COMPETENCIES

4030.6.5 : **Dimension Reduction Methods**
The graduate implements dimension reduction methods to identify significant variables.

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 should also refer to the data dictionary file for your chosen data set from the above link. 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 them, 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 principal component analysis (PCA) to analyze customer data to identify the principal variables of your customers, 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 PCA to analyze patient data to identify the principal variables of your patients, 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 by using PCA.
   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: Method Justification

B. Explain the reasons for using PCA by doing the following:
   1. Explain how PCA analyzes the selected data set. Include expected outcomes.
   2. Summarize **one** assumption of PCA.

Part III: Data Preparation

C. Perform data preparation for the chosen data set by doing the following:
   1. Identify the continuous data set variables that you will need to answer the PCA question proposed in part A1.
   2. Standardize the continuous data set variables identified in part C1. Include a copy of the cleaned data set.

Part IV: Analysis

D. Perform PCA by doing the following:
   1. Determine the matrix of *all* the principal components.
   2. Identify the *total* number of principal components, using the elbow rule or the Kaiser criterion. Include a screenshot of the scree plot.
   3. Identify the variance of *each* of the principal components identified in part D2.
   4. Identify the *total* variance captured by the principal components identified in part D2.
   5. Summarize the results of your data analysis.

Part V: Attachments

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

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

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