CS 210 Course Project Report

İlhan Sertelli – 30567

Introduction

In this study, the main objective is to create correlation between the health data in different timelines. Based on the correlations have been found, it is aimed to use the data in order to make predictions about how the pattern of the health data in the following periods will be like

Source of Data

The source of the data in order to determine patterns is Apple Health application. The data has been retrieved in an XML format. In the following phases of this project, the data has been parsed, filtered by using several different libraries such as pandas and BeautifulSoup in Python.

Structure of Dataset

The dataset consists of seven different features which are the total step count, distance that walked, total basal energy burnt, total active energy burnt, average loudness of the music that I listened to, average step length, and lastly average walking speed. The structure of the dataframe is given as follows

- **Total Step Count Dataframe**

	type	unit	creationDate	value
0	HKQuantityTypeIdentifierStepCount	count	2022-01-01	7993
1	HKQuantityTypeIdentifierStepCount	count	2022-01-02	859
2	HKQuantityTypeIdentifierStepCount	count	2022-01-03	6420
3	HKQuantityTypeIdentifierStepCount	count	2022-01-04	1966
4	HKQuantityTypeIdentifierStepCount	count	2022-01-05	1733

- **HealthDf (All features are merged)**

	creationDate	unit	Step Count	unit	Distance Walked	unit	Basal Energy Burnt	unit	Active Energy Burnt	unit	Average Loudness	unit	Average Step Length	unit	Average Walking Speed
0	1.0	count	9186	km	5.772930	kcal	1854.989	kcal	260.084	db	54.461460	cm	65.062500	km/hr	4.313250
1	2.0	count	8448	km	6.123411	kcal	1759.119	kcal	328.035	db	59.117510	cm	69.795455	km/hr	4.758545
2	3.0	count	10294	km	7.600093	kcal	1898.778	kcal	409.668	db	60.622271	cm	70.927536	km/hr	4.891304
3	4.0	count	4908	km	3.372441	kcal	1762.170	kcal	178.509	db	50.637400	cm	68.333333	km/hr	4.750500
4	5.0	count	8781	km	6.271012	kcal	1823.188	kcal	351.409	db	51.421989	cm	70.135135	km/hr	4.838595

Data Cleaning

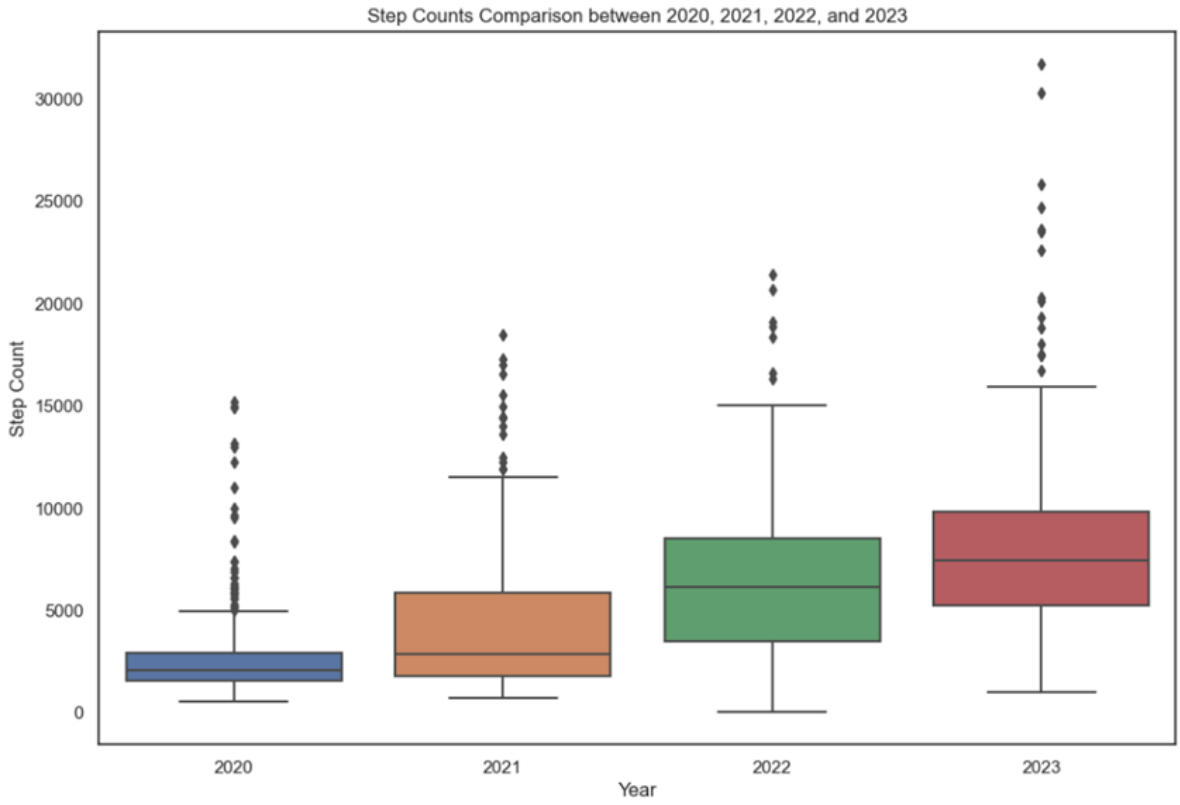
After constructing the dataframe, the data cleaning operation is done. Firstly, the rows of the dataframe was checked in case whether there are any missing values or not. If there are missing values, the "NaN" values are replaced with the mode of the columns. Subsequently, the seven features that take part in the dataset were divided into seven different dataframes. After examining the each dataframe separately, they were merged into a one single dataframe again. (As it is shown above named healthDf)

Visualizations

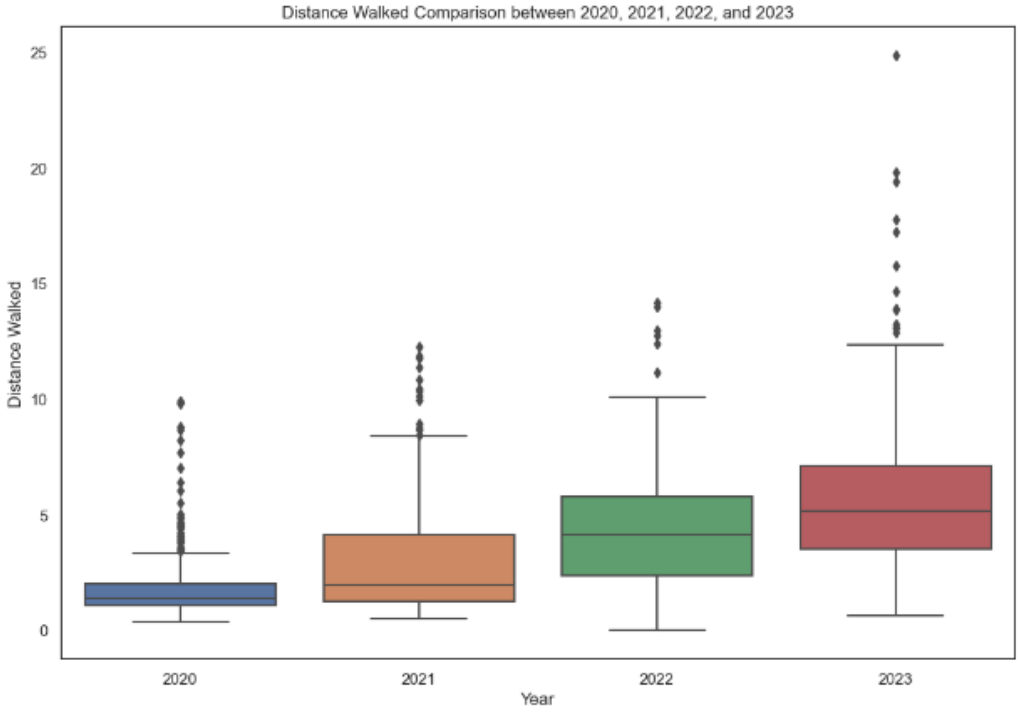
After applying the data cleaning operation, in order to check the hypothesis "The transition from high school to the university had a significant change in the pattern of my health activities", the data were visualised and represented in various models. In this phase of the project, matplotlib and seaborn tools were used in Python.

Firstly, the annual step counts and the distance walked data were compared by regarding every period 1 year by using box plots. The results of the given comparisons are presented in the below charts.

- **Step Counts Comparison Between 2020, 2021, 2022, 2023**

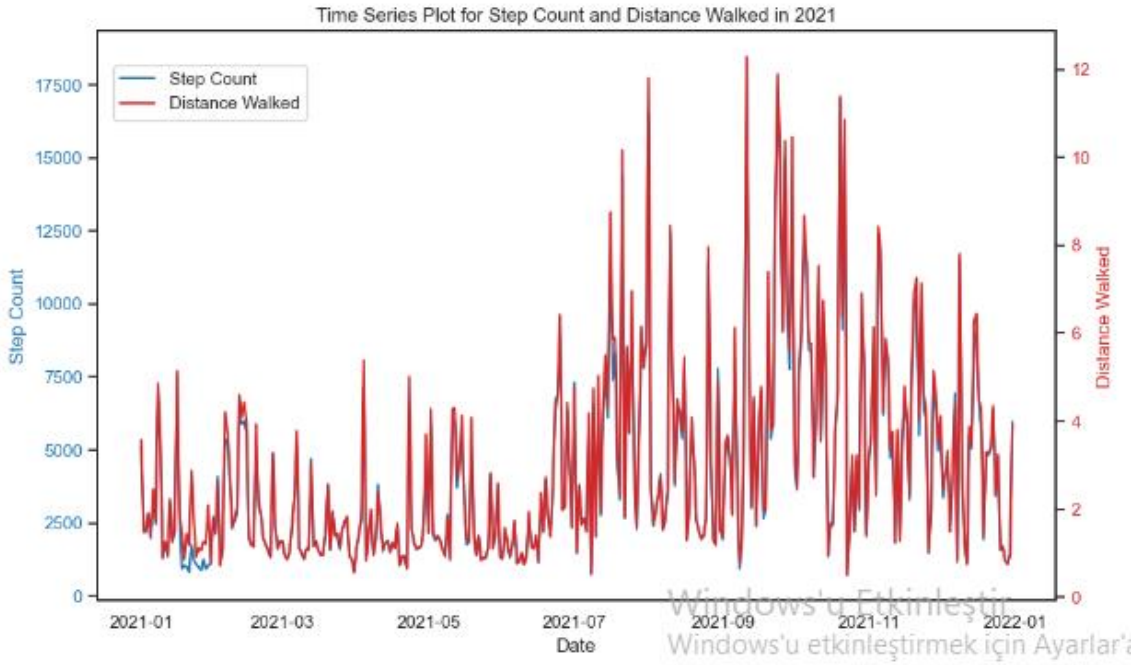


- **Distance Walked Comparison Between 2020, 2021, 2022, 2023**

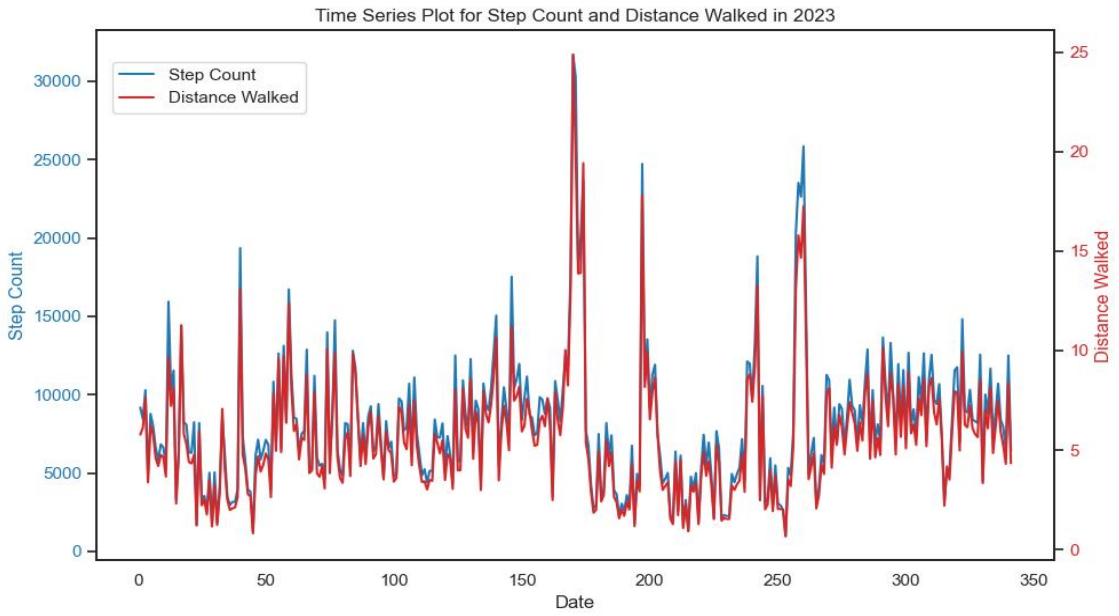


After examining the box plots, there seems to be an obvious change in the average of the step count and the distance walked data between the years 2020 and 2023. Since there is a huge similarity in the shape of the above two graphs, I combined the step count and distance walked data and plotted to observe the similarity of the patterns. One plot is created for 2021 and one plot is created for 2023.

- **The Plot of the Step Count and Distance Walked Data in 2021 (Combined)**

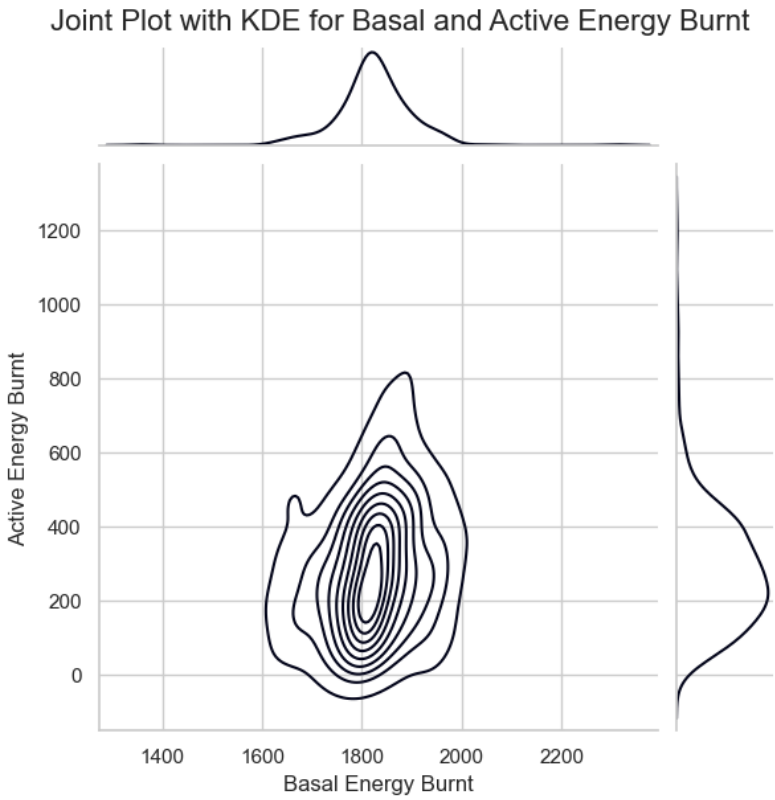


- **The Plot of the Step Count and Distance Walked Data in 2023 (Combined)**

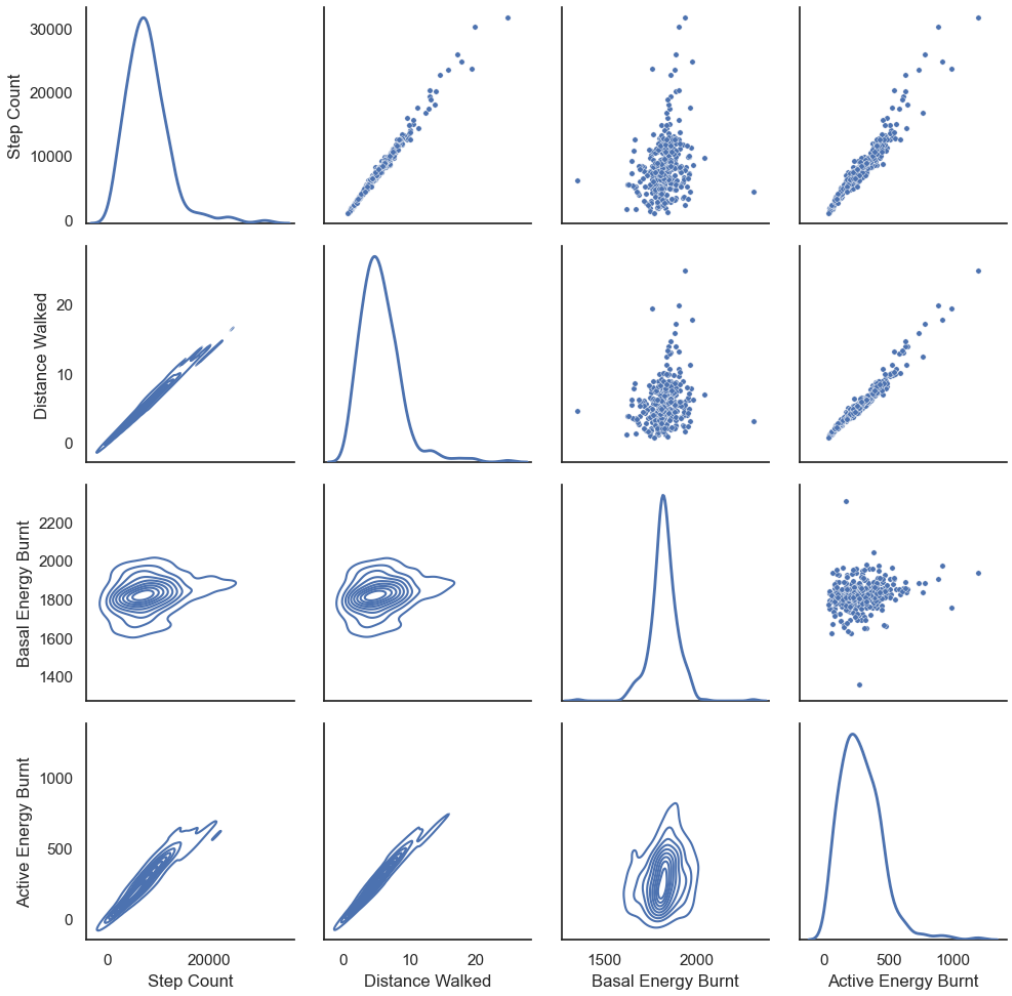


After determining an extreme similarity between the pattern step count and distance walked data, I also combined the basal energy burnt and active energy by using a joint plot with KDE.

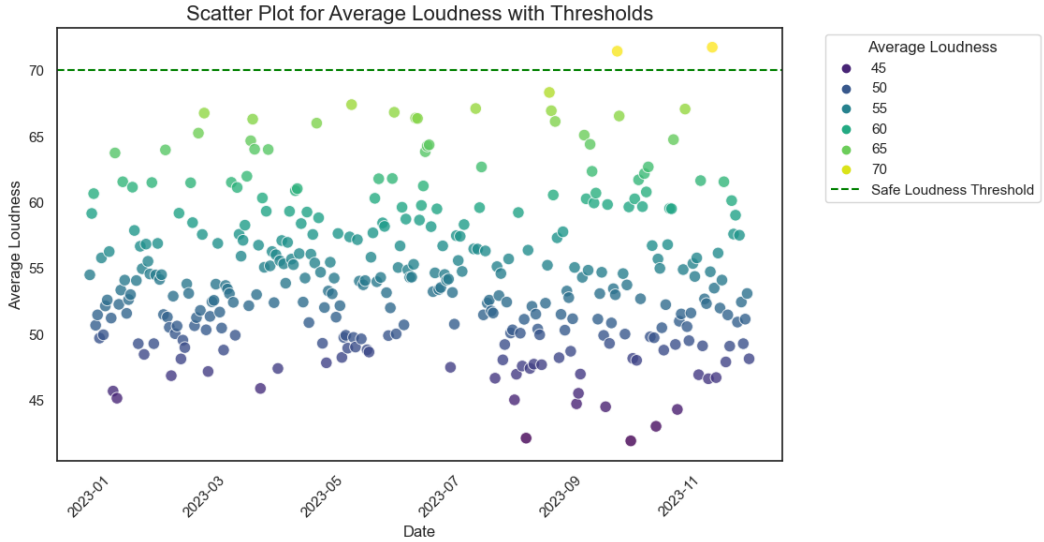
- **Joint Plot with KDE for Basal and Active Energy Burnt**

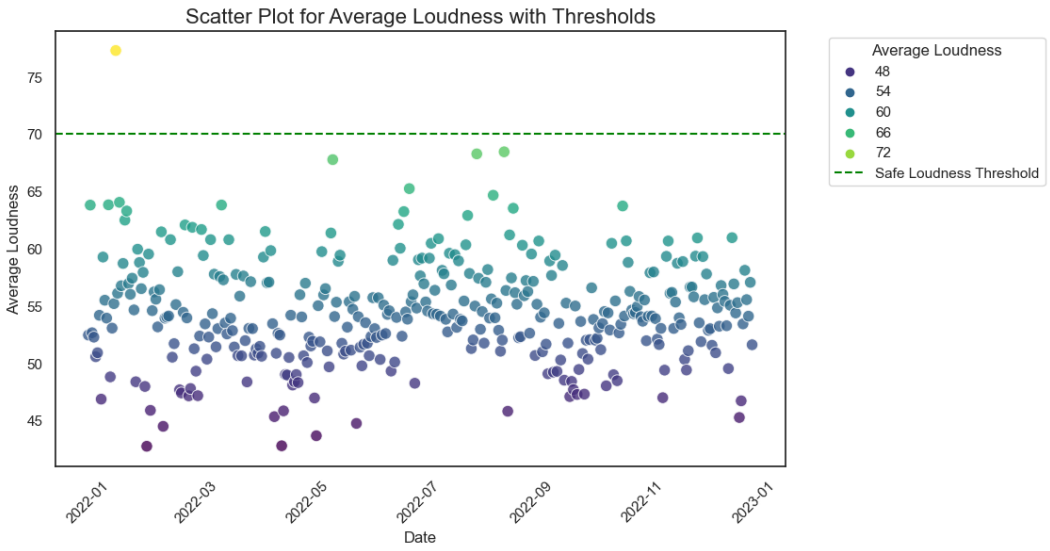


- **Pair Plot of Step Count, Distance Walked, Basal Energy Burnt, Active Energy Burnt Data**



- **Average Loudness of the music I listened to in terms of decibels (dB) in 2022 and 2023**

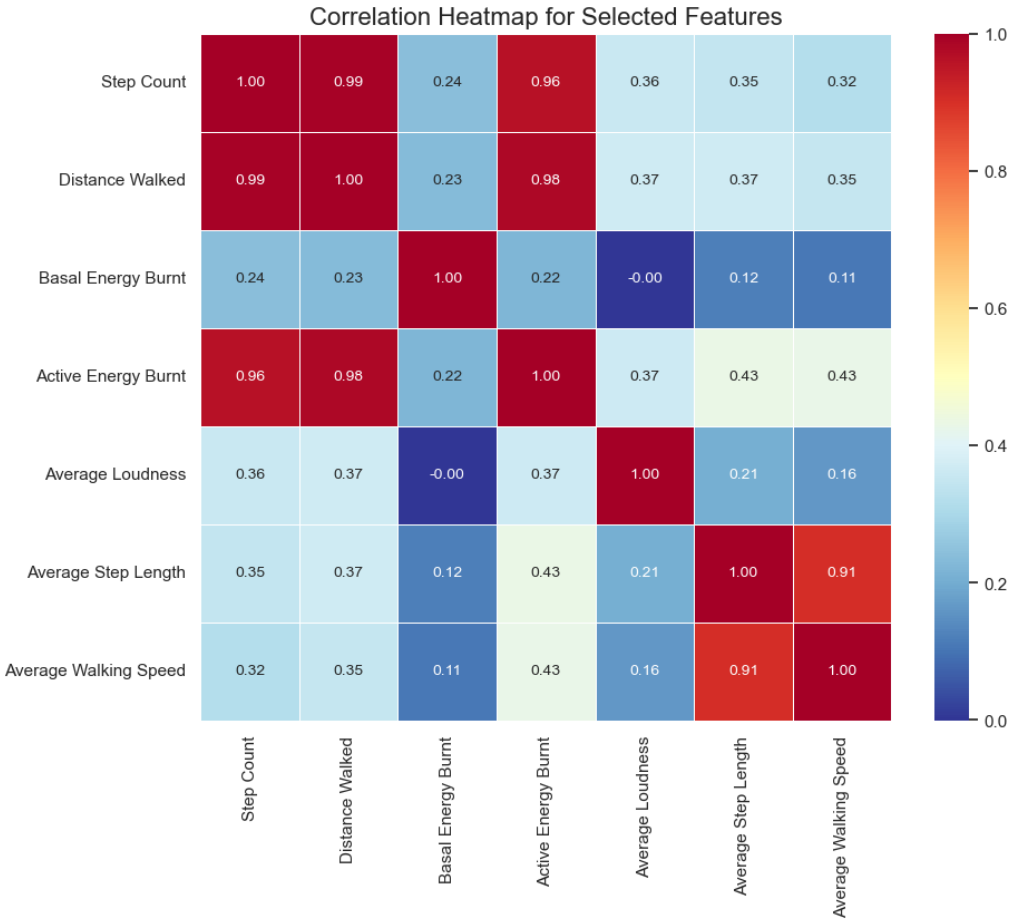




Correlations

Analyzed how the features that take part in the dataframe are correlated with each other by using a heatmap.

- **Correlations Heatmap**



Key Insights

Based on the given charts, it can be aseerted that:

- Distance Walked and Step Count are strongly correlated
- Basal Energy Burnt and Active Energy Burnt are strongly correlated
- Distance Walked and Energy Burnt (Both basal and active) are strongly correlated
- Step Count and Energy Burnt (Both basal and active) are strongly correlated
- Between 2020 and 2023, there is a considerable change in the quantities of the data. (To be tested in the Hypothesis Testing part)

Hypothesis Testing

In order to check my hypothesis "The transition from high school to the university had a significant change in the pattern of my health activities", I applied t-test on the "Distance Walked" data of 2020 and 2023. After the application on t-test, the result is presented below:

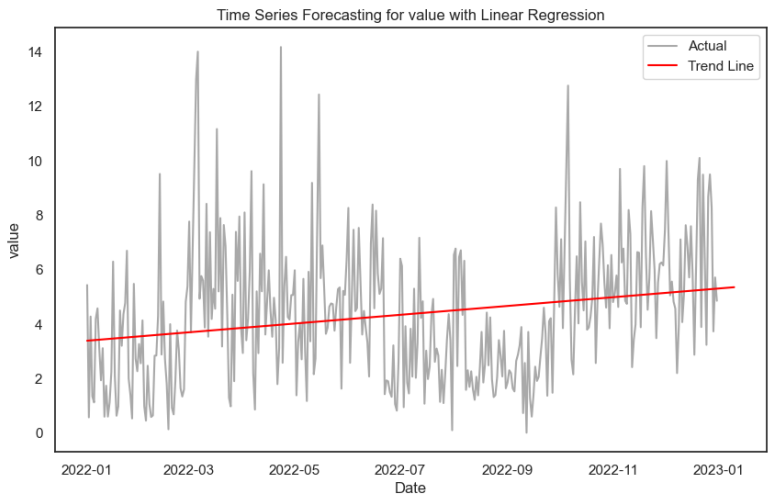
```
T-test for Step Count: t-statistic = -20.706214823344116, p-value = 9.068183860408197e-75
```

Result: As a result of the t-test, the p-value is found extremely less than the threshold (0.05) which implies that the transition from high school to the university had a significant effect on the pattern of my health activities.

Machine Learning Techniques

I lastly did a machine learning operation in order to measure how accurate does the program make predictions about the patterns of the activity. In order to do that, I firstly chose linear regression model to apply. The regression model has been used on the "Distance Walked" data. Moreover, the data is splitted for training and testing. 20% of the data is allocated for testing and 80% of the data is allocated for training.

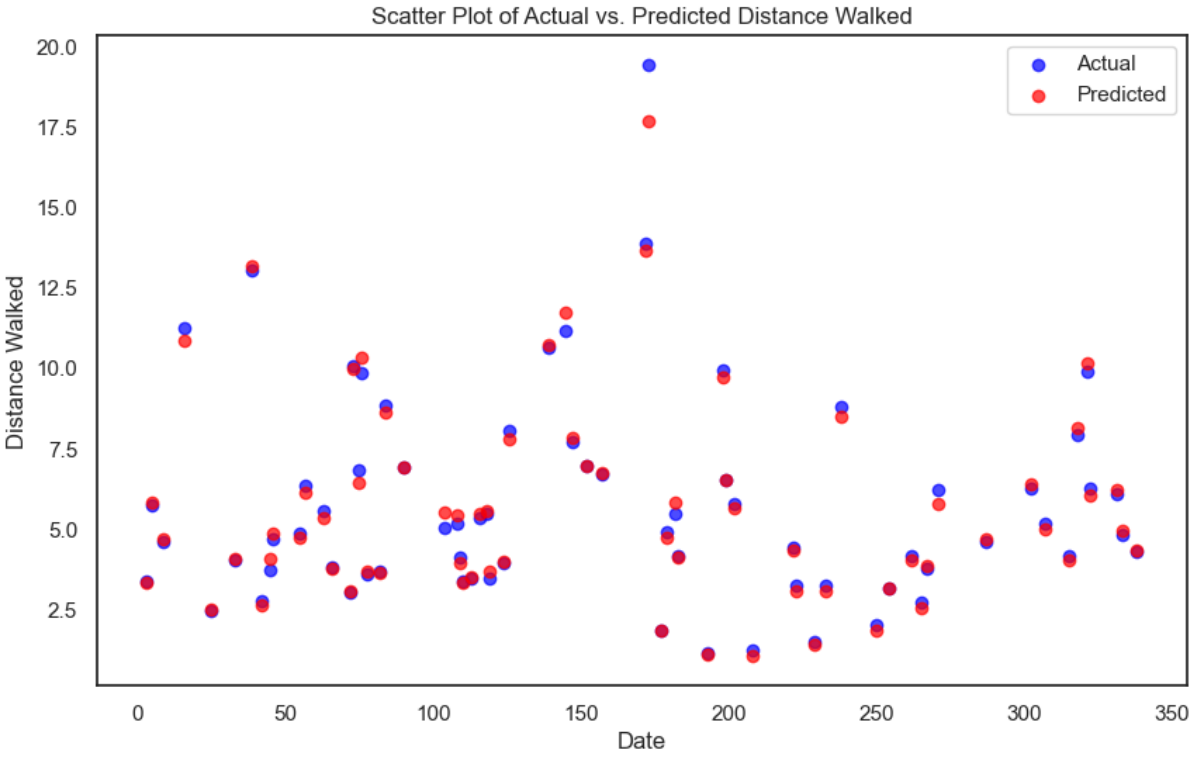
- **The linear regression model:**



Predictions and Comparison

After the training process, the performance of the model has been mesaured by making comparisons between the prediction data and the actual data. In addition, mean squared error technique is used in this operation.

- **The Scatterplot of the Actual and Predicted Data**




```
Mean Squared Error: 0.08485120859432689
```

Conclusion

To summarize, according to the observed patterns and applied tests on the data, it can be concluded that the transition from high school to university had an obvious effect on the pattern of my health activites. While conducting this project, the limitations are

- Limited type of data
- Only usage of personal data is allowed
- Process of finding an obvious correlation between data

References