INTRODUCTION TO DATA MANAGEMENT – INT217

(Project Semester August-December 2022)

HEALTHCARE COST ANALYSIS PROJECT

Submitted by

Gibran Khan Tareen

Registration No: **12100173**

Programme: **P132-L (BTech CSE)**

Section: **K20BG**

Course Code: **INT217**

Under the Guidance of

Veerpal Mam (UID: 25909)

Assistant Professor

Discipline of CSE/IT

Lovely School of Computer Science & Engineering

Lovely Professional University, Phagwara

Declaration

I hereby declare that I have completed my project on the topic **HEALTHCARE COST ANALYSIS.** I declare that I have worked with full dedication for this project and it is part of my course **INT217 – Introduction to Data Management** and for completion of my degree of **Bachelor of Technology (Computer Science & Engineering),** Lovely Professional University, Phagwara.

(Signature of Student)
Gibran Khan Tareen
Registration Number: 12100173

Date: 06 November 2022

Acknowledgement

I had a great experience learning in this project and even got to learn plenty of new skills. First, I would like to thank Almighty Allah who made it possible for me to do anything. After that I would like to thank my Late Grandfather- Mohd Nayeem Khan, my Gradma, my Parents, my little sister, and my maternal Grandparents because without their kind support and help I would not have been able to complete my project. I would also like to express my sole gratitude to my teacher **Veerpal Mam** who gave me this golden opportunity to do this training. The project making helped me to learn how to do proper Research and I learned about many new things.

(Signature of Student)
Gibran Khan Tareen
Registration Number: 12100173

Date: 06 November 2022

Table of Contents

Introduction

1.1 Context

This project has been done as part of course INT217 – Introduction to Data Management of B.Tech CSE at Lovely Professional University.

1.2 Motivations

Since very start, I have been extremely interested in everything which is related to Data Science. Extracting information from huge junks of data, this has been one of my favourites. This summer training was a great occasion to give me the time to learn and confirm my interest for this field. The fact that I can use DS in the field of Healthcare Cost analysis by using data science concepts helped me more to enhance my interest in the field of Data Science. That's why I decided to make my project around this topic of Data Science.

1.3 Idea

Back in Kashmir, few relatives of mine run a hospital named Khanam's Hospital. Many a times when I went to their home, I often found them debating on hospital costs, which age groups has most patients and many other things. These parameters helped them to efficiently track the revenue of the hospital and to analyse which wards (infants or adults) needs to be increased in numbers when they analyse the number of visits of the age groups of patients so that there is no deficit in number of wards available to patients. This was the moment at which the idea struck my mind that I should program such a Realtime system that completely analyses the data of the hospital with help of using the concepts of Data Science. The aim of my project was to develop such a system that helps all hospitals to analyse their data and costs so that they could easily understand their data and then do the necessary actions (like increase the number of beds for infant patients if frequency number of infant patients is high etc). So thus I chose to take this Healthcare Cost Analysis as my project.

Objectives of the Analysis

Topic: Healthcare Cost Analysis

Description:

1. Background and Objective: A nationwide survey of hospital costs conducted by the US Agency for Healthcare consists of hospital records of inpatient samples. The given data is restricted to the city of Wisconsin and relates to patients in the age group 0-17 years. The agency wants to analyse the data to research on healthcare costs and their utilization.

2. Domain: Cost Analysis in Healthcare

3. Goals to accomplish:

a) To Analyse Frequency of Patient visits based on all Age categories of people (to find the age group with maximum patients).

b) To Analyse total expenditures based on all Age categories of people and find age group which has maximum expenditure.

c) To analyse and find the diagnosis-related group that has maximum hospitalization (patient visits).

d) To analyse and find the diagnosis-related group that has maximum expenditure.

e) To Analyze the severity of the hospital costs by age and gender (For analysing the proper allocation of resources)

Source of Dataset

The dataset which I have used throughout my project is taken this dataset from Kaggle. It's a free to use Dataset which anyone can use. This dataset is **already filtered** and is ready to be directly used. Link to the dataset is:
https://www.kaggle.com/datasets/ibnshagufta/hospital-costs-dataset

Dataset Description:

Attribute	Description
Age	Age of the patient discharged
Female	A binary variable that indicates if the patient is female
Los	Length of stay in days
Race	Race of the patient (specified numerically)
Totchg	Hospital discharge costs
Aprdrg	All Patient Refined Diagnosis Related Groups

Table 1. Dataset Description

ETL Process

General Information

All our data is taken via our Dataset. It was already filtered so we directly imported it as an external data source.

Step 1: Import the necessary data into Excel

The first thing to do is to bring data into Microsoft Excel. If the data doesn't pre-exist in Excel, then we will have to import it to our Excel.

There are multiple ways to do so:

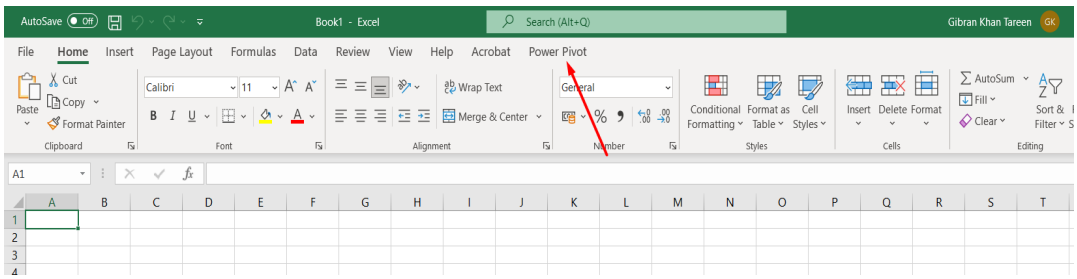
- Simply Copy and paste it to Excel
- Use any 3rd party API like Supermetrics or Open Database Connectivity etc
- Microsoft Power Query, which is an Excel add-in

The most suitable way will fully depend on your data file type, and you may have to research the best ways to import data into Excel.

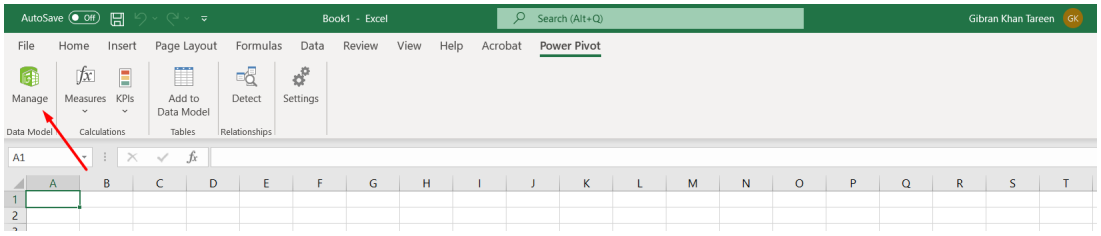
Step 2: Import the Dataset as "External Data Source" into Excel

Since our dataset already exist as an Excel file, we will import it as **"External Data Sources"** using PowerPivot steps.

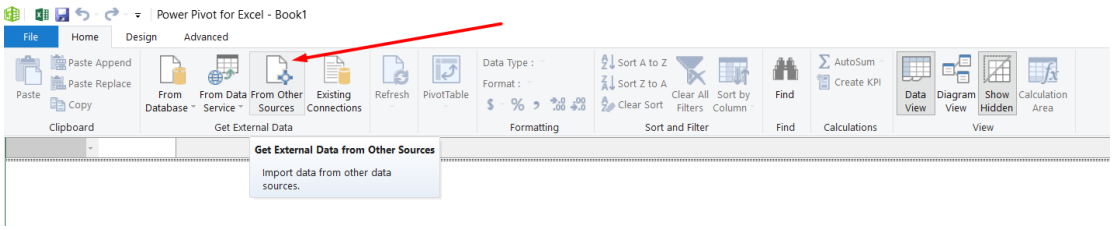
1. Click on PowerPivot



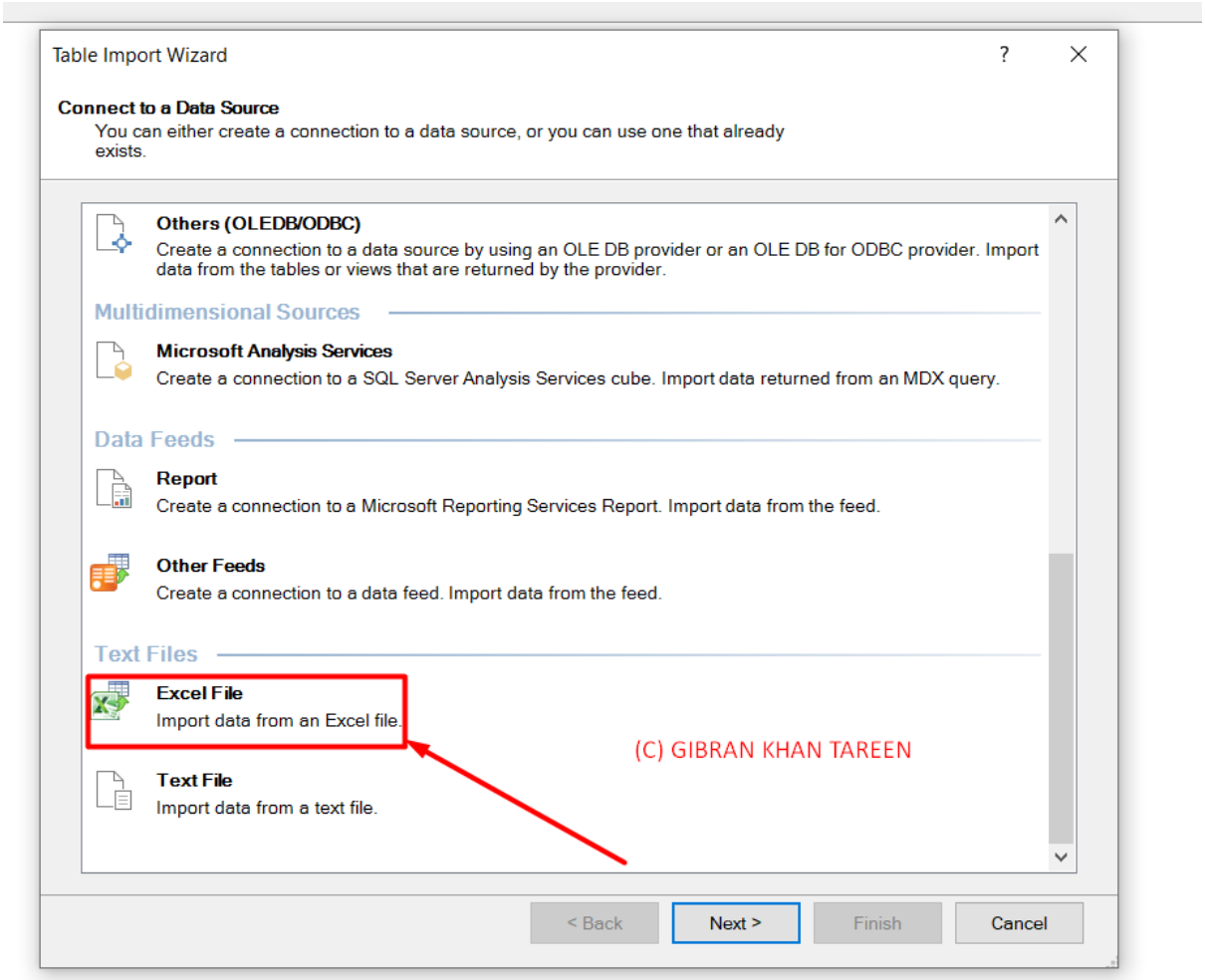
2. Now click on **Manage**



3. Now click on "**From Other Sources**" in ***GET EXTERNAL DATA*** section

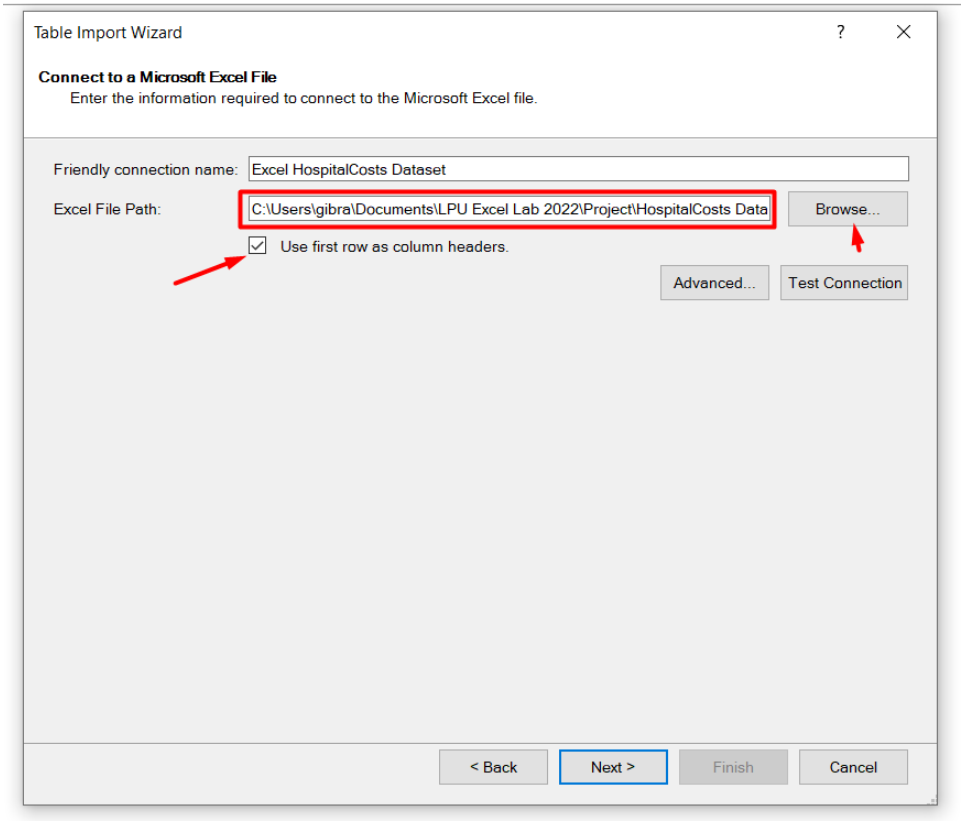


4. Now a new popup window will open. Select **Excel File,** then click Next

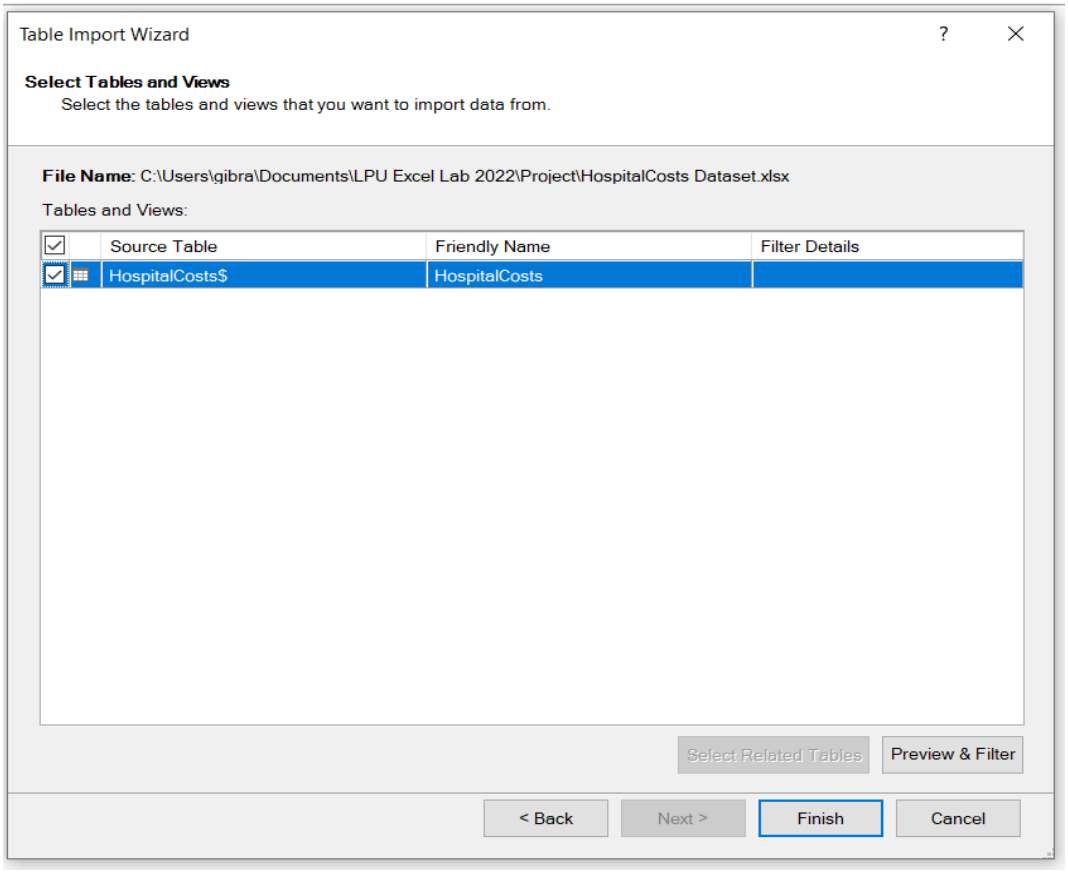


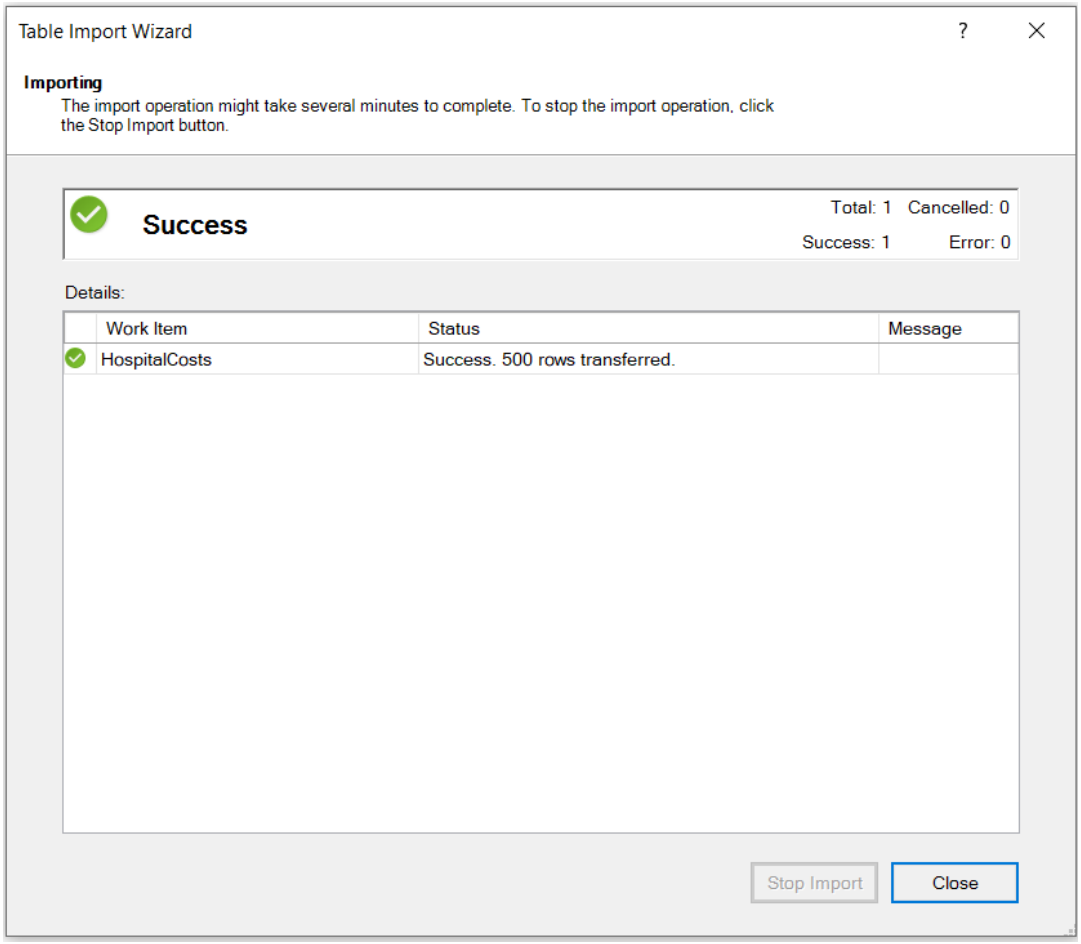
5. Now click on **Browse** and select your Dataset file. After selecting click on **Use First row as column headers.** Then click Next

(Please Turn Over)

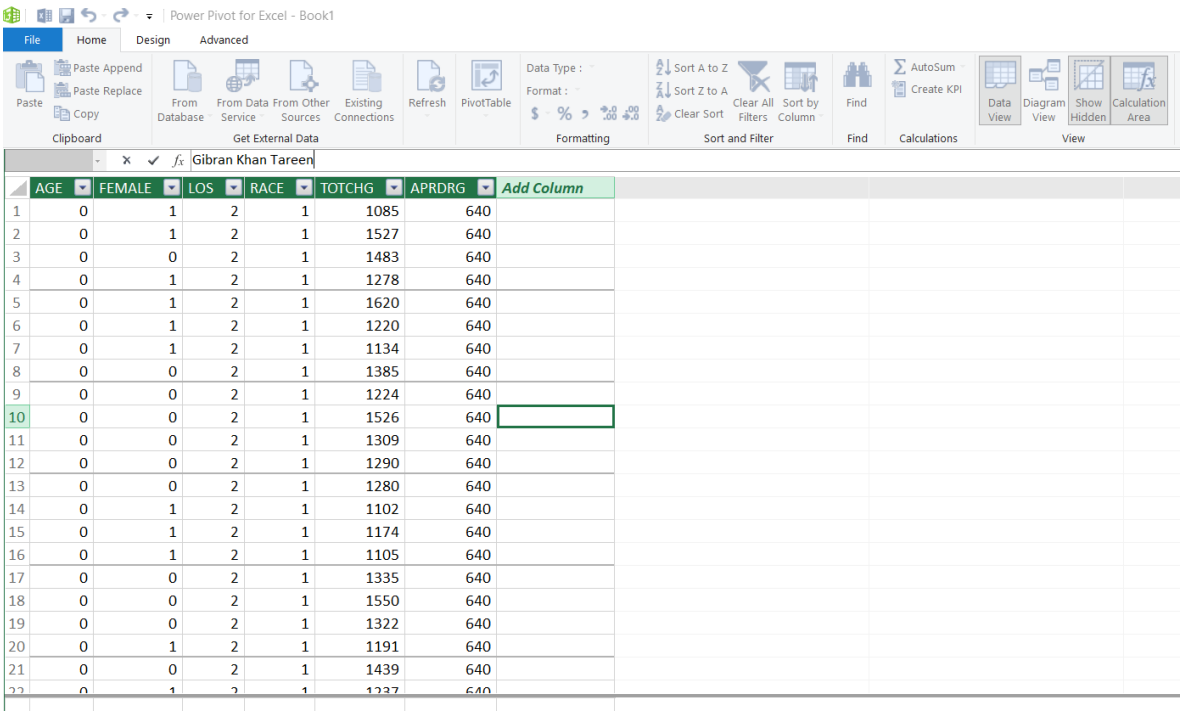


6. After import, Now click on **Finish,** then **Close**



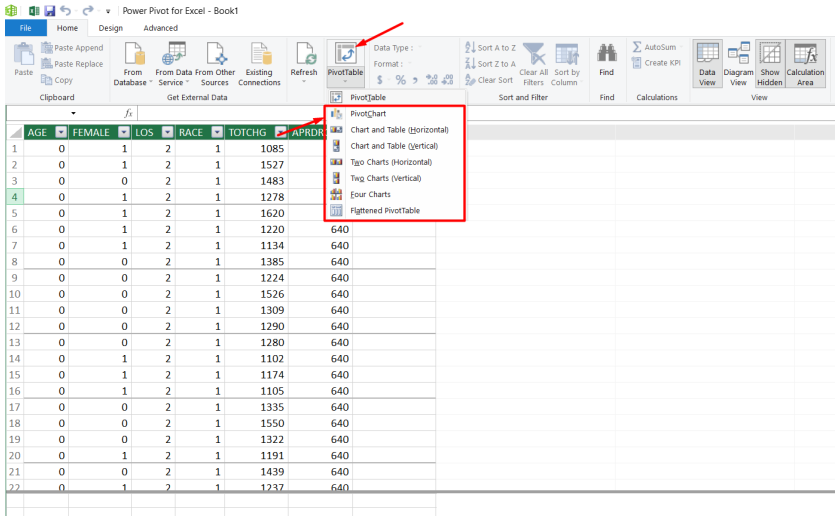


7. After that our data will be successfully imported in Excel

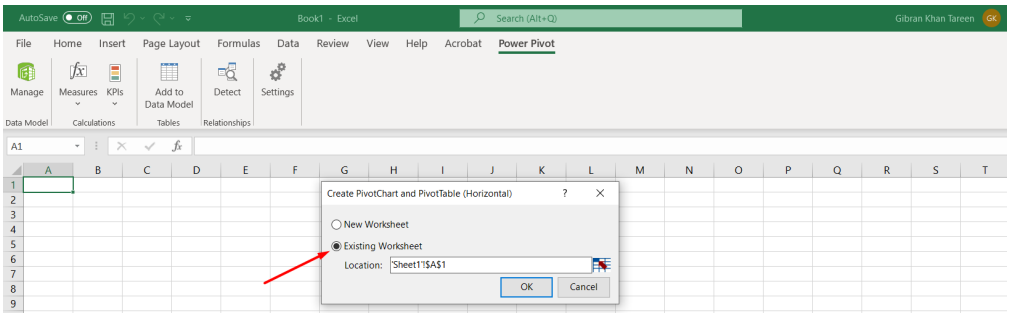


	AGE	FEMALE	LOS	RACE	TOTCHG	APRDRG	Add Column
1	0	1	2	1	1085	640	
2	0	1	2	1	1527	640	
3	0	0	2	1	1483	640	
4	0	1	2	1	1278	640	
5	0	1	2	1	1620	640	
6	0	1	2	1	1220	640	
7	0	1	2	1	1134	640	
8	0	0	2	1	1385	640	
9	0	0	2	1	1224	640	
10	0	0	2	1	1526	640	
11	0	0	2	1	1309	640	
12	0	0	2	1	1290	640	
13	0	0	2	1	1280	640	
14	0	1	2	1	1102	640	
15	0	1	2	1	1174	640	
16	0	1	2	1	1105	640	
17	0	0	2	1	1335	640	
18	0	0	2	1	1550	640	
19	0	0	2	1	1322	640	
20	0	1	2	1	1191	640	
21	0	0	2	1	1439	640	

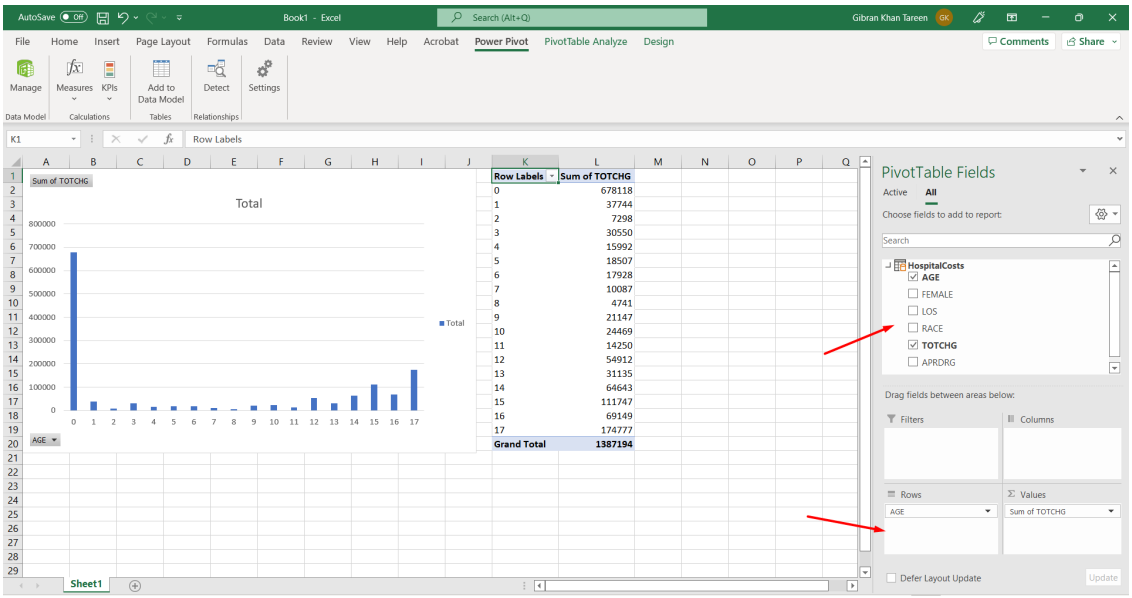
8. Now click on **Pivot Table,** then select the required option. We will go for **Chart and Table**.



9. Now select where you want your chart and table to be placed. I went with Existing



10. Now by selecting the desired values for pivot table and pivot chart filters, you will get your desired charts for analysis



Analysis on Dataset

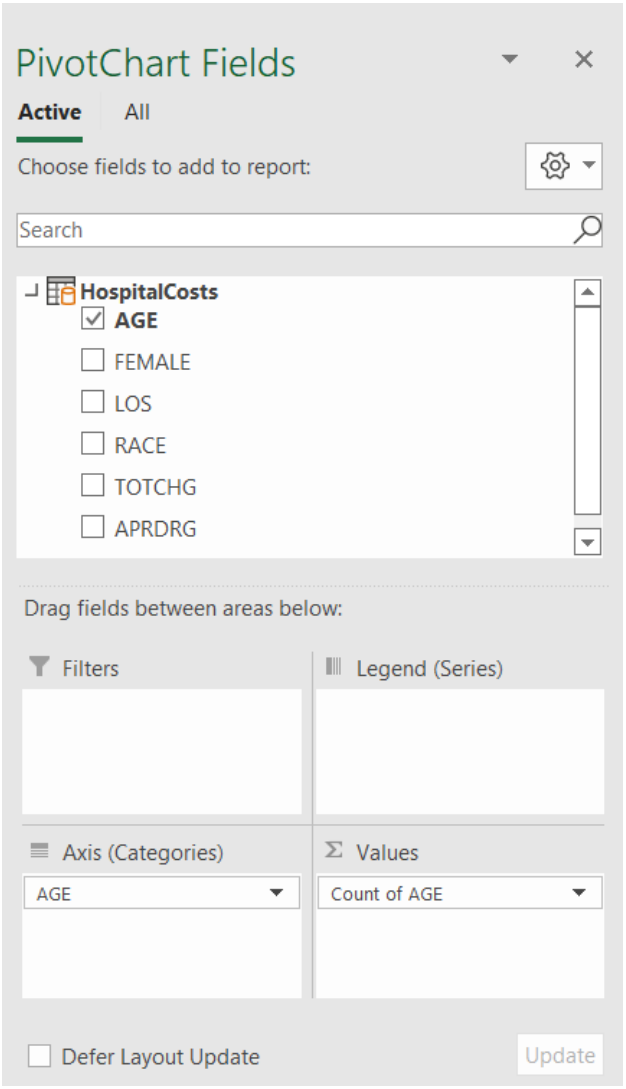
Goal 1

Introduction:

In our Goal #1, we need to analyse frequency of patient visits based on all Age categories of people (To record the patient statistics).

Specific Requirements:

To accomplish our goal, we need to import the dataset as “external source” file and make Pivot table and then thus make Pivot chart from it by keeping AGE as Category and SUM OF TOTCH as values.



Analysis Results:

We find that **Age Group 0** has highest number of patient visits of 307

Visualization:

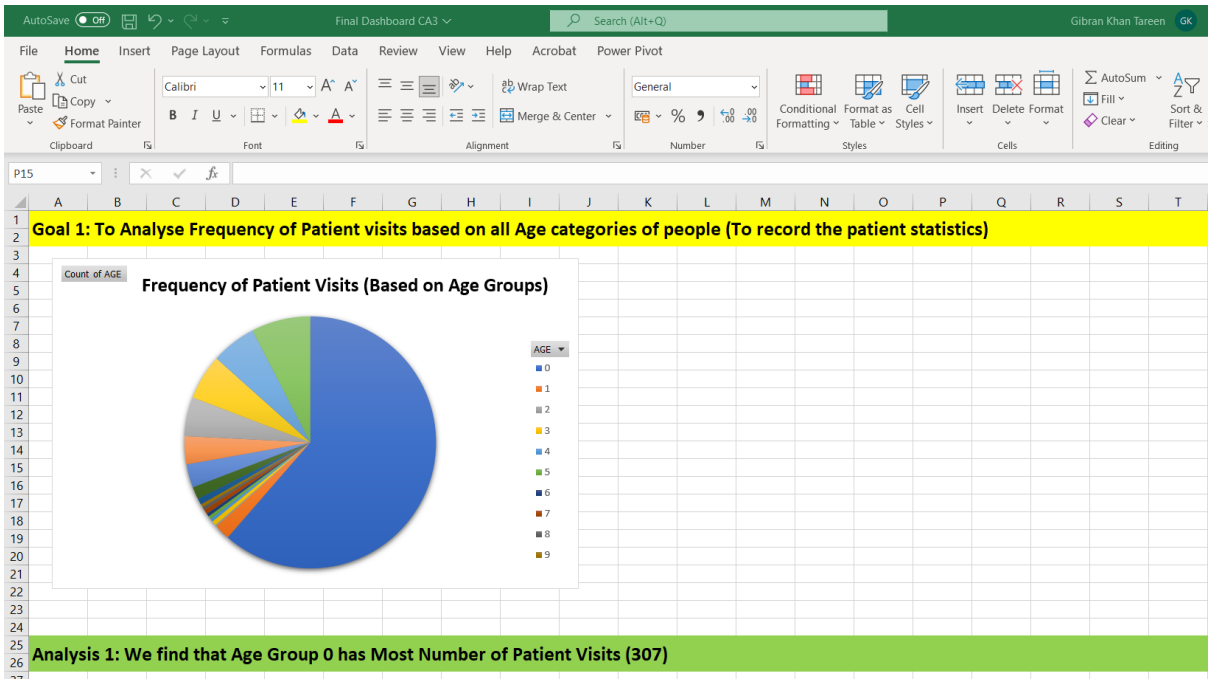


Figure 1: Result of Analysis 1

(Please Turn Over)

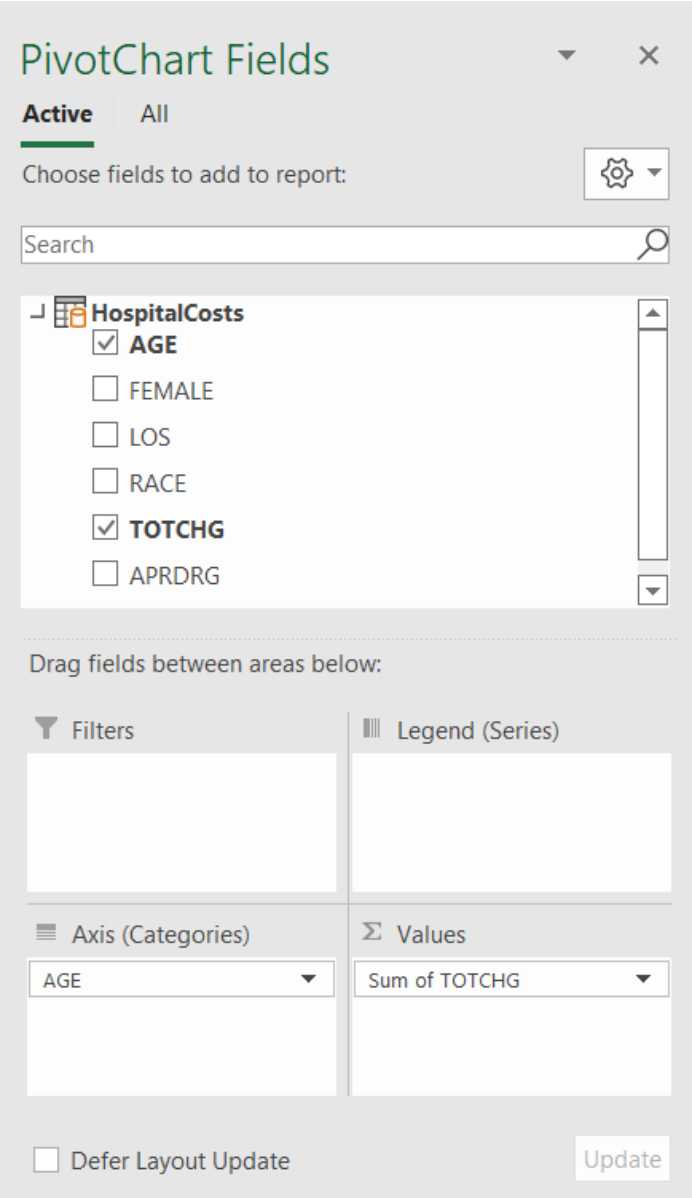
Analysis for Goal 2

Introduction:

In our Goal #2, we need to analyse total expenditures based on all Age categories of people (To record the patient statistics)

Specific Requirements:

To accomplish our goal, we need to use the same imported dataset and make Pivot table and then make Pivot chart from it by keeping AGE as Category and SUM OF TOTCH as values.



Analysis Results:

- We find out that the age group that has maximum expenditure is **Group 0** having total expenditure of 678118.

Visualization:

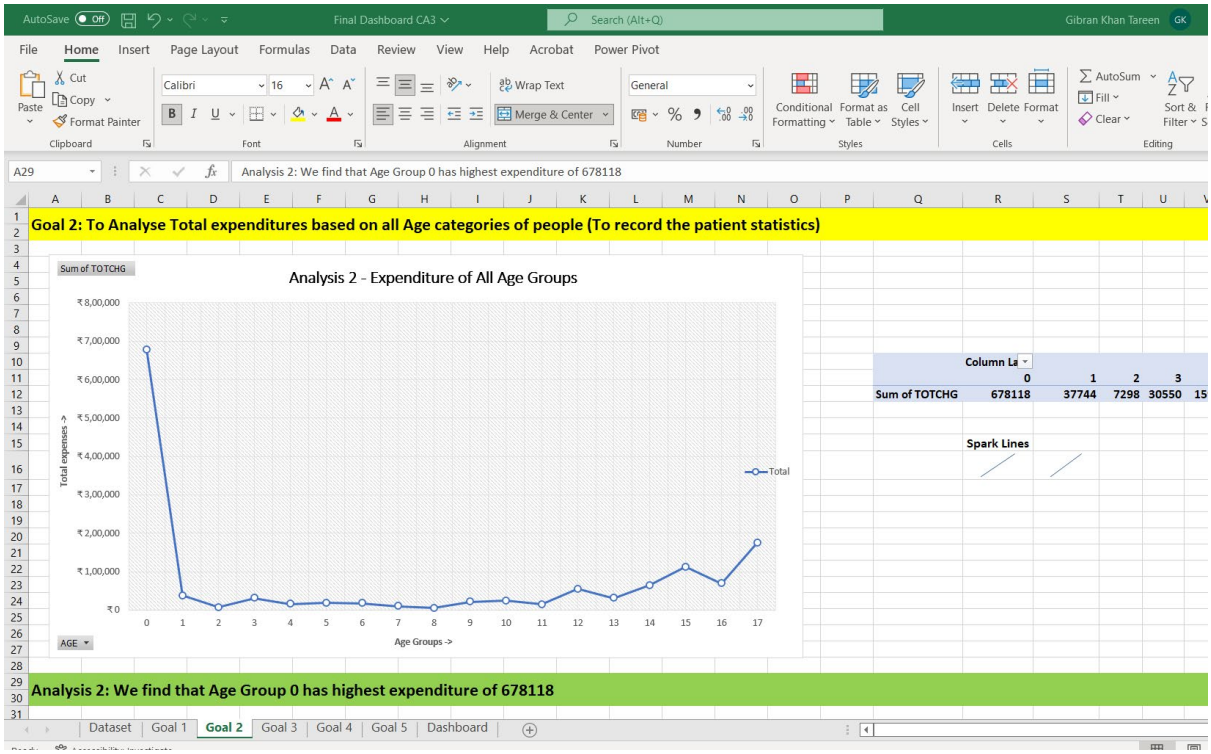


Figure 2: Result of Analysis 2

(Please Turn Over)

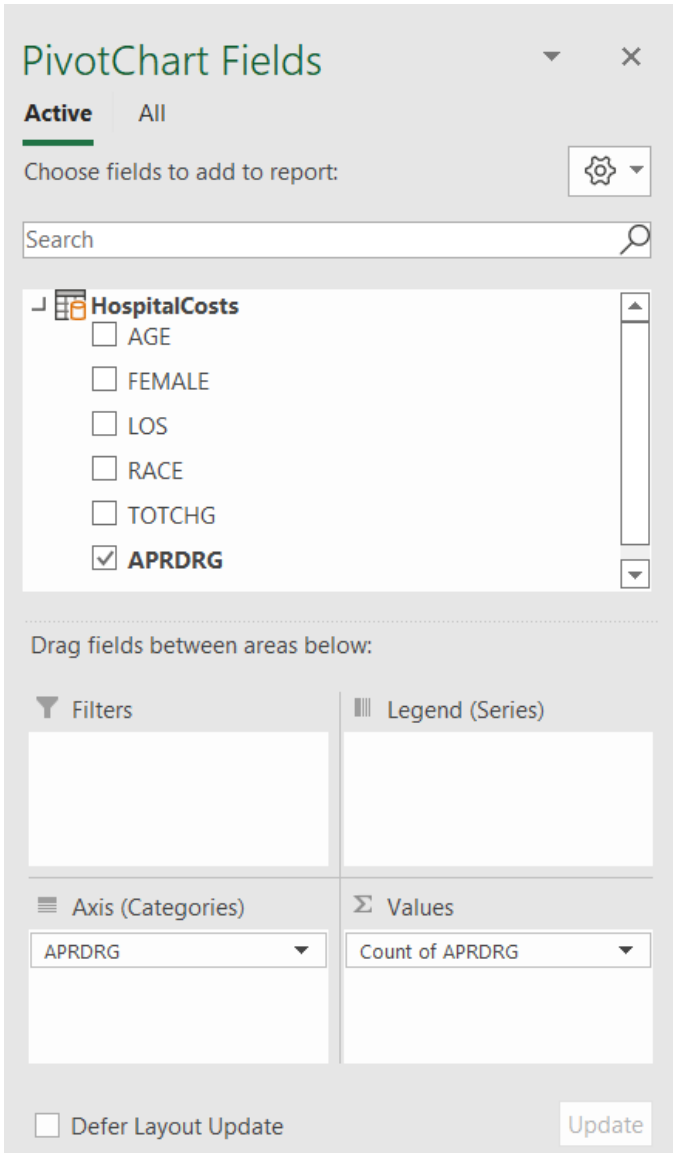
Analysis for Goal 3

Introduction:

In our Goal #3, we need to analyse the diagnosis-related group that has maximum hospitalization (patient visits).

Specific Requirements:

To accomplish our goal, we need to use the same imported dataset and make Pivot table and then make Pivot chart from it by keeping **APDRG** as Category and **COUNT OF APDRG** as values.



Analysis Results:

- We find out that the diagnosis related group that has maximum patient visits is the **Group 640** having total 267 visits.

Visualization:

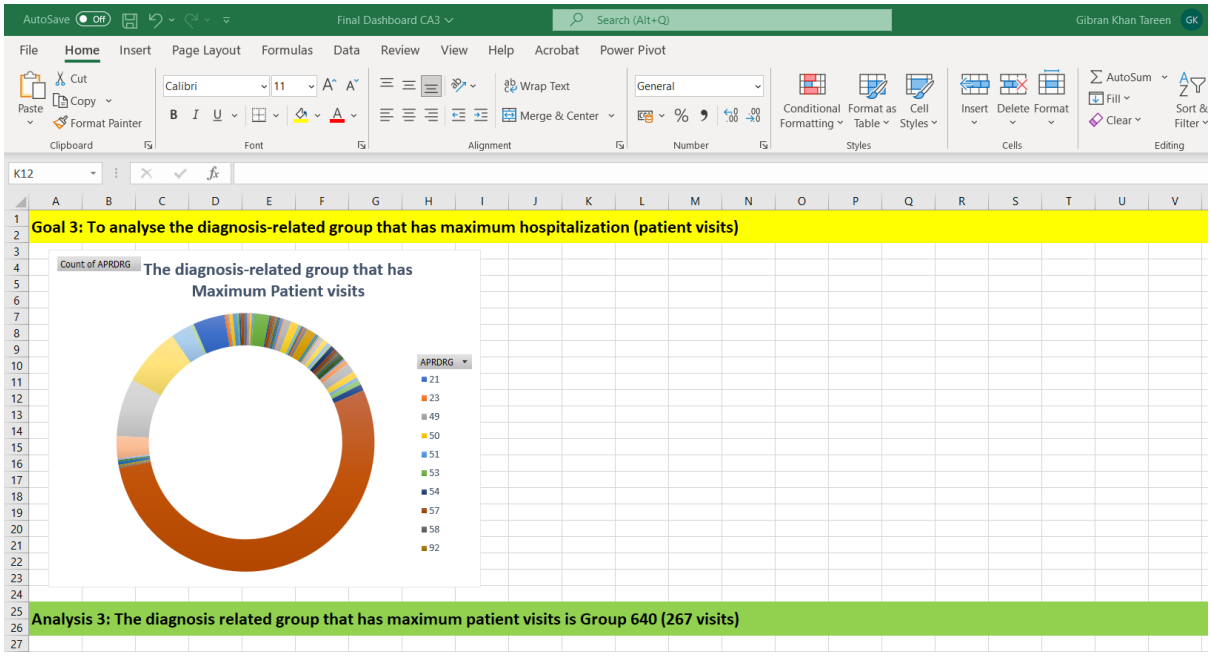


Figure 3: Result of Analysis 3

(Please Turn Over)

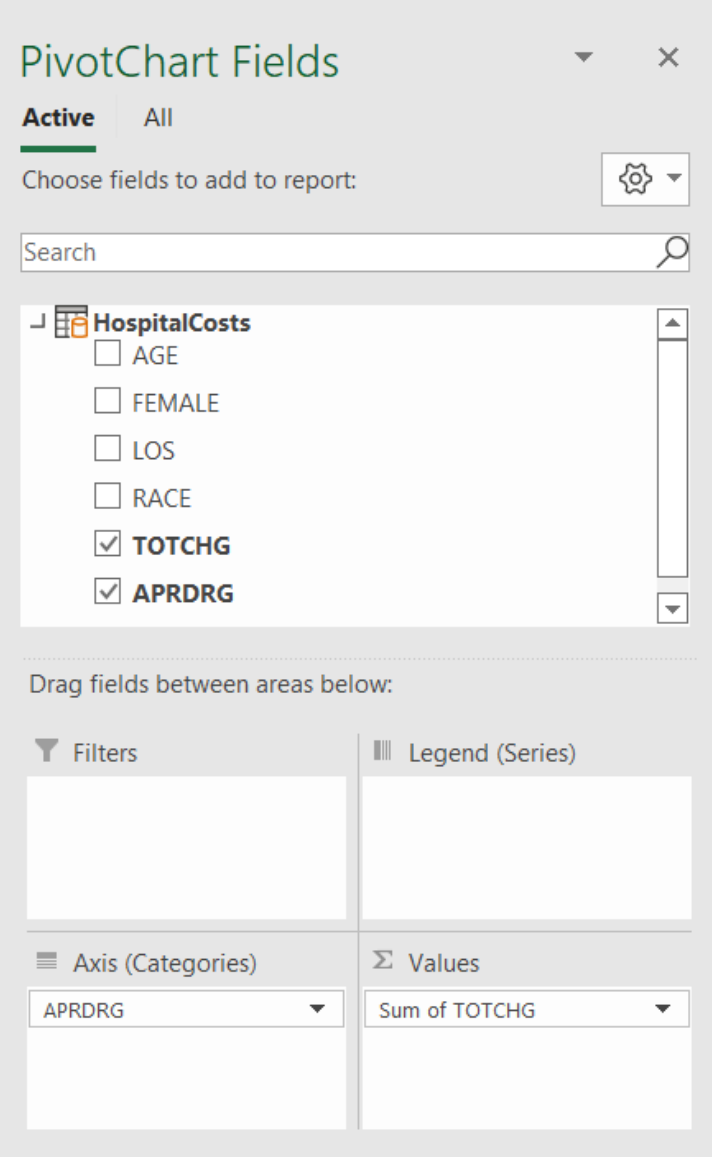
Analysis for Goal 4

Introduction:

In our Goal #4, we need to analyse the diagnosis-related group that has maximum expenditure

Specific Requirements:

To accomplish our goal, we need to use the same imported dataset and make Pivot table and then make Pivot chart from it by.



Analysis Results:

- The diagnosis-related group with Maximum expenditure is **Group 640** (having the expenditure as 437978).

Visualization:

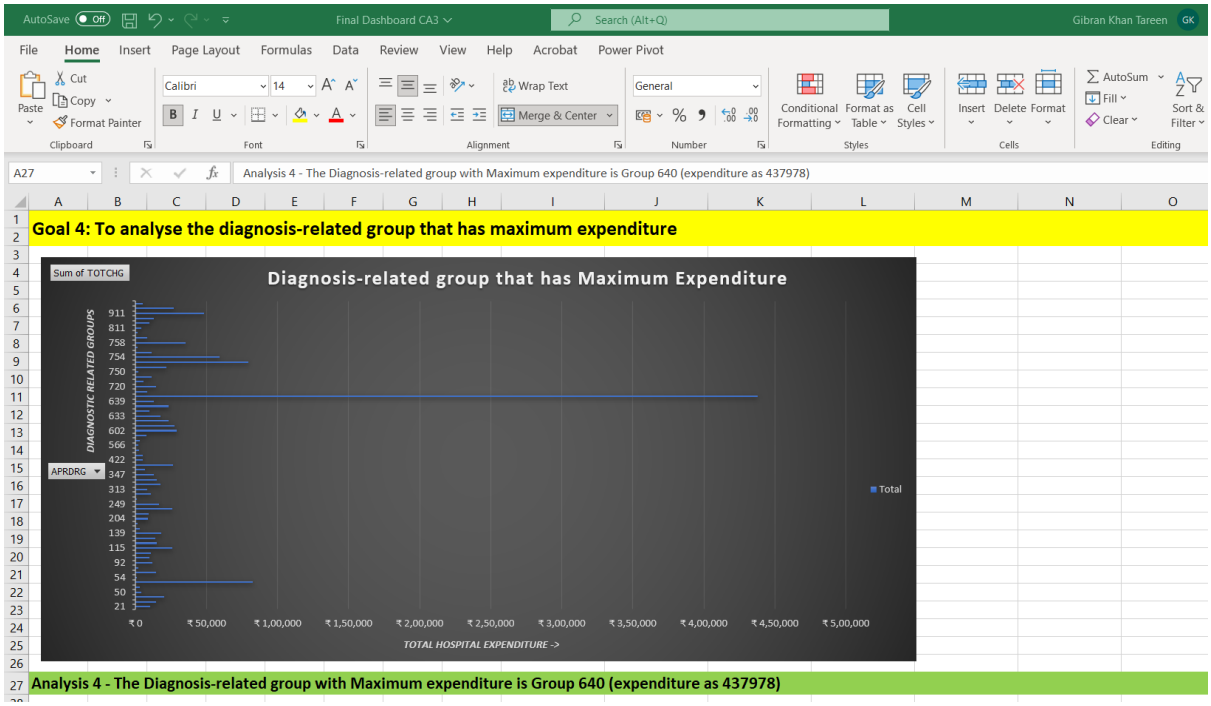


Figure 4: Result of Analysis 4

(Please Turn Over)

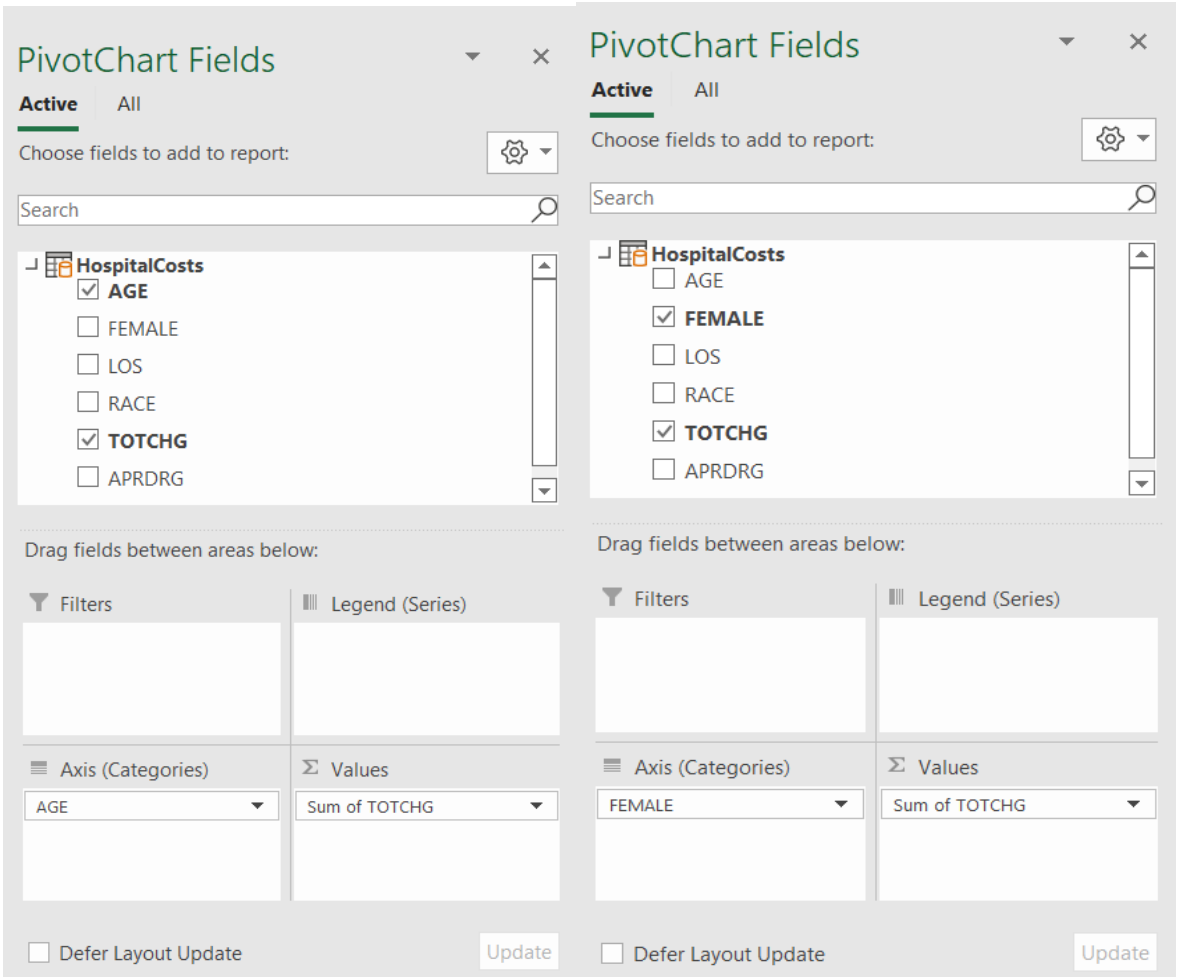
Analysis for Goal 5

Introduction:

In our Goal #4, we need to analyse the severity of the hospital costs by age and gender (to analyse the hospital's data for the proper allocation of resources).

Specific Requirements:

To accomplish our goal, we need to use the same imported dataset and make Pivot table and then make Pivot chart from it. Then we need to create slicer and connect both the charts in it.



Left Pivot table for Chart 1 & Right pivot table for Chart 2

(Please Turn Over)

Analysis Results:

- On an average, we analyse that Females incur lesser hospital costs than males.
- On analysing we can see that AGE has more impact on severity of hospital costs than gender.

Visualization:

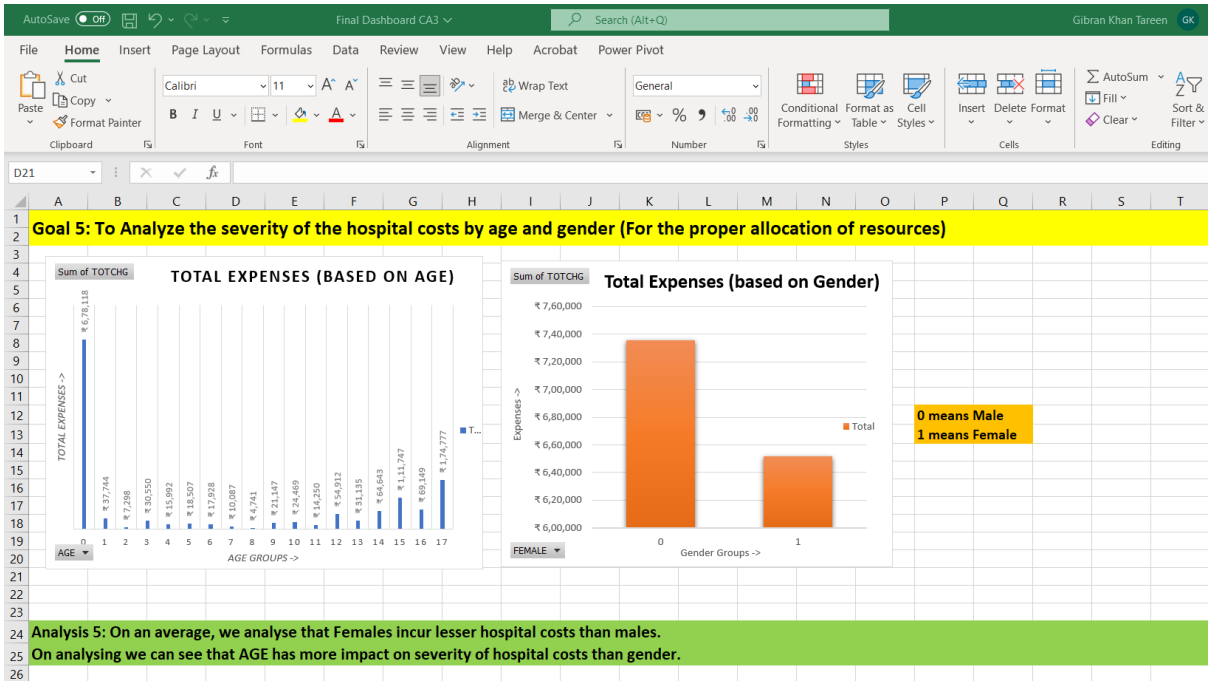


Figure 5: Result of Analysis 5

List of Analysis with Results

Analysis 1	Result: We find that Age Group 0 has highest expenditure of 678118	**Page No. 15**
Analysis 2	Result: Age Group 0 has Most Number of Patient Visits (307)	**Page No. 17**
Analysis 3	Result: 1. The diagnosis related group that has maximum patient visits is Group 640 having total 267 visits	**Page No. 19**
Analysis 4	Result: 2. The Diagnosis-related group with Max expenditure is Group 640 having total expenditure as 437978	**Page No. 21**
Analysis 5	Result: 1. On an average, we analyse that Females incur lesser hospital costs than males. 2. On analysing we can see that AGE has more impact on severity of hospital costs than gender.	**Page No. 23**

Screenshots of Dashboard

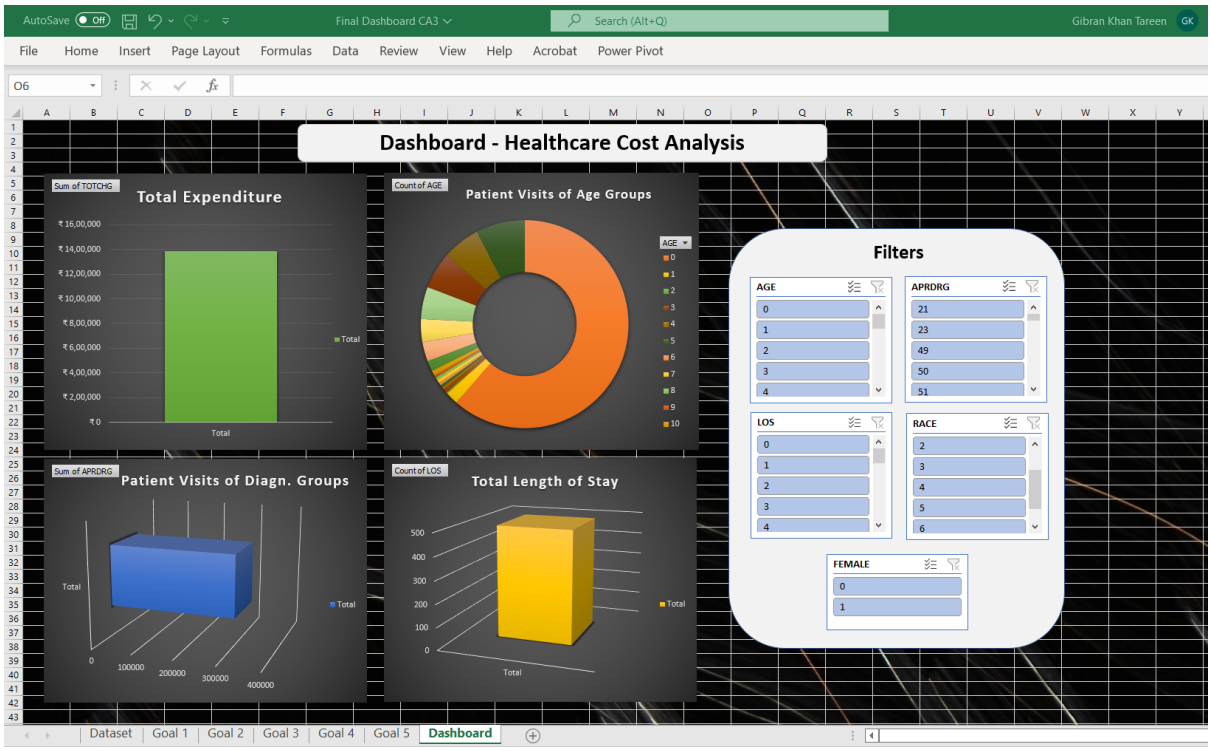


Figure: Complete Dashboard

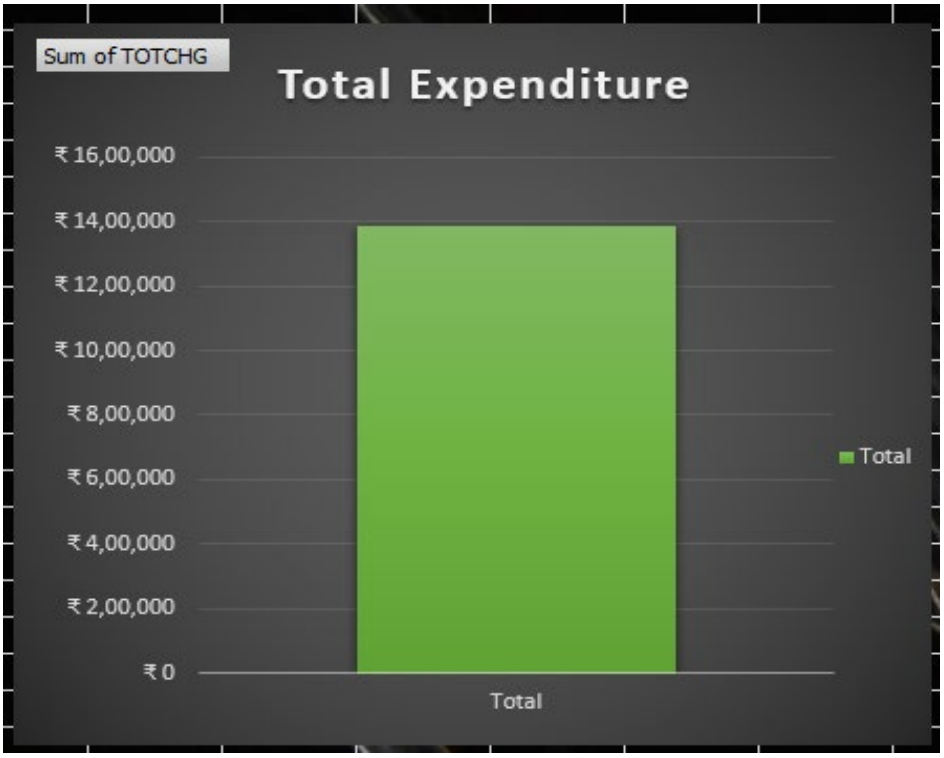


Chart 1: Total Expenditure

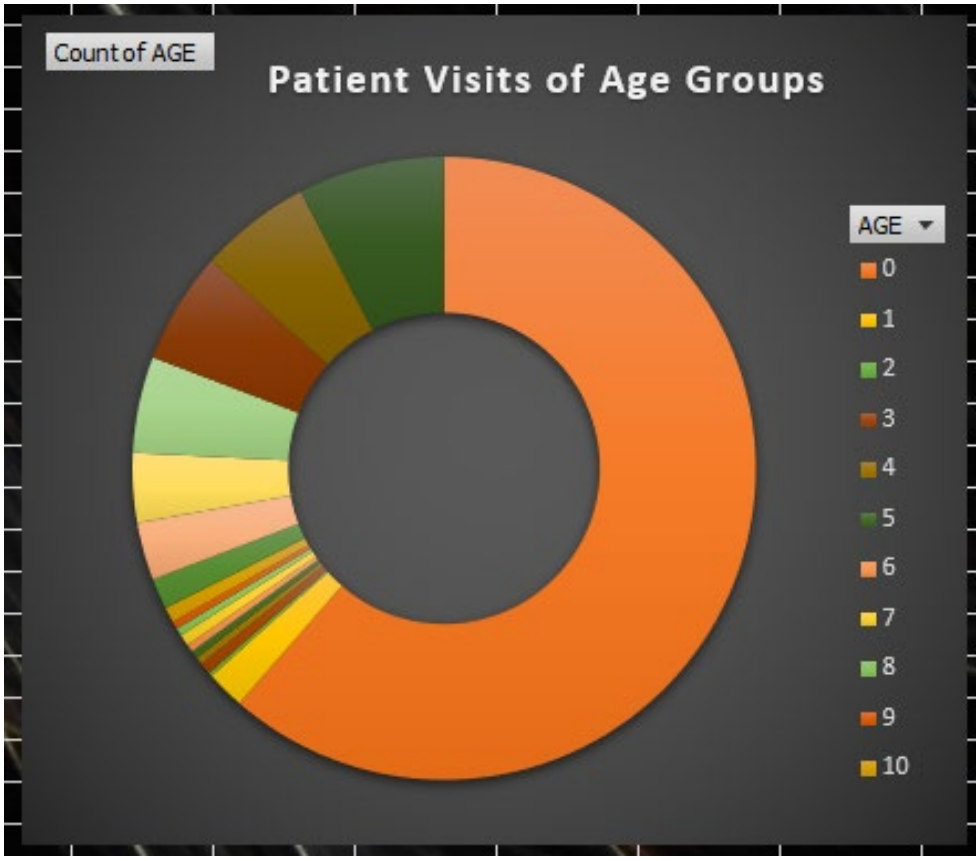


Chart 2: Patient Visits of Age Groups

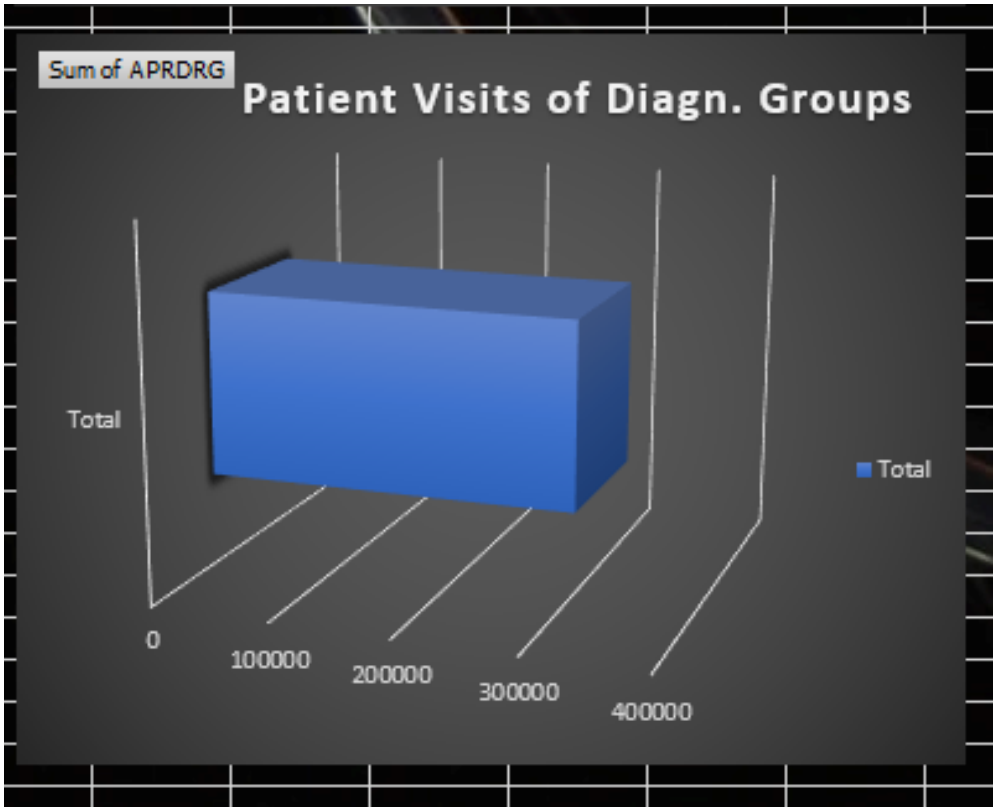


Chart 3: Patient Visits of Gender Groups

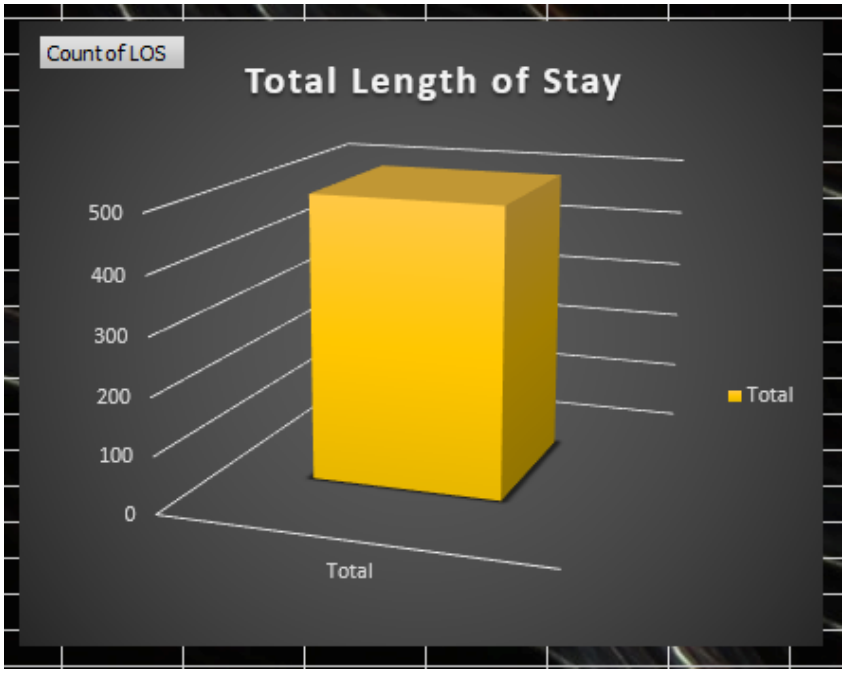


Chart 4: Length of Stay

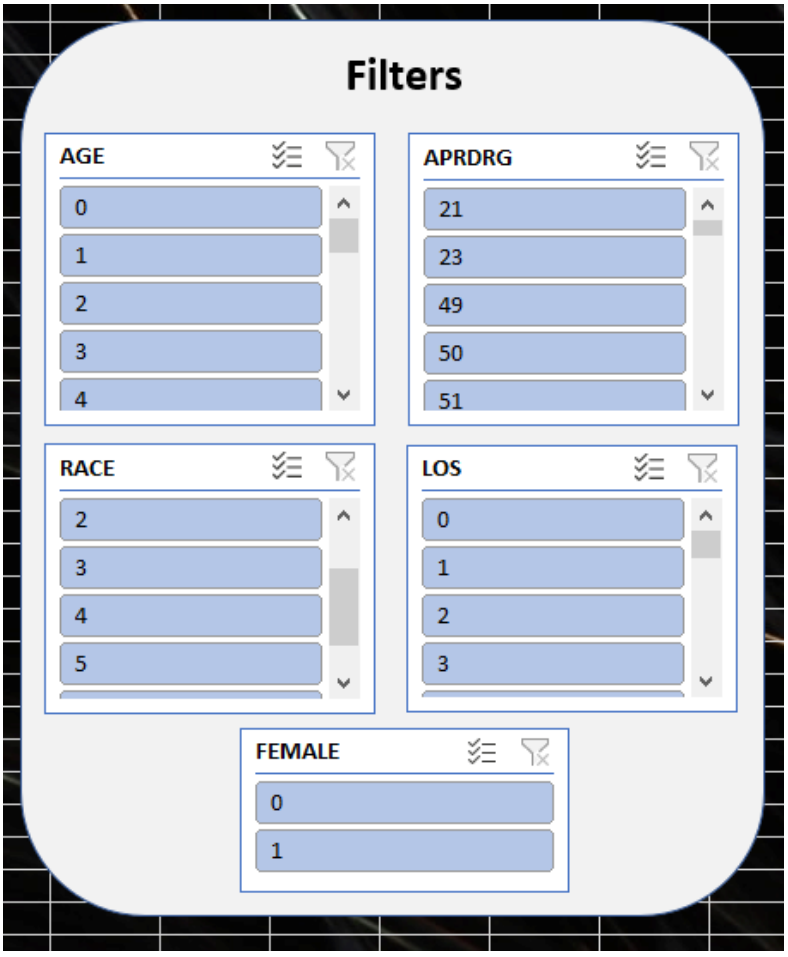


Figure: Filters used in the Dashboard

References

1. https://support.microsoft.com/en-us/office/create-and-share-a-dashboard-with-excel-and-microsoft-groups-ad92a34d-38d0-4fdd-b8b1-58379aae746e/

2. https://www.youtube.com/watch?v=aC009Px4tEg

3. https://www.youtube.com/watch?v=Xg8_iSkJpAE

4. https://www.analyticsvidhya.com/blog/2021/11/a-comprehensive-guide-on-microsoft-excel-for-data-analysis/

5. https://www.youtube.com/watch?v=n-y432o5C9o

6. https://www.youtube.com/watch?v=7K_KqD1S2M0

Bibliography

1. Hadley Wickham and Garrett Gorlemund, R for Data Science (Latest Edition), **2017**, pp. 345–372.

2. John A. Rice, Mathematical Statistics and Data Analysis (3rd Edition), **2007**, pp. 329–420.