Cyclistic Bike-Share Analysis.

Hello, I would like to thank you for reading about this project, which belongs to the Google Data Analytics Professional Certificate. I've been working on it for a few months, and this is the culmination of what I learned during these months through the seven-course path taught by Google Careers on the Coursera online learning platform.

A Divvy bike station at Addison Street and Drake Avenue. Credit: Photo by Max Herman - The Chicago Reporter -

SQL CODE (GITHUB) **PROJECT PRESENTATION**

Scenario

I am a junior data analyst working for a fictional bike-share company called "Cyclistic" in the marketing analyst team. The company was launched in 2016 and, over the years, has grown and now has a fleet of 5824 bikes and 693 docking stations across Chicago. The bikes are geotracked and locked into docking network stations, and the users can unlock them from one station and return them to any other network station.

1. Ask Phase: Diving Deep into the Company's Goal, Identifying the Stakeholders, and Business Tasks.

Stakeholders

* **Cyclistic** is a company that is developing a bike-share program in Chicago.

* **Lily Moreno** is the marketing analytics team manager and the person who leads the campaigns and initiatives to promote the bike-share program.

* The **marketing analytics department** is a team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide the company's marketing strategy.

* The **executive team** is detail-oriented and will decide whether to approve the recommended marketing program.

Marketing Strategy and Company´s Goal.

From the beginning, the marketing strategy focused on reaching all kinds of customers by implementing a flexible pricing plan: single-ride passes, full-day passes, and annual memberships. Customers who purchase single-ride or full-day passes are mentioned as "casual riders", while those who purchase annual memberships are called "members".

The company's finance department considers that members are more profitable than casual riders for the business; thus, Lily Moreno thinks that marketing campaigns should target converting casual riders into members rather than making new customers.

In order to do that, Moreno and her analyst team are interested in identifying trends and patterns in the historical trip data about how casual riders and members differ.

This analysis will help executives to make data-driven decisions about marketing programs and strategies to convert casual riders into riders with annual memberships.

My Task

Lily Moreno has tasked me with answering the following question: How do annual members and casual riders use Cyclistic bikes differently? You may have noticed that this is a SMART question because it is Specific, Measurable, Action-oriented, Relevant, and Time-bound.

I have to produce a report with the following deliverables:

* A clear statement of the business task

* A description of all data sources used

* Documentation of any cleaning or manipulation of data

* A summary of my analysis

* Supporting visualizations and key findings

* My top three recommendations based on the analysis.

2. Prepare Phase: Decide what data I have to collect in order to answer my question and how to organize it to be useful.

The last 12 months' historical trip data from Cyclistic was used to analyze and identify trends by working with a public dataset that has been made available by Motivate International Inc. (under this license) and was downloaded from this link.

Now I have to know if the data is confident and of good quality to support the analysis. We could use the ROCCC process to determine that our data is Reliable, Original, Comprehensive, Current, and Cited. As the data meets the aforementioned criteria, we can affirm that it is quality data.

After downloading the dataset (complete for the year 2022), the data was stored on my computer in a main folder called "Dataset_Cyclistic". The CSV (Comma, Separate, Values) files were renamed as follows into a subfolder called "Original_Files":

Cyclistic_2022_01 for January, Cyclistic_2022_02 for February, Cyclistic_2022_03 for March, and so on until December.

A backup of the data was stored on an external hard drive as well.

The original CSV files have 13 columns per month: ride_id, rideable_type, started_at, ended_at, start_station_name, start_station_id, end_station_name, end_station_id, start_lat, start_lng, end_lat end_lng, member_casual.

After considering some options, I decided to use MySQL to prepare and clean the data due to I feel confident working with it. Because all of the csv files contain the same data and columns, I thought that the best option was to start by creating a new SQL database by formatting the columns and data types. Once the main database is created, I am ready to merge all the CSV files into my main SQL database. I may be able to show you now a summary of my SQL working process; however, you can check out the complete version in my GITHUB repository.

```
# STEP 1: Create Database "Main_Bike_Share"

CREATE DATABASE IF NOT EXISTS Main_Bike_Share;

# STEP 2: Create 2022_original Table (to be loaded into the database as CSV
files)

CREATE TABLE IF NOT EXISTS 2022_original (
        ride_id TEXT,
          rideable_type TEXT,
          started_at TIMESTAMP,
          ended_at   TIMESTAMP,
          start_station_name TEXT,
          start_station_id TEXT,
          end_station_name TEXT,
          end_station_id TEXT,
          start_lat DECIMAL(10,8),
          start_lng DECIMAL(11,8),
          end_lat DECIMAL(10,8),
          end_lng DECIMAL(11,8),
          member_casual TEXT
          );

# STEP 3: Load the CSV files into the table previously created (loaded month
by month into a unique table called 2022_original, previously created).

LOAD DATA INFILE 'C:\\ProgramData\\MySQL\\MySQL
Server8.0\\Uploads\\Cyclistic_2022_01.csv'
INTO TABLE 2022_original
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 ROWS;
```

* This SQL code was repeated for all.csv files from January to December 2022.

#STEP 4: Check if the table has been loaded correctly and how many fields are in it.

```
SELECT * FROM 2022_original;
        # After the Query, I know that there are 5.547.140 rows in the original Dataset.
```

#STEP 5: To keep track of the PREPARE PHASE, I saved the 2022_original table as a CSV file.

```
SELECT *
INTO
OUTFILE 'C:\\ProgramData\\MySQL\\MySQLServer8.0\\Uploads\\2022_original.csv'
FIELDS TERMINATED BY ','
OPTIONALLY ENCLOSED BY '"'
ESCAPED BY '\\'
LINES TERMINATED BY '\n'
FROM 2022_original;
```

After downloading the 2022_original table from MySQL, I moved it to a Dataset_Cyclistic subfolder called "Prepare_Phase_Tracking".

Now I am ready to start with the Process Phase.

3. Process Phase: Now I am going to clean and transform the data.

Now I have to identify possible errors, inaccuracies, or inconsistencies and clean them up in order to get quality data to be analyzed. Some calculated fields should be created as well in order to perform the dataset, and I may remove other unnecessary fields.

#STEP 6: Check if there are duplicate fields.

```
SELECT DISTINCT (ride_id) FROM 2022_original;
        # After running the query, I know that there are 4.777.936 rows, which means that there are duplicated fields.
# STEP 7: Now I am going to check if there are null or empty fields in all the
columns.

SELECT *
FROM 2022_original
WHERE (ride_id IS NULL OR TRIM (ride_id)='')
OR (rideable_type IS NULL OR TRIM (rideable_type)='')
OR (started_at IS NULL OR TRIM (started_at)='')
OR (ended_at IS NULL OR TRIM (ended_at)='')
OR (start_station_name IS NULL OR TRIM (start_station_name)='')
OR (start_station_id IS NULL OR TRIM (start_station_id)='')
OR (end_station_name IS NULL OR TRIM (end_station_name)='')
OR (end_station_id IS NULL OR TRIM (end_station_id)='')
OR (start_lat IS NULL OR TRIM (start_lat)='')
OR (start_lng IS NULL OR TRIM (start_lng)='')
OR (end_lat IS NULL OR TRIM (end_lat)='')
OR (end_lng IS NULL OR TRIM (end_lng)='')
OR (member_casual IS NULL OR TRIM (member_casual)='');
            # There are 1.227.638 null or empty fields in the 2022_original
Table.

# STEP 8: Due to I have a backup of the table (refer step 5) I am going to
remove Null and Empty fields from 2022_original table.

DELETE FROM 2022_original
WHERE (ride_id IS NULL OR TRIM (ride_id)='')
OR (rideable_type IS NULL OR TRIM (rideable_type)='')
OR (started_at IS NULL OR TRIM (started_at)='')
OR (ended_at IS NULL OR TRIM (ended_at)='')
OR (start_station_name IS NULL OR TRIM (start_station_name)='')
OR (start_station_id IS NULL OR TRIM (start_station_id)='')
OR (end_station_name IS NULL OR TRIM (end_station_name)='')
OR (end_station_id IS NULL OR TRIM (end_station_id)='')
OR (start_lat IS NULL OR TRIM (start_lat)='')
OR (start_lng IS NULL OR TRIM (start_lng)='')
OR (end_lat IS NULL OR TRIM (end_lat)='')
OR (end_lng IS NULL OR TRIM (end_lng)='')
OR (member_casual IS NULL OR TRIM (member_casual)='');
         # 1.227.638 fields are been removed from 2022_original table.
```

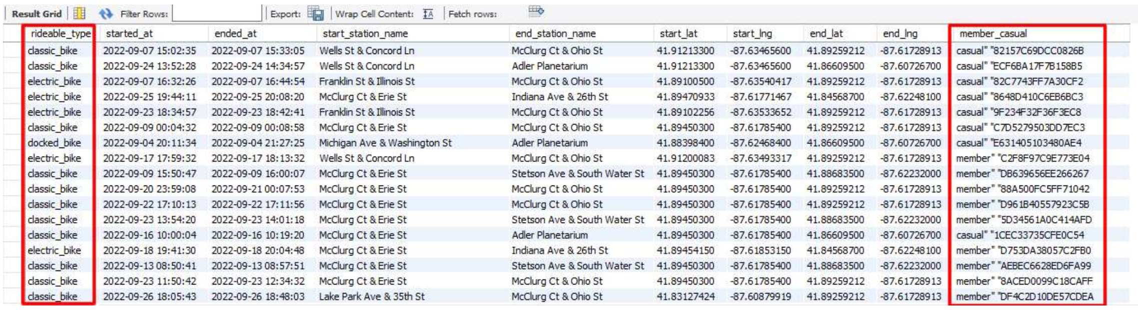
```
#STEP 9: I have to remove duplicate values (refer to step 6). My plan is to
make a copy of the original table, insert a query into it by using "DISTINCT"
clause, delete the original table, and rename the 2022_copy table as
2022_original.

CREATE TABLE 2022_copy LIKE 2022_original;

INSERT INTO 2022_copy
SELECT DISTINCT (ride_id),rideable_type,started_at,ended_at,start_station_name,start_station_id,end_station_name,end_station_id,start_lat,start_lng,end_lat,end_lng,member_casual
FROM 2022_original;
        # Query has returned 3.699.152 fields.

DROP TABLE 2022_original;

ALTER TABLE 2022_copy RENAME TO 2022_original;

#STEP 10: I am going to continue with the cleaning process. Since I realized
that I don´t need some columns for my analysis, those columns should be
removed.

ALTER TABLE 2022_original DROP COLUMN ride_id;
ALTER TABLE 2022_original DROP COLUMN start_station_id;
ALTER TABLE 2022_original DROP COLUMN end_station_id;
        # Three columns are been deleted from table 2022_original


# STEP 11: I am going to remove leading and trailing spaces from a string
with the TRIM function. I will repeat the idea I worked with in step 9.


CREATE TABLE 2022_copy LIKE 2022_original;

INSERT INTO 2022_copy
SELECT TRIM (rideable_type),
TRIM (started_at),
TRIM (ended_at),
TRIM (start_station_name),
TRIM (end_station_name),
TRIM (start_lat),
TRIM (start_lng),
TRIM (end_lat),
TRIM (end_lng),
TRIM (member_casual)
```

```
FROM 2022_original;
```

```
DROP TABLE 2022_original;
```

```
ALTER TABLE 2022_copy RENAME TO 2022_original;
```

I have to continue the cleaning process because I realized that the customer-member column has to be corrected. Noticed that according to the image below, the names are inconsistent: it should be "Member" or "Casual." Because putting "bike" after the bike type sounds redundant, the rideable_type column could be text formatted according to best practices. For that reason, I will remove "-bike" and capitalize the first word in that column as well.



rideable_type	started_at	ended_at	start_station_name	end_station_name	start_lat	start_lng	end_lat	end_lng	member_casual
classic_bike	2022-09-07 15:02:35	2022-09-07 15:33:05	Wells St & Concord Ln	McClurg Ct & Ohio St	41.91213300	-87.63465600	41.89259212	-87.61728913	casual" "82157C69DCC0826B
classic_bike	2022-09-24 13:52:28	2022-09-24 14:34:57	Wells St & Concord Ln	Adler Planetarium	41.91213300	-87.63465600	41.86609500	-87.60726700	casual" "ECF6BA17F7B158B5
electric_bike	2022-09-07 16:32:26	2022-09-07 16:44:54	Franklin St & Illinois St	McClurg Ct & Ohio St	41.89100500	-87.63540417	41.89259212	-87.61728913	casual" "82C7743FF7A30CF2
electric_bike	2022-09-25 19:44:11	2022-09-25 20:08:20	McClurg Ct & Erie St	Indiana Ave & 26th St	41.89470933	-87.61771467	41.84568700	-87.62248100	casual" "8648D410C6EB6BC3
electric_bike	2022-09-23 18:34:57	2022-09-23 18:42:41	Franklin St & Illinois St	McClurg Ct & Ohio St	41.89102256	-87.63533652	41.89259212	-87.61728913	casual" "9F234F32F36F3EC8
classic_bike	2022-09-09 00:04:32	2022-09-09 00:08:58	McClurg Ct & Erie St	McClurg Ct & Ohio St	41.89450300	-87.61785400	41.89259212	-87.61728913	casual" "C7D5279503DD7EC3
docked_bike	2022-09-04 20:11:34	2022-09-04 21:27:25	Michigan Ave & Washington St	Adler Planetarium	41.88398400	-87.62468400	41.86609500	-87.60726700	casual" "E631405103480AE4
electric_bike	2022-09-17 17:59:32	2022-09-17 18:13:32	Wells St & Concord Ln	McClurg Ct & Ohio St	41.91200083	-87.63493317	41.89259212	-87.61728913	member" "C2F8F97C9E773E04
classic_bike	2022-09-09 15:50:47	2022-09-09 16:00:07	McClurg Ct & Erie St	Stetson Ave & South Water St	41.89450300	-87.61785400	41.88683500	-87.62232000	member" "DB639656EE266267
classic_bike	2022-09-20 23:59:08	2022-09-21 00:07:53	McClurg Ct & Erie St	McClurg Ct & Ohio St	41.89450300	-87.61785400	41.89259212	-87.61728913	member" "88A500FC5FF71042
classic_bike	2022-09-22 17:10:13	2022-09-22 17:11:56	McClurg Ct & Erie St	McClurg Ct & Ohio St	41.89450300	-87.61785400	41.89259212	-87.61728913	member" "D961B40557923C5B
classic_bike	2022-09-23 13:54:20	2022-09-23 14:01:18	McClurg Ct & Erie St	Stetson Ave & South Water St	41.89450300	-87.61785400	41.88683500	-87.62232000	member" "5D34561A0C414AFD
classic_bike	2022-09-16 10:00:04	2022-09-16 10:19:20	McClurg Ct & Erie St	Adler Planetarium	41.89450300	-87.61785400	41.86609500	-87.60726700	casual" "1CEC33735CFE0C54
electric_bike	2022-09-18 19:41:30	2022-09-18 20:04:48	McClurg Ct & Erie St	Indiana Ave & 26th St	41.89454150	-87.61853150	41.84568700	-87.62248100	member" "D753DA38057C2FB0
classic_bike	2022-09-13 08:50:41	2022-09-13 08:57:51	McClurg Ct & Erie St	Stetson Ave & South Water St	41.89450300	-87.61785400	41.88683500	-87.62232000	member" "AEBEC6628ED6FA99
classic_bike	2022-09-23 11:50:42	2022-09-23 12:34:32	McClurg Ct & Erie St	McClurg Ct & Ohio St	41.89450300	-87.61785400	41.89259212	-87.61728913	member" "8ACED0099C18CAFF
classic_bike	2022-09-26 18:05:43	2022-09-26 18:48:03	Lake Park Ave & 35th St	McClurg Ct & Ohio St	41.83127424	-87.60879919	41.89259212	-87.61728913	member" "DF4C2D10DE57CDEA

```
#STEP 12: Transforming the values of the rideable_type column.
```

```
UPDATE 2022_original SET rideable_type='Electric' WHERE rideable_type
='electric_bike';
```

```
UPDATE 2022_original SET rideable_type='Classic' WHERE rideable_type=
'classic_bike';
```

```
UPDATE 2022_original SET rideable_type='Docked' WHERE rideable_type
='docked_bike';
```

```
#STEP 13: The first word in the casual-member column should be capitalized,
and inconsistent values should be removed. Repeat steps 9 and 11 procedures.
```

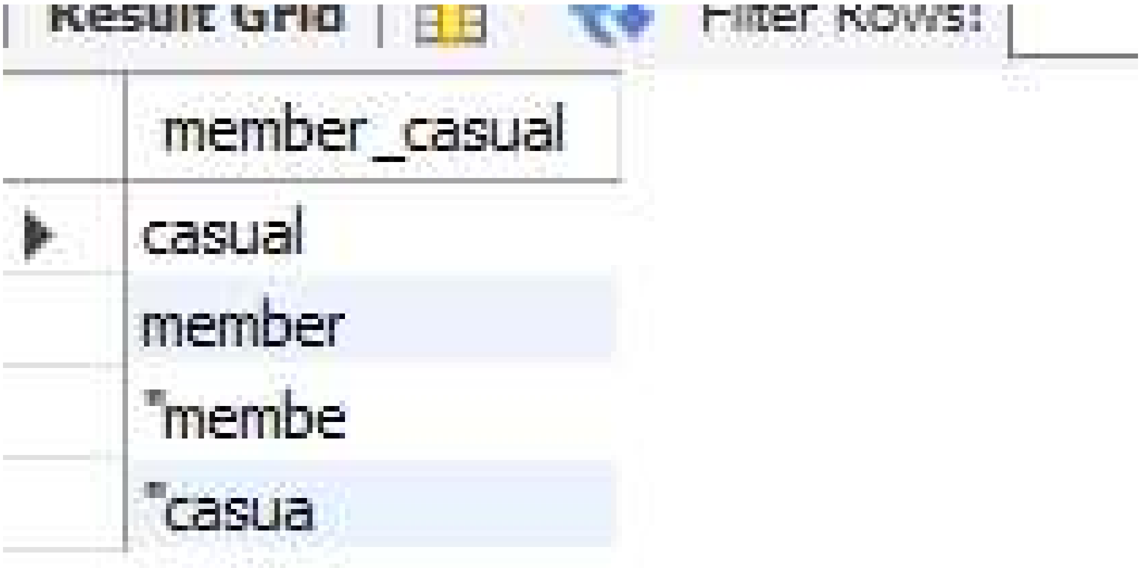
```
CREATE TABLE 2022_copy LIKE 2022_original;
```

Because "member" and "casual" have the same number of letters, I decided to use the LEFT function to extract the first 6 letters from the string.

```
INSERT INTO 2022_copy
SELECT rideable_type,
started_at,ended_at,start_station_name,end_station_name,start_lat,start_lng,end_lat,end_lng,LEFT(member_casual,6) AS member_casual
FROM 2022_original;
```

```
SELECT DISTINCT(member_casual) FROM
2022_copy;
```

After running the query above to check for distinct values in the member_casual column, I saw that there were still different values due to the fact that some words have " before the first letter.



member_casual
casual
member
"membe
"casua

```
UPDATE 2022_copy SET member_casual='Casual' WHERE member_casual
='"casua';
```

```
UPDATE 2022_copy SET member_casual='Member' WHERE member_casual='"membe';
```

Finally, I can drop the old table and rename 2022_copy as 2022_original.

```
DROP TABLE 2022_original;
```

```
ALTER TABLE 2022_copy RENAME TO 2022_original;
```

#STEP 14 To keep track of the PROCESS PHASE, I saved the 2022_original table as a CSV file.

```
SELECT *
INTO OUTFILE 'C:\\ProgramData\\MySQL\\MySQL Server
8.0\\Uploads\\2022_clean.csv'
FIELDS TERMINATED BY ','
OPTIONALLY ENCLOSED BY '"'
ESCAPED BY '\\'
LINES TERMINATED BY '\n'
FROM 2022_clean;
```

After downloading the 2022_clean table from MySQL, I moved it to a Dataset_Cyclistic subfolder called "Clean_Phase_Tracking".

Now that my dataset is clean, the next step is to create calculated fields in order to prepare for data analysis in the next phase. I am going to create some fields such as:

* Ride_Length by using the TIMEDIFF function

* Day_of_Week by using the DAYNAME function

* Part_of_Day by using the CASE statement as follow:

    - Morning (6 TO 11.59 AM)
    - Afternoon (12 to 6.59 PM)
    - Evening (7 to 10 PM)
    - Else Night.

* Month by using the MONTH function.

* Season by using the CASE statement as follow:

    - Winter (December, January, February)
    - Spring (March, April, May)
    - Summer (June, July, August)
    - Else Autumn.

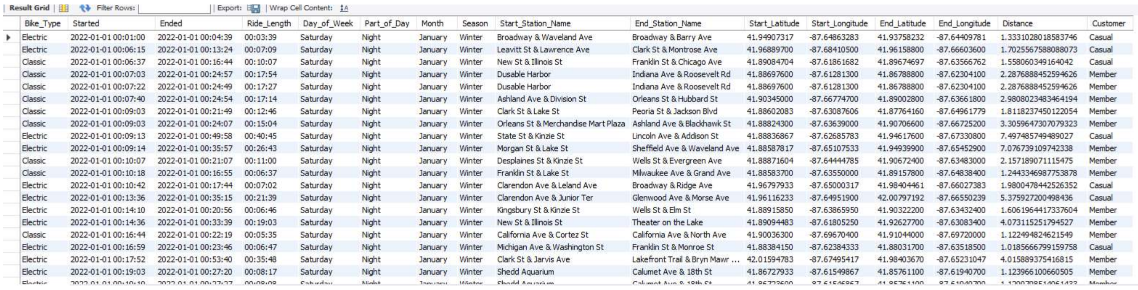
* Distance: A formula with latitude and longitude fields was used to calculate distance using the earth's radius (6378) as a reference in kilometers.

All these fields were created in a temporary table that was later exported as a CSV file and filtered in order to remove the outliers.

#STEP15: Creating a temporary table and removing outliers filtered by ride_length and distance Criteria: The ride must last at least more than 3 minutes and no more than 2 hours (7200 seconds), and the distance must be at least 1 kilometer and no more than 50 kilometers.

```
CREATE TEMPORARY TABLE 2022_Final SELECT rideable_type AS
Bike_Type, started_at AS Started, ended_at AS
Ended,TIMEDIFF (ended_at,started_at) AS
Ride_Length, DAYNAME (started_at) AS Day_of_Week,
CASE
 WHEN TIME (started_at) BETWEEN '19:00:00' AND
'22:00:00' THEN "Evening"
 WHEN TIME (started_at) BETWEEN '06:00:00' AND
'11:59:59' THEN "Morning"
 WHEN TIME (started_at) BETWEEN '12:00:00' AND
'18:59:59' THEN "Afternoon"
 ELSE "Night"
 END AS Part_of_Day, MONTHNAME (started_at) AS Month,
 CASE
 WHEN started_at BETWEEN '2022-12-01' AND '2022-12-
31' THEN "Winter"
 WHEN started_at BETWEEN '2022-01-01' AND '2022-01-
31' THEN "Winter"
 WHEN started_at BETWEEN '2022-02-01' AND '2022-02-
28' THEN "Winter"
 WHEN started_at BETWEEN '2022-03-01' AND '2022-03-
31' THEN 'Spring'
 WHEN started_at BETWEEN '2022-04-01' AND '2022-04-
30' THEN 'Spring'
 WHEN started_at BETWEEN '2022-05-01' AND '2022-05-
31' THEN 'Spring'
 WHEN started_at BETWEEN '2022-06-01' AND '2022-06-
30' THEN 'Summer'
 WHEN started_at BETWEEN '2022-07-01' AND '2022-07-
31' THEN 'Summer'
 WHEN started_at BETWEEN '2022-08-01' AND '2022-08-
31' THEN 'Summer'
 ELSE "Autumn"
END AS Season, start_station_name AS Start_Station_Name,
end_station_name AS   End_Station_Name, start_lat AS Start_Latitude,
start_lng AS Start_Longitude, end_lat AS   End_Latitude,end_lng AS
End_Longitude,
(ACOS(SIN(RADIANS(start_lat)) * SIN(RADIANS(end_lat)) +
COS(RADIANS(start_lat)) * COS(RADIANS(end_lat)) *
COS(RADIANS(start_lng) - RADIANS(end_lng))) * 6378) AS
Distance, CONCAT(UPPER(LEFT(member_casual,1)), LOWER(SUBSTR(
member_casual,2))) AS Customer
```

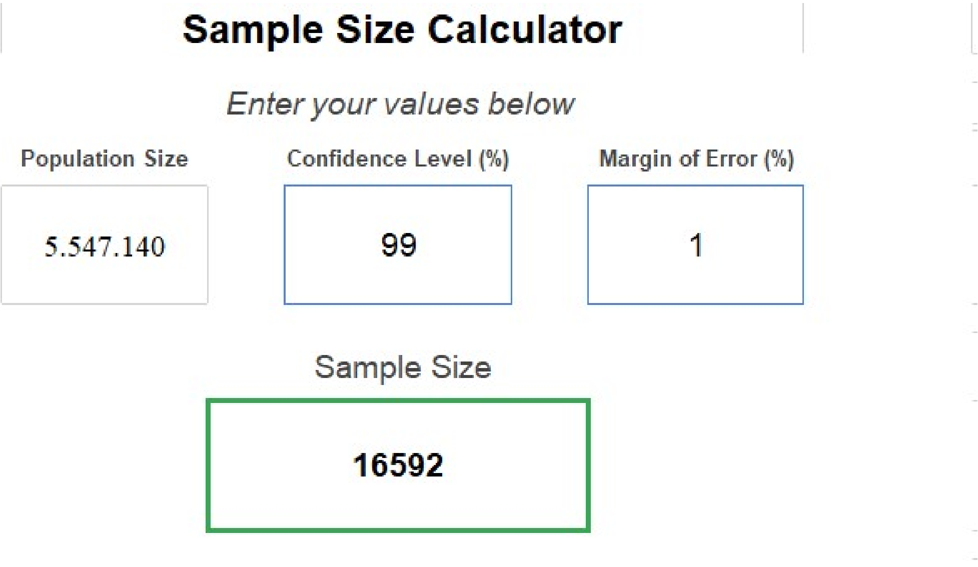
```
FROM 2022_clean
WHERE TIMEDIFF(ended_at,started_at) BETWEEN 300
AND 7200 AND (ACOS(SIN(RADIANS(start_lat))
* SIN(RADIANS(end_lat)) +
COS(RADIANS(start_lat)) * COS(RADIANS(end_lat)) *
COS(RADIANS(start_lng) - RADIANS(end_lng))) *
6378) BETWEEN 1 AND 50
ORDER BY Started;
```



Bike_Type	Started	Ended	Ride_Length	Day_of_Week	Part_of_Day	Month	Season	Start_Station_Name	End_Station_Name	Start_Latitude	Start_Longitude	End_Latitude	End_Longitude	Distance	Customer
Electric	2022-01-01 00:01:00	2022-01-01 00:04:39	00:03:39	Saturday	Night	January	Winter	Broadway & Waveland Ave	Broadway & Barry Ave	41.94907317	-87.64863283	41.93758232	-87.64409781	1.3331028018583746	Casual
Electric	2022-01-01 00:06:15	2022-01-01 00:13:24	00:07:09	Saturday	Night	January	Winter	Leavitt St & Lawrence Ave	Clark St & Montrose Ave	41.96889700	-87.68410500	41.96158800	-87.66603600	1.7025567588088073	Casual
Classic	2022-01-01 00:06:37	2022-01-01 00:16:44	00:10:07	Saturday	Night	January	Winter	New St & Illinois St	Franklin St & Chicago Ave	41.89084704	-87.61861682	41.89674697	-87.63566762	1.558060349164042	Casual
Classic	2022-01-01 00:07:03	2022-01-01 00:24:57	00:17:54	Saturday	Night	January	Winter	Dusable Harbor	Indiana Ave & Roosevelt Rd	41.88697600	-87.61281300	41.86788800	-87.62304100	2.2876888452594626	Member
Classic	2022-01-01 00:07:22	2022-01-01 00:24:49	00:17:27	Saturday	Night	January	Winter	Dusable Harbor	Indiana Ave & Roosevelt Rd	41.88697600	-87.61281300	41.86788800	-87.62304100	2.2876888452594626	Member
Classic	2022-01-01 00:07:40	2022-01-01 00:24:54	00:17:14	Saturday	Night	January	Winter	Ashland Ave & Division St	Orleans St & Hubbard St	41.90345000	-87.66774700	41.89002800	-87.63661800	2.9808023483464194	Member
Classic	2022-01-01 00:09:03	2022-01-01 00:21:49	00:12:46	Saturday	Night	January	Winter	Clark St & Lake St	Peoria St & Jackson Blvd	41.88602083	-87.63087606	41.87764160	-87.64961779	1.8118237450122054	Member
Classic	2022-01-01 00:09:03	2022-01-01 00:24:07	00:15:04	Saturday	Night	January	Winter	Orleans St & Merchandise Mart Plaza	Ashland Ave & Blackhawk St	41.88824300	-87.63639000	41.90706600	-87.66725200	3.3059647307079323	Member
Electric	2022-01-01 00:09:13	2022-01-01 00:49:58	00:40:45	Saturday	Night	January	Winter	State St & Kinzie St	Lincoln Ave & Addison St	41.88836867	-87.62685783	41.94617600	-87.67330800	7.497485749489027	Casual
Electric	2022-01-01 00:09:14	2022-01-01 00:35:57	00:26:43	Saturday	Night	January	Winter	Morgan St & Lake St	Sheffield Ave & Waveland Ave	41.88587817	-87.65107533	41.94939900	-87.65452900	7.076739109742338	Member
Classic	2022-01-01 00:10:07	2022-01-01 00:21:07	00:11:00	Saturday	Night	January	Winter	Desplaines St & Kinzie St	Wells St & Evergreen Ave	41.88871604	-87.64444785	41.90672400	-87.63483000	2.157189071115475	Member
Classic	2022-01-01 00:10:18	2022-01-01 00:16:55	00:06:37	Saturday	Night	January	Winter	Franklin St & Lake St	Milwaukee Ave & Grand Ave	41.88583700	-87.63550000	41.89157800	-87.64838400	1.2443346987753878	Member
Electric	2022-01-01 00:10:42	2022-01-01 00:17:44	00:07:02	Saturday	Night	January	Winter	Clarendon Ave & Leland Ave	Broadway & Ridge Ave	41.96797933	-87.65000317	41.98404461	-87.66027383	1.9800478442526352	Casual
Electric	2022-01-01 00:13:36	2022-01-01 00:35:15	00:21:39	Saturday	Night	January	Winter	Clarendon Ave & Junior Ter	Glenwood Ave & Morse Ave	41.96116233	-87.64951900	42.00797192	-87.66550239	5.375927200498436	Casual
Electric	2022-01-01 00:14:10	2022-01-01 00:20:56	00:06:46	Saturday	Night	January	Winter	Kingsbury St & Kinzie St	Wells St & Elm St	41.88915850	-87.63865950	41.90322200	-87.63432400	1.6061964417337604	Member
Electric	2022-01-01 00:14:36	2022-01-01 00:33:39	00:19:03	Saturday	Night	January	Winter	New St & Illinois St	Theater on the Lake	41.89094483	-87.61805250	41.92627700	-87.63083400	4.073115251794527	Member
Classic	2022-01-01 00:16:44	2022-01-01 00:22:19	00:05:35	Saturday	Night	January	Winter	California Ave & Cortez St	California Ave & North Ave	41.90036300	-87.69670400	41.91044000	-87.69720000	1.122494824621549	Member
Electric	2022-01-01 00:16:59	2022-01-01 00:23:46	00:06:47	Saturday	Night	January	Winter	Michigan Ave & Washington St	Franklin St & Monroe St	41.88384150	-87.62384333	41.88031700	-87.63518500	1.0185666799159758	Casual
Electric	2022-01-01 00:17:52	2022-01-01 00:53:40	00:35:48	Saturday	Night	January	Winter	Clark St & Jarvis Ave	Lakefront Trail & Bryn Mawr ...	42.01594783	-87.67495417	41.98403670	-87.65231047	4.015889375416815	Member
Electric	2022-01-01 00:19:03	2022-01-01 00:27:20	00:08:17	Saturday	Night	January	Winter	Shedd Aquarium	Calumet Ave & 18th St	41.86727933	-87.61549867	41.85761100	-87.61940700	1.123966100660505	Member
Electric	2022-01-01 00:19:19	2022-01-01 00:27:27	00:08:08	Saturday	Night	January	Winter	Shedd Aquarium	Calumet Ave & 18th St	41.86723600	-87.61546867	41.85761100	-87.61940700	1.1200708514061433	Member

The new table has 16 columns and 2.551.272 rows of data. I filtered my table to be as accurate as possible. There were some values in the Ride_Length column with negative or zero values that needed to be removed; because I wanted to work with more realistic data, I decided to filter the field between 300 seconds (5 minutes) and 7200 seconds (2 hours) too. In the Distance column, there were some incorrect values as well, such as zero kilometers and even more than 50 kilometers. All those incorrect values were most likely the result of a human error when entering data into the file.

* Noticed that we now have **2.551.272 rows of data**, whereas the **original file had** more than 5 million rows (**5.547.140** rows). Is there enough data in the new table to support a reliable and high-quality analysis? I think there should be enough data, but in order to be sure, I want to check out my hypothesis by using a **sample calculator**.



Sample Size Calculator

Enter your values below

Population Size	Confidence Level (%)	Margin of Error (%)
5.547.140	99	1

Sample Size

16592

After checking to see if my final dataset contains enough data, I can see that, in fact, we only need 16592 rows to be a good sample with a great confidence level and only a 1% margin of error. However, I will use my complete dataset for the analysis.

After creating the 2022_Final temporary table, I ran a query that returned to me all the fields of the table and exported the query as a CSV file, which I moved to a Dataset_Cyclistic subfolder called "Final_Table."

So let´s start with the analysis!

4. Analyze Phase: Now I have to identify trends and relationships within the dataset in order to answer the business question. How do annual members and casual riders use Cyclistic bikes differently?

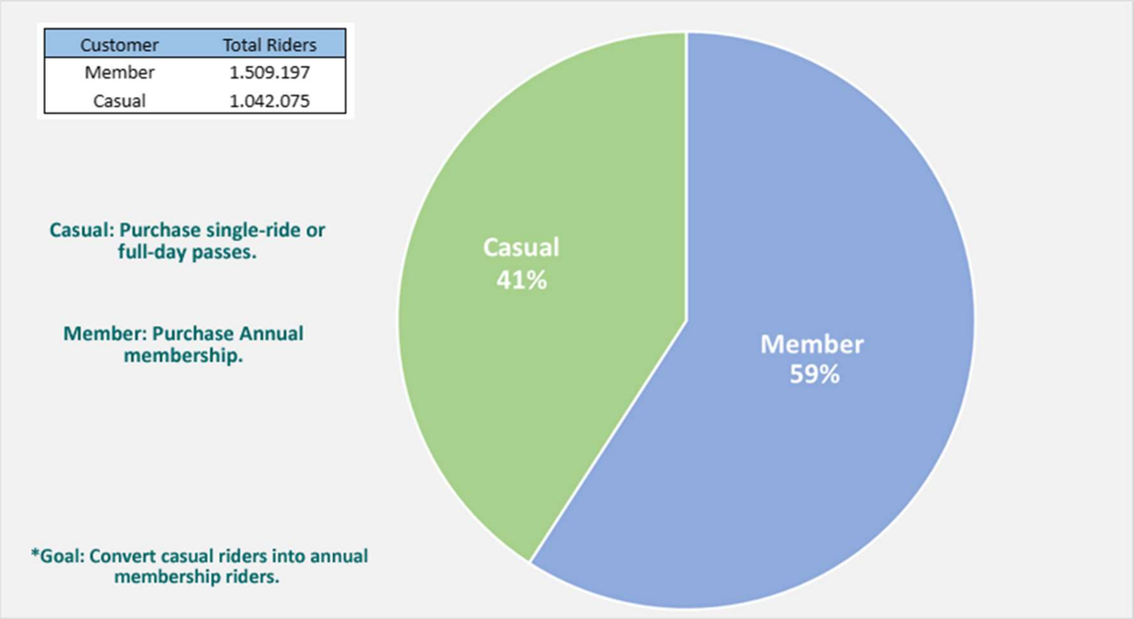
In order to analyze all the data, SQL queries were used and exported to a CSV file. CSV files were opened with Excel and generated graphics to reinforce my explanation.

The first question I'd like to address is the number of casual and member riders, as well as their percentage of the total.

```
SELECT COUNT (Customer)
FROM 2022_Final
WHERE Customer='Member';
            # There are 1.509.197 members
```

```
SELECT COUNT (Customer)
FROM 2022_Final
WHERE Customer='Casual';
                # There are 1.042.075 casual riders.
```

According to the information returned from both queries, we can see that there are more members than casual customers.



Customer	Total Riders
Member	1.509.197
Casual	1.042.075

Casual: Purchase single-ride or full-day passes.

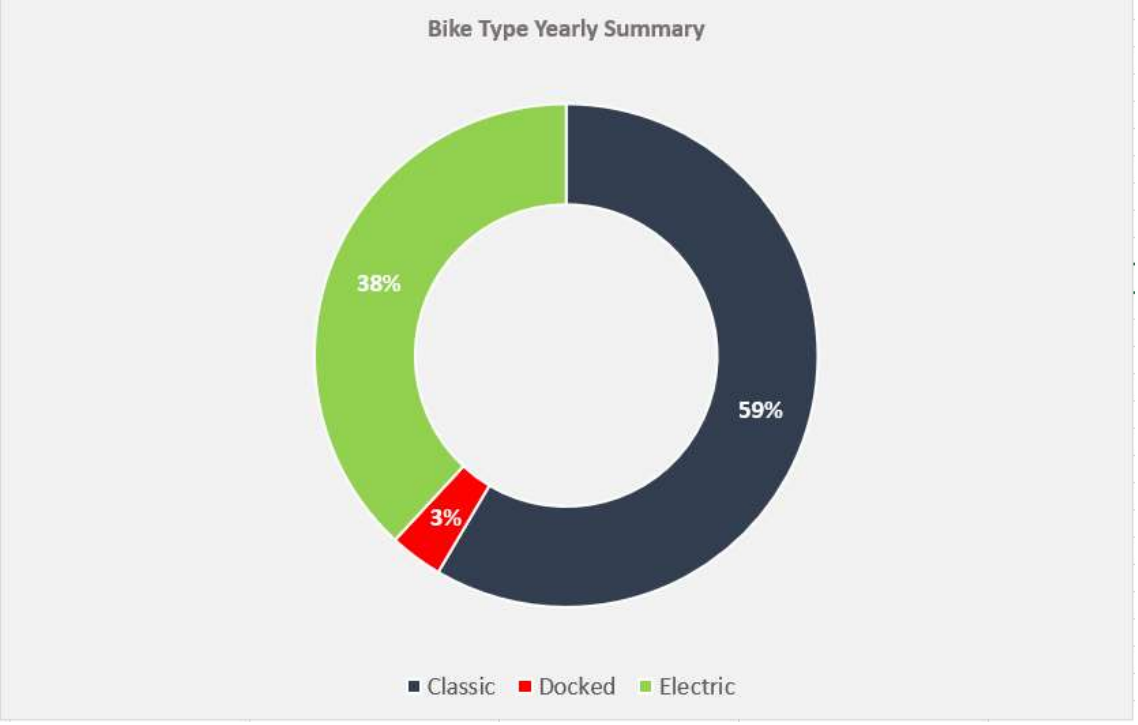
Member: Purchase Annual membership.

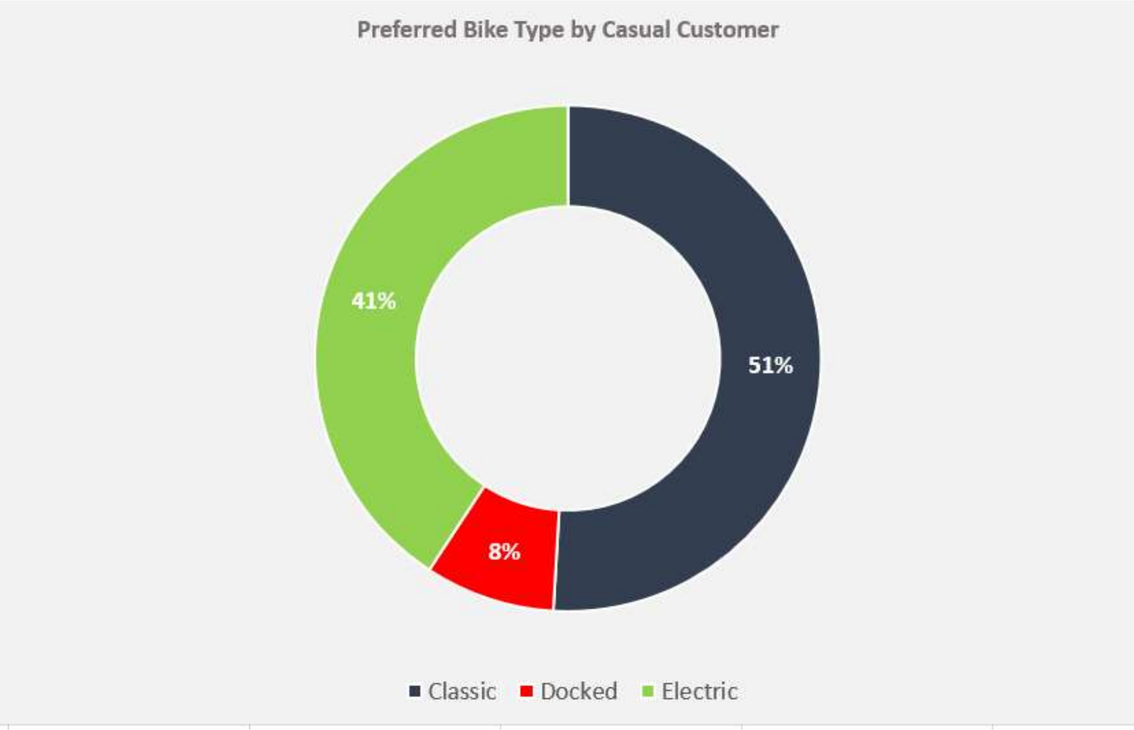
*Goal: Convert casual riders into annual membership riders.

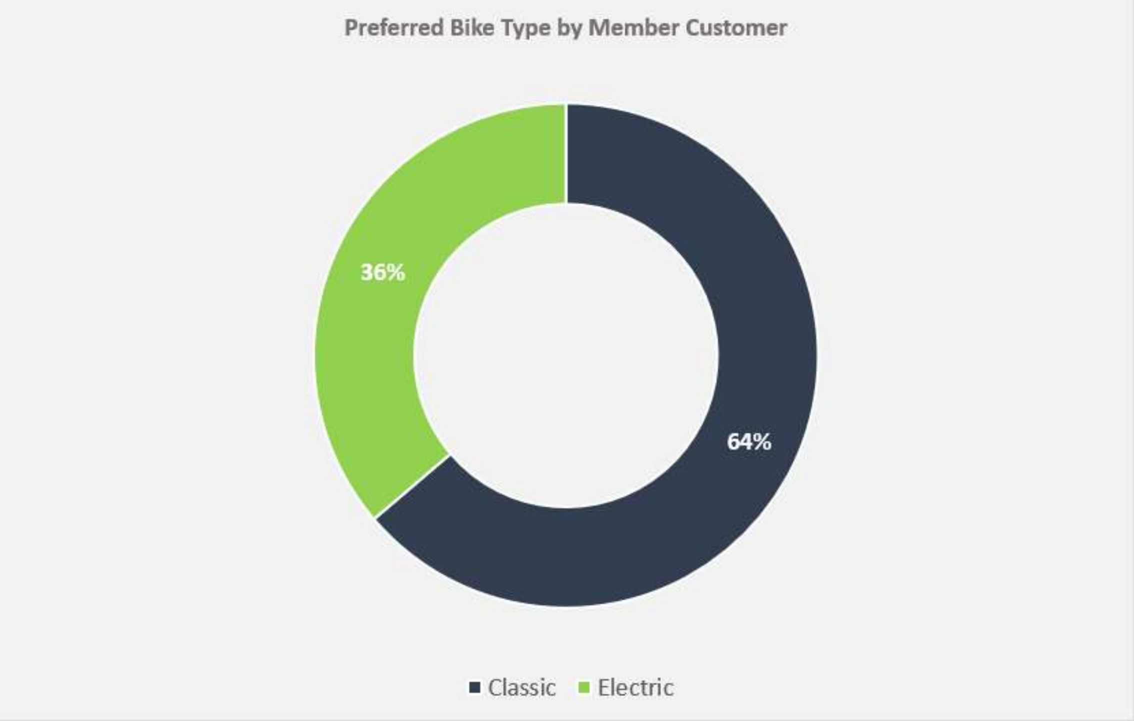
The second step of my analysis is going in the direction of figuring out what kind of bicycle is mostly preferred by the customers. I began by running a query that revealed the most popular bike used by all riders.

```
SELECT Customer,Bike_Type, Count(Bike_Type)
FROM 2022_Final
GROUP BY Bike_Type,Customer;
```

From the query above, I made a pivot table in Excel and dug deep into the information in order to get three different graphics, as you can see below:


Bike Type Yearly Summary
38%
59%
3%
■ Classic ■ Docked ■ Electric


Preferred Bike Type by Casual Customer
41%
51%
8%
■ Classic ■ Docked ■ Electric

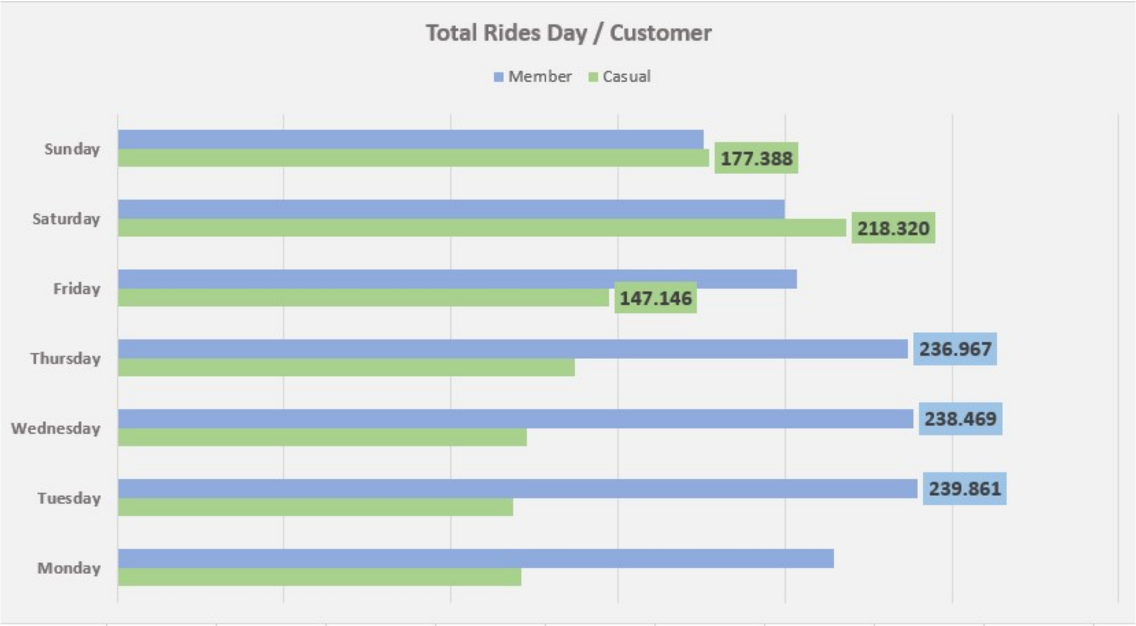


Clearly, we can see that **the preferred bike for both casual and member riders is the Classic** one. Also, both casual users and members use Electric bikes as well. However, member customers didn´t use Docked ones, and only some casual customers—a lower percentage—preferred this kind of bike.

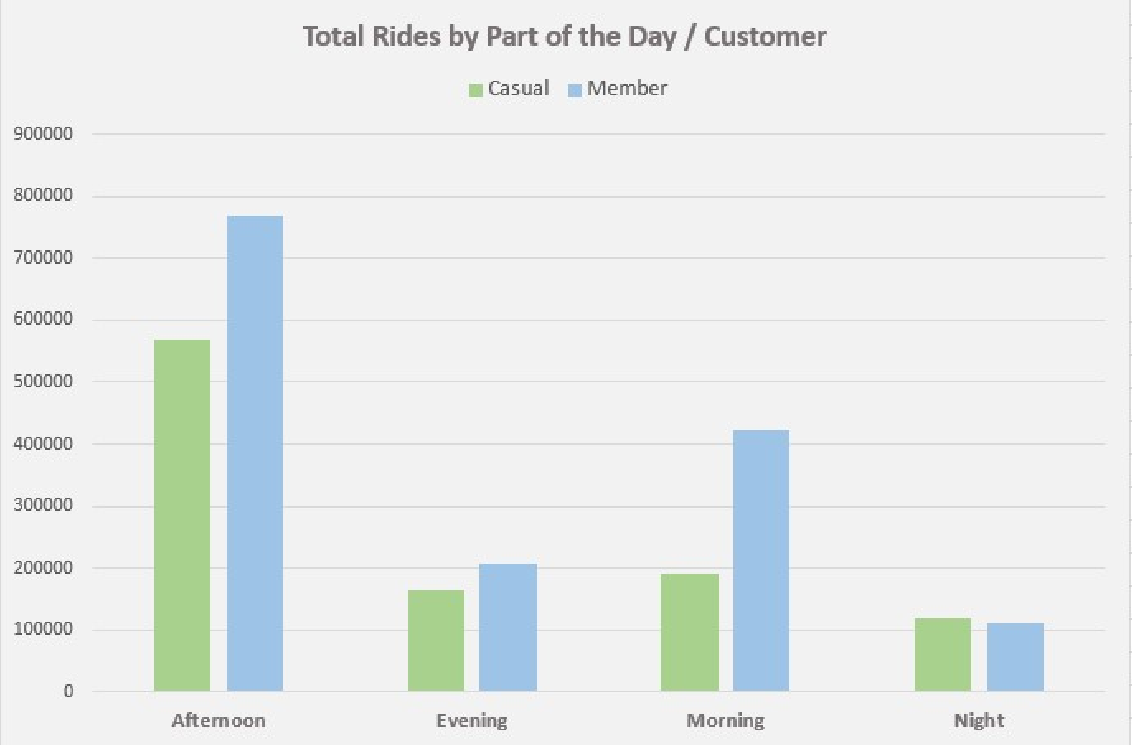
Let's start with t**otal rides broken down by day, month, and season** and make a comparison between the member and casual customer statistics.

```
SELECT Customer, Day_of_Week, COUNT(Started)
FROM 2022_Final
GROUP BY Day_of_Week, Customer;
        #Total Rides by Day of Week

SELECT Customer,Part_of_Day,COUNT(Started)
FROM 2022_Final
GROUP BY Part_of_Day,Customer;
        #Total Rides by Part of Day

SELECT Customer, Month, COUNT(Started)
FROM 2022_Final
GROUP BY Month, Customer;
         # Total Rides by Month

SELECT Customer, Season, COUNT(Started)
FROM 2022_Final
GROUP BY Season,Customer;
```

#Total Rides by Season / Customer



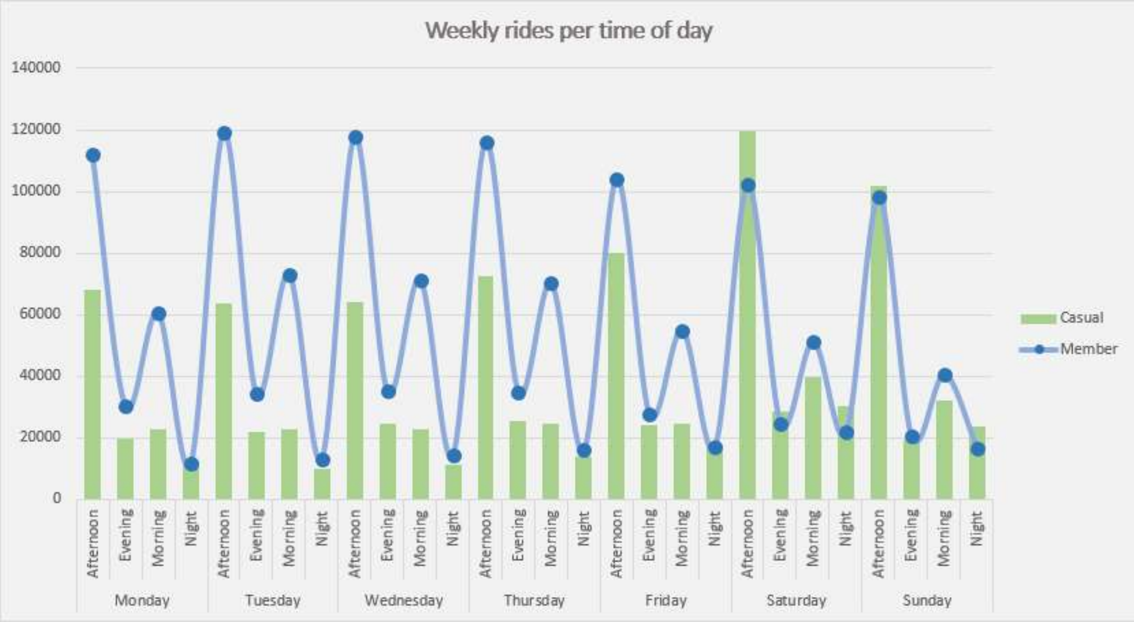
The first evidence of the differences between casual and member usage patterns is that while **casuals prefer weekends** to enjoy bicycle trips, **members usually ride in the middle of the week**. We can see according to the graphic that the top 3 most frequent days for casuals are Friday, Saturday, and Sunday. The members' most frequent rides are on Tuesday, Wednesday, and Thursday instead. Maybe mid-week ride preferences for members show us that they could be people who use the bike to go to work or as a locomotive medium. Casuals, on the other hand, could be people who ride for leisure or belong to cycling clubs or communities and enjoy or train on weekends.



We can see that for both casuals and members, the most usual time of the day for a ride is the afternoon, followed by the morning. The distinction between members and casual morning riders should be highlighted. It could lend credence to the theory that members could use the bike for transportation.

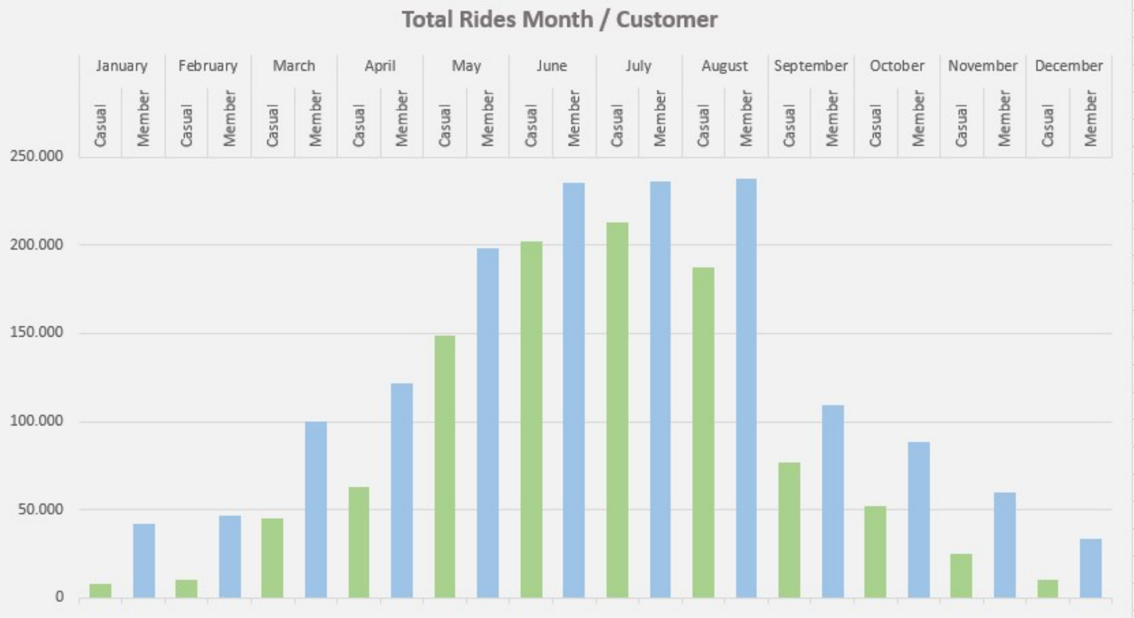
We can dig deep into the data and check the number of weekly rides by the time of day.

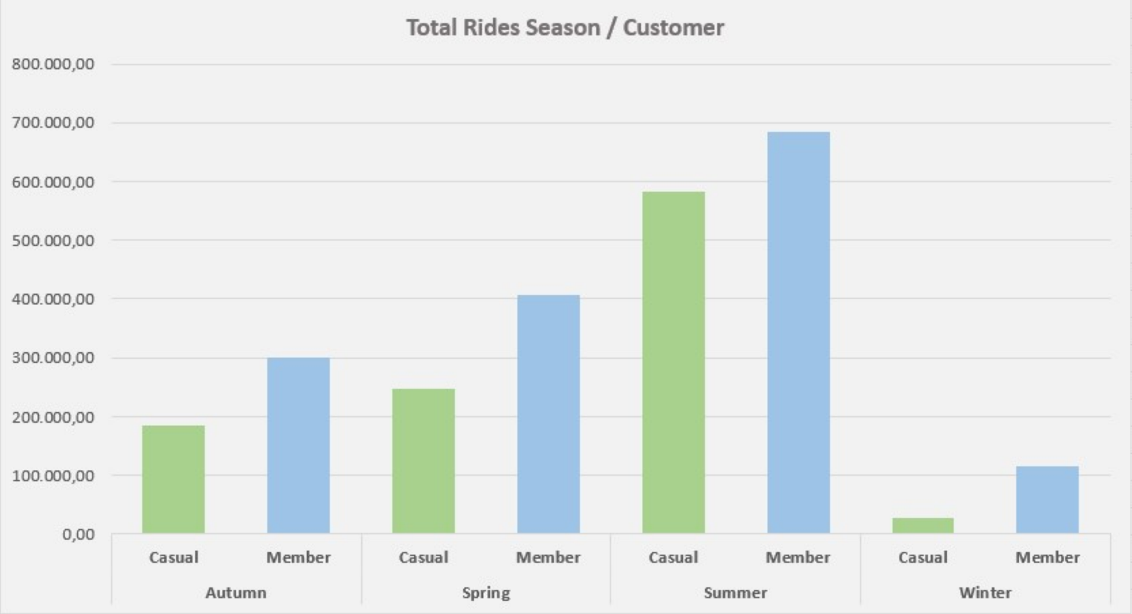
```
SELECT Customer, Part_of_Day, Day_of_Week ,COUNT (Started)
FROM 2022_Final
GROUP BY Day_of_Week, Part_of_Day, Customer;
```



According to the weekly visual insights, we can conclude that member customers are people who usually ride in the middle of the week with constant usage from the morning until the end of the afternoon. The usage decreases in the evening.

The casual rider's patterns show us that mornings and evenings have similar ride patterns for casual; however, the evening rides in casual are longer than those for members.

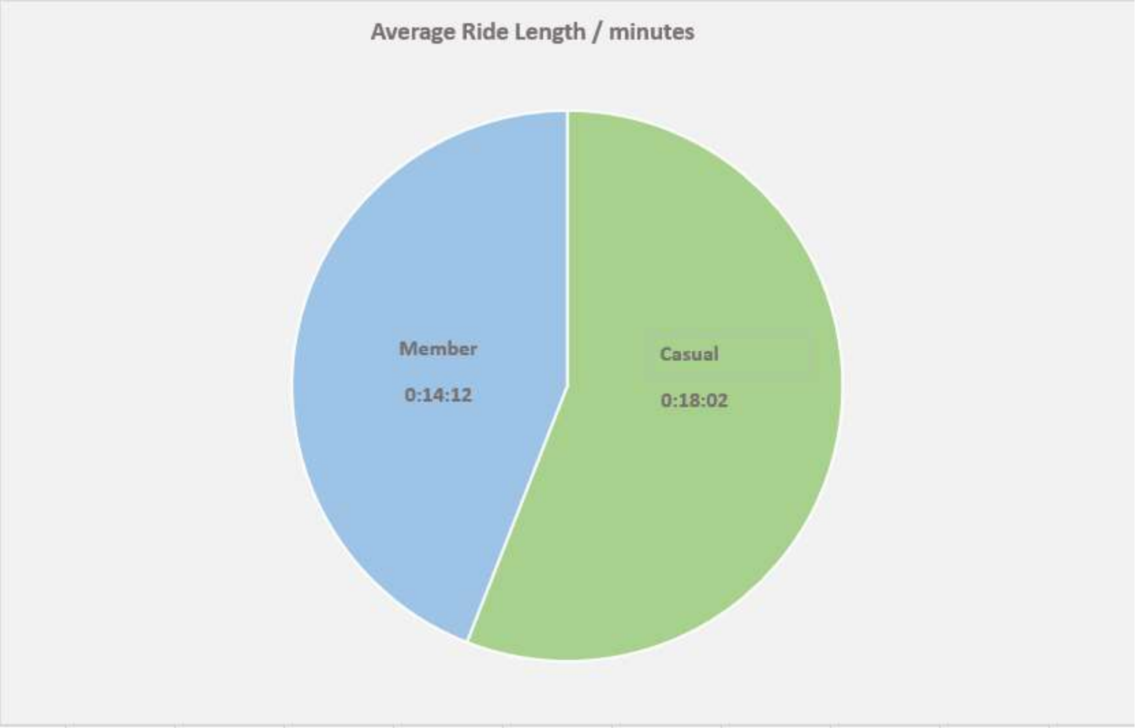




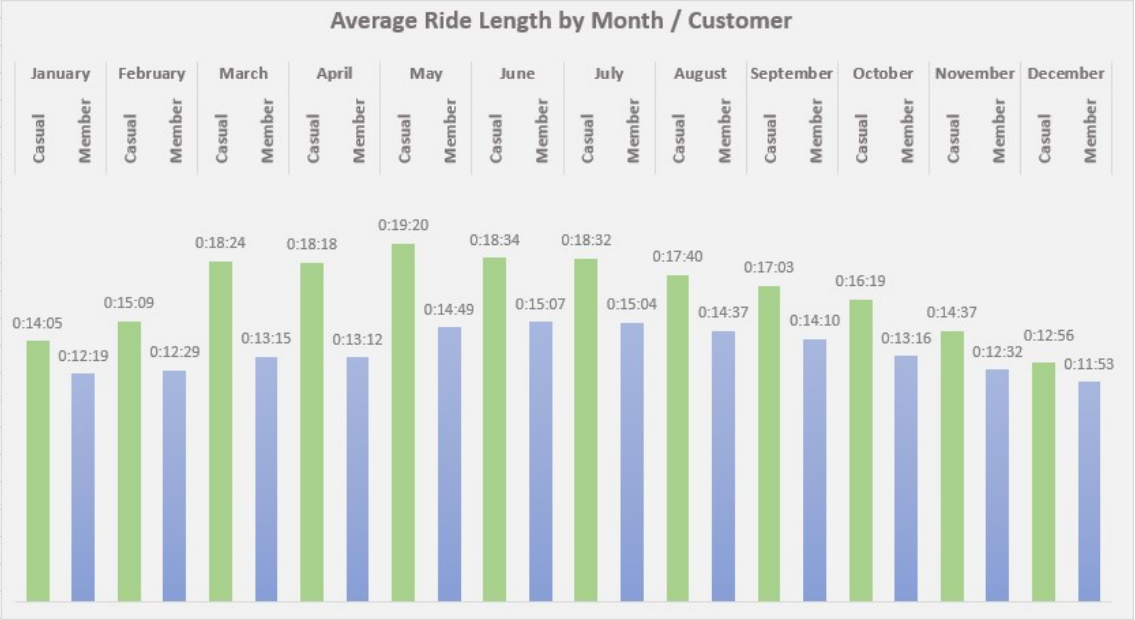
According to the monthly and seasonal statistics, we appreciate that for both casual and member clients, the preferred season is summer. The months of June, July, and August are the ones with the most demand. However, in **winter, casual customers decline at a higher rate than members**.

Let´s go deep into the information, and I'll keep an eye on the average ride length data. In order to do this, we must now run some queries to determine the **differences in the average ride length** between casual and member customers duration of rides.

```
SELECT
Customer, SEC_TO_TIME(AVG(TIME_TO_SEC (Ride_Length)))
FROM 2022_Final
GROUP BY Customer;
        #Average Member Ride Length is 14 min 12 sec
        #Average Casual Ride Length is 18 min 02 sec
```

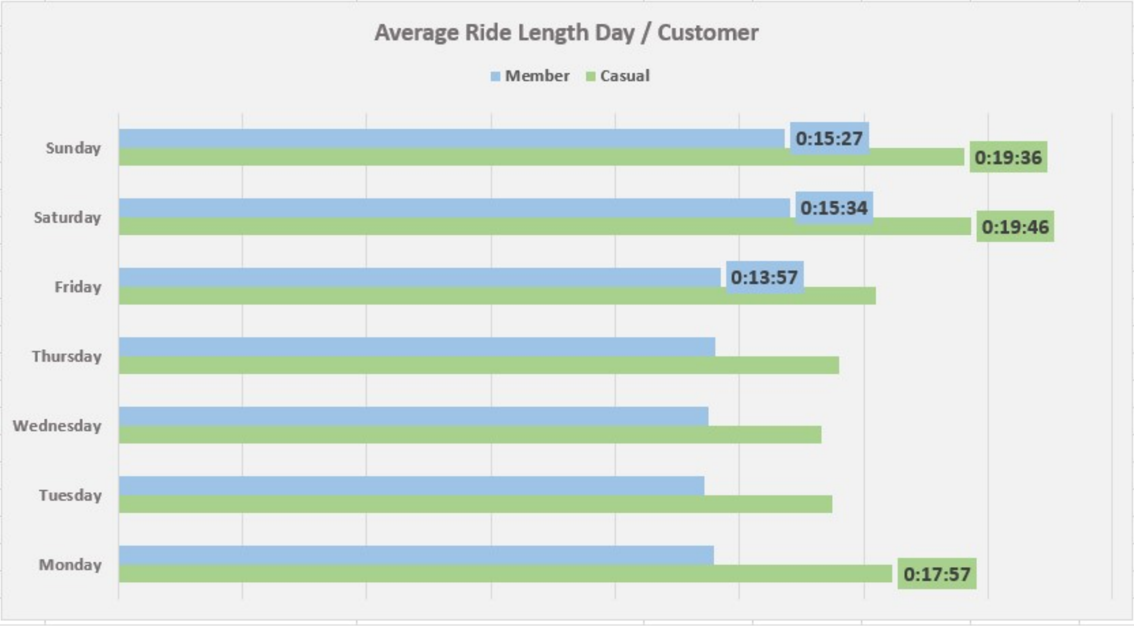


```
SELECT
Customer, Month, SEC_TO_TIME(AVG(TIME_TO_SEC (Ride_Length)))
AS 'Ride Length'
FROM 2022_Final
GROUP BY Customer, Month
ORDER BY 'Ride Length' DESC;
            # Average Ride Length by Month
```

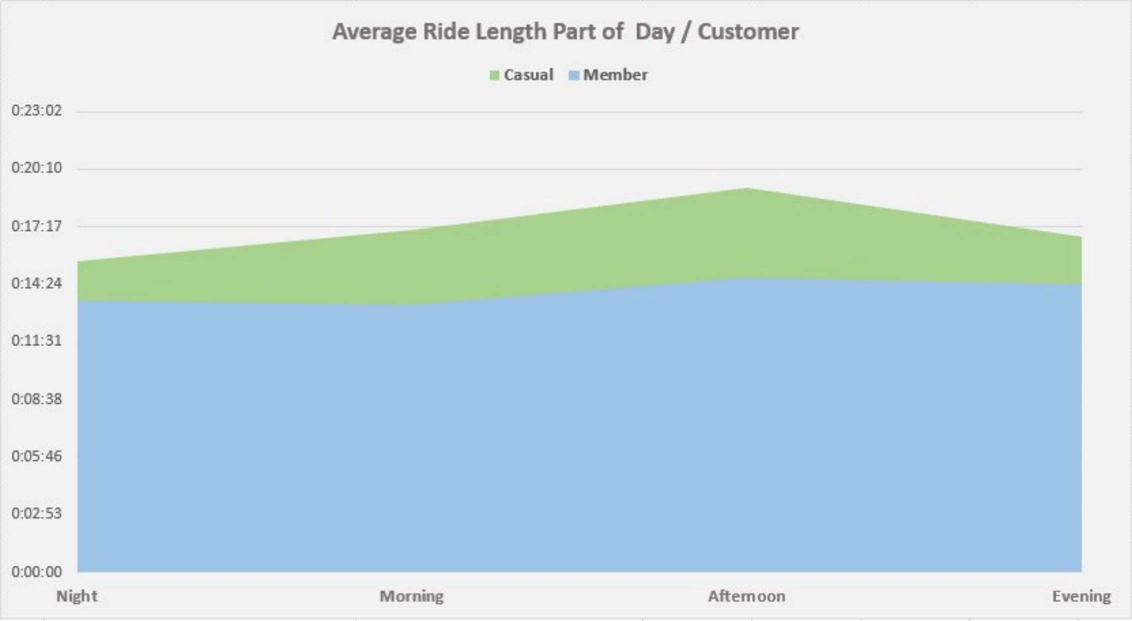


```
SELECT
```

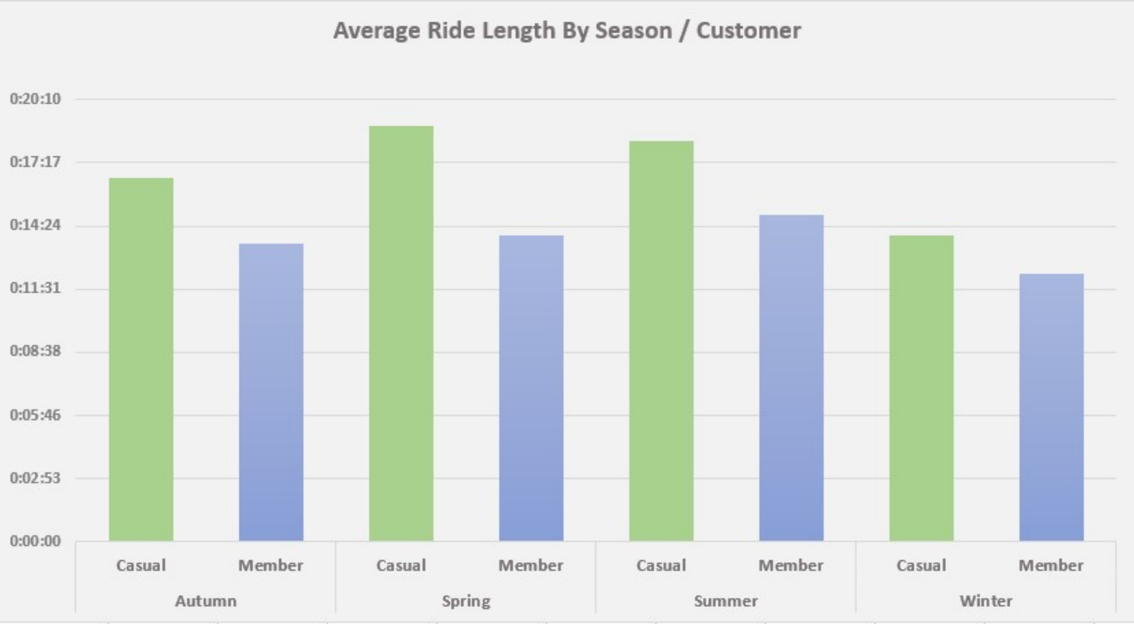
```
Customer,
Day_of_Week, SEC_TO_TIME(AVG(TIME_TO_SEC (Ride_Length)))    AS
'Ride Length'
FROM 2022_Final
GROUP BY Day_of_Week, Customer;
        # Average Ride Lenght by Day
```



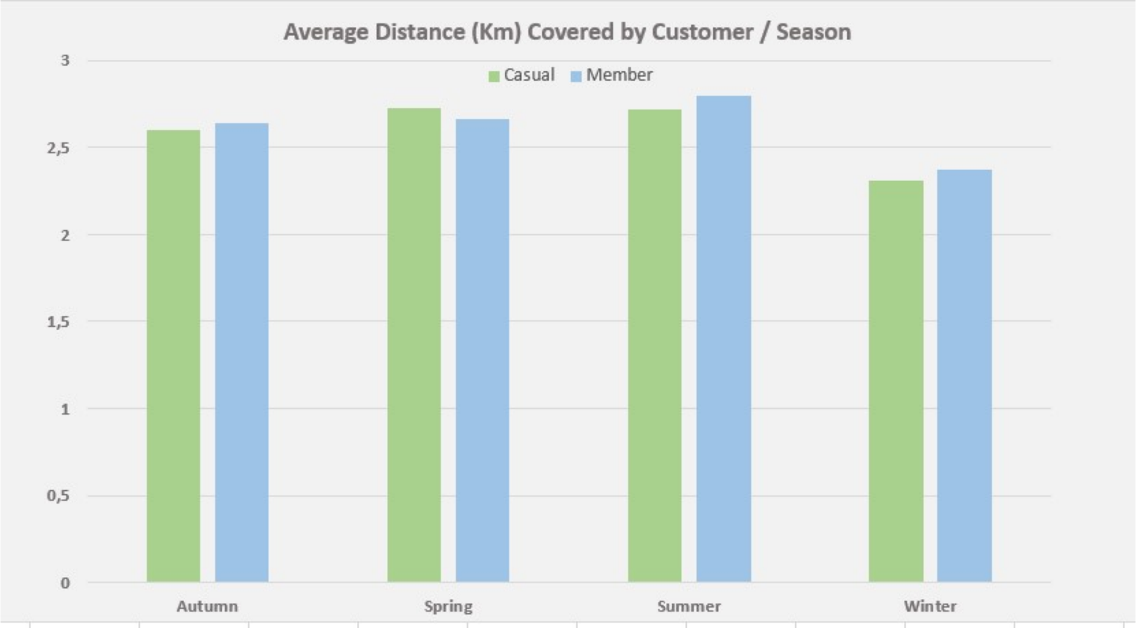
```
SELECT
Customer,
Part_of_Day, SEC_TO_TIME(AVG(TIME_TO_SEC (Ride_Length))) AS
'Ride Length'
FROM 2022_Final
GROUP BY Part_of_Day, Customer;
        # Average Ride Lenght by Part of Day
```



```
SELECT
Customer, Season, SEC_TO_TIME(AVG(TIME_TO_SEC (Ride_Length)))
AS 'Ride Length'
FROM 2022_Final
GROUP BY Customer, Season
ORDER BY 'Ride Length' DESC;
         # Average Ride Length by Season
```



Before we draw any conclusions, let's look at the distance traveled by season to reinforce our average length ride analysis.



The average ride length information shows us that casual riders spend more time on their trips across the different stations. We can reach the **same conclusion** as total ride statistics: casual customers may use the bike for leisure and enjoy it at their own pace.

Both casual and member customers spend more time on their trips in the afternoons, followed by the mornings.

However, despite what we expected to find based on total rides, average ride length patterns reveal two interesting aspects to consider about casual riders:

1. Monday´s data presents a longer average ride length than Fridays. This data invites us to ask the following question: Why do casual riders spend more time on their trips on Mondays than Fridays if we know that there are many more rides on Fridays?

2. Casual riders' average length is higher in the spring than in the summer. This hypothesis could be reinforced with the average distance covered by season, which would show us that casual riders traveled more distance in the spring season.

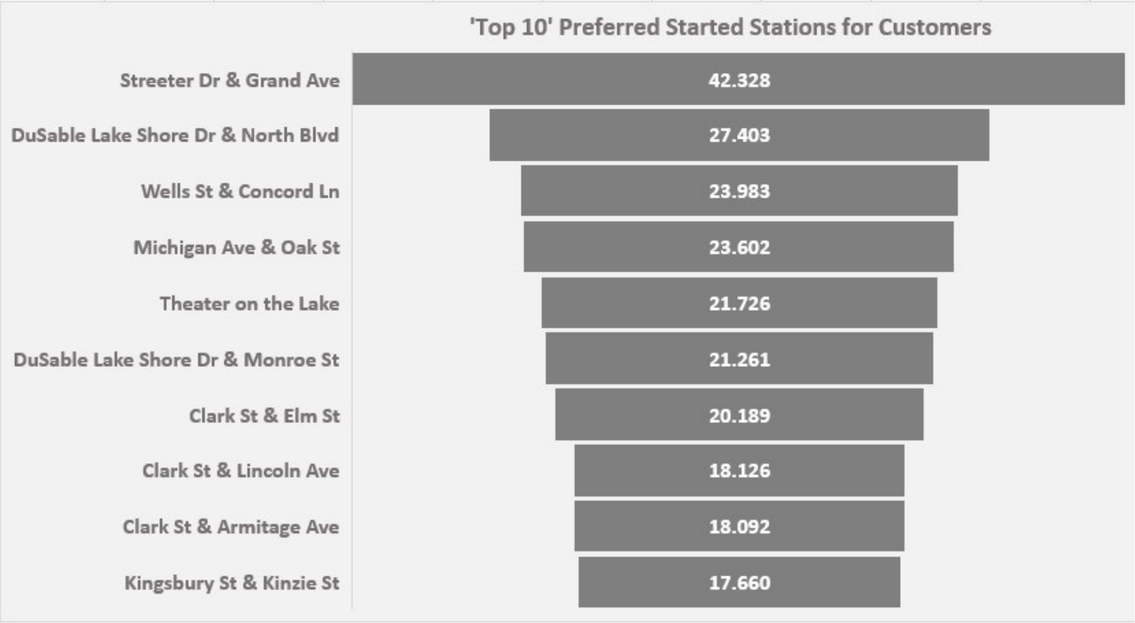
As well as casual riders' analysis, when we compare the average ride length patterns of the members to the total ride statistics, we notice a significant difference:

1. Member riders spend more time on weekends, while there are more rides in the middle of the week.

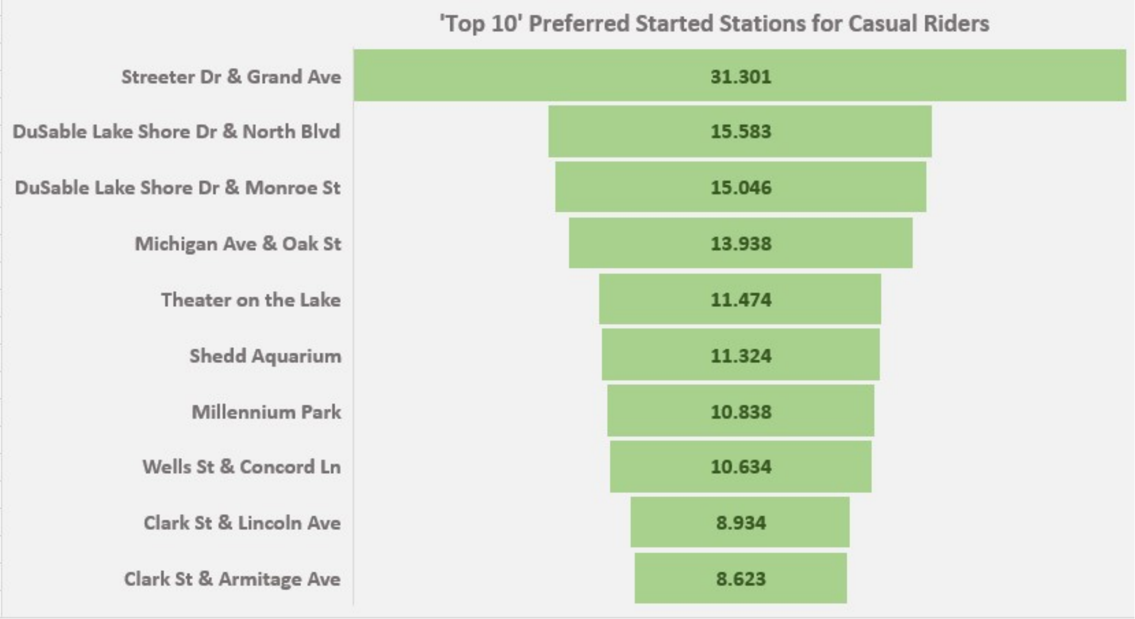
The average ride length analysis could lead us to conclude that casual travelers are mostly people who enjoy trips as leisure or with cycling clubs for training or health habit purposes, while member customers are people who use the bike frequently both during the week and enjoy it on weekends too as a hobby or for training purposes, as well as with cycling clubs or family and friends.

Now I intend to analyze which are the **start and end stations most used** by the different types of customers.

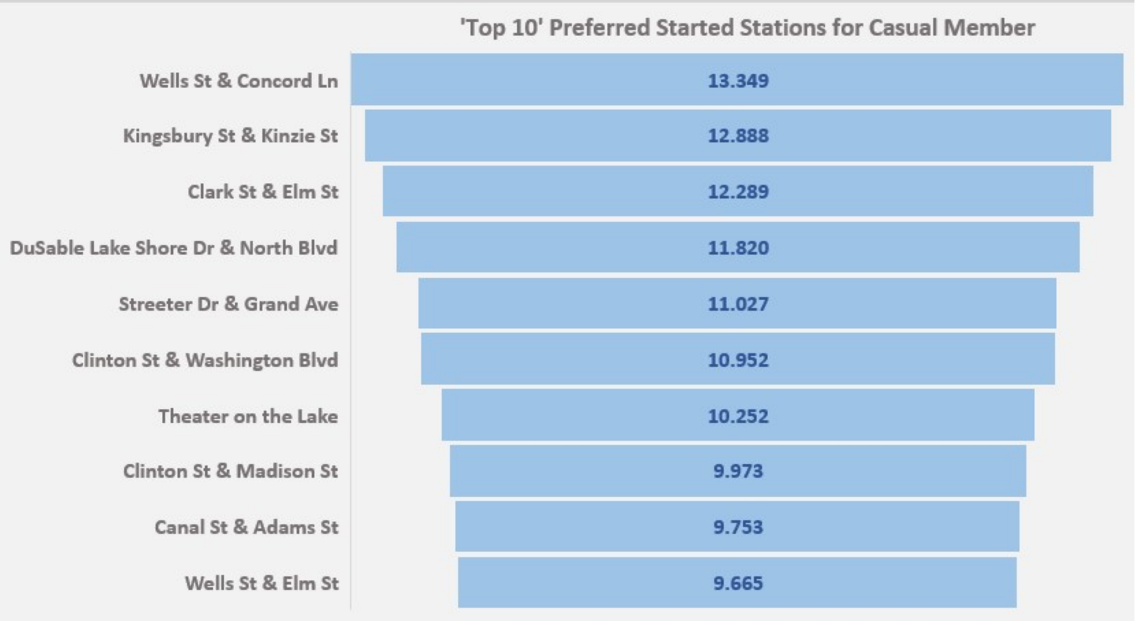
```
SELECT Start_Station_Name, COUNT(Start_Station_Name)              AS
Start_Station_Ranking
FROM 2022_Final
GROUP BY Start_Station_Name
ORDER BY Start_Station_Ranking DESC
LIMIT 10;
```



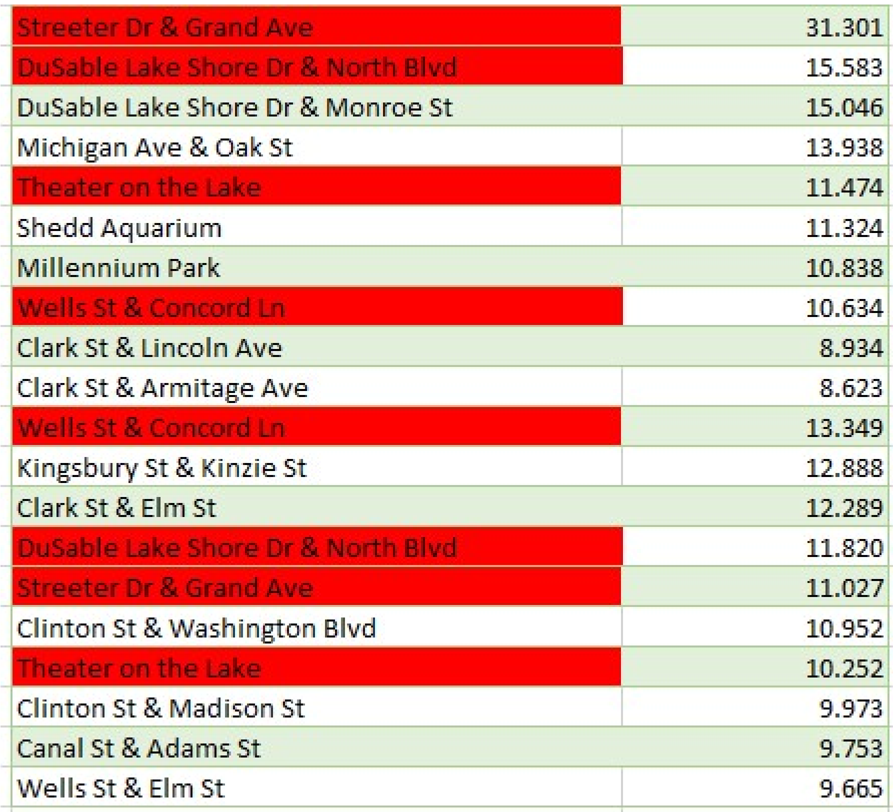
```
SELECT Start_Station_Name, COUNT(Start_Station_Name)              AS
Start_Station_Ranking
FROM 2022_Final
WHERE Customer ='Casual'
GROUP BY Start_Station_Name
ORDER BY Start_Station_Ranking DESC
LIMIT 10;
```



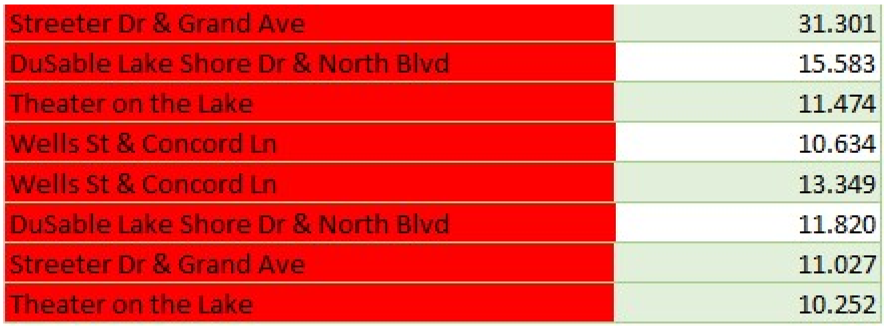
```
SELECT Start_Station_Name, COUNT(Start_Station_Name) AS
Start_Station_Ranking
FROM 2022_Final
WHERE Customer ='Member'
GROUP BY Start_Station_Name
ORDER BY Start_Station_Ranking DESC
LIMIT 10;
```



We got the list of preferred stations for casual and member customers, but I want to know which stations are **shared as preferred in the Top 10 casual and member preferences**. To accomplish this, I combined two tables into a single Excel table. Then I used conditional values to mark the duplicates in the table, removed the distinct ones, and finally removed duplicates.

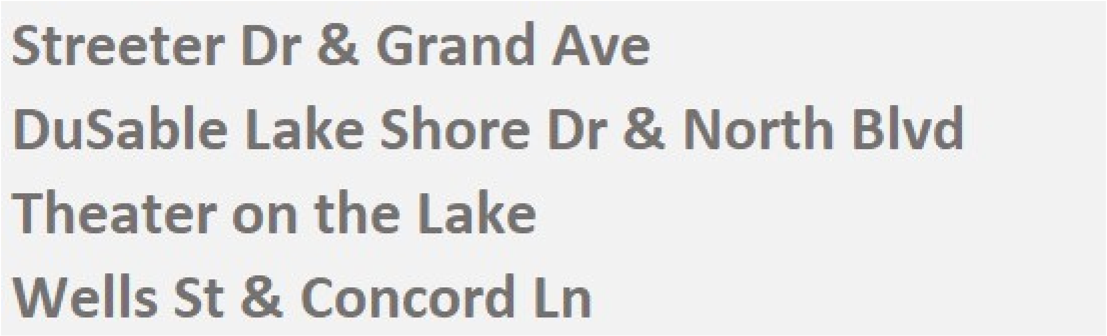


Streeter Dr & Grand Ave	31.301
DuSable Lake Shore Dr & North Blvd	15.583
DuSable Lake Shore Dr & Monroe St	15.046
Michigan Ave & Oak St	13.938
Theater on the Lake	11.474
Shedd Aquarium	11.324
Millennium Park	10.838
Wells St & Concord Ln	10.634
Clark St & Lincoln Ave	8.934
Clark St & Armitage Ave	8.623
Wells St & Concord Ln	13.349
Kingsbury St & Kinzie St	12.888
Clark St & Elm St	12.289
DuSable Lake Shore Dr & North Blvd	11.820
Streeter Dr & Grand Ave	11.027
Clinton St & Washington Blvd	10.952
Theater on the Lake	10.252
Clinton St & Madison St	9.973
Canal St & Adams St	9.753
Wells St & Elm St	9.665



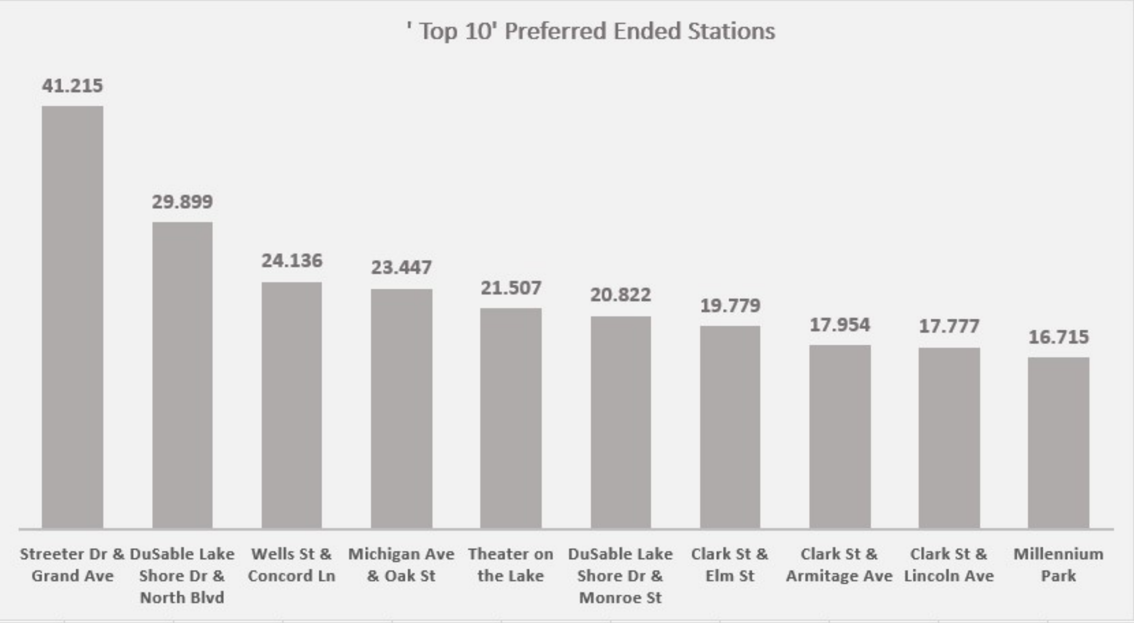
Streeter Dr & Grand Ave	31.301
DuSable Lake Shore Dr & North Blvd	15.583
Theater on the Lake	11.474
Wells St & Concord Ln	10.634
Wells St & Concord Ln	13.349
DuSable Lake Shore Dr & North Blvd	11.820
Streeter Dr & Grand Ave	11.027
Theater on the Lake	10.252

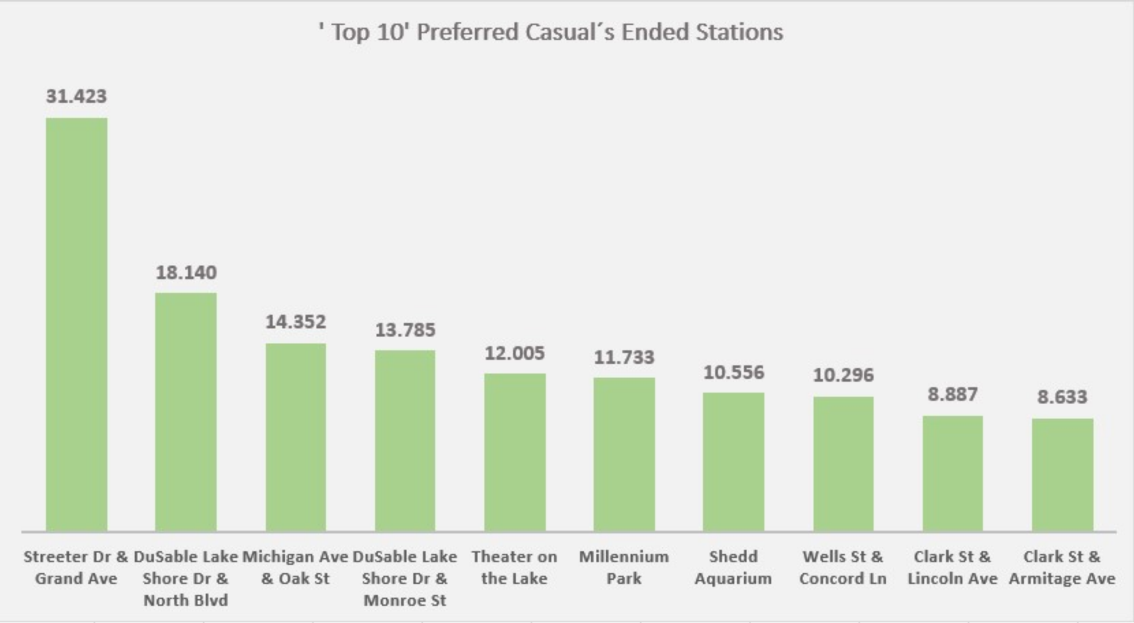
Matching values in the preference start station lists of members and casuals

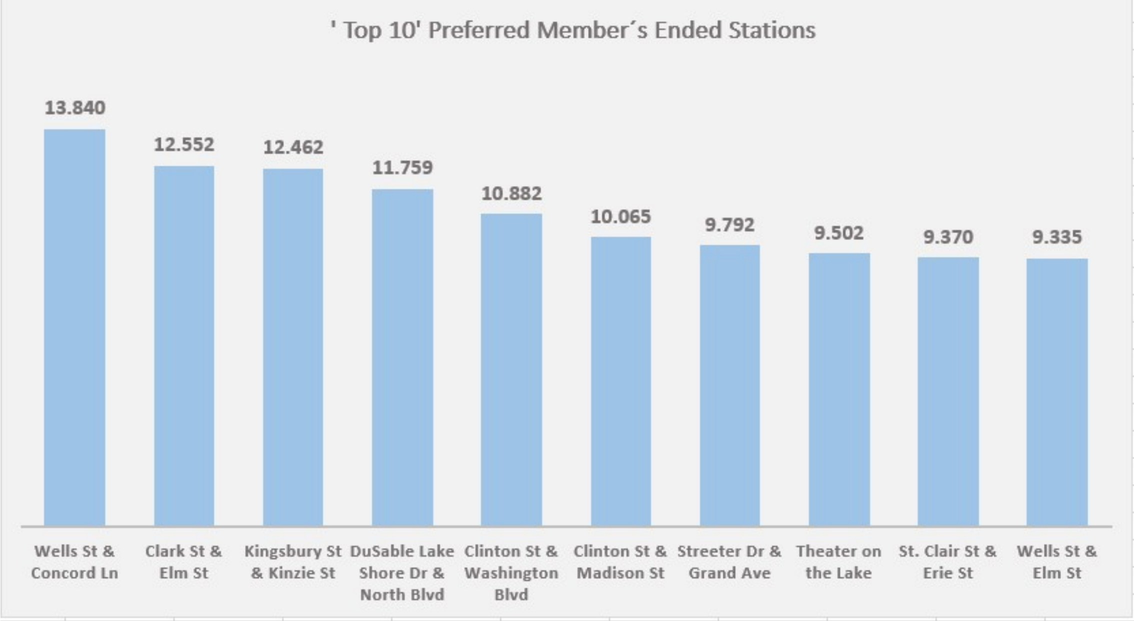


Streeter Dr & Grand Ave
DuSable Lake Shore Dr & North Blvd
Theater on the Lake
Wells St & Concord Ln

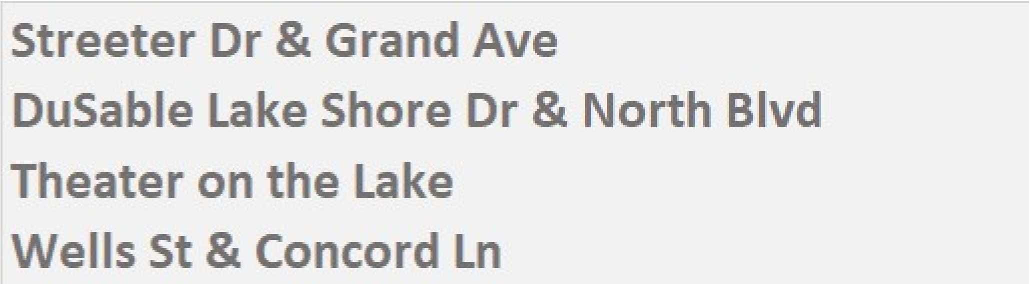
The procedure was repeated to obtain the preferences of the end stations' statistics. Refer to the same SQL queries above, but with End_Station_Name instead of Start_Station_Name.

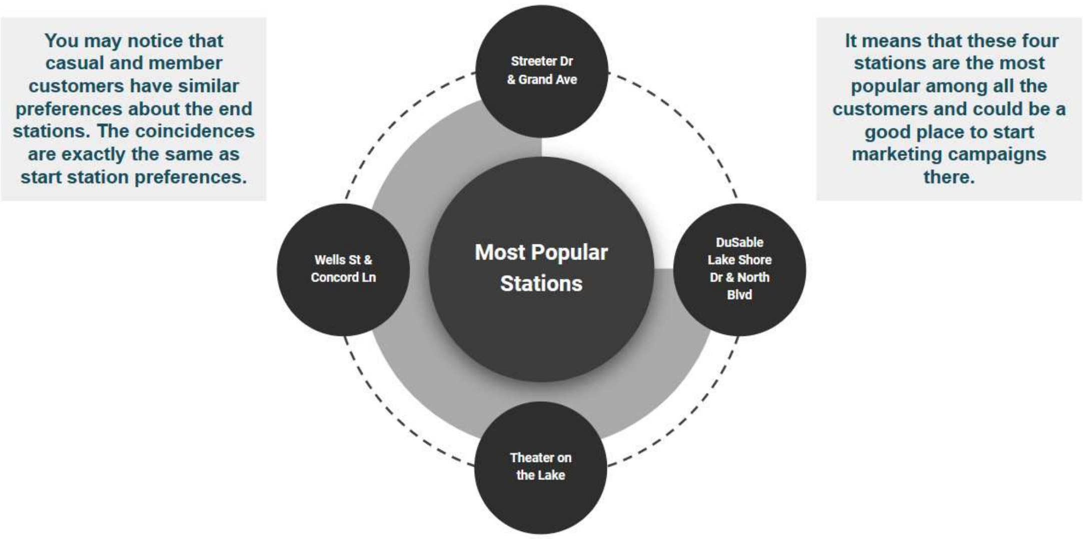






Matching values in the preference end-station lists of members and casuals

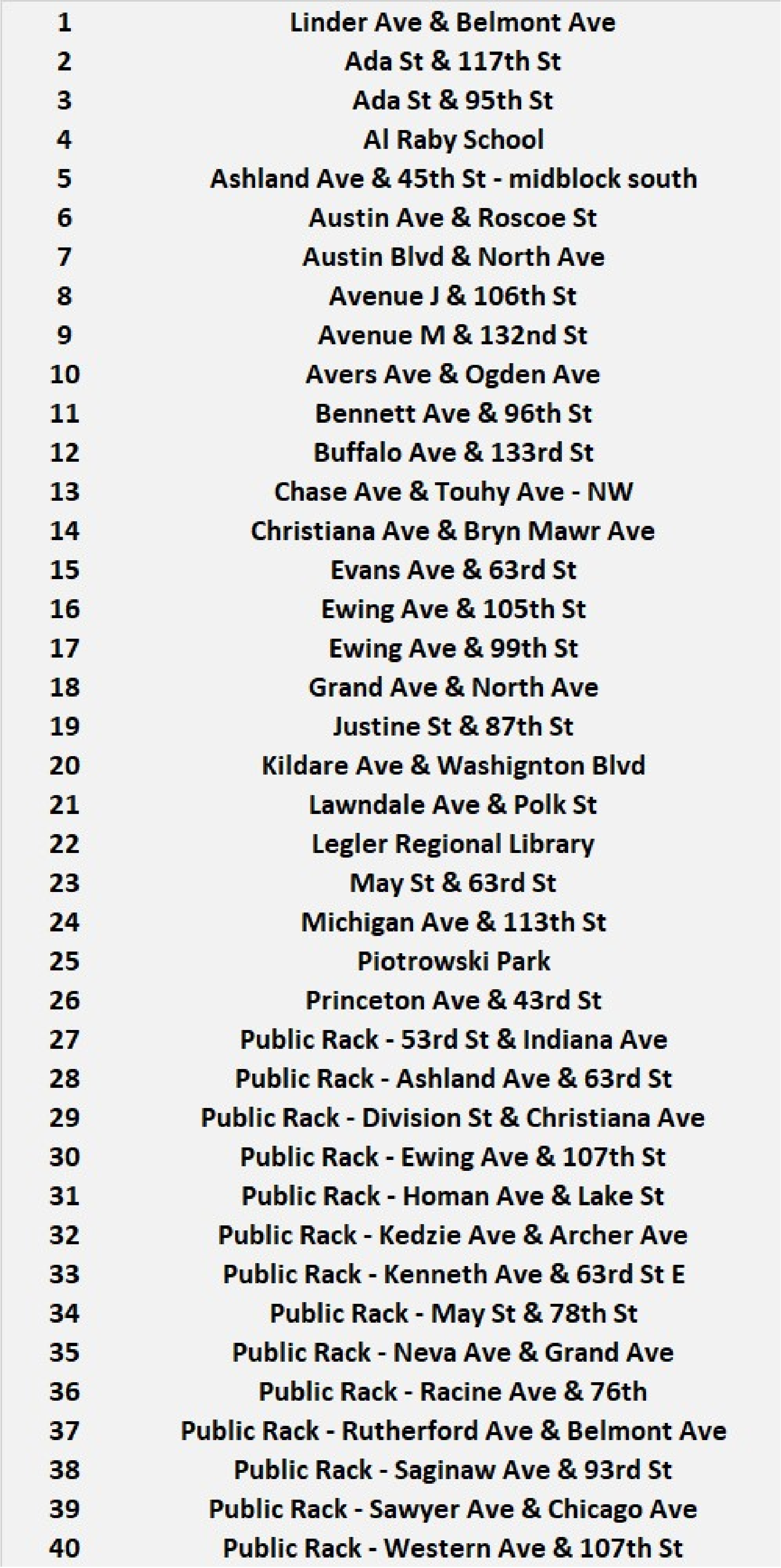

Streeter Dr & Grand Ave
DuSable Lake Shore Dr & North Blvd
Theater on the Lake
Wells St & Concord Ln


You may notice that casual and member customers have similar preferences about the end stations. The coincidences are exactly the same as start station preferences.

It means that these four stations are the most popular among all the customers and could be a good place to start marketing campaigns there.

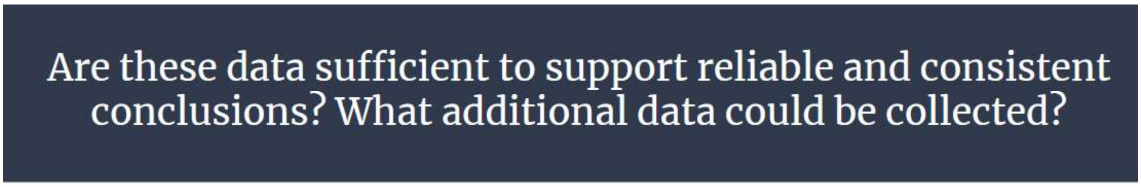
You may notice that casual and member customers have similar preferences about the end stations. The **coincidences are exactly the same as start station preferences**, so it means that these four stations are the most popular among all the customers and could be a good place to start marketing campaigns there.

The last part of my analysis was focused on the detection of disused stations. I ran two separate queries to identify those stations (starting and ending stations) with a lower abundance of riders (less than 2 riders). After running the queries, I exported the results into an Excel spreadsheet and merged them into one unique table. Duplicated values indicate stations that are no longer in use for both starting and ending points, allowing us to consider moving them to another location. A separate study is required. There are 40 disused stations.

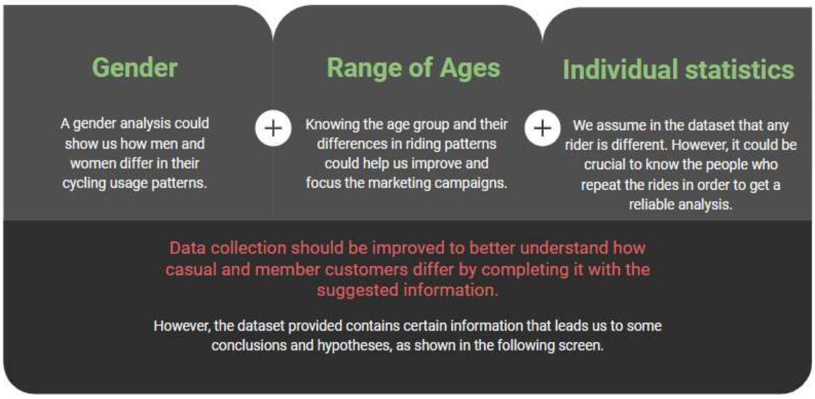


1	Linder Ave & Belmont Ave
2	Ada St & 117th St
3	Ada St & 95th St
4	Al Raby School
5	Ashland Ave & 45th St - midblock south
6	Austin Ave & Roscoe St
7	Austin Blvd & North Ave
8	Avenue J & 106th St
9	Avenue M & 132nd St
10	Avers Ave & Ogden Ave
11	Bennett Ave & 96th St
12	Buffalo Ave & 133rd St
13	Chase Ave & Touhy Ave - NW
14	Christiana Ave & Bryn Mawr Ave
15	Evans Ave & 63rd St
16	Ewing Ave & 105th St
17	Ewing Ave & 99th St
18	Grand Ave & North Ave
19	Justine St & 87th St
20	Kildare Ave & Washignton Blvd
21	Lawndale Ave & Polk St
22	Legler Regional Library
23	May St & 63rd St
24	Michigan Ave & 113th St
25	Piotrowski Park
26	Princeton Ave & 43rd St
27	Public Rack - 53rd St & Indiana Ave
28	Public Rack - Ashland Ave & 63rd St
29	Public Rack - Division St & Christiana Ave
30	Public Rack - Ewing Ave & 107th St
31	Public Rack - Homan Ave & Lake St
32	Public Rack - Kedzie Ave & Archer Ave
33	Public Rack - Kenneth Ave & 63rd St E
34	Public Rack - May St & 78th St
35	Public Rack - Neva Ave & Grand Ave
36	Public Rack - Racine Ave & 76th
37	Public Rack - Rutherford Ave & Belmont Ave
38	Public Rack - Saginaw Ave & 93rd St
39	Public Rack - Sawyer Ave & Chicago Ave
40	Public Rack - Western Ave & 107th St

5. Share Phase: Now is the time to share, through a visualization, all the insights that we found in the analysis.



Are these data sufficient to support reliable and consistent conclusions? What additional data could be collected?



Gender

A gender analysis could show us how men and women differ in their cycling usage patterns.

+

Range of Ages

Knowing the age group and their differences in riding patterns could help us improve and focus the marketing campaigns.

+

Individual statistics

We assume in the dataset that any rider is different. However, it could be crucial to know the people who repeat the rides in order to get a reliable analysis.

Data collection should be improved to better understand how casual and member customers differ by completing it with the suggested information.

However, the dataset provided contains certain information that leads us to some conclusions and hypotheses, as shown in the following screen.

I shared my conclusion with the stakeholders. I summarized the conclusions as follows:

Summary

- Electric bikes are preferred by both members and casual riders. Docked bikes are in progressive disuse.

- Member customers have a high number of rides in the middle of the week, while casuals prefer weekends.

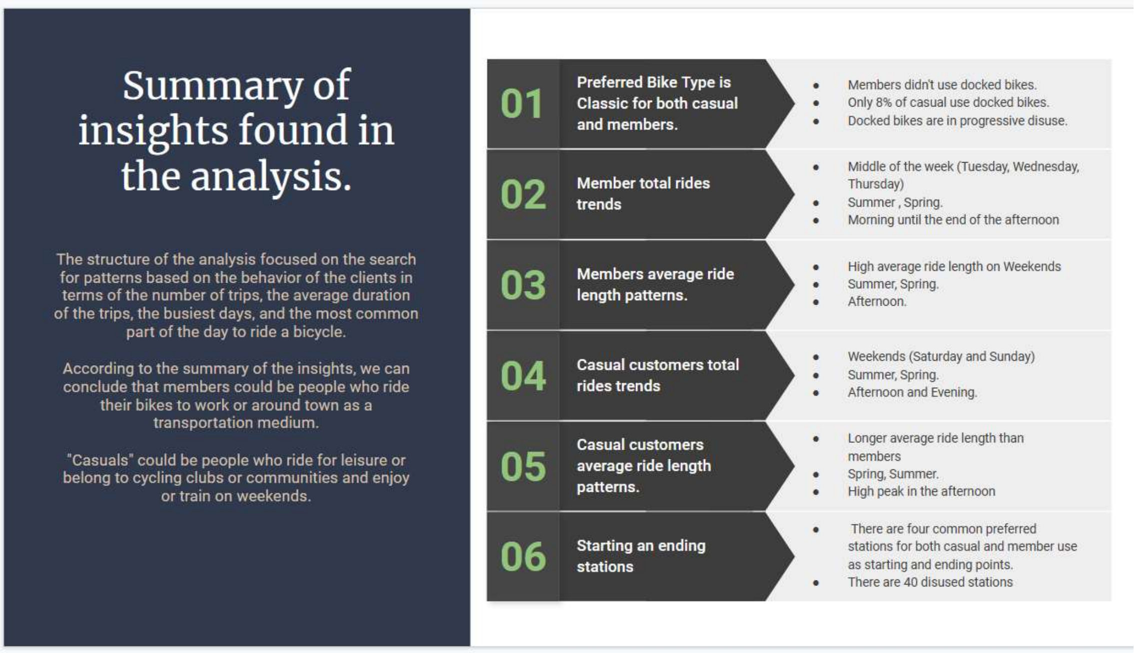
- For both, the most usual time of day to ride is the afternoon; however, we have to pay attention to the significant difference in favor of members in the morning. Casuals have a higher number of rides than members in the evenings instead.

- Casual customers spend more time riding than members.

- Most cyclists (both casual and members) prefer the summer; however, casual cyclists spend more time on average riding and cover more distance in the spring.

- There are four stations with high demand for both the starting and ending points, which are shared among casual and member customers' preferences.

- There are 40 disused stations, which means they add no value to the business.



Summary of insights found in the analysis.

The structure of the analysis focused on the search for patterns based on the behavior of the clients in terms of the number of trips, the average duration of the trips, the busiest days, and the most common part of the day to ride a bicycle.

According to the summary of the insights, we can conclude that members could be people who ride their bikes to work or around town as a transportation medium.

"Casuals" could be people who ride for leisure or belong to cycling clubs or communities and enjoy or train on weekends.

01 **Preferred Bike Type is Classic for both casual and members.**
- Members didn't use docked bikes.
- Only 8% of casual use docked bikes.
- Docked bikes are in progressive disuse.

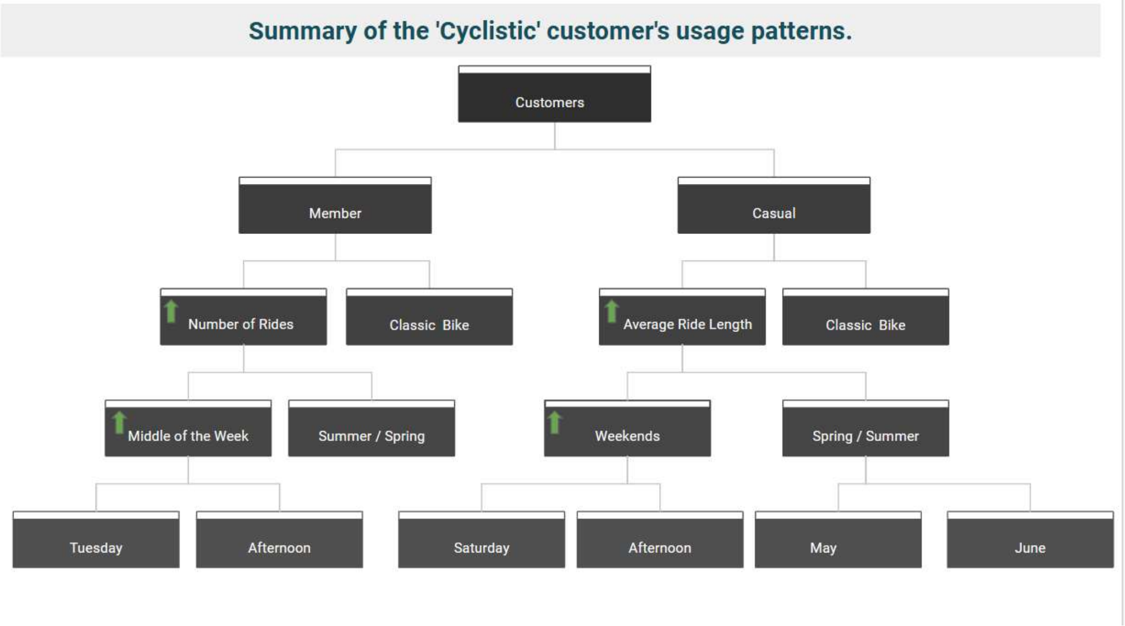
02 **Member total rides trends**
- Middle of the week (Tuesday, Wednesday, Thursday)
- Summer , Spring.
- Morning until the end of the afternoon

03 **Members average ride length patterns.**
- High average ride length on Weekends
- Summer, Spring.
- Afternoon.

04 **Casual customers total rides trends**
- Weekends (Saturday and Sunday)
- Summer, Spring.
- Afternoon and Evening.

05 **Casual customers average ride length patterns.**
- Longer average ride length than members
- Spring, Summer.
- High peak in the afternoon

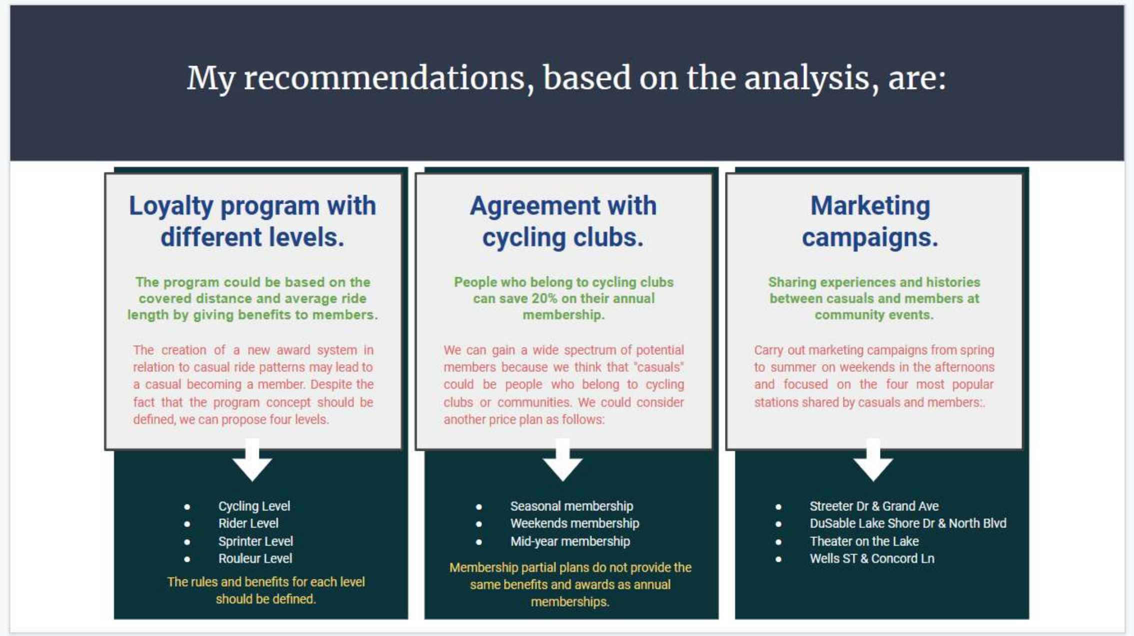
06 **Starting an ending stations**
- There are four common preferred stations for both casual and member use as starting and ending points.
- There are 40 disused stations



Summary of the 'Cyclistic' customer's usage patterns.

6. Act Phase: I have to include my three top recommendations based on the analysis.

According to the analyzed data and summary of the information found there, I would like to recommend the following actions in order to achieve the goal of converting casual customers into members:



My recommendations, based on the analysis, are:

Loyalty program with different levels.

The program could be based on the covered distance and average ride length by giving benefits to members.

The creation of a new award system in relation to casual ride patterns may lead to a casual becoming a member. Despite the fact that the program concept should be defined, we can propose four levels.

- Cycling Level
- Rider Level
- Sprinter Level
- Rouleur Level

The rules and benefits for each level should be defined.

Agreement with cycling clubs.

People who belong to cycling clubs can save 20% on their annual membership.

We can gain a wide spectrum of potential members because we think that "casuals" could be people who belong to cycling clubs or communities. We could consider another price plan as follows:

- Seasonal membership
- Weekends membership
- Mid-year membership

Membership partial plans do not provide the same benefits and awards as annual memberships.

Marketing campaigns.

Sharing experiences and histories between casuals and members at community events.

Carry out marketing campaigns from spring to summer on weekends in the afternoons and focused on the four most popular stations shared by casuals and members:.

- Streeter Dr & Grand Ave
- DuSable Lake Shore Dr & North Blvd
- Theater on the Lake
- Wells ST & Concord Ln

* Creating a new awards program based on the covered distance and average ride length by giving benefits to members This action may lead to a casual becoming a member. We could suggest the names of the different levels based on cycling slang. The loyalty program mechanism should be defined.

* Agreement with cycling clubs. People who belong to cycling clubs can save 20% on their annual membership. We can gain a wide spectrum of potential members because we think that "casuals" could be people who belong to cycling clubs or communities. We could consider an additional price plan as follows: seasonal membership, weekends membership, or mid-year membership. However, Membership partial plans do not provide the same benefits and awards as annual memberships.

* Carry out marketing campaigns from spring to summer on weekends in the afternoons and focused on the four most popular stations shared by casuals and members: Streeter Dr & Grand Ave, DuSable Lake Shore Dr & North Blvd, Theater on the Lake and Wells St & Concord Ln.

I'd like to thank you for taking the time to read about my work process. It will be great if you want to share with me your feedback in order to improve and continue learning. Feel free to contact me at any time to discuss the work method.

David Sánchez.