**Scalable Database (CS673)**

**Midterm Project**

**Rutik Kothwala (**[**rk29015n@pace.edu**](mailto:rk29015n@pace.edu)**)**

**Chandani Thumar (**[**ct26393n@pace.edu**](mailto:ct26393n@pace.edu)**)**

**Problem Definition:**

Task of the midterm project is Exploratory Data analysis.

1. To do an in depth analysis of the Mushroom data so we will be able to know how different features are related to each other and which are the most important features in order to classify which Mushroom is Edible or Poisonous .
2. To Join the unstructured data in form of images with the Mushroom data set and connect it to postgres database.

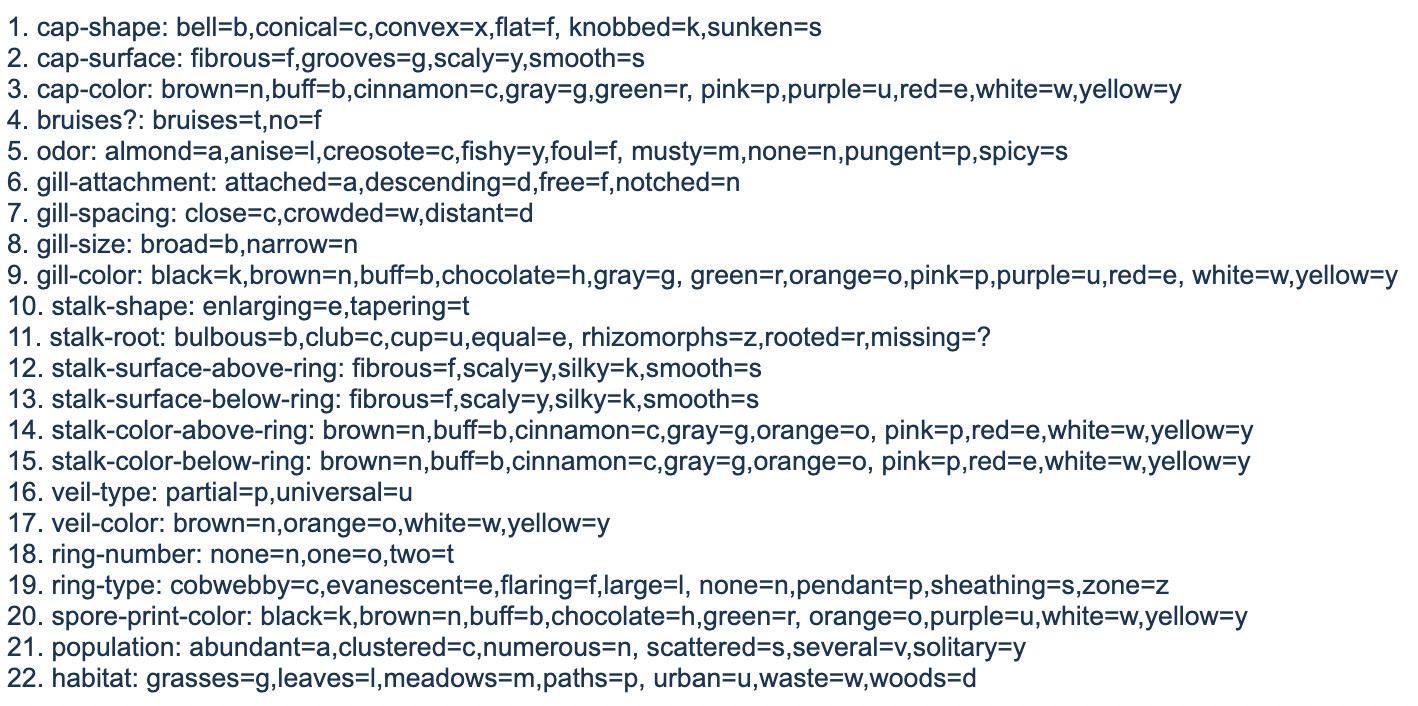
The dataset we are using is of Mushrooms. The Mushrooms dataset contains 8000 rows and 22 columns. Each species is defined in two categories which are Edible and poisonous .

In Total there are 13 species of Mushrooms in our dataset.

We are using images of these 13 species in the form of unstructured data.

**Dataset Description :**

Mushrooms Datasets Columns :





**Operations of EDA:**

1. Loading data into DataFrames. Integration of SQL and Python

:- Loaded the DataFrame of Mushroom Dataset using pandas in python. For Integration with Postgresql, extracted 13 mushroom species which are Edible and Poisonous depending open their habitat. Performed code in python to create a Unstructured database with images in Postgresql.

2. Check the Data Types of your data columns.

:- Checked the Data types of columns in Mushroom Dataset and found all column types are Object.

3. Drop any NULL, missing values or unwanted columns.

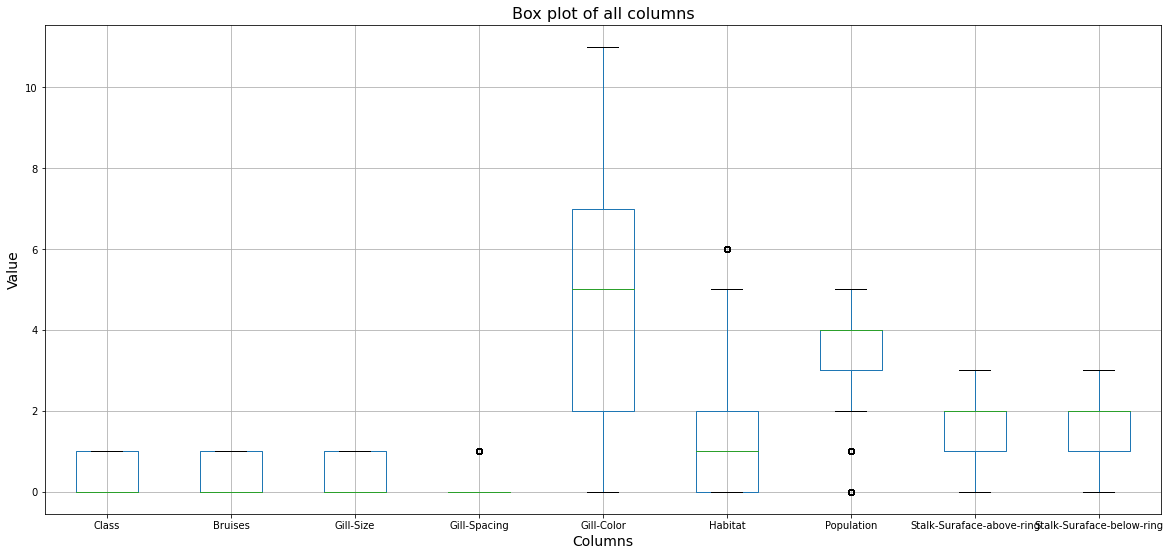
:- Performed one operation to check NULL and missing value in our dataset and found out there are no NULL values but found 2480 missing value in Stalk- root which was written as ‘?’ (Question Mark).

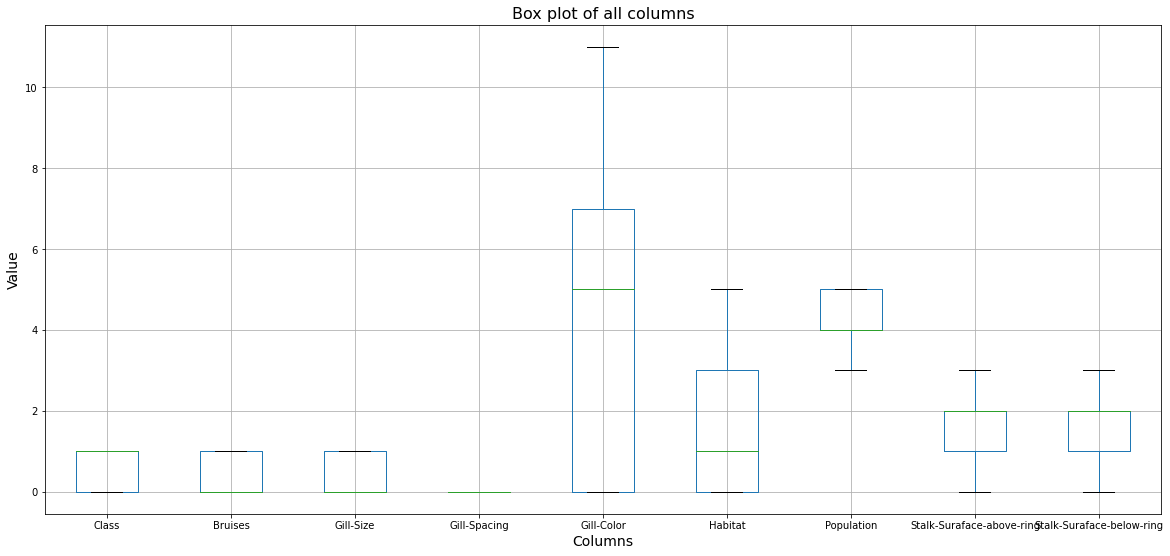
Performed Countplot operation from seaborn library on Stalk root columns and observed that ‘b’ category have more count in that column so according to the mean idea, replaced all the Missing value of ‘?’ as the ‘b’ category which is bulbous.

4. Drop duplicate values.

:- There are no duplicate values in Mushroom Datasets.

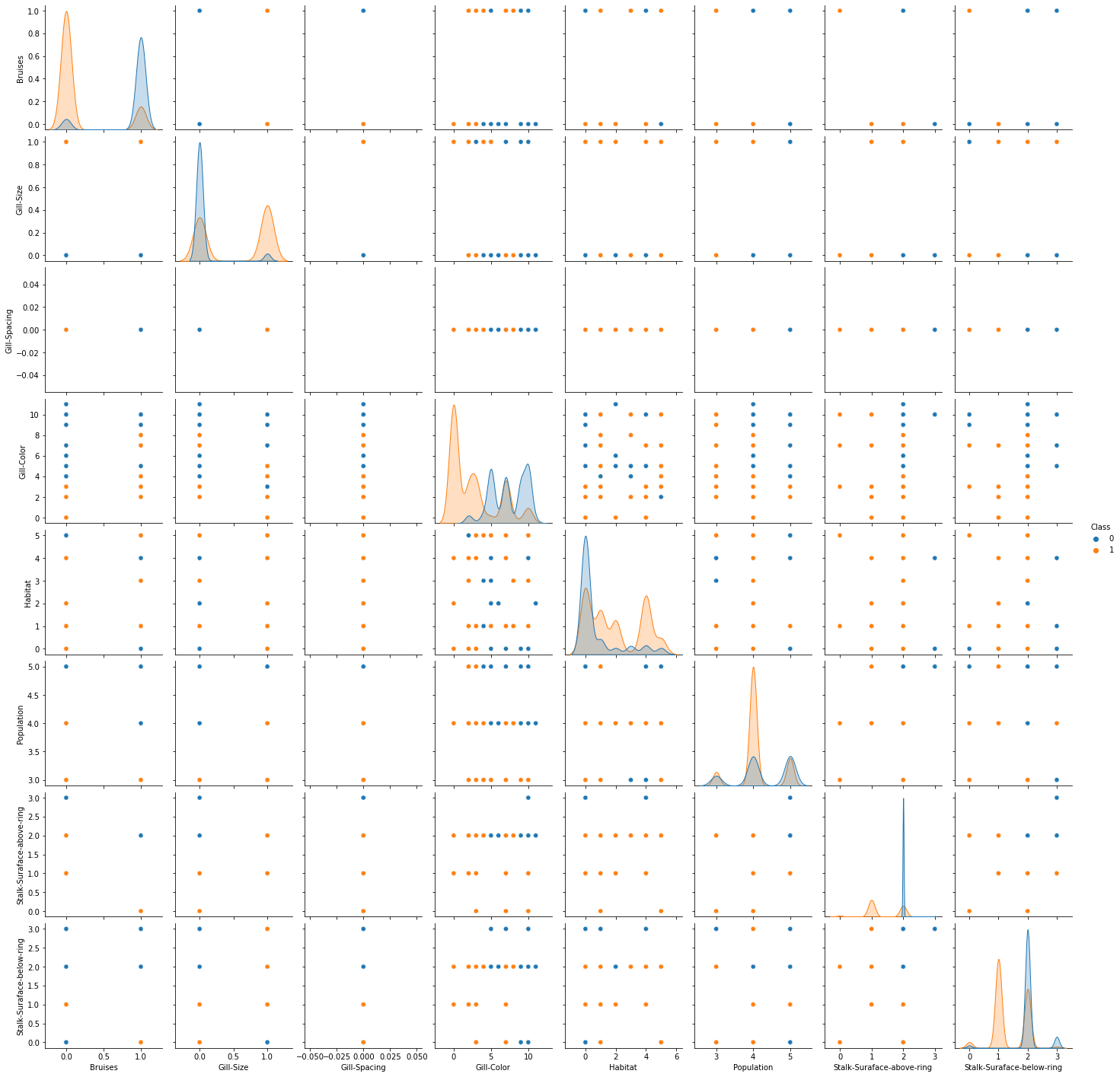
5. Check for outliers using a box plot or histogram.

:- Checking the outliers for Mushroom\_new\_dataset and found outliers in 3 columns that are Gill-Spacing,Habitat,Population. Removed outliers from all columns using boxplot and left with the 6232 rows.



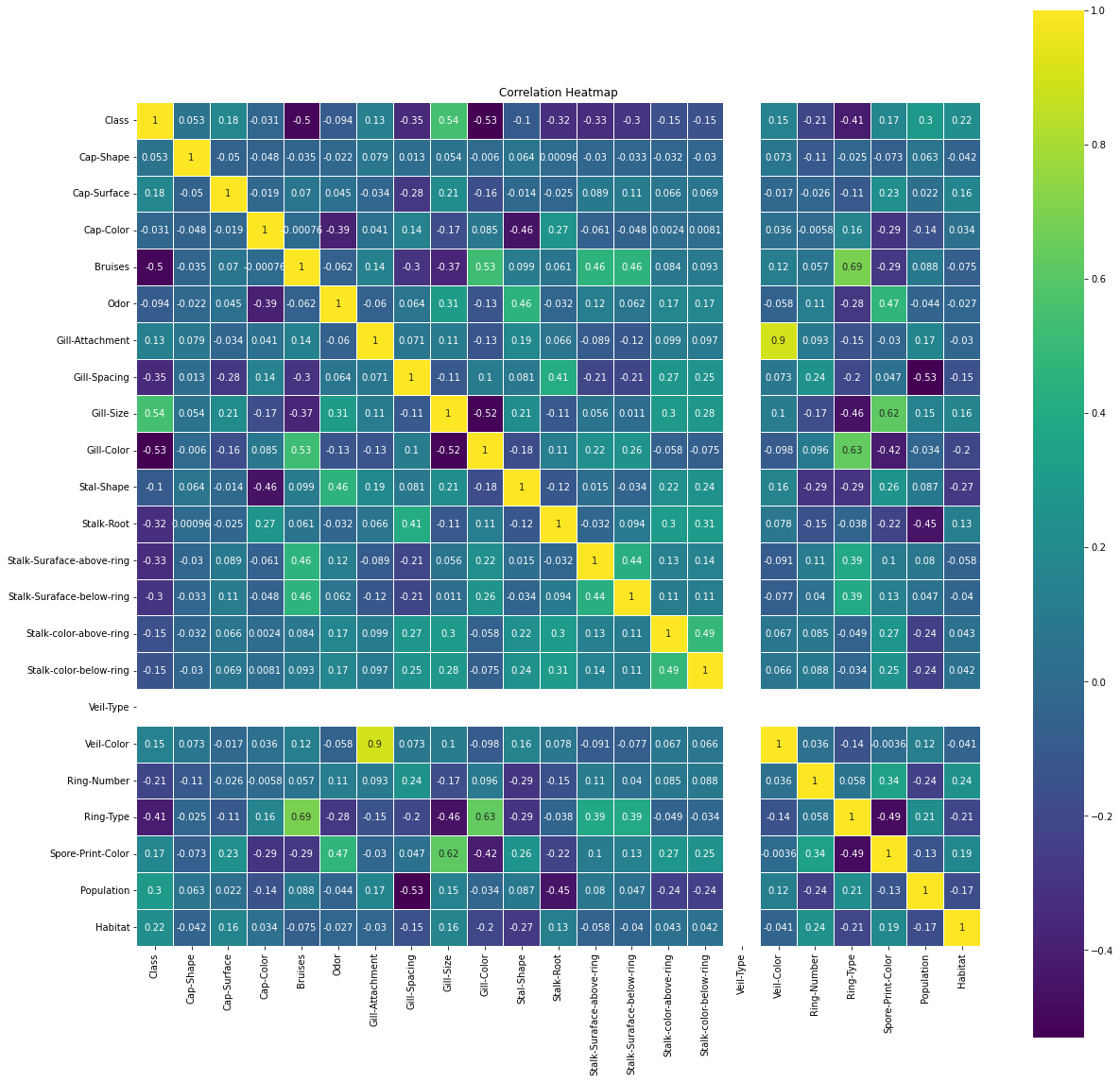
6. Plot features against each other using a pair plot.

:- Performed a pairplot function on Mushroom\_new\_dataset and we can observe the variation in every graph. The plots are in the form of matrix where the row name represent in the x axis and columns name represent in the y axis



7. Use a HeatMap for finding the correlation between the features(Feature to Feature).

:- Performed code for Heatmap to check the Correlation between all columns and after that removed the unwanted columns by checking correlation between the columns with their target column.



8. Use a scatter plot to show the relationship between 2 variables.

:- Performed scatter plot between class and Habitat.



9 Merging two Data Frames

:- Merged our 13 species of Unstructured data of mushroom with original dataset of mushroom which contains more than 8000 rows with the use of pandas join function.

10 Slicing Data of a particular column value (like year, month, filter values depending on the categorical data)

:- Created a new dataset of Edible mushroom from the original dataset.

11:- Representing data in matrix form.

:- Dataset of Edible mushroom , converted into the matrix form and after that stored in the array form.

12:-Upload data to Numerical Python (NumPy)

:- 13 species of images converted to numpy arrays and stored as a dataframe.

13. Select a slice or part of the data and display

:- We sliced the data of the column Class, Gill-color and Gill size and performed Slicing operation on that dataframe

14. Use conditions and segregate the data based on the condition (like show data of a feature(column) >,<,= a number)

:- Used condition and segregate function given below:

Equal : To find gill-color of mushroom which is equal to 10(Red)

And : To find where class and gill size are 0(Poisonous) and 1(Broad).

15. Use mathematical and statistical functions using libraries.

:- Mathematical functions which are used are given below.

Mean, Sum, Median, ValueCount, Describe

16. Select data based on a category(categorical data based)

:- All the features of mushroom Dataset are in the categorical form.

17. Libraries expected to try(minimum 4 required): Pandas, Numpy, Seaborn, Matplotlib

:- The following libraries used are mentioned below.

Pandas, Numpy, Seaborn, Matplotlib, Sweetviz, Pandas Profiling, Psycopg2, Os, CSV

18. Write your own functions and handle exceptions in the functions.

:- Used handle exception function which is given below:

**try:**

**cur.execute(query)**

**except (Exception, psycopg2.DatabaseError) as error:**

**print("Error: %s" % error)**

**cur.close()**

**return 1**

19. Use of \*arg and \*\*kwargs

:- Filter the data to show only edible mushrooms with Habitat as Urban

20. Use of data functions.

:- Used data functions such as describe and info.

**Thank You!!**