VISUALIZATION AND PREDICTING HEART DISEASES WITH AN INTERACTIVE DASH BOARD

NALAIYA THIRAN PROJECT BASED LEARNING

ON

PROFESSIONAL READINESS FOR INNOVATION, EMPLOYABILITY AND ENTREPRENEURSHIP

A PROJECT REPORT

P.MohanVamsi	410119104032
Pavan Arikapudi	410119104040
SantoshKumar S.P	410119104301
Vijay.S	410119104302

BACHELOR OF ENGINEERING

IN

ELECTRONICS AND COMMUNICATION ENGINEERING

ADHI COLLEGE OF ENGINEERING AND TECHNOLOGY

KANCHEEPURAM – 631 605

NOVEMBER 2022

ADHI COLLEGE OF ENGINEERING AND TECHNOLOGY

6, Munu Adhi Nagar, Sankarapuram, Near Walajabad,Kancheepuram – 631 605

INTERNAL MENTOR

Ms.Bhanupriya.K

Assistant Professor

Department of Electronics and Communication Engineering

Adhi College of Engineering and Technology,

Kancheepuram – 631 605

INDUSTRY MENTOR

MR. MAHIDHAR,

MS. SAUMYA

IBM

ABSTRACT

Analysis (EDA) detects mistakes, finds appropriate data, checks assumptions and determines the correlation among the explanatory variables. In the context, EDA is considered as analyzing data that excludes inferences and statistical modelling. Analytics is an essential technique for any profession as it forecast the future and hidden pattern. Data analytics is considered as a cost-effective technology in the recent past and it plays an essential role in healthcare which includes new research findings, emergency situations and outbreaks of disease. The use of analytics in healthcare improves care by facilitating preventive care and EDA is a vital step while analyzing data. In this paper, the risk factors that causes heart disease is considered and predicted using K-means algorithm and the analysis is carried out using a publicly available data for heart disease. The dataset holds 209 records with 8 attributes such as age, chest pain type, blood pressure, blood glucose level, ECG in rest, heart rate and four types of chest pain. To predict the heart disease, K-means clustering algorithm is used along with data analytics and visualization tool. The paper discusses the pre-processing methods, classifier performances and evaluation metrics. In the result section, the visualized data shows that the prediction is accurate.

TABLE OF CONTENTS

1. INTRODUCTION

A study in 2016 found that human beings are collectively generated data more than ten exabytes, or 5x1018 bytes from various sources (Lyman and Varian 2003). Exploratory Data Analysis (EDA) is a method to analyze data using advanced techniques to expose hidden structure, enhances the insight into a given dataset, identifies the anomalies and builds parsimonious models to test the underlying assumptions. Exploratory Data Analysis (EDA) is classified into Graphical or non-graphical and Univariate or multivariate Univariate data consider one data column at a time while multivariate method considers more than two variables while analyzing. The diagnostic methods of diseases are of two types namely, Invasive and Non-invasive Invasive diagnostic method includes incise procedures in which instruments are used to cut the skin, mucus membrane and connective tissues. In contrast, non-invasive methods are used to diagnose diseases without opening the skin. Some of the machine learning algorithms based on non-invasive methods are Support Vector Machine (SVM), K- means clustering, K-Nearest neighbor (KNN), Artificial Neural Network (ANN), Naive Bayes, Logistic Regression and rough set [15]. Predicting and diagnosing heart disease is the biggest challenge in the medical industry and it is based on factors like physical examination, symptoms and signs of the patient [1-3]. Factors which influence heart diseases are cholesterol level of the body, smoking habit, and obesity, family history of diseases, blood pressure and working environment. Machine learning algorithms play a vital and accurate role in predicting heart disease [4]. The advancement of technologies allows machine language to pair with big data tools to handle unstructured and exponentially growing data [5].

2. OBJECTIVE

- The objective of this project is to check whether the patient is likely to be diagnosed with any cardiovascular heart disease based on their medical attributes such as gender, age, chest pain, fasting sugar level, etc.
- A dataset is selected from the UCI repository with patient's medical history and attributes.
- To predict the heart disease, K-means clustering algorithm is used along with data analytics and visualization tool. The paper discusses the pre-processing methods, classifier performances and evaluation metrics.
- In the result section, the visualized data shows that the prediction is accurate.

3. IDEATION PHASE

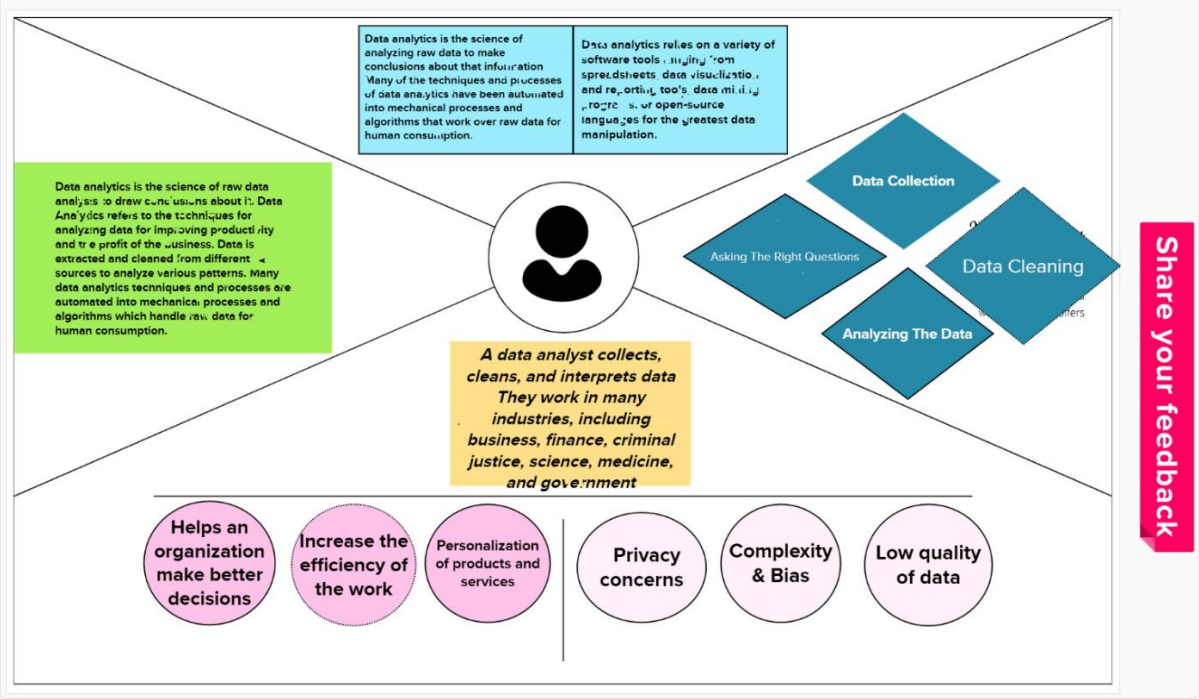
3.1 Literature Survey:

Heart Disease is the one of major causes of death globally. Around17.9millionpeopledie each year. Cardiovascular diseases include disorders of the heart and blood vessels. Four out of five cardiovascular disease deaths are due to heart attacks. One-third of these deaths occur prematurely under the age of seventy. The major number of deaths have occurred in developing countries. India is one of them. For heart disease diagnosis we need cardiologists, which are in limited number in developing countries. Also, the tests for cardiovascular diseases are quite expensive; sometimes out of the budget for common people. Early detection is important in case of heart disease with less expensive prediction techniques. As we know, now-a-days Machine Learning algorithms are used for predicting various diseases. They are also used for predicting heart disease. This paper deals with the survey of Machine Learning algorithms used for predicting heart disease, the importance of attributes to predict the disease and selection of important attributes for prediction.

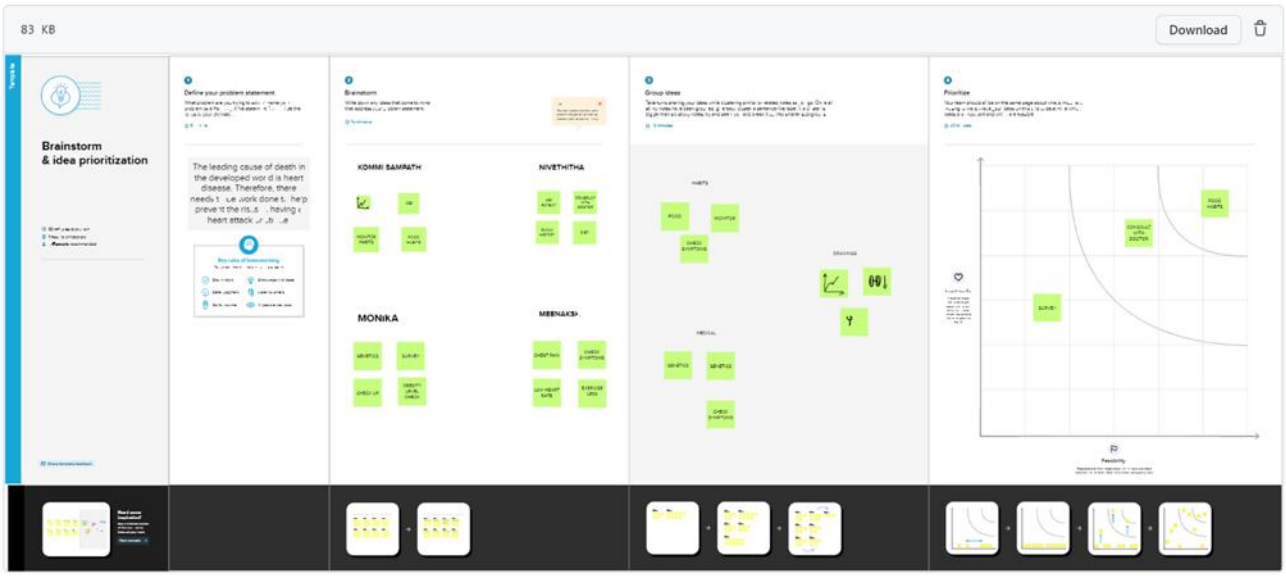
Machine learning (ML) is the subdomain of artificial intelligence (AI). Today we are using MLindaytodaylife.ML based computer programs can access data and use it to learn themselves. It means past experience is used for prediction in ML. ML algorithms are of four types: Supervised Learning in which direct supervision is involved developer label the dataset restricts the boundaries of algorithm, Unsupervised Learning supervision is not required, semi supervised machine learning both types supervised and unsupervised used in combine format and Reinforcement Learning exploration of thing one by one

fist event take as input for next event. In this paper the focus is on supervised machine learning algorithms.

3.2 Empathy Map:



3.3 Ideation:



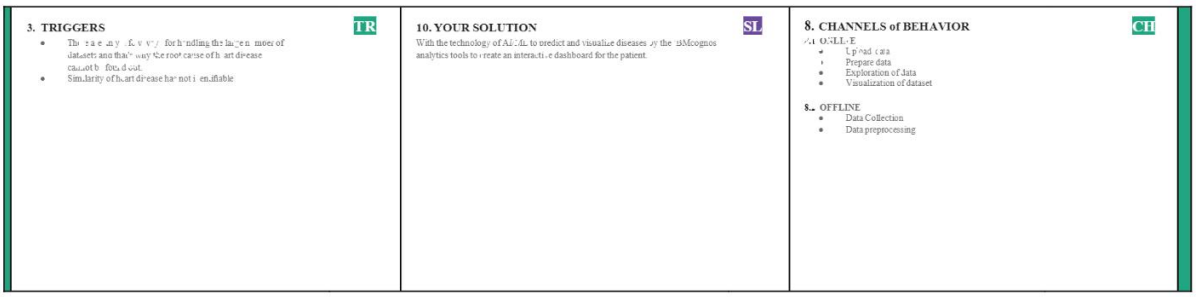
4. PROJECT DESIGN PHASE 1

4.1 Proposed Solution:

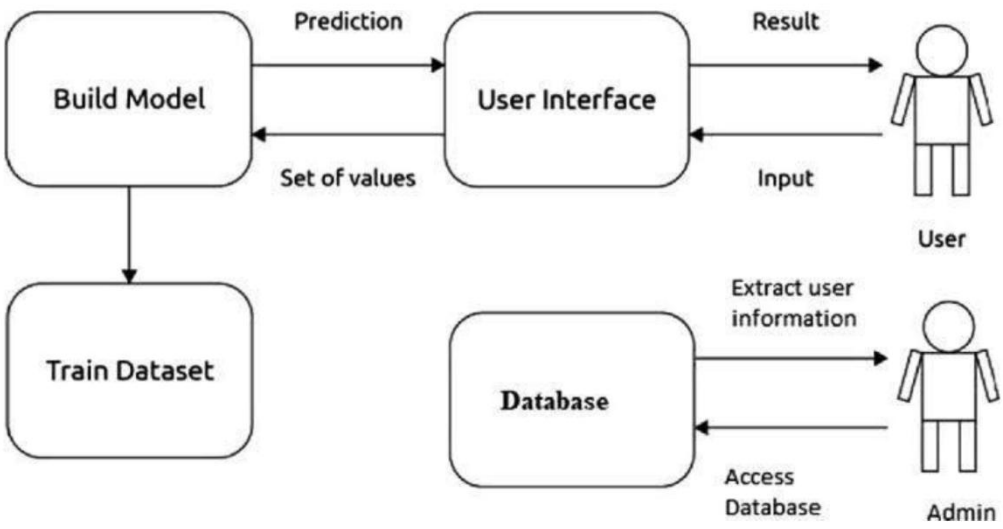
Proposed Solution Template:

S.No.	Parameter	Description
•	Problem Statement (Problem to be solved)	The leading cause of death in the developed world is heart disease. Therefore, there needs to be work done to help prevent the risks of having a heart attack or stroke. Use this dataset to predict which patients are most likely to suffer from a heart disease in the near future using the features given.
•	Idea / Solution description	In Our Project, we are Planning to develop a dashboard using IBM Cognos Analytics Which express our talent as our Outcomes. We are Using Python Language for backend Database Connection. The Data will be stored in the IBM Cloud Platform.
•	Novelty / Uniqueness	IBM Cognos Analytics is the platform that we are going to build the Dashboard is something Unique.For visualizing it, we will require the following data: * Sex * Age * Chest Pain
•	Social Impact / Customer Satisfaction	By predicting the Heart Diseases where the public will have a knowledge on how they are affecting with heart diseases
•	Business Model (Revenue Model)	By our dashboard patients can analyze and predict weather they are going to effect with heart diseases regarding the health factor they are facing, which will be more use for hospitals for genrating the revenue.
•	Scalability of the Solution	We will produce exact information to the patients through Our Project and the functionality of our project will be best in the market.

4.2 Problem Solution Fit:



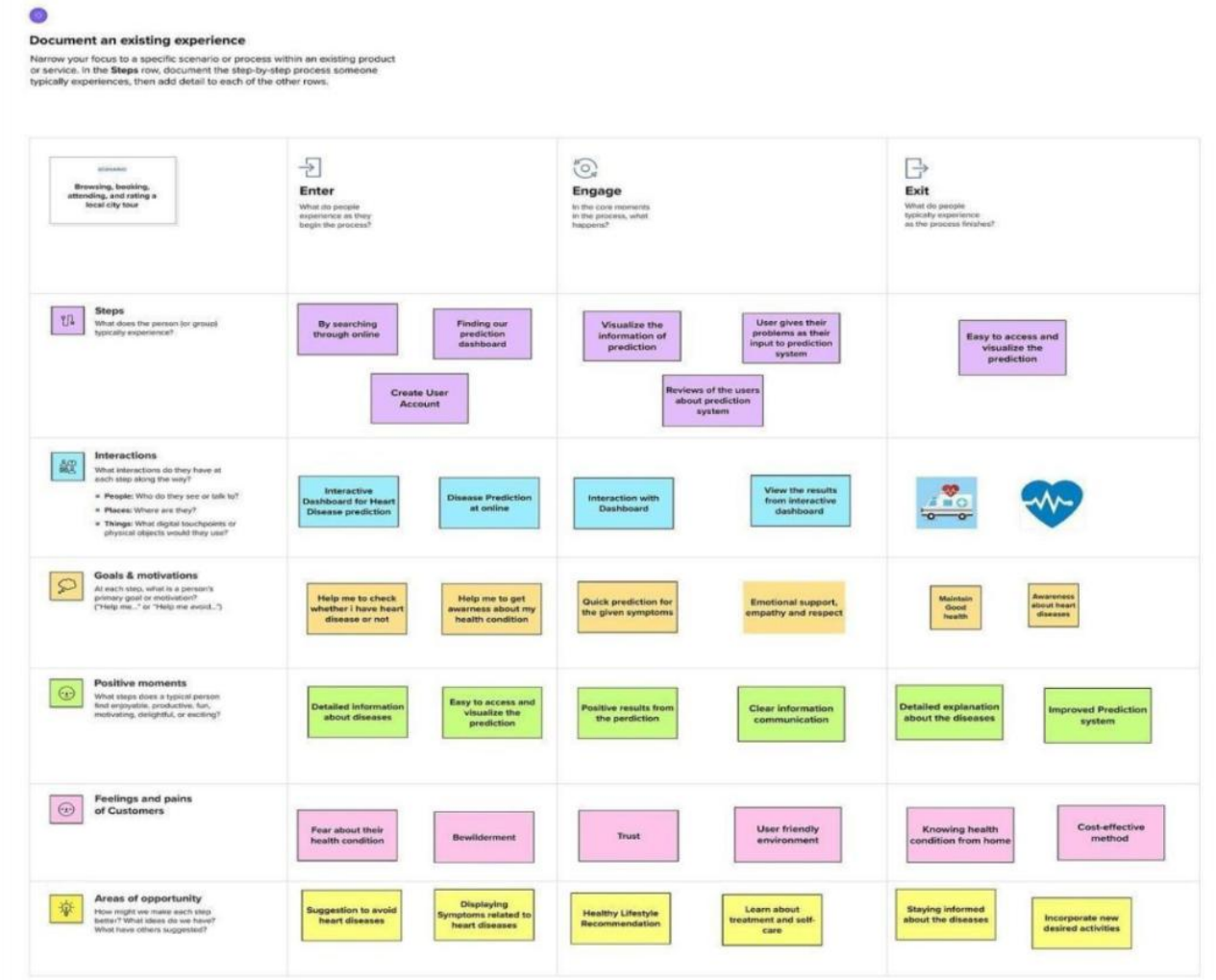
4.3 Solution Architecture:



5. PROJECT DESIGN PHASE 2

5.1 Customer Journey Map:

The customer journey map is a visual representation of the steps a customer takes to complete a specific action, such as signing up for a product trial or subscribing to a newsletter .The more steps involved to complete the specific action, the more detailed the customer journey map will be



5.2 Solution Requirements:

Functional Requirements:

Following are the functional requirements of the proposed solution.

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	Enables user to make registration for the application through Gmail
FR-2	User Confirmation	Once after registration, the user will get confirmation via Email
FR-3	Visualizing Data	User can visualize the trends on the heart disease through Dashboard created using IBM Cognos Analytics
FR-4	Generating Report	User can view his/her health report and can make decisions accordingly

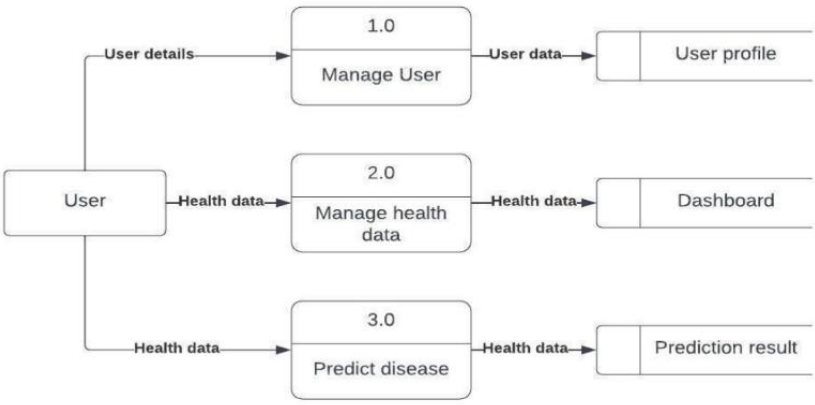
Non-Functional Requirements:

NFR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	The application will have a simple and userfriendlygraphical interface. Users will be able to understand and use all the features of the application easily. Any action has to be performedwith just a few clicks
NFR-2	**Security**	For security of the application the technique knownas database replication should be used so that all the important data should be kept safe. Incase of crash, the system should be able to backup and recover the data
NFR-3	**Reliability**	The application has to be consistent at every scenario and has to work without failure in anyenvironment
NFR-4	**Performance**	Performance of the application depends on the response time and the speed of the data submission. The response time of the application is direct and faster which depends on the efficiency of implemented algorithm
NFR-5	**Availability**	The application has to be available 24 x 7 for users without any interruption
NFR-6	**Scalability**	The application can withstand the increase in the no. of users and has to be able to develop higherversions

5.3 Data Flow Diagram:

A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. A neat and clear DFD can depict the right amount of the system requirement graphically. It shows how data enters and leaves the system, what changes the information, and where data is stored

Data Flow Diagram for Heart Disease Prodi tion Dashboard:



Flow:

1) User creates an account in the application.
2) User enters the medical records in the dashboard.
3) User can view the visualizations of trends in the form of graphs and charts for his/her medical records with the trained dataset.
4) User can view the accuracy of probability of occurrence of heart disease in the dashboard.

User Stories:

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Webuser)	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	I can access my account / Dashboard	High	Sprint-1
		USN-2	As a user, I will receive confirmation email once I have registered for the application	I can receive confirmation email & click confirm	High	Sprint-1
	Login	USN-3	As a user, I can log into the application by entering email & password	I can access my account / Dashboard whenlogged in	High	Sprint-1
Customer (Webuser)	Dashboard	USN-4	User can view his/her complete medical analysis and accuracy of disease prediction	I can view my medical analysis in the dashboard	High	Sprint-2
		USN-5	User can view the accuracy of occurrence of heart disease	I can view the accuracy of heart disease in the dashboard	High	Sprint-2
Customer Care Executive	Helpdesk	USN-6	As a customer care executive, he/she can view the customer queries.	I can post my queries in the dashboard	Medium	Sprint-3
		USN-7	As a customer care executive, he/she can answer the customer queries.	I can get support from helpdesk	High	Sprint-3

Administrator	User Profile	USN-8	As an admin, he/she can update the health details of users.	I can view my updated health details.	High	Sprint-4

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
		USN-9	As an admin, he/she can add or delete users.	I can access my account / Dashboard whenlogged in	High	Sprint-4
		USN-10	As an admin, he/she can manage the user details.	I can view the organized data of myself.	High	Sprint-4

5.4 Technology Stack:

Table-1: Components & Technologies:

S.No	Component	Description	Technology
1.	User Interface	How user interacts with application e.g. Web UI, Mobile App, Chatbot etc.	HTML, CSS, JavaScript / Angular Js / React Js etc.
2.	Application Logic-1	Logic for a process in the application	Java / Python
3.	Application Logic-2	Logic for a process in the application	IBM Watson STT service
4.	Application Logic-3	Logic for a process in the application	IBM Watson Assistant
5.	Database	Data Type, Configurations etc.	MySQL, NoSQL, etc.
6.	Cloud Database	Database Service on Cloud	IBM DB2, IBM Cloudant etc.
7.	File Storage	File storage requirements	IBM Block Storage or Other Storage Service or Local Filesystem
8.	External API-1	Purpose of External API used in the application	IBM Weather API, etc.
9.	External API-2	Purpose of External API used in the application	Aadhar API, etc.
10.	Machine Learning Model	Purpose of Machine Learning Model	Object Recognition Model, etc.
11.	Infrastructure (Server / Cloud)	Application Deployment on Local System / Cloud Local Server Configuration: Cloud Server Configuration :	Local, Cloud Foundry, Kubernetes, etc.

Table-2: Applications Characteristics:

S.No	Characteristics	Description	Technology
1.	Open-Source Frameworks	List the open-source frameworks used	Technology of Opensource framework
2.	Security Implementations	List all the security / access controls implemented, use of firewalls etc.	e.g. SHA-256, Encryptions, IAM Controls, OWASP etc.
3.	Scalable Architecture	Justify the scalability of architecture (3 – tier, Micro-services)	Technology used
4.	Availability	Justify the availability of application (e.g. use of load balancers, distributed servers etc.)	Technology used
5.	Performance	Design consideration for the performance of the application (number of requests per sec, use of Cache, use of CDN's) etc.	Technology used

6. PROJECT PLANNING PHASE

6.1 Prepare Milestone and Activity List:

PROJECT PLANNING PHASE

PROJECT MILESTONE

DATE		22 October 2022	
TEAM ID		PNT2022TMID37721	
PROJECT NAME		Visualizing and Predicting Heart diseases with an Interactive dashboard	
MAXIMUM		4 Marks	
S.NO	ACTIVITY TITLE	ACTIVITY DESCRIPTION	DURATION
1	Understanding the project requirement	Create a repository and assign team members utilising Github, give them the task, all individuals teach students how to use, open, and work on the Github, career at IBM education.	1 WEEK
2	Starting of project	Encourage the students to enrol in IBM portal classes and conceive of create a rough depiction based on project detailing and group of details about IBM and IOT task and team leader delegate a task every participant of the undertaking.	2 WEEKS

3	Attend class	Team members and the team captain must observe and discover from the classes offered from IBM and NALAYATHIRAN and must advance entry to MIT permit for their project.	4 WEEKS
4	Budget and scope of project	Data Analytics eliminates guess work and manual tasks. Analyse the project's budget and Data Analytics use and speak of using a team for budget forecast to foresee the favourableness of the client to buy.	1 WEEK

6.2 Sprint Delivery Plan:

Use the below template to create product backlog and sprint schedule

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	2	High	Sampath K
Sprint-1	Confirmation	USN-2	As a user, I will receive confirmation email once I have registered for the application	1	High	Meenakshi P
Sprint-2		USN-3	As a user, I can register for the application through Facebook	2	Low	Monika M
Sprint-1		USN-4	As a user, I can register for the application through Gmail	2	Medium	Nivethitha S
Sprint-1	Login	USN-5	As a user, I can log into the application by entering email & password	1	High	Sampath K
Sprint-1	User Interface	USN-6	As a user, I should not need any pre requisites to handle the UI	1	Medium	Monika M
Sprint-1	Dashboard		As a user, will use the templates and resources of the dashboard effectively	2	High	Nivethitha S
Sprint-1	Present data		As a user, will present the data in the IBM cognos analytics platform	2	High	Meenakshi P
Sprint	**Functional Requirement (Epic)**	**User Story Number**	**User Story / Task**	**Story Points**	**Priority**	**Team Members**
Sprint-1	EDA		As a user, will perfom the Exploratory Data Analytics(EDA) in a correct manner	2	High	Sampath K
Sprint-1	Visualization		As a user, data visualization will be performed effectively	2	High	Nivethitha S
Sprint-2	Report		As a user, I will take responsibility that a report will be finally made by our team	2	High	Monika M

PROJECT TRACKER:

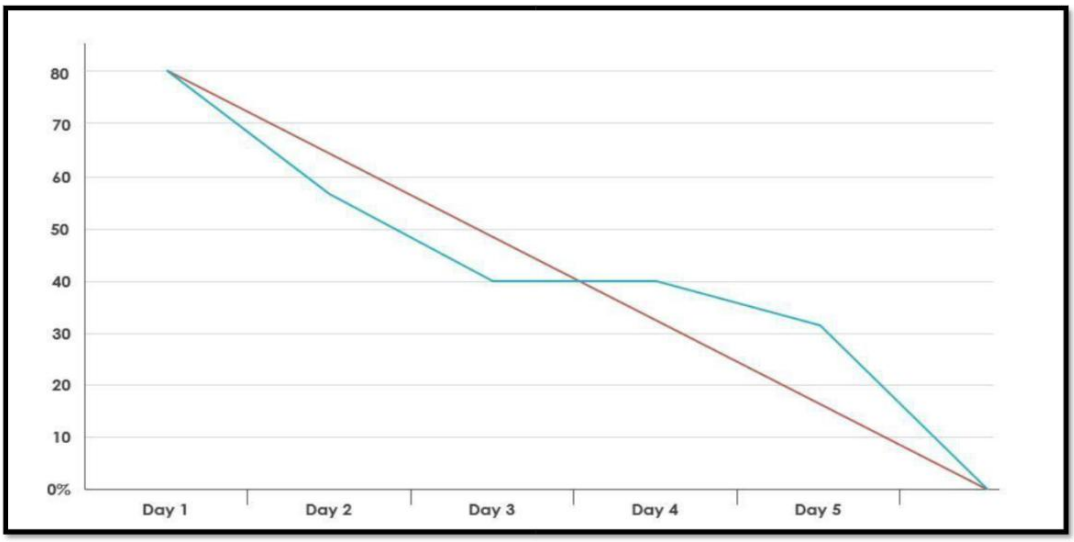
Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	20	6 Days	24 Oct 2022	29 Oct 2022	20	29 Oct 2022
Sprint-2	20	6 Days	31 Oct 2022	05 Nov 2022	30	30 Oct 2022
Sprint-3	20	6 Days	07 Nov 2022	12 Nov 2022	49	06 Nov 2022
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022	50	07 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 point per day)

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

Burndown chart:

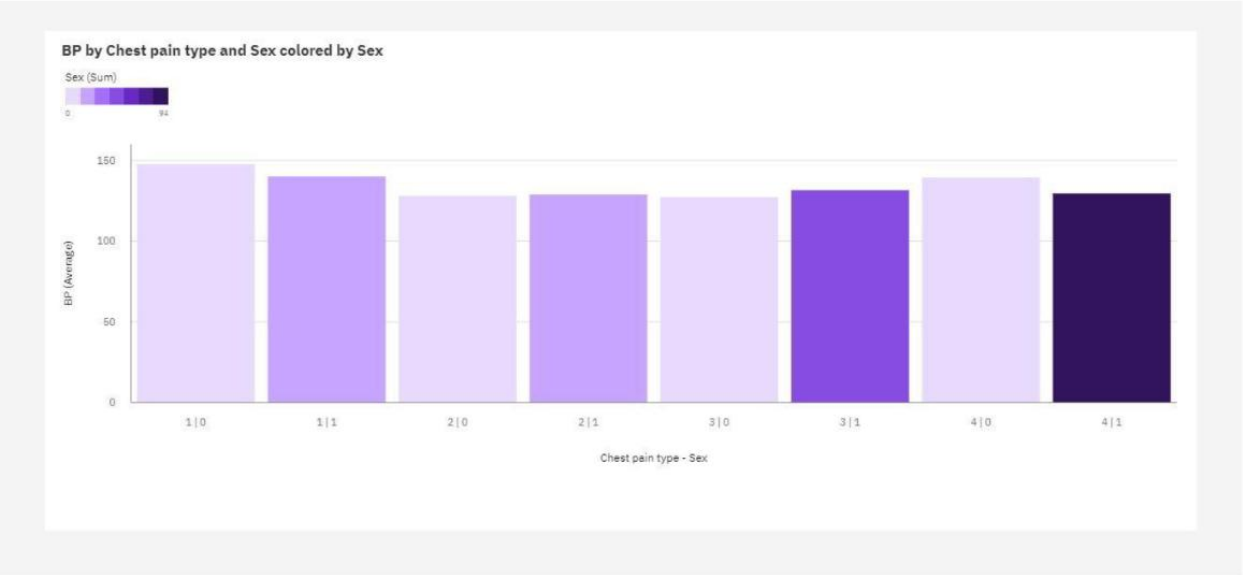


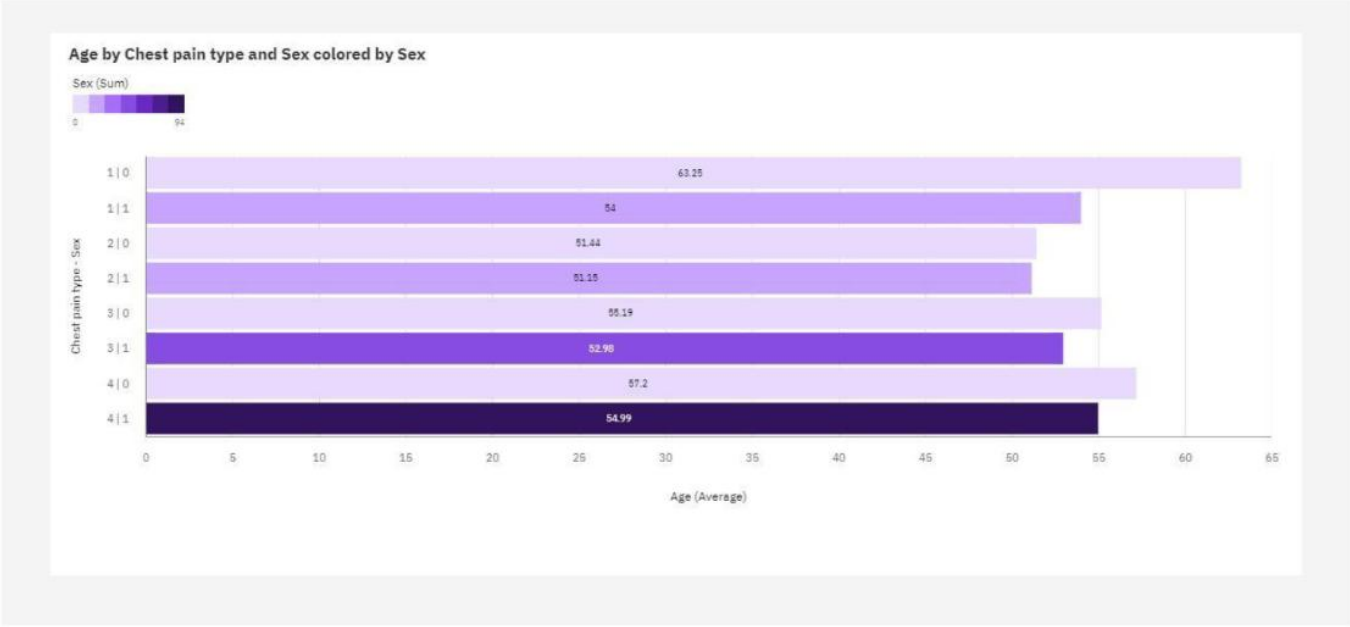
7.PROJECT DEVELOPMENT PHASE

7.1 Delivery of sprint 1: Heart Disease Prediction

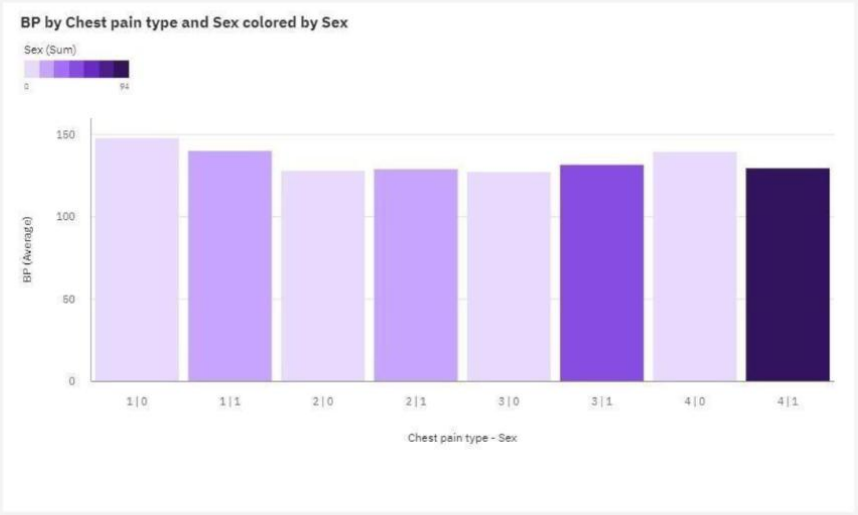
	Age	Sex	Chest pain type	BP	Cholesterol	FBS over 120	EKG results	Max HR	Exercise angina	ST depression	Slope of ST	Number of vessels fluro	Thallium	Heart Disease
1	70	1	4	130	322	0	2	109	0	2.4	2	3	3	Presence
2	67	0	3	115	564	0	2	160	0	1.6	2	0	7	Absence
3	57	1	2	124	261	0	0	141	0	0.3	1	0	7	Presence
4	64	1	4	128	263	0	0	105	1	0.2	2	1	7	Absence
5	74	0	2	120	269	0	2	121	1	0.2	1	1	3	Absence
6	65	1	4	120	177	0	0	140	0	0.4	1	0	7	Absence
7	56	1	3	130	256	1	2	142	1	0.6	2	1	6	Presence
8	59	1	4	110	239	0	2	142	1	1.2	2	1	7	Presence
9	60	1	4	140	293	0	2	170	0	1.2	2	2	7	Presence
10	63	0	4	150	407	0	2	154	0	4	2	3	7	Presence
11	59	1	4	135	234	0	0	161	0	0.5	2	0	7	Absence
12	53	1	4	142	226	0	2	111	1	0	1	0	7	Absence
13	44	1	3	140	235	0	2	180	0	0	1	0	3	Absence
14	61	1	1	134	234	0	0	145	0	2.6	2	2	3	Presence
15	57	0	4	128	303	0	2	159	0	0	1	1	3	Absence
16	71	0	4	112	149	0	0	125	0	1.6	2	0	3	Absence
17	46	1	4	140	311	0	0	120	1	1.8	2	2	7	Presence
18	53	1	4	140	203	1	2	155	1	3.1	3	0	7	Presence
19	64	1	1	110	211	0	2	144	1	1.8	2	0	3	Absence
20	40	1	1	140	199	0	0	178	1	1.4	1	0	7	Absence
21	67	1	4	120	229	0	2	129	1	2.6	2	2	7	Presence
22	48	1	2	130	245	0	2	180	0	0.2	2	0	3	Absence
23	43	1	4	115	303	0	0	181	0	1.2	2	0	3	Absence
24	47	1	4	112	204	0	0	143	0	0.1	1	0	3	Absence
25	54	0	2	132	288	1	2	159	1	0	1	1	3	Absence
26	48	0	3	130	275	0	0	139	0	0.2	1	0	3	Absence
27	46	0	4	138	243	0	2	152	1	0	2	0	3	Absence
28	51	0	3	120	295	0	2	157	0	0.6	1	0	3	Absence
29	58	1	3	112	230	0	2	165	0	2.5	2	1	7	Presence
30	71	0	3	110	265	1	2	130	0	0	1	1	3	Absence
31	57	1	3	128	229	0	2	150	0	0.4	2	1	7	Presence
32	66	1	4	160	228	0	2	138	0	2.3	1	0	6	Absence
33	37	0	3	120	215	0	0	170	0	0	1	0	3	Absence
34	59	1	4	170	326	0	2	140	1	3.4	3	0	7	Presence
35	50	1	4	144	200	0	2	126	1	0.9	2	0	7	Presence
36	48	1	4	130	256	1	2	150	1	0	1	2	7	Presence
37	61	1	4	140	207	0	2	138	1	1.9	1	1	7	Presence
38	59	1	1	160	273	0	2	125	0	0	1	0	3	Presence
39	42	1	3	130	180	0	0	150	0	0	1	0	3	Absence
40	48	1	4	122	222	0	2	186	0	0	1	0	3	Absence
41	40	1	4	152	223	0	0	181	0	0	1	0	7	Presence
42	62	0	4	124	209	0	0	163	0	0	1	0	3	Absence
43	44	1	3	130	233	0	0	179	1	0.4	1	0	3	Absence
44	46	1	2	101	197	1	0	156	0	0	1	0	7	Absence
45	59	1	3	126	218	1	0	134	0	2.2	2	1	6	Presence
46	58	1	3	140	211	1	2	165	0	0	1	0	3	Absence

Delivery of Sprint 1: Working with database:





Exploration Of Max Heart Rate During The Chest Pain

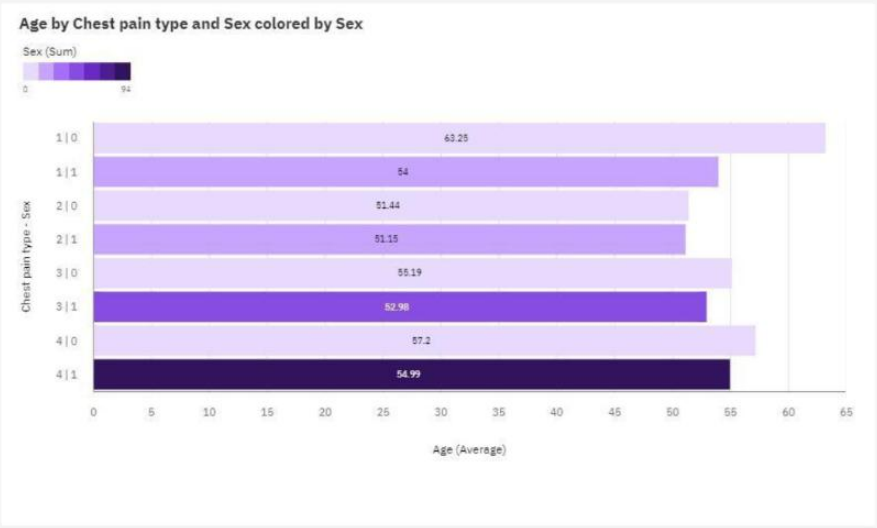


Details

Over all **chest pain type - sexes**, the average of **BP** is 0.6778.

The average values of **BP** range from 0, occurring when **Chest pain type - Sex** is 1|0, to 1, when **Chest pain type - Sex** is 1|0.

The most common value of **Chest pain type - Sex** is 2|1, occurring 129 times, which is 47.8 % of the total.

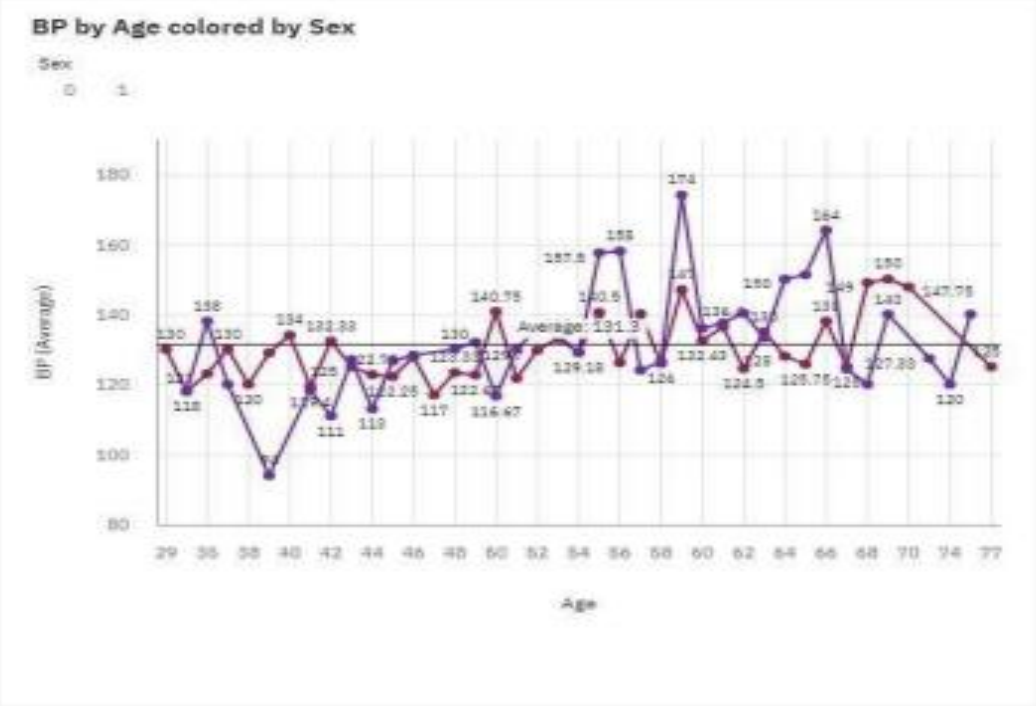


Details

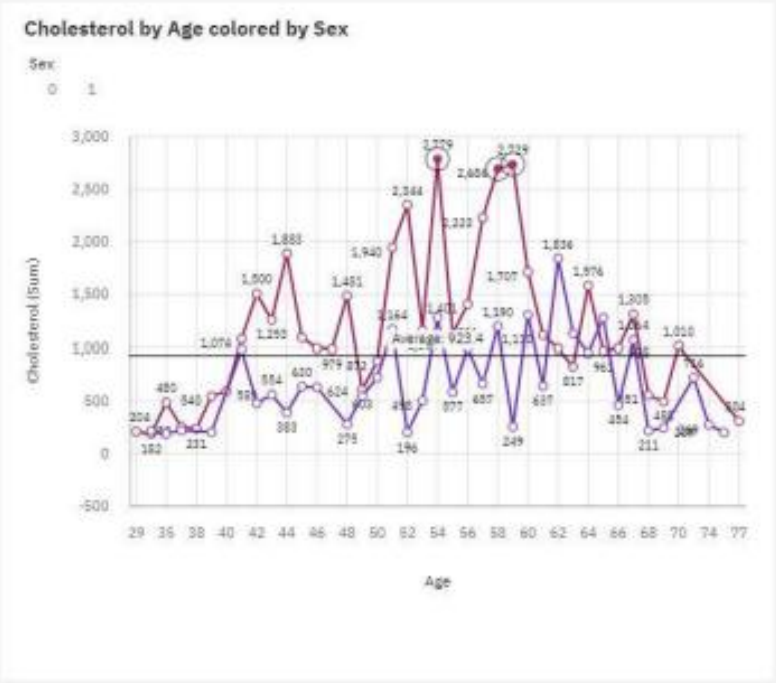
The most common value of **Chest pain type - Sex** is 2|1, occurring 129 times, which is 47.8 % of the total.

Over all **chest pain type - sexes**, the average of **Age** is 0.6778.

The average values of **Age** range from 0, occurring when **Chest pain type - Sex** is 1|0, to 1, when **Chest pain type - Sex** is 1|0.



Exploration Of Chorestrol By Age And Gender



Details

For **Cholesterol**, the most significant value of **Sex** is 1, whose respective **Cholesterol** values add up to over 44 thousand, or 65.8 % of the total.

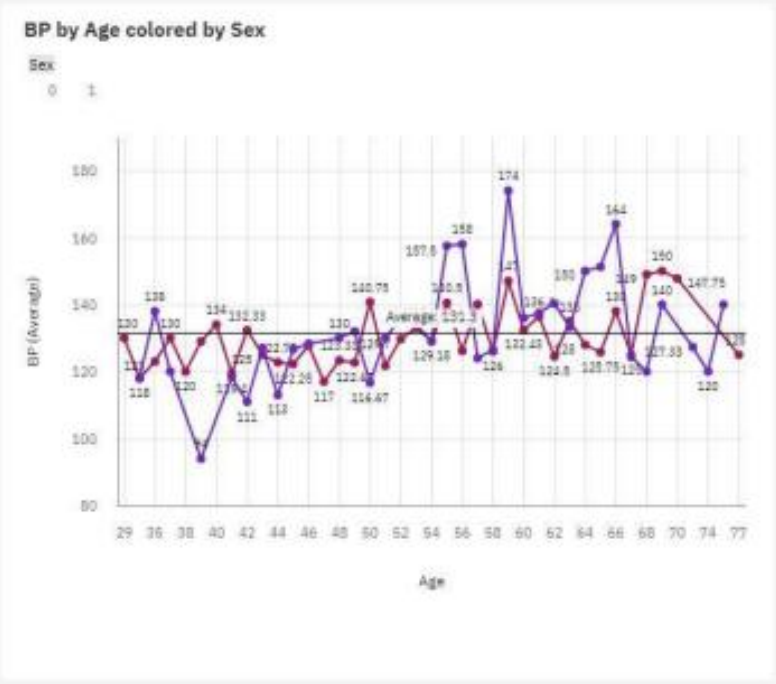
Across all **ages** and **sexes**, the sum of **Cholesterol** is over 67 thousand.

For **Cholesterol**, the most significant values of **Age** are 54 and 58, whose respective **Cholesterol** values add up to nearly eight thousand, or 11.8 % of the total.

The summed values of **Cholesterol** range from 182 to nearly three thousand.

Cholesterol is unusually high when the combinations of **Age** and **Sex** are 54 and 1, 59 and 1 and 58 and 1.

Cholesterol is unusually high when **Age** is 54 and 58.



Details

The most common values of **Age** are 54 (5.9 %) and 58 (5.6 %), together occurring 31 times, which is 11.5 % of the total.

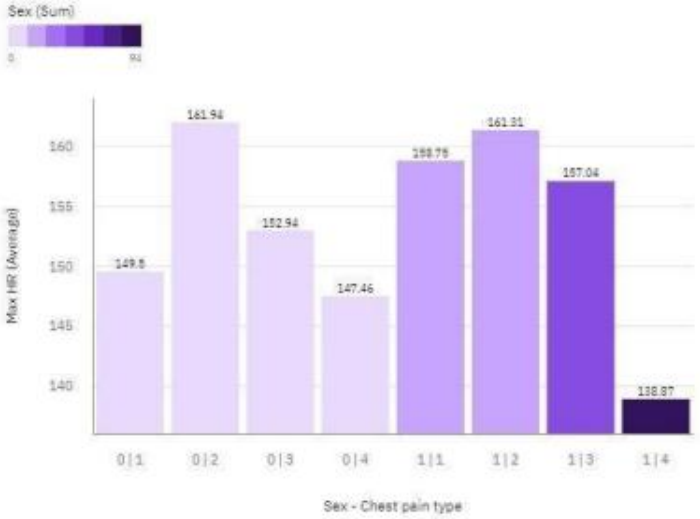
BP is unusually high when **Age** is 59.

The most common value of **Sex** is 1, occurring 183 times, which is 67.8 % of the total.

Over all **ages** and **sexes**, the average of **BP** is 131.3.

The average values of **BP** range from 94 to 174.

Max HR by Sex and Chest pain type colored by Sex



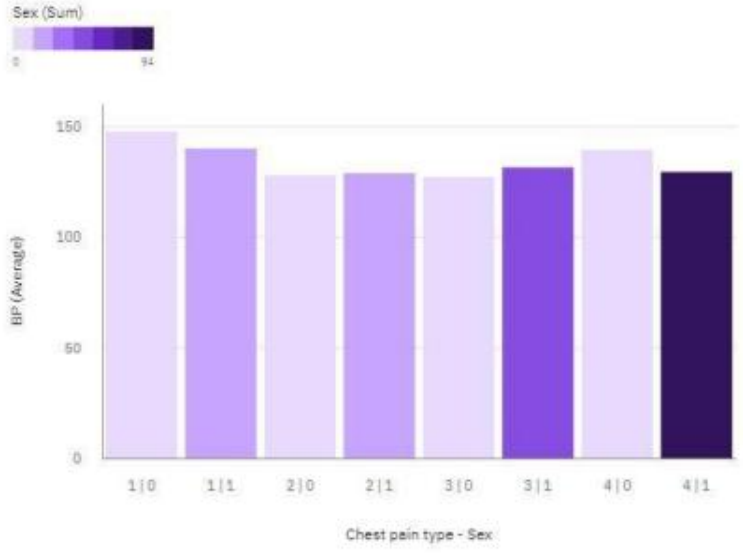
Details

Over all **sex - chest pain types**, the average of **Max HR** is 2.174.

The average values of **Max HR** range from 0, occurring when **Sex - Chest pain type** is 0|1, to 3, when **Sex - Chest pain type** is 0|1.

The most common value of **Sex - Chest pain type** is 0|2, occurring 183 times, which is 67.8 % of the total.

BP by Chest pain type and Sex colored by Sex



Details

Over all **chest pain type - sexes**, the average of **BP** is 0.6778.

The average values of **BP** range from 0, occurring when **Chest pain type - Sex** is 1|0, to 1, when **Chest pain type - Sex** is 1|0.

The most common value of **Chest pain type - Sex** is 2|1, occurring 129 times, which is 47.8 % of the total.

Delivery of Sprint 1: Working with Dataset

In [60]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

In [61]:
```
df = pd.read_csv('Heart_Disease_Prediction.csv')
```

In [62]:
```
df
```

Out[62]:

	Age	Sex	Chest pain type	BP	Cholesterol	FBS over 120	EKG results	Max HR	Exercise angina	ST depression	Slope of ST	Number of vessels fluro	Thallium	Heart Disease
0	70	1	4	130	322	0	2	109	0	2.4	2	3	3	Presence
1	67	0	3	115	564	0	2	160	0	1.6	2	0	7	Absence
2	57	1	2	124	261	0	0	141	0	0.3	1	0	7	Presence
3	64	1	4	128	263	0	0	105	1	0.2	2	1	7	Absence
4	74	0	2	120	269	0	2	121	1	0.2	1	1	3	Absence
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
265	52	1	3	172	199	1	0	162	0	0.5	1	0	7	Absence
266	44	1	2	120	263	0	0	173	0	0.0	1	0	7	Absence
267	56	0	2	140	294	0	2	153	0	1.3	2	0	3	Absence
268	57	1	4	140	192	0	0	148	0	0.4	2	0	6	Absence
269	67	1	4	160	286	0	2	108	1	1.5	2	3	3	Presence

270 rows × 14 columns

In [63]:
```
df.isnull().any()
```

Out[63]:
```
Age                     False
Sex                     False
Chest pain type         False
BP                      False
Cholesterol             False
FBS over 120            False
EKG results             False
Max HR                  False
Exercise angina         False
ST depression           False
Slope of ST             False
Number of vessels fluro False
Thallium                False
Heart Disease           False
dtype: bool
```

In [64]:
```
df.isnull().sum()
```

Out[64]:
```
Age                     0
Sex                     0
Chest pain type         0
BP                      0
Cholesterol             0
FBS over 120            0
EKG results             0
Max HR                  0
Exercise angina         0
ST depression           0
Slope of ST             0
Number of vessels fluro 0
Thallium                0
Heart Disease           0
dtype: int64
```

In [65]:
```
df.isna().sum()
```

Out[65]:
```
Age                     0
Sex                     0
Chest pain type         0
BP                      0
Cholesterol             0
FBS over 120            0
EKG results             0
Max HR                  0
Exercise angina         0
ST depression           0
Slope of ST             0
Number of vessels fluro 0
Thallium                0
Heart Disease           0
dtype: int64
```

In [66]:
```
df.corr()
```

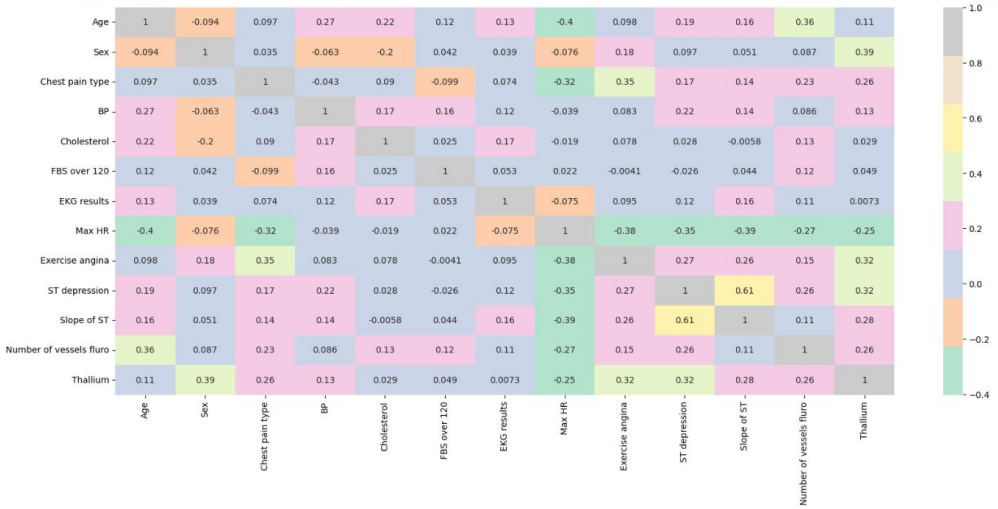
```
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_2440\1134722465.py:1: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In
a future version, it will default to False. Select only valid columns or specify the value of numeric_only to silence this warning.
  df.corr()
```

Out[66]:

	Age	Sex	Chest pain type	BP	Cholesterol	FBS over 120	EKG results	Max HR	Exercise angina	ST depression	Slope of ST	Number of vessels fluro	Thallium
Age	1.000000	-0.094401	0.096920	0.273053	0.220056	0.123458	0.128171	-0.402215	0.098297	0.194234	0.159774	0.356081	0.10610
Sex	-0.094401	1.000000	0.034636	-0.062693	-0.201647	0.042140	0.039253	-0.076101	0.180022	0.097412	0.050545	0.086830	0.39104
Chest pain type	0.096920	0.034636	1.000000	-0.043196	0.090465	-0.098537	0.074325	-0.317682	0.353160	0.167244	0.136900	0.225890	0.26265
BP	0.273053	-0.062693	-0.043196	1.000000	0.173019	0.155681	0.116157	-0.039136	0.082793	0.222800	0.142472	0.085697	0.13204
Cholesterol	0.220056	-0.201647	0.090465	0.173019	1.000000	0.025186	0.167652	-0.018739	0.078243	0.027709	-0.005755	0.126541	0.02883
FBS over 120	0.123458	0.042140	-0.098537	0.155681	0.025186	1.000000	0.053499	0.022494	-0.004107	-0.025538	0.044076	0.123774	0.04923
EKG results	0.128171	0.039253	0.074325	0.116157	0.167652	0.053499	1.000000	-0.074628	0.095098	0.120034	0.160614	0.114368	0.00733
Max HR	-0.402215	-0.076101	-0.317682	-0.039136	-0.018739	0.022494	-0.074628	1.000000	-0.380719	-0.349045	-0.386847	-0.265333	-0.25339
Exercise angina	0.098297	0.180022	0.353160	0.082793	0.078243	-0.004107	0.095098	-0.380719	1.000000	0.274672	0.255908	0.153347	0.32144
ST depression	0.194234	0.097412	0.167244	0.222800	0.027709	-0.025538	0.120034	-0.349045	0.274672	1.000000	0.609712	0.255005	0.32433
Slope of ST	0.159774	0.050545	0.136900	0.142472	-0.005755	0.044076	0.160614	-0.386847	0.255908	0.609712	1.000000	0.109498	0.28367
Number of vessels fluro	0.356081	0.086830	0.225890	0.085697	0.126541	0.123774	0.114368	-0.265333	0.153347	0.255005	0.109498	1.000000	0.25564
Thallium	0.106100	0.391046	0.262659	0.132045	0.028836	0.049237	0.007337	-0.253397	0.321449	0.324333	0.283678	0.255648	1.00000

In [67]:
```
plt.figure(figsize=(20,8))
sns.heatmap(df.corr(), cmap="Pastel2", annot=True)
plt.show()
```

```
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_2440\4025539413.py:2: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In
a future version, it will default to False. Select only valid columns or specify the value of numeric_only to silence this warning.
  sns.heatmap(df.corr(), cmap="Pastel2", annot=True)
```



Out[92]:

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	0.520833	1.0	0.333333	0.924528	0.500000	0.0	1.0	0.946565	0.0	0.000000	0.0	0.333333	1.0
1	0.500000	1.0	1.000000	0.273585	0.496269	0.0	0.0	0.183206	1.0	0.322581	0.5	0.666667	1.0
2	0.291667	0.0	1.000000	0.358491	0.716418	1.0	1.0	0.496183	1.0	0.483871	0.5	0.000000	1.0
3	0.791667	0.0	0.666667	0.547170	0.477612	0.0	0.0	0.770992	0.0	0.000000	0.0	0.333333	0.0
4	0.645833	1.0	0.666667	0.433962	0.134328	0.0	1.0	0.641221	0.0	0.483871	0.5	0.000000	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
184	0.583333	1.0	0.333333	0.283019	0.417910	0.0	0.0	0.534351	0.0	0.048387	0.0	0.000000	1.0
185	0.250000	0.0	0.333333	0.301887	0.585821	0.0	0.0	0.702290	0.0	0.000000	0.0	0.000000	0.0
186	0.250000	1.0	0.666667	0.169811	0.376866	0.0	0.0	0.824427	0.0	0.000000	0.0	0.000000	0.0
187	0.604167	1.0	1.000000	0.056604	0.317164	0.0	0.0	0.648855	0.0	0.016129	0.0	0.333333	1.0
188	0.416667	0.0	1.000000	0.339623	0.447761	0.0	0.0	0.702290	0.0	0.000000	0.0	0.000000	0.0

189 rows × 13 columns

In [93]:
```
X_train.shape, X_test.shape
```

Out[93]:
```
((189, 13), (81, 13))
```

In [94]:
```
rfm = RandomForestClassifier()
rfm.fit(X_train, y_train)
```

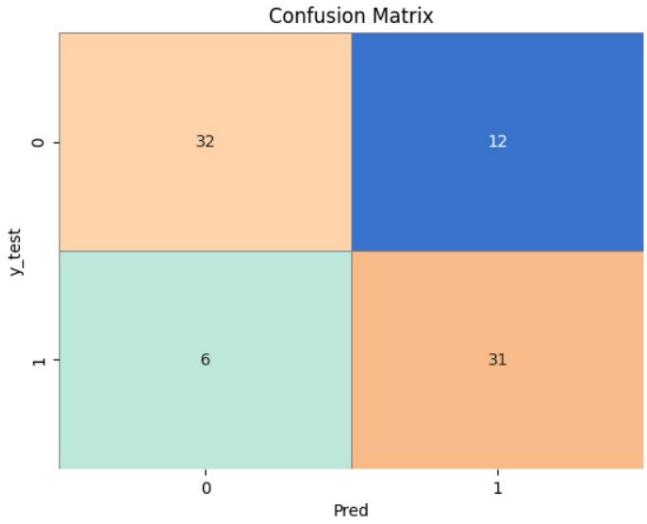
Out[94]:
```
RandomForestClassifier()
```
In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [95]:
```
pred = rfm.predict(X_test)
```

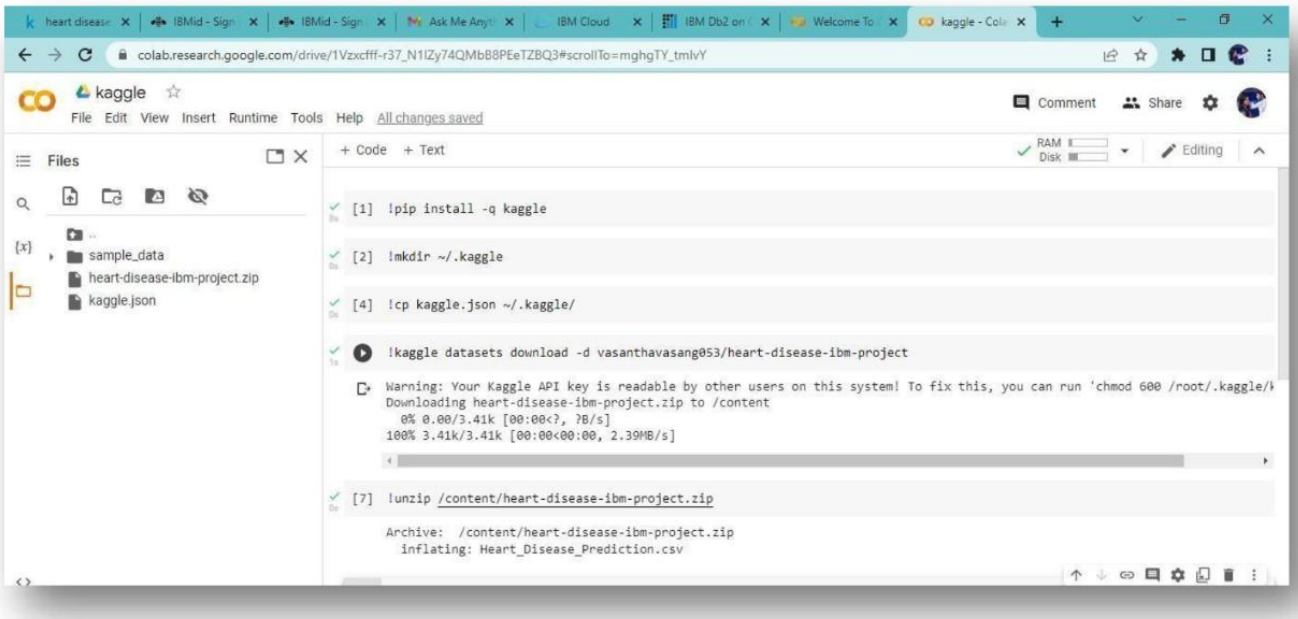
In [96]:
```
accuracy_score(y_test, pred)
```

In [98]:
```
sns.heatmap(cm, annot = True, fmt = 'g', cbar = False, cmap = 'icefire', linewidths= 0.5, linecolor= 'grey')
plt.title('Confusion Matrix')
plt.ylabel('y_test')
plt.xlabel('Pred')
```

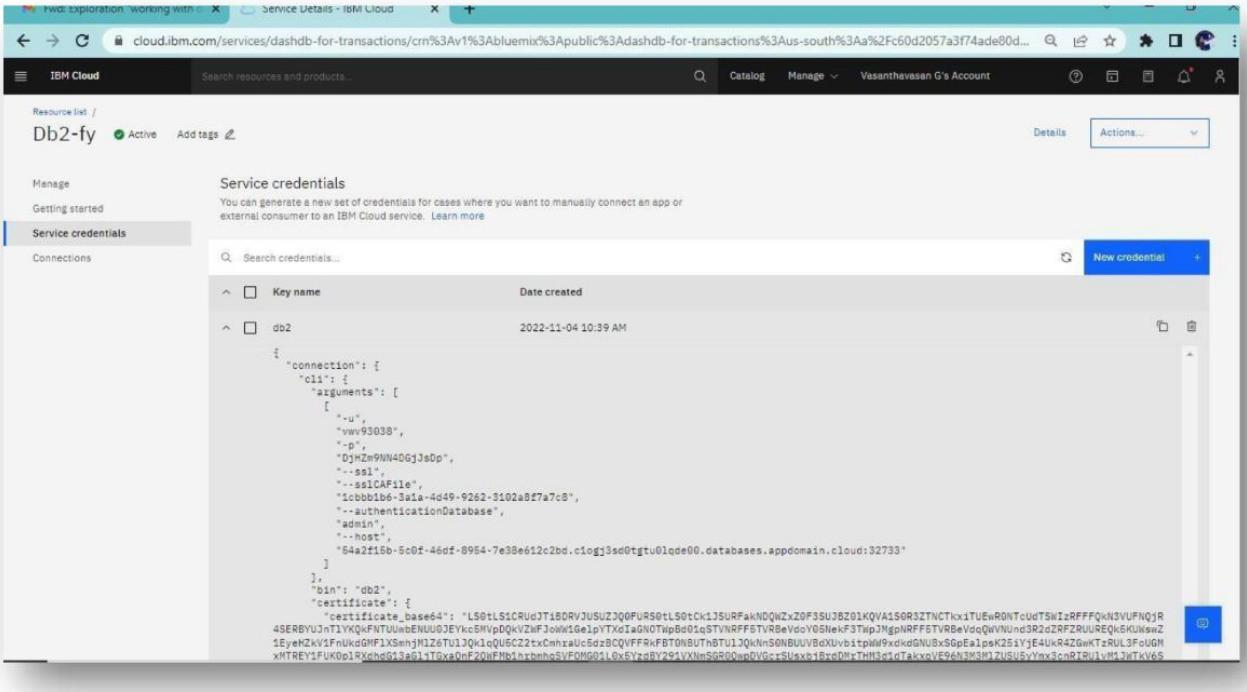
Out[98]:
```
Text(0.5, 23.52222222222222, 'Pred')
```

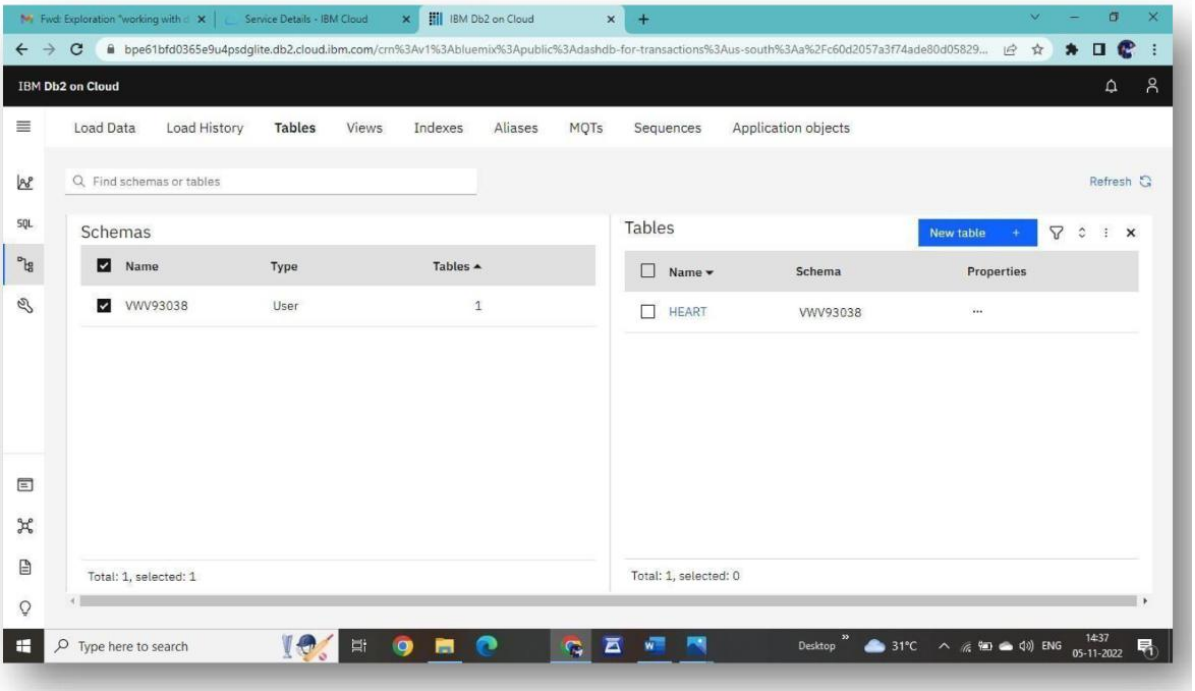


7.2 Delivery of Sprint 2: Working with Dataset

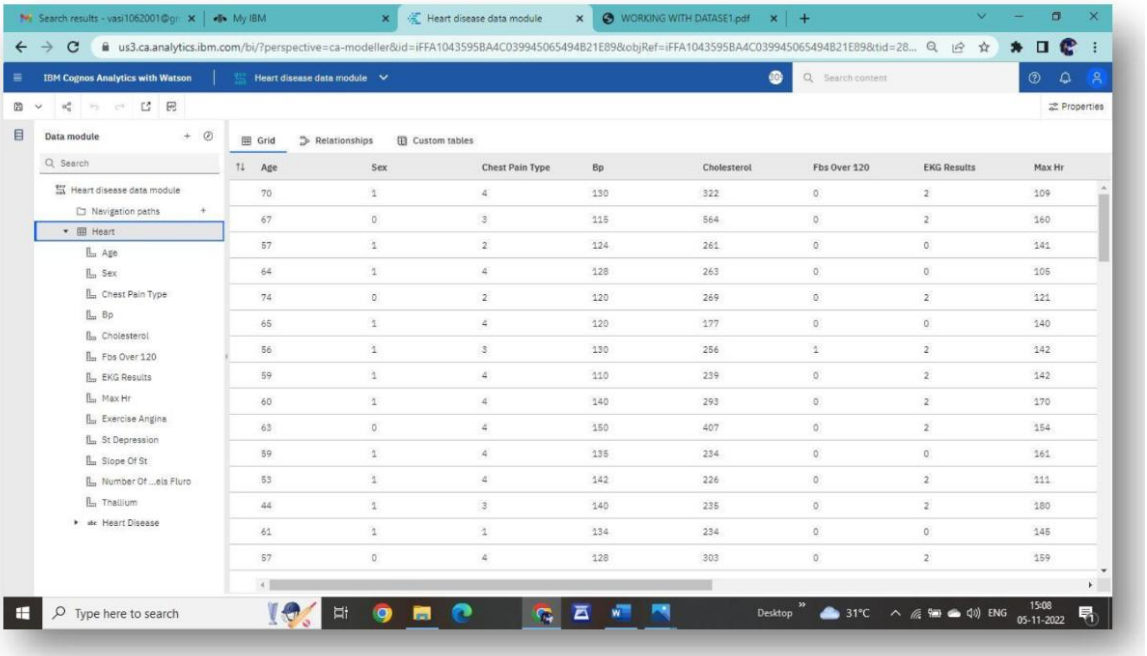


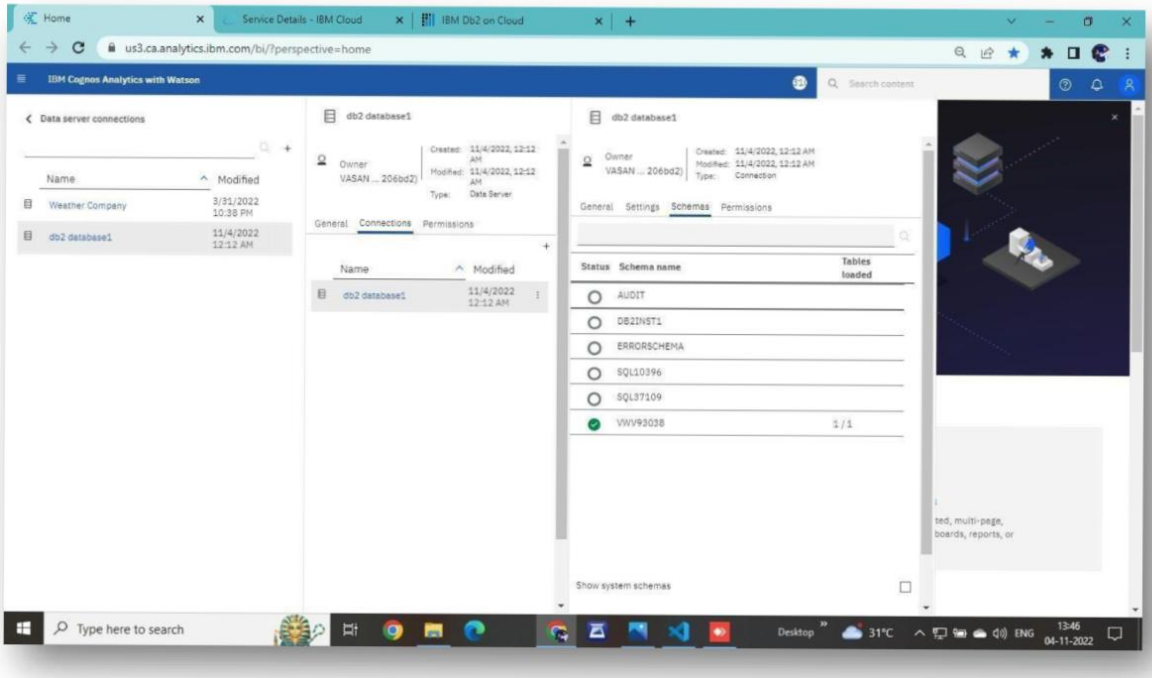
Successfully created Db2 Service Credential





Successfully connected IBM Cloud Db2 to Cognos Analytics



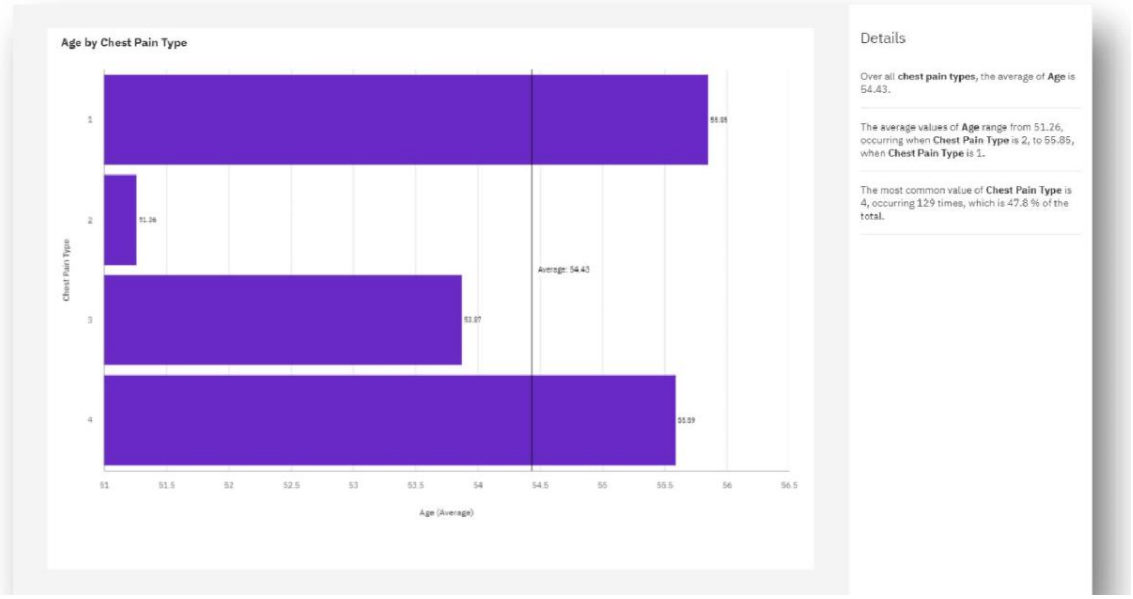


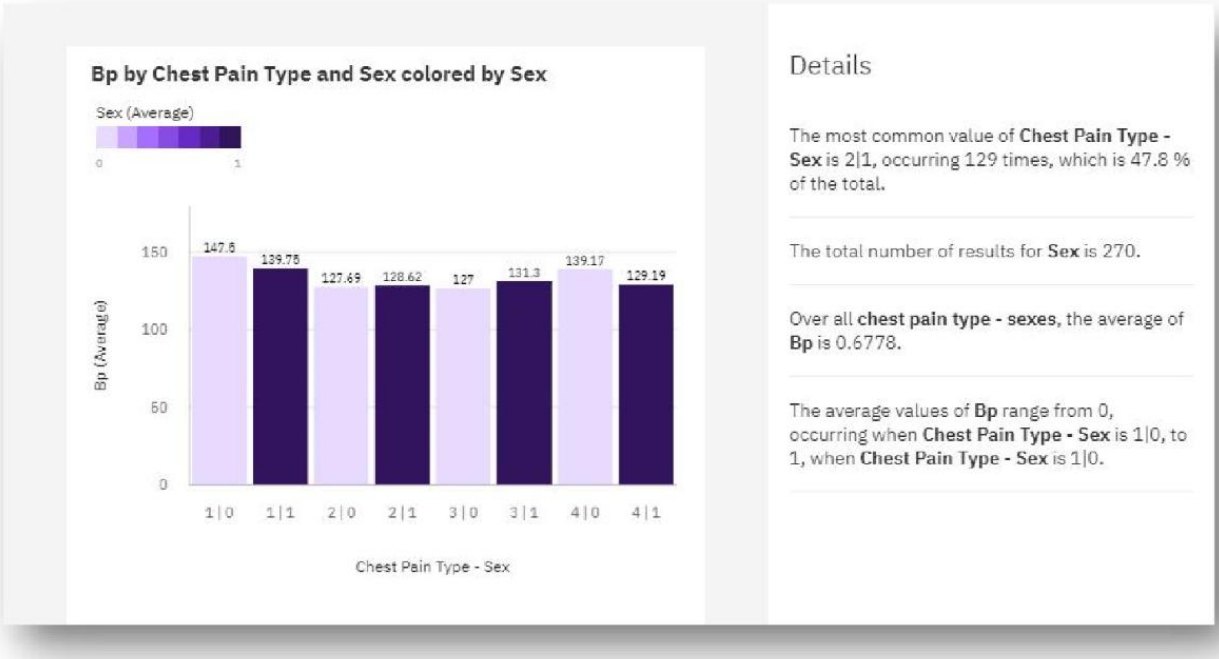
Data Preparation (Data Module)

Exploration of

Data:

Age by Chest pain type

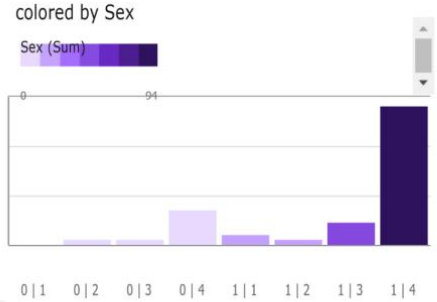




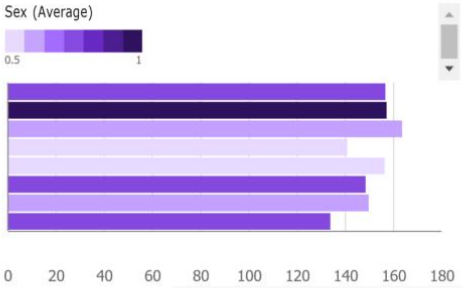
7.3 Delivery of Sprint 3: Data Visualization

Tab 10

Exercise Angina by Sex and Chest Pain Type colored by Sex

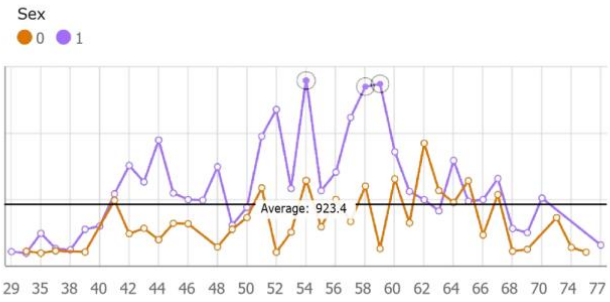


Max Hr by Chest Pain Type and Exercise Angina colored by Sex

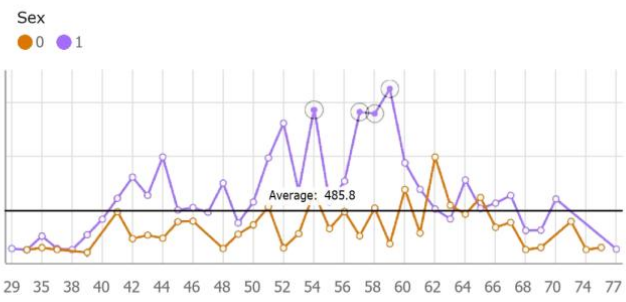


Heart Disease by Sex

Cholesterol by Age colored by Sex

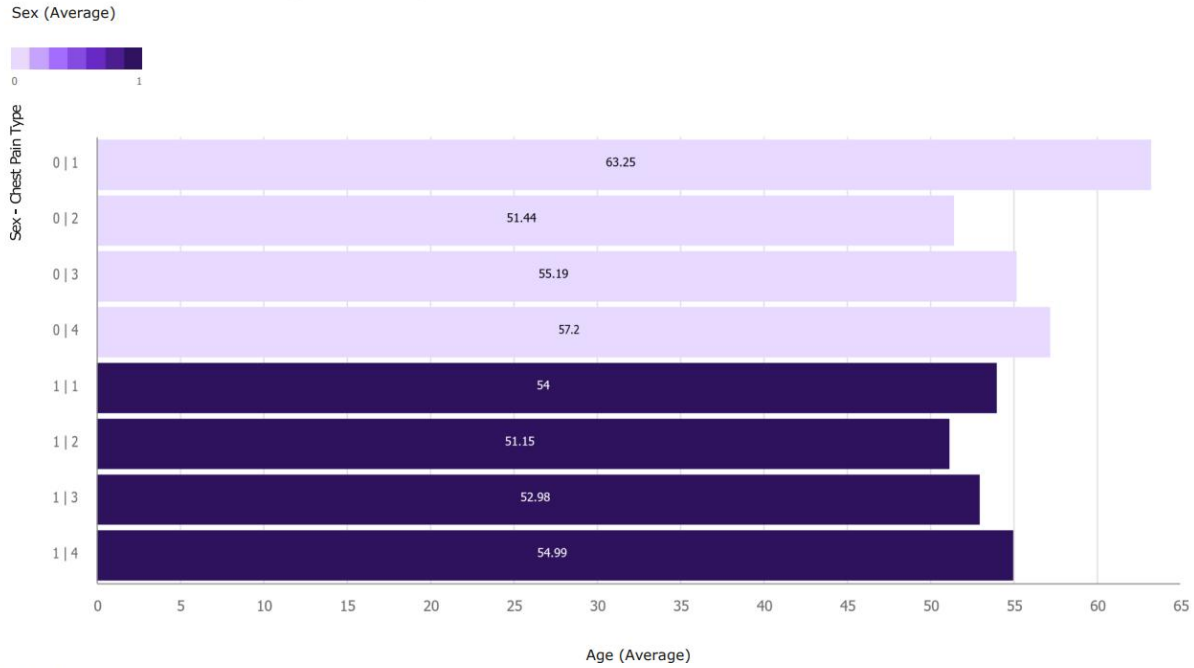


Bp by Age colored by Sex

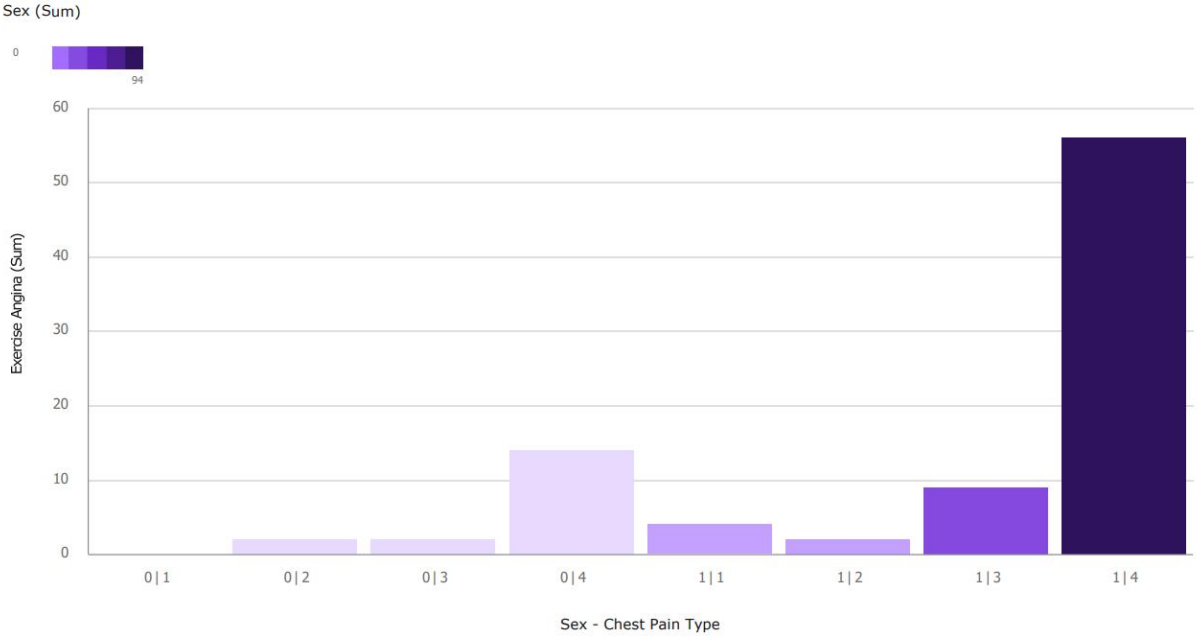


Tab 1

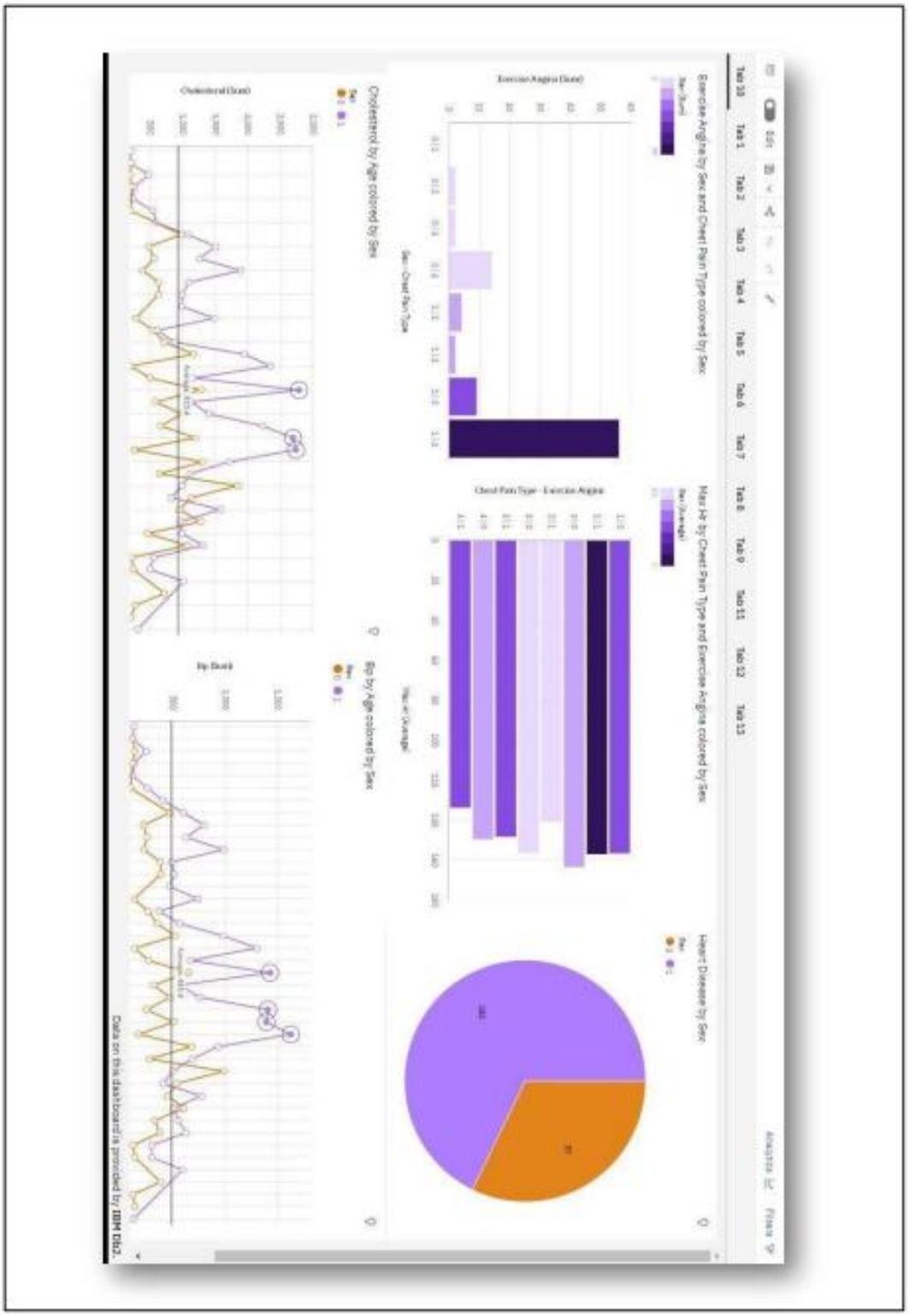
Age by Sex and Chest Pain Type colored by Male and Female



Exercise Angina by Sex and Chest Pain Type colored by Sex



7.4 Delivery of Sprint 4: Dashboard



8.CONCLUSION

Heart stroke and vascular disease are the major cause of disability and premature death. Chest pain is the key to recognize the heart disease. In this work, the heart diseases are predicted by considering major factors with four types of chest pain. K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. Here the datasets are clustered and based upon the clusters the happening of chest pain is predicted. The role of exploratory data using tableau provided a visual appealing and accurate clustering experience.

9. REFERENCE

[1] V. Manikantan & S. Latha, “Predicting the Analysis of Heart Disease Symptoms Using Medicinal Data Mining Methods”, International Journal on Advanced Computer Theory and Engineering, Volume-2, Issue-2, pp.5-10, 2013.

[2] Dr. A. V. Senthil Kumar, “Heart Disease Prediction Using Data Mining preprocessing and Hierarchical Clustering”, International Journal of Advanced Trends in Computer Science and Engineering, Volume-4, No.6, pp.07-18, 2015.

[3] Uma. K, M. Hanumathappa, “Heart Disease Prediction Using Classification Techniques with Feature Selection Method”, Adarsh Journal of Information Technology, Volume-5, Issue-2, pp.22-29, 2016

[4] Himanshu Sharma, M. A. Rizvi, “Prediction of Heart Disease using Machine Learning Algorithms: A Survey”, International Journal on Recent and Innovation Trends in Computing and Communication, Volume5, Issue-8, pp. 99-104, 2017.

[5] S. Suguna, Sakthi Sakunthala. N, S. Sanjana, S. S. Sanjana, “A Survey on Prediction of Heart Disease using Big data Algorithms”, International Journal of Advanced Research in Computer Engineering & Technology, Volume-6, Issue-3, pp. 371-378,2017.