Project for Ciclistics - Analysis

Igor Vysochanskyy

2024-04-30

This R Markdown analyzes differences between members & casual bike users.

Install R packages & their libraries to enable subsequent operations.

Depending on your RStudio version, packages may be pre-installed, or you might need to install them manually or via a clickable link that appears above the script

```
library("tidyverse")
```

```
library("skimr")
```

Upload 12 datasets for each month of 2023

```
for (i in 301:312) {
  load(paste0("RData_CleanBike\\CleanBike", i, ".RData"))
}
```

Bind 12 months dataset into a single data frame

```
data2023 <- bind_rows(data301, data302, data303, data304, data305, data306, data307, data308, da
ta309, data310, data311, data312)
```

Detailed combined dataset observation

```
skim_without_charts(data2023)
```

Data summary

Name	data2023
Number of rows	5380725
Number of columns	6
_____	
Column type frequency:	
character	4
numeric	1
POSIXct	1

Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
ride_id	0	1	16	16	0	5380725	0
rideable_type	0	1	12	13	0	2	0
member_casual	0	1	6	6	0	2	0
day_of_week	0	1	3	3	0	7	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
ride_length	0	1	14.83	20.24	2	5.9	9.9	17	719.4

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
started_at	0	1	2023-01-01 00:02:06	2023-12-31 23:58:55	2023-07-21 16:24:33	4082465

Plot 1. Average trip duration for members vs casuals combined

Group by month, member_casual, rideable_type & calc. average trip duration

```
avg_dur1 <- data2023 %>%
  mutate(month = format(started_at, "%B")) %>%  # Extract month from started_at column
  group_by(month, member_casual) %>%
  summarise(avg_dur1 = mean(ride_length), .groups = 'drop')
```

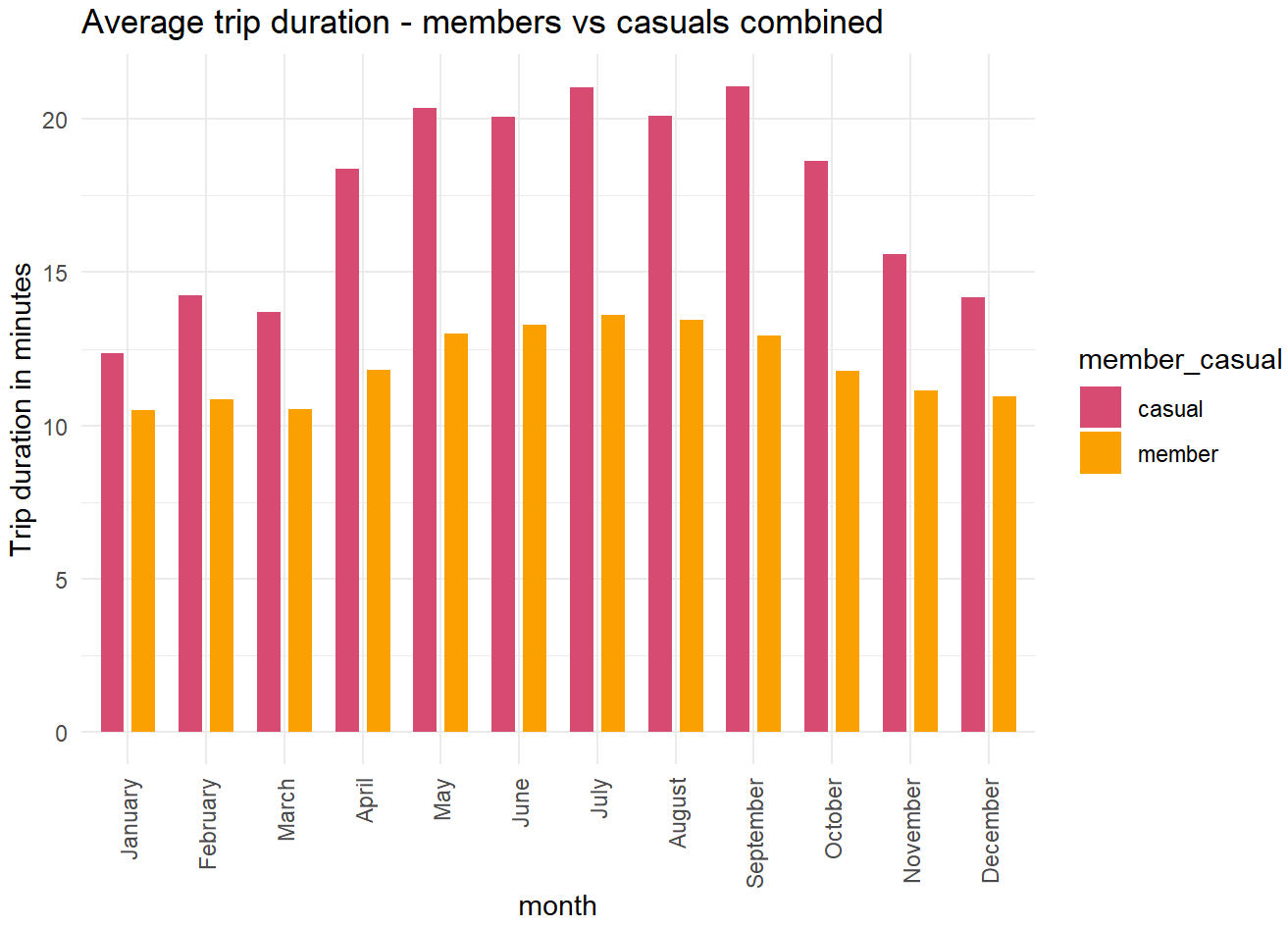
Define colors for each value of member_casual

```
color_palette1 <- c("#D74B76", "#FAA300")
```

Convert ‘month’ to factor with custom levels in the desired order

```
avg_dur1$month <- factor(avg_dur1$month, levels = c(
  "January", "February", "March", "April", "May", "June",
  "July", "August", "September", "October", "November", "December"
))
```

```
ggplot(avg_dur1, aes(x = month, y = avg_dur1, fill = member_casual)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8), width = 0.6) +
  scale_fill_manual(values = color_palette1) +
  labs(y = "Trip duration in minutes", title = "Average trip duration - members vs casuals combi
ned") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Plot2. Average trip duration & all types of bike users.

Group by the combination of all values from member_casual and rideable_type

```
avg_dur2 <- data2023 %>%
  mutate(month = format(started_at, "%B")) %>%
  group_by(month, type_of_users = paste(member_casual, rideable_type)) %>%
  summarise(avg_dur2 = mean(ride_length), .groups = 'drop')
```

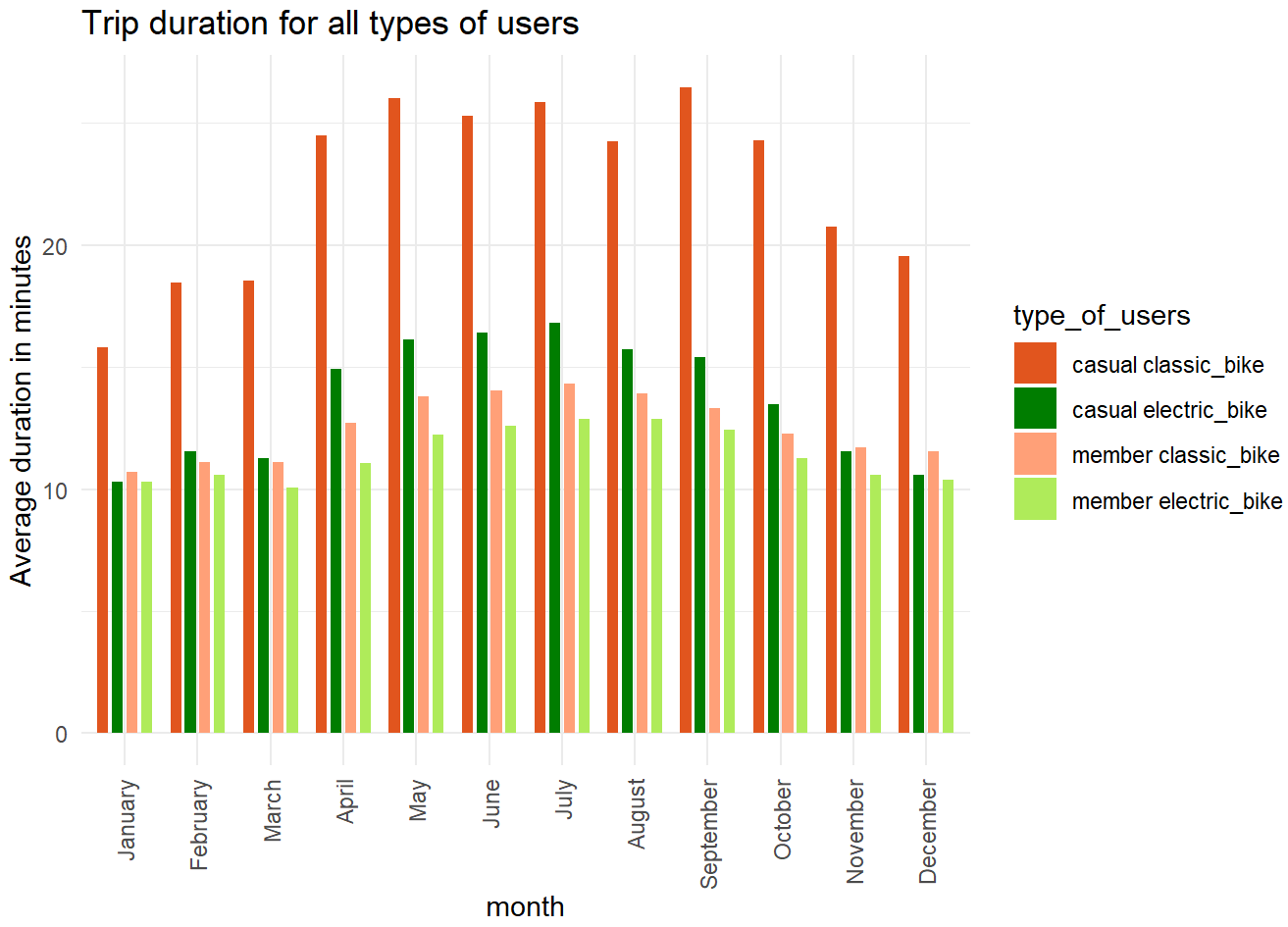
Define colors for each combination of member_casual and rideable_type

```
color_palette2 <- c("#e25822", "#008000", "#ffa07a", "#b2ec5d")
```

Convert ‘month’ to factor with custom levels in the desired order

```
avg_dur2$month <- factor(avg_dur2$month, levels = c(
  "January", "February", "March", "April", "May", "June",
  "July", "August", "September", "October", "November", "December"))
```

```
ggplot(avg_dur2, aes(x = month, y = avg_dur2, fill = type_of_users)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8), width = 0.6) +
  scale_fill_manual(values = color_palette2) +
  labs(y = "Average duration in minutes", title = "Trip duration for all types of users") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Plot3. Average trip duration for classic_bike between members & casuals.

Group by the combination of classic_bike with member_casual

```
avg_dur3 <- data2023 %>%
  mutate(month = format(started_at, "%B"),
         classic_bike_users = case_when(
           member_casual == "member" & rideable_type == "classic_bike" ~ "member_classic",
           member_casual == "casual" & rideable_type == "classic_bike" ~ "casual_classic")) %>%
  group_by(month, classic_bike_users) %>%
  summarise(avg_dur3 = mean(ride_length), .groups = 'drop') %>%
  filter(!is.na(classic_bike_users))
```

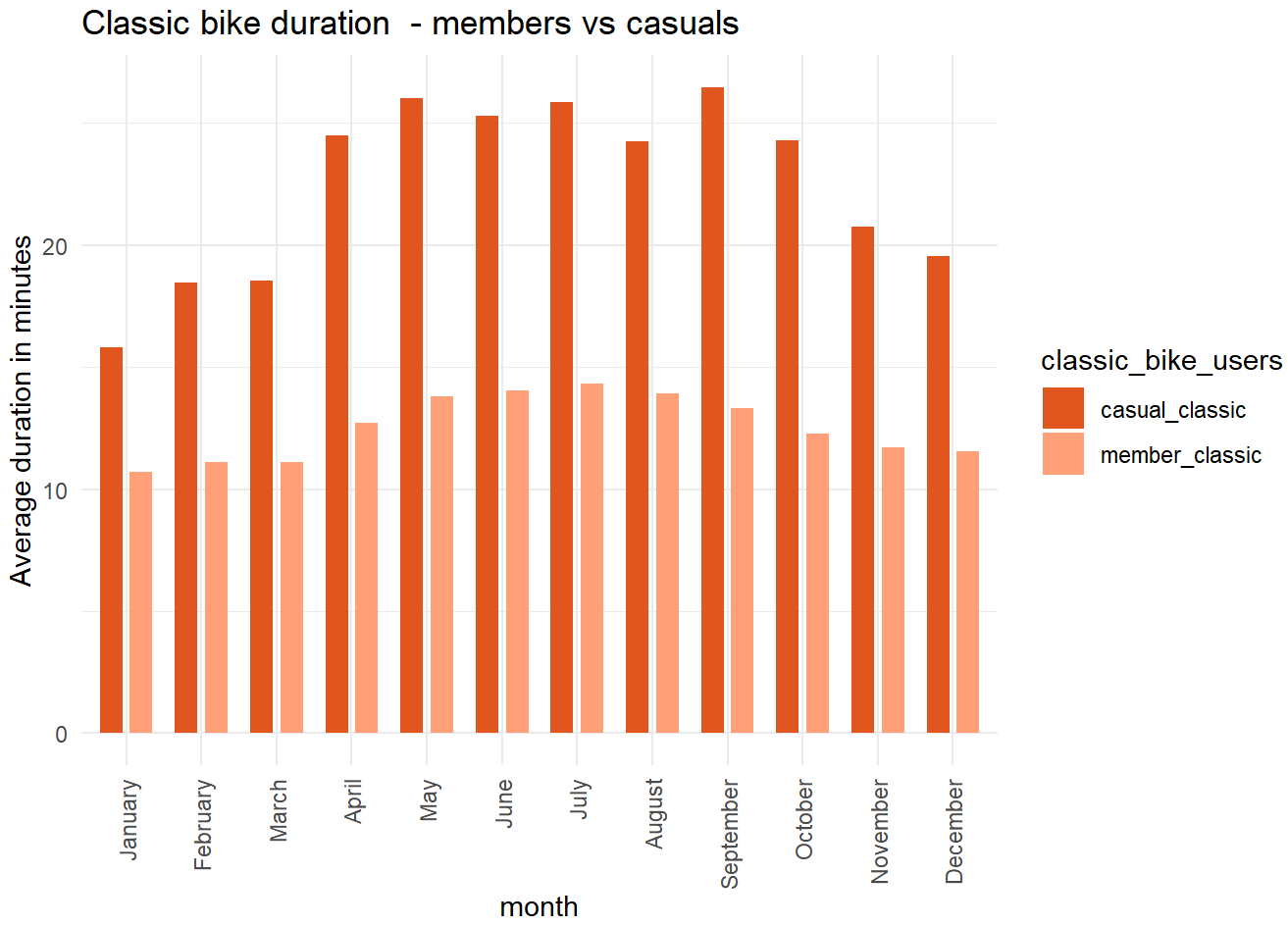
Define color palette

```
color_palette3 <- c("#e25822", "#ffa07a")
```

Convert ‘month’ to factor with custom levels in the desired order

```
avg_dur3$month <- factor(avg_dur3$month, levels = c(
  "January", "February", "March", "April", "May", "June",
  "July", "August", "September", "October", "November", "December"))
```

```
ggplot(avg_dur3, aes(x = month, y = avg_dur3, fill = classic_bike_users)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8), width = 0.6) +
  scale_fill_manual(values = color_palette3) +
  labs(y = "Average duration in minutes", title = "Classic bike duration  - members vs casuals")
+
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Plot4. Average trip duration for electric_bike between members & casuals.

Group by the combination of electric_bike with member_casual

```
avg_dur4 <- data2023 %>%
  mutate(month = format(started_at, "%B"),
         electric_bike_users = case_when(
           member_casual == "member" & rideable_type == "electric_bike" ~ "member_electric",
           member_casual == "casual" & rideable_type == "electric_bike" ~ "casual_electric")) %
>%
  group_by(month, electric_bike_users) %>%
  summarise(avg_dur4 = mean(ride_length), .groups = 'drop') %>%
  filter(!is.na(electric_bike_users))
```

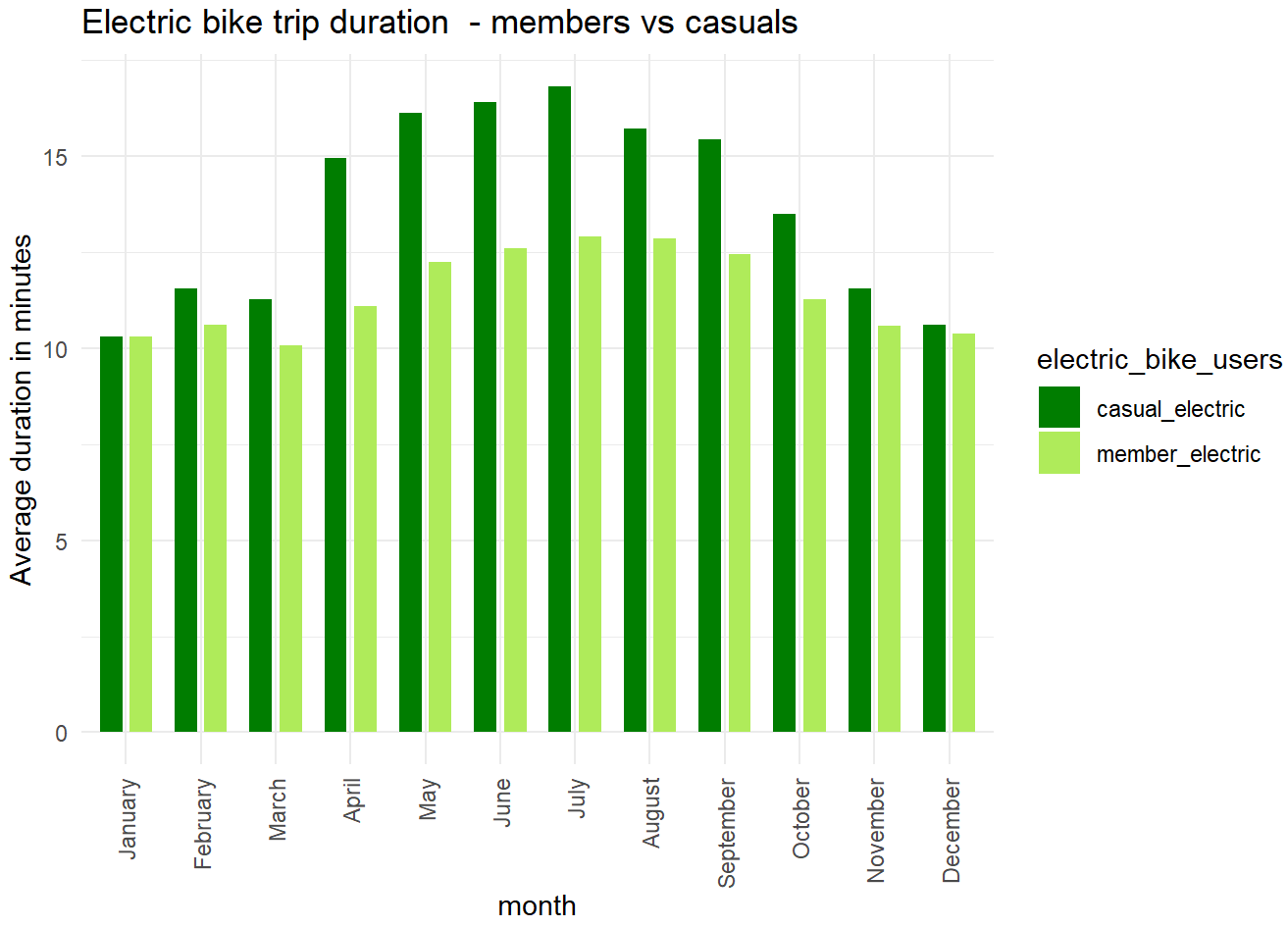
Define color palette

```
color_palette4 <- c("#008000", "#b2ec5d")
```

Convert ‘month’ to factor with custom levels in the desired order

```
avg_dur4$month <- factor(avg_dur4$month, levels = c(
  "January", "February", "March", "April", "May", "June",
  "July", "August", "September", "October", "November", "December"))
```

```
ggplot(avg_dur4, aes(x = month, y = avg_dur4, fill = electric_bike_users)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8), width = 0.6) +
  scale_fill_manual(values = color_palette4) +
  labs(y = "Average duration in minutes", title = "Electric bike trip duration  - members vs cas
uals") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Plot5. Trip count for classic_bike between members & casuals.

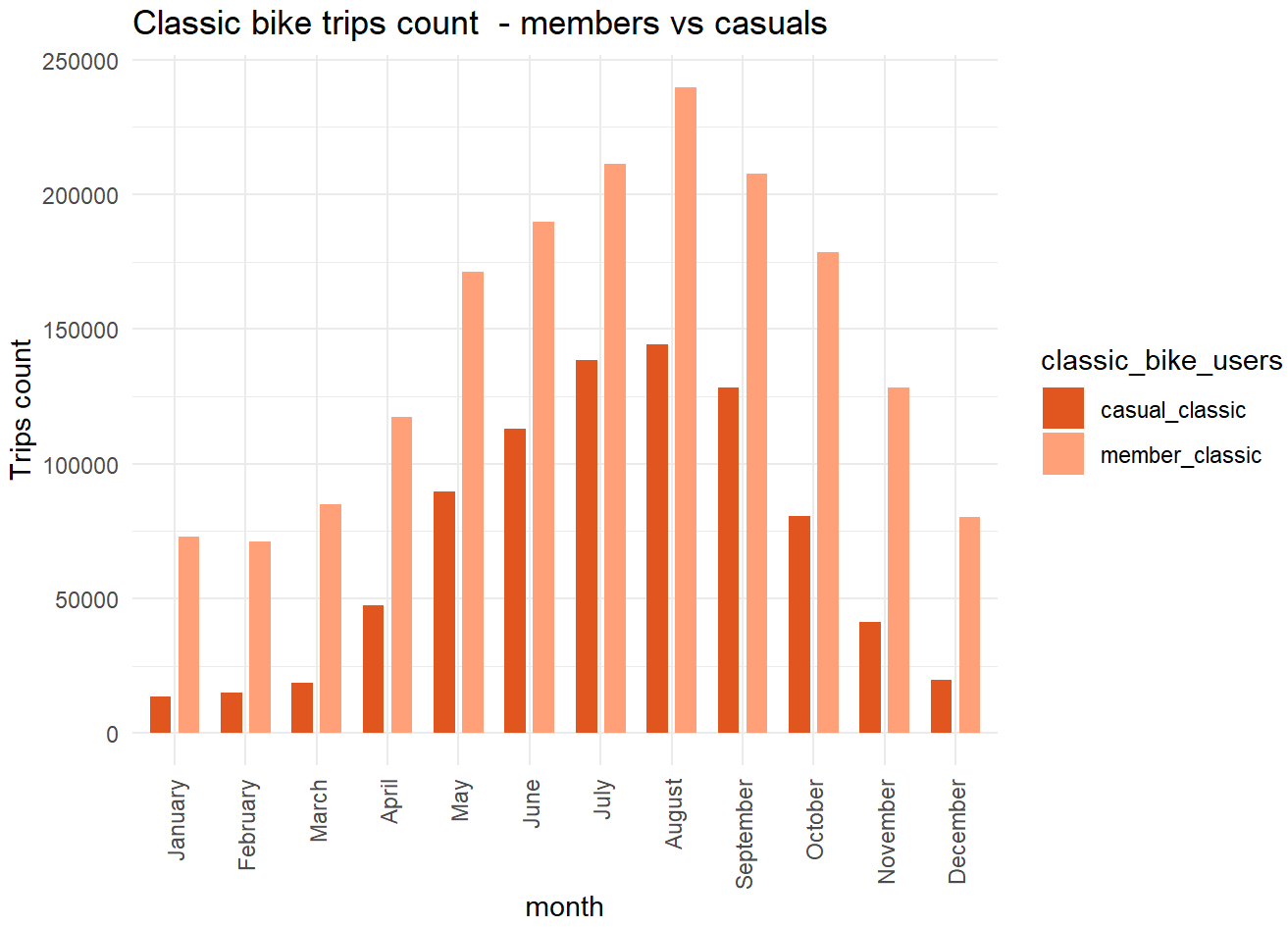
Group by the combination of classic_bike with member_casual

```
count_trips1 <- data2023 %>%
  mutate(month = format(started_at, "%B"),
         classic_bike_users = case_when(
           member_casual == "member" & rideable_type == "classic_bike" ~ "member_classic",
           member_casual == "casual" & rideable_type == "classic_bike" ~ "casual_classic")) %>%
  group_by(month, classic_bike_users) %>%
  summarise(count_trips1 = n(), .groups = 'drop') %>%
  filter(!is.na(classic_bike_users))
```

Convert ‘month’ to factor with custom levels in the desired order

```
count_trips1$month <- factor(count_trips1$month, levels = c(
  "January", "February", "March", "April", "May", "June",
  "July", "August", "September", "October", "November", "December"))
```

```
ggplot(count_trips1, aes(x = month, y = count_trips1,
                         fill = classic_bike_users)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8),
  width = 0.6) +
  scale_fill_manual(values = color_palette3) +
  labs(title = "Classic bike trips count  - members vs casuals", y = "Trips count") +
  theme_minimal() +
   theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Plot6. Trip count for electric_bike between members & casuals.

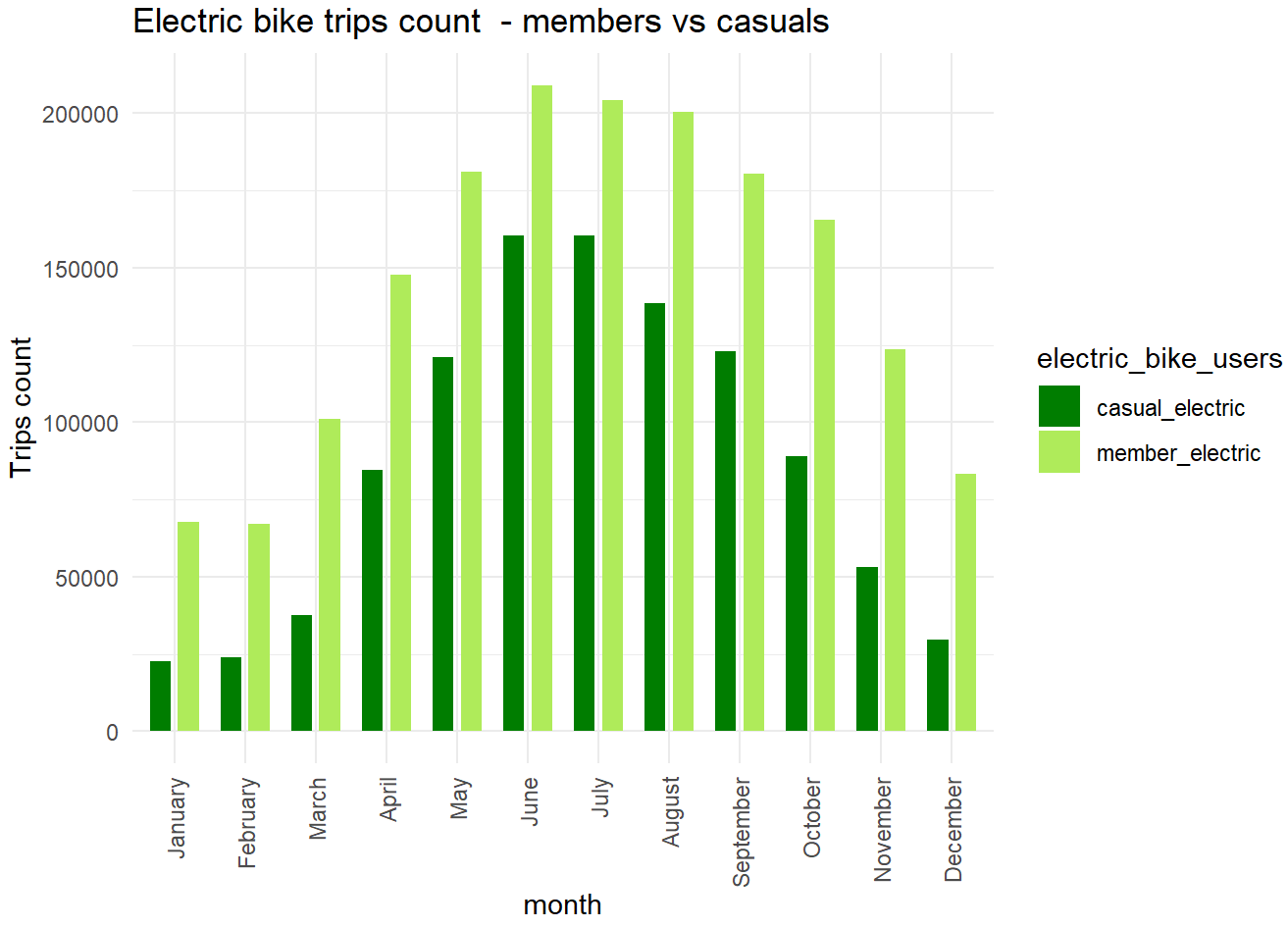
Group by the combination of electric_bike with member_casual

```
count_trips2 <- data2023 %>%
  mutate(month = format(started_at, "%B"),
         electric_bike_users = case_when(
           member_casual == "member" & rideable_type == "electric_bike" ~ "member_electric",
           member_casual == "casual" & rideable_type == "electric_bike" ~ "casual_electric")) %
>%
  group_by(month, electric_bike_users) %>%
  summarise(count_trips2 = n(), .groups = 'drop') %>%
  filter(!is.na(electric_bike_users))
```

Convert ‘month’ to factor with custom levels in the desired order

```
count_trips2$month <- factor(count_trips2$month, levels = c(
  "January", "February", "March", "April", "May", "June",
  "July", "August", "September", "October", "November", "December"))
```

```
ggplot(count_trips2, aes(x = month, y = count_trips2,
                         fill = electric_bike_users)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8),
  width = 0.6) +
  scale_fill_manual(values = color_palette4) +
  labs(title = "Electric bike trips count  - members vs casuals", y = "Trips count") +
  theme_minimal() +
   theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Plot 7. Trip count for weekdays between members & casuals.

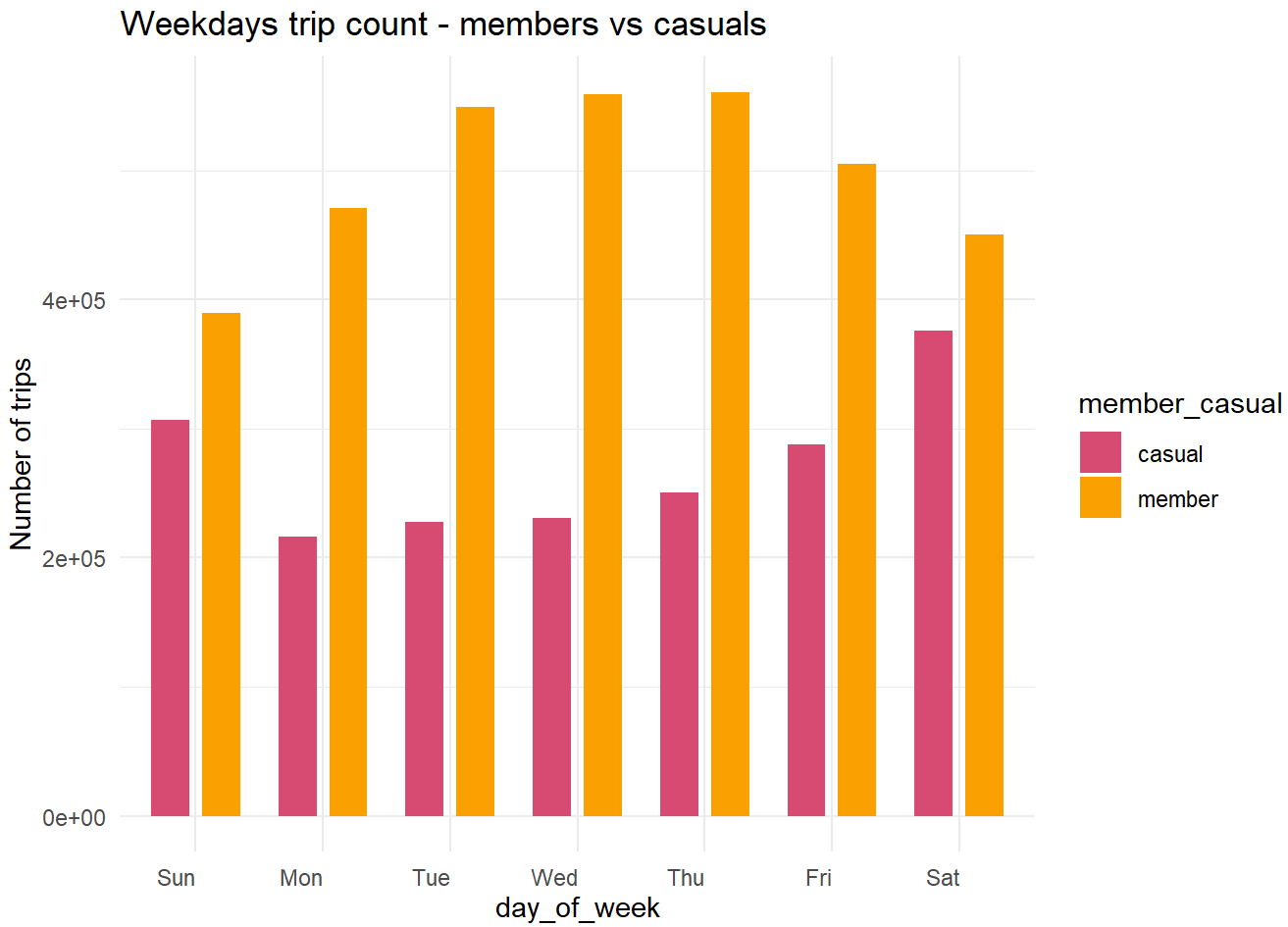
Group by the combination of day_of_week with member_casual

```
count_trips7 <- data2023 %>%
  group_by(day_of_week, member_casual) %>%
  summarise(count_trips7 = n(), .groups = 'drop') %>%
  filter(!is.na(member_casual))
```

Convert ‘day_of_week’ to factor with custom levels in the desired order

```
count_trips7$day_of_week <- factor(count_trips7$day_of_week, levels = c(
  "Sun", "Mon", "Tue", "Wed", "Thu", "Fri", "Sat"))
```

```
ggplot(count_trips7, aes(x = day_of_week, y = count_trips7,
                         fill = member_casual)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8),
  width = 0.6) +
  scale_fill_manual(values = color_palette1) +
  labs(title = "Weekdays trip count - members vs casuals", y = "Number of trips") +
  theme_minimal() +
   theme(axis.text.x = element_text(angle = 0, vjust = 0.5, hjust=1))
```



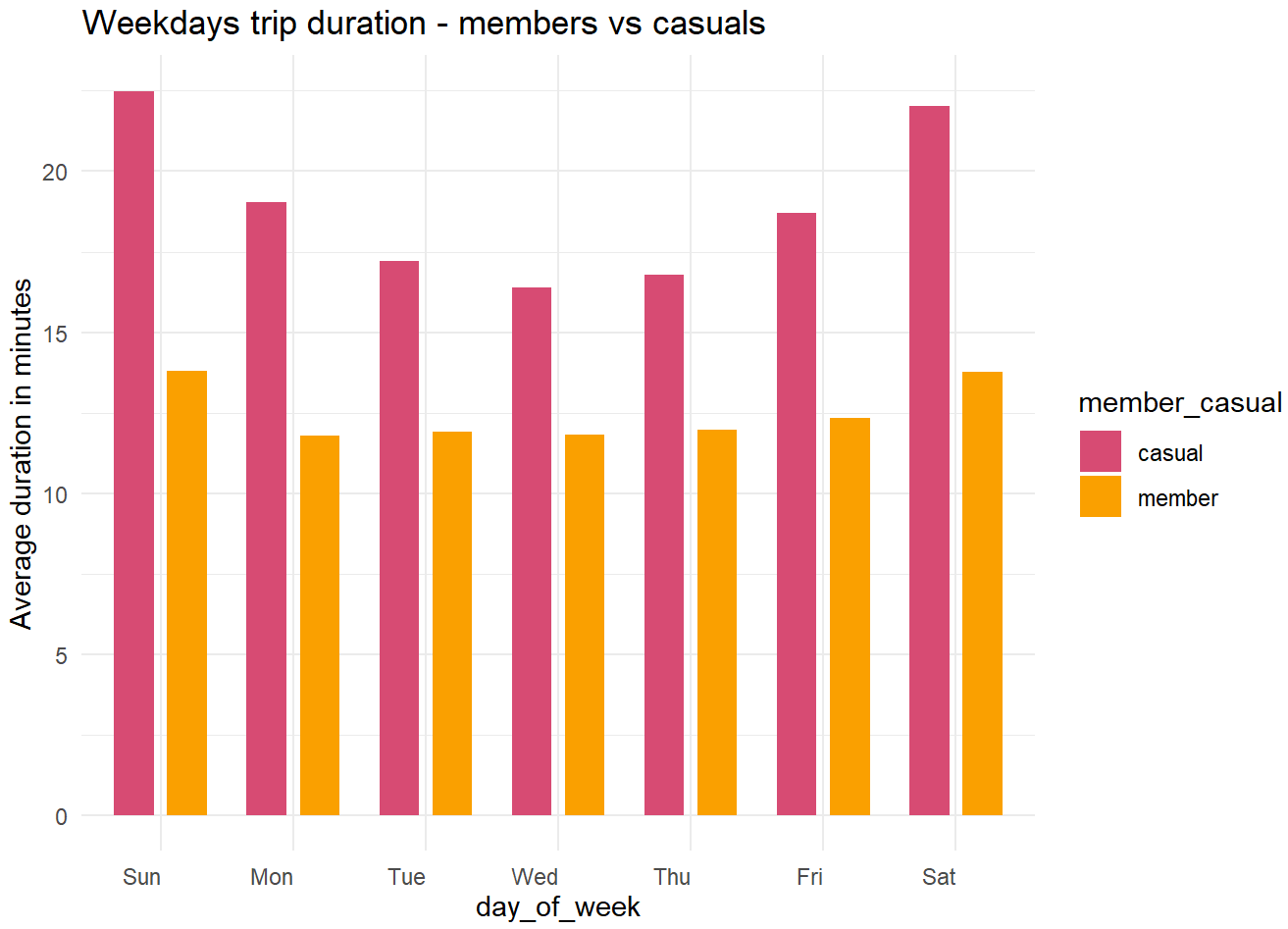
Plot 8. Duration for weekdays between members & casuals.

```
day_dur8 <- data2023 %>%
  group_by(day_of_week, member_casual) %>%
  summarise(day_dur8 = mean(ride_length), .groups = 'drop') %>%
  filter(!is.na(member_casual))
```

Convert ‘day_of_week’ to factor with custom levels in the desired order

```
day_dur8$day_of_week <- factor(day_dur8$day_of_week, levels = c(
  "Sun", "Mon", "Tue", "Wed", "Thu", "Fri", "Sat"))
```

```
ggplot(day_dur8, aes(x = day_of_week, y = day_dur8,
                         fill = member_casual)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8),
  width = 0.6) +
  scale_fill_manual(values = color_palette1) +
  labs(title = "Weekdays trip duration - members vs casuals", y = "Average duration in minutes")
+
  theme_minimal() +
   theme(axis.text.x = element_text(angle = 0, vjust = 0.5, hjust=1))
```

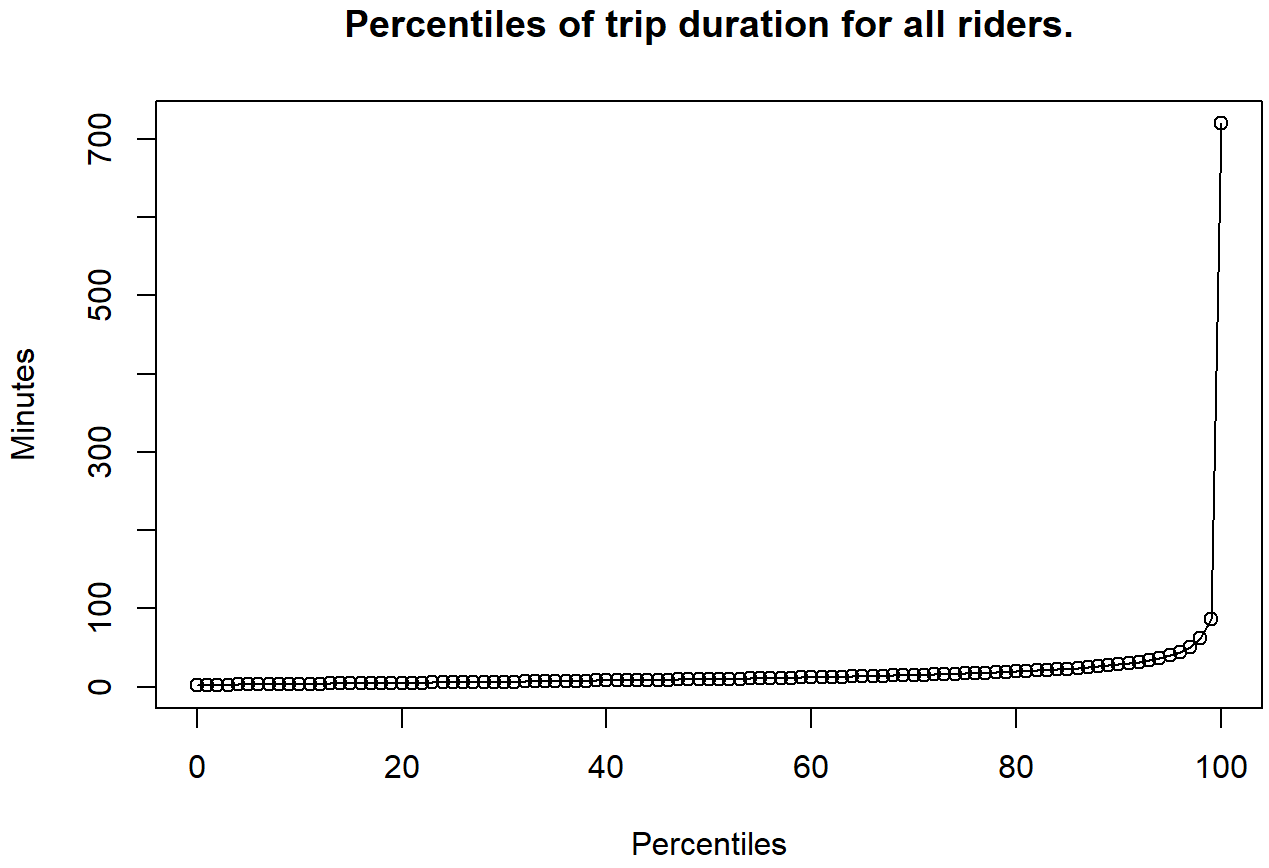


Plot 9. Percentiles of trip duration for all riders.

Define the percentiles with smaller increments

```
percentiles1 <- seq(0, 1, by = 0.01)
values1 <- quantile(data2023$ride_length, probs = percentiles1)
```

```
plot(percentiles1 * 100, values1, type = "o",
     xlab = "Percentiles", ylab = "Minutes",
     main = "Percentiles of trip duration for all riders.")
```



Actual percentiles of trip duration for casuals:

Define the percentiles with smaller increments

```
percentiles2 <- seq(0.97, 1, by = 0.005)
values2 <- quantile(subset(data2023, member_casual == "casual")$ride_length,
                    probs = percentiles2)
```

Create a data frame with percentiles and values

```
perc_casual2 <- data.frame(Percentile = percentiles2, Value = values2)
```

Print the data frame for casuals

```
print(perc_casual2)
```

```
##       Percentile Value
## 97%        0.970  75.1
## 97.5%      0.975  81.4
## 98%        0.980  89.7
## 98.5%      0.985 101.1
## 99%        0.990 118.5
## 99.5%      0.995 153.2
## 100%       1.000 719.4
```

Percentiles of trip duration for casual riders, from 60% to 100%.”

```
percentiles3 <- seq(0.6, 1, by = 0.05)
values3 <- quantile(subset(data2023, member_casual == "casual")$ride_length,
                    probs = percentiles3)
```

Create a data frame with percentiles and values

```
perc_casual3 <- data.frame(Percentile = percentiles3, Value = values3)
```

Print the data frame for casual riders

```
print(perc_casual3)
```

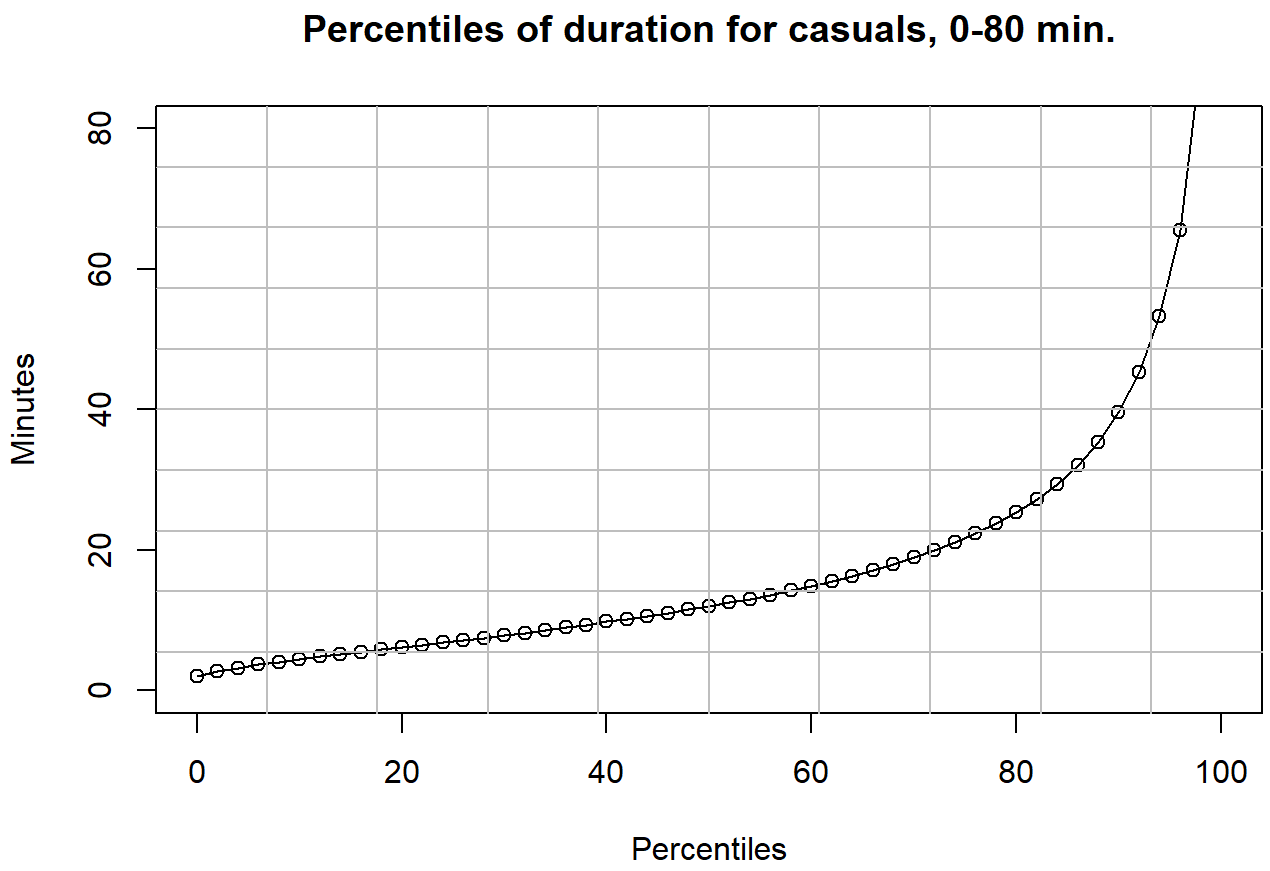
```
##      Percentile Value
## 60%        0.60  14.9
## 65%        0.65  16.7
## 70%        0.70  19.0
## 75%        0.75  21.8
## 80%        0.80  25.4
## 85%        0.85  30.7
## 90%        0.90  39.6
## 95%        0.95  58.6
## 100%       1.00 719.4
```

Plot 10. Trip duration 0-80 min. in percentiles for casuals

Define the percentiles with smaller increments

```
percentiles4 <- seq(0, 1, by = 0.02)
values4 <- quantile(subset(data2023, member_casual == "casual")$ride_length,
                    probs = percentiles4)
```

```
plot(percentiles4 * 100, values4, type = "o",
     xlab = "Percentiles", ylab = "Minutes",
     main = "Percentiles of duration for casuals, 0-80 min.",
     ylim = c(0, 80))
grid(nx = 10, ny = 10, lty = 1, col = "gray", lwd = 1)
```



Actual percentiles of trip duration for members:

Define the percentiles with smaller increments

```
percentiles5 <- seq(0.97, 1, by = 0.005)
values5 <- quantile(subset(data2023, member_casual == "member")$ride_length,
                    probs = percentiles5)
```

Create a data frame with percentiles and values

```
perc_member5 <- data.frame(Percentile = percentiles5, Value = values5)
```

Print the data frame for members

```
print(perc_member5)
```

```
##       Percentile Value
## 97%        0.970  37.5
## 97.5%      0.975  39.6
## 98%        0.980  42.0
## 98.5%      0.985  45.5
## 99%        0.990  52.1
## 99.5%      0.995  70.0
## 100%       1.000 718.1
```

Plot 11. Trip duration with data labels for all riders by quarters.

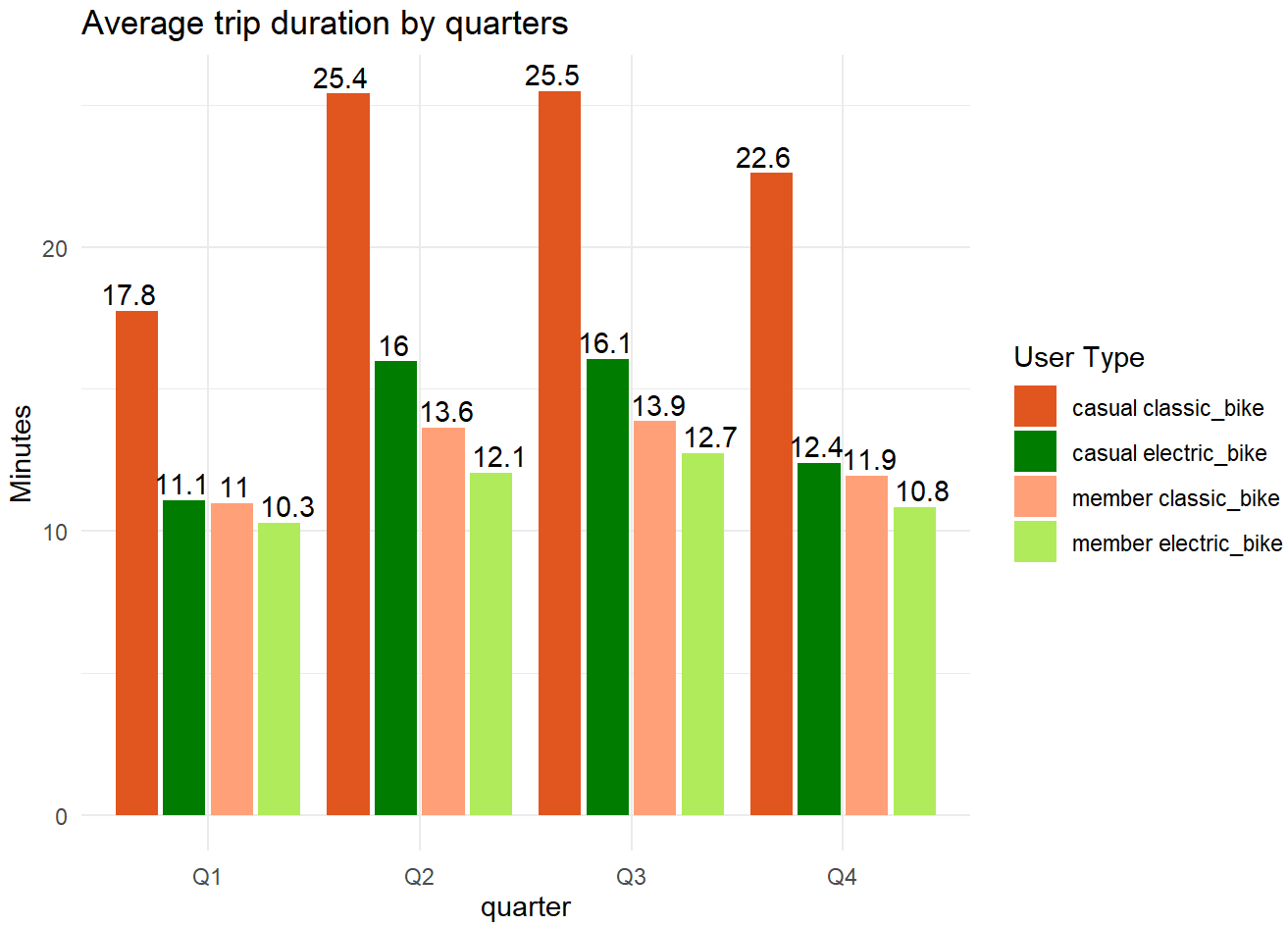
Define a function to convert month to quarter

```
month_to_quarter <- function(month) {
  case_when(
    month %in% c("January", "February", "March") ~ "Q1",
    month %in% c("April", "May", "June") ~ "Q2",
    month %in% c("July", "August", "September") ~ "Q3",
    TRUE ~ "Q4"
  )
}
```

Aggregate data by quarter

```
avg_dur_quarter <- data2023 %>%
  mutate(month = format(started_at, "%B")) %>%
  mutate(quarter = month_to_quarter(month)) %>%
  group_by(quarter, type_of_users = paste(member_casual, rideable_type)) %>%
  summarise(avg_dur_quarter = mean(ride_length), .groups = 'drop')
```

```
ggplot(avg_dur_quarter, aes(x = quarter, y = avg_dur_quarter, fill = type_of_users, label = roun
d(avg_dur_quarter, 1 ))) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.9), width = 0.8) +
  geom_text(position = position_dodge(width = 1), vjust = -0.3) +
  scale_fill_manual(values = color_palette2) +
  labs(y = "Minutes", title = "Average trip duration by quarters") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, vjust = 0.5)) +
  guides(fill = guide_legend(title = "User Type"))
```

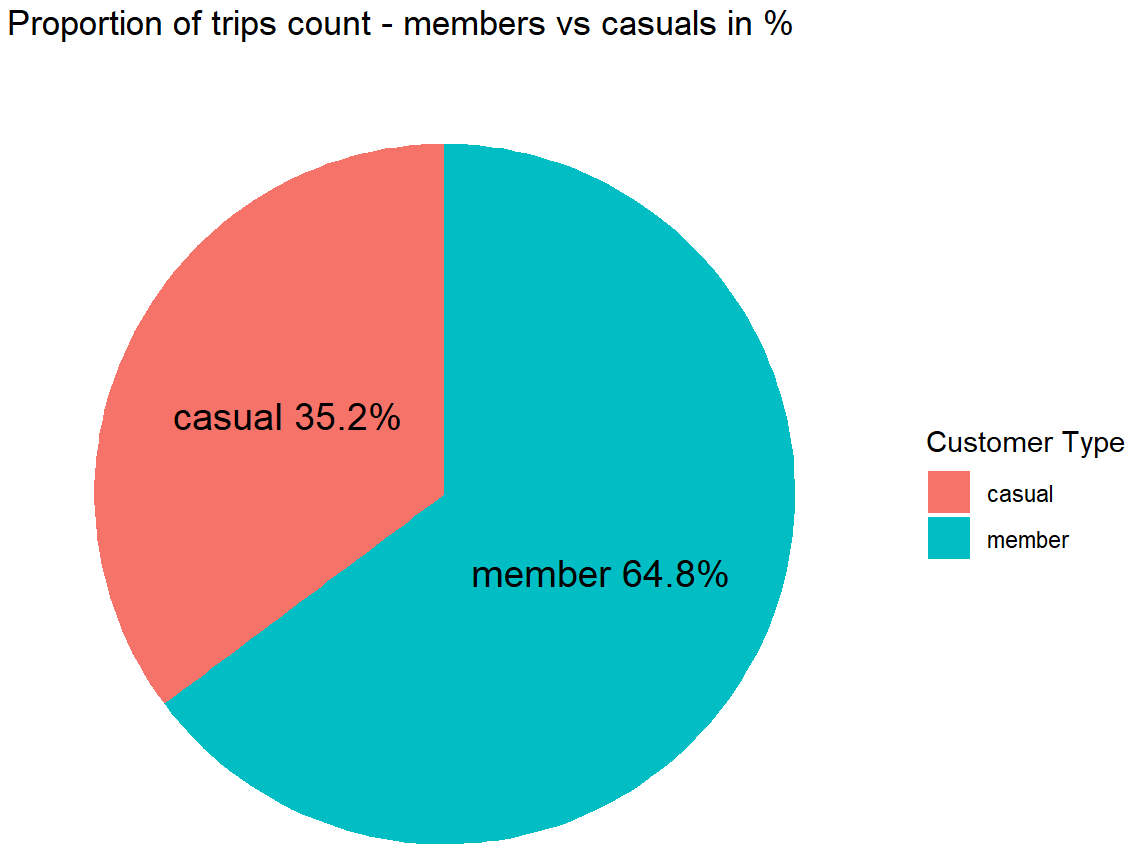


Plot 12. Pie chart for trip number of members vs casuals in %.

Grouping & aggregation

```
data2023_pie1 <- data2023 %>%
  group_by(member_casual) %>%
  summarise(count = n()) %>%
  mutate(percentage = count / sum(count) * 100)
```

```
ggplot(data2023_pie1, aes(x = "", y = percentage, fill = member_casual)) +
  geom_bar(width = 1, stat = "identity") +
  geom_text(aes(label = paste(member_casual, sprintf("%.1f%%", percentage))), position = positio
n_stack(vjust = 0.5), size = 5) +
  coord_polar(theta = "y") +
  labs(fill = "Customer Type") +
  ggtitle("Proportion of trips count - members vs casuals in %") +
  theme_minimal() +
  theme(legend.position = "right",
        axis.line = element_blank(),
        axis.text = element_blank(),
        axis.title = element_blank(),
        panel.grid = element_blank())
```

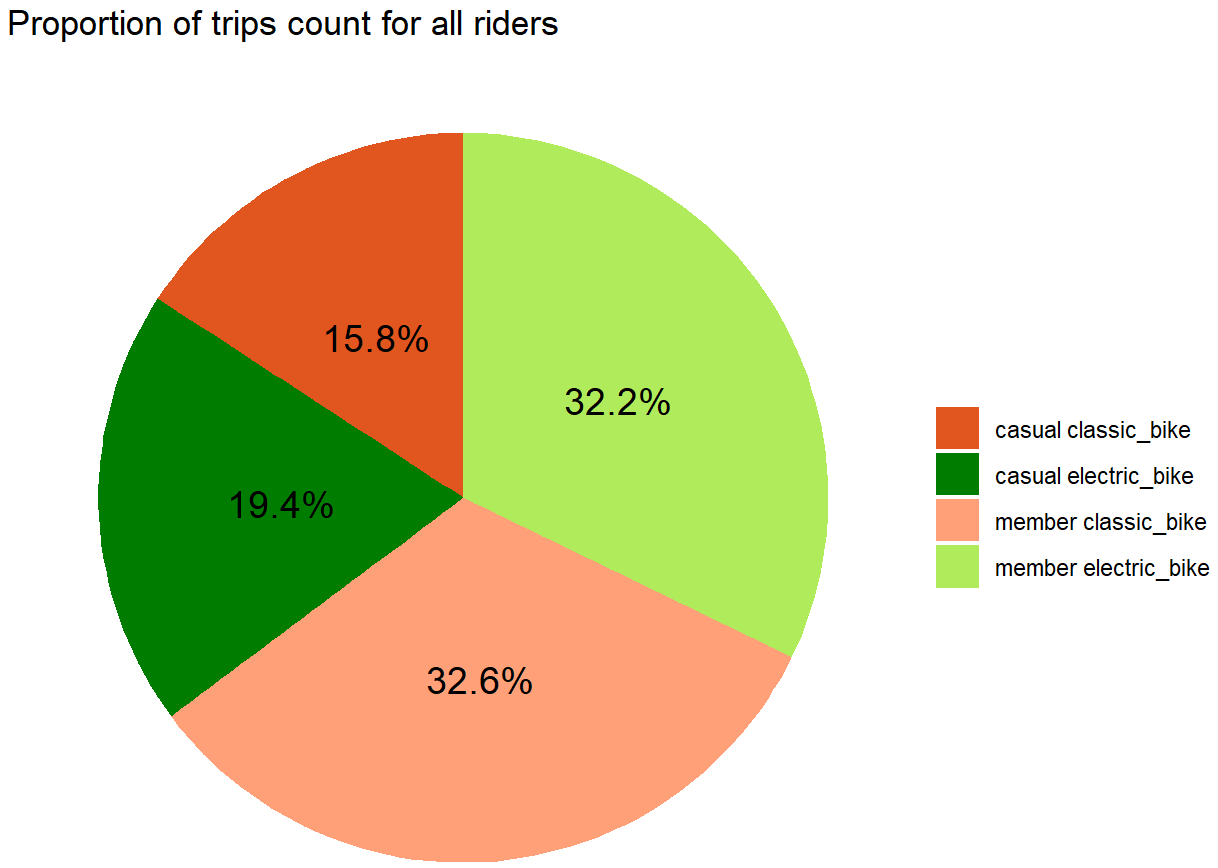


Plot 13. Pie chart for trip number of all riders in %.

Grouping & aggregation

```
trips_count <- data2023 %>%
  group_by(type_of_users = paste(member_casual, rideable_type)) %>%
  summarise(count = n(), .groups = 'drop')
```

```
ggplot(trips_count, aes(x = "", y = count, fill = type_of_users, label = sprintf("%.1f%%", coun
t/sum(count)*100))) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar(theta = "y") +
  scale_fill_manual(values = color_palette2) +
  labs(fill = "User Type") +
  ggtitle("Proportion of trips count for all riders") +
  theme_void() +
  geom_text(position = position_stack(vjust = 0.5), size = 5) +
  theme(legend.position = "right",
        legend.title = element_blank())
```



Plot 14. Trip duration with data labels for all riders by seasons.

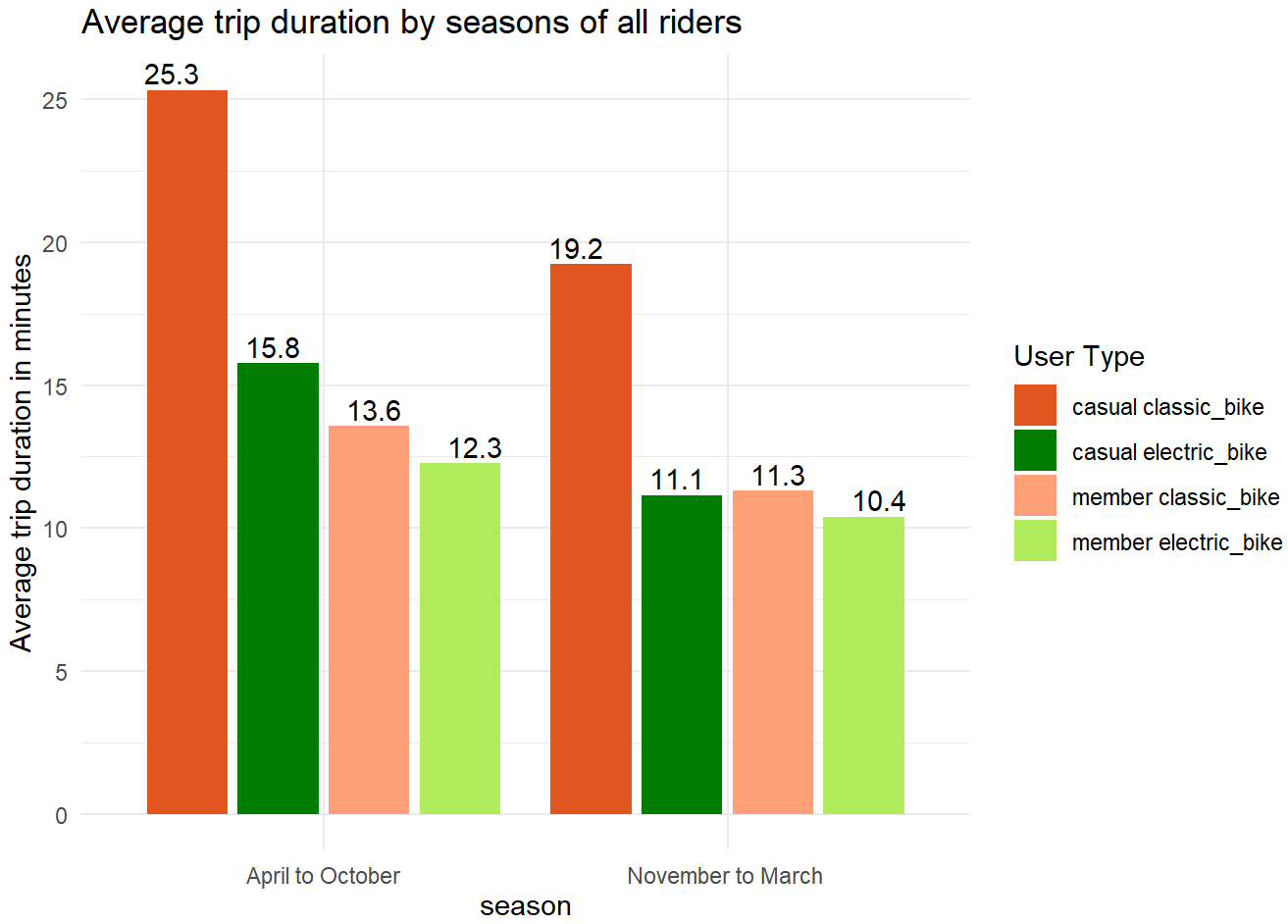
Define a function to convert month to season

```
month_to_season <- function(month) {
  case_when(
    month %in% c("January", "February", "March", "November", "December") ~ "November to March ",
        TRUE ~ "April to October"
  )
}
```

Aggregate data by seasons

```
avg_dur_season <- data2023 %>%
  mutate(month = format(started_at, "%B")) %>%
  mutate(season = month_to_season(month)) %>%
  group_by(season, type_of_users = paste(member_casual, rideable_type)) %>%
  summarise(avg_dur_season = mean(ride_length), .groups = 'drop')
```

```
ggplot(avg_dur_season, aes(x = season, y = avg_dur_season, fill = type_of_users, label = round(a
vg_dur_season, 1 ))) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.9), width = 0.8) +
  geom_text(position = position_dodge(width = 1), vjust = -0.3) +
  scale_fill_manual(values = color_palette2) +
  labs(y = "Average trip duration in minutes", title = "Average trip duration by seasons of all
riders") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, vjust = 0.5)) +
  guides(fill = guide_legend(title = "User Type"))
```



Plot 15. Trip count with data labels for all riders by seasons.

Define a function to convert month to season

```
month_to_season <- function(month) {
  case_when(
    month %in% c("January", "February", "March", "November", "December") ~ "November to March",
        TRUE ~ "April to October"
  )
}
```

Aggregate data by seasons

```
tr_count_season <- data2023 %>%
  mutate(month = format(started_at, "%B")) %>%
  mutate(season = month_to_season(month)) %>%
  group_by(season, type_of_users = paste(member_casual, rideable_type)) %>%
  summarise(tr_count_season = n(), .groups = 'drop')
```

```
ggplot(tr_count_season, aes(x = season, y = tr_count_season, fill = type_of_users, label = round
(tr_count_season, 0 ))) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.9), width = 0.8) +
  geom_text(position = position_dodge(width = 1.1), vjust = -0.4) +
  scale_fill_manual(values = color_palette2) +
  labs(y = "Trips count", title = "Trips count by seasons for all riders") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, vjust = 0.5)) +
  guides(fill = guide_legend(title = "User Type"))
```

```
## Warning: `position_dodge()` requires non-overlapping x intervals.
```