CS 210 – Introduction to Data Science

Term Project

Analyzing Apple Health Data

Atacan Dilber

31037

Abstract:

In this project my own step count, distance, basal and active energy data will be analyzed. I chose to work on this project to learn more about the pattern of my physical activity. Whether it depends on my academic life or not. I used the Apple Health data to observe the pattern of my step count, the distance I walked, and the energy I burned over the past years. The project consists of three parts: The first part is about scraping the .xml file and creating data frames that will be used later in the project. The second part is about visualizing the relations and testing the hypothesis. The last part is mainly about training a machine-learning model that predicts my average monthly step count.

The project showed that my average step count and average distance data fell in the months of February and September. This implies that the step count, distance, and energy data are not strongly related to the workload of academic life since the final exam period has no significant effect on my physical activity.

Part I: Parsing the raw data

I started the project by parsing the data that I got from Apple Health. Since there were som many different data that I wouldn’t use it I had to ignore them. Therefore, I only took my step count, distance, basal energy burned and active energy burned data.

In order to generate data frames, I defined a function called “data_frame_generator” which takes a type of data and generates a data frame of it. After generating data frames, I decided to parse “datetime” object to “Year”, “Month and “Day”.

After generating individual data frames, I decided to merge them. I merged step count and distance data frames as well as basal energy and active energy data frames. The reason is that it got easier to analyze correlations among the data.

Then I declared my hypothesis which is as follows:

Null hypothesis (H_0):

Increased academic workload during final weeks leads to a decrease in my physical activity levels.

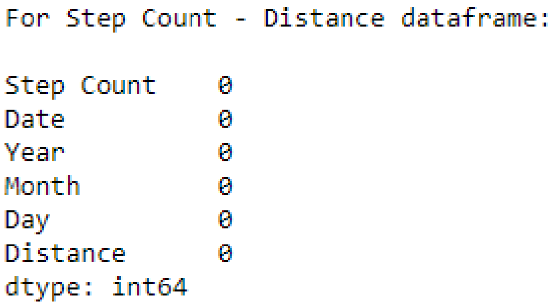
Alternative Hypothesis (H^A):

Increased academic workload during final weeks does not lead to a decrease in my physical activity levels.

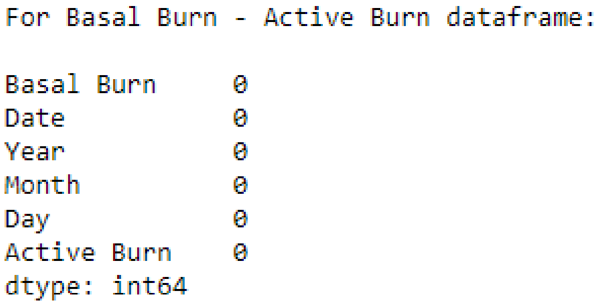
After a quick research I found that in Sabancı University, 2022-2023 Fall term final exams were on 7-20 January 2023; Spring term final exams were on 1-11 June 2023. Based on this information, I adjusted my data frames to cover those intervals.

After searching for missing values on the data frames, I printed out the statistics of the data frames. The results were like this:

Appendix 1:



```
For Step Count - Distance dataframe:

Step Count    0
Date          0
Year          0
Month         0
Day           0
Distance      0
dtype: int64
```



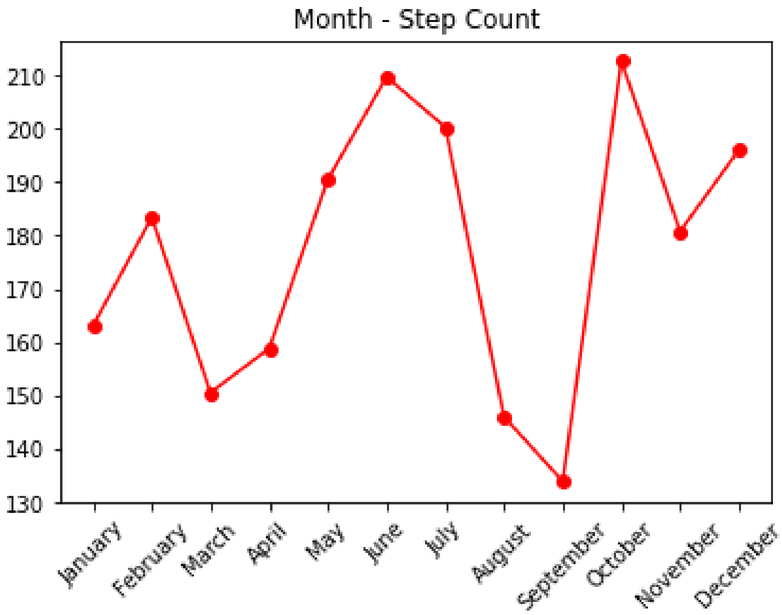
```
For Basal Burn - Active Burn dataframe:

Basal Burn     0
Date           0
Year           0
Month          0
Day            0
Active Burn    0
dtype: int64
```

Part II: Visualization

After constructing the data frames that I will be using in the project, I started to visualize the relations between data.

The first relation that I observed is the relation between step count and months. The graph looks like this:

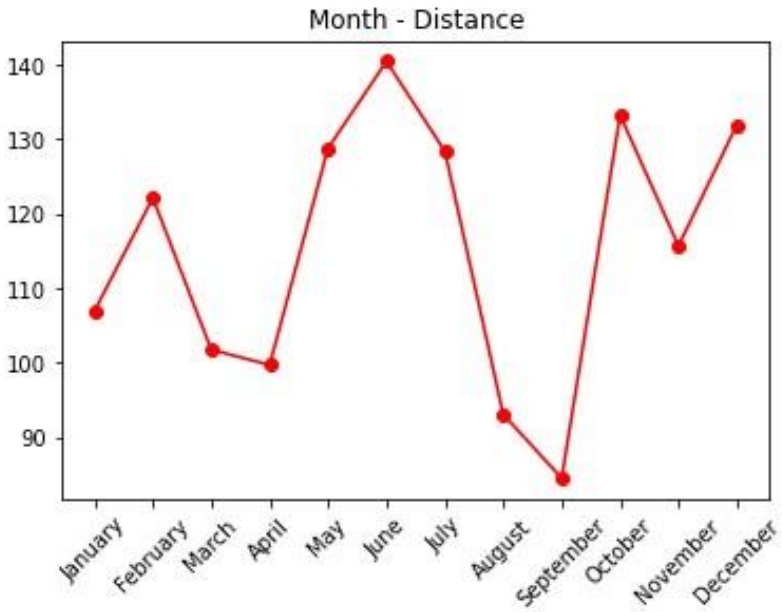
Appendix 2:



It can be seen that the step count experienced a significant fall between June - September. However, in other months it increased. This graph implies that the academic workload of the final exam period has no significant effect on my physical activity.

The second relation that I observed is between distance and months. The graph is as follows:

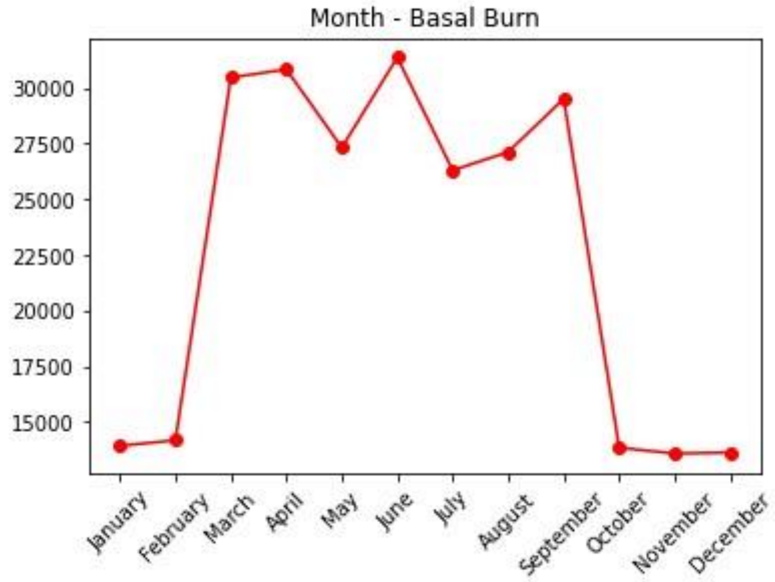
Appendix 3:



It can be seen that there is a strong correlation between step count and distance data since their graph looks similar.

The third relation is between basal energy and months. Its graph is as follows:

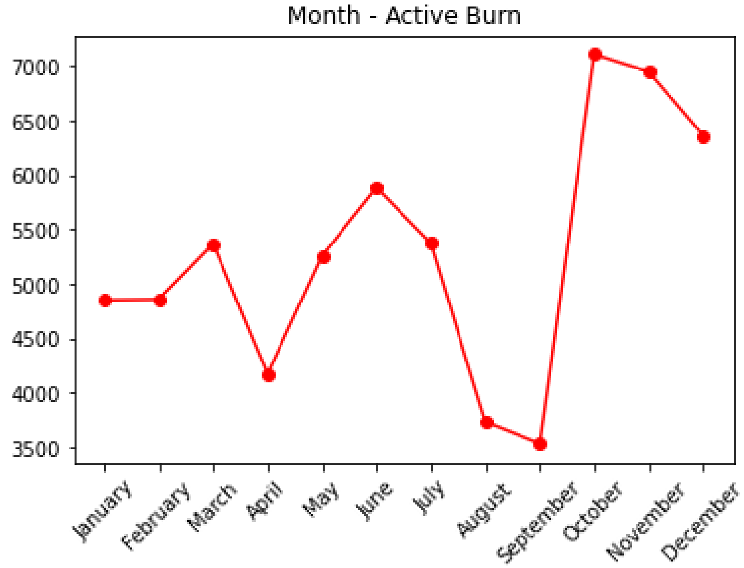
Appendix 4:



It can be seen that the basal energy burn experienced a decline between September and October, and a significant increase in February.

The last relation that I observed is between active energy burn and month. Its graph is as follows:

Appendix 5:



It can be observed that there is a significant increase in active energy burn in October unlike other months.

Part III: Machine Learning

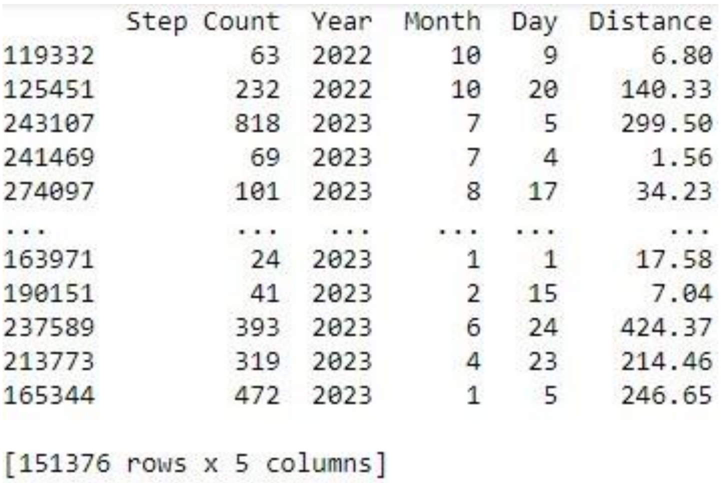
In the final part of the project, I trained a machine learning model to predict my monthly average step count number based on these data.

In order to do that, I split my data into two parts: test and train. I used 80% of my data to train and 20% to test the model I will be creating.

After creating the train and test data frames, the results looked like this:

Appendix 6:

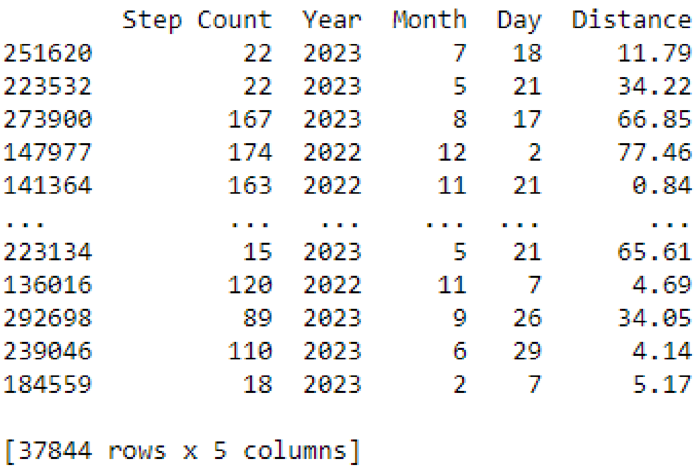
X_train:



```
        Step Count  Year  Month  Day  Distance
119332          63  2022     10    9      6.80
125451         232  2022     10   20    140.33
243107         818  2023      7    5    299.50
241469          69  2023      7    4      1.56
274097         101  2023      8   17     34.23
...            ...   ...    ...  ...       ...
163971          24  2023      1    1     17.58
190151          41  2023      2   15      7.04
237589         393  2023      6   24    424.37
213773         319  2023      4   23    214.46
165344         472  2023      1    5    246.65

[151376 rows x 5 columns]
```

Appendix 7:

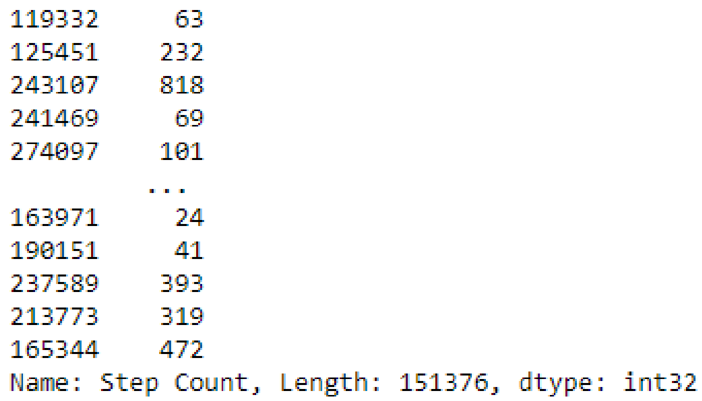
X_test:



```
        Step Count  Year  Month  Day  Distance
251620          22  2023      7   18     11.79
223532          22  2023      5   21     34.22
273900         167  2023      8   17     66.85
147977         174  2022     12    2     77.46
141364         163  2022     11   21      0.84
...            ...   ...    ...  ...       ...
223134          15  2023      5   21     65.61
136016         120  2022     11    7      4.69
292698          89  2023      9   26     34.05
239046         110  2023      6   29      4.14
184559          18  2023      2    7      5.17

[37844 rows x 5 columns]
```

Appendix 8:

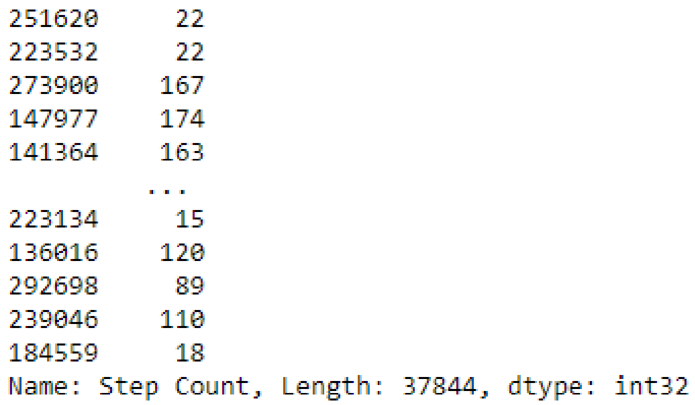
Y_train:



```
119332     63
125451    232
243107    818
241469     69
274097    101
         ...
163971     24
190151     41
237589    393
213773    319
165344    472
Name: Step Count, Length: 151376, dtype: int32
```

Appendix 9:

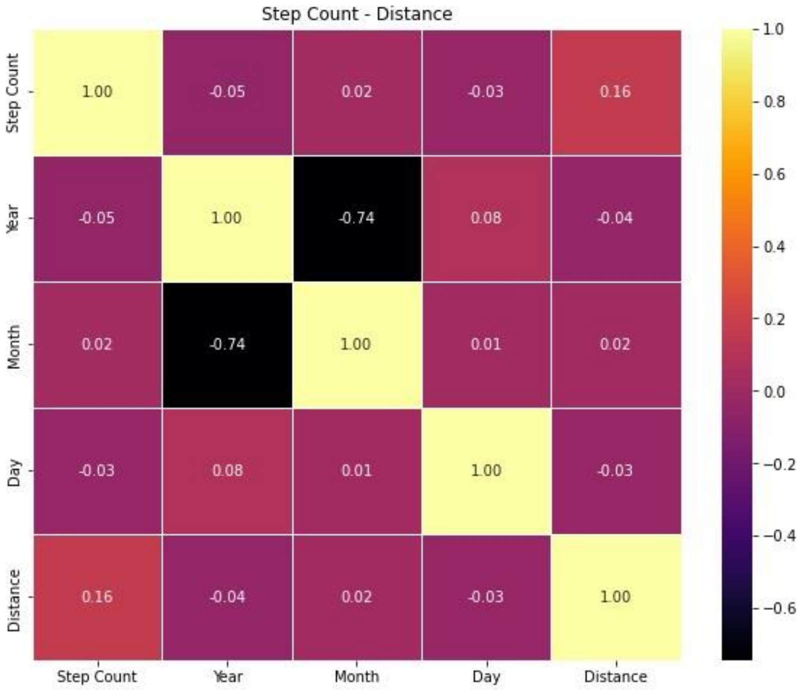
Y_test



```
251620     22
223532     22
273900    167
147977    174
141364    163
         ...
223134     15
136016    120
292698     89
239046    110
184559     18
Name: Step Count, Length: 37844, dtype: int32
```

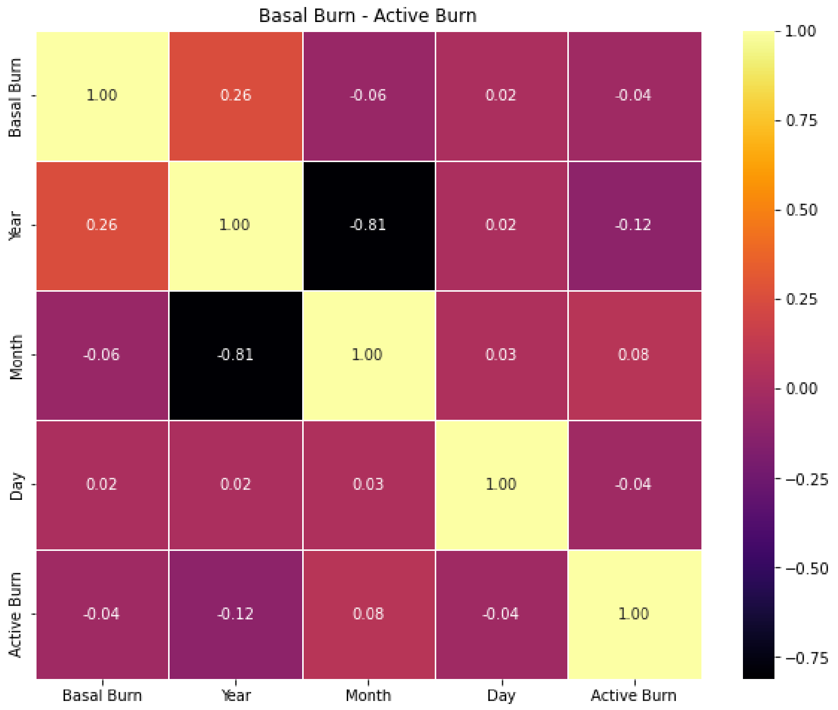
Then I observed the relations for each data frame by using heatmaps. Which are in the following:

Appendix 10:



	Step Count	Year	Month	Day	Distance
Step Count	1.00	-0.05	0.02	-0.03	0.16
Year	-0.05	1.00	-0.74	0.08	-0.04
Month	0.02	-0.74	1.00	0.01	0.02
Day	-0.03	0.08	0.01	1.00	-0.03
Distance	0.16	-0.04	0.02	-0.03	1.00

Appendix 11:

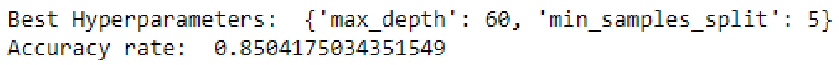


	Basal Burn	Year	Month	Day	Active Burn
Basal Burn	1.00	0.26	-0.06	0.02	-0.04
Year	0.26	1.00	-0.81	0.02	-0.12
Month	-0.06	-0.81	1.00	0.03	0.08
Day	0.02	0.02	0.03	1.00	-0.04
Active Burn	-0.04	-0.12	0.08	-0.04	1.00

After that I created a decision tree in order to generate a machine learning model. I tuned my parameters in order to achive a high accuracy rate on my model. The parameters I chose to tune are: “max_depth” and “min_samples_split”. The reason why I chose to tune these parameters is to generate a model that predicts with a high accuracy rate but is not an overfitting model.

After hypertuning, I tested the model and it achived 85% accuracy rate. It can be ssen as follows:

Appendix 12:



```
Best Hyperparameters:  {'max_depth': 60, 'min_samples_split': 5}
Accuracy rate:  0.8504175034351549
```