VIA 502E Data Mining Homework

Q1- Take the ames data from tidymodels package

```
#Q1

data(ames)

ames_split <- initial_split(ames, prop = 0.8, strata = Sale_Price)

ames_train <- training(ames_split)
ames_test <- testing(ames_split)
```

- *Using prop and strata commands, ames_train and ames_test are splitted according to distribution of data points in Sale_Price.*

Data	
ames	2930 obs. of 74 variables
ames_split	Large mc_split (4 elements, 1 MB)
ames_test	584 obs. of 74 variables
ames_train	2346 obs. of 74 variables

Q2- Set `Sale_Price` column as output and following features as input variables:

- **MS_SubClass MS_Zoning Lot_Frontage Lot_Area Street Alley Lot_Shape Land_Contour Utilities Lot_Config**

Q3- Fit a linear model using all the input variables listed above

```
# Q2 and Q3

linear_model <- lm(Sale_Price ~ MS_SubClass + MS_Zoning + Lot_Frontage + Lot_Area + Street+ Alley + Lot_Shape +
                   Land_Contour + Utilities + Lot_Config,data = ames_train)
summary(linear_model)
```

- *lm command is used for constructing a linear model in R. summary() fucntion made it possible to observe the model statistics.*

```
Residuals:
    Min      1Q  Median      3Q     Max 
-263376  -37298  -10035   24152  469195 

Coefficients:
                                                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                         7.216e+04  2.512e+04   2.872  0.00411 ** 
MS_SubClassOne_Story_1945_and_Older                -5.504e+04  6.633e+03  -8.298  < 2e-16 ***
MS_SubClassOne_Story_with_Finished_Attic_All_Ages  -1.902e+02  2.819e+04  -0.007  0.99462    
MS_SubClassOne_and_Half_Story_Unfinished_All_Ages  -4.392e+04  1.595e+04  -2.754  0.00594 ** 
MS_SubClassOne_and_Half_Story_Finished_All_Ages    -2.645e+04  5.023e+03  -5.267 1.52e-07 ***
MS_SubClassTwo_Story_1946_and_Newer                 5.098e+04  3.675e+03  13.871  < 2e-16 ***
MS_SubClassTwo_Story_1945_and_Older                -2.862e+03  7.266e+03  -0.394  0.69363    
MS_SubClassTwo_and_Half_Story_All_Ages              3.415e+04  1.520e+04   2.246  0.02477 *  
MS_SubClassSplit_or_Multilevel                     -1.340e+04  6.996e+03  -1.916  0.05551 .  
MS_SubClassSplit_Foyer                             -2.763e+04  1.035e+04  -2.671  0.00762 ** 
MS_SubClassDuplex_All_Styles_and_Ages              -3.302e+04  7.114e+03  -4.642 3.64e-06 ***
MS_SubClassOne_Story_PUD_1946_and_Newer             4.486e+04  5.908e+03   7.593 4.51e-14 ***
MS_SubClassOne_and_Half_Story_PUD_All_Ages         -2.612e+04  6.265e+04  -0.417  0.67674    
MS_SubClassTwo_Story_PUD_1946_and_Newer            -3.806e+03  7.524e+03  -0.506  0.61301    
MS_SubClassPUD_Multilevel_Split_Level_Foyer        -1.419e+04  1.803e+04  -0.787  0.43142    
MS_SubClassTwo_Family_conversion_All_Styles_and_Ages -4.439e+04  9.688e+03  -4.582 4.85e-06 ***
MS_ZoningResidential_High_Density                  -4.741e+04  1.628e+04  -2.913  0.00362 ** 
MS_ZoningResidential_Low_Density                   -3.237e+04  7.314e+03  -4.426 1.00e-05 ***
MS_ZoningResidential_Medium_Density                -4.643e+04  8.062e+03  -5.759 9.56e-09 ***
MS_ZoningA_agr                                     -2.038e+05  4.485e+04  -4.543 5.83e-06 ***
MS_ZoningC_all                                     -7.999e+04  1.720e+04  -4.650 3.50e-06 ***
MS_ZoningI_all                                     -1.260e+04  9.204e+04  -0.137  0.89109    
Lot_Frontage                                        5.071e+02  4.297e+01  11.803  < 2e-16 ***
Lot_Area                                            2.056e+00  2.058e-01   9.991  < 2e-16 ***
StreetPave                                          5.965e+04  2.178e+04   2.739  0.00621 ** 
AlleyNo_Alley_Access                                7.615e+03  7.068e+03   1.077  0.28144    
AlleyPaved                                          1.093e+04  1.137e+04   0.961  0.33662    
Lot_ShapeSlightly_Irregular                         2.083e+04  3.205e+03   6.500 9.81e-11 ***
Lot_ShapeModerately_Irregular                       1.788e+04  8.865e+03   2.017  0.04384 *  
Lot_ShapeIrregular                                 -2.473e+04  1.903e+04  -1.300  0.19388    
Land_ContourHLS                                     6.192e+04  9.164e+03   6.757 1.78e-11 ***
Land_ContourLow                                     1.104e+04  1.144e+04   0.965  0.33479    
Land_ContourLvl                                     9.927e+03  6.844e+03   1.450  0.14706    
UtilitiesNoSeWa                                    -5.571e+04  6.241e+04  -0.893  0.37219    
UtilitiesNoSewr                                    -6.654e+04  6.319e+04  -1.053  0.29249    
Lot_ConfigCulDSac                                   2.584e+04  6.494e+03   3.980 7.12e-05 ***
Lot_ConfigFR2                                      -1.053e+04  8.264e+03  -1.274  0.20265    
Lot_ConfigFR3                                       1.367e+04  1.758e+04   0.778  0.43683    
Lot_ConfigInside                                    2.051e+03  3.464e+03   0.592  0.55376    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 62140 on 2307 degrees of freedom
Multiple R-squared:  0.4038,	Adjusted R-squared:  0.394 
F-statistic: 41.12 on 38 and 2307 DF,  p-value: < 2.2e-16
```

Q4 – Use tidymodel for necessary preprocessing steps as you see fit. (normalization, transformation, etc.)

```
# Q4

simple_ames <-
  recipe(Sale_Price ~ MS_SubClass + MS_Zoning + Lot_Frontage + Lot_Area + Street+ Alley + Lot_Shape + Land_Contour + Utilities + Lot_Config,
         data = ames_train) %>%
  step_dummy(all_nominal()) %>%
  prep(training = ames_train)


train_new <- bake(simple_ames, new_data = NULL)
test_new <- bake(simple_ames, new_data = ames_test)
```

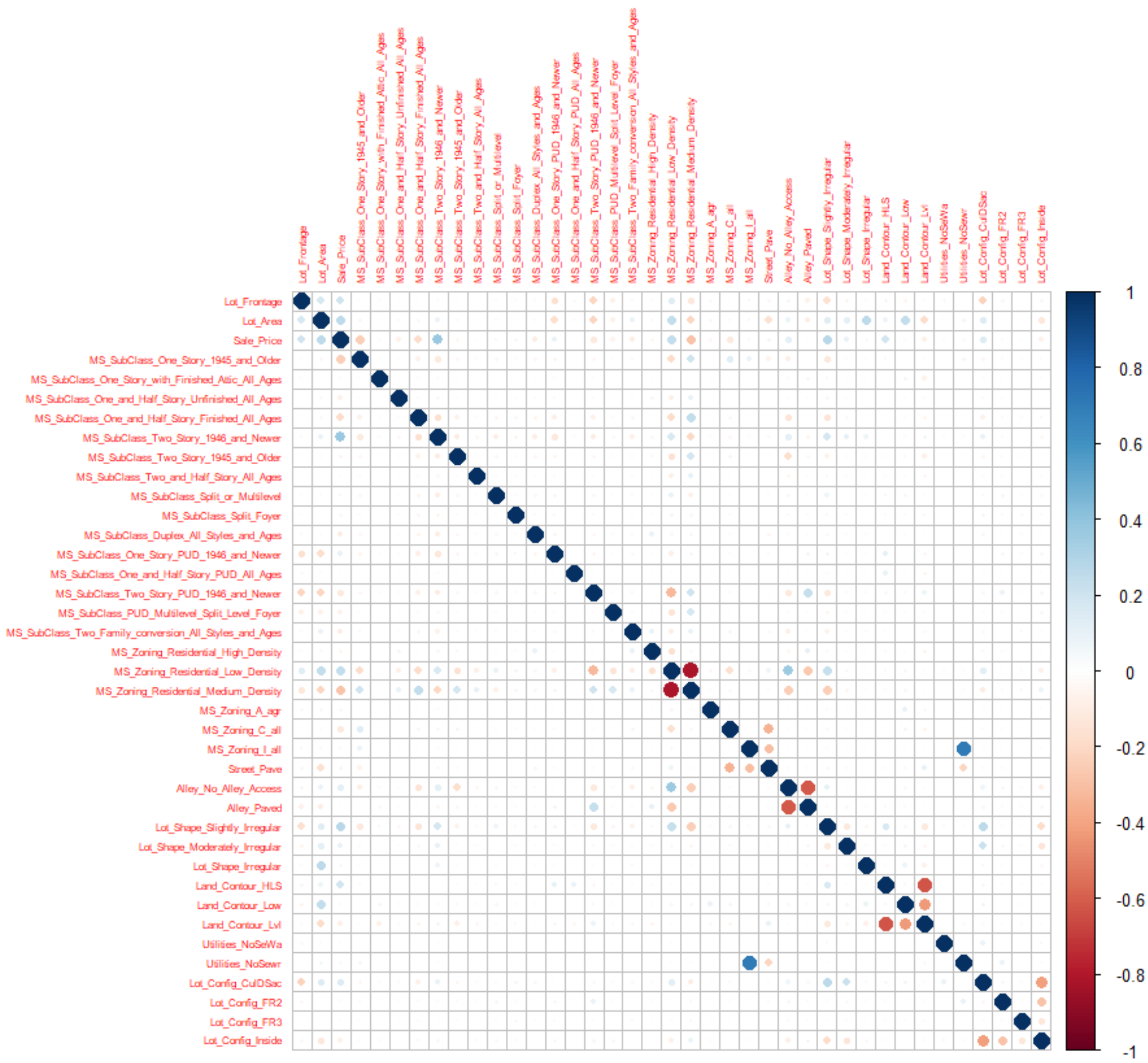
- *Like cooking a dinner; Recipe, Prep and Bake is used for preprocessing. Since varibles are categorical i used step_dummy function and transformed categorical variables to nominal.*

Q5- Look at the correlation values, can you see multicollinearity if yes, remove necessary variables. (You can use `corrplot()` function from corrplot package)

```r
# Q5 Getting correlations between variables

correlations <- cor(train_new, method="pearson")
correlations
corrplot(correlations, method = "circle", tl.cex = 0.5)
```

- *Used pearson method to observe correlations between variables in preprocessed train_new data.*



As we can see from the plot, there are some highly correlated variables:

MS_Zoning_Residential_Medium_Density	*MS_Zoning_Residential_Low_Density*
Alley_No_Alley_Access	*Alley_No_Paved*
Land_Contour_Lvl	*Land_Contour_HLS*

- *Removed one variable from each pair to prevent multicollinearity.*

```
train_new <- select(train_new, -'Alley_No_Alley_Access', -'MS_Zoning_Residential_Medium_Density', -'Land_Contour_HLS')
test_new <- select(test_new, -'Alley_No_Alley_Access', -'MS_Zoning_Residential_Medium_Density', -'Land_Contour_HLS')
```

Q6- Look at the p_values in linear regression if you see statistically insignificant values, remove them.

```
# Q6

pvalues <- lm(Sale_Price ~ MS_SubClass + MS_Zoning + Lot_Frontage + Lot_Area + Street+ Alley
              + Lot_Shape + Land_Contour + Utilities + Lot_Config, data = ames)
summary(pvalues)
```

- *If the p value of a variable is greater than 0.05, we can say that it is insignificant. R really helps us to find the significant values, indicating them with stars. So i eliminated the insignifiant ones and tested the model like that.*

```
Coefficients:
                                                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                        8.278e+04  2.304e+04   3.594 0.000332 ***
MS_SubClassOne_Story_1945_and_Older               -5.539e+04  6.126e+03  -9.042  < 2e-16 ***
MS_SubClassOne_Story_with_Finished_Attic_All_Ages -8.331e+03  2.593e+04  -0.321 0.747975    
MS_SubClassOne_and_Half_Story_Unfinished_All_Ages -4.007e+04  1.518e+04  -2.639 0.008357 ** 
MS_SubClassOne_and_Half_Story_Finished_All_Ages   -2.581e+04  4.583e+03  -5.633 1.95e-08 ***
MS_SubClassTwo_Story_1946_and_Newer                4.698e+04  3.308e+03  14.202  < 2e-16 ***
MS_SubClassTwo_Story_1945_and_Older               -4.809e+03  6.416e+03  -0.750 0.453576    
MS_SubClassTwo_and_Half_Story_All_Ages             3.409e+04  1.354e+04   2.518 0.011855 *  
MS_SubClassSplit_or_Multilevel                    -1.496e+04  6.120e+03  -2.445 0.014549 *  
MS_SubClassSplit_Foyer                            -2.670e+04  9.313e+03  -2.867 0.004177 ** 
MS_SubClassDuplex_All_Styles_and_Ages             -3.109e+04  6.379e+03  -4.874 1.15e-06 ***
MS_SubClassOne_Story_PUD_1946_and_Newer            4.189e+04  5.265e+03   7.956 2.53e-15 ***
MS_SubClassOne_and_Half_Story_PUD_All_Ages        -2.514e+04  6.320e+04  -0.398 0.690812    
MS_SubClassTwo_Story_PUD_1946_and_Newer           -6.087e+03  6.878e+03  -0.885 0.376291    
MS_SubClassPUD_Multilevel_Split_Level_Foyer       -1.492e+04  1.596e+04  -0.935 0.349964    
MS_SubClassTwo_Family_conversion_All_Styles_and_Ages -4.161e+04  8.581e+03  -4.849 1.31e-06 ***
MS_ZoningResidential_High_Density                 -5.368e+04  1.367e+04  -3.927 8.82e-05 ***
MS_ZoningResidential_Low_Density                  -3.484e+04  6.538e+03  -5.328 1.07e-07 ***
MS_ZoningResidential_Medium_Density               -5.029e+04  7.290e+03  -6.898 6.45e-12 ***
MS_ZoningA_agr                                    -2.015e+05  4.511e+04  -4.466 8.26e-06 ***
MS_ZoningC_all                                    -8.273e+04  1.482e+04  -5.581 2.61e-08 ***
MS_ZoningI_all                                    -1.260e+05  5.281e+04  -2.386 0.017115 *  
Lot_Frontage                                       5.013e+02  3.915e+01  12.804  < 2e-16 ***
Lot_Area                                           1.986e+00  1.769e-01  11.225  < 2e-16 ***
StreetPave                                         4.634e+04  1.991e+04   2.327 0.020014 *  
AlleyNo_Alley_Access                               9.525e+03  6.454e+03   1.476 0.140065    
AlleyPaved                                         9.107e+03  1.051e+04   0.866 0.386365    
Lot_ShapeSlightly_Irregular                        2.311e+04  2.877e+03   8.031 1.39e-15 ***
Lot_ShapeModerately_Irregular                      2.081e+04  7.905e+03   2.632 0.008523 ** 
Lot_ShapeIrregular                                -2.009e+04  1.634e+04  -1.229 0.218990    
Land_ContourHLS                                    6.237e+04  8.427e+03   7.402 1.75e-13 ***
Land_ContourLow                                    5.285e+03  1.058e+04   0.500 0.617454    
Land_ContourLvl                                    1.319e+04  6.190e+03   2.131 0.033174 *  
UtilitiesNoSeWa                                   -5.514e+04  6.300e+04  -0.875 0.381497    
UtilitiesNoSewr                                   -1.801e+04  5.216e+04  -0.345 0.729849    
Lot_ConfigCulDSac                                  2.398e+04  5.876e+03   4.081 4.61e-05 ***
Lot_ConfigFR2                                     -6.268e+03  7.525e+03  -0.833 0.404946    
Lot_ConfigFR3                                      1.182e+04  1.709e+04   0.692 0.489267    
Lot_ConfigInside                                   3.979e+03  3.157e+03   1.260 0.207677    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
train_excluded <- select(train_new, -'MS_SubClass_One_Story_with_Finished_Attic_All_Ages', -'MS_SubClass_Two_Story_1945_and_Older',
                         -'MS_SubClass_One_and_Half_Story_PUD_All_Ages', -'MS_SubClass_Two_Story_PUD_1946_and_Newer',
                         -'MS_SubClass_PUD_Multilevel_Split_Level_Foyer',  -'Alley_Paved', -'Lot_Shape_Irregular', -'Land_Contour_Low',
                         -'Utilities_NoSeWa', -'Utilities_NoSewr', -'Lot_Config_FR2', -'Lot_Config_FR3', -'Lot_Config_Inside' )

test_excluded <- select(test_new, -'MS_SubClass_One_Story_with_Finished_Attic_All_Ages', -'MS_SubClass_Two_Story_1945_and_Older',
                        -'MS_SubClass_One_and_Half_Story_PUD_All_Ages', -'MS_SubClass_Two_Story_PUD_1946_and_Newer',
                        -'MS_SubClass_PUD_Multilevel_Split_Level_Foyer',  -'Alley_Paved', -'Lot_Shape_Irregular', -'Land_Contour_Low',
                        -'Utilities_NoSeWa', -'Utilities_NoSewr', -'Lot_Config_FR2', -'Lot_Config_FR3', -'Lot_Config_Inside' )
```

Q7- Report your final model and its performance on the testing data

```
mdl <- lm(Sale_Price ~ . , data = train_excluded)

summary(mdl)
glance(mdl)
tidy(mdl)

predict(mdl, test_excluded)
```

- *Used glance and tidy functions to observe model.*
- *Predicted Sale_Price for the test data is given with the last command.*

```
> glance(mdl)
# A tibble: 1 x 12
  r.squared adj.r.squared  sigma statistic   p.value    df  logLik    AIC    BIC deviance df.residual  nobs
      <dbl>         <dbl>  <dbl>     <dbl>     <dbl> <dbl>   <dbl>  <dbl>  <dbl>    <dbl>       <int> <int>
1     0.401         0.395 62108.      62.1 2.16e-236    25 -29208. 58469. 58625.  8.95e12        2320  2346

> tidy(mdl)
# A tibble: 26 x 5
   term                                           estimate std.error statistic  p.value
   <chr>                                             <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)                                     81408.    23460.       3.47 5.30e- 4
 2 Lot_Frontage                                      516.       41.4     12.4  1.81e-34
 3 Lot_Area                                            2.01      0.191   10.5  2.25e-25
 4 MS_SubClass_One_Story_1945_and_Older           -53868.     6305.      -8.54 2.31e-17
 5 MS_SubClass_One_and_Half_Story_Unfinished_All_Ages -42791. 15785.     -2.71 6.76e- 3
 6 MS_SubClass_One_and_Half_Story_Finished_All_Ages -25534.   4661.      -5.48 4.77e- 8
 7 MS_SubClass_Two_Story_1946_and_Newer            51194.     3563.      14.4  6.46e-45
 8 MS_SubClass_Two_and_Half_Story_All_Ages         34902.    14959.       2.33 1.97e- 2
 9 MS_SubClass_Split_or_Multilevel                -13030.     6926.      -1.88 6.00e- 2
10 MS_SubClass_Split_Foyer                        -26676.    10291.      -2.59 9.60e- 3
# ... with 16 more rows

> predict(mdl, test_excluded)
        1         2         3         4         5         6         7         8         9        10        11        12 
167566.90 256094.67 250395.96 216694.38 186095.86 195800.18 216469.78 141279.11 227702.31 239839.54 196443.80 207945.12 
       13        14        15        16        17        18        19        20        21        22        23        24 
215486.64 228724.78 207968.59 207909.57 122009.11 265766.81 266027.16 165192.51 165192.51 231605.85 176497.06 146540.72 
       25        26        27        28        29        30        31        32        33        34        35        36 
124150.91 157517.98 178882.16 181798.54 176605.96 196027.56 220222.30  80995.33  98236.48  92412.73 128507.35 147876.00 
       37        38        39        40        41        42        43        44        45        46        47        48 
153557.75 154821.85 130887.67 171891.77 166651.45 166715.92 121451.45 167700.20 181210.64 193928.64 284402.21 146656.49 
       49        50        51        52        53        54        55        56        57        58        59        60 
222337.85 167553.44 140438.49 156510.91 170438.22 143498.08 168611.64 214971.52 159811.50 129230.33  99851.66 124291.84 
       61        62        63        64        65        66        67        68        69        70        71        72 
246587.88  93811.19  90419.85 226004.16 145336.03 125129.90 115981.84 136638.73 181798.54 143039.10 238812.20 173218.88 
       73        74        75        76        77        78        79        80        81        82        83        84 
172398.95 204970.43 134102.19 156512.22 162381.12 166491.32 115558.73 180619.56 134530.03 134560.26 202807.66 251772.36 
       85        86        87        88        89        90        91        92        93        94        95        96 
204625.48 200487.17 224569.40 247124.09 212716.93 222686.08 192232.90 192232.90 242388.59 175279.61 206546.14 216049.97 
       97        98        99       100       101       102       103       104       105       106       107       108 
151974.11 243800.87 223412.06 227653.01 247570.30 226779.43 235109.19 204922.92 268331.00 273684.16 172282.40 223061.16 
      109       110       111       112       113       114       115       116       117       118       119       120 
168262.02 178817.69 165396.01 162490.14 240705.40 163754.79 229206.22 163340.17 178011.77 152477.82 207618.42 246368.12 
      121       122       123       124       125       126       127       128       129       130       131       132 
191527.65 179623.61 179579.50 178300.62 185425.29 175298.89 164666.90 231839.99 167700.20 155770.01 111651.21 168838.25 
      133       134       135       136       137       138       139       140       141       142       143       144 
223053.85 124331.49 115661.60 113529.39 151623.53 154361.55  92920.22 117627.41 114437.86 114437.86  90655.68 152958.48 
      145       146       147       148       149       150       151       152       153       154       155       156 
194803.62  90228.63 127743.69 144101.94 142528.92 179513.86 140827.34 191009.97 171766.40 171766.40 108152.78 186617.69 
      157       158       159       160       161       162       163       164       165       166       167       168 
193838.91 145811.03 251124.89 125047.25 101029.78 194187.40 170241.19 208663.47 181100.89 164621.58 230818.06 243138.73 
      169       170       171       172       173       174       175       176       177       178       179       180 
192069.53 177551.07 155762.93 143852.52 156510.91 242353.59 167004.46 156021.75 209991.11 162151.53 214971.52 177222.07 
      181       182       183       184       185       186       187       188       189       190       191       192 
100608.09 162864.67 159164.96 156665.68 212160.21 116968.45 153268.26 210032.71 149059.73 182004.69 247195.07 115149.72 
      193       194       195       196       197       198       199       200       201       202       203       204 
151885.46 234755.94 239707.41 321073.27 279750.63 196911.84 190361.46 190154.89 230818.06 230979.24 165136.74 115558.73 
      205       206       207       208       209       210       211       212       213       214       215       216 
254738.17 199147.82 221552.28 269693.21 235384.45 258045.16 200972.07 140916.44 171590.25 228178.13 252007.58 184232.56 
      217       218       219       220       221       222       223       224       225       226       227       228 
206828.83 214618.06 203737.68 205673.06 267569.41 267922.00 216374.96 232941.59 247980.63 205573.46 207909.57 161172.24 
      229       230       231       232       233       234       235       236       237       238       239       240 
186031.47 169567.40 164418.83 194449.24 226176.80 188564.38 176198.45 183771.99 196166.47 161714.44 122208.43 114484.08 
      241       242       243       244       245       246       247       248       249       250       251       252 
164981.91 175594.00 179526.48 177096.09 162279.00 171836.28 168751.93 200944.74 160884.34 139214.45 111672.89 124635.51 
      253       254       255       256       257       258       259       260       261       262       263       264 
 87297.14 154041.30 145460.35  87306.61 139971.81 166141.84 112614.55  22430.29 109778.78 216297.50  99206.41 162402.32 
      265       266       267       268       269       270       271       272       273       274       275       276 
113383.34 171860.45 153312.71 194623.98 152291.11 144188.04 272769.82 169800.04 129651.58 131152.24 236810.83 190915.03 
      277       278       279       280       281       282       283       284       285       286       287       288 
188071.38 184029.17 216555.68 262009.97 234918.19 253310.00 155762.93 191284.37 218567.19 263125.71 167961.07 218378.01 
      289       290       291       292       293       294       295       296       297       298       299       300 
174972.70 156492.78 297008.72 137496.57 144755.50 109390.31 163547.58 133754.79 162420.58 181287.41 214500.90 141757.47 
      301       302       303       304       305       306       307       308       309       310       311       312 
244851.82 109362.29 183825.64 190881.41 170108.11 251968.89 161644.86 168768.68 123397.21 106716.16 116151.08 115389.48 
      313       314       315       316       317       318       319       320       321       322       323       324 
```