1. Spend some time exploring the data and show or discuss what you find. What are the types of data quality issues will you have to consider?

1. Determine what fields are required and what tolerance of missing rows is acceptable.
2. Search for duplicate
3. Decide actions for fields with no tolerance for missing data, such as what is required for a match key. Example: patient_id, DOB, first name, last name, gender. Options include rejecting data set ingest and alerting data staff, allowing data ingest and alerting data staff. Set checks for invalid data values. i.e. 2 in a boolean field.
4. Normalize data types in normalized tables i.e. diagnostic codes always varchar; set boolean values for male and diabetes; set integers and float data types accordingly.
5. Set lower and upper bounds for values such as SYSBP to reduce typo errors (i.e. a HEARTRATE value of 2000 would reject the import or trigger an alert). Be alert for edge cases involving outliers
6. Remove null columns. Patient_note.csv only has two fields, patient_id and note. The remaining columns are null. Is that intentional or are we missing data?
7. Verify incoming dataset has consistent headers, columns, and upload frequency
8. I don't see metadata attached to the database tables. i.e. I'm not sure what the values in patient_demographics.education refer to.

Follow tidy data rules: where each variable has its own column, each observation has its own row, and each value has its own cell.

Consider a data model to parse data into smaller, separate tables, maintaining a standard way of mapping or normalizing data:

- Parse patient_clinical into patient_clinical and diagnosis_codes tables for easier querying since many patients will have multiple diagnostic codes
- One table for individual information (patient id, gender, age, education, dob, first name, last name
- Consider data models to map separate tables for medical categories. Ie glucose_monitoring table (patient id, glucose, diabetes boolean field),

***2. Medical Device Company A* comes to us and wants to find out how many patients with diabetes are under 75, have the following diagnostic codes: 408850009, 232063007, 232053004, a total Cholesterol reading between 185 and 230, and a diastolic blood pressure reading of over 100. How many patients meet this criteria? How would you report this information back to *Medical Device Company A*?**

See Minerich_Verana.ipynb for SQL code

Dear Medical Company A,

We have completed the analysis of the patient data based on the criteria provided. The number of patients meeting the specified conditions is as follows:

Diagnostic Codes: 408850009, 232063007, 232053004
Total Cholesterol: Between 185 and 230 inclusive
Diastolic Blood Pressure: Greater than 100.0 (excluding 100.0)
Age: Less than 75 (exclude age 75)
Diabetes Status: Confirmed (true)
Total count: 436

Please let us know if you require any further details or additional analysis.

Best regards,

***3. Pharma Co. Z* has a product on the market to prevent heart attacks. They want to study patients who complain of signs of heart attacks to their doctors. Specifically, they are interested in knowing how many patients complain of pain, fluttering, pressure, or tightness in their chest have that documented in the notes of their record. Write code for and provide counts of how many unique patients match this criteria. Write a short summary to the client communicating what you did. If we wanted to recommend a more advanced text search, what would you suggest doing?**

> *See Minerich_Verana.ipynb for Python and SQL code. Final_heart_dataset.csv is included*

Dear Pharma Co. Z,

As requested, we have processed your patient note csv file to filter out specific patient notes containing keywords related to heart conditions. We read the CSV file, removed any columns that contained only null values, and defined a set of keywords related to heart conditions as follows:

'heart attacks'
'pain in chest'
'fluttering'
'pressure in chest'
'tightness in chest'
'chest pain'
'pressure in the chest'

We then filtered the dataset to include only those rows where the 'note' column contained any of these keywords. Please find the results attached, filtered_heart_dataset.csv. The count of rows in the filtered dataset is **805** patients. The results are consistent with a SQL query in the Snowflake data warehouse.

We can expand this search to include “shortness of breath” and “pain in the arm” For a more advanced text search, we recommend implementing Natural Language Processing (NLP) techniques. Specifically, we can discuss methods such as Term Frequency-Inverse Document

Frequency and Named Entity Recognition to provide a more robust and comprehensive text search capability. Let me know if you would like to schedule a call to discuss this.

Best regards,

***4. Pharma Co. Z* has a couple of follow-up questions after seeing the results of their previous inquiry.**

A. How many patients are male, have diagnoses: 232065000 or H35.52, *and* the physician noted the patient complains of pain, fluttering, pressure or tightness in their chest?

Query Result: 5 *See Minerich_Verana.ipynb for SQL code.*

B. How many patients with either of the two diagnoses might have experienced those symptoms, but cannot be confirmed using this data? Write 1-2 sentences for *Pharma Co. Z* about why this might be the case.

This might be the case because the patient notes do not explicitly mention the keywords we searched for, or the symptoms were documented in a different manner not captured by our keyword list, or symptoms are not reported to the patient's doctor.

As for why we cannot confirm the diagnoses using this data, it is possible that the symptoms mentioned in the patient_note table are not specific enough to make a definitive diagnosis. Additionally, the patient_clinical table may not contain all the necessary information to confirm the diagnoses, such as lab test results or imaging studies. Therefore, we can only use the patient_note table to find patients who have symptoms related to the two diagnoses, but cannot confirm the diagnoses using this data.

5. Assess the distribution of TenYearCHD among men and women. Run a statistical test to determine significance and interpret the results.

See Minerich_Verana.ipynb for Python code

I chose a chi-square test because we have categorical independent data, gender and presence or absence of TenYearCHD. I first created separate gender dataframes. I then determined the distribution of tenyearchd value counts per gender. I then created a matrix (shown below) and used Python scipy library to run the chi-square test and P-value. The P-value is <0.05 and we reject the null hypothesis that there is no relation between gender and tenyearchd.

matrix

tenyearchd	0	1
male		
0	2116	299
1	1474	342

Chi-Square Test Statistic: 33.06
P-Value: 8.92e-09

6. Is there a relationship between the number of cigarettes smoked per day and the prevalence of diabetes? Run a statistical test to determine significance, interpret the results and create a visualization to display this.

See Minerich_Verana.ipynb for Python code

Chi-Square test for cigs per day and diabetes: 19.89
P-Value for cigs per day and diabetes: 0.95
The P-value is above 0.05 and we accept the null hypothesis that there is no association between the number of cigarettes per day and diabetes.

I chose Plotly to visualize the data because Plotly provides more graph interaction for the user. The figure could be better presented with cig count groups (ie: 0 cigs per day; 1-10; 9-20; etc)