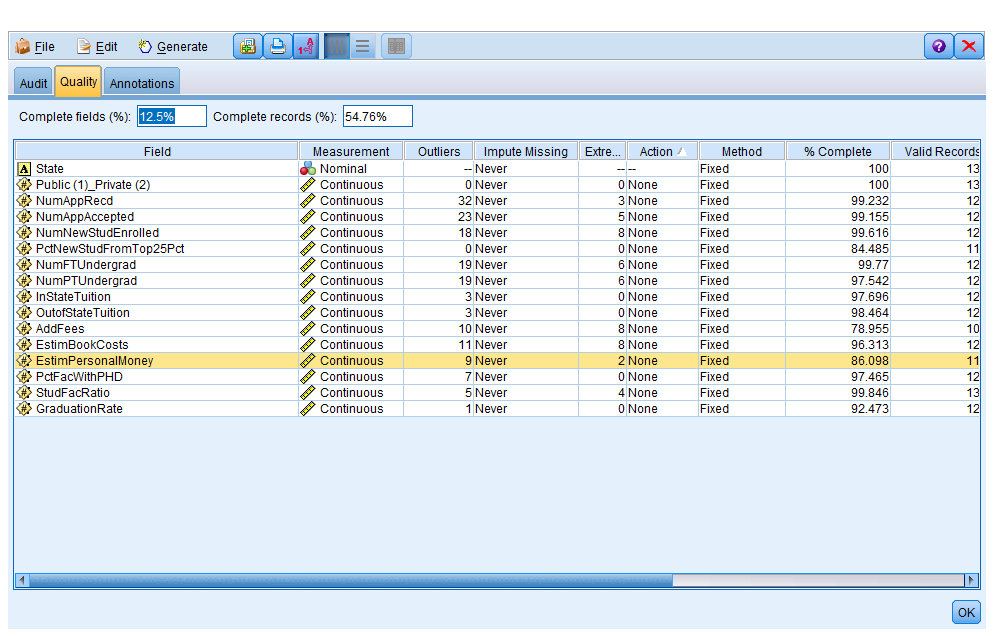
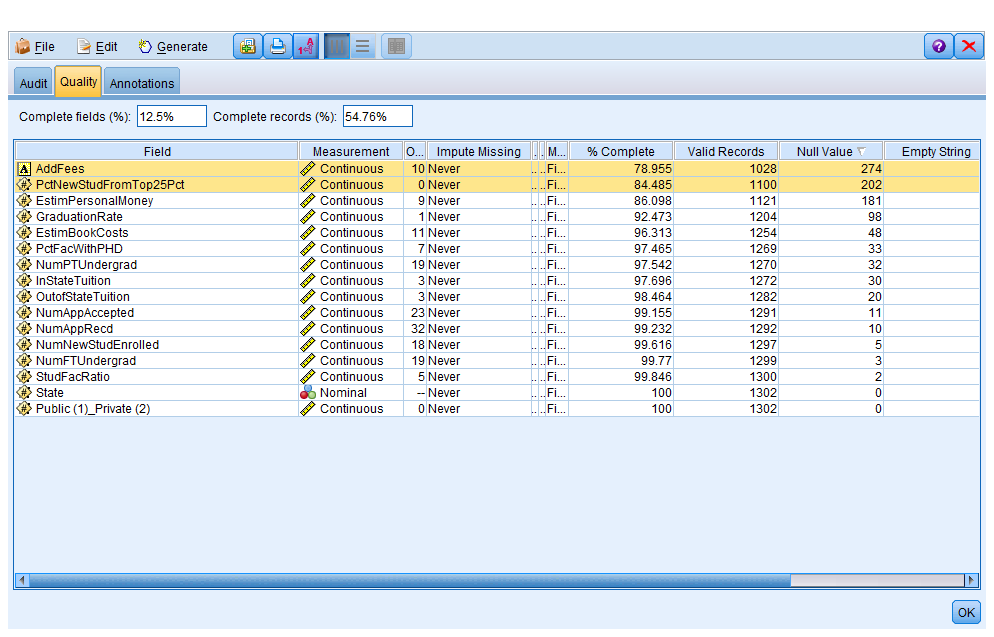
1. **Import the dataset Universities.xls.** 
   1. Use the Data Audit to examine the variables.
      1. What is the % of complete fields (variables) and % of complete records?

12.5% and 54.76 %

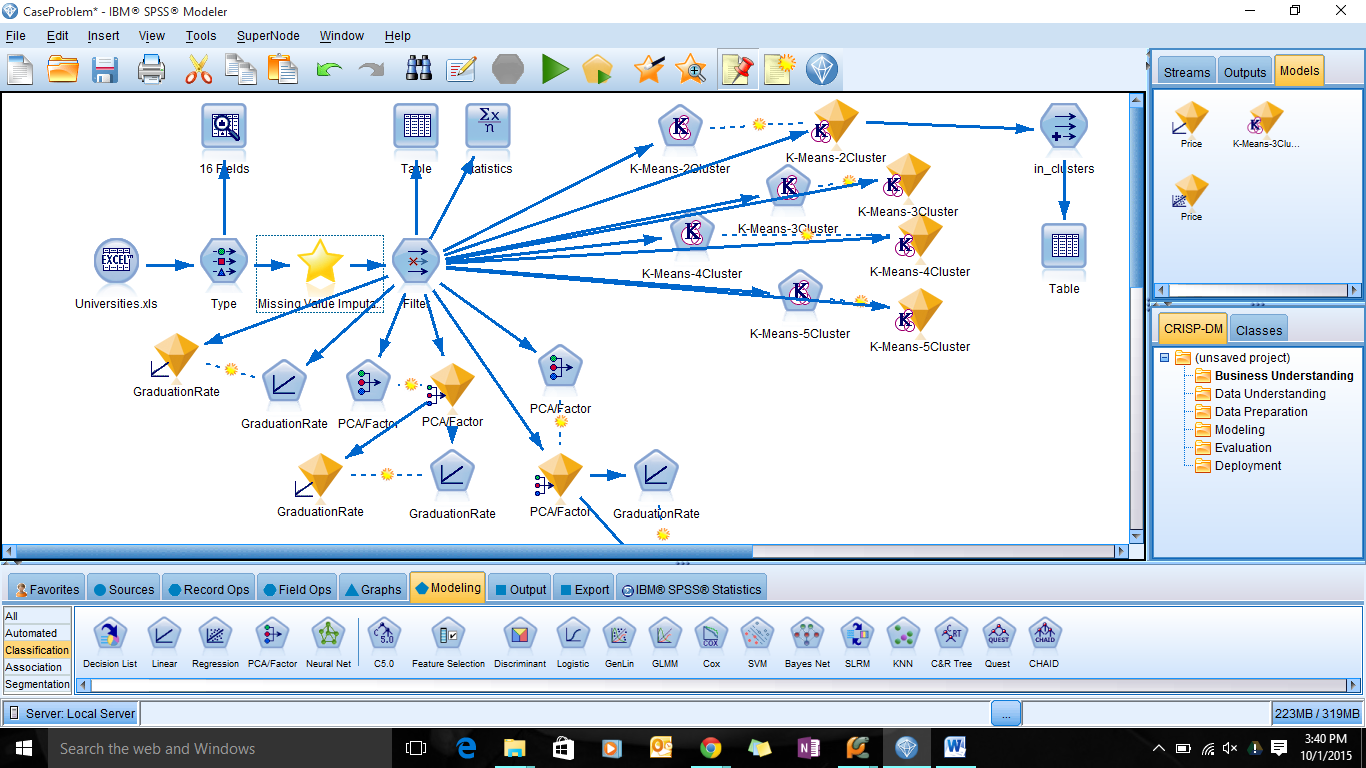


* + 1. What are the two variables with the highest number of missing values?

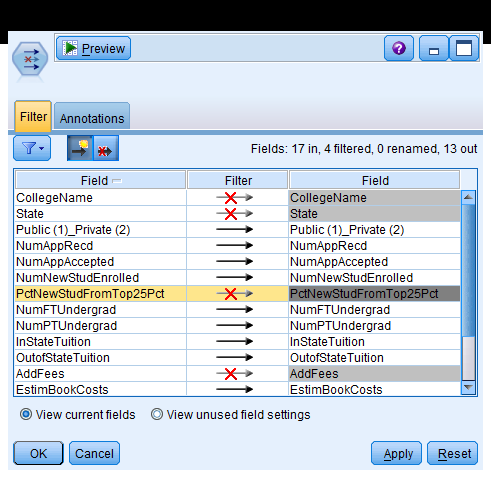
AddFees and PctNewStudFromTop25pct



* + 1. For the remainder of the variables (besides those 2) impute missing using a method of your choice. This is done under the Data Audit Quality tab. It will look something like this. Then choose Generate, Missing Value SuperNode and insert after your Type Node.



* 1. Use a type node “remove” these 2 variables along with state and CollegeName from any further analysis. Also make sure the Public/Private is a Flag variable.



1. **Cluster the Colleges.** 
   1. How many clusters do you think are a reasonable choice based on this data? Should we standardize/normalize/transform the data? Note – this causes us issues with interpretation. You could try it with and without and try to determine how much difference there is. If not much difference, keeping the original variables is easier to interpret!

There are total 13 fields after removal. From the understanding of data, I think we start by trying 4 initial clusters to see how is data spread over clusters. I think, we should normalize the data as there are lot of variables with skewed distribution and clustering algorithm requires the data to be transformed. However, we can compare the outputs to see if we achieve better results ,if we have transformed the variables.

* 1. Use cluster analysis and the k-means algorithm. Try initial 2, 3 and 4 initial clusters. Which do you think is the “best” # of clusters? Examine the cluster analysis to make your decision.

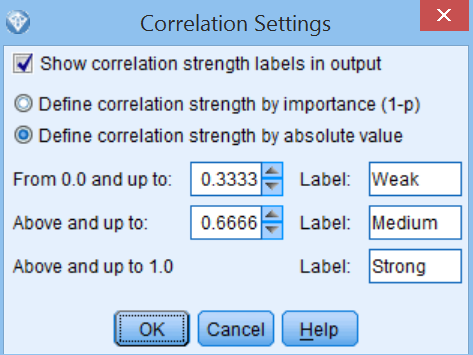
I think 3 is the best number of clusters as it provides better quality and separation between the clusters.

* 1. Interpret and “label/describe” each cluster for the analysis with your final cluster choices. You will use this information later in your analysis, you continue your stream from your chosen KMeans nugget.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Public/Private** | **NumFTUndergrad** | **In State Tuition** | **NumAPPAccepted** | **NumAppRecvd** | **Outofstate Tuition** | **Num PTundergard** |
| Cluster 1 | Private | Low | High | Low | Low | High | Low |
| Cluster 2 | Public | Medium | Low | Medium | Medium | Low | Medium |
| Cluster 3 | Public | HigSh | Low | High | High | Medium | High |

* 1. Use the statistics node to find the correlations among the variables excluding graduation rate and Public Private (and the variables you removed earlier).

Edit the Correlation Settings as shown here:



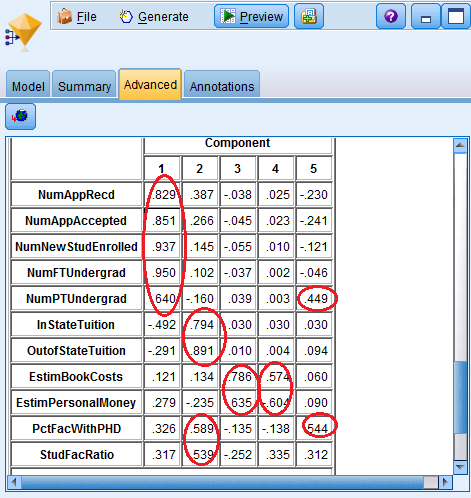
Which variables are strongly correlated?

**NumappRecd,NumAppAccepted,NumNewStudentEnrolled,NumFTUndergrad**

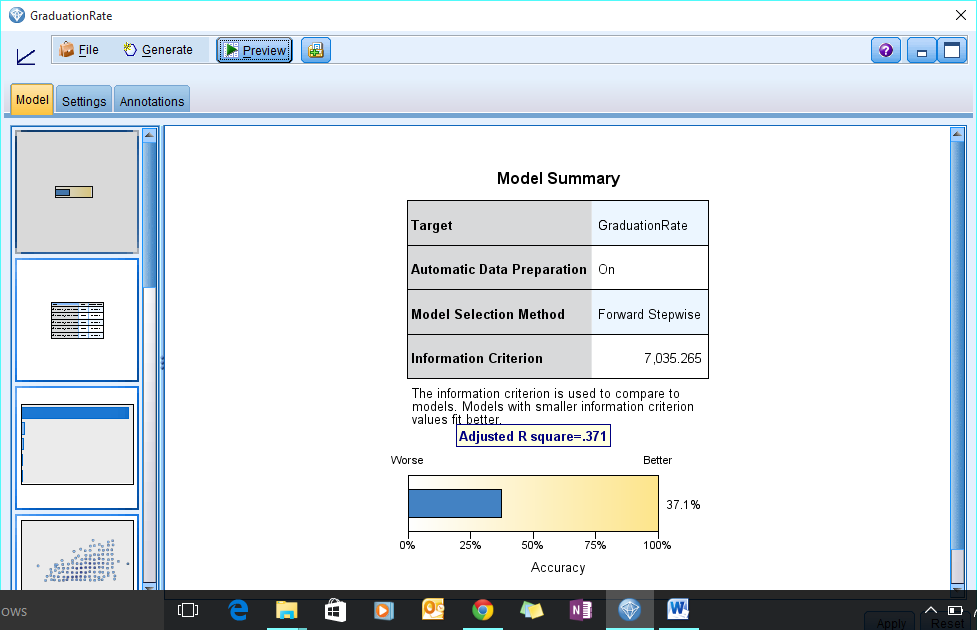
**And Outofstatetuition and instatetuition**

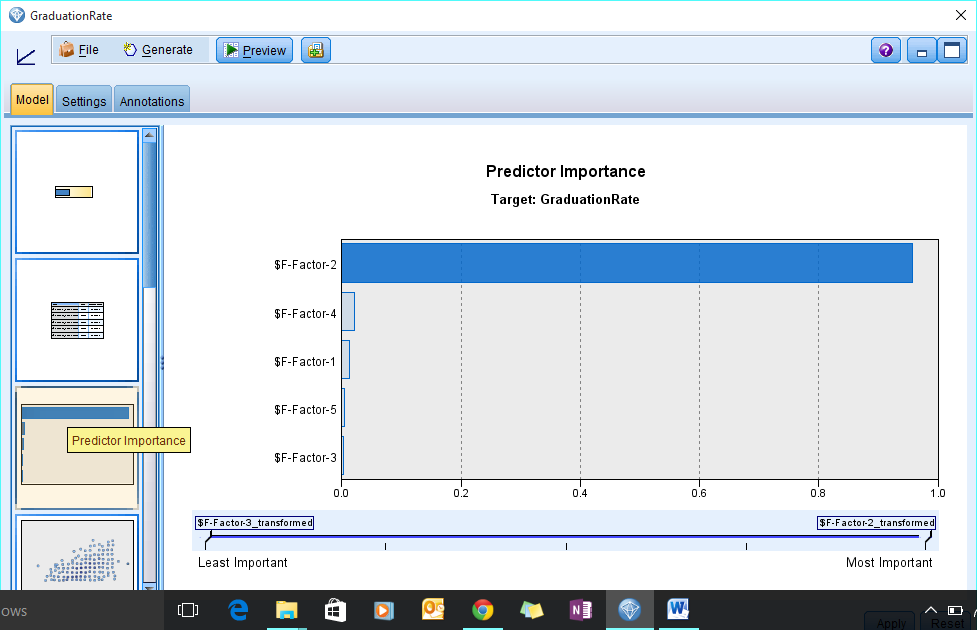
* 1. Since we are going to have issues with multicollinearity, let’s perform principal components analysis on this dataset. DO NOT include the Flag/categorical variable “Public/Private” or “graduation rate.” (That should leave you 11 variables.) How many principal components do you need to explain at least 85% of the variance in this data? Describe each component based on which variables load the most on it. Let’s use these as we continue in our analysis.

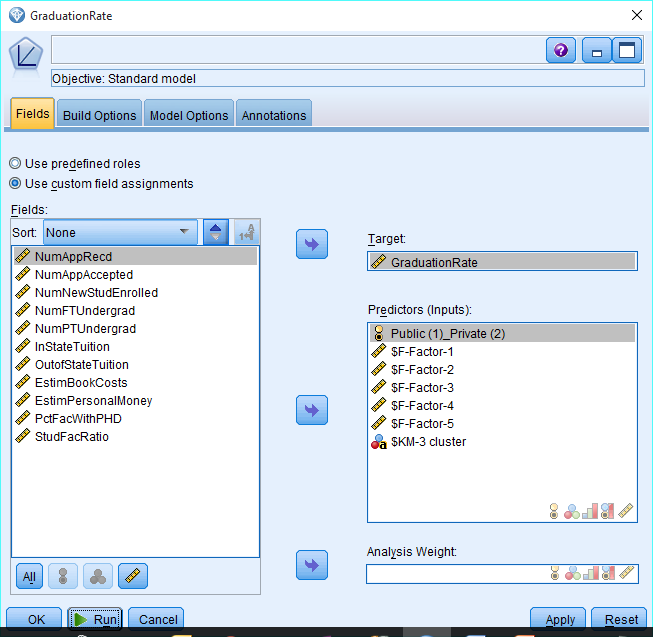
Top 5 components describes 85% of variance in the data. I have highlighted the top loads for each component in the below screenshot.

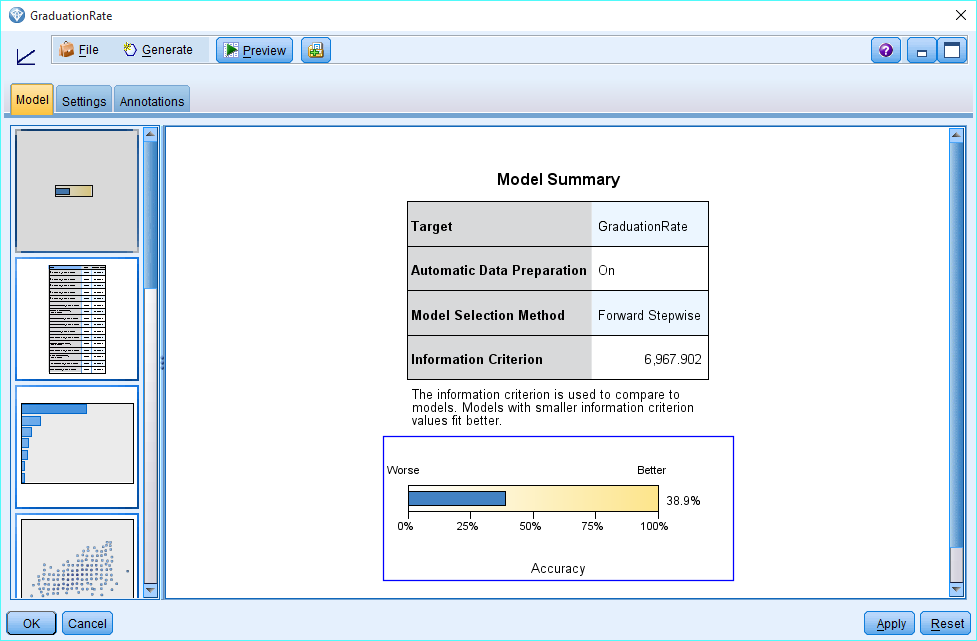


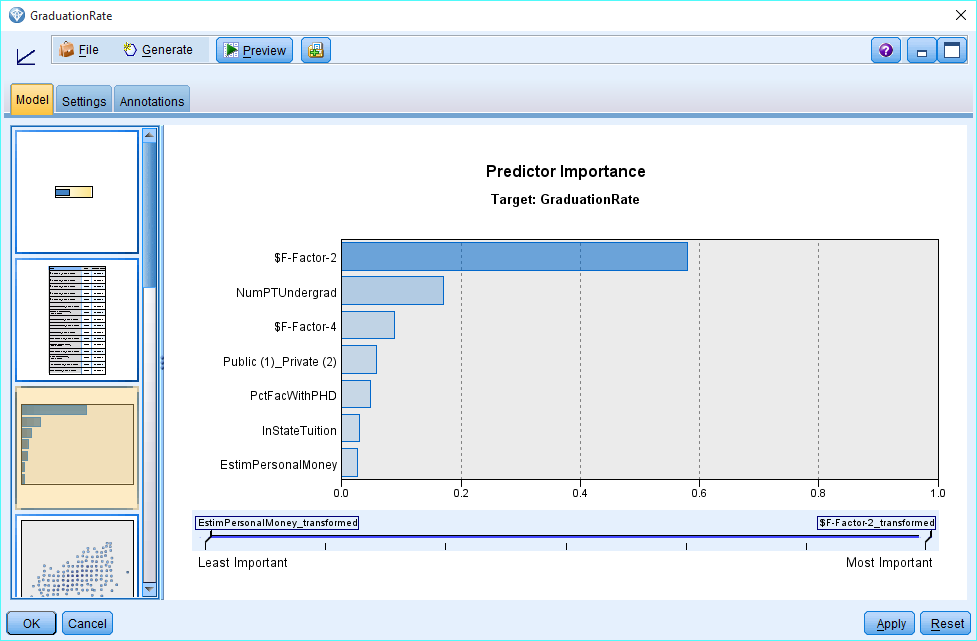
* 1. Fit three regression models (use the Node: Linear) predicting graduation rate
     1. Use the Private/Public variable and your principal components. What is the adjusted R2 value? Insert a screenshot of the Predictor Importance Output.



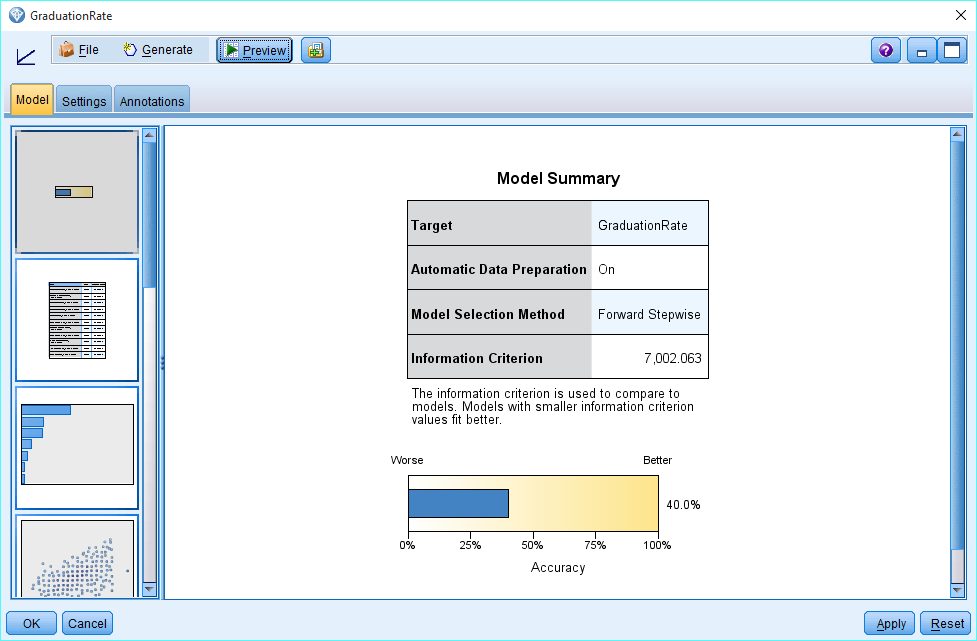


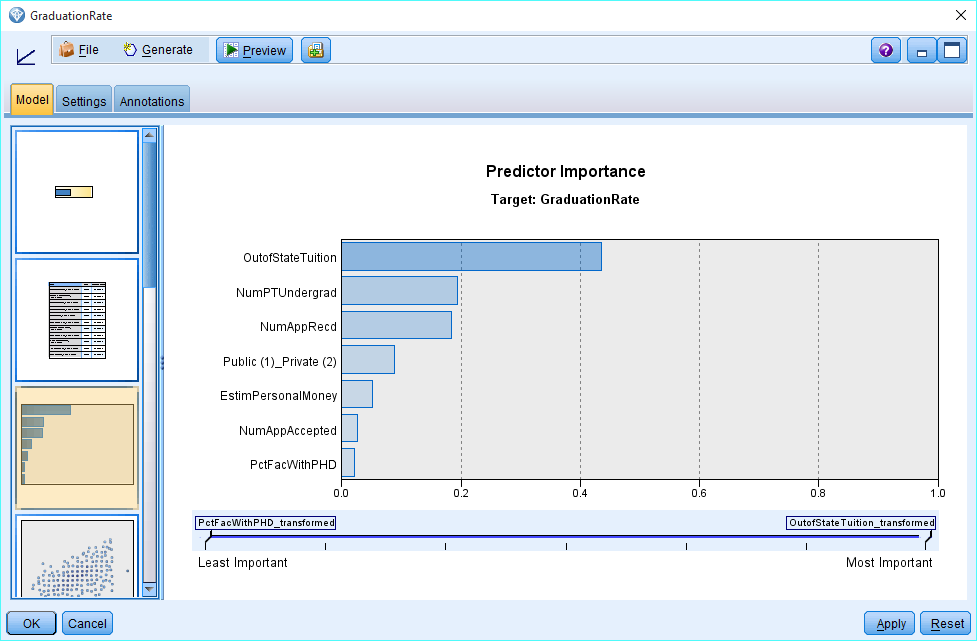
* + 1. Use the Public/Private variable and your principal components plus your cluster variable. What is the adjusted R2 value? Insert a screenshot of the Predictor Importance Output.
  1. 



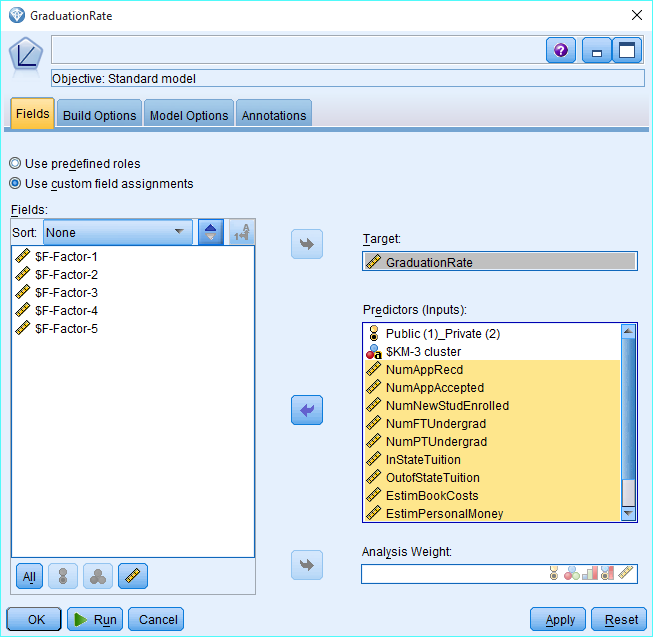


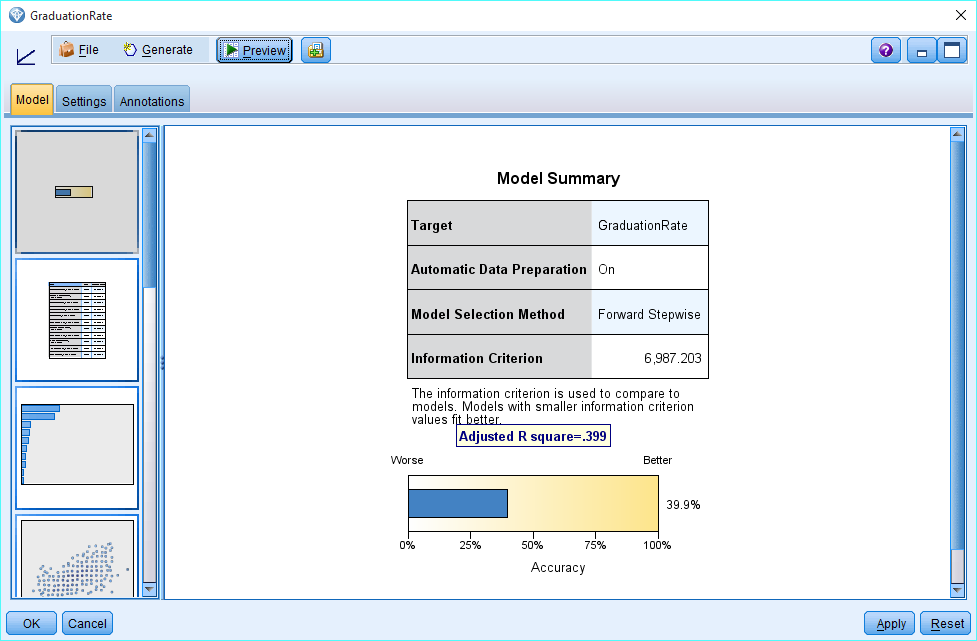
* + 1. Use the Public/Private variable and your 11 continuous variables. What is the adjusted R2 value? Insert a screenshot of the Predictor Importance Output.

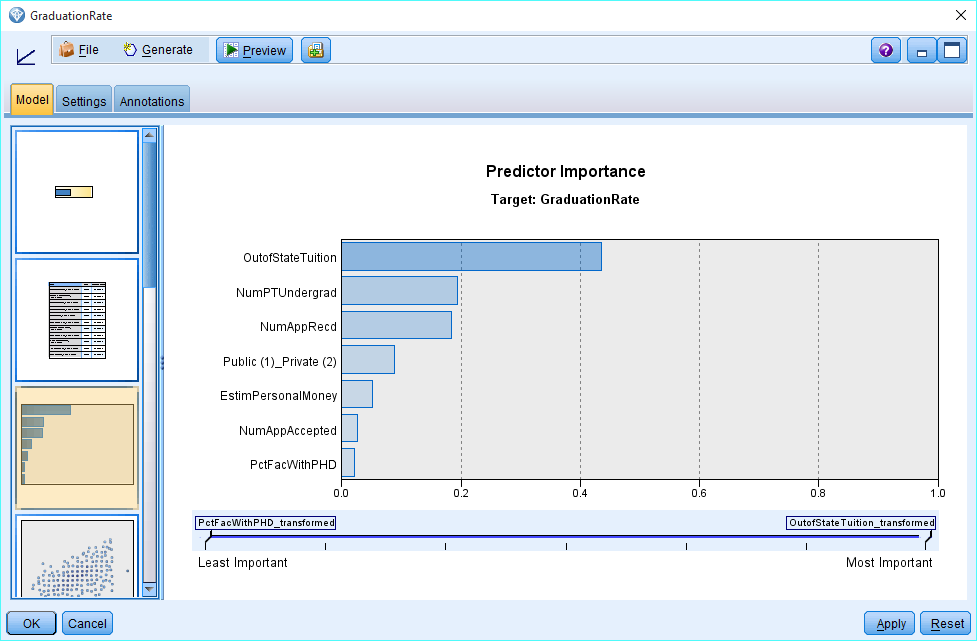




* + 1. Use the Public/Private variable and your 11 continuous variables and your cluster variable. What is the adjusted R2 value? Insert a screenshot of the Predictor Importance Output.





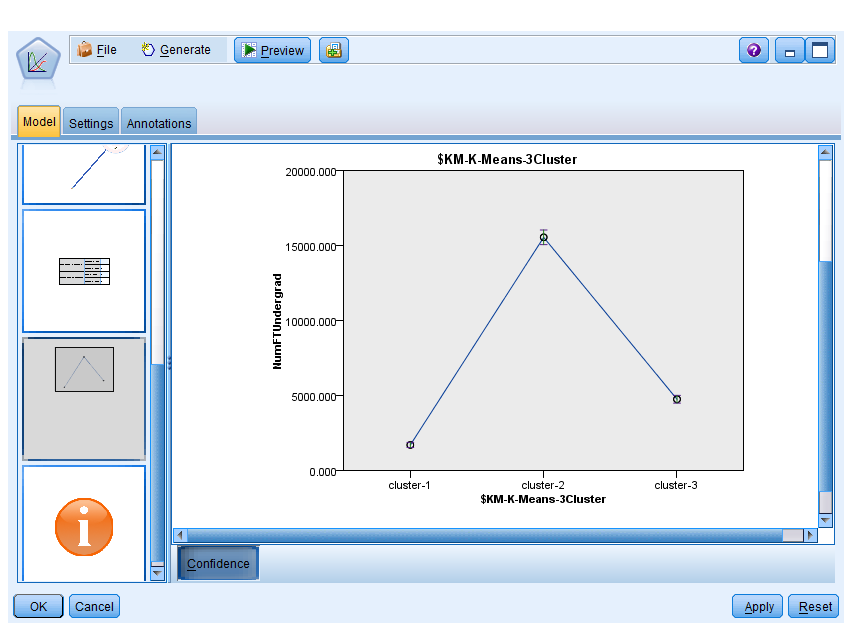
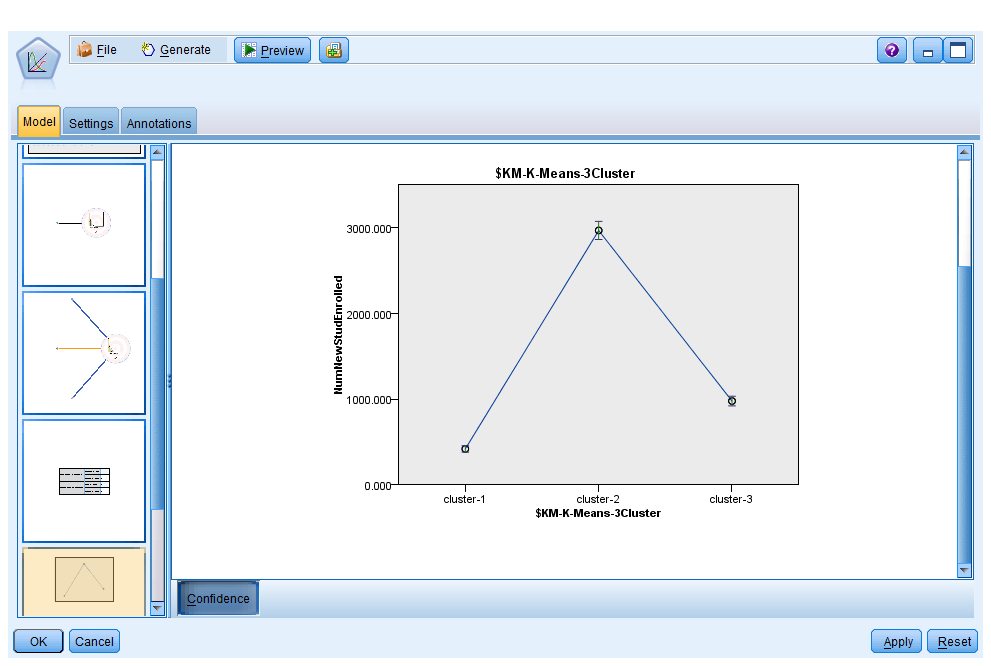
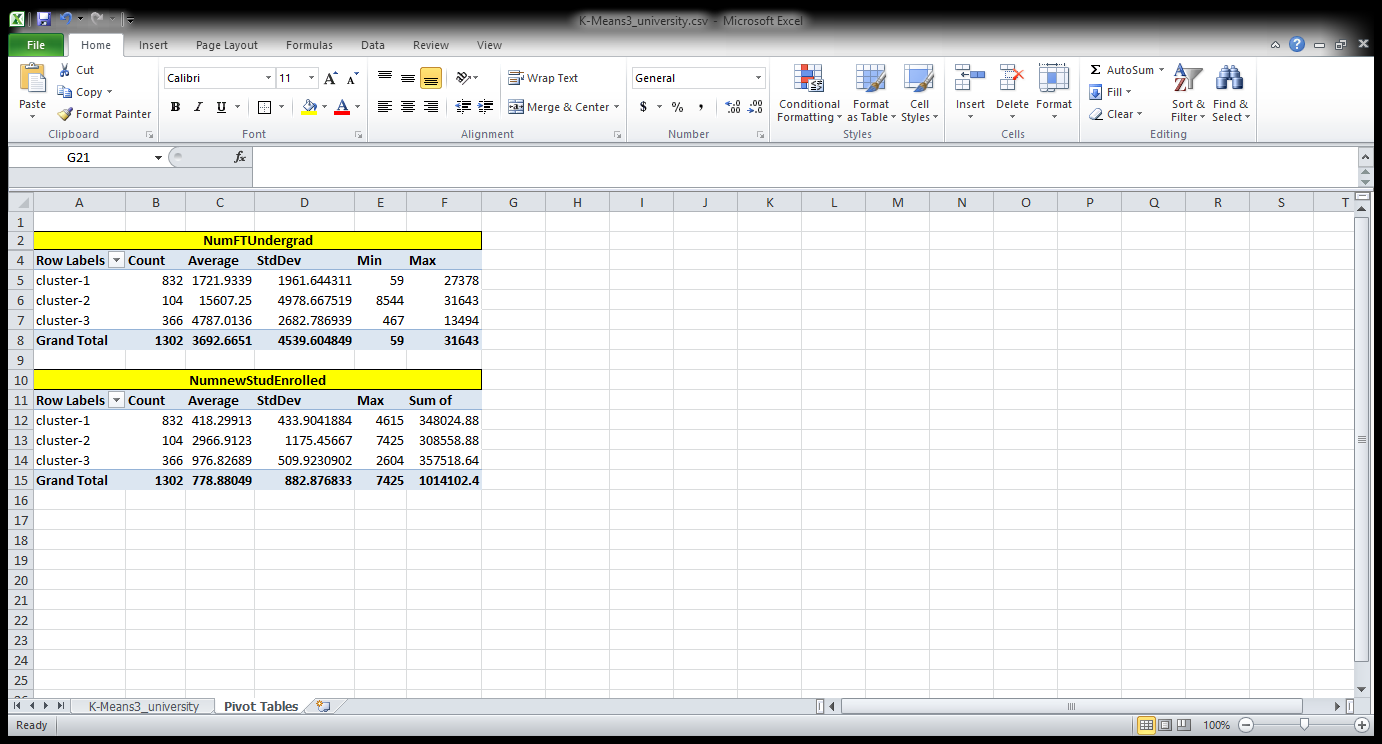


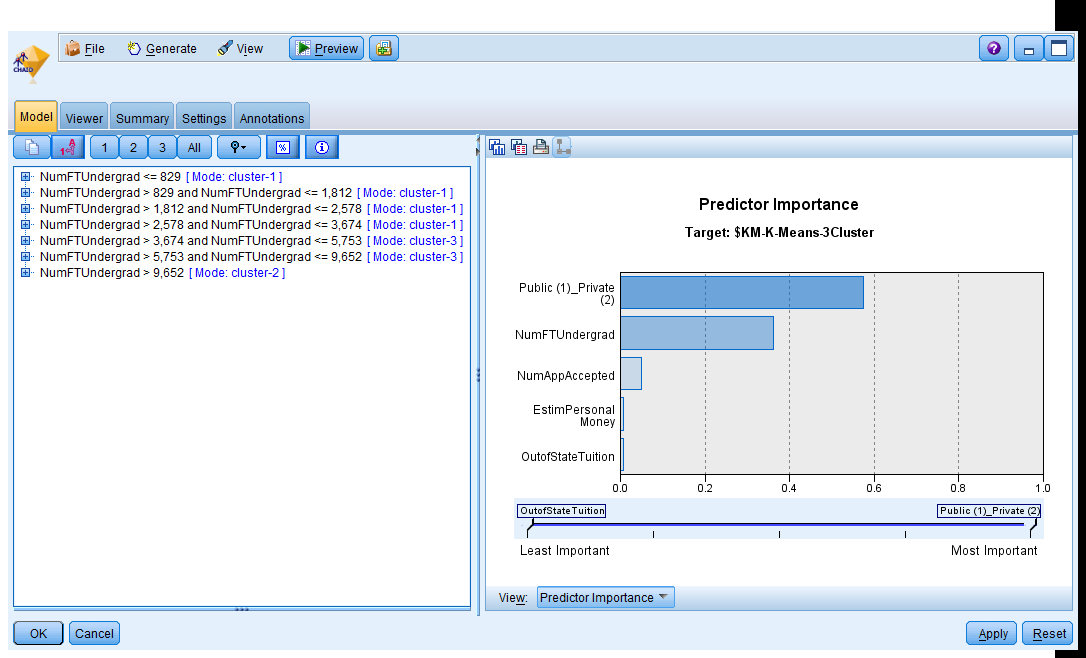
* + 1. None of these do very good – but which do you think you might pick as the ‘best.’ Do you have any other suggestions that might improve the prediction?

We get the best performance if we add all the continuous variables and public\_private variable. Max accuracy we received is 40%. So I think it is better to keep the original variables because if we use principal components then even though it would do the dimensionality reduction but it would be difficult to interpret the results. Therefore , I will choose the model with all original variables.

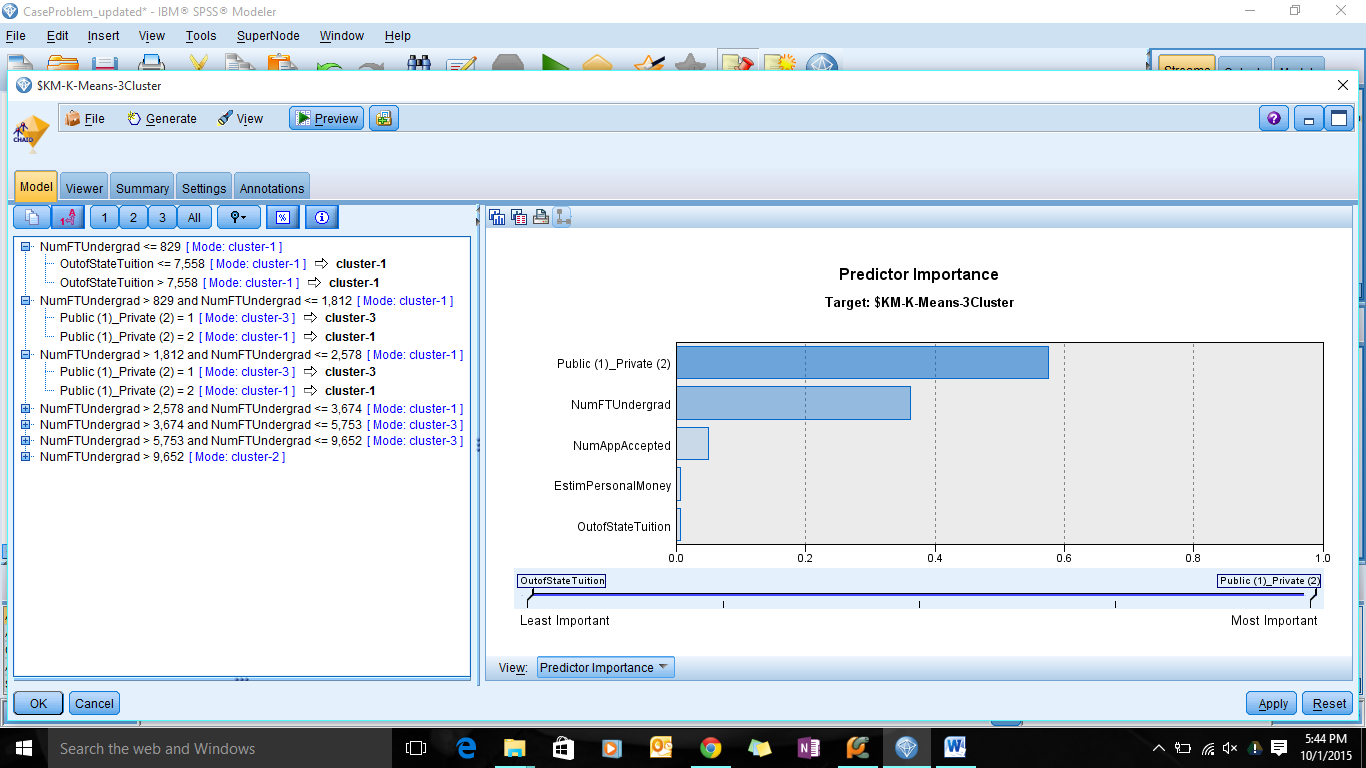
**Submit your completed Word file and your stream file on Canvas. In the comments, put the names of the folks who worked on this assignment.**

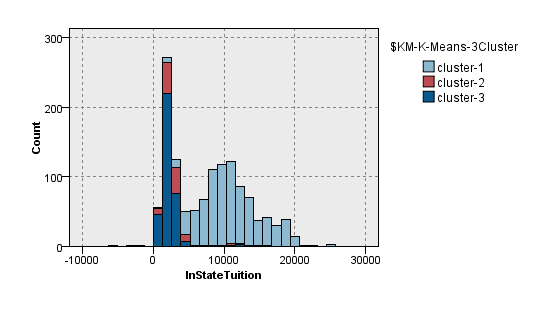
**Assignment work for M1L6:**

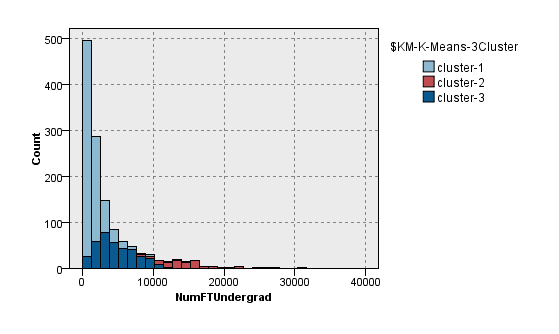
* Add from Lesson 6 notes an ANOVA comparison of at least 2 variables in your clusters.
* 
* Provide an Excel Pivot table of at least 2 variables using Excel.
  + 
* Use these to help with your labelling/interpretation of your clusters. Provide your interpretations.

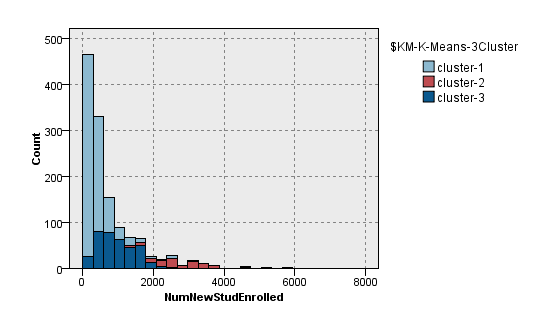
****

We can see from the decision tree that cluster 1 has NUMFTUndergard between 829 and 2578 along with other conditions as described below: For example if public\_private variable has value 2 then it is part of cluster 1, however if it has value 1 then it is part of cluster 3.



****

****

****