OFM3 — OFM3 TASK 2: DIMENSIONALITY REDUCTION METHODS

DATA MINING II — D212
PRFA — OFM3

TASK OVERVIEW | SUBMISSIONS | EVALUATION REPORT

COMPETENCIES

4030.06.5 : Dimensionality 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. In doing so, you must select one of the data dictionary and data set files to use for your report from the following link: Data Sets and Associated Data Dictionaries.

You should also refer to the data dictionary file for your chosen dataset from the provided 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 its customers. When a company can better understand its customer 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 principal component analysis (PCA) to analyze customer data to identify the principal variables of your customers, ultimately allowing 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 its patients and the conditions leading to hospital admissions. When a hospital can better understand its patients' characteristics, it is better able to target treatment to patients, resulting in 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 allowing 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 originality 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 this data mining report by doing the following:
   1. Propose **one** question relevant to a real-world organizational situation that you will answer by using principal component analysis (PCA).
   2. Define **one** goal of the data analysis. Ensure that your goal is reasonable within the scope of the 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 dataset by doing the following:
   1. Identify the continuous dataset variables that you will need in order to answer the PCA question proposed in part A1.
   2. Standardize the continuous dataset variables identified in part C1. Include a copy of the cleaned dataset.

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.

File Restrictions

File name may contain only letters, numbers, spaces, and these symbols: ! - _ . * ' ( )
File size limit: 200 MB
File types allowed: doc, docx, rtf, xls, xlsx, ppt, pptx, odt, pdf, txt, qt, mov, mpg, avi, mp3, wav, mp4, wma, flv, asf, mpeg, wmv, m4v, svg, tif, tiff, jpeg, jpg, gif, png, zip, rar, tar, 7z

RUBRIC

A1:PROPOSAL OF QUESTION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not propose 1 question answered using PCA.	The submission proposes 1 question answered using PCA that is not relevant to a real-world organizational situation.	The submission proposes 1 question answered using PCA that is relevant to a real-world organizational situation.

A2:DEFINED GOAL

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not define 1 goal for data analysis.	The submission defines 1 goal for data analysis, but the goal is not reasonable, is not within the scope of the scenario, or is not represented in the available data.	The submission defines 1 reasonable goal for data analysis that is within the scope of the scenario and is represented in the available data.

B1:EXPLANATION OF PCA

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT

The submission does not explain how PCA analyzes the selected dataset.	The submission does not logically explain how PCA analyzes the selected dataset or includes inaccurate expected outcomes.	The submission logically explains how PCA analyzes the selected dataset and includes accurate expected outcomes.

B2:PCA ASSUMPTION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not summarize 1 assumption of PCA.	The submission inadequately summarizes 1 assumption of PCA.	The submission adequately summarizes 1 assumption of PCA.

C1:CONTINUOUS DATASET VARIABLES

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not identify the continuous dataset variables needed to answer the PCA question from part A1.	The submission inaccurately identifies the continuous dataset variables needed to answer the PCA question from part A1.	The submission accurately identifies the continuous dataset variables needed to answer the PCA question from part A1.

C2:STANDARDIZATION OF DATASET VARIABLES

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not standardize the continuous dataset variables.	The submission inaccurately standardizes the continuous dataset variables identified in part C1 or does not include a cleaned dataset.	The submission accurately standardizes the continuous dataset variables identified in part C1 and includes a cleaned dataset.

D1:PRINCIPAL COMPONENTS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not determine the matrix of *all* the principal components.	The submission inaccurately determines the matrix of 1 or more of the principal components.	The submission accurately determines the matrix of *all* of the principal components.

D2:IDENTIFICATION OF TOTAL NUMBER OF COMPONENTS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not identify the *total* number of principal components.	The submission inaccurately identifies the *total* number of principal components, or it does not use the elbow rule or the Kaiser criterion or does not include a screenshot.	The submission accurately identifies the *total* number of principal components, and the submission uses the elbow rule or the Kaiser criterion and includes a screenshot.

D3:TOTAL VARIANCE OF COMPONENTS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not identify the variance of *each* of the principal components identified in part D2.	The submission inaccurately identifies the variance of 1 or more of the principal components identified in part D2.	The submission accurately identifies the variance of *each* of the principal components identified in part D2.

D4:TOTAL VARIANCE CAPTURED BY COMPONENTS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not identify the *total* variance captured by the principal components identified in part D2.	The submission inaccurately identifies the *total* variance captured by the principal components identified in part D2.	The submission accurately identifies the *total* variance captured by the principal components identified in part D2.

D5:SUMMARY OF DATA ANALYSIS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not summarize the results of the data analysis.	The submission inadequately summarizes the results of the data analysis.	The submission adequately summarizes the results of the data analysis.

E:SOURCES FOR THIRD-PARTY CODE

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not record web sources used to acquire data or segments of third-party code.	The submission records 1 or more unreliable web sources	The submission records *all* web sources used to acquire data or segments of third-party code, and the web sources are reliable.

	used to acquire data or segments of third-party code.	

F:SOURCES

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not include both in-text citations and a reference list for sources that are quoted, paraphrased, or summarized.	The submission includes in-text citations for sources that are quoted, paraphrased, or summarized and a reference list; however, the citations or reference list is incomplete or inaccurate.	The submission includes in-text citations for sources that are properly quoted, paraphrased, or summarized and a reference list that accurately identifies the author, date, title, and source location as available.

G:PROFESSIONAL COMMUNICATION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
Content is unstructured, is disjointed, or contains pervasive errors in mechanics, usage, or grammar. Vocabulary or tone is unprofessional or distracts from the topic.	Content is poorly organized, is difficult to follow, or contains errors in mechanics, usage, or grammar that cause confusion. Terminology is misused or ineffective.	Content reflects attention to detail, is organized, and focuses on the main ideas as prescribed in the task or chosen by the candidate. Terminology is pertinent, is used correctly, and effectively conveys the intended meaning. Mechanics, usage, and grammar promote accurate interpretation and understanding.

WEB LINKS

Data Sets and Associated Data Dictionaries
If you have trouble with the link, copy and paste the link directly into your web browser.

Panopto Access
Sign in using the "WGU" option. If prompted, log in with your WGU student portal credentials, which should forward you to Panopto's website. If you have any problems accessing Panopto, please contact Assessment Services at assessmentservices@wgu.edu. It may take up to two business days to receive your WGU Panopto recording permissions once you have begun the course.

Panopto How-To Videos