D206 – Data Cleaning
Course Guide

NOT SURE WHERE TO START IN THE COURSE?? You are in the right place!
Your D206 instructor team has put together the following guide full of resources and course tips to help you get the most out of this course and to help you pass the course in the most efficient way possible!

Welcome to D206 – Data Cleaning!

Data Cleaning continues building proficiency in the data analytics life cycle with data preparation skills. This course addresses exploring, transforming, and imputing data as well as handling outliers. Learners write code to manipulate, structure, and clean data as well as to reduce features in data sets.

D206 RECOMMENDED TIPS AND REMINDERS!

- **Course Guidance:** Your course instructors are here to help and encourage you with the course content. Email or schedule an appointment with *any* course instructor if you have any questions about the course content. We want to help you succeed in this course!

- **Cohorts are available for this course:** Select Explore Cohort Offerings near the bottom of the course page to register for the Study Group.

- **Course Resources:** This course contains a primary learning resource--WGU Courseware Materials, which infuses two other resources:

    - The textbook: *Data Science using Python and R*. It is suggested to bookmark the textbook and/or download the textbook by chapters because it will be useful as a supplemental resource in several other classes. You will be prompted within the course when to utilize the specific chapters from this textbook.

    - DataCamp Videos: To best complete this course, it is highly recommended to use the step-by-step six-week Course Pacing Guide below.

  Other supplemental resources (e.g. instructor video recordings) are embedded in the six-week pacing plan (study guide) below.

- **D206 Textbook Data Files:** The data files for this textbook can be found within the WGU course materials and/or can all be accessed here.

- **Frequently Asked Questions:** You are encouraged to review frequently asked questions (FAQ) regarding the course and the performance assessment.

- **Pacing Guide:** To best navigate through this course, it is best that you utilize the six-week pacing guide (study plan) below. The pacing guide includes tips and other useful resources to help you navigate through the course.

- **Performance Assessment Guidance:** Before you begin working on your performance assessment, review all of the resources listed in the six-week pacing plan (study guide), below.

- **Select your programming language, either Python or R**: Review R or Python for some side-by-side comparisons. (You will continue to learn both R and Python through the MSDA program).

PRO TIP: **SAVE YOUR CODE FREQUENTLY!!** Be sure to save a copy of your code to your written document after completing each section. You can do this by retyping your code into your document, or by emailing yourself a copy of your code and then pasting the code into your document.

D206 COURSE PACING GUIDE

The amount of time any course will take you to complete depends on many factors, including your background, previous experience with the course material, and the amount of time you can devote to studying. Many students can complete this course in 5-7 weeks (see pacing options). To help you stay motivated while making progress toward your goals, your course instructor group has put together a **45-day Challenge** for this course. To meet the challenge, we suggest following the recommended pacing plan below. You can also review Completed Days-in-Course (background information).

Take the D206 Data Cleaning 45-Day Challenge!

CHALLENGE ACCEPTED

SIX-WEEKS (45-DAYS) PACING GUIDE / STUDY PLAN

Week 1

Read First: This week, you will begin your D206 Data Cleaning journey. You are encouraged to watch the introduction video, as it provides a high-level overview of the course. Additionally, this week you will begin configuring your workspace by selecting an interactive development environment (IDE) to use Python and R. It is recommended to install Anaconda Navigator as it will give you access to multiple environments (PyCharm, Jupyter Lab, R Studio, etc.) to work with both Python and R. Remember, you are not required to use both Python and R for the performance assessment, however, you do need to begin getting familiar with both environments. Therefore, this week, be certain to review all resources for both languages. All data files are accessible through the WGU Course materials.

1. Watch: Instructor-Led Videos

Video Links	PowerPoint Slides
▪ Getting Started with D206 | Course Introduction	Slides
▪ Getting Started with D206 | Data Cleaning Essentials	Slides

2. Read: WGU Course Materials

To access WGU Courseware Materials go to the D206 Course and click on the yellow button "Go To Course Materials".

Python vs. R

Lesson 1: Introduction to R and Python

Lesson 2: How to Install Python: In lesson 2, you will also import a data file. The example within the lesson assumes that you have set your working directory. Therefore, you have two options: Set your working directory or Specify the file path of the file you desire to import. Remember, your path file and the name of the file you are importing must be exact.

Lesson 3: How to Install R:

3. Watch: DataCamp

Lesson 1: Introduction to Python Recommended: Prioritize all chapters for review.

Lesson 2: Introduction to R Recommended: Prioritize all chapters for review.

Lesson 3: Introduction to Importing Data in Python: Recommended: Prioritize chapters 1 and 2 for review.

Lesson 4: Introduction to Importing Data in R Recommended: Prioritize chapters 1, 2, and 3 for review.

General Guidance for Using DataCamp

- Follow the recommendations listed in the course guide next to each DataCamp lesson.

- **Location of DataCamp Videos**
  All DataCamp videos can be located under *Course Materials > Welcome to Data Cleaning > Learning Resources > DataCamp Content*

- **DataCamp: Data Files**
  Do the following to access the data files for the resources in DataCamp.
  1. From the custom track in DataCamp (i.e. the landing page), select a course title.
  2. You will find the data files for that course at the bottom right corner of the page. Python data files are in CSV format. R data files are in FST (fast storage) format. These FST files require the fst package and use of read_fst().

- **DataCamp: PDF of Slides**
  You can download a PDF of the slides for a DataCamp chapter by selecting the page icon in the upper right corner of any of the chapter's videos. Having these slides available will make your studies more efficient because you will not need to search online for syntax help as you complete the demonstration portion after each video. You can also view the slides on the Slides tab next to the Console in the exercises. However, this view is quite small and challenging to use.

4. Supplemental Resources

Overview of Anaconda Navigator

Jupyter Notebook Tutorial | Introduction To Jupyter Notebook | Python Jupyter Notebook |

Week 2

Read First: This week you will continue practicing using both Python and R. Therefore, be certain to read, review and complete all practice with both languages. By the time you complete this week, however, you should decide if you will use either R or Python for the Performance Assessment, as it would be most helpful to focus on your language of preference beginning in Week 3.

1. **Read: WGU Course Materials:**
   To access WGU Courseware Materials go to the D206 Course and click on the yellow button "Go To Course Materials".

   Lesson 4: The Basics of Python and R: For this lesson, you will read Chapter 2 of the book *Data Science Using Python and R* to gain a clearer understanding of the basics of Python and R. Follow along with the examples for both Python in R to get experience with both languages.

   Lesson 4: Data Preparation: For this lesson, you will read Chapter 3 of the book Data Science Using Python and R to gain a clearer understanding of how to prepare data for analysis. Follow along with the examples for both Python in R to get experience with both languages.

2. **Watch: DataCamp**

   Lesson 5: Cleaning Data in Python: Recommended: Prioritize chapters 1, 2, and 3 for review.

   Lesson 6: Cleaning Data in R: Recommended: Prioritize chapters 1, 2, and 3 for review.

Week 3

Read First: This week you will learn how to detect and treat common data quality issues (i.e., duplicates, missing values, outliers). This week, it is encouraged that you select the one language that you will use for the performance assessment, and it is highly recommended to focus on that language within the reference lessons below. It is also recommended that you stick with your choice of language through D208, switch to the other language in D209 to experiment with it, then decide which language you want to continue using for the remainder of the courses. This approach is most recommended.

1. **Read: WGU Course Material**

   To access WGU Courseware Materials go to the D206 Course and click on the yellow button "Go To Course Materials"

   Lesson 5: Missing Data

   Lesson 6: Outliers

2. **Watch: DataCamp**

   Lesson 7: Dealing with Missing Data in Python Recommended: Prioritize all chapters for review.

   Lesson 8: Dealing with Missing Data in R Recommended: Prioritize all chapters for review.

3. **Watch: Instructor-Led Videos**

Instructor Video Links	PowerPoint Slides
▪ Getting Started with D206 | Detecting and Treating Duplicates	Slides
▪ Getting Started with D206 | Detecting and Treating Missing Values	Slides
▪ Getting Started with D206 | Detecting and Treating Outliers	Slides
▪ Getting Started with D206 | Re-expression of Categorical Variables	Slides

4. **Supplemental Resources**

   Python -- Tamboli, N. (Oct., 2021). All you need to know about different types of missing data values and how to handle it. Analytics Vidhya. Available at https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/

   R -- Nguyen, M. (2022). Chapter 11: Imputation (Missing Data) in *A guide on data analysis*. Available at https://bookdown.org/mike/data_analysis/imputation-missingdata.html

Week 4

Read First: This week you will be introduced to an Unsupervised Learning method, Principal Component Analysis (PCA). It is worth mentioning, that PCA is not considering a data-cleaning activity, however, similarly to the concepts of data wrangling (i.e., re-expression of categorical variables), this concept is being introduced in D206 to provide early exposure. PCA, along with other Unsupervised Learning methods will be discussed in greater detail in D212.

1. **Read: WGU Course Material**
   To access WGU Courseware Materials go to the D206 Course and click on the yellow button "Go To Course Materials".

   Lesson 7: Principal Component Analysis

2. **Watch: DataCamp**

   Lesson 9: Dimensionality Reduction in Python Recommended: Prioritize chapters 1 and 4 for review.

   Lesson 10: Unsupervised Learning In R Recommended: Prioritize chapter 3 for review.

   Lesson 11: Advanced Dimensionality Reduction in R (optional)

3. **Watch: Instructor-Led Video**

Video Links	PowerPoint Slides
▪ Getting Started with D206 | Principal Component Analysis (PCA)	Slides

4. **Supplemental Resources**
   Brems (2017) has a great overview of the PCA process if you need more detail.
   https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis5582fb7e0a9c

   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.

Weeks 5 and 6

Read First: This week, you will begin working on your Performance Assessment for submission. To complete this assessment, you will use either Python or R to clean data. To aid you in this process, it is critical that you utilize the resources provided. As you work on the performance assessment, you should consider cleaning the data first (including PCA), then completing the written report. Furthermore, be reminded that the intended purpose of the performance assessment is to assess and evaluate your technical abilities to clean data using Python and R and your writing ability to discuss the data-cleaning process performed.

D206 Performance Assessment (PA): When working towards completing the Performance Assessment it is highly recommended that you use review and use the following resources below:

1. **Watch: Instructor-Led Videos**

Instructor Video Links	PowerPoint Slides
▪ Getting Started with D206 | The Performance Assessment	Slides

2. **Review Other Helpful Performance Assessment Resources**

- D206 Performance Assessment (PA) Step-by-Step Guide: **PLEASE READ BEFORE BEGINNING THE PA**. This **critical resource** provides detailed rubric explanations about this requirement of the performance assessment task. Use this guide to help you maximize your chances of passing the PA.
- General Tips for Writing a Performance Assessment: Provides general guidance when working on a performance assessment.
- FAQ (Frequently Asked Questions): Commonly asked questions related to the course and performance assessment.
- **REMEMBER: Do not use the data from a previous class. Download the D206 data and data dictionary for D206.**
    - Go to the D206 course page.
    - Select View Task under Assessments at the bottom center of the page.
    - Select D206 Definitions and Data Files under Scenario on the Task.
    - Select the link for the dataset you will be using.
    - Unzip the downloaded folder.
    - The data file is in CSV format.
    - The data dictionary is in PDF format. Ignore the Scenario on page 1 of the PDF.
    - The Scenario is not relevant to D206.

3. **Submission Requirements for the Performance Assessment:** The PA submission should include the following below:

- **Written Report:** Addressing ALL the requirements and good use of Professional Communication: APA Format, References and Free from Grammatical Errors.
- **Cleaned Dataset:** Extract the cleaned data from the R or Python environment. Note: CSV files are accepted.
- **Panopto Video** (10-15 mins) Discuss the Programming Environment and execute error-free code.
- **Submit a copy of all your annotated code (Note: .ipynb and R files are accepted)**

If you do not pass on your first attempt, don't panic! View your evaluator's feedback by clicking on the View Task link in the Assessments section of your course page, and then clicking on the Evaluation tab on the next screen. Read your evaluator's comments and note what they ask you to change in each section that is not marked 'Competent' in green.

Where to Get Help for Each Section of the Assessment!

A

Part I: Research Question

- D206 Performance Assessment Step-by-Step Guide

B

Part I: Variables

- D206 Performance Assessment Step-by-Step Guide

C

Part II: Data Cleaning Plan! (Detection)

WGU Course Ware Materials

Lesson 5: Missing Data

Lesson 6: Outliers

Data Camp Videos

Lesson 7: Dealing with Missing Data in Python

Lesson 8: Dealing with Missing Data in R

Instructor-Led Videos

Getting Started with D206 | Detecting and Treating Duplicates

Getting Started with D206 | Detecting and Treating Missing Values

Getting Started with D206 | Detecting and Treating Outliers

Getting Started with Re-expression of Categorical Variables

PA Step-by-Step Guide

D206 Performance Assessment Step-by-Step Guide

D

Part III: Data Cleaning! (Treatment)

WGU Course Ware Materials
Lesson 5: Missing Data

Lesson 6: Outliers

Data Camp Videos
Lesson 7: Dealing with Missing Data in Python

Lesson 8: Dealing with Missing Data in R

Instructor-Led Videos
Getting Started with D206 | Detecting and Treating Duplicates

Getting Started with D206 | Detecting and Treating Missing Values

Getting Started with D206 | Detecting and Treating Outliers

Getting Started with Re-expression of Categorical Variables

PA Step-by-Step Guide
D206 Performance Assessment Step-by-Step Guide

E

Part IV: PCA!

WGU Course Ware Material:
Lesson 7

Data Camp Videos
Lesson 9: Dimensionality Reduction in Python
Lesson 10: Unsupervised Learning In R

Instructor-Led Video
Getting Started with D206 | Principal Component Analysis (PCA)

PA Step-by-Step Guide
D206 Performance Assessment Step-by-Step Guide

F

Creating Your Panopto Video!

- Online Panopto resources:
  - How to Create a Video Using Panopto
  - How to Submit Your Performance Assessment including Video Link
- **Contact Assessment Services at 877-HELP-WGU Option 2 if you:**
  - Need Panopto access / no CREATE button.
  - Can't adjust your video sharing settings / evaluator cannot access video.
  - Can't find the D206 Student Assignment's folder to put your video into

G

Using & Citing Sources!

- Every source in section I must be cited within your document.
- Student Writing Center
  - How to document sources (article)
  - How to Cite Webpages (article)
  - How to Cite Webpages (video)

H

Professional Communication!

- www.grammarly.com
- WGU Student Writing Center
- I Need Help with Professional Communication, which includes links to Writing Center resources on writing, grammar, and more!