```r
library("tidyverse")
library("here")
library("janitor")
library("ggplot2")
library("geosphere")
```

Introduction

Capital Bikesshare is a company that provides bike sharing services and is located in Washington DC By analysing their data we could uncover trends that could be beneficial to the company to improve their efficiency provide high bike usage among locals at peak hours and encourage environment for sustainable transportation solutions

Data Cleaning

The data consist of duration, types of members,total trips covered, station name and the distance covered given in the form of longitude and latitude as well as type of bikes available at Capital Bikes Share.

```r
# Loading rides data and making the necessary changes

bike_data <- read_csv(here("data", "rides_2020_2021_extract.csv"), na = "N/A") %>%


  # Renaming columns to make it more readable


  rename(total_time = Duration) %>%
  rename(start_date = `Start date`) %>%
  rename(end_date = `End date`) %>%
  rename(start_station_name = `Start station name`) %>%
  rename(end_station_name = `End station name`) %>%

  # Convert the data type to an integer

  mutate(total_time = as.integer(total_time)) %>%

  # Convert date columns to POSIXct

  mutate(start_date = as.POSIXct(start_date, origin = "1970-01-01", tz = "UTC")) %>%
  mutate(end_date = as.POSIXct(end_date, origin = "1970-01-01", tz = "UTC")) %>%
  mutate(member_casual = tolower(member_casual)) %>%

  # Conveeting longitiude and lanfitr to double data type

  mutate(start_lng = as.double(start_lng)) %>%
  mutate(start_lat = as.double(start_lat)) %>%
  mutate(end_lng = as.double(end_lng)) %>%
  mutate(end_lat = as.double(end_lat)) %>%

  # Clean column is used to remove any special character

  clean_names()



print(bike_data)
```

```
## # A tibble: 1,986,564 × 16
##    total_time start_date          end_date            start_station_id
##         <int> <dttm>              <dttm>              <chr>
##  1        411 2020-03-20 13:33:17 2020-03-20 13:40:08 31623
##  2        753 2021-08-29 11:48:52 2021-08-29 12:01:25 31320
##  3        382 2021-12-09 16:57:40 2021-12-09 17:04:02 31272
##  4        698 2020-02-25 08:10:41 2020-02-25 08:22:19 31113
##  5        375 2021-11-20 11:23:50 2021-11-20 11:30:05 31626
##  6        503 2021-07-31 09:51:08 2021-07-31 09:59:31 31212
##  7        581 2021-03-30 11:47:20 2021-03-30 11:57:01 31126
##  8        188 2020-10-02 16:07:31 2020-10-02 16:10:39 31125
##  9       2681 2021-10-10 11:55:35 2021-10-10 12:40:16 31261
## 10        796 2021-07-15 18:36:53 2021-07-15 18:50:09 31276
## # ℹ 1,986,554 more rows
## # ℹ 12 more variables: start_station_name <chr>, end_station_id <chr>,
## #   end_station_name <chr>, bike_number <chr>, member_casual <chr>,
## #   ride_id <chr>, rideable_type <chr>, start_lat <dbl>, start_lng <dbl>,
## #   end_lat <dbl>, end_lng <dbl>, is_equity <chr>
```

Question 1

What are the busy hours among casual and member riders?

To operationalize this question, we defined a variable and converted the start date and time to hours. We then get the hours of start time for each ride, which makes it easy to categorize the data.

Question 2

What is the most preferred bike for users traveling long distances?

To operationalize the question to find the most popular bike to be used for a long journey we first calculated the distance of each ride. We made use of a distHaversine function that converts longitude and latitude to meters later dividing it by 1000 to get the measurements in kilometers. We calculate the mean distance for each bike. This gives us a high probability of which type of bike is used for longer routes.

Question 3

What percentage of rides are taken during weekends compared to weekdays?

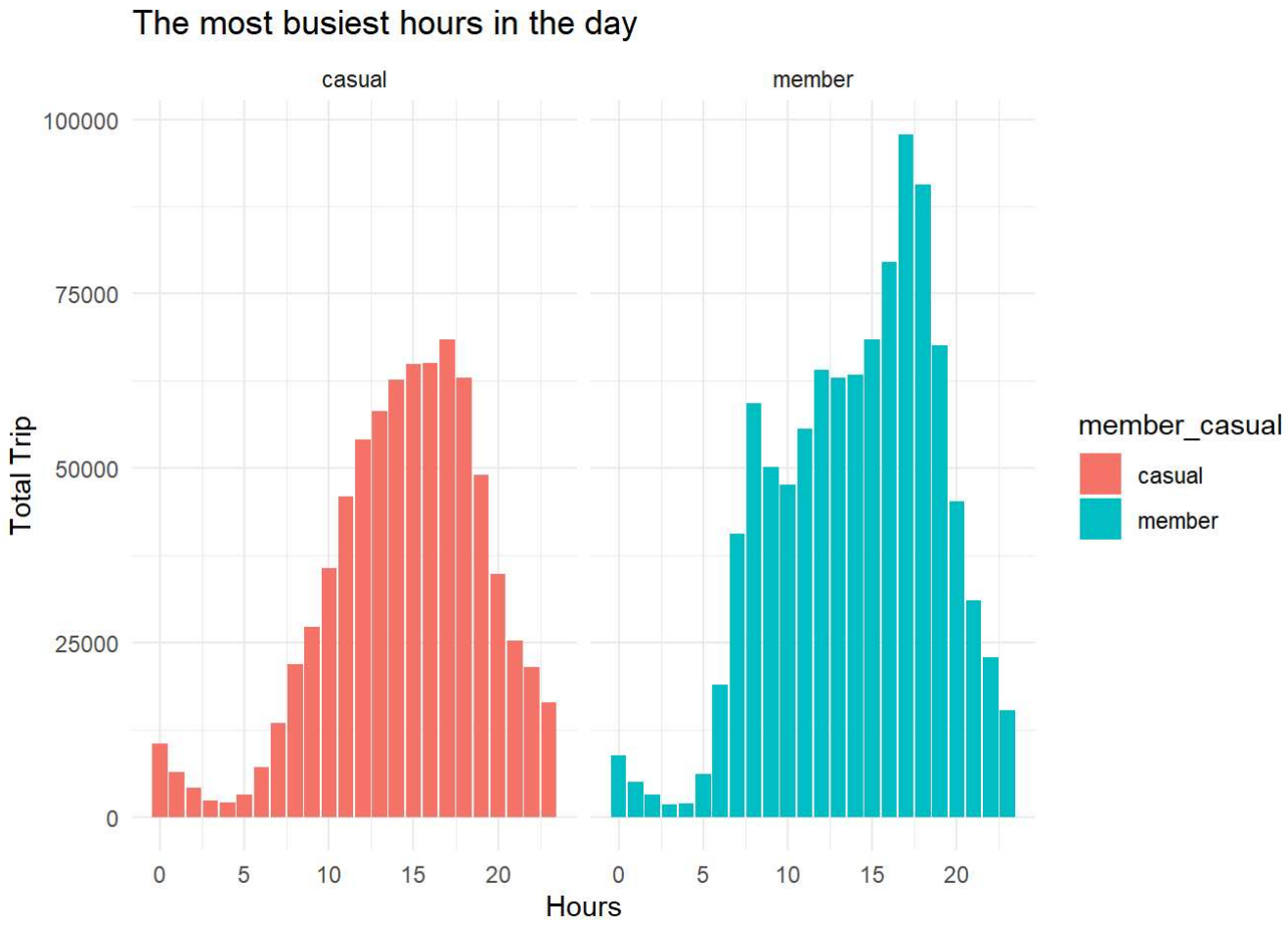
To operationalize this question we made two functions wday() to calculate the day of the week for each start date based on each bike trip. Next, we utilize another function called ifelse to categorise if it’s a weekday or a weekend And calculate the percentage of total rides.

Answer to Question 1

The research determines the peak hours for bike utilisation among casual and member riders. The data shows that the busiest hours are 5 PM (18th hour) for casual users and 6 PM (17th hour) for member users. This implies a larger number of bike rentals among members during these peak periods, implying that many members utilise the service to commute in the evenings.

The chart below depicts the overall number of rides each hour for both casual and member users, showing these patterns.

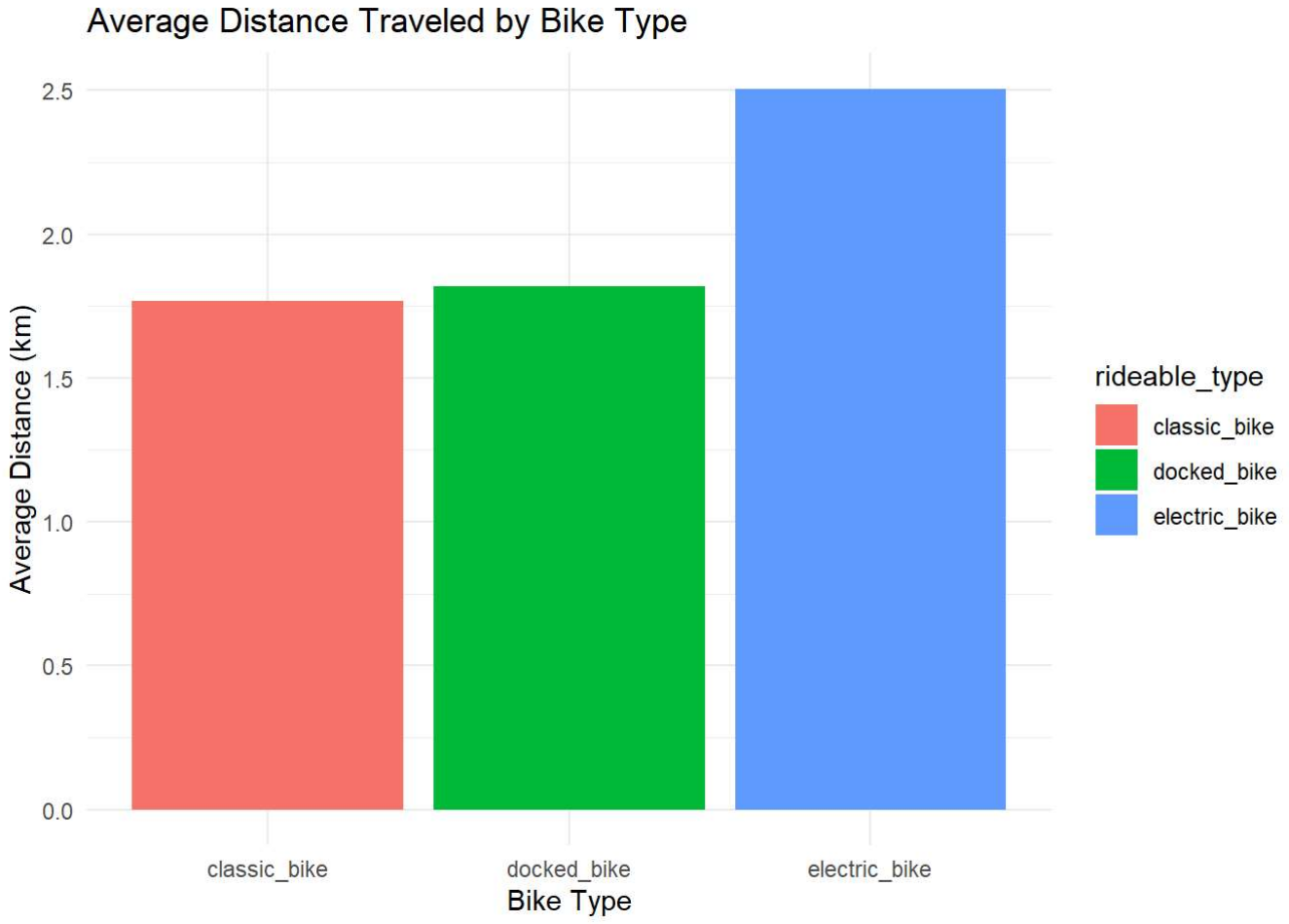
```
# Calculate total trips in hours

total_trips <- bike_data %>%
  # Grouping the necessary data

  mutate(start_hours = hour(start_date)) %>%
  group_by(member_casual, ride_id, start_hours) %>%
  # Summarize the data

  summarize(total_rides = n_distinct(ride_id), .groups = 'drop')

#Create a graph by using ggplot libraries =


ggplot() +
  geom_col(
    data = total_trips,
    mapping = aes(x = start_hours , y = total_rides , fill = member_casual)
  ) +
  facet_wrap(~ member_casual) +
  labs(title = "The most busiest hours in the day", x = "Hours", y = "Total Trip") +
  theme_minimal()
```



Answer to Question 2

As per the analysis, users prefer electric bikes for long distances as it is way more convenient for them to travel. By calculating the average distance traveled, we found that electric bikes are the preferred choice.

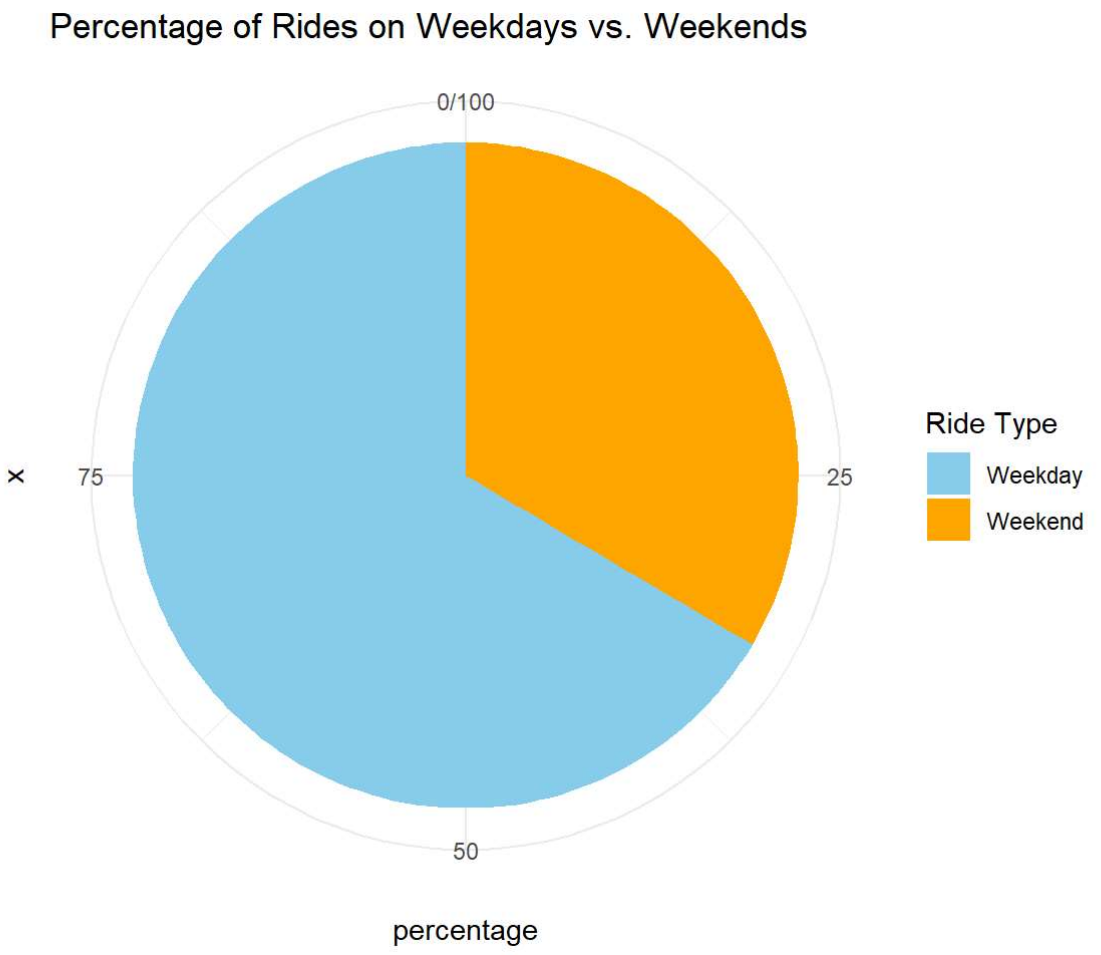
```
bike_distance <- bike_data %>%
  #To remove data that is emply or NA

  filter(!is.na(start_lng) &
           !is.na(start_lat) & !is.na(end_lng) &
           !is.na(end_lat)) %>%

  # Converting longitude and latitude  to kilometre by usimg the Haversine formula

  mutate(distance_km = distHaversine(matrix(c(
    start_lng , start_lat
  ), ncol = 2), matrix(c(end_lng , end_lat), ncol = 2)) / 1000) %>%
  group_by(rideable_type) %>%
  summarise(
    distance_km = mean(distance_km),
    na.rm = TRUE,
    .groups = "drop"
  )

ggplot() +
  geom_col(data = bike_distance,
           aes(x = rideable_type, y = distance_km, fill = rideable_type)) +
  labs(title = "Average Distance Traveled by Bike Type", x = "Bike Type",
       y = "Average Distance (km)") +
  theme_minimal()
```



Answer to question 3

As per the data we have customers riding the bike on weekdays way more often compared to weekends. This also implies that most users who use the service on weekdays are most probably commuting to the office. As the below chat illustrates 75% of customers ride on weekdays and the rest ride on weekends.

```
ride_categories <- bike_data %>%

  # Extracting data from start date

  mutate(day_of_week = wday(start_date, label = TRUE, abbr = FALSE)) %>%

  # To differentiate between Weekdays and Weekend

  mutate(ride_type = ifelse(day_of_week %in% c("Saturday", "Sunday"), "Weekend", "Weekday"))
%>%
  group_by(ride_type) %>%
  summarize(total_rides = n(), .groups = 'drop')

#Calculating the percentage

ride_categories <- ride_categories %>%
  mutate(percentage = (total_rides / sum(total_rides)) * 100)

ggplot(ride_categories, aes(x = "", y = percentage, fill = ride_type)) +
  geom_bar(stat = "identity", width = 1) + # Use a bar chart to create the pie chart
  coord_polar(theta = "y") +  # Transform to pie chart
  labs(title = "Percentage of Rides on Weekdays vs. Weekends", fill = "Ride Type") +
  scale_fill_manual(values = c("Weekday" = "skyblue", "Weekend" = "orange")) +
  theme_minimal()
```



Conclusion

The main conclusions on user behaviour and usage trends are summed up in this study of the Capital Bikeshare data. According to the data, the busiest hours for renting bikes are from 5 to 6 p.m., when more members than non-members ride. This implies that a large number of members commute using the service. Longer-distance riders have preferences for electric bike models. Additionally, The comparison of weekday and weekend travels clearly demonstrates distinct usage trends, indicating specific areas for targeted advertising and service enhancements.

The results prompt further questions into how external factors, such as local events or weather, affect bike usage.

Use of Generative AI

I acknowledge the use of Chat GPT and https://chatgpt.com/ (https://chatgpt.com/) to generate materials for background research and independent study and/or that I have adapted to include within the work submitted for assessment. I confirm that all use of AI content is acknowledged and referenced appropriately.

The following prompts were input into :can you calculate the distance from latitude and longitude.

The output obtained was: Yes, you can calculate the distance between two points given their latitude and longitude using the Haversine formula. Example usage: Calculate the distance between two locations # Location 1: Latitude and Longitude of New York City lat1 <- 40.7128 lon1 <- -74.0060 # Location 2: Latitude and Longitude of London lat2 <- 51.5074 lon2 <- -0.1278 # Calculate the distance distance <- haversine_distance(lat1, lon1, lat2, lon2)

Full detail of how the output was adapted : I did look at the example and tried to code i did get few errors put it into chagtgpt and it told me to install library(“geosphere”) package. And after few errors i got the desired output