

# NUM4 – NUM4 TASK 1: DATA CLEANING

DATA CLEANING – D206

PRFA – NUM4

Preparation

**Task Overview**

Submissions

Evaluation Report

## COMPETENCIES

### 4030.3.1 : Predicting Obstacles in Data Analysis

The graduate predicts potential obstacles in data analysis based on the quality of data provided.

### 4030.3.2 : Preparing Data for Analysis

The graduate prepares data for analysis to address organizational needs.

### 4030.3.3 : Manipulating Data for Analysis

The graduate writes reusable code to manipulate and clean data in preparation for analysis.

## INTRODUCTION

In a previous course, you used Structured Query Language (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.

## 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](#) 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](#) 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](#).

*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.*

*Note: All visualizations are created with Python or R and should be embedded in the report. Do not use CAD programs because attachments will be too large.*

Select **one** of the data files and its associated dictionary file from the "[D206 Definitions and Data Files](#)" web link, then do the following:

### Part I: Research Question

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

- A. In a document file, describe your research problem by doing the following:
  1. Describe **one** question or decision that could be addressed using the data set you chose. The summarized question or decision must be relevant to a realistic organizational need or situation.
  2. Describe *all* variables in the data set (regardless of the research question) and indicate the data type for *each* variable. Use examples from the data set to support your claims.

### Part II: Data-Cleaning 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 task prompts must be provided in a document file. Unless otherwise specified, responses to PA requirements that are included in a python or R studio notebook will not be accepted.*

- B. In a document file, 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 assess the quality of the data in the data set.
  2. Justify your approach for assessing the quality of the data, including the following:
    - characteristics of the data being assessed
    - the approach used to assess the quality of the data

3. Justify your selected programming language and any libraries and packages that will support the data-cleaning process.

### Part III: Data Cleaning

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

- C. In a document file, summarize the data-cleaning process by doing the following:
  1. Describe the findings for the data quality issues found from the implementation of the data-cleaning plan from Part II.
  2. Justify your methods for mitigating the data quality issues in the data set.
  3. Summarize the outcome from the implementation of *each* data-cleaning step.
  4. Summarize the limitations of the data-cleaning process.
  5. Discuss how the limitations summarized in part C4 could affect the question analysis or decision from part A1.
- D. Provide the detection and mitigated code you used for the data-cleaning process by doing the following:
  1. Provide the annotated code you will use to mitigate the data quality issues—including anomalies—in the data set in an executable script file such as Python or R.
  2. Provide a copy of the cleaned data set as a CSV file.
- E. Apply principal component analysis (PCA) to identify the significant features of the data set and provide the following in a document file:
  1. Identify the total number of principal components and provide the output of the principal components loading matrix.
  2. Justify the reduced number of the principal components and include a screenshot of a scree plot.
  3. Describe how the organization would benefit from the use of PCA.

### Part IV: Supporting Documents

- F. Provide a Panopto video recording that includes the presenter and a vocalized demonstration of the functionality of the code used for the analysis of 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 NUMx | D206 \(Student Creators\) \[assignments\]](#)." 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. Acknowledge web sources, using in-text citations and references, for segments of third-party code used 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.