ASSIGNMENT SUBMISSION COVER SHEET

Programme Title:	**MSc Data Analytics**
Module Code and Title:	**B9DA109 Machine Learning and Pattern Recognition**
Assessment Title:	**Individual Report: House Prices Prediction with Linear Regression**

Student Name	**Saima Khan**
Student ID	**20001871**
Word Count	**515**

In the group project, my main contribution was focused on the data preparation phase, specifically data pre-processing and exploratory data analysis (EDA). In terms of data pre-processing, I played a key role in handling missing values, removing duplicate rows, dealing with outliers, and ensuring data consistency.

Regarding missing values, we conducted a thorough analysis using Python code and found that there were no missing values in any of the features or rows of the dataset. This allowed us to proceed without the need for missing value imputation techniques. Similarly, we applied Python code to check for duplicate rows, and the result showed that none of the rows possessed any duplicates, eliminating the need for duplicates removal techniques.

Next, we focused on dealing with outliers. To detect outliers, we developed a Python function that utilized the Boxplot technique. By calculating quartiles and interquartile range (IQR), we identified features with significant percentages of outliers, such as 'price', 'sqft_lot', 'view', 'grade', and 'sqft_lot15'. To address these outliers, we utilized the Z-score method of outlier removal.

In terms of data consistency, I conducted various checks. Firstly, I examined the data types of all the features to ensure they aligned with the data they represented, and found that the data types were consistent. Additionally, I checked the unique values present in the categorical variables to identify any mistakes in the entries. For example, I discovered decimal figures in the 'bathroom' and 'floors' features, which should not be the case. To rectify this, I rounded the values to the nearest whole numbers and changed the data types to int64.

Moving on to the EDA phase, I contributed significantly to descriptive statistics and data visualization. For descriptive statistics, I generated a summary statistics table that provided important statistical values for the features in the dataset, including count, mean, standard deviation, and the five Boxplot points. By studying the values in the table, we confirmed that there were no errors in the data.

In terms of data visualization, I focused on univariate analysis to understand individual elements in the dataset. I utilized histograms, bar charts, and box plots to gain insights into various variables. For example, analyzing the price column through a histogram, I observed a slight negative skew, indicating that a larger portion of the prices in the dataset leaned towards higher values. I also used bar charts to examine the number of houses built by decades, which revealed trends and patterns in house construction over time.

Furthermore, I conducted bivariate analysis to explore relationships between variables. By utilizing box plots, I investigated the relationships between categorical variables and the target variable (price). For instance, I observed that higher grade scores were associated with higher house prices. Additionally, I examined the impact of factors like condition, view, waterfront presence, and number of floors on house prices.

Overall, my contribution to the group project involved handling missing values, removing duplicates, addressing outliers, ensuring data consistency, and conducting exploratory data analysis. This experience has enhanced my understanding of data preparation techniques and the importance of thorough analysis in deriving meaningful insights from the data.