***Rahul Tamhankar***

***Position of - USF Health Data Management Analyst***

**Report: Data Analysis and Dataset Creation**

**Task 1: Create a Single Data File for Each Patient**

1. **Data Preparation:**
   * Loaded the three datasets (patient, encounter, and diagnosis) into separate pandas DataFrames which are in CSV format
2. **Data Manipulation:**
   * Merged the patient and encounter DataFrames on the "Patient\_id" column to include demographic information for all encounters.
   * Convert date columns to date format
   * Filtered the merged DataFrame to include only encounters that occurred in 2019.
   * Merge the patient and encounter DataFrames to diagonsis df.
   * Grouped the filtered DataFrame by "Patient\_id" and selected the most recent encounter for each patient in 2019.
   * Merged the resulting DataFrame with the diagnosis DataFrame on both "Patient\_id" and "Encounter\_id" columns to include condition codes.
   * Pivoted the merged DataFrame to transform the condition codes into separate columns for each condition diagnosed in the encounter.
   * Created a final DataFrame with the required columns: "Patient\_id," "Sex," "Year\_of\_birth," "Encounter\_id," "Encounter\_date," and the condition code columns.
3. **Data Export:**
   * Saved the final DataFrame as a CSV file.

**Task 2: General Information about software and population**

**Software and Libraries Used:**

* + Python: A programming language for data analysis and manipulation.
  + pandas: A Python library used for data manipulation and analysis.
  + Matplotlib: A Python library for creating visualizations, including histograms(seaborn can be used too).

1. **Population Age Histogram:**
   * Loaded the patient dataset into a pandas DataFrame.
   * Calculated the age of each patient based on the "Year\_of\_birth" column and the current year.
   * Created a histogram using Matplotlib or Seaborn to visualize the distribution of ages in the population.
2. **Five Most Common Conditions:**
   * Loaded the diagnosis dataset into a pandas DataFrame.
   * Counted the occurrence of each condition code to determine the five most common conditions.

Found the following –

Z79.899 2778

I10 1172

K21.9 1000

Z87.891 917

Z79.82 890

1. **Conclusion:**
   * The tasks were successfully completed using Python, specifically the pandas library, for data manipulation and analysis.
   * The final dataset, as requested, includes the following columns: "Patient\_id," "Sex," "Year\_of\_birth," "Encounter\_id," "Encounter\_date," and separate columns for each condition code diagnosed in the most recent encounter of each patient in 2019.
   * The dataset has been saved as a CSV file and will be emailed back to you.