# CS210 TERM PROJECT REPORT

# An Analysis on My Health Data

Prepared by Selin Imren- 31194

## Motivation

The goal of this project is to analyze health data, specifically daily steps and energy expenditure, to derive insights into the relationship between physical activity and energy burn. The dataset used contains information about daily steps, basal energy burned, and active energy burned. We aim to explore patterns, correlations, and trends within this data. When I examined my health data, I realized that my physical activity was limited to walking. Because when I looked at my other activity data, I came across very low numbers. That's why I decided to compare my step counts with my daily energy expenditure and examine the correlation between them. That's why I determined my hypothesis that my daily energy burn is related to my step count. If the two parameters I stated in my hypothesis are highly dependent, this will show that I do not do any other sports and my physical activity is limited to my step count.

There is a significant positive correlation between the number of daily steps taken and the daily energy expenditure, suggesting that an increase in daily steps is associated with a higher daily energy burn.

## Data Source

I received my health data from the apple health application. Then I started reading my health data from the health_data csv file. I scrapped the data which is needed from the file by looking the type names. These are the number of steps, the basal energy burned and the active energy burned ,recorded over a specific time period. The 'StartDate' and 'EndDate' columns were transformed into datetime format to facilitate temporal analysis. Subsequently, the data was filtered to focus on the year 2023, ensuring relevance to the research objectives.

## Data Analysis:

To facilitate a more comprehensive analysis, feature engineering was conducted. The 'Total_energy' column was engineered to represent the sum of basal and active energy, providing a holistic measure of daily energy expenditure. Additionally, a 'Value_steps' column was created to capture the daily steps metric. These engineered features served as key variables for subsequent analyses. Then I find the correlation between the total energy burned in a day and the number of steps in a day, then I received a high correlation between the variables so my hypothesis cannot be rejected.

```
Correlation between daily steps and active energy burned: 0.9713620805454196
```

. To further comprehension, I create 7-rolling averages of daily number of steps and burned total energy and the control the correlation between them because I would like to analyze the data for longer period than a day. It will give more reliable correlation. As a result of the second correlation, I also got a high correlation between them. Further, I find the Pearson correlation it is very close to 1 and p-value is low so we can understand the high correlation between variables.

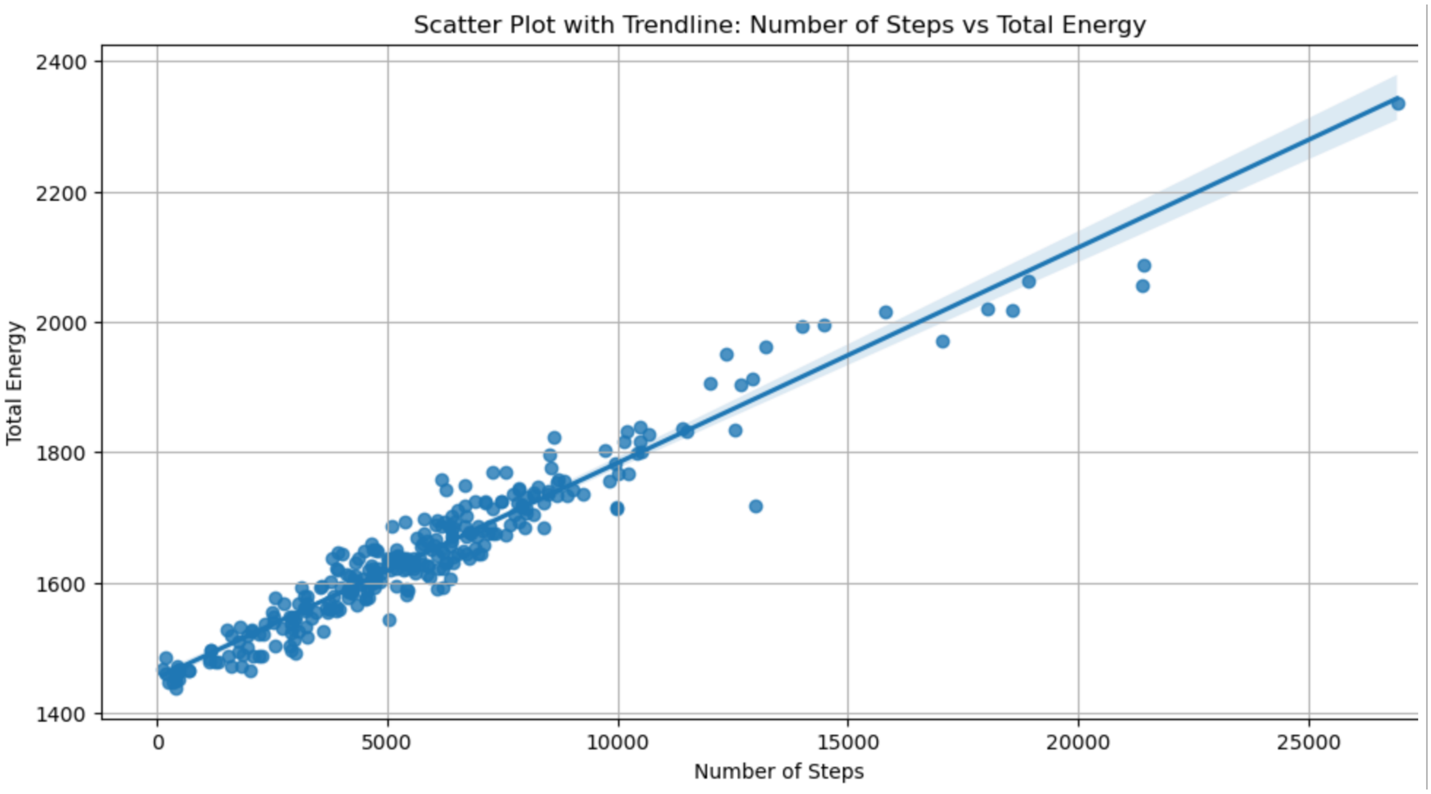
```
Correlation with Daily Number of Steps and Burned Total Energy: 0.9713620805454196

Correlation with 7-day Rolling Number of Steps and 7-day Rolling Burned Total Energy:
0.9743846249655004
```

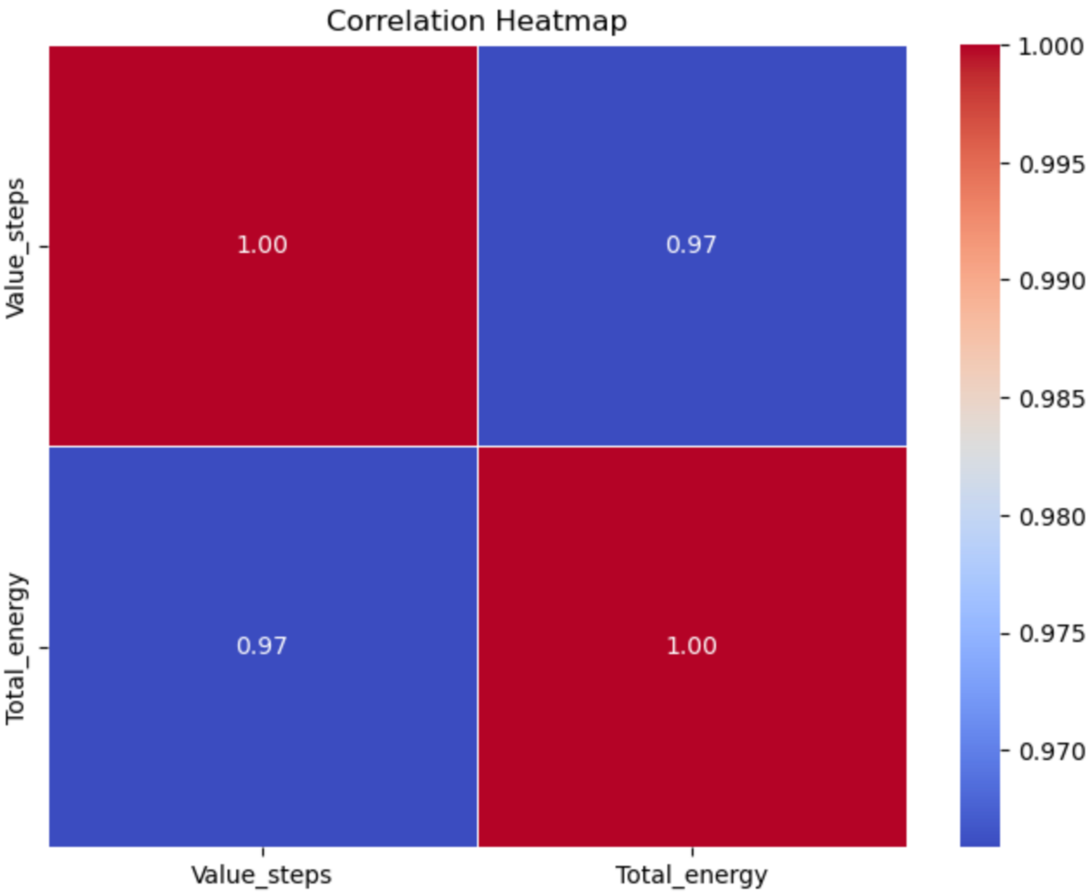
```
Pearson Correlation Coefficient: 0.9713620805454198

P-value: 8.535375989025902e-187
```

To visualize this correlation , I draw a scatter plot chart with a trendline:

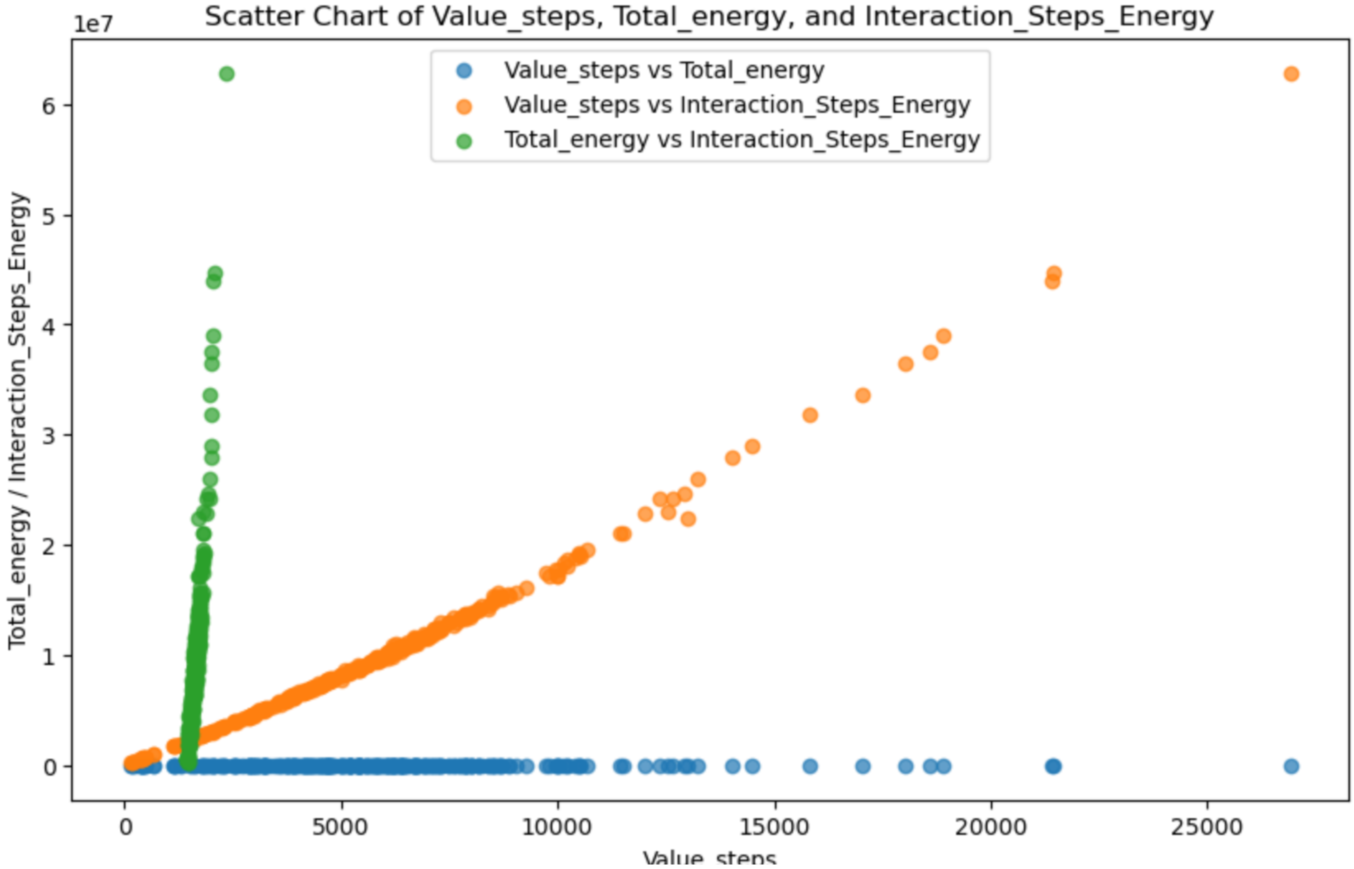


Here is the heat map which shows the correlation:

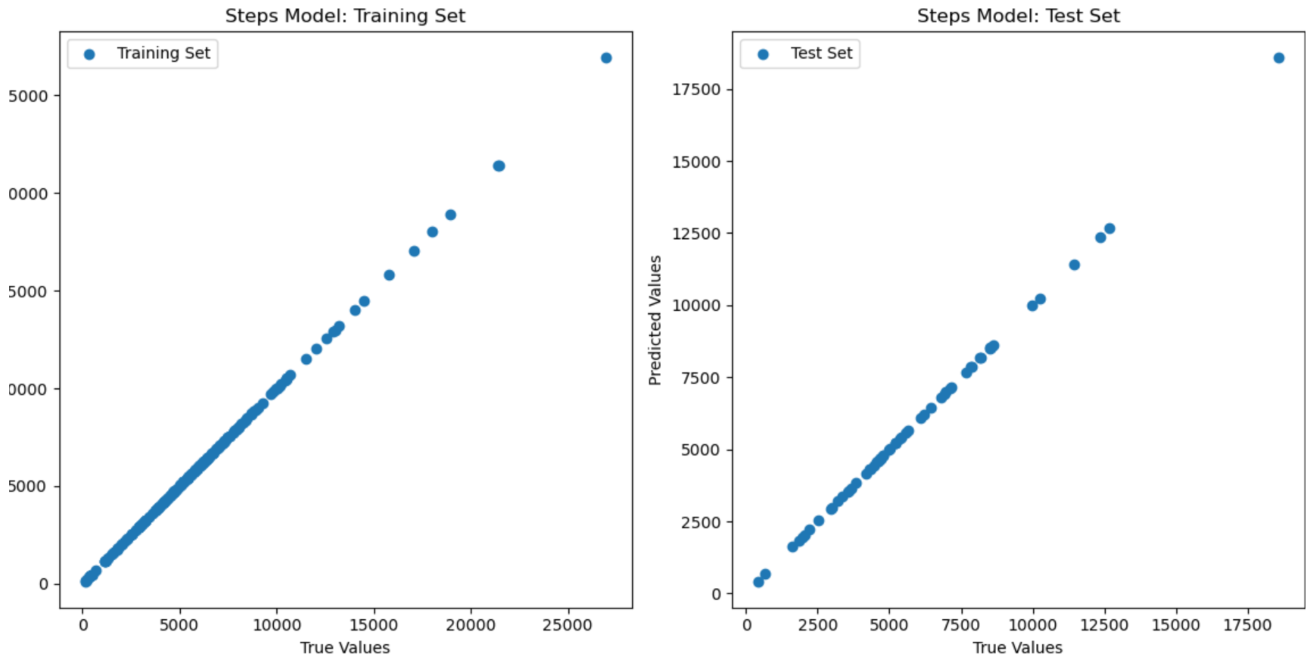


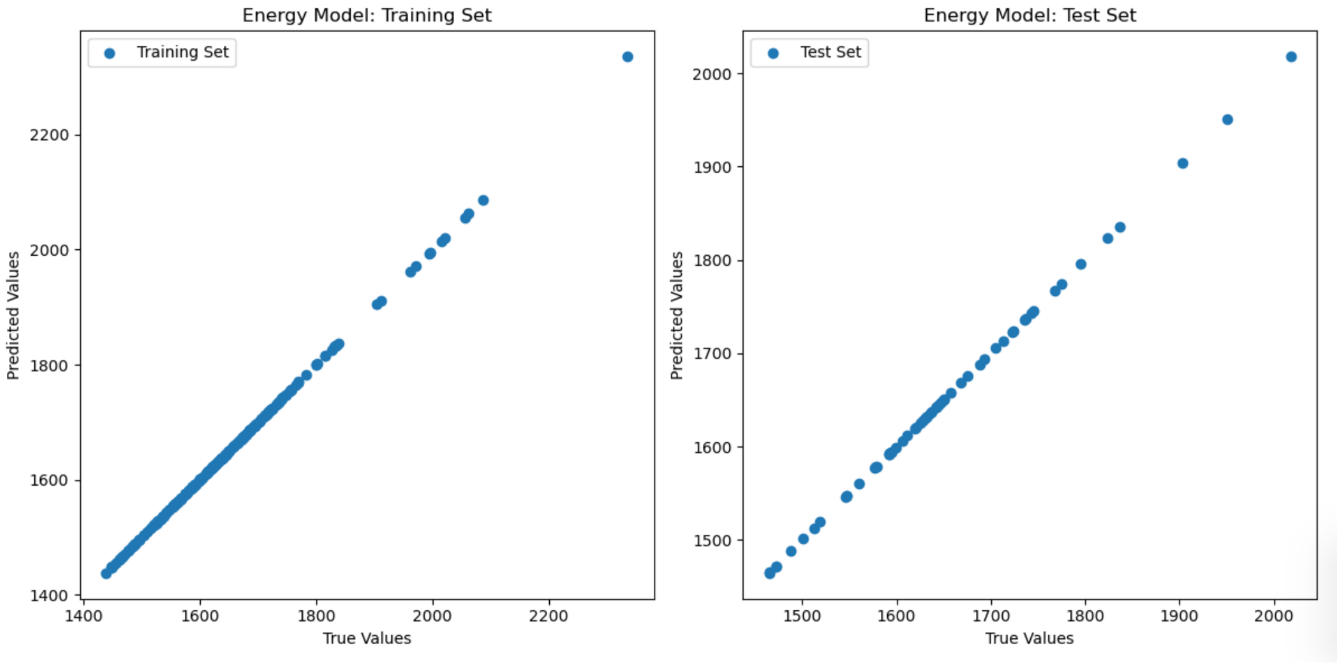
I created a new feature a the 'Interaction_Steps_Energy' is the result of multiplication, a higher value could suggest that the impact of steps on energy is magnified, indicating a potential nonlinear relationship. It explores potential synergies or dependencies between daily steps and overall energy burn, shedding light on whether certain levels of physical activity influence the overall energy expenditure differently.  I made some analysis for the new feature.

A chart is plotted for visualize the our features and their correlation:



I try to train separate linear regression models for predicting steps and energy expenditure and also evaluate model performance using metrics such as MSE, RMSE, and R-squared. I visualize the predictions and actual datas:





# FINDINGS

As a result of my project, the fact that my step count was so closely related to my total energy burn showed me that my physical activity was limited only to my daily walking. In my opinion, this result does not indicate a very healthy life. I think that living a physically inactive life may cause diseases in the long run. Thanks to this project, I realized an important fact for me to live a healthier life. I should start doing sports to increase my physical activities. If I do sports regularly, not just walking, I can see the effects of regular sports in my health data after a significant period of time. That's why, maybe if I can manage to exercise regularly within a year, I would like to do another project to see its differences from today's project.

# LIMITATIONS and FUTURE WORK

As I mentioned in my findings I would like to conduct another project with my new health data from a more active physical activity, because I can see the differences more clearly when I have another data it can be comparable with the this project.

I had a hard time capturing my health data because I was inexperienced. At the same time, I realized that some of my data was not stored correctly, and I could not access it. Now, I have given limited permissions to applications that I did not allow, in order to access my own data later. If I had more data, I could prepare a more comprehensive project. I prepared such a project only with the data I had, and the output of the project was noteworthy for me. If I were more knowledgeable in Python, I could use it more appropriately for my purpose.