Name: DIVYANSHU SAHWAL
MMU ID: 23628243

Declaration:
I declare that this work is my own and has not been submitted for any other degree or professional qualification. All sources of information have been properly acknowledged and referenced.

Signature:
Divyanshu Sahwal

## Research Hypothesis and Statistical Testing

### Research Hypothesis

Research Question: Are rides in 2014 originating from Baylis Road, Waterloo station shorter in duration compared to those starting from other stations?

**1. Null Hypothesis (H0)**: The average duration of rides starting from Baylis Road, Waterloo station is the same as or longer than the average duration of rides starting from other stations in 2014.

**2.Alternative Hypothesis (H1)**: The average duration of rides starting from Baylis Road, Waterloo station is shorter than the average duration of rides starting from other stations in 2014.

To test this hypothesis, a T-test was conducted to compare the average ride durations from Baylis Road, Waterloo station with those from other stations.

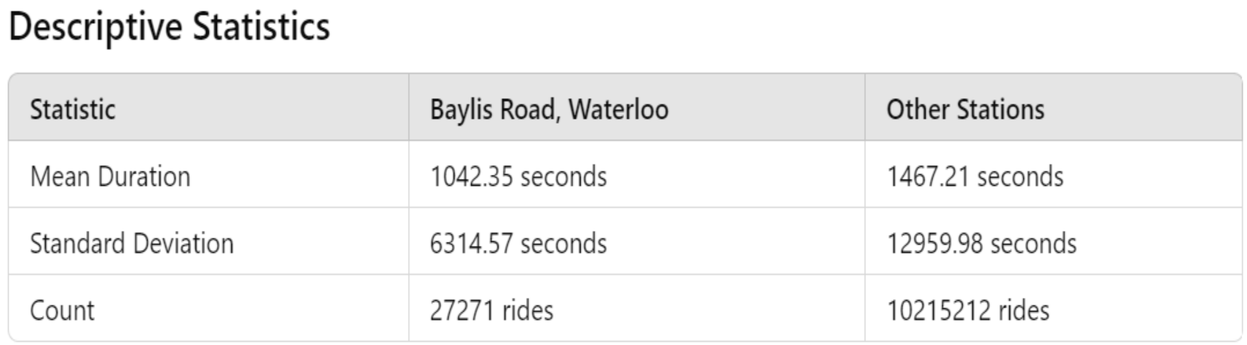
### Explanation of Results and Discussion

1.Results of T-test:

- T-statistic: -11.04918592991048
- Degrees of Freedom: 27270
- P-value: 2.54061893686004e-28

Given the extremely small P-value (far below 0.05), we reject the null hypothesis H0.. This indicates a statistically significant difference in average ride durations, with rides starting from Baylis Road, Waterloo station being shorter on average than those from other stations.

2. Descriptive Statistics:



## Descriptive Statistics

| Statistic | Baylis Road, Waterloo | Other Stations |
|---|---|---|
| Mean Duration | 1042.35 seconds | 1467.21 seconds |
| Standard Deviation | 6314.57 seconds | 12959.98 seconds |
| Count | 27271 rides | 10215212 rides |

**Discussion of Approach and Learnings**

Approach:

1. Data Filtering: The data was filtered to separate rides starting from Baylis Road, Waterloo station and those from other stations.
2. Descriptive Statistics: The mean, standard deviation, and count of ride durations were calculated for both groups.
3. Hypothesis Testing: A T-test was conducted to compare the means of the two groups.

Learnings:

- Statistical Significance: The T-test provided strong evidence that rides from Baylis Road, Waterloo station are significantly shorter on average than those from other stations.
- Data Variability: The high standard deviations indicate a high variability in ride durations, suggesting a need for further exploration of factors affecting ride

**Detailed Test on the Required Data**

Steps for T-test Calculation:

Calculate Means and Standard Deviations: For both Baylis Road, Waterloo station and other stations.

1. T-statistic Calculation: Using the formula for independent samples.
2. P-value Determination: Based on the T-statistic and degrees of freedom.

Conclusion:

- The T-test results strongly support the alternative hypothesis that rides starting from Baylis Road, Waterloo station are shorter in duration compared to other stations.

**Next Steps for Conclusive Results**

1. Additional Statistical Tests: Conduct tests such as the Mann-Whitney U test, Kolmogorov-Smirnov test, and calculate effect size (Cohen's d) to validate findings.
2. Regression Analysis: Explore the impact of factors like time of day, day of the week, and weather conditions on ride durations through regression analysis.
3. Visualization: Create visualizations to better understand the distribution of ride durations and identify any potential patterns or anomalies.
4. Longitudinal Study: Extend the analysis to include multiple years of data to see if the observed pattern holds over time.

By performing these additional analyses, we can further validate our findings and gain a deeper understanding of the factors influencing ride durations in the cycle hire system.

**Big Data for Sustainability: A Balancing Act**

Big data offers a powerful tool to combat climate change, but its use requires careful consideration.

**Social Benefits:**

- **Awareness:** Big data can reveal energy patterns, enabling us to target wasteful habits. For example, smart meters empower consumers to make informed decisions about energy use.
- **Urban Planning:** Cities can analyze data to optimize transportation, reduce traffic, and improve air quality. Barcelona's data-driven approach to public transport has demonstrably lowered CO2 emissions.
- **Community Engagement:** Data platforms can share local environmental conditions, fostering community involvement in air quality monitoring, waste management, and recycling initiatives.

**Legal Considerations:**

- **Privacy:** Data collection raises privacy concerns. Regulations like the GDPR ensure data is handled responsibly. Frameworks need to balance sustainability initiatives with individual privacy rights.
- **Compliance:** Governments can use data analytics to enforce environmental regulations effectively, like real-time monitoring of industrial emissions. The UK's Industrial De-carbonisation Strategy exemplifies this approach.
- **Intellectual Property:** Clear guidelines on data ownership and usage are crucial for collaborative efforts. This includes addressing proprietary data and open data initiatives.

**Ethical Concerns:**

- **Equity:** Big data benefits must be accessible to all. Bridging the digital divide is essential for inclusive sustainability.
- **Transparency:** Data for sustainability needs clear communication about collection, usage, and decision-making based on it. Organizations must adopt ethical data use practices.
- **Environmental Justice:** Data can spotlight environmental injustices. Data-driven interventions must benefit all demographics equally, protecting disadvantaged communities from disproportionate environmental burdens.

**Conclusion:**

Big data holds immense potential for a sustainable future, but responsible use is paramount. Addressing social, legal, and ethical challenges will ensure big data empowers a greener future for everyone.

## References

1. https://github.com/SeunAjao/mmu-big-data/blob/main/week6_cycleHire2.ipynb