Objectives of the project

We want to identify the differences between casual riders and annual members and any trends from historical trip data to build a targeted marketing campaign to convert casual riders into annual members.

By understanding how casual riders and annual members use the bikes, we can identify any opportunities for casual riders who could potentially be convinced to join the membership.

Stakeholders

- Executive team who will decide whether to approve our recommendations.
- Our manager, the director of marketing, Lily Moreno.
- My team and I who will collect, analyze and report the data.

Prepare

The monthly data collected from Divvy Bikes are from June 2023 to May 2024. As the volume of data exceeds 5 million rows, I will use PostgreSQL to store and clean the entire dataset.

Problems encountered when importing datetime values.

- Could not import datetime values
    a. I created the 'started_at' and 'ended_at' columns as **date** datatype. Turns out I was supposed to store them as **timestamp** datatype.
- Could not convert date datatype to timestamp datatype.

```
ALTER TABLE public.cyclistic_fy_data
    ALTER COLUMN started_at type timestamp without time zone;
```

```
Messages
ERROR:  column "started_at" cannot be cast automatically to type timestamp without time zone
HINT:  You might need to specify "USING started_at::timestamp without time zone".
```

Solution

I had to convert the columns to varchar datatype first before changing it to timestamp datatype.

```
ALTER TABLE public.cyclistic_fy_data
    ALTER COLUMN started_at type varchar ;
```

```
ALTER TABLE public.cyclistic_fy_data
    ALTER COLUMN started_at type timestamp without time zone
    USING started_at::timestamp without time zone;
```

Process

I will be keeping the null values for station IDs and names as those columns can still provide important information such as start and end time, type of bicycle used and membership type.

Data checks

Check for no duplicate ride record --- passed

```
select count(*) as num_of_trips
from cyclistic_fy_data
group by ride_id
having count(*) >1;
```

Check ride id for no null --- passed

```
select *
    from cyclistic_fy_data
    where ride_id is null;
```

Create VIEW for filtered data

```
create table cleaned_data as
select * from cyclistic_fy_data
where (ended_at - started_at) > interval '10 seconds'

create or replace view time_data as (
    select ride_id,
    (ended_at - started_at) as ride_length,
    to_char(started_at, 'Day') as day_of_week,
    date(started_at) as date
    from cleaned_data);
```

- Filter out bike trips that last less than 10 seconds as we assume the initial 10 seconds are for registering the bike. We are also excluding negative interval data.

Data

Export data in batches according to month

```
select c.ride_id,
    c.member_casual,
c.rideable_type,
t.ride_length,
t.day_of_week,
t.date
from cleaned_data as c
inner join time_data as t
on c.ride_id = t.ride_id
    where date between '20231201' and '20231231'
order by date desc
```

	ride_id character varying (16)	member_casual character varying (20)	rideable_type character varying (15)	ride_length interval	day_of_week text	date date
1	CE4F251F4FE85152	member	electric_bike	00:04:00	Sunday	2023-12-31
2	E814607042E96408	casual	electric_bike	00:05:00	Sunday	2023-12-31
3	A0E882028592 53C9	member	electric_bike	00:06:00	Sunday	2023-12-31
4	400A7973C75148F6	member	electric_bike	00:07:00	Sunday	2023-12-31
5	BAAE2B03A4837475	member	classic_bike	00:15:00	Sunday	2023-12-31
6	0C0F561C27E3044D	casual	classic_bike	00:05:00	Sunday	2023-12-31
7	11E0E2B2D20BD0CA	member	classic_bike	00:04:00	Sunday	2023-12-31

Analyze

Question 1: Which group of riders ride longer and have more trips?

- Casual riders ride longer on average.
- Annual members have more trips in terms of volume.

```sql
select c.member_casual,
    avg(t.ride_length) as avg_ride_length,
    count(*) as num_of_trips
from cleaned_data as c
inner join time_data as t
on c.ride_id = t.ride_id
group by member_casual
```

	member_casual character varying (20)	avg_ride_length interval	num_of_trips bigint
1	casual	00:28:23.802342	2034296
2	member	00:13:01.461486	3664348

- Members do have a greater number of trips but do not ride longer on average.

Question 2: What is the average trip duration between casual riders and annual members grouped by bicycle type?

```sql
select c.member_casual,
    c.rideable_type,
    avg(t.ride_length) as avg_ride_length
from cleaned_data as c
inner join time_data as t
on c.ride_id = t.ride_id
group by member_casual, c.rideable_type
order by c.member_casual, avg_ride_length desc
```

	member_casual character varying (20)	rideable_type character varying (15)	avg_ride_length interval
1	casual	docked_bike	03:26:41.426622
2	casual	classic_bike	00:34:31.236627
3	casual	electric_bike	00:14:35.379966
4	member	classic_bike	00:14:35.279207
5	member	electric_bike	00:11:22.625571

- Casual riders spent the most time on docked bikes at an average of 3 and a half hours.
- **Member riders do not have any trips with docked bikes.**
- Casual riders spend 20 minutes longer (> 100%) on classic bikes than annual members.
- Causal riders spend 3 minutes (> 30%) longer on electric bikes.

Question 3: Which bicycles do casual riders and annual members ride the most?

```sql
select c.member_casual,
    c.rideable_type,
    count(*) as num_of_trips
from cleaned_data as c
inner join time_data as t
on c.ride_id = t.ride_id
group by member_casual, c.rideable_type
order by c.member_casual, num_of_trips desc
```

	member_casual character varying (20)	rideable_type character varying (15)	num_of_trips bigint
1	casual	electric_bike	1049796
2	casual	classic_bike	935330
3	casual	docked_bike	49170
4	member	classic_bike	1879898
5	member	electric_bike	1784450

- Electric bikes and classic bikes are the bicycles with the most number of trips for each membership group.
- Number of docked bike trips are very low.

Conclusion: Based on the data from question 2 and question 3, the marketing campaign should target casual riders who **ride electric bikes and classic bikes.**

Question 4: What are the days with the highest number of trips?

```sql
select c.member_casual,
    t.day_of_week,
    count(*) as num_of_trips
from cleaned_data as c
inner join time_data as t
on c.ride_id = t.ride_id
group by member_casual, day_of_week
order by c.member_casual, num_of_trips desc
```

	member_casual character varying (20)	day_of_week text	num_of_trips bigint
1	casual	Saturday	423082
2	casual	Sunday	332913
3	casual	Friday	309282
4	casual	Thursday	258936
5	casual	Wednesday	240606
6	casual	Monday	237178
7	casual	Tuesday	232299

	member_casual character varying (20)	day_of_week text	num_of_trips bigint
1	member	Thursday	592301
2	member	Wednesday	584956
3	member	Tuesday	562162
4	member	Friday	534442
5	member	Monday	506971
6	member	Saturday	480613
7	member	Sunday	402903

- Ridership among casual riders is higher closer to the end of the week.
- Ridership among members is higher during the weekdays

Conclusion: Members may be using the bikes for commuting to work as ridership numbers are lowest during the weekends. Therefore, our marketing campaign will be targeted towards **casual riders who commute to work during the weekdays.**

Question 5: Which month has the highest number of trips?

```
select to_char(date, 'Month') as month_of_year,
    count(*) as num_of_trips
from cleaned_data as c
inner join time_data as t
on c.ride_id = t.ride_id
group by to_char(date, 'Month')
order by num_of_trips desc
```

	month_of_year text	num_of_trips bigint
1	August	767142
2	July	762740
3	June	708494
4	September	662851
5	May	605393
6	October	534440
7	April	412101
8	November	358228
9	March	299886
10	February	222117
11	December	221438
12	January	143814

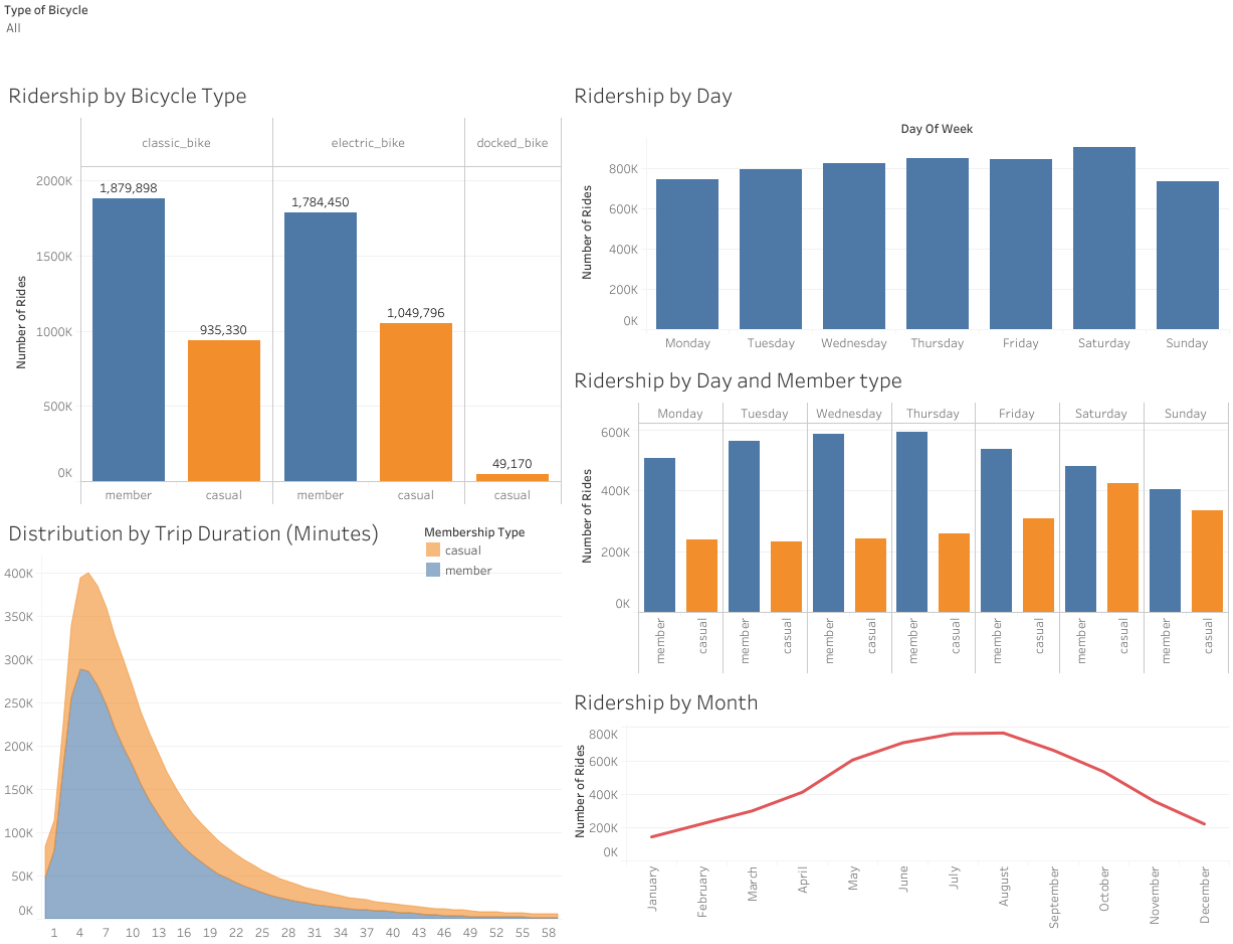
- The **top 4 months with the most number of trips** are August, July, June and September.
    - These are closer towards the second half of the year. This could be **due to the summer season.**
- The bottom 4 months with the least number of trips are March, February, December and January
    - These are close to or are at the start of the year. This could be due to the Winter season.
- Ridership numbers start picking up after Winter.

Summary

Based on what we have gathered from our analysis, the marketing strategy should target certain key points:

1. Target casual riders who ride the classic and electric bikes the most.
2. Target casual riders who use the bike to commute to work on weekdays (To mimic pattern of annual members).
3. Deploy the marketing campaign after the end of Winter.

Share + Act



Key Findings:

- Classic and electric bikes are the most used.
- Most rides take within 20 minutes to complete.
- Casual riders ride more as the week ends while annual members ride more when the week starts.
- Ridership begins to pick up rapidly in April while it starts to decline in August. (Spring to Summer)

Insights and Recommendations:

- Based on the information we have gathered; our marketing campaigns should be targeted towards working age adults who commute to work in the classic and electric bicycles. And it should be held at the end of Spring in May as ridership begin to increase rapidly.

Additional deliverables that would be helpful to include for further exploration:

- Cleaned data on the engagement rates of riders and the channels they were engaged from.
- Past yearly data to make y-o-y, q-o-q and m-o-m comparison.