

Student's Name

Professor Name

University Affiliation

Course Number

Date of Submission

2:37





Assignment Details

INFO 5770 Section 021 - Introduction to Health Data Analy...

Submission & Rubric



## Description

### Aim

The purpose of assignment 2 is to exercise skills related to data preprocessing. This assignment is somehow related to the term project, but you need to use the given dataset about asthma ([asthma\\_data\\_2016.csv](#)). The assignment requires you to submit three files. The first one is the Jupyter Notebook file, where you need to write your own code based on "week6\_data\_processing.ipynb" and "week7\_data\_preprocessing.ipynb". You need a little modification to complete data preprocessing and answer questions. Please note that these Jupyter notebooks are processing 2009 data, and you need to process 2016 data. The second file is an MS word document where you need to provide your answers to 10 questions. The last file is a CSV file to which data preprocessing is applied. Please note that this assignment should be individual work. More details will be explained below.

Submit Assignment



2:38



## Assignment Details

INFO 5770 Section 021 - Introduction to Health Data Analy...

## What to Do for the Assignment 2

1. Open the file (asthma\_data\_2016.csv) by using pandas' read\_csv() method.
2. What are the number of rows (tuples) and columns (attributes) of this file? (10 points)
3. What is the value of row index 1494 and column index 229? (10 points)
4. Select the following variables: ["DUPERSID", "SEX", "ASPRIN53", "ADAPPT42", "ADHECR42", "AGE16X", "BMINDX53", "CHBMIX42", "FAMINC16", "WAGEP16X", "TTLP16X", "UNEMP16X", "RACETHX", "TOTEXP16"]. Then, assign the selected attributes to a new data frame, called asthma\_selected. (10 points)
5. Show (i.e., print) the values of DUPERSID and SEX in rows from 150 to 160. Here, numbers indicate row indexes. (10 points)
6. By using describe method(), examine the distribution of data and answer two questions.
  - Which attribute does seem to have the most missing values? (5 points)
  - List the names of categorical variables. Explain why you think they are categorical variables. (5 points)
7. Change missing values of ADHECR42 to the mean value of this attribute. (10 points)
8. Remove tuples if more than three attributes'

Submit Assignment



Dashboard



Calendar



To Do



Notifications



Inbox

2:38



## Assignment Details

INFO 5770 Section 021 - Introduction to Health Data Analy...

- asthma\_selected. (10 points)
5. Show (i.e., print) the values of DUPERSID and SEX in rows from 150 to 160. Here, numbers indicate row indexes. (10 points)
  6. By using describe method(), examine the distribution of data and answer two questions.
    - Which attribute does seem to have the most missing values? (5 points)
    - List the names of categorical variables. Explain why you think they are categorical variables. (5 points)
  7. Change missing values of ADHECR42 to the mean value of this attribute. (10 points)
  8. Remove tuples if more than three attributes' values are below  $Q1 - 1.5 \times IQR$  or above  $Q3 + 1.5 \times IQR$ . In other words, remove outliers. How many tuples are left? (10 points)
  9. Combine BMINDX53 (for adults, >17) and CHBMIX42 (for children, 0-17) into one variable called BMINDX. (10 points)
  10. We need to avoid redundancy in attributes. We can use the correlation coefficient for numeric data and the Chi-square test for categorical data. After applying the correlation coefficient, what attribute would you drop? Explain your rationale. (10 points)
  11. Apply Max-Min normalization to numerical attributes. (10 points)
  12. Save the file as "asthma\_data\_processed.csv".

Submit Assignment



Dashboard



Calendar



To Do



Notifications



Inbox

2:38



 **Assignment Details**  
INFO 5770 Section 021 - Introduction to Health Data Analy...

11. Apply Max-Min normalization to numerical attributes. (10 points)
12. Save the file as "asthma\_data\_processed.csv".

## How to write

For the assignment 2, I will not put constraints on the format for the MS word report. You simply need to provide answers and rational if asked. Please include the title, your name, and EUID. Up to this part, you can write the report in a **MS word file**. You also need to submit your **Jupyter notebook** you used for data preprocessing and the final data file (csv).






## What to include

Your submission of the report should include:

- Your name
- EUID
- Title
- What I asked you in the "How to Write"

Please attach a .docx file for the report and .ipynb and .csv files for the data processing. Again, this is an individual assignment.

**Submit Assignment**

 Dashboard  Calendar  To Do  Notifications  Inbox