Data Science Competition Report

Isaac Kistler and Enyu Li

April 3, 2022

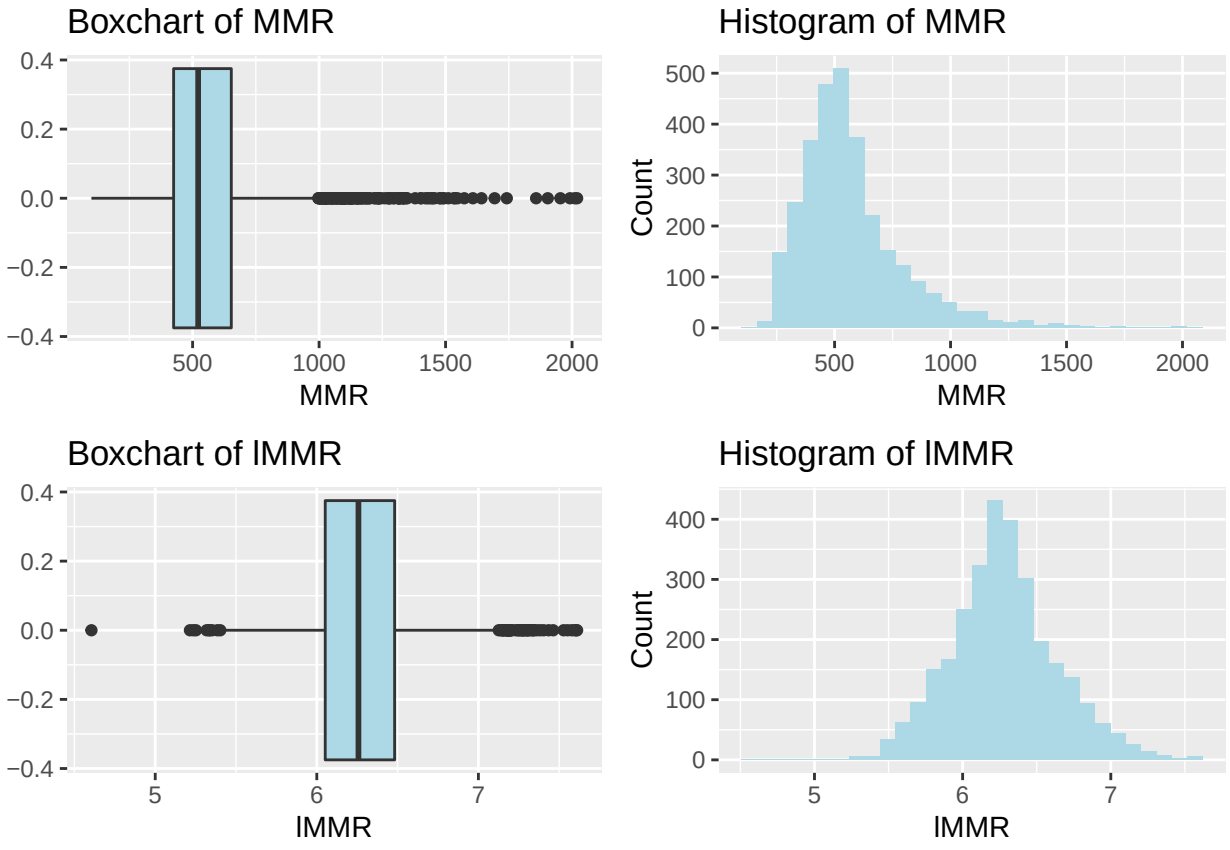
Introduction

Rent prices are something that are of much importance to college graduate students, who most often live in rentals as they work to finish their degrees. As such, it is of interest to find the factors that influence rent price. In this report, we will use data from the American Communities Survey (ACS) to determine which factors affect the monthly median rent, or MMR, for a one bedroom unit, averaged over the past 5 years.

Data Analysis

First, we must evaluate our outcome variable and get a general idea of distribution of the data. By viewing the first set of statistics below, along with the two upper plots, we can see that MMR is heavily right skewed.

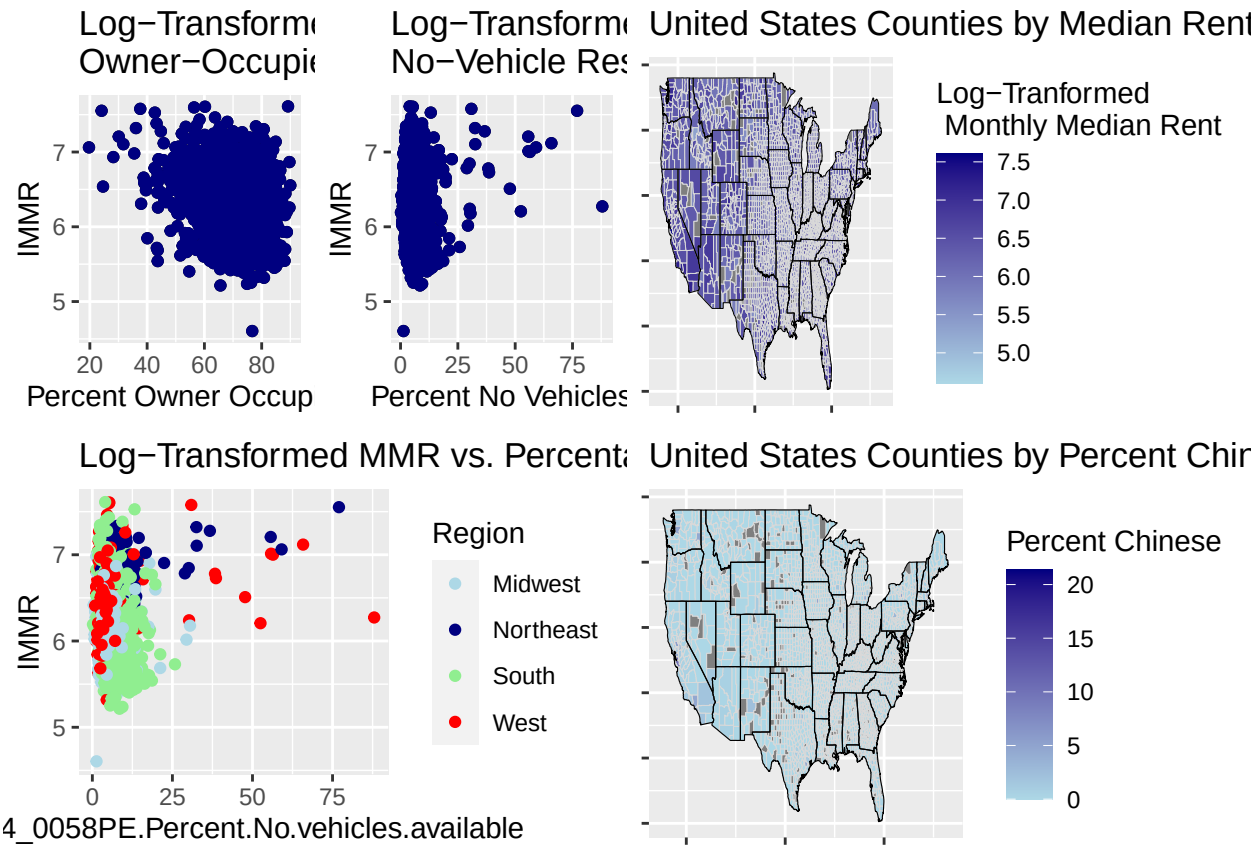
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. Std.Dev.
##    100.0    425.0    522.0    569.7    653.0   2020.0    225.3
```




```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. Std.Dev.
##   4.6052   6.0521   6.2577   6.2781   6.4816   7.6109   0.3594
```

We next create a log-transformed monthly median rent (lMMR) variable. This is preferred because it normalizes the distribution, which allows linear modeling assumptions to be met and also condenses the range of the data. This is shown in the two plots on the second row, along with the statistics below them.

After exploring and transforming MMR, the next step is to see what relationships exist between the transformed response variable and its potential predictors. An initial step for this is to produce scatterplots, such as the two in the first row below, in order to spot any obvious visual trends. For example, the percentage of owner occupied residences has a negative relationship with lMMR, while there does not appear to be a strong relationship between lMMR and the percentage of residences with no vehicles available. Additionally, it is good to see how these variables interact with each other in different categories. Factor variables were created to classify the counties into categories based on geographical region, as well as other variables which can be found in the appendix. The scatterplot in the second row above demonstrates the relations of this geographical region to lMMR and the percentage of residences with no vehicles available in a county. From this we see that the midwest and south tend to have lower lMMR's, while the the west and northeast tend to be have higher lMMRs.



In order to find any geographic trends, we created a county-level map of log-transformed monthly median rent, as seen in the first map above. From this we were able to see that the more expensive (darker) areas that would typically be considered more urbanized, such as the coast of California, the New England area, and other areas that contain large cities. In addition, although it is not a clear trend it does appear to be more expensive in areas closer to the west coast. In an effort to see how this compares to our potential predictors, we created similar maps according to several

different variables. The one shown in the second map, above, shows a heat map of the percentage of the population that is Chinese, while several others can be seen in the appendix. The map of Chinese population percentage shows generally low percentages, darker areas can be seen, which do match up with some of the higher rent areas found in the US.

To select the most important variables to fit into the final model, we use ridge regression to find the optimal λ value via cross validation. We then start building the regression model from the variables whose coefficients are not zero. We first do a single linear regression on this and removed the two variables that are not statistically significant. We next check for collinearity by choosing one variable from each of the categories and rerunning the least square linear regression model. After that, we check the multicollinearity of the variables by checking the VIF values.

After checking the VIF of each variable and running a correlation matrix of each explanatory variable with other variables, we found the variables bachelor, Units, Internet, and Non_family income are highly correlated with each other. Thus, we decided to drop the variable Units and to add Bachelor and Internet together to create a linear regression model which ignores unnecessary variables without significantly affecting the overall performance.

After also looking for interaction using the FSA method in R, the last step is evaluate the model and check whether all assumptions are met. The assumptions of linearity, normality, and independence are all satisfied, but the assumption of equal variance is not. Though several methods were tried, including using log transformation of the dependent variable Y and using weighted linear regression models, the assumption of equal variance is violated and the bptest score always show there exists Heteroscedasticity. The final model, along with various other plots and models, are included in the appendix.

Results and Conclusion

Although the model did not meet all assumptions, it can still be used as an indicator to evaluate which factors impact rent prices .The primary factors for single bedroom apartments tended to involve percentage of apartments with more than one occupant, Vietnamese and Italian populations, and income by non-family households. These all make sense for different reasons. When looking for low rent apartments, it is best to find a small apartment which only has room for one occupant in an area with lower Vietnamese and Italian populations and less income by non-family units.

References

Business Insider. 2018. Even the US government can't agree on how to divide up the states into regions. Available at https://www.businessinsider.com/regions-of-united-states-2018-5

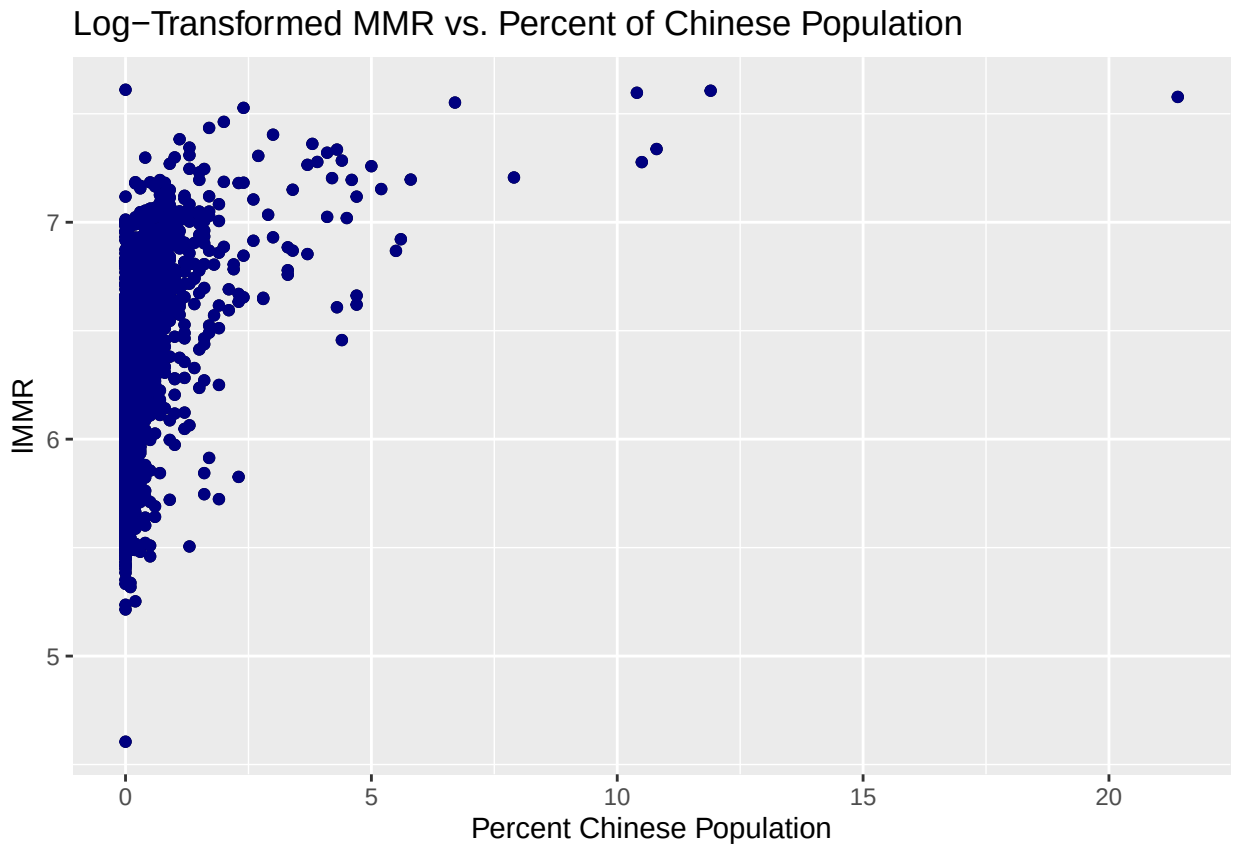
US Census Bureau. 2022. American Community Survey (ACS). Available at https://www.census.gov/programs-surveys/acs

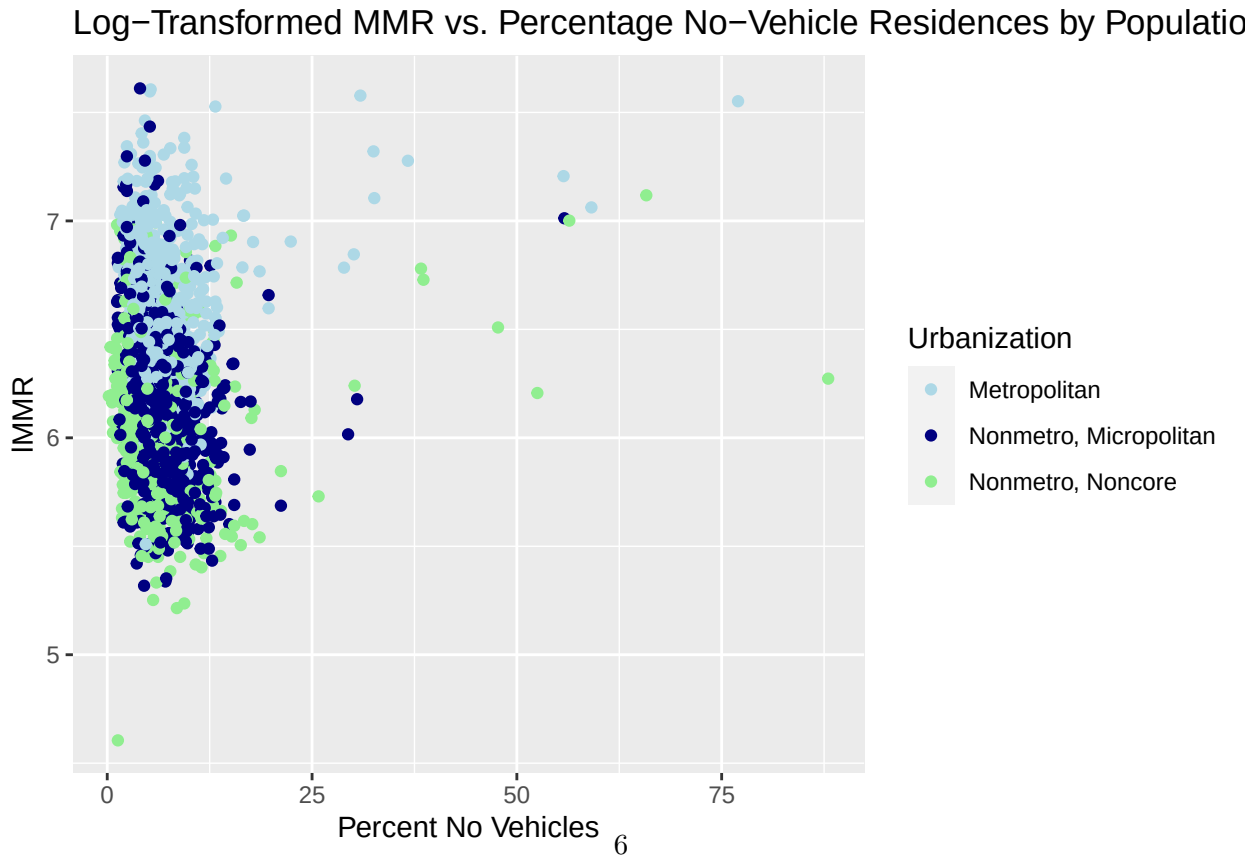
USDA ERS. 2019. What is Rural? Available at https://www.ers.usda.gov/topics/rural-economy-population/rural-classifications/what-is-rural/

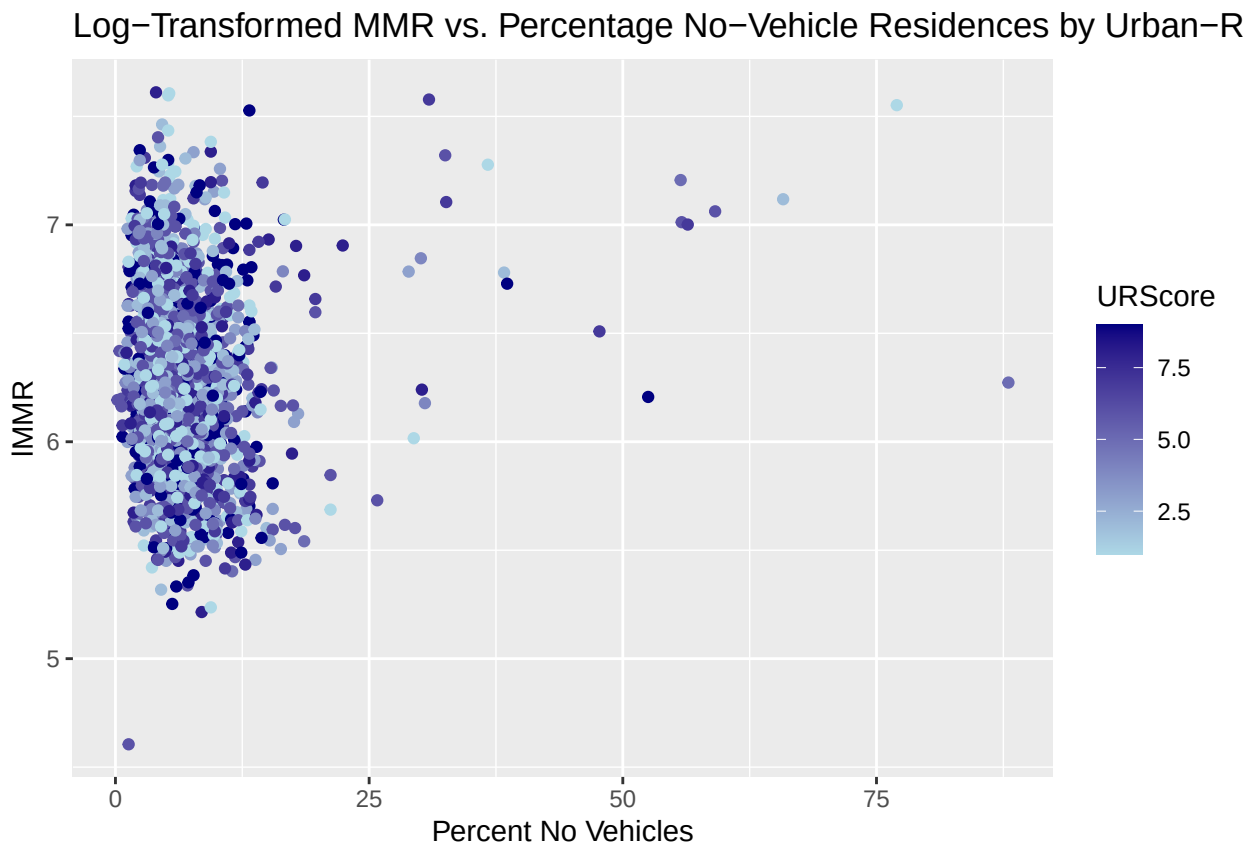
Code Used

https://github.com/IsaacSKistler/Data-Science-Competition

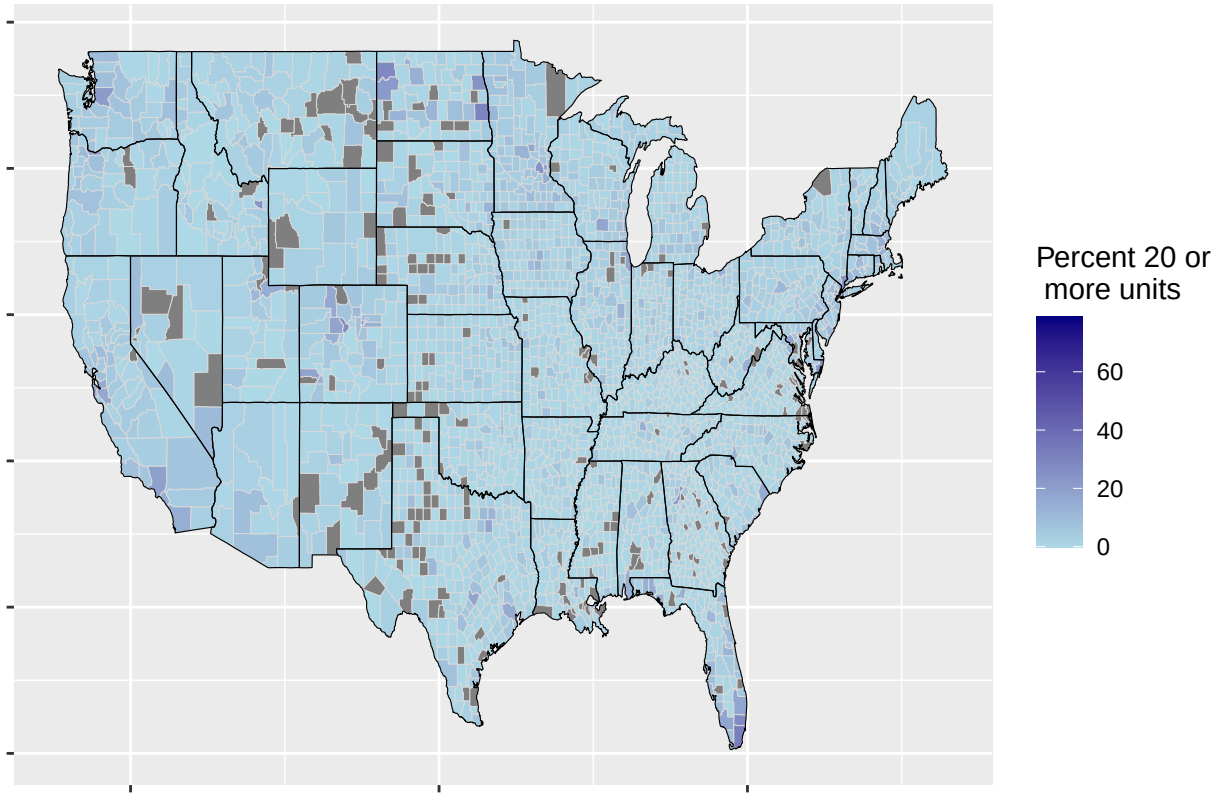
Appendix: Alternative Graphs and Models



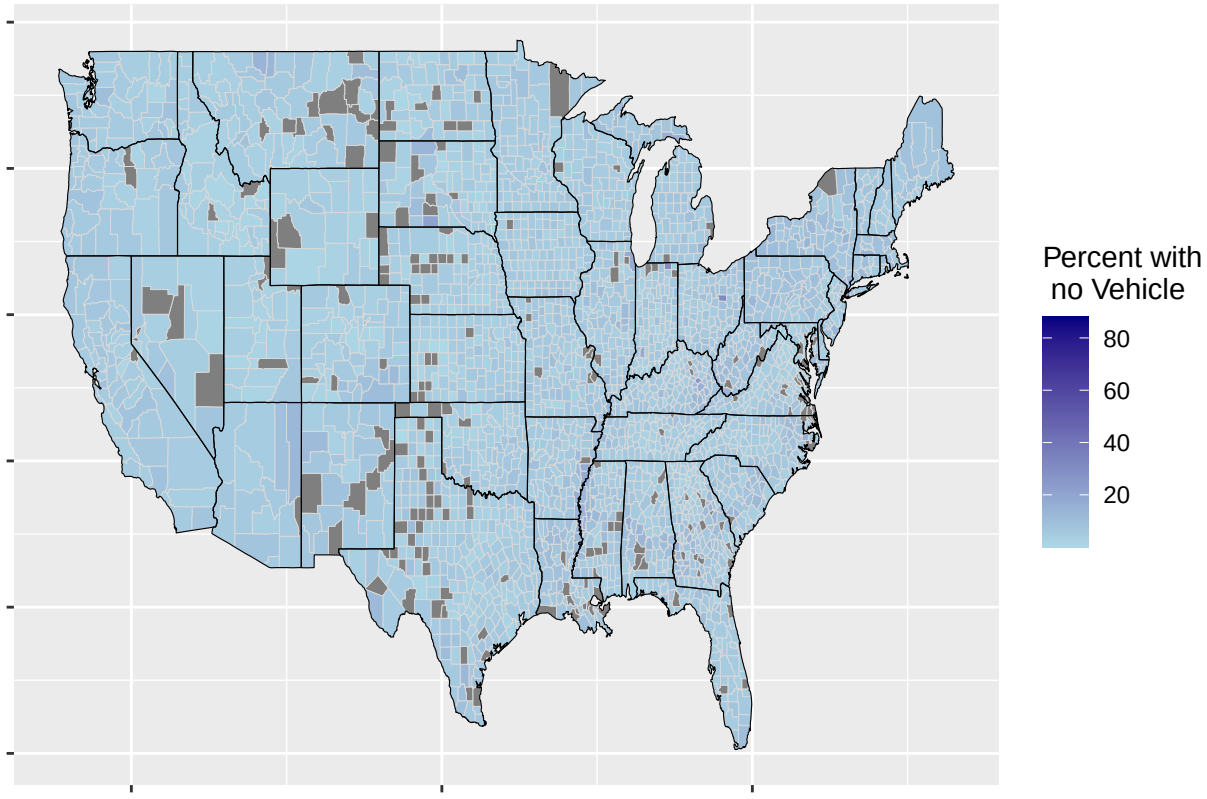


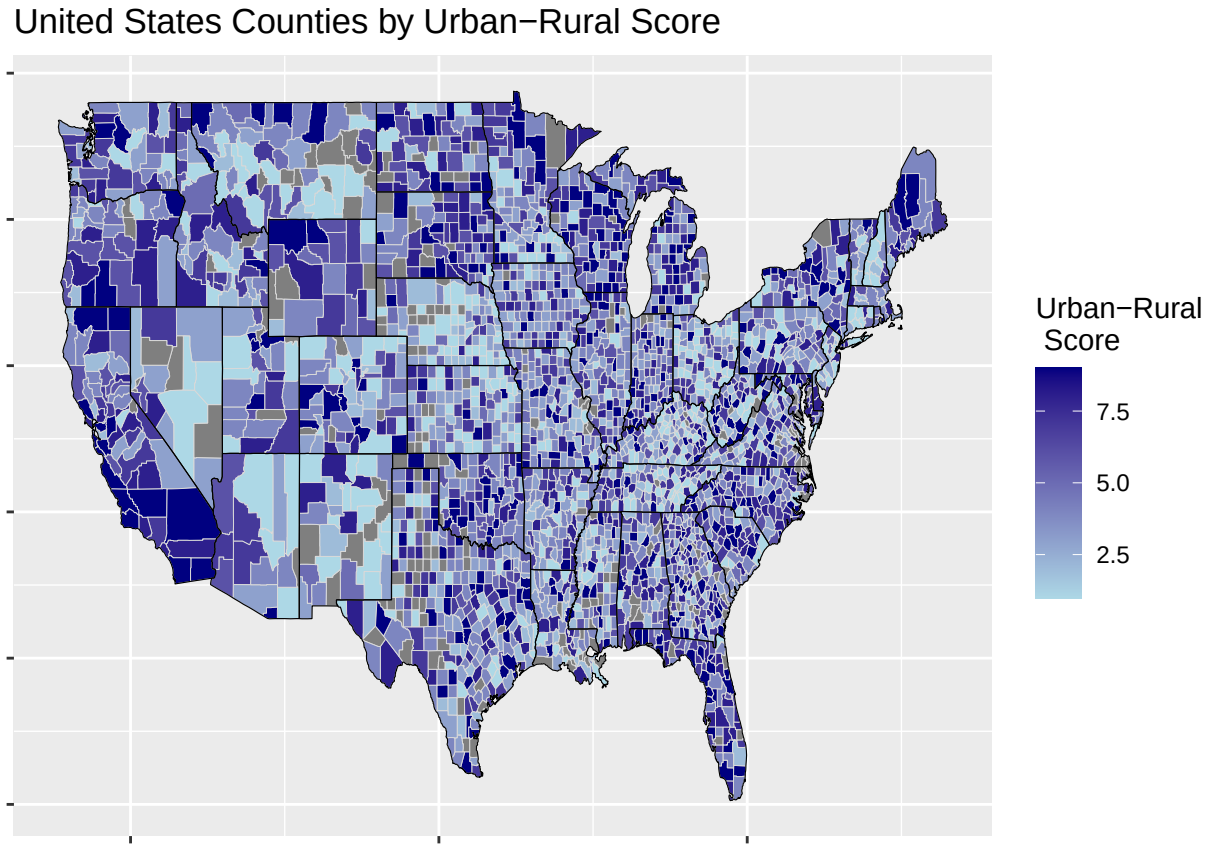

Log−Transformed MMR vs. Percentage No−Vehicle Residences by Urban−R
IMMR
7
6
5
0
25
50
75
Percent No Vehicles
URScore
7.5
5.0
2.5

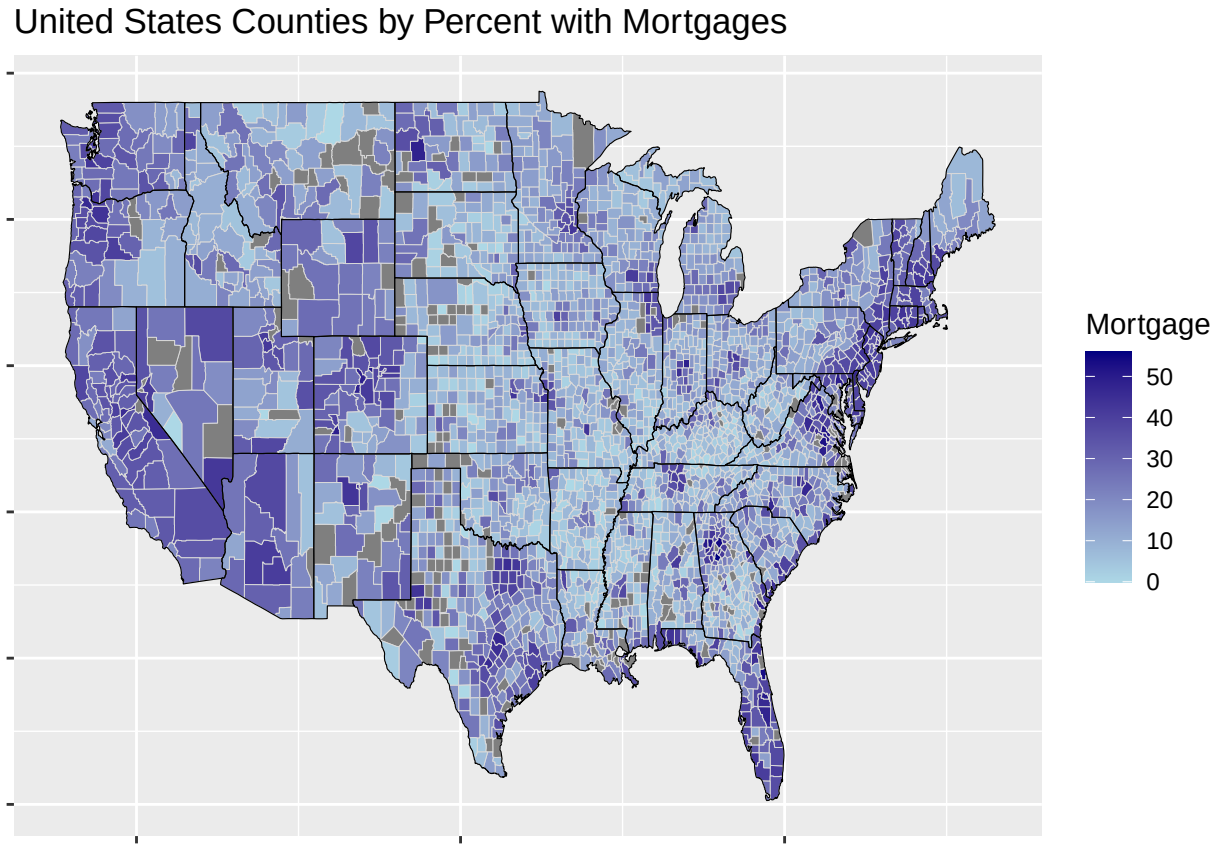
United States Counties by Percent of Housing Structures with 20 or More Units



United States Counties by Percent with No Vehicle



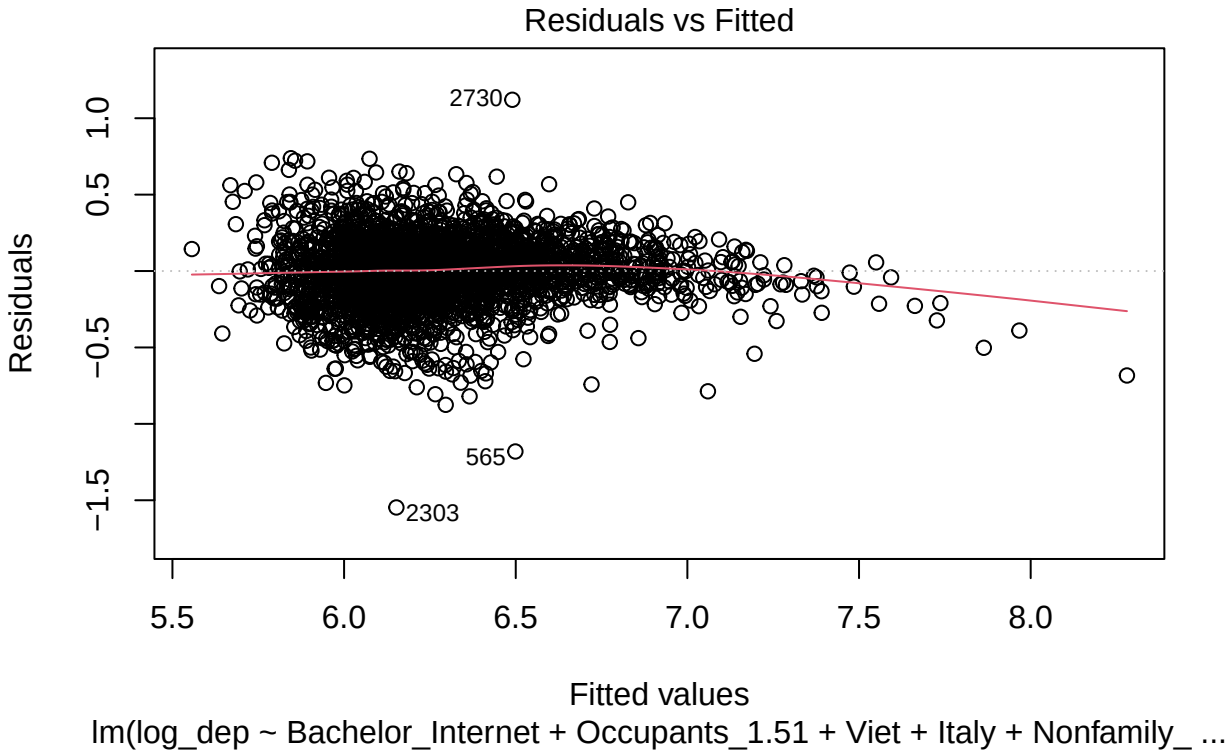

United States Counties by Urban−Rural Score
Urban−Rural Score
7.5
5.0
2.5

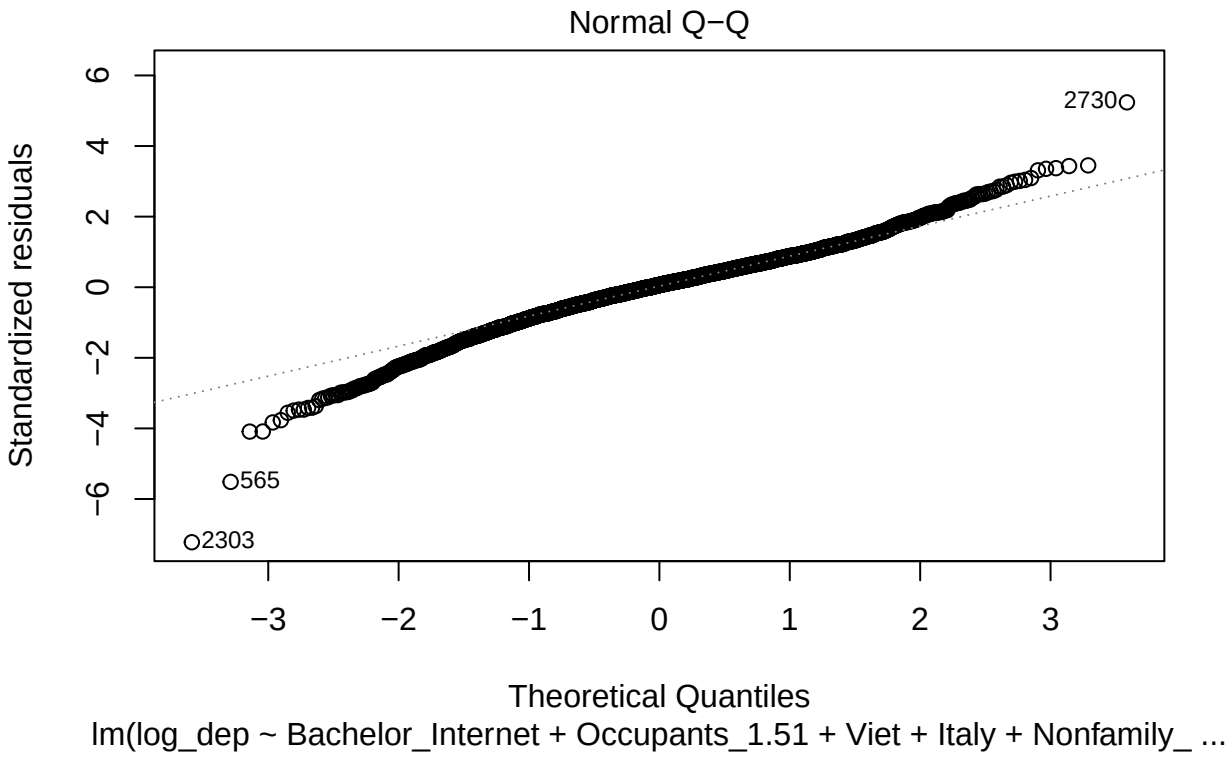


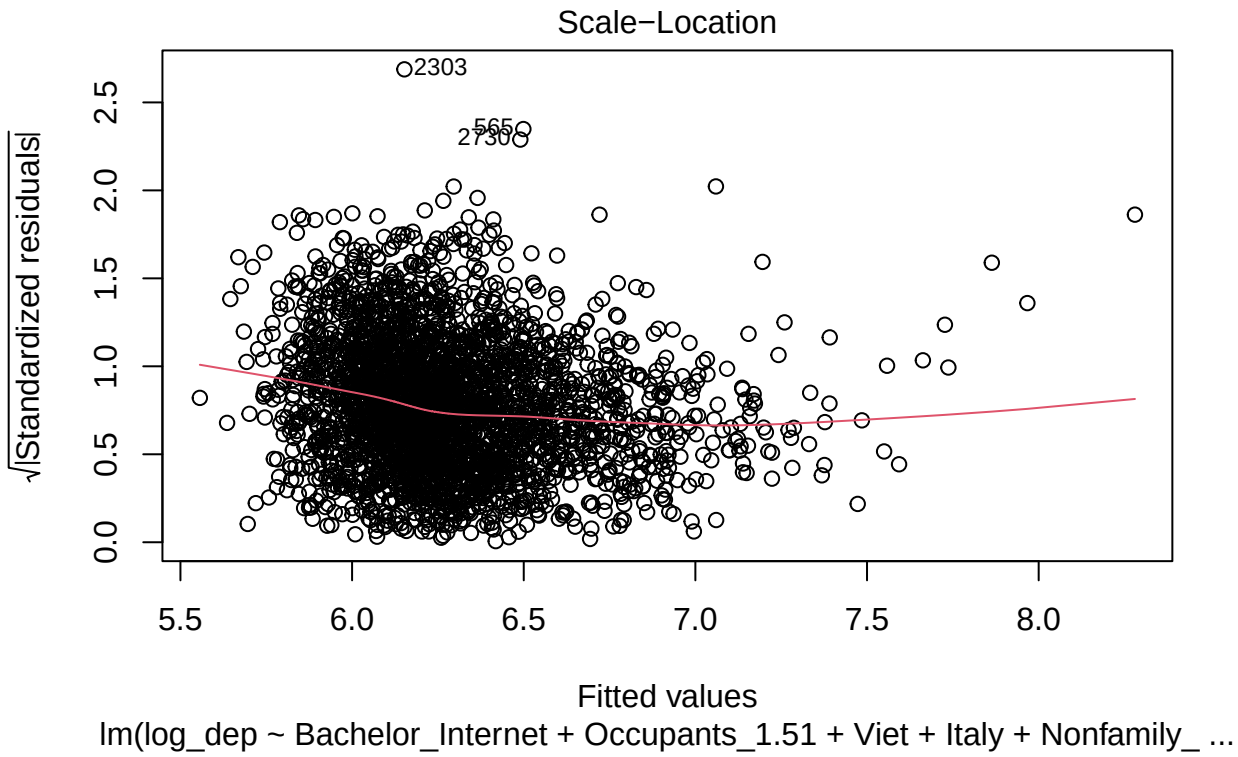

```
##  [1] "bachelor"           "Units"              "Occupants_1.51"
##  [4] "Percent_Gross_Rent" "Viet"               "Italy"
##  [7] "Internet"           "Nonfamily_Income"   "log_dep"
## [10] "Bachelor_Internet"  "Units_Internet"
```

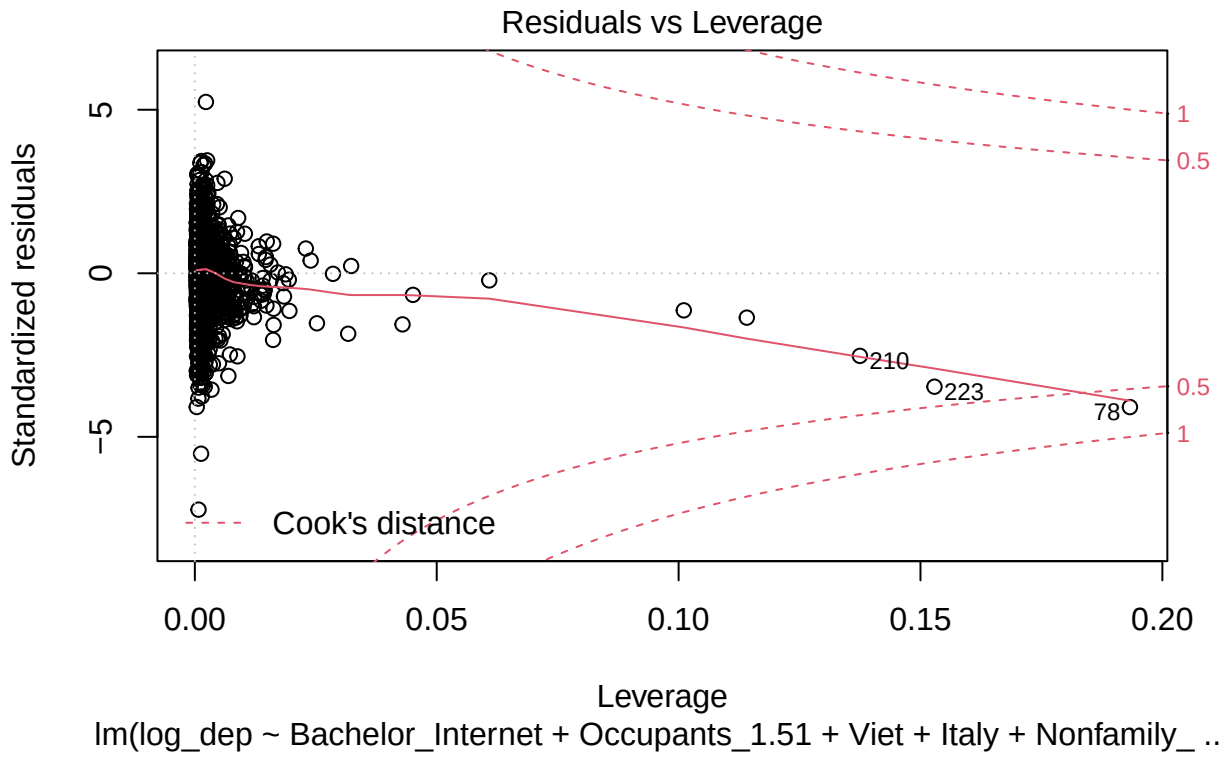
```
##
## Call:
## lm(formula = log_dep ~ Bachelor_Internet + Occupants_1.51 + Viet +
##     Italy + Nonfamily_Income, data = M2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.54714 -0.11609  0.01359  0.12937  1.12078
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.926e+00  2.610e-02 188.721   <2e-16 ***
## Bachelor_Internet 8.613e-03  3.871e-04  22.252   <2e-16 ***
## Occupants_1.51    4.933e-02  3.844e-03  12.831   <2e-16 ***
## Viet              1.224e-01  1.294e-02   9.458   <2e-16 ***
## Italy             1.364e-02  1.329e-03  10.266   <2e-16 ***
## Nonfamily_Income  9.866e-06  5.557e-07  17.753   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2142 on 2972 degrees of freedom
## Multiple R-squared:  0.6453, Adjusted R-squared:  0.6447
## F-statistic:  1082 on 5 and 2972 DF,  p-value: < 2.2e-16
```



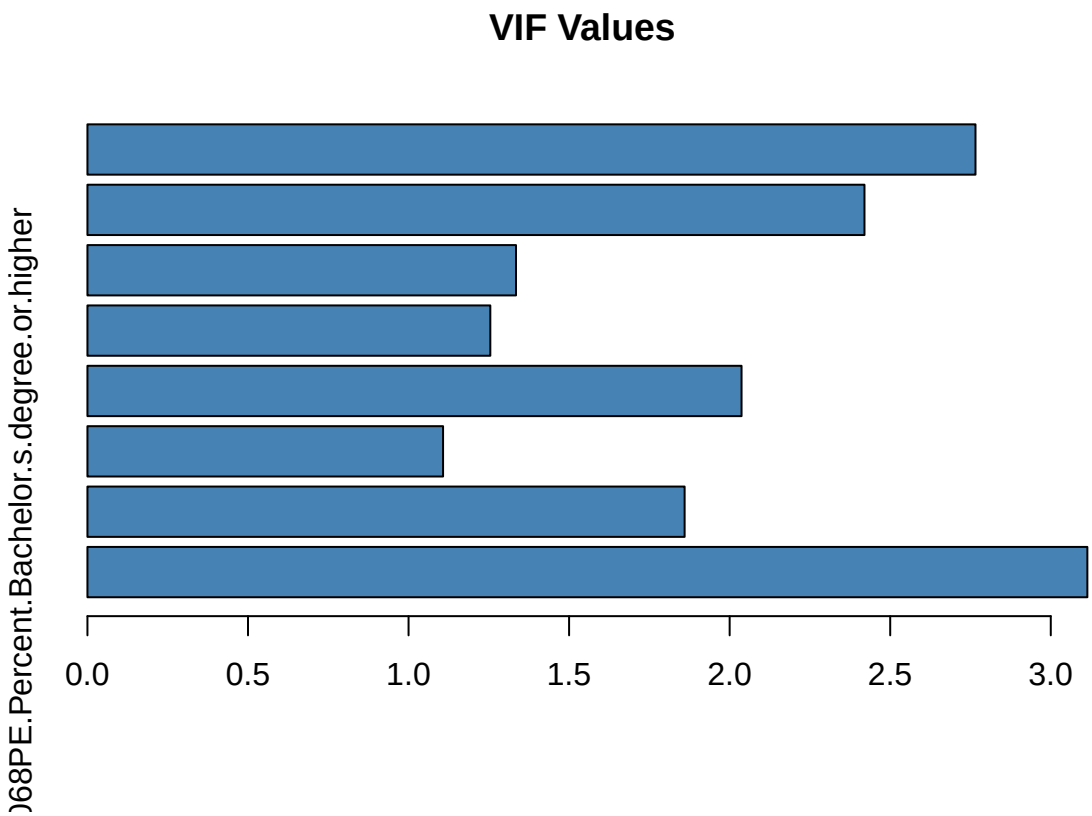

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log_dep ~ Bachelor_Internet + Occupants_1.51 + Viet + Italy + Nonfamily_ ...
2730
565
2303
6
4
2
0
−2
−4
−6
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
2303
565
27300
5.5
6.0
6.5
7.0
7.5
8.0
Fitted values
lm(log_dep ~ Bachelor_Internet + Occupants_1.51 + Viet + Italy + Nonfamily_ ...




```
## [[1]]
##
## Call:
## fitfunc(formula = as.formula(original$formula), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.40006 -0.09817  0.01384  0.10989  1.09323
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.239e+00  2.799e-02 187.163  < 2e-16 ***
## X1          3.660e-02  3.557e-03  10.291  < 2e-16 ***
## X2          8.033e-06  5.151e-07  15.594  < 2e-16 ***
## X3          4.481e-03  4.064e-04  11.027  < 2e-16 ***
## X4          1.502e-02  2.181e-03   6.884 7.05e-12 ***
## X5          9.368e-03  1.238e-03   7.565 5.15e-14 ***
## X6          8.324e-02  1.216e-02   6.845 9.26e-12 ***
## X7          9.562e-03  4.276e-04  22.362  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1954 on 2970 degrees of freedom
```

```
## Multiple R-squared:  0.7051, Adjusted R-squared:  0.7044
## F-statistic:  1014 on 7 and 2970 DF,  p-value: < 2.2e-16
##
##
## [[2]]
##
## Call:
## fitfunc(formula = as.formula(form), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.40119 -0.09661  0.01658  0.10885  1.10342
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.249e+00  2.807e-02 187.005  < 2e-16 ***
## X1          2.967e-02  4.003e-03   7.412 1.62e-13 ***
## X2          7.830e-06  5.169e-07  15.147  < 2e-16 ***
## X3          4.491e-03  4.055e-04  11.074  < 2e-16 ***
## X4          1.453e-02  2.181e-03   6.661 3.23e-11 ***
## X5          6.011e-03  1.526e-03   3.938 8.41e-05 ***
## X6          8.040e-02  1.216e-02   6.612 4.47e-11 ***
## X7          9.637e-03  4.272e-04  22.561  < 2e-16 ***
## X1:X5       6.606e-03  1.763e-03   3.746 0.000183 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.195 on 2969 degrees of freedom
## Multiple R-squared:  0.7065, Adjusted R-squared:  0.7057
## F-statistic: 893.2 on 8 and 2969 DF,  p-value: < 2.2e-16
```




```
##                       bachelor      Units Occupants_1.51 Percent_Gross_Rent
## bachelor            1.00000000 0.62997293    -0.02584906         0.60610036
## Units               0.62997293 1.00000000     0.09260814         0.48932224
## Occupants_1.51     -0.02584906 0.09260814     1.00000000         0.07709596
## Percent_Gross_Rent  0.60610036 0.48932224     0.07709596         1.00000000
## Viet                0.32717545 0.38867275     0.11356465         0.29261336
## Italy               0.42046839 0.19803284    -0.06486362         0.41789024
## Internet            0.67594047 0.45985354    -0.09194953         0.61133556
## Nonfamily_Income    0.72025676 0.47934071     0.09969836         0.61789595
## log_dep             0.68861650 0.55164488     0.15633292         0.72645195
## Bachelor_Internet   0.92405512 0.59978074    -0.06251723         0.66466487
## Units_Internet      0.66703703 0.99490985     0.08292486         0.51474080
##                           Viet       Italy    Internet Nonfamily_Income
## bachelor            0.32717545  0.42046839  0.67594047       0.72025676
## Units               0.38867275  0.19803284  0.45985354       0.47934071
## Occupants_1.51      0.11356465 -0.06486362 -0.09194953       0.09969836
## Percent_Gross_Rent  0.29261336  0.41789024  0.61133556       0.61789595
## Viet                1.00000000  0.09248492  0.26205850       0.36951577
## Italy               0.09248492  1.00000000  0.40277301       0.39184376
## Internet            0.26205850  0.40277301  1.00000000       0.68226345
## Nonfamily_Income    0.36951577  0.39184376  0.68226345       1.00000000
## log_dep             0.38760453  0.43250591  0.64397943       0.73061312
## Bachelor_Internet   0.32354681  0.45003378  0.90631458       0.76691756
```

```
## Units_Internet     0.40512693  0.22092502  0.50776467       0.52043914
##                       log_dep Bachelor_Internet Units_Internet
## bachelor           0.6886165        0.92405512     0.66703703
## Units              0.5516449        0.59978074     0.99490985
## Occupants_1.51     0.1563329       -0.06251723     0.08292486
## Percent_Gross_Rent 0.7264519        0.66466487     0.51474080
## Viet               0.3876045        0.32354681     0.40512693
## Italy              0.4325059        0.45003378     0.22092502
## Internet           0.6439794        0.90631458     0.50776467
## Nonfamily_Income   0.7306131        0.76691756     0.52043914
## log_dep            1.0000000        0.72891570     0.58052583
## Bachelor_Internet  0.7289157        1.00000000     0.64588646
## Units_Internet     0.5805258        0.64588646     1.00000000

##
## Call:
## lm(formula = log_depvar ~ DP02_0068PE.Percent.Bachelor.s.degree.or.higher +
##     DP04_0012PE.Percent.10.to.19.units + DP04_0013PE.Percent.20.or.more.units +
##     DP04_0040PE.Percent.1.bedroom + DP04_0079PE.Percent.1.51.or.more +
##     DP04_0129PE.Percent..1.000.to..1.499 + DP04_0130PE.Percent..1.500.to..1.999 +
##     DP04_0131PE.Percent..2.000.to..2.499 + DP04_0133PE.Percent..3.000.or.more +
##     DP05_0050PE.Percent.Vietnamese + DP02_0136PE.Percent.Italian +
##     DP02_0153PE.Percent.With.a.broadband.Internet.subscription +
##     DP03_0091E.Estimate.Mean.nonfamily.income..dollars., data = numeric_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.37905 -0.08861  0.01609  0.10574  1.12519
##
## Coefficients:
##                                                                  Estimate Std. Error
## (Intercept)                                                     5.334e+00  4.043e-02
## DP02_0068PE.Percent.Bachelor.s.degree.or.higher                 2.577e-03  6.656e-04
## DP04_0012PE.Percent.10.to.19.units                              9.786e-03  2.431e-03
## DP04_0013PE.Percent.20.or.more.units                            3.168e-03  1.321e-03
## DP04_0040PE.Percent.1.bedroom                                   3.552e-03  1.272e-03
## DP04_0079PE.Percent.1.51.or.more                                2.501e-02  3.646e-03
## DP04_0129PE.Percent..1.000.to..1.499                            1.026e-02  4.865e-04
## DP04_0130PE.Percent..1.500.to..1.999                            3.385e-03  1.237e-03
## DP04_0131PE.Percent..2.000.to..2.499                            1.098e-02  2.675e-03
## DP04_0133PE.Percent..3.000.or.more                              1.642e-02  3.498e-03
## DP05_0050PE.Percent.Vietnamese                                  4.137e-02  1.235e-02
## DP02_0136PE.Percent.Italian                                     6.057e-03  1.253e-03
## DP02_0153PE.Percent.With.a.broadband.Internet.subscription      5.270e-03  6.312e-04
## DP03_0091E.Estimate.Mean.nonfamily.income..dollars.             4.129e-06  5.939e-07
##                                                                 t value Pr(>|t|)
## (Intercept)                                                     131.925  < 2e-16 ***
## DP02_0068PE.Percent.Bachelor.s.degree.or.higher                   3.872 0.000110 ***
```

```
## DP04_0012PE.Percent.10.to.19.units                                4.025 5.84e-05 ***
## DP04_0013PE.Percent.20.or.more.units                              2.398 0.016536 *
## DP04_0040PE.Percent.1.bedroom                                     2.792 0.005269 **
## DP04_0079PE.Percent.1.51.or.more                                  6.860 8.37e-12 ***
## DP04_0129PE.Percent..1.000.to..1.499                             21.099  < 2e-16 ***
## DP04_0130PE.Percent..1.500.to..1.999                              2.736 0.006262 **
## DP04_0131PE.Percent..2.000.to..2.499                              4.103 4.19e-05 ***
## DP04_0133PE.Percent..3.000.or.more                                4.694 2.80e-06 ***
## DP05_0050PE.Percent.Vietnamese                                    3.351 0.000816 ***
## DP02_0136PE.Percent.Italian                                       4.836 1.39e-06 ***
## DP02_0153PE.Percent.With.a.broadband.Internet.subscription        8.349  < 2e-16 ***
## DP03_0091E.Estimate.Mean.nonfamily.income..dollars.               6.953 4.38e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1905 on 2964 degrees of freedom
## Multiple R-squared:  0.7203, Adjusted R-squared:  0.7191
## F-statistic: 587.2 on 13 and 2964 DF,  p-value: < 2.2e-16
```