Statistical Analysis of Housing Affordability in California

By: Ananya Ratakonda, Dung Hoang, and Jamie Hui

From the lively and beautiful environment that California has, it's no wonder people steer towards a residency in California. However, the affordability of living in this state continues to be a crucial consideration and concern for both current residents and prospective movers. As housing prices continue to rise, people begin to wonder where the affordable areas in California may be to relocate. To assist individuals considering a new residency or continuing residency in California with this concern, our objective in this research is to understand possible factors that may influence the cost of living in California through applying statistical learning applications. We shall explore factors like median housing values, median income, proximity to the coast, and distance from metropolitan areas to shape our analysis on the cost of living in California.

Our analysis will focus on the following questions:

- Is there a correlation between median household income and housing values in California?
- Is there a relationship between the proximity of certain metropolitan areas and Medium Housing Values?

In our project, we examine a dataset on California's housing prices that pertains median house prices for various California districts from the 1990 census which we will use to extend our analysis to a diverse range of audience. This dataset captures data points of houses in California from 1990 along with 14 additional variables: Median House Value, Median Income, Median Age, Total Rooms, Total Bedrooms, Population, Households, Latitude, Longitude, Distance to Coast, Distance to Los Angeles, Distance to San Diego, Distance to San Jose, and Distance to San Francisco.

While the data may be dated, our analysis on this dataset is still important to current and prospective residents, since understanding how various factors affect housing prices can help people make informed decisions about their housing and lifestyle preferences. By learning more about the economic determinants of California, the audience of our research will be able to have a deeper understanding of the state's cost of home owning.

Our Hypothesis

Primary challenges in our analysis arise when delving into California's housing market with various indicators affecting housing costs in our data, particularly proximities to urban cities and one's median income. Preventing the large scope and maintaining our examination focused on the housing dynamics within California, our hypotheses center around examining the relationship between California households' median income and housing value, as well as evaluating the correlation of different California metropolitan areas on housing prices.

Data Processing

Understanding that our analysis will mainly focus on the correlation between median income for households and median house values in California as well as how geographical differences impact the median house values, we prepared our dataset for these subsequent analyses. After removing missing rows and removing outliers from the Median House Value variable, we subsetted our data to capture the relation between Median House Value to Median Income and Median House Value to different metropolitan areas. We removed outliers that were greater than 500,000 for Median House Value because there were far too many outliers which caused the distribution to be very skewed and ruin our Gamma distribution. For the Median Income, since the original data set values were represented in the 10,000s, we multiplied all Median Income values by 10,000. For example, one of our values was 8.3252 and we changed it to 83,252 in order to convert the values so they will be in terms of 10k. This change would allow our audience to easily decipher the data without having to recalculate the values themselves.

We decided to have two different subsets: one with just Median Income and Median House Value, and another with Median House Value, Distance to Coast, Distance to Los Angeles, Distance to San Diego, Distance to San Jose, and Distance to San Francisco. The first subset was used to analyze the correlation between Median House Value and Median Income and whether Median Income had an effect on Median House Value. The second subset was used to see if there is a correlation between the Median House Value and the different distances. Also to see which Distance had the most effect on the Median House Value.

After changing our dataset this was how the summary looked like for both sub-datasets:

```
 Median_House_Value Median_Income
 Min.   : 14999     Min.   :  4999
 1st Qu.:116475     1st Qu.: 25263
 Median :173600     Median : 34490
 Mean   :192055     Mean   : 36764
 3rd Qu.:247900     3rd Qu.: 45825
 Max.   :499100     Max.   :150001
 Median_House_Value Distance_to_coast  Distance_to_LA      Distance_to_SanDiego Distance_to_SanJose Distance_to_SanFrancisco
 Min.   : 14999     Min.   :   120.7   Min.   :    420.6   Min.   :    484.9    Min.   :   569.4    Min.   :   456.1
 1st Qu.:116475     1st Qu.:  9851.1   1st Qu.:  33053.6   1st Qu.: 158890.6    1st Qu.:117875.9    1st Qu.:120402.5
 Median :173600     Median : 21327.1   Median : 177200.7   Median : 223113.6    Median :458025.4    Median :524571.7
 Mean   :192055     Mean   : 41967.7   Mean   : 271815.9   Mean   : 399800.7    Mean   :349802.7    Mean   :387223.1
 3rd Qu.:247900     3rd Qu.: 53250.1   3rd Qu.: 528960.0   3rd Qu.: 707671.8    3rd Qu.:517889.8    3rd Qu.:585659.7
 Max.   :499100     Max.   :333804.7   Max.   :1018260.1   Max.   :1196919.3    Max.   :836762.7    Max.   :903627.7
```

Figure 1: Summaries of both Sub-Datasets

Looking at Figure 1 we can see that Median House Value has a median of around 173,600 and the mean house value that California households pay for a house is about $192,055. Therefore, we can also assume that the median housing value in California is around $173,600. Based on these summaries, the range for how much California households spend on a house is around $116,475 to $247,900, which was very expensive for 1990. Moving onto the Median Income, we can see that a median California household earns around $34,490 and the mean is around $36,764. A good estimate of how much California residents earn in order to afford living here was around $25,263 to $45,825. Now, looking at the distances we can see that out of all distances, the distance to coast has the lowest max value so if one were to live in California at worst case an estimate of how far away you are from the coast would be 333804.7 meters. However it seems out of all the distances, Distance to Coast has the lowest distance, meaning

that no matter where you live in California, you will be somewhat close to the coast. All other distances seem to be further away and also keeping in mind that California is the second largest state in America, so depending on where you live it might be closer to south california with Los Angeles and San Diego or be closer to north california with cities like San Jose and San Francisco, as calculated by the original dataset. If our audience is interested in learning more information about the dataset, we have provided a citation of the original set at the end of this report.

Even though this dataset is from 1990 and incorporates housing median prices from that time period, the information can still be used by potential buyers for California homes today. Looking at the increasing living costs in California, our 1990 dataset provides a baseline in the overall housing market for our audience to reference and make inferences as they continue on with their personal house hunting in the present day. They can see how geographical location and median income plays a role in affecting the median house values of those areas, and decide which location may be more profitable to reside in long-term.

3. Visualization and Methodology

3.1 Median Housing Value vs Median Income

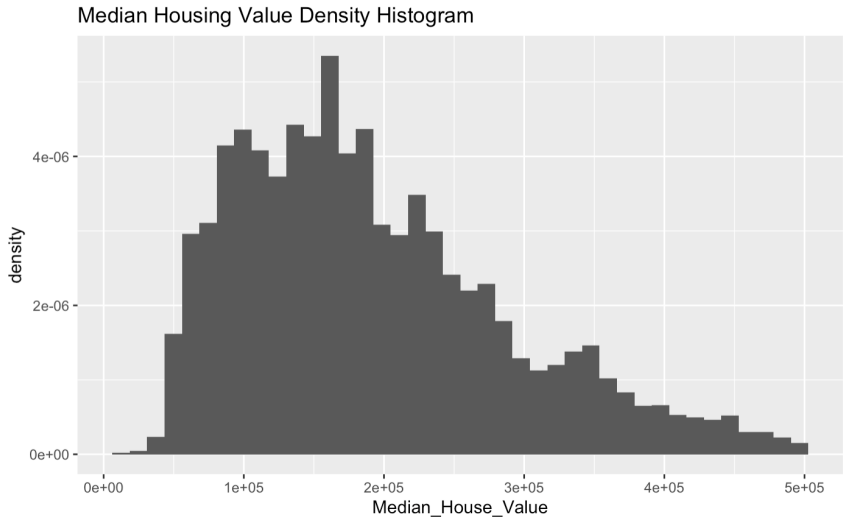

Figure 2: Histogram of Median House Value

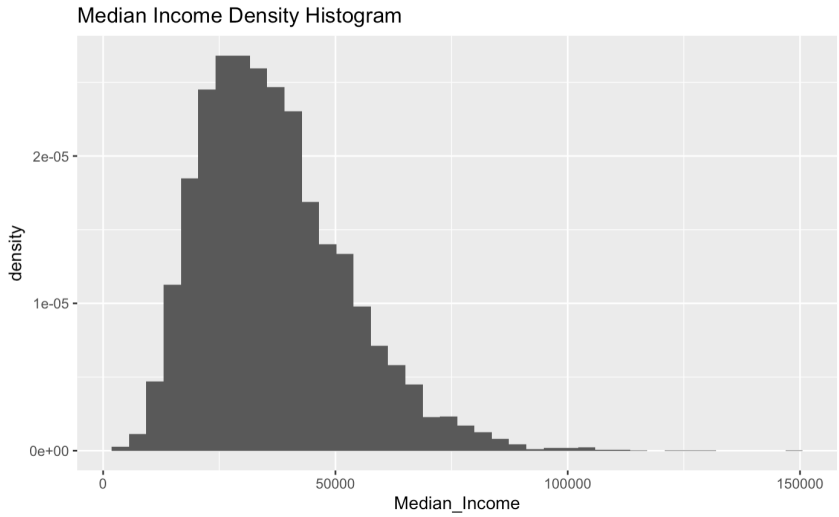

Figure 3: Histogram of Median Income

Our early understanding of the correlation between Median House Value and Median Income is through utilizing histograms to observe possible distributional patterns. We observed that Figure 2 appears to have a Normal-Gamma distribution with a slightly right skewness and Figure 3 appears more apparently skewed to the right, indicating a Gamma distribution. Furthermore our correlation coefficient (0.6467) indicated a moderately strong positive relationship between the two variables. The observations made thus far informed our decision to use a gamma distribution with a log link function to fit our simple linear regression model as it best reflects how our data is right-skewed with continuous positive values. We discovered that the effect of Median Income on Median Housing Value is that if Median Income increases by one then log of Median Housing Value increases by 0.00002008.

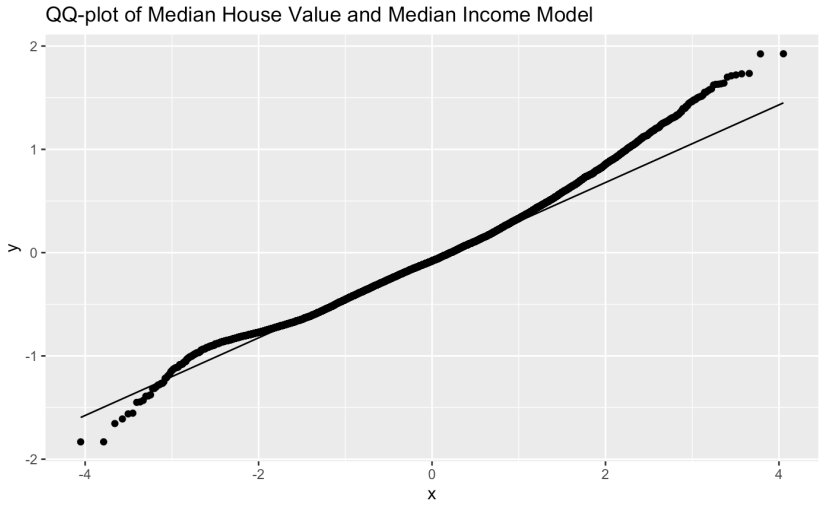

Figure 4: QQ-Plot of Median House Value and Median Income

While performing linear regression we decided that using a gamma distribution would be the best option because our data is right skewed with continuous variables. Poisson could have also

been an option however our data is not discrete so we decided not to use it. For the link after debating between identity and log, we decided to use log as the QQ-Plot looked more aligned with the line. Figure 4's QQ-Plot serves as an important way to assess the normality of our linear regression. As we can see the data points on the plots are all gathered very closely to the line. However, the data points at the top right corner do appear to have slightly deviated from the line, indicating slight right-skewness. Continuing our analysis, we examine two types of model where one model includes the Median Income variable and one model does not include the Median Income predictor. Comparing these two models will help us better understand if the Median Income predictor will better explain the variability in Median House Value that we see. Through testing with the AIC and BIC model selection criteria, it is clear that the model with the Median House Value predictor is more favorable, because both of it's AIC (493997.7) and BIC (494021.3) values are lower. A lower AIC and BIC value means that the model with Median House Value is a better fit to explaining the variability in Median House Value and overall generalization of the data.

Through these analyses, we can gather that people with higher median income move to areas with higher median house values. The potential demand from higher-income households in these geographical areas may influence the appreciation of their property values, which can become a challenge for lower median income households who are unable to meet the housing market's pricing demands. The socioeconomic commentary here may be of importance to our audience as they use this information to observe best fitted areas according to how their income aligns and ultimately decide whether California is the state to relocate to. However, this information alone does not determine relocation recommendations or contributing factors to the overall housing price variation in California. Deeper exploration of the dataset is essential to developing a more holistic understanding of the factors that impact California housing prices.

3.2 Housing Prices vs. Distance from Metropolitan Cities

In this exploration, we look at California's housing market, specifically seeing how location-related factors can influence the Median House Value. Looking at the relationship between the Median House Value and the distance from Metropolitan areas, this section focuses on the impact of the geographic proximity on the Median House Value, seeing the relevant predictors that shape the Median House Value in California. To explore this, multiple linear regression models were used in order to compare a full model that employs all of the potential predictors with simpler alternative models. Overall, to picture which factors greatest impact on the Median House Value.

Full Model

The Full Model, incorporates all of the available predictors such as distance to the coast, Los Angeles, San Diego, San Jose, and San Francisco. Each coefficient in the full model assumes the extent to which proximity to different cities influences the median house values in California. The polarity of these coefficients reveals the nature of the correlation observed between the predictor variables (distance to a city) and the response variable (Median House Value), giving insights on what influences California's median house values. For example, a positive coefficient presented in the regression model reveals a positive correlation, indicating that as the distance of a specific city decreases, the Median House Value increases too. On the other hand, a negative coefficient portrays a negative correlation suggesting that as the distance to a city decreases, the

Median House Value tends to decrease too. A full model is with all of these available predictors – Distance_to_coast, Distance_to_LA, Distance_to_SanDiego, Distance_to_SanJose, and Distance_to_SanFrancisco – regressed against Median_House_Value. The regression model would look like:
Median House Value = β0 + β1*Distance_to_coast + β2*Distance_to_LA + β3*Distance_to_SanDiego + β4*Distance_to_SanJose + β5*Distance_to_SanFrancisco.

For the Full Model, the summary shows as:

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.79271  -0.33919  -0.08473   0.21289   2.64683

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.292e+01  3.161e-02 408.811  < 2e-16 ***
Distance_to_coast        -5.280e-06  7.394e-08 -71.416  < 2e-16 ***
Distance_to_LA           -6.690e-07  4.379e-08 -15.275  < 2e-16 ***
Distance_to_SanDiego     -1.777e-07  5.671e-08  -3.134  0.00172 **
Distance_to_SanJose       2.497e-07  1.396e-07   1.789  0.07360 .
Distance_to_SanFrancisco -1.076e-06  1.522e-07  -7.068 1.62e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.1837657)

    Null deviance: 5227.8  on 19647  degrees of freedom
Residual deviance: 3462.8  on 19642  degrees of freedom
AIC: 495071

Number of Fisher Scoring iterations: 6
```

Figure 5: Results for the Full Model of Housing and Distances

Looking at the results from the full model linear regression analysis we can see that all the estimates except for Distance to San Jose and the intercept had a negative number. This would mean that as Distance to coast increases by one then Median House Value decreases by 5.280e-06. As Distance to LA increases by one then Median House Value decreases by 6.690e-07, Distance to San diego increases by one then Median House Value decreases by 1.777e-07, Distance to San Francisco increases by one then Median House Value decreases by 1.076e-06. The only distance that the Median House Value increases is when Distance to San Jose increases by one then Median House Value increases by 2.497e-07. For the Null hypothesis of $\beta i = 0$, where i would be all the different distance variables. The Alternative Hypothesis would be that $\beta i \neq 0$. The distances that are significant right now at alpha 0.001 would be Distance to coast, Distance to LA, and Distance to San Francisco. Therefore we can reject the null hypothesis for Distance to coast, Distance to LA, and Distance to San Francisco. This would mean that Distance to coast, Distance to LA, and Distance to San Francisco do have a significant relationship with Median House Value.

Anova Model for the Full Model

```
Analysis of Deviance Table

Model: Gamma, link: log

Response: Median_House_Value

Terms added sequentially (first to last)


                         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                     19647     5227.8
Distance_to_coast         1  1595.52     19646     3632.2 < 2.2e-16 ***
Distance_to_LA            1     7.01     19645     3625.2 6.485e-10 ***
Distance_to_SanDiego      1    84.82     19644     3540.4 < 2.2e-16 ***
Distance_to_SanJose       1    68.13     19643     3472.3 < 2.2e-16 ***
Distance_to_SanFrancisco  1     9.49     19642     3462.8 6.694e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6: ANOVA (Analysis of Variance) table which shows all of the distance to _ predictor variables

Looking at the results of the ANOVA table above, it describes the overall significance of each variable and how the distance to the coast, Los Angeles, San Diego, San Jose, San Francisco differs as being a significant indicator of a median house value in California. All of the predictor variables have low p-values which emphasizes that each variable is significant and contributes to the Median Housing Value in California. If we were to perform a F-test: The Null hypothesis would be $\beta1 = \beta2 = \beta3 = \beta4 = \beta5 = 0$. The Alternative Hypothesis would be $\beta1 \neq 0$ or $\beta2 \neq 0$ or $\beta3 \neq 0$ or $\beta4 \neq 0$ or $\beta5 \neq 0$. Looking at Figure 6, the Anova results for the full model we can reject the null hypothesis at alpha equals 0.001 because all the Pr(>Chi) values are less than alpha. This means that distances do have a significant relationship with Median House Value.

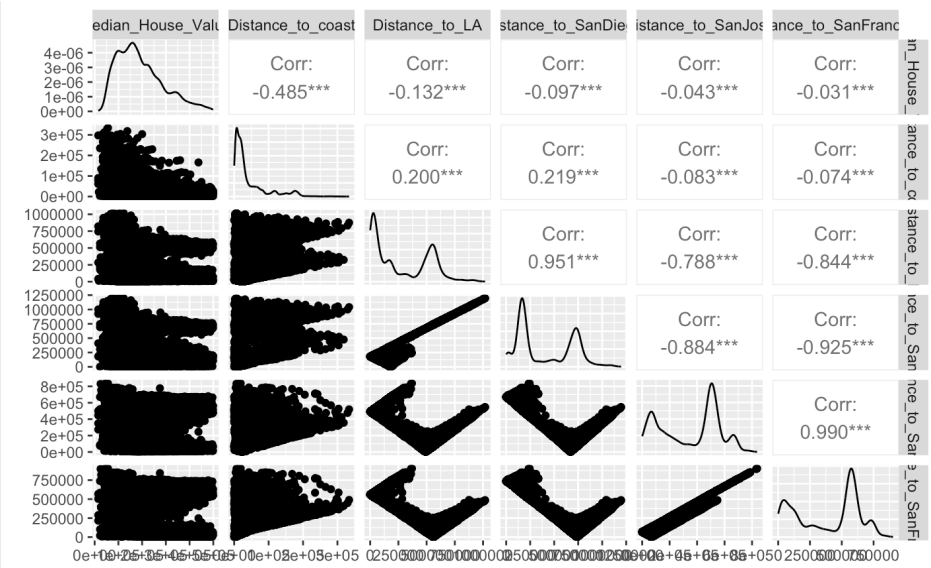

Figure 7: Ggpairs of median house value vs distance to different metropolitan areas

Looking at Figure 7, it represents a visual summary plot that assesses how the median house values and the distance to all of the different metropolitan areas have a pattern. Each of the scatterplots reveal how the distances to a specific city changes, seeing if there is a linear pattern.

It seems that all correlations with the Median Housing Values and the Distances are all negative values. Additionally, we started our analysis by comparing the Full Model with all the predictor variables with a Bad Model, where the intercept is only added to give insight on how the Full Model gives an overall accurate representation on the relationship between the proximity to cities with Median House Values in California.

Reduced Model

Using a reduced model that has a simplified version of the full model, includes fewer predictor variables. A reduced model represented by the smallerModel in the data, includes the predictor variables distances to the coast, Los Angeles, and San Francisco because these were the variables that were significant from the full model. Comparing the reduced model with the full model that has all of the predictor variables, the reduced model is the better model with the two predictor variables.

```
[1] "The AIC for Bad Model: 503447.345495"
[1] "The BIC for Bad Model: 503463.116957"
[1] "The AIC for Full Model: 495071.431407"
[1] "The BIC for Full Model: 495126.631523"
[1] "The AIC for Reduced Model: 495078.435130"
[1] "The BIC for Reduced Model: 495117.863784"
```

Figure 8: Full Model vs. Reduced Model Analysis of Deviance Comparison (AIC vs. BIC)

AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are both measures that critique models in their goodness of fit. AIC and BIC are overall important in our California housing data analysis because it gives model selection insights. According to our data analysis, we saw that the Bad Model that has only the Median Housing Value and the dataset's intercept is the least favorable model by looking at the AIC and BIC since it has the largest value. The Smaller Model (predictors – distance to the coast, Los Angeles, and San Francisco) would be the most favorable since noted in Figure 8 above, this model has the lowest BIC despite the AIC being greater than the AIC for the full model, because we decided to favor the BIC. Through these findings, the Smaller Model is the most accurate model that predicts the Median Housing Value the most effectively. This means that the model with predictors – distance to the coast, Los Angeles, and San Francisco – is the best model to see the interplay between proximity to cities and the Median House Value.

Anova Model for the Full Model vs Reduced Model

```
Analysis of Deviance Table

Model 1: Median_House_Value ~ Distance_to_coast + Distance_to_LA + Distance_to_SanFrancisco
Model 2: Median_House_Value ~ (Distance_to_coast + Distance_to_LA + Distance_to_SanDiego +
    Distance_to_SanJose + Distance_to_SanFrancisco)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     19644     3464.7
2     19642     3462.8  2   1.8846 0.005929 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 9: Full Model vs. Reduced Model Anova Results

If we were to perform a test using this result we would get: H0: $\beta1 = \beta2 = \beta5 = 0$, HA: $\beta1 \neq 0$ or $\beta2 \neq 0$ OR or $\beta5 \neq 0$. We cannot reject the null hypothesis at alpha 0.001 because the Pr(>Chi) value is greater than 0.001 therefore we choose the small model and get rid of the full model.

Full Model vs. Reduced Model QQ Plots

Looking at the following two QQ (Quantile - Quantile) plots in Figure 10 and 11, these plots compare the quantiles of the Full Model with the response variable, Median House Value, and all of location distance predictor variables. The Reduced Model has the response variable, Median House Value, and three of the location distance predictor variables.

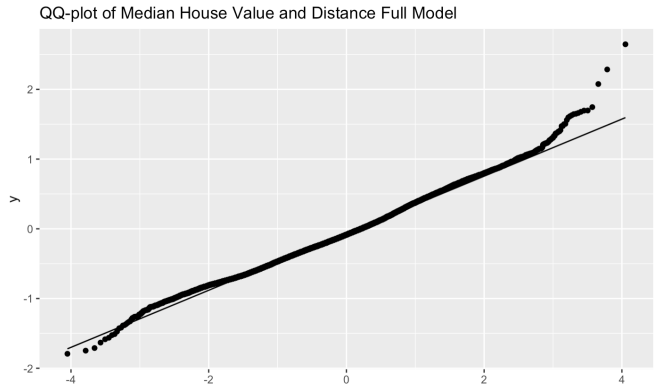

Figure 10: Full-Model QQ Plot of Median House Value and Distance to Coast, Los Angeles, San Diego, San Jose, San Francisco

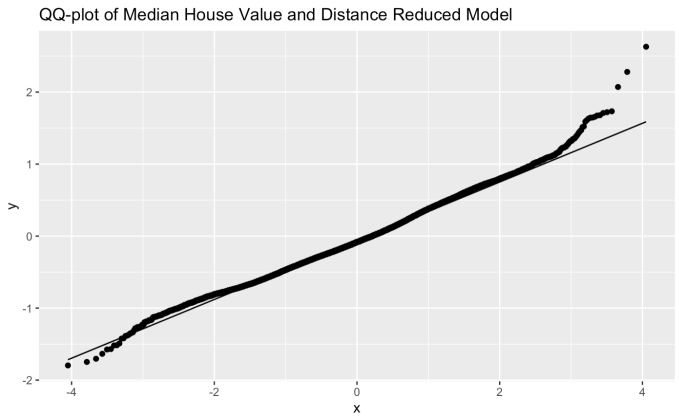

Figure 11: Reduced-Model QQ Plot of Median House Value and Distance to coast, Los Angeles, San Francisco

Looking at the residuals in the QQ Plots, it's evident to see that both the Full Model and the Reduced Model have a similar pattern, which is a roughly straight line. Figure 10 and Figure 11 both have very similar QQ plots and have three notable outliers, which show a departure from the assumption of normality. Comparing Figures 11 and 10's QQ plot to the QQ-Plot of Median House Value and Median Income (Figure 4), Figures 10 & 11 have a more robust normal distribution with less deviations from the linear line than Figure 4's plot. This emphasizes how comparing Median House Value to distances to metropolitan California locations is estimated to be more reliable in explaining overall Median House Values in California.

The main evaluation in choosing the Reduced Model as the better model is through comparing the AIC, BIC and Anova. Overall, when comparing the AIC and BIC of each model like done in the section before and also the Anova model, it emphasizes the Reduced Model as the most suitable predictor model between California housing and the proximity of metropolitan areas.

4. Conclusion

Overall, multiple factors have an effect on housing prices in California to an extent. From the relationship of Median Income on Median House Values and Distance from Metropolitan Cities on Median House Values, we found that these variables play an important role in determining the median house values in California. For one, we found that households with higher median income could live in areas with higher median value houses. This explains the lack of affordability variations in those higher median value housing areas, because people with a higher median income are willing to pay more for these houses which can cause the housing prices in those areas to adjust to the higher demands. As for Distance from Metropolitan Cities on Median House Values, we found that a metropolitan area significantly influences a higher median housing area specifically looking at how being in proximity to California's most known cities, San Francisco and Los Angeles, increases one's median house value. That explains the significant impact of geographical location on housing affordability in California. This project ultimately highlights the significance that Median Income and Distance from Metropolitan areas plays in affecting the Median House Value in California. With today's socioeconomic dynamics in mind, owning a home in California is not impossible but it does require a certain level of income or a willingness to adjust one's lifestyle, like choosing to live further from a metropolitan area in order to afford home ownership in California.

Dataset we used:
https://www.kaggle.com/datasets/fedesoriano/california-housing-prices-data-extra-features

Contributions:
Ananya Ratakonda - All Code, Data Processing, Anova testing, Full model hypothesis testing.
Dung Hoang - Introduction, Data Processing, 3.1 Median Housing Value vs. Median Income section, Conclusion
Jamie Hui - Hypothesis, QQ Plot for Full vs. Reduced Model, Full Model + Reduced Model Section