CASE_STUDY_FINAL

Tomasz Mulka

2025-03-01

1. Introduction

This report presents an in-depth analysis of a bike-sharing dataset using R. The objective is to explore key trends, seasonal patterns, and user behaviors to understand the factors influencing bike rentals. The analysis includes data preprocessing, visualization, and statistical insights to derive meaningful conclusions.

2. Data Preparation

The dataset consists of records of bike rentals, including attributes such as date, time, and user type. Understanding these variables is essential for analyzing rental trends and optimizing service operations. The data was examined for completeness and consistency before conducting further analysis.

Dataset upload

```
X2019_RIDES1 <- read_delim("2019_RIDES1.csv", delim = ";", col_types = cols())
```

3. Data Preprocessing

Prior to analysis, the dataset underwent preprocessing, including handling missing values, filtering inconsistencies, and transforming categorical data where necessary. Data was also aggregated where needed to facilitate visualization and interpretation.

Changing incorrect format for columns "start_time" & "end_time"

```
X2019_RIDES1$start_time <- as.POSIXct(X2019_RIDES1$start_time, format="%m/%d/%y %H:%M", tz="UTC")
X2019_RIDES1$end_time <- as.POSIXct(X2019_RIDES1$end_time, format="%m/%d/%y %H:%M", tz="UTC")
```

Changing incorrect format of ride_lenght column

```
X2019_RIDES1$ride_lenght <- as.character(X2019_RIDES1$ride_lenght)
```

Convert ride_lenght into Seconds

```
time_to_seconds <- function(time_str) {
  parts <- unlist(strsplit(time_str, ":"))
  if (length(parts) == 3) {
    return(as.numeric(parts[1]) * 3600 + as.numeric(parts[2]) * 60 + as.numeric(parts[3]))
  } else {
    return(NA)  # Handle incorrect formats
  }
}

# Apply conversion function
X2019_RIDES1$ride_lenght_seconds <- sapply(X2019_RIDES1$ride_lenght, time_to_seconds)

# Convert to minutes for easier readability
X2019_RIDES1$ride_lenght_minutes <- X2019_RIDES1$ride_lenght_seconds / 60
```

Remove Outliers

```
threshold <- quantile(X2019_RIDES1$ride_lenght_seconds, 0.99, na.rm = TRUE)
X2019_RIDES1 <- X2019_RIDES1 %>% filter(ride_lenght_seconds <= threshold)
```

Extract start hour from start_time column

```
X2019_RIDES1$start_hour <- format(X2019_RIDES1$start_time, "%H")
```

Extract hour from start time for customers

```
rental_hours_customers <- X2019_RIDES1 %>%
  filter(usertype == "Customer") %>%
  group_by(start_hour) %>%
  summarise(count = n(), .groups = "drop")
```

Extract hour from start time for subscribers

```
rental_hours_subscriber <- X2019_RIDES1 %>%
  filter(usertype == "Subscriber") %>%
  group_by(start_hour) %>%
  summarise(count = n(), .groups = "drop")
```

3. Data analysis

Exploratory Data Analysis (EDA) was conducted to identify trends in bike rentals over time. Various visualizations were created to analyze daily and monthly usage patterns, peak hours, and differences between registered users and casual users. These insights help in understanding user behavior and demand fluctuations.

Average Ride Duration

```r
ride_duration_avg <- X2019_RIDES1 %>%
  group_by(usertype) %>%
  summarise(avg_duration_min = mean(ride_lenght_minutes, na.rm = TRUE))

print(ride_duration_avg)
```

```
## # A tibble: 2 x 2
##   usertype   avg_duration_min
##   <chr>                 <dbl>
## 1 Customer               17.8
## 2 Subscriber             10.4
```

Gender Distribution

```r
gender_count <- X2019_RIDES1 %>%
  group_by(usertype, gender) %>%
  summarise(count = n(), .groups = "drop") %>%
  pivot_wider(names_from = gender, values_from = count, values_fill = list(count = 0))

print(gender_count)
```

```
## # A tibble: 2 x 3
##   usertype   Female   Male
##   <chr>       <int>  <int>
## 1 Customer     1445   3387
## 2 Subscriber  64580 272463
```

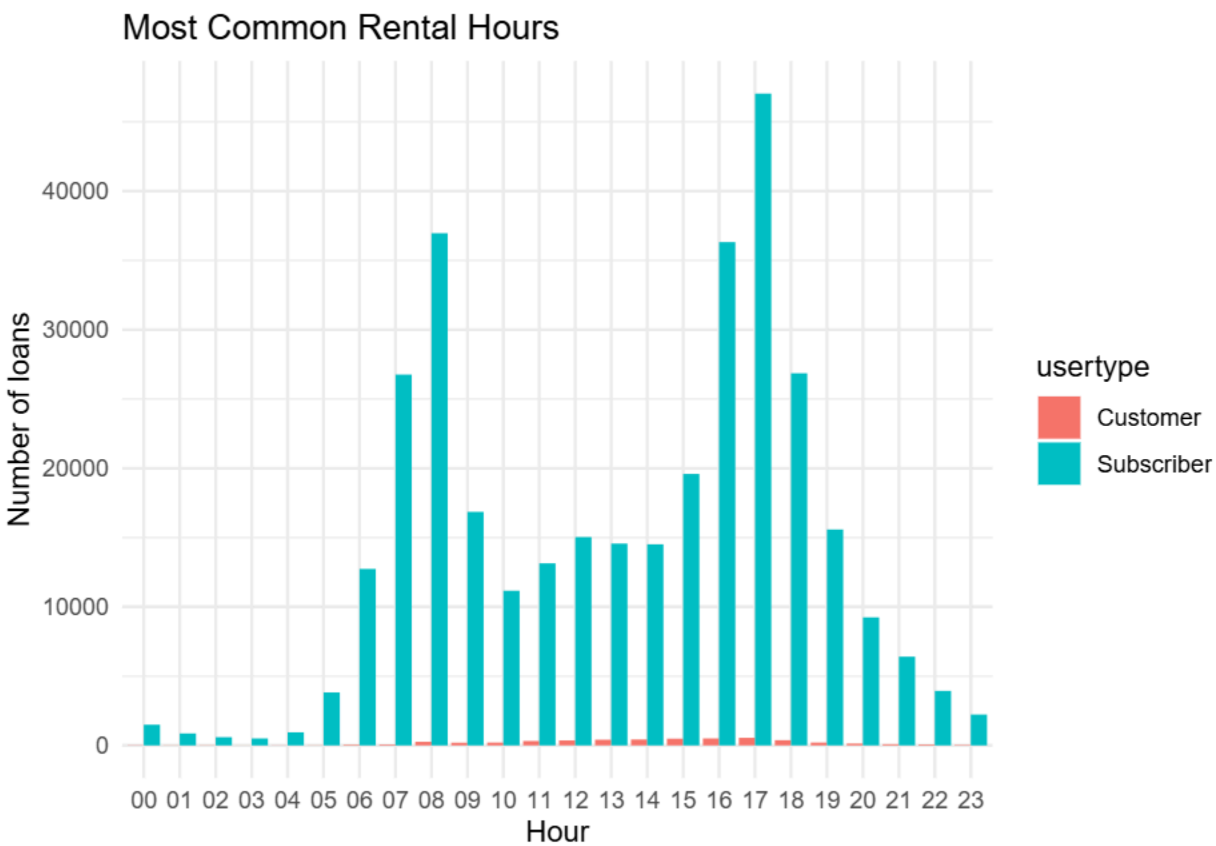
Most Common Rental Hours

Hourly analysis of bike rentals reveals peak usage times, particularly during commuting hours. Identifying these periods is crucial for managing bike availability and optimizing service operations to meet demand.

```
rental_hours <- X2019_RIDES1 %>%
  group_by(usertype, start_hour) %>%
  summarise(count = n())
```

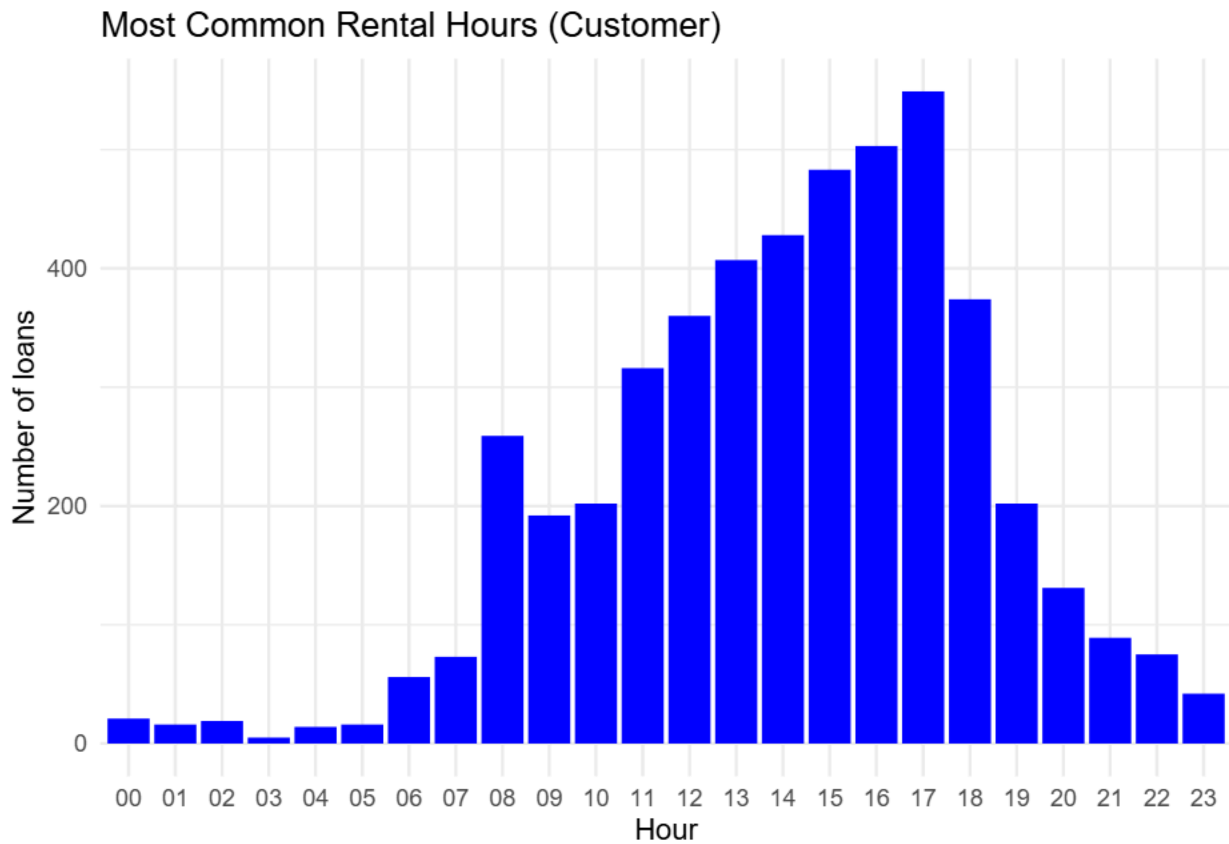
```
## `summarise()` has grouped output by 'usertype'. You can override using the
## `.groups` argument.
```

```
ggplot(rental_hours, aes(x = start_hour, y = count, fill = usertype)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Most Common Rental Hours", x = "Hour", y = "Number of loans") +
  theme_minimal()
```



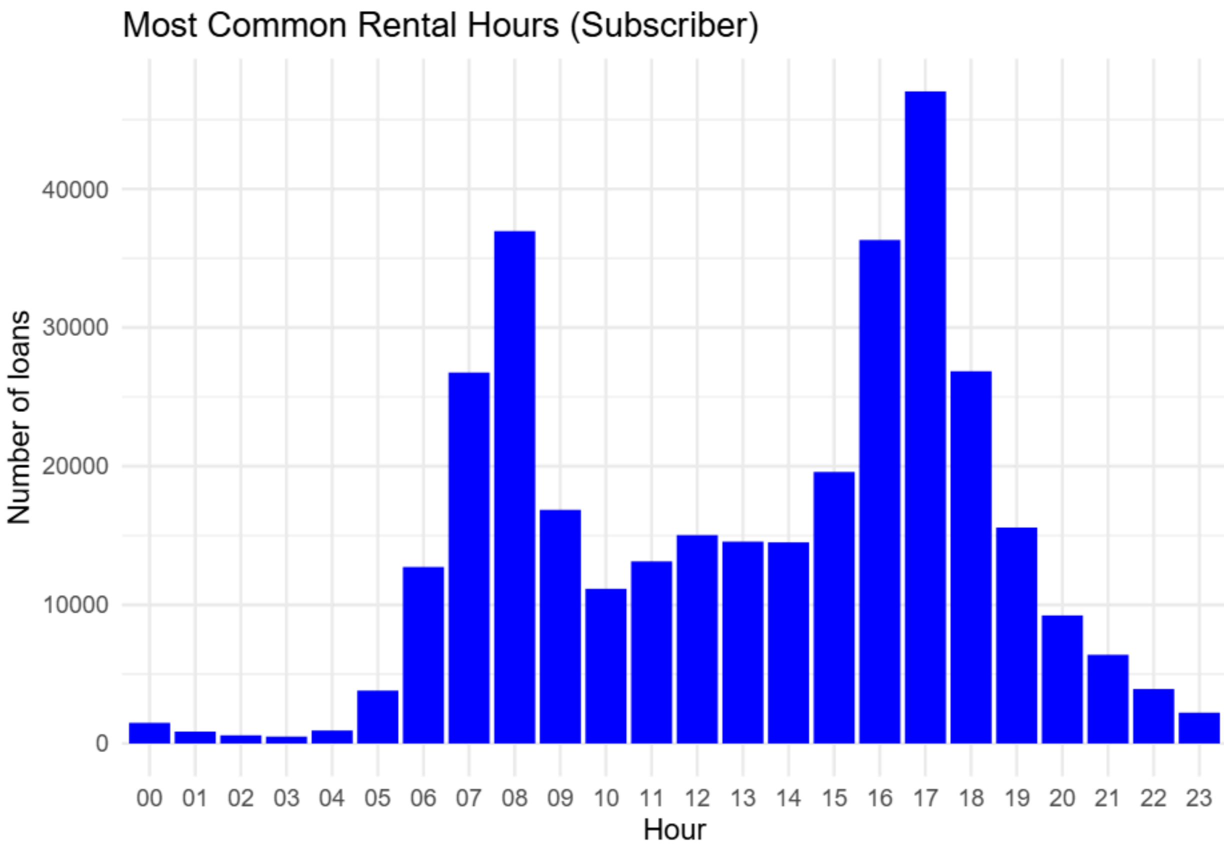
Peak Rental Hours for Customers

```r
ggplot(rental_hours_customers, aes(x = start_hour, y = count)) +
  geom_bar(stat = "identity", fill = "blue") +  # Możesz zmienić kolor
  labs(title = "Most Common Rental Hours (Customer)",
       x = "Hour",
       y = "Number of loans") +
  theme_minimal()
```



Peak Rental Hours for Subscribers

```r
ggplot(rental_hours_subscriber, aes(x = start_hour, y = count)) +
  geom_bar(stat = "identity", fill = "blue") +  # Możesz zmienić kolor
  labs(title = "Most Common Rental Hours (Subscriber)",
       x = "Hour",
       y = "Number of loans") +
  theme_minimal()
```

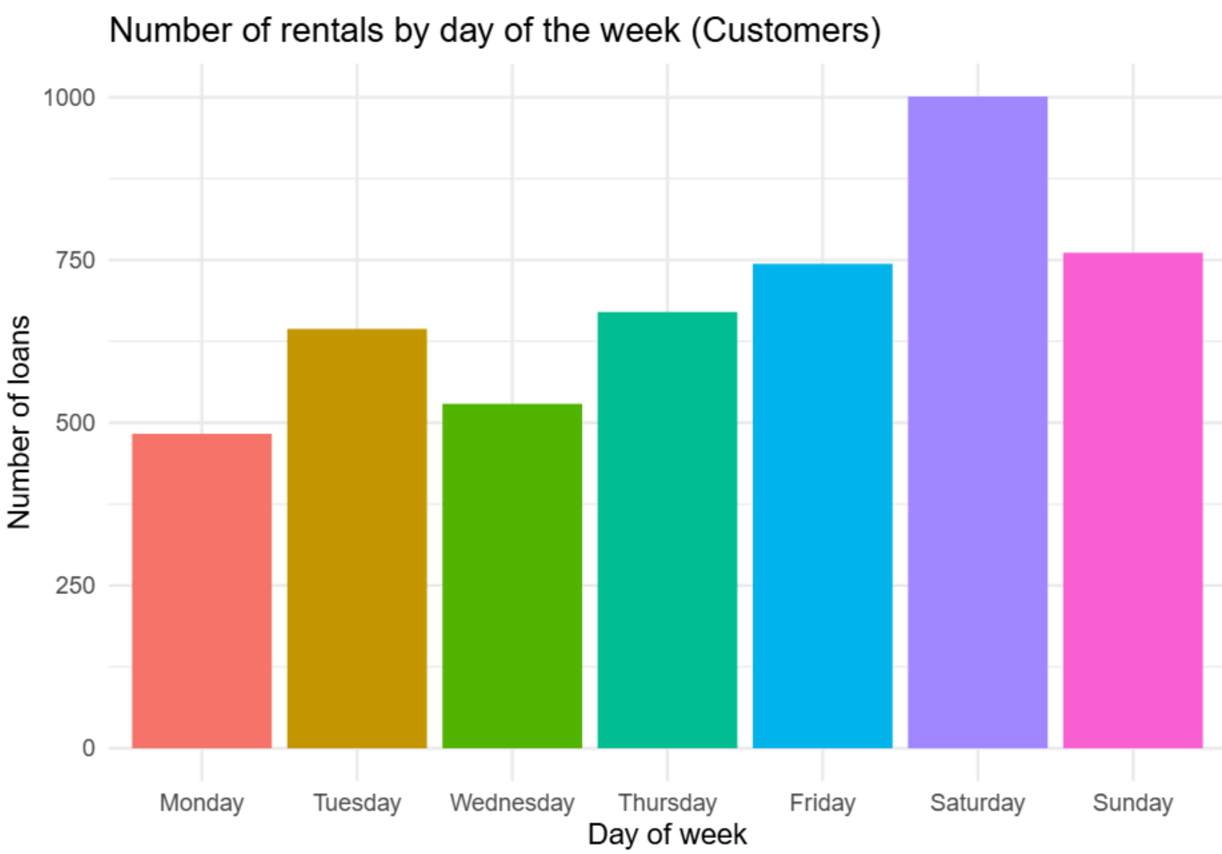


Dependecy for day of week for Customer

```
rental_weekday_customers <- X2019_RIDES1 %>%
  filter(usertype == "Customer") %>%
  group_by(day_of_week) %>%
  summarise(count = n())

# Swap the order of the days of the week (optional if the days are in the wrong order)
rental_weekday_customers$day_of_week <- factor(rental_weekday_customers$day_of_week,
                                               levels = c("Monday", "Tuesday", "Wednesday", "Thursday",

# Creating a chart for Customers
ggplot(rental_weekday_customers, aes(x = day_of_week, y = count, fill = day_of_week)) +
  geom_bar(stat = "identity") +
  labs(title = "Number of rentals by day of the week (Customers)",
       x = "Day of week",
       y = "Number of loans") +
  theme_minimal() +
  theme(legend.position = "none")  # Hides the legend because the colors correspond to the days
```

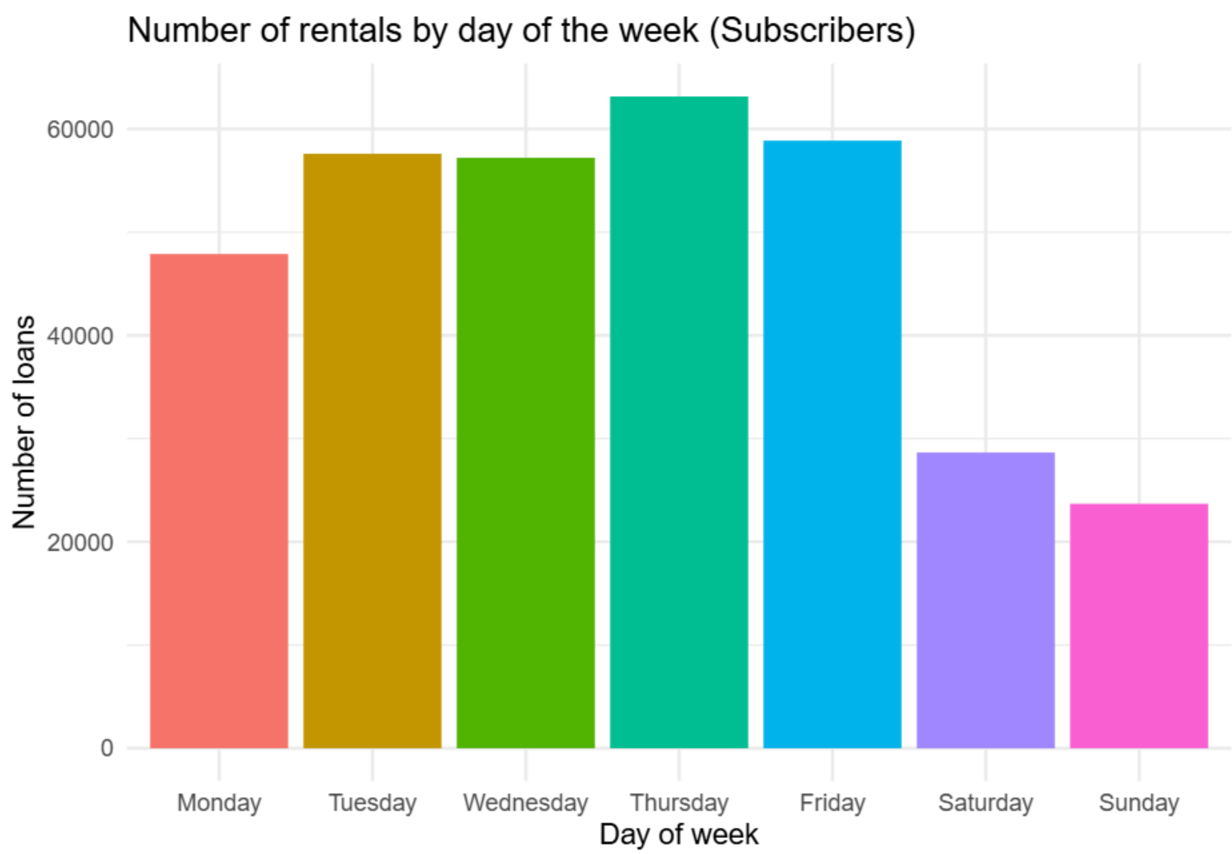


Dependecy for day of week for Subscriber

```
rental_weekday_subscribers <- X2019_RIDES1 %>%
  filter(usertype == "Subscriber") %>%
  group_by(day_of_week) %>%
  summarise(count = n())

# Swap the order of the days of the week (optional if the days are in the wrong order)
rental_weekday_subscribers$day_of_week <- factor(rental_weekday_customers$day_of_week,
                                                levels = c("Monday", "Tuesday", "Wednesday", "Thursday",

# Creating a chart for Subscriber
ggplot(rental_weekday_subscribers, aes(x = day_of_week, y = count, fill = day_of_week)) +
  geom_bar(stat = "identity") +
  labs(title = "Number of rentals by day of the week (Subscribers)",
       x = "Day of week",
       y = "Number of loans") +
  theme_minimal() +
  theme(legend.position = "none")  # Hides the legend because the colors correspond to the days
```



Conclusion

This analysis has highlighted key patterns in bike rentals, including seasonal trends, peak hours, and differences in user behavior. The findings can support strategic decisions to enhance service efficiency and user experience. Future research could involve predictive modeling to anticipate demand and optimize bike distribution

Trip duration

-Customers take significantly longer rides on average than Subscribers, suggesting a recreational usage pattern.
-Subscribers have shorter trip durations, which indicates commuting behavior.

To attract Customers into becoming Subscribers, Company should emphasize the benefits of a subscription for frequent and short-distance trips.

Usage Patterns by Day of the Week

-Subscribers primarily use bikes during weekdays, especially during morning and evening peak hours, reinforcing the idea that they use bikes for work commutes. -Customers ride more frequently on weekends, indicating they use bikes for leisure and social activities.

Promote subscription plans highlighting cost savings for weekday commuting. Introduce special weekend discounts or family plans to attract Customers who ride recreationally.

Usage Patterns by Hour

-Subscribers usage spikes during rush hours (7-9 AM and 5-7 PM), confirming their reliance on bikes for daily commutes. -Customers ride mostly in the afternoon (12 PM - 6 PM), further supporting the idea that they ride for leisure or errands.

Offer morning perks or priority docking for Subscribers to improve the commuter experience.Introduce promotional campaigns in the afternoon to attract Customers with incentives like "subscribe and get unlimited afternoon rides