**Big Data Assignment**

**Jamie Atiyah**

The entirety of this work is my own


Faculty of Science and Engineering
Manchester Metropolitan University
United Kingdom
16.05.2023

## I. Introduction

Constant advancements in data storage capacity along with the reduction size of memory devices have provided the opportunity to store more data than ever before. These advancements are driven by the increasing generation of data along with its growing acceptance as a powerful tool for businesses. Data can be received in many volumes, velocities, varieties such as unstructured and semi-structured, value, and veracity (the quality of data). Big data solutions which can deal with these issues are vital as the need for quicker and more efficient processing and exploration is growing with the demand from insights guided by data ever-increasing. Big Data frameworks such as Apache Spark (Zaharia et al. 2016) have been developed and are becoming increasingly popular. Big data solutions are being utilised in a wide range of scenarios, from healthcare (Ebada et al. 2022) (Sujitha and Paramasivan 2022) (Albaldawi, Almuttairi, and Manaa 2022), to agricultural applications (Cravero et al. 2022).

This report will use Spark 3.3.2 (Spark 2023a) to test two hypotheses which relate to bike rentals. The first hypothesis is that in 2014 people ride for longer in autumn than in spring. The second hypothesis is that more people will rent a bike whilst traffic is low. Both hypothesis will utilise the transport for London (tfl) cycle hire usage stats for 2014 (London 2015). The second hypothesis will also utilise the road traffic statistics relating to London from the Department for Transport (dft) (Transport 2015).

## II. Methodology

### A. Experimental Setup

The experiments were carried out in Spark 3.3.2 (Spark 2023a). Spark 'local' mode was set to run on 2 cores. The integrated SparkSQL (Spark 2023b) functions are utilised for data manipulation. The cloud-based jupyter notebook environment Google Colaboratory is used to write and execute the code.

### B. Stated Hypothesis

The first hypothesis is stated below 1.

**Hypothesis 1 (H1):** *In 2014 people ride for longer in autumn than in spring.*

To suitably test this hypothesis, definitions for the time periods that span autumn and spring needed to be defined. Guided by insights from the Met Office, spring was determined to be between the 1st March and 31st May (Office 2014b) whilst autumn was determined to be between 1st September to 31st November (Office 2014a).

It is theorised that traffic volumes may have a direct effect on the number of bike rentals. Therefore, the second hypothesis is.

**Hypothesis 2 (H2):** *In 2014, there are more bike rentals whilst traffic is low*

All bike rentals and vehicle traffic data in the year 2014 are used to test the second hypothesis.

### C. Data Collection

*1) Transport for London:* The tfl data is obtained using the wget function to obtain https://cycling.data.tfl.gov.uk/usage-stats/cyclehireusagestats-2014.zip and the zipped file is unzipped. There are 20 csv files, covering various time periods within the year 2014. The schema of each file was checked to provide verification that they have identical schema. Each file has the columns rental id, duration, bike id, end date, endstation id, endstation name, start date, start station id, and start station name. Some files however have 3 additional columns _c9, _c10, and _c11. Assessing the distinct values in each of these 3 additional columns shows that only null values exist in each. Therefore, the schema of each file was deemed suitable to one another.

```
!wget https://cycling.data.tfl.gov.uk/usage-
    ↪ stats/cyclehireusagestats-2014.zip
!unzip cyclehireusagestats-2014.zip
```

For the first hypothesis, the data is filtered to obtain bike rentals that occurred within the spring and autumn seasons.The files 2,3,4,5,6 span a time period of 03 February 2014 to 21 June 2014. Files 9b, 10a, 10b, 11a, 11b, 12a, 12b span a time period of 01 September 2014 to 06 December 2014. By amending the column 'Start Date' to a timestamp data type, bike rentals that occurred outside the respective time periods, outlined in section II-B, can be filtered out. For the second hypothesis, all files were loaded and any bikes rented in years outside of 2014 were filtered out of the data.

*2) Department for Transport:* The dft data is obtained using the wget function to obtain https://storage.googleapis.com/dft-statistics/road-traffic/downloads/rawcount/region_id/dft_rawcount_region_id_6.csv.

```
!wget https://storage.googleapis.com/dft-
    ↪ statistics/road-traffic/downloads/
    ↪ rawcount/region_id/
    ↪ dft_rawcount_region_id_6.csv
```

Amending the date column to timestamp allows for any traffic data that relates to years that are not 2014 to be filtered.

### D. Data Inspection and Understanding

*1) TFL Data:* In the spring and autumn bike rental data sets there are 2390431 and 257923 rows respectively. In the data frame that consists of all bike rentals in 2014, there are 9895412 rows. Table I shows the columns in the tfl data as well as the data types. Many columns share similar properties. For example, end station id denotes the unique id for each end station name. The top 5 rows are displayed in table II. For the second hypothesis, only 3 of the columns are kept. The variables kept were rental Id, start, and duration.

Table I
TFL VARIABLES

| Variable Name | Description | Data Type |
|---|---|---|
| Rental Id | Unique identifier of the bike rental | String |
| Duration | The duration of the bike rental in seconds | Integer |
| Bike Id | The Id of the bike that has been rented | String |
| End Date | The date when the rental ended | String |
| End Station Id | The Id of the station where the rental ended | String |
| End Station Name | The name of the station where the rental ended | String |
| Start Date | The date when the rental started | String |
| Start Station Id | The Id of the station where the rental started | String |
| Start Station Name | The name of the station where the rental started | String |
| Start | The timestamped transform start date | Timestamp |

*2) DFT Data:* The dft data consists of vehicle counts recorded at many count points. After filtering to dates in 2014, there are 12360 rows with 32 columns. Many of the columns are not required for this project, such as the direction of travel, the region id and name, local authority id and name, road name and type, and the starting and ending junction road name. Table III shows the variables that were selected along with their data types. Table IV shows the top 5 rows in the dft data set.

### E. Data Processing

*1) Missing Values:* To ensure the completeness and cleanliness of the data sets, missing values and outliers are processed. In the spring data set, there are 20 bike rentals that have a null or Nan value in the EndStation Id and End Station Name variables. There was little correlation between the bike rentals that had null values. Therefore, null values were dropped from the spring data frame. There were no null values in the data set used for the second hypothesis.

*2) Outliers:* To be able to suitably test both project hypotheses, an establishment of what a valid bike ride needs to be defined. Bike rides that last in excess of a day can be questioned as being valid rides whilst rides that last an incredibly small amount of time could be rides that have been mistakenly rented. To detect outliers, the inter quartile range is used (IQR). The 25th and 75th percentiles of the duration columns are detected for each bike rental data set used. Table V shows the upper and lower quartiles of duration for each of the data sets, along with the calculated IQR. As the lower bound ($lower\_quartile - 1.5 \times IQR$) is negative, the lower bound is manually set to 30 seconds. Bike rentals that had a duration that were either less than 30 seconds or exceeded the upper bound ($upper\_quartile + 1.5 \times IQR$) were removed from each data set using their respective quartiles and IQR's.

Table III
DFT VARIABLES

| Variable Name | Description | Data Type |
|---|---|---|
| Count Date | The date when the vehicles were counted | Timestamp |
| Hour | The hour when the vehicles were counted | Integer |
| Pedal Cycles | The number of pedal cycles counted | Integer |
| All Motor Vehicles | The number of motor vehicles counted | Integer |

Table IV
5 ROWS OF DFT DATA

| Count Date | Hour | Pedal Cycles | All Motor Vehicles |
|---|---|---|---|
| 2014-09-23 00:00:00 | 13 | 3 | 719 |
| 2014-09-23 00:00:00 | 14 | 6 | 668 |
| 2014-09-23 00:00:00 | 15 | 7 | 687 |
| 2014-09-23 00:00:00 | 16 | 10 | 823 |
| 2014-09-23 00:00:00 | 17 | 18 | 762 |

Table V
UPPER AND LOWER QUARTILES OF DURATION

| Data Set | 25% | 75% | IQR |
|---|---|---|---|
| Spring | 480 | 1320 | 840 |
| Autumn | 480 | 1260 | 780 |
| 2014 | 480 | 1320 | 840 |

*3) Erroneous Values:* The timestamp variable count date required amending to date type as all values of hour, minute and second in the timestamp values were 0. However, these values are clear errors as values of hour are already provided. The minute and seconds were not required for this project so the count date variable was amended to a date variable.

### F. Data Merging

For the second hypothesis 2, the tfl and dft data sets were joined by using a combination of aggregation and merging. The number of bike rides, the average bike ride duration, the minimum and maximum bike ride duration per hour were calculated. The total number of counted motor vehicles per hour was calculated. The two aggregated data frames were inner joined by the date and hour.

Table II
5 ROWS OF TFL DATA

| Rental Id | Duration | Bike Id | End Date | EndStation Id | EndStation Name | Start Date | StartStation Id | StartStation Name | Start |
|---|---|---|---|---|---|---|---|---|---|
| 33253206 | 1140 | 11439 | 27/05/2014 00:59 | 695 | Islington Green, ... | 27/05/2014 00:40 | 81 | Great Titchfield ... | 2014-05-27 00:40:00 |
| 33345935 | 660 | 10792 | 30/05/2014 18:00 | 326 | Graham Street, Angel | 30/05/2014 17:49 | 203 | West Smithfield R... | 2014-05-30 17:49:00 |
| 33385892 | 1560 | 6172 | 31/05/2014 19:10 | 430 | South Parade, Che... | 31/05/2014 18:44 | 286 | St. John's Wood R... | 2014-05-31 18:44:00 |
| 33340654 | 1380 | 9506 | 30/05/2014 16:19 | 86 | Sancroft Street, ... | 30/05/2014 15:56 | 430 | South Parade, Che... | 2014-05-30 15:56:00 |
| 33353596 | 1620 | 11872 | 30/05/2014 21:49 | 430 | South Parade, Che... | 30/05/2014 21:22 | 440 | Kennington Oval, ... | 2014-05-30 21:22:00 |

Table VI
MERGED DATA SET FOR HYPOTHESIS 2

| Sum of Bike Ride Durations | Average Bike Ride Duration | Minimum Bike Ride Duration | Maximum Bike Ride Duration | Bike Ride Count | Date | Hour | Total Count of Motor Vehicles |
|---|---|---|---|---|---|---|---|
| 1117080 | 760.4356705241661 | 60 | 2400 | 1469 | 2014-10-24 | 7 | 5632 |
| 2027940 | 808.5885167464115 | 60 | 2400 | 2508 | 2014-10-24 | 18 | 6222 |
| 1043040 | 773.7685459940653 | 60 | 2400 | 1348 | 2014-10-24 | 13 | 6614 |
| 1550160 | 735.3700189753321 | 60 | 2400 | 2108 | 2014-10-24 | 9 | 5264 |
| 802560 | 787.5956820412168 | 60 | 2400 | 1019 | 2014-10-24 | 10 | 5114 |

### G. Testing Strategy

The testing strategy comprised of data analysis through aggregation, exploratory data analysis and statistical testing for both hypotheses. The aggregation comprised of finding relevant averages, minimum and maximums, and summations. The exploratory data analysis was performed using a pandas dataframe and use matplotlib. The statistical test employed was the scipy (Virtanen et al. 2020) one-way ANOVA test. To perform the statistical test, the data sets were merged with a column called Season being created which denotes the relevant season the bike rental belongs to.

## III. EVALUATION: HYPOTHESIS 1

### A. Aggregation

Statistics relating to the duration of cycle rides for the spring and autumn can be found using aggregation. Table VII shows the aggregated duration over the two time periods. There are more bike rides in the autumn time period. The average duration of bike rides is higher in spring compared to autumn. The standard deviation of the spring time period is higher, indicating a more dispersed distribution of duration.

Table VII
AGGREGATED DURATION OVER THE TIME PERIODS

| Time Period | Count | Mean | Standard Deviation | Minimum | Maximum |
|---|---|---|---|---|---|
| Autumn | 2483021 | 848.80 | 489.61 | 60 | 2400 |
| Spring | 2232250 | 883.75 | 520.06 | 60 | 2520 |

Group differences can be shown by grouping the data by the starting station. The three most common starting stations for each time period are displayed in table VIII and table IX. Interestingly across both time periods the most common starting stations are identical. The average bike ride duration's across all three of the starting stations are higher in spring. Belgrove Street and Waterloo Stations have a higher total of bike ride duration's and have more bike rides in the spring time period.

### B. Exploratory Data Analysis

A pandas dataframe of the the bike ride duration's was created for both the spring and autumn time periods. Figure 1 shows the distribution of bike rental duration for the spring and autumn time periods. The plot shows the duration of bike rentals that are less than 1000 seconds being higher for the autumn time period. There is not much separation between the seasons when looking at the durations between 1000 and 2400 seconds.

Table VIII
AGGREGATED DURATIONS FOR THE MOST COMMON STARTING STATION IN AUTUMN

| Start Station | Sum of Bike Ride Durations | Average Duration | Number of Bike Rides |
|---|---|---|---|
| Belgrove Street | 18290640 | 822.83 | 22229 |
| Waterloo Station | 16038420 | 777.47 | 20629 |
| Hype Park Corner | 18853020 | 1253.19 | 15044 |

Table IX
AGGREGATED DURATIONS FOR THE MOST COMMON STARTING STATION IN SPRING

| Start Station | Sum of Bike Ride Durations | Average Duration | Number of Bike Rides |
|---|---|---|---|
| Belgrove Street | 17315460 | 839.70 | 20621 |
| Waterloo Station | 14400540 | 793.82 | 18141 |
| Hype Park Corner | 22664040 | 1323.83 | 17120 |

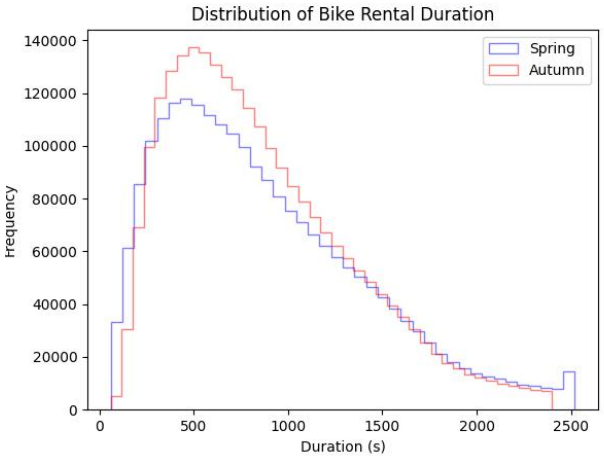


Figure 1. Bike Rental Durations

### C. Statistical Testing

A one-way ANOVA test was used to test for significance between the durations of bike rentals between autumn and spring. The test showed that there is a statistically significant difference between the groups as the p value is 0. However, whilst there is a statistical significance, there is no measure on the size of such significance.

## IV. Evaluation: Hypothesis 2

### A. Aggregation

Statistics relating to the duration of cycle rides for different traffic volumes can be found using aggregation. The total count of motor vehicles were grouped into 5 classes:

- An hour within a day where the total count of motor vehicles is between the 0 and 20th percentile are classed as 'Very Low' traffic volumes.
- An hour within a day where the total count of motor vehicles is between the 20 and 40th percentile are classed as 'Low' traffic volumes.
- An hour within a day where the total count of motor vehicles is between the 40 and 60th percentile are classed as 'Medium' traffic volumes.
- An hour within a day where the total count of motor vehicles is between the 60 and 80th percentile are classed as 'High' traffic volumes.
- An hour within a day where the total count of motor vehicles is between the 80 and 100th percentile are classed as 'Very High' traffic volumes.

Table X shows statistics of bike rentals for the different classes of traffic volume. Generally, higher durations of bike rentals occur with lower volumes of traffic. The average number of rentals is less correlated with the volume of traffic.

Table X
Bike Rental Statistics for Traffic Volumes

| **Traffic Volume** | **Average Bike Ride Duration** | **Average Number of Rentals** | **Total Bike Ride Durations** |
|---|---|---|---|
| Very Low | 878.66 | 793 | 696780 |
| Low | 842.17 | 1913.45 | 370201620 |
| High | 835.46 | 2082.72 | 1046324520 |
| Medium | 827.43 | 1908.88 | 381480420 |
| Very High | 783.29 | 1913 | 1498440 |

### B. Exploratory Data Analysis

As the count of motor vehcles and bike rentals are long-tailed distributions, as shown in figure 3, log-transformations were used to plot the variables. Figure 2 shows the distribution of the total count of motor vehicles within an hour against the total bike rental count on a log-log scale. The plot indicates that there is little correlation between the two variables.

### C. Statistical Testing

Correlation coefficients were calculated to test for correlation between bike rental statistics and the total count of motor vehicles. Table XI shows the correlation coefficients of bike rental statistics with the total count of motor vehicles. There is little correlation between all the bike rental statistics and the count of motor vehicles.

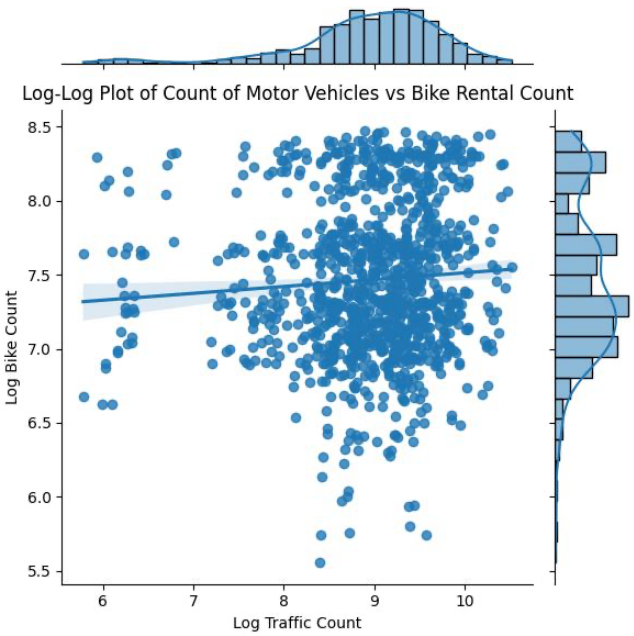


Figure 2. Log - Log plot of Count of Bike rentals vs Count of Motor Vehicles

Table XI
Correlation Coefficients of Bike Rental Statistics with the Total Count of Motor Vehicles

| Bike Rental Statistic | Correlation Coefficient |
|---|---|
| Sum of Bike Ride Duration | 0.1025 |
| Average Bike Ride Duration | 0.0185 |
| Minimum Bike Ride Duration | 0.0019 |
| Maximum Bike Ride Duration | 0.2824 |
| Number of Bike Rentals | 0.1023 |

## V. Discussion

### A. Hypothesis 1

From the findings displayed in section III, it can be concluded that hypothesis 1 1 should be rejected. This is due to the aggregated data indicating the bike ride duration is higher in spring than in autumn. The exploratory data analysis showed autumn bike rides are distributed with lower durations in comparison to the spring bike rides.

### B. Hypothesis 2

From the findings displayed in section IV, a conclusive decision on whether to accept or reject the hypothesis is not achieved. Whilst the aggregated data suggests that bike ride duration is higher when traffic volume is low, the number of bike ride rentals portrays an unclear pattern across the different traffic volumes. The exploratory data analysis and the correlation coefficients do not strongly indicate any correlation between the number of bike ride rentals and traffic volumes. Using geographical data, such as the postcode of a relevant start station, may allow for the traffic volumes to be better localised and better represent the areas traffic distribution.

## References


Albaldawi, Wafaa S, Rafah M Almuttairi, and Mehdi Ebady Manaa (2022). "Big Data Analysis for Healthcare Application using Minhash and Machine Learning in Apache Spark Framework". In: *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*. IEEE, pp. 1–7.

Cravero, Ania et al. (2022). "Challenges to Use Machine Learning in Agricultural Big Data: A Systematic Literature Review". In: *Agronomy* 12.3, p. 748.

Ebada, Ahmed Ismail et al. (2022). "Applying apache spark on streaming big data for health status prediction". In.

London, Transport for (2015). *Cycle Hire Usage Stats - 2014*. URL: https://cycling.data.tfl.gov.uk (visited on 05/02/2023).

Office, Met (2014a). *2014: A year in weather*. URL: https://blog.metoffice.gov.uk/2014/12/31/2014-a-year-in-weather/ (visited on 05/02/2023).

– (2014b). *Spring has sprung*. URL: https://blog.metoffice.gov.uk/2014/03/05/spring-has-sprung/ (visited on 05/02/2023).

Spark, Apache (2023a). *Spark Release 3.3.2: Apache Spark*. URL: https://spark.apache.org/releases/spark-release-3-3-2.html (visited on 05/02/2023).

– (2023b). *SQL*. URL: https://spark.apache.org/sql/ (visited on 05/02/2023).

Sujitha, R and B Paramasivan (2022). "Classification of Healthcare Data Using T-BMSVM in Big Data". In: *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, pp. 1–7.

Transport, Department for (2015). *Road Traffic Statistics - London region*. URL: https://roadtraffic.dft.gov.uk/regions/6 (visited on 05/02/2023).

Virtanen, Pauli et al. (2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17, pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

Zaharia, Matei et al. (2016). "Apache spark: a unified engine for big data processing". In: *Communications of the ACM* 59.11, pp. 56–65.

# APPENDIX A
# GRAPHS

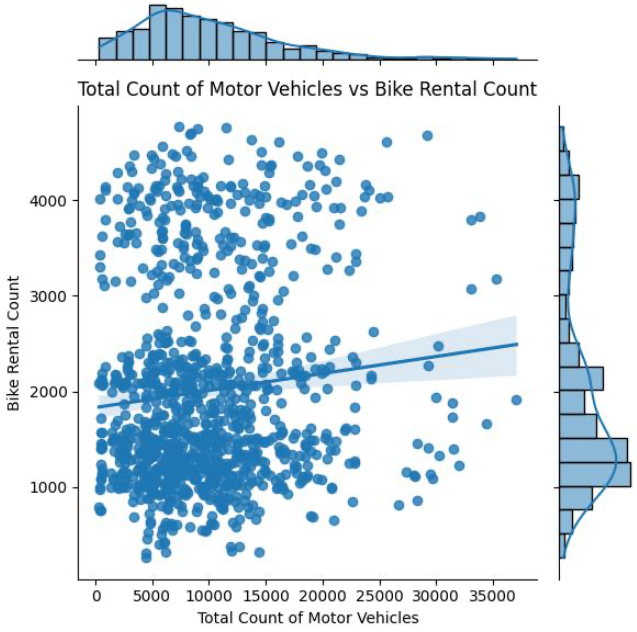


Figure 3. Count of Bike rentals vs Count of Motor Vehicles