TCN1 — TCN1 TASK 1: DATA CLEANING AND PROFILING

DATA PREPARATION AND EXPLORATION — D599
PRFA — TCN1

COMPETENCIES

4159.1.1 : Profiles Data
The learner interprets a data dictionary to understand the data set.

4159.1.2 : Interprets Statistics and Visualization
The learner interprets probability, descriptive and inferential statistics, and visualization.

4159.1.3 : Wrangles Data
The learner wrangles data to ensure accuracy, format, and integrity relevant to the task being performed.

INTRODUCTION

Throughout your career in data analytics, you will prepare data according to business and data analytic needs. You will interpret data dictionaries to understand a dataset, identify the type of scale for each variable, analyze outliers and numeric variables, and identify duplicate data and missing values to clean data.

In this task, you will review the raw dataset and accompanying data dictionary provided to prepare the dataset file for further analysis according to business needs.

SCENARIO

You are a data analyst at a large multinational technology firm. To sustain success in the highly competitive tech industry, your company relies on the availability and retainment of highly skilled workers. Due to increasing employee turnover (i.e., the total number of workers who leave a company over a certain period, either voluntarily or involuntarily), the senior partners think it is time to formally evaluate this issue through data analysis. The company would like to understand which employees are at greater risk of leaving their positions so it can design smarter employee retention strategies and reduce employee turnover.

You have been asked to profile and clean the raw dataset to optimize it for future analysis and model building. Your goal is to gain insights into the characteristics and quality of the data, identify any anomalies or inconsistencies, clean the data, and prepare a data cleaning report outlining your findings for the firm.

Refer to the most recent company data provided in the "Employee Turnover Dataset" and "Employee Turnover Considerations and Dictionary" supporting documents to inform your work.

Note: The IDE for this assessment is either Anaconda or RStudio, depending on which language you decide to use to complete the task. Please use the "WGU Virtual Lab Environment" web link below.

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).*

Note: Written responses need to be submitted through EMA.

Part I: Data Profiling

*Note: Your responses to the following task prompts must be provided in a document file. Unless otherwise specified, responses to PA requirements that are included in a Python or RStudio notebook will **not** be accepted.*

A. Profile data by doing the following:
   1. Review the data dictionary in the attached "Employee Turnover Considerations and Dictionary" document and do the following:
      a. Describe the general characteristics of the initial dataset (e.g., rows, columns).
      b. Indicate the data type and data subtype for *each* variable.
      c. Provide a sample of observable values for *each* variable.

Part II: Data Cleaning and Plan

*Note: You may use Python or R for implementing your coding solutions, manipulating the data, and creating visual representations. However, your responses to the following task prompts must be provided in a document file. Unless otherwise specified, responses to PA requirements that are included in a Python or RStudio notebook will **not** be accepted.*

B. Inspect the dataset through data cleaning techniques for *all* duplicate entries, missing values, inconsistent entries, formatting errors, and outliers and do the following:
   1. Explain how you inspected the dataset for *each* of the quality issues listed in part B.
   2. List your findings for *each* quality issue listed in part B.

C. Discuss which data cleaning techniques you used to correct *all* the data quality issues you identified by doing the following:
   1. Describe how you modified the dataset after identifying *each* quality issue listed in part B.

2. Discuss why you chose the specific data cleaning techniques you used to clean the quality issues listed in part B.
3. Describe **two** or more advantages to your data cleaning approach specified in part C1.
4. Discuss **two** or more limitations to your data cleaning approach specified in part C1.

Part III: Submission

D. Submit your findings by doing the following:

1. Provide a data cleaning report as a document file that includes responses to task prompts.
2. Provide the annotated code you used to detect and mitigate the data quality as an executable script file. R files and Python script files are accepted.
3. Provide a copy of the cleaned dataset as a CSV file.
4. Provide a Panopto video recording that includes a screen share of the presenter demonstrating the functionality of the code used and a discussion commenting on 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 Preparation and Exploration TCN1 | D599 (Student Creators)." 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.

Sources

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

Professional Communication

F. 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

A1A:PROFILE DATA

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT

A description of the dataset is not provided in a document file.	The description of the dataset in the document file is provided but is incomplete or inaccurate.	The description in the document file includes complete and accurate information about the dataset.

A1B:VARIABLE DATA TYPES

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
An indication of data types or data subtypes for variables is not provided in a document file.	The submitted document file does not include *all* the variables in the dataset. Or the document file does not indicate the specific type of data being described for *each* variable.	The submitted document file includes *all* variables in the dataset and indicates the specific type of data being described for *each* variable.

A1C:OBSERVABLE VALUES

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
A sample of observable values is not provided in a document file.	The submitted document file provides some observable values but not for *all* variables. Or the document file does not provide a sample of values for *each* variable.	The submitted document file includes a sample of observable values for *each* dataset variable.

B1:DATASET QUALITY ISSUES

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
An explanation about how the dataset was inspected is not provided in a document file.	The submitted document file explains how the dataset was inspected, but the explanation is illogical, inaccurate, or incomplete.	The submitted document file explains how the dataset was inspected and is logical, accurate, and complete.

B2:LIST OF QUALITY ISSUES

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
A list of quality issues is not provided in a document file.	The submitted document file provides a list of quality issues, but the list is inaccurate or incomplete.	The submitted document file provides a list of quality issues that is accurate and complete.

C1:DATASET MODIFICATION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
A discussion about how the dataset was modified is not provided in a document file.	The submitted document file discusses how the dataset was modified, but the discussion is illogical, inaccurate, or incomplete.	The submitted document file discusses how the dataset was modified and is logical, accurate, and complete.

C2:DATA CLEANING TECHNIQUES

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
A discussion of why specific data cleaning techniques were used is not provided in a document file.	The submitted document file describes data cleaning techniques, but the description does not include why specific techniques were used. Or the discussion is illogical, inaccurate, incomplete, or not aligned with the quality issues listed in part B.	The submitted document file describes data cleaning techniques, including why specific techniques were used. The discussion is logical, accurate, complete, and aligned with the quality issues listed in part B.

C3:TECHNIQUE ADVANTAGES

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
A description of advantages is not provided in a document file.	The submitted document file describes only 1 advantage of the data analysis. Or the advantages described are illogical, inaccurate, incomplete, or not applicable to the analysis.	The submitted document file describes *at least* 2 advantages of the data analysis that are logical, accurate, complete, and applicable to the analysis.

C4:TECHNIQUE LIMITATIONS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
A discussion of limitations is not provided in a document file.	The submitted document file discusses limitations of the data analysis, but 1 or more of the limitations provided are illogical,	The submitted document file discusses *at least* 2 limitations of the data analysis, and the discussion is logical, accurate, com-

	inaccurate, incomplete, or not applicable to the analysis.	plete, and applicable to the analysis.

D1:DATA CLEANING REPORT

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
A data cleaning report is not provided in a document file.	The submitted document file is provided, but it is illogical, inaccurate, or incomplete.	The submitted document file includes a logical, accurate, and complete data cleaning report.

D2:ANNOTATED CODE

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
No annotated code is provided.	The submission provides annotated code that has warnings or errors or does not accurately use 1 of the listed cleaning techniques in part C2 to analyze the data.	The submission includes warning- and error-free annotated code to accurately analyze the dataset using 1 of the listed techniques in part C2.

D3:CLEAN DATASET CSV

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
No dataset is provided.	The submitted document file provides a dataset that contains data quality issues that should have been mitigated. Or the dataset is missing variables from the provided dataset. Or the file is not in a CSV format.	The submitted document file provides the cleaned dataset in a CSV format. The submission is created from raw data that is free of *all* data quality issues, and it includes the complete list of variables from the provided dataset.

D4:PANOPTO VIDEO

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
A Panopto video recording is not provided.	The Panopto video recording provided is missing either the demonstration of the functionality of the code or the summary of the tools used. Or either the	The Panopto video recording provided includes *both* a screen share of the presenter demonstrating the functionality of the code used and a discussion commenting on the programming en-

	demonstration or the summary is inaccurate.	vironment. *Both* the demonstration and the summary are accurate and complete.

E: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, or the submission states no sources were used.

F: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

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 FAQs

Panopto How-To Videos

WGU Virtual Lab Environment

SUPPORTING DOCUMENTS

Employee Turnover Considerations and Dictionary.docx

Employee Turnover Dataset.xlsx