Cyclistic_CaseStudy

Leonardo Özuluca

2023-12-26

Case study: How does a bike-share navigate speedy success?

Scenario

You are a junior data analyst working on the marketing analyst team at Cyclistic, 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 Cyclistic bikes differently. From these insights, your team 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 offering 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 the bikes 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's mission and business goals, as well as how you, as a junior data analyst, can help Cyclistic achieve them.
- Cyclistic executive team: The notoriously detail-oriented executive team will decide whether to approve the recommended marketing program.

About the company

In 2016, Cyclistic launched a successful bike-share offering. Since then, the program has grown to a fleet of 5,824 bicycles that are geotracked and locked into a network of 692 stations across Chicago. The bikes can be unlocked from one station and returned to any other station in the system anytime.

Until now, Cyclistic's marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the flexibility of its pricing plans: single-ride passes, full-day passes, and annual memberships. Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members.

Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps Cyclistic attract more customers, Moreno believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, Moreno believes there is a solid opportunity to convert casual riders into members. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs.

Moreno has set a clear goal: Design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the team needs to better understand how annual members and casual riders differ, why casual riders would buy a membership, and how digital media could affect their marketing tactics. Moreno and her team are interested in analyzing the Cyclistic historical bike trip data to identify trends.

Ask

Three questions will guide the future marketing program:

1. How do annual members and casual riders use Cyclistic bikes differently?
2. Why would casual riders buy Cyclistic annual memberships?
3. How can Cyclistic use digital media to influence casual riders to become members?

Your task is to answer the first one and give suggestions for the rest.

Start of the Case Study

In this case study I am going to try to answer all of the questions in an informed and data driven way. But let's start at the beginning.

Data Source and structure.

The Data that was used was the Divvy_Trips_2020_Q1 Dataset. Upon receiving the data, it contained 426 887 rows, each of them representing one bike trip and is made publicly available here by Motivate International Inc. under the following licence.

It contains the following rows:

- ride_id: A unique identifier for each ride
- ridable_type: The type of bike that was used
- started_at: The time at which the trip has started
- ended_at: The time at which the trip has ended
- start_station_name: The name of the station where the trip has started
- start_station_id: The id of the station where the trip has started
- end_station_name: The name of the station where the trip has ended
- end_station_id: The id of the station where the trip has ended
- start_lat, start_lng: coordinates of the starting station.
- end_lat, end_lng: coordinates of the ending station.
- member_casual: This value is "member" if the trip was made by an subscribed member and "casual" if not.

Data cleaning

The first things I did was looking for problematic start/end times. I added a new row "duration" to the dataset that contained the duration of each ride.

Out of the 426 887 trips, 210 had a duration of <=0. This indicated a problem with either start or stop time for me. I took these rows out of the dataset.

Furthermore I wanted to make sure that only "real" rides are in the data. A ride that only went a couple of seconds is more likely to be an unlocking and relocking of the bike to make sure the bike is properly locked than an actual ride. I filtered out every ride that was shorter than 30 seconds.

Some of the formats in the start_lng column were broken, so I fixed them to make sure the formats are consistent.

```
Trips<- read_csv("Divvy_Trips_2020_Q1.csv")
```

```
## Rows: 426887 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (5): ride_id, rideable_type, start_station_name, end_station_name, memb...
## dbl  (6): start_station_id, end_station_id, start_lat, start_lng, end_lat, e...
## dttm (2): started_at, ended_at
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
Trips<-mutate(Trips, trips_duration=ended_at-started_at)
Trips <- Trips %>% filter(trips_duration>=30)
```

There were no dublicates, broken formats or other problems left that I found. So I proceeded to the next step. The dataset was now left with 230 283 rows.

Data preparation

I knew that I was not allowed to use spacial information, so I took them out from the beginning. I took out the columns Start_lat, start_lng, end_lat and end_lng. I also knew that the station names would not be necessary for this analysis, so I dropped them as well. I selected only the rest.

```
Trips <- Trips %>% select(ride_id, rideable_type, started_at, ended_at, start_station_id, end_station_id
```

After dropping unnecessary columns, the next step was to precalculate information that would be needed. I wanted to include the weekday of the trip into the data, as well as a marker if it is a weekend day or a workday.

```
wd <- Trips %>% mutate(weekday=weekdays(started_at))
Trips <- rbind(wd %>%  filter(weekday=="Samstag"|weekday=="Sonntag") %>% mutate(weekend="weekend"), wd %
```

Afterwards I also wanted information about the time that the trips were made at. The started_at column contains information about day, time and seconds. I only need the hour. But I also wanted to create buckets of 6h timeframes to analyse later, in which timeframes how many rides were done.

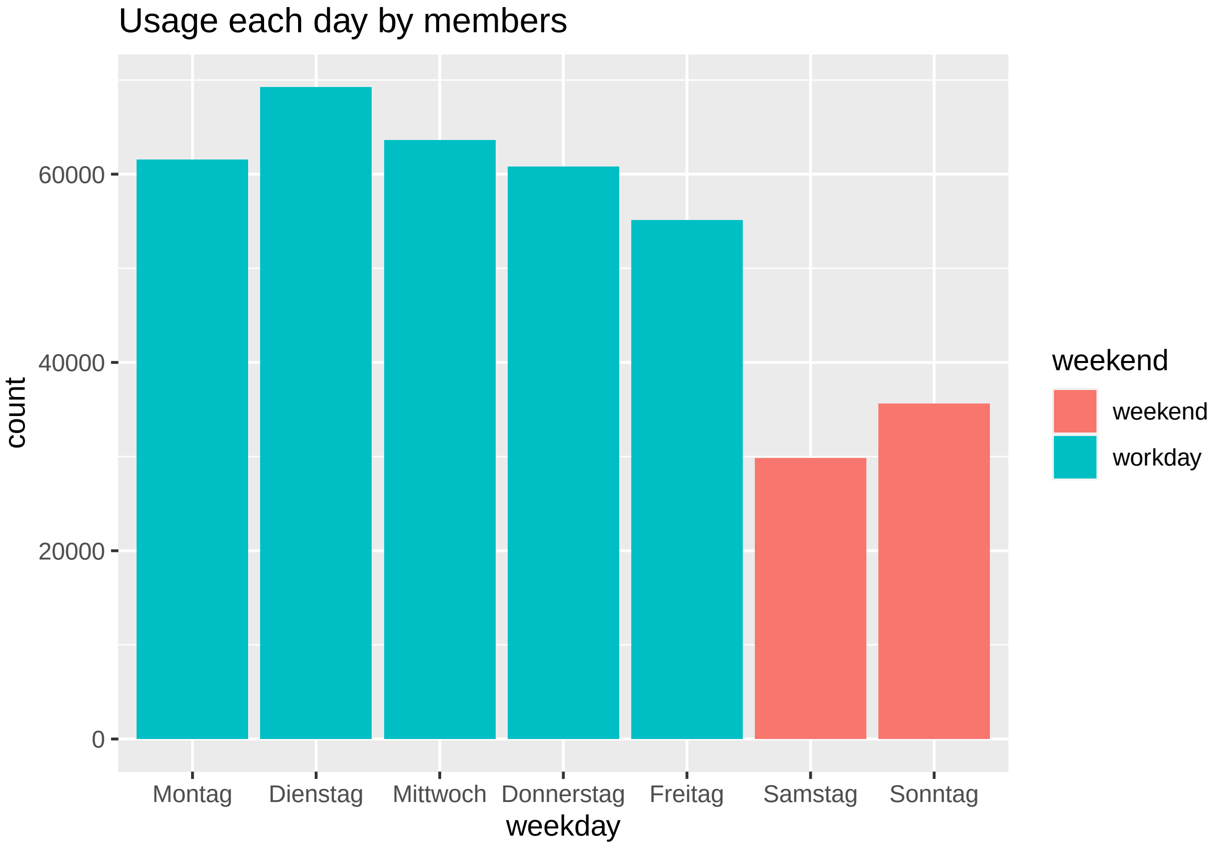
```
wehours<-mutate(Trips, hour=format(started_at, format = "%H"))
Trips<- wehours %>% mutate(hourbin=floor(as.numeric(hour)/6))
```

Analysis

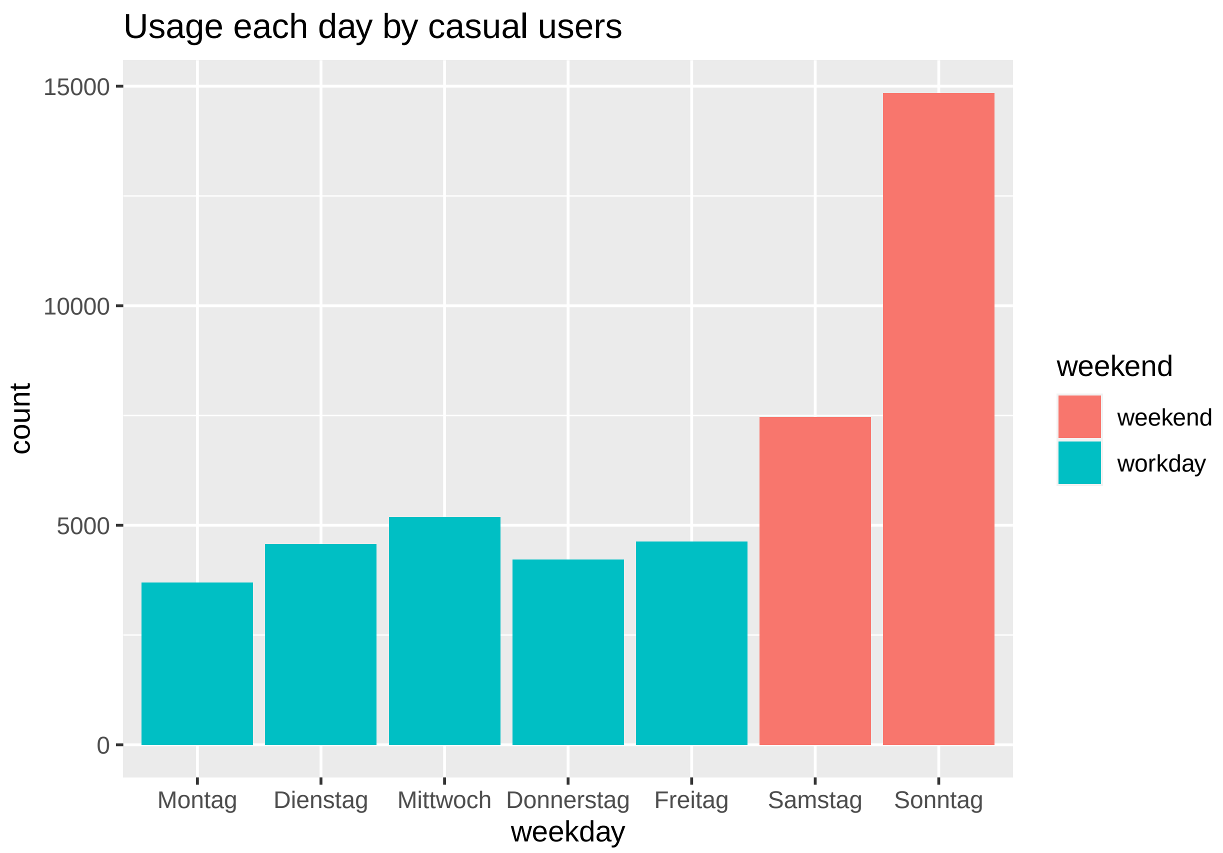
Now that the Data is clean and prepared, the next step is the actual analysis. Just as a reminder, the question we try to answer is:

How do annual members and casual riders use Cyclistic bikes differently?

The first thing I wanted to look at was on which days the bikes are used. I thought this might give us a hint on how they are used. I looked at the members first.



This shows very clearly that the members use the Cyclistic service more during the week and less during the weekend. Then I had a look at the casual users.



Here the picture was also very clear. The Cyclistic service is used by casual users significantly more during the weekend, especially the Sunday than during the workdays. This alone is not enough to form any conclusions, but it was enough to form a working theory:

=>members use Cyclistic more for regular commuting to work while casual users use the service more for leisure purposes

Now that I had a working theory, I tried to refute or disprove it. Thinking about how to do that I came to the following possibilities:

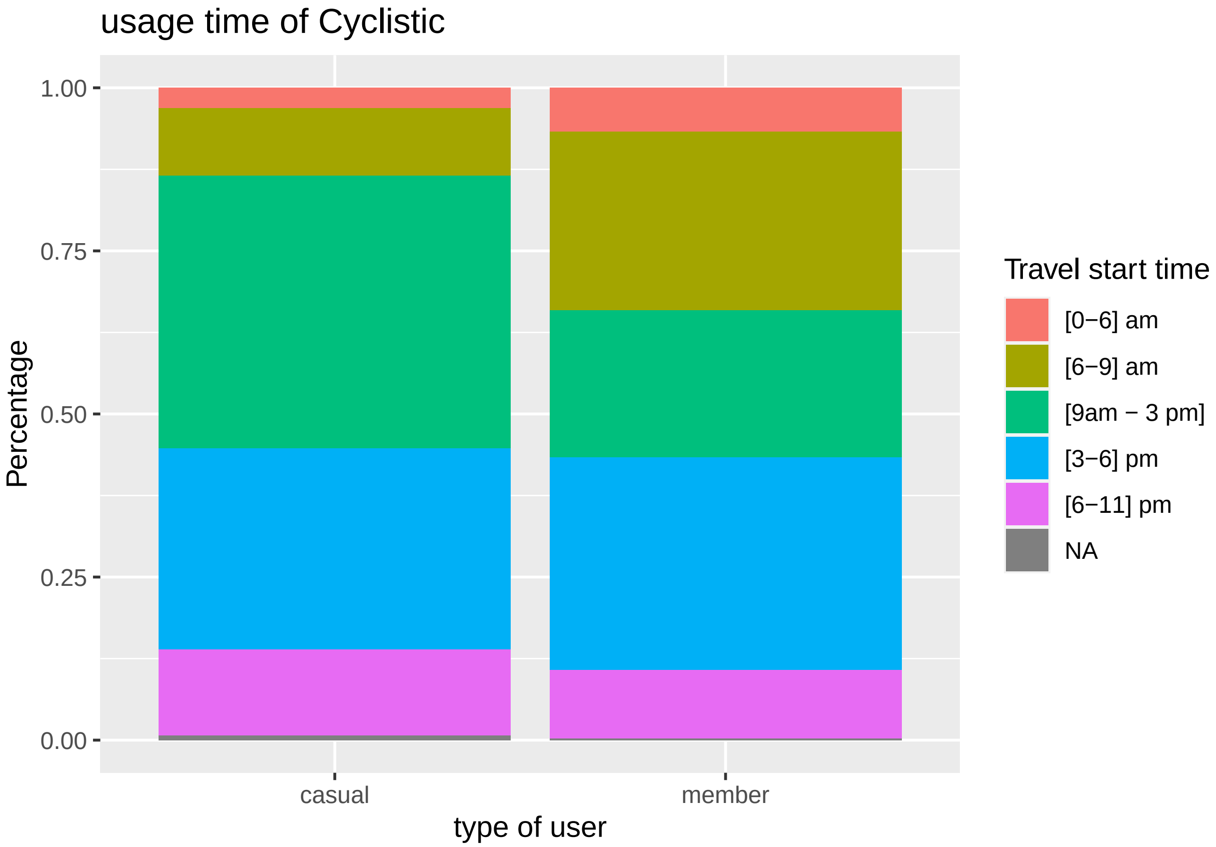
- Time: If the timeframe in which the service is used the most is during the normal working hours, then maybe commuting to work is not the reason for the usage. If they would be used a lot outside of working hours but only after work, maybe a lot of people use them for after work excursions or for other leisure reasons.
- Duration: If the trips are very long, it might be unlikely that they commuting but rather excursions or bike-tours. It might be more reasonable for short trips to be trips to "get somewhere" like work, school or university while long trips are probably more for fun and leisure purposes. A lot of long trips during the work days could refute our working theory.
- start and end: If we want to refute our theory that members use the bikes for commuting to work we might look at the start and end station. If a ride ends at the same station as it started, it is a round tour and was probably not used for commuting. If it ends somewhere else, then Cyclistic was probably used to get somewhere.

Let's go through them in an ordered manner:

1. Time

As explained above, we might be able to refute our theory if we could show that the Cyclistic members use the service the most in the normal working hours. This might indicate that Cyclistic was not used to commute to work. Another option would be if the usage is high after working hours but not before working hours. Then it might be that the bikes were used for leisure purposes and not for commuting purposes.

To do this, I first filtered the data for workdays (Monday-Friday) so that we only get the relationship between working time and Cyclistic usage time. As working times and therefore commuting times differ from person to person I decided to take the 6h timeframe from 9am to 3pm as "working time" while the 3h timeframes from 6-9am and from 3-6pm I considered normal times for commuting. This is the result I got:

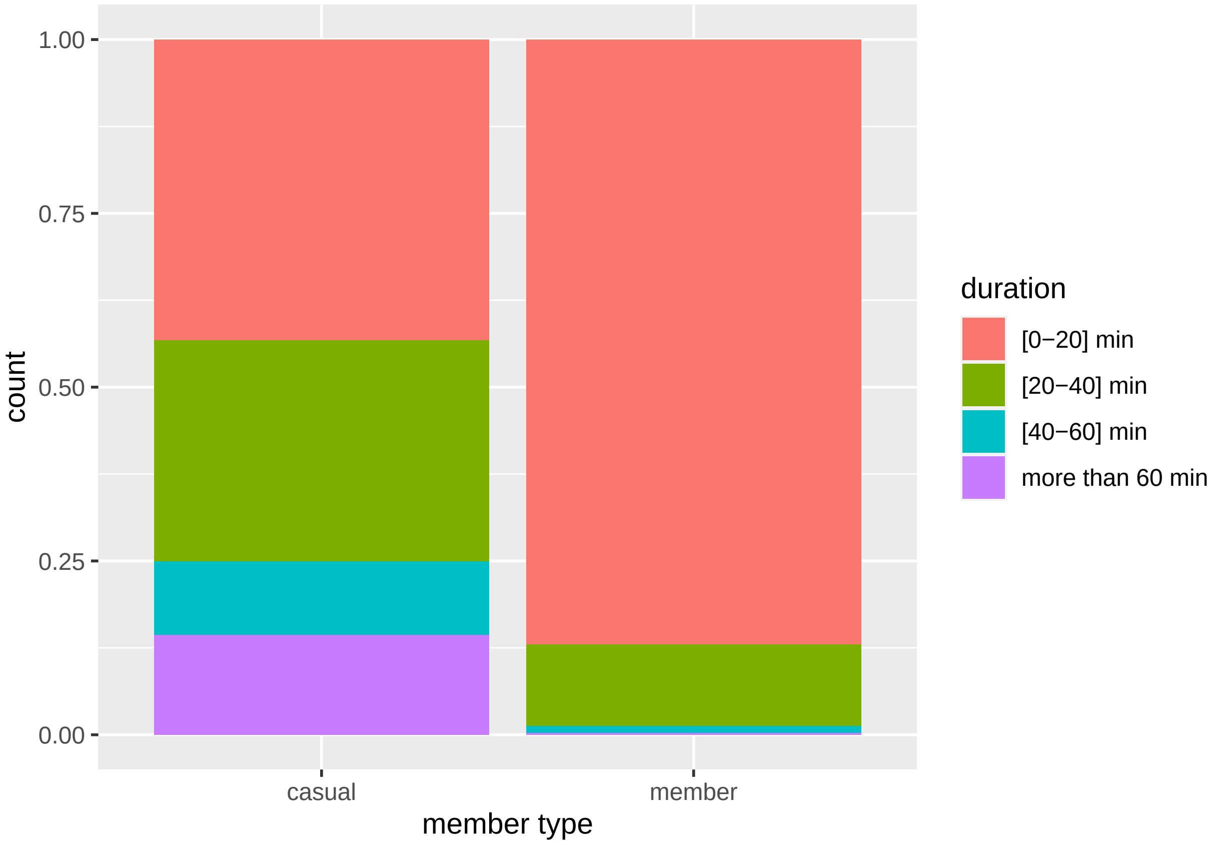


In the result it is visible, that casual users and members use the Cyclistic service differently. For the casual users we see that a big portion is during the normal worktime. The usage before and after worktime is very asymetric. There are significantly more users that use Cyclistic after work than users that use it before work. This might indicate that the bikes are not used for commuting but for after work leisure.

For the members we can see a very different picture. The usage during normal working time is rather low, the highest usage is both before and after working time. This might be because Cyclistinc is used for commuting. Therefore we are not able to refute our theory so we proceed to the next point.

2. Duration

The next thing I want to have a look at is the duration of the rides. Most people will not commute by bike for more than an hour or do a bike-tour that is shorter than 20 minutes. Different usages usually take different durations. So if we have a look at the duration of the rides, we might be able to infer a difference in usage. To do this I divided the durations in four categories. The short rides up two 20 minutes, a medium category [20-40] min and two categories for long rides [40-60] min and more than 60 minutes.



As we can see in the graph, the usage of members is mostly within 20 minutes. It alone makes roughly 75% of all of the rides. Including the medium long rides of [20-40] minutes will fill up nearly all of the rides.

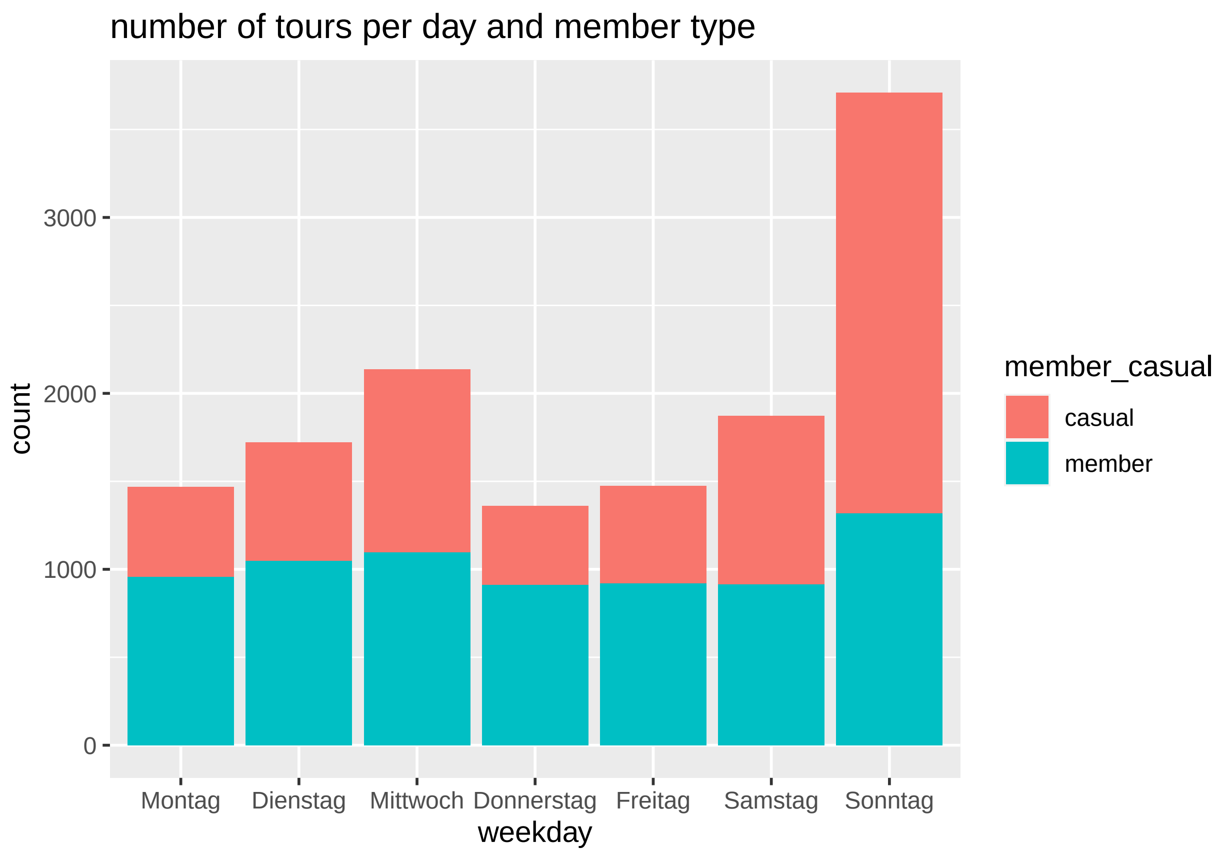
For the casual rides this looks very differently. not only do the long rides of over 40 minutes take up nearly a quarter of all of the casual rides, but also the [20-40] minute long rides are more. Casual rides seem to use Cyclistic significantly less for short distances and more for long.

All in all I would say that we can not infer anything new from this. According to our working theory members use the Cyclistic service mostly for commuting, which fits to their usage durations, while casual members use it more for leisure and fun purposes and less for commuting. Therefore the longer durations on average fit to our current working theory.

3. Start and End

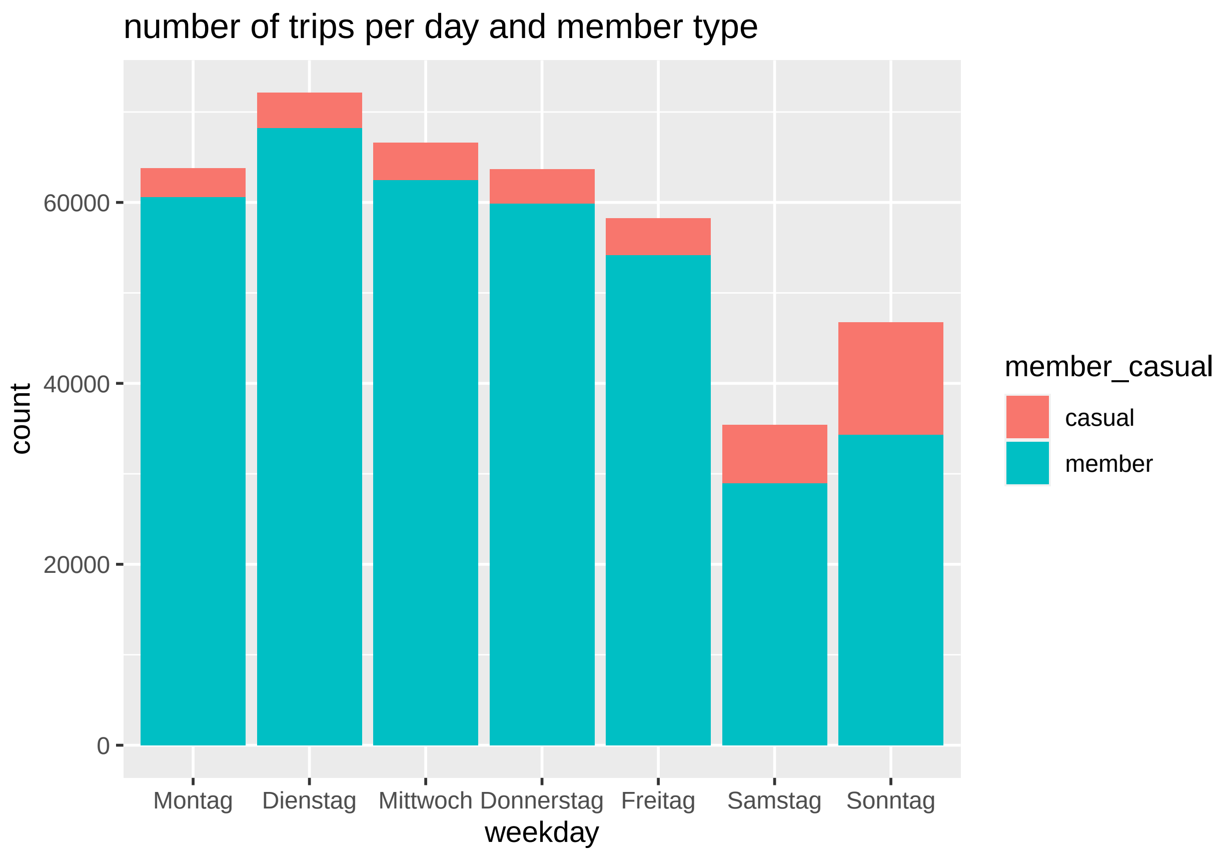
According to our theory members use Cyclistic more for commuting while casual users use it more for leisure purposes. If we look at the start and end stations we can infer if a ride was a tour (start and end are the same) or a trip to get somewhere (start and end station are different). If we think that members mostly use Cyclistic for commuting, then most of their trips should be trips to to get somewhere, while most of the tours should be from casual members.

First I decided to filter out the rides that have the same start and end station. So I filtered out the tours.



We can see in that especially on the Sunday the number of trips is very high. This was to be expected as we thought that the number of trips correlate with rides for leisure purposes. If we have a look now at the different member types we see that casual members are slightly more present than members, but not enough for a clear argumentation.

So the reasonable thing to do would be turning it around. If we don't look if a member type is under present in tours and instead look if they are over present in trips, we might get more answers.



We can clearly see that the members are significantly overpresent in the trips while the casual users are under represented. So most of the rides that are trips to get somewhere are done by members and not by casual users.

Conclusion

The Question that we tried to answer was:

How do annual members and casual riders use Cyclistic bikes differently?

As far as our data shows, Cyclistic members seem to use the bikes more for normal commuting to work, university and other destinations, mostly during the workdays. They use them mostly between 6am and 9am as well as between 3pm and 6pm. They mostly use it for short distances with usually less than 20 minutes duration.

The casual users use the Cyclistic service more during the weekend or after work. Their trips take longer on average and are more likely to be tours (rides that end where they started).

With regards to the Cyclistic advertisement campaign I would suggest trying to convince the casual users of the benefits of Cyclistic for normal commuting. If they start using it regularly to go to work, to school, to university or wherever they might need to go, Cyclistic could become a solid part of their life. If we were able to do this, we could increase their chances switching to an annual membership.