EDA on Ride Sharing Dataset

No.	Name	Student ID	Program
1	MUHAMMAD BASYERUL HAKIMI BIN AFFANDI	22003506	CS
2	MUHAMMAD QUDRI ELZAM BIN SAMSUL BAHRI	22007523	CS
3	MUHAMMAD FAKHRUL HARIZ BIN SHAZLI	22005265	CS
4	MUHAMMAD SYAMIL AIMAN BIN ANUAR	22005733	CS

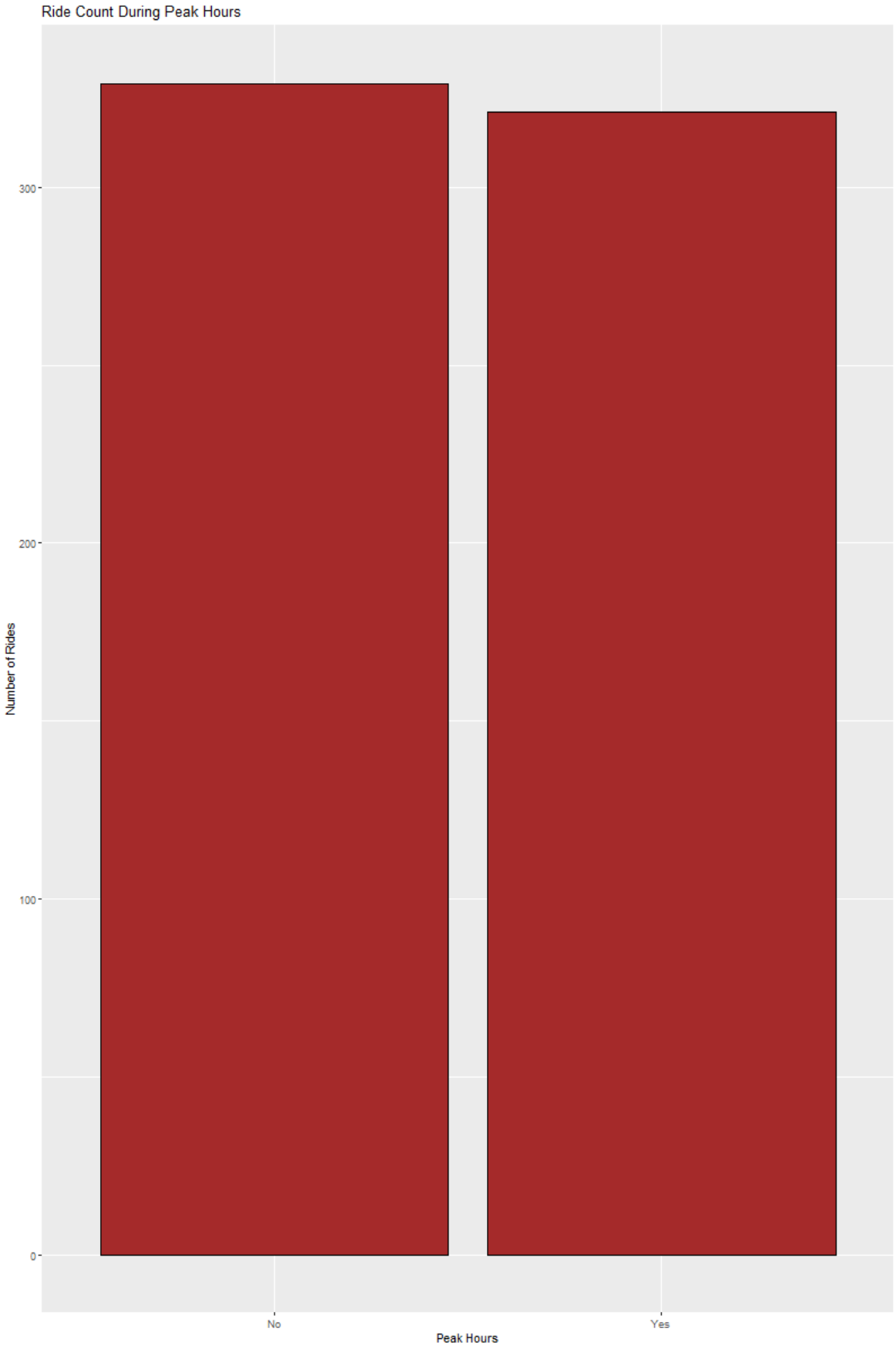
R code:

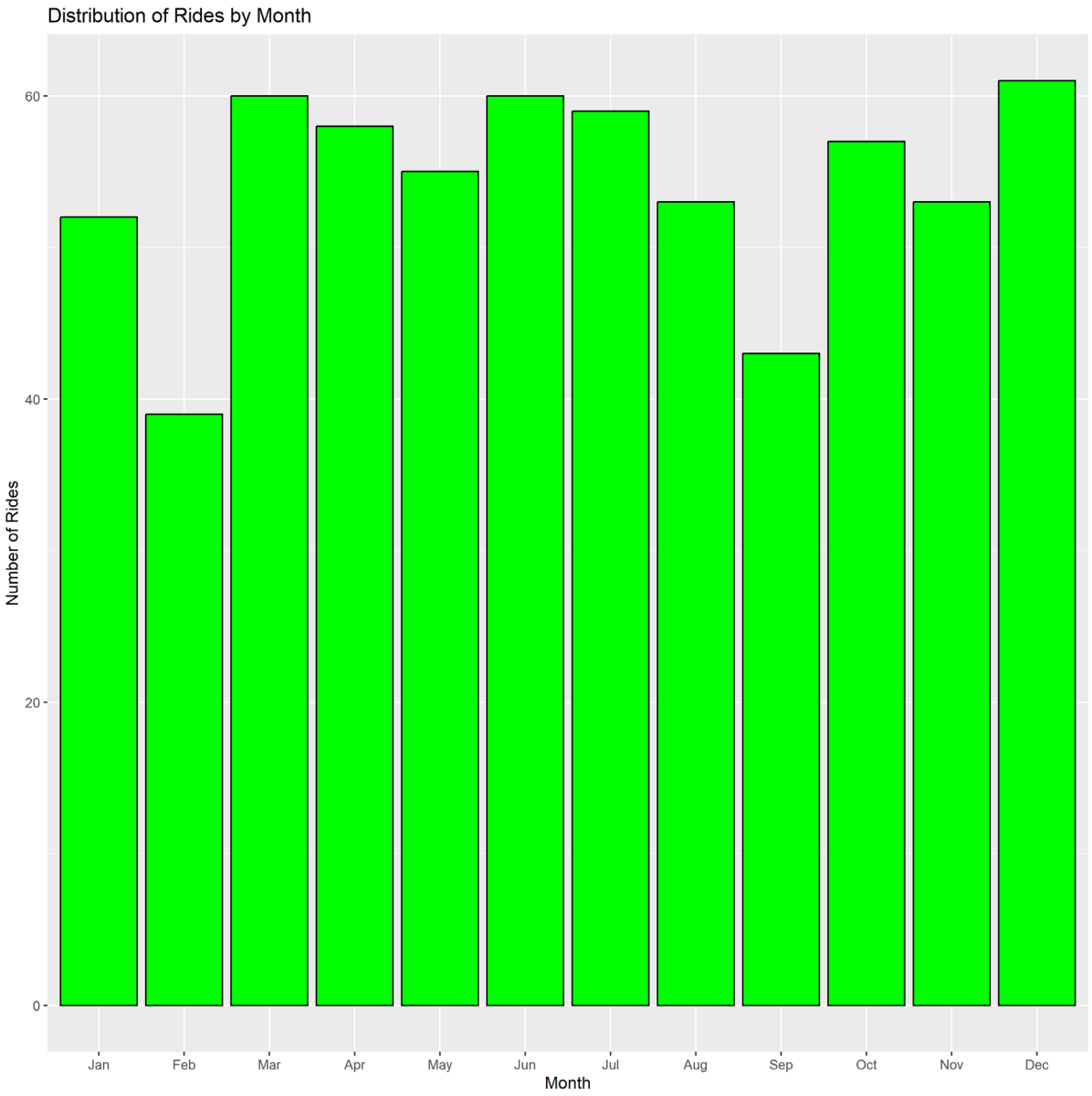
```
# Load necessary libraries

if (!require(readxl)) install.packages('readxl', dependencies=TRUE)

if (!require(dplyr)) install.packages('dplyr', dependencies=TRUE)

if (!require(ggplot2)) install.packages('ggplot2', dependencies=TRUE)

if (!require(lubridate)) install.packages('lubridate', dependencies=TRUE)


library(readxl)

library(dplyr)

library(ggplot2)

library(lubridate)


# Load the dataset

ride_data <- read_excel("C:/Users/Msi/Desktop/R folder/Group Work/EDA Ride Sharing/Ride Sharing Dataset.xlsx")


# Display the first few rows of the dataset

print(head(ride_data))


# Summary of the dataset
```

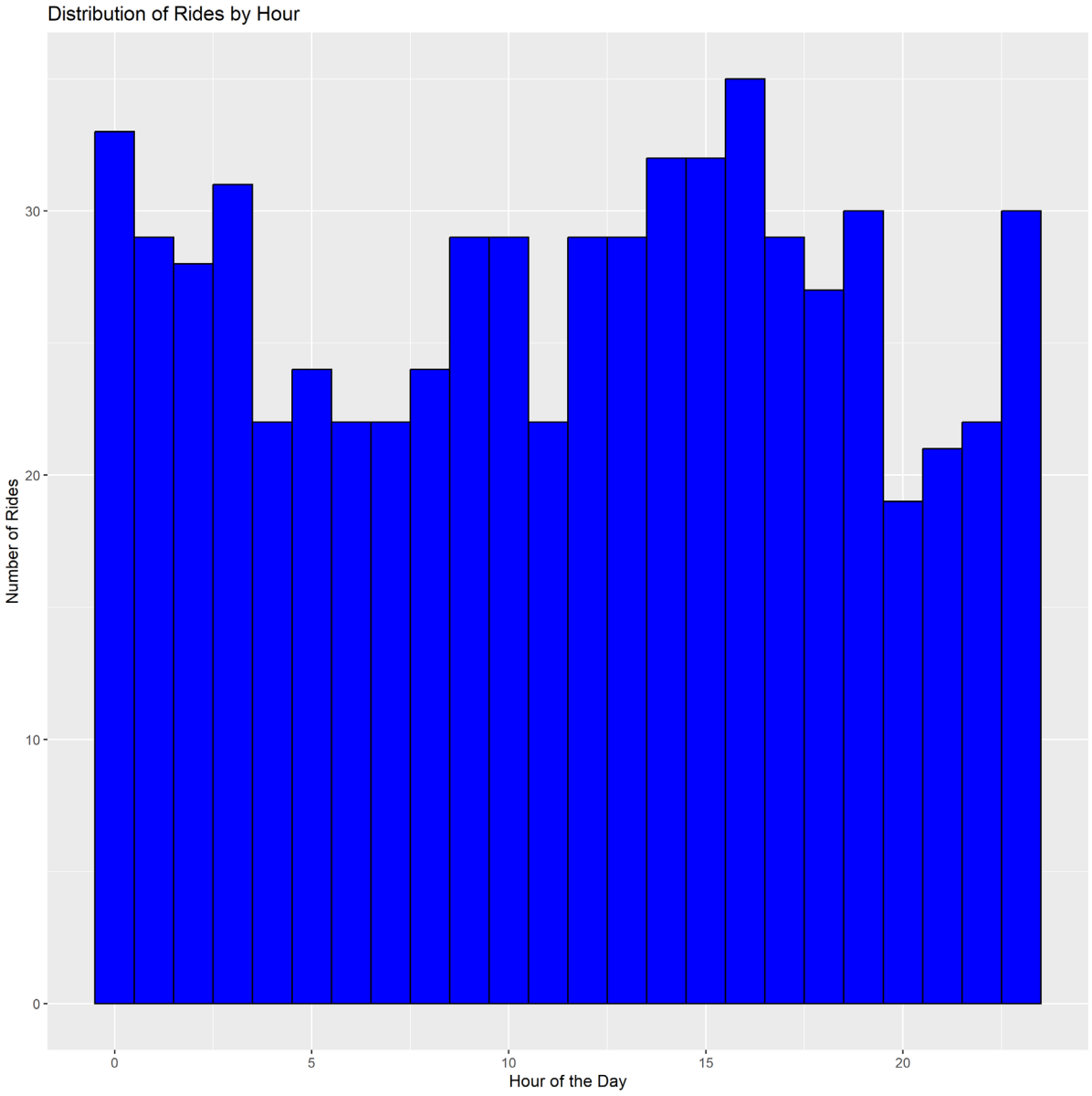
```
print(summary(ride_data))

# Check for missing values
print(sum(is.na(ride_data)))

# Convert 'Request Time' to a proper datetime format
ride_data$`Request Time` <- as.POSIXct(ride_data$`Request Time`, origin="1970-01-01")

# Extract additional time-based features
ride_data$Hour <- hour(ride_data$`Request Time`)
ride_data$Day <- wday(ride_data$`Request Time`, label=TRUE)
ride_data$Month <- month(ride_data$`Request Time`, label=TRUE)

# Distribution of rides by hour
p1 <- ggplot(ride_data, aes(x=Hour)) +
  geom_histogram(binwidth=1, fill="blue", color="black") +
  labs(title="Distribution of Rides by Hour", x="Hour of the Day", y="Number of Rides")
print(p1)

# Distribution of rides by day of the week
p2 <- ggplot(ride_data, aes(x=Day)) +
  geom_bar(fill="orange", color="black") +
  labs(title="Distribution of Rides by Day of the Week", x="Day of the Week", y="Number of Rides")
print(p2)

# Distribution of rides by month
p3 <- ggplot(ride_data, aes(x=Month)) +
```

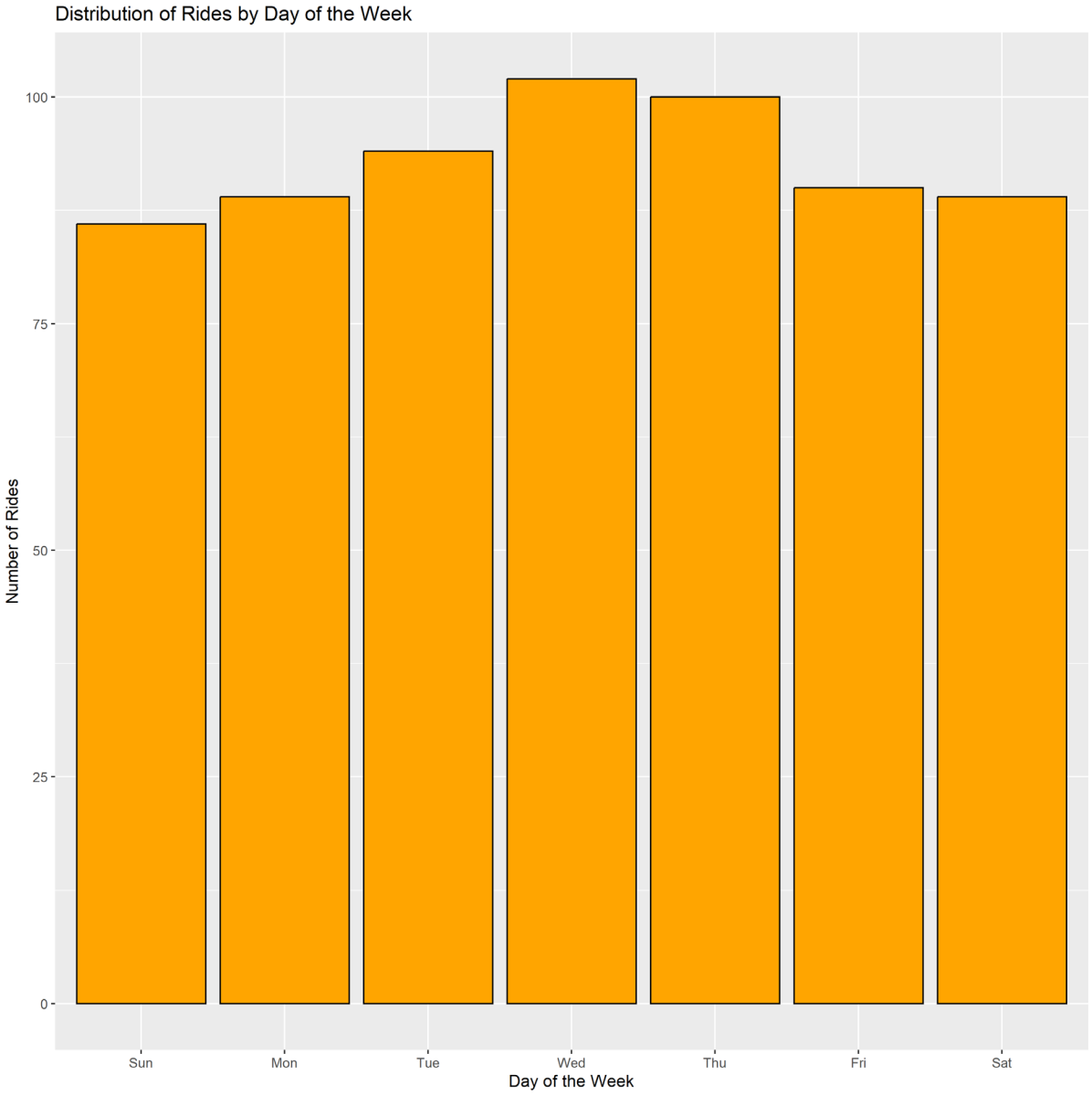
```
  geom_bar(fill="green", color="black") +
  labs(title="Distribution of Rides by Month", x="Month", y="Number of Rides")
print(p3)

# Average fare amount by vehicle type
p4 <- ride_data %>%
  group_by(`Vehicle Type`) %>%
  summarise(Average_Fare = mean(`Fare Amount (in $)`)) %>%
  ggplot(aes(x=`Vehicle Type`, y=Average_Fare)) +
  geom_bar(stat="identity", fill="purple", color="black") +
  labs(title="Average Fare Amount by Vehicle Type", x="Vehicle Type", y="Average Fare Amount ($)")
print(p4)

# Relationship between ride distance and fare amount
p5 <- ggplot(ride_data, aes(x=`Ride Distance (in miles)`, y=`Fare Amount (in $)`)) +
  geom_point(color="red") +
  geom_smooth(method="lm", color="blue") +
  labs(title="Relationship Between Ride Distance and Fare Amount", x="Ride Distance (miles)", y="Fare Amount ($)")
print(p5)

# User ratings distribution
p6 <- ggplot(ride_data, aes(x=`User Rating`)) +
  geom_histogram(binwidth=0.5, fill="yellow", color="black") +
  labs(title="Distribution of User Ratings", x="User Rating", y="Count")
print(p6)

# Payment method preferences
```

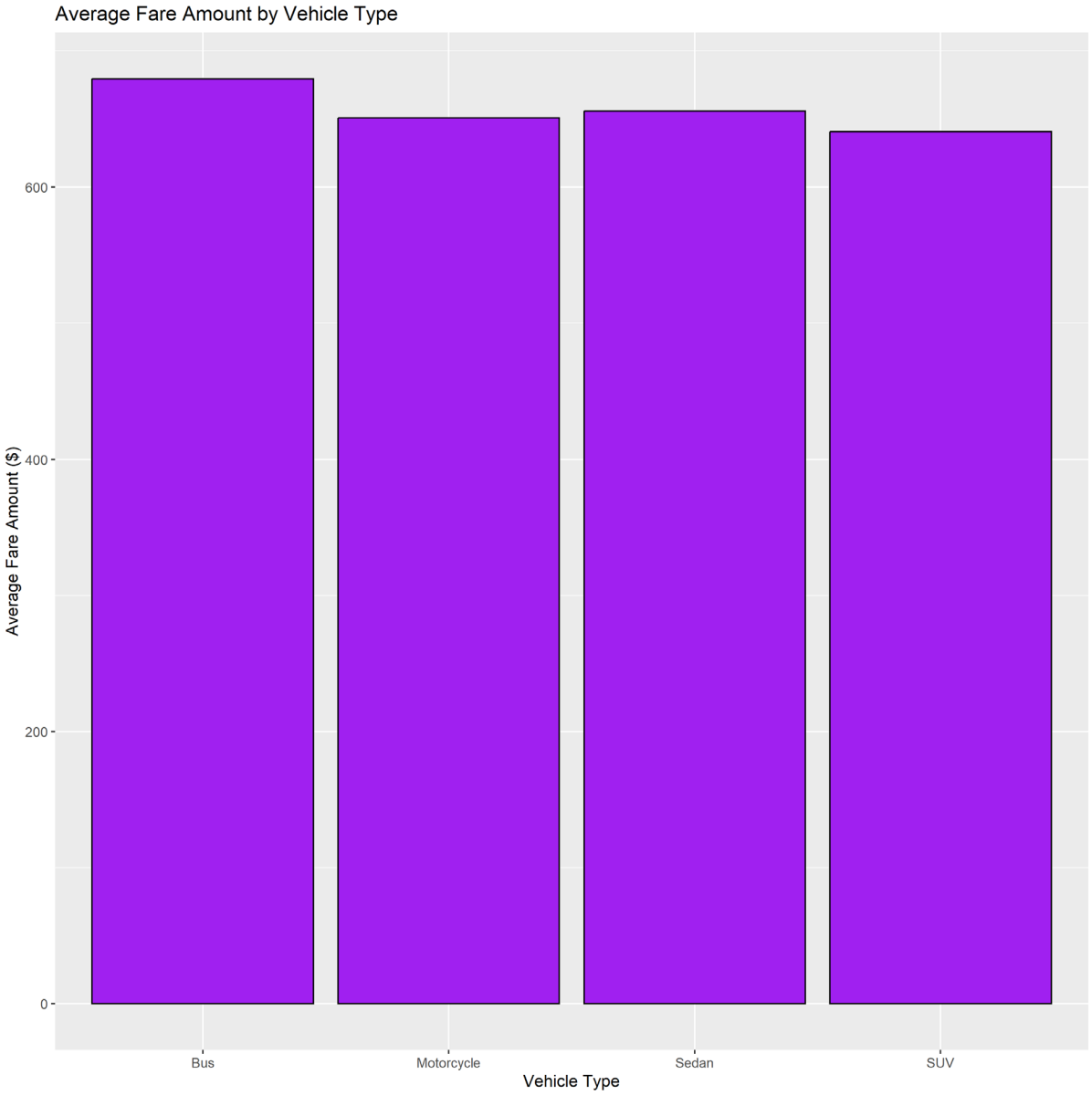
```
p7 <- ride_data %>%
  group_by(`Payment Method`) %>%
  summarise(Count = n()) %>%
  ggplot(aes(x=`Payment Method`, y=Count)) +
  geom_bar(stat="identity", fill="cyan", color="black") +
  labs(title="Payment Method Preferences", x="Payment Method", y="Count")
print(p7)

# Traffic conditions impact on ride distance
p8 <- ride_data %>%
  group_by(`Traffic Condition`) %>%
  summarise(Average_Distance = mean(`Ride Distance (in miles)`)) %>%
  ggplot(aes(x=`Traffic Condition`, y=Average_Distance)) +
  geom_bar(stat="identity", fill="pink", color="black") +
  labs(title="Impact of Traffic Conditions on Ride Distance", x="Traffic Condition",
y="Average Ride Distance (miles)")
print(p8)

# Peak hours analysis
p9 <- ride_data %>%
  group_by(`Peak Hours`) %>%
  summarise(Count = n()) %>%
  ggplot(aes(x=`Peak Hours`, y=Count)) +
  geom_bar(stat="identity", fill="brown", color="black") +
  labs(title="Ride Count During Peak Hours", x="Peak Hours", y="Number of Rides")
print(p9)
```

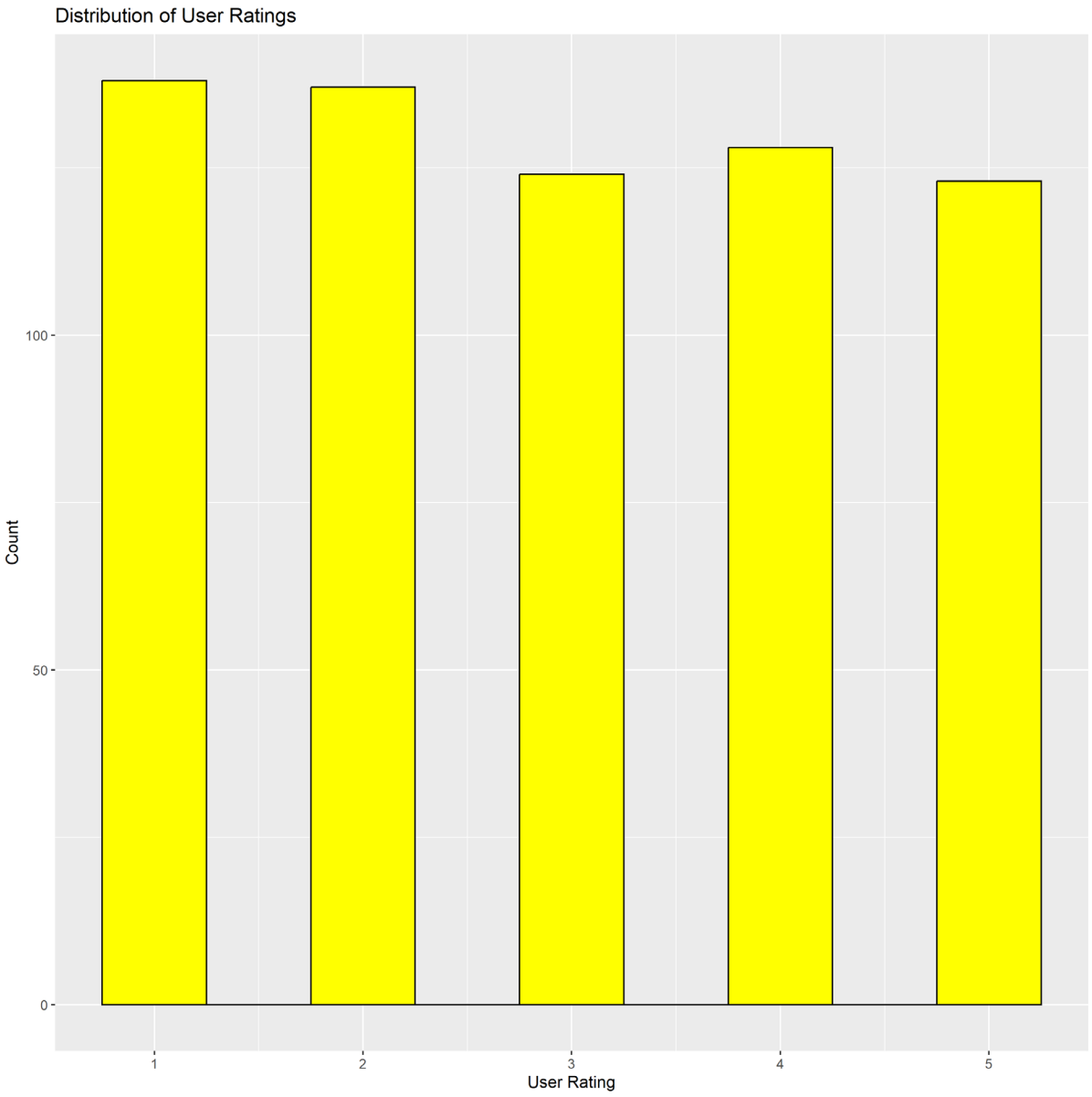

Ride Count During Peak Hours
300
200
100
0
Number of Rides
No
Yes
Peak Hours

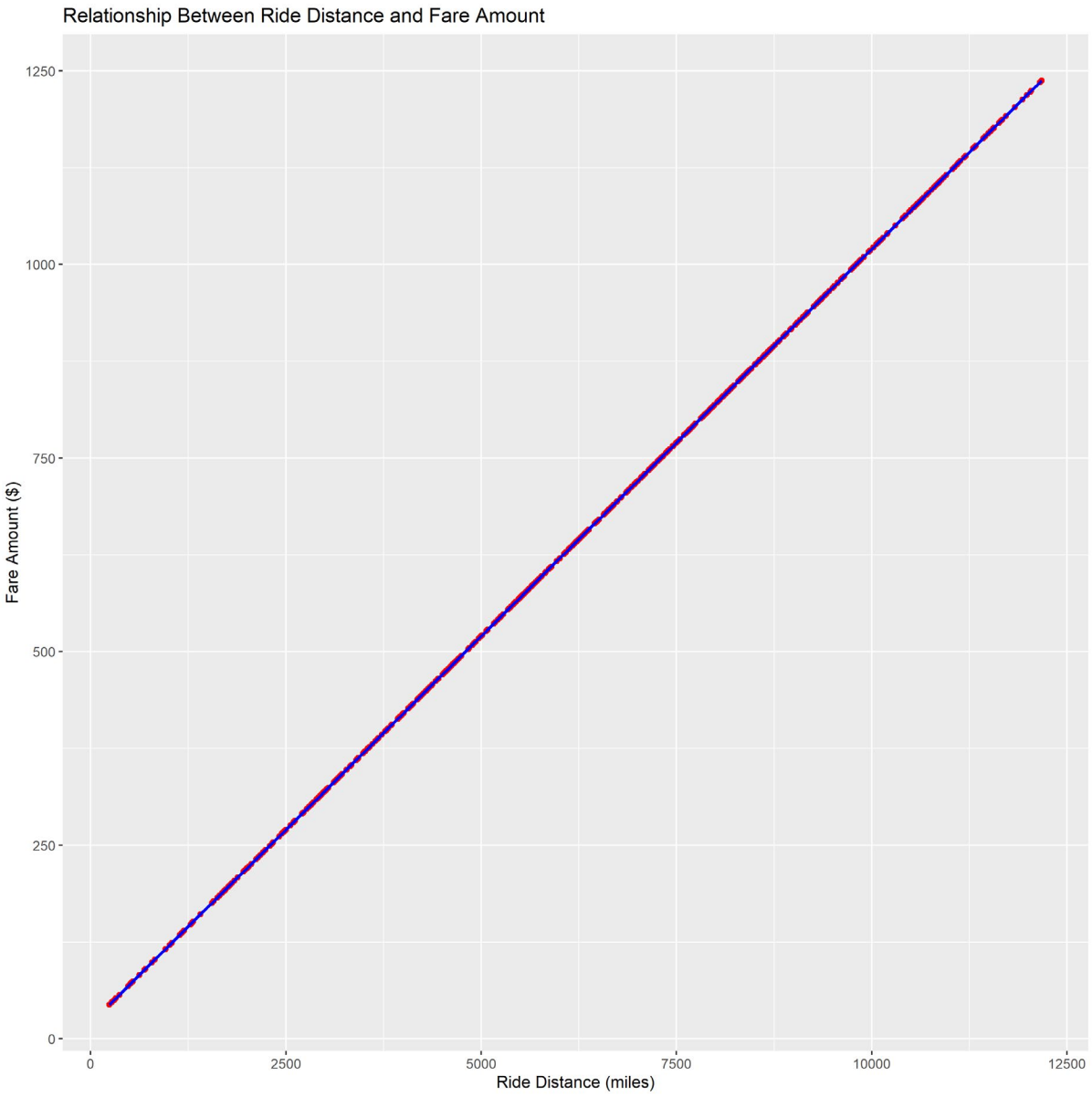

Distribution of Rides by Month
Number of Rides
60
40
20
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month

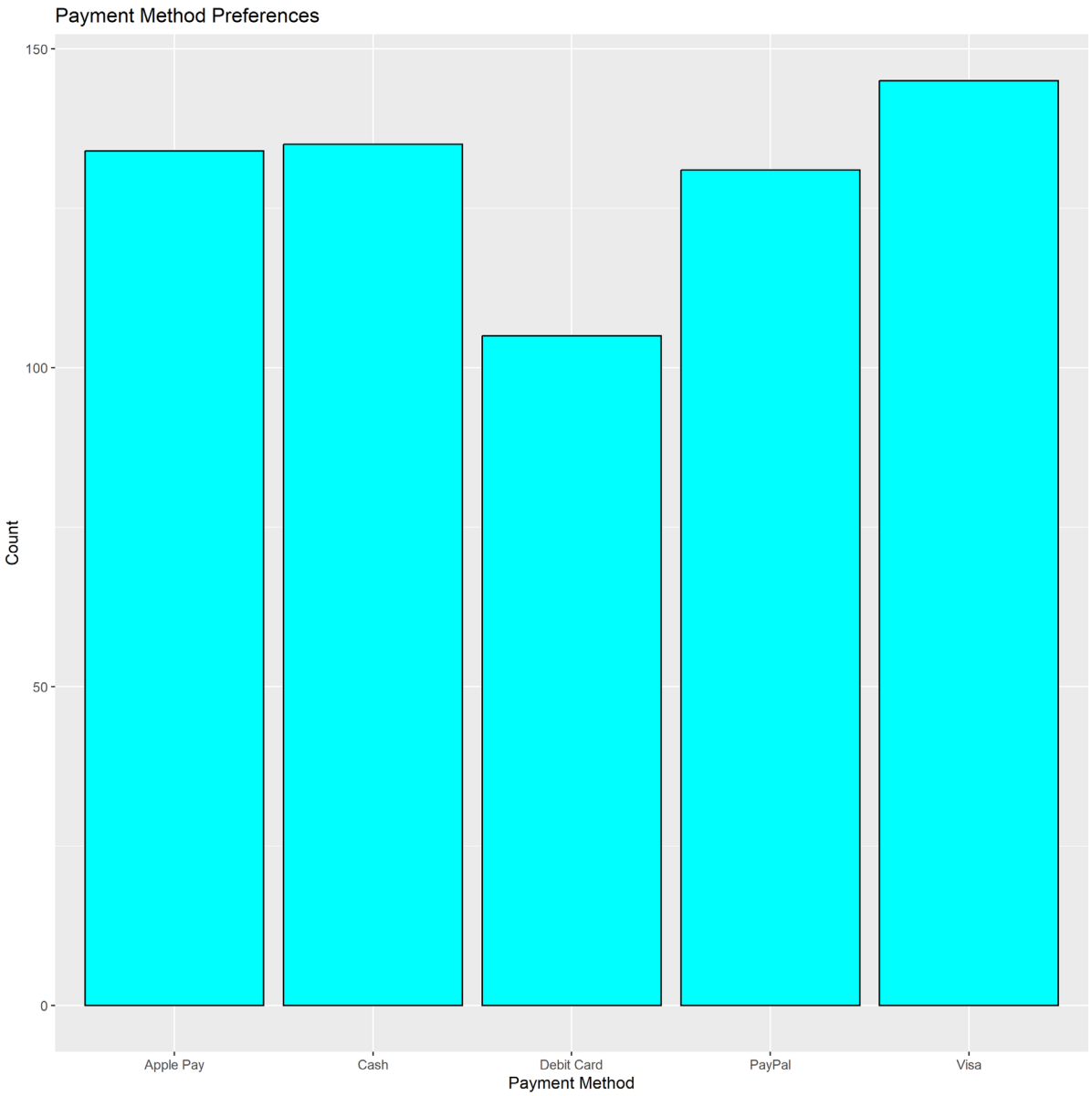

Distribution of Rides by Hour
Number of Rides
0
10
20
30
0
5
10
15
20
Hour of the Day

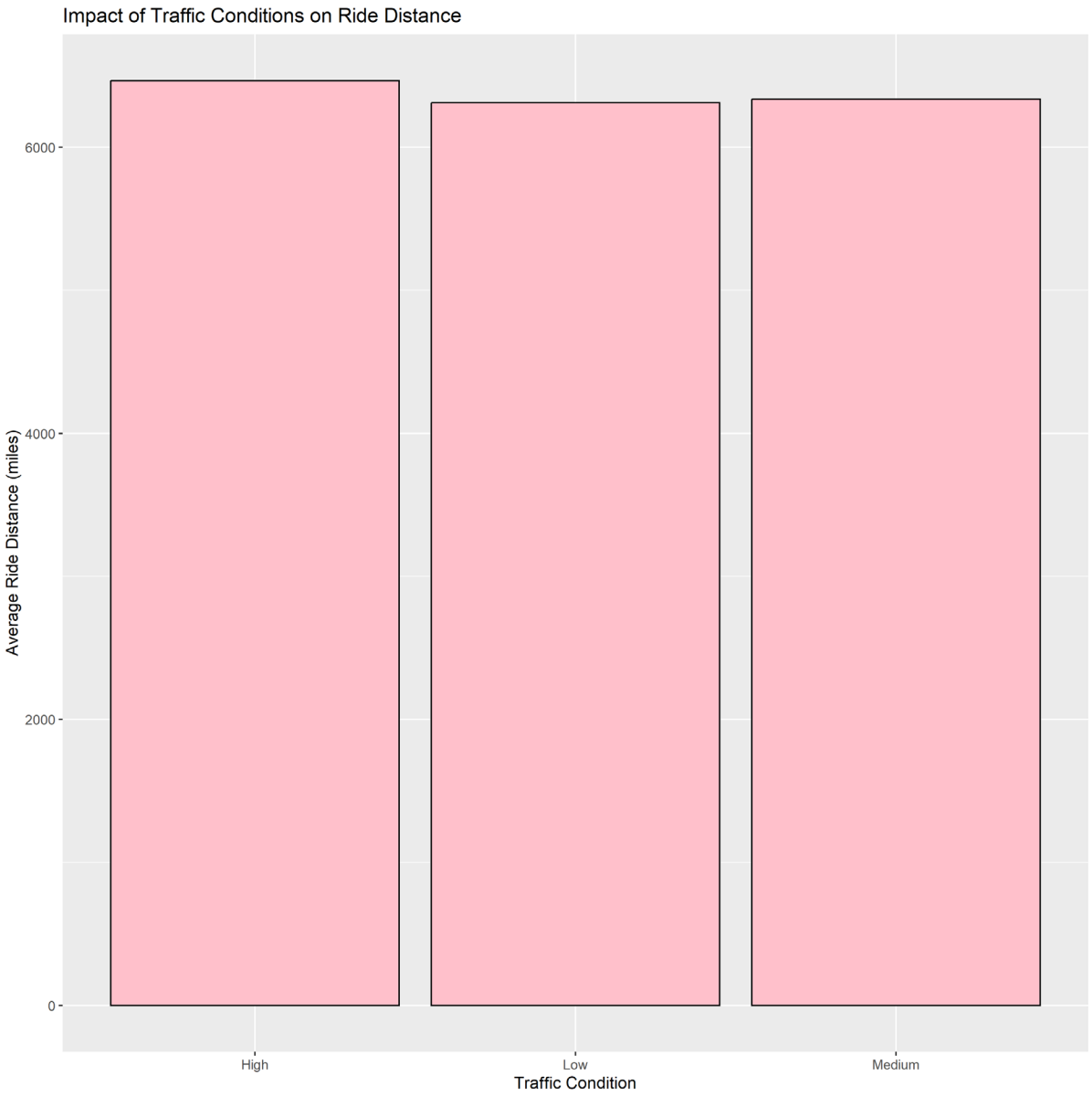

Distribution of Rides by Day of the Week
Number of Rides
100
75
50
25
0
Sun
Mon
Tue
Wed
Thu
Fri
Sat
Day of the Week


Average Fare Amount by Vehicle Type
Average Fare Amount ($)
600
400
200
0
Bus
Motorcycle
Sedan
SUV
Vehicle Type


Distribution of User Ratings
Count
100
50
0
1
2
3
4
5
User Rating


Relationship Between Ride Distance and Fare Amount
Fare Amount ($)
1250
1000
750
500
250
0
0
2500
5000
7500
10000
12500
Ride Distance (miles)


Payment Method Preferences
150
100
50
0
Count
Apple Pay
Cash
Debit Card
PayPal
Visa
Payment Method


Impact of Traffic Conditions on Ride Distance
Average Ride Distance (miles)
6000
4000
2000
0
High
Low
Medium
Traffic Condition