Casual Riders vs Annual Members use of Bikes

Google Data Analytics Capstone Project :

By: Ali-Hussain Momin

Introduction

- In this project, I will be using Cyclist data to analyze and understand how casual riders and annual members use Cyclist bikes differently; in order to add more members.
- I will be going through the following steps of data analysis process:
  - Ask
  - Prepare
  - Process
  - Analyze
  - Share
  - Act
- Each step will consist of guiding questions with answers, key tasks, code, and deliverables.

Scenario:

You are a junior data analyst working in the marketing analyst team at Cyclist, a bike-share company in Chicago. The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore, your team wants to understand how casual riders and annual members use Cyclist bikes differently. From these insights, your team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclist executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations.

Ask

Guiding Questions:

- What is the problem you are trying to solve ?
  - A way to build a marketing strategy to turn casual riders into annual members.

- How can your insights drive business decisions ?
  - The marketing team will have an idea what to do in order to add more members by seeing the difference between casual and annual members use of bikes.

Key Tasks:

- Identify the business task
- Consider key stakeholders

Deliverables:

- A clear statement of the business task:
  - Find key differences between casual and member riders and how digital media could influence their decision

Prepare

Guiding Questions:

- Where is your data located ?
  - The data is provided by Google and saved onto Kaggle.
- How is the data organized ?
  - The data is separated by month on each csv file.
- Are there issues with bias or credibility in this data ? Does your data ROCCC (Reliable, Original, Comprehensive, Current, Cited) ?
  - The dataset is from its own clients, therefore fully credible and ROCCC.
- How are you addressing licensing, privacy, security, and accessibility ?
  - The company has their own license over the dataset which does not contain any personal information about the riders.
- How did you verify the data's integrity ?
  - By verifying if the type of data in column is correct and the consistency of the column in the dataset.
- How does it help you answer your question ?
  - The dataset can hold key insights about the use of bikes between both casual and annual members.
- Are there any problems with the data ?
  - More information about the riders and the bike station can be useful in order to dive deeper into the problem.

Prepare continued...

Key Tasks:

- Download data and store it properly
- Identify how it's organized
- Sort and filter the data
- Determine the credibility of the data

Deliverable:

- A description of all data sources used

Process

Guiding Questions

- What tools are you choosing and why ?
  - I am using R for this project because it works well with large datasets and visualizations.
- Have you ensured your data's integrity?
  - The data has been checked and it is consistent throughout.
- What steps have you taken to ensure that your data is clean ?
  - Correcting the format of the columns and removing any duplicates within the data.
- How can you verify that your data is clean and ready to analyze ?
  - My notebook in Kaggle that was used to work on this project.
- Have you documented your cleaning process so you can review and share those results ?
  - Everything is documents with in the R notebook.

Key Task:

- Check the data for errors
- Choose your tools
- Transform the data so you can work with it effectively
- Document the cleaning process

Deliverables:

- Documentation of any cleaning or manipulation of the data

Code

Tidyverse will be the main dependency.

[1]:
```
library(tidyverse)
```

```
── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

✔ ggplot2 3.3.5     ✔ purrr   0.3.4
✔ tibble  3.1.5     ✔ dplyr   1.0.7
✔ tidyr   1.1.4     ✔ stringr 1.4.0
✔ readr   2.0.2     ✔ forcats 0.5.1

── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
```

Next, I will be concatenating (combining) all of the datasets into one.

[2]:
```
trip202012 <- read_csv("../input/cyclist/data/202012-divvy-tripdata.csv")
trip202101 <- read_csv("../input/cyclist/data/202101-divvy-tripdata.csv")
trip202102 <- read_csv("../input/cyclist/data/202102-divvy-tripdata.csv")
trip202103 <- read_csv("../input/cyclist/data/202103-divvy-tripdata.csv")
trip202104 <- read_csv("../input/cyclist/data/202104-divvy-tripdata.csv")
trip202105 <- read_csv("../input/cyclist/data/202105-divvy-tripdata.csv")
trip202106 <- read_csv("../input/cyclist/data/202106-divvy-tripdata.csv")
trip202107 <- read_csv("../input/cyclist/data/202107-divvy-tripdata.csv")
trip202108 <- read_csv("../input/cyclist/data/202108-divvy-tripdata.csv")
trip202009 <- read_csv("../input/cyclist/data/202009-divvy-tripdata.csv")
trip202010 <- read_csv("../input/cyclist/data/202010-divvy-tripdata.csv")
trip202011 <- read_csv("../input/cyclist/data/202011-divvy-tripdata.csv")
df1 <- read.csv("../input/cyclist/data/202009-divvy-tripdata.csv")
df2 <- read.csv("../input/cyclist/data/202010-divvy-tripdata.csv")
df3 <- read.csv("../input/cyclist/data/202011-divvy-tripdata.csv")
df4 <- read.csv("../input/cyclist/data/202012-divvy-tripdata.csv")
df5 <- read.csv("../input/cyclist/data/202101-divvy-tripdata.csv")
df6 <- read.csv("../input/cyclist/data/202102-divvy-tripdata.csv")
df7 <- read.csv("../input/cyclist/data/202103-divvy-tripdata.csv")
df8 <- read.csv("../input/cyclist/data/202104-divvy-tripdata.csv")
df9 <- read.csv("../input/cyclist/data/202105-divvy-tripdata.csv")
df10 <- read.csv("../input/cyclist/data/202106-divvy-tripdata.csv")
df11 <- read.csv("../input/cyclist/data/202107-divvy-tripdata.csv")
df12 <- read.csv("../input/cyclist/data/202108-divvy-tripdata.csv")
```

Continue...

[4]:
```
bike_trips <- rbind(trip202012, trip202101, trip202102, trip202103
                , trip202104, trip202105, trip202106
                , trip202107, trip202108, trip202009
                , trip202010, trip202011, df1, df2, df3, df4, df5, df6, df7, df8, df9, df10, df11, df12)
```

Removing duplicates

[14]:
```
new_bike_trips <- bike_trips[!duplicated(bike_trips$ride_id), ]
print(paste("Removed", nrow(bike_trips) - nrow(new_bike_trips), "duplicated rows"))
```

```
[1] "Removed 4913281 duplicated rows"
```

Parsing datetime columns

[15]:
```
new_bike_trips$started_at <- as.POSIXct(new_bike_trips$started_at, "%Y-%m-%d %H: %M:%S")
```

Continue... (Manipulating Data)

Adding new columns to help with data analysis when calculating time in the future.

Ride_time_m rep. Total time of bike ride in minutes

[17]:
```
new_bike_trips <- new_bike_trips %>%
    mutate(ride_time_m = as.numeric(new_bike_trips$ended_at - new_bike_trips$started_at) / 60 )
summary(new_bike_trips$ride_time_m)
```

```
    Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-29049.97      7.18     12.80     21.14     23.27  55944.15
```

Separate year and month into one column

[19]:
```
new_bike_trips <- new_bike_trips %>%
    mutate(year_month = paste(strftime(new_bike_trips$started_at, "%Y"), "-",
                              strftime(new_bike_trips$started_at, "%m"), paste("(",
                              strftime(new_bike_trips$started_at, "%b"), ")", sep = "")))
unique(new_bike_trips$year_month)
```

'2020 - 12 (Dec)' · '2021 - 01 (Jan)' · '2021 - 02 (Feb)' · '2021 - 03 (Mar)' · '2021 - 04 (Apr)' · '2021 - 05 (May)' · '2021 - 06 (Jun)' · '2021 - 07 (Jul)' · '2021 - 08 (Aug)' · '2020 - 09 (Sep)' · '2020 - 10 (Oct)' · '2020 - 11 (Nov)'

Weekday will be used to find patterns during the week

[21]:
```
new_bike_trips <- new_bike_trips %>%
    mutate(weekday = paste(strftime(new_bike_trips$ended_at, "%u"), "-", strftime(new_bike_trips$ended_at, "%a")))
unique(new_bike_trips$weekday)
```

'7 - Sun' · '5 - Fri' · '2 - Tue' · '4 - Thu' · '6 - Sat' · '1 - Mon' · '3 - Wed'

Continue...

start_hour for daily intra analysis

[11]:
```
new_bike_trips <- new_bike_trips %>%
    mutate(start_hour = strftime(new_bike_trips$ended_at, "%H"))
unique(new_bike_trips$start_hour)
```

'12' · '17' · '15' · '16' · '13' · '09' · '23' · '11' · '10' · '14' · '19' · '08' · '00' · '22' · '18' · '07' · '21' · '20' · '01' · '06' · '05' · '02' · '04' · '03'

Saving the cleaned data as csv

[12]:
```
new_bike_trips %>%
    write.csv("bike_trips_clean.csv")
```

Analyze

Guiding Questions

- How should you organize your data to perform analysis on it ?
  - The data has been organized into a single CSV concatenating all the files from the dataset.
- What trends or relationships did you find in the data ?
  - There are more members than casuals in the dataset.
  - There are more data points in the last semester of 2020.
  - There are more of a difference between the flow of members/casual from midweek to weekends.
  - Members use bikes on schedules that differs from casual.
  - Members have less riding time.
  - Members tend to prefer docked bikes.
- How will theses insights help answer your business questions?
  - These insights helps to build a profile for members.

Key Tasks:

- Aggregate your data so it's useful and accessible.
- Organize and format your data.
- Perform calculations.
- Identify trends and relationships.

Deliverables:

- Identify trends and relationships.

Code

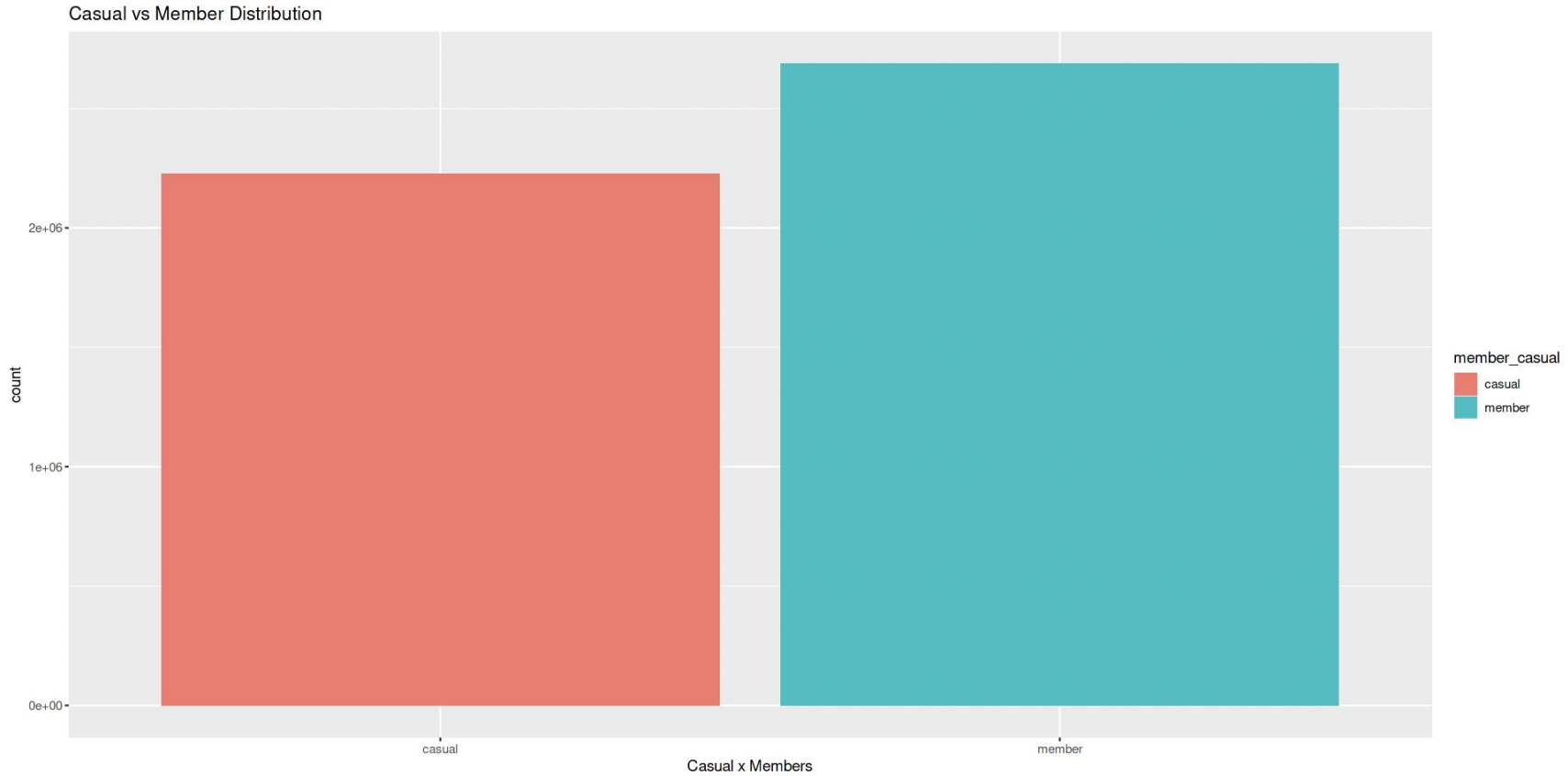
Data distribution between casual and annual members.

[15]:
```
new_bike_trips %>%
    group_by(member_casual) %>%
    summarise(count = length(ride_id), '%' = (length(ride_id) / nrow(new_bike_trips)) * 100)
```

A tibble: 2 × 3

member_casual	count	%
<chr>	<int>	<dbl>
casual	2225064	45.29058
member	2687799	54.70942

[16]:
```
fig(16, 8)
ggplot(new_bike_trips, aes(member_casual, fill = member_casual)) +
    geom_bar() +
    labs(x = "Casual x Members", title = "Casual vs Member Distribution")
```

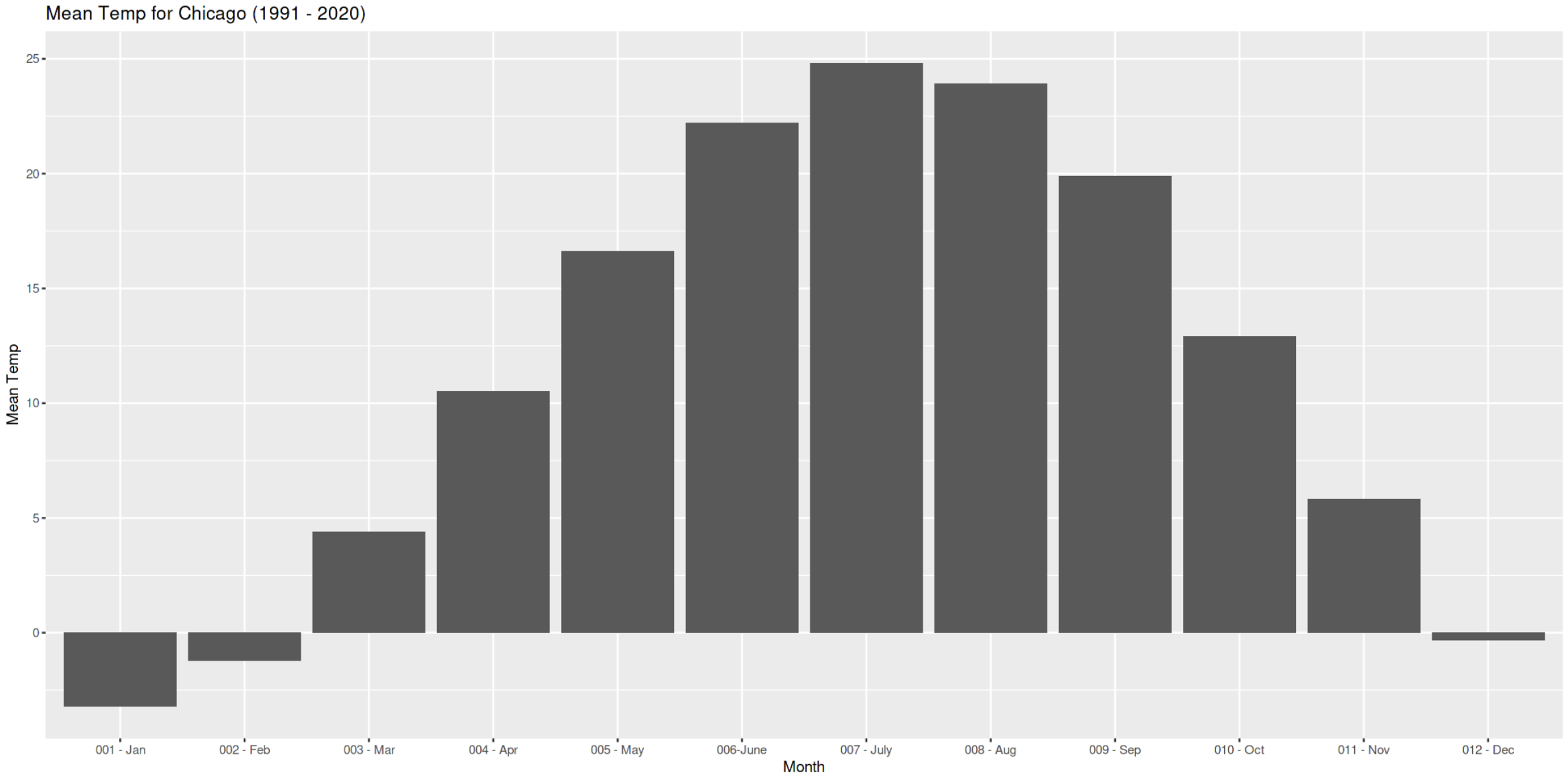


Continue...

[25]:

```
chicago_mean_temp <- c(-3.2, -1.2, 4.4, 10.5, 16.6, 22.2, 24.8, 23.9, 19.9, 12.9, 5.8, -0.3)
month <- c("001 - Jan", "002 - Feb", "003 - Mar", "004 - Apr", "005 - May", "006-June", "007 - July",
           "008 - Aug", "009 - Sep", "010 - Oct", "011 - Nov", "012 - Dec")

data.frame(month, chicago_mean_temp) %>%
    ggplot(aes(x = month, y = chicago_mean_temp)) +
    labs(x = "Month", y = "Mean Temp", title = "Mean Temp for Chicago (1991 - 2020)") +
        geom_col()
```



Continue...

Distibution by month

[34]:

```
new_bike_trips %>%
    group_by(year_month) %>%
    summarise(count = length(ride_id), '%' = (length(ride_id) / nrow(new_bike_trips)) * 100,
              'members_p' = (sum(member_casual == "member") / length(ride_id)) * 100,
              'casual_p' = (sum(member_casual == "casual") / length(ride_id)) *100,
              'Member x Casual Diff.' = members_p - casual_p)
```

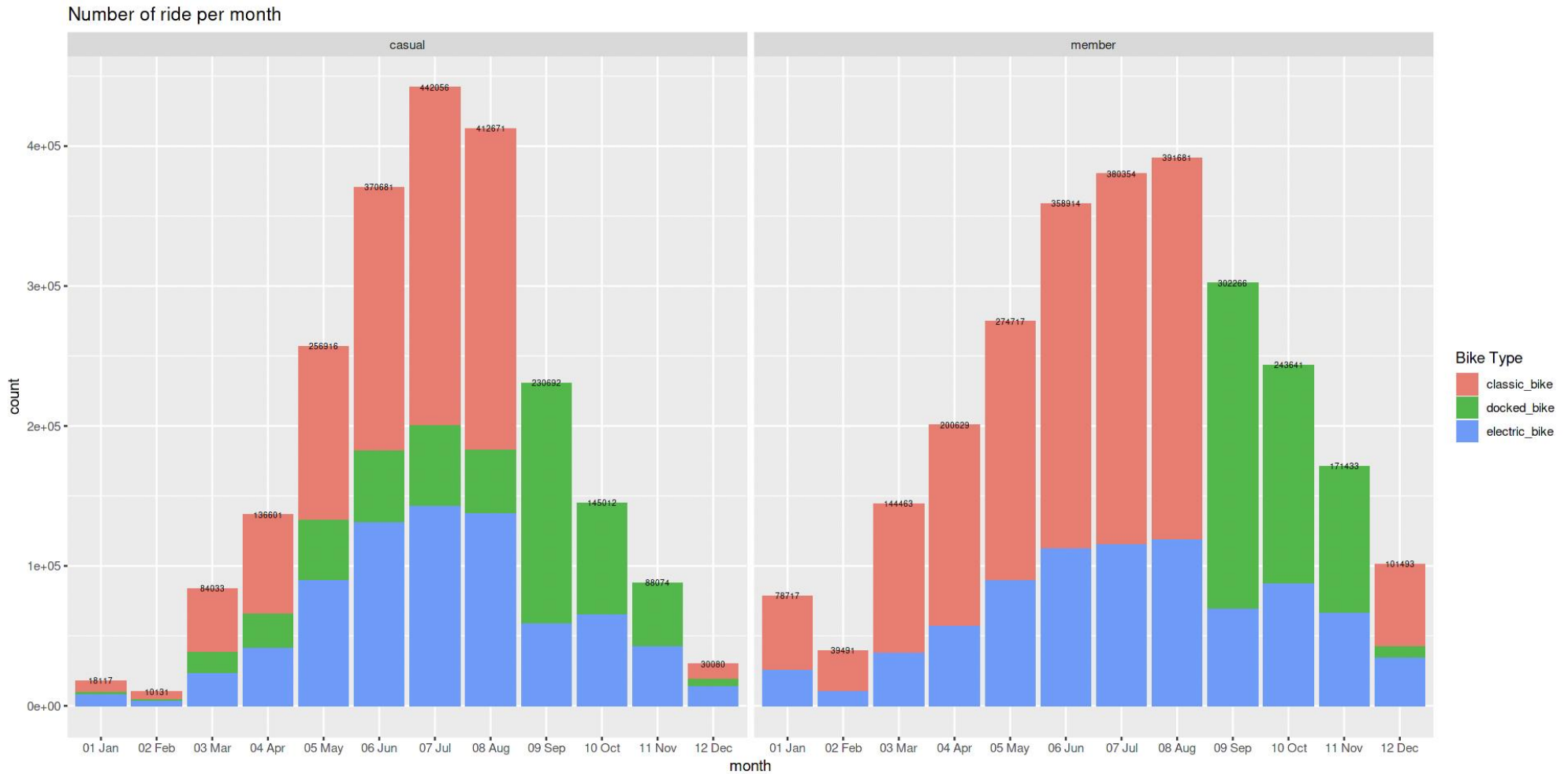
A tibble: 12 × 6

year_month	count	%	members_p	casual_p	Member x Casual Diff.
<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
2020 - 09 (Sep)	532958	10.848216	56.71479	43.28521	13.429576
2020 - 10 (Oct)	388653	7.910927	62.68857	37.31143	25.377136
2020 - 11 (Nov)	259507	5.282195	66.06103	33.93897	32.122062
2020 - 12 (Dec)	131573	2.678133	77.13817	22.86183	54.276333
2021 - 01 (Jan)	96834	1.971030	81.29066	18.70934	62.581325
2021 - 02 (Feb)	49622	1.010042	79.58365	20.41635	59.167305
2021 - 03 (Mar)	228496	4.650974	63.22343	36.77657	26.446852
2021 - 04 (Apr)	337230	6.864226	59.49322	40.50678	18.986448
2021 - 05 (May)	531633	10.821246	51.67418	48.32582	3.348362
2021 - 06 (Jun)	729595	14.850709	49.19359	50.80641	-1.612813
2021 - 07 (Jul)	822410	16.739934	46.24871	53.75129	-7.502584
2021 - 08 (Aug)	804352	16.372368	48.69522	51.30478	-2.609554

Continue...

[75]:

```
ggplot(new_bike_trips, aes(month)) +
  geom_bar(mapping= aes(fill = rideable_type)) +
  stat_count(geom = "text", aes(label = ..count..), size = 2, position=position_stack(vjust=1)) +
  labs(title ="Number of ride per month", fill = "Bike Type") +
  facet_grid(~member_casual)
```

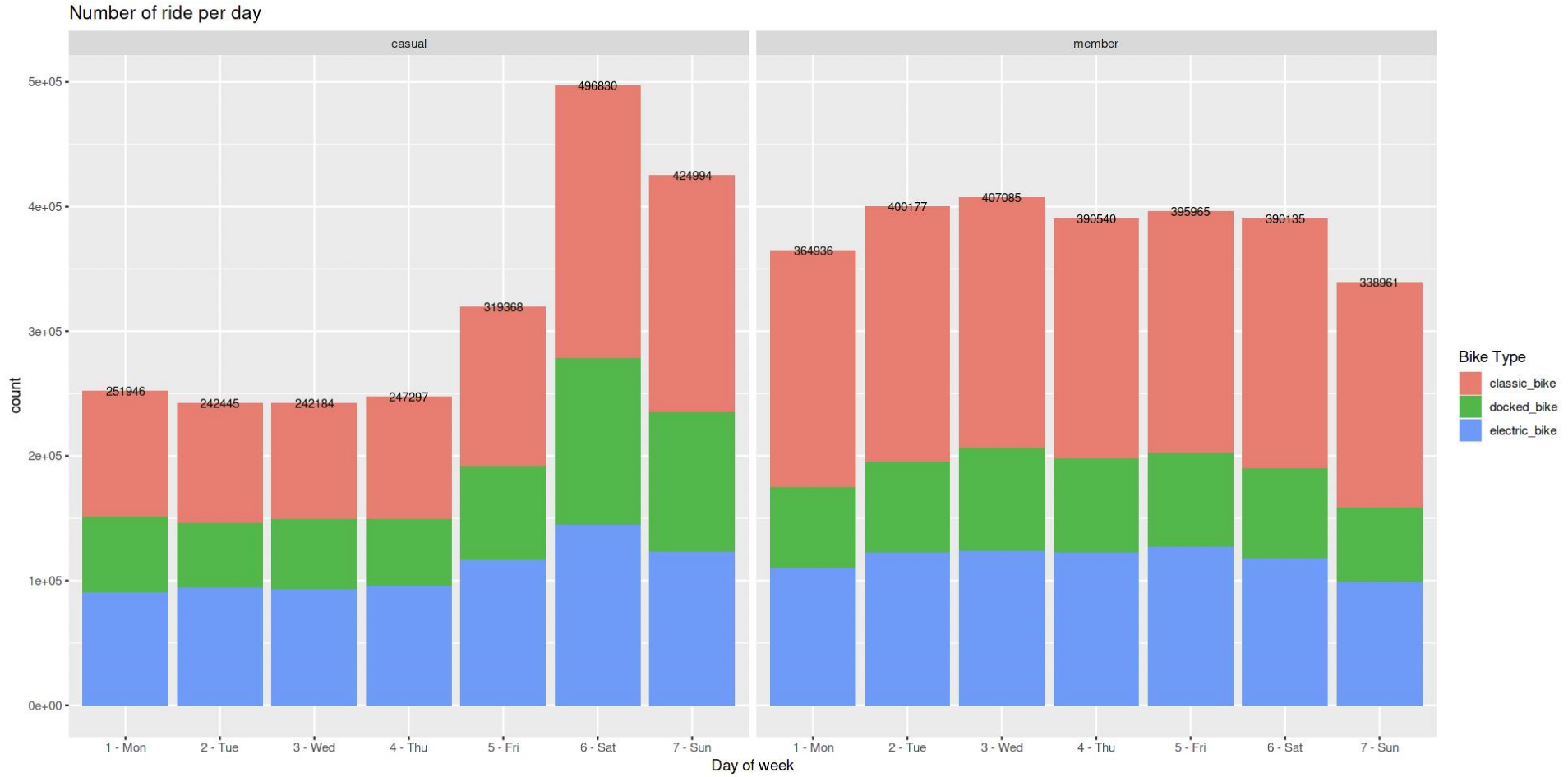


Continue...

Distribution by weekdays

[28]:

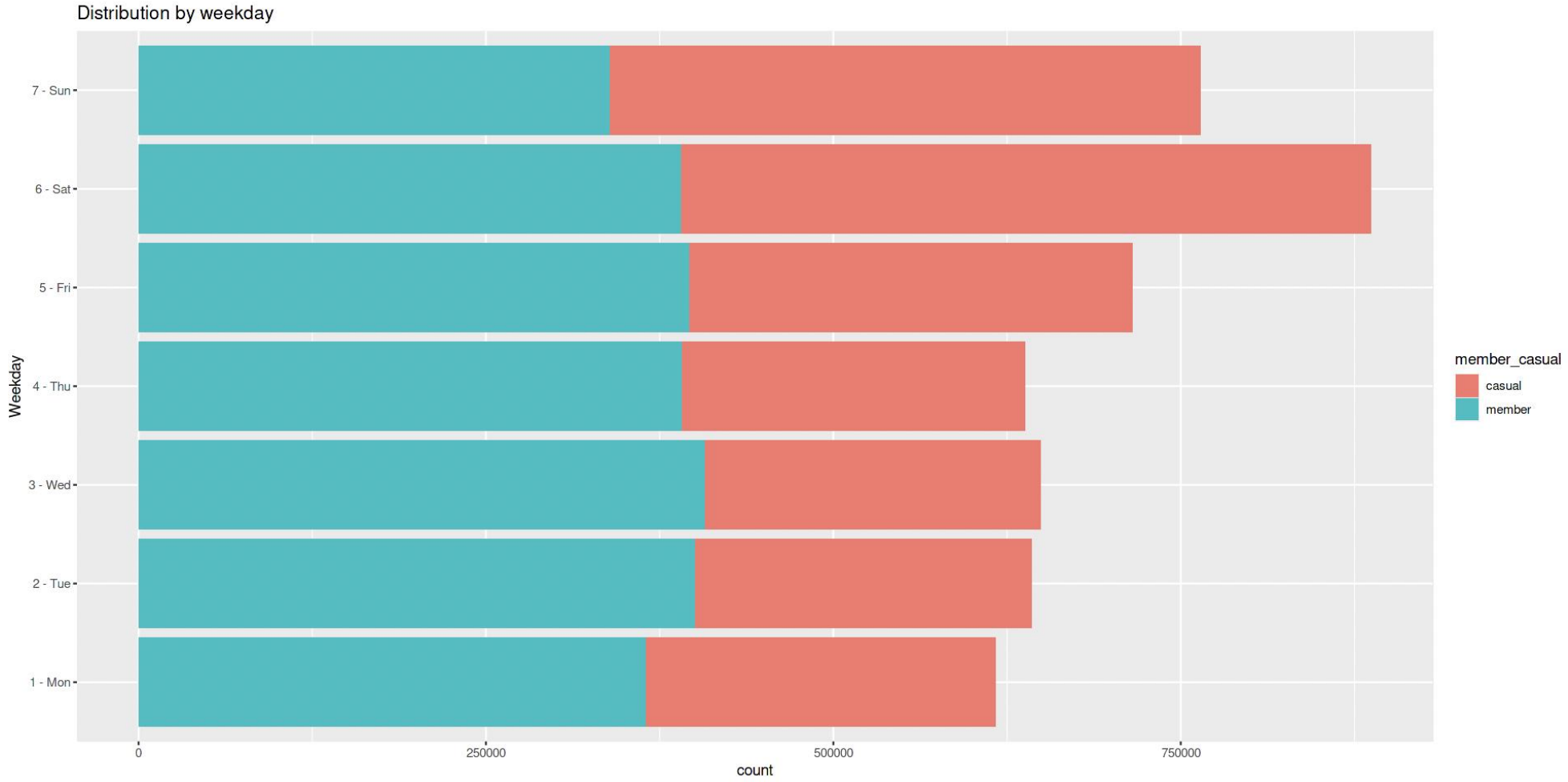
```
ggplot(new_bike_trips, aes(weekday)) +
  geom_bar(mapping= aes(fill = rideable_type)) +
  stat_count(geom="text",aes(label=..count..), size =3, position = position_stack(vjust = 1)) +
  facet_grid(~member_casual) +
  labs( title ="Number of ride per day", x = "Day of week", fill="Bike Type")
```



Continue...

[20]:

```
ggplot(new_bike_trips, aes(weekday, fill = member_casual)) +
    geom_bar() +
    labs(x = "Weekday", title = "Distribution by weekday") +
    coord_flip()
```



Continue...

Distribution by hour of the day

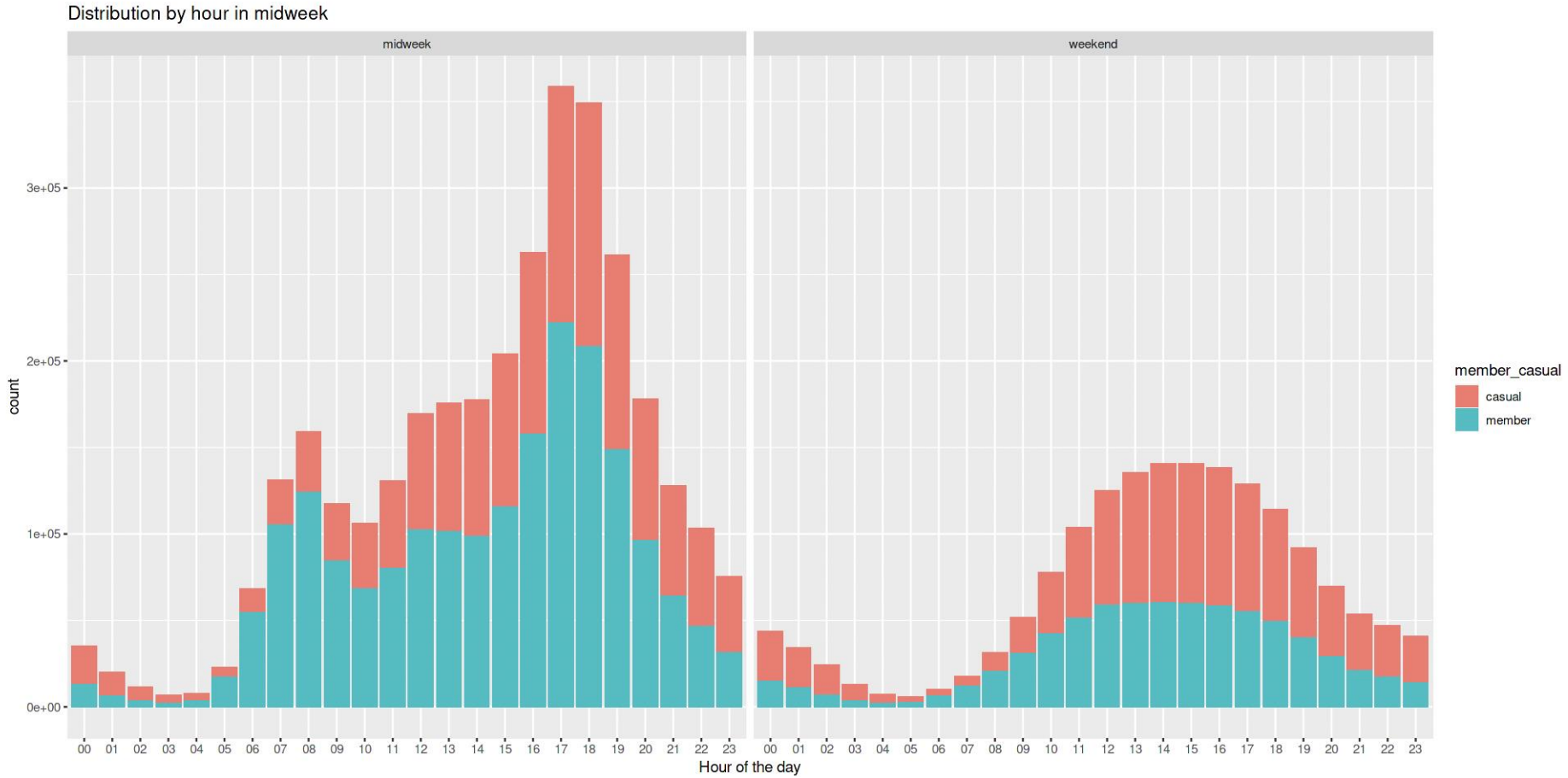
[36]:

```
new_bike_trips %>%
    ggplot(aes(start_hour, fill=member_casual)) +
    geom_bar() +
    labs(x="Hour of the day", title="") +
    facet_wrap(~weekday)
```

Continue...

Weekday vs Weekend

```
new_bike_trips %>%
    mutate(type_of_weekday = ifelse(weekday == '6 - Sat' | weekday == '7 - Sun',
                                    'weekend', 'midweek')) %>%
    ggplot(aes(start_hour, fill = member_casual)) +
    labs(x = "Hour of the day", title = "Distribution by hour in midweek") +
geom_bar() +
facet_wrap(~type_of_weekday)
```

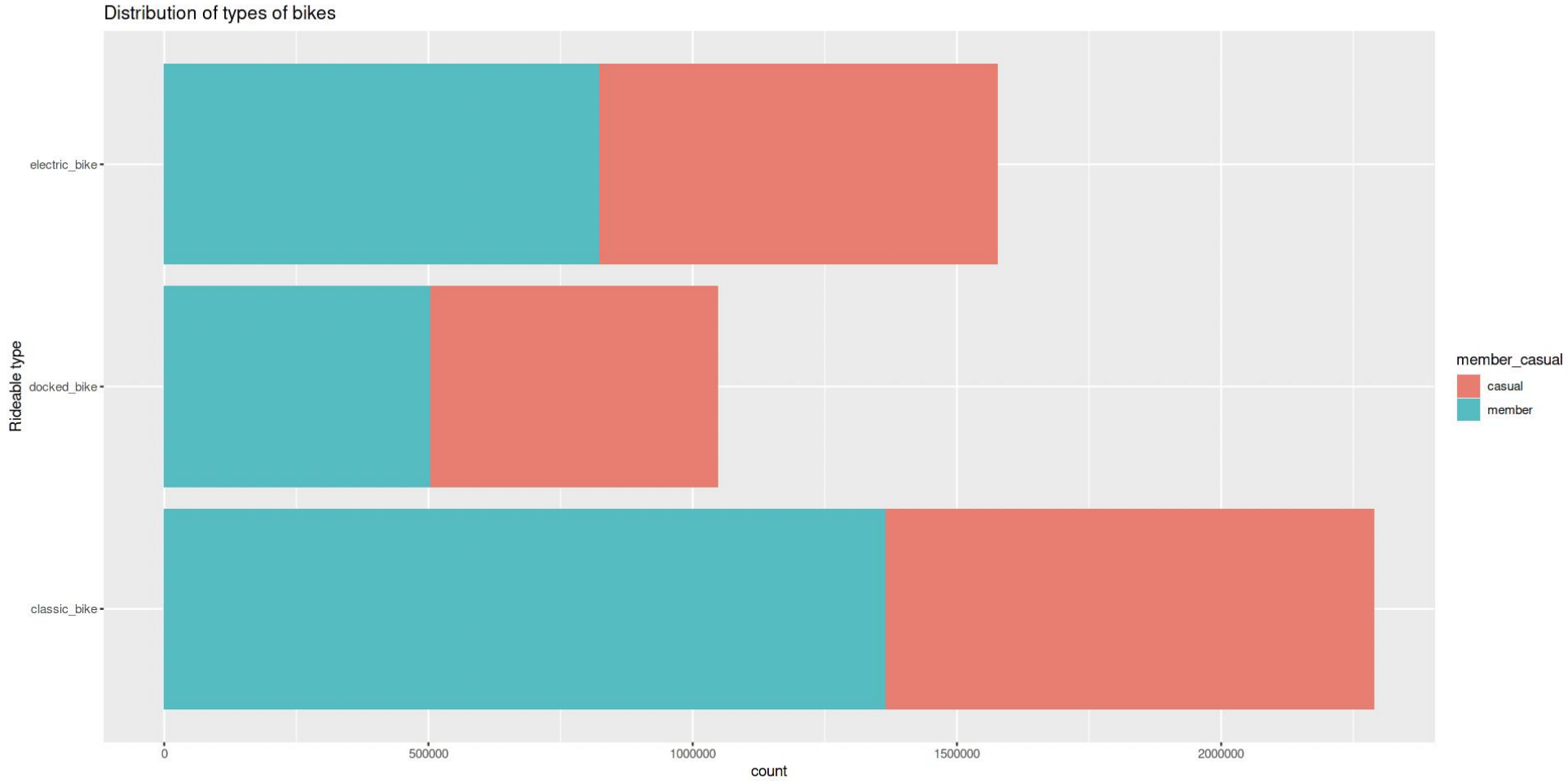


Continue...

Rideable type of bikes distribution

[30]:

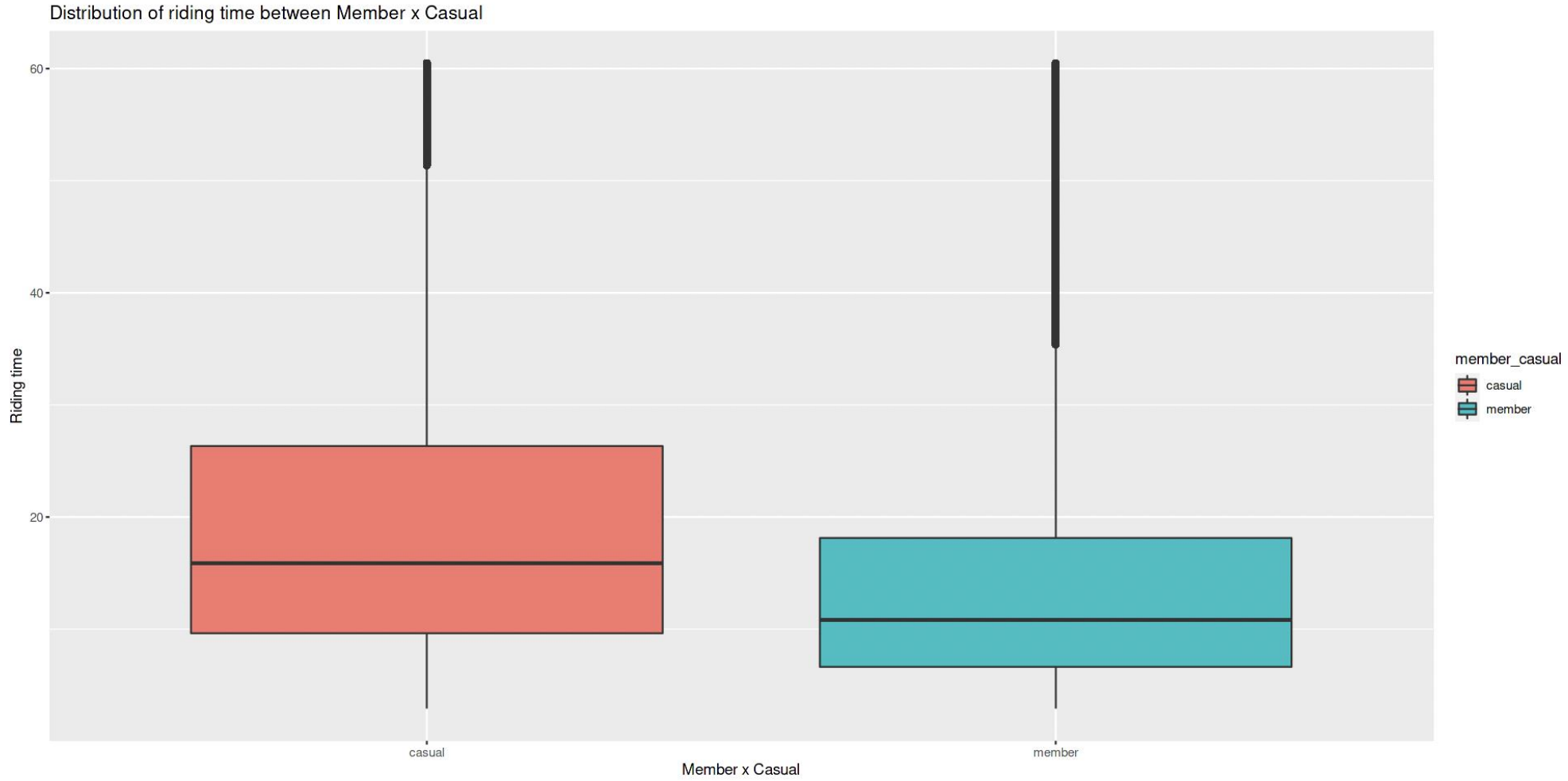
```
ggplot(new_bike_trips, aes(rideable_type, fill = member_casual)) +
    labs(x = "Rideable type", title = "Distribution of types of bikes") +
    geom_bar() +
    coord_flip()
```



Continue...

Length of ride between Causal vs Member

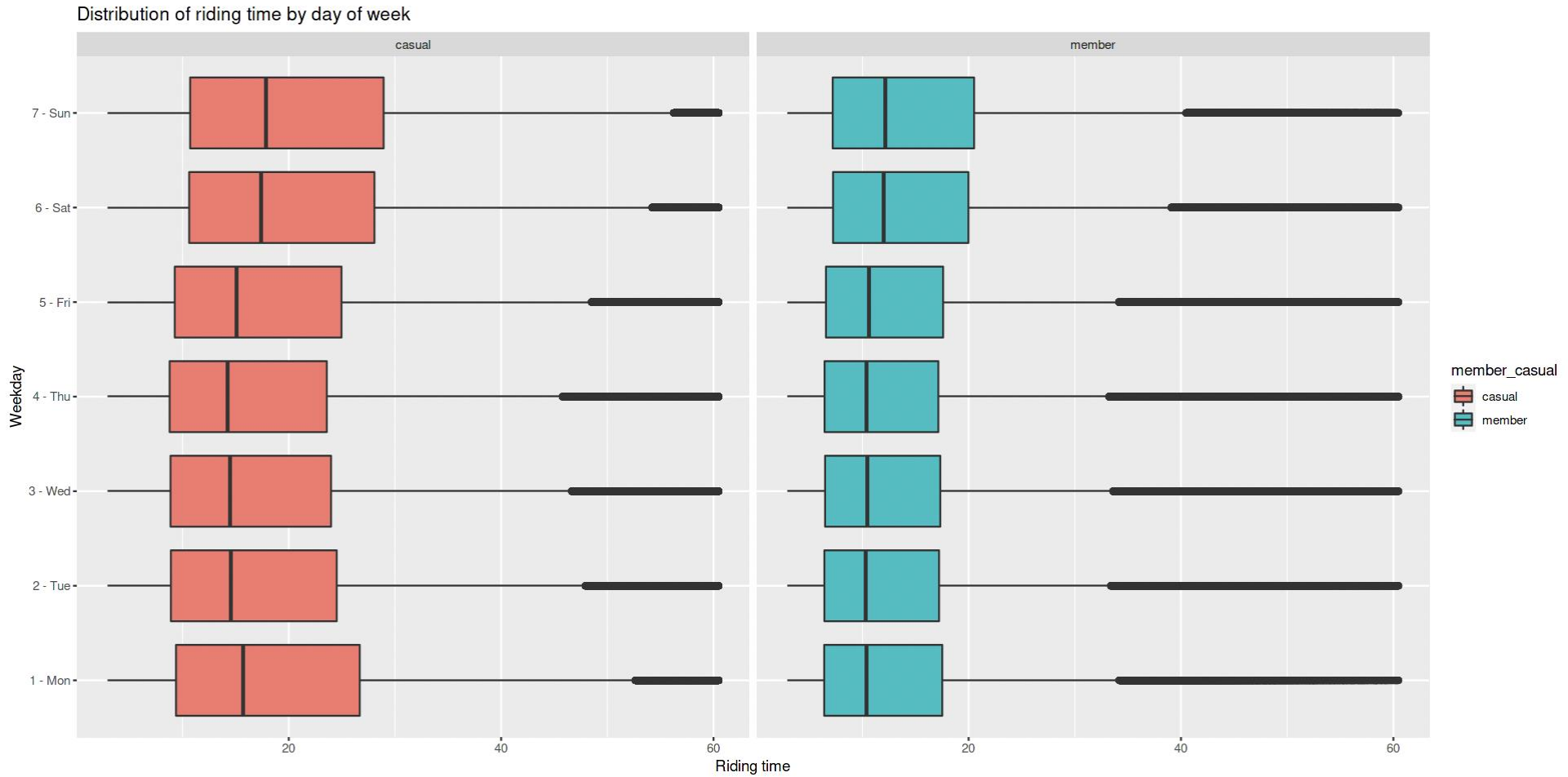
[39]:
```
ggplot(new_bike_trips_no_outliners, aes(x = member_casual, y = ride_time_m, fill = member_casual)) +
    labs(x = "Member x Casual", y = "Riding time", title = "Distribution of riding time between Member x Casual") +
    geom_boxplot()
```



Continue...

[41]:

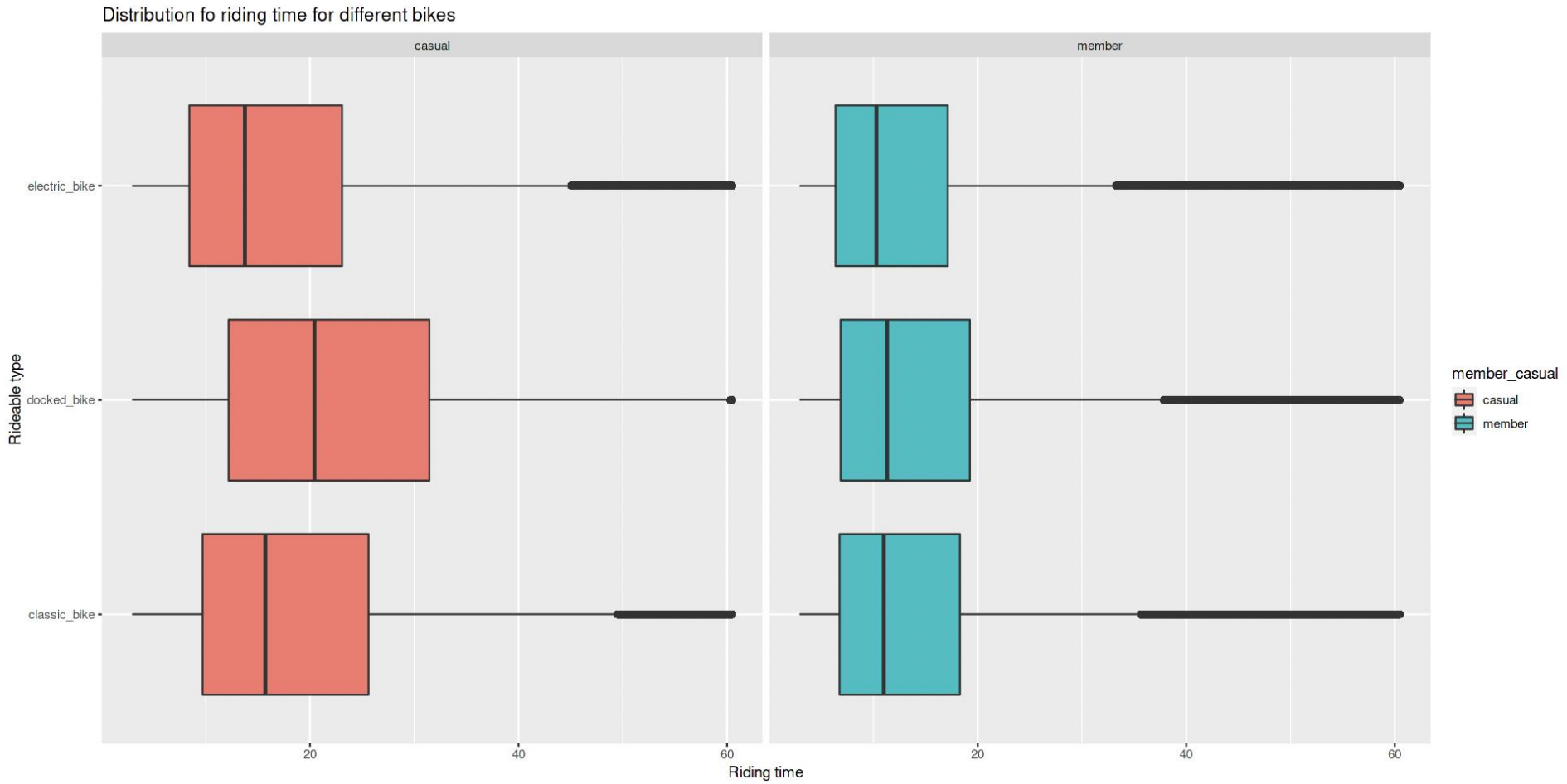
```
ggplot(new_bike_trips_no_outliners, aes(x = weekday, y = ride_time_m, fill = member_casual)) +
    geom_boxplot() +
    facet_wrap(~member_casual) +
    labs(x = "Weekday", y = "Riding time", title = "Distribution of riding time by day of week") +
    coord_flip()
```



Continue...

[43]:

```
ggplot(new_bike_trips_no_outliners, aes(x = rideable_type, y = ride_time_m, fill = member_casual)) +
    geom_boxplot() +
    facet_wrap(~member_casual) +
    labs(x = "Rideable type", y = "Riding time", title = "Distribution fo riding time for different bikes" ) +
    coord_flip()
```



Share

Let's go through the main finds and try to arrive at a conclusion:

What we know about the dataset:

- Members have the biggest proportion of the dataset, 19% bigger than casuals.
- There's more data points at the last semester of 2020.
- The month with the biggest count of data points was August with ~18% of the dataset.
- In all months we have more members' rides than casual rides.
- The difference of proportion of member x casual is smaller in the last semester of 2020.
- Temperature heavily influences the volume of rides in the month.
- The biggest volume of data is on the weekend.
- There's a bigger volume of bikers in the afternoon.

Continue...

Now for how members differs from casuals:

- Members may have the biggest volume of data, besides on Saturday. On this weekday, casuals take place as having the most data points.
- Weekends have the biggest volume of casuals, starting on Friday, a ~20% increase.
- We have more members during the morning, mainly between 5am and 11am. And more casuals between 11pm and 4am.
- There's a big increase of data points in the midweek between 6am to 8am for members. Then it fell a bit. Another big increase is from 5pm to 6pm.
- During the weekend we have a bigger flow of casuals between 11am to 6pm.
- Members have a bigger preference for classic bikes, 56% more.
- Casuals have more riding time than members.
- Riding time for members keeps unchanged during the midweek, increasing during weekends.
- Casuals follow a more curve distribution, peaking on Sundays and volleying on Wednesday/Thursday.

What we can take from this information is that members have a more fixed use for bikes besides casuals. Their uses is for more routine activities, like:

- Go to work.
- Use it as an exercise.

This can be proven we state that we have more members in between 6am to 8am and at 5pm to 6pm. Also, members may have set routes when using the bikes, as proven by riding time for members keeps unchanged during the midweek, increasing during weekends. The bikes is also heavily used for recreation on the weekends, when riding time increases and casuals take place.

Members also have a bigger preference for classic bikes, so they can exercise when going to work.

Concluding:

- Members use the bikes for fixed activities, one of those is going to work.
- Bikes are used for recreation on the weekends.
- Rides are influenced by temperature.

Continue...

Guiding Questions

- Were you able to answer the question of how annual members and casual riders use Cyclist bikes differently?

Yes. The data points to several differences between casuals and members.

- What story does your data tell?

The main story the data tells is that members have set schedules, as seen on chart 06 on key timestamps. Those timestamps point out that members use the bikes for routine activities, like going to work. Charts like 08 also point out that they have less riding time, because they have a set route to take.

- How do your findings relate to your original question?

The findings build a profile for members, relating to "Find the keys differences between casuals and annual riders", also knowing when they use the bikes helps to find "How digital media could influence them".

- Who is your audience? What is the best way to communicate with them?

The main target audience is my cyclist marketing analytics team and Lily Moreno. The best way to communicate is through a slide presentation of the findings.

- Can data visualization help you share your findings?

Yes, the main core of the finds is through data visualization.

- Is your presentation accessible to your audience?

Yes, the plots were made using vibrant colors, and corresponding labels.

Key tasks:

- Determine the best way to share your findings.
- Create effective data visualizations.
- Present your findings.
- Ensure your work is accessible.

Deliverables:

- Supporting visualizations and key findings

Act

Guiding questions

- What is your final conclusion based on your analysis?

Members and casual have different habits when using the bikes. The conclusion is further stated on the share phase.

- How could your team and business apply your insights?

The insights could be implemented when preparing a marketing campaign for turning casual into members. The marketing can have a focus on workers as a green way to get to work.

- What next steps would you or your stakeholders take based on your findings?

Further analysis could be done to improve the findings, besides that, the marketing team can take the main information to build a marketing campaign.

- Is there additional data you could use to expand on your findings?

Mobility data.

Improved climate data.

More information members.

Key tasks:

- Create your portfolio.
- Add your case study.
- Practice presenting your case study to a friend or family member.

Deliverables:

- Your top three recommendations based on your analysis

1 Build a marketing campaign focusing on show how bikes help people to get to work, while maintaining the planet green and avoid traffic. The ads could be show on professional social networks.

2 Increase benefits for riding during cold months. Coupons and discounts could be handed out.

3 As the bikes are also used for recreations on the weekends, ads campaigns could also be made showing people using the bikes for exercise during the weeks. The ads could focus on how practical and consistent the bikes can be.