COMPETENCIES

4030.6.4 : **Clustering Techniques**
The graduate applies clustering techniques to accurately predict outcomes of interest.

INTRODUCTION

In this task, you will act as an analyst and create a data mining report. You must select one of the data dictionary and data set files to use for your report from the following web link: "Data Sets and Associated Data Dictionaries."

You will use Python or R to analyze the given data and create a data mining report in a word processor (e.g., Microsoft Word). Throughout the submission, you must visually represent each step of your work and the findings of your data analysis.

Note: All algorithms and visual representations used need to be captured either in tables or as screenshots added into the submitted Word document. A separate Microsoft Excel (.xls or .xlsx) document of the cleaned data should be submitted along with the written aspects of the data mining report.

SCENARIO

Scenario 1
One of the most critical factors in customer relationship management that directly affects a company's long-term profitability is understanding the customers. When a company understands its customers' characteristics, it is better able to target products and marketing campaigns for customers, resulting in better profits for the company in the long term.

You are an analyst for a telecommunications company that wants to better understand the characteristics of its customers. You have been asked to use clustering techniques to analyze customer data to identify groups of customers with similar characteristics, ultimately enabling better business and strategic decision-making.

Scenario 2
One of the most critical factors in patient relationship management that directly affects a hospital's long-term cost-effectiveness is understanding the patients and the conditions leading to hospital admissions. When a hospital understands its patients' characteristics, it is better able to target treatment to patients, resulting in a more effective cost of care for the hospital in the long term.

You are an analyst for a hospital that wants to better understand the characteristics of its patients. You have been asked to use clustering techniques to analyze patient data to identify groups of patients with similar characteristics, ultimately enabling better business and strategic decision-making for the hospital.

REQUIREMENTS

Your submission must be your original work. No more than a combined total of 30% of the submission and no more than a 10% match to any one individual source can be directly quoted or closely paraphrased from sources, even if cited correctly. The similarity report that is provided when you submit your task can be used as a guide.

You must use the rubric to direct the creation of your submission because it provides detailed criteria that will be used to evaluate your work. Each requirement below may be evaluated by more than one rubric aspect. The rubric aspect titles may contain hyperlinks to relevant portions of the course.

*Tasks may **not** be submitted as cloud links, such as links to Google Docs, Google Slides, OneDrive, etc., unless specified in the task requirements. All other submissions must be file types that are uploaded and submitted as attachments (e.g., .docx, .pdf, .ppt).*

Part I: Research Question

A. Describe the purpose of your data mining report by doing the following:

1. Propose **one** question relevant to a real-world organizational situation that you will answer using **one** of the following clustering techniques:
   - *k*-means, using only continuous variables
   - hierarchical
2. Define **one** goal of the data analysis. Ensure your goal is reasonable within the scope of the selected scenario and is represented in the available data.

Part II: Technique Justification

B. Explain the reasons for your chosen clustering technique from part A1 by doing the following:

1. Explain how the clustering technique you chose analyzes the selected data set. Include expected outcomes.
2. Summarize **one** assumption of the clustering technique.
3. List the packages or libraries you have chosen for Python or R, and justify how *each* item on the list supports the analysis.

Part III: Data Preparation

C. Perform data preparation for the chosen data set by doing the following:

1. Describe **one** data preprocessing goal relevant to the clustering technique from part A1.
2. Identify the initial data set variables you will use to perform the analysis for the clustering question from part A1, and label *each* as continuous or categorical.
3. Explain *each* of the steps used to prepare the data for the analysis. Identify the code segment for *each* step.
4. Provide a copy of the cleaned data set.

Part IV: Analysis

D. Perform the data analysis, and report on the results by doing the following:
   1. Determine the optimal number of clusters in the data set, and describe the method used to determine this number.
   2. Provide the code used to perform the clustering analysis technique.

Part V: Data Summary and Implications

E. Summarize your data analysis by doing the following:
   1. Explain the quality of the clusters created.
   2. Discuss the results and implications of your clustering analysis.
   3. Discuss **one** limitation of your data analysis.
   4. Recommend a course of action for the real-world organizational situation from part A1 based on the results and implications discussed in part E2.

Part VI: Demonstration

F. Provide a Panopto video recording that includes the presenter and a vocalized demonstration showing all code used, the code being executed, and the results of all code used in the task.
   1. Include the presenter and a vocalized demonstration describing the programs used to complete this task in the Panopto video recording.

   Note: The audiovisual recording should feature you visibly presenting the material (i.e., not in voiceover or embedded video) and should simultaneously capture both you and your multimedia presentation.

   Note: For instructions on how to access and use Panopto, use the "Panopto How-To Videos" web link provided below. To access Panopto's website, navigate to the web link titled "Panopto Access," and then choose to log in using the "WGU" option. If prompted, log in using your WGU student portal credentials, and then it will forward you to Panopto's website.

   To submit your recording, upload it to the Panopto drop box titled "Data Mining II – OFM4" Once the recording has been uploaded and processed in Panopto's system, retrieve the URL of the recording from Panopto and copy and paste it into the Links option. Upload the remaining task requirements using the Attachments option.

G. Record the web sources you used to acquire data or segments of third-party code to support the analysis. Ensure the web sources are reliable.

H. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

I. Demonstrate professional communication in the content and presentation of your submission.