Housing Availability in Denver Neighborhoods

Adelle Price

University of Colorado Denver

MATH 5387

Introduction

This study seeks to explore the impact of various neighborhood demographics on the number of vacant living spaces in Denver, Colorado. In an essay regarding "Unequal Growth in Housing"- published by the US Department of Housing and Urban Development- Denver is noted to be a High-Opportunity City with high median income and high job growth. Impeding Denver's development, however, is its low numbers of "single family permitting housing".[1] According to a Denver Post article, the number of new home listing in Denver decreased 11.5% from 2020 to 2021.[2] Additionally, 9 News reports median home buying costs in Denver were seen to rise 12.6% from 2020 to 2021.[3] Being a High-Opportunity City, it is vital public policy officials explore possible causal factors for limited available housing within Denver. To aide in doing so, this study used data from the American Community Survey; provided to the public by the Denver Open Data Catalog. [4] The information described in the data was initially collected at the census level tract and then summarized into neighborhoods. Generally speaking, this data reflects a 5 year average of an observational study conducted from 2013 to 2018. The data describes 78 neighborhoods within the City and County of Denver and contains 146 observations per neighborhood. 7 of these observations, including the number of available housing units, were analyzed in this study.

Methods of Multiple Linear Regression were used to evaluate the relationships between the number of available housing units and the remaining 6 variables of percent of two or more races, median age per neighborhood, total housing units, number of family households, median gross rent, median home value, and percent of families in poverty. In order to use this method, any neighborhoods with missing data points were excluded; 76 neighborhoods remained and were subsequently analyzed. The resulting model defining causal variables(the regressors) and the number of available housing units (the response) was evaluated for its features of predictive accuracy, model structure, and error assumptions. The study is concluded with model interpretations and recommendations for potential policy changes in Denver, Colorado to address its increasing lack of available housing.

Results

Data Exploration

Before beginning model selection, the distribution of each variable was investigated. In figures 1-7, it is shown that the variables *Total Housing Units, Median Home Value, Number of Family Households, Percent of Two or More Races,* and *Vacant Housing* are right skewed and thus a square root transformation of each is necessary. Numerical summaries of all variables after the transformations is provided in Table 1.

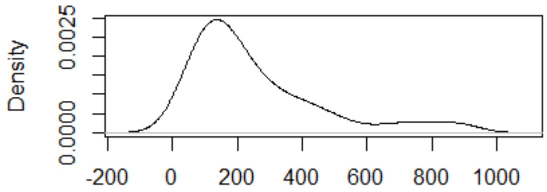


Figure 1. Density plot of *Vacant Housing*.

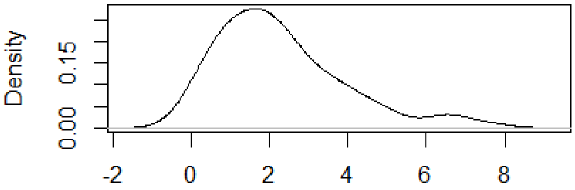


Figure 2. Density plot of *Percent of Two or More Races*.

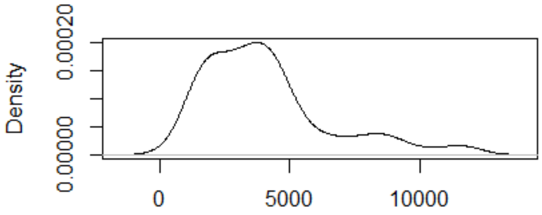


Figure 3. Density plot of *Total Housing Units*

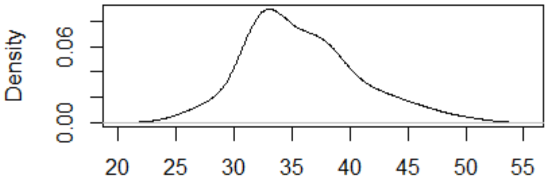


Figure 4. Density plot of *Median Age*.

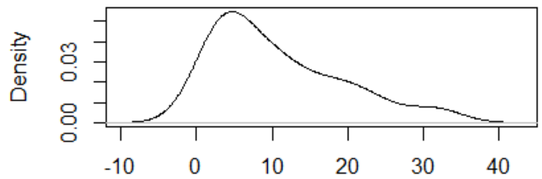


Figure 5. Density plot of *Median Gross Rent.*

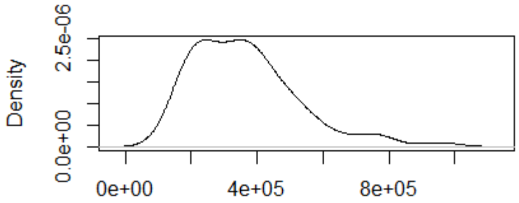


Figure 6. Density plot of *Median Home Value.*

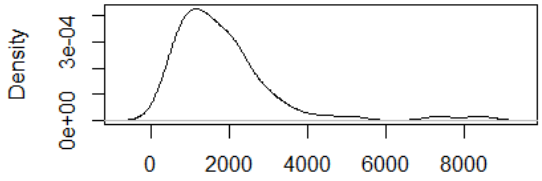


Figure 7. Density plot of *Number of Family Households*.

	Min.	Median	Max.
sqrt(Vacant Housing Units)	**4.899**	**13.528**	**30.166**
sqrt(Percent of Individuals of Two or more Races)	**0.000**	**1.924**	**7.563**
Median Age	**25.30**	**35.10**	**50.10**
sqrt(Total Housing Units)	**24.98**	**60.11**	**108.23**
Number of Family Households	**308.0**	**1511.0**	**8403.0**
Median Gross Rent	**772**	**1160**	**2143**
sqrt(Median Home Value)	**378.9**	**590.9**	**978.5**
sqrt(Percent of Families in Poverty)	**0.00**	**2.91**	**5.76**

Table 1. Numerical summaries of each observation investigated after necessary transformations were performed.

After a square root transformation of *Total Housing Units, Median Home Value, Number of Family Households, Percent of Two or More Races,* and *Vacant Housing*, a scatterplot was observed to evaluate pairwise correlation. In Figure 8, we see notable positive correlation between *Number of Family Households* and *Total Housing Units*, *Vacant Housing* and *sqrt(Total Housing Units), Median Gross Rent* and *sqrt(Median Home Value)*. We also see notable negative correlation between *sqrt(Percent of Family Poverty)* and *sqrt(Median Home Value)* as well as *sqrt(Percent of Family Poverty)* and *Median Age*. We did not adjust our model according to Figure 8 alone- however the highly correlated variables were considered for elimination from the model during investigation of collinearity.

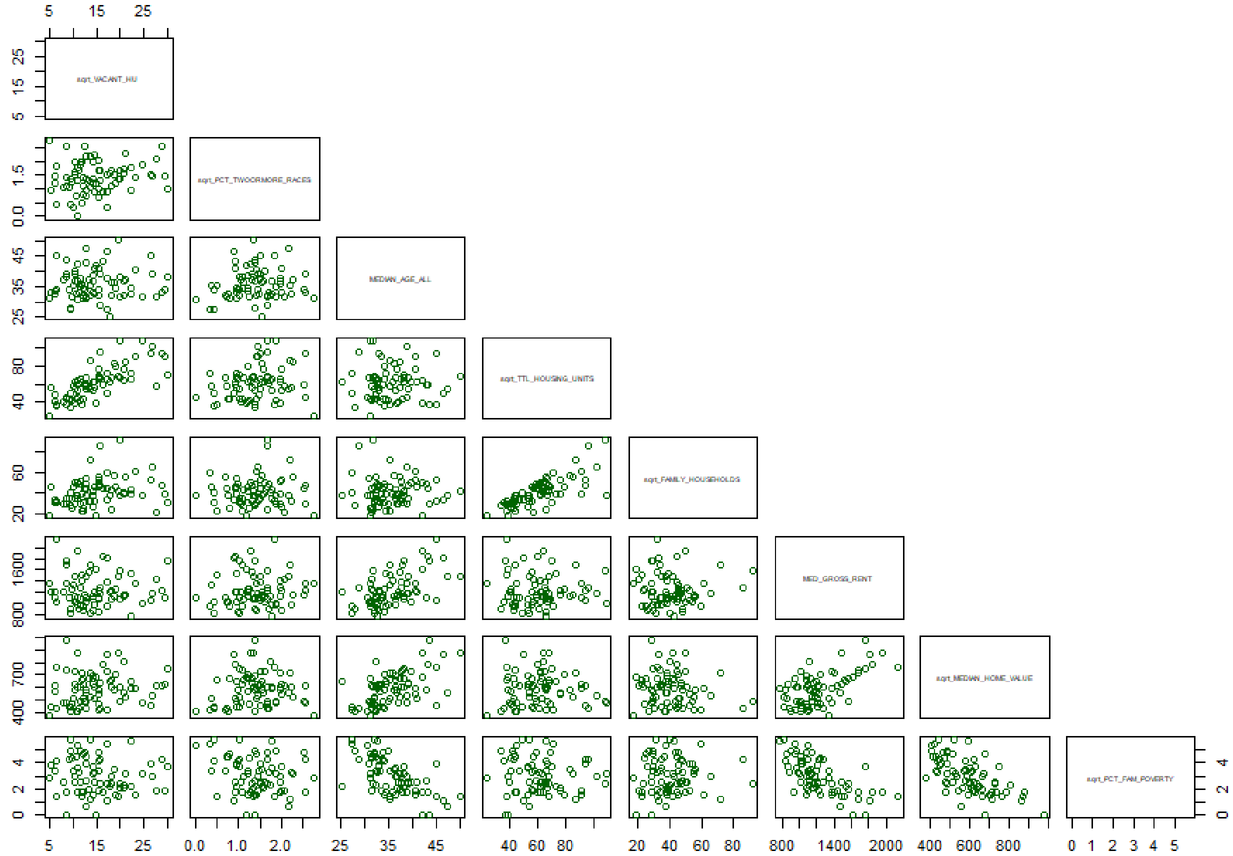


Figure 8. Scatterplot of pairwise correlation between all observed variables.

Variable Selection

Beginning with the initial model of $sqrt(Vacant\ Housing)\sim\ sqrt(Percent\ of\ Two\ or\ More\ Races) + Median\ Age + sqrt(Total\ Housing\ Units) + sqrt(Family\ Households) + Median\ Gross\ Rent + sqrt(Median\ Home\ Value) + sqrt(Percent\ Family\ Poverty)$, regressors were eliminated from the model based on their Variance Inflation Factors(VIFs) and Condition Indexes which evaluated variable collinearity. The three regressors remaining after removing collinear variables were $sqrt(Total\ Housing\ Units)$, $Median\ Gross\ Rent$, and $sqrt(Median\ Home\ Value)$.

The performance of models with one, two, or three regressors was then evaluated. The models' performance was compared using Akaike's Information Criterion (AIC) and the adjusted R^2 of each respective model. Figure 9 shows models with two and three regressors have competitively low AIC values, indicating they have a similar balance between model fit and model complexity. Figure 10 shows the adjusted R^2 value for the model using all three regressors is the greatest. The model with three regressors was then selected for further evaluation.

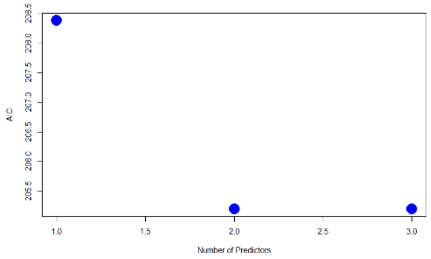


Figure 9. Plot of the number of model predictors versus the associated AIC.

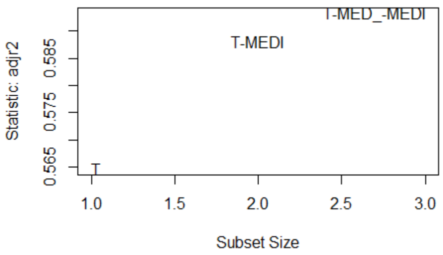


Figure 10. A plot of the number of model predictors versus the respective adjusted R^2 value where *sqrt(Total Housing Units)= T, Median Gross Rent=MED_,* & *sqrt(Median Home Value)=MEDI.*

Model Evaluation

With the model of $sqrt(Vacant\ Housing) \sim sqrt(Total\ Housing\ Units) + Median\ Gross\ Rent + sqrt(Median\ Home\ Value)$ selected, the following attributes were investigated:

- if the adjusted R^2 value used to select the three regressors was an appropriate measure of goodness of fit,
- the structural integrity of the model (residual plots, marginal model plots, and added variable plots), and
- potential influential observations in the chosen model.

In Figure 11, we see that y and yhat are positively correlated and the fitted line is straight. This indicates that the adjusted R^2 value used to measure the goodness of fit of the model with three regressors is indeed an appropriate measurement. Figure 12- plots of the model residuals versus the regressors and fitted values- confirms that the condition of residual mean equal to zero is met in each situation. Additionally, there is no evidence of systematic curvature in any relationship and the variance is reasonably constant in each plot.

The marginal model plots in Figure 13 indicate that there are no concerns in our model's predictive accuracy as the fitted line for model predictions closely follows the fitted lines for the observed data. When examining the added variable plots in Figure 14, we do not see any notable non-linear relationships between the regressors and response after accounting for the effect of the other regressors in the model.

Though we did not identify structural problems in Figures 11-14, we did identify model outliers. In each figure, there are several residuals that have a noticeably larger variance when compared to others. We further investigated the influence of these points on our model.

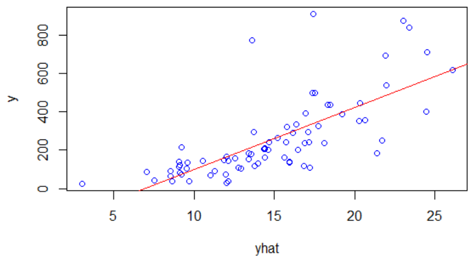


Figure 11. A plot of the estimated values the response (yhat) versus the observed values of the response (y).

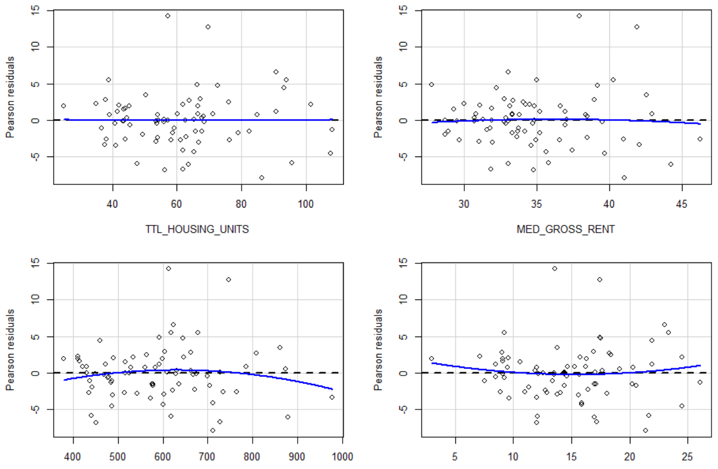


Figure 12. Residual plots for $sqrt(Vacant\ Housing) \sim sqrt(Total\ Housing\ Units) + Median\ Gross\ Rent + sqrt(Median\ Home\ Value)$.

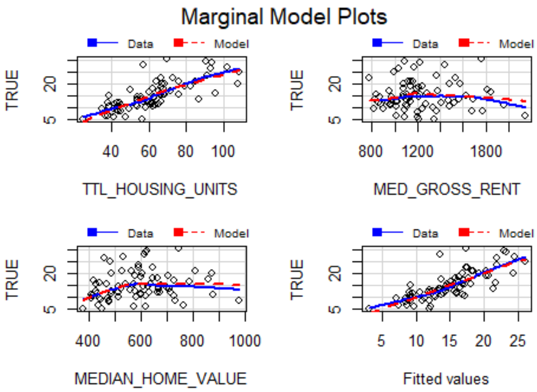


Figure 13. Marginal model plots for $sqrt(Vacant\ Housing) \sim sqrt(Total\ Housing\ Units) + Median\ Gross\ Rent + sqrt(Median\ Home\ Value)$.

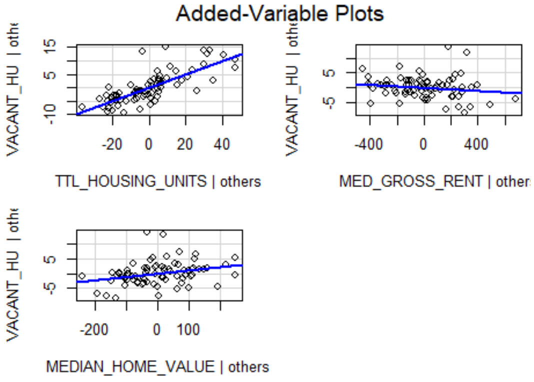


Figure 14. Added variable plots for $sqrt(Vacant\ Housing) \sim sqrt(Total\ Housing\ Units) + Median\ Gross\ Rent + sqrt(Median\ Home\ Value)$.

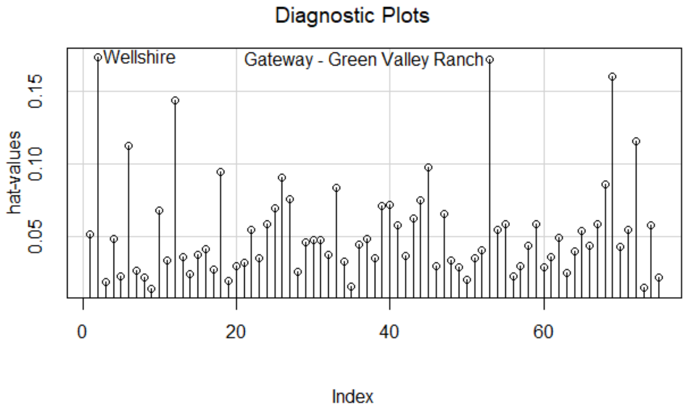


Figure 15. An index plot of leverage values for the selected model.

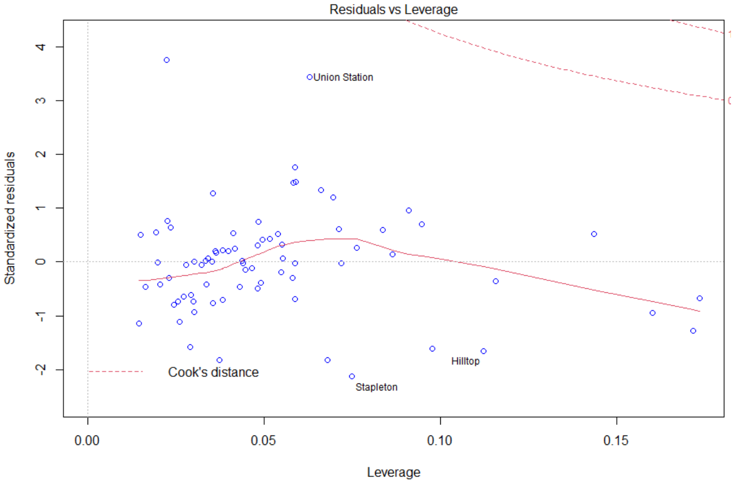


Figure 16. Model leverage points versus model residual values with corresponding Cook's distances.

Figure 15 identifies the model leverage points of "Wellshire" and "Gateway-Green Valley Ranch". In Figure 16, however, the most influential leverage point in the model is "Union Station" and has a Cook's distance value less than .5 (a common threshold for influential leverage points). Therefore we concluded this leverage point does not have a significant influence on our model performance.

In our chosen model, we also investigated whether the assumption that the model errors are normally distributed was met. In Figure 17, we see that the model residuals are not normally distributed, but instead right skewed as there are two observations lying outside of the regions for q-q plot normalcy. In order to address the lack of normally distributed errors, we would consider performing robust linear regression in a complementary project. In this study, however, we relied on the Central Limit Theorem which indicates if we were to increase our sample size, our findings would become increasingly accurate.

We investigated for correlated errors in Figure 18. We see that there is no discernable serial correlation between any two successive residuals as the points in Figure 18 are randomly scattered above and below $y = 0$. The lack of correlation between errors was confirmed using a Durbin-Watson test which yielded a p-value of 0.2737- indicating the errors of our chosen model are not correlated.

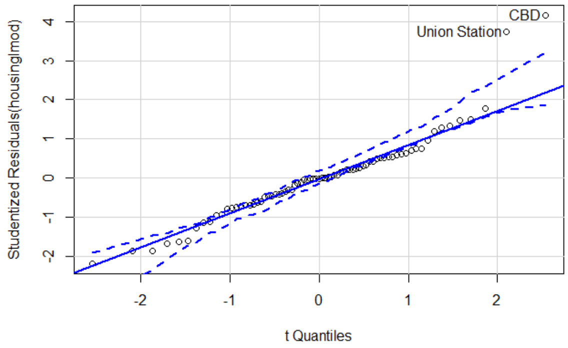


Figure 17. A quantile-quantile plot of the standardized model residuals.

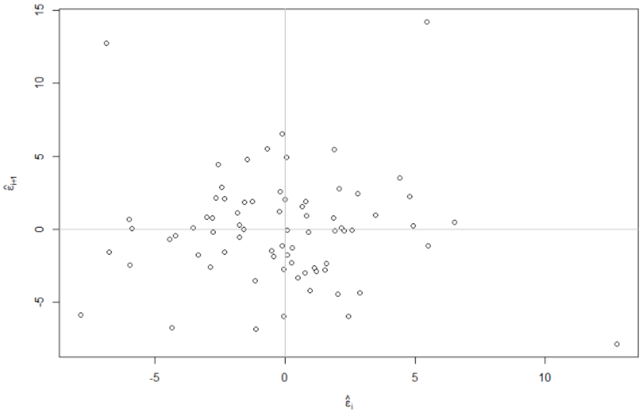
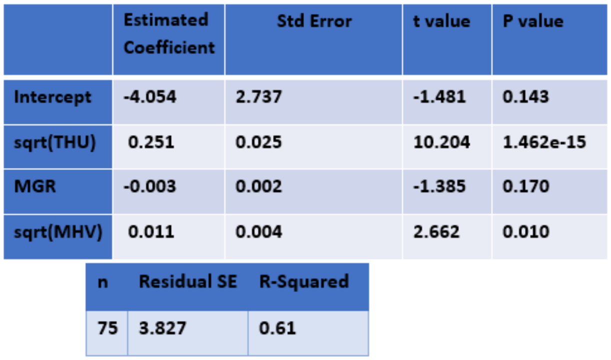


Figure 18. A plot of successive pairs of residuals: $\hat{\varepsilon}_i$ versus $\hat{\varepsilon}_{i+1}$.

Model Summary

Table 2 shows the estimated regression coefficient of the $sqrt(Total\ Housing\ Units)$ as 0.251. This indicates that a single unit increase in $Total\ Housing\ Units$ corresponds with a .251 unit increase in $Vacant\ Housing$ units. Likewise, the estimated regression coefficient of $sqrt(Median\ Home\ Value)$ corresponds to a 0.011 unit increase in $Vacant\ Housing$ per a single unit increase in $Median\ Home\ Value$. The p-values associated with both of the aforementioned estimated regression coefficients are below 0.05, and therefore we conclude they are statistically significant estimates. Conversely, the estimated regression coefficient of $Median\ Gross\ Rent$ that corresponds to a 2*(0.003) = 0.006 unit decrease in $Vacant\ Housing$ per a single unit increase in $Median\ Gross\ Rent$ has a p-value greater than 0.05. Thus, we conclude this estimate is not statistically significant and will not be used for inference.

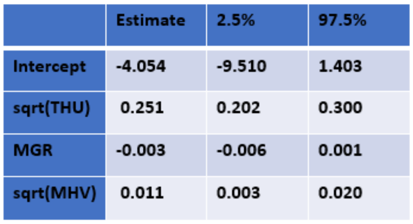
As a measure of the goodness of fit of our selected model, the R^2 value of 0.61 is regarded as reasonably strong. The residual standard error for this model is shown to be 3.827. This indicates our model will predict the actual number of available housing units with an approximately 4 unit error on average.



	Estimated Coefficient	Std Error	t value	P value
Intercept	-4.054	2.737	-1.481	0.143
sqrt(THU)	0.251	0.025	10.204	1.462e-15
MGR	-0.003	0.002	-1.385	0.170
sqrt(MHV)	0.011	0.004	2.662	0.010

n	Residual SE	R-Squared
75	3.827	0.61

Table 2. Summary of $sqrt(Vacant\ Housing) \sim sqrt(Total\ Housing\ Units) + Median\ Gross\ Rent + sqrt(Median\ Home\ Value)$



	Estimate	2.5%	97.5%
Intercept	-4.054	-9.510	1.403
sqrt(THU)	0.251	0.202	0.300
MGR	-0.003	-0.006	0.001
sqrt(MHV)	0.011	0.003	0.020

Table 3. 95% Confidence Intervals for the estimated regression coefficients.

In Table 3, the 95% Confidence Intervals for the estimated regression coefficients describe the intervals in which the actual regression coefficent value will lie in, 95% of the time. We observe that the true coefficieint of $sqrt(Total\ Housing\ Units)$ will lie between 0.201 and 0.301, 95% of the time and the true coefficient of $sqrt(Median\ Home\ Value)$ will lie between 0.002 and 0.021, 95% of the time. We also see that the confidence intervals for the model intercept and $Median\ Gross\ Rent$ contain 0, and thus we cannot make any inferences about the actual values of either parameter.

Conclusions

Per our fitted model, an increase in total housing units corresponds to an increase in vacant housing and an increase in median home value corresponds to an increase in vacant housing. Thus, our model indicates that the more expensive a housing unit in Denver, Colorado is, the more likely it is to be available. If this trend is seen to continue, impoverished families and community members may not have sufficient access to affordable housing.

Public policy changes to address this issue involve implementing housing development projects that focus on the creation of affordable housing units. Specifically, Denver's Community Planning and Development committee has put together a project entitled "Expanding Housing Affordability". This project looks to allow [housing development] projects to build taller buildings if more affordable units are included, update the city's linkage fee (this requires all new developments to either include affordable housing or pay a fee that supports Denver's affordable housing fund), and to change state law on inclusionary housing (requirements that cities and states can establish for new for-sale or for-rent developments). [5]

References

1. *A Comprehensive Look at Housing Market Conditions Across America's Cities,* Cityscape , 2020, Vol. 22, No. 2, Two Essays on Unequal Growth in Housing (2020), pp. 111-132.

2. *Metro Denver housing market selling homes faster and faster,* https://www.denverpost.com/2021/03/03/denver-housing-market-high-costs-low-inventory/

3. *Latest Denver-area housing market stats: Median single-family home price reaches $560K*, https://www.9news.com/article/money/markets/real-estate/denver-metro-housing-market-stats/73-18e4f86a-1a06-48d6-9212-72ef24a4bd35

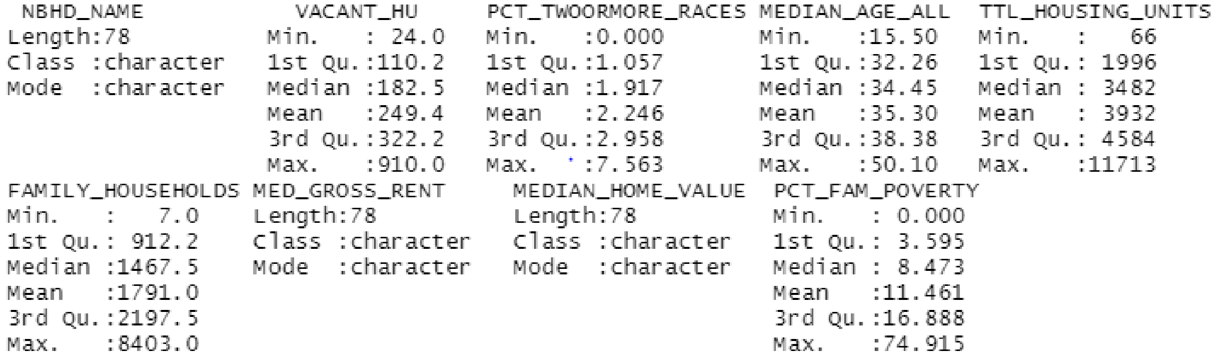
4. *Expanding Housing Affordability*, https://www.denvergov.org/Government/Departments/Community-Planning-and-Development/Denver-Zoning-Code/Text-Amendments/Affordable-Housing-Project#section-2

5. *American Community Survey Nbrhd (2013-2017),* https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-american-community-survey-nbrhd-2013-2017

Appendix

Code and console output used within report:

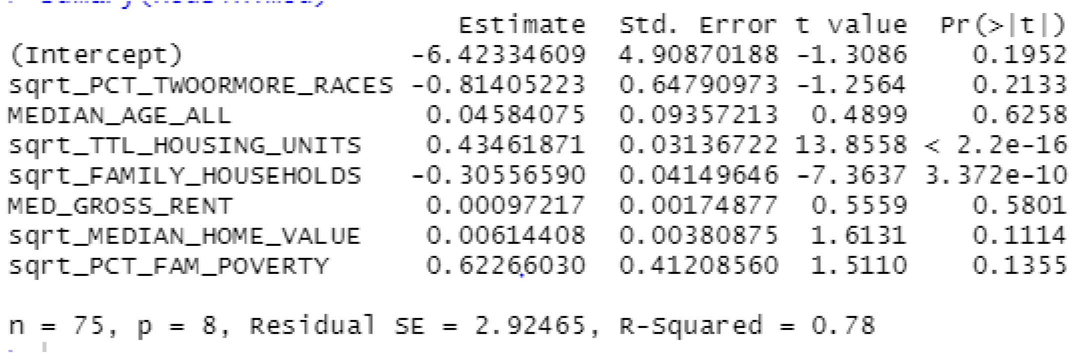
```
library(faraway)
housing <-
read.csv("C:/Users/adell/Desktop/american_community_survey_nbrhd_2013_2017.cs
v", header = TRUE,)
filthousing = subset(housing, select = c(NBHD_NAME, VACANT_HU,
PCT_TWOORMORE_RACES, MEDIAN_AGE_ALL, TTL_HOUSING_UNITS, FAMILY_HOUSEHOLDS,
MED_GROSS_RENT, MEDIAN_HOME_VALUE, PCT_FAM_POVERTY))

rownames(filthousing) <- filthousing$NBHD_NAME
summary(filthousing)
```



```
   NBHD_NAME            VACANT_HU     PCT_TWOORMORE_RACES MEDIAN_AGE_ALL  TTL_HOUSING_UNITS
 Length:78          Min.   : 24.0   Min.   :0.000       Min.   :15.50   Min.   :   66    
 Class :character   1st Qu.:110.2   1st Qu.:1.057       1st Qu.:32.26   1st Qu.: 1996    
 Mode  :character   Median :182.5   Median :1.917       Median :34.45   Median : 3482    
                    Mean   :249.4   Mean   :2.246       Mean   :35.30   Mean   : 3932    
                    3rd Qu.:322.2   3rd Qu.:2.958       3rd Qu.:38.38   3rd Qu.: 4584    
                    Max.   :910.0   Max.   :7.563       Max.   :50.10   Max.   :11713    
 FAMILY_HOUSEHOLDS MED_GROSS_RENT      MEDIAN_HOME_VALUE  PCT_FAM_POVERTY 
 Min.   :   7.0    Length:78          Length:78          Min.   : 0.000  
 1st Qu.: 912.2    Class :character   Class :character   1st Qu.: 3.595  
 Median :1467.5    Mode  :character   Mode  :character   Median : 8.473  
 Mean   :1791.0                                          Mean   :11.461  
 3rd Qu.:2197.5                                          3rd Qu.:16.888  
 Max.   :8403.0                                          Max.   :74.915  
```

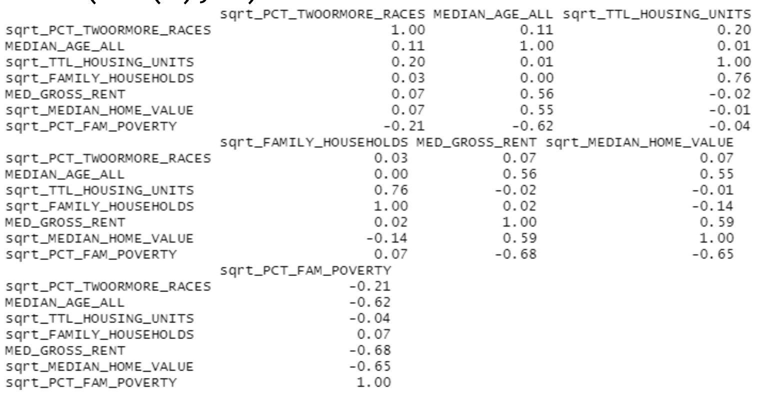
```
filthousing$MEDIAN_HOME_VALUE = as.numeric(filthousing$MEDIAN_HOME_VALUE)
filthousing$MED_GROSS_RENT = as.numeric(filthousing$MED_GROSS_RENT)

filthousing  = na.action = na.exclude(filthousing)
#Initial Data Exploration
plot(density(filthousing$VACANT_HU))
plot(density(filthousing$PCT_TWOORMORE_RACES))
plot(density(filthousing$MEDIAN_AGE_ALL))
plot(density(filthousing$TTL_HOUSING_UNITS))
plot(density(filthousing$FAMILY_HOUSEHOLDS))
plot(density(filthousing$MED_GROSS_RENT))
plot(density(filthousing$MEDIAN_HOME_VALUE))
plot(density(filthousing$PCT_FAM_POVERTY))
#Make Transformations
filthousing$sqrt_PCT_FAM_POVERTY = sqrt(filthousing$PCT_FAM_POVERTY)
filthousing$sqrt_VACANT_HU = sqrt(filthousing$VACANT_HU)
filthousing$sqrt_MEDIAN_HOME_VALUE = sqrt(filthousing$MEDIAN_HOME_VALUE)
filthousing$sqrt_MED_GROSS_RENT = sqrt(filthousing$MED_GROSS_RENT)
filthousing$sqrt_FAMILY_HOUSEHOLDS = sqrt(filthousing$FAMILY_HOUSEHOLDS)
filthousing$sqrt_PCT_TWOORMORE_RACES = sqrt(filthousing$PCT_TWOORMORE_RACES)
filthousing$sqrt_TTL_HOUSING_UNITS = sqrt(filthousing$TTL_HOUSING_UNITS)
##Data exploration
library(car)
library(perturb)
pairs(~sqrt_VACANT_HU + sqrt_PCT_TWOORMORE_RACES + MEDIAN_AGE_ALL+
sqrt_TTL_HOUSING_UNITS+ sqrt_FAMILY_HOUSEHOLDS+ MED_GROSS_RENT+
sqrt_MEDIAN_HOME_VALUE+ sqrt_PCT_FAM_POVERTY, data = filthousing, col =
"darkgreen", upper.panel=NULL, cex.labels = .5)
par(mar=c(1,1,1,1))
dens <- density(filthousing$sqrt_VACANT_HU)
plot(dens, frame = FALSE, col = "darkgreen",
     main = "sqrt(Vacant Housing in Denver)")
polygon(dens, col = "darkgreen")
```

```
#Assess for collinearity and amputate variables as needed
housinlmod <- lm(sqrt_VACANT_HU ~ sqrt_PCT_TWOORMORE_RACES + MEDIAN_AGE_ALL+
sqrt_TTL_HOUSING_UNITS+ sqrt_FAMILY_HOUSEHOLDS+ MED_GROSS_RENT+
sqrt_MEDIAN_HOME_VALUE+ sqrt_PCT_FAM_POVERTY, data = filthousing)
sumary(housinlmod)
```



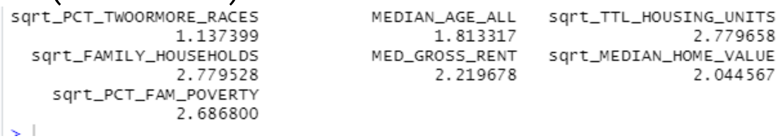
```
                            Estimate  Std. Error t value  Pr(>|t|)
(Intercept)              -6.42334609  4.90870188 -1.3086    0.1952
sqrt_PCT_TWOORMORE_RACES -0.81405223  0.64790973 -1.2564    0.2133
MEDIAN_AGE_ALL            0.04584075  0.09357213  0.4899    0.6258
sqrt_TTL_HOUSING_UNITS    0.43461871  0.03136722 13.8558 < 2.2e-16
sqrt_FAMILY_HOUSEHOLDS   -0.30556590  0.04149646 -7.3637 3.372e-10
MED_GROSS_RENT            0.00097217  0.00174877  0.5559    0.5801
sqrt_MEDIAN_HOME_VALUE    0.00614408  0.00380875  1.6131    0.1114
sqrt_PCT_FAM_POVERTY      0.62266030  0.41208560  1.5110    0.1355

n = 75, p = 8, Residual SE = 2.92465, R-Squared = 0.78
```

```
x = model.matrix(housinlmod)
x = x[,-1]
round(cor(x), 2)
```



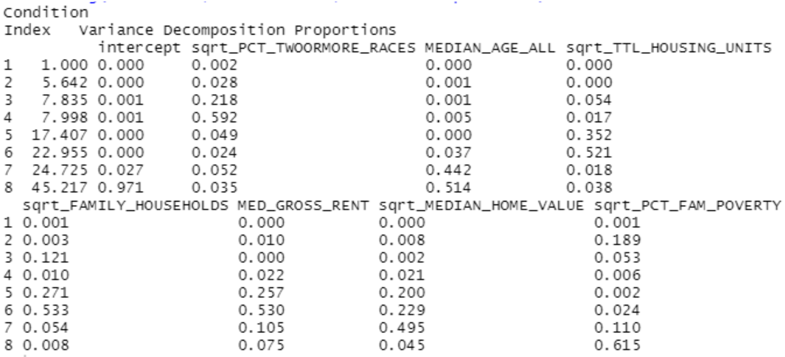
```
                         sqrt_PCT_TWOORMORE_RACES MEDIAN_AGE_ALL sqrt_TTL_HOUSING_UNITS
sqrt_PCT_TWOORMORE_RACES                     1.00           0.11                   0.20
MEDIAN_AGE_ALL                               0.11           1.00                   0.01
sqrt_TTL_HOUSING_UNITS                       0.20           0.01                   1.00
sqrt_FAMILY_HOUSEHOLDS                       0.03           0.00                   0.76
MED_GROSS_RENT                               0.07           0.56                  -0.02
sqrt_MEDIAN_HOME_VALUE                       0.07           0.55                  -0.01
sqrt_PCT_FAM_POVERTY                        -0.21          -0.62                  -0.04
                         sqrt_FAMILY_HOUSEHOLDS MED_GROSS_RENT sqrt_MEDIAN_HOME_VALUE
sqrt_PCT_TWOORMORE_RACES                   0.03           0.07                   0.07
MEDIAN_AGE_ALL                             0.00           0.56                   0.55
sqrt_TTL_HOUSING_UNITS                     0.76          -0.02                  -0.01
sqrt_FAMILY_HOUSEHOLDS                     1.00           0.02                  -0.14
MED_GROSS_RENT                             0.02           1.00                   0.59
sqrt_MEDIAN_HOME_VALUE                    -0.14           0.59                   1.00
sqrt_PCT_FAM_POVERTY                       0.07          -0.68                  -0.65
                         sqrt_PCT_FAM_POVERTY
sqrt_PCT_TWOORMORE_RACES                -0.21
MEDIAN_AGE_ALL                          -0.62
sqrt_TTL_HOUSING_UNITS                  -0.04
sqrt_FAMILY_HOUSEHOLDS                   0.07
MED_GROSS_RENT                          -0.68
sqrt_MEDIAN_HOME_VALUE                  -0.65
sqrt_PCT_FAM_POVERTY                     1.00
```

```
vif(housinlmod)
```



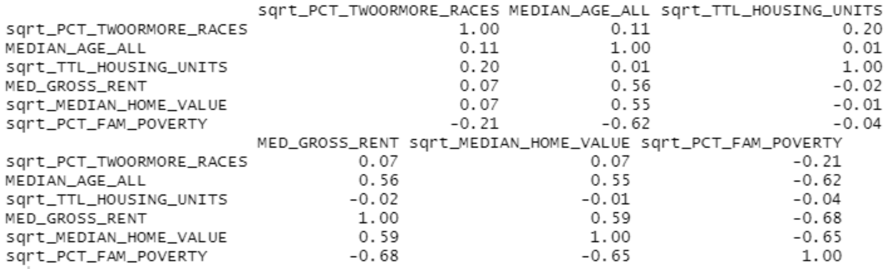
```
sqrt_PCT_TWOORMORE_RACES           MEDIAN_AGE_ALL   sqrt_TTL_HOUSING_UNITS
                1.137399                 1.813317                 2.779658
  sqrt_FAMILY_HOUSEHOLDS           MED_GROSS_RENT   sqrt_MEDIAN_HOME_VALUE
                2.779528                 2.219678                 2.044567
    sqrt_PCT_FAM_POVERTY
                2.686800
>
```

```
colldiag(housinlmod, scale = TRUE, add.intercept = TRUE)
```



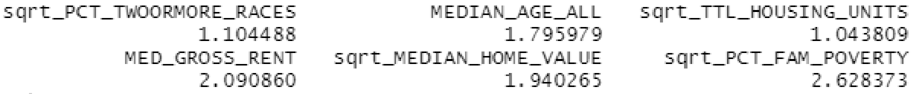
```
Condition
Index   Variance Decomposition Proportions
         intercept sqrt_PCT_TWOORMORE_RACES MEDIAN_AGE_ALL sqrt_TTL_HOUSING_UNITS
1   1.000 0.000     0.002                    0.000          0.000
2   5.642 0.000     0.028                    0.001          0.000
3   7.835 0.001     0.218                    0.001          0.054
4   7.998 0.001     0.592                    0.005          0.017
5  17.407 0.000     0.049                    0.000          0.352
6  22.955 0.000     0.024                    0.037          0.521
7  24.725 0.027     0.052                    0.442          0.018
8  45.217 0.971     0.035                    0.514          0.038
  sqrt_FAMILY_HOUSEHOLDS MED_GROSS_RENT sqrt_MEDIAN_HOME_VALUE sqrt_PCT_FAM_POVERTY
1 0.001                  0.000          0.000                  0.001
2 0.003                  0.010          0.008                  0.189
3 0.121                  0.000          0.002                  0.053
4 0.010                  0.022          0.021                  0.006
5 0.271                  0.257          0.200                  0.002
6 0.533                  0.530          0.229                  0.024
7 0.054                  0.105          0.495                  0.110
8 0.008                  0.075          0.045                  0.615
```

```
##remove sqrt_FAMILY_HOUSEHOLDS
housinlmod2 = update(housinlmod, .~. -sqrt_FAMILY_HOUSEHOLDS)
x = model.matrix(housinlmod2)
x = x[,-1]
round(cor(x), 2)
```



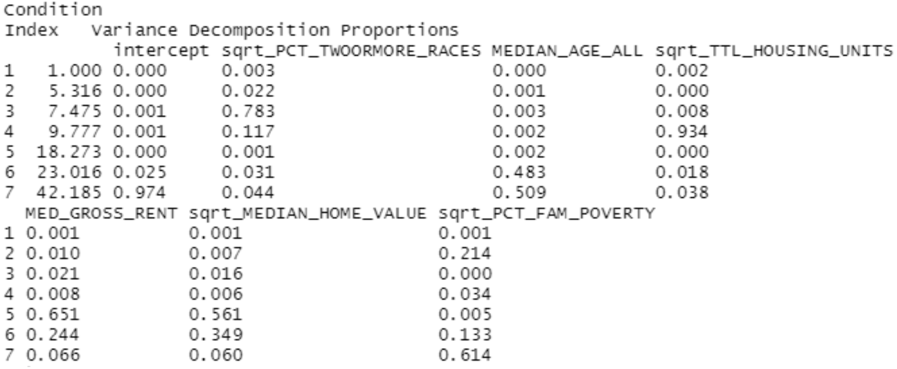
```
                         sqrt_PCT_TWOORMORE_RACES MEDIAN_AGE_ALL sqrt_TTL_HOUSING_UNITS
sqrt_PCT_TWOORMORE_RACES                     1.00           0.11                   0.20
MEDIAN_AGE_ALL                               0.11           1.00                   0.01
sqrt_TTL_HOUSING_UNITS                       0.20           0.01                   1.00
MED_GROSS_RENT                               0.07           0.56                  -0.02
sqrt_MEDIAN_HOME_VALUE                       0.07           0.55                  -0.01
sqrt_PCT_FAM_POVERTY                        -0.21          -0.62                  -0.04
                         MED_GROSS_RENT sqrt_MEDIAN_HOME_VALUE sqrt_PCT_FAM_POVERTY
sqrt_PCT_TWOORMORE_RACES           0.07                   0.07                -0.21
MEDIAN_AGE_ALL                     0.56                   0.55                -0.62
sqrt_TTL_HOUSING_UNITS            -0.02                  -0.01                -0.04
MED_GROSS_RENT                     1.00                   0.59                -0.68
sqrt_MEDIAN_HOME_VALUE             0.59                   1.00                -0.65
sqrt_PCT_FAM_POVERTY              -0.68                  -0.65                 1.00
```

```
vif(housinlmod2)
```



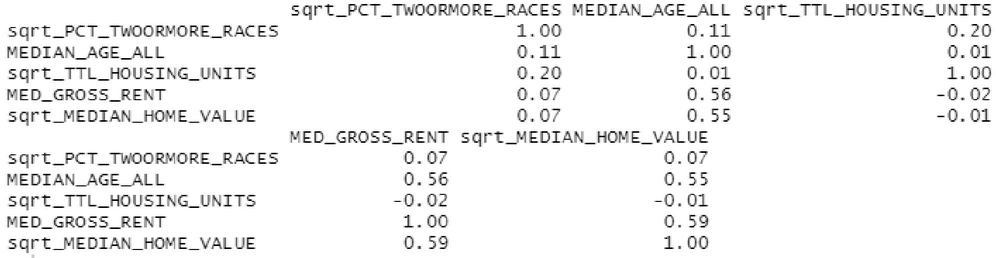
```
sqrt_PCT_TWOORMORE_RACES           MEDIAN_AGE_ALL   sqrt_TTL_HOUSING_UNITS
                1.104488                 1.795979                 1.043809
          MED_GROSS_RENT   sqrt_MEDIAN_HOME_VALUE     sqrt_PCT_FAM_POVERTY
                2.090860                 1.940265                 2.628373
```

```
colldiag(housinlmod2, scale = TRUE, add.intercept = TRUE)
```



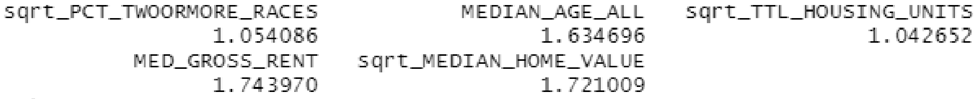
```
Condition
Index   Variance Decomposition Proportions
         intercept sqrt_PCT_TWOORMORE_RACES MEDIAN_AGE_ALL sqrt_TTL_HOUSING_UNITS
1   1.000 0.000     0.003                    0.000          0.002
2   5.316 0.000     0.022                    0.001          0.000
3   7.475 0.001     0.783                    0.003          0.008
4   9.777 0.001     0.117                    0.002          0.934
5  18.273 0.000     0.001                    0.002          0.000
6  23.016 0.025     0.031                    0.483          0.018
7  42.185 0.974     0.044                    0.509          0.038
  MED_GROSS_RENT sqrt_MEDIAN_HOME_VALUE sqrt_PCT_FAM_POVERTY
1 0.001          0.001                  0.001
2 0.010          0.007                  0.214
3 0.021          0.016                  0.000
4 0.008          0.006                  0.034
5 0.651          0.561                  0.005
6 0.244          0.349                  0.133
7 0.066          0.060                  0.614
```

```
##remove sqrt_PCT_FAM_POVERTY
housinlmod3 = update(housinlmod2, .~. -sqrt_PCT_FAM_POVERTY)
x = model.matrix(housinlmod3)
x = x[,-1]
round(cor(x), 2)
```



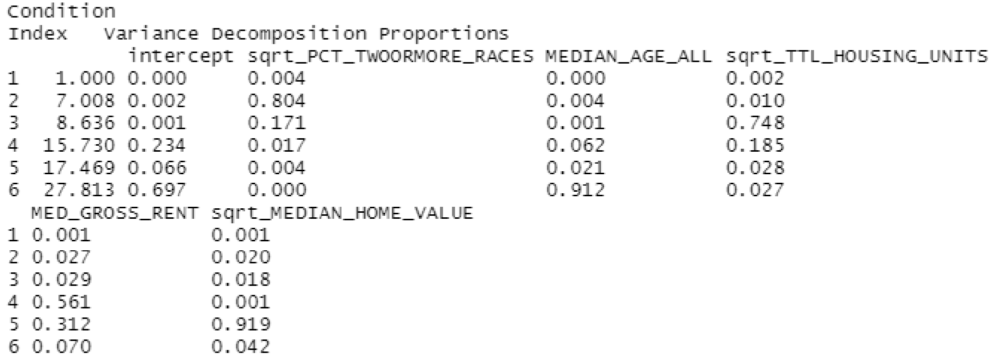
```
                         sqrt_PCT_TWOORMORE_RACES MEDIAN_AGE_ALL sqrt_TTL_HOUSING_UNITS
sqrt_PCT_TWOORMORE_RACES                     1.00           0.11                   0.20
MEDIAN_AGE_ALL                               0.11           1.00                   0.01
sqrt_TTL_HOUSING_UNITS                       0.20           0.01                   1.00
MED_GROSS_RENT                               0.07           0.56                  -0.02
sqrt_MEDIAN_HOME_VALUE                       0.07           0.55                  -0.01
                         MED_GROSS_RENT sqrt_MEDIAN_HOME_VALUE
sqrt_PCT_TWOORMORE_RACES           0.07                   0.07
MEDIAN_AGE_ALL                     0.56                   0.55
sqrt_TTL_HOUSING_UNITS            -0.02                  -0.01
MED_GROSS_RENT                     1.00                   0.59
sqrt_MEDIAN_HOME_VALUE             0.59                   1.00
```

```
vif(housinlmod3)
```



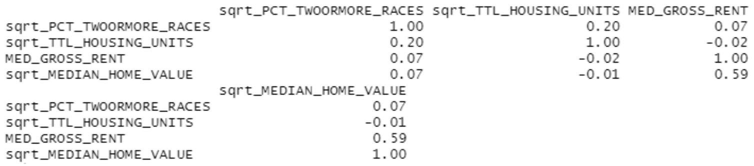
```
sqrt_PCT_TWOORMORE_RACES           MEDIAN_AGE_ALL   sqrt_TTL_HOUSING_UNITS
                1.054086                 1.634696                 1.042652
          MED_GROSS_RENT   sqrt_MEDIAN_HOME_VALUE
                1.743970                 1.721009
```

```
colldiag(housinlmod3, scale = TRUE, add.intercept = TRUE)
```



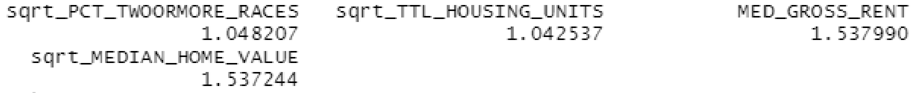
```
Condition
Index   Variance Decomposition Proportions
          intercept sqrt_PCT_TWOORMORE_RACES MEDIAN_AGE_ALL sqrt_TTL_HOUSING_UNITS
1   1.000 0.000     0.004                    0.000          0.002
2   7.008 0.002     0.804                    0.004          0.010
3   8.636 0.001     0.171                    0.001          0.748
4  15.730 0.234     0.017                    0.062          0.185
5  17.469 0.066     0.004                    0.021          0.028
6  27.813 0.697     0.000                    0.912          0.027
  MED_GROSS_RENT sqrt_MEDIAN_HOME_VALUE
1 0.001          0.001
2 0.027          0.020
3 0.029          0.018
4 0.561          0.001
5 0.312          0.919
6 0.070          0.042
```

```
##remove MEDIAN_AGE_ALL
housinlmod4 = update(housinlmod3, .~. -MEDIAN_AGE_ALL)
x = model.matrix(housinlmod4)
x = x[,-1]
round(cor(x), 2)
```



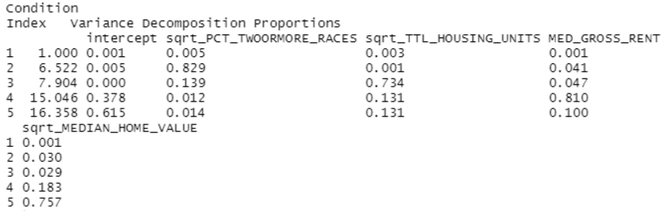
```
                         sqrt_PCT_TWOORMORE_RACES sqrt_TTL_HOUSING_UNITS MED_GROSS_RENT
sqrt_PCT_TWOORMORE_RACES                     1.00                   0.20           0.07
sqrt_TTL_HOUSING_UNITS                       0.20                   1.00          -0.02
MED_GROSS_RENT                               0.07                  -0.02           1.00
sqrt_MEDIAN_HOME_VALUE                       0.07                  -0.01           0.59
                         sqrt_MEDIAN_HOME_VALUE
sqrt_PCT_TWOORMORE_RACES                   0.07
sqrt_TTL_HOUSING_UNITS                    -0.01
MED_GROSS_RENT                             0.59
sqrt_MEDIAN_HOME_VALUE                     1.00
```

```
vif(housinlmod4)
```



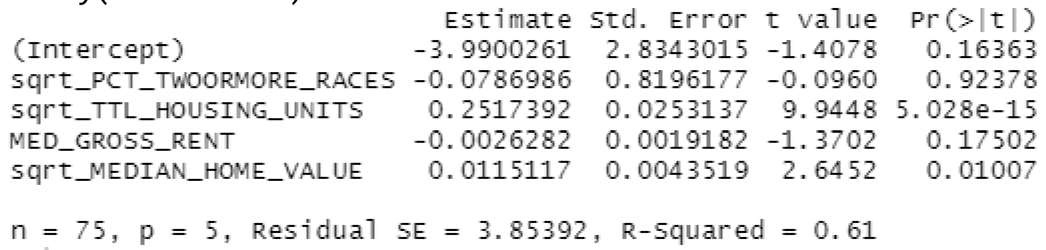
```
sqrt_PCT_TWOORMORE_RACES   sqrt_TTL_HOUSING_UNITS           MED_GROSS_RENT
                1.048207                 1.042537                 1.537990
  sqrt_MEDIAN_HOME_VALUE
                1.537244
```

```
colldiag(housinlmod4, scale = TRUE, add.intercept = TRUE)
```



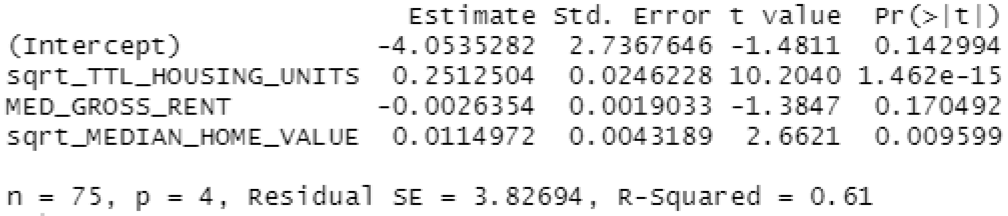
```
Condition
Index   Variance Decomposition Proportions
          intercept sqrt_PCT_TWOORMORE_RACES sqrt_TTL_HOUSING_UNITS MED_GROSS_RENT
1   1.000 0.001     0.005                    0.003                  0.001
2   6.522 0.005     0.829                    0.001                  0.041
3   7.904 0.000     0.139                    0.734                  0.047
4  15.046 0.378     0.012                    0.131                  0.810
5  16.358 0.615     0.014                    0.131                  0.100
  sqrt_MEDIAN_HOME_VALUE
1 0.001
2 0.030
3 0.029
4 0.183
5 0.757
```

```
####Variable selection

# use alpha_crit = 0.05
sumary(housinlmod4)
```



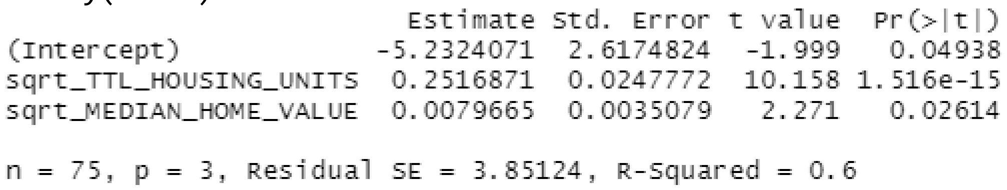
```
                           Estimate Std. Error t value  Pr(>|t|)
(Intercept)              -3.9900261  2.8343015 -1.4078   0.16363
sqrt_PCT_TWOORMORE_RACES -0.0786986  0.8196177 -0.0960   0.92378
sqrt_TTL_HOUSING_UNITS    0.2517392  0.0253137  9.9448 5.028e-15
MED_GROSS_RENT           -0.0026282  0.0019182 -1.3702   0.17502
sqrt_MEDIAN_HOME_VALUE    0.0115117  0.0043519  2.6452   0.01007

n = 75, p = 5, Residual SE = 3.85392, R-Squared = 0.61
```

```
lmod2 <- update(housinlmod4, . ~ . -sqrt_PCT_TWOORMORE_RACES)
sumary(lmod2)
```



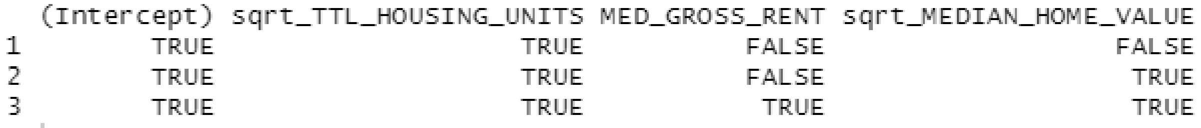
```
                         Estimate Std. Error t value  Pr(>|t|)
(Intercept)            -4.0535282  2.7367646 -1.4811  0.142994
sqrt_TTL_HOUSING_UNITS  0.2512504  0.0246228 10.2040 1.462e-15
MED_GROSS_RENT         -0.0026354  0.0019033 -1.3847  0.170492
sqrt_MEDIAN_HOME_VALUE  0.0114972  0.0043189  2.6621  0.009599

n = 75, p = 4, Residual SE = 3.82694, R-Squared = 0.61
```

```
lmod3 <- update(lmod2, . ~ . - MED_GROSS_RENT)
sumary(lmod3)
```



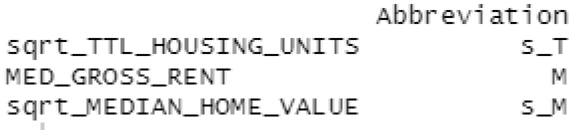
```
                         Estimate Std. Error t value  Pr(>|t|)
(Intercept)            -5.2324071  2.6174824  -1.999   0.04938
sqrt_TTL_HOUSING_UNITS  0.2516871  0.0247772  10.158 1.516e-15
sqrt_MEDIAN_HOME_VALUE  0.0079665  0.0035079   2.271   0.02614

n = 75, p = 3, Residual SE = 3.85124, R-Squared = 0.6
```

```
# model selection in terms of AIC
library(leaps)
b <- regsubsets(sqrt_VACANT_HU ~ sqrt_TTL_HOUSING_UNITS + MED_GROSS_RENT +
sqrt_MEDIAN_HOME_VALUE, data = filthousing)
rs <- summary(b)
rs$which
```



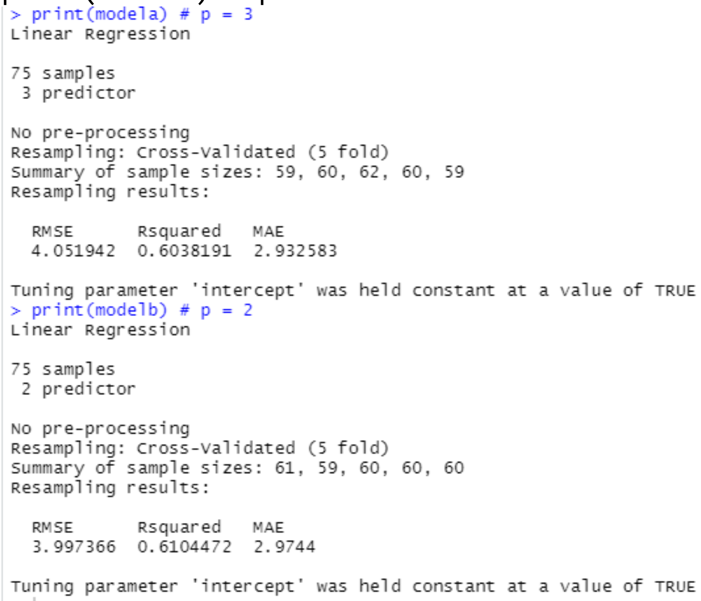
```
  (Intercept) sqrt_TTL_HOUSING_UNITS MED_GROSS_RENT sqrt_MEDIAN_HOME_VALUE
1        TRUE                   TRUE          FALSE                  FALSE
2        TRUE                   TRUE          FALSE                   TRUE
3        TRUE                   TRUE           TRUE                   TRUE
```

```
n = nobs(lmod2)
AIC <- n*log(rs$rss/n) + (2:4)*2
plot(AIC ~ I(1:3), ylab="AIC", xlab="Number of Predictors", col = "blue", cex
= 3, pch = 16)

# Construct adjusted R^2 plot
which.max(rs$adjr2)
subsets(b, statistic = "adjr2", legend = FALSE, col = "blue")
```



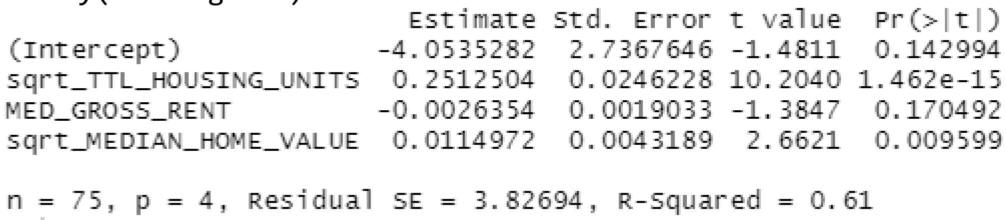
```
                       Abbreviation
sqrt_TTL_HOUSING_UNITS          s_T
MED_GROSS_RENT                    M
sqrt_MEDIAN_HOME_VALUE          s_M
```

```
## compare with model performance with 3 vs with 2 regressors:
library(caret)
f1 = sqrt_VACANT_HU ~ sqrt_TTL_HOUSING_UNITS + MED_GROSS_RENT +
sqrt_MEDIAN_HOME_VALUE
f2 = sqrt_VACANT_HU ~ sqrt_TTL_HOUSING_UNITS + sqrt_MEDIAN_HOME_VALUE
cv_5fold = trainControl(method = "cv", number = 5) # 5-fold crossvalidation
train/test data
modela = train(f1, data = filthousing, trControl = cv_5fold,
               method = "lm")
modelb = train(f2, data = filthousing, trControl = cv_5fold,
               method = "lm")
# compare mse (rmse) for the three models using 5-fold cv
print(modela) # p = 3
print(modelb) # p = 2
```



```
> print(modela) # p = 3
Linear Regression

75 samples
 3 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 59, 60, 62, 60, 59
Resampling results:

  RMSE      Rsquared   MAE
  4.051942  0.6038191  2.932583

Tuning parameter 'intercept' was held constant at a value of TRUE
> print(modelb) # p = 2
Linear Regression

75 samples
 2 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 61, 59, 60, 60, 60
Resampling results:

  RMSE      Rsquared   MAE
  3.997366  0.6104472  2.9744

Tuning parameter 'intercept' was held constant at a value of TRUE
```

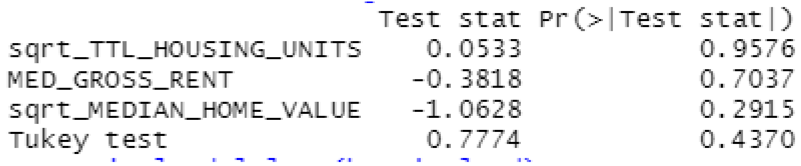
```
###Model with three regressors performs better.

housinglmod <- lm(sqrt_VACANT_HU ~ sqrt_TTL_HOUSING_UNITS + MED_GROSS_RENT +
sqrt_MEDIAN_HOME_VALUE, data = filthousing)
sumary(housinglmod)
```



```
                          Estimate Std. Error t value  Pr(>|t|)
(Intercept)             -4.0535282  2.7367646 -1.4811  0.142994
sqrt_TTL_HOUSING_UNITS   0.2512504  0.0246228 10.2040 1.462e-15
MED_GROSS_RENT          -0.0026354  0.0019033 -1.3847  0.170492
sqrt_MEDIAN_HOME_VALUE   0.0114972  0.0043189  2.6621  0.009599

n = 75, p = 4, Residual SE = 3.82694, R-Squared = 0.61
```

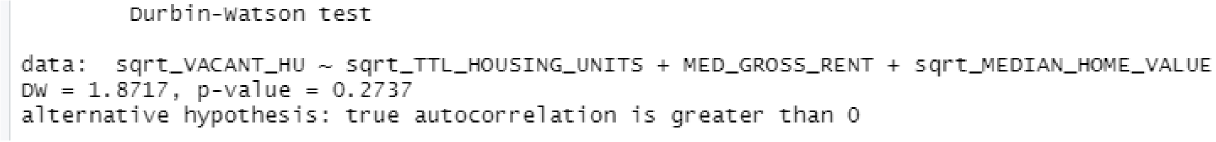
```
#plot yhat vs y
plot(predict(housinglmod),filthousing$VACANT_HU,
     xlab="yhat",ylab="y", col = "blue")
abline(l <- lm(filthousing$VACANT_HU~ predict(housinglmod)), col = "red")

##investigating structure
par(mar=c(1,1,1,1))
```

```
residualPlots(housinglmod)
```



```
                       Test stat Pr(>|Test stat|)
sqrt_TTL_HOUSING_UNITS    0.0533           0.9576
MED_GROSS_RENT           -0.3818           0.7037
sqrt_MEDIAN_HOME_VALUE   -1.0628           0.2915
Tukey test                0.7774           0.4370
```

```
marginalModelPlots(housinglmod)
avPlots(housinglmod, id = FALSE)

h <- hatvalues(housinglmod)
neighborhoods <- filthousing$NBHD_NAME
halfnorm(h, nlab = 2, labs = neighborhoods, ylab = "leverage")
infIndexPlot(housinglmod, vars = "hat")

halfnorm(cook, n=1, labs = neighborhoods, ylab = "Cook's Distances", col =
"blue")
#obtain leverage vs. standardized residuals in plot 4:
plot(housinglmod, col = "blue")
qqPlot(housinglmod)

n = nobs(housinglmod)
plot(tail(residuals(housinglmod), n - 1) ~ head(residuals(housinglmod), n -
1),
     xlab = expression(hat(epsilon)[i]),
     ylab = expression(hat(epsilon)[i + 1]))
abline(h = 0, v = 0, col = grey(0.75))


library(lmtest)
dwtest(sqrt_VACANT_HU ~ sqrt_TTL_HOUSING_UNITS + MED_GROSS_RENT +
sqrt_MEDIAN_HOME_VALUE, data = filthousing)
```



```
	Durbin-Watson test

data:  sqrt_VACANT_HU ~ sqrt_TTL_HOUSING_UNITS + MED_GROSS_RENT + sqrt_MEDIAN_HOME_VALUE
DW = 1.8717, p-value = 0.2737
alternative hypothesis: true autocorrelation is greater than 0
```