CALIFORNIA HOUSING: LINEAR REGRESSION AND ANALYSIS

CALDER KATYAL

S&DS 230: SPRING 2024

PROFESSOR JONATHAN REUNING-SCHERER

INTRODUCTION

In this project we take a look at the California Housing Dataset, a collection of information concerning homes in California derived from the 1990 U.S. Census. In the Census, California is divided up into numerous "blocks," a geographical unit used to group together houses. The dataset includes the following attributes, which are measured with respect to a given block group:

Type	Base Unit	Attribute	Information
Continuous; numeric	100,000 USD (1990)	MedInc	Median income in block group
Discrete; integer	Years	HouseAge	Median house age in block group
Continuous; numeric	Room	AveRooms	Average number of rooms per household
Continuous; numeric	Bedroom	AveBedrms	Average number of bedrooms per household
Discrete; integer	Person	Population	Block group population
Continuous; numeric	Person	AveOccup	Average number of household members
Continuous; numeric	Decimal Degrees	Latitude	Block group latitude
Continuous; numeric	Decimal Degrees	Longitude	Block group longitude
Continuous; numeric	100,000 USD (1990)	MedHouseVal	Median house value in block group

It is a common task in Machine Learning to predict this dataset's target variable **MedHouseVal** using the eight predictors. Such a model is generally very complicated. In this project, we explore and visualize the data and see if we can fit a decent linear regression model. The task turns out to be extremely difficult due to the lack of normality in predictors and the target, the inherent nonlinear nature of the data, and the very unfortunate fact discovered throughout the process that our target **MedHouseVal** is capped at a certain threshold and thus threatens the quality of our model.

DATA CLEANING AND NEW VARIABLE CREATION

Data cleaning for this project is comparatively mild due to the numeric nature of the data. The dataset lacks missing values and is in overall strong condition. Cleaning mostly entailed ensuring that the columns were treated as numerics and verifying the structure of our data. In this verification, it was discovered that **MedHouseVal** was capped.

Due to the lack of data cleaning required, and the lack of categorical predictors in the model, additional predictors were created. They are detailed below:

LONG_BIN, LAT_BIN, AND COAST

Because it is well-known that wealth in California is strongly correlated to geographic location, and **longitude** and **latitude**, being continuous, are not particularly great geospatial predictors in this instance, we decided to create three new variables: **Long_bin**, **Lat_bin**, and **Coast**. Only the latter was used in any model. To create a categorical geospatial predictor we use a technique called *binning*. We create 10 "bins" corresponding to equal ranges of latitude and longitude respectively and calculate the average house value for all blocks located inside a given bin. Our bins are of the form $B(X,Y) = \{(X,Y) \mid 1 \leq X \leq 10,\ 1 \leq Y \leq 10,\ (X,Y) \in \mathbb{Z}\}$, where (1, 1) corresponds with the bottom-left corner of California if it were to be extended to a bounding rectangle. We created a visual of the bins via *ggplot2*, *maps*, and *dplyr*. Each bin is colored with a darkness corresponding to its average house value:

We then create the indicator factor variable **Coast** by creating a new column **Coast** in our dataframe and assigning a value of "Yes" if (**Lat_bin**, **Long_bin**) ∈ {(10, 10), (9, 9), (8, 1), (7, 1), (7, 2), (6, 2), (6, 3), (5, 2), (5, 3), (4, 3), (4, 4), (3, 4), (3, 5), (2, 5), (2, 6), (2, 7), (1, 6), (1, 7), (1, 8)} and assign it to "No" otherwise (these are the bins that either touch California's coast or an island near the coast). The hope is that our new predictor **Coast** will add additional information to our model.

INCOME_LEVEL

During the course of attempting to fit a linear regression model it became apparent that the predictor **MedInc** had a very irregular distribution that threatened the quality of most models. This distribution, combined with the fact that MedHouseVal was capped at an arbitrary level, somewhat compromises its ability as a predictor. Applying simple transformations such as *log* and *sqrt* did little to improve its quality. While we use the predictor unchanged in some of our models, we also create a new factor variable **IncomeLevel** as follows by breaking up **Med_Inc** into three quartiles "Low" "Medium" and "High." While some information is lost in the process, the new variable can be a valuable predictor in simple models for our target.

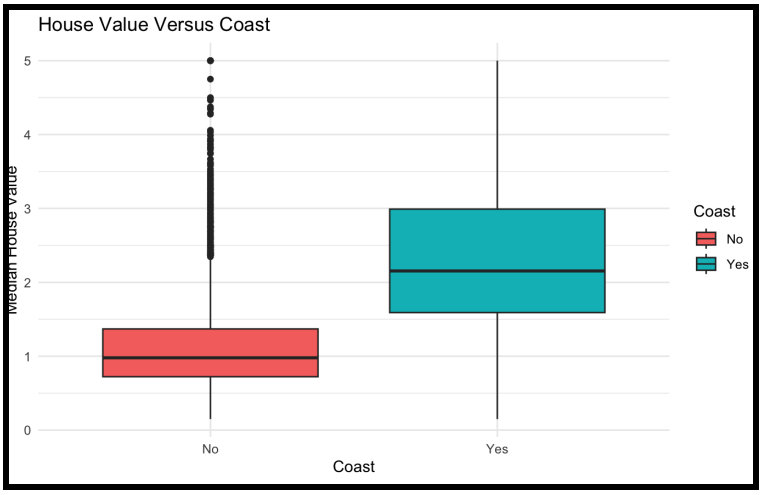
Thus, we now have two new predictors:

Categorical (factor)	N/A	Coast	"Yes" for coastal; "No" for non-coastal
Categorical (factor)	N/A	IncomeLevel	Splits MedHouseVal into three equal intervals "Low," "Medium," and "High"

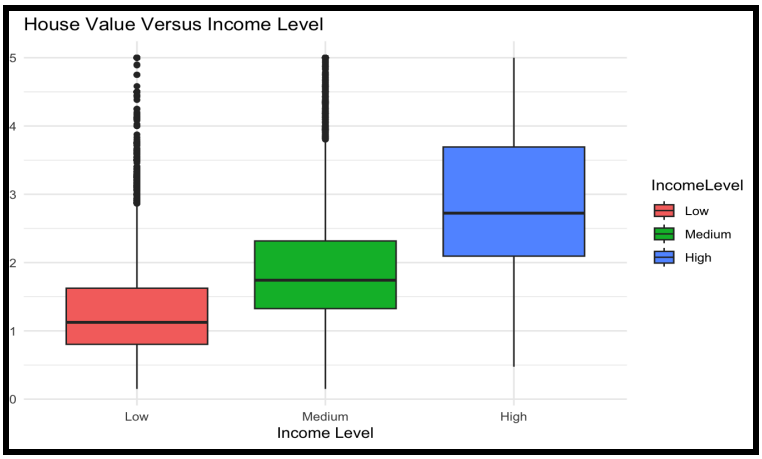
PRELIMINARY PLOTTING

We first investigate the nature of our data via boxplots, scatterplots, and histograms. The preliminary plotting will help us realize the inherent complexity of the data and begin to understand the relationship between the predictors and the target.

BOX PLOTS



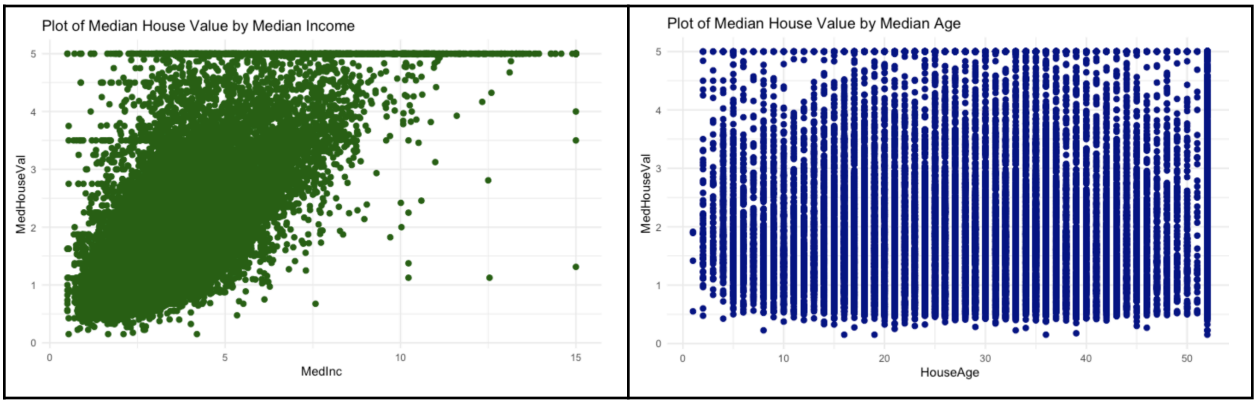
This boxplot validates our assumption that blocks near the coast generally correspond with higher **MedHouseVal**. In the no coast plot, we see a narrow IQR and many outliers, which is indicative of significant variance within no coast blocks. Conversely, the IQR of the coast boxplot is large and we don't see outliers.

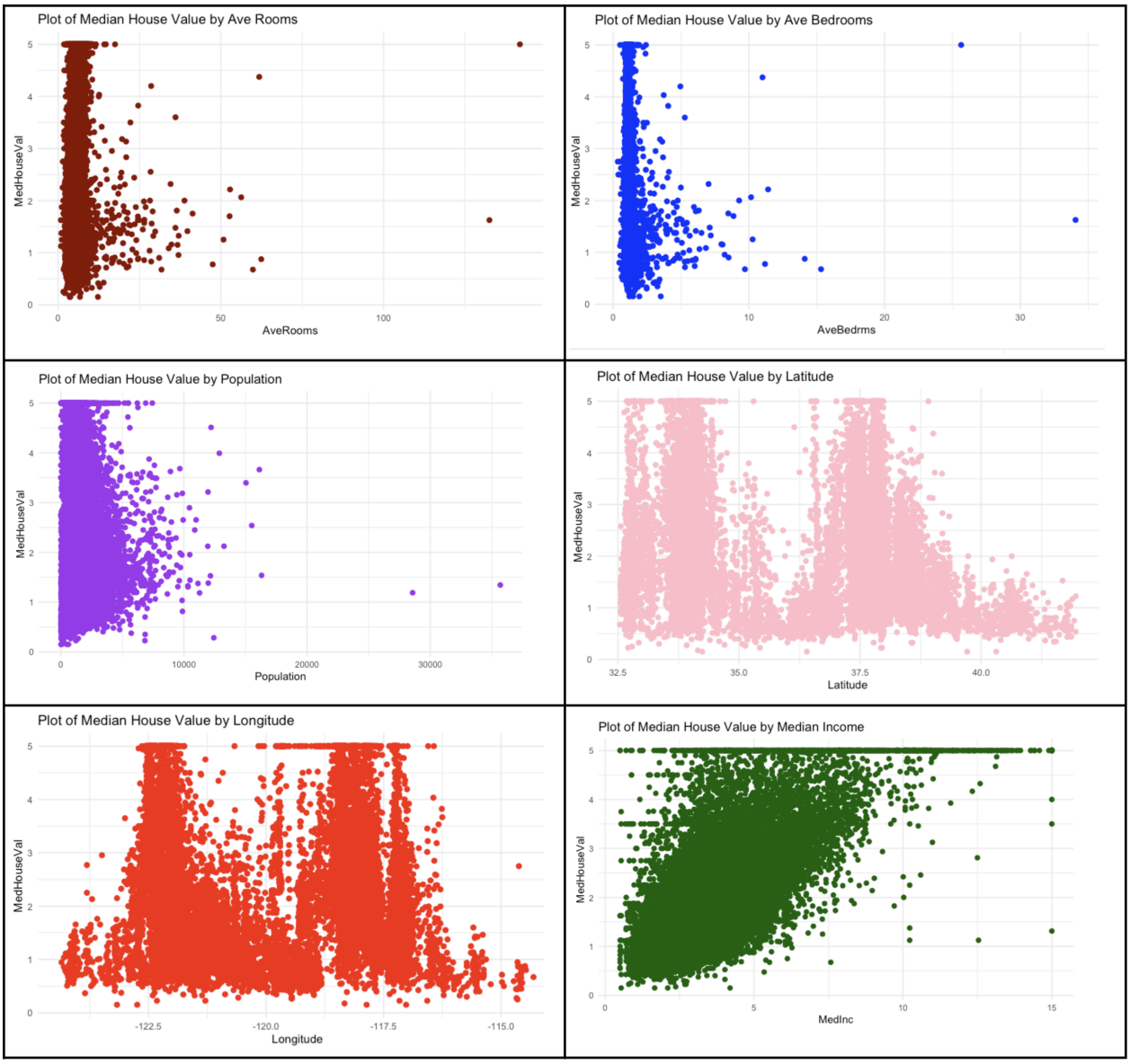


We see that median house value significantly rises with income level as expected. However, we have a large number of outliers for "Low" and "Medium", suggesting a lack of normality. We further observe that the income level "High" has a large IQR, indicative of the large variance in wealth at the upper end of the spectrum.

SCATTER PLOTS

We provide scatter plots for the continuous predictors:

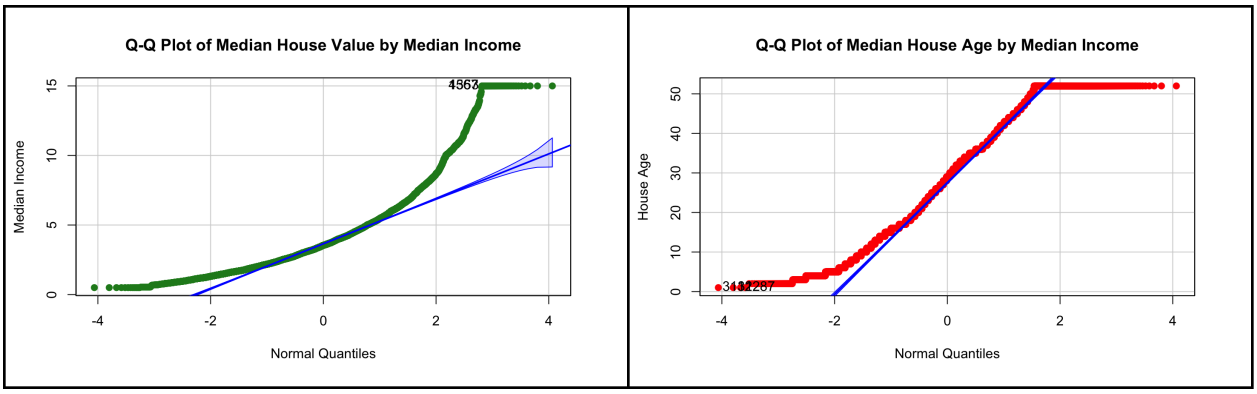




None of the scatterplots show any observable trends except for possibly Median House Value vs. Median Income (the plot in dark green) which shows a mildly linear relationship between the two variables. The main takeaway from these plots is that our data is extremely far from normality and the simple transformations are unlikely to help.

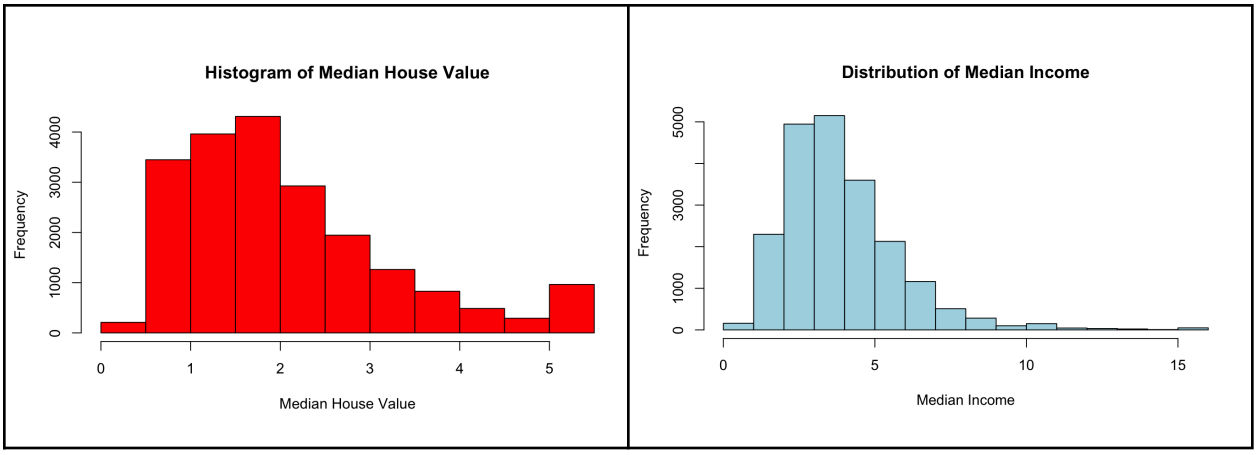
NORMAL QUANTILE PLOTS

We investigate how far our data is off from normality. We provide two result here; the rest exhibit a similar lack in normality:



As seen above, the data quickly curves away from the blue envelope of normality. This indicates that our data is significantly skewed. Although normality of the predictors is not expected for our linear model, the entropy of the data is a harbinger of the difficulties to come.

HISTOGRAMS



We first create a histogram of the target **MedHouseVal**. The data is significantly skewed to the right but lacks many values at the end signifying lowest median house value. Thus, our target is not normally distributed. It is also worth looking more into the distribution of **MedInc**, as it was the only predictor that we saw in the scatter plots to be vaguely linear, and it also intuitively would seem to correlate with **MedHouseVal**. Its histogram is significantly right skewed due to the presence of a few blocks with extreme wealth. Thus **MedInc** lacks normality, and the presence of such significant outliers will complicate our modeling.

BASIC TESTING

We conduct preliminary tests for our data to assess the nature of our predictors. Further testing will follow once a basic model has been fit.

T-TEST

We employ a Welch Two Sample t-test to assess whether there is a significant difference in the mean of **MedHouseVal** under the **Coast** factor:

```
Welch Two Sample t-test

data:  MedHouseVal by Coast
t = -109, df = 18452, p-value <2e-16
alternative hypothesis: true difference in means between group No and group Yes is not equal
to 0
95 percent confidence interval:
 -1.3 -1.3
sample estimates:
 mean in group No mean in group Yes
              1.1               2.4
```

The null hypothesis for our t-test is that there is no significant difference in the means of **MedHouseVal** under the groups coast and no coast. As we get a mean of 1.1 for the no coast group and a mean of 2.4 for the coast group with p < 2e-16 $< \alpha = 0.05$, we can reject the null hypothesis and conclude that coast significantly impacts the mean of **MedHouseVal**.

CORRELATION TEST

We first provide a basic correlation matrix for the continuous predictors.

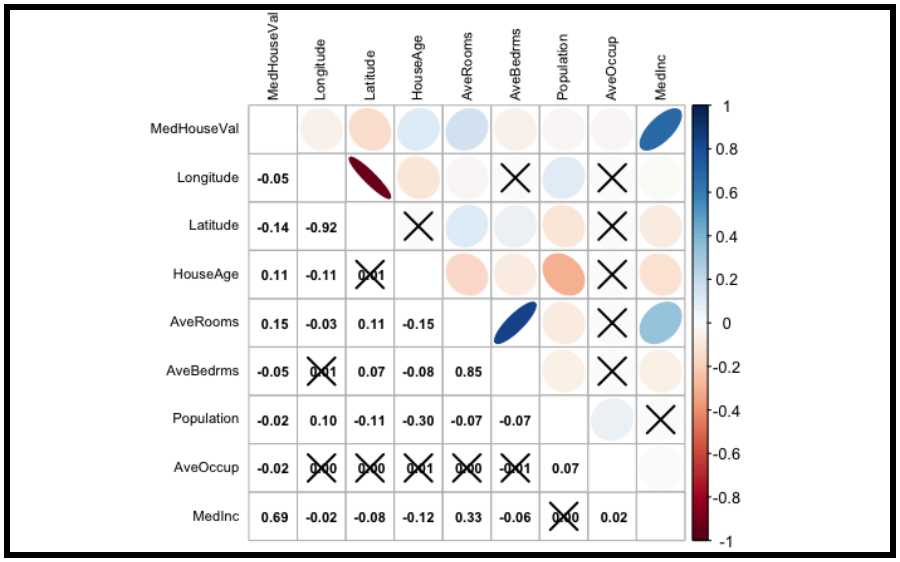
	MedHouseVal	Longitude	Latitude	HouseAge	AveRooms	AveBedrms	Population	AveOccup	MedInc
MedHouseVal	1.00	-0.05	-0.14	0.11	0.15	-0.05	-0.02	-0.02	0.69
Longitude	-0.05	1.00	-0.92	-0.11	-0.03	0.01	0.10	0.00	-0.02
Latitude	-0.14	-0.92	1.00	0.01	0.11	0.07	-0.11	0.00	-0.08
HouseAge	0.11	-0.11	0.01	1.00	-0.15	-0.08	-0.30	0.01	-0.12
AveRooms	0.15	-0.03	0.11	-0.15	1.00	0.85	-0.07	0.00	0.33
AveBedrms	-0.05	0.01	0.07	-0.08	0.85	1.00	-0.07	-0.01	-0.06
Population	-0.02	0.10	-0.11	-0.30	-0.07	-0.07	1.00	0.07	0.00
AveOccup	-0.02	0.00	0.00	0.01	0.00	-0.01	0.07	1.00	0.02
MedInc	0.69	-0.02	-0.08	-0.12	0.33	-0.06	0.00	0.02	1.00

The only predictor that appears to be highly correlated with **MedHouseVal** is **MedInc**, which has a positive correlation coefficient of 0.69. Note that some of the predictors are correlated with each other, which suggests that we may have issues with multicollinearity in the future. We next provide p-values for our pairwise correlations as well as lower and upper confidence intervals:

```
$p
            MedHouseVal Longitude Latitude HouseAge AveRooms AveBedrms Population AveOccup  MedInc
MedHouseVal     0.0e+00   3.9e-11  2.9e-96  2.8e-52 7.6e-107   1.9e-11    4.0e-04  6.5e-04 0.0e+00
Longitude       3.9e-11   0.0e+00  0.0e+00  8.6e-55  7.6e-05   5.5e-02    8.1e-47  7.2e-01 2.9e-02
```

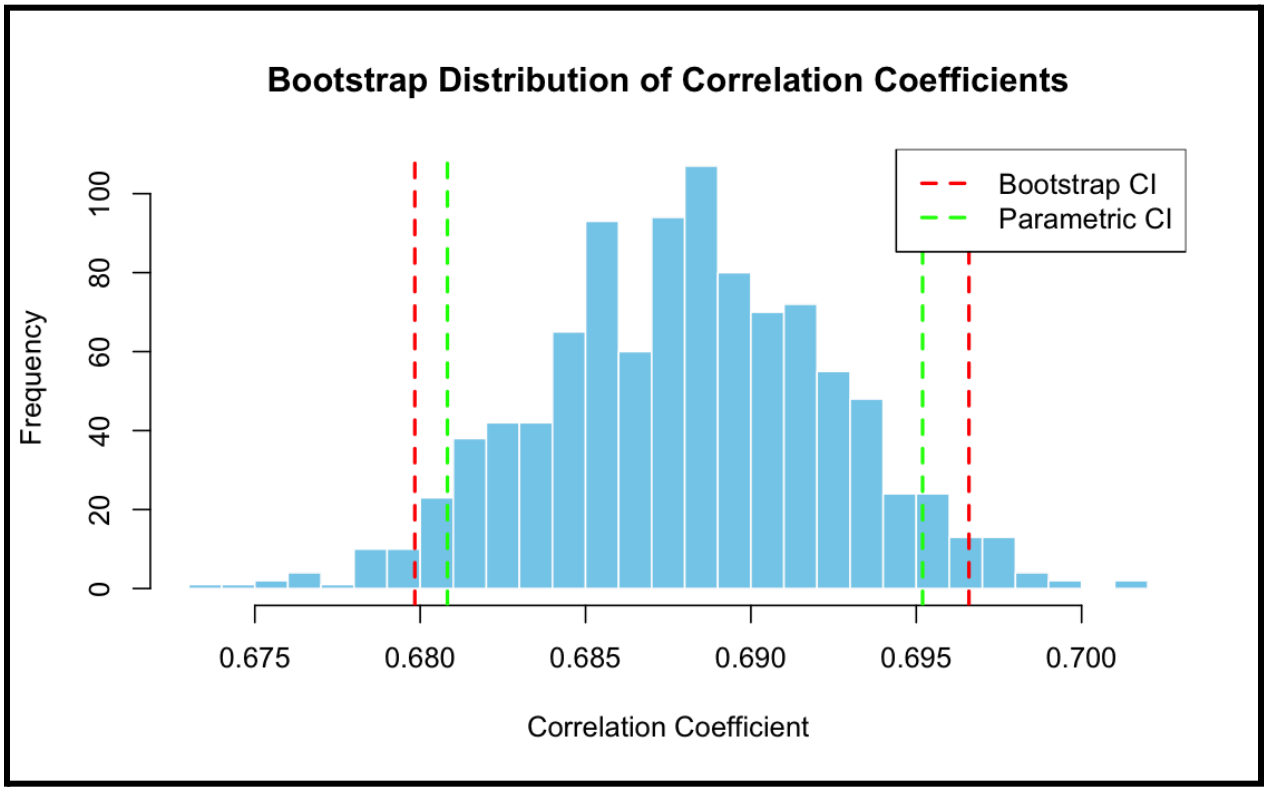
```
Latitude        2.9e-96   0.0e+00   0.0e+00   1.1e-01   5.0e-53   1.1e-23    2.3e-55   7.3e-01 1.6e-30
HouseAge        2.8e-52   8.6e-55   1.1e-01   0.0e+00 1.0e-108   4.8e-29    0.0e+00   5.8e-02 5.2e-66
AveRooms       7.6e-107   7.6e-05   5.0e-53 1.0e-108   0.0e+00   0.0e+00    2.8e-25   4.9e-01 0.0e+00
AveBedrms       1.9e-11   5.5e-02   1.1e-23   4.8e-29   0.0e+00   0.0e+00    1.7e-21   3.7e-01 4.6e-19
Population      4.0e-04   8.1e-47   2.3e-55   0.0e+00   2.8e-25   1.7e-21    0.0e+00   9.3e-24 4.9e-01
AveOccup        6.5e-04   7.2e-01   7.3e-01   5.8e-02   4.9e-01   3.7e-01    9.3e-24   0.0e+00 7.0e-03
MedInc          0.0e+00   2.9e-02   1.6e-30   5.2e-66   0.0e+00   4.6e-19    4.9e-01   7.0e-03 0.0e+00


$lowCI
            MedHouseVal Longitude Latitude HouseAge AveRooms AveBedrms Population AveOccup  MedInc
MedHouseVal       1.000   -0.0596  -0.1575  0.09211    0.139   -0.0603    -0.0383 -0.03737  0.6808
Longitude        -0.060    1.0000  -0.9266 -0.12166   -0.041   -0.0003     0.0862 -0.01117 -0.0288
Latitude         -0.157   -0.9266   1.0000 -0.00247    0.093    0.0561    -0.1222 -0.01128 -0.0934
HouseAge          0.092   -0.1217  -0.0025  1.00000   -0.167   -0.0913    -0.3086 -0.00045 -0.1325
AveRooms          0.139   -0.0412   0.0929 -0.16657    1.000    0.8437    -0.0858 -0.01849  0.3147
AveBedrms        -0.060   -0.0003   0.0561 -0.09129    0.844    1.0000    -0.0798 -0.01982 -0.0756
Population       -0.038    0.0862  -0.1222 -0.30864   -0.086   -0.0798     1.0000  0.05627 -0.0088
AveOccup         -0.037   -0.0112  -0.0113 -0.00045   -0.018   -0.0198     0.0563  1.00000  0.0051
MedInc            0.681   -0.0288  -0.0934 -0.13246    0.315   -0.0756    -0.0088  0.00512  1.0000

$uppCI
            MedHouseVal Longitude Latitude HouseAge AveRooms AveBedrms Population AveOccup  MedInc
MedHouseVal       1.000   -0.0323   -0.131    0.119   0.1652   -0.0331     -0.011 -0.0101  0.6952
Longitude        -0.032    1.0000   -0.923   -0.095  -0.0139    0.0270      0.113  0.0161 -0.0015
Latitude         -0.131   -0.9227    1.000    0.025   0.1199    0.0833     -0.095  0.0160 -0.0662
HouseAge          0.119   -0.0947    0.025    1.000  -0.1399   -0.0642     -0.284  0.0268 -0.1056
AveRooms          0.165   -0.0139    0.120   -0.140   1.0000    0.8514     -0.059  0.0088  0.3390
AveBedrms        -0.033    0.0270    0.083   -0.064   0.8514    1.0000     -0.053  0.0075 -0.0484
Population       -0.011    0.1133   -0.095   -0.284  -0.0586   -0.0526      1.000  0.0834  0.0185
AveOccup         -0.010    0.0161    0.016    0.027   0.0088    0.0075      0.083  1.0000  0.0324
MedInc            0.695   -0.0015   -0.066   -0.106   0.3390   -0.0484      0.018  0.0324  1.0000
```

All p-values are less than $\alpha = 0.05$ and so suggest the statistical significance of our correlations. We finally provide a correlation plot to summarize the information:



BOOTSTRAP CONFIDENCE INTERVALS

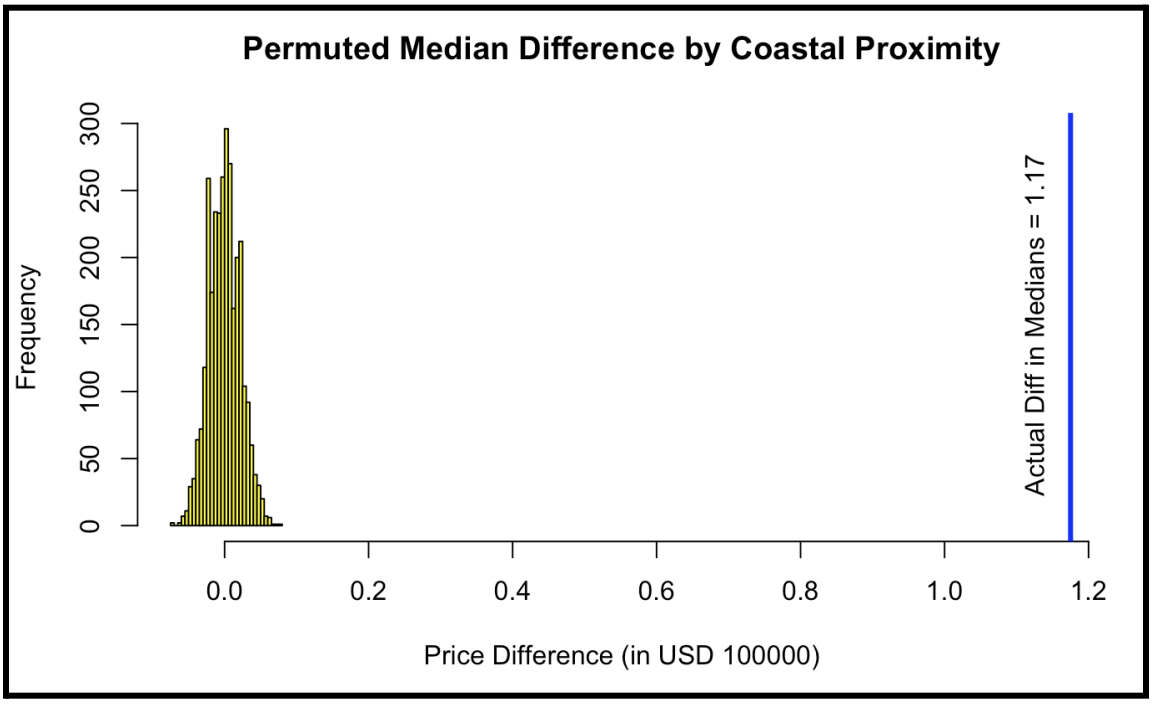
We use a bootstrapped (eg. “sampled”) confidence interval for the correlation between **MedHouseVal** and **MedInc**. This is suggested because the usual parametric confidence interval makes assumptions about normality which likely do not hold in our case. This confidence interval is the result of repeatedly sampling rows from our data frame and calculating the correlation coefficient between **MedHouseVal and MedInc** and then creating a 95% quantile. The parametric and bootstrapped confidence intervals are plotted together in the following histogram:



Both confidence intervals are extremely close to each other (the particular distribution of the data makes the bootstrapped confidence interval marginally wider, but this is negligible). Thus our results confirm the suggested correlation between the two variables.

PERMUTATION TEST

While not strictly necessary due to the prior significant results, it is interesting to see whether **MedHouseVal**, when grouped by the categorical variable **Coast**, significantly deviates from the overall median value of MedHouseVal calculated across the entire dataset irrespective of coast categorization.



As we can see, the actual difference in medians is significantly higher than all of the values in our vector of permuted differences; thus we have a p-value of 0 for the test. The permutation test works by forming a large number random binary permutation of the values in **Coast**. For each permutation we take the difference in the medians of **MedHouseVal** for **Coast** = "Yes" and **Coast** = "No" and put it into a vector; finally, we plot this vector as a histogram in yellow and overlay a line representing the difference in medians of **MedHouseVal** at **Coast** = "Yes" and **Coast** = "No" for the whole dataset. The significant difference between the expected and calculated values suggests that **Coast** is far from random and approximately bifurcates **MedHouseVal**.

LINEAR REGRESSION

Now we employ three methods to attempt to fit a linear regression model for our target. After the first model is fit, we make the unfortunate discovery that our target **MedHouseVal** is capped at a maximum value of 5.00001 with a count of 965 blocks. This was extremely problematic for our models; even if we attempted to remove these capped values from our data frame, we would still face a similar issue as our target could not reach its natural continuum of values. This results in a seeming lack of normality of our residuals. We provide our three models below:

METHOD: BEST SUBSETS REGRESSION

We employ best subsets of linear regression to our data, which compares all possible models based on a set of predictors. For this model we include only the original predictors provided in our dataset as well as our indicator variable **Coast**; we do not include **IncomeLevel** as there are already a plethora of highly skewed, irregular continuous variables in our set of predictors. **Lat_bin** and **Long_bin** are not considered to prevent overfitting. A summary of models considered by best subsets regression are provided below, where "*" indicates that a predidctor is used in the model:

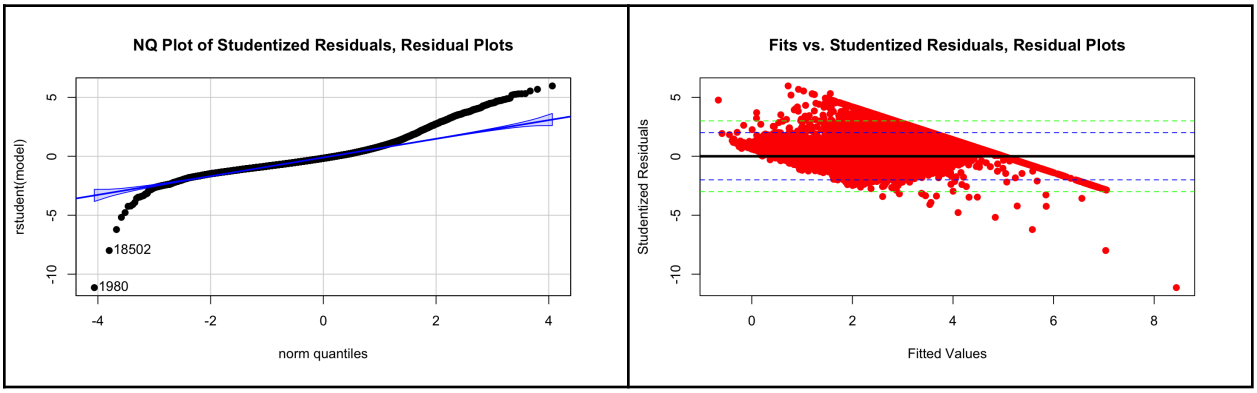
```
  MedInc HouseAge AveRooms AveBedrms Population AveOccup Latitude Longitude CoastYes
1   "*"     " "      " "       " "        " "       " "      " "       " "       " "
2   "*"     " "      " "       " "        " "       " "      " "       " "       "*"
3   "*"     "*"      " "       " "        " "       " "      " "       " "       "*"
4   "*"     " "      " "       " "        " "       " "      "*"       "*"       "*"
5   "*"     "*"      " "       " "        " "       " "      "*"       "*"       "*"
6   "*"     "*"      " "       "*"        " "       " "      "*"       "*"       "*"
7   "*"     "*"      "*"       "*"        " "       " "      "*"       "*"       "*"
8   "*"     "*"      "*"       "*"        " "       "*"      "*"       "*"       "*"
9   "*"     "*"      "*"       "*"        "*"       "*"      "*"       "*"       "*"
```

Clearly the best predictor is **MedInc**, which is expected upon looking back at our preliminary scatter plots. Next, we identify the best model generated by best subsets regression under the Bayesian Information Criterion (BIC):

```
Call:
lm(formula = MedHouseVal ~ ., data = housing_temp_bic)

Residuals:
   Min     1Q Median     3Q    Max
-5.325 -0.471 -0.127  0.312  4.060

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.61e+01   8.80e-01  -29.67  < 2e-16 ***
MedInc       3.61e-01   3.06e-03  117.95  < 2e-16 ***
HouseAge     8.69e-03   4.26e-04   20.41  < 2e-16 ***
AveRooms     1.68e-02   2.28e-03    7.36  1.9e-13 ***
AveOccup    -3.47e-03   4.86e-04   -7.13  1.0e-12 ***
Latitude    -2.75e-01   1.09e-02  -25.30  < 2e-16 ***
Longitude   -3.00e-01   1.05e-02  -28.49  < 2e-16 ***
CoastYes     4.41e-01   2.01e-02   21.91  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.72 on 20632 degrees of freedom
Multiple R-squared:  0.605,  Adjusted R-squared:  0.605
F-statistic: 4.52e+03 on 7 and 20632 DF,  p-value: <2e-16
```

Our model attains a R-squared and Adjusted R-squared of 0.605, suggesting that it accounts for approximately 60% of the variance of the target (which is quite decent). The model is significant with a p-value of less than 2e-16. Unfortunately, this model includes

eight predictors, comprising seven of our original predictors (**population** is excluded) as well as our categorical predictor **Coast**. All are statistically significant; **MedInc**, **HouseAge**, **AveBedrms**, and **Coast** have positive coefficients while the others are negative. In general this makes sense as all these predictors intuitively align with greater wealth and thus higher **MedHouseVal**. **Latitude** and **Longitude** have negative coefficients due to the geographical wealth distribution in California; the other negative predictors **AveRooms** and **AveOccup** have coefficients with very small magnitude and so their seemingly arbitrary direction is not too problematic. The non-intercept predictors with the greatest magnitude coefficients are **MedInc**, **AveBedrms**, **Latitude**, **Longitude**, and **Coast**. Having eight predictors in our model means that it is quite complex and does not align nicely with our intuition. It furthermore indicates that we might have difficulty attaining normality for the residuals, a condition often required of linear models. They are plotted here:



From the first plot it is clear that the residuals are not approximately normal as points at the tails curve away from the blue envelope of normality. In the second plot we see that, instead of being randomly dispersed from the interval (-2, 2) as often desired, our studentized residuals form a distinctive sharped-edged parallelogram shape with many outliers. While this suggests heteroscedacity (non-constant variance of the residuals), we must not forget that our target is capped. This accounts for the sharp edges in our residuals and their distinctive shape. Analyses of similar datasets with uncapped targets tend to produce approximately normal distributions for residuals, especially when considering a variable such as **MedHouseVal** which is in many cases approximately normal itself. It is for these reasons that it is acceptable to further analyze the model.

Nevertheless, it is worth seeing if we can improve the quality of the residuals. To do this, we utilize the Box-Cox Transformation.

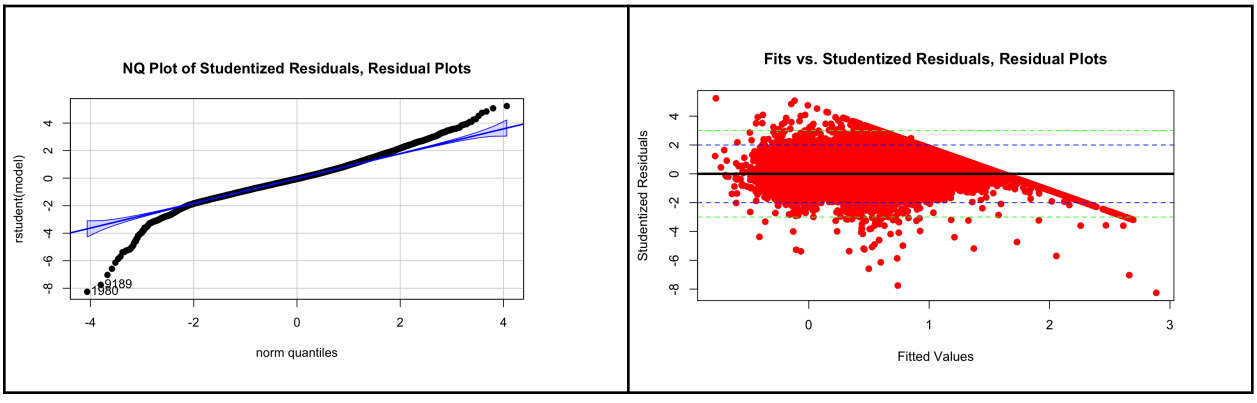
METHOD: BOX-COX TRANSFORMATION

As noted in our discussion of the **MedHouseVal** histogram, our target variable lacks

normality. Applying a Box-Cox transformation to our previous model can rectify this problem. We calculate the λ associated with the Box-Cox Transform to be $0.1414141 \approx 0$, which signifies it is worth attempting a logarithmic transformation to our target. To this end, we create **LogMedHouseVal**. We thus form the following model:

```
Call:
lm(formula = LogMedHouseVal ~ ., data = housing_temp_boxcox)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5637 -0.2169 -0.0189  0.2019  1.7087

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.35e+01   4.15e-01  -32.46  < 2e-16 ***
MedInc       1.54e-01   1.44e-03  106.92  < 2e-16 ***
HouseAge     1.22e-03   2.01e-04    6.08  1.2e-09 ***
AveRooms     1.12e-02   1.08e-03   10.44  < 2e-16 ***
AveOccup    -1.41e-03   2.29e-04   -6.16  7.4e-10 ***
Latitude    -1.35e-01   5.13e-03  -26.39  < 2e-16 ***
Longitude   -1.50e-01   4.97e-03  -30.07  < 2e-16 ***
CoastYes     3.96e-01   9.50e-03   41.66  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.34 on 20632 degrees of freedom
Multiple R-squared:  0.639,  Adjusted R-squared:  0.639
F-statistic: 5.22e+03 on 7 and 20632 DF,  p-value: <2e-16
```

The model has the same significant predictors as our prior model and a similar R-squared and p-value. We proceed to plot the residuals:



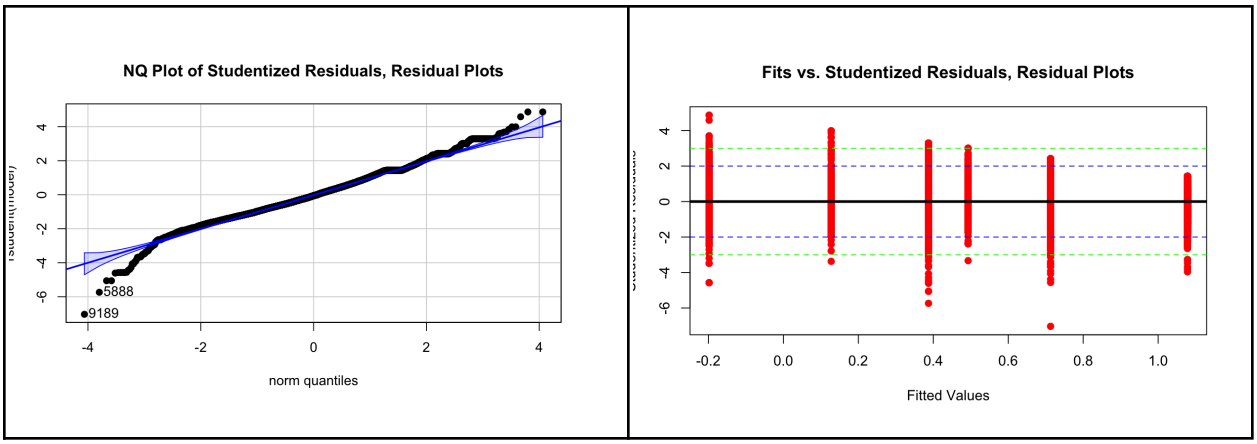
With no apparent improvement in the normality of the residuals, we neglect to further consider this model. Applying a logarithmic transformation to the target but not to any of our continuous predictors (some of which are already scaled) does not align with our intuition and might distort our conclusions.

METHOD: CATEGORICAL-ONLY MODELING

If we want to attempt to attain normality of the residuals (which has debatable significance due to the capped target variable), a method that includes a continuous predictor will likely fail. This is because any association between the continuous predictor and the target will be immediately severed when the target reaches its capped value. By using categorical predictors, we can somewhat reduce this effect, while slightly compromising model quality. We still use **LogMedHouseVal** as our target in an effort to ensure normality of the residuals. A model containing our two categorical predictors **Coast** and **IncomeLevel** is given here:

```
Call:
lm(formula = LogMedHouseVal ~ IncomeLevel + Coast, data = housing)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6102 -0.2574 -0.0228  0.2413  1.8076

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.19821    0.00573   -34.6   <2e-16 ***
IncomeLevelMedium  0.32575    0.00641    50.8   <2e-16 ***
IncomeLevelHigh    0.69115    0.00663   104.2   <2e-16 ***
CoastYes           0.58547    0.00612    95.7   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.37 on 20636 degrees of freedom
Multiple R-squared:  0.574,  Adjusted R-squared:  0.574
F-statistic: 9.26e+03 on 3 and 20636 DF,  p-value: <2e-16
```

We have that both predictors are significant as well as the model with associated p-values of less than 2.2e-16. However, we now have a slightly decreased multiple and adjusted R-squared of 0.574, suggesting that our model now captures less than 60% of the variance of the target. This is not necessarily bad, especially when considering the capped nature of the target. Plots of the residuals are below:

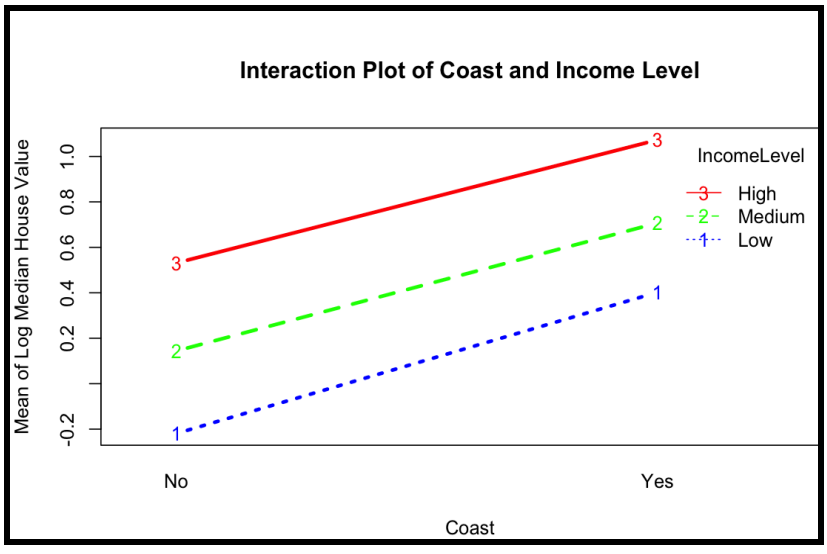


We have finally obtained approximate normality of the residuals to a modest degree. While the residuals certainly are not perfectly normally distributed (the left tail of the normal quantile plot curves away from the blue envelope of normality), this is a considerable and sufficient degree of normality considering the capped target. Although there are several outliers, the vast majority of studentized residuals fall in the range (-4,4), of which most are in the range (-2,2).

PREDICTOR INTERACTION AND ANOVA

We now seek to better understand the significance of our categorical predictors to the target, which will help us assess the strength of our categorical-only model. We start by analyzing the interaction between **Coast** and **IncomeLevel**. We will continue to use the transformed target **LogMedHouseVal**

PREDICTOR INTERACTION



From our interaction plot we see that there is significant interaction between Coast and Income Level at all levels. This justifies the notion that their interaction might be a significant predictor in our model.

TWO-WAY ANOVA

Next, we conduct a two-way ANOVA to decompose the total variability of the target into the categoricals **Coast** and **IncomeLevel** as well as their interaction **Coast:IncomeLevel**.

```
Anova Table (Type III tests)

Response: LogMedHouseVal
                  Sum Sq    Df F value  Pr(>F)
(Intercept)          138     1  1000.6 < 2e-16 ***
Coast                642     1  4658.5 < 2e-16 ***
IncomeLevel          401     2  1453.9 < 2e-16 ***
Coast:IncomeLevel      4     2    15.2 2.6e-07 ***
Residuals           2846 20634
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

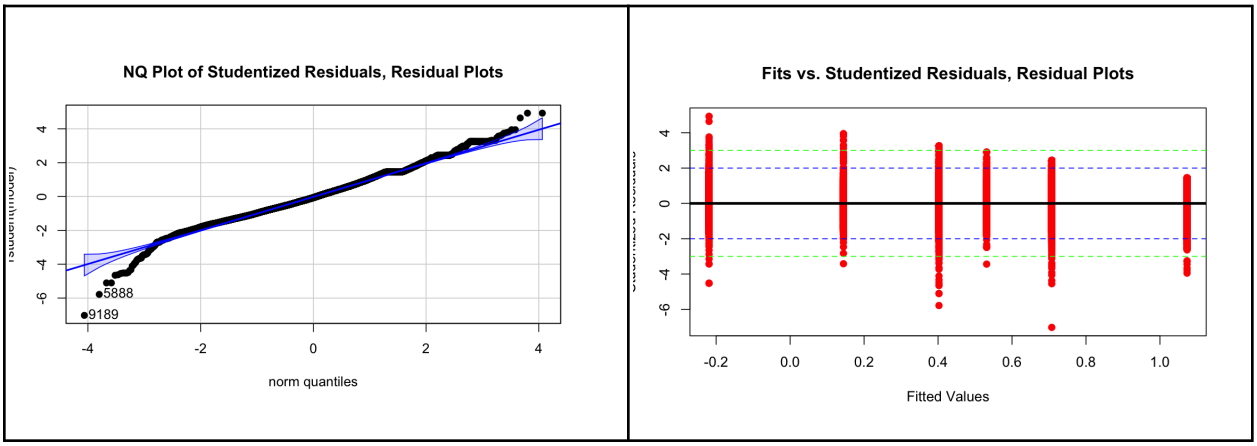
We see that both categorical variables as well as their interaction are all significant with associated p-values of less than $\alpha = 0.05$. We have a sum of squares of 2846, suggesting that a considerable (but not overwhelming) portion of the variance in **MedHouseVal** is not explained by the predictors.

Thus, we can improve our simple categorical-only model from before by adding in an interaction term. The results are summarized below:

```
Call:
lm(formula = LogMedHouseVal ~ Coast + IncomeLevel + Coast * IncomeLevel)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6046 -0.2564 -0.0238  0.2405  1.8286

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              -0.21916    0.00693  -31.63  < 2e-16 ***
Coast1                    0.62168    0.00911   68.25  < 2e-16 ***
IncomeLevelMedium         0.36340    0.01126   32.29  < 2e-16 ***
IncomeLevelHigh           0.75056    0.01482   50.66  < 2e-16 ***
Coast1:IncomeLevelMedium -0.05849    0.01371   -4.27  2.0e-05 ***
Coast1:IncomeLevelHigh   -0.07982    0.01666   -4.79  1.7e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.37 on 20634 degrees of freedom
Multiple R-squared:  0.574,  Adjusted R-squared:  0.574
F-statistic: 5.57e+03 on 5 and 20634 DF,  p-value: <2e-16
```

The model attains similar R-squared and residual standard error as our previous simple model. While the interaction effects are statistically significant, their associated coefficients are extremely small, thus not largely influencing the quality of our model.

Now we check the assumptions of ANOVA. We start by checking normality of the residuals:



The residuals are approximately as normal as in our original categorical-only model. While they are not perfectly normal (the left tail of the normal quantile plot again curves away from normality), the residuals are close enough to normal for our analysis to be of considerable significance.

We try to satisfy ANOVA's homogeneity of variance assumption. We calculate the ratio of the maximum and minimum standard deviations across all interaction groups:

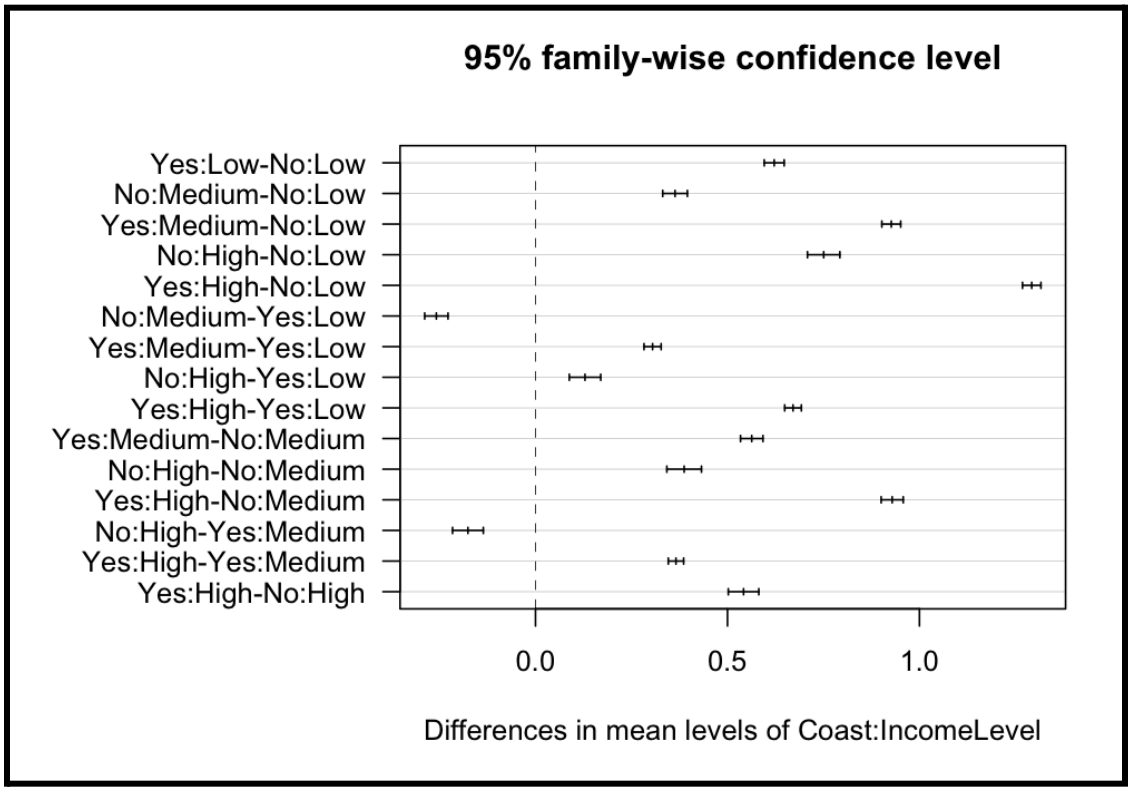
```
Largest SD: 0.42
Smallest SD: 0.32
Ratio of Largest to Smallest SD: 1.3
```

As this ratio is smaller than 2, we have evidence that the variances of the different groups are approximately equal.

Finally, we perform Tukey's HSD test on our model. We have the following results:

```
 Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = LogMedHouseVal ~ Coast + IncomeLevel + Coast * IncomeLevel)

$Coast
       diff  lwr  upr p adj
Yes-No 0.76 0.75 0.78     0

$IncomeLevel
```

```
            diff  lwr  upr p adj
Medium-Low  0.30 0.28 0.31     0
High-Low    0.64 0.62 0.65     0
High-Medium 0.34 0.33 0.36     0

$`Coast:IncomeLevel`
                       diff    lwr   upr p adj
Yes:Low-No:Low         0.62  0.596  0.65     0
No:Medium-No:Low       0.36  0.331  0.40     0
Yes:Medium-No:Low      0.93  0.902  0.95     0
No:High-No:Low         0.75  0.708  0.79     0
Yes:High-No:Low        1.29  1.268  1.32     0
No:Medium-Yes:Low     -0.26 -0.289 -0.23     0
Yes:Medium-Yes:Low     0.30  0.283  0.33     0
No:High-Yes:Low        0.13  0.088  0.17     0
Yes:High-Yes:Low       0.67  0.649  0.69     0
Yes:Medium-No:Medium   0.56  0.534  0.59     0
No:High-No:Medium      0.39  0.342  0.43     0
Yes:High-No:Medium     0.93  0.900  0.96     0
No:High-Yes:Medium    -0.18 -0.216 -0.14     0
Yes:High-Yes:Medium    0.37  0.346  0.39     0
Yes:High-No:High       0.54  0.502  0.58     0
```



The Tukey analysis reveals significant differences in **LogMedHouseVal** values across various combinations of coast and income levels. Particularly notable is the combination

of high income and coastal, which shows a much higher median house value compared to low income and non-coastal, with a mean difference of 1.29. As none of the intervals in the plot cross the zero line, we have a statistically significant difference in means of **MedHouseVal** across all interaction groups.

CONCLUSION

We have reached the end of our analysis for this dataset. In our exploration, we have come to discover why this dataset usually requires complicated machine learning models—the target is capped, none of the variables are normally distributed, and nonlinear terms might be needed to get a high accuracy model. Despite these limitations, however, we have fit three statistically significant models all with R-squared values of around 0.6. We created our own factor variables to fit a model with approximately normal residuals, allowing us to perform ANOVA and discover interaction effects between our predictors. While we ultimately suggest that a more robust model is needed or an enhanced dataset that does not have a capped target, we nevertheless believe that our models have successfully uncovered significant patterns and relationships that provide valuable insights into the dataset.