

CS210- Introduction to Data Science- Course Project

Table of Contents

CS210- Introduction to Data Science- Course Project	1
Introduction – Motivation	2
Dataset Description	3
Data Analysis	5
Conclusion	11

Introduction – Motivation

From the Course CS 210, I am only one of the students who are deeply interested in understanding personal activity patterns and how they differentiate across different phases and locations of my life, I embarked on an exciting journey of data exploration and statistical analysis. The primary focus of this analysis was to examine my step count data, which the data is collected from my Iphone's Health Data, which is only used by myself. The data includes various periods, including my Erasmus term in Hamburg, Germany, Normal daily life in Istanbul and holidays such as Copenhagen, Milano, Amsterdam, and Zadar locations. I will come to analysis later in this report.

The motivation behind this project arises from a blend of curiosity and a drive for self-improvement. By analyzing my step count data, I aimed to gain insights into my physical activity patterns during different times and locations. This understanding is crucial as it not only reflects my lifestyle and mobility but also helps in making informed decisions regarding health and time management and furthermore, it can give me insights for "Which city is more tiring for visiting?"

The analysis began with the collection and organization of step count data, from my Iphone Health Data, and I have converted that file from xml to csv. Moreover, the data was then categorized into different samples corresponding to specific time periods and locations. Each sample was subjected to a thorough statistical examination, including calculating descriptive statistics to understand the central tendencies and variabilities of my daily activities for each location and occasion.

To deepen the analysis, I performed hypothesis testing and ANOVA (Analysis of Variance) tests. These tests were crucial in comparing the mean step counts across different periods, providing a statistical basis to understand whether my activity levels significantly varied during these times.

The project not only offered valuable personal insights but also served as an excellent opportunity to apply statistical concepts and data analysis techniques in a real-world context. The findings from this analysis are helped and guided me in making lifestyle adjustments and in setting more tailored and achievable physical activity goals and give me the power of knowledge on my activity levels in different occasions.

Dataset Description

The dataset used for this analysis generally consists of count of my steps data historically. This dataset represents a rich source of information about my daily physical activity, quantified in the number of steps taken each day. The data was sourced from my Iphone's health application, ensuring a high level of accuracy and reliability. The structure of the dataset that I have used is as following.

- 1) Date and Time: Each entry in the dataset includes a timestamp indicating when the steps were recorded. This granularity allowed for detailed analysis of daily and even hourly activity patterns. I will be using them mostly "daily"
- 2) Step Count: This is the primary target of this project and dataset, it represents the number of steps taken by me during the recorded intervals. These values vary significantly, reflecting the level of physical activity on each day varying due to many features such as weather, location, busyness, holidays etc. I will be focusing on the holidays mostly.

- Sample Segmentation:

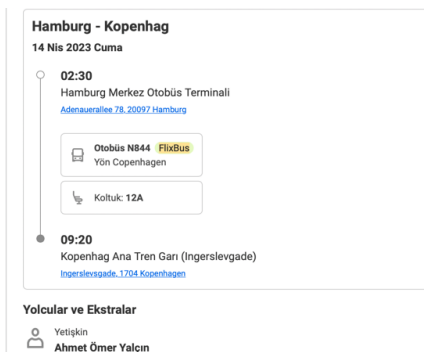
The dataset was divided into several samples as previously mentioned, each corresponding to a specific time frame and location. For the verifying manner, I will be also providing my tickets in my holidays as well!

These samples include:









Erasmus Period: This sample comprises 134 days of my Erasmus period, between 9th of April and 20th of August 2023. it shows relatively higher activity levels than my daily life back in İstanbul with a significant number of steps per day.

Istanbul Period: This sample includes again 134 days for the sake of the statistical measures and calculations I will run with using these two main samples. I have fairly chosen the period between 26th of December 2022 and 8th of April 2023. Indicated a different pattern in the analysis, possibly reflecting a more routine or sedentary lifestyle compared to the Erasmus period data.











Copenhagen Period: This subset corresponds to a short, specific timeframe, giving a snapshot of my activities during my stay in Copenhagen. (Me in Copenhagen, 😊)



Milano Period: Reflecting the days spent in Milano, it shows the variation in my physical activity in a different holiday, the dates can be seen below.

Your past flights status	Your past flights status
 Hamburg to Milan Bergamo 03 May 2023 • 08:25 - 10:15 • FR3550 	 Milan Bergamo to Hamburg 10 May 2023 • 06:10 - 07:55 • FR3549 
Passengers & Products	Passengers & Products
 Ahmet Ömer Yalçın	 Ahmet Ömer Yalçın
FLIGHT PRODUCTS 	FLIGHT PRODUCTS 

Amsterdam Period: This part of the dataset encompasses time spent in Amsterdam, offering a perspective on my activity patterns in a vacation in Amsterdam, the dates are below.

Münster - Amsterdam	Amsterdam - Hamburg
30 Haz 2023 Cuma	02 Tem 2023 Pazar
 09:55 Münster Ana Tren Garı (Hafenstraße) Hafenstraße 34, 48153 Münster  Otobüs N69  Yön Schiphol Havalimanı  Koltuk: 2D 13:45 Amsterdam Sloterdijk Piaarcooplein, 1043 DW Amsterdam	 20:55 Amsterdam Sloterdijk Piaarcooplein, 1043 DW Amsterdam  Otobüs N844  Yön Copenhagen  Koltuk: 25C 3 Tem 02:20 Hamburg Merkez Otobüs Terminali Adenauerallee 78, 20097 Hamburg
Yolcular ve Ekstralar	Yolcular ve Ekstralar
 Yetişkin Ahmet Ömer Yalçın	 Yetişkin Ahmet Ömer Yalçın

Zadar Period: Representing a shorter duration, it gives an insight into my activities during a brief visit to Zadar, which was more in the middle of the summer, dates can be seen below.

RYANAIR NON-PRIORITY	RYANAIR NON-PRIORITY
Boarding pass TURKEY • P/U2-2987 • S	Boarding pass TURKEY • P/U2-2987 • S
NON-PRIORITY Q HAM - ZAD FR8462 AHMET OMER YALCIN	NON-PRIORITY Q ZAD - FMO FR1851 AHMET OMER YALCIN
Boarding Front Seat 08E Booking ref MEGVXV Seq 66	Boarding Back Seat 22C Booking ref UM4D7N Seq 96
Hamburg T2 Date 10. Aug 2023 Gate Closes 08:45	Zadar Date 15. Aug 2023 Gate Closes 18:00
Zadar Date 10. Aug 2023 Gate Closes 08:45	Munster Date 15. Aug 2023 Gate Closes 18:30

Each of these samples was analyzed to extract descriptive statistics such as mean, median, standard deviation, minimum, and maximum step counts. These statistics provided an overview of my activity levels, highlighting days with unusually high or low activity.

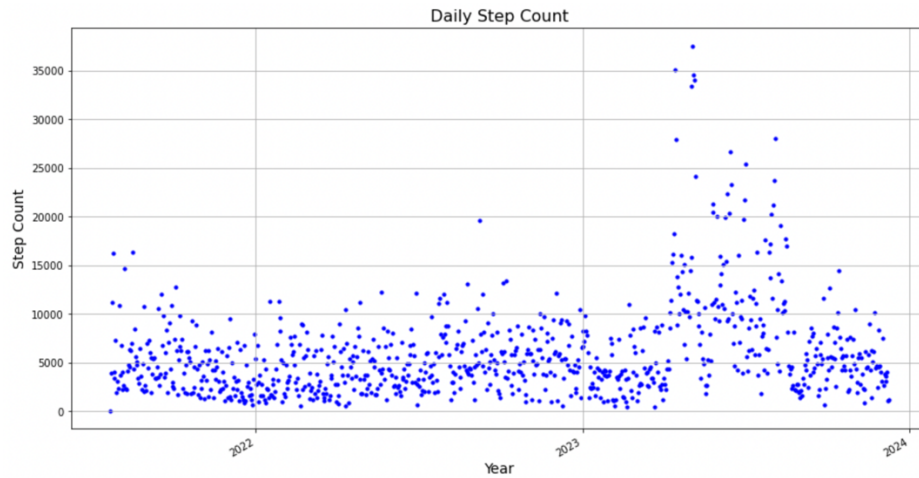
For instance, a day with an exceptionally high step count might indicate extensive exploration or physical activities, while a day with a very low count could suggest a more sedentary routine. By comparing these statistics across different samples, I could observe how my activity levels changed with location and time.

Data Analysis

The primary goal of the analysis was to explore and understand the variations in my daily step counts across different time periods and locations. The analysis involved several stages, each focusing on a specific aspect of the data.

For each sample previously mentioned (Erasmus, Istanbul, Copenhagen, Milano, Amsterdam, Zadar), I have calculated descriptive statistics, including mean, median, standard deviation, minimum, and maximum step counts. These statistics provided a foundational understanding of my activity patterns.

For the starting, I have calculated descriptive statistics of the whole data, and



Which is showing my daily step data, from starting the time I bought my phone until today. I had the following descriptive statistics:

Number of days: 870

Mean: 5923

Standard Deviation: 4747

Min: 65

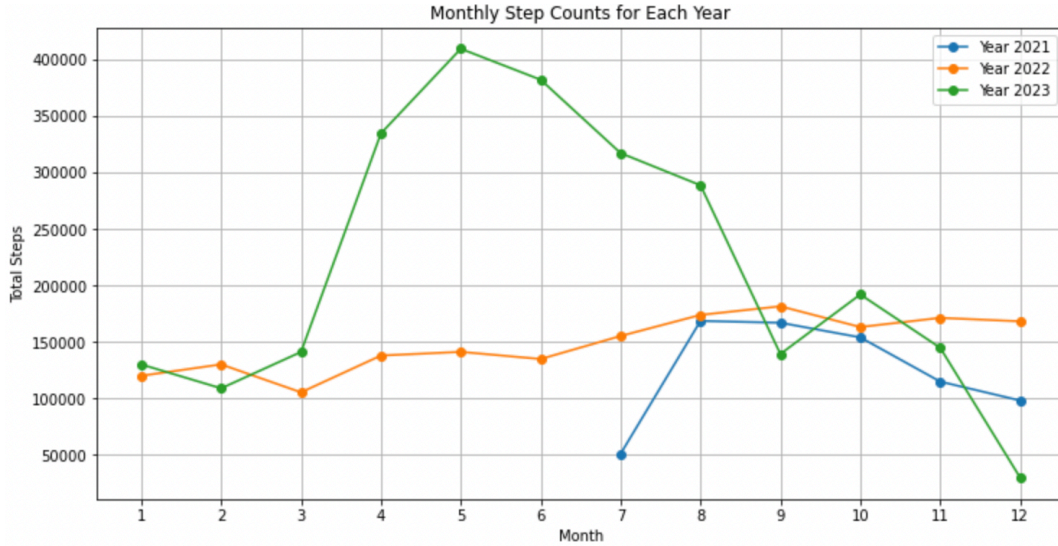
Max: 37574

Furthermore, I wanted to learn what are the dates that I had my top performance each year.

	date	value	year	month
0	2021-08-17	16348	2021	8
1	2021-07-26	16219	2021	7
2	2021-08-08	14673	2021	8
3	2021-10-04	12740	2021	10
4	2021-09-18	12023	2021	9
5	2021-07-25	11207	2021	7
6	2021-08-02	10910	2021	8
7	2021-09-29	10872	2021	9
8	2021-08-29	10816	2021	8
9	2021-09-15	10508	2021	9
10	2022-09-08	19635	2022	9
11	2022-10-08	13443	2022	10
12	2022-10-04	13178	2022	10
13	2022-08-25	13032	2022	8
14	2022-05-21	12193	2022	5
15	2022-12-03	12124	2022	12
16	2022-06-29	12108	2022	6
17	2022-07-30	12074	2022	7
18	2022-09-11	12057	2022	9
19	2022-07-26	11607	2022	7
20	2023-05-04	37574	2023	5
21	2023-04-14	35109	2023	4
22	2023-05-05	34608	2023	5
23	2023-05-06	34049	2023	5
24	2023-05-03	33397	2023	5
25	2023-08-04	28038	2023	8
26	2023-04-15	27914	2023	4
27	2023-06-15	26620	2023	6
28	2023-07-02	25359	2023	7
29	2023-05-07	24136	2023	5

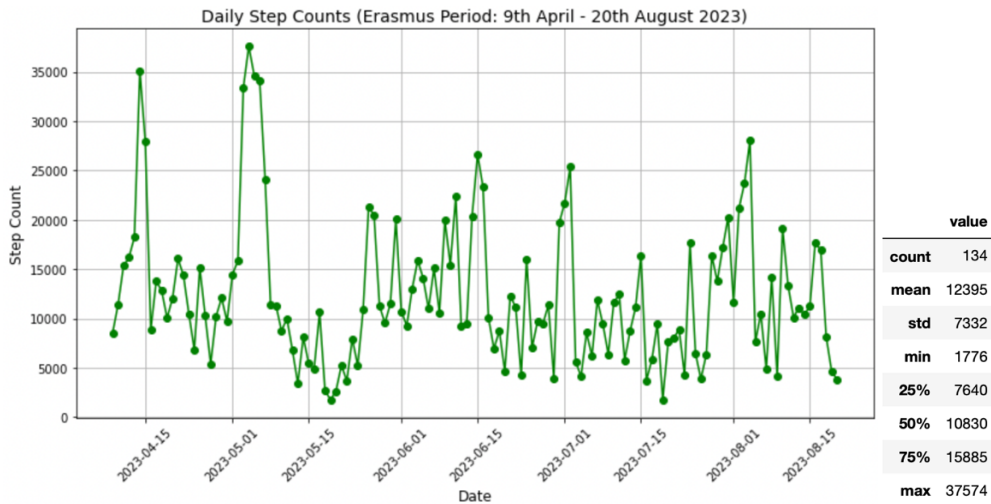
The values above are corresponding to the maximum top 10 values each year, they are mostly in the summer for every year.

Furthermore, wanted to go month specific, to see if there are patterns or not in my data more clearly, it can be observed from the below graph that, my values are similar to each other outside the Erasmus period, than this derived me to curiosity of these period's comparison.



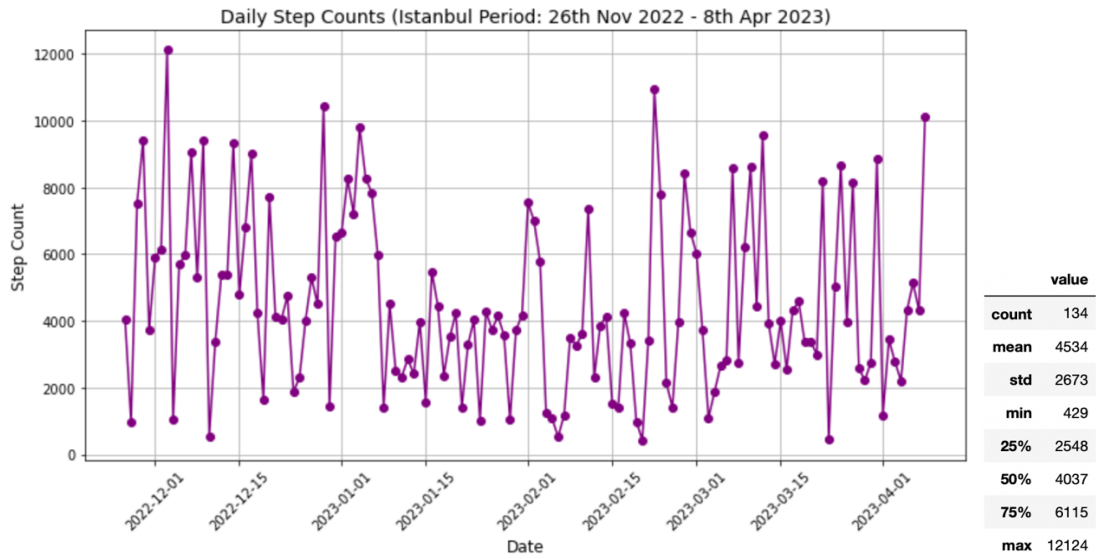
- **Sample Analysis**

As the previous graph made me curious about the comparison of my Erasmus period and normal daily life in İstanbul, I have created samples which I have previously explained, and worked on those samples.



The graph above shows the line chart of my daily step count in my Erasmus period and descriptive statistics of the sample, it includes max value of my datasets, which means I had my top performance in my Erasmus period, which falls in my Milano vacation.

Moreover, I have also did the same calculations on my İstanbul sample, and it resulted as follows.



Descriptive statistics above are for the period of my daily life in İstanbul for the same sample size with the Erasmus period.

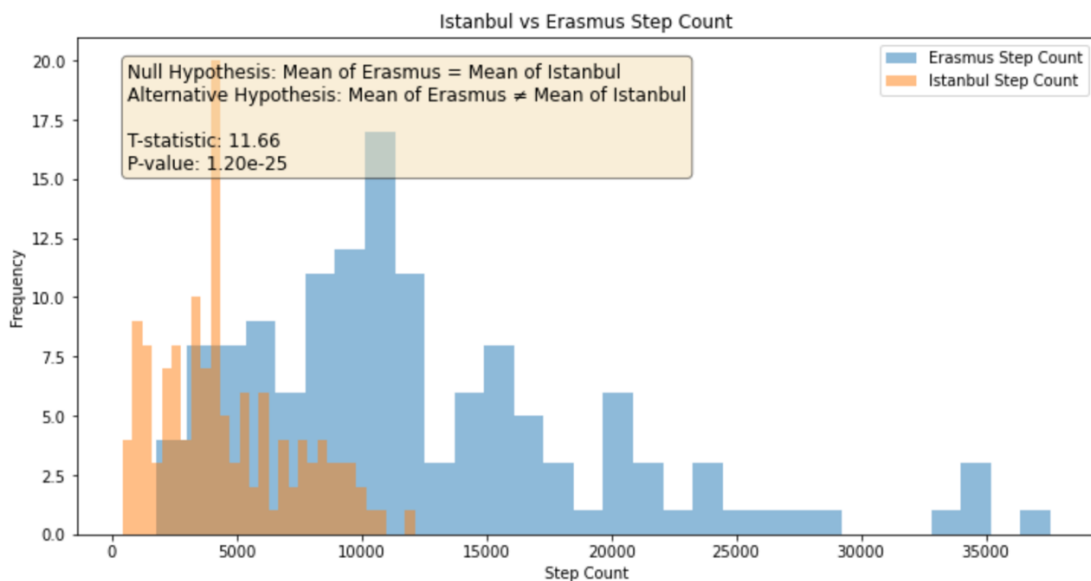
- **2-Sample T-TEST:**

Then, for the comparison of these two samples and statistically proving there is a difference between their means, I have conducted a 2-Sample t-test, with the following hypothesis.

Null Hypothesis (H0): The mean of step count during the Erasmus period equals the mean step count during the İstanbul period.

Alternative Hypothesis (H1): The mean step count during the Erasmus period is not equal to the mean step count during the İstanbul period.

The following calculation resulted as the below graph.



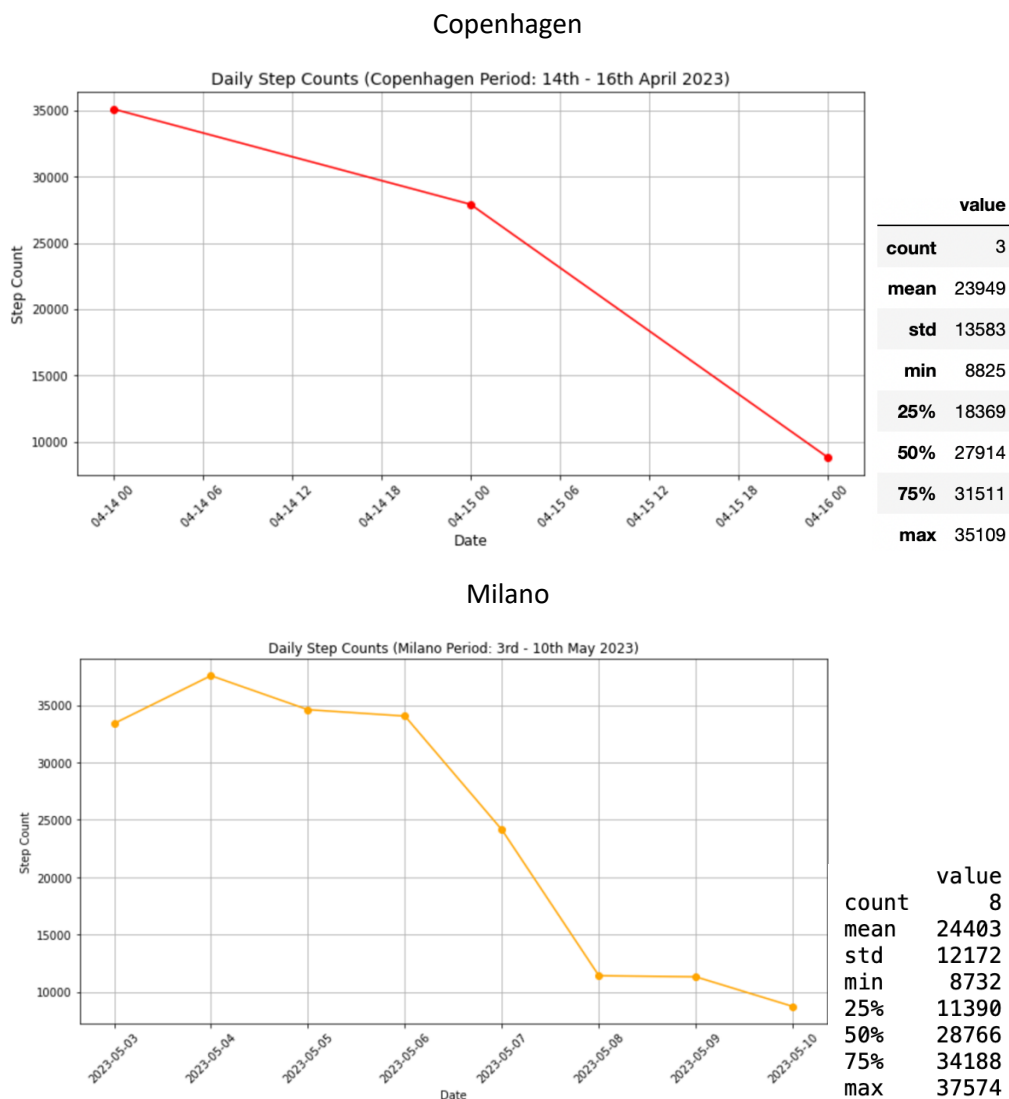
(11.65882534602826, 1.1992331895738105e-25)

The t-test results showed a significantly low p-value, leading to the rejection of the null hypothesis as the P value is below our threshold of 0,05. **This indicates a statistically significant difference in the mean step counts between these two periods**, so, P value is corresponding to **the risk of stating** mean values of both periods to be different, as the P value is too low, we state it is risk that we can take, and we accept the alternative hypothesis.

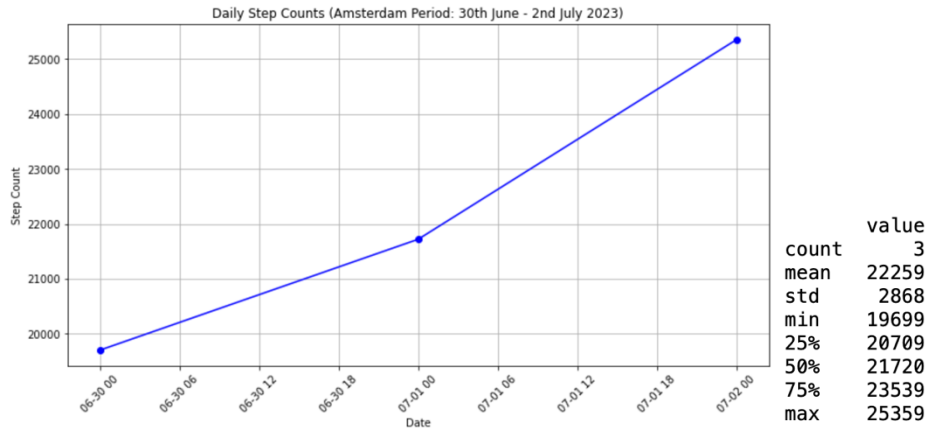
- **Holiday Comparison - ANOVA**

Furthermore, I wanted to derive another test on my dataset, and created new samples which are including the count of steps that I had in specific time periods. The periods are the vacations that I went to in my Erasmus period, and I have explained those variables previously. With this analysis, I wanted to inspect on how a location changes my walking habits and which city is a no go if you are looking for a chilling vacation without burning too much calory by walking.

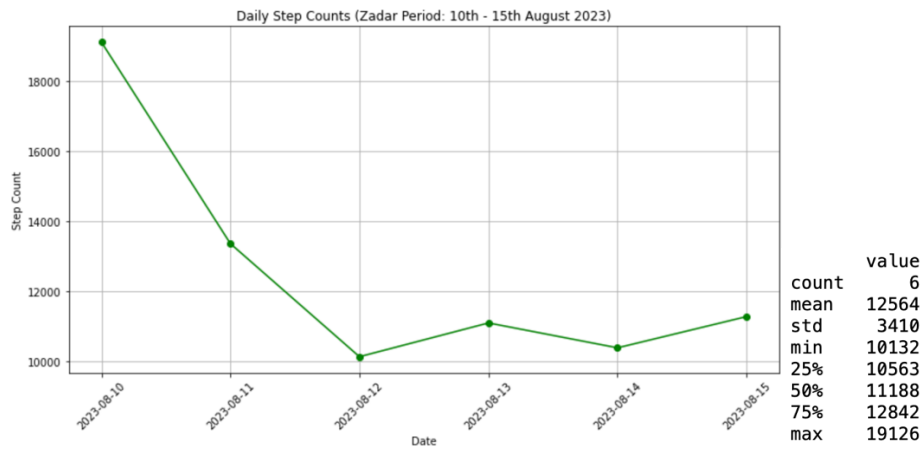
Individual Graphs:



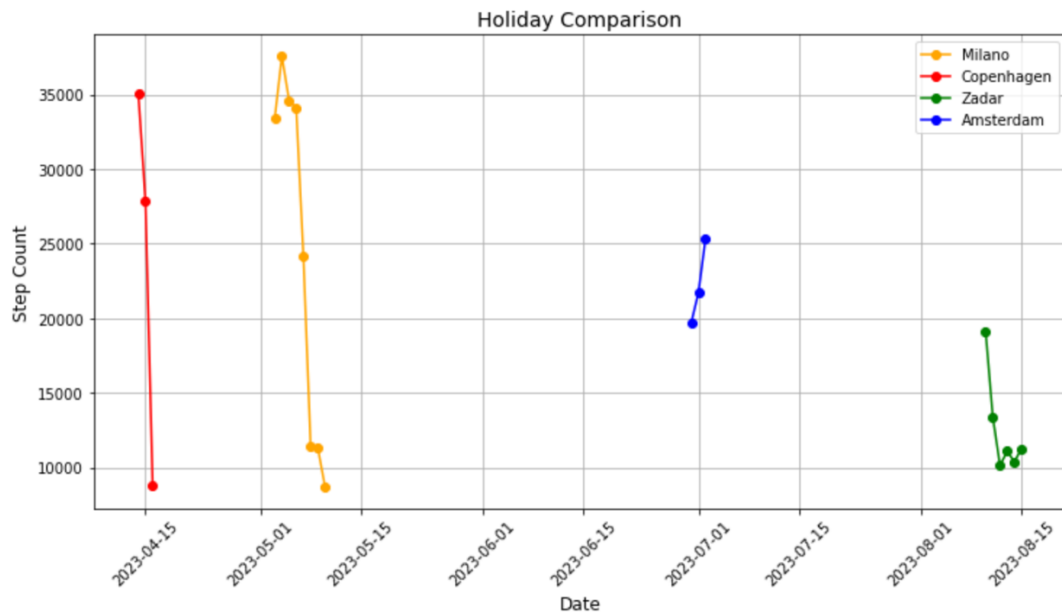
Amsterdam



Zadar



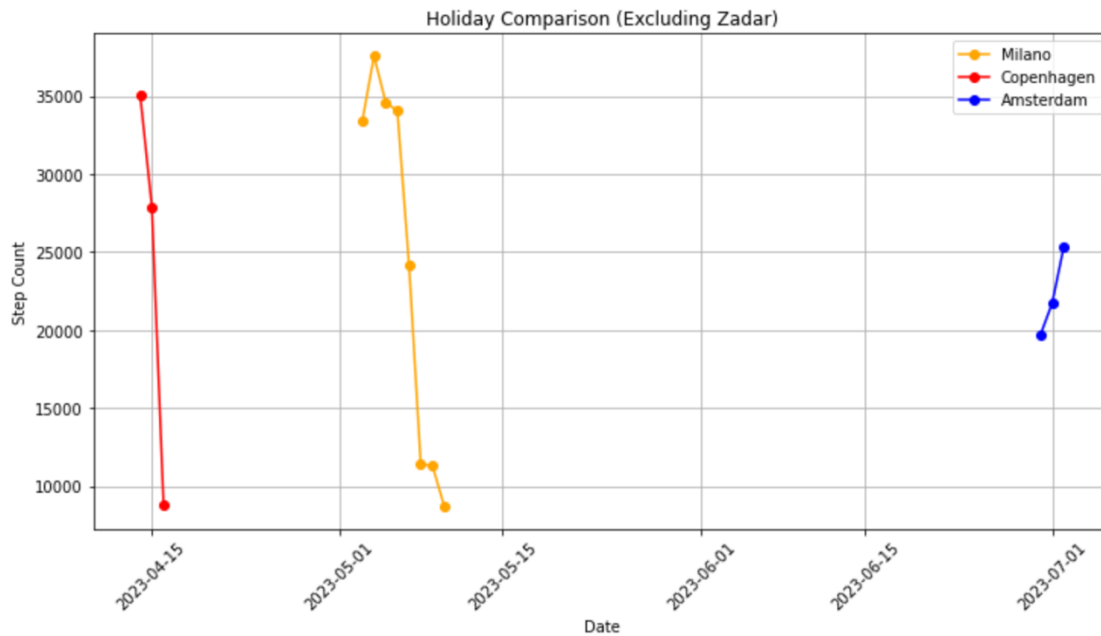
To compare the mean step counts across all periods, I have conducted a ANOVA test this time, and it resulted as;



F_onewayResult(statistic=1.9619876989427987, pvalue=0.1604004938270146)

P value of the ANOVA, which is 0.16, indicates that H_0 can be accepted, ($0.16 > 0.05$), which corresponds to accept that they share the same mean values. In explanation, the statistics proves that in every holiday, I have the same mean for the count of steps.

But furthermore, as I was trying to learn which holiday was less tiring, I have realized that Zadar seems to have a less count of steps compared to others, so I have excluded Zadar and conducted a new ANOVA test, which resulted as,



ANOVA Result (Excluding Zadar): F-statistic = 0.03893630374893996, P-value = 0.961943916051738

The P value significantly increased to 0.96, which is a very strong indicator that Milano, Copenhagen and Amsterdam vacations were sharing the same means, and this leads me to conclude that Zadar was the vacation where I had more chill and relaxed time.

Eventually, the main conclusion that I have ended up is, you can go to Zadar for a chilling vacation proven statistically. Furthermore, the statistical analysis, supported by both hypothesis testing and ANOVA, revealed significant differences in my physical activity across different periods and locations. This suggests that my daily step counts are influenced by the environment, lifestyle changes, and possibly other factors like weather, social engagements, and academic or work schedules.

Conclusion

Finally, for a total overview of this report, I have conducted some tests over my step count data from my Iphone's health application, these insights not only illuminate patterns in physical activity but also reflect lifestyle adaptations to different environments. Findings can be listed as:

- 1) Variability in Activity Levels: One of the important findings is the significant variability in daily step counts across different periods. Each location and time frame, such as Erasmus, Istanbul, Copenhagen, Milano, Amsterdam, and Zadar, exhibited distinct activity patterns. This variability indicates that my physical activity levels are sensitive to environmental and situational changes.
- 2) Similar Patterns out of Erasmus Period: In the first graphs I have provided, the daily life values seems to be similar and a overlapping graph for the months such as September, October and November for both 2021 & 2022, this seems to be explaining the school factor and the same busy calendars in different time periods are resulting in the same level of physical activity
- 3) Erasmus vs. Istanbul Comparison: The two-sample t-test between the Erasmus and Istanbul step count data showed a significant difference in mean step counts. And with the lower P value from the 0,05 threshold, the HA is accepted. This indicates that my activity levels were substantially different in these two locations, with the Erasmus period generally showing higher activity levels. Additionally, the statistical rejection of the null hypothesis in this comparison highlights how my lifestyle and daily activities varied considerably between an exchange program environment and my time in Istanbul.
- 4) Holiday Comparison: The ANOVA tests conducted for all holidays that I went to in 2 different scenarios (including and excluding Zadar) reinforced the finding of significant differences in mean step counts across various time frames and locations. This further substantiates the idea that my physical activity levels were not consistent but rather influenced by the specific circumstances of each period and holidays plan.
- 5) Impact of Specific Locations: Locations like Copenhagen, Milano, Amsterdam, and Zadar, each associated with specific dates, showed unique activity trends. These variations could be attributed to factors like the purpose of the visit (e.g., tourism, study, work), the geographical layout of the location, cultural factors, and personal preferences or habits during the stay. This can be a subject for a further analysis as the dataset that I was working with was my own data, it did not allow me to compare the results in these manner.

- 6) General Lifestyle Insights: The analysis provided a quantitative backing to subjective experiences, transforming my understanding of how active I was during different phases. Beyond the statistical results, this analysis served as a catalyst for self-reflection. It encouraged a data-driven approach to understanding my habits and lifestyle, potentially guiding future decisions regarding health, time management, and even choices about travel and living locations in the way that I want to do them, as I have emphasized in the last paragraph on my introduction part.

In conclusion, the findings from this research offer an overall view of my physical activity patterns. They underline the complex interplay between environment, lifestyle, and personal habits in shaping daily activity levels and our power as data enthusiasts to use a basic daily data which can be reached from our phone can derive us too many conclusions and reflections about our lives and living habits. These insights not only enrich personal understanding but also have wider implications for health and lifestyle optimization.