NUM2 — NUM2 TASK 1: DATA CLEANING

DATA CLEANING — D206
PRFA — NUM2

TASK OVERVIEW SUBMISSIONS EVALUATION REPORT

COMPETENCIES

4030.03.01 : Predicting Obstacles to Cleaning Data
The graduate predicts potential obstacles in data analysis based on the quality of the data provided.

4030.03.02 : Preparing Data for Analysis
The graduate prepares data for analysis to address a business need.

4030.03.03 : Writing code to Clean data
The graduate writes reusable code to manipulate and clean data in preparation for analysis.

INTRODUCTION

In a previous course you used SQL methods to collect data for analysis and to support decision-making processes. The next step involves preparing the data for analysis, a process known as data cleaning. You will explore various graphs and statistics to identify outliers, consider various methods to handle missing data, such as imputation, and explore a basic use of principal component analysis (PCA) for data reduction of a set of variables.

To complete this assessment, you will use raw data from the industry of your choice and prepare the data set for analysis. You will also create visualizations and deliver a clean data set ready for exploratory analysis.

SCENARIO

For this task, you will select one of the Data Dictionary and Data Set files from the following link:

D206 Definitions and Data files

You will review the Data Dictionary to understand the needs of the company and to prepare to clean the data. In this assessment, you will analyze the .csv data file, also referred to as the data set.

Note: This assessment may require you to submit pictures, graphics, and/or diagrams. Each file must be an attachment no larger than 30 MB in size. Diagrams must be original and may be hand-drawn or drawn using a graphics program. Do not use CAD programs because attachments will be too large.

REQUIREMENTS

Help

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., .csv, .docx, .pdf, .ppt).*

Part I: Research Question

A. Describe **one** question or decision that you will address using the data set you chose. The summarized question or decision must be relevant to a realistic organizational need or situation.

B. Describe the variables in the data set and indicate the specific type of data being described. Use examples from the data set that support your claims.

Part II: Data-Cleaning Plan

Note: You may use Python, R, or any other programming language for implementing your coding solutions, manipulating the data, and creating visual representations.

C. Explain the plan for cleaning the data by doing the following:
   1. Propose a plan that includes the relevant techniques and specific steps needed to identify anomalies in the data set.
   2. Justify your approach for assessing the quality of the data, include:
      - characteristics of the data being assessed,
      - the approach used to assess the quality.
   3. Justify your selected programming language and any libraries and packages that will support the data-cleaning process.
   4. Provide the code you will use to identify the anomalies in the data.

Part III: Data Cleaning

D. Summarize the data-cleaning process by doing the following:
   1. Describe the findings, including all anomalies, from the implementation of the data-cleaning plan from part C.
   2. Justify your methods for mitigating each type of discovered anomaly in the data set.
   3. Summarize the outcome from the implementation of *each* data-cleaning step.
   4. Provide the code used to mitigate anomalies.
   5. Provide a copy of the cleaned data set.
   6. Summarize the limitations of the data-cleaning process.
   7. Discuss how the limitations in part D6 affect the analysis of the question or decision from part A.

E. Apply principal component analysis (PCA) to identify the significant features of the data set by doing th following:

1. List the principal components in the data set.
2. Describe how you identified the principal components of the data set.
3. Describe how the organization can benefit from the results of the PCA

Part IV. Supporting Documents

F. Provide a Panopto recording that demonstrates the warning- and error-free functionality of the code used to support the discovery of anomalies and the data cleaning process and summarizes the programming environment.

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 Cleaning – NUM2 \ D206" 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. Reference the web sources used to acquire segments of third-party code to support the application. Be sure 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, 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

A:QUESTION OR DECISION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
A description is not provided.	The description does not state a question or decision that can be addressed through analysis of the chosen data set. Or the question or decision is not rele-	The description states a question or decision that can be addressed through analysis of the chosen data set. The question or decision is relevant to a realistic organizational need or situation.

Help

vant to a realistic organizational need or situation.

B:REQUIRED VARIABLES

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
A description of the data set is not provided.	The description only one of the variables in the data set, or does not indicate the specific type of data being described. Or the description does not include examples from the data set to support claims.	The description includes the variables in the data set and indicates the specific type of data being described, and includes examples from the data set to support claims.

C1:PLAN TO FIND ANOMALIES

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
A proposal is not provided.	The proposal includes techniques and steps needed for identifying anomalies, but either the techniques or the steps are not relevant or not appropriate for identifying the anomalies in the data set.	The proposal includes a detailed description of the techniques and steps needed for identifying anomalies in the selected data set.

C2:JUSTIFICATION OF APPROACH

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
A justification is not provided.	The justification includes the characteristics of the data being assessed, but the justification does not reference the approach used to assess the quality of the data. Or the justified approach is not aligned with the selected data set.	The justification includes the characteristics of the data being assessed and references the approach used to assess the quality of the data. The justified approach aligns with the selected data set.

C3:JUSTIFICATION OF TOOLS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
A justification is not provided.		

Help

	The justification describes the benefits of using the programming language, including any libraries and packages used to clean the data, but does not include specific examples of how these tools are ideal in this scenario as opposed to other available tools.	The justification describes the benefits of using the programming language, including any libraries and packages used to clean the data, and includes specific examples of how these tools are ideal in this scenario as opposed to other available tools.

C4:PROVIDE THE CODE

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not provide any code.	The submission provides an incomplete or non-executable code. Or the code provided could not be used to identify anomalies in the data set.	The submission provides the complete and executable code, which could be used to identify anomalies in the data set.

D1:CLEANING FINDINGS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
A description is not provided.	The description includes the findings, including some of the anomalies found, by running the code from part C4, but the description contains inaccuracies, or the description does not include *all* of the anomalies.	The description accurately includes *all* of the anomalies found by running the code from part C4.

D2:JUSTIFICATION OF MITIGATION METHODS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The justification is not provided.	The justification includes mitigation methods that are not specific to the anomalies listed in part D1. Or the methods could not be used to mitigate the anomalies.	The justification includes the specific mitigation methods for each type of anomaly listed in part D1.

D3:SUMMARY OF THE OUTCOMES

Help

NOT EVIDENT

A summary is not provided.

APPROACHING COMPETENCE

The summary details the outcome from the implementation of some but not all the data-cleaning steps. Or the summarized outcomes are not plausible given the interventions.

COMPETENT

The summary details the outcome from the implementation of each data-cleaning step. The summarized expected outcomes are plausible given the interventions.

D4:MITIGATION CODE

NOT EVIDENT

The submission does not provide any code.

APPROACHING COMPETENCE

The submission provides an incomplete or non-executable code. Or the code provided could not be used to mitigate the anomalies.

COMPETENT

The submission provides complete and executable code that could be used to mitigate the anomalies.

D5:CLEAN DATA

NOT EVIDENT

The submission does not provide a data set.

APPROACHING COMPETENCE

The submission includes a data set that still contains anomalies that should have been mitigated. Or the provided data set is missing variables from the chosen data set in part A

COMPETENT

The submission includes a clean data set created from the raw data. The provided data set includes the complete list of variables from the chosen data set in part A.

D6:LIMITATIONS

NOT EVIDENT

The submission does not summarize *any* limitations.

APPROACHING COMPETENCE

The submission summarizes the limitations of the implemented data-cleaning process, but the summary includes either inaccuracies or misses obvious limitations of the process.

COMPETENT

The submission accurately summarizes the limitations of the implemented data-cleaning process.

D7:IMPACT OF THE LIMITATIONS

NOT EVIDENT

COMPETENT

Help

The submission does not discuss the impact of *any* limitations.	**APPROACHING COMPETENCE** The submission includes a discussion of the impact of the limitations from part D6. But the discussion does not logically align with the question or decision from part A.	The submission includes a discussion of the impact of the limitations from part D6. The discussion logically aligns with the question or decision from part A.

E1:PRINCIPAL COMPONENTS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not list *any* principal components of the data set.	The submission includes a partial list of the principal components of the data set.	The submission lists *all* principal components of the data set.

E2:CRITERIA USED

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
A description of how the principal components of the data set were identified is not provided	The description of how the principle components of the data set were identified is inaccurate or incomplete.	The description of how the principle components of the data set were identified is accurate and complete.

E3:BENEFITS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
A description is not provided.	The description of how the organization would benefit from the results of the PCA is illogical or inaccurate.	The description of how the organization can benefit from the results of the PCA is logical and accurate.

F:VIDEO

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
A Panopto video is not provided.	The Panopto video recording is missing the demonstration of the functionality of the code	The Panopto video recording demonstrates the warning-and error-free functionality of the code used to support the di

Help

	used to support discovery of anomalies or the data cleaning process, or it is missing an accurate summary of the programming environment, or it is missing both the demonstration and the summary, or the code functionality is not warning- or error-free.	ery of anomalies and the data cleaning process. An accurate summary of the programming environment is provided in the video.

G:SOURCES FOR THIRD-PARTY CODE

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The web sources are not provided.	The submission records only some of the web sources used to acquire data or third-party code. Or the web sources are not reliable.	The submission records all web sources used to acquire data or third-party code and all of 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.

Help

WEB LINKS

Churn Data Dictionary and Data Set
If you have trouble with the link, copy and paste the link directly into your web browser.

Medical Data Dictionary and Data Set
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