The Exercise:

Submitted by Yoav Kan-Tor

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.

For 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)

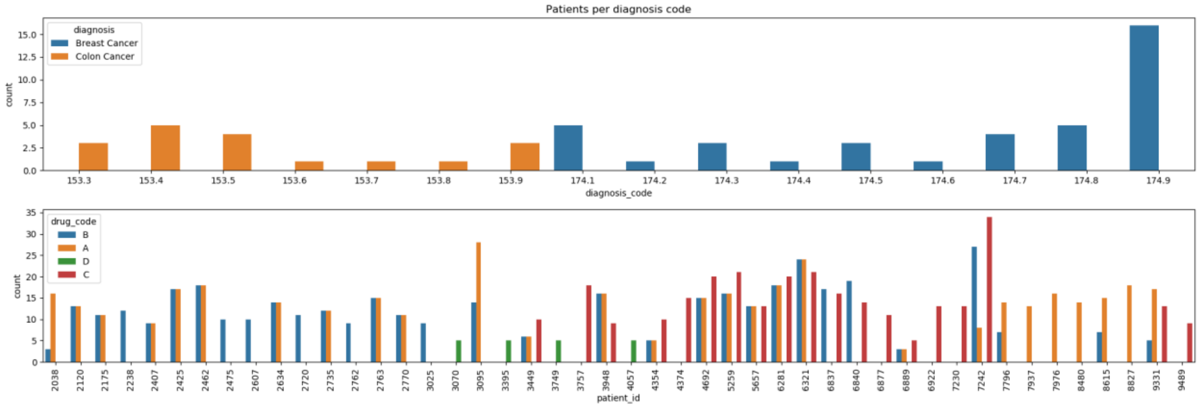
General Questions:

1. When presented with a new dataset or database, what steps do you generally take to evaluate it prior to working with it?

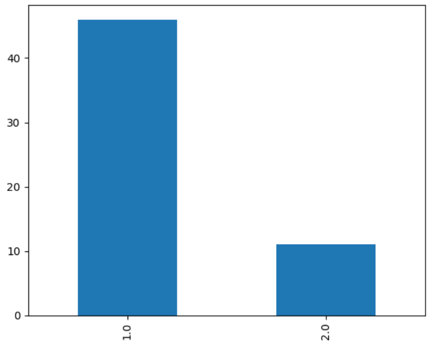
    - The first thing is to see if the data has a meaningful unique index. In this case in the we don't have. We can't use patient id (we have patients which are diagnosed with multiple cancers, and each patient is treated multiple time)
    - Search missing data points first. In this exercise I am told that there is no missing data.
    - Assign data type for each column, category, continuous (if so which value) Also I will look at the relation between the tables (if all patients have treatment and vice versa)

2. 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?

    - What is the number of patients with each treatment
    - what is the number of treatments per patient
      see the following report (fig 1 at code)



- How many patients were diagnosed with two cancers, we can see it in the following figure:



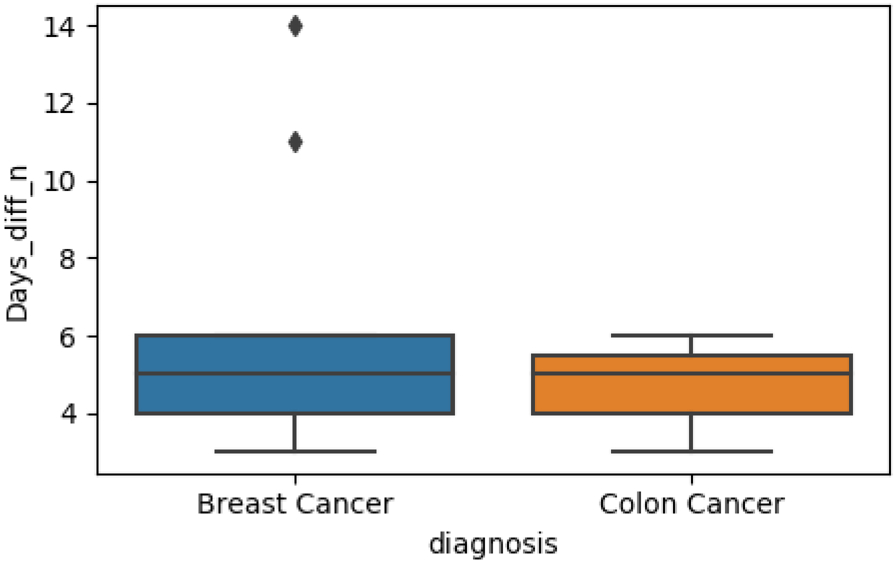
Data analysis questions:

1. First, the clinic would like to know the distribution of cancer types across their patients. Please provide the clinic with this information.

It is visible in the figure above.

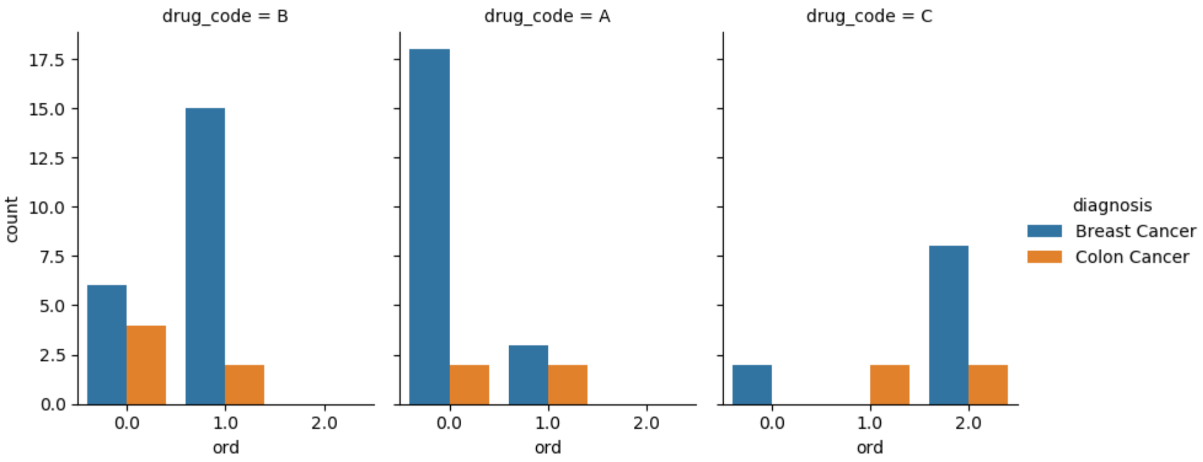
2. The clinic wants to know how long it takes for patients to start therapy after being diagnosed, which they consider to be helpful in understanding the quality of care for the patient. How long after being diagnosed do patients start treatment?

   To do this I calculated when was the first treatment, I removed outliers (negative values and above the 98th percentile) the data is displayed in the attached box plot. We can see that most patients are treated in less than 6 days and that there is not a big difference between groups.

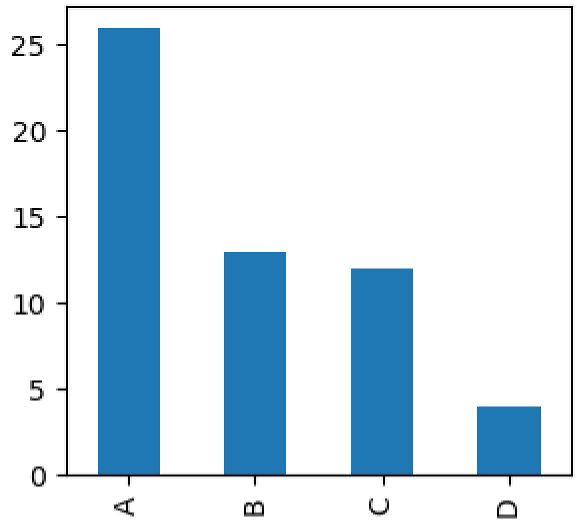


3. Which treatment regimens [i.e., drug(s)] do you think would be indicated to be used as first-line of treatment for breast cancer? What about colon cancer? (For more information between first-line and second-line treatments (applicable between chemotherapy drugs as well as chemo v immuno therapies), please reference https://www.cancer.gov/publications/dictionaries/cancer-terms?cdrid=346494)

   To answer this question, I decided to ask what the order is of giving a drug, for how many patients it was given first second etc. To differentiate between Breast cancer to colon cancer I took only patients with single diagnose. It is clear that drug A is usually the first (marked with zero) for breast cancer and drug B is usually first for colon cancer.

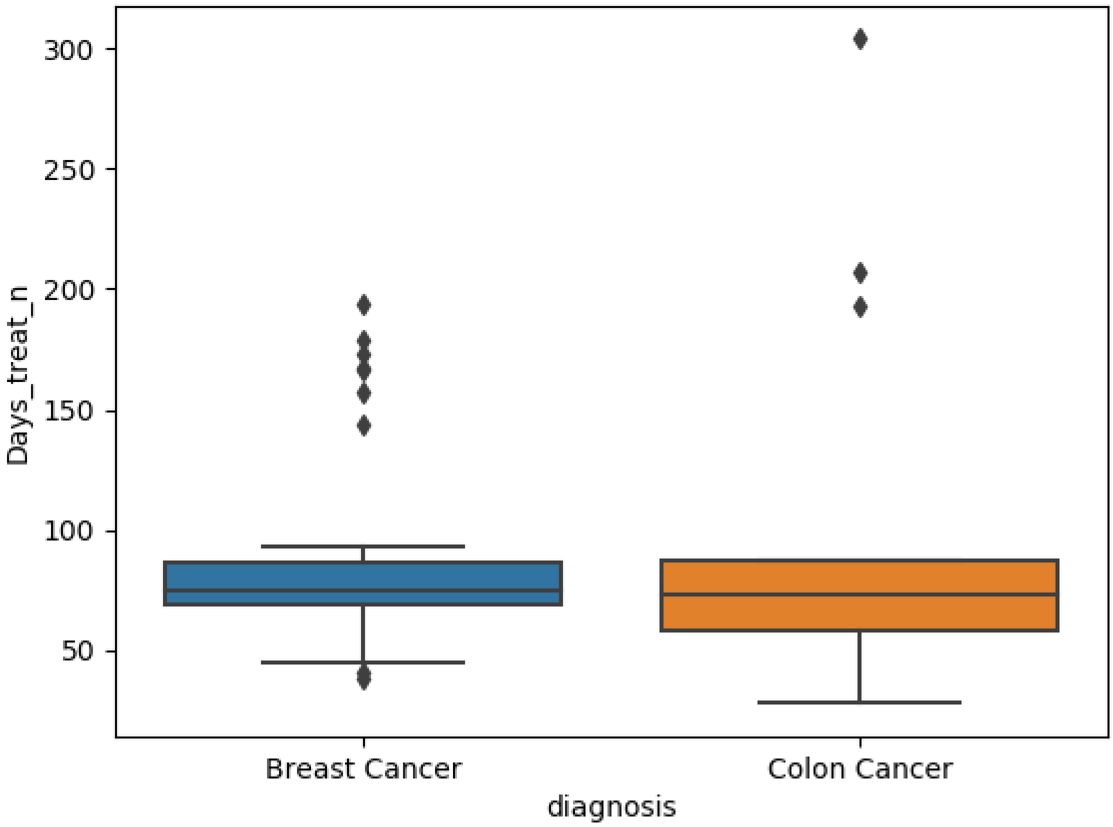


The following is a simpler plot which shows how many times each treatment was the first.



4. Do the patients taking Regimen A vs. Regimen B as first-line therapy for breast cancer vary in terms of duration of therapy? Please include statistical tests and visualizations, as appropriate.

First I would look at a box plot of the time of treatment, we can see that If we look at 95% of the data the differences are not bit between the quantiles and average. Ttest is an additional indication that there is not difference. So I will conclude that there is not significance difference.S


300
250
200
150
100
50
Days_treat_n
Breast Cancer
Colon Cancer
diagnosis