**COLUMBIA ASIA HOSPITAL PROJECT**

**OBJECTIVE QUESTIONS**

1. **In analyzing the hospital dataset with Power BI, ensure data cleaning to address inconsistencies and missing values before further analysis.**

 **Load Data:** I began by loading both the Doctor\_Patients\_Data and Hospital\_ER datasets into Power BI. This was done by selecting "Get Data" from the "Home" tab, choosing the appropriate file format, and importing the datasets to initiate my analysis.

 **Open Power Query:** I accessed the Power Query Editor by clicking on "Transform Data" in the "Home" tab of Power BI. Power Query served as the platform for executing all the necessary data cleaning tasks.

 **Handle Missing Values:** I carefully examined each column for missing values. For instance, any blank entries in the patient\_sat\_score column were replaced with 0 to prevent nulls from distorting the results. In cases where the Doctor Name was missing, I substituted it with "Unknown" to ensure all records were complete. I used the "Replace Values" function to make these replacements quickly.

 **Standardize Text Fields:** To maintain consistency, I standardized text-based fields like patient\_race and department\_referral. In Power Query, I selected these columns and used the "Transform" > "Format" > "Capitalize Each Word" option to ensure uniform formatting, making the data more coherent and consistent in visual reports.

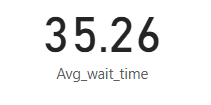
 **Ensure Uniform Date Format:** I reviewed the date column and set its data type to "Date/Time" to ensure a consistent format across all date entries. This step helped prevent any potential issues related to varying date formats.

 **Apply and Load:** After completing the necessary data cleaning steps, I clicked "Close & Apply" to save and load the cleaned data back into Power BI, ready for further analysis and visualizations in the report tabs.

1. **Assess the Average Waiting Time: Analyse the patient wait times to identify the average duration a patient spends before receiving care.**

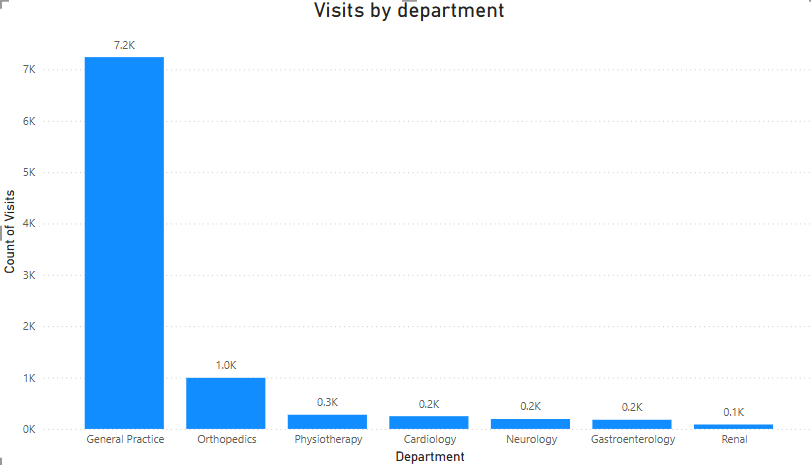
I calculated the average patient wait time using DAX by creating a measure. This measure derives the average directly from the patient\_waittime column in the Hospital ER table, accurately representing patient wait experiences.

**Formula : **

****

1. **Visits by Department Referral: Calculate the total number of visits to each department based on referrals to understand which departments are most frequently visited.**

**Visualization**

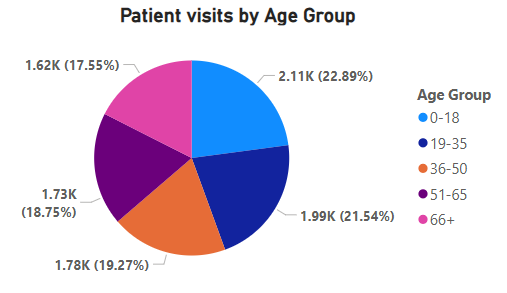
****

**Insights**

* The visual illustrates the distribution of patients across departments based on referrals.
* The General Practice department receives the highest number of patients (7.2K).
* The Renal department has the lowest patient count (0.1K).

1. **Patient Visits by Age Group: Segregate patient visits according to different age groups to see which demographics utilize healthcare services the most.**

**Visualization:**

****

**Insights:**

* The visual provides insights into the patient count across different age groups.
* The highest number of patients belong to the 0-18 age group, with a total of 2,110.
* The 66+ age group has the lowest patient count, totalling 1,617.

**Approach Used:**

* + The **Age Group** column was created using the **Custom Column** feature in Power BI, categorizing patients based on their age into predefined groups.
  + The formula used for the custom column:

Age Group =

SWITCH(

    TRUE(),

    'Hospital ER (1)'[patient\_age] <= 18, "0-18",

    'Hospital ER (1)'[patient\_age] <= 35, "19-35",

    'Hospital ER (1)'[patient\_age] <= 50, "36-50",

    'Hospital ER (1)'[patient\_age] <= 65, "51-65",

    "66+"

)

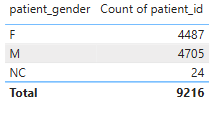
* A **Pie Chart** was created using the **Age Group** column and **Patient ID** to visualize the distribution of patients across different age groups.

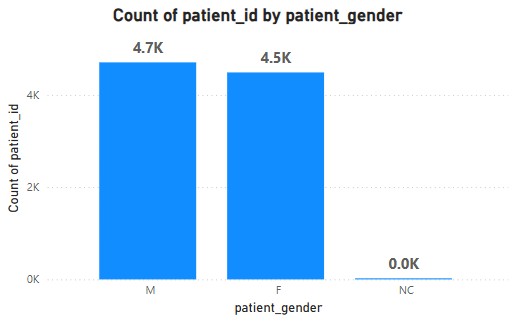
1. **Were there any Null values in the data? What would be the best way to handle these Null values and which approach have you opted for?**

* The **"Patient Sat Score"** column had null values, and I needed to decide the best way to handle them.
* There are different ways to deal with missing data depending on the scenario. For example, if the missing values were in a numerical column like **Age**, I could have used the **average age** to fill them.
* The most straightforward approach is to remove rows with null values, but that would mean losing important data, which isn’t always ideal.
* In this case, since a large portion of the **"Patient Sat Score"** column had null values, calculating an accurate **mean or average** wouldn’t be reliable. Instead of deleting the rows, I replaced the null values with **0** to maintain data integrity and avoid errors in analysis.

1. **Is there any relation between the number of visits and the Gender of the patients?**

**Visualization:**

****

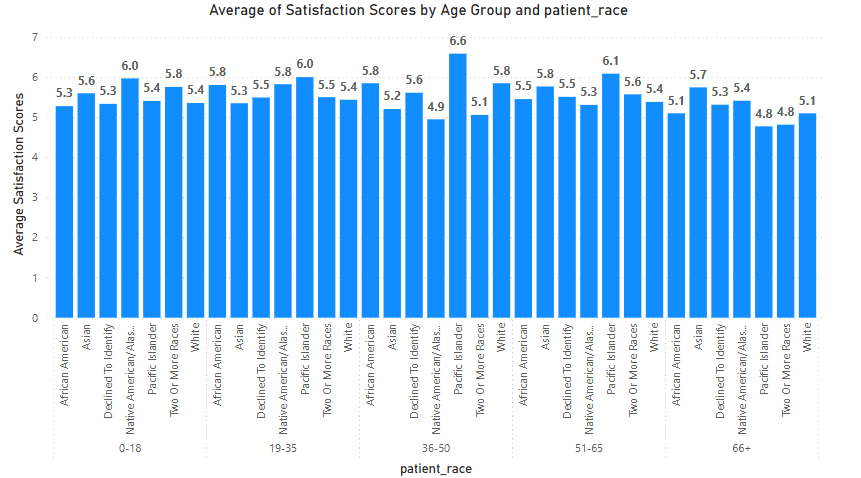
****

**Insights:**

* The visual indicates that **gender does not impact the number of visits**.
* Both **Male (4.5K)** and **Female (4.7K)** patients have a similar number of visits, with no significant difference.
* The category **NC (24)** is not explicitly defined in the schema, but based on my judgment, it likely represents patients who did not specify their gender. Given the low count, it does not have a meaningful impact on the overall trend.
* Therefore, there is **no clear relationship between gender and the number of visits**.

1. **Average Satisfaction by Demographics: Determine the relationship between patient satisfaction scores, their age groups, and racial backgrounds to pinpoint areas for improvement in patient experience.**

**Visualization:**

****

**Insights:**

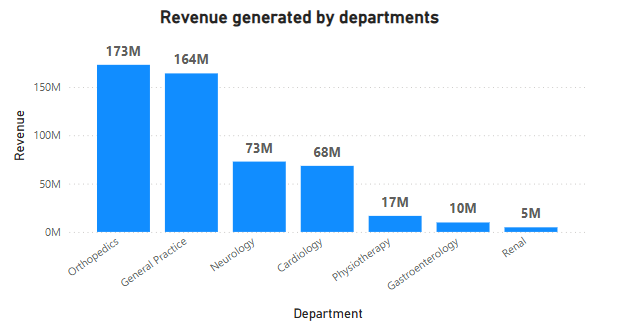
* Since a significant portion of the dataset contained **null values**, calculating the average **satisfaction score** directly would not provide an accurate representation.
* To address this, a **measure** was created to calculate the average satisfaction score while **excluding null values (0)**. Including all ratings would have resulted in an **artificially low average (around 1-2)**, which could negatively impact the hospital’s reputation.
* The dataset indicates that the **satisfaction rating scale ranges from 1 to 10**, with an **average rating of 5.47**, which is quite **neutral**.
* The visual helps in understanding the relationship between **patient satisfaction scores, age groups, and racial backgrounds**, enabling insights to improve the **patient experience**.

**Approach Used:**

* The **previously created Age Group and Race columns** were used to build a **Clustered Column Chart**, incorporating the **average satisfaction score measure**.
* This visualization helps in identifying patterns and relationships between **patient satisfaction scores, different age groups, and racial backgrounds**.

1. **The hospital's managing director seeks to evaluate the revenue of each department to understand how much revenue is generated by each.**

**Visualization:**

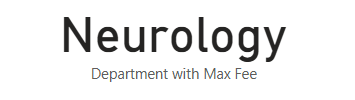
****

**Insights:**

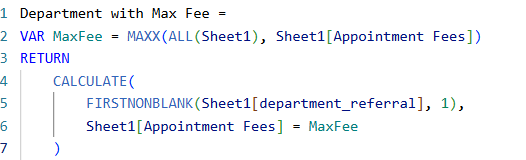
* + The visual provides insights into the **revenue generated by each department**.
  + The **Orthopaedics department** contributes the **highest revenue**, totalling **$173M**.
  + The **Renal department** has the **lowest revenue**, amounting to **$5M**.

1. **Which department is charging the highest appointment fees in general? Use an aggregation DAX function to solve this question.**

**Visualization:**

****

**Formula Used:**

****

**FIRSTNONBLANK(Sheet[department\_referral], 1)** ensures that the department name is returned when the highest appointment fee is found.

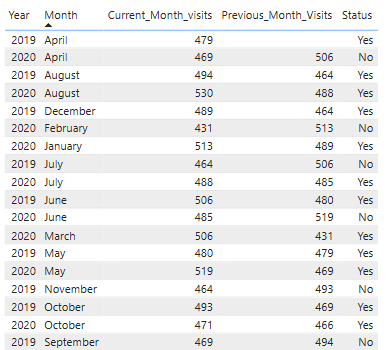
**CALCULATE()** filters the table to find the department associated with the maximum fee.

**Insights:**

* Neurology department has the highest appointment fees (i.e., 1500).

1. **Create a tabular visualization in the Report view which consists of Month-wise total visits in the hospital.  Add a third column in the table that consists of the previous month’s total visits for each month’s row. Also, include a column that states whether the visits in a month are greater than that of the previous month's visits.**

**Visualization:**

****

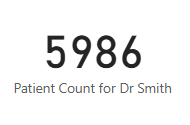
**Insights:**

* The visual displays the **patient visit trends**, showing the number of visits for both the **current month** and the **previous month**.
* Based on these visit counts, the **Status** column indicates whether the number of visits has **increased or decreased** compared to the previous month.

1. **Using ‘Calculate’ and a row iteration DAX function calculate the total number of patients who have visited Dr. Smith.**

Patient count for Dr Smith = 5986

**Visualization:**

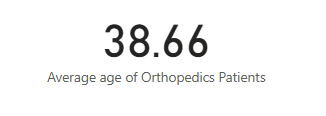
****

**Formula Used:**

****

1. **Calculate the average age of the patients who visit the Orthopedics department. Will the approach used to calculate this metric be different if the requirement had been all departments’ average age?**

**Visualization:**

****

**Formula Used:**

****

**Insights:**

* + The **average age** of **Orthopedics** patients is **38.66**.
  + If the goal was to calculate the **average age across all departments**, the **CALCULATE** function wouldn't be necessary since there are no specific conditions to filter.
  + In this case, the average age could be obtained directly using the **AVERAGE** function with the following formula: **Average age of departments = AVERAGE('Hospital ER(1)'[patient\_age])**

1. **Were there any data format issues in the data, and if there were/are how you handle them?**

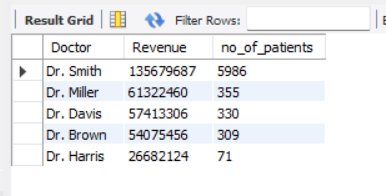
* After analysing the dataset, **no data format issues** were identified.
* The **patient wait time** was stored as a numerical value, which, according to the schema, represented **minutes**, so no corrections were needed.
* Additionally, **new columns** were created for analysis, and appropriate **data types** were assigned to ensure consistency.
* If any data format issues had been found, such as **incorrect date and time formats**, they would have been corrected by updating the data types to ensure **accurate analysis** and prevent errors.

1. **When we add a column in Power Query what’s the code that comes in M language in the formula bar? What do you know about M-query?**

* **M-Query** is the scripting language used in **Power Query** within **Power BI** for data transformation.
* It is a **functional language**, meaning it processes data through a series of functions and transformations.
* When a **new column** is added in **Power Query**, the corresponding **M code** is automatically generated in the **formula bar**. This code represents the **transformation steps**, allowing users to track and manage changes efficiently.

1. **Identify the top 5 doctors who generated the most revenue but had the fewest patients. (SQL)**

**Visualization:**

****

**SQL Code:**

select `Doctor Name` as Doctor, sum(`Total Bill`) as Revenue,

count(distinct patient\_id) as no\_of\_patients

from doctor

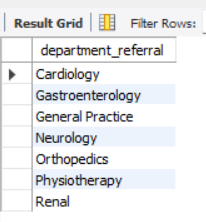
group by Doctor

order by Revenue desc , no\_of\_patients asc

limit 5;

1. **Find the department where the average waiting time has decreased over three consecutive months. (SQL)**

**Visualization:**

****

**SQL Code:**

with cte as

(select department\_referral, `date`,

round(avg(patient\_waittime), 2) as average\_wait\_time

from hospital\_er

group by department\_referral, `date`),

cte2 as

(select department\_referral, `date`, average\_wait\_time,

lag(average\_wait\_time, 1) over (partition by department\_referral order by `date`) as previous\_month\_average,

lag(average\_wait\_time, 2) over (partition by department\_referral order by `date`) as previous\_2\_month\_average,

lag(average\_wait\_time, 3) over (partition by department\_referral order by `date`) as previous\_3\_month\_average

from cte)

select department\_referral

from cte2

where average\_wait\_time < previous\_month\_average

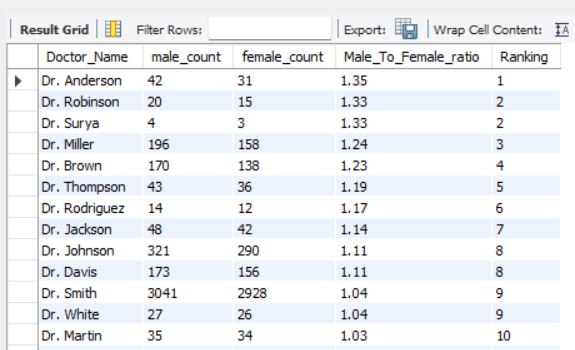
and previous\_month\_average < previous\_2\_month\_average

and previous\_2\_month\_average < previous\_3\_month\_average

group by department\_referral;

1. **Determine the ratio of male to female patients for each doctor and rank the doctors based on this ratio. (SQL)**

**Visualization:**

****

**SQL Code:**

with cte as

(select d.`Doctor Name` as Doctor\_Name,

sum(case when h.patient\_gender = "M" then 1 else 0 end) as male\_count,

sum(case when h.patient\_gender = "F" then 1 else 0 end) as female\_count

from hospital\_er h

inner join doctor as d on h.patient\_id = d.patient\_id

group by Doctor\_Name),

cte2 as

(select Doctor\_Name, male\_count, female\_count,

round((male\_count/female\_count),2) as Male\_To\_Female\_ratio

from cte)

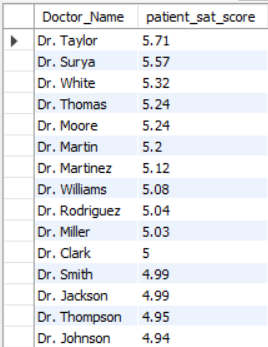
select Doctor\_Name, male\_count, female\_count, Male\_To\_Female\_ratio,

dense\_rank() over(order by Male\_To\_Female\_ratio desc) as Ranking

from cte2;

1. **Calculate the average satisfaction score of patients for each doctor based on their visits. (SQL)**

**Visualization:**

****

**SQL Code:**

select d.`Doctor Name` as Doctor\_Name,

round(avg(case when h.patient\_sat\_score = "" then 5 else h.patient\_sat\_score end),2) as patient\_sat\_score

from hospital\_er h

inner join doctor d

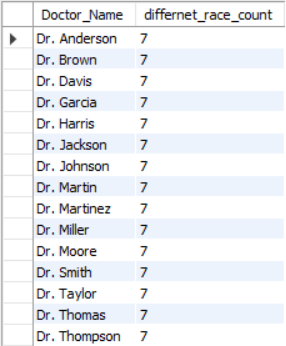
on h.patient\_id = d.patient\_id

group by Doctor\_Name

order by patient\_sat\_score desc;

1. **Find doctors who have treated patients from different races and calculate the diversity of their patient base. (SQL)**

**Visualization:**

****

**SQL Code:**

select d.`Doctor Name` as Doctor\_Name,

count(distinct h.patient\_race) as differnet\_race\_count

from hospital\_er h

inner join doctor d

on h.patient\_id = d.patient\_id

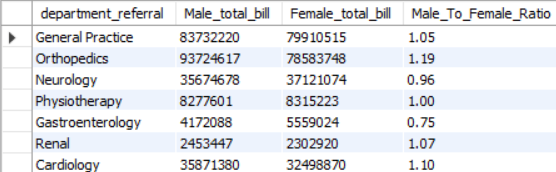
group by Doctor\_Name

having count(distinct h.patient\_race) > 1

order by differnet\_race\_count desc;

1. **Calculate the ratio of total bills generated by male patients to female patients for each department. (SQL)**

**Visualization:**

****

**SQL Code:**

select h.department\_referral,

sum(case when patient\_gender = "M" then d.`Total Bill` end) as Male\_total\_bill,

sum(case when patient\_gender = "F" then d.`Total Bill` end) as Female\_total\_bill,

round(sum(case when patient\_gender = "M" then d.`Total Bill` end) / sum(case when patient\_gender = "F" then d.`Total Bill` end),2) as Male\_To\_Female\_Ratio

from hospital\_er h

inner join doctor d

on h.patient\_id = d.patient\_id

group by h.department\_referral

1. **Update the patient satisfaction score for all patients who visited the "General Practice" department and had a waiting time of more than 30 minutes. Increase their satisfaction score by 2 points, but ensure that the satisfaction score does not exceed 10. (SQL)**

UPDATE hospital\_er

SET patient\_sat\_score = LEAST(patient\_sat\_score + 2, 10)

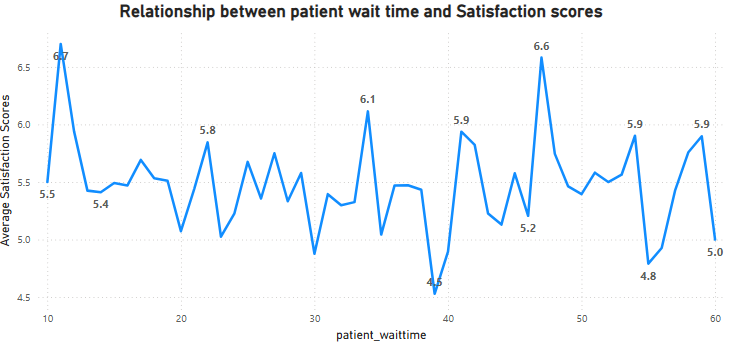
WHERE LOWER(department\_referral) = "general practice"

AND patient\_waittime > 30;

**SUBJECTIVE QUESTIONS**

1. **What is the relation between patient wait time and satisfaction scores?**

**Visualization:**

****

**Insights:**

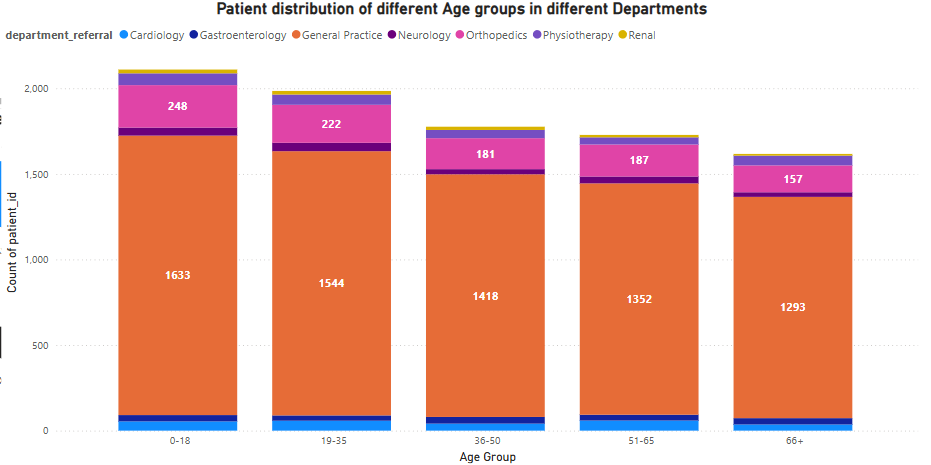
* The visual illustrates the **relationship between patient wait time and satisfaction score**.
* The **satisfaction score** remains relatively consistent across different wait times, but the **highest score (6.70)** is observed at a **10-minute wait time**, while the **lowest score (4.53)** occurs around **40 minutes**.
* Since **75% of the ratings are missing**, the analysis is based on only **25% of the available data**, which may not fully reflect patient sentiment. To ensure a more **accurate assessment**, the hospital should encourage patients to provide feedback.

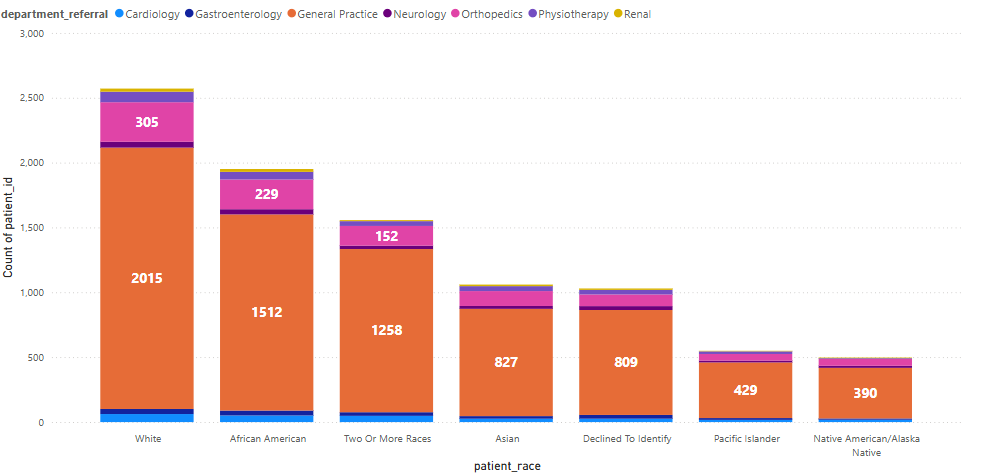
**Approach Used:**

* A **line chart** was created using the **Wait Time** column and the **Average Satisfaction Score** measure.
* The **Average Satisfaction Score** measure was used because **75% of the ratings data is missing**, so the analysis was conducted using the remaining **25%** to ensure a more **reliable interpretation**.

1. **How do patient demographics affect the frequency of visits to different departments?**

**Visualization:**

****

****

**Insights:**

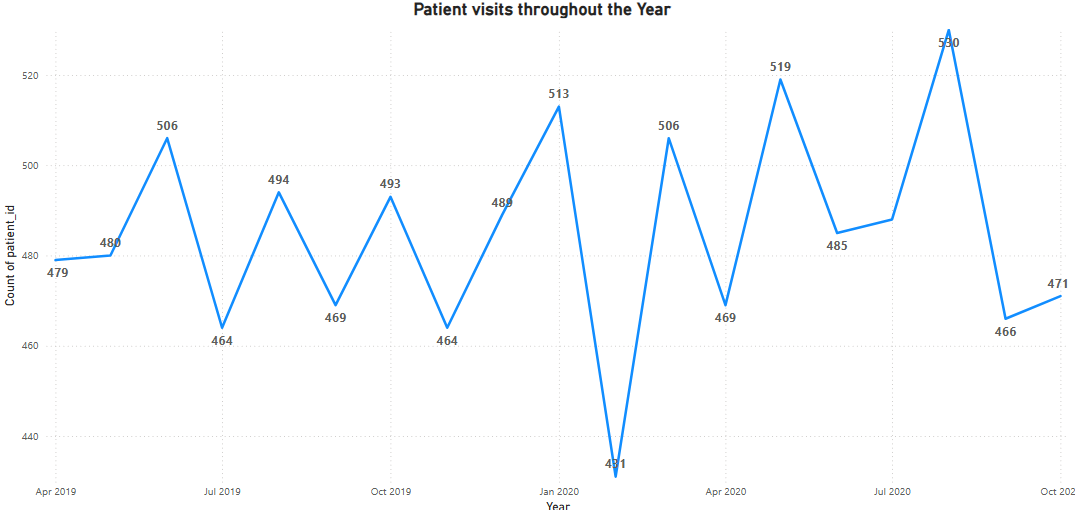
* + The visuals provide insights into the **distribution of patients across departments** based on **age and race**.
  + This analysis helps in understanding **patient demographics** more effectively.
  + The findings can be used to **enhance facilities and services** based on observed trends.

**Approach Used:**

* + A **stacked column chart** was created to visualize the **distribution of patients**.
  + The chart includes **Department Referral, Age Group, Race,** and the **Patient ID count** to represent the data effectively.

1. **Is there a noticeable trend in the volume of patient visits throughout the year?**

**Visualization:**

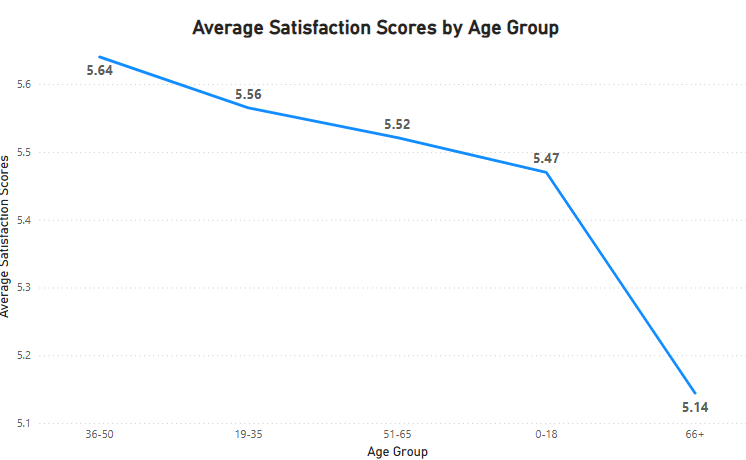
****

**Insights:**

* + The visual showcases **patient visits throughout the year**.
  + The **highest number of visits** occurred in **August 2020 (530)**, while the **lowest** was recorded in **February 2020 (430)**.
  + However, due to **inconsistencies in the data**, identifying a clear trend is challenging, which impacts the accuracy of the analysis.

1. **Which age groups report the highest and lowest satisfaction scores?**

**Visualization:**

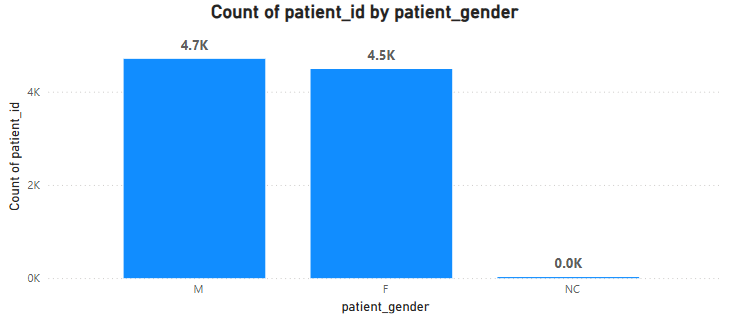
****

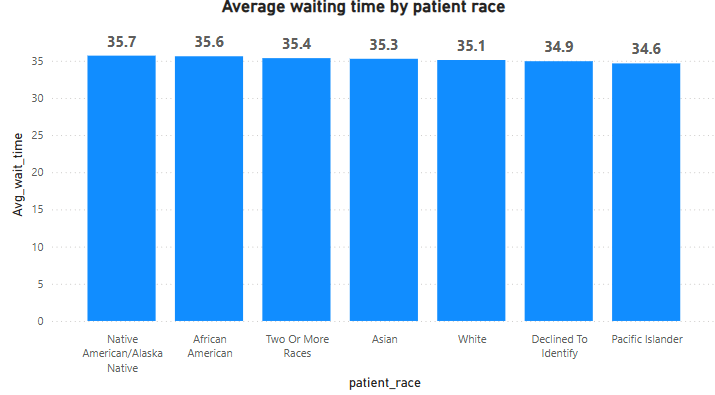
**Insights:**

* + The visual represents satisfaction scores across different age groups.
  + A line chart was created using the **average satisfaction score** measure and the **age group** column.
  + The **36-50 age group** has the highest satisfaction score of **5.64**.
  + The **66+ age group** has the lowest satisfaction score of **5.14**.

1. **Say someone outside of the hospital claims that there is racial or gender-based discrimination in the hospital, how will you identify whether the claim was right or not?**

**Visualization:**

****

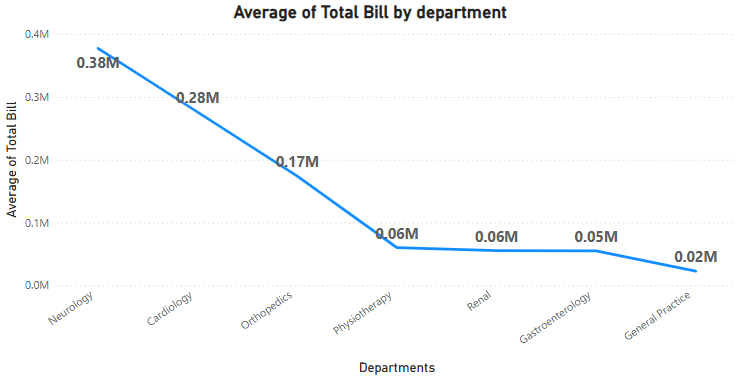
****

**Insights:**

* The visual on **patient distribution** shows that the number of male and female patients is nearly equal, with no significant difference between them.
* The **average waiting time by patient race** indicates that all races experience similar waiting times, suggesting fair treatment for all.
* This highlights that medical care is provided **equally**, regardless of **gender or race**, with no signs of discrimination or bias in the hospital.

1. **The hospital management intends to offer discounts to patients. How should these offers/discounts be assigned to patients, on what basis, and why?**

**Visualization:**

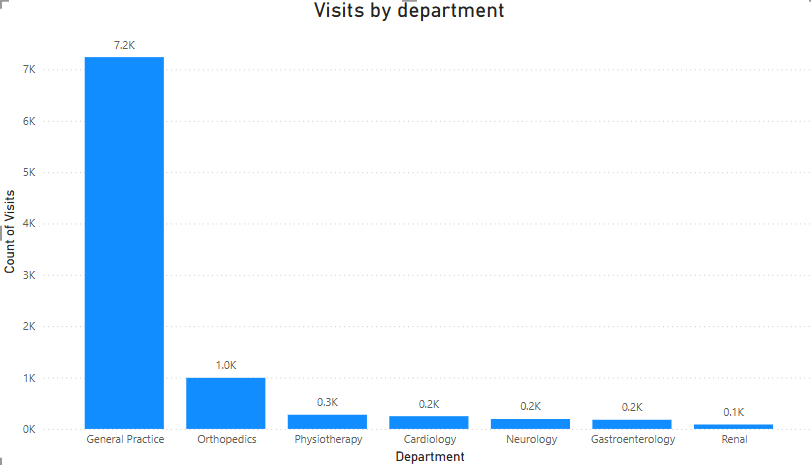
****

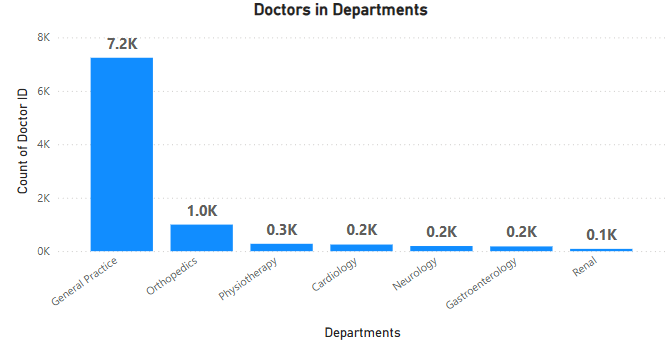
**Insights:**

* + The visual represents the **revenue generated** by each department.
  + **Neurology (0.38M)** and **Cardiology (0.28M)** are the top revenue-generating departments, indicating a higher number of patient visits.
  + Previous analysis showed that **Neurology** has the highest appointment fee of **$1500**.
  + Considering these factors, offering **discounts** to patients in the **Neurology** and **Cardiology** departments could be beneficial.

1. **The hospital has a budget to hire 2-3 new doctors. They have asked for your suggestions on which departments they should hire.**

**Visualization:**

****

****

**Insights:**

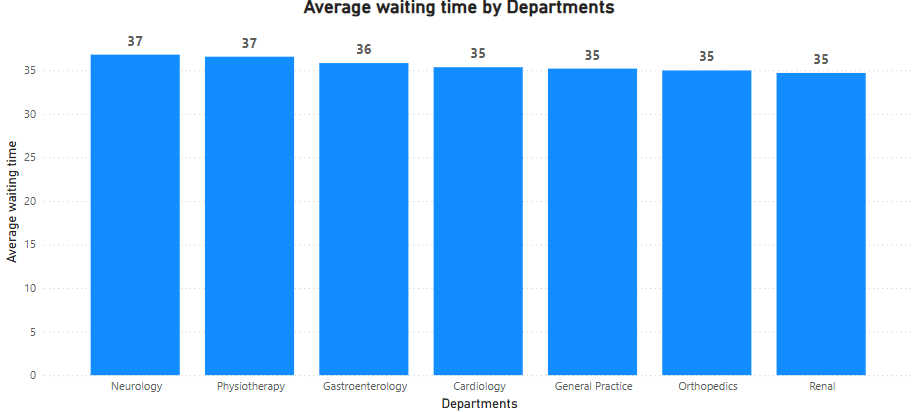
* + The visual illustrates the **number of patients** across different departments.
  + **General Practice (7.2K)** has the highest patient count, followed by **Orthopedics (1.0K)**.
  + The **Renal department (0.1K)** has the lowest patient count.
  + To manage the high patient load, **additional doctors** should be hired in departments with **higher patient numbers**.
  + Specifically, **General Practice, Orthopedics, and Physiotherapy** require more doctors.
  + **General Practice** currently has only **three doctors** managing **7.2K patients**, making it a priority for new hires.

1. **Is the hospital profitable? How will you determine the profitability?**

* **Essential data**, including doctor salaries, maintenance costs, and faculty salaries, is not available, making it difficult to perform a complete analysis.
* While the **revenue generated** can be calculated, the **operational costs** cannot be determined due to the missing data.
* Without these key financial details, the **overall profitability and cost-effectiveness** of the departments cannot be accurately analyzed.
* As a result, the **analysis remains incomplete** and may not provide actionable insights.

1. **Any Department for which the waiting time is oddly large?**

**Visualization:**

****

**Insights:**

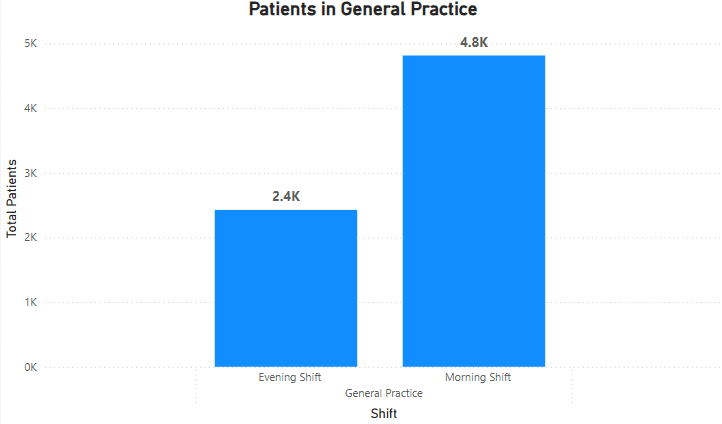
* The visual represents the average wait time across different departments.
* The waiting times are fairly consistent across all departments.
* No department shows an unusually long wait time compared to others.

1. **Come up with strategies to provide discounts to the patients.**

* As mentioned earlier in **Question 6**, **Neurology** and **Cardiology** are the highest revenue-generating departments and have **high appointment fees**.
* These departments are ideal candidates for offering **discounts**.
* Introducing **medical campaigns** with **discounted rates on specific days** could attract more patients.
* This strategy would **increase hospital profitability** while making healthcare more **affordable and accessible** for patients.

1. **Say you need to align the doctors of the “General Practice” department to work in one of the two shifts, how will you identify what will these two shifts' timings be, and how will you divide the doctors in these two shifts? And also will this 2 shift policy be helpful for the hospital?**

**Visualization:**

****

**Insights:**

* The visual represents the **distribution of patients** based on shift timings—**morning or evening**.
* The **morning shift (4.8K patients)** has significantly more patients than the **evening shift (2.4K patients)**.
* Previous analysis showed that the **General Practice department** has only **three doctors**, leading to **high workload and inefficiency**.
* To improve efficiency, **additional doctors should be hired** and **distributed across different shifts**.
* Implementing a **two-shift policy** would ensure doctors get adequate rest, enhancing their **performance and patient care**.
* In the medical field, even **minor negligence** can have serious consequences, so ensuring **well-rested doctors** is crucial for **quality healthcare**.

1. **What do you understand by PowerBI gateway? What are its use cases?**

* A **Power BI Gateway** acts as a **secure bridge** between **on-premises data sources** and the **Power BI service (cloud)**, enabling **seamless data transfer**.
* It allows users to **access on-premises data** without the need to migrate it to the cloud.
* The gateway ensures **secure integration** and supports **automatic data refreshes**, keeping cloud-based **Power BI reports and dashboards** updated with the latest data from local servers.

**Use Cases:**

* + **Scheduled Data Refresh** – Automates data updates at predefined intervals.
  + **Real-Time Data Access (DirectQuery)** – Enables live data retrieval without importing it.
  + **Access to On-Premises Data** – Connects cloud-based Power BI to local databases securely.
  + **Sharing & Collaboration** – Facilitates working with on-premises data while maintaining security and control.

1. **How would you approach this problem, if the objective and subjective questions weren't given?**

If **guidelines were not provided**, my approach would be as follows:

* **Data Cleaning** – I would begin by ensuring that all columns have **appropriate data types** and are free from inconsistencies.
* **Handling Missing Values** – I would identify and address **missing data** using suitable techniques while also creating **separate date and time columns** for better efficiency.
* **Optimizing Measures** – A separate table would be created to store all **calculated measures**, improving **performance and organization**.
* **Basic Metrics & Analysis** – I would establish **key metrics**, analyze them, and use the insights for **further exploration**.
* **Chart Creation** – Based on business requirements, I would generate **visuals** such as **doctor distribution across departments, revenue by department**, and more.
* **Report Compilation** – Finally, I would bring together all **essential visuals** to create a **comprehensive and insightful report**.

1. **Can you analyse and write the type of relationship between the doctor id and department, is it one-to-one?**

* **No,** the relationship between **Doctor ID** and **Department** is **not one-to-one**.
* Instead, it follows a **one-to-many** relationship.
* This is because multiple **doctors** can belong to the same **department** (e.g., General Practice), while each **doctor** is typically assigned to **only one department** at a time.