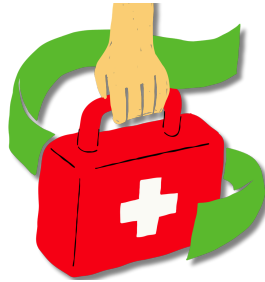# Health Data Analysis in Pakistan

by Abdul Raffay Naeem

I. About the Dataset

The dataset contains 2000 rows and 15 columns, representing health-related data from various hospitals in Pakistan. Each row corresponds to a patient's visit to a hospital, including personal and medical information. Here are the details of each column:

Column Name	Column Description
Hospital Name	Name of the hospital where the patient was treated.
Patient Name	Full name of the patient.
Age	Age of patient in years.
Gender	Gender of the patient (Male or Female).
Date of Admission	Date when the patient was admitted to the hospital (YYYY-MM-DD format).
Date of Discharge	Date when the patient was discharged from the hospital (YYYY-MM-DD format).
Disease	The primary disease or condition for which the patient was treated.
Treatment Cost	Cost of the treatment in PKR (Pakistani Rupees).
Insurance	Whether the patient had insurance coverage (Yes or No).
Blood Group	Blood group of the patient.
Height (cm)	Height of the patient in centimeters.
Weight (kg)	Weight of the patient in kilograms.
Contact Number	Contact number of the patient.
City	City where the patient resides.
Province	Province where the patient resides.

II. Aims of this Analysis

- Analyze patient demographics to understand the age, gender, height, and weight distribution across various hospitals.

- Evaluate disease prevalence to identify the most common diseases treated in hospitals and their distribution across different cities and provinces.

- Examine healthcare costs by assessing the treatment costs associated with different diseases and hospitals, and understanding the financial burden on patients with and without insurance.

- Study hospital performance by comparing patient volume, treatment costs, and other key metrics across hospitals.

- Identify patterns in hospital admissions and discharges to determine peak times and optimize resource allocation and management.

- Conduct an insurance coverage analysis to understand the percentage of patients with insurance coverage and its impact on treatment costs.

- Explore the distribution of blood groups among patients and its relevance to healthcare services.

- Gain geographical insights into patient distribution and healthcare needs across different cities and provinces in Pakistan.

- Investigate potential correlations between patient demographics (e.g., age) and treatment costs.

- Enable data-driven decision making for healthcare providers, policymakers, and researchers to improve healthcare services and outcomes in Pakistan.

III. Data Wrangling and Cleaning

- Contact number was identified as the primary key as it's unique for each entry (contained no duplicates).

- Each column was checked for blank spaces and spelling mistakes using filters.

- The data types of columns were adjusted (Data types of Date of admission and Date of discharge were changed to date, Data type of Treatment Cost was changed to currency.)

- The blank spaces were also checked by selecting our columns -> Selecting "Find and Select" -> Clicking "Go to Special" -> Select "Blanks"-> Press "Okay."

- A new column named "Correlation b/w Patient Age and Treatment Cost" was made using the =CORREL function to determine whether correlation between the two variable exist or not.

- Two New columns named "Month of Admission" and "Month of Discharge" were made using the =MONTH function to aid in analyzing peak times of admission and discharge.

- Created "Age Group" column using =IFS function to help in analysis of age distribution.
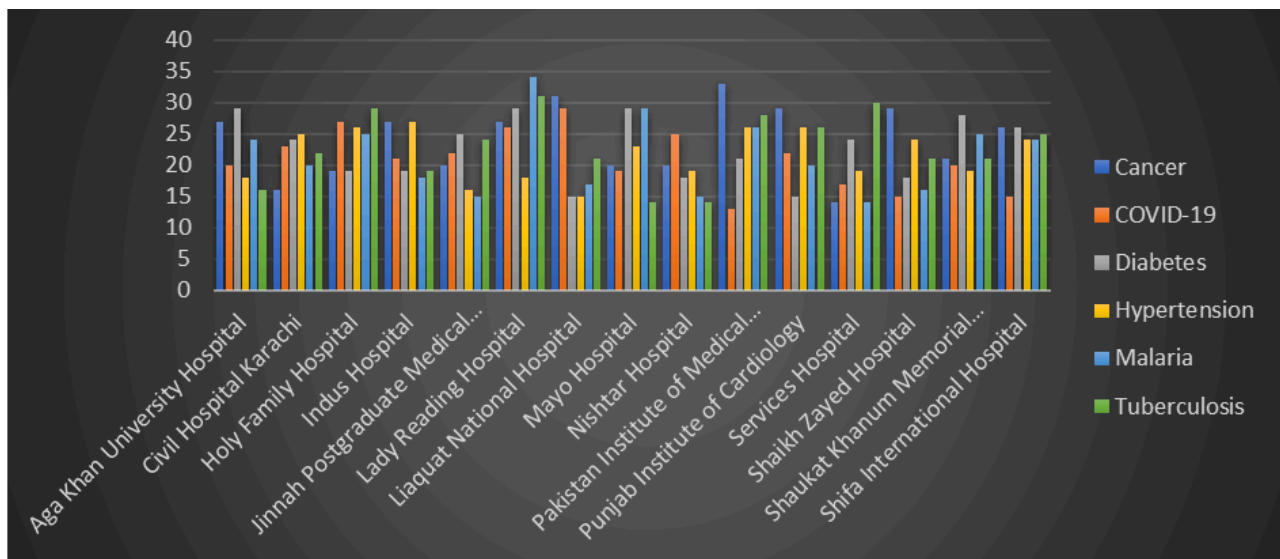
IV. Data Analysis

- **Hospital Statistics:**

Distribution of patients across various hospitals is given as
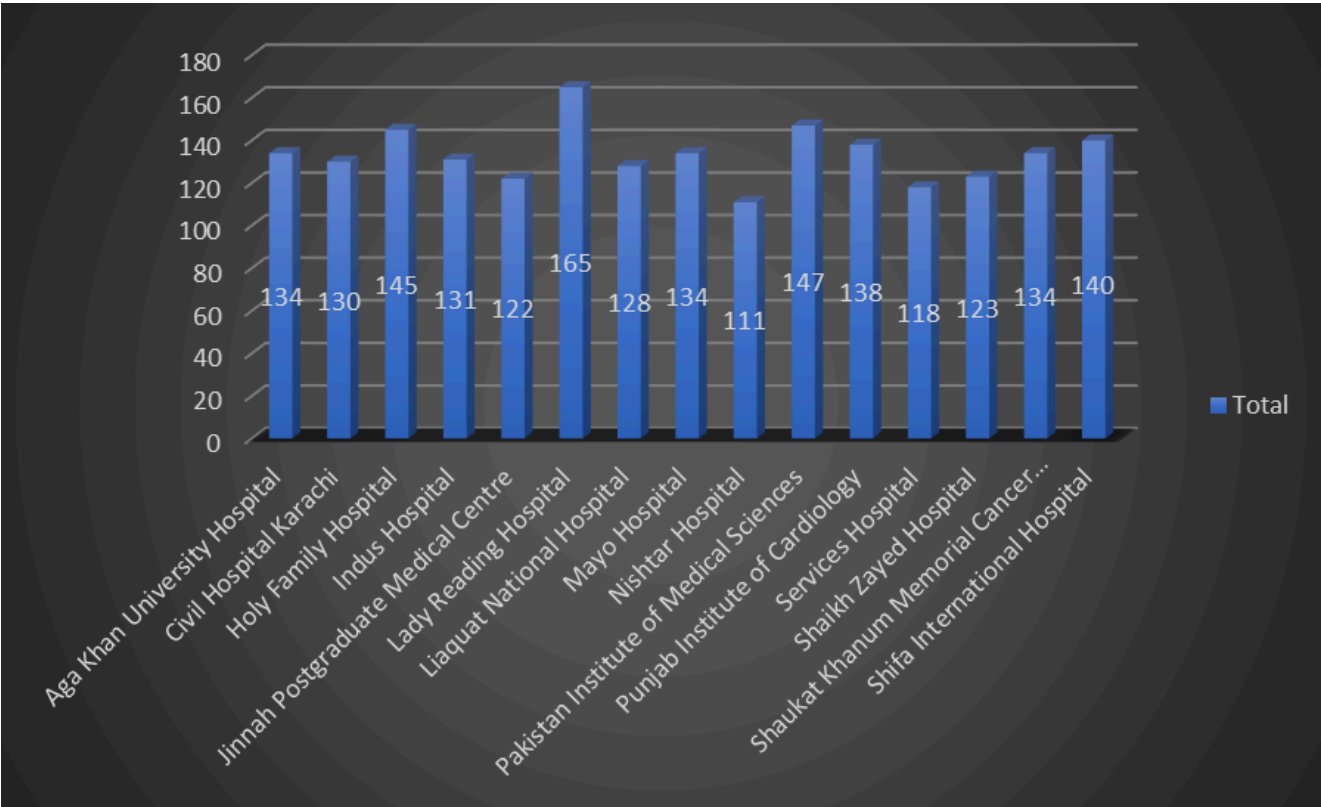
It was noticed that Lady Reading Hospital had the highest number of patients (165) whereas Nishtar Hospital had the lowest amount (111).

- **Disease Frequency:**

Although the total amount of diseases reported in these hospitals are similar, the most common ones include Cancer (359 patients in total) and Tuberculosis (341 patients in total).

| Disease Distribution	Diseases					
Hospital Names	Cancer	COVID	Diabet	Hyper	Malaria	Tuberculosis
Aga Khan University Hospital	27	20	29	18	24	16
Civil Hospital Karachi	16	23	24	25	20	22
Holy Family Hospital	19	27	19	26	25	29
Indus Hospital	27	21	19	27	18	19
Jinnah Postgraduate Medical Centre	20	22	25	16	15	24
Lady Reading Hospital	27	26	29	18	34	31
Liaquat National Hospital	31	29	15	15	17	21
Mayo Hospital	20	19	29	23	29	14
Nishtar Hospital	20	25	18	19	15	14
Pakistan Institute of Medical Sciences	33	13	21	26	26	28
Punjab Institute of Cardiology	29	22	15	26	20	26
Services Hospital	14	17	24	19	14	30
Shaikh Zayed Hospital	29	15	18	24	16	21
Shaukat Khanum Memorial Cancer Hospital	21	20	28	19	25	21
Shifa International Hospital	26	15	26	24	24	25
Grand Total	359	314	339	325	322	341

- **Age Distribution:**
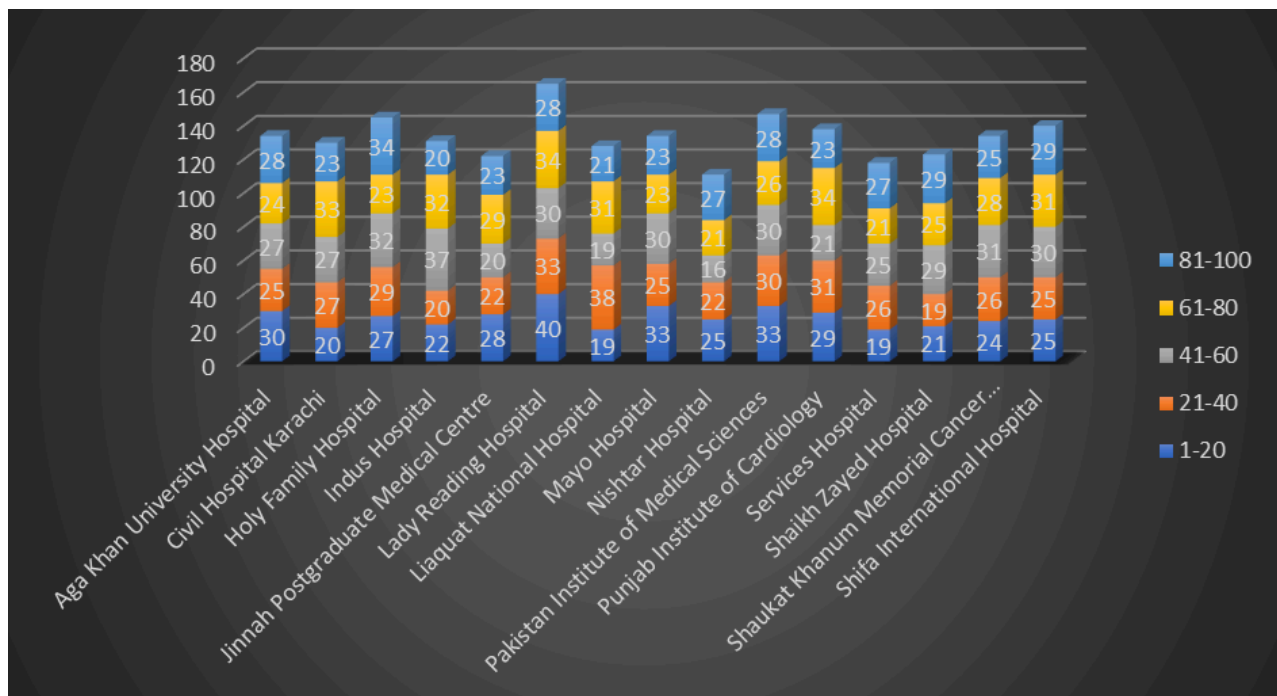
It's observed that the age group of the majority patients coming into these hospitals is 61-80.

| Age Distribution	Age Groups				
Hospital Names	1-20	21-40	41-60	61-80	81-100
Aga Khan University Hospital	30	25	27	24	28
Civil Hospital Karachi	20	27	27	33	23
Holy Family Hospital	27	29	32	23	34
Indus Hospital	22	20	37	32	20
Jinnah Postgraduate Medical Centre	28	22	20	29	23
Lady Reading Hospital	40	33	30	34	28
Liaquat National Hospital	19	38	19	31	21
Mayo Hospital	33	25	30	23	23
Nishtar Hospital	25	22	16	21	27
Pakistan Institute of Medical Sciences	33	30	30	26	28
Punjab Institute of Cardiology	29	31	21	34	23
Services Hospital	19	26	25	21	27
Shaikh Zayed Hospital	21	19	29	25	29
Shaukat Khanum Memorial Cancer Hospital	24	26	31	28	25
Shifa International Hospital	25	25	30	31	29
Grand Total	395	398	404	415	388

- **Gender Ratio:**
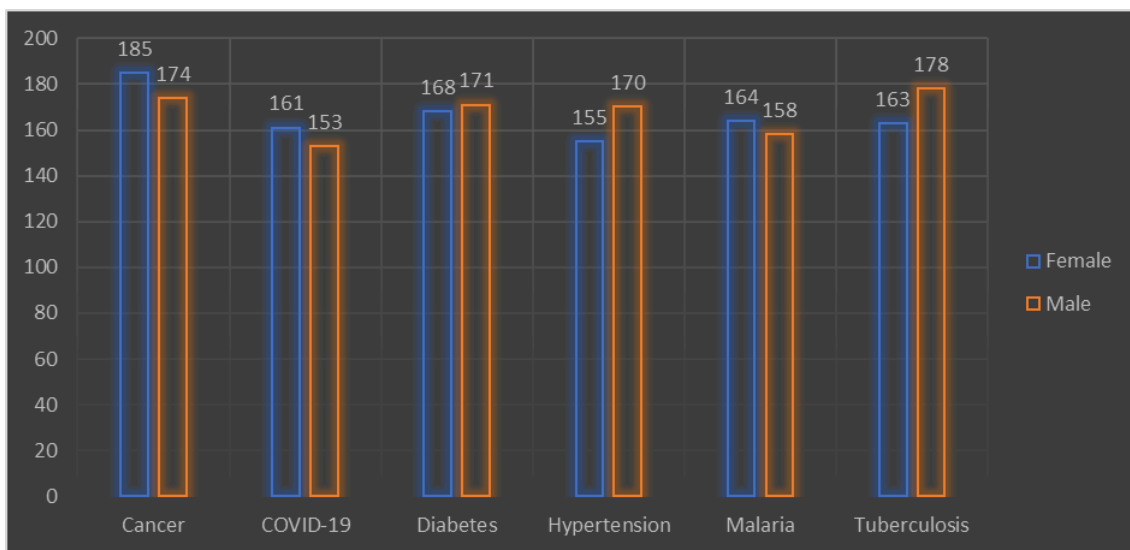
When diseases and genders were analyzed together it showed that the most frequent disease female patients had was Cancer (185 cases) where as for males it's Tuberculosis (178 cases). It was also observed that the overall count of male patients (1004) was higher then female patients (996).
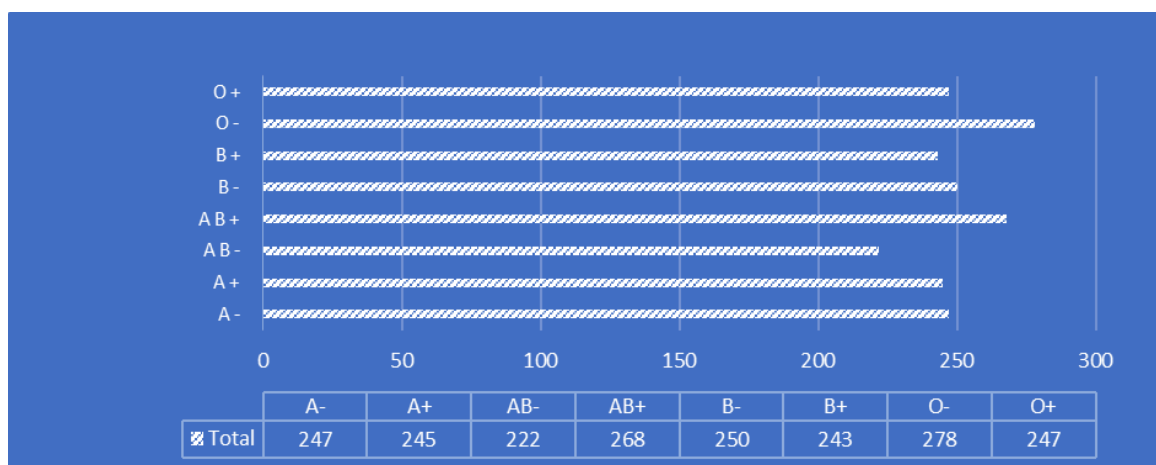
- **Treatment Cost Analysis:**
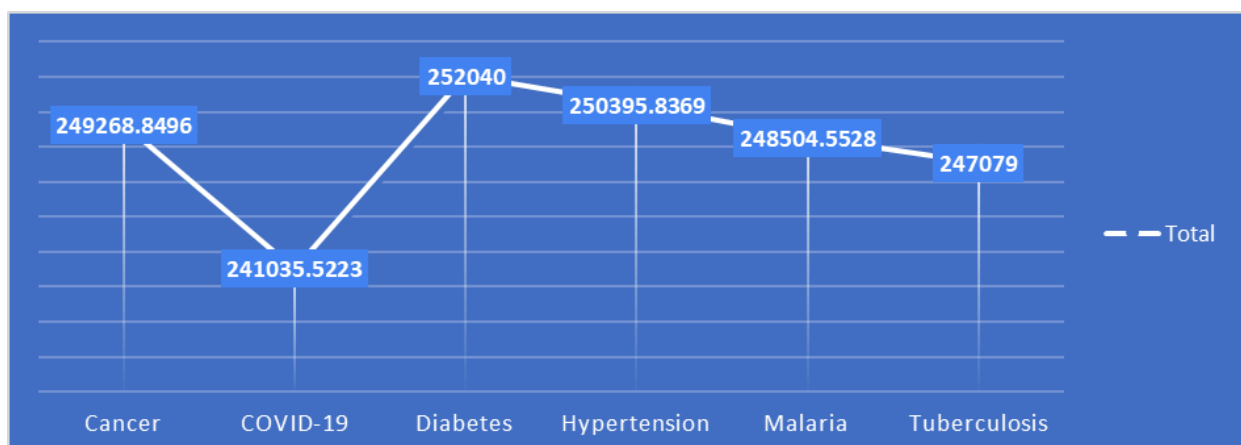
Diabetes has the highest avg treatment cost of Rs 252040 whereas COVID-19 has the lowest avg treatment cost of Rs 241036 (rounded off).

- **Insurance Coverage:**
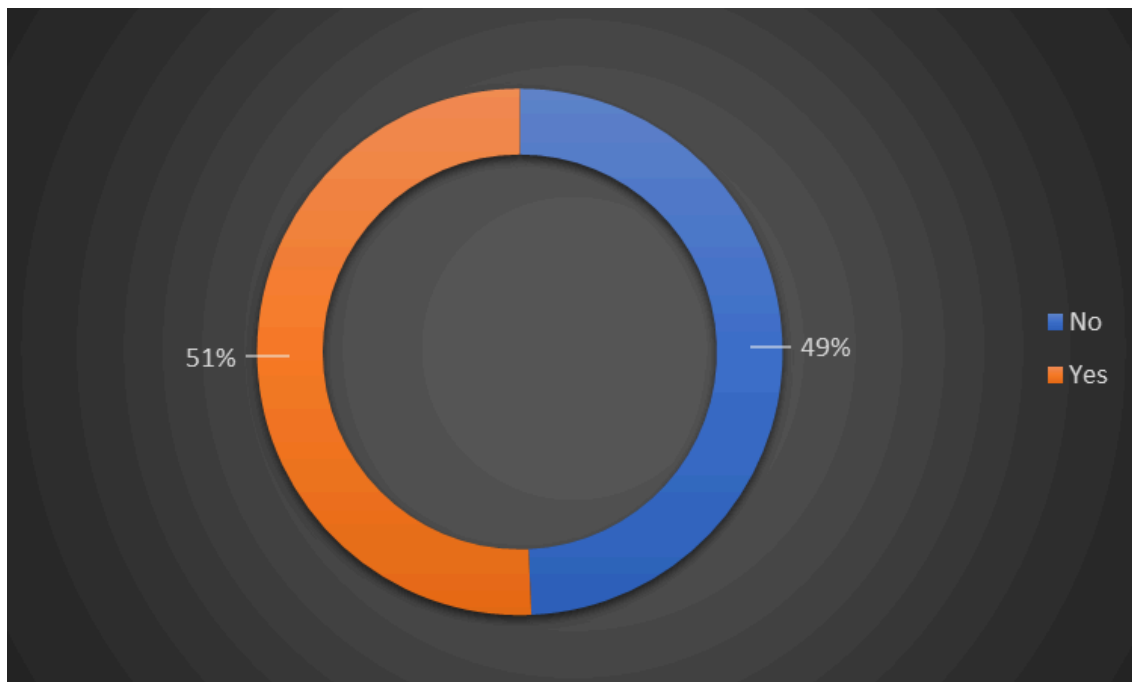
Only 51% patients had insurance whereas the other 49% did not have their expenses covered by insurance.
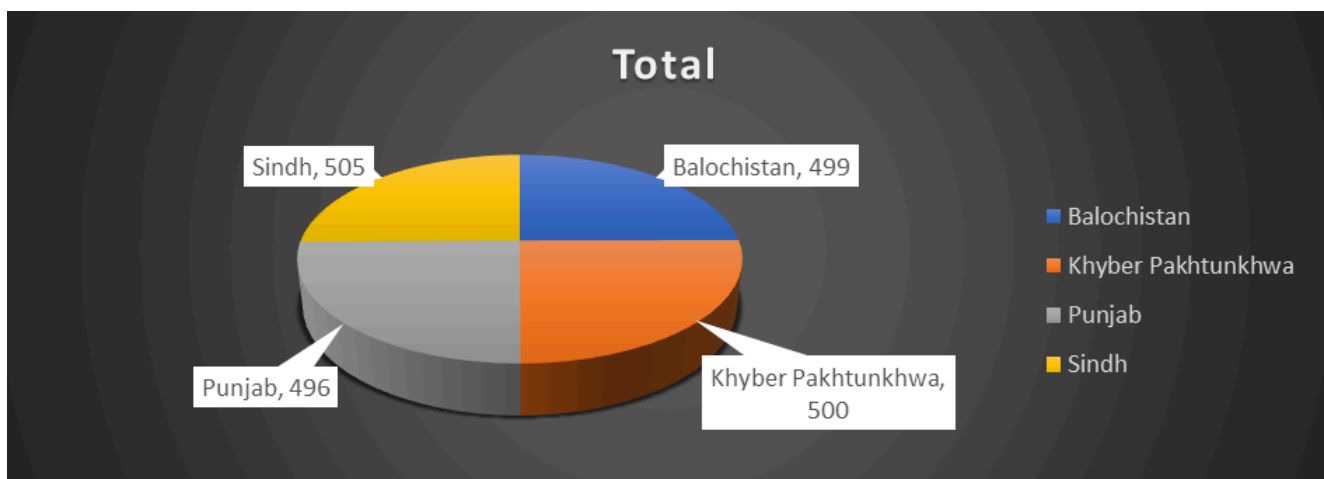
- **Blood Group Distribution:**

	A-	A+	AB-	AB+	B-	B+	O-	O+
Total	247	245	222	268	250	243	278	247

Patients with O- blood group were recorded the most (278) whereas the lowest recorded blood group was AB-

- **City-wise Patient Count:**
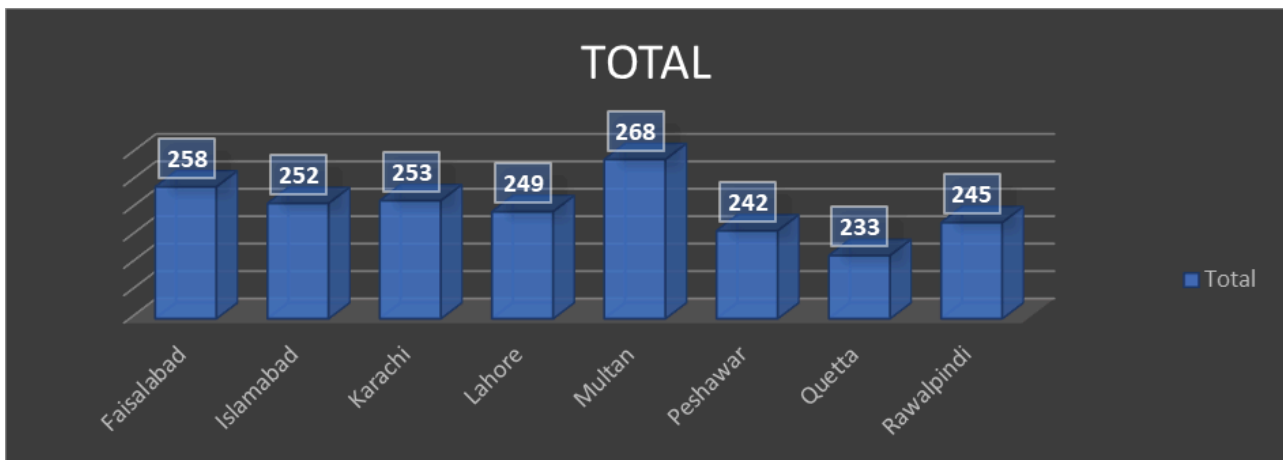
On analyzing city wise patient count, we noticed that city Multan had the highest number of patients reported (268).
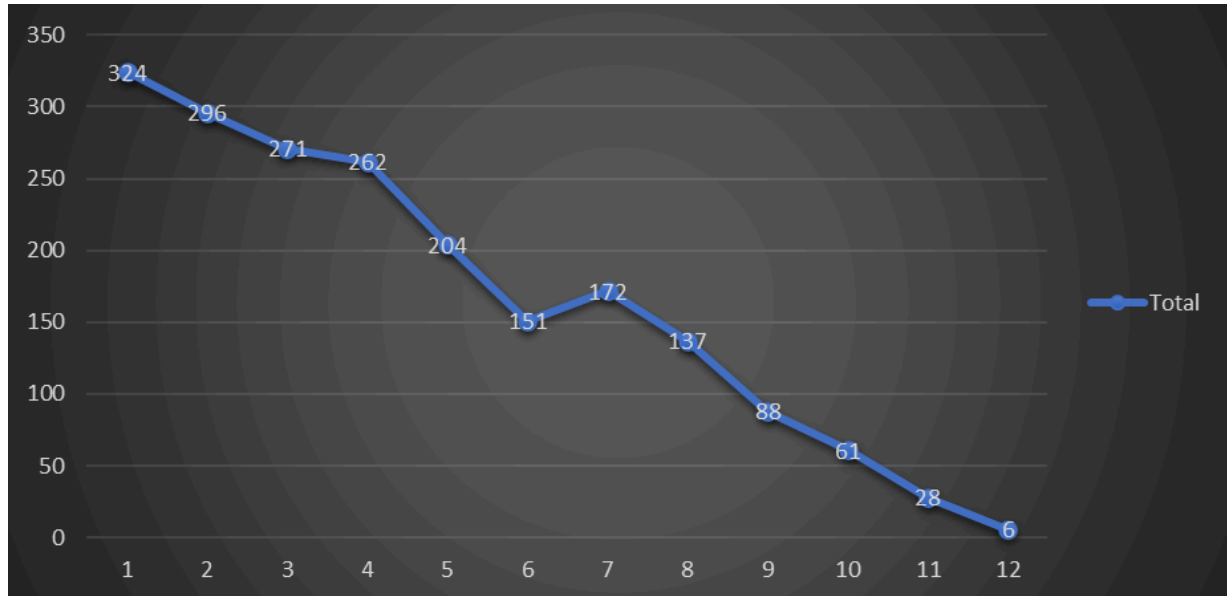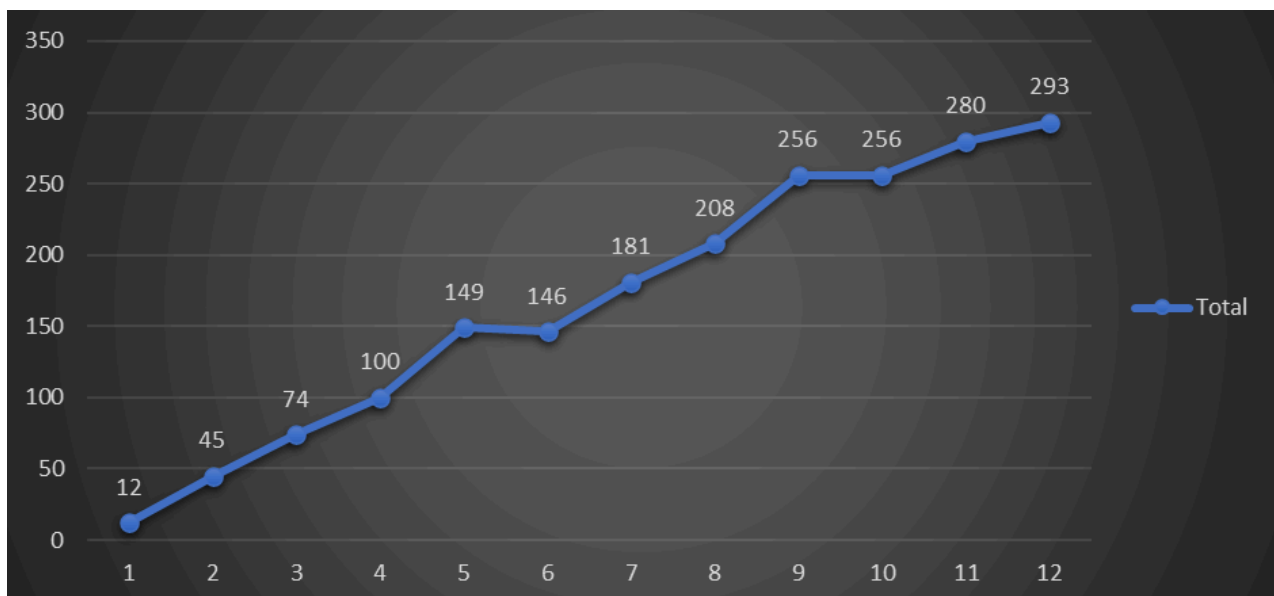
- **Province-wise Patient Count:**

All the provinces had comparable number of patients with Sindh having the maximum count (505).

- **Admission and Discharge Dates:**

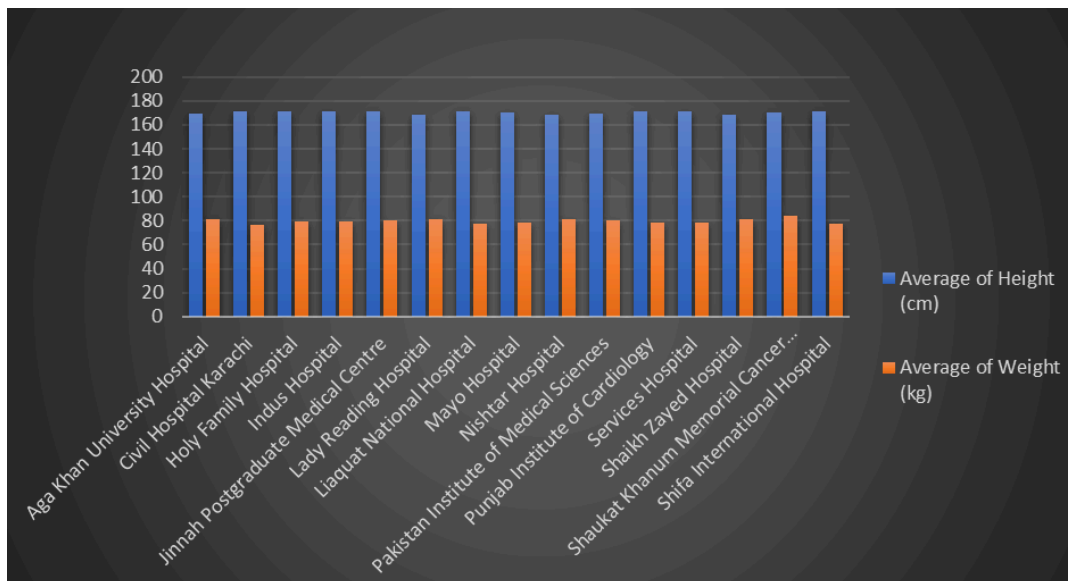On analyzing admission dates, the following trend was obtained

The first two months show peak admissions in hospitals. Afterwards, analysis of discharge dates resulted in
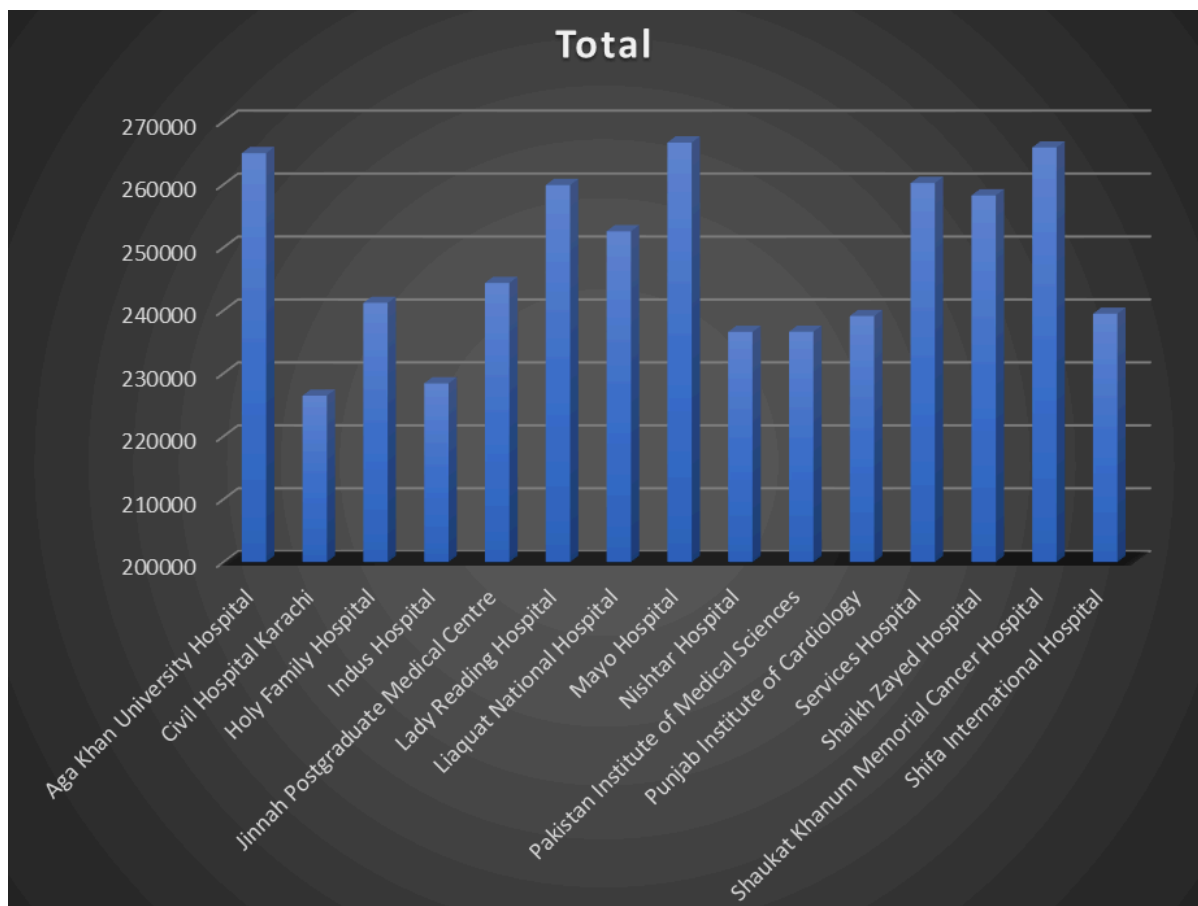

It was observed that the last two months show peak discharge in hospitals.

- **Height and Weight Statistics:**

The analysis of height and weight statistics produced the following insights
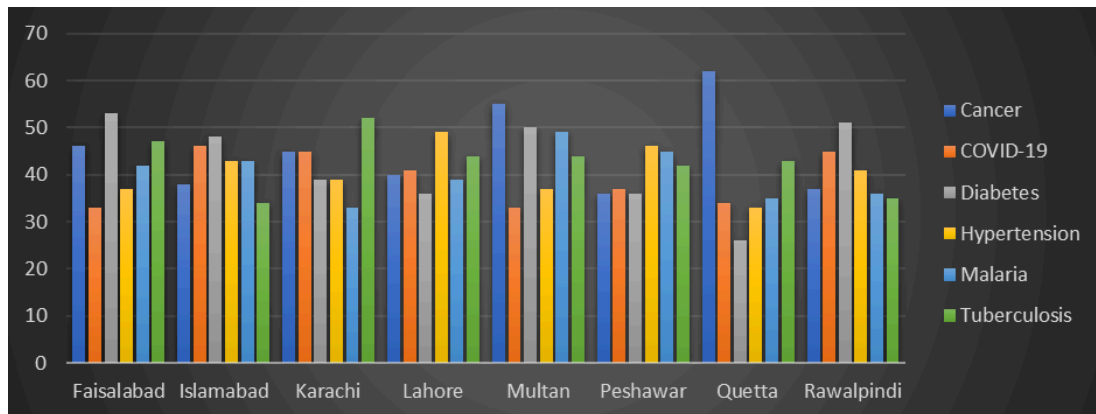

- **Treatment Cost by Hospital:**

When the avg treatment cost of each hospital was analyzed, it was observed that Mayo Hospital had the highest (266666) followed by Shaukat Khanum Memorial Cancer Hospital (265875) and Agha Khan University Hospital (264980).

- **Patient Age vs Treatment Cost:**

On calculating correlation coefficient between patient age and treatment cost using the =CORREL function, we obtained a value of -0.009975865 which indicated an extremely weak negative linear relationship between the two sets of data.

- **Disease by City:**

Diabetes was found to be the most common disease in 3 cities (Faisalabad, Islamabad and Rawalpindi). Hypertension in 2 cities (Lahore and Peshawar). Cancer in 2 cities (Multan and Quetta). Tuberculosis in 1 city (Karachi).

- **Hospital Performance:**
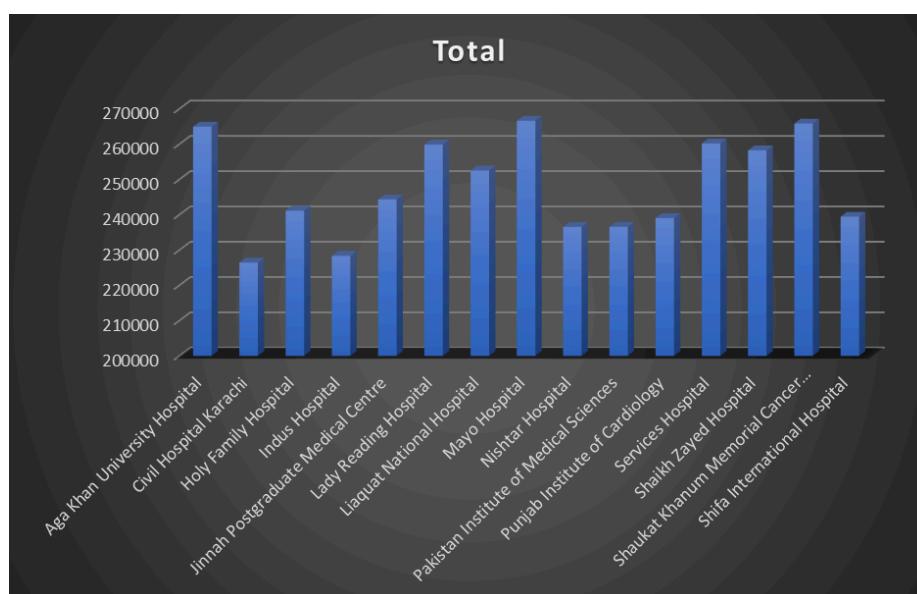
When the avg treatment cost of each hospital was analyzed, it was observed that Mayo Hospital had the highest (266666) whereas Civil Hospital Karachi had the lowest (226452).