Task

**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 the 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.

Rubric

A: Question or Decision

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.

B: Required variables

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

The proposal includes a detailed description of the techniques and steps needed for identifying anomalies in the selected data set.

C2: Justification of Approach

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

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

The submission provides the complete and executable code, which could be used to identify anomalies in the data set.

D1: Cleaning Findings

The description accurately includes all of the anomalies found by running the code from part C4.

D2: Justification of Mitigation Methods

The justification includes the specific mitigation methods for each type of anomaly listed in part D1.

D3: Summary of Outcomes

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

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

D5: Clean Data

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

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

D7: Impact of the Limitations

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

The submission lists *all* principal components of the data set.

E2: Criteria Used

The description of how the principal components of the data set were identified is accurate and complete.

E3: Benefits

The description of how the organization can benefit from the results of the PCA is logical and accurate.

F: Video

The Panopto video recording demonstrates the warning-and error-free functionality of the code used to support the discovery 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

The submission records all web sources used to acquire data or third-party code and all of the web sources are reliable.

H: Sources

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

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.