Big Data Analysis of TfL Cycle Hire for Global Environmental Sustainability

I. Introduction

This research delves into cycling patterns within London, utilizing the 2014 Transport for London (TfL) cycle hire scheme dataset. We specifically investigate whether rides commencing from Baylis Road, Waterloo Station were shorter in duration than those starting elsewhere.

To analyse the extensive TfL dataset, we employed Apache Spark, a tool renowned for efficiently managing large datasets. Apache Spark enabled us to process the data swiftly and conduct parallel computations, crucial for handling the dataset's scale and complexity effectively

II. Research Hypothesis

In this study, we will explore the following hypotheses:
H0 (Null Hypothesis): The average duration of bike rides originating from Baylis Road, Waterloo Station in 2014 is not significantly different from those starting at other TfL cycle hire stations.
H1 (Alternative Hypothesis): The average duration of bike rides originating from Baylis Road, Waterloo Station in 2014 is significantly shorter compared to those starting at other TfL cycle hire stations.

III. Data Analysis

The data preparation phase is crucial for ensuring the quality and reliability of the results derived from any data analytics project. This phase involved several steps to clean, transform, and organize the 2014 Transport for London (TfL) cycle hire scheme dataset for efficient analysis using Apache Spark. Our preparation activities included the following:

A. Environment Setup:

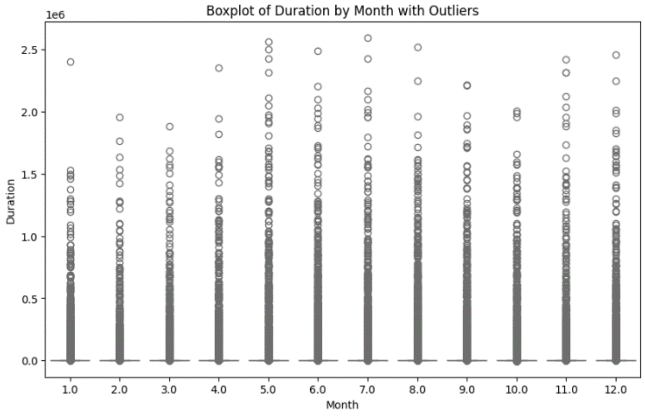
Initialized Apache Spark with the custom application name 'Bikes_london'. We configured the environment with 8 GB of memory for both the driver and the executors and set the number of cores for the executors to 2.

B. Data Ingestion:

Data Source: https://cycling.data.tfl.gov.uk/usage-stats/cyclehireusagestats-2014.zip
From the data source, we have loaded a total of 11,481,596 cycling trip records into a Spark Data Frame from multiple CSV files. The data contained fields such as Rental ID, Duration, Bike ID, Start Date, End Date, End Station Id, EndStation Name, StartStation Id, StartStation Name of each trip.

C. Initial Data Exploration:

The boxplot provides an overview of ride durations across different months.



The spread of data within the interquartile ranges remains relatively similar month to month, suggesting little variation in ride duration over time.
Notable outliers are present each month, representing instances of significantly longer rides. The distributions are right-skewed, indicating that longer rides are relatively infrequent.

summary	Duration
count	10242483
mean	1466.081447242822
stddev	12946.838147599005
min	-3360
max	2596560

An examination of the dataset has revealed a total of 10,242,483 recorded durations. The average duration is calculated to be approximately 1,467 minutes. A considerable standard deviation of about 12,947 minutes indicates a broad variation in ride durations. The presence of a negative minimum duration value, -3,360 minutes, has been identified, suggesting inaccuracies in the data. Additionally, the maximum recorded duration of 2,596,560 minutes has been observed, which likely signifies the existence of extreme outliers or erroneous entries.

D. Data Cleaning

This critical phase was essential for ensuring the accuracy and integrity of our dataset for analysis. The steps taken included

1) ***Type Conversion****:* The 'Duration' field, essential for our analysis, was initially of the type string. This field was transformed into an integer data type, which is crucial for any subsequent analysis.

2) ***Date Normalization****:* We reformatted the 'Start Date' from a string to a date type using to_date Method. This adjustment allowed us to accurately extract the 'Month' from each entry, which was imperative for the analysis of temporal trends within the data.

*3) **Handling Missing Values & Duplicate Entries**:* The Analysis of the data revealed that there were 1,239,112 records with missing values across all fields. This exact number also corresponded to the count of duplicate records identified, signaling potential data recording discrepancies. In response, these records were removed to enhance the dataset's coherence and to ensure the integrity of the analysis.

```
{col:cyclehire.filter(cyclehire[col].isNull()).count() for col in cyclehire.columns}
```

```
{'Rental Id': 1239113,
 'Duration': 1239113,
 'Bike Id': 1239113,
 'End Date': 1239113,
 'EndStation Id': 1239245,
 'EndStation Name': 1239245,
 'Start Date': 1239113,
 'StartStation Id': 1239113,
 'StartStation Name': 1239113}
```

*4) **Handling Incorrect Values:***

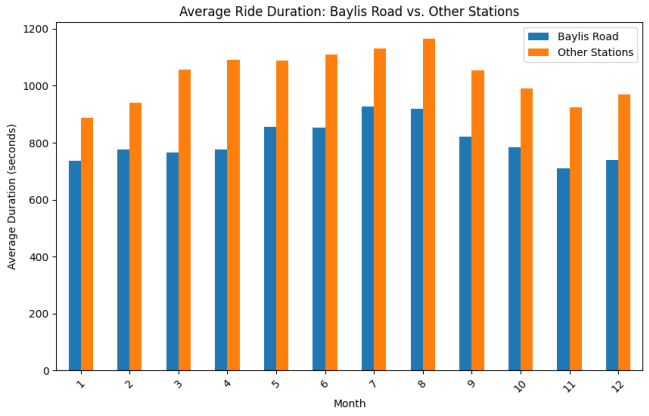
In the data cleansing process, entries with durations less than or equal to zero, as well as those with null values, were deemed incorrect and have been excluded from the dataset. Additionally, the dataset was further refined by removing outliers; specifically, records where the duration was below 90 seconds or exceeded 7200 seconds were filtered out. These thresholds were set to exclude durations that are unlikely to represent genuine rentals, with the lower limit removing potential false starts and the upper limit removing likely cases of bikes not returned on time.

IV. INSIGHTS

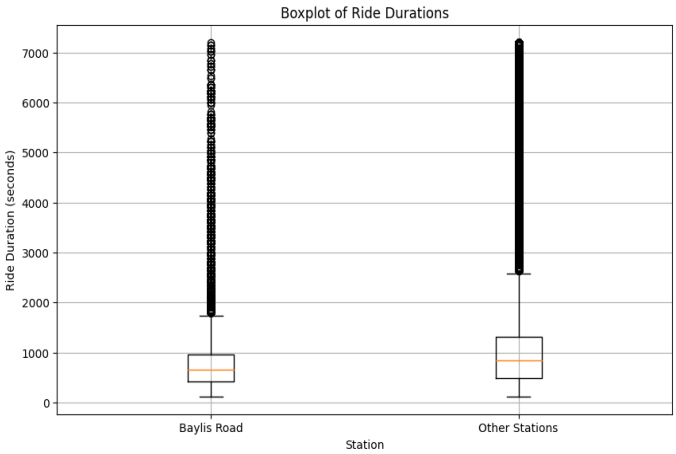
Subsequent to data cleansing, the dataset was partitioned into two subsets for comparative purposes: one comprising rides originating from Baylis Road and the other encompassing all other stations. This sub setting was crucial to ensure a focused analysis relevant to the hypothesis.
Following the sub setting of the data, Visualization was created for clear comparison.

*A. **Average Ride Duration Comparison:***

The bar graph titled "Average Ride Duration: Baylis Road vs. Other Stations" illustrates the monthly comparison of average ride durations. The graph distinctly shows that Baylis Road (Blue) consistently has shorter average ride durations compared to other stations (orange) for each month.



*B. **Distribution of Ride Durations:***



The ride duration distributions for Baylis Road in comparison to other stations are shown as a boxplot. While the wider distributions for other stations show a larger variance in trip lengths, the more centered interquartile range for Baylis Road indicates shorter central travel times. There is strong visual evidence that rides beginning at Baylis Road have shorter average durations.

After the visual comparisons were showcased, the T-test results were subsequently analyzed to provide statistical validation for the observed differences:

*C. **T-Test:***

The T-test, a statistical technique applied to evaluate the significance of the mean differences between two distinct groups. This evaluation was performed to thoroughly examine the monthly average ride durations from Baylis Road in comparison to other stations. This phase is crucial as it provides statistical support to the earlier visual findings, establishing a firm foundation for the hypothesis that rides starting from Baylis Road consistently have shorter durations.

*D. **T-test Results Overview:***

The T-test results for each month clearly indicate a significant difference in ride durations between Baylis Road and other stations.

Monthly Analysis Highlights:

January: With a T-statistic of -10.17 and a p-value well below 0.0001, the significant difference in ride durations is evident right from the start of the year.

February to December: Similar patterns persist, with each month showing a significant difference in ride durations, as denoted by the negative T-statistics and negligible p-value.

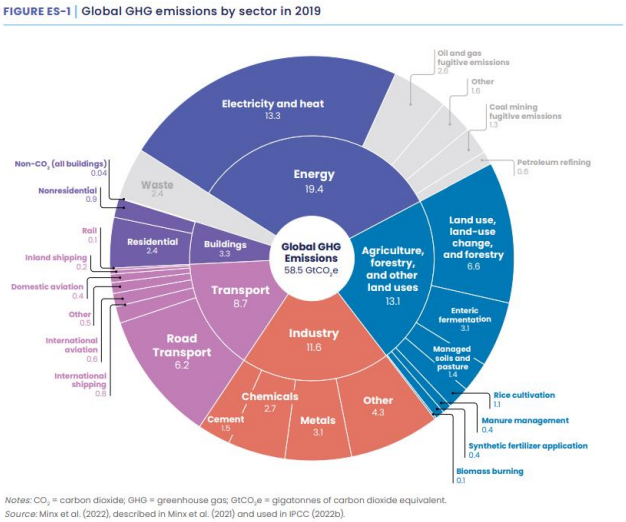
```
Month: 1, T-statistic: -10.165251645928988, P-value: 1.6016057794460003e-23
Month: 2, T-statistic: -10.270432476698222, P-value: 6.1862243392770346e-24
Month: 3, T-statistic: -21.477724339487512, P-value: 4.5849281857983135e-92
Month: 4, T-statistic: -24.911824215337262, P-value: 3.745975235693533e-120
Month: 5, T-statistic: -15.000383834128451, P-value: 1.1025104589107031e-48
Month: 6, T-statistic: -19.36017518006243, P-value: 2.9969920556755436e-78
Month: 7, T-statistic: -13.691833188617627, P-value: 1.9712812243143243e-41
Month: 8, T-statistic: -15.393231043718451, P-value: 2.723840201820637e-51
Month: 9, T-statistic: -18.64836689743181, P-value: 3.32440907131376666e-73
Month: 10, T-statistic: -16.393214318974408, P-value: 2.850898952373374e-57
Month: 11, T-statistic: -18.247514204991752, P-value: 5.77616448632208e-69
Month: 12, T-statistic: -18.323290447350928, P-value: 3.79254239966668e-70
```

E. Correlation with Visual Data:

The statistical findings align with the visual observations from the earlier boxplot and bar graph analyses. The visual data had already suggested shorter ride durations for Baylis Road, which is now statistically verified through the T-test results.

V. CRITICAL REFLECTION:

Carbon dioxide levels are higher now than they have been in the last three million years. This spike is largely due to human activities that emit greenhouse gases (GHGs) like burning fossil fuels, deforestation, and intensive agriculture. These activities not only increase CO2 but also trap heat in the atmosphere, leading to a rise in average temperatures and a surge in extreme weather events.



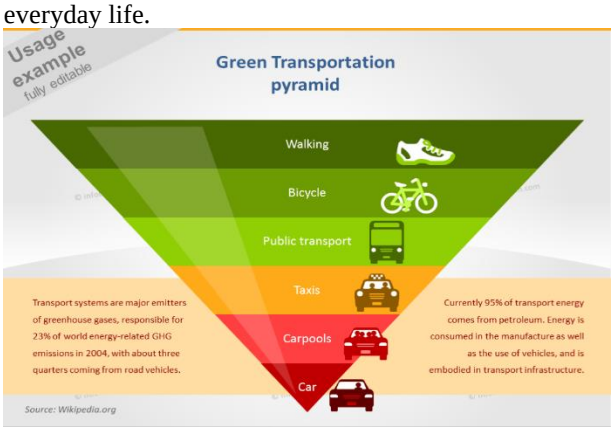
The provided emissions chart shows that the energy industry, agriculture, and transport are major sources of greenhouse gases. To combat climate change, we need to target these areas.

.

Big Data Analytics stands out as a transformative digital technology with the potential to drive decarbonization in these key sectors. It holds the capacity to dissect the vast arrays of data they generate, identifying emission hotspots and informing strategies for energy optimization, waste reduction, and overall emission cuts. It can sift through massive amounts of information from these sectors to help us understand where and how emissions are happening. By analyzing energy use, Big Data can help make power plants more efficient. It can also help factories and farms reduce waste and emissions by pinpointing exactly where changes need to be made. Big Data's predictive analytics can also forecast future emission trends, allowing for preemptive strategies in energy production and agricultural practices, thus minimizing environmental impact before it occurs.

A. Social Perspective:

Big Data Analytics goes beyond system optimization—it sparks societal transformation. By highlighting how activities like energy use, farming, and transportation lead to significant greenhouse gas emissions, it guides smarter choices. Easy access to this data nudges the public towards eco-friendlier travel, such as walking or biking, in line with the ideals of the green transportation pyramid. These data-driven insights can shift societal norms towards public transit, lessening the dependence on cars and lowering emissions. This approach promotes a community-centered fight against climate change, building a culture that embraces sustainability as a collective commitment woven into everyday life.



B. Legal Perspective:

For the legal perspective on Big Data in transportation, it is clear that challenges like data access and ownership stand in the way of full utilization. The study by Sánchez-Martínez & Munizaga (2016) outlines the concern that transportation data, often held by various organizations, is not readily shared due to competition and the commercial value of the data.
The challenges in leveraging Big Data within the transportation sector primarily stem from issues related to data access, ownership, and privacy concerns:

Data Access and Sharing: Different organizations, including private companies and government agencies, collect vast amounts of transportation data. The reluctance to share this data is often due to competitive fears or the commercial value attributed to proprietary information.

Data Ownership: Transportation data owners, such as operators and agencies managing Automated Vehicle Location (AVL), Automated Fare Collection (AFC), and Automatic Passenger Counting (APC) systems, may resist sharing their data. They worry it could be exploited by competitors or lead to unwanted exposure.

Privacy Concerns: Legal challenges arise around protecting individuals' privacy, particularly with data detailing trip origins, destinations, and fare transactions. Adhering to privacy laws while promoting data sharing for the public good is a delicate balance.

C. Ethical Perspective:

In examining the ethical perspective of transportation through the lens of industry, government intervention, and individual behavior, several nuanced challenges emerge.
L**abor Impacts:** The effect of automation and digital technologies on employment within the transportation industry, including job displacement and the need for re-skilling.

Surveillance: Utilizing big data in environmental observation may inadvertently lead to surveillance-like issues, risking personal autonomy. It's vital to navigate this ethically, safeguarding individual liberties while pursuing communal environmental benefits.
Industry Concerns: Deregulation raises questions about private companies' social responsibilities versus their commercial goals. Issues include fare affordability and service accessibility, particularly in low-demand areas.
Government Intervention: Ethical dilemmas involve balancing economic efficiency with social equity and environmental sustainability. Decisions about investment and new infrastructure need to consider territorial justice, impacting regional development.

VI. Conclusion

To conclude, this analysis of TfL cycle hire data has revealed the shorter ride durations from Baylis Road, demonstrating the impact of Big Data Analytics on understanding transportation patterns. This study teaches us how Big Data can help us understand travel habits better and show us ways to make transport greener. By thinking about how we use data and making sure it's fair and safe, Big Data can be a big help in fighting climate change. This work shows that using data wisely can lead us to smarter choices for our planet.

References:

[1] European Environment Agency, 2017. Greenhouse gas emissions from transport. [online] Available at: www.eea.europa.eu/data-and-maps/indicators/transport-emissions-of-greenhousegases/transport-emissions-of-greenhouse-gases-10 [Accessed 27 March 2024].

[2] Xia, Y., Chen, J., Lu, X., Wang, C. & Xu, C., 2016. Big traffic data processing framework for intelligent monitoring and recording systems. Neurocomputing, 181, pp.139–146. Available at: https://doi.org/10.1016/j.neucom.2015.07.140.

[3] Amazon, 2023. Carbon Methodology [pdf]. Available at: https://sustainability.aboutamazon.com/carbon-methodology.pdf [Accessed 27 March 2024].

[4] Jain, B., 2023. The role of big data and artificial intelligence in sustainability. Voices. [blog] Times of India. Available at: https://timesofindia.indiatimes.com/blogs/voices/the-role-of-big-data-and-artificial-intelligence-in-sustainability/ [Accessed 27 March 2024].

[5] United Nations, n.d. The Development Agenda. United Nations Sustainable Development. Available at: https://www.un.org/sustainabledevelopment/development-agenda/ [Accessed 27 March 2024].

[6] Corbett, C.J., 2018. How Sustainable Is Big Data? Production and Operations Management, 27(9), pp.1685-1695. Available at: https://doi.org/10.1111/poms.12837

[7] World Economic Forum, 2022. How digital solutions can reduce global emissions. Davos Agenda. Available at: https://www.weforum.org/agenda/2022/05/how-digital-solutions-can-reduce-global-emissions/ [Accessed 27 March 2024].

[8] UK Government, 2023. Transport and Environment Statistics 2023. Available at: https://www.gov.uk/government/statistics/transport-and-environment-statistics-2023/transport-and-environment-statistics-2023#greenhouse-gases-journey-emission-comparisons [Accessed 27 March 2024].

[9] Neethirajan, S., 2024. Net Zero Dairy Farming—Advancing Climate Goals with Big Data and Artificial Intelligence. Climate, 12, p.15.

[10] ODBMS, 2018. Big Data in Transportation [pdf]. Available at: https://www.odbms.org/wp-content/uploads/2018/10/Big-Data-in-Transportation.pdf [Accessed 27 March 2024].

[11] Hashem, I.A.T., Chang, V., Anuar, N.B., Adewole, K., Yaqoob, I., Gani, A., Ahmed, E. and Chiroma, H., 2016. The role of big data in smart city. International Journal of Information Management, 36(5), pp.748-758. Available at: https://doi.org/10.1016/j.ijinfomgt.2016.05.002 [Accessed 27 March 2024].