BI Project First Draft – Group 3

Synthetic Healthcare Dataset on Kaggle:

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

Introduction

This project explores how Business Intelligence can enhance healthcare operations using UnitedHealthcare as the case company. A synthetic dataset replicating real healthcare data from Kaggle is analyzed to model insights related to patient outcomes, billing efficiency, and resource optimization.

Company Overview and BI Project Context

One of the top health insurance companies in the US, UnitedHealthcare provides coverage to people, businesses, and government programs including Medicare and Medicaid. In order to enhance healthcare results and access, the organization handles intricate patient and claims data.

Challenges and Goals

High administrative expenses, fraud detection, gaps in care coordination, and unused data for operational efficiency are some of the issues UnitedHealthcare faces. Their objectives are to improve fraud prevention, lower expenses, improve patient outcomes, and use analytics to support value-based care.

How Business Intelligence Helps

BI tools make it possible to analyze patient demographics, claims, and service usage in order to spot fraud and inefficiencies, build dashboards for in-the-moment decision-making, and promote better patient care and resource allocation.

Software Tools Used

- Excel for data cleaning
- SQL for data querying
- Python for advanced analytics
- Power BI and Tableau for visualization and dashboards

Data Modeling: Star Schema vs. Snowflake Schema

Star Schema

A star schema uses a central fact table linked directly to dimension tables. It is simple and optimized for fast querying.

Fact Table	Description
Claims Fact	Stores individual claim details, including patient ID, provider ID, service date, diagnosis and procedure codes, billed amount, and claim status.

Dimension Tables	Description
Patient Dimension	Patient demographics: age, gender, location, and insurance type.
Provider Dimension	Information on doctors, specialists, and medical facilities.
Time Dimension	Date information allowing analysis by day, month, quarter, and year.
Service Dimension	Medical procedures, diagnosis codes, and treatment details.
Payer Dimension	Insurance plan types and payer information.

Snowflake Schema

A snowflake schema normalizes dimension tables into related sub-tables to reduce redundancy and improve data integrity but complicates queries.

Fact Table	**Same as in the star schema.**

Dimension Tables	**Normalized Structure Example**
Patient Dimension	Split into Patient (ID, Name, InsuranceID) and Insurance (InsuranceID, Type, Coverage).
Provider Dimension	Divided into Provider (ID, Name, SpecialtyID) and Specialty (SpecialtyID, Description).
Time Dimension	Can be kept denormalized or split further if needed.
Service Dimension	Divided into Service (ServiceID, Name, CategoryID) and Category (CategoryID, Description).
Payer Dimension	Can also be normalized if necessary.

Summary Table: Star vs. Snowflake

Schema Type	Advantages	Disadvantages
Star Schema	Simpler design, faster query speed	More data redundancy possible
Snowflake Schema	Saves storage, enforces normalization	Increased complexity, slower queries

Recommendation

Because of its ease of use and effectiveness for reporting and dashboarding, the star schema is suggested for the UnitedHealthcare BI project that analyzes insurance and healthcare claims data. The snowflake schema might be taken into consideration, nevertheless, if reducing redundancy and maintaining data consistency are the top priorities.