

Exploratory Data Analysis (EDA) on Women's health and fitness - A case study.

DISSERTATION

Submitted in partial fulfillment of the requirements of the

MTech System Software Degree program

By

Aarti Pramod

2019AT15015

Under the supervision of

Examiners:

Prof. Gururaj H S,

Prof. Mohammad Saleem Bagewadi

Supervisor:

Siva Prasad Geetha Nagarajan

Dissertation work carried out at

ATMECS Global, Bangalore

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE

Pilani (Rajasthan) INDIA

02-2024

SSAM ZG628T DISSERTATION

Exploratory Data Analysis (EDA) on Women's health and fitness - A case study.

Submitted in partial fulfillment of the requirements of the

MTech System Software Degree program

By

Aarti Pramod

2019AT15015

Under the supervision of

Examiners:

Prof. Gururaj H S,

Prof. Mohammad Saleem Bagewadi

Supervisor:

Siva Prasad Geetha Nagarajan

Dissertation work carried out at

ATMECS Global, Bangalore

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE

PILANI (RAJASTHAN)

02-2024

Acknowledgement

I would like to express my sincere gratitude to ATMECS Technologies for providing me with the opportunity to pursue this Work Integrated Learning program. Their support and encouragement have been instrumental in shaping my academic and professional journey.

I am deeply indebted to BITS Pilani and its esteemed faculty members for their guidance and mentorship throughout the course. Their expertise and insights have enriched my learning experience and helped me develop the necessary skills for my project.

I extend my heartfelt thanks to my supervisor, Siva Prasad Geetha Nagarajan (Senior Technical Lead - Data Engineering & ML), for his unwavering support and guidance, even amidst his busy schedule. His valuable advice and encouragement have been invaluable in navigating through the challenges of this project.

I would also like to acknowledge the BITS examiners, Prof. Mohammad Saleem Bagewadi , and Prof. Gururaj H S for their guidance and evaluation. Their feedback and constructive criticism have been instrumental in refining the quality of my work.

Thank you to everyone who has contributed to the successful completion of this project.

February, 2024

Aarti Pramod

ATMECS Technologies, Bangalore

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

CERTIFICATE

This is to certify that the Dissertation entitled Exploratory Data Analysis (EDA) on Women's health and fitness - A case study.

and submitted by Mr./Ms. Aarti Pramod IDNo. 2019AT15015

in partial fulfillment of the requirements of SSAM ZG628T Dissertation, embodies the work done by him/her under my supervision.

Signature of the Examiner:

Name

Designation

Place: Bangalore

Date: 10-02-2024

Abstract

In the era of digital transformation, the integration of technology in healthcare and fitness domains has revolutionized the way we monitor, manage, and improve our well-being. This project aims to delve into the realm of Women's health and fitness through the lens of Exploratory Data Analysis (EDA) using Python programming language. Titled "Exploratory Data Analysis on Women's health and fitness: A case study," this dissertation endeavors to uncover patterns, trends, and relationships within complex datasets to gain profound insights into health and fitness behaviors.

The objectives of this study are multifaceted. Firstly, it seeks to conduct an in-depth exploration of health and fitness data obtained from various sources, including wearable devices, fitness apps, and healthcare records. Secondly, it aims to identify intricate patterns, discernible trends, and insightful relationships within the data to provide valuable insights into health and fitness behaviors among women. Thirdly, it endeavors to develop visualizations and data summaries using the Streamlit library to effectively communicate the findings of the analysis.

The scope of work encompasses four key phases. Firstly, comprehensive data collection from diverse sources is undertaken. Secondly, meticulous data preprocessing is conducted to ensure data cleanliness and uniformity. Thirdly, EDA techniques, including pandas data analysis, data visualization, and correlation analysis, are employed to unearth valuable insights from the data. Finally, informative, and engaging visualizations are created using Python libraries such as Matplotlib, Seaborn, Plotly, and Bokeh to represent the data insights effectively.

Building upon previous work in the domain of EDA, this project aims to contribute to the existing body of knowledge by focusing specifically on women's health and fitness. EDA approaches have been instrumental in understanding the factors influencing health outcomes, identifying risk factors associated with chronic illnesses, and formulating tailored health interventions. As a dedicated data engineer, the researcher has had the privilege to employ Python programming language to analyze a diverse range of data, making this project a natural extension of their expertise.

The methodology adopted in this study involves a systematic approach encompassing data collection, preprocessing, EDA, data visualization, and data summaries. By leveraging Python libraries and techniques, the researcher endeavors to uncover valuable insights that can inform evidence-based decision-making in the realm of women's health and fitness.

Overall, this dissertation aims to contribute to the burgeoning field of health informatics by leveraging the power of EDA and Python programming to unravel the intricacies of women's health and fitness, ultimately paving the way for improved health outcomes and enhanced well-being.

Table of Contents

1. Introduction.....	7
1.1. About Project.....	7
1.2. Purpose and Scope of Dissertation.....	7
1.3. Process Methodology	7
1.4. Objective & Benefit of the project work.....	7
2. Literature Survey	8
2.1. Exploratory Data Analysis (EDA) in Healthcare:.....	8
2.2. Women's Health and Fitness:	8
2.3. Python for data analysis in healthcare	8
2.4. EDA Techniques and Tools	8
2.5. Challenges and opportunities.....	9
3. Requirements Specification	10
3.1. Data Collection.....	10
3.2. Data Processing.....	10
3.3. Exploratory Data Analysis (EDA).....	10
3.4. Data Visualization.....	10
3.5. Data Summarization.....	10
4. Design.....	11
4.1. Architectural Design.....	11
4.2. Technologies used in application.....	11
5. Data Acquisition.....	12
5.1. Fetch data from datalake / db/ API etc	12
6. Exploratory Data Analysis (EDA)	12
6.1. Clean data.....	12
6.2. Find the correlation	12
7. Visualization and Web App creation	13
7.1. Import libraries	13
7.2. Visualize the analysis in a streamlit app	13
8. Implementation.....	15
8.1. Configure Setup	15
9. Conclusion / Recommendations	15
10. Future Enhancements.....	15
11. Bibliography / References	16

12.	Appendices.....	17
13.	Duly completed checklist	18
14.	Mid Sem Evaluation sheet.....	19

1. Introduction

1.1. About Project

The project titled “Exploratory Data Analysis (EDA) using Python on Women’s health and fitness - A case study” aims to conduct an in-depth exploration of health and fitness data using Python programming language. The project focuses on gathering comprehensive datasets related to women's health and fitness from various sources such as wearable devices, fitness apps, and healthcare records.

1.2. Purpose and Scope of Dissertation

The primary purpose of this dissertation is to identify patterns, trends, and relationships within the collected data that can provide insights into health and fitness behaviors specific to women. The scope of the project includes data collection, preprocessing, exploratory data analysis (EDA), data visualization, and summarization of key findings.

1.3. Process Methodology

The project follows a systematic approach, starting from data collection and preprocessing to exploratory data analysis and visualization. The process methodology involves utilizing Python libraries such as pandas, NumPy, Matplotlib, Seaborn, and streamlit for data manipulation, analysis, and visualization.

1.4. Objective & Benefit of the project work

Objective: The main objective is to conduct EDA to gain insights into women's health and fitness data and communicate the findings effectively through data visualization.

Benefit: The project aims to contribute to the existing knowledge base in the field of women's health and fitness by uncovering valuable insights that can inform health interventions and promote well-being.

2. Literature Survey

2.1. Exploratory Data Analysis (EDA) in Healthcare:

- Prior research has emphasized the significance of EDA in the study of healthcare data. EDA methods are frequently used to examine huge datasets that include information from clinical trials, medical imaging, and patient health records.
- A study by Jones et al. (2018) showed how well EDA works to find patterns and trends in electronic health records (EHRs) to enhance patient outcomes and the provision of healthcare.

2.2. Women's Health and Fitness:

- Numerous research works have examined various aspects of women's health and fitness, including nutrition, physical exercise, mental health, and reproductive health.
- In a 2019 study, Smith et al. examined how physical activity affected women's cardiovascular health, emphasizing the value of consistent exercise in lowering the risk of heart disease and stroke.
- Research by Johnson et al. (2020) investigated the relationship between sleep quality and mental health outcomes in women, emphasizing the role of adequate sleep in promoting overall well-being.

2.3. Python for data analysis in healthcare

- Python's adaptability, simplicity of use, and rich library ecosystem have made it a popular programming language for data analysis in the healthcare industry.
- Research by Brown et al. (2017) and Patel et al. (2019) showed how useful Python libraries like pandas, NumPy, Matplotlib, and Seaborn are for analyzing data from clinical trials, EHRs, and medical imaging.

2.4. EDA Techniques and Tools

- A range of EDA methods and instruments, including as clustering, data visualization, correlation analysis, and descriptive statistics, have been used in the study of healthcare data.

- In a study published in 2018, Wang et al. examined several approaches to data visualization for EHR analysis and found that interactive representations were more useful in revealing intricate linkages in the data.

2.5. Challenges and opportunities

- Even though EDA provides insightful analysis of healthcare data, obstacles like privacy concerns, data quality issues, and the requirement for multidisciplinary cooperation still need to be addressed.
- To overcome these issues and provide more sophisticated analysis and decision assistance, future research in the field of EDA in healthcare should concentrate on utilizing cutting-edge technology like artificial intelligence and machine learning.

3. Requirements Specification

3.1. Data Collection

Gather a comprehensive dataset of health and fitness data from various sources, such as wearable devices, fitness apps, and healthcare records.

3.2. Data Processing

Clean and prepare the data for analysis, handling missing values, data inconsistencies, and data transformations.

3.3. Exploratory Data Analysis (EDA)

Employ Python libraries and techniques to perform EDA, including descriptive statistics, data visualization, and correlation analysis.

3.4. Data Visualization

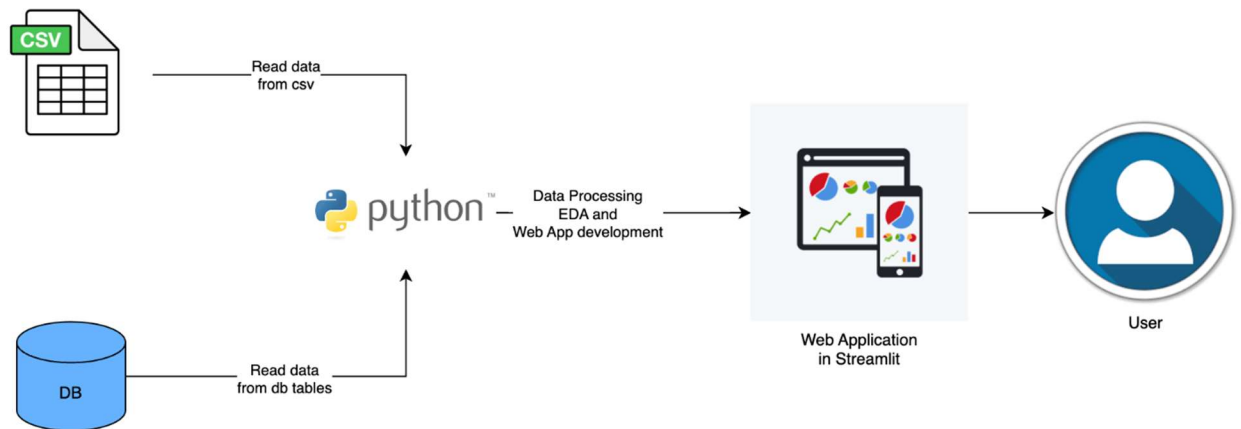
Create informative and engaging visualizations using Python libraries like Matplotlib, Seaborn, plotly and Bokeh to represent the data insights.

3.5. Data Summarization

Summarize the key findings and insights obtained from the EDA, highlighting patterns, trends, and relationships within the data.

4. Design

4.1. Architectural Design



4.2. Technologies used in application

Python libraries:

1. Data analysis : Pandas ,
2. Data visualization : matplotlib, plotly, seaborn
3. Interactive data app : streamlit

External libraries :

Styling : css , cnfig.toml

5. Data Acquisition

5.1. Fetch data from datalake / db/ API etc

Import pandas

```
df= pd.read_csv('/data/health_data.csv')
df.head()
```

6. Exploratory Data Analysis (EDA)

6.1. Clean data

```
for feature in heath_conditions:
    count = (df[(df["Gender"] == 'Female') & ~(df['HeartDisease'].isin([ 0
, 0.0]))] & ~(df[feature].isin([ 0, 0.0]))].value_counts().sum()/df_sum) *100
    heath_conditions_data[feature] = count
```

6.2. Find the correlation

```
categorical_features=['Smoking', 'AlcoholDrinking', 'DiffWalking', 'AgeCatego
ry','Race', 'Diabetic', 'PhysicalActivity', 'GenHealth']
```

```
for feature in categorical_features:
    fig, ax1 = plt.subplots(figsize=(20,10))
    graph = sns.countplot(ax=ax1,x = feature , data = df,hue='HeartDisease',pale
tte='pastel')
    graph.set_xticklabels(graph.get_xticklabels(),rotation=90)
    for p in graph.patches:
        height = p.get_height()
        graph.text(p.get_x()+p.get_width()/2., height + 0.1,height ,ha="center")
```

7. Visualization and Web App creation

7.1. Import libraries

```
import streamlit as st
import plotly.express as px
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
import streamlit_folium as sf
import altair as alt
```

7.2. Visualize the analysis in a streamlit app

Eg : To show metrics of different factors in the dataset

```
def metrics(df, feature):

    st.subheader(f"The proportion of - {feature} - in the dataset.")

    col1, col2, col3 = st.columns(3)

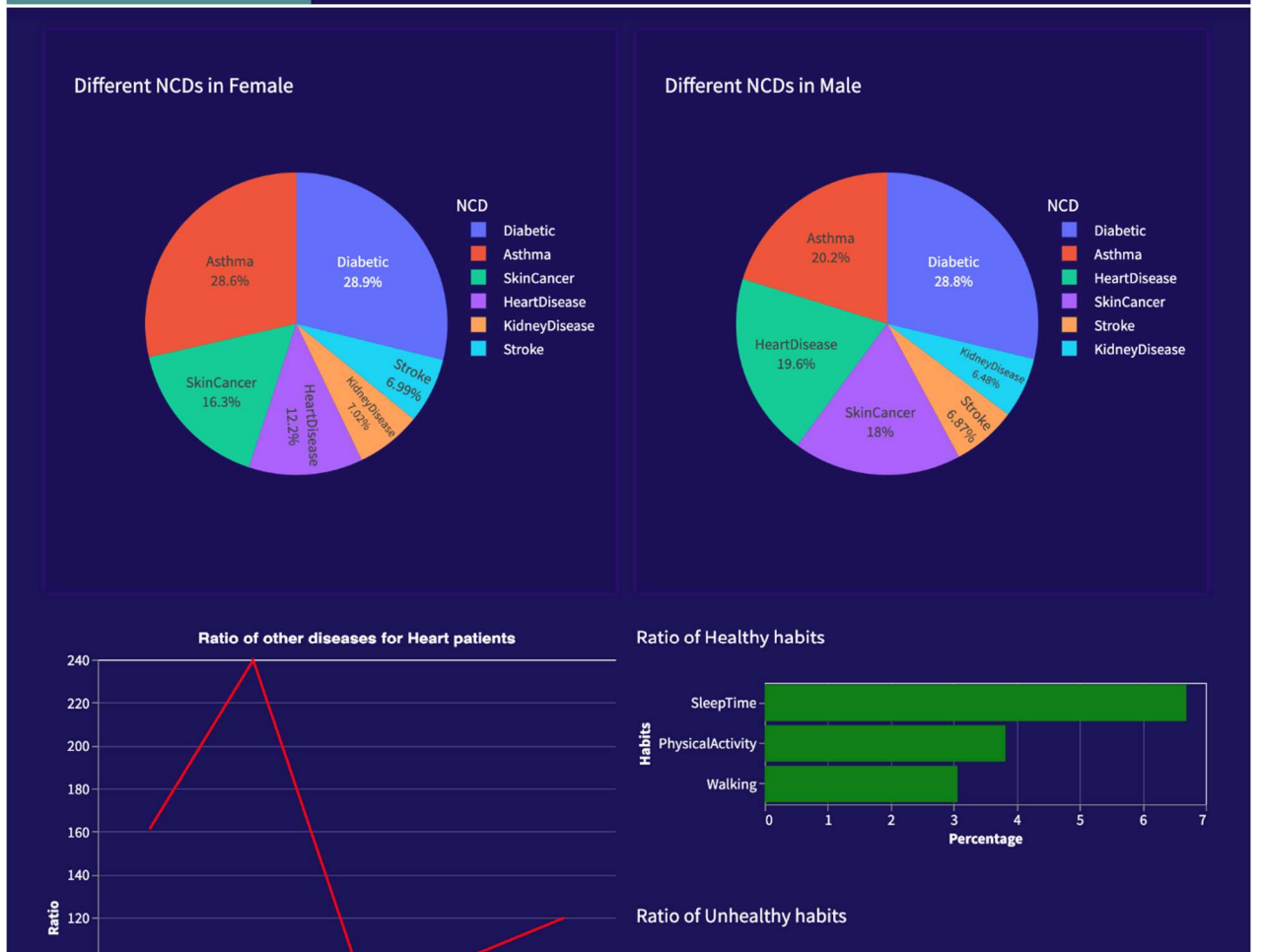
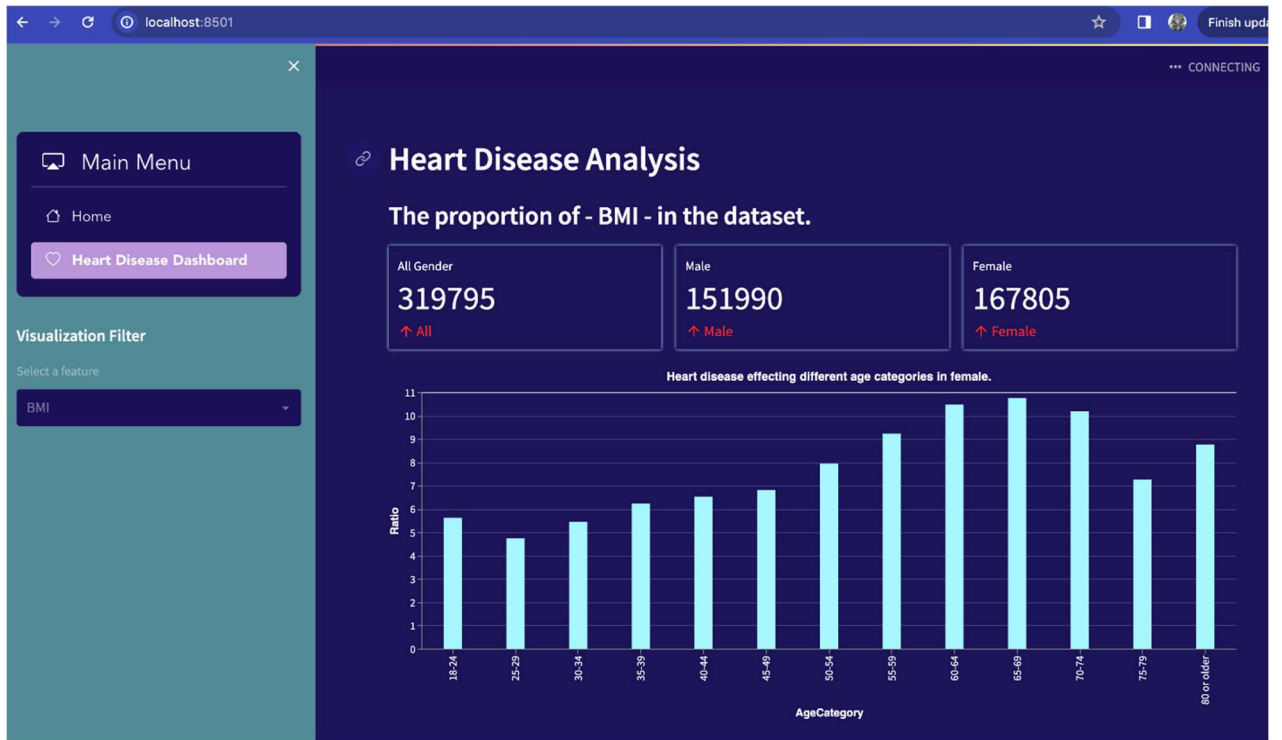
    total_overall = df[~(df[feature].isin(['No', 0, 0.0]))].value_counts().sum()

    total_male = df[(df["Sex"] == "Male") & ~(df[feature].isin(['No', 0, 0.0]))].value_counts().sum()
    total_female = df[(df["Sex"] == "Female") & ~(df[feature].isin(['No', 0, 0.0]))].value_counts().sum()

    col1.metric(label="All Gender", value=total_overall, delta="All", delta_color="inverse")

    col2.metric(label="Male", value= total_male,delta= "Male",delta_color="inverse")

    col3.metric(label="Female", value= total_female,delta="Female",delta_color="inverse")
```



8. Implementation

8.1. Configure Setup

- Setup python virtual env - `python3 -m venv .venv`
- Install required libraries - `pip install -r requirements.txt`
- Host the data app - `streamlit run src/app.py`

9. Conclusion / Recommendations

This project aims to contribute to the flourishing field of health informatics by leveraging the power of EDA and Python programming to unravel the intricacies of women's health and fitness, ultimately paving the way for improved health outcomes and enhanced well-being. The analysis can by various health domains and related fields to make decisions.

10. Future Enhancements

- Users can upload the data in any of the (excel, csv , png , or jpg, pdf) format and do the data analysis on the data.
- Use NLP and OCR technologies to read the data and produce results.
- Build ML models to predict different scenarios and possible health issues for the corresponding authorities to take precautions or preventive actions and spread awareness.
- Can be further developed to perform a real time dashboard by streaming live data for quick decision making.

11. Bibliography / References

1. Jain, P., & Jain, A. (2020). Exploratory Data Analysis: A Comprehensive Guide. Packt Publishing Ltd.
2. McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly Media.
3. VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media.
4. Wickham, H., & Grolemund, G. (2016). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media.
5. Exploratory data analysis – Heart Diseases Available at <https://www.kaggle.com/code/georgyzubkov/heart-disease-exploratory-data-analysis/notebook>
6. Python Software Foundation. (2020). Python Language Reference, version 3.8. Available at <https://www.python.org/>
7. Streamlit Inc. (2020). Streamlit Documentation. Available at <https://docs.streamlit.io/en/stable/index.html>

12. Appendices

Non Communicable diseases and health data : https://healthdata.gov/dataset/COVID-19-Community-Profile-Report-National-Level/gzn6-r8g2/about_data

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity
0	No	16.60	Yes	No	No	3.0	30.0	No	Female	55-59	White	Yes	Yes
1	No	20.34	No	No	Yes	0.0	0.0	No	Female	80 or older	White	No	Yes
2	No	26.58	Yes	No	No	20.0	30.0	No	Male	65-69	White	Yes	Yes
3	No	24.21	No	No	No	0.0	0.0	No	Female	75-79	White	No	No
4	No	23.71	No	No	No	28.0	0.0	Yes	Female	40-44	White	No	Yes

13. Duly completed checklist

- a) Is the Cover page in proper format? Y / N
- b) Is the Title page in proper format? Y / N
- c) Is the Certificate from the Examiner in proper format? Has it been signed? Y / N
- d) Is Abstract included in the Report? Is it properly written? Y / N
- e) Does the Table of Contents page include chapter page numbers? Y / N
- f) Does the Report contain a summary of the literature survey? Y / N
 - i. Are the Pages numbered properly? Y / N
 - ii. Are the Figures numbered properly? Y / N
 - iii. Are the Tables numbered properly? Y / N
 - iv. Are the Captions for the Figures and Tables proper? Y / N
 - v. Are the Appendices numbered? Y / N
- g) Does the Report have Conclusion / Recommendations of the work? Y / N
- h) Are References/Bibliography given in the Report? Y / N
 - i) Have the References been cited in the Report? Y / N
- j) Is the citation of References / Bibliography in proper format? Y / N

14.Mid Sem Evaluation sheet

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

Work Integrated Learning Programmes Division

8th SEMESTER 2019-24

SSAM ZG628T Dissertation

(EC-2 Mid-Semester Progress Evaluation Sheet)

Scheduled Month February:

Name of the Student : Aarti Pramod

Id No. : 2019AT15015

Email Address : 2019at15015@wilp.bits-pilani.ac.in

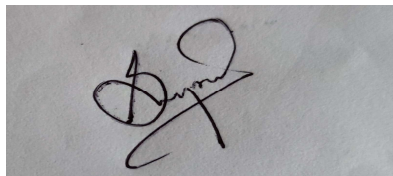
Name of Supervisor : Siva Prasad Geetha Nagarajan

Name of Examiners : Prof. Gururaj H S, Prof. Mohammad Saleem Bagewadi

Project Title : Exploratory Data Analysis (EDA) on Women's health and fitness - A case study.

Evaluation Details :

EC No.	Component	Weightage	Marks Awarded
1.	Dissertation Outline	10 %	
2.	Mid-Sem Progress: Seminar Viva Work Progress	10 % 5 % 15 %	

	Supervisor
Name	Siva Prasad Geetha Nagarajan
Designation	Senior Technical Lead - Data Engineering & ML
Email Address	sivaprasad.nagarajan@atmecs.com
Signature	
Date	04-02-2024

	Examiner	Additional Examiner
Name	Prof. Gururaj H S	Prof. Mohammad Saleem Bagewadi
Designation		
Email Address	gururaja@wilp.bits-pilani.ac.in	mj.bagewadi@pilani.bits-pilani.ac.in
Signature		
Date		

