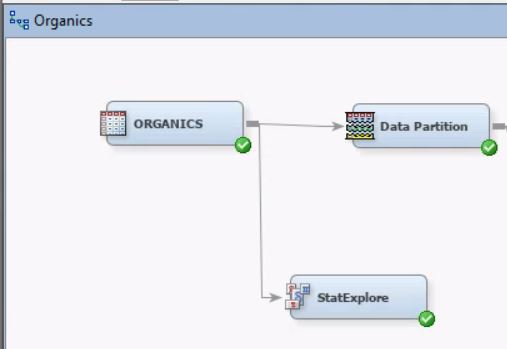
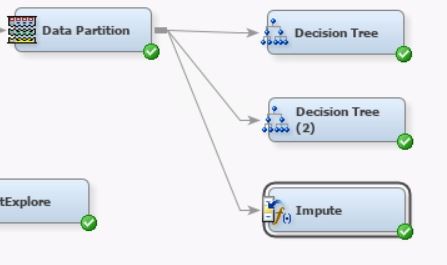
Exercise 16

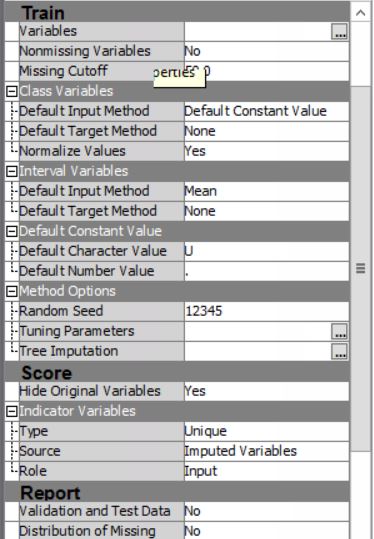
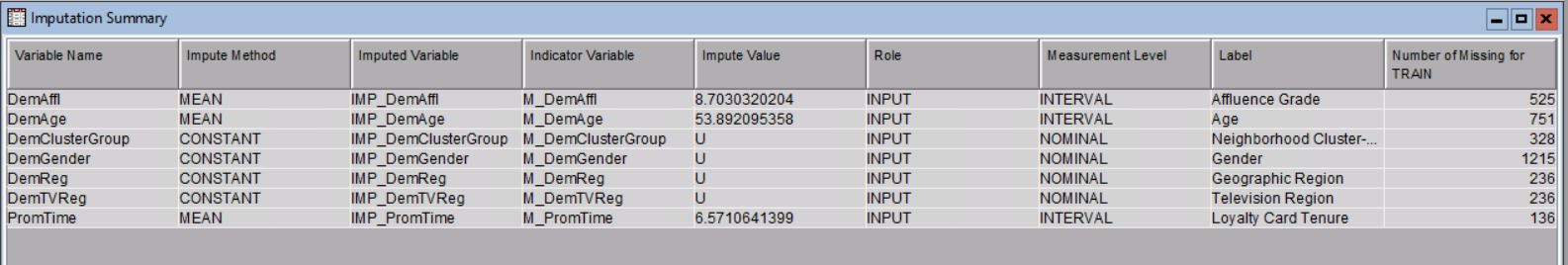
Problem 1

1. Return to the Chapter 3 Organics Diagram. Attach StatExplore tool to the **ORGANICS** data source and run it.

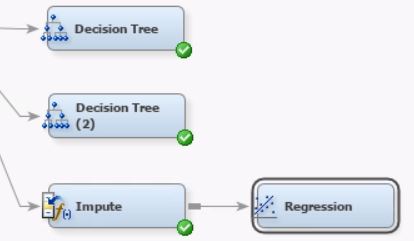


1. In preparation for regression, is any missing values imputation needed? If yes, should you do this imputation before generating the decision tree models? Why or why not?
   * Yes, because if variables in the data have too many missing values in can cause errors in the analysis and interpretation. The imputation should be done AFTER generating the decision tree models because the decision trees can identify similarities and correlations among the attributes.
2. Add an **Impute** node to the diagram and connect it to the **Data Partition** node.

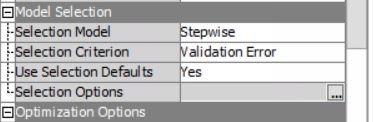




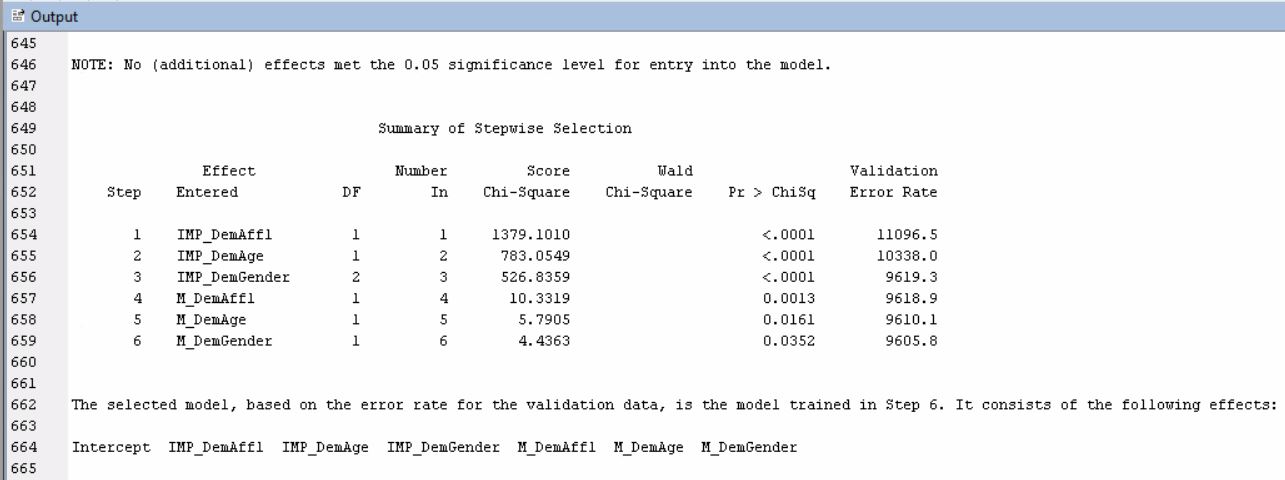
1. Add a **Regression** node to the diagram and connect it to the **Impute** node.



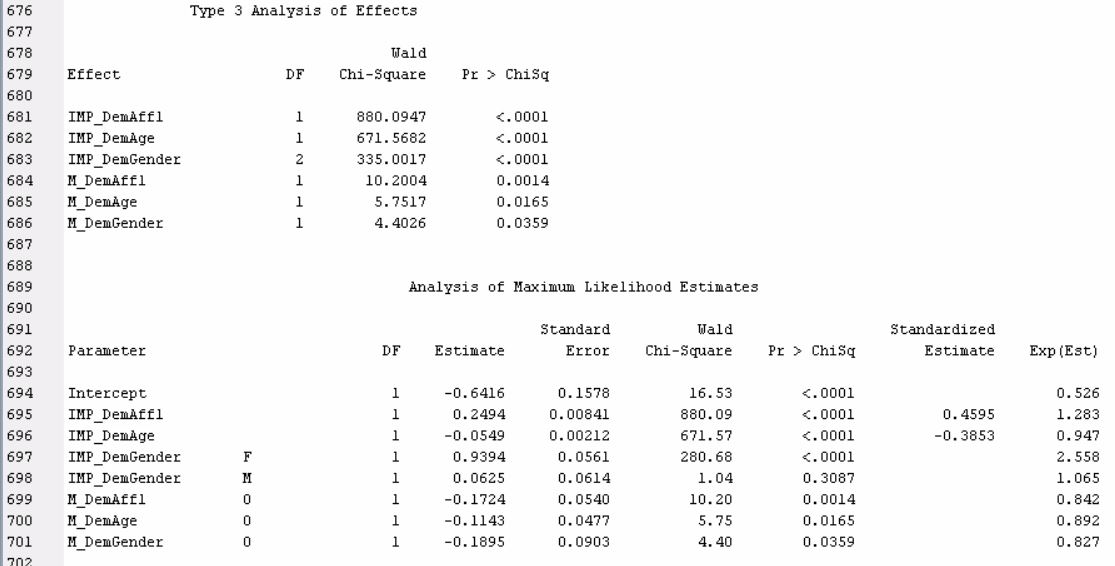
1. Choose **Stepwise** as the selection model and **Validation Error** as the selection criterion.



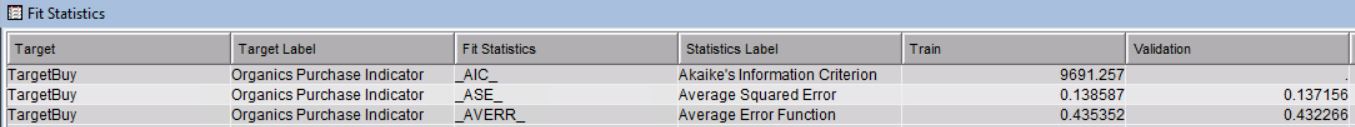
1. Run the Regression node and view the results.
   * Which variables are included in the final model?
     + IMP\_DemAffl, IMP\_DemAge, IMP\_DemGender, M\_DemAffl, M\_DemAge, M\_DemGender



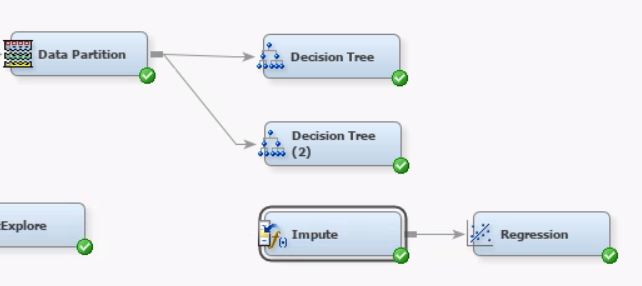
* + Which variables are important in this model?
    - The Type 3 Analysis of Effects summarizes the significance of the individual model effects. The Analysis of Maximum Likelihood Estimates (MLE), ranks the importance of inputs in the model. Based on the p-values in the MLE, all inputs are important except for IMP\_DemGender = M.



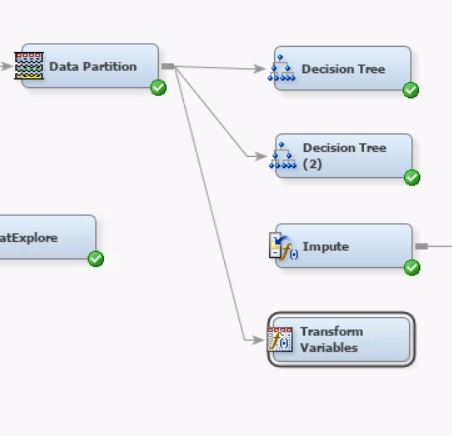
* + What is the validation ASE?
    - Average Square Error – 0.137156



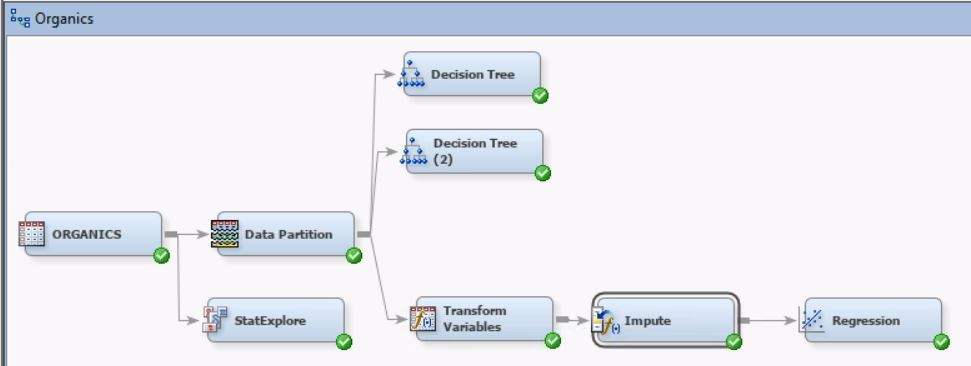
1. In preparation for regression are any transformations of the data warranted? Why or why not?
   * Yes, because some variables have skewed distributions that affect the usefulness of the model.
2. Disconnect the **Impute** node from the **Data Partition** node.



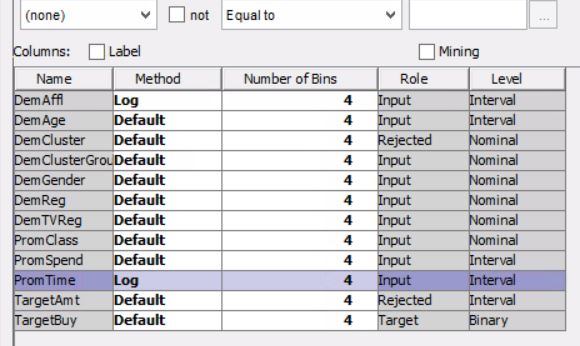
1. Add a **Transform Variables** node to the diagram and connect it to the **Data Partition** node.



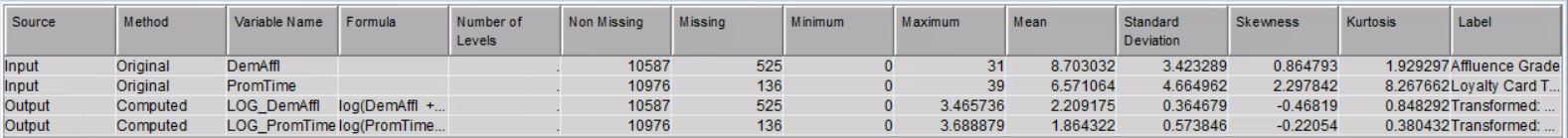
1. Connect the **Transform Variables** node to the **Impute** node.



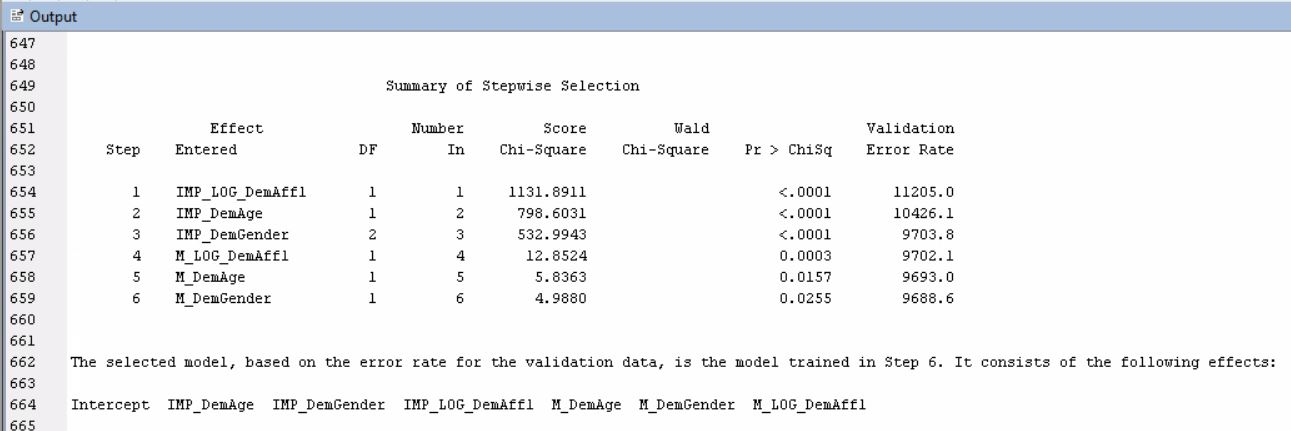
1. Apply a log transformation to the **DemAffl** and **PromTime** inputs.

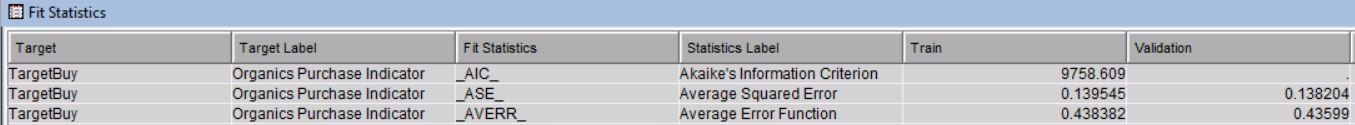


1. Run the **Transform Variables** node. Explore the exported training data. Did the transformations result in less skewed distributions?
   * Yes. The original variables are highly skewed and the log transformation of the skewed variables normalized the distributions and made the curves more symmetrical. (Based on skewness)

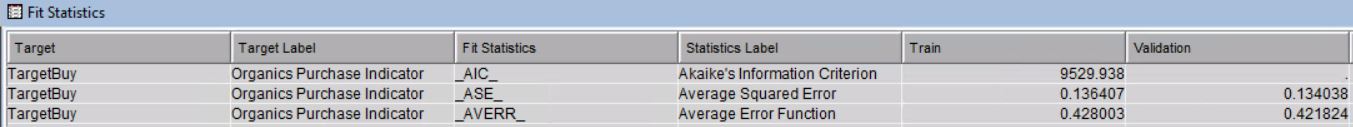


1. Rerun the **Regression** node. Do the selected variables change? How about the validation ASE?
   * Yes, the selected variables changed. The new variables are IMP\_LOG\_DemAffl, IMP\_DemAge, IMP\_DemGender, M\_LOG\_DemAffl, M\_DemAge, M\_DemGender
   * The validation ASE also changed to 0.138204





1. Create a full second-degree polynomial model. How does the validation average squared error for the polynomial mode 1 compare to the original model?
   * The ASE of the polynomial model is smaller than the original model.



1. Save the project as **Exercise 16**.