**Analyzing Influential Factors in Boston Housing: Uncovering Patterns and Disparities**

Neha Vunnam (BL60852)

Poornima Kolasani (AL79610)

Santosh Kumar Velgapuri (YS79974)

**M.P.S Data Science, University of Maryland Baltimore County**

**DATA 690 - Statistical Analysis & Visualization with Python Term Paper**

**Professor Ergun Kacar**

**Dec 2023**

**Background: Recognizing Key Elements of Boston Housing**

The Boston housing dataset offers a thorough look at the different characteristics of cities and suburbs in the Boston Standard Metropolitan Statistical Area (SMSA) in 1970. Our goal is to explore the depths of the dataset's richness of data in order to identify the variables that have a major impact on median house values, therefore contributing to our quest to understand the complexities of the local housing market. With the goal of providing real estate industry participants, urban planners, and policymakers with important information, our study looks for trends and differences in the housing market in Boston. The collection includes a wide range of characteristics, such as the amount of nitric oxides present, land zoning percentages, crime rates, and the distance between homes and the Charles River. Finding out the complex links between these predictor factors and the median values of owner-occupied houses is what spurs this dataset's investigation.

Through statistical analysis and hypothesis testing, we aim to identify the critical elements that significantly influence house prices, providing insight into the workings of the Boston housing market in 1970. Throughout this investigation, we will make use of visual aids like histograms to clarify how housing attributes are distributed among various demographic groups. The dataset will then be divided into smaller groups so that we may perform statistical analysis and hypothesis testing, which is similar to the approach used in clinical research.

Hypothesis 1: There is no significant difference in median home values between areas with high and low average numbers of rooms.

Hypothesis 2: There is no significant difference in median home values between areas with high and low crime rates.

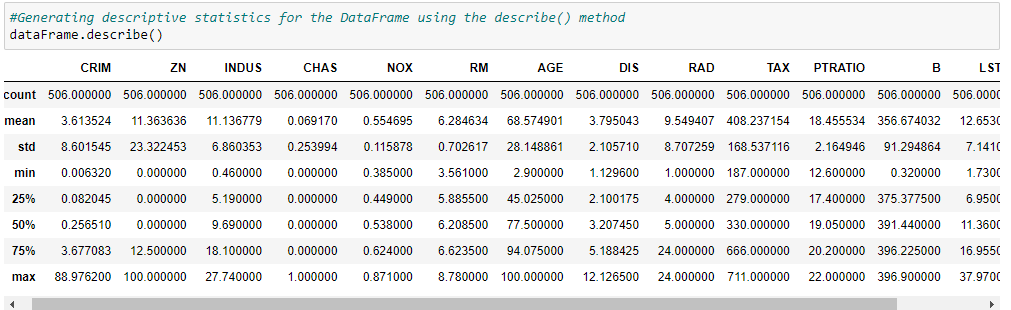
Hypothesis 3: There is no significant difference in median home values between areas with high and low distances to employment centers.

**Data:**

The Boston Housing Prices dataset, commonly employed in machine learning and statistical analysis, offers insights into factors influencing housing prices across various suburbs of Boston. Each row in the dataset corresponds to a distinct suburb and includes features such as the per capita crime rate (CRIM), the proportion of residential land zoned for large lots (ZN), and the proportion of non-retail business acres per town (INDUS). Other features encompass aspects like the presence of the Charles River (CHAS), nitric oxides concentration (NOX), average number of rooms per dwelling (RM), and the proportion of owner-occupied units built prior to 1940 (AGE). Additionally, variables like weighted distances to employment centers (DIS), accessibility to radial highways (RAD), full-value property tax rates (TAX), and pupil-teacher ratios (PTRATIO) provide a comprehensive view of the characteristics of each suburb.

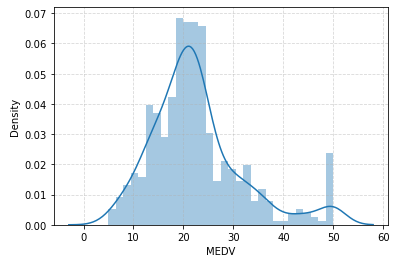
The dataset also includes indices reflecting racial diversity (B) and the percentage of lower-status population (LSTAT). The target variable, median value of owner-occupied homes (MEDV), serves as the focal point for regression analysis and predictive modeling, making this dataset a valuable resource for understanding the dynamics of housing markets.

**SUMMARY:**

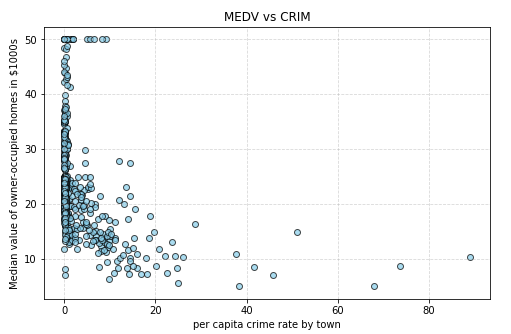


**VISUALIZATIONS:**

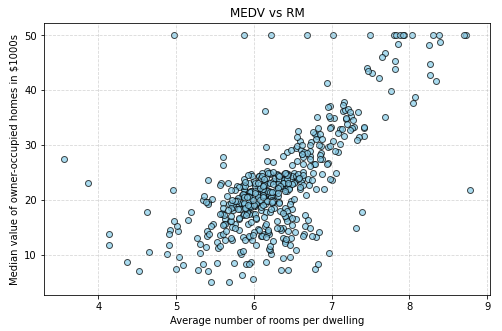
The below distributed plot for density of Median value of owner-occupied homes (MEDV) shows that on an average the distribution of the house prices are between 10,000 and 50,000 dollars.



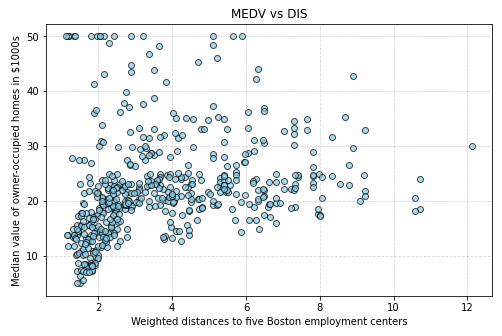
The scatter plot below says that if a neighborhood is safe (less crime), the houses there are more expensive, like between $12,000 and $40,000. But if a place has more crime, houses are cheaper. So, lower crime means costlier houses, and higher crime means cheaper houses.



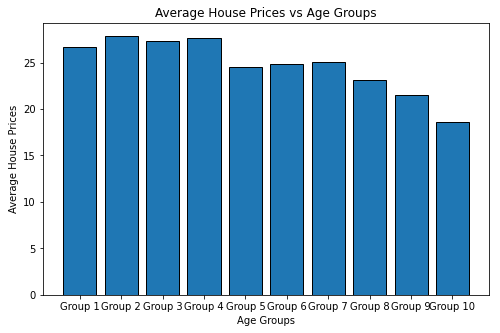
The scatter plot below shows that most houses in the dataset cost around 20,000 to 30,000 usd and have about 6 to 7 rooms. There are a few really big and expensive houses with more than 7 rooms, costing above 30,000, even up to 50,000 dollars.



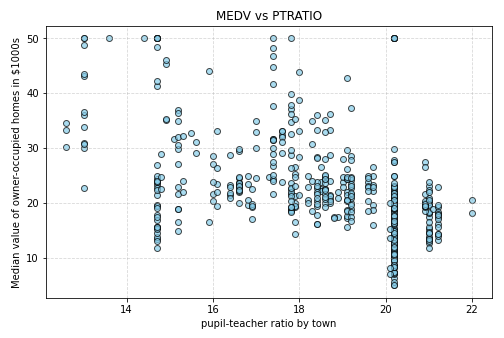
The scatter plot below shows that houses that cost between 5,000 and 30,000 $ in our data are usually not too far, like 1 to 6 units, from the 5 Boston job centers. If a house is close to where people work, it's more likely to be cheaper.



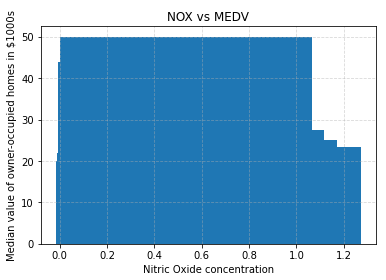
The bar graph below shows that If a house is old, like built before 1940, it tends to be cheaper, usually around $ 10,000 to $ 20,000. Most of these older houses have an age proportion between 80% and 100%



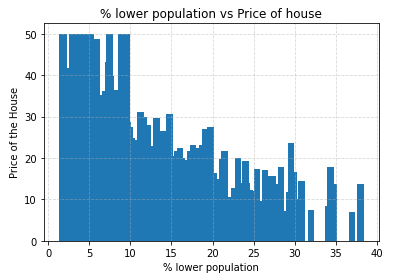
The plot below shows that houses where the pupil-teacher ratio by town is high the majority of the house prices range between $5-25.



The bar graph below depicts that if the nitric oxide concentration increases the rate of house decreases.



From the graph below, we can observe that if the lower status population is low price then the houses are high.



**Methods:**

First, we started off with collecting our dataset; the Boston Housing Prices dataset sourced from Kaggle. The dataset comprises various features that might influence the median value of owner-occupied homes (MEDV) in Boston suburbs. These features include crime rate (CRIM), zoning information (ZN), proximity to the Charles River (CHAS), and others. Then we performed data preprocessing. Upon loading the dataset using the Pandas library, headers were added to the DataFrame, and any missing values were addressed.   
 Descriptive statistics and information about the dataset were obtained, and a correlation matrix was computed to understand the relationships between different numerical variables. Then Exploratory Data Analysis was conducted to visually inspect the relationships between key features and the target variable (MEDV). Scatter plots were generated for variables such as crime rate (CRIM), average number of rooms (RM), distance to employment centers (DIS), and others. Additionally, a distribution plot was created to visualize the density of median home values.

Next, Hypothesis tests were performed to investigate the impact of specific predictor variables on median home values. For example, hypothesis tests were conducted on the average number of rooms (RM) and the percentage of lower-status population (LSTAT). The null hypothesis assumed no significant difference in median home values between different groups, while the alternative hypothesis posited a significant difference. Various visualization techniques, such as scatter plots and bar charts, were employed to provide a clear understanding of the relationships between predictor variables and median home values. These visualizations aided in identifying trends and patterns in the data.

A linear regression model was selected to predict median home values based on the chosen predictor variables. The dataset was split into training and test sets, with 80% used for training and 20% for testing. The scikit-learn library's LinearRegression class was utilized for model training, and performance metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared were employed to evaluate the model's accuracy on both the training and test sets.

Statistical analysis was employed to interpret the results of hypothesis tests and model performance metrics. The significance of relationships between variables and their impact on median home values was assessed.

**Results:**

Our exploration of the Boston Housing Prices dataset uncovered key insights into the dynamics influencing median home values in suburban Boston. The dataset, encompassing diverse features such as crime rates, zoning details, and environmental factors, set the stage for a comprehensive analysis. A distribution plot of median home values revealed a concentration between $20,000 and $30,000, providing an initial grasp of pricing trends. Notably, our scatter plots illuminated impactful patterns, particularly in the inverse correlation between crime rates and housing prices. Safer neighborhoods tended to exhibit higher prices, while areas with higher crime rates had more affordable housing options. Additionally, the positive correlation between the average number of rooms per dwelling and housing prices was evident, with larger and more expensive houses characterized by over 7 rooms.

Proximity to employment centers emerged as a significant factor, with houses priced between $5,000 and $30,000 typically situated closer to Boston's job hubs. Further, our exploration into the age of houses unveiled a consistent trend—older houses built prior to 1940 tended to be more affordable, falling within the $10,000 to $20,000 range. The creation of age groups reinforced this finding, emphasizing the inverse relationship between age and median prices. In parallel, an examination of the pupil-teacher ratio showcased a correlation with more affordable housing options, especially in areas with higher pupil-teacher ratios.

Moving beyond exploratory analysis, we applied a linear regression model to predict housing prices. The model performed well on both the test and train datasets. For the test data, the model displayed a mean squared error of 11.32 and a mean absolute error of 2.57, indicating low prediction errors and high accuracy. The R2 score of 0.76 highlighted the model's strong explanatory power. Similarly, on the train data, the model showed good performance with a mean squared error of 18.38, a mean absolute error of 3.03, and an R2 score of 0.79. Overall, these results suggest that the linear regression model effectively captured and explained the underlying patterns in the data, making it reliable for accurate predictions.

Hypothesis testing further strengthened our findings, revealing significant differences in median home values based on specific predictor variables. The first hypothesis on the influence of average number of rooms (RM) revealed a significant difference in median home values between areas with high and low average numbers of rooms. The second hypothesis on Impact of Crime Rate (CRIM) revealed that a substantial difference in median home values was identified between areas characterized by high and low crime rates. Finally, the third hypothesis on Neighborhood Disparities Based on Proximity to Employment Centers (DIS) revealed that a noteworthy difference in median home values was observed between areas with high and low distances to employment centers.

**Conclusion:**

In summary, our comprehensive analysis of the Boston housing dataset reveals key determinants significantly impacting median home values. The average number of rooms in a dwelling emerged as a critical factor, with areas boasting a higher room count exhibiting substantially elevated home values. Additionally, the crime rate proved to be a substantial influence, as regions with higher crime rates experienced notably diminished median home values. Proximity to employment centers stood out as another pivotal factor, with areas in close vicinity showcasing significantly higher home values.

These findings offer valuable insights for homebuyers and real estate stakeholders, emphasizing the importance of considering housing characteristics and neighborhood features in property evaluations. Moreover, policymakers can leverage these insights for informed urban planning and development strategies, contributing to the enhancement and growth of communities. This conclusive understanding of influential factors provides a nuanced perspective that is integral for stakeholders in the real estate and urban planning domains.

|  |
| --- |
| #Reading data from dataset file  path = 'C:**\\**Users**\\**NEHA**\\**Downloads**\\**housing.csv'  dataFrame = pd.read\_csv(path) |

|  |
| --- |
| #Adding headers to the dataframe  headers = ['CRIM', 'ZN', 'INDUS', 'CHAS', 'NOX', 'RM', 'AGE', 'DIS', 'RAD', 'TAX', 'PTRATIO', 'B', 'LSTAT', 'MEDV']  dataFrame= pd.read\_csv(path, sep='\s+', names= headers)  **print**(dataFrame) |

|  |
| --- |
| dataFrame.shape |

#Creating a scatter plot for the relationship between average number of rooms and housing price

plt.figure(figsize=(**8**, **5**))

plt.scatter(dataFrame.RM, dataFrame.MEDV, color='skyblue', marker='o', edgecolors='black', alpha=**0.7**)

plt.xlabel("Average number of rooms per dwelling")

plt.ylabel("Median value of owner-occupied homes in $1000s")

plt.title("MEDV vs RM")

plt.grid(True, linestyle='--', alpha=**0.5**)

|  |
| --- |
| #Checking and summing null values in each column of the DataFrame.  dataFrame.isnull().sum() |

|  |
| --- |
| #Displaying information about the DataFrame, including data types and non-null counts.  dataFrame.info() |

|  |
| --- |
| #Generating descriptive statistics for the DataFrame using the describe() method  dataFrame.describe() |

|  |
| --- |
| #Finding correlation for all numeric values  dataFrame.corr() |

|  |
| --- |
| #Plotting a distribution plot for the 'MEDV' column in the DataFrame  sns.distplot(dataFrame['MEDV'], **30**)  plt.grid(True, linestyle='--', alpha=**0.5**)  plt.show() |

|  |
| --- |
| #Creating a scatter plot for the relationship between per capita crime rate and housing price  plt.figure(figsize=(**8**, **5**))  plt.scatter(dataFrame.CRIM, dataFrame.MEDV, color='skyblue', marker='o', edgecolors='black', alpha=**0.7**)  plt.xlabel("per capita crime rate by town")  plt.ylabel("Median value of owner-occupied homes in $1000s")  plt.title("MEDV vs CRIM");  plt.grid(True, linestyle='--', alpha=**0.5**) |

|  |
| --- |
| #Creating a scatter plot for the relationship between weighted distances to employment centers and housing prices  plt.figure(figsize=(**8**, **5**))  plt.scatter(dataFrame.DIS, dataFrame.MEDV, color='skyblue', marker='o', edgecolors='black', alpha=**0.7**)  plt.xlabel("Weighted distances to ﬁve Boston employment centers")  plt.ylabel("Median value of owner-occupied homes in $1000s")  plt.title("MEDV vs DIS")  plt.grid(True, linestyle='--', alpha=**0.5**) |

|  |
| --- |
| #Creating age groups based on the 'AGE' column and calculating the average 'MEDV' for each group  bins = [**0**, **10**, **20**, **30**, **40**, **50**, **60**, **70**, **80**, **90**, **100**]  labels = [f'Group {i}' **for** i **in** range(**1**, **11**)]  dataFrame['AgeGroup'] = pd.cut(dataFrame['AGE'], bins=bins, labels=labels, right=False)  #Calculating the average 'MEDV' for each age group  grouped\_data = dataFrame.groupby('AgeGroup')['MEDV'].mean().reset\_index()  #Plotting a bar chart to visualize the relationship between age groups and average house prices  plt.figure(figsize=(**8**, **5**))  plt.bar(grouped\_data['AgeGroup'], grouped\_data['MEDV'], edgecolor='black')  plt.xlabel("Age Groups")  plt.ylabel("Average House Prices")  plt.title("Average House Prices vs Age Groups")  plt.show() |

|  |
| --- |
| #Creating a scatter plot for the relationship between pupil-teacher ratio and housing prices  plt.figure(figsize=(**8**, **5**))  plt.scatter(dataFrame.PTRATIO, dataFrame.MEDV, color='skyblue', marker='o', edgecolors='black', alpha=**0.7**)  plt.xlabel("pupil-teacher ratio by town")  plt.ylabel("Median value of owner-occupied homes in $1000s")  plt.title("MEDV vs PTRATIO")  plt.grid(True, linestyle='--', alpha=**0.5**) |

|  |
| --- |
| #Creating a bar chart to visualize the relationship between nitric oxide concentration and house prices  plt.bar(dataFrame.NOX, dataFrame.MEDV)  plt.xlabel('Nitric Oxide concentration')  plt.ylabel('Median value of owner-occupied homes in $1000s')  plt.title('NOX vs MEDV')  plt.grid(True, linestyle='--', alpha=**0.5**)  plt.plot() |

|  |
| --- |
| #Creating a bar chart to visualize the relationship between the percentage of lower status population and house prices  plt.bar(dataFrame.LSTAT, dataFrame.MEDV)  plt.xlabel('% lower population')  plt.ylabel('Price of the House')  plt.title('% lower population vs Price of house')  plt.grid(True, linestyle='--', alpha=**0.5**)  plt.plot() |

|  |
| --- |
| #Dropping rows with missing values and removing the 'AgeGroup' column from the DataFrame  dataFrame = dataFrame.dropna()  dataFrame = dataFrame.drop('AgeGroup', axis = **1**) |

|  |
| --- |
| #Splitting the data into features (X) and target variable (y)  **from** **sklearn.model\_selection** **import** train\_test\_split  x = dataFrame.drop("MEDV", axis = **1**) #Features excluding the target variable  y = dataFrame["MEDV"] #Target variable  #Splitting the data into training set (80%) and test set (20%)  x\_train, x\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size=**0.2**, random\_state=**42**)  #Displaying the shapes of the training and test sets  **print**("Train set shape : ", x\_train.shape, y\_train.shape)  **print**("Test set shape : ", x\_test.shape, y\_test.shape) |

|  |
| --- |
| #Creating a Linear Regression model and training it on the training data  lrModel = LinearRegression()  lrModel.fit(x\_train, y\_train)  #Making predictions on the training set  y\_pred\_train = lrModel.predict(x\_train)  #Making predictions on the test set  y\_pred\_test = lrModel.predict(x\_test) |

|  |
| --- |
| #Making predictions on the test set  y\_pred\_test = lrModel.predict(x\_test)  #Evaluating the model's performance on the test set  test\_accuracy = lrModel.score(x\_test, y\_test) \* **100**  #Evaluating the model's overall accuracy  model\_accuracy = r2\_score(y, lrModel.predict(x)) \* **100**  **print**("Testing accuracy is:", test\_accuracy)  **print**("Model accuracy is:", model\_accuracy) |

|  |
| --- |
| #Evaluating the model's performance  mse\_test = mean\_squared\_error(y\_test, y\_pred\_test)  mae\_test = mean\_absolute\_error(y\_test, y\_pred\_test)  r2\_test = r2\_score(y\_test, y\_pred\_test)  mse\_train = mean\_squared\_error(y\_train, y\_pred\_train)  mae\_train = mean\_absolute\_error(y\_train, y\_pred\_train)  r2\_train = r2\_score(y\_train, y\_pred\_train)  **print**("Linear Regression Model Performance on test data : ")  **print**("Mean Squared Error : ", mse\_test)  **print**("Mean Absolute Error : ", mae\_test)  **print**("R2 Score : ", r2\_test)  **print**("**\n**Linear Regression Model Performance on train data : ")  **print**("Mean Squared Error : ", mse\_train)  **print**("Mean Absolute Error : ", mae\_train)  **print**("R2 Score : ", r2\_train) |

|  |
| --- |
| **from** **scipy.stats** **import** ttest\_ind  #Hypothesis 1: Influence of Average Number of Rooms (RM) on Median Home Values  highRM = dataFrame[dataFrame['RM'] > dataFrame['RM'].median()]['MEDV']  lowRM = dataFrame[dataFrame['RM'] <= dataFrame['RM'].median()]['MEDV']  tstat\_RM, pvalue\_RM = ttest\_ind(highRM, lowRM)  **print**("**\n**Hypothesis Test for RM : ")  **print**(f"T-statistic : {tstat\_RM}, P-value : {pvalue\_RM}")  #Interpretation for 'RM': If p-value is less than 0.05, reject the null hypothesis  **if** pvalue\_RM < **0.05** :  **print**("There is a significant difference in median home values between areas with high and low average number of rooms.")  **else** :  **print**("There is no significant difference in median home values between areas with high and low average number of rooms.")  #Hypothesis 2: Impact of Crime Rate (CRIM) on Median Home Values  highCRIM = dataFrame[dataFrame['CRIM'] > dataFrame['CRIM'].median()]['MEDV']  lowCRIM = dataFrame[dataFrame['CRIM'] <= dataFrame['CRIM'].median()]['MEDV']  tstat\_CRIM, pvalue\_CRIM = ttest\_ind(highCRIM, lowCRIM)  **print**("**\n**Hypothesis Test for CRIM : ")  **print**(f"T-statistic: {tstat\_CRIM}, P-value: {pvalue\_CRIM}")  #Interpretation for 'CRIM': If p-value is less than 0.05, reject the null hypothesis  **if** pvalue\_CRIM < **0.05** :  **print**("There is a significant difference in median home values between areas with high and low crime rates.")  **else**:  **print**("There is no significant difference in median home values between areas with high and low crime rates.")  #Hypothesis 3: Neighborhood Disparities Based on Proximity to Employment Centers (DIS)  highDIS = dataFrame[dataFrame['DIS'] > dataFrame['DIS'].median()]['MEDV']  lowDIS = dataFrame[dataFrame['DIS'] <= dataFrame['DIS'].median()]['MEDV']  tstat\_DIS, pvalue\_DIS = ttest\_ind(highDIS, lowDIS)  **print**("**\n**Hypothesis Test for LSTAT:")  **print**(f"T-statistic: {tstat\_DIS}, P-value: {pvalue\_DIS}")  # Interpretation for 'DIS': If p-value is less than 0.05, reject the null hypothesis  **if** pvalue\_DIS < **0.05** :  **print**("There is a significant difference in median home values between areas with high and low distances to employment centers.")  **else**:  **print**("There is no significant difference in median home values between areas with high and low distances to employment centers.") |

**References:**

Vikrishnan, V. (n.d.). Boston House Prices. [Dataset]. Kaggle.

[**https://www.kaggle.com/datasets/vikrishnan/boston-house-prices**](https://www.kaggle.com/datasets/vikrishnan/boston-house-prices)