DATATHON SPRINT 1 REPORT

Team 2 : Data Explorers

CG Macros : a scientific dataset for personalized nutrition and diet monitoring

Objective: Uncover meaningful relationships between dietary habits, glucose responses, physical activity, and gut health using CG Macros dataset and microbiome profiling.

Tools & Language used:

- Python (Jupyter Notebook)
- Excel
- Zoom, Slack, WhatsApp

Data Extraction:

The raw data for the project was provided in multiple CSV files distributed across various folders. We used Python libraries- **pandas** and **os** to read and load the data into memory.

- **Participant Data**: We programmatically navigated through the subfolders of each participant using the os library and read their CGM readings into individual data frames.
- **Bio and Gut Health and Microbial Data:** The bio.csv, gut_health_test.csv, and microbes.csv files were read into separate data frames.

Data Transformation:

Several transformation steps were applied using **NumPy** and **Pandas** libraries to consolidate the data, ensure consistency and prepare the dataset for analysis:

- **Merging Data frames:**
  - **Participants Data merge:** Individual participant CGM readings data frames were merged into a single unified data frame named **merged_participants_df** using the pd.merge() function.
  - **Bio, GutHealth and Microbes data merge:** The bio_df and guthealth_df were merged into one data frame named **merged_bio_guthealth_df** based on participant identifiers. Then some transformations were done using the microbe's data (which is explained in detail under the data cleaning step) and it was merged with the merged_bio_guthealth_df and named **merged_bio_gutHealth_microbes_df**.
- **Standardizing Column Names:**
  Some columns contained leading/trailing spaces, which caused inconsistencies and duplications after merging. To resolve this, all column names were standardized using the **str.strip()** method to remove unnecessary whitespace.
- **Meal Type Normalization:**
  As per the accompanying Physionet documentation, participants recorded three main meals (breakfast, lunch, dinner) and occasionally consumed snacks (water or sugar-free coffee). However, the Meal Type column included inconsistent entries such as 'Lunch', 'Dinner', 'Snacks', 'Breakfast', 'dinner', 'snack', 'lunch', 'breakfast', 'snack 1', 'Snack'.
  To standardize, all entries were converted to lowercase.
  - Variants like "snack 1", "snacks" were mapped to "snack".
  - The cleaned values were then capitalized using **.str.capitalize()** for readability, resulting in a uniform set of values: "Breakfast", "Lunch", "Dinner", and "Snack".
- **Decimal Precision Adjustment:**
  - In merged_participants_df, columns like Libre_GL and Calories (Activity) had values with up to six decimal places.
  - Similarly, the BMI column in merged_bio_gutHealth_df had excessive precision.
  - The above-mentioned columns were rounded to two decimal places using the .round(2) method to enhance readability, minimize noise, and maintain uniformity.

- **Timestamp Decomposition:**
  The merged participants dataset included a Timestamp column combining both date and time information. To enable more granular time-based analysis (e.g. trends by day or time of day), we first converted the Timestamp values into python datetime objects using **pd.to_datetime()**. Subsequently, we extracted and created two separate columns:
    - Date: containing only the calendar date
    - Time: containing only the time of day.

  This transformation facilitated easier grouping, filtering, and time-based comparisons during the analysis phase.

Data Cleaning:

- **Removal of Redundant and Inconsistent Columns:**
  After merging all 45 participants' data into **merged_particpants_df,** we identified five columns that were inconsistently recorded:
    - **Unnamed: 0**: Present in 9 participants' data, served as redundant row numbering.
    - **Steps and Record Index**: Unique to Participant 7.
    - **Intensity**: Unique to Participant 11.
    - **Sugar**: Unique to Participant 13.

  Since these columns were not consistently present across all participants and were not part of the standardized data dictionary (DataDictionary_CGmacros-00X.csv), they were dropped from the dataset.
- **Unprocessed Columns Due to Ambiguity:**
  The dataset also included two additional columns—Amount_Consumed and Image_Path — which contained a significant number of inconsistencies, such as missing image paths for the labelled meal type and errors in percentage of meal consumed readings. Due to the unclear utility of these columns in our current analysis scope and time constraints, we decided to retain them in their original form without further transformation or cleaning. These columns may be revisited in future iterations of the project if they are found to contribute meaningful insights.
- **Removal of missing rows:**
  When we merged the bio.csv and gut_health_test.csv files into a single data frame - **merged_bio_gutHealth_df**, we noticed Subject(participants) - 24 and 25 were missing in the given bio.csv file but are present in the given gut_health_test.csv. When we used the merge function with inner join, they were excluded in the merged file (intentional).
- **Consistency in Column Naming:**
  Column names were reviewed and renamed where necessary to ensure alignment with the naming conventions outlined in the data dictionary. This step ensured seamless downstream analysis.
- **Handling Missing Values:**
  Missing values in the Libre_GL and Dexcom_GL columns from the **merged_participants_df** were imputed using participant-specific averages. For each participant, we calculated the mean glucose level and used it to fill the corresponding missing values. This approach preserved the individual variability in glucose readings and improved the completeness of the dataset without introducing bias from global averages.
- **Handling erroneous values for subject 12:**
  For Subject 12, the data dictionary indicated that the LDL and VLDL values were erroneous. While recalculated values were obtained using standard clinical formulas, the resulting values still appeared abnormal. To ensure data integrity, we imputed the LDL and VLDL values for this participant using the median of the recalculated values across all participants.

- **Microbes data categorization:**
  The microbes.csv file contained 1,979 distinct microbe names, posing challenges for direct analysis due to its high dimensionality and granularity. We divided the list of microbes among our team to derive meaningful insights, and each microbe was categorized into one of three health impact groups—**Good**, **Moderate**, or **Bad**, based on established research literature and external references. The details for the classification are mentioned in the DataDefinitions Excel sheet **microbes_HealthImpactCategory**. Following this classification, we computed the count of microbes in each category for every participant. This transformation enabled dimensionality reduction while preserving biologically significant patterns, facilitating more interpretable downstream analysis. The categorized microbe summary was subsequently merged with the merged_bio_guthealth_data dataset.
- **Final Cleaned Datasets:**
  The resulting datasets saved as csv files -
  02DataExplorers_Cleaned_Merged_Participants_Data.csv,
  02DataExplorers_Cleaned_MergedBioGuthealthMicrobesData.csv contain only the standardized and relevant columns shared across all 45 participants.
  A detailed summary of these changes is presented in **Tables 1, 2 and 3.**

TABLE 1: Cleaned & Combined Participants Data: **merged_participants_data**

From	To
Subject	ParticipantID
Timestamp	Added 2 new columns Date & Time using Timestamp
Libre GL	Libre_GL
Dexom GL	Dexcom_GL
HR	Heart_Rate
Calories(Activity)	Activity_Calories_Per_Minute
MET'S	METs
Meal Type	Meal_Type
Calories	Meal_Calories
Carbs	Carbs
Protein	Protein
Fat	Fat
Fiber	Fiber

Amount Consumed	Amount_Consumed(%)
Image path	Image_path

TABLE 2: Rename Columns for better understanding: **merged_bio_guthealth_data**

From	**To**
subject	ParticipantID
Body weight	Weight
Self-identify	Ethnicity
A1c PDL (Lab)	HbA1c
Fasting GLU - PDL (Lab)	Fasting Glucose
LDL (Cal)	LDL
VLDL (Cal)	VLDL
Cho/HDL Ratio	Cho/HDL_Ratio
Collection time PDL (Lab)	Fasting _Lab_Collection _Time
#1 Contour Fingerstick GLU	Fingerstick_Glucose_1
Time (t)	Fingerstick _Glucose_1_Time
#2 Contour Fingerstick GLU	Fingerstick_Glucose_2
Time (t).1	Fingerstick_Glucose_2 _Time
#3 Contour Fingerstick GLU	Fingerstick_Glucose_3
Time (t).2	Fingerstick_Glucose_3_Time
Gut Lining Health	Gut_Lining_Health
LPS Biosynthesis Pathways	LPS_Biosynthesis_Pathways
Biofilm, Chemotaxis, and Virulence Pathways	Biofilm_Chemotaxis_and_Virulence_Pathways
TMA Production Pathways	TMA_Production_Pathways
Ammonia Production Pathways	Ammonia_Production_Pathways
Metabolic Fitness	Metabolic_Fitness
Active Microbial Diversity	Active_Microbial_Diversity

Butyrate Production Pathways	Butyrate_Production_Pathways
Flagellar Assembly Pathways	Flagellar_Assembly_Pathways
Putrescine Production Pathways	Putrescine_Production_Pathways
Uric Acid Production Pathways	Uric_Acid_Production_Pathways
Bile Acid Metabolism Pathways	Bile_Acid_Metabolism_Pathways
Inflammatory Activity	Inflammatory_Activity
Gut Microbiome Health	Gut_Microbiome_Health
Digestive Efficiency	Digestive_Efficiency
Protein Fermentation	Protein_Fermentation
Gas Production	Gas_Production
Methane Gas Production Pathways	Methane_Gas_Production_Pathways
Sulfide Gas Production Pathways	Sulfide_Gas_Production_Pathways
Oxalate Metabolism Pathways	Oxalate_Metabolism_Pathways
Salt Stress Pathways	Salt_Stress_Pathways
Microbiome-Induced Stress	Microbiome_Induced_Stress

TABLE 3: Transformations on existing/new columns:
02DataExplorers_Cleaned_MergedBioGuthealthMicrobesData.csv

Column	**Existing/ New**	**Step(s) Applied**
BMI	Existing	Rounded values to two decimal points
BMI_Classification	New	Underweight(< 18.5) , Normal(18.5 - 24.9) , Overweight (25.0 - 29.9), Obese (>=30.0)
HbA1c_Classification	New	Normal (<5.7), Prediabetes (5.7-6.4), Diabetes (>6.4)
Fasting_Glucose_Classification	New	Normal (<100), Prediabetes (100-125), Diabetes (>= 126)
Fasting Lab Collection Time	Existing	- Renamed to Fasting_Lab_Collection_Time - change format from HH:MM:SS AM/PM to HH:MM (24 hour)

LDL_Calculated	New	- Calculating with formula: LDL = Cholesterol - HDL - (Triglycerides / 5) - Rounded as Integer
VLDL_Calculated	New	- Calculating with formula: VLDL = Triglycerides / 5 - Rounded as Integer
Cho/HDL_Ratio_Calculated	New	- Calculating with formula: Cho/HDL Ratio = Cholesterol / HDL - Rounded values to 1 decimal point
Gut_Lining_Health_Desc	New	Mapped 1 = Not Optimal, 2 = Average, 3 = Good
LPS_Biosynthesis_Pathways_Desc	New	Mapped 1 = Not Optimal, 2 = Average, 3 = Good
Biofilm_Chemotaxis_and_Virulence_Pathways_Desc	New	Mapped 1 = Not Optimal, 2 = Average, 3 = Good
TMA_Production_Pathways_Desc	New	Mapped 1 = Not Optimal, 2 = Average, 3 = Good
Ammonia_Production_Pathways_Desc	New	Mapped 1 = Not Optimal, 2 = Average, 3 = Good
Metabolic_Fitness_Desc	New	Mapped 1 = Not Optimal, 2 = Average, 3 = Good
Active_Microbial_Diversity_Desc	New	Mapped 1 = Not Optimal, 2 = Average, 3 = Good
Butyrate_Production_Pathways_Desc	New	Mapped 1 = Not Optimal, 2 = Average, 3 = Good
Flagellar_Assembly_Pathways_Desc	New	Mapped 1 = Not Optimal, 2 = Average, 3 = Good
Putrescine_Production_Pathways_Desc	New	Mapped 1 = Not Optimal, 2 = Average, 3 = Good
Uric_Acid_Production_Pathways_Desc	New	Mapped 1 = Not Optimal, 2 = Average, 3 = Good
Bile_Acid_Metabolism_Pathways_Desc	New	Mapped 1 = Not Optimal, 2 = Average, 3 = Good
Inflammatory_Activity_Desc	New	Mapped 1 = Not Optimal, 2 = Average, 3 = Good
Good_Microbes_Count	New	Total number of good health impact microbes
Moderate_Microbes_Count	New	Total number of moderate health impact microbes
Bad_Microbes_Count	New	Total number of bad health impact microbes