Programme	Digital Technology Solutions Specialist Software Engineering (Integrated Degree Apprenticeship)
Module name	Data Engineering for Software Developers
Schedule term	Sept-2025
Student Reference Number (SRN)	2C7AGCYYX
Report / Assignment Title	London Bike Share August 2023 Analysis
Date of Submission (Please attach the confirmation of any extension received)	07/01/2026

Declaration of Original Work:

- I hereby declare that I have read and understood BPP's regulations on plagiarism and that this is my original work, and that I have researched, undertaken, completed and submitted in accordance with the requirements of BPP School of Technology.
- I declare that where I have used any AI tools, it was for the following reasons (highlight as appropriate):
    - To create an assignment plan
    - To create a draft
    - To correct language errors
    - Other (please describe)
        - I have not employed any generative AI tools on this assignment.
- I have copied & retained for BPP University's reference, all AI prompts used in the creation of AI content and all AI-generated responses in support of my assignment and attached relevant evidence in the appendices. I understand that I may be required to participate in a viva voce, where I will be questioned on any aspect of my assignment, including key concepts, theories, examples used, & any sources included.
- The word count, excluding contents table, bibliography and appendices, is 2177 words.

Student Reference Number: 2C7AGCYYX Date: 07/01/2026

Table of Contents

ii. How do Trip Durations and Pickup Frequencies Differ for Classic Bikes Versus Ebikes?
iii. Route Popularity Revisited
iv. Is There Any Correlation Between Bike Number and Trip Duration that Could Indicate Consistent Defects?

Project Submission

To run the project submission please upload index.ipynb from the zip file submitted & follow the instructions contained.

Data Subject Choice & Analysis Direction

Relevance to NatWest Group

NatWest Group (NWG) is currently on a journey to become more “customer focused” which is described internally as attempting to look at a ‘human scale’ consideration of the data. To avoid using real bank data we can examine a bikeshare system which can be seen as a microcosm for economies one could find across society; individual choices are influenced by numerous variables, conscious and subconscious, influenceable by the individual, as well as wider systemic influences (Corcoran et al., 2018).

By analysing a dataset like this through a human-centered lens we can begin to ask more accurate questions about what these variables are, what they say about the specific service (the fleet of bikes and their corresponding infrastructure like stands), and wider systemic considerations such as the provision of safe cycle routes away from high speeds of volumes of motor traffic, moving away from common persisting myths about barriers. (Johnson, Pearce and Schultz, 2019)

Relevance of Geospatial Data Analysis and Visualisation

Our software engineering team (SE) has identified a gap in the bank’s analysis capabilities where geospatial data is concerned. The team believes that much more could be done to incorporate location-based data, visualisations, and analysis to give a richer view of macro-level behaviors and trends, helping us to move to a more customer-sympathetic position versus looking at ‘flat’ numbers.

Chosen Datasets

To supply bikeshare data the dataset “kalacheva/london-bike-share-usage-dataset” on Kaggle was chosen. This dataset contains origin-destination (OD) data for 776,527 trips from August 2023 on the main cycle hire scheme in London.

Another Kaggle-hosted dataset, “zongaobian/london-weather-data-from-1979-to-2023”, which contains various weather metrics for the whole of London. A more granular dataset would have been preferred, for example broken down by borough, but the only sources were of unsuitable quality.

Data Quality

The primary bikeshare dataset was found to be very high quality while the weather dataset contained values of varying accuracy, providing a corresponding column for every data column denoting its accuracy.

Data Quality Dimensions

The Dimensions of Data Quality provide a framework for evaluating the quality of the datasets.

Missing & Inaccurate Values

Within the weather dataset, there are three accuracy numbers, described as follows:

- `0`: Valid data
- `1`: Suspect data
- `9`: Missing data

Additionally, there is a note that values of -9999 may appear for missing data.

Testing the entire dataset the following results were found, potentially presenting issues using the temperature and snow fields.

Key	Description	Entries with ‘0’	Entries with ‘1’	Entries with 9	% not ‘0’
Q_TX	Daily maximum temperature in 0.1°C.	321657	23499	0	7.31%
Q_TN	Daily minimum temperature in 0.1°C.	339822	5334	0	1.57%
Q_TG	Daily mean temperature in 0.1°C.	321048	23499	609	7.51%
Q_SS	Daily sunshine duration in 0.1 hours.	345156	0	0	0.00%
Q_SD	Daily snow depth in 1 cm.	322581	0	22575	7.00%
Q_RR	Daily precipitation amount in 0.1 mm.	345156	0	0	0.00%
Q_QQ	Daily global radiation in W/m².	343476	0	525	0.15%
Q_PP	Daily sea level pressure in 0.1 hPa.	345072	0	84	0.02%
Q_HU	Daily relative humidity in %.	343959	0	1197	0.35%
Q_CC	Daily cloud cover in oktas.	344736	42	378	0.12%

Fortunately for the purposes of this analysis, between and including the dates 01/08/2023 00:00:00 and 31/08/2023 23:59:59 the entire dataset was quality ‘0’, meaning no issues at all.

Completeness & Validity

None of the values found are of the wrong type, there were no missing values.

Uniqueness & Timeliness

There were no noted instances of duplicated rows. All dates appear to be valid, no instances were noted of end dates being before start dates.

Accuracy & Consistency

All columns were of a consistent, machine-readable format except for the readable station names, however this is not significant as we also have supplied a unique station number identifier.

It was noted that ride durations are accurate to the millisecond while the timestamps are accurate only to the minute, meaning a slight discrepancy.

Questions Posed

What Portion of Trips are False Starts?

Sometimes faults are not noticed until the rental starts as the bike is locked until then. These ‘false starts’ may show up in the data as very short rides with the same start and end station.

Do We Note Any Correlation Between Weather Events and Trip Characteristics?

Weather conditions typical of their region are less of a determinant in that region, i.e. hot weather countries are less affected by heat, very wet countries are less affected by rain, etc (Bean et al., 2021).

What effect do particular weather events have on ridership? What is the difference across various routes between these changes? Can we colocate this with the provision of safe cycle infrastructure?

Methodology

Mapping Station Names to Coordinates

To visualise the data in a geospatially meaningful sense it was necessary to join the station names / numbers to additional information, namely coordinates.

A series of files prefixed with setup_{number} perform each stage of this data query and cleaning sequentially, to allow stages to be re-run without unnecessary API calls.

The TFL Unified API

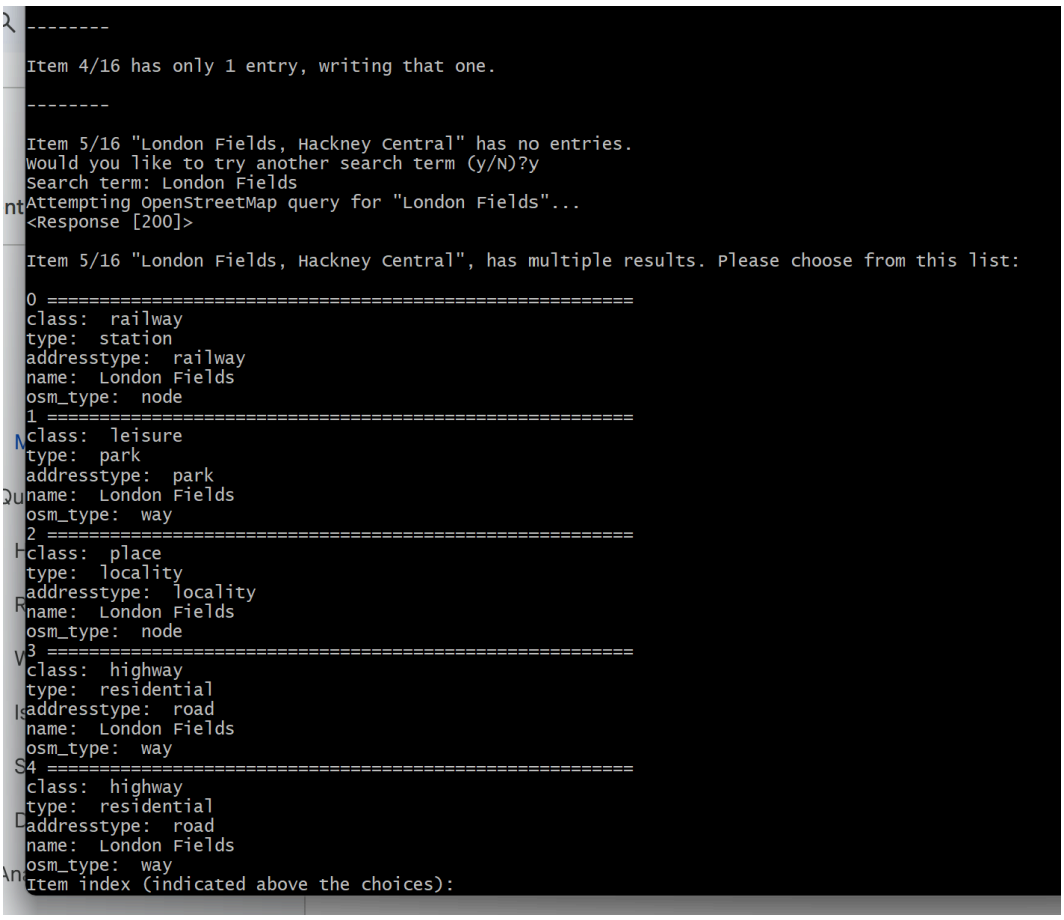
Various options to populate the station data exist, for the highest accuracy the TFL Unified API was chosen as the preferred primary source.

Resolving Unfound Stations with the OSM API

To supplement the TFL API's lack of historic data, the Open Street Map (OSM) API provides a similarly high quality of location data, with the one caveat that it may return more than one result. The file setup_3_attempt_requery_empty_station queries the OSM API for missing entries from the TFL query.

CLI to Select from Multiple Stations

The step 2 and 4 files respectively filter the responses from the TFL and OSM APIs, presenting the user with a CLI to guide the choice for selecting stations when multiple options are found.



Handling Null Stations

For any values still missing two additional files allow the user to manually enter a search term for either TFL or OSM. If the results still cannot be reconciled the station is omitted from the final combined list of all stations.

User Error Selecting Substitutions

One drawback of involving user intervention to reconcile the data and potentially bolsters an argument to simply exclude stations which the TFL API cannot confidently reconcile.

Working with Time

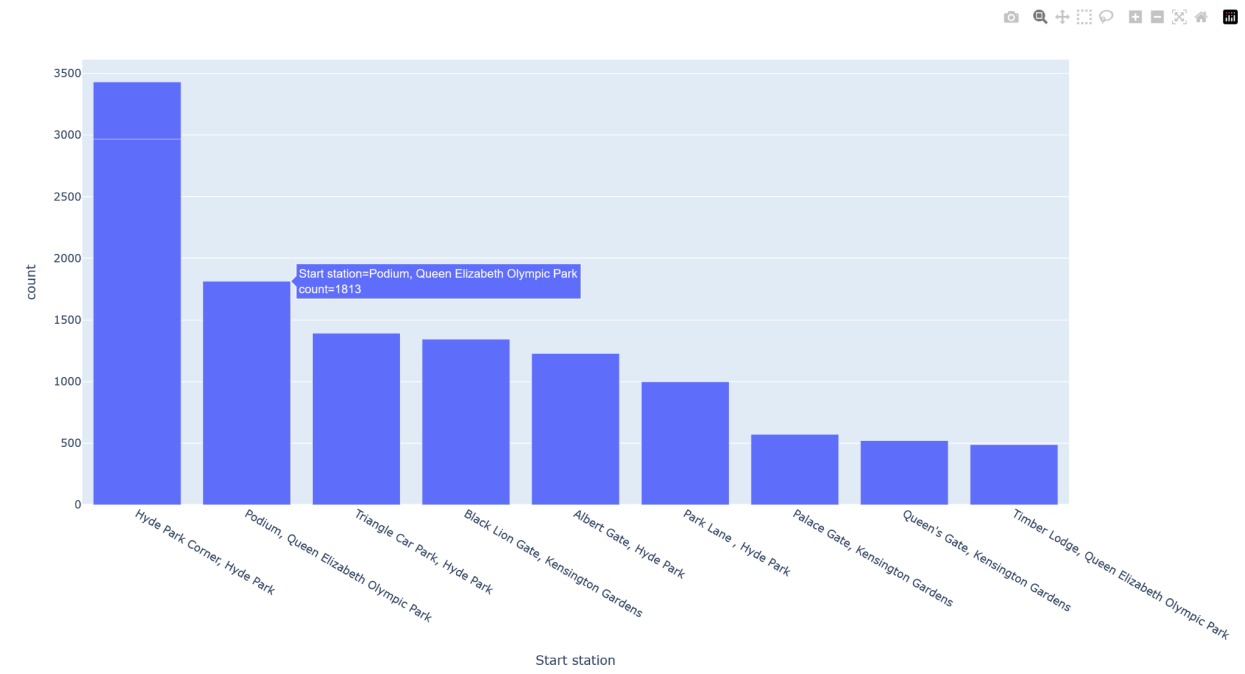
The dataset is very fine-grained containing timestamps that are accurate to the second meaning that for many types of analysis some kind of date time aggregation is required to work effectively with the data.

```
# Create a new column for a datetime timestamp on the weather data.
df_weather['date_formatted'] = pd.to_datetime(df_weather['DATE'], format='%Y%m%d')

# Ensure Start date is parsed to datetime
df_bike_data['Start date'] = pd.to_datetime(df_bike_data['Start date'])

# Extract date (day-level)
df_bike_data['date_hour'] = df_bike_data['Start date'].dt.floor('h')  # type: ignore
df_bike_data['date_day'] = df_bike_data['Start date'].dt.floor('d')  # type: ignore

df_merged = pd.merge(
    left=df_bike_data,
    right=df_weather,
    left_on=['date_day'],
    right_on=['date_formatted'],
)
```

Most Common Repeated Trips

By aggregating over the dataset using a combination of Panda's 'agg' function and calculations applied after a 'groupby', visualisations were produced showing the top ten routes, defined by origin-destination station, quantified by frequency. 8 of the top 10 (13 of the top 20) are trips with the same start-end station.

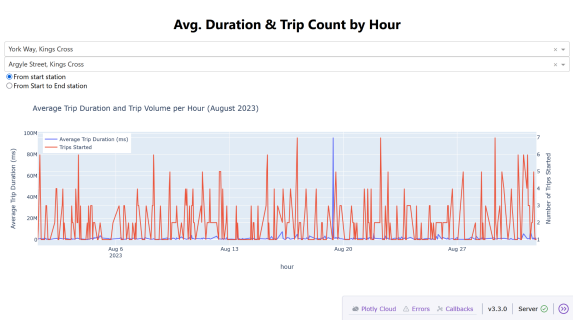
```python
# Create a new df which contains a list of all unique trips (origin-destination pairs) and the quantity counts for each.
df_unique_od = (
    df.groupby(['Start station', 'End station'])
    .size()
    .reset_index()
    .rename(columns={0: 'count'})
)

df_sorted = df_unique_od.sort_values('count', ascending=False).reset_index().head(x_axis_quant)
print(df_sorted)

fig = px.bar(df_sorted, x='Start station', y='count')

fig.show()
```

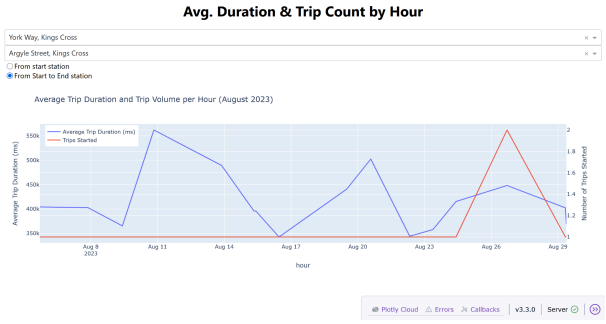


Next, a modification on this aggregation was used to total the number of trips started, per start-station.

Trip Duration Average by Day or Hour

Using a Dash app the average trip aggregation and number of pickups against each other.

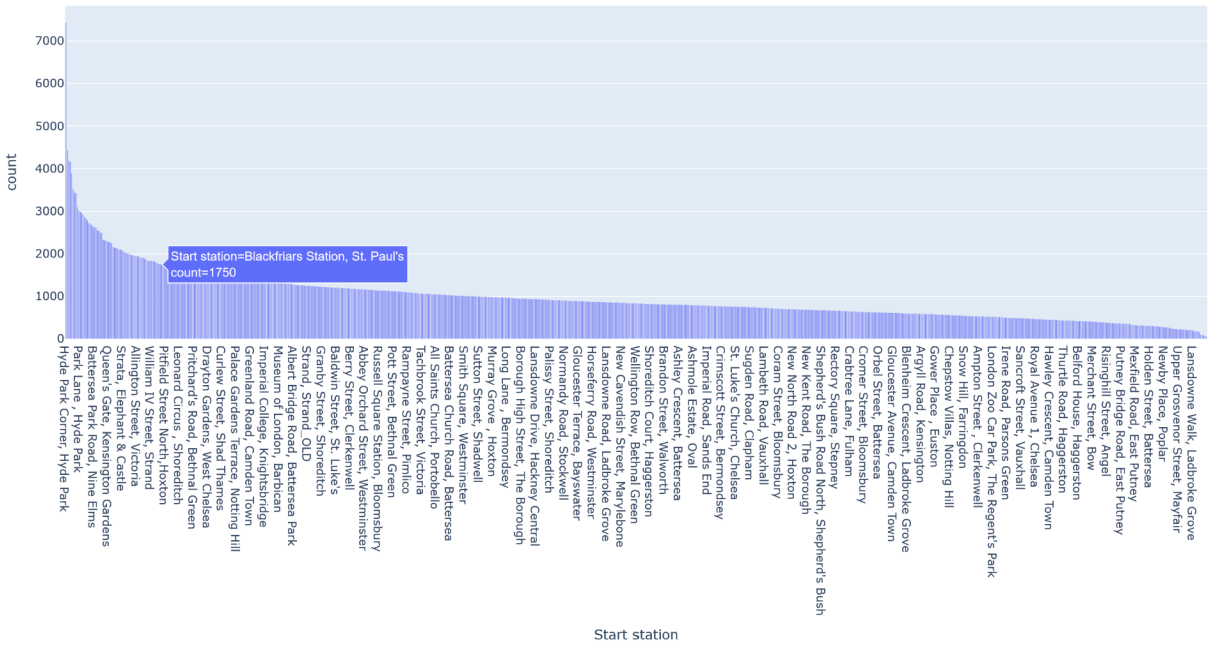




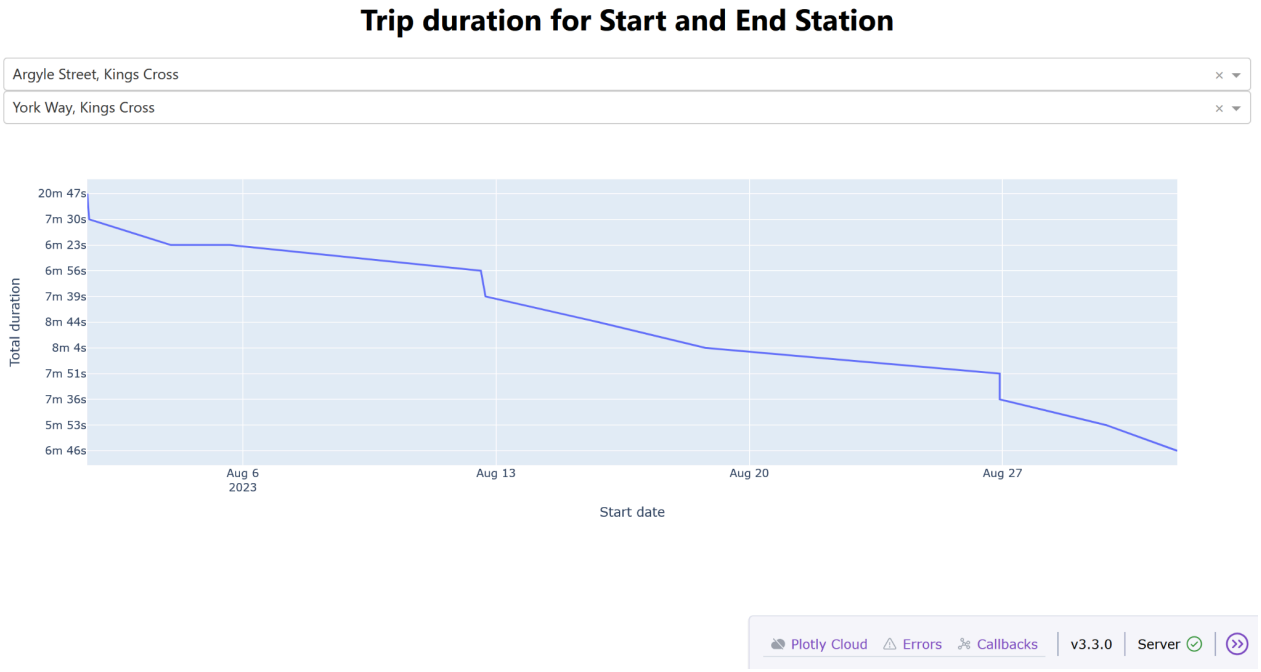
Stations Ranked by number of Pickups

Producing a bar graph for all start stations shows a smooth discrepancy in station popularity with some dramatic outliers.

```python
# Create a new df which contains a list of all unique trips (origin-destination pairs) and the quantity counts for each.
df_sorted = (
    df.groupby(['Start station'])
    .size()
    .reset_index()
    .rename(columns={0: 'count'})
    .sort_values('count', ascending=False)
    .reset_index()
)
print(df_sorted)

fig = px.bar(df_sorted, x='Start station', y='count')

fig.show()
```



Following on from this, an interactive dashboard built with Dash provided a way to view the average duration of trips, by start and optional end station.



Circular Trips and False Starts

Running a check for items with the same start and end station revealed that 796 station pairs, nearly the entire station list, occurred in the dataset. Running the same check against the entire dataset revealed 39,790 items, just over 5% of all trips. Could these all be ‘false starts’?

Taking the example again of York Way, Kings Cross, an aggregation to filter for any trip less than 60 seconds results in only 2 entries.

```python
kx = "York Way, Kings Cross"

# Find a subset of data matching YW, Kings Cross as a start station.
# This will for the basis of later queries examining the validity of this subset.
print("All from KX: ")
print(df.loc[df["Start station"] == kx])

# We assume that trips under 60 seconds are potential "false starts".
# Select how many fall into this category.
print("All from KX under 1 minute: ")
print(
    df.loc[df["Start station"] == kx]["Total duration (ms)"]
    .apply(lambda x: x < 60000)
    .sum()
)
```

```
All from KX:
         Number       Start date  Start station number          Start station  ... Bike number  Bike model Total duration  Total duration (ms)
542     132825746    8/1/2023 5:43                300235  York Way, Kings Cross  ...       57468     CLASSIC         7m 29s               449042
987     132826192    8/1/2023 6:36                300235  York Way, Kings Cross  ...       11656     CLASSIC         12m 9s               729784
1068    132826274    8/1/2023 6:41                300235  York Way, Kings Cross  ...       60335  PBSC_EBIKE        11m 51s               711064
3418    132828670    8/1/2023 8:00                300235  York Way, Kings Cross  ...       60497  PBSC_EBIKE        13m 21s               801433
3727    132828991    8/1/2023 8:07                300235  York Way, Kings Cross  ...       21448     CLASSIC        19m 44s              1184724
...           ...             ...                   ...                    ...  ...         ...         ...             ...                  ...
768763  133616580  8/31/2023 17:54                300235  York Way, Kings Cross  ...       50867     CLASSIC         7m 57s               477486
769168  133616993  8/31/2023 18:02                300235  York Way, Kings Cross  ...       57724     CLASSIC         4m 20s               260472
769650  133617484  8/31/2023 18:11                300235  York Way, Kings Cross  ...       30178     CLASSIC        10m 35s               635495
772541  133620434  8/31/2023 19:23                300235  York Way, Kings Cross  ...       53634     CLASSIC         33m 8s              1988306
774603  133622557  8/31/2023 21:00                300235  York Way, Kings Cross  ...       54843     CLASSIC         8m 10s               490528

[761 rows x 11 columns]
All from KX under 1 minute:
2
```

However, is this generalisable to the whole dataset? Running the same aggregation for trips less than 60 seconds on the 39,790 same OD entries yielded 4660 entities, just under 12%.

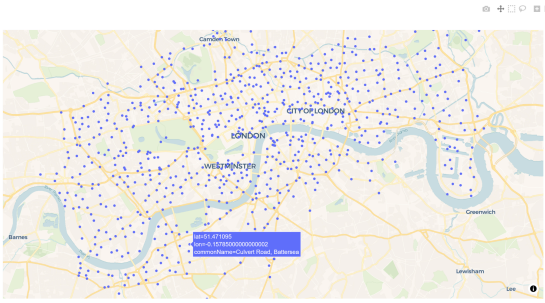
```python
# How many trips are circular in the entire dataset?
print("Full DF Start station and End station are the same: ")
print(df.loc[df["Start station"] == df["End station"]].reset_index())

# Of this, how many are potential "false starts"?
print("Full DF Start station and End station are the same and less than 60 seconds: ")
print(
    df.loc[df["Start station"] == df["End station"]]["Total duration (ms)"]
    .apply(lambda x: x < 60000)
    .sum()
)
```

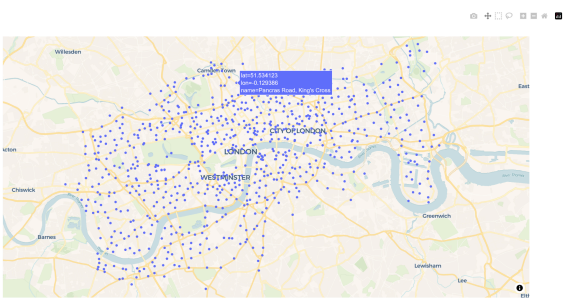
```
Full DF Start station and End station are the same:
       index     Number       Start date  Start station number  ... Bike number  Bike model  Total duration Total duration (ms)
0         18  132825207    8/1/2023 0:05                200215  ...       21222     CLASSIC           8m 3s              483259
1         33  132825225    8/1/2023 0:09                  1020  ...       55324     CLASSIC         17m 25s             1045195
2         51  132825243    8/1/2023 0:14                300079  ...       60612  PBSC_EBIKE          9m 11s              551569
3         53  132825245    8/1/2023 0:14                  1108  ...       53790     CLASSIC         36m 24s             2184240
4         70  132825260    8/1/2023 0:22                300021  ...       52777     CLASSIC         15m 43s              943350
...      ...        ...             ...                   ...  ...         ...         ...             ...                 ...
39785  776486  133624535  8/31/2023 23:53                  1151  ...       23119     CLASSIC          4m 23s              263052
39786  776498  133624545  8/31/2023 23:54                  1062  ...       40332     CLASSIC         25m 48s             1548335
39787  776504  133624548  8/31/2023 23:55                  1062  ...       30713     CLASSIC         25m 35s             1535701
39788  776507  133624554  8/31/2023 23:56                200195  ...       53293     CLASSIC         13m 18s              798708
39789  776514  133624562  8/31/2023 23:57                300252  ...       56854     CLASSIC     13h 34m 32s            48872637

[39790 rows x 12 columns]
Full DF Start station and End station are the same and less than 60 seconds:
4660
```

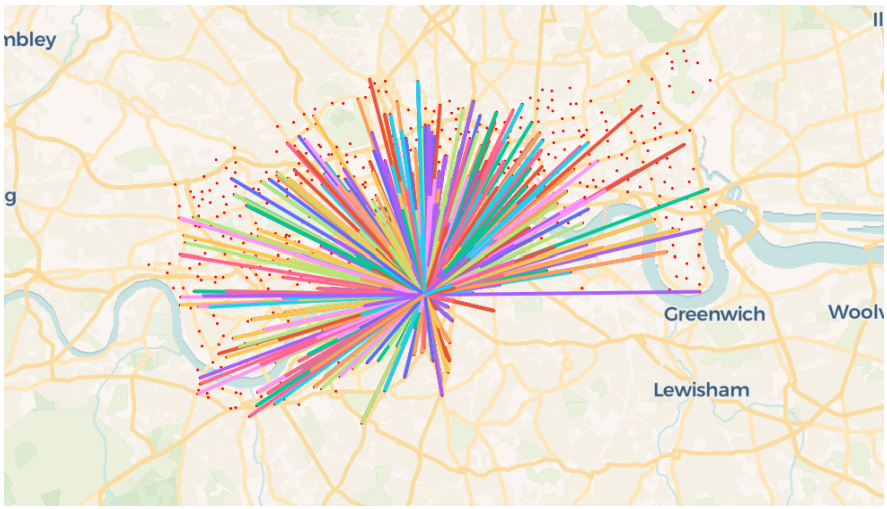
Visualising Stations & Trips Geospatially

Map Options

Using Plotly's scatter_map function a trace was added to draw lines between stations representing the trips taken but the mass of lines was so dense it was unusable. Switching to visualise only one start station helped but it was clear more was needed to narrow the granularity of the data.

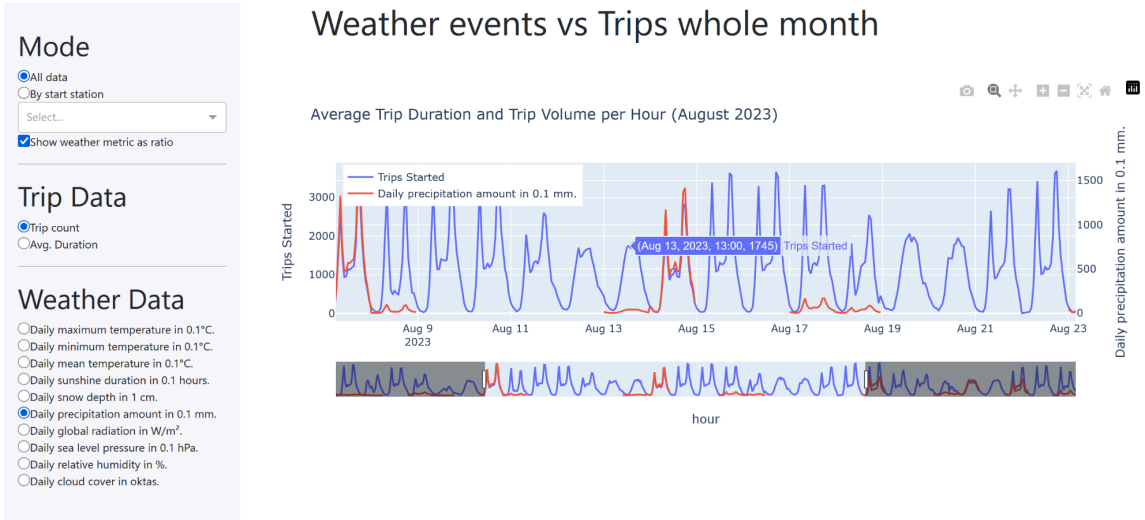






Weather Data Join & Dash App

A Dash application was produced to plot the trip quantity or average ride duration against data from the weather dataset, again with the option to view the entire dataset or to choose a start station.



Sliding Time Filter

A custom time slider was introduced, calibrated to days, which allowed multiple visualisations to be used. The map was re-introduced for start-station mode which became more useful when a short time period, for example the days of a working week.

Logistic regression was used to compare various weather events to a set of selected stations, chosen from the previous lists of most to least popular.

Weather events vs Trips whole month

Mode

- All data
- By start station

Euston Road, Euston

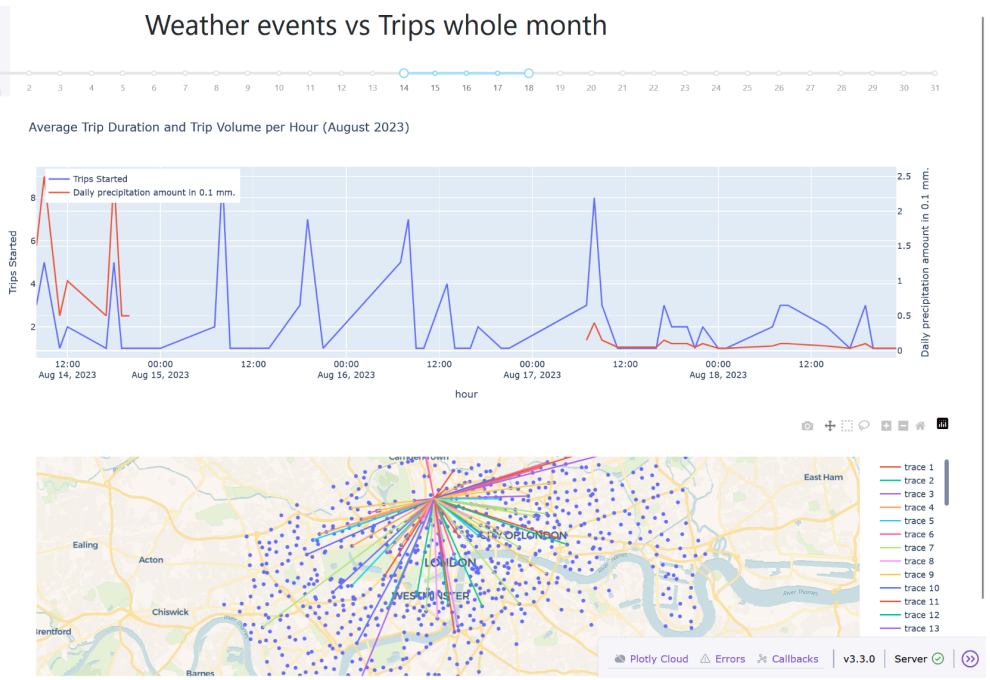
Show weather metric as ratio

Trip Data

- Trip count
- Avg. Duration

Weather Data

- Daily maximum temperature in 0.1°C.
- Daily minimum temperature in 0.1°C.
- Daily mean temperature in 0.1°C.
- Daily sunshine duration in 0.1 hours.
- Daily snow depth in 1 cm.
- Daily precipitation amount in 0.1 mm.
- Daily global radiation in W/m².
- Daily sea level pressure in 0.1 hPa.
- Daily relative humidity in %.
- Daily cloud cover in oktas.



Weather events vs Trips whole month

Mode

- All data
- By start station

Import Dock, Canary Wharf

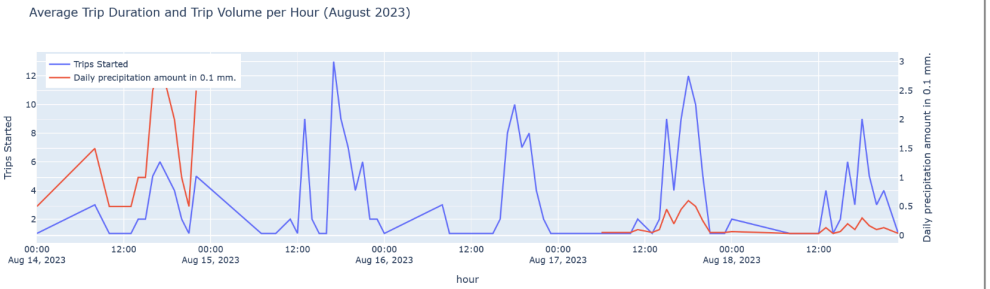
Show weather metric as ratio

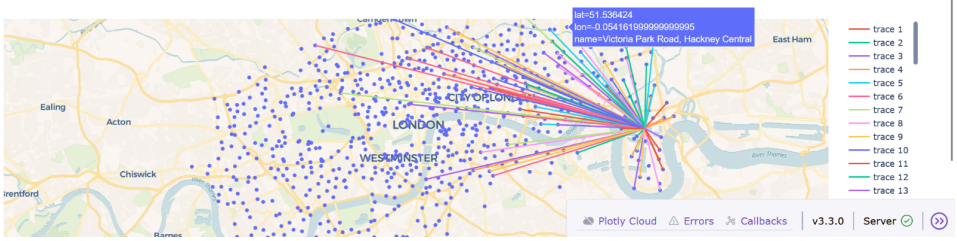
Trip Data

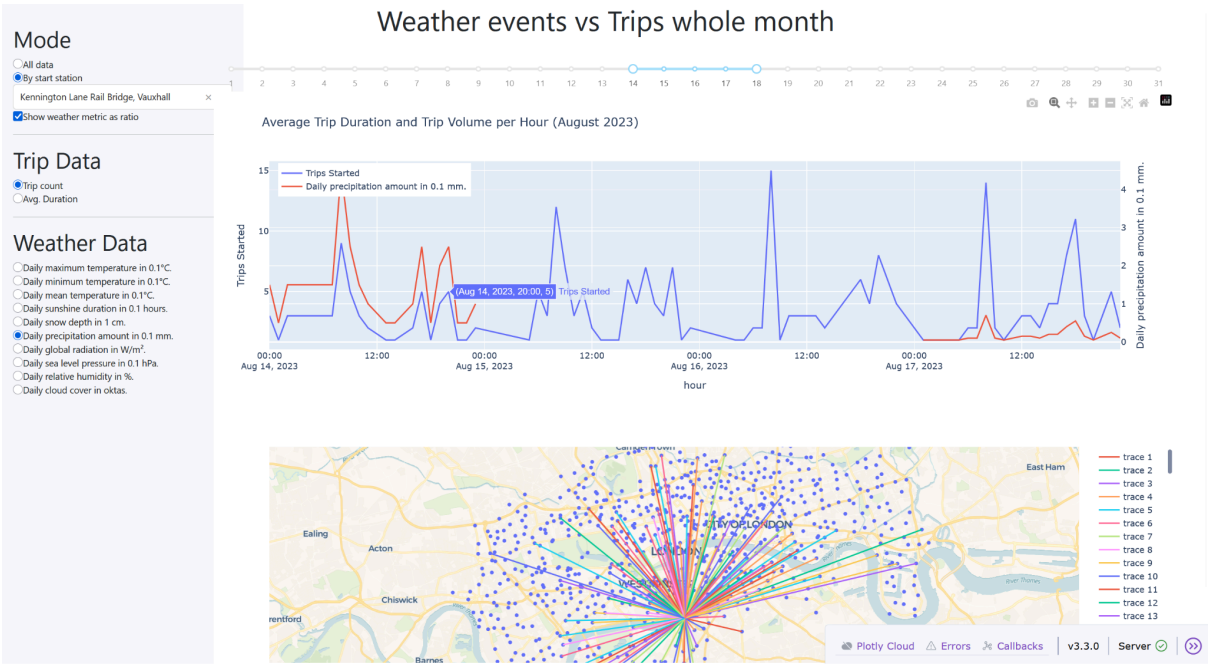
- Trip count
- Avg. Duration

Weather Data

- Daily maximum temperature in 0.1°C.
- Daily minimum temperature in 0.1°C.
- Daily mean temperature in 0.1°C.
- Daily sunshine duration in 0.1 hours.
- Daily snow depth in 1 cm.
- Daily precipitation amount in 0.1 mm.
- Daily global radiation in W/m².
- Daily sea level pressure in 0.1 hPa.
- Daily relative humidity in %.
- Daily cloud cover in oktas.







Results

Circular Trips

It is a reasonable conclusion that a decent proportion of these trips are potential false starts, however it is safe to say that the rest are recreational rides or other self-contained rides, perhaps rides where the parking function is used to keep a bike while you visit a location, until it's needed again.

Trip Frequencies

For the majority of stations selected and across the entire dataset, a trend merges with trip quantities peaking in the middle of the working week in a bell curve.

Average trip durations also fluctuate between about 1 and 4km) but with a far less pronounced trend, slightly increasing during commuting hours. Curiously, large spikes are seen during the nights, with the average duration often spiking to 5+km or in some cases rising more dramatically, for example 17.4km on the 28th at 03:00.

This could be explained by two effects; on the one hand a lower ridership number overnight will allow some longer trips to move the average more significantly than during the day, on the other hand, these could be trips that would otherwise be taken on public transport which is limited at night.

This capacity of the bikeshare system to absorb missing capacity from public transport at night could give further justification to designing cycle infrastructure for nighttime use specifically.

Weather Impacts on Travel

It was found that the larger precipitation events had somewhat on the system's utilisation, visible by comparing changes to the mid-week bell curve. However, this trend varies significantly by looking at individual stations, with many being unaffected or even increasing in utilisation.

This same effect can be seen when comparing the three temperature metrics, solar radiation, and humidity, with some stations being affected negatively and others increasing or showing no notable effect.

This supports existing evidence that suggests the relationship between rain and decreasing cycling rates is affected by more factors than simple correlation.

Pickup seemed to frequently increase with increased cloud cover which would seem to contradict the sentiment that cycling is only correlated with 'good' weather.

Ethical Considerations

Station Density

Whereas a commute by rail will take an individual through stations used by many thousands of people, the granularity of bikeshare stations means that their chosen stops are more identifiable with their ultimate start and end destinations, more comparable to a bus stop.

As can be seen on the map of stations, the density and consistency of stations within the operational area is fairly consistent. This means that preference for a particular start station could be narrowed down to a very small radius for an individual trip.

Repeat Trips De-anonymising Data

With some of the analysis performed in this report clear repeating trends are immediately obvious. While anonymised, these represent metadata pertaining to real people, combined with other metadata it may be possible to identify individuals and / or reveal sensitive details of their daily patterns.

Conclusion

Further Analysis

Can We Correlate Any Other Markers of Deprivation to Utilisation Rates?

By joining other datasets such as census data and Areas of Multiple Deprivation, can we find more health and inequality indicators which impact bikeshare utilisation (*24/62 Active Travel facilitators and barriers within different populations*, 2024), (Malden *et al.*, 2024).

How do Trip Durations and Pickup Frequencies Differ for Classic Bikes Versus Ebikes?

It has been demonstrated that personal ebikes rapidly change the accessibility and appeal of cycling to those not already engaged (Castro et al., 2019). How does this trend transfer to sharebikes?

Route Popularity Revisited

Further work could be done to group and compare routes and stations at popular locations. Could we then co-locate our findings with cycle infrastructure.

Is There Any Correlation Between Bike Number and Trip Duration that Could Indicate Consistent Defects?

If there are a significant number of false starts, is there any trend between stations or bike numbers that could indicate repeat defects?

References

Corcoran, J. *et al.* (2014) ‘Spatio-temporal patterns of a Public Bicycle Sharing Program: the effect of weather and calendar events’, *Journal of Transport Geography*, 41, pp. 292–305. Available at: https://doi.org/10.1016/j.jtrangeo.2014.09.003.

Malden, S. *et al.* (2024) ‘Identifying Barriers and Facilitators to Active Travel Infrastructure Usage Amongst Under-Represented Population Groups in the United Kingdom: A Rapid Systematic Review’, *Active Travel Studies*, 4(1). Available at: https://doi.org/10.16997/ats.1510.

Bean, R., Pojani, D. and Corcoran, J. (2021) ‘How does weather affect bikeshare use? A comparative analysis of forty cities across climate zones’, *Journal of Transport Geography*, 95, p. 103155. Available at: https://doi.org/10.1016/j.jtrangeo.2021.103155.

Johnson, H., Pearce, M. and Schultz, R. (2019) ‘Common Misconceptions of Active Travel Investment A Review of the Evidence LCWIP Strategic Support’. Sustrans. Available at: https://www.walkwheelcycletrust.org.uk/media/5224/common-misconceptions-of-active-travel-investment.pdf.

24/62 Active Travel facilitators and barriers within different populations (2024). Available at: https://www.nihr.ac.uk/2462-active-travel-facilitators-and-barriers-within-different-populations (Accessed: 4 November 2025).

Castro, A. *et al.* (2019) ‘Physical activity of electric bicycle users compared to conventional bicycle users and non-cyclists: Insights based on health and transport data from an online survey in seven European cities’, *Transportation Research Interdisciplinary Perspectives*, 1, p. 100017. Available at: https://doi.org/10.1016/j.trip.2019.100017.