**Experiment 1**

**Student Name:** Sahil Kaundal **UID:** 21BCS8197

**Branch:** BE CSE (Lateral Entry) **Section/Group:** 616/A

**Semester:** 5th **Date of Performance:** 23/08/2022

**Subject Name:** ML Lab **Subject Code:** 20CSP-317

1. **Aim/Overview of the practical:**

Implement Exploratory Data Analysis on any data set.

**2. Apparatus / Simulator Used:**

1. Windows 7 or above.
2. Google Collab.

**Heatmap:** It is a plot of rectangular data as color encoded matrix as parameters. It takes 2D datasheet. It is the best way to analyze the data because it can show the relation between variables.

**Pair plot:** When you want to analyze only two variables where variables can be continuous, Boolean as well. It is basically a grade of plot for each variable in your dataset.

**3. Program / Commands:**

#Sahil Kaundal

#21BCS8197

import pandas as pd

import seaborn as sns

#Load the Data

data=pd.read\_csv('/content/sample\_data/california\_housing\_test.csv')

#View the data

data.head()

data.tail()

data.shape

corelation=data.corr()

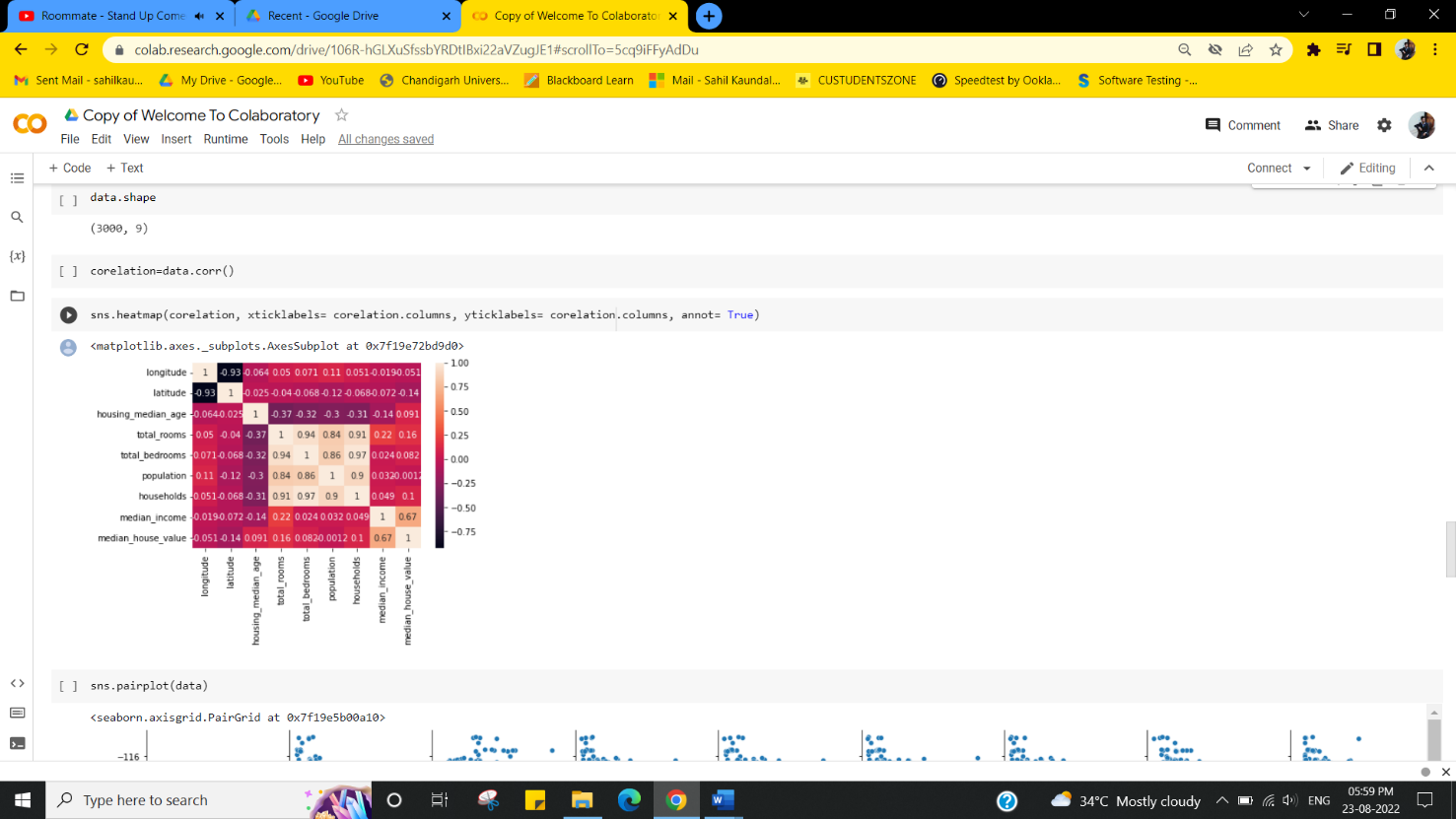
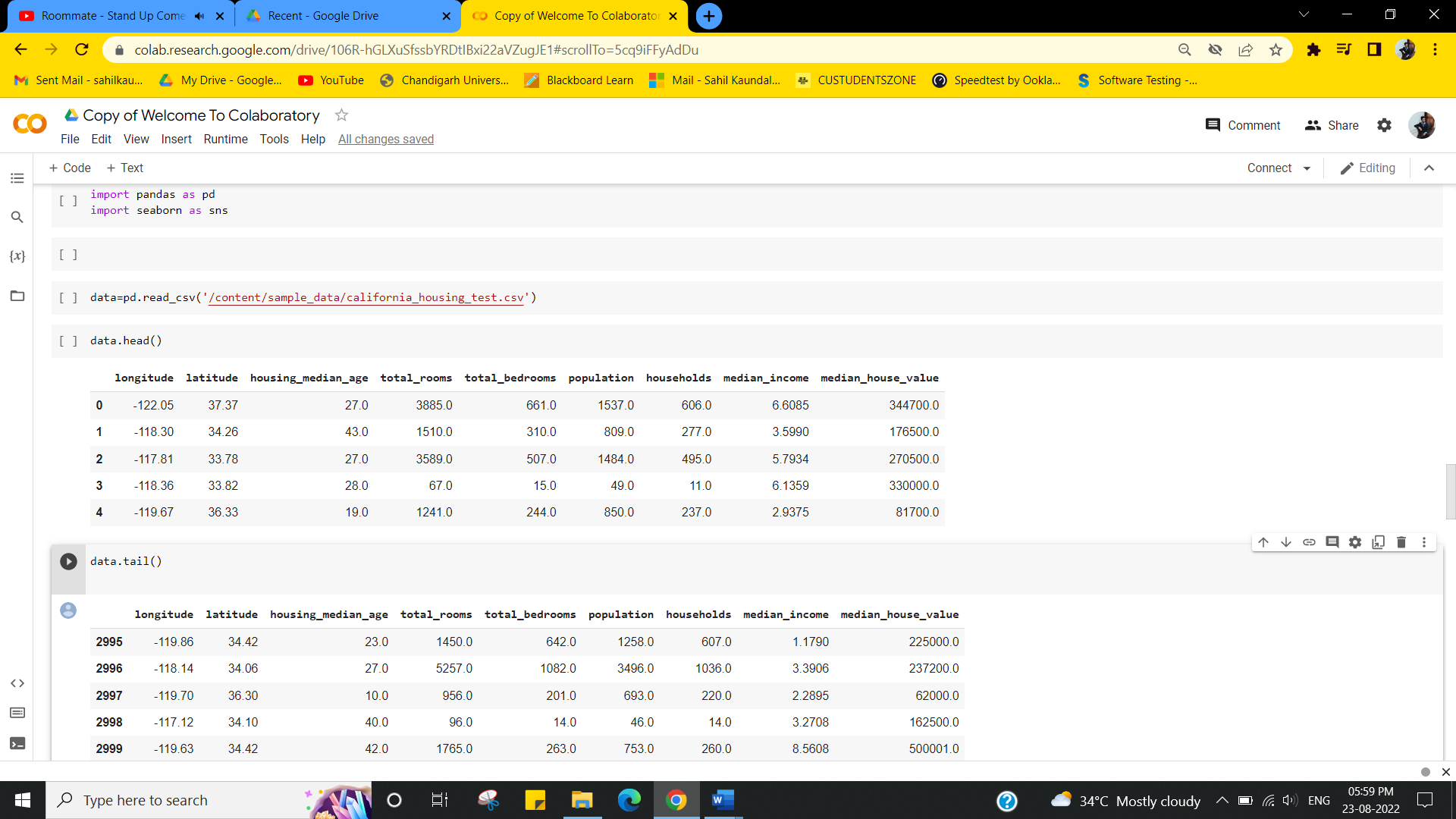
sns.heatmap(corelation, xticklabels= corelation.columns, yticklabels= corelation.columns, annot= True)

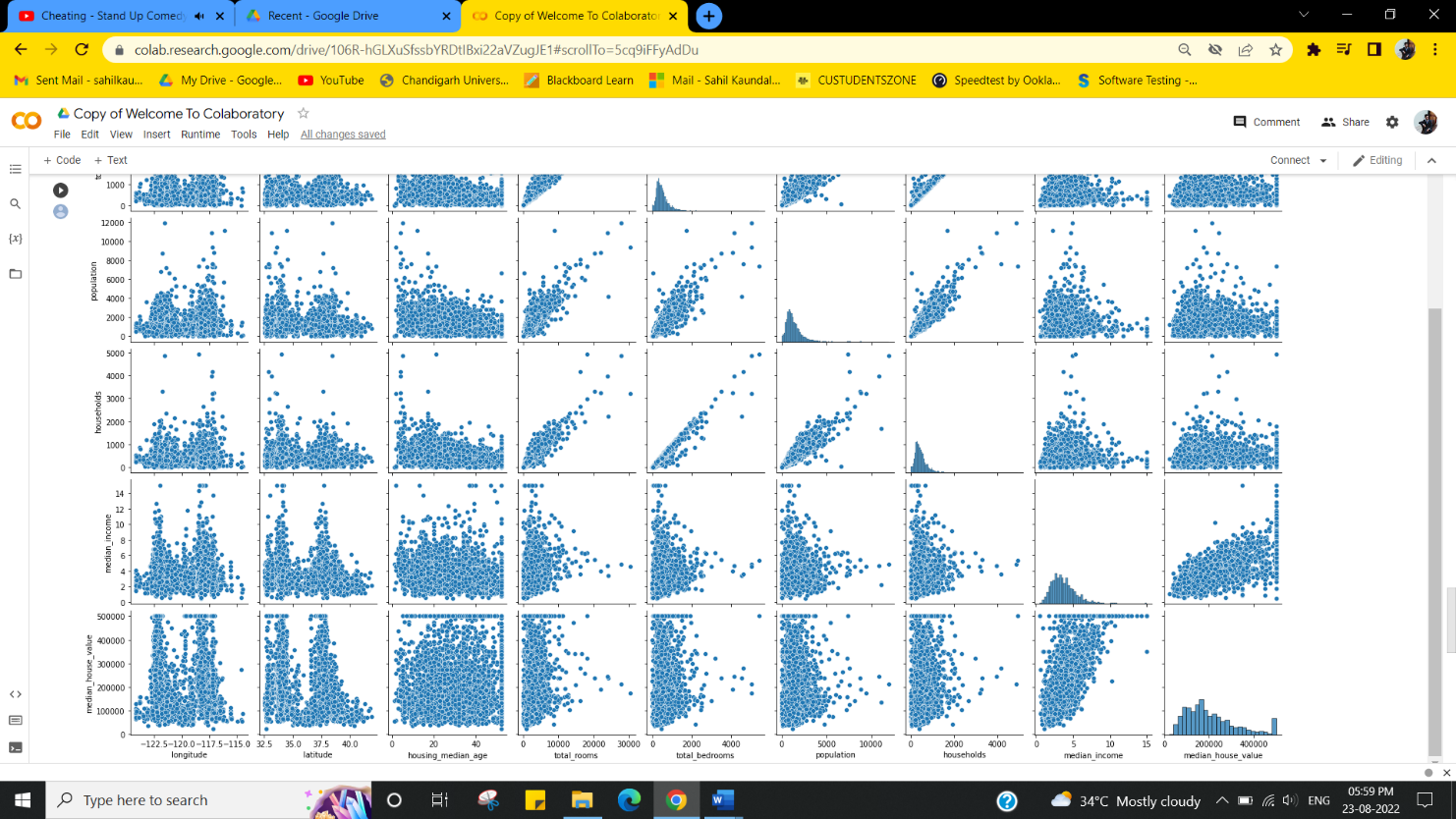
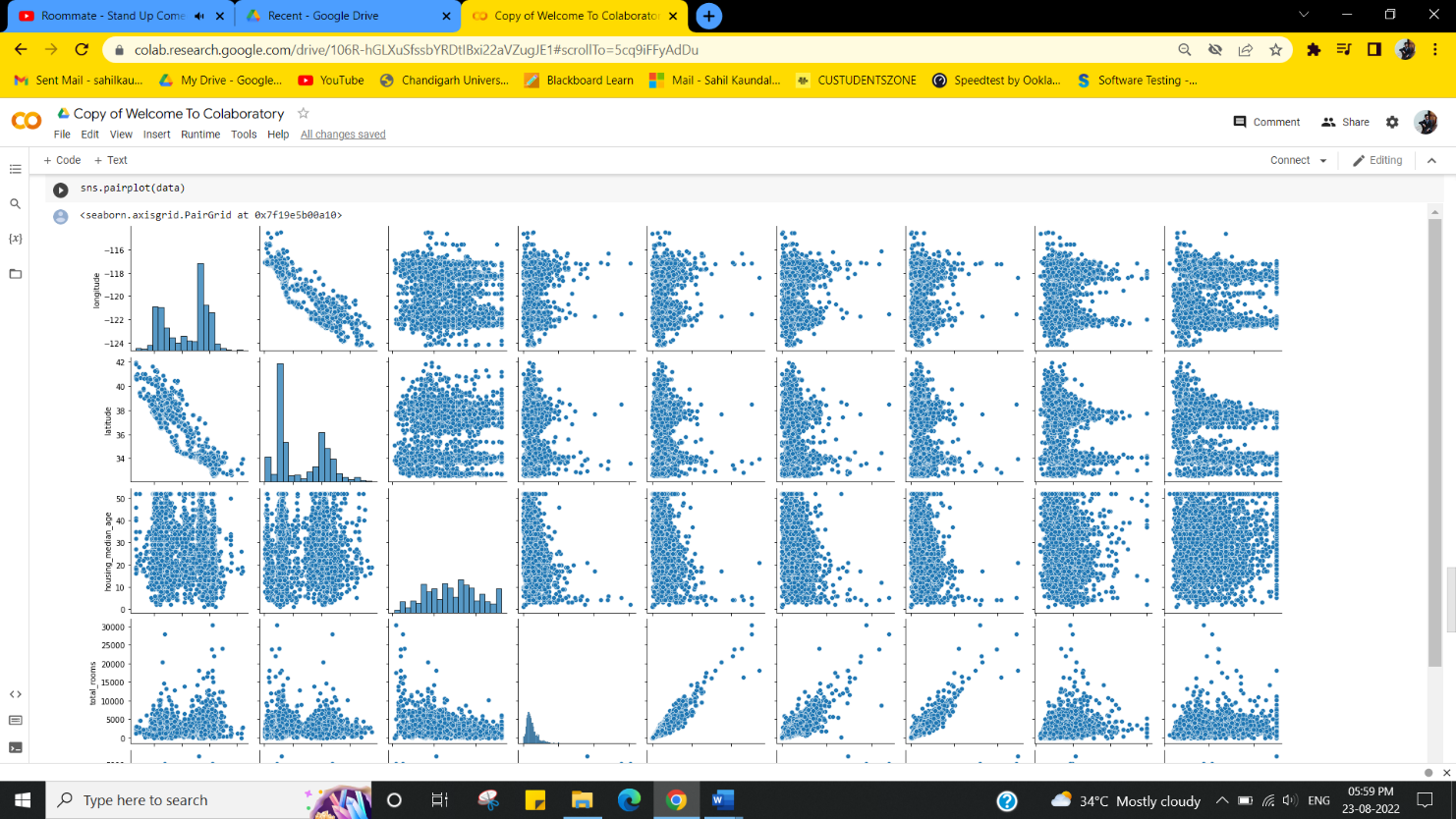
sns.pairplot(data)

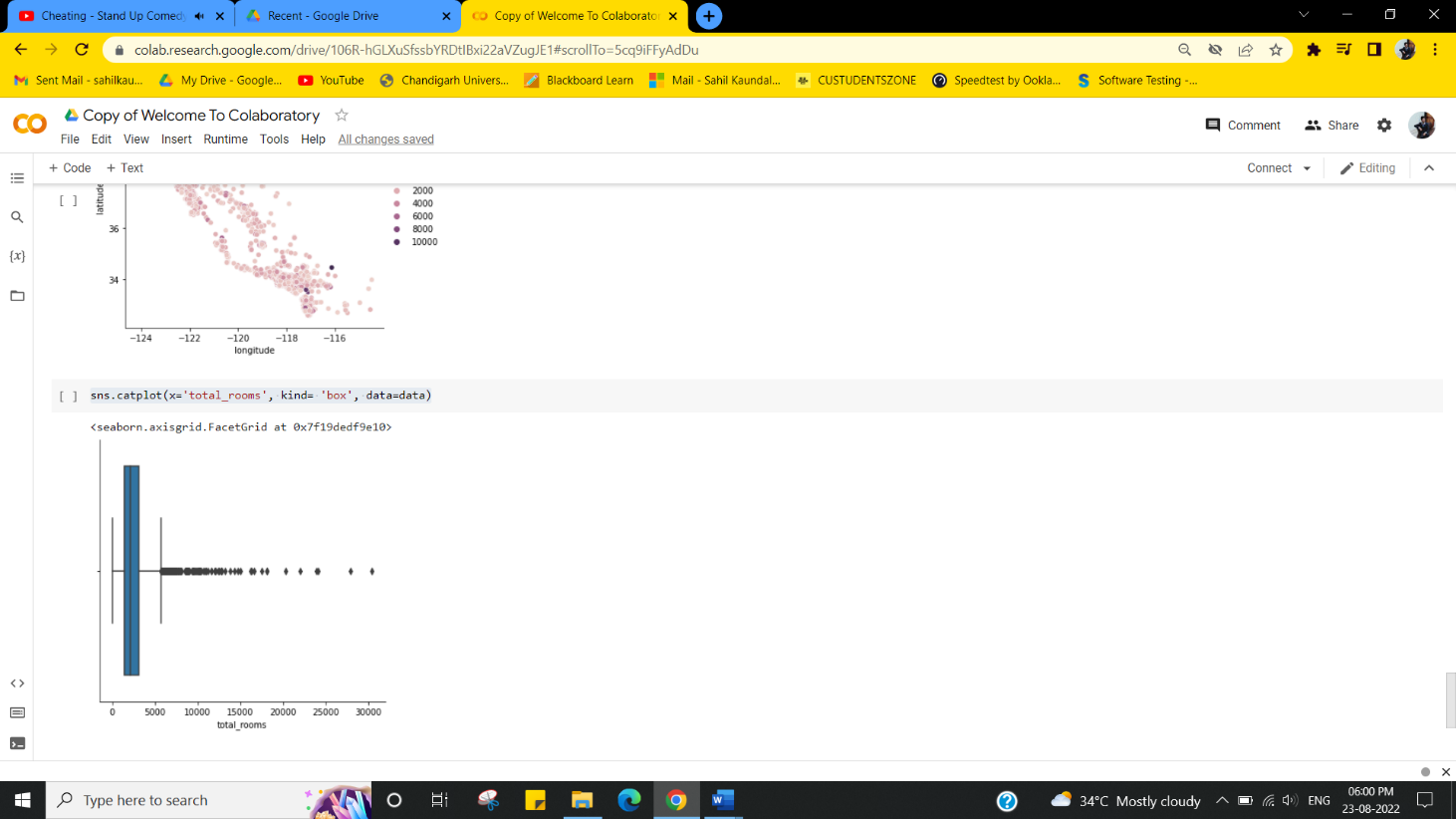
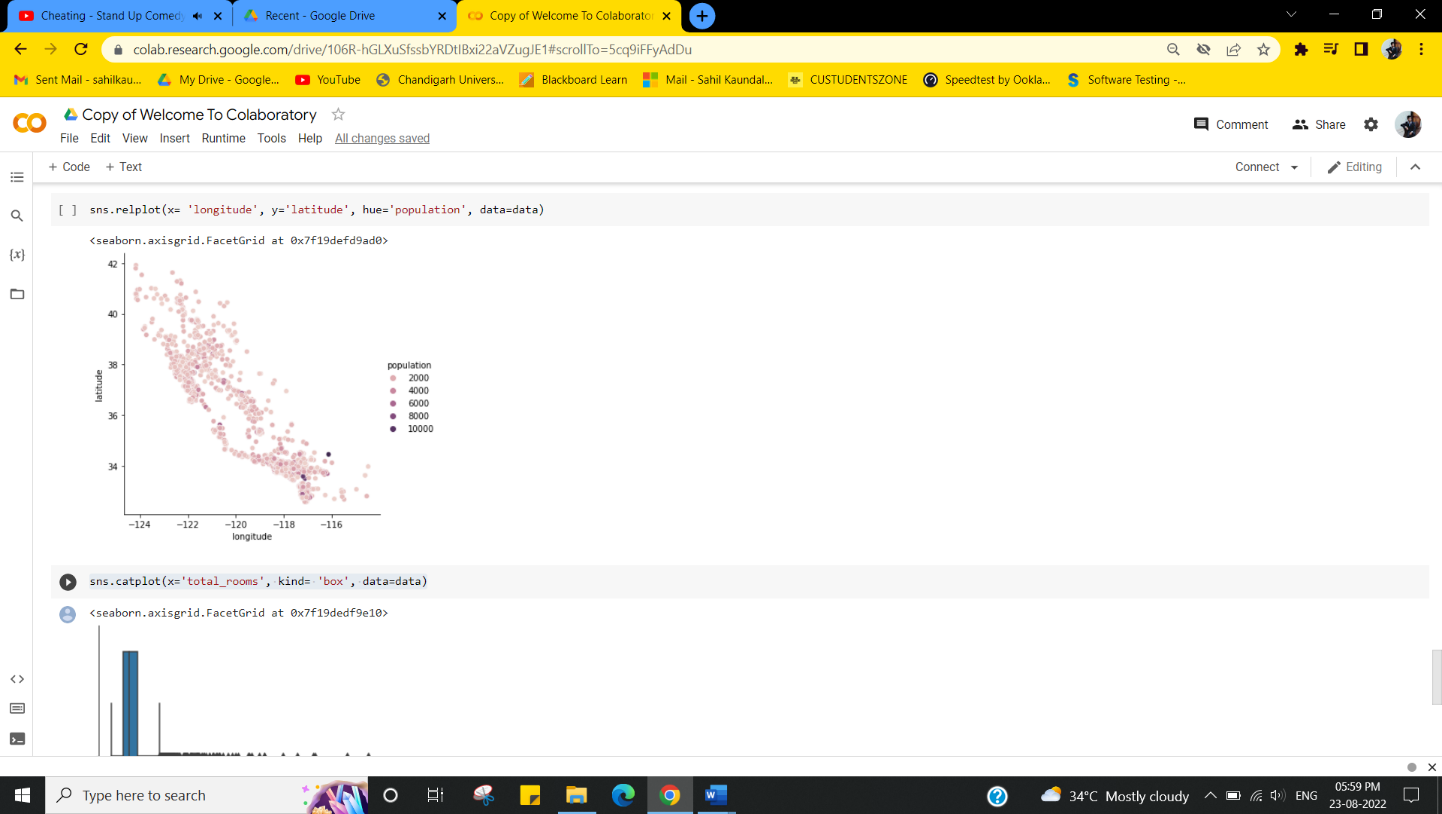
sns.relplot(x= 'longitude', y='latitude', hue='population', data=data)

sns.catplot(x='total\_rooms', kind= 'box', data=data)

1. **Result/Output/Writing Summary:**



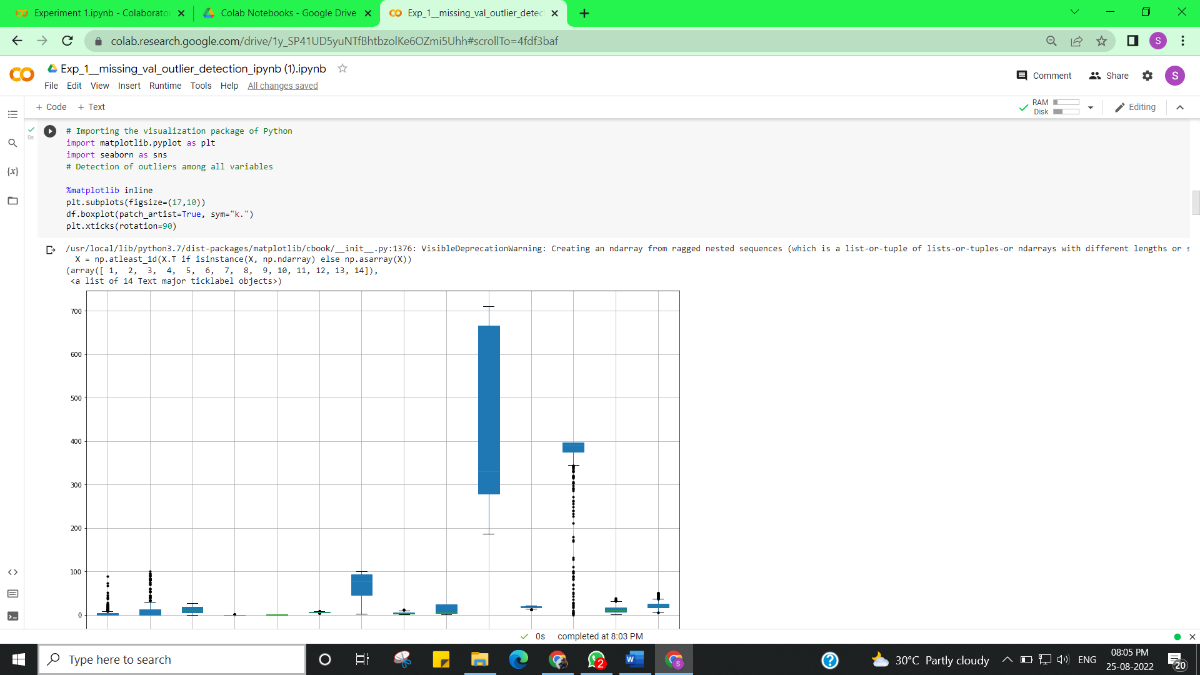
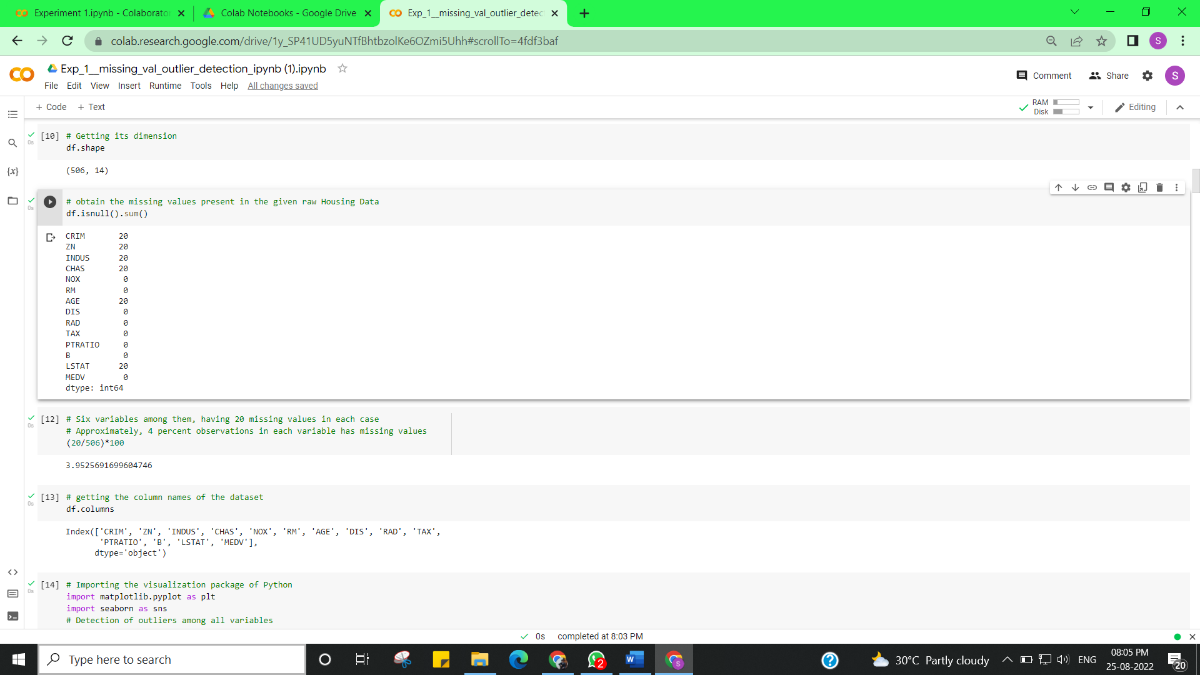
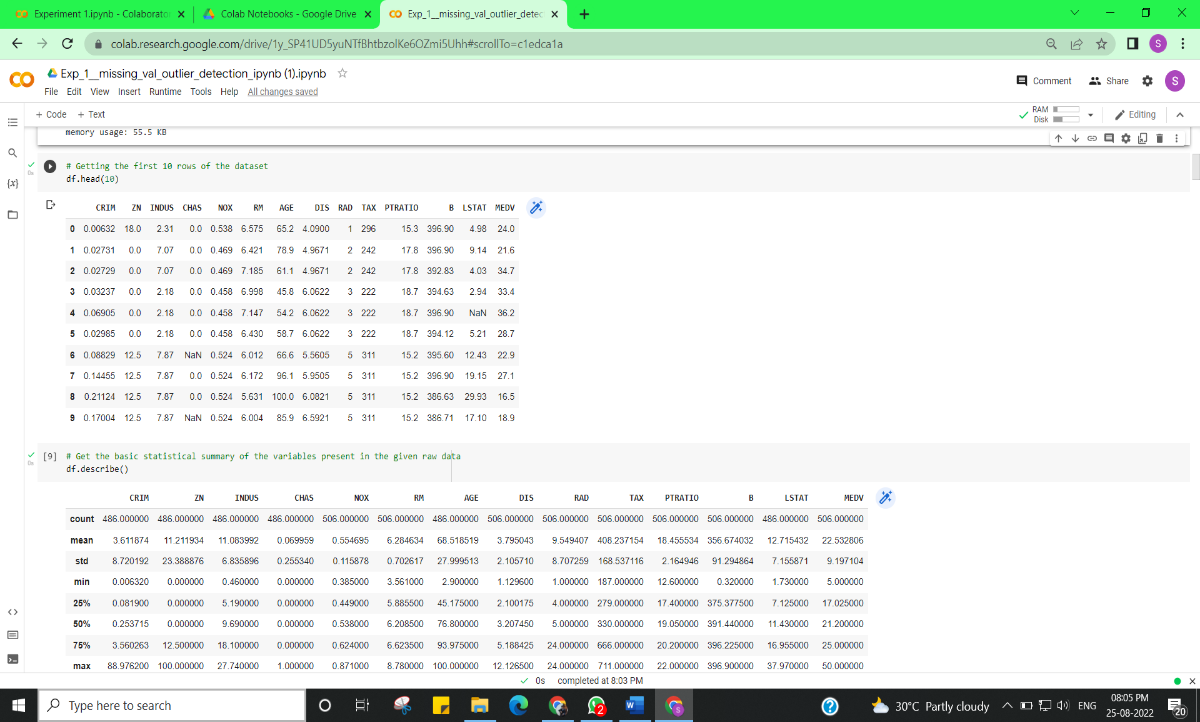




EXPLORATORY DATA ANALYSIS in Python

**Step 1** Treatment of Missing Values and Outliers (Preparation of the dataset for analysis by the treatment of the irregularities.)

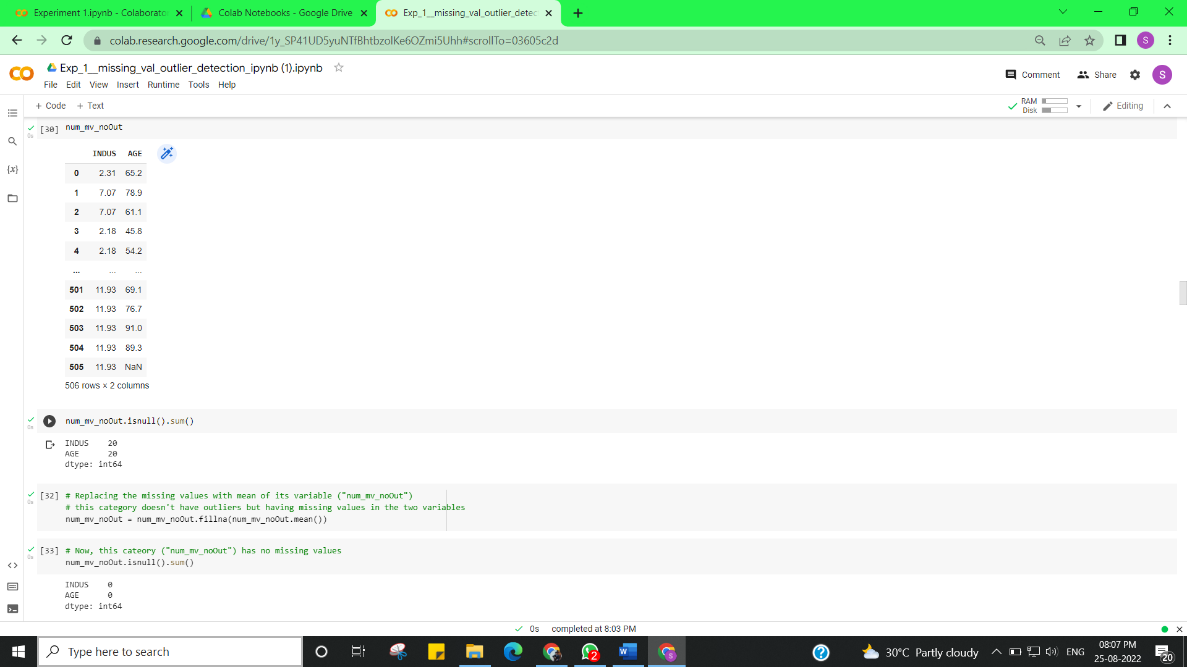
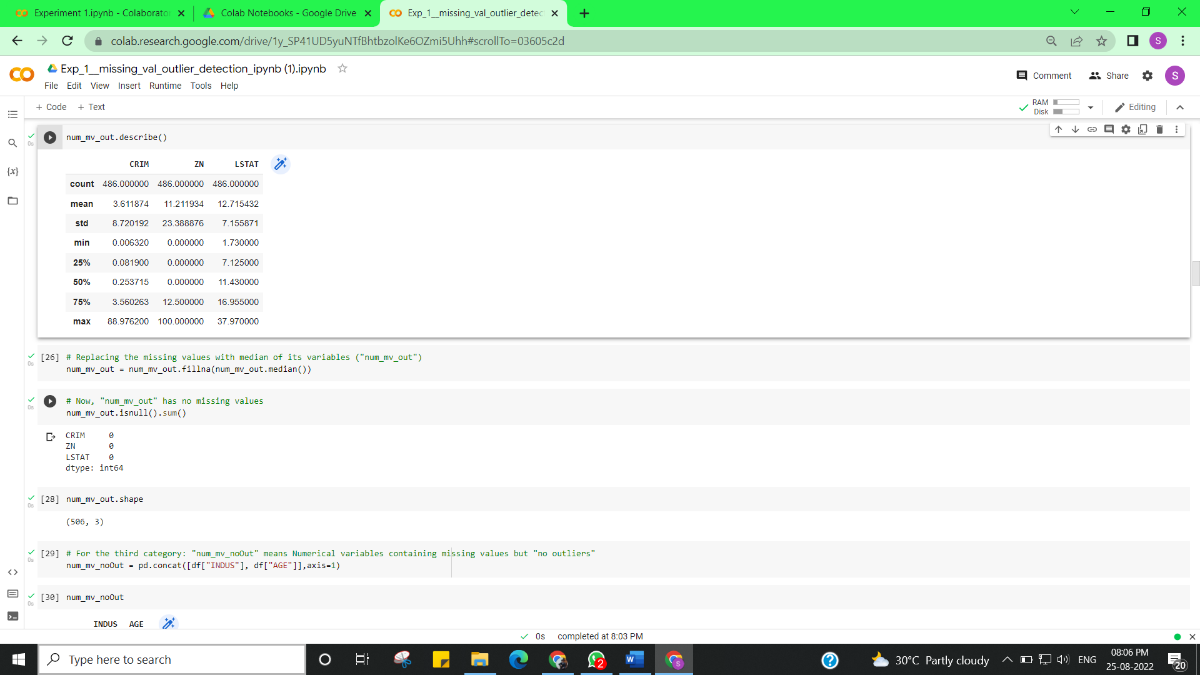
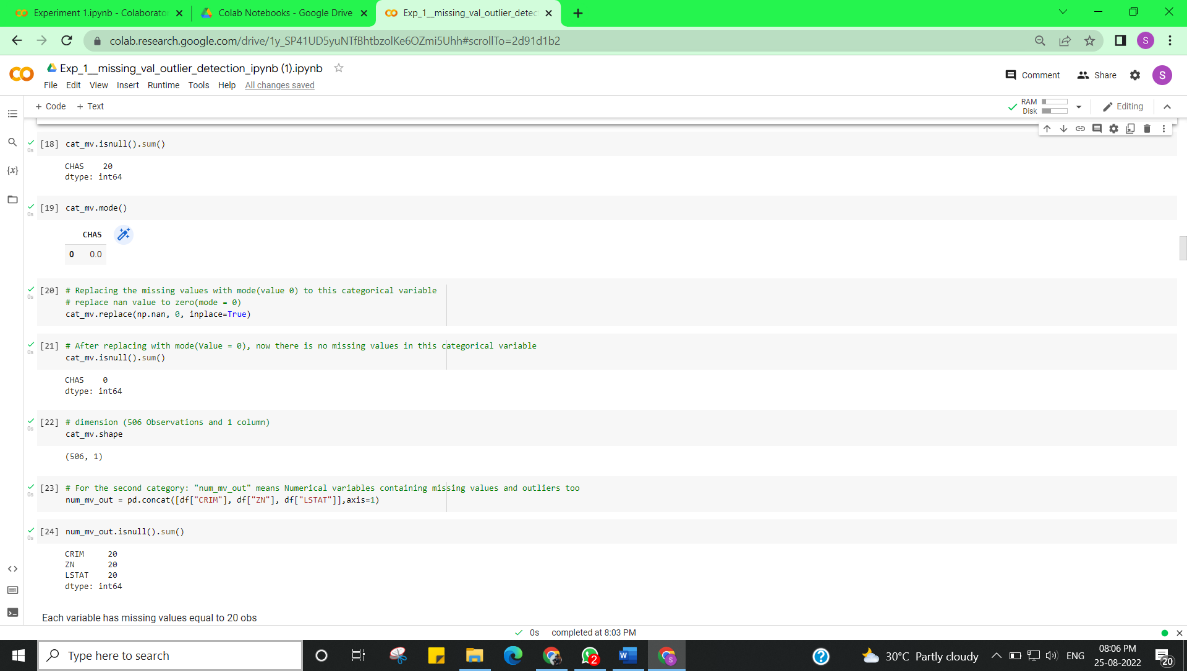
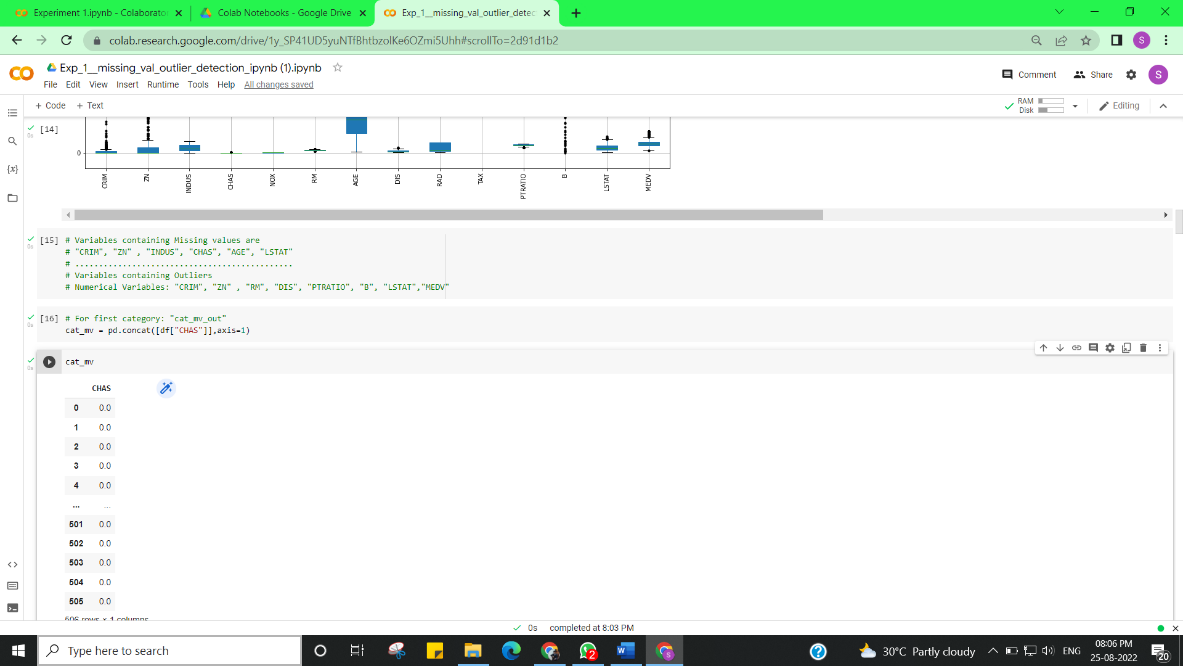
**Step 2** Draw meaningful patterns and insights with the help of data visualization to summarize their main characteristics.

** **

**PHASE 1: TREATMENT OF MISSING VALUES**

**Separate all six variables contain missing values into three groups on the basis of the presence of outliers.**

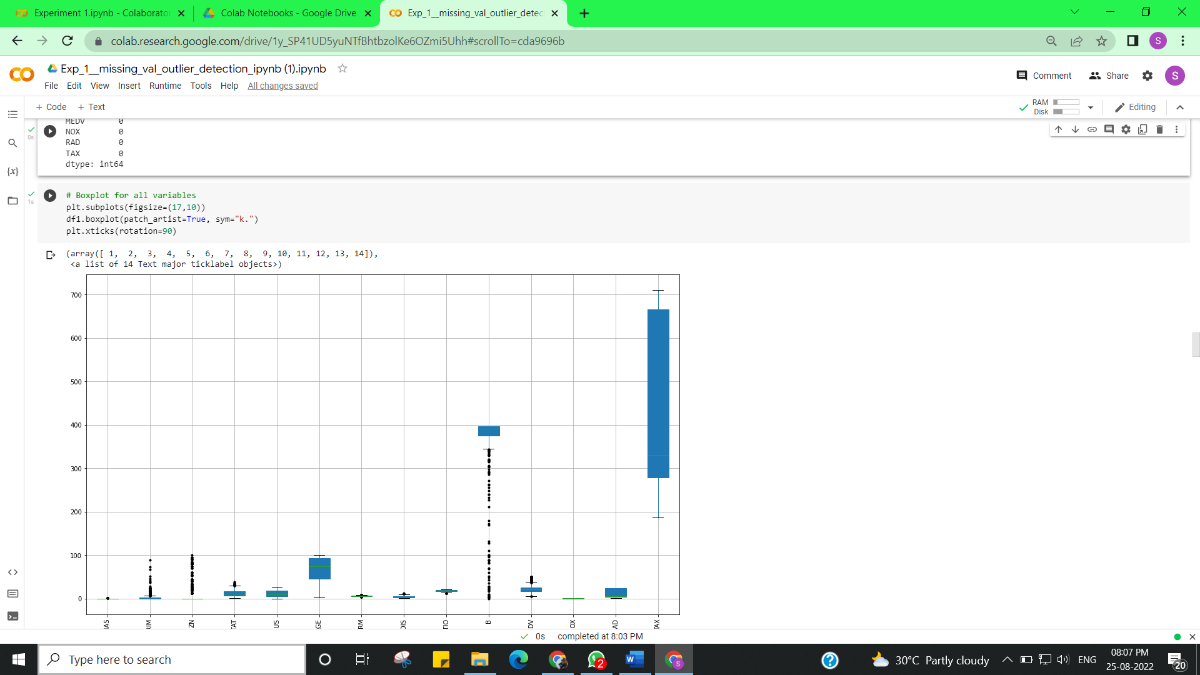
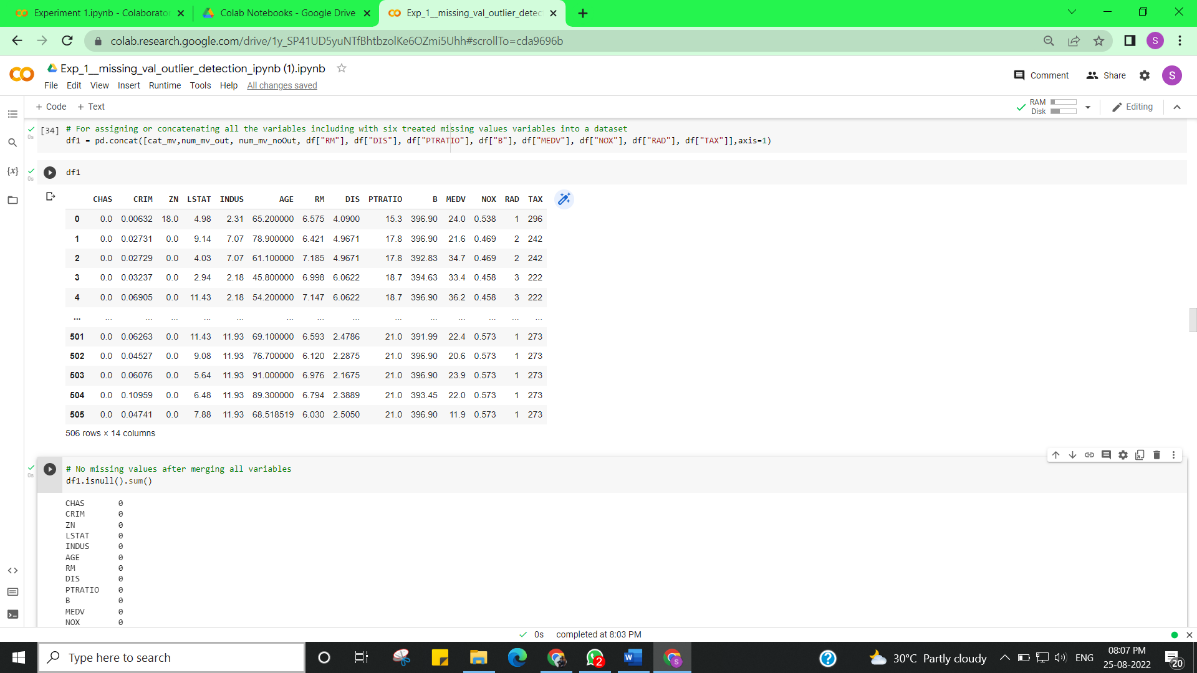
1. cat\_mv = Categorical variable conatining missing values (Missing values will be treated with **mode**)--- "CHAS"
2. num\_mv\_out = Numerical variables containing missing values and outliers too (Missing values will be treated with **median**)--- "CRIM", "ZN" ,"LSTAT"
3. num\_mv\_noOut = Numerical variables containing missing values but "no outliers" (Missing values will be treated with **mean**)--- "INDUS", "AGE"

**Initiate the treatment of Missing Values**

**PHASE 2: TREATMENT OF OUTLIERS**

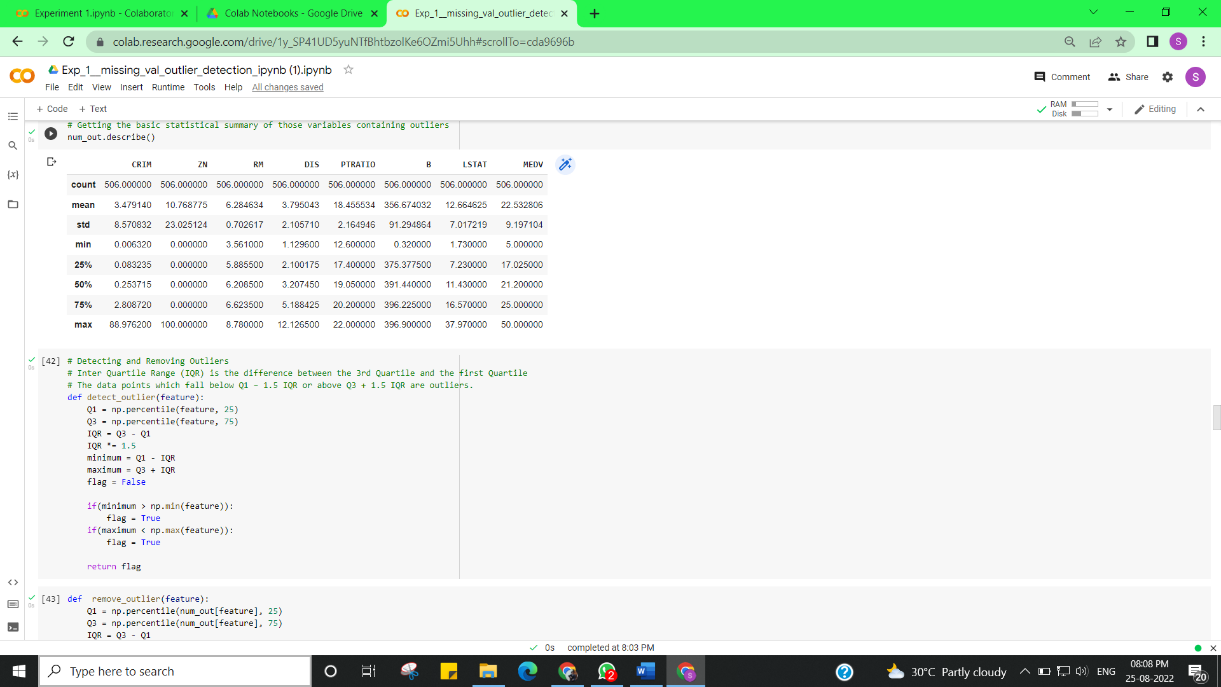
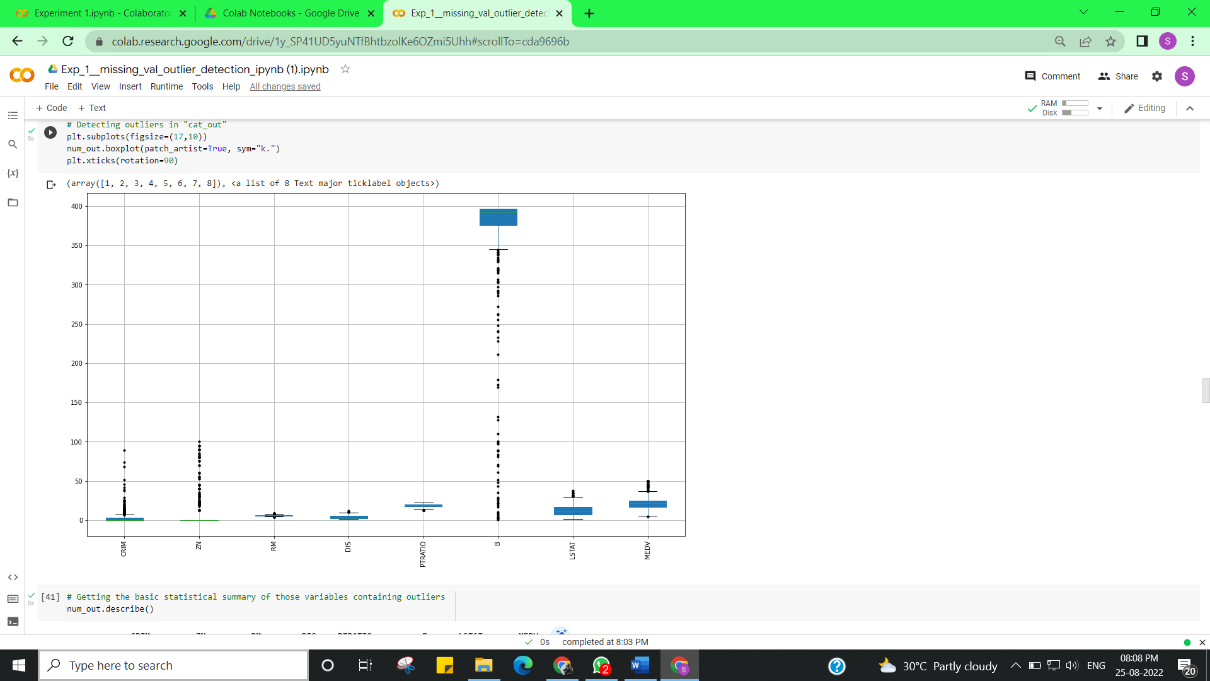
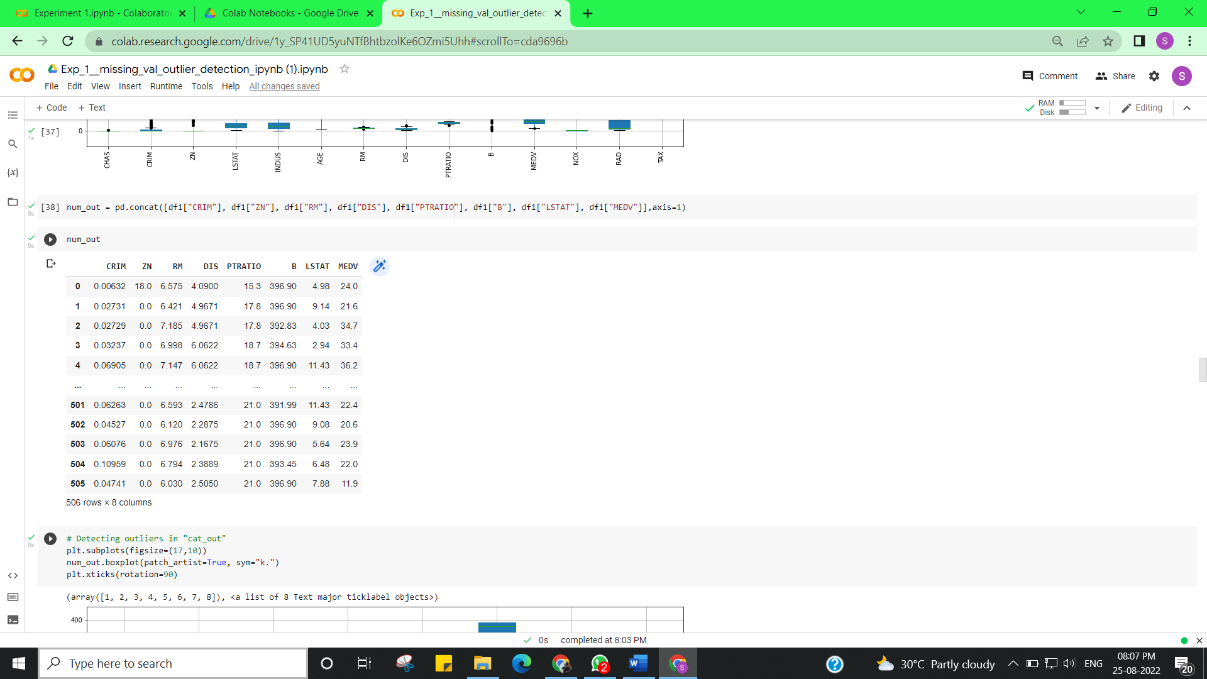
**After treatment of missing values, the dataset will have only outliers problems. So, the next treatment will be for outliers. Now, assign a dataset that will contain all 14 variables including the above three category ("Treated Missing Values" Variables). Finally, split this dataset into three categories. But the thing is, Only the first category will be focussed here because the first category contains outliers. The second and third categories have no outliers.**

1. num\_out = Numerical variables containing outliers (Missing values will be treated with **median**)--- "CRIM", "ZN", "RM", "DIS", "PTRATIO", "B", "LSTAT", "MEDV"
2. num\_noOut = Numerical variables containing "no outliers" (Missing values will be treated with **mean**)--- "INDUS", "NOX", "AGE", "RAD", "TAX"
3. cat\_out = Categorical variable conatining no outliers --- "CHAS"---- In this variable, the observation is either 1 or 0

****

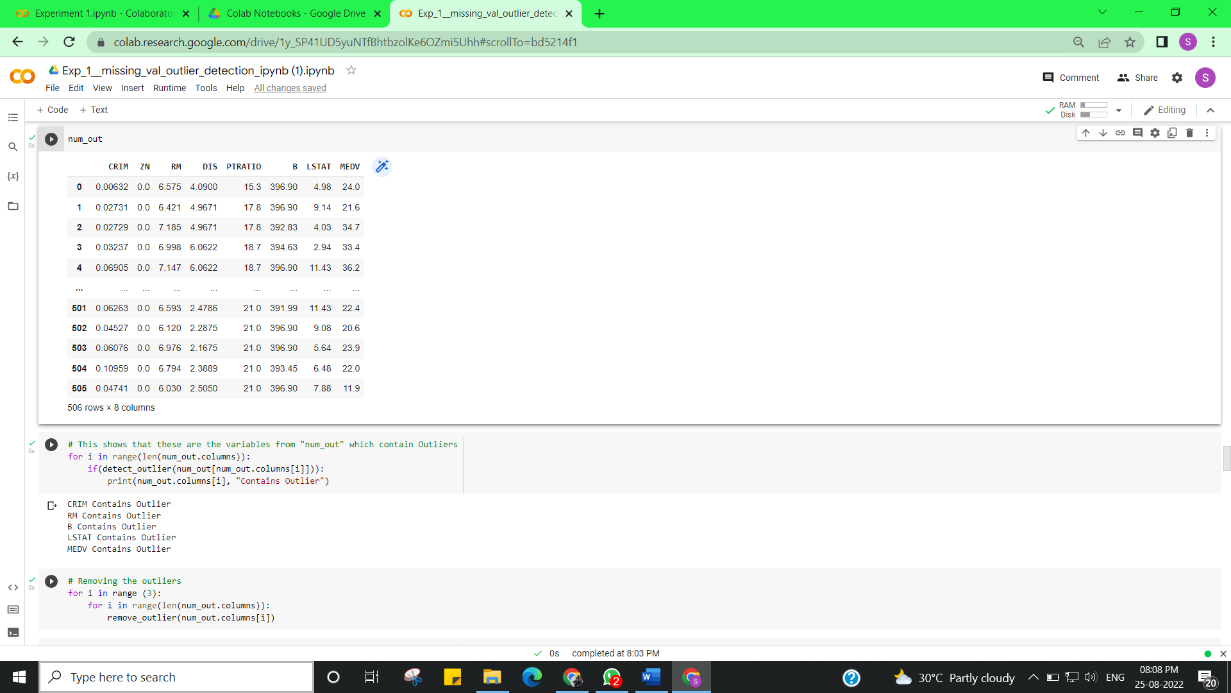
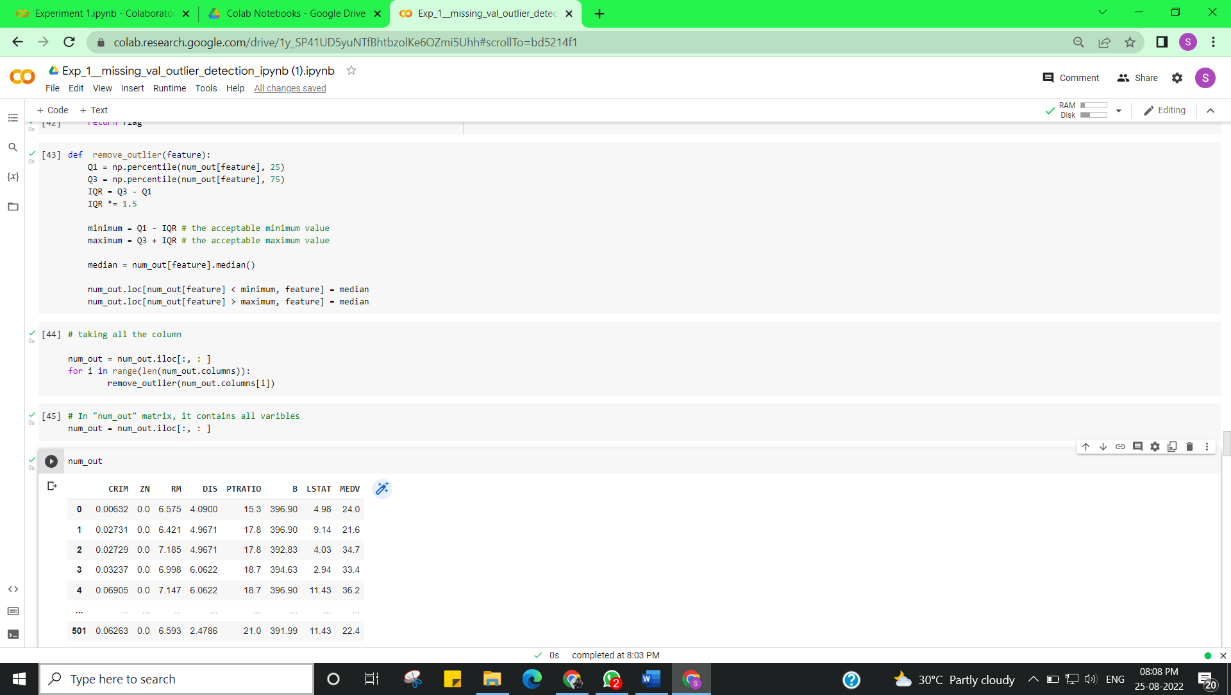
**Now, It's time for treatment of outliers**

1. num\_out = Numerical variables containing outliers (Missing values will be treated with **median**)--- "CRIM", "ZN", "RM", "DIS", "PTRATIO", "B", "LSTAT", "MEDV"

****

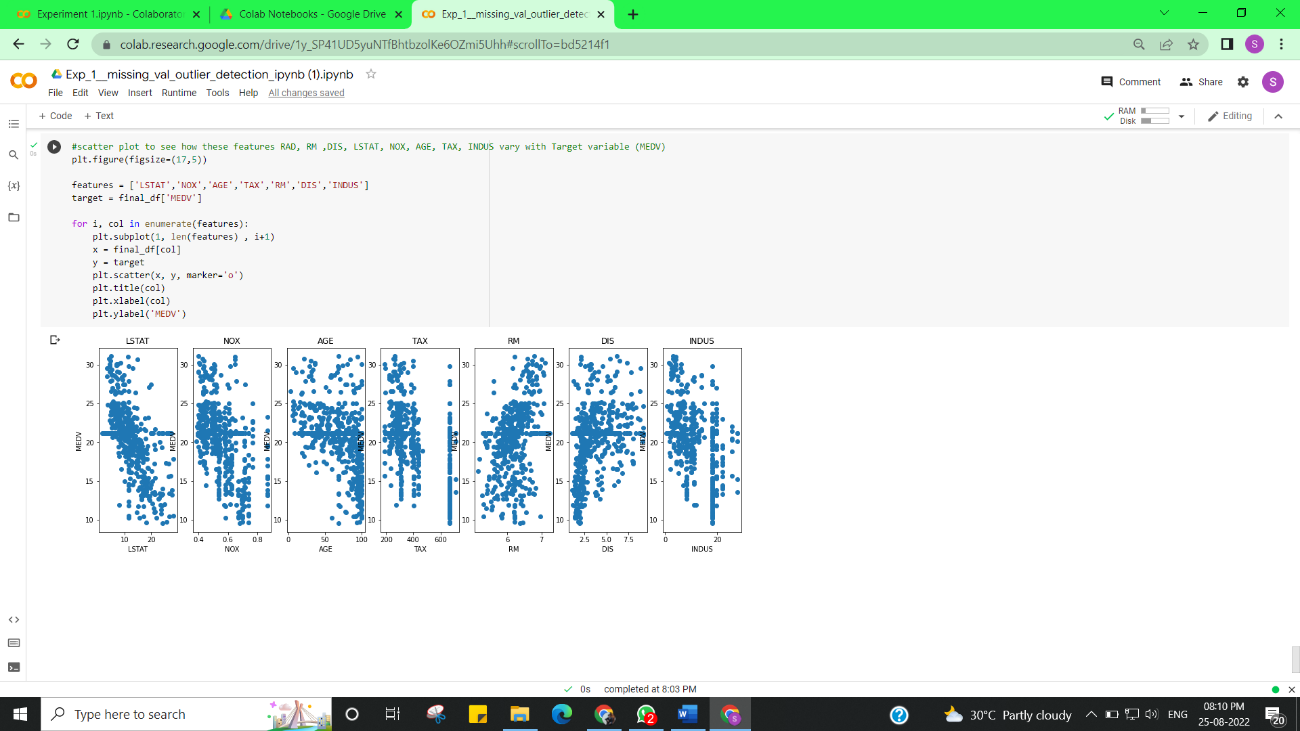
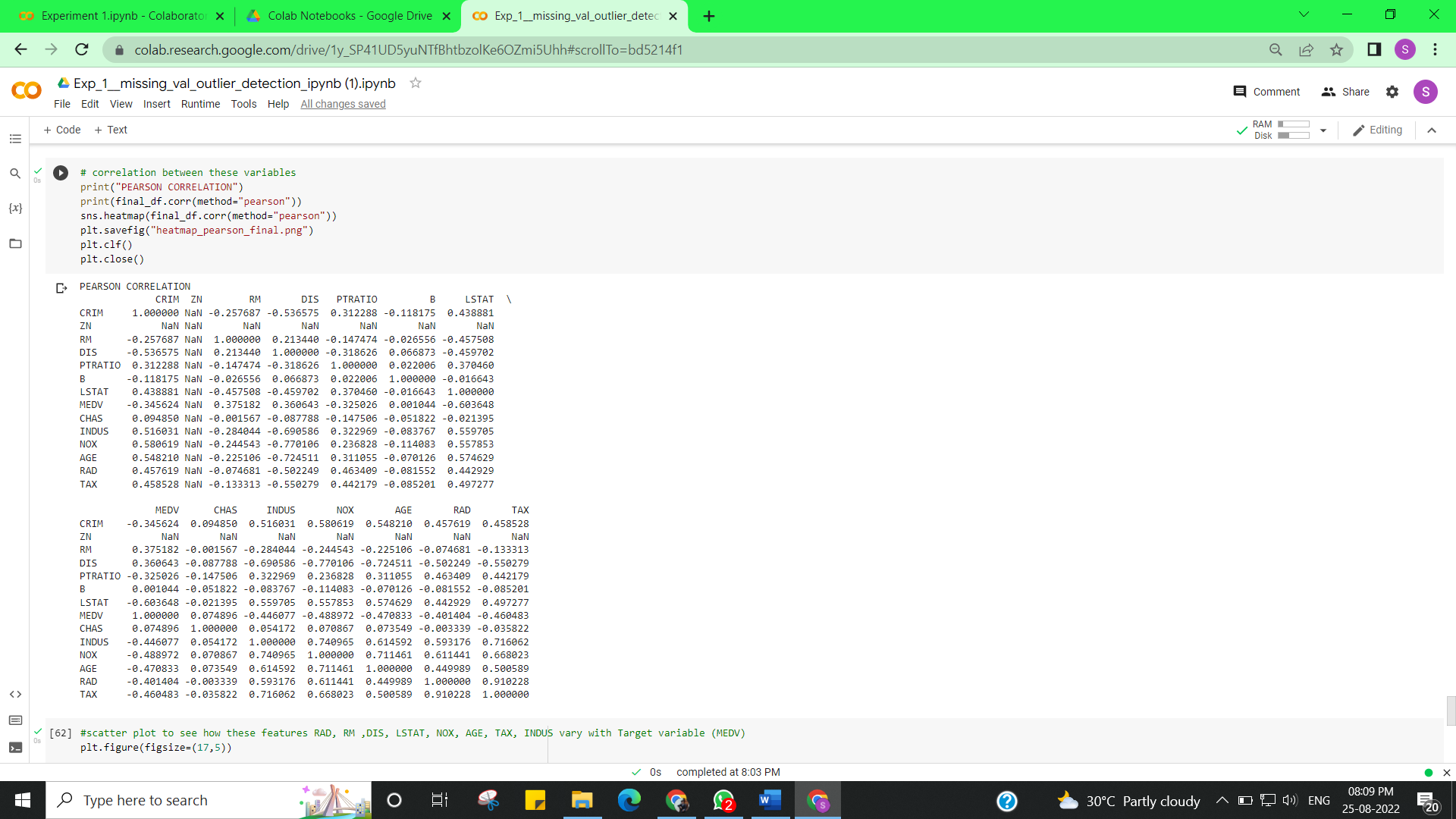
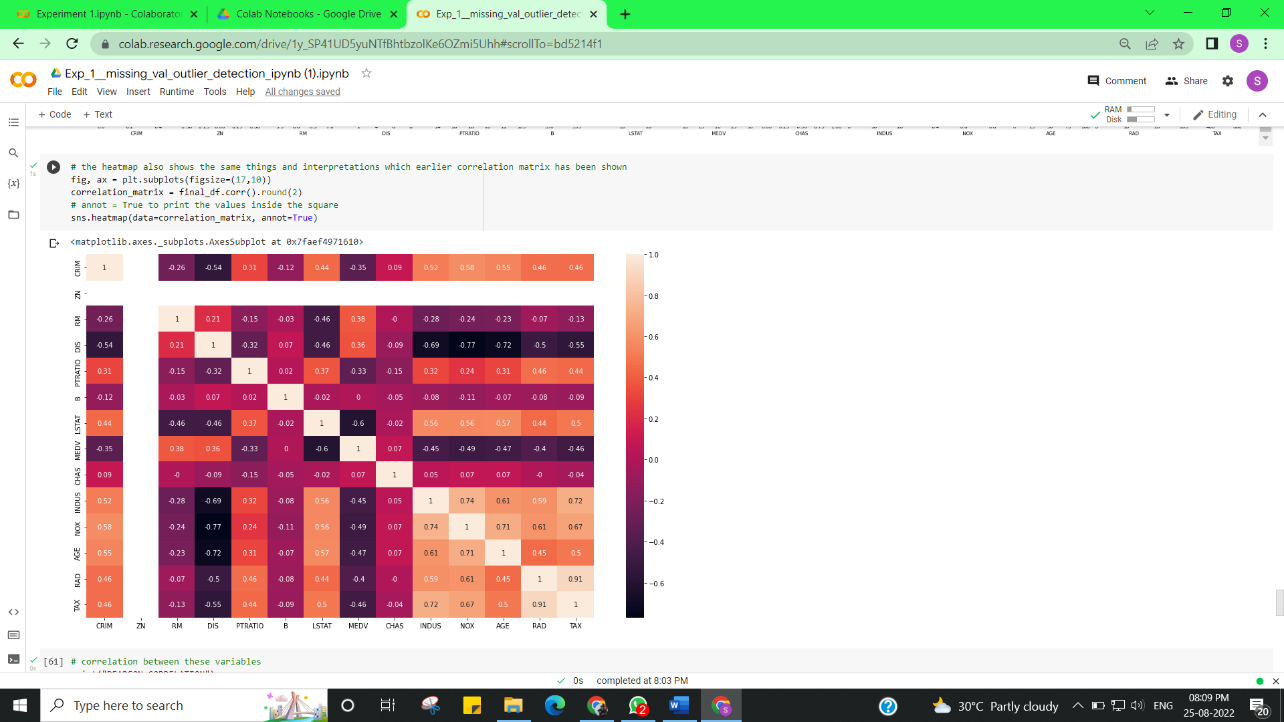
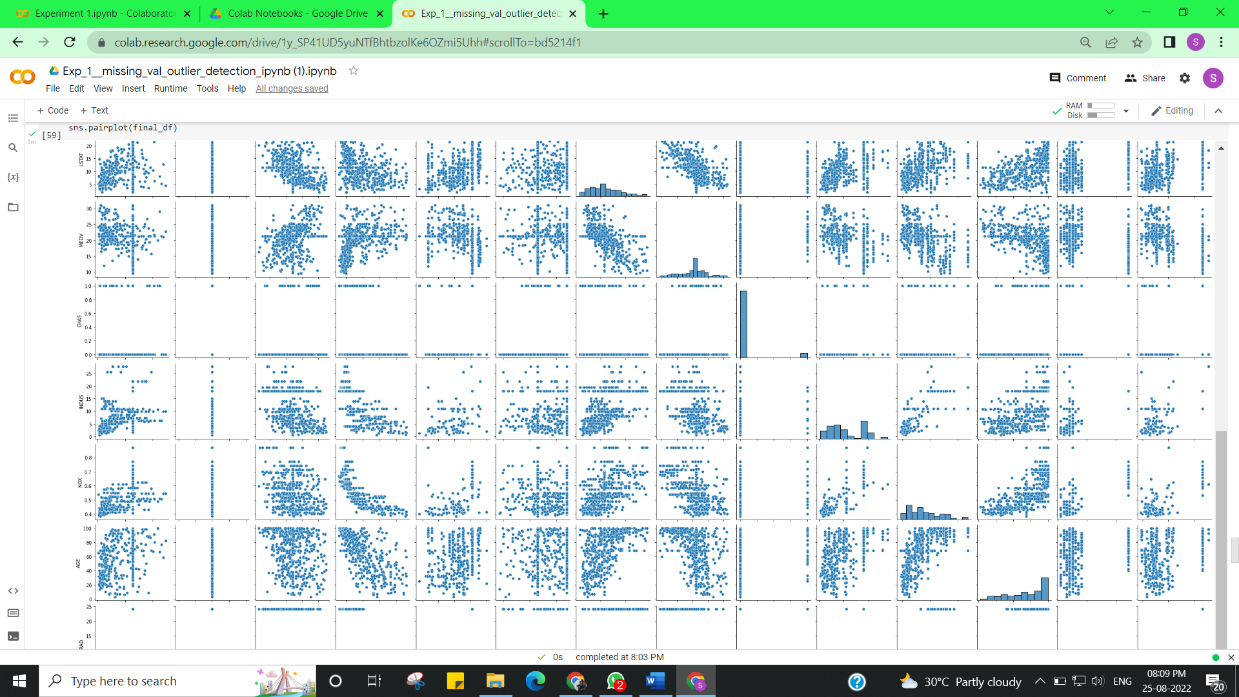
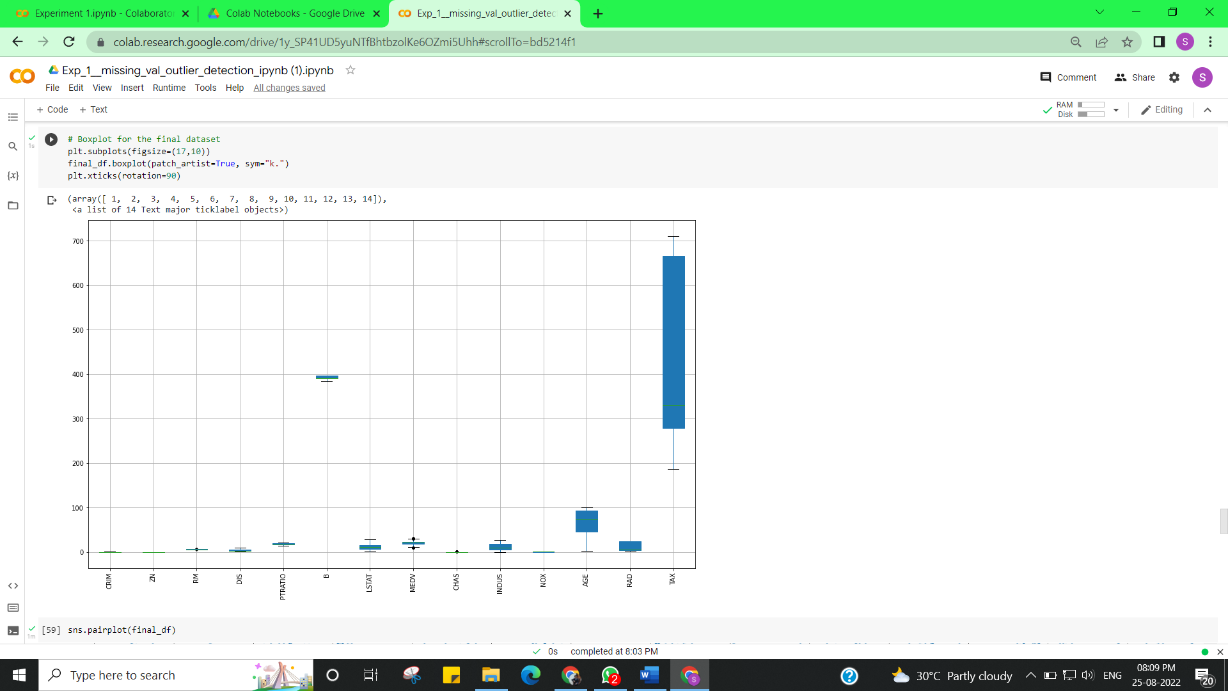
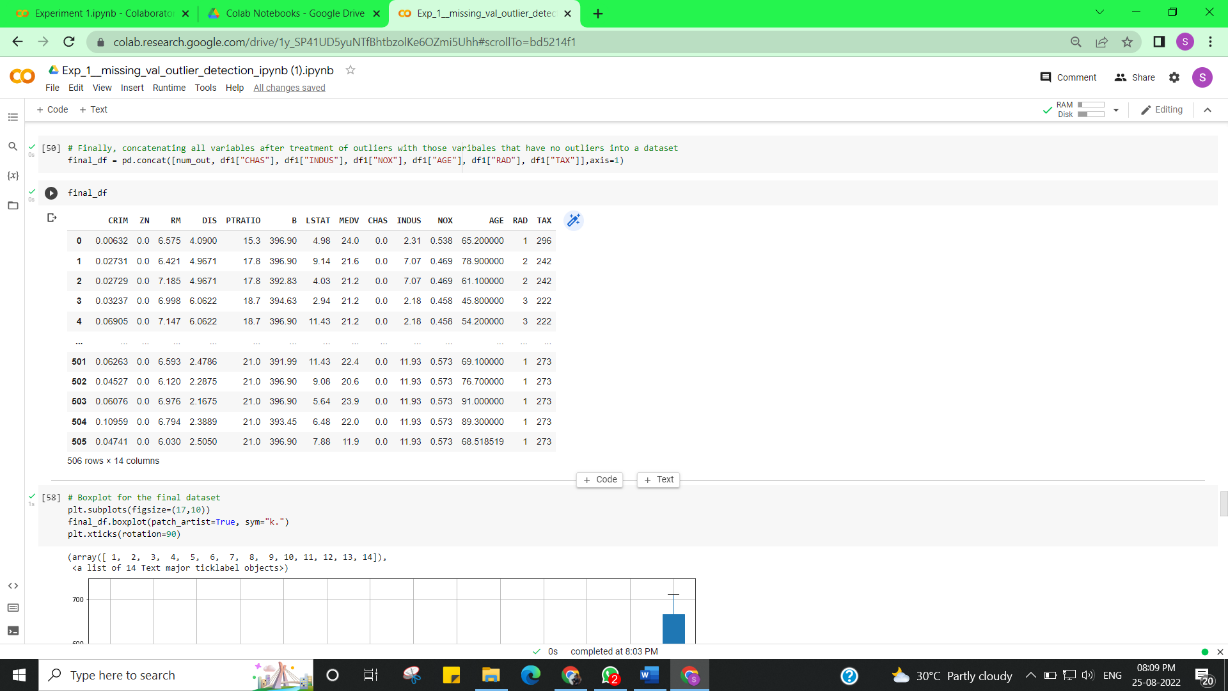
Using tukey method to remove outliers. Whiskers are set at 1.5 times Interquartile Range (IQR). Any value beyond the acceptance range are considered as outliers.

1. **Replacing the outliers with the median value of that feature**
2. **Why replacing with median value?**
3. As the mean value is highly influenced by the outliers, it is advised to replace the outliers with the median value.

****

# After treatment of missing values as well as outliers

# The dataset is now ready for further analysis

****

**Learning outcomes (What I have learnt):**

Implement Exploratory Data Analysis on any data set.

**Evaluation Grid (To be created as per the SOP and Assessment guidelines by the faculty):**

|  |  |  |  |
| --- | --- | --- | --- |
| Sr. No. | Parameters | Marks Obtained | Maximum Marks |
| 1. |  |  |  |
| 2. |  |  |  |
| 3. |  |  |  |
|  |  |  |  |