

Assignment 3

Objective(s) - The student will demonstrate a synthesis of the knowledge gained in this course by producing a mini data project that reflects knowledge of patient data, analysis and softwares used.

1. Healthcare/Medical Data
 - a. The dataset needs to be healthcare, medical or even research data that is related to medicine.
 - b. Need to have at least 100 rows of data, and 5 columns. You can take two different datasets and make relevant comparisons/correlations, or grab from different sources and combine them.
 - c. Provide a brief description of your data.
2. Data Collection:
 - a. Provide source(s)/methods of your data/data collection
 - i. Links like Kaggle, CDC, WHO, etc. If text mined, or used an API or NLP, provide a brief summary.
3. Data Cleaning/Wrangling/Data "Manipulation"
 - a. Provide a brief description how you cleaned, and wrangled the data such as removing columns not needed, combining two datasets, standardizing different formats, etc. If data is already cleaned, state that it is so.
 - b. Creation of variables/indicators, etc. The created variables/indicators can be a part of the required 5 columns.
4. Data Dictionary
 - a. Your variables and their meaning.
5. Data Analysis
 - a. Provide **10 relevant or important** questions/hypotheses that your data can answer.
 - b. Provide answers to your given questions using statistics/visualizations/charts.
 - c. Provide statistical methods/machine learning algorithms you used to analyze data.
 - d. State any software you used (R, SAS, Excel, Python, SQL, etc)

What you will submit/Deliverables:

- A zip file containing the following:
 - **Need to have your dataset approved first.**
 - A word document/pdf that has the following:
 - A brief description of your data and your reason(s) for choosing your dataset(s),
 - A description of your data collection methods/sources.
 - A paragraph describing your data cleaning/wrangling process.
 - 10 questions you formulated and the corresponding answers/charts/visuals to your own questions about your data, including machine learning algorithms or statistical analysis methods you applied to get your answers, and what software(s) you used.
 - A separate page for your data dictionary

- Your citation/sources such as links, etc.
- Your dataset in an excel/csv/sql file.
- Your code in .r, .sql, .py or even in a txt format. If you used plain excel, copy paste formulas and copy/paste on a notepad.

Failure to submit a project with an instructor approved dataset, will result in a grade of 0. Please submit your dataset to be approved via the Assignment 3 Dataset Submission discussion by the posted deadline.

Grading Rubric:

Assignment 3:	Points
Type of data is Healthcare/Medical/Medicine Research Data including a copy of data	20
Data Collection and sources	10
Data Wrangling/Cleaning and Code/Formulas	15
Data Analysis: Methods and Relevant Questions/Answers	50
Data Dictionary	5
Total	100/100