The exercise:

A cancer clinic wants to understand how four antineoplastic (e.g., anti-cancer) drugs are being given. Drugs A and B are chemotherapy drugs (sometimes given in combination) and Drugs C and D are immunotherapy drugs. The clinic has provided us with two datasets: one gives diagnoses by patient and the other dataset gives treatment dates for these patients for the drugs of interest. None of the patients in this cohort have died to date, and no data is missing.

Please answer the following questions using the provided dataset for this exercise and in your response to each question, please include:

* Your code
* Results of your code
* Your thought process or any necessary explanation for each question (the hiring team will review all responses)

General questions

* When presented with a new dataset or database, what steps do you generally take to evaluate it prior to working with it?
* Based on the information provided above and the attached dataset, what three questions would you like to understand prior to conducting any analysis of the data?

Data analysis questions

* First, the clinic would like to know the distribution of cancer types across their patients. Please provide the clinic with this information.
* The clinic wants to know how long it takes for patients to start therapy after being diagnosed, which they consider to be helpful in understanding the quality of care for the patient. How long after being diagnosed do patients start treatment?
* Which treatment regimens [i.e., drug(s)] are typically prescribed for the initial treatment for breast cancer? What about colon cancer?

Data format

*Please see provided dataset for the actual data*

Patient\_Diagnosis

patient\_id - patient identifier; each patient has a unique patient\_id diagnosis\_date - date of diagnosis; YYYY-MM-DD format

diagnosis\_code - an ICD9CM diagnosis code (see [here HYPERLINK "http://www.cdc.gov/nchs/icd/icd9cm.htm"](http://www.cdc.gov/nchs/icd/icd9cm.htm) for more background) diagnosis - a diagnosis description used for reporting purposes

Patient\_Treatment

patient\_id

treatment\_date - date of treatment; YYYY-MM-DD format drug\_code - an internal drug identifier