**CST8502 - MACHINE LEARNING**

**“BICYCLE BURGLARS - PREDICTING THEIR END FOR SAFER TORONTO”**

**Assignment -03 - Project Proposal for Final Project**

by

Joshua (Kishor) Ayyasamy - 041133904



Project Proposal Report

submitted to

**Dr. Anu Thomas Ph.D.**

in satisfaction of the requirements

for the Assignment-03

November 17, 2023

Table of Contents

[A brief note on Machine Learning 3](#_Toc701)

[1. Introduction 4](#_Toc30426)

[2. Objective 5](#_Toc21960)

[3. Business Understanding 6](#_Toc3935)

[4. Data Understanding 6](#_Toc1761)

[4.1 Collecting the data 7](#_Toc32057)

[4.2 Statistical analysis of the numerical data 9](#_Toc26252)

[4.3 Checking the total number of null values in the data 9](#_Toc23727)

[4.4 Checking the total number of unique values in the data 9](#_Toc24161)

[4.5 Checking the co-relation in the data 10](#_Toc32167)

[4.6 Understanding the visualization in the data 11](#_Toc180)

[5. Data Preparation 16](#_Toc11038)

[5.1 Selection of Data and Cleaning Unwanted Data 16](#_Toc20128)

[5.2 Filtering out the data only for the latest year (2022) 18](#_Toc911)

[5.3 Handling the missing or NULL values 19](#_Toc12112)

[5.4 Converting the Continuous data attribute into categorical data 20](#_Toc4494)

[5.5 Encoding the categorical data and splitting the model into train and test models:- 21](#_Toc10946)

[5.6 Resultant data frame - bicycleTheft01 after encoding 23](#_Toc3602)

[6. Conclusion 23](#_Toc19276)

# A brief note on Machine Learning

Machine learning is one area in Artificial Intelligence (AI) and is part of computer science, involving data to be analysed and algorithms to be used such that machines can slowly replicate how real people learn to get better in precision. Basing on this, ‘machine learning is a super developer! Because it means developing one sample of data information for a particular computer programme and a better result over time. Thus, the aim of machine learning is to teach the computers how to learn on their own so that they can mimic smartness that is characteristic of man in any aspect of life. Why it is being used today in different places.

For instance, Machine Learning may be applied in various services such as movie suggestions provided via digital platforms, voice recognition in virtual assistants, autonomous cars, search engines, robotics, medical diagnostics, in the financial system, and so forth.

Machine learning is another very large world in which various types of applications are worked using different methodologies and approaches. For instance, there are decision trees, regression models, classification models, clustering techniques among many others. However, the most vital method that has been adopted within the past decade globally is none other but the artificial neural networks technique. You will need to learn a programming language such as Python or R to use it as a tool for creating your neural networks or other approaches. The two are statistical-big data manipulative approaches.

There are four approaches to Machine Learning process:-

* Supervised Machine Learning
* Unsupervised Machine Learning
* Semi-Supervised Machine Learning
* Reinforcement Learning

# Introduction

The final project in Machine Learning encompasses and covers all the topics covered in the course. As per the instructions, the dataset used in the project is “Toronto Bicycle Theft”, which reports all the bicycle theft cases in a specific time frame, in the different areas of the city of Toronto. The cases enlists the occurrence date and time of the theft followed by the reporting date and time and the status (whether it is stolen or recovered). For using the brief data instead of using 31970 instances, the data belonging to the year 2022 is extracted and hence, the project is worked on 2923 instances of bicycle theft.



**Question framed for the project:-**

**“Which are the types of bikes most commonly stolen in Toronto in 2022? Predict the areas that are more prone to bicycle theft and the timings when the risk of stolen bicycles was at the peak.”**

Technology used for the project:- For the project, **Python** language is used to develop the whole solution and to implement all the steps of the CRISP-DM process involving Data Understanding, Data Preparation, Modelling and discussion of results. The interface to develop the whole solution used is **PyCharm Community Edition 2023**. It is an open-source software which handles all the coding parts and provides accurate results based on the given code. Further, the report is documented using **Microsoft Word**.

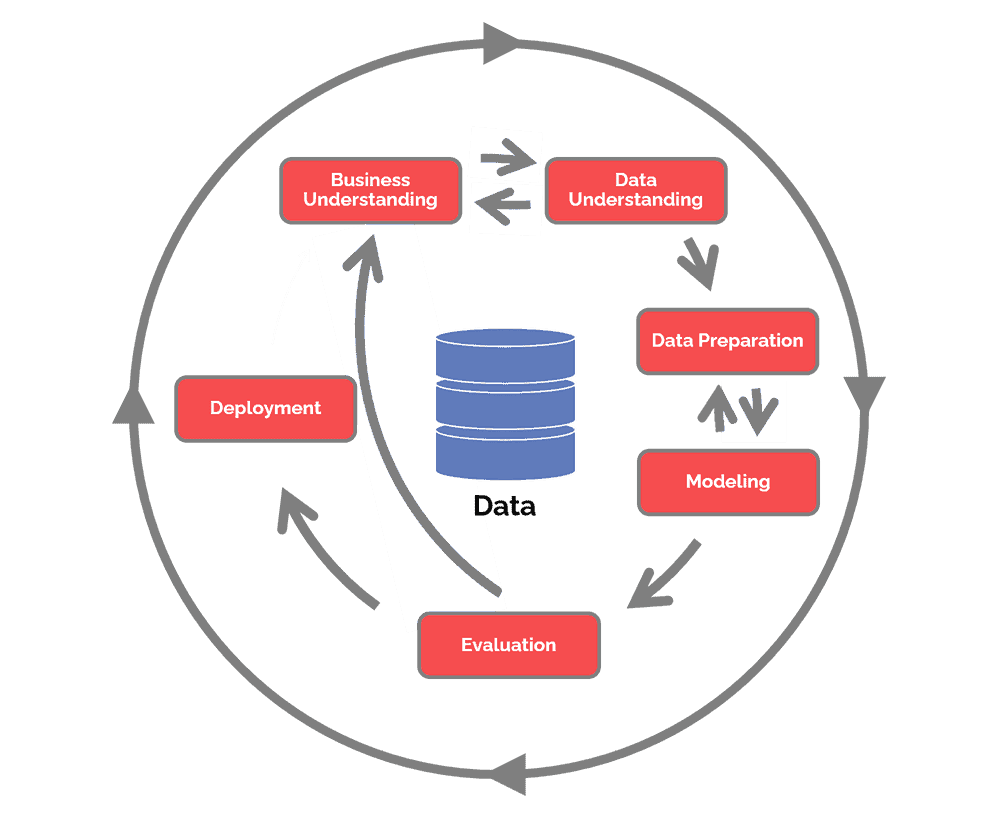
In Data Understanding, the data is completely explored and analysed to prepare it for classification purposes. The continuous and categorical attributes are completely understood and along with that, the quality of data is also checked by looking into the outliers and missing values. In order to understand the data, several operations were performed such as finding the data types of the attributes, analysing the statistical aspects (count, max, min, mean, median, mode), checking the co-relation between all the numerical attributes, checking for the number of unique values so as to check the Cardinality of the attributes, checking for the NULL or the missing values. It is finally followed by understanding the data visually using the bar graphs and histograms. This helped in developing relationships which made data preparation easier.

Looking towards the Data Preparation part, the data set is first loaded in the Python program and stored in a data frame. The necessary attributes are then selected and all the unwanted attributes are excluded. This step is done on the basis of the need of attributes for the solution i.e. question demand, higher cardinality, higher number of missing values or outliers in the data. After that, all the null values for the included attributes were filled and binning was performed in order to convert the continuous feature into categorical. At last, encoding is done for the categorical data for ease in classification purposes and finally, the model was split into 80% of train and 20% of test data.

# Objective

The goal of this assignment is to explore and analyze the factors that contributed to the stealing of bicycles in various areas in Toronto. It encompasses the relationship of bikes stolen with the type of bikes, the timing of the occurrence, and the type of area where the incident happened. Further, continuing from here to perform classification using several models such as Decision Trees, kNN; outlier detection, and clustering which form the part of the final project followed by the project presentation. We use Python to accomplish the tasks undertaken in this assignment. Herewith, we undergo the following steps in a sequenced manner using CRISP-DM methodology.

**CRISP-DM Methodology:-**

It stands for Cross Industry Standard Process for Data Mining. This methodology is widely used among Data Scientists and Data Analysts to solve all the problems related to Machine Learning and to develop algorithms for the same. The CRISP-DM methodology provides a structured approach to planning a data mining project. It is a robust and well-proven methodology. This model is an idealized sequence of events. In practice, many of the tasks can be performed in a different order and it will often be necessary to backtrack to previous tasks and repeat certain actions.[[1]](#footnote-1) The CRISP-DM approach helps ensure that the business goals remain at the center of the project throughout. CRISP-DM provides an iterative approach, including frequent opportunities to evaluate the progress of the project against its original objectives.[[2]](#footnote-2)

Steps involved in the methodology:- Figure: CRISP-DM stage cycle[[3]](#footnote-3)

* Business Understanding
* Data Understanding
* Data Preparation
* Data Modelling
* Evaluation
* Deployment

# Business Understanding

* This dataset contains Bicycle Theft occurrences by reported date and details regarding the stolen item where available. It uses machine learning to analyze the Toronto bicycle theft dataset and develop predictive models that can enhance proactive measures and community awareness to reduce bicycle theft incidents.[[4]](#footnote-4)
* The Toronto Bicycle Theft dataset includes all bicycle theft occurrences reported to the Toronto Police Service, including those where the location has not been able to be verified. As a result, coordinate fields may appear blank.[[5]](#footnote-5)
* This also includes occurrences where the coordinate location is outside the City of Toronto but it is out of scope for the project. The whole solution is developed taking into account the data of the latest year i.e. 2022.
* As per the model, it is addressing the concerns of the weather conditions, the brand, model and cost of a vehicle affecting the frequency of bike theft cases across Toronto.
* It's crucial to collaborate with local authorities, law enforcement, and relevant organizations to access accurate and up-to-date data for a more comprehensive analysis. This approach can contribute to effective preventive measures and community awareness initiatives.
* For predicting high-risk areas and peak timings of bicycle theft, a data-driven approach is essential. Utilize machine learning algorithms and statistical analysis on historical crime data.
* We obtained information on stolen bikes, including their make, model, and features. This could involve collaborating with law enforcement, utilizing crime databases, or engaging with local bike shops and communities to gather insights.
* During this phase of the project, we implement predictive models that can identify potential risk factors and highlight areas and times with increased vulnerability to bike theft. One of the main objectives of our project is to predict and identify geographic areas where bicycle theft is more likely to occur based on the data given for the year 2022.

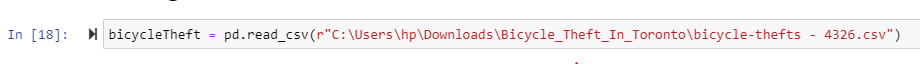
# Data Understanding

A crucial initial step in the journey involves loading the datasets into our analysis environment - Python. This action sets the stage for **exploration, enabling us to delve into the attributes, examine patterns, and uncover relationships that underpin wine quality**. These datasets encompass a rich tapestry of 28 columns, such as the 28 attributes mentioned below, where each related to the theft occurred in the area. With a total of 31970 instances, this collection of data holds the promise of predicting the type of area, bicycle, and time of occurrences. The attributes in the datasets given are of a real numerical type, object type (nominal). So understanding their type and format is beneficial for data preparation which will be done in next step. Let’s broadly discuss the steps below -

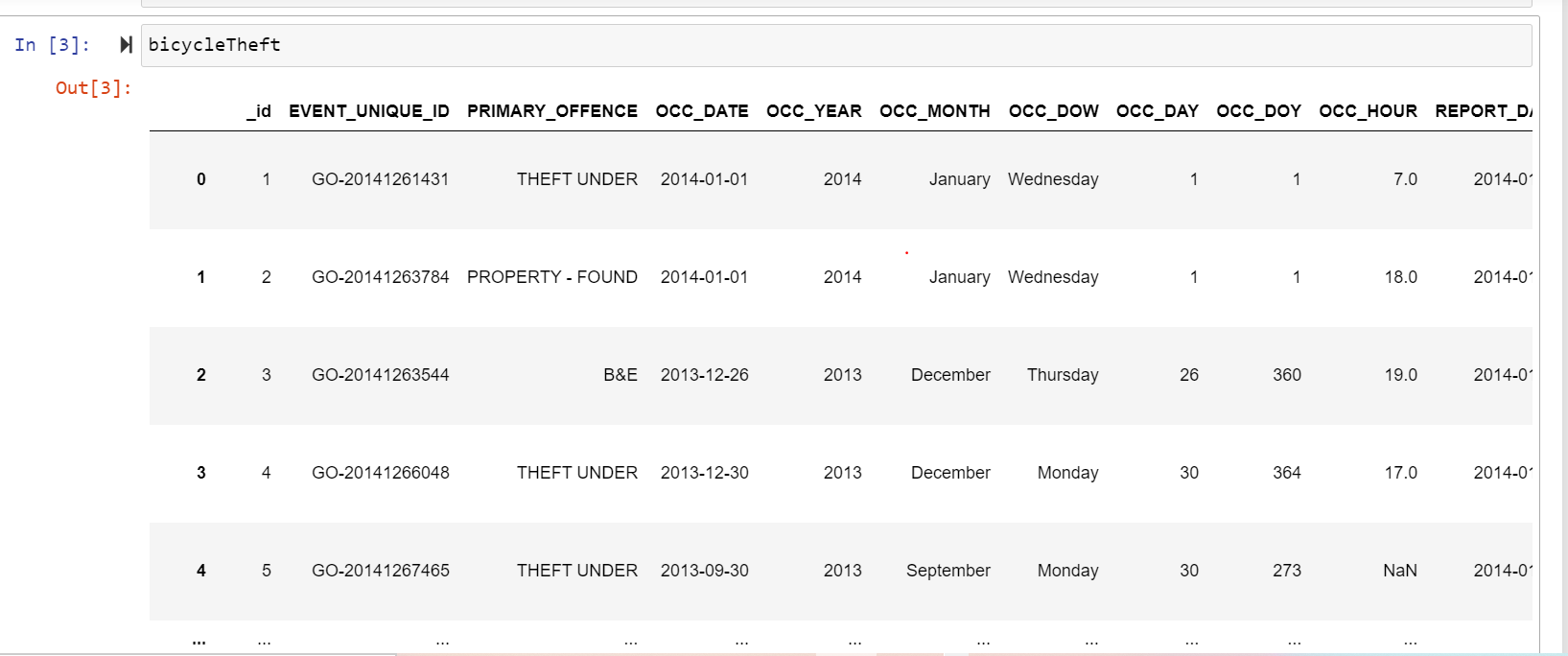
### 4.1 Collecting the data

The dataset in the .csv file is loaded in the Python program and is stored in a data frame named “bicycle Theft”.

Notice that r is used here to keep the UTF encoding as it is. The r is a prefix that indicates a raw string literal in Python. When dealing with file paths in Windows, backslashes are often used in paths, and a single backslash is an escape character in Python strings. Using a raw string literal (indicated by the r prefix) allows you to include backslashes in a string without interpreting them as escape characters.



Hence the bicycleTheft is displayed as follows -

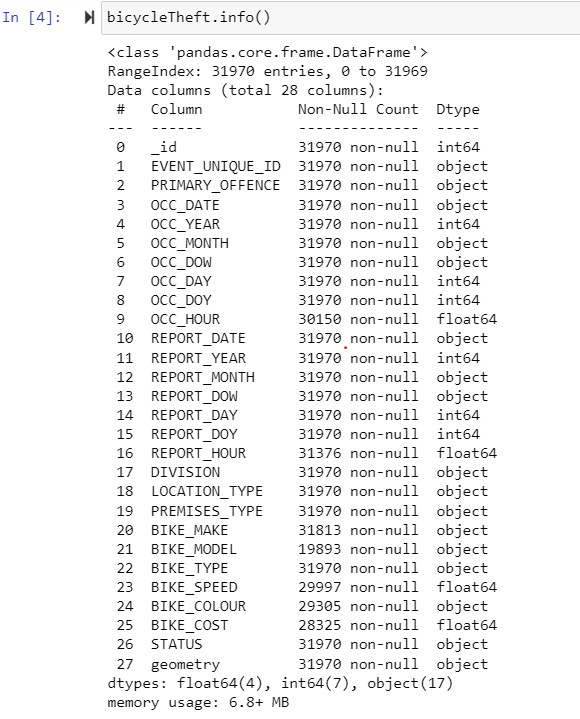


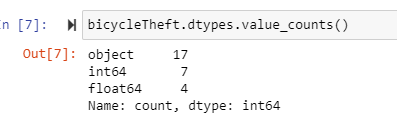
Therefore, the characteristics of the loaded data are described and, the below picture shows the basic information of the dataset.

* Total Number of Attributes:- 28
* Total Number of Instances:- 31970

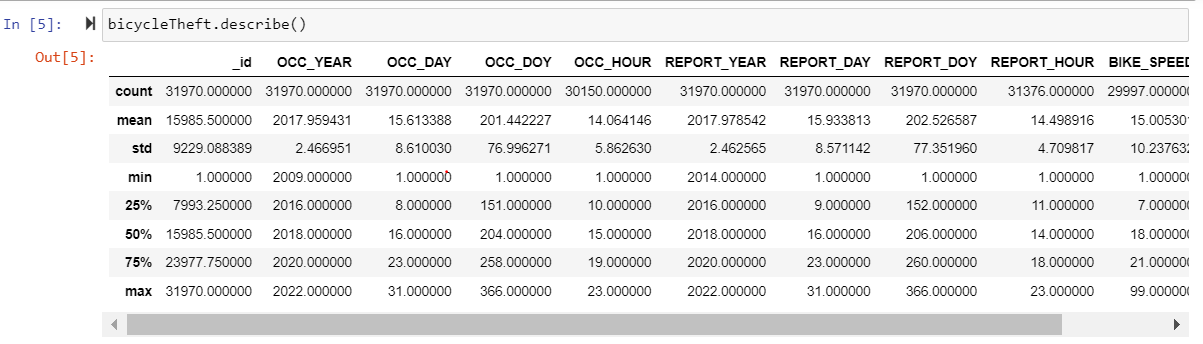
|  |  |
| --- | --- |
| **Attributes** | **Description** |
| id | Unique row identifier for Open Data database |
| EVENT\_UNIQUE\_ID | Offence Number |
| PRIMARY\_OFFENCE | Primary Offence Type |
| OCC\_DATE | Date Of Offence |
| OCC\_YEAR | Year of Offence |
| OCC\_MONTH | Month of Offence |
| OCC\_DOW | Day of the week offence occurred |
| OCC\_DAY | Day of the month offence occurred |
| OCC\_DOY | Day of the year offence occurred |
| OCC\_HOUR | Hour of the offence |
| REPORT\_DATE | Date of reporting the offence |
| REPORT\_YEAR | Reporting year of offence |
| REPORT\_MONTH | Reporting month of offence |
| REPORT\_DOW | Reporting Day of week of offence |
| REPORT\_DAY | Reporting Day of the month of offence |
| REPORT\_DOY | Reporting Day of the year of offence |
| REPORT\_HOUR | Reporting Hour of the offence |
| DIVISION | Police Division where the offence occurred |
| LOCATION\_TYPE | Type of Location where the offence occurred |
| PREMISES\_TYPE | Type of Premises/ area where the offence occurred |
| BIKE\_MAKE | Make of Bicycle |
| BIKE\_MODEL | Model of Bicycle |
| BIKE\_TYPE | Type of Bicycle |
| BIKE\_SPEED | Speed of the bicycle |
| BIKE\_COLOUR | Colour of the bicycle |
| BIKE\_COST | Cost of the bicycle |
| STATUS | Status of the offence |
| GEOMETRY | Describes the latitude and longitude of the area |

*Table 1*: *Description of Attributes*



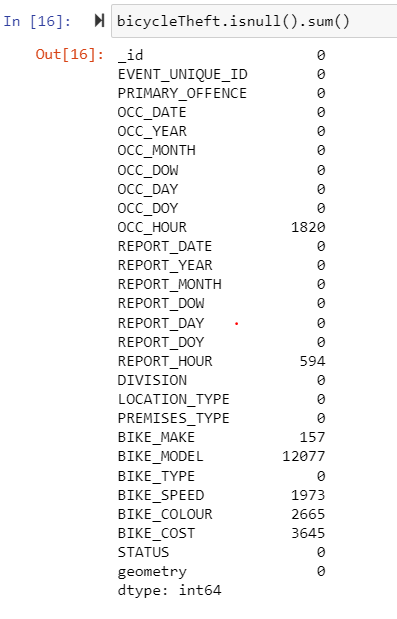


### 4.2 Statistical analysis of the numerical data



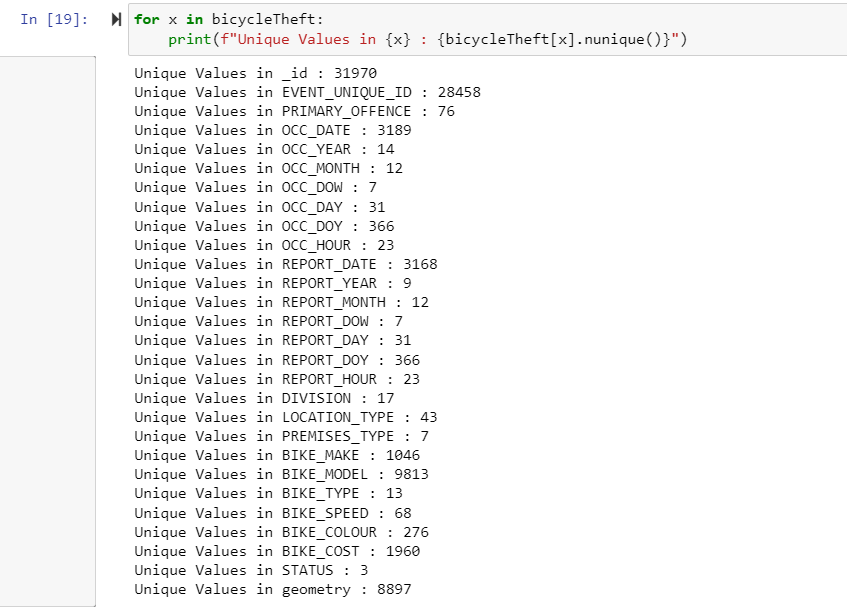
### 4.3 Checking the total number of null values in the data

Now to understand the data with more depth, the total number of null or missing values are analysed. Attributes with a large number of null/missing values should be excluded in order to maintain the data quality.



### 4.4 Checking the total number of unique values in the data

Data is also analysed by checked the unique values of the attributes. This helps in selecting attributes on the basis of their cardinality. Higher the cardinality results in model over fitting.

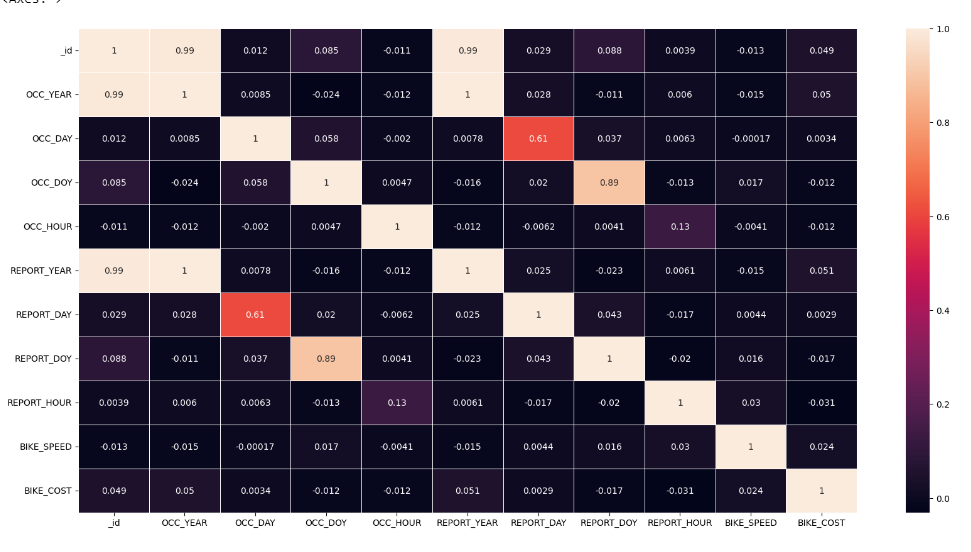


4.5 Checking the co-relation in the data

The correlation of data refers to the statistical relationship or association between two or more variables. In other words, it measures how changes in one variable are associated with changes in another variable. It is done via first creating a new data frame that stores all the columns with numeric values and then finding the co-relation of the data.

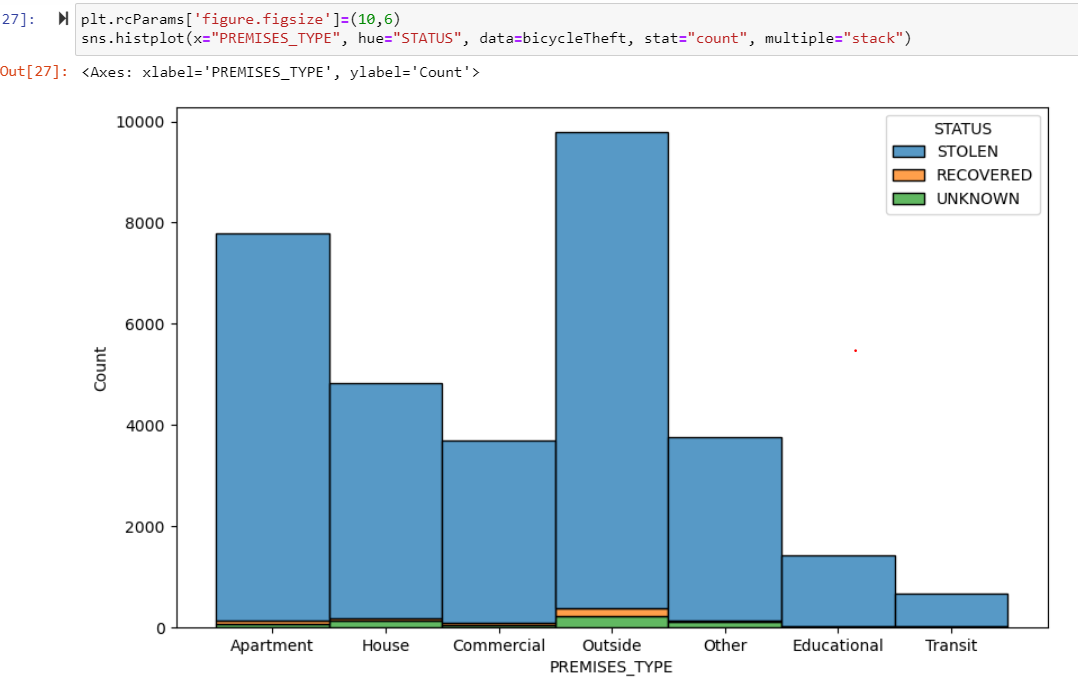
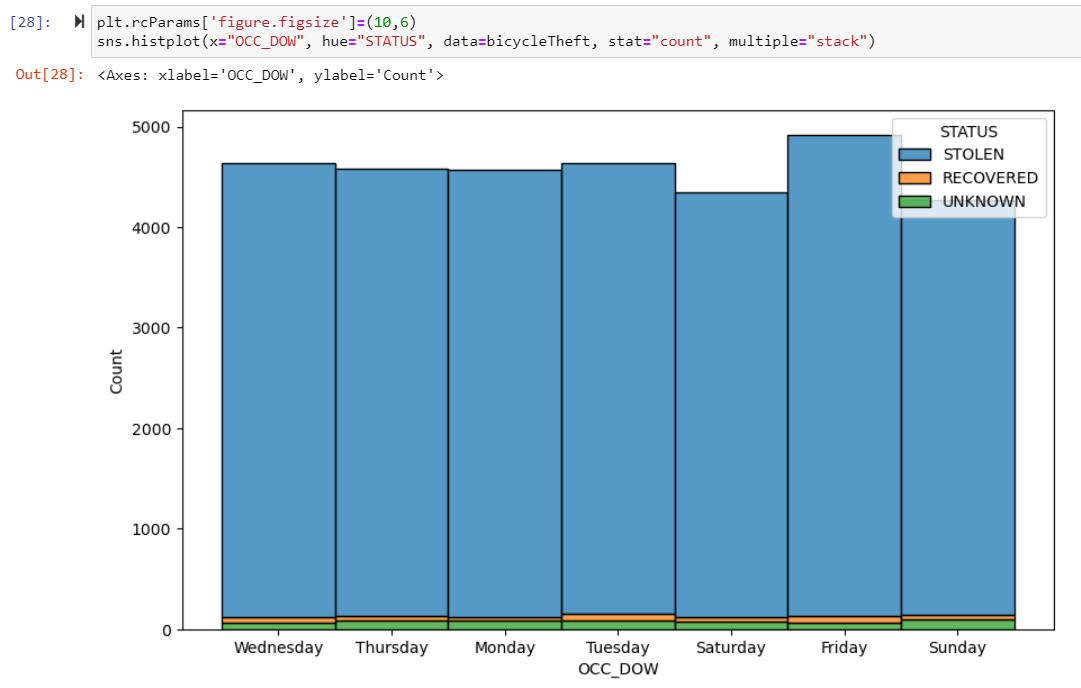
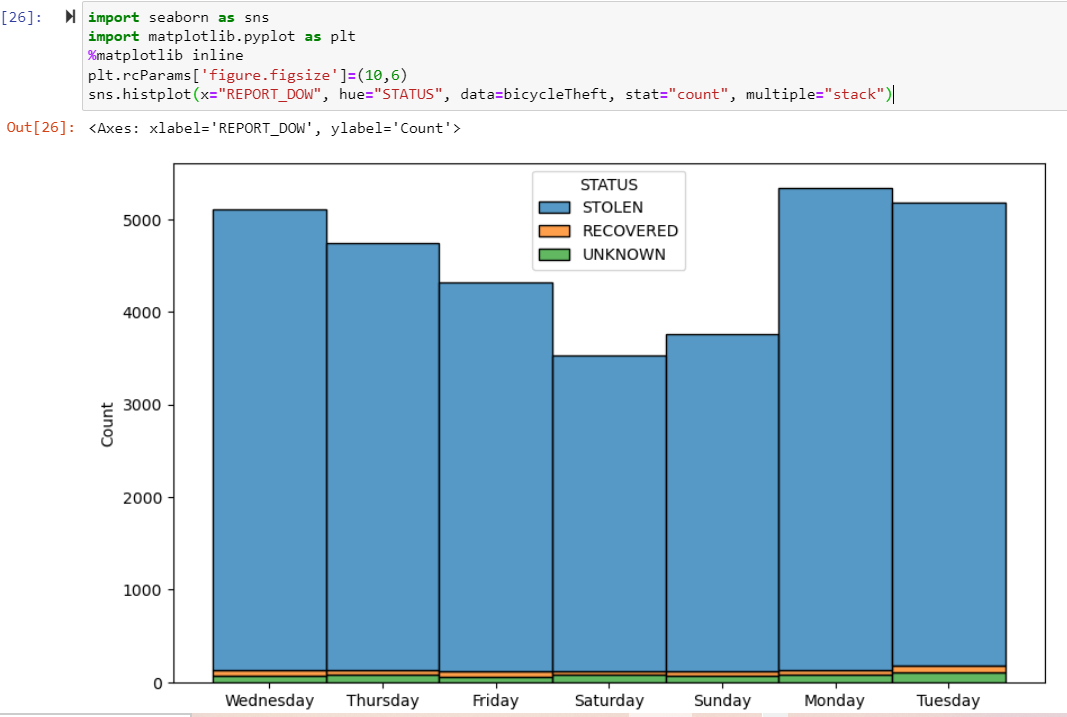
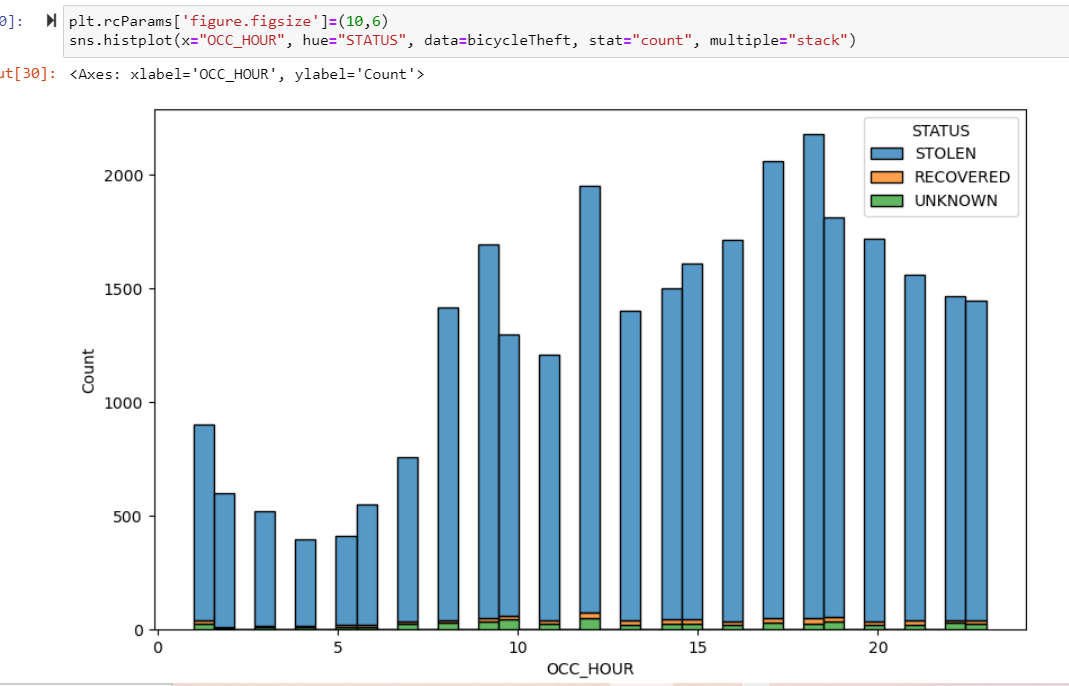
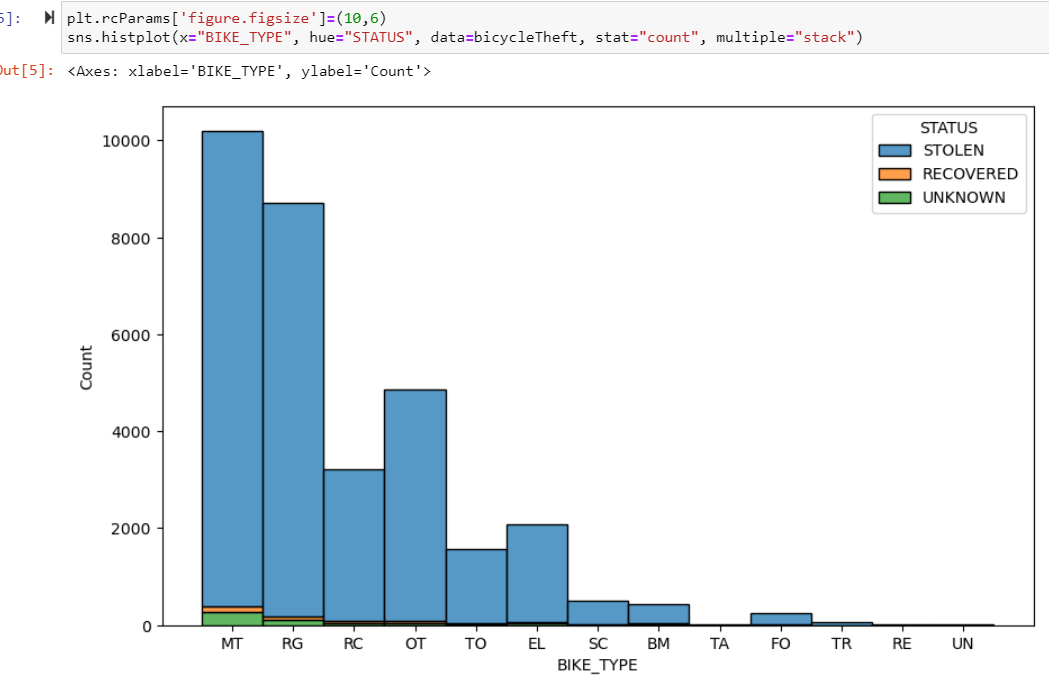
A screenshot of a computer screen

Description automatically generated



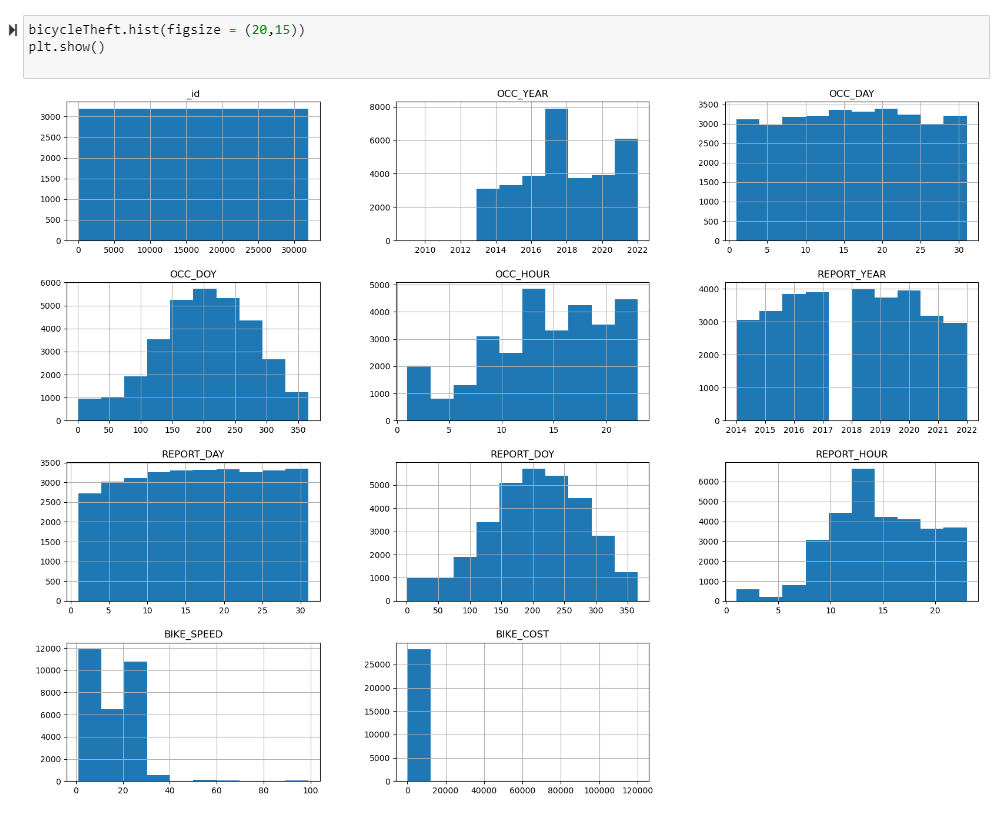
### 4.6 Understanding the visualization in the data

Data is also understood by looking at the visualizations in order to get a clearer picture of the relationships between the data which helps to solve machine learning algorithms in an easier way.

* **Plot depicting the type of premises where occurrences happened** It shows that a lot of bike thefts had occurred in **outside** premises.
* **Plot depicting the days when (days of the week) the occurrences happened.**It shows that most of the cases of bikes stolen were recorded on Friday.
* **Plot depicting the days of the week the incidents reported**It shows that maximum cases of bicycles stolen were reported on Monday with least during the weekends.
* **Plot depicting the hour of the day occurrences were evident:-**It shows that the maximum number of bikes were stolen during the evening hours.
* **Plot depicting the type of bike stolen at the most:-**

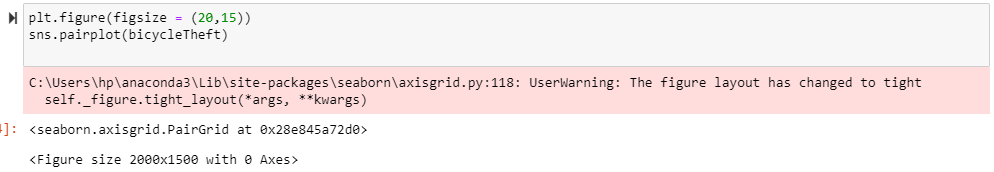
Several other plots are also created (please refer to the Python code attached with the deliverable).

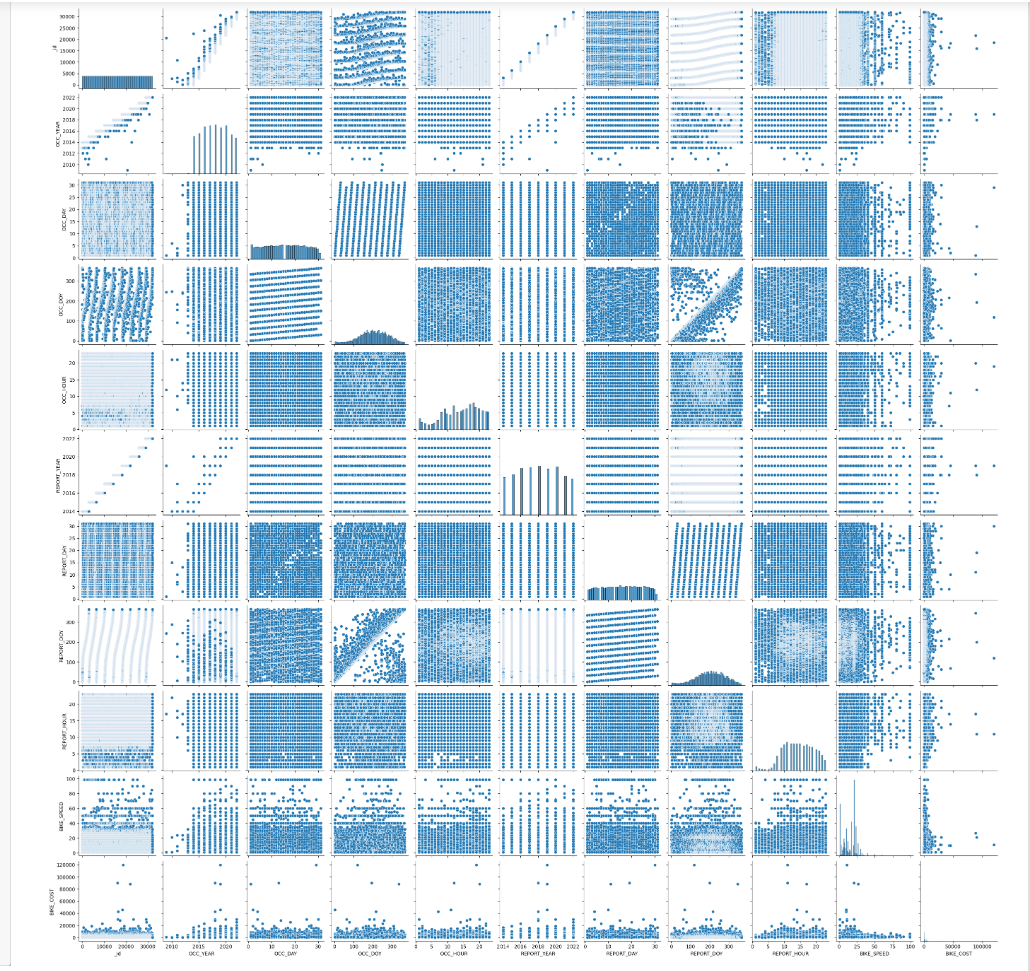
**Histograms for the frequency count of each column: -**



**Pair plot of the data :-**

A pair plot is a grid of scatter plots and histograms that allows you to visualize relationships between numerical variables in a dataset. Each scatter plot shows the relationship between two variables, and histograms show the distribution of each individual variable. The pair plot function will generate scatter plots for all numerical variables in the dataset, and histograms along the diagonal for each individual variable.



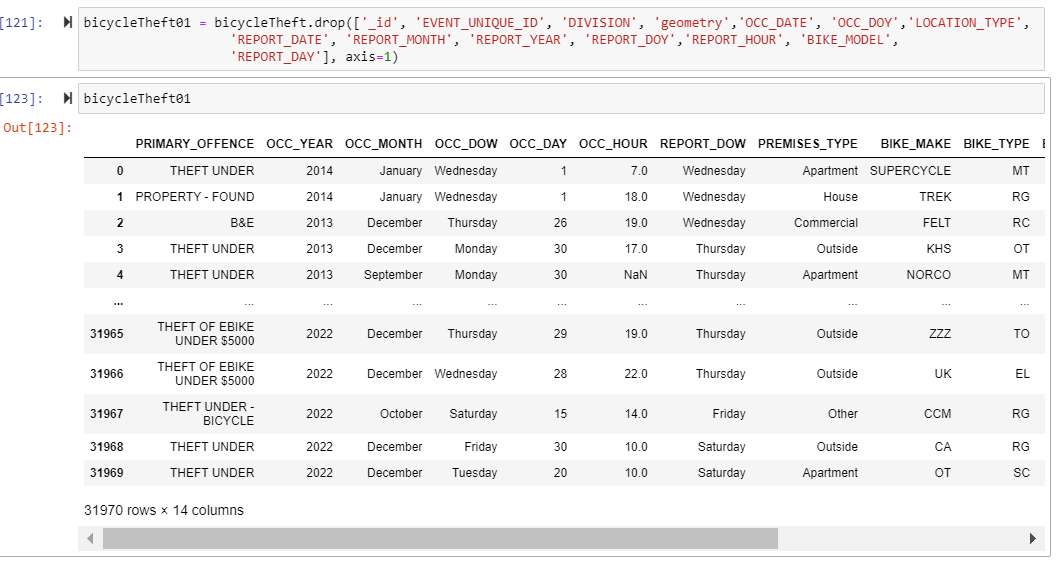


# Data Preparation

It involves steps such as **Selection of data and Cleaning unwanted data** under which the relevant attributes to be used for classification of the dataset and removing all the unwanted attributes are decided. It is made sure that the data selected is free of unwanted values, outliers, missing values, too many unique values in order to have the accurate data classification models. Data is then **integrated** and **formatted** to form a new updated data frame by performing operations such as binning etc.

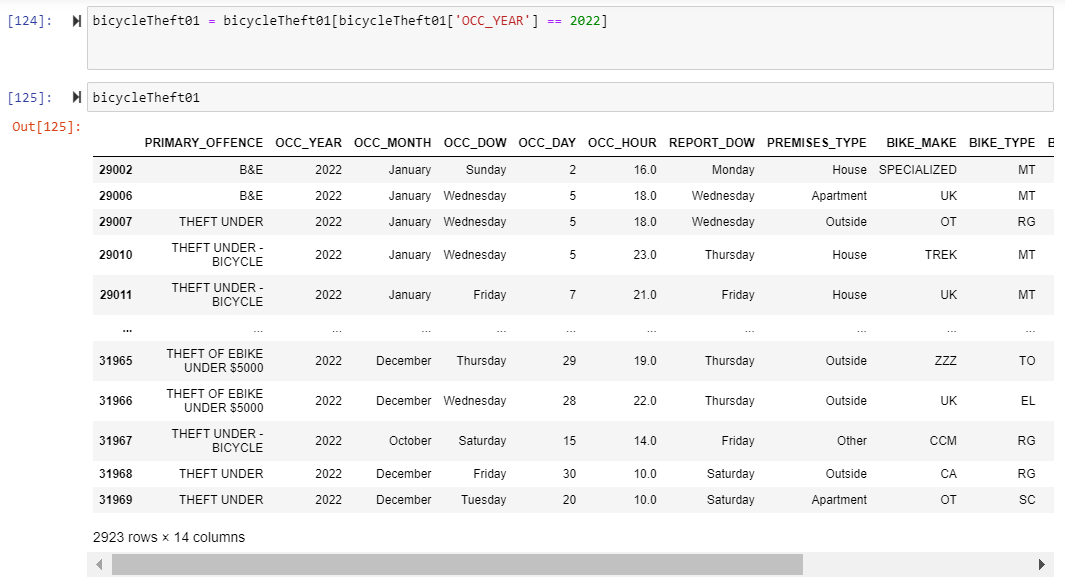
### 5.1 Selection of Data and Cleaning Unwanted Data

|  |  |
| --- | --- |
| **Attributes** | **Included / Excluded** |
| \_id | Excluded - Not needed for the project’s scope. |
| EVENT\_UNIQUE\_ID | Excluded - Not needed for the project’s scope |
| PRIMARY\_OFFENCE | Included - type of the offence |
| OCC\_DATE | Excluded - Not needed for the classification purposes. We have used the individual attributes depicting the day, month and year. |
| OCC\_YEAR | Excluded - as the whole machine learning process is done for the data of 2022 |
| OCC\_MONTH | Included - in order to classify the data as per the time when thefts occured |
| OCC\_DOW | Included - in order to classify the data according the day of the week. |
| OCC\_DAY | Included - in order to classify by looking the number of cases on a particular day of month. |
| OCC\_DOY | Excluded - Not needed for the project’s scope |
| OCC\_HOUR | Included - In order to classify the data according to the timing of the offence occured |
| REPORT\_DATE | Excluded - Not needed according to the question as we are dealing with the time of offence occurred instead of timing of offence reported |
| REPORT\_YEAR | Excluded - Not needed according to the question as we are dealing with the time of offence occurred instead of timing of offence reported |
| REPORT\_MONTH | Excluded - Not needed according to the question as we are dealing with the time of offence occurred instead of timing of offence reported |
| REPORT\_DOW | Included- In order to perform classification further based on the days when the incidents reported |
| REPORT\_DAY | Excluded - Not needed according to the question as we are dealing with the time of offence occurred instead of timing of offence reported |
| REPORT\_DOY | Excluded - Not needed according to the question as we are dealing with the time of offence occurred instead of timing of offence reported |
| REPORT\_HOUR | Excluded - Not needed according to the question as we are dealing with the time of offence occurred instead of timing of offence reported |
| DIVISION | Excluded- Not needed for the project’s scope, the demand of question does not include the police division where the offence reported |
| LOCATION\_TYPE | Excluded - Not useful as it is containing a lot of unique values which increases the cardinality making the data difficult to handle |
| PREMISES\_TYPE | Included - In order to classify the type of area where offence occurred. |
| BIKE\_MAKE | Included - In order to classify the type of bike on the basis of bike make. |
| BIKE\_MODEL | Excluded - Containing a lot of missing values and the cardinality is higher which does not contribute to efficient model (around 1377 unique values out of 2923) |
| BIKE\_TYPE | Included - to classfy the data on the basis of bike type |
| BIKE\_SPEED | Included -has a good effect on the algorithm accuracy |
| BIKE\_COLOUR | Included - to further distinguish the bike theft cases on the basis of the bike colour |
| BIKE\_COST | Included - to further distinguish the bike theft cases on the basis of the bike cost |
| STATUS | Included - class of the dataset |
| geometry | Excluded - not needed at all |

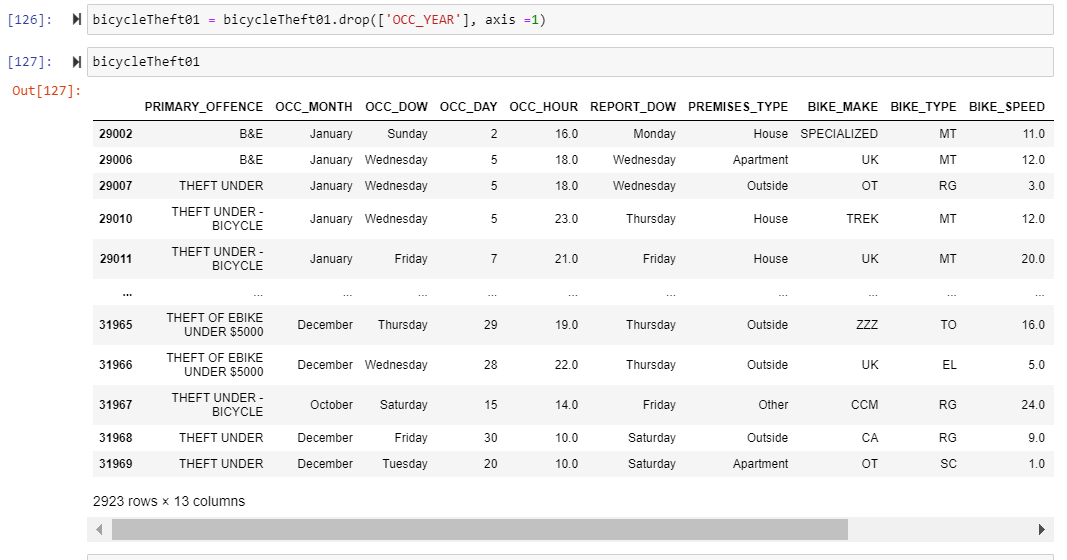


5.2 Filtering out the data only for the latest year (2022) **-**

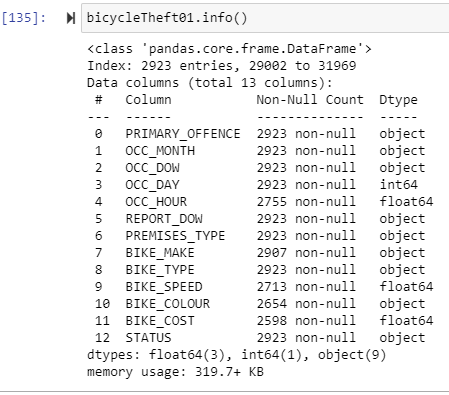
If we have a lot of data from multiple years, we can consider data from the latest year-



And then dropping that column as well:-

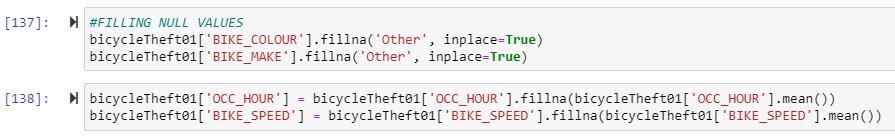


So, the resultant data frame we are working on with:-



### 5.3 Handling the missing or NULL values

The data columns which are included to perform the further classification are also checked with any missing and null values. Therefore, we ran the python code segment that fills missing values in case of BIKE\_COLOUR and BIKE\_MAKE to “Other” and OCC\_HOUR and “BIKE\_SPEED” by taking the average value of all the instances as they fall under continuous data.



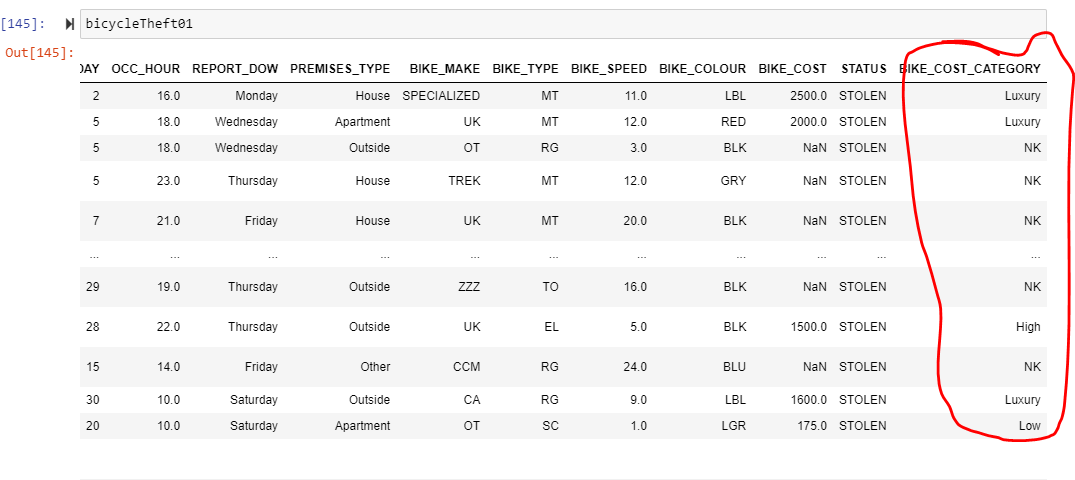
### 5.4 Converting the Continuous Data Attributes into Categorical Data;

This step is known as binning which is done on the cost of the bike to depict what type of class the bike fits among Low, Average, High, or Luxury car and the unknown or NULL cost is marked as NK.

Here the Python code segment is generated that performs a custom binning that generates 4 bins (**Low, Average, High, Luxury**) of equal width (which was done by calculating the quantile), and the NaN or the null values are classified as **NK.**



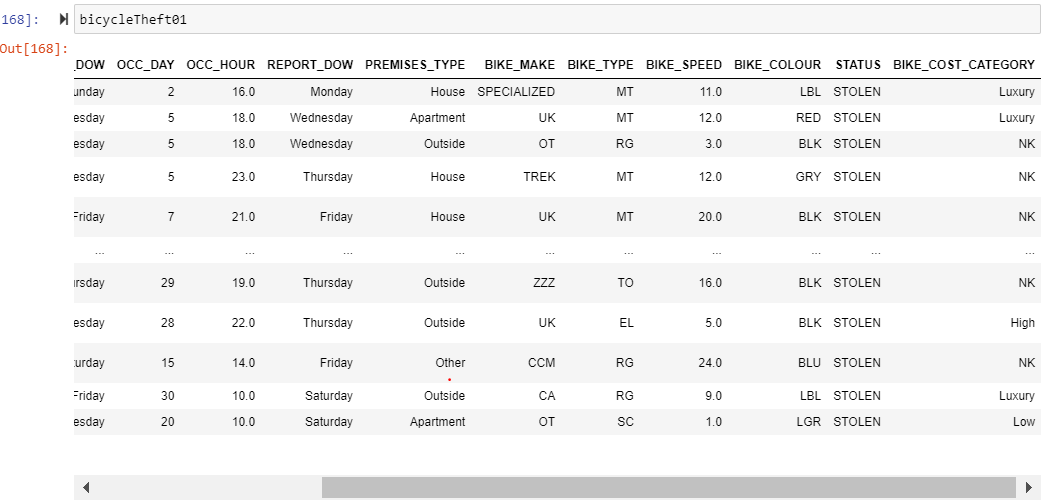
This creates a new attribute called BIKE\_COST\_CATEGORY which stores the values according to the bike cost. Here we can drop the BIKE\_COST attribute now.

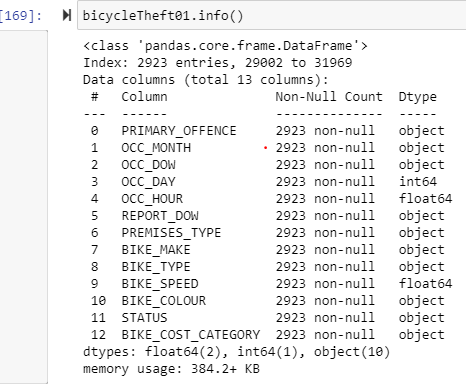


Deleting the attribute BIKE\_COST:-



Hence the data frame looks like:-



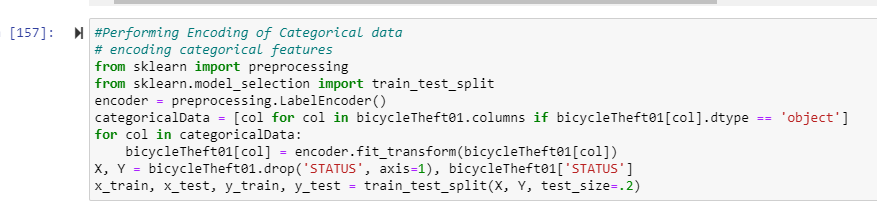


### 5.5 Encoding the categorical data and splitting the model into train and test models:-

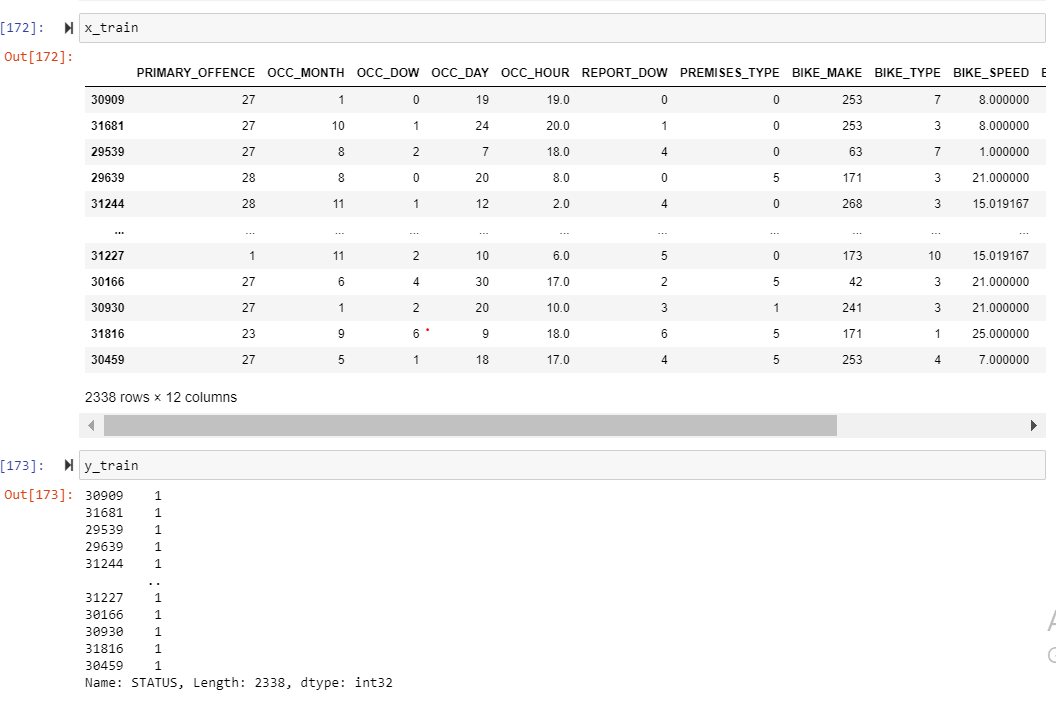
Here using the encoder class imported from pre-processing, the categorical values are encoded so as to completely convert the whole table into Categorical data and having the encoded values making it easier for the classification models.

After that, for the data modelling, there are two separate types of models generated known as train and test data comprising 80% and 20% of data respectively. When working with supervised machine learning, it's common to split the dataset into features (often denoted as X) and labels (often denoted as y). The reason for this separation is rooted in the nature of supervised learning tasks, where the goal is to train a model to predict labels (target variables) based on input features.

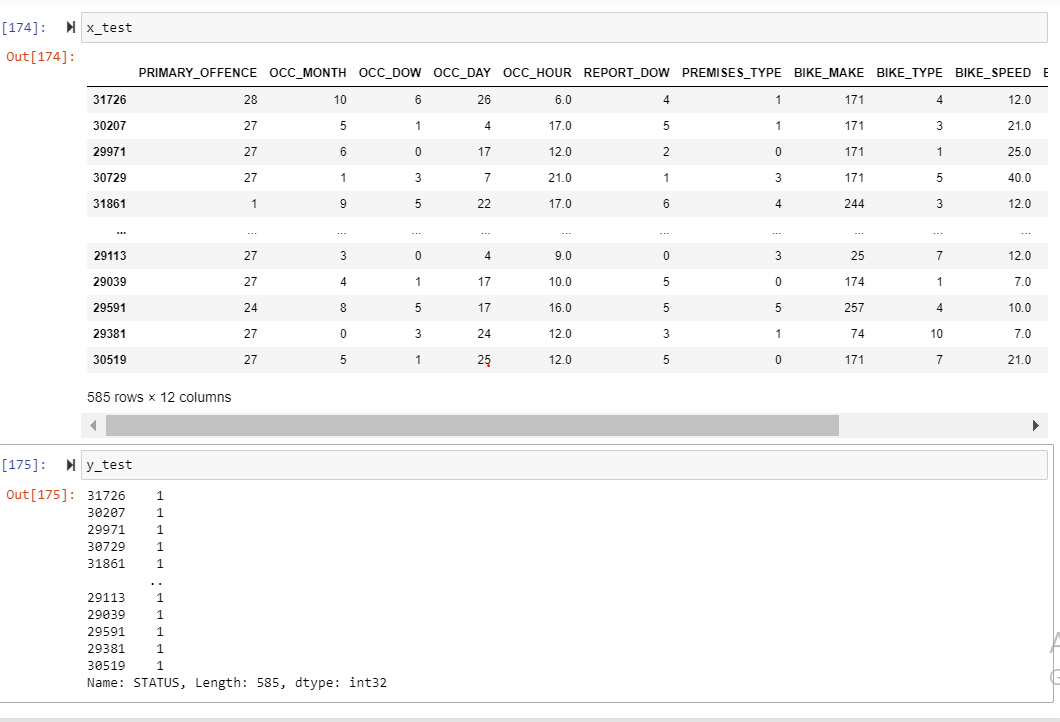
Hence, the models are classified into training and test sets as x\_train, y\_train, x\_test, y\_test.



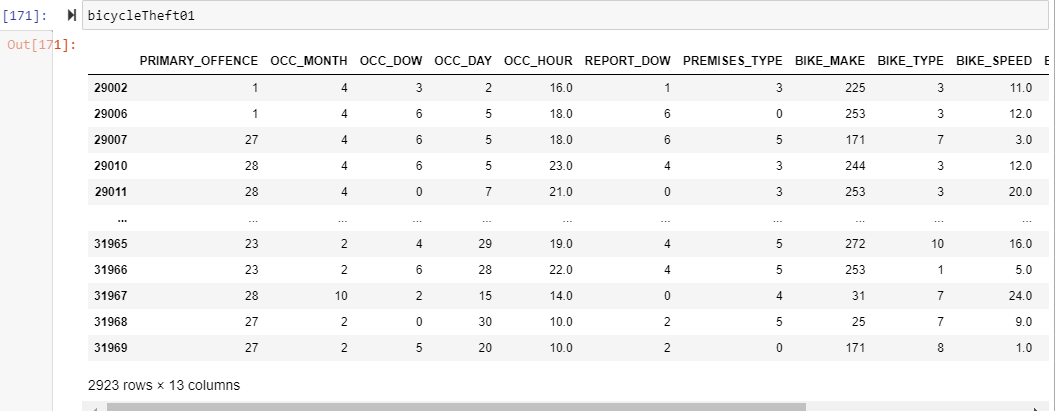
**X\_train and y\_train**



**x\_test and y\_test**



### 5.6 Resultant data frame - bicycleTheft01 after encoding



Now the data is completely ready for classification, outlier detection, kNN, clustering and further modelling processes.

# Conclusion

To summarize, upon looking for a perfect predictive model for data analysis of the theft cases reported in Toronto and to determine the highest geographic area prone to such criminal activities, the type of bikes being the victims of burglary and the peak times for the same, the data is completely prepared and free of any errors and null or missing values. Out of all the steps of CRISP-DM, this assignment includes Business Understanding - talking out the business objective of this dataset and project; Data Understanding - understanding the relationship of data with each other, determining the null and unique values and analysing the data set statistically as well as visually; and Data Preparation- preparing the data by cleaning the unwanted data and constructing a new categorical model to accurately build models such as kNN, Decision Tree, Clustering as well as performing outlier detection etc.

1. <https://www.sv-europe.com/crisp-dm-methodology/#:~:text=CRISP%2DDM%20stands%20for%20cross,robust%20and%20well%2Dproven%20methodology.> [↑](#footnote-ref-1)
2. <https://thinkinsights.net/data/crisp-dm/#:~:text=The%20CRISP%2DDM%20approach%20helps,project%20against%20its%20original%20objectives.> [↑](#footnote-ref-2)
3. <https://www.datascience-pm.com/crisp-dm-2/> [↑](#footnote-ref-3)
4. <https://open.toronto.ca/dataset/bicycle-thefts/> [↑](#footnote-ref-4)
5. <https://towardsdatascience.com/demystifying-bicycle-theft-cases-in-toronto-which-neighborhoods-should-get-more-attention-1ff273115474> [↑](#footnote-ref-5)