Data Team Live Coding Exercise

Summary

The purpose of this exercise is to evaluate your ability to comfortably work through a problem in your environment/language of choice and perform basic data processing/transformation work.

Background

Meseeks Inc. is a global pharmaceutical company working with Reify Health. Meseeks is currently recruiting patients at multiple clinical sites across the United States for a clinical trial codenamed ANDROMEDA. ANDROMEDA is a phase III clinical trial for the FDA to evaluate the effectiveness of a new cancer drug. Reify Health has provided StudyTeam to clinical sites to help Meseeks recruit and enroll new patients into ANDROMEDA more effectively. In return, Meseeks has provided us with two data files: one has information about the clinical sites involved in the ANDROMEDA trial and the other has information about patients recruited and processed at those sites. We would like to perform some basic data processing on these files.

Provided Data

andromeda_site_information_report_2018.csv

Information about clinical sites related to the ANDROMEDA trial, including:

- **Site Number**: Integer identifier of a specific clinical trial site
- **Site Status**: The current status of the site in terms of involvement in the trial. Sites can be “**Selected**” for involvement, “**Initiated**” into the trial, or actively “**Enrolling**” patients
- **Site Activation Date**: The date a site was initiated into the trial
- **State/Country**: The state and country of the specific clinical site

andromeda_ivrs_enrollment_data_2018.csv

Information about patients being recruited at clinical sites involved in ANDROMEDA, including:

- **Country**: The country the site is located in
- **Site Number**: Integer identifier of a specific clinical trial site
- **Investigator Name**: Name of the investigator overseeing trial recruitment at the site
- **Subject ID**: The internal integer identifier of a given patient being evaluated
- **Screen Date**: The date a patient was initially screened for eligibility in the trial
- **Screen Failure Date**: If present, date patient was determined to be ineligible for the trial
- **Rescreen Date**: If present, date patient was re-screened to determine eligibility for trial
- **Randomization Date**: If present, date a screened patient was randomly enrolled into a specific treatment group
- **Randomization Number**: Integer uniquely identifying patient among other randomized patients
- **Patient Type**: The current status of the patient. Can be “**Randomized**”, “**In-Screening**”, or **“Screen Failure**”

Goals

The goal of this exercise is to successfully ingest the files, perform one or more data processing tasks, and export the results in a single CSV file. It is not required to make a CLI interface unless you prefer it (i.e. you can run the code dynamically or from a REPL).

Ingest the Data Files [Required]

Successfully ingest the data from both CSV files into memory in your application.

Export a Single Data File [Required]

Export a single CSV file, with the following columns, sorted by site number in ascending order:

- **Site Number**
- **State**
- **Country**

Perform Additional Data Processing

In addition to the two required tasks above, you must perform at least **one (1)** of the following processing tasks. The result of each task can appear as a separate column in addition to the three required columns indicated above. If you have additional time remaining, please attempt additional tasks until the time has expired.

Total Patients

An integer representing the total number of patients screened or randomized for a given site.

Total Patients Randomized

An integer representing the total number of patients randomized at a given site.

Total Patients Randomized Between January 1st and July 1st (Inclusive)

An integer representing the total number of patients randomized between January 1st and July 1st, ignoring the calendar year.

Average Number of Days Between Screen Date and Randomization Date

A number indicating the average number of days elapsed between a patient’s screening date and randomization date for a given site. Non-randomized patients can be ignored.

Average Number of Patients Screened Per Month

A number indicating the average number of patients that were screened each month (from earliest date a patient was screened at a site to the latest date a patient was screened at the site).