

Data Cleaning

Webinar 1: Getting Started with D206



Welcome to
“Getting Started
with D206”

*To provide an overview of the course,
course concepts and helpful tips related to
the performance assessment.*



Course Instructors



Course Resources



Course Pacing



Course Context and Content

Topics for today's discussion

- Please use the chat to ask questions. We will do our best to address questions.
- Please stay muted throughout the webinar, as some webinars are recorded.



Course Instructors

Meet Your Course Instructors



Dr. Eric
Straw

If your assigned Course Instructor (CI) is not available, feel free to schedule time with any (CI) that supports this course.



Dr. Keiona
Middleton

Visit “Announcements” within the course to book with any CI.



Course Resources

Course Learning Resources

1. WGU Courseware Material Lessons 1-7 (uses the following text- *Data science using Python and R*)

2. Data Camp Videos – (Access Data Camp Video under ‘Learning Resources)

3. D206 Webinars (Each Tuesday – Recorded sessions are located under Course Tips)

- Webinar 1 – Getting Started with D206
- Webinar 2 – Getting Started with Missing Data and Outliers
- Webinar 3 – Getting Started with Re-expression of Categorical Variables
- Webinar 4 – Getting Started with PCA

Learning

Welcome to Data Cleaning!

This course will prepare you to demonstrate achievement of three competencies!

You will demonstrate competency through a performance assessment.

There are two prerequisites for this course:

- The Data Analytics Journey (D204)
- Data Acquisition (D205)

[GO TO COURSE MATERIAL](#)

INSTRUCTOR ▼

ANNOUNCEMENTS ▼

COURSE TIPS ▼

COURSE SEARCH

COURSE CHATTER

HOW TO ACCESS THE DATA CAMP VIDEOS!

WELCOME TO DATA
CLEANING

COURSE OVERVIEW

 Welcome to Data
Cleaning

Learning Resources

Tips for Success

SECTION 1: INTRODUCTION
TO R AND PYTHONSECTION 2: DATA SCIENCE
USING PYTHON AND RSECTION 3: MISSING DATA,
OUTLIER DETECTION, AND
PRINCIPAL COMPONENT
ANALYSIS (PCA)

Welcome to Data Cleaning

Course Overview

Learning Resources

Automatically Enrolled Resources

Throughout this course, you will be directed to learning resources housed in WGU's library. You may be prompted to log in to the WGU student portal to access these resources.

Courseware Resources

Content for the courseware is drawn primarily from the following text.

Larose, C. D., & Larose, D. T. (2019). Data science using Python and R. ISBN-13: 978-1-119-52684-1.

DataCamp Content


In addition to the included content, this course also provides access to [DataCamp \(opens new tab\)](#) to better explore the concepts of Data Cleaning.

The [DataCamp learning resource \(opens new tab\)](#) provides a highly interactive overview of data cleaning. Throughout the course you will find a variety of interactive elements that are designed to challenge and deepen your comprehension of the topics presented. It is important to note that these elements are not meant to reveal any characteristics about the format or design of the final assessment. Instead, they are designed specifically to help you learn, and are offered as tools for you to use to your advantage as you work through the course.

This course consists of **eleven** [DataCamp courses \(opens new tab\)](#) comprising 42 hours of content:


1. Introduction to Python (4 hours)
2. Introduction to R (4 hours)
3. Introduction to Importing Data in Python (3 hours)
4. Introduction to Importing Data in R (3 hours)
5. Cleaning Data in Python (4 hours)
6. Cleaning Data in R (4 hours)
7. Dealing with Missing Data in Python (4 hours)
8. Dealing with Missing Data in R (4 hours)
9. Dimensionality Reduction in Python (4 hours)
10. Dimensionality Reduction in R (4 hours)

Course Resource Guide

Webinars	WGU Courseware Resources	 DataCamp Courses for Python		
			Data Camp Chapters Priority: High	Data Camp Chapters Priority: Med/Low
Webinar 1	Lessons 1, 2 & 3	Lesson 1: Introduction to Python	All Chapters	-
Webinar 2	Lesson 4	Lesson 3: Introduction to Importing Data in Python	Chapters 1 & 2	Chapter 3
Webinar 2 and 3	Lessons 5 & 6	Lesson 5: Cleaning Data in Python	Chapters 1, 2 & 3	Chapter 4
		Lesson 7: Dealing with Missing Data in Python	All Chapters	-
Webinar 4	Lesson 7	Lesson 9: Dimensionality Reduction in Python	Chapters 1 & 4	Chapters 2 & 3

Priority High: Chapters immediately necessary for the Performance Assessment

Priority Med/Low: Chapters not immediately necessary for the Performance Assessment, but necessary for future courses.

Webinars	WGU Courseware Resources	 DataCamp Courses for R		
		Lessons	Data Camp Chapters Priority: High	Data Camp Chapters Priority: Med/Low
Webinar 1	Lessons 1, 2 & 3	Lesson 2: Introduction to R	All Chapters	-
		Lesson 4: Introduction to Importing Data in R	Chapters 1, 2 & 3	Chapter 4
Webinar 2	Lesson 4	Lesson 6: Cleaning Data in R	Chapters 1, 2, & 3	Chapter 4
Webinar 2 and 3	Lessons 5 & 6	Lesson 8: Dealing with Missing Data in R	All Chapters	-
Webinar 4	Lesson 7	Lesson 10: Unsupervised Learning In R	Chapter 3 Only	-
		Lesson 11: Advanced Dimensionality Reduction in R	None	Chapters 1-4

Priority High: Chapters immediately necessary for the Performance Assessment

Priority Med/Low: Chapters not immediately necessary for the Performance Assessment, but necessary for future courses.

Resource Reminders!

- Review Course Announcements, regularly
- Review all information found under course tips.
 - Webinar Recordings
 - Dr. Straw's Tips for Success ([click here](#))
 - Frequently Asked Questions ([click here](#))
 - Additional Supplemental Resources
 - Performance Assessment Helpful Guide / Tips ([click here](#))



Course Pacing

Course Pacing Options

60-Days

- Reduce stress and allow for life's interruptions.

45-Days

- Requires more daily time commitment and has less flexibility than the Relaxed Pace.

30-Days

- Requires significant daily commitment, does not easily accommodate life's interruptions, and may increase stress.



60-Days Relaxed Pace

Time Allocation	Tasks
25 Days	Study course material for 25 days
15-28 Days	Work on your task submission for 15-28 days Note: Work on task revisions if needed
7 days	Celebrate finishing with time to spare



45-Days

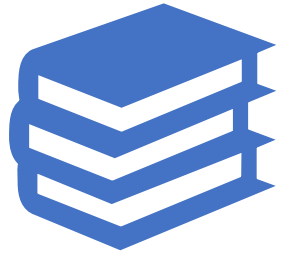
Comfortable Pace

Time Allocation	Tasks
20 days	Study course material for 20 days.
20-23 days	Work on your task submission for 20-23 days Note: Work on task revisions if needed
2 days	Celebrate finishing with time to spare!



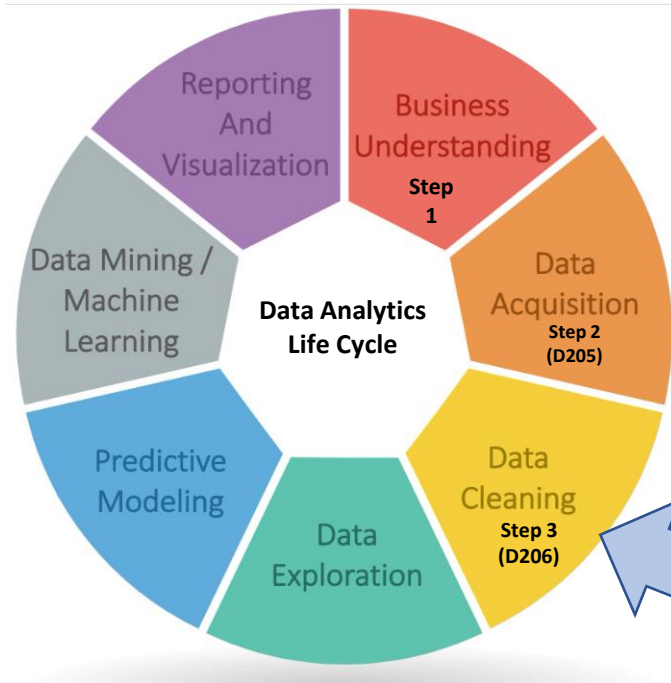
30-Days Tight Pace

Time Allocation	Tasks
10 days	Study course material for 10 days
10-17 days	Work on your task submission for 10-15 days Note: Work on task revisions, if needed
3 days	Celebrate finishing just in time



Course Content

About The Course (D206 Data Cleaning)



- Data cleaning is the process of removing incomplete, irrelevant, corrupt, or otherwise faulty – or dirty – data from a data set.
- Therefore, your goal in this course is to *Clean the dataset*.
- In this course, you will be introduced to both Python or R, and encouraged to employ these tools to facilitate common data cleaning techniques.
- Your competence will be demonstrated by the successful completion of the performance assessment.

YOU ARE HERE!

D206 Course Concepts:

1. This course will introduce three elements of Data Cleaning.

- Duplicates
- Missing data (N/As)
- Outliers

While this is not an exhaustive list of issues related to dirty data, these are some of the more common issues data scientists experience (Duplicates, variable name uniformity)

2. This course will introduce basic applications of Data Wrangling.

- The process of restructuring raw data in ways beneficial for analysis, such as recoding values
- Example: Re-expression of Categorical Field Value

[Note: You will learn more about this in covered in future courses. Some of the techniques discussed in this course can be considered advanced, but this course will introduce this concept.]

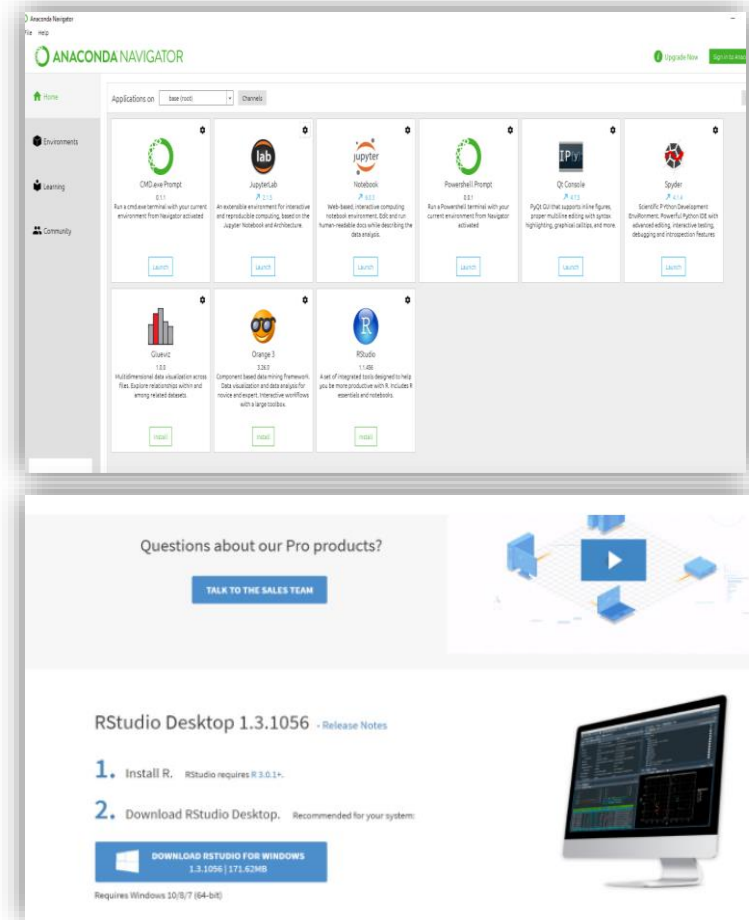
3. This course will introduce one element of Feature Engineering.

- The process of transforming data to better represent their characteristics or features, such as data reduction (reducing data down to meaningful parts)
- For example, PCA(Principal Component Analysis)

Lesson 2 & Lesson 3:

Installation of R/Python

- Download Anaconda Navigator as suggested in the course materials.
- Anaconda Navigator is a graphical interface for launching common Python programs without having to use command lines, to install packages and manage your environments.
- Anaconda Navigator provides the following applications by default :
 - *JupyterLab*
 - *Jupyter Notebook (Python)*
 - *Spyder*
 - *Orange 3 App*
 - *R Studio*
 - *And many*



Getting Started with Python and R

- Install Anaconda Navigator (if you are using R, you can just install R Studio)

<https://www.youtube.com/watch?v=92A6px29Hfk>

<https://www.youtube.com/watch?v=YU7ZGgPKSsA>

<https://www.rstudio.com/products/rstudio/download/>

- Import necessary packages, modules and libraries you will need

Package	Definition of Terms	Common Libraries
Python	https://realpython.com/lessons/scripts-modules-packages-and-libraries/#:~:text=Modules%20are%20Python%20files%20that,to%20achieve%20a%20common%20goal.	https://www.kdnuggets.com/2021/03/top-10-python-libraries-2021.html
R Studio	https://www.tutorialspoint.com/r/r_packages.htm	https://support.rstudio.com/hc/en-us/articles/201057987-Quick-list-of-useful-R-packages

- Import your data file

Python: <https://www.listendata.com/2017/02/import-data-in-python.html#Import-CSV-files>

R Studio: <http://r-tutorial.nl/ch7.html>

- Navigating Jupyter and R Studio

Jupyter (Python):

<https://www.youtube.com/watch?v=1A7tea9LSEk>

<https://www.youtube.com/watch?v=7wfPqAyYADY>

R Studio:

<https://www.youtube.com/watch?v=FlrsOBy5k58>

Lesson 4:

Data Science Using Python and R

- Review section 3.1-3.6 in the course textbook *“Data Science Using Python and R”*
- Utilize the Training Data Set (can be found in the WGU courseware material)
- You will be introduced to both R and Python (you are encouraged to use both for the exercises; once you complete these exercises,
- When you complete Lesson 4, select a language that you will utilize for the remainder of the course, select a language that you will utilize for the remainder of the course.

Lesson 5:

Missing Data

Missing Data are values in a dataset that seems missing (blanks) and/or null value (N/A), research or data collection.

Missing data could adversely impact analysis.

Methods of Treating Missing Data

Deletion

- Removing Variables (Columns)
- Removing Observations (Rows)

Imputation

- Univariate Statistical Imputation (Mean, Median, Mode)
- Backward/Forward Fill
- Multiple Imputation by Chained Equations (MICE)
(usually for R users)
- Iterative Imputer *(usually for Python users)*
- K-Nearest Neighbor (kNN) Imputation

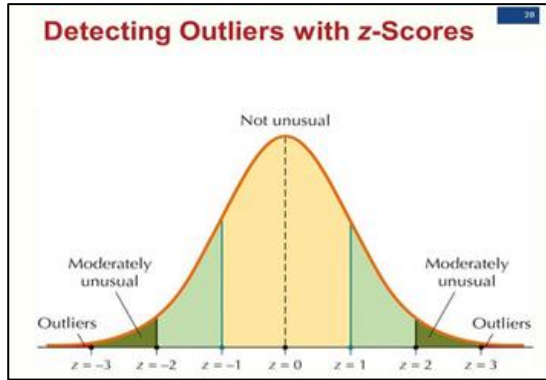
Lesson 6:

Outliers

- Outliers are points that are far from the others (typically in an abnormal way).
- In general, they can create analytic challenges by distorting individual measures or relationships and potentially leading to mistaken conclusions.
- Some of the most common causes of outliers in a data set include:
 - ✂ data entry/processing errors (human errors)
 - ✂ measurement errors (instrument errors)
 - ✂ sampling errors (extracting or mixing data from wrong or various sources)
 - ✂ natural (not an error, novelties in data)
- Thus, it is important to understand their presence in data sets. Next, we will review three methods discussed in this course to identify outliers.

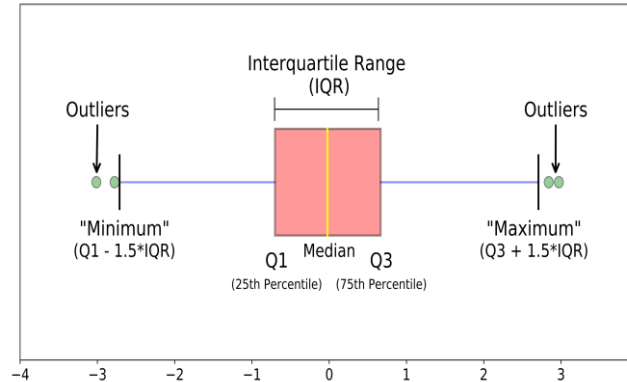
Detecting and Identifying Outliers

Z-Scores



A Z-score is a numerical measurement that describes a value's relationship to the mean. The further away an observation's Z-score is from zero, the more unusual it is. A standard cut-off value for finding outliers are Z-scores of ± 3 or further from zero.

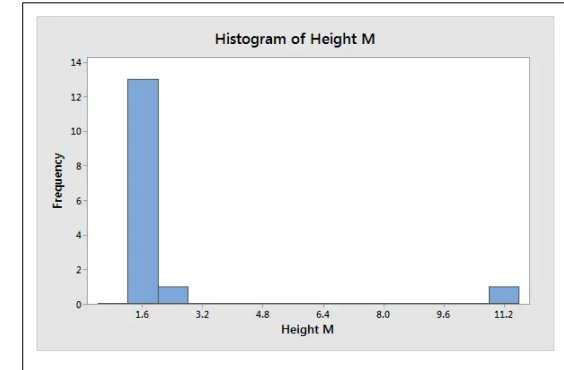
Boxplots



Box plots are useful as they provide a visual summary of the data to quickly identify mean values, the dispersion of the data set, signs of skewness, and outliers.

Values that fall outside the two inner fences (minimum and maximum) are outliers.

Histograms



A histogram is the most used graph to show frequency distributions.

Histograms emphasize the existence of outliers. Look for isolated bars.

Treating Outliers

- If you think it is a legitimate factual error, then you can remove it.
- If you find out that it is not an error or a legitimate entry, then you can retain the outlier, but note that it exists.
- If you find out that it is not an error or a legitimate entry, then you can exclude the outlier, from your dataset.
- You can replace the outlier with median values- In this technique, we replace the extreme values with median values. It is advised to not use mean values as they are affected by outliers.



Lesson 7:

Principal Component Analysis

- Principal Component Analysis (PCA) is a statistical techniques used to reduce the dimensionality of the data.
- It involves the process of finding linear combinations of variables that best explain the covariation (correlated or connection among variables).
- Principal components represents the combination of variables based on correlation/connection/shared properties/likeness (*New Set of Variables (PCs) from old set/initial of variables*)

PCA Related Terms

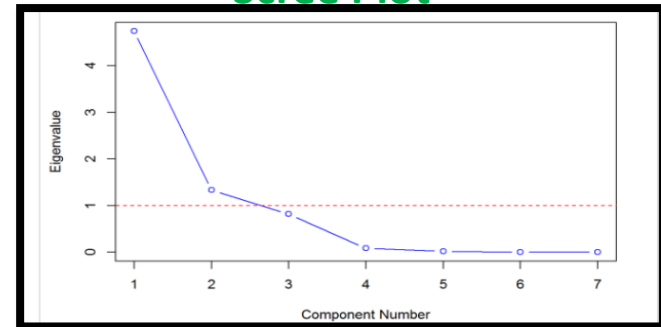
- **PCA Loadings:** Loadings are interpreted as the coefficients of the linear combination of the initial variables from which the principal components are constructed.
- **Eigenvalue:** It is a number assigned to a component to represent how good it is at explaining the variance in a given dataset. Greater than 1 means the principal component (grouping) is better at explaining the variance in the dataset.
- **Scree Plots:** A scree plot is a graphical tool used in the selection of the number of relevant to be considered in a principal component analysis.

PCA Loadings Output

	PC1	PC2	PC3	PC4	PC5	PC6
mpg	-0.3931477	0.02753861	-0.22119309	-0.006126378	-0.3207620	0.72015586
cyl	0.4025537	0.01570975	-0.25231615	0.040700251	0.1171397	0.22432550
disp	0.3973528	-0.08888469	-0.07825139	0.339493732	-0.4867849	-0.01967516
hp	0.3670814	0.26941371	-0.01721159	0.068300993	-0.2947317	0.35394225
drat	-0.3118165	0.34165268	0.14995507	0.845658485	0.1619259	-0.01536794
wt	0.3734771	-0.17194306	0.45373418	0.191260029	-0.1874822	-0.08377237
qsec	-0.2243508	-0.48404435	0.62812782	-0.030329127	-0.1482495	0.25752940
gear	-0.2094749	0.55078264	0.20658376	-0.282381831	-0.5624860	-0.32298239
carb	0.2445807	0.48431310	0.46412069	-0.214492216	0.3997820	0.35706914

	PC7	PC8	PC9
mpg	-0.38138068	-0.12465987	0.11492862
cyl	-0.15893251	0.11032177	0.16266295
disp	-0.18233095	-0.06416707	-0.66190812
hp	0.69620751	-0.16573993	0.25177306
drat	0.04767957	0.13505066	0.03809096
wt	-0.42777608	-0.19839375	0.56918844
qsec	0.27622581	0.35613350	-0.16873731
gear	-0.08555707	0.31636479	0.04719694
carb	-0.20604210	-0.10832772	-0.32045892

Scree Plot



Regardless of your research question, you should ...

Detect/Identify and Treat Duplicates

Detect/Identify and Treat Missing Data

Detect/Identify and Treat Outliers

Re-express Categorical Variables, if possible

Perform PCA (with numerical variables only)

Consider cleaning the data first (including PCA), then complete the written report.



The Performance Assessment

Overview of the Performance Assessment

1. Written Report (Addressing ALL the requirements)

- Word documents are helpful to use.
- Use headers (helps with reviewing the assessment).
- Professional Communication: APA Format, References and Free from Grammatical Errors].
- Provide meaningful visualizations.

2. Cleaned Dataset

- Extract the cleaned data from the R or Python environment.

3. Panopto Video (10-15 minutes)

- Programming Environment
- Execution (Code)
- Discuss your process/plan

4. Submit a copy of all your code (very highly recommended) several components of your code will be required in the submission of your report.

Note: .ipynb and R files are accepted.

Extracting Your Data frame



Using Python:

`df.to_csv(r'Path where you want to store the exported CSV file\File Name.csv')`

Using R:

`write.csv(Your DataFrame,"Path to export the DataFrame\\File Name.csv")`

- Must use “\\” between all folders in the path

Professional Communication

Visit the writing center to provide assistance with Professional Communication

<https://my.wgu.edu/success-centers/writing-center>

Attend Live Workshops and/or view recorded sessions:

- Writing for Performance Assessments: Before You Begin ([click here](#) for recording)
- Writing for Performance Assessments: Get Started ([click here](#) for recording)

*It is encouraged for you to utilized this document prior
to starting the performance assessment*

[Click here:](#)
[D206 Performance Assessment](#)
[Helpful Tips and Guide](#)

Words of Encouragement!

1. Focus on the process and functions for cleaning data. You will have the entire duration of this program and your lifetime to learn Python and R.
2. Feel free to use other resources to get acclimated with Python and R.
3. Work on the performance assessment, one requirement at a time.
4. Reach out to your Community of Care (Program Mentor / Course Instructor) if you have any questions!

Questions?