PERSONAL PROJECT - California Housing Price Prediction

Natalie Tran

2026-01-12

Introduction

The dataset, referred to as the California Housing Data, originates from the 1990 U.S. Census and was first introduced by R. Kelly Pace and Ronald Barry (1997) in their paper Sparse Spatial Autoregressions (Statistics & Probability Letters, 33(3), 291–297). It contains 20,640 observations on California housing block groups, each including demographic, economic, and geographic information. The response variable in this project is the median house value for each block group, measured in U.S. dollars. The remaining variables will be treated as potential predictors of housing prices.

Below is a brief description of the main variables used in this project:

- **median_house_value**: Median house value in the block group (U.S. dollars) — *response variable.*
- **median_income**: Median income of households in the block group (in tens of thousands of dollars).
- **housing_median_age**: Median age of houses in the block group (years).
- **total_rooms**: Total number of rooms in all houses in the block group.
- **total_bedrooms**: Total number of bedrooms in all houses in the block group.
- **population**: Total population in the block group (number of people).
- **households**: Number of households in the block group (occupied housing units).
- **longitude**: Longitude coordinate of the block group (degrees).
- **latitude**: Latitude coordinate of the block group (degrees).
- **ocean_proximity**: Categorical variable indicating proximity to the ocean (e.g., “INLAND”, “NEAR OCEAN”).

This project focuses on both modeling and prediction. First, I will use exploratory data analysis (EDA) to understand how individual variables behave and how they are related to median house value. Then, I will fit regression models to study the relationship between housing prices and selected predictors, and compare a linear model with nonlinear alternatives.

The main research questions are:

1. Which demographic and housing characteristics have the strongest relationship with median house value?
2. How does income relate to housing prices? Is it linear?
3. Are there regional differences in housing prices across California?
4. How is the distribution of house prices shaped?
5. Which functional form best fits the data?

Exploratory Data Analysis

This project aims to explore how socioeconomic and geographic factors influence housing prices across California.

A check for missing data (Appendix A) showed that the variable *total_bedrooms* contained 207 missing values, while all other variables were complete. Since the amount of missingness was small (about 1% of the dataset), I handled it by imputing the median value in those missing spots, ensuring a complete dataset for modeling.

1. Which demographic and housing characteristics have the strongest relationship with median house value? To identify which variables are most strongly associated with housing prices, I computed the correlation between *median_house_value* and each numeric predictor. The results are shown in Table 1 below.

Table 1: Correlation between Median House Value and each numerical predictor

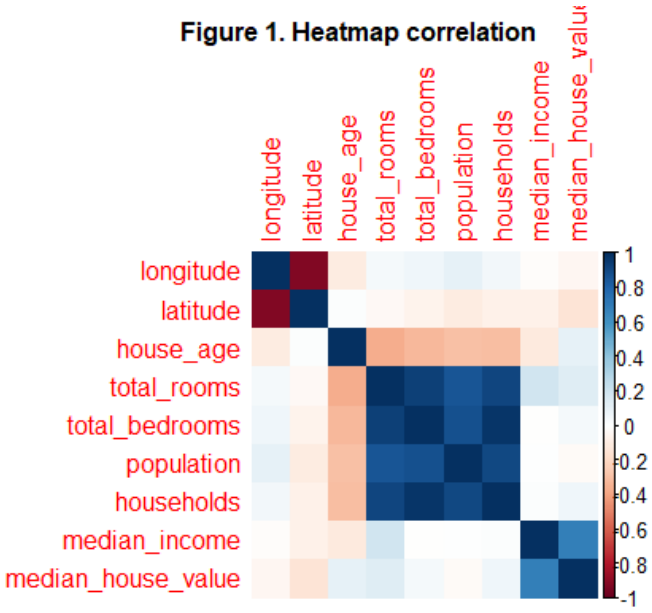
	Median House Value
Longitude	-0.04539822
Latitude	-0.14463821
Median House Age	0.10643205
Total rooms	0.13329413
Total bedrooms	0.04968618
Population	-0.02529973
Households	0.06489355
Median Income	**0.68835548**

Among all variables examined, Median Income has by far the strongest correlation with house value ($r = 0.69$). In other words, higher-income areas tend to have significantly higher home prices. On the other hand, other demographic variables, such as population, households, and total bedrooms, exhibit very weak correlations with house prices ($|r| < 0.1$).

Notice that three predicted variables have a negative correlation with the house prices. The population has a very small negative correlation ($r = -0.025$), reflecting that densely populated block groups are not necessarily the most expensive areas. The negative relationship between longitude/latitude and prices likely shows underlying geographic effects (for example, differences between coastal and inland areas), but a deeper exploration of location-based variation will be conducted in Research Question 3.

A correlation heatmap (Figure 1) further illustrates these relationships: income shows the darkest (strongest) correlation with house value, whereas most other predictors appear much lighter. The heatmap also highlights a cluster of variables that are strongly correlated with each other (*total_rooms*, *total_bedrooms*, *population*, and *households*), suggesting possible redundancy that will be addressed during model selection.

Figure 1. Heatmap correlation



Overall, the table and heatmap indicate that *median income is the strongest single predictor* of median house value, with substantially weaker relationships for the remaining demographic and housing variables.

2. How does income relate to housing prices? Is it linear? To examine the relationship between income and housing prices, I plotted *median_house_value* against *median_income* and combined both a linear regression line and a quadratic (second-degree polynomial) curve. The scatterplot (Figure 2) shows a clear positive association between income and house value, indicating that higher-income areas tend to have more expensive homes. The overall pattern suggests a strong upward trend with slight curvature at higher income levels, where the relationship begins to flatten. This flattening is partly due to the dataset's upper-censoring at $500,000.

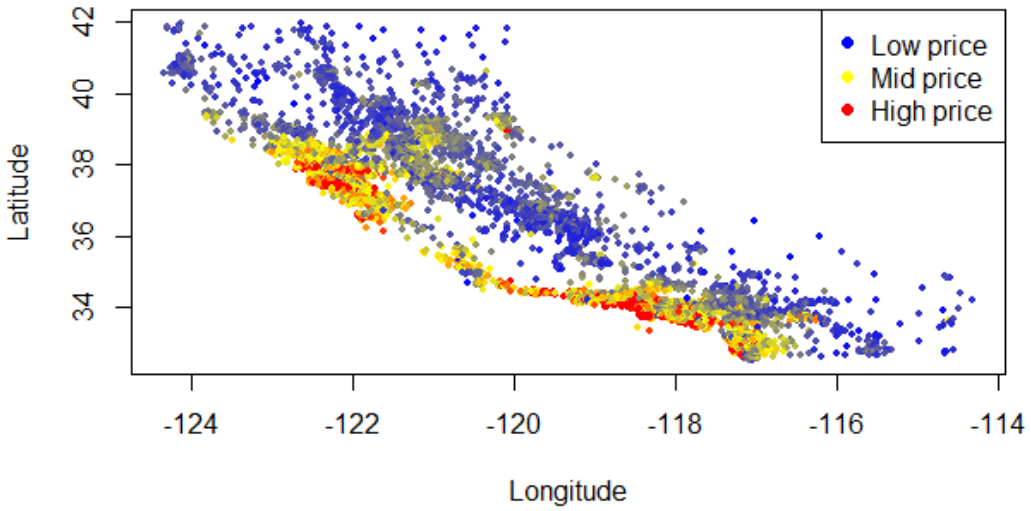
Figure 2. Median House Value vs. Median Income and two regression lines

I fit two models to quantify this relationship. The simple linear regression explains approximately 47.3% (Adjusted $R^2 = 0.4734$) of the variation in house value (*Appendix*). Although an R^2 in the 0.47 range may not seem extremely high, this value is actually quite strong for socioeconomic and housing data, where many factors should influence prices, and single predictors rarely explain large proportions of variance. The highly significant coefficient for median income (p-value $<$ 2e-16) further confirms that income is a key driver of housing value.

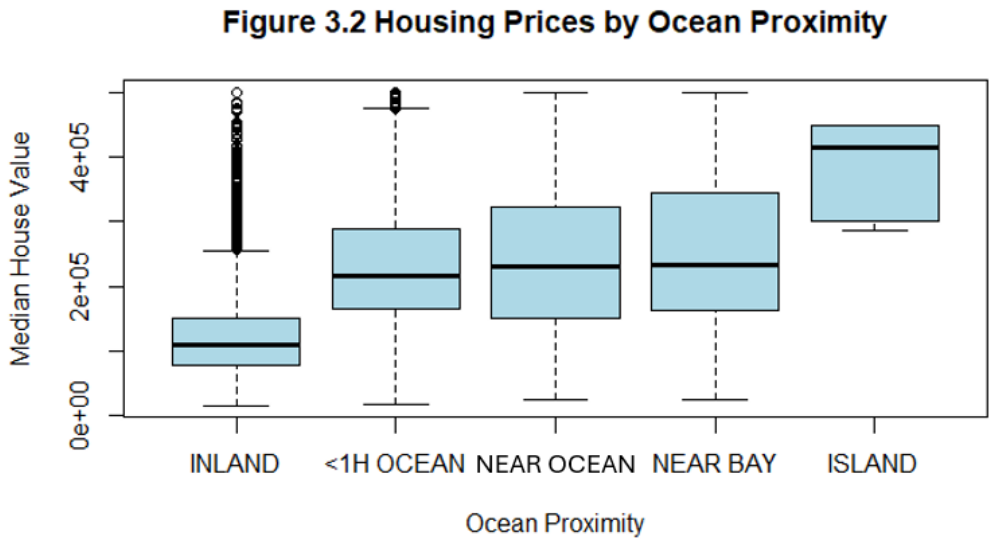
Afterwards, I fit a quadratic model by adding a quadratic term. This model produces a slightly higher adjusted R^2 of about 0.478, a very small improvement. This suggests that although there is some curvature, consistent with the visual pattern in the scatterplot, the relationship between income and housing prices is primarily linear.

3. Are there regional differences in housing prices across California? To examine whether housing prices vary across different regions of California, I explored the geographic distribution of median house values using longitude and latitude coordinates. Figure 3.1 displays each housing block group as a point on the California map, colored according to its relevant price level. The plot reveals a clear and consistent spatial trend: higher housing prices are concentrated along the coast, especially near the Los Angeles region and the San Francisco Bay Area. In contrast, inland regions such as the Central Valley exhibit mostly lower-priced housing, shown by dense clusters of blue points.

Figure 3.1 Geographic Distribution of Housing Prices



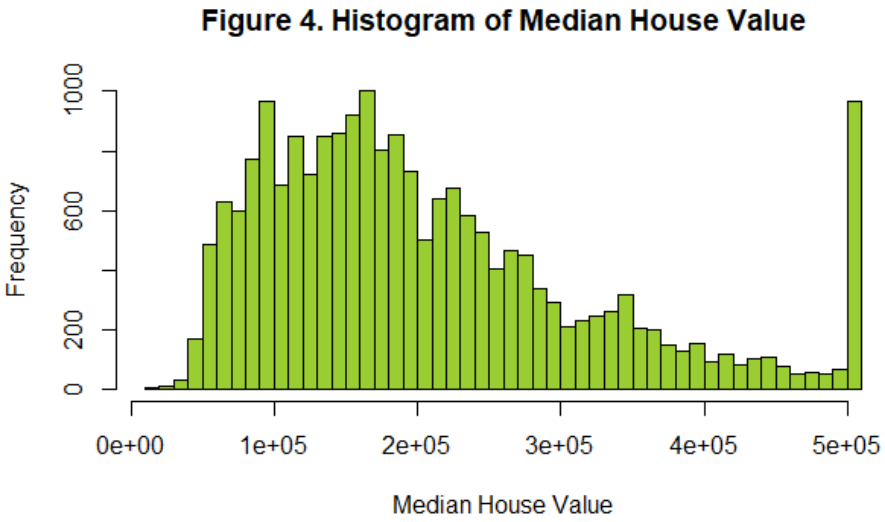
To be more specific, Figure 3.2 presents a boxplot of median house value grouped by the *ocean_proximity* category. The <1H OCEAN, NEAR OCEAN, and NEAR BAY categories all show higher prices, with noticeably higher medians and broader spreads, reflecting the diversity and desirability of coastal neighborhoods. Although the ISLAND category contains very few observations (only 5), it exhibits some of the highest values in the dataset.



Recall from Table 1, the negative relationship between latitude and house price (r = -0.145) also means that as latitude increases (moving North), housing prices tend to decrease slightly. This aligns with the known geographic pattern that Southern California, which lies at lower latitudes, tends to have higher housing costs.

Together, these visual analyses confirm that regional factors strongly influence housing prices; therefore, they are likely to be important components of the regression model.

4. How is the distribution of house prices shaped? Before constructing a regression model, it is important to understand the distribution of the response variable, *median_house_value*. A histogram (Figure 4) of the housing values reveals a right-skewed distribution with a long upper tail, as well as a noticeable spike at the maximum value of $500,001. This spike occurs because many homes in the dataset were top-coded at this amount in the 1990 Census, meaning values above this threshold were recorded as the same number. Such censoring creates a concentration at the upper end of the distribution that may affect model fitting and interpretation.



Besides, the skewness in the distribution suggests that the variance of housing prices is not constant and that the relationship between predictors and the response may benefit from transformation.

Model construction and analysis

Based on the explanatory analysis, I estimated three regression models: a standard linear model, a log-linear model, and a double-log model. These models were chosen to compare different functional forms and determine which specification best captures how income, housing characteristics, and location influence housing prices. Table 2 reports the results.

Model Predictions prior to Estimation The EDA suggested that the log-linear model would perform well. Housing values were heavily right-skewed (Figure 4) and top-coded at $500,001, indicating that a log transformation of the dependent variable could stabilize variance and improve model behavior. The scatterplot of income versus house value (Figure 2) showed a mostly linear trend with only a mild curvature, suggesting that the linear model would be reasonable but not optimal, and a quadratic model wouldn't be much better either. Moreover, it is also worth trying the double-log form, which is common in economics for elasticity interpretation.

Model (1): Linear Specification Model (1) shows that median income, longitude, latitude, and ocean proximity are strongly significant and align with expectations from EDA. Several housing-size variables (total rooms and total bedrooms) are statistically significant but inconsistent in sign due to multicollinearity within this group. The Adjusted R^2 of 0.645 means that the model explains a substantial portion of the variation but leaves room for improvement.

Model (2): Log-Linear Specification In Model (2), the dependent variable is log-transformed. This improves the Adjusted R^2 to 0.665, consistent with the expectation that logging a skewed variable enhances fit. With this new function, a one-unit increase in median income (the predictor) will induce a 16.7% increase in the predicted housing value.

Model (3): Double-log Specification Model (3) applies a log transformation to both the dependent variable and median income. While the original motivation for the double-log model was based on economic elasticity interpretation, the histogram of median income (Appendix C) shows that income is also right-skewed.

Consistent with these expectations, the double-log model, having the highest Adjusted R^2, explains 67.1% of the median house price variation, outperforming both linear and log-linear models. With this model, a 1% change in median income will induce a 0.693% change in the median house prices, ceteris paribus. Geographic variables and ocean proximity categories remain strongly significant, reinforcing the importance of location identified earlier in EDA.

Table 2. Multiple Regression Models

	Dependent variable:		
	median_house_value	log(median_house_value)	
	(1)	(2)	(3)
log(median_income)			0.693***
			(0.010)
longitude	-26,430.440***	-0.160***	-0.152***
	(1,036.878)	(0.006)	(0.006)
latitude	-25,173.280***	-0.156***	-0.148***
	(1,043.372)	(0.006)	(0.006)
house_age	1,057.816***	0.002***	0.003***
	(51.412)	(0.0002)	(0.0002)
total_rooms	-4.731***	-0.00001	-0.00001**
	(1.223)	(0.00001)	(0.00001)
total_bedrooms	71.345***	0.0003***	0.0003***
	(10.022)	(0.00004)	(0.00005)
population	-39.287***	-0.0002***	-0.0002***
	(4.639)	(0.00002)	(0.00002)
households	77.804***	0.0004***	0.0003***
	(14.931)	(0.0001)	(0.0001)
median_income	38,760.450***	0.167***	
	(526.662)	(0.003)	
ocean_proximity< 1H OCEAN	39,766.400***	0.310***	0.303***
	(1,633.009)	(0.010)	(0.009)
ocean_proximityNEAR OCEAN	44,525.150***	0.278***	0.280***
	(2,398.798)	(0.013)	(0.013)
ocean_proximityNEAR BAY	36,069.000***	0.272***	0.271***
	(2,485.138)	(0.013)	(0.012)
ocean_proximityISLAND	195,832.100***	0.911***	0.882***
	(36,424.760)	(0.108)	(0.121)
Constant	-2,272,858.000***	-2.500***	-1.997***
	(87,013.100)	(0.479)	(0.481)
Observations	20,640	20,640	20,640
R^2	0.645	0.665	0.671
Adjusted R^2	0.645	0.665	0.671
Residual Std. Error (df = 20627)	68,730.970	0.330	0.326
F Statistic (df = 12; 20627)	3,129.289***	3,411.541***	3,511.941***
Note:			*p<0.1; **p<0.05; ***p<0.01

Discussion

The regression analysis shows clear and consistent patterns about what drives housing prices in California. Income and geographic location emerge as the most influential factors, while many housing-size variables contribute little. The performance of the models reflects the structure of the data: both housing values and income are right-skewed (Appendix C), the log-based models handle the distribution more effectively than the linear form (Table 2). The double-log model performs slightly better than the log-linear version, suggesting that proportional changes in income are closely associated with proportional changes in house values. In summary, the double-log model explains 67.1% of the median household variation, and the coefficient of log(median_income) indicates a 1% change in median income will induce a 0.693% change in the median house prices, ceteris paribus.

Although the models fit reasonably well, there are several limitations to note. The data contain a substantial number of top-coded house values at $500,001, which compresses the upper tail and understates variation in high-value coastal areas. Additionally, some predictors, particularly the size-related variables, are highly correlated (Figure 1). Moreover, the dataset omits many real-world factors that might influence housing markets, such as school quality, crime rates, environmental risks, and neighborhood amenities.

Despite these limitations, the results provide a coherent picture: income and proximity to coastal regions are the strongest determinants of housing prices. The double-log model offers the clearest and most stable interpretation, making it the most appropriate specification among those considered.

References

Dataset: California housing data, derived from the 1990 U.S. Census and first introduced by Pace, R. K., & Barry, R. (1997). Sparse Spatial Autoregressions. Statistics & Probability Letters, 33(3), 291–297. https://doi.org/10.1016/S0167-7152(96)00140-X

Siegler, Mark V. *Introduction to Data Analysis and Econometrics.* Forthcoming 2026, W. W. Norton & Company, pp. 261–269.

Appendix

```
#Dataset
ca_housing <- read.csv("Cali_House_Price.csv", header = T)
colnames(ca_housing)[3] <- "house_age"

#Check for NA values
sum(is.na(ca_housing)) #207
```

Appendix A

```
## [1] 207
```

```
colSums(is.na(ca_housing)) #All 207 NAs are in total bedrooms column
```

```
##          longitude           latitude         house_age        total_rooms
##                  0                  0                  0                  0
##     total_bedrooms         population         households      median_income
##                207                  0                  0                  0
##    ocean_proximity median_house_value
##                  0                  0
```

```r
ca_housing$total_bedrooms[is.na(ca_housing$total_bedrooms)] <- median(ca_housing$total_bedrooms, na.rm =
str(ca_housing)
```

```
## '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 ...
##  $ house_age         : int  41 21 52 52 52 52 52 52 42 52 ...
##  $ total_rooms       : int  880 7099 1467 1274 1627 919 2535 3104 2555 3549 ...
##  $ total_bedrooms    : int  129 1106 190 235 280 213 489 687 665 707 ...
##  $ population        : int  322 2401 496 558 565 413 1094 1157 1206 1551 ...
##  $ households        : int  126 1138 177 219 259 193 514 647 595 714 ...
##  $ median_income     : num  8.33 8.3 7.26 5.64 3.85 ...
##  $ ocean_proximity   : chr  "NEAR BAY" "NEAR BAY" "NEAR BAY" "NEAR BAY" ...
##  $ median_house_value: int  452600 358500 352100 341300 342200 269700 299200 241400 226700 261100 ..
```

```r
summary(ca_housing)
```

```
##    longitude         latitude       house_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.: 297.0   1st Qu.:  787   1st Qu.: 280.0   1st Qu.: 2.5634
##  Median : 435.0   Median : 1166   Median : 409.0   Median : 3.5348
##  Mean   : 536.8   Mean   : 1425   Mean   : 499.5   Mean   : 3.8707
##  3rd Qu.: 643.2   3rd Qu.: 1725   3rd Qu.: 605.0   3rd Qu.: 4.7432
##  Max.   :6445.0   Max.   :35682   Max.   :6082.0   Max.   :15.0001
##  ocean_proximity    median_house_value
##  Length:20640       Min.   : 14999
##  Class :character   1st Qu.:119600
##  Mode  :character   Median :179700
##                     Mean   :206856
##                     3rd Qu.:264725
##                     Max.   :500001
```

```r
#Question 1 - Correlation
numeric_vars <- ca_housing[sapply(ca_housing,is.numeric)]
cor_matrix <- cor(numeric_vars, use= "complete.obs")
cor_matrix #Table 1
```

Appendix B

```
##                      longitude    latitude   house_age total_rooms
## longitude           1.00000000 -0.92466443 -0.10819681  0.04456798
## latitude           -0.92466443  1.00000000  0.01117267 -0.03609960
```

```
## house_age          -0.10819681  0.01117267  1.00000000 -0.36126220
## total_rooms         0.04456798 -0.03609960 -0.36126220  1.00000000
## total_bedrooms      0.06911970 -0.06648391 -0.31902633  0.92705820
## population          0.09977322 -0.10878475 -0.29624424  0.85712597
## households          0.05531009 -0.07103543 -0.30291601  0.91848449
## median_income      -0.01517587 -0.07980913 -0.11903399  0.19804965
## median_house_value -0.04596662 -0.14416028  0.10562341  0.13415311
##                    total_bedrooms   population  households median_income
## longitude             0.069119698  0.099773223  0.05531009  -0.015175865
## latitude             -0.066483906 -0.108784747 -0.07103543  -0.079809127
## house_age            -0.319026332 -0.296244240 -0.30291601  -0.119033990
## total_rooms           0.927058197  0.857125973  0.91848449   0.198049645
## total_bedrooms        1.000000000  0.873534861  0.97436629  -0.007616874
## population            0.873534861  1.000000000  0.90722227   0.004834346
## households            0.974366294  0.907222266  1.00000000   0.013033052
## median_income        -0.007616874  0.004834346  0.01303305   1.000000000
## median_house_value    0.049456862 -0.024649679  0.06584265   0.688075208
##                    median_house_value
## longitude                  -0.04596662
## latitude                   -0.14416028
## house_age                   0.10562341
## total_rooms                 0.13415311
## total_bedrooms              0.04945686
## population                 -0.02464968
## households                  0.06584265
## median_income               0.68807521
## median_house_value          1.00000000
```

```
cor(ca_housing$median_house_value, ca_housing$median_income)
```

```
## [1] 0.6880752
```

```
#Figure 1
#install.packages("corrplot")
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.3
```

```
## corrplot 0.95 loaded
```

```
corrplot(cor_matrix, method = "color")
title("Figure 1. Heatmap correlation", cex.main=1, line=3)
```



```
#Question 2 - Model for Median House Value and Median Income
#Linear Model
inc_house <- lm(median_house_value ~ median_income, ca_housing)
summary(inc_house)
```

```
##
```

```
## Call:
## lm(formula = median_house_value ~ median_income, data = ca_housing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -540697  -55950  -16979   36978  434023
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    45085.6     1322.9   34.08   <2e-16 ***
## median_income  41793.8      306.8  136.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 83740 on 20638 degrees of freedom
## Multiple R-squared:  0.4734, Adjusted R-squared:  0.4734
## F-statistic: 1.856e+04 on 1 and 20638 DF,  p-value: < 2.2e-16
```

```
plot(ca_housing$median_income, ca_housing$median_house_value,
     col = "steelblue",
     pch = 1,
     xlab = "Median Income (ten of thousands of dollars)",
     ylab = "Median House Value (US dollar)",
     cex.main = 0.9,
     main = "Figure 2. Median House Value vs. Median Income and two regression lines")

abline(inc_house,
       col = "red",
       lwd = 2)

#Quadratic Model
inc_house_quad <- lm(median_house_value ~ median_income + I(median_income^2), ca_housing)
summary(inc_house_quad)
```

```
##
## Call:
## lm(formula = median_house_value ~ median_income + I(median_income^2),
##     data = ca_housing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -450948  -55535  -16549   37400  453242
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        20421.11    2257.64   9.045   <2e-16 ***
## median_income      53211.27     902.09  58.987   <2e-16 ***
## I(median_income^2) -1050.42      78.09 -13.451   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 83370 on 20637 degrees of freedom
## Multiple R-squared:  0.478,  Adjusted R-squared:  0.478
## F-statistic:  9450 on 2 and 20637 DF,  p-value: < 2.2e-16
```

```r
order_id <- order(ca_housing$median_income)

lines(x = ca_housing$median_income[order_id],
      y = fitted(inc_house_quad)[order_id],
      col = "black",
      lwd = 2)
```

```r
#Question 3
#Scatterplot - Geographic Distribution of Housing Prices
price_col <- colorRampPalette(c("blue", "yellow", "red"))(100)
value_rank <- cut(ca_housing$median_house_value,
                  breaks = 100,
                  labels = FALSE)

plot(ca_housing$longitude, ca_housing$latitude,
     col = price_col[value_rank],
     pch = 16, cex = 0.6,
     xlab = "Longitude",
     ylab = "Latitude",
     main = "Figure 3.1 Geographic Distribution of Housing Prices")
legend("topright",
       legend = c("Low price", "Mid price", "High price"),
       col = c("blue", "yellow", "red"),
       pch = 16,
       pt.cex = 1)
```

```r
#Boxplot - Housing Prices by Ocean Proximity
par(mar = c(5, 4, 4, 0.1))
ca_housing$ocean_proximity <- factor(
  ca_housing$ocean_proximity,
  levels = c("INLAND", "<1H OCEAN", "NEAR OCEAN", "NEAR BAY", "ISLAND"))
boxplot(median_house_value ~ ocean_proximity,
        data = ca_housing,
        col = "lightblue",
        xlab = "Ocean Proximity",
        ylab = "Median House Value",
        main = "Figure 3.2 Housing Prices by Ocean Proximity")
```

```r
#Question 4 - Histogram of Median House Value
hist(ca_housing$median_house_value,
     col = "olivedrab3",
     breaks = 40,
     main = "Figure 4. Histogram of Median House Value",
     xlab = "Median House Value")
```

```r
#Median Income is right-skewed which indicates a need for a log transformation
hist(ca_housing$median_income,
     col = "deeppink",
     breaks = 40,
     main = "Histogram of Median Income",
     xlab = "Median Income")
```

Appendix C

```r
#Model construction and analysis
#Linear
m1 <- lm(median_house_value~., ca_housing)
summary(m1)
```

```
##
## Call:
## lm(formula = median_house_value ~ ., data = ca_housing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -550613  -42739  -10602   28750  794919
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -2.273e+06  8.646e+04 -26.287  < 2e-16 ***
## longitude                  -2.643e+04  1.014e+03 -26.068  < 2e-16 ***
## latitude                   -2.517e+04  9.998e+02 -25.178  < 2e-16 ***
## house_age                   1.058e+03  4.371e+01  24.203  < 2e-16 ***
## total_rooms                -4.731e+00  7.706e-01  -6.139 8.48e-10 ***
## total_bedrooms              7.134e+01  5.932e+00  12.027  < 2e-16 ***
## population                 -3.929e+01  1.064e+00 -36.928  < 2e-16 ***
## households                  7.780e+01  6.659e+00  11.685  < 2e-16 ***
## median_income               3.876e+04  3.322e+02 116.670  < 2e-16 ***
## ocean_proximity<1H OCEAN    3.977e+04  1.736e+03  22.904  < 2e-16 ***
## ocean_proximityNEAR OCEAN   4.453e+04  2.240e+03  19.878  < 2e-16 ***
## ocean_proximityNEAR BAY     3.607e+04  2.326e+03  15.507  < 2e-16 ***
## ocean_proximityISLAND       1.958e+05  3.083e+04   6.351 2.18e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68730 on 20627 degrees of freedom
## Multiple R-squared:  0.6455, Adjusted R-squared:  0.6452
## F-statistic:  3129 on 12 and 20627 DF,  p-value: < 2.2e-16
```

```r
#Log-linear
m2 <- lm(log(median_house_value) ~., data = ca_housing)
summary(m2)
```

```
##
## Call:
## lm(formula = log(median_house_value) ~ ., data = ca_housing)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3585 -0.1990 -0.0089  0.1911  3.4319
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -2.500e+00  4.145e-01  -6.031 1.66e-09 ***
## longitude                  -1.604e-01  4.861e-03 -32.999  < 2e-16 ***
## latitude                   -1.562e-01  4.794e-03 -32.588  < 2e-16 ***
## house_age                   2.454e-03  2.095e-04  11.709  < 2e-16 ***
## total_rooms                -8.305e-06  3.695e-06  -2.248   0.0246 *
## total_bedrooms              2.703e-04  2.844e-05   9.503  < 2e-16 ***
## population                 -1.779e-04  5.101e-06 -34.872  < 2e-16 ***
## households                  3.628e-04  3.192e-05  11.366  < 2e-16 ***
## median_income               1.670e-01  1.593e-03 104.855  < 2e-16 ***
## ocean_proximity<1H OCEAN    3.099e-01  8.324e-03  37.231  < 2e-16 ***
## ocean_proximityNEAR OCEAN   2.781e-01  1.074e-02  25.891  < 2e-16 ***
## ocean_proximityNEAR BAY     2.725e-01  1.115e-02  24.436  < 2e-16 ***
## ocean_proximityISLAND       9.115e-01  1.478e-01   6.166 7.14e-10 ***
## ---
## 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
```

```
#Double-log
m3 <- lm(log(median_house_value) ~ log(median_income) + house_age + total_rooms + total_bedrooms + popul
summary(m3)
```

```
##
## Call:
## lm(formula = log(median_house_value) ~ log(median_income) + house_age +
##     total_rooms + total_bedrooms + population + households +
##     longitude + latitude + ocean_proximity, data = ca_housing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4092 -0.2031 -0.0162  0.1874  3.2751
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -1.997e+00  4.110e-01  -4.858  1.2e-06 ***
## log(median_income)          6.933e-01  6.433e-03 107.765  < 2e-16 ***
## house_age                   2.982e-03  2.082e-04  14.325  < 2e-16 ***
## total_rooms                -1.276e-05  3.661e-06  -3.485 0.000494 ***
## total_bedrooms              3.168e-04  2.825e-05  11.212  < 2e-16 ***
## population                 -1.656e-04  5.070e-06 -32.674  < 2e-16 ***
## households                  2.789e-04  3.163e-05   8.817  < 2e-16 ***
## longitude                  -1.520e-01  4.824e-03 -31.512  < 2e-16 ***
## latitude                   -1.483e-01  4.757e-03 -31.171  < 2e-16 ***
## ocean_proximity<1H OCEAN    3.031e-01  8.250e-03  36.732  < 2e-16 ***
## ocean_proximityNEAR OCEAN   2.803e-01  1.063e-02  26.360  < 2e-16 ***
```

```
## ocean_proximityNEAR BAY    2.708e-01  1.104e-02  24.521  < 2e-16 ***
## ocean_proximityISLAND      8.824e-01  1.464e-01   6.027  1.7e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3263 on 20627 degrees of freedom
## Multiple R-squared:  0.6714, Adjusted R-squared:  0.6712
## F-statistic:  3512 on 12 and 20627 DF,  p-value: < 2.2e-16
```

```
#Table 2 - Stargazer
#install.packages("AER")
library(AER)
```

```
## Loading required package: car
```

```
## Warning: package 'car' was built under R version 4.3.3
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.3.3
```

```
## Loading required package: lmtest
```

```
## Warning: package 'lmtest' was built under R version 4.3.3
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.3.3
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
## Warning: package 'sandwich' was built under R version 4.3.3
```

```
## Loading required package: survival
```

```
rob_se <- list(sqrt(diag(vcovHC(m1, type = "HC1"))),
               sqrt(diag(vcovHC(m2, type = "HC1"))),
               sqrt(diag(vcovHC(m3, type = "HC1"))))

#install.packages("stargazer")
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```r
#install.packages("tinytex")
#tinytex::install_tinytex()
stargazer(m1,m2,m3,
          digits = 3,
          header = FALSE,
          type = "html",
          se = rob_se,
          out = "prj2.html",
          title = "Table 2. Multiple Regression Models",
          model.numbers = FALSE,
          column.labels = c("(1)", "(2)", "(3)"))
```

```
##
## <table style="text-align:center"><caption><strong>Table 2. Multiple Regression Models</strong></capt
## <tr><td colspan="4" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left"
## <tr><td></td><td colspan="3" style="border-bottom: 1px solid black"></td></tr>
## <tr><td style="text-align:left"></td><td>median_house_value</td><td colspan="2">log(median_house_valu
## <tr><td style="text-align:left"></td><td>(1)</td><td>(2)</td><td>(3)</td></tr>
## <tr><td colspan="4" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left"
## <tr><td style="text-align:left"></td><td></td><td></td><td>(0.010)</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td style="text-align:left">longitude</td><td>-26,430.440<sup>***</sup></td><td>-0.160<sup>***</s
## <tr><td style="text-align:left"></td><td>(1,036.878)</td><td>(0.006)</td><td>(0.006)</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td style="text-align:left">latitude</td><td>-25,173.280<sup>***</sup></td><td>-0.156<sup>***</su
## <tr><td style="text-align:left"></td><td>(1,043.372)</td><td>(0.006)</td><td>(0.006)</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td style="text-align:left">house_age</td><td>1,057.816<sup>***</sup></td><td>0.002<sup>***</sup
## <tr><td style="text-align:left"></td><td>(51.412)</td><td>(0.0002)</td><td>(0.0002)</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td style="text-align:left">total_rooms</td><td>-4.731<sup>***</sup></td><td>-0.00001</td><td>-0
## <tr><td style="text-align:left"></td><td>(1.223)</td><td>(0.00001)</td><td>(0.00001)</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td style="text-align:left">total_bedrooms</td><td>71.345<sup>***</sup></td><td>0.0003<sup>***</s
## <tr><td style="text-align:left"></td><td>(10.022)</td><td>(0.00004)</td><td>(0.00005)</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td style="text-align:left">population</td><td>-39.287<sup>***</sup></td><td>-0.0002<sup>***</su
## <tr><td style="text-align:left"></td><td>(4.639)</td><td>(0.00002)</td><td>(0.00002)</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td style="text-align:left">households</td><td>77.804<sup>***</sup></td><td>0.0004<sup>***</sup>
## <tr><td style="text-align:left"></td><td>(14.931)</td><td>(0.0001)</td><td>(0.0001)</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td style="text-align:left">median_income</td><td>38,760.450<sup>***</sup></td><td>0.167<sup>***
## <tr><td style="text-align:left"></td><td>(526.662)</td><td>(0.003)</td><td></td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td style="text-align:left">ocean_proximity< 1H OCEAN</td><td>39,766.400<sup>***</sup></td><td>0
```

```
## <tr><td style="text-align:left"></td><td>(1,633.009)</td><td>(0.010)</td><td>(0.009)</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td style="text-align:left">ocean_proximityNEAR OCEAN</td><td>44,525.150<sup>***</sup></td><td>0
## <tr><td style="text-align:left"></td><td>(2,398.798)</td><td>(0.013)</td><td>(0.013)</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td style="text-align:left">ocean_proximityNEAR BAY</td><td>36,069.000<sup>***</sup></td><td>0.2
## <tr><td style="text-align:left"></td><td>(2,485.138)</td><td>(0.013)</td><td>(0.012)</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td style="text-align:left">ocean_proximityISLAND</td><td>195,832.100<sup>***</sup></td><td>0.91
## <tr><td style="text-align:left"></td><td>(36,424.760)</td><td>(0.108)</td><td>(0.121)</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td style="text-align:left">Constant</td><td>-2,272,858.000<sup>***</sup></td><td>-2.500<sup>***
## <tr><td style="text-align:left"></td><td>(87,013.100)</td><td>(0.479)</td><td>(0.481)</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td colspan="4" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left"
## <tr><td style="text-align:left">R<sup>2</sup></td><td>0.645</td><td>0.665</td><td>0.671</td></tr>
## <tr><td style="text-align:left">Adjusted R<sup>2</sup></td><td>0.645</td><td>0.665</td><td>0.671</td
## <tr><td style="text-align:left">Residual Std. Error (df = 20627)</td><td>68,730.970</td><td>0.330</t
## <tr><td style="text-align:left">F Statistic (df = 12; 20627)</td><td>3,129.289<sup>***</sup></td><td
## <tr><td colspan="4" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left"
## </table>
```