APPLIED DATA SCIENCE -1
ASSIGNMENT - 2: CLUSTERING AND FITTING

Name: Praneet Sivakumar

Student Id: 23095964

GitHub Link:
https://github.com/Prani8/Clustering-and-fitting.git

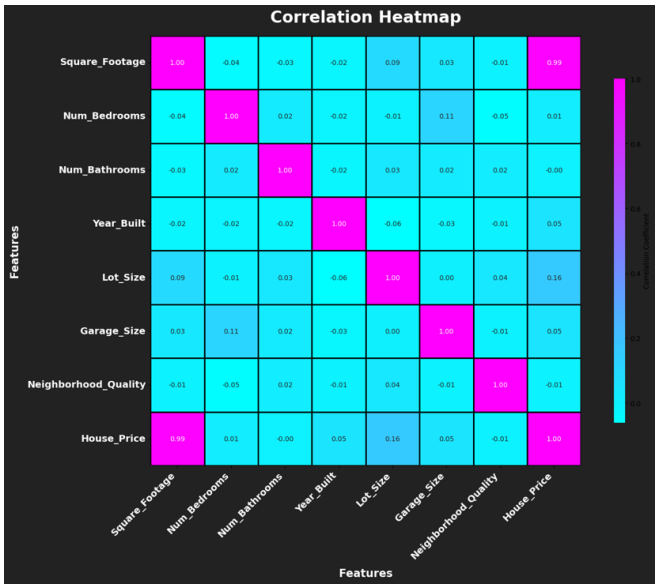
Dataset Link:
https://www.kaggle.com/datasets/prokshitha/home-value-insights

Abstract:

Using cutting-edge data analysis and visualisation techniques, this research examines the connections between home prices and important property characteristics. The Elbow Method is used to optimise a K-means clustering strategy that divides the dataset into three significant categories. As a crucial predictor, the correlation heatmap demonstrates the substantial positive association between square footage and home price. Additional information about cluster distributions and inter-feature relationships may be found using pairwise plots. According to the Enhanced Regression Fit plot, which shows few prediction errors, a linear regression model is highly accurate at predicting home values. By confirming the use of clustering and regression for property analysis and highlighting square footage as the most significant feature, these results provide a thorough comprehension of the dataset.

Introduction

A dataset comprising property attributes and their associated home values is the subject of the analysis. To determine how they relate to home values, important characteristics such square footage, lot size, number of bedrooms, and neighbourhood quality are assessed. To find patterns and trends, a variety of data science techniques are used, such as clustering, regression, and exploratory data analysis. A linear regression model assesses the predictability of home prices based on available features, while the K-means clustering technique groups qualities into meaningful clusters. We seek to understand the elements that influence home prices and derive actionable insights from the data using visualisations such as correlation heatmaps, scatter plots, and pair plots.

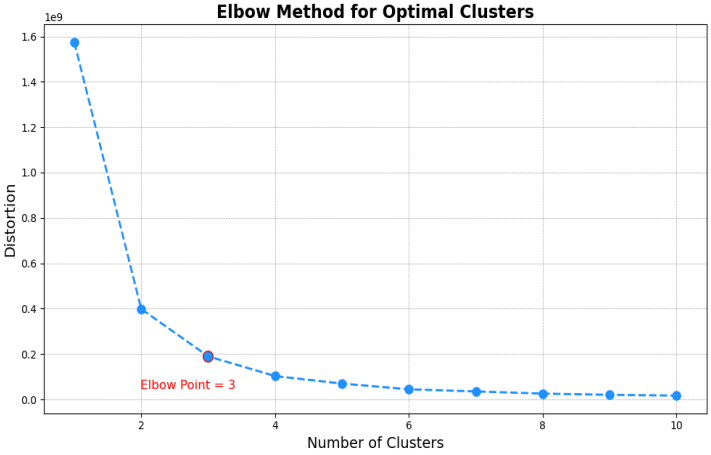


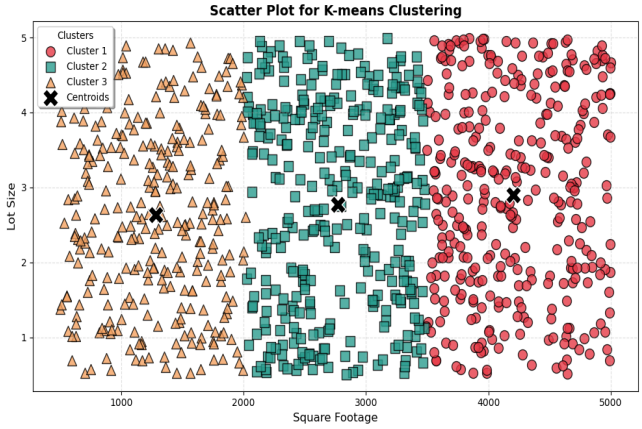
Heatmap:

Important correlations between the dataset's properties are highlighted in this heatmap. The most important predictor is Square Footage, which has an almost perfect positive correlation with House Price (0.99). Garage Size (0.05) and Lot Size (0.16) have little effect, indicating that interior space is more important than exterior characteristics. It's interesting to note that Num Bedrooms and Num Bathrooms have virtually little relationship with price, suggesting that the quantity of rooms has no bearing on costs. A little positive association (0.05) between Year Built and price suggests that newer homes might sell for a little more. Due to its subjective nature, Neighborhood_Quality does not correspond with any of the criteria. With its vivid gradient, the heatmap successfully highlights these trends, highlighting Square Footage as the main factor influencing home values and laying the groundwork for additional research.

Elbow Method:

This diagram presents the Elbow Method, it is a standard method for finding the right number of clusters for the application of K-means. For example, the number of clusters corresponds to the x-axis, and the y-axis contains the distortion (or inertia), which includes the sum of squared distances from each data point to the cluster centroid assigned to that data point. At one end, distortion is high: one cluster produces a great deal of distortion at one end. Because of an increase in the number of clusters, the distortion reduces sharply. This is mainly due to the points coming much closer to the centroids. After a certain number, reduction in distortion is very small and it forms a distinct elbow in the curve. In this graph, the elbow point is three clusters, which has been highlighted in red. It is the point beyond which adding more clusters gives minimal improvement on reducing distortion. The elbow symbolizes this trade-off between having either few clusters (underfitting) or too many clusters (overfitting). It means that a choice of three clusters permits a development focusing all real patterns of the data without overloading. The dashed blue line with an annotated elbow point makes sense and easy to grasp for the plot, serving as a guide in selecting the best number of clusters for the dataset.



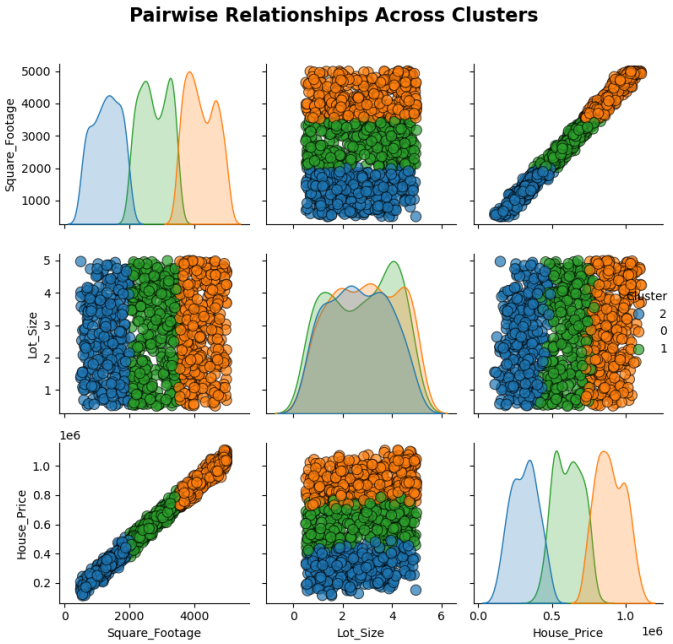


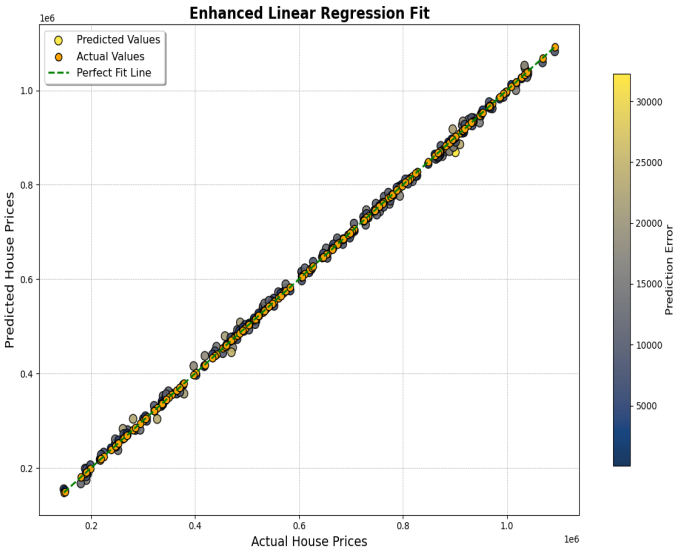
Scatter Plot:

Only such a scatter plot demonstrates K-means clustering per plot points grouped into three clusters according to their Square Footage and Lot Size for representation: Cluster 1 in red circles, Cluster 2 in green squares, and Cluster 3 in orange triangles, although the colour and marker are used to identify the clusters more easily. The central markers at X in black are for the centroid of each cluster; that is, the central point that minimizes the distance between itself and all the points that fall into its cluster. It collects results from the K-means algorithm, which carried out iterative assignments of points to clusters for optimal data segmentation. The distribution of points indicates that Cluster 1 (red) contained properties with larger square footage, Cluster 3(orange) with smaller square footage, and Cluster 2 (green) with points in between. This segmentation follows naturally groups inside the dataset based on the two endpoints being analysed. The plot legend allows the easy identification of the clusters, while the axes clearly depict the relationship between the two variables: lot size and square footage. This visualization demonstrates the power of K-means clustering to create meaningful differentiations and segments data into logical groups for further analysis.

Pair Plot:

This pair plot, which highlights three separate clusters found by the K-means algorithm, offers a straightforward and understandable method to comprehend correlations between important variables in the dataset. To make them easy to distinguish, each cluster is colour-coded: orange for Cluster 3, green for Cluster 2, and blue for Cluster 1. Smooth density curves (KDE) are used in diagonal plots to display feature distributions; for example, the Square Footage segmentation across clusters is evident. Relationships between attributes are shown by scatter plots. For example, there is a substantial positive correlation between Square Footage and House Price, indicating that larger homes (Cluster 3, in orange) have higher prices. On the other hand, Lot Size shows a lesser correlation with price, suggesting that land size is less important. The clusters are generally discrete and non-overlapping, demonstrating how well K-means segments properties.





Linear Regression

This video shows improved linear regression fits with respect to actual house prices (x-axis) and predicted house prices (y-axis). Every data point on the scatter plot is a house, and their position indicates how close the prediction from the model is to the actual value. The yellow points are the actual ones, while the darker gradient circles are the predicted ones with the colour intensity of each point indicating the prediction error (colour bar is reference). Lighter colour points indicate greater error while darker shades indicate points are much closer. The dashed green line represents the perfect fit line which would indicate if the model is really good, where actual values would be and compared to predicted values. Most of the points are closely clustered along this line, indicating that there is a very high linear relationship and quite a high accuracy in the regression model. The colour bar to the right is pretty good in scale for reading the prediction error, going from low to high. It graphically captures also the performance of the model, having most points closely huddled indicating highly reliable prediction by the model in terms of house pricing. All these indices present a graph summarizing that this model is quite capable, and at the same time, it flashes the prediction power to the reader. Further, it is apparent and easy to recognize the outlier and gain insight into where the model fails. Thus, one of the important differences between these graphs is for the value judgment of performance for the regression model.

Conclusion

The findings from this analysis reveal that house prices are influenced by key factors, the most significant among which are those relevant to the Square Footage feature defined herein. This has a near perfect correlation with house prices. K-means clustering effectively provides three clusters for the dataset in question, and hence emphasizes meaningful groupings along dimension characteristics size and price. The pairwise plots show up very well how features correlate with one another; for instance, between square footage and house price, it is indeed linear. Yet, the regression model proves better regarding these relationships' reliability in prediction of house prices. Overall, the methods employed-clustering, correlation analysis, regression-demonstrate their efficacy in uncovering patterns in data on property. These results can direct the decision towards informed ones on real estate prices as well as investments.