**IS471 Semester Project Final Report**

Project Group 1

Alexandru Diloreanu

Lee Business School, University of Nevada - Las Vegas

Dr. Michael Lee

**Introduction**

For the project, I analyzed the Housing dataset. The dataset includes information on various socioeconomic, geographic, and structural features of neighborhoods in the area. The main goal was to uncover the most significant factors influencing housing prices and determine whether the common belief that “location, location, location” truly holds up. Through my analysis, I explored the relationships between multiple features and the median value of homes, ultimately aiming to build a predictive model for housing prices.

The key areas I focused on were Correlation Analysis, Regression Modeling, and Feature Importance. I also evaluated the broader implications of these findings for policymakers, homebuyers, and real estate professionals.

The preliminary investigation revealed that while location does play an important role in determining housing prices, factors such as house size, neighborhood status, and environmental quality often have a more pronounced impact. In particular, I found strong positive relationships between housing prices and features such as the number of rooms, while variables like crime rate and population status were negatively correlated with value.

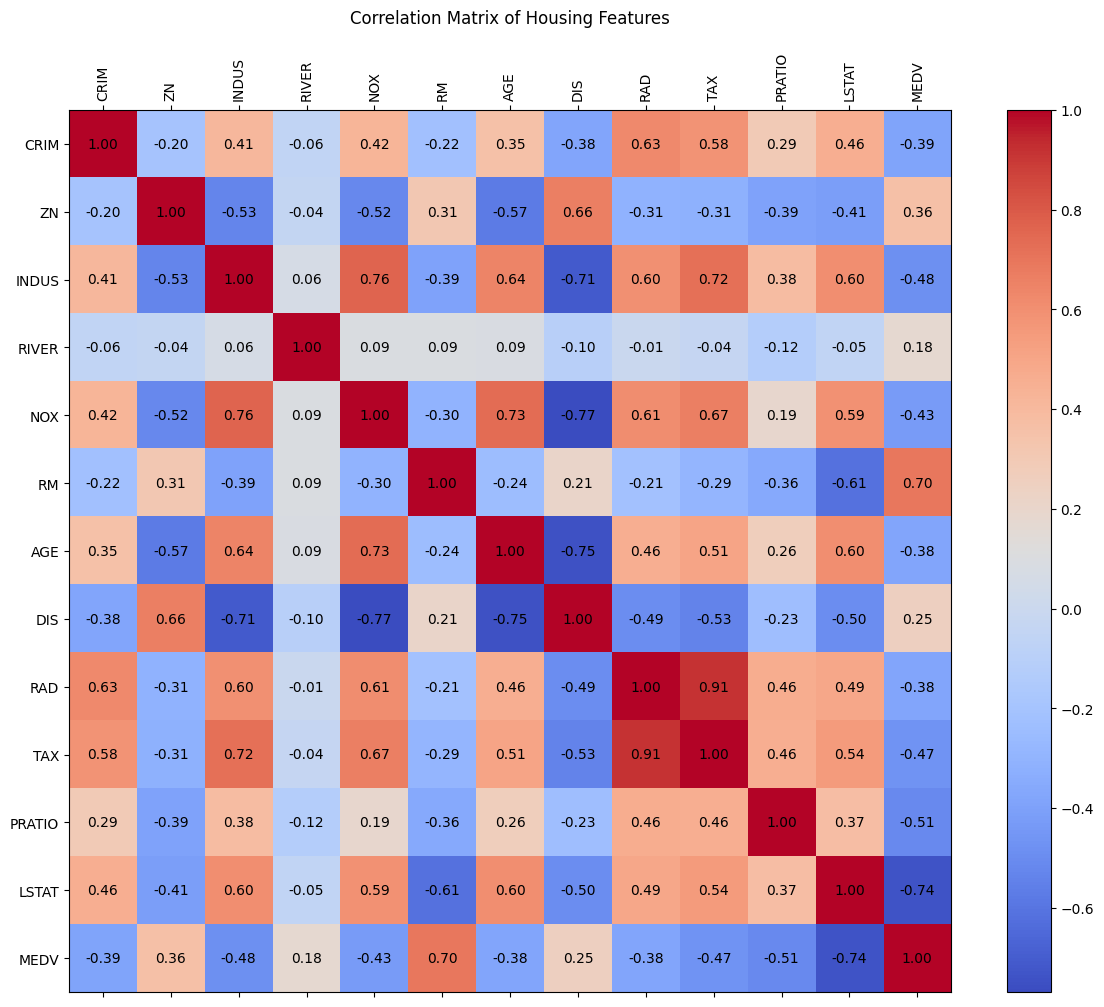
**Methodology & Technical Implementation**

I used Python as my tool for data exploration, visualization, and modeling. I utilized libraries such as Pandas to clean and visualize the data, and Scikit-learn to build and evaluate our linear regression model. The dataset consists of 506 records with 13 features, including both numerical and categorical variables. I started by examining the distribution and relationships between variables using correlation matrices and heatmaps. This helped me identify which features were most strongly related to median house prices and guided my selection of predictors for modeling. I then built a multiple linear regression model to estimate MEDV using several explanatory variables. The RIVER feature, which indicates whether a tract borders the River, was encoded as a binary variable with 1 for “Yes” and 0 for “No.” I split the dataset into 80% training and 20% testing subsets and evaluated the model’s performance using mean squared error and R-squared.

The correlation matrix revealed several important relationships. The average number of rooms per dwelling had a strong positive correlation of 0.70 with housing prices. The percentage of lower-status population was negatively correlated at -0.74, suggesting that neighborhoods with higher proportions of disadvantaged residents tended to have lower house prices. Other negatively correlated variables included NOX, TAX, and PTRATIO, which all showed negative associations with MEDV.

**Findings & Insights**

**Correlation Analysis**

****

My correlation analysis showed that RM and LSTAT are the most influential features affecting housing prices. A higher average number of rooms is strongly associated with increased property values, while a higher proportion of lower-status residents is strongly associated with lower property values. Environmental and education-related variables such as NOX, TAX, and PTRATIO also demonstrated negative relationships with price, though not as strongly as RM and LSTAT. This suggests that larger, more exclusive homes in cleaner, better-educated neighborhoods tend to command higher prices. This supports some elements of the “location” hypothesis while also highlighting the impact of structural and socioeconomic characteristics.

**Linear Regression Modeling**

****

****

I developed a linear regression model using features identified as impactful from our correlation analysis. The performance metrics were promising, with a mean squared error of 28.99 and an R-squared value of 0.684. These results indicate that our model explains approximately 68.4% of the variance in housing prices, which is considered strong for a basic regression approach.

The regression coefficients provided additional insight into the influence of each feature. The RM variable had a coefficient of +3.81, indicating that each additional room increased the median home value by approximately $3,810. LSTAT had a coefficient of -0.55, meaning that for every 1% increase in lower-status population, the median house value dropped by about $550. NOX had a significant negative coefficient of -15.45, while DIS and PTRATIO had coefficients of -1.39 and -0.97, respectively. Interestingly, the RIVER variable had a positive coefficient of +2.90, suggesting that proximity to the River adds value to nearby properties.

**Interpretation**

My results show that the assumption that “location” is the main driver of housing value is only partially true. While proximity to desirable features like water access and city centers does matter, our analysis indicates that housing structure and neighborhood demographics have the most substantial impact on price. This shifts some focus away from geographic location alone and toward a more holistic view of property value.

**Conclusion & Recommendations**

While location-related features like river access and proximity to city centers do matter, our findings show that housing structure and neighborhood demographics have the most substantial impact. I recommend that policymakers focus on improving education quality and reducing pollution in disadvantaged areas, as these factors have a measurable effect on home values. Additionally, collecting new data such as school ratings, crime trends over time, or public transit access could further enhance the model’s predictive power.

Overall, this project helped me gain a deeper understanding of data analysis. As well as how data analysis can be used for real-world applications outside of just a classroom setting.