Case Study #1

Cyclistic, a bike-share company

About:

This project is a case study presented in "Google Data Analytics" course. This case study follows the bike share services of a company named “CYCLISTIC” that operates in Chicago. The data consists of users with annual memberships and casual users and their bike ride preferences that were collected from July 2022 – June 2023. The variety of parameters analyzed include bike types (classic, electric, docked), ride duration, start and end date and time of ride, and station names. Also, the data includes other complementary fields such as ride IDs, station IDs, and starting and ending points displayed as latitude and longitude.

The report follows the following data analysis process: Ask, prepare, process, analyze, share, and act.

Ask phase:

- Marketing team objective: convert casual riders into annual members.
- **Business task: Evaluate the differences between casual users (“casuals”) and annual members (“members”) of Cyclistic.**
- Key stakeholders: executive team and the director of marketing.

Prepare phase:

This public data is shared by CYCLISTIC company (https://Cyclisticbikes.com/how-it-works) that provides alternative transportation options in the way of bicycle sharing services. It is organized chronologically from the year 2020-2023, however, the analysis will be carried out on the last 12 months only, from July 2022 to June 2023.

The data is owned by the city of Chicago, it is up to date (06-2023), and it is divided into clear categories in a table format.

The data is public, there are no identifying details that can be traced back to a specific person, the data is unexclusive and open to everyone, and anyone is welcomed to use it in a lawful manner. The data is provided according to the Cyclistic Data License Agreement and released on a monthly schedule.

Process phase - cleaning and manipulating

All cleaning and manipulation were done in Excel and SSMS.

- **Creating duration columns**

A duration column was created and named “ride_length_min”. This column would hold the value of ride duration in minutes. Also “ride_length_sec” column was created, and its purpose was to hold the exact ride length as a whole number without any modifications.

First, “started_at” and “ended_at” columns (ride starting and ending time and date respectively) were converted from datetime data type to time data type via the following calculation:

= time(hour(started_at), minute(started_at), second(started_at))

The columns were named “ended_at_time” and “started_at_time”.

Next, the following formula was used to calculate duration in minutes:

F2 – ended_at_time column.

E2 – started_at_time column.

= IF(F2>=E2, (F2-E2)*24*60, (1-E2+F2)*24*60)

The same formula was used to calculate the ride duration in seconds but multiplied by another 60.

To simplify the evaluation, the decimal number was rounded to the closest integer, so that the analysis could be done in full minutes.

= ROUND (cell, 0)

This procedure made an insignificant deviation in ride length of ± 30 seconds.

After both columns were created and their values were calculated, with the help of the filter tool, the records who held zero values in both ride length columns (in minutes and seconds), were filtered out and deleted.

- **Creating “Day of the week” column**

This column represents the day on which the ride started.

The following function was used:

= WEEKDAY(C2,1)

C2 is the column that holds the start date.

1- represents Sunday.
2- Monday and so on.

Note: to avoid future complications in the analysis or in data migration, all formulas' outputs were copied to new columns as mere values and their formats were changed to the preferable data type.

- **Checking for duplicates**

Data -> Remove duplicates.

- **Checking for primary key (unique ride id)**

This step was skipped because there were too many records to manage it in excel, therefore it was completed later in SQL server. However, in case of less records, it could be evaluated in the following manner:

Data -> Sort&Filter: Advanced -> copy unique values to new location (the unique values of "ride_id" column would be copied to new location). Next, both columns' values count can be compared. If both columns have the exact count, all values in "ride_id" column are unique.

- **Checking formats**

Values were checked for format consistency.

- **Looking for unusual values (via filter option)**

There was a single case of an unusual finding. In September 2022, there was a negative value in "ride_length_min" column. The end date was smaller than the start date. Because it was not clear if these values were misplaced or if the whole record is faulty, it was decided to remove the record.

Creating a single table (SSMS)

After all files were uploaded to SSMS, a single table was created to hold all data.

First, there was a problem with data type compatibility. The columns "start_station_id" and "end_station_id" were recognized by the server as of nvarchar data type or as of float data type. All station id columns were changed to nvarchar data type.

```
ALTER TABLE [dbo].['202207-processed$']
ALTER COLUMN [start_station_id] nvarchar(255)
```

```
ALTER TABLE [dbo].['202209-Cyclistic-processed$']
ALTER COLUMN [end_station_id] nvarchar(255)

ALTER TABLE [dbo].['202210-Cyclistic-processed$']
ALTER COLUMN [start_station_id] nvarchar(255)

ALTER TABLE [dbo].['202211-Cyclistic-processed$']
ALTER COLUMN [start_station_id] nvarchar(255),
ALTER COLUMN [end_station_id] nvarchar(255)

ALTER TABLE [dbo].['202212-Cyclistic-processed$']
ALTER COLUMN [end_station_id] nvarchar(255)

ALTER TABLE [dbo].['202303-Cyclistic-processed$']
ALTER COLUMN [start_station_id] nvarchar(255)

ALTER TABLE [dbo].['202305-Cyclistic-processed$']
```

When the compatibility issue was solved, a consolidated table was created:

```
SELECT *
INTO BikeTrips
FROM (
                SELECT * FROM [dbo].['202207-processed$']
                UNION ALL
                SELECT * FROM [dbo].['202208-Cyclistic-processed$']
                UNION ALL
                SELECT * FROM [dbo].['202209-Cyclistic-processed$']
                UNION ALL
                SELECT * FROM [dbo].['202210-Cyclistic-processed$']
                UNION ALL
                SELECT * FROM [dbo].['202211-Cyclistic-processed$']
                UNION ALL
                SELECT * FROM [dbo].['202212-Cyclistic-processed$']
                UNION ALL
                SELECT * FROM [dbo].['202301-Cyclistic-processed$']
                UNION ALL
                SELECT * FROM [dbo].['202302-Cyclistic-processed$']
                UNION ALL
                SELECT * FROM [dbo].['202303-Cyclistic-processed$']
                UNION ALL
                SELECT * FROM [dbo].['202304-Cyclistic-processed$']
                UNION ALL
                SELECT * FROM [dbo].['202305-Cyclistic-processed$']
                UNION ALL
                SELECT * FROM [dbo].['202306-Cyclistic-processed$']
)t
```

Cleaning in SSMS - Unique values in “ride_id” column

It was important to understand if all values in “ride_id” column are unique or were there any repeating IDs. This step was skipped in Excel due to an incomputable number of records.

The total number of values were counted:

```
SELECT COUNT(*)
FROM [dbo].[BikeTrips]
```

Result

num_values
5,779,054

Next, unique values were counted in column “ride_id”:

```
SELECT COUNT(DISTINCT [ride_id]) num_unique_values
FROM [dbo].[BikeTrips]
```

Result

num_unique_values
5,778,987

There was a difference between the total count and “ride_id” count. There were 67 repeating values. One possible explanation was that there were NULL values in “ride_id” column.

```
SELECT COUNT(*) null_count
FROM [dbo].[BikeTrips]
WHERE [ride_id] IS NULL
```

Result

null_count
67

COUNT(*) counts all cells including NULL values, whereas COUNT([ride_id]) counts all cells except NULL values. The difference between the two is the number of NULL values. Because 67 values lack ride id, as shown in the above query, these records’ validity is questionable.

These records were removed:

```
DELETE FROM [dbo].[BikeTrips]
WHERE [ride_id] IS NULL
```

Result

(67 rows affected)

To verify the result, the unique values were compared with the total count.

```
SELECT COUNT(*) total_count
FROM [dbo].[BikeTrips]

SELECT COUNT(DISTINCT [ride_id]) num_unique_values
FROM [dbo].[BikeTrips]
```

Results

total_count	num_unique_values
5,778,982	5,778,982

Now, both queries result in an exact count.

Analysis SQL (SSMS)

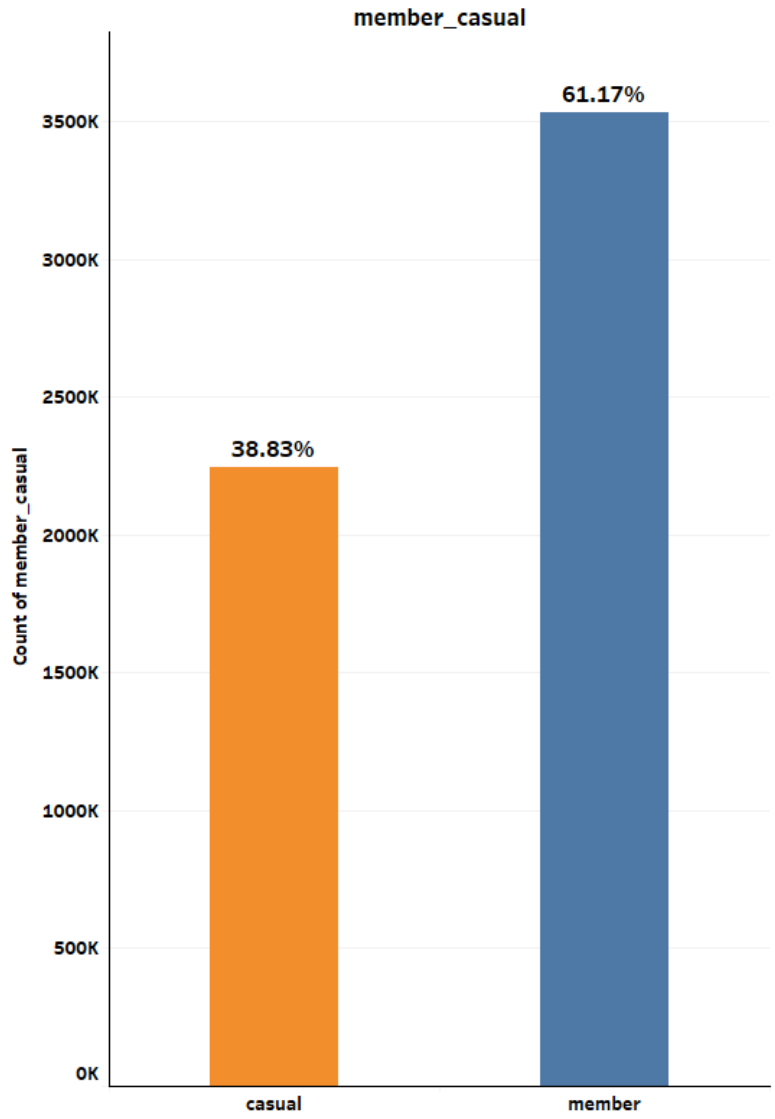
1. Compare casual riders count to annual members count.

```
SELECT
        SUM(CASE WHEN [member_casual] = 'casual' THEN 1 ELSE 0 END) casuals_count,
        SUM(CASE WHEN [member_casual] = 'member' THEN 1 ELSE 0 END) member_count
FROM [dbo].[BikeTrips]
```

Results

casuals_count	member_count
2,244,046	3,534,936

Members Count vs. Casual Users Count



The percentage of annual users (61.17%) is greater than the percentage of casual users (38.83%).

2. How many of the bikes used by annual and casual users are classic, electric, and docked bikes?

Members

```
SELECT
        SUM(CASE WHEN [rideable_type] = 'classic_bike' THEN 1 ELSE 0 END) AS Member_Classic,
        SUM(CASE WHEN [rideable_type] = 'electric_bike' THEN 1 ELSE 0 END) AS Member_Electric,
        SUM(CASE WHEN [rideable_type] = 'docked_bike' THEN 1 ELSE 0 END) AS Member_docked
FROM [dbo].[BikeTrips]
```

```
WHERE [member_casual] = 'member'
```

Results

Member_Classic	Member_Electric	Member_docked
1,690,904	1,844,032	0

Bike Type Among Members

Electric bikes were the more popular choice among annual users (52.17%) compared to classic bikes (47.83%).

Surprisingly, annual users did not use docked bikes at all.

Note: Docked bikes – bikes that must be returned to a docking station (can be either classic or electric bike).

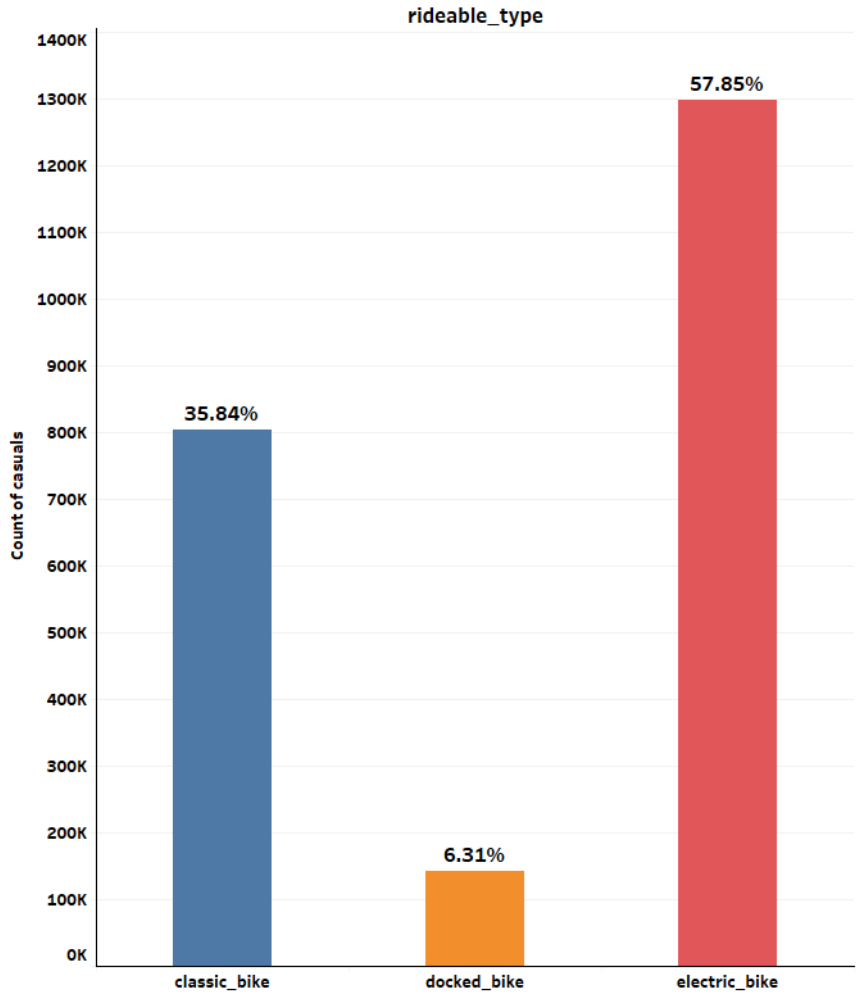
Casuals

```
SELECT
        SUM(CASE WHEN [rideable_type] = 'classic_bike' THEN 1 ELSE 0 END) AS Casual_Classic,
        SUM(CASE WHEN [rideable_type] = 'electric_bike' THEN 1 ELSE 0 END) AS Casual_Electric,
        SUM(CASE WHEN [rideable_type] = 'docked_bike' THEN 1 ELSE 0 END) AS Casual_docked
FROM [dbo].[BikeTrips]
WHERE [member_casual] = 'casual'
```

Results

Casual_Classic	Casual_Electric	Casual_docked
804,340	1,298,176	141,530

Bike Type Among Casuals



Casual users use more electric bikes (57.85%) than classic (35.84%) and docked bikes (6.31%). Docked bikes are the unpopular choice among causal users with less than 10 percent usage.

Comparing Bike Types in Members and Casuals

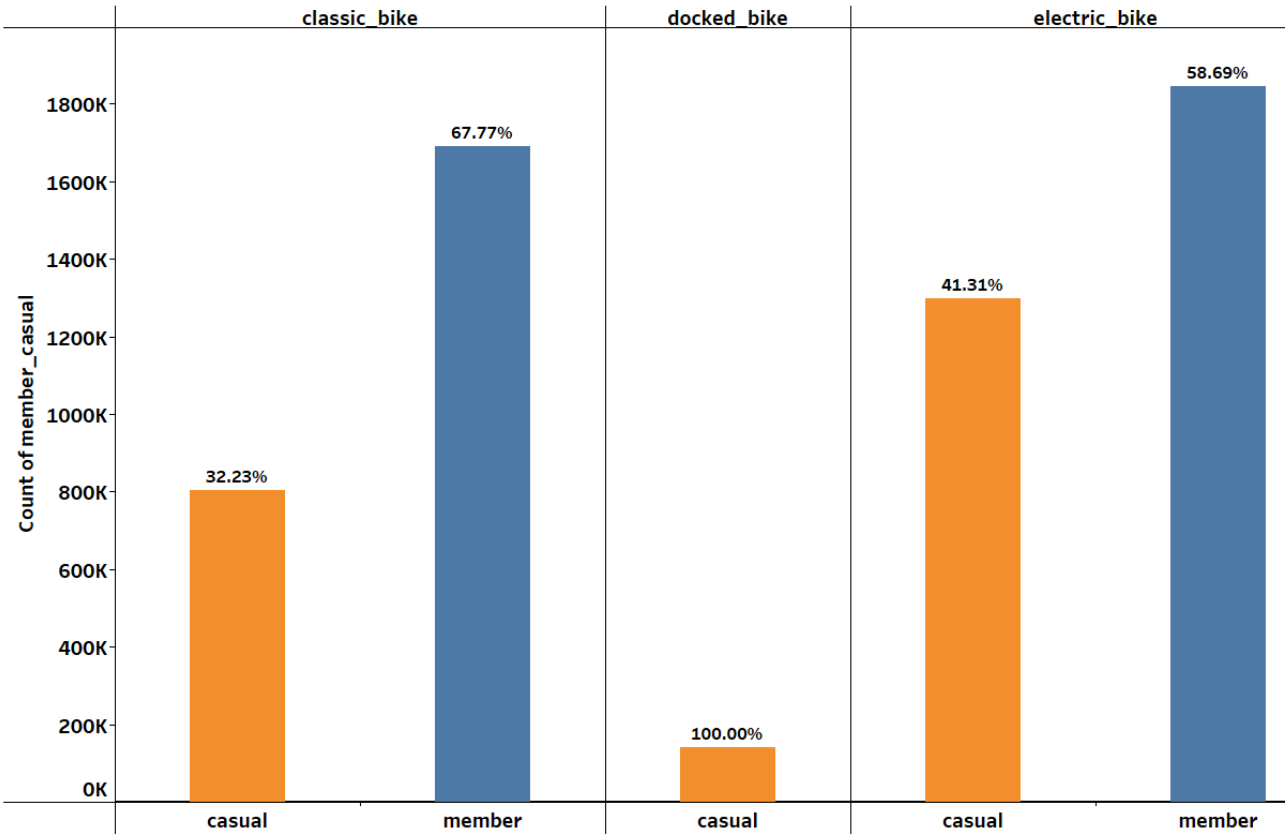
```
SELECT
        SUM(CASE WHEN [rideable_type] = 'classic_bike' THEN 1 ELSE 0 END) AS total_Classic,
        SUM(CASE WHEN [rideable_type] = 'electric_bike' THEN 1 ELSE 0 END) AS total_Electric,
        SUM(CASE WHEN [rideable_type] = 'docked_bike' THEN 1 ELSE 0 END) AS total_docked
FROM [dbo].[BikeTrips]
```

Results

total_Classic	total_Electric	total_docked
2,495,244	3,142,208	141,530

Summarizing table

Classic		Electric		Docked	
Members	Casuals	Members	Casuals	Members	Casuals
67.77%	32.23%	58.69%	41.31%	0%	100%



The percentage of members who use classic and electric bikes is greater than the percentage of casuals in these categories. However, only casual users use the docked type of bikes.

3. What is the busiest day using Cyclistic transportation?

Members

All types of bikes

```
SELECT [day_of_week], COUNT([ride_id]) AS 'rides_count'
FROM [dbo].[BikeTrips]
WHERE [member_casual] = 'member'
GROUP BY [day_of_week]
ORDER BY 'rides_count' DESC
```

Note: 1 - Sunday, 7 - Saterday

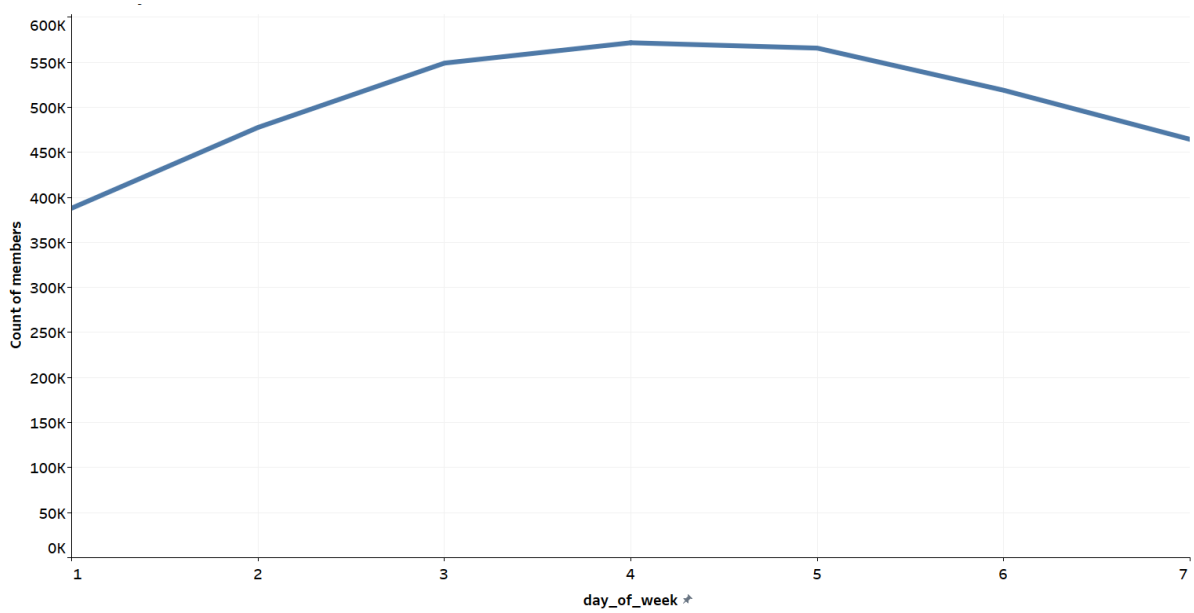
Results

MAX

MIN

day_of_week	rides_count
4	571,593
5	565,639
3	548,992
6	518,840
2	477,594
7	464,364
1	387,914

Preferable Days for Bike Share Use in Members



Regardless of bike type, members use Cyclistic transportation the most on Wednesday and the least on Sunday. This line chart shows that members use Cyclistic transportation less on weekends and it can be reinforced with Saturday having the next lowest ride counts in the summarizing table.

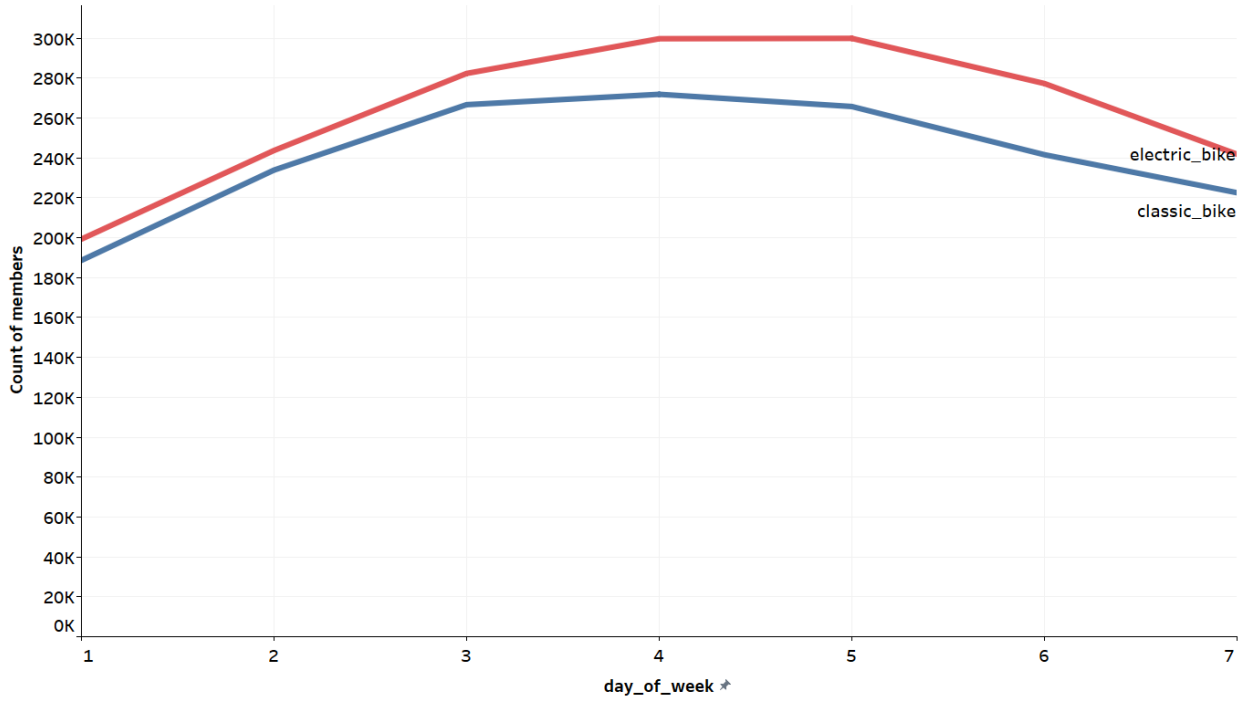
This line chart creates an overall view of members' bike share usage per day of the week from July 2022 to June 2023.

Segregation to bike types

```
SELECT [day_of_week],
        SUM(CASE WHEN [rideable_type] = 'classic_bike' THEN 1 ELSE 0 END) AS member_classic,
        SUM(CASE WHEN [rideable_type] = 'electric_bike' THEN 1 ELSE 0 END) AS member_electric
FROM [dbo].[BikeTrips]
WHERE [member_casual] = 'member'
GROUP BY [day_of_week]
ORDER BY 'member_classic' DESC
```

day_of_week	member_classic	member_electric
4	271,864	299,729
3	266,668	282,324
5	265,740	299,899
6	241,577	277,263
2	233,886	243,708
7	222,539	241,825
1	188,630	199,284

Preferable Days for Bike Share Use in Members Divided to Bike Types



As per bike type, the line charts for electric and classic bikes display similar trends during the week. The charts start with the lowest number of rides on Sundays, proceed to the peaks (Thursday for electric bikes and Wednesday for classic bikes), and then they decrease gradually until Saturday. This result proves that members who use classic bikes or/and electric bikes, use the different bikes in the same manner when it comes to different days of the week.

Casuals

All types

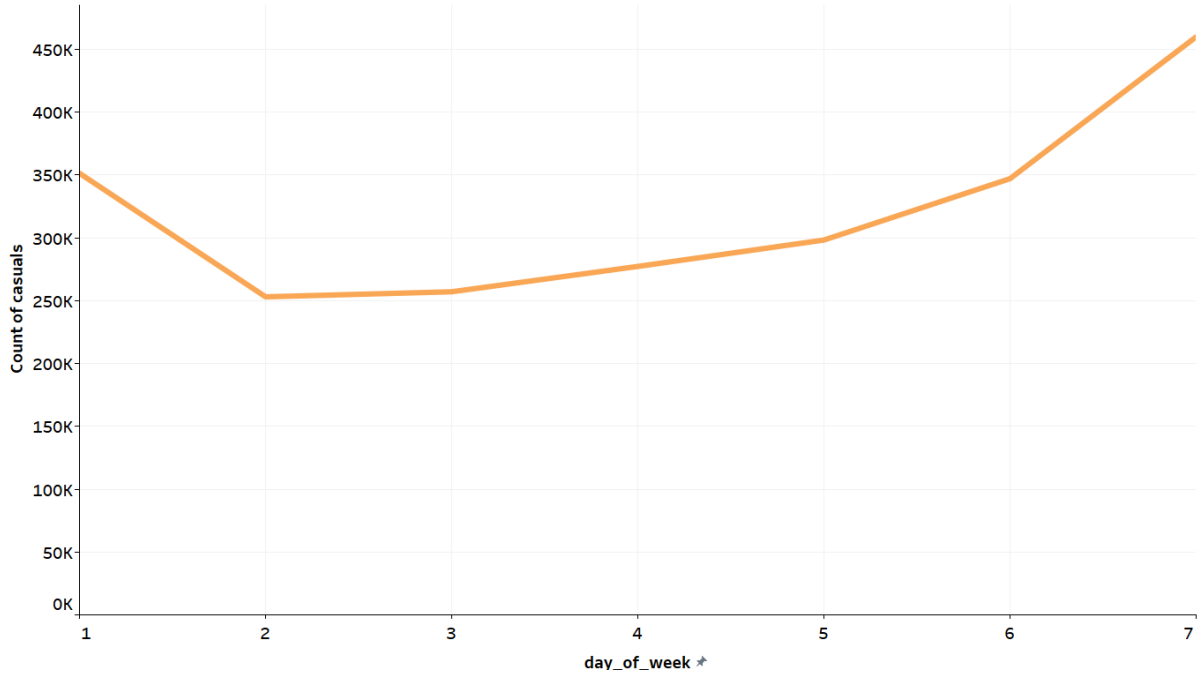
```
SELEC [day_of_week], COUNT([ride_id]) AS 'rides_count'
FROM [dbo].[BikeTrips]
WHERE [member_casual] = 'casual'
GROUP BY [day_of_week]
ORDER BY 'rides_count' DESC
```

Results

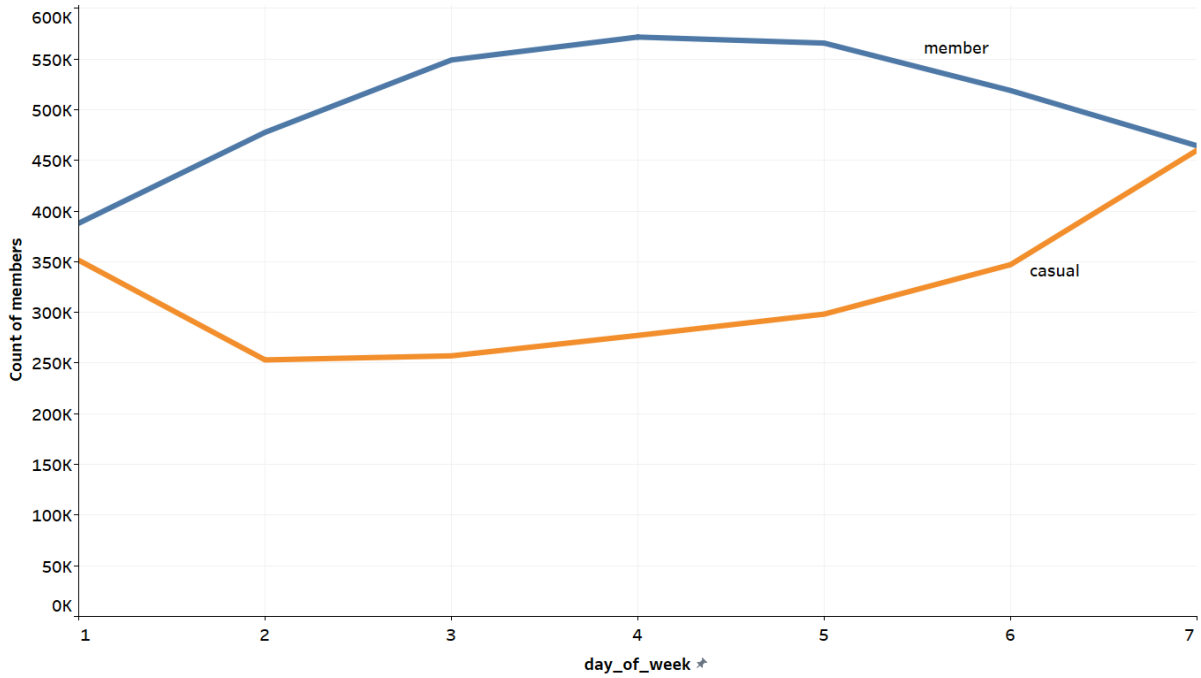
day_of_week	rides_count
7	459,936
1	351,382
6	347,074
5	298,257
4	277,246
3	257,084

2	253,067

Preferable Days for Bike Share Use in Casuals



The most popular day for casual users is Saturday and the least popular day is Monday. In contrast to annual users(members), casual users ride more on the weekends and less on the weekdays.



When combined, members and casuals exhibit opposite trends.

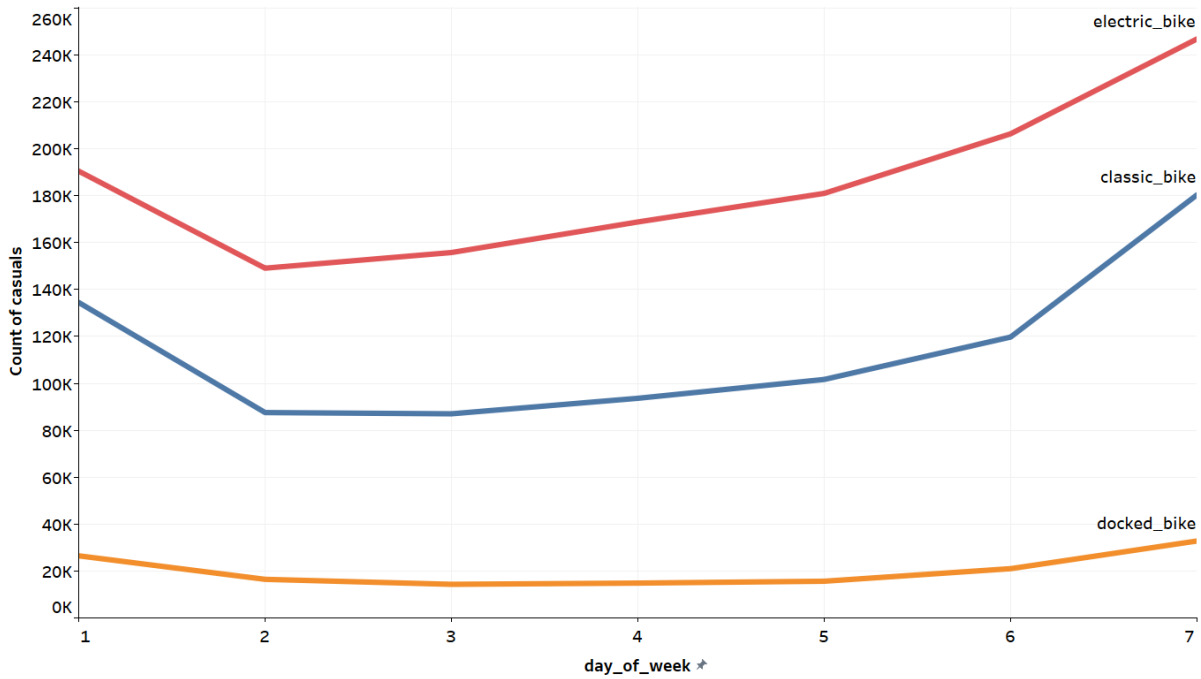
Segregation to bike types

```
SELECT [day_of_week],
        SUM(CASE WHEN [rideable_type] = 'classic_bike' THEN 1 ELSE 0 END) AS casual_classic,
        SUM(CASE WHEN [rideable_type] = 'electric_bike' THEN 1 ELSE 0 END) AS casual_electric,
        SUM(CASE WHEN [rideable_type] = 'docked_bike' THEN 1 ELSE 0 END) AS casual_docked
FROM [dbo].[BikeTrips]
WHERE [member_casual] = 'casual'
GROUP BY [day_of_week]
ORDER BY 'casual_classic' DESC
```

Results

day_of_week	casual_classic	casual_electric	casual_docked
7	180,340	246,800	32,796
1	134,465	190,444	26,473
6	119,721	206,359	20,994
5	101,651	180,963	15,643
4	93,620	168,778	14,848
2	87,537	149,081	16,449
3	87,006	155,751	14,327

Preferable Days for Bike Share Use in Casuals Divided to Bike Types



As for the different types of bikes, when grouped by the days of the week, the most popular day for casual users is Saturday, and the least popular day is Tuesday for classic and docked bikes and Monday for electric bikes. When compared to each other, the line charts display the same behavior regardless of bike type, which is the weekend being with the highest demand for bike share transportation in casual users.

Distribution of Rides per Month and Day

```
SELECT MONTH([started_at]) month_num, [day_of_week], COUNT([ride_id]) rides_count
FROM [dbo].[BikeTrips]
WHERE [member_casual] = 'member'
-- WHERE [member_casual] = 'casuals'
GROUP BY [day_of_week], MONTH([started_at])
ORDER BY month_num, rides_count DESC
```

Results

	Members		Casuals	
month_num	**day_of_week**	**rides_count**	**day_of_week**	**rides_count**
1	3	29377	3	6903
1	4	24743	1	6376
1	2	22648	4	5978
1	5	22644	2	5698
1	6	20106	5	5021
1	1	15989	7	5017
1	7	14780	6	5012

	Members		Casuals	
month_num	**day_of_week**	**rides_count**	**day_of_week**	**rides_count**
2	3	28630	1	9954
2	2	24391	7	6881
2	4	21214	3	6853
2	1	20335	2	6788
2	5	18139	4	4577
2	7	17404	6	4249
2	6	17308	5	3712

	Members		Casuals	
month_num	**day_of_week**	**rides_count**	**day_of_week**	**rides_count**
3	4	39750	4	11303
3	5	35902	5	10125
3	3	32122	6	9361
3	6	29153	3	9217

3	2	25388	7	7986
3	1	17189	1	7384
3	7	16961	2	6817

	Members		Casuals	
month_num	**day_of_week**	**rides_count**	**day_of_week**	**rides_count**
4	5	48315	7	32773
4	6	43580	6	25922
4	4	43072	5	23504
4	3	40487	1	18764
4	7	40364	4	18368
4	2	33924	3	15824
4	1	29545	2	12115

	Members		Casuals	
month_num	**day_of_week**	**rides_count**	**day_of_week**	**rides_count**
5	3	67675	1	42121
5	4	67511	7	40751
5	5	55684	3	33378
5	2	48891	4	31390
5	6	46298	2	29539
5	7	43315	6	28708
5	1	41236	5	28269

	Members		Casuals	
month_num	**day_of_week**	**rides_count**	**day_of_week**	**rides_count**
6	6	74729	7	69288
6	5	74265	6	56415
6	7	61329	5	42914
6	4	61083	1	41621
6	2	53012	4	33745
6	3	51482	2	32265
6	1	42450	3	24954

	Members		Casuals	
month_num	day_of_week	rides_count	day_of_week	rides_count
7	7	68864	7	95222
7	6	61642	1	78245
7	5	61151	6	56498
7	4	59608	5	47787
7	1	58778	2	43969
7	3	57520	4	42850
7	2	49847	3	41450

	Members		Casuals	
month_num	day_of_week	rides_count	day_of_week	rides_count
8	3	76710	7	66183
8	4	76612	6	56860
8	2	62598	3	51504
8	6	58697	4	51486
8	5	57506	1	48151
8	7	51883	2	42359
8	1	42968	5	42350

	Members		Casuals	
month_num	day_of_week	rides_count	day_of_week	rides_count
9	5	76582	7	64066
9	6	72326	6	56380
9	4	61504	5	45834
9	3	57028	1	36253
9	7	54014	4	33502
9	2	47417	2	31048
9	1	35737	3	29584

	Members		Casuals	
month_num	day_of_week	rides_count	day_of_week	rides_count
10	2	58286	7	52213
10	7	58133	1	44676
10	1	50363	2	27233
10	5	49234	6	25975
10	4	48887	5	22583
10	6	45017	4	20590
10	3	39741	3	15692

	Members		Casuals	
month_num	day_of_week	rides_count	day_of_week	rides_count
11	4	47447	5	17981
11	3	46114	4	17779

11	5	39125	3	15808
11	2	32305	6	14538
11	6	30324	1	12497
11	1	21207	7	11869
11	7	20440	2	10300

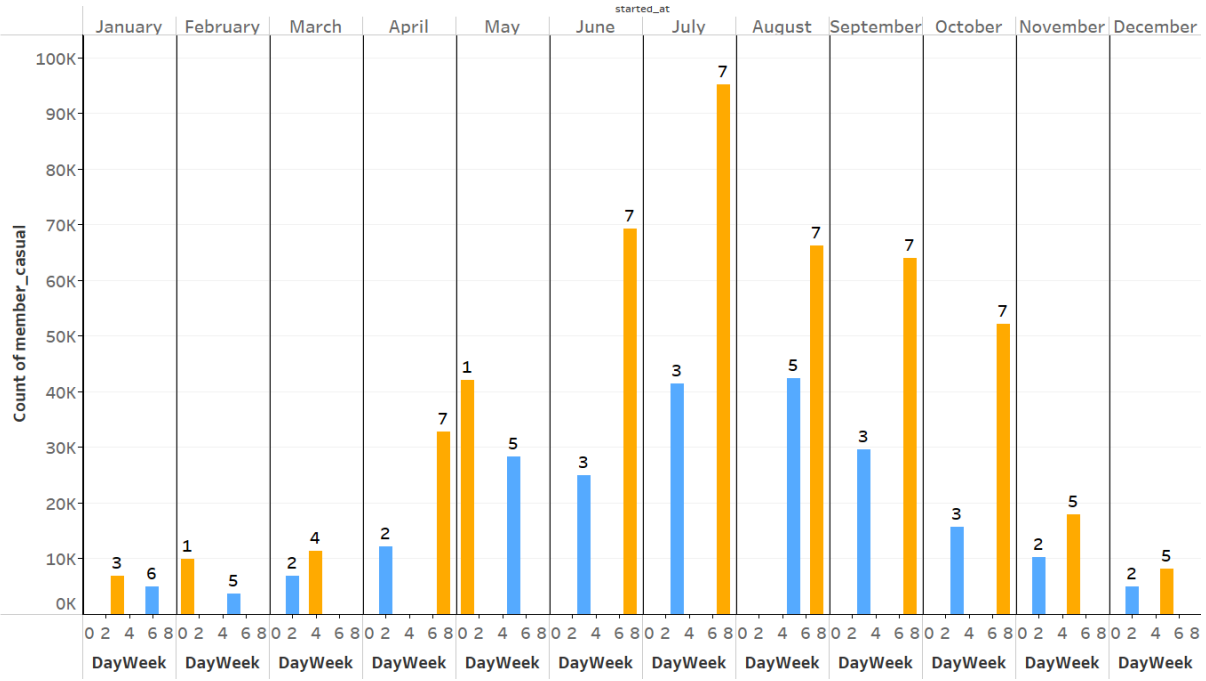
	Members		Casuals	
month_num	**day_of_week**	**rides_count**	**day_of_week**	**rides_count**
12	5	27092	5	8177
12	3	22106	7	7687
12	4	20162	6	7156
12	6	19660	3	5917
12	2	18887	4	5678
12	7	16877	1	5340
12	1	12117	2	4936

Summarizing table

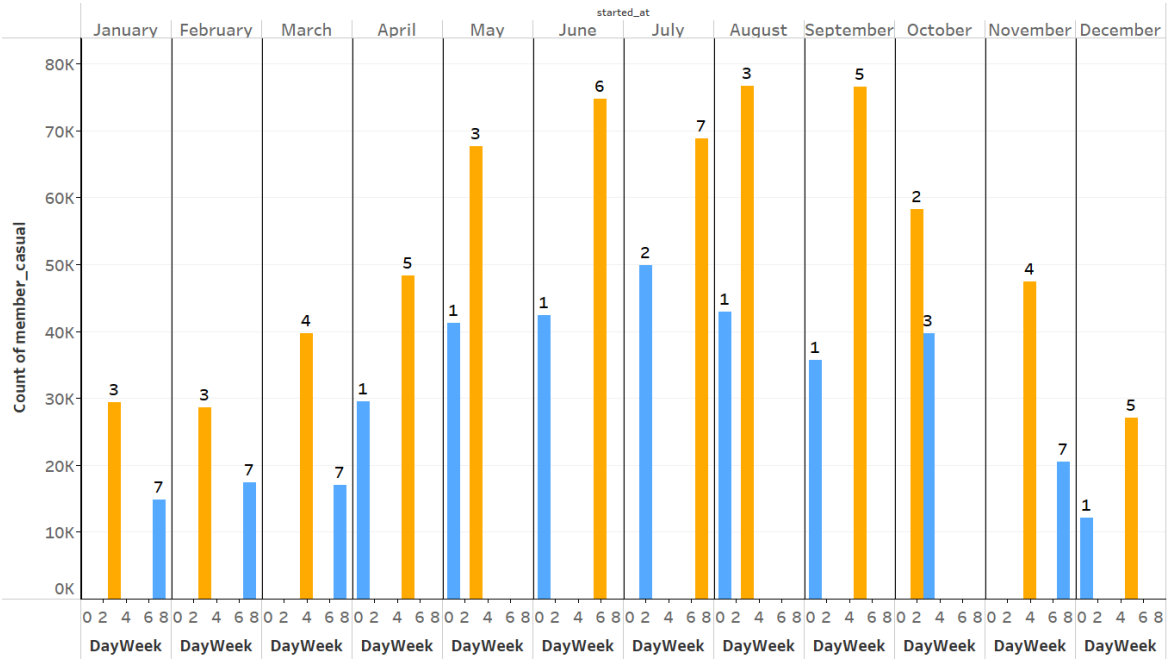
	Members				Casuals			
Month	**Most Popular Day**	**Rides Count**	**Least Popular Day**	**Rides Count**	**Most Popular Day**	**Rides Count**	**Least Popular Day**	**Rides Count**
1	3	29,377	7	14,780	3	6,903	6	5,012
2	3	28,630	6	17,308	1	9,954	5	3,712
3	4	39,750	7	16,961	4	11,303	2	6,817
4	5	48,315	1	29,545	7	32,773	2	12,115
5	3	67,675	1	41,236	1	42,121	5	28,269
6	6	74,729	1	42,450	7	69,288	3	24,954
7	7	68,864	2	49,847	7	95,222	3	41,450
8	3	76,710	1	42,968	7	66,183	5	42,350
9	5	76,582	1	35,737	7	64,066	3	29,584

10	2	58,286	3	39,741	7	52,213	3	15,692
11	4	47,447	7	20,440	5	17,981	2	10,300
12	5	27,092	1	12,117	5	8,177	2	4,936

Most/Least Popular Days of Using Bike Share in Casuals



Most/Least Popular Days of Using Bike Share in Members



This summarizing table holds the highest and the lowest ride counts per month for annual and casual users. Among members, there is no consistency regarding the most popular day to use Cyclistic transportation, however, when examining the unpopular days, the weekends stand out (Saturday and Sunday). In contrast to members, casual users use Cyclistic transportation the most on the weekends (Sunday and Saturday), however, as for the least popular days, there is no consistency except for being one of the weekdays. Bike share used by all users during the whole week, however, in an overall perspective, members use it more on weekdays and casuals use it more on weekends. It can be stated that if the user is a casual user, there is a bigger chance for him to use bike share on the weekends instead of the weekdays. It can only be **speculated** that casual users usually use Cyclistic bikes for different activities than annual users. For example, it is possible that most annual users register for annual membership so they can use bike share transportation to get to or from work, while most casual users use bike share for recreational activities or tourism.

Above all, it is important to point out the limitations of the data. Except for speculating for the possible reason for least/most popular days to use Cyclistic transportation, there is no way of knowing the true reason for these days' popularity. Only a questionnaire can reveal the common cause behind these choices. Also, it is important to point out that the table and the charts only summarize the highest and the lowest ride count records, however there are many other occasions in which the difference between the ride counts among different days is nearly insignificant. Moreover, the charts display cumulative amounts of rides, and therefore it masks the behavior of individual weeks within the examined period.

Ride Count Divided by Seasons

```
SELECT month_num, SUM(rides_count_member) total_count_member,
SUM(rides_count_casual)total_count_casual
FROM(SELECT MONTH([started_at]) month_num, [day_of_week],
        SUM(CASE WHEN [member_casual] = 'member' THEN 1 ELSE 0 END) rides_count_member,
        SUM(CASE WHEN [member_casual] = 'casual' THEN 1 ELSE 0 END) rides_count_casual
FROM [dbo].[BikeTrips]
GROUP BY MONTH([started_at]), [day_of_week]
)t
GROUP BY month_num
```

Results

Winter (December, January, February)

Spring (March, April, May)

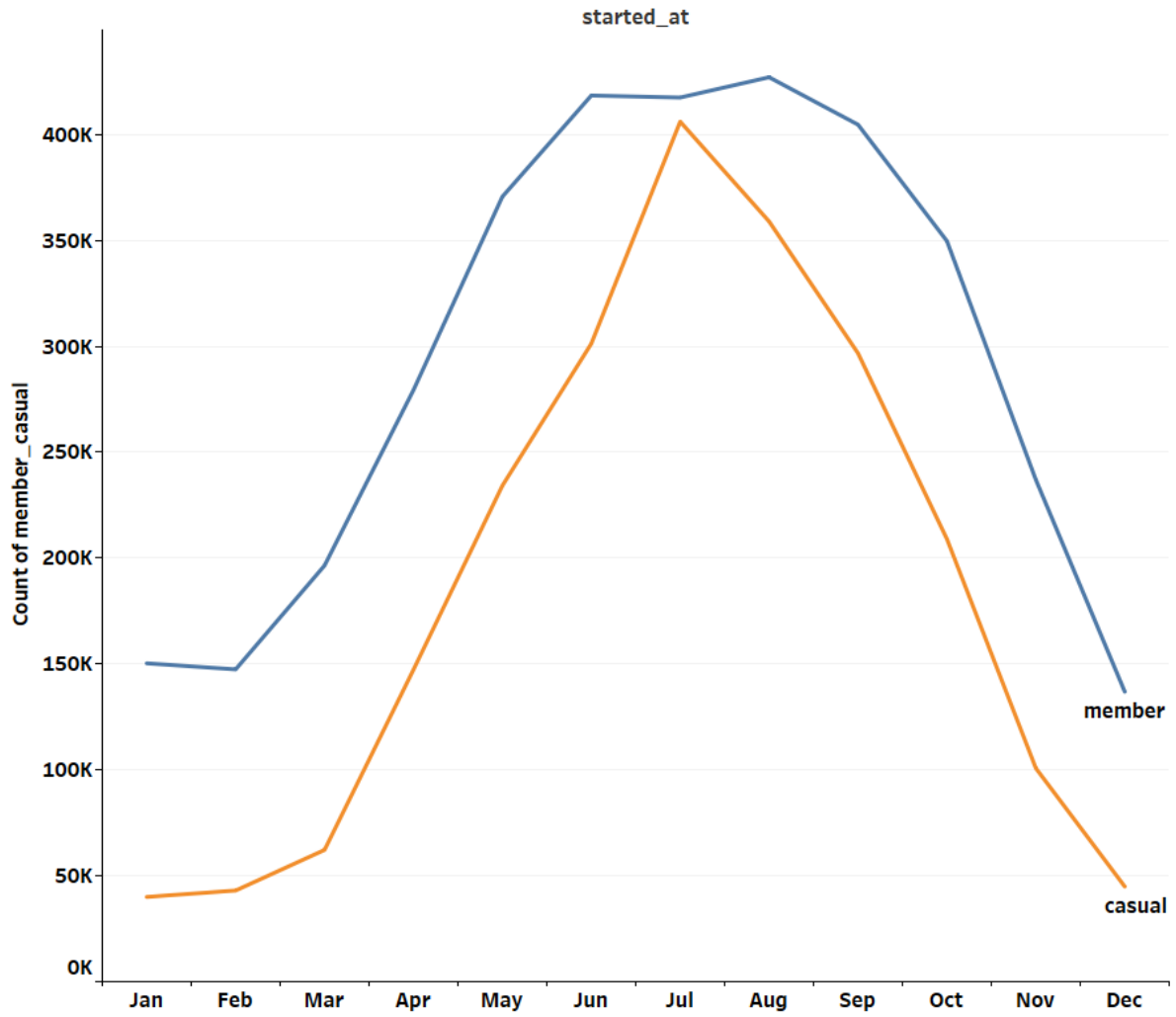
Summer (June, July, August)

Autum (September, October, November)

month_num	total_count_member	total_count_casual
12	136,901	44,891

1	150,287	40,005
2	147,421	43,014
3	196,465	62,193
4	279,287	147,270
5	370,610	234,156
6	418,350	301,202
7	417,410	406,021
8	426,974	358,893
9	404,608	296,667
10	349,661	208,962
11	236,962	100,772

Seasonal Change Effect on Ride Counts



The summarizing table and the line chart show cyclical behavior throughout the year. For both members and casuals, the total ride counts grow and drop seasonally. There is a gradual rise in ride counts from winter (January) to spring (March), reaching their peak in summertime.

Members' ride counts stay high for all summer months (June-August), while casuals' ride counts have a clear peak in July that transcends June and August. However, both eventually drop gradually in September and down to their lowest numbers in the winter month of December. It can be speculated that this behavior is due to weather change, however, to show true connection to seasonal change, more than one year should be analyzed.

Minimum Ride Duration for Casuals and Members

Reminder: ride duration was calculated as follows:

=ROUND(IF(F2>=E2, (F2-E2)*24*60, (1-E2+F2)*24*60),0)

The round function rounds the value to the nearest integer (i.e., ±30 seconds).

```
SELECT [member_casual], MIN([ride_length_min]) min_ride_duration_minutes
FROM [dbo].[BikeTrips]
GROUP BY [member_casual]
```

Results

member_casual	min_ride_duration_minutes
member	0
casual	0

Minimum rides count

```
SELECT [member_casual], count([ride_id]) count_min_rides_minutes
FROM [dbo].[BikeTrips]
WHERE [ride_length_min] = (SELECT MIN([ride_length_min]) FROM [dbo].[BikeTrips] WHERE
[member_casual] = 'casual') AND [member_casual] = 'casual'
GROUP BY [member_casual]

UNION

SELECT [member_casual], count([ride_id])
FROM [dbo].[BikeTrips]
WHERE [ride_length_min] = (SELECT MIN([ride_length_min]) FROM [dbo].[BikeTrips] WHERE
[member_casual] = 'member') AND [member_casual] = 'member'
GROUP BY [member_casual]
```

Results

member_casual	count_min_rides_minutes
casual	34,170
member	61,719

The result sums up to 95,889 records, which is 1.66% of all records. Meaning, 1.66% of all records have up to 30-second-long ride.

It could be determined arbitrarily that 60 seconds long rides were considered as potentially false starts or users that re-docked their bike to ensure it was secure. “ride_length_sec” column was used in order to understand how significant the percentage of records who met this condition. “ride_length_sec” column holds the true value of ride length without any manipulations (the numbers are not rounded).

```
SELECT count(*) up_to_60_sec
FROM [dbo].[BikeTrips]
where [ride_length_sec] < 60
```

Results

up_to_60_sec
148,817

2.56% of all data has up to 60 seconds ride duration. This percentage is small and can be considered insignificant, however on demand, it could have been removed. For the sake of this specific analysis and report, these records were kept.

Maximum ride length for casuals and members

```
SELECT [member_casual], MAX([ride_length_min]) max_ride_duration_minutes
FROM [dbo].[BikeTrips]
GROUP BY [member_casual]
```

Results

member_casual	max_ride_duration_minutes
member	1560
casual	41387

These two values show 2 cases of more than one day usage. It is not clear if these numbers were inserted incorrectly into the report or if these numbers are legitimate. It is possible for users to return the bike to a station whenever they decide, however, these numbers probably don’t represent users’ behavior in general. To understand better the most common ride duration for most users it is better to examine different ranges of time rides.

Examining Different Ranges of Time Rides

10 minutes

```
SELECT
	SUM(CASE WHEN [ride_length_min] >= 0 AND [ride_length_min] < 10 AND [member_casual]
= 'member' THEN 1 ELSE 0 END) first_10_min_members,
	SUM(CASE WHEN [ride_length_min] >= 0 AND [ride_length_min] < 10 AND [member_casual]
= 'casual' THEN 1 ELSE 0 END) first_10_min_casuals
FROM [dbo].[BikeTrips]
```

Results

first_10_min_members	first_10_min_casuals	Total
1,950,683	882,514	2,833,197

30 minutes

Expanding the range:

```
SELECT
	SUM(CASE WHEN [ride_length_min] >= 0 AND [ride_length_min] < 30 AND [member_casual]
= 'member' THEN 1 ELSE 0 END) first_30_min_members,
	SUM(CASE WHEN [ride_length_min] >= 0 AND [ride_length_min] < 30 AND [member_casual]
= 'casual' THEN 1 ELSE 0 END) first_30_min_casuals
FROM [dbo].[BikeTrips]
```

Results

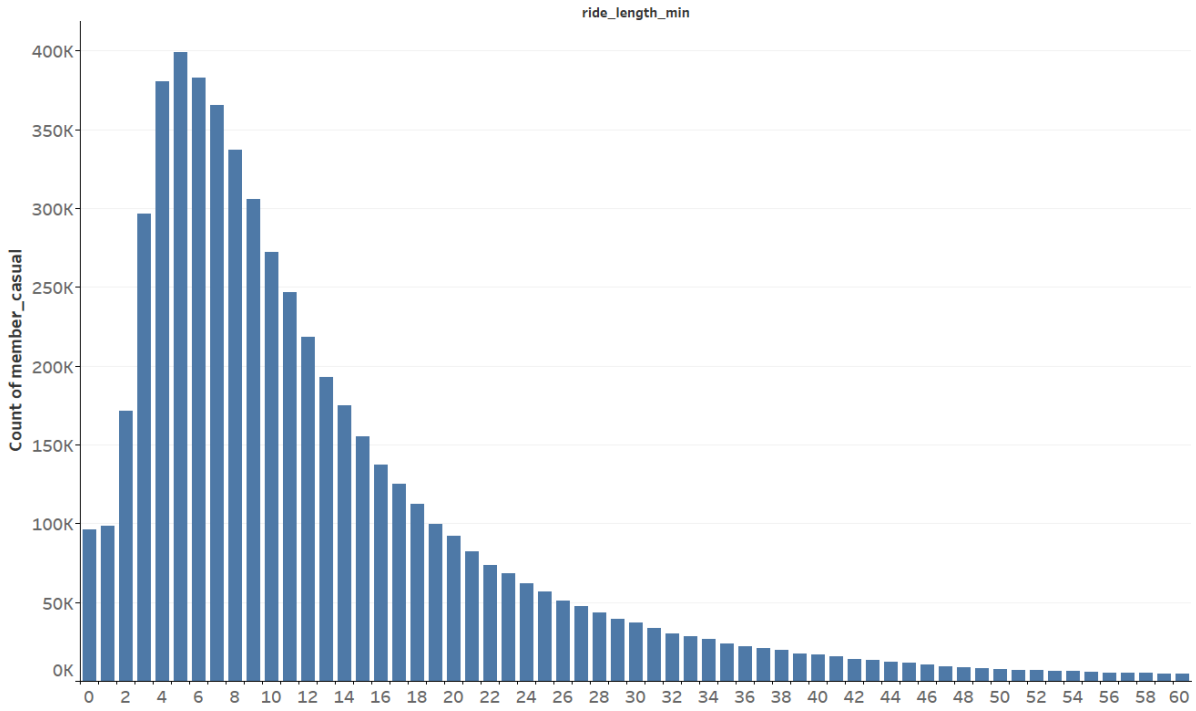
first_30_min_members	first_30_min_casuals	Total
3,321,288	1,864,211	5,185,499

1 hour

hour_ride_members	hour__ride_casuals	Total
3,509,534	2,116,605	5,626,139

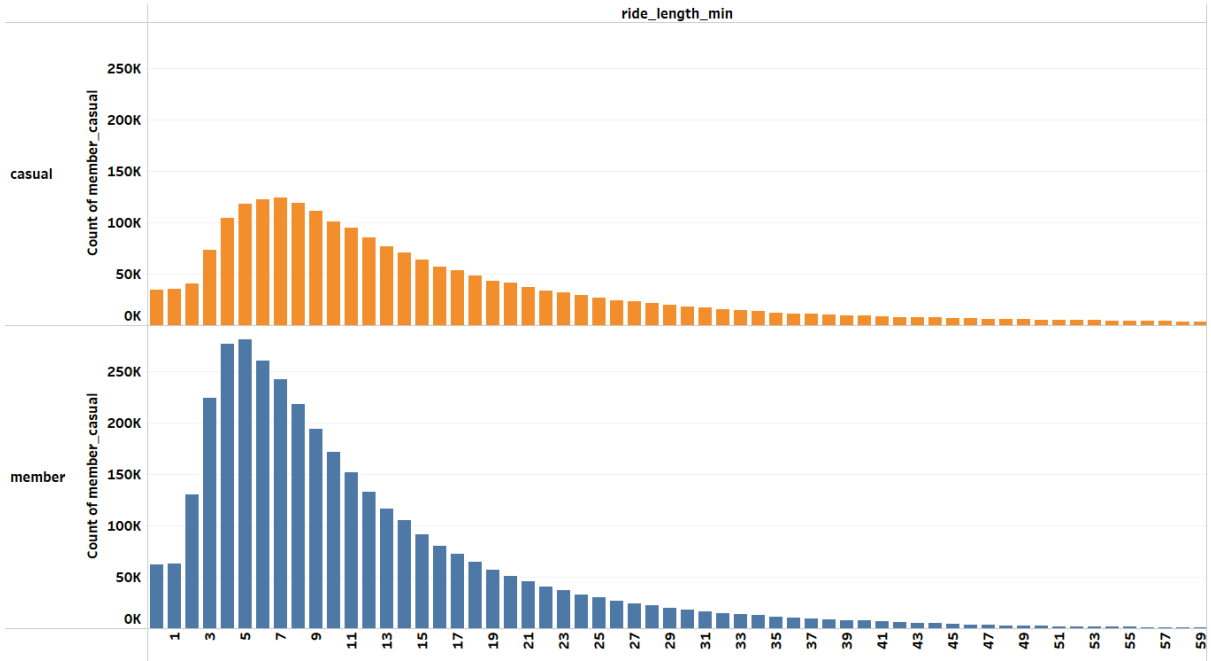
The total count is 5,626,139 which is almost the total number of rides the dataset contains (5,778,982). This proves that approximately all rides were up to 1 hour long.

Distribution of Ride Duration for All Users



The following distribution encapsulates ride duration for all users. As was mentioned above, most users ride for up to 1 hour and therefore, the x-axis was limited to 60 minutes. It can be concluded from the distribution that the most common travel time is 5 minutes. Also, it is clear from the chart that right after the summit, the distribution exhibits a gradual decrease, minute by minute, in the number of rides.

Distribution of Ride Duration Divided by User Type



When divided into user types, casuals and members exhibit similar behavior as shown in the chart above. Casuals' and members' most common travel times are 7 and 5 minutes respectively.

Checking Different Time Intervals

Ride Distribution in 15-Minute-Long Intervals (Members)

To gain another perspective on ride distribution among members, rides were divided into 15-minute-long intervals.

```
SELECT
	SUM(CASE WHEN [member_casual] = 'member' AND [ride_length_min] >=0 AND
[ride_length_min] < 15 THEN 1 ELSE 0 END) intvl_1,
	SUM(CASE WHEN [member_casual] = 'member' AND [ride_length_min] >=15 AND
[ride_length_min] < 30 THEN 1 ELSE 0 END) intvl_2,
	SUM(CASE WHEN [member_casual] = 'member' AND [ride_length_min] >=30 AND
[ride_length_min] < 45 THEN 1 ELSE 0 END) intvl_3,
	SUM(CASE WHEN [member_casual] = 'member' AND [ride_length_min] >=45 AND
[ride_length_min] < 60 THEN 1 ELSE 0 END) intvl_4
FROM [dbo].[BikeTrips]
```

Results

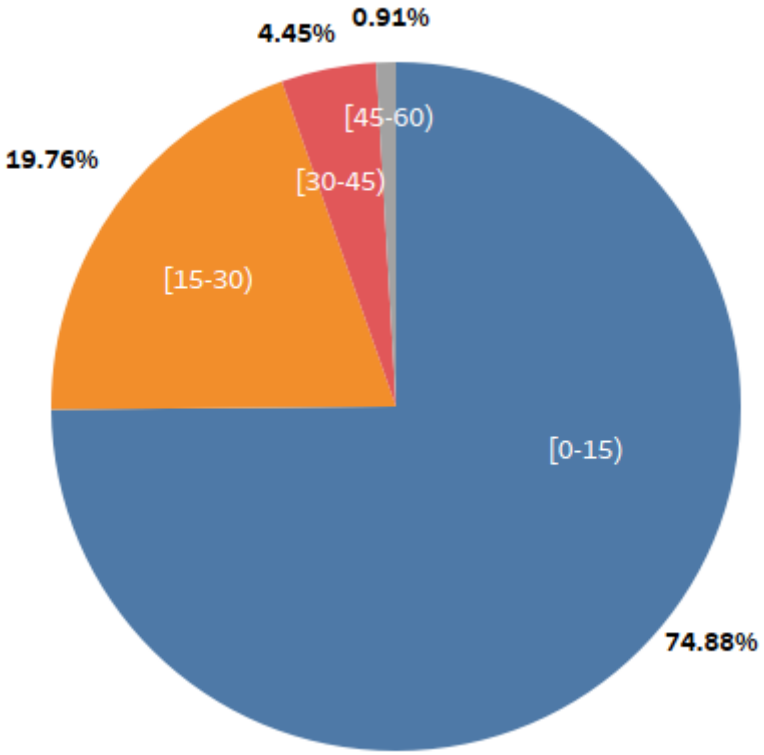
Intvl_1 [0 – 15)	Intvl_2 [15-30)	Intvl_3 [30-45)	Intvl_4 [45-60)
2,627,854	693,434	156,228	32,018
74.88%	19.76%	4.45%	0.91%

The total number of rides in the first hour is: 3,509,534

The first 15 minutes [0-15) hold 74.88% of rides. 19.76% of rides last to the second interval [15-30), 4.45% of rides last to the third interval [30-45), and only 0.91% of rides last to the final interval [45-60).

Annual users use Cyclistic bikes mostly for up to 15-minute-long rides, and as the ride duration increases, the number of rides drops drastically.

The examined time intervals can be viewed in the following pie chart:



Ride Distribution in 15-Minute-Long Intervals (Casuals)

```
SELECT
        SUM(CASE WHEN [member_casual] = 'casual' AND [ride_length_min] >=0 AND
[ride_length_min] < 15 THEN 1 ELSE 0 END) intvl_1,
        SUM(CASE WHEN [member_casual] = 'casual' AND [ride_length_min] >=15 AND
[ride_length_min] < 30 THEN 1 ELSE 0 END) intvl_2,
        SUM(CASE WHEN [member_casual] = 'casual' AND [ride_length_min] >=30 AND
[ride_length_min] < 45 THEN 1 ELSE 0 END) intvl_3,
        SUM(CASE WHEN [member_casual] = 'casual' AND [ride_length_min] >=45 AND
[ride_length_min] < 60 THEN 1 ELSE 0 END) intvl_4
FROM [dbo].[BikeTrips]
```

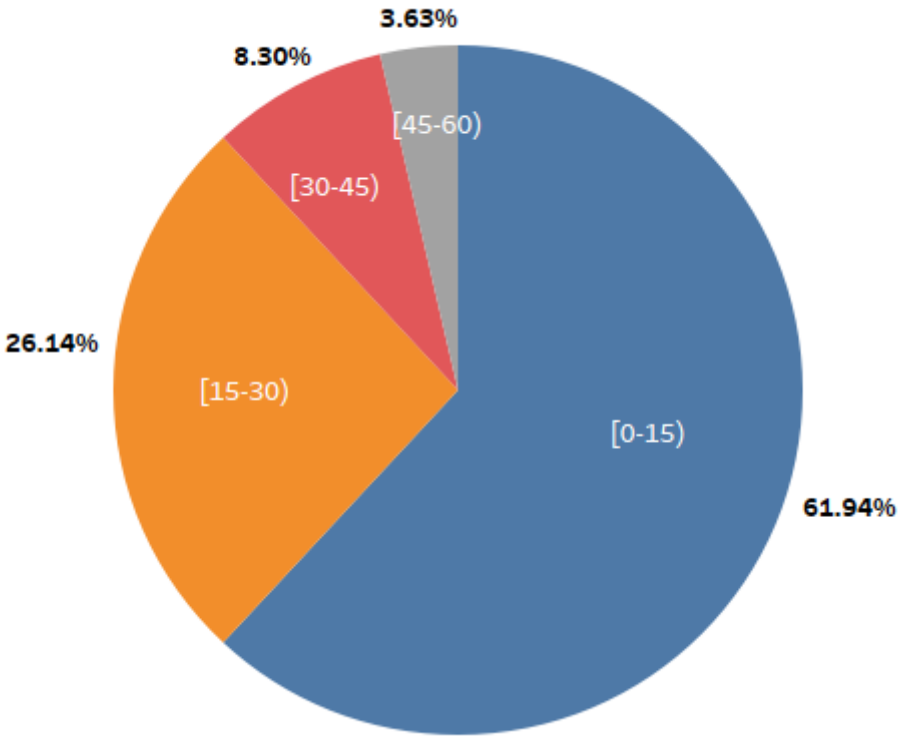
Results

intvl_1 [0 – 15)	intvl_2 [15 – 30)	intvl_3 [30 – 45)	intvl_4 [45 – 60)
1,311,017	553,194	175,629	76,765
61.94%	26.14%	8.30%	3.63%

The total number of rides in the first hour is: 2,116,605

The first 15 minutes [0-15) hold 61.94% of rides. 26.14% of rides last to the second interval [15-30), 8.3% of rides last to the third interval [30-45), and only 3.63% of rides last to the final interval [45-60).

Similar effect happens both in casuals and members, the longer the ride duration the smaller the ride count. As shown in the following pie chart, most casual users use Cyclistic bikes for up to 15-minute-long rides.



Members and Casuals - Average Trip Duration

```
SELECT
	ROUND(AVG(CASE WHEN [member_casual] = 'member' THEN [ride_length_min] END), 2)
avg_ride_members,
	ROUND(AVG(CASE WHEN [member_casual] = 'casual' THEN [ride_length_min] END), 2)
avg_ride_casuals
FROM [dbo].[BikeTrips]
```

Results

avg_ride_members	avg_ride_casuals

12.4	27.76

On average, members' travel time is lower than casuals' travel time.

4. What is the rush hour of members and casuals?

For members

```
SELECT TOP 10 FORMAT([started_at],'HH:mm') ride_start_time, COUNT([ride_id])
count_start_time_member
FROM [dbo].[BikeTrips]
WHERE [member_casual] = 'member'
GROUP BY FORMAT([started_at],'HH:mm')
ORDER BY count_start_time_member DESC
```

Results

ride_start_time	count_start_time_member
17:06	6,879
17:08	6,879
17:10	6,852
17:16	6,786
17:11	6,759
17:17	6,736
17:14	6,711
17:09	6,708
17:05	6,697
17:15	6,672

For casuals

```
SELECT TOP 10 FORMAT([started_at],'HH:mm') ride_start_time, COUNT([ride_id])
count_start_time_casual
FROM [dbo].[BikeTrips]
WHERE [member_casual] = 'casual'
GROUP BY FORMAT([started_at],'HH:mm')
ORDER BY count_start_time_casual DESC
```

Results

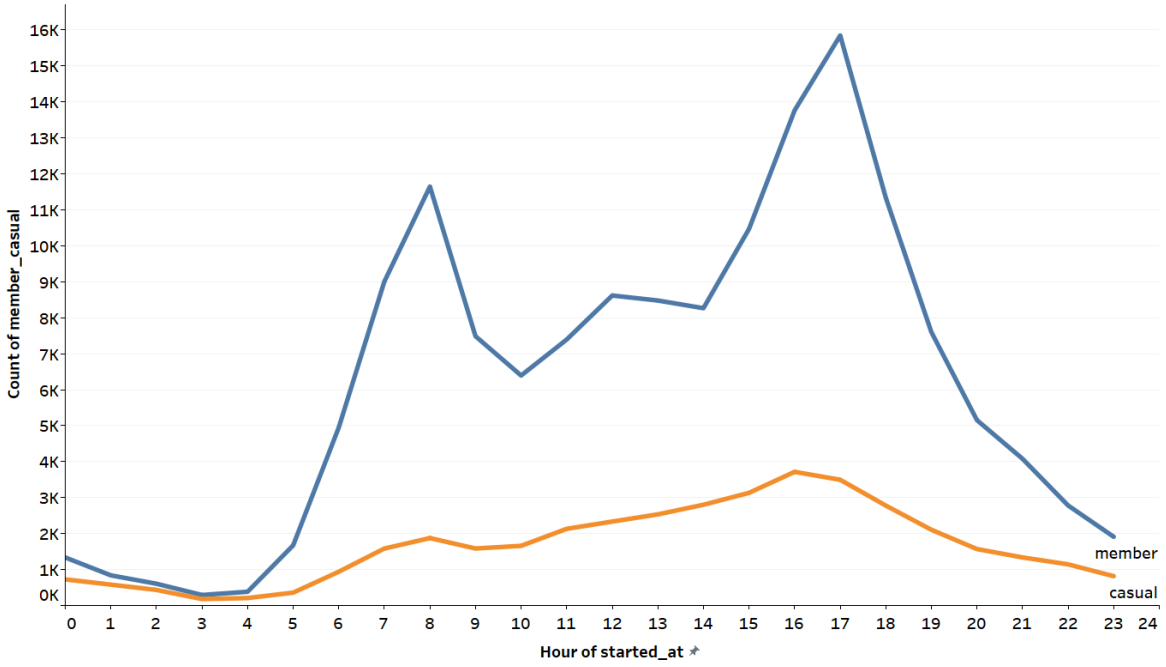
ride_start_time	count_start_time_casual
17:06	3,838
17:07	3,822
17:12	3,795

17:08	3,788
17:19	3,770
17:10	3,768
17:16	3,749
17:23	3,746
17:18	3,740
17:05	3,733

Members and casuals have their rush hours approximately at the same time, around 5 o'clock in the afternoon. However, this result is an overall view of the whole year. Therefore, time distributions will be a better representative of users' behavior during the day.

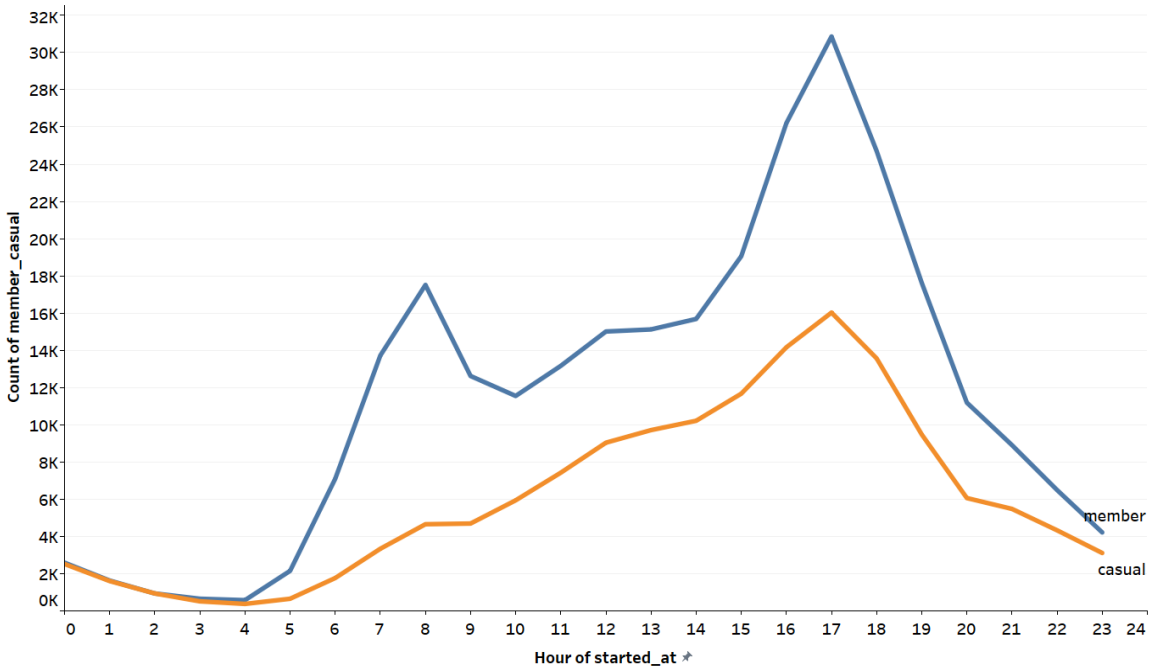
For this purpose, 4 different months were chosen; January, April, July, and October. These months were chosen for being exactly mid-season.

Start Time Distribution – January



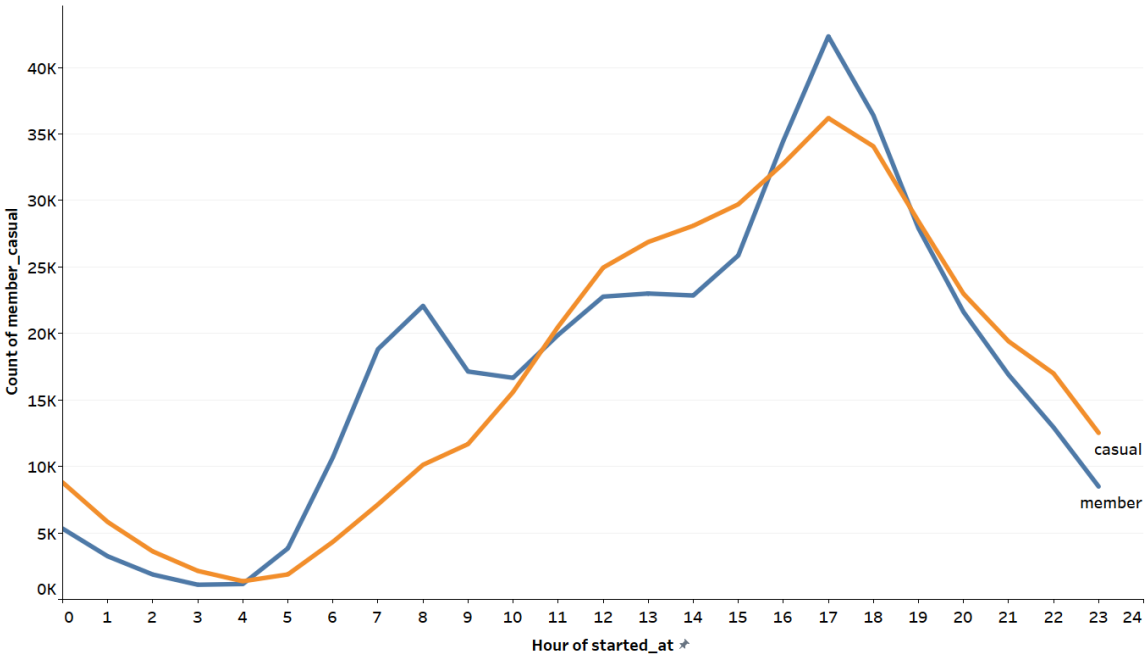
Members and casuals both have peaks at 8. Members have another lower clear peak at 12 followed by a mild decrease. Afterwards, the chart keeps on climbing back to its highest peak at 17. And then there is a drastic drop till the end of the day. As for the casuals' chart, after the first peak at 8, the chart exhibits an insignificant drop, followed by a gradual increase from 10 to 16 where its highest peak can be spotted. Afterwards, there is a gradual decrease till the end of the day. Overall, members and casuals follow the same trend and have similar behaviors.

Start Time Distribution – April



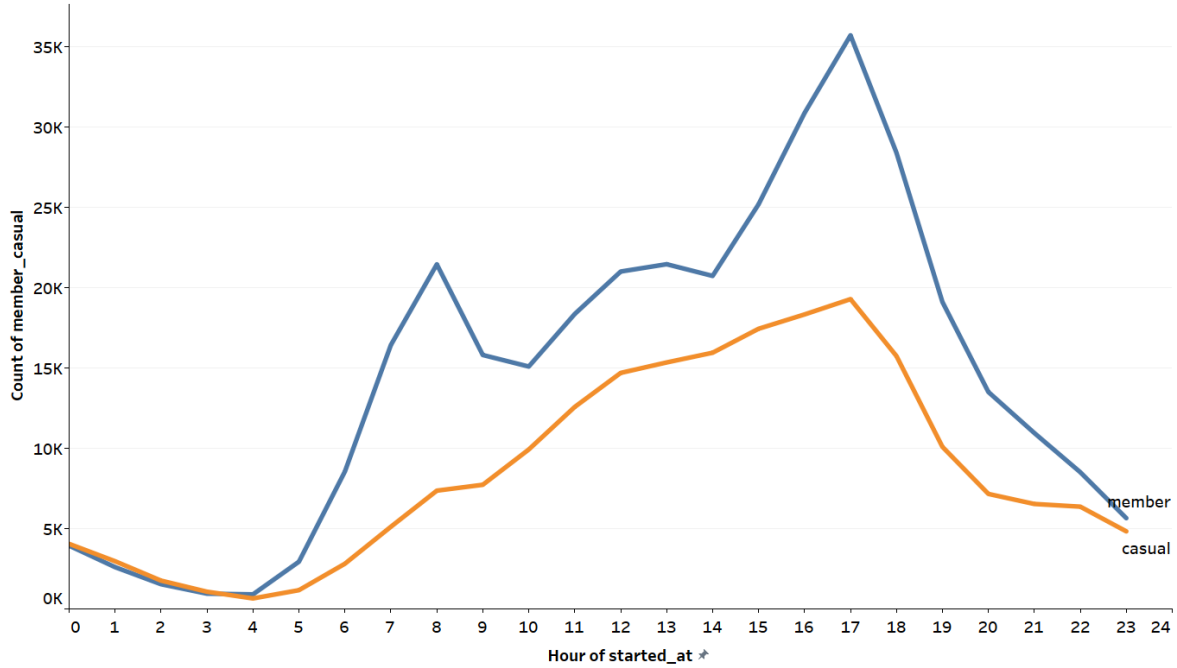
In April, members exhibit the same pattern as in January by having visible peaks at 8, 12, and 17 followed by a drastic drop. As for the casuals' chart, there is a peak at 8 and a gradual increase until it reaches an inflection point at 12 and its highest peak at 17 which is also, as in members, followed by a drastic drop. Overall, in April as in January, members and casuals follow the same trend and have similar behaviors.

Start Time Distribution – July



In July, members exhibit the same pattern as in January and April by having visible peaks at 8, 12, and 17 followed by a drastic drop. Casuals on the other hand, exhibit a gradual increase from 4 with two clear inflection points at 8 and 12 and a peak at 17, also followed by a drastic drop. Overall, in July, members and casuals also follow the same trend and have similar behaviors as spotted in January and April.

Start Time Distribution – October

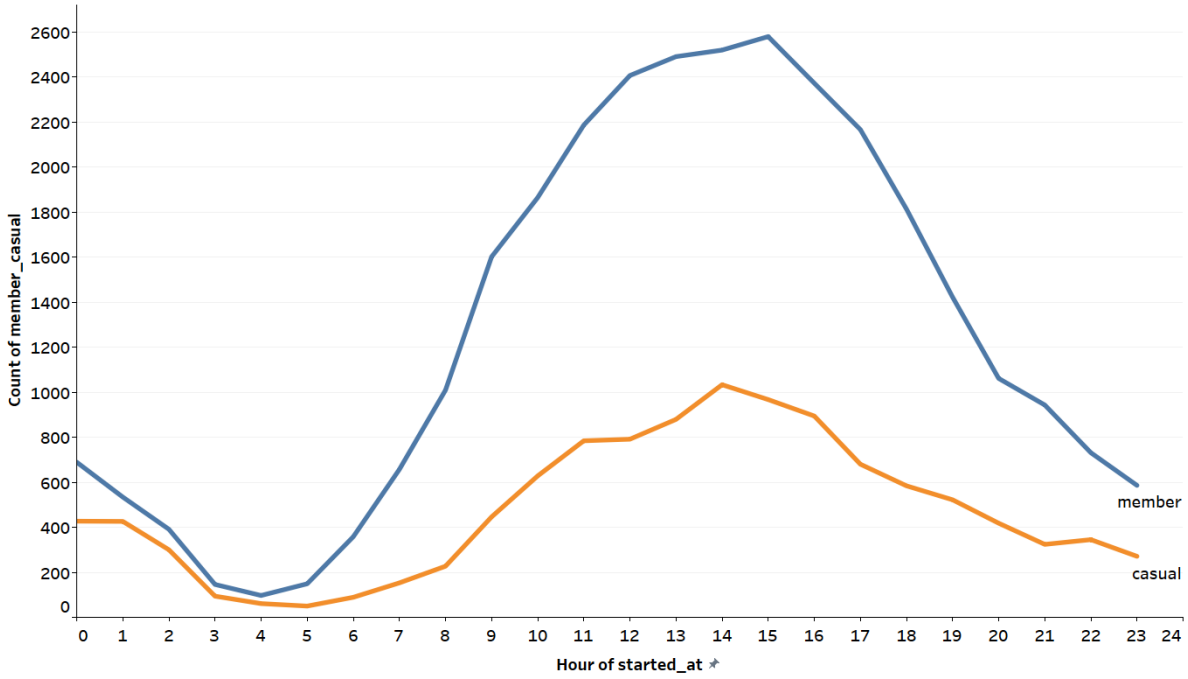


In October, members exhibit the same pattern as in previous charts by having visible peaks at 8, 12-13, and 17 followed by a drastic drop. Casuals, as in the month of July, exhibit a gradual increase from 4 with two clear inflection points at 8 and 12 and a peak at 17 followed by a drastic drop. Overall, in October, members and casuals also follow the same trend and have similar behaviors as spotted in previous charts.

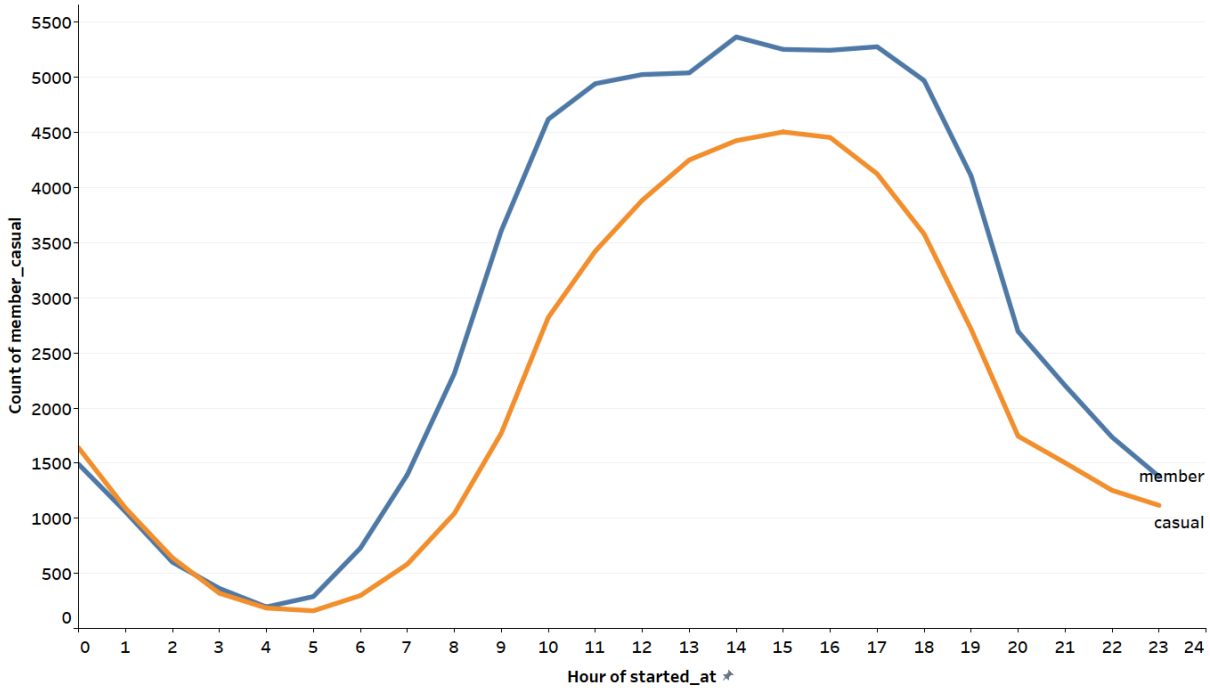
In conclusion, there are several repeated time points that stand out from all charts; 8, 12, and 17 o'clock when 17 o'clock exhibits the highest peak in all charts, regardless of user's type. However, let us not forget that these charts present a cumulative sum that adds up ride counts for each hour during the whole month. Therefore, individual trends (e.g., over the weekends) are probably masked and not represented.

The following charts show the rush hour trends over the weekends (Saturday and Sunday)

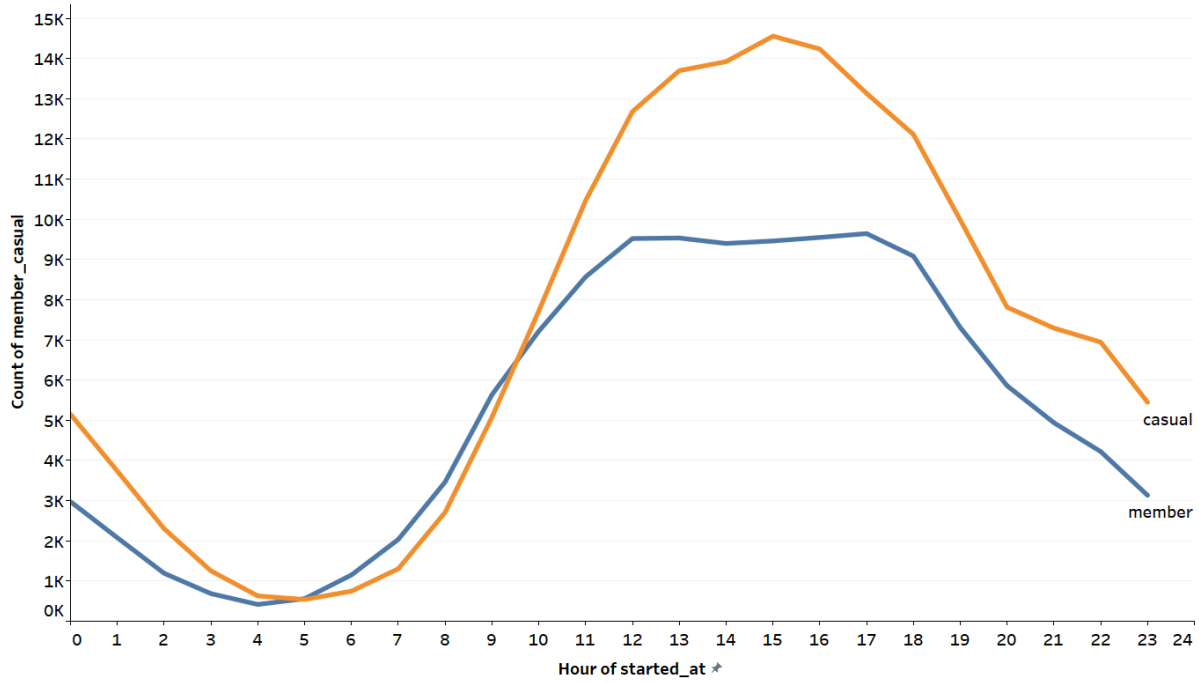
January



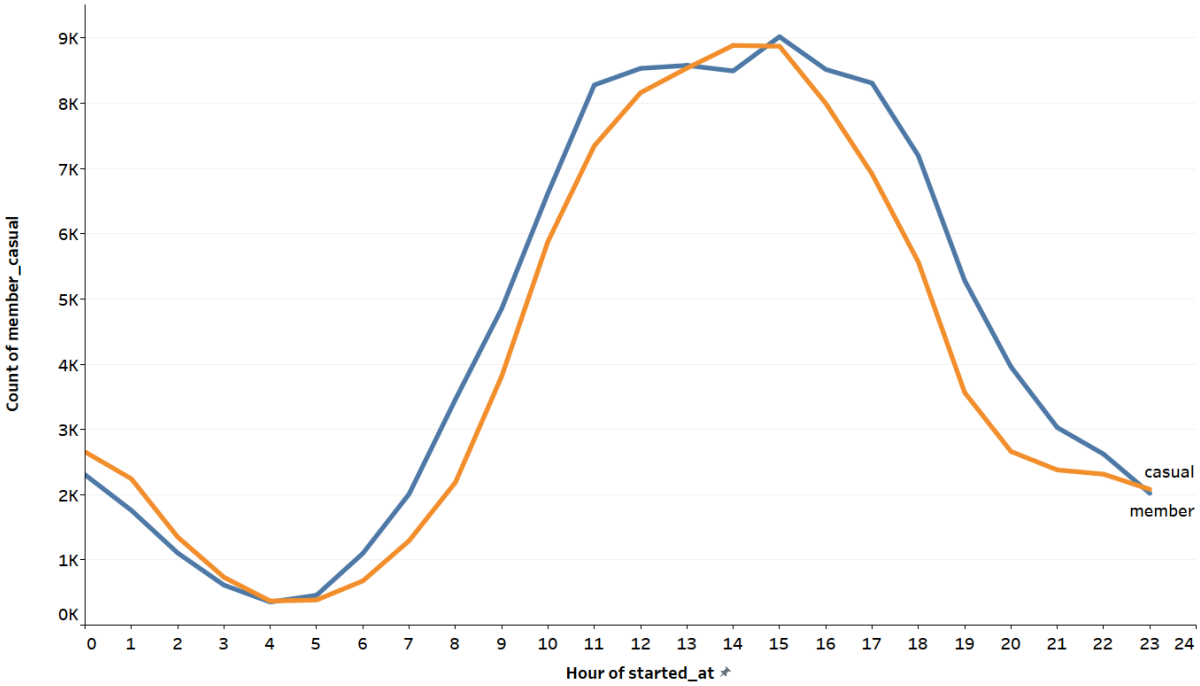
April



July



October



When exploring the trends over the weekends only, they are completely different from the ones presented in previous charts and yet, all distributions are similar to one another. All distributions resemble more of a normal distribution shape where the peak is reached in the

afternoon hours. In overall perspective, members and casuals follow the same trend and display similar behaviors over the weekends.

5. What are the most popular bike share stations?

```
SELECT TOP 10 [start_station_name], COUNT([member_casual]) count_start_st
FROM [dbo].[BikeTrips]
GROUP BY [start_station_name]
ORDER BY count_start_st DESC
```

Results

start_station_name	count_start_st
NULL	857,853
Streeter Dr & Grand Ave	70,030
DuSable Lake Shore Dr & Monroe St	40,130
Michigan Ave & Oak St	38,678
DuSable Lake Shore Dr & North Blvd	37,755
Wells St & Concord Ln	36,775
Clark St & Elm St	35,510
Kingsbury St & Kinzie St	34,439
Millennium Park	33,126
Theater on the Lake	31,544

There is a significant number of NULL values (14.84% of all data) that could not be removed or manipulated during the cleaning process. Therefore, the most common start station name cannot be evaluated properly either for members or for casual.

Final Conclusions

The marketing team's objective was to increase the percentage of annual users by revealing the differences between annual and casual users.

As suspected, annual users ("members") and casual users ("casuals") were found to be similar in some inspected areas however different in others.

The first finding examined the different bike types used by members and casuals. It was found that members used more electric bikes than classic bikes, however the difference was relatively small (52.17% and 47.83% respectively). Casuals also used more electric bikes than classic and docked bikes (57.85%, 35.84%, and 6.31% respectively) however with a greater difference between them. Overall, members had a greater percentage of ride counts in both classic and electric bikes. Yet, for some unknown reason, docked bikes were only used by

casuals, and still in a very small percentage. Based on these findings, it seemed that docked bikes are the least favorite choice for both types of users.

When members and casuals were examined in relation to different days of the week, they exhibited opposite trends. However, if divided into bike types, both members and casuals displayed the same trends of usage inside their group of users.

Another finding showed similar behaviors during the year, in relation to ride durations and rush hours. However, the average ride times were different; 12.4 minutes for members and 27.8 minutes for casuals.

The last field that was examined is the busiest start station. Unfortunately, many of the start station names were NULL values (14.84% of records didn't have start station name) that could not be manipulated during the cleaning phase of the analysis, and therefore the busiest station for members and casuals could not be determined.

In conclusion, the reason for the lack of use of docked bikes in members is unclear and should be examined further. Also, the reason for the opposite trends members and casuals exhibit towards different days of the week is unclear, can be only speculated and should be further examined (by a questionnaire or more precise data) to reveal the cause behind these behaviors.

Above all, it is important to point out the limitations of this data. For example, when looking at long time periods it can mask other short-term behaviors. These limitations should be taken into consideration when deciding to act upon a particular implication that was suggested by the data.

Recommendations

1. Replace all docked bikes with electric or classic bikes. Granting greater priority to electric bikes.
2. Conduct a survey regarding the purpose of the ride.
3. Collect more data about the start station names.

If docked bikes were replaced with other bike types, it would enlarge the pool of preferable bike types for both members and casuals which can help to increase the number of users in general. Once the preferable bikes are enlarged, Cyclistic transportation will be more available which can play a role in casual's decision to become a member. Secondly, if the purpose of the rides is revealed, more bikes can be stationed near strategic locations such as subway/train/bus station. Once again it might make bike share transportation more available and encourage casual users to use it in the same way members do. Lastly, it is crucial to understand which stations are the busiest to enlarge their bikes amount. This move will also help with the availability of Cyclistic transportation and could potentially increase the annual users' percentage due to the increase in the availability it suggests.