Sebastián Pastor Ferrari

Data Mining – Assignment 02

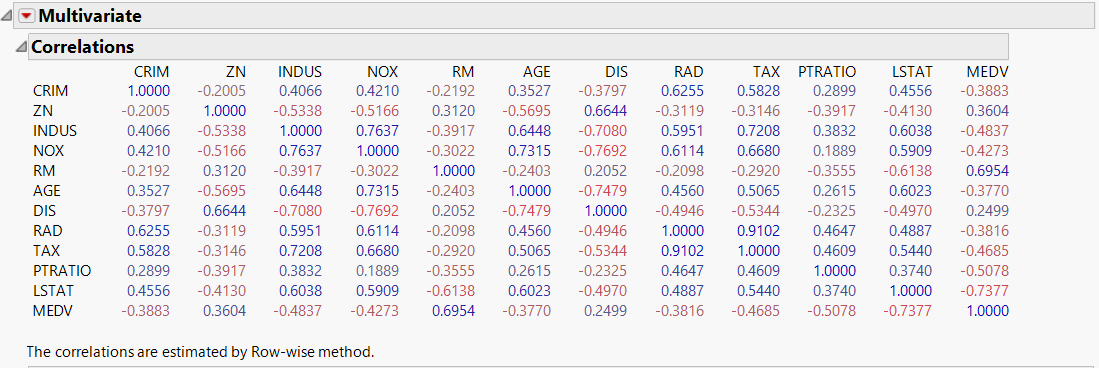
02.21.2020

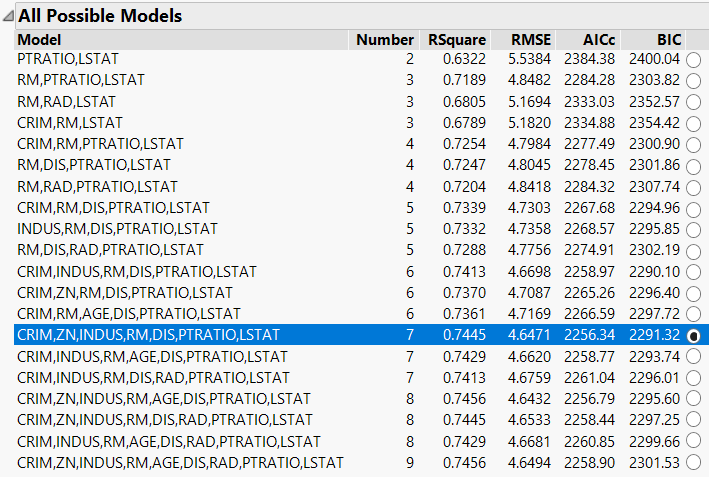
**Question 01:**

1. **Which predictors are likely to be measuring the same thing among the entire set of predictors? Discuss the relationships among INDUS, NOX, and TAX.**

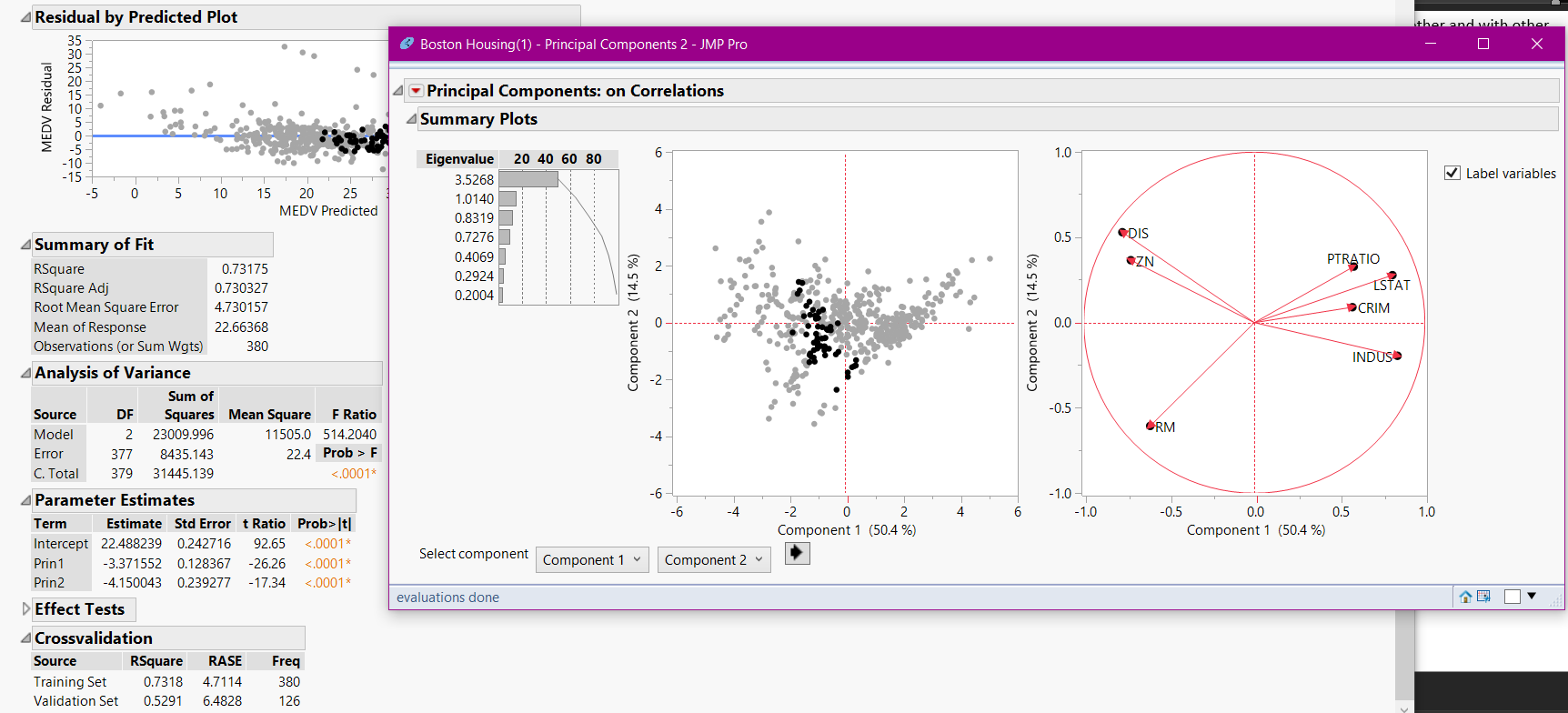
RAD-TAX which have a .91 correlation coefficient. INDUS-NOX have a high coefficient (0.763), as well as INUDS-TAX (0.721), as well as NOX-TAX (0.668).

1. **Compute the correlation table for the numerical predictors and search for highly correlated pairs. These have potential redundancy and can cause multicollinearity. Choose which ones to remove based on this table.**

I will remove TAX and NOX, given they are highly correlated with each other and with other predictors.

1. **Use an exhaustive search (All Possible Models) to reduce the remaining predictors as follows: First, choose the top three models. Then run each of these models and compare their predictive accuracy for the validation set. Compare RMSE, 𝐶𝑝, 𝐴𝐼𝐶𝑐 , and Validation RSquare. Finally, describe the best model.**

The best model I found was: MEDV = CRIM,ZN,INDUS,RM,DIS,PTRATIO,LSTAT, it has a predictive value (Adj. R2 ) = 0.7445.

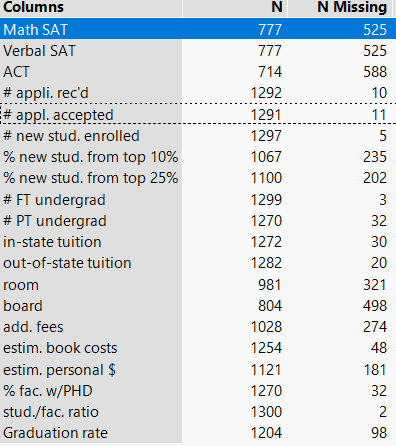
1. **Run a PCA and use only a few principal components in a regression model to predict MEDV. Compare this model with the one you obtained in part iii.**

After running a PCA with 7 predictors (CRIM,ZN,INDUS,RM,DIS,PTRATIO,LSTAT), I saved 2 components as columns and then ran a Principal Components Regression Model (PCR) to predict MEDV.

My new R2 = 0.73, slightly less than the 0.7445 that I obtained from the original Stepwise Regression.

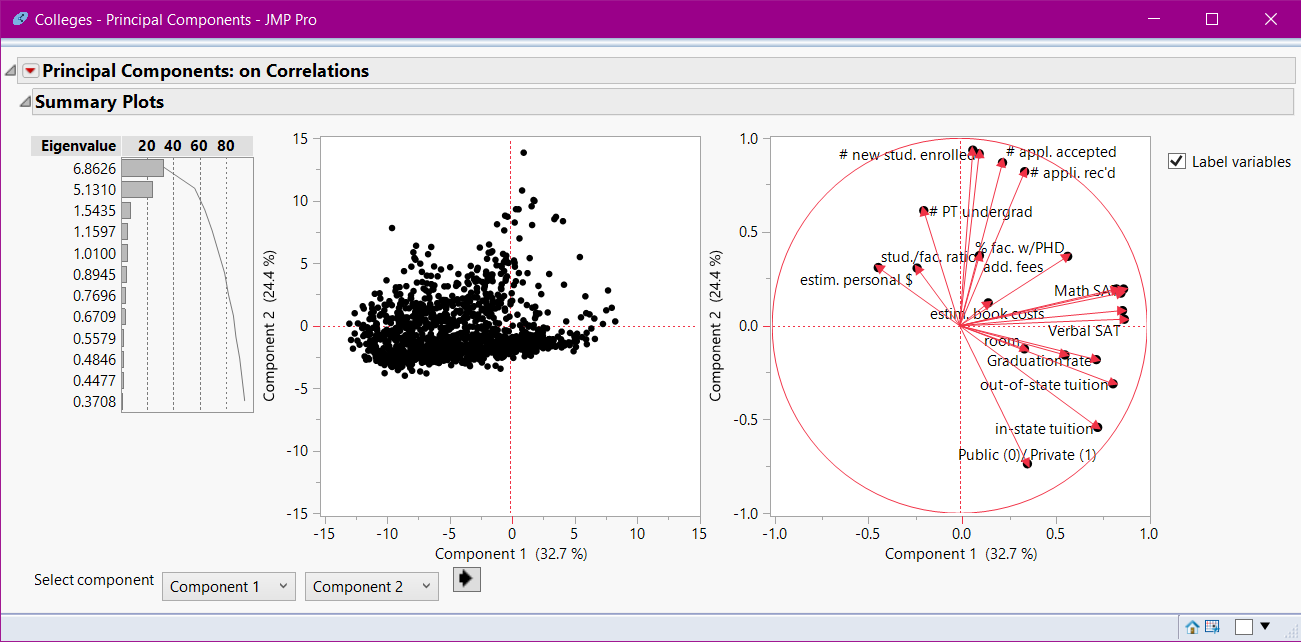
**Question 02:**

1. **Make sure the variables are coded correctly in JMP (Nominal, Ordinal, or Continuous), then use the Columns Viewer to summarize the data. Are there any missing values? How many Nominal columns are there?**

-Public/Private must be coded to ordinal, I also recoded it into {Public:0,Private:1}.

- There is a ton of missing data, notably in standardized testing. This is curious because it alludes to the trend of making standardized testing optional by American colleges.

* There are 20 nominal columns.

1. **Conduct a principal components analysis on the data and comment on the results. Recall that, by default, JMP will conduct the analysis on correlations rather than covariances. Is this necessary? Do the data need to be normalized in this case? Discuss key considerations in this decision.**

Things such as standardized testing scores max out at certain values, and they are being weighted in comparison with things like costs, out/in state tuition rates, personal $, and % features (scaled 0-100). This is not an accurate PCA because all the features have entirely different scales, which are not accounted in their weights by JMP. The issue with different numerical scales, and different measures means that it is probably best to normalize the data to a certain scale and run the PCA based on their covariances.