# MNGT 379 - Business Analytics

## Module 4 Descriptive Data Mining Homework

This document describes the steps required to complete this assignment; read each question ***carefully*** and make sure that you follow each directionprecisely as described. When you have finished each instruction in this document, you will upload the completed document to the Assignment on D2L. Don’t forget to save your copy of this document (I suggest the Documents folder) early and often! Don’t risk losing your work!

**Problem 1**

The file BigTen.xlsx contains data on a number of university athletic programs in the United States. Follow the instructions below to complete a *k*-Means Clustering analysis of the data.

*Problem 1.1* In Step 1 of ASP’s *k*-Means Clustering, select StadiumCapacity, Latitude, Longitude, Endowment ($000), and Enrollment as your variables. Select . In Step 2, check the box for Normalize input data, set # Clusters to 10, # Iterations to 50, and Random starts to 10. Click Finish. Take a screenshot (using the Snipping Tool, Grab, or a keyboard shortcut) of the **Inputs** table (Range B10:J31) and paste it in the space below:

*Problem 1.2* ASP will generate two new worksheets named  and ; rename them to  and  respectively. Take a screenshot of the worksheet list and paste it in the space below:

*Problem 1.3* On the  worksheet, scroll down to the **Cluster Centers** area (row 165). By default, Excel will display the data in the **Cluster Centers**, **Inter-Cluster Distances**, and **Cluster Summary** tables to a large number of decimal places, but this is somewhat messy. Use the Decrease Decimal Tool  to reduce the number of displayed decimal places to two e.g. . Apply this formatting to all three aforementioned tables.

*Problem 1.4* Next, apply Conditional Formatting, Blue – White – Red Color Scales format to the **Cluster Centers** and **Inter-Cluster Distances** tables. Make sure that you apply the formatting to each table separately, so each uses its own scale. Take a screenshot of both tables (separately) and paste them in the space below each table name below:

 *Cluster Centers Table*

*Inter-Cluster Distances Table*

*Problem 1.5* Similarly, apply the same Conditional Formatting to the Average Distance column of the **Cluster Summary** table. Be sure to NOT include the Size column or it will affect your results. Take a screenshot of the table and paste it in the space below:

*Problem 1.6* Characterize Cluster 1 by describing the cluster centers for each variable. In other words, describe each variable in Cluster 1 by if it the average value for each variable is average, below average, or above average, and by how much (similar to how I did in my explanatory video). E.g. Cluster 1 has a slightly above-average Stadium Capacity, moderately below-average Latitude, and so on. Enter your response in the space below.

*Problem 1.7* What is the largest (either most positive or most negative) z-value in the **Cluster Centers** chart? Report the value of the z-value itself, to which variable it relates, and in what cluster the value appears.

 *Z-Value*:

 *Associated Variable*:

 *Associated Cluster*:

*Problem 1.8* Look at the red values in the **Inter-Cluster Distances** table; why is there a string of values that are zeros – is that strange?

*Problem 1.9* Examine the **Cluster Summary** table; one of the clusters is actually *impossible* to occur – something must be wrong, but what? To answer this question, you must first identify which cluster is the problem. Once you have an answer, go to the  sheet and use the filter button  in the Cluster column to display only the problem cluster (uncheck the Select All option, check the box next to the suspected cluster, then click OK). Note the Record ID numbers showing on the row labels, then return to the  worksheet and inspect the schools matching the Record ID numbers you noted previously (in other words, if Record 10 appears in the first column of your filtered table, then look at the school with ID #10, which turns out to be the Kansas State Wildcats). What seems to be the problem? Enter your response in the answer below:

*Problem 1.10* Run the *k*-Means Clustering procedure again, but this time do NOT Normalize Input Data. As before, take a screenshot of the **Inputs** table (Range B10:J31) and paste it in the space below:

*Problem 1.11* What is the biggest difference between the results from this version from the Normalized version? You may wish to use the formatting guidelines you implemented on the previous version to aid you in your search. *Hint: the answer is NOT that all the numbers are bigger*.

**Problem 2**

The file AppleCart.xlsx contains data on 2000 transactions from the local Apple Store involving iPads. Follow the instructions below to complete an Association analysis of the data.

*Problem 2.1* In Step 1 of ASP’s Association procedure, make sure that the correct Data range is selected and that the First Row Contains Headers box is checked. Also verify that the Input Data Format is set to Data in binary matrix format. Finally, set the Minimum confidence (%) to 45, then select . Take a screenshot of the **Rules** table (Range B37:J59) and paste it in the space below:

*Problem 2.2* Describe the first Association Rule in “if-then customers are likely to” format.

*Problem 2.3* Use the Filter button  to sort the chart in order of confidence from Largest to Smallest. What do you notice about the first half-dozen rules? Do they have something in common? What does this suggest about customers’ prioritization of the accessories and features?

*Problem 2.4* Use the Filter button  again, this time to sort the chart by Lift-Ratio. Again consider the first few rules; do they suggest any interesting relationships? *Hint: don’t just describe a particular Association Rule, but look for a trend across Association Rules.*