**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)

minmaxDiamond <- data.frame(variables…)

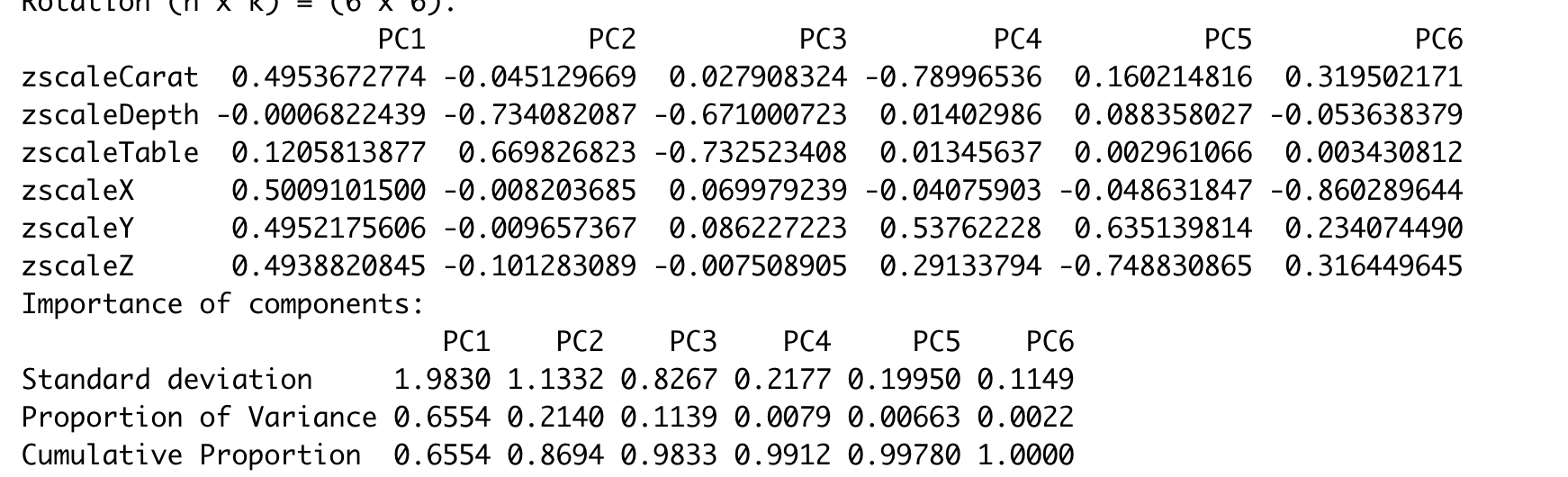
1. **Use PCA and forward feature selection and backward feature selection to select the 5 best features of the data**

**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.

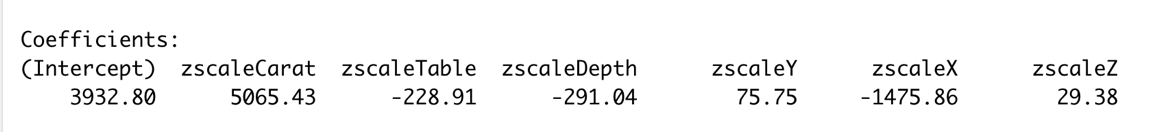
**PCA:**

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.

****

**Forward feature selection:**

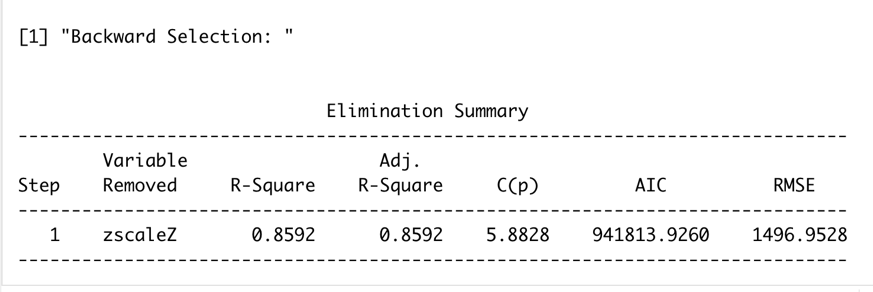
In R using stepAIC(zDiamondFM, direction = "forward"), we find Z is the least important features of the data, so our model would include Carat, Table, Depth, Y, and X.



(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.

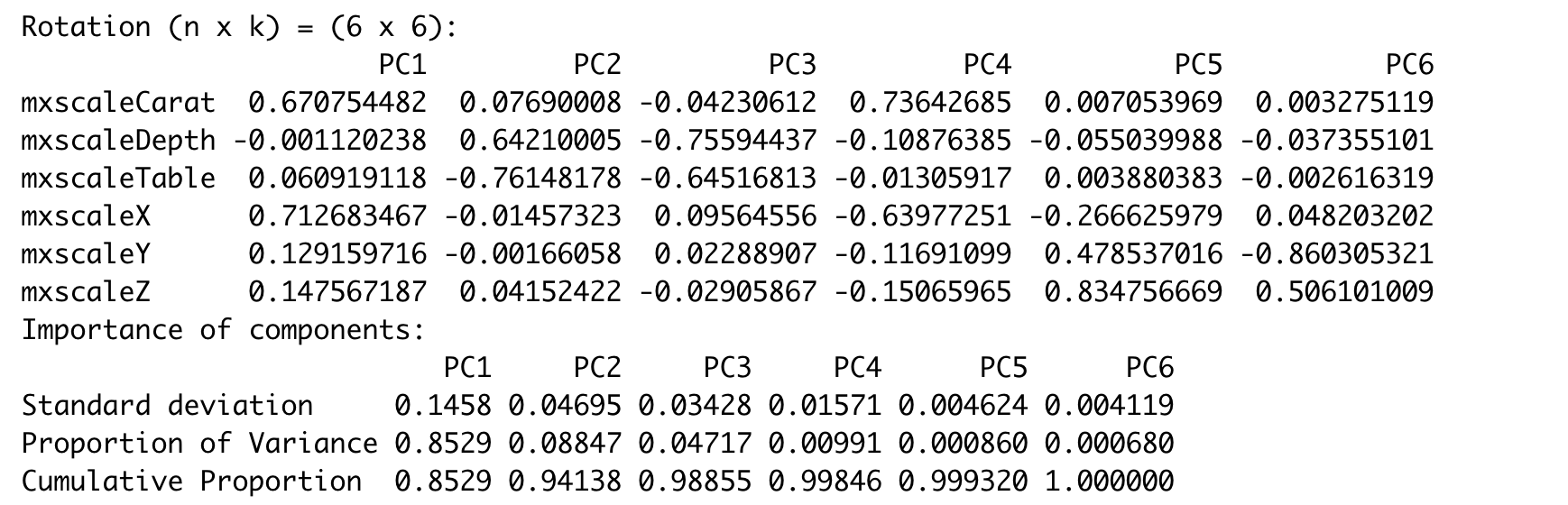


(output for backward feature selection, shows that zscalez was dropped)

**Min-Max Scaling:**

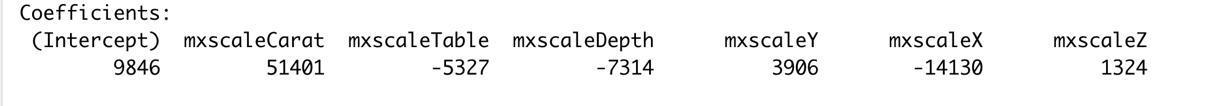
**PCA:**

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.

****

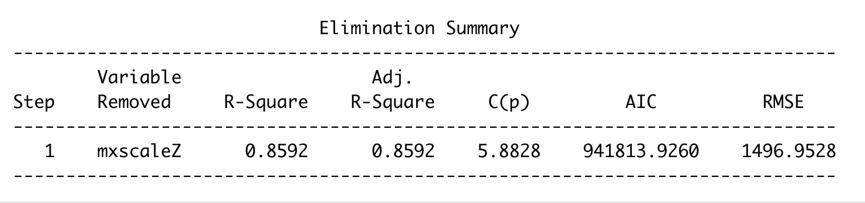
**Forward feature selection:**

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.

****

**Backward feature selection:**

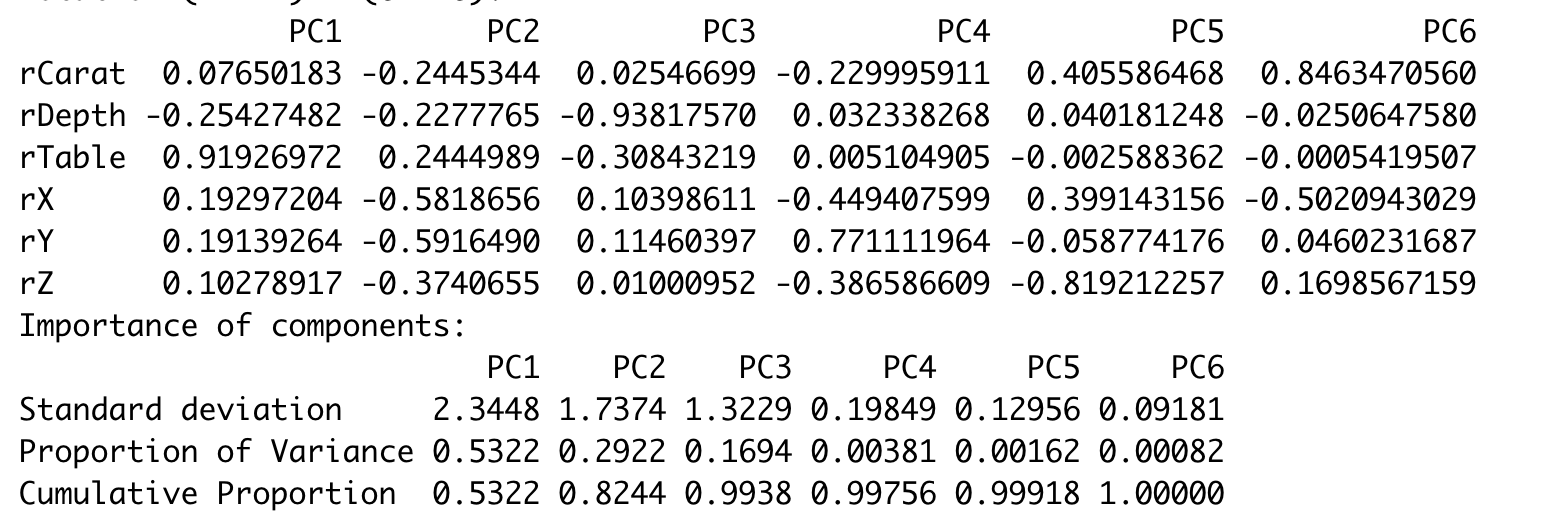
In R using ols\_step\_backward\_p(mxDiamondFM), we find Z is the least important feature of the data and is dropped, so our model would include Carat, Table, Depth, Y, and X.



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

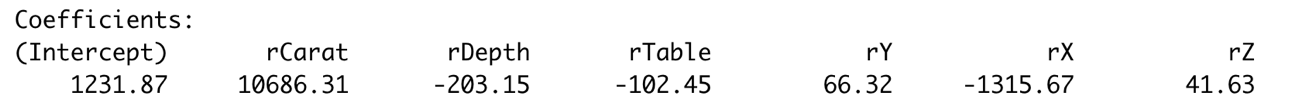
**scaler option – Robust Scaler:**

**PCA:**

****

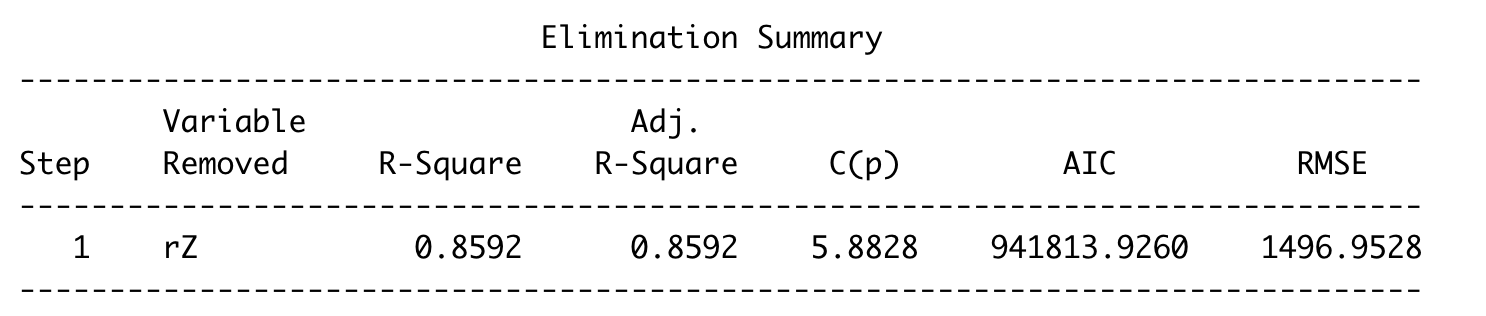
With robust scaling, we see that PC1 and PC2 accounts for 98% of the variance in the data, so we will focus on those three. Focusing on those PCs we find that Z impacts the data the least, so we can drop it and our model will include Carat, Table, Depth, Y, and Z.

**forward feature selection:**

****

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.

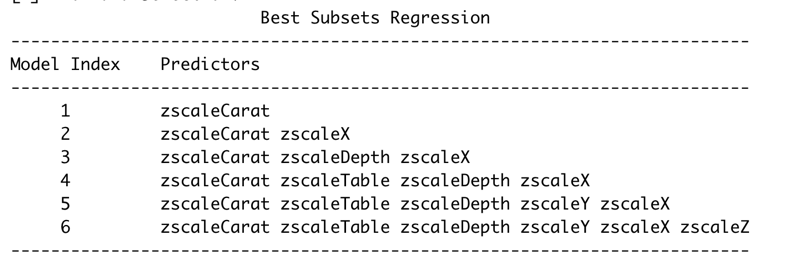
**backward feature selection:**

****

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.

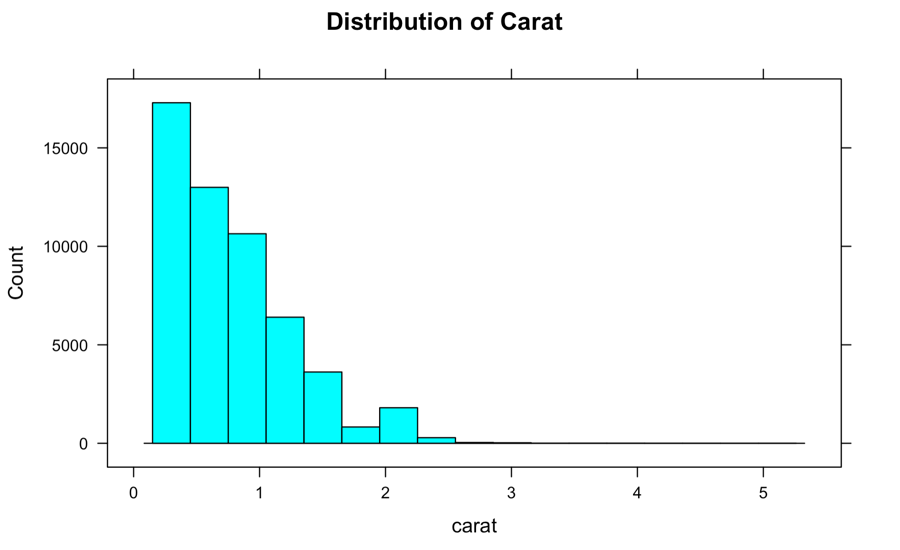
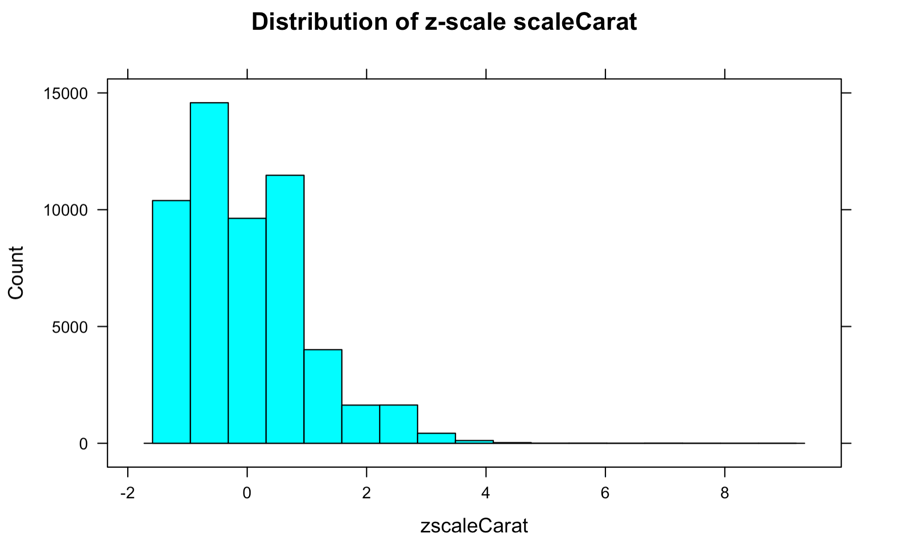
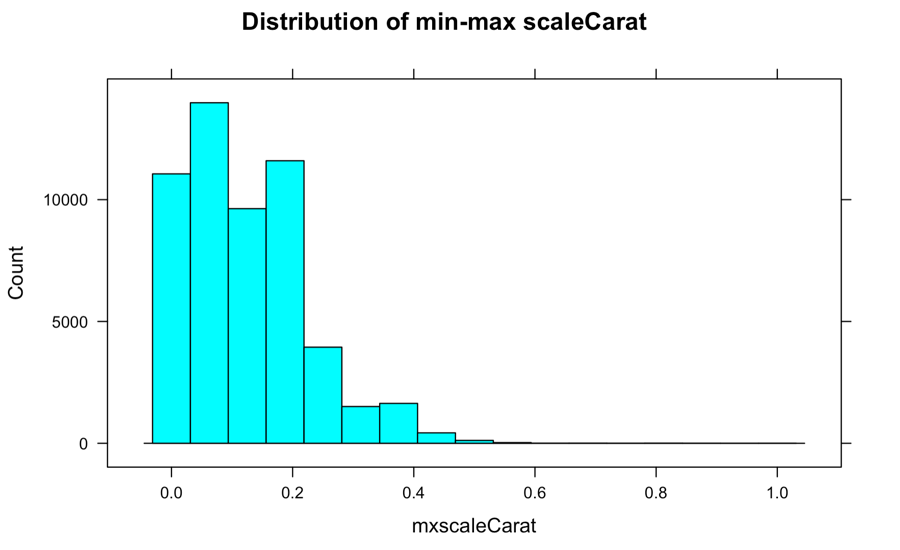
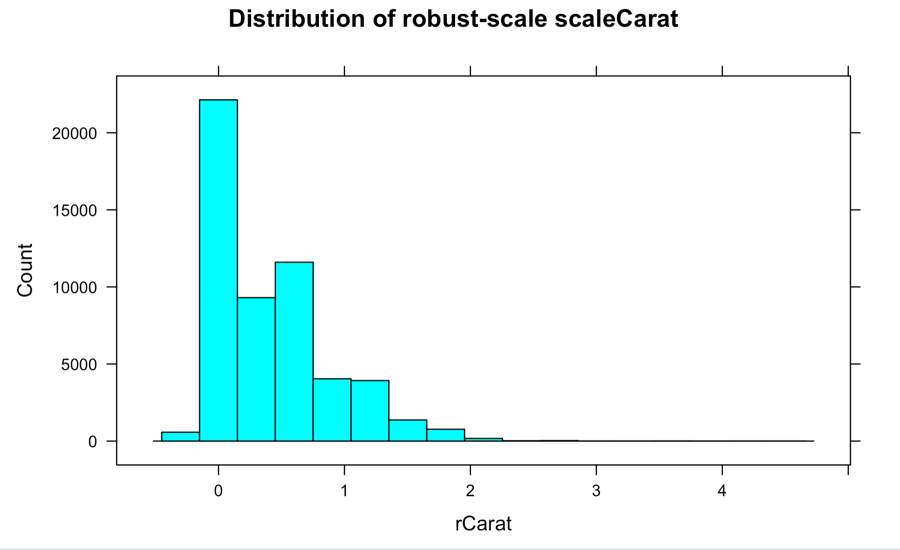


1. **Discuss your findings**
2. **what scaling methods you used and what results they gave. Are they different? How different are they? Include a screenshot of the results as proof.**

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.

****

1. **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.

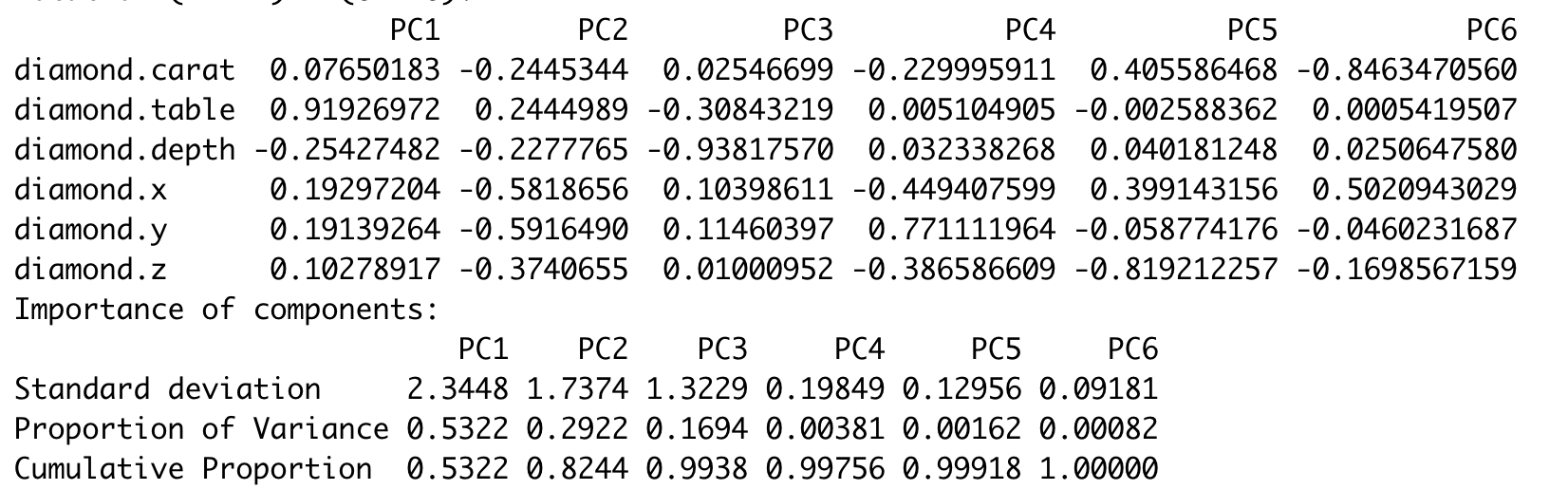
1. **If you do not use the scaling methods, how different do the results become for step 2? Include a screenshot of the results as proof.**

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:**

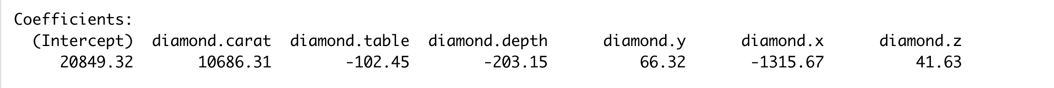
**PCA:**

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.

****

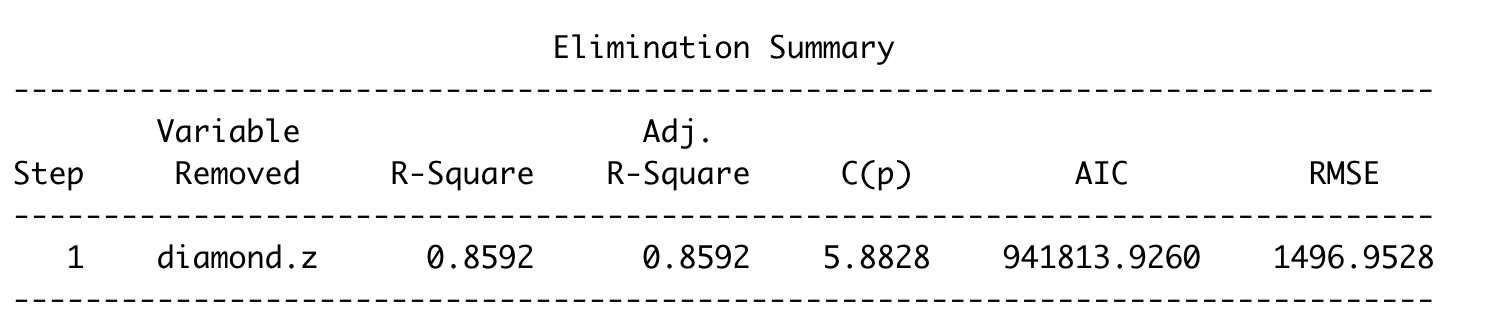
**Forward feature selection:**

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

****

**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.

****