7CCMMS61 - Statistics for Data Analysis
Analysis of Californian Residential Block Housing (2001)

Alexander Smerdon
Department of Informatics
23031306

December 2023

Contents

1 Introduction **3**

2 Statistical Methods **3**

3 Exploratory Data Analysis Results **3**

4 Forecasts **7**

5 Conclusion **16**

Summary/Abstract

1 Introduction

The data-set is about the median house price of residential block housing in the US state of California, and was collected in 2001. A block is a geographic descriptor of an area (typically the intersection between roads), which in this context contains homes. These homes can vary from single-family detached homes to large apartment complexes.

This data-set has data on 20640 residences, and has information about their location (including their proximity to the ocean), their age, the total rooms (and bedrooms), the population of a block, the number of households on a block, the median income of all the houses on a block, and finally the median house value.

The data-set was presumably collected as a general census, and could be used to get information on income and wealth distributions in California, as well as the general value of properties across the state.

This report is trying to address the impact of different variables on the median house value. It will first cover the methods used in RStudio to analyse the data during exploratory data analysis, and what this means for building a statistical model. The report will then cover a linear regression model built to try and model the impact of different variables on the median house value, and evaluate whether or not this model is successful. Finally, it will discuss the key insights we can get from this model, and the limitations in the analysis.

2 Statistical Methods

Several statistical methods have been used to analyse the data. This includes getting the summary information about the dataset, displaying this as a box plot, getting the descriptive statistics of each variable in the dataset for analysis, and using scatter plots of variables to assess relationships between any variables. A linear regression model has also been used to try and predict the median house value in California. I felt these were appropriate because it gives a good overview of all the variables in dataset, and allows for testing to make sure that the linear regression model is accurate.

3 Exploratory Data Analysis Results

Figure 1 (below) shows the summary information about the data-set. This includes a five-number summary of each relevant variable. A five-number summary contains information on each variable's minimum value, it's IQR, and it's maximum value. The IQR are the values at the 25th percentile, median (middle

value), and 75th percentile. The summary in Figure 1 also contains the mean, which all the values summed, and divided by the amount.

Figure 1. Five number summary (plus mean) for California Housing Variables.

House Median Age (Years)	
Min.	1.00
1st Qu.	18.00
Median	29.00
Mean	28.64
3rd Qu.	37.00
Max.	52.00

Table 1. House Median Age per Block

Total Rooms	
Min.	2
1st Qu.	1448
Median	2127
Mean	2636
3rd Qu.	3148
Max.	39320

Table 2. Total Rooms per Block

Total Bedrooms	
Min.	1.0
1st Qu.	296.0
Median	435.0
Mean	537.9
3rd Qu.	647.0
Max.	6445.0

Table 3. Total Bedrooms per Block

Population	
Min.	3
1st Qu.	787
Median	1166
Mean	1425
3rd Qu.	1725
Max.	35682

Table 4. Population per Block

Households	
Min.	1.0
1st Qu.	280.0
Median	409.0
Mean	499.5
3rd Qu.	605.0
Max.	6082.0

Table 5. Households per Block

Median Income	
Min.	0.4999
1st Qu.	2.5634
Median	3.5348
Mean	3.8707
3rd Qu.	4.7432
Max.	15.0001

Table 6. Median Income per Block

For Figure 1, it's important to highlight the range in the number of total rooms (Table 2), total bedrooms (Table 3), and households per block (Table 5), as it demonstrates the large amount of variation between the sizes of blocks. These statistical outliers will need to be kept in mind during further exploratory data analysis, and during regression modelling. Any variable relating to location has been omitted here.

These summaries can also be shown in box plots (Figure 2) below:

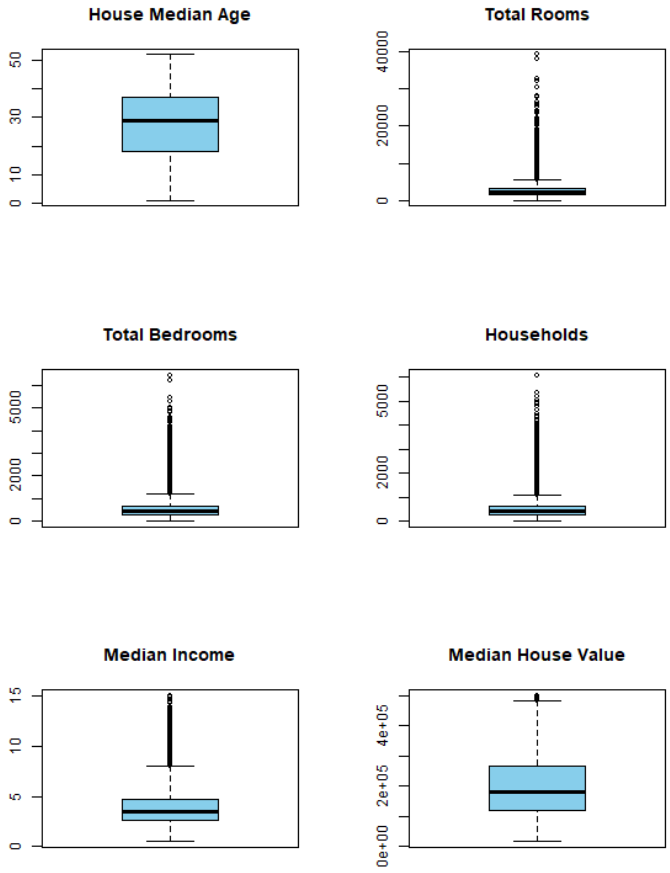

Figure 2: **B**ox-plots of variables in Dataset.

The Median Income box plot shows that there are a lot of very high-income (outside of the third-quartile) earners in California. This makes sense given the high-earning industries California is known for (film and technology).

In Figure 2, the IQR of Total Rooms, Total Bedroom and Households is very small. This is because all three cases there are many outliers that are much larger than the IQR. This may reflect the differences in block sizes in California, and should be kept in mind during subsequent modelling.

Table 7: Descriptive Statistics

Variable	Mean	Median	SD	Skewness	Kurtosis
longitude	-119.570	-118.490	2.004	-0.298	-1.330
latitude	35.632	34.260	2.136	0.466	-1.118
house_median_age	28.639	29.000	12.586	0.060	-0.801
total_rooms	2635.763	2127.000	2181.615	4.147	32.619
total_bedrooms	537.871	435.000	421.385	3.459	21.977
population	1425.477	1166.000	1132.462	4.935	73.528
households	499.540	409.000	382.330	3.410	22.050
median_income	3.871	3.535	1.900	1.646	4.950
median_house_value	206855.817	179700.000	115395.616	0.978	0.327

The following are some descriptive statistics for each of the variables in the California housing data set (taken from Table 7). This can help us assess the normality of the data (which we need to do before building a linear regression model):

Longitude: negative skewness (-0.298) indicates a very slight leftwards skew, whilst a Kurtosis value of -1.330 suggests a Platykurtic shape (lighter tails and a flatter peak compared to a normal distribution).

Latitude: positive skewness (0.466) suggests a very slight rightward skew, whilst a Kurtosis value of -1.118 suggests a similar shape to the longitude.

House Median Age: a very small positive skewness (0.060) suggests that it's very close to a normal distribution, whilst a Kurtosis value of -0.801 suggests a Platykurtic shape.

Total Rooms: positive skewness (4.147) suggests a rightward skew, whilst a Kurtosis value of 32.619 suggests a strong Leptokurtic shape (larger peak than a normal distribution, with longer tails).

Total Bedrooms: positive skewness (3.459) suggests a rightward skew, whilst a Kurtosis value of 21.977 suggests a strong Leptokurtic shape.

Population: positive skewness (4.935) suggests a rightward skew, whilst a Kurtosis value of 73.528 suggests a very strong Leptokurtic shape.

Households: positive skewness (3.410) suggests a rightward skew, whilst a Kurtosis value of 22.050 suggests a strong Leptokurtic shape.

Median Income: positive skewness (1.646) suggests a rightward skew, whilst a Kurtosis value of 4.950 suggests a Leptokurtic shape.

Median House Value: positive skewness (0.978) suggests a rightward

skew, whilst a Kurtosis value of 0.327 suggests a Platykurtic shape.

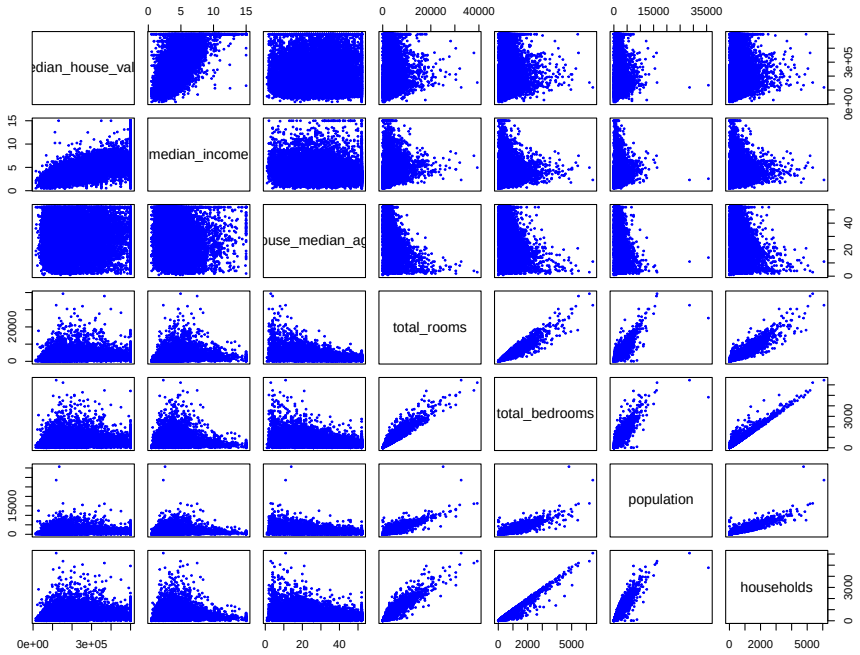

Figure 3: Scatter plot matrix of variables. These plots are constructed to see the relationships of the variables within the dataset.

Figure 3 shows that there is a strong correlation between several variables, including between total rooms and total bedrooms, between total rooms and population, and between population and households. This indicates dependency between the variables. Although this would typically mean that these variables should be omitted from any linear regression model, only total bedrooms will be omitted for the sake of testing out different models in the Forecasts section.

4 Forecasts

The Multiple Linear Regression model can be represented as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

In this model, Y is the median house value, β_0 is the intercept, and $\beta_1, \ldots, \beta_n$ are the coefficients. ε is the error term.

$$
\begin{aligned}
\text{median\_house\_value} = & -3506059.6384 \\
& + 38348.8292 \cdot \text{median\_income} \\
& + 1125.2551 \cdot \text{house\_median\_age} \\
& - 1.7719 \cdot \text{total\_rooms} \\
& - 43.2018 \cdot \text{population} \\
& + 149.3997 \cdot \text{households} \\
& - 42073.6718 \cdot \text{longitude} \\
& - 42289.7110 \cdot \text{latitude} \qquad (1)
\end{aligned}
$$

This equation means that for each unit of change in the coefficient, it impacts the median house value by n amount provided no other covariates change. For example, a change in the unit of median income ($1000) would increase the median house value by $38348.829.

Table 8: Residuals

Statistic	**Min**	**1Q**	**Median**	**3Q**	**Max**
Residuals	-540492	-44503	-11764	30753	869685

Table 9: Linear Regression Model Coefficient Results

Variable	**Estimate**	**Std. Error**	**t value**	**Pr($>\|t\|$)**
(Intercept)	-3506059.638	62714.743	-55.90	< 0.0000000000000002 ***
median_income	38348.829	316.526	121.16	< 0.0000000000000002 ***
house_median_age	1125.255	43.406	25.92	< 0.0000000000000002 ***
total_rooms	-1.772	0.689	-2.57	0.01 *
population	-43.202	1.048	-41.24	< 0.0000000000000002 ***
households	149.400	4.343	34.40	< 0.0000000000000002 ***
longitude	-42073.672	716.169	-58.75	< 0.0000000000000002 ***
latitude	-42289.711	677.404	-62.43	< 0.0000000000000002 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 70000 on 20632 degrees of freedom
Multiple R-squared: 0.632, Adjusted R-squared: 0.632
F-statistic: 5.07×10^3 on 7 and 20632 DF, p-value: < 0.0000000000000002

Table 8 shows the five-number summary for the residuals of the linear regression model. A residual is the difference between observed value of the dependant variable and the value predicted by the model. It is calculated using this formula: $r_i = y_i - \hat{y}_i$, where r_i is the residual for an observation, y_i is the actual value (observed) and $\hat{y}_i$ is the predicted value.

Table 9 shows information about the coefficient of each variable in the linear model. As explained before, an estimate represents the change in the dependant variable (median house price) for each unit change of the coefficient (all covariates being constant). The standard error measures the variability of each coefficient (with smaller values being more accurate). The t-value is used to test the null hypothesis of whether a predictor is significant (a larger t-value indicates a smaller p-value), and a p-value is needed to test significance (with a smaller p-value suggesting an indicator is significant at predicting the dependant variable). An R-squared value is an indication of how well the model fits, with 1 being a perfect fit and 0 being no fit. A value of 0.632 indicates that the model is somewhat accurate at predicting the median house value. The tests on these variables to check whether they should have been included in the model (and whether the models is accurate at predicting the median house price) can be seen below:

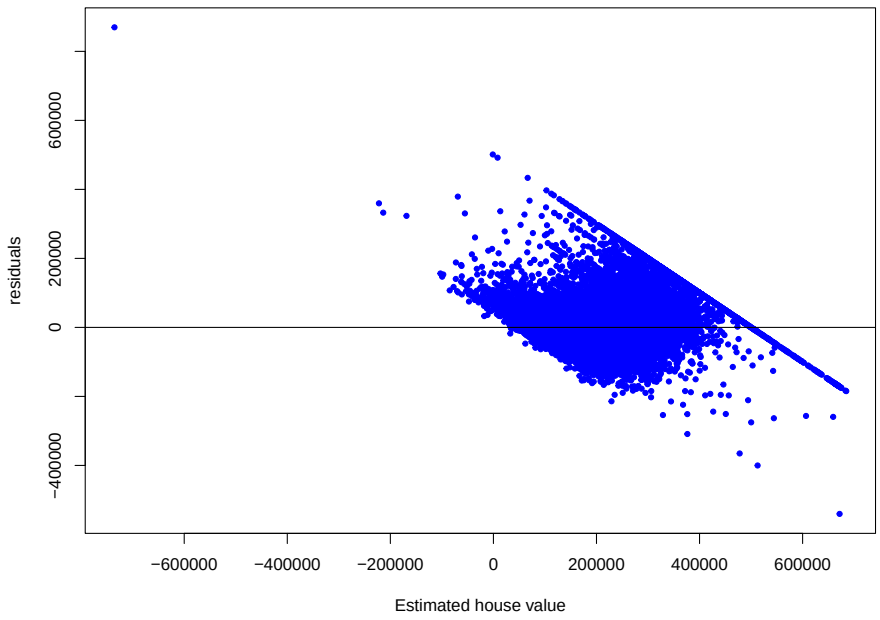

Figure 4: Residuals fitted against fitted median house values.

Figure 4 shows a residuals plot. A residuals plot should have random scatter, no clear patterns and show no trends. Figure 4 clearly shows that there is an issue with the linear regression model. It shows all the data clustered into the center right of the graph, with a straight line of points. To investigate why this must be the case, there needs to be a test of the individual predictors.

The following is a test of the median income predictor:

H_0: There is no linear effect between "median_house_value" and "median_income".
H_1: There is a linear effect between "median_house_value" and "median_income".

Test Statistic: The t-value under H_0 (as seen in Table 7) is 121.16 with degrees of freedom being 20632.

The p-value (0.0000000000000002) is much lower than any level of significance.

This would lead to the rejection of H_0. There is statistically significant evidence that there is in fact a linear relationship between the "median_house_value" and "median_income".

Doing the same hypothesis tests with the rest of the individual predictors leads to p-values lower than significance levels of 0.001, indicating that all the predictors have a linear relationship with the response variable (median house value). Figure 4 shows however that there must be something wrong with the model, so more investigation is needed. Below (Figure 5) shows the individual predictors plotted against the residuals.

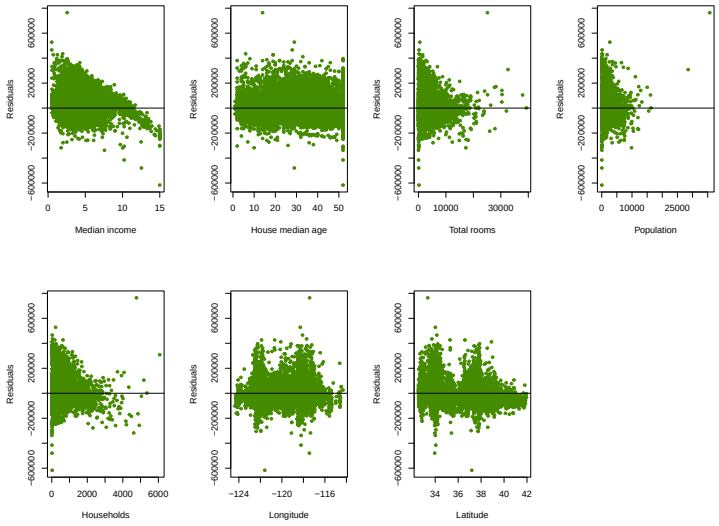

Figure 5: Individual coefficients vs residuals plot.

The most immediate issue that can be seen in Figure 5 is that there has been an inclusion of count data in the model. Count data should not be included

in a linear regression model. The count data includes the Total Rooms, the Population, and the Households. These coefficients should be removed.

The Longitude and Latitude vs residuals plots (in Figure 5) show similar patterns with two clusters in similar areas. Because there should be no clear patterns in a linear regression model (assumptions not met), these predictors are also going to be removed from the model. This leaves just the median income and the median house age as the predictors for the linear model.

Finally, the straight line of points shown in Figure 4 is going to be removed by taking a subset of the data-set that doesn't include any median house price results over $500,000. This is because there are many data-points with this value, which may cause issues in the linear regression model. The result of this linear model can be seen below (Table 9 and Table 10).

Table 10: Residuals

Statistic	**Min**	**1Q**	**Median**	**3Q**	**Max**
Residuals	-577384	-48861	-12265	36050	373830

Table 11: Linear Regression Model Coefficient Results

Variable	**Estimate**	**Std. Error**	**t value**	**Pr($>\|t\|$)**
(Intercept)	-7706.6	1925.6	-4.0	0.000063 ***
median_income	42395.5	331.6	127.8	< 0.0000000000000002 ***
house_median_age	1547.2	41.7	37.1	< 0.0000000000000002 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 71600 on 19645 degrees of freedom
Multiple R-squared: 0.456, Adjusted R-squared: 0.456
F-statistic: 8.25×10^3 on 2 and 19645 DF, p-value: < 0.0000000000000002

The most immediate observation to be made is that the R-squared value is less than what it was on the previous linear regression model (0.632 compared to 0.456). This indicates that this linear model is actually less accurate at predicting the median house value. The p-values on the other hand are similarly very small, which indicates that the independent variables do have an impact on the median house value.

To test the accuracy of the model, the residuals are again plotted against the median house value, as seen on Figure 6 (below):

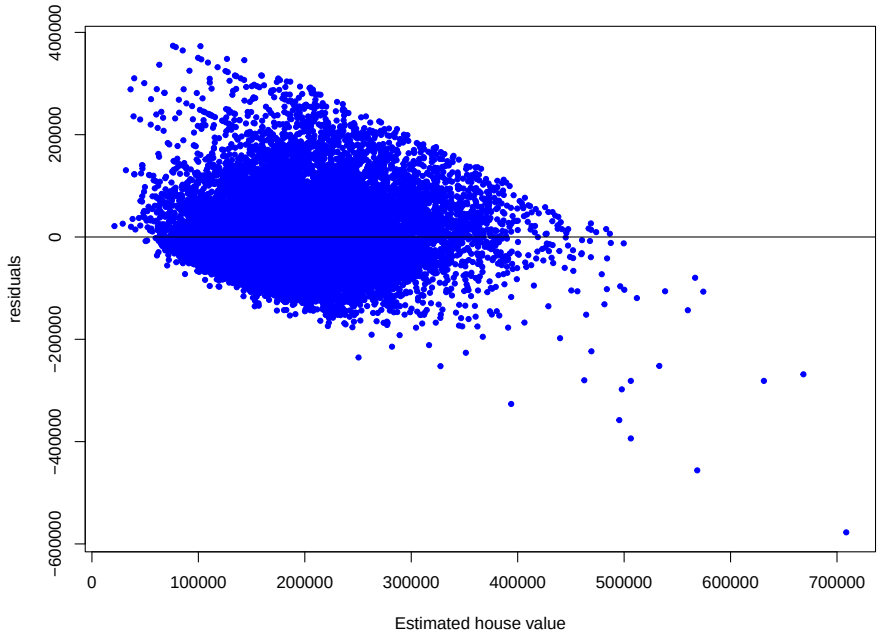

Figure 6: Residuals fitted against fitted median house values (altered data-set).

Unfortunately, it looks as though there is still linearity present in the residuals, which indicates the assumptions of the data have not been met. As mentioned before, a residuals plot should have random scatter, no clear patterns and show no trends (homoscedasticity). Once again the individual coefficients will be plotted against the residuals (Figure 7 below):

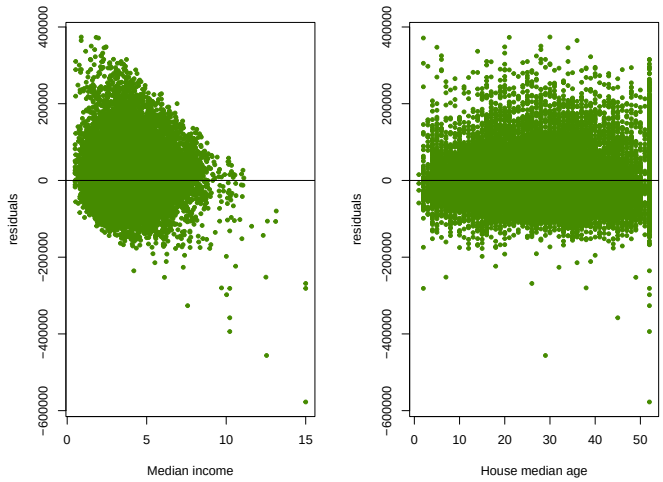

Figure 7: Individual coefficients vs residuals plot (altered data-set).

As shown in Figure 7, the median income vs residuals plot shows a cluster of data generally in the top left of the plot. This implies that there is a pattern in the data, and means the the underlying assumptions have not been met. In both tables there are several outlier values. These may be impacting the accuracy of the model, and have not been dealt with properly. Figure 8 (below) shows a QQ plot testing the distribution of the two variables:

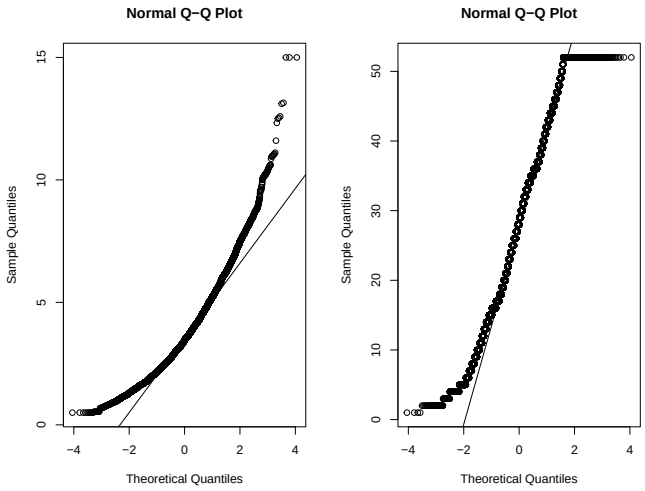

Figure 8: QQ plot for median income and median house age to check whether data is normally distributed (left is median income, right is median house age)

Figure 8 shows that neither the median income variable or the median house age variable follow a normal distribution, as they do not follow the plotted line (median income follows a curve shape, and median house age has a lot of points to the right of the line). Again, this shows the underlying assumptions about the model have not been met.

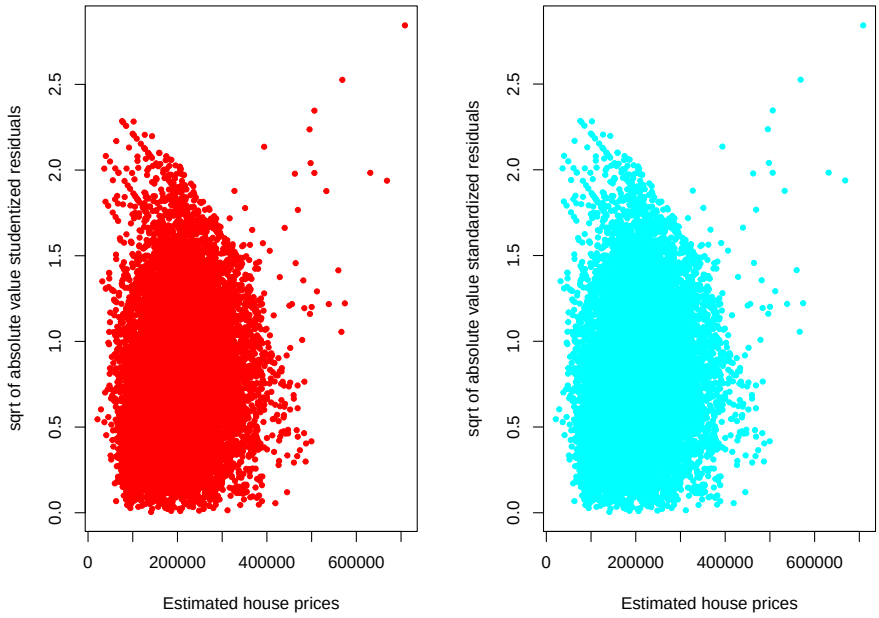

Figure 9: Studentized and Standardized Residuals plotted against Estimated house prices.

Figure 9 shows the studentized and standardized residuals plotted against the dependant variable. This is a way of detecting outliers. As shows in the two plots, there are many outliers (to the right of each plot). This indicates that there are a lot of observed median house prices that are a lot higher than what is being predicted by the model. Again, this shows that the linear model is not particularly good at predicting the median house price.

One way that might improve the model is to transform the dependant variable by square rooting it. This should hopefully get the issues relating to homoscedasticity in the previous residual plots (as in the residuals following a pattern). Table 12 gives an overview of the model after transforming the dependant variable:

Table 12: Residuals

Statistic	**Min**	**1Q**	**Median**	**3Q**	**Max**
Residuals	-644.5	-55.1	-8.1	47.0	377.7

Table 13: Linear Regression Model Coefficient Results

Variable	**Estimate**	**Std. Error**	**t value**	**Pr($>\|t\|$)**
(Intercept)	200.9423	2.1616	93.0	< 0.0000000000000002 ***
median_income	47.9743	0.3723	128.9	< 0.0000000000000002 ***
house_median_age	1.6588	0.0468	35.5	< 0.0000000000000002 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 80.4 on 19645 degrees of freedom
Multiple R-squared: 0.46, Adjusted R-squared: 0.46
F-statistic: 8.36×10^3 on 2 and 19645 DF, p-value: < 0.0000000000000002

Looking at the R-squared value, it very slightly improves the fit of the model (0.46 vs the previous 0.456), however this still isn't a good fit. Figure 10 (below) shows the residuals vs fits plot after the tranformation:

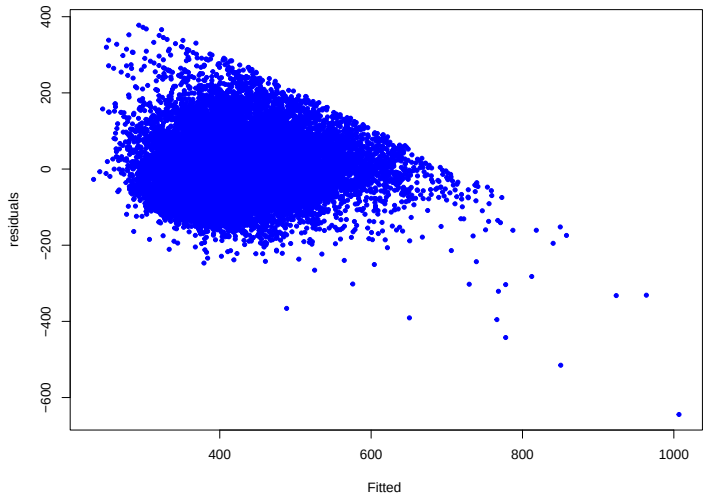

Figure 10: Fitted vs residuals plot after square root transformation of dependant variable.

Figure 10 shows a change in the scale when compared to the non transformed

plot (Figure 6), however it shows much of the same shape, which indicated the transformation has not been successful in improving the linear regression model.

5 Conclusion

In conclusion, it seems that a linear regression model is not the best fit for predicting the median house prices in California. There are several limitations with the data which make the model inaccurate - the main one being that the variables don't seem to follow a normal distribution. The presence of outliers in that dataset (which were perhaps not dealt with properly) may have also contributed to the accuracy of the model.

If we were to base some results off of the model (using the most recent transformed model), we could say that for each increase in unit ($1000) of median income, there is a $47,974.30 increase in the price of the median house value, and that for each increase in unit of median house age (so a year), it increases the price of the median house value by $1,658. Having an incredibly small p-value of 0.0000000000000002 for both independent variables indicates very strongly that they both have an effect on the dependent variable, however again due to the underlying assumptions about the data not being met, this is very likely inaccurate.

Other areas, such as location, were also not taken into consideration when creating the linear regression model, which would likely have an impact on price given that California is a coastal state. To improve the model, this should be taken into consideration.