Group 1: Happy Uber Cycling – Thoughtworks Project Final Report

1. Introduction

In July 2019, the Department for Transport released data on the type and number of journeys within the UK per year. The annual number of car journeys consistently climbed from 2015 to 2018, resulting in a need for considering alternative transportation methods. As a result, the Mayor of London introduced the Mayor's Transport Strategy in 2018.

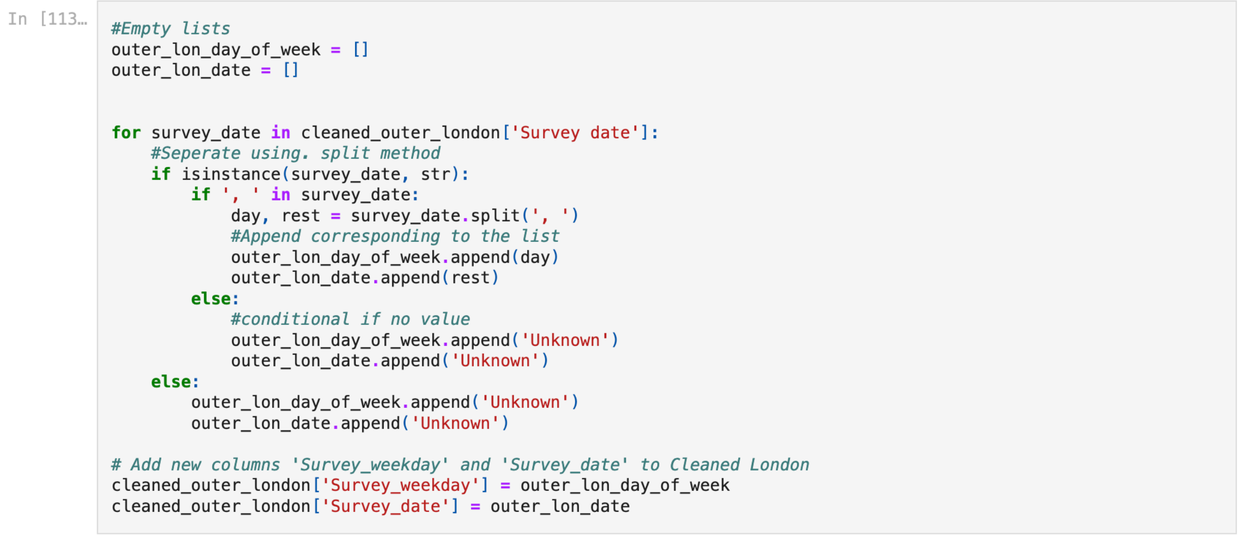
The Mayor and London Assembly are looking to evaluate the successfulness of the 2018 Mayor's Transport Strategy and advise if any changes should be made to the strategy. The main focus of this report is to investigate the factors that affect the number of journeys completed by bike, and provide recommendations to increase the numbers of journeys completed on bike going forward.

2. Development Process & Patterns, Trends & Insights Created

Cleaning the data set:

The cleaning process across datasets followed a uniform strategy, involving the formatting of the datasets, removal or replacement of columns/rows with NaN values and examination of duplicates.

The split method was utilized to separate concatenated categorical day/datetime entries into distinct columns for the 'survey date' column and 'Period' columns. Additionally, the strip method was applied to remove redundant bracketed time from the 'Period' column, because time already existed on a separate column.



```
In [113…
#Empty lists
outer_lon_day_of_week = []
outer_lon_date = []


for survey_date in cleaned_outer_london['Survey date']:
    #Seperate using. split method
    if isinstance(survey_date, str):
        if ', ' in survey_date:
            day, rest = survey_date.split(', ')
            #Append corresponding to the list
            outer_lon_day_of_week.append(day)
            outer_lon_date.append(rest)
        else:
            #conditional if no value
            outer_lon_day_of_week.append('Unknown')
            outer_lon_date.append('Unknown')
    else:
        outer_lon_day_of_week.append('Unknown')
        outer_lon_date.append('Unknown')

# Add new columns 'Survey_weekday' and 'Survey_date' to Cleaned London
cleaned_outer_london['Survey_weekday'] = outer_lon_day_of_week
cleaned_outer_london['Survey_date'] = outer_lon_date
```

In [115…

```
#Drop the bracketed time zones
cleaned_inner_london['Period'] = cleaned_inner_london['Period'].str.split('(').str[0].str.strip()
```

For consistency in the dataset for cross-reference analysis, the 'start hour' and 'start minute' columns were concatenated. Leveraging the zipped method, the hour and minute values were merged into a single column.

In [113…

```
from datetime import time


#Apply datetime module
cleaned_outer_london['Start time'] = cleaned_outer_london.apply(
    #Concatinate the start hour and minutes
    lambda row: time(row['Start hour'], row['Start minute']).strftime('%H:%M'),
    axis=1
)

cleaned_outer_london = cleaned_outer_london.drop(['Start hour','Start minute'], axis=1)
```

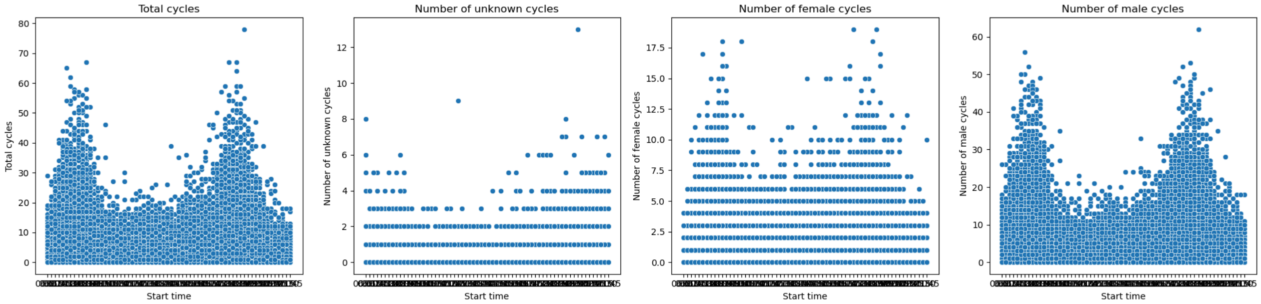
The primary concern in the dataset was the numerous weather types in the column. Employing a for loop and the enumerate method, a list of the different types was created and mapped to their respective weather categories, reducing the number of categories from 138 to six.

In [114…

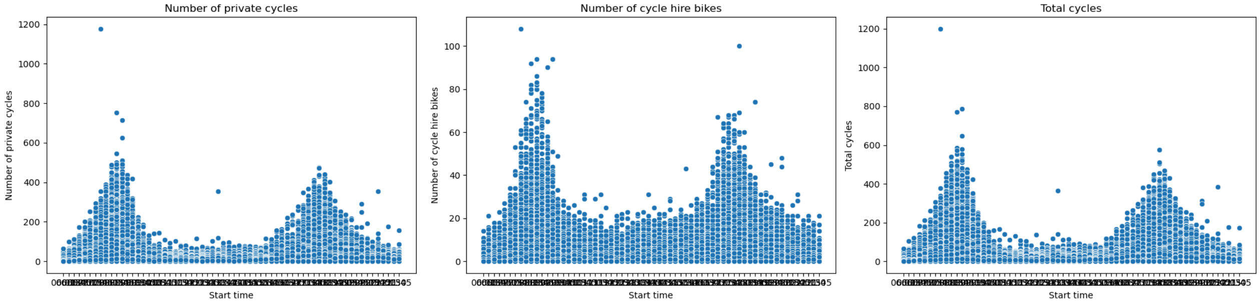
```
outer_london_weather_mapping = {
    'Dry': ['dry', 'dry chill', 'sunny', 'sunny overcast', 'sunny/cloudy', 'dry dark', 'fine', 'good', 'dry very
    'Wet': ['showers', 'cloudy/rain/sunny', 'rain/showers', 'showery', 'intermitent showers', 'short hail shower'
    'Sunny': ['cloudy', 'cloudy sunny', 'cloudy + sunny', 'sunny + cloudy', 'cloudy/sunny', 'cloudy/dry', 'rain &
    'Mixed':[ 'cold/sunny', 'generally overcast brief shower'],
    'Cold':[ 'cold/cloudy',],
    'Windy':[ 'cloudy/windy',],
    'unknown': ['unknown', 'n/a','sun setting', 'dry/cold', 'dry cold', 'windy', 'bright + cloudy', 'dark/dry', '
}

#get the key: value pair from the dictionay
for key, values in outer_london_weather_mapping.items():
    #replace value == Key in the data set.
    cleaned_outer_london['Weather'] = cleaned_outer_london['Weather'].replace(values, key)
```

Lastly, regarding outliers, a scatter plot was employed for numerical data to assess their significance. It was determined that none of the outliers were significant enough to be excluded from the data, as they all fell within reasonable bounds and were retained in the dataset.

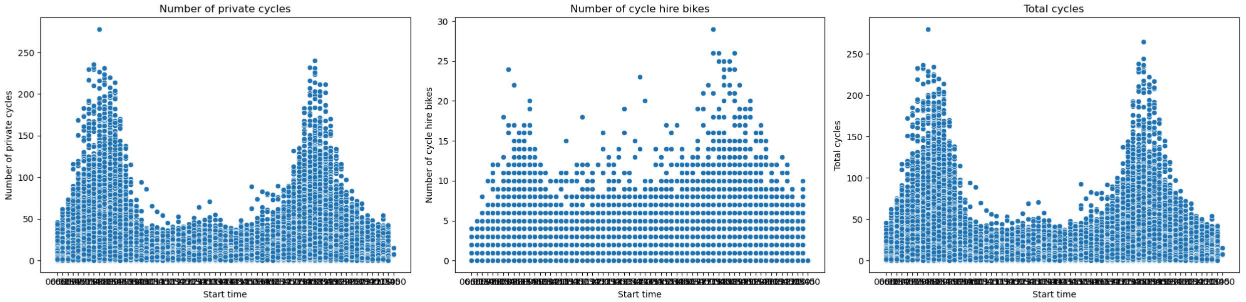
Outer London



Central London



Inner London



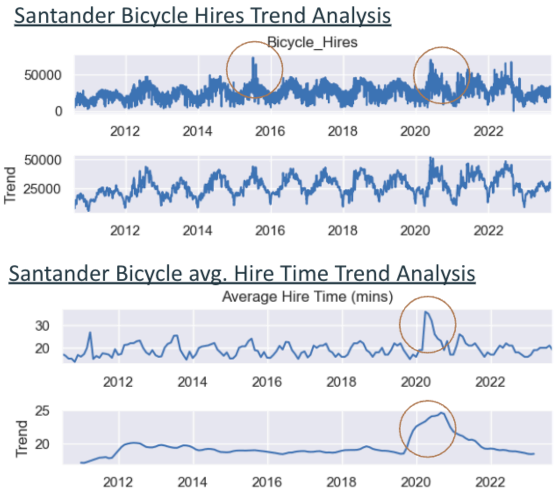
3. Approach, Analysis & Recommendations

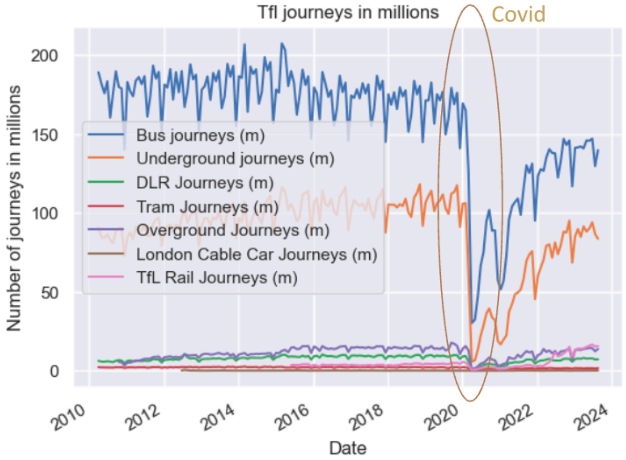
Based on the initial data cleaning & exploration detailed above, the following initial hypotheses were deduced;

- Having good availability and distribution of safe and affordable hire bikes in London will increase the numbers of journeys completed by bike
- More journeys are completed by bike in dry weather than rain
- More journeys are completed by bike in the summer months than in the winter months
- The time of day has an impact on the number of journeys completed by bike, with a large proportion of journeys completed by bike falling within commuter hours
- Residents of 'deprived' areas of London complete fewer journeys on bike than those in 'wealthy' areas

Patterns and trends in the data were then investigated, and various visualization techniques were used.

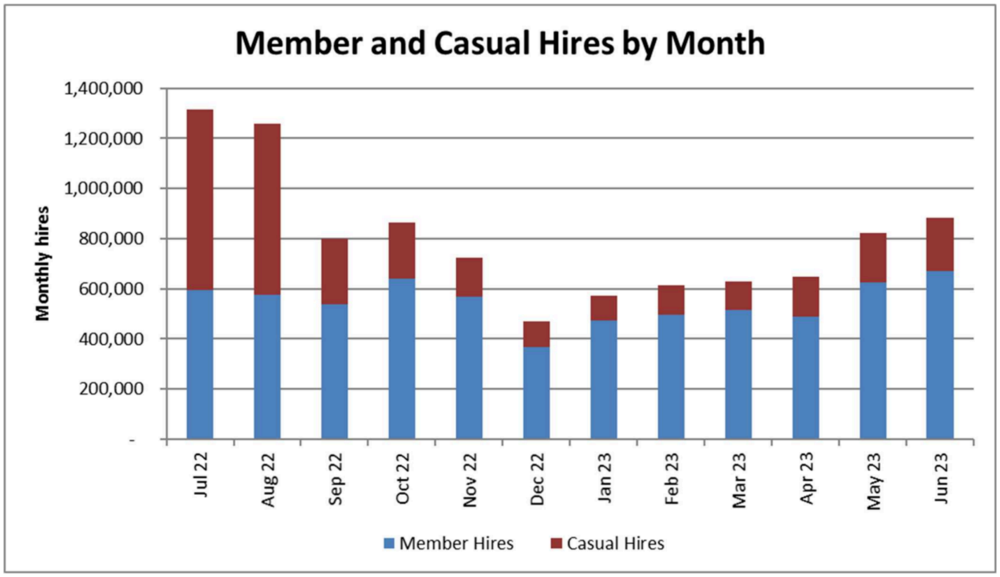
To understand the effects of Covid, a time series analysis on Santander bike hiring data was completed, along with the average hiring time spanning from 2011 to 2023. To achieve this in Python, "from statsmodels.tsa.seasonal import seasonal_decompose" was used to decompose the time series data. The time index was set, allowing Python to generate trend and seasonal data. This showed a peak in both the number of journeys and journey distance by bike in COVID, and as a result it was decided that time-series data would be avoided where possible due to the anomalies caused by COVID.



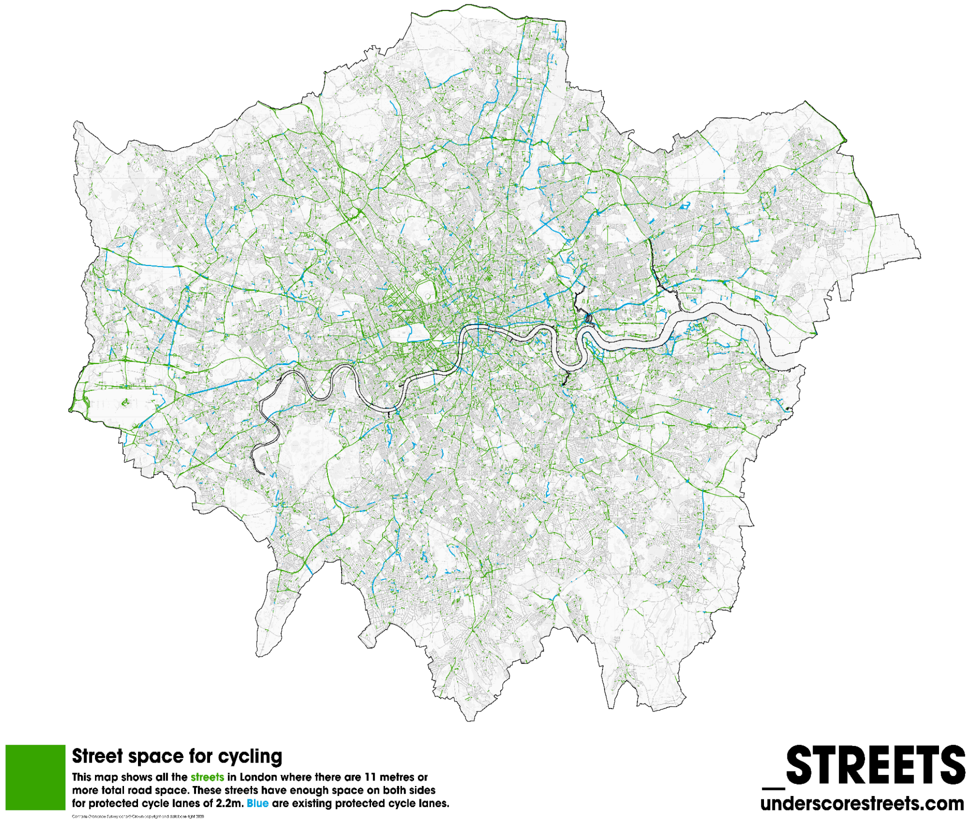


Data taken from Santander bikes showed that approximately 2/3 of casual users of hire bikes live in London. This was against the initial findings which showed peaks of use in holiday months, suggesting that the influence of tourists drives the hires. Bar graphs were used for these visualizations as it clearly showed the changes over time whilst also separating member and casual hires using colors. Following further analysis into the Santander bike hire data, it was recommended to the client that more hire bikes were made available in London. The locations of Hackney & Islington were chosen due to the fact they had a high number of private bike journeys but a low number of hire bike journeys. This demonstrated that the demand was there, but the availability of hire bikes was insufficient. It was recommended that £10 million was spent supplying 588 new bikes, which would generate a return of investment of 12%

based on the income produced at £2/hire.

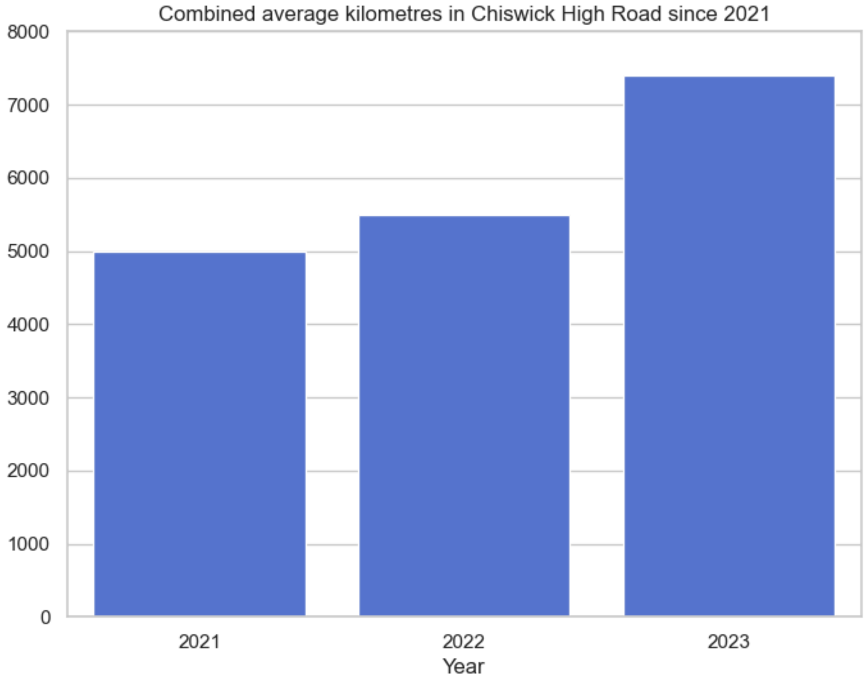


Data comparing the average numbers of bike journeys in central, inner & outer London were then compared. It was clear that more journeys were completed in central London to Inner or Outer London. Further research suggested that this was due to the better infrastructure in central London to the other regions. It was therefore suggested that bike lanes were prioritized for increasing the number of journeys completed by bike.


Street space for cycling
This map shows all the streets in London where there are 11 metres or more total road space. These streets have enough space on both sides for protected cycle lanes of 2.2m. Blue are existing protected cycle lanes.
Contains Ordnance Survey data © Crown copyright and database right 2020
_STREETS
underscorestreets.com

Two methods across three locations were suggested for installing bike lanes. The first method was building new separated bike lanes in Kensington & Camden at a cost of £15 million, producing an increase of 30% of the number of cyclists using the roads. Coloured column and bar graphs were used to present the data justifications, as well as labeled bike path maps to show the areas where this was most required. This visually showed the lack of bike paths in these regions.

The second method included closing a lane of traffic from Brent Cross Station to Regents Park and replacing it with a bike lane. Similar had been completed for the C9 Segregated cycleways (Chiswick): which was established in 2020 as an emergency solution in response to the COVID pandemic. Subsequently, in September 2023, an official public vote solidified their permanence, showing that locals enjoy the segregated cycleways.

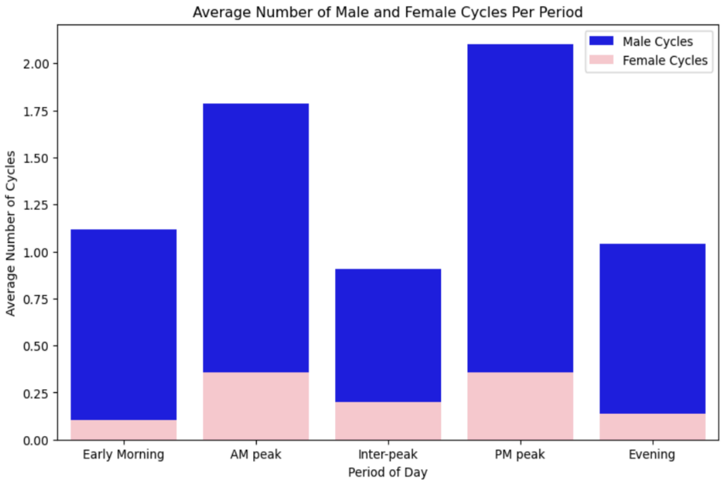


Since 2021, a 47% increase has been found in the distance traveled on Chiswick High Road. Further insights from a survey studying barriers to cycling reveal that the perception of safety, particularly in relation to segregated bike lanes, encourages people to cycle. Additionally, a Boston study demonstrated that two and one-way cycle tracks not only promote cycling, but also significantly enhance safety by showing lower crime and crash rates.

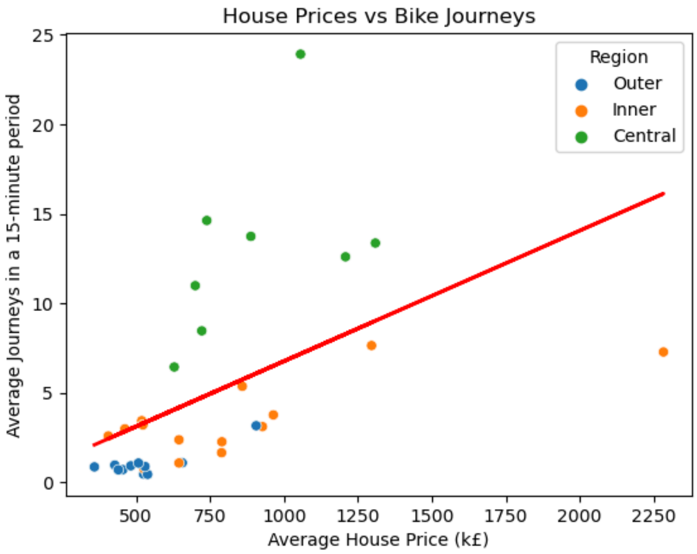
The graphs were designed to look consistent throughout the presentation, and so similar sizes and colors were used. The colors were chosen to be easy to read and stand out on a screen, as the presentations were delivered remotely, whilst also differentiating between the different groups. The size of the charts were designed to fill the screen and be easy to read.

Following the analysis completed and the graphs below, it was also noted that further research in the following areas would be beneficial;

1. Investigating COVID's impact on cycling trends during and post-pandemic, including long-term effects.
2. Exploring gender differences in cycling, focusing on safety perceptions and habits, to inform targeted initiatives.
3. Further analysis into how wealth influences bike usage as a means of transport - does wealth have a direct influence or do other factors create the impression of influence?



There are large discrepancies in the number of journeys completed by bike depending on gender. Further analysis into initiatives on how to encourage more females to use bikes as a form of transport would be beneficial.

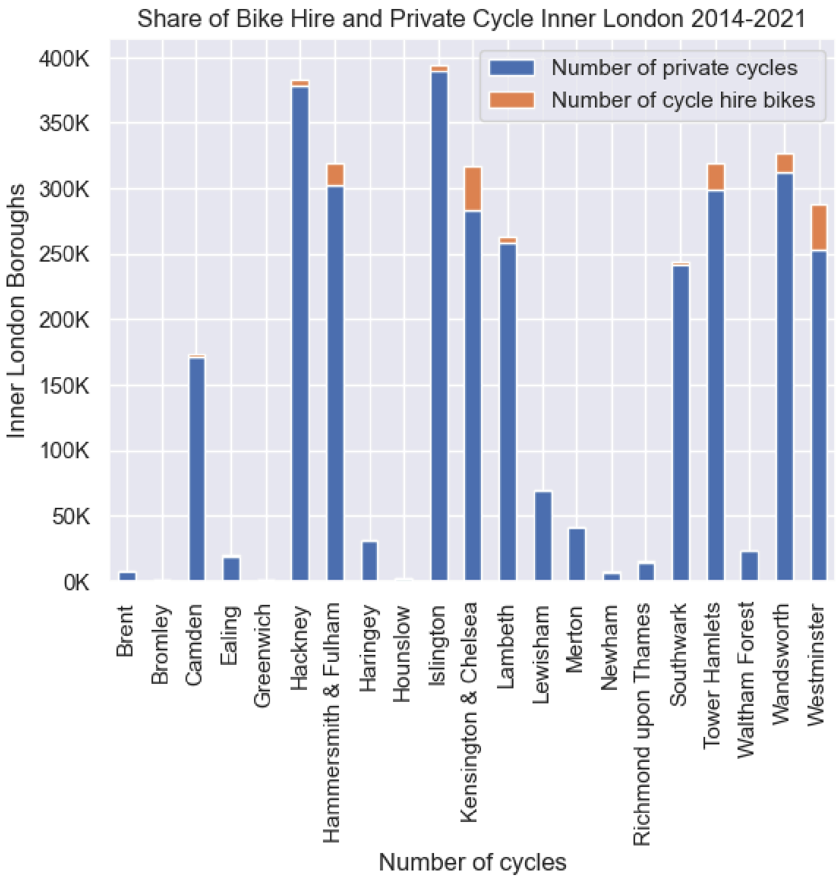


A linear regression model was completed to investigate whether average house price had an effect on the number of bike journeys completed. Further analysis into whether wealth has an influence, or if location & infrastructure are the cause of the increased bike journeys could be justified.

4. Technical Overview of the Code

Python was chosen as the primary analysis tool due to its versatility, user-friendly interface, and extensive library. Pandas, Numpy, sklearn, MatPlotLib, Seaborn, and statsmodels were all used. Individual files, stored on GitHub, were used to avoid misalignments and code misplacements. GitHub served as our platform of choice for version control and collaboration, essential when working as a team on large datasets. The repository was set as 'Public' for ease of use within the group. After the data cleaning process outlined above, clean files were uploaded to Google Drive to allow easy access for all team members.

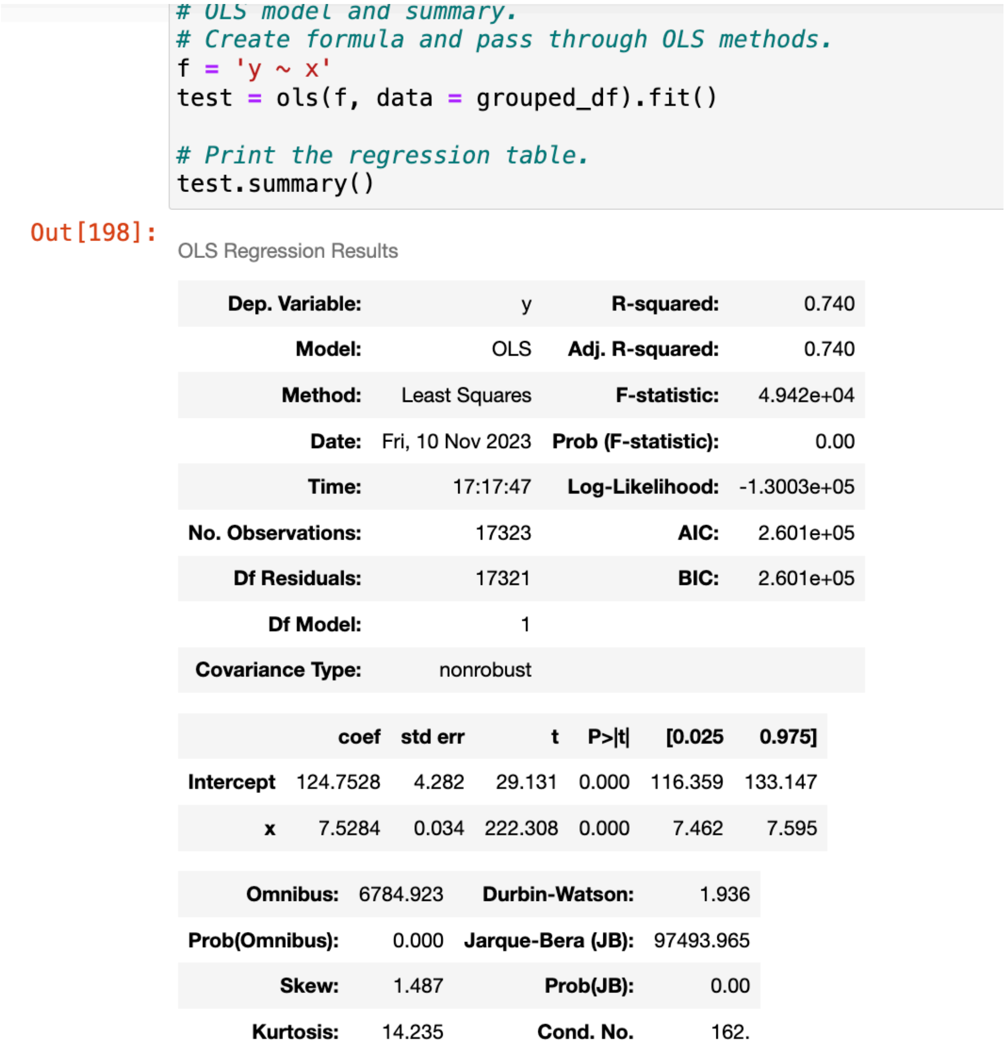
In the Santander bike hire analysis, cleaned data from Tfl and additional data were loaded as dataframes, decomposed using statsmodel into time-series to identify initial trends. Python pandas commands were leveraged for tasks such as renaming columns, merging, joining based on a primary key (location ID), grouping, and creating visual outputs for our presentation.



Through regression analysis, initial correlations between bike hires and total private cycles were explored. This involved assigning an independent variable (x) and a dependent variable (y). Utilizing OLS and linear model imports, an initial OLS regression test was conducted. The coefficients obtained were then used to predict the regression line and visualize a scatter plot for our linear regression. The analysis revealed that bike hires significantly impact overall bike hire demand, with an R-squared value of 0.74. However, due to the impact of Covid, further datapost-Covid is needed to contribute to the development of more robust linear or multi-linear regression models.

Bar, stacked-bar, scatter plots, linear plots and others were used to present the analysis. It was important that the visuals had the appropriate labels and scale. Where units needed displaying in thousands, the following code line was used:

plt.gca().yaxis.set_major_formatter(FuncFormatter(lambda x, _: '{:.0f}K'.format(x / 1000)))



```
# OLS model and summary.
# Create formula and pass through OLS methods.
f = 'y ~ x'
test = ols(f, data = grouped_df).fit()

# Print the regression table.
test.summary()
```

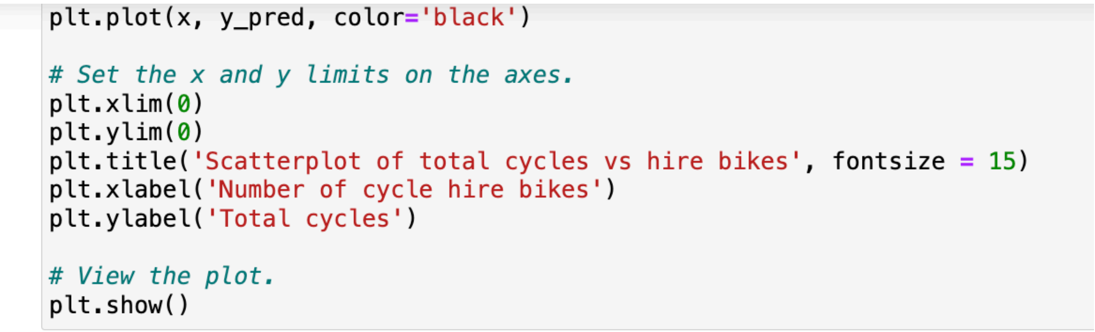
Out[198]:

OLS Regression Results

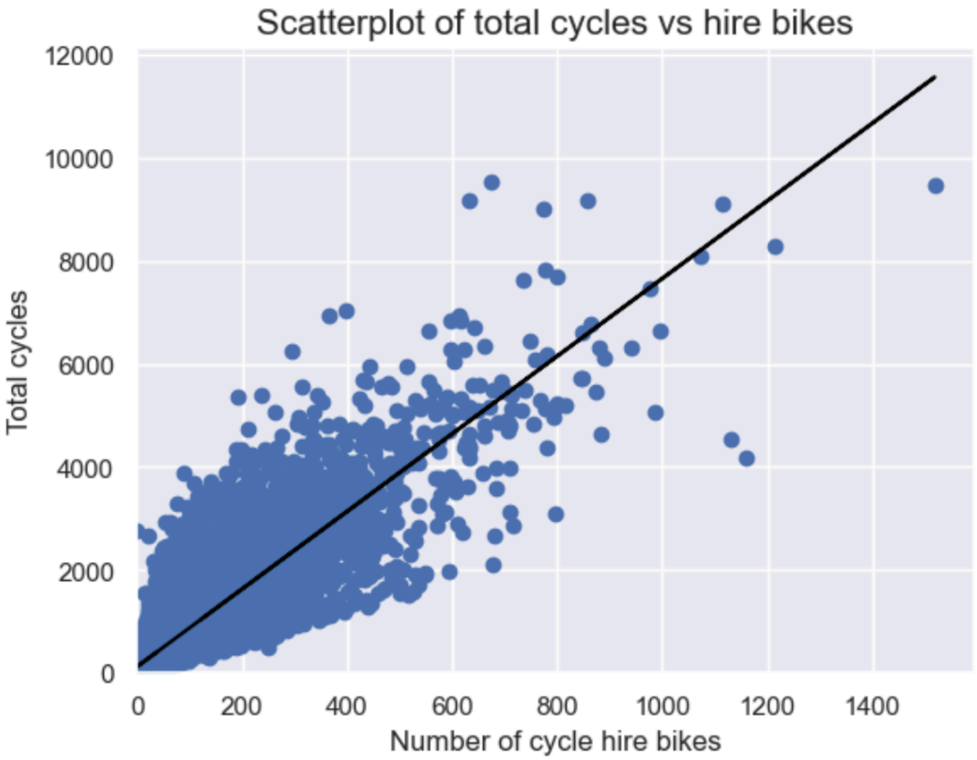
Dep. Variable:	y	**R-squared:**	0.740
Model:	OLS	**Adj. R-squared:**	0.740
Method:	Least Squares	**F-statistic:**	4.942e+04
Date:	Fri, 10 Nov 2023	**Prob (F-statistic):**	0.00
Time:	17:17:47	**Log-Likelihood:**	-1.3003e+05
No. Observations:	17323	**AIC:**	2.601e+05
Df Residuals:	17321	**BIC:**	2.601e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	124.7528	4.282	29.131	0.000	116.359	133.147
x	7.5284	0.034	222.308	0.000	7.462	7.595

Omnibus:	6784.923	**Durbin-Watson:**	1.936
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	97493.965
Skew:	1.487	**Prob(JB):**	0.00
Kurtosis:	14.235	**Cond. No.**	162.



```
plt.plot(x, y_pred, color='black')

# Set the x and y limits on the axes.
plt.xlim(0)
plt.ylim(0)
plt.title('Scatterplot of total cycles vs hire bikes', fontsize = 15)
plt.xlabel('Number of cycle hire bikes')
plt.ylabel('Total cycles')

# View the plot.
plt.show()
```



5. Conclusion

To conclude, a thorough analysis was completed to investigate what factors affected the number of journeys completed by bike. Various factors, including demographic variations in location and gender, natural changes in weather and season and the availability of safe and affordable hire bikes all influence the number of journeys completed by bike in London.

It was therefore recommended to the client that £10 million is invested in more hire bikes across the city, along with a further £19 million to be spent improving the infrastructure in London, specifically by building more separated bike lanes from the traffic, which makes cyclists safer, hence increasing the uptake of bikes as a means of transport.