# Uncovering Demand Drivers in Bike-Sharing: Forecasting Ridership and Strategic Insights

Submitted by:

Al Rafi | AXR230065

**Date:** December 2, 2024
**Course:** MIS 6356.004 - Business Analytics With R

## Executive Summary

This report analyzes the key factors influencing demand in a bike-sharing system using advanced data analysis methods. Leveraging a comprehensive dataset, we uncover insights into seasonal patterns, user behaviors, and environmental influences. The analysis provides actionable strategies for optimizing fleet management, pricing, and marketing efforts. Key findings include the impact of temperature and seasonality on demand, distinct usage patterns among casual and registered users, and clustering insights that highlight user segmentation.

## Project Motivation

Bike-sharing systems have emerged as a sustainable urban mobility solution, addressing congestion, environmental concerns, and the demand for flexible transportation. However, optimizing these systems requires a deep understanding of demand dynamics to ensure efficient operations and enhance user satisfaction.

This project seeks to uncover the drivers of bike rentals through data-driven analysis, enabling actionable insights for better decision-making. By leveraging user segmentation, temporal patterns, and environmental factors, we aim to deliver strategies that elevate the efficiency and effectiveness of bike-sharing systems.

To achieve this, the project focuses on four primary objectives:

1. Predict Demand: Analyze demand fluctuations to enhance resource planning and anticipate peak usage periods.
2. Optimize Pricing: Develop dynamic pricing strategies tailored to peak and off-peak times, maximizing revenue while encouraging usage.
3. Improve Fleet Usage: Ensure optimal bike availability, reduce idle inventory, and minimize operational inefficiencies.
4. Design Targeted Promotions: Create customized marketing campaigns for casual and registered users based on their distinct usage trends.

By addressing these objectives, this project aspires to empower bike-sharing operators with the tools and insights necessary for sustainable growth and improved customer satisfaction.

## 1.0 Dataset Description

We have selected the following dataset from the UCI repository. The dataset contains 2 years of daily and hourly data on Bike Rentals, as well as data for casual and registered bike renters. In addition, we have 13 features that strongly correlate with each other, including:

- Temporal variables (hour, day, month, year)
- Seasonal indicators
- Environmental factors (temperature, humidity, windspeed)
- User segmentation (casual vs. registered)

This dataset enables a holistic analysis of demand patterns, revealing both macro (seasonal trends) and micro (hourly usage) dynamics.

## 2.0 Data Pre-Processing and Exploratory Insights

### 2.1 Pre-Processing

1. We have merged the two CSV files to have one single point of data frame with which to work.
2. We have looked for NULL/Missing values to which we did not find any hence we were spared of any sort of data cleaning.

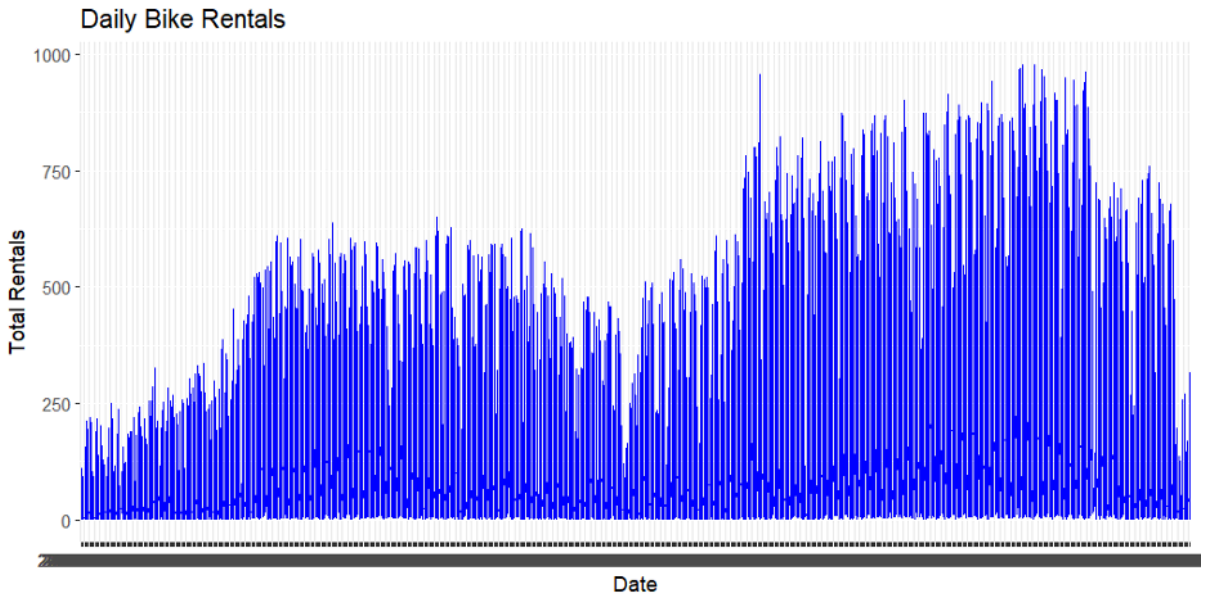
### 2.2 Exploratory Insights/Rough Findings

#### 2.2.1 Summary Statistics

Here are some surface-level insights extracted from the summary statistics of the dataset that we have gathered after merging the daily and hourly datasets.

1. There are 5 temporal variables available in the dataset, along with seasonal information, and environmental factors such as ‘temp’, ‘atemp’ etc.
2. In addition to that, from the ‘cnt’ data we gather two user types: Casual’ and ‘Registered’. A mean of 35.68 for Casual users and 153.8 for registered users indicates that registered users contribute more towards the user base all throughout the year.

### 2.2.2 Total counts of Rentals per day

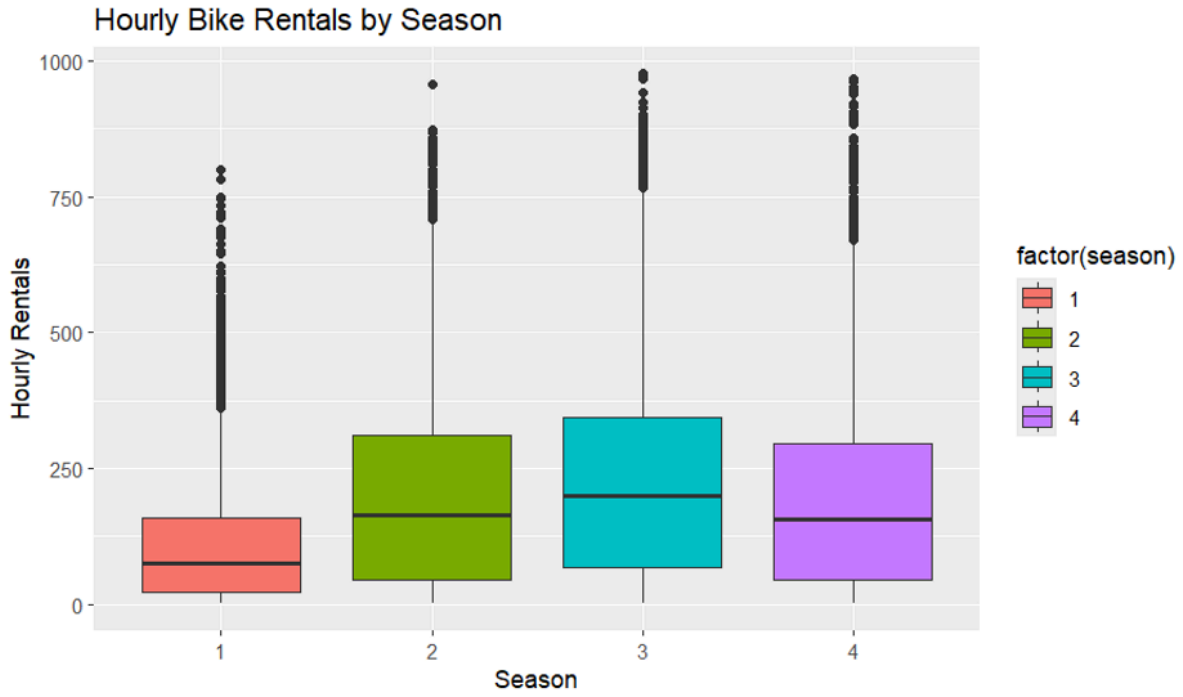



We plotted the total counts of rental per day using a line graph and extrapolated the following insights:

1. Seasonal Trend: The rentals appear to have an upward trend over time, particularly from the start of the data collection period. This might indicate an increasing user base or growing popularity of the bike-sharing program.
2. Cyclic Patterns: There are cyclical fluctuations in the rental numbers, possibly corresponding to Seasonal changes (eg: winter/summer) and Weekly trends (eg: higher usage on weekdays, weekends, or holidays).
3. Rentals near the beginning of the dataset are low but increase substantially in later periods.

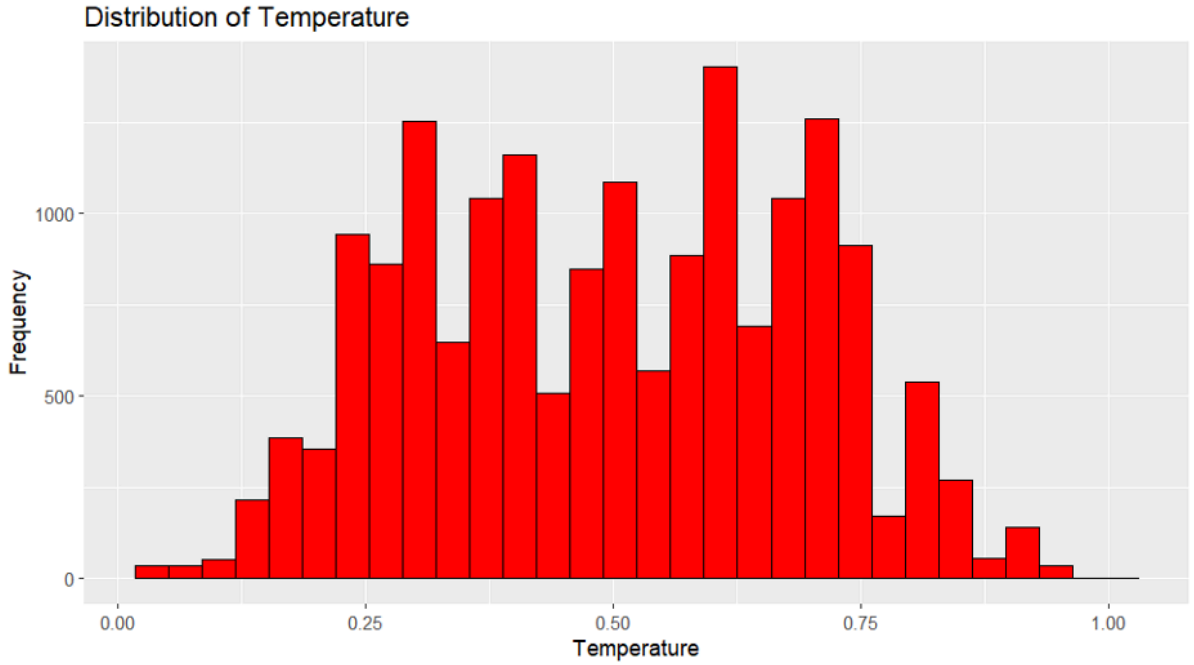
### 2.2.3 Hourly Rentals by Season

We used boxplot to plot hourly bike rentals by season and derived the following:




1. Seasonal trend: The rentals appear to have an upward trend over time, particularly from the start of the data collection period. This might indicate an increasing user base or growing popularity of the bike-sharing program.
2. Cyclic Patterns: There are cyclical fluctuations in the rental numbers, possibly corresponding to Seasonal changes (eg: winter/summer) and Weekly trends (eg: higher usage on weekdays or weekends or holidays).
3. Rentals near the beginning of the dataset are low but increase substantially in later periods.

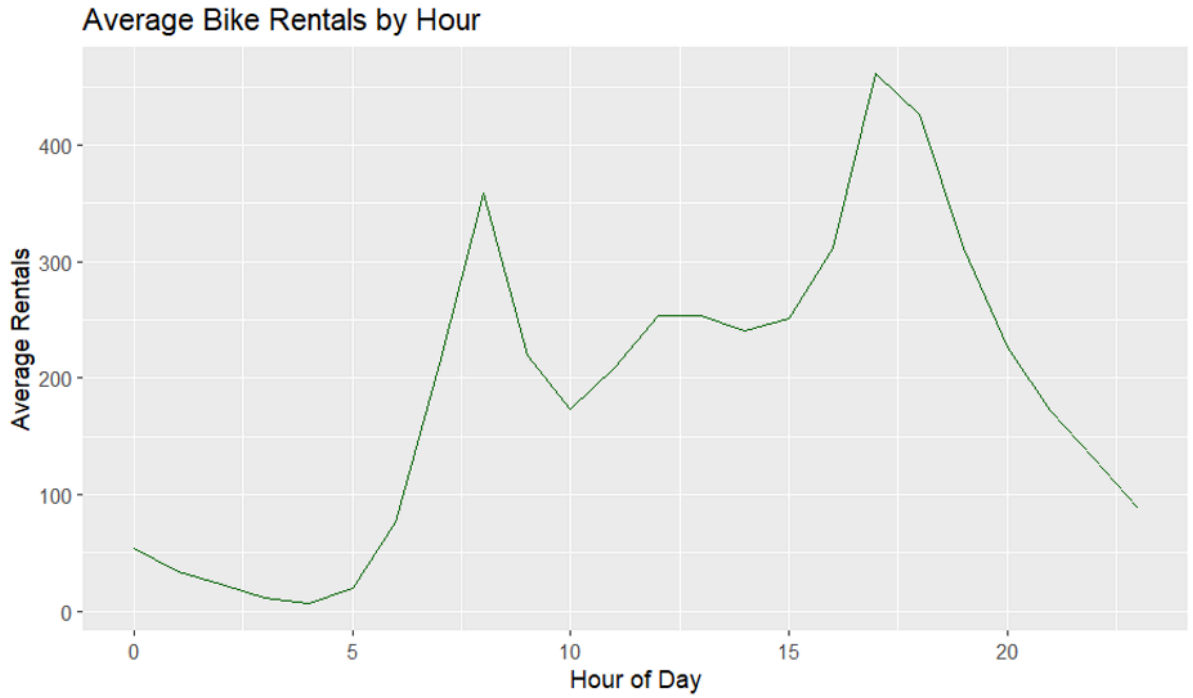
### 2.2.4 Distribution of temperatures against frequency




This histogram seemingly helps us understand and validate the relationship between temperature and the frequency of bike rentals. Here are some insights that are extracted from here:

1. The histogram shows temperature data is centered between 0.3 and 0.7 which if translated to normal temperature measures circles around 15 Celsius to 27 Celsius. These temperatures are likely conducive to higher bike rentals.
2. When temperatures are extreme or below 0.25 and above 0.75 we notice the frequency drops substantially.

### 2.2.5 Usage patterns by hour of the day




The line graph plot will now help us get a granular grasp of what the rental pattern looks like during specific hours of the day.

1. Low Rentals (Late Night and Early Morning): Rentals are minimal between 12 AM (midnight) and 6 AM, likely due to low activity during late-night and early-morning hours.
2. Morning Peak (Commuting Hours): There is a sharp increase in bike rentals between 6 AM and 9 AM, peaking around 8 AM. This corresponds to morning commute hours, indicating that many users likely use bikes to travel to work or school.
3. Midday Decline: Rentals decline after the morning peak and remain relatively stable between 10 AM and 3 PM. This suggests reduced bike usage during non-commuting hours, possibly for leisure or casual purposes.
4. Evening Peak (Commuting Hours): Rentals rise sharply again in the late afternoon, peaking around 5-6 PM. This corresponds to the evening commute, mirroring the morning peak.
5. Decline at Night: Rentals drop steadily after 6 PM, reaching very low levels by 10 PM, as activity slows down for the day.

### 2.3 EDA Findings

Based on the exploratory methods described here we also plotted a correlation matrix and we have received similar insights that can be summarized to suggest a few business avenues that we can explore.

1. Temporal Usage Trends along with Seasonal and Weekly cycles can be identified to optimize certain business processes such as Fleet availability, Maintenance Scheduling, etc.
2. The majority of users are registered users who make up the majority of the pie. Knowing this information, allows us to develop strategies such as loyalty programs to retain these users. Simultaneously, allows budgeting for promotional activities and optimizes the allocation of relevant resources.
3. The environmental sensitivity insights tell us, that offering incentives during unfavorable weather or encouraging users to do maintenance during low-demand periods, or developing different pricing strategies for different weather situations are possible use cases for analyzing this data.
4. Predictive models based on the data will allow us to optimize bike redistribution, reduce idle inventory, forecast off-peak demands and allow us to see through a lot more lenses than before.

To summarize, the EDA gives us the idea that we can use such kind of data to optimize a lot of the business process and take it up a notch by developing certain predictive models to predict demand, and forecast certain aspects of the business to elevate business processes.

### 3.0 Data Mining Objective

Primary Objective

The primary objective of this research is to develop predictive models that accurately forecast bike rental demand by leveraging temporal, seasonal, and environmental data. These models aim to provide actionable insights for real-life business applications, including:

1. Demand Prediction: Understanding fluctuations in bike rental demand across different times and conditions.
2. Pricing Optimization: Use regression models to understand price sensitivity based on environmental and temporal factors. Develop dynamic pricing models for peak and off-peak hours.
3. Fleet and Resource Optimization: Enhancing bike availability and operational efficiency during peak hours and seasons.
4. Promotion Design: Creating targeted marketing strategies based on user behavior, seasonal trends, and environmental conditions.

These insights will enable bike-sharing businesses to make data-driven decisions, maximize resource utilization, and improve overall service efficiency.

## 4.0 BI Model

### 4.1 Linear Regression Model:

In our analysis, we are implementing three distinct linear regression models to gain comprehensive insights into bike rental patterns:

1. Total Bike Rentals Model
2. Casual Users Model
3. Registered Users Model

These models allow us to understand the factors influencing overall demand as well as the specific behaviors of casual and registered users.

Our approach of using separate models for total, casual, and registered users aligns with the work of Faghih-Imani et al. (2014), who developed distinct models for different user types in Montreal's bike-sharing system3. This segmentation allows for a more nuanced understanding of user behavior and helps in tailoring strategies for different user groups.

Bike Rental Model:

**`linear_model <-** ***lm(cnt ~ temp + hum + windspeed + season + weathersit + holiday + hr*workingday + mnth, data = train.df)`***

In the case of, the casual and registered bike rental model we substitute the target variable with the casual and the registered user count.

#### 4.1.1 Key Observations: Total Bike Rental Count

1. Temperature is the most important predictor with a large positive coefficient (290.5596) and a highly significant ($p < 0.001$), indicating that an increase in temperature positively impacts rentals.
2. Humidity is a strong negative predictor with a large negative coefficient (-118.35), and it shows that high humidity decreases bike rentals.
3. Rentals vary by season, with higher rentals in warmer seasons. The positive coefficients confirm seasonality as a relevant factor. Spring (39.21), Summer (28.97), and Fall (61.65): All significantly increase rentals compared to Winter.
4. Some specific hours on working days significantly affect bike rentals. Morning hours (e.g., hr9 to hr12): Positive coefficients suggest high demand during these periods on working days. Late evening/night hours (e.g., hr1, hr2): Negative coefficients reflect low usage.

#### 4.1.2 Key Observations: Casual Bike Rental Count

1. Casual users are more influenced by holidays, weekends, and weather conditions.
2. Significant Predictors: Temperature, Humidity, Season, Workingday and Hour interaction
3. The RMSE value for the casual model (27.9) is low.

### 4.1.3 Key Observations: Casual Bike Rental Count

1. Registered users exhibit more stable behavior tied to working hours and commuting patterns. Hence Working day and hour-specific effects are more critical for registered users.
2. Significant Predictors: Temperature, Humidity, Season, Workingday and Hour interaction
3. The RMSE value for the registered model (73.88) is low

### 4.2 Time Series

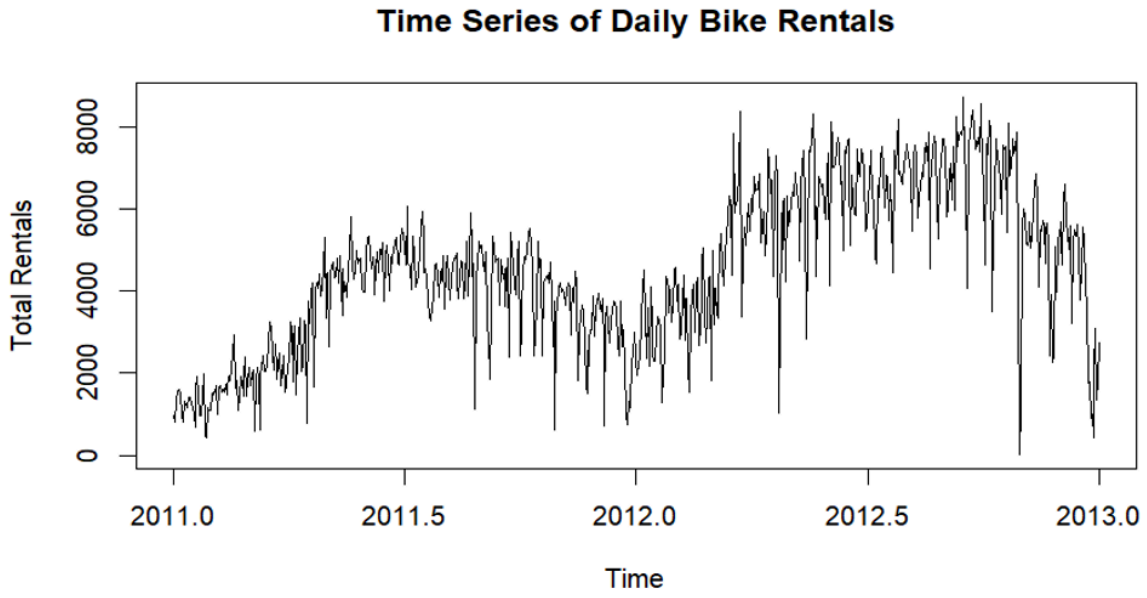
The primary objective of this time series analysis is to develop a robust forecasting model for bike-sharing demand. By accurately predicting future rental patterns, the aim is to:
Enhance operational efficiency by optimizing bike allocation and maintenance schedules., improve user experience by ensuring adequate bike availability during peak demand periods, and support strategic decision-making for expansion and marketing initiatives.

The use of ARIMA models for bike-sharing demand forecasting is well-supported in the literature. For instance, Ashqar et al. (2019) demonstrated the effectiveness of ARIMA models in predicting bike-sharing demand, highlighting their ability to capture both seasonal and trend components.

Our time series analysis methodology for forecasting bike rental demand involves:

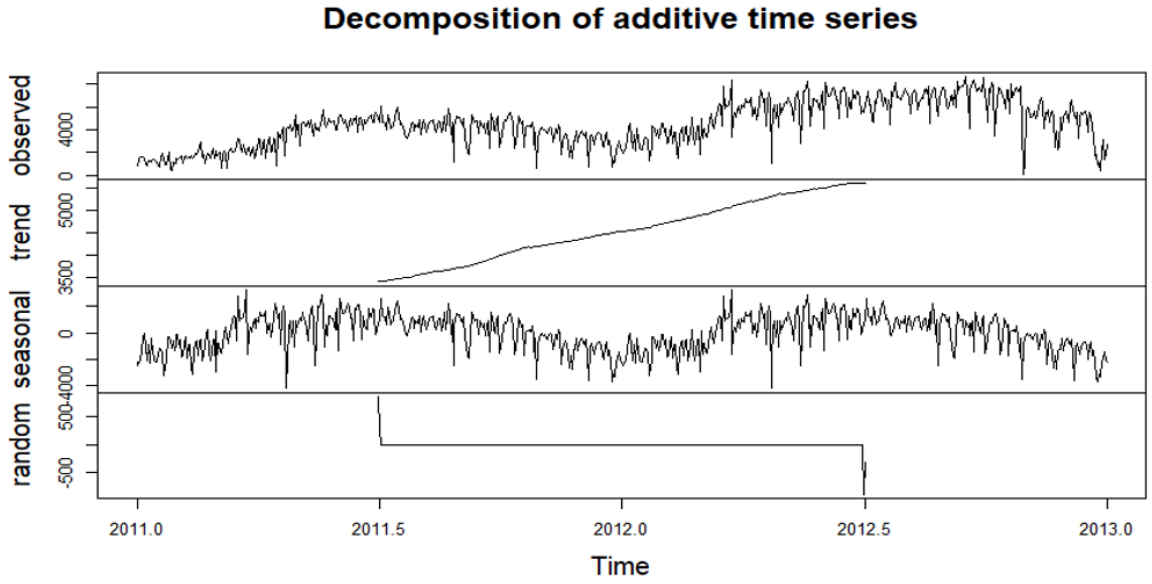
- Data Aggregation: Consolidating hourly data into daily totals.
- Time Series Creation: Constructing a time series object with annual seasonality (frequency = 365), starting from January 1, 2011.
- Visualization: Plotting the data to identify trends and patterns.
- Decomposition: We decompose the time series into its constituent components - trend, seasonality, and residuals - using both classical decomposition and STL (Seasonal and Trend decomposition using Loess) methods.
- ARIMA Modeling: We utilize the auto.arima() function to automatically select and fit an optimal ARIMA (AutoRegressive Integrated Moving Average) model to our time series data.
- Forecasting: Generating 100-day predictions using the fitted ARIMA model. A 100-day prediction is done to show the growth trend more prominently.

### 4.2.1 Key Observations



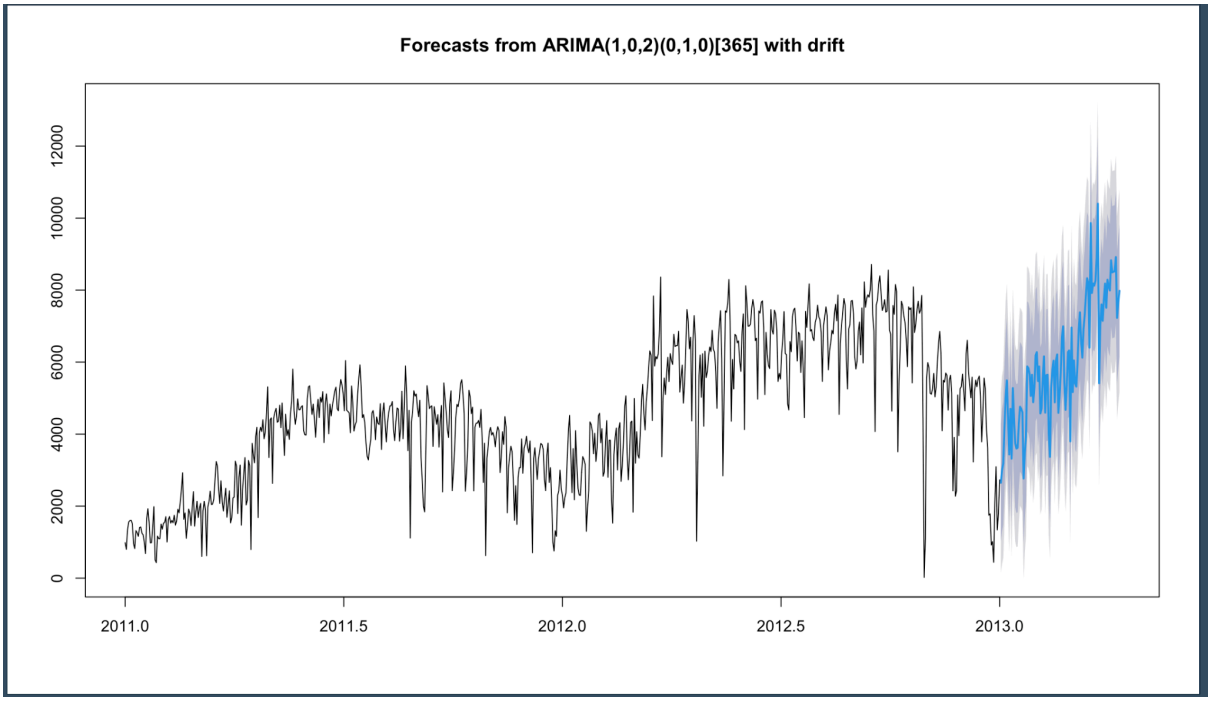

1. The upward trend highlights the growing adoption of the bike-sharing program.
2. Seasonal fluctuations suggest that bike usage heavily depends on weather and time of year. The demand at the beginning of the year 2011 is quite similar to the beginning of the year 2012 and 2013;

To get a deeper understanding, we will decompose the time series into trend, seasonal, and random components.




3. Observed Component (Top Panel): Original time series. It reflects a combination of trend, seasonality, and random variations. The upward trend, seasonal peaks, and variability observed here are broken down in the subsequent panels.

4. Trend Component (Second Panel): The upward trend highlights the increasing popularity of bike rentals, while the slight decline in late 2012 may be due to holidays or some form of unwarranted issues. Identifying these sorts of drops can give solid business insights.
5. Seasonal Component (Third Panel): The seasonal component is stable and consistent, indicating that demand is predictably higher in certain seasons.
6. Random/Residual Component (Bottom Panel): Significant random fluctuations suggest the need to examine external factors like weather, holidays, or special events

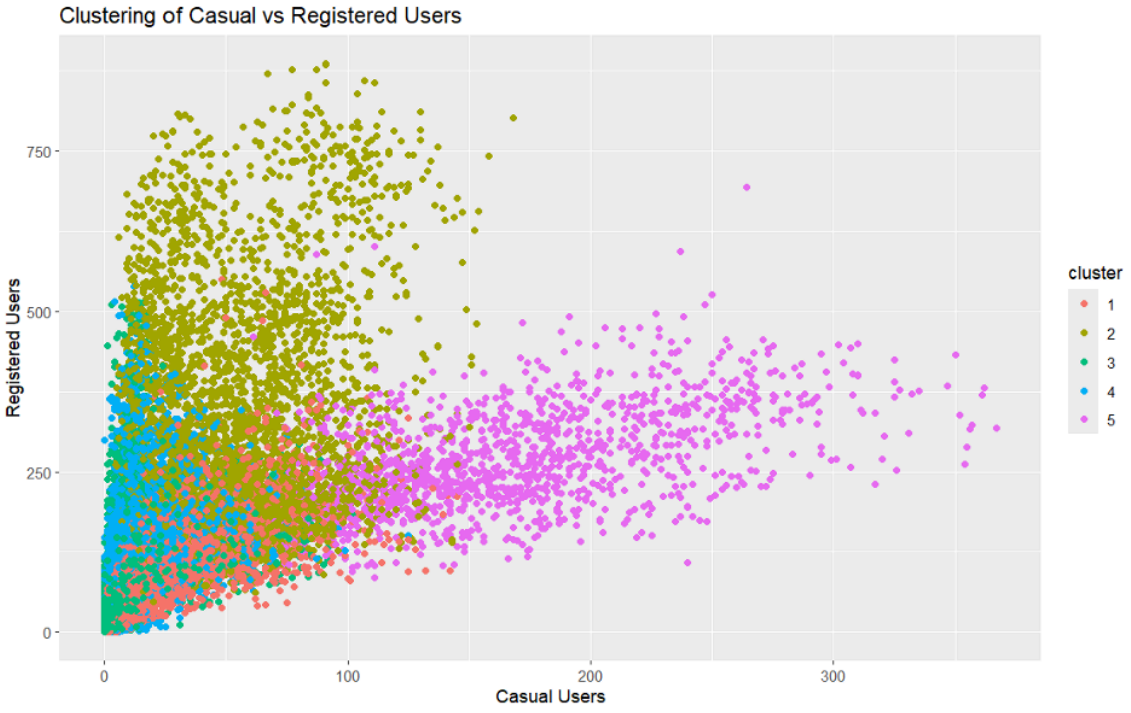


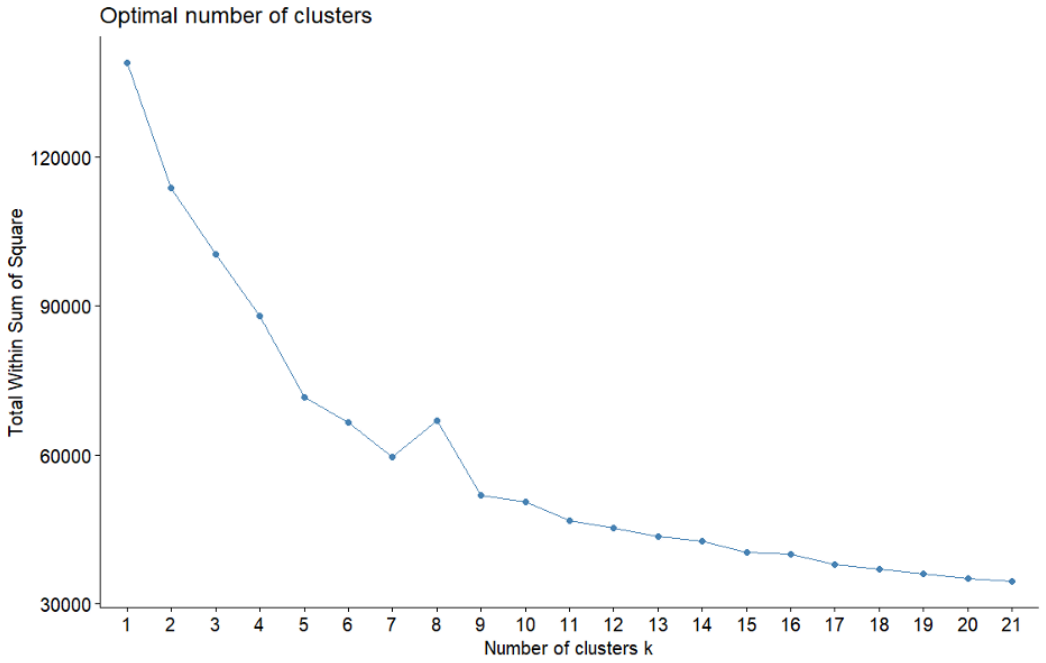

7. Observing the 100-day forecast shows us an upward trend in demand in mid-2013; just like we have observed in 2011 and 2012. However, the shaded gray area representing the confidence intervals tells us to be weary and careful about the predictions. Because, as time progressed it is noticeable that CI is wider. And this denotes uncertainty. Hence, predictions closer to the recent data trend are more reliable and also show tight confidence intervals.

### 4.3 Clustering Analysis

In the analysis of bike-sharing demand patterns, clustering techniques to segment users and identify distinct groups based on their rental behaviors are something that will provide granular insights about the two types of users that we have. Clustering techniques have been widely used in bike-sharing analysis. For instance, Guo et al. (2017) applied K-means clustering to identify spatial-temporal patterns in bike-sharing usage, demonstrating its effectiveness in uncovering distinct usage behaviors.

#### 4.3.1 K-means Clustering



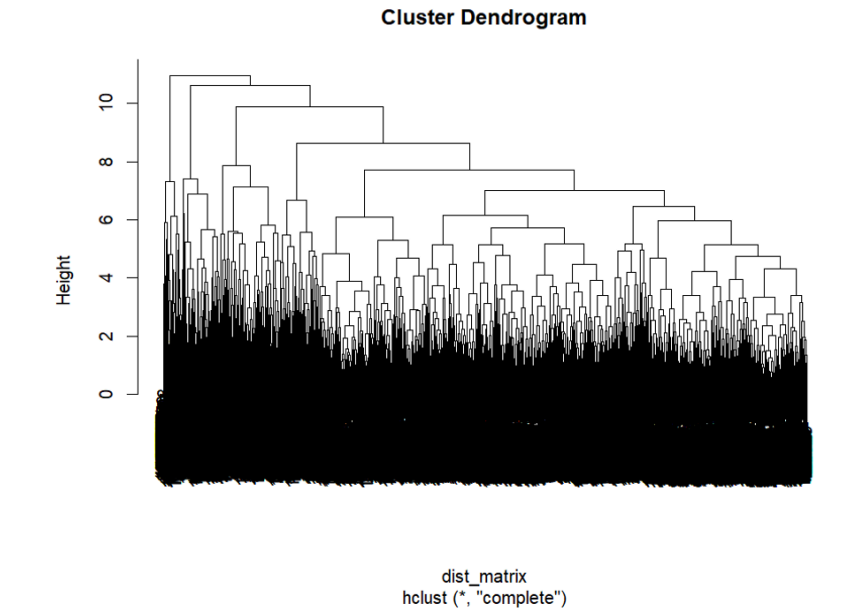
**Cluster 1 (Red)**
Low casual users, low registered users: This cluster represents periods with minimal bike usage by both casual and registered users. Likely corresponds to off-peak hours, adverse weather conditions, or low-demand seasons (e.g., winter).
**Cluster 2 (Yellow)**
Moderate casual users, moderate registered users: This cluster includes times with balanced usage from both casual and registered users. Likely occurs during weekends or favorable weather conditions when both groups are active.
**Cluster 3 (Cyan/Greenish-Blue)**
Low casual users, moderate registered users: Registered users dominate in this cluster, while casual users remain minimal. Likely corresponds to weekdays or commuting periods when registered users rely on bikes for regular travel.
**Cluster 4 (Pink)**
High casual users, moderate-to-high registered users: Casual users are more active in this cluster, alongside moderate or high registered usage. This could represent leisure activities during weekends or holidays, where casual users increase their activity significantly.
**Cluster 5 (Blue)**
Moderate casual users, high registered users: Registered users dominate in this cluster, but casual users are also moderately active. Likely occurs during peak usage times, such as weekdays with favorable conditions or busy commuter hours.

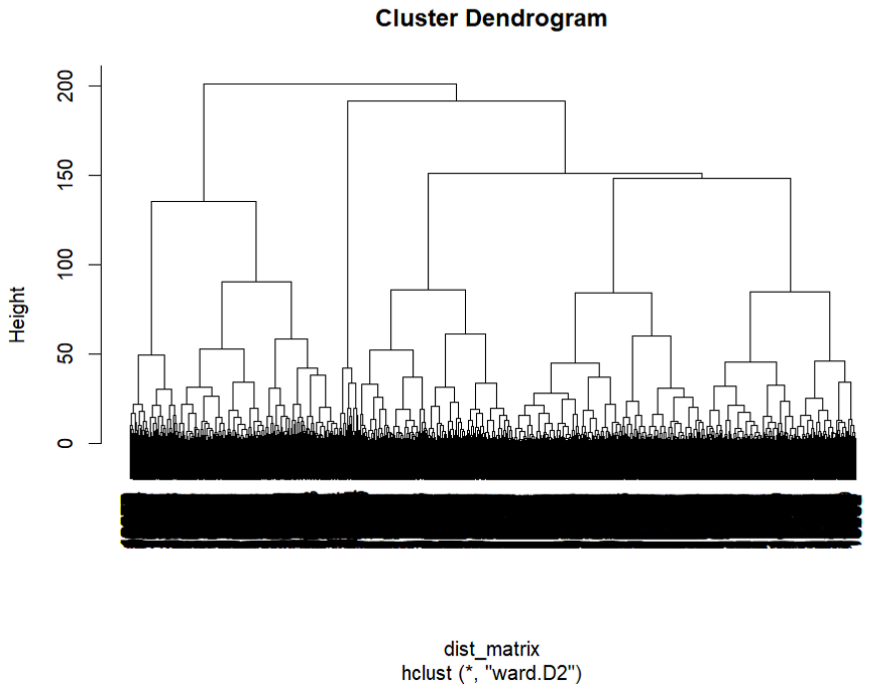
### 4.3.2 Key Observations

1. Cluster 1 represents low overall bike rentals, likely during high humidity periods. This could correspond to early mornings, off-peak hours, or days with suboptimal weather (e.g., moderate temperature but high humidity).
2. Cluster 2 represents periods with high registered user rentals and moderate casual user rentals. Likely corresponds to weekday peak commuting periods with favorable weather conditions (high temperature, moderate humidity).
3. Cluster 3 likely corresponds to cold weather days or seasons (low temperature), discouraging both casual and registered users. This likely corresponds to cold weather days or seasons (low temperature), discouraging both casual and registered users
4. Cluster 4 represents low overall rentals, possibly during humid and low windspeed conditions. This could occur during non-commuting periods or weather conditions less favorable for biking.
5. Cluster 5 represents high activity from both casual and registered users. Likely occurs during weekends, holidays, or warm, pleasant weather conditions with moderate humidity and windspeed.

### 4.3.3 Hierarchical Clustering

Hierarchical clustering: using complete and Ward.D2(Ward's method with squared distance) linkages.




1. The largest vertical gap between horizontal merges (around the middle of the dendrogram) supports the choice of 5 clusters.
2. Each of the 5 clusters is distinct, as evident from their hierarchical separation at different heights.

## 5.0 Findings

Based on the Regression Models, Time-Series Model, and Clustering methods insight we can flesh out several actionable business insights and suggestions that can be helpful for a bike rental business as such. Recalling the primary objectives to provide optics into the findings more precisely; discussing the four major areas of the objective seems like an apt approach:

1. Demand Prediction: During the Summer, Fall, and Spring seasons a high demand can be observed from the linear regressions and also the time series model. There is also a temperature interval of 15 to 25 degrees Celsius which can be labeled as optimal for renting bikes from the analysis. On top of that, it was successfully predicted and established that there are more registered users than casual users. For casual users holidays, weekends, weather conditions and for registered users working hours, weather conditions are major influencers.
   Based on these highlights, a business can predict or forecast seasonal demands and allocate more bikes for those seasons.
2. Price Optimization: Since the regression analysis identified the key drivers (temp, season) a business can develop a dynamic pricing strategy that will reduce prices during unfavorable weather to attract more users. They can also introduce surge pricing during peak hours. Based on the seasonal trend they can design special pricings that will attract both casual and registered users.
3. Fleet and Resource Optimization: Time-series trends and regression analysis ensure fleet allocation during peak demand hours and seasons. We see the evidence in ARIMA decomposition and in the regression analysis of hour-specific effects. If a business, knows the peak and lowest demand it can optimize resources allocated for maintenance.
   Along with that, they can plan fleet mobilization in the future coupled with traffic and geo-location data to get specific numbers of vehicles to be deployed in specific areas. This will significantly help reduce the cost of operations.
   With that data, a business can reduce the wastage or idle hours for a vehicle and optimize the usage of all the vehicles.
4. Promotional Campaign Design: From the time series and regression analysis we know that casual user constitutes a small portion of the total user base. Since there is a growing trend of users, the marketing team can cohort their effort towards the casual users with discounts and specially designed promotional offers during the weekend which is the highest usage time of casual users.
   Along with that, the business can also design promotional activity for the registered users by giving them special prices or discounts from 9 AM-12 PM of a working day. This increases the chance of retention and improves the overall daily active user rate.
   From the clustering analysis, it is noticeable, that cluster 1 has low casual and low registered users. Since registered users are more likely to avail of the service, targeted campaigns can be designed for them. On the flip side, cluster 5 has high casual and high registered users. Based on the same principle, registered users can be offered

loyalty programs, special bonus programs, and referral programs to increase retention of these users.

## 6.0 Conclusion

This research provides a detailed framework for forecasting bike rental demand and optimizing operational efficiency in bike-sharing businesses. By utilizing temporal, seasonal, and environmental data, the models developed offer actionable insights into user behavior and demand patterns, enabling data-driven decision-making to improve service quality and resource allocation.

The linear regression models reveal the critical factors influencing demand, such as temperature, humidity, seasonality, and working-day commuting patterns. These insights support dynamic pricing strategies, fleet optimization, and marketing campaigns tailored to specific user segments, including casual and registered users. Clustering analysis further segments users based on behavioral patterns, providing granular insights into low-demand and high-demand periods, while time-series analysis forecasts demand trends and seasonal variations, helping businesses anticipate and prepare for fluctuating usage patterns.

The findings highlight the importance of aligning bike availability and operational efforts with peak seasons, favorable weather conditions, and commuting hours, ensuring efficient resource utilization and an improved user experience. With strategic implementation of these insights, bike-sharing businesses can maximize efficiency, enhance customer satisfaction, and drive long-term growth.

In conclusion, this research equips bike-sharing operators with the tools and knowledge needed to adapt to changing demand patterns and make informed, strategic decisions, ensuring sustainable and efficient operations in a competitive market.