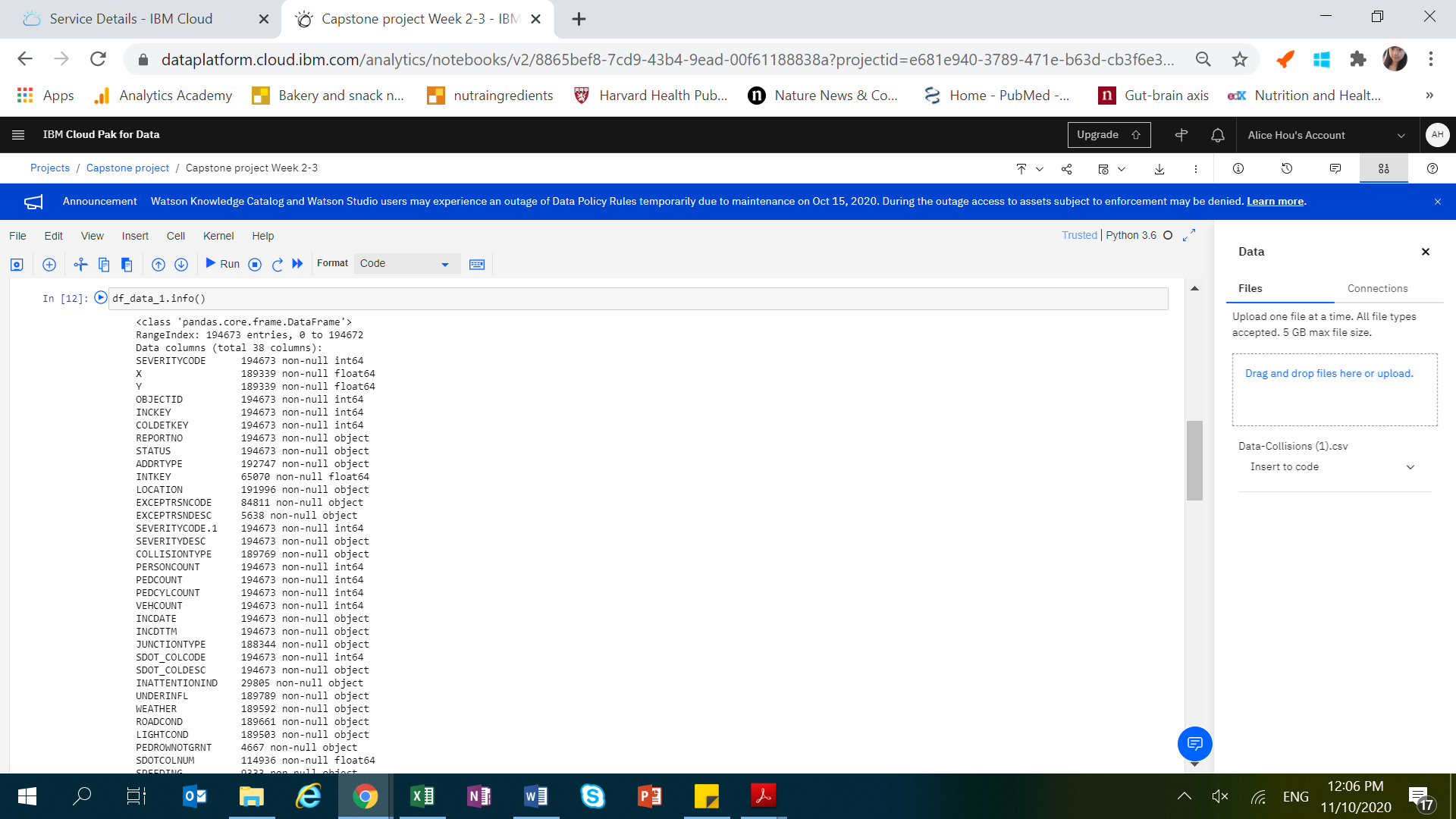
The target variable is SEVERITYCODE which takes the value from 0, 1, 2, 2b or 3. The definition of the severity codes are provided in the metadata and shown below.

|  |  |
| --- | --- |
| Severity Code | Meaning |
| 0 | Unknown |
| 1 | Prop damage |
| 2 | Injury |
| 2b | Serious injury |
| 3 | Fatality |

As shown below, the data consists of 38 columns.



From the data information description, we can see that some variables are categorical (data type as object).

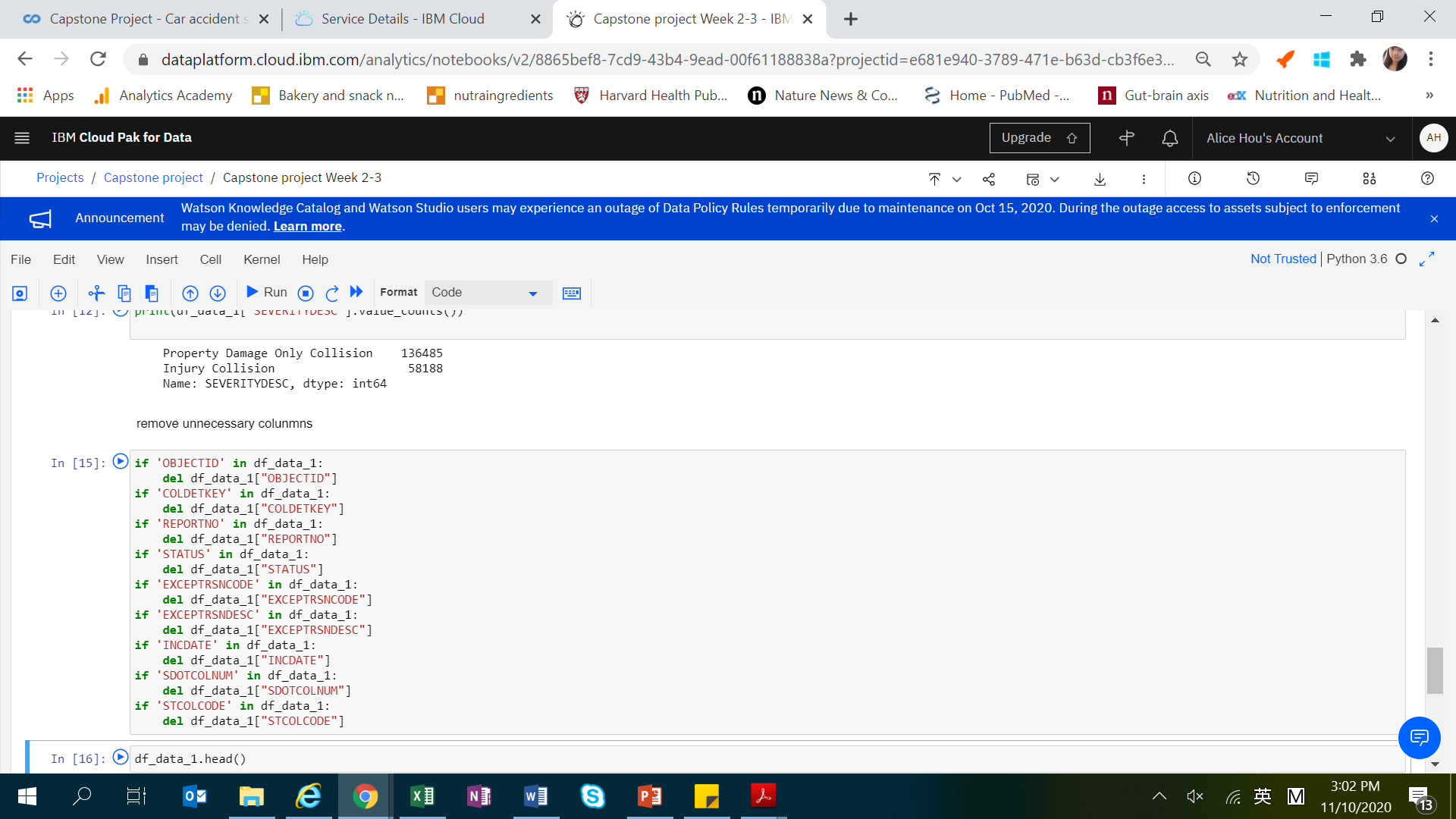


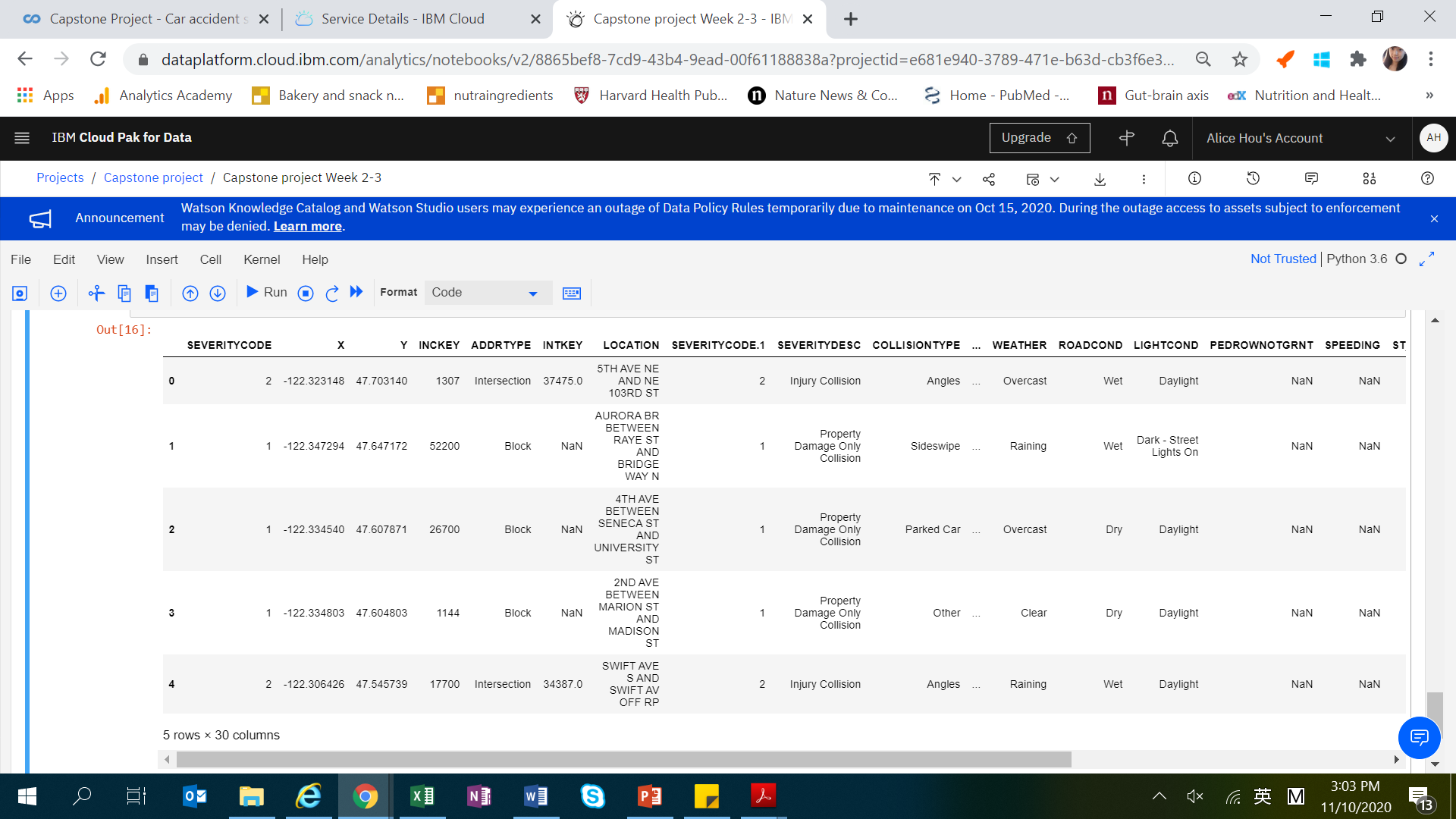
**Methodology**

Data clean-up

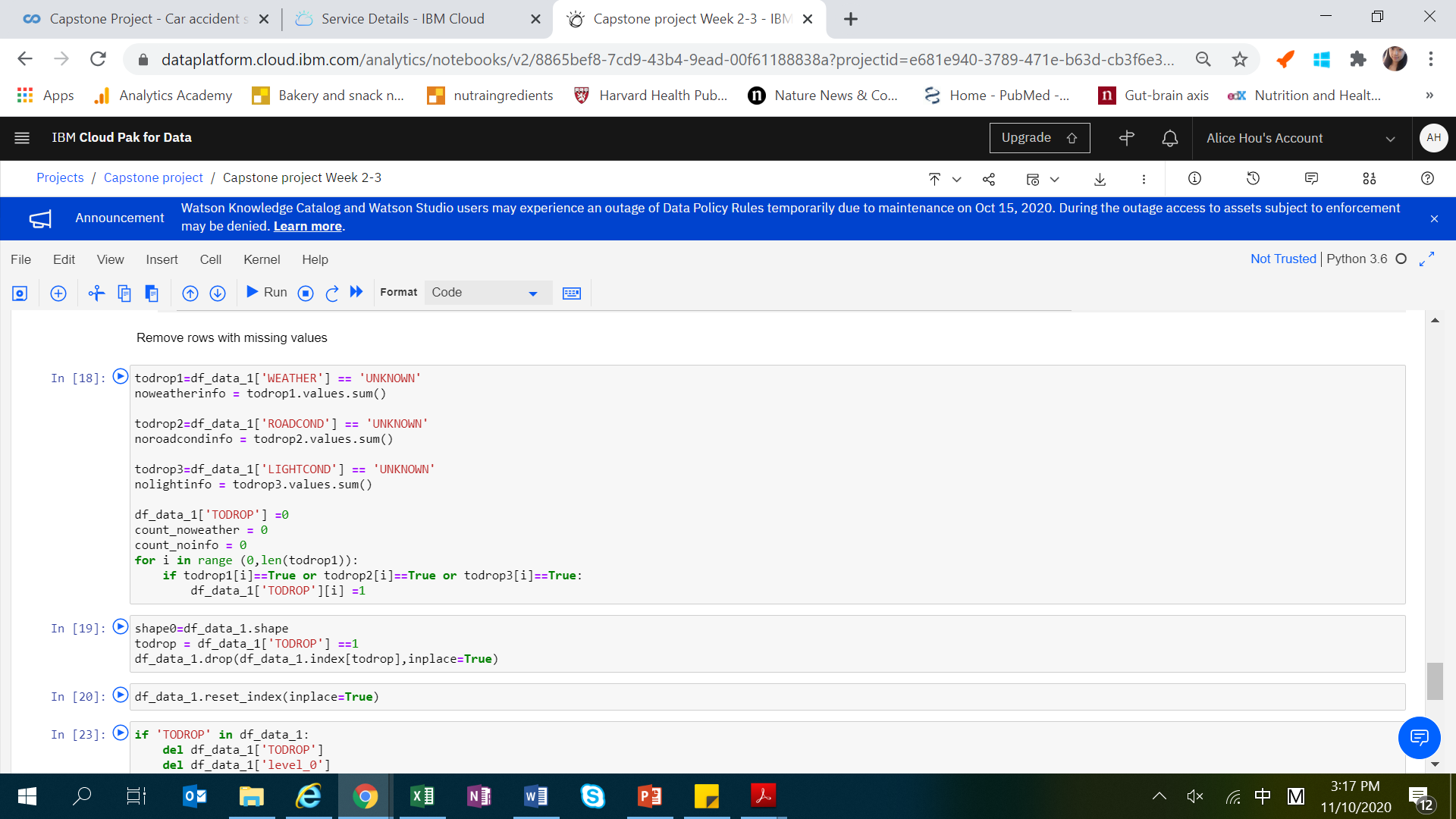
In the original dataset, there are a number of missing values, for example severity code, weather condition etc. Hence this section aims to clean up the data to remove the missing values.

We removed 8 columns which contain unnecessary/redundant information.

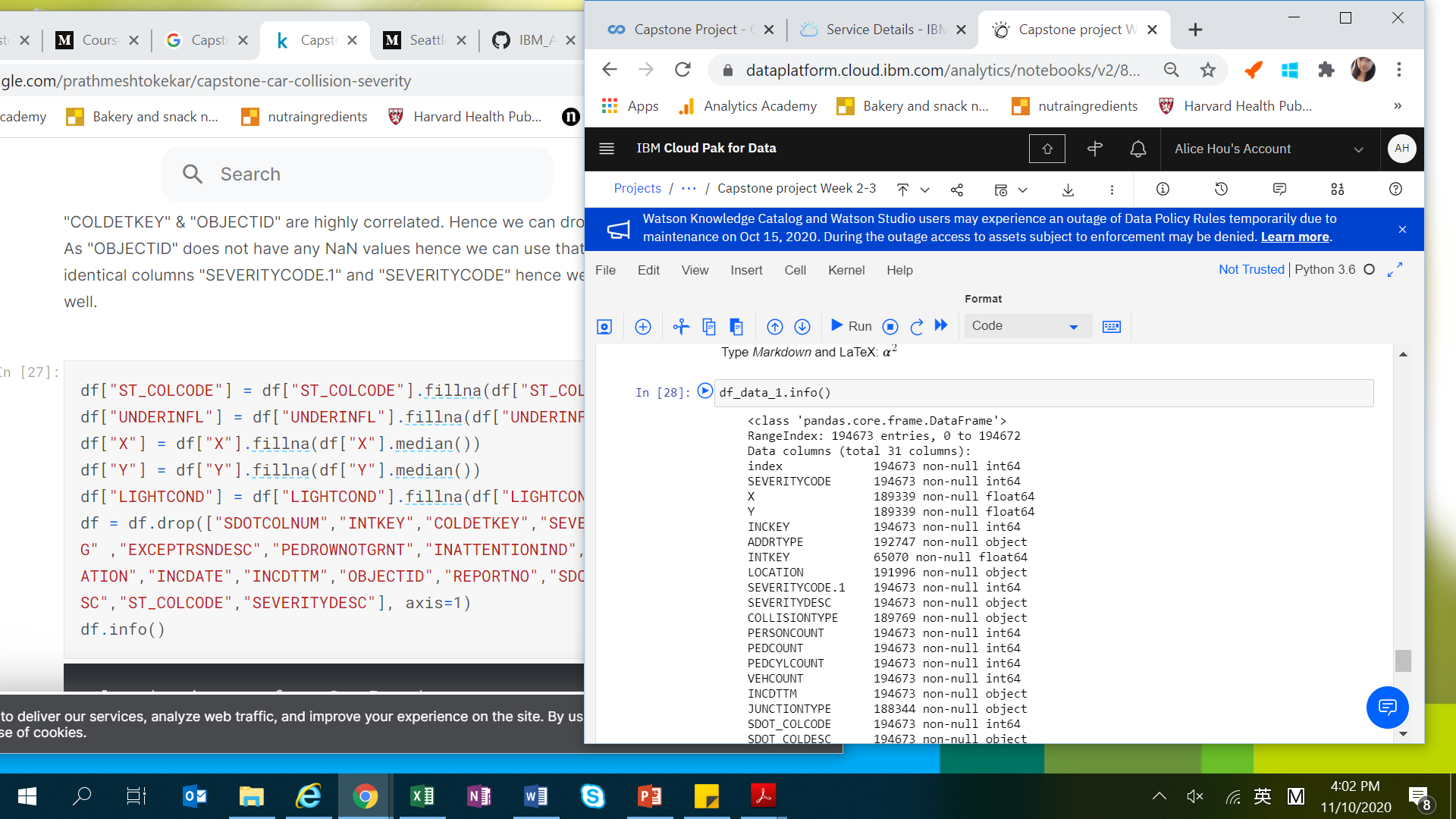




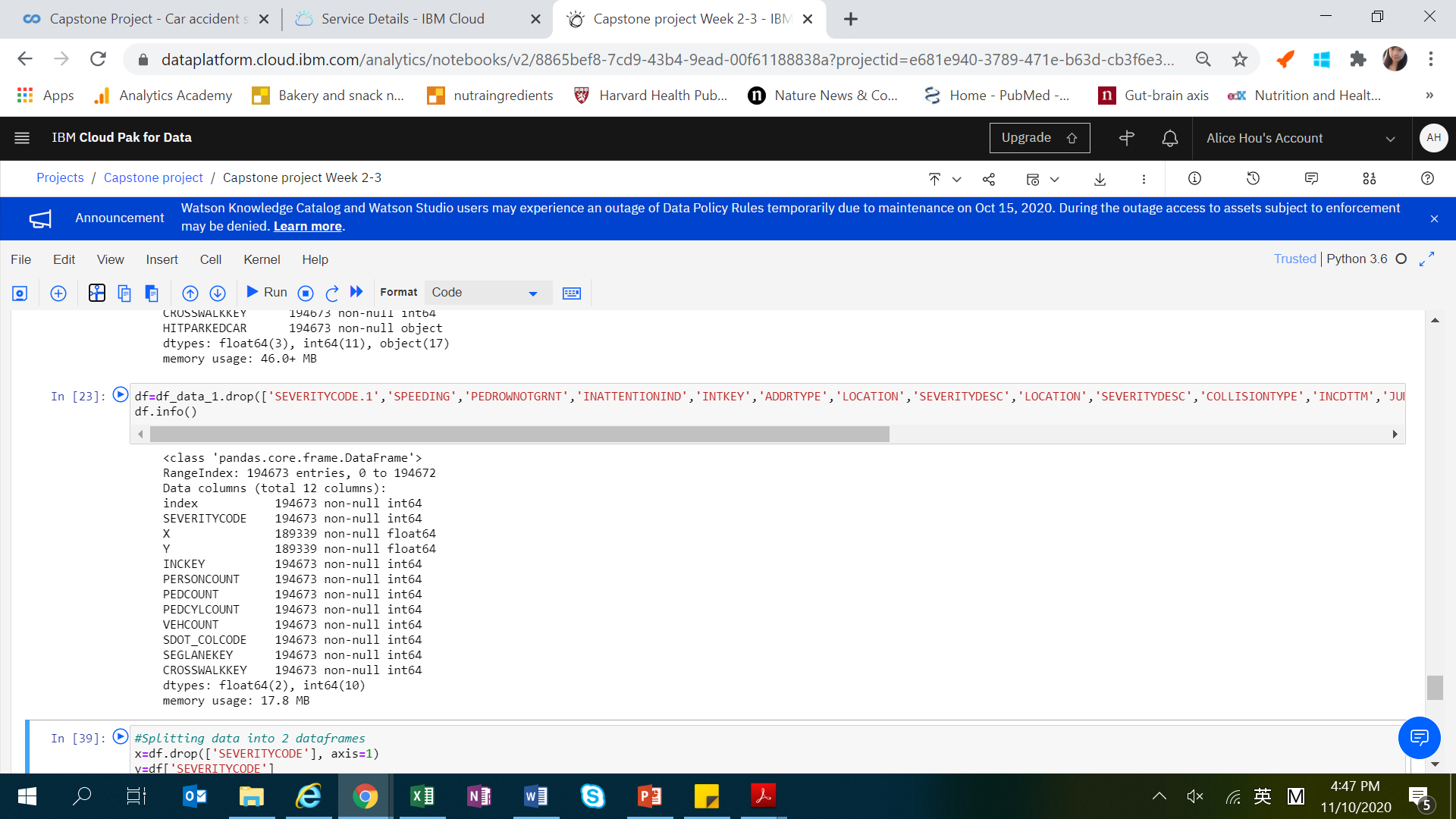
Next we set out to remove rows with missing values.



Next we check for missing values in columns and drop those columns with high number of missing values.



There are categorical variables which are not suitable for next phase modelling. Hence we removed categorical variable columns.



**Results**

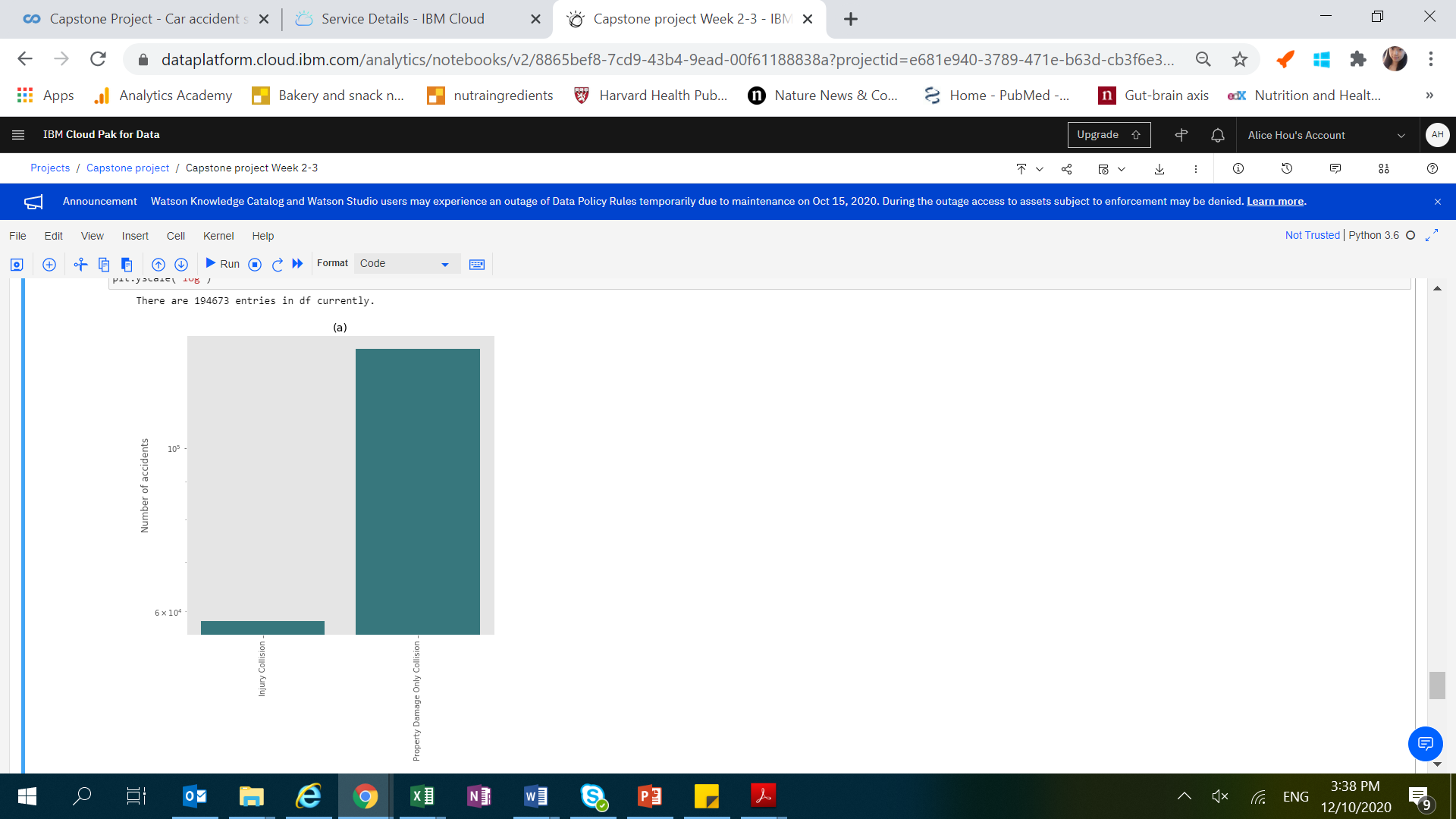
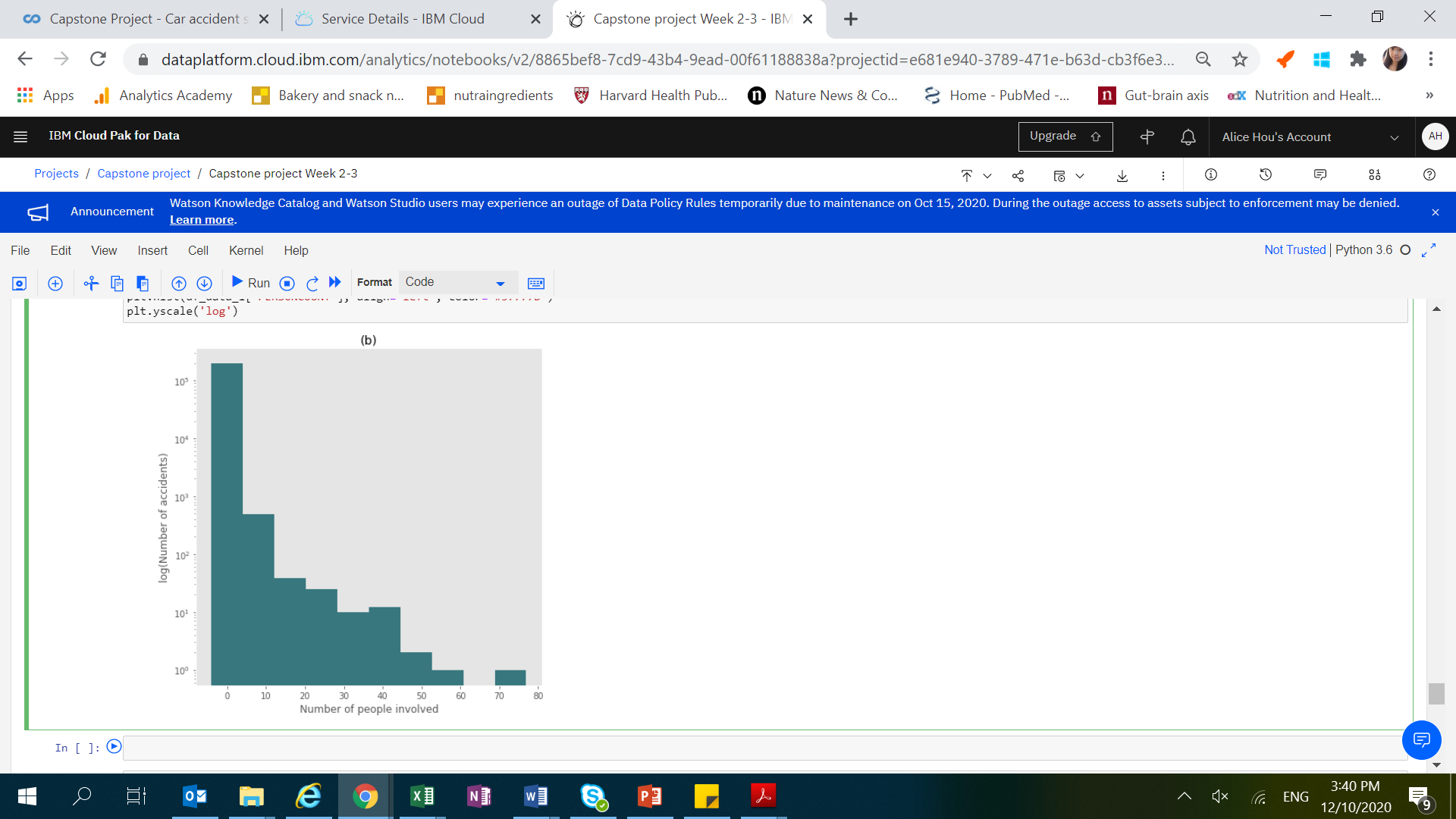
**Data visualization**

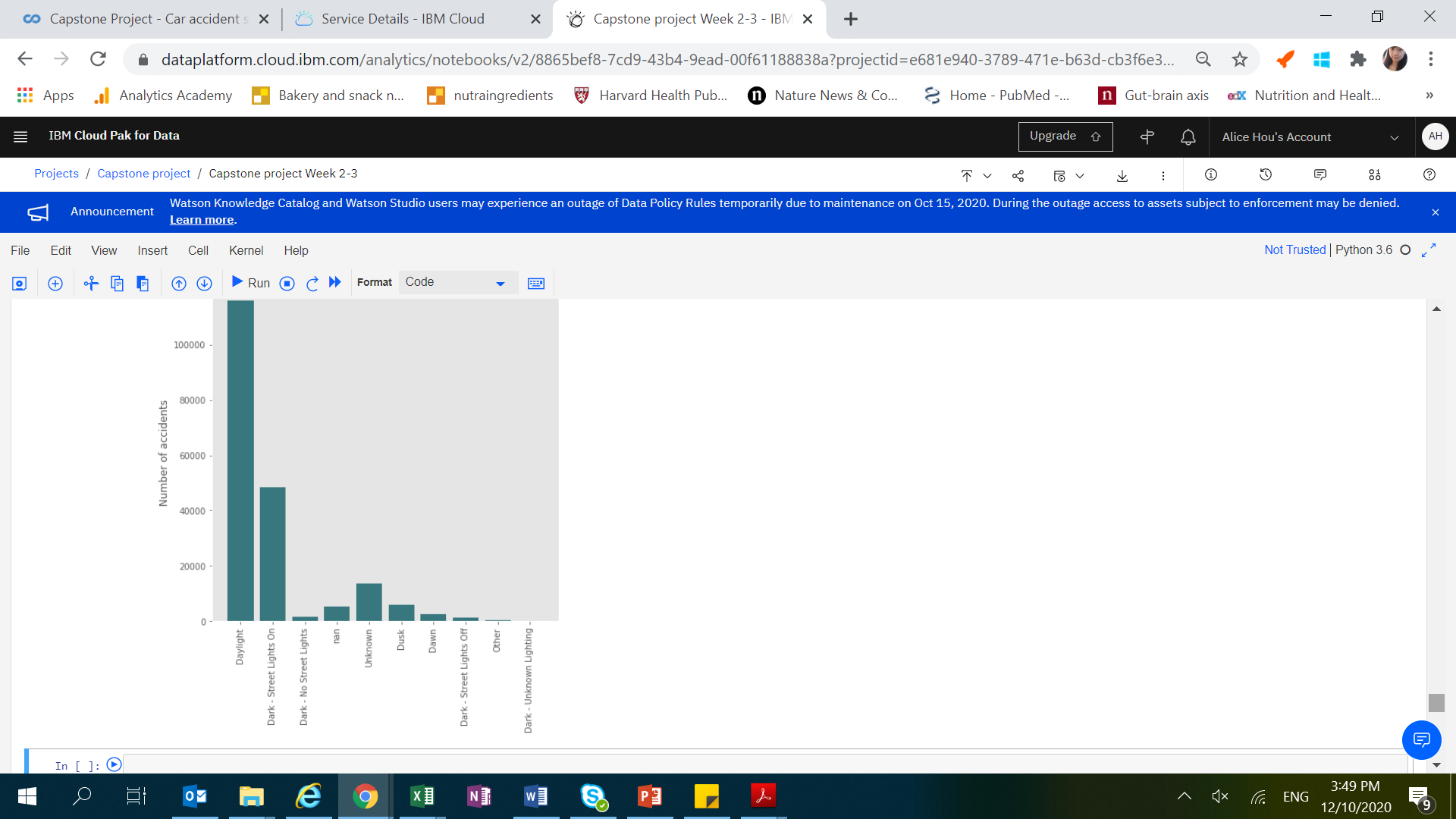
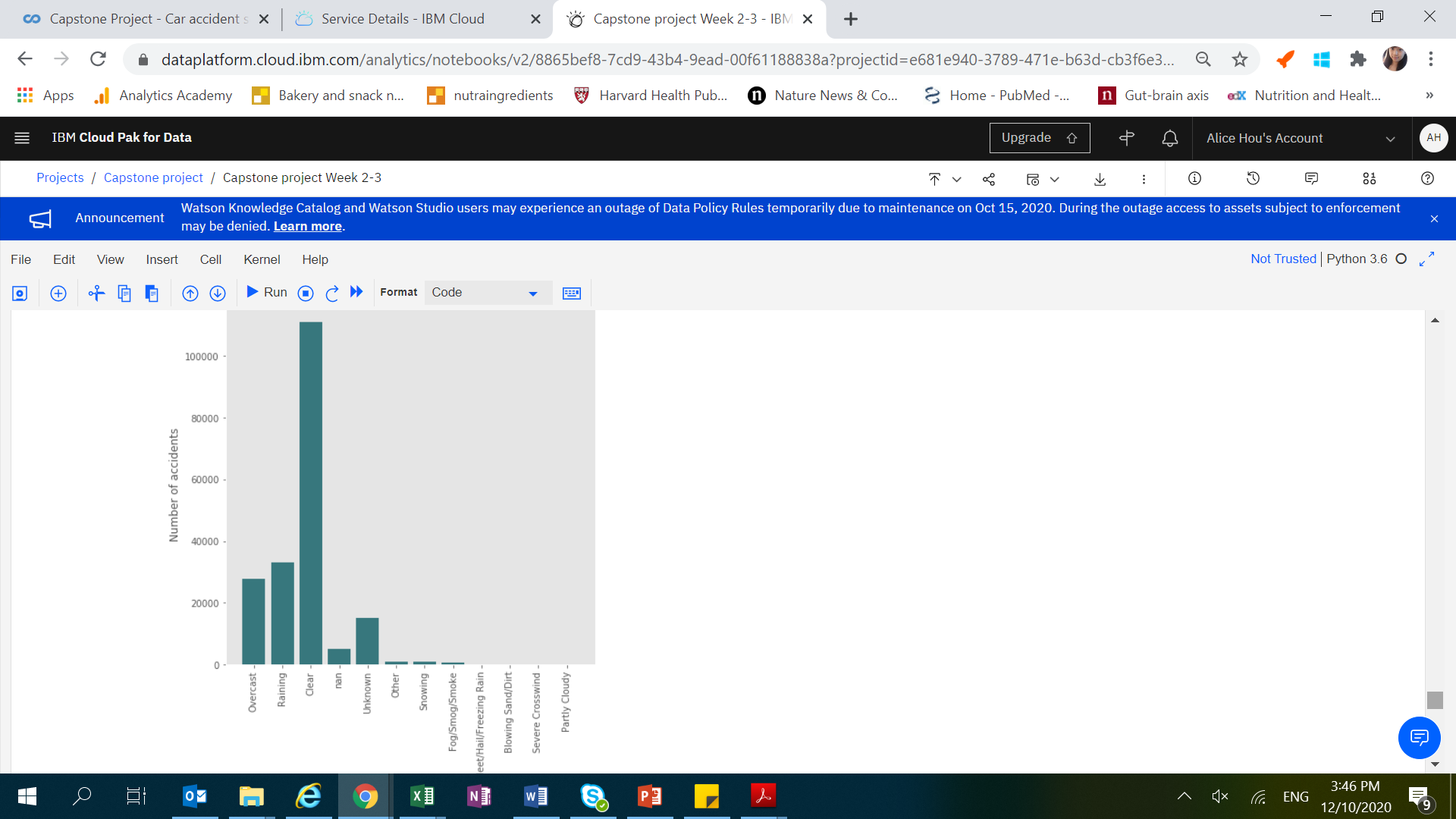
From data visualization, we could see that majority of accidents are property damage only collisions.

Number of people involved in accidents range from 0-5 (most) to 60.

Most accidents happen at clear weather condition.

Most accidents happen at daylight or at night with street light on.



**Discussion**

Contrary to what we have envisioned that more accidents will happen during wet weather and dark conditions, we have found that most accidents happened at daylight or with streetlight, and most accidents occur at dry conditions.

Due to the complexity of multiple factors involved, prediction of car accident severity may depend on various machine leaning factors and can be a combined effect include traffic condition, location, road conditions etc.

**Conclusion**

Through the Capstone project, I have learned to handle a real-world data with complexity. Various techniques from previous modules have been applied and I have benefited tremendously from doing this project.