Data Assessment and Analysis Report

Introduction:

In this report, I will outline the process of assessing and analyzing a dataset consisting of various healthcare-related data. The tasks include data assessment to identify quality issues and data analysis to derive meaningful insights from the dataset. I will provide detailed explanations of the steps taken and the findings obtained.

Data Assessment:

1. Missing Columns and Mapping Strategies:

Upon initial examination of the datasets, I identified missing columns in each dataset and proposed mapping strategies to address these gaps. Here's a breakdown of the missing columns and corresponding mapping strategies for each dataset:

- condition.csv: Missing columns include condition_id, recorded_date, status, condition_type, and others. Mapping strategies involve creating unique identifiers, approximating

missing dates, and adding constant values for missing information.

- encounter.csv: Missing columns include claim_id and data_source. Strategies involve generating unique IDs and adding constant values.

- medication.csv: Missing columns include medication_id, source_code_type, and others. Strategies involve creating unique identifiers and extracting information from existing columns.

- symptoms.csv: Missing columns include observation_id, encounter_id, and others. Strategies involve creating unique identifiers and obtaining information from other datasets.

2. Merging Datasets and Finding Missing Column Values:

I explored the possibility of merging datasets to fill in missing information. For instance, I proposed merging condition.csv and encounter.csv based on common identifiers like PATIENT and ENCOUNTER. Additionally, I discussed strategies for finding missing column values, such as approximating dates and extracting information from related datasets.

Data Analysis:

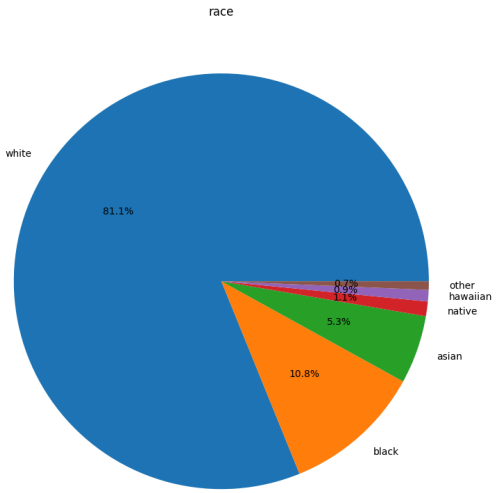
1. Distinct Patients:

To determine the number of distinct patients in the dataset, I extracted the patient IDs and calculated the count of unique values. This provides insight into the patient population represented in the dataset.

Answer: 10000

3. Racial and Gender Distribution:

I created a pie chart to visualize the percentage of patients across each racial category and gender. This provides an overview of the demographic composition of the patient population.



4. Symptom Categories:

I calculated the percentage of patients who have all four symptom categories greater than or equal to 30. This metric helps assess the prevalence of multiple symptoms among patients in the dataset.

Answer: 26.8336019172761 45% have all 4 symptoms & its not greater than 30%

Conclusion:

In conclusion, the data assessment and analysis process involved identifying data quality issues, proposing mapping strategies, and conducting various analyses to derive insights from the dataset. The findings provide valuable information for further exploration and decision-making in the healthcare domain.

DataPipeline structure