# D206 Performance Assessment
# Step-by-Step Guide

**Read! Important Reminders!!**

1. Read this guide thoroughly to minimize performance assessment revisions.
2. You are required to clean the entire dataset, regardless of your research question. Do not drop any columns/variables from your dataset as a result of your research question.
3. Allow a 3-day turnaround for the evaluation of your performance assessment.
4. Follow appropriate APA formatting guidelines for the cover page, table of contents, headers, etc.
5. Annotate your code (use comments). If sources were used, you must include an in-text citation of the sources within the annotation. You will then include a reference entry in section (H) of your performance assessment.
6. When working on the content of your paper, keep a list of all sources used (videos, course content, etc.). Whenever a source is used, create an in-text citation. You will then include a reference entry in section (G) of your performance assessment.
7. There is no page requirement for the Performance Assessment. No page requirement.
8. (*) Indicates an area of the PA that is often most challenging to students. Utilize the *step-by-step* guide below to ensure you respond to all requirements accurately and thoroughly.

## 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.**

**REMINDER!** For the PA, consider making your research relatively broad, and not too narrow. As an example, if you have a large dataset of student factors (age, GPA, number of hours of study time, standardized test scores, etc.), forming a research question that states, "Does study time and hours of rest influence GPA" is too narrow. A much better approach is to ask, "What student factors influence GPA."

**REMINDER!** 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).

**REMINDER!** Your research question does not impact task requirements B, C, D1 - D6, or E. In future classes, you will write a research question that impacts your entire task.

**REMEMBER!** Do not use the data from a previous class. Download the D206 data and data dictionary for D206.

1. Go to the D206 course page.
2. Select View Task under Assessments at the bottom center of the page
3. Select D206 Definitions and Data Files under Scenario on the Task
4. Select the link for the dataset you will be using.
5. Unzip the downloaded folder.
6. The data file is in CSV format.
7. The data dictionary is in PDF format. Ignore the Scenario on page 1 of the PDF. The Scenario has nothing to do with your work in this class.

**B:**

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

- **Variable name**
- **Data type** (either quantitative or qualitative) Do not provide the data type from your respective environment (i.e., float, string)**.**
- **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 of a value directly from the dataset for each variable.**

**REMEMBER:** When determining rather a variable is quantitative or qualitative be certain to refer to the definition of the variable. If a categorical variable has been re-expressed numerically, it is still a categorical/qualitative variable). Review Dr Straw breakdown of data types.pdf

**REMEMBER:** The following variables should be described (Requirement I.B) but should not be cleaned (CaseOrder is a sequence number; Interaction and UID are encoded customer ID information; Address related (i.e. State, County, Zip, and City) – In a corporate setting these address related items would be cleaned based on master lookup files; Lat and Lng – In a corporate setting the latitude and longitude would be cleaned based on a lookup database

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

**C1.**

**Discuss what methods (functions/commands) were used to detect the following data quality issues (at minimum):**

- Duplicates
- Missing values
- Outliers
- Re-expression of categorical variables (if needed), and any other data quality issues.

**REMEMBER!** In this section, do not discuss how you treated the data quality issues. Only discuss what approaches were used to **detect** data quality issues.

**REMEMBER!** In this section discuss re-expressed variables. Pay close attention to the encoding of the variables for medical conditions in the medical data set. Two of the medical condition variables are encoded differently than the rest of the medical condition variables and should be re-encoded to match the method used in the other medical condition variables.

**REMEMBER!** You must detect outliers for all quantitative variables.

**C2:**

**Discuss why you used the (functions/commands) discussed in C1 to detect the following (at minimum):**

- Duplicates
- Missing values
- Outliers
- Re-expression of categorical variables (if needed), and any other data quality issues.

**C3:**

**Discuss what** programming language you used to clean your data and **why**. Also, discuss what libraries and packages and why.

**REMEMBER!** You are encouraged 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

**C4:**

**Provide the detection code (input code only; no output as this will be included in D1) and attach a copy of your code.**

> **REMEMBER!** Your detection code would include the methods you discussed in C1 to detect all data quality issues (duplicates, missing values, outliers, etc.).
>
> **REMEMBER!** Attach your annotated and executable (upload) your Python .py; Python via Jupyter Notebook .ipynb; R .r. These file types are accepted.
>
> **REMEMBER:** Be certain to state, *"See code attached."*

## Part III: Data Cleaning (Treatment)

**D1.**

**Discuss what you found after you checked for duplicates, missing values, outliers, etc.**

- Discuss how many duplicates you found.
- Discuss which variables did you find missing values and how many values were missing.
- Discuss what variables you found outliers, how many outliers you found, and the values of those outliers. You must detect outliers for all quantitative variables.

> **REMEMBER!** Do not discuss treatment or how you fixed what you found in this section. This should be discussed in D2.

**D2.**

**Discuss what you did to treat the duplicates, missing values, and outliers, that you referenced in D1 and why you used the treatment methods you selected.**

- Discuss how you treated the duplicates and why you used that treatment method.
- Discuss how you treated each variable that contained missing values and why you utilized the specific treatment method for each variable.
- Discuss how you treated the outliers for each variable, and why you utilized the specific treatment method for each variable.
- Be very specific and thorough on why and how you treated any other data quality concerns (including the re-expression of categorical variables).

**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 treatment code (input code only; no output as this will be included in D2) and attach a copy of your code.**

> **REMEMBER!** Your treatment code would include the methods you discussed in D2 to treat the data quality issues you detected with duplicates, missing values, outliers, etc.
>
> **REMEMBER!** Attach your annotated and executable (upload) your Python .py; Python via Jupyter Notebook .ipynb; R .r. These file types are accepted.
>
> **REMEMBER:** Be certain to state, *"See code attached."*

**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")

> **REMEMBER!** Review your extracted data to ensure the data is clean, the values are permissible for the variable. For example, be certain "age" is not listed as 25.2342 but rounded as a whole number 25 or negative values for a variable in which negative values cannot exist.
>
> **REMEMBER!** If you use the "exclusion treatment method" for outliers, be certain to include that dataset with your Performance Assessment.

**D6.**

**Discuss the disadvantages of the methods you used to clean/treat your data (duplicates, missing values, and outliers) that you referenced in D2. (*)**

> **REMEMBER!** This discussion should include the disadvantages of the methods used for both 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. How might the data cleaning limitations impact the pursuit of an answer to your research question? (*)**

## Part IV: PCA

**E1:**

**Perform PCA, list the variables used for PCA and provide a screenshot of the PCA loadings matrix.**

> **REMINDER!** Use as many quantitative (continuous) variables from the dataset. Note: PCA is not an appropriate method for categorical variables. Thus, do not include the categorical variables even if they are encoded to numbers.
>
> **REMEMBER!** PCA is most meaningful when using only continuous variables. This is because PCA relies on variance. Continuous data has values that are not fixed. and have an infinite number of possible values (i.e., temperature, weight)

**E2:**

**Discuss which PCs should be retained and why (visualizations such as a scree plot are helpful to include with eigenvalues plotted).**

> **REMEMBER!** You need to state which principal components are the most important and why. The Kaiser rule is one method. It states that you keep all the principal components with an eigenvalue equal to or greater than 1. The scree plot test is another method. It states that you look for the elbow in the plot and select the principal components that have the highest eigenvalues (i.e., at the elbow). Whatever method you choose you must write a paragraph describing the method used and listing the most important principal components (e.g., PC1, PC2, and PC3).
>
> **REMEMBER!** The scree plot output using the plot() function in Python's matplotlib package begins counting principal components at 0. Thus, PC1 is 0 on the scree plot, PC2 is 1 on the scree plot, PC3 is 2, etc. For example, you will retain PC1 and PC2 if the scree plot elbow is at 2.
>
> **REMEMBER!** Your PCs are not your original variables. Your PCs represent the combination of variables that are related, based on your loadings matrix.

**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.**

> **REMEMBER!** Check out *The Benefits of PCA* section at https://www.bigabid.com/what-is-pca-andhow-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, Jupyter 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.
>
> **REMEMBER!** You must be visible in the Panopto video, you must be heard, and your computer screen must be visible in the Panopto video. Recommended to keep you video to 10-15 minutes.
>
> **REMEMBER:** There are three links at the bottom of the task overview page: (1) Panopto Access; (2) Panopto FAQs; and (3) Panopto how-to videos. I encourage you to Ensure you have Panopto access as soon as possible. This access allows you to place your completed video in the respective D206 course folder. You must still upload your Panopto video link with your task submission.

**G.**

**Third-Party Code References : In this section, cite the sources you used to assist with the CODE of your work as a reference. (*)**

> *Reference Entry (Full Citation) Example:*
>
> Venkatachalam, S. (2022, September 21). *Data Science Made Simple*. Selecting Variables. Retrieved June 30, 2023, from https://www.datasciencemadesimple.com/select-variables-columns-r-using-dplyr-select-function/
>
> **REMEMBER!** Using APA citation, list all sources you used to help you with the code. If no were used, please state "no third-party code references were used".

**REMEMBER!** All sources listed here must also have **in-text citations** in the annotation of your code. See below for an example of how to include an in-text citation within the annotation (comment) of your code:

*Example of In-text citation for code*
**# Select columns of the data frame [In-Text Citation: (Venkatachalam, n.d.)]**
cars = cars_data [['mpg','cyl','disp','hp','drat','wt','qsec','gear','carb']]

**H.**

**References In this section, cite the source used to assist you with the CONTENT of your work. (*)**

**REMEMBER!** Using APA citation, list all sources you used to help you with the content (include videos, course materials, etc.). If no sources were used, please state, " No references were used."

**REMEMBER!** All sources listed here must also have **in-text citations** within the content of the paper. For suggestions on how to write in-text citations see the first two links in the Create In-Text Citations section at https://cm.wgu.edu/t5/Writing-Center-Knowledge-Base/I-Need-Help-with-APA-Style/ta-p/33524