# FIND CHEAP HOMES IN CALIFORNIA!!

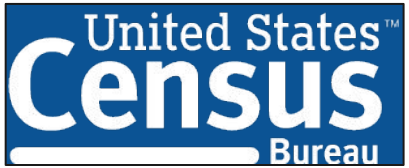
STATS 112
- Dara Hashemi (**705097381**)
- Joshua Park (505091304)
- Oscar Monroy (**305360444**)
- William Foote (**305134696**)

# The Research Question

Since 2017, what has affected Real Estate prices in California?

# The Dataset

- Data is collected from various publicly available 2017-2019 data sets separated by California counties.
  - **Voter Registration**
    - California Secretary of State
  - **Median House Price**
    - California Association of Realtors
  - **Demographic Data/GDP**
    - US Census Bureau


# The Multiple Linear Regression Model

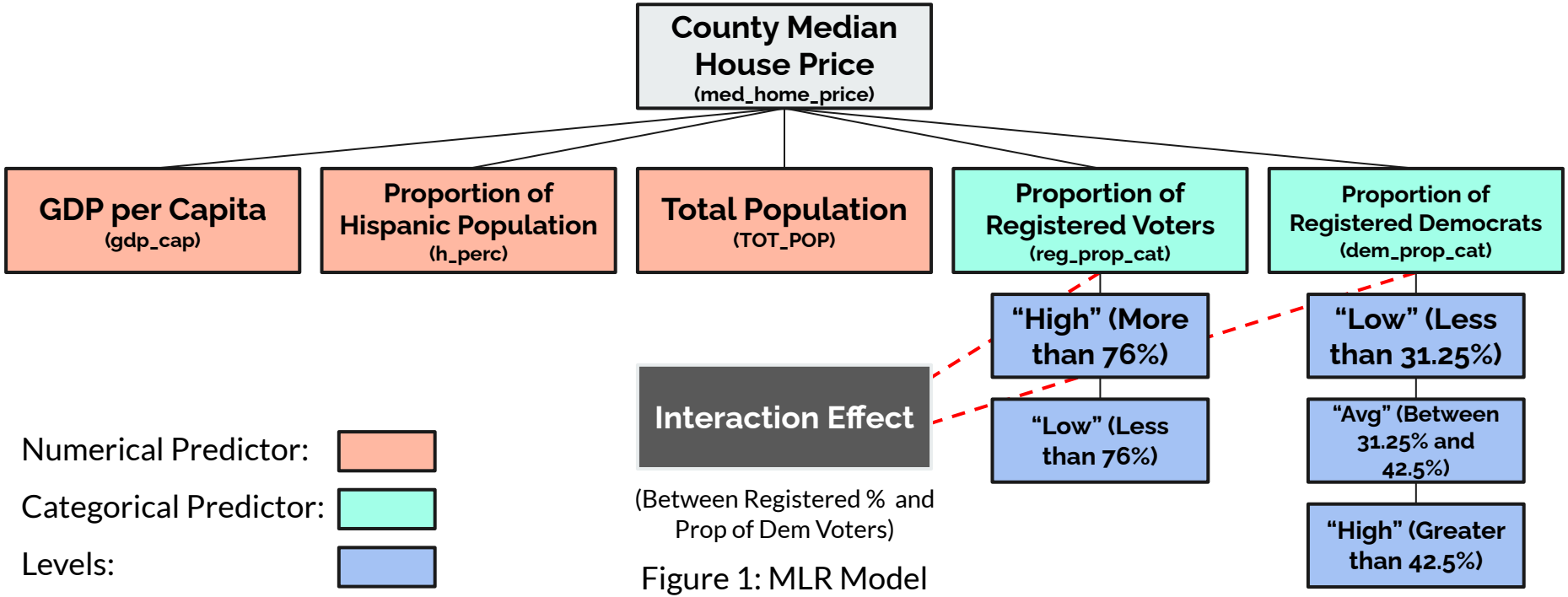
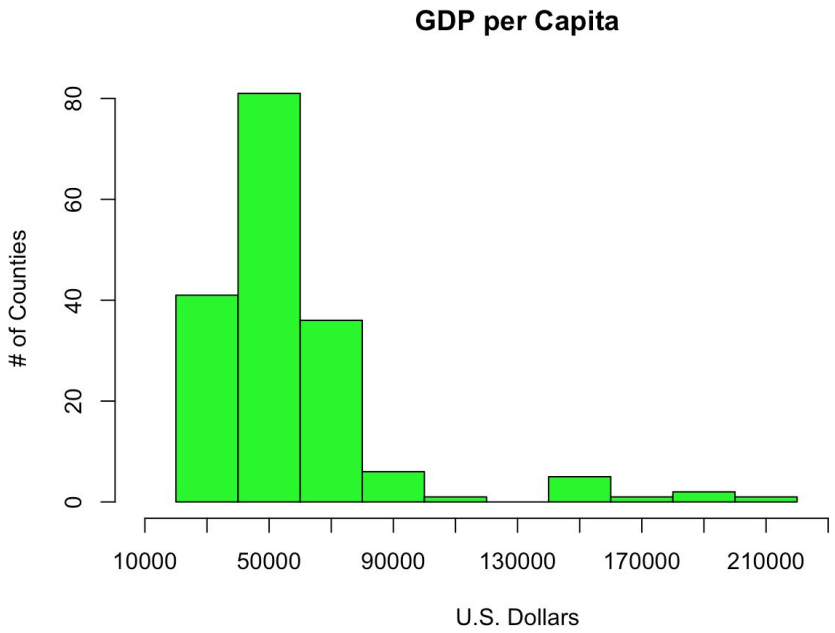
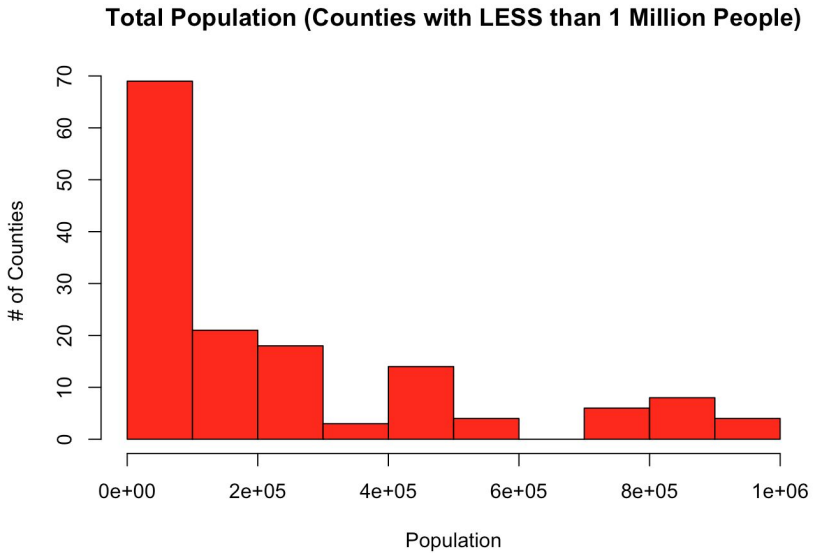


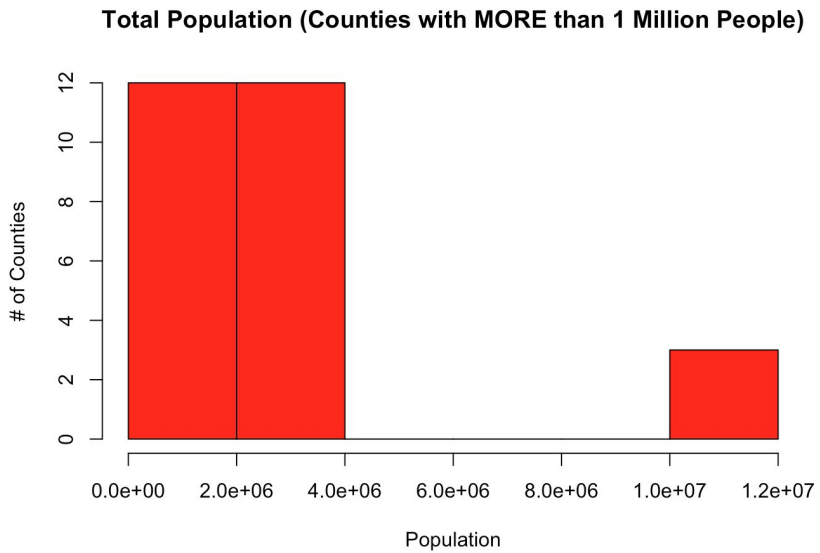

Figure 1: MLR Model

# Exploratory Analysis - Numerical Variables

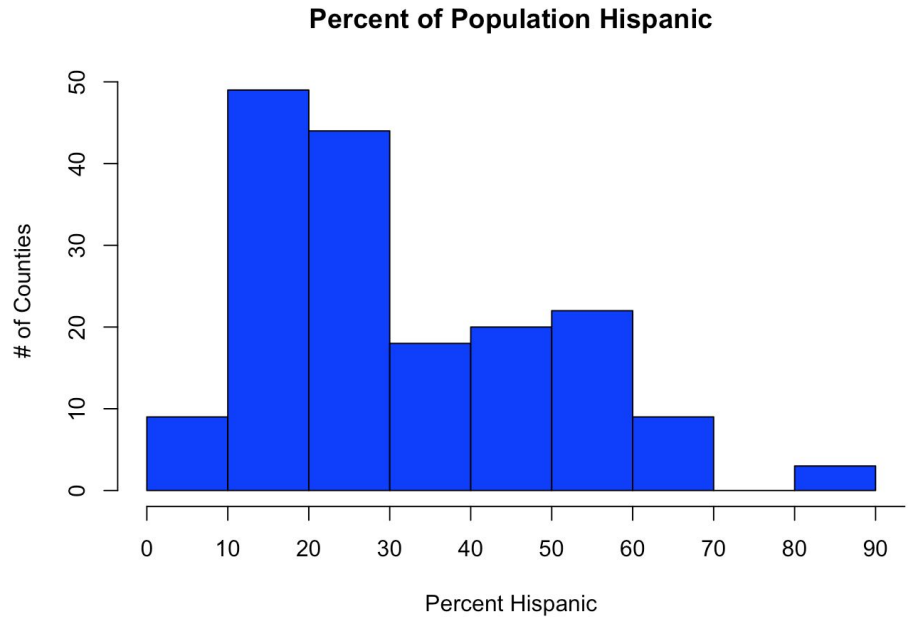


# Exploratory Analysis - Numerical Variables

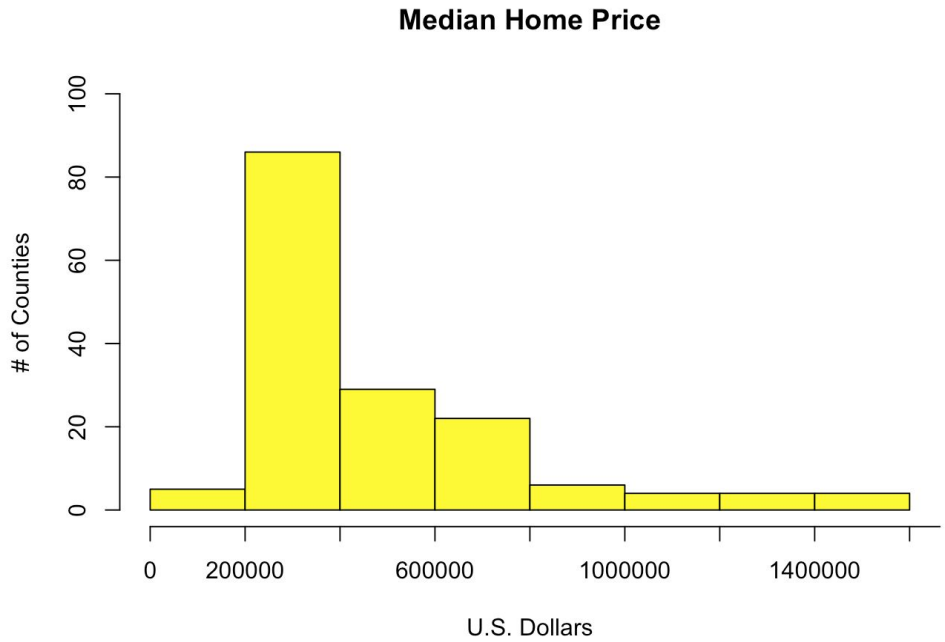


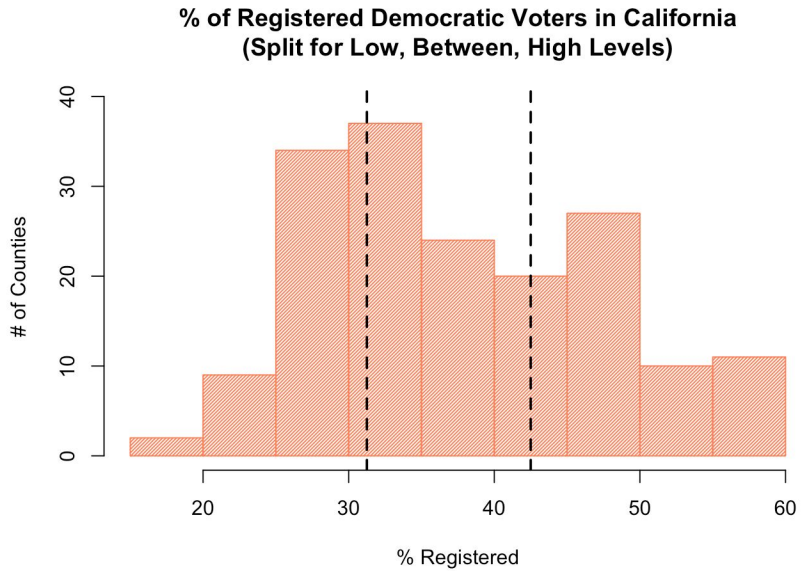
# Exploratory Analysis - Numerical Variables



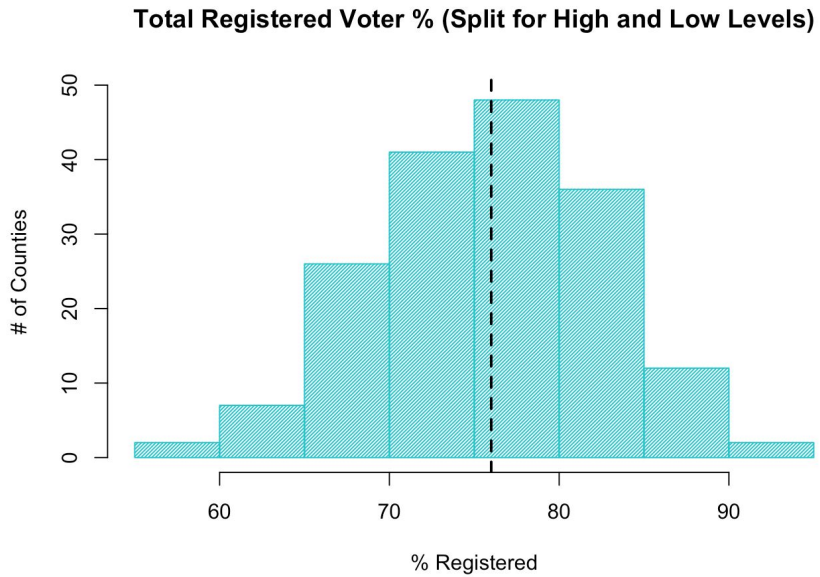
# Exploratory Analysis - Numerical Variables



# Exploratory Analysis - Categorical Variables



# Exploratory Analysis - Categorical Variables



# Exploratory Analysis - Categorical Variables (cont.)

## Factor + Contingency Tables

**Proportion of Total Registered Voters (# of Counties)**

| Low | High |
|---|---|
| 87 | 87 |

**Proportion of Registered Democrat Voters (# of Counties)**

| Low | Average | High |
|---|---|---|
| 56 | 60 | 58 |

# Exploratory Analysis - Categorical Variables (cont.)

## Factor + Contingency Tables

**Proportion of Registered Democrat Voters vs. Proportion of Total Registered Voters (# of Counties)**

| | | Prop. Total Registered Voters | |
|---|---|---|---|
| | | Low | High |
| Prop. Registered Democrat Voters | Low | 30 | 26 |
| | Average | 20 | 40 |
| | High | 37 | 21 |

# Exploratory Analysis - Scatterplot Matrix

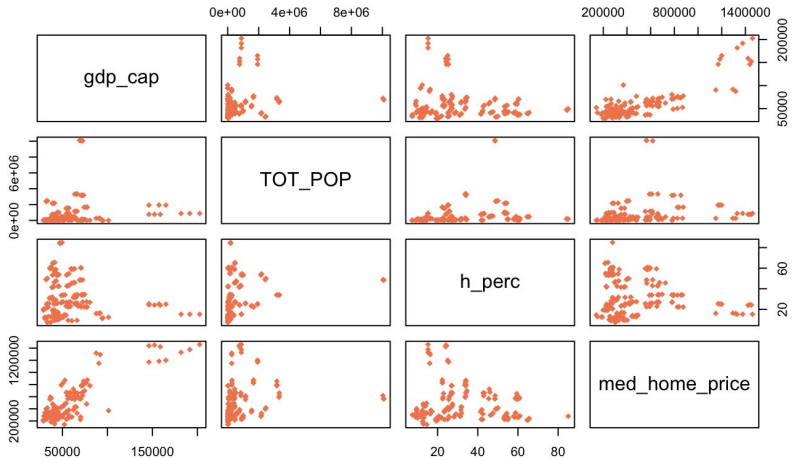



Figure 7: Scatterplot Matrix of 4 Numerical Variables

# Exploratory Analysis - Correlation Matrix

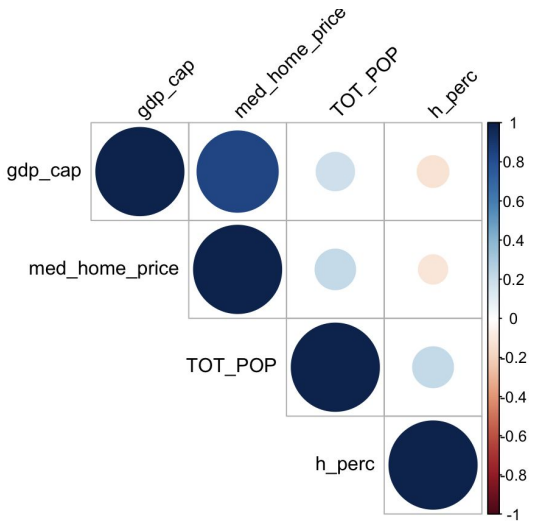



Figure 8: Correlation Matrix of Four Numerical Variables

# Final Linear Regression Model

| | Coefficients | Standard Error | T-Stat | P-value | |
|---|---|---|---|---|---|
| (Intercept) | 86,600 | 48,670 | 1.779 | 0.0772 | . |
| Total Population | -0.002485 | 0.007677 | -0.324 | 0.7466 | |
| Hispanic Percentage | -1,183 | 790.7 | -1.496 | 0.1368 | |
| GDP per Capita | 7.017 | 0.4167 | 16.841 | 0.0000 | *** |
| as.factor(Prop. Dem. Voters)high | 28,200 | 40,880 | 0.690 | 0.4913 | |
| as.factor(Prop. Dem. Voters)low | -105,400 | 39,840 | -2.646 | 0.0090 | ** |
| as.factor(Prop. Tot. Voters)low | 69,340 | 40,420 | 1.715 | 0.0884 | . |
| as.factor(Prop. Dem. Voters)high : as.factor(Prop. Tot. Voters)low | 91,390 | 56,160 | 1.627 | 0.1057 | |
| as.factor(Prop. Dem. Voters)low : as.factor(Prop. Tot. Voters)low | -11,600 | 55,340 | -0.210 | 0.8343 | |

## Interpreting Findings

**R^2:** 80.1% (Adjusted R^2 is 79.63%)

**Regression Statistics**

| | |
|---|---|
| Multiple R Squared | 0.8065 |
| Adjusted R Squared | 0.7963 |
| Standard Error | 134,900 |
| Observations (Degrees Freedom) | 151 |

# The MLR Model - Formula

**MedianHomePrice** = 86600 - 0.002485**TotalPopulation**

-1183**HispanicPercent** + 7.017**(GDP per Capita)**

+ 3182 **{DemHigh}**

-105400**{DemLow}**

– 240693 **{Dem<30}**

– 69340**{RegLow}**

+ 91390**{DemHigh * RegLow}**

- 11600 **{DemLow * RegLow}**

# Coefficients Interpretation (cont.)

**Intercept: 86,660**

If total population, the proportion of Hispanics, and GDP per Capita, of a given county were all zero, its proportion of Democrats of the registered voting population was between 31.25% and 42.5%, and its voter registration rate was greater than 76%, we would expect the median home price to be $86,660.

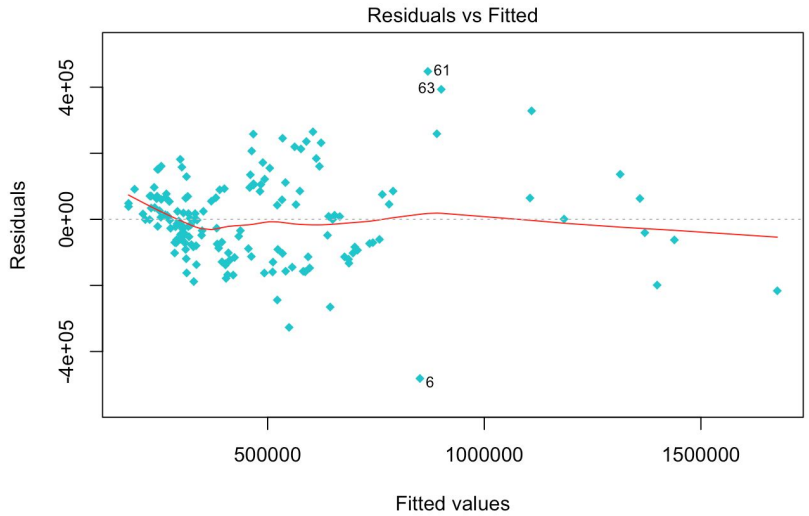
**GDP per Capita:  7.017**

For every one dollar increase in GDP per capita, we expect the median house price to increase by $7, holding all other variables constant.

**DemLow: -105,400**

Holding all other variables constant, the average median house price for counties where the percentage of Democrats of the registered voting population is less than 31.25% is $105,400 less than counties where the percentage of Democrats of the registered voting population is between 31.25% and 42.5%.

# Testing Assumptions - Linearity

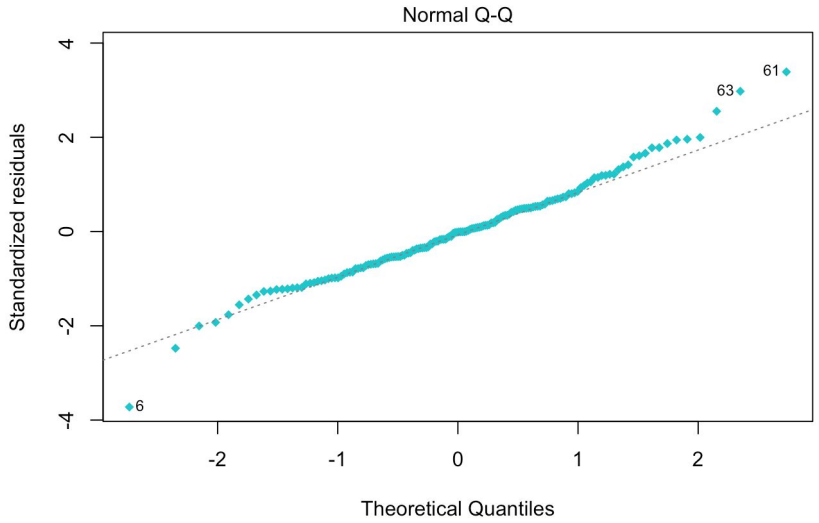



Checking for Equality of the Variance of Residuals

## How does model validity hold up?

The assumption seems to be met as there is no major pattern to the data and the points seem evenly spread around the mean of residuals (0).

# Testing Assumptions - Normal Q-Q

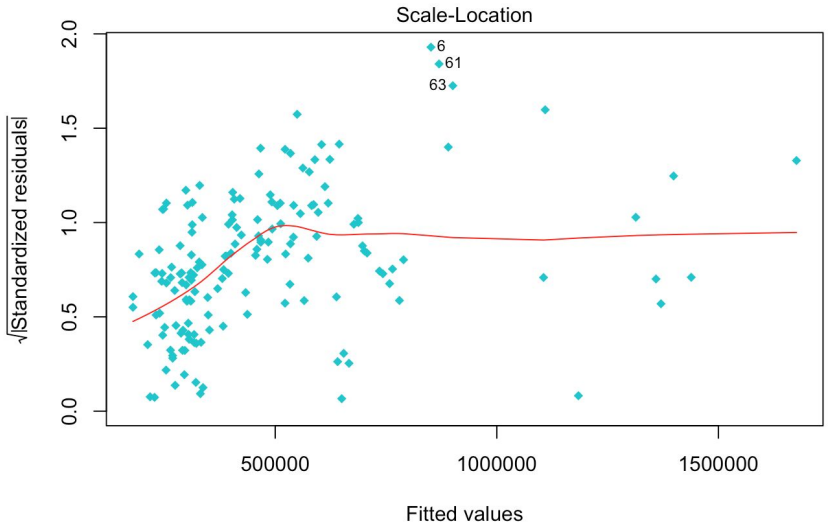



Checking for Normality

## How does model validity hold up?

Most of the points are on the line and show that we are not violating the assumption of normality of the outcome variable.

# Testing Assumptions - Standardized Residuals vs. Y^

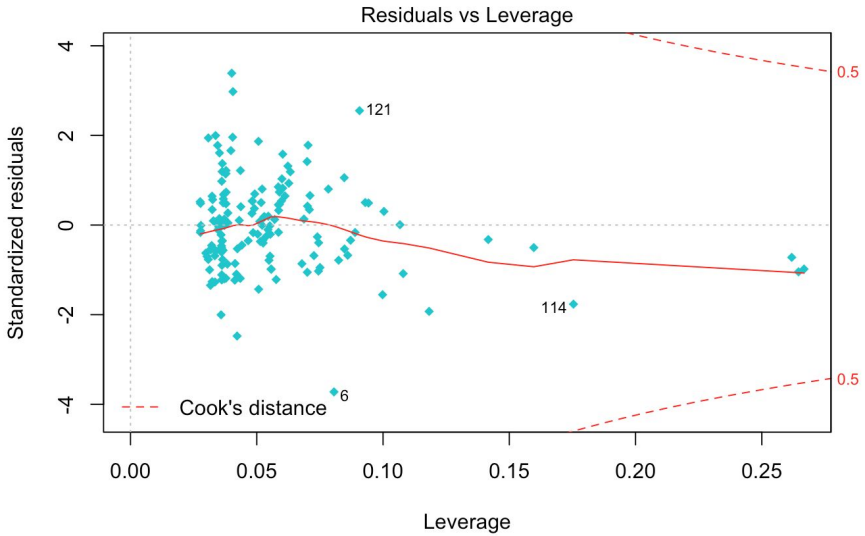



Checking for Constant Variance

## How does model validity hold up?

This assumption does not seem to be as good as the other, as the mean of the plot is not centered around zero and there is a fan shape following the points. As Y^ increases, variance of the residuals are not constant for the first section of values, but then eventually become constant.

# Testing Assumptions - Outliers




Checking for Outliers

## How does model validity hold up?

To find outliers, we must look for points above 3 and below -3 as a rule of thumb. We see from this plot there is only 1 outlier under -3, which will not have much effect on our coefficients due to the number of observations in our model.

# Variance Inflation Factor

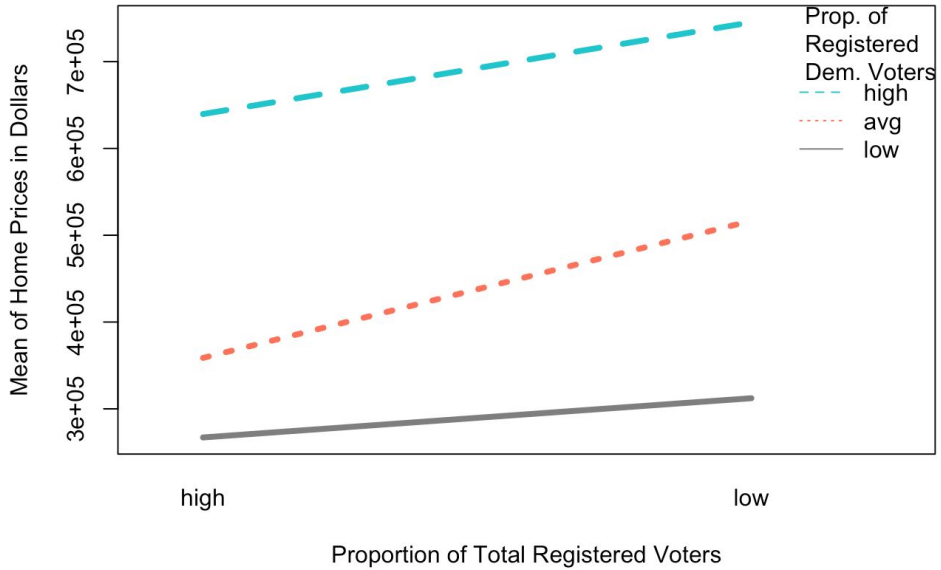
Look at Final Column, and apply the general “rule of thumb” for regular VIF. Source.

**Variance Inflation Factor (VIF)**

| | GVIF | Df | $GVIF^{(1 / (2*Df))}$ |
|---|---|---|---|
| Total Population | 1.176906 | 1 | 1.084853 |
| Hispanic Percentage | 1.580369 | 1 | 1.257127 |
| GDP per Capita | 1.434112 | 1 | 1.197544 |
| as.factor(Prop. Dem. Voters) | 6.585142 | 2 | 1.601922 |
| as.factor(Prop. Tot. Voters) | 3.587472 | 1 | 1.894062 |
| as.factor(Prop. Dem. Voters) : as.factor(Prop. Tot. Voters) | 13.339302 | 2 | 1.911099 |

From our analysis, we can conclude that our predictor variables are not highly correlated with each other, due to our VIF’s being below 5 for each predictor.

# Interaction Effect




**Interpretation:**

The Average and Low levels of the Proportion of Registered Democrat Voters seem to have a very minimal interaction effect with the Proportion of Total Registered Voters.

However, these lines for the most part are **parallel**, indicating the **absence of an interaction effect** and that the effect of the proportion of Registered Democrat Voters does not depend on the proportion of Total Registered Voters.

# Overall Conclusions

- Counties' GDP per Capita and the percentage of registered Democrat voters being less than 31.25% are **highly correlated** to Housing Prices in counties across California.
- There was **no correlation** to more diverse counties (Hispanic specifically) having an effect on Housing Prices.
- In more Democratic populated areas, housing prices are higher compared to less Democratic populated areas.
- Counties with lower proportions of registered voters tend to have higher housing prices.


**Housing prices in California from 2017-2019 can be explained by our model with roughly 80% strength ($R^2$).**

# Appendix: Non-Significant Coefficient Interpretations

**Total Population: -0.002485**
For every one unit increase in Total Population, we expect the median house price to decrease by 0.002485, holding all other variables constant.

**Hispanic Percentage: -1,183**
For every one unit increase in Proportion of Hispanic Population, we expect the median house price to decrease by $1,183, holding all other variables constant.

**DemHigh: 28,200**
Holding all other variables constant, the average median house price for counties where the percentage of Democrats of the registered voting population is greater than 42.5% is $28,200 more than counties where the percentage of Democrats of the registered voting population is between 31.25% and 42.5%.

**TotalLow: 69,340**
Holding all other variables constant, the average median house price for counties where the percentage of the total registered voting population is less than 76% is $69,340 more than counties where the percentage of the total registered voting population is greater than 76%.

**Interactions**

**DemHigh : TotalLow: 91,390**
Holding all other variables constant, the average median house price for counties where the percentage of Democrats of the registered voting population is greater than 42.5% and the total registered voting population is less than 76% is $91,390 more than counties where the percentage of Democrats of the registered voting population is between 31.25% and 42.5% and total registered voting population is above 76%.

**DemLow : TotalLow: -11,600**
Holding all other variables constant, the average median house price for counties where the percentage of Democrats of the registered voting population is less than 31.25% and the total registered voting population is less than 76% is $11,600 less than counties where the percentage of Democrats of the registered voting population is between 31.25% and 42.5% and total registered voting population is above 76%.

# Thank You!