GitHub link: https://github.com/shrijith2002/clustering-and-fitting

Student id: 23017322

Dataset link: https://www.kaggle.com/datasets/camnugent/california-housing-prices/data

Insights and Analysis of California Housing Dataset:

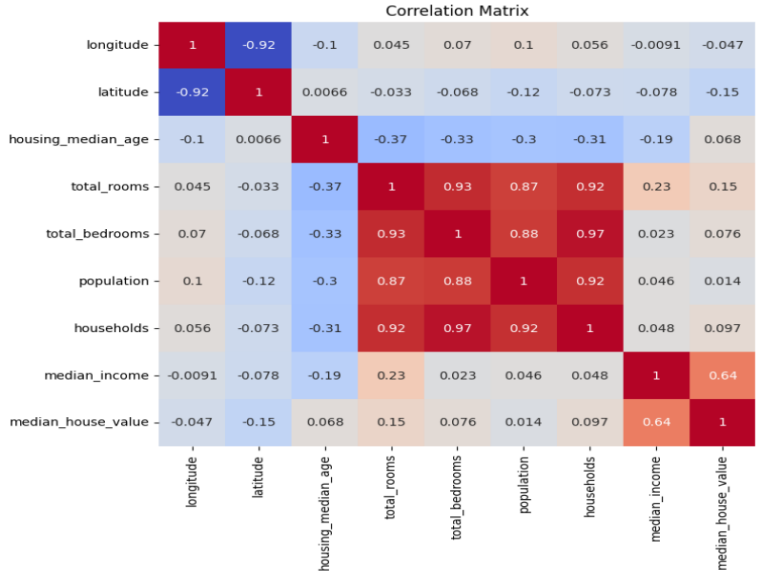
The dataset from the 1990 California census offers a rich resource for exploring housing dynamics and building predictive models. This report delves into various aspects of the dataset, including data exploration, visualization, missing value handling, clustering, and predictive modeling using linear regression.

Major Moments:

	Mean	Median	Standard Deviation	Skewness
longitude	-119.569704	-118.4900	2.003532	-0.297801
latitude	35.631861	34.2600	2.135952	0.465953
housing_median_age	28.639486	29.0000	12.585558	0.060331
total_rooms	2635.763081	2127.0000	2181.615252	4.147343
total_bedrooms	536.852229	435.0000	419.390765	3.481072
population	1425.476744	1166.0000	1132.462122	4.935858
households	499.539680	409.0000	382.329753	3.410438
median_income	3.870671	3.5348	1.899822	1.646657
median_house_value	206855.816909	179700.0000	115395.615874	0.977763

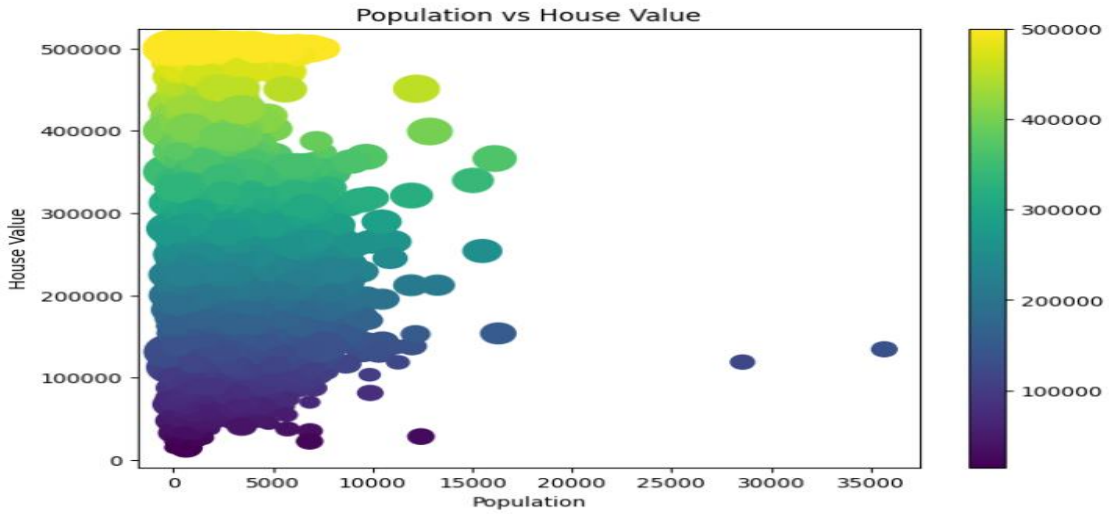
	Kurtosis
longitude	-1.330152
latitude	-1.117760
housing_median_age	-0.800629
total_rooms	32.630927
total_bedrooms	22.242583
population	73.553116
households	22.057988
median_income	4.952524
median_house_value	0.327870

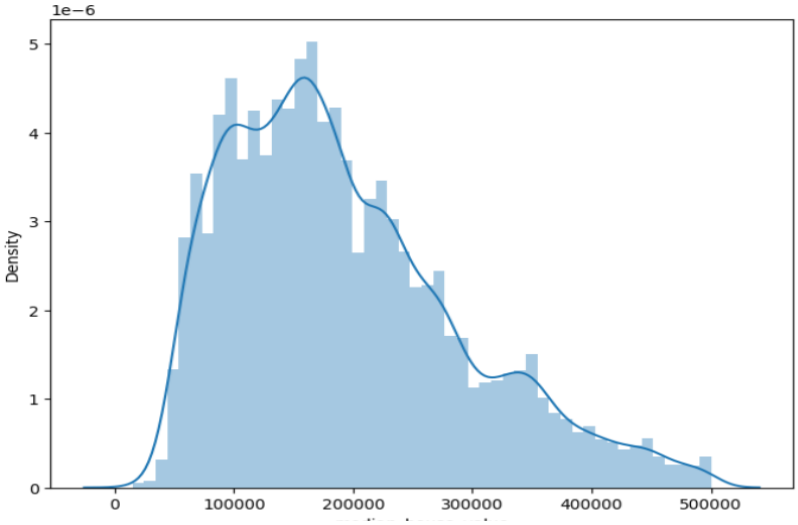
The mean longitude and latitude values indicate that the dataset covers a broad geographical area across California. The housing median age, with a slightly positive skewness and low kurtosis, suggests a relatively balanced distribution of housing age, albeit with a slight right skew. However, the total number of rooms, bedrooms, population, and households exhibit significant positive skewness and high kurtosis, indicating heavy-tailed distributions with a concentration of data towards lower values but with some extreme outliers on the higher end.



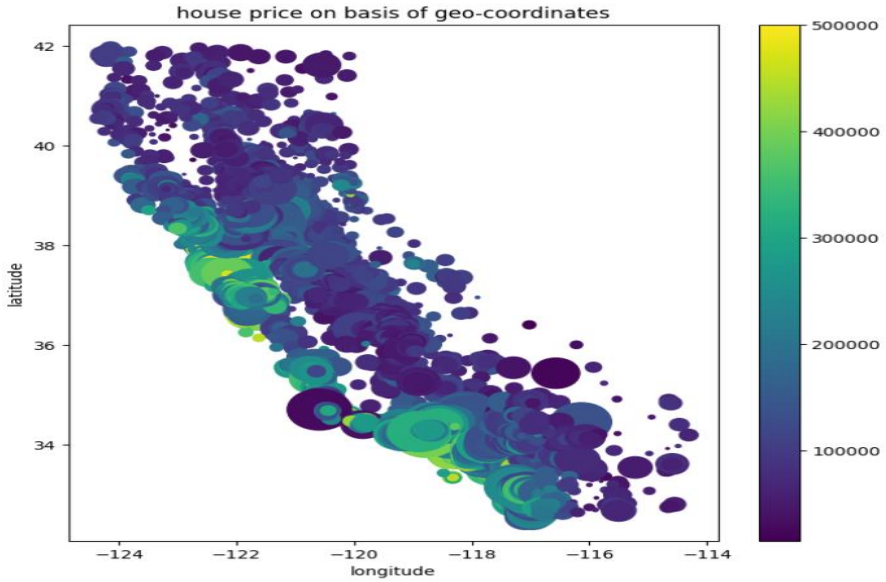
The heatmap of the correlation matrix reveals the relationships between different features in the dataset. Strong correlations (positive or negative) between variables can provide insights into potential predictors of median house value and guide feature selection for predictive modeling.

The scatter plot visualizes the relationship between population and median house value. It suggests that areas with higher population densities tend to exhibit a wider range of median house values, with some areas showing notably higher values.

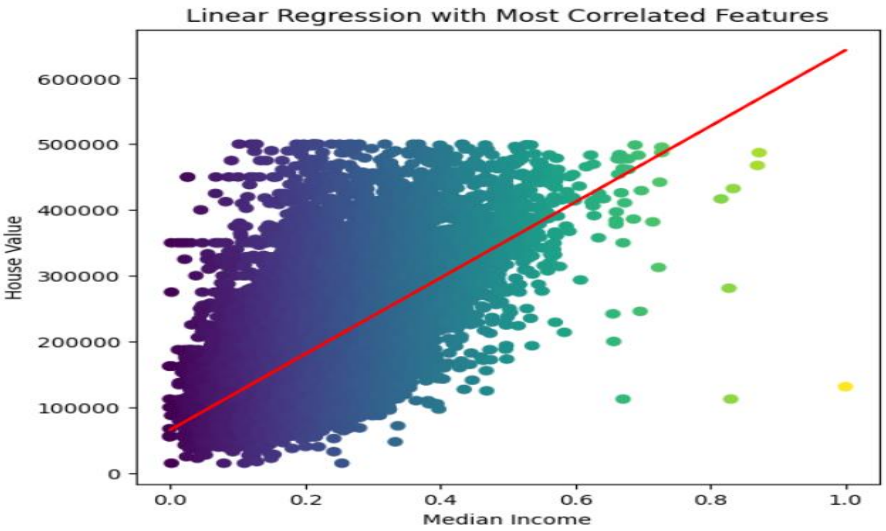




After filtering out high median house values and extreme population values, the distribution plot provides a clearer view of the remaining data. It allows for a more focused analysis of median house values within a specific range.

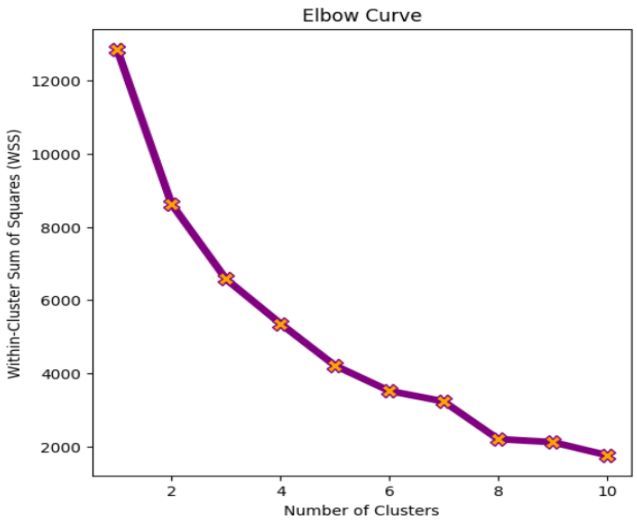


The scatter plot illustrates the geographical distribution of house prices based on latitude and longitude. It enables visualizing how house prices vary across different geographic regions, with warmer colors indicating higher median house values. This visualization helps identify spatial patterns and hotspots of high-value properties.

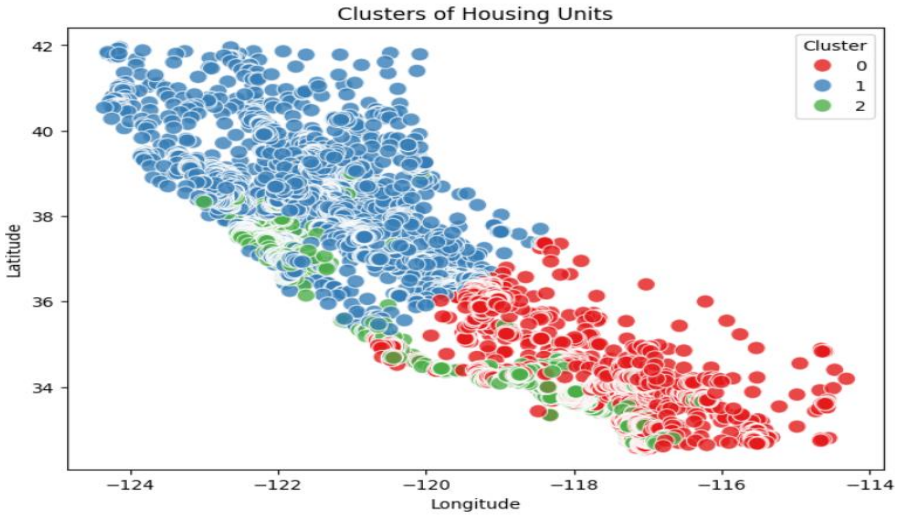


Linear regression models were employed to predict median house values based on various features. The evaluation of model performance through root mean squared error (RMSE) and residual plots provided an assessment of predictive accuracy and insights into model behavior.

.



The elbow curve method helped determine the optimal number of clusters, with three clusters being identified as the most suitable choice. Visualizing these clusters provided insights into the spatial distribution of housing units across different regions.



The visualization of clusters of housing units based on location (longitude and latitude) reveals the spatial distribution of the clusters. Each data point represents a housing unit, and the color of the point indicates the cluster to which the housing unit belongs.