VIT Vidyalankar Institute of Technology Accredited A+ by NAAC	DEPARTMENT OF COMPUTER ENGINEERING

Collaborative Project Report

Semester	B.E. Semester VII – Computer Engineering
Subject	Big Data Analytics
Subject Professor In-charge	Prof. Pankaj Vanvari

Roll Number	Name of Students
21102B0044	Suyash Kamath
21102B0053	Atharv Deshmukh
21102B0036	Atish Limje
21102B0042	Vivek Yadav

Name of the Project:

Data Visualization in R to get inference (Data set: Cyclistic Case Study)

Dataset Link :

https://www.kaggle.com/code/tendobosa/cyclistic-in-what-ways-can-bike-sharing-improve/input

Dataset Size : 900 MB

Dataset Rows : 1,55,04,123

Github : https://github.com/Suyash-Kamath/Semester-7-Projects/tree/main/BDA

Project Details:

1. **Dataset Source**
   The dataset is obtained from Divvy, a bike-sharing program in Chicago, which provides detailed records of bike trips, including information about trip duration, start and end locations, and user demographics.
2. **Dataset Description**
   This dataset contains information on bike trips taken over several years. The goal is to analyze bike usage patterns, understand user behavior, and derive insights for improving urban mobility.
3. **Attributes and Features**
   The dataset comprises approximately 3 million instances (rows), with 12 key attributes (columns):

Attribute	Type	Description
ride_id	Categorical	Unique identifier for each bike ride.
rideable_type	Categorical	Type of bike used (e.g., Classic, Electric).
start_time	Timestamp	Timestamp when the ride started.
end_time	Timestamp	Timestamp when the ride ended.
start_station_id	Numeric	Unique identifier for the starting station.
start_station_name	Categorical	Name of the starting station.
end_station_id	Numeric	Unique identifier for the ending station.
end_station_name	Categorical	Name of the ending station.
duration	Numeric	Duration of the ride in seconds.
user_type	Categorical	Type of user (Subscriber or Customer).
gender	Categorical	Gender of the user (Male, Female, or Other).
birthyear	Numeric	Year of birth of the user, used for age calculations.

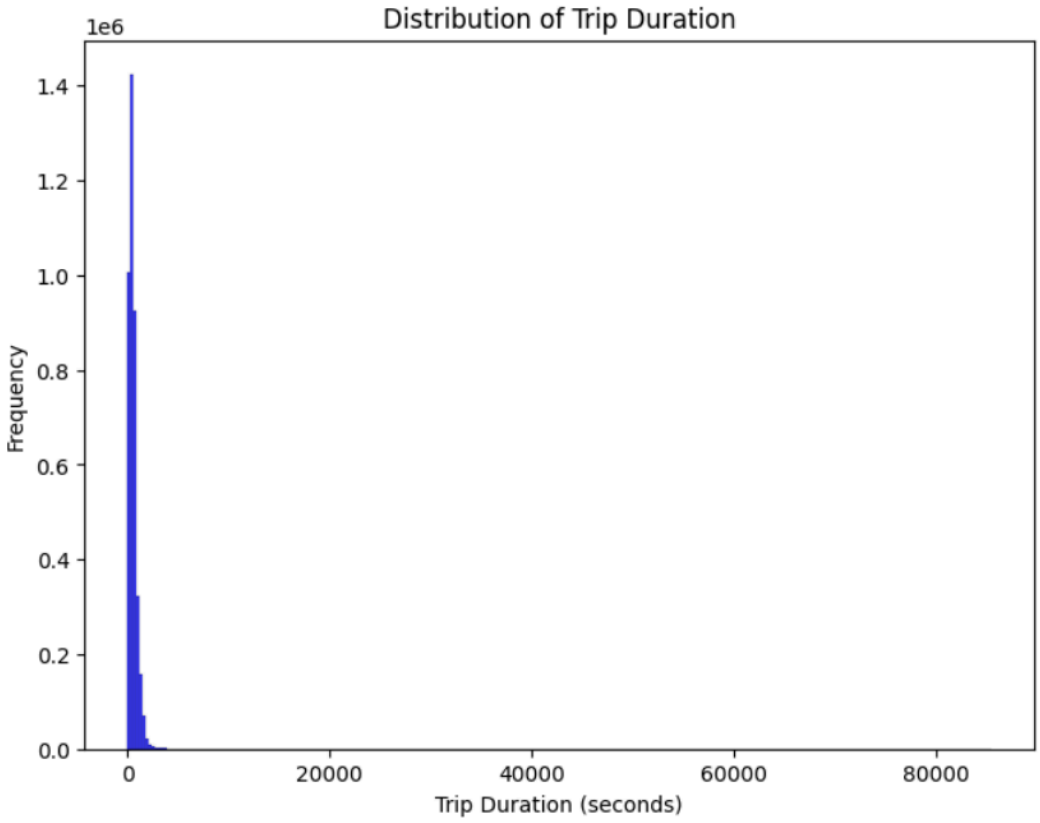
Output Screenshots:

```r
# Load necessary libraries
library(tidyverse)
library(lubridate)

# Load the dataset into R
Combined_Divvy_Trips <- read.csv("Combined_Divvy_Trips.csv", stringsAsFactors = FALSE)

# Check if the dataset loaded correctly
print(head(Combined_Divvy_Trips))

# Convert trip duration to numeric, if necessary
Combined_Divvy_Trips$tripduration <- as.numeric(Combined_Divvy_Trips$tripduration)

# Convert start_time to datetime
Combined_Divvy_Trips$start_time <- ymd_hms(Combined_Divvy_Trips$start_time)

# Drop rows with NA in essential columns to prevent errors in plots
Combined_Divvy_Trips <- Combined_Divvy_Trips %>%
  drop_na(tripduration, usertype, gender, birthyear, from_station_name, to_station_name)
```

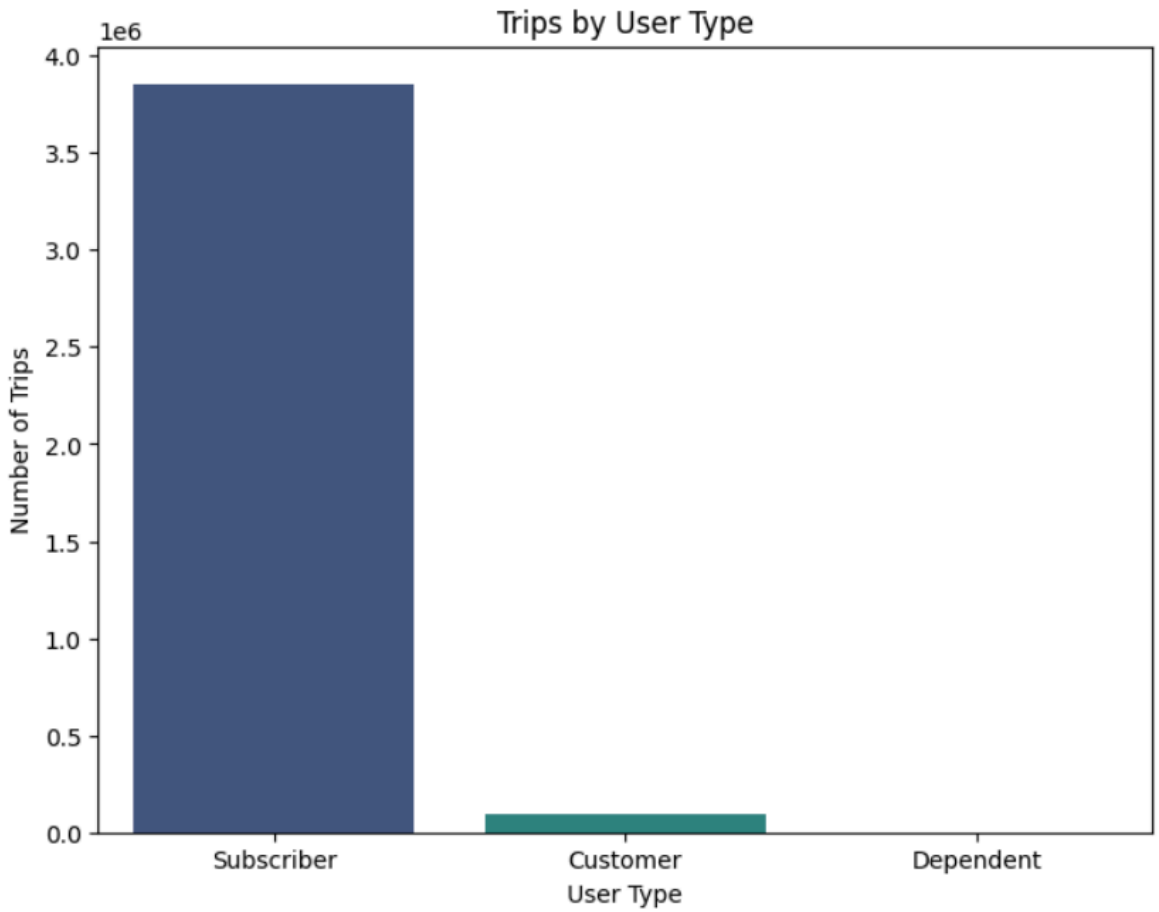
```
    trip_id          start_time            end_time  bikeid tripduration  \
0  16734065  9/30/2017 23:59:58  10/1/2017 00:05:47    1411           349
1  16734064  9/30/2017 23:59:53  10/1/2017 00:05:47    3048           354
2  16734063  9/30/2017 23:59:06  10/1/2017 00:02:52    2590           226
3  16734062  9/30/2017 23:58:56  10/1/2017 00:07:37     551           521
4  16734061  9/30/2017 23:58:47  10/1/2017 00:07:37    1287           530

   from_station_id             from_station_name  to_station_id  \
0              216  California Ave & Division St            259
1              216  California Ave & Division St            259
2              141        Clark St & Lincoln Ave            144
3               96   Desplaines St & Randolph St            217
4               96   Desplaines St & Randolph St            217

                  to_station_name    usertype  gender  birthyear
0     California Ave & Francis Pl  Subscriber    Male     1985.0
1     California Ave & Francis Pl  Subscriber    Male     1979.0
2       Larrabee St & Webster Ave  Subscriber    Male     1993.0
3  Racine Ave (May St) & Fulton St    Customer     NaN        NaN
4  Racine Ave (May St) & Fulton St  Subscriber  Female     1994.0
```

Histogram :

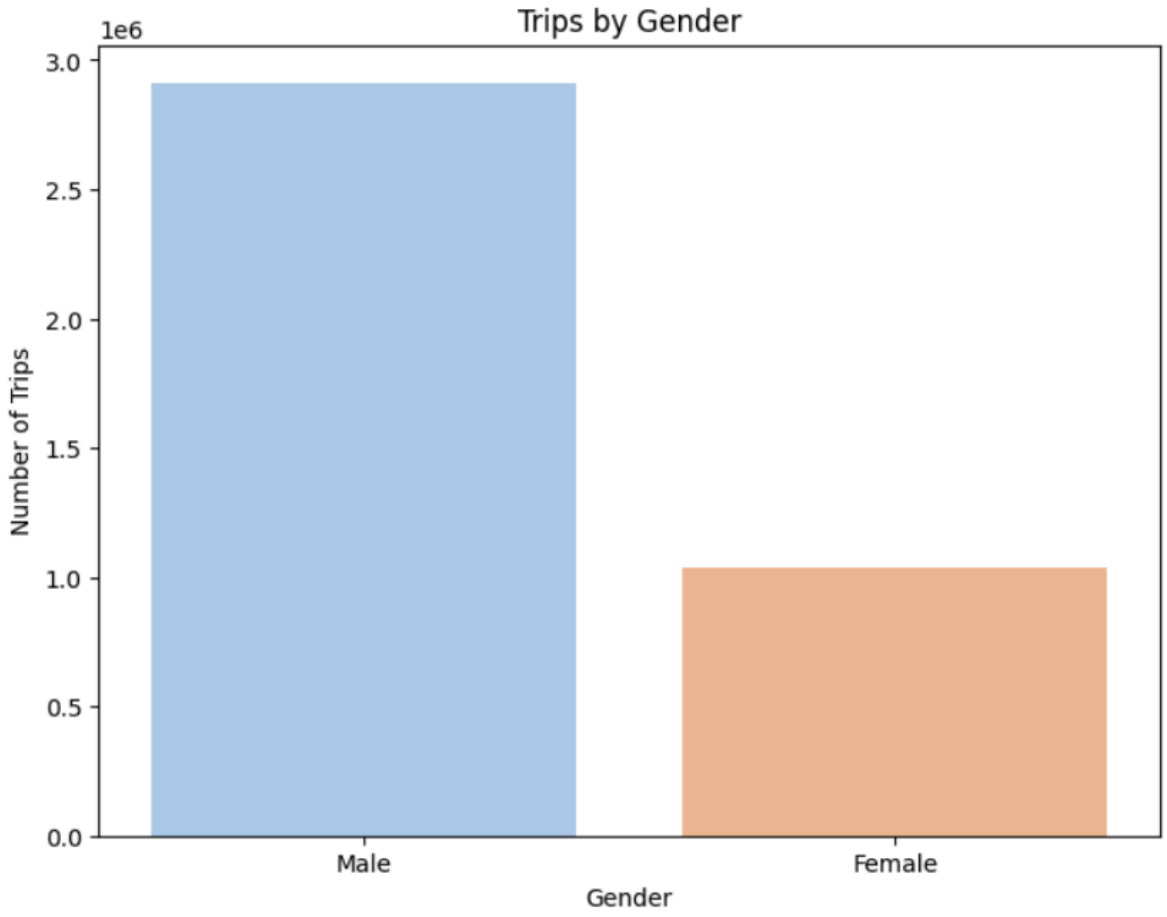
```r
# 1. Histogram of Trip Duration
ggplot(Combined_Divvy_Trips, aes(x = tripduration)) +
  geom_histogram(binwidth = 300, fill = "blue", color = "black") +
  ggtitle("Distribution of Trip Duration") +
  xlab("Trip Duration (seconds)") +
  ylab("Frequency")
```



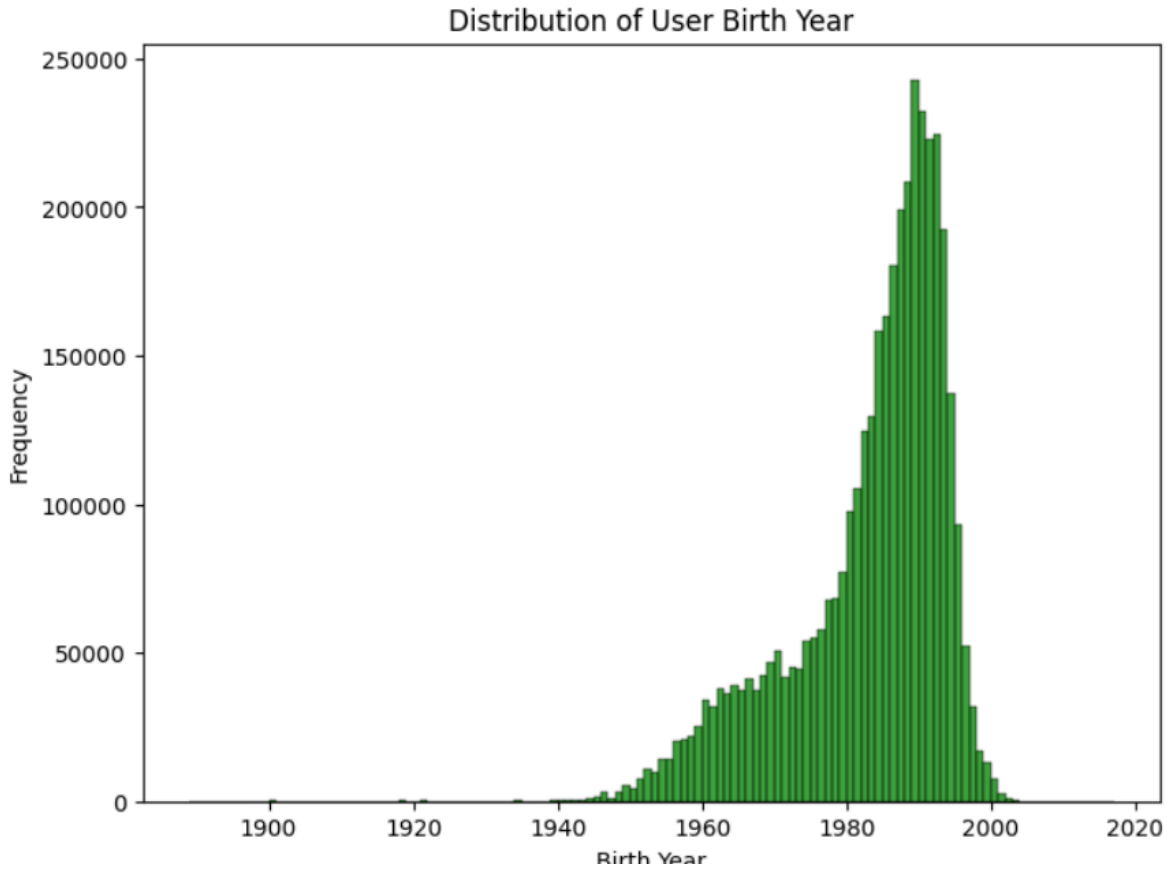

```
# 2. Trips by User Type (Subscriber vs. Customer)
ggplot(Combined_Divvy_Trips, aes(x = usertype)) +
  geom_bar(fill = "viridis") +
  ggtitle("Trips by User Type") +
  xlab("User Type") +
  ylab("Number of Trips")
```



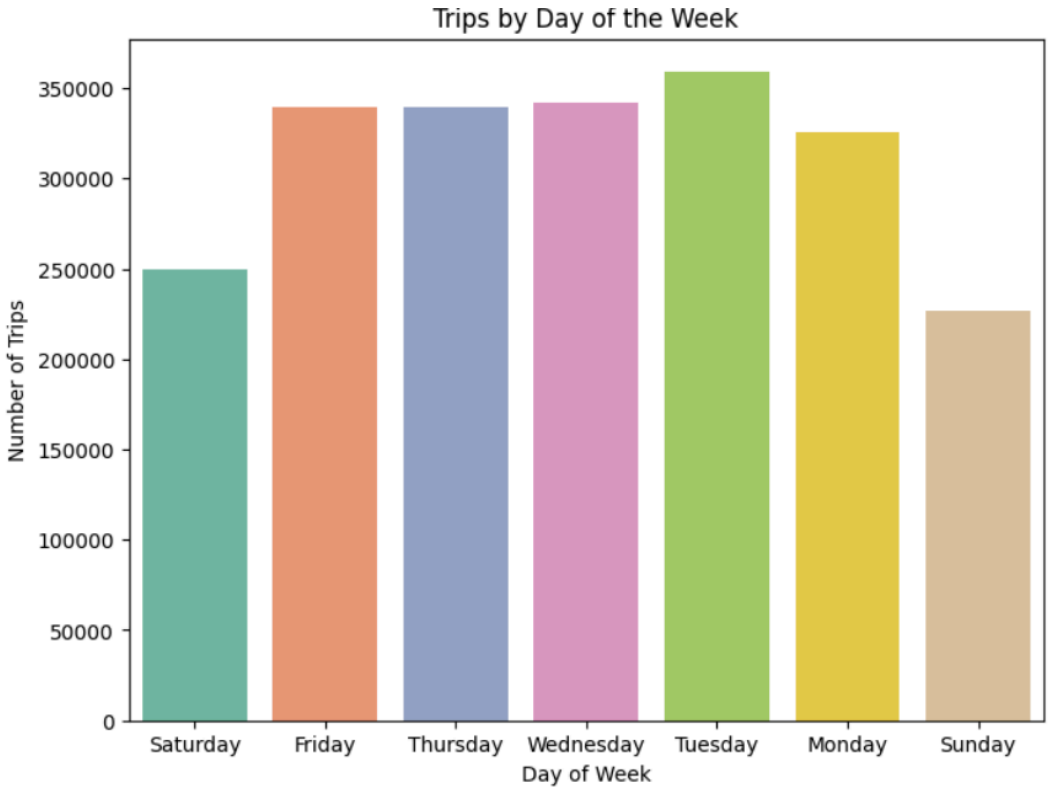

```
# 3. Trips by Gender
ggplot(Combined_Divvy_Trips, aes(x = gender)) +
  geom_bar(fill = "pastel") +
  ggtitle("Trips by Gender") +
  xlab("Gender") +
  ylab("Number of Trips")
```



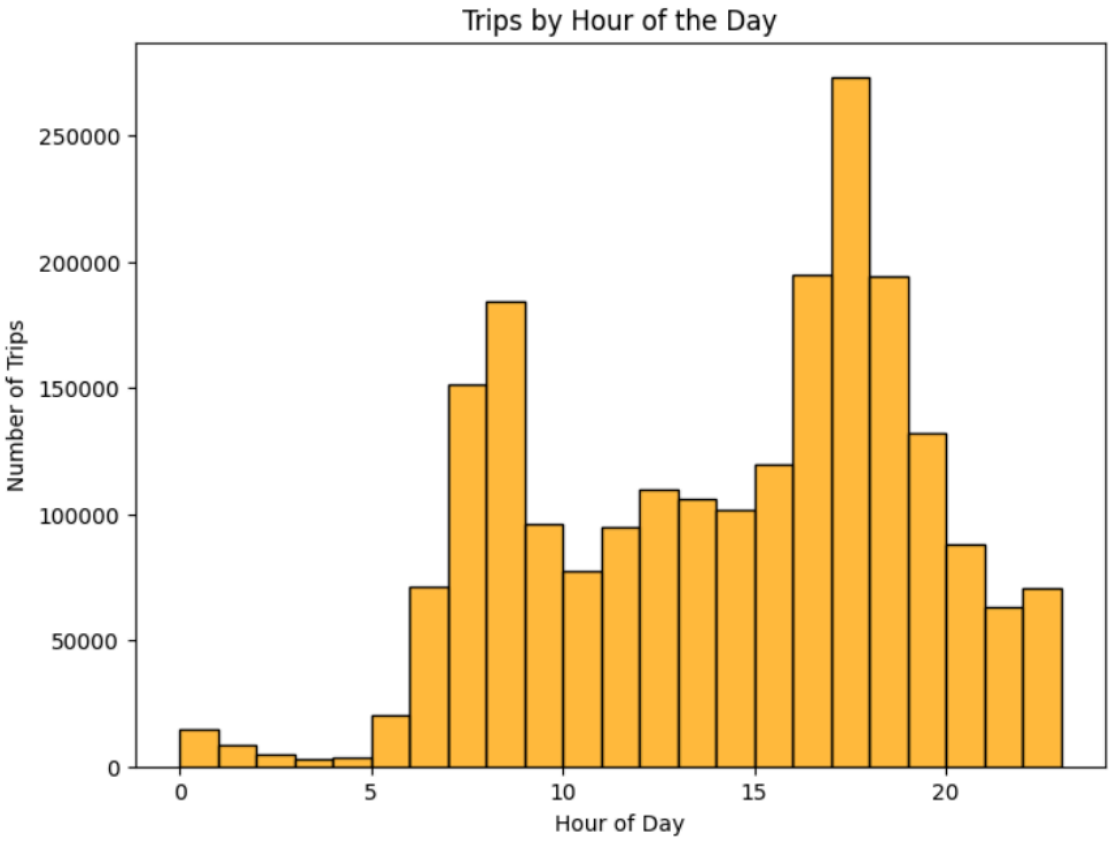

```r
# 4. Trips by Birth Year (Ignoring NA values)
ggplot(Combined_Divvy_Trips %>% filter(!is.na(birthyear)), aes(x = birthyear)) +
  geom_histogram(binwidth = 1, fill = "green", color = "black") +
  ggtitle("Distribution of User Birth Year") +
  xlab("Birth Year") +
  ylab("Frequency")
```



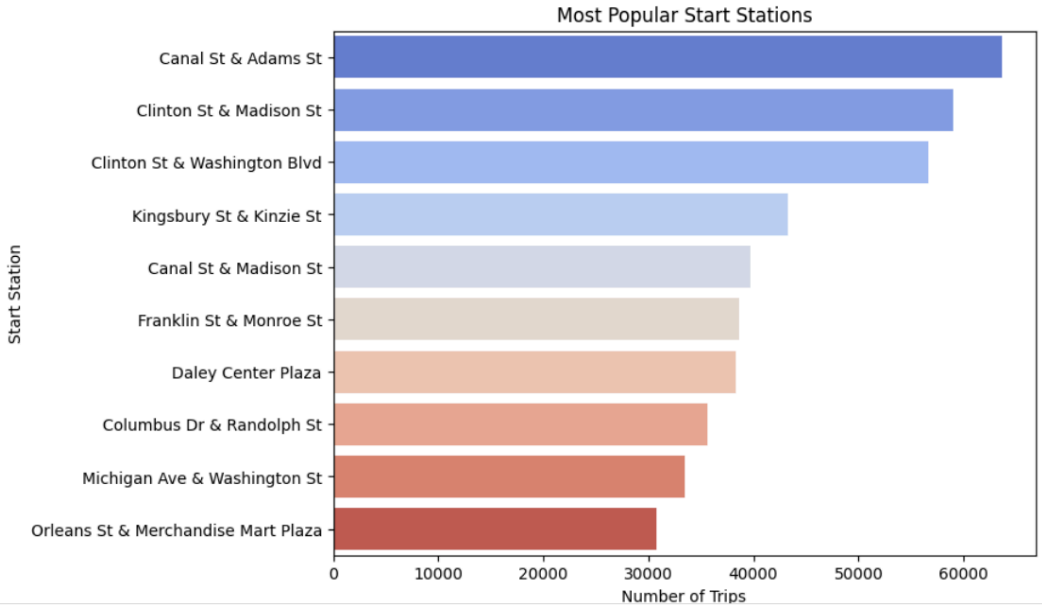

```
# 5. Trips Over the Days of the Week
Combined_Divvy_Trips$day_of_week <- weekdays(Combined_Divvy_Trips$start_time)

ggplot(Combined_Divvy_Trips, aes(x = day_of_week)) +
  geom_bar(fill = "Set2") +
  ggtitle("Trips by Day of the Week") +
  xlab("Day of Week") +
  ylab("Number of Trips")
```



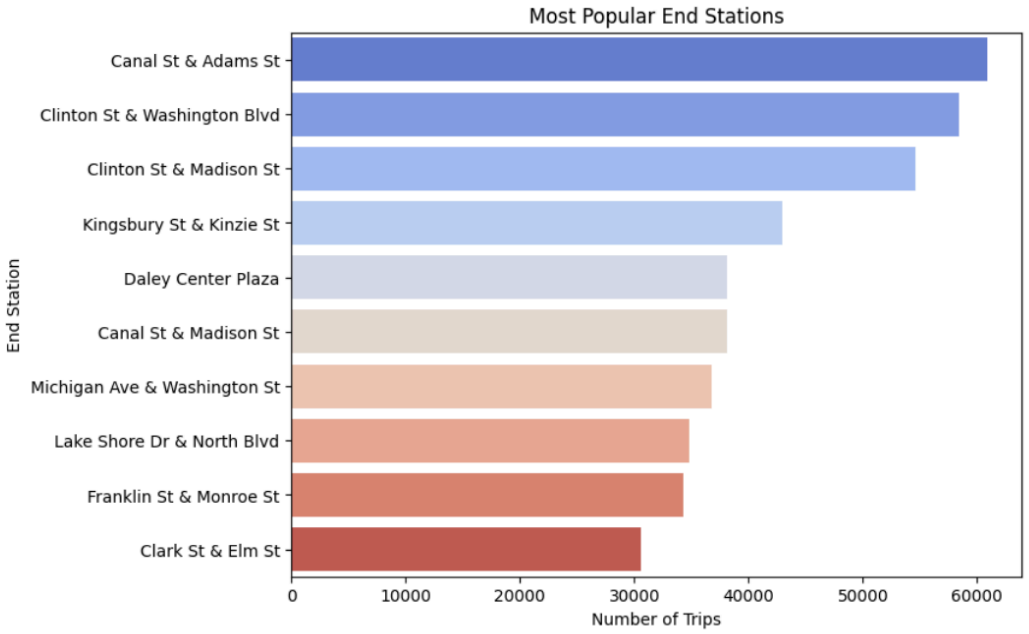

```r
# 6. Trips by Hour of the Day
Combined_Divvy_Trips$hour <- hour(Combined_Divvy_Trips$start_time)

ggplot(Combined_Divvy_Trips, aes(x = hour)) +
  geom_histogram(binwidth = 1, fill = "orange", color = "black") +
  ggtitle("Trips by Hour of the Day") +
  xlab("Hour of Day") +
  ylab("Number of Trips")
```




```r
# 7. Most Popular Start Stations
ggplot(Combined_Divvy_Trips %>% count(from_station_name) %>% top_n(10, n), aes(y = reorder(from_station_name, n), x = n)) +
  geom_bar(stat = "identity", fill = "coolwarm") +
  ggtitle("Most Popular Start Stations") +
  xlab("Number of Trips") +
  ylab("Start Station")
```

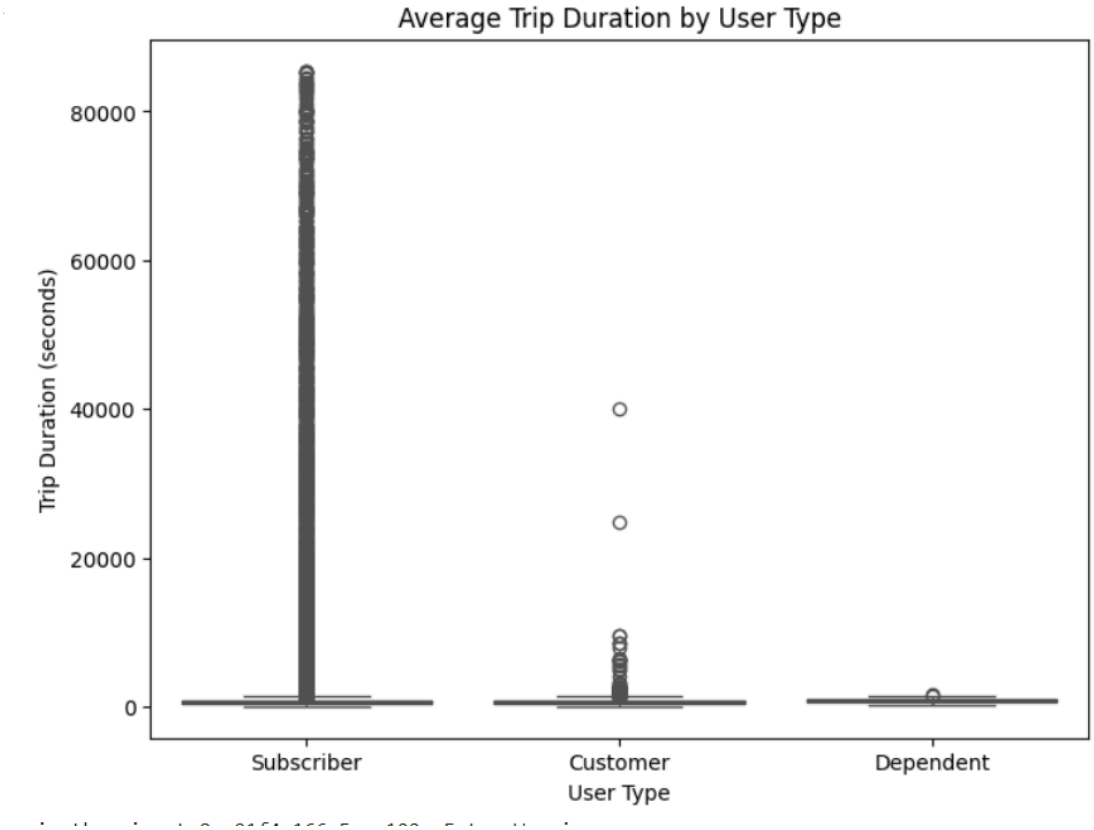


```
# 8. Most Popular End Stations
ggplot(Combined_Divvy_Trips %>% count(to_station_name) %>% top_n(10, n), aes(y = reorder(to_station_name, n), x = n)) +
  geom_bar(stat = "identity", fill = "coolwarm") +
  ggtitle("Most Popular End Stations") +
  xlab("Number of Trips") +
  ylab("End Station")
```

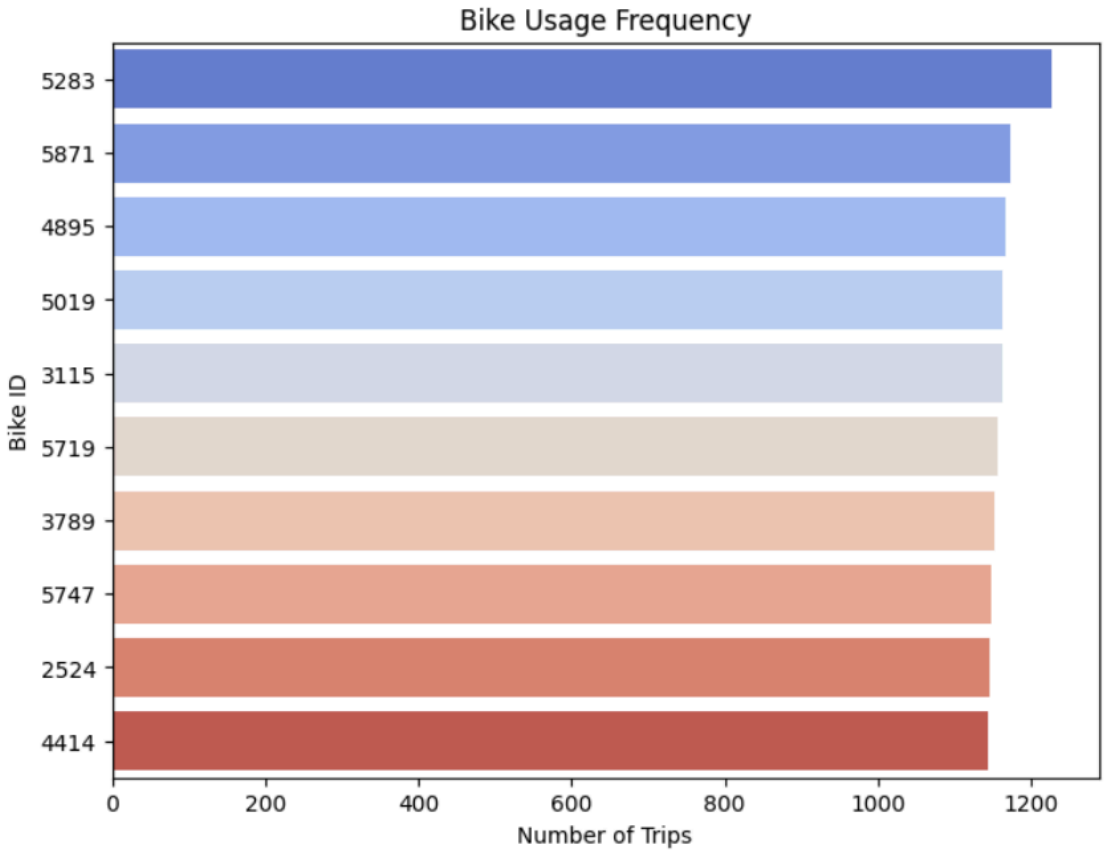


<ipython-input-8-c91f4c166a5a>:94: FutureWarning:

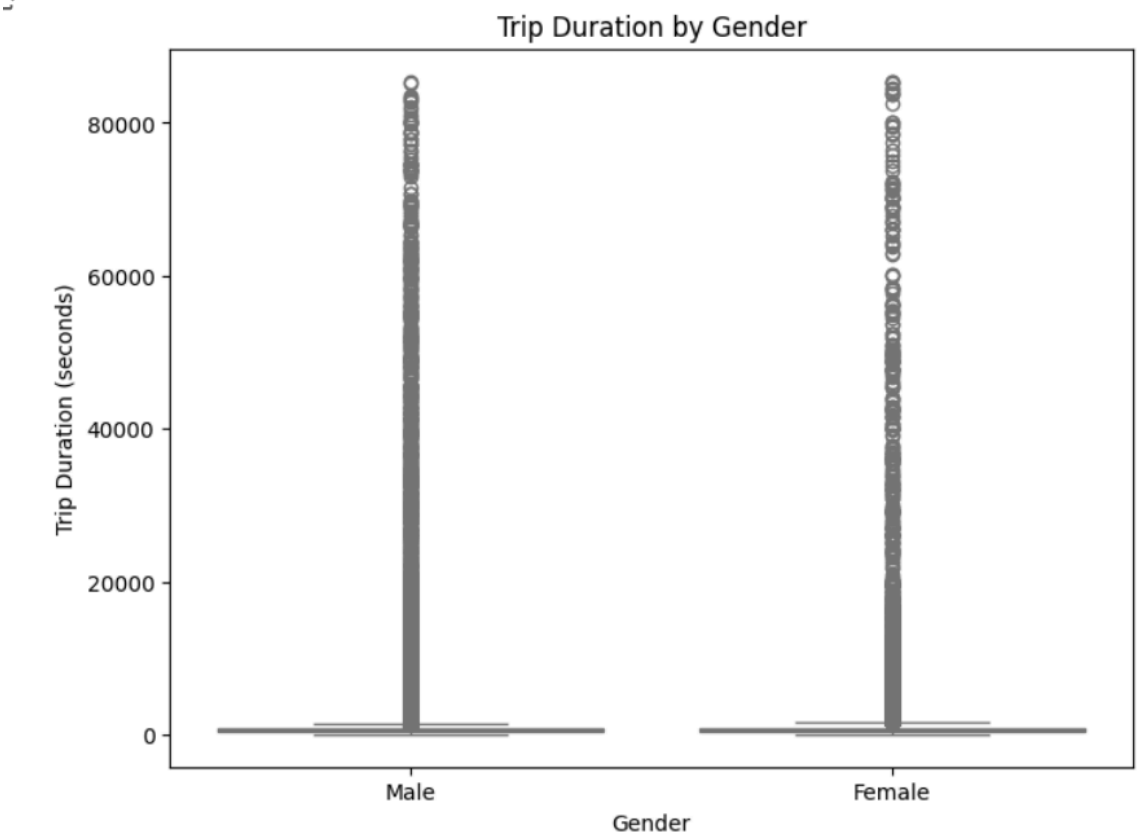

```
# 9. Average Trip Duration by User Type
ggplot(Combined_Divvy_Trips, aes(x = usertype, y = tripduration)) +
  geom_boxplot(fill = "muted") +
  ggtitle("Average Trip Duration by User Type") +
  xlab("User Type") +
  ylab("Trip Duration (seconds)")
```



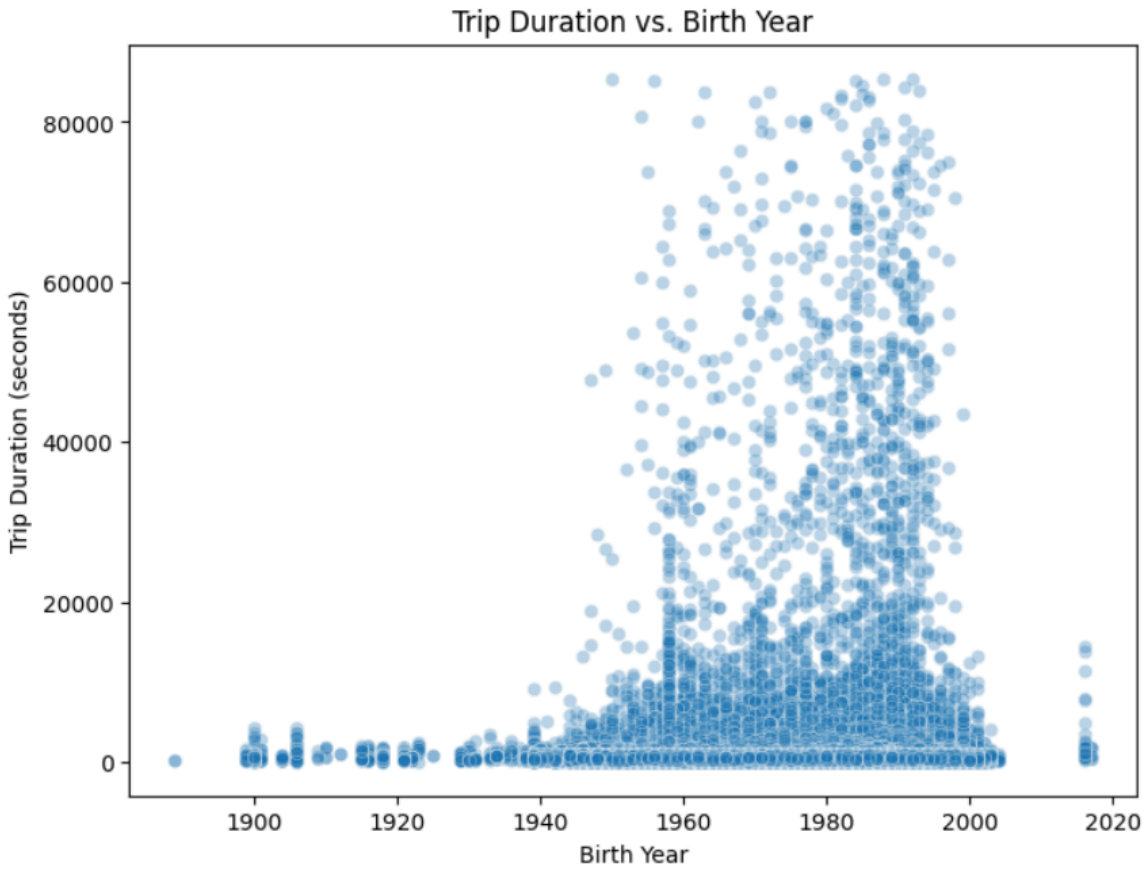

```
# 10. Bike Usage Frequency
ggplot(Combined_Divvy_Trips %>% count(bikeid) %>% top_n(10, n), aes(y = reorder(bikeid, n), x = n)) +
  geom_bar(stat = "identity", fill = "coolwarm") +
  ggtitle("Bike Usage Frequency") +
  xlab("Number of Trips") +
  ylab("Bike ID")
```



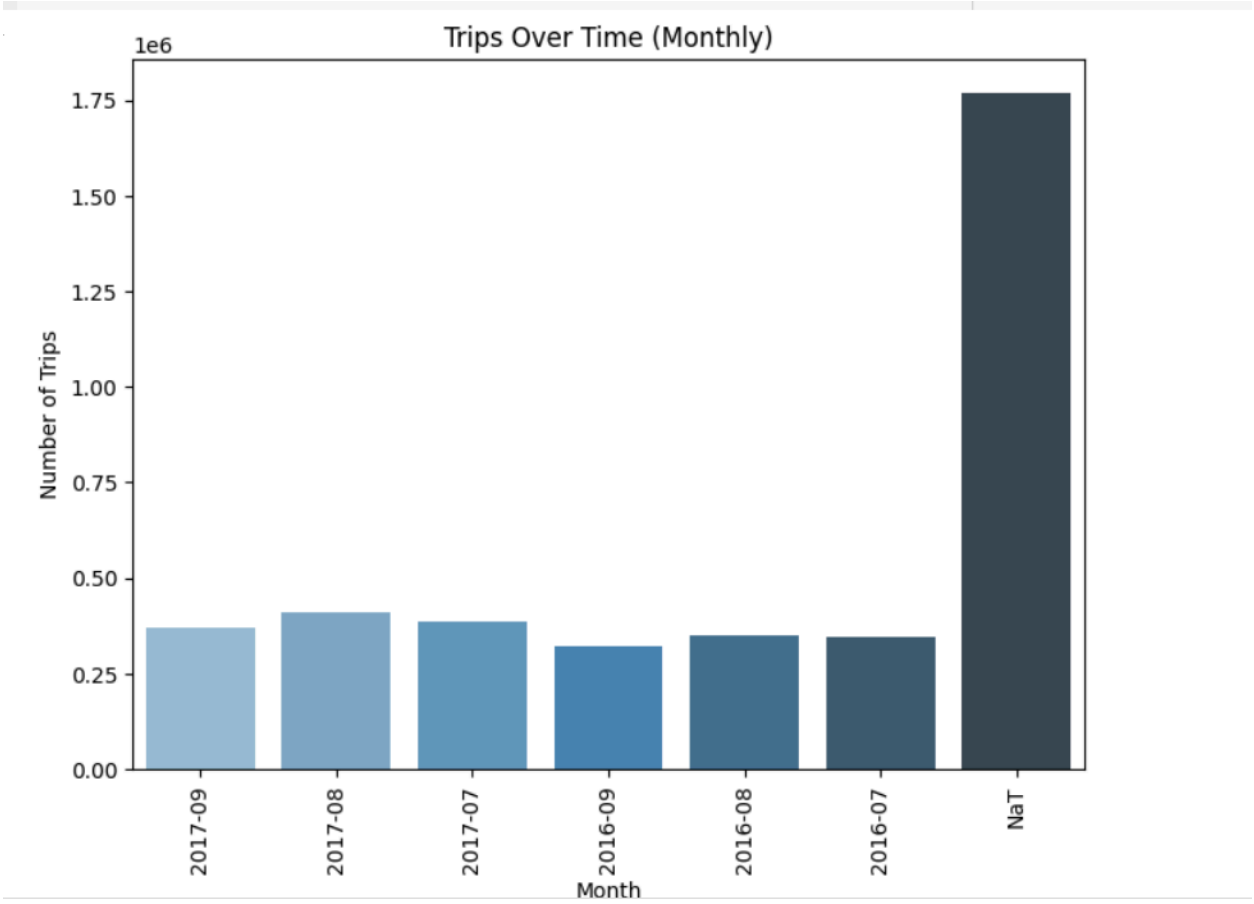

```
# 11. Trip Duration by Gender
ggplot(Combined_Divvy_Trips, aes(x = gender, y = tripduration)) +
  geom_boxplot(fill = "pastel") +
  ggtitle("Trip Duration by Gender") +
  xlab("Gender") +
  ylab("Trip Duration (seconds)")
```



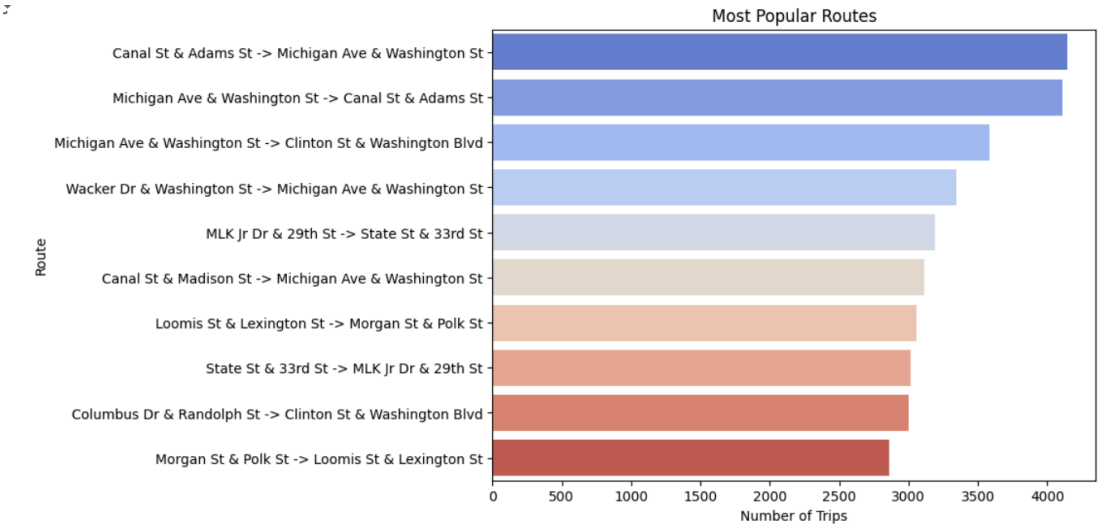

```r
# 12. Trip Duration by Birth Year (Ignoring NA values)
ggplot(Combined_Divvy_Trips %>% filter(!is.na(birthyear)), aes(x = birthyear, y = tripduration)) +
  geom_point(alpha = 0.3) +
  ggtitle("Trip Duration vs. Birth Year") +
  xlab("Birth Year") +
  ylab("Trip Duration (seconds)")
```



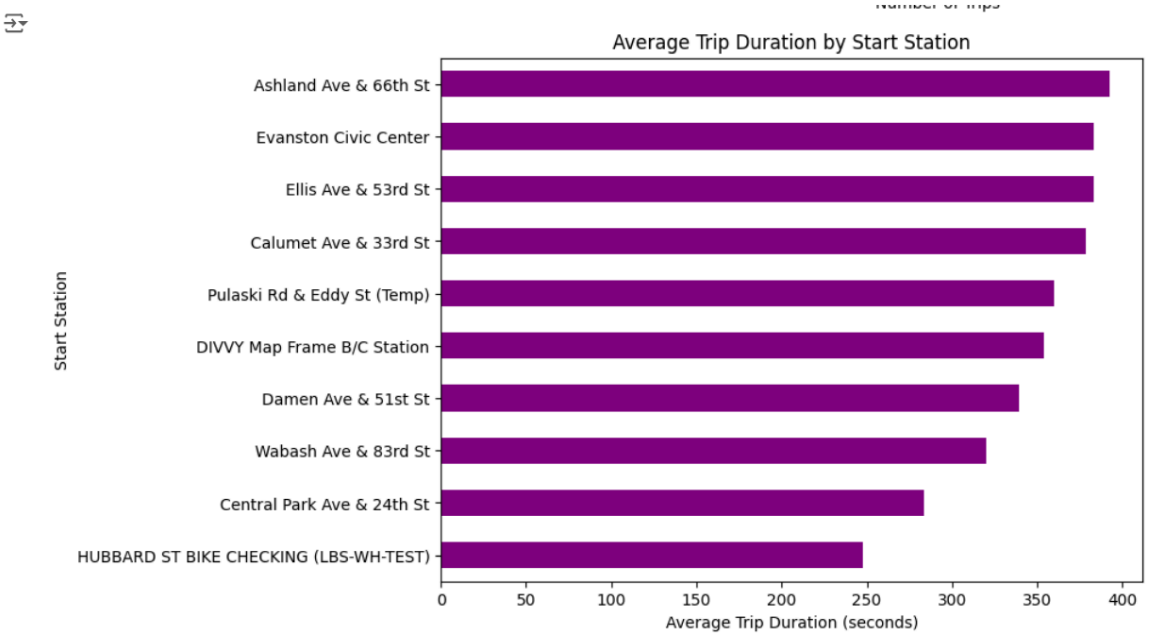

```r
# 13. Trips Over Time (Monthly)
Combined_Divvy_Trips$month <- floor_date(Combined_Divvy_Trips$start_time, "month")

ggplot(Combined_Divvy_Trips, aes(x = month)) +
  geom_bar(fill = "Blues") +
  ggtitle("Trips Over Time (Monthly)") +
  theme(axis.text.x = element_text(angle = 90)) +
  xlab("Month") +
  ylab("Number of Trips")
```



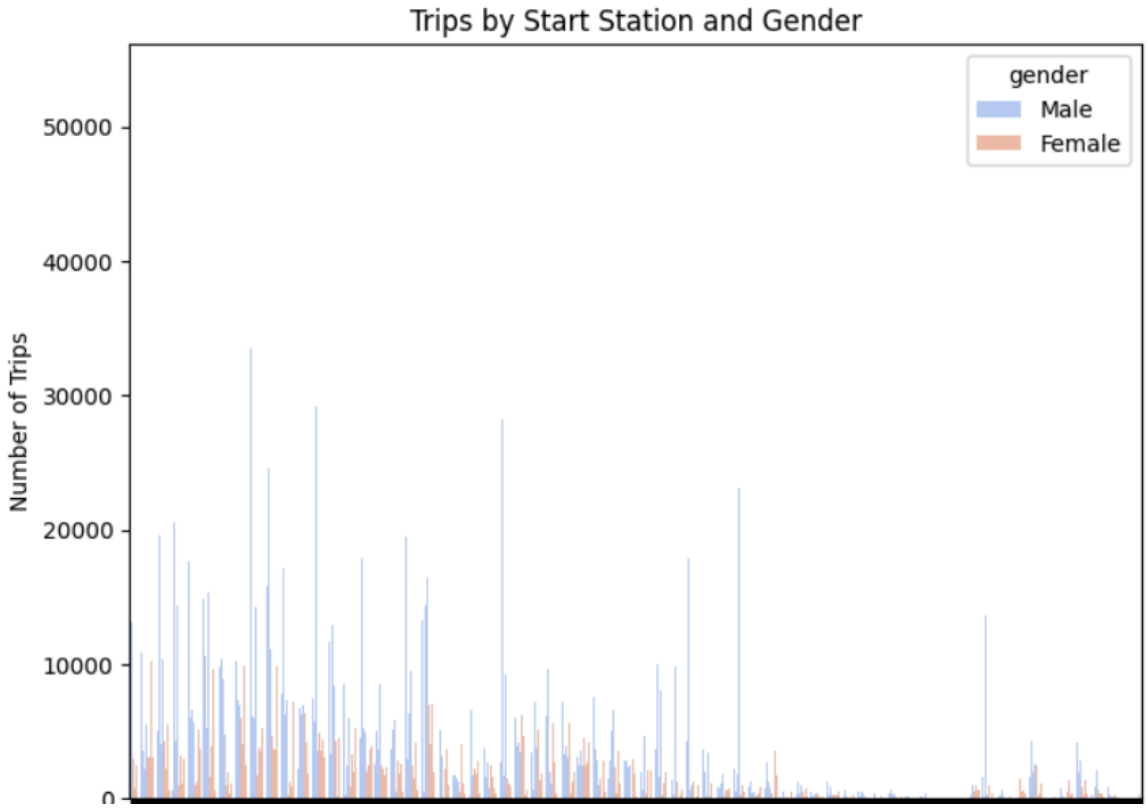

```r
# 14. Trips by Start and End Station
Combined_Divvy_Trips$route <- paste(Combined_Divvy_Trips$from_station_name, "->", Combined_Divvy_Trips$to_station_name)

ggplot(Combined_Divvy_Trips %>% count(route) %>% top_n(10, n), aes(y = reorder(route, n), x = n)) +
  geom_bar(stat = "identity", fill = "coolwarm") +
  ggtitle("Most Popular Routes") +
  xlab("Number of Trips") +
  ylab("Route")
```



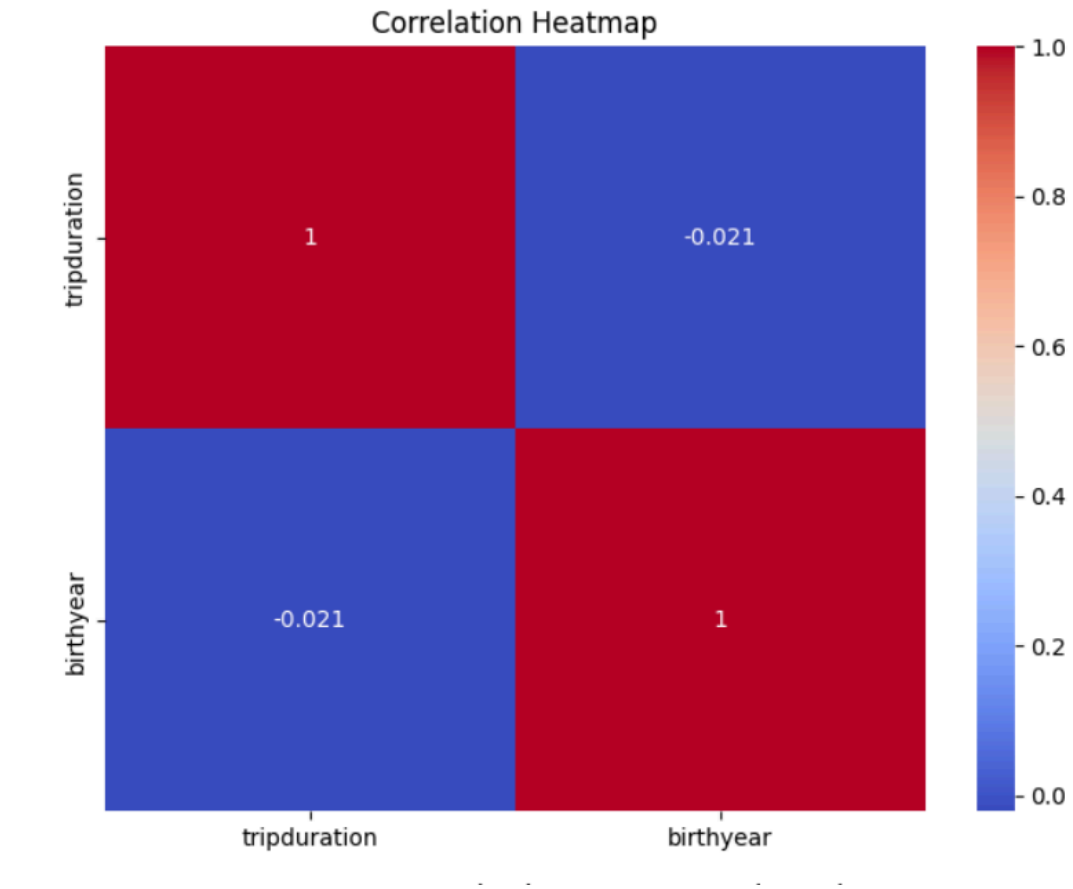
```
# 15. Average Trip Duration by Start Station
Combined_Divvy_Trips %>%
  group_by(from_station_name) %>%
  summarize(mean_duration = mean(tripduration, na.rm = TRUE)) %>%
  top_n(10, mean_duration) %>%
  ggplot(aes(y = reorder(from_station_name, mean_duration), x = mean_duration)) +
  geom_bar(stat = "identity", fill = "purple") +
  ggtitle("Average Trip Duration by Start Station") +
  xlab("Average Trip Duration (seconds)") +
  ylab("Start Station")
```



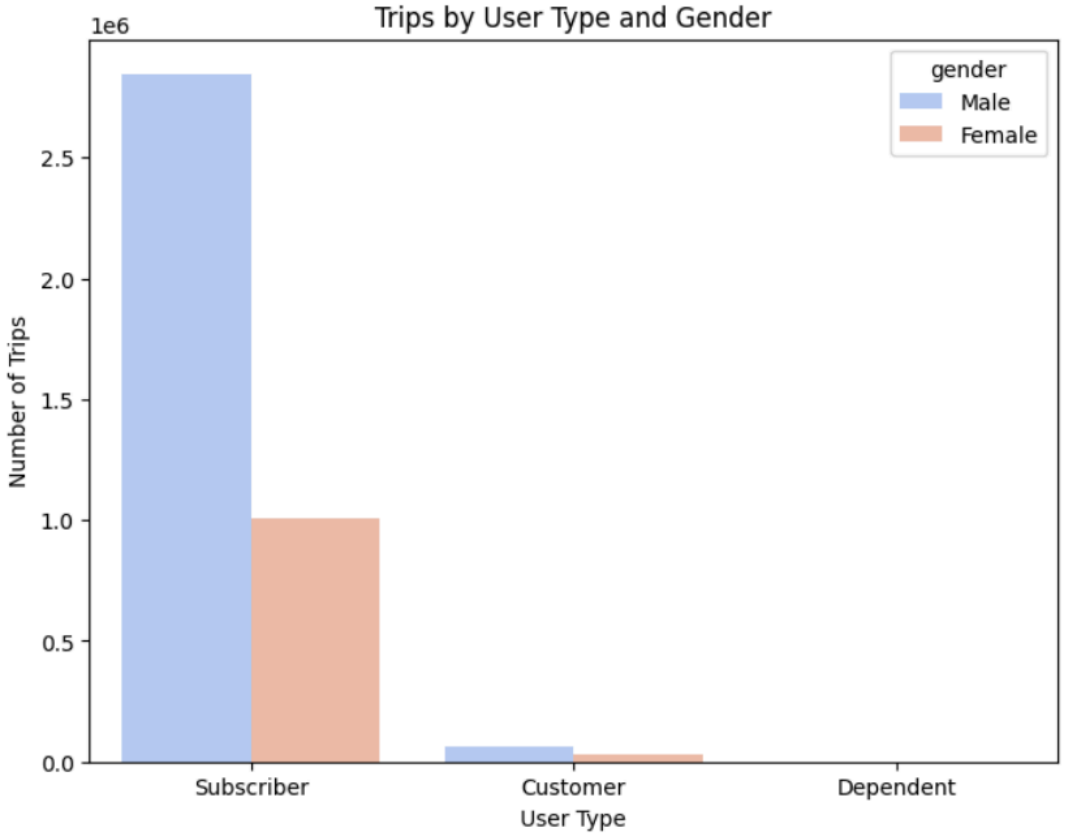

```
# 16. Trips by Start Station and Gender
ggplot(Combined_Divvy_Trips, aes(x = from_station_name, fill = gender)) +
  geom_bar() +
  ggtitle("Trips by Start Station and Gender") +
  xlab("Start Station") +
  ylab("Number of Trips") +
  theme(axis.text.x = element_text(angle = 90))
```



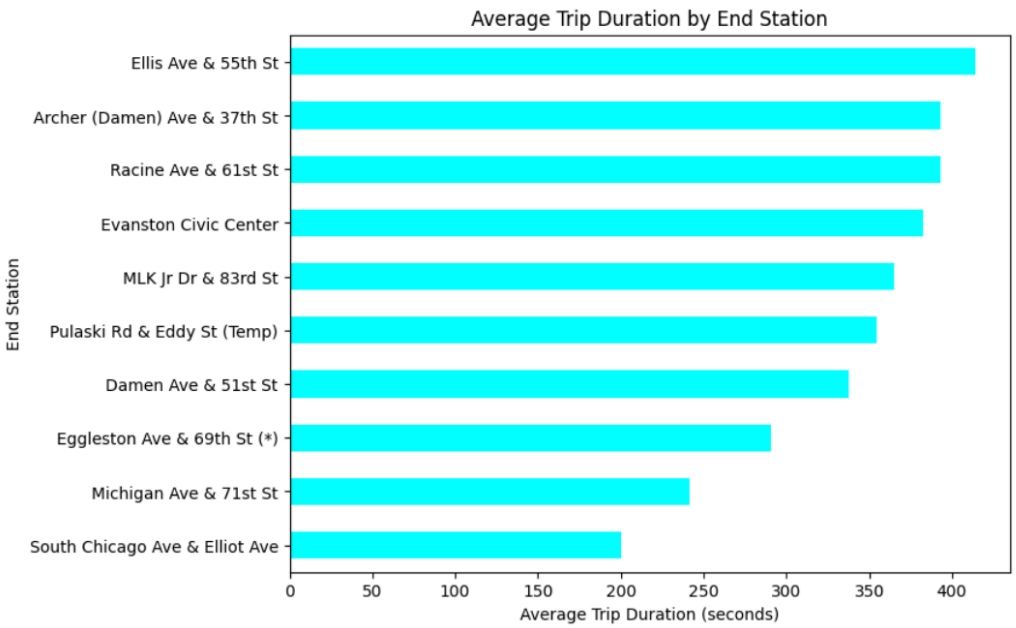

```r
# 17. Correlation Matrix Heatmap
corr_matrix <- cor(Combined_Divvy_Trips %>% select(tripduration, birthyear), use = "complete.obs")

ggplot(melt(corr_matrix), aes(Var1, Var2, fill = value)) +
  geom_tile() +
  geom_text(aes(label = round(value, 2))) +
  scale_fill_gradient2() +
  ggtitle("Correlation Heatmap") +
  theme_minimal()
```



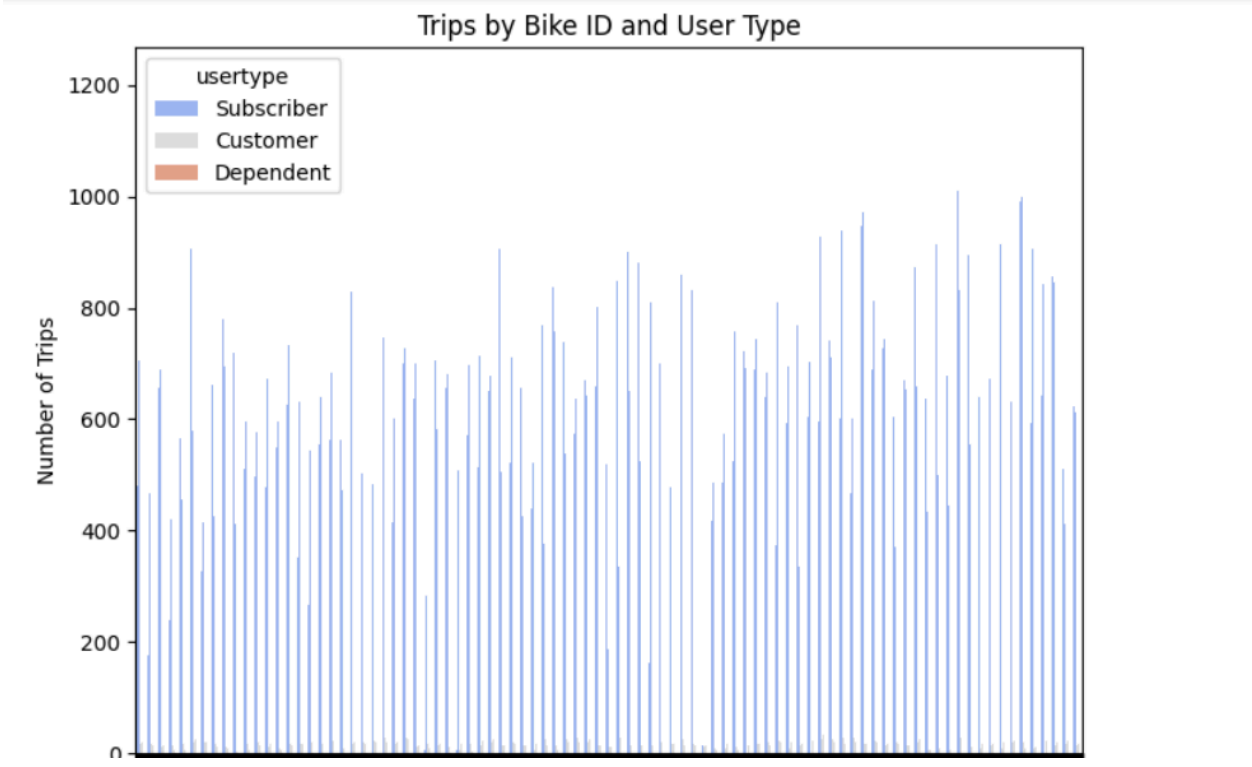

```
# 18. Trips by User Type and Gender
ggplot(Combined_Divvy_Trips, aes(x = usertype, fill = gender)) +
  geom_bar() +
  ggtitle("Trips by User Type and Gender") +
  xlab("User Type") +
  ylab("Number of Trips")
```




```
# 19. Average Trip Duration by End Station
Combined_Divvy_Trips %>%
  group_by(to_station_name) %>%
  summarize(mean_duration = mean(tripduration, na.rm = TRUE)) %>%
  top_n(10, mean_duration) %>%
  ggplot(aes(y = reorder(to_station_name, mean_duration), x = mean_duration)) +
  geom_bar(stat = "identity", fill = "cyan") +
  ggtitle("Average Trip Duration by End Station") +
  xlab("Average Trip Duration (seconds)") +
  ylab("End Station")
```




```
# 20. Trips by Bike ID and User Type
ggplot(Combined_Divvy_Trips, aes(x = bikeid, fill = usertype)) +
  geom_bar() +
  ggtitle("Trips by Bike ID and User Type") +
  xlab("Bike ID") +
  ylab("Number of Trips") +
  theme(axis.text.x = element_text(angle = 90))
```




```
print("Everything is done")
```

```
Everything is done
```

Inference and Observation

1. Trip Duration Distribution

- The histogram of trip duration reveals that most trips are relatively short, with a significant peak at lower durations. This suggests that many users opt for quick rides, which may reflect the usage patterns of the bike-sharing service.

2. User Type Analysis

- The bar plot shows a clear distinction between subscribers and customers, with subscribers dominating the number of trips. This indicates that regular users are more likely to utilize the service than one-time customers, highlighting the importance of subscriber retention strategies.

3. Gender Distribution

- The gender-based trip distribution indicates a disparity between male and female users. A higher number of trips by one gender suggests potential marketing opportunities to engage the underrepresented gender.

4. Age Analysis via Birth Year

- The histogram of user birth years indicates the age demographics of the users, which may suggest that the bike-sharing service is particularly popular among younger individuals. This could influence targeted marketing campaigns or service enhancements.

5. Weekly Trip Patterns

- The analysis of trips by day of the week shows fluctuations in usage, with certain days attracting more riders. This information can help in optimizing bike availability and service offerings on peak days.

6. Hourly Trip Trends

- The histogram of trips by hour indicates peak times for bike usage, likely corresponding with commuting hours. Understanding these patterns can aid in resource allocation and scheduling maintenance.

7. Popular Start and End Stations

- The visualizations for popular start and end stations highlight the key areas of bike usage. These insights can inform strategic planning for station placements and potential expansions.

8. Trip Duration by User Type

- The boxplot comparing average trip durations between user types suggests that one group may consistently take longer trips than the other. This could prompt further investigation into the reasons behind these differences.

9. Bike Usage Frequency

- The bike usage frequency plot showcases which bikes are most popular. This data can inform maintenance schedules and inventory decisions for the fleet.

10. Trips Over Time

- The monthly trip trend visualization indicates seasonality in bike usage. A significant increase or decrease in specific months could guide promotional activities and inventory management.

11. Route Popularity

- The most popular routes visualization identifies key pathways for users. Understanding these routes can enhance route planning for better efficiency.

12. Average Trip Duration by Station

- Analyzing the average trip duration by start station reveals stations where users may spend more time riding. This information can aid in assessing station performance and rider preferences.

13. Correlation Insights

- The correlation matrix heatmap indicates relationships between trip duration and birth year. Noting any significant correlations can drive targeted user engagement strategies.

14. User Type and Gender Interactions

- The plot illustrating trips by user type and gender reveals insights into gender behavior within different user categories. This could lead to tailored services for different user segments.

Conclusion

These visualizations collectively provide valuable insights into user behavior, trip patterns, and operational effectiveness within the Divvy bike-sharing system. Stakeholders can utilize this data to enhance user experience, optimize resources, and tailor marketing strategies to specific demographics or usage trends.