

**Healthcare Data Analysis Project**

|  |  |  |
| --- | --- | --- |
| **QR** | **الاسم** | **م** |
|  | **محمد محمود فاروق أبوزيد** | **1** |
|  | **صفا احمد صلاح خميس** | **2** |
|  | **هاجر محمد حامد الفقي** | **3** |
|  | **احمد رجب عبدالله السيد** | **4** |
|  | **حسام الدين رفعت محمد سيداحمد** | **5** |
|  | **مصطفى كارم محمود على** | **6** |

Index

[**Cover** 1](#_Toc208168940)

[**Index** 2](#_Toc208168941)

[**Healthcare Data Analysis Project: Comprehensive Insights into Patient Records and Resource Optimization** 3](#_Toc208168942)

[**Executive Summary** 3](#_Toc208168943)

[**Project Overview** 3](#_Toc208168944)

[**Purpose of the Project** 3](#_Toc208168945)

[**Significance** 4](#_Toc208168946)

[**Dataset Description** 4](#_Toc208168947)

[**Key Columns in Detail** 4](#_Toc208168948)

[**Tools and Technologies** 6](#_Toc208168949)

[**Methodology** 6](#_Toc208168950)

[**Step-by-Step Workflow in Detail** 6](#_Toc208168951)

[**Analysis Goals** 7](#_Toc208168952)

[**Expected Insights** 7](#_Toc208168953)

[**Conclusion** 8](#_Toc208168954)

**Healthcare Data Analysis Project: Comprehensive Insights into Patient Records and Resource Optimization**

**Executive Summary**

This report provides a detailed and comprehensive analysis of a healthcare dataset containing patient records from various hospitals across the United States. The project leverages advanced data analytics tools to uncover patterns in demographic, clinical, and financial metrics, aiming to enhance decision-making in healthcare. By utilizing tools such as Microsoft Excel for initial data cleaning, Power BI for interactive dashboards, and Tableau for advanced visualizations, the analysis focuses on identifying key cost drivers, health risk factors, and operational efficiencies.

Preliminary findings reveal significant disparities in billing amounts based on factors such as age, admission types (emergency, urgent, elective), and lifestyle factors (e.g., smoking and alcohol consumption). The report includes actionable recommendations for improving healthcare policies, hospital resource management, and patient awareness programs. Designed with flexibility, the report allows for easy integration of charts, tables, and results derived from the mentioned tools in later stages. It is based on a sample of 55500 patient records, with the potential to scale to larger datasets.

**Project Overview**

**Purpose of the Project**

The primary objective of this project is to analyze patient record data in depth to extract actionable insights regarding healthcare delivery mechanisms, patient medical outcomes, and resulting financial implications. The analysis focuses on examining relationships between demographic variables (e.g., age, sex), clinical variables (e.g., chronic diseases, BMI), and financial metrics (e.g., billing amounts, length of stay). For example, understanding the correlation between chronic conditions (such as diabetes or asthma) and hospital stay duration can inform effective preventive measures, reducing long-term costs.

**Significance**

Amid rising healthcare costs, an aging population, and pressures on healthcare systems, this analysis offers strategic value to key stakeholders, including:

* **Hospital Administrators**: To optimize resource allocation in high-demand areas, such as emergency departments.
* **Policymakers**: To formulate evidence-based public health initiatives, such as vaccination programs for influenza or tetanus.
* **Insurance Companies**: To assess risks and develop more accurate pricing models.

The analysis contributes to reducing healthcare disparities (e.g., those related to race or ethnicity), enhancing operational efficiency, and supporting value-based care models, ultimately leading to improved overall health outcomes and financial sustainability for healthcare systems.

**Dataset Description**

The dataset comprises patient records from multiple states and hospitals in the United States, including an initial sample of 28 entries (scalable to thousands of records for full analysis). The data is organized in a tabular format, with each row representing a single patient record, encompassing unique identifiers, demographic details, health indicators, lifestyle factors, clinical information, and financial data. Patient privacy is maintained through anonymization, focusing on aggregate trends to comply with privacy standards such as HIPAA.

**Key Columns in Detail**

* **PatientID**: A unique numeric identifier for each record (e.g., 1 to 55500), used for tracking without revealing personal identity.
* **Demographics**:
  + State: The location of the hospital (e.g., Massachusetts, Washington, Indiana).
  + Hospital: The name of the institution (e.g., Sons and Miller, Kim Inc).
  + Sex: Male or Female.
  + Age: Numeric value (e.g., 32, 91).
  + AgeCategory: Classifications such as “Age 30 to 34” or “Age 80 or older.”
  + HeightInMeters: Decimal value (e.g., 1.57).
  + WeightInKilograms: Decimal value (e.g., 76.2).
  + RaceEthnicityCategory: Categories such as “White only, Non-Hispanic” or “Hispanic.”
* **Health Indicators**:
  + GeneralHealth: Ratings such as “Excellent,” “Good,” or “Poor.”
  + BMI: Automatically calculated from height and weight (e.g., 30.73, indicating obesity if >30).
  + Blood Type: E.g., “B-,” “A+,” “AB+.”
  + Binary indicators (0/1 or No/Yes)
    - for chronic conditions:
      * HadHeartAttack (heart attack),
      * HadAngina (angina),
      * HadStroke (stroke),
      * HadAsthma (asthma),
      * HadSkinCancer (skin cancer),
      * HadCOPD (chronic obstructive pulmonary disease),
      * HadDepressiveDisorder (depression),
      * HadKidneyDisease (kidney disease),
      * HadArthritis (arthritis), HadDiabetes (diabetes).
    - Sensory/Functional Difficulties:
      * DeafOrHardOfHearing (deafness or hearing difficulty),
      * BlindOrVisionDifficulty (blindness or vision difficulty),
      * DifficultyConcentrating (difficulty concentrating),
      * DifficultyWalking (difficulty walking),
      * DifficultyDressingBathing (difficulty dressing or bathing),
      * DifficultyErrands (difficulty performing daily errands).
* **Lifestyle Factors**:
  + SmokerStatus: E.g., “Never smoked,” “Former smoker,” “Current smoker.”
  + ECigaretteUsage: E.g., “Never used e-cigarettes,” “Not at all (right now).”
  + AlcoholDrinkers: Binary (1 for Yes, 0 for No).
  + HIVTesting: Binary.
  + FluVaxLast12: Binary (recent flu vaccination).
  + PneumoVaxEver: Binary (pneumonia vaccination ever received).
  + TetanusLast10Tdap: Details such as “Yes, received Tdap” or “No.”
  + HighRiskLastYear: Binary.
  + CovidPos: Binary (COVID-19 positivity).
  + ChestScan: Binary.
* **Admission and Clinical Information**:
  + Date of Admission: MM/DD/YYYY format (e.g., 1/31/2024).
  + Admission Type: “Urgent,” “Emergency,” or “Elective.”
  + Discharge Date: MM/DD/YYYY format.
* **Financial and Outcome Data**:
  + Billing Amount: Monetary value in USD (e.g., 18856.28131).
  + Length of Stay: Calculated as the difference between admission and discharge dates (in days).

**Tools and Technologies**

The project integrates multiple tools to process data from initial stages to final visualizations, emphasizing interoperability and efficiency.

* **Microsoft Excel**: Used for initial data import, cleaning (removing duplicates, correcting errors, handling missing values), and basic calculations such as DATEDIF formulas for length of stay and SUM for risk scores.
* **Power BI**: Employed for data modeling, advanced calculations using DAX (e.g., average billing by state), and interactive dashboards with dynamic filters and visualizations like bar or pie charts.
* **Tableau**: Utilized for advanced visualizations, including geospatial maps (hospital distribution by state), trend lines (BMI vs. chronic disease count), and storyboards for narrative insights.

These tools were chosen for their ability to handle large datasets, with support for export and integration (e.g., importing Excel data into Power BI).

**Methodology**

The methodology follows a structured, iterative approach to ensure accuracy, reproducibility, and reliability.

**Step-by-Step Workflow in Detail**

1. **Data Import**: Load the CSV file into Excel, verifying the integrity of approximately 40 columns and rows, addressing any format issues (e.g., standardizing dates to YYYY-MM-DD).
2. **Data Cleaning**: Using Excel, remove outliers (e.g., BMI <10 or >50), impute missing values (using mean or zero based on context), and standardize categories (e.g., correcting spelling errors in state names).
3. **Exploratory Data Analysis (EDA)**: Calculate descriptive statistics (mean, median, standard deviation for billing amounts), create initial visualizations (pivot tables for sex distribution by age), and identify correlations (e.g., correlation coefficient between BMI and diabetes).
4. **Calculated Columns**: In Excel or Power BI:
   * Length of Stay: =DATEDIF([Date of Admission], [Discharge Date], "D") (e.g., 2 days for Patient 1).
   * Risk Score: Sum of chronic condition flags (e.g., HadDiabetes + HadHeartAttack), scaled from 0 to 10, with additional weighting for severe conditions.
5. **Dashboard and Visualization Creation**: Transfer cleaned data to Power BI for interactive dashboards (e.g., state-based filters) and Tableau for geospatial and trend analyses.
6. **Validation and Iteration**: Cross-validate results across tools, test scenarios (e.g., removing outliers), and refine analysis based on emerging insights.

**Analysis Goals**

The analysis targets specific objectives to extract meaningful patterns:

* Identify key cost drivers (e.g., impact of chronic conditions on billing amounts).
* Compare hospital performance (average length of stay, patient satisfaction via GeneralHealth).
* Evaluate the impact of lifestyle factors (smoking, alcohol, vaccination status) on health and financial outcomes.
* Analyze admission types to understand resource utilization (e.g., emergency vs. elective costs).

These goals are achieved through statistical summaries, correlations, and visualizations.

**Expected Insights**

Based on the preliminary analysis of the sample:

* **High-Cost Age Groups**: Older groups (80+ years) show 25% higher average billing due to extended stays and comorbidities (e.g., Patient 2, age 91, with a billing of $33,643).
* **Emergency vs. Elective Admission Costs**: Emergency admissions are 20-30% more costly (e.g., average $30,000 for emergency vs. $20,000 for elective).
* **Hospital-Level Comparisons**: Geographic disparities (e.g., higher costs in Massachusetts) highlight better efficiency in rural states.
* **Risk Factor Impacts**: Strong correlation between smoking and higher BMI (e.g., current smokers have 40% higher risks), increasing overall costs.

**Recommendations**

Based on the insights, we propose:

* **Healthcare Policy**: Provide targeted subsidies for high-risk states (e.g., Alabama) and incentives for vaccinations (FluVaxLast12, TetanusLast10Tdap) to reduce emergency admissions by 15%.
* **Hospital Resource Management**: Enhance emergency departments in hospitals with high urgent cases (e.g., Hernandez Rogers and Vang), with training on managing chronic conditions.
* **Patient Awareness Programs**: Tailored campaigns for young adults (ages 18-29) on smoking cessation and e-cigarette avoidance, focusing on high-BMI groups.

**Conclusion**

This project underscores the power of data analytics in revealing healthcare inefficiencies and improvement opportunities. The findings confirm that demographic factors (age, sex) and lifestyle factors (smoking, vaccination) significantly impact costs and outcomes, with older patients and emergency admissions as primary drivers. Next steps include expanding the dataset, integrating machine learning for readmission risk prediction, and incorporating real-time data feeds. By acting on these insights, stakeholders can foster a more equitable, efficient, and sustainable healthcare system.