PROJECT REPORT

Date	**17 November, 2022**
Team ID	**PNT2022TMID21554**
Project Name	**Project -Analytics For Hospital'sHealth-Care Data**

Analytics For Hospitals' Health-Care Data

1. INTRODUCTION

1.1 Project Overview:

Researchers faces issues when they are dealing with large datasets as there is Depicting a diversity of opinions and experiences embedded within patient-generated information (not standard data)

Health Researchers and Students are not able to Extract useful Information's due to lack of data made available publicly as Many hospitals are not sharing health care data being mindful with patients' privacy.

Issues with system functionality, including poor user interfaces and fragmented displays, delayed care delivery. Issues with system access, system configuration, and software updates also delayed care.

1.2 Purpose:

The goal is to accurately predict the Length of

Stay for each patient on a case-by-case basis so that the Hospitals can use this information for optimal resource allocation and better functioning.

2.Literature Survey:

S.NO	TITLE OF THE PAPER	AUTHOR	METHODS	OBSERVATION
1.	Data analytics in healthcare: promise and potential	Wullianallur Raghupathi And Viju Raghupathi	The paper describes the nascent field of big data analytics in healthcare, discusses the benefits, outlines an architectural framework and	Health data volume is expected to grow dramatically in the years ahead. Comparative effectiveness research to determine more clinically relevant and cost-effective ways to

			methodology, describes examples reported in the literature, briefly discusses the challenges, and offers conclusions.	diagnose and treat patients. Big data analytics in healthcare is evolving into a promising field for providing insight from very large data sets and improving outcomes while reducing costs. Its potential is great; however there remain challenges to overcome. The paper provides a broad overview of big data analytics for healthcare researchers and practitioners.
2.	Big Data Analytics in Healthcare:	Ashwin Belle,Raghuram Thiagarajan ,Fatemeh Navidiand Kayvan Najarian	The rapidly expanding field of big data analytics has started to play a pivotal role in the evolution of healthcare practices and research. It has provided tools to accumulate, manage, analyze, and assimilate large volumes of disparate, structured, and unstructured data produced by current healthcare systems. Big data analytics has been recently applied towards aiding the process of care delivery and disease exploration.	Big data analytics which leverages legions of disparate, structured, and unstructured data sources is going to play a vital role in how healthcare is practiced in the future. One can already see a spectrum of analytics being utilized, aiding in the decision making and performance of healthcare personnel and patients. Here we focused on three areas of interest: medical image analysis, physiological signal processing, and genomic data processing. The exponential growth of the volume of medical images forces computational scientists to come up with innovative solutions to

				process this large volume of data in tractable timescales. Medical image analysis, signal processing of physiological data, and integration of physiological and "-omics" data face similar challenges and opportunities in dealing with disparate structured and unstructured big data sources.
3.	Big data analytics in healthcare: a systematic literature review.	Sayantan khanra, Amandeep Dhir and A.K.Ajmul Islam.	The current study performs a systematic literature review (SLR) to synthesise prior research on the applicability of big data analytics (BDA) in healthcare. The SLR examines the outcomes of 41 studies, and presents them in a comprehensive framework. The findings from this study suggest that applications of BDA in healthcare can be observed from five perspectives, namely, health awareness among the general public, interactions among stakeholders in the healthcare ecosystem, hospital	The current study intended to address four research questions related to the application of BDA in healthcare. These questions have been answered following a standard protocol for reviewing resources from key databases. The study has identified the gaps in the existing literature and provided an actionable research agenda for future research on the utilisation of big data in the healthcare sector. However, despite the significant contributions of this current study, it suffers from three main limitations: first, book chapters, magazine articles, and thesis studies have been excluded from the scope of this study; second, journal articles and conference studies not available in English

			management practices, treatment of specific medical conditions, and technology in healthcare service delivery.	were not considered; third, studies not available in the four databases were not reviewed unless they appeared in the forward and backward searches. Future research is invited to overcome these limitations.

2.3Problem Statement Definition:

Public hospitals has some main challenges such as deficient in infrastructure, deficient in manpower, unmanageable patient load and etc.,so peoples can be benefited if these problems are solved adhering to certain software or some notes to maintain all. Govt. Hospitals facing data management due to lack of IT trained staffs.

Private/Small Health sectors cannot store and analyzed large data set it consumes lots of money and time.

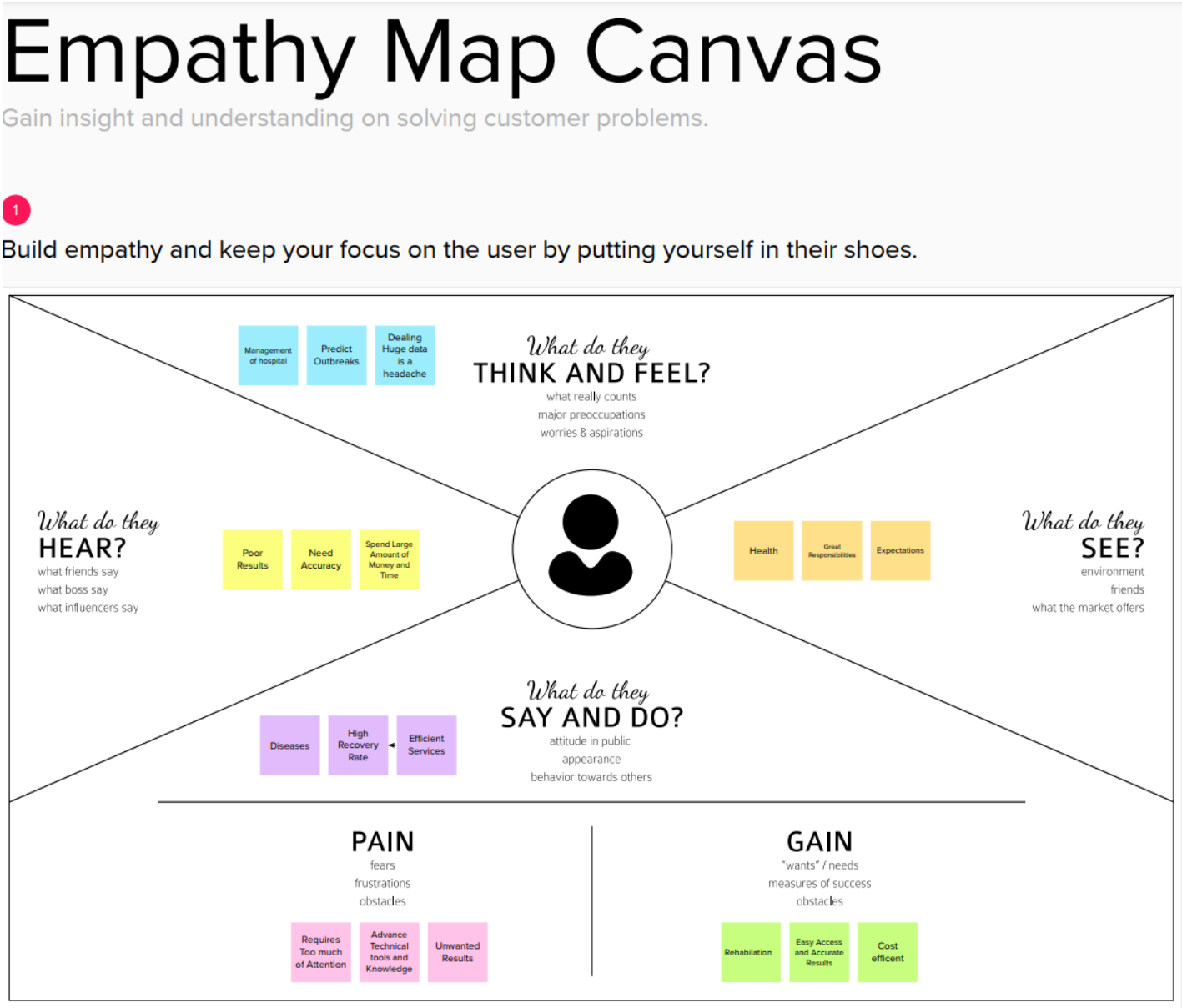
Researchers faces issues when they are dealing with large datasets as there is Depicting a diversity of opinions and experiences embedded within patient-generated information(not standard data)

Health Researchers and Students are not able to Extract useful Information's due to lack of data's made available publicly as Many hospitals are not sharing health care data being mindful with patients privacy.

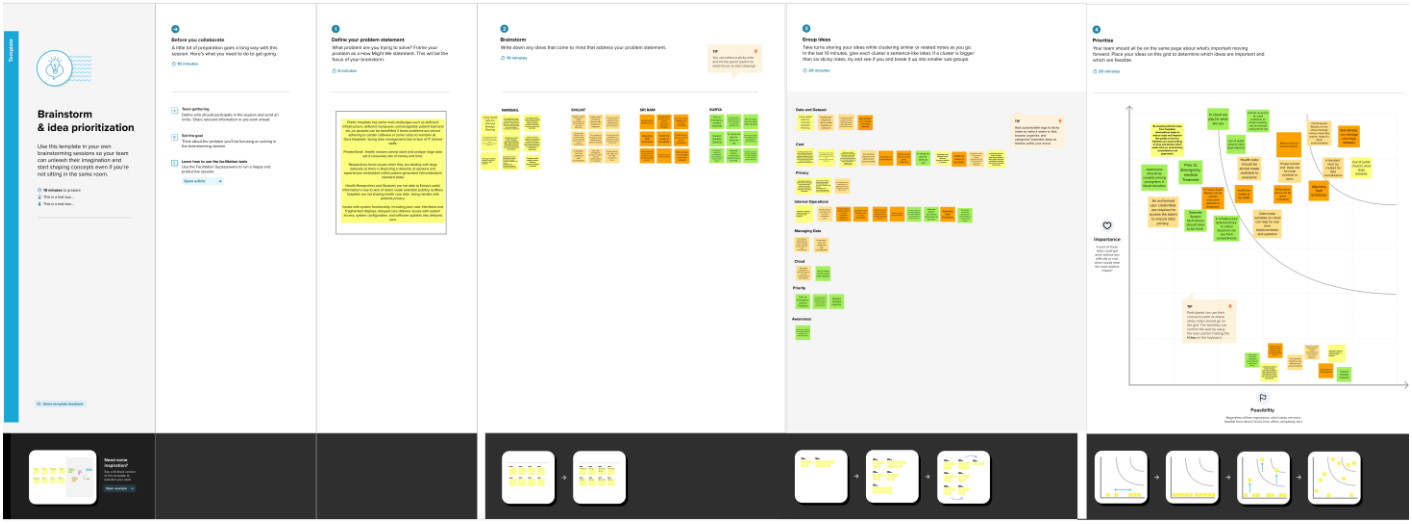
Issues with system functionality, including poor user interfaces and fragmented displays, delayed care delivery. Issues with system access, system configuration, and software updates also delayed care.

3.IDEATION & PROPOSED SOLUTION:

3.1 Empathy Map Canvas:



3.2Ideation & Brainstorming:



3.3Proposed Solution:

SNo.	Parameter	Description
1.	Problem Statement (Problem to be solved)	Analytics for Hospitals Health-Care Data: Hospitals have some main challenges such as deficient infrastructure, deficient manpower, unmanageable patient load, etc., so people can benefit if these problems are solved by adhering to certain software or some notes to maintain them all. The goal is to accurately predict the Length of Stay for each patient on a case-by-case basis so that the Hospitals can use this information for optimal resource allocation and better functioning. The length of stay is divided into 11 different classes ranging from 0-10 days to more than 100 days.
2.	Idea/Solution description	We are able to predict the length of stay of patients with data from the movement they entered the hospital and are diagnosed with an accuracy of ~70%. Such a model has the ability to profoundly improve hospital management and patient well-being.

		Also, we can predict the LOS with big data analytic tools within a Python interface such as Spark, AWS clusters, SQL query optimization, and dimensionality reduction techniques.
3.	Novelty/Uniqueness	Length of stay in the hospital differs based upon the critical in their health situation it can range between 2 to 3 days or even upto 10- 20 days so based on the exploratory analysis of various patients we can accurately predict the length of stay of patients and can allocate optimum resource allocation
4.	Social Impact/Customer satisfaction	With Exploratory analysis using different methods to predict the length of stay creates a way to our patients to know the vacancy of beds in the hospitals and also paved a way in their critical times to secure their better life
5.	Business Model (Revenue Model)	Using this model, the usage of length of stay of patients in the hospitals has increased among the people and it is free of cost to get the details about the vacancy. It doesn't affect the revenue model.
6.	Scalability of the Solution	It is a easily scalable method using dataset of previous patients we can able to predict the LOS • Increased productivity among the users • Decreased stress level • Possibility of getting the detailed list of vacancy

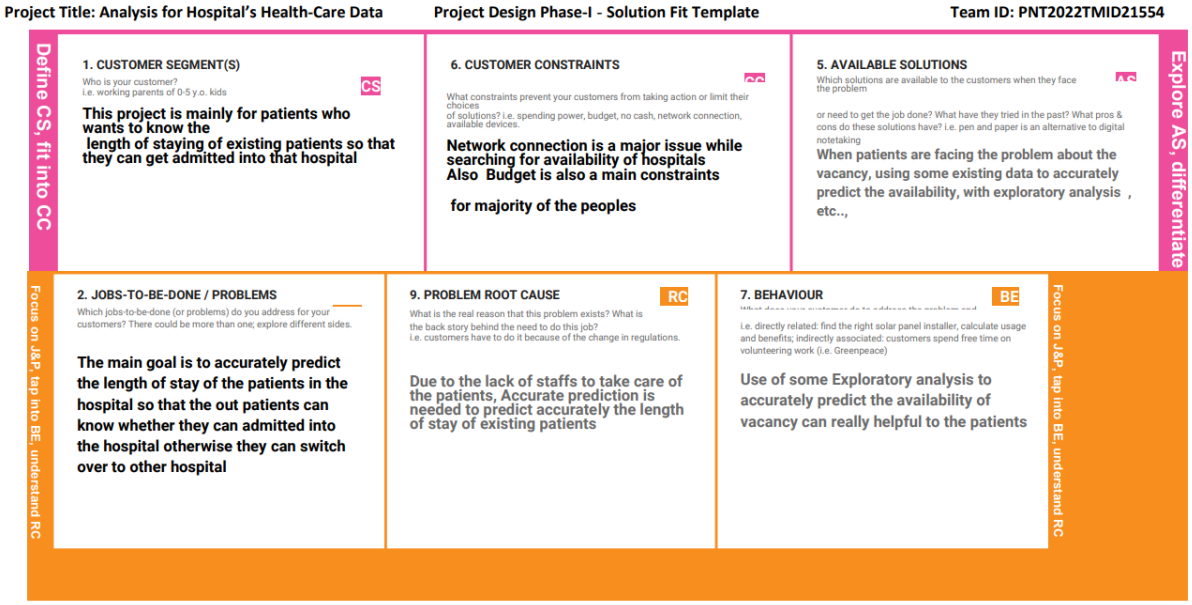
SNo	Parameter	Description
1	Problem Statement (Problem to be solved)	Bioinformatic It is a powerful technology to manage, query, and analyze big data in life sciences. Here The sequence of issues are faced such as the data problems such as representation , storage and retrieval , analysis (statistics, artificial intelligence, optimization, etc.) and biology problems such as sequence analysis, structure or function prediction, data mining, etc.
2	Idea/Solution description	We can modify the database with different data categorizing on the basis of different properties from genotype to phenotype. Sequence alignment Database similarity search Motif finding(Gene finding Comparative genomics DNA methylation)

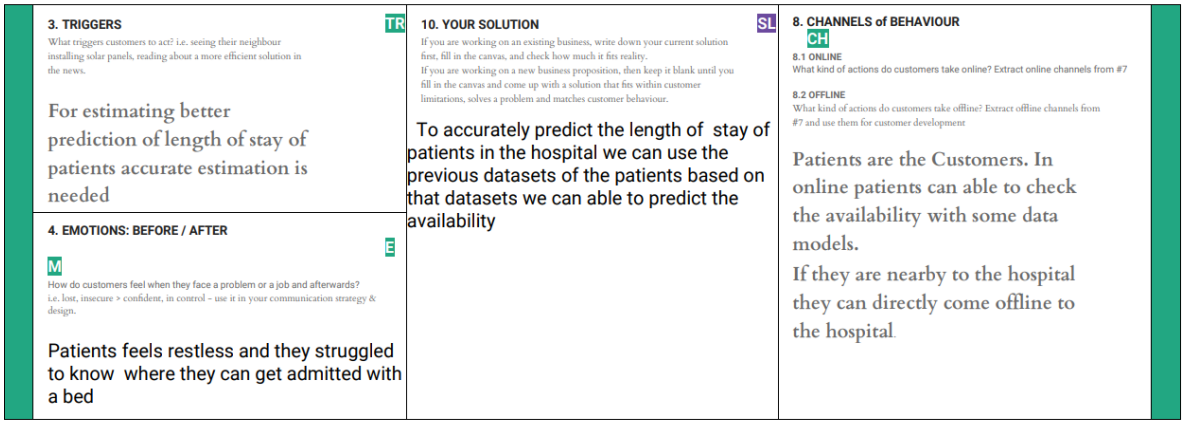
3	Novelty/Uniqueness	Can Form Molecular Networks: Protein interaction networks Transcription regulation networks Metabolic & signaling networks
4	Social Impact/Customer satisfaction	Easy Manage Data Collect,Store ,Ensure Security. Interpret Data: Create Data Models Enhance Interoperability
5	Business Model (Revenue Model)	The main objectives of this technique is a top-down, holistic, data-driven, genome-wide, and systems approach that generates new hypotheses, finds new patterns, and discovers new functional elements and with all those features it fits best in business implementation.
6	Scalability of the Solution	• Limited Resources better results. • Boost productivity in results. • Much more user friendly and can further be improved in future.

SNo	Parameter	Description
1.	Problem Statement (Problem to be solved)	Peoples from all over the world who were busy at their work who needs a way to maintain their health, Analysis of level of stay in hospitals for various treatments, so that the users can be benefited in their busy schedule
2	Idea/Solution description	Health sector has improved in many factors. Nowadays people can maintain their health at their place of stay. Like smartphones, smart watches and many more gadgets came to make our lives easier. We can monitor our blood pressure, no of distance walked and etc.
3	Novelty/Uniqueness	Health gadgets have come in a large amount and it is handy to the users. From their place of stay they can monitor their health like breathing capacity, heart beat rate and many more
4	Social Impact/Customer satisfaction	It has impacted from rich to poor, from educated to common peoples all got to know the usage of gadgets and it is really helpful to the peoples who are far away from the hospital.

		Now they can easily monitor their heart beat rate, blood pressure, etc.
5	Business Model (Revenue Model)	In the modern world. the usage of healthcare gadgets has increased and peoples are more likely to buy these gadgets and so it has increased the revenue model in the market
6	Scalability of the Solution	Scalability is up to the usage of peoples by getting their commands about the gadgets. We can scale the gadgets to their usage in terms of speedness, accuracy in predicting the results and even more.

3.4Proposed Solution Fit:





4.REQUIREMENT ANALYSIS:

Functional Requirements:

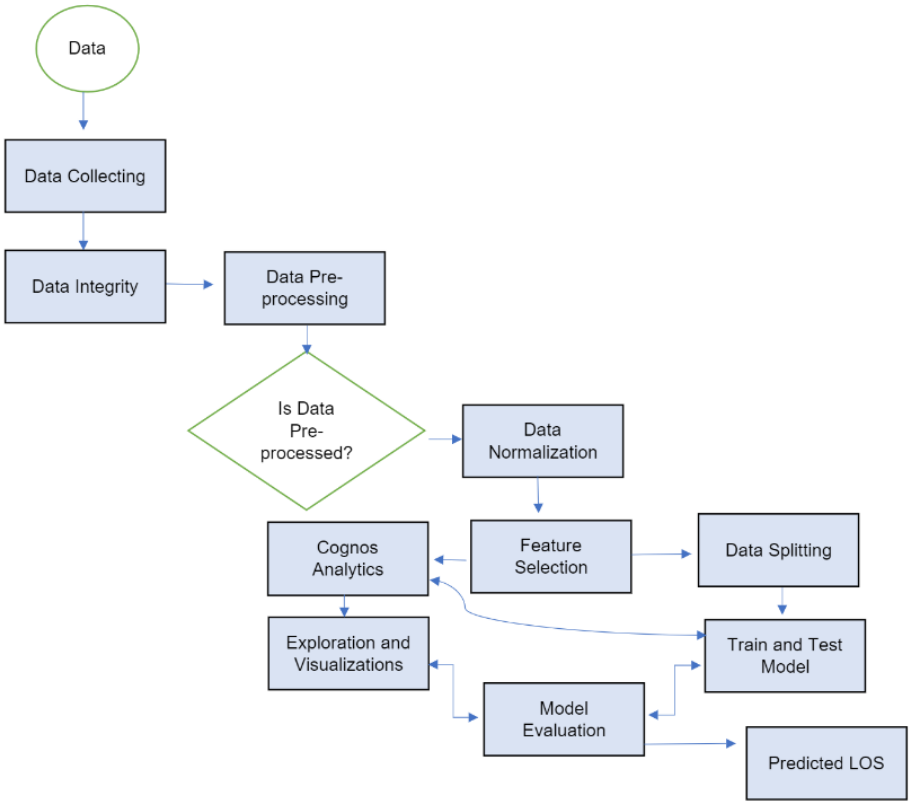
FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	The User can have own ID to get registered in the portal or Dashboard
FR-2	Analyzing the Hospital's data	The user can analyse the data related to hospitals such as availability of beds Number of existing patients All the users can analyze through the hospital's portal
FR-3	Prediction of length of stay	After analysing the data of the particular Hospital's we can able to predict the length of stay of each and every patients in terms with their severity of diseases
FR-4	Get the user response	After the prediction of Length of stay of each patients We can improve the prediction accuracy by obtaining feedback from the users
FR-5	Monitoring user response	All the responses will then be stored in the database for future reference We can store the data and can visualize through charts like bar chart , pie chart end etc..
FR-6	Monitoring System accuracy	System should be monitored periodically to prevent errors in this way we can keep our system in robotic manner

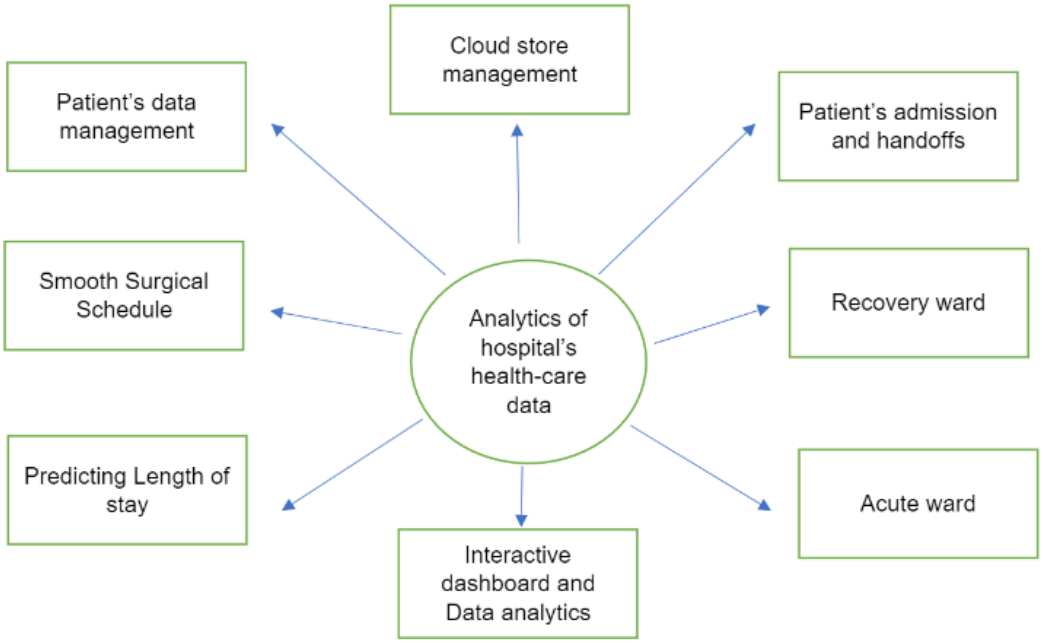
Non-functional Requirements:

FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	The goals of the users are easily accomplished quickly by interactive design and less error.
NFR-2	**Security**	The dataset is accessed only by the administrators and the user's input is encrypted and it is protected.
NFR-3	**Reliability**	It works without a failure at the prediction time because of less bugs in the code it is because of using good trained data.
NFR-4	**Performance**	It supports at most 1000 patients queries at a time and after prediction is done it will be fastly communicated to the users.
NFR-5	**Availability**	The application should be available 24/7.
NFR-6	**Scalability**	The application should support all browser types and it can handle maximum users.

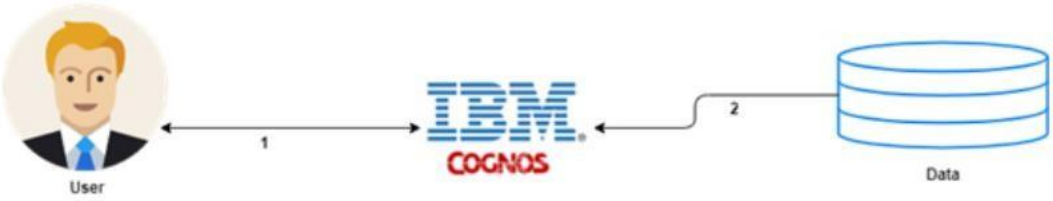
5.PROJECT DESIGN:

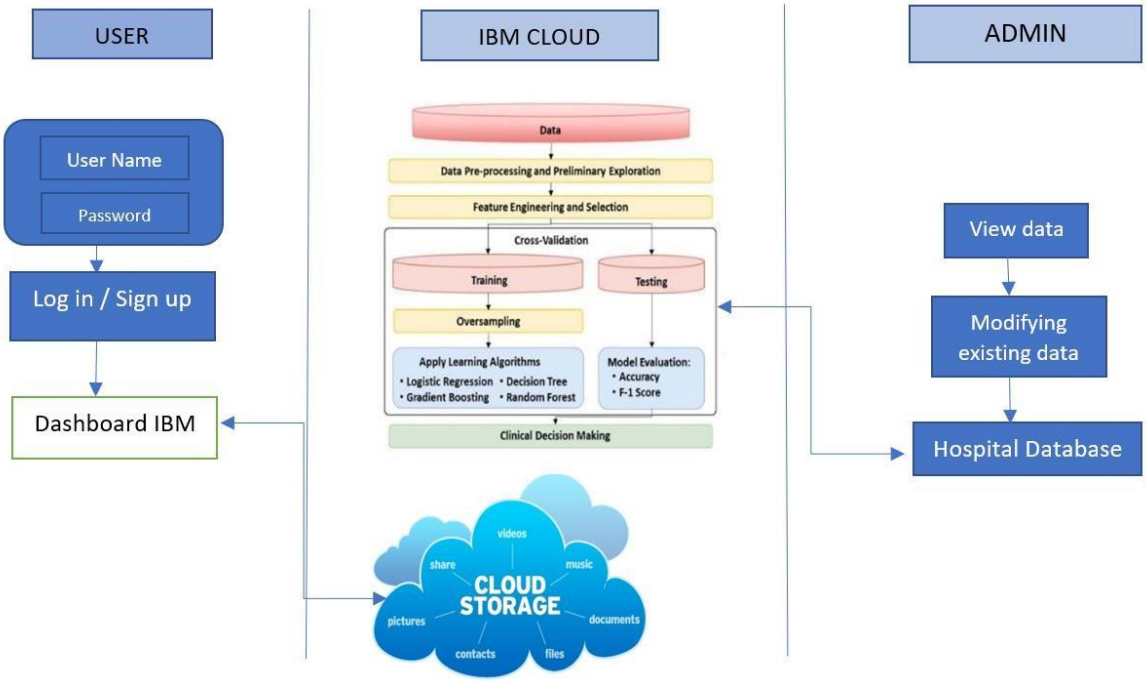
5.1Data Flow Diagrams:





5.2Solution & Technical Architecture:





5.3User Stories:

User Stories

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Mobile user)	Registration	USN-1	As a user, I can sign in for the application entering my email, password	I can access my account dashboard	High	Sprint-1
			As a user, I can sign up for the application through google		High	Sprint-1
		USN-2	As a user, I will receive confirmation email	I can click confirmation email	High	Sprint-1
		USN-3	As a user, I can register for the application through Instagram	I can access the dashboard with Instagram Login	Low	Sprint-2
		USN-4	As a user, I can register for the application to know the length of stay of patients	I can access patient's data based on their severeness	High	Sprint-1
	Login	USN-5	As a user, I can log into the application by entering user id & password		High	Sprint-1
		USN-6	As a user I can explore the all the details regarding to hospital in my dashboard		Medium	Sprint-2
Administrator	Updating data	USN-7	As a user I can collect the data of all the patients from their attendees and I can store it	I can check the gathered data and can store it	High	Sprint-1
		USN-8	As a Administrator I can categorize the patients based on their risk factors		High	Sprint-1
Customer (Web User)	Accessing the resources	USN-9	As a user I can get all the informations in the dashboard	These resources cannot be accessed by others but only me	High	Sprint -1
Customer tools	Tools	USN -10	As a user I can explore the data through data visualization tools like Cognos analytics		High	Sprint 2

6. PROJECT PLANNING & SCHEDULING:

6.1 Sprint Planning & Estimation:

Product Backlog, Sprint Schedule, and Estimation (4 Marks)

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Analysing , Visualizing , and Data Preparation Hospital health care data	USN-1	As a user, I want to collect the details regarding to hospitals data As a patient, I want to visualize the hospital health care data.	10	Medium	SHUJAT HUSSAIN
Sprint-1		USN-2	As a patient, I want to load the data, and data has to be prepared	5	High	SRI RAM PRASAD S
Sprint-2	Exploration of data	USN-3	As a patient/user I want to explore all the details in the given in the dataset	5	High	SURYA S
Sprint -2		USN-4	As a user, I want to visualize all the details in the dataset in different formats	5	Medium	NAMGAIL DORJAY
Sprint-3	Prediction of LOS	USN-5	As a patient/user I want an interactive dashboard to understand the data easily	5	High	SRI RAM PRASAD S
		USN - 6	As a patient, I want to predict length of stay in the hospitals	8	High	SHUJAT HUSSAIN
Sprint-4	An website to Know the Details about the hospitals	USN -7	As a Patient, I want an website to know the details about the hospitals, availability of beds and etc.. through that website	10	High	SRI RAM PRASAD S, SURYA S
	Admin Dashboard	USN-8	As an admin I want to create a report.	5	Medium	NAMGAIL DORJAY

6.2 Sprint Delivery Schedule

Project Tracker, Velocity & Burndown Chart: (4 Marks)

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	15	6 Days	24 Oct 2022	29 Oct 2022	10	
Sprint-2	10	6 Days	31 Oct 2022	05 Nov 2022		
Sprint-3	13	6 Days	07 Nov 2022	12 Nov 2022		
Sprint-4	15	6 Days	14 Nov 2022	19 Nov 2022		

Velocity:

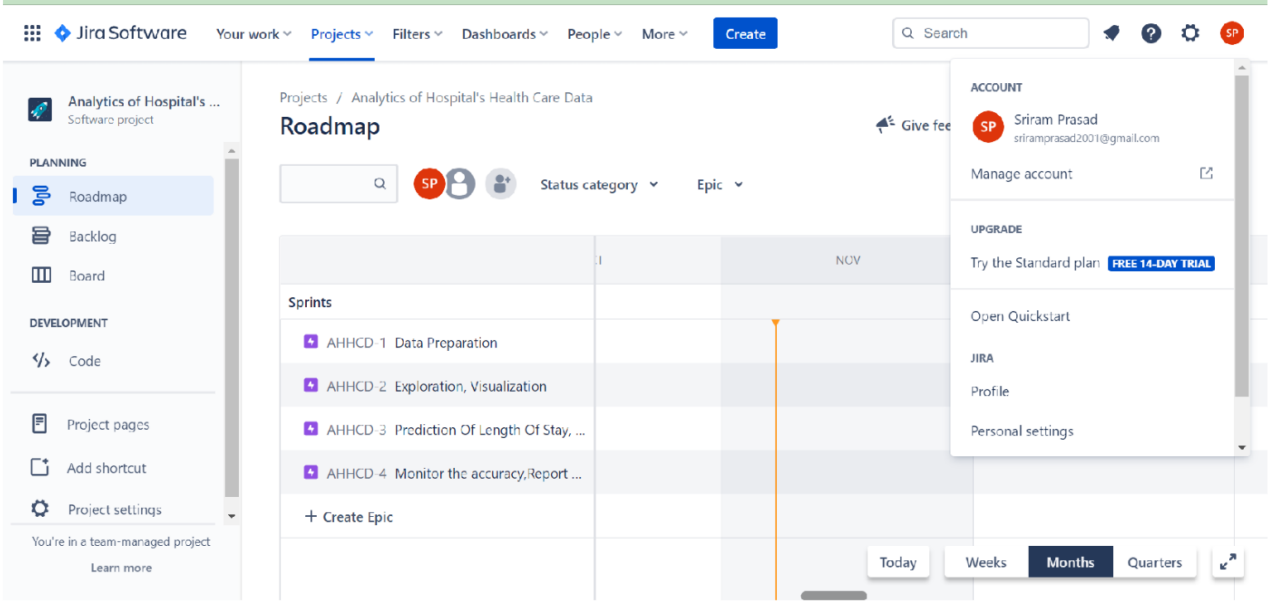
Imagine we have a 10-day sprint duration, and the velocity of the team is 20 (points per sprint). Let's calculate the team's average velocity (AV) per Iteration unit (story points per day)

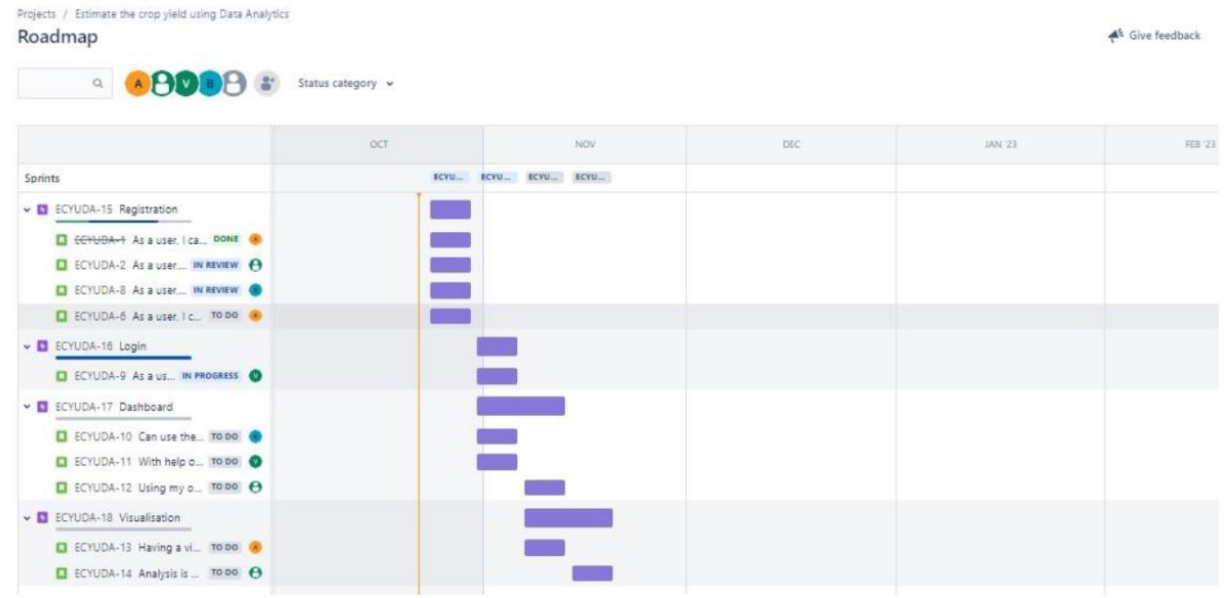
$$AV = \frac{sprint\ duration}{velocity} = \frac{20}{10} = 2$$

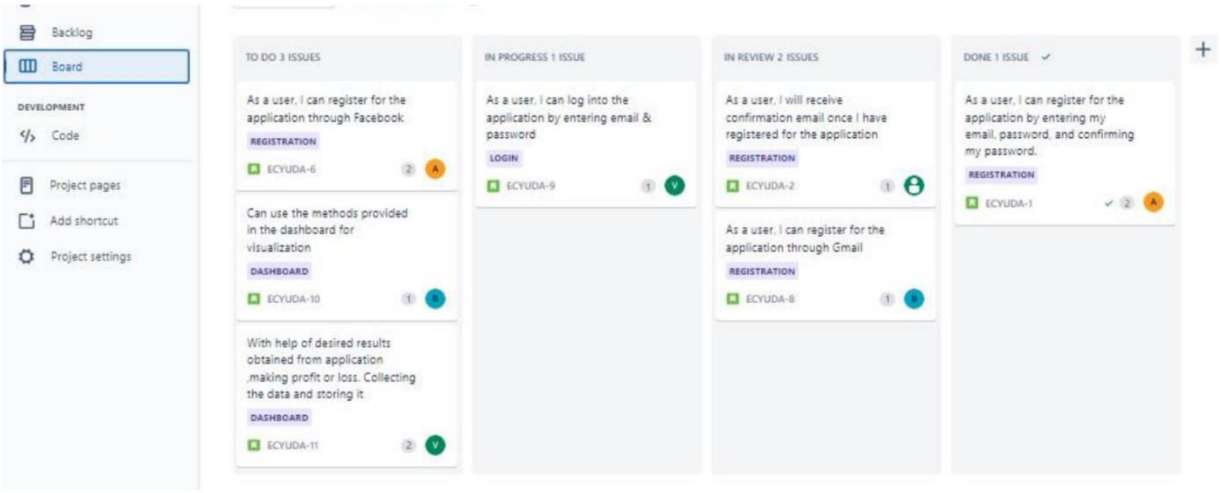
Sprint	Total Story points	Sprint duration	Average velocity
Sprint -1	15	6 days	15/6=2.5
Sprint -2	10	6 days	10/6=1.67
Sprint -3	13	6 days	13/6=2.16
Sprint -4	15	6 days	15/6=2.5

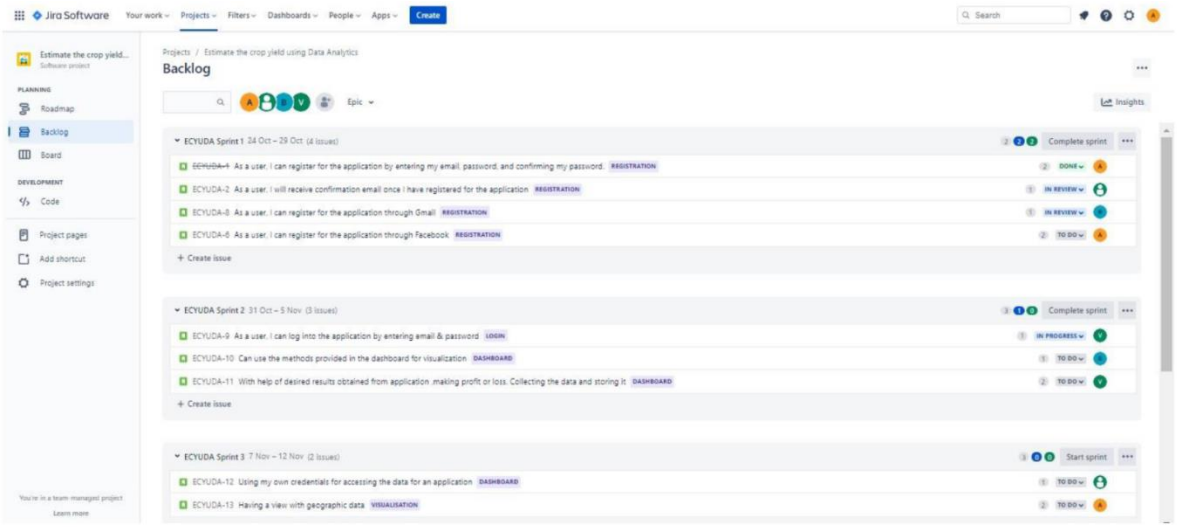
6.3Reports from JIRA

TOOL USED: JIRA SOFTWARE OF ATLASSIAN






Backlog
Board
DEVELOPMENT
Code
Project pages
Add shortcut
Project settings
TO DO 3 ISSUES
As a user, I can register for the application through Facebook
REGISTRATION
ECYUDA-6
Can use the methods provided in the dashboard for visualization
DASHBOARD
ECYUDA-10
With help of desired results obtained from application ,making profit or loss. Collecting the data and storing it
DASHBOARD
ECYUDA-11
IN PROGRESS 1 ISSUE
As a user, I can log into the application by entering email & password
LOGIN
ECYUDA-9
IN REVIEW 2 ISSUES
As a user, I will receive confirmation email once I have registered for the application
REGISTRATION
ECYUDA-2
As a user, I can register for the application through Gmail
REGISTRATION
ECYUDA-8
DONE 1 ISSUE
As a user, I can register for the application by entering my email, password, and confirming my password.
REGISTRATION
ECYUDA-1

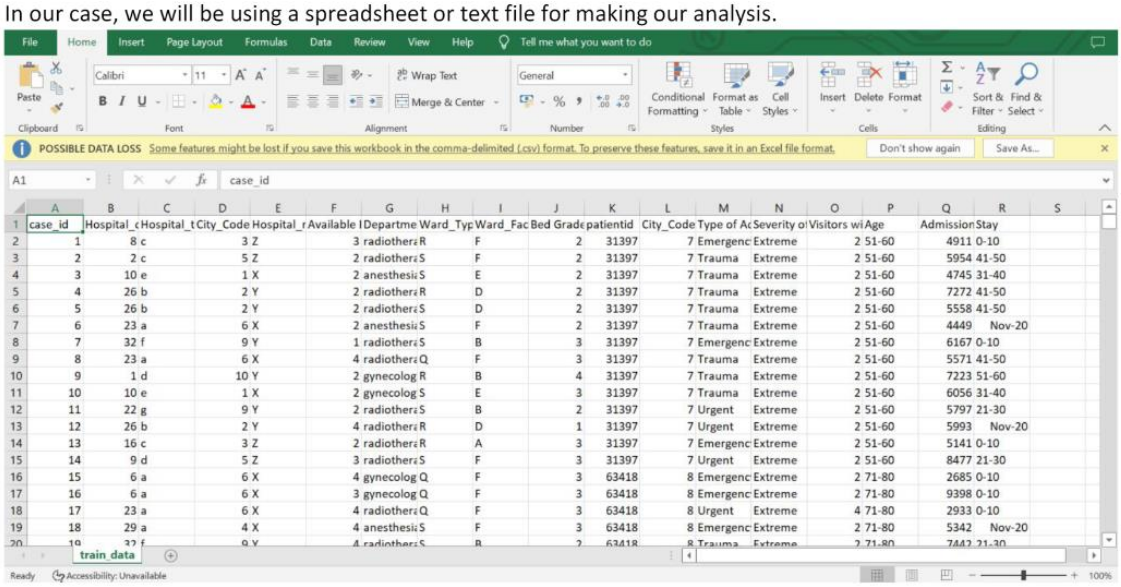

Jira Software
Your work
Projects
Filters
Dashboards
People
Apps
Create
Search
Estimate the crop yield...
Software project
PLANNING
Roadmap
Backlog
Board
DEVELOPMENT
Code
Project pages
Add shortcut
Project settings
You're in a team-managed project
Learn more
Projects / Estimate the crop yield using Data Analytics
Backlog
Epic
Insights
ECYUDA Sprint 1 24 Oct – 29 Oct (4 issues)
Complete sprint
ECYUDA-4 As a user, I can register for the application by entering my email, password, and confirming my password. REGISTRATION DONE
ECYUDA-2 As a user, I will receive confirmation email once I have registered for the application REGISTRATION IN REVIEW
ECYUDA-8 As a user, I can register for the application through Gmail REGISTRATION IN REVIEW
ECYUDA-6 As a user, I can register for the application through Facebook REGISTRATION TO DO
+ Create issue
ECYUDA Sprint 2 31 Oct – 5 Nov (3 issues)
Complete sprint
ECYUDA-9 As a user, I can log into the application by entering email & password LOGIN IN PROGRESS
ECYUDA-10 Can use the methods provided in the dashboard for visualization DASHBOARD TO DO
ECYUDA-11 With help of desired results obtained from application ,making profit or loss. Collecting the data and storing it DASHBOARD TO DO
+ Create issue
ECYUDA Sprint 3 7 Nov – 12 Nov (2 issues)
Start sprint
ECYUDA-12 Using my own credentials for accessing the data for an application DASHBOARD TO DO
ECYUDA-13 Having a view with geographic data VISUALISATION TO DO

7.CODING & SOLUTIONING

7.1Feature 1:

Exploratory Data Analysis:

In our case, we will be using a spreadsheet or text file for making our analysis.



case_id	Hospital_c	Hospital_t	City_Code	Hospital_r	Available I	Departme	Ward_Typ	Ward_Fac	Bed Grade	patientid	City_Code	Type of Ac	Severity o	Visitors wi	Age	Admission	Stay
1	8	c	3	Z	3	radiothera	R	F	2	31397	7	Emergenc	Extreme	2	51-60	4911	0-10
2	2	c	5	Z	2	radiothera	S	F	2	31397	7	Trauma	Extreme	2	51-60	5954	41-50
3	10	e	1	X	2	anesthesia	S	E	2	31397	7	Trauma	Extreme	2	51-60	4745	31-40
4	26	b	2	Y	2	radiothera	R	D	2	31397	7	Trauma	Extreme	2	51-60	7272	41-50
5	26	b	2	Y	2	radiothera	S	D	2	31397	7	Trauma	Extreme	2	51-60	5558	41-50
6	23	a	6	X	2	anesthesia	S	F	2	31397	7	Trauma	Extreme	2	51-60	4449	Nov-20
7	32	f	9	Y	1	radiothera	S	B	3	31397	7	Emergenc	Extreme	2	51-60	6167	0-10
8	23	a	6	X	4	radiothera	Q	F	3	31397	7	Trauma	Extreme	2	51-60	5571	41-50
9	1	d	10	Y	2	gynecolog	R	B	4	31397	7	Trauma	Extreme	2	51-60	7223	51-60
10	10	e	1	X	2	gynecolog	S	E	3	31397	7	Trauma	Extreme	2	51-60	6056	31-40
11	22	g	9	Y	2	radiothera	S	B	2	31397	7	Urgent	Extreme	2	51-60	5797	21-30
12	26	b	2	Y	4	radiothera	R	D	1	31397	7	Urgent	Extreme	2	51-60	5993	Nov-20
13	16	c	3	Z	2	radiothera	R	A	3	31397	7	Emergenc	Extreme	2	51-60	5141	0-10
14	9	d	5	Z	3	radiothera	S	F	3	31397	7	Urgent	Extreme	2	51-60	8477	21-30
15	6	a	6	X	4	gynecolog	Q	F	3	63418	8	Emergenc	Extreme	2	71-80	2685	0-10
16	6	a	6	X	3	gynecolog	Q	F	3	63418	8	Emergenc	Extreme	2	71-80	9398	0-10
17	23	a	6	X	4	radiothera	Q	F	3	63418	8	Urgent	Extreme	4	71-80	2933	0-10
18	29	a	4	X	4	anesthesia	S	F	3	63418	8	Emergenc	Extreme	2	71-80	5342	Nov-20

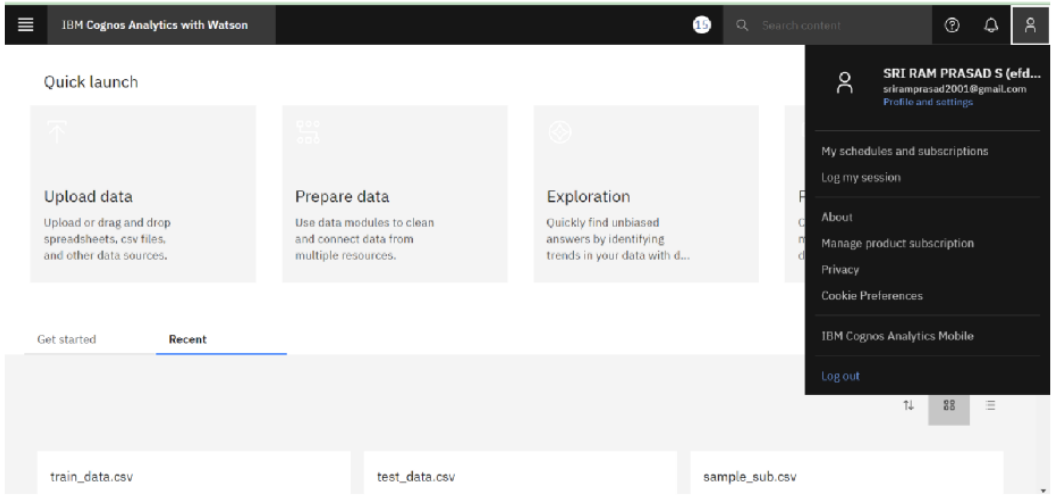
UNDERSTANDING THE DATASET

TEAM ID: PNT2022TMID21554

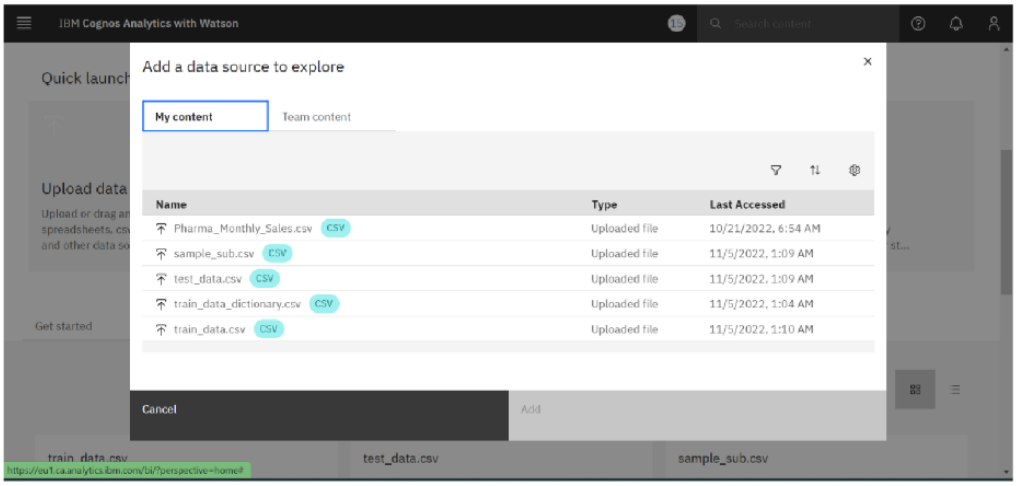
Column	Description
case_id	CaseID registered in Hospital
Hospital_code	Unique code for t he Hospital
Hospital_type_code	Unique code for the type of Hospital
City_Code_Hospital	City Code of the Hospital
Hospital_region_code	Region Code of the Hospital
Available Extra Rooms in Hospital	Number of Extra rooms available in the Hospital
Department	Department overlooking the case
Ward_Type	Code for the Ward type
Ward_Facility_Code	Code for the Ward Facility
Bed Giade	Condition of Bed in the Ward

Once uploaded the data, we need to prepare it.

Step-01: choose the .csv file for preparation

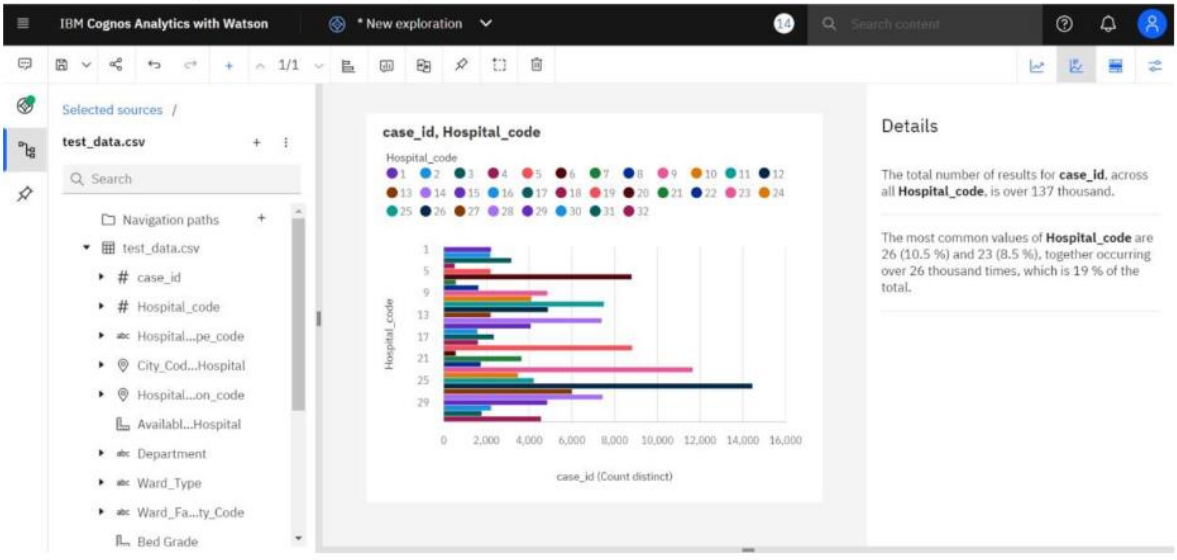


Step-02: then upload & then prepared file will like this,

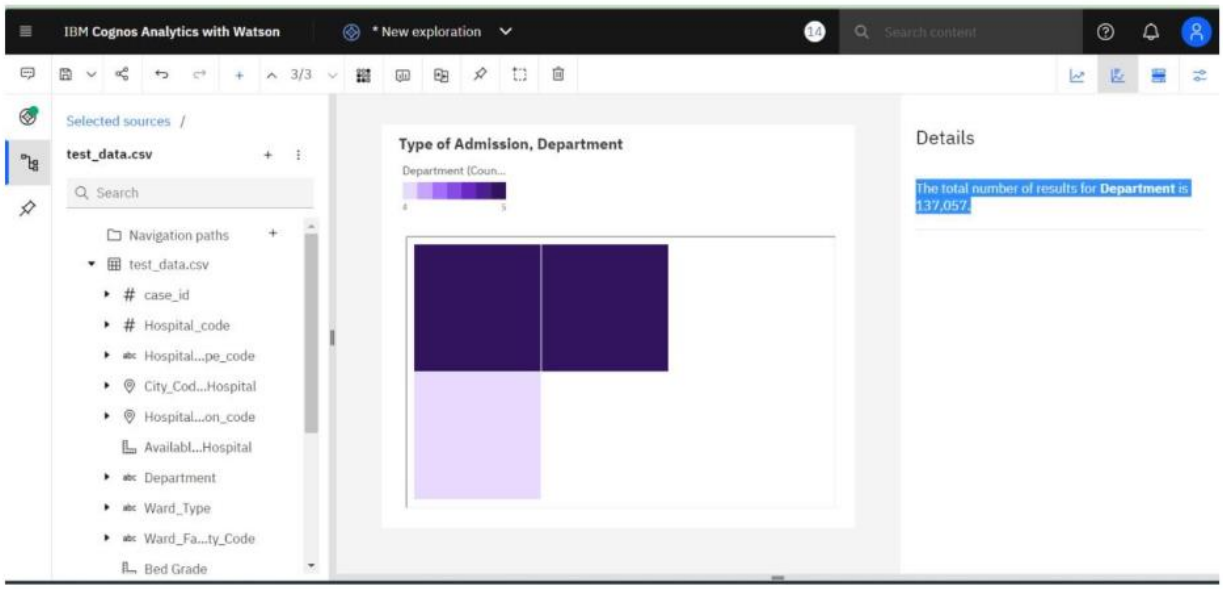


Data Visualisation

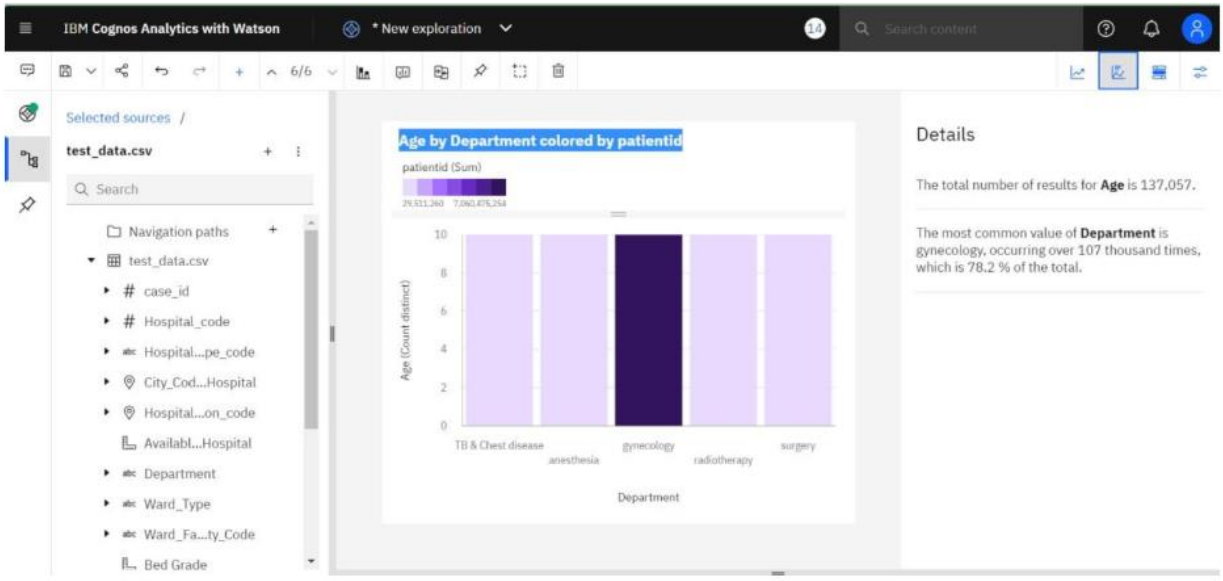
Case ID of each patient with Hospital Code:



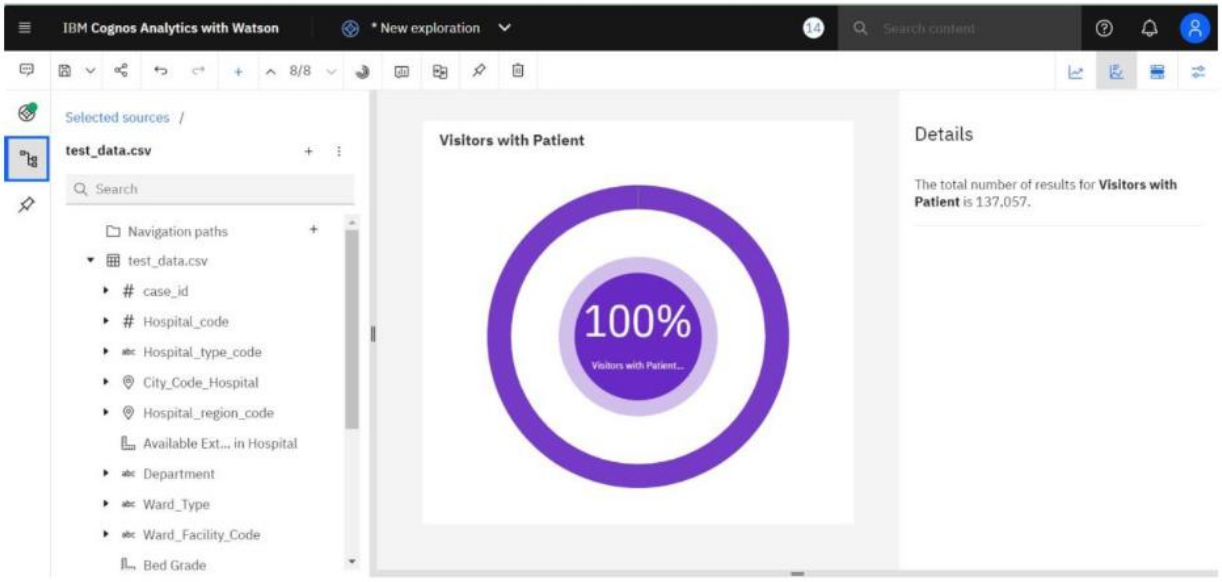
The total number of results for **Department** is 137.057.



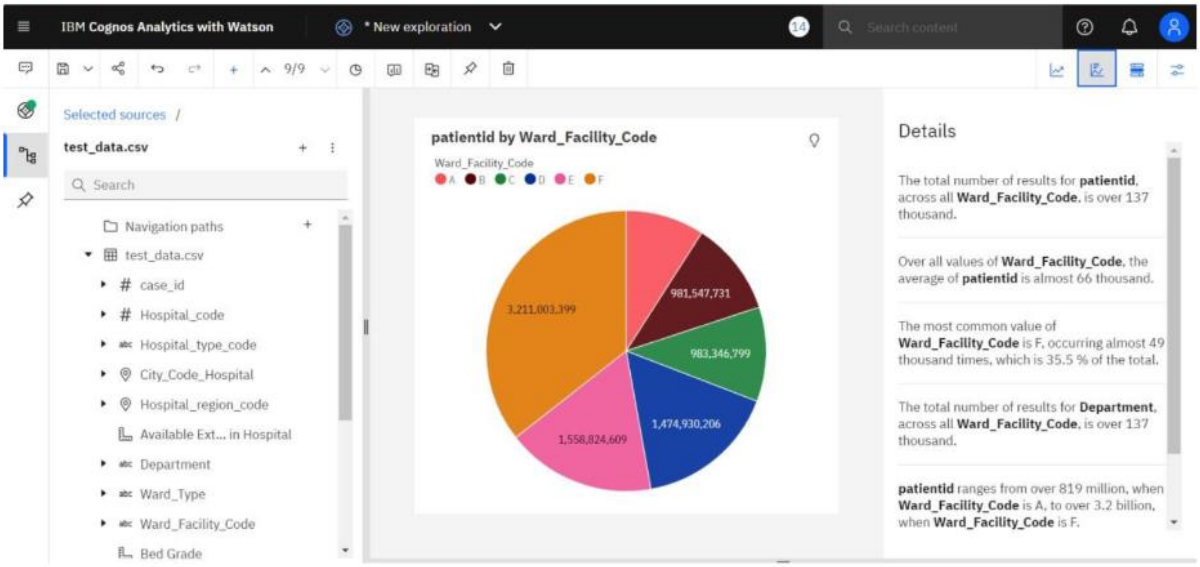
Age by Department colored by patientid



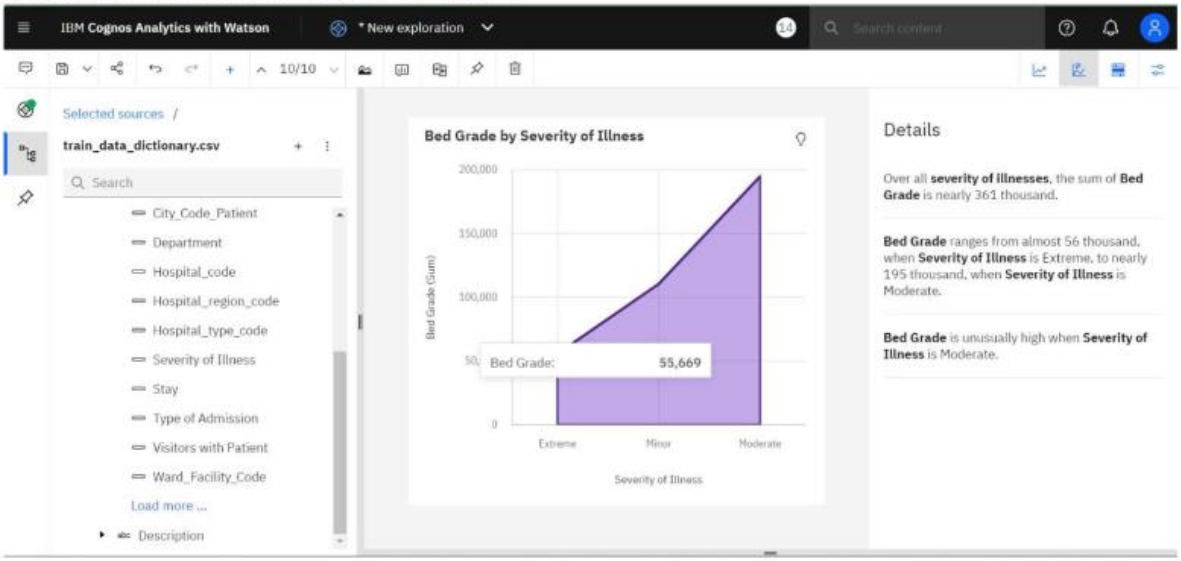
Visitors with Patient:



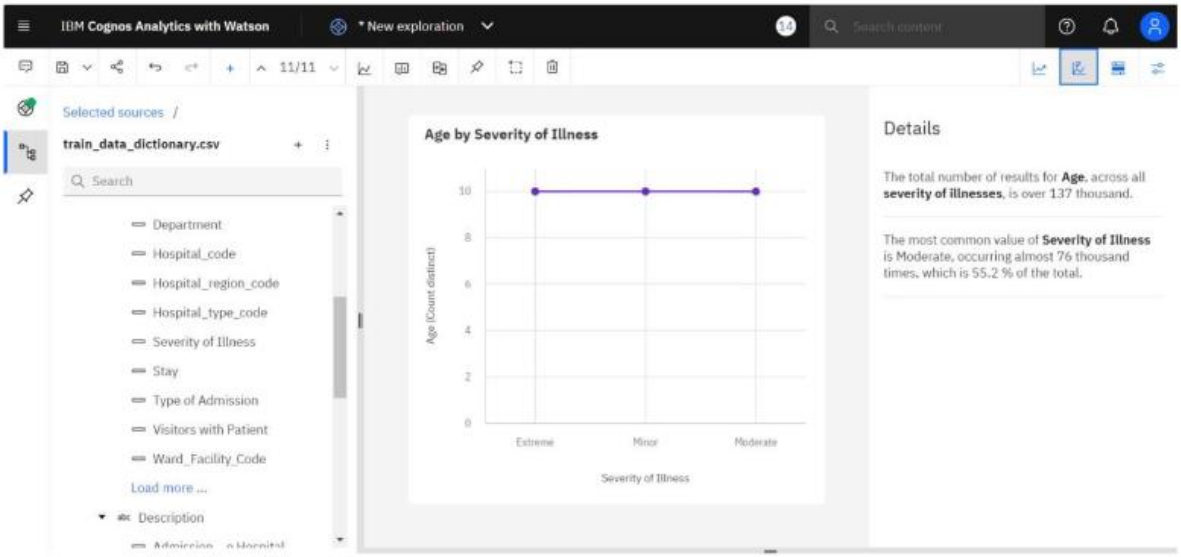
Patientid by Ward_Facility_Code



Availability of Beds by Severity of Illness:



Age by Severity of Illness



7.2 Feature 2:

Algorithm & Metric Evaluation:

In [2]:
```
from google.colab import files


uploaded = files.upload()
```

Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Saving train_data.csv to train_data.csv

In [3]:
```
import pandas as pd
import io

df = pd.read_csv(io.BytesIO(uploaded['train_data.csv']))
print(df)
```

```
        case_id  Hospital_code Hospital_type_code  City_Code_Hospital  \
0             1              8                  c                   3
1             2              2                  c                   5
2             3             10                  e                   1
3             4             26                  b                   2
4             5             26                  b                   2
...         ...            ...                ...                 ...
318433   318434              6                  a                   6
318434   318435             24                  a                   1
318435   318436              7                  a                   4
318436   318437             11                  b                   2
318437   318438             19                  a                   7

       Hospital_region_code  Available Extra Rooms in Hospital    Department  \
0                         Z                                 3  radiotherapy
1                         Z                                 2  radiotherapy
2                         X                                 2    anesthesia
3                         Y                                 2  radiotherapy
4                         Y                                 2  radiotherapy
...                     ...                               ...           ...
318433                    X                                 3  radiotherapy
318434                    X                                 2    anesthesia
318435                    X                                 3    gynecology
318436                    Y                                 3    anesthesia
318437                    Y                                 5    gynecology
```

In [10]:
```
df.Department.duplicated()
```

Out[10]:
```
0         False
1          True
2         False
3          True
4          True
          ...
318433     True
318434     True
318435     True
318436     True
318437     True
Name: Department, Length: 318438, dtype: bool
```

Counting duplicates and non-duplicates

In [11]:
```
df.Department.duplicated().sum()
```

Out[11]: 318433

number of non-duplicates

In [12]:
```
(~df.duplicated()).sum()
```

Out[12]: 318438

Dropping duplicate rows

In [13]:
```
df.drop_duplicates()
```

Dropping duplicate rows

In [13]:
```
df.drop_duplicates()
```

Out[13]:

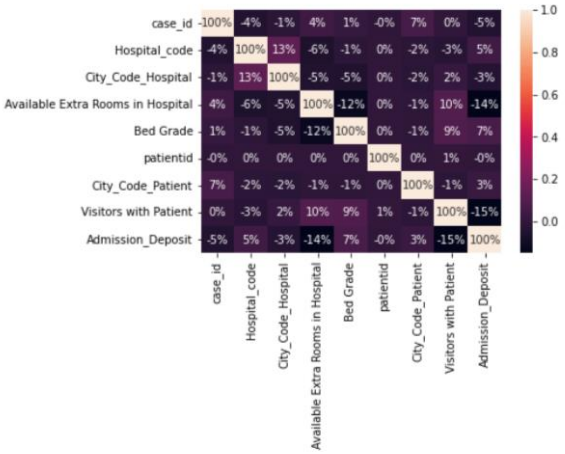
	case_id	Hospital_code	Hospital_type_code	City_Code_Hospital	Hospital_region_code	Available Extra Rooms in Hospital	Department	Ward_Type	Ward_Facility_Code	Bed Grade	patientid	City_Code_
0	1	8	c	3	Z	3	radiotherapy	R	F	2.0	31397	
1	2	2	c	5	Z	2	radiotherapy	S	F	2.0	31397	
2	3	10	e	1	X	2	anesthesia	S	E	2.0	31397	
3	4	26	b	2	Y	2	radiotherapy	R	D	2.0	31397	
4	5	26	b	2	Y	2	radiotherapy	S	D	2.0	31397	
...	...	...	...	...	...	...	...	...	...	...	...	...
318433	318434	6	a	6	X	3	radiotherapy	Q	F	4.0	86499	
318434	318435	24	a	1	X	2	anesthesia	Q	E	4.0	325	
318435	318436	7	a	4	X	3	gynecology	R	F	4.0	125235	
318436	318437	11	b	2	Y	3	anesthesia	Q	D	3.0	91081	
318437	318438	19	a	7	Y	5	gynecology	Q	C	2.0	21641	

318438 rows × 18 columns

In [15]:
```
import seaborn

seaborn.heatmap(df.corr(), annot=True, fmt='.0%')
```

Out[15]:



In [18]:
```
import numpy as np
import pandas as pd

# Input data files are available in the "../input/" directory.
# Let's input them into a Pandas DataFrame
train = pd.read_csv("train_data.csv")
```

SPRINT 3

TEAM ID - PNT2022TMID21554

PREDICTION OF LENGTH OF STAY

In [1]:
```
from google.colab import files


uploaded = files.upload()
```

Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Saving train_data.csv to train_data.csv

In [3]:
```
import pandas as pd
import io

df = pd.read_csv(io.BytesIO(uploaded['train_data.csv']))
print(df)
```

```
        case_id  Hospital_code Hospital_type_code  City_Code_Hospital  \
0             1              8                  c                   3
1             2              2                  c                   5
2             3             10                  e                   1
3             4             26                  b                   2
4             5             26                  b                   2
...         ...            ...                ...                 ...
318433   318434              6                  a                   6
318434   318435             24                  a                   1
318435   318436              7                  a                   4
318436   318437             11                  b                   2
318437   318438             19                  a                   7

       Hospital_region_code  Available Extra Rooms in Hospital    Department  \
0                         Z                                  3  radiotherapy
1                         Z                                  2  radiotherapy
2                         X                                  2    anesthesia
3                         Y                                  2  radiotherapy
4                         Y                                  2  radiotherapy
...                     ...                                ...           ...
318433                    X                                  3  radiotherapy
318434                    X                                  2    anesthesia
318435                    X                                  3    gynecology
318436                    Y                                  3    anesthesia
318437                    Y                                  5    gynecology
```

	Ward_Type	Ward_Facility_Code	Bed Grade	patientid	City_Code_Patient
0	R	F	2.0	31397	7.0
1	S	F	2.0	31397	7.0
2	S	E	2.0	31397	7.0
3	R	D	2.0	31397	7.0
4	S	D	2.0	31397	7.0
...	...	...	...	...	...
318433	Q	F	4.0	86499	23.0
318434	Q	E	4.0	325	8.0
318435	R	F	4.0	125235	10.0
318436	Q	D	3.0	91081	8.0
318437	Q	C	2.0	21641	8.0

	Type of Admission	Severity of Illness	Visitors with Patient	Age
0	Emergency	Extreme	2	51-60
1	Trauma	Extreme	2	51-60
2	Trauma	Extreme	2	51-60
3	Trauma	Extreme	2	51-60
4	Trauma	Extreme	2	51-60
...	...	...	...	...
318433	Emergency	Moderate	3	41-50
318434	Urgent	Moderate	4	81-90
318435	Emergency	Minor	3	71-80
318436	Trauma	Minor	5	11-20
318437	Emergency	Minor	2	11-20

	Admission_Deposit	Stay
0	4911.0	0-10
1	5954.0	41-50
2	4745.0	31-40
3	7272.0	41-50
4	5558.0	41-50
...	...	...
318433	4144.0	11-20
318434	6699.0	31-40
318435	4235.0	11-20
318436	3761.0	11-20
318437	4752.0	0-10

[318438 rows x 18 columns]

Install pyspark libraries

In [4]:
```
!pip install -q findspark
!pip install pyspark
!pip install matplotlib
!pip install seaborn
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: pyspark in /usr/local/lib/python3.7/dist-packages (3.3.1)
Requirement already satisfied: py4j==0.10.9.5 in /usr/local/lib/python3.7/dist-packages (from pyspark) (0.10.9.5)
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: matplotlib in /usr/local/lib/python3.7/dist-packages (3.2.2)
Requirement already satisfied: numpy>=1.11 in /usr/local/lib/python3.7/dist-packages (from matplotlib) (1.21.6)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.7/dist-packages (from matplotlib) (0.11.0)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in /usr/local/lib/python3.7/dist-packages (from matplotlib) (3.0.9)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.7/dist-packages (from matplotlib) (1.4.4)
Requirement already satisfied: python-dateutil>=2.1 in /usr/local/lib/python3.7/dist-packages (from matplotlib) (2.8.2)
Requirement already satisfied: typing-extensions in /usr/local/lib/python3.7/dist-packages (from kiwisolver>=1.0.1->matplotlib) (4.1.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/dist-packages (from python-dateutil>=2.1->matplotlib) (1.15.0)
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: seaborn in /usr/local/lib/python3.7/dist-packages (0.11.2)
Requirement already satisfied: matplotlib>=2.2 in /usr/local/lib/python3.7/dist-packages (from seaborn) (3.2.2)
Requirement already satisfied: numpy>=1.15 in /usr/local/lib/python3.7/dist-packages (from seaborn) (1.21.6)
Requirement already satisfied: pandas>=0.23 in /usr/local/lib/python3.7/dist-packages (from seaborn) (1.3.5)
Requirement already satisfied: scipy>=1.0 in /usr/local/lib/python3.7/dist-packages (from seaborn) (1.7.3)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in /usr/local/lib/python3.7/dist-packages (from matplotlib>=2.2->seaborn) (3.0.
9)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.7/dist-packages (from matplotlib>=2.2->seaborn) (0.11.0)
Requirement already satisfied: python-dateutil>=2.1 in /usr/local/lib/python3.7/dist-packages (from matplotlib>=2.2->seaborn) (2.8.2)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.7/dist-packages (from matplotlib>=2.2->seaborn) (1.4.4)
Requirement already satisfied: typing-extensions in /usr/local/lib/python3.7/dist-packages (from kiwisolver>=1.0.1->matplotlib>=2.2->seaborn) (4.1.1)
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.7/dist-packages (from pandas>=0.23->seaborn) (2022.6)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/dist-packages (from python-dateutil>=2.1->matplotlib>=2.2->seaborn) (1.15.0)
```

Ensure spark is set up and running.

In [5]:
```
import findspark
findspark.find()
```

Out[5]:
```
'/usr/local/lib/python3.7/dist-packages/pyspark'
```

In [6]:
```
from pyspark.sql import SparkSession
import seaborn as sns
import matplotlib.pyplot as plt

spark = SparkSession.builder.master('local')\
.appName("Predicting LOS for High Risk Patient")\
.getOrCreate()
```

In [7]:
```
spark
```

In [7]:
```
spark
```

Out[7]: **SparkSession - in-memory**

SparkContext

Spark UI

Version	v3.3.1
Master	local
AppName	Predicting LOS for High Risk Patient

In [11]:
```
print(f"Counts of rows/samples: {df.count()}")
print(f"Counts of columns/features: {len(df.columns)}")
```

```
Counts of rows/samples: case_id                             318438
Hospital_code                       318438
Hospital_type_code                  318438
City_Code_Hospital                  318438
Hospital_region_code                318438
Available Extra Rooms in Hospital   318438
Department                          318438
Ward_Type                           318438
Ward_Facility_Code                  318438
Bed Grade                           318325
patientid                           318438
City_Code_Patient                   313906
Type of Admission                   318438
Severity of Illness                 318438
Visitors with Patient               318438
Age                                 318438
Admission_Deposit                   318438
Stay                                318438
dtype: int64
Counts of columns/features: 18
```

In [13]:
```
df
```

Out[13]:

	case_id	Hospital_code	Hospital_type_code	City_Code_Hospital	Hospital_region_code	Available Extra Rooms in Hospital	Department	Ward_Type	Ward_Facility_Code	Bed Grade	patientid	City_Code_
0	1	8	c	3	Z	3	radiotherapy	R	F	2.0	31397	
1	2	2	c	5	Z	2	radiotherapy	S	F	2.0	31397	
2	3	10	e	1	X	2	anesthesia	S	E	2.0	31397	
3	4	26	b	2	Y	2	radiotherapy	R	D	2.0	31397	
4	5	26	b	2	Y	2	radiotherapy	S	D	2.0	31397	
...	...	...	...	...	...	...	...	...	...	...	...	...
318433	318434	6	a	6	X	3	radiotherapy	Q	F	4.0	86499	

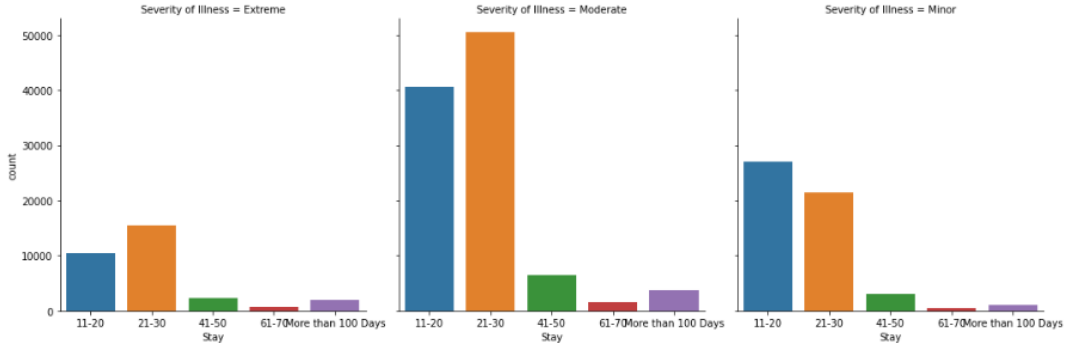
```
selected_list = ["11-20","21-30","41-50","61-70", "More than 100 Days"]

def bivariate_analysis(dataframe, dependent_variable, independent_variable, selected_list):
    g = sns.catplot(dependent_variable, col=independent_variable, col_wrap=3,\
    data=dataframe,kind="count", height=5, aspect=1, order=selected_list
    )
```

Random Forest

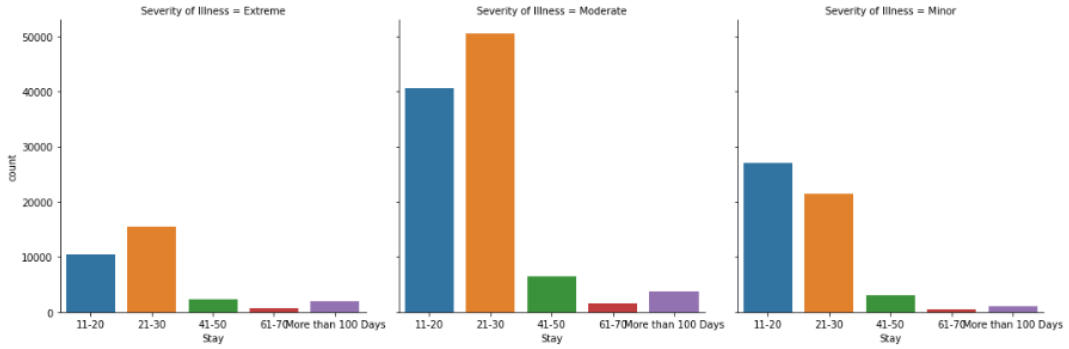
In [44]:
```
bivariate_analysis(df, "Stay", "Severity of Illness", selected_list)
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, th
e only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  FutureWarning
```



Decision Tree

In [47]:
```
bivariate_analysis(df, "Stay", "Severity of Illness", selected_list)
```



8.TESTING

8.1 Test Cases:

Test case ID	Feature Type	Component	Test Scenario	Pre-Requisite	Steps To Execute	Test Data	Expected Result
dataset to IBM	Functional	IBM CLOUD	Verify user is able to upload the data in IBM Cognos	Dataset	Admin should upload the Dataset	Dataset Uploaded	Working as expected
Responsive Login Page	Login Page	DASHBOARD	Verify whether the user is able to create a dashboard	IBM COGNOS ANALYTICS	Upload the dataset to create a interactive dashboard	Login button with orange colour	Working as expected
Report	Functional	IBM CLOUD	Verify user is able to create a Report and can view it	Account to login or data to signup	In IBM dashboard page, Reports can be created	COGNOS	Working as Expected
Home Page UI	Functional	Home page	Verify user is able to log into application with Valid credentials	Account to login or data to signup	3.Enter InValid username/email in Email button	https://63749d149eda8.site123	Working as expected
Hospital's Page	Functional	Home page	Verify user is able to view the details of the hospital's	Login credentials	3.Enter Valid username/email in Email button	https://63749d149eda8.site123	Working as Expected
Contact Page	Functional	Home page	Verify user is able to contact	CONTACT NUMBER / EMAIL	3.Enter InValid username/email in	https://63749d149eda8.site123	Working as Expected

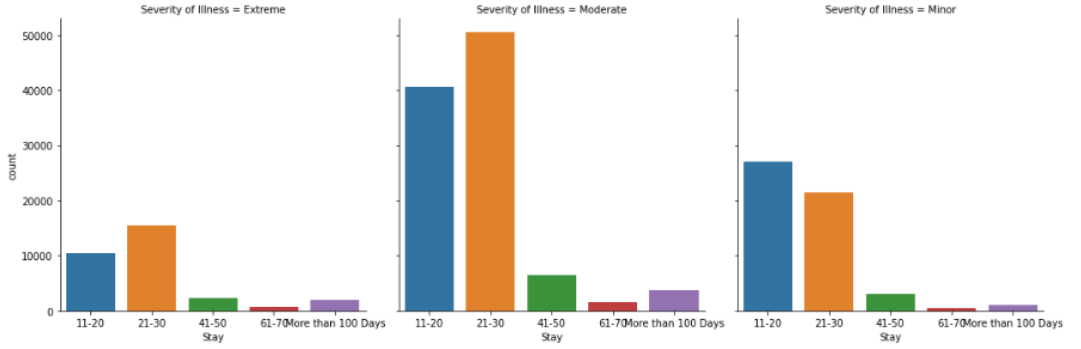
```
selected_list = [ '11-20' , '21-30' , '41-50' , '61-70' , 'More than 100 Days' ]

def bivariate_analysis(dataframe, dependent_variable, independent_variable, selected_list):
    g = sns.catplot(dependent_variable, col=independent_variable, col_wrap=3,\
    data=dataframe,kind="count", height=5, aspect=1, order=selected_list
    )
```

Random Forest

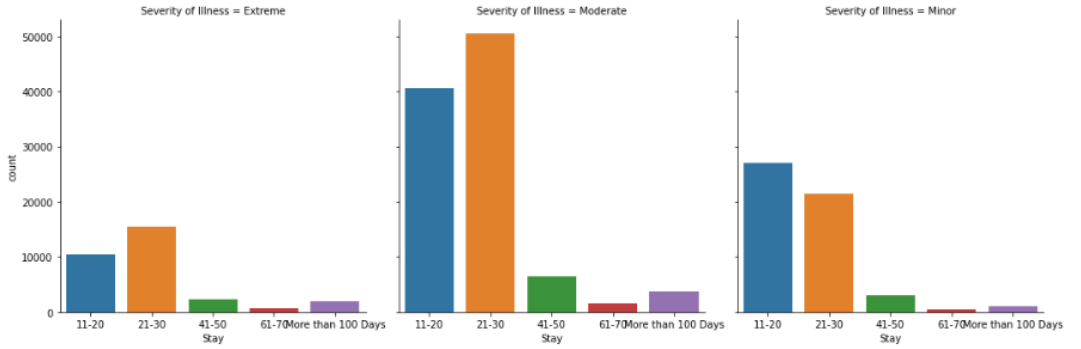
In [44]:
```
bivariate_analysis(df, "Stay", "Severity of Illness", selected_list)
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, th
e only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  FutureWarning
```



Decision Tree

In [47]:
```
bivariate_analysis(df, "Stay", "Severity of Illness", selected_list)
```



8.2 User Acceptance Testing::

Defect Analysis

This report shows the number of resolved or closed bugs at each severity level, and how they were resolved

Resolution	Severity 1	Severity 2	Severity 3	Severity 4	Subtotal
By Design	6	3	1	0	10
Duplicate	1	0	0	1	2
External	1	4	1	2	8
Fixed	5	0	6	6	17
Not Reproduced	1	1	0	1	3
Skipped	1	1	0	0	2
Won't Fix	0	1	2	1	4
Totals	15	10	10	11	46

3.Test Case Analysis

This report shows the number of test cases that have passed, failed, and untested

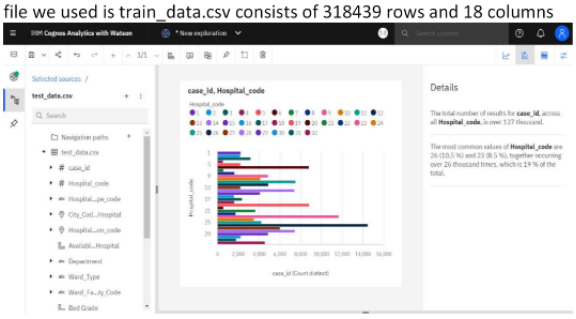
Section	Total Cases	Not Tested	Fail	Pass
design	5	0	0	5
dashboard	15	0	0	15
responsiveness	10	0	0	10
Exception Reporting	17	0	0	17
Final Report Output	13	0	0	13

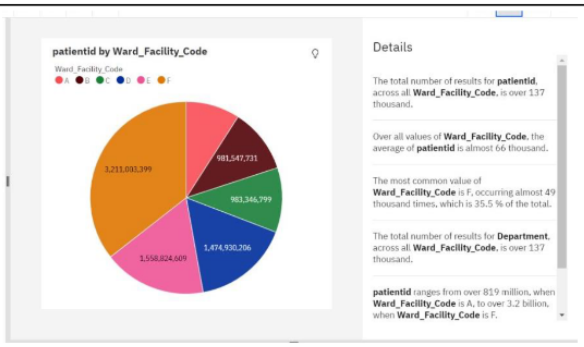
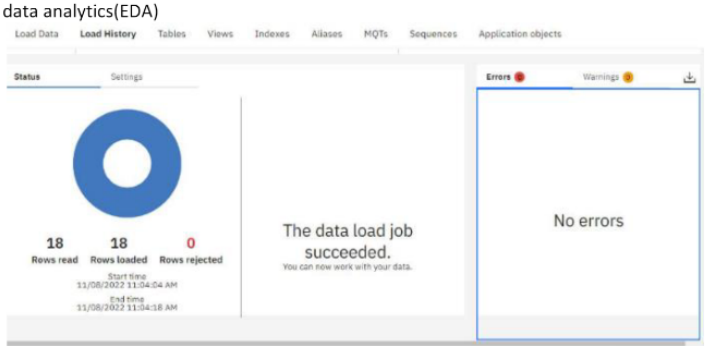
9.RESULTS

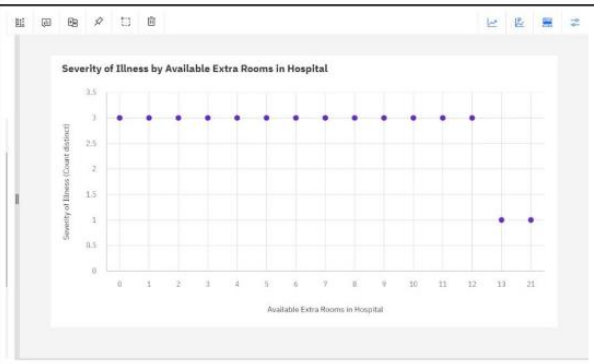
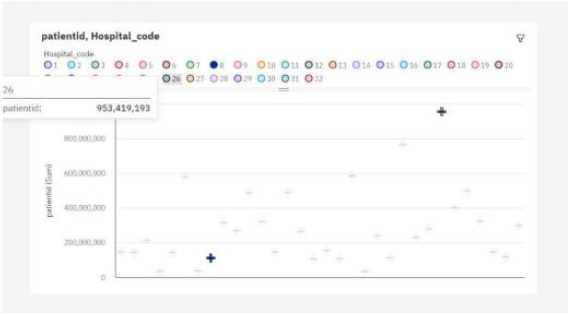
9.1 Performance Test:

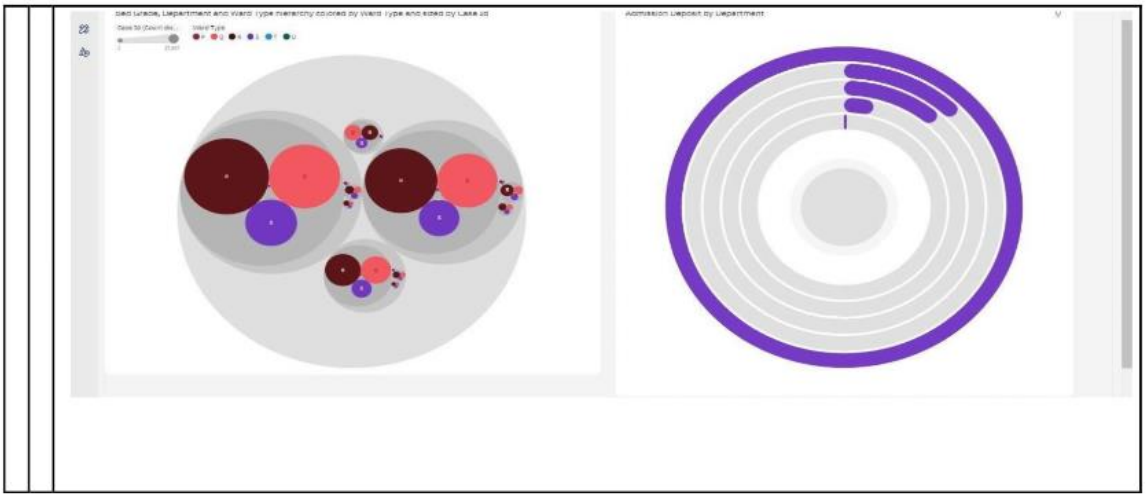
Model Performance Testing:

Project team shall fill the following information in the model performance testing template.

S.No.	Parameter	Screenshot / Values
1.	Dashboard design	No of Visualizations / Graphs – 35
2.	Data Responsiveness	Yes, the website is responsive completely, that is by resizing the browser window size as per the test scenario. **Healthcare PRODUCTION DATASET** The Data is spread across 3 data files and one dictionary file is enclosed. The primary file we used is train_data.csv consists of 318439 rows and 18 columns 

		
3.	Amount Data to Rendered (DB2 Metrics)	**To connect IBM Db2 database cloud with cognos analytics:** By using IBM Db2 to create Dashboard, Report, Story, Visualization and Exploratory data analytics(EDA) 
4.	Utilization of Data Filters	Utilization of data filters – 12 Without Filter

		 With filter 
5.	Effective User Story	No of Scene Added – 6 To Create Dashboard on IBM Cognos To create the Registration page of the Website To create the Login page of the Website To create the Dashboard page of the Website To work on the given dataset, Understand the Dataset Load the dataset to Cloud platform then Build the required
6.	Descriptive Reports	No of Visualizations / Graphs – 5

- 

10. ADVANTAGES & DISADVANTAGES

Advantages

- **We can predict the length of stay patient.**
- **Availability of beds can be checked easily using our dashboard**
- **no report will be delayed from reaching the hands of the authorities.**
- **Improved productivity and cost-effectiveness**

Disadvantages

- **In some algorithms accuracy is less**
- **Data's are not secured**

11.CONCLUSION

This Analytics for Hospitals' Health-Care Data is a quite the reliable and is proven on many stages. All the basic requirements of the hospital are provided in the hospital in order to manage it perfectly and large amount of data can also be stored. It gives many facilities like searching for the detail of patient, billing facilities as well as the creation of test reports. So, it's an important system for modern days.

12.FUTURE SCOPE

Next, we will work on the large deployment of this system, also we will focus to create a more futuristic user interface so that it will be easily accessible to every user to accurately predict the Details regarding to the hospitals

APPENDIX

Sourcecode:https://github.com/IBM-EPBL/IBM-Project-26696-1660033752/tree/main/FINAL%20DELIVERABLES

Git hub link : https://github.com/IBM-EPBL/IBM-Project-26696-1660033752

Project demo link:

https://drive.google.com/file/d/19_L3fqQRT7wes_h2xoIi7-iU287epwms/view?usp=sharing