CYCLISTIC RIDE-SHARING ANALYSIS

Differentiating casual riders from annual members to inform the upcoming marketing campaign.

OBJECTIVE

Cyclistic is a fictitious alias for a bicycle ridesharing company in Chicago. The goal of this mock is to help inform the marketing team, whose campaign is aimed towards converting "casual riders" into annual members. Towards this aim, I identify key differences between the groups and general patterns in casual riding to provide opportunities for effective marketing.

DATA

The data used was provided in csv files by month. We only consider the last 12 months in our analysis.

DATA

- The 12 separate csv files were loaded into dataframes and subsequently concatenated together for a master dataframe with ~5.75 million rows and 13 columns:
  - Ride ID
  - Rideable type: "classic", "electric", or "docked"
  - Datetimes of when ride started and ended (2 separate columns)
  - Station names and IDs for starting station and ending station (4 cols)
  - Latitude and longitude coordinates for each starting and ending station (4 cols)
  - Membership status: "member" or "casual"

DATA CLEANING

- 140 rides had the start time later than the end time, resulting in negative time spent on the bike.
- **Solution:** I swapped the start and end times for these rides. Due to the small relative amount and impossibility of accuracy, I could've dropped these observations instead but chose to potentially save as much data as possible.

- 790,000 observations were missing starting station names while 843,000 observations were missing ending station names, with a good deal of overlap. 4,766 observations were missing ending latitude and longitude coordinates, all these observations overlapped with those missing ending station names.
- **Solution:** There's no way to know where the 4,766 rides ended so those observations were dropped. For the rest of the missing data, a dictionary was created matching station names with known latitude and longitude coordinates and then using this dictionary to determine the missing station names based on their coordinates.
  - Some bias may be introduced as all the 4,766 dropped rides used electric bikes

ANALYSIS

There are **four** key differences and/or patterns that were identified:

1. **What**
2. **When**
3. **How Long**
4. **Where**

WHAT

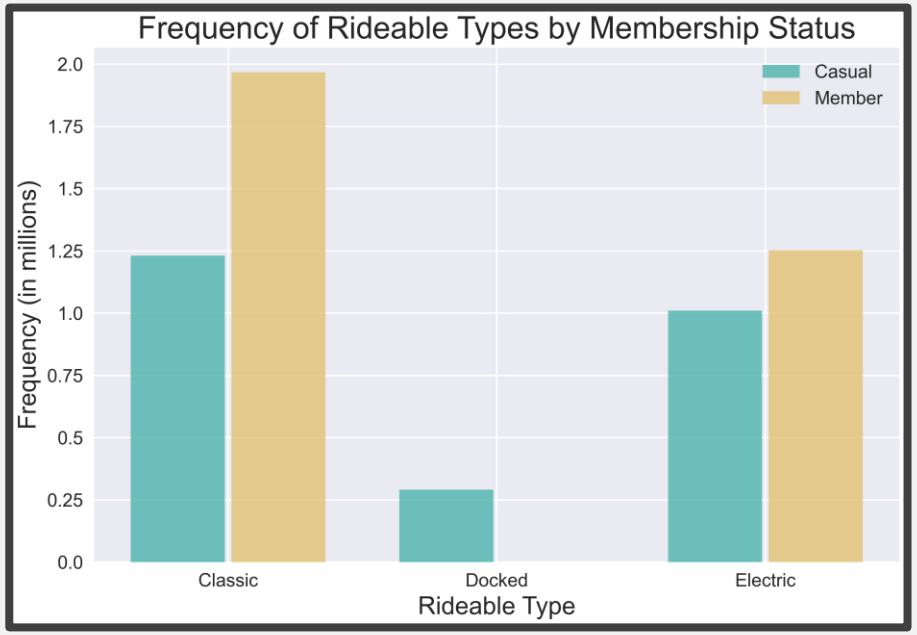
Cyclistic offers 3 rideables:

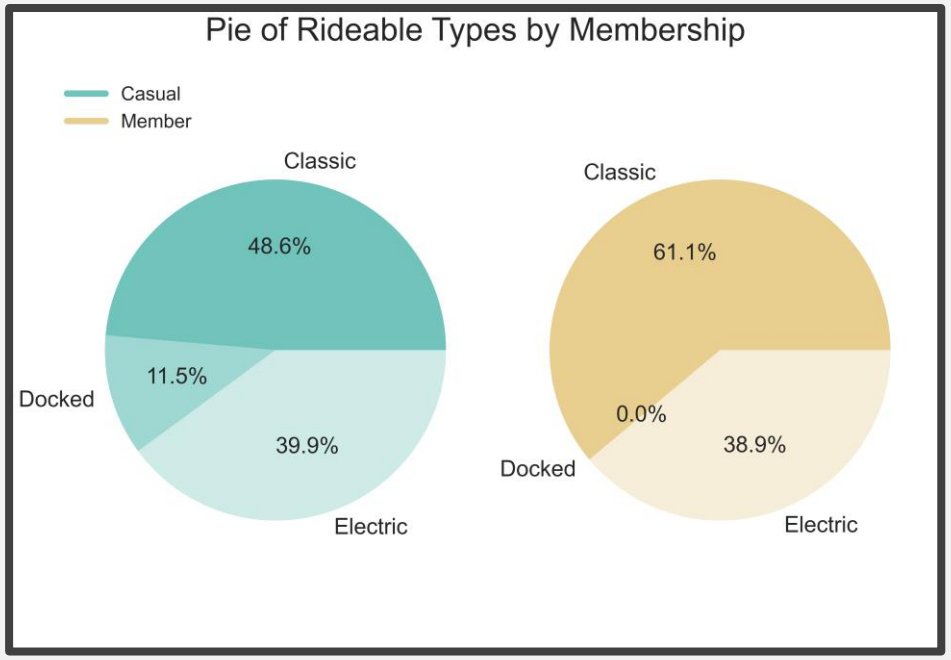
- Classic bike
- Electric bike
- Docked bike

How are member and casual rides distributed over these?

WHAT

- While members and casuals ride electric at the same rate, only one member ride was on a docked bike this past year. The difference is made up in classic bike rides.





ANALYSIS

There are **four** key differences and/or patterns that were identified:

1. What
2. **When**
3. **How Long**
4. **Where**

WHEN

When do casual riders and members ride bikes?

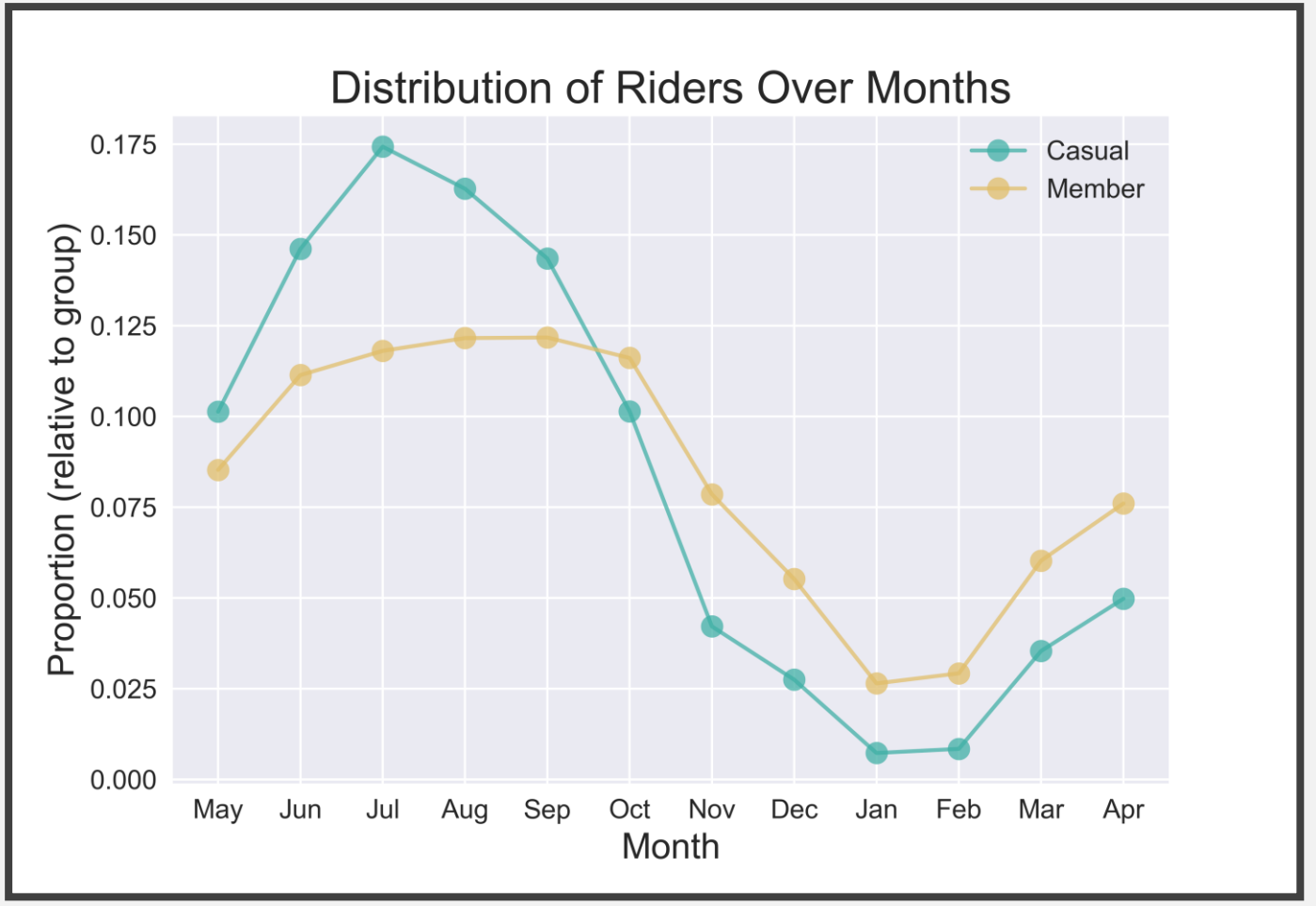
There are 3 levels of "when":

- Yearly scale over months
- Weekly scale over days
- Daily scale over hours/minutes

WHEN

There are 3 levels of "when":

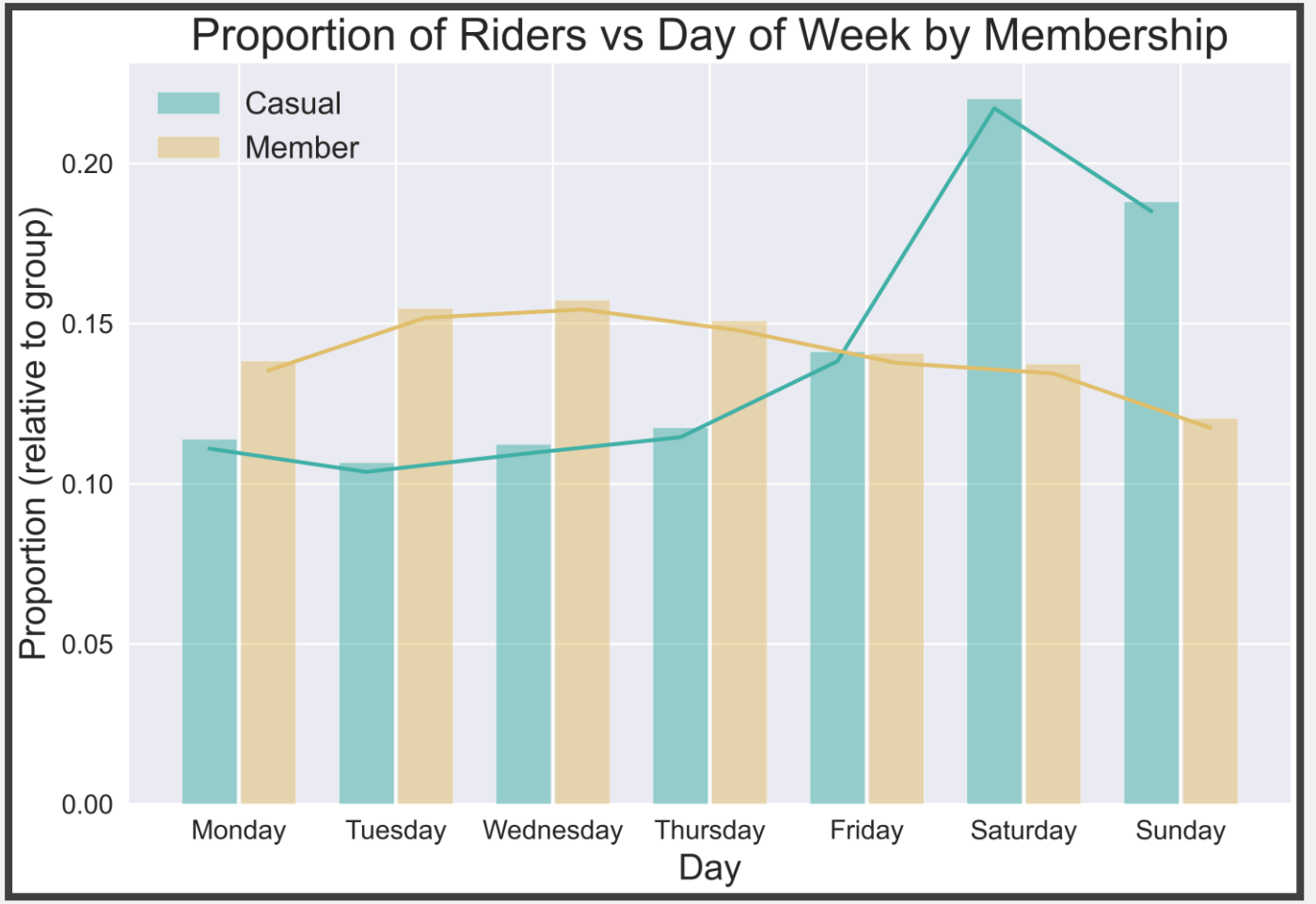
- **Yearly scale over months**
- Weekly scale over days
- Daily scale over hours/minutes



WHEN

There are 3 levels of "when":

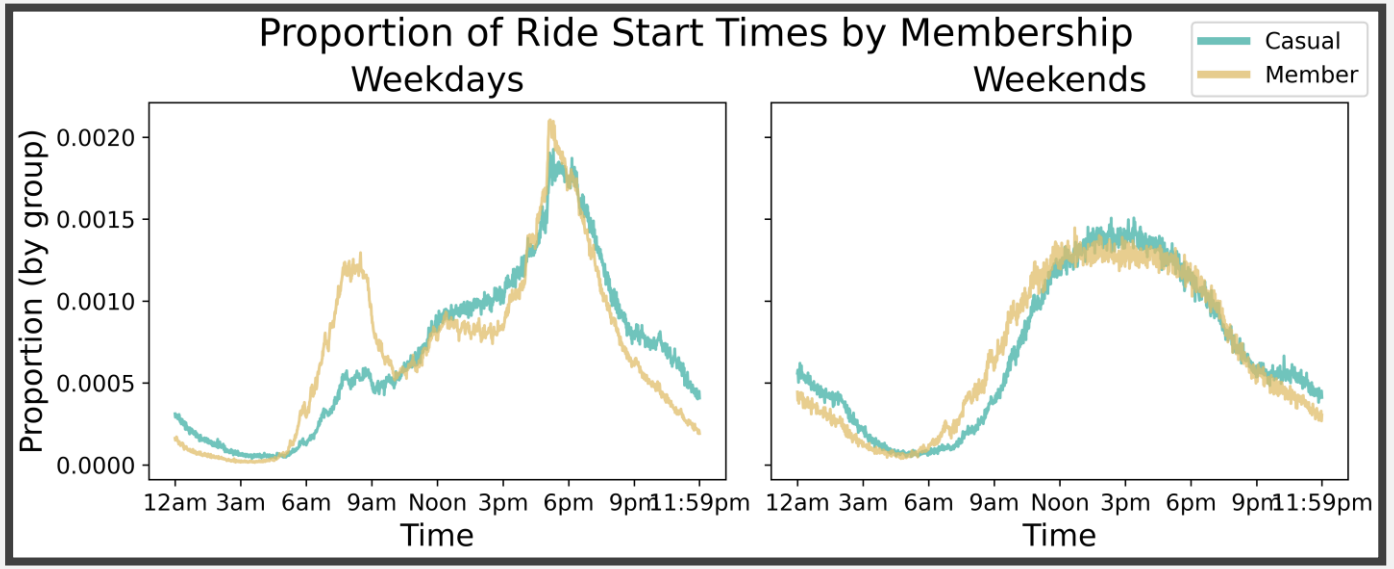
- Yearly scale over months
- **Weekly scale over days**
- Daily scale over hours/minutes



WHEN

There are 3 levels of "when":

- Yearly scale over months
- Weekly scale over days
- **Daily scale over hours/minutes**



ANALYSIS

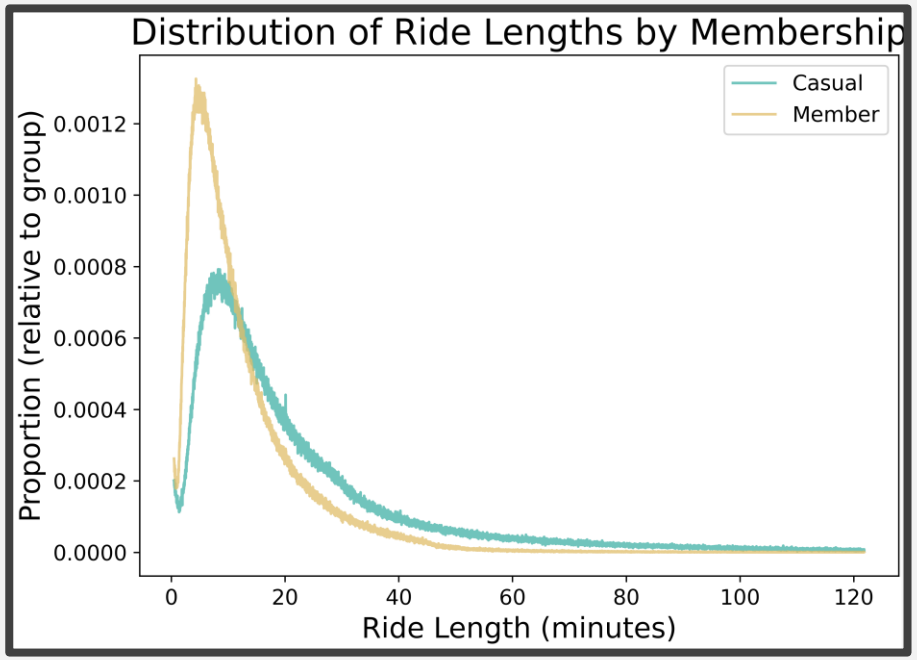
There are **four** key differences and/or patterns that were identified:

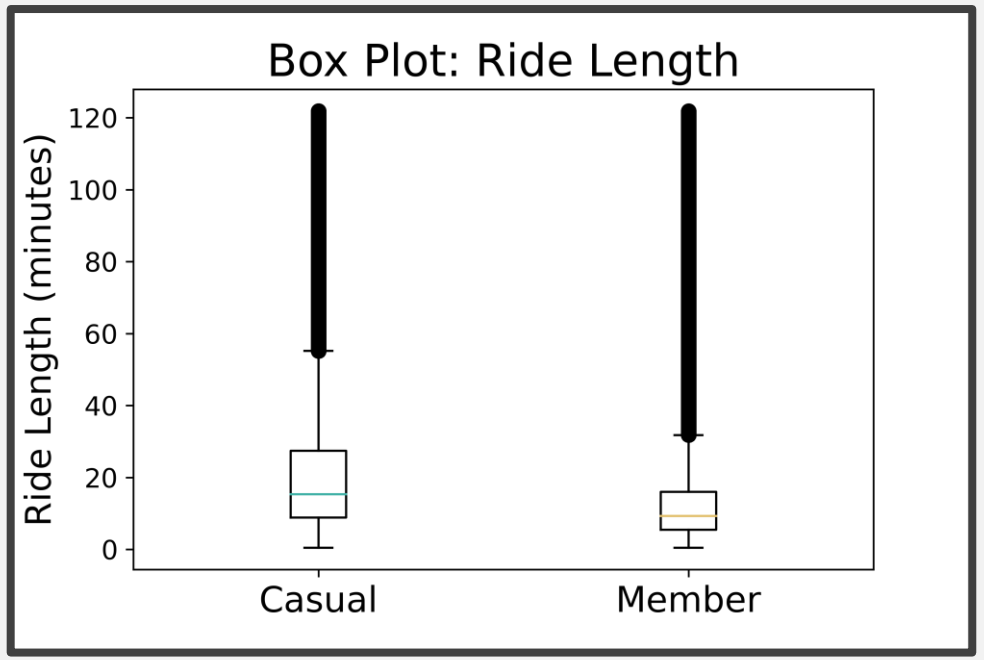
1. What
2. When
3. **How Long**
4. **Where**

HOW LONG

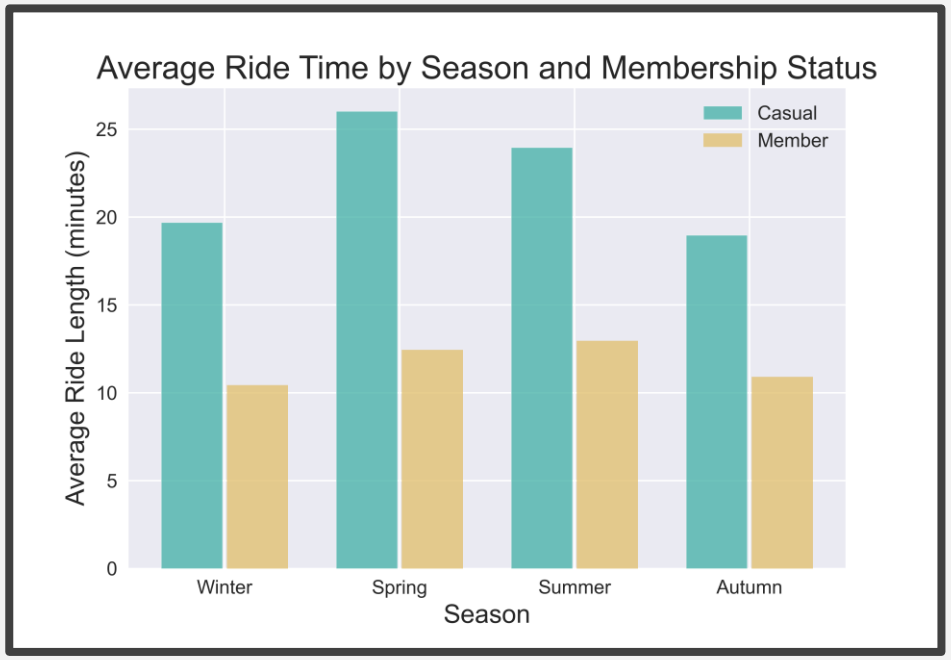
- The *ride_length* column was created by taking the difference between start and end times, kept in minutes.
- There are some large outliers in the data here (e.g., times well over 30 days) so 1% trimmed statistics were used
- The *ride_length* data was analyzed against multiple categorical variables

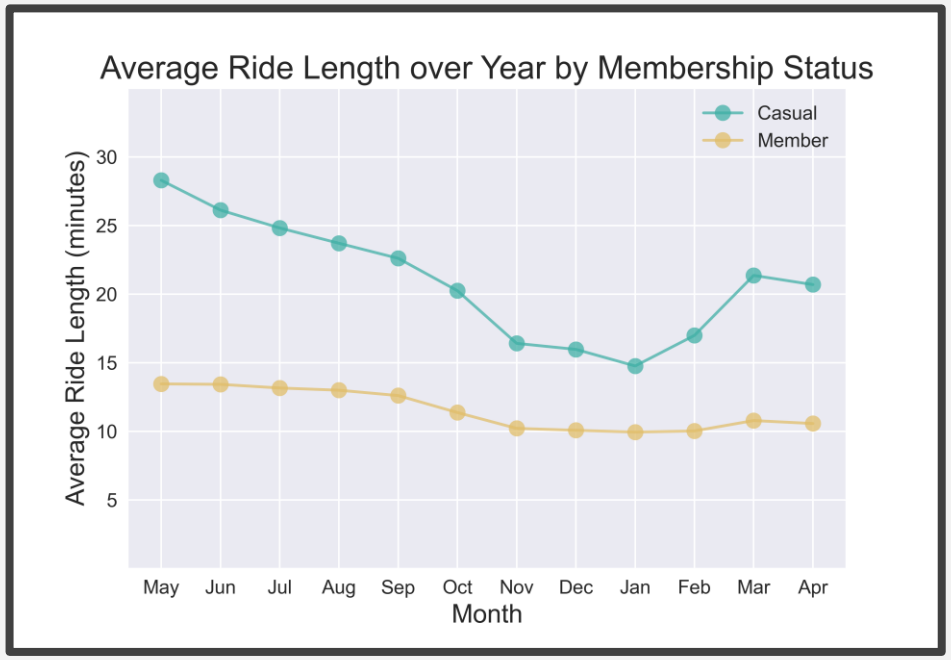
HOW LONG


Distribution of Ride Lengths by Membership
Casual
Member
Proportion (relative to group)
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
0
20
40
60
80
100
120
Ride Length (minutes)

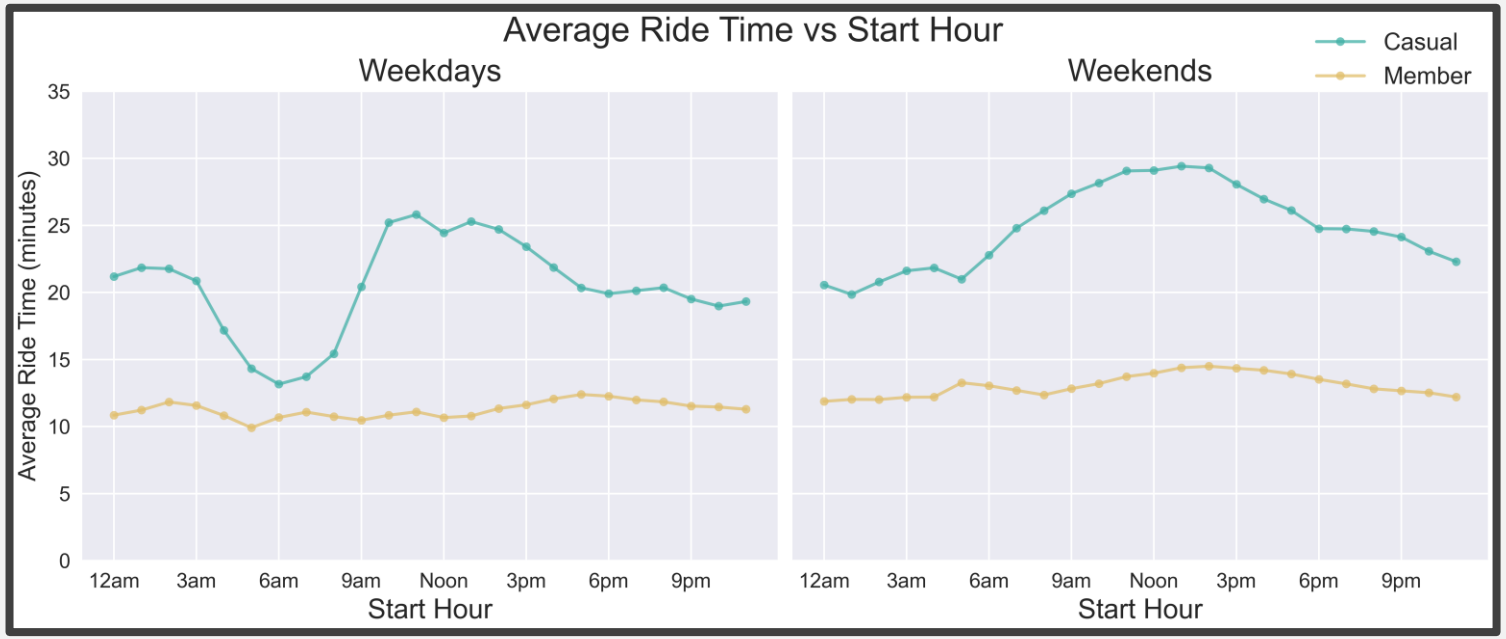

Box Plot: Ride Length
Ride Length (minutes)
120
100
80
60
40
20
0
Casual
Member

HOW LONG


Average Ride Time by Season and Membership Status
Casual
Member
Average Ride Length (minutes)
0
5
10
15
20
25
Winter
Spring
Summer
Autumn
Season


Average Ride Length over Year by Membership Status
Casual
Member
Average Ride Length (minutes)
5
10
15
20
25
30
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
Month

HOW LONG


Average Ride Time vs Start Hour
Weekdays
Weekends
Casual
Member
Average Ride Time (minutes)
0
5
10
15
20
25
30
35
12am
3am
6am
9am
Noon
3pm
6pm
9pm
Start Hour
Start Hour

ANALYSIS

There are **four** key differences and/or patterns that were identified:

1. What
2. When
3. How Long
4. **Where**

WHERE

What areas were riders in most often?

We analyzed which stations casual riders and members interacted with most frequently to inform possible physical marketing locations.

We used machine learning to separate stations into 3 areas:

- University of Chicago
- Navy Pier
- North Side

WHERE

To analyze the spread of station interactions, we create a distribution of the Euclidean Norms from the average location of all interactions to each station interaction. We choose standard deviation as a measure of spread.

The distribution for members has a standard deviation almost **4x** as large as the standard deviation for casual riders. Using different norms leads to various differences between the standard deviations.

In every case, the standard deviation of the members distribution is greater than that of the casuals → Casual Riders are more condensed around the center of the distribution (near Navy Pier)

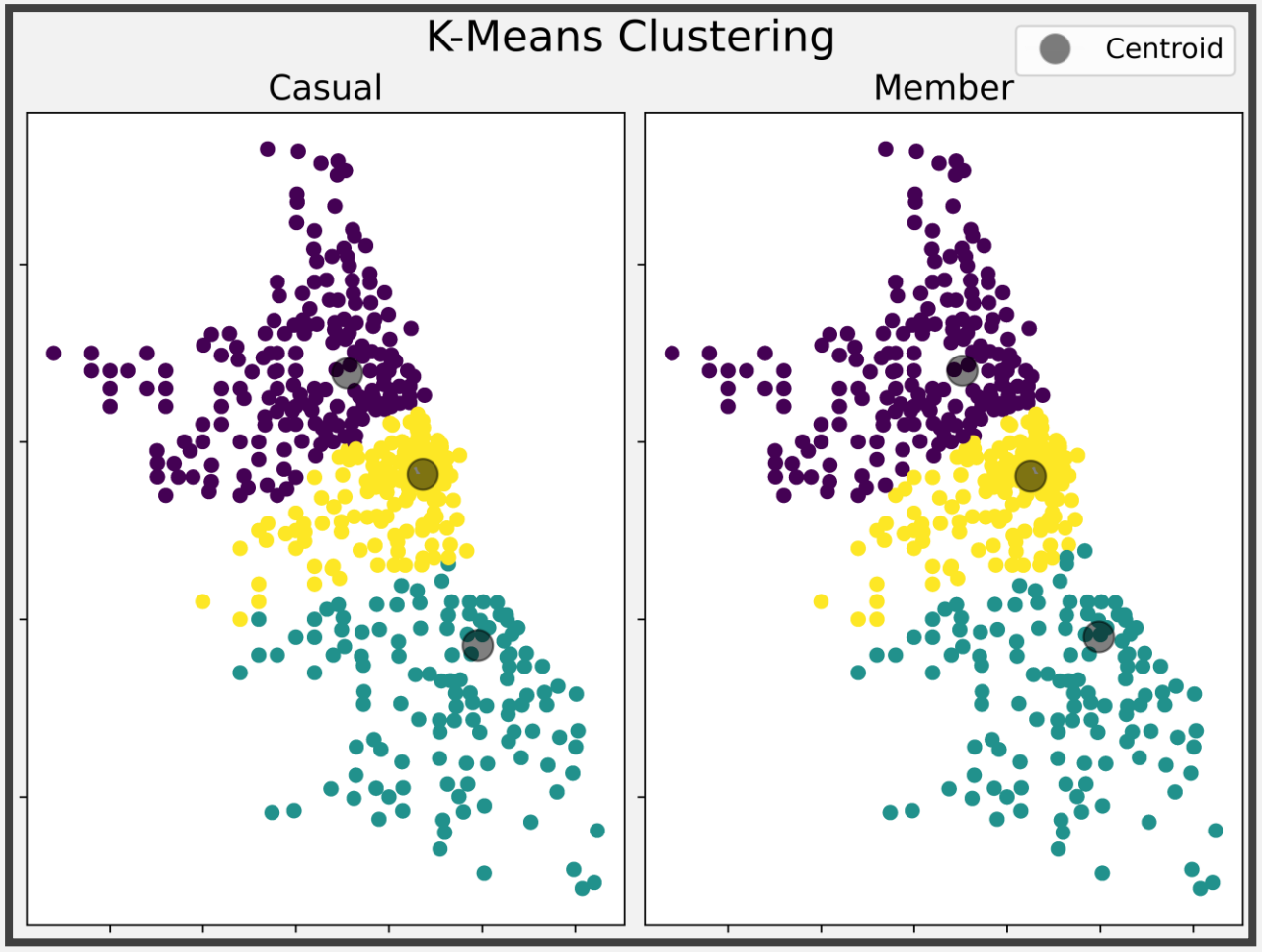
WHERE

Elbow method with the sum of squares was used to determine the optimal number of clusters

- 3 for casual riders and members

K-Means clustering algorithm grouped the stations similarly, but not identically.

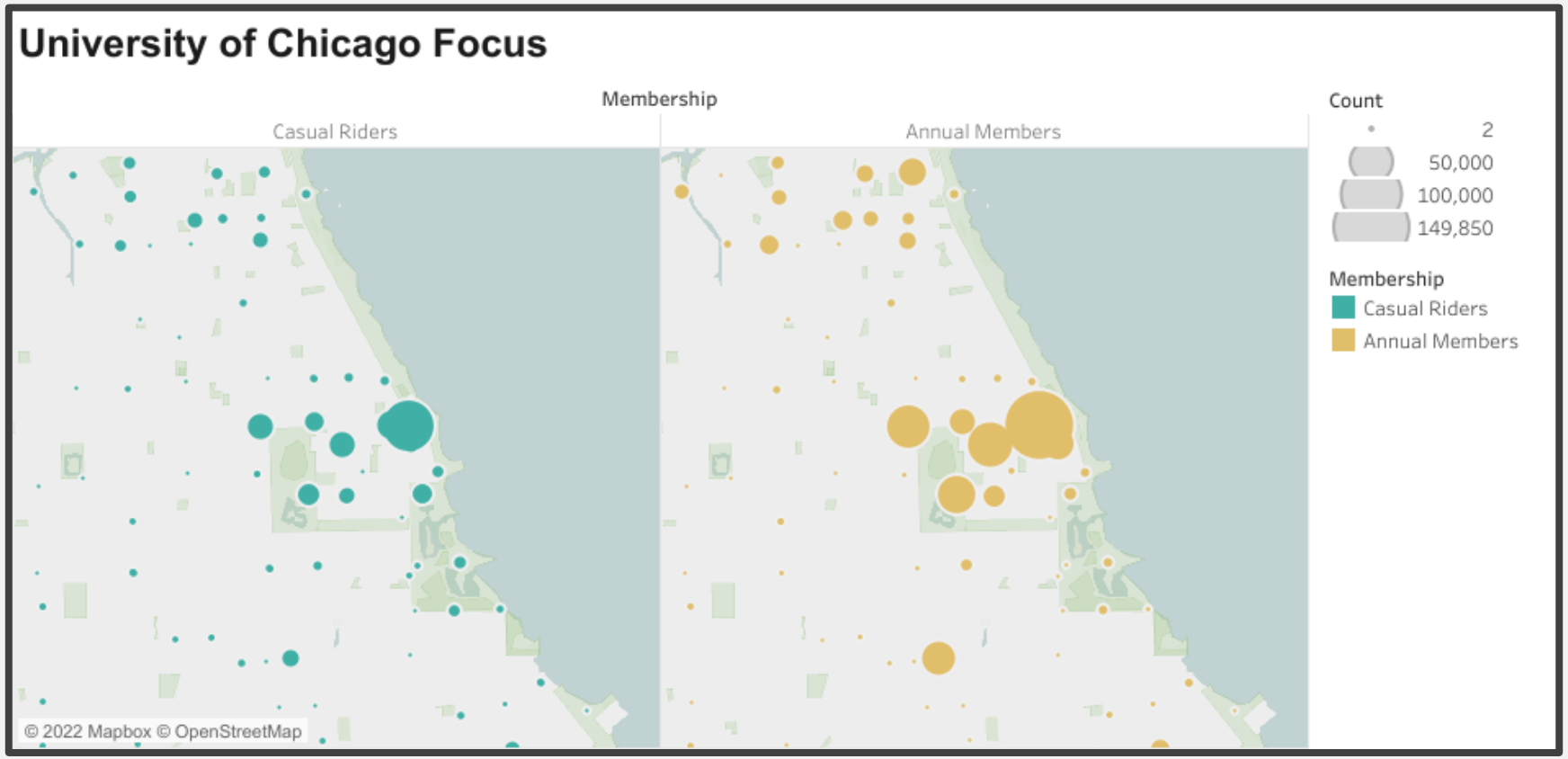
Member cluster sizes are more equivalent in size, matching the expected behavior from being more spread out



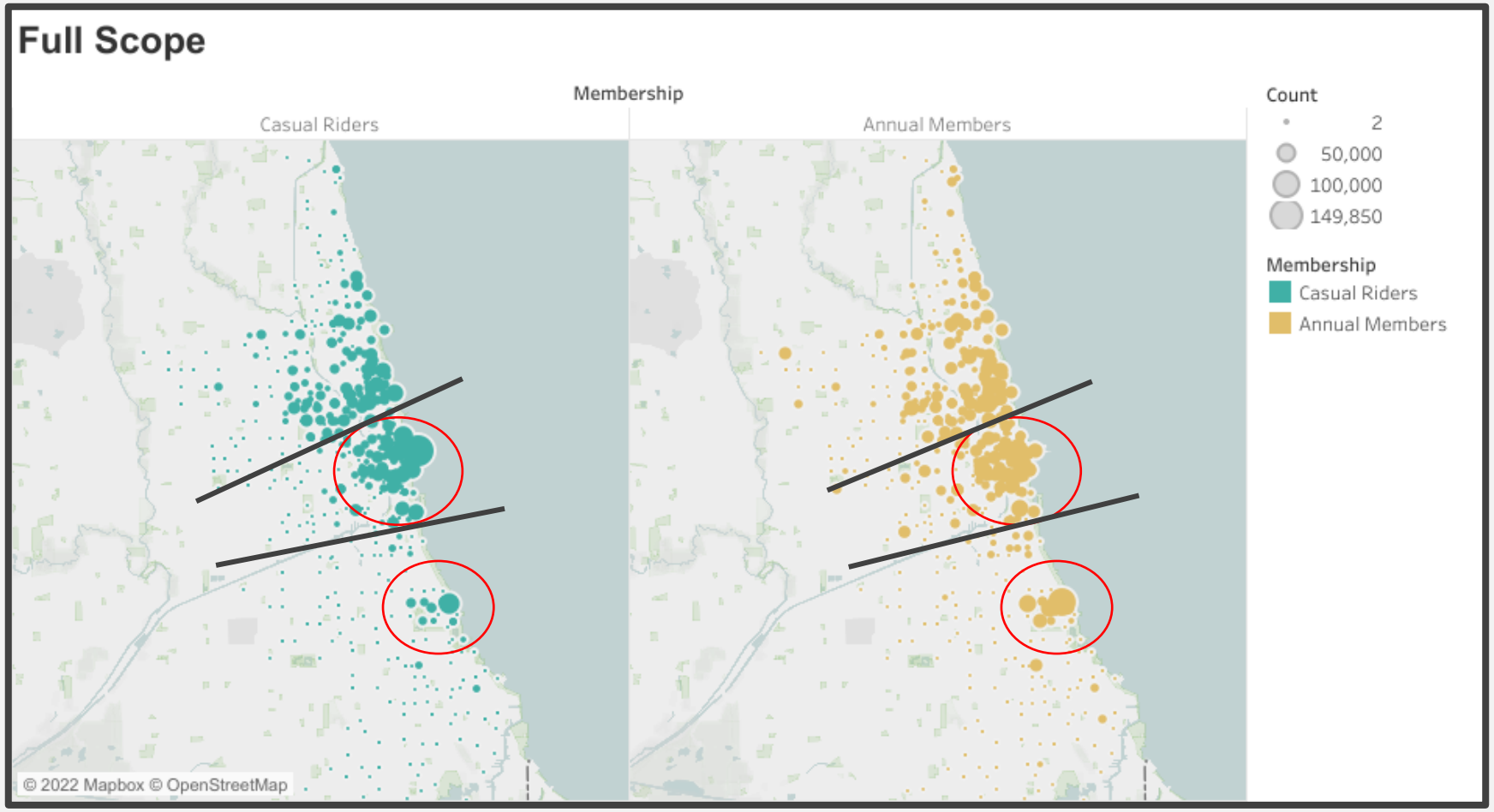
WHERE


Full Scope
Membership
Casual Riders
Annual Members
Count
2
50,000
100,000
149,850
Membership
Casual Riders
Annual Members
© 2022 Mapbox © OpenStreetMap

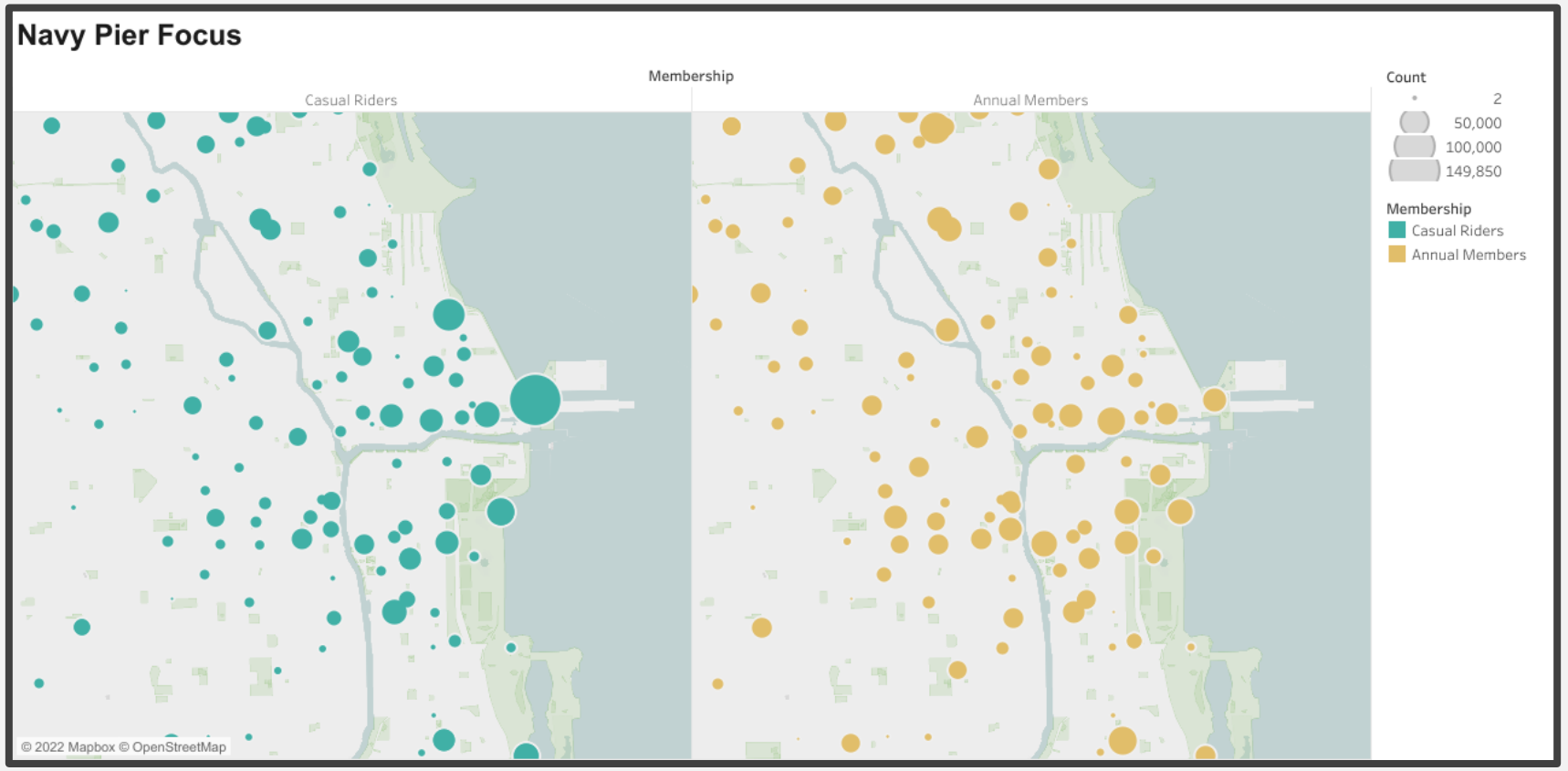
WHERE


University of Chicago Focus
Membership
Casual Riders
Annual Members
Count
2
50,000
100,000
149,850
Membership
Casual Riders
Annual Members
© 2022 Mapbox © OpenStreetMap

WHERE


Full Scope
Membership
Casual Riders
Annual Members
Count
2
50,000
100,000
149,850
Membership
Casual Riders
Annual Members
© 2022 Mapbox © OpenStreetMap

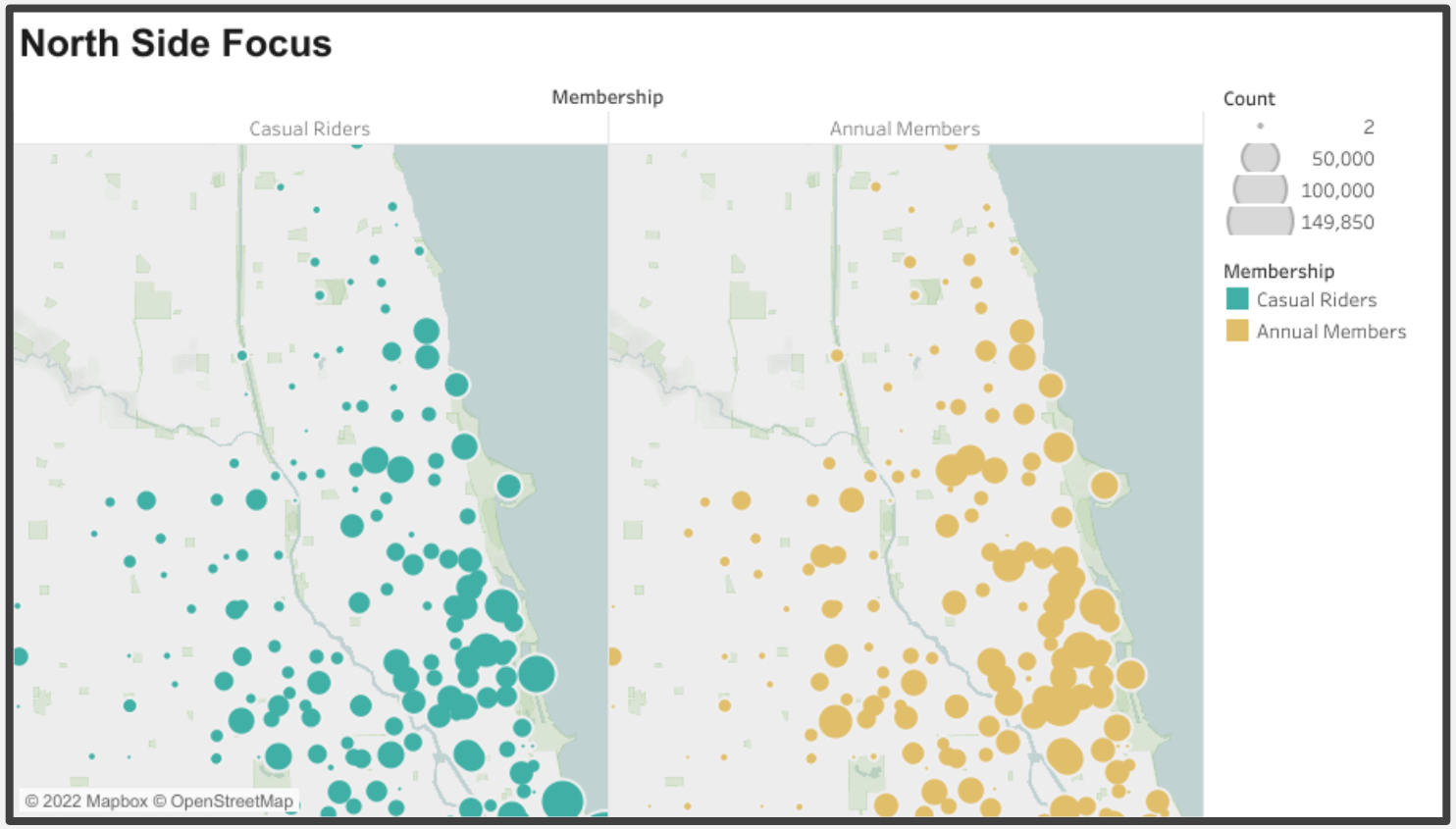
WHERE


Navy Pier Focus
Membership
Casual Riders
Annual Members
Count
2
50,000
100,000
149,850
Membership
Casual Riders
Annual Members
© 2022 Mapbox © OpenStreetMap

WHERE


Full Scope
Membership
Casual Riders
Annual Members
Count
2
50,000
100,000
149,850
Membership
Casual Riders
Annual Members
© 2022 Mapbox © OpenStreetMap

WHERE


North Side Focus
Membership
Casual Riders
Annual Members
Count
2
50,000
100,000
149,850
Membership
Casual Riders
Annual Members
© 2022 Mapbox © OpenStreetMap