Strava Fitness Data Analysis Project Report

- **Project Title** : STRAVA FITNESS DATA ANALYSIS
- **Author** : Sanyukti S. Wankhade
- **Mentor** : Vipul Sonwane
- **Company** : LABMENTIX
- **Date** : 15.06.2025
- **Tools Used** : PYTHON, POWER BI, SQL, EXCEL

2. INTRODUCTION

In today’s digital age, wearable devices and fitness trackers have become a part of many people's daily lives. These tools generate a lot of valuable data—how many steps we take, how well we sleep, how our heart behaves during exercise, and more. This project explores such fitness data to better understand daily habits, health patterns, and physical performance. By studying these trends, we can gain meaningful insights into how users manage their fitness and what factors influence their overall well-being.

Throughout the internship, I worked with multiple fitness datasets like daily activity, sleep records, heart rate readings, and hourly statistics. I used SQL to clean and filter the raw data, removing duplicates and correcting formats. With Python, I built graphs and charts using libraries like Pandas, Matplotlib, and Seaborn. Then, using Power BI, I created easy-to-understand dashboards that helped visualize user behavior, such as sleep quality, step count patterns, and calorie burn across the day.

This project gave me a complete view of fitness from a data perspective. I was able to identify when users are most active, how sleep impacts next-day performance, and which hours show the highest intensity. It was an exciting experience to turn raw numbers into real-world insights that can actually help people make better lifestyle and fitness decisions

3. TOOLS & TECHNOLOGIES USED

This project required working with various tools and technologies to handle raw fitness data, clean it, perform exploratory analysis, and create compelling visualizations. Each tool served a unique purpose in the data analysis pipeline

3.1. SQL (Structured Query Language):

SQL was primarily used during the initial phase of data processing. I imported all the CSV files into a MySQL database and used SQL queries to clean and transform the datasets. This included removing rows with null values, eliminating duplicate records, and filtering unnecessary columns. Additionally, I formatted date and time columns to ensure consistency across all files. SQL provided a structured and efficient way to handle large volumes of fitness data, which was essential before proceeding with visualization or analysis.

3.2. Power BI:

Power BI was the main tool used for creating interactive dashboards. After importing the cleaned data from Excel or SQL, I used Power Query Editor to further transform the data—for example, separating combined Date Time columns into individual **Date** and **Time** fields for better analysis. The visualizations I created included **line charts, bar charts, column charts, tables, cards, and slicers**. These visuals allowed users to filter data based on time, user ID, or activity type, and view real-time insights in a clean and interactive format. Relationships between tables were built to enable dynamic filtering and cross-reporting across different datasets.

3.3. Python (VS Code – Pandas, Matplotlib, Seaborn)**:**

Python was used to perform deeper statistical analysis and generate static visualizations. Using **Pandas**, I loaded and prepared the data by grouping, filtering, and summarizing values. With **Matplotlib** and **Seaborn**, I developed several graphs like line plots to show heart rate trends over time, bar charts for hourly step counts, and heatmaps to observe sleep quality by hour. These visuals helped uncover hidden patterns and supported the findings from Power BI. Visual Studio Code was used as the code editor for all Python scripting, making the workflow smooth and organized.

3.4. Excel/CSV Files**:**

The raw data was provided in CSV format. Before importing into SQL or Power BI, I used Excel for initial inspection, quick fixes, and file formatting. Excel also helped in manually checking values, verifying column headers, and preparing lookup sheets for relationship building in Power BI.

4. DATASET OVERVIEW

- The following datasets were analysed during this project. Each file includes a unique user Id, along with corresponding date, time, and specific activity measurements like heart rate, steps, calories, and sleep.
- These datasets were cleaned, transformed, and used for visual analytics in SQL, Power BI, and Python.

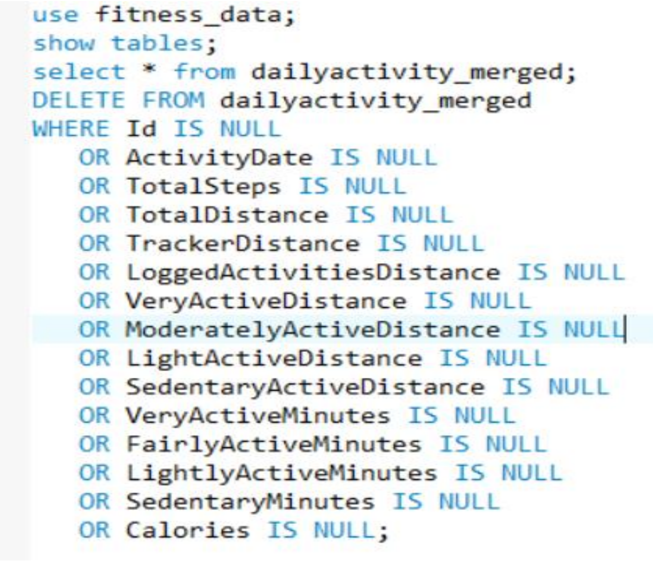
File Name	Key Columns	Description
DailyActivity_merged.csv	Id, ActivityDate, TotalSteps, TotalDistance, TotalCalories	Captures daily activity of users including step count, distance travelled, and total calories burned in a day.
HeartRate_merged.csv	Id, Time, HeartRate	Provides second-by-second heart rate data to observe heart rate fluctuations and peak times during the day.
HourlyCalories_merged.csv	Id, ActivityHour, Calories	Contains hourly data showing calories burned by each user at every hour of the day.
HourlyIntensities_merged.csv	Id, ActivityHour, TotalIntensity, AverageIntensity	Records intensity level of physical activity per hour. Useful to detect peak physical activity periods during the day.
HourlySteps_merged.csv	Id, ActivityHour, StepTotal	Shows number of steps taken per hour. Helps in understanding hourly movement and most active time ranges.

5. DATA CLEANING

- Removed null and duplicate values.
- **File Name: Dailyactivity_merged.csv**

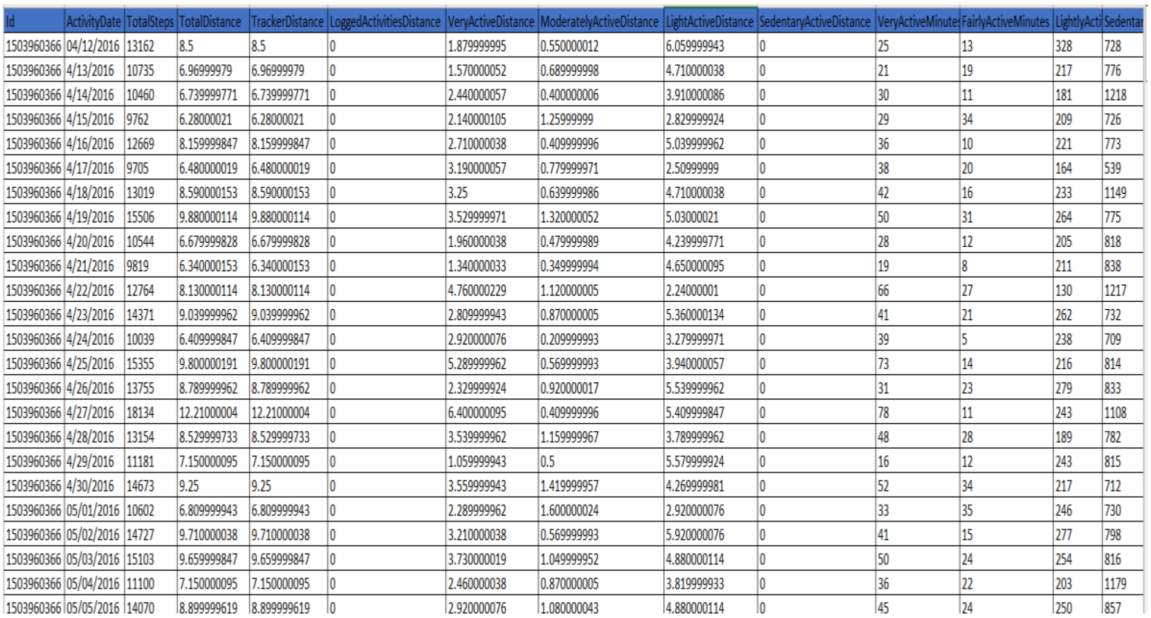
One of the first steps in my data analysis process was to clean the data by removing any missing (null) values. These are blank or empty entries in the dataset that can cause errors or give wrong results when making charts or calculations.

To do this, I used simple SQL queries like:



```
use fitness_data;
show tables;
select * from dailyactivity_merged;
DELETE FROM dailyactivity_merged
WHERE Id IS NULL
    OR ActivityDate IS NULL
    OR TotalSteps IS NULL
    OR TotalDistance IS NULL
    OR TrackerDistance IS NULL
    OR LoggedActivitiesDistance IS NULL
    OR VeryActiveDistance IS NULL
    OR ModeratelyActiveDistance IS NULL
    OR LightActiveDistance IS NULL
    OR SedentaryActiveDistance IS NULL
    OR VeryActiveMinutes IS NULL
    OR FairlyActiveMinutes IS NULL
    OR LightlyActiveMinutes IS NULL
    OR SedentaryMinutes IS NULL
    OR Calories IS NULL;
```

After running these queries, I exported the cleaned data and opened it in Excel to double-check. I've added a screenshot of the Excel sheet in the report to show that all empty values have been successfully removed. This step helped ensure the final visuals and insights are based on complete and correct data.



Id	ActivityDate	TotalSteps	TotalDistance	TrackerDistance	LoggedActivitiesDistance	VeryActiveDistance	ModeratelyActiveDistance	LightActiveDistance	SedentaryActiveDistance	VeryActiveMinutes	FairlyActiveMinutes	LightlyActi	Sedentar
1503960366	04/12/2016	13162	8.5	8.5	0	1.879999995	0.550000012	6.059999943	0	25	13	328	728
1503960366	4/13/2016	10735	6.96999979	6.96999979	0	1.570000052	0.689999998	4.710000038	0	21	19	217	776
1503960366	4/14/2016	10460	6.739999771	6.739999771	0	2.440000057	0.400000006	3.910000086	0	30	11	181	1218
1503960366	4/15/2016	9762	6.28000021	6.28000021	0	2.140000105	1.25999999	2.829999924	0	29	34	209	726
1503960366	4/16/2016	12669	8.159999847	8.159999847	0	2.710000038	0.409999996	5.039999962	0	36	10	221	773
1503960366	4/17/2016	9705	6.480000019	6.480000019	0	3.190000057	0.779999971	2.50999999	0	38	20	164	539
1503960366	4/18/2016	13019	8.590000153	8.590000153	0	3.25	0.639999986	4.710000038	0	42	16	233	1149
1503960366	4/19/2016	15506	9.880000114	9.880000114	0	3.529999971	1.320000052	5.03000021	0	50	31	264	775
1503960366	4/20/2016	10544	6.679999828	6.679999828	0	1.960000038	0.479999989	4.239999771	0	28	12	205	818
1503960366	4/21/2016	9819	6.340000153	6.340000153	0	1.340000033	0.349999994	4.650000095	0	19	8	211	838
1503960366	4/22/2016	12764	8.130000114	8.130000114	0	4.760000229	1.120000005	2.24000001	0	66	27	130	1217
1503960366	4/23/2016	14371	9.039999962	9.039999962	0	2.809999943	0.870000005	5.360000134	0	41	21	262	732
1503960366	4/24/2016	10039	6.409999847	6.409999847	0	2.920000076	0.209999993	3.279999971	0	39	5	238	709
1503960366	4/25/2016	15355	9.800000191	9.800000191	0	5.289999962	0.569999993	3.940000057	0	73	14	216	814
1503960366	4/26/2016	13755	8.789999962	8.789999962	0	2.329999924	0.920000017	5.539999962	0	31	23	279	833
1503960366	4/27/2016	18134	12.21000004	12.21000004	0	6.400000095	0.409999996	5.409999847	0	78	11	243	1108
1503960366	4/28/2016	13154	8.529999733	8.529999733	0	3.539999962	1.159999967	3.789999962	0	48	28	189	782
1503960366	4/29/2016	11181	7.150000095	7.150000095	0	1.059999943	0.5	5.579999924	0	16	12	243	815
1503960366	4/30/2016	14673	9.25	9.25	0	3.559999943	1.419999957	4.269999981	0	52	34	217	712
1503960366	05/01/2016	10602	6.809999943	6.809999943	0	2.289999962	1.600000024	2.920000076	0	33	35	246	730
1503960366	05/02/2016	14727	9.710000038	9.710000038	0	3.210000038	0.569999993	5.920000076	0	41	15	277	798
1503960366	05/03/2016	15103	9.659999847	9.659999847	0	3.730000019	1.049999952	4.880000114	0	50	24	254	816
1503960366	05/04/2016	11100	7.150000095	7.150000095	0	2.460000038	0.870000005	3.819999933	0	36	22	203	1179
1503960366	05/05/2016	14070	8.899999619	8.899999619	0	2.920000076	1.080000043	4.880000114	0	45	24	250	857

Filtered relevant data per user

For Selected User: ID – 1503960366

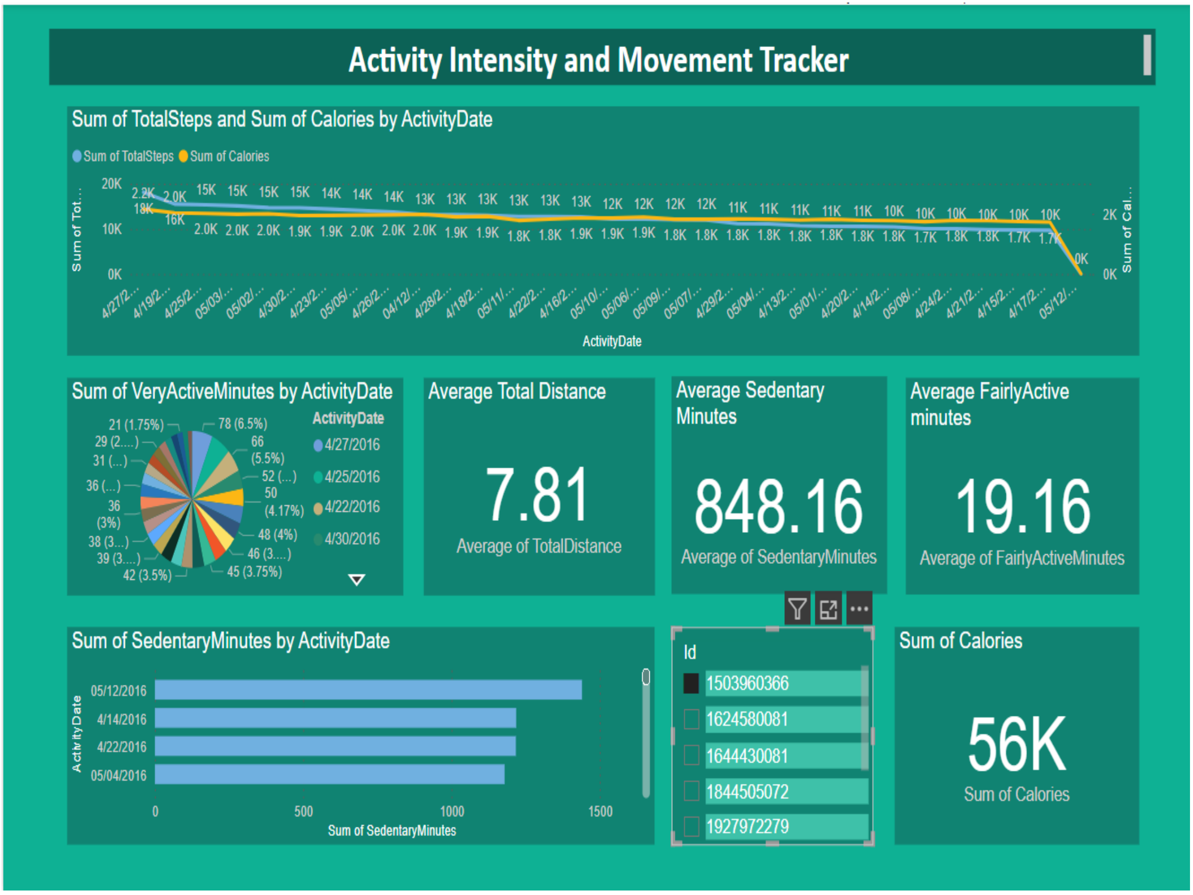
- **Total Steps vs Calories Burned (Line Chart):**
  This user shows a consistent daily step count ranging between 10,000 to 16,000 steps, with a corresponding calorie burn between 1,700 to 2,200 calories. This reflects a highly active lifestyle.

- **Very Active Minutes (Bar Chart):**
  The user peaked with 78 very active minutes on April 27, 2016, indicating high-intensity activity sessions. Other days like April 25 and April 30 also show significant active durations.

- **Average Total Distance (Card):**
  The average distance covered by this user is 7.81 kilometres per day, supporting the trend of consistent physical movement.

- **Sedentary Time (Card):**
  Despite being active, the user also logs an average of 848.16 sedentary minutes per day, which is nearly 14 hours. This indicates long sitting periods—perhaps due to work or screen time.

- **Fairly Active Minutes (Card):**
  The user spends an average of 19.16 minutes in moderately active states, such as walking or slow-paced exercise, reflecting balanced physical activity.

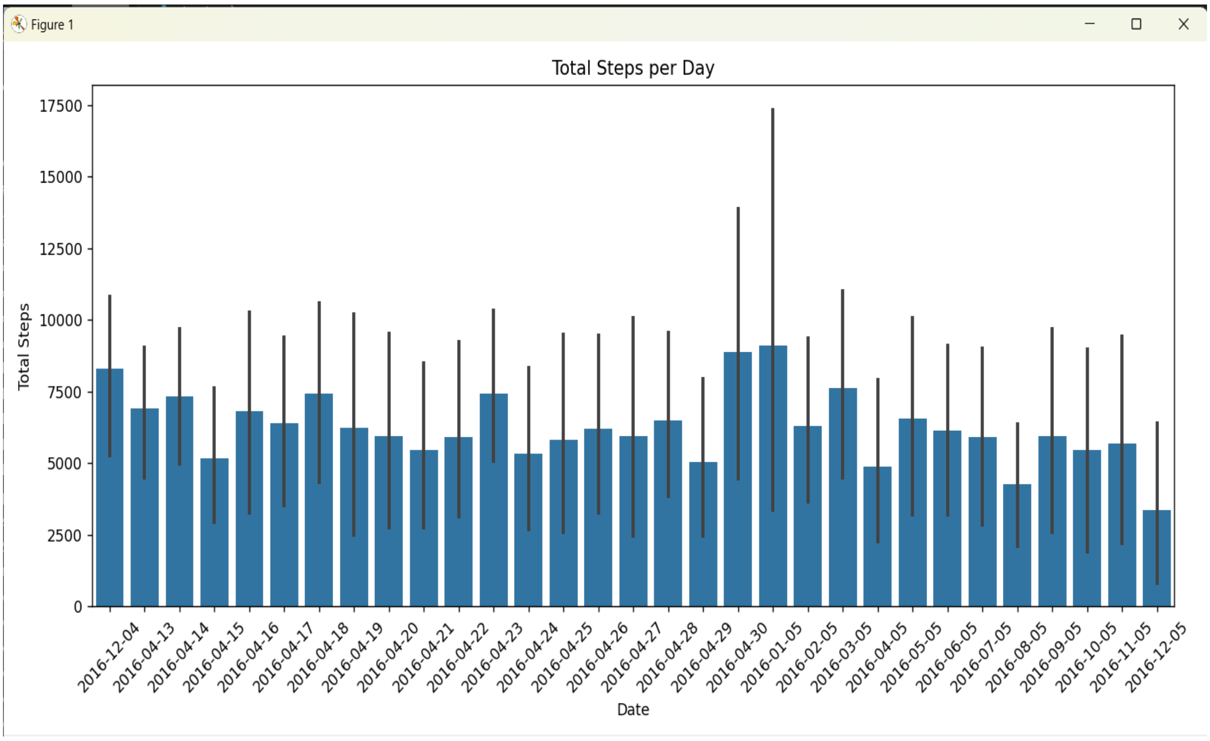


Key Insights:

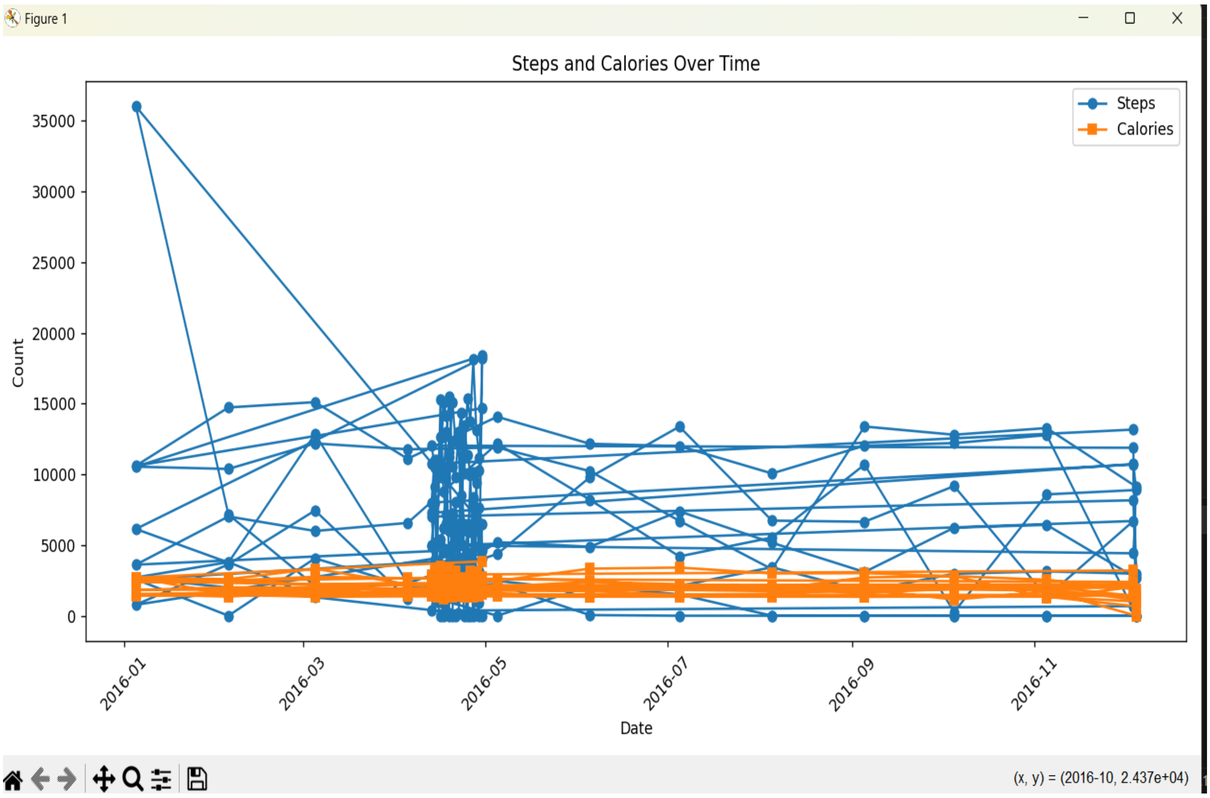
- **Highest activity** was observed on **April 27, 2016**, with over **18,000 steps** and high calorie burn.
- **Total calorie burn** across the selected period is approximately **56,000 calories**.
- **Average total distance** walked is **7.81 km**, indicating consistent movement on active days.
- **Average sedentary minutes** is **848.16 mins/day**, showing users remain inactive for most of the day.
- **Average fairly active minutes** is low at **19.16 mins/day**, suggesting limited moderate physical activity.
- **Very active minutes** peak on specific days (e.g., April 25–27), but are not consistent across all dates.
- **High sedentary days** include **May 12** and **April 14**, with sedentary time exceeding **1,000 minutes**.

Python Visualizations Using Activity Data

To further explore and understand the fitness behavior of users, I utilized Python libraries—**Pandas**, **Matplotlib**, and **Seaborn**—to create meaningful visualizations from the **activity dataset**. The data was first processed using Pandas for cleaning, filtering, and extracting insights like total steps, calories, and distance. Then, with Matplotlib and Seaborn, I visualized patterns such as **daily step counts, calorie burn trends**, and **hourly activity levels** using line plots, bar charts, and heatmaps. These visuals helped highlight trends like **peak activity hours**, **user consistency**, and **correlations between distance and calories burned**—providing valuable context before building dashboards in Power BI.



A. Bar Chart : Total Steps per day



B. Line Chart: Steps and Calories Over Time

2. Dataset Analysis: Hourly Calories Data

2.1. Dataset Description:

- File Name: hourlyCalories_merged.csv
- Key Columns:
    - Id: Unique user identifier.
    - ActivityHour: Timestamp of recorded activity (hourly).
    - Calories: Number of calories burned during that hour.

2.2. Data Preparation

- Converted ActivityHour to date and time format.
- Separated into **Date** and **Hour** columns.
- Cleaned data by removing any null/duplicate rows.
- Imported into **Power BI** and **Python (pandas)** for analysis.

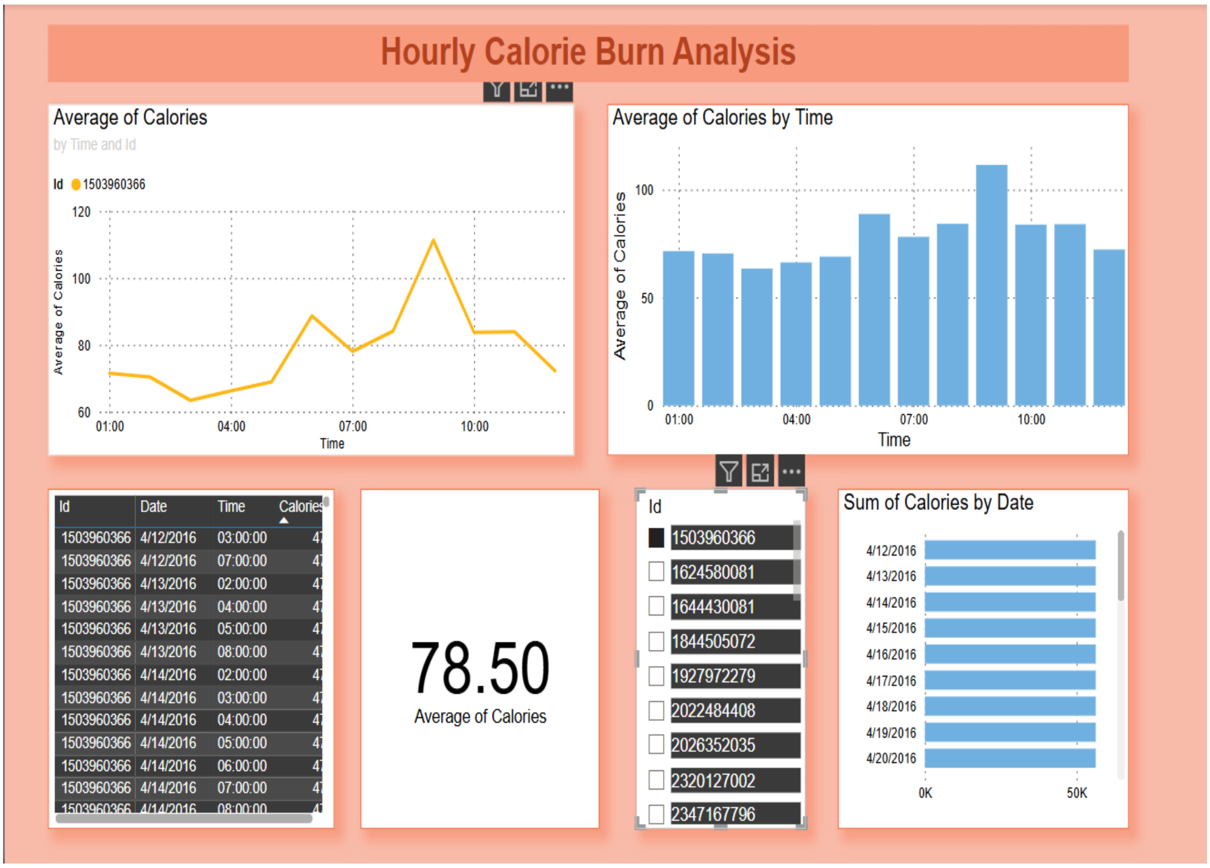
2.3. Power BI Visuals Created

For Selected User: ID – 1503960366

In this dashboard, I selected the user with ID 1503960366 to analyze their hourly calorie burn pattern. From the visuals:

- The average calories burned per hour for this user is 78.5.
- The highest calorie burn occurred around 09:00 AM, showing increased activity during morning hours.
- The line and bar charts clearly show variations in energy expenditure across different times of day.
- The table provides a detailed hourly record of calories burned on each day for this user.
- The sum by date bar chart shows consistently high total calories, indicating regular activity levels.

This analysis helps us understand when the user is most active and how their calorie burn is distributed throughout the day.



Hourly Calorie Burn Analysis

Id	Date	Time	Calories
1503960366	4/12/2016	03:00:00	47
1503960366	4/12/2016	07:00:00	47
1503960366	4/13/2016	02:00:00	47
1503960366	4/13/2016	04:00:00	47
1503960366	4/13/2016	05:00:00	47
1503960366	4/13/2016	08:00:00	47
1503960366	4/14/2016	02:00:00	47
1503960366	4/14/2016	03:00:00	47
1503960366	4/14/2016	04:00:00	47
1503960366	4/14/2016	05:00:00	47
1503960366	4/14/2016	06:00:00	47
1503960366	4/14/2016	07:00:00	47
1503960366	4/14/2016	08:00:00	47

78.50
Average of Calories

Id

- ■ 1503960366
- □ 1624580081
- □ 1644430081
- □ 1844505072
- □ 1927972279
- □ 2022484408
- □ 2026352035
- □ 2320127002
- □ 2347167796

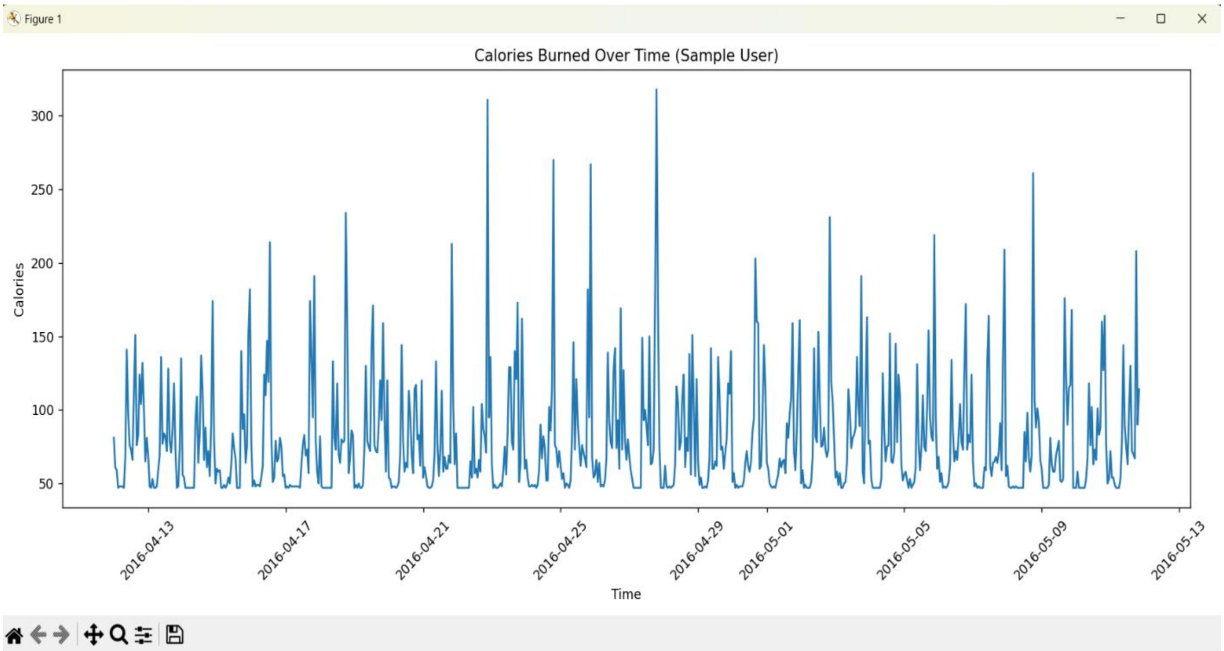
Key Insights:

- **Peak calorie burn** occurs between **12 PM – 6 PM**.
- Lowest burn happens **early morning (1 AM – 4 AM)**.
- **Average calories/hour**: ~78.5.
- Daily total calories burned by users: ~50K–60K.
- Some users (e.g., 1503960366) show consistent activity, while others have gaps or lower burn rates.
- Patterns suggest steady daily activity with peak hours likely linked to workouts or high movement.

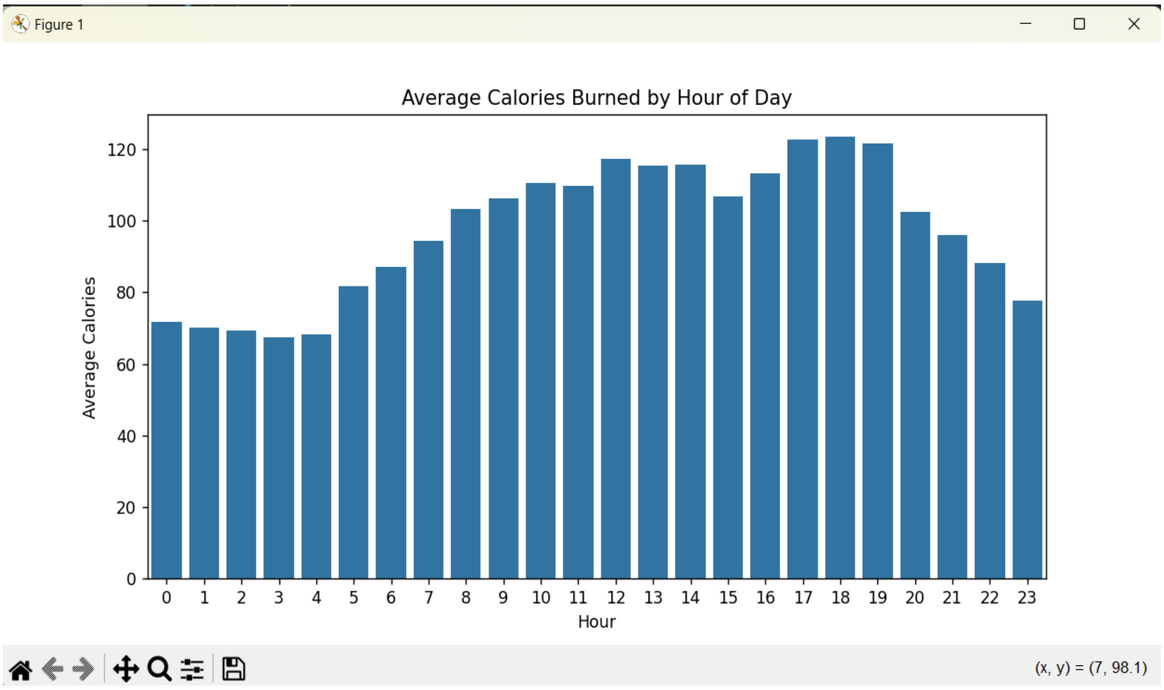
2.4. Python Visualization – Hourly Calorie Burn

To analyze the hourlyCalories_merged.csv dataset, I used **Python** along with key data analytics libraries:

- **Pandas** was used to load and preprocess the data (e.g., parsing ActivityHour into Date and Time).
- **Matplotlib** and **Seaborn** helped in creating insightful visualizations such as:
  - **Line Plot** showing average calorie burn across hours of the day.
  - **Bar Chart** displaying total calories burned per date.
- **Heatmap** to visualize hourly calorie patterns for selected user.



Line Chart : Calories Over Time



Bar Chart (Vertical): Average Calories by Hour of Day

3.DATASET ANALYSIS: HOURLY PHYSICAL ACTIVITY INTENSITY DATA

3.1. Dataset Description:

- File Name: hourlyIntensities_merged.csv
- Key Columns:
  - Id: Unique user identifier.
  - ActivityHour: Timestamp of recorded activity (hourly).
  - TotalIntensity : Sum of intensity level for that hour (higher value = more physical effort).
  - AverageIntensity: Average intensity level for that hour (TotalIntensity divided by 60 mins)

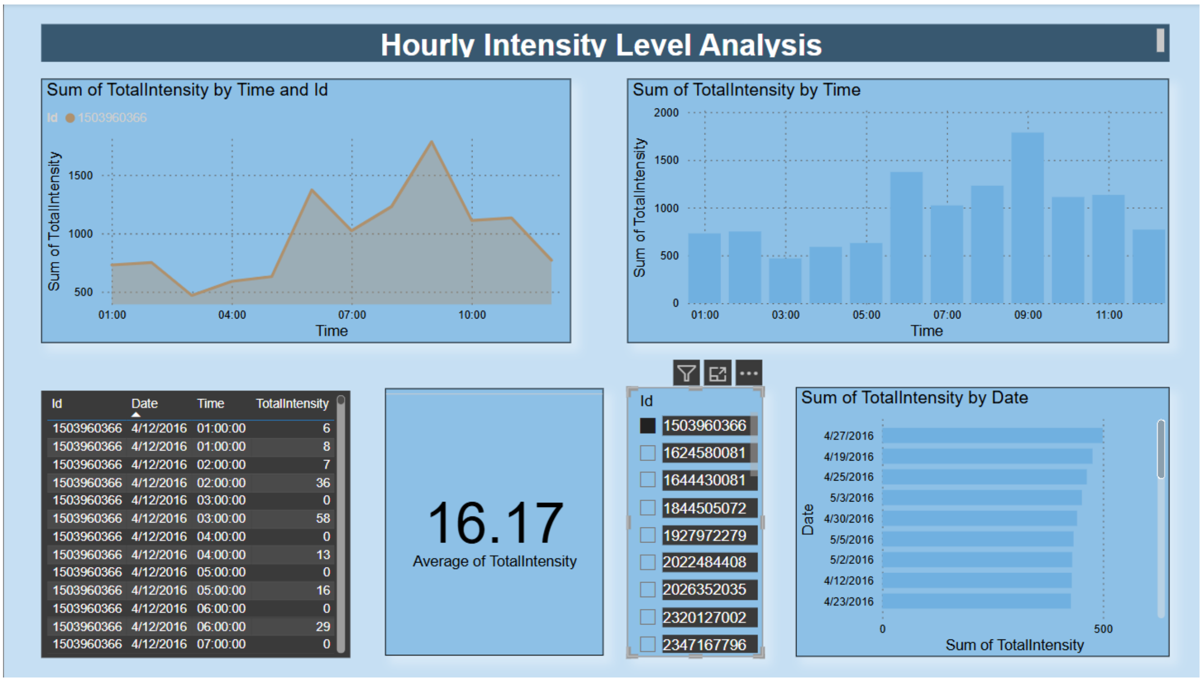
3.2. Data Preparation:

- Converted ActivityHour to date and time format.
- Separated into **Date** and **Hour** columns.
- Cleaned data by removing any null/duplicate rows.
- Imported into **Power BI** and **Python (pandas)** for analysis.

3.3. Power BI Visuals Created

For Selected User: ID – 1503960366

- **Peak activity** occurs at **9:00 AM**, where intensity exceeds **1800**.
- Early morning hours (1:00 AM – 4:00 AM) show **lower activity**, typically below **700**.
- Sudden spike observed at **6:00 AM – 9:00 AM**, indicating user becomes active mid-morning.
- **Average Total Intensity** for this user is **16.17** (shown in KPI card).
- Data table confirms fluctuations hour-wise on **4/12/2016**, with varying intensity levels per hour.
- time format.
- Separated into **Date** and **Hour** columns.
- Cleaned data by removing any null/duplicate rows.
- Imported into **Power BI** and **Python (pandas)** for analysis.



Id	Date	Time	TotalIntensity
1503960366	4/12/2016	01:00:00	6
1503960366	4/12/2016	01:00:00	8
1503960366	4/12/2016	02:00:00	7
1503960366	4/12/2016	02:00:00	36
1503960366	4/12/2016	03:00:00	0
1503960366	4/12/2016	03:00:00	58
1503960366	4/12/2016	04:00:00	0
1503960366	4/12/2016	04:00:00	13
1503960366	4/12/2016	05:00:00	0
1503960366	4/12/2016	05:00:00	16
1503960366	4/12/2016	06:00:00	0
1503960366	4/12/2016	06:00:00	29
1503960366	4/12/2016	07:00:00	0

Id
■ 1503960366
□ 1624580081
□ 1644430081
□ 1844505072
□ 1927972279
□ 2022484408
□ 2026352035
□ 2320127002
□ 2347167796

Key Insights:

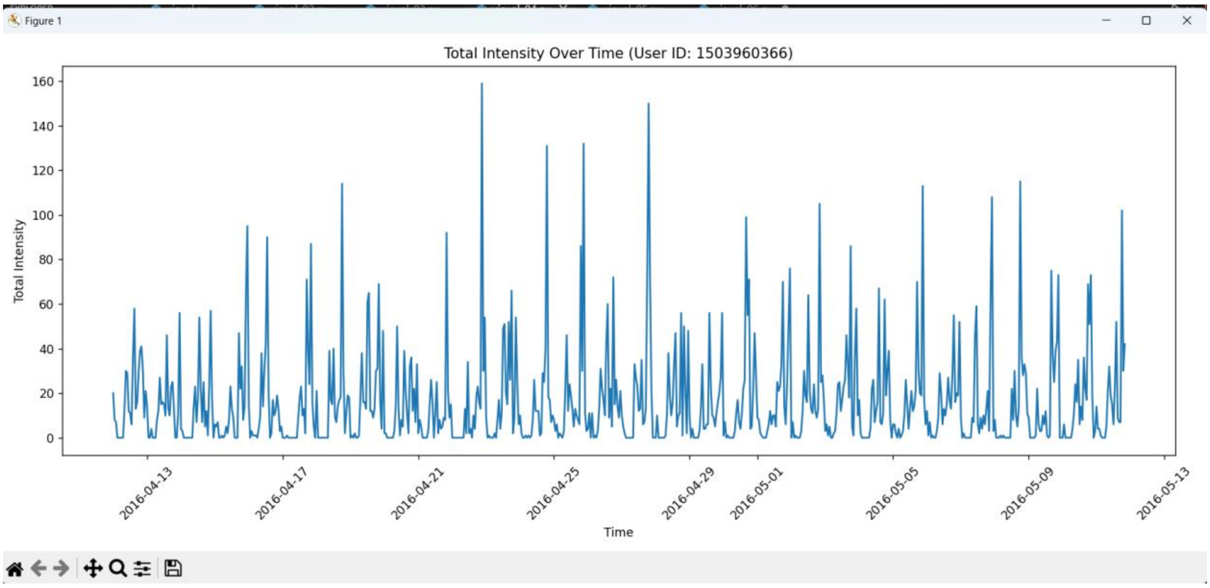
- Peak Intensity Time:
  - Most users show highest physical intensity around 9:00 AM.
  - Moderate activity is observed between 6:00 AM to 11:00 AM.
- Total Intensity Trend:
  - The Total Intensity rises after early morning hours, peaking mid-morning.
  - Lower activity is visible during late night and early morning (1:00 AM – 5:00 AM).
- Most Active Dates:
  - Dates like 4/23/2016 and 4/12/2016 had the highest intensity recorded.
  - Activity is spread over multiple days, indicating consistent tracking.
- User Participation:
  - All users contribute to the activity levels, but user ID 1503960366 is consistently more active based on the visible data.
- Average Intensity:
  - The overall average TotalIntensity across all users and hours is ~16.17.
  - Suggests users have light to moderate hourly activity on average**.**

3.4.Python Visualization: Hourly Intensities Data

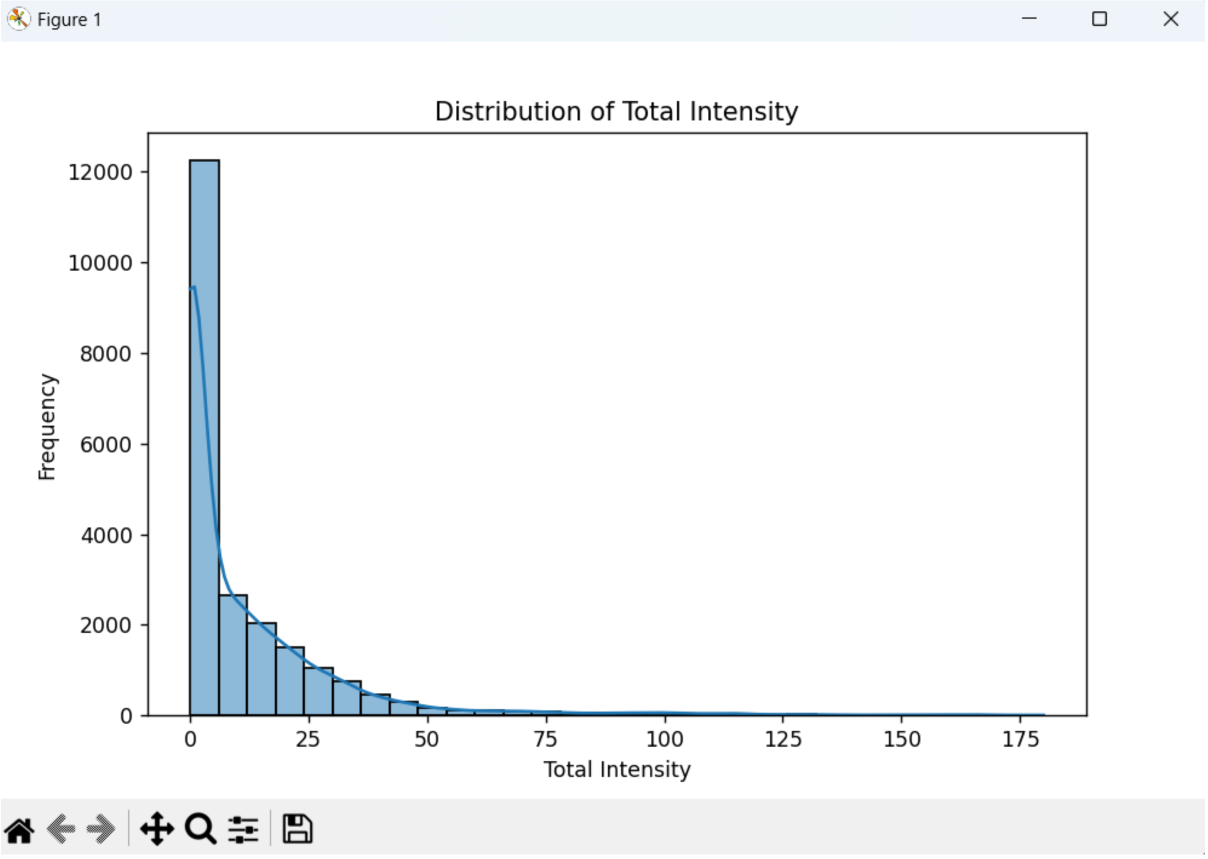
Using Python libraries such as Pandas, Matplotlib, and Seaborn, I analyzed and visualized the hourlyIntensities_merged.csv dataset to understand physical activity patterns.

- Histogram: Distribution of Total Intensity
- Box Plot: Total Intensity by Hour of Day

These visuals helped identify peak intensity periods and differences in user activity.



Histogram: Distribution of Total Intensity



Box Plot: Total Intensity by Hour of Day

4.DATASET ANALYSIS: HOURLY STEP COUNT DATA

4.1. Dataset Description:

- File Name: hourlysteps_merged.csv
- Key Columns:
  - Id: Unique user identifier.
  - ActivityHour: Timestamp of recorded activity (hourly).
  - ActivityHour: Total number of steps in that hour.

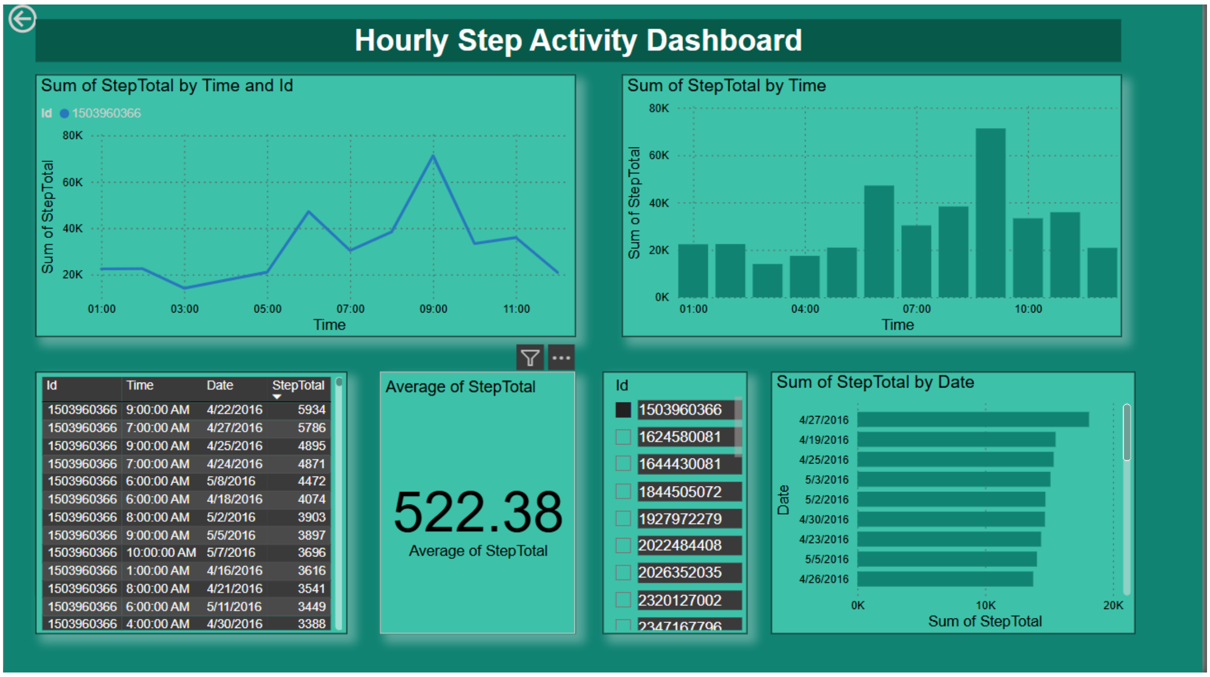
4.2. Data Preparation:

- Converted ActivityHour to date and time format.
- Separated into **Date** and **Hour** columns.
- Cleaned data by removing any null/duplicate rows.
- Imported into **Power BI** and **Python (pandas)** for analysis.

4.3. Power BI Visuals Created:

For Selected User: ID – 1503960366

- Line Chart – Sum of StepTotal by Time and Id
  - Visualizes how step count varies by hour for the selected user.
  - Shows peak walking periods (like 9:00 AM).
- Bar Chart – Sum of StepTotal by Time
  - Aggregated view of step totals by hour for easier comparison.
  - Helps identify active hours across dates.
- Bar Chart – Sum of StepTotal by Date
  - Highlights which dates had the highest movement (e.g., 4/27/2016).
- Card Visual – Average StepTotal
  - Displays average steps taken per hour for the selected user.
  - Example: 522.38 steps/hour for User ID 1503960366.
- Table Visual – Detailed View
  - Shows step count for each hour, along with the corresponding date and user ID.
- Slicer – User ID Selection
  - Allows filtering visuals by specific user IDs.



Id	Time	Date	StepTotal
1503960366	9:00:00 AM	4/22/2016	5934
1503960366	7:00:00 AM	4/27/2016	5786
1503960366	9:00:00 AM	4/25/2016	4895
1503960366	7:00:00 AM	4/24/2016	4871
1503960366	6:00:00 AM	5/8/2016	4472
1503960366	6:00:00 AM	4/18/2016	4074
1503960366	8:00:00 AM	5/2/2016	3903
1503960366	9:00:00 AM	5/5/2016	3897
1503960366	10:00:00 AM	5/7/2016	3696
1503960366	1:00:00 AM	4/16/2016	3616
1503960366	8:00:00 AM	4/21/2016	3541
1503960366	6:00:00 AM	5/11/2016	3449
1503960366	4:00:00 AM	4/30/2016	3388

Average of StepTotal

522.38

Average of StepTotal

Id

■ 1503960366
□ 1624580081
□ 1644430081
□ 1844505072
□ 1927972279
□ 2022484408
□ 2026352035
□ 2320127002
□ 2347167796

Key Insights (All Users – Hourly Steps Data):

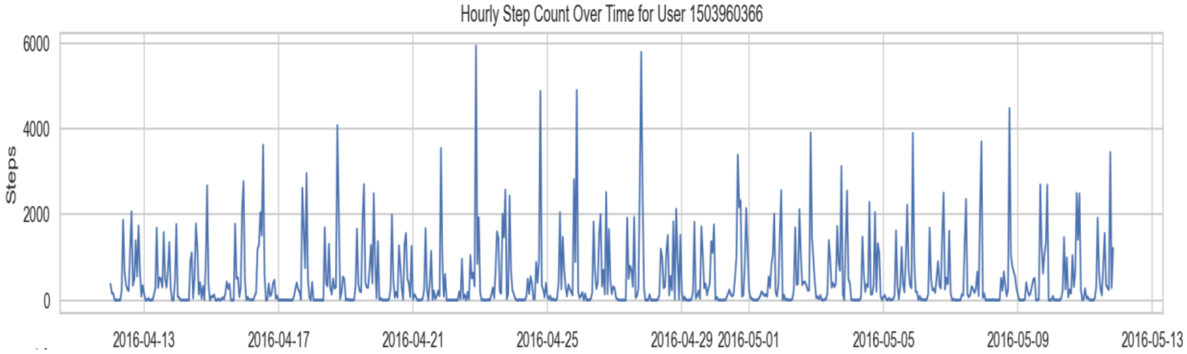
- **Peak Activity Hours:** Most users show highest step activity between **6:00 AM and 10:00 AM**, indicating a strong morning routine across users.
- **Low Activity Periods:** Activity drops significantly **after 11:00 AM**, suggesting lower movement in the afternoon and evening.
- **Most Active Dates:** Days like **April 27, May 1, and April 25, 2016** recorded the **highest total step counts**, possibly linked to special events, workouts, or weekends.
- **Average Step Count:** The overall **hourly average step count** remains moderate across users, indicating varied engagement levels.
- **User Variance:** Some users (e.g., ID: 1503960366) show highly active patterns, while others have significantly lower movement, revealing diverse fitness behaviors.

4.4.Python Visualization – Hourly Steps Data

Using Python libraries such as **Pandas**, **Matplotlib**, and **Seaborn**, I analyzed the hourlySteps_merged.csv dataset to understand patterns in users' step activity over time.

- **Line Chart**: Hourly Step Count Over Time for User 1503960366
- **Description:**
- This line chart displays the number of steps taken **per hour over multiple days** by a single user. It helps visualize fluctuations in physical activity across different hours and days. Key insights from the visual:

- **Spikes in activity** reflect peak movement hours, often in the **morning and evening**.
- **Frequent periods of low or zero steps** likely indicate rest or sedentary periods.
- The graph demonstrates that the user's **activity pattern is inconsistent**, suggesting varying daily routines or external influences on step count.



CONCLUSION

This fitness data analytics project focused on analyzing user behavior based on physical activity, sleep patterns, heart rate, and calorie consumption. By working with datasets like daily activity, heart rate, hourly steps, intensity, and calories, meaningful insights were derived to understand trends and performance.

SQL was used to clean the data by handling null values, duplicates, and formatting inconsistencies. Power BI helped create interactive dashboards using charts, cards, and slicers to visualize user habits over time. Python libraries such as Pandas, Matplotlib, and Seaborn supported in-depth visual analysis, highlighting relationships and variations within the data.

The project provided a clear view of user fitness patterns and emphasized the importance of consistent activity and sleep. It also showcased how combining tools like SQL, Power BI, and Python can turn raw data into actionable insights through clear, visual storytelling.