Assignment 3: Final Report

LSE Employer Project

22 August 2022

Team 4: Megan Bilas, Hazel Chan, Akira Leyow, Ted Malumbe, Li Chen Zhou

GitHub Link to scripts:
https://github.com/hazz292/Cycling_in_London_Team_4

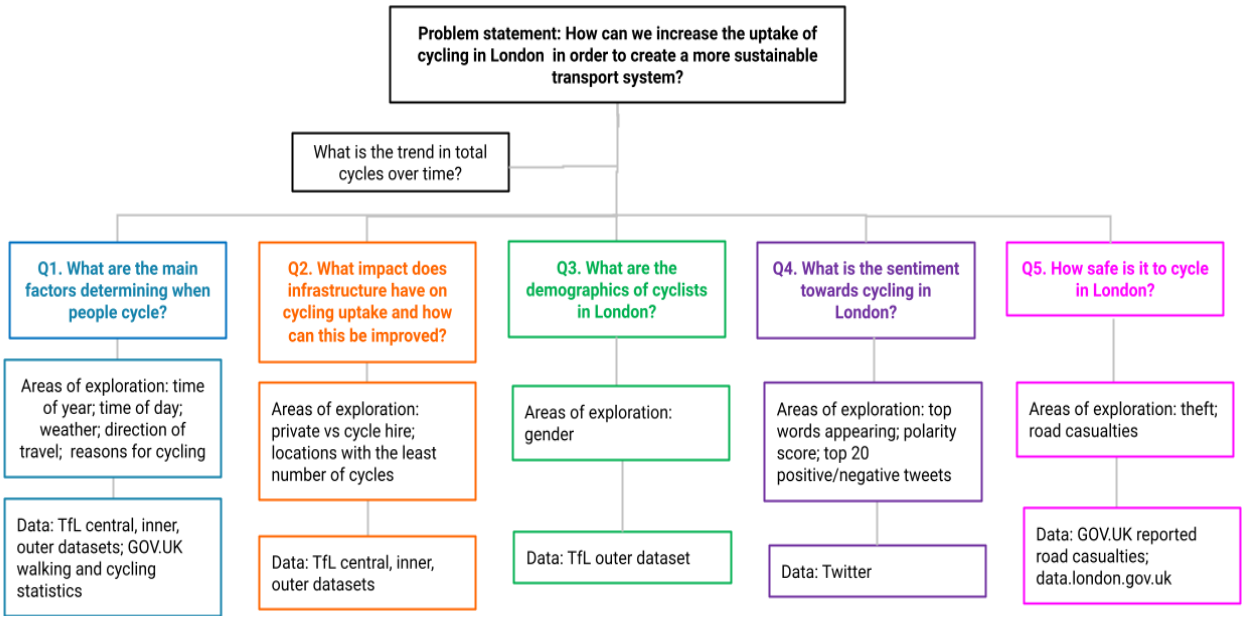
Table of contents

Background/context

According to the Mayor's Transport Strategy 2018, the Greater London Authority aims for 80% of all journeys in London to be made by foot, cycling or public transport by 2041 with a vision to create a fairer, greener, healthier and more prosperous city.[1]

To address the cycling component of this strategy, we have examined the business questions outlined in Figure 1.

Figure 1. Business case

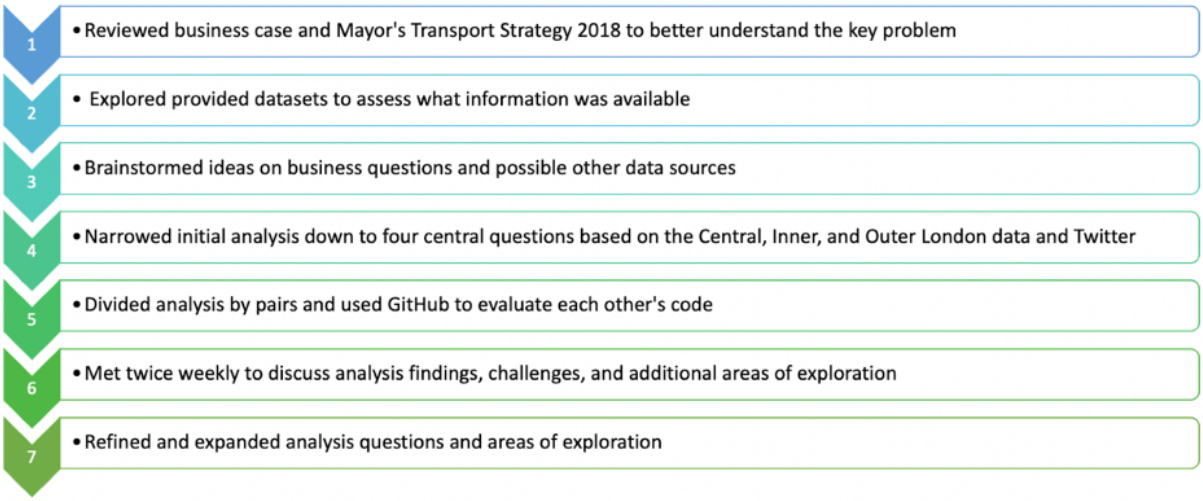


Project development

The development process is summarised in Figure 2 and described in more detail below.

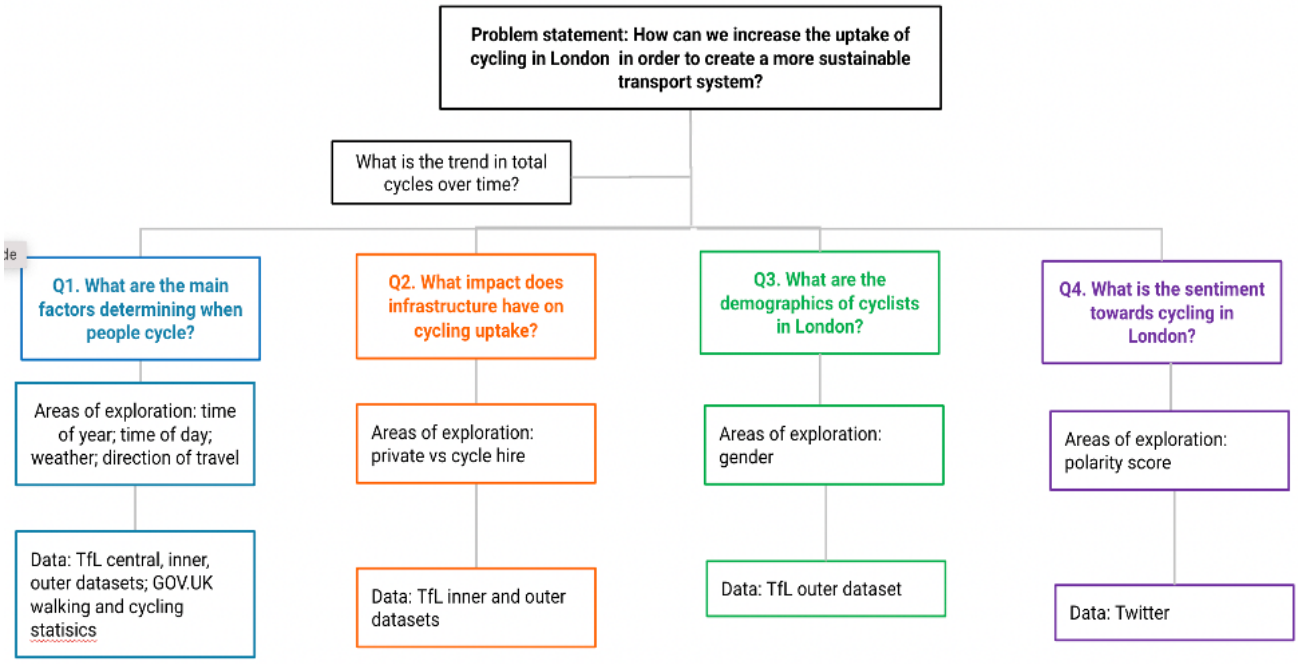
[1] Mayors Transport Strategy 2018

Figure 2. Development process



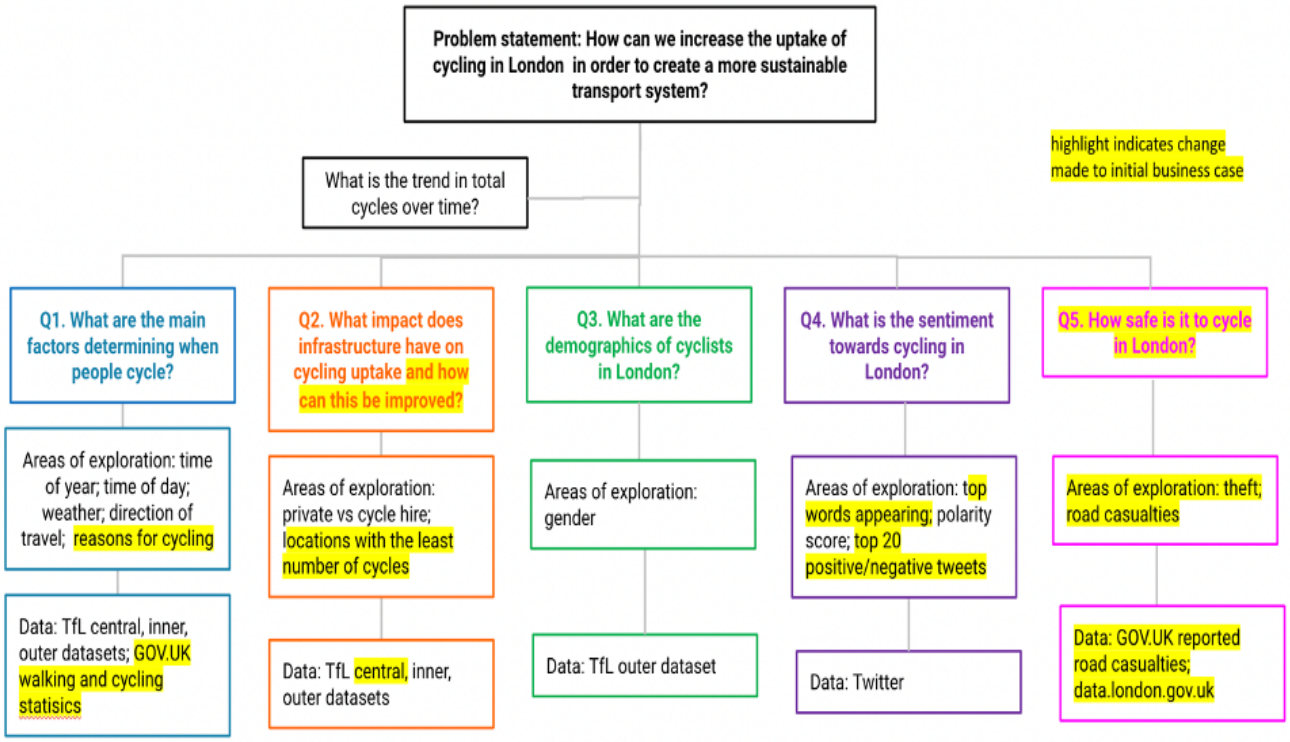
After reviewing the Mayor's Transport Strategy 2018 and exploring the London datasets to understand what information was available, we began looking into the initial business questions outlined in Figure 3. These questions were chosen based on feasibility (e.g. data availability and timelines) and to provide us with sufficient background information on the current state of cycling in London.

Figure 3. Initial business case



Following analysis of the initial business questions, we expanded our scope into additional areas of exploration, investigated external data sources, and added another business question about safety. This question was added after our analysis showed that there were significantly more male than female cyclists, which we hypothesised could be due to safety concerns. Changes are highlighted in Figure 4.

Figure 4. Expanded business case



Challenges

We encountered challenges throughout our analysis process but used our bi-weekly team and weekly stakeholder meetings to brainstorm solutions. Details on challenges and associated decisions/solutions are provided in Figure 5.

Figure 5. Challenges and decisions/solutions

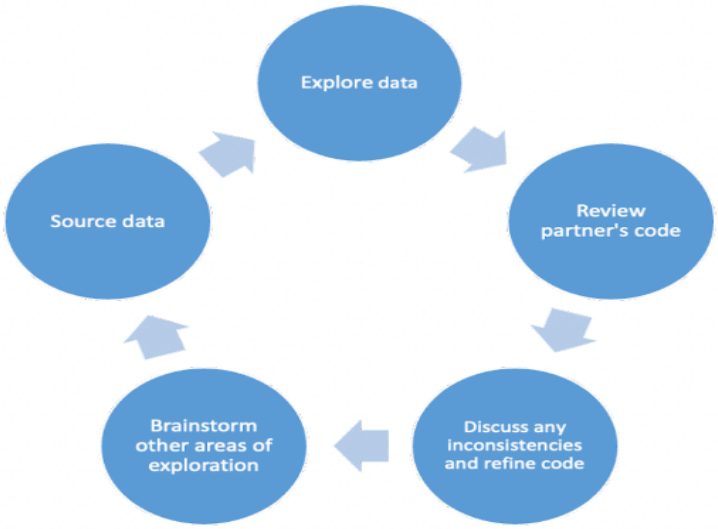
Business Question	Challenges	Decisions/Solutions
Q1. What are the main factors determining when people cycle?	When analysing the aggregated data (central, inner, outer London), the trends are thrown off by the central London dataset due to the higher number of observations in this dataset in comparison to inner and outer London. Data was available on reasons for cycling but the format of the file could not easily be read into Python.	Decided to split the analysis by the three regions in London This is in line with TfL's spatial strategy of looking at each of the regions individually due to the different concerns in each region. Reorganised data into a simplified csv file in Excel.
Q2. What impact does cycling infrastructure have on cycling uptake?	Only inner and central london has data on private vs cycle hires, unable to compare with outer london. Lack of data on categorising location to draw further correlation.	Conducted analysis separately for each area of london to examine cycles between private vs hire. Used a common variable "total cycles" to identify roads used least for cycling as areas for infrastructure improvement.
Q3. What are the demographics of cyclists in London?	The dataset did not have much demographic data beyond gender and it would have been better to also leverage income, age, employment status and race.	Carried out as much bivariate analysis as the dataset would allow. Since even attempts to gather external data were met with paywalls.
Q4. What is the sentiment towards cycling in London?	Geolocation can only be used to find trends per city. When searching for "#cycle", worldwide results were produced rather than those specific to London, England.	An alternative method for searching for several keywords was used (Tweepy). The search criteria was expanded to "london uk cycle" to narrow results to London, England.

Business Question	Challenges	Decisions/Solutions
Q5. How safe is it to cycle in London?	Data was available on cyclist accidents but the format of the file could not easily be read into Python.	Reorganised data into a simplified csv file in Excel.

Analysis process and quality assurance

We identified coding errors as a high priority risk early on in project development. To counter this risk and ensure equal distribution of coding responsibilities, we adapted a pairwise approach in which partners would analyse the same question and use GitHub to compare scripts and outputs. This approach aided us in cleaning decisions and helped to pinpoint errors. Our analysis approach is summarised in Figure 6.

Figure 6. Analysis process



Technical overview of the code / Visualisation considerations

We agreed to use the same coding language (Python) to facilitate script sharing and easily identify any differences in analytical procedures. An explanation of the libraries and packages we used is provided in Figure 7.

Figure 7. Libraries and packages used

Library/Package	Use	Analyses
Pandas	Data manipulation and analysis; explore datasets; understand data structure, data types, and variables available	All
Numpy	Statistical calculations	All
Matplotlib	Plotting and visualisation	All
Seaborn	Plotting and visualisation	All
Statsmodels	Perform time-series analysis and forecast Used to get ordinary least squares (OLS)	Time-series forecast Q1 – regression analysis
Sklearn	Perform time-series analysis and forecast Used to calculate simple linear regression	Time-series forecast Q1 – regression analysis
NLTK	Used to convert text scraped from Twitter into tokens and to generate a frequency distribution of the most-used words	Q5
Tweepy	Used to scrape tweets off of Twitter	Q5
WordCloud	Used to generate word cloud from tweets	Q5
TextBlob	Used to process text from Twitter and perform sentiment analysis	Q5
import warnings	To ignore any errors that may have arised	Q1, Q3, Q4, Q5
Math	To calculate square root	Time-series forecast
Pylab	Determine best fit parameters for SARIMA model	Time-series forecast
Itertools	Perform time-series analysis and forecast	Time-series forecast
Joblib	Perform time-series analysis and forecast	Time-series forecast

In order to maximise the amount of data available, cleaning was conducted on a question-by-question basis. Missing values were filtered out of visualisations only after the key variables had been isolated.

Certain analyses required additional considerations. These are outlined by business questions below.

What is the trend in total cycles over time?

Total cycles per year in London

Data for central, inner, and outer London were concatenated and then segmented by hue in order to plot them on the same chart. This choice was made in order to help the viewer quickly see the differences between the datasets over time.

Predictions

Time-series analysis was conducted on all three datasets in order to apply the SARIMA model for forecasting. Due to the pandemic, the trends shown in the datasets showed drastic changes for total cycles. Therefore, training and testing datasets have been split by pre-pandemic and pandemic time periods. Next, the seasonal decompose method, determining rolling statistics, and Augmented Dickey-Fuller Test were used to show that all three datasets showed an overall slight increasing trend, seasonality by months, and stationarity, which satisfy the requirements for the SARIMA model. A grid-search parameter was applied to determine the best parameter option and maximise the accuracy of the model.

Q1. What are the main factors determining when people cycle?

Time of year, time of day, weather, direction of travel

The different regions of London were plotted on the same chart when examining time of year, time of day, and direction of travel to allow the viewer to quickly see the differences between the regions in terms of total cycle count and scale.

There were many different values for 'Weather' across all datasets. Therefore, the data was filtered by weather conditions with 1000 or more observations. Due to differences in top weather conditions present in each of the regional datasets, separate plots were generated for each of the regions.

Reasons why people cycle

The data for this analysis was drawn from the GOV.UK Walking and Cycling Statistics page[2] (file CW0302). The original file was converted from an ODS file to a simplified CSV file using Excel and then read into Python for analysis. A simple line chart was produced to show changes over time and convergence in leisure and travel from 2019-2020.

Ordinary Least Squares (OLS) analysis

Regression tables were produced using ordinary least squares (OLS) to determine the statistical significance (measured by the p-value) and variance in total cycles (measured by the adjusted R-squared) that could be explained by key factors in the London datasets. Due to the more

[2] https://www.gov.uk/government/statistical-data-sets/walking-and-cycling-statistics-cw

technical nature of this analysis, results were included in output appendices but excluded from the main presentation/report.

Q2. What impact does infrastructure have on cycling uptake and how can this be improved?

Pandas was used to concatenate inner and central dataset to ensure a consistent way of cleaning. Then, subsets were created using groupby() and summed all cycle counts. Afterward, sort_values() was used to draw insights based on the highest number of private, cycle hires, total cycles in each direction and lowest number of cycles in each location. Seaborn was used to create bar plot and line plot as visualisations based on aggregated tables to compare between areas. Theme and font size were adjusted larger and removed x and y axis labels to eliminate clutter for final slides presentation.

Q3. What are the demographics of cyclists in London?

The biggest challenge was quantifying variables, such as weather, against gender to accurately draw conclusions. This was similar with variables such as Location and SiteID. The use of external data was considered but the datasets were communicating different years with different time buckets that would have made the visualisation and recommendations incorrect.

Q4. What is the sentiment towards cycling in London?

Due to challenges in isolating tweets by geocode (this could be done for city trends but not for tweets), the search criteria was altered to isolate tweets specific to “london uk cycle.” It is worth noting that there may be other ways of isolating tweets by geocode that we did not have time to investigate within the timeframe of this analysis.

Duplicate and re-tweets were filtered out of the analysis in order to get a more varied view on tweets.

Multiple visualisations were produced - including a word cloud, sentiment plots, a frequency plot, and tables of the top 20 positive/negative tweets.

Q5. How safe is it to cycle in London?

Casualties

Information on cyclist accidents was collected from the GOV.UK Reported road accidents, vehicles and casualties tables for Great Britain (file RAS30043)[3]. The original file was converted from an ODS file to a simplified CSV file using Excel and then read into Python for analysis. Simple line charts were produced to show change over time.

Bicycle theft

Data for bicycle thefts across London boroughs was collected from the London Datastore - Recorded Crime Geographic breakdown file which covered 24 months of recorded crime[4]. The data was provided in csv format and fed into Python for analysis after brief data exploration on Excel. Bar charts were used to compare monthly bicycle thefts across London.

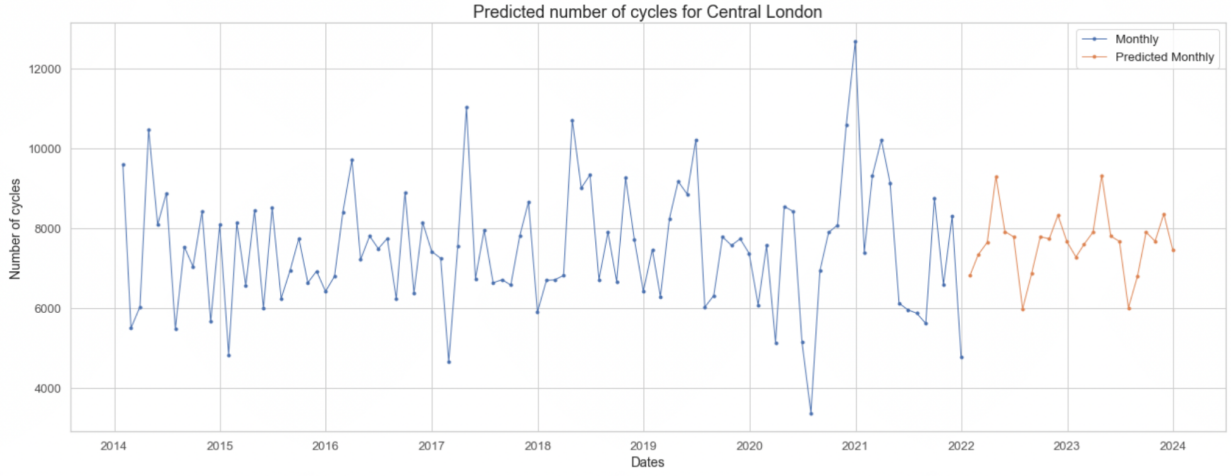
Patterns, trends, insights, and recommendations

Important patterns, trends, insights, and recommendations are grouped by business question below.

What is the trend in total cycles over time?

For central, inner and outer London, a slight increasing trend can be seen until 2020. Due to the pandemic, data from 2020 to 2022 shows a more drastic change, with a much greater increase during spring and summer times and greater decrease during autumn and winter times compared to the previous years. By applying the SARIMA model for forecasting, there isn't much of a change for trends over the next two years. This could be due to a potential bottleneck and certain changes that need to be applied in order to overcome it.

Figure 8. Predicted cycles for Central London



[3] https://www.gov.uk/government/statistical-data-sets/reported-road-accidents-vehicles-and-casualties-tables-for-great-britain#casualties-in-reported-road-accidents-ras30

[4] http://data.london.gov.uk

Figure 9. Predicted cycles for inner London

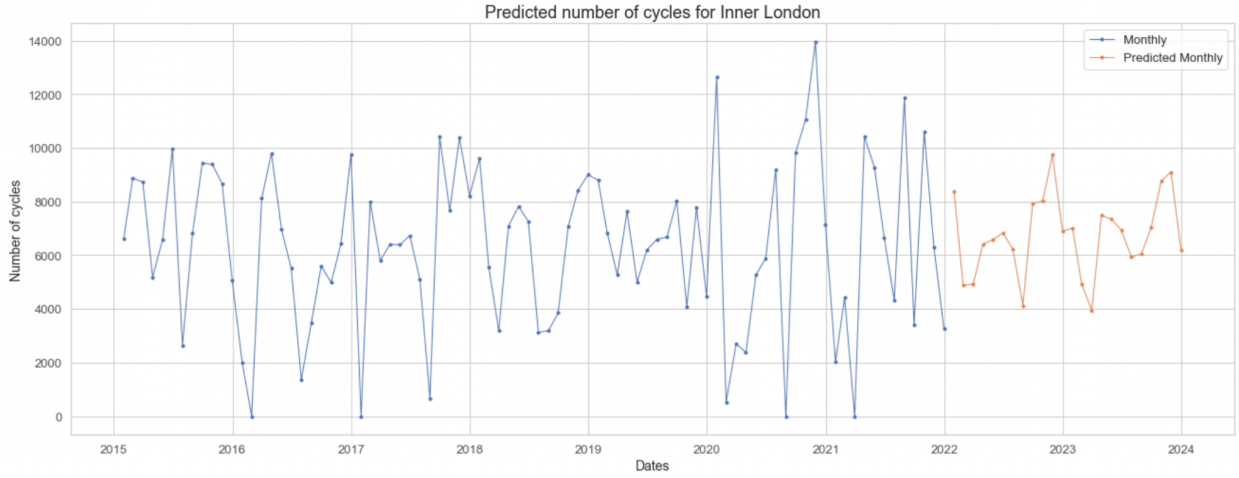
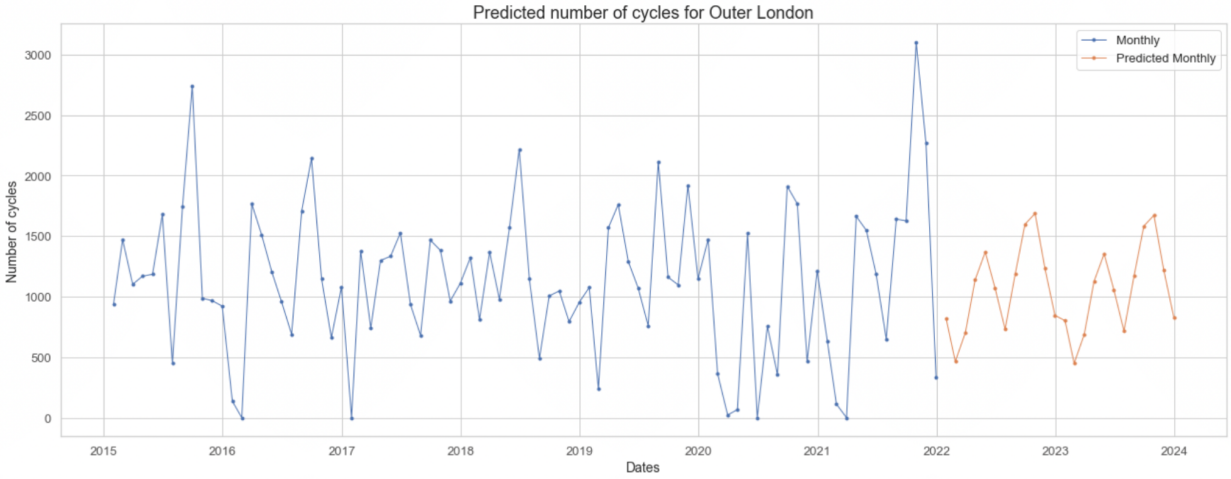


Figure 10. Predicted cycles for outer London



Q1. What are the main factors determining when people cycle?

As expected, morning and afternoon peak hours have the highest number of cycles throughout London, with the only exception being outer London, which has a higher number recorded during the day over morning peak hours. People generally tend to cycle in good weather conditions, and during the spring and summer (Q2 and Q3). Lastly, people living in north and south London cycle more than those living east and west. This could be due to north and south London being more residential, and it's worth investigating this in the future. The time of day when most people cycle and the directionality correspond with the finding that most people tend to cycle for travel-related reasons.

Figure 11. Total cycles by time of day

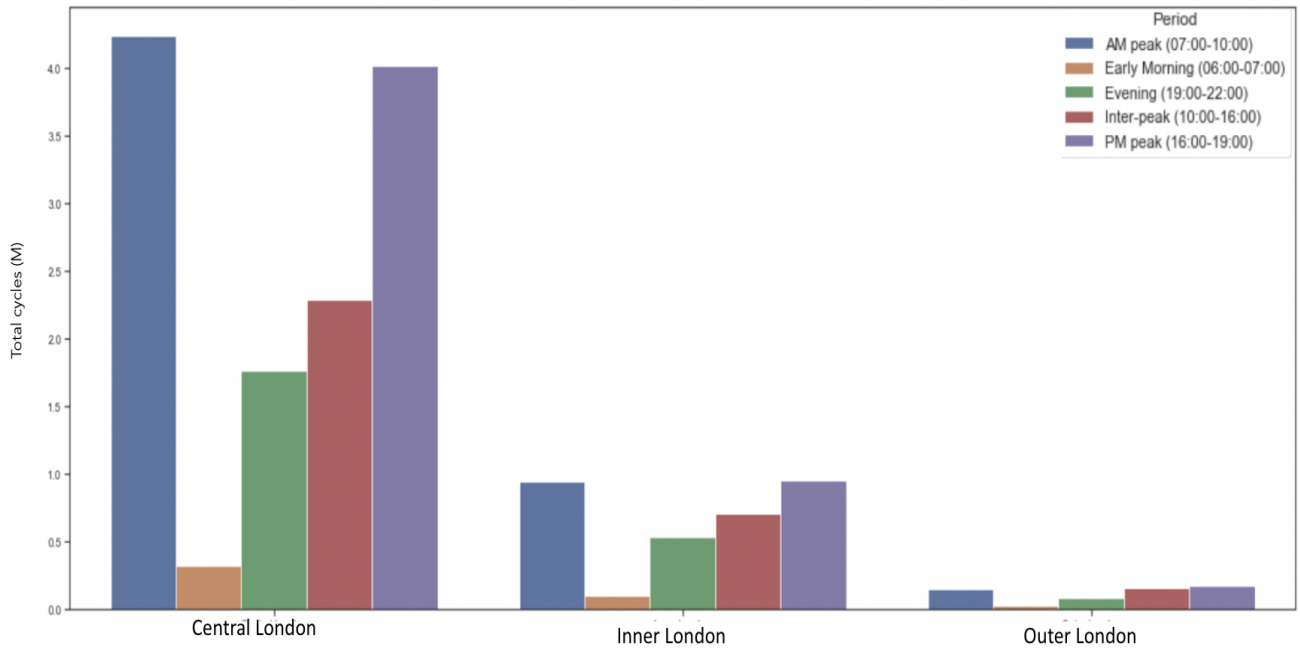


Figure 12. Total cycles by weather conditions

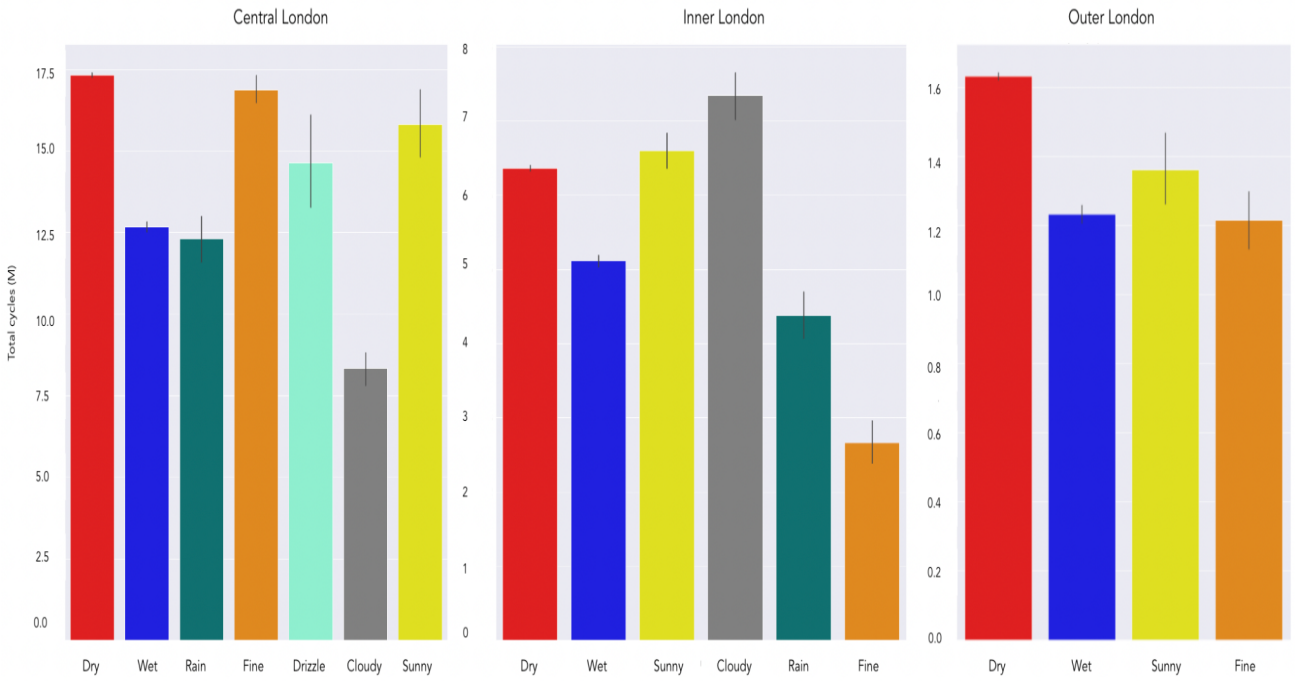


Figure 13. Total cycles by quarter

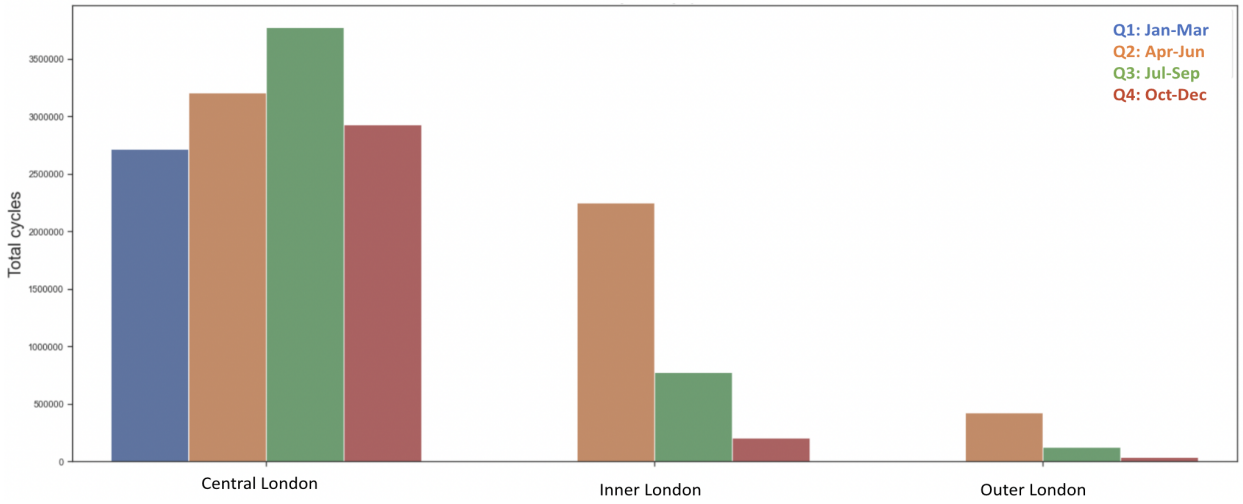


Figure 14. Total cycles by direction of travel

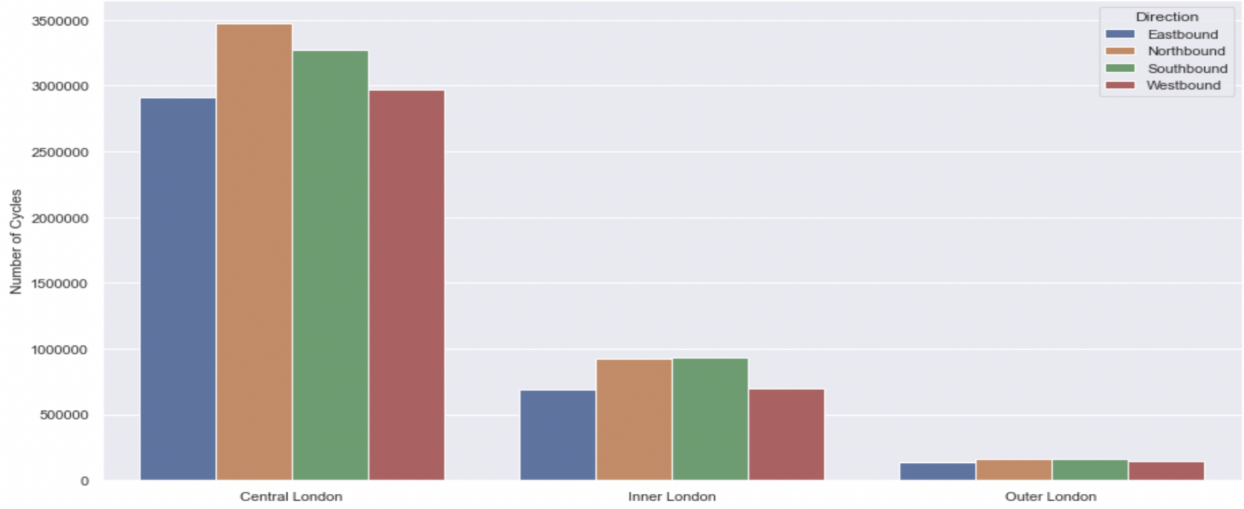
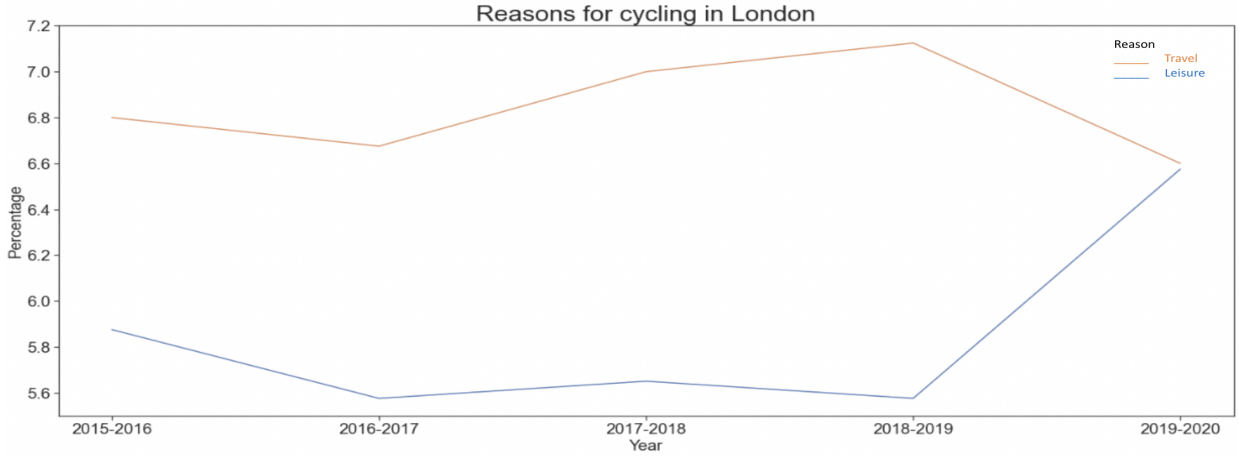


Figure 16. Reasons for cycling



Q2. What impact does infrastructure have on cycling uptake and how can this be improved?

In Inner and Central London, 90% of the cycling population use their own bikes while 10% hire cycles consistently for the last 7 years. This signifies that people who cycle prefer using their own bikes. Also, less people cycle towards east and westbound which may be due to longer journeys. Therefore, TFL is recommended to implement a cycle scheme to promote purchase of private cycles targeted towards east and west London to increase the uptake of cycling. In addition, TFL should improve bike storage facilities on public transport to encourage long journey travellers to cycle once they arrive in central london. Finally, TFL should examine the top 5 locations in each area with the least number of cycles to improve infrastructure, such as cycling roads or cycle hire docking stations.

Figure 17. Private cycles vs cycle hires by year

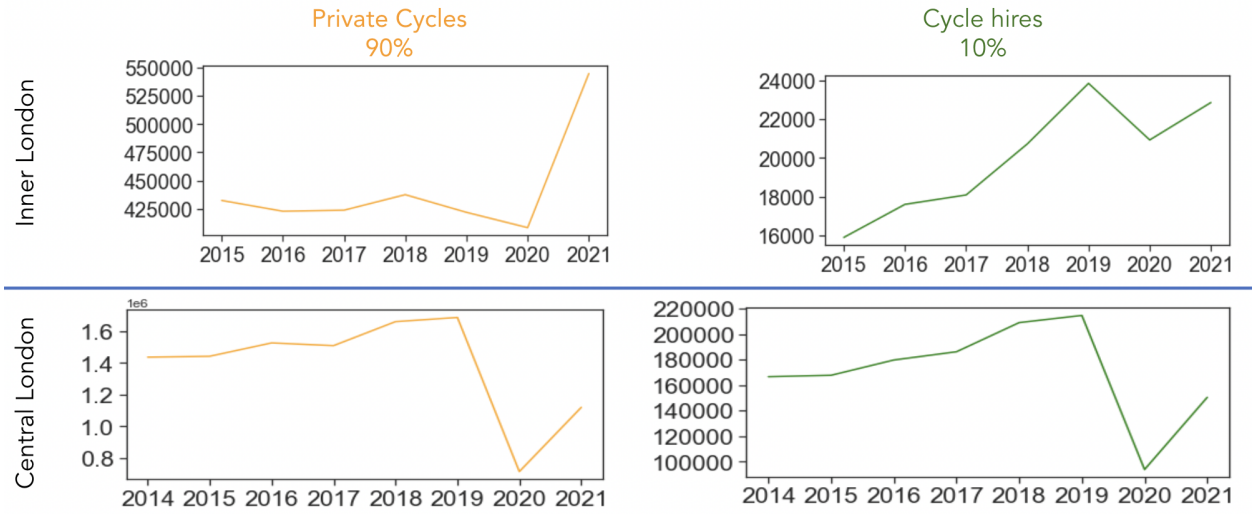
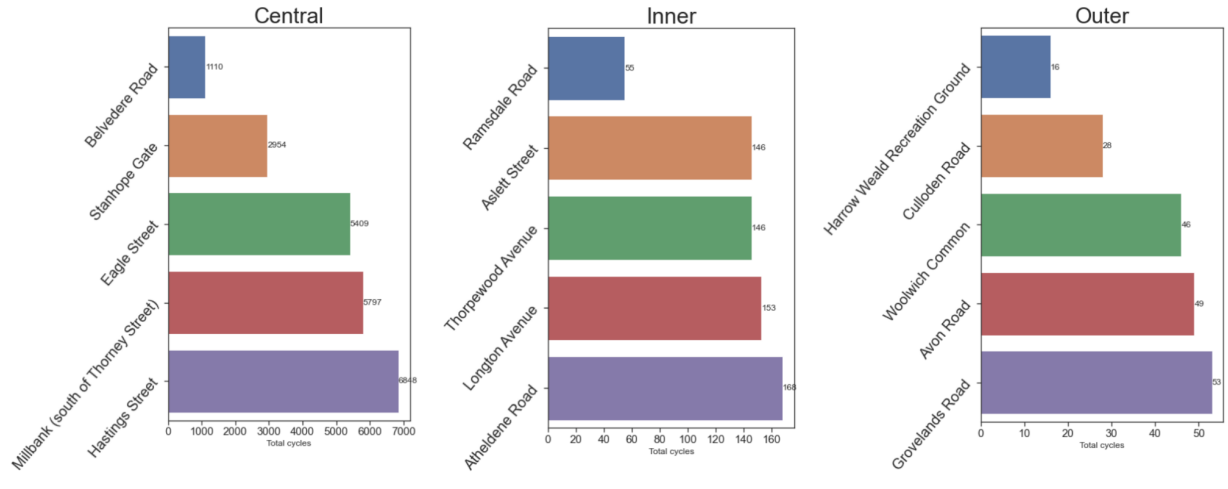


Figure 18. Locations with the least number of cycles



Q3. What are the demographics of cyclists in London?

We observe from the data that unless there is a structural change or policy change, men will often be the higher gender of total cycles. We can see that men represent overall cyclists and this trend seems unlikely to change unless the Government addresses the barriers preventing women from cycling.

Cycling between both genders is carried out predominantly between the morning and hours. This directly affects the risks associated with cycling because it coincides with high vehicle traffic. Despite this the Government must continue its efforts in making cycling safe for both genders despite the vehicle traffic.

Figure 19. Gender split

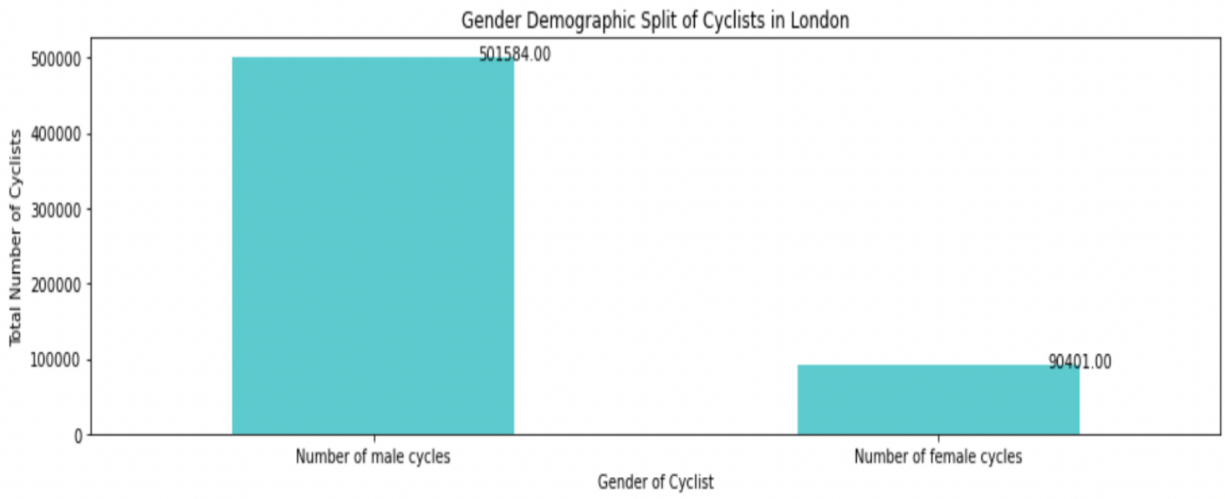
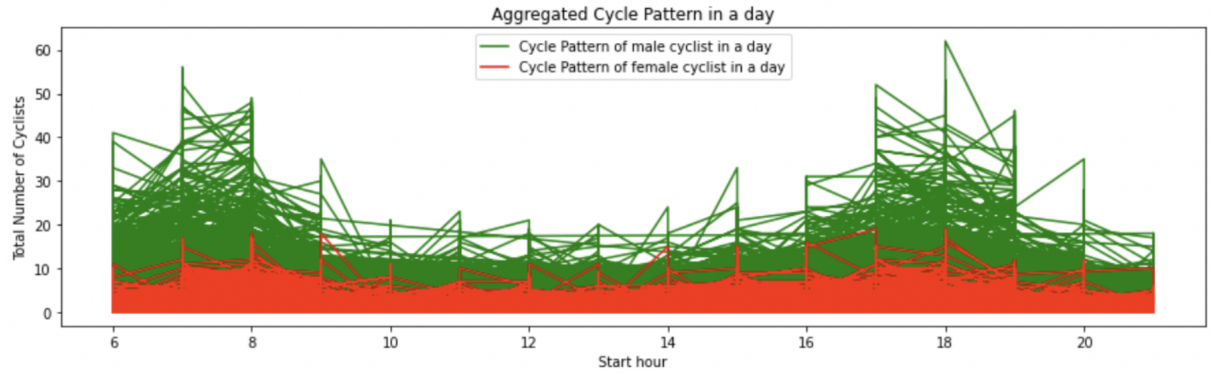


Figure 20. Gender cycle pattern in a day



Q4. What is the sentiment towards cycling in London?

Most tweets associated with “london uk cycle” tended to be neutral to positive. An analysis of the top 20 positive and negative tweets indicated good perception of cycling events, like cycle to work day and initiatives to increase cycling among women. Meanwhile, negative tweets

indicated that more work could be done to increase cyclist safety, to increase the number of cycle lanes available, and to encourage greater use of cycle lanes.

Possible influencers (e.g. thejeremyvine, wearecyclinguk, hounslowcycling) appeared in the word cloud analysis and could be worth investigating as potential organisations to partner with in the future to promote new cycling infrastructure/initiatives.

Figure 21. Sentiment score polarity

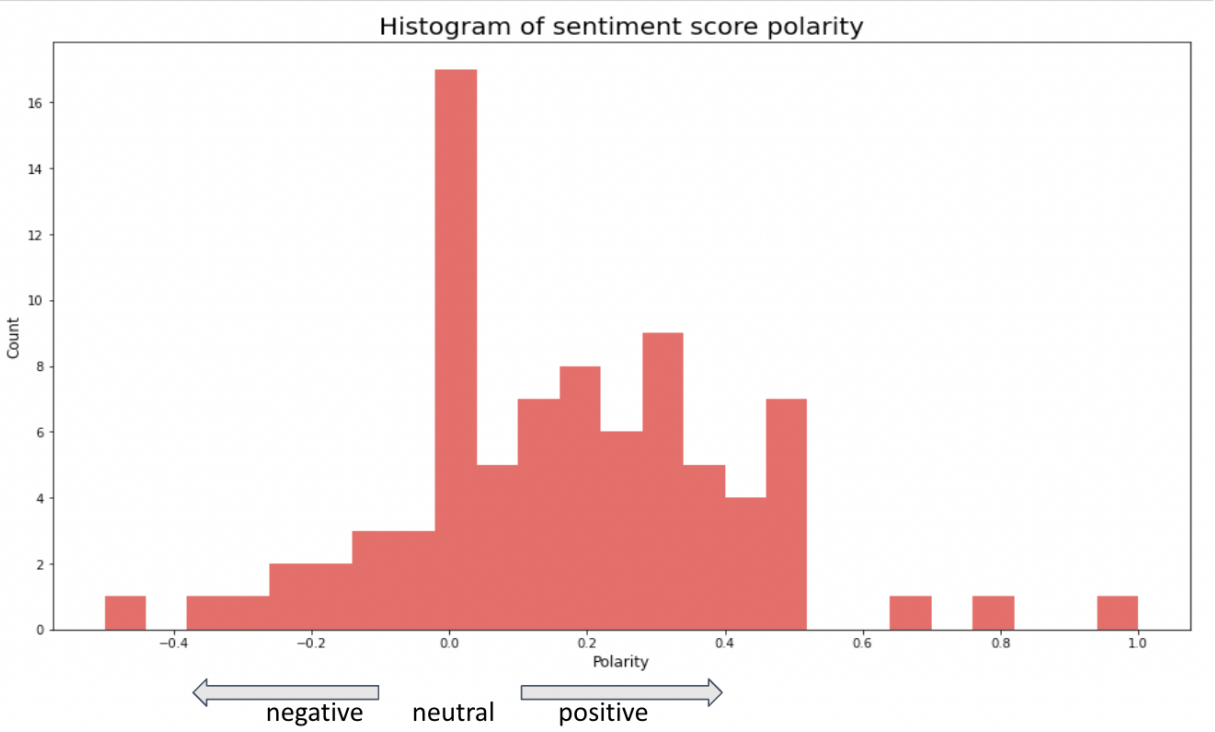


Figure 22. Word cloud

Q5. How safe is it to cycle in London?

Casualties

Cycling casualties decreased from 2014 but began to steadily rise from 2016. The number of serious/fatal casualties also declined from 2014 but began rising from 2017 and is nearing the peak level recorded in 2013. It is highly recommended that TfL investigate reasons for the decline and later increase, especially in regard to serious/fatal casualties.

Figure 23. Cycling casualties in London - all

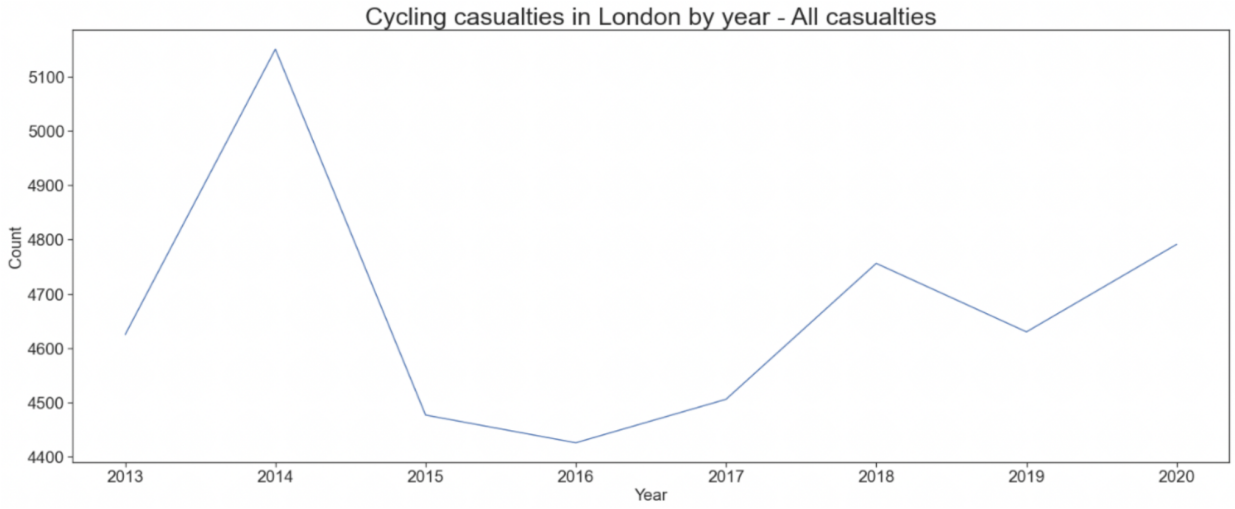
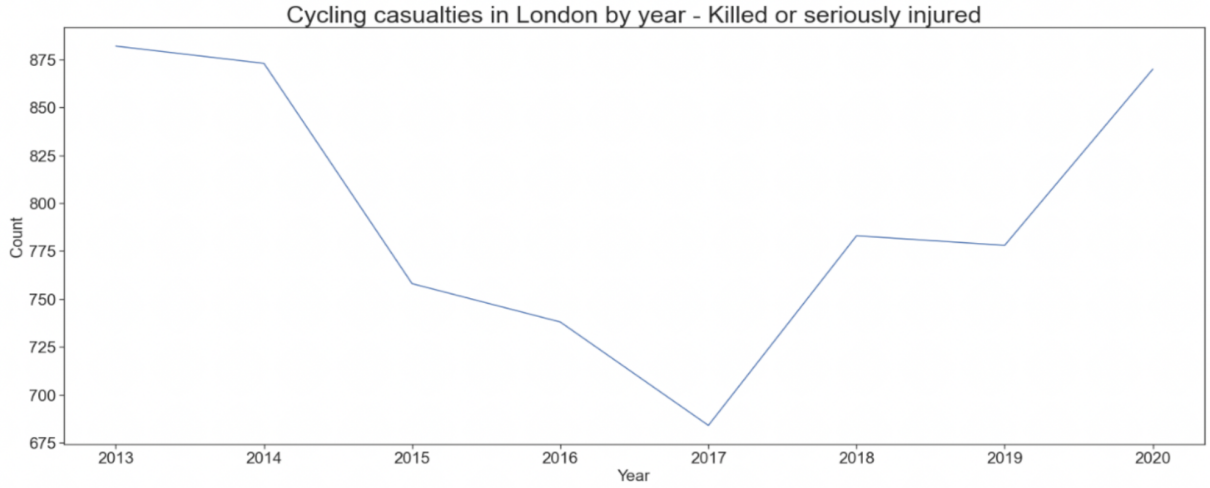


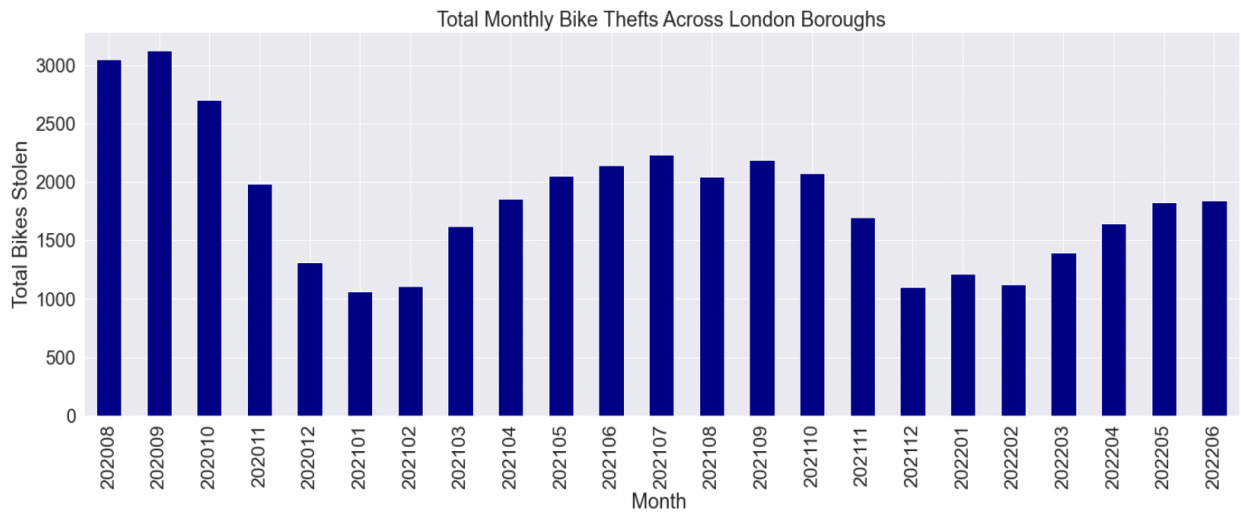
Figure 24. Cycling casualties in London - killed or seriously injured



Bicycle theft

In general, there has been a decrease in bicycle thefts over time. Peaks tend to coincide with seasonality, with greater thefts during the warm-weather months. It is recommended that TfL look into ways of building theft surveillance during the spring and summer.

Figure 25. Monthly bicycle thefts



Conclusion

In response to the main problem statement of how to increase cycling in London, there are five key avenues that TfL should explore as outlined in Figure 26. This is important as the population continues to grow and places additional burdens on infrastructure and public places.

Figure 26. Key recommendations

Appendix A. Coding scripts and data sources

Full scripts can be found on GitHub (https://github.com/hazz292/Cycling_in_London_Team_4) and are organised by folders based on the business questions below.

Business question	Scripts used	Data sources
What is the trend in total cycles over time	Q1_Factors_RZ_Final.ipynb	Central London, Inner London, Outer London
Q1. What are the main factors determining when people cycle?	Q1_Factors_RZ_Final.ipynb Q1_Factors (Reasons why people cycle)_MB_Final.ipynb Q1_Factors (Time of year, Time of Day, Weather, Direction)_MB_Final.ipynb Q1_Factors (Regression analysis)_MB_Final.ipynb	Central London, Inner London, Outer London GOV.UK walking and cycling statistics page (file: CW0302) modified csv uploaded as: 'Q1_Reasons for cycling.csv'
Q2.What impact does infrastructure have on cycling uptake and how can this be improved?	Q2_Infrastructure_HC_Final.ipynb	Central London, Inner London, Outer London
Q3. What are the demographics of cyclists in London?	Q3_Demographics_TM_Final.ipynb	Outer London
Q4. What is the sentiment towards cycling in London?	Q4_Sentiment_MB_Final.ipynb	Twitter (accessed 11-Aug-2022)
Q5. How safe is it to cycle in London?	Q5_Safety (Road accidents and casualties)_MB_Final.ipynb Q5_Safety (Bicycle Theft)_AL_Final.ipynb	GOV.UK website: Reported road accidents, vehicles and casualties tables for Great Britain (file: RAS30043)

Business question	Scripts used	Data sources
		modified csv uploaded as: 'Q5_Cyclist Casualties.csv' data.london.gov.uk:Recorded Crime Geographic Breakdown uploaded as: 'Q5 MPS Borough Level Crime (most recent 24 months)'