Data Audit Report

This **Data Audit Report** provides a technical and procedural overview of the data pipeline. It evaluates the data lifecycle from raw Excel ingestion, Exploratory data analysis to the finalized SQLite storage, focusing on data quality, security, and the logic used to derive clinical insights.

Table of Contents

Exploratory Data Analysis

Health Dataset 1 Analysis:

- *Data Structure and Quality*:

Variable	Position	Variable Label	Value Labels	Measurement Level
Patient_Number	1	Patient Number	Not Applicable	Nominal
Blood_Pressure_Abnormality	2	Blood Pressure Abnormality	0 = Normal 1 = Abnormal	Nominal
Level_of_Hemoglobin	3	Level of Hemoglobin (g/dl)	Not Applicable	Ratio
Genetic_Pedigree_Coefficient	4	Genetic Pedigree Coefficient*	Not Applicable	Ratio
Age	5	Age	Not Applicable	Ratio
BMI	6	BMI	Not Applicable	Ratio
Sex	7	Sex	0 = Male 1 = Female	Nominal
Pregnancy	8	Pregnancy	0 = No 1 = Yes	Nominal
Smoking	9	Smoking	0 = No 1 = Yes	Nominal
salt_content_in_the_diet	10	Salt content in the diet (mg/per day)	Not Applicable	Ratio
alcohol_consumption_per_day	11	Alcohol consumption per day (ml/day)	Not Applicable	Ratio
Level_of_Stress	12	Level of Stress (Cortisol Secretion)	1 = Low 2 = Normal 3 = High	Ordinal
Chronic_kidney_disease	13	Chronic kidney disease	0 = No 1 = Yes	Nominal
Adrenal_and_thyroid_disorders	14	Adrenal and thyroid disorders	0 = No 1 = Yes	Nominal

**Genetic Pedigree Coefficient (GPC) of an individual for a particular disease is a continuum between 0 and 1, where:*
*GPC **closer to 0** indicates very **distant occurrence** of that disease in her/his pedigree, and*
*GPC **closer to 1** indicates very **immediate occurrence** of that disease in her/his pedigree]*

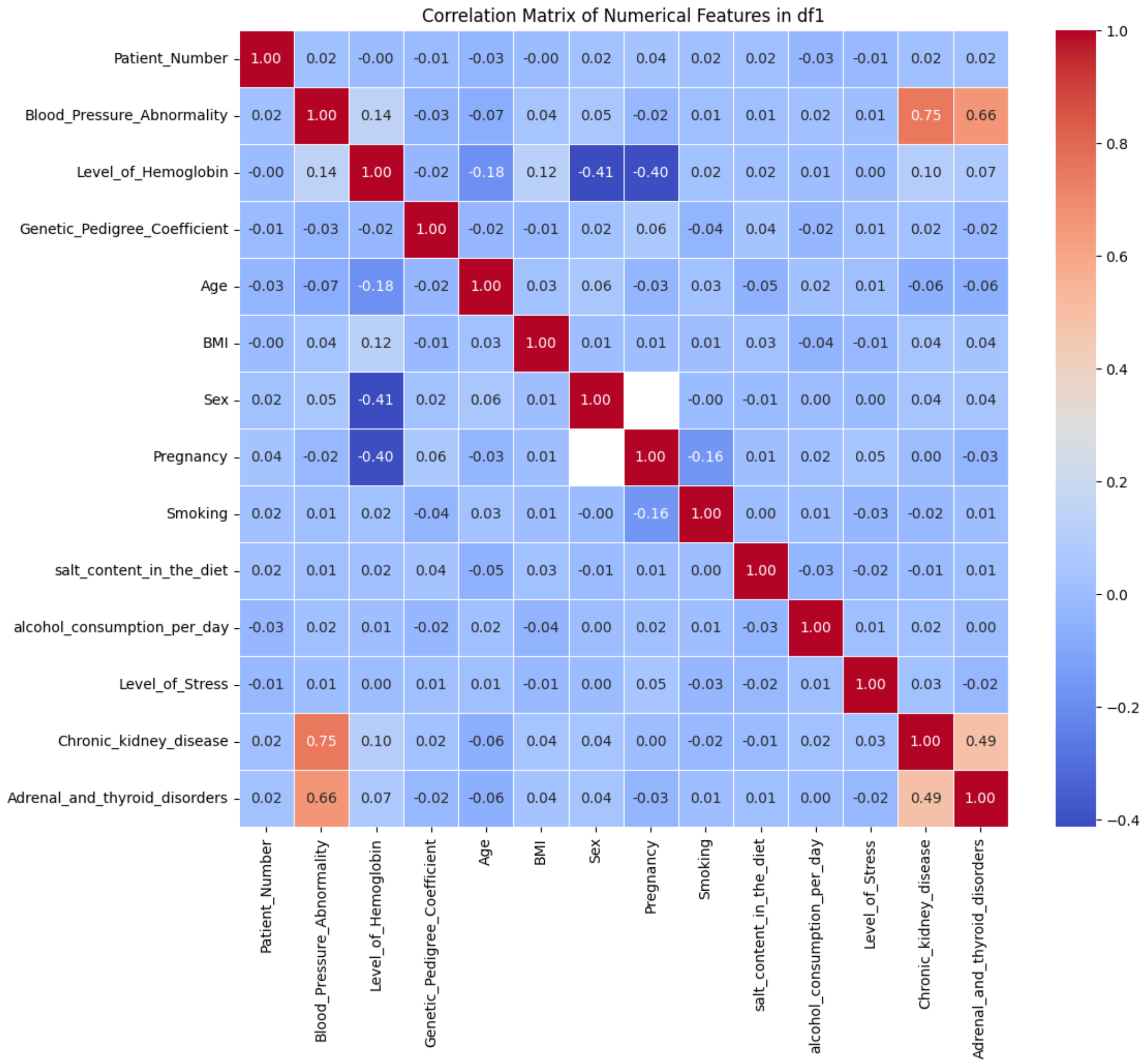
  - The dataset contains 2000 rows and 14 columns.
  - Significant missing values were identified in Pregnancy (77.9%), alcohol_consumption_per_day (12.1%), and Genetic_Pedigree_Coefficient (4.6%). Other columns are complete.

- *Descriptive Statistics and Distributions*:

  - Categorical columns such as Blood_Pressure_Abnormality (1013 '0' vs 987 '1') and Sex (1008 '0' vs 992 '1') are relatively balanced.
  - Chronic_kidney_disease (1287 '0' vs 713 '1') and Adrenal_and_thyroid_disorders (1404 '0' vs 596 '1') are skewed, indicating a higher prevalence of '0' (absence of the condition).

- Numerical columns' distributions and potential outliers were visually inspected using histograms and box plots.

- *Correlations*:
    - A correlation matrix and heatmap for numerical features were generated to identify relationships between variables. Specific correlations were not detailed in the execution result, but the analysis laid the groundwork for further investigation.



Health Dataset 2 Analysis:

- *Data Structure and Quality*:

Variable	Position	Variable Label	Value Labels	Measurement Level
Patient_Number	1	Patient Number	Not Applicable	Nominal
Day_Number	2	Day Number	Not Applicable	Nominal
Physical_activity	3	Physical activity (no. of steps/day) in the last 10 days	Not Applicable	Ratio

- The dataset contains 20000 rows and 3 columns.
- The Physical_activity column has a notable 19.205% of its values missing (3841 missing entries).
- Patient_Number and Day_Number columns are complete with no missing values.

- *Descriptive Statistics and Distributions*:
    - Patient_Number ranges from 1 to 2000, indicating multiple entries per patient across different days.
    - Day_Number shows a perfectly even distribution, with each day from 1 to 10 having exactly 2000 entries.
    - Physical_activity has a wide range (628 to 49980) with a mean of approximately 25353.5, and its distribution and potential outliers were visually inspected.

Insights

- *Address Missing Data*: The high percentage of missing values in Pregnancy (~78%) and Physical_activity (20%) suggests careful consideration of its utility or imputation strategy.
- *In-depth Relationship Analysis*: Further investigate the correlations identified in df1's heatmap to understand the strength and direction of relationships between key health indicators, potentially leading to predictive modeling or hypothesis generation.

Data Ingestion Pipeline

System Configuration & Environment

- **Database Engine:** SQLite 3.x with **Write-Ahead Logging (WAL)** enabled to support concurrent reads and optimized disk I/O.
- **Audit Metadata:** Every record is tagged with an _ingestion_time (ISO 8601 UTC) to ensure temporal traceability.

- **Logging:** A centralized simple_logger tracks row counts ($rows\_in$ vs $rows\_out$), transformation errors, and schema violations.

Ingestion & Schema Integrity

The pipeline enforces strict structural rules before data reaches the persistence layer.

- **Column Normalization:** All headers undergo a sanitization process: stripping whitespace, lowercasing, and replacing non-alphanumeric characters with underscores.
- **Validation:** Tables are rejected if the required_columns (e.g., Patient_Number) are missing.
- **Type Casting:** The system forces strict typing (Int64, Float, Datetime, or String) to prevent "dirty data" from causing downstream failures in the SQL engine.

Data De-identification (Security)

To maintain HIPAA-like privacy standards, a heuristic-based de-identification layer is applied to PII (Personally Identifiable Information).

- **Identifier Masking:** The Patient_Number is masked using a partial-reveal strategy (12****78) for 6+ digit numbers.
- **Fallback Anonymization:** For irregular identifiers, a SHA-1 hash prefixed with ANON_ is generated to maintain referential integrity without exposing raw data.
- **Selective Exposure:** The pii_columns list in the configuration explicitly targets sensitive fields for masking.

Clinical Transformation & Feature Engineering

Raw data is converted into semantically meaningful categories based on health standards.

Missing Value Handling

- **Impute as 0** for columns alcohol_consumption, physical_activity
- Nan is assigned as “data not available” for **Pregnancy column**

Categorical Re-encoding

- **Binary Flags:** Columns like Smoking, Pregnancy, and Chronic_kidney_disease are converted from 0/1 to no/yes.
- **Demographic Labels:** The Sex column is remapped from 0/1 to male/women.
- **Stress Assessment:** Numeric codes are translated into qualitative labels: 1: low, 2: normal, 3: high.

Derived Clinical Metrics

The pipeline uses the following logic for automated feature generation:

- **BMI Category:** Implements a standard clinical binning: Underweight (<18.5), Normal (18.5-25), Overweight (25-30), Obese Class I (30-35), and Obese Class II+ (>35).
- **Hemoglobin Normalcy:** A sex-aware transformation where "normal" ranges are defined differently for males ([14, 18] g/dL) and females ([12, 16] g/dL).
- **Age Bracketing:** Bins ages into seven distinct groups ranging from <18 to 70+.

Summary of Physical Activity Aggregation

The health_dataset_2_agg table provides a longitudinal view of patient engagement derived from daily activity logs.

Metric	Calculation Logic
Total Physical Activity	Sum of all recorded values, treating NaNs as 0.
Mean/Max/Min	Statistical distribution of activity, ignoring null values.
Active Days	Count of records where activity is both non-null and non-zero.
Missed Days	Count of records where activity is either null or explicitly zero.

Storage Strategy

- **Persistence:** Tables are written using the method="multi" approach for efficient batching (chunk size 500).
- **Indexing:** To optimize query performance, indexes are automatically created on Patient_Number, Age, and Sex.
- **Idempotency:** The pipeline supports upsert_keys, allowing it to update existing records rather than duplicating them if the same data is re-processed.