Explore Bike Share Data

For this project, your goal is to ask and answer three questions about the available bikeshare data from Washington, Chicago, and New York. This notebook can be submitted directly through the workspace when you are confident in your results.

You will be graded against the project [Rubric (https://review.udacity.com/#!/rubrics/2508/view)](https://review.udacity.com/#!/rubrics/2508/view) by a mentor after you have submitted. To get you started, you can use the template below, but feel free to be creative in your solutions!

Bike Share Data for Marketing Strategy

Our bikeshare company has collected data from Washington, Chicago, and New York to inform a new marketing campaign aimed at maximizing profitability. This project will guide us in identifying key areas and demographics for targeting. First, we'll analyze the data to determine the most popular bikeshare location. Next, we'll explore demographic details, including the predominant age and gender of our users. Finally, we'll examine seasonal trends to select the optimal month for launching the campaign.

In [58]:

```
ny = read.csv('new_york_city.csv')
wash = read.csv('washington.csv')
chi = read.csv('chicago.csv')
```

In [59]:
```
head(ny)
```

X	Start.Time	End.Time	Trip.Duration	Start.Station	End.Station	User.Type	Gend
5688089	2017-06-11 14:55:05	2017-06-11 15:08:21	795	Suffolk St & Stanton St	W Broadway & Spring St	Subscriber	Male
4096714	2017-05-11 15:30:11	2017-05-11 15:41:43	692	Lexington Ave & E 63 St	1 Ave & E 78 St	Subscriber	Male
2173887	2017-03-29 13:26:26	2017-03-29 13:48:31	1325	1 Pl & Clinton St	Henry St & Degraw St	Subscriber	Male
3945638	2017-05-08 19:47:18	2017-05-08 19:59:01	703	Barrow St & Hudson St	W 20 St & 8 Ave	Subscriber	Fema
6208972	2017-06-21 07:49:16	2017-06-21 07:54:46	329	1 Ave & E 44 St	E 53 St & 3 Ave	Subscriber	Male
1285652	2017-02-22 18:55:24	2017-02-22 19:12:03	998	State St & Smith St	Bond St & Fulton St	Subscriber	Male

In [60]:
```
head(wash)
```

X	Start.Time	End.Time	Trip.Duration	Start.Station	End.Station	User.Type
1621326	2017-06-21 08:36:34	2017-06-21 08:44:43	489.066	14th & Belmont St NW	15th & K St NW	Subscriber
482740	2017-03-11 10:40:00	2017-03-11 10:46:00	402.549	Yuma St & Tenley Circle NW	Connecticut Ave & Yuma St NW	Subscriber
1330037	2017-05-30 01:02:59	2017-05-30 01:13:37	637.251	17th St & Massachusetts Ave NW	5th & K St NW	Subscriber
665458	2017-04-02 07:48:35	2017-04-02 08:19:03	1827.341	Constitution Ave & 2nd St NW/DOL	M St & Pennsylvania Ave NW	Customer
1481135	2017-06-10 08:36:28	2017-06-10 09:02:17	1549.427	Henry Bacon Dr & Lincoln Memorial Circle NW	Maine Ave & 7th St SW	Subscriber
1148202	2017-05-14 07:18:18	2017-05-14 07:24:56	398.000	1st & K St SE	Eastern Market Metro / Pennsylvania Ave & 7th St SE	Subscriber

In [61]:
```
head(chi)
```

X	Start.Time	End.Time	Trip.Duration	Start.Station	End.Station	User.Type	Gend
1423854	2017-06-23 15:09:32	2017-06-23 15:14:53	321	Wood St & Hubbard St	Damen Ave & Chicago Ave	Subscriber	Male
955915	2017-05-25 18:19:03	2017-05-25 18:45:53	1610	Theater on the Lake	Sheffield Ave & Waveland Ave	Subscriber	Fema
9031	2017-01-04 08:27:49	2017-01-04 08:34:45	416	May St & Taylor St	Wood St & Taylor St	Subscriber	Male
304487	2017-03-06 13:49:38	2017-03-06 13:55:28	350	Christiana Ave & Lawrence Ave	St. Louis Ave & Balmoral Ave	Subscriber	Male
45207	2017-01-17 14:53:07	2017-01-17 15:02:01	534	Clark St & Randolph St	Desplaines St & Jackson Blvd	Subscriber	Male
1473887	2017-06-26 09:01:20	2017-06-26 09:11:06	586	Clinton St & Washington Blvd	Canal St & Taylor St	Subscriber	Male

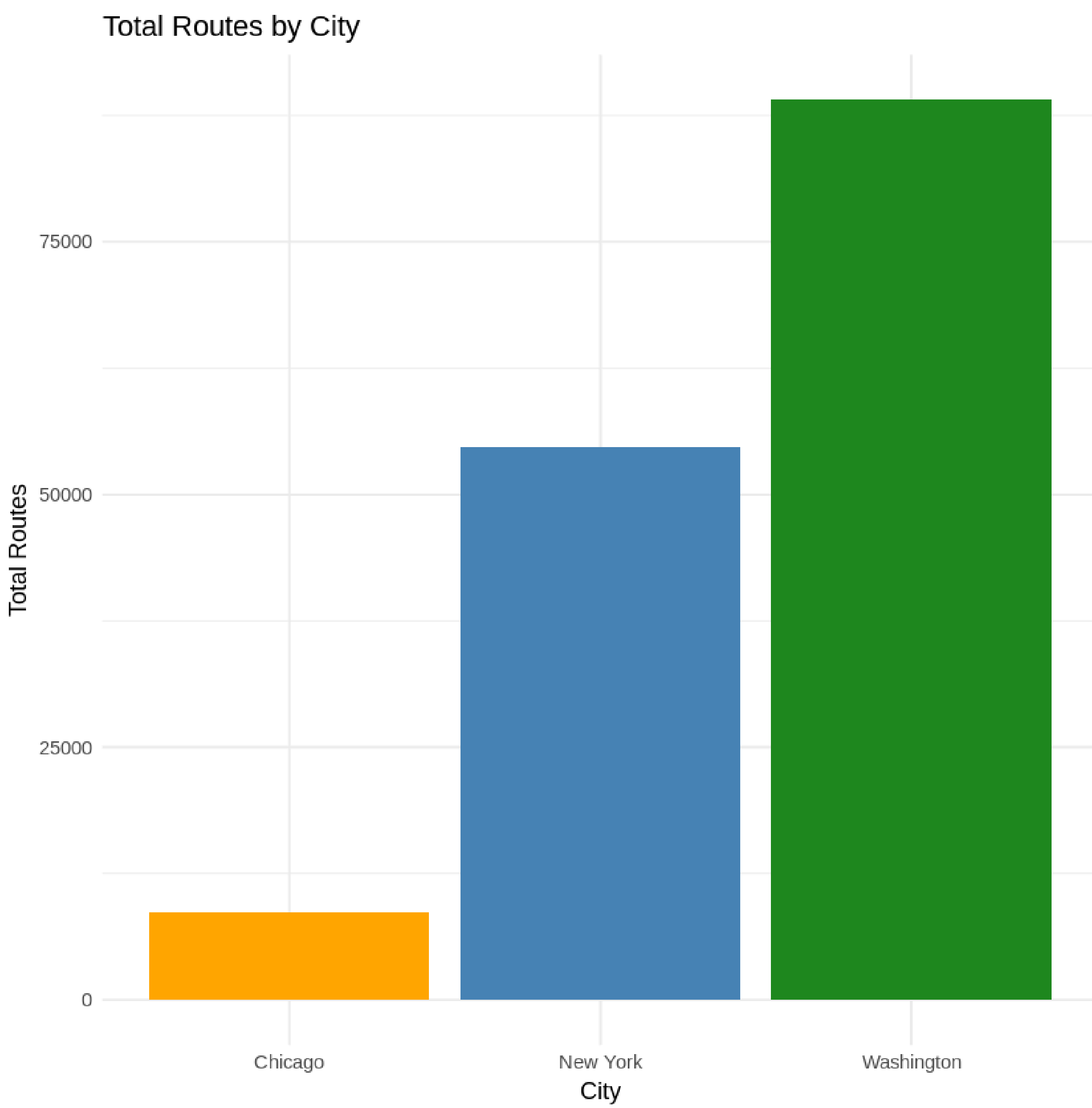
Question 1

To begin our marketing campaign we would like to find where the most routes are located, this will help us decide what location we would like to focus on.

In [62]:
```
# Calculate the total number of routes for each city
ny_total_routes <- nrow(ny)
wash_total_routes <- nrow(wash)
chi_total_routes <- nrow(chi)

# Display the results
cat("Total routes in New York:", ny_total_routes, "\n")
cat("Total routes in Washington:", wash_total_routes, "\n")
cat("Total routes in Chicago:", chi_total_routes, "\n")
```

```
Total routes in New York: 54770
Total routes in Washington: 89051
Total routes in Chicago: 8630
```

In [63]:
```
# Load necessary library
library(ggplot2)

# Create a data frame for the totals
route_totals <- data.frame(
  City = c("New York", "Washington", "Chicago"),
  Total_Routes = c(54770, 89051, 8630)
)

# Create a bar plot to compare total routes
ggplot(route_totals, aes(x = City, y = Total_Routes, fill = City)) +
  geom_bar(stat = "identity") +
  labs(title = "Total Routes by City", x = "City", y = "Total Routes") +
  theme_minimal() +
  theme(legend.position = "none") +
  scale_fill_manual(values = c("New York" = "steelblue", "Washington" = "fores
tgreen", "Chicago" = "orange"))
```

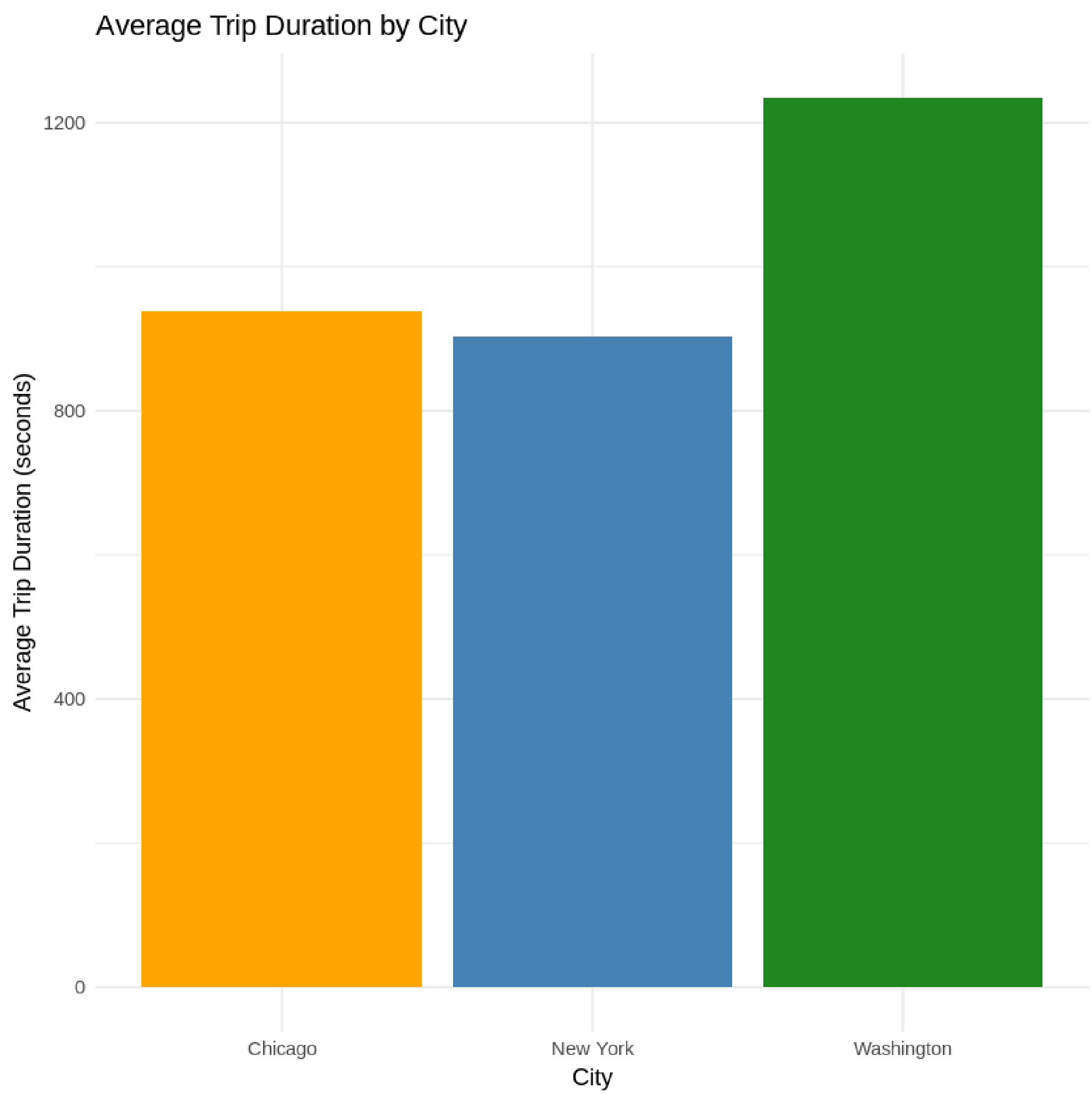


In [64]:

```
# Calculate average trip duration for each city
ny_avg_duration <- mean(ny$Trip.Duration, na.rm = TRUE)
wash_avg_duration <- mean(wash$Trip.Duration, na.rm = TRUE)
chi_avg_duration <- mean(chi$Trip.Duration, na.rm = TRUE)

# Display the results
cat("Average Trip Duration in New York:", ny_avg_duration, "seconds\n")
cat("Average Trip Duration in Washington:", wash_avg_duration, "seconds\n")
cat("Average Trip Duration in Chicago:", chi_avg_duration, "seconds\n")
```

```
Average Trip Duration in New York: 903.6147 seconds
Average Trip Duration in Washington: 1233.953 seconds
Average Trip Duration in Chicago: 937.1728 seconds
```

In [65]:
```
# Create a data frame for average trip durations
trip_duration_data <- data.frame(
  City = c("New York", "Washington", "Chicago"),
  Avg_Trip_Duration = c(ny_avg_duration, wash_avg_duration, chi_avg_duration)
)

# Create a bar plot
ggplot(trip_duration_data, aes(x = City, y = Avg_Trip_Duration, fill = City))
+
  geom_bar(stat = "identity") +
  labs(title = "Average Trip Duration by City", x = "City", y = "Average Trip
Duration (seconds)") +
  theme_minimal() +
  theme(legend.position = "none") +
  scale_fill_manual(values = c("New York" = "steelblue", "Washington" = "fores
tgreen", "Chicago" = "orange"))
```



Summary for Question 1 Part: 1

In our analysis of bikeshare data from New York, Washington, and Chicago, we aimed to identify the most promising location for a marketing campaign by examining both the total number of routes and the average trip duration in each city. The findings revealed that Washington has the highest total number of routes at 89,051, followed by New York with 54,770 routes, and Chicago with only 8,630 routes. However, Washington also leads with an average trip duration of 1,233.95 seconds, while Chicago has a slightly longer average trip duration of 937.17 seconds compared to New York's 903.61 seconds.

After realizing that Chicago had longer trip durations, it made me question if Chicago would be more profitable. To explore this, I calculated the estimated revenue by multiplying the total number of trips by the average duration and the price charged per hour. One last test before we make the final decision on the location to pursue.....

In [66]:
```
# Set the price charged per hour
price_per_hour <- 2.50  # Example price per hour

# Convert average trip durations from seconds to hours
ny_avg_duration_hours <- ny_avg_duration / 3600
wash_avg_duration_hours <- wash_avg_duration / 3600
chi_avg_duration_hours <- chi_avg_duration / 3600

# Calculate estimated revenue for each city
ny_revenue <- ny_avg_duration_hours * 54770 * price_per_hour
wash_revenue <- wash_avg_duration_hours * 89051 * price_per_hour
chi_revenue <- chi_avg_duration_hours * 8630 * price_per_hour

# Display the results
cat("Estimated Revenue in New York: $", round(ny_revenue, 2), "\n")
cat("Estimated Revenue in Washington: $", round(wash_revenue, 2), "\n")
cat("Estimated Revenue in Chicago: $", round(chi_revenue, 2), "\n")
```

```
Estimated Revenue in New York: $ 34368.73
Estimated Revenue in Washington: $ 76308.87
Estimated Revenue in Chicago: $ 5616.53
```

Summary for Question 1: Part 2

The results for the estimated revenues showed Washington still being the top selection at $76,309.87 followed by 34,368.73 for New York, and only 5,616.53 for Chicago. This puts Washington ranking higher in all areas compared to the other two locations.

However, after reviewing these calculations Washington does have higher profitability followed by New York I am also taking into consideration that there is additional demographic data (gender and birth year) available for New York further which strengthens its position as the optimal focus area for our marketing campaign. Therefore, New York emerges as the best choice, balancing route availability, average trip duration, and valuable customer insights for effective marketing targeting.

Question 2

What should our target demographics be for our marketing campaign based on gender and birth year from the New York bikeshare data?

In [67]:
```
# Count the number of rentals by Gender
gender_counts <- ny %>%
  group_by(Gender) %>%
  summarise(Total_Rentals = n())

# Display the results for Gender
cat("Total Rentals by Gender in New York:\n")
print(gender_counts)
```

```
Total Rentals by Gender in New York:
# A tibble: 3 x 2
  Gender Total_Rentals
  <fct>          <int>
1 ""              5410
2 Female         12159
3 Male           37201
```

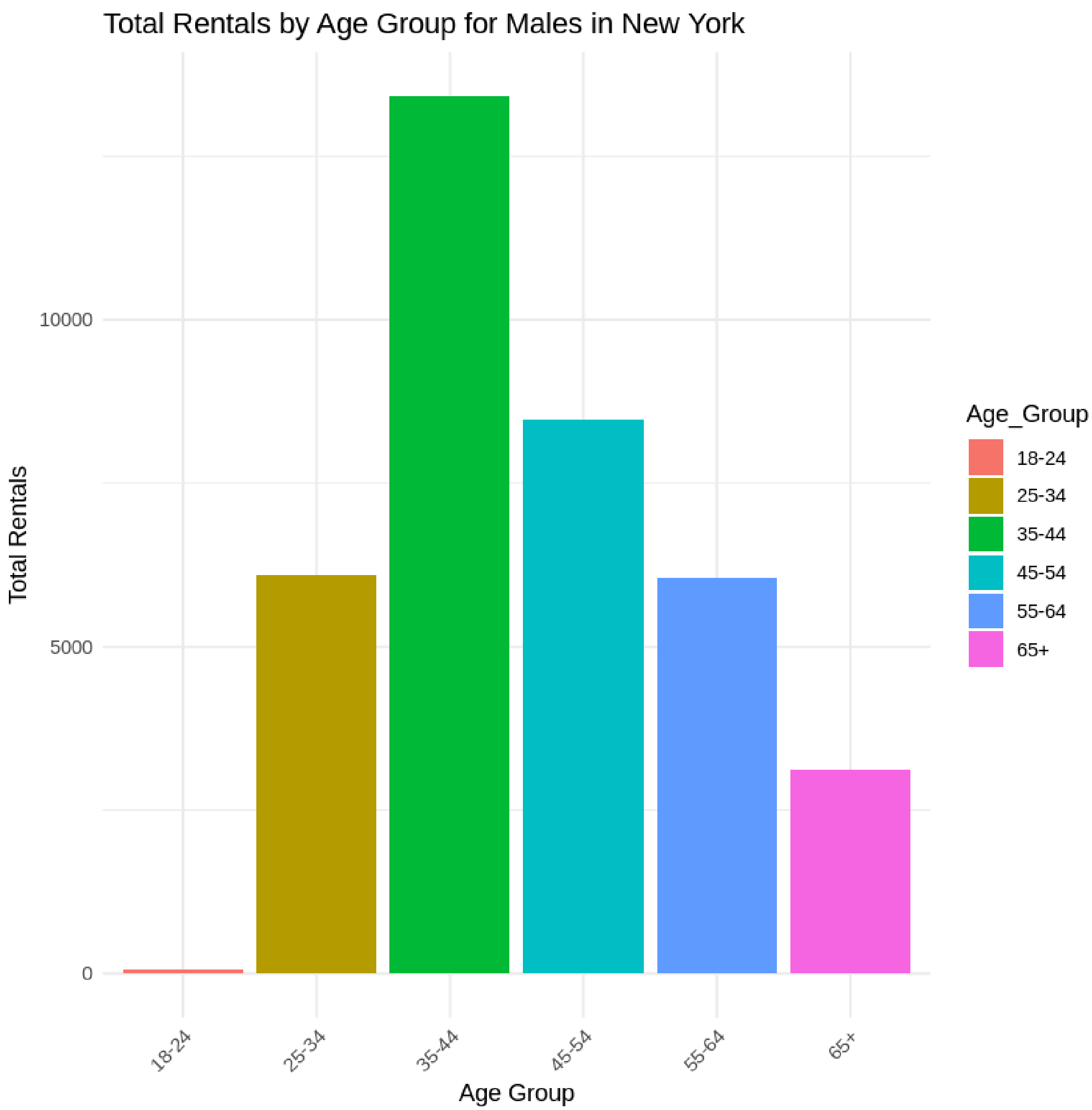
In [68]:
```
# Current year for age calculation
current_year <- 2024

# Create a new column for Age
ny <- ny %>%
  mutate(Age = current_year - Birth.Year)

# Calculate average age of renters by Gender
average_age_by_gender <- ny %>%
  group_by(Gender) %>%
  summarise(Average_Age = mean(Age, na.rm = TRUE))

# Display the average age results
cat("\nAverage Age of Renters by Gender in New York:\n")
print(average_age_by_gender)
```

```
Average Age of Renters by Gender in New York:
# A tibble: 3 x 2
  Gender Average_Age
  <fct>        <dbl>
1 ""            47.5
2 Female        44.9
3 Male          46.1
```

In [33]:
```
# Filter for male renters only
ny_men <- ny %>% filter(Gender == "Male")

# Create age groups for male renters
ny_men <- ny_men %>%
  mutate(Age_Group = case_when(
    Age < 18 ~ "Under 18",
    Age >= 18 & Age < 25 ~ "18-24",
    Age >= 25 & Age < 35 ~ "25-34",
    Age >= 35 & Age < 45 ~ "35-44",
    Age >= 45 & Age < 55 ~ "45-54",
    Age >= 55 & Age < 65 ~ "55-64",
    Age >= 65 ~ "65+",
    TRUE ~ "Unknown"
  ))

# Count the number of rentals by Age Group for males
age_group_counts_men <- ny_men %>%
  group_by(Age_Group) %>%
  summarise(Total_Rentals = n()) %>%
  arrange(desc(Total_Rentals))

# Display the results for Male Age Groups
cat("\nTotal Rentals by Age Group for Males in New York:\n")
print(age_group_counts_men)
```

```
Total Rentals by Age Group for Males in New York:
# A tibble: 6 x 2
  Age_Group Total_Rentals
  <chr>             <int>
1 35-44             13415
2 45-54              8464
3 25-34              6100
4 55-64              6048
5 65+                3113
6 18-24                61
```

In [69]:
```
# Create a bar chart for total rentals by age group
ggplot(age_group_counts_men, aes(x = Age_Group, y = Total_Rentals, fill = Age_
Group)) +
  geom_bar(stat = "identity") +
  labs(title = "Total Rentals by Age Group for Males in New York",
       x = "Age Group",
       y = "Total Rentals") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) # Adjust x-axis tex
t for better readability
```



Summary for question 2:

In our analysis of the bikeshare data from New York, we aimed to identify key demographic insights to guide our marketing campaign. We first examined rental patterns by gender and discovered that male renters significantly outnumbered female renters, with a total of 37,201 rentals by men compared to 12,159 by women. This disparity highlights a potential focus area for our marketing efforts.

Next, we explored the age distribution of male renters to determine which age groups contributed most to the rental trends. The results indicated that the average age of renters falls within a diverse range, but a more detailed look at male renters revealed that the 35-44 age group was the most active demographic, accounting for 13,415 rentals. This suggests that marketing initiatives targeting men aged 35-44 may yield the highest return on investment.

In summary, the data indicates that our marketing campaign should primarily target male renters, particularly those aged 35-44, to maximize engagement and profitability.

Question 3

What month should we launch the marketing campaign?

When deciding whether to market in a month with already high profits or focus on increasing sales in slower months we need to consider the following factors.

- Market Saturation vs. Groth Potential: Marketing during periods of high profits may capatalize on existing demand but the market may be saturated making it harder to stand out and attract new customers. Concentrating efforts on slower months can stimulate demand and encurage consistent usage throughout the year.

-Customer Behavior:- If certain months show a consistent drop in rentals, we may need to identify the reasons and tailor marketing campaigns to address those issues. Consider seasonal promotions or discounts to attract customers during slower months. Offering incentives can drive traffic and increase awareness.

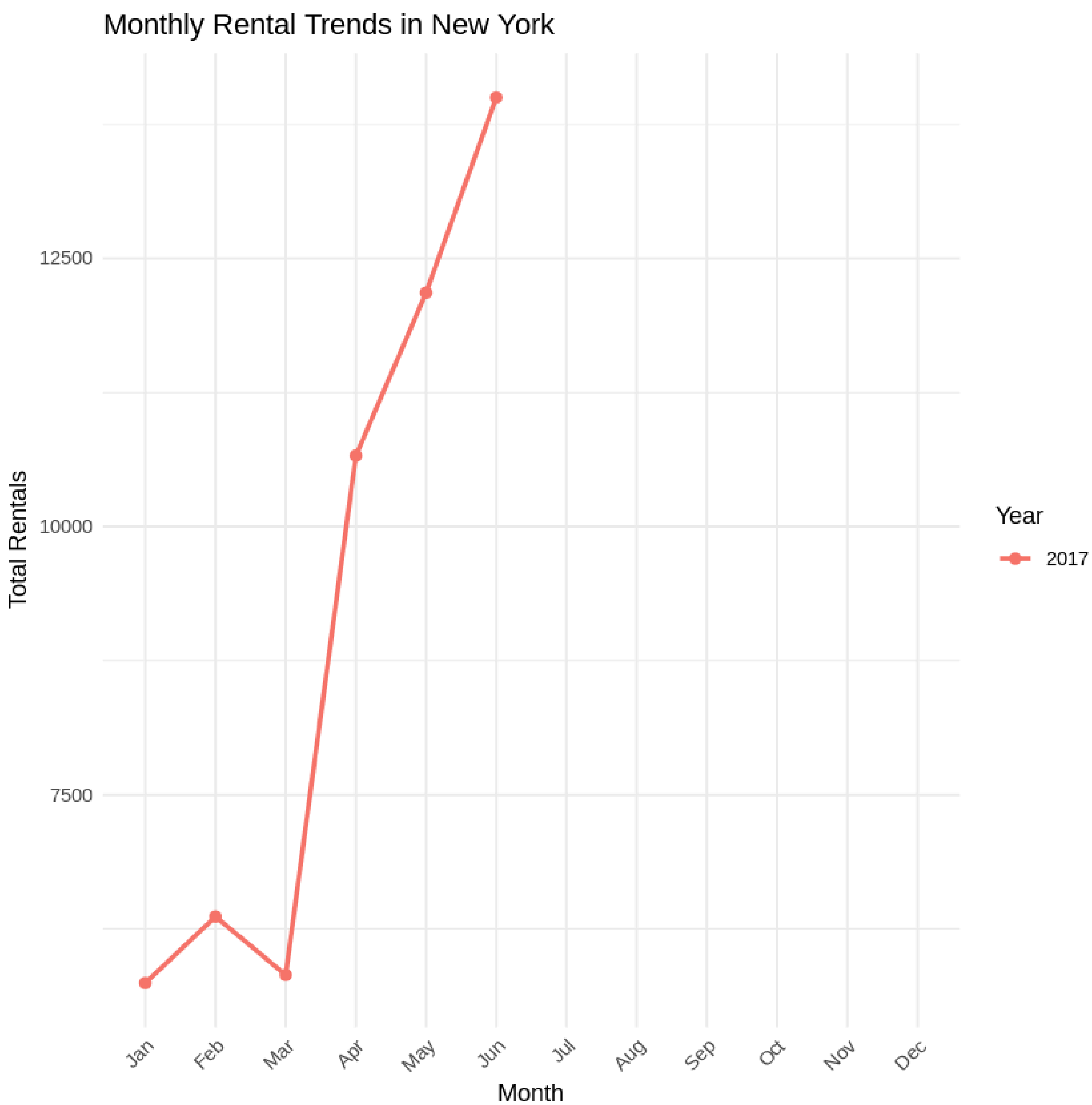
-Long-term vs. Short-term Stragegy: Long term growth may focus on increasing sales in slower months can contribute to sustained growth, creating a stronger brand presence and customer base over time. With short-term Gains marketing during high-profit months might provide immediate financial benefits, but it may not lead to long-term customer retention.

Ultimately a balanced approach may be the most effective. Running targeted campaigns during high-profit months to maximize short-term gains while simultaneously implementing strategies to boost awareness and rentals during slower months.

In [70]:
```
# Convert Start.Time to Date-Time format
ny$Start.Time <- as.POSIXct(ny$Start.Time)

# Find month and year
ny <- ny %>%
  mutate(Month = format(Start.Time, "%b"),  # Abbreviated month name
         Year = format(Start.Time, "%Y"))   # Full year

# Count the number of rentals by month and year
monthly_rentals <- ny %>%
  group_by(Year, Month) %>%
  summarise(Total_Rentals = n()) %>%
  ungroup()

# Display the results
print(monthly_rentals)
```

```
# A tibble: 6 x 3
  Year  Month Total_Rentals
  <chr> <chr>         <int>
1 2017  Apr           10661
2 2017  Feb            6364
3 2017  Jan            5745
4 2017  Jun           14000
5 2017  Mar            5820
6 2017  May           12180
```

In [71]:
```
# Plot monthly rentals with a line graph
ggplot(monthly_rentals, aes(x = Month, y = Total_Rentals, group = Year, color
= Year)) +
  geom_line(size = 1) +  # Line thickness
  geom_point(size = 2) + # Points on the line for each month
  labs(title = "Monthly Rental Trends in New York",
       x = "Month",
       y = "Total Rentals") +
  theme_minimal() +
  scale_x_discrete(limits = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul",
"Aug", "Sep", "Oct", "Nov", "Dec")) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis la
bels for better readability
```



Summary of question 3:

In our analysis, we first converted the rental data to include a clear date format and extracted each record’s month and year to observe monthly rental trends in New York. We then calculated total rentals per month across our dataset, revealing six months' worth of data with peak rentals occurring in June and May. To visualize these trends, we created a line chart showing fluctuations in monthly rentals. However, given the limited timeframe, it may be premature to launch a targeted marketing campaign based solely on these results. Collecting a full year's data would provide a more complete seasonal picture and allow for a more informed marketing strategy.

Final Thoughs

After an in-depth analysis of the bikeshare data, we identified New York as the most promising location for maximizing profitability and refined our target demographic to be males aged 35-44, allowing us to tailor our marketing strategy more effectively. While we have a solid foundation for the location and audience, we recommend collecting at least an additional six months of data to capture a full year of seasonal trends. This expanded dataset will provide better insight into the most opportune month(s) for launching a comprehensive marketing campaign, ensuring an optimal return on investment. In the interim, a smaller promotional push in March could help boost rentals during a historically slower month, positioning the brand for sustained growth while final data is collected for a broader, data-driven campaign launch.

Finishing Up

> Congratulations! You have reached the end of the Explore Bikeshare Data Project. You should be very proud of all you have accomplished!
>
> **Tip**: Once you are satisfied with your work here, check over your report to make sure that it is satisfies all the areas of the rubric (https://review.udacity.com/#!/rubrics/2508/view).

Directions to Submit

> Before you submit your project, you need to create a .html or .pdf version of this notebook in the workspace here. To do that, run the code cell below. If it worked correctly, you should get a return code of 0, and you should see the generated .html file in the workspace directory (click on the orange Jupyter icon in the upper left).
>
> Alternatively, you can download this report as .html via the **File** > **Download as** submenu, and then manually upload it into the workspace directory by clicking on the orange Jupyter icon in the upper left, then using the Upload button.
>
> Once you've done this, you can submit your project by clicking on the "Submit Project" button in the lower right here. This will create and submit a zip file with this .ipynb doc and the .html or .pdf version you created. Congratulations!

In [ ]:
```
system('python -m nbconvert Explore_bikeshare_data.ipynb')
```