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 represent your original work and understanding of the course material. Most performance assessment submissions are automatically scanned through the WGU similarity checker. Students are strongly encouraged to wait for the similarity report to generate after uploading their work and then review it to ensure Academic Authenticity guidelines are met before submitting the file for evaluation. See [Understanding Similarity Reports](https://cm.wgu.edu/t5/Frequently-Asked-Questions/Understanding-Similarity-Reports/ta-p/252) for more information.    
  
**Grammarly Note:**   
Professional Communication will be automatically assessed through Grammarly for Education in most performance assessments before a student submits work for evaluation. Students are strongly encouraged to review the Grammarly for Education feedback prior to submitting work for evaluation, as the overall submission will not pass without this aspect passing. See [Use Grammarly for Education Effectively](https://cm.wgu.edu/t5/Academic-Coaching-Center/Use-Grammarly-for-Education-Effectively/ta-p/52276) for more information.    
  
**Microsoft Files Note:**  
Write your paper in Microsoft Word (.doc or .docx) unless another Microsoft product, or pdf, is specified in the task directions. Tasks may not be submitted as cloud links, such as links to Google Docs, Google Slides, OneDrive, etc.  All supporting documentation, such as screenshots and proof of experience, should be collected in a pdf file and submitted separately from the main file. For more information, please see [Computer System and Technology Requirements.](https://cm.wgu.edu/t5/WGU-Student-Policy-Handbook/Computer-System-and-Technology-Requirements/ta-p/78)    
 *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.*

**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)*](https://wgu.hosted.panopto.com/Panopto/Pages/Sessions/List.aspx#folderID=%22302aa83e-e437-40f5-b21a-b1bd00ec8d67%22)*." 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.