Assignment 3

YeHao Zheng, Id 1005064635

December 1, 2020

I. Data Wrangling

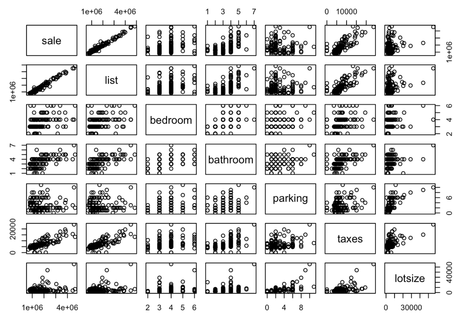
```
##   [1]  27 154 149  34 104  51 190 118  62 133  77  80  88  75  54 102  44 187
##  [19] 100 107  15  37 108  49 176  99  93  73  72 165 229  16 166  17  14  66
##  [37] 161 122 105  47 115  50 168  83  97  41  95 114 194  92  52 205  64 183
##  [55]  12 148 156  90  58 132 186  35 119 169 157 189  53  94  55 153  87  24
##  [73]  61  11 172 142 227  60  63 150 160  84 163   9  81  76 152 164  42 138
##  [91] 110 196  48  89   8  33  74 140  40  23 162 212  85 182  21 174 111 146
## [109]   7  29  20 141 126  28 207  25 159 167 218   1 155 143 134 180  69 139
## [127]  67 116 136  65 193  13  86  10  46 173 151  45   5 158   6  30   2 195
## [145] 179  19  70 109 201   4
```

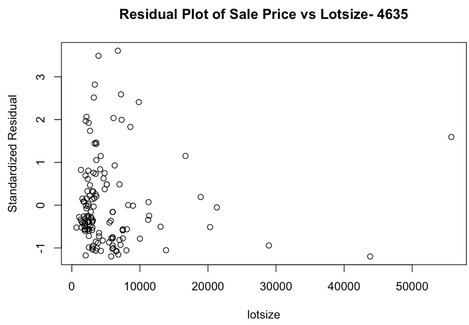
(c)
After observing the box plot of each variables, I find that there's a case where a house has 12 parking spaces. I think it's quite unusual for a house to have as many parking spaces, and for the houses in this data, the average parking space for a house is 3; so, there might be a error when inputting the data of this house. Another case that I reomved is the house with 8 bathrooms. In this case, the house only has 5 bedrooms. Even if each bedroom has 1 bathroom, and the living room has 1 as well, that's only 6 bathrooms. Therefore, I have removed this case too.
The predictor I removed is the 'maxsqfoot' because it has so many missing values. The result will not be so convincing if the data has many NAs. Also, the new variable 'lotsize' has pretty much the same function as 'maxsqfoot', we can probably know how big is the house based on its lotsize.
After I removed 'maxsqfoot', I have also removed all the other rows that contain NAs in the rest of the data.

II. Exploratory Data Analysis

```
##             sale  list bedroom bathroom parking taxes lotsize
## sale       1.000 0.985   0.378    0.596  -0.008 0.799   0.299
## list       0.985 1.000   0.389    0.606   0.034 0.791   0.310
## bedroom    0.378 0.389   1.000    0.544   0.313 0.347   0.241
## bathroom   0.596 0.606   0.544    1.000   0.280 0.554   0.316
## parking   -0.008 0.034   0.313    0.280   1.000 0.281   0.637
## taxes      0.799 0.791   0.347    0.554   0.281 1.000   0.571
## lotsize    0.299 0.310   0.241    0.316   0.637 0.571   1.000
```

```
##     list    taxes bathroom  bedroom  lotsize  parking
##    0.985    0.799    0.596    0.378    0.299   -0.008
```





(a)
sale: continuous
list: continuous
bedroom: discrete
bathroom: discrete
parking: discrete
maxsqfoot: continuous
taxes: continuous
lotwidth: continuous
lotlength: continuous
lotsize : continuous
location: categorical
(b)
Ranking each quantitative predictor for sale price, in therm of correlation coefficient, from highest to lowest is "list," "taxes," "bathroom," "bedroom," "lotsize," "parking."
(c)
Based on the scatterplot matrix, we can see that the plot of lotsize vs sale price violates the assumption of constant variance strongly because it's showing an increasing variance. Then, the plot of the standardized residuals is also showing a clear pattern of an increasing variance.

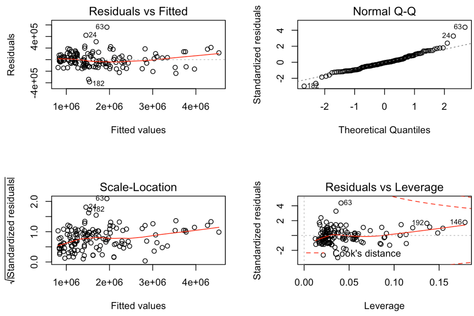
III. Methods and Model

i.

Values	Estimate	P-Value
Intercept	4.482e+04	0.437723
list	8.162e-01	< 2e-16
bedroom	5.365e+03	0.720695
bathroom	2.029e+04	0.163124
parking	-1.504e+04	0.085626
taxes	2.354e+01	0.000132
location	1.084e+05	0.006841
lotsize	1.780e+00	0.486668

From the table, we see that only the predictors "list," "taxes," "locationT" are significant because their p-value is smaller than significance level of 5% Then, we see that when list price increases by one dollar, the expected sale price will increases by 0.816 dollar when everything else stays the same. The price of taxes increases by one dollar, and then the expected sale price will increases by 0.235 dollar when everything else stays the same. Also, for the predictor "locationT", we see that for the properties in Toronto, the expected sale price will be 108400 dollar higher comparing to the properties in Mississauga, when everything else stays the same.
ii. After the backward elimination with AIC, the fitted model is
$\hat{sale} = 49570 + 0.816list + 21350bathroom - 11800parking + 25.08taxes + 108600locationT$. Here we have 5 predictor, but the result in part i shows that only "list," "taxes," "locationT" are significant, so the results are not consistent.
iii. After the backward elimination with BIC, the fitted model is
$\hat{sale} = 49570 + 0.816list + 21350bathroom - 11800parking + 25.08taxes + 108600locationT$. The result is also not consistent with the result in part i, but it's consistent with the result in part ii.

IV. Discussions and Limitations



For the residuals vs fitted plot and the scale-location plot, we can say that the linearity and constant variance assumption is basically valid because most of the dots are lying on a straight line. However, there are still rooms for improvement.
For noraml QQ plot, even though there are some dots at the two ends which are not lying on the dash line, most of the dots in the middle are lying on the dash line. Therefore, the normality assumption is basically valid, but there are still some improvement we can make.
In order to find a more valid final model, we can first do transformation on Y for the model, so that we can get a more valid normality assumption. Then we can also do transformation on X, so the linearity will be better. Finally, we'll do weighted least squares on the model, so we'll have a more constant variance.