Healthcare Data

Analytics Report & Findings

by Asher J. Frank

June 15, 2025

CAPSTONE PROJECT

PREPARED FOR

QUICKSTART LABS • UC SANTA BARBARA
PROFESSIONAL & CONTINUING EDUCATION

DATA ANALYTICS AND VISUALIZATION BOOTCAMP

help@extension.ucsb.edu
https://quickstart.professional.ucsb.edu/
1-805-893-4200
Santa Barbara, CA 93106

PREPARED BY

ASHER J. FRANK
DATA & THINGS, INC.

June 15, 2025

asher.data1@gmail.com
https://www.linkedin.com/in/asher-frank-1a1b16356/
1-213-671-3650
Reseda, CA 91335

OVERVIEW

OBJECTIVE

Demonstrate end-to-end data analytics skills by performing data cleaning and transformation, applying calculations and aggregations, developing insightful visualizations, and designing an interactive dashboard.

Analyze a complex healthcare dataset to uncover patterns and correlations using modern analytical techniques and tools such as Semantic Models, SQL, Microsoft Fabric, Dataflows, and Power BI.

THE DATA

All data used in this project is sourced from the publicly available *Healthcare Dataset with Multi-Category Classification Problem*, available on Kaggle:

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

METHODOLOGY

EXTRACT

The healthcare dataset was downloaded as a .csv file from the provided public domain source. The file contained 55,500 rows of patient records, with attributes such as Name, Age, Gender, Blood Type, Medical Condition, Admission and Discharge Dates, Doctor, Hospital, Insurance Provider, Billing Amount, Room Number, Admission Type, Medications, and Test Results.

Microsoft Fabric was chosen for its integrated, end-to-end data pipeline capabilities, offering scalable and secure support for data ingestion, transformation, modeling, and visualization. A **Lakehouse** was created to store the flat file, and the dataset was loaded as a new table.

During validation, the upload failed due to column names containing white spaces. This was resolved by replacing spaces with underscores (e.g., `Blood_Type`) in Microsoft Excel. The updated file was then reuploaded and successfully imported into the Lakehouse.

Next, a **Dataflow Gen2** was created by connecting to the Lakehouse and selecting the imported dataset as the source.

TRANSFORM

With the Dataflow established, the built-in **Power Query** editor in Fabric's Dataflow framework was used to clean, structure, and model the dataset into a star schema. Initial data cleaning steps included formatting names to proper case, validating that no null or anomalous values were present, assigning correct data types to each column, and ensuring date fields were properly formatted.

To normalize the data, the main table was duplicated and segmented into several dimension tables by grouping related categorical fields. For instance, `Name`, `Age`, `Gender`, and `Blood_Type` were isolated to create a `dim_patient` table. Similar steps were repeated to create dimensions for doctors, medical conditions, admission types, medications, test results, and dates.

The values of each dimension table were de-duplicated and assigned a unique index column to serve as the primary key. These keys replaced the original descriptive attributes in the central fact table—now named `fact_healthcare`—which retained quantitative and transactional data, such as billing amounts.

A special approach was taken for date fields. Since both admission and discharge dates were present, each was extracted as a helper query, appended into a single column, de-duplicated, and used to build a `dim_date` table. However, to better support and simplify the underlying

TRANSFORM CONT'D

code required for date-specific calculations (e.g., length of stay), both original date columns were retained in the fact table alongside their corresponding date IDs.

To verify that normalization preserved data integrity, a helper query was written using **M Code** to count the rows in the fact table. This served as a check against accidental data loss or duplication. This validation flagged an issue: joining the `dim_patient` table had nearly doubled the row count. Investigation revealed that patients with multiple admissions (e.g., in different years and with different ages) caused unintended duplications. To resolve this, `Age` was moved back into the fact table to treat each hospital visit as a distinct event while preserving a single unique ID number per individual patient name in the dimension table.

Lastly, calculated columns were added to support analysis and improve visualization. These included month and quarter extraction from full dates, hospital stay duration, and age group classifications (e.g., 0–18, 19–35). A numerical sorting key was also assigned to age groups to ensure consistent ordering in visuals.

LOAD

Once the fact table was finalized and dimension tables were properly constructed, the Dataflow was published to the Fabric **Workspace**, which populated the completed tables into the Lakehouse. A new **Semantic Model** was then created, with relationships defined between the fact and dimension tables using one-to-many cardinality and single-direction cross-filtering—forming a classic star schema structure.

To protect sensitive information not intended for end-user visibility (such as patient and doctor names), **Row-Level Security (RLS)** was configured in the model view to hide these identifiers. However, since **Object-Level Security (OLS)**—which would allow entire columns to be hidden—is not yet natively supported in Microsoft Fabric, a workaround was implemented. Sensitive fields were set to display values only when the column equaled `"FALSE()"`, a condition that would never be met, effectively masking the data without needing to switch platforms. Although user groups were not available at the time of RLS creation, roles were configured and are ready for assignment when needed.

With these measures complete, the semantic model was ready for use in Power BI and other downstream tools.

VISUALIZATION

POWER BI

To maximize flexibility and development control, **Power BI Desktop** was used to design the dashboard, though the process could have been completed entirely within Microsoft Fabric for full platform continuity. The dataset was brought into Power BI using the built-in Microsoft Fabric connector, establishing a **live connection** to the published Semantic Model.

The dashboard was structured across five pages, each dedicated to a primary analytical theme: **Patients**, **Doctors**, **Hospitals**, **Conditions**, and **Treatments**. This layout allowed for focused exploration of each domain and highlighted the unique patterns within the dataset.

For a cohesive and user-friendly experience, each page followed a consistent layout:

- **Left panel**: four key metrics displayed in card visuals (e.g., totals, averages, maxima).
- **Main body**: one large chart focused primarily on trends over time (e.g., admissions by month or quarter).
- **Upper right**: three smaller supporting charts.
- **Top controls**: a synced **dropdown slicer** for filtering by year, quarter, or month.
- **Left margin**: vertical page navigator with custom isometric icons for visual appeal.

Visualization types were selected for both readability and analytical depth:

- **Area charts** were used in the main panel to illustrate trends and comparisons over time.
- **Pie and donut charts** summarized categorical proportions in a compact, visual format.
- **Stacked bar/column charts** enabled category comparisons across multiple values.
- **Treemaps** handled large categorical distributions more effectively than pie charts.

Each visual was made fully **interactive**, allowing users to filter, drill down, and compare across multiple dimensions. Slicer syncing and cross-filtering behavior were carefully configured to support dynamic, ad hoc exploration.

Custom **DAX measures** were written as needed for precise calculations, such as rankings, conditional totals, and filtered aggregations. **Visual-**

POWER BI CONT'D

level filters were also applied to tailor outputs (e.g., Top N charts based on specific rules).

Lastly, the completed Power BI Desktop report was published back to the Microsoft Fabric Workspace, making it accessible for collaboration and distribution. The final result is a highly interactive, visually coherent dashboard that facilitates deep insight into the dataset's medical, financial, and operational dimensions.

Leveraging the dashboard's interactivity and visual design enabled the discovery of several important trends and relationships, outlined below.

DATA-DRIVEN INSIGHTS

PATIENTS

- The average patient age was **51.5 years**, with the **0–18 age group** representing the **smallest demographic**.
- **AB+** was the most common blood type, though **blood types were otherwise evenly distributed.**
- **Gender distribution** was nearly **equal** between male and female patients.
- Patients stayed an average of **15.5 days** per hospital visit.
- **February** consistently had the **lowest number of patient admissions**.

DOCTORS

- The dataset included **50,000 unique doctor IDs**.
- There were **more doctors than individual patients** across all records.
- On average, there was **approximately one doctor per hospital**.
- Doctors were **evenly distributed** across **admission types**, **medical conditions**, and **insurance providers**.
- **Doctor counts closely mirrored patient admissions** over time, suggesting a proportional staffing pattern.

HOSPITALS

- Nearly **40,000 distinct hospitals** were represented in the dataset.
- **Insurance providers** and **admission types** were **evenly distributed** across hospitals.
- Each hospital billed an **average of $35,550** across all records.
- The **highest cumulative billing** for a single hospital exceeded **$363,000**.
- The hospital with the **most admissions** did **not** correspond to the one with the **highest total billing**.

CONDITIONS

- **Arthritis** was the **most frequently treated** condition.
- **Obesity** generated the **highest total billing** among all conditions.
- **Asthma** was associated with the **longest average hospital stay**.
- **Medical conditions** were **evenly distributed** across **admission types**.
- **Patient counts per condition** were **uniformly distributed**, varying by less than **one percentage point**.

TREATMENTS

- **Lipitor** was the **most frequently prescribed** medication among the **five tracked**.
- **Medication prescriptions** were **evenly distributed** across conditions, with less than **one percentage point variation**.
- **"Abnormal"** was the **most common** test result.
- **Elective admissions** were the most frequent, though **admission types were generally balanced** across records.
- **Average billing** did **not consistently align** with **length of stay** over time.

TAKEAWAYS

Demographic and clinical patterns revealed interesting disparities, with younger age groups (0-18) being notably underrepresented.

Operational metrics revealed unusual 1:1 relationships across entities, such as a near-equal number of doctors to hospitals and doctors to patients.

Cost and care duration do not follow a predictable correlation, showing a potential inconsistent alignment between billing models, treatment types, or data structure.

Synthetic data can enable structural analysis but may lack real-world nuance, with relationships highlighting the limitations of mock-data when looking for real-world implications.

CONCLUSION

PROJECT SUMMARY

This project demonstrated a complete end-to-end data analytics workflow, leveraging Microsoft Fabric as a unified platform for data ingestion, transformation, modeling, and visualization, from extracting and transforming raw healthcare records to designing a fully interactive Power BI dashboard.

By implementing a star schema, applying data cleaning and normalization techniques, and using both Power Query M code and DAX for calculated logic, the data was successfully structured into a semantic model optimized for analysis and insight generation.

The resulting dashboard provides meaningful visualizations across key healthcare categories—patients, doctors, hospitals, conditions, and treatments—uncovering patterns in demographics, billing practices, and care delivery.

While the dataset's synthetic nature introduced some unrealistic correlations, it served as a valuable tool for developing technical proficiency and showcasing the capabilities of modern data platforms like Microsoft Fabric and Power BI.

Ultimately, this exercise highlights not only the importance of strong data modeling and design but also the need for critical thinking when interpreting findings, especially when working with simulated or unfamiliar datasets.