Clinical Trail Data Analysis

Learning outcomes of this assessment

- Provide a broad overview of the general field of 'Data Science '
- Developing Specialised knowledge in areas that demonstrate the interaction and synergy between ongoing research and practical deployment of this field of study.

Key Skills to be assessed

- The usages of common data engineering tools and techniques
- Your ability to implement a standard data analysis process
  - Loading the data
  - Cleansing the data
  - Analysis
  - Visualisations / Reporting
- Use of Python, SQL and Data Science Libraries

Task

You will be given a dataset and a set of problem statements. You are required to implement your solution to each problem

General instructions

You will follow a typical data analysis process:

1. Load / ingest the data to be analysed
2. Prepare / clean the data
3. Analyse the data
4. Visualise results / generate report

For steps 1, 2 and 3 you will use environments that have been used within this module. The data necessary for this assignment will be downloadable as .csv files.

The .csv files have a header describing the file's contents. They are:

1. clinicaltrial <year>.csv: each row represents an individual clinical trial, identified by an Id, listing the sponsor (Sponsor), the status of the study at time of the file's download (Status), the start and completion dates (Start and Completion respectively), the type of study (Type), when the trial was first submitted (Submission), and the lists of conditions the trial concerns (Conditions) and the interventions explored (Interventions). Individual conditions and interventions are separated by commas. (Source: ClinicalTrials.gov)
2. mesh.csv: the conditions from the clinical trial list may also appear in a number of hierarchies. The hierarchy identifiers have the format [A-Z][0-9]+(˙[A-Z][0-9]+)* (such as, e.g., D03.633.100.221.173) where the initial letter and number combination designates the root of this particular hierarchy (in the example, this is D03) and each "." descends a level down the hierarchy. The rows of this file contain condition (term), hierarchy identifier (tree) pairs. (Source: U.S. National Library of Medicine.)
3. pharma.csv: the file contains a small number of a publicly available list of pharmaceutical violations. For the puposes of this work, we are interested in the second column, Parent Company, which con- tains the name of the pharmaceutical company in question. (Source: https://violationtracker. goodjobsfirst.org/industry/pharmaceuticals)

When creating tables for this work, you must name them as folows:

- clinicaltrial 2021 (and clinicaltrial 2019, clinicaltrial 2020 for the sample data)
- mesh
- pharma

The data

You will be using clinical trial datasets in this work and combining the information with a list of pharma- ceutical companies and condition hierarchy information. You will be given the answers to the questions, for a basic implementation, for two historical datasets, so you can verify your basic solution to the prob- lems. Your final submission will need to consist of results executed on the third, 2021, release of the data. All data will be available with task.

Problem statements

You are a data analyst / data scientist whose client wishes to gain further insight into clinical trials. You are tasked with answering these questions, using visualisations where these would support your conclusions.
You should address the following problem statements. You should use the solutions for historical datasets (available on Blackboard) to test your implementation.

1. The number of studies in the dataset. You must ensure that you explicitly check distinct studies.
2. You should list all the types (as contained in the Type column) of studies in the dataset along with the frequencies of each type. These should be ordered from most frequent to least frequent.
3. The top 5 conditions (from Conditions) with their frequencies.
4. Each condition can be mapped to one or more hierarchy codes. The client wishes to know the 5 most frequent roots (i.e. the sequence of letters and numbers before the first full stop) after this is done.
   To clarify, suppose your clinical trial data was:
   NCT01, ... ,"Disease_A,Disease_B",
   NCT02, ... ,Disease_B,

   And the mesh file contained:
   Disease_A A01.01 C23.02
   Disease_B B01.34.56

   The result would be

   B01 2
   A01 1
   C23 1
5. Find the 10 most common sponsors that are not pharmaceutical companies, along with the number of clinical trials they have sponsored. Hint: For a basic implementation, you can assume that the Parent Company column contains all possible pharmaceutical companies.
6. Plot number of completed studies each month in a given year – for the submission dataset, the year is 2021. You need to include your visualization as well as a table of all the values you have plotted for each month.

Report

A 3000 word report that documents your solution should be included with your submission. In this module, a background, literature review or citations are not required.