DS 4320 Midterm Exam: Take-home

Instructions

- Put your solutions for all problems in a google colaboratory notebook. Create a level-1 header for each problem and then present the solution in subsequent cells. The header must contain the problem number followed by "confidence x" where x is an integer from 1 to 10 with 10 representing high confidence. For full credit you must show your work. The submission is made via a link on UVACanvas.
- Include an honor pledge at the start. Type your full name as your digital signature.
- You are allowed to access all resources as you normally would during a homework assignment with one exception. You must work individually, you are not allowed to consult other people.
- The assignment due date is shown on canvas. It is estimated to take 3 hours, you are allowed to take as much time during the release window as you like.

Data

- For Question 1 the given data can be downloaded from GitHUb here `https://github.com/UVADS/DS-4320/blob/main/question-1.csv` or from sharepoint here `https://bit.ly/4r07UZV`. It is a csv file with a size of 40 KB.
- For Question 2 the "raw" data is the course information for UVA. You may access it however you wish, for example via Lou's List here `https://hooslist.virginia.edu/`.
- For Question 3 we are using the same data as in homework set 5, Synthea. The data is available from the synthea downloads site `https://synthea.mitre.org/downloads`. Use the file linked behind "1K Sample Synthetic Patient Records, CSV ". It is a zip file that is 56 MB.

1 Create and Evaluate Synthetic Data (10ψ)

The goal for this assignment is to make a synthetic data generator to mimic the data in the given csv file, documentation on how to use the generator, plots demonstrating a good match, and a statistical test demonstrating a good match.

Specifications

1. Data Generator
   (a) Written in python
   (b) A function that takes one argument, n, with a default value of 10
   (c) Returns n rows of synthetic data that mimic the properties of the given csv
2. Documentation
   (a) Explain what your function does.
   (b) HINT: Use standard numpy documentation as a guide (e.g. numpy.random.randn)
3. Plots

(a) Use your generator to produce a histogram with 1,000 trials for each column in the data set.

(b) Show your code to produce the histograms.

(c) Demonstrate the quality of your synthetic data by plotting your histograms on top of histograms generated from the given csv. (Make sure all are visible and the legend clearly indicates which is which).

4. Statistical Test

   (a) Conduct a statistical test to assess how well your synthetic data mimics the given data.

   (b) Show the code.

   (c) Write a paragraph explaining your thinking and result for the test.

HINT: Study the csv first using your exploratory data analysis skills. Hypothesize how it was generated. Then start coding. Then iterate.

2 Create a Dataset with a Data Dictionary (10ψ)

The goal for this assignment is to create a dataset with a data dictionary, sample code, and plots.

Specifications

1. The dataset

   (a) is in csv format, stored in the cloud, and linked in your submission.

   (b) The dataset has the following features: Course Mnemonic, Course Title, Instructor, Room, Day(s) of the week, Start time, End time

   (c) Contains all courses at UVA with a DS mnemonic from this semester

   (d) Show any code you used in the creation of the dataset

2. The data dictionary is in markdown format

3. The sample code is in python, and performs the following tasks

   (a) imports necessary packages

   (b) loads the data into a dataframe

   (c) converts times into datetime object or equivalent

   (d) plots a histogram showing the start times of classes

   (e) plots a histogram showing the classes by level (1xxx, 2xxx, 3xxx, etc.)

   (f) **REMINDER:** Include your plots in the submitted notebook

REMINDER: Don't forget to check the permissions on your dataset file so that we can grade it.

3 Working with the Relational Data Model (10ψ)

The goal of this exercise is to load the Synthea data into a MySQL database, query it to generate a finding, and report the finding in a press release.

REMINDER: Use the MySQL process from DS 2022. Alternatively you may use DuckDB. SQLite is prohibited.

HINT: This question requires creative exploration on your part. Use the documentation for the dataset to aid your exploration. You are required to come up with a finding from the dataset on your own.

1. Creating the database
   (a) Create a MySQL database
   (b) Load the csv files into the database as tables
   (c) Show your SQL code
2. Querying the database
   (a) Query the database to produce your finding
   (b) Show your SQL code
   (c) Use at least 4 tables
3. Plot for press release
   (a) Show your code to generate your plot
4. Press Release
   (a) Headline
   (b) Motivation
   (c) Investigation details
   (d) Supporting plot
   (e) Call to action