Assignment 2
CIS 335 - 01
Monica Klosin

1. **Use min-max scaling (range 0-1 for all vars)and z-score scaling (0 =mean, 1= std for all vars) to transform the data**

To calculate the z-score scaling for the data in R:

```
zscaleVariable <- scale(diamond$variable) #(for all variables)
zscaleDiamond <- data.frame(variables…… )
```

To calculate the min-max scaling for the data in R:

Created a function minmax:

```
minmax <- function(x)
        {
          return((x- min(x)) /(max(x)-min(x)))
        }
```

then:

```
mxscaleVariable <- minmax(diamond$variable) #(for all
variables)
```

2. **Use PCA and forward feature selection and backward feature selection to select the 5 best features of the data**

Z-scaling:

We had to take our scaled data, stored in zscaleDiamond, and create a full model in R using the function lm (linear model). Our new variable for this is zDiamondFM.

PCA:

With PCA selection, we find that PC1, PC2, and PC3 hold ~90% of the variance for the data. with this information, we see that zscalez has the lowest variance among all the PCs we are intrested in, so we would drop Z and keep Carat, Table, Depth, Y, and X in the Model.

```
Rotation (n x k) = (6 x 6):
                      PC1          PC2          PC3         PC4          PC5          PC6
zscaleCarat  0.4953672774 -0.045129669  0.027908324 -0.78996536  0.160214816  0.319502171
zscaleDepth -0.0006822439 -0.734082087 -0.671000723  0.01402986  0.088358027 -0.053638379
zscaleTable  0.1205813877  0.669826823 -0.732523408  0.01345637  0.002961066  0.003430812
zscaleX      0.5009101500 -0.008203685  0.069979239 -0.04075903 -0.048631847 -0.860289644
zscaleY      0.4952175606 -0.009657367  0.086227223  0.53762228  0.635139814  0.234074490
zscaleZ      0.4938820845 -0.101283089 -0.007508905  0.29133794 -0.748830865  0.316449645
Importance of components:
                          PC1    PC2    PC3    PC4     PC5    PC6
Standard deviation     1.9830 1.1332 0.8267 0.2177 0.19950 0.1149
Proportion of Variance 0.6554 0.2140 0.1139 0.0079 0.00663 0.0022
Cumulative Proportion  0.6554 0.8694 0.9833 0.9912 0.99780 1.0000
```

Forward feature selection:

In R using stepAIC(zDiamondFM, direction = "forward"), we find Z is the least important features of the data, so our model would include Carat, Table, Depth, Y, and X.

```
Coefficients:
(Intercept)  zscaleCarat  zscaleTable  zscaleDepth      zscaleY      zscaleX      zscaleZ
    3932.80      5065.43      -228.91      -291.04        75.75     -1475.86        29.38
```

(output for forward feature selection, we see that zscaleZ has the lowest coefficient, showing it has the smallest impact on the variable price)

Backward feature selection:

In R using ols_step_backward_p(zDiamondFM), we find Z is the least important features of the data, so our model would include Carat, Table, Depth, Y, and X.

```
[1] "Backward Selection: "

                                  Elimination Summary
-------------------------------------------------------------------------------
        Variable                  Adj.
Step    Removed     R-Square    R-Square     C(p)        AIC           RMSE
-------------------------------------------------------------------------------
   1    zscaleZ       0.8592      0.8592    5.8828    941813.9260    1496.9528
-------------------------------------------------------------------------------
```

(output for backward feature selection, shows that zscalez was dropped)

Min-Max Scaling:

PCA:

With min-max scaling, we see that PC1, PC2, and PC3 hold ~97% of the variance of the data, PC1 and PC2 along hold ~93% of the variance of the data. With this information, we see that mxscaleY has the lowest variance long the PCs we are intrested in, so we would drop Y and keep Carat, Table, Depth, X, and Z in the Model.

```
Rotation (n x k) = (6 x 6):
                      PC1         PC2         PC3         PC4          PC5          PC6
mxscaleCarat  0.670754482  0.07690008 -0.04230612  0.73642685  0.007053969  0.003275119
mxscaleDepth -0.001120238  0.64210005 -0.75594437 -0.10876385 -0.055039988 -0.037355101
mxscaleTable  0.060919118 -0.76148178 -0.64516813 -0.01305917  0.003880383 -0.002616319
mxscaleX      0.712683467 -0.01457323  0.09564556 -0.63977251 -0.266625979  0.048203202
mxscaleY      0.129159716 -0.00166058  0.02288907 -0.11691099  0.478537016 -0.860305321
mxscaleZ      0.147567187  0.04152422 -0.02905867 -0.15065965  0.834756669  0.506101009
Importance of components:
                          PC1     PC2     PC3     PC4      PC5      PC6
Standard deviation     0.1458 0.04695 0.03428 0.01571 0.004624 0.004119
Proportion of Variance 0.8529 0.08847 0.04717 0.00991 0.000860 0.000680
Cumulative Proportion  0.8529 0.94138 0.98855 0.99846 0.999320 1.000000
```

Forward feature selection:

In R using stepAIC(mxDiamondFM, direction = "forward"), we find Z is the least important features of the data, so our model would include Carat, Table, Depth, Y, and X.

Coefficients:

(Intercept)	mxscaleCarat	mxscaleTable	mxscaleDepth	mxscaleY	mxscaleX	mxscaleZ
9846	51401	-5327	-7314	3906	-14130	1324

Backward feature selection:

In R using ols_step_backward_p(mxDiamondFM), we find Z is the least important feature of the data and is dropped, so our model would include Carat, Table, Depth, Y, and X.

Elimination Summary

Step	Variable Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	mxscaleZ	0.8592	0.8592	5.8828	941813.9260	1496.9528

3. **Try one other scaler and one other feature selection**

scaler option – Robust Scaler:

PCA:

	PC1	PC2	PC3	PC4	PC5	PC6
rCarat	0.07650183	-0.2445344	0.02546699	-0.229995911	0.405586468	0.8463470560
rDepth	-0.25427482	-0.2277765	-0.93817570	0.032338268	0.040181248	-0.0250647580
rTable	0.91926972	0.2444989	-0.30843219	0.005104905	-0.002588362	-0.0005419507
rX	0.19297204	-0.5818656	0.10398611	-0.449407599	0.399143156	-0.5020943029
rY	0.19139264	-0.5916490	0.11460397	0.771111964	-0.058774176	0.0460231687
rZ	0.10278917	-0.3740655	0.01000952	-0.386586609	-0.819212257	0.1698567159

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	2.3448	1.7374	1.3229	0.19849	0.12956	0.09181
Proportion of Variance	0.5322	0.2922	0.1694	0.00381	0.00162	0.00082
Cumulative Proportion	0.5322	0.8244	0.9938	0.99756	0.99918	1.00000

With robust scaling, we see that PC1 and PC2 accounts for 98% of the variance in the data, so we will focus on those three. Focusing on those PCs we find that Z impacts the data the least, so we can drop it and our model will include Carat, Table, Depth, Y, and Z.

forward feature selection:

Coefficients:

(Intercept)	rCarat	rDepth	rTable	rY	rX	rZ
1231.87	10686.31	-203.15	-102.45	66.32	-1315.67	41.63

With robust scaling, we see that X has the lowest coefficient, meaning it has the smallest impact on the data. We drop X in this case.

backward feature selection:

Elimination Summary

Step	Variable Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	rZ	0.8592	0.8592	5.8828	941813.9260	1496.9528

Backward feature selection with robust scaling drops Z, so our model will include Carat, Table, Depth, Y, and Z.

Another feature - Best Subset Regression:

This analysis helps determine which variables are the most useful, if you could only use 1 variable, 2,3, etc. With what R shows, Z is the last value added, which means it is the least important variable and we would drop it from the model.

Best Subsets Regression

Model Index	Predictors
1	zscaleCarat
2	zscaleCarat zscaleX
3	zscaleCarat zscaleDepth zscaleX
4	zscaleCarat zscaleTable zscaleDepth zscaleX
5	zscaleCarat zscaleTable zscaleDepth zscaleY zscaleX
6	zscaleCarat zscaleTable zscaleDepth zscaleY zscaleX zscaleZ

4. **Discuss your findings**

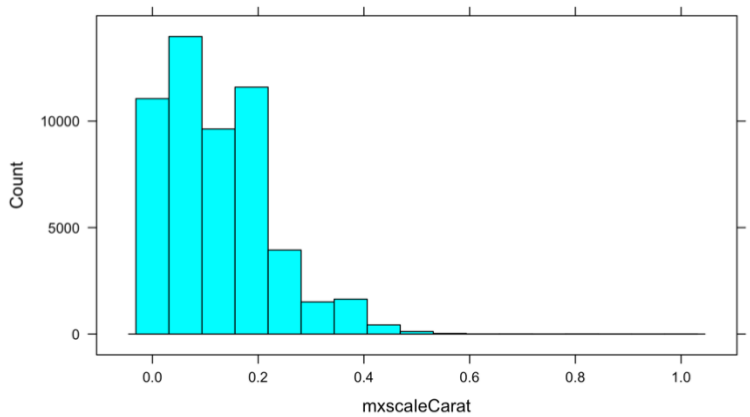
   1. **what scaling methods you used and what results they gave. Are they different? How different are they? Include a screenshot of the results as proof.**

      I used Robust scaling, Max-min scaling, and z-score scaling. All types of scaling tied with the types of selection I used determined that Z, or depth was the least useful in helping fit the data. Carat was the most important variable in fitting the data. This makes sense since the weight of a diamond corelates with its size, and the bigger the diamond, the more expensive it is. Depth is just how deep a diamond is, which can vary from size of width and height, so it is not a very good indicator of how big a diamond is, and therefor how expensive it is.

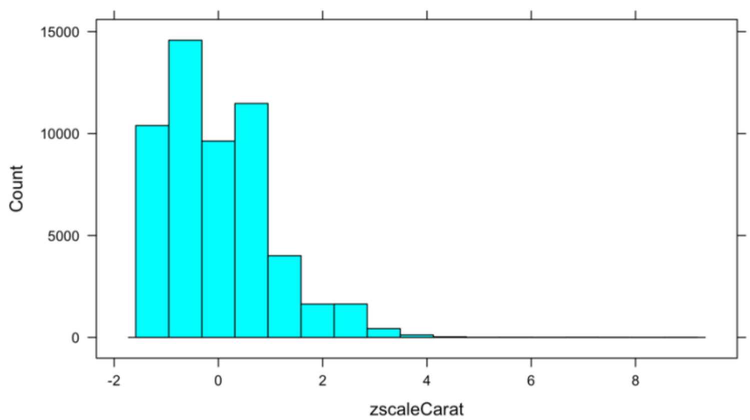
      Using PCA, Forward feature selection, backward feature selection, and best subset regression on all forms of the data (Scaled and unscaled) we find that the 5 best variables to keep in the model are carat, table, depth, y, and x.

      Because the data is so skewed, scaling the data doesn't impact it by that much. Below is a graph showing the distribution of the value carat in the data for all scaled types, and by in large, the data has stayed the same by being very right skewed.

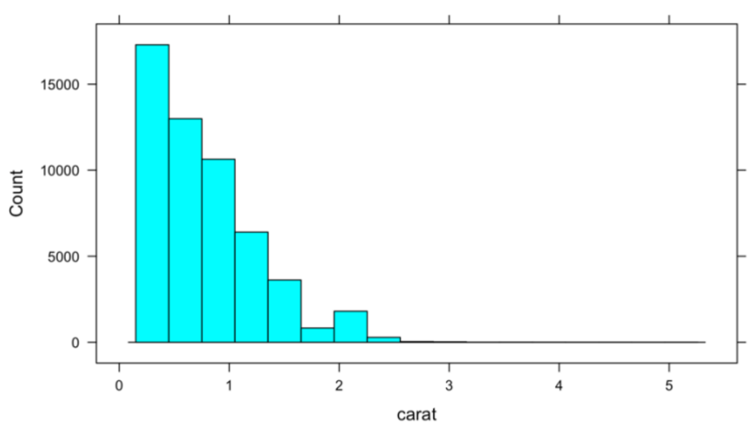
Distribution of min-max scaleCarat



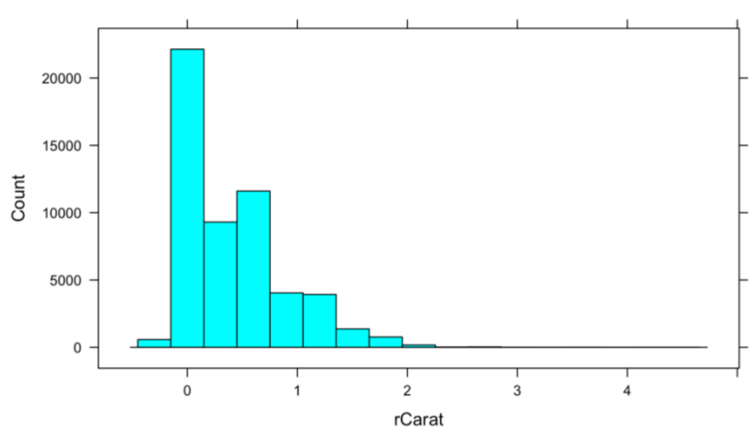
Distribution of z-scale scaleCarat



Distribution of Carat



Distribution of robust-scale scaleCarat



2. **Describe the feature selection techniques that you used. How different are they from each other? How consistent are the results? Include a screenshot of the results as proof.**

   Using all feature selection techniques, (PCA, forward, backward, and subset) all showed that Z, depth, was the least impactful variable – and that it can be dropped. Screenshots above show that with the analysis of all selection techniques, Z was the most dropped variable in the dataset.

3. **If you do not use the scaling methods, how different do the results become for step 2? Include a screenshot of the results as proof.**

   Using unscaled data, we still find that Z is the best variable to drop, since it impacts the data the least. The results from determining which variable to drop using unscaled data are the same as the results from the various scaled data.

With unscaled data:

PCA:

In unscaled PCA, PC1, PC2, and PC3 accounts for ~98% of the variance of the data. Looking at those PCs, we see that diamond depth or *z* is insignificant, and we can drop variable z.

```
                      PC1        PC2         PC3          PC4          PC5           PC6
diamond.carat  0.07650183 -0.2445344  0.02546699 -0.229995911  0.405586468 -0.8463470560
diamond.table  0.91926972  0.2444989 -0.30843219  0.005104905 -0.002588362  0.0005419507
diamond.depth -0.25427482 -0.2277765 -0.93817570  0.032338268  0.040181248  0.0250647580
diamond.x      0.19297204 -0.5818656  0.10398611 -0.449407599  0.399143156  0.5020943029
diamond.y      0.19139264 -0.5916490  0.11460397  0.771111964 -0.058774176 -0.0460231687
diamond.z      0.10278917 -0.3740655  0.01000952 -0.386586609 -0.819212257 -0.1698567159
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6
Standard deviation     2.3448 1.7374 1.3229 0.19849 0.12956 0.09181
Proportion of Variance 0.5322 0.2922 0.1694 0.00381 0.00162 0.00082
Cumulative Proportion  0.5322 0.8244 0.9938 0.99756 0.99918 1.00000
```

Forward feature selection:

We find Z is the least important features of the data, so our model would include Carat, Table, Depth, Y, and X.

```
Coefficients:
  (Intercept)  diamond.carat  diamond.table  diamond.depth      diamond.y      diamond.x      diamond.z
     20849.32       10686.31        -102.45        -203.15          66.32       -1315.67          41.63
```

Backward feature selection:

We find Z is what R drops by using backward feature selection, so our model would include Carat, Table, Depth, Y, and X.

```
                            Elimination Summary
-----------------------------------------------------------------------------
        Variable                  Adj.
Step    Removed     R-Square    R-Square     C(p)         AIC          RMSE
-----------------------------------------------------------------------------
   1    diamond.z     0.8592      0.8592    5.8828    941813.9260    1496.9528
-----------------------------------------------------------------------------
```