**California Housing Prices Project**

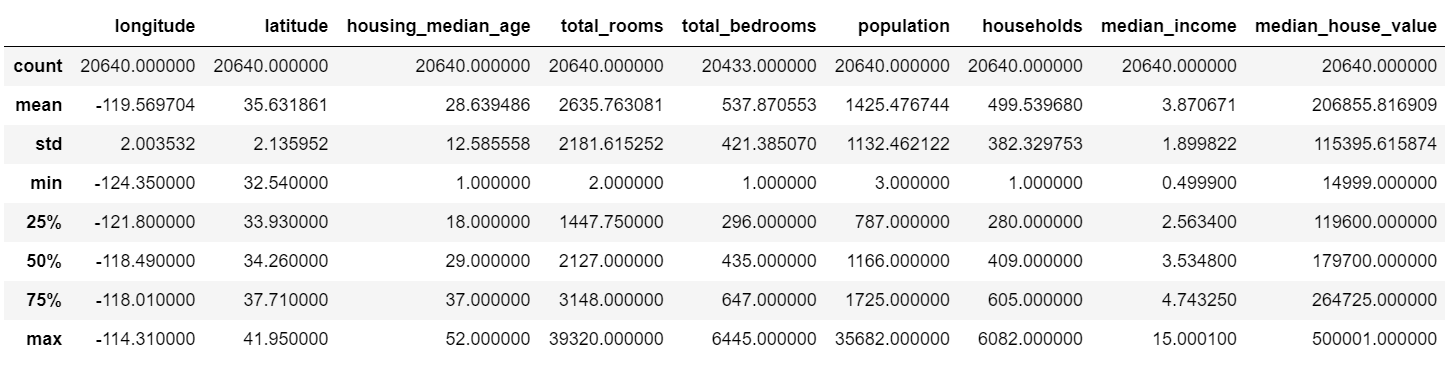
**Introduction**

Understanding the data set is the first step in dealing with it. The data we have primarily represents houses found in California during the 1990 census (Glynnis, 2020). There are ten features with 20,640 observations. Each observation is a single block within California (about 1500 people per block). Nine of these features are independent variables that could impact our dependent variable, which is the house price. Besides ocean proximity, all features are continuous. There are five categories under ocean proximity; those are <1H OCEAN (less than one hour drive to the ocean), INLAND, ISLAND, NEAR BAY, and NEAR OCEAN. The longitude and latitude indicate how far west and north the block is. Moreover, we should know that the median age is counted in years, the median income (10,000 USD) and the median house value in USD (Glynnis, 2020). The total number of bedrooms, total rooms, population, and households reflect the number of people and housing units in each block.

Our project will be divided into two parts. First, we will dive into the data and clean it up, put it in a formatted way, and then try to make the best possible predictions on the house prices, by using machine learning models and creating new features.

**Describing the data**

For columns where there are nine numerical variables, we can produce descriptive statistics which publish the number of entries, the mean value, the standard deviation (spread of the data), and values for various percentiles. This can be done using the pandas library.



According to the number of entries, it is evident that the data for the total number of rooms has some missing values. This is not surprising since usually the real-world data contains noise and missing values, and it is in a format that cannot be directly used for machine learning. Thus, it is necessary to clean the data and prepare it for a machine learning model. This will also increase the accuracy and efficiency of a machine learning model.

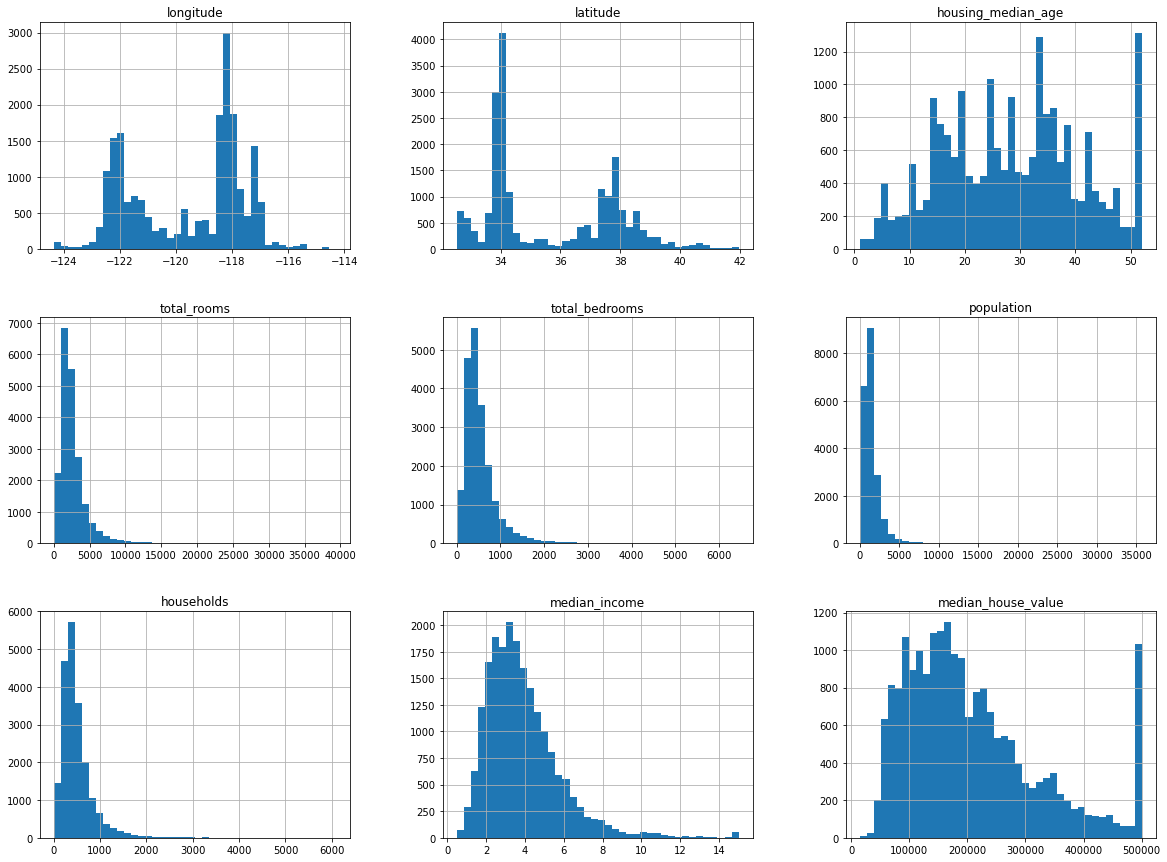
**Missing Values and Duplicates**

We calculated the exact number of missing values in each column. It turned out that there are exactly 207 missing values in the total number of bedrooms. We can easily drop the rows that have them. It is possible to delete the whole column that has those null values; however, this method is only used when the column is mostly NaN. In our case, there are only 207 missing values out of 20640 rows, so this would not be the right choice for us. Imputing missing values using K-nearest neighbours is also an effective way of dealing with missing values, and it is being widely used to replace traditional imputation techniques (Chowdhury, 2020). Traditional imputation meaning that we fill the missing values with the mean or median value or fill with zero or -999 or some other number that does not occur in the data, so that the machine will understand that the data is not genuine (Eddie, 2021). K-nearest neighbours identifies a sample with one or more missing values, then it identifies the K most similar samples among all the training data. Based on a distance metric, the similarity between samples is calculated for this method (GUPTA, 2022). In cases where all the predictors are numerical, Euclidean distance is usually used to determine similarity (GUPTA, 2022). Based on distance computations, we identify the K closest samples to the one with a missing value, and we calculate the average of the predictor of interest. The average value is then used to replace the missing value of the sample. If it is a categorical variable, we can fill it with the mode, but this might occasionally diminish the accuracy value (Eddie, 2021).

After we are done dealing with the missing values there is another check that needs to be done, which is checking for duplicates. In this data set there are no duplicates.

**Frequency Distribution**

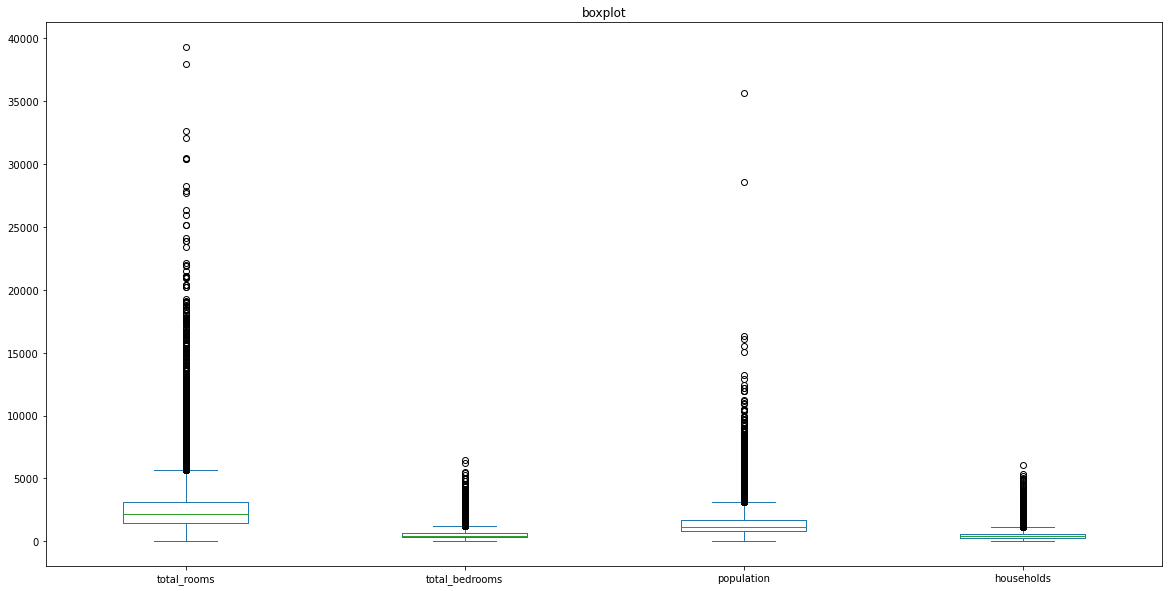
Checking the frequency distribution of the features by plotting their histograms.



The data is positively skewed, meaning that the data is the median is lower than the mean. There is a significant spike in the distribution of median housing prices around $5000000. This indicates that California has a lot of luxuries and expensive houses. For the house median age, there are a lot of local peaks (all are quite gradual) but one odd peak at the maximum value stands out, but we will not do anything with it because we are not going to use the median age in our analysis, we will focus on other factors that have a greater impact on the house price. However, the odd peak in the house median value is considered an outlier and it can skew the mean price, and it could miss our model prediction, so we will drop it. Moreover, there might be some other outliers in other features. We will determine them by plotting a box plot for each feature.

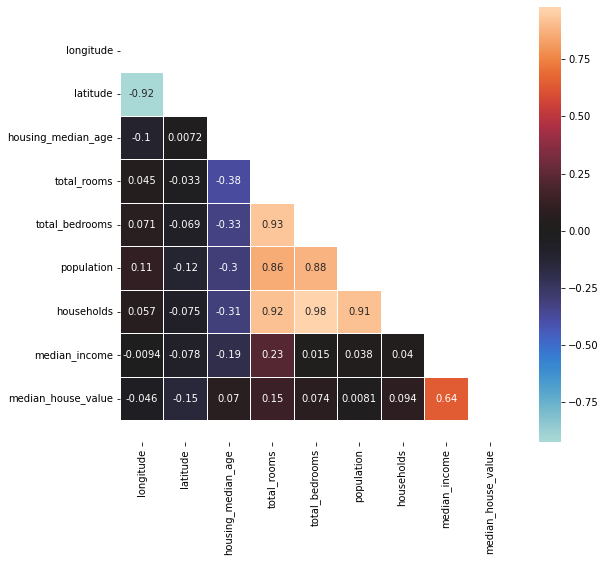
**Outliers**

Producing a boxplot is a great way to check outliers in numerical variables. We noticed that there are so many outliers, we tried to remove all of them at first, but this just decreased the data we have from 20433 to 4123 rows, and it didn’t change the correlation between the variables, and when we thought about it we found that the boxplot just gives us half of the picture here, like we can’t say that those are actually outliers since their values are really close to each other. Therefore, we decided just to remove the extreme outliers, so, we now have 19461 rows.

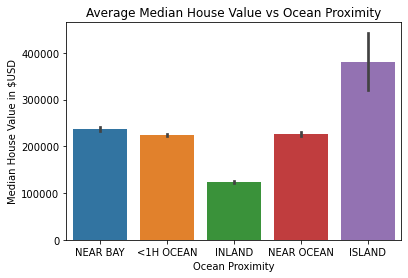


**Correlation**

The way we checked the correlation is by producing a correlation heat map that measures nullity correlation between columns of the dataset. It shows how strongly the presence or absence of one feature affects the other. Nullity correlation ranges from (-1 to 1), -1 means if one attribute is present, the other is certainly absent, zero means there is no dependence between the attributes, and one means if one attribute is present, the other is also certainly present (Wikipedia). We produced the correlation heat map; it looks like the only variable that has a strong relationship with the house value is income.

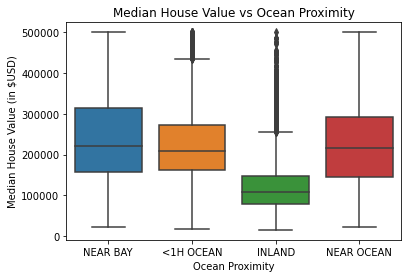


**Ocean proximity**



There are only five houses in the island data. According to the law of small numbers, small random changes have a large apparent effect on the analysis of the data. This is shown by the large error bar of the values of the island data, marked by the black line in the graph. Such a small dataset of five homes means that if even one home was an outlier, then that one outlier would largely affect this whole dataset. For now, we will remove the island data, as this data is unreliable with such a small data set.

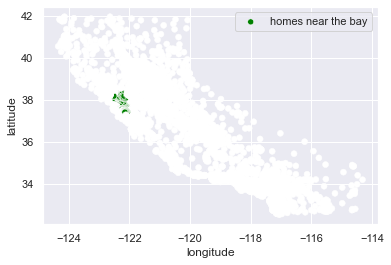
Now, we need to scan for further outliers with the rest of the data using a box plot. It will mark outliers with a dot (since they lie further than 1.5 multiplied by the interquartile range).



Excluding houses in the inland, there are not too many outliers present. We can forgive the inland data for having outliers. This is since there were over 6500 homes, which means there will be plenty of individual homes which could be outliers in this large sample. Furthermore, the regions in California considered 'inland' are vast and quite different, so the house prices will be very varied. For example, you cannot compare the house prices in the built-up urban Sacramento (the capital of California), to a small country town in the Mojave Desert except they are both classified as 'inland'.

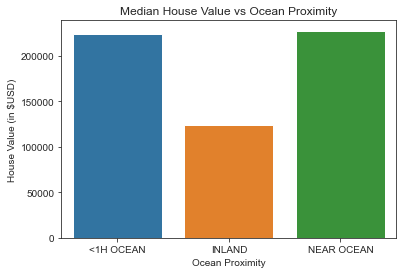
**Analysis of Data**

We see that the homes located near the bay have a higher median housing value in 1990 compared to the homes in any of the other areas (about $9552 more than homes near the ocean, despite both being near the ocean). Now, this makes sense, as any homes which are near the bay are in the San Francisco Bay area.



It is obvious that the homes near the 'bay' are all in a similar location. That location is the San Francisco Bay area. From 2014 to 2020, San Francisco was ranked the most expensive city to live in in the United States (Bote, 2022). Even in 1998 (close to the year the data was collected), a suburb called Atherton in San Francisco was the most expensive zip code in the United States (DePietro, 2018). Therefore San Francisco data was singled out, since its homes would be significantly expensive, compared to the rest of California, and maybe outliers which may skew the data. However, we see that the difference in the median value of homes in San Francisco compared to other homes near the ocean is quite small ($9552), we need to consider that if we adjust for inflation ($1 in 1990 is now $2.21 today ((CPI inflation calculator, n.d.)), we see that the difference in the median house value is $18,750. This does still show that even in 1990, San Francisco had more expensive housing on average than the rest of the state. Factoring how the US also had an interest rate of 8.31% compared to 0.75% today (Macrotrends, 2022), this shows that this difference in average price would cost more in interest then compared to today.

San Francisco is near the ocean, as such homes near the bay are near the ocean. It would be interesting to see whether the general trend for homes near the ocean is more expensive than homes inland excluding the San Francisco data. This would include homes all over the coast not just in one area.



Even in homes not in San Francisco, there is a general trend. As homes get further away from the ocean, the average price of those homes being cheaper. It is clear from the graph that inland homes are significantly less expensive than homes near the coast ($103,000 less than the homes that are near the coast, and $100,000 less on average than homes less than an hour from the coast).

**Possible demand factor; The Climate**

In summer San Francisco (which is near the ocean) has an average temperature of a pleasant 20C (US Department of Commerce [USDOC]). This can be compared to the temperature in the Mojave Desert (the most inland portion of California) in summer, that area is a boiling 36C on average (USDOC, 2015). San Francisco and areas near the ocean in general, have a nicer climate than areas further inland. Therefore, there will be a higher demand for living in areas near the ocean, which will thus drive-up housing prices.

**Possible Supply Factors**

Fewer homes have been built then have been demanded in California's coastal areas. There have been fewer homes being built in California's coastal areas than even in the inland areas of California. This is shown by this graph...

**Fig:**  Annual Growth in Housing Units in California vs Rest of US (Legislative Analyst's Office [LAO], 2015)

Chart, line chart

Description automatically generated

Land Prices are Expensive. The cost to buy land to build housing in California's coastal areas is one of the most expensive in the United States (LAO, 2015). Therefore, there is not much housing built because of the huge cost to buy the land to build it on. Land prices in the inland regions of California, however, are at or below the US average (LAO, 2015)

**Population**

Here we want to determine where most people congregate and how is this related to median house value. We started by plotting a scatter plot, using longitude and latitude and our x and y axis, the population as dots. We can see that the north and centre of California have less population than the south. This can be because California is along the coast, and there are two significant cities around the Bay, those are Los Angeles and San Francisco, both of which have good infrastructures and pleasant weather, so people tend to congregate there.

We expected that with a higher population, there would be a higher demand for homes (with more people needed to live in housing), and as such driving up housing prices (Mulder, 2018). However, is this really the case? We will investigate this, by calculating a test score for the regression line of best fit for median house value over ocean proximity and population.

𝑦̂ = 𝛽0 + 𝛽1x+ 𝛽2x + 𝛽3x + 𝛽4x = 224479.08 - 100311.88x + 12727.78x + 3799.90x - 0.59x

The intercept is $224479.08, which is the average value if the house is <1H Ocean; and the Inland house's value decreases by $100311.88; and the Near Bay house's value increases by $12727.78; and the Near Ocean house's value increases $3799.90 than the <1H Ocean's house; and the model shows that when one population increases, the median house value decreases by $0.59 on average.

A test score of 0.25, shows us that the population has only a weak positive correlation with the median house price based on the knowledge of the home’s ocean proximity. Therefore, the population is not related to why homes near the ocean have a higher house value than homes inland.

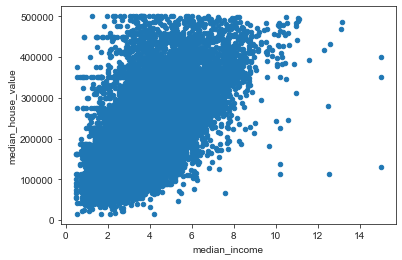
Chart, scatter chart

Description automatically generated

**Median Income**

According to the correlation heatmap, the only variable that has a strong relationship with the house value is income. We can also see a similarity in the shape of their histograms. It is demonstrated that the median income for households within most blocks is around $20000 to $40000, and there is a spike for incomes over $80000. While looking at the median house value histogram, it is between $100000 and $200000

Because both variables are continuous, a scatterplot is the most appropriate plot to have here. So, we used seaborn to make a scatter plot. The scatterplot displays a moderately strong, positive relationship. As median income increases, the median house value also tends to increase. This demonstrates that families with higher incomes tend to purchase more expensive houses.

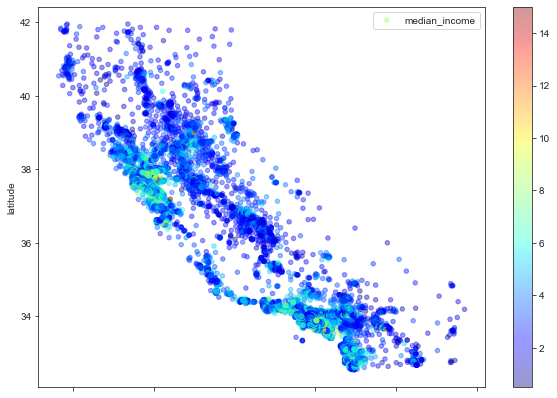


To investigate the relationship between these two factors in more depth, we constructed a regression model. A regression model is a statistical model (a supervised machine learning approach) that examines the linear relationship between two or more variables - a dependent variable and an independent variable. Therefore, the regression model is used when we want to predict a continuous dependent variable from one or more independent variables. Here we want to predict the median house value from the median income.

The regression equation is: 𝑦̂ = 45591.9+ 39918.96𝑥.

Here 45591.9 is the intercept, which is the average value when a block’s family’s median income is 0 (Although it does not make sense); and for every $10000 increase in the income, the median house value increases by $39918.96.

The training score and the testing score are both 0.41



The graph above shows the relationship between the location and the families’ median income. The median income decreases as the homes are away from the ocean. And we mentioned how housing prices decrease as you move further inland, which supports the positive relationship between income and house value.

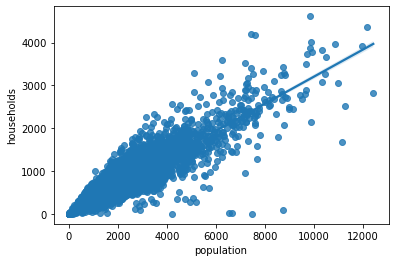
**Total Rooms, Total Bedrooms and Number of households**

We wanted to dig deeper into the data, so we checked if the number of rooms and bedrooms affects the house value, but basically, the total number of rooms in a particular district is not much related to the house price if the population is not specified. The more relevant variable for the house price is the number of people in a house. Therefore, created four new attributes population per household, room per household, bedroom per household, and bedroom per room to make sure that the information about rooms, bedrooms and population are in the same “unit” as our target.

**In California, Do People Prefer to Live Alone or With Family Members?**

To answer the question, we used a reggplot and a regression model, to produce a graph depicting the relationship between households and the population, and we calculated the correlation coefficient between them using a correlation heatmap. As shown in the graph, there is a strong positive relationship with a coefficient of 0.91 between the population and the number of households, therefore, as the population grows, so does the number of households. The training and testing scores are quite close (0.831, 0.825)

But that is not enough evidence to answer the question that people prefer to live alone or with their families in the big cities like Los Angeles or San Francisco. The data on households and total bedrooms is then used. The correlation coefficient between them is 0.97 when we use the regression model indicates a strong positive relationship between households and population. The training and testing scores are quite close (0.956, 0.954).

Chart, scatter chart

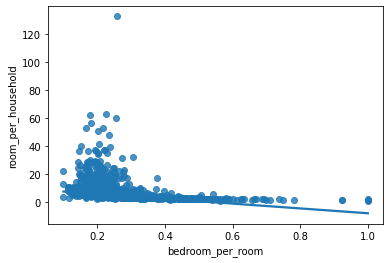
Description automatically generated

It is possible to argue that in densely populated areas such as Los Angeles and San Francisco, most individuals prefer to live with their family or share a home with someone else. One of the causes is the extraordinarily high cost of living in large cities such as Los Angeles and San Francisco. California is the second-least affordable state, behind Hawaii. (The Affordability Rankings of U.S. News & World Report)

Besides the price of housing in California, as it is called the Golden State, California offers several good options to live in, including some of the world's greatest universities, such as Stanford University, the California Institute of Technology, and the University of California. Since many significant-tech companies' headquarters are in California, such as Apple, H, and Facebook, which offers more high-paying jobs after graduation. Silicon Valley in San Francisco is a global centre of technology innovation. The area was named after the key component of computer microprocessors. Hundreds of important technology, software, and internet companies call Silicon Valley home. Not only are there prospects in the technological industry, but also film industry, as Hollywood is located in Los Angeles. More demand, but limited land supply, is driving California's real estate market insane. Many parents have chosen to settle in California because they want their children to have greater future opportunities. Other factors make California an attractive place to settle, besides the abundant opportunities for the future, such as their climates, their landscapes, and their diversity in culture.

**How large is the House in California?**

The link between bedroom per room and room per household is used to calculate the house size in California. According to the heatmap, there is a negative link between them, with a coefficient of -0.43. We may deduce that as the population grows, the room size shrinks. It is possible to infer that people in densely populated places prefer to live in apartments. It helps in space saving. Renting in LA is 35 percent less expensive than buying a home (Livabl, BuzzBuzzHome, 2021). We can conclude that people in California prefer to live in apartments for a temporary time, with the primary reason being to work in the city or to study.



**Conclusion**

The cost of housing in California is influenced by several factors. Based on the relationships that we found that houses are more expensive on average as they are closer to the ocean in California, the median income is a significant factor in this development, but the population is not. There are other supply factors including the difference in climate between two areas, education and job opportunities, unique cultures, and entertainment activities. From this data set we can tell how critical pre-processing the data is, and we know how some outliers can mess up our models. Pandas, Seaborn, and Matplotlib libraries with the linear regression are great tools to analyse the data.

**Reference List**

Foley Glynnis. (2020, May 19). Exploratory Data Analysis of the California Housing Market. RStudio. <https://rstudio-pubs-static.s3.amazonaws.com/617841_020b1c3834334c1c8c3e0ec67645775b.html>

ANKIT GUPTA. (2022, April 15). Advance Data Pre-processing. Kaggle. <https://www.kaggle.com/code/nkitgupta/advance-data-preprocessing>

[Kaushik Roy Chowdhury](https://www.analyticsvidhya.com/blog/author/kaushikrch/). (2020, July 14). KNNImputer: A robust way to impute missing values (using Scikit-Learn). Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2020/07/knnimputer-a-robust-way-to-impute-missing-values-using-scikit-learn/>

Eddie. (2021, May 19). Dealing With Missing Values in Python – A Complete Guide. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2021/05/dealing-with-missing-values-in-python-a-complete-guide/>

Wikipedia. (2022 April 20). Correlation. Wikipedia. <https://en.wikipedia.org/wiki/Correlation>

CPI inflation calculator. (n.d.). *$1 in 1990 → 2022 | Inflation Calculator*. Www.in2013dollars.com. Retrieved May 21, 2022, from   <https://www.in2013dollars.com/us/inflation/1990?amount=1#:~:text=Value%20of%20%241%20from%201990>

Effectivioology. (2014). *The Law of Small Numbers: Overestimating the Representativeness of Small Samples – Effectiviology*. Effectiviology. https://effectiviology.com/law-of-small-numbers/

Mulder, C. (2018). *Sixty-ninth session of the UNECE Committee on Housing and Land Management Keynote Presentation The relationship between population and housing*. [https://unece.org/fileadmin/DAM/hlm/archive/Key%20note%20population%20and%20housing.pdf](https://www.in2013dollars.com/us/inflation/1990?amount=1#:~:text=Value%20of%20%241%20from%201990)

Legislative Analyst's Office. (2015, March 17). *California’s High Housing Costs: Causes and Consequences*. Ca.gov. [https://lao.ca.gov/reports/2015/finance/housing-costs/housing-costs.aspx](https://www.in2013dollars.com/us/inflation/1990?amount=1#:~:text=Value%20of%20%241%20from%201990)

DePietro, A. (2018, July 31). *Housing 1998-2018: America’s Most Expensive Zip Codes, Then and Now*. Forbes. [https://www.forbes.com/sites/andrewdepietro/2018/07/31/housing-1998-2018-most-expensive-zip-codes/?sh=7a49f9e41aea](https://www.in2013dollars.com/us/inflation/1990?amount=1#:~:text=Value%20of%20%241%20from%201990)

Chappelow, J. (2019, September 29). *Law of Supply and Demand*. Investopedia. [https://www.investopedia.com/terms/l/law-of-supply-demand.asp#:~:text=The%20law%20of%20demand%20says](https://www.in2013dollars.com/us/inflation/1990?amount=1#:~:text=Value%20of%20%241%20from%201990)

admin. (2020, November 29). *Factors Affecting Housing Market Supply and Demand in Australia*. SuburbsFinder. [https://www.suburbsfinder.com.au/resources/factors-affecting-housing-market-supply-and-demand-in-australia/#:~:text=An%20increase%20in%20income%20means](https://www.in2013dollars.com/us/inflation/1990?amount=1#:~:text=Value%20of%20%241%20from%201990)

Bote, J. (2022, March 2). *SF dethroned as most unaffordable housing market in America*. SFGATE. [https://www.sfgate.com/realestate/article/San-Francisco-not-most-unaffordable-city-16971811.php](https://www.in2013dollars.com/us/inflation/1990?amount=1#:~:text=Value%20of%20%241%20from%201990)

US Department of Commerce, N. (2022, May 21). *Climate*. Www.weather.gov. [https://www.weather.gov/wrh/Climate?wfo=mtr](https://www.in2013dollars.com/us/inflation/1990?amount=1#:~:text=Value%20of%20%241%20from%201990)

Macrotrends. (2022). *Federal Funds Rate - 62 Year Historical Chart*. Macrotrends.net. <https://www.macrotrends.net/2015/fed-funds-rate-historical-chart>

(2021, December). Cost of Living in California. Sofi. <https://www.sofi.com/cost-of-living-in-california/#:~:text=Average%20Cost%20of%20Living%20in%20California%3A%20%2446%2C636%20per%20year&text=According%20to%202020%20data%20from,living%20in%20California%20is%20%2446%2C636>.

TROY SEGAL. (2022, March 15). Silicon Valley. Investopedia

<https://www.investopedia.com/terms/s/siliconvalley.asp#:~:text=4-,What%20Is%20 Silicon%20Valley%20 Famous%20For%3F,of%20the%20world's%20richest%20people>.

Kelsey Pudloski. (2021, Mar 2). Renting is nearly 35% cheaper than owning a home in LA: Realtor.com. Livable. <https://www.livabl.com/2021/03/renting-cheaper-than-owning-la-2021.html>