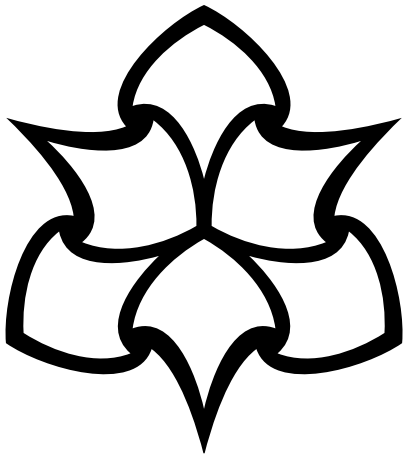

Manchester Metropolitan University

Big Data (6G7V0025_2223_1F)

Student ID: 22547202

Neha Alam Hussain

22/23 Semester 2

Contents

1. Introduction

This report aims to explore two research hypotheses related to the Transport for London (TfL) cycle hire usage dataset for 2014, focusing on cycling behavior in London. The first research hypothesis investigates ride durations in Autumn compared to Spring, while the second hypothesis will be formulated based on an additional dataset from the National Centers for Environmental Information (NCEI), which contains daily weather summaries for 2014. The report will present a detailed analysis of both hypotheses, including data preprocessing, visualization, and statistical testing.

2. Dataset Description

Transport for London (TfL) Cycle Hire Usage Dataset
The primary dataset for this study, provided by TfL, focuses on London's cycle hire usage in 2014, with details on individual journeys, docking station locations, and durations (Transport for London (TfL), 2014).

NCEI Daily Weather Summary Dataset
The additional dataset used for the second research hypothesis is obtained from the National Centers for Environmental Information (NCEI). This dataset contains daily weather summaries for 2014, including various weather-related measurements, such as temperature, precipitation, and wind speed. The combination of these two datasets will enable us to investigate the second research hypothesis by exploring the relationships between cycle hire usage patterns and weather conditions in London during 2014 (National Centers for Environmental Information, 2014).

3. Research Hypothesis 01:

3.1 Hypothesis Statement: In 2014, people ride for longer in Autumn than in Spring.

3.2. Methodology

3.2.1. Data Preprocessing

Data from multiple files were loaded into a single data frame using the Apache Spark environment. The schema of the data was automatically inferred, revealing a total of 9 features: rental id, duration, bike id, end date, end station id, end station name, start date, start station id, and start station name. The dataset comprised 11,481,596 rows.

An analysis of the entire dataset revealed that dates were initially in string formats, which were then converted to the correct timestamp format. Null values were inspected in each column, and a significant number of them were found. To focus the analysis on specific time periods of interest, the dataset was filtered to include only records from the Spring (March, April, May, June) and Autumn (September, October, November, and December) months of 2014.

Upon checking for null values in the Spring and Autumn datasets, it was observed that the Spring dataset had a total of 3,466,010 rows, with 32 rows containing null values. These null values constituted a very small fraction of the dataset (less than 0.001%), so the decision was made to remove the rows with null values as they were unlikely to have a significant impact on the results. Similarly, in the Autumn dataset, the fraction of null values was very small, so they were dropped as well. This ensured the completeness and accuracy of the data.

3.2.2. Feature Engineering
Additional features are derived from the data, such as converting the "Duration" column from seconds to minutes and extracting the month, year, day of the week, and hour of the day from the "Start Date" column. These features will be useful in further analysis.

3.2.3. Descriptive Statistics
The average ride duration for both Spring and Autumn months is calculated. The results show that the average ride duration in Spring is 25.50 minutes, while the average ride duration in Autumn is 22.70 minutes.

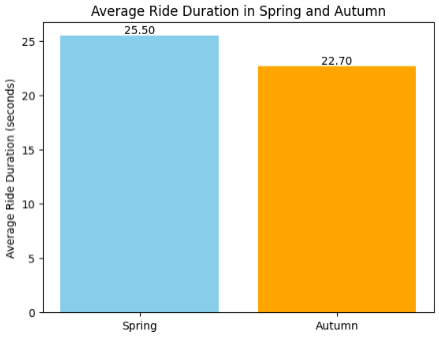

Figure 1

3.2.4. Hypothesis Testing
A statistical test was performed to determine if a significant difference exists in ride durations between Spring and Autumn. The t-statistic measures the difference between the means of the two groups relative to the variability of the data, while the p-value indicates the probability that the observed difference occurred by chance alone. A two-sample independent t-test was conducted to ascertain if a significant difference in ride durations between Spring and Autumn existed. The t-test produced a t-statistic of 16.63 and a p-value of 4.22e-62, highlighting a significant difference.

3.3 Result Interpretation
Based on the analysis of the provided dataset and the results obtained, the initial research hypothesis, "In 2014, people ride for longer in Autumn than in Spring," is not supported. The findings reveal that the average ride duration in Spring (25.50 minutes) was higher than in Autumn (22.70 minutes). The t-test results (t-statistic: 16.63, p-value: 4.21e-62) indicate a statistically significant difference between the ride durations in the two seasons. The very low p-value (much less than the typical significance threshold of 0.05) suggests that the observed difference is not due to random chance.

3.4. Refining the hypothesis
Based on the results obtained, the average ride duration in spring is longer than in autumn, which contradicts our initial research hypothesis. In this case, we can refine the hypothesis to make it more specific, and testable, and consider other factors that might influence ride durations.

3.4.1. Weekend Rides
The average weekend ride duration is calculated for both Spring and Autumn. The results show that the average weekend ride duration in Spring is 34.88 minutes, while the average weekend ride duration in Autumn is 30.25 minutes. This suggests that people ride for longer during weekends in Spring than in Autumn.

3.4.2. Time of Day Analysis
Upon analyzing ride durations during different times of the day in Spring and Autumn, it has been observed that ride durations are generally longer in Spring than in Autumn across all times of the day. The most noticeable differences are seen in the afternoon and evening, while the night and morning show relatively smaller differences.

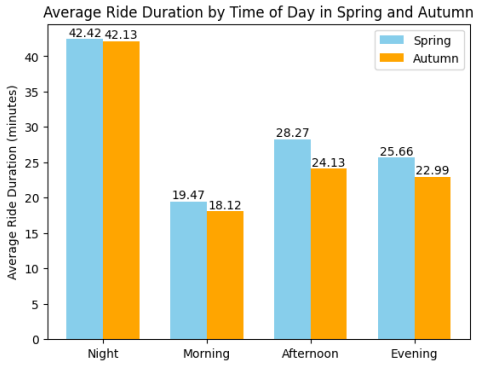

Figure 2

3.5. Conclusion
Based on the analysis of the TfL dataset, the first research hypothesis that people ride for longer in Autumn than in Spring in 2014 is not supported by the data. Instead, the analysis demonstrates that people ride for longer durations in Spring than in Autumn. The difference in ride durations between the two seasons is statistically significant. Additional analysis of weekend rides and rides during specific time periods further supports the finding that people tend to ride for longer durations in Spring than in Autumn.

4. Research Hypothesis 02:

4.1. Hypothesis Statement: Higher temperatures and lower precipitation levels in Spring 2014 led to longer bike ride durations as compared to Autumn 2014

4.2. Methodology

4.2.1. Data Preprocessing

The second hypothesis required the bike journey data to be merged with the weather data. The preprocessing stage was initiated by converting the date columns in the bike journey dataset to an appropriate timestamp format. It was noticed that a large number of null values were present across various columns in the bike journey dataset. Consequently, the null count for each column was computed to understand the extent of the problem.

The weather dataset was then reduced to only the 'DATE', 'TEMP', and 'PRCP' columns, allowing for a concise focus on date, temperature, and precipitation information. This decision was guided by the requirements of the hypothesis, which necessitated data relevant to weather conditions.

The date columns in both datasets were then converted to the same format to prepare for merging. The bike journey data and the weather data were then merged on the basis of the 'Start Date' and 'DATE', respectively. The merging resulted in a comprehensive dataset with 9,895,412 rows in total.

To meet the conditions of the second hypothesis, the merged dataset was filtered to include only data from the Spring and Autumn months of 2014. This filtering was facilitated by breaking down the 'Start Date' column into 'month' and 'year' columns.

Once filtered, null values in the Spring and Autumn datasets were checked. A small number of null values were found in the 'EndStation Id' and 'EndStation Name' columns. Considering their negligible proportion, rows with these null values were dropped, ensuring the integrity and cleanliness of the data.

Finally, to enhance the interpretability of our data, we converted the duration of the bike rides from seconds to minutes. To align with international measurement standards and make the data more relatable, we also converted the temperature from Fahrenheit to Celsius.

This comprehensive data preprocessing stage laid the groundwork for our subsequent data analysis, ensuring that our findings would be based on accurate, complete, and relevant data.

4.2.2. Feature Engineering

To analyze the influence of weather conditions on ride duration, feature engineering was conducted on both the Spring and Autumn datasets. This involved the generation of new categorizations for the variables of temperature and precipitation.

A new column, termed 'TEMP_RANGE', was created for temperature. This column divided the temperature into different categories: '0-10', '10-20', '20-30', and '30+' (measured in degrees Celsius). The determination of the range was accomplished by evaluating if the 'TEMP_C' value fell within the specified limits.

A similar categorization was applied to precipitation, resulting in a new column labeled 'PRCP_RANGE'. This column identified four categories: 'No Precipitation', 'Light Precipitation', 'Moderate Precipitation', and 'Heavy Precipitation'. The 'PRCP' values were used as a basis for this categorization, with 'No Precipitation' corresponding to a 'PRCP' value of 0, 'Light Precipitation' to 'PRCP' values greater than 0 but less than or equal to 0.1, 'Moderate Precipitation' to 'PRCP' values greater than 0.1 but less than or equal to 0.3, and 'Heavy Precipitation' to 'PRCP' values above 0.3.

These newly introduced categories enabled a more nuanced exploration of how temperature and precipitation could affect bike journey durations.

4.2.3. Descriptive Statistics

In the Spring dataset, it was found that the longest bike rides were experienced within the temperature range of '10-20' degrees Celsius when no precipitation was observed. The shortest rides, on the other hand, were observed within the '0-10' degrees Celsius range during heavy precipitation (see Appendix A).

In the Autumn dataset, the longest journey durations were observed under light precipitation within the '0-10' degrees Celsius range. The shortest durations, however, were recorded in the same temperature range, but with moderate precipitation (see Appendix B).

Longer rides are typically observed under milder and drier conditions, while shorter rides are more common in colder and wetter conditions.

4.2.4. Statistical Testing

Statistical tests were conducted on the Spring and Autumn datasets. The independent variables used in the tests were the temperature range and precipitation range, while the dependent variable was the total duration of bike rides. Two-way ANOVA tests were performed to determine the influence of the independent variables on the dependent variable.

The ANOVA results for Spring showed that both the temperature range (F=6.13, p=0.002956) and precipitation range (F=4.08, p=0.008543) significantly affected the total ride duration. For Autumn, only the temperature range (F=23.86, p=0.000003) significantly affected the total ride duration, while the precipitation range did not (F=0.94, p=0.424274) (see Appendix C).

4.3 Result Interpretation

The results of the statistical tests confirmed the initial hypothesis that higher temperatures and lower precipitation levels led to longer bike ride durations. Specifically, in the Spring dataset, longer durations were associated with temperatures in the '10-20' degrees Celsius range and light or no precipitation. In the Autumn dataset, significant differences in ride durations were observed between '0-10' and '10-20' degrees Celsius, indicating that higher temperatures still contributed to longer rides, despite the lack of significant differences in precipitation levels.

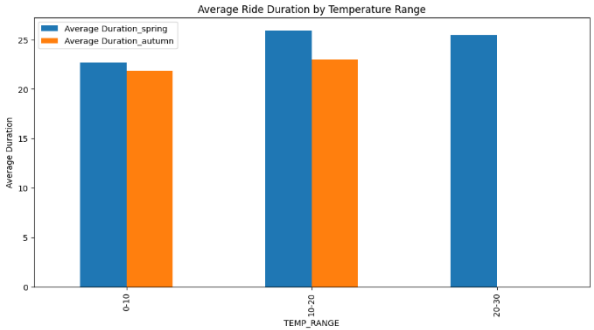

Figure 3

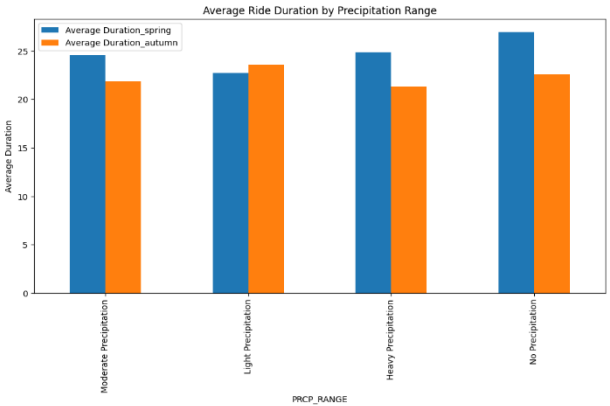

Figure 4

4.5. Conclusion

From the data analysis, it can be concluded that temperature and precipitation are significant factors affecting bike ride durations. The hypothesis that higher temperatures and lower precipitation levels in Spring led to longer bike rides compared to Autumn was supported by the findings. The analysis suggests that temperature may be a more influential factor on bike ride duration than precipitation. These insights could be valuable in planning and managing bike-sharing schemes, particularly in anticipating seasonal changes in demand and usage patterns. This study also demonstrates the potential of utilizing weather data in analyzing and predicting human mobility patterns.

Appendixes

Appendix A

TEMP_RANGE	PRCP_RANGE	Average Duration
10-20	Heavy Precipitation	26.063421702515523
0-10	Light Precipitation	20.283229822573475
0-10	Moderate Precipit...	22.163550790953067
0-10	Heavy Precipitation	16.11090696509103
10-20	Moderate Precipit...	25.01357619339655
0-10	No Precipitation	24.883793922776146
10-20	No Precipitation	27.24760708985301
10-20	Light Precipitation	23.194083675699332
20-30	No Precipitation	25.425725323839103

Appendix B

TEMP_RANGE	PRCP_RANGE	Average Duration
10-20	Heavy Precipitation	20.8503439404955
10-20	Moderate Precipit...	23.704578180041654
10-20	No Precipitation	23.228643712600576
10-20	Light Precipitation	22.61058765650678
0-10	Light Precipitation	25.88487979696564
0-10	Moderate Precipit...	17.058664627930682
0-10	Heavy Precipitation	21.912048633330286
0-10	No Precipitation	19.96184609726947

Appendix C

Anova Test Result for Spring

	sum_sq	df	F	PR(>F)
C(TEMP_RANGE)	1.017154e+12	2.0	6.126526	0.002956
C(PRCP_RANGE)	1.016019e+12	3.0	4.079790	0.008543
Residual	9.629430e+12	116.0	NaN	NaN

Anova Test Result for Autumn

	sum_sq	df	F	PR(>F)
C(TEMP_RANGE)	1.472554e+12	1.0	23.860454	0.000003
C(PRCP_RANGE)	1.738423e+11	3.0	0.938948	0.424274
Residual	7.220685e+12	117.0	NaN	NaN

References

National Centers for Environmental Information, 2014. *Global Summary of the Day (GSOD).* [Online]
Available at: https://www.ncei.noaa.gov/data/global-summary-of-the-day/access/2014/03770099999.csv
[Accessed Monday May 2023].

Transport for London (TfL), 2014. *Cycle Hire Usage Stats.* [Online]
Available at: https://cycling.data.tfl.gov.uk/usage-stats/cyclehireusagestats-2014.zip
[Accessed Monday May 2023].