Gurpinder Singh

STAT 108

12/9/2022

The research questions is to predict the price of California homes given the quality of the associated school district.

Load all the followin library

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## Loading required package: carData
##
##
## Attaching package: 'car'
##
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
##
## The following object is masked from 'package:purrr':
##
##     some
```

The following data was collected from https://www.cde.ca.gov/ which provides accurate data about california Schools. The following variables: Absentness/Reason for absent, chronic Absentee, stability of student, suspension count, dropout rate, nation test results. Additionally the information for housing data was obtained from zillow. The data is of home sales which have occured inside of California inside the year 2021. This entire data was comiled into one using python script under the data section. All individual data can be refrenced directly.

The response variable is housing price.

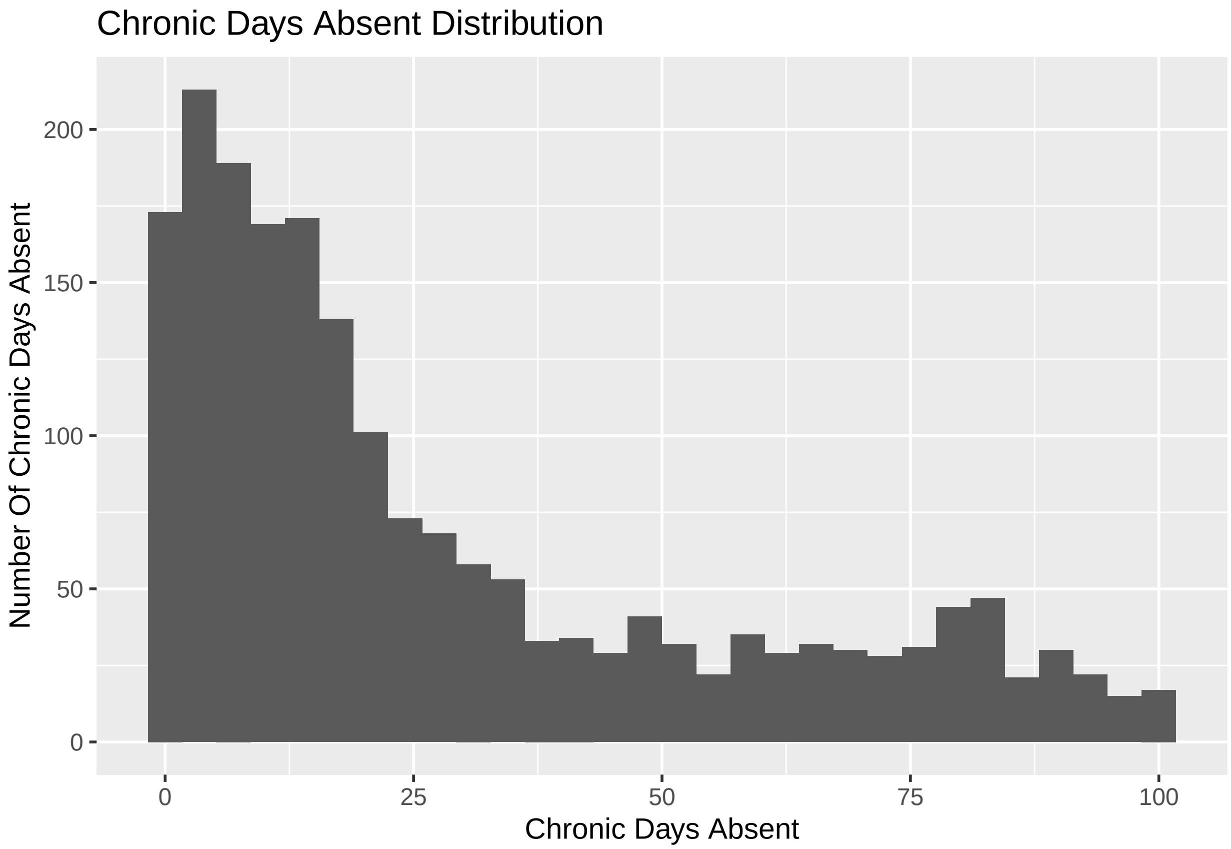
```
## Rows: 1978 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): School Name
## dbl (11): County Code, District Code, School Code, Zip Code, Average Days Ab...
```

```
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
## Rows: 1,978
## Columns: 12
## $ ‘County Code‘                   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ ‘District Code‘                 <dbl> 10017, 10017, 10017, 31617, 61119, 6111~
## $ ‘School Code‘                   <dbl> 112607, 130419, 136101, 131763, 106401,~
## $ ‘School Name‘                   <chr> "Envision Academy for Arts & Technology~
## $ ‘Zip Code‘                      <dbl> 94612, 94544, 94587, 94538, 94501, 9450~
## $ ‘Average Days Absent‘           <dbl> 21.775000, 55.054545, 9.740000, 11.1388~
## $ ChronicAbsenteeismRate          <dbl> 36.3571429, 83.7111111, 0.9105263, 11.4~
## $ ‘Dropout (Rate)‘                <dbl> 8.257143, 62.075000, 12.050000, 2.51428~
## $ ‘Non-Stability Rate (percent)‘ <dbl> 10.3866667, 52.1888889, 7.5052632, 7.17~
## $ ‘Suspension Rate (Total)‘       <dbl> 0.00000000, 0.00000000, 0.00000000, 0.0~
## $ ‘Mean Scale Score‘              <dbl> 2451.167, 2472.000, 2570.371, 2401.963,~
## $ average2021                     <dbl> 718260.8, 793237.2, 1145040.2, 1168105.~
```

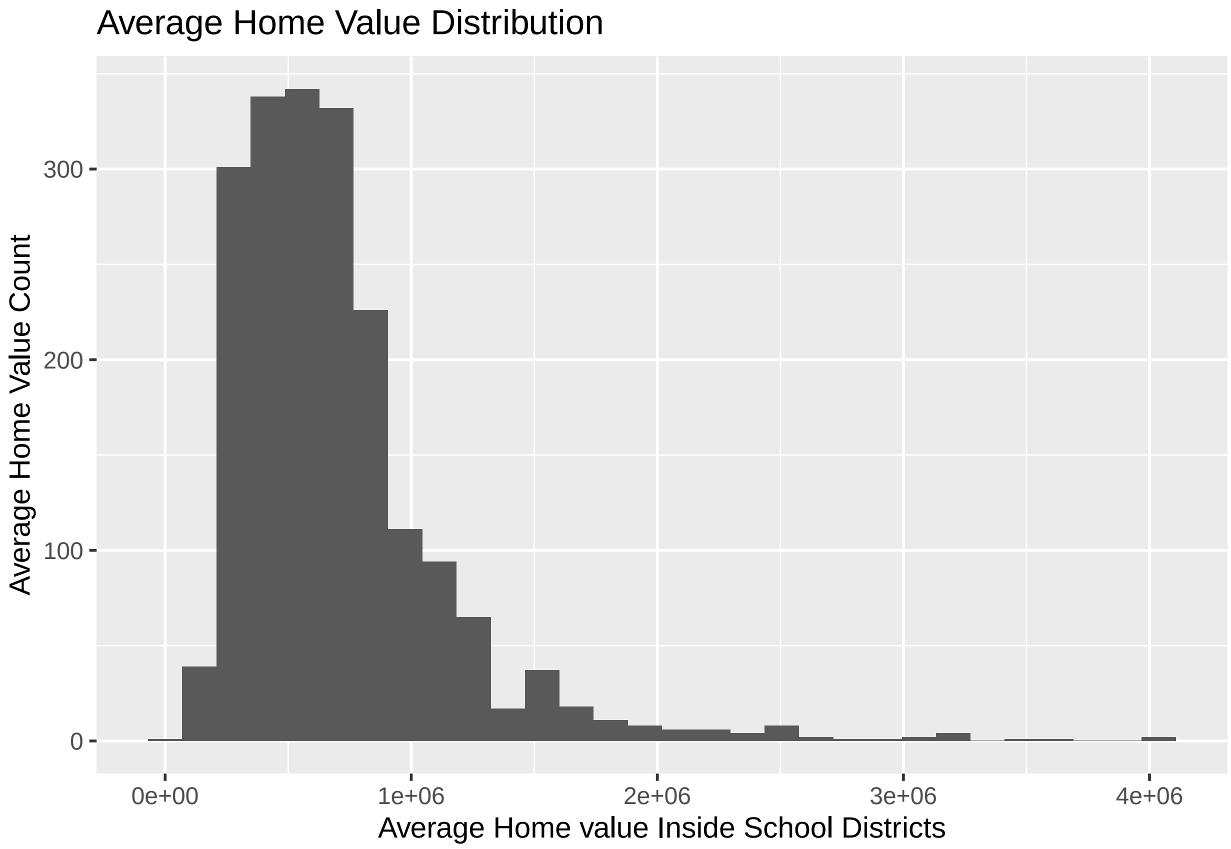
The following are some plots which I believed to be most interesting to showcase the relation of the variables with the response variable Housing price. All of the values were plotted but only a few were shown to preserve space.

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

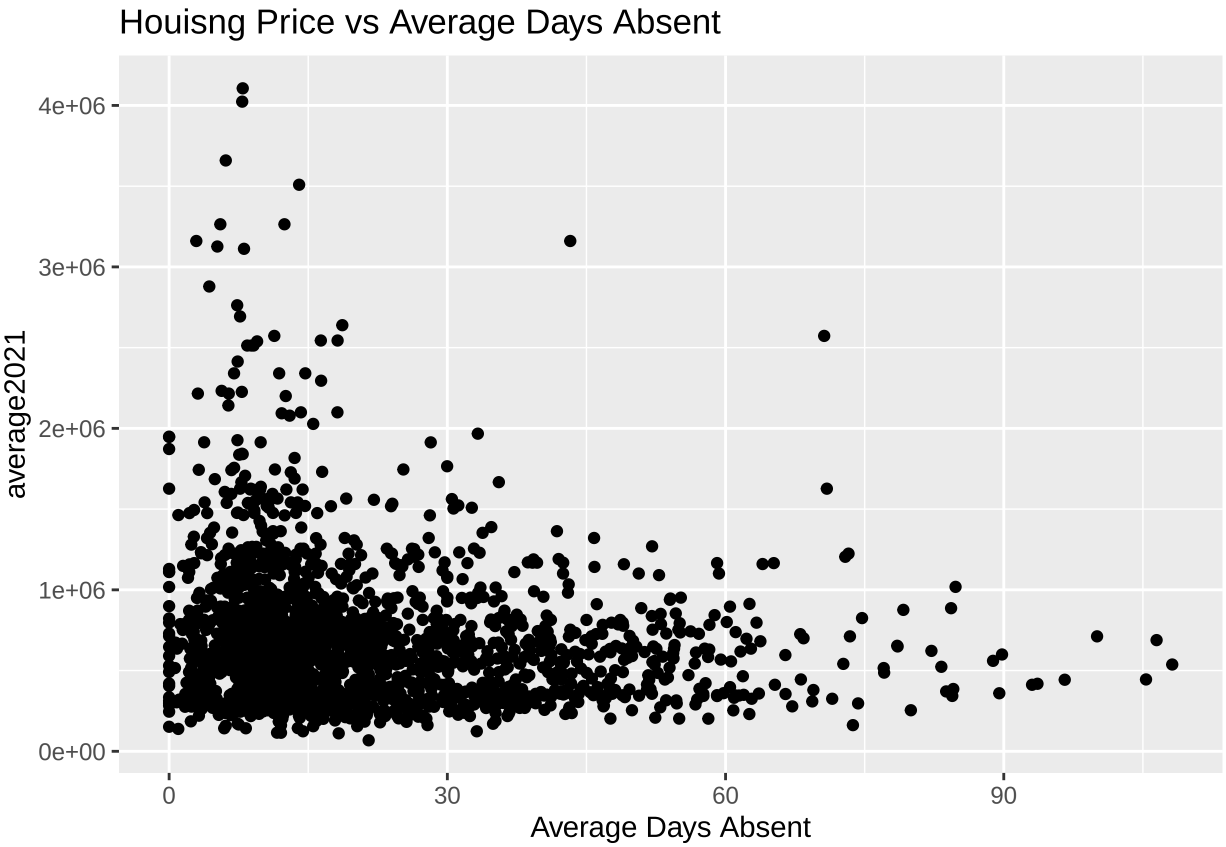


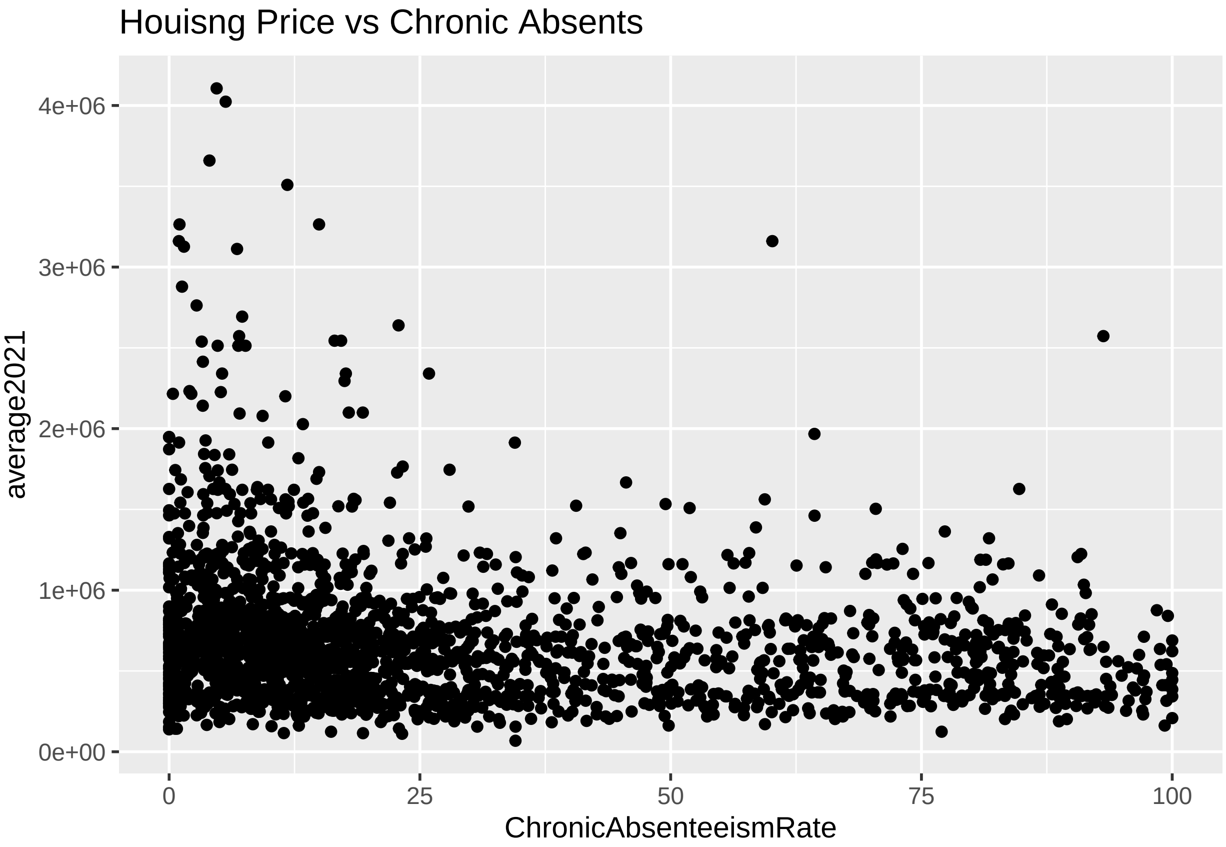
```
## # A tibble: 1 x 5
##     max   min  mean   med    sd
##   <dbl> <dbl> <dbl> <dbl> <dbl>
## 1   100     0  28.4  17.4  27.6
```

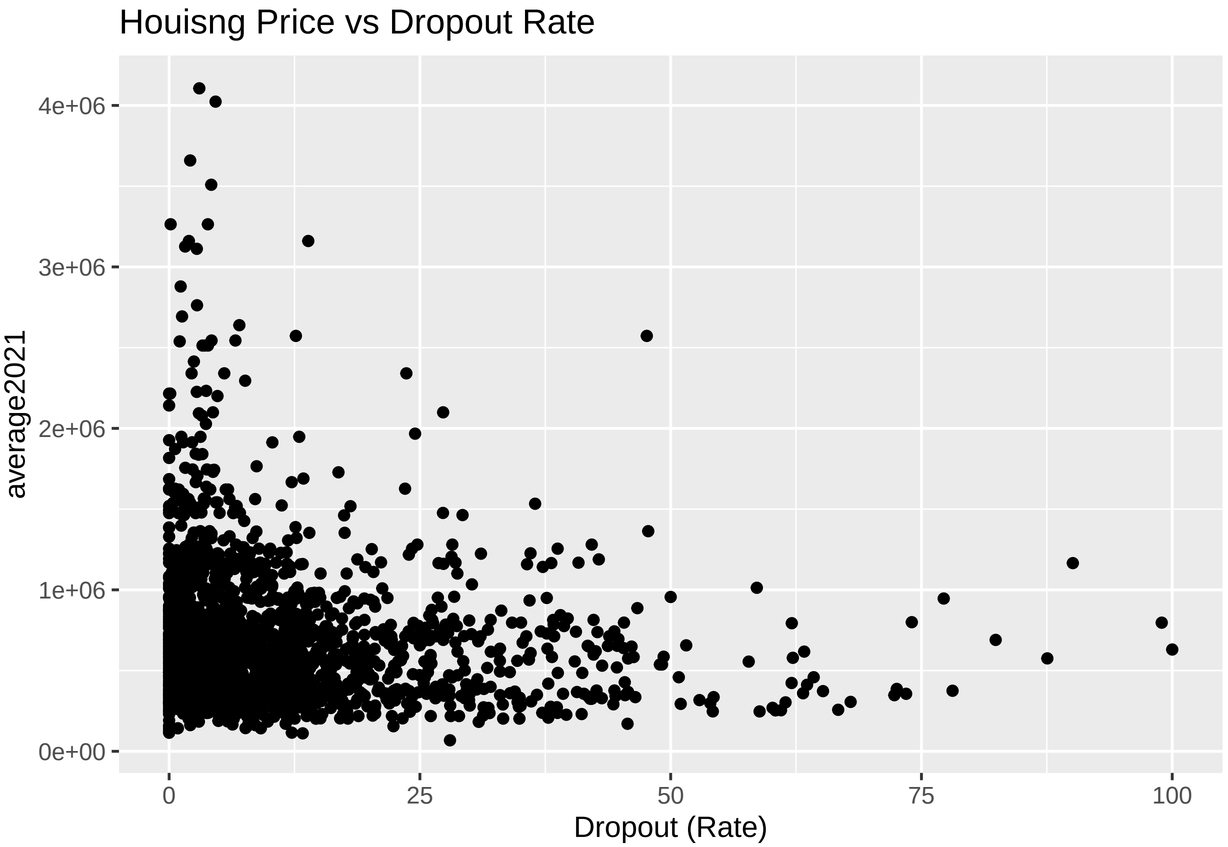
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

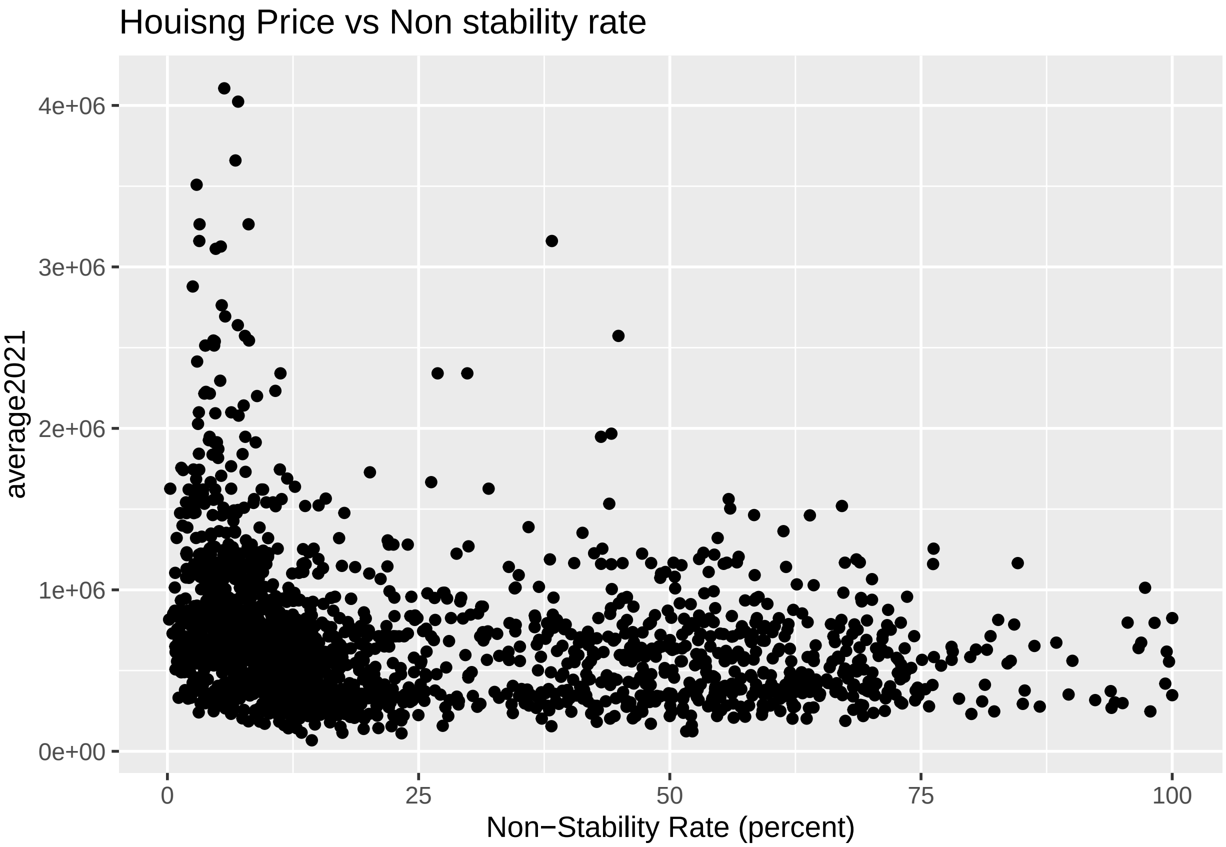


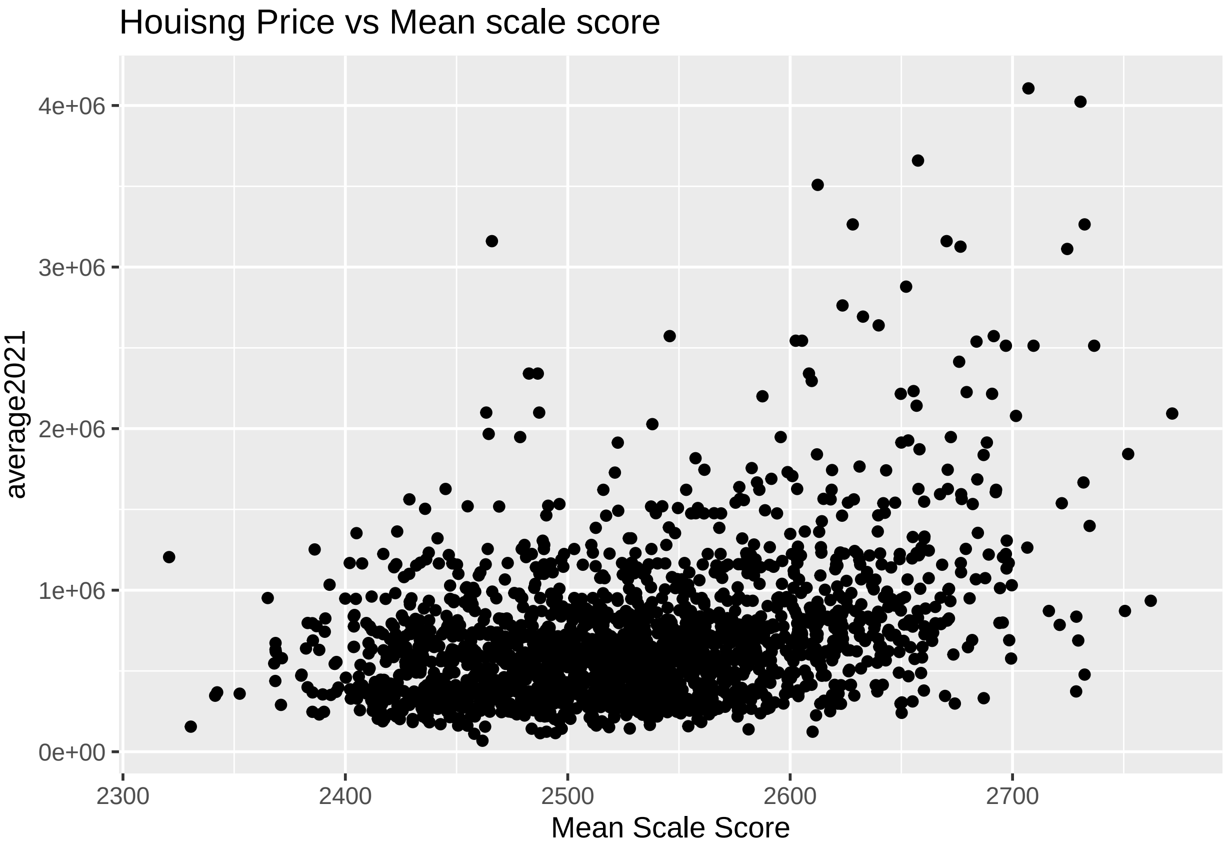
```
## # A tibble: 1 x 5
##        max    min    mean     med      sd
##      <dbl>  <dbl>   <dbl>   <dbl>   <dbl>
## 1 4105966. 68073. 697944. 616592. 437134.
```


Houisng Price vs Average Days Absent
average2021
4e+06
3e+06
2e+06
1e+06
0e+00
0
30
60
90
Average Days Absent


Houisng Price vs Chronic Absents
average2021
4e+06
3e+06
2e+06
1e+06
0e+00
0
25
50
75
100
ChronicAbsenteeismRate


Houisng Price vs Dropout Rate
average2021
4e+06
3e+06
2e+06
1e+06
0e+00
0
25
50
75
100
Dropout (Rate)


Houisng Price vs Non stability rate
average2021
4e+06
3e+06
2e+06
1e+06
0e+00
0
25
50
75
100
Non−Stability Rate (percent)



During the folloowing section I scale the value of the houisng price by taking the log of the value.

In the following place we create three models. The first model is all of the values. The next one is without non stability rate. The next one is suspension rate.

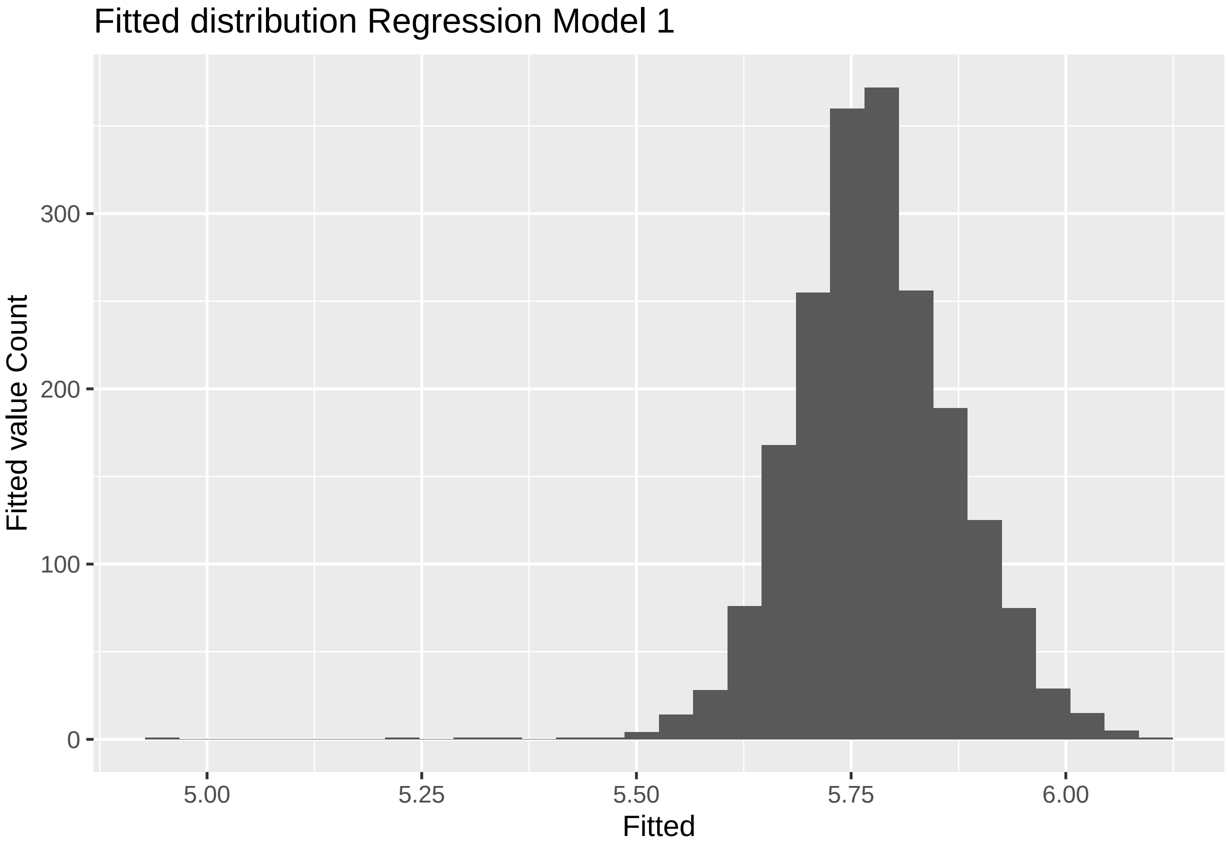
term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	5.777	0.005	1180.298	0.000	5.768	5.787
`Average Days Absent`	0.039	0.012	3.268	0.001	0.016	0.063
ChronicAbsenteeismRate	-0.026	0.014	-1.914	0.056	-0.053	0.001
`Dropout (Rate)`	0.015	0.006	2.408	0.016	0.003	0.028
`Non-Stability Rate (percent)`	-0.001	0.008	-0.148	0.882	-0.017	0.015
`Suspension Rate (Total)`	-0.027	0.005	-5.366	0.000	-0.037	-0.017
`Mean Scale Score`	0.098	0.007	14.522	0.000	0.085	0.111

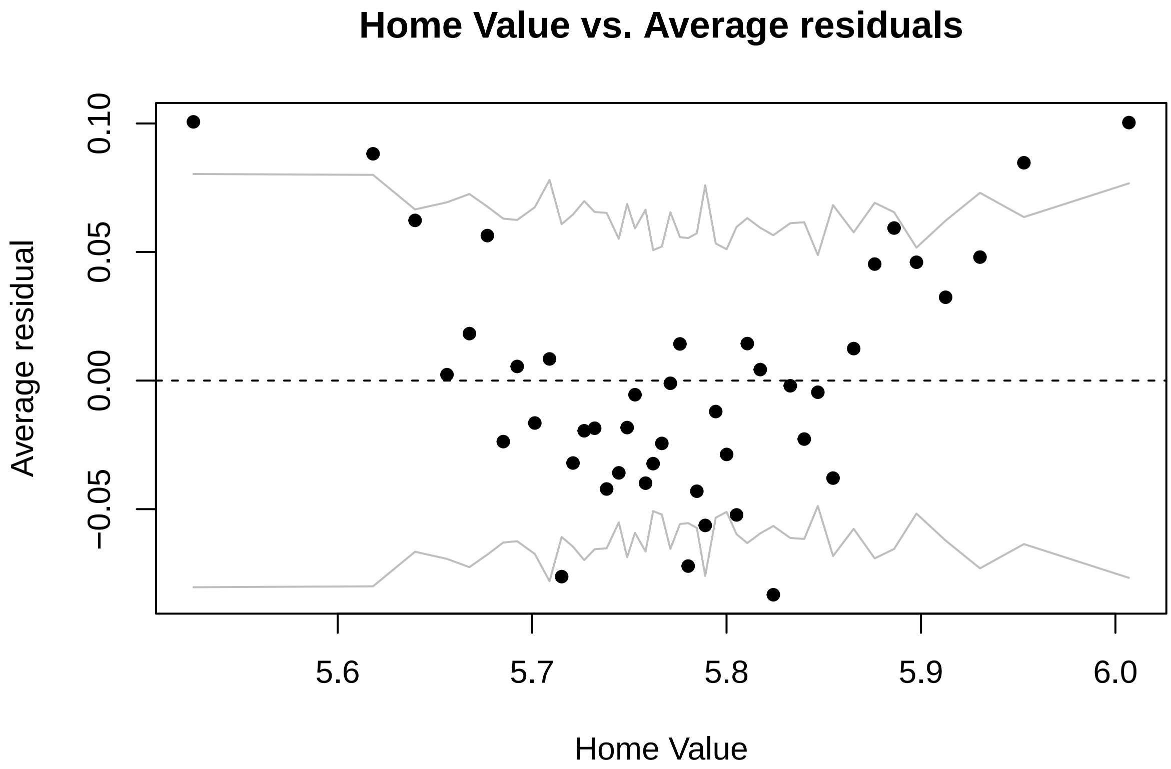
term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	5.777	0.005	1180.591	0.000	5.768	5.787
`Average Days Absent`	0.040	0.012	3.425	0.001	0.017	0.062
ChronicAbsenteeismRate	-0.027	0.012	-2.214	0.027	-0.051	-0.003
`Dropout (Rate)`	0.015	0.006	2.521	0.012	0.003	0.027
`Suspension Rate (Total)`	-0.027	0.005	-5.389	0.000	-0.037	-0.017
`Mean Scale Score`	0.098	0.006	15.277	0.000	0.086	0.111

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	5.777	0.005	1172.066	0.000	5.768	5.787
`Average Days Absent`	0.049	0.012	4.102	0.000	0.025	0.072
ChronicAbsenteeismRate	-0.036	0.014	-2.642	0.008	-0.062	-0.009
`Dropout (Rate)`	0.017	0.006	2.569	0.010	0.004	0.029
`Non-Stability Rate (percent)`	-0.004	0.008	-0.498	0.619	-0.020	0.012
`Mean Scale Score`	0.098	0.007	14.461	0.000	0.085	0.112

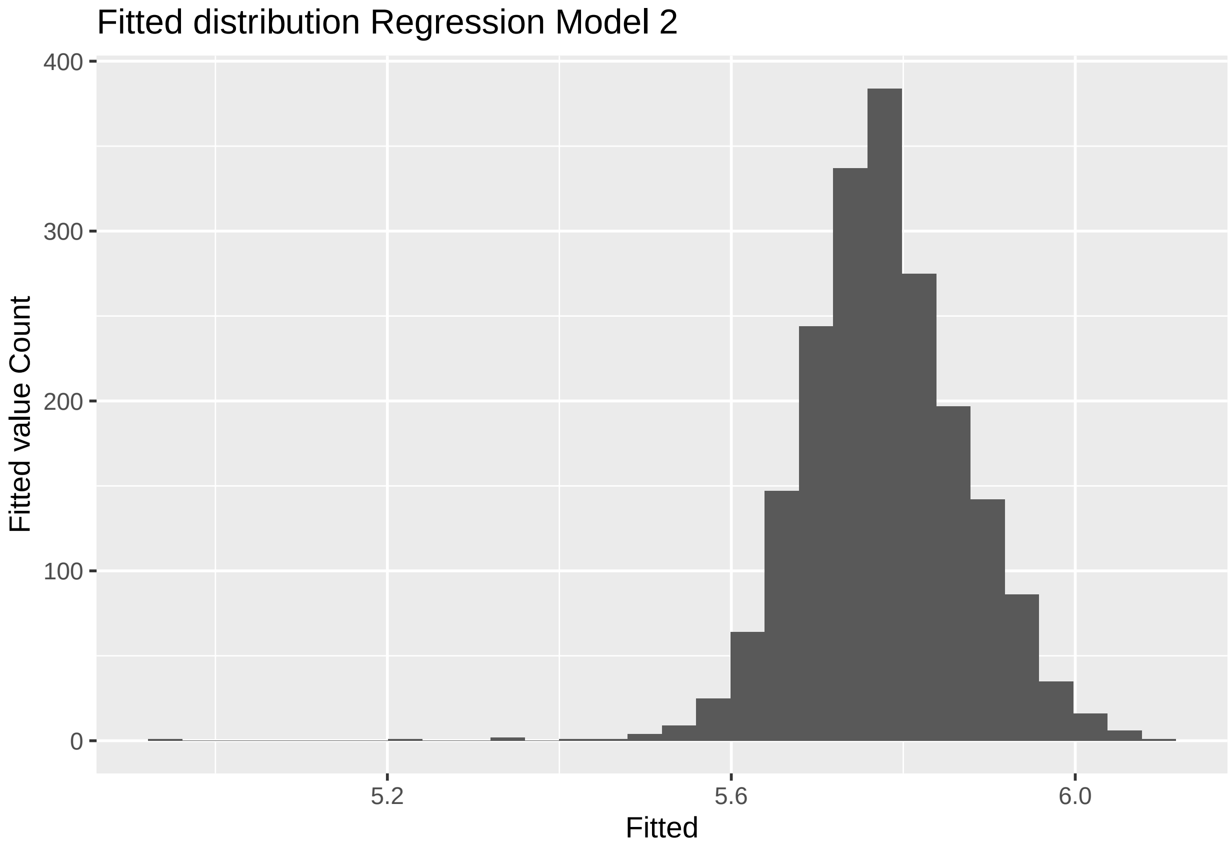
The following is to check Normality and linearity. As you can see all but regression model 1 accepts linearity condition and All follow Normality condition. Independence is already accepted by how we collected the data.

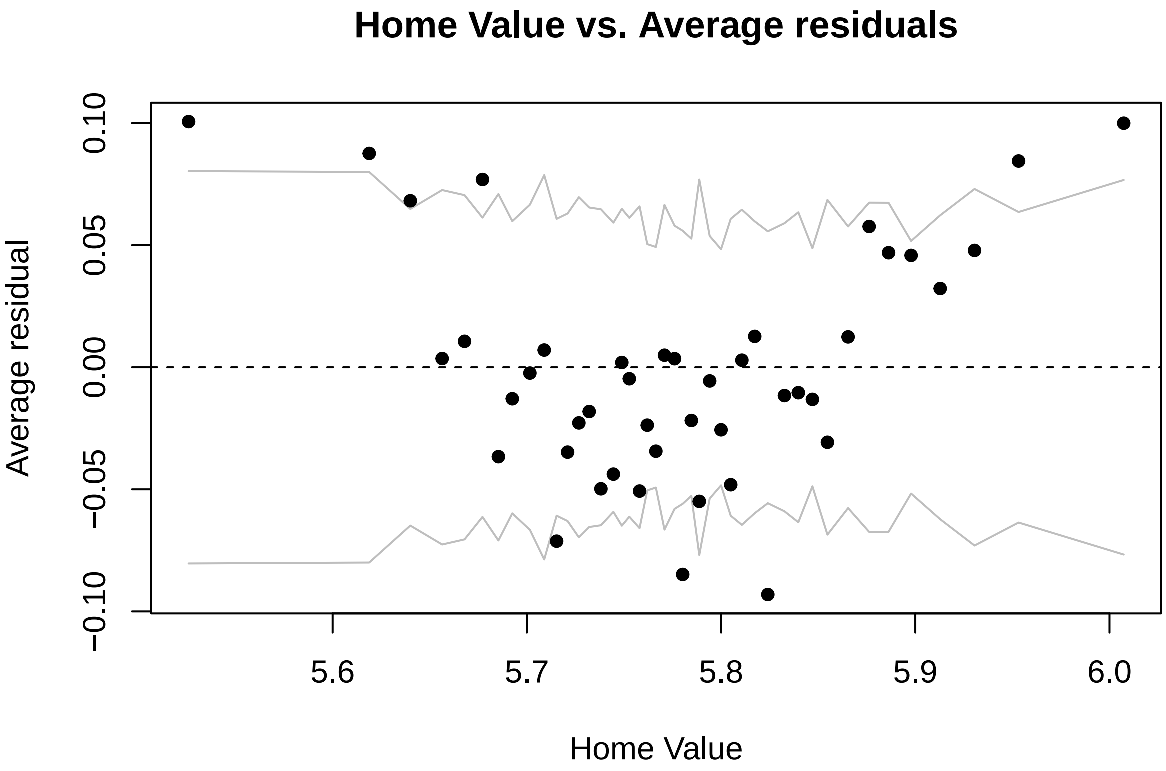
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



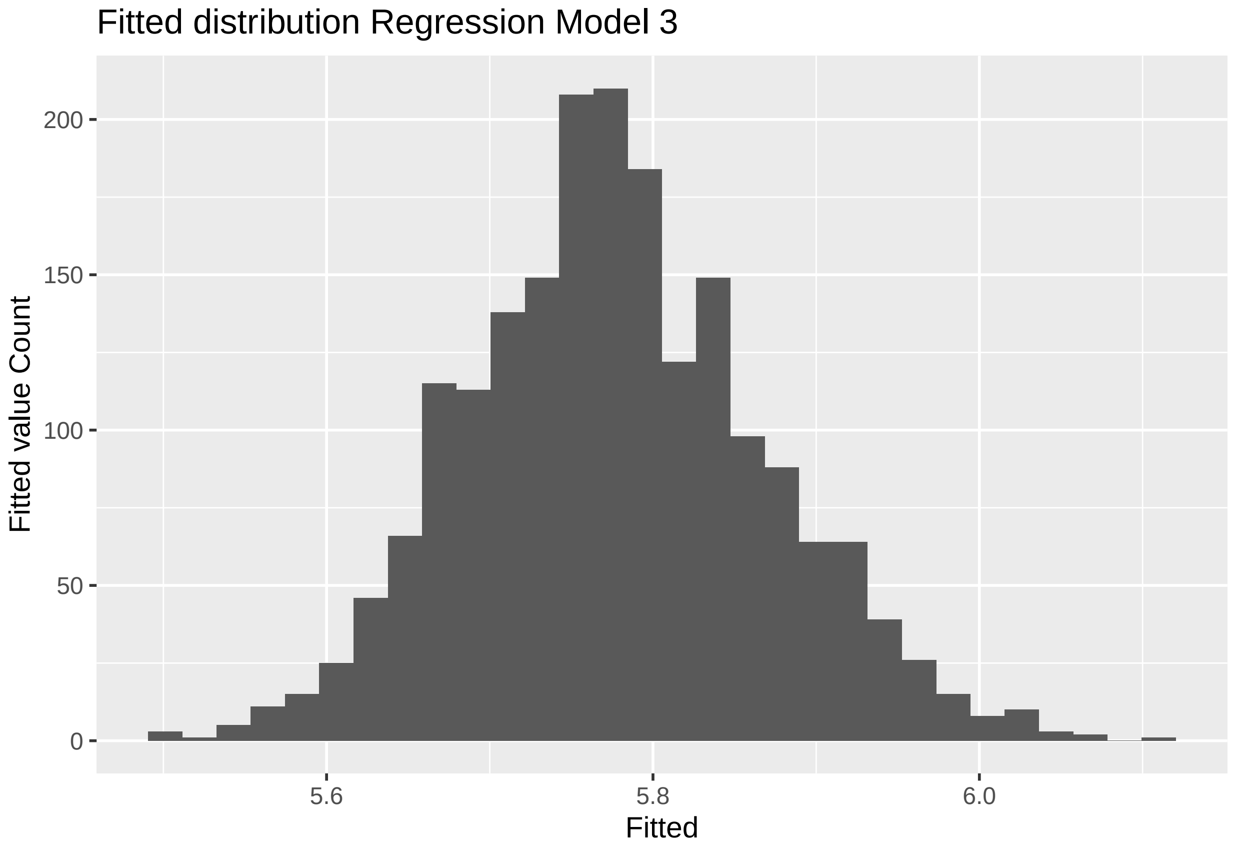


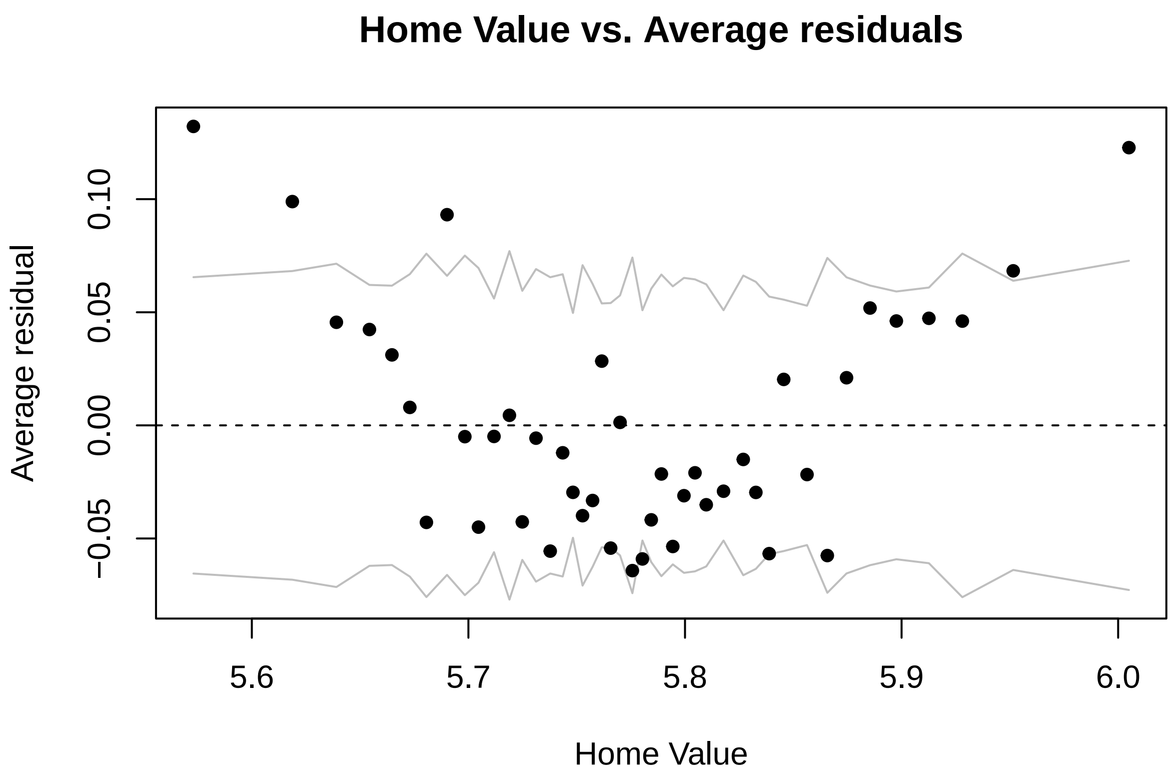
```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```


Fitted distribution Regression Model 2
Fitted value Count
400
300
200
100
0
5.2
5.6
6.0
Fitted



```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


Fitted distribution Regression Model 3
Fitted value Count
0
50
100
150
200
5.6
5.8
6.0
Fitted



Regression model 2 gives the following equation: 5.7710^x0+.04(x1)+ -.027x2+.015$x3$-$.027$x4+.098*x5. I preformed a couple of tests and say the summary statsitics for the model such as the r square as well as these showcase that values between 1 to 5 are somewhat corelated. Values 5+ are highly and 0 to 1 are not that correlated.

```
## Analysis of Variance Table
##
## Model 1: average2021 ~ ‘Average Days Absent‘ + ChronicAbsenteeismRate +
##     ‘Dropout (Rate)‘ + ‘Suspension Rate (Total)‘ + ‘Mean Scale Score‘
## Model 2: average2021 ~ ‘Average Days Absent‘ + ChronicAbsenteeismRate +
##     ‘Dropout (Rate)‘ + ‘Non-Stability Rate (percent)‘ + ‘Suspension Rate (Total)‘ +
##     ‘Mean Scale Score‘
##   Res.Df    RSS Df Sum of Sq Pr(>Chi)
## 1   1972 93.408
## 2   1971 93.407  1 0.0010422   0.8821
```

```
## Analysis of Variance Table
##
## Model 1: average2021 ~ ‘Average Days Absent‘ + ChronicAbsenteeismRate +
##     ‘Dropout (Rate)‘ + ‘Non-Stability Rate (percent)‘ + ‘Mean Scale Score‘
## Model 2: average2021 ~ ‘Average Days Absent‘ + ChronicAbsenteeismRate +
##     ‘Dropout (Rate)‘ + ‘Non-Stability Rate (percent)‘ + ‘Suspension Rate (Total)‘ +
##     ‘Mean Scale Score‘
##   Res.Df    RSS Df Sum of Sq  Pr(>Chi)
## 1   1972 94.772
## 2   1971 93.407  1    1.3647 8.037e-08 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
## 
## Call:
## lm(formula = average2021 ~ ‘Average Days Absent‘ + ChronicAbsenteeismRate + 
##     ‘Dropout (Rate)‘ + ‘Suspension Rate (Total)‘ + ‘Mean Scale Score‘, 
##     data = data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.78714 -0.15120  0.00108  0.13974  0.95356 
## 
## Coefficients:
##                            Estimate Std. Error  t value Pr(>|t|)    
## (Intercept)                5.777310   0.004894 1180.591  < 2e-16 ***
## ‘Average Days Absent‘      0.039549   0.011546    3.425 0.000626 ***
## ChronicAbsenteeismRate    -0.026874   0.012141   -2.214 0.026973 *  
## ‘Dropout (Rate)‘           0.015139   0.006005    2.521 0.011774 *  
## ‘Suspension Rate (Total)‘ -0.026927   0.004997   -5.389 7.94e-08 ***
## ‘Mean Scale Score‘         0.098471   0.006446   15.277  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 0.2176 on 1972 degrees of freedom
## Multiple R-squared:  0.1594, Adjusted R-squared:  0.1572 
## F-statistic: 74.76 on 5 and 1972 DF,  p-value: < 2.2e-16
```

```
##     ‘Average Days Absent‘    ChronicAbsenteeismRate          ‘Dropout (Rate)‘ 
##                  5.563722                  6.151946                  1.504836 
## ‘Suspension Rate (Total)‘        ‘Mean Scale Score‘ 
##                  1.042037                  1.734102 
```

Doing a Chi square test on the Data where Null hypothesis is that the variable removed is not a predictor and alternative hypothesis is that it is a predictor we see that the chi square test we see that we accept the null hypothesis and remove it from the test. Meaning the the second model is best.

```
## [1] "average residual for the model is"
```

```
## [1] -3.356257e-17
```

There are a couple of limitations which should be noticed. One is that there are only no enough observations to cover every school district inside of california. The could make the model that we use not applicable to the entirity of California. However a a large portion of california is covered. Additionaly, many times it can be assumed that schools arent represented are ones in lower income brackets. This could cause a bit of bias onto the model itself. I additionally would also like to add more variables which are applicable to the study I believe that could greatly improve the accuracy of the model.