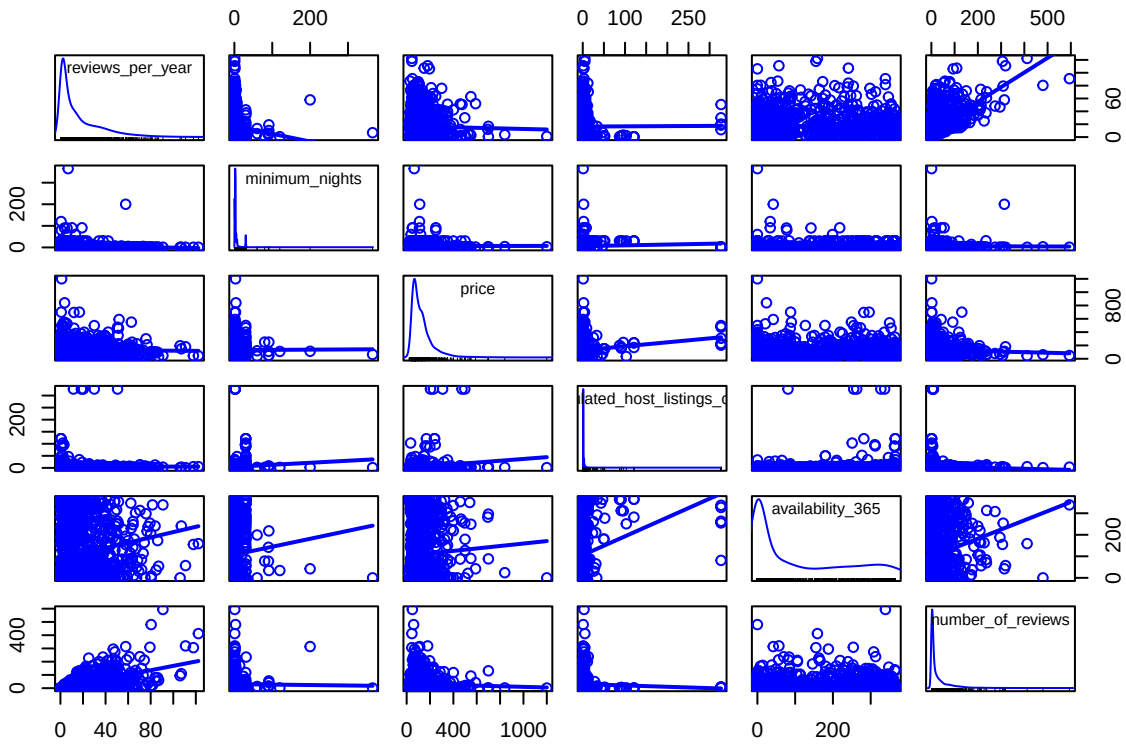
MA4710 Project Part 2

Cody

2024-05-31

An interpretation of the scatterplot matrix

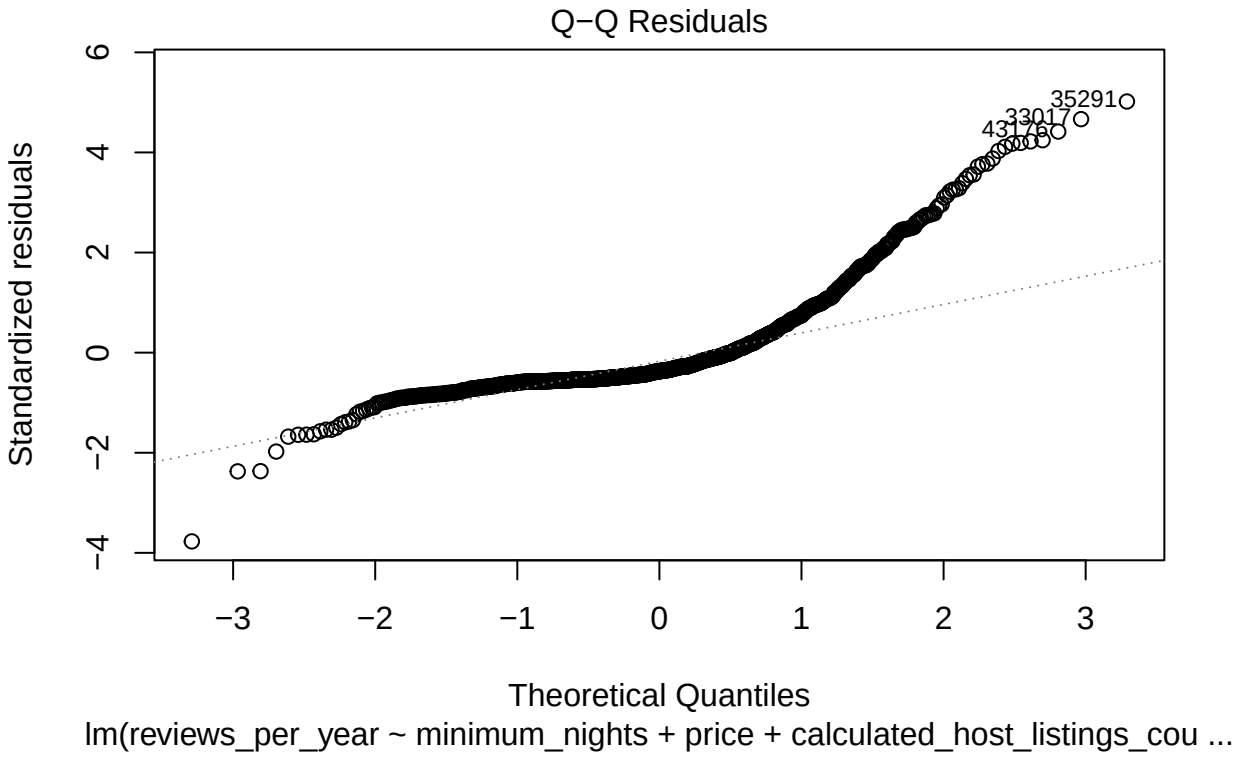
```
Loading required package: carData
Warning: package 'olsrr' was built under R version 4.3.1

Attaching package: 'olsrr'
The following object is masked from 'package:datasets':

    rivers
```



- The scatter plot matrix shows the distribution of individual predictors, in this case each predictor appears like it is skewed towards the higher end.
- The scatter plot matrix also shows relationships that might exist between a predictor and other predictors in the model. There does appear to be relationships in the data that we might want to

explore later on.

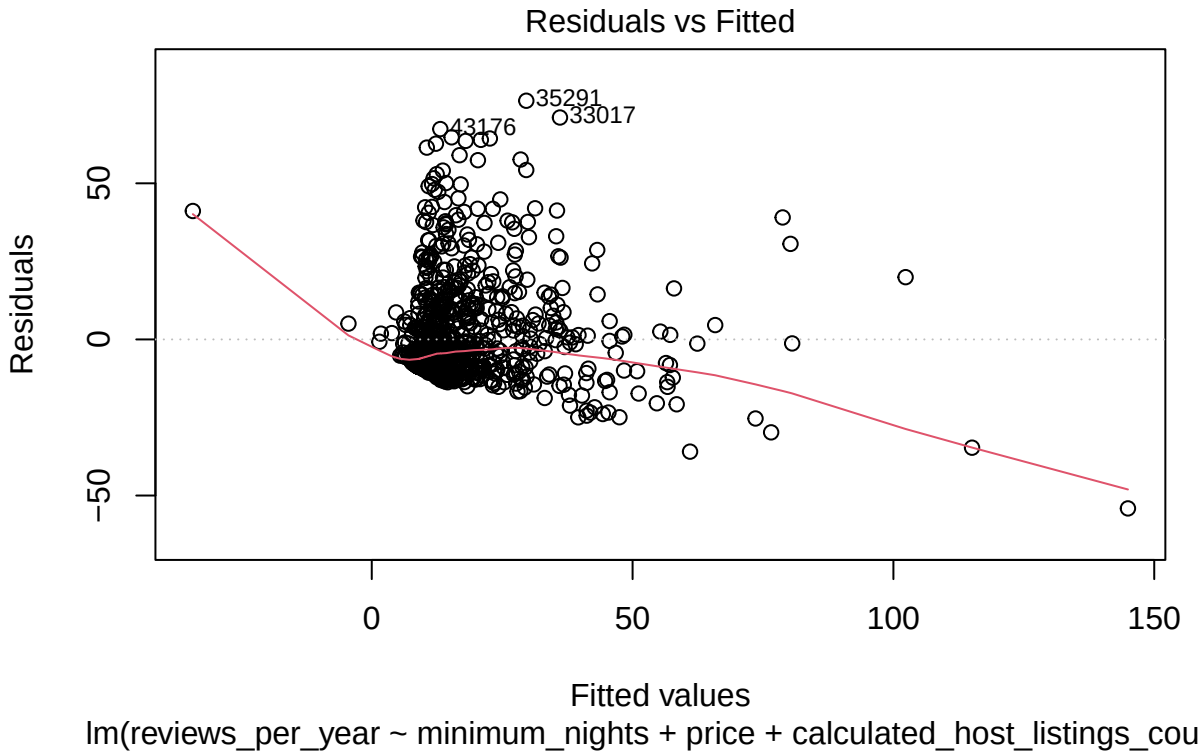
An assessment of the normality assumption, which refers to appropriate graphs and the Shapiro-Wilk p-value




```
    Shapiro-Wilk normality test

data:  rstandard(NYC.lm)
W = 0.78492, p-value < 2.2e-16
```

- The normality assumption appears to be violated. The QQ plot shows that the data skews higher than anticipated as the theoretical upper quantiles are reached.
- The Shapiro Wilks results tell us that is unlikely that the data is coming from a normally distributed population because the p value for a null hypothesis of normality is very low.

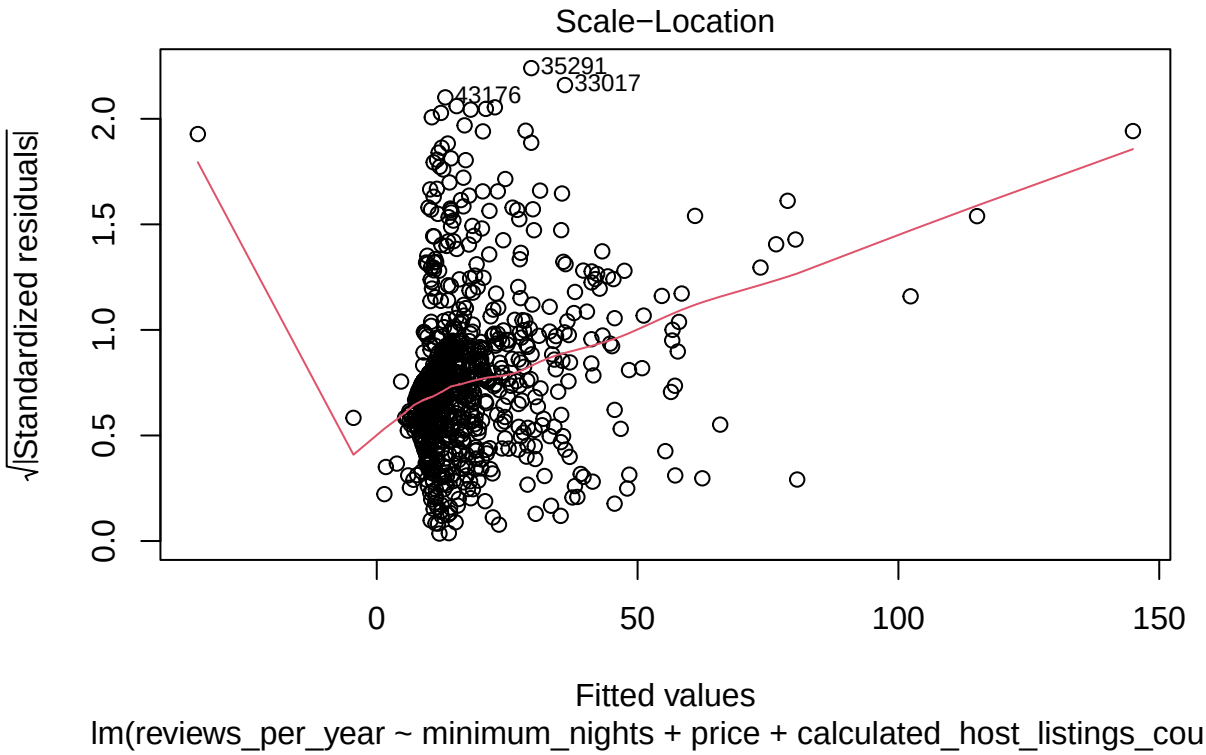
An assessment of the linearity assumption, which refers to the appropriate graphs you created in the program




```
The R squared value of the data is:

0.3919233
```

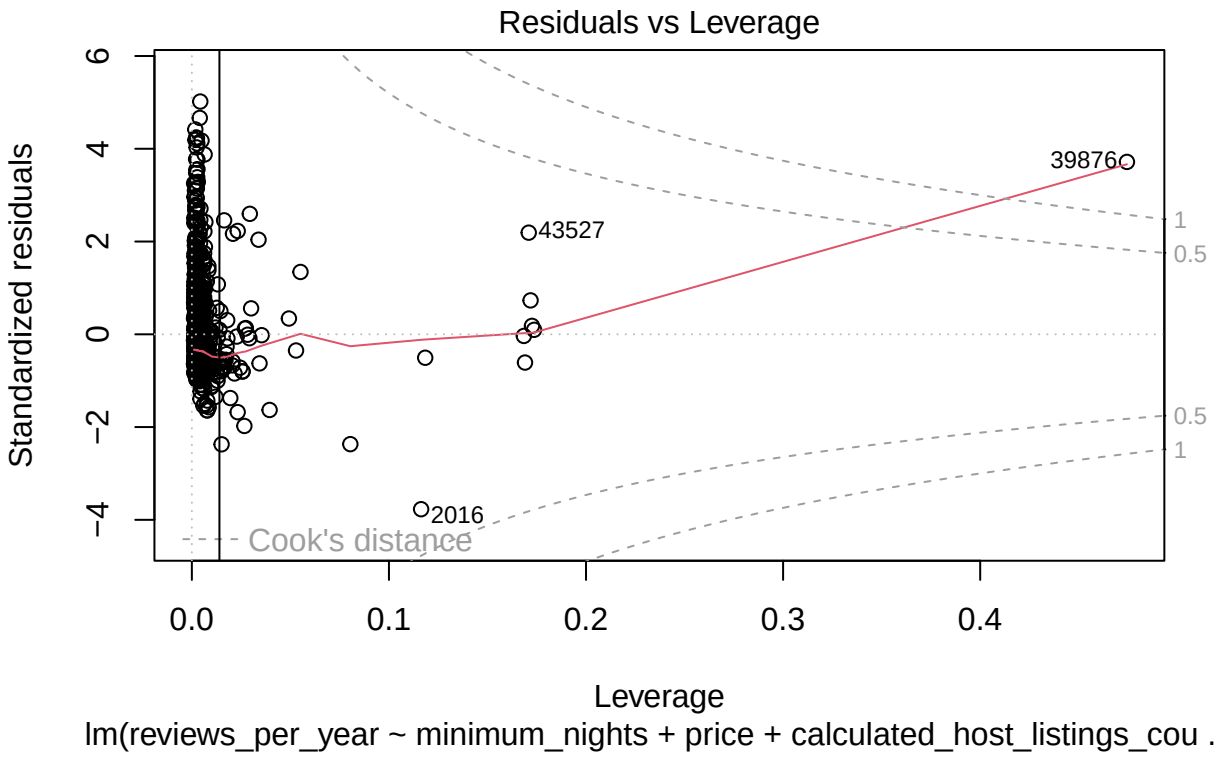
- If the data was linear, the residuals would be scattered approximately evenly as the predicted values increase. The trend line for the residuals would be a horizontal line approximately centered at 0. This does not appear to be the case with this data, as the red trend line shows some nonlinear deviation.
- The R squared value is another indicator of linearity. A value of 0.39 as is in this case indicates that there is a weak positive relationship between the linear fitted values and the actual values.

An assessment of the homoscedasticity assumption, which refers to appropriate graphs you created



- The data does not appear to meet the homoscedasticity assumption. There is a clear upwards trajectory in the trend line, this indicates that the variance fans out and increases as the predicted values increase.

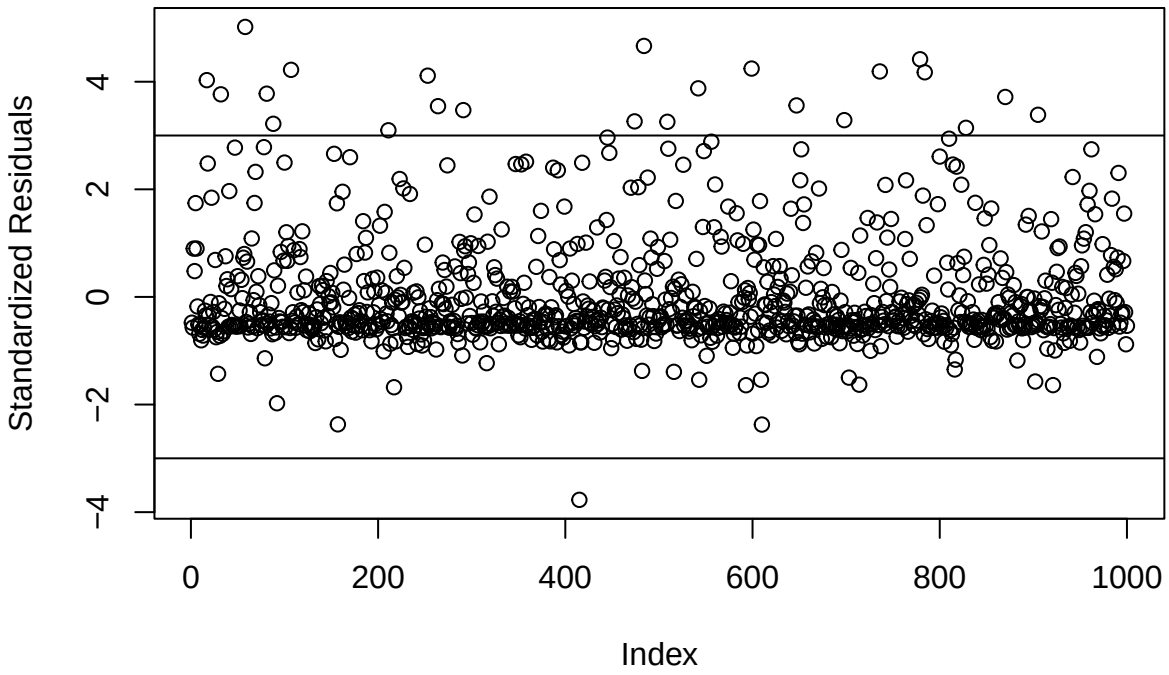
A list of points which appear to be high-leverage points




```
Indicies of leverage points greater than 0.014:

13 25 27 66 70 92 121 157 163 170 217 221 223 245 271 289 323 327 364 369 382 415 416 442 471 475 478

482 507 513 526 535 610 618 640 646 688 708 714 717 757 764 813 843 845 870 888 892 918 942 990
```

- The leverage vs residual plot is a good way to visualize if certain points have high leverage.
- A vertical line is drawn in at the value (2(p+1)/n), leverage values that fall to the right of this vertical line are considered to be high leverage points.
- A rule of thumb to determine if a point is high leverage considers the formula (2*(p+1)/n)). For easier display purposes, I listed leverage points that were a little higher than this value to limit the number of indices that showed up.

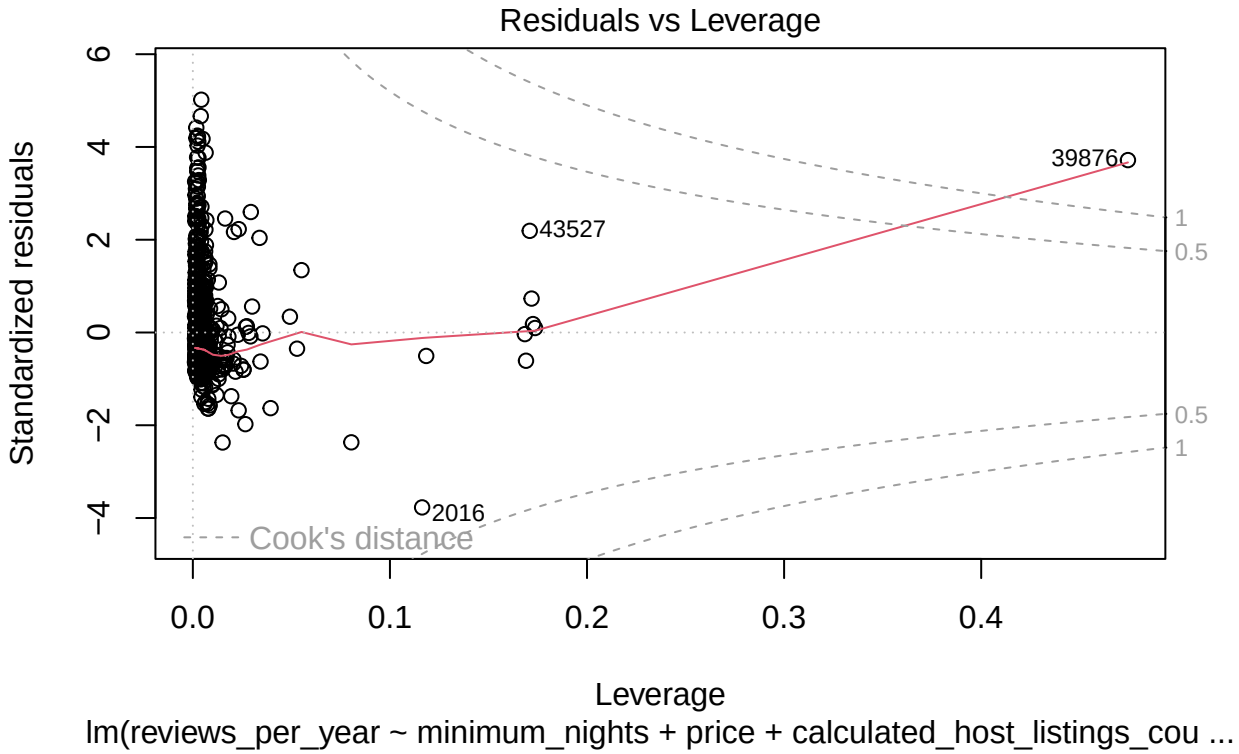
A list of points which appear be outliers or



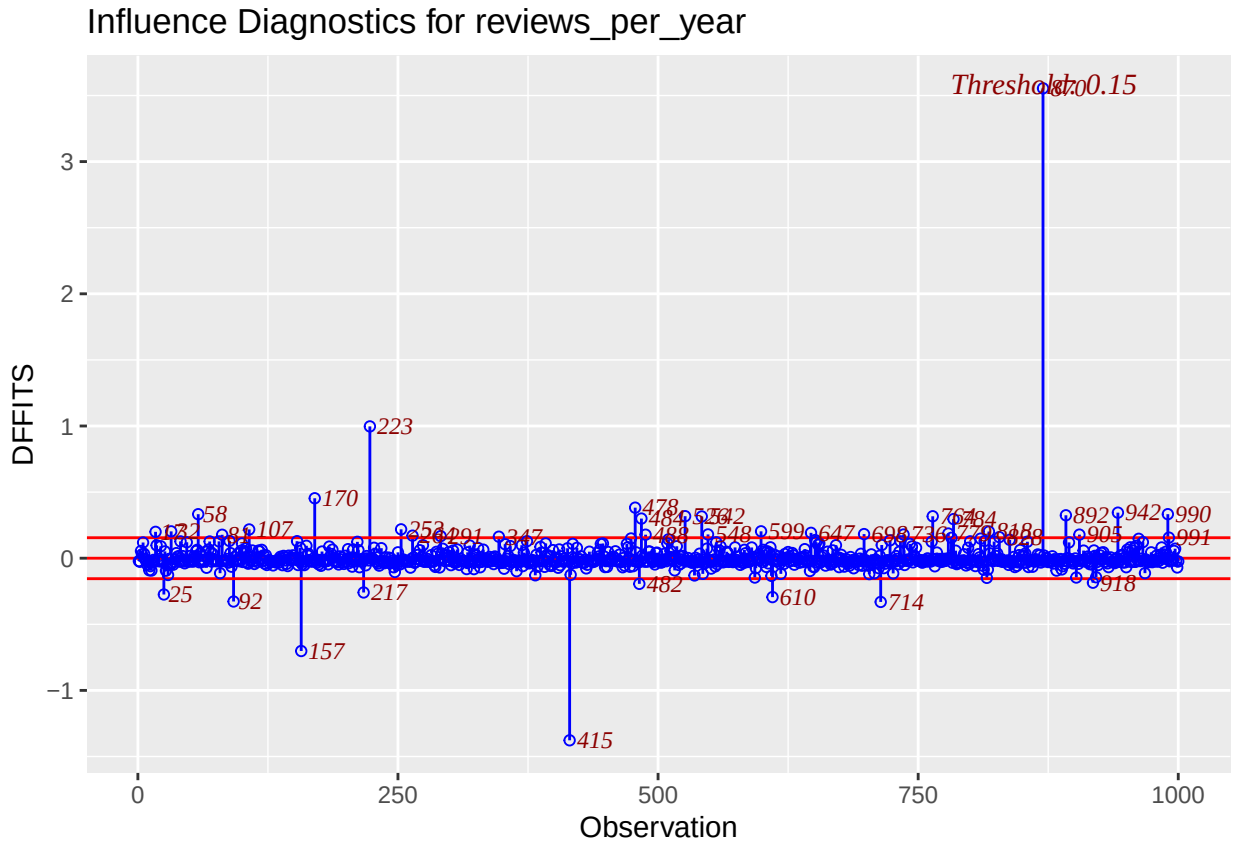

```
Indicies of standardized residuals greater than 3:

17 32 58 81 88 107 211 253 264 291 474 484 509 542 599 647 698 736 779 784 828 870 905
```

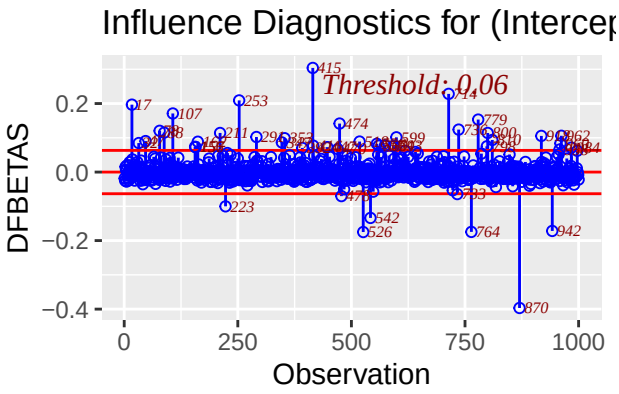
- I am considering outliers as points having a standardized residual with an absolute value greater than 3. 99% of observations in a normally distributed population should fall within 3 standard deviations of the mean, so about only 1% of data points should be this extreme.
- Listed above are the indices of review frequency data points that have standardized residuals greater than 3.

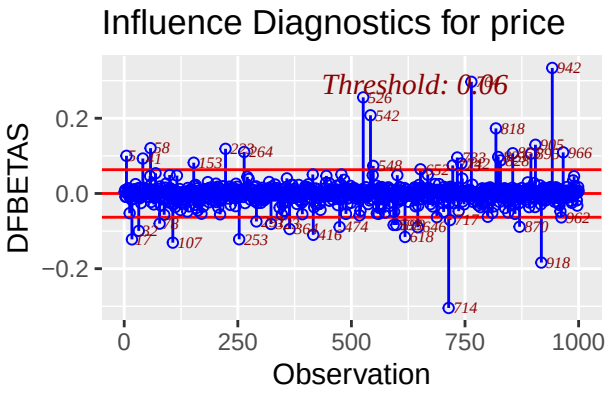
A list of points which appear to be influential on various aspects of the model fit describe the reasoning you use, by referring to appropriate graphs, when coming to a conclusion as to whether or not an observation is influential

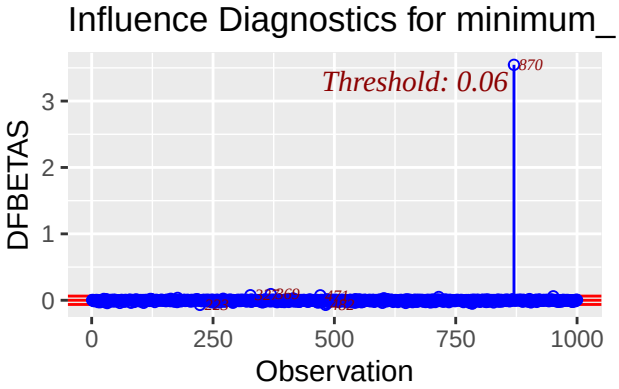


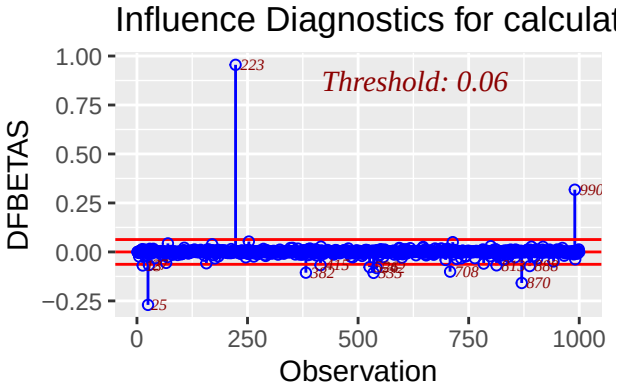

```
points which have a Cook distance greater than 0.1:
223 415 870
```

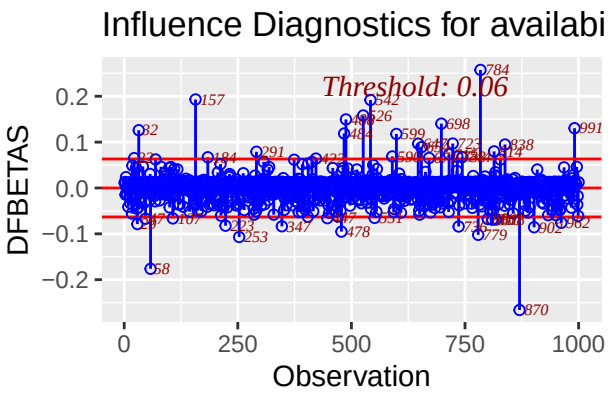

Influence Diagnostics for reviews_per_year
Threshold: 0.15
DFFITS
Observation
3
2
1
0
−1
0
250
500
750
1000
223
170
58
107
25
92
157
217
415
482
478
610
714
599
647
892
942
990
991
905
918

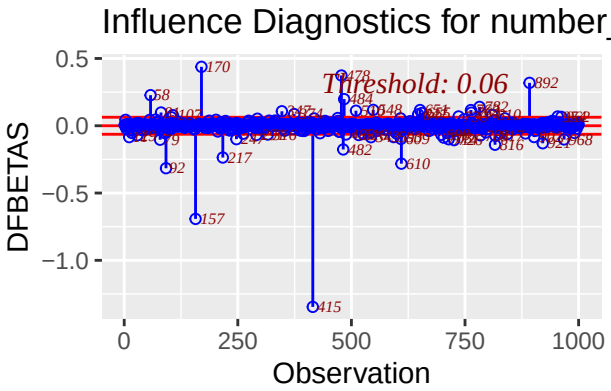

Influence Diagnostics for (Intercep
Threshold: 0.06
DFBETAS
Observation


Influence Diagnostics for price
Threshold: 0.06
DFBETAS
Observation


Influence Diagnostics for minimum_
Threshold: 0.06
DFBETAS
Observation

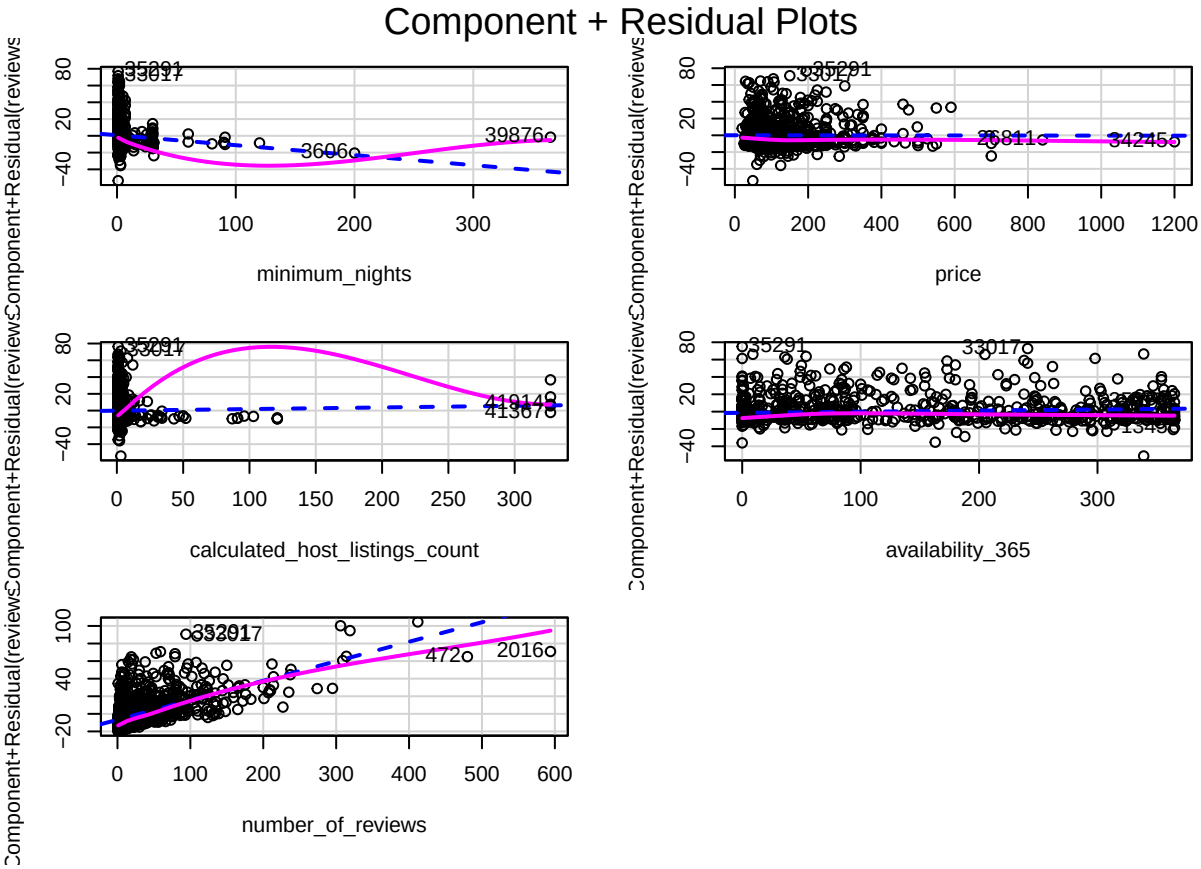

Influence Diagnostics for calculat
Threshold: 0.06
DFBETAS
Observation





- Point 870 is clearly influential on the model as its cook's distance is greater than 1.
- Points 213 and 445 are higher influence points because they are higher leverage points with fairly large standardized residuals, however their cook's distance is still less than 1 so they are less influential than point 870.

An interpretation of the residual plus component plots in regards to whether or not any variables add anything of value to the model in the presence of the other predictors



- Total number of reviews shows a clear positive linear relationship with frequency of reviews.
- Minimum reviews appears to have relatively strong negative linear relationship with the frequency of reviews. There is a high leverage, high influence point (Point 39876) that is pulling the slope of the line characterizing this relationship up.