STAT 420: Data Analysis Project

Amandeep,Kumar Gaurav and Vishal

08/03/2020

Table of Contents

Team Engineers

Name	NetID
Amandeep Takhar	atakhar2
Kumar Gaurav	Kgdubey2
Vishal Agarwal	vishala2

Introduction

Title - **California Housing Price Prediction** An analysis on factors contributing to determine housing price in California

Dataset background

The data pertains to the houses found in a given California district and some summary stats about them based on the 1990 census data.The dataset contains 20640 records and 9 predictors. Our goal is to explore correlation between given variables like total bedroom

,population ,ocean proximity etc in determining the price of housing in a given area.In the process we would also like to divide the dataset into test and train and test the behavior of our model.

Source Dataset

Reading data

```
data = read.csv("housing.csv")

head(data, 10)

##    longitude latitude housing_median_age total_rooms total_bedrooms
population
## 1    -122.23    37.88                 41         880            129
322
## 2    -122.22    37.86                 21        7099           1106
2401
## 3    -122.24    37.85                 52        1467            190
496
## 4    -122.25    37.85                 52        1274            235
558
## 5    -122.25    37.85                 52        1627            280
565
## 6    -122.25    37.85                 52         919            213
413
## 7    -122.25    37.84                 52        2535            489
1094
## 8    -122.25    37.84                 52        3104            687
1157
## 9    -122.26    37.84                 42        2555            665
1206
## 10   -122.25    37.84                 52        3549            707
1551
##    households median_income median_house_value ocean_proximity
## 1         126        8.3252             452600        NEAR BAY
## 2        1138        8.3014             358500        NEAR BAY
## 3         177        7.2574             352100        NEAR BAY
## 4         219        5.6431             341300        NEAR BAY
## 5         259        3.8462             342200        NEAR BAY
## 6         193        4.0368             269700        NEAR BAY
## 7         514        3.6591             299200        NEAR BAY
## 8         647        3.1200             241400        NEAR BAY
## 9         595        2.0804             226700        NEAR BAY
## 10        714        3.6912             261100        NEAR BAY

str(data)

## 'data.frame':    20640 obs. of  10 variables:
##  $ longitude         : num  -122 -122 -122 -122 -122 ...
##  $ latitude          : num  37.9 37.9 37.9 37.9 37.9 ...
```

```
##  $ housing_median_age: num  41 21 52 52 52 52 52 52 42 52 ...
##  $ total_rooms       : num  880 7099 1467 1274 1627 ...
##  $ total_bedrooms    : num  129 1106 190 235 280 ...
##  $ population        : num  322 2401 496 558 565 ...
##  $ households        : num  126 1138 177 219 259 ...
##  $ median_income     : num  8.33 8.3 7.26 5.64 3.85 ...
##  $ median_house_value: num  452600 358500 352100 341300 342200 ...
##  $ ocean_proximity   : chr  "NEAR BAY" "NEAR BAY" "NEAR BAY" "NEAR BAY"
...
```

Description about the variables

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 building

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 (USD Response variable)

10. oceanProximity: Location of the house w.r.t ocean/sea

Method

Missing Data

As a first step in data quality , we will look for missing data.

```
sum(is.na(data))

## [1] 207
```

We see 207 missing values, which we plan to remove in the below step.

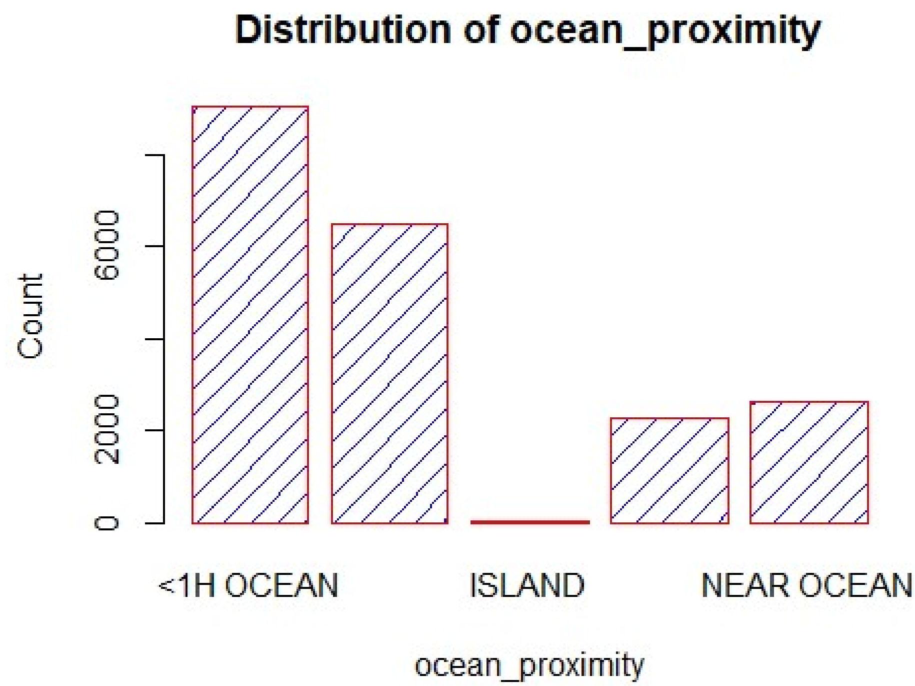
```
data = na.omit(data)
str(data)
```

```
## 'data.frame':    20433 obs. of  10 variables:
##  $ longitude         : num  -122 -122 -122 -122 -122 ...
##  $ latitude          : num  37.9 37.9 37.9 37.9 37.9 ...
##  $ housing_median_age: num  41 21 52 52 52 52 52 52 42 52 ...
##  $ total_rooms       : num  880 7099 1467 1274 1627 ...
##  $ total_bedrooms    : num  129 1106 190 235 280 ...
##  $ population        : num  322 2401 496 558 565 ...
##  $ households        : num  126 1138 177 219 259 ...
##  $ median_income     : num  8.33 8.3 7.26 5.64 3.85 ...
##  $ median_house_value: num  452600 358500 352100 341300 342200 ...
##  $ ocean_proximity   : chr  "NEAR BAY" "NEAR BAY" "NEAR BAY" "NEAR BAY"
...
##  - attr(*, "na.action")= 'omit' Named int [1:207] 291 342 539 564 697 739
1098 1351 1457 1494 ...
##   ..- attr(*, "names")= chr [1:207] "291" "342" "539" "564" ...
```

Categorical Variables

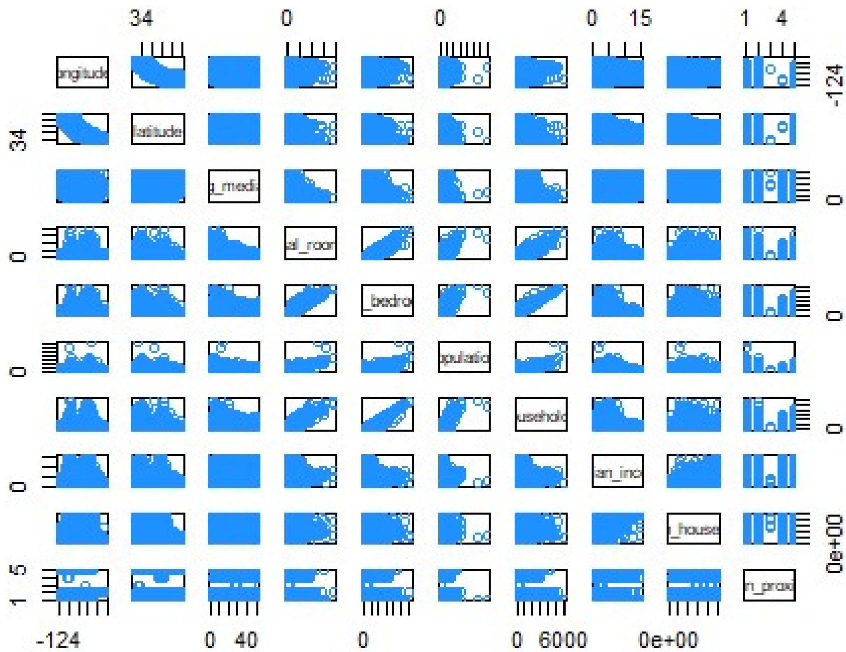
On taking an in depth look at each variable, we decided to make ocean_proximity as a categorical variable, we can see below that it is broadly classified into 5 values.

```
is.factor(data$ocean_proximity)

## [1] FALSE

data$ocean_proximity = as.factor(data$ocean_proximity)
levels(data$ocean_proximity)

## [1] "<1H OCEAN"  "INLAND"     "ISLAND"     "NEAR BAY"   "NEAR OCEAN"

barplot(table(data$ocean_proximity), main="Distribution of ocean_proximity",
          xlab="ocean_proximity",
          ylab="Count",
          border="red",
          col="blue",
          density=10)
```



The distribution depicts that “island” has the least count and “1H OCEAN” has the maximum count. This data also make practical sense.

```
pairs(data, col = "dodgerblue")
```




```r
kable(t(cor(data[,-10])))
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
longitude	1.0000000	-0.9246161	-0.1093565	0.0454802	0.0696080	0.1002703	0.0565128	-0.0155502	-0.0453982
latitude	-0.9246161	1.0000000	0.0118991	-0.0366668	-0.0669828	-0.1089973	-0.0717742	-0.0796263	-0.1446382
housing_median_age	-0.1093565	0.0118991	1.0000000	-0.3606283	-0.3204510	-0.2957873	-0.3027680	-0.1182777	0.1064320
total_rooms	0.0454802	-0.0366668	-0.3606283	1.0000000	0.9303795	0.8572813	0.9189915	0.1978815	0.1332941
total_bedrooms	0.069608	-0.06	-0.320451	0.9303795	1.0000000	0.877746	0.979728	-0.0077	0.0496862

	0	69828	0			7	3	228	
population	0.1002703	-0.1089973	-0.2957873	0.8572813	0.8777467	1.0000000	0.9071859	0.0050866	-0.0252997
households	0.0565128	-0.0717742	-0.3027680	0.9189915	0.9797283	0.9071859	1.0000000	0.0134339	0.0648935
median_income	-0.0155502	-0.0796263	-0.1182777	0.1978815	-0.0077228	0.0050866	0.0134339	1.0000000	0.6883555
median_house_value	-0.0453982	-0.1464382	0.1064320	0.1332941	0.0496862	-0.0252997	0.0648935	0.6883555	1.0000000

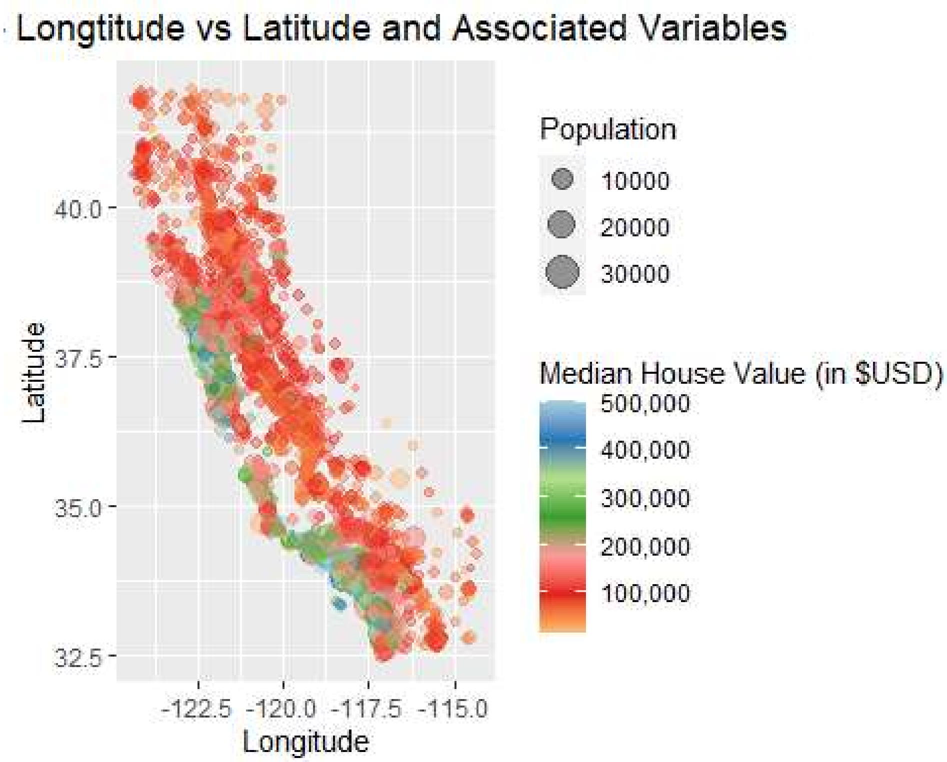
We noticed there is collinearity between (households and total_bedrooms) & (households and total_rooms). We will keep this in mind and explore the data further

Training and Test Data

We took 80% of the data as training data and used seed to be consistent with the results.

```
set.seed(100)
totalnrows = nrow(data)

x = sample(totalnrows, floor(totalnrows * .80) )
train_data = data[x, ]
test_data = data[-x, ]

plot_map = ggplot(train_data,
                  aes(x = longitude, y = latitude, color =
median_house_value,
                      hma = housing_median_age, tr = total_rooms, tb =
total_bedrooms,
                      hh = households, mi = median_income)) +
              geom_point(aes(size = population), alpha = 0.4) +
              xlab("Longitude") +
              ylab("Latitude") +
              ggtitle("Data Map - Longtitude vs Latitude and Associated
Variables") +
              theme(plot.title = element_text(hjust = 0.5)) +
              scale_color_distiller(palette = "Paired", labels = comma) +
```

```
                labs(color = "Median House Value (in $USD)", size =
"Population")
plot_map
```



The graph above shows distribution of Median house value based on population and Latitude. It gives us fair distribution of values across geographical area.

Additive Model

```
#Training additive Model
model_add = lm(median_house_value ~ ., data = train_data)
summary(model_add)

##
## Call:
## lm(formula = median_house_value ~ ., data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -554770  -42731  -10480   28801  761094
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -2.274e+06  9.846e+04 -23.096  < 2e-16 ***
## longitude                 -2.681e+04  1.140e+03 -23.512  < 2e-16 ***
## latitude                  -2.540e+04  1.123e+03 -22.609  < 2e-16 ***
## housing_median_age         1.102e+03  4.885e+01  22.557  < 2e-16 ***
```

```
## total_rooms               -5.850e+00  8.771e-01  -6.670 2.64e-11 ***
## total_bedrooms             9.931e+01  7.737e+00  12.835  < 2e-16 ***
## population                -3.732e+01  1.183e+00 -31.533  < 2e-16 ***
## households                 4.817e+01  8.405e+00   5.731 1.02e-08 ***
## median_income              3.905e+04  3.740e+02 104.386  < 2e-16 ***
## ocean_proximityINLAND     -3.966e+04  1.954e+03 -20.295  < 2e-16 ***
## ocean_proximityISLAND      1.531e+05  3.068e+04   4.990 6.09e-07 ***
## ocean_proximityNEAR BAY   -4.041e+03  2.122e+03  -1.904  0.05691 .
## ocean_proximityNEAR OCEAN  5.578e+03  1.744e+03   3.199  0.00138 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68490 on 16333 degrees of freedom
## Multiple R-squared:  0.6471, Adjusted R-squared:  0.6469
## F-statistic:  2496 on 12 and 16333 DF,  p-value: < 2.2e-16
```

```
summary(model_add)$adj.r.squared
```

```
## [1] 0.6468567
```

By analyzing p-value of all Beta variable in Additive model, we can say that we fail to reject that Null Hypothesis that Beta value of any variable is Zero. Hence all variables are playing important role in prediction of House Median Income. And Adjusted R squared value of Model is 64.6%

Interaction Model

```
model_int = lm(median_house_value ~ . ^ 2, data = train_data)
summary(model_int)$adj.r.squared
```

```
## [1] 0.7025208
```

In interaction model we can see an increment of Model performance by Adjusted R Squared which is 70.3%

Testing Interaction model with respect to Additive Model

```
anova(model_int, model_add)
```

```
## Analysis of Variance Table
##
## Model 1: median_house_value ~ (longitude + latitude + housing_median_age +
##     total_rooms + total_bedrooms + population + households +
##     median_income + ocean_proximity)^2
## Model 2: median_house_value ~ longitude + latitude + housing_median_age +
##     total_rooms + total_bedrooms + population + households +
##     median_income + ocean_proximity
##   Res.Df        RSS  Df    Sum of Sq      F    Pr(>F)
## 1  16277 6.4322e+13
## 2  16333 7.6621e+13 -56 -1.2299e+13 55.575 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P-value of test is 2.2e-16 which is very less hence we can consider Interactive models is better than additive model

Model Improvement Using AIC and BIC

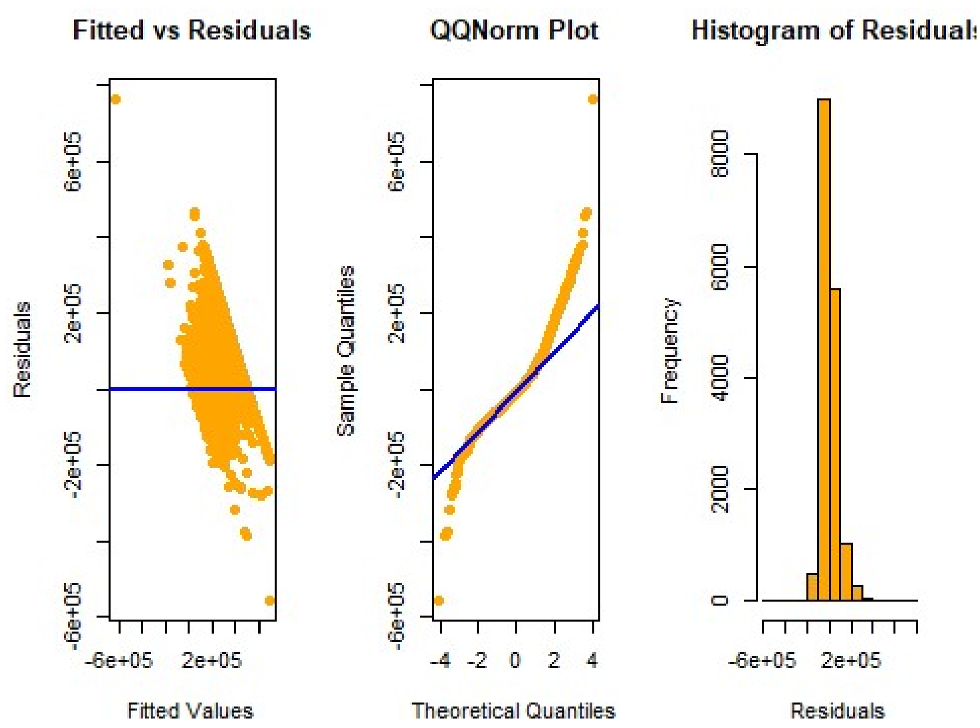
```
model_add_aic = step(model_add, direction = "backward", trace = 0)
summary(model_add_aic)$adj.r.squared

## [1] 0.6468567

model_add_bic = step(model_add, direction = "backward", trace = 0, k =
log(nrow(train_data)))
summary(model_add_bic)$adj.r.squared

## [1] 0.6468567

model_int_aic = step(model_int, direction = "backward", trace = 0)
summary(model_int_aic)$adj.r.squared

## [1] 0.7025212

model_int_bic = step(model_int, direction = "backward", trace = 0, k =
log(nrow(train_data)))
summary(model_int_bic)$adj.r.squared

## [1] 0.7019587

beginning_mods_results = data.frame(
  "Total Predictors" =
    c("Additive Model" = extractAIC(model_add)[1],
      "Interaction Model" = extractAIC(model_int)[1],
      "AIC_additive Model" = extractAIC(model_add_aic)[1],
      "AIC_Int Model" = extractAIC(model_int_aic)[1],
      "BIC_additive Model" = extractAIC(model_add_bic)[1],
      "BIC_Int Model" = extractAIC(model_int_bic)[1]),
  "AIC" =
    c("Additive Model" = extractAIC(model_add)[2],
       "Interaction Model" = extractAIC(model_int)[2],
      "AIC_additive Model" = extractAIC(model_add_aic)[2],
      "AIC_Int Model" = extractAIC(model_int_aic)[2],
       "BIC_additive Model" = extractAIC(model_add_bic)[2],
      "BIC_Int Model" = extractAIC(model_int_bic)[2]),
  "Adj R-Squared" =
    c("Additive Model" = summary(model_add)$adj.r.squared,
      "Interaction Model" = summary(model_int)$adj.r.squared,
      "AIC_additive Model" =summary(model_add_aic)$adj.r.squared,
      "AIC_Int Model" = summary(model_int_aic)$adj.r.squared,
       "BIC_additive Model" =summary((model_add_bic))$adj.r.squared,
      "BIC_Int Model" = summary(model_int_bic)$adj.r.squared))


kable(beginning_mods_results, align = c("c", "r"))
```

	Total.Predictors	AIC	Adj.R.Squared
Additive Model	13	364021.2	0.6468567
Interaction Model	69	361273.2	0.7025208
AIC_additive Model	13	364021.2	0.6468567
AIC_Int Model	64	361268.2	0.7025212
BIC_additive Model	13	364021.2	0.6468567
BIC_Int Model	56	361291.2	0.7019587

We see that the model with the best (i.e., lowest) AIC is Interaction Model, with a score of 361268.2. But we will work further to enhance performance of model.

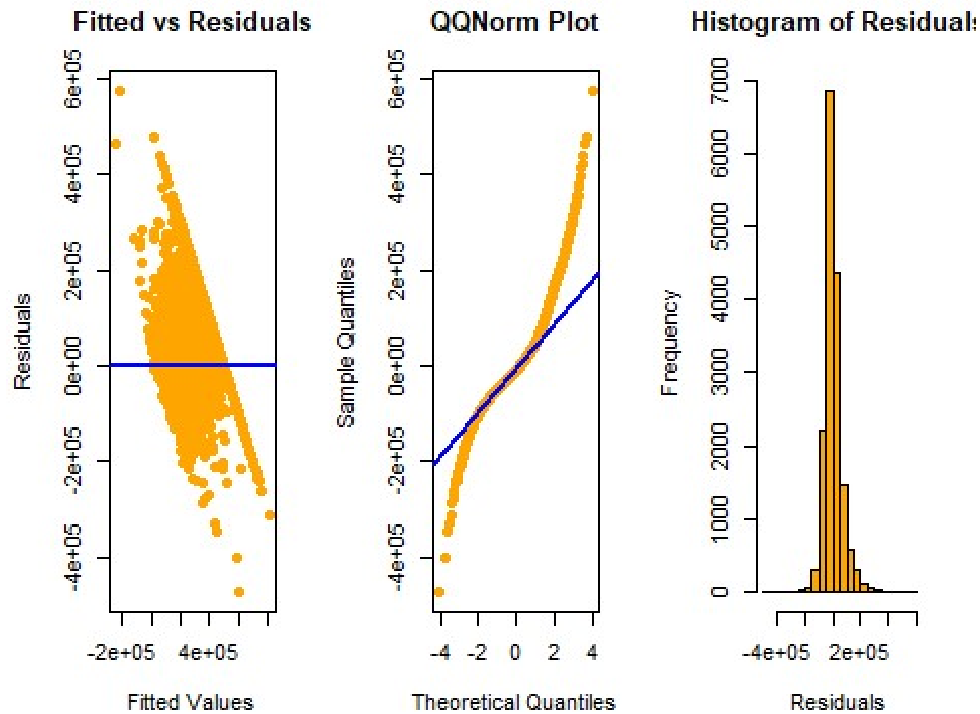
```
diagnostics = function(model, alpha = .05, pointcol = "orange", linecol =
"blue", plots = TRUE, tests = TRUE, pointtype = 16) {
    if (plots == TRUE) {
        par(mfrow = c(1, 3))
        plot(
                fitted(model),
                resid(model),
                pch = pointtype,
                xlab = "Fitted Values",
                ylab = "Residuals",
                main = "Fitted vs Residuals",
                col = pointcol
            )
        abline(h = 0, lwd = 2, col = linecol)

        qqnorm(
                resid(model),
                pch = pointtype,
                main = "QQNorm Plot",
                col = pointcol
            )
        qqline(
                resid(model),
                lwd = 2,
                col = linecol
                )
        hist(
            resid(model),
            main = "Histogram of Residuals",
            col = pointcol,
            xlab = "Residuals",
            ylab = "Frequency"
            )
    }
    if (tests == TRUE) {
        ks_test = ks.test(resid(model),y='pnorm',alternative='two.sided')
```

```
        bp_test = bptest(model)
        test_results = data.frame(
          "Kolmogorov-Smirnov  Test" =
            c("Test Statistic" = round(ks_test$statistic, 5),
              "P-Value" = ks_test$p.value,
              "Result" = ifelse(ks_test$p.value < alpha, "Reject", "Fail To
Reject")),
          "Breusch-Pagan Test" =
            c("Test Statistic" = round(bp_test$statistic, 5),
              "P-Value" = bp_test$p.value,
              "Result" = ifelse(bp_test$p.value < alpha, "Reject", "Fail To
Reject")))

        kable(t(test_results), col.names = c("Test Statistic", "P-Value",
"Decision"))
    }
}

diagnostics(model_add)
```



	Test Statistic	P-Value	Decision
Kolmogorov.Smirnov..Test	0.5802	0	Reject
Breusch.Pagan.Test	813.50284	2.10341410745741e-166	Reject

```
diagnostics(model_int)
```

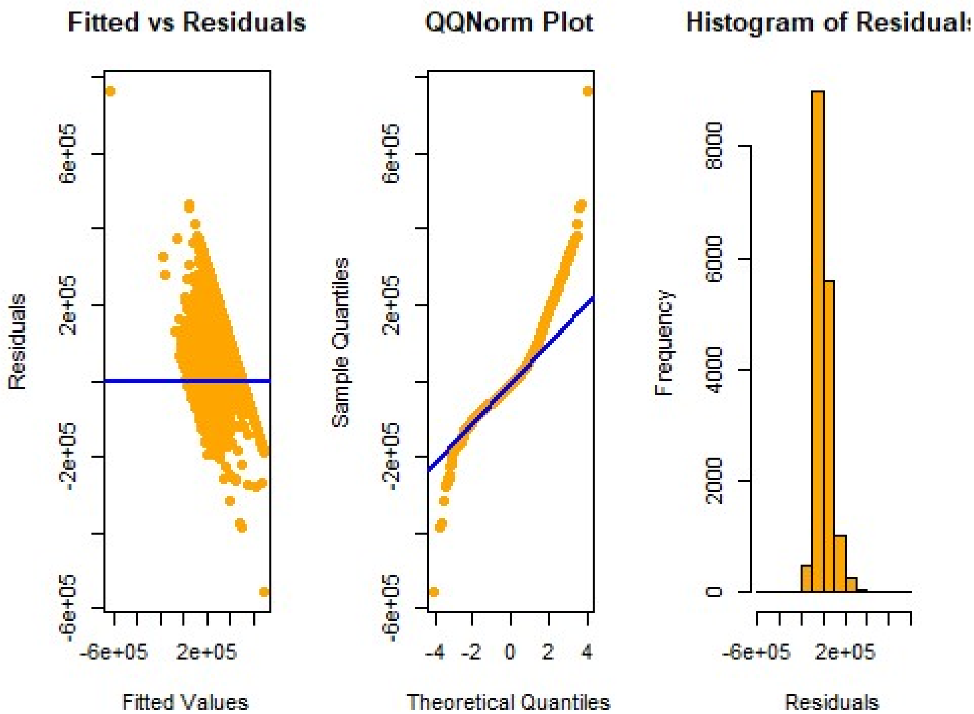


	Test Statistic	P-Value	Decision
Kolmogorov.Smirnov..Test	0.57806	0	Reject
Breusch.Pagan.Test	1698.68853	7.46002686123318e-310	Reject

```r
diagnostics(model_add_aic)
```

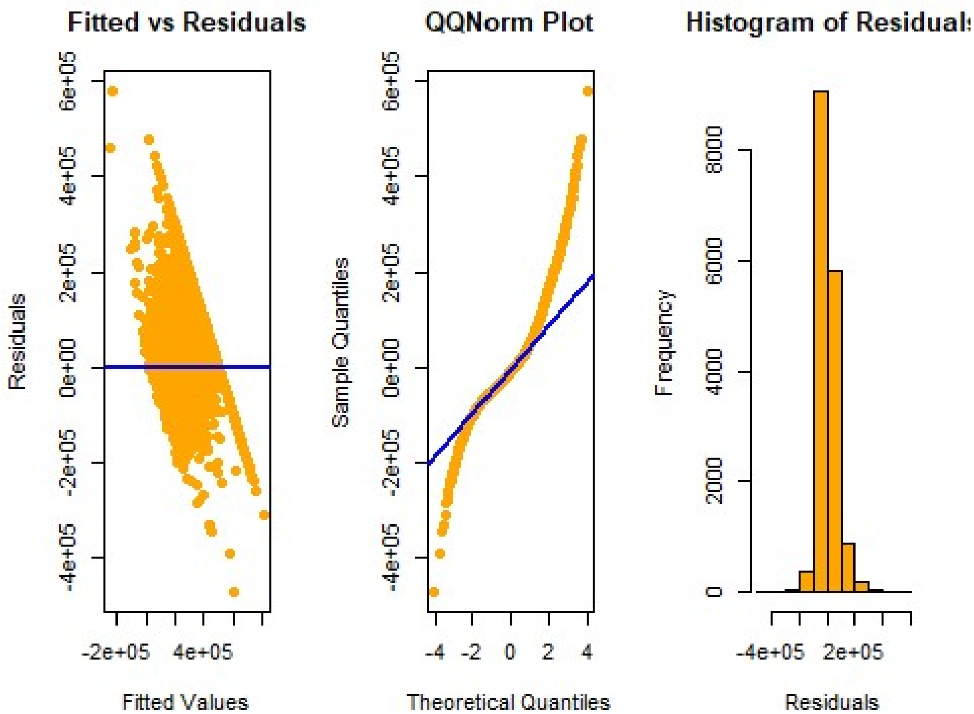
	Test Statistic	P-Value	Decision
Kolmogorov.Smirnov..Test	0.5802	0	Reject
Breusch.Pagan.Test	813.50284	2.10341410745741e-166	Reject

```r
diagnostics(model_add_bic)
```



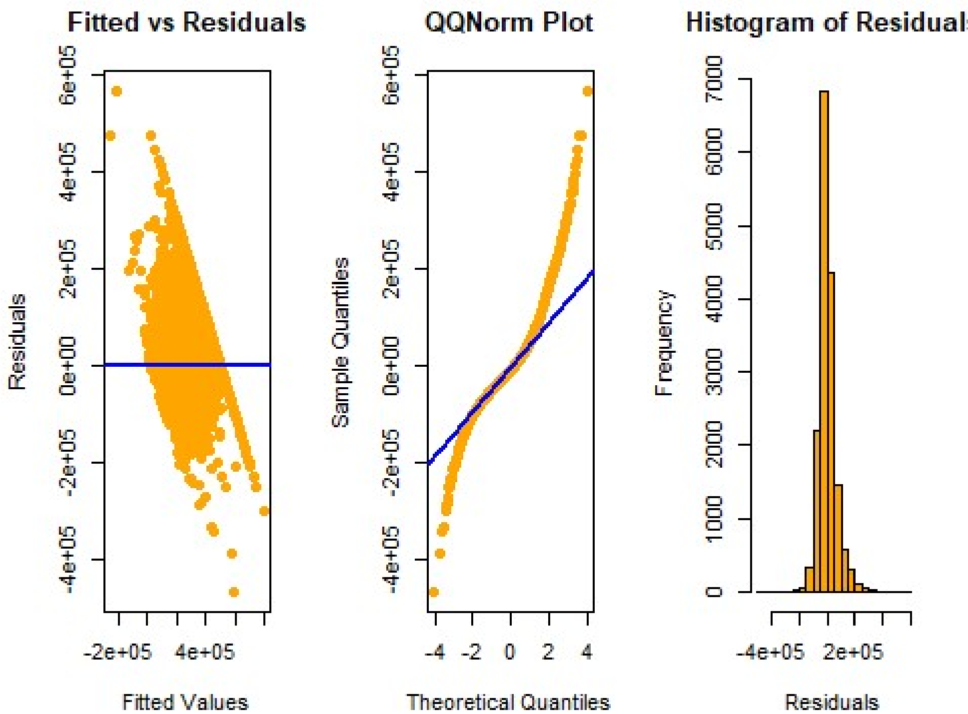
	Test Statistic	P-Value	Decision
Kolmogorov.Smirnov..Test	0.5802	0	Reject
Breusch.Pagan.Test	813.50284	2.10341410745741e-166	Reject

```
diagnostics(model_int_aic)
```



	Test Statistic	P-Value	Decision
Kolmogorov.Smirnov..Test	0.57873	0	Reject
Breusch.Pagan.Test	1589.03891	1.76669099023126e-290	Reject

```
diagnostics(model_int_bic)
```



	Test Statistic	P-Value	Decision
Kolmogorov.Smirnov..Test	0.57776	0	Reject
Breusch.Pagan.Test	1436.29723	3.15291874921026e-264	Reject

```
x = ks.test(x=rnorm(10^4),y='pnorm',alternative='two.sided')

x$p.value

## [1] 0.9685713
```

We can see that all above models do not have Equal variance and residual in Normal form. Hence we need to improve model.

Kolmogorov–Smirnov test- In statistics, the Kolmogorov–Smirnov test is a nonparametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution, or to compare two samples. Note- We tried using shapiro.test first, but the test that did not work considering the size of the dataset.

Model Improvement

Now, we will calculate the cooks distance and will remove outliers and high influential values.

```
value = cooks.distance(model_add)
sum(value > 4 / length(resid(model_add)))
```

```
## [1] 885
```

```r
model_new_add =  lm(median_house_value ~ ., data = train_data, subset = value
<= (4 / nrow(train_data)))

model_new_int =  lm(median_house_value ~ .^2, data = train_data, subset =
value <= (4 / nrow(train_data)))

model_new_add_AIC = step(model_new_add, direction = "backward", trace = 0)

model_new_int_AIC = step(model_new_int, direction = "backward", trace = 0)
```

Based on the new data data values, we will again train the models and calculate ADJ R Squared and LOOCV values (Leave-One-Out Cross-Validation)

Results

When we initially calculated the AdjustedR2 value the results were not very convincing as we had low ADJ R Squared value for all the models. However,when we remove the outliers and high influential values using the cooks distance we got better results.

```r
Result = data.frame(
        "Additive Model" =c("LOOCV" = sqrt(mean((resid(model_new_add) / (1 -
hatvalues(model_new_add))) ^ 2)),
              "ADJ R Squared" = summary(model_new_add)$adj.r.squared,
              "Test RMSE" = sqrt(mean((test_data$median_house_value -
predict(model_new_add, newdata = test_data))^2)),
               "SE" = summary(model_new_add)$sigma),

        "Interaction Model" = c( "LOOCV" = sqrt(mean((resid(model_new_int) /
(1 - hatvalues(model_new_int))) ^ 2)),
              "ADJ R Squared" = summary(model_new_int)$adj.r.squared,
              "Test RMSE" = sqrt(mean((test_data$median_house_value -
predict(model_new_int, newdata = test_data))^2)),
              "SE" = summary(model_new_int)$sigma),

        "Additive Model AIC" = c( "LOOCV" =
sqrt(mean((resid(model_new_add_AIC) / (1 - hatvalues(model_new_add_AIC))) ^
2)),
              "ADJ R Squared" = summary(model_new_add_AIC)$adj.r.squared,
              "Test RMSE" = sqrt(mean((test_data$median_house_value -
predict(model_new_add_AIC, newdata = test_data))^2)),
              "SE" = summary(model_new_add_AIC)$sigma),

        "Interaction Model AIC" = c( "LOOCV" =
sqrt(mean((resid(model_new_int_AIC) / (1 - hatvalues(model_new_int_AIC))) ^
2)),
              "ADJ R Squared" = summary(model_new_int_AIC)$adj.r.squared,
              "Test RMSE" = sqrt(mean((test_data$median_house_value -
predict(model_new_int_AIC, newdata = test_data))^2)),
```

```r
              "SE" = summary(model_new_int_AIC)$sigma)
)
 kable(t(Result))
```

	LOOCV	ADJ R Squared	Test RMSE	SE
Additive.Model	53496.07	0.7424312	69798.42	53477.11
Interaction.Model	49852.24	0.7847754	64271.04	48884.06
Additive.Model.AIC	53496.07	0.7424312	69798.42	53477.11
Interaction.Model.AIC	49773.83	0.7847864	64266.58	48882.81

Based on the results, we can say that Interaction.Model.AIC is having better ADJ R Squared(0.7847864) among all model and hence can be considered best among the given model. Also, this is also better than the previous all models discussed(without removal of outliers") where the max adjusted R2 value was 0.7025212 for"AIC_Int Model"

Discussion

As shown above table, our selected model "model_new_int_AIC" (AIC of Interaction Model) has lowest LOOCV RMSE in all models i.e 49773.83 and better Adjusted R squared around 78.5%. We have an average Standard Error 48882.21 that means on average, our model's predicted housing price will be ± 48882.21 in comparison to the actual price.

Above table also shows Model performance on Test Data. "Test RMSE" columns shows root squared error for Test Data and "model_new_int_AIC" showed lowest RMSE in all i.e. 64266.58.

Our aim was to predict Housing price for California Region and based on above observation we can conclude that No individual predictor determines the cost of the house however interaction of predictor make up better prediction model.

Appendix

- Names of Team : Team Engineer
- Original Data :

```r
head(data, 5)

##   longitude latitude housing_median_age total_rooms total_bedrooms
population
## 1   -122.23    37.88                 41         880            129
322
## 2   -122.22    37.86                 21        7099           1106
2401
## 3   -122.24    37.85                 52        1467            190
496
## 4   -122.25    37.85                 52        1274            235
```

```
558
## 5   -122.25    37.85                 52        1627            280
565
##   households median_income median_house_value ocean_proximity
## 1        126        8.3252             452600        NEAR BAY
## 2       1138        8.3014             358500        NEAR BAY
## 3        177        7.2574             352100        NEAR BAY
## 4        219        5.6431             341300        NEAR BAY
## 5        259        3.8462             342200        NEAR BAY
```

- Outlier and high influence points removal by Cook's Distance
- Best Model

```
summary(model_new_int_AIC)

##
## Call:
## lm(formula = median_house_value ~ longitude + latitude +
housing_median_age +
##     total_rooms + total_bedrooms + population + households +
##     median_income + ocean_proximity + longitude:latitude +
longitude:housing_median_age +
##     longitude:total_rooms + longitude:total_bedrooms +
longitude:households +
##     longitude:median_income + longitude:ocean_proximity +
latitude:housing_median_age +
##     latitude:total_rooms + latitude:total_bedrooms +
latitude:median_income +
##     latitude:ocean_proximity + housing_median_age:total_rooms +
##     housing_median_age:total_bedrooms + housing_median_age:population +
##     housing_median_age:households + housing_median_age:median_income +
##     housing_median_age:ocean_proximity + total_rooms:population +
##     total_rooms:households + total_rooms:median_income +
total_rooms:ocean_proximity +
##     total_bedrooms:population + total_bedrooms:households +
total_bedrooms:median_income +
##     total_bedrooms:ocean_proximity + population:households +
##     population:median_income + population:ocean_proximity +
households:median_income +
##     median_income:ocean_proximity, data = train_data, subset = value <=
##     (4/nrow(train_data)))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -237366  -30345   -5336   25048  380747
##
## Coefficients:
##                                                  Estimate Std. Error t value
## (Intercept)                                    -2.750e+06  8.339e+05  -3.298
## longitude                                      -9.984e+03  7.428e+03  -1.344
```

```
## latitude                                         2.239e+05  2.512e+04   8.913
## housing_median_age                              -7.573e+04  7.008e+03 -10.805
## total_rooms                                      1.336e+03  1.811e+02   7.379
## total_bedrooms                                  -5.715e+03  9.617e+02  -5.942
## population                                      -1.851e+01  5.009e+00  -3.696
## households                                      -1.420e+03  4.243e+02  -3.347
## median_income                                   -9.931e+05  6.630e+04 -14.980
## ocean_proximityINLAND                           -1.816e+04  2.252e+05  -0.081
## ocean_proximityNEAR BAY                         -1.746e+07  1.133e+06 -15.405
## ocean_proximityNEAR OCEAN                       -1.123e+06  2.803e+05  -4.005
## longitude:latitude                               1.489e+03  1.929e+02   7.717
## longitude:housing_median_age                    -9.418e+02  8.100e+01 -11.627
## longitude:total_rooms                            1.650e+01  2.146e+00   7.690
## longitude:total_bedrooms                        -7.631e+01  1.122e+01  -6.799
## longitude:households                            -1.211e+01  3.605e+00  -3.360
## longitude:median_income                         -1.226e+04  7.841e+02 -15.635
## longitude:ocean_proximityINLAND                  2.115e+03  2.668e+03   0.793
## longitude:ocean_proximityNEAR BAY               -1.751e+05  1.004e+04 -17.436
## longitude:ocean_proximityNEAR OCEAN             -1.081e+04  3.349e+03  -3.229
## latitude:housing_median_age                     -1.049e+03  7.975e+01 -13.150
## latitude:total_rooms                             1.772e+01  2.213e+00   8.006
## latitude:total_bedrooms                         -9.453e+01  1.161e+01  -8.144
## latitude:median_income                          -1.247e+04  8.105e+02 -15.385
## latitude:ocean_proximityINLAND                   5.402e+03  2.770e+03   1.950
## latitude:ocean_proximityNEAR BAY                -1.033e+05  7.674e+03 -13.456
## latitude:ocean_proximityNEAR OCEAN              -4.645e+03  3.505e+03  -1.325
## housing_median_age:total_rooms                  -6.179e-01  7.637e-02  -8.091
## housing_median_age:total_bedrooms                4.127e+00  8.110e-01   5.088
## housing_median_age:population                   -1.545e+00  1.102e-01 -14.023
## housing_median_age:households                    3.626e+00  9.027e-01   4.017
## housing_median_age:median_income                 2.722e+02  2.586e+01  10.527
## housing_median_age:ocean_proximityINLAND         5.413e+02  1.343e+02   4.032
## housing_median_age:ocean_proximityNEAR BAY      -7.071e+02  1.472e+02  -4.805
## housing_median_age:ocean_proximityNEAR OCEAN -1.586e+02  1.222e+02  -1.298
## total_rooms:population                          -7.594e-03  1.117e-03  -6.798
## total_rooms:households                           2.191e-02  3.402e-03   6.438
## total_rooms:median_income                        3.430e+00  3.098e-01  11.071
## total_rooms:ocean_proximityINLAND               -1.709e+01  3.284e+00  -5.205
## total_rooms:ocean_proximityNEAR BAY              1.336e+01  3.486e+00   3.832
## total_rooms:ocean_proximityNEAR OCEAN            5.971e+00  2.592e+00   2.303
## total_bedrooms:population                        2.371e-02  7.014e-03   3.381
## total_bedrooms:households                       -1.449e-01  1.549e-02  -9.353
## total_bedrooms:median_income                     1.636e+01  4.962e+00   3.296
## total_bedrooms:ocean_proximityINLAND             2.877e+01  1.821e+01   1.580
## total_bedrooms:ocean_proximityNEAR BAY          -5.653e+01  1.664e+01  -3.397
## total_bedrooms:ocean_proximityNEAR OCEAN        -3.727e+01  1.301e+01  -2.865
## population:households                            2.690e-02  5.100e-03   5.275
## population:median_income                        -2.564e+00  7.097e-01  -3.613
## population:ocean_proximityINLAND                 2.349e+01  2.552e+00   9.206
## population:ocean_proximityNEAR BAY              -7.087e+00  4.641e+00  -1.527
```

```
## population:ocean_proximityNEAR OCEAN          1.764e+00  2.993e+00   0.590
## households:median_income                     -2.377e+01  5.783e+00  -4.110
## median_income:ocean_proximityINLAND           5.920e+03  1.278e+03   4.630
## median_income:ocean_proximityNEAR BAY        -3.243e+03  1.202e+03  -2.697
## median_income:ocean_proximityNEAR OCEAN      -9.153e+02  9.691e+02  -0.944
##                                              Pr(>|t|)
## (Intercept)                                  0.000977 ***
## longitude                                    0.178903
## latitude                                      < 2e-16 ***
## housing_median_age                            < 2e-16 ***
## total_rooms                                  1.67e-13 ***
## total_bedrooms                               2.87e-09 ***
## population                                   0.000220 ***
## households                                   0.000818 ***
## median_income                                 < 2e-16 ***
## ocean_proximityINLAND                        0.935727
## ocean_proximityNEAR BAY                       < 2e-16 ***
## ocean_proximityNEAR OCEAN                    6.24e-05 ***
## longitude:latitude                           1.26e-14 ***
## longitude:housing_median_age                  < 2e-16 ***
## longitude:total_rooms                        1.56e-14 ***
## longitude:total_bedrooms                     1.09e-11 ***
## longitude:households                         0.000781 ***
## longitude:median_income                       < 2e-16 ***
## longitude:ocean_proximityINLAND              0.427824
## longitude:ocean_proximityNEAR BAY             < 2e-16 ***
## longitude:ocean_proximityNEAR OCEAN          0.001246 **
## latitude:housing_median_age                   < 2e-16 ***
## latitude:total_rooms                         1.27e-15 ***
## latitude:total_bedrooms                      4.13e-16 ***
## latitude:median_income                        < 2e-16 ***
## latitude:ocean_proximityINLAND               0.051162 .
## latitude:ocean_proximityNEAR BAY              < 2e-16 ***
## latitude:ocean_proximityNEAR OCEAN           0.185183
## housing_median_age:total_rooms               6.37e-16 ***
## housing_median_age:total_bedrooms            3.65e-07 ***
## housing_median_age:population                 < 2e-16 ***
## housing_median_age:households                5.91e-05 ***
## housing_median_age:median_income              < 2e-16 ***
## housing_median_age:ocean_proximityINLAND     5.57e-05 ***
## housing_median_age:ocean_proximityNEAR BAY   1.56e-06 ***
## housing_median_age:ocean_proximityNEAR OCEAN 0.194304
## total_rooms:population                       1.10e-11 ***
## total_rooms:households                       1.24e-10 ***
## total_rooms:median_income                     < 2e-16 ***
## total_rooms:ocean_proximityINLAND            1.97e-07 ***
## total_rooms:ocean_proximityNEAR BAY          0.000128 ***
## total_rooms:ocean_proximityNEAR OCEAN        0.021267 *
## total_bedrooms:population                    0.000723 ***
## total_bedrooms:households                     < 2e-16 ***
```

```
## total_bedrooms:median_income                     0.000982 ***
## total_bedrooms:ocean_proximityINLAND             0.114236
## total_bedrooms:ocean_proximityNEAR BAY           0.000684 ***
## total_bedrooms:ocean_proximityNEAR OCEAN         0.004174 **
## population:households                            1.35e-07 ***
## population:median_income                         0.000304 ***
## population:ocean_proximityINLAND                  < 2e-16 ***
## population:ocean_proximityNEAR BAY               0.126741
## population:ocean_proximityNEAR OCEAN             0.555492
## households:median_income                         3.98e-05 ***
## median_income:ocean_proximityINLAND              3.68e-06 ***
## median_income:ocean_proximityNEAR BAY            0.007008 **
## median_income:ocean_proximityNEAR OCEAN          0.344960
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48880 on 15404 degrees of freedom
## Multiple R-squared:  0.7856, Adjusted R-squared:  0.7848
## F-statistic:  1008 on 56 and 15404 DF,  p-value: < 2.2e-16
```