DA 6223 Exercise 16

Problem 1
Predictive Modeling Using Regression

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

b. 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?

c. Add an **Impute** node to the diagram and connect it to the **Data Partition** node. Set the node to impute U for unknown class variable values and the overall me an for unknown interval variable values Create imputation indicators for all imputed inputs.

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

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

f. Run the Regression node and view the results.

   (1) Which variables are included in the final model?

   (2) Which variables are important in this model?

   (3) What is the validation ASE?

g. In preparation for regression are any transformations of the data warranted? Why or why not?

h. Disconnect the **Impute** node from the **Data Partition** node.

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

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

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

l. Run the **Transform Variables** node. Explore the exported training data. Did the transformations result in less skewed distributions?

m. Rerun the **Regression** node. Do the selected variables change? How about the validation ASE?

n. Create a full second degree polynomial model. How does the validation average squared error for the polynomial mode l compare to the original model?

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