California Housing prices

The purpose of this project is to analyze the California housing prices dataset from Kaggle and construct a linear regression model to predict house prices. The dataset contains information about houses in California districts based on the 1990 census.

Data Cleaning and Exploration

The dataset has 20640 observations and 10 variables. It was read into R and a summary of each variable was generated. Out of the 10 variables, 9 are quantitative and 1 variable (ocean_proximity) is a categorical variable.

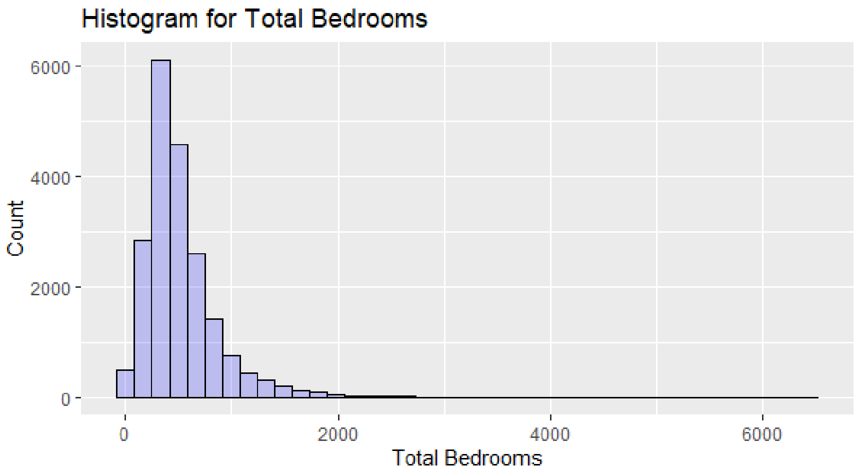
```
library(tidyverse)
caH = read_csv("http://www.utdallas.edu/~vds190000/housing.csv")
summary(caH)
```

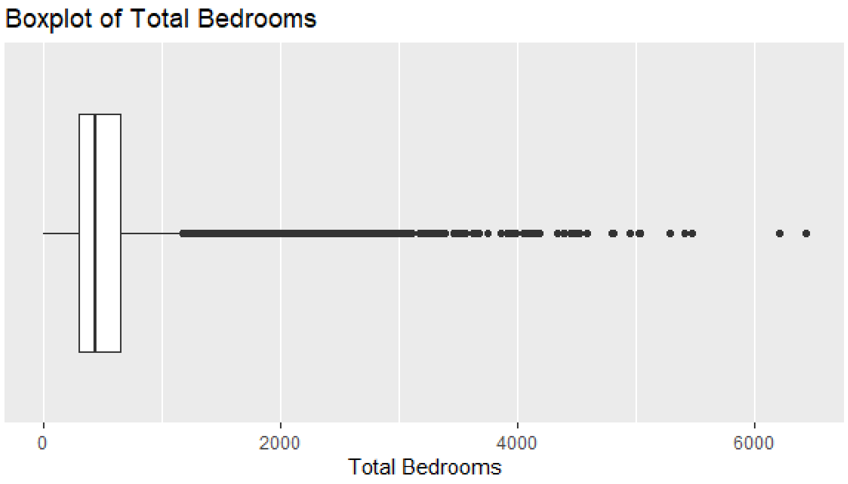
```
longitude         latitude     housing_median_age  total_rooms
 Min.   :-124.3   Min.   :32.54   Min.   : 1.00      Min.   :    2
 1st Qu.:-121.8   1st Qu.:33.93   1st Qu.:18.00      1st Qu.: 1448
 Median :-118.5   Median :34.26   Median :29.00      Median : 2127
 Mean   :-119.6   Mean   :35.63   Mean   :28.64      Mean   : 2636
 3rd Qu.:-118.0   3rd Qu.:37.71   3rd Qu.:37.00      3rd Qu.: 3148
 Max.   :-114.3   Max.   :41.95   Max.   :52.00      Max.   :39320

 total_bedrooms     population      households     median_income
 Min.   :   1.0   Min.   :    3   Min.   :   1.0   Min.   : 0.4999
 1st Qu.: 296.0   1st Qu.:  787   1st Qu.: 280.0   1st Qu.: 2.5634
 Median : 435.0   Median : 1166   Median : 409.0   Median : 3.5348
 Mean   : 537.9   Mean   : 1425   Mean   : 499.5   Mean   : 3.8707
 3rd Qu.: 647.0   3rd Qu.: 1725   3rd Qu.: 605.0   3rd Qu.: 4.7432
 Max.   :6445.0   Max.   :35682   Max.   :6082.0   Max.   :15.0001
 NA's   :207

 median_house_value ocean_proximity
 Min.   : 14999     Length:20640
 1st Qu.:119600     Class :character
 Median :179700     Mode  :character
 Mean   :206856
 3rd Qu.:264725
 Max.   :500001
```

The summary above shows that the total_bedrooms variable has 207 NA or missing values. To further examine the distribution of total_bedrooms, a histogram and a boxplot were constructed. The histogram of total_bedrooms below shows that the distribution is right skewed. This is also evident from the summary statistics where the mean is greater than the median. The boxplot of total_bedrooms shows the presence of many outliers. Instead of dropping the observations with NAs, the median can be used to impute the missing values. The mean would not be a good choice to use for imputation because it is easily affected by the presence of outliers.

```
#histogram of total bedrooms
ggplot(data=caH, aes(x=caH$total_bedrooms)) +
  geom_histogram(bins = 40, col="black", fill="blue",alpha = .2) +
  labs(title="Histogram for Total Bedrooms", x="Total Bedrooms", y="Count")
```

```
#boxplot for total_bedrooms
ggplot(caH,aes("var",total_bedrooms))+
  geom_boxplot()+
  xlab(" ")+
  ylab("Total Bedrooms")+
  scale_x_discrete(breaks=NULL)+
  labs(title = "Boxplot of Total Bedrooms")+
  coord_flip()
```





To deal with the 207 missing values for total_bedrooms, the median of total_bedrooms was used to impute data points.

```
#imputing missing values with median
caH$total_bedrooms[is.na(caH$total_bedrooms)] = totalBedroomMedian
```

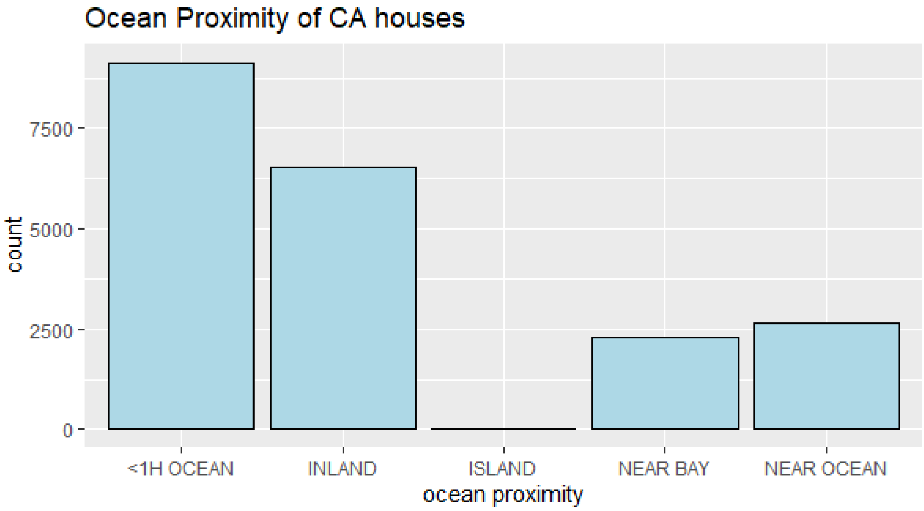
Next, the categorial variable ocean_proximity was converted into a factor variable with 5 levels.

```
#converting categorical variable into a factor
caH$ocean_proximity = as.factor(caH$ocean_proximity)
levels(caH$ocean_proximity)
```

```
[1] "<1H OCEAN"  "INLAND"     "ISLAND"     "NEAR BAY"
[5] "NEAR OCEAN"
```

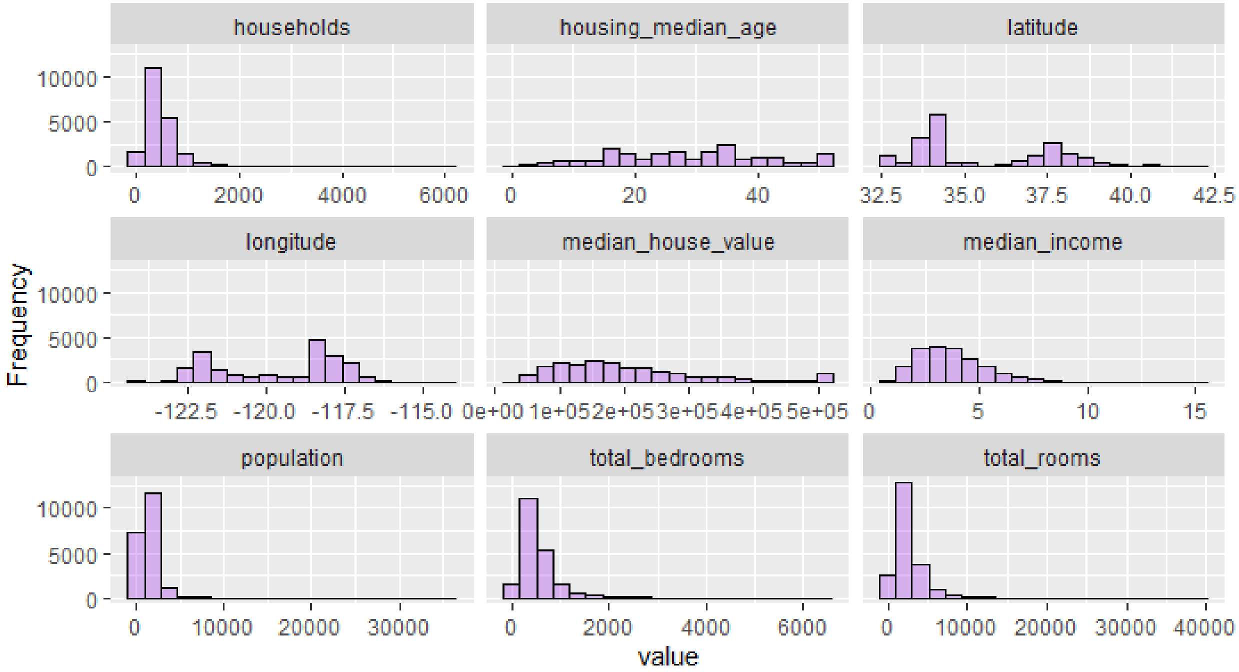
The barplot of ocean_proximity shows that only a few observations are part of the islands level as compared to the other levels.

```
#barplot for categorial variable
ggplot(data = caH,mapping = aes(x=factor(ocean_proximity)))+
  geom_bar(color="black", fill="light blue")+
  labs(title="Ocean Proximity of CA houses", x="ocean proximity")
```



After filling in the missing values, the histograms of all the quantitative variables were examined to obtain a better sense of the distributions.

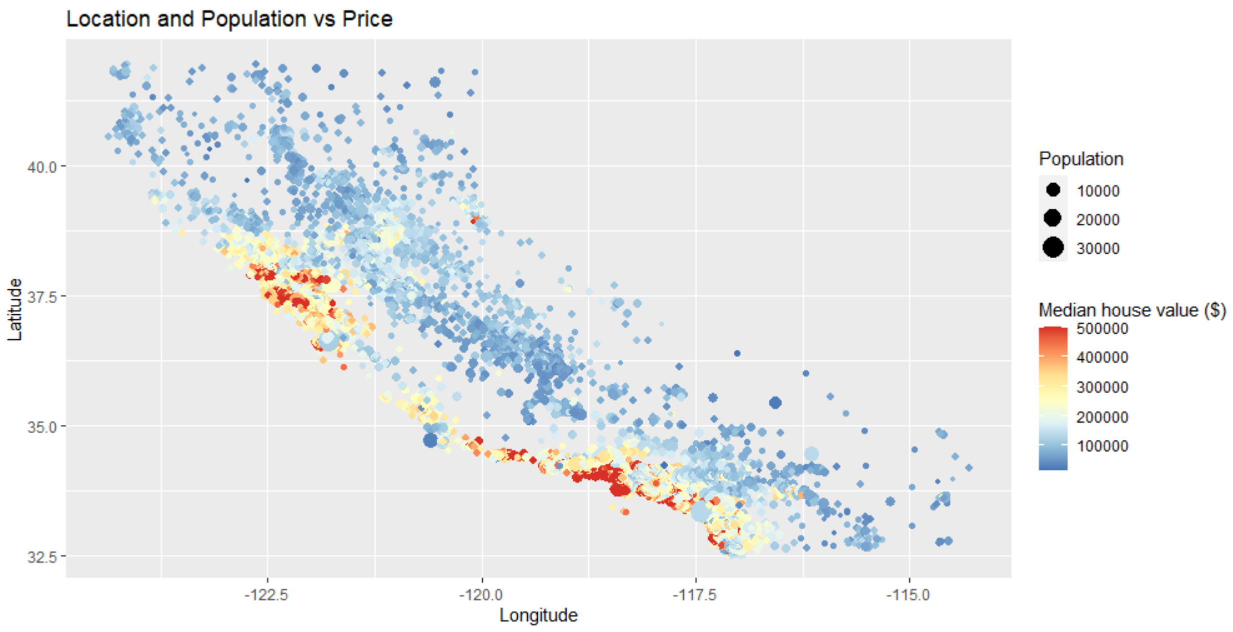
```
#histograms for quantitative variables
data1 = subset(caH , select=-c(ocean_proximity))
data1 %>% gather() %>% head()
ggplot(gather(data1), aes(value)) +
  geom_histogram(bins = 20, color="black", fill="purple", alpha=0.3) +
  facet_wrap(~key, scales = 'free_x')+
  labs(y="Frequency")
```


households
housing_median_age
latitude
longitude
median_house_value
median_income
population
total_bedrooms
total_rooms
Frequency
value
10000
5000
0
0
2000
4000
6000
0
20
40
32.5
35.0
37.5
40.0
42.5
-122.5
-120.0
-117.5
-115.0
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
0
5
10
15
0
10000
20000
30000
0
2000
4000
6000
0
10000
20000
30000
40000

Location is a known factor to influence real estate prices. The median house values in the dataset were plotted against longitude and latitude to understand the degree of this effect. The graph below shows that the large densely populated metropolitan areas of Los Angeles and San Francisco have the highest median house values. California's largely agricultural Central Valley have some of the lowest median house values.

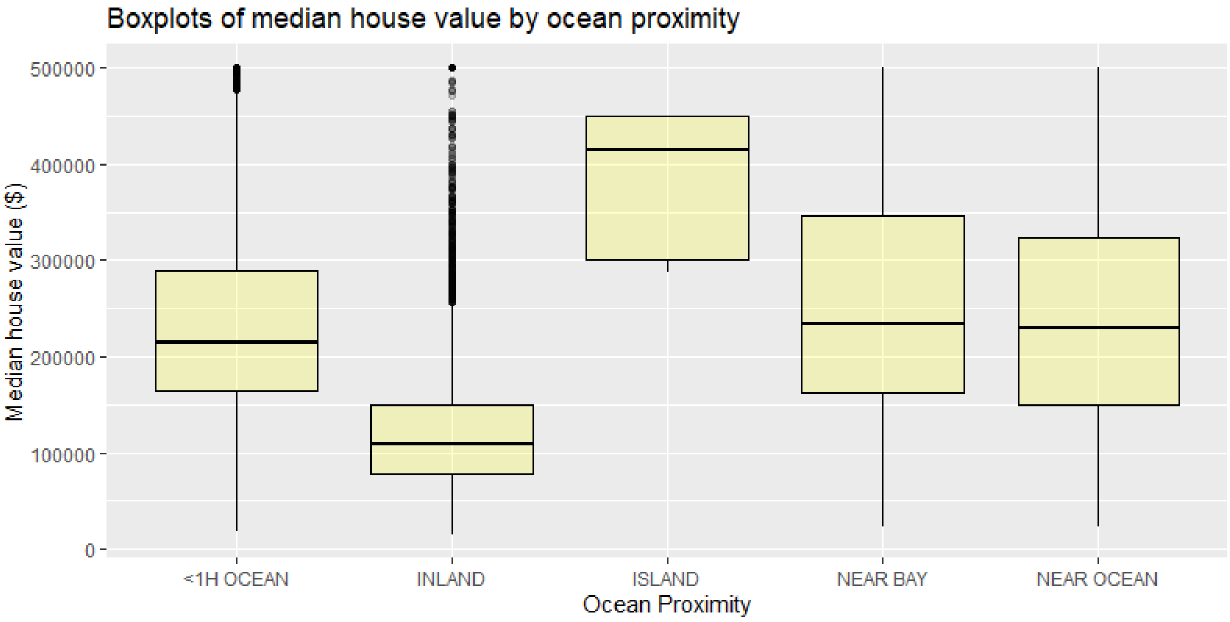
On average houses closer to the ocean have higher median prices whereas those inland tend to have lower median prices. But it is interesting to note that houses in certain coastal regions like in the remote towns of California's northern most coast have lower median house values. On the other hand, certain inland locations like Sacramento and the Lake Tahoe region have higher median house values. This illustrates that ocean_proximity is important in predicting house values, but it does not tell the whole story. Other factors do play a role in determining median house value.

```
#map with location, population and price
require(scales)
plot_map1 = ggplot(caH,
                  aes(x = longitude, y = latitude))+
  geom_point(aes(color=median_house_value, size=population))+
  scale_color_distiller(palette = "RdYlBu")+
  theme_pander()
  labs(title="Location and Population vs Price",color="Median house value
($)", size="Population", x="Longitude", y="Latitude")

plot_map1
```



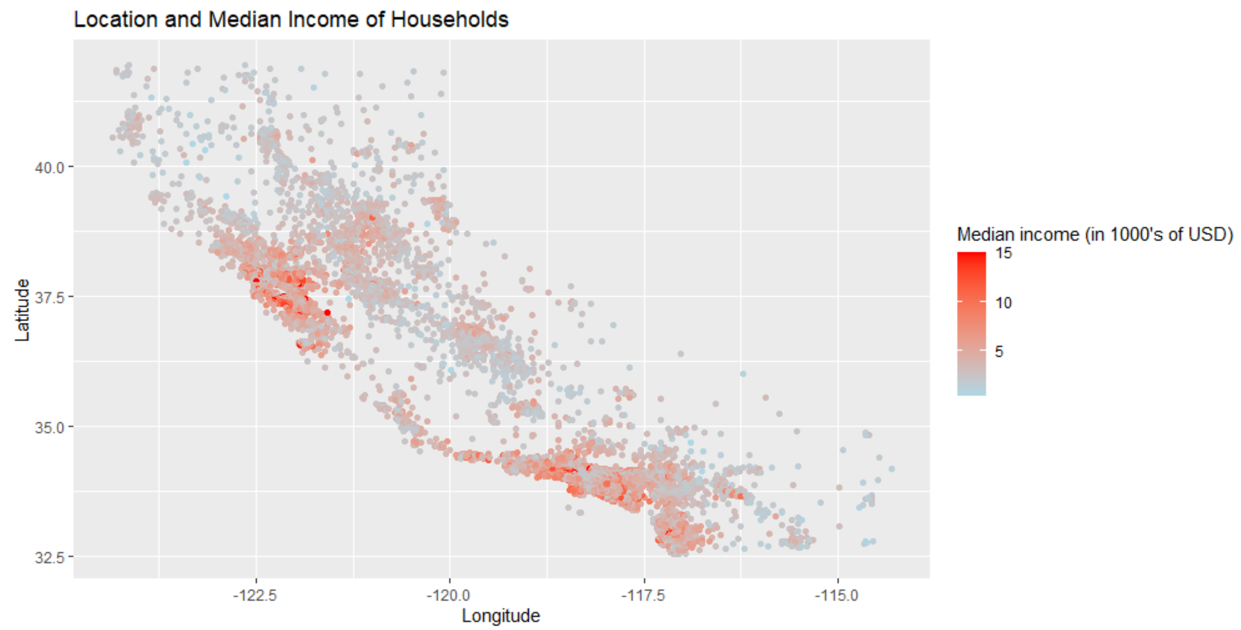
The boxplots below show the variation in the data among the different levels of ocean_proximity. There is a wide range in median house values of ocean-adjacent regions. As we saw in the map above, remote coastal towns have lower median house values compared to coastal towns near the metropolitan areas. The inland level contains a lot of outliers with high median house values which could correspond to the Sacramento area for example.

```
#boxplots of price and ocean proximity
ggplot(caH, aes(x=ocean_proximity, y=median_house_value))+
  geom_boxplot(color="black", fill="yellow", alpha=0.2)+
  labs(title="Boxplots of median house value by ocean proximity", x="Ocean
Proximity", y="Median house value ($)")
```



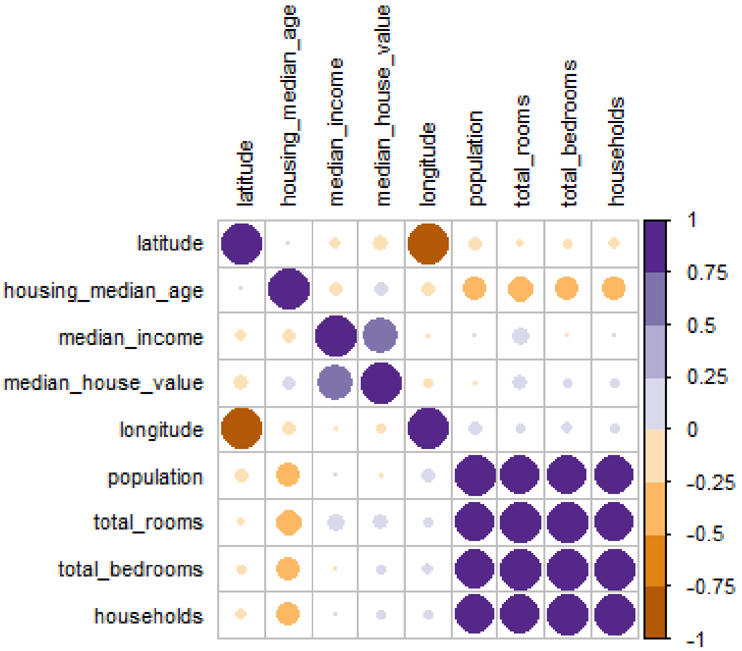
Median income also tends to be higher on average in large metropolitan areas of Los Angeles and San Francisco. This correlates with the higher median house values in these areas as seen in the graph above. Median household income is therefore a good predictor of median house values and can be included in the linear model.

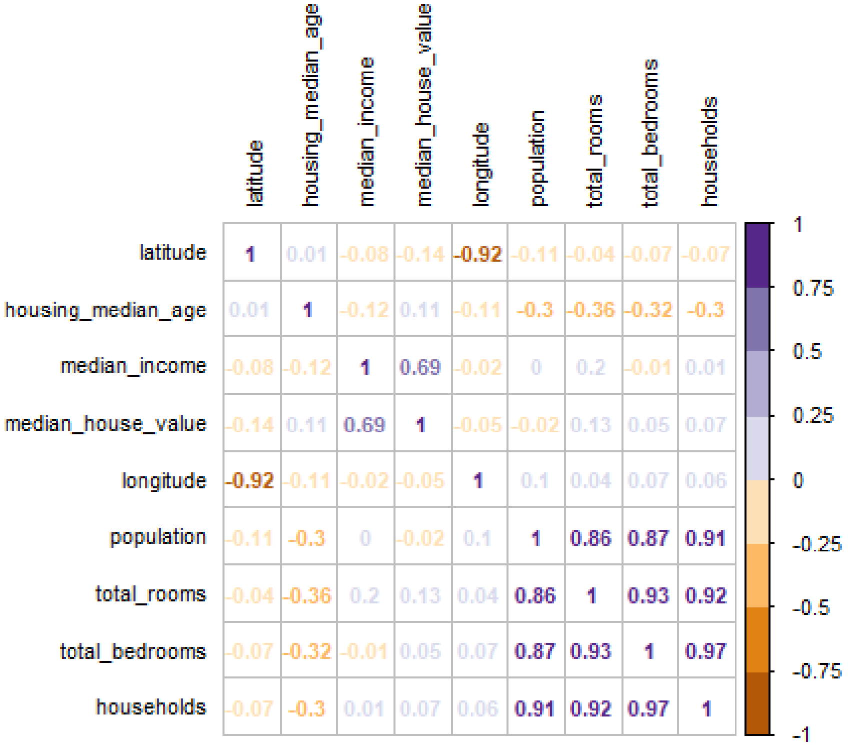
```
#map with location and median income
plot_map2 = ggplot(caH,
                   aes(x = longitude, y = latitude, color=median_income))+
  geom_point()+
  scale_color_gradient(low="light blue", high="red")+
labs(title="Location and Median Income of Households",color="Median income
(in 1000's of USD)",x="Longitude", y="Latitude")
plot_map2
```


Location and Median Income of Households
Latitude
40.0
37.5
35.0
32.5
-122.5
-120.0
-117.5
-115.0
Longitude
Median income (in 1000's of USD)
15
10
5

Corrplots was generated to understand the relationships among variables. As expected, median income has a positive correlation (r=0.69) with median house value. Housing_median_age and total_rooms are very weakly correlated with median house value. However, there is also collinearity between population and total_rooms, population and total_bedrooms, and population and households. Although these predictor variables have correlations among them, they are typically influential in predicting house values. Bigger houses with more bedrooms for example tend to be more expensive.

```
require(corrplot)
require(RColorBrewer)
corrplot(cor(caH[,1:9]),tl.cex = 0.7, tl.col = "black",cl.cex = 0.7,cl.ratio
= 0.2,cl.align.text = "l",order = "hclust",col = brewer.pal(n=8,name="PuOr"))
corrplot(cor(caH[,1:9]),method="number",number.cex=0.7,tl.cex = 0.7, tl.col =
"black",cl.cex = 0.7,cl.ratio = 0.2,cl.align.text = "l",order = "hclust",col
= brewer.pal(n=8,name="PuOr"))
```





	latitude	housing_median_age	median_income	median_house_value	longitude	population	total_rooms	total_bedrooms	households
latitude	1	0.01	-0.08	-0.14	-0.92	-0.11	-0.04	-0.07	-0.07
housing_median_age	0.01	1	-0.12	0.11	-0.11	-0.3	-0.36	-0.32	-0.3
median_income	-0.08	-0.12	1	0.69	-0.02	0	0.2	-0.01	0.01
median_house_value	-0.14	0.11	0.69	1	-0.05	-0.02	0.13	0.05	0.07
longitude	-0.92	-0.11	-0.02	-0.05	1	0.1	0.04	0.07	0.06
population	-0.11	-0.3	0	-0.02	0.1	1	0.86	0.87	0.91
total_rooms	-0.04	-0.36	0.2	0.13	0.04	0.86	1	0.93	0.92
total_bedrooms	-0.07	-0.32	-0.01	0.05	0.07	0.87	0.93	1	0.97
households	-0.07	-0.3	0.01	0.07	0.06	0.91	0.92	0.97	1

Model Building

Based on the explanatory data analysis above, it seems that median_income and ocean_proximity have the strongest relationship with median_house_value. If we wanted to predict median_house_value, it seems that we should at least include these two predictors. To begin with, a multiple regression model was fitted with log(median_house_value) as the response and median_income and ocean_proximity as predictors. Log transformation of the response variable increased the adjusted R^2 value. The null hypothesis in this case states that there is no relation between any of the predictors and the response variable. The null hypothesis and the significance of the model is tested with the F statistic.

```
fit1 = lm(log(caH$median_house_value)~caH$median_income +caH$ocean_proximity,
data=caH)
summary(fit1)
```

```
Call:
lm(formula = log(caH$median_house_value) ~ caH$median_income + 
    caH$ocean_proximity, data = caH)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.30539 -0.22853 -0.01999  0.20937  1.96266 

Coefficients:
                              Estimate Std. Error  t value
(Intercept)                  11.591963   0.006747 1718.103
caH$median_income             0.166577   0.001334  124.869
caH$ocean_proximityINLAND    -0.515529   0.005881  -87.655
caH$ocean_proximityISLAND     0.780848   0.158096    4.939
caH$ocean_proximityNEAR BAY   0.058064   0.008259    7.030
caH$ocean_proximityNEAR OCEAN 0.040012   0.007794    5.134
                             Pr(>|t|)    
(Intercept)                   < 2e-16 ***
caH$median_income             < 2e-16 ***
caH$ocean_proximityINLAND     < 2e-16 ***
caH$ocean_proximityISLAND    7.91e-07 ***
caH$ocean_proximityNEAR BAY  2.12e-12 ***
caH$ocean_proximityNEAR OCEAN 2.86e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3534 on 20634 degrees of freedom
Multiple R-squared:  0.6145,	Adjusted R-squared:  0.6145 
F-statistic:  6580 on 5 and 20634 DF,  p-value: < 2.2e-16
```

A high value for the F statistic and a very low p-value (< 2.2e^{-16}) implies that the null hypothesis can be rejected. The coefficient estimates are highly statistically significant with very low p-value. This implies that there is a relationship between the response and predictor variables. However, the adjusted R^2 value of 0.6145 indicates that only 61.45% of the variability in the data can be explained by the model.

It is possible that including more information and using additional predictors can improve our model. Next, a multiple regression model was fitted with all the predictors.

```
fit2 = lm(log(caH$median_house_value)~., data=caH)
summary(fit2)
```

```
Call:
lm(formula = log(caH$median_house_value) ~ ., data = caH)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.3585 -0.1990 -0.0089  0.1911  3.4319 

Coefficients:
                            Estimate Std. Error t value
(Intercept)               -2.190e+00  4.194e-01  -5.222
longitude                 -1.604e-01  4.861e-03 -32.999
latitude                  -1.562e-01  4.794e-03 -32.588
housing_median_age         2.454e-03  2.095e-04  11.709
total_rooms               -8.305e-06  3.695e-06  -2.248
total_bedrooms             2.703e-04  2.844e-05   9.503
population                -1.779e-04  5.101e-06 -34.872
households                 3.628e-04  3.192e-05  11.366
median_income              1.670e-01  1.593e-03 104.855
ocean_proximityINLAND     -3.099e-01  8.324e-03 -37.231
ocean_proximityISLAND      6.016e-01  1.475e-01   4.077
ocean_proximityNEAR BAY   -3.743e-02  9.138e-03  -4.096
ocean_proximityNEAR OCEAN -3.188e-02  7.492e-03  -4.254
                          Pr(>|t|)    
(Intercept)               1.79e-07 ***
longitude                  < 2e-16 ***
latitude                   < 2e-16 ***
housing_median_age         < 2e-16 ***
total_rooms                 0.0246 *  
total_bedrooms             < 2e-16 ***
population                 < 2e-16 ***
households                 < 2e-16 ***
median_income              < 2e-16 ***
ocean_proximityINLAND      < 2e-16 ***
ocean_proximityISLAND     4.57e-05 ***
ocean_proximityNEAR BAY   4.22e-05 ***
ocean_proximityNEAR OCEAN 2.11e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3295 on 20627 degrees of freedom
Multiple R-squared:  0.665,	Adjusted R-squared:  0.6648 
F-statistic:  3412 on 12 and 20627 DF,  p-value: < 2.2e-16
```

Again, a high value for the F statistic and a very low p-value (< 2.2e^{-16}) implies that the null hypothesis can be rejected. There is a potential relationship between the predictors and the response variable. The adjusted R^2 value increased slightly compared to the previous model. 66.48% of the variance in the data can be explained by the new model.

The first model with only two predictors was compared with this model using the anova function in R:

```
anova(fit1,fit2)
```

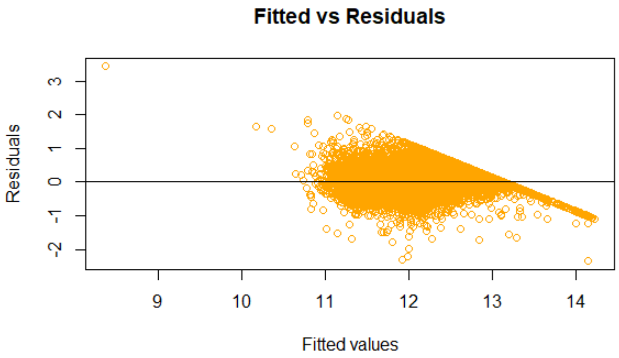
```
Analysis of Variance Table

Model 1: log(caH$median_house_value) ~ caH$median_income + caH$ocean_proximit
y
Model 2: log(caH$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  20634 2576.9                                  
2  20627 2239.8  7    337.02 443.38 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The low p-value of <2.2e-06 indicates that the inclusion of more predictors did lead to a significant improvement over just using median_income and ocean_proximity.

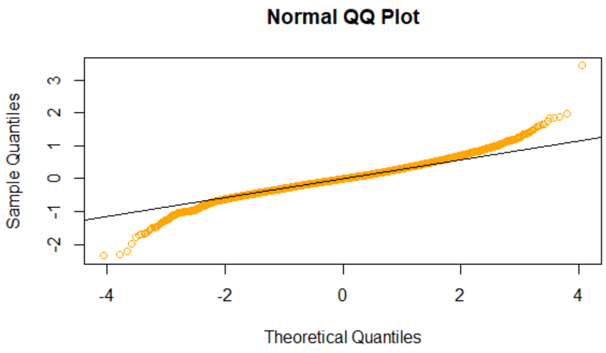
Model Evaluation

One of the assumptions of linear regression is that there is a linear relationship between the predictors and the response. The linearity of the model was verified via a residual plot of fitted values vs. the residuals. A non-random pattern would suggest that a linear model is not appropriate. Most points tend to cluster towards the center of the plot below. However, there is some pattern towards the right and so we cannot claim that the linearity assumption is fully satisfied. We also cannot verify the assumption that residuals have constant variance.

```
#residual plot and QQ plot

plot(fitted(fit2),resid(fit2), xlab ="Fitted values", ylab="Residuals",
main="Fitted vs Residuals", col="orange")
abline(h=0)
qqnorm(resid(fit2), col="orange", main = "Normal QQ Plot")
qqline(resid(fit2))
```



Another assumption of linear regression is that residuals are normally distributed. To verify this assumption, a normal QQ plot was used. For normally distributed data, observations lie approximately on a straight line. Since the points in the QQ plot below are scattered roughly along the straight diagonal line, the normality assumption for residuals is satisfied.



Based on the exploratory data analysis, median income and ocean proximity had positive correlations with median house value. Therefore, the initial model only included these variables as predictors. However, the resulting model had a low adjusted R-squared value and high residual standard error.

To improve the model, additional predictors were included. Although there was some collinearity between predictor variables such as total_bedrooms and total_rooms, they were included in the model because these variables are typically influential in predicting house prices. This model with all the predictors had a lower residual standard error and a higher adjusted R-squared value.

$$\begin{aligned} log(median\_house\_value) &= \beta_0 + \beta_1 longitude + \beta_2 latitude + \beta_3 housing\_median\_age \\ &+ \beta_4 total\_rooms + \beta_5 total\_bedrooms + \beta_6 population + \beta_7 households \\ &+ \beta_8 median\_income + \beta_9 ocean\_proximityInland \\ &+ \beta_{10} ocean\_proximityIsland + \beta_{11} ocean\_proximityNearBay \\ &+ \beta_{12} ocean\_proximityNearOcean \end{aligned}$$

$$\begin{aligned} log(median\_house\_value) &= -2.190 + -1.604e^{-1}\, longitude + -1.562e^{-1} latitude \\ &+ 2.454e^{-3} housing\_median\_age + -8.305e^{-6} total\_rooms \\ &+ 2.703e^{-4} total\_bedrooms + -1.779e^{-4} population \\ &+ 3.628e^{-4} households + 1.670e^{-1} median\_income \\ &+ -3.099e^{-1} ocean\_proximityInland + 6.016e^{-1} ocean\_proximityIsland \\ &+ -3.743e^{-2} ocean\_proximityNearBay \\ &+ -3.188e^{-2} ocean\_proximityNearOcean \end{aligned}$$

References

Grolemund, Garrett, and Hadley Wickham. “R For Data Science.” *R For Data Science*, O'Reilly, r4ds.had.co.nz/index.html.