Modeling the price of a diamond

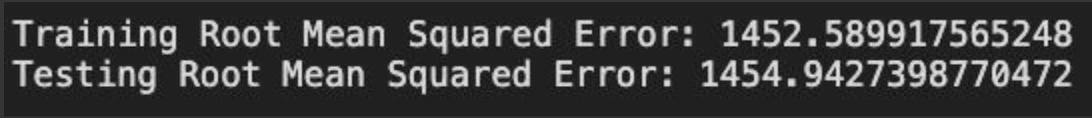
Kenny Oh

Given the following attributes of a diamond, can we predict its price?

- price price in US dollars ($326--$18,823)
- carat weight of the diamond (0.2--5.01)
- cut quality of the cut (Fair, Good, Very Good, Premium, Ideal)
- color diamond colour, from J (worst) to D (best)
- clarity a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
- x length in mm (0--10.74)
- y width in mm (0--58.9)
- z depth in mm (0--31.8)
- depth total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)
- table width of top of diamond relative to widest point (43--95)

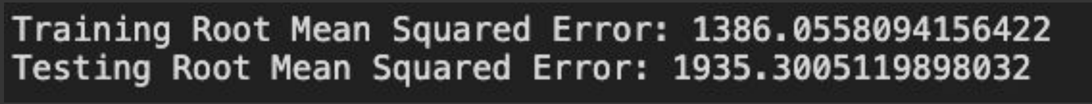
The results:

Good (simple linear regression on all features)



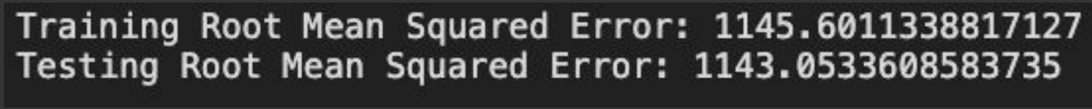
```
Training Root Mean Squared Error: 1452.589917565248
Testing Root Mean Squared Error: 1454.9427398770472
```

Better (after generating polynomial features)



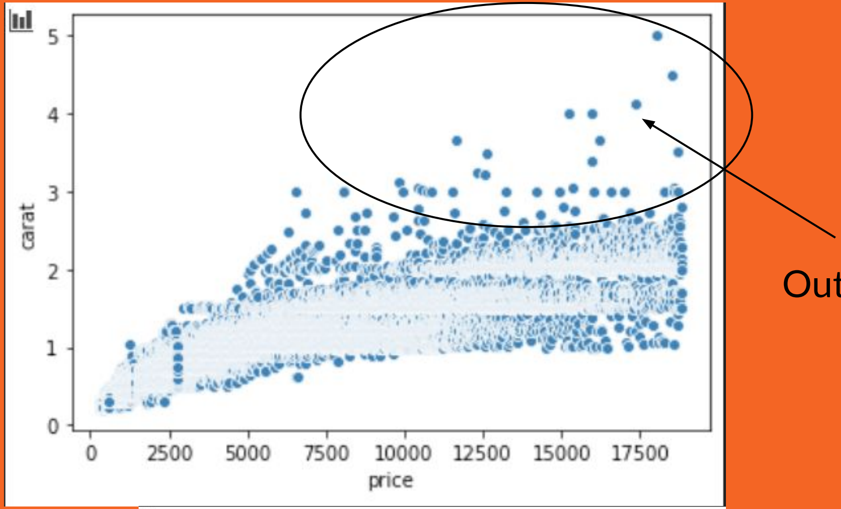
```
Training Root Mean Squared Error: 1386.0558094156422
Testing Root Mean Squared Error: 1935.3005119898032
```

Best (after using k best feature selection method and generating dummy variables for categorical data [color, clarity, cut])

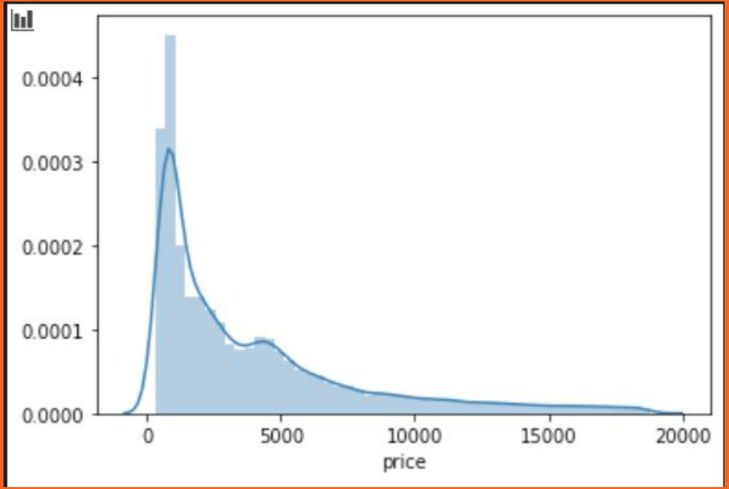


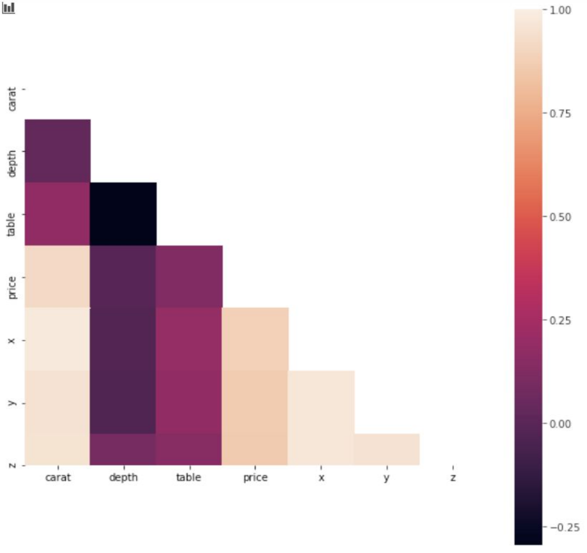
```
Training Root Mean Squared Error: 1145.6011338817127
Testing Root Mean Squared Error: 1143.0533608583735
```

Looking at the data...



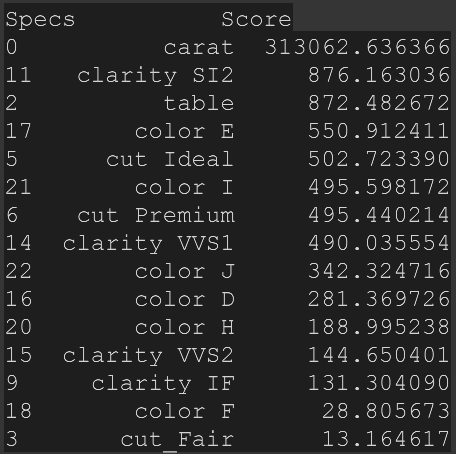
Outliers



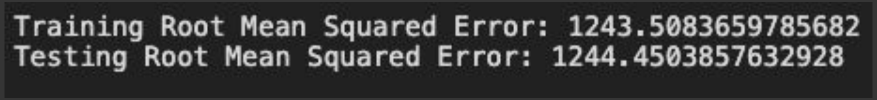


SelectKBest using different scoring functions

f_regression

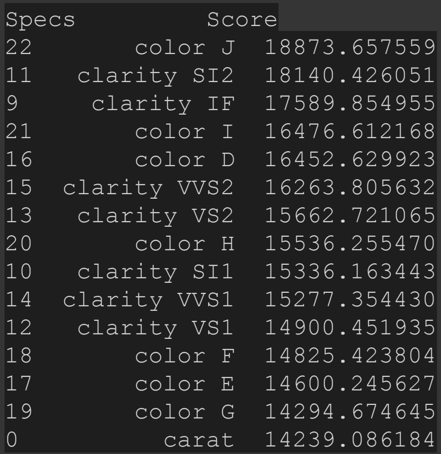


```
Specs          Score
0          carat  313062.636366
11   clarity SI2     876.163036
2          table     872.482672
17       color E     550.912411
5      cut Ideal     502.723390
21       color I     495.598172
6    cut Premium     495.440214
14  clarity VVS1     490.035554
22       color J     342.324716
16       color D     281.369726
20       color H     188.995238
15  clarity VVS2     144.650401
9     clarity IF     131.304090
18       color F      28.805673
3       cut_Fair      13.164617
```

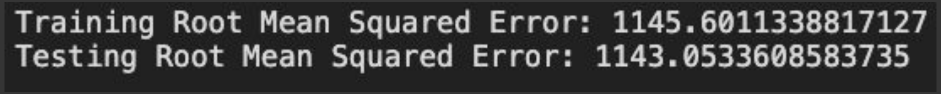


```
Training Root Mean Squared Error: 1243.5083659785682
Testing Root Mean Squared Error: 1244.4503857632928
```

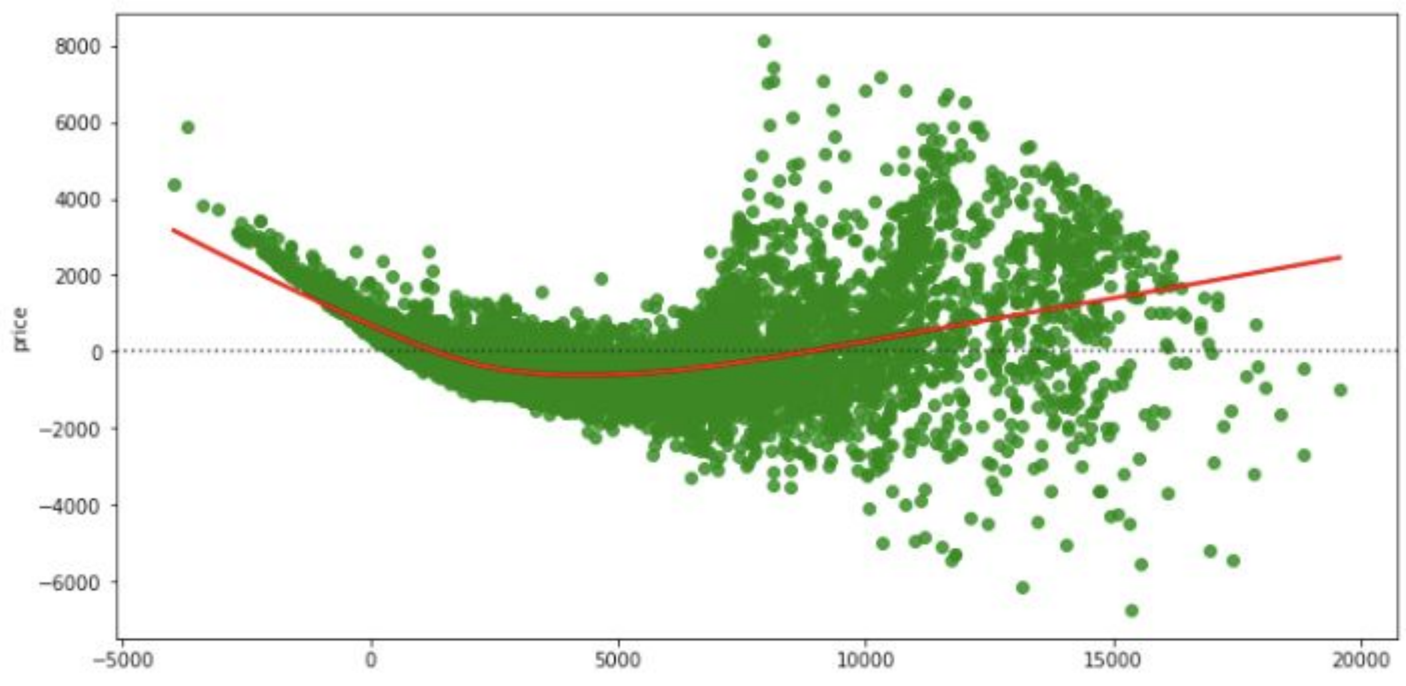
chi_squared



```
Specs         Score
22       color J  18873.657559
11   clarity SI2  18140.426051
9     clarity IF  17589.854955
21       color I  16476.612168
16       color D  16452.629923
15  clarity VVS2  16263.805632
13   clarity VS2  15662.721065
20       color H  15536.255470
10   clarity SI1  15336.163443
14  clarity VVS1  15277.354430
12   clarity VS1  14900.451935
18       color F  14825.423804
17       color E  14600.245627
19       color G  14294.674645
0          carat  14239.086184
```



```
Training Root Mean Squared Error: 1145.6011338817127
Testing Root Mean Squared Error: 1143.053608583735
```


price
8000
6000
4000
2000
0
−2000
−4000
−6000
−5000
0
5000
10000
15000
20000

Thanks!