|  |
| --- |
| **Important Reminders!**   1. Follow appropriate APA formatting guidelines for the cover page, table of contents, and headers, etc. 2. (\*) Indicates an area of the PA that often present the most challenge to students. Utilize the Performance Assessment Rubric to ensure you respond to all requirements accurately and thoroughly. 3. Allow a 3-day turnaround for the evaluation of your performance assessment. |

# **Part I: Research Question and Variables** **A:** Discuss your research question. Be sure to devise a research question related to the dataset you have selected and related to all variables in the dataset. **NOTE: For the purpose of the PA, consider making your research relatively broad, and not too narrow. For example, instead of saying "Does A, B, and C cause D?" consider asking "What causes D?“** **NOTE: Regardless of your research question, you will need to examine and clean the entire dataset for missing values, outliers, etc. (do not drop any variables as a result of your research question).** You are not responsible for answering the research question in this course. Your primary responsibility is to clean the entire dataset (minimum: addressing duplicates, nulls, and outliers).

**B:** List **every** variable in the dataset, regardless of your research question, which includes: (\*)

* Variable name
* Data type (ex: quantitative/qualitative)
* Description/definition of the variable (Note: use the data dictionary for assistance). Do not copy from the data dictionary provided, as this will cause originality concerns.
* An example directly from the dataset for each variable.

**Part II: Data-Cleaning Plan (Detection)**

**C1.** Discuss **what** **methods (functions/commands)** were used to **detect** duplicates, missing values, outliers, and any other data quality issues.

* Note: In this section discuss re-expressed variables (if used/applicable. Remember if you have variables with the inconsistent presentation of categorical variables (Yes/No) for some and (0/1) for others, detection/treatment is necessary).
* Note: Detect outliers for all quantitative variables.

**C2:** Discuss **why** you selected the **(methods, functions, commands)** discussed in C1. Use credible sources (i.e., course textbook) to support your why.

**C3:** Discuss **what** programming language you used to clean your data and **why.** Also, discuss **what** libraries and packages and **why. [Note: I would encourage you to use the following link to assist you when justifying the use of Python or R** [<https://www.wgu.edu/online-it-degrees/programming-languages/r-or-python.html>](https://www.wgu.edu/online-it-degrees/programming-languages/r-or-python.html)**]**

**C4: Provide the code (input code only, no screenshots, no output as the output should be included in D1) used to detect duplicates, missing values, outliers** (all numeric values that are related to customer/patient), and any other data cleaning matters. Remember, you need to detect outliers for all quantitative variables.If you are uploading a copy of your code (which is highly recommended), state, “see code/script attached” in addition to providing your code within the document itself).

**Part III: Data Cleaning (Treatment)**

**D1.** Discuss what you found after you checked for **duplicates, missing values, outliers**, etc. For example, which variables did you find missing values? How many values were missing? What variables did you find outliers? How many outliers did you find? What were the values of those outliers?

**D2.** Discuss **what** you did to treat the **duplicates, missing values, and outliers**, that you referenced in D1. Discuss **why** you used the treatment methods you selected. Use credible sources (i.e., course textbook) to support your why. Be very specific on why and how you treated each variable in which you found data quality concerns. Use visualizations as the evaluators are only reviewing the Panopto video to ensure code functionality.

**D3.** Summarize all the work that was performed. Discuss how your data looks now that it has been treated (**duplicates, missing values, and outliers**). **Provide evidence and visualizations confirming the data is cleaned.**

**D4.** Provide the code you used to **treat your data (input code only, no screenshots, no output as this should have been included in D2)**. If you are uploading a copy of your code (which is highly recommended), state, “see code/script attached” in addition to providing your code within the document itself.).

**D5.** Provide a CSV file of your clean data (extract the file from your environment). CSV Files are accepted.

* **To extract your clean data using Python:**  
  df.to\_csv(r'Path where you want to store the exported CSV file\File Name.csv’)
* **To extract your clean data Using R:**   
  *Write.csv (Your DataFrame,"Path to export the DataFrame\\File Name.csv”)*

**D6.** Discuss the disadvantages of the **methods** you used to clean/treat your data **(duplicates, missing values, and outliers)** that you referenced in D2. This discussion should include the disadvantages of the methods used for the detection and treatment of duplicates, missing values, and outliers, at minimum). (\*)

**D7.** Discuss what challenges a data analyst may encounter if they were to use your now-cleaned data for analysis. (\*)

**Part IV: PCA**

**E1:** Include the variables used for PCA **and** the output (a screenshot) of the PCA loadings matrix. Note, use as many variables as possible and be mindful to only select variables that meet the appropriate data type permissible for PCA.

**E2:** Discuss which PCs should be **retained** and **why** (visualizations such as a scree plot is helpful to include with eigenvalues plotted). The Kaiser rule is one method introduced in this course.

**E3.** Describe how the organization can benefit from the results of the PCA. You can answer this question by answering how any organization could benefit from any PCA. Your answer does not have to address the specific principal components you created in your PCA. However, you should use your results as an example in your answer to E3.

Check out *The Benefits of PCA* section at [https://www.bigabid.com/what-is-pca-andhow-can-i-use-it/,](https://www.bigabid.com/what-is-pca-and-how-can-i-use-it/) which provides a very succinct list of why we might perform a PCA and how an organization might benefit from a PCA.

**F.  Provide a Panopto recording (\*)**

When creating your video be certain to include the following:

1. Discuss the programming environment used (e.g, PyCharm, Juptyer lab, etc.)
2. Execute every line of code (showing code is free from errors). Remember, you do not need to type every line of code, again, just execute your existing code used for the Performance Assessment.

Check to verify you have access to the Panopto folder (if you do not have access email

[assessmentservices@wgu.edu](mailto:assessmentservices@wgu.edu)).

**G. Third-Party Code References (\*)**

In this section, cite the sources you used to assist with the **CODE** of your work.

Note: Any reference entry listed here must have an **in-text citation** in either C4 or D4. Use

APA Citation. If no additional sources were used, please state that as such.

**H. References (\*)**

In this section, cite the source used to assist you with the **CONTENT** of your work.   
  
Note: Any reference entry listed here must have an **in-text citation** in the report above. Use APA Citation. If no additional sources were used, please state that as such.