# Week 8 Assignment

**Assignment:** You can complete one of the two parts to this assignment, or both. If you do both parts, you can obtain extra points for this assignment. You are expected to complete at least one part successfully.

**Data Files:** Excel File <*CaliforniaCabernet.xlsx*> -> contain over 13K reviews of California Cabernet Sauvignon. The reviews are in the column labeled 'description'. The full data dictionary is:

review: A number unique for each review (an ID)
description: The actual review (text)
year: Year the wine was bottled. This is missing for some wines.
points: The points assigned by the reviewer to the wine. These range from 80 to 100. Better reviews have higher points.
price: The retail price for a bottle of the wine ($0-$3000).
winery: The winery where the wine was bottled. (a text label)
Region: Region of California (text) where wine was produced.

There are no outliers in these data, but many of the years are missing. For this assignment there is no need to impute missing values.

**Part 1:** Create a SAS EM project. Store the Excel data file for this assignment in your DataSources directory. When you import this file into your project, ensure the "Role" of the column "description" is set to "Text" with "Level" as "Nominal". These are the wine reviews we will be analyzing.

1. Conduct a text classification analysis of these data following the analytical process described in class.
2. Use POS, stop words and stemming for building the term/doc matrix.
3. Use TF-IDF weighting for the term-doc matrix.
4. Use text-cluster node to develop exactly 9 clusters for the wine reviews.
5. Develop a table showing the average points and average price for wines in each cluster. Include the 15 words that describe the clusters.

Solution Summary:

1. Report a screen shot of the diagram, the property windows for all text analytics nodes in the diagram.
2. Identify and describe the topic groups.
3. Report a Table of average points and price for each topic group.
4. Provide a professional summary describing 1) The data and your goals for this analysis, 2) How the data were analyzed, 3) The topic groups you discovered and 4) your professional opinion of this analysis.

**Part 2:** Do the same assignment as Part 1 using Python.

Follow the process described in this week's notes. Use pandas to read the file. Use LDA to identify the top 9 topics.

Analysis:

1. Conduct a text classification analysis of these data following the analytical process described in class.
2. Use POS, stop words and stemming for building the term/doc matrix. These three actions are the default conditions in *TextAnalytics.analyzer*.
3. Use TF-IDF weighting for the term-doc matrix.
4. Use LDA to develop exactly 9 clusters for the wine reviews.
5. Develop a table showing the average points and average price for wines in each cluster. Include the 15 words that describe the clusters.

Solution Summary:

1. The Python code (.py file) for text and topic analysis.
2. Show all output from running the code.
3. Show Table of average points and price for each topic group.
4. Provide a professional summary describing 1) The data and your goals for this analysis, 2) How the data were analyzed, 3) The topic groups you discovered and 4) your professional opinion of this analysis.