Group 5 Final Report

1. Introduction

With the new crown epidemic, more and more people choose to buy a house instead of renting. Thus, "What factors affect the price of houses in major cities?" become a question that plagues many homebuyers in the present.

To answer this question, our group decided to refine this further to the question “What are the principal factors that affect housing prices in California?” Thus, our target audience is the college-educated general public who wants or is ready to buy a house in California. In order to attract this audience, we planned to choose the dataset of housing prices in California which includes several factors that may affect housing prices, such as distance between house and ocean, income for households, etc. Since geography has changed little in the last 30 years, we hope this information will also help current home buyers understand what factors impact home prices and provide a reference for homebuyers.

2. Data

2.1 Data Description

The original data is from the paper *‘Sparse spatial autoregressions’*. This dataset contains housing prices in California given the exact directions, based on 1990 California census data. Since the dataset is large and not cleaned, we need to implement a simple EDA to figure out the features of our data in the following steps. Here are descriptions of each variable in that dataset(from Kaggle directly):

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. housingMedianAge: Median age of a house within a block; a lower number is a newer one
4. totalRooms: Total number of rooms within a block
5. totalBedrooms: 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. medianIncome: Median income for households within a block of houses (measured in tens of thousands of US Dollars)
9. medianHouseValue: Median house value for households within a block (measured in US Dollars)
10. oceanProximity: Location of the house w.r.t ocean/sea

2.2 Exploratory Data Analysis (EDA)

```
> summary(CAHousing)
   longitude         latitude     housing_median_age  total_rooms    total_bedrooms
 Min.   :-124.3   Min.   :32.54   Min.   : 1.00      Min.   :    2   Min.   :   1.0
 1st Qu.:-121.8   1st Qu.:33.93   1st Qu.:18.00      1st Qu.: 1448   1st Qu.: 296.0
 Median :-118.5   Median :34.26   Median :29.00      Median : 2127   Median : 435.0
 Mean   :-119.6   Mean   :35.63   Mean   :28.64      Mean   : 2636   Mean   : 537.9
 3rd Qu.:-118.0   3rd Qu.:37.71   3rd Qu.:37.00      3rd Qu.: 3148   3rd Qu.: 647.0
 Max.   :-114.3   Max.   :41.95   Max.   :52.00      Max.   :39320   Max.   :6445.0
                                                                    NA's   :207
   population      households     median_income     median_house_value ocean_proximity
 Min.   :    3   Min.   :   1.0   Min.   : 0.4999   Min.   : 14999     Length:20640
 1st Qu.:  787   1st Qu.: 280.0   1st Qu.: 2.5634   1st Qu.:119600     Class :character
 Median : 1166   Median : 409.0   Median : 3.5348   Median :179700     Mode  :character
 Mean   : 1425   Mean   : 499.5   Mean   : 3.8707   Mean   :206856
 3rd Qu.: 1725   3rd Qu.: 605.0   3rd Qu.: 4.7432   3rd Qu.:264725
 Max.   :35682   Max.   :6082.0   Max.   :15.0001   Max.   :500001
```

Firstly, we found that the total_bedrooms variable has 207 NA values when summarizing each variable. Then, we deleted those NA variables. To better understand the skewness in ocean_proximity, we changed the category variable ocean_proximity from a categorical variable to a factorial variable.

```
> summary(CAHousing$ocean_proximity)
 <1H OCEAN      INLAND      ISLAND    NEAR BAY NEAR OCEAN
      9034        6496           5        2270        2628
```

Further insight into ocean_proximity, we figured out that the ISLAND has only five rows, while the other levels have more than 2000 rows. Due to possible problems with the model fit, we decided to accept the risk and remove this level from ocean_proximity.

In addition, we choose numerical values to quantify our categorical variables to help with our model selection process, namely ocean_less1h (<1H OCEAN); ocean_inland (INLAND); near_bay (NEAR BAY); and near_ocean (NEAR OCEAN). These dummy variables will use binary variables 1 or a 0 to indicate a yes or no answer.

2.3 Data Visualization

We built a shiny app for visualization. Click here. This shiny app contains 4 plots, location vs. price, location vs. population, house age vs. price, distance_to_ocean vs. price. We used it to gain a preliminary understanding of our data.

3. Methodology

3.1 Introduction to Model

Since our data contain many explanatory variables and the goal is to explore the principal factors affecting California housing prices, we decided to use the regression model. Two different regression models were subsequently implemented in this project, ordinal regression and multi-linear regression.

3.2 Ordinal Regression Model

Beautiful beaches in California would affect the housing prices. From the geographical graph (in the shiny app, location vs. price), the closer houses to the ocean, the higher median house values,

while houses inland have lower median house values. It tells us that the ocean proximity variable would significantly impact the median house value in California.

Hence, our group started building the model with only one explanatory variable, which is "ocean_proximity", composed of 4 categorical factors and a response variable ("median house value"). The reason why we choose the ordinal regression model is because we transformed the response variable into the ordinal variable having three different factors (low, medium, high) by grouping. Grouping the median house values are based on the one-third and second-thirds points of all values so that the frequency within each group has the same number compared with other groups. As a result, predicting the probability of house values being in a specific group is not biased. Our ordinal regression model and numeric summary are shown below:

$$P(y_i \leq k) = r_{ik} = \exp(\theta_k - \eta_i) / (1 + \exp(\theta_k - \eta_i))$$

```
                      Value Std. Error     t value      p value
ocean_proximity1 -2.51310998 0.03596993 -69.8669621 0.000000e+00
ocean_proximity2  0.16896054 0.04505484   3.7501082 1.767583e-04
ocean_proximity3  0.04105888 0.04275116   0.9604157 3.368460e-01
```

The table shows that each factor has a significant impact on "median house value". The significance depends on factors though. Only "ocean_proximity1" is negatively correlated with "median house value", because the factor represents inland, while the basis "ocean_proximity0" represents the houses less than an hour from the ocean. The others represent "Near Bay", and "Near Ocean", respectively, which coincides with the relatively small values and large p-values since "Less than an hour", "Near Bay", and "Near Ocean" are closer to the ocean than "inland".

In addition, our group added "housing_median_age" as another variable in the model. We only considered adding one variable since too many variables would make this model complex.

```
Coefficients:
                    Value Std. Error t value
housing_median_age -0.01249   0.001172 -10.662
ocean_proximity1   -2.58747   0.036804 -70.304
ocean_proximity2    0.27086   0.046214   5.861
ocean_proximity3    0.04331   0.042881   1.010

Intercepts:
    Value    Std. Error
0|1  -1.9884   0.0427
1|2  -0.1589   0.0396
```

This model's deviance is 38095.49, less than that of the first model with 38210.12, which means adding "housing_median_age" gives a better explanation of the median house values. Also, the variable's negative impact on "median housing values" is observed from the above tables. For example, if houses within a block are inland and 50 years on average, the log-odds for the median house values to be within the "low" class range is the following:

log-odds = -1.9884 + 2.58747 + 50 * 0.01249, where the probability is 0.7727

```
> new_obs <- data.frame(ocean_proximity = "1", housing_median_age = 50 )
> predict(hp_lm, type = "probs", newdata = new_obs)
        0         1         2
0.7727050 0.1822169 0.0450781
```

These log-odds and probability can be checked via the "predict" function in the MASS package. From the ordinal regression model, our group could observe that both a distance from the ocean and "housing_median_age" are negatively related to median housing values in California.

3.3 Multi-linear Regression Model

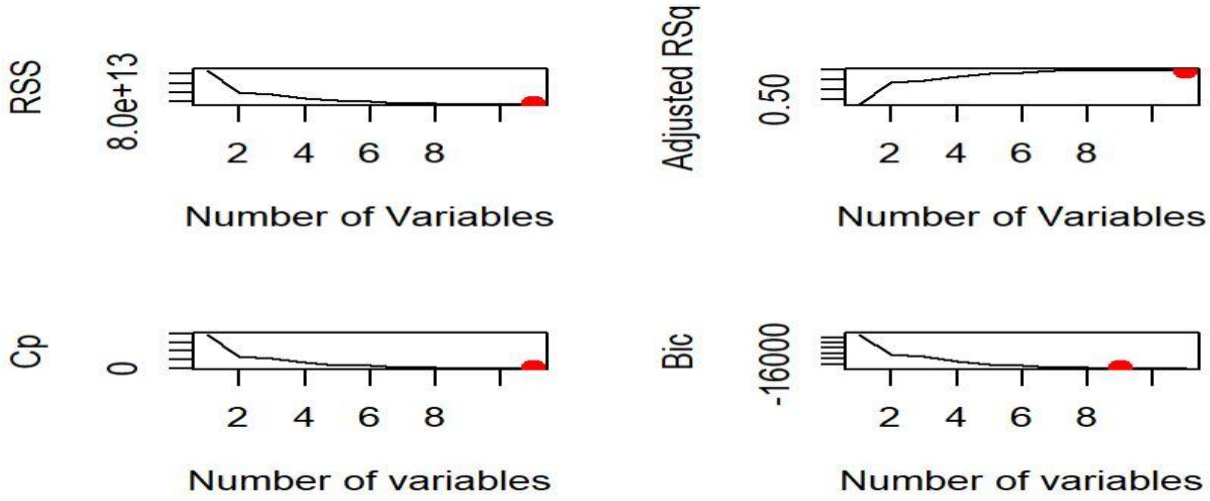
As we mentioned above, we found that some explanatory variables may have significant impacts on the median price of a house. But in the previous discussion, we only considered no more than two explanatory variables. Then the natural idea is to consider more explanatory variables and start with the simplest model, the linear regression model. Because if a simple model can fit quite well, then there's no need to waste time on complicated models.

To fit the best multi-linear regression model, we use the "best model selection" method by *regsubsets* in R package *leaps*. The main idea of the "best model selection" is searching for all the possible models. For example, if we have n variables, which means we have 2^n possible linear models, and then pick the best one by using model selection criteria like BIC or RSS.

```
regfit.full=regsubsets(median_house_value~.,data =
data3[train,], nvmax = 11)
reg.summary=summary(regfit.full)
```

Before starting, we considered splitting our data first. To test our result, we split our data into two parts: 80% for training data and 20% for test data. We train on the training data and test on the test data. Because we have the categorical variable "ocean_proximity", we also need to keep the distribution of the label the same on both training and test data.

First, we selected by rss, Adjusted R-Square, Mallows' Cp statistics, and Schwartz's information criterion, BIC. We drew all of our results in the graph below, the red point in each graph means the best model we select. The index of the model we choose are 11,11,11 and 9.

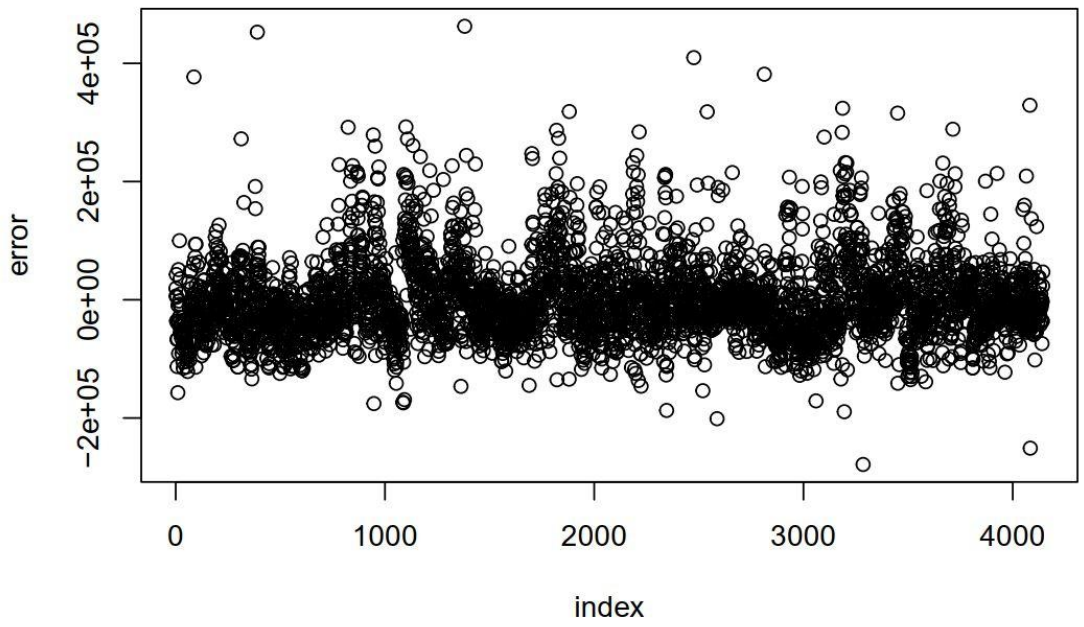


After predicting the test set and computing the error, we calculate the MSE and pick the one with the least MSE. The index number is 11. In this case, we mainly selected between the number 9 model and the number 11 model. To further make a decision, we utilized the k-fold cross-validation to check our model. We picked k=10 here, and the main idea was to divide all the data into 10 folds, then compute for 10 rounds. In each round, we used 9 folds as training data and the rest as the test set. At last, we averaged all the errors in each round as the final error for the model.

The model with the least cross-validation error is the number 11 model. In this case, this model was picked up as our final result. The coefficients of the model on our training data are shown below. We regarded these as a measure of climate change for longitude and latitude. For example, northern places would be colder in winter than southern places, which might be a factor that affects housing prices.

(Intercept)	longitude	latitude
-2.265255e+06	-2.678295e+04	-2.551030e+04
housing_median_age	total_rooms	total_bedrooms
1.088585e+03	-5.649179e+00	9.978323e+01
population	households	median_income
-3.654846e+01	4.431026e+01	3.906886e+04
ocean_proximityINLAND	ocean_proximityNEAR BAY	ocean_proximityNEAR OCEAN
-3.940828e+04	-4.034668e+03	3.369864e+03

And there is the error in the test data, the MSE is 4638962473, which is quite huge.



Our model seems to fit not so well, and one possible reason is that our model may not be so suitable for the linear model. For example, the four factors of “ocean_proximity” in the linear

model containing all the variables are not statistically significant(as shown in the graph below). But the "best model selection" method shows that models with "ocean_proximity" are still better than the models without it.

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)

ocean_proximityNEAR BAY   -4.035e+03  2.146e+03  -1.880   0.0601 .
ocean_proximityNEAR OCEAN  3.370e+03  1.752e+03   1.924   0.0544 .
```

4. Conclusion

In this project, we analyzed the principal factors affecting housing prices based on data from 1990 California and built two different models to explain housing prices.

Firstly, we found that the distance from the ocean and the median age of a house are negatively correlated with housing prices by analyzing the ordinal regression model, which means newer houses or houses near the ocean would probably have higher prices.

Secondly, we built a multiple linear regression model, splitting our data set to 80% training and 20% testing. We used "best model selection" to search from all possible MLR models and pick the best one. However, the MSE of the selected best model is huge, indicating that the multiple linear regression model cannot explain housing prices well.

There is an interesting point that based on the two different models, the relationship between house age and house price is opposite. In the ordinal model, it is negatively related, while it is positively correlated in the multiple linear model. We guess the effect of grouping variables might cause it. We can interpret it intuitively from the shiny app house age vs price plot. It is not a monotone function of price in respect to the house age. Therefore, it is possible to inverse the conclusion by grouping housing prices.

Generally, some of our explanatory variables, like ocean proximity and median income, are the principal factors that significantly correlate with California housing prices. Buyers and sellers are more likely to value houses near the ocean or rich blocks at higher prices. But there is still some part of housing prices that our data and models cannot explain. We are looking forward to finding more detailed data and building models to explain housing prices better.

5. Reference

Nugent, Cam. "California Housing Prices." Kaggle, 24 Nov. 2017, https://www.kaggle.com/camnugent/california-housing-prices?select=housing.csv.

Kelley Pace, R., and Ronald Barry. "Sparse Spatial Autoregressions." *Statistics & Probability Letters*, vol. 33, no. 3, 1997, pp. 291–297., https://doi.org/10.1016/s0167-7152(96)00140-x.