Insight into Modelling Consumer Preferences

Introduction

This project aims to scrutinize the discourse inside the world of data analysis using Python based tools to understand and possibly answer questions related to the housing market of California. At the end of our analysis, we should be able to answer questions that can help a prospective buyer to provide his preferences and constraints and get answers to help him decide on buying his/her dream house.

Housing prices is an intriguing yet familiar topic which is touched upon time and again in an attempt to grasp and derive the relationship between the mechanism of pricing houses and factors affecting these very prices. Often, the results drawn from this analysis turn out to be skewed and non-representative of the original population. With the aid of the dataset provided below named 'California Housing Prices in 1990', we wish to re-analyse the data provided within the dataframe with a macro level view of perspective and depict the resulting correlation.

Our objective is to ascertain how various independent variables such as the Age of the House, Ocean Proximity, and Average Occupancy(rooms per person) steer a consumer’s decision to adjust his budget into purchasing a house. We wish to do the same by visualizing each of our handpicked independent as well as dependent variables and plotting them against each other using meaningful visualization techniques and graphical tools to derive relationships that can help us in achieving our objective. We will also try to recreate them on a 2D map using various python libraries in order to understand the variables on a broader geographic scale, widening our scope of analysis.

I) Describing the Variables:

The rationale behind picking the independent variables we have chosen is associated with the research question we wish to answer at the end of our analysis. Since we would like to establish the dynamics of consumer budget with respect to the Price of a house, we need to study how prices vary given the Age of the House, Rooms Per Person, Ocean Proximity and Average Income of the population living in a particular region.

Each of the variables stated above perform the following roles:

- Ocean Proximity describes how far the house is from the nearest water body and can take upto 5 values. This variable gains importance due to the fact that the general notion associated with the reality market is that houses near to the ocean are relatively more expensive than the ones further away. We want to understand this and verify what kind of a relationship the proximity to the ocean has with the price of a house. A person having budget constraints may not want to consider houses near the ocean in case of a positive correlation.

- **Age of the House** refers to the median age of houses within that location that are bought by consumers. Age of the House can be a deciding factor for a house buyer in case his/her preferences for younger houses play a major role in his final decision. We want to understand if older houses demand a higher price compared to the newer houses or is it vice versa. Alternatively, we also would like to know if they are not related at all. If there is a budget constraint then the knowledge that switching to houses based on age can help find a more suitable property, it would help the user a lot.
- **Average occupancy(rooms per person)** is a derived variable which is calculated by dividing the total rooms with the households and displays the average occupancy. It indicates how many houses are available in a particular location given its longitude and latitude. It is a crucial variable because in the absence of the total houses data, there is no way to calculate the actual average rooms per person since we don't have a clue as to the number of rooms shown in the dataset belonging to how many houses. A higher value suggests more houses are available in that location.
- **Average Income** of a region was chosen as a variable in order to ascertain whether systemmic differences in average household income between different counties made any difference to house prices. An intuitive guess would be that a richer neighbourhood will boast of higher property prices and vice versa. Hence it becomes a very important factor to decide which counties fit into the user budget and which do not provided there is a positive correlation between them.

II) Literature Review:

We went through a series of research papers that delved into similar topics of either predicting the house prices or finding the relationship of dependent factors with the price of a house. We could not find a specific literature that has touched upon the relationship of the factors provided in our given dataset and evaluated via a linear regression, a model analysing relationship between these factors and the price of a house. In this context, hence my analysis takes importance as it clearly demonstrates the importance of the factors (independent variables) such as the Age of the House, Proximity to the Ocean Front etc. and helps a prospective to consider these factors as part of his/her budget calculations.

In *Truong, Q. D., Nguyen, M. T., Dang, H., & Mei, B. (2020). Housing Price Prediction via Improved Machine Learning Techniques*, the authors reiterate our hypothesis that the very commonly used factor House Price Index (HPI) to estimate the changes in housing price is incorrect since housing price is strongly correlated to other factors such as location, area, population, it requires information apart from HPI to predict individual housing price. But their work is based on data for Beijing and not California and hence our research becomes pertinent in this context.

In *Gallatin, N., Hill, D. (2022, October 9). California Housing Markets: A Tale Twice Told*, the author aims to examine the current health of the California residential real estate market through an analysis that is threefold. In the first portion of our analysis, they try to observe the relationship between home price and real estate specific measures that are widely considered predictive of home price appreciation. These five variables are: the traditional Housing Affordability Index (HAI), Unsold Inventory Index (UII), Median Time on Market (TOM), Building Permits, and Ten-Year Treasury Note Yield. They employ the traditional HAI from the California Association of Realtors (CAR), which is calculated as the proportion of people that can afford a median priced home. This paper does not consider the 5 variables that my work has built upon, hence the importance of the factors I have considered are ignored in this research paper.

In *Zixu Wu. Prediction of California House Price Based on Multiple Linear Regression. Academic Journal of Engineering and Technology Science (2020) Vol. 3*, the author although affirms that the multivariate linear regression deployed in my research and analysis of the California is the correct approach to attach the given research question, the data the paper uses seems to be of Boston notwithstanding the title of the paper that says California. Also, it uses only a few of the factors considered by me namely the number of rooms. Hence the value of my research and findings is different and probably more valuable than that presented by the author.

III) Summary Findings:

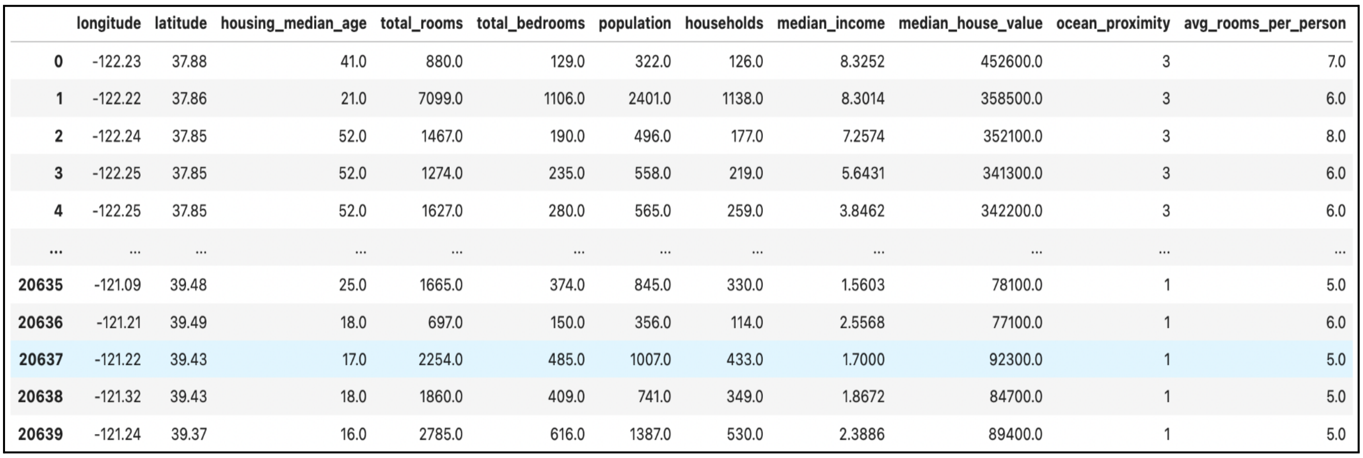
Our main goal is to analyse and deduce notable facts and figures which were not so evident to the public eye by viewing the numerous rows of data presented in the California housing market dataset in order to understand their effects on consumers purchasing behaviour. Below are some of the most noteworthy that will help a buyer to use our analysis in order to narrow down on a prospective house.

- When we considered the effect of **Ocean Proximity** on the pricing of the house, the frequency of houses found under the categories of 'Inland' and 'Near the Ocean' far exceeded those of the other proximities provided, thus indicating a higher demand for houses along the coast among consumers.
- Upon a careful analysis of the **Income distribution** across counties when considering consumers individually as well as clubbing them into income groups led us to believe that Income is positively skewed against the backdrop of Price. Similiarly, we notice a negatively skewed graph when comparing the frequency of consumers against each income group.
- **Age of the House** may have a considerably weaker impact on the consumer's decision to purchase a home but it does have a significant relationship with Ocean Proximity. A majority of the consumers prefer to live in older houses either along the coastline or inland.
- The **Average Occupancy** values provided across different coordinates suggests that there is a mix of a surplus and shortage of houses depending upon the location due to an unstable and inconsistent demand and supply of houses.

Finally, in order to build up on our research and provide a detailed as well as effective approach to implement our analysis, we have looked beyond and found many factors that might influence a buyer's buying mindset but are not available in the current dataset. Many of these factors are available on different web portals and need to be examined first to analyse if they are meaningful in our context and then figure out how we can get the relevant data off of the web portal. One of the ways to do this is web scraping. But in order to do so, the data should be in a format that can easily be scraped off. It's possible that the data that is present, is relevant but it's impossible to scrape it off and use in adjacency to the given dataset. In our analysis below, we have successfully scraped off data from a wikipedia webpage and aligned it to our existing dataset which has produced significant results. Furthermore, we have integrated Regression and Machine Learning to our analysis by creating various models which can enable us in using various factors and variables provided within the dataset to examine the effects of a change on our dependent variable by altering our independent variables. Let’s analyse all of these one by one.

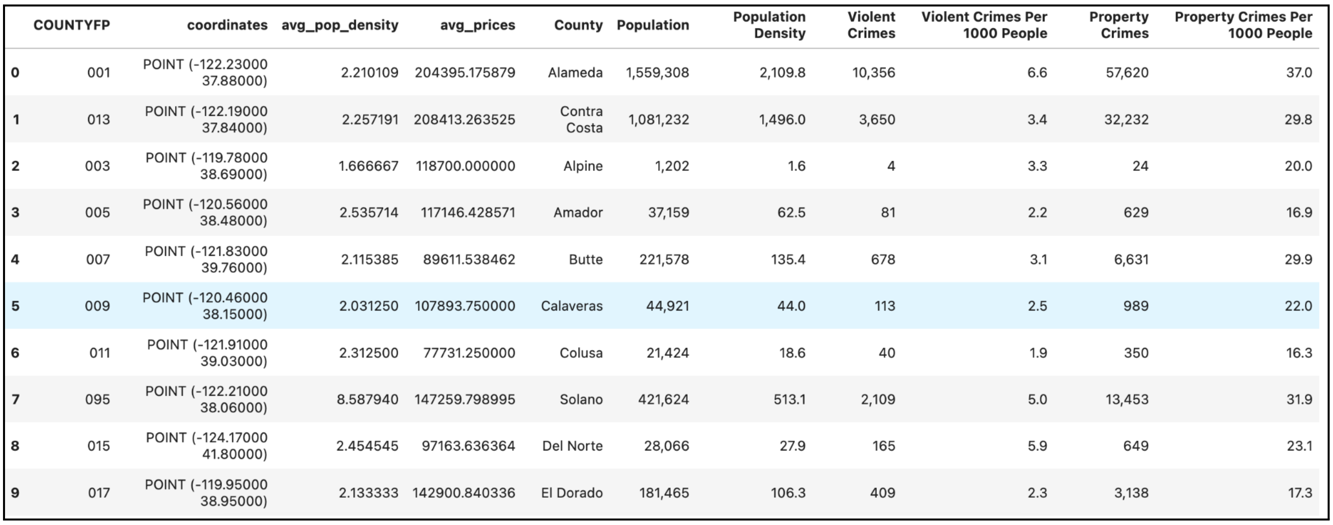
Data

Our original dataset with which we started off our analysis was picked up from one of kaggle’s many dataframes and consisted of 20,640 observations in total where each row or observation represented a particular set of coordinates, i.e., latitudes and longitudes. These coordinates were locations of individual homes at different locations spread across the state. We were also provided variables such as Median Income, Median House Value, Age of the House, Ocean Proximity, Total Rooms and Bedrooms, Population and Households. This data was in its raw and unstructured form. Our first task required us to clean it up and filter it to our needs. This included changing data types of certain variables, creating dummies, replacing missing values, and finally creating new variables using the existing ones. The table below shows the new and filtered dataset we obtained after data cleaning.



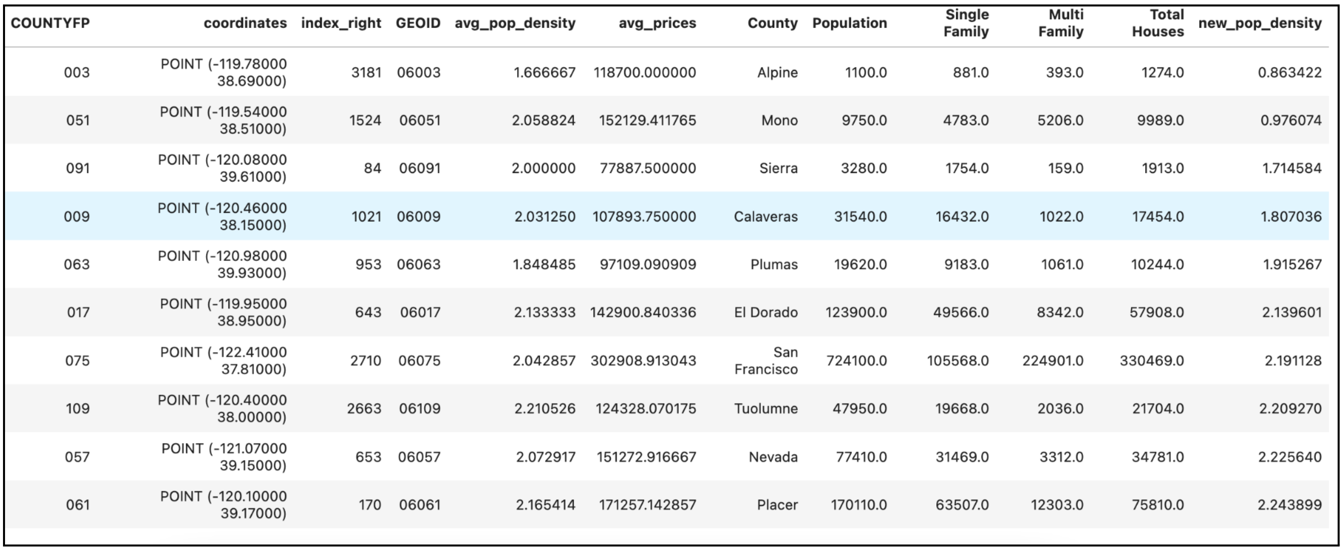
	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity	avg_rooms_per_person
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	3	7.0
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	3	6.0
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	3	8.0
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	3	6.0
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	3	6.0
...	...	...	...	...	...	...	...	...	...	...	...
20635	-121.09	39.48	25.0	1665.0	374.0	845.0	330.0	1.5603	78100.0	1	5.0
20636	-121.21	39.49	18.0	697.0	150.0	356.0	114.0	2.5568	77100.0	1	6.0
20637	-121.22	39.43	17.0	2254.0	485.0	1007.0	433.0	1.7000	92300.0	1	5.0
20638	-121.32	39.43	18.0	1860.0	409.0	741.0	349.0	1.8672	84700.0	1	5.0
20639	-121.24	39.37	16.0	2785.0	616.0	1387.0	530.0	2.3886	89400.0	1	5.0

Most of our graphical analysis was done using this cleaned and filtered dataset from interpreting relationships between Income and House Value to mapping the trend of average price levels in each county. We gathered more valuable data by either web-scraping or scrounging for additional datasets available publicly. For instance, we used California's government official web portal to access data for all public and private educational institutions present in the state. Similarly, we used wikipedia to gather information about crime rates prevailing within the state on a county wide scale. Merging this data helped us in creating another new dataset which is shown below consisting of additional variables such as violent crimes, property crimes, crimes per 1000 people etc. The dataset below shows the first 10 observations of the entire dataframe.



	COUNTYFP	coordinates	avg_pop_density	avg_prices	County	Population	Population Density	Violent Crimes	Violent Crimes Per 1000 People	Property Crimes	Property Crimes Per 1000 People
0	001	POINT (-122.23000 37.88000)	2.210109	204395.175879	Alameda	1,559,308	2,109.8	10,356	6.6	57,620	37.0
1	013	POINT (-122.19000 37.84000)	2.257191	208413.263525	Contra Costa	1,081,232	1,496.0	3,650	3.4	32,232	29.8
2	003	POINT (-119.78000 38.69000)	1.666667	118700.000000	Alpine	1,202	1.6	4	3.3	24	20.0
3	005	POINT (-120.56000 38.48000)	2.535714	117146.428571	Amador	37,159	62.5	81	2.2	629	16.9
4	007	POINT (-121.83000 39.76000)	2.115385	89611.538462	Butte	221,578	135.4	678	3.1	6,631	29.9
5	009	POINT (-120.46000 38.15000)	2.031250	107893.750000	Calaveras	44,921	44.0	113	2.5	989	22.0
6	011	POINT (-121.91000 39.03000)	2.312500	77731.250000	Colusa	21,424	18.6	40	1.9	350	16.3
7	095	POINT (-122.21000 38.06000)	8.587940	147259.798995	Solano	421,624	513.1	2,109	5.0	13,453	31.9
8	015	POINT (-124.17000 41.80000)	2.454545	97163.636364	Del Norte	28,066	27.9	165	5.9	649	23.1
9	017	POINT (-119.95000 38.95000)	2.133333	142900.840336	El Dorado	181,465	106.3	409	2.3	3,138	17.3

Quite similarly, we were successful at finding 2 more publicly available datasets online, the first one being the population dataset for the state of California for the years 1970 all the way to 2018 and the second dataset consisting of the number of houses sold to various types of families in California for a specified time interval. By combining these two new datasets with our original filtered one, we were able to find significant differences between the self inferred population density variable of ours and the new population density provided in the new dataframe. Further, we could interpret and graphically analyse these differences.

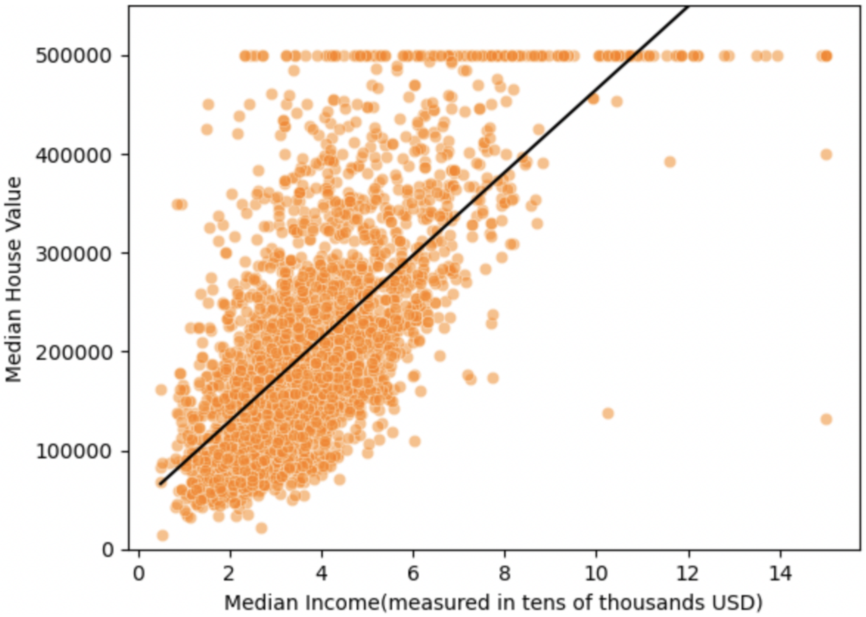


COUNTYFP	coordinates	index_right	GEOID	avg_pop_density	avg_prices	County	Population	Single Family	Multi Family	Total Houses	new_pop_density
003	POINT (-119.78000 38.69000)	3181	06003	1.666667	118700.000000	Alpine	1100.0	881.0	393.0	1274.0	0.863422
051	POINT (-119.54000 38.51000)	1524	06051	2.058824	152129.411765	Mono	9750.0	4783.0	5206.0	9989.0	0.976074
091	POINT (-120.08000 39.61000)	84	06091	2.000000	77887.500000	Sierra	3280.0	1754.0	159.0	1913.0	1.714584
009	POINT (-120.46000 38.15000)	1021	06009	2.031250	107893.750000	Calaveras	31540.0	16432.0	1022.0	17454.0	1.807036
063	POINT (-120.98000 39.93000)	953	06063	1.848485	97109.090909	Plumas	19620.0	9183.0	1061.0	10244.0	1.915267
017	POINT (-119.95000 38.95000)	643	06017	2.133333	142900.840336	El Dorado	123900.0	49566.0	8342.0	57908.0	2.139601
075	POINT (-122.41000 37.81000)	2710	06075	2.042857	302908.913043	San Francisco	724100.0	105568.0	224901.0	330469.0	2.191128
109	POINT (-120.40000 38.00000)	2663	06109	2.210526	124328.070175	Tuolumne	47950.0	19668.0	2036.0	21704.0	2.209270
057	POINT (-121.07000 39.15000)	653	06057	2.072917	151272.916667	Nevada	77410.0	31469.0	3312.0	34781.0	2.225640
061	POINT (-120.10000 39.17000)	170	06061	2.165414	171257.142857	Placer	170110.0	63507.0	12303.0	75810.0	2.243899

Visualization

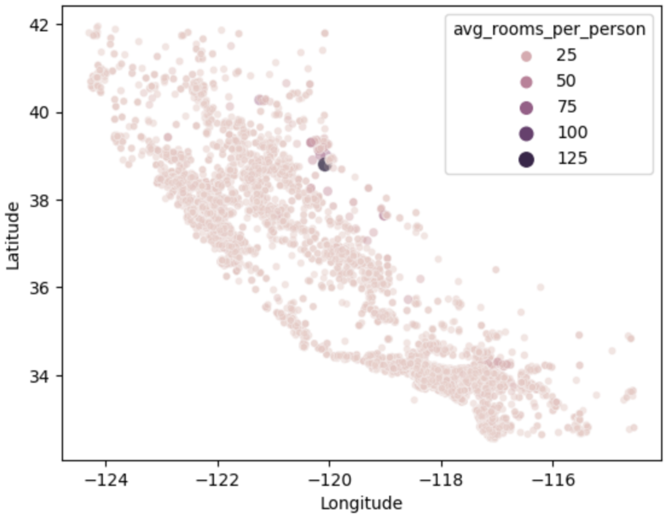
Once acquiring all the data that we described above, we moved onto analysing it. The most effective way to communicate your findings and main message is in the form of graphs and maps. Thus, for each important analysis that pertained to answering our original query stated earlier, we represented its results using a variety of graphical tools.

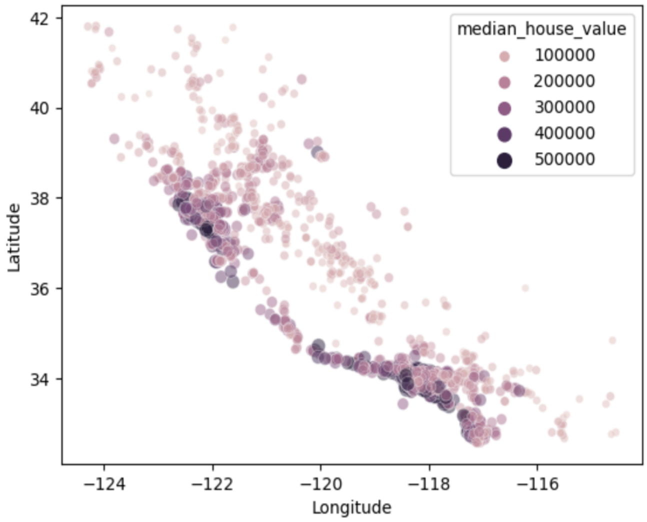
For instance, we were able to successfully establish that there indeed existed a linear relationship between Median Income and Average House Value. The figure below depicts a scatterplot showcasing the exact relationship that we described above. We can observe that the scatter transitions upwards, rising from the bottom left to the upper right indicating a positive and linear relationship as well. Another way to confirm the above result is to draw a straight line through the scatter, if the resulting points lie around the line we can conclude a positive and linear relationship.



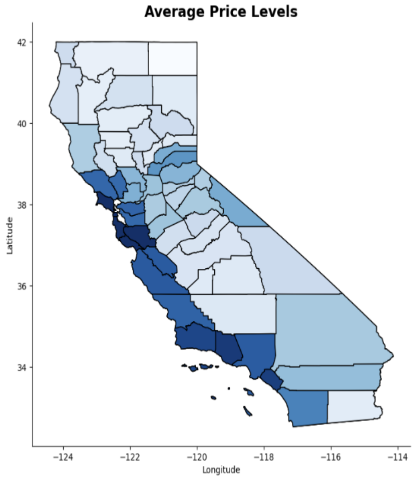
Quite similarly, we were also able to interpret the relationship between some of our other crucial variables such as Average Rooms Per Person on Location and Location on Average House Value. The graphs below depict the same. The first scatter plot depicts a strong linear and negative relationship between Location and Average Rooms Per Person. It was unfortunately unable to identify areas with a higher Average Rooms Per Person statistic, and thus we assumed that a normal family may require a maximum of 2-3 rooms per person (we are considering the whole house which does not just include bedrooms only), implying all numbers higher than 3 are indicative of the fact that there are houses available in that region.

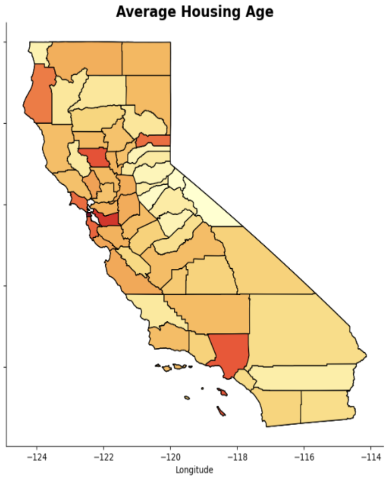
Our second scatterplot below shows how prices of houses change from locality to locality measured by the longitude and latitude of a specific block in the state of California. We were able to deduce a negative linear relationship from the graph below as the scatter extends from the top right of the figure all the way down to the bottom right. We were able to locate places within California with different price levels; places such as San Diego, Rockland, Los Angeles, San Francisco will usually have houses priced at the higher end while those found in small towns like El Monte, Imperial County would be cheaper.

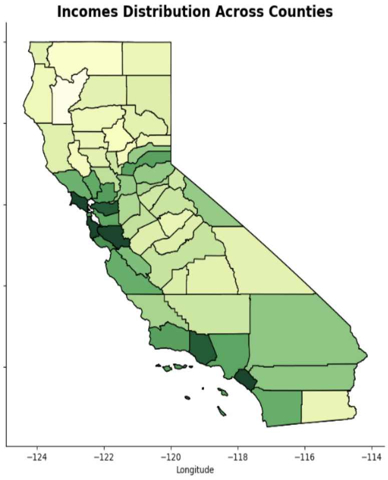




Another key result we derived using our mapping analysis was the relationship between factors such as Housing Age, Income Distribution and Average House Values based on their values in different counties across the state of California. In the plot below, we used a color coordinated scheme to highlight significant differences in the above factors. We can clearly see that all counties along the coast are darker in color in both maps 1 as well as 3, implying that these counties are not only well off but expensive as well. Intuitively, if we think about it, higher Income of a consumer raises his/her purchasing power, which enables him to buy higher priced houses. Thus, counties with higher incomes should have more pricey houses as well as shown below. The Housing Age map indicates that a similiar relationship with price exists as well, where houses found or built near the ocean as well as some places inland are much more old and aged increasing their worth and value, thereby prices being higher for those counties.



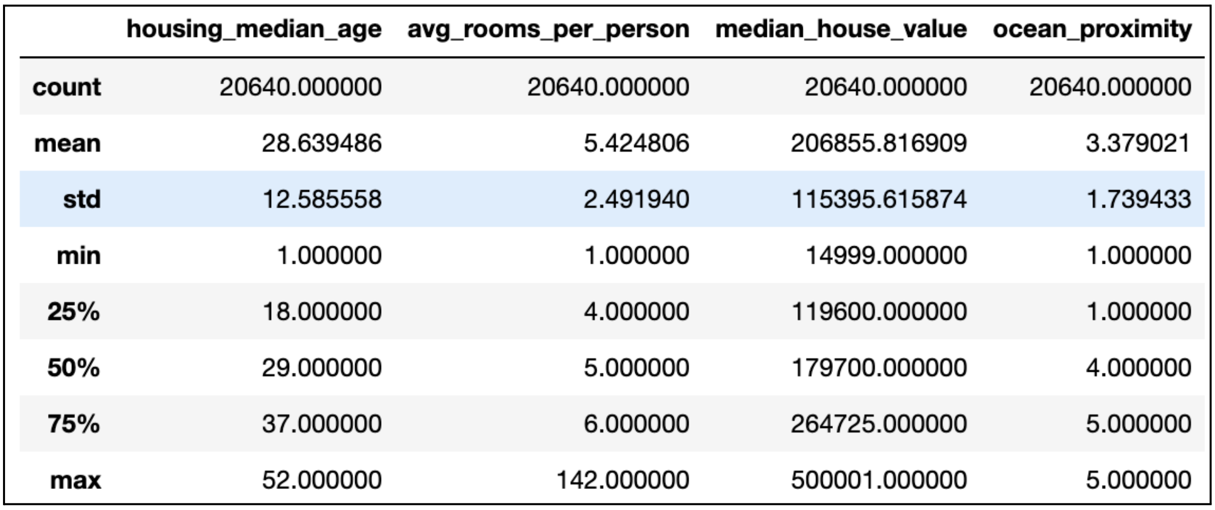




Summary Statistics

In our initial findings, we mainly used 4 different variables or factors that could potentially influence or alter decisions made by consumers to purchase a house. These were mainly the Age of the House, Average Rooms Per Person, Price of the House and Ocean Proximity. We have already discussed and described their role in our analysis earlier during the introduction. Given below is the table that briefly summarises all the above variables with respect to their approximately 20,600 observations in our dataset. Some of the key findings we made during our analysis were as follows:

- The mean or average age of houses provided within our dataframe is between 28-29 years which signifies that a majority of the houses are quite old. Similarly, the mean Average Rooms Per Person indicating how many houses are available in a particular location was roughly around 2 which indicated that many houses were empty for purchase across the state.
- We also observed that 68% of the houses in the dataset have an age which ranges from 16 years all the way to 40 years. This improved our understanding of the houses in California from an age perspective and indicated that houses both new and old gave a choice to the housing market buyer to cater to his preferences of purchasing both types.
- The lower percentile for Housing Age indicates that 25% or less of the total houses available are less than 18 years old but 75% of the houses are either 37 years old or fairly newer with half of the total houses available being 29 years older or less.
- Half of the total number of the houses are available at a value of 179,000 USD or below with only 25% houses being more expensive than 264,000 USD. Infact, 25% or less of the total houses lie below 119,000 USD. We could infer that affordable housing is certainly available in California.
- The housing age variable varies from 1 to 52 years showing that California provides a prospective home buyer a wide variety of options to choose from ranging from almost brand new houses to some that are almost a half a century old. This allows buyers to prioritize their buying preferences.



	housing_median_age	avg_rooms_per_person	median_house_value	ocean_proximity
count	20640.000000	20640.000000	20640.000000	20640.000000
mean	28.639486	5.424806	206855.816909	3.379021
std	12.585558	2.491940	115395.615874	1.739433
min	1.000000	1.000000	14999.000000	1.000000
25%	18.000000	4.000000	119600.000000	1.000000
50%	29.000000	5.000000	179700.000000	4.000000
75%	37.000000	6.000000	264725.000000	5.000000
max	52.000000	142.000000	500001.000000	5.000000

Regression Results

In our analysis, we created 5 different regression models for different variables as can be seen in the regression table below to primarily examine the effect of changes in explanatory variables on the dependent variable. Our dependent variable is the Average House Value for all 5 regressions and we have used a suite of over 4 explanatory variables in different ways to regress upon the Price Level of Houses. These are the Age of the House, Ocean Proximity, Median Income and Population Density. All 5 regressions can be summarised in the following way:

I) Regression Model 1:

The first regression we created was a multiple regression consisting of Income and Ocean Proximity as explanatory variables and Average House Value as the dependent variable. This model could also be written as ***HouseValue* = β_0 + β_1*MedianIncome* + β_2*OceanProx* + ε** where **β_1** and **β_2** represent slope coefficients of each **x variable**. **β_0** is the intercept of linear trend line determining the base price of a house which over here has no interpretation since we would have to set all x variables as 0 making it improbable to do so. **β_1** represents the slope coefficient of Income and it can be interpreted as the marginal effect of a 10,000 USD increase in Median Income on the Average House Value after controlling for Ocean Proximity is a rise in the price of a house by 38,000 USD, lastly **β_2** represents the slope coefficient of Ocean Proximity which is a dummy variable and can take values from 1-5. We can interpret this as we move towards the coast or ocean by increasing the value of this variable, the Average House Value increases by approximately 18,000 USD after controlling for Median Income. **ε** is simply the random error term which depicts all the deviations of observations from the trend line due to factors not included in the model but affecting the Average House Value.

Also, we can observe that all the above values are significant, that is, they are both economically and statistically significant at a conventional 99% confidence level. The **R^2** signifies that 55% of the proportion of variation in the Price of Houses is explained by the variation in Median Income and Ocean Proximity. To test whether the overall model is significant, **R^2** is not the sole tool to rely on. Even though it is 55%, we need to consider the F test statistic which represents the overall statistical significance of the model. We can see that the F test, which is a big value, is highly statistically significant at a 99% confidence level as the P-value lies in the interval 0-0.01. This means that this model does help in providing an effective prediction of the price of a house given a certain Income and distance of the house from the ocean.

II) Regression Model 2:

The second regression we created was another multiple regression consisting of Median Income and Population Density as explanatory variables and Average House Value as the dependent variable. This model could also be written as ***HouseValue = β_0 + β_1MedianIncome + β_2PopDensity + ε*** where **β_1** and **β_2** represent slope coefficients of each **x variable**. **β_0** is the intercept of linear trend line which cannot be interpreted as the base/starting price of a house since it would require us to set all x variables to 0 which would be logically impossible to do so as a Population Density of 0 would mean a family with 0 members which is highly unlikely. **β_1** represents the slope coefficient of Income of the consumer and it can be interpreted as the average effect of a 10,000 USD increase in Income of a consumer will be a roughly 42,000 USD rise in the price of the house after controlling for Population Density, and **β_2** which represents the slope coefficient of Population Density can be interpreted as on average the price of a house decreases by approximately 310 USD after an additional family member is added to the household by controlling for Median Income. Intuitively, we know that adding another member to the house usually requires more space which means the size of the house would need to increase causing the prices to rise as well. However, since this happens to a multiple regression model, we cannot solely rely on this negative association between these variables and will need to construct a correlation matrix to find whether there really exists a negative relationship or not. **ε** is simply the random error term which depicts all the deviations of observations from the trend line due to factors not included in the model but affecting the Average House Value.

R^2 stands at an impressive 49% describing the fact that about 49% of the proportion of variation in the price of houses is explained by the variation in Income of a consumer and Population Density. The F test on the other hand produces an overall statistical significance at a 99% confidence level with a value of 1421 implying that the P-value will be in the range of 0-0.01. This means that this model does help in providing an effective prediction of the price of a house given a certain Income and Population Density of a household.

III) Regression Model 3:

The third regression we created was a model estimate consisting of Median Income and Age of the House as explanatory variables and Average House Value as the dependent variable. This model could also be written as ***HouseValue = β_0 + β_1Age + β_2MedianIncome + ε*** where **β_1** and **β_2** represent slope coefficients of each **x variable**. **β_0** which is the intercept of the OLS line cannot be interpreted since to interpret it would mean to hold all x variables 0 such that the Income of a consumer will be 0 USD which is highly unlikely. **β_1** which represents the slope coefficient of the Age of the House can be interpreted as that on average as the Age of the House becomes older by each year, the price of the house increases by about 1600 USD after controlling for Median Income, and lastly **β_2** representing the slope coefficient of the Median Income of a consumer can be interpreted as the average effect of a 10,000 USD increase in Income of a consumer will be a roughly 43,000 USD rise in the price of the house after controlling for Age of the House. **ε**

is simply the random error term which depicts all the deviations of observations from the trend line due to factors not included in the model but affecting the Average House Value.

R^2 stands strong at 51% which means that about 51% of the proportion of variation in the price of houses is explained by the variation in the Age of the House as well as the Income of a consumer. The F test on the same hand produces an overall statistical significance at a 99% confidence level with a value of 1583. A high F test statistic suggests that the regression value is also statistically significant at a 0.1% significance level and further indicates that the following model fits our data accurately and will enable us in predicting the Average House Value effectively.

IV) Regression Model 4:

The fourth regression we created was another model estimate consisting of Population Density, Median Income and Age of the House as explanatory variables while the Average House Value was considered as the dependent variable. This model could also be written as ***$HouseValue = \beta_0 + \beta_1 PopDensity + \beta_2 MedianIncome + \beta_3 Age + \varepsilon$*** where **$\beta_1$**, **$\beta_2$**, and **$\beta_3$** represent slope coefficients of each **x variable**.

β_0 which is the intercept of the OLS line cannot be interpreted since to do that would mean substituting all x variables with 0. It is impossible to imagine any consumer not earning any Income and wanting to purchase a house or a family having a 0 Population Density which is why it would be improbable to make an interpretation. **β_1** which represents the slope coefficient of Population Density can be interpreted as with each additional family member introduced in the house, the price of the house will decrease by approximately 337 USD after controlling for Income and Age of the House. However, an intuitive thought might remind us that more family members in a house translates to more rooms which means that the family would require a bigger house to accomodate more people causing the price to increase, so there seems to be a contradiction. We must remember though that since we are in a multiple regression model, each of these variables do not independently influence the Average House Value. So, we cannot take this negative relationship result to word and would require a correlation matrix to further enhance our analysis.

β_2 represents the slope coefficient of Median Income and can be interpreted the same way as we did before, i.e, with a 10,000 USD increase in Income of a given consumer or family, the Price of the House will increase on average by roughly 43000 USD after controlling for Age of the House and Population Density which is quite intuitive again since house is a luxury good with a higher income elasticity, which means that as the Incomes rise the ability or demand of a consumer to buy more expensive goods increase as well. Lastly, **β_3** represents the slope coefficient of the Age of the House which can be interpreted as on average as the Age of the House becomes older by each year, the price of the house increases by about 1600 USD after controlling for Median Income and Population Density. **ε** is simply the random error term which depicts all the deviations of observations from the trend line due to factors not included in the model but affecting the Average House Value.

R^2 stands at a good 52% which means that about 52% of the proportion of variation in the price of houses are explained by the variation in Median Income, Population Density and Age of the House. The F test produces an overall statistical significance at a 99% confidence level with a value of roughly 1074. A high F test statistic suggests that the regression value is also statistically significant at a 0.1% significance level and further indicates that the following model fits our data accurately and will enable us in predicting the Average House Value effectively.

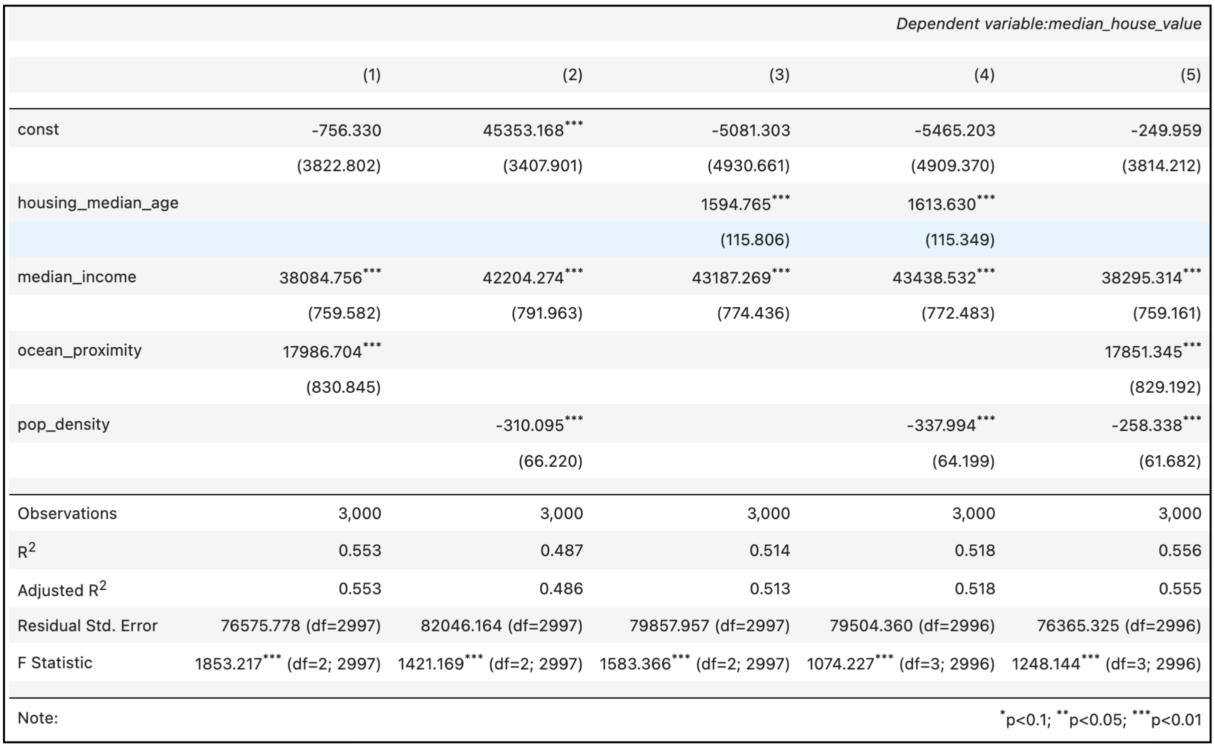
V) Regression Model 5:

The final regression model of ours consists of Median Income, Population Density and Ocean Proximity as explanatory variables while the Average House Value is considered as the dependent variable. This model could also be written as ***HouseValue = β_0 + β_1PopDensity + β_2 MedianIncome + β_3OceanProx + ε*** where **β_1**, **β_2**, and **β_3** represent slope coefficients of each **x variable**.

β_0 which is the intercept of the trend line cannot be interpreted since to do that would mean substituting all x variables with 0. We cannot possibly assume that any household or consumers will have a 0 Income or that any house will have no family members residing within it. **β_1** represents the slope coefficient of Population Density and can be interpreted as with each additional family member introduced in the house, the price of the house will decrease by approximately 258 USD after controlling for Income and Ocean Proximity. As we explained earlier, intuitively one might think that more family members in a house translates to more rooms which means that the family would require a bigger house to accomodate more people causing the price to increase. However, we know that since we are in a multiple regression model, each of these variables do not independently influence the Average House Value and so we would require a correlation matrix to make an effective conclusion.

β_2 which represents the slope coefficient of Median Income can be interpreted here the same way we did earlier, which is the average effect of a 10,000 USD increase in Income of a consumer will be a 38000 USD rise in the Average Price Level of a house after controlling for Ocean Proximity and Population Density. Lastly, **β_3** which represents the slope coefficient of Ocean Proximity which is a dummy variable can be interpreted as we move closer to ocean/coast by increasing the value of the dummy variable, the price of the house increases by approximately 18,000 USD after controlling for Income and Population Density. This makes intuitive sense since houses on the coast are more expensive to the ones located Inland.

The **R^2** signifies that about 57% of the proportion of variation in the price of houses are explained by the variation in Median Income, Population Density and Ocean Proximity. To test whether the overall model is significant, we need to consult the F test statistic as well. We can see that the F test which is about 1248 is highly statistically significant at a 99% confidence level implying that the P-value lies in the interval 0-0.01. This means that this model does help in providing an effective prediction of the price of a house given a certain Income, Population Density and proximity of the household to the ocean.



					Dependent variable:median_house_value
	(1)	(2)	(3)	(4)	(5)
const	-756.330	45353.168***	-5081.303	-5465.203	-249.959
	(3822.802)	(3407.901)	(4930.661)	(4909.370)	(3814.212)
housing_median_age			1594.765***	1613.630***	
			(115.806)	(115.349)	
median_income	38084.756***	42204.274***	43187.269***	43438.532***	38295.314***
	(759.582)	(791.963)	(774.436)	(772.483)	(759.161)
ocean_proximity	17986.704***				17851.345***
	(830.845)				(829.192)
pop_density		-310.095***		-337.994***	-258.338***
		(66.220)		(64.199)	(61.682)
Observations	3,000	3,000	3,000	3,000	3,000
R^2	0.553	0.487	0.514	0.518	0.556
Adjusted R^2	0.553	0.486	0.513	0.518	0.555
Residual Std. Error	76575.778 (df=2997)	82046.164 (df=2997)	79857.957 (df=2997)	79504.360 (df=2996)	76365.325 (df=2996)
F Statistic	1853.217*** (df=2; 2997)	1421.169*** (df=2; 2997)	1583.366*** (df=2; 2997)	1074.227*** (df=3; 2996)	1248.144*** (df=3; 2996)
Note:					*p<0.1; **p<0.05; ***p<0.01

Conclusion

The aim of our research cum analysis was twofold viz:

- Finding out which factors affect the price level of a house in the state of California the most and further may influence the decisions of consumers in purchasing a house. Our study was spread across all the counties in the state. These factors were selected from the various possible variables available in the sample dataset provided. We will call them the **Physical Factors**.
- In addition, the study also tried to find other important factors that a prospective house owner could consider that could influence his/her decision to buy a house in a certain area. These factors were not part of the sample dataset and were found externally by analysing literature, journals etc. We will call these **Social Factors**.

Physical Factors: We found 4 factors that could contribute to the house prices that I would like to group them under physical factors. These are Proximity to the Ocean, Age of the House, Average occupancy in an area and Average Income of a particular region. It was observed that more than 50% of the houses are less than 1 hour away from the ocean with their median price range being greater than 180,000

USD. It was evident that there are only a very few island houses but probably being very few they demand a high price of in excess of 300,000 USD.

Over half the houses in the state of California were built almost 30 - 35 years back. It tells us that construction of new houses is not something that is happening and people are contending to live in older houses. They may renovate, undertake repairs to suit their living standard or personal preferences but not many new constructions have been taking place. Another factor that plays an important role is the Average Income of the community where the property is located.

The study realized the hypothesis we started with was in fact true and that the Average Income has a direct positive correlation with the average price of a house in a certain area. The higher the Average Income of the neighbourhood was, the higher the price of the house vis-a-vis its counterpart in a slightly less richer neighbourhood. This also makes a lot of commercial sense as rich neighbourhoods typically have higher per square foot price, have bigger and more glamorous properties and therefore demand a much higher premium compared to some of their lesser cousins in other communities.

Social Factors: Additionally, we have analysed and scrutinized two other factors that could enable a buyer to examine prospective houses and neighbourhoods in a more objective manner. These are the Crime Rates in a county and the number of schools in a locality. We find that these two factors have direct correlation with the house pricing. People prefer to choose their house in a relatively less crime infested area with good school districts close by showing a positive correlation.

The study engaged in understanding the relationship between the narrowed down impact factors and the house price by using different mathematical and statistical tools such as histograms, bar charts, scatter plots and overlay of data onto physical state map of California, Ordinary Least Squares linear regression methodology and finally using Machine Learning regression tree algorithms to re-iterate the relationships. Notwithstanding the method of analysis, each method kept building on top of the previous technique thereby reimposing the faith on our conclusion.

Everything though was not perfect in our study and there is a lot of scope for improvement in future. To start with, the dataset is well short of data that can more objectively help to find out important factors to consider when looking to buy a house in California. Factors such as Crime Rate, Schooling Data, Public Transport Facilities, Distance from Office Downtown District, Ethinicity are some of the ones that immediately come to one’s mind. Moreover, some of the data even though present could not be interpreted since it lacked the depth and was dependent on certain other data which was blatantly missing from the dataset.

References

1) *Truong, Q. D., Nguyen, M. T., Dang, H., & Mei, B. (2020). Housing Price Prediction via Improved Machine Learning Techniques*

2) *Gallatin, N., Hill, D. (2022, October 9). California Housing Markets: A Tale Twice Told*
3) *Zixu Wu. Prediction of California House Price Based on Multiple Linear Regression. Academic Journal of Engineering and Technology Science (2020) Vol. 3*