Healthcare Dataset - Phase 0 Profiling Report

Name: Ahmed Niazy, Ahmed Ragheb, Ibraheem Atef

ID: 202300852, 202301566, 202301058

Sec: 3,3,1

Dataset name: Healthcare Dataset

Dataset link: https://www.kaggle.com/datasets/prasad22/healthcare-dataset/data

# Dataset Description

This dataset contains 55,500 patient records from various hospitals and includes a wide range of healthcare-related fields, such as patient demographics, admission details, medical conditions, medications, and billing information. It is structured to represent realistic hospital operations with potential applications in predictive healthcare modeling, medical analytics, or cost analysis.  
Each record includes fields like Name, Age, Gender, Blood Type, Medical Condition, and Test Results, along with hospital-level information such as Doctor, Hospital, Room Number, Admission Type, and Billing Amount.

# Column Profiling Table

|  |  |  |  |
| --- | --- | --- | --- |
| Column Name | Column Profiling | Column Dependencies | Notes |
| Name | Type: Object, Unique IDs (mostly), No nulls | - | Could be anonymized for privacy. |
| Age | Min: 0, Max: 100, Mean: ~48.3, Type: Int, No nulls | Medical Condition, Blood Sugar | Distribution looks realistic; normal for hospitals. |
| Gender | Categories: Male, Female, Type: Categorical | - | Balanced gender split across records. |
| Blood Type | Categories: A+, A-, B+, B-, AB+, AB-, O+, O-, Type: Categorical | Age, Medical Condition | Useful for matching blood donors or genetic conditions. |
| Medical Condition | Categories: Cancer, Obesity, Diabetes, etc., Type: Categorical | Age, Gender, Blood Type | High frequency of chronic illnesses. |
| Date of Admission | Format: YYYY-MM-DD, Type: Object | Admission Type, Discharge Date | Should ideally be converted to datetime type. |
| Doctor | Type: Object, ~200 unique doctors | Hospital | Names indicate which doctor saw which patient. |
| Hospital | ~50 unique hospitals, Type: Object | Doctor | Helps group outcomes by institution. |
| Insurance Provider | Categories: Aetna, Medicare, Blue Cross, etc., Type: Categorical | Billing Amount | Insurance type influences cost. |
| Billing Amount | Min: ~5,000, Max: ~60,000, Mean: ~27,500, Type: Float | Age, Insurance, Admission Type | Could indicate cost of treatment or hospital type. |
| Room Number | Range: 1–500, Type: Int | Hospital | No apparent patterns, mostly random. |
| Admission Type | Categories: Emergency, Elective, Urgent, Type: Categorical | Billing, Discharge Date | Emergency tends to be most frequent. |
| Discharge Date | Format: YYYY-MM-DD, Type: Object | Date of Admission, Admission Type | Could be used to calculate length of stay. |
| Medication | Categories: Paracetamol, Ibuprofen, Aspirin, Penicillin, etc. | Medical Condition | No missing values; linked to diagnosis. |
| Test Results | Categories: Normal, Abnormal, Inconclusive, Type: Categorical | Age, Medical Condition, Genetics | Diagnostic interpretation. Used in patient evaluation. |

Profiling is done using YData Profiling code:

from ydata\_profiling import ProfileReport

import pandas as pd

# Load the dataset

df = pd.read\_csv('/content/healthcare\_dataset.csv')

# Generate profile

profile = ProfileReport(df, title="Healthcare Data Profiling Report")

profile.to\_file("ydata\_report.html")