HUMBER INSTITUTE OF TECHNOLOGY

AND ADVANCED LEARNING

(HUMBER COLLEGE)

**Machine Learning 1 - BIA-5302-0GA**

ASSIGNMENT: (“Unveiling Boston: A Comprehensive Analysis of Housing Trends and Patterns”)

Submitted to: Dr. **Salam Ismaeel**

Submission Date: October 04, 2023

**TABLE OF CONTENTS**

**Introduction …..................................................................................................................** **3**

**Understanding the Data …...............................................................................................** **4**

**Tools and Techniques …...................................................................................................** **5**

**Analysis and Interpretations …........................................................................................** **5**

**Conclusions….....................................................................................................................**

**INTRODUCTION**

In this report, a project that entails analyzing the “BostonHousing” dataset, a collection of data with range of variables and characteristics, will be discussed. The analysis will start with the dataset being carefully organized and cleaned so that it is ready to produce trustworthy results. The foundation for a thorough and perceptive analysis will be laid by methodically reviewing and refining each data point, attribute, and variable.

After the dataset preparation phase, the analysis will delve into the field of data analysis, utilizing the PYTHON programming language as a reliable companion. A variety of in-built functions, proficient in handling both qualitative and quantitative data, will be employed for numerous tasks and analyses. PYTHON's computational capabilities will be utilized to investigate the intricacies of the dataset, unveiling concealed patterns, correlations, and trends. To offer a comprehensive perspective, the analysis will incorporate a variety of visual representations, including diverse types of plots like boxplots, heatmaps, histograms, and others. Each visualization will be thoughtfully selected to illuminate various aspects of the data, facilitating effective communication of the findings.

As the analysis progresses, the goal will be to extract meaningful interpretations from the abundance of statistical analyses and visualizations. The objective is to reveal hidden insights, identify trends, and highlight relationships essential for informed decision-making. The journey is not limited to numerical computations but encompasses an understanding of the context in which these numbers reside, transforming raw data into actionable knowledge.

At the end of this report, brief conclusions will be drawn that would enable good decision - making. The summaries and interpretations will represent the core of the analysis, providing a practical perspective that can guide decision makers. Through the “BostonHousing” dataset, this report will take a journey of data exploration, analysis and enlightenment and highlight the true potential of the data.

**UNDERSTANDING THE DATA**

For this project, a dataset named “BostonHousing” is being taken from *http://lib.stat.cmu.edu/datasets/boston* (Harrison Jr., David, & Daniel L. Rubinfeld[1978])

The data file is in the Excel format with 11 columns representing variables and attributes. Following is the description of all the variables / attributes:

|  |  |
| --- | --- |
| CRIM | per capita crime rate by town |
| ZN | proportion of residential land zoned for lots over 25,000 sq.ft. |
| INDUS | proportion of non-retail business acres per town |
| CHAS | Charles River dummy variable (= 1 if tract bounds river; 0 otherwise) |
| NOX | nitric oxides concentration (parts per 10 million) |
| RM | average number of rooms per dwelling |
| AGE | proportion of owner-occupied units built prior to 1940 |
| DIS | weighted distances to five Boston employment center's |
| RAD | index of accessibility to radial highways |
| TAX | full-value property-tax rate per $10,000 |
| PTRATIO | pupil-teacher ratio by town |

Although the data in the spreadsheet might appear to be simple, it is in a rough and chaotic form with problems including missing numbers and data entry mistakes. Therefore, it is essential to analyze this data before conducting a complete study, and then a variety of tools and procedures will be used to get useful findings.

**TOOLS AND TECHNIQUES**

In our data analysis project, we employed a range of tools and techniques to thoroughly explore and analyze the Boston Housing dataset.

* For data preprocessing tasks, we harnessed the power of **Python**.
* Utilized the `**pandas**` library for data manipulation, which allowed us to seamlessly handle missing data, descriptive statistics, including mean, median, min, max, and standard deviation.
* The identification of outliers in the PTRATIO predictor was facilitated by Python's data visualization libraries.
* `**matplotlib**`, enabled us to create scatter plots and box plots for outlier detection. For visual exploration, histograms for each quantitative variable were generated.
* Additionally, to gain insights into variable relationships, we crafted a side-by-side box plot, comparing two variables visually with the aid of `**seaborn**`, also to compute the correlation matrix and visualized it as a heatmap.

These tools and techniques allowed us to conduct a comprehensive analysis of the dataset, enabling us to draw meaningful conclusions and insights from our findings.

**ANALYSIS AND INTERPRETATIONS**

**Missing Data and Outlines,**

Managing missing data in a dataset is a crucial step during the data preprocessing phase. It is essential not only to account for typical missing values but also to consider outliers in the data. Here are several techniques for addressing missing data and outliers:

**Deleting Rows or Columns with Missing Data and Outliers:**

This straightforward approach involves removing rows or columns containing missing data or outliers. However, this method can result in a loss of valuable information if a massive portion of the data is deleted. To solve this, we can replace the missing data with a NULL or NaN value.

To replace missing data with NaN in pandas DataFrame, the replace () function can be utilized. This function allows us to replace a specific value with NaN, which can effectively handle missing data.

**Imputation with Outlier Handling:**

Imputation replaces missing values with estimates derived from the available data, while also addressing outliers.

To identify outliers in the PTRATIO predictor, we employ the Interquartile Range (IQR) method, which utilizes the lower\_bound and upper\_bound values. Here is how it works:

**Calculate the IQR (Interquartile Range):** We begin by computing the IQR, a measure of the spread of data. It is determined by finding the difference between the third quartile (Q3) and the first quartile (Q1).

**Establish Lower and Upper Bounds:** With the IQR in hand, we set lower and upper bounds for outlier detection. The lower bound is set as Q1 minus 1.5 times the IQR, while the upper bound is set as Q3 plus 1.5 times the IQR.

**Identify Outliers:** Any data point falling below the lower bound or exceeding the upper bound is considered a potential outlier. These values lie significantly outside the typical range of the data distribution and warrant special attention.

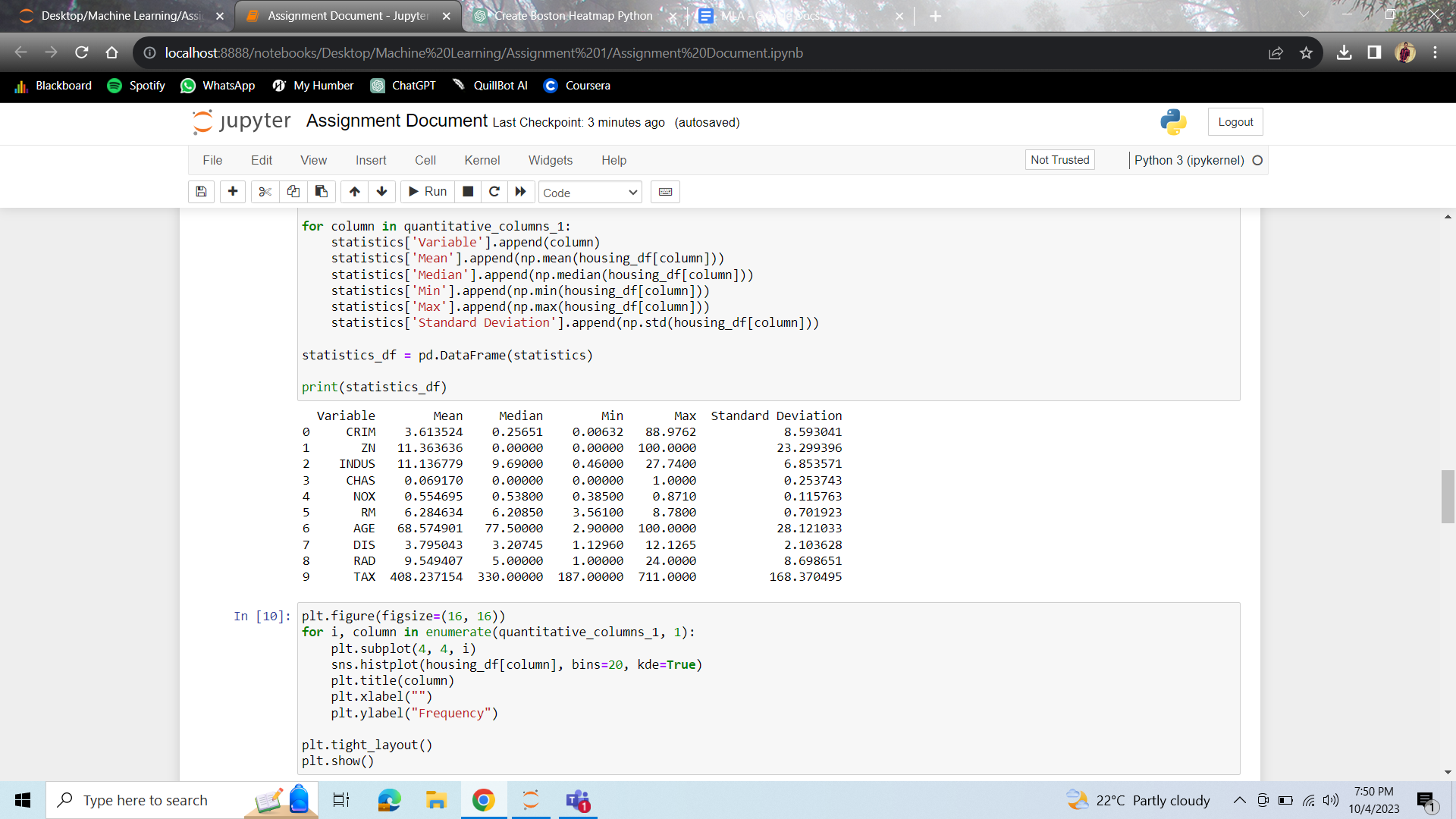
By utilizing these lower and upper bounds, the IQR method enables us to systematically pinpoint potential outliers in the PTRATIO predictor, enhancing our ability to identify data points that may have a notable impact on our analysis.

**Omission with Outlier:**

When dealing with missing data, one approach is to exclude observations or variables that contain missing values. In pandas, we can achieve this using the dropna() function. These panda functions provide versatile tools for handling missing data in the dataset, enabling tailoring the approach based on a specific data preprocessing requirement.

**Statistical summaries,**

Statistical summaries are essential in data analysis because they provide a condensed depiction of important characteristics such as central tendencies, variability, and distributions. They support better decision-making, hypothesis testing, and a deeper comprehension of the underlying patterns in the data by assisting researchers in identifying patterns, outliers, and trends. Here, PYTHON has an in-built function for each of these measures he is following is the

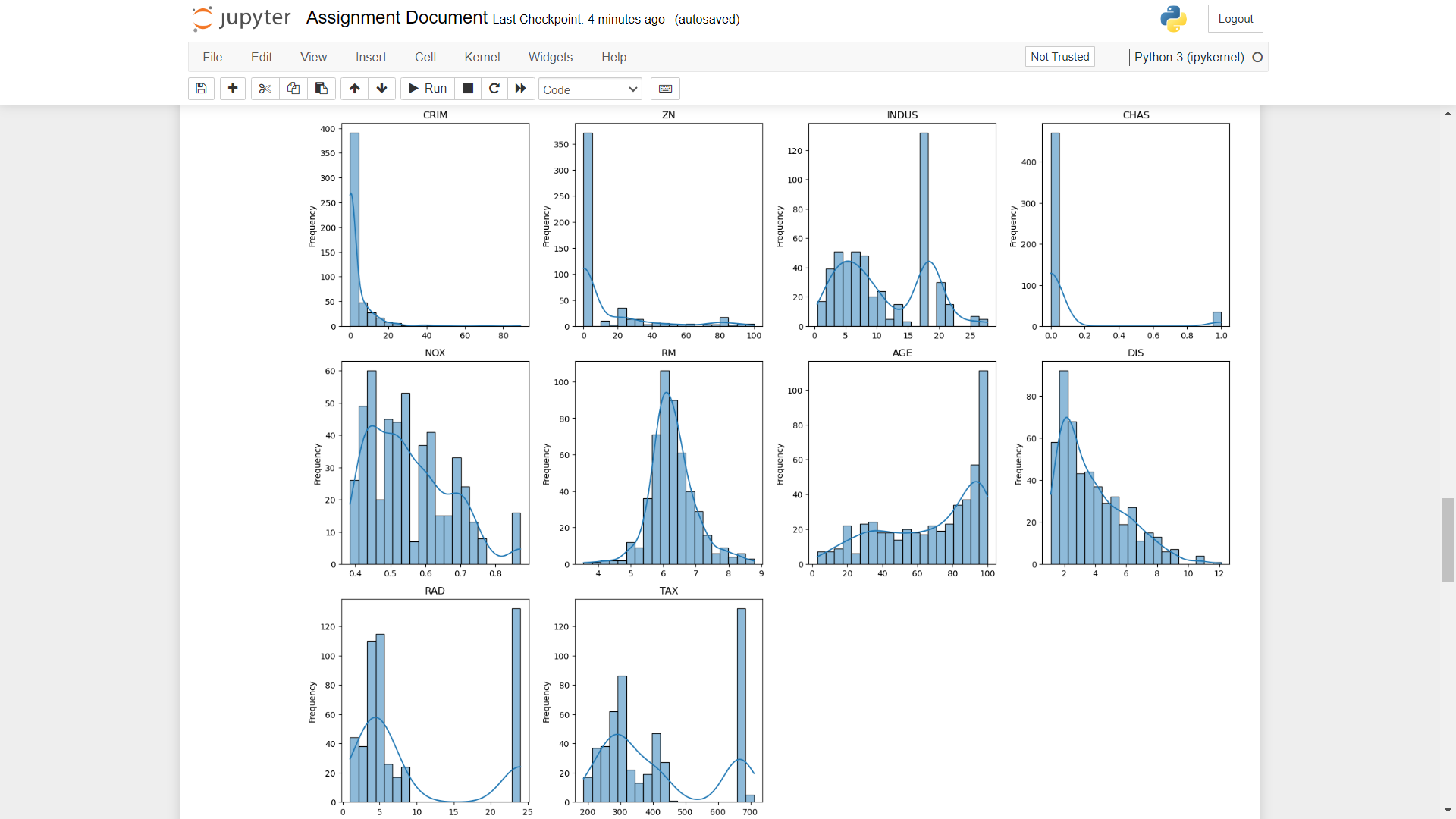
**Output:**  
 

**Statistical summaries interpretation,**

The table provides key statistics for various housing-related variables. Notably, the crime rate (CRIM) has a wide range, with a mean of 3.61, indicating varying levels of safety in the area. The proportion of residential land (ZN) is typically low, with a mean of 11.36, suggesting limited residential zones.

**Visual Representation – Histogram,**

An alternative method for statistically examining the data involves using visual representations to elucidate the distribution, presence of outliers, and variability across various variables. Here, histograms have been generated for each variable.

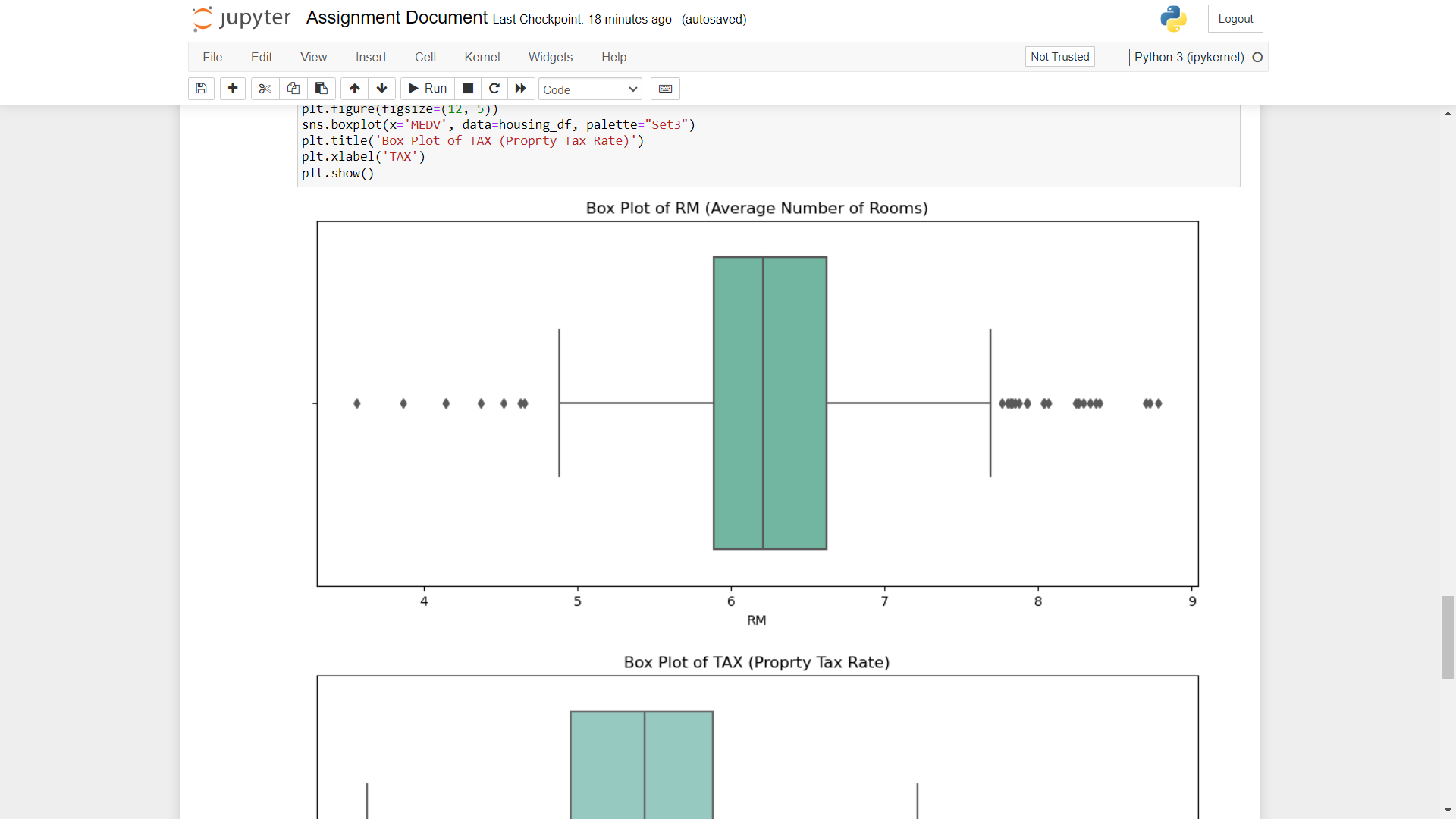


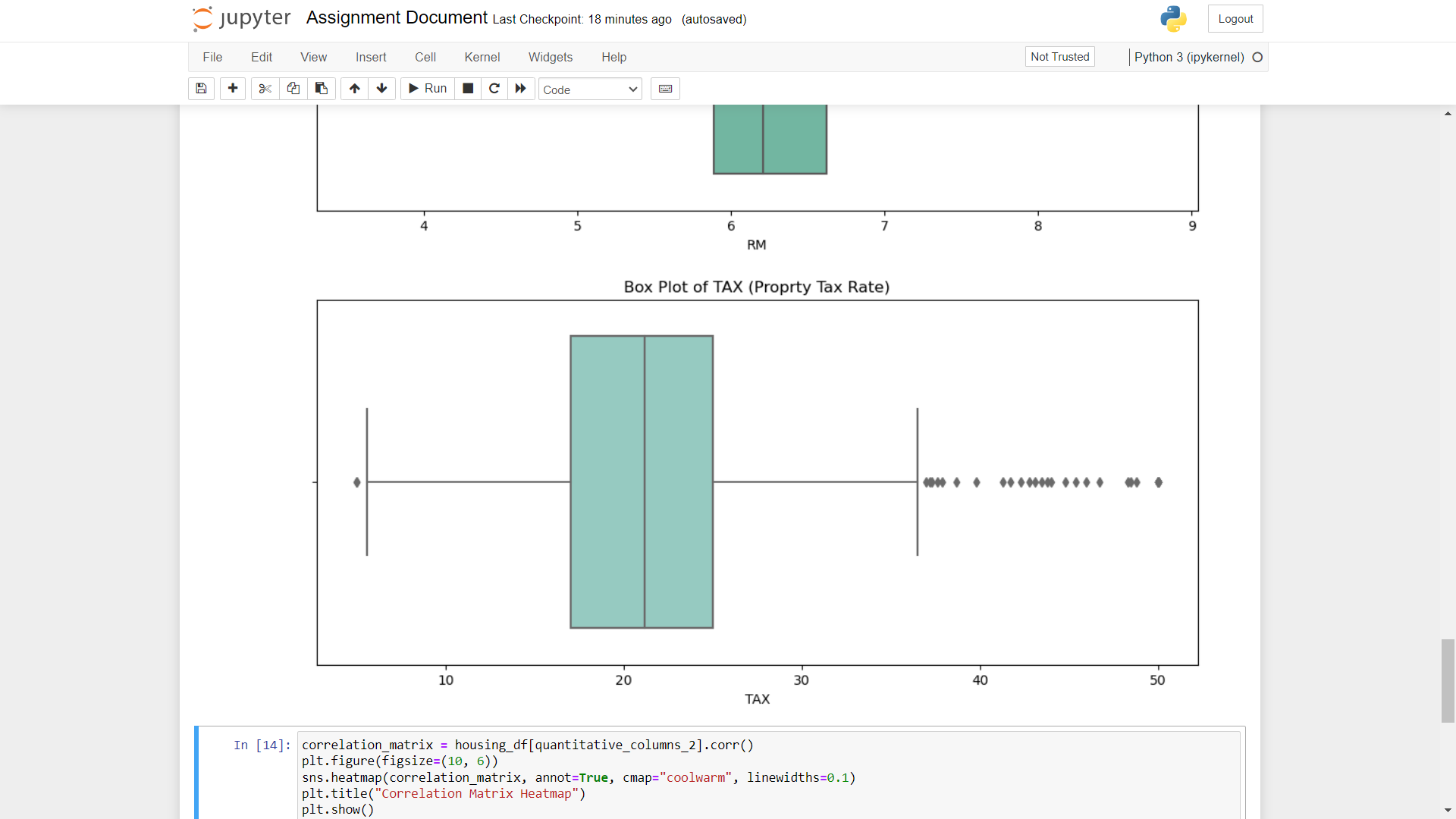
**Histogram interpretation,**

Here we can observe that all variables except CRIM and RM show greater variability as the distribution of histogram for these two variables are tall and narrow. Variables CRIM and DIS strongly depict a left skewness however, AGE is right skewed. In RAD and TAX there are extreme outliers on the right side of the graph.

**Visual Representation – Box Plot,**

A Box plot can quickly identify differences and similarities between any two variables by examining the relative positions, spreads and shape of the boxes and whiskers. It demonstrates key characteristics such as medians, quartiles, and outliers. Here, we have considered variables RM and TAX variables.



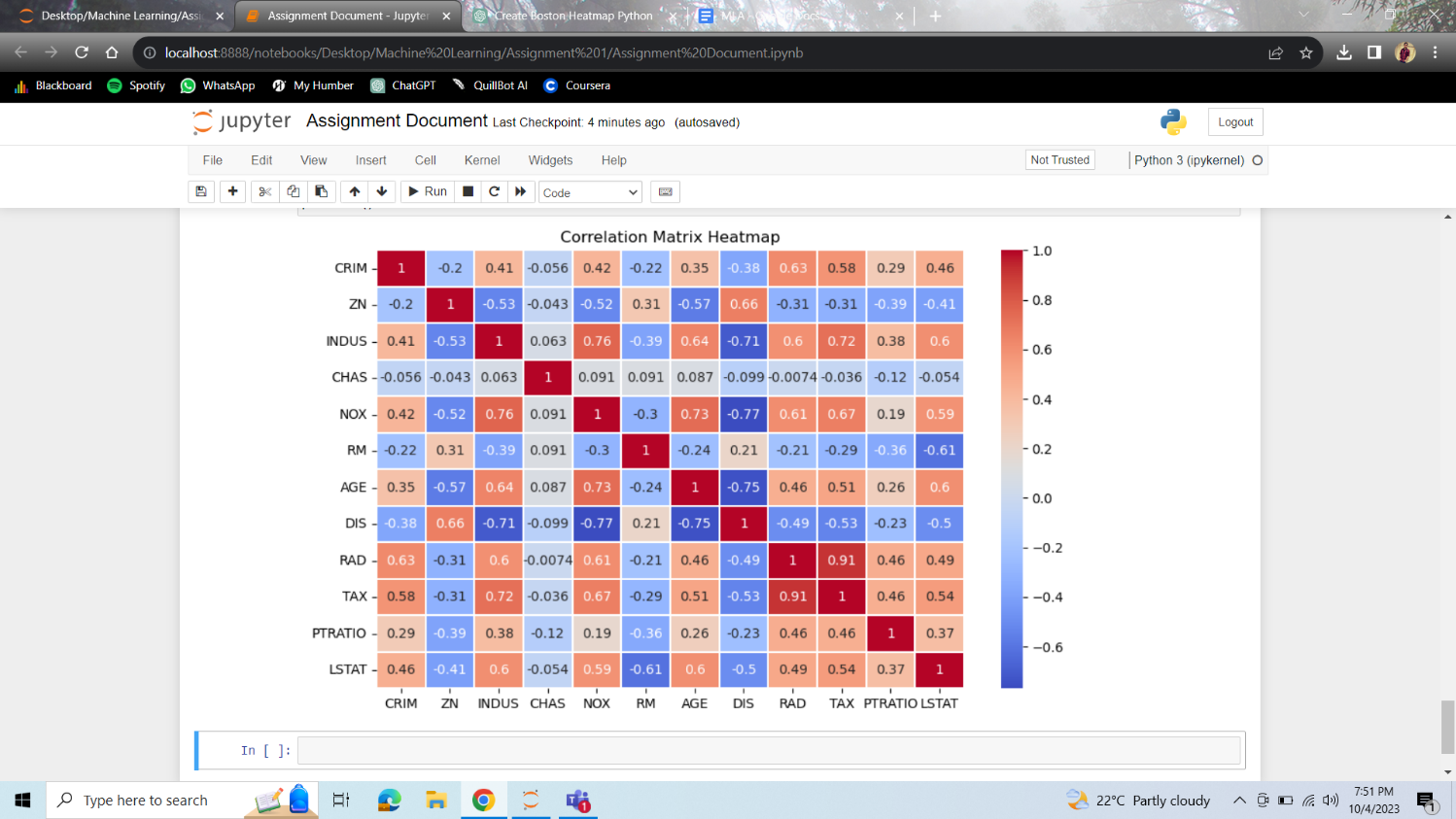


**Box Plot Interpretation,**

In the boxplot comparison between "RM" (average number of rooms per dwelling) and "TAX" (property tax rate) in the Boston Housing dataset, it is evident that there is no apparent relationship between the number of rooms and property tax rates. The boxplots show varying tax rates across properties with different room counts, indicating a lack of significant correlation or pattern in this regard.

**Visual Representation – Heatmap,**

Correlation analysis is significant because it quantifies the strength and direction of the relationship between two variables. It helps identify patterns and dependencies, aiding in decision-making, predictions, and understanding cause-and-effect relationships between 2 variables. Here we have more variables, which is why Heatmap is best suited for this analysis.



**Heatmap interpretation,**

Here, we can observe that variables INDUS and NOX are positively correlated with value 0.76 and variable DIS is negatively correlated with NOX and AGE with values -0.77 and -0.75 respectively.

Here, we can see that some variables are less correlated, meaning variables whose correlation values are between –0.7 to 0.7 can be eradicated from the analysis. However, model testing is required after the removal of variables to check that this does not affect the model significantly.

***However, if we normalize the data, following changes can be observed:***  
 **Magnitude Preservation:** While maintaining the direction of correlations (positive or negative), linear adjustments like Min-Max scaling or Z-score normalization may alter their magnitudes. If variables were initially scaled differently, correlations might become less significant.

**Nonlinear Transformations:** Nonlinear normalization, such as exponential or logarithmic transformations, can change correlations' amplitude and axis. Weakly correlated variables could become strongly correlated or vice versa.

**Outlier Sensitivity:** Some normalizing techniques are susceptible to outliers. Extreme values have the potential to significantly alter the normalized data and alter correlation coefficients.

**Data Distribution Influence:** The effects of normalization are influenced by data distribution. The degree to which skewed data or patterns are altered by normalization can affect the correlations that result.

**CONCLUSION**

Finally, our experience working on this project using the "BostonHousing" dataset has been a worthwhile learning opportunity. We have discovered and learned the following major insights:

The Foundation of Any Successful Data Analysis Project Is Thorough Data Preparation: We discovered the importance of meticulous data preparation. It's imperative to clean, organize, and handle missing data to guarantee the accuracy of the findings.

Python is a Powerful Tool: Python has shown to be incredibly useful for handling and analyzing complex datasets, especially when combined with libraries like pandas, matplotlib, seaborn, and others. We were able to swiftly complete a variety of data manipulation and visualization jobs thanks to these tools.

Outliers and Missing Data: Accurate analysis requires that outliers and missing data be addressed. We investigated a number of methods, including imputation and outlier handling, to preserve data integrity while making the most of the information at our disposal.

Insights Are Provided by Statistical Summaries and Visualizations: We were aware of the value of visual representations and statistical summaries in data analysis. They aid in the discovery of trends, outliers, and patterns that help make complex data easier to comprehend.

Decision-Making is Guided by Correlation Analysis: Understanding the connections between variables has shown to be a useful use of correlation analysis. It helps with prediction, cause-and-effect analysis, and informed decision-making.

Normalization Can Impact Interpretation: We recognized that data normalization can alter the magnitude and patterns of correlations, which is an important consideration when interpreting results and building models.

Data Analysis Transforms Raw Data into Knowledge: Ultimately, this project demonstrated the power of data analysis techniques in transforming raw data into actionable knowledge. It showed how data analysis can provide valuable insights for decision-making in domains like housing and urban planning.

In essence, this project has enriched our understanding of data analysis and its practical applications. It emphasized the importance of rigor, tools, and techniques in extracting meaningful information from data. As we move forward in our academic and professional journeys, the skills and knowledge acquired during this project will continue to serve us well in various analytical endeavors.

**REFERENCE**

**Harrison Jr., David**, and **Daniel L. Rubinfeld**. "*Hedonic housing prices and the demand for*  *clean air.*" Journal of Environmental Economics and Management 5.1 (1978): 81-102. Retrieved from *http://lib.stat.cmu.edu/datasets/boston*