**DATA SCIENCE TOOLBOX: PYTHON PROGRAMING**

**PROJECT REPORT**

(Project Semester January-April 2025)

***Visualization and Analysis of Health Activity***

Submitted by:Adarsh Yadav Registration No: 12321805

Programme and Section B.Tech (CSE) K23EP

Course Code INT375

A logo with orange and black text

AI-generated content may be incorrect.

Under the Guidance of

**Dr. Tamina Thakur**

**Discipline of CSE/IT**

**Lovely School of Computer Science**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that Adarsh Yadav bearing Registration no. 12321805 has completed INT375 project titled, **“**Visualization and Analysis of Health Activity**”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Dr. Tamina Thakur**

**Faculty**

**School of Computer Science**

Lovely Professional University

Phagwara, Punjab.

Date:

**DECLARATION**

I, Adarsh Yadav, student of B.Tech CSE under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 12/04/2025 Signature

Registration No. 12321805 Adarsh Yadav

**TABLE OF CONTENTS**

| **S. No.** | **Content** | **Page No.** |
| --- | --- | --- |
| 1 | Introduction | [Introduction](#Intro) |
| 2 | Source of Dataset | [Source](#dataset) |
| 3 | Data Pre Processing | [Data Preprocessing](#dataP) |
| 4 | Exploratory Data Analysis (EDA) Process | [EDA](#EDA) |
| 5 | Analysis on Dataset | [Analysis](#Objective) |
| 6 | Conclusion | [Conclusion](#conslu) |
| 7 | Future Scope | [Future](#FutureScope) |
| 8 | References | [Reference](#re) |
| 9 | LinkedIn Post | [Post](#LI) |

Introduction

This data science project is centred around analysing a health activity dataset using Python to extract meaningful insights and patterns related to physical wellness and lifestyle habits. The objective is to work with raw structured data and apply fundamental data science techniques to interpret trends in metrics such as steps taken, calories burned, active minutes, heart rate, and sleep duration. By using Python libraries such as Pandas for data manipulation and Matplotlib and Seaborn for visualizations, the project demonstrates how raw health data can be translated into understandable and actionable information.

Visualizations are used to support the findings and to present the data in a clear and informative manner. Line plots, bar charts, histograms, and correlation heatmaps help communicate the results effectively, making it easier to interpret patterns in daily activities, health behavior, and wellness trends. Overall, this project showcases the practical application of Python in personal health data analysis and highlights how structured programming and basic data science techniques can be used to promote healthier lifestyle decisions through data-driven insights.

Source of Dataset

https://github.com/adarshyadav62/healty-activity-data/blob/main/health\_activity\_data.csv

Data Preprocessing

As part of the data preprocessing process in this project, specific transformations were applied to make the dataset more suitable for analysis and visualization. One of the first steps included standardizing column names for consistency and converting date-related fields into a uniform format to allow for accurate time-based analysis. Any timestamps were simplified to just the date or extracted into separate components like day, month, and year to support daily and weekly trend analysis.

Missing values were addressed by either imputing with suitable averages (such as mean sleep duration or average step count) or removing incomplete entries when necessary. Additionally, numerical fields such as step count, calories burned, and heart rate were checked for anomalies and outliers that could distort the analysis. These values were treated through filtering or normalization to ensure a consistent range for visual interpretation.

Categorical variables, where present (such as activity types or intensity levels), were converted into appropriate formats using label encoding or mapping schemes. The dataset was also grouped and aggregated by time intervals (daily, weekly, or monthly) to make trend analysis more insightful and readable.

These preprocessing steps played a critical role in preparing the health activity dataset for accurate analysis and meaningful visualizations that reflect personal wellness and activity patterns.

A computer screen shot of colorful text

AI-generated content may be incorrect.

**Data Visualization**

Data visualizations played a key role in making these insights interpretable. A **count plot** was used to visualize the gender distribution in the dataset, helping to identify whether the data was balanced between male and female participants. This helped ensure that further comparisons between genders were meaningful and not biased due to unequal representation.

Other visualizations included **line plots** to track daily and weekly trends in steps and calories burned, **bar charts** to compare average activity levels, **histograms** to show the distribution of sleep durations and heart rate values, and **heatmaps** to explore correlations among key health indicators. These visualizations were created using Python libraries such as Matplotlib and Seaborn, which provided flexibility in customizing and interpreting the results.

Through these visual representations, we were able to highlight patterns in physical activity, detect variations across different demographics, and uncover relationships between different health behaviors. These findings support the development of healthier lifestyle strategies by making complex data both accessible and

A screen shot of a computer

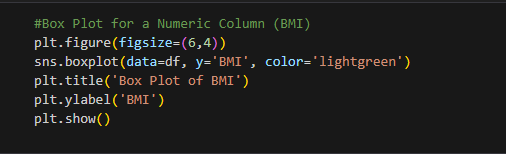
AI-generated content may be incorrect.

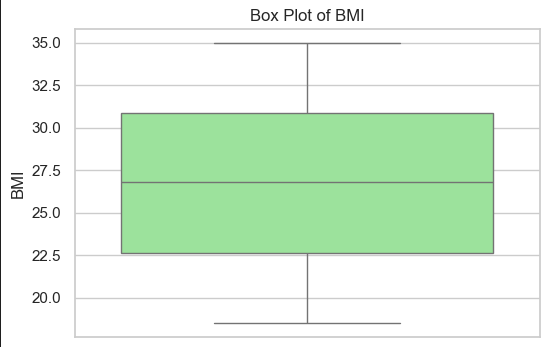
A computer screen with colorful text

AI-generated content may be incorrect.

A graph of heart rate

AI-generated content may be incorrect.





A computer screen with text

AI-generated content may be incorrect.

A screen shot of a graph

AI-generated content may be incorrect.

**Analysis on dataset**

1) The Relationship Between Age and BMI:

**Introduction:**  
This section of the project uses data visualization techniques to explore the relationship between age and Body Mass Index (BMI). Understanding this relationship helps in identifying how body composition changes with age and whether any trends or patterns exist that could impact health recommendations.

**Specific Requirements:**  
The specific goal is to visualize the relationship between age and BMI using a scatter plot. This helps to observe whether there is a positive, negative, or no correlation between these two variables across different age groups.

**Analysis Result:**  
The x-axis of the scatter plot represents the age of individuals, while the y-axis shows their BMI values. The distribution of points helps to identify trends or clusters that may indicate health patterns related to age.

**Visualization:**

A computer screen shot of a program

AI-generated content may be incorrect.

A diagram of age and bmi distribution

AI-generated content may be incorrect.

2) Daily Steps in Relation to Heart Disease Status:

**Introduction:**  
This section of the project uses data visualization techniques to analyze how daily step count varies between individuals with and without heart disease. This comparison helps identify whether physical activity levels are linked to heart health.

**Specific Requirements:**  
The specific goal is to visualize daily step distribution by heart disease status using a box plot. This provides a clear picture of the range, median, and variability in step counts for both groups—those diagnosed with heart disease and those without.

**Analysis Result:**  
The x-axis of the box plot represents the heart disease status (Yes or No), while the y-axis shows the number of daily steps. The spread and median of each box indicate the typical step count and how it differs between the two groups.

A graph showing a diagram

AI-generated content may be incorrect.A computer screen shot of a program

AI-generated content may be incorrect.Visualization:

3) The Relationship Between Sleep Duration and Heart Rate:

**Introduction:**  
This section of the project focuses on understanding the relationship between the number of hours slept and the average heart rate. Sleep is a vital component of overall health, and this analysis aims to explore whether longer or shorter sleep durations have any visible impact on heart rate patterns.

**Specific Requirements:**  
The specific goal is to sort the dataset by sleep duration and visualize the trend between sleep hours and heart rate using a line plot or scatter plot. This helps to detect potential correlations or irregularities that may reveal how rest affects cardiovascular health.

**Analysis Result:**  
The x-axis of the plot represents the number of hours slept, while the y-axis displays the corresponding heart rate values. Sorting the data by sleep hours ensures that trends are easily visible, helping to interpret whether more sleep generally corresponds to a higher or lower resting heart rate.

A graph showing a sleep hours and heart rate

AI-generated content may be incorrect.A computer screen shot of a program code

AI-generated content may be incorrect.Visualization:

4) The Relationship Between Smoking Habits and Heart Disease Prevalence:

**Introduction:**

This section of the project aims to investigate the potential association between smoking habits and the prevalence of heart disease within the dataset. Understanding this relationship is crucial for identifying risk factors and informing public health initiatives.

**Specific Requirements:**

The specific goal is to calculate the percentage of individuals with heart disease within the smoker and non-smoker groups. This will then be visualized using a bar plot, where one bar represents the percentage of heart disease among smokers and another bar represents the percentage among non-smokers.

**Analysis Result:**

The bar plot will have two distinct bars along the x-axis, labeled "Smoker" and "Non-Smoker". The height of each bar on the y-axis will represent the percentage of individuals within that group who have been diagnosed with heart disease. This visual comparison will allow for a clear interpretation of whether smoking is associated with a higher or lower prevalence of heart disease in this dataset.

A graph of a heart disease

AI-generated content may be incorrect.A screenshot of a computer screen

AI-generated content may be incorrect.Visualization:

5) The Overall Prevalence of Heart Disease in the Dataset

**Introduction:**

This section of the project focuses on visualizing the overall proportion of individuals within the dataset who have been diagnosed with heart disease compared to those who have not. This provides a general understanding of the disease prevalence in the studied population.

**Specific Requirements:**

The specific goal is to count the number of individuals with heart disease and the number of individuals without heart disease. These counts will then be used to calculate the percentage of each group relative to the total number of individuals in the dataset. This will be visualized using a suitable chart, such as a pie chart or a bar plot with two categories.

**Analysis Result:**

The visualization will clearly display the percentage of individuals with heart disease and the percentage of individuals without heart disease. If a pie chart is used, each slice will represent one of the two groups, with the size of the slice proportional to the percentage.

Visualization:

A computer screen shot of text

AI-generated content may be incorrect.

A green circle with red triangle and red triangle

AI-generated content may be incorrect. Conclusion

In conclusion, this project provided valuable insights into various aspects of health activity through effective data analysis and visualization. By exploring the dataset, we were able to identify trends related to sleep duration and heart rate, as well as the prevalence of heart disease in relation to smoking habits and overall. Visual tools like line plots, scatter plots, and bar plots made complex health data more accessible and highlighted potential correlations and proportions. These findings not only enhance our understanding of the relationships between different health factors but also lay the groundwork for more advanced analytical applications such as risk factor identification and population health monitoring. Overall, this analysis demonstrates the power of data-driven decision-making in the healthcare domain and opens the door for further innovation and strategic planning for health and wellness initiatives.

**Future Scope**

Building upon the insights gained from this initial visualization and analysis of health activity data, several promising avenues for future research and development emerge:

* **Predictive Modeling:** The identified relationships and trends can be leveraged to build predictive models. For example, machine learning algorithms could be trained to predict the likelihood of heart disease based on factors like sleep duration and smoking habits. Similarly, models could be developed to forecast potential heart rate patterns based on sleep schedules.
* **Longitudinal Analysis:** Expanding the analysis to include time-series data would allow for the examination of how health activities and indicators change over time for individuals. This could reveal valuable insights into the long-term effects of lifestyle choices and the progression of health conditions. Visualizations could then depict these temporal trends.
* **Integration of Additional Health Metrics:** Incorporating other relevant health metrics, such as physical activity levels, dietary information, stress levels, and blood pressure, would provide a more comprehensive understanding of overall health. Visualizing the interplay between these multiple factors could uncover more nuanced relationships.
* **Interactive Dashboards for Personalized Insights:** Developing interactive dashboards that allow users to explore their own health data and visualize personalized trends could empower individuals to take a more proactive role in managing their well-being. These dashboards could present summaries of key metrics, highlight potential areas of concern, and track progress towards health goals.
* **Comparative Analysis Across Demographics:** Analyzing the data across different demographic groups (e.g., age, gender, geographical location) could reveal disparities in health activities and disease prevalence, informing targeted public health interventions. Visualizations could highlight these differences and inequalities.
* **Real-time Monitoring and Alerts:** With the integration of wearable technology and real-time data streams, future work could focus on developing systems that provide real-time monitoring of health indicators and generate alerts for potentially concerning patterns, enabling timely interventions.
* **Causal Inference:** While this initial analysis focuses on correlations, future research could explore causal relationships between health activities and outcomes using more advanced statistical techniques. Understanding causality can lead to more effective interventions and recommendations.
* **Development of Recommendation Systems:** Based on the identified patterns and predictive models, recommendation systems could be developed to suggest personalized interventions or lifestyle adjustments to improve health outcomes. For instance, a system could recommend optimal sleep durations or smoking cessation programs based on an individual's profile.
* **Integration with Electronic Health Records (EHRs):** Linking this type of analysis with EHRs could provide healthcare professionals with valuable insights into patient health behaviors and trends, supporting more informed clinical decision-making. Visualizations could summarize key patient information for easier interpretation.

Reference

* Pandas Documentation: <https://pandas.pydata.org/docs>
* Matplotlib Documentation: <https://matplotlib.org/stable/contents.html>
* Seaborn Documentation: <https://seaborn.pydata.org>
* Python Official Website: <https://www.python.org>
* YouTube: <https://youtu.be/tjIWRqqMDaw?si=4sAZtKVb1d3kQSvp>

LinkedIn Link:-

<https://www.linkedin.com/posts/adarsh-yadav-697a6728b_datascience-python-healthanalytics-activity-7317244056118390785-i6YP?utm_source=share&utm_medium=member_desktop&rcm=ACoAAEZ1QRMBco1tD4X9hQvb1klmcTvovEPCvYU>

Github Link:-

https://github.com/adarshyadav62/Visualization-and-Analysis-of-Health-Activity.git