# The Correlation Between the Weather and Step Counts

CS210 Fall 23-24 Term Project

Sinan Özkan
28966

## Abstract:

This report aims explain the outcomes of a project which aims to investigate the impact of weather conditions on daily step count. The results of hypothesis testing aim to determine whether weather conditions have a significant effect on students’ (Sinan Özkan) daily step counts. The research seeks to develop an understanding of how daily physical activity may vary under various weather conditions.

## Introduction:

Physical activities and their influence on our lives, cannot be underestimated. Most of the research shows that the physically active life, leads to various benefits into the individuals’ well-being. It is assumed that various factors have an impact on the daily activity levels, such as working schedule, mental situations, and: weather conditions.

This study aims to investigate the correlation between weather conditions and daily step counts of a student who lives in İstanbul. The null hypothesis (H0) is suggesting that the step counts do not have any correlation with weather conditions, and the alternative hypothesis (H1) suggests that there is a relation between these two elements. By analyzing this relationship, project aims to understand the impact of the environmental factors into the physical activity.

# Theoretical Framework:

With the intent of understanding the dynamics of how external factors have an influence on daily physical activities, many different aspects have been handled. Bronfenbrenner, who is a Russian American psychologist known by his ecological systems theory, suggests that people are influenced by an interconnected system, and weather is contained by the macrosystem which represent the immediate surroundings (Bronfenbrenner, 1979). By handling these conditions with another aspect, research asserts that weather has a direct influence on moods of the individuals, and these mental directions has a relationship with the desire of being physically active (Lozano et al., 2021). In the light of these kind of research, the project aims to analyze the connections between the weather conditions and the measure of being physically active.

# Methodology:

To analyze the correlation, these following steps were conducted:

**1. Acquire and Clean the Data**

Firstly, a CSV file containing health data such as daily step counts, calories burned, distance walked, and bicycle distance from Apple's Health application is generated (Appendix 1). After that, a weather data from an online source, encompassing various climate details daily is acquired (Appendix 2). To facilitate merging and analysis, the file formatted the 'datetime' (weather data) and 'startDate' (step count data) columns appropriately filtered (Appendix 3), then merged the datasets together.

2. Data Analysis

To analyze the data, several visualization methods are used such as boxplot, correlation heatmap, line plot over time, heat map of step count over time, etc. To be able to use all these visualizations, seaborn library is used, and it enabled to make a comparison between the step counts of different weather conditions. In the analysis phase, by using different methods, an examination of the seasonal variations and their influence on step counts had been made.

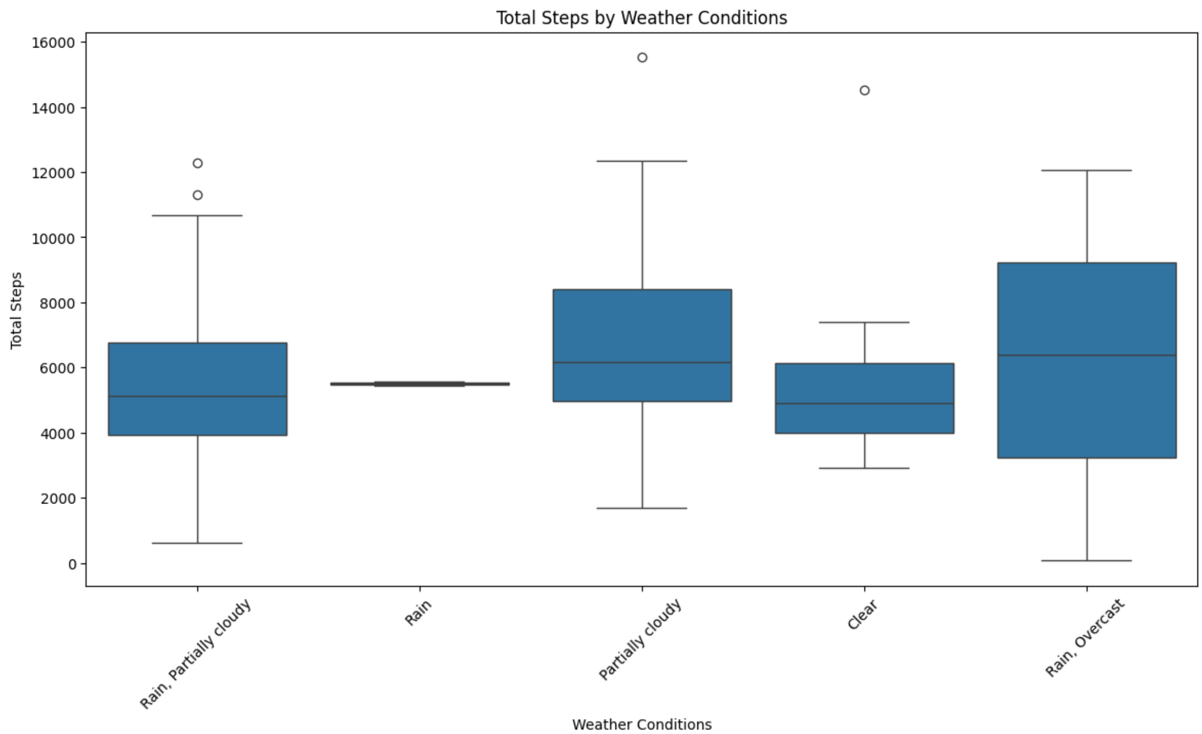
3. Statistical Analysis

This phase aims to examine the correlation in more details. Creating a correlation matrix and heatmap allows to see the details clearer and enables to discuss whether there is a direct relationship. Also, KMeans clustering analysis is employed to identify similar combinations of step counts and weather conditions, revealing distinct behavioral patterns within the dataset. In addition, a regression analysis is held to determine the correlation in a greater detail which means it is aimed to analyze the whole dataset from different perspectives with visual and quantitative methods.

# Data & Analysis

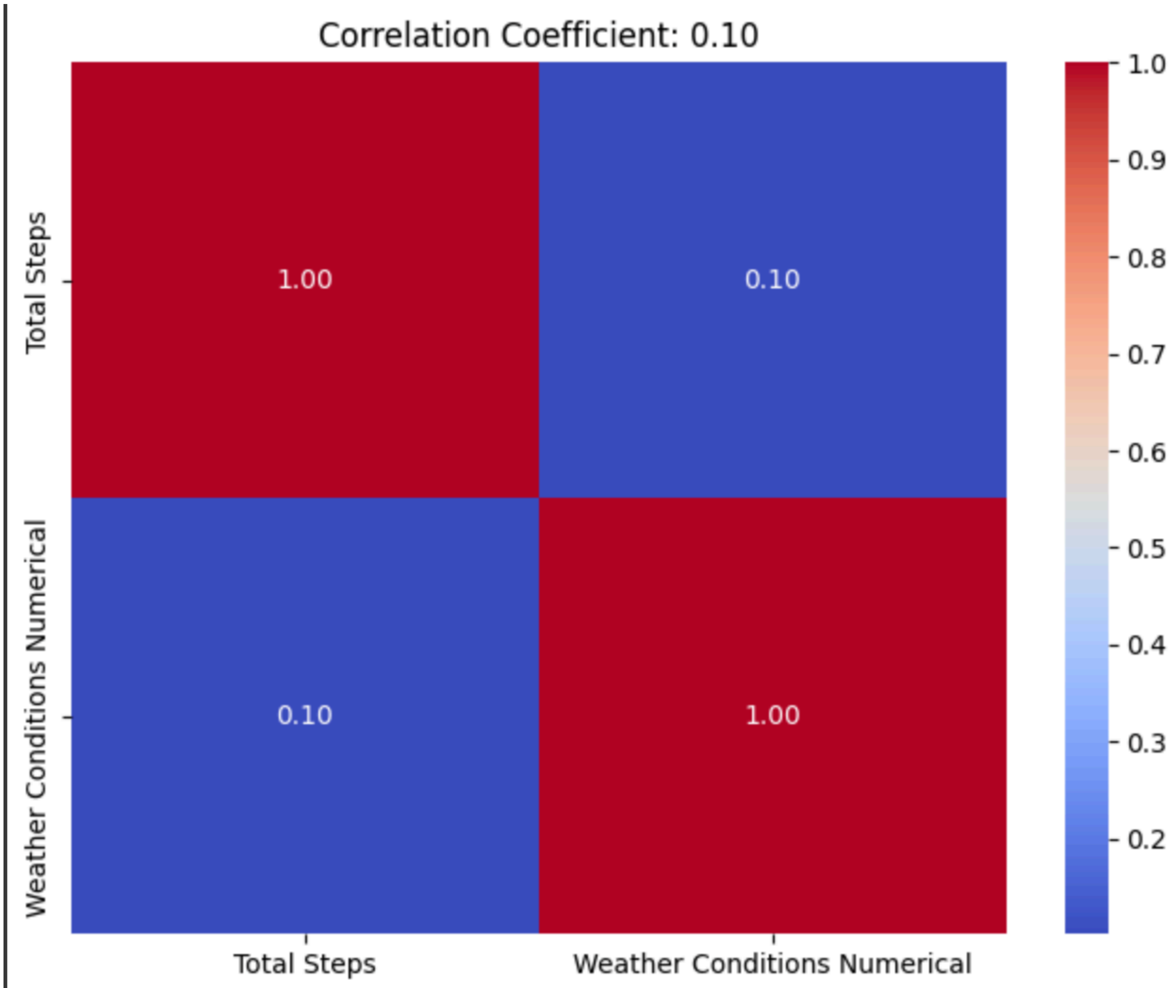
In this part of the project, it is aimed to explain the outputs which are received from the implemented code. As stated, various visualization methods are used and they will be directly seen in this part, without appendix part usage.

First, a boxplot is composed to see the variation of step counts in terms of different weather categorizations. This kind of visualization is chosen because it enables to analyze the middle point of step counts for each condition, with the lines called median placed in the boxes. Also, depending on the heights of each box, it is possible to determine the variety of step counts in several conditions. Below graph shows that the rainy over casted weather has the widest range. In addition, there are whiskers that shows dispersion and some dots that shows anomalous values which cannot be seen ordinarily, so they are not commonly used to analyze. The choice of boxplot as a visualization method facilitates a quick comparison between the central values of step counts for each weather condition. The median values within the boxes offer a robust measure of the middle point, aiding in the interpretation of the overall trend and in below graph the comparison of median values shows that the over casted rainy days are the most physically active days, and partially cloudy days are following it. Overall, it can be assumed that the boxplot is an important tool to summarize a complex step count data across diverse weather conditions.




**The Boxplot of Total Steps by Weather Conditions**

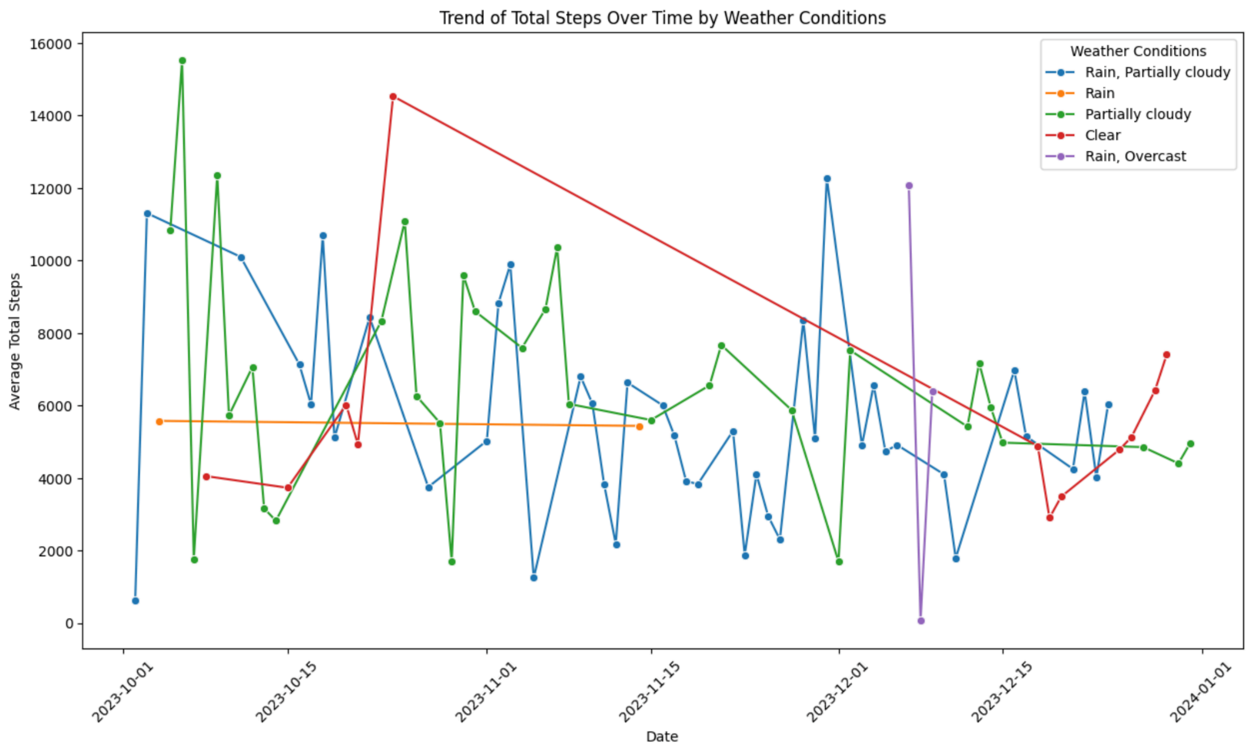
Thereafter, as a different method of visualization, a heatmap is created. This heatmap shows the correlation between the two main factors of the project and helps to determine whether the impact of the weather is significantly effective with the step counts of Sinan Özkan. Higher the correlation coefficient, higher impact would be assumed, but as the below graph and the value beneath shows that the correlation is significantly low, with a value of nearly 0.10. So, it can be assumed that there is a correlation between the weather and the step counts, but the impact might not be a big deal. Sinan Özkan may be active in a rainy day as well as a clear or cloudy day, according to the heatmap that created.




```
0.10178557607268593
```

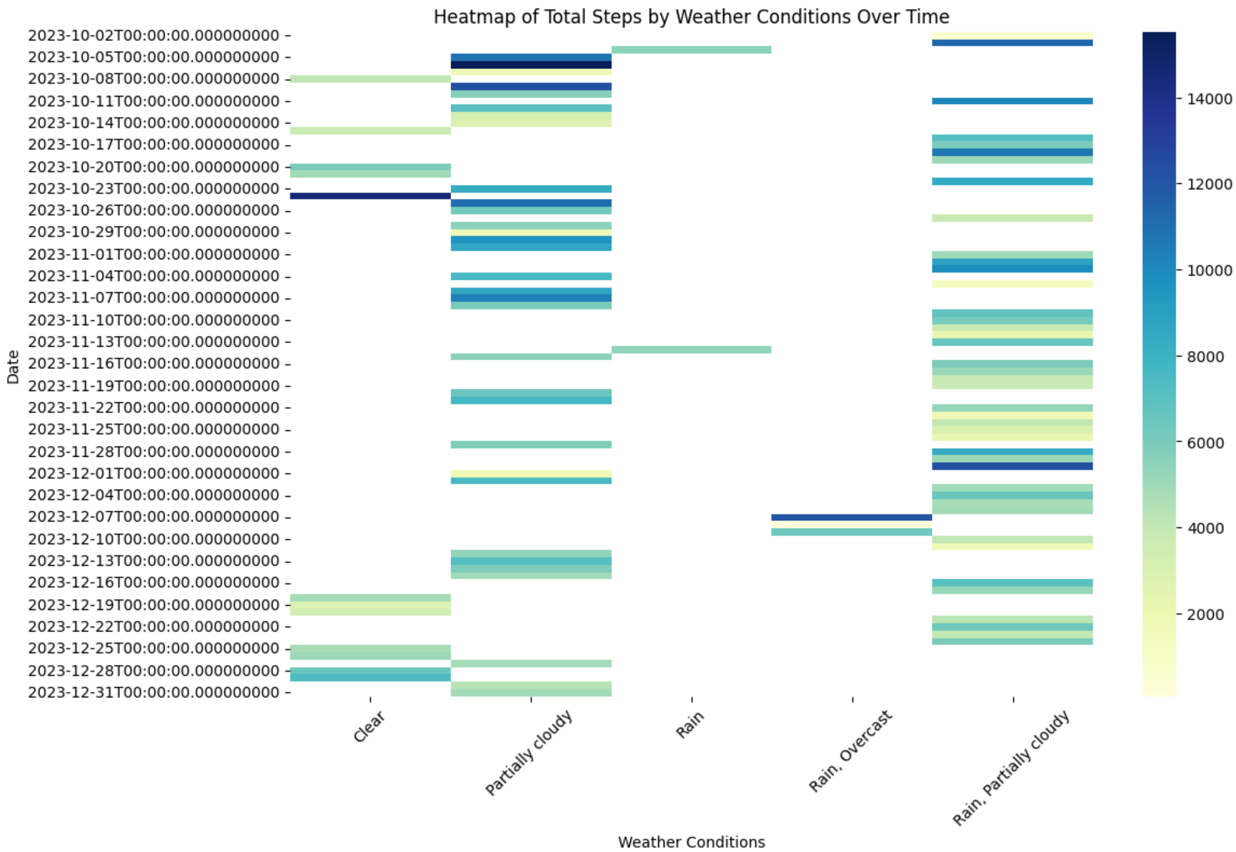
**Heatmap of Correlation**

Another method to be able to see the correlation is to create a trendline for each weather condition, following to the number of steps day within intervals of 2 weeks. By using this kind of visualization method, it is possible to see the variety of each condition, independent from the period. As it can be seen, the variety in each condition is very frequent which means a perfect correlation is quite unfair to be set. Another feature of this graph is to determine which conditions has the higher count of steps. For instance, we can observe that step counts are generally higher on "Partially Cloudy" days. This analysis shows that weather conditions have an impact on daily step counts. However, for a more precise determination of this relationship and to understand cause-and-effect relationships, a more detailed analysis of different effects, or a broader dataset may be better to detailly look into it.



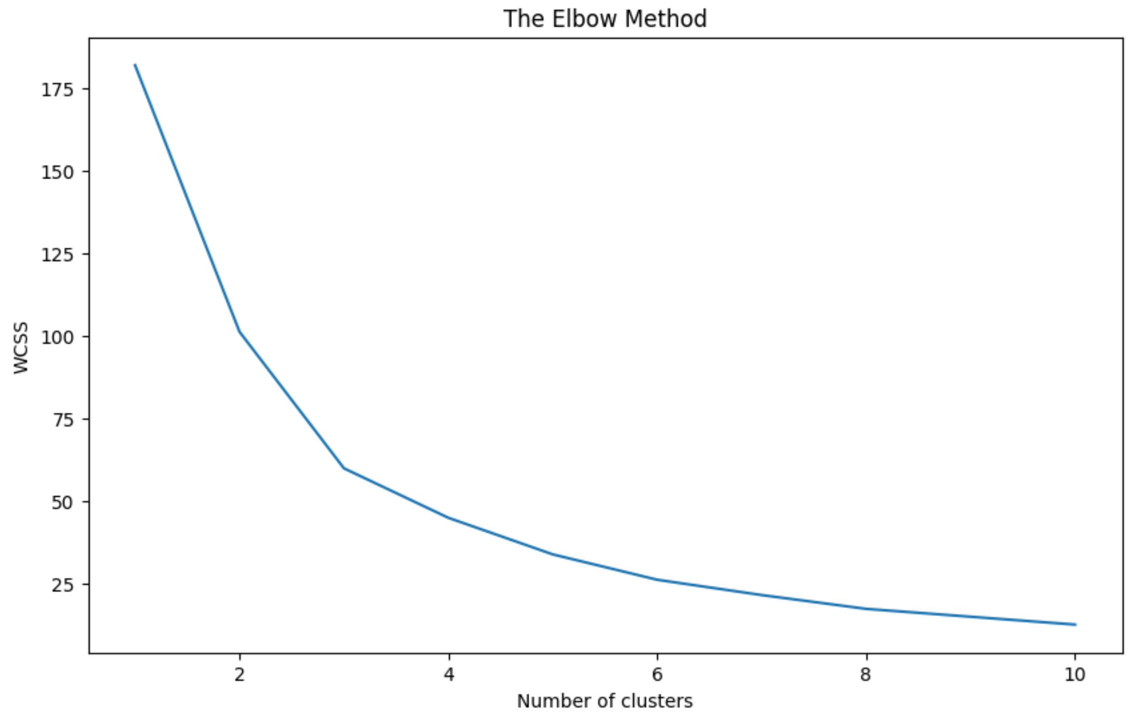

**Trendline for Total Steps by Weather Conditions**

Another type of heatmap visualization is also valid in terms of being able to see the dispersion between the weather conditions. This heatmap aims to show to density of step counts by using colors, with the denser colors, it can be assumed that the corresponding value is more likely to occur when the specified condition has arrived, and this information is clearly shown at the right hand-side of the graph.




**Heatmap Over Time**

Lastly, to evaluate the hypothesis that have been asserted, a KMeans clustering algorithm is utilized by employing the "Elbow Method" to determine the optimal number of clusters. By using this method, it is possible to calculate the WCSS values for each cluster count and illustrates how these values change. By examining the graph, we can observe a decrease in WCSS as the number of clusters increases. The elbow point is the juncture where the increase in the number of clusters does not significantly reduce WCSS. It is used to identify the optimal number of clusters. From the graph, we can observe that this elbow point falls between 2 and 4 clusters, although, as it is stated in previous parts, usage of more detailed data can affect the results in a different way.




As a final analysis, linear regression values should be handled properly:

**Coefficients:** The coefficient, found to be -273.05549872, this value suggests a negative impact of weather conditions on step counts.

**Intercept:** It has a value of 6177.465984654732, and it represents the estimated starting point for step counts under default weather conditions.

**$R^2$ Score:** It is -0.17004840487449924, indicating a potential poor fit or a lack of a linear relationship between step counts and weather conditions.

**Manual $R^2$ Score:** A manually calculated $R^2$ score of 0.062143369089733946 suggests a weak indicator for predicting step counts based on weather conditions.

**Mean Absolute Error (MAE):** The MAE value of 3095.9075245658905 indicates that, on average, the model's predictions are approximately 3096 steps away from the real values.

**Mean Squared Error (MSE):** The MSE value of 18072986.8436868 shows that the model's predictions generally contain errors compared to the real values.

**Root Mean Squared Error (RMSE):** With an RMSE value of 4251.233567293945, the model's average predictions are about 4251 steps away from the real values.

According to these values that have been found that the linear regression model does not effectively explain the data, and weather conditions can be assumed as a weak predictor for step counts and physical activity. Considering a more complex model or adding some extra features may lead for better predictions. If the aim of the project would lead to a more detailed analysis, accounting for other factors in the prepared dataset, could provide further insights.

# Discussion

The idea that can be put forward based on the obtained data and conducted analyses in the previous parts of the project is that there might be a correlation between weather conditions and walking distances. However, maybe due to the lack of robust data fully supporting this hypothesis, it cannot be entirely accepted. To accept this hypothesis more convincingly, further research encompassing a broader time frame, utilizing different datasets, and deepening the analysis with more detailed information would provide stronger conditions. In the meantime, it should be assumed, for the owner of the dataset (Sinan Özkan), there is not a certain relation between these **two factors.**

# Conclusion

In conclusion, the investigation into the impact of weather conditions on daily step counts has done by using an individual's step count data and local weather condition data. While some of the visualizations have suggested a connection between weather and the number of steps, some of the analyses have indicated that this relationship might be weaker than it was assumed. On the other hand, it is important to be aware of the limitations of the current study, including the reliance on a single individual's data (Sinan Özkan). Moreover, employing more people's data into the dataset, adding some extra factors, focusing on more complex weather conditions, adding some advanced modeling techniques might offer a better understanding of the complex interplay between environmental factors and step count. But according to the current dataset, it should be assumed that although the weather conditions seem to play a role in influencing daily step counts, a more comprehensive investigation is needed.

# References


**Bronfenbrenner, U. (1979). The ecology of human development: Experiments by nature and design. Harvard University Press.**

**Champion, V. L., & Skinner, C. S. (2008). The health belief model. In K. Glanz, B. K. Rimer, & K. Viswanath (Eds.), Health behavior: Theory, research, and practice (pp. 45-65). Jossey-Bass.**

**Engel, G. L. (1977). The need for a new medical model: A challenge for biomedicine. Science, 196(4286), 129-136.**

**Lozano, H., Park, S., Young, S., & Rozanova, L. (2021). Effects of weather on mood and physical activity in a real-life setting. PLoS ONE, 16(3)**

# Appendices

| sep=, | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ype,sourceName,sourceVersion,productTyp | e,device,startDate,endDate,unit,value | | | | | | | | |
| HKQuantityTypeIdentifierStepCount,Sinan | iPhone’u,16.3,"iPhone12,1","<<HKDevice: | 0x283371090>, | name:iPhone, | manufacturer:Apple | Inc., | model:iPhone, | hardware:iPhone12,1, | software:16.3>",2023-10-02 | 20:16:36 |
| HKQuantityTypeIdentifierStepCount,Sinan | iPhone’u,16.3,"iPhone12,1","<<HKDevice: | 0x283371090>, | name:iPhone, | manufacturer:Apple | Inc., | model:iPhone, | hardware:iPhone12,1, | software:16.3>",2023-10-02 | 20:27:27 |
| HKQuantityTypeIdentifierStepCount,Sinan | iPhone’u,16.3,"iPhone12,1","<<HKDevice: | 0x283371090>, | name:iPhone, | manufacturer:Apple | Inc., | model:iPhone, | hardware:iPhone12,1, | software:16.3>",2023-10-02 | 20:42:39 |
| HKQuantityTypeIdentifierStepCount,Sinan | iPhone’u,16.3,"iPhone12,1","<<HKDevice: | 0x283371090>, | name:iPhone, | manufacturer:Apple | Inc., | model:iPhone, | hardware:iPhone12,1, | software:16.3>",2023-10-03 | 06:05:44 |
| HKQuantityTypeIdentifierStepCount,Sinan | iPhone’u,16.3,"iPhone12,1","<<HKDevice: | 0x283371090>, | name:iPhone, | manufacturer:Apple | Inc., | model:iPhone, | hardware:iPhone12,1, | software:16.3>",2023-10-03 | 06:16:30 |
| HKQuantityTypeIdentifierStepCount,Sinan | iPhone’u,16.3,"iPhone12,1","<<HKDevice: | 0x283371090>, | name:iPhone, | manufacturer:Apple | Inc., | model:iPhone, | hardware:iPhone12,1, | software:16.3>",2023-10-03 | 06:26:33 |
| HKQuantityTypeIdentifierStepCount,Sinan | iPhone’u,16.3,"iPhone12,1","<<HKDevice: | 0x283371090>, | name:iPhone, | manufacturer:Apple | Inc., | model:iPhone, | hardware:iPhone12,1, | software:16.3>",2023-10-03 | 06:37:14 |
| HKQuantityTypeIdentifierStepCount,Sinan | iPhone’u,16.3,"iPhone12,1","<<HKDevice: | 0x283371090>, | name:iPhone, | manufacturer:Apple | Inc., | model:iPhone, | hardware:iPhone12,1, | software:16.3>",2023-10-03 | 07:00:10 |
| HKQuantityTypeIdentifierStepCount,Sinan | iPhone’u,16.3,"iPhone12,1","<<HKDevice: | 0x283371090>, | name:iPhone, | manufacturer:Apple | Inc., | model:iPhone, | hardware:iPhone12,1, | software:16.3>",2023-10-03 | 07:25:57 |
| HKQuantityTypeIdentifierStepCount,Sinan | iPhone’u,16.3,"iPhone12,1","<<HKDevice: | 0x283371090>, | name:iPhone, | manufacturer:Apple | Inc., | model:iPhone, | hardware:iPhone12,1, | software:16.3>",2023-10-03 | 07:38:05 |
| HKQuantityTypeIdentifierStepCount,Sinan | iPhone’u,16.3,"iPhone12,1","<<HKDevice: | 0x283371090>, | name:iPhone, | manufacturer:Apple | Inc., | model:iPhone, | hardware:iPhone12,1, | software:16.3>",2023-10-03 | 07:48:28 |
| HKQuantityTypeIdentifierStepCount,Sinan | iPhone’u,16.3,"iPhone12,1","<<HKDevice: | 0x283371090>, | name:iPhone, | manufacturer:Apple | Inc., | model:iPhone, | hardware:iPhone12,1, | software:16.3>",2023-10-03 | 09:26:52 |
| HKQuantityTypeIdentifierStepCount,Sinan | iPhone’u,16.3,"iPhone12,1","<<HKDevice: | 0x283371090>, | name:iPhone, | manufacturer:Apple | Inc., | model:iPhone, | hardware:iPhone12,1, | software:16.3>",2023-10-03 | 09:55:37 |
| HKQuantityTypeIdentifierStepCount,Sinan | iPhone’u,16.3,"iPhone12,1","<<HKDevice: | 0x283371090>, | name:iPhone, | manufacturer:Apple | Inc., | model:iPhone, | hardware:iPhone12,1, | software:16.3>",2023-10-03 | 10:15:41 |
| HKQuantityTypeIdentifierStepCount,Sinan | iPhone’u,16.3,"iPhone12,1","<<HKDevice: | 0x283371090>, | name:iPhone, | manufacturer:Apple | Inc., | model:iPhone, | hardware:iPhone12,1, | software:16.3>",2023-10-03 | 10:30:41 |
| HKQuantityTypeIdentifierStepCount,Sinan | iPhone’u,16.3,"iPhone12,1","<<HKDevice: | 0x283371090>, | name:iPhone, | manufacturer:Apple | Inc., | model:iPhone, | hardware:iPhone12,1, | software:16.3>",2023-10-03 | 10:55:35 |

(Appendix 1)

| name | datetime | tempmax | tempmin | temp | feelslikemax | feelslikemin | feelslike | dew | humidity | precip | precipprob | precipcover | preciptype | snow | snowdepth | windgust | windspeed | winddir | sealevelpressure |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Istanbul,Turkey | 2023-09-01 | 29.1 | 22.9 | 26.1 | 31.1 | 22.9 | 26.9 | 20 | 69.5 | 0 | 0 | 0 | | 0 | 0 | 38.2 | 19.7 | 18.4 | 1014.8 |
| Istanbul,Turkey | 2023-09-02 | 28.9 | 23.9 | 26.2 | 30.8 | 23.9 | 26.9 | 20.6 | 72.1 | 0 | 0 | 0 | | 0 | 0 | 44.3 | 29.7 | 39.1 | 1016.2 |
| Istanbul,Turkey | 2023-09-03 | 28.3 | 23.5 | 25.7 | 29.8 | 23.5 | 26.2 | 18.9 | 66.4 | 207 | 100 | 8.33 | rain | 0 | 0 | 45 | 29.7 | 33.1 | 1015.5 |
| Istanbul,Turkey | 2023-09-04 | 27.8 | 21.2 | 24.5 | 28.7 | 21.2 | 24.7 | 16.9 | 62.9 | 0 | 0 | 0 | | 0 | 0 | 43.6 | 25.5 | 47.1 | 1010.9 |
| Istanbul,Turkey | 2023-09-05 | 28 | 22.1 | 24.7 | 28.8 | 22.1 | 24.7 | 19.1 | 71.6 | 1.024 | 100 | 8.33 | rain | 0 | 0 | 44.6 | 37.1 | 46.4 | 1011.1 |
| Istanbul,Turkey | 2023-09-06 | 27.4 | 23.8 | 25.1 | 29.3 | 23.8 | 25.4 | 19.7 | 72.2 | 166 | 100 | 8.33 | rain | 0 | 0 | 51.1 | 36.5 | 46.7 | 1015.8 |
| Istanbul,Turkey | 2023-09-07 | 27.9 | 23.6 | 25.3 | 28.7 | 23.6 | 25.4 | 19.1 | 68.9 | 0 | 0 | 0 | | 0 | 0 | 57.6 | 38.7 | 48.7 | 1016.5 |
| Istanbul,Turkey | 2023-09-08 | 27.2 | 23 | 25 | 27.8 | 23 | 25 | 17.3 | 62.8 | 0 | 0 | 0 | | 0 | 0 | 49.7 | 29.7 | 45.1 | 1015.9 |
| Istanbul,Turkey | 2023-09-09 | 26.1 | 20.6 | 23.5 | 26.1 | 20.6 | 23.5 | 15.3 | 60.2 | 0 | 0 | 0 | | 0 | 0 | 44.3 | 29.3 | 44.8 | 1015.6 |
| Istanbul,Turkey | 2023-09-10 | 26.3 | 19.5 | 23.1 | 26.3 | 19.5 | 23.1 | 15.3 | 61.9 | 1.545 | 100 | 4.17 | rain | 0 | 0 | 55.4 | 30.9 | 52.8 | 1015.8 |
| Istanbul,Turkey | 2023-09-11 | 24.6 | 19.5 | 22 | 24.6 | 19.5 | 22 | 14 | 60.5 | 0 | 0 | 0 | | 0 | 0 | 42.8 | 27.6 | 47.9 | 1015.6 |
| Istanbul,Turkey | 2023-09-12 | 25.4 | 18.5 | 22.1 | 25.4 | 18.5 | 22.1 | 14.5 | 62.6 | 0 | 0 | 0 | | 0 | 0 | 39.2 | 22.9 | 54.3 | 1016.7 |
| Istanbul,Turkey | 2023-09-13 | 25.8 | 18.1 | 22.2 | 25.8 | 18.1 | 22.2 | 14.8 | 63.7 | 0 | 0 | 0 | | 0 | 0 | 33.8 | 17.4 | 50.8 | 1017.7 |
| Istanbul,Turkey | 2023-09-14 | 28.9 | 17 | 22.3 | 27.9 | 17 | 22.3 | 14.4 | 62.5 | 0 | 0 | 0 | | 0 | 0 | 30.2 | 15.4 | 43.2 | 1016.5 |
| Istanbul,Turkey | 2023-09-15 | 26 | 17 | 22 | 26 | 17 | 22 | 16.5 | 71.5 | 0 | 0 | 0 | | 0 | 0 | 34.2 | 18.1 | 33.6 | 1016.7 |
| Istanbul,Turkey | 2023-09-16 | 26.4 | 19.2 | 22.9 | 26.4 | 19.2 | 22.9 | 17.7 | 73.2 | 0 | 0 | 0 | | 0 | 0 | 42.8 | 25.1 | 49.1 | 1018.2 |
| Istanbul,Turkey | 2023-09-17 | 26.2 | 20.8 | 23.2 | 26.2 | 20.8 | 23.2 | 17.4 | 70.4 | 0 | 0 | 0 | | 0 | 0 | 42.8 | 26.8 | 47.3 | 1018.5 |
| Istanbul,Turkey | 2023-09-18 | 24.7 | 19.6 | 22.1 | 24.7 | 19.6 | 22.1 | 14.1 | 60.7 | 0 | 0 | 0 | | 0 | 0 | 59.4 | 30 | 55.6 | 1017.5 |
| Istanbul,Turkey | 2023-09-19 | 26 | 19.4 | 22.5 | 26 | 19.4 | 22.5 | 14.7 | 61.6 | 0 | 0 | 0 | | 0 | 0 | 39.2 | 29.5 | 47 | 1015.9 |

(Appendix 2)

filtered_step_count

| startDate | value |
|---|---|
| 2023-10-02 | 621.0 |
| 2023-10-03 | 11309.0 |
| 2023-10-04 | 5580.0 |
| 2023-10-05 | 10851.0 |
| 2023-10-06 | 15525.0 |
| 2023-10-07 | 1774.0 |
| 2023-10-08 | 4058.0 |
| 2023-10-09 | 12359.0 |
| 2023-10-10 | 5734.0 |
| 2023-10-11 | 10102.0 |
| 2023-10-12 | 7065.0 |
| 2023-10-13 | 3166.0 |
| 2023-10-14 | 2826.0 |
| 2023-10-15 | 3735.0 |
| 2023-10-16 | 7140.0 |
| 2023-10-17 | 6035.0 |
| 2023-10-18 | 10689.0 |

filtered_weather_data

| datetime | conditions |
|---|---|
| 2023-09-01 | Partially cloudy |
| 2023-09-02 | Partially cloudy |
| 2023-09-03 | Rain, Partially cloudy |
| 2023-09-04 | Partially cloudy |
| 2023-09-05 | Rain, Partially cloudy |
| 2023-09-06 | Rain, Partially cloudy |
| 2023-09-07 | Partially cloudy |
| 2023-09-08 | Partially cloudy |
| 2023-09-09 | Partially cloudy |
| 2023-09-10 | Rain, Partially cloudy |
| 2023-09-11 | Partially cloudy |
| 2023-09-12 | Partially cloudy |
| 2023-09-13 | Clear |
| 2023-09-14 | Clear |
| 2023-09-15 | Clear |
| 2023-09-16 | Partially cloudy |
| 2023-09-17 | Partially cloudy |

(Appendix 3)