PREDICTING LONDON BIKE USAGE USING MACHINE LEARNING

Kishore Rajendra

34812636

MATH6185: Operation Research and Data Science Case Study - 1

Dr Xiang Song

Mathematical Science

University of Southampton

July 25, 2024

Contents

List of Figures

List of Tables

1. Introduction

Cycling is being sustainable form of transportation, presents many advantages, reducing Carbon emission, greenhouse gases, decrease in parking demands, and promotion of overall Health and exercise (Cervero et al., 2019). Bicycle sharing is a decisive policy intervention integrated into many urban transportation networks worldwide to promote cycling (Heydari et al., 2021). In 2007, London faced significant challenges to large traffic congestion, results in high levels of pollution and long travel times. The urban transport system introduced a bicycle rental scheme known as "The Santander Cycles scheme" in 2010, offering over 12,000 bikes at 750 docking stations throughout central London, which enables 24-hour bike rentals which was inspired by similar programs in Paris and Brussels.

Ensuring that sufficient number of bicycles are available in high-demand areas at right time can enhance the system's efficiency and user convenience (*Tfl.gov.uk*, n.d.). Therefore, predicting bike usage is crucial for urban mobility which ensures bikes are always available to use. However, so far, few studies have been conducted on the impact of weather on the frequency of public bicycle rental (Feng & Wang, 2017). This case study explores different data science methodologies to address these challenges. This study is intended to provide valuable information for effective resource allocation and management within the bike-sharing system. This case study analyses urban mobility patterns and cycling preferences in London through extensive data collection, cleaning, and preparation, followed by exploratory data analysis, geospatial and multivariate analysis, and predictive modelling using machine learning to enhance the bike-sharing system's efficiency and user satisfaction.

2. Literature Review

The analysis presented in this case study regarding the prediction of bicycle usage in London, insights are synthesized from a variety of academic papers to guide the methodology and provide a contextual framework for the results obtained. The factors influencing the demand for bike-sharing services are diverse, encompassing elements such as environmental circumstances, consumer attitudes, and temporal trends (Eren & Uz, 2020). Our examination of the bike-sharing system in London is consistent with these observations, uncovering notable fluctuations in utilization based on factors like the time of day, day of the week, and weather conditions.

The utilization of machine learning methodologies, particularly random forests and multiple linear regression, has been shown to be effective in projecting the demand for bicycle rentals (Feng & Wang, 2017). In our investigation, similar techniques are adopted, utilizing these models to anticipate patterns of bike usage in London. The superior performance of random forest models, as evidenced by (Gu & Lin, 2023) with an R-squared value of 0.95, corroborates our own results, where ensemble methods like random forest and gradient boosting exhibited high precision in predicting bike demand. The versatility of Python in the realm of machine learning, as underscored by (Raschka & Mirjalili, 2019), has played a crucial role in our analysis. We made use of tools such as scikit-learn and pandas for tasks like data preprocessing, model training, and assessment, demonstrating the efficacy of Python in managing intricate machine learning assignments. Our spatiotemporal analysis methodology draws inspiration from studies in analogous domains, such as the prediction of taxi demand (Yao et al., 2018). By integrating external variables like weather conditions and time-related attributes, our objective was to refine the accuracy of our forecasts regarding bike usage, akin to the strategies employed in forecasting taxi demand. The exhaustive comparative evaluation of short-term passenger flow prediction theories conducted (Zhai et al., 2018) informed our approach to feature engineering. We took into account cyclic variations and seasonal dynamics in patterns of bike usage, aligning with the principles of sophisticated models such as Prophet and Gradient Boosting Decision Trees.

To sum up, our investigation into predicting bicycle usage in London builds upon a robust foundation of research on bike-sharing systems, applications of machine learning, and patterns of urban mobility. By amalgamating these insights and adapting them to the specific context of London's bike-sharing program, we contribute to the continual endeavours aimed at optimizing urban transportation systems and improving sustainable mobility solutions.

3. Analysing Urban Mobility patters and cycling preferences in London's Diverse population

This study shows the patterns of urban mobility and the choices for bike sharing in London through an analysis of the data on bicycle-sharing obtained from Transport for London (TfL) in June 2023.It delves into customer behaviour, temporal trends, and spatial characteristics related to bike-sharing. The main aim is to grasp the influence of various factors on bike usage to improve the system and elevate customer satisfaction.

3.1 Data Collection and Preparation:

Collecting data and preparing for analysis involves strategic venture that demands planning, execution and supervision (*Data Migration and Cleaning*, 2024). The dataset for the Bike rentals was sourced from Transport for London (TfL) https://cycling.data.tfl.gov.uk/ database for the month June 2023.

- `373JourneyDataExtract05Jun2023-11Jun2023.csv`
- `374JourneyDataExtract12Jun2023-18Jun2023.csv`
- `375JourneyDataExtract19Jun2023-30Jun2023.csv`

The columns in the dataset include:

- **Number:** A unique identifier (Trip ID).
- **Start Date:** The date and time when trip started.
- **Start Station Number:** The identifier for the starting station.
- **Start Station:** Name of starting station.
- **End Date:** The date and time when trip ended.
- **End Station Number:** The identifier for the ending station.
- **End Station:** Name of ending station.
- **Bike Number:** A unique identifier for the bicycle used.
- **Bike Model:** The model of the bicycle used.
- **Total Duration:** Total time duration of the trip (in a human-readable format).
- **Total Duration (ms):** Total time duration of the trip in milliseconds.

3.2 Data Loading and Concatenation:

The data frame is loaded into a python using Pandas. Each “csv” file is read into separate data frame and then these data frames are concatenated into single data frame “df” for further analysis.

3.2.1 Data Cleaning:

To ensure intergrity if analysis, duplicate data is removed from the data frame using ‘drop_duplicates()’ function. Additionally, the relevant data columns are converted from string to datetime format using ‘pd.to_datetime()’ function to facilitate time based analysis. The total duration of each bike is converted into timedelta format using ‘pf.to_timedalta()’ function for easy manipulation. This conversion is for straightforward calculation of total and average rental duration.

3.2.2 Filtering and Resampling data:

To ensure that only June 2023 data is used for the analysis, filtering the data frame based on month and year of the ‘Start Date’ column is done. To analyse rental patters over time, the data is resampled to 3-days interval using ‘resample()’ function. This allows the calculation of total and average rental duration, as well as count bike rental withing each interval.

3.3 Customer Behaviour Analysis:

3.3.1 Rental Duration Analysis:

To gain insights of the customer behaviour over trip duration, the data is divided into 3-days interval period. This analysis focus on examining the total trip duration and average duration over June 2023. This analysis reveals interesting patters in bike usage across the month.

3.3.1.1 Total Rental Duration Analysis:

The analysis of total rental duration uncovers significate variation of bike usage for June 2023.

- **Peak Usage -** The highest total trip duration was observed from June 22nd to 23rd.
- **Low Usage -** Significate decrease in trip duration is observed from 14th to 17th of June.

This indicates that variability in bike usage over a month, suggest potential external factors influence bike usage during this period. Fig. 1 below illustrates the Total rental analysis over the month of June.

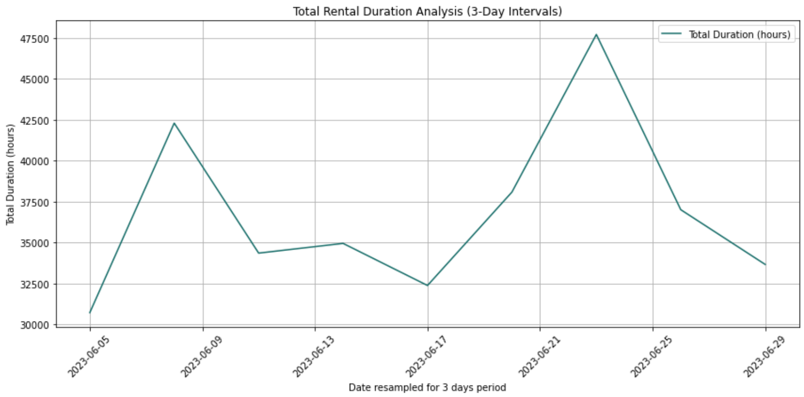

Fig. 1 Total rental analysis over the month of June 2023.

3.3.1.2 Average Rental Duration Analysis:

The analysis for Average rental duration provides further insights for the customer behaviour analysis.

- **Longest Average Trips**: The highest average rental duration is noted from June 27th to 29th closely followed by the period 20th to 23rd of June.
- **Shortest Average Trips**: Notable decline in average rental duration is captured from June 7th to 14th.

These variation in average trip length could be an attribute to various factors, such as change in weather special events in city or shifts in demographics during the month. Further analysis is needed to identify the cause of these fluctuations. Fig. 2 below illustrates the Average rental analysis over the month of June.

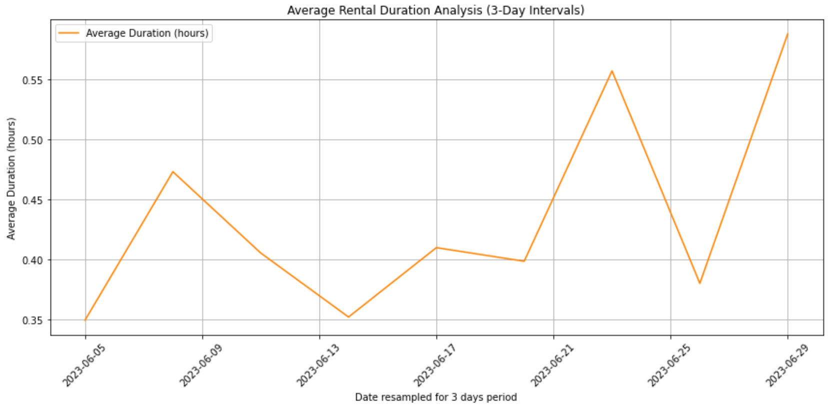

Fig. 2 Average rental analysis over the month of June 2023.

3.3.2 Frequency Analysis:

To further understand the customer behaviour, frequency analysis was conducted on the bike rental data for the month June 2023. This analysis focused on daily rental frequency throughout the month, providing insights and trends of the bike trip. The frequency analysis begins with the count of the number of rentals by grouping them by bike numbers and station to see some patterns. And it is visualized to see the rental frequency using line plot.

- **Irregular pattern:** Overall trend is irregular, indicates that various factors may influence in the bike usage.
- **Drastic drop:** There is noticeable drop in the rental frequency from 15th June, reaching it to lowest point on 18th of June.
- **Sudden Increase:** Following the dip on 18th, the rental frequency sharply increases.
- **Peak Usage**: Maximum rental frequency is observed on 21st of June.
- **End of Month decline:** Towards the end of the month the rental frequency shows decreasing trend.

These Fluctuations suggest that some external factors may influence in the rental frequency pattern, such as special events, weather conditions or other variables. This indicates that deeper investigations required to know their fluctuations in their pattern. Fig. 3 below illustrates the Daily rental frequency over the month of June.

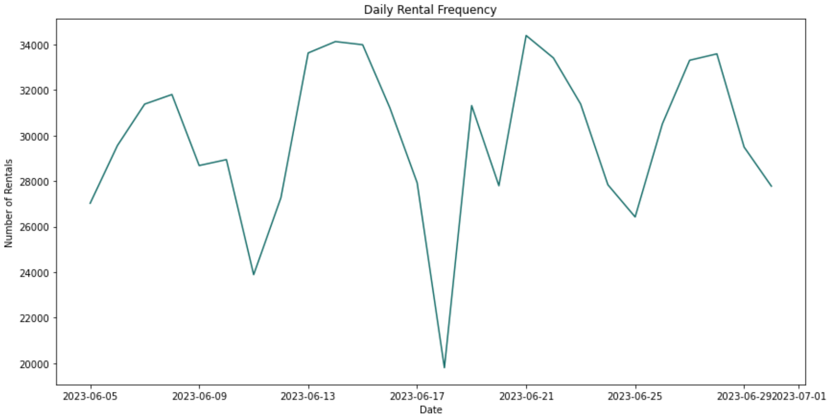

Fig. 3 Daily Rental Frequency

3.3.2.1 Time Series Analysis:

To gain deeper insights into the temporal patters of bike rental, a time series analysis was conducted on the daily rental frequency data. This analysis aimed to identify trends, seasonality and other patterns in bike rental data for June 2023. This analysis also includes examination of autocorrelation functions.

The stationarity of time series was tested using Augmented Dickey- Fuller (ADF) test. Which helps determining if time series is stationary, which is crucial step in time series analysis.

- **ADF Statistic**: -2.7962694141188105
- **p-value:** 0.05880863483788623
- **Critical Values:**
  - 1%: -3.9240193847656246
  - 5%: -3.0684982031250003
  - 10%: -2.67389265625

The results suggest that time series is close to being non stationary.

3.3.2.2 Seasonal Decomposition:

The seasonal decomposition of the time series was performed to separate trends, seasonal and residual components.

- **Trends Analysis**: The trend component of the time series appeared to be relatively horizontal, suggesting there is no significant upward or downward trend over a period.
- **Seasonality:** There is clear seasonal pattern, which peaks at Wednesday and reached lowest point on Sunday. This pattern shows the strong influence of the day of the week on rental behaviour.
- **Residuals:** The residual pattern captures the random noise in the data.

Fig. 4 below shows the Time series seasonal decomposition plot over the month of June.

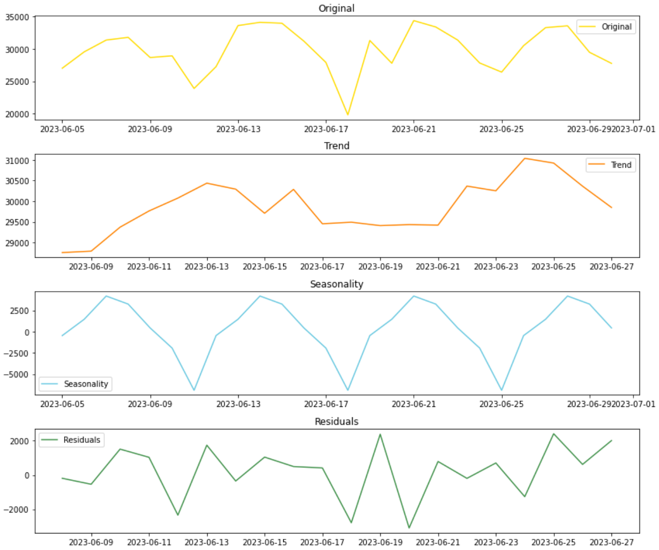

Fig. 4 Time series Seasonality Decomposition

3.3.2.3 Autocorrelation Function and Partial Autocorrelation Function Analysis:

The Autocorrelation Function (ACF) and Partial Auto correlation (PACF) plots are used to analyse correlation within the time series data.

- **ACF Plot:** There is positive spike at lag 7, which indicates weekly pattern, and negative spike at lag 3 and 10.
- **PACF Plot:** There is significant spike at lag 1, with negative spike were noted at lag 3 and 8.

These results indicates that the bike rental data is both trend and seasonal patterns, with notable weekly cycle. This analysis provides foundation for developing predictive models and improve bike sharing services. Fig. 5 below shows the Time series Autocorrelation Function and Partial Autocorrelation Function plot over the month of June.

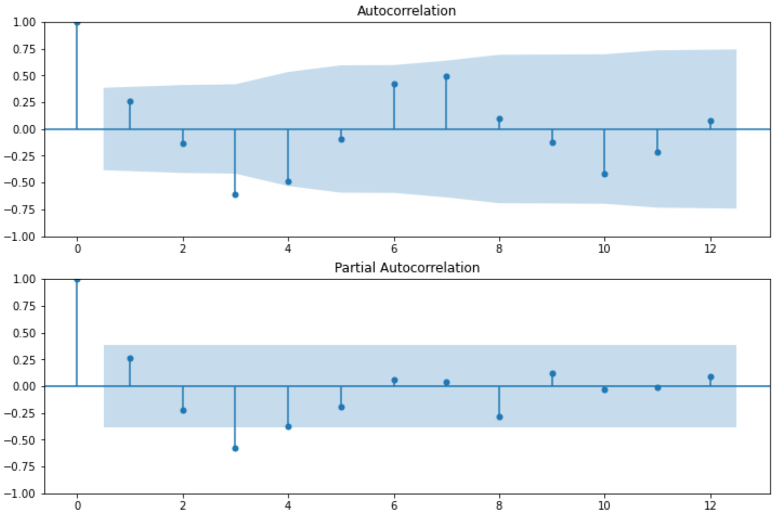

Fig. 5 Autocorrelation Function and Partial Autocorrelation Function

3.4 Geospatial Analysis:

The Geospatial analysis for bike rental provides valuable insights into the spatial patterns, which help to identify the most popular routes, start station, and end station. This analysis helps to understand the spatial distribution of bike usage and can inform decision about where allocation of resource and expanding of infrastructure is needed.

3.4.1 Popular Start Stations:

The analysis of most popular start stations revealed the Top 5 locations where bike rentals are initiated most frequency. Fig. 6 Shows the Top 5 Popular Start stations

1. **Hyde Park Corner, Hyde Park:** 5,508 trips
2. **Waterloo Station 3, Waterloo:** 4,299 trips
3. **Black Lion Gate, Kensington Gardens:** 4,051 trips
4. **Albert Gate, Hyde Park:** 3,365 trips
5. **Waterloo Station 1, Waterloo:** 3,339 trips

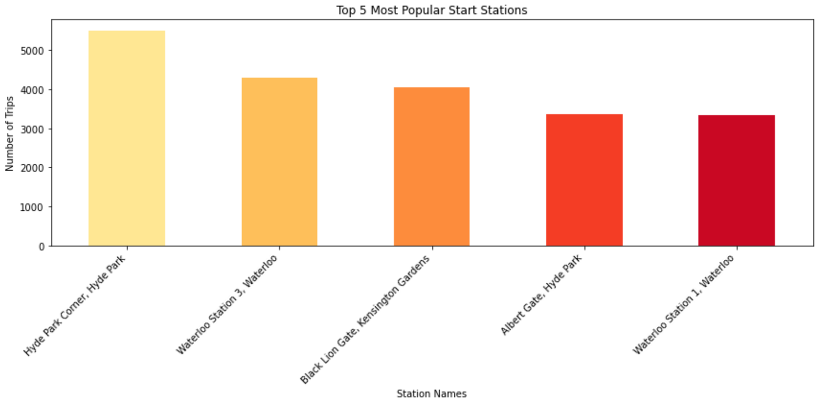

Fig. 6 Top 5 Popular Start Stations

3.4.2 Popular End Stations:

The analysis of most popular End stations revealed the Top 5 locations where bike rentals were concluded. Fig. 7 Shows the Top 5 Popular End stations.

1. **Hyde Park Corner, Hyde Park:** 5,532 trips
2. **Waterloo Station 3, Waterloo:** 4,677 trips
3. **Black Lion Gate, Kensington Gardens:** 4,026 trips
4. **Hop Exchange, The Borough:** 3,750 trips
5. **St. James's Square, St. James's:** 3,690 trips

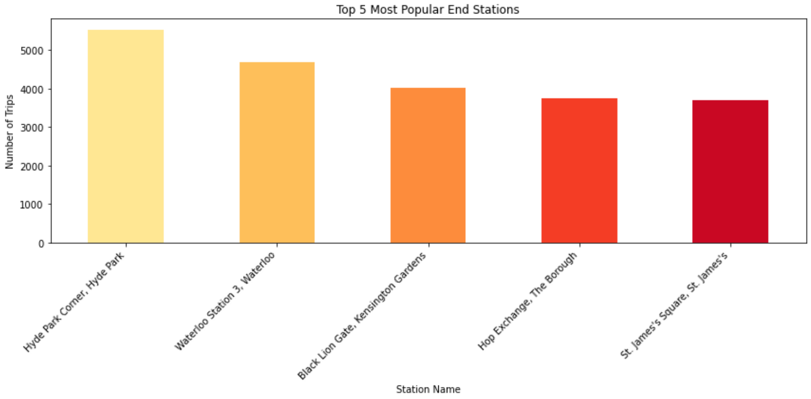

Fig. 7 Top 5 Popular End Stations

3.4.3 Popular Routes:

The analysis also identifies the Top 5 Most popular routes, defined as trip between specific start station and end station. Fig. 8 Shows the Top 5 Popular Routes.

1. **Hyde Park Corner, Hyde Park to Hyde Park Corner, Hyde Park:** 1,615 trips
2. **Podium, Queen Elizabeth Olympic Park to Podium, Queen Elizabeth Olympic Park:** 1,302 trips
3. **Black Lion Gate, Kensington Gardens to Black Lion Gate, Kensington Gardens:** 931 trips
4. **Albert Gate, Hyde Park to Albert Gate, Hyde Park:** 704 trips
5. **Triangle Car Park, Hyde Park to Triangle Car Park, Hyde Park:** 550 trips

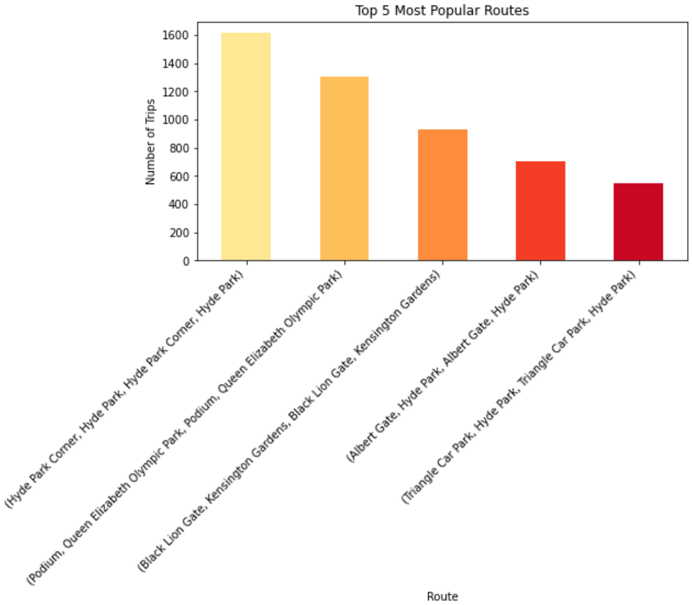

Fig. 8 Top 5 Popular Routes

The data was visualised to provide clear understanding of number of trips for the Top start and End stations, as well as the popular routes. Bar plot was created to illustrate these patterns which help to identify the key areas of bike rental activities and their popular routes in the city.

3.5 Multivariate Analysis:

3.5.1 Trip Duration vs Time of the Day:

To gain much more deeper insights into how trip duration varies by time of the day and day of the week, a detailed multivariant analysis is conducted. The analysis helps in understanding the behaviour patterns and identify peak usage times, which can be crucial for resource allocation and planning.
Fig. 9 Shows the visualisation of the distribution of trip duration by hour of the day and day of the week. The analysis involves in plotting trips, with different colours representing different days of the week. The visualisation provides several key insights.

- **Peak Hours:** Trip durations are generally higher during the early morning from (8 AM to 11 AM) and late afternoon to early evening (2 PM to 7PM) hours. This pattern suggest that many users are utilizing the bike rental service for commuting.
- **Day of the week:** There is noticeable variation in trip duration across different days of the week. Weekdays, especially Monday to Friday, shows more consistent pattern of higher trip duration during peak commuting hours.
- **Hourly Distribution:** The plot shows tend to shorten during the midnight, which could be shorter, non-commuting trips.

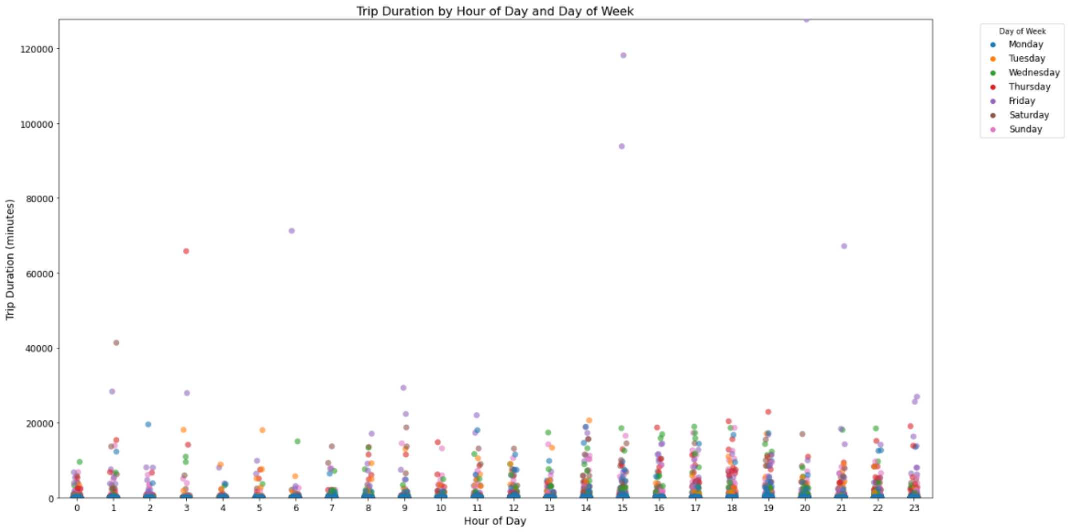

Fig. 9 Trip Duration by Hour of day and Day of Week

This analysis provides valuable insights into temporal patters of bike rentals. By understanding when users are likely to rent bike and for how long, the bike sharing system can be better managed to ensure bike availability and meet user needs. During identified peak hours, more bike can be made available at high demand station.

3.5.2 Heat Map Analysis of Trip Duration:

To further explore the relationship between the trip duration and the time of day, heatmaps were created to visualise trip duration by hour of the day and day of week.

3.5.2.1 Heat Map for Trip Duration by Hour of Day:

The heat map reveals distinct patterns in the bike usage. The analysis indicates the peak hours for bike rentals occurs primarily in the early morning at 8 AM. Followed by the significant activity in the late afternoon around 5 PM to 6PM. Conversely, shows lower levels during the early morning hours from midnight to 5AM. Fig. 10 Shows the heatmap of trip duration by Hour of Day

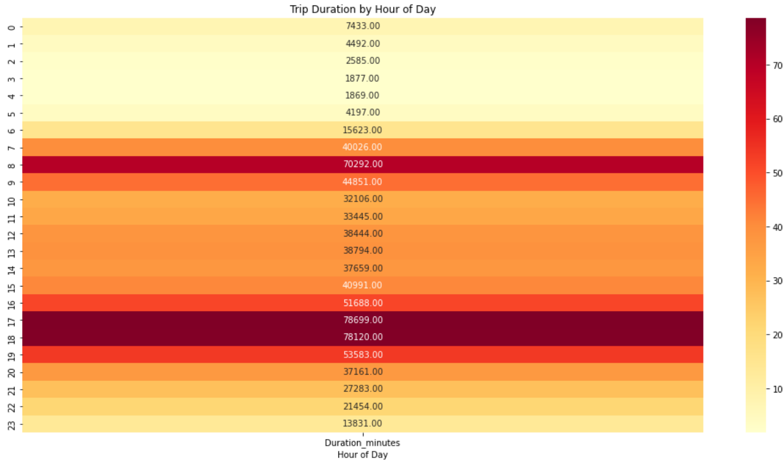

Fig. 10 Heat Map for Trip Duration by Hour of Day

3.5.2.2 Heat Map for Trip Duration by Day of the Week:

The heatmap shows the trip duration by day of week for further enhance understanding. The heatmap revels weekdays generally have higher trip duration than at weekends, indicates that many users are likely use the bike sharing service to commuting during workweeks. It can be noted that Wednesday is stands 1st where most of the users use bike sharing service compared to all other weekdays. This variation in trip duration across different days highlights the importance of considering daily patterns when planning bike distribution. Fig. 11 Shows the heatmap of trip duration by Day of the Week.

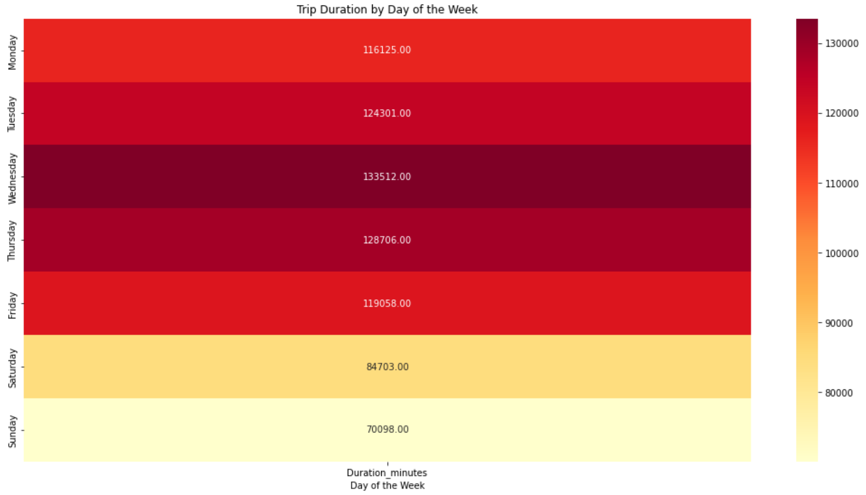

Fig. 11 Heat Map for Trip Duration by Day of the Week

3.5.2.3 Heat Map for Trip Duration by Hourly Rentals by Day of the Week:

To clearly capture the relationship between the trip frequency, time of day, and day of the week. Additional Heatmap is plotted which will visualise hourly rentals across different days of week. Fig. 12 Shows the heatmap of trip by hourly rentals by day of the week.

1. **Weekday Pear Hour:**
   - **Morning Peak:** There is distinct surge in rentals at 8 AM on weekdays, likely corresponding to morning commutes.
   - **Evening Peak:** Another prominent peak occurs from 5 PM to 8PM on weekdays, aligning with evening commute times and post work activities.

2. **Weekend Patterns:**
   - Weekends shows markedly different patterns with fewer overall trips compared to weekdays.
   - weekend activities are more evenly distributed throughout the day, without the sharp peaks seen on weekdays.

3. **Low Activity Hours:**
   - Across all days, the period from midnight to 5AM has lowest rental activity.
   - This low activity period is particularly pronounced on weekdays and weekends.

4. **Midday Usage:**
   - Weekdays show moderate usage during midday hours (10 AM to 4PM)

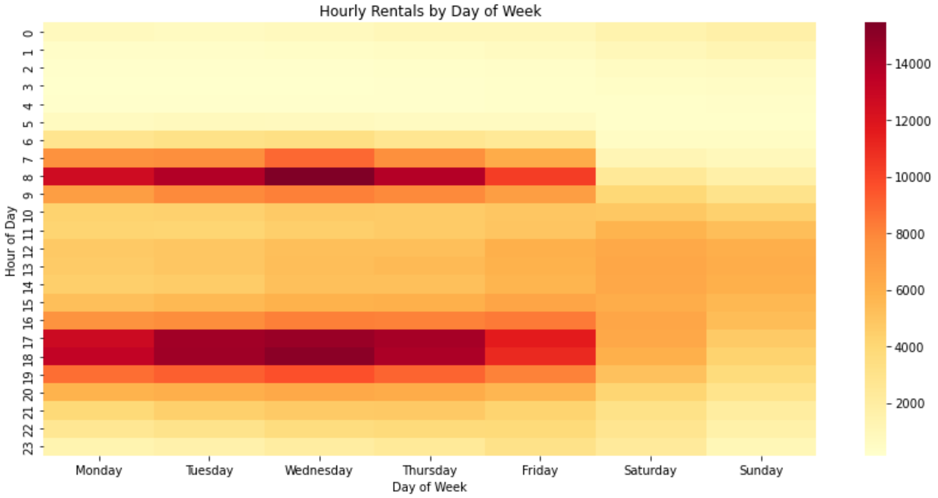

Fig. 12 Heat Map for Trip by hourly rentals by day of the week

3.5.3 Station Popularity vs. Day of Week:

This analysis revels the pattern for the bike rental across different days of week, shows the popularity of top 10 stations. This analysis helps us identify the station having most frequently used and how their usages throughout the week. Fig. 13 Shows the heatmap of Station Popularity vs. Day of Week.

The Heatmap reveals several key patters:

1. **Hyde Park Corner, Hyde Park:**
   - This station shows the highest usage on weekends, with peak activity on Saturday and Sunday.
   - The pattern shown for this station suggest that this is the popular station for leisure activities during the weekends.

2. **Waterloo Station:**
   - Waterloo station 3 and waterloo station 1 shows high usage during the weekdays, particularly on Tuesday, Wednesday and Thursday.
   - This indicates that these stations are heavily used by commuters during the workweek.

3. **Other station:**

- Other popular stations like Black Lion Gate, Kensington Gardens and Albert Gate, Hyde Park, etc also shows significant activities, but their usage is more evenly distributed throughout the week.

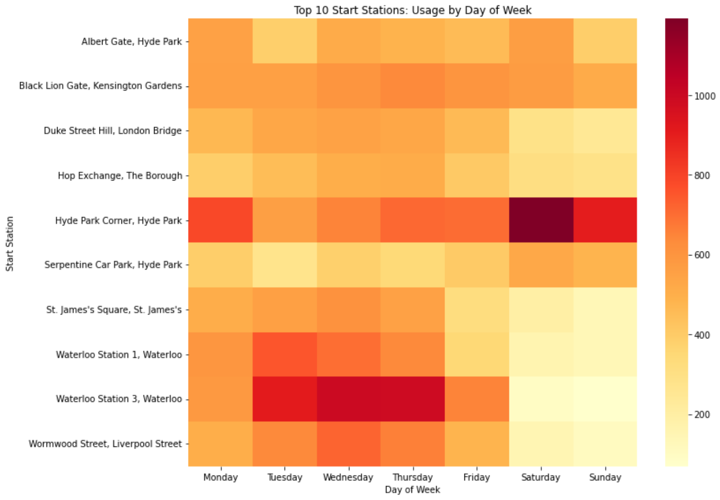

Fig. 13 Heat Map for Station Popularity vs. Day of Week

3.5.4 Normalized Station Popularity Analysis:

To get more nuanced understanding of station usage patterns, normalized analysis for top 10 stations was conducted across different days of week. This analysis allows us to compare stations on relative basis, highlighting their patterns independent of their overall popularity. The heatmap visualize the percentage of weekly usage for each station across different days. Fig. 14 Shows the heatmap for Normalized Station Popularity Analysis

The normalized heatmap reveal several distinct patterns:

1. **Commuter Heavy Stations:**
   - Waterloo Station 3 and Waterloo Station 1 shows the highest relative usage during weekdays.
   - These stations experience significant drop in usage during weekends, suggesting they are primarily serving commuter traffic.

2. **Leisure oriented Stations:**
   - Hyde Park Corner, Hyde Park and Serpentine Car Park, Hyde Park stations exhibit peak usage during weekends.

- This pattern indicated that these stations are popular for recreational activities, particularly on Saturday and Sunday.

3. **Balance Usage Stations:**
    - Several other stations in the Top 10 shows relatively consistent usage throughout the week.
    - This indicates that the users commute and leisure or potentially areas with both residential and commercial activities.

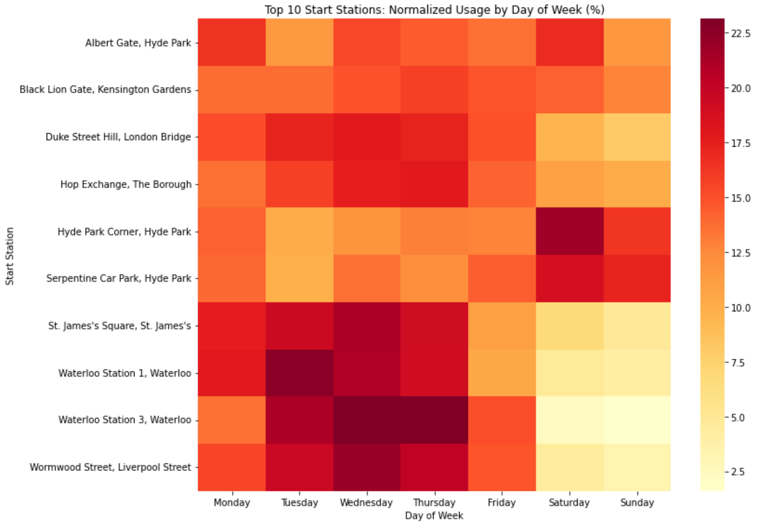

Fig. 14 Heat Map for Normalized Station Popularity Analysis

3.5.5 Daily Usage Trends for Top 10 Start Stations:

The analysis for Daily Usage Trends for Top 10 Start Stations further elucidates the usage patterns for most popular stations, the line plot shows the daily trends for top stations. The line plot directly compares the usage pattern between the stations across different days. Fig. 15 Shows the Line plot of Daily usage trends for Top 10 Stations. Fig. 15 Shows the line plot for Daily Usage Trends for Top 10 Start Stations

The line plot reveals distinct usage patterns among the Top 10 start stations:

1. **Weekend Oriented Stations:**
    - Hyde Park Corner, Hyde Park show unique patter with significant increase in usage during weekends, particularly on Saturday and Sundays.

- This trend suggests that Hyde Park corner is popular destination for recreational activities and weekend outings.

2. **Weekday Dominant stations:**
   - Several stations show clear weekday centric usage patterns, with higher trips from Monday to Friday and low during the weekends:
     a) Waterloo Station 3, Waterloo
     b) St. James's Square, St. James's
     c) Waterloo Station 1, Waterloo
     d) Wormwood Street

This pattern indicates these stations primarily serving commuter traffic.

3. **Stations with Consistent Usage:**
   - Remaining Top stations shows slightly undulating pattern throughout the week.
   - This shows more balanced mix of commuter and leisure usage.

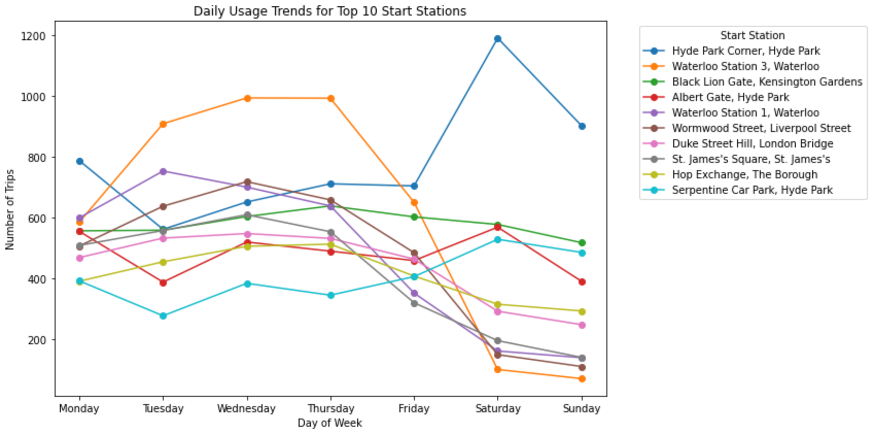

Fig. 15 Line plot for Daily Usage Trends for Top 10 Start Stations

3.6 Findings and Recommendations:

3.6.1 Findings:

The examination of the bike-sharing data sourced from Transport for London (TfL) for June 2023 has provided numerous crucial insights into urban mobility trends and cycling preferences within London. The outcomes are succinctly outlined as follows:

1. **Customer Behaviour Analysis:**
    - **Examination of Rental Durations:** Significantly varied bike usage patterns were identified, with the highest activity noted from June 22nd to 23rd and the lengthiest average trips occurring between June 27th to 29th.
    - **Frequency Assessment:** The daily rental frequency exhibited an erratic trend, with a noticeable decline from June 15th to 18th, succeeded by a sharp upsurge, reaching its peak on June 21st.
    - **Time Series Evaluation:** The Augmented Dickey-Fuller (ADF) test suggested the time series was nearly non-stationary. Seasonal decomposition unveiled a horizontal trend with distinct weekly seasonality, reaching its peak on Wednesdays and declining on Sundays.

2. **Geospatial Analysis:**
    - **Identification of Popular Start and End Stations:** The primary start and end stations were predominantly situated at crucial commuter nodes and recreational spots, such as Hyde Park Corner and Waterloo Station.
    - **Analysis of Popular Routes:** The frequently traversed routes typically initiated and terminated at the same stations, indicating a high volume of intra-station traffic.

3. **Multivariate Analysis:**
    - **Comparison of Trip Duration and Time of Day:** Peak trip durations were witnessed during early mornings and late afternoons, aligning with peak commute hours. Weekdays exhibited more consistent rental patterns in contrast to weekends.
    - **Heat Mapping Analysis:** Heat maps illustrated heightened rental activity at 8 AM and 5-6 PM on weekdays, with reduced activity from midnight to early morning. Weekdays showcased longer trip durations than weekends, particularly on Wednesdays.
    - **Station Popularity in relation to Day of the Week:** Stations like Hyde Park Corner displayed elevated usage levels on weekends, whereas commuter-centric stations like Waterloo experienced peak usage on weekdays.

3.7 Conclusion:

The thorough examination of bike-sharing data for June 2023 has yielded valuable insights into the urban mobility patterns and cycling preferences in London. Through the comprehension of customer behaviour, temporal trends, and spatial characteristics, enhancements can be made to the bike-sharing system to better cater to its users' requirements. The application of the proposed strategies will not only optimize the system's efficiency but also enhance overall user satisfaction, thereby contributing to a more sustainable and user-friendly urban transportation network.

3.8 Future Recommendations:

Based on the insights obtained from these analyses, a multitude of recommendations can be put forward to enhance user satisfaction and optimize the bike-sharing system:

1. **Resource Allocation:**

   Peak Hour Management: It is advised to increase the number of bikes available at stations that face high demand during peak hours (8 AM and 5-6 PM) to accommodate commuter traffic.

   Weekend Strategies: Ensuring sufficient bike availability at popular leisure destinations like Hyde Park Corner on weekends is crucial to meet the demand for recreational activities.

2. **Service Enhancements:**

   Station Expansion: Expanding docking stations in areas with high demand can help reduce congestion and improve system accessibility.

   Dynamic Pricing Implementation: Introducing dynamic pricing mechanisms can encourage bike rentals during off-peak hours and effectively distribute the system's workload.

3. **User Engagement:**

   Customized Promotions: Developing customized promotions and incentives can increase bike usage during periods of low activity and in regions with lower rental rates.

   Collection of Customer Feedback: Regularly collecting and analysing customer feedback is vital for pinpointing areas of improvement and addressing customer concerns in the bike-sharing service.

4. **Operational Efficiency:**

Real-Time Data Utilization: Employing real-time data analytics for continuous monitoring of bike availability and demand patterns allows for timely adjustments in bike distribution and maintenance schedules.

Maintenance Schedule Optimization: Aligning maintenance schedules with low-activity periods ensures a maximum number of bikes are available during peak hours.

4. Predictive Model Using Machine Learning

4.1 Data Collection and Preparation:

For this predictive modelling study, the data was collected and prepared data from two primary sources: bike usage data from Transport for London (TfL) and weather data from Visual Crossing.

4.1.1 Bike Usage Data:

The bike usage data was sourced from TfL's cycling data portal https://cycling.data.tfl.gov.uk/ for June 2023. Three separate CSV files were used:

1. `373JourneyDataExtract05Jun2023-11Jun2023.csv`
2. `374JourneyDataExtract12Jun2023-18Jun2023.csv`
3. `375JourneyDataExtract19Jun2023-30Jun2023.csv`

These files were concatenated into a single Data Frame using pandas' `concat` function.

4.1.1.1 Data Cleaning and Preparation:

1. Date columns ('Start date' and 'End date') were converted to datetime format by using the function 'pd.to_datetime' in python.
2. Removed Duplicate entries using the 'df.drop_duplicates()' function.
3. 'Total duration' was converted to timedelta format using 'pd.to_timedelta' function.
4. 'Total duration (ms)' was converted to minutes for easier interpretation.
5. 'Start station' and 'End station' were split into separate columns for station name and city using the `str.split()` function.
6. Missing values in city columns were filled with 'Unknown'.

4.1.2 Weather Data:

Weather data for London in June 2023 was obtained from Visual Crossing https://www.visualcrossing.com/weather/weather-data-services and was named as 'Weather_data_June.csv'. The original dataset included the following columns:

name, datetime, temp, feelslike, dew, humidity, precip, precipprob, preciptype, snow, snowdepth, windgust, windspeed, winddir, sealevelpressure, cloudcover, visibility, solarradiation, solarenergy, uvindex, severerisk, conditions, icon, stations

For our analysis, the following relevant columns were selected:

- **datetime:** Date and time of the weather observation
- **temp:** Temperature in degree centigrade
- **precip:** Precipitation amount
- **humidity:** Relative humidity
- **windspeed:** Wind speed
- **cloudcover:** Cloud cover percentage
- **visibility:** Visibility distance
- **conditions:** Weather conditions description
- **solarradiation:** Solar radiation
- **uvindex:** UV index

4.1.3 Data Merging:

To prepare the data for predictive modelling:

- "Hyde Park Corner" Station was considered for detailed analysis.
- Bike start times were rounded to the nearest hour using the `dt.round('H')` function.
- The bike usage data and weather data were merged using pandas' `merge_asof` function, which allows for joining on nearest key rather than exact matches.

4.2 Exploratory Data Analysis (EDA):

4.2.1 Frequency of Top 10 Destinantion Stations for Different Bike Models:

For this analysis, bike model and Destination station were grouped and counted the occurrence of each combination for top 10 destination stations for each bike model based on the frequency and visualize using bar plot. This helps to identify most popular destinations for different bike models and provide insights into user preferences. Fig. 16 Shows the Bar plot for Top 10 Destinantion Stations having classic bike model. Fig. 17 Shows the Bar plot for Top 10 Destinantion Stations having PBSC_EBike bike model.

The exploratory data analysis revealed several interesting patterns:

- The phenomenon of the Black Lion Gate being the most visited destination station for the classic bike model underscores its significance as a prominent terminus for journeys utilizing classic bikes within the bike-sharing network.

- The prominence of Harvard Mews as the primary destination station for the PBSC bike model suggests a preference or greater accessibility of PBSC bikes for journeys culminating at Harvard Mews.

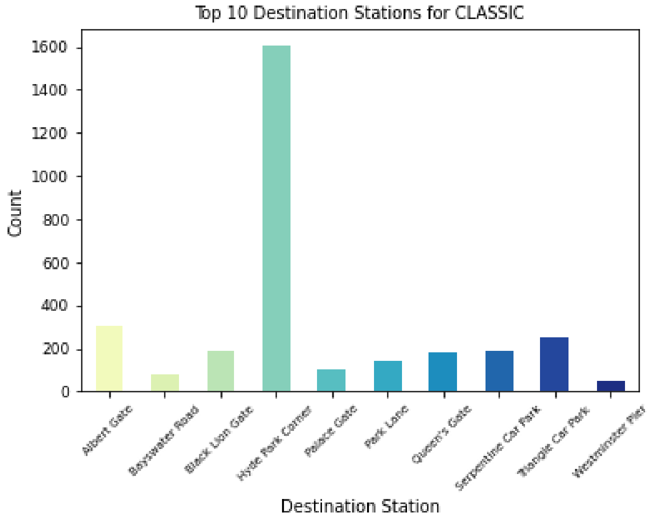

Fig. 16 Bar plot for Top 10 Destinantion Stations having classic bike model

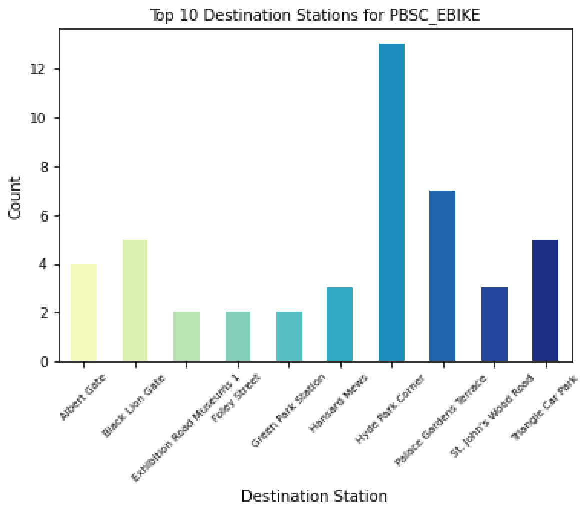

Fig. 17 Bar plot for Top 10 Destinantion Stations having PBSC_EBike bike model

4.2.2 Trip Frequency by Hour of the Day and Day of the Week:

In order to acquire a more profound understanding of the temporal trends in bicycle utilization, an examination was conducted on the frequency of trips based on both the hour of the day and the day of the week. This investigation plays a crucial role in pinpointing the peak periods of usage and specific days, a fundamental aspect for the effective distribution of resources and strategic operational scheduling. Fig. 18 Shows the heatmap for Trip Frequency by Hour of the Day and Day of the Week.

The heatmap reveals several key patterns in trip frequency:

4.3 Peak Hours:

- The highest trip frequency is observed on Saturdays at 5 PM, followed closely by 6 PM on the same day. This suggests that Saturday evenings are particularly popular for bike rentals.
- Another significant peak is seen on Mondays from 1 PM to 4 PM, indicating a high usage period during the early afternoon.

4.4 Low Activity Periods:

- There is minimal to no activity from 11 PM to 7 AM across all days of the week. This indicates that bike usage is very low during late-night and early-morning hours.

4.5 Weekend vs. Weekday Patterns:

- Saturdays show the highest overall trip frequency, particularly in the evening, suggesting recreational or leisure activities.
- Sundays also show high usage, though slightly less than Saturdays.
- Weekdays exhibit a more consistent pattern of usage, with notable peaks during specific hours such as Monday afternoons.

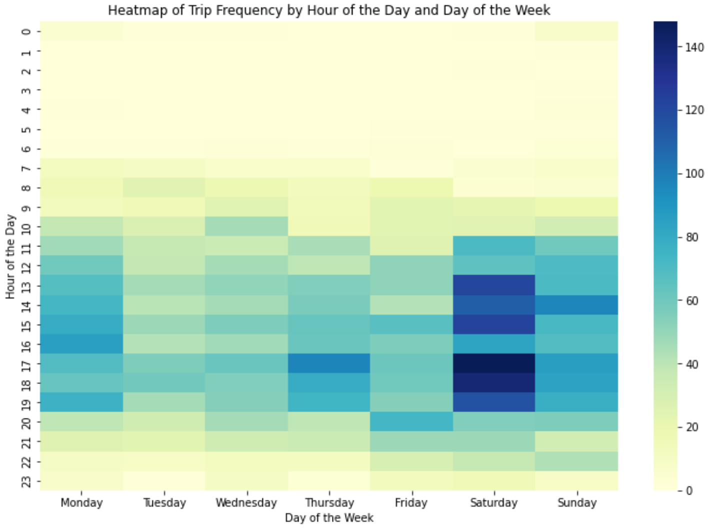

Fig. 18 Heatmap for Trip Frequency by Hour of the Day and Day of the Week.

4.5.1 Trip Frequency for Different Weather Conditions:

During the phase of exploratory data analysis, an inquiry was carried out to establish a correlation between weather conditions and the frequency of trips. This investigation provides valuable insights into the influence of weather patterns on the utilization of bicycle rentals, a crucial aspect for predicting demand and improving service provision. Fig. 19 Illustrates the Bar plot to show Trip Frequency for Different Weather Conditions.

The bar graph illustrates several significant trends in trip frequency across various weather conditions:

1. **Partly Cloudy:** This particular weather condition demonstrates the highest frequency of trips, indicating that moderate cloud cover does not significantly discourage bicycle usage.
2. **Clear:** Weather conditions characterized as clear represent the second most common category for bicycle trips, implying that favourable weather conditions promote bicycle rentals.
3. **Other Conditions:** Weather conditions like overcast, rain, and combinations such as "rain, partly cloudy" and "rain, overcast" exhibit noticeably lower frequencies of trips.

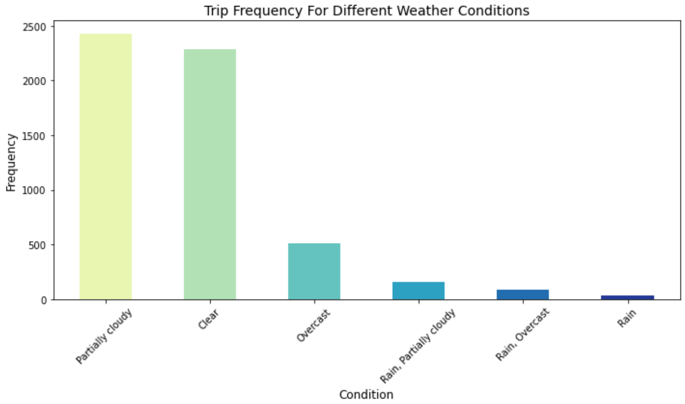

Fig. 19 Bar plot to show Trip Frequency for Different Weather Conditions.

4.5.2 Scatter Plot and Correlation Heatmap:

During the phase of exploratory data analysis, scatter diagrams and a correlation heatmap were produced in order to gain insights into the connections among numerical attributes within the dataset, thus facilitating the endeavours of predictive modelling.

1. **Scatter Plot Matrix:** Depicts the pairwise associations between numerical variables, thus shedding light on potential linear or non-linear relationships.
2. **Correlation Heatmap:** Exhibits the correlation coefficients, thereby pinpointing robust correlations that can function as predictive factors.

This approach aids in the process of selecting features by pinpointing pivotal attributes with significant correlations for predictive modelling. It facilitates the development of models by providing guidance on the selection of algorithms and fine-tuning of parameters, while also offering practical insights for decision-making regarding the distribution and upkeep of bicycles based on weather conditions. Fig. 20 Illustrates the Scatter Plot and Fig. 21 shows the Correlation using Heatmap.

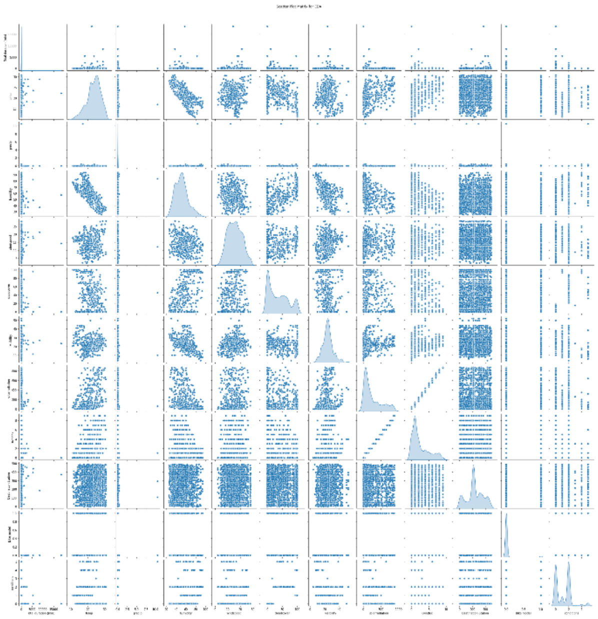

Fig. 20 Scatter Plot for merged data

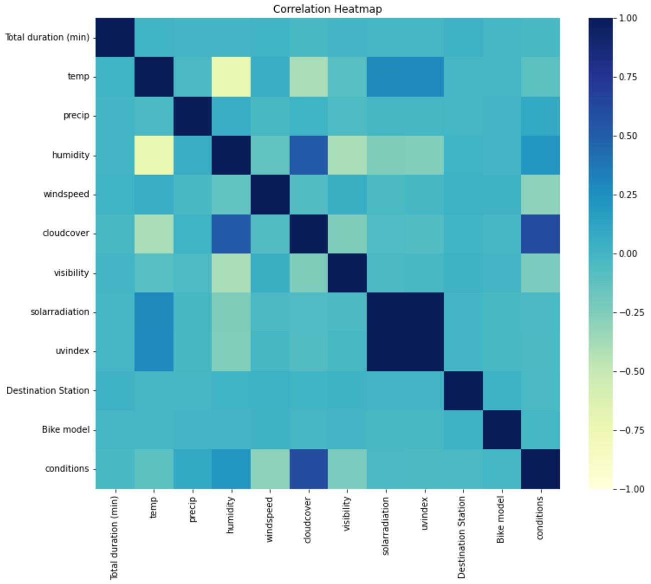

Fig. 21 Correlation Heatmap for merged data

4.6 Feature Engineering:

Feature engineering plays a pivotal role in the predictive modelling procedure, wherein novel features are generated from the available data to enhance the efficacy of machine learning models. The subsequent section delineates the procedures and computations entailed in the feature engineering process utilizing the combined dataset comprising information on bike utilization and weather conditions.

4.6.1 Time-Based Features:

1. **Hour of the Day:** Extracted the hour from the 'Start date' to capture the time of day when the trip started.
2. **Day of the Week:** Extracted the day name from the 'Start date' to identify the day of the week.
3. **Is Weekend:** Created a binary feature indicating whether the trip started on a weekend (Saturday or Sunday).
4. **Time of Day Categories:** Categorized the hour into four periods: morning (5 AM to 12 PM), afternoon (12 PM to 5 PM), evening (5 PM to 9 PM), and night (9 PM to 5 AM).

4.6.2 Weather-Related Features:

1. **Apparent Temperature:** Calculated the apparent temperature using the formula:

$$\text{apparent\_temp} = \text{temp} + 0.33 \times \text{humidity} - 0.70 \times \text{windspeed} - 4.00$$

2. **Solar Radiation and UV Index Interaction:** Created a feature representing the interaction between solar radiation and UV index:

$$\mathbf{solar\_uv\_interaction = solarradiation \times uvindex}$$

3. **Clear Sky Indicator:** Created a binary feature indicating clear sky conditions based on cloud cover and visibility:

$$\mathbf{clear\_sky = (cloudcover < 20) \& (visibility > 10)}$$

4. **Humidity Category:** Categorized humidity into three levels: Dry (0-30%), Comfortable (30-60%), and Humid (60-100%).

5. **Weather Comfort Index:** Calculated a composite weather comfort index:

$$\boldsymbol{weather\_comfort\_index = temp - 0.55 \times \left(1 - \frac{humidity}{100}\right) \times (temp - 14.5) - 0.2 \times windspeed + 0.1 \times cloudcover - \alpha \times precip}$$

4.6.3 Trip Duration Features:

1. **Trip Duration Category:** Categorized trip duration into six bins: <5 min, 5-15 min, 15-30 min, 30-60 min, 1-2 hours, and >2 hours.
2. **Interaction with Time of Day and Day of Week:** Calculated the average trip duration for each time of day and day of the week.

4.6.4 Data Encoding and Preparation

1. **Label Encoding:** Converted categorical columns to numerical values using `LabelEncoder`.
2. **Correlation Matrix:** Calculated the correlation matrix for the engineered features to understand their relationships.

Fig. 22 Shows the Correlation Heatmap for the new Features based on which the predictive models using machine learning is built.

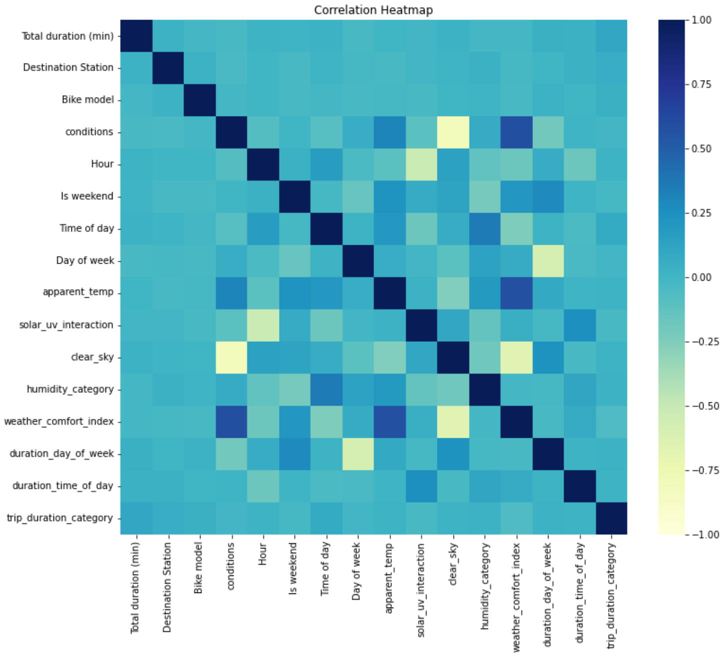

Fig. 22 Correlation Heatmap for the new Features

4.7 Machine Learning Model Building:

4.7.1 Improving resource allocation:

The process of predictive modelling encompassed the development of a sturdy model for predicting bicycle demand and enhancing the allocation of resources. This segment delineates the actions carried out, such as data preprocessing, feature manipulation, model fitting, cross-validation, and comparing outcomes.

4.7.1.1 Data Preparation:

The dataset was prepared to facilitate the development of an accurate predictive model. The following steps were undertaken:

- **Calculate Trip Counts:** The number of trips starting and ending at each station for each hour was calculated.
- **Merge Data:** The calculated trip counts were merged with the original dataset to include features for demand prediction.

4.7.1.2 Feature Engineering:

1. **Time-Based Features:**
   - Hour of the Day: Extracted from the 'Start date'.
   - Day of the Week: Extracted from the 'Start date'.
   - Is Weekend: Binary feature indicating if the trip started on a weekend.
   - Time of Day Categories: Categorized into morning, afternoon, evening, and night.

2. **Weather-Related Features:**
   - Conditions, humidity category, and weather comfort index were included.

3. **Data Encoding and Scaling:**
   - Categorical features were encoded using `OneHotEncoder`.
   - Numerical features were scaled using `StandardScaler`.

4.7.1.3 Model Training and Testing:

The data was split into training and testing sets:

```
X = merged_data[features]
y = merged_data[target]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

4.7.1.4 Model Selection and Cross-Validation

Several regression models were evaluated using pipelines and cross-validation to identify the best-performing model:

1. **Pipelines:** Pipelines were defined for each model, incorporating pre-processing steps and the estimator.
2. **Hyperparameter Tuning:** Hyperparameter grids were defined, and `RandomizedSearchCV` was used for hyperparameter tuning.

3. **Cross-Validation:** Each model was evaluated using 3-fold cross-validation, optimizing for negative mean squared error.

4.7.1.5 Results and Comparison

Table 1. Comparison of Models based on RMSE, R^2 score and Time taken(s) for Improving resource allocation machine learning model

Model	RMSE	R^2	Time Taken(s)
Ridge Regression	11.31	0.10	0.72
Lasso Regression	11.29	0.10	0.72
Decision Tree	10.69	0.19	3.52
Random Forest	10.57	0.21	371.36
Gradient Boosting	10.53	0.22	99.01
XGBoost	10.51	0.22	16.56
LightGBM	10.53	0.22	8.73

- **Best Performance:** XGBoost and LightGBM (RMSE: 10.51, R^2 Score : 0.22)
- **Strong Performance:** Gradient Boosting (close to XGBoost and LightGBM)
- **Quick Training:** Ridge and Lasso Regression (shortest training times, lower accuracy)
- **Long Training:** Random Forest (longest training time, moderate performance)
- **Recommendations:**
  1. Preferred: XGBoost for best balance of accuracy and training time
  2. Others: Gradient Boosting, LightGBM for quick and accurate predictions
  3. Future Work: Fine-tuning and feature engineering for improved performance

4.7.2 Predict Bike Model:

The process of predictive modelling encompassed the development of a sturdy model for predicting bicycle model and enhancing the allocation of the specific bike models. This segment delineates the actions carried out, such as data preprocessing, feature manipulation, model fitting, cross-validation, and comparing outcomes.

4.7.2.1 Data Preparation:

The dataset was prepared to facilitate the development of an accurate predictive model. The following steps were undertaken:

1. **Feature and Target Definition:**
   - Features: 'Destination Station', 'conditions', 'Time of day', 'weather_comfort_index'
   - Target: 'Bike model'
2. **Handling Missing Values:**
   - Filled numerical columns with the mean.
   - Filled categorical columns with the mode.
3. **Encoding Categorical Variables:**
   - Used `ColumnTransformer` and `OneHotEncoder` to encode categorical features.
   - Encoded the target variable 'Bike model' using `LabelEncoder`.
4. **Data Splitting:**
   - Split the dataset into training and testing sets with an 80-20 ratio.

4.7.2.2 Model Training and Evaluation:

Bike model predictive using machine learning models is performed using different classification models using pipelines and hyperparameter tuning through Randomized Search. The models include:

- Ridge Classifier
- Logistic Regression (Lasso)
- Decision Tree Classifier
- Random Forest Classifier
- Gradient Boosting Classifier
- XGBoost Classifier
- LightGBM Classifier

4.7.2.3 Model Pipelines and Hyperparameter Tuning:

Each model was integrated into a pipeline that included preprocessing steps and the classifier itself. Hyperparameter grids were defined for each model, and Randomized Search was used to find the best parameters.

4.7.2.4 Results and Comparison:

Table 2. Comparison of Models based on Accuracy and Time taken(s) for Predict Bike Model using machine learning model

Model	Accuracy	Time Taken(s)
Ridge Regression	0.9782	0.53
Lasso Regression	0.9764	0.49
Decision Tree	0.9746	1.16
Random Forest	0.9773	41.66
Gradient Boosting	0.9782	45.45
XGBoost	0.9791	12.43
LightGBM	0.9782	8.94

- **Accuracy:** XGBoost Classifier: Highest accuracy (0.9791) and Ridge Classifier, Gradient Boosting, LightGBM: Similar accuracy (~0.9782)
- **Training Time:**
  Fastest: Logistic Regression (Lasso) and Ridge Classifier (under 1 second)
  Longest: Random Forest and Gradient Boosting (over 40 seconds)
- **Model Selection:**
  Recommended: XGBoost for highest accuracy
  Competitive Options: Ridge Classifier and LightGBM for faster training times

4.7.3 Predicting Hourly Demand:

This process of prediction of hourly bike rental demand utilizing various machine learning regression models. The following steps comprehensively detail the processes involved in data preparation, model training, and evaluation.

4.7.3.1 Data Preparation:

1. **Aggregate Data:** Aggregated the data to create demand metrics by grouping the data by 'Destination Station' and 'Hour' to calculate the demand.

2. **Merge Demand Data:** Merged the demand metrics with the original dataset to include the demand as a target variable.

3. **Handling Missing Values:**
   - Filled numerical columns with the mean.
   - Filled categorical columns with the mode.

4. **Encoding Categorical Variables:** Used `ColumnTransformer` and `OneHotEncoder` to encode categorical features.

5. **Data Splitting:** Split the dataset into training and testing sets with an 80-20 ratio.

4.7.3.2 Model Training and Evaluation:

Predicting hourly demand using machine learning models is performed using different classification models using pipelines and hyperparameter tuning through Randomized Search. The models include:

- Ridge Classifier
- Logistic Regression (Lasso)
- Decision Tree Classifier
- Random Forest Classifier
- Gradient Boosting Classifier
- XGBoost Classifier
- LightGBM Classifier

4.7.3.3 Model Pipelines and Hyperparameter Tuning:

Each model was integrated into a pipeline that included preprocessing steps and the regressor itself. Hyperparameter grids were defined for each model, and Randomized Search was used to find the best parameters.

4.7.3.4 Results and Comparison

Table 3. Comparison of Models based on RMSE, R^2 score and Time taken(s) for Improving resource allocation machine learning model

Model	RMSE	R^2	Time Taken(s)
Ridge Regression	20.90	0.89	1.00
Lasso Regression	21.05	0.88	0.73
Decision Tree	1.25	1.00	3.49
Random Forest	0.95	1.00	419.52
Gradient Boosting	1.03	1.00	77.23
XGBoost	0.99	1.00	24.31
LightGBM	2.01	1.00	8.12

1. **Accuracy:**
   - **Top Performers:** Random Forest, Gradient Boosting, XGBoost, and LightGBM models achieved perfect R^2 scores of 1.00.
   - **Decision Tree:** Also performed well with an R^2 score of 1.00 and an RMSE of 1.25.

2. **Training Time:**
   - **Fastest Models:** Lasso Regression and Ridge Regression trained in under one second.
   - **Longest Training:** Random Forest exceeded 419 seconds.

3. **Model Selection:**
   - **Recommended Models:** XGBoost and LightGBM for their high accuracy and reasonable training times.
   - **Quick Predictions:** Ridge Regression and Lasso Regression for faster, slightly less accurate predictions.

4.8 Conclusion:

The analysis of bike-sharing data from Transport for London (TfL) in June 2023 through predictive modelling revealed valuable insights regarding urban mobility and the effectiveness of different machine learning models. The integrity of the dataset was ensured through a thorough process of comprehensive data collection and precise cleaning. Through Exploratory Data Analysis (EDA), user preferences, peak usage times (evenings and weekends), and the impact of weather on bike usage were identified. Predictions were enhanced by incorporating time-based and weather-related attributes through feature engineering. XGBoost and LightGBM demonstrated high accuracy and efficiency in training for resource allocation, with XGBoost leading in bike model prediction, followed closely by Ridge Classifier and LightGBM. In terms of hourly demand forecasting, Random Forest, Gradient Boosting, XGBoost, and LightGBM achieved outstanding R^2 scores, with XGBoost and LightGBM being recommended for their efficiency.

4.9 Future Recommendation:

Based on the findings from this study, several recommendations are proposed to further enhance the bike-sharing system in London:

1. **Resource Allocation:**
   - **Peak Hour Management:** Increase the number of bikes available at high-demand stations during peak hours (8 AM and 5-6 PM) to accommodate commuter traffic.
   - **Weekend Strategies:** Ensure sufficient bike availability at popular leisure destinations like Hyde Park Corner on weekends to meet recreational demand.

2. **Service Enhancements:**
   - **Station Expansion:** Expand docking stations in high-demand areas to reduce congestion and improve system accessibility.
   - **Dynamic Pricing:** Implement dynamic pricing mechanisms to encourage bike rentals during off-peak hours, effectively distributing the system's workload.

3. **User Engagement:**

- **Customized Promotions:** Develop targeted promotions and incentives to increase bike usage during low-activity periods and in regions with lower rental rates.
- **Customer Feedback:** Regularly collect and analyze customer feedback to identify areas of improvement and address user concerns.

4. **Operational Efficiency:**

- **Real-Time Data Utilization**: Employ real-time data analytics for continuous monitoring of bike availability and demand patterns, allowing for timely adjustments in bike distribution and maintenance schedules.
- **Maintenance Schedule Optimization:** Align maintenance schedules with low-activity periods to ensure maximum bike availability during peak hours.

5. **Model Enhancement:**

- **Further Tuning:** Conduct additional fine-tuning and evaluation of models to optimize performance.
- **Feature Engineering:** Incorporate additional features and more extensive datasets to enhance model accuracy and predictive power.

6. Code snippets

6.1 Data Preparation

1. Import Libraries: Utilized `pandas`, `seaborn`, and `matplotlib` for data handling and visualization.
2. Read Data: Loaded CSV files for June 2023 bike usage data.
3. Concatenate Data: Combined multiple dataframes into one unified dataframe.
4. Remove Duplicates: Ensured data integrity by removing duplicate entries.
5. Datetime Conversion: Converted `Start date` and `End date` columns to datetime format.
6. Duration Conversion: Converted `Total duration` to a `timedelta` format for easier manipulation.
7. Filter Data: Filtered the dataframe to include only data from June 2023.
8. Resampling: Aggregated data at 3-day intervals to analyze rental patterns over time.

6.2 Customer Behaviour Analysis

1. Rental Duration Analysis:
   - Plotted total and average rental durations over time.
   - Identified peak and low usage periods in June 2023.
2. Frequency Analysis:
   - Visualized daily rental frequency.
   - Noted irregular patterns and significant fluctuations in bike usage.
3. Time Series Analysis:
   - Performed Augmented Dickey-Fuller (ADF) test for stationarity.
   - Conducted seasonal decomposition to identify trends and seasonality.
   - Created Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots to analyse correlations within the data.

6.3 Geospatial Analysis:

1. Popular Routes:
   - Identified top start and end stations, and the most popular routes.
   - Visualized these patterns using bar plots.
2. Station Popularity:
   - Analysed station usage patterns across different days of the week.
   - Created heatmaps and line plots to show daily trends for top start stations.

6.4 Multivariate Analysis

1. Trip Duration vs. Time of Day:
   - Generated strip plots to visualize trip durations by hour and day of the week.
2. Heat Maps:
   - Created heatmaps to illustrate trip duration by hour of the day and day of the week.
   - Analysed station popularity by day of the week.

6.5 Predictive Model Using Machine Learning

1. Data Collection and Preparation:
    - Merged bike usage data with weather data.
    - Performed feature engineering to create new time-based and weather-related features.
2. Model Training and Evaluation:
    - Evaluated various regression models for predicting hourly demand and total trip duration.
    - Used pipelines and cross-validation for model selection and hyperparameter tuning.
    - Compared models based on RMSE, R² score, and training time.
3. Predict Bike Model:
    - Evaluated classification models for predicting bike models.
    - Compared models based on accuracy and training time.

References

1. Cervero, R., Denman, S., & Jin, Y. (2019). Network design, built and natural environments, and bicycle commuting: Evidence from British cities and towns. 74. https://doi.org/10.1016/J.TRANPOL.2018.09.007

2. Tfl.gov.uk. (n.d.). Tfl.gov.uk. Transport for London. https://www.tfl.gov.uk/corporate/about-tfl/what-we-do

3. Ngo, T. T., Pham, H. T., Acosta, J. G., & Derrible, S. (2022). Predicting Bike-Sharing Demand Using Random Forest. https://doi.org/10.58845/jstt.utt.2022.en65

4. Heydari, S., Konstantinoudis, G., & Behsoodi, A. W. (2021). Effect of the COVID-19 pandemic on bike-sharing demand and hire time: Evidence from Santander Cycles in London. 16(12). https://doi.org/10.1371/JOURNAL.PONE.0260969

5. Feng, Y., & Wang, S. (2017). A Forcast for Bicycle Rental Demand Based on Random Forests and Multiple Linear Regression. https://ieeexplore.ieee.org/document/7959977

6. Data Migration and Cleaning. (2024, July 24). Data Migration and Cleaning. https://doi.org/10.1007/979-8-8688-0230-0_5

7. Colin Cameron, A. and Windmeijer, F. A. G., “An R-squared measure of goodness of fit for some common nonlinear regression models,” J. Econometrics, 77(2), 329–342(1997). https://doi.org/10.1016/S0304-4076(96)01818-0 Google Scholar

8. DecisionBrain. "Bike Sharing." Available at: https://decisionbrain.com/bike-sharing/ (Accessed: 22 May 2024).

9. Ding, C., Wang, D., Ma, X. and Li, H., “Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees,” Sustain., 8(11), (2016). Google Scholar

10. Eberly, L. E., [Topics in Biostatistics], Humana Press, Totowa, 165–187(2007).Eberly, L. E., [Topics in Biostatistics], Humana Press, Totowa, 165–187(2007).

11. Eren, E. and Uz, V. E., “A review on bike-sharing: The factors affecting bike-sharing demand,” Sustain. Cities Soc., 54, (2020). https://doi.org/10.1016/j.scs.2019.101882 Google Scholar

12. Feng, Y. and Wang, S., “A forecast for bicycle rental demand based on random forests and multiple linear regression,” Proc. IEEE/ACIS International Conference on Computer and Information Science, ICIS, 101–105(2017).

13. Feasibility study for a central London cycle hire scheme, Final report, November 2008, Transport for London

14. Gu, K.Y., Lin, Y., Prediction for bike-sharing demand in London using multiple linear regression and random forest. Proceedings Volume 12803, Fifth International Conference on Artificial Intelligence and Computer Science (AICS 2023); 128031I (2023) https://doi.org/10.1117/12.3009514

15. London Datastore. "Number of Bicycle Hires," created July 2010.

16. Marill MD, K. A., "Advanced statistics: Linear regression, Part II: Multiple linear regression," Acad. Emerg. Med., 11(1), 94–102(2008). Google Scholar

17. Raschka, S. and Mirjalili, V., [Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2], Packt Publishing, Brimingham, (2019).

18. Transport for London. Feasibility study for a central London cycle hire scheme, Final report, November 2008.

19. Xu, Y., "Research and implementation of improved random forest algorithm based on Spark," Proc. IEEE International Conference on Big Data Analysis, ICBDA, 499–503(2017).

20. Yao, H., Wu, F., Ke, J., Tang, X., Jia, Y., Lu, S., Gong, P., Ye, J. and Li, Z., "Deep multi-view spatial-temporal network for taxi demand prediction," Proc. AAAI, 32(1), (2018). Google Scholar

21. Zhai, H., Cui, L., Nie, Y., Xu, X. and Zhang, W., "A comprehensive comparative analysis of the basic theory of the short term bus passenger flow prediction," Symmetry, 10(9), (2018). https://doi.org/10.3390/sym10090369 Google Scholar