**Android Malware Detection Using Machine Learning**

**Milestone - 1**

**Research Goal**

The main aim of this writing is, analyzing the android malware using the machine learning techniques.

**Abstract**

From metropolitan cities to rural villages the usage of mobiles and electronic devices are surged. The data security and safety is collapsing by loss or leakage of confidential information. The android applications are affecting in the form malware and virus attacks. The malware is software which targets phones and electronic devices that are accessed using internet. The Android Malware Dataset from kaggle has feature vectors extracted from android applications, including API signature, intent, command signatures, manifest permissions. The dataset has two data files and one of the files data in binary format. The dataset has 215 attributes which are extracted from 15,036 applications (5,560 malware apps and 9,476 benign apps). The goal of the project is evaluating android malware detection using machine learning techniques. This report looks into a dataset of Android malware to create better ways of using machine learning to spot and prevent these attacks.

**Research Questions**

The below questions which will be answered by this project:

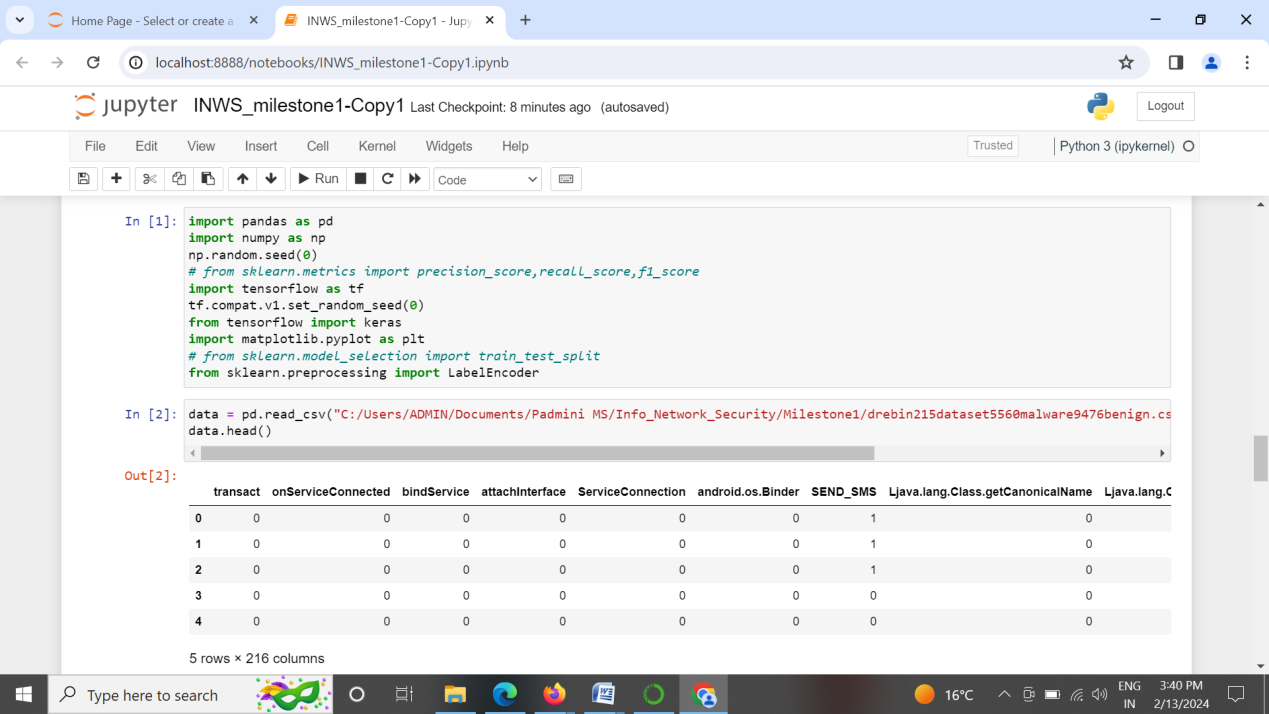
1. What makes an Android app show up like malware?
2. What is the distribution of features in the Android malware dataset?
3. Are there any significant correlations between different features in the dataset?

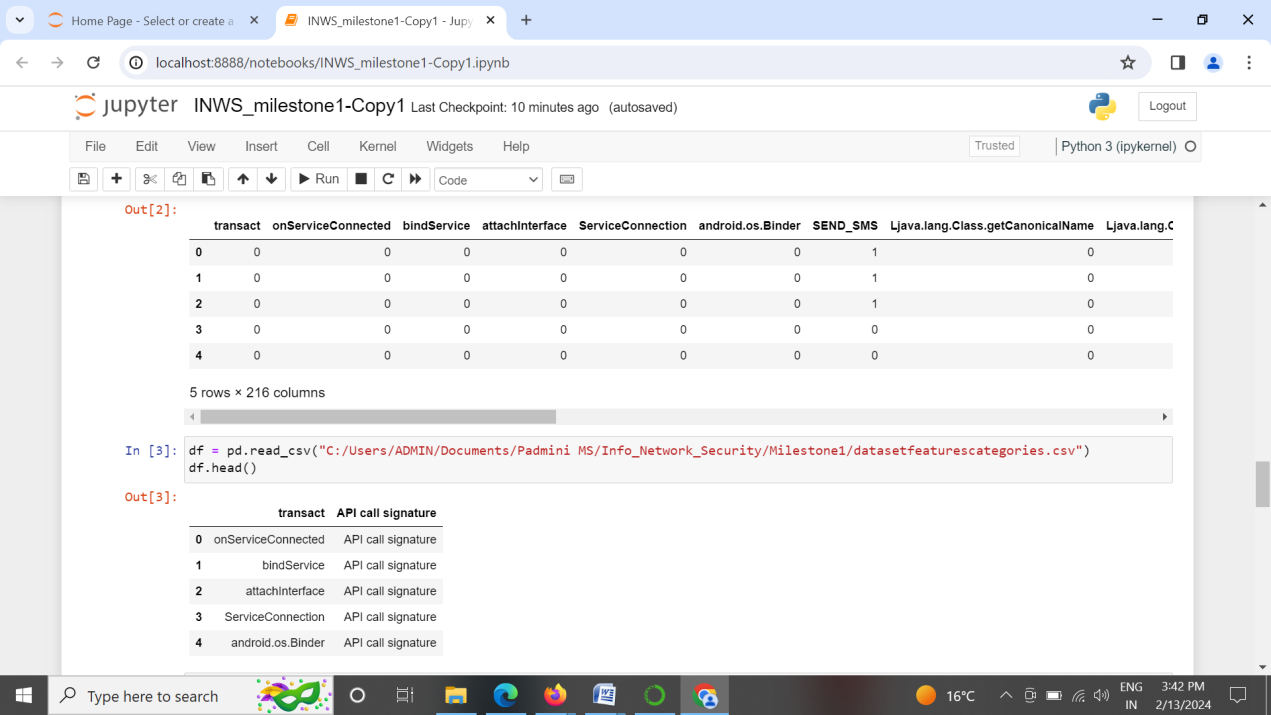
**Introduction**

This report focuses on analyzing an Android malware dataset to develop effective machine learning models for malware detection. Identifying applications that mimic malware behavior, distinguishing between malicious and benign apps. The initial step starts with reading the data or collecting data files.

**Data Collection:**

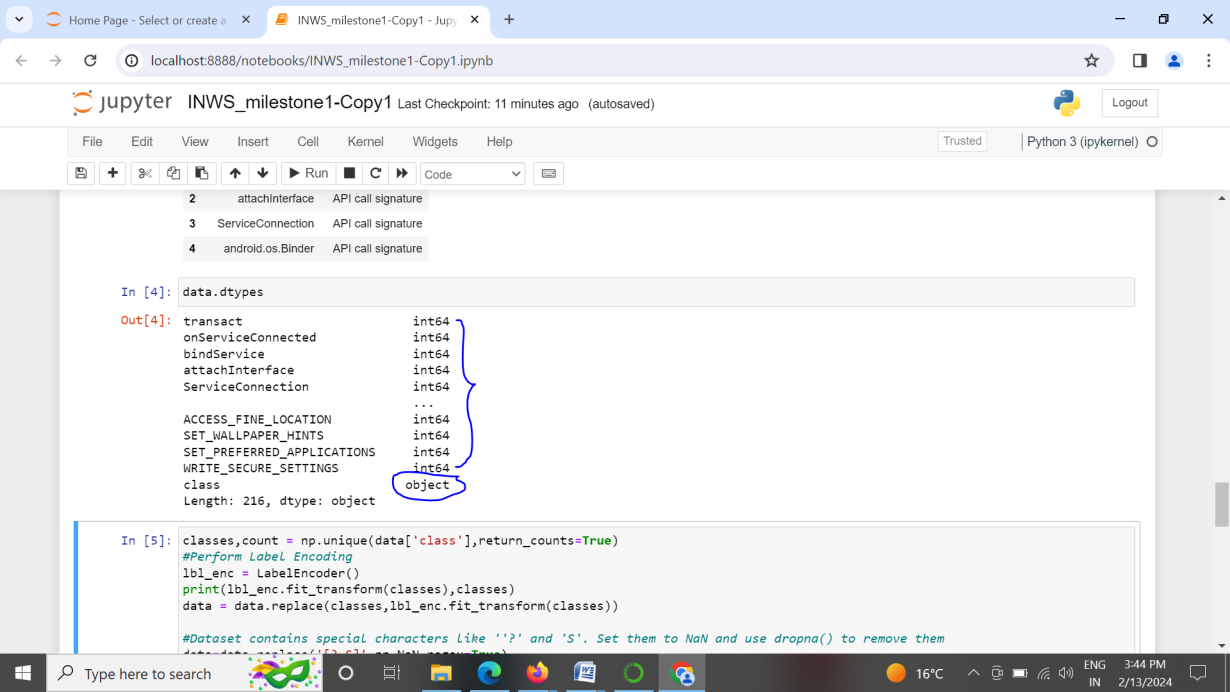
Importing the required python libraries (example: pandas, sklearn, scipy is for visualization) which are required for data loading, preprocessing and visualization. In the below two screenshots, the two data files are loaded into data frames using (read\_csv).



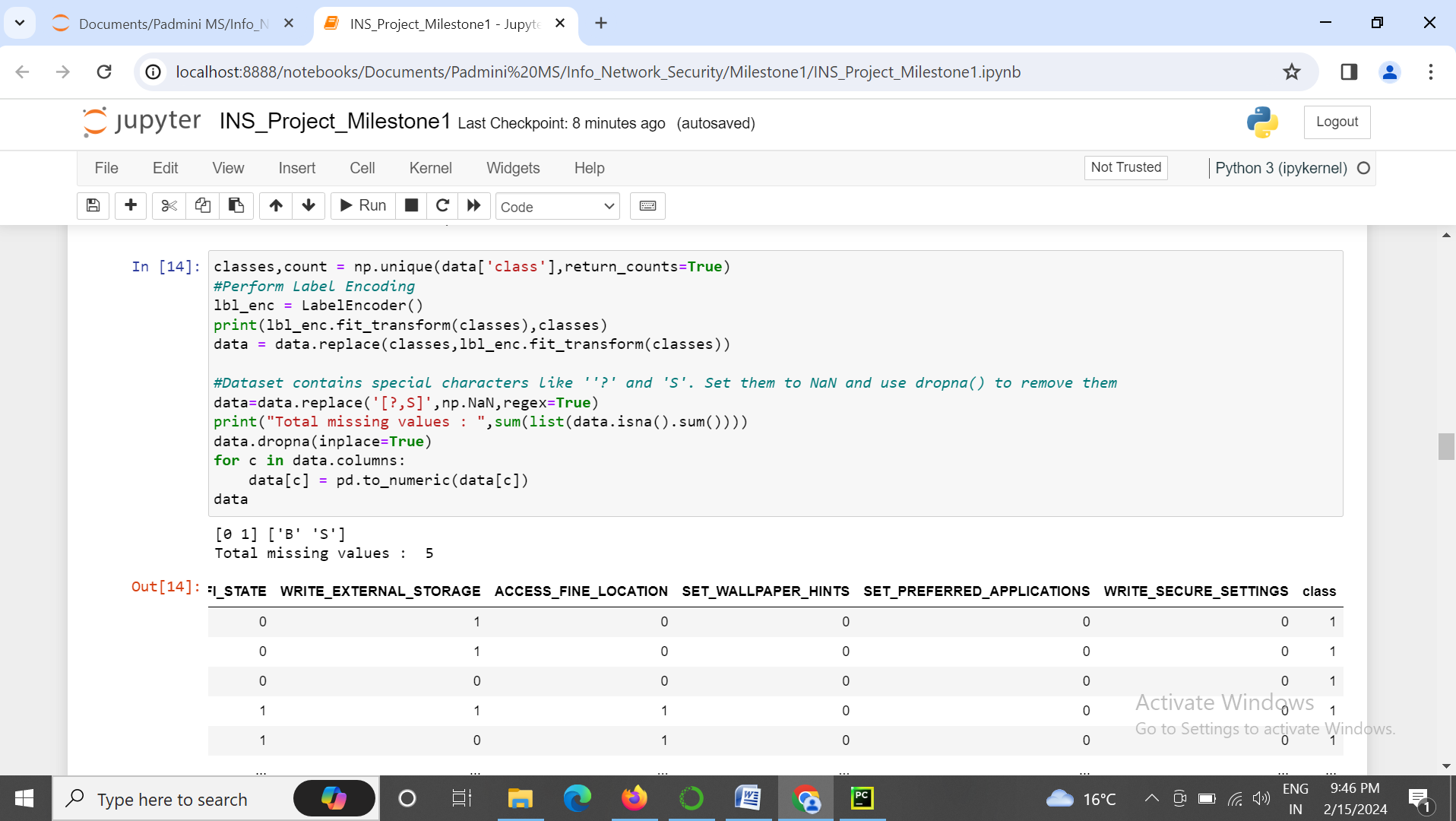


**Data Preprocessing**

1. As part of preprocessing there are various methods to process the data like eliminating duplicates, finding missing values, discarding null values, changing data format as per requirement. This dataset has no null values.
2. As part of this dataset out of 216 attributes only one attribute ‘class’ with data type **object.**

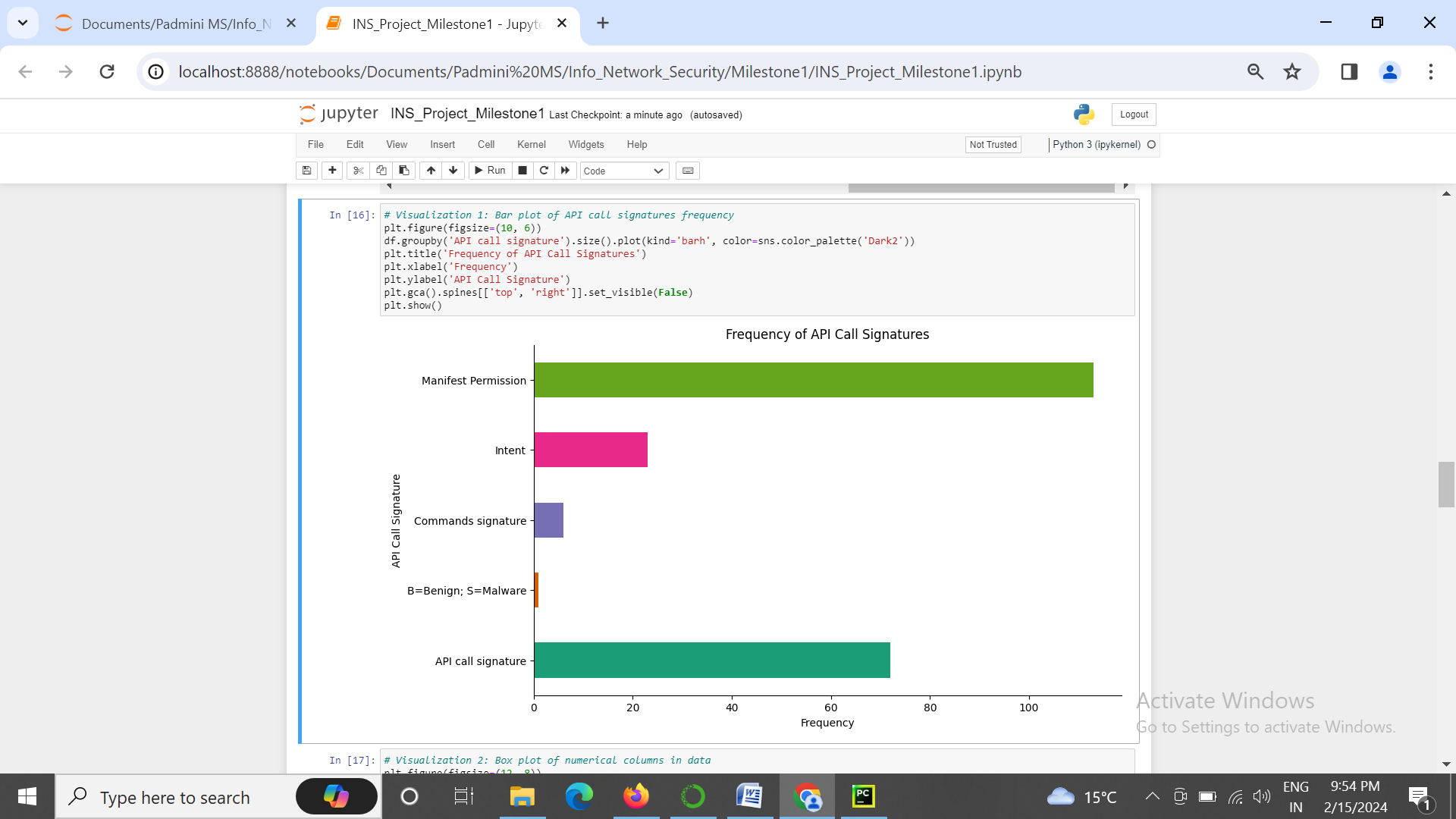
****

1. In below screenshot following operations have been performed:
2. The data & data type in ‘class’ attribute is converted to ‘int’ as (B-0,B-1) from ‘object’ to make analysis more precise. The reason to convert the data types is to eliminate any incompatibility between the attributes for the analysis.
3. Finding the ‘missing’ values. Also, having missing values may cause problem further like data inconsistency, incomplete analysis which can leads to incomplete conclusions in the project.
4. The special character, missing values, are setting them to NaN and are eliminated using ‘dropna()’.

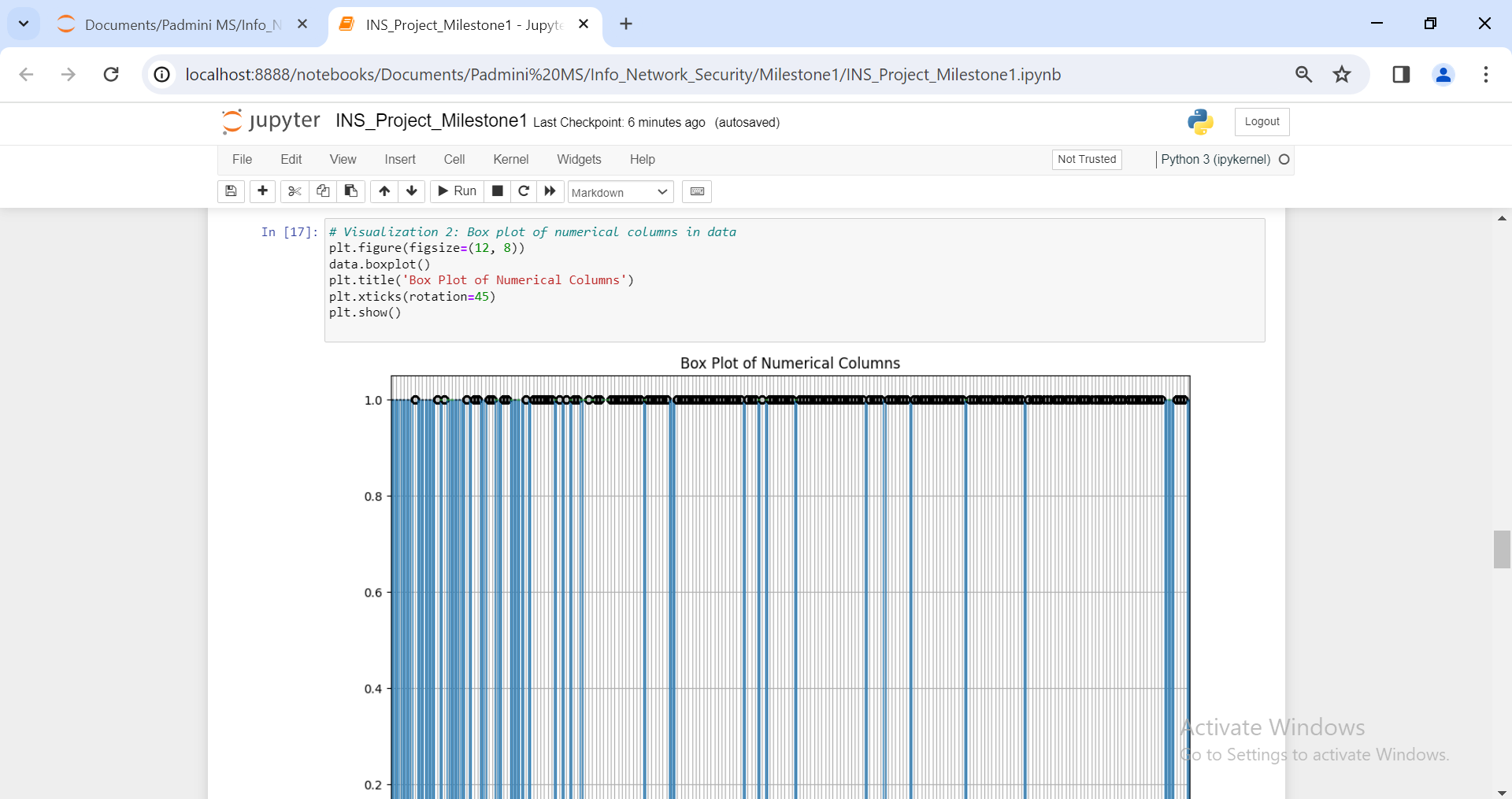


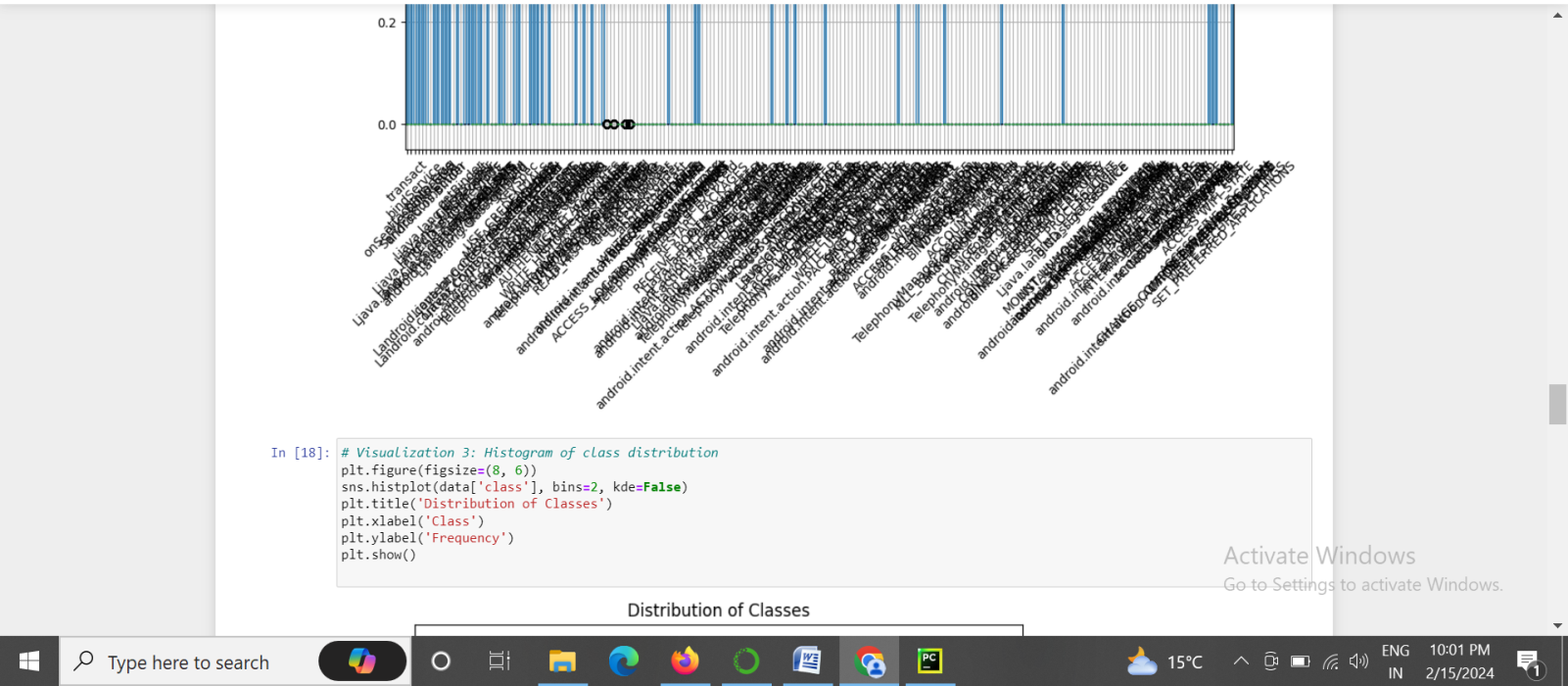
**Data Visualization**

Visualization 1 : Below graph is API call signatures frequency which show the highest frequency in Manifest Permission, and (Benign, Malware has the least frequency)

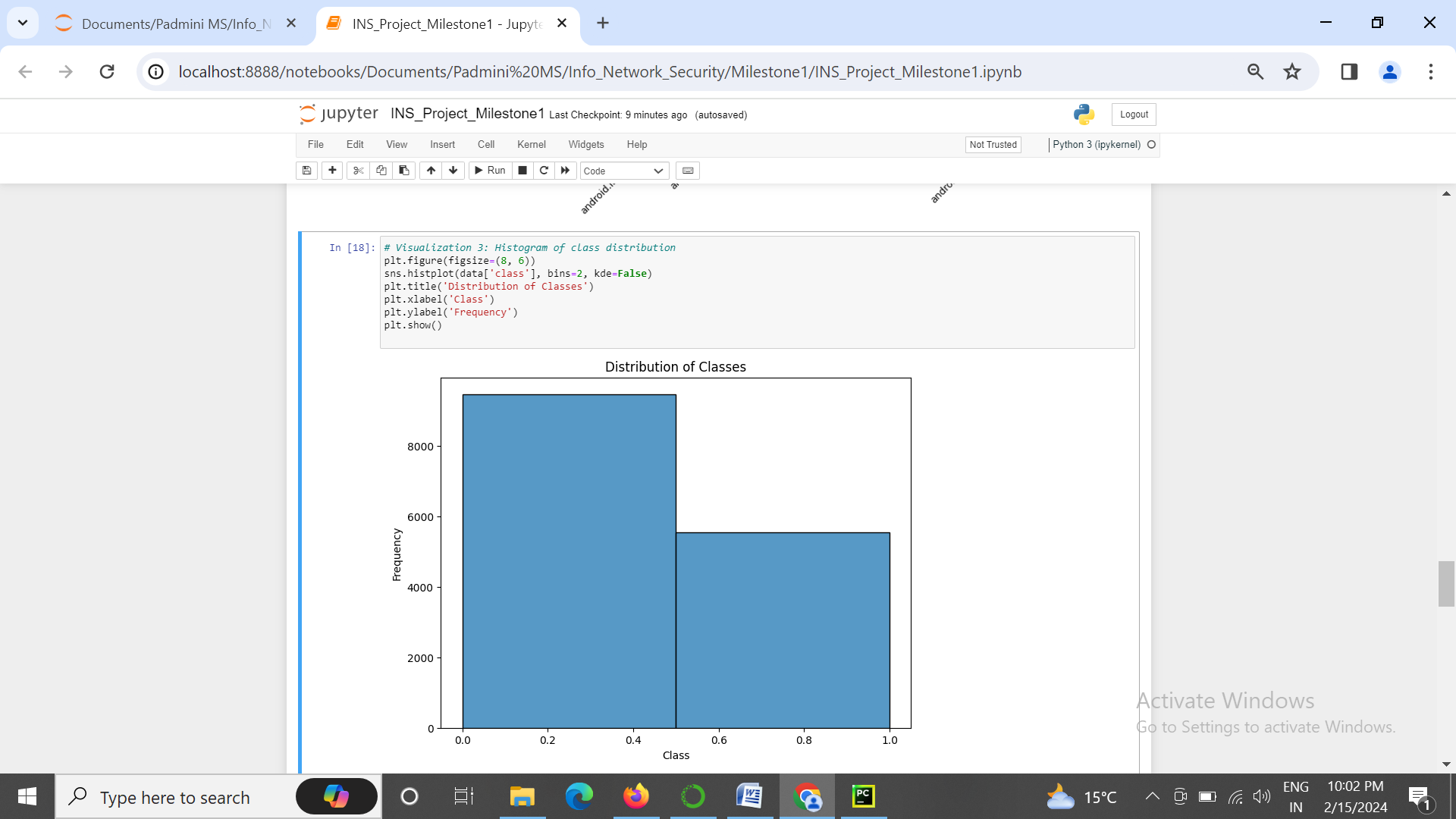
****

Visualization 2: Below graph is for numerical columns in dataframe **(data).**  On x-axis we have attributes and in Y-axis we have numerical accuracy values.

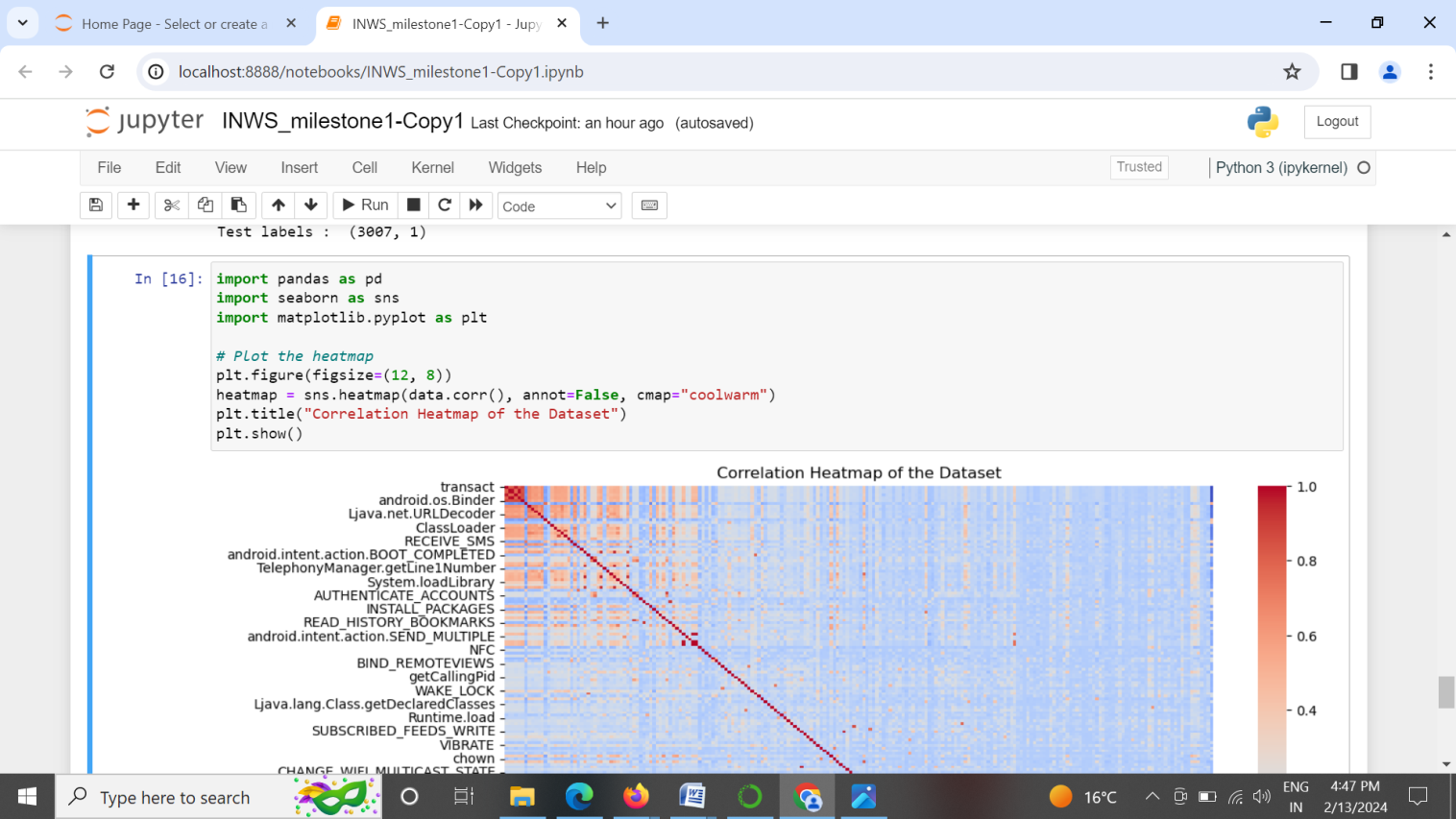
****

****

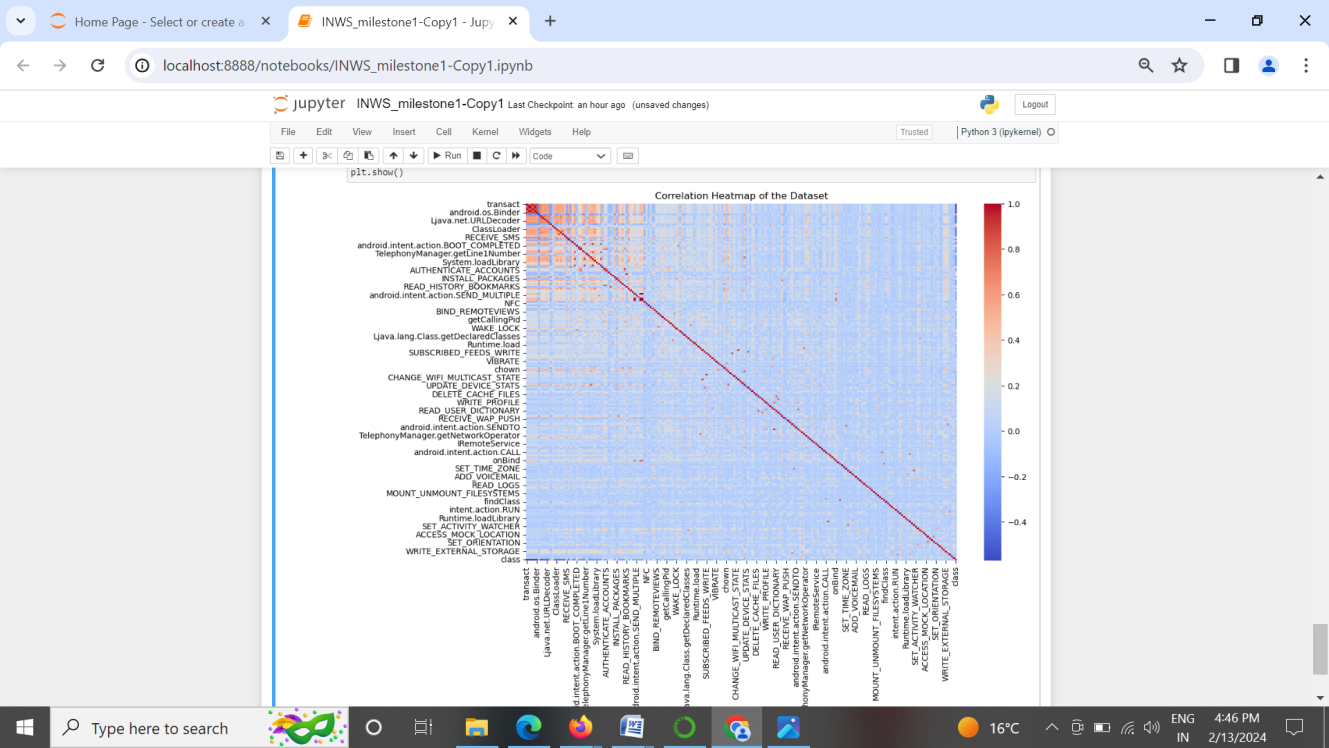
Visualization 3: Below graph is class distribution for data frame **(data).** The class is an attribute, (Benign and Malware) are its values and we have them on x-axis. On x-axis from 0.0 to 0.5 is Benign, 0.5 to 1.0 is Malware. On y-axis we have frequency which refers as number of records in data file for class attribute. This plot is representation differentiation on records for (Benign and Malware) values.

****

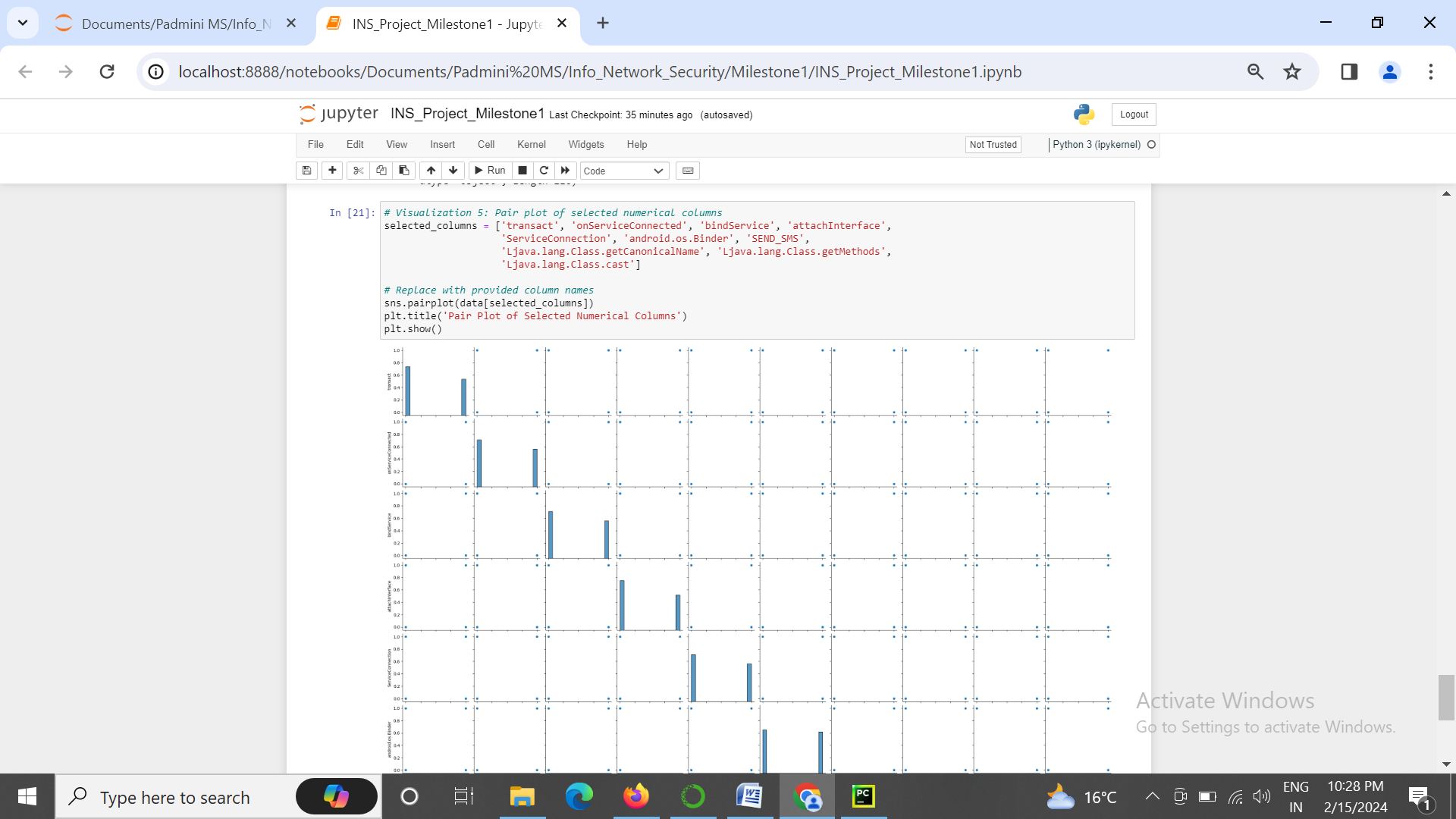
Visualization 4: The below code for building the heat map visualization which represents relationship between features in the dataset. We call them attributes from the .csv files these are about the special privileges that apps must ask for user approval to when they want to access the sensitive information like location, contacts, recording, images etc.

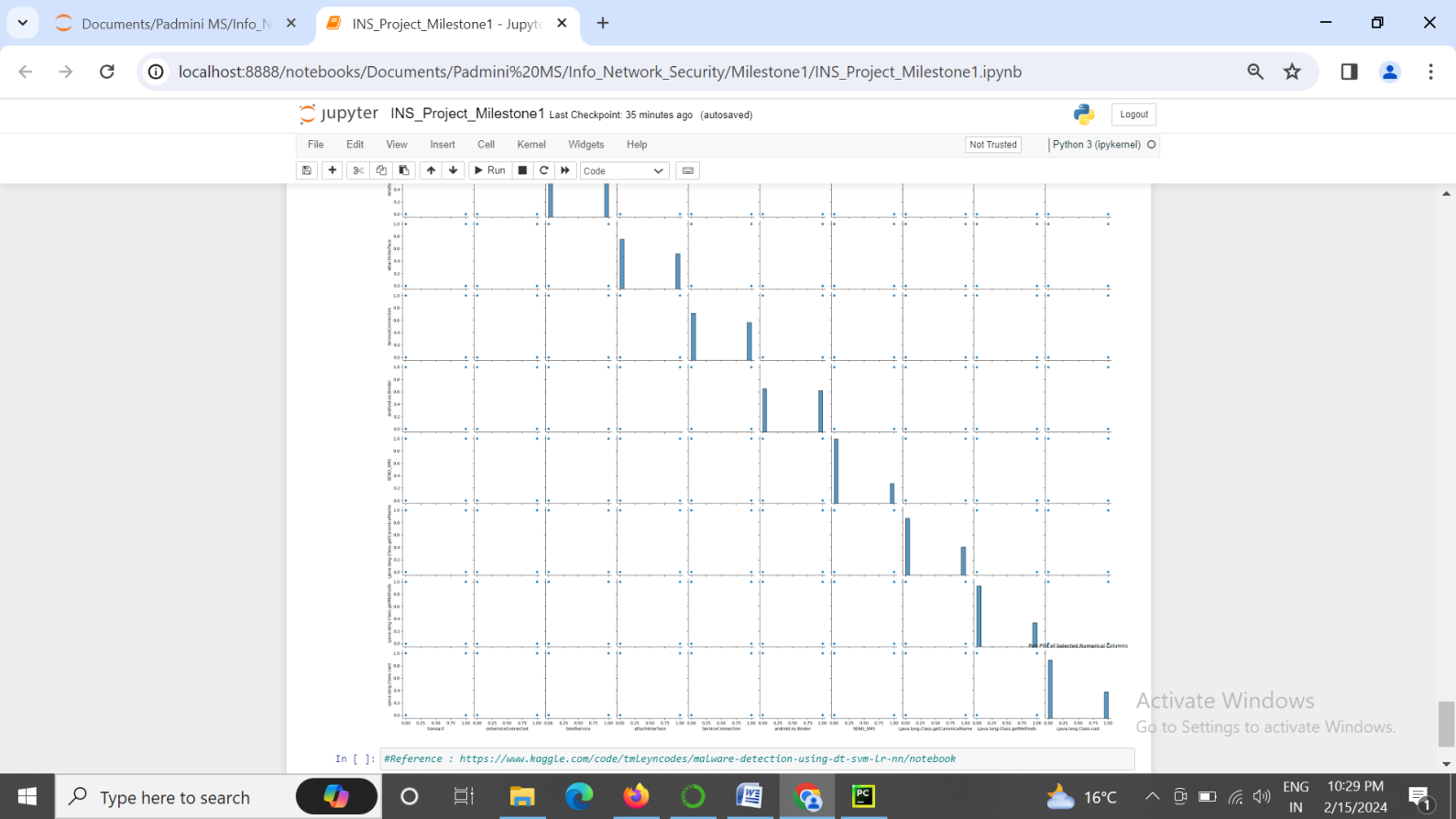
****

From this heat map understanding the dark red color indicates the strong relationship among attributes, blue color indicates the negative relationship and white color indicates the no significant in relationship. However, the heat map scale on the right side, which shows numerical values for correlation values from -1 to 1 where correlation is near to 1 is positive correlation and -1 is opposite for positive. These positive correlated attributes values can be used for further analysis to develop more insight on the dataset.

****

Visualization 5: Below plot shows the relation among selected numerical columns.

****

****

For the upcoming milestone, we need to understand the correlation between the attributes in detail like, how are they strongly correlated, and why there is no significant relationship between some of the attributes etc. In a nut shell the dataset is all about the malicious attacks on android applications mobiles and electronic devices and how they are affected by the attacks and how the confidential information is leaking.

Code References: <https://www.kaggle.com/code/tmleyncodes/malware-detection-using-dt-svm-lr-nn/notebook>

**Team Members - Section - 030**

Padmini Aripirala - 999903261

Keerthi Reddy Chimalapati - 999903554

Sai Kiran Duvvuri - 999902946

**Participant Activities:**

**Padmini Aripirala**: From data collection to data preprocessing.

**Keerthi Reddy Chimalapati**: Data visualization, Visualization 1 and Visualization 2

**Sai Kiran Duvvuri**: Data visualization, Visualization 3, Visualization 4, Visualization 5.