**CS210 Project Report**

**Title: Analyzing the Relationship Between Step Counts and Temperature**

**Mohamad Samer Bader  
30437**

**Introduction**

This project explores the relationship between daily step counts and weather conditions, particularly temperature. The hypothesis is that higher temperatures correlate with increased physical activity. By analyzing personal step data alongside weather data, the goal is to uncover patterns and insights into how environmental factors influence daily movement.

**Motivation**

This project stems from a curiosity to better understand how environmental factors, such as weather and temperature, influence daily physical activity. Step counts serve as a key indicator of health and well-being, and exploring the relationship between weather conditions and activity levels provides insights into behavioral patterns. By identifying how weather impacts movement, I aim to improve my personal fitness habits and plan activities more effectively. Additionally, this analysis contributes to a broader understanding of how environmental factors influence lifestyle choices, which could have implications for public health recommendations.

**Data Overview**

The analysis utilizes two datasets:

1. **Step Count Data** – Daily recorded step counts spanning 827 days, including the date and day of the week.
2. **Weather Data** – Daily temperature, and weather descriptions recorded for Istanbul, Beirut and Antalya, covering the same timeframe as the step data.

**Methodology**

* Health app data were exported from my iphone’s health app.
* Step counts were, then, extracted and separated from the rest of the health app data using a python code.
* Weather data, depending on the city I was in at the time, was extracted using “Visual Crossing Weather” API, where I extracted 24 temperature recordings per day, for each day, and then took the average temperature of that day.
* Weather data was merged with the step count data based on matching dates.
* Visualizations and statistical analyses were conducted to identify potential correlations between temperature and step counts.

**Preliminary Results**

Initial descriptive statistics of step count data reveal:

* **Mean Daily Steps:** 9044
* **Median Daily Steps:** 8447
* **Standard Deviation:** 4338
* **Minimum Steps:** 28 (after further investigation, I noticed that this was the day I bought my new phone)
* **Maximum Steps:** 28461
* **Interquartile Range (25%-75%):** 5988 - 11644

These statistics indicate a broad distribution of physical activity, suggesting the presence of both highly active and less active days. The next stage of the analysis investigates how these patterns align with temperature variations.

**Visualization of Step Counts:** The histogram below illustrates the distribution of daily step counts over the recorded period.

**A diagram of a step

Description automatically generatedAnalysis:**

* The histogram indicates a right-skewed distribution, with most daily step counts clustering between 5000 and 12000 steps.
* The peak frequency falls around 8000-10000 steps, reflecting a typical daily activity level.
* There are notable outliers with step counts exceeding 20000, which could correspond to days involving higher physical activity or outdoor events.
* The presence of low-step days suggests occasional inactivity, potentially influenced by weather or personal circumstances.

**Time Series Analysis:** The plot below shows the trend of daily step counts over time.

**A graph showing a blue line

Description automatically generatedAnalysis:**

* The time series plot reveals fluctuations in daily step counts, with periods of high activity interspersed with days of lower step counts.
* Several spikes exceeding 20000 steps indicate occasional bursts of high physical activity.
* A gradual decline in average step counts appears noticeable towards the later months of the year, potentially suggesting seasonal changes, fatigue, or other influencing factors.
* Periods of low activity might coincide with colder months or adverse weather conditions, reinforcing the hypothesis of weather impacting physical activity.

**Average Steps by Day of the Week:** The bar plot below shows the average step count for each day of the week.

**A graph of different colored bars

Description automatically generatedAnalysis:**

* The highest average step counts occur on Fridays, while Wednesdays have the lowest average.
* Weekends (Saturday and Sunday) show relatively high activity levels, suggesting more physical activity during leisure days.
* The variation in step counts throughout the week may reflect differences in routine, with midweek dips likely influenced by work or study schedules.

**Daily Steps with 7-Day Rolling Average:** The plot below shows daily step counts alongside a 7-day rolling average.

**A graph showing a number of red and blue lines

Description automatically generatedAnalysis:**

* The rolling average smooths out short-term fluctuations, revealing underlying trends in physical activity.
* Periods of sustained high activity are evident, aligning with seasons or lifestyle patterns.
* The rolling average highlights seasonal dips, suggesting lower activity levels during certain months, potentially due to colder weather or less outdoor activity.
* This visualization emphasizes long-term patterns that may correlate with external factors like weather conditions or holidays.

**Outlier Analysis:** Outliers were identified by defining thresholds based on the 5th and 95th percentiles of daily step counts.

* **Low Threshold (5th percentile):** Days with fewer than approximately 2600 steps.
* **High Threshold (95th percentile):** Days with more than approximately 18900 steps.

Key Observations:

* A total of 80 days were identified as outliers.
* Outlier days with extremely high step counts (e.g., exceeding 22000) often occur on weekends or holidays, suggesting increased outdoor activities.
* Extremely low step count days (e.g., below 1000) were primarily linked to periods of illness or examination days, reflecting reduced physical activity due to external commitments. (I made this conclusion after checking what I was doing on these days through my gallery app)
* The distribution of outliers provides insight into the variability of physical activity and potential influencing factors.

**Data Merging Process:**

* Both datasets were aligned by converting the date columns to datetime format.
* An inner merge was performed to ensure that only days with both step count and weather data were included.
* The merged dataset includes variables such as temperature, precipitation, humidity, and weather descriptions.

**Snapshot of Merged Data:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Date** | **Day** | **Steps** | **Temperature** | **Precipitation** | **Humidity** | **Description** |
| 2022-09-18 | Sunday | 6478 | 73.9 | 0.006 | 63.0 | Rain, Partially cloudy |
| 2022-09-19 | Monday | 9078 | 67.7 | 0.000 | 54.0 | Partially cloudy |
| 2022-09-20 | Tuesday | 5593 | 65.2 | 0.002 | 66.6 | Rain, Partially cloudy |
| 2022-09-21 | Wednesday | 7347 | 64.9 | 0.005 | 58.6 | Rain, Partially cloudy |
| 2022-09-22 | Thursday | 6275 | 58.4 | 0.738 | 78.7 | Rain, Partially cloudy |

**Analysis:**

* The merged dataset allows for comprehensive analysis by directly comparing step counts to daily weather metrics.
* Initial observations indicate that rainy days with higher humidity might correlate with lower step counts, suggesting weather may play a significant role in daily activity.
* Further analysis will involve statistical tests and visualizations to explore these relationships in depth.

**Average Steps by Temperature Range and Weather Description:**

* **By Temperature Range:**
  + Low (below 50°F): (7999.43 steps) Lower step counts indicate reduced physical activity during colder temperatures.
  + Moderate (50°F - 70°F): (8567.92 steps) Increased step counts suggest that moderate weather encourages outdoor activities.
  + High (70°F - 90°F): (10392.92 steps) The highest average activity reflects comfortable weather conditions for physical exertion.
  + Very High (above 90°F): NaN, this was excluded because the temperature never reaches this much where I live.
* **By Weather Description:**
  + Clear (9920 steps): Clear weather fosters higher activity levels.
  + Partially cloudy (9694 steps): Physical activity remains high, though slightly lower than clear conditions.
  + Rain, Overcast (6511 steps): The lowest step counts highlight the adverse effect of rain and overcast skies on physical activity.
  + Snow, Rain, Overcast (5766 steps): Snow and rain further reduce activity levels, likely due to discomfort and unsafe walking conditions.
  + Snow, Rain, Partially cloudy (8874 steps): Mixed conditions show higher activity compared to overcast weather but still below clear days.

**Scatter Plot Analysis: Steps vs Temperature**

**A graph with different colored dots

Description automatically generated**

* **General Trend:**
  + The scatter plot shows a positive relationship between temperature and step counts. As temperatures rise from 40°F to around 80°F, step counts tend to increase, peaking between 60°F and 80°F.
  + Beyond 80°F, step counts slightly decline, suggesting extreme heat may reduce activity.
* **Weather Impact:**
  + Clear and Partially Cloudy Days: Higher step counts, clustered towards the upper range.
  + Rain and Overcast: Lower step counts, particularly below 60°F.
  + Snow and Rain Mix: Consistently low step counts, reinforcing the idea that adverse weather discourages activity.
* **Distribution Insights:**
  + Significant variability in step counts at similar temperatures suggests factors beyond weather influence activity.
  + Highest step counts (20000+) occur between 60°F and 80°F under clear or partially cloudy skies.
* **Outliers:**
  + Exceptional activity days (25000+ steps) in moderate temperatures suggest events or outdoor activities.
  + Low-temperature days with high steps indicate unavoidable outdoor movement.

**Boxplot Analysis: Steps by Weather Description**

**A diagram of a graph

Description automatically generated with medium confidence**

* **Clear and Partially Cloudy:**
  + Highest median step counts with wide distribution. Outliers above 20000 steps indicate high activity days.
* **Rain and Overcast:**
  + Lower median step counts and less variability. Fewer high-activity outliers.
* **Overcast and Mixed Conditions:**
  + Lowest median step counts with minimal variation. Consistent low activity levels.
* **Key Insight:**
  + Clear weather leads to higher activity, while rain and overcast conditions reduce step counts. Occasional high outliers reflect non-weather influences.

**Correlation Analysis: Temperature vs Step Counts**

* **Correlation Coefficient: 0.19**
  + The correlation between temperature and step counts is 0.19, indicating a weak positive relationship.
  + This suggests that while higher temperatures are associated with slightly higher step counts, the effect is minimal.
* **Interpretation:**
  + The weak correlation implies that temperature alone is not a strong predictor of physical activity. Other factors, such as personal schedules, weather conditions (rain, snow), or lifestyle, likely play a more significant role.
  + The variability observed in the scatter plot further supports this, showing a broad spread of step counts across similar temperatures.

**Analysis: Regression Line for Steps vs Temperature**

**A diagram of a graph

Description automatically generated**

* **Trend Observation:**
  + The scatter plot with a regression line shows a slight upward slope, indicating a positive but weak relationship between temperature and step counts.
  + The spread of data points around the regression line highlights significant variability, reinforcing that while temperature may influence activity, it is not the sole determining factor.
* **Slope Interpretation:**
  + The gentle slope suggests that increases in temperature lead to marginal increases in step counts. However, the impact is minimal, aligning with the previously calculated correlation of 0.19.
  + Clusters of points above and below the regression line reflect days where step counts diverged from the expected trend, likely influenced by weather conditions, personal habits, or external events.

**Conclusion**

* **Hypothesis: "Higher temperatures are associated with increased physical activity (step counts)."**
* **Result:**
  + The analysis reveals a weak positive correlation (0.19) between temperature and step counts. While there is a slight upward trend, the relationship is not strong enough to conclude that temperature is a significant driver of physical activity.
  + Boxplots and scatter plots consistently show greater step counts in favorable weather (clear or partially cloudy conditions), suggesting that weather descriptions play a larger role than temperature alone.
  + Regression analysis and visualizations confirm that activity patterns are highly variable, indicating that personal schedules, lifestyle factors, and precipitation may have more substantial effects on daily step counts.
* **Decision on Hypothesis:**
  + Reject the hypothesis that higher temperatures alone significantly increase physical activity.
  + Accept a modified hypothesis: Favorable weather conditions (clear skies, moderate temperatures) encourage higher physical activity, but temperature alone is not a strong predictor of step counts.

**Limitations and Future Work**

**While this project provides valuable insights, there are some limitations that could be addressed in future work:**

* Limited Scope of Data: The analysis focuses solely on step counts and basic weather parameters like temperature and weather descriptions. Including additional variables, such as wind speed, air quality, or sunshine hours, could enhance the understanding of how weather impacts physical activity.
* Activity Context: The step count data does not differentiate between different types of activities (e.g., walking, running, indoor vs. outdoor). Incorporating contextual data, such as activity type or purpose, would provide a more nuanced view of how weather influences behavior.
* Seasonal Trends: While temperature and weather descriptions were analyzed, seasonal variations were not explicitly explored. Future work could focus on comparing activity levels across seasons.

**Future Plans:**

* Incorporating Additional Data Sources: Adding wearable device data, such as heart rate or GPS tracking, could offer a richer understanding of physical activity patterns.
* Building Predictive Models: Using machine learning techniques to predict step counts based on weather conditions and personal factors could provide actionable recommendations for activity planning.
* Developing a Weather-Activity Tracker: Creating an app or tool that integrates weather forecasts with personalized activity suggestions would make the findings of this project practical for everyday use.