**Project Proposal Instructions**

Respond to the following in the three project sections:

1. **Predictive Modeling Discussion**. Do not state the specific models you will use, but you MUST state:
   1. Are you working on a REGRESSION or CLASSIFICATION problem?
   2. Which variables are inputs?
   3. Which variables are responses/outputs/outcomes/targets?
   4. Did you need to DERIVE the responses of interest by SUMMARIZING the available data?
   5. If so, what summary actions did you perform?
   6. Which variables are identifiers and should NOT be used in the models?
   7. Important: Answer this question after completing parts C and D. Return to this predictive modeling discussion section to answer the following:
      1. Which of the inputs do you think influence the response, based on your exploratory visualizations? Which exploratory visualization helped you identify potential input-to-output relationships? (If you are not sure which inputs seem to influence the response, it is okay to say so.)

NOTE: Module 1’s Course Goals item included the Data Science Supervised Learning Overview, which discusses the differences between REGRESSION and CLASSIFICATION.

1. **Exploratory Data Analysis**. Complete these general requirements:
   1. You must read in the data associated with your project.
   2. You must perform the ESSENTIAL Exploration activities:
      1. Display the number of rows and columns.
      2. Display the column names and their associated data types.
      3. Display the number of missing values for each column.
      4. Display the number of unique values for each column.
   3. You must state whether you want to effectively treat a numeric column as a non-numeric for exploration purposes.
   4. You do NOT need to display the COUNTS for categorical variables because you will visually display the variables. However, you may display/print COUNTS if it helps you.
   5. You must visualize the MARGINAL distributions for ALL variables in your data.
      1. Continuous variables: Histograms or density plots.
      2. Categorical variables: Bar charts.
   6. You must decide appropriate visualizations to show:
      1. Categorical-to-categorical relationships (COMBINATIONS).
      2. Categorical-to-continuous relationships.
      3. Continuous-to-continuous relationships.
   7. You must decide appropriate visualizations to show if RELATIONSHIPS change across GROUPS. This is especially important to continuous-to-continuous relationships. Do not forget though that categorical-to-continuous relationships can also be GROUPED BY a secondary categorical variable!
   8. If you are working on a REGRESSION problem, you must:
      1. Visualize the relationship between the continuous response and the continuous inputs using scatter plots and trend plots.
      2. Group the relationships by categorical variables (if appropriate).
      3. Summarize the response for each unique value of the categorical inputs using boxplots, violin plots, and point plots.
   9. If you are working on a CLASSIFICATION problem, you must:
      1. Visualize conditional distributions of the continuous inputs GROUPED BY the response (outcome) unique values.
      2. Visualize relationships between continuous inputs GROUPED BY the response (outcome) unique values.
      3. Visualize the counts of combinations between the response (outcome) and categorical inputs.
2. **Cluster Analysis**.

NOTE: This is an initial attempt at cluster analysis to help you explore the data. You will also perform cluster analysis for the final project submission. The approach you use for the final project might be different from what you try now, but that is okay! Remember that this is just an initial attempt. See what you learn! If it does not seem to reveal anything “useful,” remind yourself that that is okay, and part of the learning process!

You will NOT apply cluster analysis using ALL variables. Cluster analysis is UNSUPERVISED. It does NOT make distinctions between INPUTS and OUTPUTS. However, the projects are PREDICTIVE ANALYTICS problems with INPUTS and OUTPUTS. Therefore, you will need to decide the class of variables on which you will focus in your cluster analysis.

Complete the following setup steps before the requirements that follow.

* 1. Select variables. You MUST state which class of variables you used in your cluster analysis. For example, you could:
     1. Cluster using CONTINUOUS inputs.
     2. Cluster using CONTINUOUS responses (if you have multiple outputs).
     3. Cluster using multiple summary statistics derived from inputs. For example, if your data includes MANY categorical variables consider GROUPING BY them and SUMMARIZING the OTHER inputs. Summary statistics to consider: number of unique values, mean, standard deviation, median, etc.
        1. If you do this, do NOT start by creating COMBINATIONS of ALL categorical inputs. Select a few which provide many COMBINATIONS.
     4. Cluster using multiple summary statistics derived from the outputs. For example, if your data includes MANY categorical variables consider GROUPING BY them and SUMMARIZING the outputs. Summary statistics to consider: number of unique values, mean, standard deviation, median, etc.
        1. If you do this, do NOT start by creating COMBINATIONS of ALL categorical inputs. Select a few which provide many COMBINATIONS.
  2. Consider the following for your cluster analysis:
     1. How many variables are you using relative to the number of observations?
     2. Are the variables you are using “Gaussian-like” or are the distributions “odd looking”? You must visualize the MARGINAL histograms of the variables IF you are using summary statistics derived from INPUTS or OUTPUTS. If you are using the INPUTS or OUTPUTS directly, you already visualized the MARGINAL distributions.
     3. Are the variables you are using highly correlated to each other? You must visualize the relationships between the variables you are using. If you are using the INPUTS/OUTPUTS directly then you already visualized those relationships.
     4. How should you handle missing values? If you drop all rows with at least one missing value, how many observations would you still have (thus, how many complete cases do you have)?

Now that you have completed the setup steps, complete the following requirements:

* 1. You must use KMeans to execute the Cluster analysis.
  2. First, use 2 clusters and:
     1. Count the number of observations per identified cluster. Are the clusters balanced?
     2. Visualize the cluster results between 2 of the variables used for the cluster analysis.
  3. Then, identify the optimal number of clusters, and rerun KMeans for the optimal number.
     1. Count the number of observations per identified cluster.
     2. Visualize the cluster results between 2 of the variables used for the cluster analysis.
  4. Interpret:
     1. Do the OPTIMAL number of clusters ALIGN with CATEGORIES of KNOWN grouping variables in your data?
        1. If you are working on a CLASSIFICATION problem, are the identified clusters consistent with the outcome categories?
        2. If you are working on a REGRESSION problem, are the identified clusters consistent with categories of a categorical input?
     2. What are the CONDITIONAL distributions of the variables you used for clustering GIVEN the identified clusters?
        1. Use appropriate visualizations to study the CONDITIONAL distributions of the variables used for clustering GROUPED BY the identified clusters.
        2. HINT: you are creating CONDITIONAL distributions using the DERIVED cluster groupings rather than using a “real” categorical variable.
        3. Describe the clusters based on the CONDITIONAL distributions.