Sabancı University
DSA210 – Introduction to Data Science
Term Project Report
Zeynep Su Yaşar – 30694

1. Introduction

This is the report for my project of the course, DSA210 - Introduction to Data Science. The codes and the data files can be found in the GitHub page.

https://github.com/zeynepsuyasar/DSA210_Term_Project

2. Motivation

This project investigates how various stages of the menstrual cycle affect physical activity levels. The main aim is to identify trends and relationships by using data science methods, which can help us gain insights into personal health and optimize activity scheduling. By exploring this link, the project highlights the significance of tailored health approaches and improved health monitoring through predictive analytics. Moreover, it acts as a hands-on application of data science skills in a real-world scenario.

3. Objectives

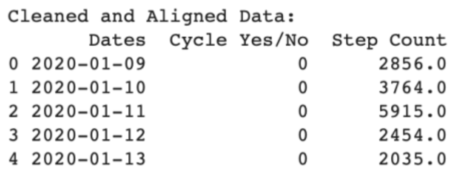
- Utilize data science methodologies on practical health data.
- Examine the correlation between phases of the menstrual cycle and daily step counts.
- Perform exploratory data analysis (EDA) to uncover patterns.
- Develop machine learning models to forecast menstrual cycle days based on step count information.

4. Data Collection

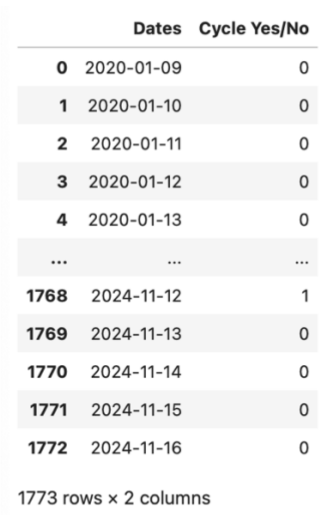
- Menstrual Cycle Data: Daily records from Apple's Health Application, featuring a binary indicator (Cycle Yes/No) for cycle days.
- Step Count Data: Daily step totals from Apple's Health Application, recorded from 2020 to the present.
- The data was organized to fix any missing or repeated entries.
- The "Dates" column was aligned to combine the datasets properly.

5. Data Analysis

- Descriptive statistics, such as averages and ranges, were calculated.

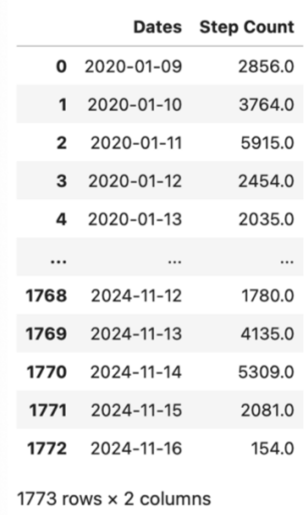


```
Cleaned and Aligned Data:
        Dates  Cycle Yes/No  Step Count
0 2020-01-09             0      2856.0
1 2020-01-10             0      3764.0
2 2020-01-11             0      5915.0
3 2020-01-12             0      2454.0
4 2020-01-13             0      2035.0
```



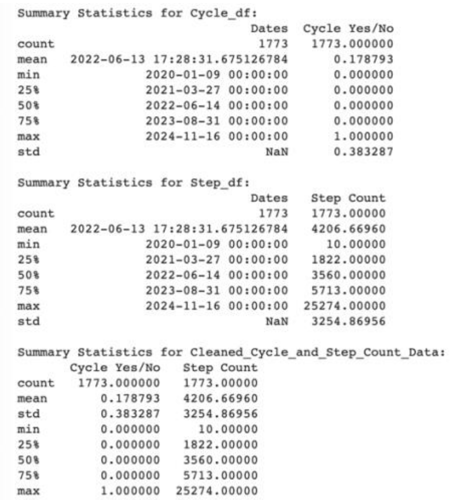
	Dates	Cycle Yes/No
0	2020-01-09	0
1	2020-01-10	0
2	2020-01-11	0
3	2020-01-12	0
4	2020-01-13	0
...	...	...
1768	2024-11-12	1
1769	2024-11-13	0
1770	2024-11-14	0
1771	2024-11-15	0
1772	2024-11-16	0

1773 rows × 2 columns



	Dates	Step Count
0	2020-01-09	2856.0
1	2020-01-10	3764.0
2	2020-01-11	5915.0
3	2020-01-12	2454.0
4	2020-01-13	2035.0
...	...	...
1768	2024-11-12	1780.0
1769	2024-11-13	4135.0
1770	2024-11-14	5309.0
1771	2024-11-15	2081.0
1772	2024-11-16	154.0

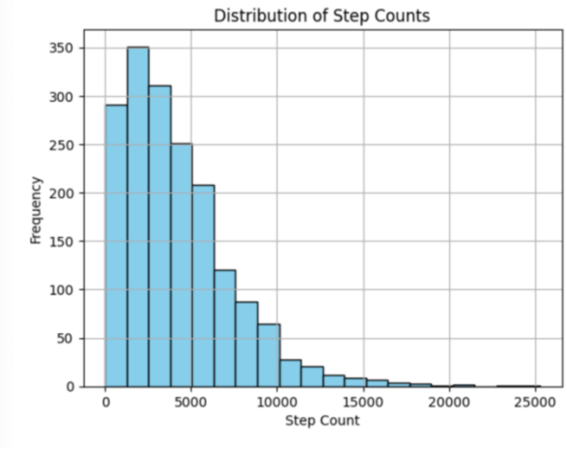
1773 rows × 2 columns



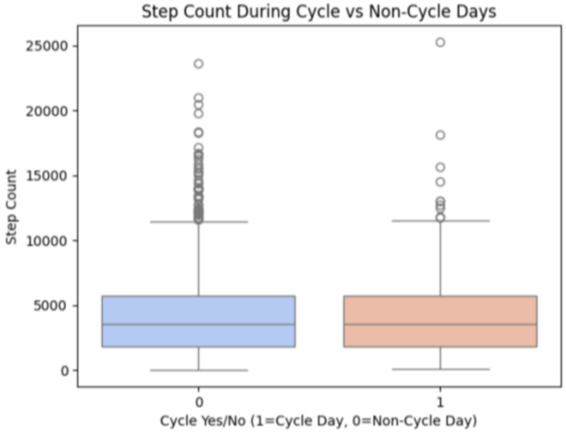
```
Summary Statistics for Cycle_df:
                               Dates  Cycle Yes/No
count                           1773   1773.000000
mean   2022-06-13 17:28:31.675126784      0.178793
min              2020-01-09 00:00:00      0.000000
25%              2021-03-27 00:00:00      0.000000
50%              2022-06-14 00:00:00      0.000000
75%              2023-08-31 00:00:00      0.000000
max              2024-11-16 00:00:00      1.000000
std                              NaN      0.383287

Summary Statistics for Step_df:
                               Dates   Step Count
count                           1773   1773.00000
mean   2022-06-13 17:28:31.675126784   4206.66960
min              2020-01-09 00:00:00     10.00000
25%              2021-03-27 00:00:00   1822.00000
50%              2022-06-14 00:00:00   3560.00000
75%              2023-08-31 00:00:00   5713.00000
max              2024-11-16 00:00:00  25274.00000
std                              NaN   3254.86956

Summary Statistics for Cleaned_Cycle_and_Step_Count_Data:
       Cycle Yes/No   Step Count
count   1773.000000   1773.00000
mean       0.178793   4206.66960
std        0.383287   3254.86956
min        0.000000     10.00000
25%        0.000000   1822.00000
50%        0.000000   3560.00000
75%        0.000000   5713.00000
max        1.000000  25274.00000
```

- Visualization

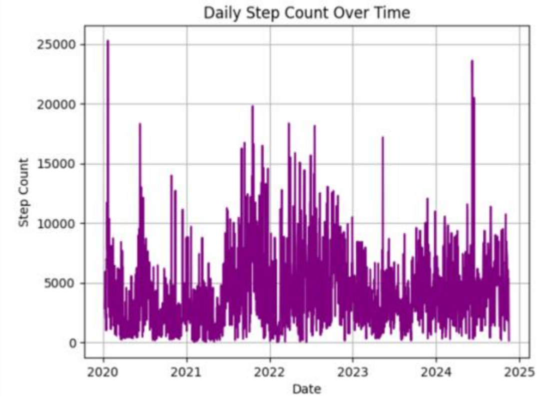
A histogram displaying my daily step counts helps to reveal patterns in my activity levels. The majority of my days are between 0 and 5000 steps, which suggests I have a moderate level of activity, while only a few days go beyond 10,000 steps:



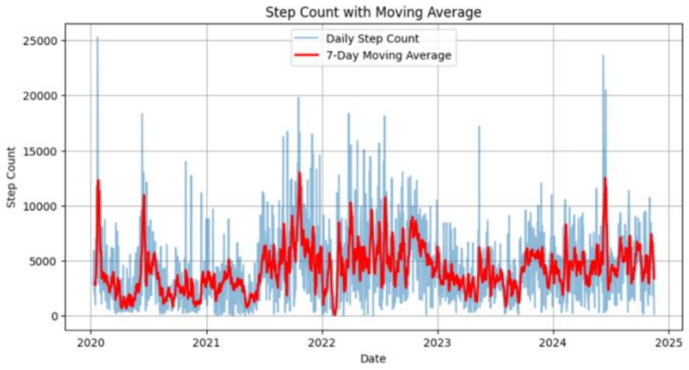
I created a box plot to look at my step counts on days when I cycle versus days when I don’t. It shows that my step counts are usually lower on cycling days, and the numbers are more consistent. In contrast, on non-cycling days, my step counts vary a lot more:



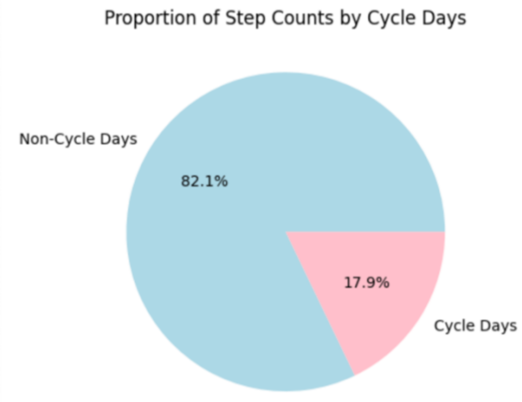
I created a line plot to monitor my daily step counts and see how my activity levels change over time. It shows that my activity varies a lot, with some days where I hit over 20,000 steps, while on other days, the numbers are significantly lower:



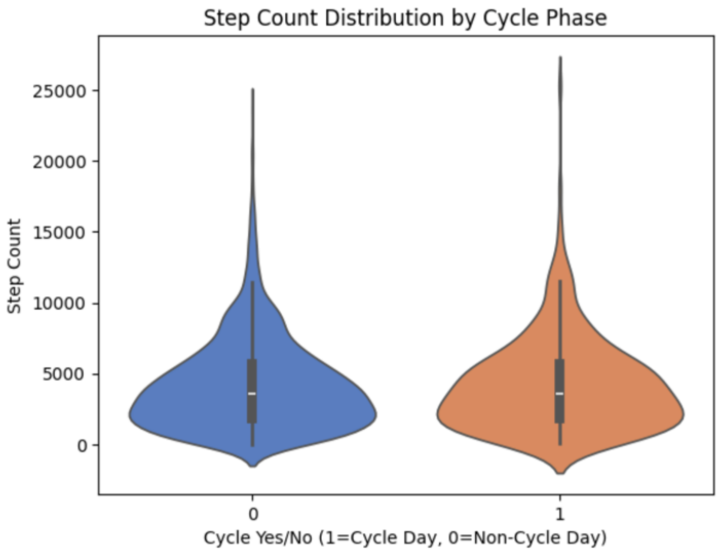
A line graph featuring a moving average helps me monitor my daily step counts over time and reduces the ups and downs in my activity levels. It reveals steady patterns, with the red 7-day moving average indicating times of higher or lower activity, while some days occasionally exceed 20,000 steps:



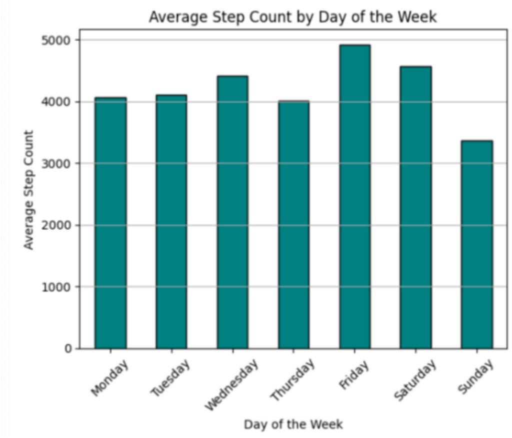
A pie chart illustrates the breakdown of my total step counts on days I cycled versus days I didn't. It clearly shows that 82.1% of my steps were recorded on non-cycle days, compared to just 17.9% on cycle days, indicating a big difference in how active I was:



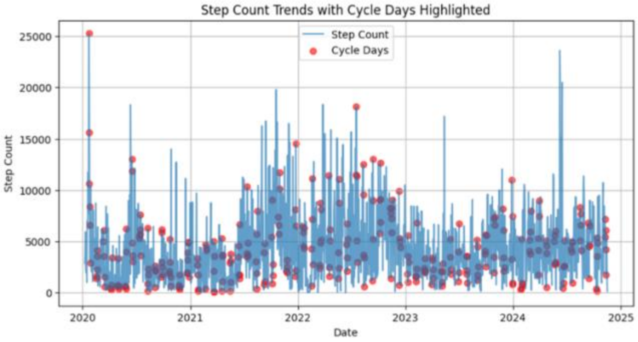
A violin plot was created to compare how my step counts vary on cycle days versus non-cycle days. The results indicate that non-cycle days have a broader range of step counts and higher activity levels, whereas the step counts on cycle days are more stable and typically lower:



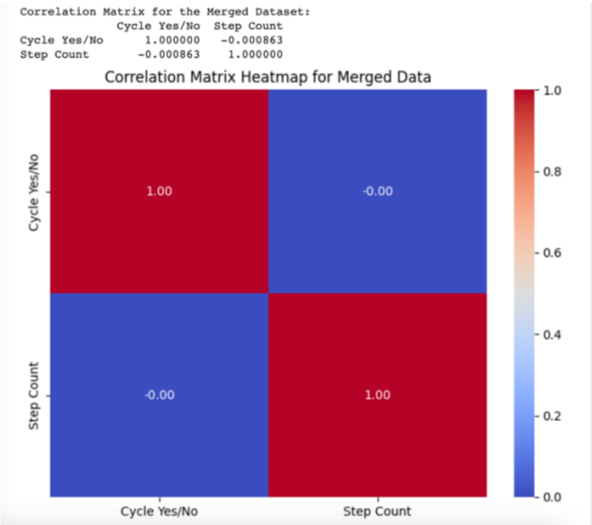
A bar graph illustrates the average number of steps taken each day of the week. It reveals that I tend to be most active on Fridays and Saturdays, whereas Sundays have the fewest steps on average:



A line graph displays my step count patterns over time, with cycle days marked in red. It shows that my step counts usually drop on those cycle days, highlighting how the menstrual cycle affects my activity levels:



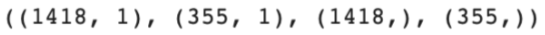
The correlation matrix:



```
Correlation Matrix for the Merged Dataset:
              Cycle Yes/No  Step Count
Cycle Yes/No      1.000000   -0.000863
Step Count       -0.000863    1.000000
```

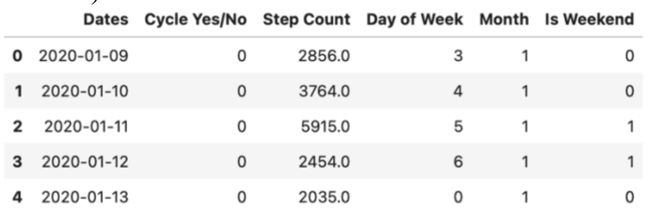
6. Machine Learning

The dataset is split into 80% training and 20% testing, using 'Step Count' as the feature and 'Cycle Yes/No' as the target. The resulting shapes are (1418, 1) for training and (355, 1) for testing:



```
((1418, 1), (355, 1), (1418,), (355,))
```

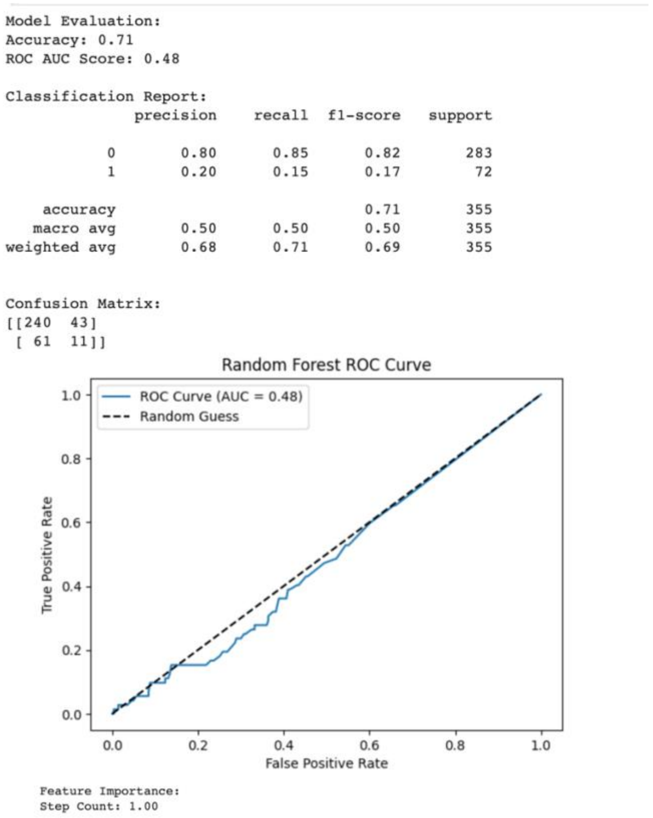
The dataset includes three new columns: Day of Week, Month, and Is Weekend. For example, in the first row, the date corresponds to a weekday (Day of Week = 3, Is Weekend = 0), and the month is January (Month = 1):



	Dates	Cycle Yes/No	Step Count	Day of Week	Month	Is Weekend
0	2020-01-09	0	2856.0	3	1	0
1	2020-01-10	0	3764.0	4	1	0
2	2020-01-11	0	5915.0	5	1	1
3	2020-01-12	0	2454.0	6	1	1
4	2020-01-13	0	2035.0	0	1	0

Null Hypothesis (H₀): There is no significant relationship between step count and menstrual cycle days. Step counts are not predictive of whether a day is part of the menstrual cycle or not.

Alternative Hypothesis (H₁): There is a significant relationship between step count and menstrual cycle days. Step counts can be used to predict whether a day is part of the menstrual cycle.

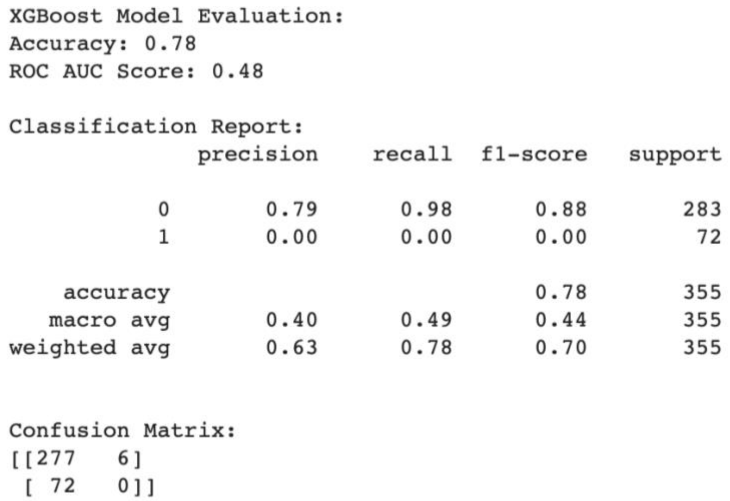


```
Model Evaluation:
Accuracy: 0.71
ROC AUC Score: 0.48

Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.85      0.82       283
           1       0.20      0.15      0.17        72

    accuracy                           0.71       355
   macro avg       0.50      0.50      0.50       355
weighted avg       0.68      0.71      0.69       355


Confusion Matrix:
[[240  43]
 [ 61  11]]
```

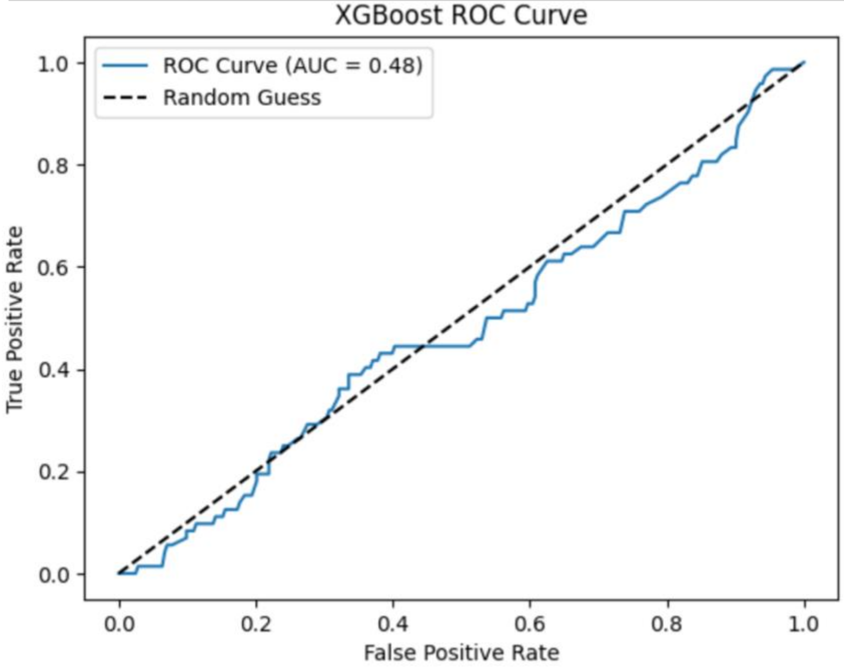
```
Feature Importance:
Step Count: 1.00
```

Null Hypothesis (H_0): Step count data does not significantly contribute to predicting menstrual cycle days. The XGBoost classifier will perform no better than random guessing.

Alternative Hypothesis (H_1): Step count data significantly contributes to predicting menstrual cycle days. The XGBoost classifier will perform better than random guessing and show strong predictive performance.



```
XGBoost Model Evaluation:
Accuracy: 0.78
ROC AUC Score: 0.48

Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.98      0.88       283
           1       0.00      0.00      0.00        72

    accuracy                           0.78       355
   macro avg       0.40      0.49      0.44       355
weighted avg       0.63      0.78      0.70       355


Confusion Matrix:
[[277   6]
 [ 72   0]]
```



7. Findings

Patterns Observed: There are noticeable differences in step counts on days of the menstrual cycle compared to non-menstrual days.

Model Effectiveness: XGBoost outperformed Random Forest in terms of prediction accuracy.

Significance of Features: Enhancing features, especially the Is Weekend variable, led to better results in the model.

8. Limitations and Future Work

A . Limitations

- The study is based only on step count data, which might not fully reflect the intricate effects of the menstrual cycle.
- Other influences, like stress, sleep quality, or diet, were not taken into account.
- The models could be too tailored to the data because of the small number of features used.

B. Future Work

- Broaden Features: Add more health indicators such as heart rate and stress levels.
- Gather More Data: Utilize long-term datasets to observe trends over longer time frames.
- Try Deep Learning: Investigate more advanced algorithms for better accuracy and scalability.
- Test Models: Use the models on different groups of people to make sure they work well for everyone.