MedTourEasy

Traineeship Program 2024

Final Project Report

Visualization in Tableau

By-

Prachi Nandkumar Wathore

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to all those who have contributed to the successful completion of this project on "Data Visualisation Project in Tableau - US Patients." This endeavour would not have been possible without the support, guidance, and cooperation of several individuals and organizations.

First and foremost, I extend my deepest appreciation to the MedTourEasy team who generously provided access to the necessary data for this analysis. Their commitment to advancing healthcare has been instrumental in shaping the insights presented in this report.

I am profoundly thankful to my project supervisor and mentor, Mr. Ankit Hasija, for his valuable guidance, constructive feedback, and continuous support throughout the project.

Additionally, I express my gratitude to the authors of the various libraries, frameworks, and tools used in the data analysis process. The open-source community's commitment to advancing data science and analytics has significantly enhanced the efficiency and robustness of our methodology.

Lastly, I want to thank my family and friends for their unwavering support and encouragement throughout this project. Their understanding and encouragement have been a source of motivation during challenging phases.

In conclusion, this project has been a collective effort, and I am deeply appreciative of everyone who played a role, directly or indirectly, in its completion. Thank you for being a part of this journey.

TABLE OF CONTENT

ABSTRACT

The project focuses on a comprehensive analysis of US hospital performance, delving into key performance indicators such as sales, profit, discount, days of operation and the range of medical procedures conducted.

In the first phase, extensive data collection is undertaken from a diverse array of healthcare institutions across the United States, encompassing both public and private sectors. The dataset is meticulously cleaned and pre-processed to ensure accuracy and reliability, laying the foundation for robust analytical insights.

The second phase involves a detailed exploration of sales and profit dynamics within the sampled hospitals.

The project aims to identify factors influencing revenue generation and profitability, providing valuable benchmarks for hospital administrators and stakeholders. Additionally, this analysis seeks to unveil patterns in the relationship between hospital financial performance and various demographic and economic indicators.

In the third phase, the project scrutinizes the spectrum of medical procedures performed by hospitals. The goal is to discern correlations between the diversity of procedures and financial performance, offering insights into the strategic allocation of resources and potential areas for revenue growth.

The final phase of the project involves the synthesis of findings and the formulation of actionable recommendations. Hospital administrators, policymakers, and healthcare stakeholders can leverage these insights to make informed decisions, optimize resource allocation, and enhance overall hospital performance. The outcomes of this project contribute not only to the understanding of the financial dynamics within US hospitals but also to the broader discourse on improving healthcare delivery and sustainability.

INTRODUCTION

I. ABOUT THE COMPANY

MedTourEasy, a global healthcare company, provides you the informational resources needed to evaluate your global options. MedTourEasy provides analytical solutions to our partner healthcare providers globally. It helps you find the right healthcare solution based on specific health needs, affordable care while meeting the quality standards that you expect to have in healthcare.

MedTourEasy improves access to healthcare for people everywhere. It is an easy to use Platform and service that helps patients to get medical second opinions and to schedule affordable, high-quality medical treatment abroad.

II. PROJECT

The project embarks on a comprehensive exploration of the operational and financial aspects of US hospitals. The focal points of our inquiry are the key financial parameters including sales, profits, discount structures, and the operational metric of the number of days of hospital activity. Additionally, we endeavour to delve into the extensive array of medical procedures conducted by these institutions and discern the interrelations among these multifaceted factors.

This endeavour entails a systematic collection of data from a diverse spectrum of hospitals throughout the United States. Through meticulous analysis and visualization techniques, we seek to unravel the complex interactions between sales, profits, operational days, discounts, and the array of medical procedures. By presenting our findings through insightful charts and graphs, we aim to provide a nuanced understanding of the dynamics governing the performance of US hospitals.

The project is akin to assembling a mosaic, wherein each financial parameter and operational metric contributes to the comprehensive narrative of hospital performance. The objective is not only to observe these metrics in isolation but to discern the intricate connections that inform strategic decision-making within the healthcare sector. In essence, we embark on a journey to bring forth a holistic perspective that aids in fostering informed and prudent decisions for the benefit of the healthcare landscape in the United States.

III. KEY REQUIREMENTS AND DELIVERABLES

1. General Requirements
    - Dashboard size is 1250px wide by 750px tall.
    - The dashboard has a total of 5 containers (no more, no less)
    - The Filter Pane: Each filter has some padding
2. Business Requirements
    - Show four filters- Category, Sub-Category, Region, and Segment. These filters should have

- only relevant values.
- The dashboard should have the title “Executive sales”.
- The first chart should have the title “YTS KPIs” and should show the following – Total Discount, Overall Profit, Total Quantity and Total Sales.
- The second graph should have the title as “Sales” and should show monthly sales per year.
- The third graph should the title as “Profit” and should show monthly profit per year.

3. Create a Dashboard in Tableau
4. Present a final report highlighting:
    - Key problem statement
    - Entire process of project implementation
    - Data filtering and cleaning
    - Exploratory data analysis
    - Create insightful visuals helpful in drawing key information to drive decision in right directions.

IV. ASSUMPTIONS

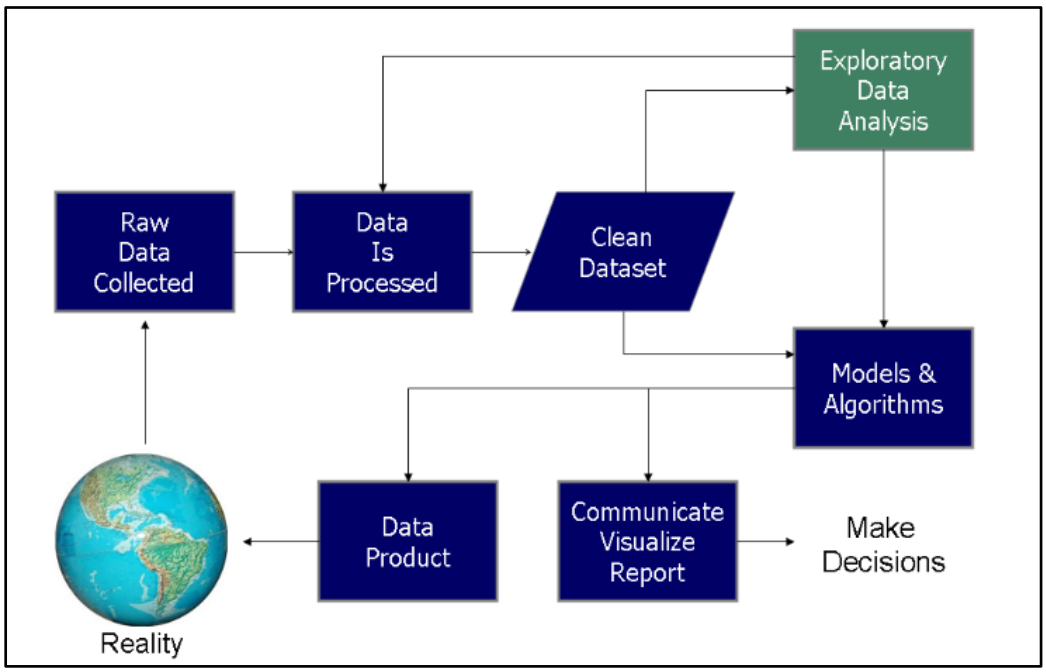
An assumption is an underlying belief or presupposition that is considered to be true, often without explicit evidence or verification. Assumptions are frequently necessary when dealing with incomplete information, helping to simplify complex situations or scenarios. However, it is important to recognize that assumptions come with a degree of uncertainty and may not always align with reality. Towards the end of project, we can verify our assumptions or discard them altogether based on insights drawn from the observations and analysis.

Following are the assumptions made:

1. The period for which is collected, and collated and being used for analysis, there were no natural disasters, calamity or medical emergency of any kind which possess the potential of distorting the information and disturbing the trends, driving us away from key insights.
2. High number of days of operation of a hospital or medical facility implies high sales.
3. Offering high discounts results in increase in sales.

METHODOLOGY

1. FLOW OF PROJECT



In a data visualization and analysis project, the workflow typically follows a systematic and iterative process. The journey begins with defining the project's objectives and establishing clear research questions. This initial phase involves understanding the stakeholders' requirements and determining the key metrics or variables of interest. Once the objectives are defined, the next step is data collection, involving the acquisition and extraction of relevant datasets.

Following data collection, the project transitions into the data cleaning and pre-processing stage. This phase focuses on handling missing values, addressing outliers, and transforming data into a format suitable for analysis.

Subsequently, the heart of the project lies in exploratory data analysis (EDA), where statistical methods and visualization techniques are employed to uncover patterns, trends, and outliers in the data. The insights gained from EDA guide further analysis and contribute to the selection of appropriate modelling or visualization techniques.

The final stages involve advanced analytics, the results of which are then translated into meaningful visualizations and communicated to stakeholders. Throughout the project, there is an iterative feedback loop, allowing for refinement of analyses based on insights and stakeholder feedback. This iterative nature ensures that the data visualization and analysis project is dynamic and responsive to emerging insights and evolving stakeholder needs.

2. USE CASE

Use case refers to a specific scenario or situation where the application of data visualization techniques can provide valuable insights or address a particular business or analytical need. It outlines how the visualization will be utilized to solve a real-world problem or support decision-making. Use cases help define the purpose and goals of the data visualization, ensuring that the visual representation of data serves a meaningful and practical purpose.

Defining clear use cases is essential in guiding the design and development of data visualizations. It ensures that the visualizations are aligned with business objectives, facilitate decision-making, and provide actionable insights to the end-users. Each use case should have specific goals, data requirements, and a clear understanding of the target audience to maximize the impact of the data visualization project.

- Primary objective:

The primary objective of this data visualization project is to comprehensively assess the financial performance of hospitals while concurrently exploring the relationship between financial metrics and the spectrum of medical procedures conducted within these institutions.

- Potential Audience:

This data visualization project is designed for stakeholders like hospital administrators, financial analysts, and healthcare policymakers who seek a holistic understanding of hospital financial performance and its correlation with the range of medical procedures conducted. The visualizations will provide actionable insights to drive informed decision-making, improve resource allocation, and ultimately contribute to the overall financial health and efficiency of healthcare institutions.

- Goals:

Visualize key financial indicators such as sales, profits, and operational costs across a diverse range of hospitals. Identify trends and patterns in financial performance over specified time periods. Explore variations in financial metrics based on hospital size, location, and ownership (public vs. private). Categorize and visualize the types and frequency of medical procedures carried out by each hospital. Examine the diversity of procedures across different hospital departments and specialties. Identify high-impact procedures contributing significantly to revenue generation. Investigate the correlation between financial performance metrics (sales, profits) and the volume and types of medical procedures performed. Examine whether certain procedures have a direct impact on financial outcomes. Explore potential relationships between hospital size, specialization, and financial success.

- Benefits:

Enable hospital administrators to identify and focus on high-impact procedures that contribute to financial success. Support evidence-based decision-making for resource allocation and strategic planning. Enhance financial sustainability by uncovering patterns that can inform revenue optimization strategies.

3. PLATFORMS AND LANGUAGE

1) Jupyter Notebook

Jupyter Notebook is an open-source, web-based application that facilitates interactive computing and data exploration through a combination of code, text, and visualizations. Jupyter supports a multitude of programming languages, but its name reflects its core trio: Julia, Python, and R. The notebook format allows users to create and share documents containing live code, equations, visualizations, and narrative text.

In the realm of data analysis and visualization, Jupyter Notebook has become a preferred tool for its seamless integration of code execution and visualization outputs. Utilizing the Python programming language along with libraries like Pandas, NumPy, and Matplotlib, analysts and data scientists can perform data manipulation, statistical analysis, and generate visual representations of their findings within a single, interactive environment. The ability to run code in chunks, known as cells, allows for step-by-step exploration and immediate visualization of data, fostering an iterative and dynamic approach to the data analysis process.

Jupyter Notebooks are instrumental in creating a narrative around data analysis, providing an accessible platform for sharing insights, code, and visualizations with collaborators. The combination of code execution, data visualization, and narrative makes Jupyter Notebook a versatile and powerful tool for both beginners and experienced data professionals engaged in the intricate journey of data exploration and storytelling.

2) SQL Server Management Studio (SSMS)

The SQL Server Management Studio (SSMS) provides a user-friendly interface for designing tables, relationships, and defining constraints. This ensures that data is organized, relationships are maintained, and database structures are optimized for efficient querying.
Microsoft SQL Server is a comprehensive relational database management system (RDBMS) developed by Microsoft. It is designed to store, retrieve, and manage vast amounts of data efficiently, providing a secure and scalable solution for businesses and organizations.
SQL Server supports the structured query language (SQL), making it compatible with various programming languages and applications. Its capabilities extend beyond traditional database functionalities, encompassing business intelligence, data warehousing, and advanced analytics.

In the realm of data cleaning and filtering, MS SQL Server plays a pivotal role in ensuring data integrity and quality. Its Transact-SQL (T-SQL) language allows users to write powerful queries to identify and rectify inconsistencies, missing values, and outliers within datasets.

Moreover, SQL Server provides robust filtering capabilities, allowing users to extract specific subsets of data based on predefined criteria. When it comes to database design, SQL Server offers a robust environment for creating, modifying, and managing databases. Overall, Microsoft SQL Server stands as a versatile platform for not only storing and retrieving data but also for implementing robust data cleaning practices and designing efficient, well-structured databases.

3) Python programming language

Python is a versatile, high-level programming language known for its readability, simplicity, and extensive ecosystem of libraries and frameworks. In the field of data, Python has become a cornerstone for data analysis, manipulation, and visualization. Its popularity in the data domain is attributed to libraries such as Pandas, NumPy, and Matplotlib, which provide powerful tools for handling and analysing structured data efficiently. Python's simplicity and readability make it an ideal choice for data scientists, analysts, and engineers, allowing them to focus on the logic of their data workflows rather than getting bogged down in complex syntax.
Additionally, Python serves as the language of choice for machine learning and artificial intelligence applications, with frameworks like TensorFlow and scikit-learn, enabling practitioners to implement sophisticated models and algorithms seamlessly. Overall, Python's versatility and accessibility make it a go-to language in the data field, facilitating tasks ranging from data cleaning and analysis to advanced machine learning applications.

Libraries used:

- Pandas:
  Pandas is a powerful Python library for data manipulation and analysis, providing easy-to-use data structures and functions essential for working with structured data. It introduces two primary data structures, namely Series and DataFrame, enabling efficient handling of diverse datasets. Pandas is widely utilized in data science, facilitating tasks such as data cleaning, filtering, aggregation, and seamless integration with other libraries for comprehensive data analysis and visualization.

- Seaborn:
  Seaborn is a Python data visualization library based on Matplotlib, offering an enhanced and aesthetically pleasing interface for creating informative statistical graphics. It simplifies the process of generating complex visualizations, including heatmaps, violin plots, and pair plots, with minimal code. Seaborn is particularly useful for exploring relationships in datasets and presenting compelling visual insights in a concise and efficient manner.

- Matplotlib:
  Matplotlib is a comprehensive Python library for creating static, interactive, and publication-quality visualizations. It offers a wide range of plotting functions

and customization options, allowing users to create a diverse array of graphs, charts, and plots. Matplotlib is extensively used in data analysis, scientific research, and data visualization projects to effectively communicate insights and findings from datasets.

- Warnings:
  The warnings library in Python provides a flexible mechanism for issuing warning messages during code execution. It enables developers to alert users about potential issues or deprecated features in a non-disruptive manner. By using the warnings module, developers can ensure that important information is communicated without interrupting the flow of the program.

4) Tableau Public

Tableau Public is a free, cloud-based data visualization platform that enables individuals to create, share, and explore interactive visualizations and dashboards. Geared towards a broad audience, Tableau Public is accessible to anyone and allows users to connect to various data sources, including Excel, CSV, or cloud databases, to bring their datasets to life. With an intuitive drag-and-drop interface, users can design compelling visualizations, apply filters, and create dynamic dashboards, making it an ideal tool for students, journalists, and data enthusiasts to tell stories through data without the need for advanced coding skills.

The primary use of Tableau Public is to democratize data by making it accessible and engaging for a wide audience. Users can publish their visualizations to Tableau Public's cloud-based platform, making them shareable and embeddable across websites and blogs. This fosters a collaborative environment where individuals can showcase their data-driven insights, learn from others' work, and contribute to the broader data visualization community. Tableau Public serves as a valuable resource for creating impactful visual narratives and fostering a culture of data literacy and storytelling on a global scale.

IMPLEMENTATION

1. PROBLEM STATEMENT

Healthcare in the United States faces the dual challenge of ensuring robust financial performance for hospitals while concurrently improving the quality and efficiency of services provided to patients. The lack of a comprehensive understanding of the interplay between financial metrics and service delivery hinders hospitals in optimizing their operations. This data visualization project aims to address this challenge by analysing US patients' data to uncover patterns and correlations between hospitals' financial performance, including sales, profits, and operational costs, and the quality of services rendered.

The project seeks to answer critical questions such as whether certain medical procedures contribute significantly to revenue generation and how even different hospitals 'locations impact financial outcomes, and whether there are discernible trends in the data generated by the hospitals. By harnessing the power of data visualization, we intend to create interactive dashboards and visual representations that provide actionable insights for hospital administrators, enabling them to make informed decisions to enhance both financial viability and service quality.

Ultimately, this project aims to bridge the gap between financial performance and service delivery in US hospitals, fostering a data-driven approach that empowers healthcare institutions to allocate resources effectively, optimize operational processes, and ultimately improve the overall patient experience. Through compelling visualizations, we aim to empower stakeholders to navigate the complex landscape of healthcare data, facilitating strategic decision-making that aligns financial success with enhanced service delivery.

2. COLLECT AND GATHER DATA

Collecting and gathering data is a foundational step in a data visualization project, involving the systematic acquisition of relevant datasets to address specific objectives. This process begins with identifying and defining the scope of the project, determining the key variables, and selecting appropriate data sources. Data may be sourced from internal databases, external repositories, or a combination of both. It is crucial to ensure data quality by addressing issues such as missing values, outliers, and inconsistencies.
Depending on the project's requirements, data may be collected through surveys, APIs, or by extracting information from existing databases. The goal is to assemble a comprehensive and clean dataset that forms the basis for meaningful visualizations and insights in subsequent stages of the project.
The data in use for this project was collected and presented in a excel file by the MTE team. Mr. Ankit Hasija, my supervisor kindly presented the data to me in the project proposal document.

3. DESIGN DATABASE

Database design in a data visualization project involves structuring and organizing data in a way that facilitates efficient retrieval and analysis. It requires defining tables, establishing relationships between them, and choosing appropriate data types for each attribute. Well-designed databases are crucial for seamless integration with visualization tools, allowing for quick and accurate querying of relevant information. A thoughtful database design ensures that data is stored in a format conducive to visualization, enabling the creation of insightful charts, graphs, and dashboards. Additionally, it contributes to the overall performance and scalability of the project by optimizing data storage and retrieval processes.

The database design for the project is as follows:

NAME OF DATABASE: DBPatientsUS

TABLES:

i. TBLPeople

ATTRIBUTE	DATATYPE	
RowID	INT	(PK)
CreatedDt	DATETIME	NULL
Region	NVARCHAR(255)	NULL
Person	NVARCHAR(255)	NULL

ii. TBLPatientInfo

ATTRIBUTE	DATATYPE	
RowID	INT	NULL
CreatedDt	DATETIME	NULL
PatientID	NVARCHAR(255)	(PK)
PatientName	NVARCHAR(255)	NULL
Region	NVARCHAR(255)	NULL
State	NVARCHAR(255)	NULL
City	NVARCHAR(255)	NULL
Country	NVARCHAR(255)	NULL
Segment	NVARCHAR(255)	NULL
PostalCode	FLOAT	NULL

iii. TBLProcedureInfo

ATTRIBUTE	DATATYPE	
RowID	INT	NULL
ATTRIBUTE	DATATYPE	NULL
CreatedDt	DATETIME	NULL
PatientID	NVARCHAR(255)	NULL

ProcedureID	NVARCHAR(255)	(PK)
ProcedureDt	DATETIME	NULL
DischargeDt	DATETIME	NULL
ShipMode	NVARCHAR(255)	NULL
TransactionID	NVARCHAR(255)	NULL
Category	NVARCHAR(255)	NULL
SubCategory	NVARCHAR(255)	NULL
Sales	FLOAT	NULL
Discount	FLOAT	NULL
Profit	FLOAT	NULL

iv. TBLReturns

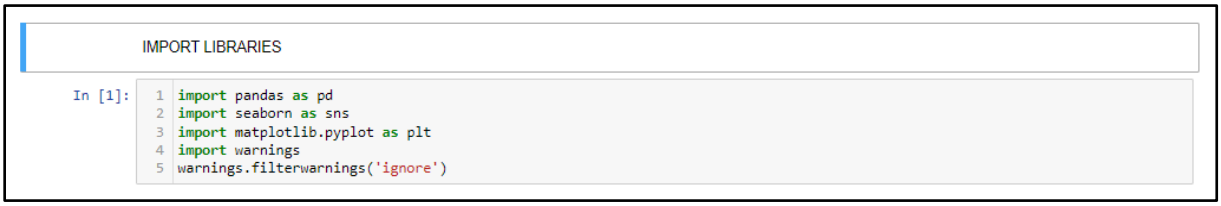
ATTRIBUTE	DATATYPE	
RowID	INT	(PK)
CreatedDt	DATETIME	NULL
PatientID	NVARCHAR(255)	NULL
ProcedureID	NVARCHAR(255)	NULL
ProcedureDt	DATETIME	NULL
Relapses	NVARCHAR(255)	NULL

4. DATA FILTERING, CLEANING AND EDA

4.1 Import the libraries.

The very first step is to import the libraries in the Jupyter notebook so as to use them efficiently to access the data, clean and filter it, there by visualise it to draw insights.

The libraries being used are: Pandas, Seaborn, Matplotlib and Warnings.



4.2 Load the Data, Filter Data & Cleaning Data

The next step is to load the data. A Comma Separated File(csv) is loaded to read information in the data set. A function “read_csv()” is used for the purpose of reading a csv file.

LOAD CSV FILE

In [7]:
```
#read_csv() is used to read data from comma seperated values file
#file name - updated-patients-united-states.csv
# 'r' is used brfore the path to convert the string to a raw sting
df = pd.read_csv(r'~\Desktop\DA DS\updated-patients-united-states.csv')
```

- CSV file:

  CSV (Comma-Separated Values) file is a plain-text file format commonly used for storing tabular data. Each row in the file represents a record, and values within each row are separated by commas or other delimiters. Python's csv module facilitates easy reading and writing of CSV files, making it a widely adopted format for data interchange due to its simplicity and compatibility with various data analysis tools and libraries.

- read_csv ():

  In Python, read_csv () is a function provided by the Pandas library that helps you read data from a CSV (Comma-Separated Values) file. Imagine a CSV file as a table with rows and columns, where each row represents a record, and values in each column are separated by commas. When you use read_csv (), Pandas reads this file and turns it into a special kind of table called a Data Frame. A Data Frame is like a supercharged table that allows you to easily manipulate and analyse your data. So, read_csv () is like a magical spell that turns a CSV file into a Data Frame, making it ready for all sorts of data operations and exploration in Python.

Now, we explore data to determine how and what to perform data filtering and cleaning upon. The various functions used in the project for this purpose are as mentioned below:

- head ()

  In Python's Pandas library, the head () method is used to display the first few rows of a DataFrame. When you have a large dataset, and you want to quickly inspect what it looks like, head () comes in handy. By default, it shows the first 5 rows of the DataFrame, giving you a snapshot of the data.

EXPLORATORY DATA ANALYSIS AND DATA CLEANING

In [8]:
```
#view first 05 entries
df.head()
```

Out[8]:

	Row ID	Procedure ID	Procedure Date (Order Date)	Discharge Date	Ship Mode	Patient ID	Segment	Country	City	State	...	Region	Category	Sub-Category	Sale
0	1	CA-2017-152156	2017-11-08 00:00:00	2017-11-11 00:00:00	Deluxe	CG-12520	Non-Insured	United States	Henderson	Kentucky	...	South	Cardiology	CABG	261.960
1	2	CA-2017-152156	2017-11-08 00:00:00	2017-11-11 00:00:00	Deluxe	CG-12520	Non-Insured	United States	Henderson	Kentucky	...	South	Cardiology	Coronary Angioplasty	731.940
2	3	CA-2017-138688	2017-06-12 00:00:00	2017-06-16 00:00:00	Deluxe	DV-13045	Insured	United States	Los Angeles	California	...	West	Orthopedics	Knee Hemiarthroplastyhroplasty	14.620
		US-2016-	2016-10-	2016-10-		SO-	Non-	United	Fort						

- tail ()

  In Python's Pandas library, the tail () method is used to display the last few rows of a DataFrame. Similar to head (), which shows the beginning of the DataFrame, tail () allows you to quickly inspect the end of your dataset. This is particularly useful when you want to check the last few records or observations in your data.

  By default, tail () shows the last 5 rows of the DataFrame.

```
In [10]:
#view last 10 entries
df.tail(10)
```

Out[10]:

	Row ID	Procedure ID	Procedure Date (Order Date)	Discharge Date	Ship Mode	Patient ID	Segment	Country	City	State	...	Region	Category	Sub-Category	Sales	(Qua
9984	8491	CA-2017-158841	2017-02-02 00:00:00	2017-02-04 00:00:00	Deluxe	SE-20110	Non-Insured	United States	Arlington	Virginia	...	South	Cardiology	Artificial pacemaker surgerys	18.690	
9985	8492	CA-2018-106824	2018-07-07 00:00:00	2018-07-11 00:00:00	Standard	AT-10735	Non-Insured	United States	Los Angeles	California	...	West	Orthopedics	Arthroscopy	5.940	
9986	8493	CA-2016-109190	2016-10-23 00:00:00	2016-10-28 00:00:00	Standard	CC-12685	Non-Insured	United States	Lubbock	Texas	...	Central	Orthopedics	Shoulder Replacement	60.736	
9987	8494	CA-2016-109190	2016-10-23 00:00:00	2016-10-28 00:00:00	Standard	CC-12685	Non-Insured	United States	Lubbock	Texas	...	Central	Neurology	Radiosurgery	479.976	

- shape

  In Pandas, the shape attribute provides a tuple representing the dimensions of a DataFrame.

- columns ()
  In Pandas, the columns method returns a list of column labels or names from a DataFrame, allowing you to quickly inspect the features present in your data. For example, df. columns() provides a list of column names in the DataFrame 'df'.

- info ()

  The Pandas info () method provides a concise summary of a DataFrame, including the data types, non-null counts, and memory usage, offering a quick overview of the dataset's structure. For instance, df.info () provides essential information about the DataFrame 'df'.

```
In [12]:
#find the dimensions of the data being worked upon
df.shape
```

```
Out[12]: (9994, 21)
```

```
In [13]:
# find out the columns in data set
df.columns
```

```
Out[13]: Index(['Row ID', 'Procedure ID', 'Procedure Date (Order Date)',
       'Discharge Date', 'Ship Mode', 'Patient ID', 'Segment', 'Country',
       'City', 'State', 'Postal Code', 'Region', 'Category', 'Sub-Category',
       'Sales', 'Days (Quantity)', 'Discount', 'Profit', 'Discount_Amount',
       'Relapses', 'Person'],
      dtype='object')
```

```
In [14]:
# find the meta data of the data set
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 21 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Row ID                      9994 non-null   int64
 1   Procedure ID                9994 non-null   object
 2   Procedure Date (Order Date) 9994 non-null   object
```

- describe ()

  This method in Pandas provides a statistical summary of numeric columns in a DataFrame, offering insights into central tendencies, spread, and distribution. It includes count, mean, standard deviation, minimum, 25th percentile, median (50th percentile), 75th percentile, and maximum values. This concise summary is instrumental in understanding the distribution and basic statistics of numerical data within the DataFrame.

```
In [33]:
#determine, if there are any outliers
# describe() give the summary statistics of numerical columns
df.describe()
```

Out[33]:

	Row ID	Procedure Date (Order Date)	Sales	Days (Quantity)	Discount	Profit	Discount_Amount
count	9994.000000	9994	9994.000000	9994.000000	9994.000000	9994.000000	9994.000000
mean	4997.500000	2017-04-30 05:17:08.056834048	229.858001	3.789574	0.156203	28.656896	32.277580
min	1.000000	2015-01-03 00:00:00	0.444000	1.000000	0.000000	-6599.978000	0.000000
25%	2499.250000	2016-05-23 00:00:00	17.280000	2.000000	0.000000	1.728750	0.000000
50%	4997.500000	2017-06-26 00:00:00	54.490000	3.000000	0.200000	8.666500	1.036800
75%	7495.750000	2018-05-14 00:00:00	209.940000	5.000000	0.200000	29.364000	14.870400
max	9994.000000	2018-12-30 00:00:00	22638.480000	14.000000	0.800000	8399.976000	11319.240000
std	2885.163629	NaN	623.245101	2.225110	0.206452	234.260108	164.025577

```
In [21]:
# some of the columns have object as data type - MEANS THEY ARE CATEGORICAL COLUMNS
# We can see how many unique values are there in each object data type columns
# this is done using describe()
# describe() is only for numbers, BUT WHEN 'incluce='object'', it gives summary statistics for for object columns
df.describe( include='object' )
```

Out[21]:

	Procedure ID	Discharge Date	Ship Mode	Patient ID	Segment	Country	City	State	Region	Category	Sub-Category	Relapses	Person
count	9994	9994	9994	9994	9994	9994	9994	9994	9994	9994	9994	800	9994
unique	5009	1334	4	793	3	1	531	49	4	3	16	2	4
top	CA-2018-100111	2016-12-16 00:00:00	Standard	WB-21850	Non-Insured	United States	New York City	California	West	Orthopedics	ACL Surgery	No	Anna Andreadi
freq	14	35	5968	37	5191	9994	915	2001	3203	6026	1989	651	3203

- unique ()

  In Pandas, this method is used to retrieve an array of unique values from a column, allowing you to quickly identify distinct elements in categorical or non-numeric data.

```
In [26]:
#determine what unique values are there in each object using a for loop
for col in df.describe( include='object' ).columns:
    print(col)
    print(df[col].unique())
    print(100*'-')
```

```
Procedure ID
['CA-2017-152156' 'CA-2017-138688' 'US-2016-108966' ... 'CA-2018-106824'
 'CA-2016-109190' 'CA-2017-143154']
----------------------------------------------------------------------------------------------------
Discharge Date
['2017-11-11 00:00:00' '2017-06-16 00:00:00' '2016-10-18 00:00:00' ...
 '2018-04-16 00:00:00' '2015-06-30 00:00:00' '2018-02-15 00:00:00']
----------------------------------------------------------------------------------------------------
Ship Mode
['Deluxe' 'Standard' 'Twin Sharing' 'Suite']
----------------------------------------------------------------------------------------------------
Patient ID
['CG-12520' 'DV-13045' 'SO-20335' 'BH-11710' 'AA-10480' 'IM-15070'
 'HP-14815' 'PK-19075' 'AG-10270' 'ZD-21925' 'KB-16585' 'SF-20065'
 'EB-13870' 'EH-13945' 'TB-21520' 'MA-17560' 'GH-14485' 'SN-20710'
 'LC-16930' 'RA-19885' 'ES-14080' 'ON-18715' 'PO-18865' 'LH-16900'
 'DP-13000' 'JM-15265' 'TB-21055' 'KM-16720' 'PS-18970' 'BS-11590'
 'KD-16270' 'HM-14980' 'JE-15745' 'KB-16600' 'SC-20770' 'DN-13690'
```

- isnull ()

  The isnull() method in Pandas returns a DataFrame of the same shape as the input, with each element indicating whether it is a null or missing value (True) or not (False). It is useful for identifying and handling missing data within a dataset.

```
In [27]: df.isnull().sum()

Out[27]: Row ID                          0
         Procedure ID                    0
         Procedure Date (Order Date)     0
         Discharge Date                  0
         Ship Mode                       0
         Patient ID                      0
         Segment                         0
         Country                         0
         City                            0
         State                           0
         Postal Code                    11
         Region                          0
         Category                        0
         Sub-Category                    0
```

- drop () and dropna ()

  The drop () method in Pandas is used to remove specified rows or columns from a DataFrame based on labels or indices.

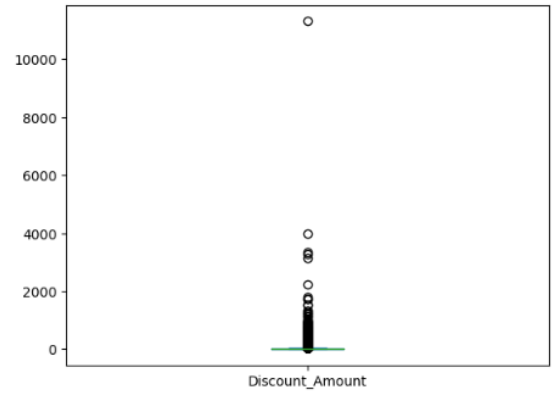
  On the other hand, dropna () is used to remove rows with missing values from a DataFrame. When applied to a DataFrame, it eliminates any row containing at least one NaN or null value.

```
In [28]: df.shape

Out[28]: (9994, 21)

In [29]: #drop Relapses as 9194 missing values are there out of 9994 records
         # drop Postal code as 11 values are missing
         df.drop(['Relapses','Postal Code'], axis = 1, inplace = True)
         # dropna() method removes the rows that contains NULL values
         df.dropna(inplace = True)
```

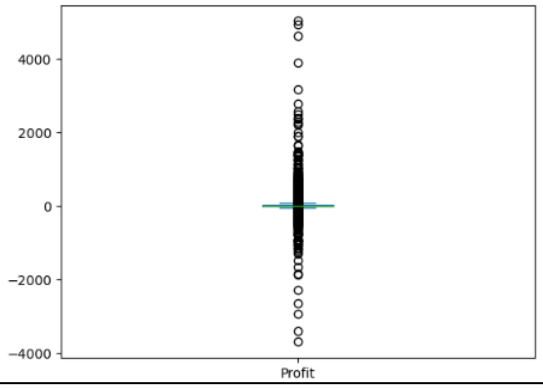
- plot ()

  In Pandas, the plot () method is used to generate basic plots, such as line charts or bar graphs, from a DataFrame.

```
In [36]: #determine how far is the outlier using box plot
         df['Discount_Amount'].plot( kind = 'box')

Out[36]: <Axes: >
```



```
In [48]: 1 df['Profit'].plot( kind = 'box')
Out[48]: <Axes: >
```



5. DASHBOARD PROTOTYPE – TABLEAU

- What is Tableau?

  Tableau is a powerful data visualization and business intelligence tool that allows users to transform raw data into understandable and visually appealing insights. It is widely used across various industries to analyse, interpret, and share complex datasets. Tableau simplifies the process of creating interactive and dynamic dashboards, charts, and graphs, making it accessible to both technical and non-technical users.

- What is the use of Tableau?

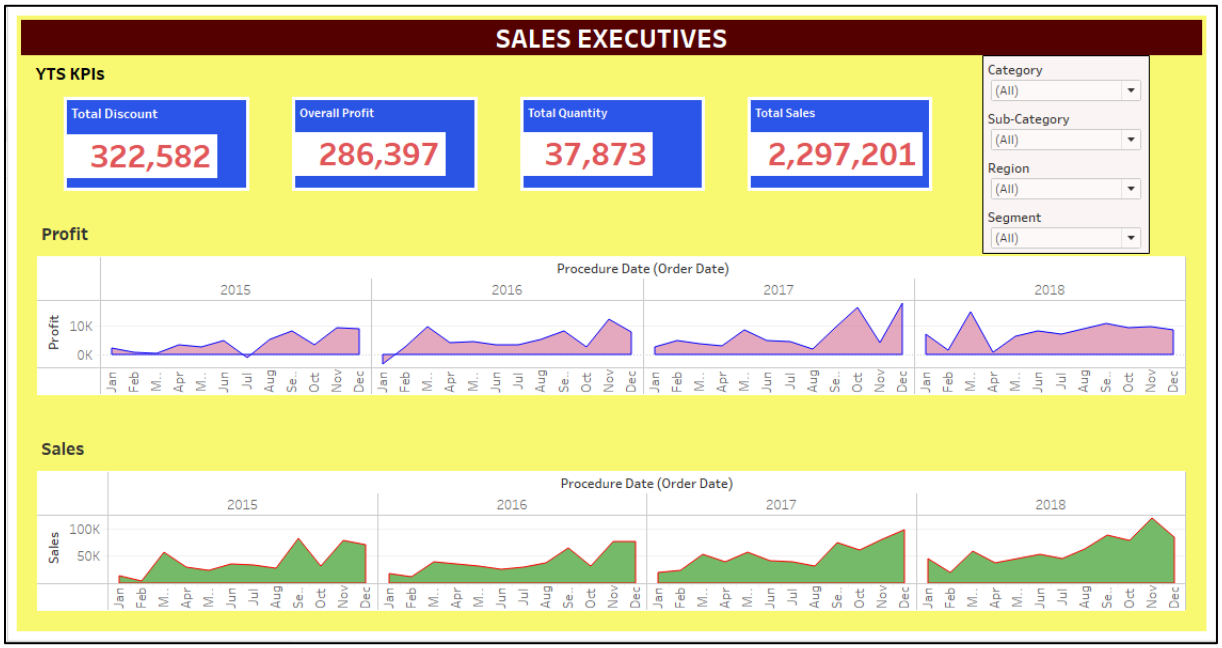
  Its primary purpose is to help organizations make data-driven decisions by providing a user-friendly platform for exploring and presenting data visually. Tableau connects to various data sources, including databases, spreadsheets, and cloud-based platforms, enabling users to bring diverse datasets together for analysis. Through a drag-and-drop interface, users can create visually compelling representations of their data without the need for extensive coding or programming skills.

- How do you use Tableau?

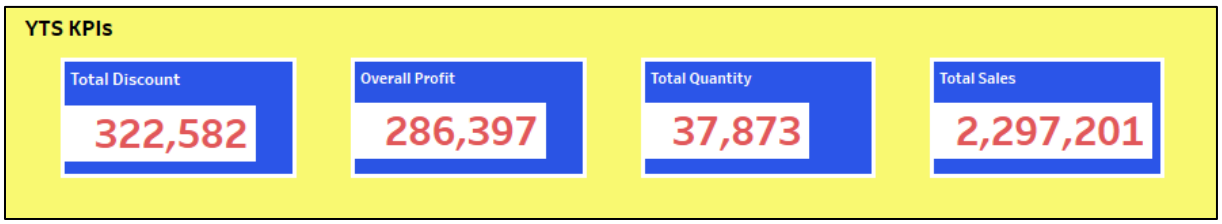
  Tableau is utilized in diverse scenarios, from creating interactive reports for business intelligence to designing interactive dashboards for executive presentations. It allows users to uncover patterns, trends, and outliers in data and provides a platform for collaborative decision-making. With features like real-time data connectivity, user-friendly interface, and a vibrant user community, Tableau has become a go-to tool for professionals seeking to extract meaningful insights and communicate them effectively.

OBSERVATION AND ANALYSIS

DASHBOARD MOCKUP



1. YTS KPIs

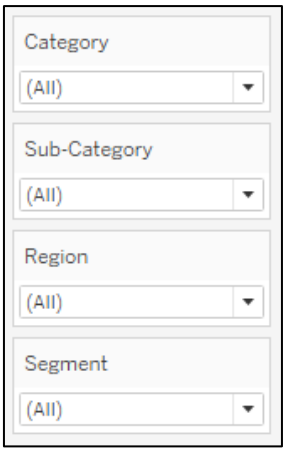


According to one of the business requirements, the above chart was implemented on the dashboard mock up for displaying the Key Performance Indicators (KPIs) of the business in the problem statement.

The chart displays: Total discount, Over Profit, Total Quantity (Days) and Total Sales.

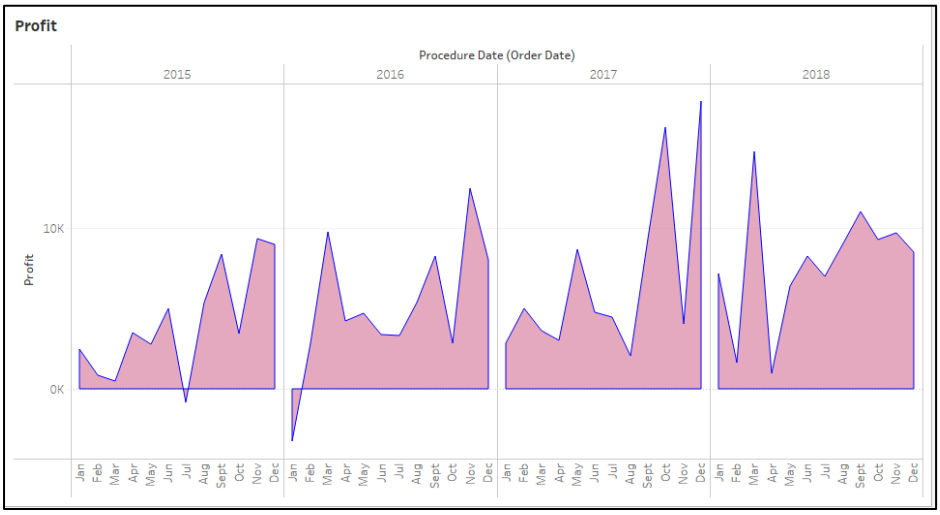
The chart is completely dynamic in respect of category and sub-category of project, region in the country and segment according to the filter panel below.

2. FILTER PANEL



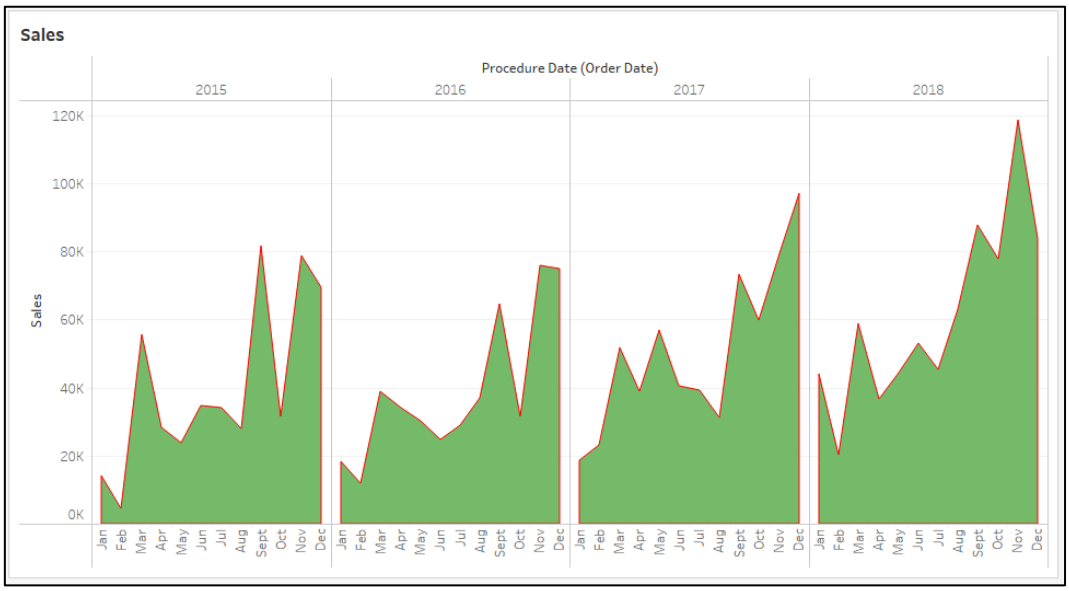
The filter panel is designed according to the business requirement and shows the relevant fields only as mentioned in the project proposal document.

3. PROFIT



The area chart displays the profit for the period of 2015 – 2018. It can be clearly deduced that the profits are higher in the year ending months of November and December in all years except, 2018, in comparison to the rest of the months in the respective years. The factors affecting the profit are further analysed as we proceed.

4. SALES



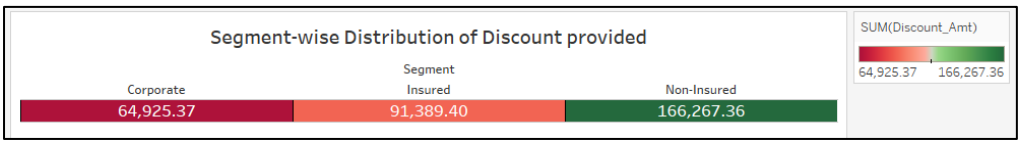
The area chart displays the sales finalised in the period of 2015-2018. As observed in the area chart for the profit, profit is higher in the year ending months, it can also be observed that the sales in the respective year ending months was also higher in comparison to the other months of the respective years. It aligns with the assumption made in the beginning that higher sales translate to higher profit.

5. ANALYSIS OF KPIs AND MAJOR ATTRIBUTES

5.1 Analysis of the effect of segment of patients on KPIs

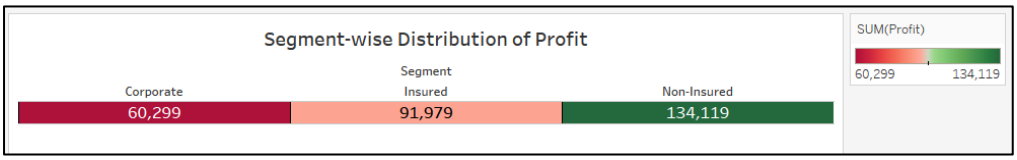
The segment mentioned in the data provided is the insurance category of the patients. From the EDA of the data in one of the previous steps, we determined that there are three segments of patients in the data collected, namely, corporate, insured and non-insured. We trying to explore the effect on KPIs in lieu of the patients from different segment.

5.1.1 Effect on Total Discount



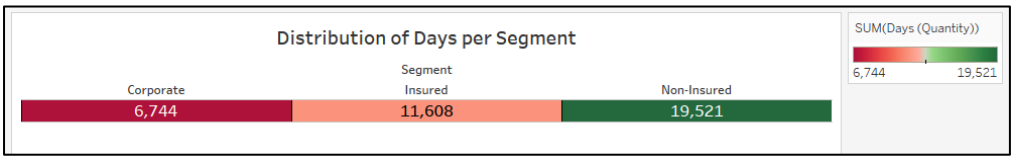
As observed from above figure, almost 50% of the discount is availed by the non-insured category of US patients.

5.1.2 Effect on Overall Profit



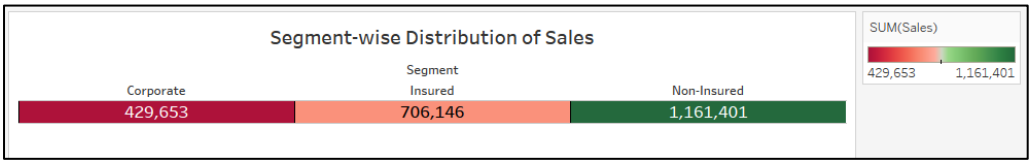
As observed from the above chart, even the profit earned from the non-insured segment of the patients is nearing 50%.

5.1.3 Effect on the Total Quantity (Days)



As observed from the above chart, the segment of non-insured patients also take up the hospital services for major amount nearing almost more than 50% in comparison to the insured and the corporate segment of patients combined.

5.1.4 Effect on the Total Sales



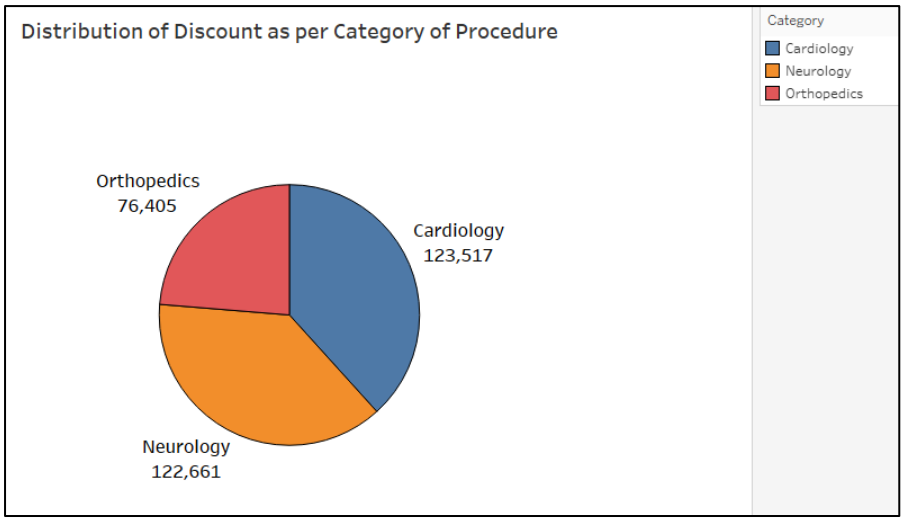
As observed from the chart above, the sales to the non-insured category of patients amount to slightly less than 50% of the total sales in the period of 2015-2018. Also, the sales of more than 50% go towards the insured and corporate sector segment of patients.

We can conclude that the non-insured patients while being served for slightly more than 50% of total time benefit almost 50% of discount and generate almost half of the profit even after less than 50% of sales are coming from them. Also, we notice that even less than 50% of sales generate almost 50% of the profit. So, we should question are assumption of higher sales translating to higher profits.

5.2 Analysis of effect of category on KPIs

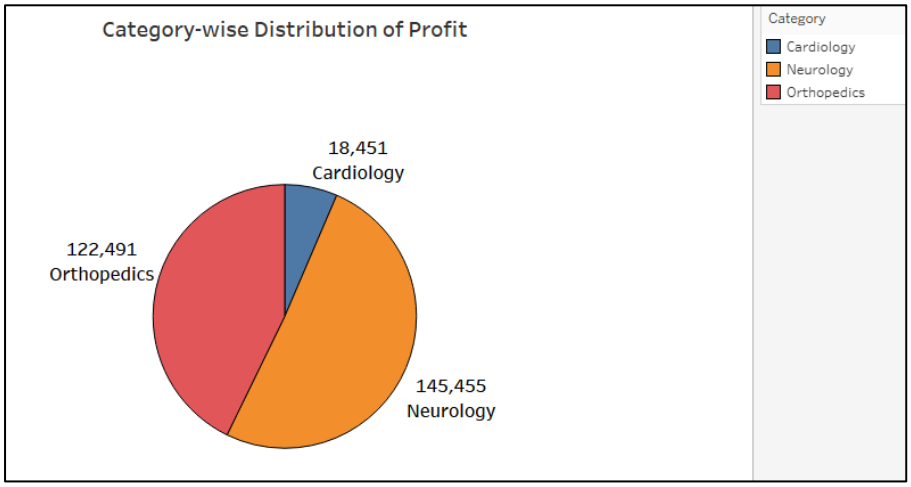
Different category of procedures is carried out at different hospitals, namely, cardiology, orthopaedics and neurology as determined from EDA on the given data. We will examine the effect on KPIs considering the different procedure categories.

5.2.1 Effect on Total Discount



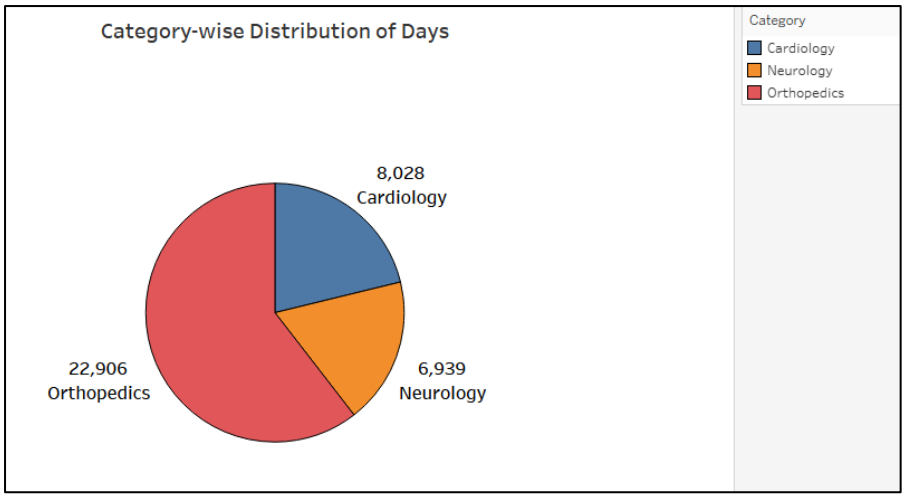
It can be observed from above pie chart that more than 75% of the discount is availed by patients undergoing procedures under the category of Cardiology and Neurology, leaving a mere less than 25% for Orthopaedics category patients.

5.2.2 Effect on Overall Profit



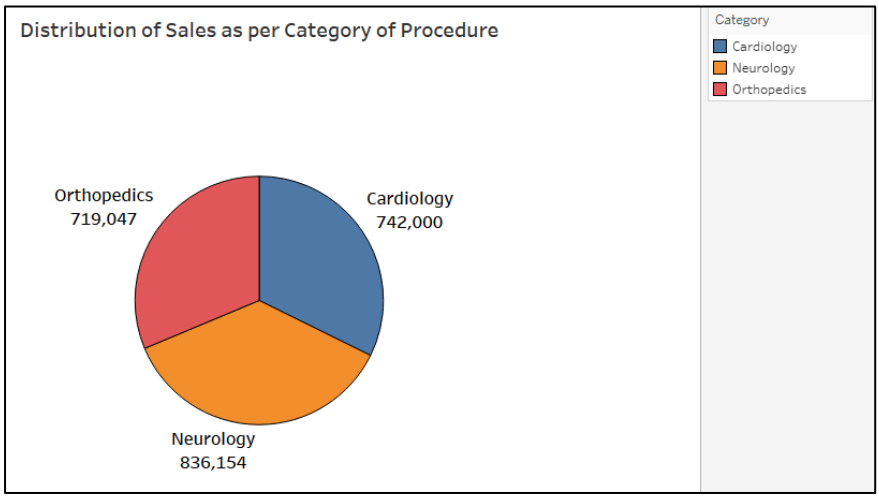
Analysis of profit presents an interesting fact that though Cardiology patients avail a large sum of discount contribute only around $1/9^{th}$ of the profit, where as orthopaedics patients contribute a large 45% of profit. So, we can conclude that the discount availed is more or less inversely proportional to profit generated.

5.2.3 Effect on the Total Quantity (Days)



This is clear from above chart that the orthopaedic patients avail around 60% of total quantity, i.e., the number of days in hospital, whereas cardiology and neurology patients each occupy a bear 20%.

5.2.4 Effect on the Total Sales



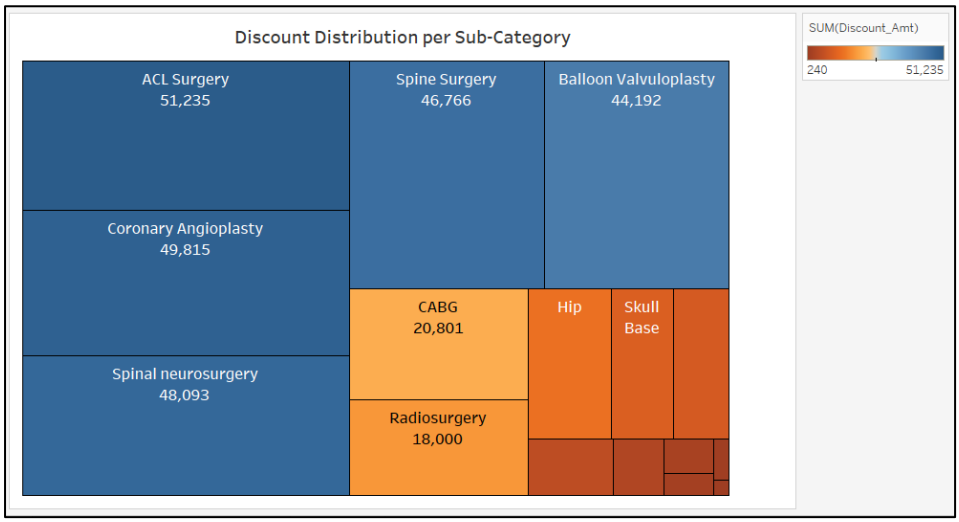
It can be clearly observed that all category of patients contributes almost equally towards sale.

The study of effect of category of procedure on performance of KPIs paint a very different picture regarding relationship of KPIs. Though, all category patients contribute equally towards sale, the profit generated by them is disproportionate, where neurology and orthopaedics generate a large chunk of profit as opposed to cardiology. It can be deduced that it is in financial health of the hospital to carry out orthopaedics category procedure.

5.3 Analysis of effect of sub-category on KPIs

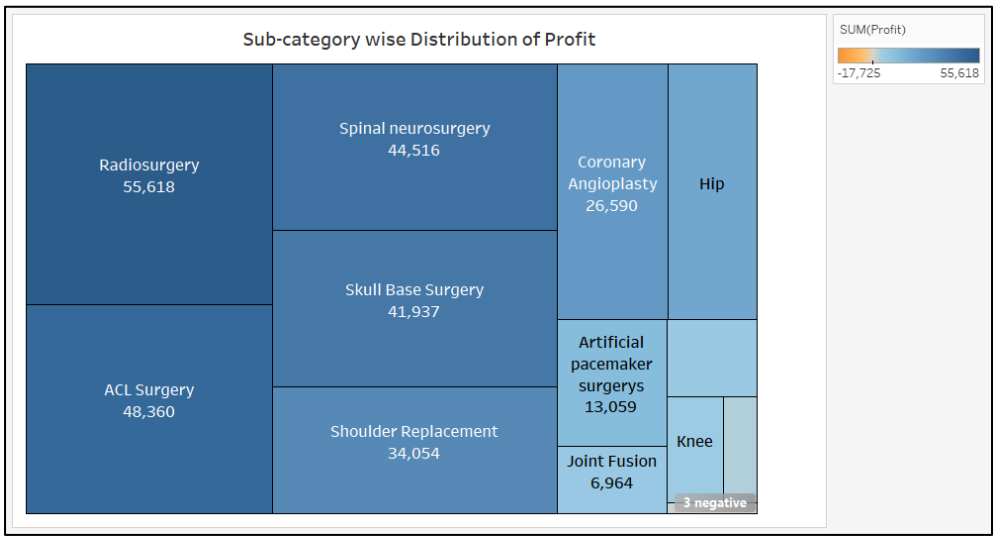
After carrying out a high-level analysis of effect on KPIs considering different categories of procedures, now we carry out analysis on KPIs a little deeper by considering the sub-categories of procedures in different hospitals. The different sub-categories are 16 in total.

5.2.1 Effect on Total Discount



It is clear from the above chart that only five of subcategories of procedure occupy nearing 80%.

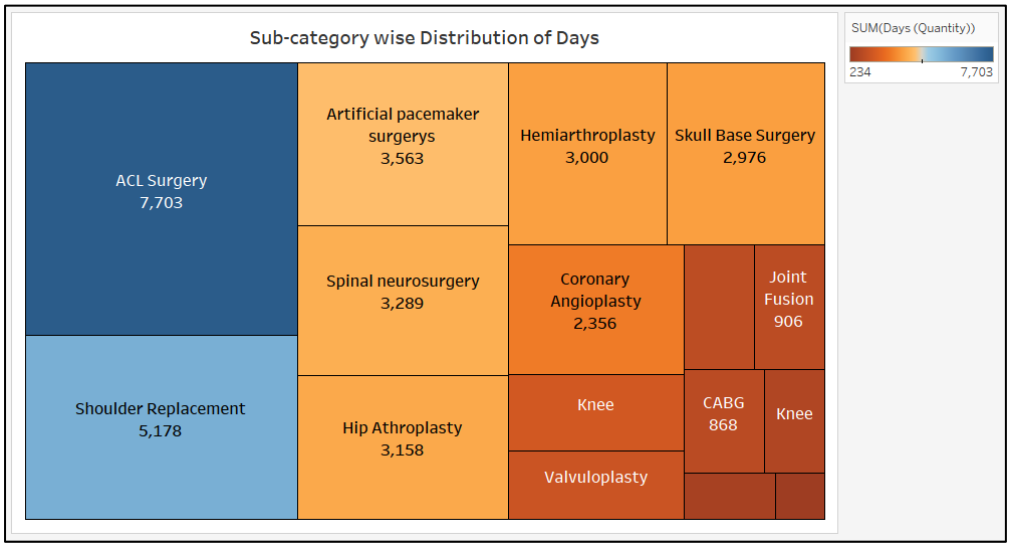
5.2.2 Effect on Overall Profit



From the above chart, the profit generated is even negative in 03 of the sub-categories. Also, the discount offered has nothing much to do with profit as except

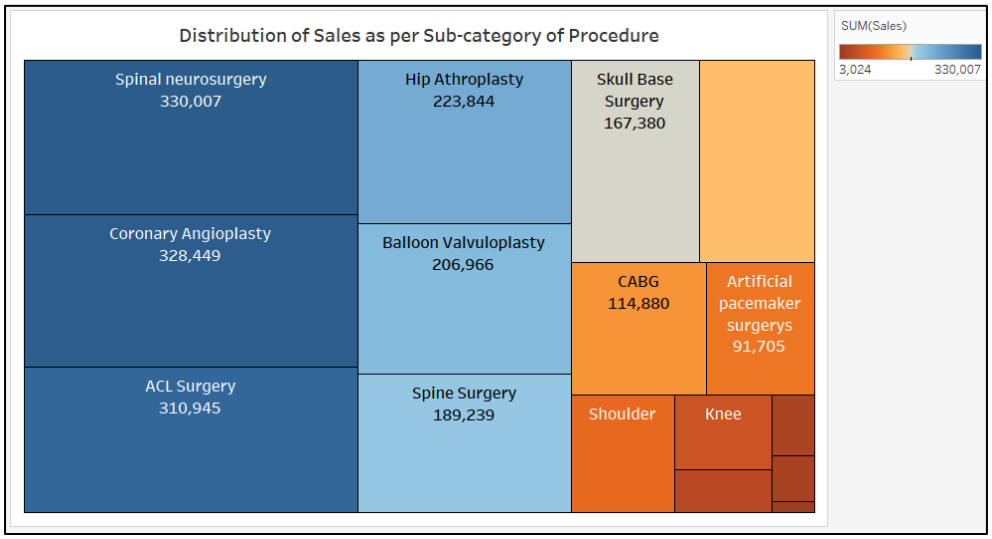
in one, all the sub-categories that generate a major chunk of profit do not avail top high discount.

5.2.3 Effect on the Total Quantity (Days)



The above chart clearly specifies that only two sub-categories of patients need more days of services and the rest fall below the median value of total quantity that is the days.

5.2.4 Effect on the Total Sales



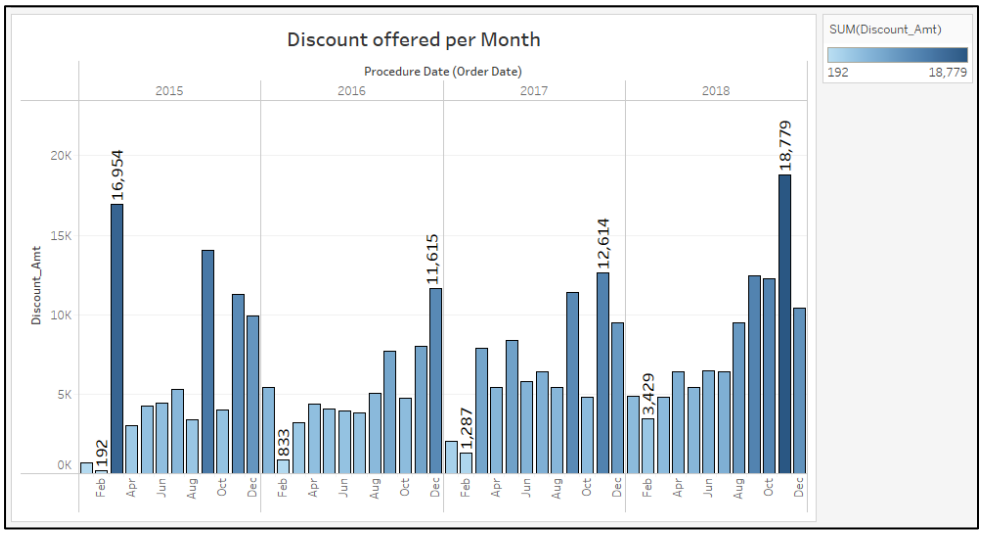
The contribution towards sales shows a high degree of divergence from different sub-categories of procedure and 06 among the total of 16 contribute the more.

The only thing that we can deduce from the studying the effect on KPIs from sub-categories is we can prioritise the 13 sub-categories which generate profit leaving the remaining 3 on the basis on urgency to maintain financial health of the hospital.

5.4 Analysis of effect of time on KPIs

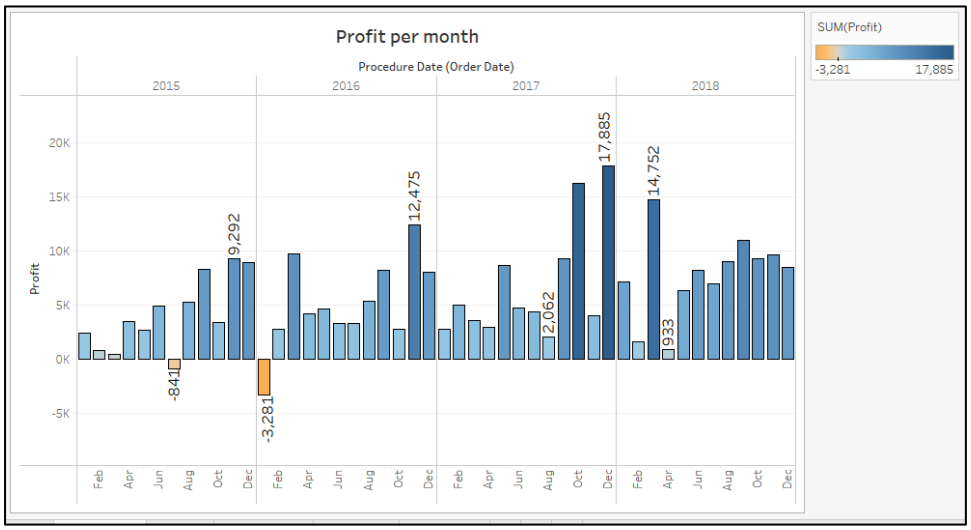
Now, we will examine the progression of KPIs with time in the period given, i.e., 2015-2018. The charts show the KPIs highlighting the minimum and maximum value of KPI attained in each year.

5.2.1 Effect on Total Discount



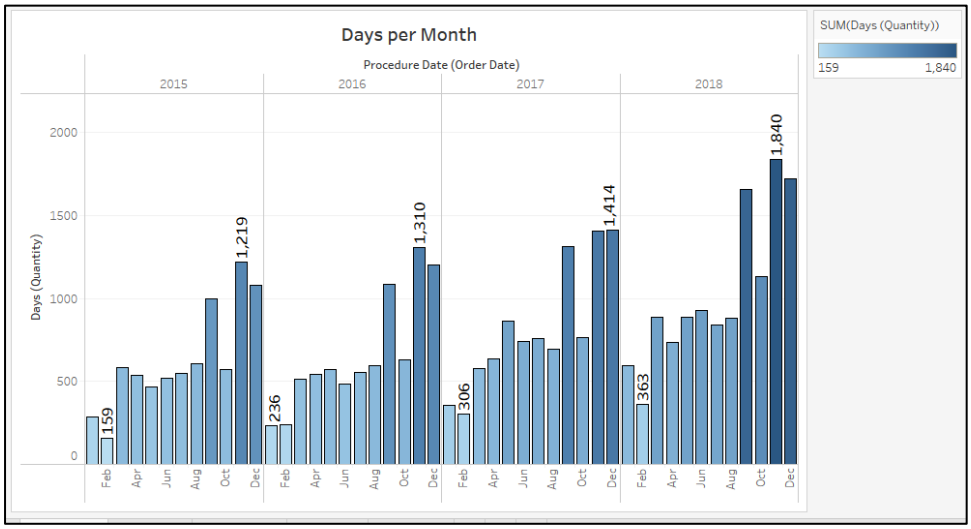
It is trending that ids observed that the ending months of year avail more discount, whereas the month of march in 2015 in an anomaly as the starting months of the year in rest of the years avail discount on the lower side of the platform.

5.2.2 Effect on Overall Profit



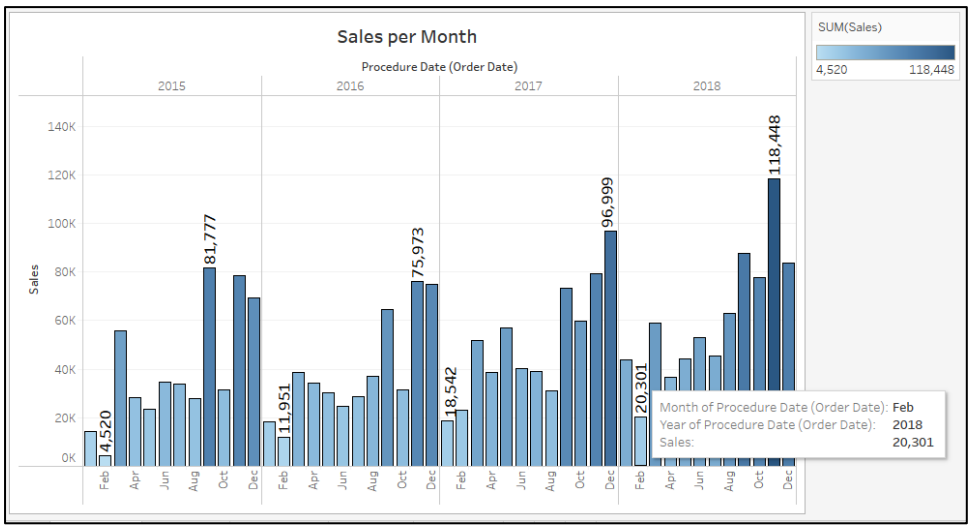
Overall, the last months of the year present maxima in profit, the complete distribution of profit is quite uneven as there is low profit generated in second half of the year as well.

5.2.3 Effect on the Total Quantity (Days)



The pattern is consistent indicating that higher number of total quantity (days) are in year ending months.

5.2.4 Effect on the Total Sales

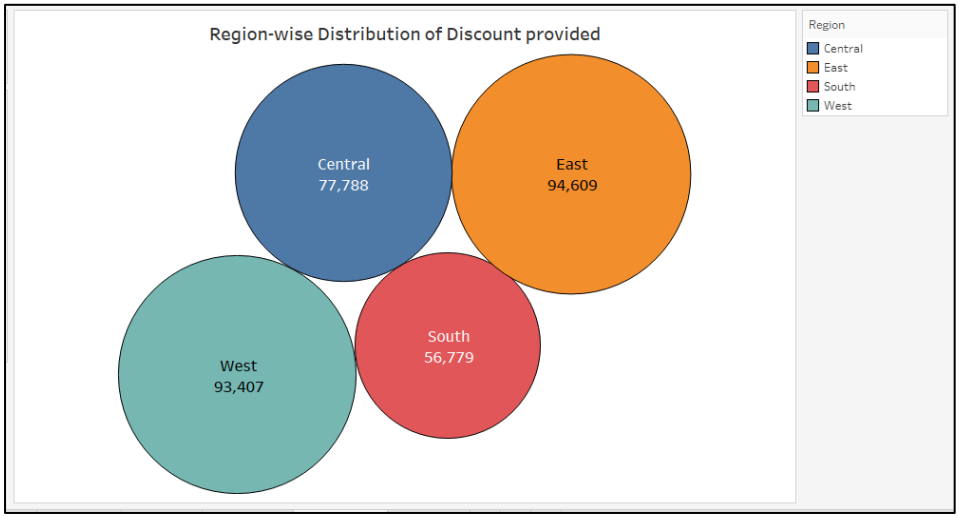


Though the last months of the years contribute more to sales, total quantity and discount, it can be deduced that the remaining months also generate comparable profit. Indicating that specific months are preferred for carrying out procedures and contribute more towards sales and profit, i.e., year-end.

5.5 Analysis of effect of spatial distribution in respect of region on KPIs

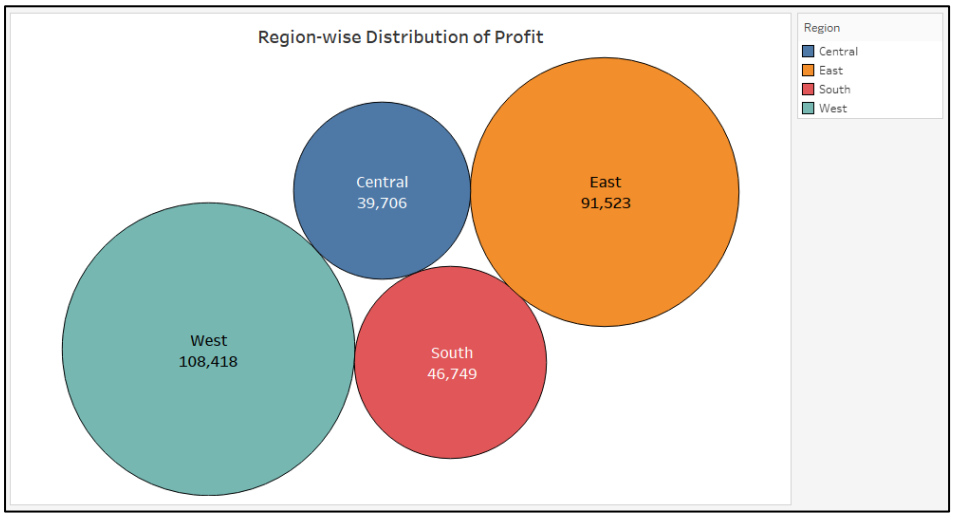
Moving forward, we will observe the performance of KPIs in the geo-spatial distribution. As per the data the information is divided in four regions of the country, namely, South, West, Central and East as known from the EDA.

5.2.1 Effect on Total Discount



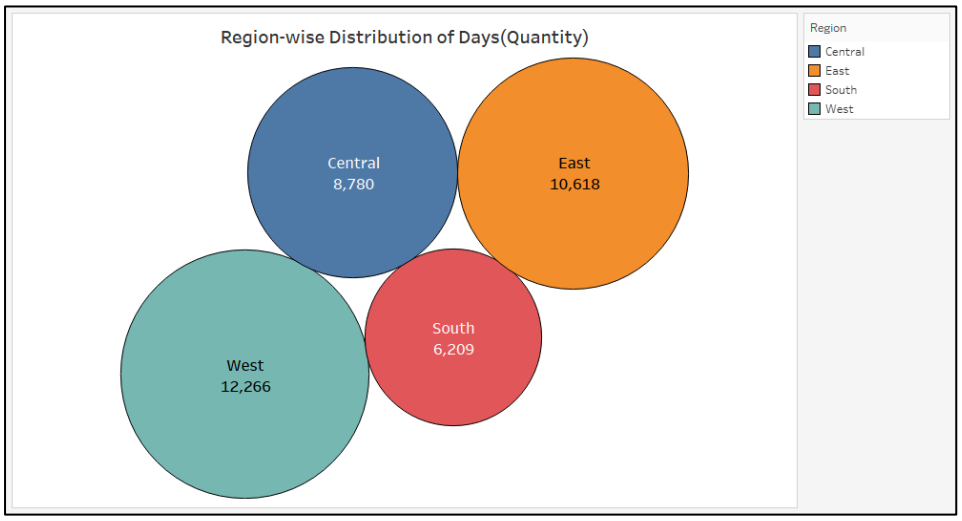
The chart above shows that except southern region, all other regions avail a large chunk of discount.

5.2.2 Effect on Overall Profit



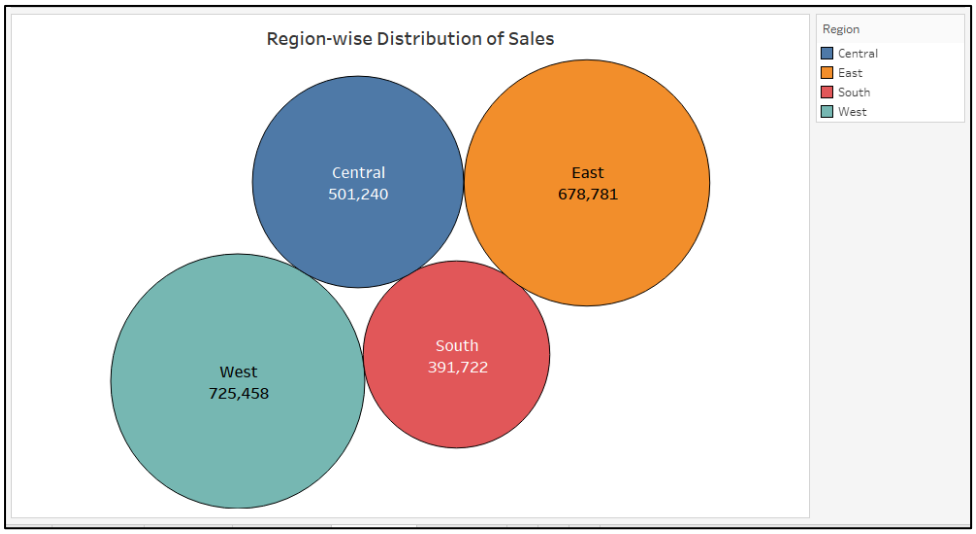
The profits generated in all regions except Central is in proportion to the discount availed. The central region contributes inversely towards profit in relation to the discount availed.

5.2.3 Effect on the Total Quantity (Days)



The above chart shows that the KPI, total quantity is in direct proportion to discount availed.

5.2.4 Effect on the Total Sales

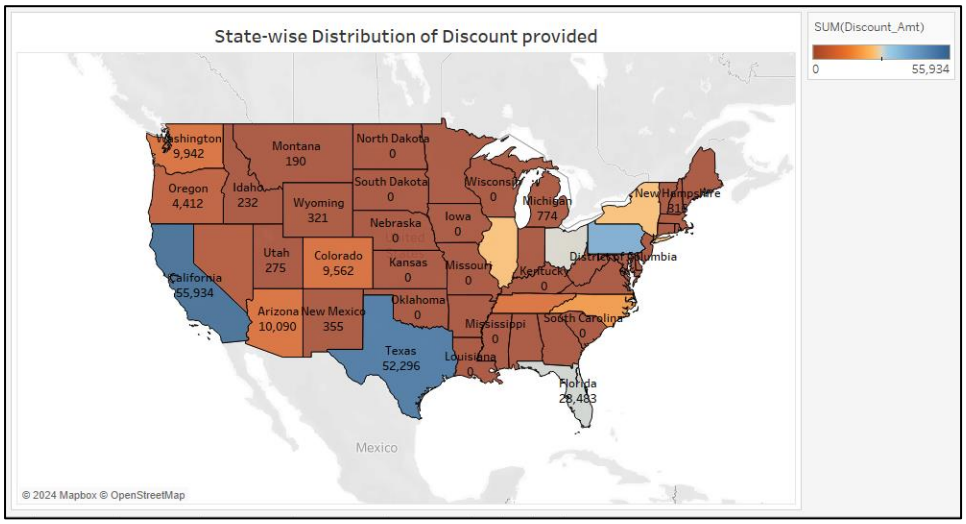


The total sales distribution in different regions is in proportion to discount and total quantity.
The conclusion that we can draw is all KPIs are in direct proportion in all regions except one which is Central region.

5.6 Analysis of effect of spatial distribution in respect of states on KPIs

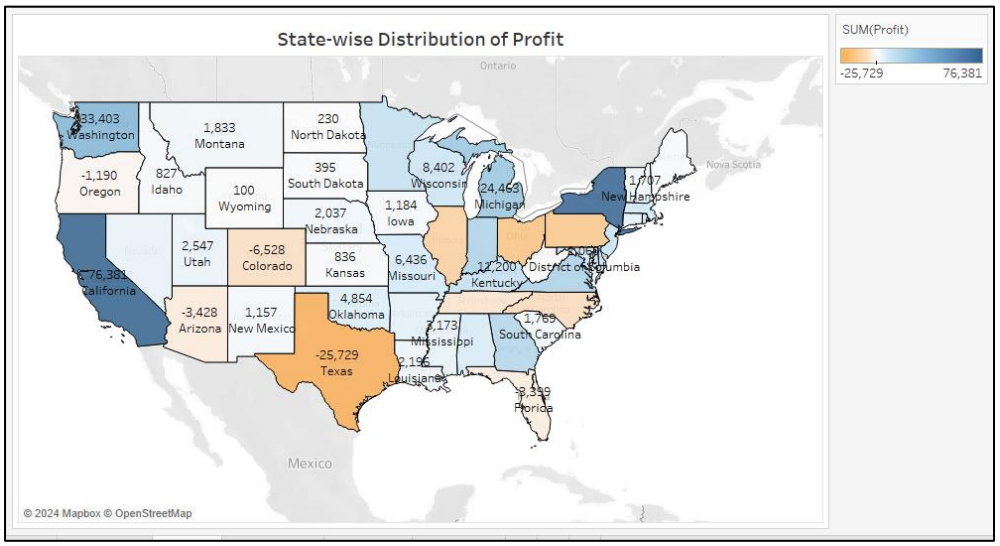
After a high-level of geo-spatial analysis, we will examine a little deeper on state level. The data provides us information of 49 states of US patients leaving just one from the 50 states of the US.

5.2.1 Effect on Total Discount



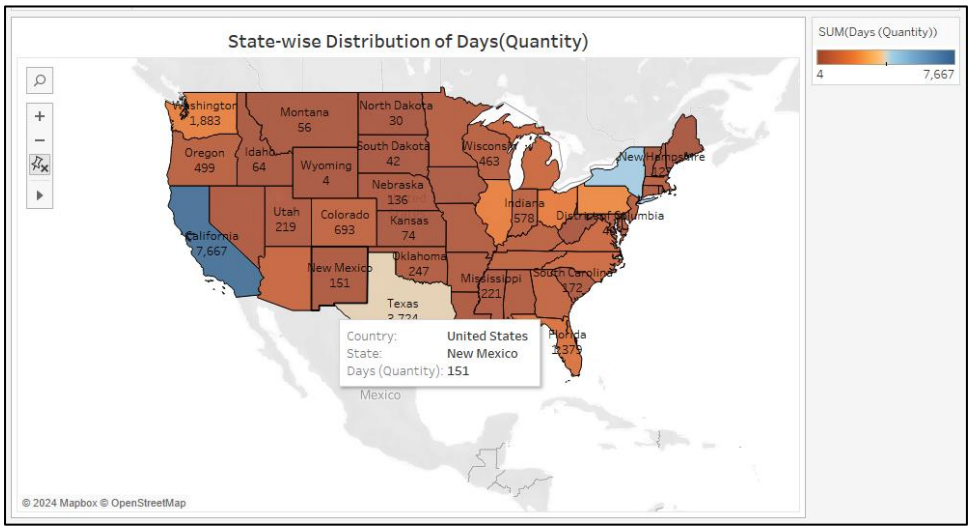
The major of the discount availed is concentrated in 05 states only, among which 3 only avail a large chunk. This shows disparity in different states.

5.2.2 Effect on Overall Profit



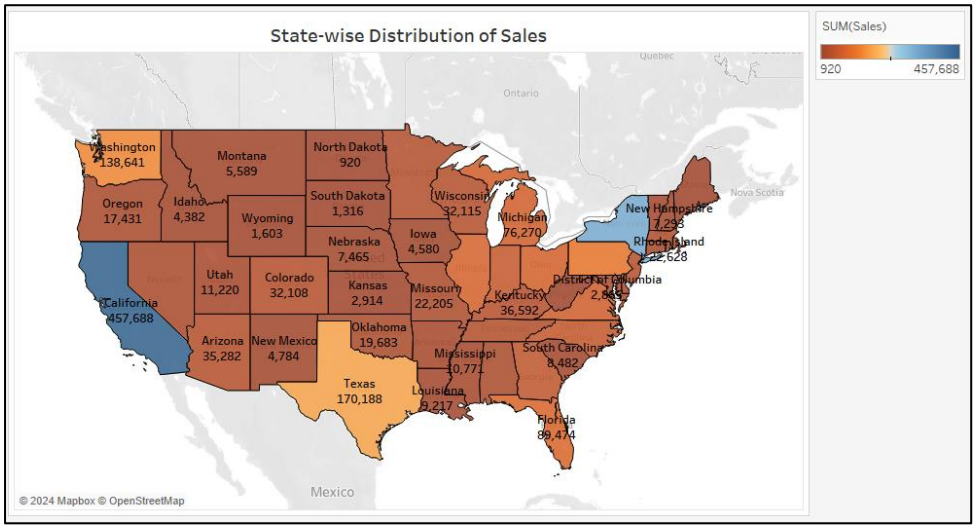
All states except 04 generate a healthy profit. Texas is inversely generating very lower profits in comparison to discount availed.

5.2.3 Effect on the Total Quantity (Days)



All except two states contribute very low to the KPI of total quantity.

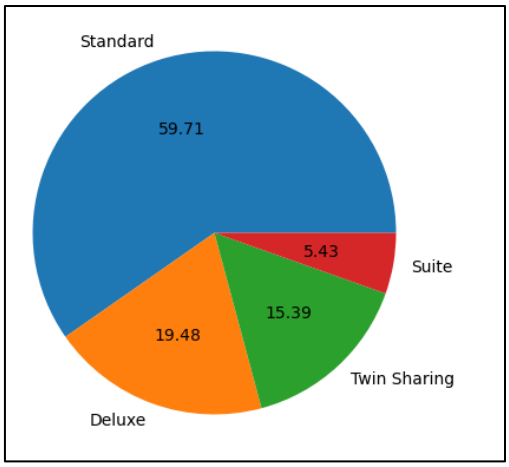
5.2.4 Effect on the Total Sales



It can be concluded that all the KPIs are in direct proportion in almost all state except in Texas.

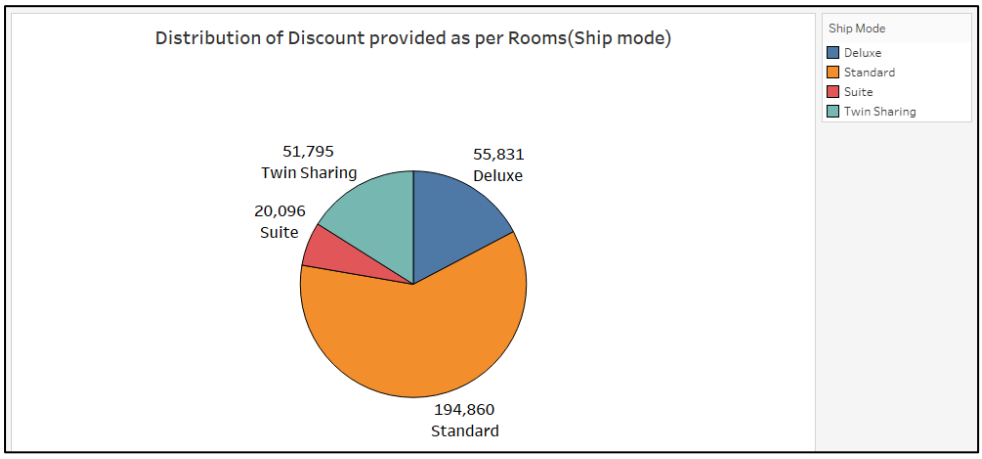
5.7 Analysis of effect of ship mode (type of room) on KPIs

The hospital provides different room categories to patients and the facilities and charges for the patient vary accordingly. So, there is a room for having an effect on KPIs due to different types of room services provided. It is mentioned as ship mode and there are four different ship modes available as per the data which are Standard, Deluxe, Twin sharing and Suite.
The distribution is as per follows in the period 2015-2018.

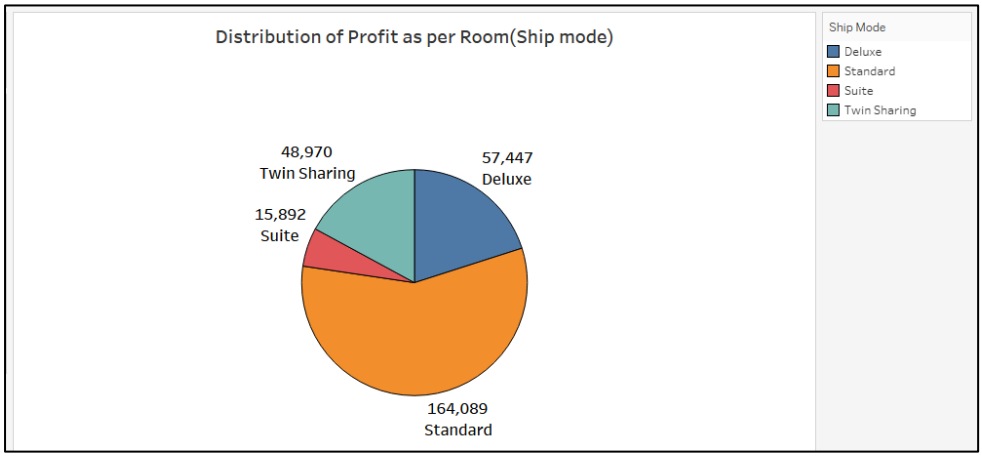


It is straightforward that 80% of the services to the patients are provide in the ship mode category of standard and deluxe. So, we can assume they hold a major weight to influence the KPIs.

5.2.1 Effect on Total Discount

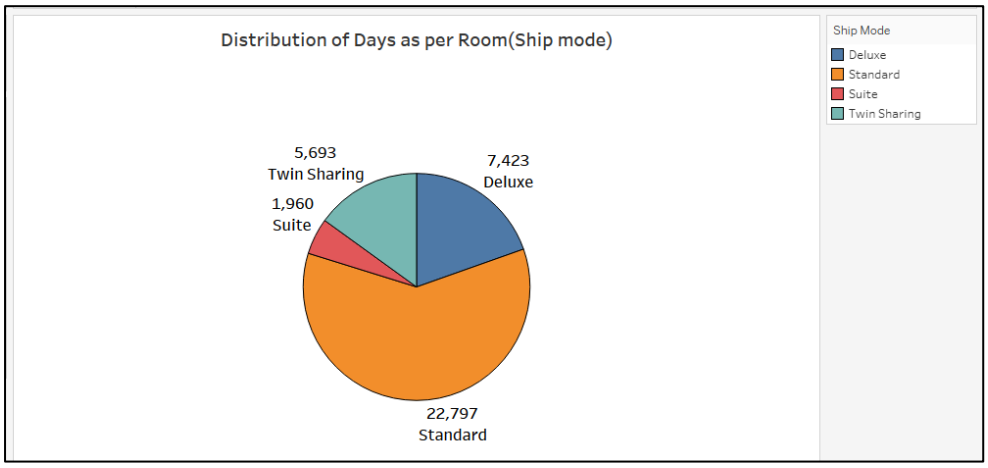


5.2.2 Effect on Overall Profit



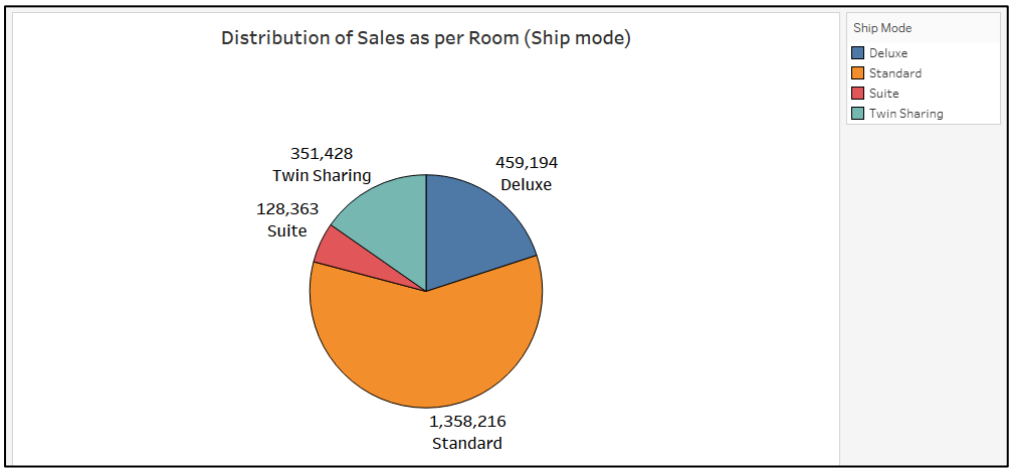
The patients in Standard ship mode generate as much as 60% profits in direct proportion to discount availed.

5.2.3 Effect on the Total Quantity (Days)



The patients in Standard ship mode contribute as much as 60% to the KPI of total quantity in direct proportion to discount availed.

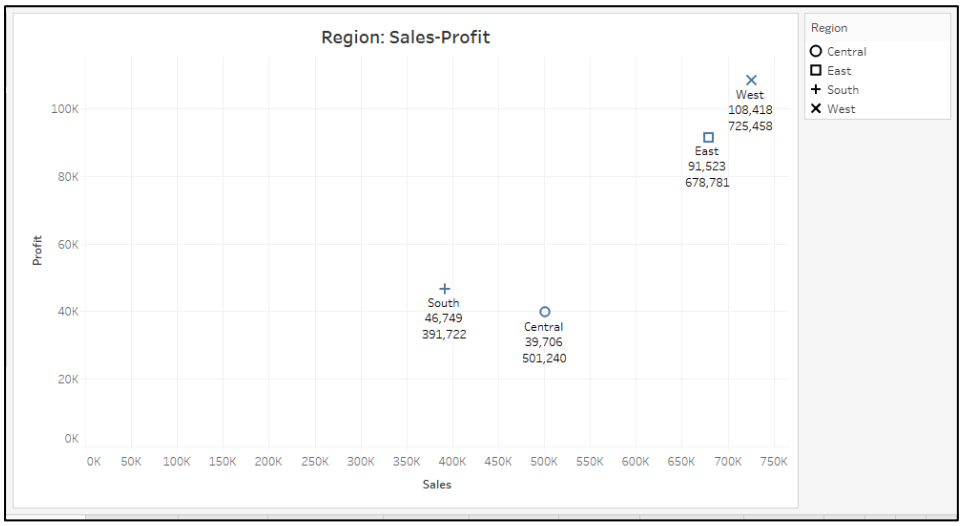
5.2.4 Effect on the Total Sales



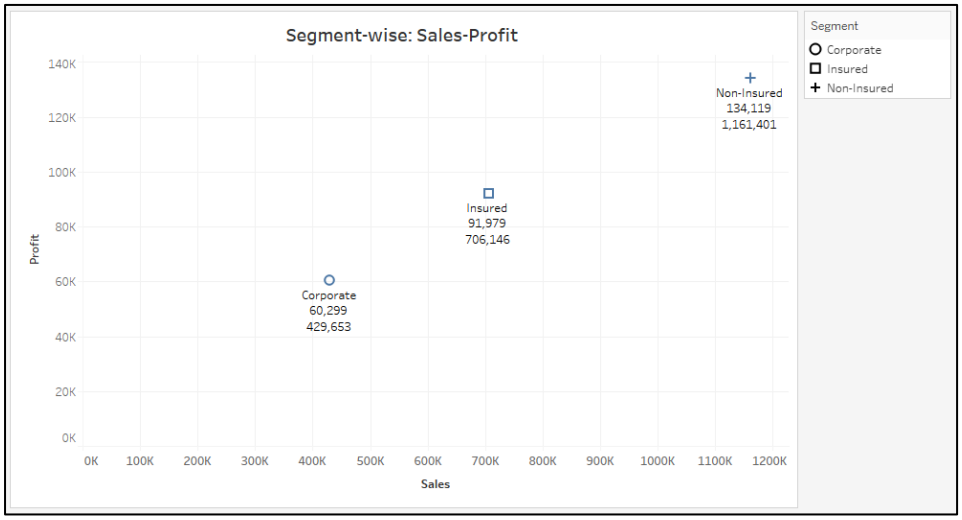
The patients in Standard ship mode contribute as much as 60% sales in direct proportion to discount availed. It can be very clearly stated that all the KPIs are directly proportional when studying variation on it due to ship mode.

5.8 Direction of Movement of KPIs in relation with each other

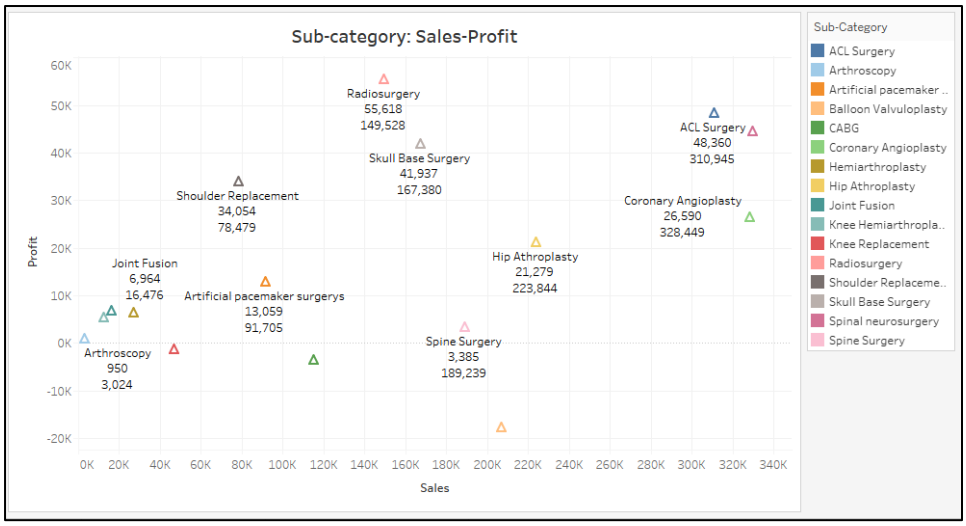
5.8.1 Direction of Movement of Profit in relation to Sales

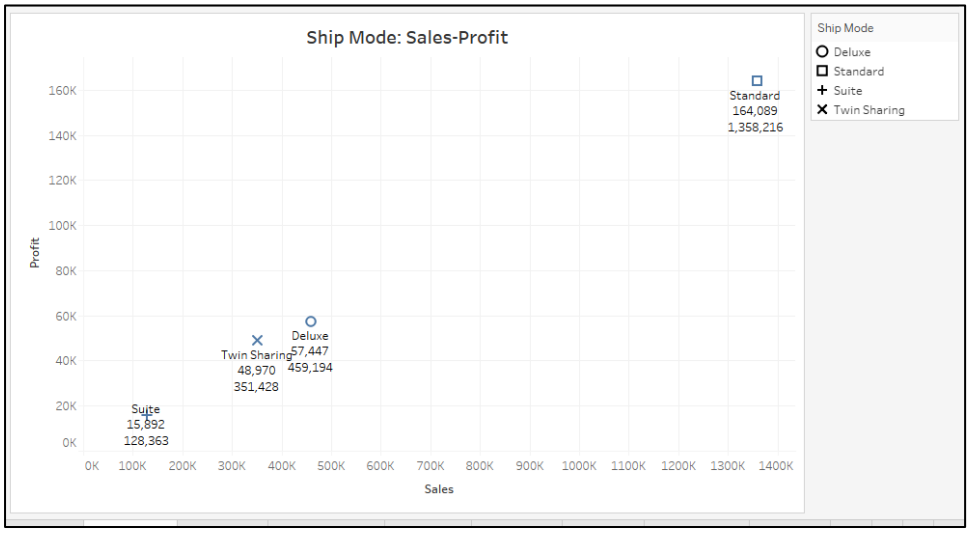


As deduced before, except in central region sales and profit are in direct proportion.



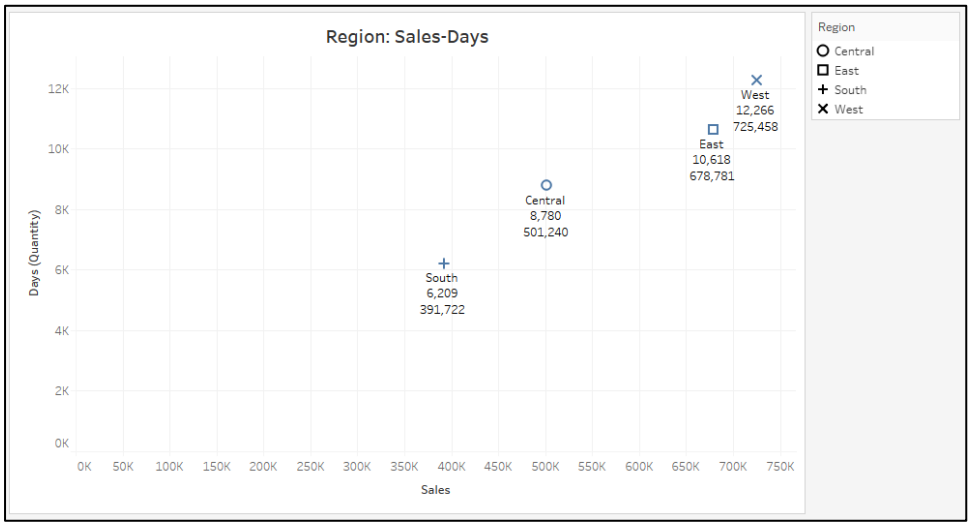
As a confirmation, higher profits are generated in case of higher sales.

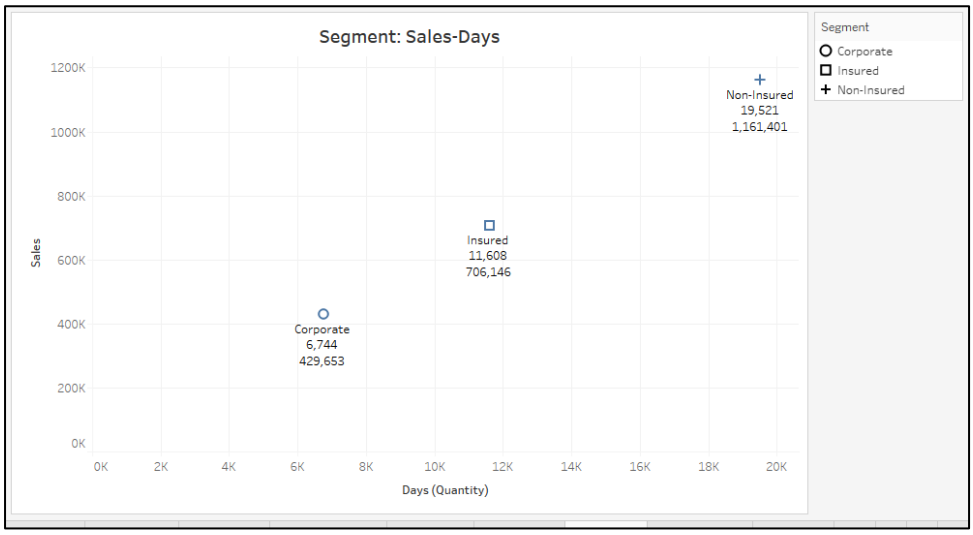


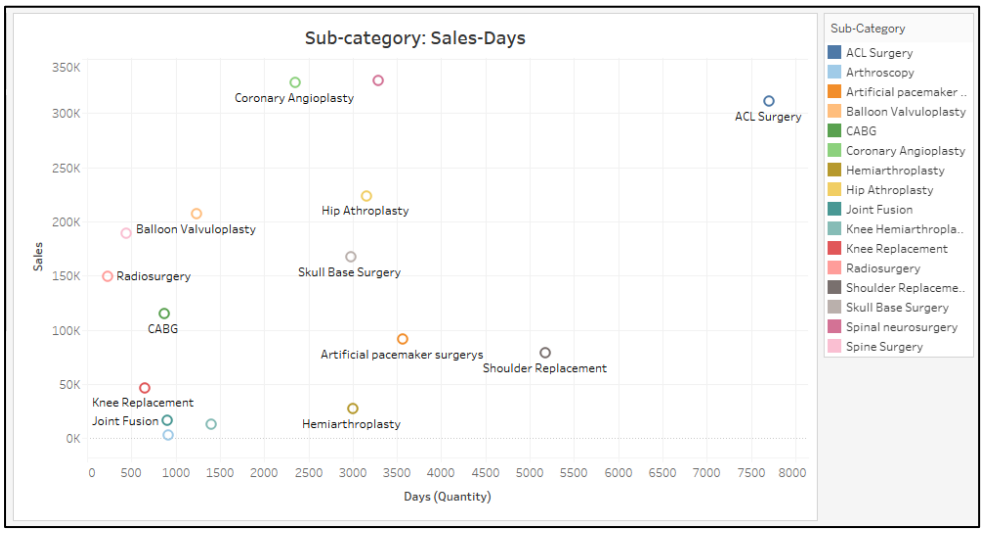


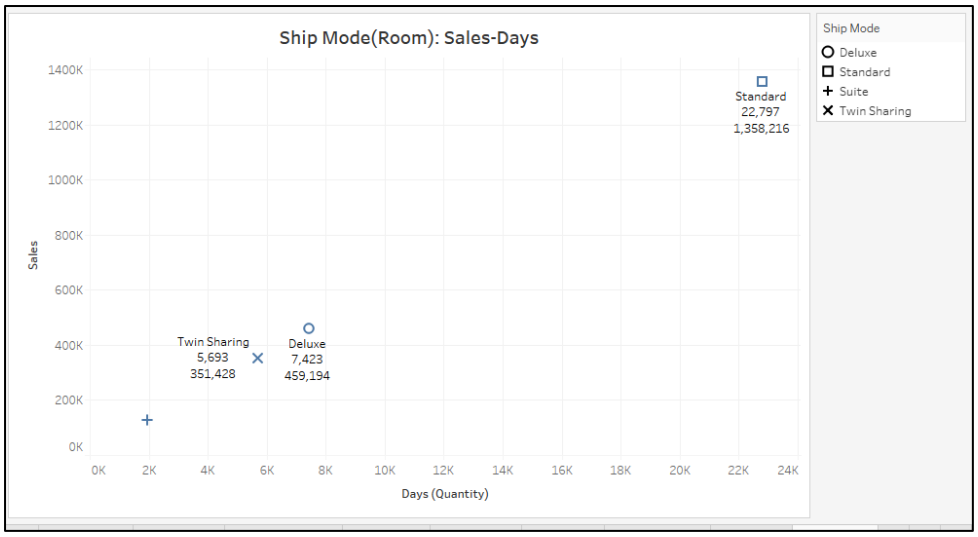
It can be deduced sales and profit are in direct proportion on a high level study, but a low level study shows that the relationship between sales and profit is majorly disturbed due to certain sub-category of procedures.

5.8.2 Direction of Movement of Sales in relation with Total quantity (Days)



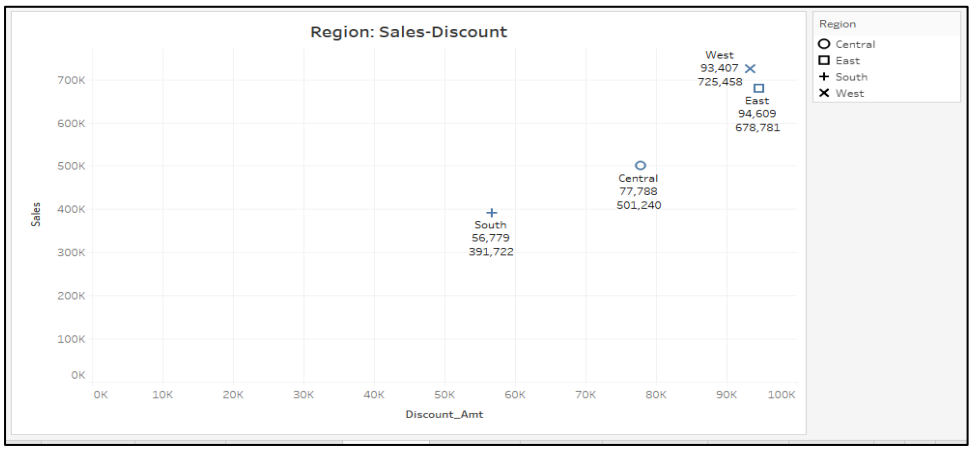


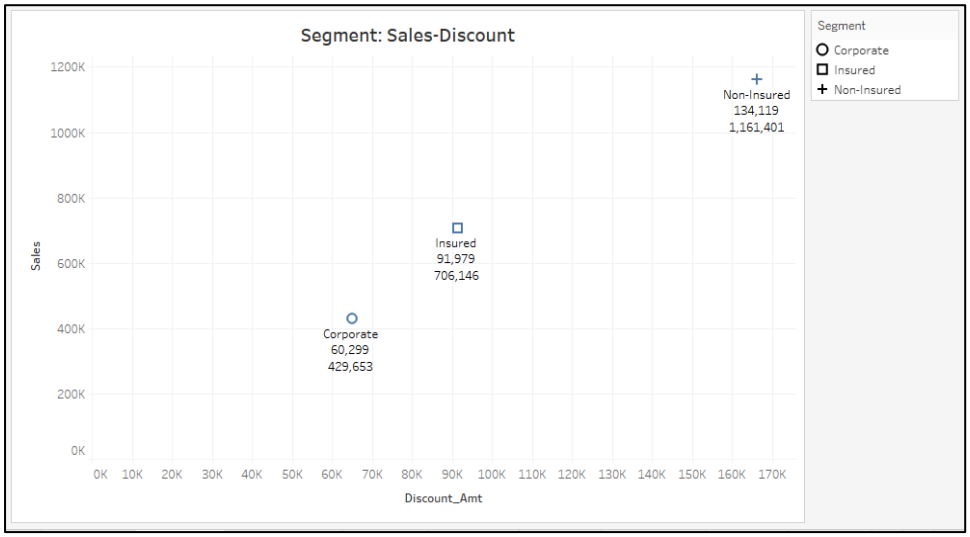


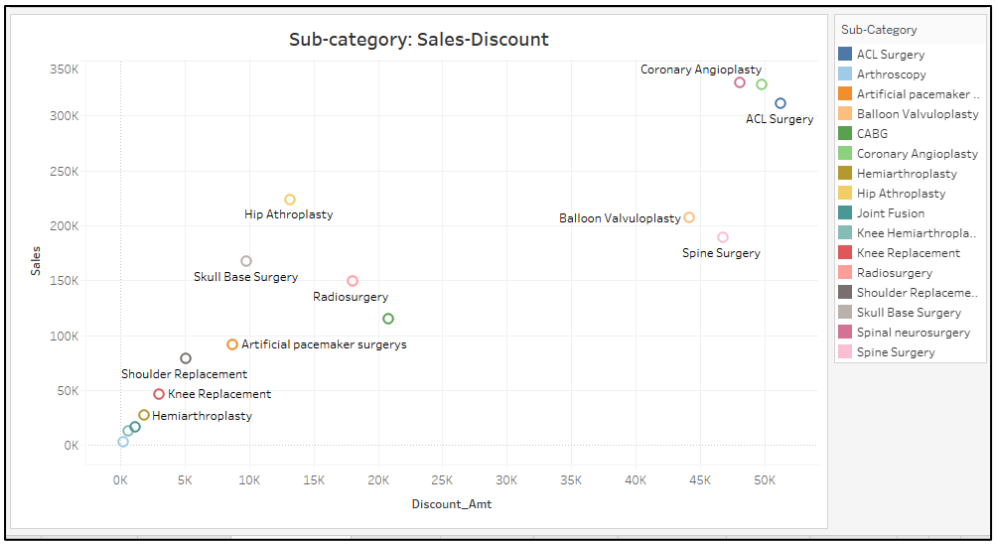


The KPI total quantity and sales are in direct proportion except when considering sub-category of procedures.

5.8.3 Direction of Movement of Sales in relation with Discount offered for the procedures

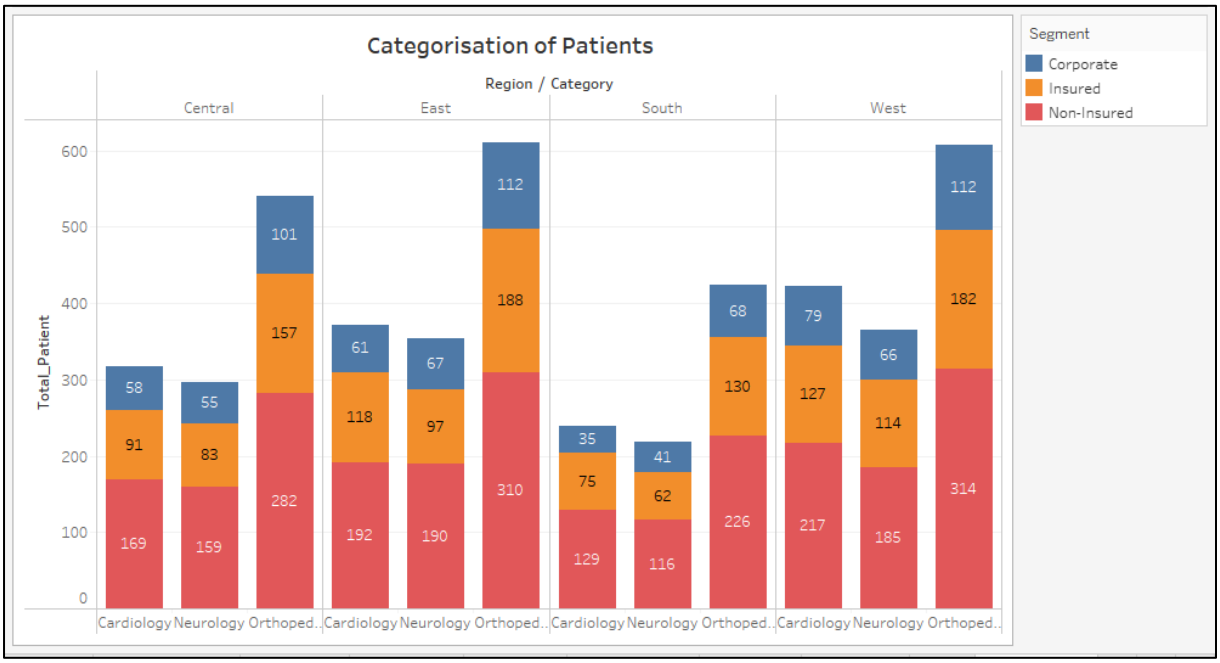






As with the relation between total quantity and sales, the discount and sales also have same relation which is directly proportional which is skewed only by sub-category of procedures.

5.9 Categorisation of Patients in the period 2015-2018



The study of above chart of distribution of various number of patients shows clear indications which are as follows.

- There is similar distribution of category of procedures in all regions.
- All regions have maximum procedures under the category of orthopaedics.
- Major patients fall in the category of non-insured segment.

SUGGESTIONS

The following suggestions are being made keeping in mind the major concern of optimising the KPIs.

1. Non-insured segment of patients contributes more to the KPIs and also occupy majority of the patient's pool. So, the policies could alter in favouring them.
2. Orthopaedics category of patients occupy majority in all regions and this category of procedure is in favour of good financial health of hospitals in all regions. They could be prioritised to boost financial health of a hospital.
3. The ending months of November and December contribute to more to all four KPIs. More discounts could be offered this time to increase sales and thereby profits.
4. All four KPIs are in direct proportion in all states and there is high amount of disparity among values of KPIs in all states. All the KPIs perform better on West coast better in one state only. Steps should be taken to remove interstate disparity.
5. It is observed that the most distorting factor in the performance of KPIs is the sub-category of procedure. It should be a focal major focal point while finalising the policies to improve the KPIs.

To enhance the financial health of the hospital without compromising service quality, consider implementing strategic cost-cutting measures through operational efficiency improvements considering above pointers could be helpful. Leverage data insights to identify areas for resource optimization and negotiate cost-effective vendor contracts. Additionally, explore innovative revenue streams, such as telehealth services or partnerships, to diversify income sources and ensure sustained financial stability.

CONCLUSION

In conclusion, this data visualization project has illuminated critical insights into the financial performance of US hospitals and the avenues for enhancing service delivery.

By analysing diverse datasets encompassing sales, profits, discount and operational days, we've discerned valuable patterns and correlations. The interactive dashboards and visual representations created provide a comprehensive view that empowers hospital administrators to make informed decisions.

This project not only contributes to optimizing financial strategies but also serves as a roadmap for improving service quality, thereby fostering a holistic approach to healthcare management in the United States. The synthesis of financial and service-related insights positions hospitals to navigate challenges effectively, ultimately leading to a more sustainable and patient-centric healthcare ecosystem.

REFERENCES

DATA AND STATISTICS

i. https://drive.google.com/file/d/1XsNol-NdHhudke8RjIaOd2Ws_yctniA3/view?usp=sharing

PROGRAMMING AND VISUALISATION

i. https://stackoverflow.com/questions/51165589/data-analytics-using-python
ii. https://realpython.com/
iii. https://www.datacamp.com/tutorial
iv. https://medium.com/analytics-vidhya/tagged/blog
v. https://pandas.pydata.org/pandas-docs/version/0.15/tutorials.html
vi. https://www.tutorialspoint.com/matplotlib/matplotlib_tutorial.pdf
vii. https://www.tutorialspoint.com/seaborn/seaborn_tutorial.pdf
viii. https://www.geeksforgeeks.org/tableau-tutorial/
ix. https://help.tableau.com/current/guides/get-started-tutorial/en-us/get-started-tutorial-next.htm