Cyclistic Project Report

Diego Sanoja

2022-11-14

Objective of the project

Determine how casual riders and annual members used Cyclistic bikes differently during a one year period to develop a marketing strategy whose goal is to turn casual riders into annual members.

Data Used

For this project, 12 csv files which contained data collected from the Cyclistic customers from the beginning of October 2021 to the end of September 2022 were downloaded from a public data set whose URL direction is https://divvy-tripdata.s3.amazonaws.com/index.html.

Each csv file contains information of the use of a Cyclistic bike like the bike id, the type of bike rented, the starting and ending time of the lease, the station where the bike was rented and where it was returned, and finally the customer type of a particular month.

The data has some limitations: it only provided information from the last 12 months, it didn't provide the prices to rent the bikes for each type of customer nor season discounts.

Changelog

Version 1.0.0 (07-11-2022)

- Created a folder called Capstone_Project to store the files related to this project.
- Added a sub folder called original_files to store the original csv files.

The following applies to the 12 csv files used in this project.

- Saved the original file in the original_files sub folder.
- Used the delete duplicates option of Excel to remove repeated rows.
- Removed the start_lat, start_lng, end_lat, end_lng columns.
- Added the following 4 columns:

i. Rent_time which displays the amount of time each bike was used in minutes.

ii. Month which contains the name of the month when the bike was used using the TEXT function.

iii. Day whose value represents the day of the week that a bike was rented using the WEEKDAY function (the values go from 1 to 7 where 1 represents the Sunday and the 7 represents the Saturday).

iv. Hour which shows the hour of the day when the bike was rented using the HOUR function.

- Removed the rows where the ride_id value was made only of numbers.
- Capitalized the values of the Member_Casual columns using the Find and Replace option.

- Removed the _bike sub string from the cells of the rideable_type column and then capitalized the strings using the Find and Replace option.

- Sorted the rows by the value of the started_at column from oldest to newest.

- Saved the file with the cleaned data using the same name plus the -cleaned string as a csv file in a sub folder called cleaned_files.

Version 1.1.0 (08-12-2022)

- Created an RStudio document to continue the cleaning and analysis of the 12 csv files. The name of this R file is Project_analysis_upgraded.

- Created a data frame containing all the data from the 12 csv files that contained the cleaned data. The name of this data frame is full_year_frame.

- Removed the rows of the data frame where the value of the ride_id column was null or the Rent_time was lower than 1.5 minutes.

- Replaced the values of the Day column by the name of their day in the following way: 1 per Sunday, 2 per Monday and so on.

- Created the following vectors using for loops:

i. year_vector which contains the number of the year where the bike was used by using the values in the Month column of the data frame.

ii. period_vector which contains the period of the day when the bike was rented by using the Hour column.

- Added the new vectors to the data frame.

- Converted the Day and Month columns into factors (data objects which are used to categorize the data and store it as levels).

- Renamed the following columns:

i. Rideable_type to Bike_used.

ii. Started_at to start_time.

iii. Ended_at to end_time.

iv. Member_Casual to Customer_type.

- Created the following sub data frames from the full_year_frame:

i. aggregated_frame_month which contains the mean rent time and the amount of bikes used in thousands per month.

ii. aggregated_frame_day which stores the mean rent time and the number of bikes rented in thousands per day.

iii. aggregated_frame_period which contains the amount of bikes used in thousands per period of the day.

iv. aggregated_frame_bike_type which stores the number of bikes rented in thousands per type of bike used.

v. max_frame_month which contains the time of the longest bike lease per month.

vi. max_frame_day which contains the time of the longest bike lease per day.

Version 1.1.1 (09-11-2022)

- Created the following sub data frames from the full_year_frame:

i. long_rents_frame_month which stores the amount of bikes used per month for long leases (leases longer than 60 minutes).

ii. long_rents_frame_day which stores the number of bikes used per day for long leases.

iii. long_rents_frame_period which stores the amount of bikes used per period of the day for long leases.

iv. short_mean_frame_month which contains the mean rent time of short leases (leases less than or equal to 1 hour) per day.

v. short_mean_frame_day which contains the mean rent time of short leases (leases less than or equal to 1 hour) per day.

- Used the data in the:

i. aggregated_frame_month to design 2 visuals:

1. A line chart that displays the number of bikes rented per month and customer type.

2. A line chart that shows the mean time that a bike was used per month and customer type.

ii. aggregated_frame_month to design 2 visuals:

1. A grouped bar chart that displays the number of bikes rented per day and customer type.

2. A grouped bar chart that displays the mean time that a bike was used per day and customer type.

iii. aggregated_frame_period to create a grouped bar chart that displays the number of bikes rented per period of the day and customer type.

iv. aggregated_frame_bike_type to create a table to display the number of bikes used per type of bike rented and customer type.

v. max_frame_month to create a dot plot that shows the time of the longest bike lease per month and customer type.

vi. max_frame_day to create a dot plot that shows the time of the longest bike lease per day and customer type.

vii. long_rents_frame_month to produce a line chart that displays the number of bikes rented per month and customer type for long rents.

viii. long_rents_frame_day to produce a line chart that shows the number of bikes rented per day and customer type for long rents.

ix. long_rents_frame_period to create a table that displays the number of bikes rented per period of the day and customer type for long rents.

x. short_mean_frame_month to design a line chart that shows the mean time that a bike was used for short leases per month and customer type.

xi. short_mean_frame_day to design a grouped bar chart that displays the mean time that a bike was used for short leases per day and customer type.

Analysis summary

For the analysis phase, the R programming language tool was used. The reasons to use R for the analysis phase were its ability to process larger amounts of data faster than excel and the packages it has with functions and codes that can create effective data visuals and reports.

After cleaning the data of the 12 csv files, they were combined in a single data frame to study all the data together and perform further data cleaning and transformations. The name of the final data frame which contains all the relevant data is full_year_frame which has 15 columns and 5668453 rows or observations.

By using the information stored in this data frame, 11 sub data frame which have aggregated data by groups were created with the purpose of answering questions regarding how the casual riders and customers with annual memberships use Cyclistic bikes differently were answered.

The first question to answer was how many bikes were used by each group of users through the time period of interest. To answer it, the data in the sub data frame aggregated_frame_month was used to designed Figure 1 shown below which displays the number of bikes rented per group of customers each month.


Figure 1: Number of Rents per Month
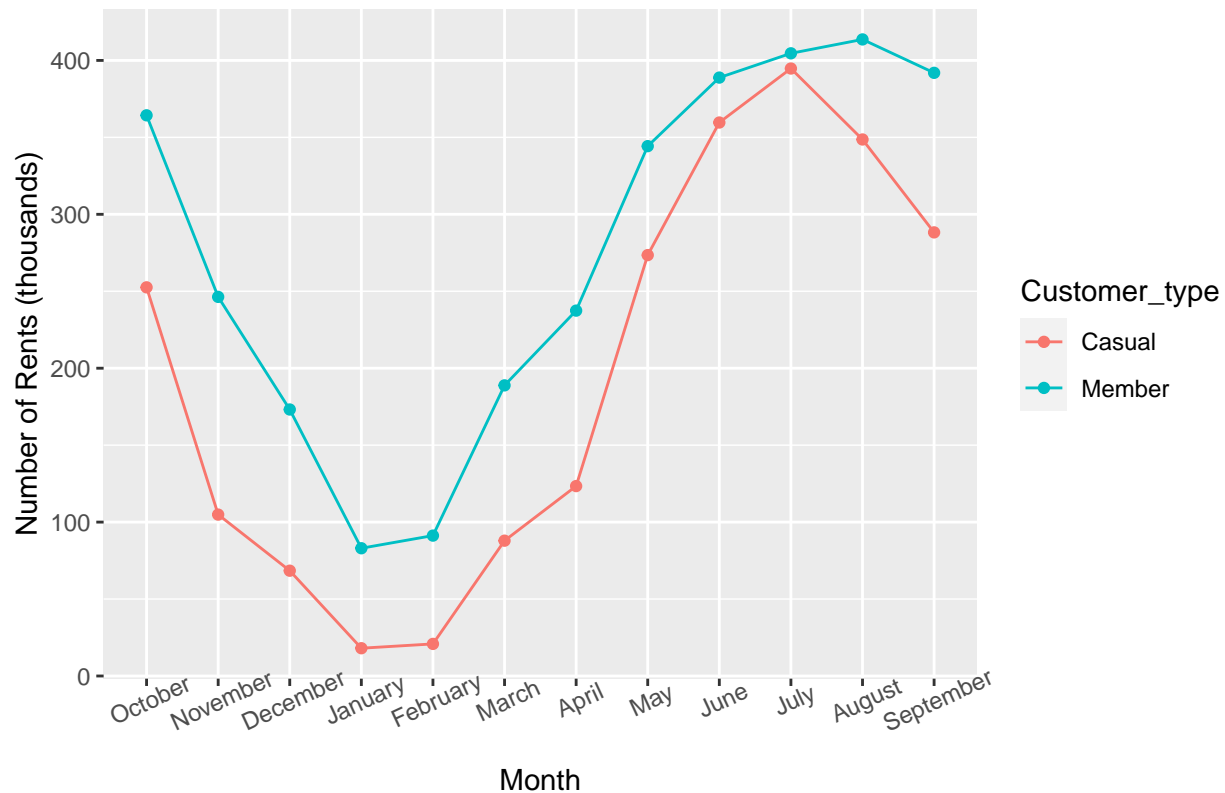
This line chart provided the following insights:

a. The months when bikes were used the most were June, July, August and September which correspond to the months of the summer season.

b. The months where bikes were used the least were December, January and February which correspond to the months of the winter season.

c. Through the year, customers with annual memberships used more bikes than casual riders every month.

To understand how both groups of clients rented bike differently through the days of the week, the sub frame aggregated_frame_day was designed. The data stored in it was used to create Figure 2 which displays the number of bikes rented per group of customers each day.

Figure 2: Number of Leases per Day
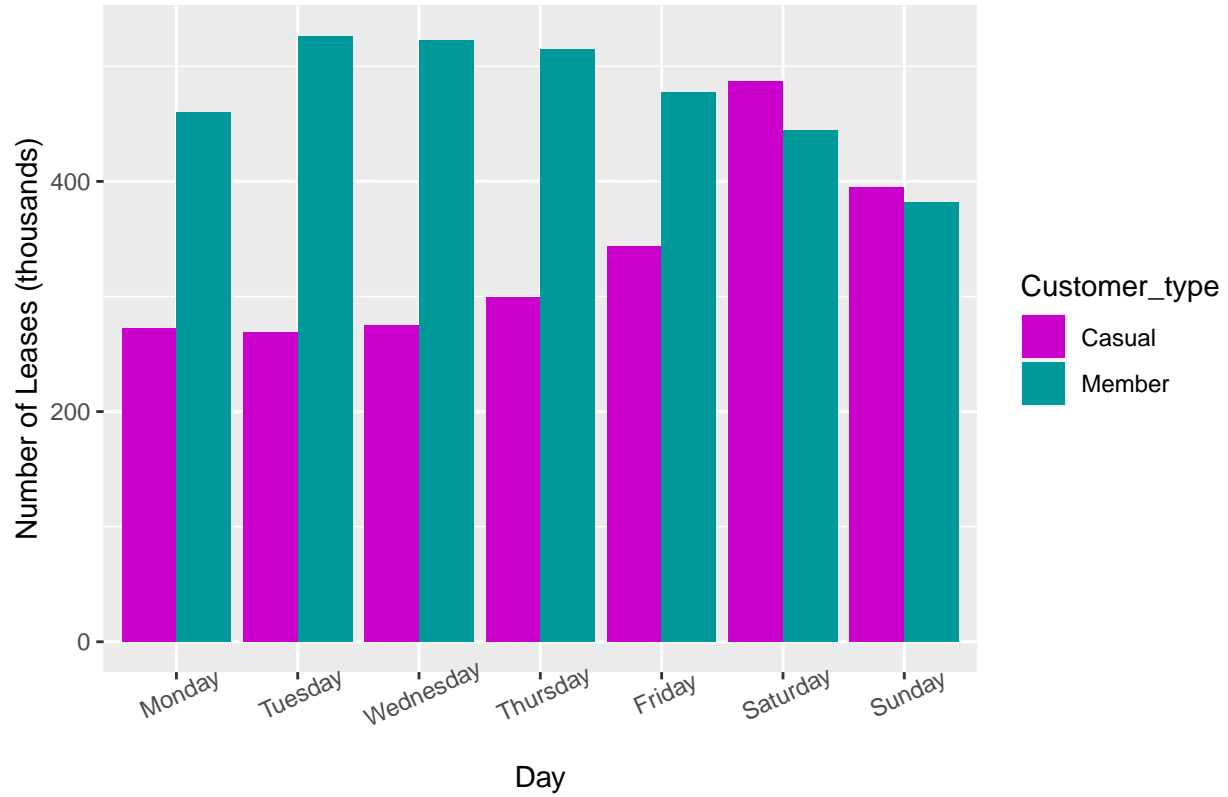
By examining Figure 2 closely, the following conclusions were reached:

a. From Monday to Friday, annual members used more bikes than casual riders. The opposite happened during the weekends.

b. The difference between the rents of each group is greater from Mondays to Fridays.

c. The day where bikes were mainly used for annual members and casual riders were Tuesday and Saturday respectively.

To discover in which moment of the day bikes were used mostly by each group of users, the sub frame aggregated_frame_period was built. The information in it was used to create Table 1 shown below where the rows and columns are represented by the customer type and the period of the day when a bike was rented respectively.

Table 1: Table 1: Leases per Group and Period

	Afternoon	Evening	Morning	Night
Casual	834603	683923	454449	367845
Member	1064847	980208	940372	342206

By studying Table 1 carefully, the following insights were discovered:

a. Annual members rented more bikes than casual riders in almost every period of the day except the night.

b. Most of the leases for each group happened during the afternoon and the evening.

The next step of the analysis was to understand which type of bike in Cyclistic was used mostly by each group of clients. To accomplish this, the aggregated_frame_bike_type was used to create Figure 3 which displays the number of bikes leased per type of bike used for each group.

Figure 3: Number of Leases per Type of Bike
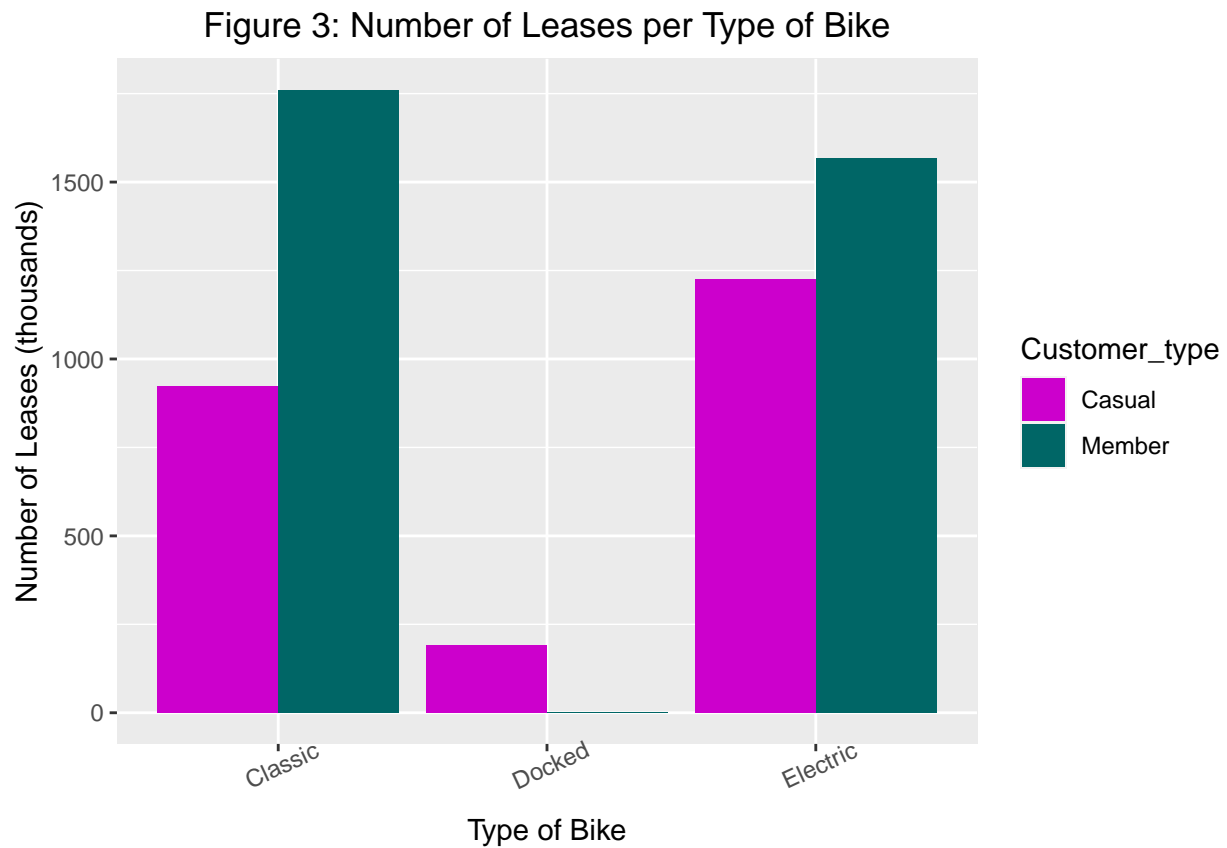
After studying Figure 3 closely, the following conclusions were reached:

a. From the 3 types of bikes, the one that was used the least was the docked bike. No annual member used a docked bike through the year.

b. Casual riders and annual members leased electric and classic bikes respectively mostly through the year.

Besides counting the number of times that bikes were leased for each group of clients, it was important to discover the average time of the rents both for each month and each day. To accomplish this, the data stored in the aggregated_frame_month and aggregated_frame_day sub frames was used to create Figures 4 and 5 respectively, each one shown next.
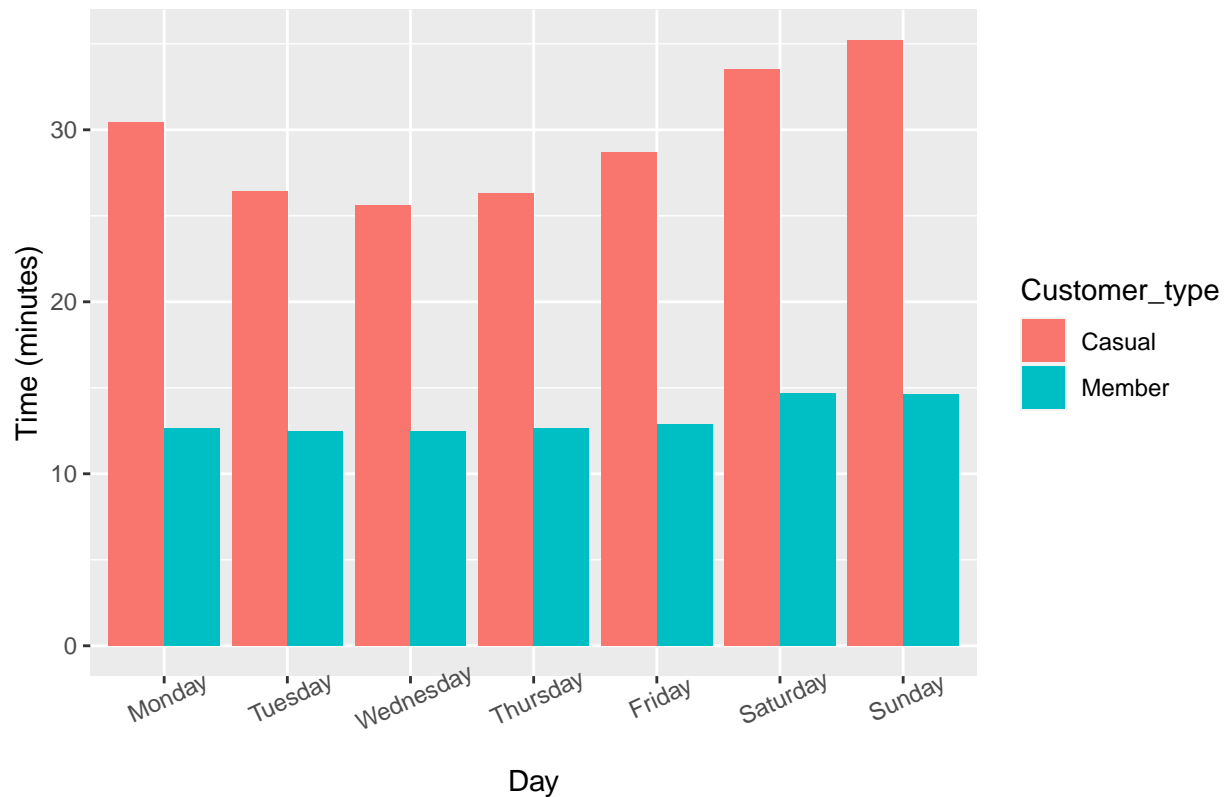
Figure 4: Average Rent Time per Month
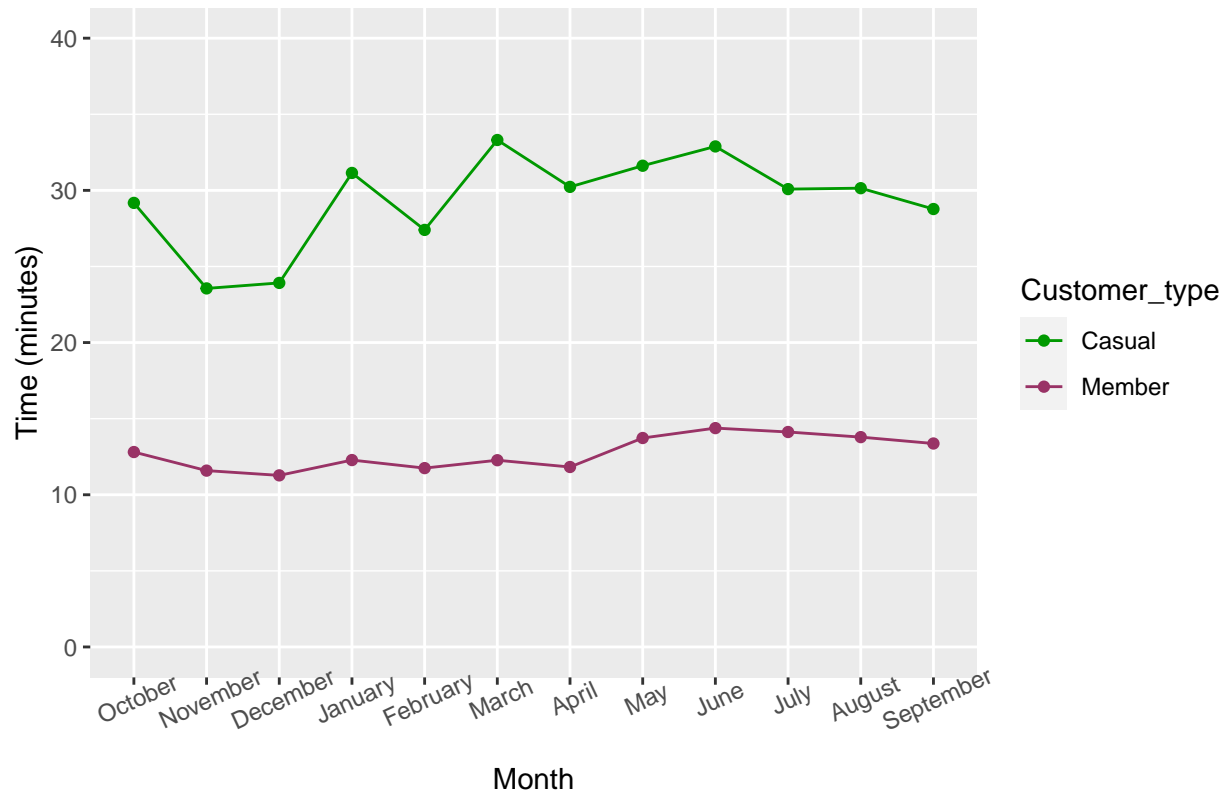
By examining Figure 4 closely, the following conclusions were reached:

a. Casual members rented bikes for longer periods of time than annual members every month in average.

b. The average rent time of casual members didn't change very much per month. Furthermore, each month had an average in the range of 11 to 15 minutes.

c. The average rent time of casual members fluctuated every month of the year. Furthermore, each month had an average in the range of 23 to 34 minutes.

d. There is no seasonality trend of increase nor decrease for both groups.

Figure 5: Average Rent Time per Day

By studying Figure 5 carefully, the following insights were discovered:

a. Casual members rented bikes for longer periods of time than annual members every day in average.

b. The average rent time of casual members didn't change very much per day. Furthermore, each month had an average in the range of 12 to 15 minutes.

c. The average rent time of casual members changed significantly almost every day. Furthermore, each month had an average in the range of 25 to 36 minutes.

d. For both groups (especially the casual riders), the rents were longer during the weekends in average.

In the same way the average rent time was analyzed, it was important to study the time of the longest bike lease for each group of riders both by month and day. To do this, the sub data frames max_frame_month and max_frame_day were designed which store the time of the longest lease in hours for each month and day respectively for both groups of clients. The data in those frames was used to build Figures 6 and 7 displayed below.
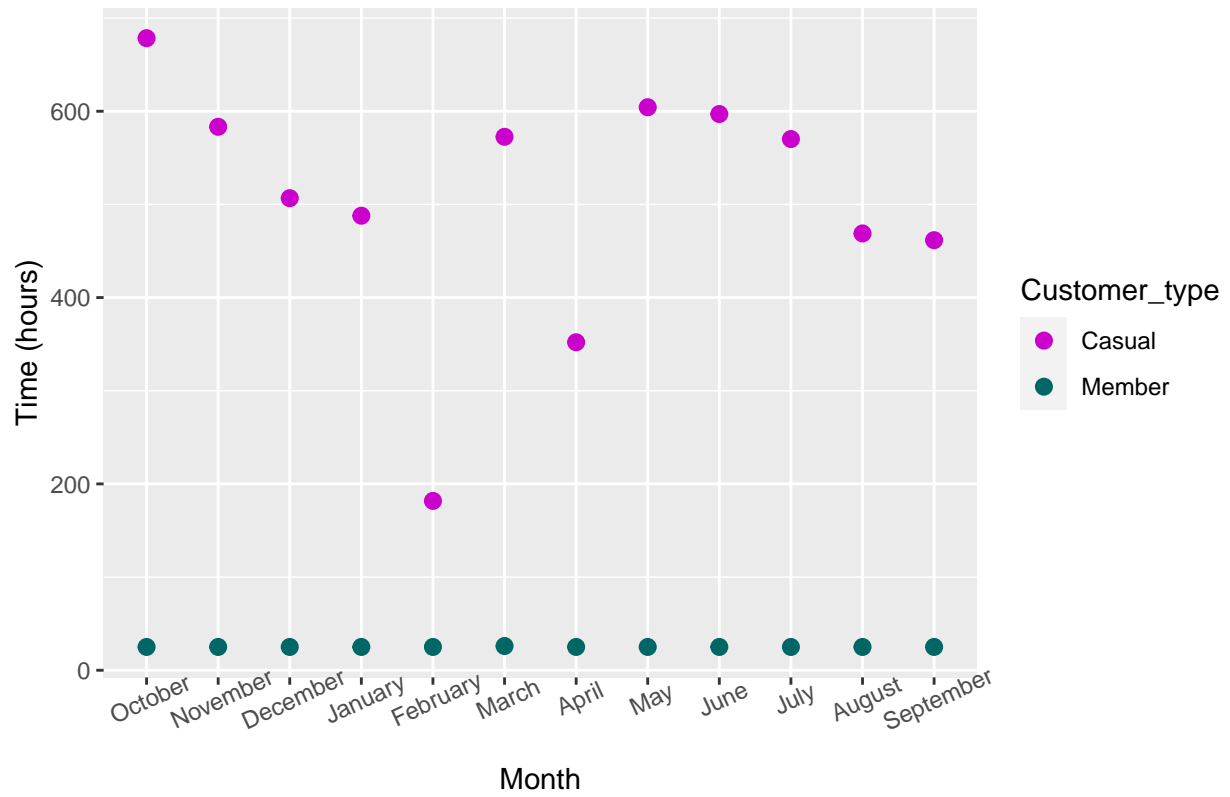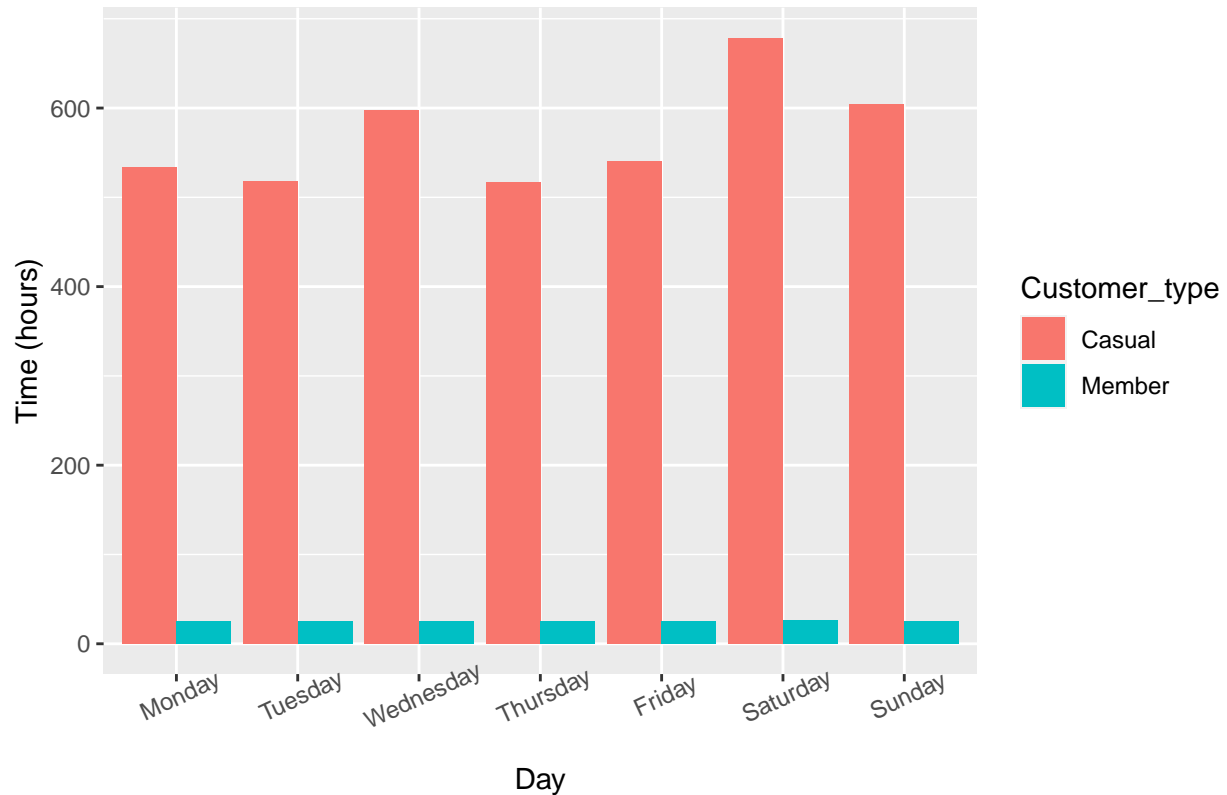
Figure 6: Longest Bike Lease per Month

Figure 7: Longest Bike Lease per Day

By examining Figures 6 and 7 closely at the same time, the following conclusions were reached:

a. The longest leases of casual members both by month and day of the year were significantly higher than the one of annual members.

b. Each month, the longest lease for casual riders was longer than 150 hours while for annual members it was about 25 hours.

c. For the days of the week, the longest leases for casual riders were longer than 500 hours while for annual members it was about 25 hours.

d. The longest lease began a Saturday in the month of October and was done by a casual rider.

Since most of the lease times in Figures 6 and 7 were higher than 20 hours, it is possible that the average rent time of Figures 4 and 5 were skewed both by those leases and others close to them. To discover if the results were skewed, the leases were divided in 2 categories, short and long. Shorts leases are the ones were the a bike was used by a client for an hour at most while long leases are the ones that lasted more than 1 hour.

The first sub data frame analyzed to check for skewness in the averages was the long_rents_frame_month which contains the amount of long bike rents of every month of the year. By using this data, Figure 8 was designed.

Figure 8: Long Rents per Month
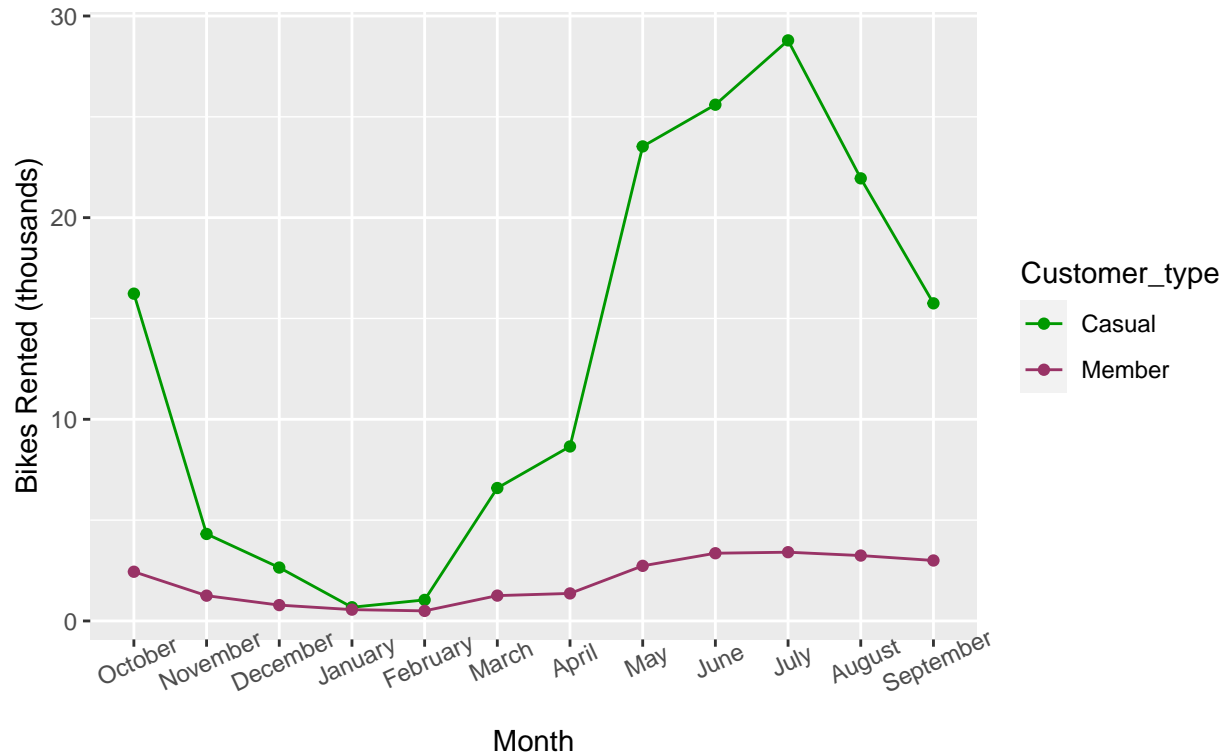
By studying Figure 8 carefully, the following insights were discovered:

a. Every month, casual riders did more long rents than annual members (especially during October of 2021 and from May to September of 2022).

b. There was a seasonality trend since the greatest and lowest amounts of long were done during summer and winter respectively.

c. The increase and decrease of long rents was smooth for annual members during the year but fluctuated significantly for casual riders.

The next frame studied to check for skewness in the averages was the long_rents_frame_day which stores the amount of long bike rents per day. Figure 9 displayed below was created using the information of this data frame.

Figure 9: Bikes Rented per Day
Rents Longer Than 60 Minutes
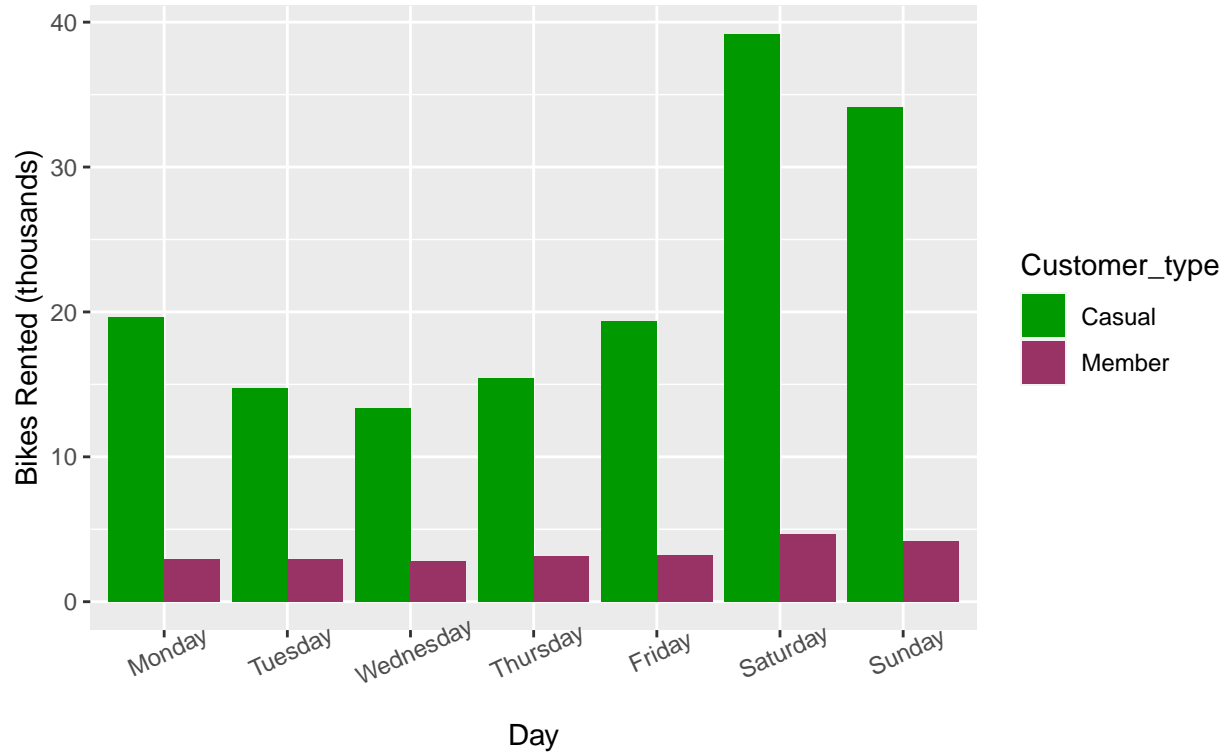
After examining Figure 9 closely, the following conclusions were obtained:

    a. Most of the long rents for both groups (especially for casual riders) were done during the weekends.

    b. Every day, most of the long rents were done by casual members.

The last important question regarding the long rents analysis was to determine how many long rents were done during each period of the day during the year. To make this possible, the long_rents_frame_period sub frame was designed which contains the number of long rents per period of the day and user type. By using this data, Table 2 shown below was created.

Table 2: Table 2: Long Rents per Group and Period

	Afternoon	Evening	Morning	Night
Casual	69676	35771	34078	16266
Member	8654	7119	5880	2256

By studying Table 2 carefully, the following insights were discovered:

    a. Most of the long rents were done during the afternoon and evening by casual riders.

    b. For each period, casual riders did more long rents than annual members.

The final part of the analysis was to examine how the average rent times per month and day changed after removing the leases that were longer than 1 hour. To do this, the short_mean_frame_month and short_mean_frame_day sub frames which store the average lease time of short leases (leases equal or shorter than 1 hour) per month and day respectively were built, analyzed and used to design Figures 10 and 11 respectively.
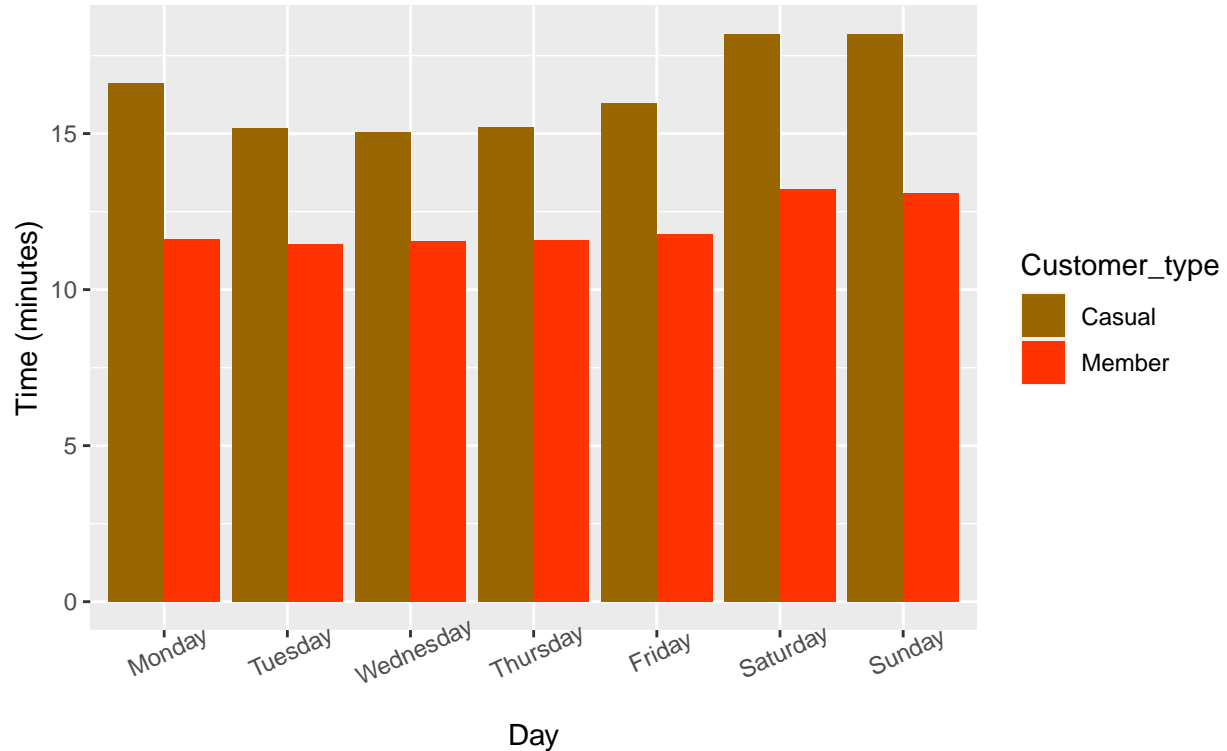
Figure 10: Average Lease Time of per Month
Rents Shorter or Equal to 60 Minutes
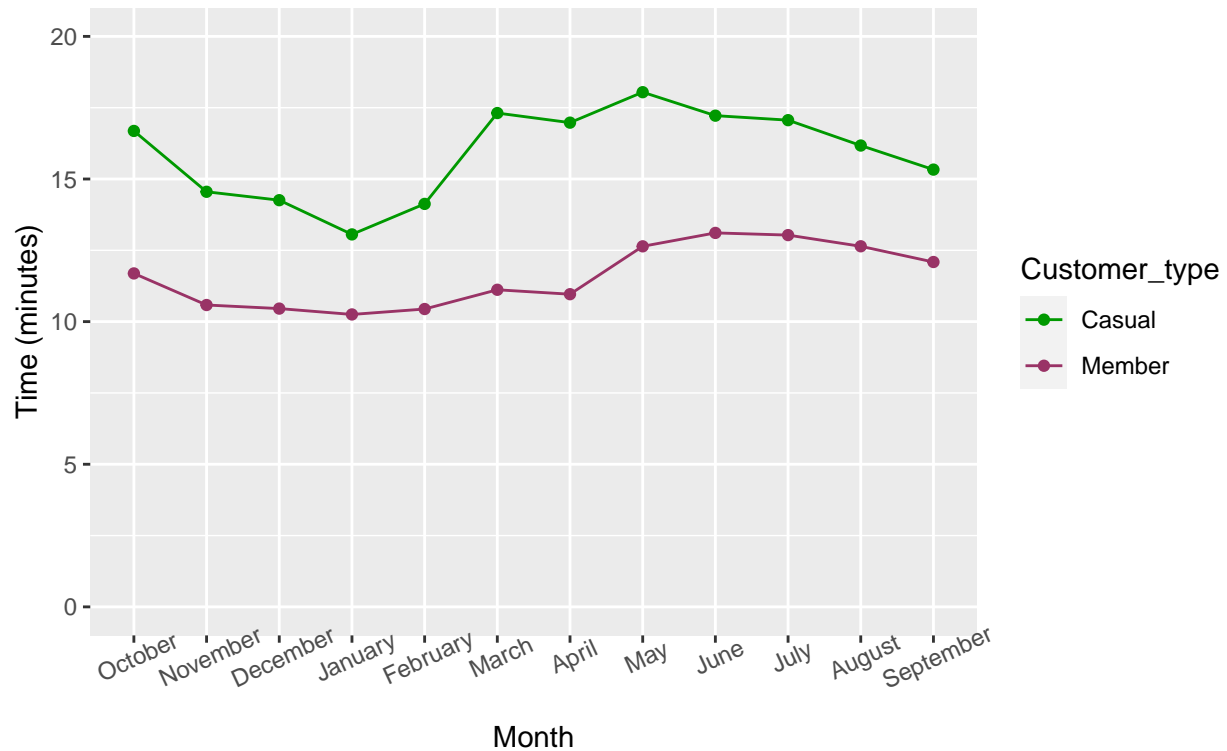
By studying Figure 10 and comparing it to Figure 4 carefully, the following insights were discovered:

a. The average lease time of both groups was reduced each month, especially for casual riders.

b. There was a seasonality trend for the annual members since the average were higher and lower during the summer and winter seasons respectively.

c. Casual riders still had higher average lease times than annual members each month but the difference was lower than before.

d. The change in average remained smooth for annual members and the fluctuations remained but were lower for casual riders.

e. The averages were now between 10 and 14 minutes and 13 and 18 minutes for annual members and casual riders respectively.

Figure 11: Average Rent Time of per Day
Rents Shorter or Equal to 60 Minutes

After analyzing Figure 11 and comparing it to Figure 5 carefully, the following conclusions were reached:

a. The average lease time of both groups was reduced each day, especially for casual riders.

b. Casual riders still had higher average lease times than annual members each day but the difference was lower than before.

c. The rents remained longer for the weekend by a lower difference than the rest of the week in average.

d. The averages were now between 11 and 13.5 minutes and 15 and 18.5 minutes for annual members and casual riders respectively.

Conclusions

- The months where bikes were used the most are the ones of the summer season (from June to September), where annual members used more bikes than casual riders.

- Casual riders used bikes mostly on weekends while customers with annual memberships used them mostly from Monday to Friday.

- Casual riders preferred to use electric bikes while members rented mainly classic bikes. Also, annual members didn't used docked bikes.

- Casual riders rented bikes by longer periods of time than annual members.

Recommendations to turn casual riders into annual members

a. Create a user name for annual members and anyone who uses a bike at least once using a piece of information like phone number or email. Then, send an email or message to casual riders which explains

them the benefits of becoming annual members.

b. Propose a discount or a special offer during the months of summer to turn casual riders into annual members.

c. Reduce the price for the membership for first-year annual members.

d. Repeat this analysis after 1 year to review the effectiveness of the campaign to improve and/or create new ideas to increase the number of annual members.