Report

Laercio Lima

Data Analyst

Phone: (083) 868 6694 || **Email:** larry.laerciolima@gmail.com

LinkedIn: https://www.linkedin.com/in/laercio-lima-larry/

My Portfolio: https://laerciolima1.github.io/myportfolio/

Dublin 8 – Ireland

Apr 2023

1. Dataset Selection

The data utilized for this project was obtained from https://data.cms.gov/, which offers relevant healthcare datasets that were carefully chosen to meet the project's objectives.

2. Business Understanding

In this hypothetical scenario, a company possesses data on hospitals, doctors, patients, and their respective locations. The objective is to gain insights into the distribution of the data across various regions in the United States.

2.1 Business Questions

1. Can you provide the total number of hospitals, doctors, patients, and specializations?
2. Which regions of the US have a higher concentration of doctors?
3. Is there a relationship between doctors' rating and experience, and their availability?
4. What is the most common time for patients to visit a hospital?

These questions are focused on understanding the availability of doctors in different US regions. Answering these questions can help the company to identify the areas where more doctors are needed and also help hospitals to better understand the needs of their patients. Additionally, knowing the preferences of patients can help hospitals to better allocate their resources and staff.

3. Data Understanding

In the data understanding stage, I utilized PostgreSQL to create tables and import data. Afterwards, I applied SQL queries to obtain a preliminary understanding of the datasets, including identifying duplicates, missing values, and the overall size of the data.

4. Feature Engineering

For feature engineering, I utilized the CASE function in PostgreSQL to generate a new feature named 'Region'. This feature was created by extracting data from the 'state_code' column of the 'locations' table.

```sql
-- creating a new feature in 'locations' to show the Regions based on the state_code:
-- 1. creating the column
ALTER TABLE locations
ADD COLUMN region VARCHAR(50);

-- 2. populating the column
UPDATE locations
SET region = CASE
    WHEN state_code IN ('ME', 'NH', 'VT', 'MA', 'RI', 'CT') THEN 'Northeast'
    WHEN state_code IN ('NY', 'PA', 'NJ') THEN 'Mid-Atlantic'
    WHEN state_code IN ('OH', 'MI', 'IN', 'IL', 'WI', 'MN') THEN 'Midwest'
    WHEN state_code IN ('MO', 'KS', 'OK', 'TX', 'AR', 'LA', 'MS') THEN 'South'
    WHEN state_code IN ('ND', 'SD', 'NE', 'KS', 'IA', 'MO') THEN 'Great Plains'
    WHEN state_code IN ('MT', 'ID', 'WY', 'UT', 'CO', 'NV', 'AZ', 'NM') THEN 'West'
    WHEN state_code IN ('WA', 'OR', 'CA', 'AK', 'HI') THEN 'Pacific'
    ELSE 'Unknown'
END;

-- 3. Saving the new version as new_locations
COPY locations TO 'C:\Users\laerc\Desktop\ci\new_locations.csv' DELIMITER ',' CSV HEADER;
```

5. Answering the Business Questions

To address the business questions, the following actions were taken:

The first question, which concerned the total numbers of hospitals, doctors, patients, and specializations, was answered using both SQL and Power BI.

SQL	Power BI

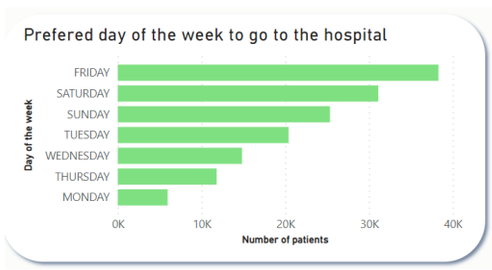
`-- Business Questions`
`--1. Can you provide the total number of hospitals, doctors, patients, and specializations?`
`SELECT COUNT(DISTINCT facility_id) AS total_number`
`FROM hospitals`
	

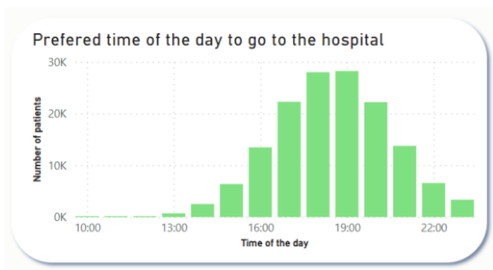
After creating the 'Region' feature using SQL, I executed queries to extract the required information regarding regions. Additionally, I included a filter to the Power BI dashboard to select specific regions.

SQL	Power BI

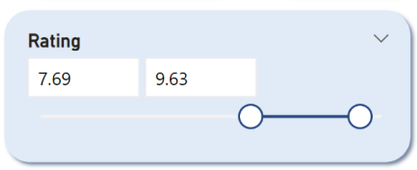
```sql
--2.    Which regions of the US have a higher concentration of doctors?
SELECT COUNT(DISTINCT doc.npi) AS number_of_doctors, loc.region
FROM doctors AS doc
INNER JOIN locations AS loc
ON CAST(doc.zip AS integer) = loc.zip
WHERE loc.region != 'Unknown'
GROUP BY loc.region
ORDER BY number_of_doctors DESC
```

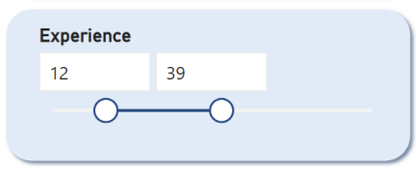
To address the question of the most common time for patients to visit a hospital, I created bar charts and filters on the second page of the dashboard.





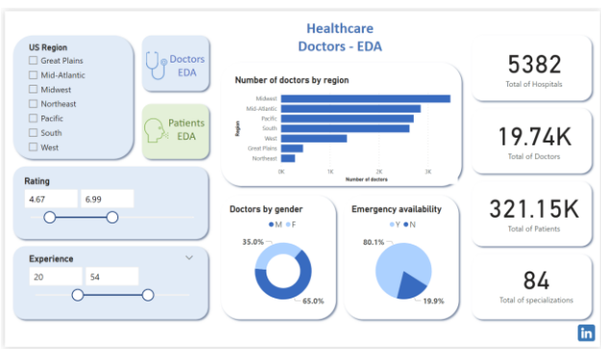
I added sliders to the Power BI dashboard, which were linked to doctors' ratings and experience to investigate their relationship with availability. The dashboard updates automatically in response to changes made to the filters.

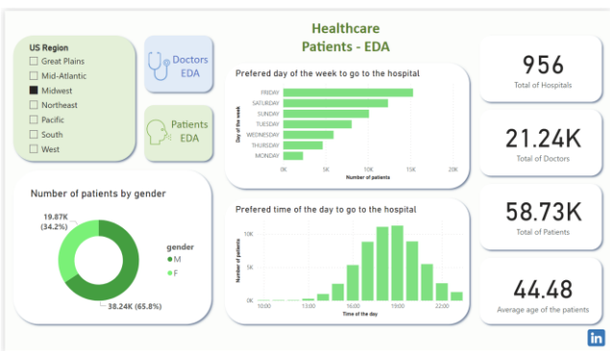




6. Creating the Dashboards

I utilized Power BI to develop the dashboard, which was organized into two pages. The first page presents a summary of doctors' information, while the second page provides patients' information. I created the background design using Power Point, and incorporated filters and buttons to facilitate navigation between the pages.





7. Conclusion and Future Steps

This project utilized Power BI and SQL to answer four key questions. Firstly, the total number of hospitals, doctors, patients, and specializations was obtained from the available datasets. Secondly, the analysis revealed that certain regions in the US had a higher concentration of doctors compared to others. Thirdly, no significant correlation was found between doctors' ratings, experience, and availability. Finally, the most common time for patients to visit a hospital was during the night-time, particularly in the early evening hours. These insights could be utilized to optimize healthcare resource allocation and improve patient outcomes.

One potential future step is to develop a system that enables patients to identify hospitals with greater doctor availability during their preferred time.