Dr. Straw's Tips for Success in D206

Please send all questions and suggestions with the subject: "D206 tips suggestion" to eric.straw@wgu.edu.

These tips provide my suggestions as well as answers to the most common student questions. Tips are sorted in alphabetically by title. You may want to search this document for key words if you are looking for a specific tip.

Data and Data Dictionary

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 Overview page
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.

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.

Data Files for Practice

The data files for the D206 textbook can be found in Dr. Straw's D206 resource folder.

The *Boston Housing Data.csv* file, which is referenced in the *Importing Data l*esson (Section 1, Lesson 2: Importing Data for Python; Section 1, Lesson 3: Importing Data for R), is also available in *Dr. Straw's D206 resource folder*.

Getting Started

You should start this course by watching Dr. Middleton's *Webinar 1: Getting started with D206*. Watch the remaining webinars as you are studying the relevant material. You can find a recording of all webinars and the slides used in each webinar on the course page. Select *Course Tips* on the right-hand side, then select *View All*. At a minimum, you should review the PowerPoint slides for each webinar.

Dr. Straw's Tips for Success in D206

Imputation

Do you need more help with imputing missing values? If so, the resources listed below can fill in many of the knowledge gaps for you.

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-missing-data.html

Imputation: Categorical Variables

The D206 task (Requirement III.D) requires you to clean all of the data. This includes replacing missing values for categorical variables (i.e., imputation of categorical variables).

Review Dr. Middleton's webinar #2 for an overview of imputation methods. The webinars can be located by selecting Course Tips on the right-hand side of the course page, then selecting View All? These webinars will save you time and prevent many of the common mistakes.

There are several ways to impute missing values for categorical variables.

1. Mode imputation: Replace the missing values with the most frequent value. This method can alter the distribution of the variable. Consider the percentage of missing values and the ratio of the categories.
2. Ratio imputation
3. MICE imputation (multivariate imputation by chained equation)

Language: Python or R

Choose either Python or R for D206, but not both. Complete the DataCamp and textbook reading for the language you choose, but not for the other language. The choice is completely up to you. You should consider what your industry and company use.

Stick with your choice 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.

Medical data set

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.

PCA

PCA requires numerical variables, and the results of a PCA are most meaningful when using only continuous variables. This is because PCA relies on variance. Variance is the sum of the squared deviation of all observations of an attribute from the mean of that attribute divided by n-1 (i.e. one less than the sample size).

Variance is in the squared units of measurement of the data. Thus, it is much easier to look at examples with standard deviation, which is the square root of variance and is in the same units of measurement as the data.

Example 1: Continuous data: A set of temperatures in Fahrenheit with a mean of 82.22 degrees and a standard deviation of 4.53 degrees. This makes intuitive sense.

Example 2: Categorical data: A count of items on three shelves in a grocery store. The three shelves can be labeled 1, 2, and 3. What does a mean of 2.25 tell you? It tells you that the count skews to shelves 2 and 3. But this information should be found with counts and percentages, which will provide a

more accurate picture. What does the standard deviation of 0.776 shelves tell you? Nothing useful. Temperature can be 87.75 degrees, but items on these shelves must either be on shelf 1, shelf 2, or shelf 3. Items cannot be on shelf 2.81. This is true for all categorical data, including binary data, multicategory data, dummy variables, label encoded variables, and geographic data like ZIP code.

Thus, you can feed categorical numeric data into the PCA process, but you should not.

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-analysis-5582fb7e0a9c

Research Question

Requirement I.A. states, "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."

Your research question should be broad enough to encompass the entire dataset. Writing, "Does Y cause X?" or "Is Y correlated with X?" are insufficient. Rather, you should write something like, "Given the available data, can I determine why X is..."

X and Y are names of attributes/variables/columns in my example dataset.

Your research question only impacts task requirement D7. 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.

Resources

D206 provides four learning opportunities. You should complete the chapters and videos for the tool (i.e. Python or R, but not both) that you have selected to use to complete the task.

- Four webinars by Dr. Middleton will save you time and prevent common mistakes. You should start this course by watching Webinar 1: Getting started with D206. Watch the remaining webinars as you are studying the relevant material. You can find a recording of all webinars and the PowerPoint slides used in each webinar on the course page. Select Course Tips on the right-hand side, then select View All. At a minimum, you should review the PowerPoint slides for each webinar.
- The Study Plan, which is found by selecting the orange Go to Course Material button in the center bottom of the course page, guides you to specific chapters in the D206 textbook, Data Science Using Python and R. Save the link to this textbook. It has chapters that will help you in future courses.
- Data Camp video series: Five for Python or six for R. Complete either Python or R, but not both. Select the DataCamp link in the Learning Resources section of the Study Plan. Once in Data Camp select the green Enroll button the first time you access these learning resources. Practice coding by duplicating all the examples (Python or R) that you study. You cannot learn to code by reading about code or watching someone else code. You must code to learn to code.
- Study Group, which is held every Tuesdays at 7:30 PM Eastern Time and is led by Dr. Middleton. Select Explore Cohort Offerings near the bottom of the course page.
- R User Group, which is held most Thursdays at 9:00 PM Eastern Time and is led by Dr. Straw. Select Explore Cohort Offerings near the bottom of the course page.

Task Requirement Overview

All requirements in the task require you to work with the whole data set.

- You must write a research question that includes the whole data set (Requirement I.A)

- You must describe all variables in the data set (Requirement I.B)
- You must develop a cleaning plan for all the data (Requirement II)
- You must clean all the data (Requirement III.D)
- You must run PCA on all the continuous variables in the data set (Requirement III.E) [Note: PCA is not an appropriate method for categorical variables. Thus, do not include the categorical variables even after they are encoded to numbers.]

There are a few exceptions to the rules above. The following variables should be described (Requirement I.B) but should not be cleaned (Requirements II and III.D).

1. CaseOrder is a sequence number and is not considered a variable in the data set
2. Interaction and UID are encoded customer ID information
3. Zip – In a corporate setting the ZIP code would be cleaned based on a master ZIP code lookup file
4. Lat and Lng – In a corporate setting the latitude and longitude would be cleaned based on a lookup database

Task Requirement D7

Task requirement D7 requires you to link the limitations of your data cleaning (which you describe in D6) back to your research question in task requirement A.

How might the data cleaning limitations impact the pursuit of an answer to your research question?

Task Requirement E2

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 eigen value of 1 or higher. 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 eigen values (i.e., at the elbow). Whatever method you choose you must write a paragraph describing the method and listing the most important principal components (e.g., PC1, PC2, and PC3).

Task Requirement E3

For task requirement E3 you need to 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-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.

Task Requirement H

You are not required to follow APA or any other strict writing guide for references and citations. However, APA provides an adequate format to emulate.

Every item listed in section H must be cited in your 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.

A few details:

• Use the last names of authors (e.g. LoDolce in the Format References: Basic Principles example via the link above) or, if the last names are not provided, use the publisher name (e.g. Obesity Action Coalition in the Format References: Basic Principles example via the link above).

• n.d. means No Date. Use either the date of the publication or use n.d. if the date is not provided by the publisher.

Textbook

Here is a direct link to the D206 textbook *Data science using Python and R* used in the Study Plan. I suggest bookmarking this textbook because it will be useful as a supplemental resource in several other classes.

The data files for this textbook can be found in Dr. Straw's D206 resource folder.

Scree Plot with Python

The scree plot output using the plot() function in Python's matplotlib package begins counting principle 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. See Dr. Middleton's webinar #4 for an example scree plot where you can compare identical results between Python (starts counting at 0) and R (starts counting at 1).