OFM4 — OFM4 Task 1: Clustering Techniques

Data Mining II — D212
PRFA — OFM4

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 understand[ing] patients and the conditions leading to hospital admissions. When a hospital understands its patients' characteristics, it is better able to target tr[eatments]

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.

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, csv, 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.	The submission proposes 1 question that is not relevant to a real-world organizational situation. Or the proposal does not identify 1 of the given clustering techniques that will be used.	The submission proposes 1 question that is relevant to a real-world organizational situation, and the proposal identifies 1 of the given clustering techniques that will be used.

A2:DEFINED GOAL

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

B1:EXPLANATION OF THE CLUSTERING TECHNIQUE

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not explain how the chosen clustering technique analyzes the selected data set.	The submission does not logically explain how the chosen clustering technique analyzes the selected data set, or the explanation does not include expected outcomes.	The submission logically explains how the chosen clustering technique analyzes the selected data set and includes expected outcomes.

B2:SUMMARY OF THE TECHNIQUE ASSUMPTION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not summarize 1 assumption of the clustering technique.	The submission inaccurately summarizes 1 assumption of the clustering technique.	The submission accurately summarizes 1 assumption of the clustering technique.

B3:PACKAGES OR LIBRARIES LIST

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not list the packages or libraries chosen for Python or R.	The submission lists the packages or libraries chosen for Python or R but does not justify how 1 or more items on the list support the analysis.	The submission lists the packages or libraries chosen for Python or R and justifies how *each* item on the list supports the analysis.

C1:DATA PREPROCESSING

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not describe 1 data preprocessing goal.	The submission describes 1 data preprocessing goal, but it is not relevant to the clustering technique from part A1.	The submission describes 1 data preprocessing goal that is relevant to the clustering technique from part A1.

C2:DATA SET VARIABLES

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not identify *any* data set variables used to perform the analysis for the clustering question from part A1.	The submission identifies the data set variables used to perform the analysis for the clustering question from part A1, but the submission inaccurately labels 1 or more variables as continuous or categorical.	The submission identifies the data set variables used to perform the analysis for the clustering question from part A1, and the submission accurately labels *each* variable as continuous or categorical.

C3:STEPS FOR ANALYSIS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not explain *each* step used to prepare the data for the	The submission inaccurately explains 1 or more steps used to prepare the data for analysis, or	The submission accurately explains *each* step used to prepare the data for analysis,

analysis, or the submission does not identify the code segment for *each* step.	the submission identifies an inaccurate code segment for 1 or more steps.	and the submission identifies an accurate code segment for *each* step.

C4:CLEANED DATA SET

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not include a copy of the cleaned data set.	The submission includes a copy of the cleaned data set, but the data set is inaccurate.	The submission includes an accurate copy of the cleaned data set.

D1:OUTPUT AND INTERMEDIATE CALCULATIONS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not determine the optimal number of clusters in the data set.	The submission determines the optimal number of clusters in the data set, but the submission inaccurately describes the methodology used or inappropriately applies the methodology.	The submission determines the optimal number of clusters in the data set and accurately describes the methodology used. The methodology is appropriately applied.

D2:CODE EXECUTION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not provide the code used to perform the clustering analysis technique.	The submission provides the code used to perform the clustering analysis technique, but 1 or more errors are evident during the execution of the code.	The submission provides the code used to perform the clustering analysis technique, and the code executes without errors.

E1:QUALITY OF THE CLUSTERING TECHNIQUE

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT

The submission does not explain the quality of the clustering technique .	The submission does not logically explain the quality of the clustering technique.	The submission logically explains the quality of the clustering technique.

E2:RESULTS AND IMPLICATIONS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not discuss *both* the results and implications of the clustering analysis.	The submission discusses *both* the results and implications of the clustering analysis, but the discussion is inaccurate.	The submission accurately discusses *both* the results and implications of the clustering analysis.

E3:LIMITATION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not discuss 1 limitation of the data analysis.	The submission discusses 1 limitation of the data analysis, but it is illogical or lacks adequate detail to support its logic.	The submission logically discusses 1 limitation of the data analysis with adequate detail.

E4:COURSE OF ACTION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not recommend a course of action for the real-world organizational situation from part A1.	The submission does not recommend a reasonable course of action for the real-world organizational situation from part A1, or the course of action is not based on the results and implications discussed in part E2.	The submission recommends a reasonable course of action for the real-world organizational situation from part A1 based on the results and implications discussed in part E2.

F:PANOPTO VIDEO OF CODE

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
A Panopto video recording of the code used is not provided, or the link provided for the video is not functional.	A Panopto video recording is provided, but a full demonstration of the code used, the code being executed, or the results of the code used in the task is not provided, or the video does not capture both the presenter and the vocalized demonstration.	A Panopto video recording is provided that includes a full demonstration of the code used, the code being executed, and the results of the code used in the task. For the duration of the presentation, the video captures both the presenter and the vocalized demonstration.

F1:PANOPTO VIDEO OF PROGRAMS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
A Panopto video recording of the programs used is not provided.	A Panopto video recording is provided, but a complete description of the programs used to complete the task is not provided, or the video does not capture both the presenter and the vocalized presentation describing the programs used to complete the task.	A Panopto video recording is provided that includes a complete description of the programs used to complete the task. For the duration of the presentation, the video captures both the presenter and the vocalized presentation describing the programs used to complete the task.

G: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 used to acquire data or segments of third-party code.	The submission records *all* web sources used to acquire data or segments of third-party code, and the web sources are reliable.

H: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.

I: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 URL 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