**Healthcare Data Analysis Project Report**

**Summary of the Project**

In this project, we took on the role of a data analytics specialist working with real-world healthcare data. The dataset included anonymized patient records sourced from a public Kaggle dataset, with various attributes such as name, age, gender, medical conditions, doctors, hospitals, insurance providers, billing amounts, and treatment outcomes.

The primary objective of the project was to clean, prepare, and structure this healthcare dataset using SQL so that it could be used for downstream analysis and visualization in Power BI. The goal was to uncover meaningful insights regarding patient care, medical costs, and treatment outcomes while ensuring the data was export-ready and analysis-friendly.

**Methodology Used**

Using **SQL data preparation and insight generation**, we followed a structured approach:

**1. Data Cleaning**

* Converted all names to lowercase for consistency.
* Removed honorifics like “Mr.”, “Mrs.”, “Dr.”, “Ms.”, and “Miss” from patient names using SQL REPLACE() functions.
* Verified the dataset for **nulls and blank values** across all major columns, including:
  + Name
  + Age
  + Gender
  + Medical Condition
  + Doctor
  + Insurance Provider
  + Billing Amount
  + Admission Type
  + Test Results
* Validated entries in categorical fields (e.g., ensuring all test results were either *Normal*, *Abnormal*, or *Inconclusive*).

**2. Key Business Questions Answered via SQL:** To drive business value and clinical insight, we developed several targeted SQL queries that answered the following:

**SQL Questions Answered:**

**Q1: What are the most common medical conditions among patients?**

* Query counted patients grouped by medical condition, ordered by frequency.

**Q2: What is the average billing amount per type of hospital admission?**

* Calculated average billing grouped by "Admission Type" (e.g., Urgent, Emergency, Elective).

**Q3: Which doctors have treated the most patients?**

* Ranked doctors by total number of patients they treated, showing the top 10.

**Q4: Which conditions most frequently result in abnormal test results?**

* Filtered for Test\_Results = 'Abnormal', then grouped by Medical\_Condition.

**Q5: What is the total billing amount handled by each insurance provider?**

* Grouped billing data by Insurance\_Provider and summed the billing amounts.

**3. Data Modeling for Power BI**

To prep the data for use in Power BI dashboards, we created tables to enable easy, flexible reporting:

* **Dimension Tables**:
  + Doctors: Unique combinations of doctors and hospitals.
  + Conditions: Distinct medical conditions, medications, and test results.
  + Insurance: All insurance providers.
  + Dates: A simplified date dimension from the admission dates (useful for time-based charts).

|  |  |
| --- | --- |
| **Predicted Target** | **Achieved Target** |
| * Clean and structure the raw healthcare dataset using SQL. | * Successfully cleaned and transformed the data, ensuring consistency and removing irrelevant text (prefixes). |
| * Answer 5+ insightful questions using SQL analysis. | * Generated insights using SQL across patient demographics, cost, outcomes, and providers. |
| * Create normalized tables ready for Power BI modeling. | * Built export-ready, normalized SQL tables that align perfectly with Power BI’s visual and relational modeling capabilities. |

4. Visualization in Power BI

* Imported .csv files from SQL Workbench
* Established relationships between the tables if there were commonalities
* Created slicer that allows you to filter down all graphs by Medical condition and can be further filtered by Age to give granular data visualizations
* Created a drill through that shows which hospital have the most of patients by Condition, this can be done using the Average of Billing Amount by Medical condition and Admission type

**Conclusion**

This project highlighted how SQL serves as the foundation for powerful, real-world healthcare analytics. By cleaning and organizing the data efficiently, we laid the groundwork for deeper insights and visualization in Power BI. The queries developed not only answered practical clinical and operational questions but also positioned the data to support more advanced visualizations and dashboards going forward.

This exercise gave us hands-on experience with data cleaning, SQL-based transformation, schema modeling, and analytical query design — all essential skills for any data professional working in healthcare or beyond.