Pauline Pan
Cyrus Shen

Housing Price Predictions

Our research question was to see what features affected housing prices in San Diego. We decided to look at gender and race for data visualization and linear regression. We also included male age between 18 and 65 for linear regression. In the beginning, we had two datasets of population and housing information of San Diego in 1970. After we explored the two datasets, we found that some data were NaN or filled in with ellipses. In order to deal with this, we first replaced the ellipses with np.NaN values and then dropped the rows with more than 5 NaN values because we wanted to focus on rows with as much information as possible. We then interpolated the data to replace any remaining NaN values, but for filling in missing city names, we decided to randomly pick and fill in the city name since the city rows were not in any particular order. Since we wanted to explore how distribution of gender and people of different races affected the housing prices, we selected columns of gender and races in the dataset of population. However, since values in these columns were of numbers of type string, we converted them to integers by replacing ‘,’ with empty space if the values in the strings are in thousands and then convert them to np.int64. Then, we summed up columns of gender and races to get the total number of males, females, and people of different races. For the dataset of housing prices, after we selected the columns of average prices, which were in type string, we stripped ‘$’, replaced ‘,’ with empty space, and split the decimal point ‘.’ to convert these values to floats that can be used for calculation. After we dealt with invalid inputs, selected columns that we wanted to use for testing our assumption, and changed data types, we merged our new datasets of population and housing prices together based on the block number, census tract number, and city.

After we cleaned and processed our data, we first decided to look at how total owner occupied average value of housing units were distributed in San Diego cities as shown in Fig1. To make this plot, we grouped the merged dataset by place names of San Diego, and then calculated means of total owner occupied average value of housing units. Then, we made the following barh plot using matplotlib.pyplot library. We can see that price was about the same except for Grossmont Mount Helix, Del Mar, and Coronado. In order to explain this, we did some research online on the most expensive San Diego cities in 2019 and found that Del Mar and Coronado were among the top 100 most expensive US cities[1].

[1] https://patch.com/california/san-diego/4-san-diego-county-zip-codes-among-100-most-expensive-u-s

Pauline Pan
Cyrus Shen

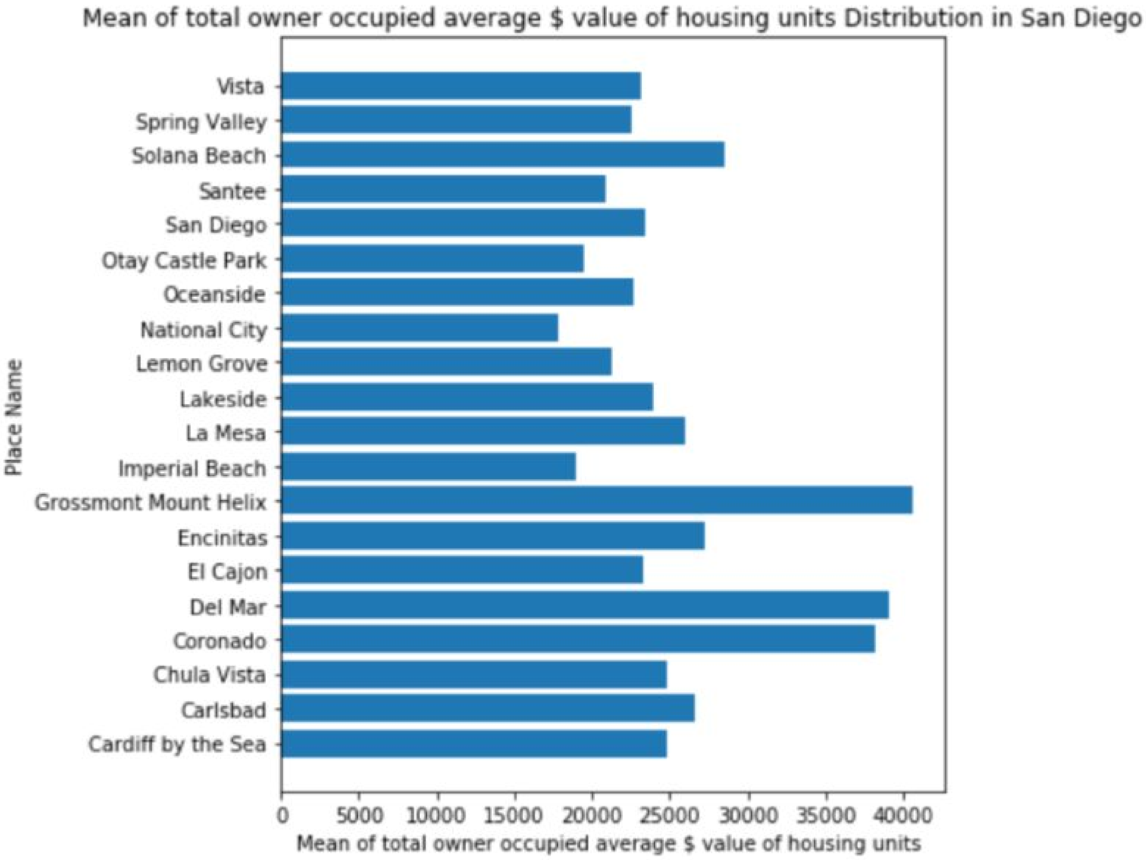


Fig 1: Mean of total owner occupied average $ value of housing units distribution in San Diego

Pauline Pan
Cyrus Shen

In order to see if housing prices were affected by the ratio of genders in the surrounding area by comparing it with Fig1, we plotted the bar plot of ratio of male to females in each San Diego city in the dataset using matplotlib.pyplot library. To get the data for Fig2, we grouped by city and then took the average value of total males and females by city. Then we got ratio by dividing the male/female values by sum of male and female. We expected to see more males in the more expensive cities since a higher priced house likely means that the owner earns more money and in 1970, it was likely that mostly males had higher paying jobs. From Fig2, we found that the ratio of male to females was actually close to 50/50 and there is not a clear dominant gender in each city.

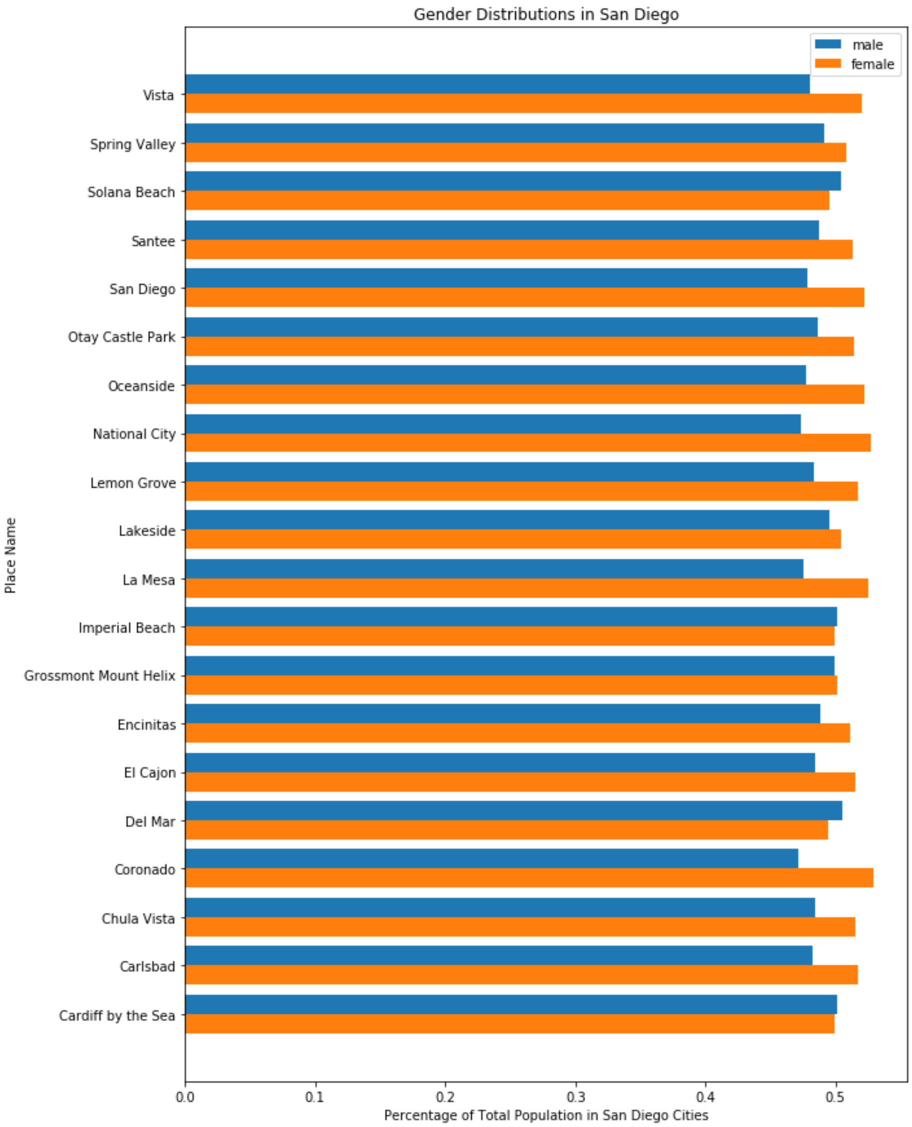


Fig2: Gender distributions per city in San Diego

Lastly, we compared race distribution percentages in San Diego cities in Fig3 with the pricing distributions in Fig1 and unfortunately could not get much meaningful insights since the city populations were mostly dominated by whites.

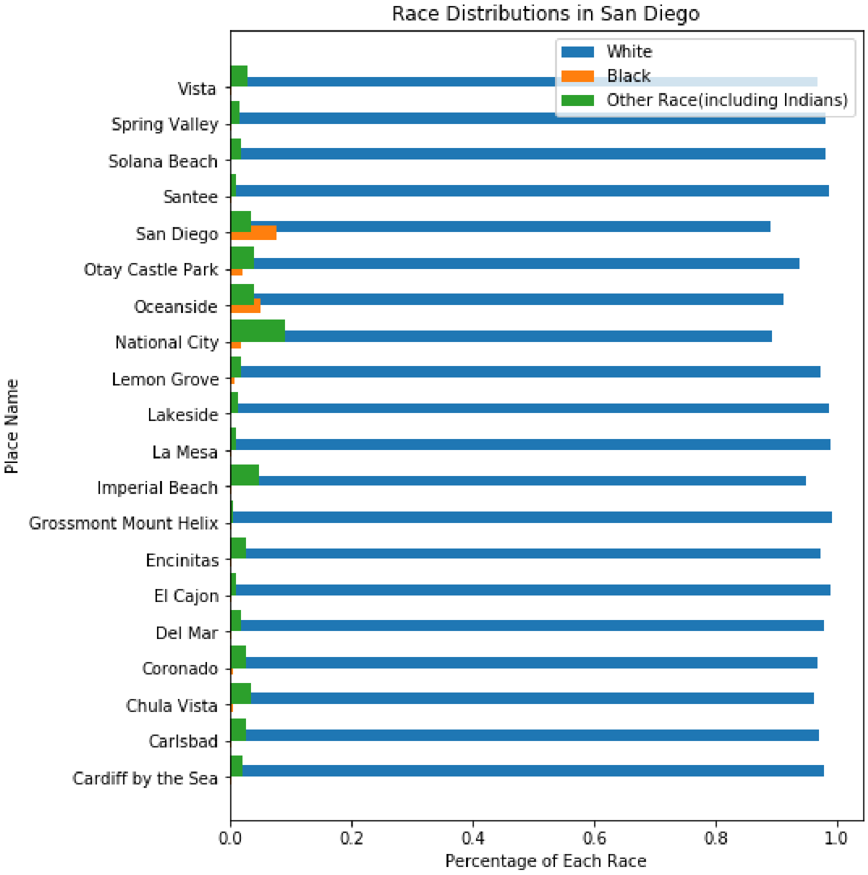
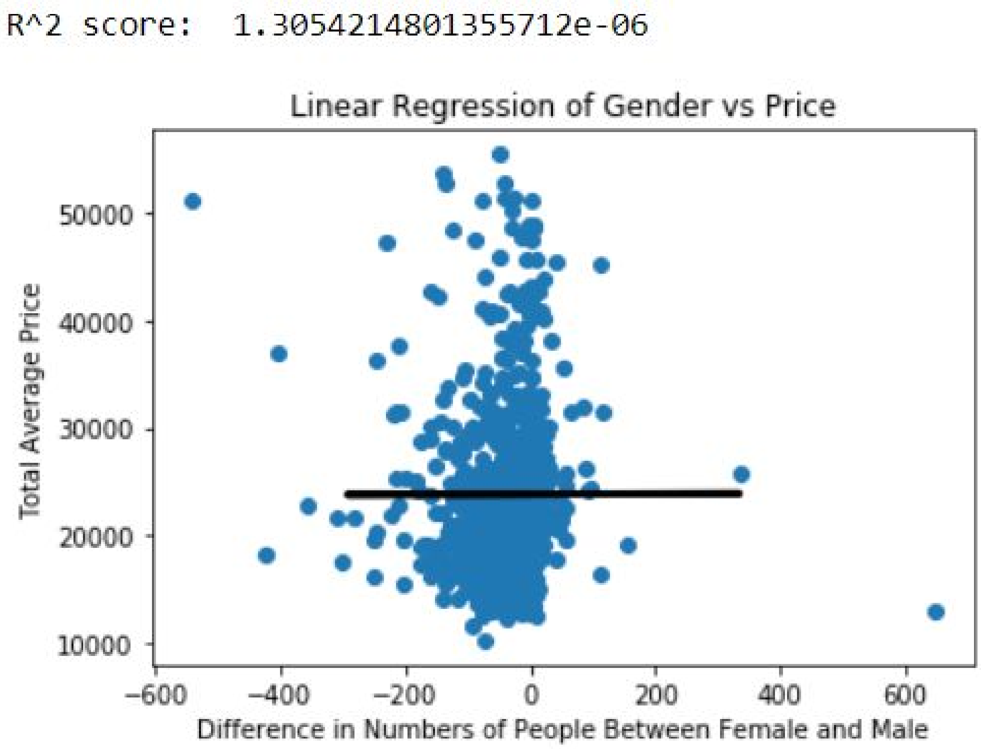


Fig3: Race distribution percentages in San Diego

In the Machine Learning Part, we wanted to employ Linear Regression model to see if there is a linear correlation between the difference in number of people between females and males and total owner occupied average prices. We thought that gender, race, and male age affected the price of the house so we created some linear regression models using those 3 variables. It turns out that they all don't correlate too well but gender performed the worst while male age had the highest R^2 value for predicting price. We first imported Linear Regression and techniques of splitting test and train sets from sklearn library. Then, we took the total males and total females columns and subtracted them from each other and fed that into the model which is why the x-axis goes from negative to positive. Then, we used part of the dataset as train data to fit the linear model, and then predicted the remaining x-values using trained Linear Regression

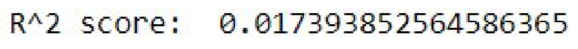
Model. Unfortunately, as we can see in Fig4, our model did not work very well. This could mean that gender is not a good predictor of house price. The R^2 value is very close to 0 which confirms that there is not much correlation between gender and price.



```
R^2 score:  1.3054214801355712e-06
```

Fig4: Linear regression of gender vs price

In Fig5, we fit a linear regression model for predicting house price using the difference between the number of whites and blacks to see if race could influence total average housing prices. Our model does better than the one in Fig4 but it is still not great since it has a R^2 score of 0.017 which is still relatively close to 0. In hindsight, taking the difference between races or even using race in the first place may not have been a great idea since the total number of whites completely dominated the other races in our dataset so our model ends up predicting house price based on the number of white people in the area.



```
R^2 score:  0.017393852564586365
```

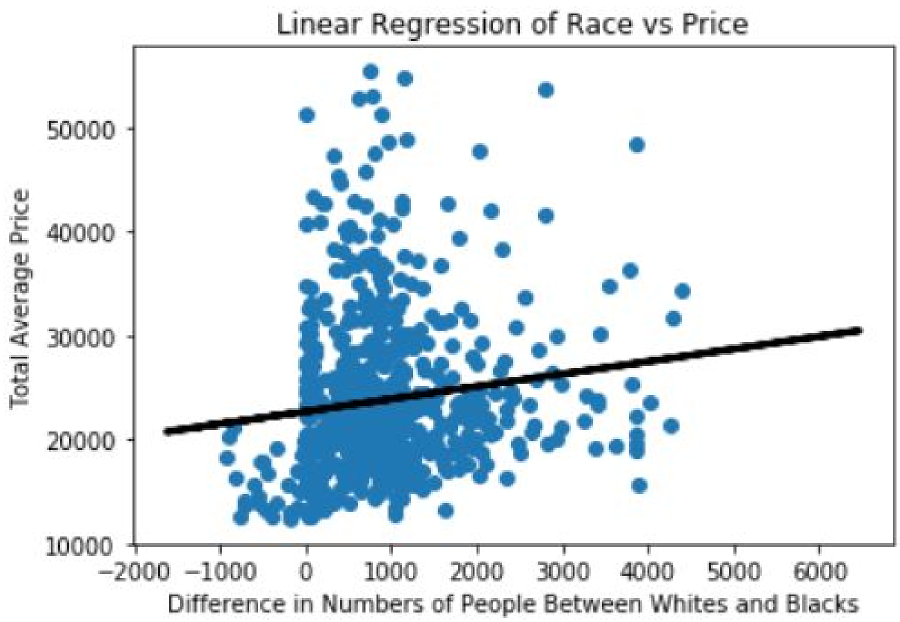
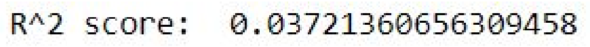


Fig5: Linear regression of race vs price

In Fig6, we fit a linear regression model for house price using the difference between the number of males aged 18-34 and males aged 35-64. The model performed better than the ones in Fig4 and Fig5 but is still not the great since it has a R^2 score of 0.037.



```
R^2 score:  0.03721360656309458
```

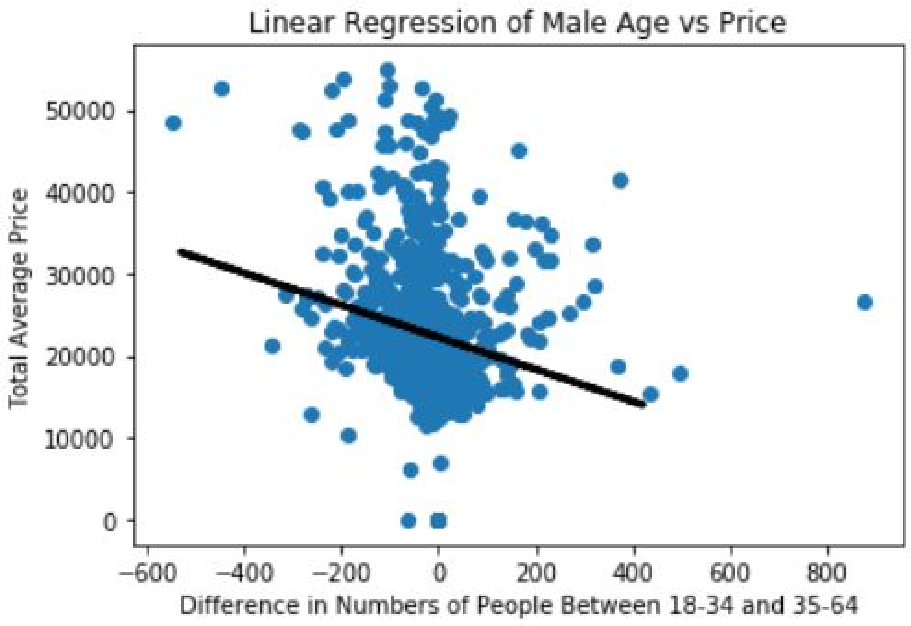
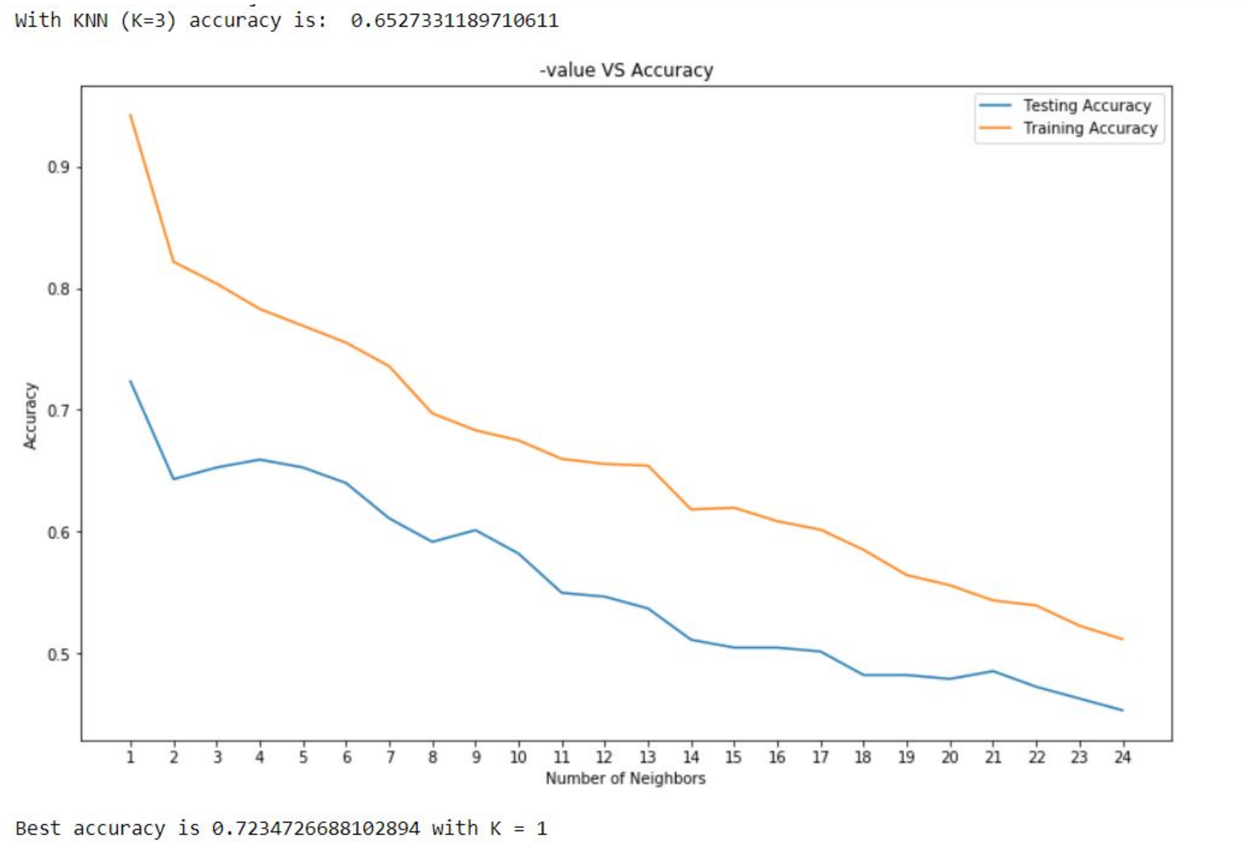


Fig6: Linear regression of male age vs price

Pauline Pan
Cyrus Shen

In Fig7, we performed a KNN classification to predict housing units with all plumbing facilities and 1.01 or more persons per room with data of occupied housing units 1-1.51 or more persons per room. We thought that there would be clusters for plumbing based on the number of people per room since more people would probably need more plumbing facilities. Our model did alright with an accuracy of 74% and n_neighbors=1, showing that there could be some clustering.



```
With KNN (K=3) accuracy is:  0.6527331189710611
```

```
Best accuracy is 0.7234726688102894 with K = 1
```

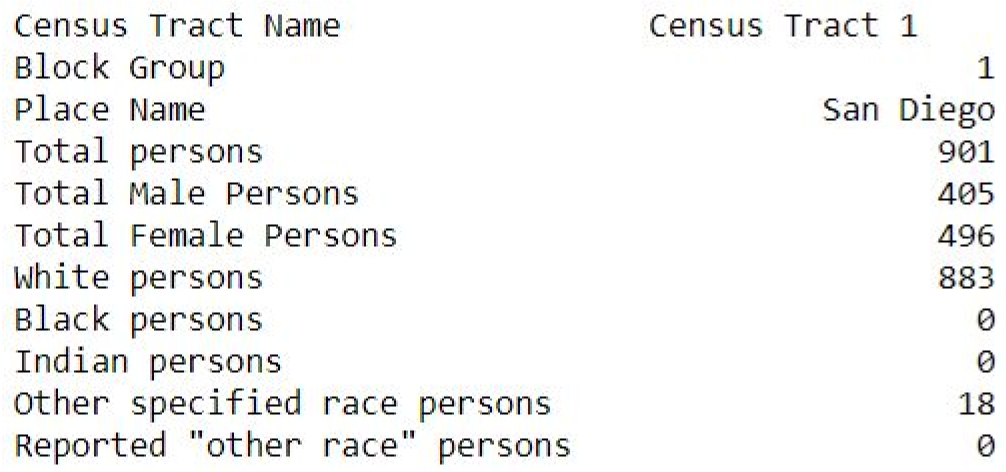
Fig7: KNN to predict housing units with all plumbing facilities and 1.01 or more persons per room with data of occupied housing units 1-1.51 or more persons per room.

Pauline Pan
Cyrus Shen

Questions:

1)

1st row of our cleaned population dataframe:



```
Census Tract Name                  Census Tract 1
Block Group                                     1
Place Name                              San Diego
Total persons                                 901
Total Male Persons                            405
Total Female Persons                          496
White persons                                 883
Black persons                                   0
Indian persons                                  0
Other specified race persons                   18
Reported "other race" persons                   0
```

2) Chula Vista is the second most common city following San Diego
3) For this question, we grouped by city and looked at ‘San Diego’ specifically. We get an average of $23436.167254.
4) Highest average price belongs to Grossmont Mount Helix.