The given project applies machine learning technique linear regression and a  K-means clustering method, to analyze and predict house prices based on property size, number of bedrooms, and proximity to amenities. This project uses regression modeling combined with clustering to provide insights into pricing trends and market segmentation. The results of the research show how well the model can predict housing prices while segmenting properties into distinct categories for better understanding. House pricing is a very complex problem that is influenced by a lot of factors, including size, location, and distance to the city center or transport hubs. Accurate prediction of housing prices will support stakeholders in making better decisions. The proposed project leverages the previous studies on regression modeling for price prediction and clustering techniques for market segmentation to present an integrated approach to the solution.

The data set includes important property features: number of bedrooms, net square meters of the house, city center distance, metro distance, floor number, age of the property, and price. Property size was most strongly related to the price, at 0.68, and then bedroom count at 0.55. After that  comes city center distance 0.42 and metro distance with 0.15. Age and Floor number had a negative correlation with price which indicates that the older a property is and how high up it is in the building takes away from its value. These insights drove choices in modeling, using regression to predict price using bedroom count, size, and proximity to metro and city center. I did not have to organize or change the data set much as it was already fit to be used. I just made a small tweak where I changed the distance to the metro and city center to negative values to reflect how far they are from the housing unit.

Two machine learning models were implemented in this project: K-means clustering and linear regression. K-means clustering was selected because it can segment properties into distinct groups, which is useful for market understanding. The algorithm identified three clusters representing small, affordable properties, medium-sized mid-priced properties, and large, expensive properties. Linear regression was chosen because it is simple and interpretable, thus being an ideal choice for price prediction. The regression model incorporated features that have a strong correlation to price, which included the number of bedrooms, the size of the property, and proximity metrics to maximize the predictive accuracy. The K-means clustering model was trained on standardized features such as property size and price. The best number of clusters was determined to be three, representing distinct market segments. Regarding linear regression, the dataset was split into training and testing sets, with 80% of the data used for training. The model was trained on features like a number of bedrooms, property size, and proximity metrics, excluding features with very little correlation to price. Validation was done by testing the model on unseen data and checking its performance using metrics like MAE and RMSE. These approaches ensured robust and generalizable models.

The K-means clustering model segmented the data into three clusters: small, low-priced; medium-sized, mid-priced; and large, high-priced. This gave great insight into market segmentation and trends. The linear regression model did a great job in prediction, as evidenced by the MAE and RMSE, showing very good accuracy. A scatter plot of actual versus predicted prices showed good alignment along the ideal diagonal, hence showing the effectiveness of the model. Overall, this analysis enabled the comprehensive study of housing prices through a combination of clustering and regression. While the project reached its goals, there are some areas for improvements. Further steps can be done by developing more complex models, such as Random Forest and Gradient Boosting, with the ability to capture non-linear relationships. The model also can be further improved with more features, such as neighborhood quality or other local economic indicators. I could deploy the model as a web application for real estate professionals and buyers to access in real time and make predictions on their own. Other directions for future work would include expanding the analysis to encompass rental markets or commercial properties.

Major stakeholders that are  involved are real estate agents, property buyers, and property sellers. Real estate agents can use the output of the clustering step in order to identify target markets or even price trends, and both buyers and sellers could also rely on the model for more accurate price prediction in order to make their decision; wrong predictions may lead them in a wrong direction. Thus, it is critical to have more model interpretability and validation. Further testing and feedback on the model by engaging stakeholders may further enhance its applicability and usability. This project has shown the use of linear regression and K-means clustering for housing price prediction and market segmentation. It combined both models to give practical insights into property valuation and market dynamics. The results confirm the appropriateness of the chosen methods and indicate a very good potential for further development and application in real-world scenarios.

# Works Cited

1. F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011. Available: https://scikit-learn.org/stable/.
2. J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90-95, 2007. Available: https://matplotlib.org/stable/.
3. "K-means clustering," Wikipedia. Available: https://en.wikipedia.org/wiki/K-means_clustering. [Accessed: Dec. 2, 2024].
4. "Linear regression," Wikipedia. Available: https://en.wikipedia.org/wiki/Linear_regression. [Accessed: Dec. 2, 2024].