PM 566 Final Project

Author: Karisa Ke

Introduction

The Californian housing market stands as a multifaceted landscape shaped by a myriad of socio-economic and geographical factors. Understanding the intricate dynamics governing property values and development across this region is pivotal, not only for real estate professionals but also for policymakers and individuals seeking insights into housing trends and patterns.

This project embarks on an in-depth exploration into the Californian housing market using a robust dataset sourced from the 1990 census. The primary objective is to unravel the intricate relationships between various factors influencing housing prices, with a particular focus on income levels, geographical coordinates, and housing age distribution.

Methods

In this research project, the primary dataset utilized is the California Housing prices dataset sourced from Kaggle datasets. This dataset encompasses median house prices across various districts in California with other related parameters, derived from the comprehensive 1990 census. The data is downloaded as a zip file and can be unzipped to a csv file. The initial phase of this study involved meticulous data cleaning procedures, which included the identification and handling of missing values (NA values) and the exclusion of potentially anomalous or suspicious data points. This process was guided by the visualization of data distributions, enabling the identification of outliers or irregularities. Following the data preparation stage, a comprehensive exploratory data analysis (EDA) was conducted. This involved employing diverse analytical techniques and visualization tools, leveraging packages such as 'ggplot' and 'corrplot.' Through these tools, an extensive examination of the dataset was carried out, aiming to answer formulated research questions and gain deeper insights into the underlying patterns, correlations, and distributions within the data.

Results

Using str() function to examine the variables in this data set:

1. longitude: A measure of how far west a house is; a higher value is farther west.
2. latitude: A measure of how far north a house is; a higher value is farther north.
3. housing_median_age: Median age of a house within a block; a lower number is a newer building.
4. total_rooms: Total number of rooms within a block.
5. total_bedrooms: Total number of bedrooms within a block.

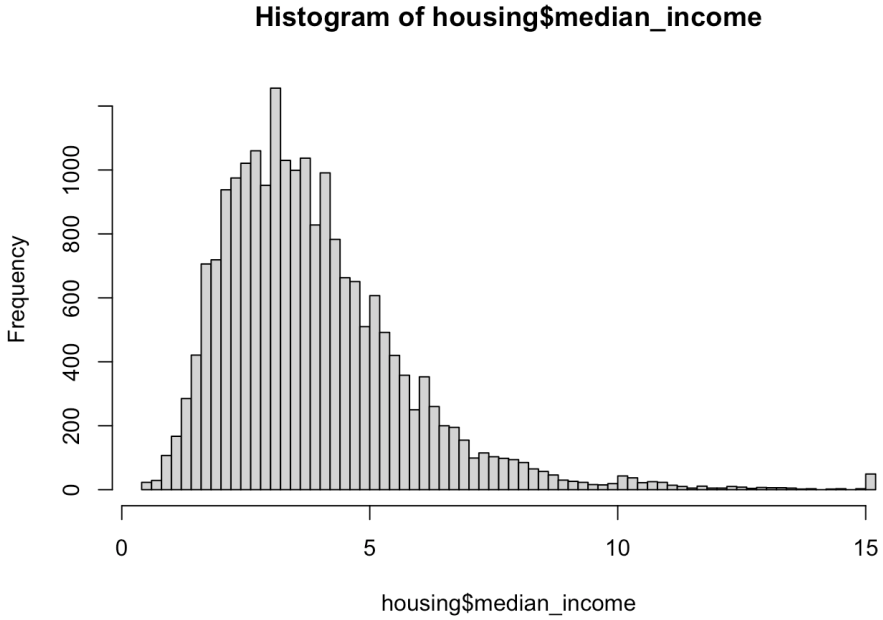
6. population: Total number of people residing within a block.
7. households: Total number of households, a group of people residing within a home unit, for a block.
8. median_income: Median income for households within a block of houses (measured in tens of thousands of US Dollars).
9. median_house_value: Median house value for households within a block (measured in US Dollars).
10. ocean_proximity: Location of the house w.r.t ocean/sea.

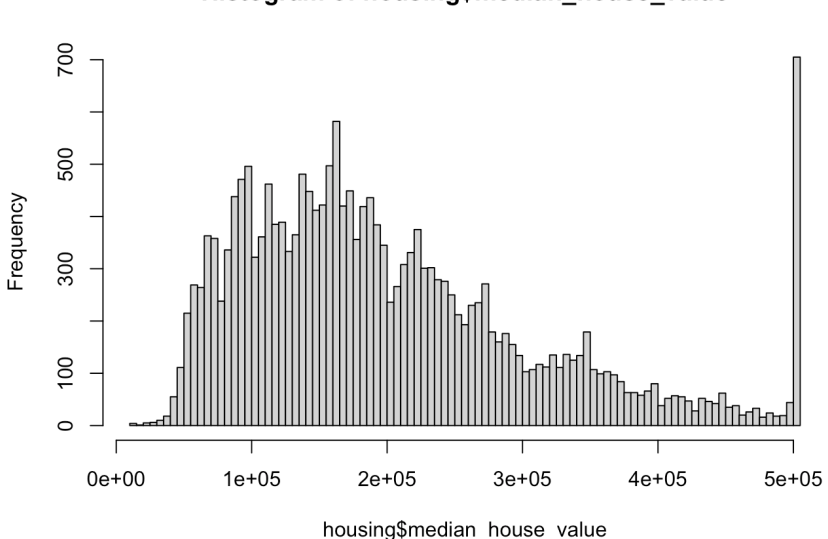
Use the summary() function to examine the basic statistics (min, max, median, quartiles) for each key variable and consider dropping the suspicious values. For example, the summary statistics for median_income is as below:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

 0.4999  2.5634  3.5348  3.8707  4.7432 15.0001
```

Use histograms to visualize the data distribution of variables, for example for median_income, the distribution is right-skewed.





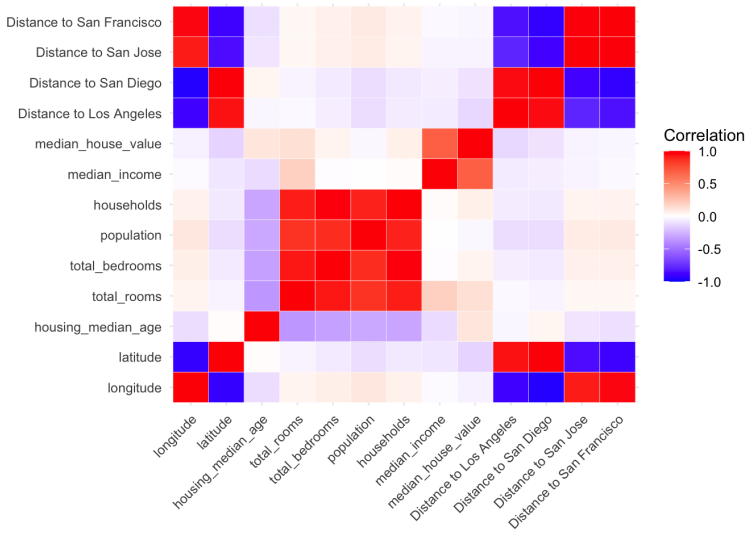
I. Correlation Between Variables

To assess the strength of correlation between the continuous variables within the dataset, an initial analysis was conducted to validate the expected relationships between specific variables. The objective was to confirm anticipated correlations that align with real-world expectations. For instance, an absence of correlation between total_bedrooms and longitude was anticipated, while a strong positive correlation between median_income and median_housing_value was expected.

To facilitate a clear visualization of these correlations, the 'corrplot()' function was utilized, producing a comprehensive correlation matrix. This matrix enabled a detailed examination of the interrelationships between variables. The observations derived from this analysis confirmed the expected patterns: a robust positive correlation was evident between median income and median house value. This finding substantiates the intuitive understanding that higher income levels tend to correspond with more expensive housing prices.

Moreover, the examination of the correlation coefficient between total_bedrooms and longitude revealed a near-zero correlation coefficient, approximately 0. This observation aligns with logical reasoning, as there's typically no inherent relationship between the total number of bedrooms in a district and its geographical longitude.

This preliminary analysis of correlations serves as a foundational step in validating the expected associations between variables and provides initial insights into the dataset's interdependencies. These confirmed correlations establish a basis for further in-depth explorations and subsequent analytical procedures within this research context.

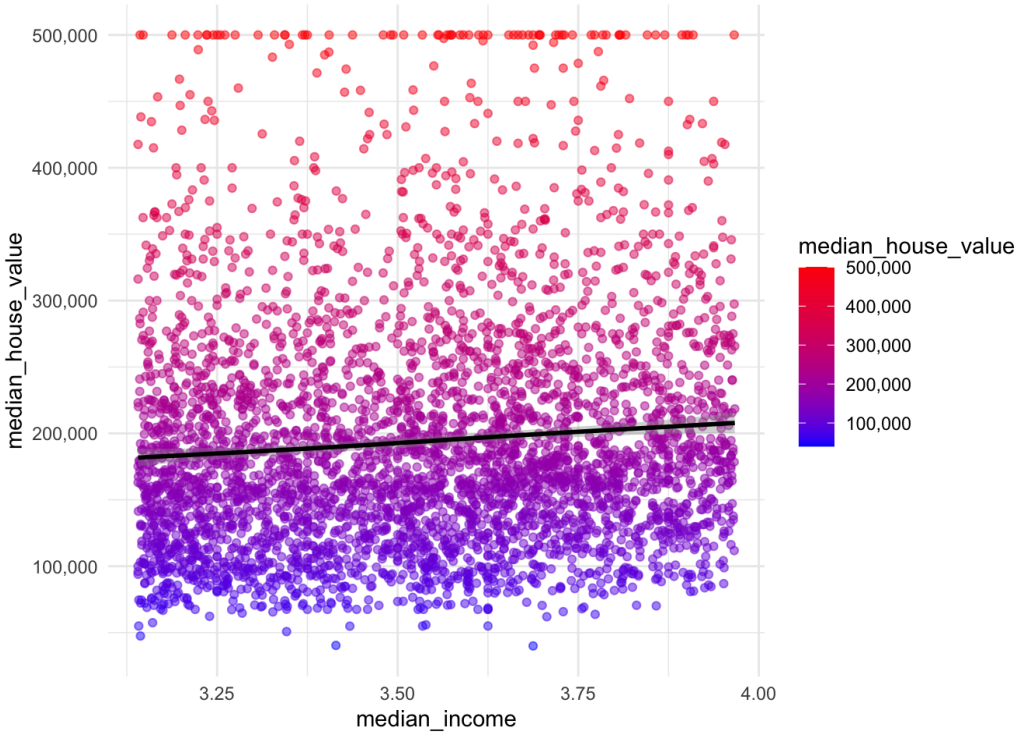


II. House Value for the Medians of Median_Income

To delve deeper into the correlation between median_income and median_house_value, a focused analysis was undertaken to validate whether individuals falling within the 40%-60% percentile range of median_income also exhibit house values within the corresponding median range of median_house_value data.

Isolating the dataset to encompass individuals within the 40%-60% percentile of median_income, which translates to an income range of $30,000 to $40,000 USD, a scatter plot

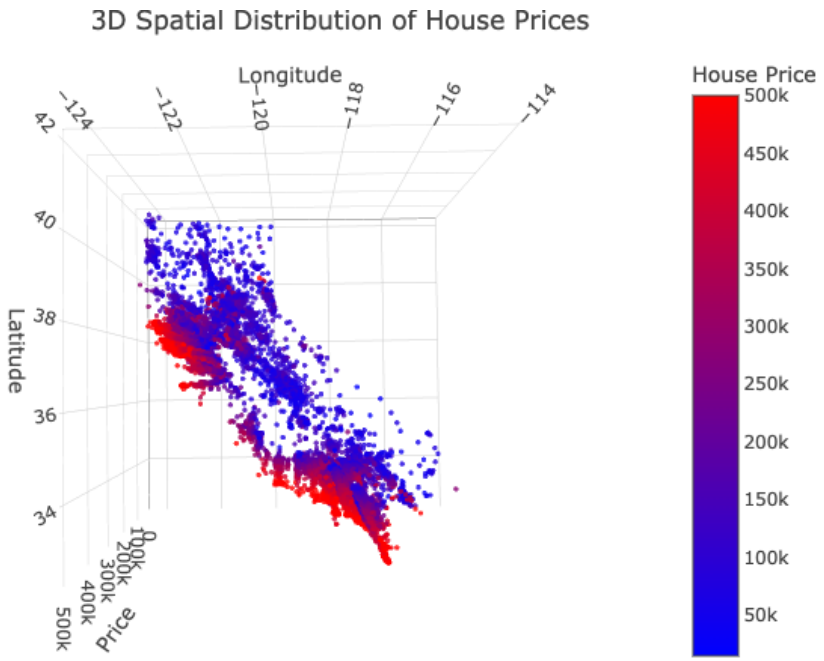
was generated to illustrate the corresponding house values for this specific income bracket. The scatter plot revealed notable variability in house values, attributed partly to the presence of extreme high values within the median_house_value dataset.
However, upon incorporating a trend line into the scatter plot, a discernible pattern emerged. It became apparent that individuals with incomes ranging between $30,000 and $40,000 exhibited house values spanning from a low of $180,000 to a high of $210,000. This observation aligns closely with the calculated 40%-60% percentile range of median_house_value, which spans from $157,300 to $209,400.
The consistency observed between the actual house values of individuals within the specified income percentile and the anticipated range within the median_house_value data further validates the correlation between median_income and median_house_value. This analysis reinforces the notion that individuals within a specific income bracket tend to possess house values that correspond closely to the expected median range within the housing price distribution, as indicated by the percentile calculations.



III. Location and House Value

Creating a three-dimensional projection using longitude, latitude, and median_house_value offers a perspective on how location influences housing prices. This interactive visualization, accessible on the final project website, presents an insightful view of the relationship between geographical coordinates and median_house_value.
From the static snapshot provided, the representation on the map of California illustrates distinct clusters of higher house values primarily centered around major urban centers such as San

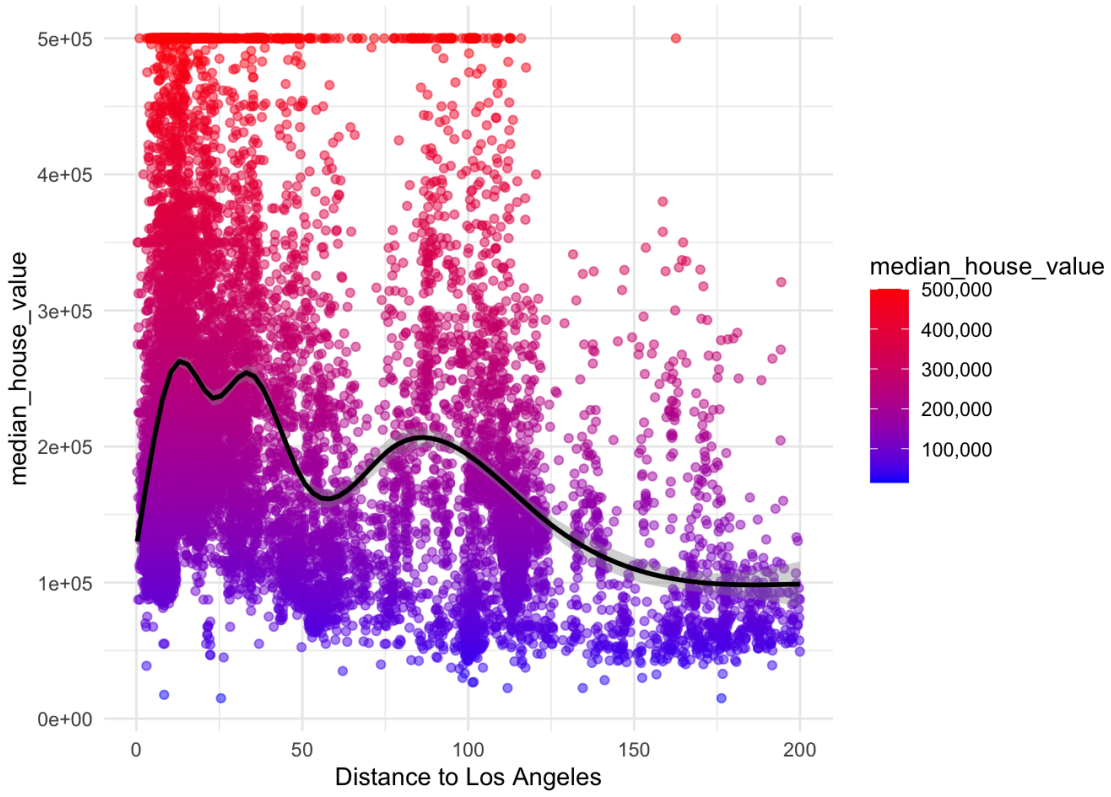
Francisco, Los Angeles, and San Diego. This observation aligns with general expectations, showcasing a concentration of elevated housing prices in metropolitan areas.
The dynamic nature of the interactive model further allows users to explore and navigate through the geographical landscape, gaining an understanding of how specific locations correspond to varying median_house_values. This visualization not only reinforces the influence of location on housing prices but also enables a visual exploration of the spatial distribution of property values across the Californian terrain.



IV. House Values Around Los Angeles Area

Analyzing the variation in house values concerning the proximity to the center of Los Angeles provides valuable insights into the spatial dynamics of property prices in the vicinity of this major urban hub.
Utilizing a scatter plot coupled with a trend line, the visual representation vividly illustrates the fluctuation in house values concerning the distance from the center of Los Angeles. The discernible pattern reveals that the highest house values are predominantly concentrated within a radius of approximately 50 miles from the city center. Within this range, property prices exhibit a notable peak, indicative of the premium associated with proximity to the urban core.
Beyond the 50-mile mark, there is a noticeable and abrupt decrease in property values, signifying a decline as one moves further away from the city center. Interestingly, the trend line depicts a subsequent rise in house values around the 80-mile mark, suggesting another peak in property prices at this distance.
Following this secondary peak, there is a gradual decrease in house values as the distance from the center of Los Angeles increases. This observation highlights a relationship between

distance from the urban center and property values, showcasing peaks at specific intervals, potentially influenced by varying regional factors or suburban developments.
This analysis offers an understanding of how property values fluctuate concerning proximity to the central hub of Los Angeles, unveiling distinct peaks and troughs in housing prices across the surrounding areas.



Note: The distance (in miles) is calculated using the Haversine formula, which is a comment I found under the Kaggle website.

$$
\begin{aligned}
d &= 2r \arcsin\left(\sqrt{\operatorname{hav}(\varphi_2 - \varphi_1) + \cos(\varphi_1)\cos(\varphi_2)\operatorname{hav}(\lambda_2 - \lambda_1)}\right) \\
&= 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)
\end{aligned}
$$

where:

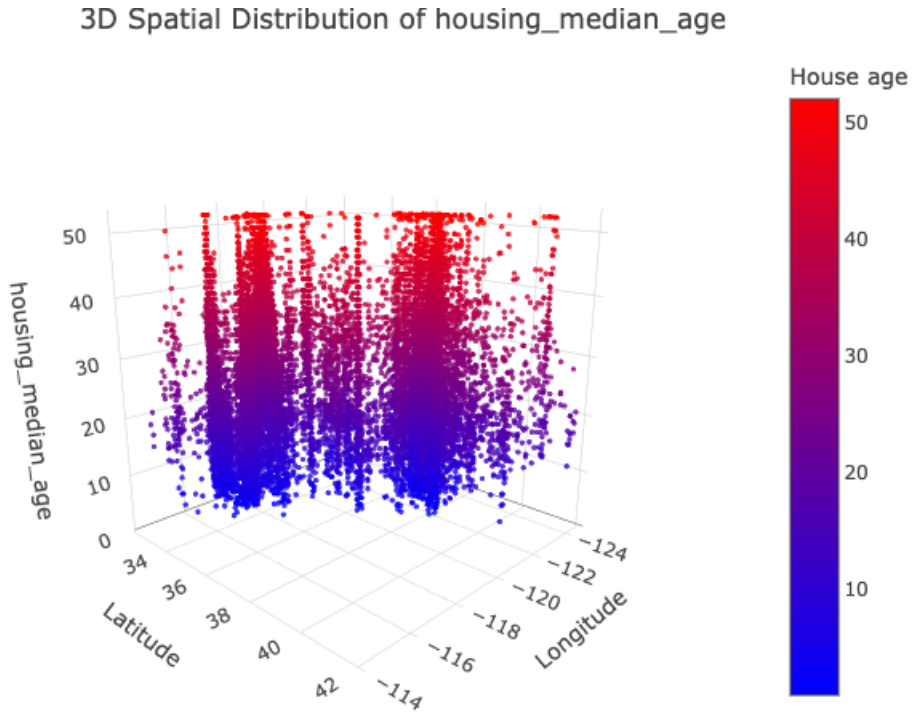
- `phi_1` and `phi_2` are the Latitudes of point 1 and point 2, respectively
- `lambda_1` and `lambda_2` are the Longitudes of point 1 and point 2, respectively
- `r` is the radius of the Earth (6371km)

V. Pattern of Housing Median Age Across California

Exploring the potential relationship between housing_median_age and geographical location through another 3D projection offers additional insights despite the initial indication of no correlation within the correlation matrix.

The visual representation showcases the coexistence of varying housing ages across different geographical coordinates. This uniform distribution of old and new houses indicates that the development and distribution of real estate properties across California are not distinctly influenced by location. Instead, it suggests a more widespread and balanced pattern of housing age distribution throughout the state.

Although the correlation matrix initially indicated no significant relationship between housing_median_age and geographical coordinates, this exploration highlights an interesting finding that showcases the consistent presence of both older and newer housing units across various locations in California. This observation suggests a more evenly spread real estate development pattern throughout the region, irrespective of specific geographical coordinates.



Conclusion and Summary

The comprehensive analysis conducted in this study offers profound insights into the intricate dynamics of California's housing market in 1990, shedding light on the multifaceted relationships between various factors influencing housing prices and development across the region.

The initial exploration validated anticipated correlations, affirming the strong relationship between median_income and median_house_value. Individuals within specific income brackets exhibited house values consistent with expected percentile ranges, solidifying the association between income levels and corresponding property values.

Furthermore, the investigation into geographical influences on housing prices revealed compelling patterns. Proximity to major urban centers, notably Los Angeles, demonstrated a distinct impact on property values, with higher prices clustered within a 50-mile radius of the city center. This analysis underscored the significance of location in shaping housing prices, with discernible peaks and troughs based on distance from urban cores.

Moreover, while the correlation matrix suggested no explicit link between housing_median_age and location, the subsequent exploration uncovered an intriguing uniformity in real estate development. The presence of both older and newer houses dispersed uniformly across geographical coordinates indicated a balanced and consistent pattern of housing age distribution throughout California, independent of specific locations.

In conclusion, this comprehensive analysis delineates the complex interplay between income levels, geographical proximity to urban hubs, and the distribution of housing ages in influencing property values across California. The findings emphasize the critical role of income and location in determining housing prices, while also revealing a surprising uniformity in housing age distribution. This nuanced understanding contributes valuable insights to comprehend the intricate dynamics of the California housing market, providing a foundation for informed decision-making in the real estate domain.