Summary

Housing market in different cities in the US fluctuates a lot. Housing price is not always related to a single factor. In this project, we uncovered patterns on housing prices in different California cities. We looked on how the city population, income, poverty, and different other factors that we get from census API like education, transportation affects the housing price. And from all these different factors, we nominated the best 5 cities to live in.

1. Understanding the Barriers of Affordable Housing

a. How Housing Prices related to other factors?

At first, we want to uncover patterns on housing prices in different California cities and want to see how the housing price is related to other factors. For this, we looked on the most important parameters for the year 2019 which are as follows:

i. Economy

Fig. 1 shows the housing price in each city with population size. This figure gives us an idea to see how housing value varies in each city.

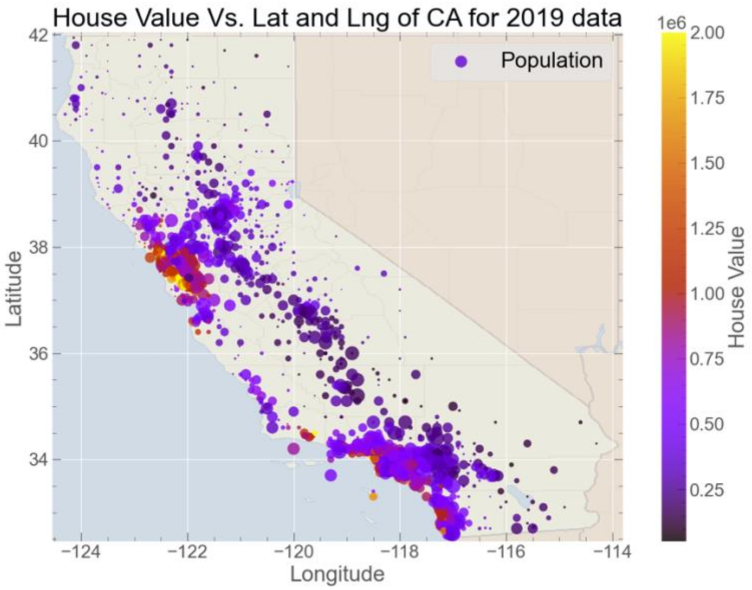


Fig. 1: Latitude and Longitude plot of housing price for each cit

We looked how the household income, monthly rent, poverty rate, and monthly owner cost effects the housing value. We saw that there is positive relation with household income, monthly owner cost, and monthly rent whereas a negative relationship with poverty as shown in Fig. 2.

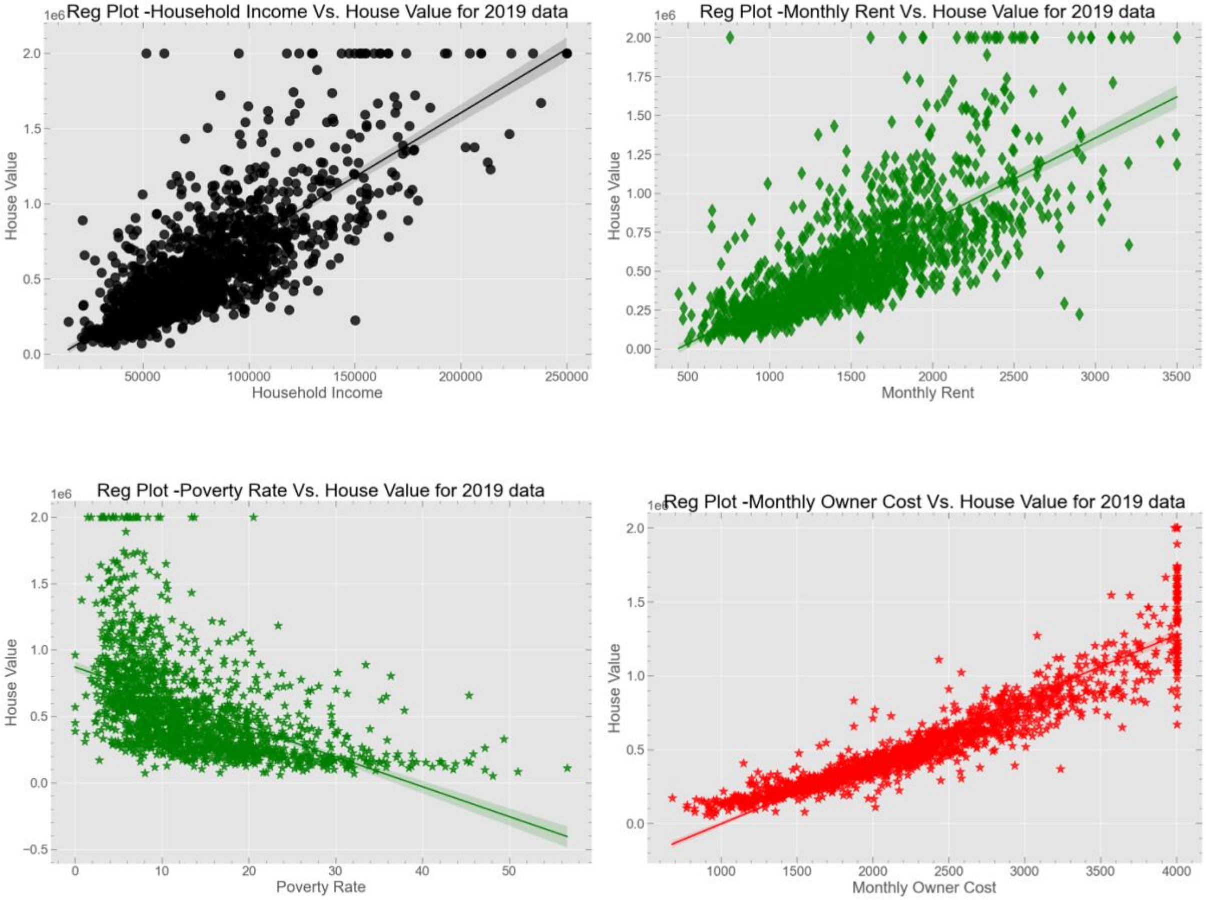


Fig. 2: Scatter plots of housing value with household income, monthly rent, poverty rate, and monthly owner cost. The straight line is the fitted data.

ii. Education

We also explored how with college rate, median age of people, high school rate, and uneducated rate effects the housing value. We saw that there is positive relationship with people having college degree and median age of people, however there is negative relationship with high school people rate and uneducated people rate. (see Fig.3).

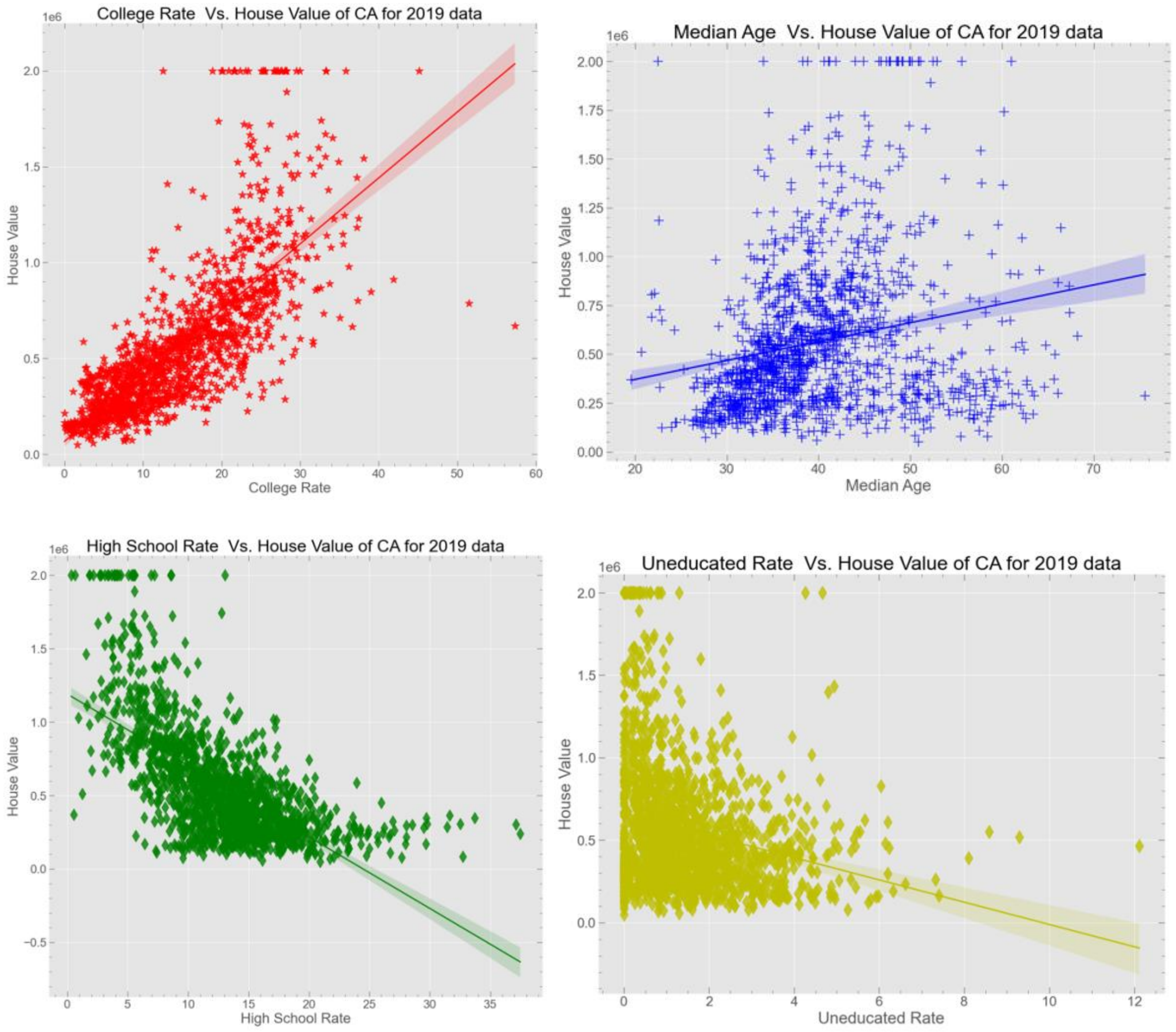


Fig. 3: Scatter plots of housing value with college rate, Median Age of people, high school rate, and college rate. The straight line is the fitted result.

iii. Transportation

For the transportation related parameters, we explored how the public transportation, personal transportation, and commute time of the city effects the housing value. We saw that there is positive relation with public transportation, i.e., city having good public transportation has a higher house value. Personal transportation and commute time has no effects on the housing value as shown in see Fig.4.

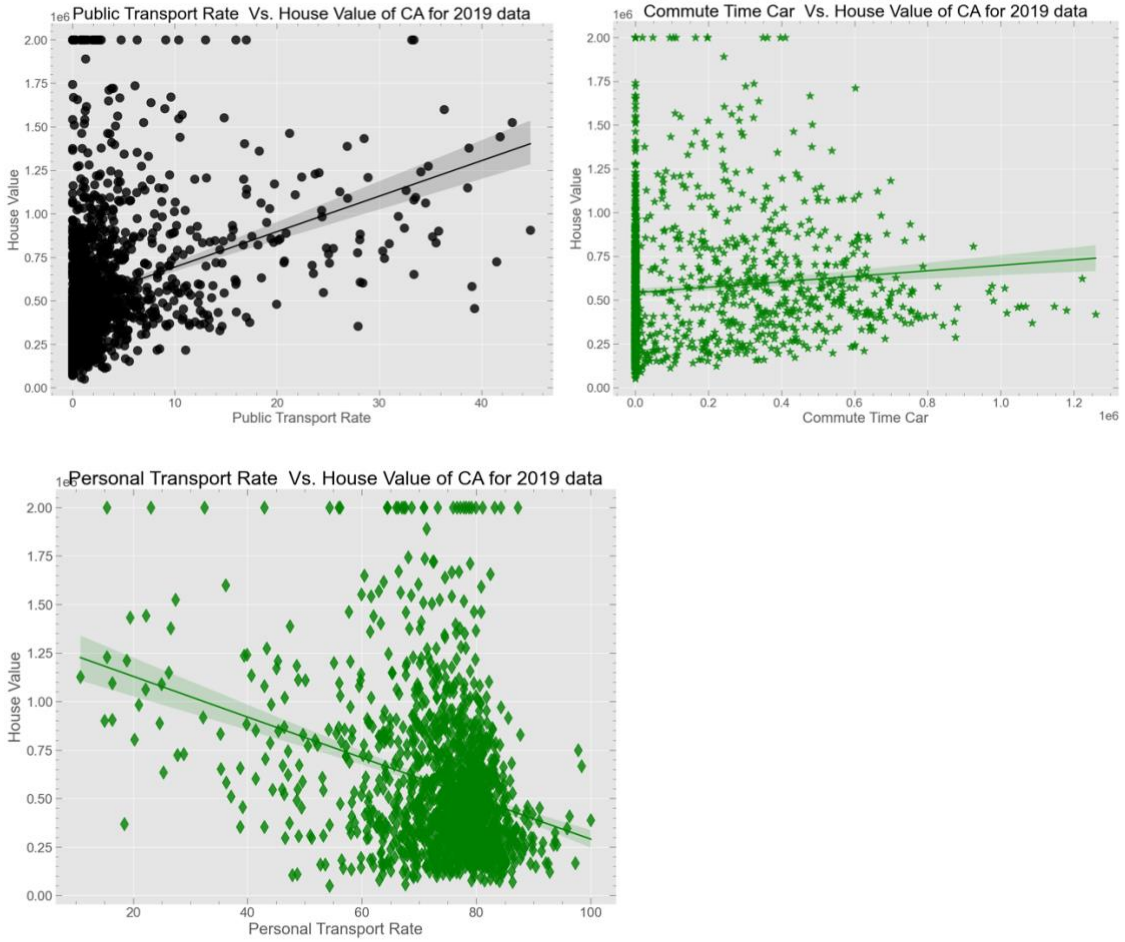


Fig. 4: Scatter plots of housing value with public transportation, personal transportation, and commute time of different cities.

iv. Race

For the race related parameters, we explored how the race of people like, White, Black, Asian, Hispanic race effects the housing value. We saw that there is positive relation with Asian Race people and negative relation with Hispanic race to the housing value. Other race has no effects on the housing value as shown in see Fig.4.

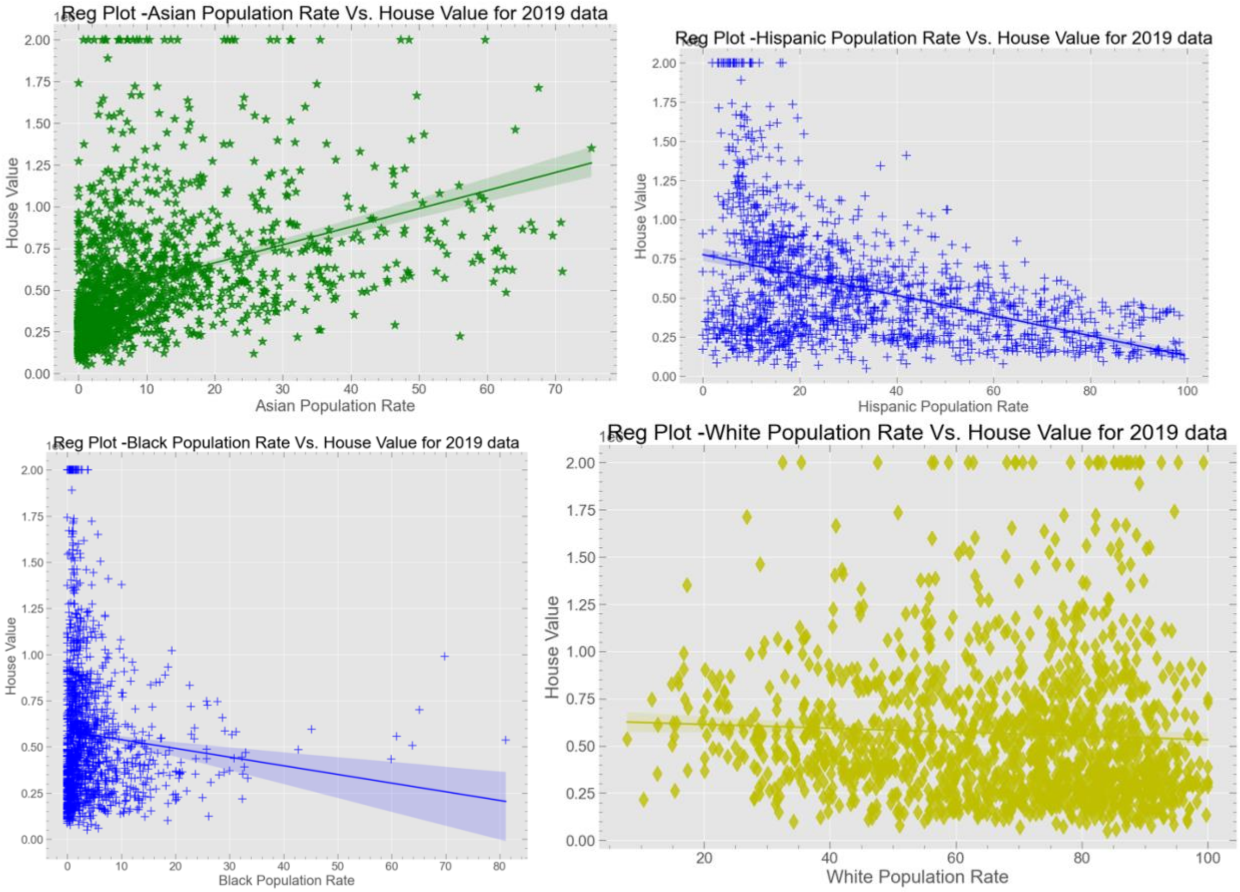


Fig. 5: Scatter plots of housing value with different race for different cities.

b. Understanding the housing price correlation between different terms.

We also explored on housing price for different years from 2012 to 2019, and see slight increase in the California county house value for each year as shown in Fig. 6. We plotted the correlation of housing price of different years data to different other factors as shown in Fig. 7. From the figure, we can see a positive correlation with household income, population, monthly owner cost, monthly rent, public transportation, median people age, per capita income, and college rate. However, there is negative correlation of housing price with poverty rate, unemployment rate, and uneducated rate. We also see an important trend for different years data. The correlation of housing price decrease for people median age and college rate for 2019 in compression to 2012 data, which means younger and less educated people are buying the house more in 2019.

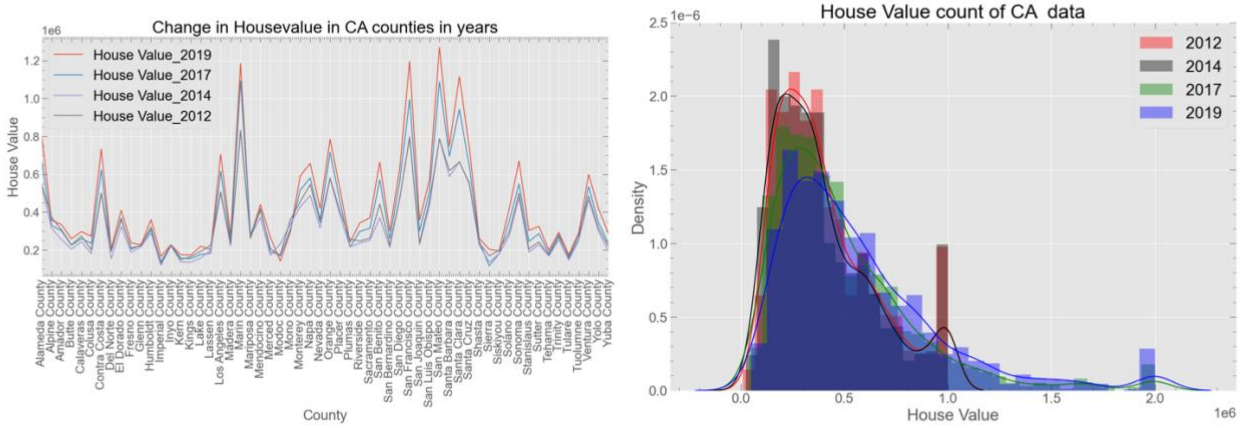
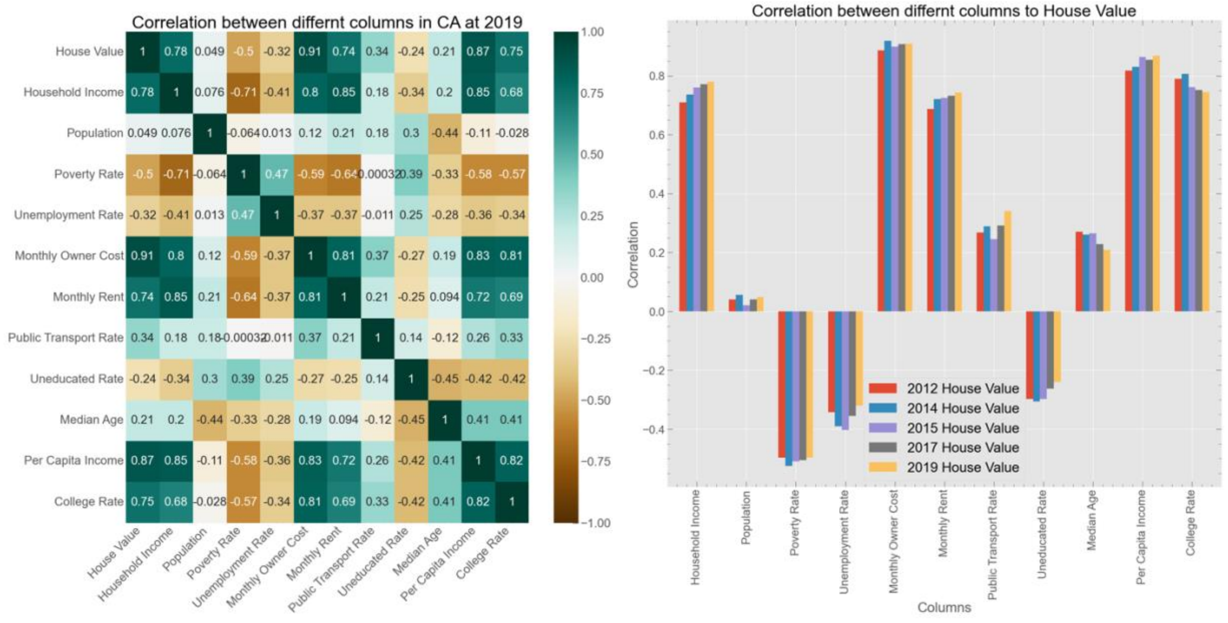


Fig. 6: a) Line plot of housing value with County for different years. b) Distplot of house value for different years.



	House Value	Household Income	Population	Poverty Rate	Unemployment Rate	Monthly Owner Cost	Monthly Rent	Public Transport Rate	Uneducated Rate	Median Age	Per Capita Income	College Rate
House Value	1	0.78	0.049	-0.5	-0.32	0.91	0.74	0.34	-0.24	0.21	0.87	0.75
Household Income	0.78	1	0.076	-0.71	-0.41	0.8	0.85	0.18	-0.34	0.2	0.85	0.68
Population	0.049	0.076	1	-0.064	0.013	0.12	0.21	0.18	0.3	-0.44	-0.11	-0.028
Poverty Rate	-0.5	-0.71	-0.064	1	0.47	-0.59	-0.64	0.00032	0.39	-0.33	-0.58	-0.57
Unemployment Rate	-0.32	-0.41	0.013	0.47	1	-0.37	-0.37	-0.011	0.25	-0.28	-0.36	-0.34
Monthly Owner Cost	0.91	0.8	0.12	-0.59	-0.37	1	0.81	0.37	-0.27	0.19	0.83	0.81
Monthly Rent	0.74	0.85	0.21	-0.64	-0.37	0.81	1	0.21	-0.25	0.094	0.72	0.69
Public Transport Rate	0.34	0.18	0.18	-0.00032	-0.011	0.37	0.21	1	0.14	-0.12	0.26	0.33
Uneducated Rate	-0.24	-0.34	0.3	0.39	0.25	-0.27	-0.25	0.14	1	-0.45	-0.42	-0.42
Median Age	0.21	0.2	-0.44	-0.33	-0.28	0.19	0.094	-0.12	-0.45	1	0.41	0.41
Per Capita Income	0.87	0.85	-0.11	-0.58	-0.36	0.83	0.72	0.26	-0.42	0.41	1	0.82
College Rate	0.75	0.68	-0.028	-0.57	-0.34	0.81	0.69	0.33	-0.42	0.41	0.82	1

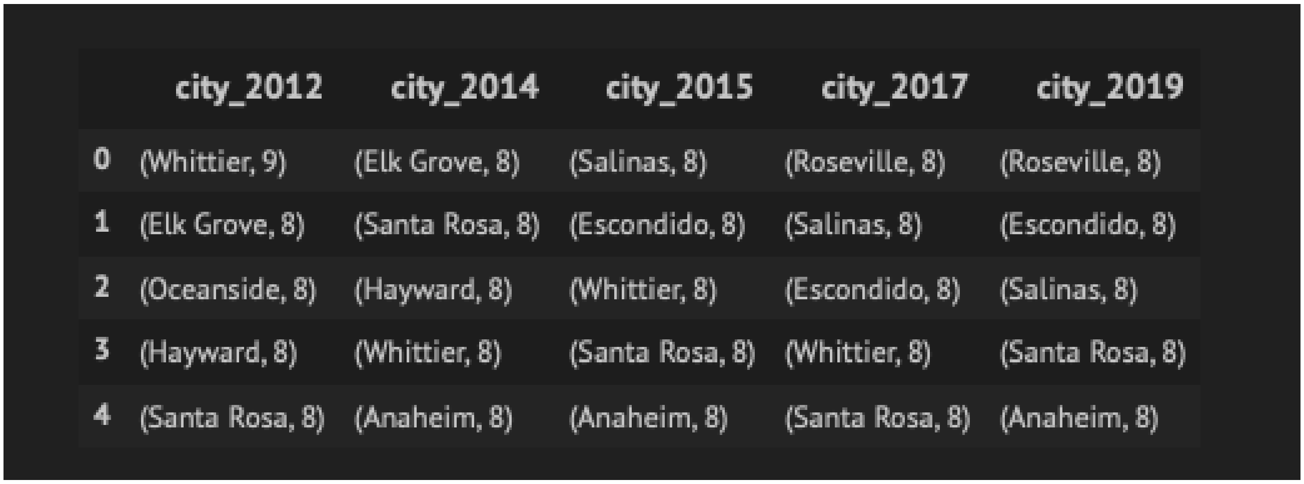
Fig. 7: Correlation of house value with other parameters.

2. Most Affordable 5 Cities of California

Our team analyzed the best cities to live in based on certain criteria, for some it could be weather, cuisine or endless entertainment for family. Hence, we chose to analyze on some of the most important and common factors an individual would consider. After exploring all these different parameters, we nominated the best cities with the most common factors which are as follows:

- **House Value**
- **Income**
- **Cost/Rent**
- **Poverty**
- **Public Transport**
- **Employment/unemployment rate**

We gave each parameter equal weight and nominated the best cities which satisfy most of the parameters above. The table for best cities for different years is as follows



	city_2012	city_2014	city_2015	city_2017	city_2019
0	(Whittier, 9)	(Elk Grove, 8)	(Salinas, 8)	(Roseville, 8)	(Roseville, 8)
1	(Elk Grove, 8)	(Santa Rosa, 8)	(Escondido, 8)	(Salinas, 8)	(Escondido, 8)
2	(Oceanside, 8)	(Hayward, 8)	(Whittier, 8)	(Escondido, 8)	(Salinas, 8)
3	(Hayward, 8)	(Whittier, 8)	(Santa Rosa, 8)	(Whittier, 8)	(Santa Rosa, 8)
4	(Santa Rosa, 8)	(Anaheim, 8)	(Anaheim, 8)	(Santa Rosa, 8)	(Anaheim, 8)

We also plotted the best city using google maps API. As we can see, most best cities are in bay area and southern part of California.

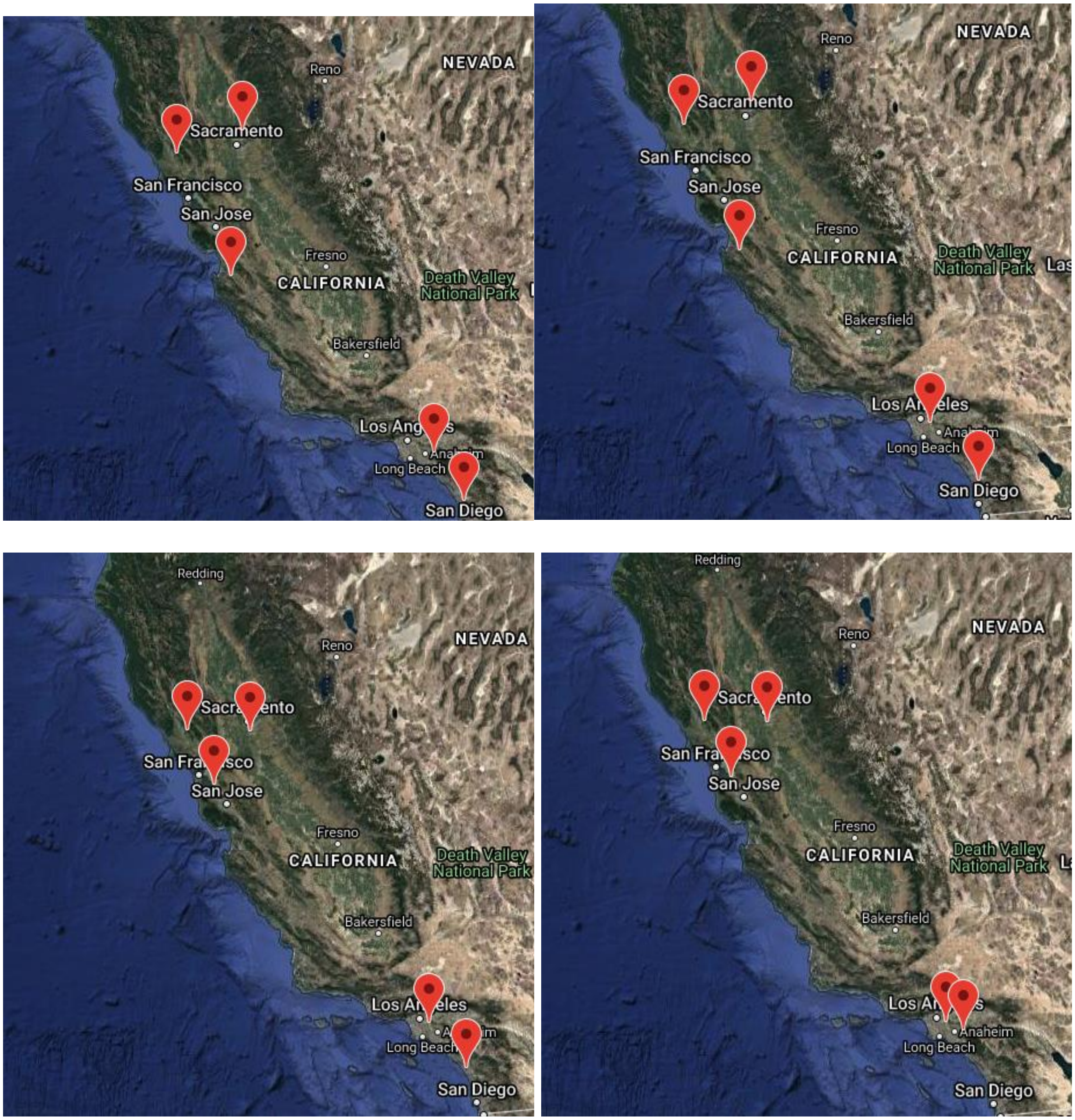


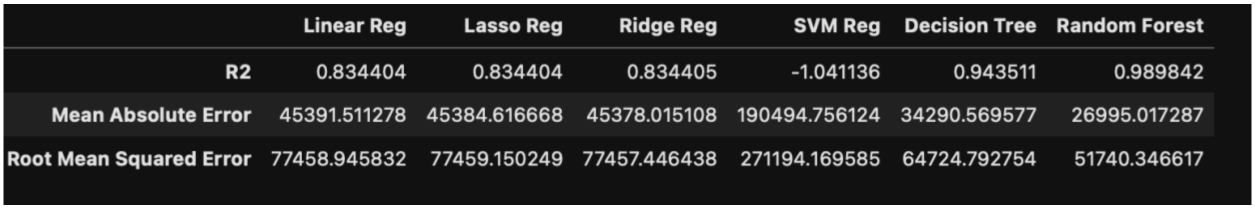
Fig. 8: Our best nominated cities for different years

3. Model Building

To predict the house value, we built different supervised linear models from scikit learn. Different models we worked on are as follows:

- **Linear Regression**
- **Lasso Regression**
- **Ridge Regression**
- **Support Vector Machine**
- **Decision Tree Regressor**
- **Random Forest Regressor**

As we can see from the evaluation metrices below, our best model is Random Forest Regressor which has a mean absolute error of around $27000.



	Linear Reg	Lasso Reg	Ridge Reg	SVM Reg	Decision Tree	Random Forest
R2	0.834404	0.834404	0.834405	-1.041136	0.943511	0.989842
Mean Absolute Error	45391.511278	45384.616668	45378.015108	190494.756124	34290.569577	26995.017287
Root Mean Squared Error	77458.945832	77459.150249	77457.446438	271194.169585	64724.792754	51740.346617

From our best model, we predicted the house value of unseen data from 2017. The actual house value and predicted house value from our best model is shown in Fig 9. So our best model successfully predict the housing value of unseen data from 2017 Census with an prediction error of around $27,000

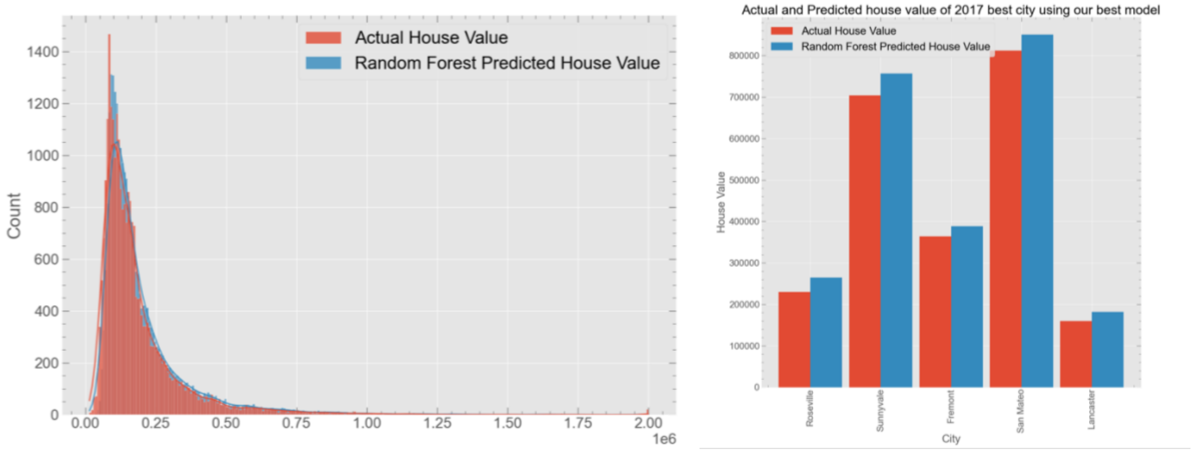


Fig. 9: Actual house value and predicted house value from our best model for 2017 Census data.

Conclusion:

We used Census API to download the 5 years of US data which has 33000 rows and 30 columns. We looked the relationship of house price to other factors like economy, education, transportation and race. We get some interesting trend of house value on different years census data. We also analyzed the data and recommended the best city of CA based on these different parameters. We build a model to predict the house value and our best model predicts the house value perfectly with an error of around $30,000. Hence, from this project, we found the most important parameters for the housing price in the city. From all these different variables, we found the best 5 CA cities to live in. So, from our project we can saw that economy, education, and transportation are the important parameters for the housing value and what recommended that these parameters other cities need to improve to be in the best city.