# **Introduction**

This is my version of the Google Data Analytics Capstone - Case Study 1. The full document to the case study can be found in the [Google Data Analytics Capstone: Complete a Case Study](https://www.coursera.org/learn/google-data-analytics-capstone) course.

For this project this steps will be followed to ensure its completion:

* It will follow the steps of the data analysis process: Ask, prepare, process, analyse, share, and act.
* Each step will follow its own roadmap with:
  + Code, if needed on the step.
  + Guiding questions with answers.
  + Key tasks, as a checklist.
  + Deliverable, as a checklist.

# **Ask**

For the ask step, first let's get some context from the cyclistic document:

Case study statement

In this case study, I am taking the data of previous 12 months from 2021-01 to 2021-12 of cyclistic bike share analysis. The director of marketing believes the company’s future success depends on maximizing the number of annual memberships. From these insights, we will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations.

Characters and teams

Cyclistic: A bike-share program that features more than 5,800 bicycles and 600 docking stations. Cyclistic sets itself apart by also reclining bikes, hand tricycles, and cargo bikes, making bike-share more inclusive to people with disabilities and riders who can’t use a standard two-wheeled bike. The majority of riders opt for traditional bikes; about 8% of riders use the assistive options. Cyclistic users are more likely to ride for leisure, but about 30% use them to commute to work each day.

Lily Moreno: The director of marketing and your manager. Moreno is responsible for the development of campaigns and initiatives to promote the bike-share program. These may include email, social media, and other channels.

Cyclistic marketing analytics team: A team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy. You joined this team six months ago and have been busy learning about Cyclistic mission and business goals — as well as how you, as a junior data analyst, can help Cyclistic achieve them.

Guiding questions

* **What is the problem you are trying to solve?**

The main objective is to determine a way to build a profile for annual members and the best marketing strategies to turn casual bike riders into annual members.

* **How can your insights drive business decisions?**

The insights will help the marketing team to increase annual members.

Key tasks

* Identify the business task
* Consider key stakeholders.

Deliverable

* A clear statement of the business task.
* Find the keys differences between casual and members riders and how digital media could influence them

# **Prepare**

Google also provided their [own link](https://divvy-tripdata.s3.amazonaws.com/index.html) with the same dataset but expanded with more years and station descriptions.

Guiding questions

* **Where is your data located?**

The data is located in a kaggle dataset.

* **How is the data organized?**

The data is separated by month, each on it's own csv.

* **Are there issues with bias or credibility in this data? Does your data ROCCC?**

Bias isn't a problem, the population of the dataset is it's own clients as bike riders. And have full credibility for the same reason. And finally, it's ROCCC because it's reliable, original, comprehensive, current and cited.

* **How are you addressing licensing, privacy, security, and accessibility?**

The company has their own licence over the dataset. Besides that, the dataset doesn't have any personal information about the riders.

* **How did you verify the data’s integrity?**

All the files have consistent columns and each column has the correct type of data.

* **How does it help you answer your question?**

It may have some key insights about the riders and their riding style

* **Are there any problems with the data?**

It would be good to have some updated information about the bike stations. Also more information about the riders could be useful.

Key tasks

* Download data and store it appropriately.
* Identify how it’s organized.
* Sort and filter the data.
* Determine the credibility of the data.

Deliverable

* A description of all data sources used

The main data source is 12 months (Between January 2021 and December2021) of riding data provided by the Cyclistic company.

**Process**

This step will prepare the data for analysis. All the csv files will be merged into one file to improve workflow.

Code

The main dependency for the project will be tidyverse

**Concatenating**

All the csvs files will be concatenated into one dataframe.

**Data cleaning**

* Removing the duplicate files.
* Parsing the datetime columns.

**Manipulating the data**

New columns will help improve calculation time in the future.

* ride\_time\_m : Represents the total time of a bike ride, in minutes.
* Year\_month : Separate the year and the month into one column.
* Start\_hour : Getting the hour of the day also may be useful for intra day analysis.
* Weekday : The weekday will be useful to determine patterns of travels in the week.
* Saving the result as a CSV file.

Guiding questions

* **What tools are you choosing and why?**

I'm using R for this project, for two main reasons: Because of the large dataset and to gather experience with the language.

* **Have you ensured your data’s integrity?**

Yes, the data is consistent throughout the columns.

* **What steps have you taken to ensure that your data is clean?**

First the duplicated values where removed, then the columns where formatted to their correct format.

* **How can you verify that your data is clean and ready to analyze?**

It can be verified by this notebook.

* **Have you documented your cleaning process so you can review and share those results?**

Yes, it's all documented in this R notebook.

Deliverable

* Documentation of any cleaning or manipulation of data.

**Analyze**

The data exploration will consist of building a profile for annual members and how they differ from casual riders.

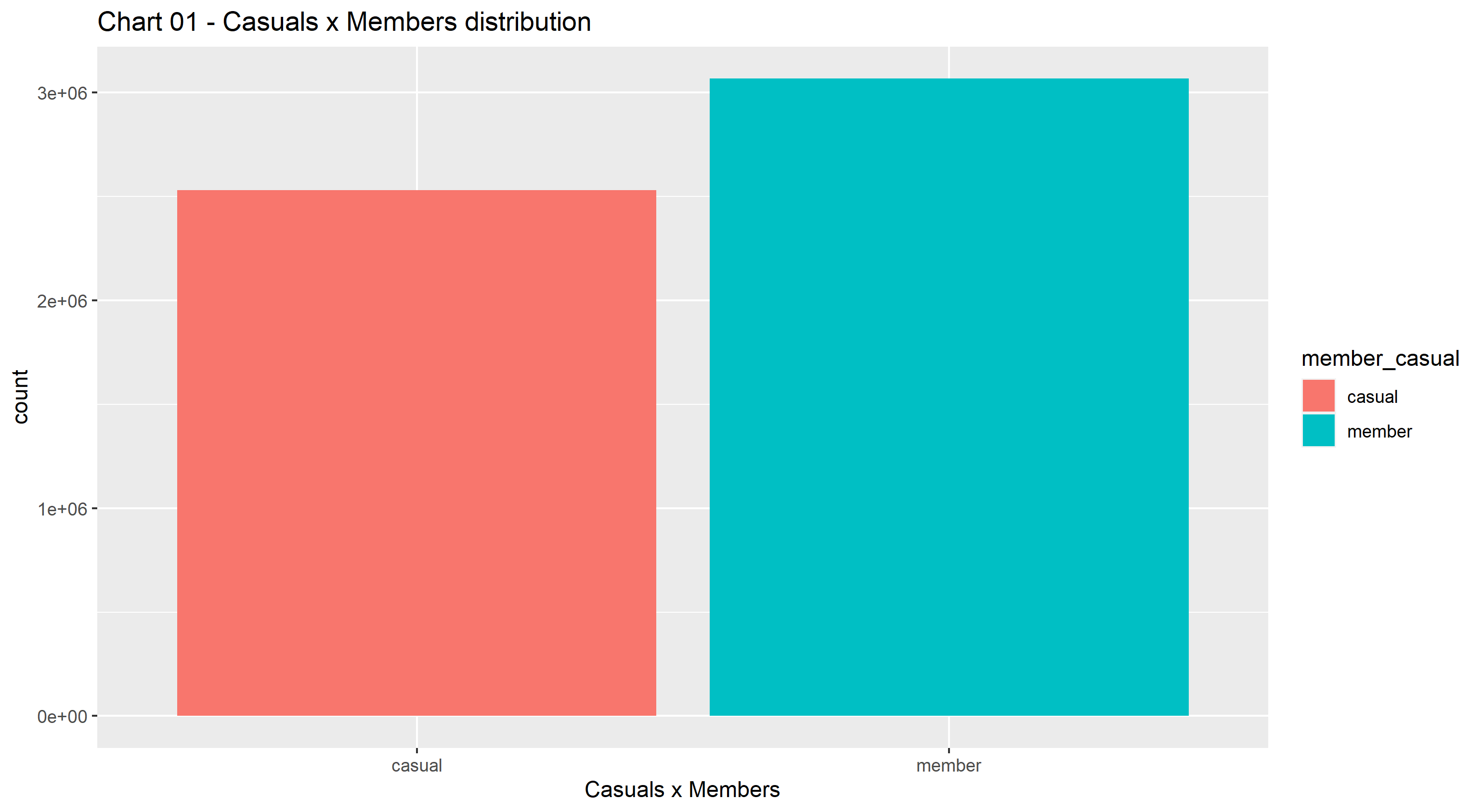
Putting in a new variable with a simpler name will help reduce some typing in the future.

**Code**

* The fig() function helps to resize the plots with variables like height and width.
* Summary() function helps to generate summary of the dataset.

### **Data distribution**

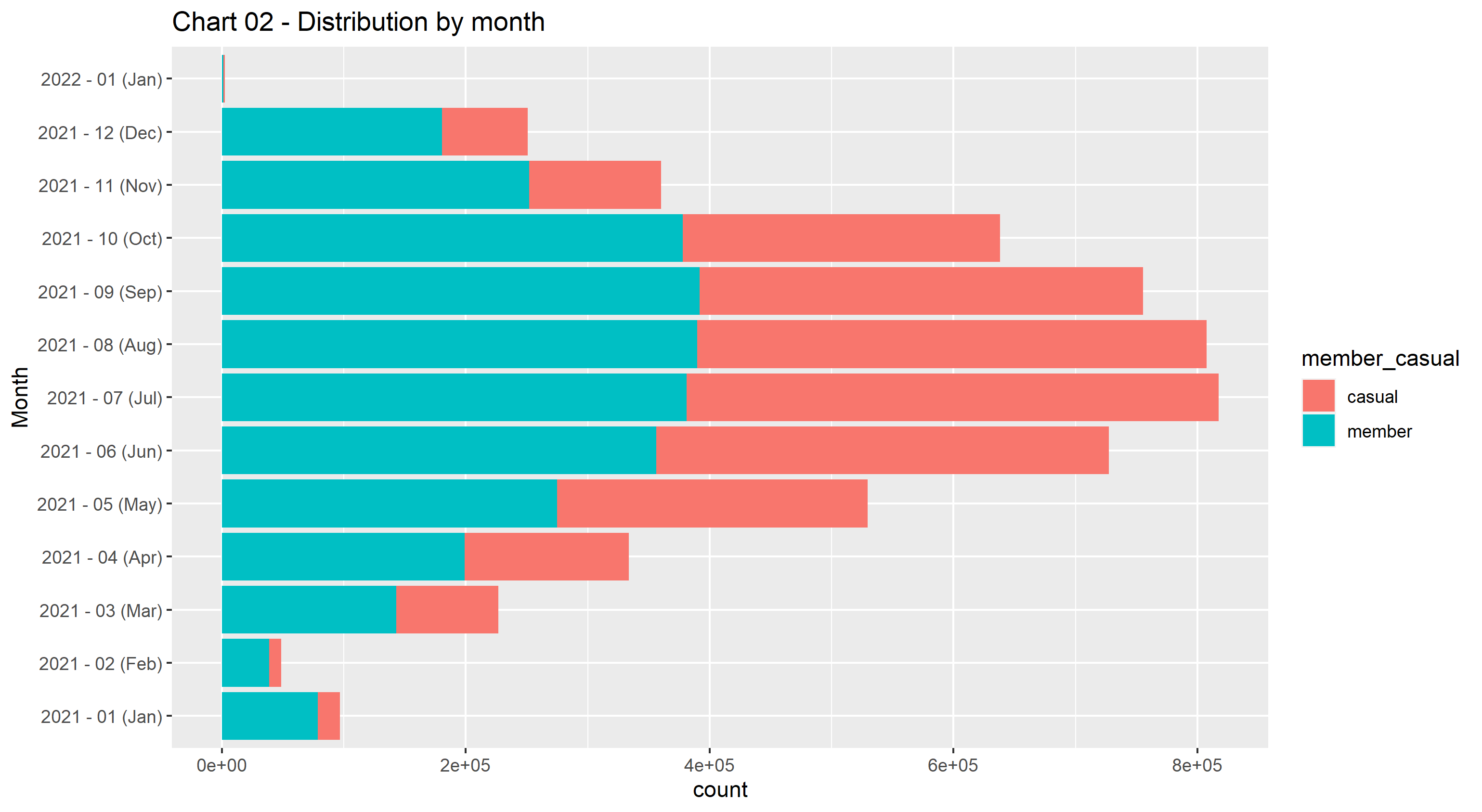
Here we can try to answer the most basic questions about how the data is distributed between casuals vs membership.



We can see members are bigger~ 55% of dataset than in casual riders.

**Month**

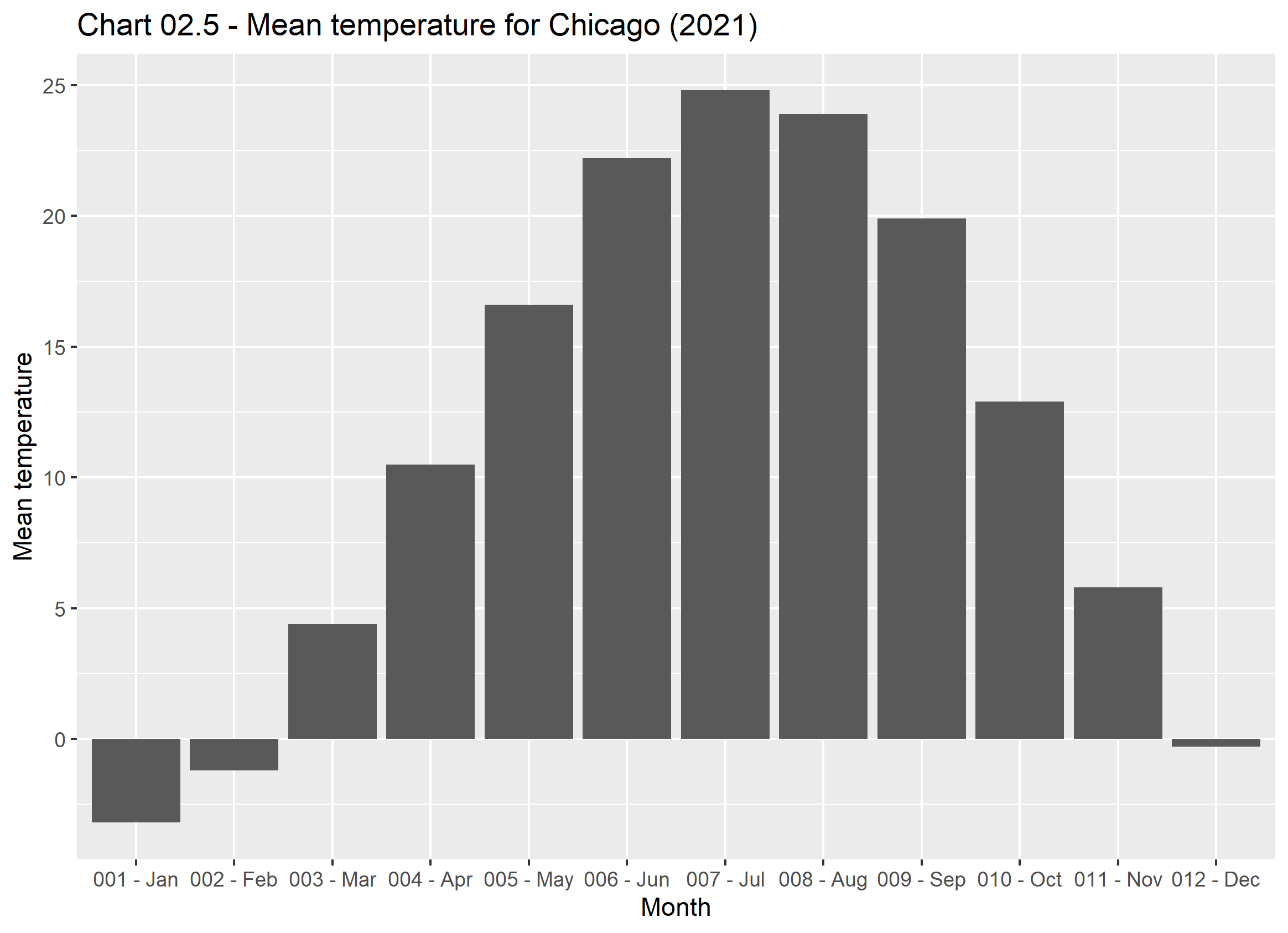
How much of the data is distributed by month?



Some considerations can be taken by this chart:

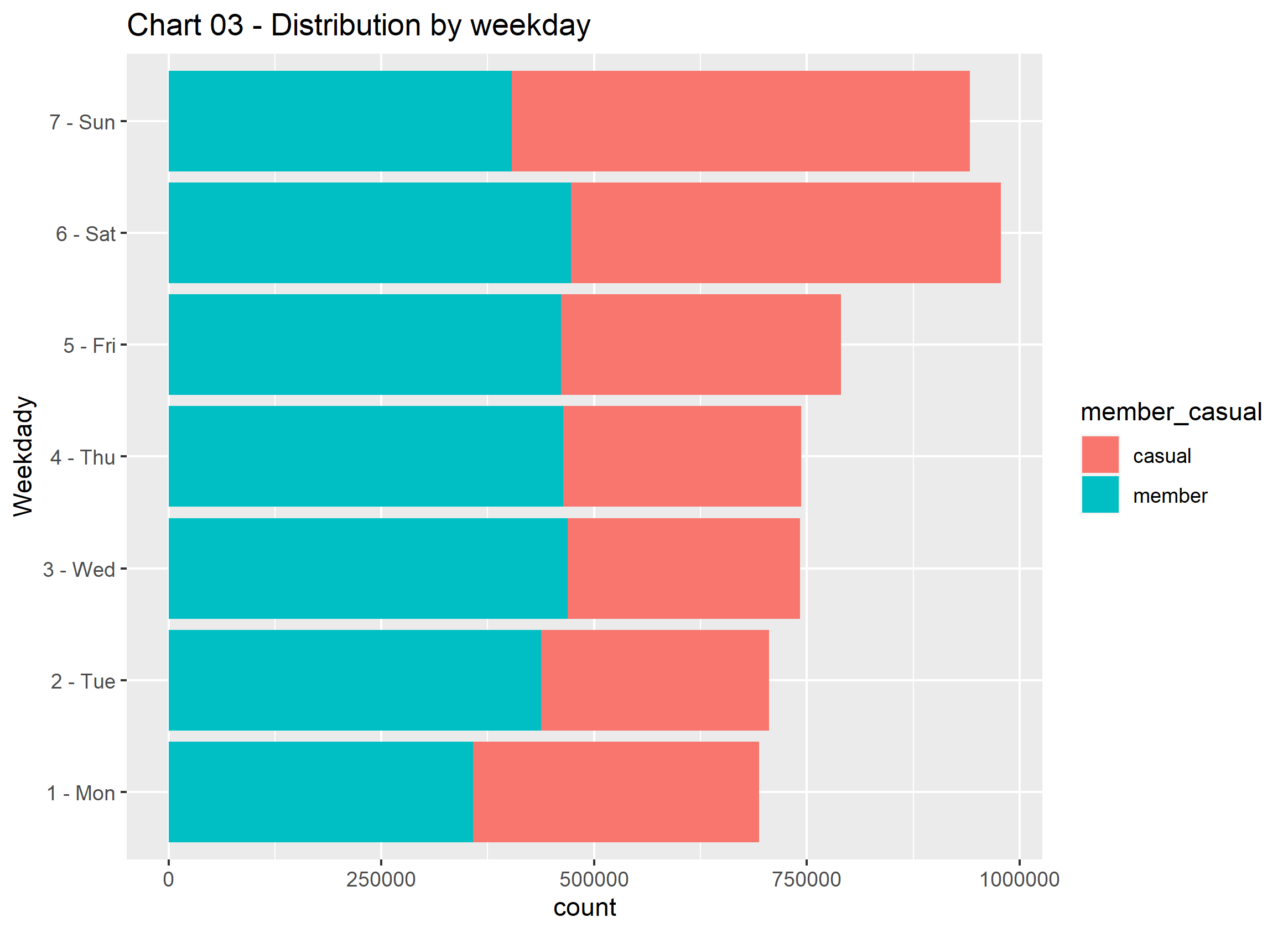
* The month with the biggest count of data points was July with ~14.6% of the dataset.
* In all months we have more members' rides than casual rides.

The data will be taken by [Climate of Chicago](https://en.wikipedia.org/wiki/Climate_of_Chicago) (Daily mean °C, 2021).

The data will be taken by [Climate of Chicago](https://en.wikipedia.org/wiki/Climate_of_Chicago) (Daily m

* Temperature heavily influence the volume of rides in the month.

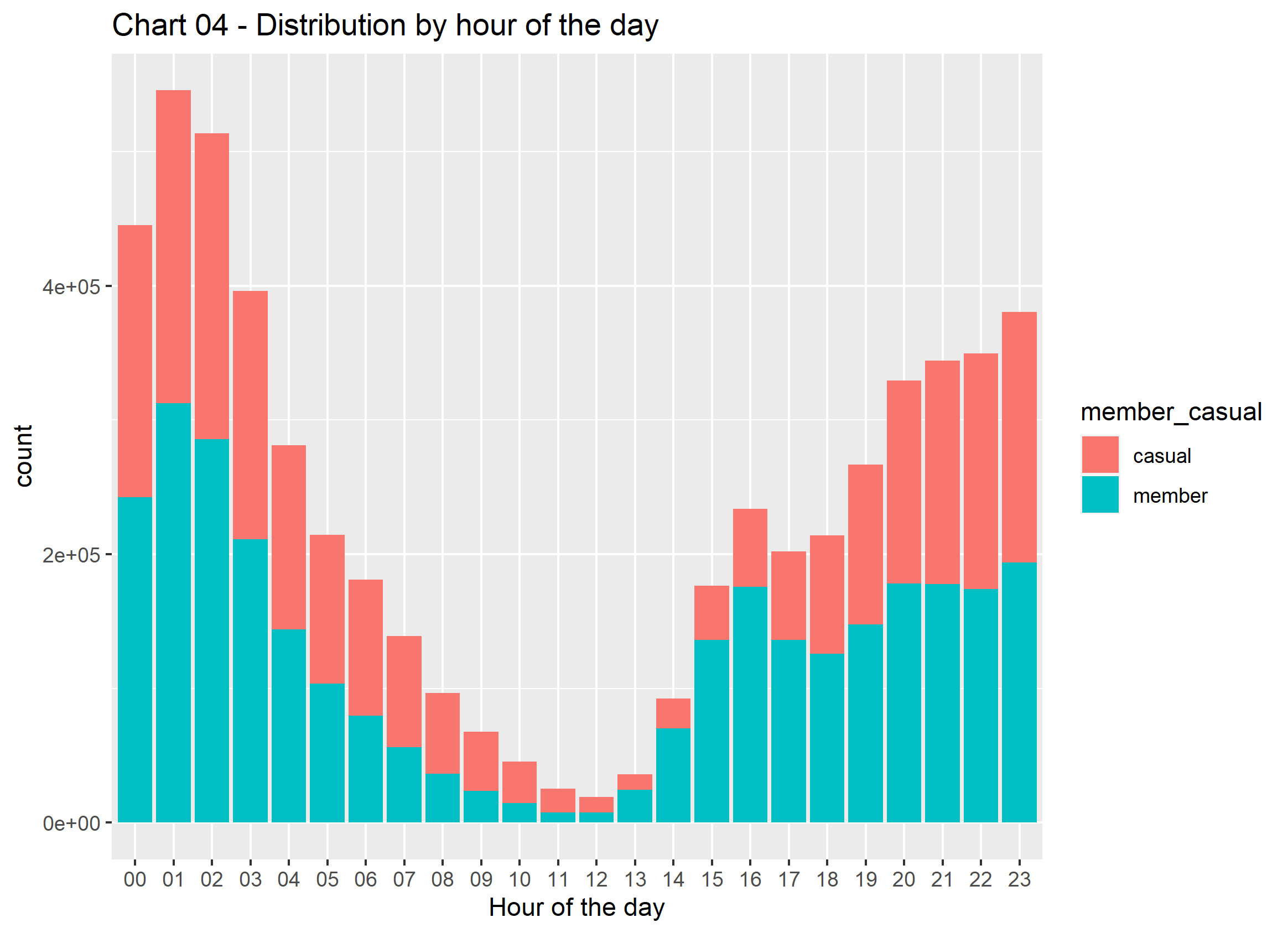
#### **Weekday**

How much of the data is distributed by weekday? 

The chart shows that

* The biggest volume of data is on the weekend.
* Saturday has the biggest data points.
* Members may have the biggest volume of data, on weekdays. On this weekend, casual take place as having most data points.

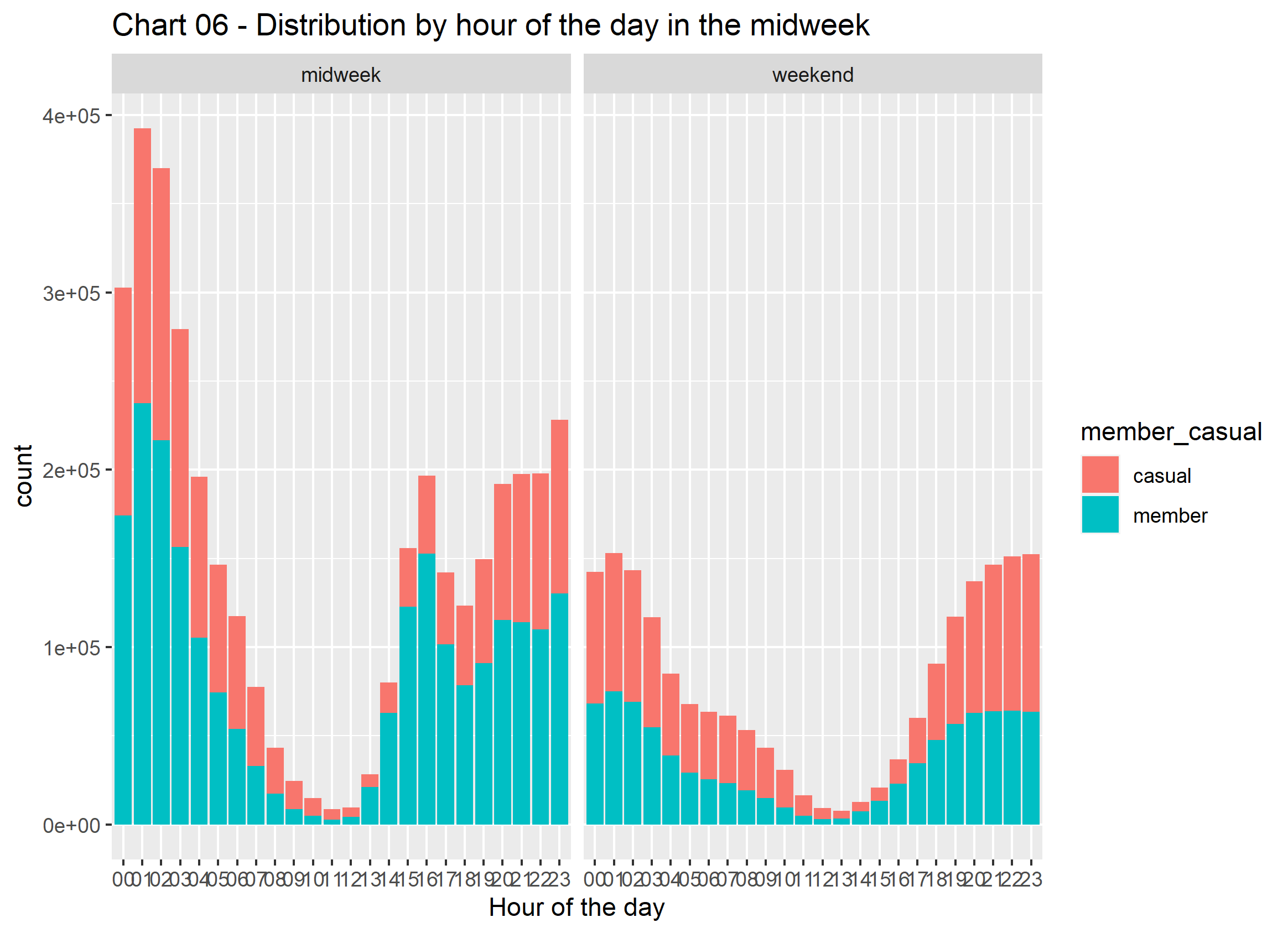
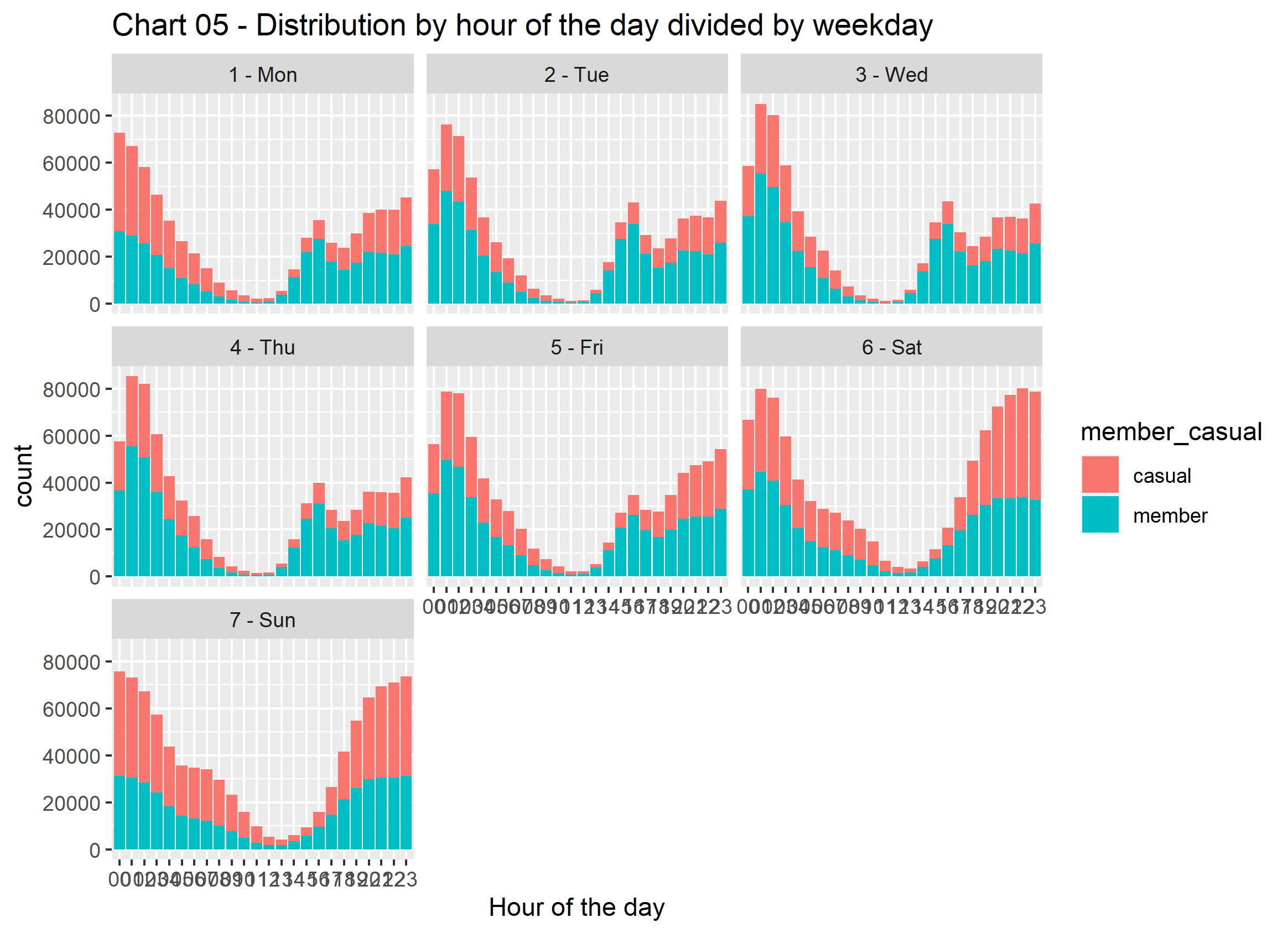
#### **Hour of the day**



From this chart, we can see:

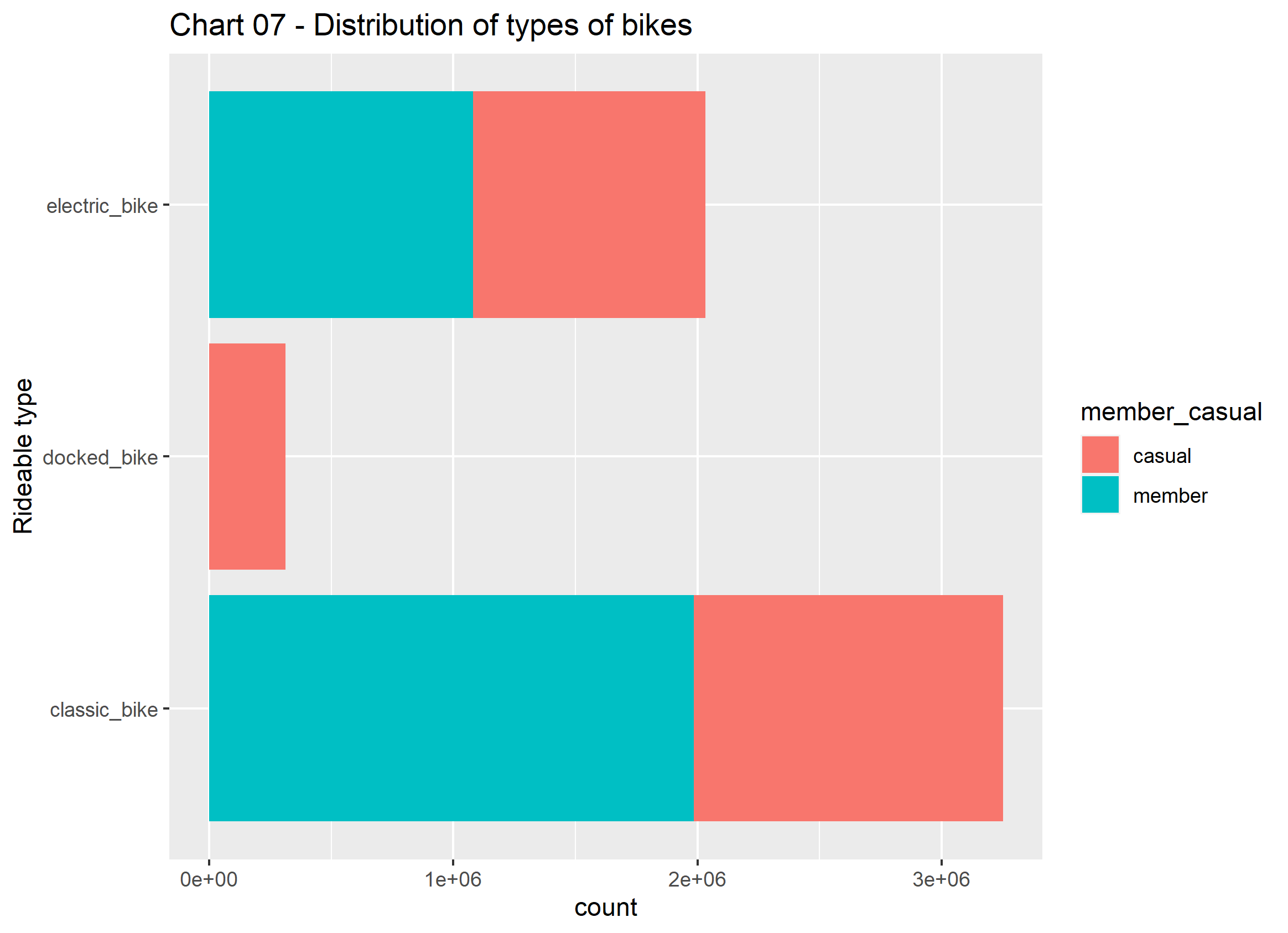
* There's a bigger volume of bikers in the morning.
* We have more members during the morning, mainly in between 01am and 4pm
* And more casuals between 10am and 11pm.

This chart can be expanded ween seen it divided by day of the week.

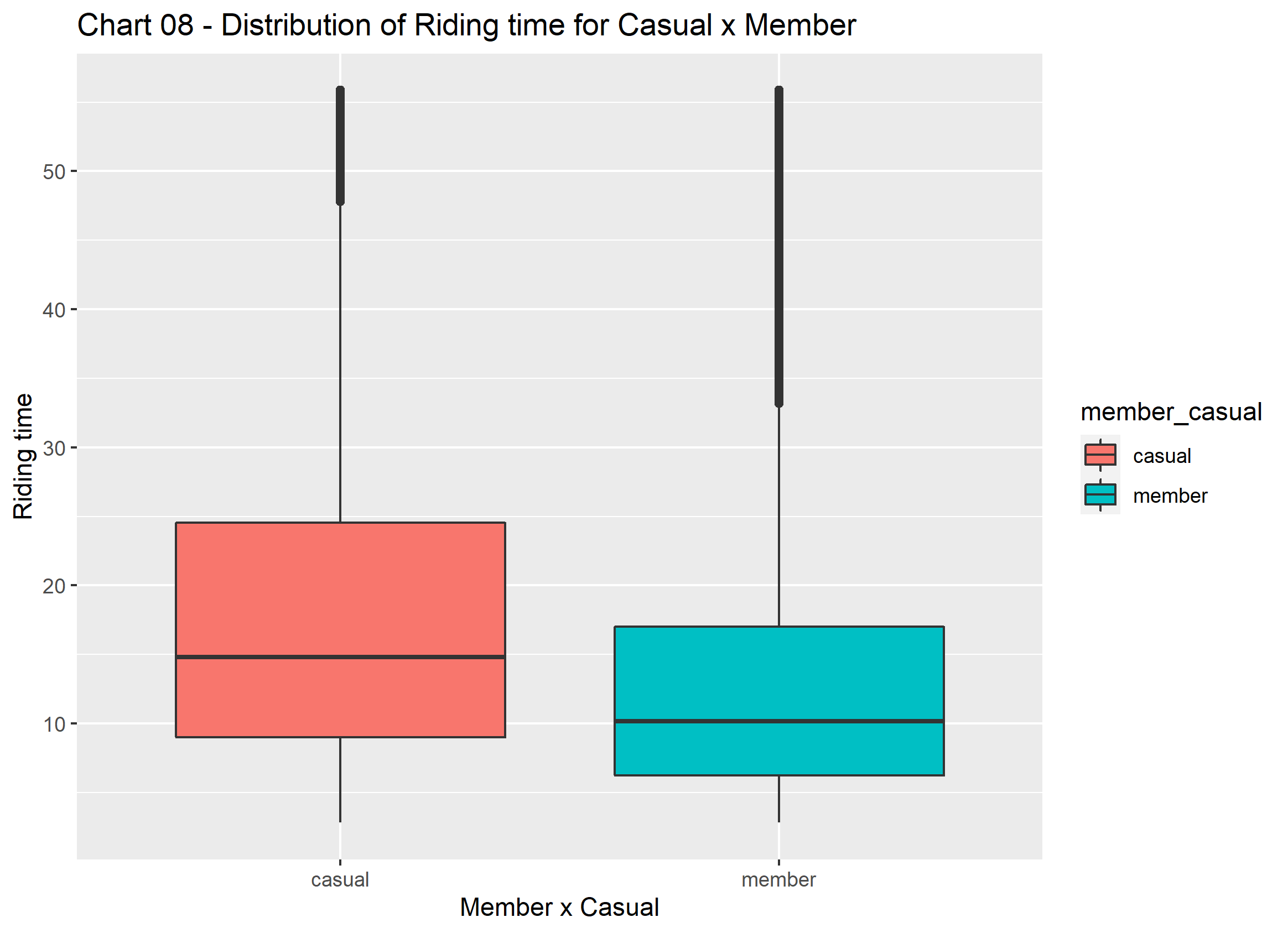
Here we can see the smooth outflow and step flow of data in charts.

* Casual riders increases from 10am to11pm in weekday
* Casual riders will be low between 11am to 1pm in weekend.

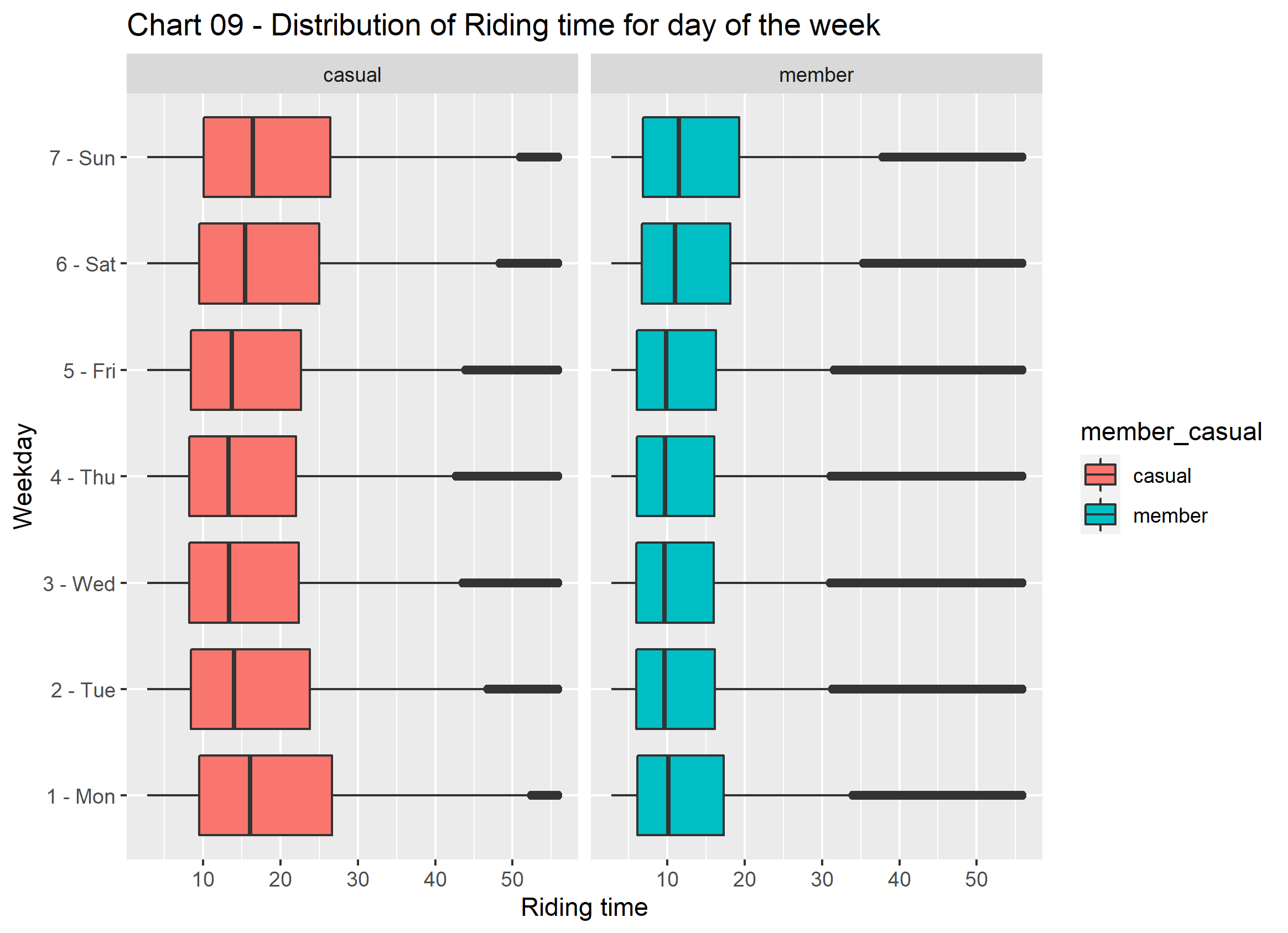
Rideable-type



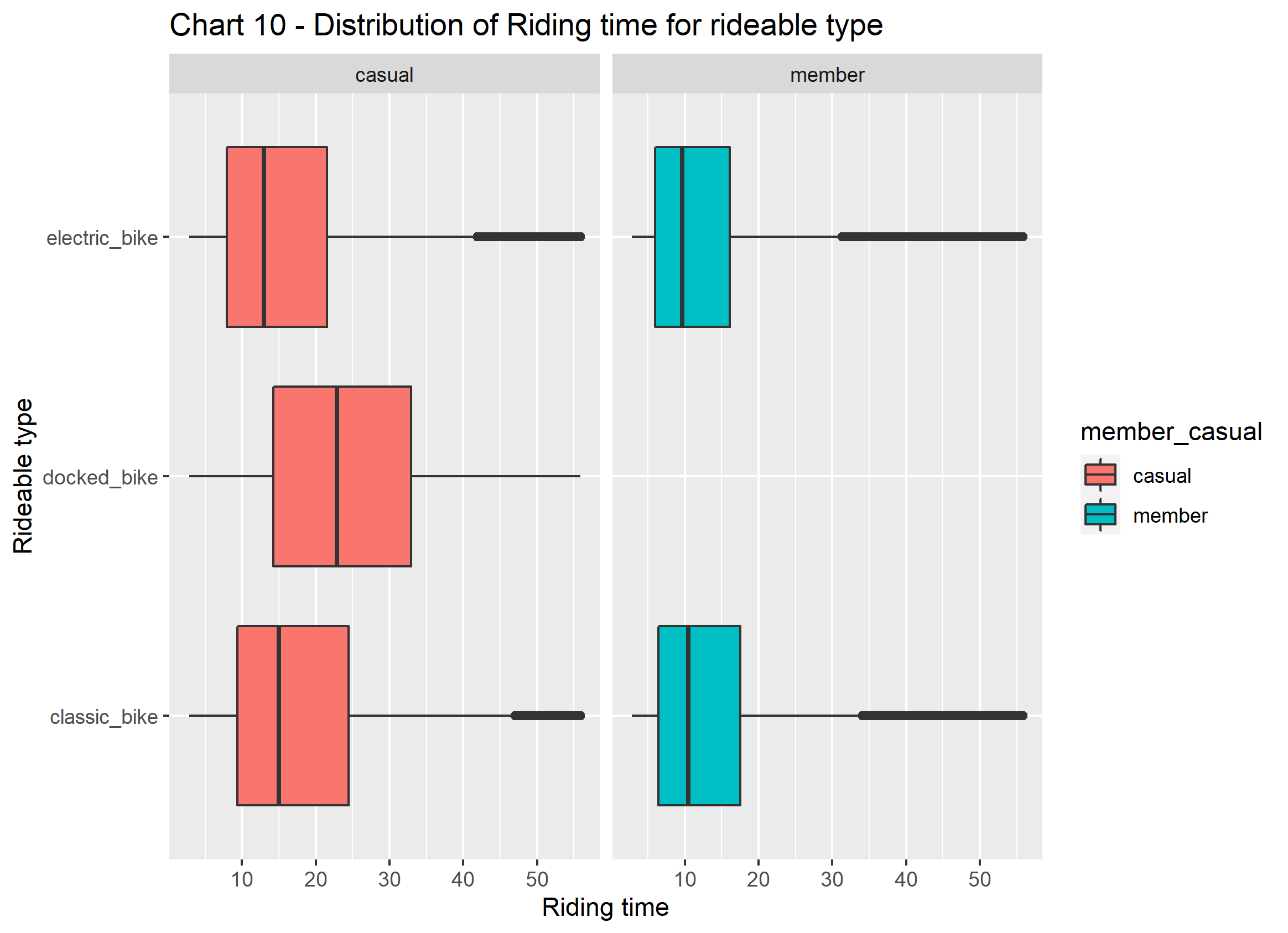
Members have high preference on classic bike 22% and also electric bike.



Casuals have higher riding time than members.



* + Casual riders peaks in Monday and valleying on Thursday/ Friday.
  + Membership riders keeps unchanged midweek and increase on weekend



* + - No members using docked bike.
    - Casual riders are using more classic\_bike and electric\_bike than members.