Google Data Analytics Capstone Project.

Case Study: Bike-share analysis.

Author: Maria Lykoudi

The following project is part of the Google Data Analytics Professional Certificate. The analysis is conducted for the fictional Cyclistic Bike-Share company. The data analysis in this project adheres to a structured approach, as taught in the course, which includes the following six phases: **ask**, **prepare**, **process**, **analyze**, **share**, and **act**.

Phase 1: Ask

1.1 Company's background:

Cyclistic was launched in Chicago in 2016. Until now, Cyclistic has demonstrated remarkable growth, expanding its fleet to a total of 5,824 bicycles with 692 stations. The bikes offered by Cyclistic can be unlocked from one station and returned to any other station within the system at any time. Enabling customers to conveniently rent and return bicycles at their desired locations.

The company provides customers with flexible pricing plans, including options such as

A. **single ride passes**
B. **full-day passes**
C. **annual memberships** (subscribers)

Customers who opt for the single-ride passes or full-day passes are categorized as **casual riders**, while those who choose the annual membership are referred to as **members**. The company offers bikes that cater to a wide range of people's needs, including standard two-wheeled bikes, reclining bikes, hand tricycles, and cargo bikes.

1.2 Stakeholders:

- **Liliy Moreno:** Director of marketing, responsible for organizing campaigns that will generate a steady stream of revenue for the company.
- **Cyclistic marketing analytics team**: By leveraging advanced analytics tools they aim to gain insights into rider behavior, patterns, and preferences.
- **Cyclistic executive team:** The team assumes leadership roles within the company, providing strategic and operational guidance.

1.3 Business task / stakeholders' expectations:

- The company aims to increase its revenue.
- The strategy is to establish a more loyal customer base by converting casual riders into annual members.

1.4 Research objective:

- To draw accurate conclusions on how subscribers and casual riders use Cyclistic bikes differently?

Phase 2: Prepare

Data source: Third-party data provided by Motivate International Inc.

Data location: The data is stored as zipped files on the publicly available Amazon Web Service (AWS) https://divvy-tripdata.s3.amazonaws.com/index.html.

Data license: Data is available by Motivate International Inc. under the following license.

Data privacy: Personal identifying information is subjected to anonymization to ensure the protection of sensitive data (financial data, address etc.).

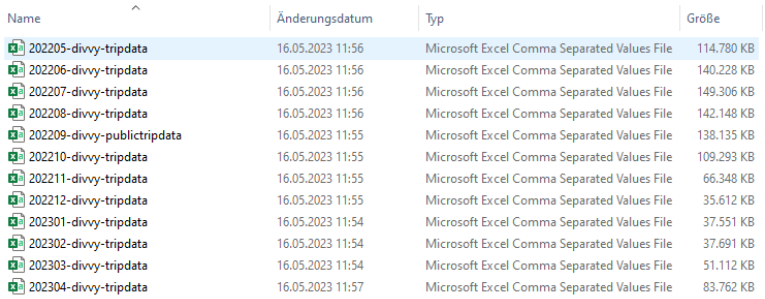
Data credibility:

I. **Reliable**: The data's high reliability is supported by the large sample size of over 5 million rows.
II. **Original**: Since the data is sourced from a third party, its originality is categorized as low.
III. **Comprehensive**: The data set is highly comprehensive because it contains all critical information to make an informed decision with respect to riders' cycling patterns.
IV. **Current**: The data is considered current as it covers the period between May 2022 and April 2023.
V. **Cited**: The data source is not cited.

Phase 3: Process

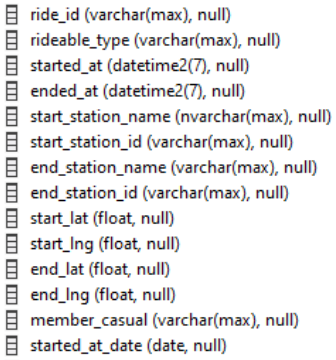
At this phase data is subjected to cleaning so to be ready for the analysis.

Data is located in 12 different files which correspond to a different month. The files are in a CSV format and all files contain the same set of attributes.



Name	Änderungsdatum	Typ	Größe
202205-divvy-tripdata	16.05.2023 11:56	Microsoft Excel Comma Separated Values File	114.780 KB
202206-divvy-tripdata	16.05.2023 11:56	Microsoft Excel Comma Separated Values File	140.228 KB
202207-divvy-tripdata	16.05.2023 11:56	Microsoft Excel Comma Separated Values File	149.306 KB
202208-divvy-tripdata	16.05.2023 11:56	Microsoft Excel Comma Separated Values File	142.148 KB
202209-divvy-publictripdata	16.05.2023 11:55	Microsoft Excel Comma Separated Values File	138.135 KB
202210-divvy-tripdata	16.05.2023 11:55	Microsoft Excel Comma Separated Values File	109.293 KB
202211-divvy-tripdata	16.05.2023 11:55	Microsoft Excel Comma Separated Values File	66.348 KB
202212-divvy-tripdata	16.05.2023 11:55	Microsoft Excel Comma Separated Values File	35.612 KB
202301-divvy-tripdata	16.05.2023 11:54	Microsoft Excel Comma Separated Values File	37.551 KB
202302-divvy-tripdata	16.05.2023 11:54	Microsoft Excel Comma Separated Values File	37.691 KB
202303-divvy-tripdata	16.05.2023 11:54	Microsoft Excel Comma Separated Values File	51.112 KB
202304-divvy-tripdata	16.05.2023 11:57	Microsoft Excel Comma Separated Values File	83.762 KB

The set of attributes are:

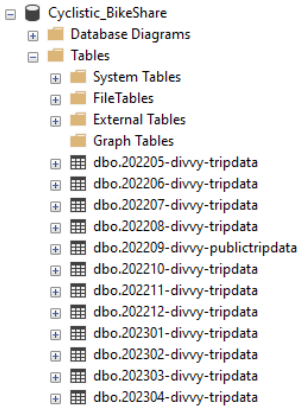


Tools used for the analysis.

1. SQL server
2. Power BI

Given the substantial size of the dataset (5,089,857 rows), I opted to perform data cleaning and processing in SQL Server. I used Power BI for data analysis and visualization, benefiting from its good integration with SQL Server.

3.1. Data importing in SQL Server:



3.2. Data Unification into one table:

3.3. Data cleaning in SQL Server:

Overall, the data was **complete**, **correct**, **relevant** to the business objectives, and did not present any significant flaws. However, there was no information regarding the type of passes purchased by casual customers. Specifically, there was no clarification on the following:

A. Single-ride passes

B. Full-day passes

- Checking for duplicates:

There were no duplications.

```
select ride_id, count(*) as DublicateValues
from TRIPDATA
Group by ride_id
Having Count(*)>1;
```

- Checking for completeness:

Null values were found in the following attributes.

```sql
select count(*) AS Num_Null
from TRIPDATA
where start_station_name is null;

select count(*) AS Num_Null
from TRIPDATA
where start_station_id is null;

select count(*) AS Num_Null
from TRIPDATA
where end_station_name is null;

select count(*) AS Num_Null
from TRIPDATA
where end_station_id is null;

SELECT count(*) AS Num_Null
from TRIPDATA
where end_lat is null;

select count(*) AS Num_Null
from TRIPDATA
where end_lng is null;
```

➢ Checking for consistency:

- There we no spelling errors or other potential errors encountered. Example:

```sql
select distinct rideable_type
from TRIPDATA
order by rideable_type;
```

	rideable_type
1	classic_bike
2	docked_bike
3	electric_bike

- Dates were formatted consistently throughout the database as datetime 2.

Column Name	Data Type
started_at	datetime2(7)
ended_at	datetime2(7)

- I controlled for extra spaces by including the Trim into the Update query, *see below*.
- In the attributes ***start_station_name*** and ***start_station_id*** and their respective end stations the id columns lack of consistency and some contain both letters and numbers:

```sql
select distinct top 10 start_station_name, start_station_id, end_station_name, end_station_id
from TRIPDATA
```

	start_station_name	start_station_id	end_station_name	end_station_id
1	Wabash Ave & Grand Ave	TA1307000117	Halsted St & Roscoe St	TA1309000025
2	DuSable Lake Shore Dr & Monroe St	13300	Field Blvd & South Water St	15534
3	Clinton St & Madison St	TA1305000032	Wood St & Milwaukee Ave	13221
4	Clinton St & Madison St	TA1305000032	Clark St & Randolph St	TA1305000030
5	Clinton St & Madison St	TA1305000032	Morgan St & Lake St	TA1306000015
6	Carpenter St & Huron St	13196	Sangamon St & Washington Blvd	13409
7	Noble St & Milwaukee Ave	13290	Wood St & Augusta Blvd	657
8	Halsted St & Wrightwood Ave	TA1309000061	Southport Ave & Clybourn Ave	TA1309000030
9	Clinton St & Madison St	TA1305000032	Clybourn Ave & Division St	TA1307000115
10	Southport Ave & Waveland Ave	13235	N Southport Ave & W Newport Ave	20257.0

- Update the format of the *rideable types* for improving readability in the visualizations.
- Replace the null values with 'Unknown' for character data types and with '0' for integer data types. This decision was made considering that the current project is a key step towards obtaining the current certification and does not reflect a real business setting with active stakeholders involved.

To ensure the removal of any potential additional spaces, I apply trimming as part of the cleaning process.

```
UPDATE TRIPDATA SET rideable_type = (case when rideable_type = 'classic_bike' then 'Classic Bike'
            when rideable_type = 'docked_bike' then 'Docked Bike'
            when rideable_type = 'electric_bike' then 'Electric Bike'
            end), start_station_name = ISNULL(start_station_name, 'Unknown'), end_station_name = ISNULL(end_station_name,'Unknown'),
            start_station_id = ISNULL(start_station_id, '0'), end_station_id = ISNULL(end_station_id, '0'),
            member_casual = TRIM(member_casual), started_at = TRIM(started_at), ended_at = TRIM(ended_at)
```

- Confirm that there are no inconsistencies between the start and end dates and ensure that no dates were flipped between the two different attributes:

```
UPDATE TRIPDATA SET started_at = (case when started_at > ended_at then ended_at else started_at end),
                    ended_at =   (case when ended_at > started_at then ended_at else started_at end)
```

- Add up data to prepare for the analysis:
    - From the started_at and ended_at columns I extracted: the date, time, day of the week, month, and quarter, season.

```
ALTER TABLE TRIPDATA ADD started_at_date date;
ALTER TABLE TRIPDATA ADD started_at_time time;
ALTER TABLE TRIPDATA ADD started_at_day varchar(10);
ALTER TABLE TRIPDATA ADD started_at_month varchar(20);
ALTER TABLE TRIPDATA ADD started_at_quarter varchar(2);

ALTER TABLE TRIPDATA ADD ended_at_date date;
ALTER TABLE TRIPDATA ADD ended_at_time time;
ALTER TABLE TRIPDATA ADD ended_at_day varchar(10);
ALTER TABLE TRIPDATA ADD ended_at_month varchar(20);
ALTER TABLE TRIPDATA ADD ended_at_quarter varchar(2);
ALTER TABLE TRIPDATA ADD ride_week_day varchar(15);

ALTER TABLE TRIPDATA ADD ride_season varchar(10);
```

    - I fill up the new columns with data:

```
UPDATE TRIPDATA set started_at_date = CAST(started_at as DATE),
                    ended_at_date = CAST(ended_at as DATE);


UPDATE TRIPDATA set started_at_time = CAST(started_at as TIME),
                    ended_at_time = CAST(ended_at as TIME);

UPDATE TRIPDATA set started_at_day = DAY(started_at),
                    ended_at_day = DAY(ended_at);

UPDATE TRIPDATA set started_at_month = MONTH(started_at),
                    ended_at_month = MONTH(ended_at);

UPDATE TRIPDATA set started_at_quarter = 'Q'+ CAST(DATEPART(quarter, started_at) as varchar),
                    ended_at_quarter = 'Q'+ CAST(DATEPART(quarter, ended_at) as varchar);

UPDATE TRIPDATA set ride_week_day = DATENAME(weekday, started_at);

UPDATE TRIPDATA SET ride_season = case when started_at_month in (12,1,2) then 'winter'
                                       when started_at_month in(3,4,5) then 'spring'
                                       when started_at_month in (6,7,8) then 'summer'
                                       else 'automn' end
```

    - I added the ride_duration column by calculating the difference between starting and ending time.

```
ALTER TABLE TRIPDATA ADD ride_length int;
```

- In cases where time crossed midnight the SQL Server calculated wrong the number of minutes and gave negative numbers. By executing the following query, the problem was fixed:

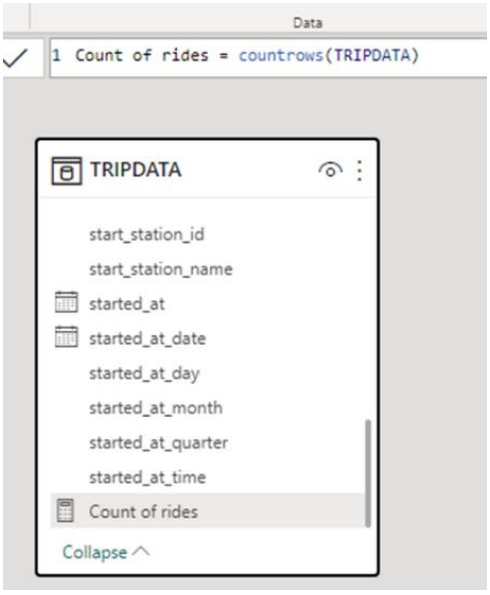
```
UPDATE TRIPDATA SET ride_length = CASE WHEN DATEDIFF(minute, started_at_time, ended_at_time) > 0
                                  THEN DATEDIFF(minute, started_at_time, ended_at_time)
                                  ELSE 1440 + DATEDIFF(minute, started_at_time, ended_at_time) end
```

Phase 4: Analysis

Following data import from SQL Server, the analysis was predominantly conducted using Power BI.

The primary key: ride_id.

For counting the rides, I created a new measure using DAX. The formula is below:



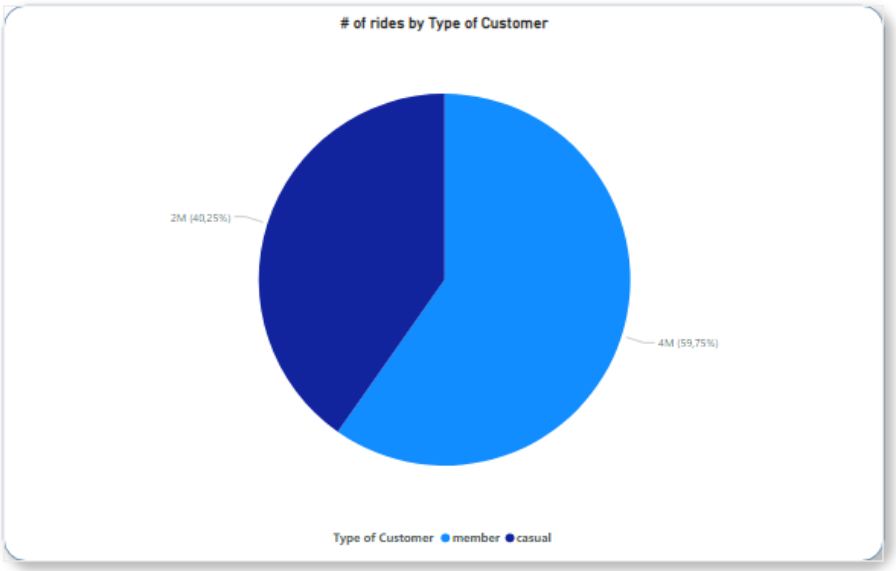
I conducted my analysis to address the following inquiries:

- Count of rides by type of customer.
- Count of rides per customer type and by season.
- Count of rides per ride length group.
- Count of rides by weekday.
- Count of rides by started at month per type of customer.
- Top 10 stations per type of customer.
- Count of rides by starting at hour and type of customer.

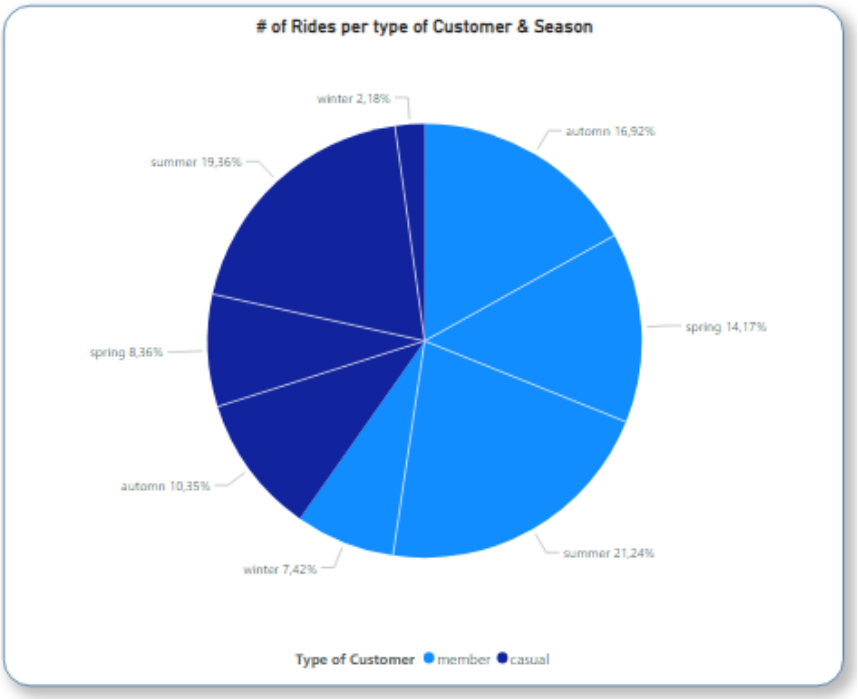
Phase 5: Share

The share phase is the culmination of the data analysis process. Visualizations accompanied by a narrative are presented in order to effectively communicate the data-driven insights.

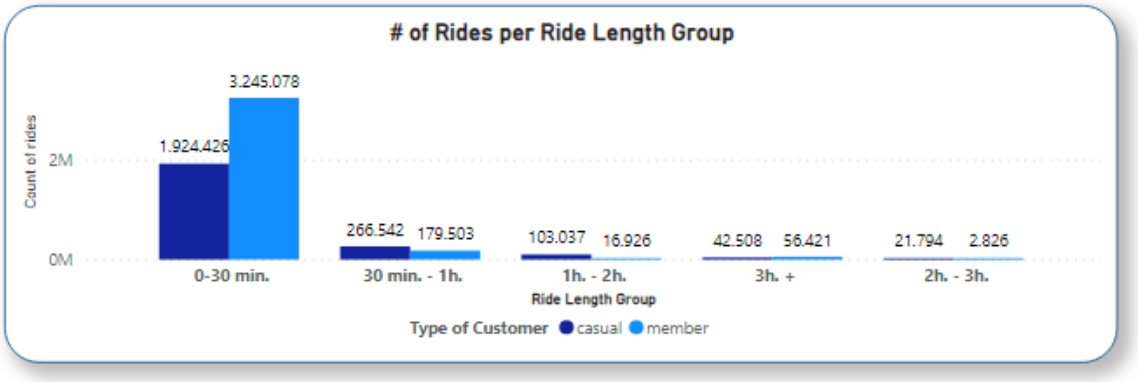
- **Members** surpass the number of casual users and make up **59,75%** of the total.



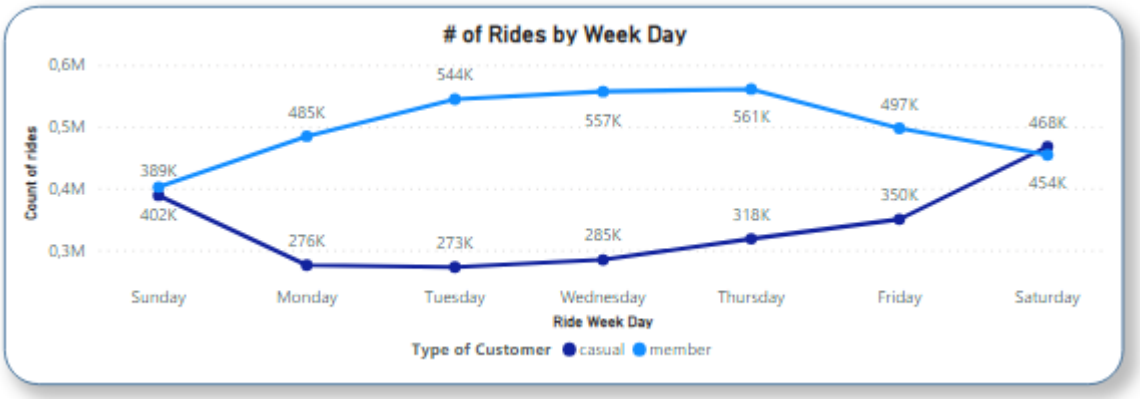
- Spring, summer, and autumn emerge as the most favored seasons for both members and casual riders. Conversely, winter proves to be the least favorable season for both customer segments.



- Both members and casual riders demonstrate a similar pattern in terms of the maximum ride length, with the majority of rides falling within the 0-30 minute range for both groups. However, the most notable difference between casual riders and members lies in the significant disparity in numbers, with members outnumbering casual riders by 1,320,652.
  When considering the remaining ride length groups, casual riders are more likely to spend more time riding compared to members. In the 30-minute to 1-hour ride length group, the rate of casual riders is 2.20 times higher than that of members. Moreover, in the 2-3-hour ride duration category, the rate of casual riders is 11.44 times higher than that of members.

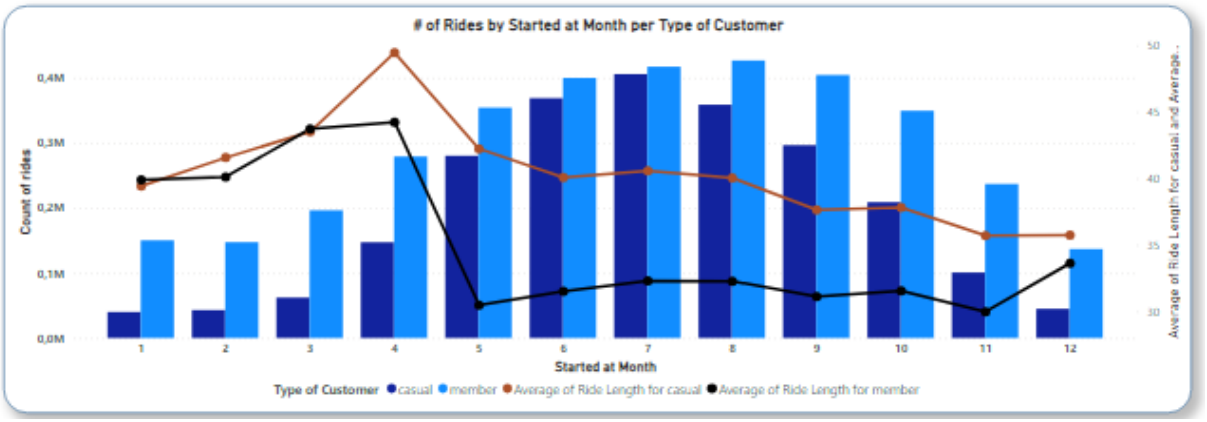


- Casual riders and members demonstrate distinct patterns in bike usage throughout the weekdays. Members exhibit an upward trend in ride counts from Sunday to Tuesday, reaching a peak on Thursday at 560.884, followed by a significant drop leading up to Sunday. On the contrary, casual riders show increased bike usage during the weekends, particularly on Saturdays, where the number of rides reaches 467,926.



- Further analysis reveals that the ride counts, when examined by month, align with the previously mentioned seasonal patterns. Both customer types exhibit a higher frequency of rides from spring to autumn, with a notable decline during winter. Moreover, a negative correlation exists between the average ride length and the

month of initiation. As the month of initiation increases, the average ride length decreases. This effect is particularly pronounced among members, who experience a more significant decline in average ride length.



- Customers and members exhibit contrasting patterns in terms of their preferred start stations. Among casual riders, the majority tend to start their rides at Streeter Dr & Grand Ave, while for members, the most popular start station is Kingsbury St & Kinzie St. Interestingly, this start station is the least popular among casual riders.

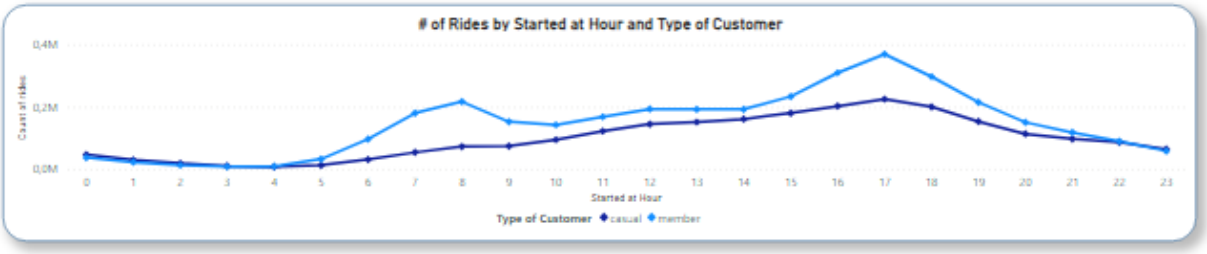
Top 10 Start Stations per Type of Customer*

*Unknown Station Names are excluded

Start Station Name	casual	member	Total
Streeter Dr & Grand Ave	57.436	17.450	**74.886**
DuSable Lake Shore Dr & Monroe St	32.039	9.357	**41.396**
Michigan Ave & Oak St	25.381	14.968	**40.349**
DuSable Lake Shore Dr & North Blvd	23.707	16.534	**40.241**
Wells St & Concord Ln	16.484	22.082	**38.566**
Clark St & Elm St	13.086	23.329	**36.415**
Millennium Park	25.287	9.719	**35.006**
Kingsbury St & Kinzie St	9.228	25.427	**34.655**
Theater on the Lake	18.485	14.883	**33.368**
Wells St & Elm St	12.459	19.884	**32.343**
Total	**233.592**	**173.633**	**407.225**

- Both casual riders and members exhibit a similar trend in terms of the starting hour of their rides, albeit with significant differences in numbers. The line graph illustrates a general upward trend from 5:00 to 17:00. However, member riders experience intermittent drops between 9:00 and 14:00, followed by a notable increase that culminates in a peak at 17:00.

In contrast, the data for casual riders shows a consistent upward trend, also peaking at 17:00. Importantly, there are no sudden spikes or significant declines observed, indicating a smooth and consistent progression over time.



Phase 6: Act

This phase involves delivering data-driven recommendations to stakeholders in order to enhance their operational efficiency and enable informed decision-making.

6.1 Unveiling the Data's Story

a. The total number of customers is 5.859.061, among which members count to 3.500.754 and casuals to 2.358.307.
b. Both types of customers highly prefer the time period between spring and autumn, while winter is the least favored. This indicates that weather significantly influences customer preferences, favoring warmer seasons.
c. The majority of total customers ride for a maximum of 30 minutes. However, casual riders are more likely to ride for longer durations than 30 minutes. This pattern leads us to conclude that members primarily use the bikes for commuting to work, while casual riders use them for leisure activities.
d. Members predominantly utilize Cyclistic bike services on weekdays, whereas casual riders tend to use the services more frequently on weekends. This observation reinforces the aforementioned conclusion that members use bikes for commuting to work, while casual riders use them for leisure and recreational purposes. Important to note that given that the data is anonymized, we lack information regarding customer addresses.
e. There is a negative correlation between the average ride length and the month of initiation. As the month of initiation progresses and becomes warmer, the average ride length tends to decrease. This effect is particularly noticeable among members, who experience a more significant decline in average ride length. The discomfort of riding in hot weather might lead people to prefer shorter bike rides. Members, who are more likely to use the bike-sharing service for daily commuting, might choose shorter rides during hot weather to avoid arriving at their destinations sweaty or fatigued. And casual riders, who may be tourists or occasional visitors, might opt for shorter rides to explore local attractions without excessive physical exertion.
f. Members and casual riders exhibit distinct preferences for different starting stations. This highlights the diversity in how members and casual riders utilize the bike-sharing service. Members seem to have regular commuting routes or specific destinations located closer to their workplaces. On the other hand, casual riders might use the

service for leisure or exploration, leading them to choose starting stations which are in proximity to popular attractions.

g. Members are more likely to ride between 05:00-09:00 and again between 14:00-17:00. In contrast, casual riders exhibit a consistent upward trend, with rides starting as early as 5:00 and peaking at 17:00. The early morning and afternoon peaks among members align with typical work commute hours. On the other hand, casual riders' increasing trend throughout the day suggests they might use the service for leisure activities or sightseeing. This observation supports the idea that casual riders might be tourists arriving early in the morning and extending their activities throughout the day.

6.2 Conclusions

To appeal to a broader range of casual customers, my recommendations are as follows:

a. Creating impactful marketing campaigns that highlight the advantages of subscribing.
b. Strategically placing advertisements at the most frequently used start stations to maximize visibility among casual customers.
c. Engage with casual customers on social media platforms to create a sense of community and showcase the advantages of being a subscriber.
d. Offering special benefits, such as discounted subscription rates, unlimited rides, free access to parks.
e. Introducing flexible subscription plans such as short-term and long-term subscriptions.
f. Collaborating with local businesses related to famous tourists' spots offering joint promotions.
g. Analyzing further the casual customers' segmentation and understanding further their needs, habits, and demographics.