CYCLISTIC'S BIKE SHARING SERVICE

- DATA ANALYSIS FINAL REPORT -

By Luca Albertini

OUTLINE

Main Question

Ask - Main Goals

Prepare - Data Loading & Description

Cleaning - Data Cleaning & Comments

Analyze - Data Analysis & expectations

Share - Data Visualization & Insights

Act - Data-driven Recommendations

Conclusions

MAIN QUESTION

How do annual members and casual ride Cyclisitc's bikes differently?

ASK - MAIN GOALS

Cyclistic's Marketing dept. believes that maximizing the number of annual members is a key component to ensure the business' future growth.
As such, the marketing strategy must aim to convert casual (i.e. sporadic) users into annual (i.e. with a subscription) ones rather than acquiring new customers. Determining a marketing strategy that will finally lead to these results is the business task that the firm must accomplish.

In this project's phase, the main goal is to determine the key differences between casual and annual members' use of Cyclistic's bikes. As such, the analysis should be driven by two related questions:

- In terms of bike rides, do casuals and members differ? How?
- Are there some common elements between casuals' and members' usage of Cyclistic service?

The first question aims to understand what conditions can come while riding as a casual that can "force" a casual rider to become a member; the second will help understand what services can be implemented to make an annual subscription even more appealing.
Insights coming from this analysis will play a crucial role in increasing the conversion (from casual to annual) rate and, as a consequence, the amount of annual members.

All the results of this analysis are going to be reported to the head of the marketing dept. and to the executive team.

PREPARE - DATA LOADING & DESCRIPTION

Data is available thanks to Motivate International Inc. Data are publicly available online. A local copy has been created to be referred to as the starting point for the analysis and the cleaning phase.
Data registers bike rides in Chicago done by using the service of docking bikes provided by Cyclistic.

Data frames were divided by quarters, and they were all joined together to create a year-long of bike ride information. In detail, The period taken into consideration goes from the 2nd quarter of 2019 (Q2_2019) to quarter 1 of 2020 (Q1_2020). Note that q1 2020 is the last data available.
To keep data anonymous, it is organized based on the ride's identification number, and it carries various information about bike use and users (such as *type, starting and ending date and time, starting and ending station, type of membership, etc*).
Data has been sorted by start date and time to get a chronological order. At this stage, no filter was applied.
Delving into data, it seems that it does not contain biases. Moreover, it is reliable (well-known agency), comprehensive (covers a lot of different aspects), and current (most recent).
There may be issues with its originality (it is provided by a third-party agency) and its citation (seems not to be cited anywhere, yet being a niche company data this could be plausible).
There are two major limitations:

- Data is collected by a third-party agency, even though it is a reliable and well-known agency.
- Data is 2 years old, thus it could be that trends shown in this analysis may have changed a little (yet, it is the last data about an entire quarter available).

CLEANING - DATA CLEANING AND COMMENTS

The entire cleaning phase was performed both in Excel and Python environments. This was due to a big weight dataset. All 4 different dfs have been cleaned, joined, and then manipulated. In the .py files, the cleaning procedure is within the "cleaning" and "processing" sections.
The final df, named "***tripsclean.csv***" has **3,872,240** observations.

All the changes made to the original df have been documented within the document named "***ChangeLog_52224.xlsx***". In it, every change has been registered along with the date and a brief description. The Excel file contains a short description for the added columns too.

Some important comments about the results of the cleaning phase:

- Rides that start in one station and end in another and are less than 60 seconds occurred almost always from the same two stations. A brief geographical check highlighted that these two stations are close, thus it is reasonable that those kinds of rides took less than 60 seconds.
- Rides longer than 60 seconds and where the undocking station is the same as the docking one have been classified as "*user_error*", i.e. a user wrongly undocked the bike, did not use it and docked it almost immediately. I opted for removing these rides from the analysis since, as errors, they are not that relevant in determining the core differences in customers' behaviors.
- There were rides whose starting time was registered after the ending time. This is impossible and could be a collecting or software mistake. As such, all the observations presenting this issue were dropped.
- If the starting and/or ending stations were "HQ QR", then the observation has been dropped. Indeed these display bikes that were taken by the headquarters. As a consequence, do not provide any information about users.

ANALYZE - DATA ANALYSIS & EXPECTATIONS

To get insight from data, data have been analyzed through different computations.
The first step was to get descriptive statistics (*mean, median, standard deviation, max, min*) for the ride length.
Then conditional statistics were computed. As such, descriptive statistics on ride length have been calculated based on:

- Membership type
- Quarter
- Month
- Day of the week
- Date
- Undocking and docking stations

Finally, it has been performed a rush hour analysis. For this purpose, rush hours are between 7.30 and 9.30 am and 5.00 and 8.00 pm Monday to Saturday (Sundays are excluded).

- EXPECTATIONS

Before performing an analysis, it is good practice to sketch some ideas about what to expect. Nevertheless, it is important to remember that these are just expectations that may or may not be confirmed by the analysis. Data may show the opposite of those. If this is the case, it will be interesting to discover why.

Furthermore, these expectations must not lead to user bias and must be treated as ideas not as facts that data should confirm.
In details, expectations are:

1. Casuals' average ride time may be longer than the members' one. Members do not pay for every unlock, thus they could/would unlock the bike even for a short trip. Not-members will probably think twice about that.
2. Correlated to point 1, it is likely that most of the short rides are done by members.
3. In the summer and spring periods, the number of rides should increase.
4. If the number of rides is higher during working days, this may mean that the service is used more to commute rather than for leisure trips. Working days may be the time when annuals ride more.
5. It is reasonable to expect the average ride length to be higher during weekends since people (independent of membership type) may have more time to bike around. These may be the days when casuals ride more.
6. If the service is used for working purposes, then rides in rush hours should be a significant amount.

- SUMMARY & MAJOR TRENDS

It may be useful to take a glance at the major trends that have been spotted after the data analysis:

- **MEMBER ANALYSIS**
  1. Members do more rides during the year than casuals
  2. Casuals' average ride length is greater than members' average one
  3. Casuals have the highest bike usage in terms of total time spent riding Cyclistic bikes.
  4. A casual user performed the longest ride of the whole year while an annual user performed the shortest one (little more than 30 seconds)
- **MONTHLY ANALYSIS**
  1. There is a peak of rides within June, July and August both for casuals and members.
  2. For most of the year, the total riding time is greater for casuals, however members display a longer total riding time in the coldest months (November, December, January).
  3. Independently of the month, on average casuals' rides last more than members' ones.
- **SINGLE-DAY ANALYSIS**
  1. The longest rides are done on Sunday (longest average length) and Saturday
  2. Tuesday experienced the highest number of rides; while the days with fewer trips are Sunday and Saturday.
  3. The day that saw the highest use of bikes in terms of total ride time is Saturday.
- **DAY & MEMBER ANALYSIS**
  1. Members prefer to ride during week-days, while casuals prefer using cyclists bikes during weekends
  2. During weekdays, there is not a lot of difference in total ride time between members and casuals, while during the weekend casuals total ride time skyrockets.
  3. Friday is when casuals have the highest ride average duration; for members is Saturday.
  4. Generally, members tend to have a lower average ride duration than casuals.
  5. For casuals, trips are almost always longer than 60 sec. It is not so for members.
- **RUSH HOUR RIDES ANALYSIS**
  1. Members ride more in rush hours than casuals during every working day
  2. The working day that sees more rush hour trips by members is Tuesday while by casuals is Saturday.
- **STATION ANALYSIS**
  1. The most used stations (both docking and undocking) are located in the city center, while the least used ones are located in the suburbs.
  2. Most used stations by members are in the city center, while the least used are in the southern suburbs. Generally speaking, this is true even for casuals.

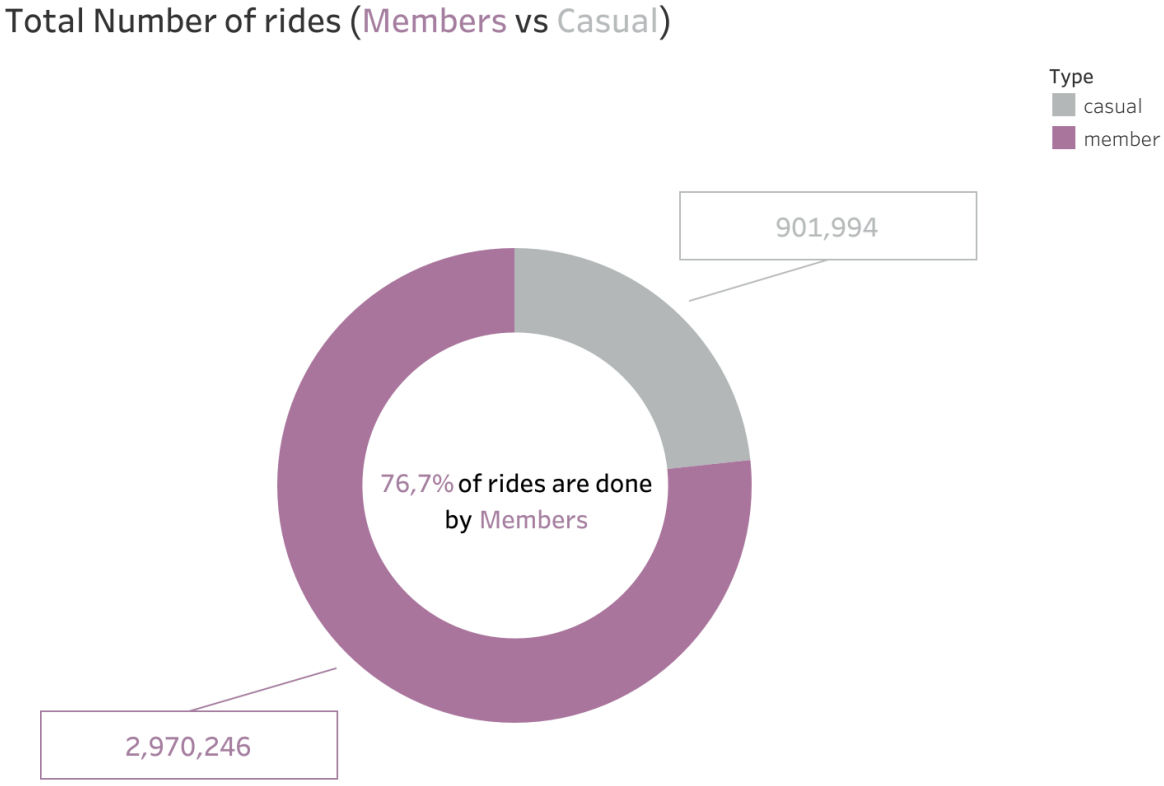
3. Just 1 station is within the 15 most used stations by members and the 15 most used by casuals. On the contrary, 8 stations figure both within the least used by members and the least used by casuals.

SHARE - VISUALIZATION & INSIGHTS

How do annual members and casuals ride differently?
Several major insights come from data that can help tailor a data-driven marketing strategy.

General Analysis

First and foremost, it is clear that casuals ride more once the bike is undocked while members undock the bike more often. The count of trips within the year and the average ride length show this important difference among users.



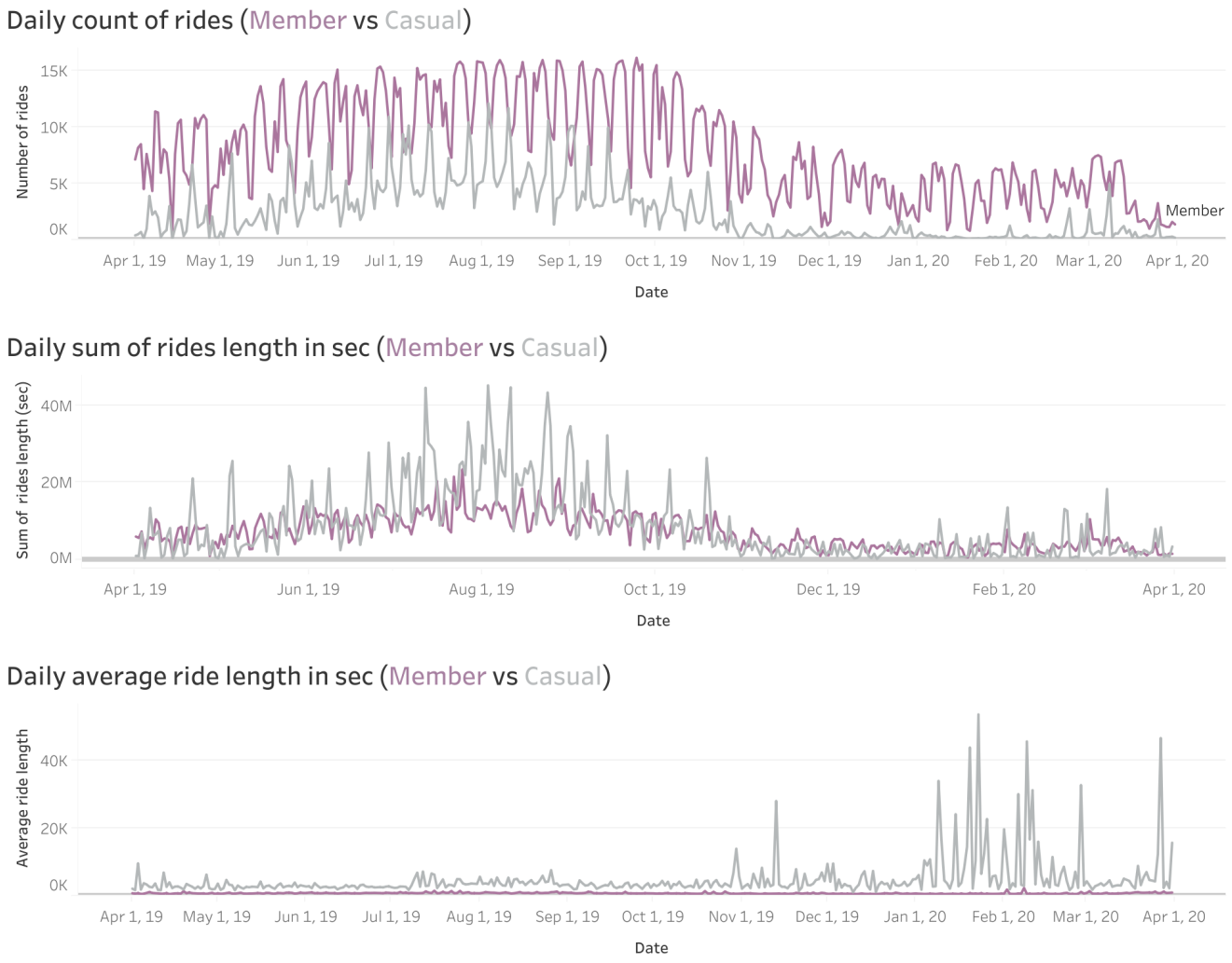
Picture 1. Total number of rides within the year

Picture 2. Average ride length (left) and total ride length (right) in seconds

As such it does not come as a surprise that the total time spent riding a bike is greater for casuals. What is interesting is what contributes mostly to such a high number.

Daily and Monthly Analysis

Considering daily rides gives another interesting point of view. Indeed, disentangling daily habits helps develop the right market strategy.



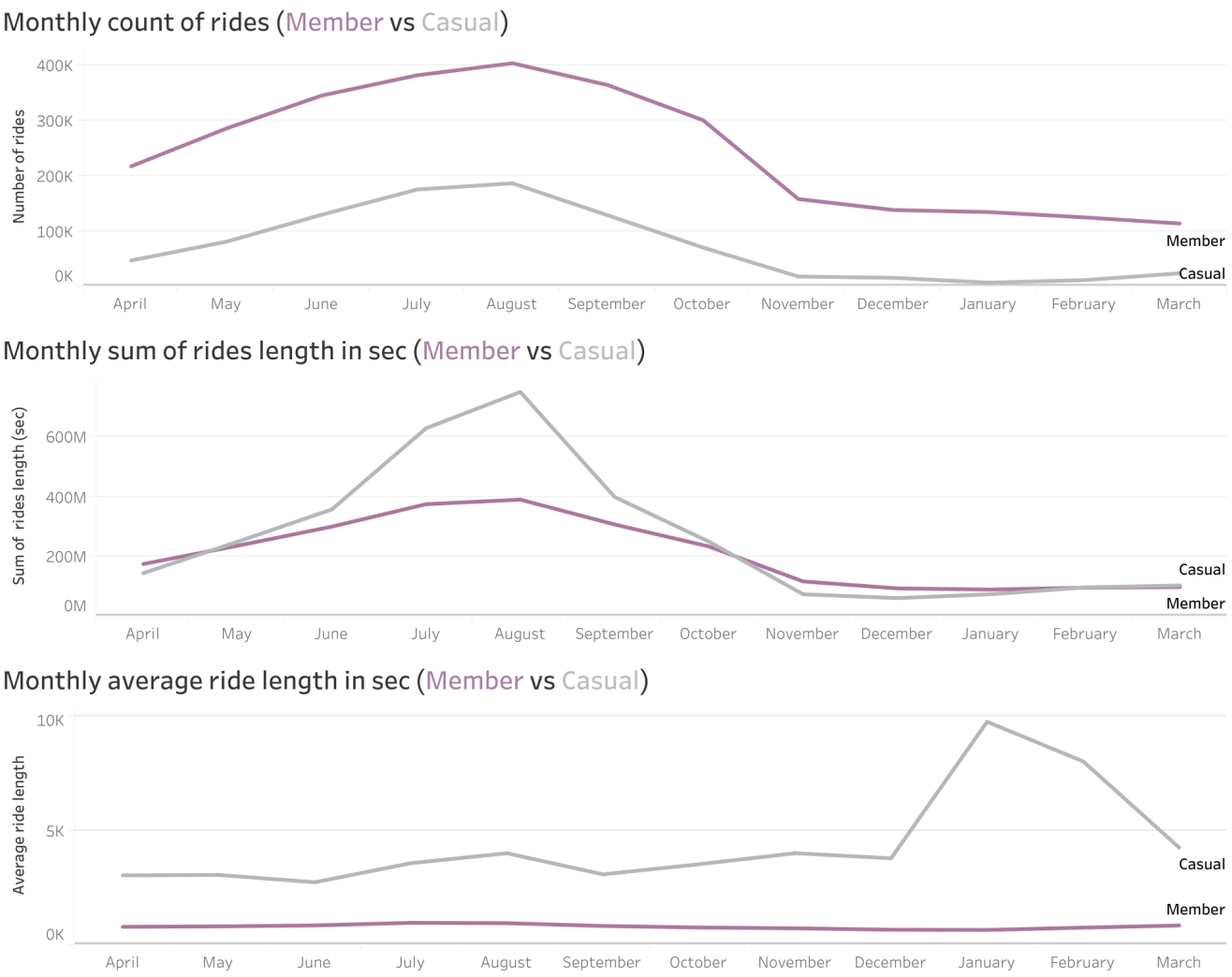
Picture 3. Count (first), sum (second) and average (third) ride length for each day from April 1 2019 to April 1, 2020.

These charts display two interesting things:

1. There are just a few days (less than 10) where members perform fewer rides than casuals
2. There is not a single day during the year where members' average ride length is greater than casuals one. Moreover, it is worth noting how casuals average ride length shows greater variability, especially in the first month of 2020. This may be due to the continuous weather changes during the day in the first months of the year.

There is not a user type that constantly rides more than the other. Interestingly, casuals have a greater total ride length during summer months while members have a bigger total ride time during the coldest month. Furthermore, what drives up members' total time spent riding is the amount of rides they take, while for casuals is either the amount of rides (during the summer) or the average ride length (at the beginning of the year).

Zooming out and displaying monthly data swift the focus on general trends by reducing volatility.

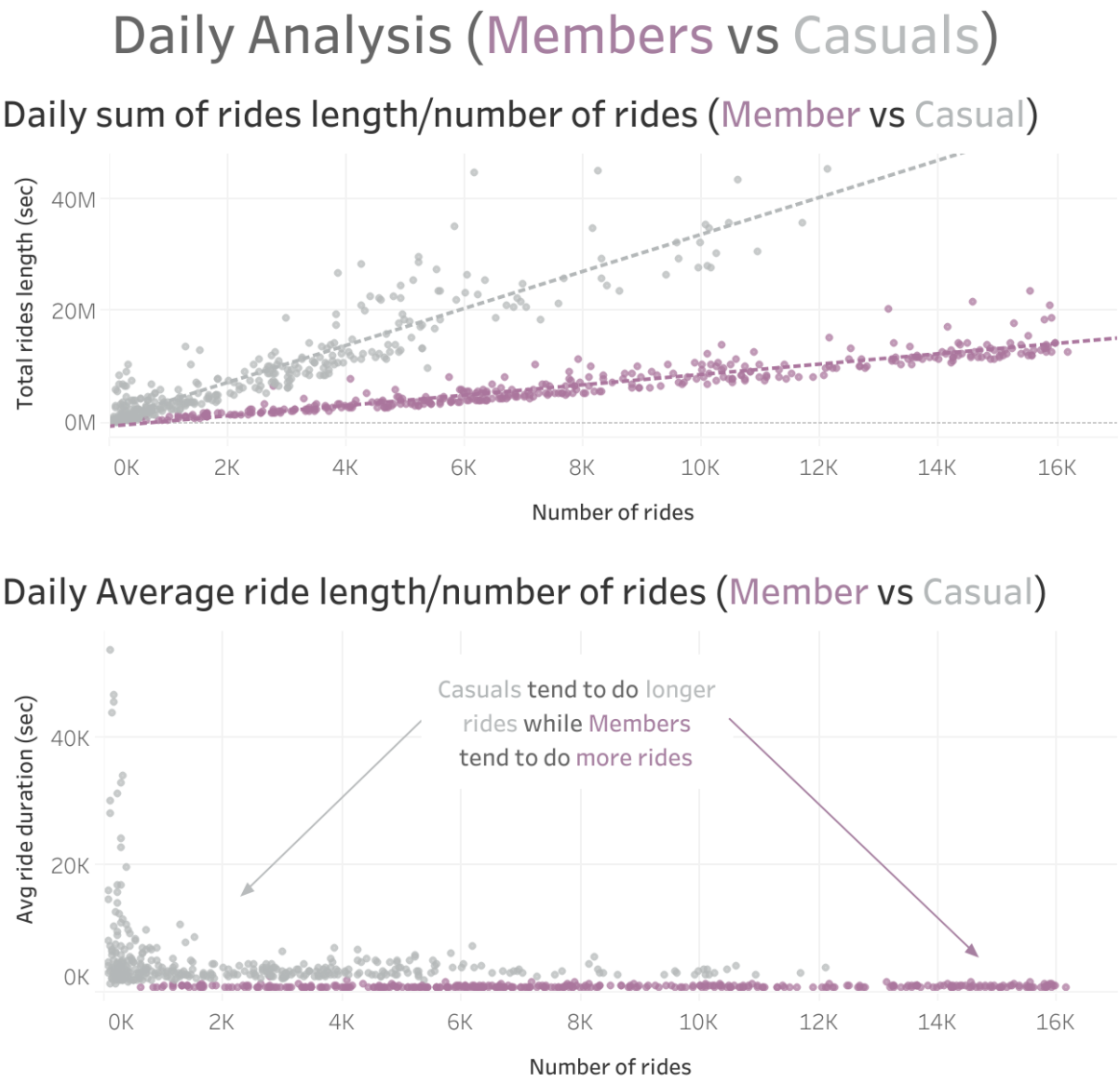


Picture 4. Count (first), sum (second) and average (third) ride length for each month from April 1 2019 to April 1 2020.

Casuals and annuals have a couple of common habits. Firstly, the ride count has an upward trend from April to August and a downward one from August to November. By November this tendency weakens. Secondly, the peak of total time spent riding is during the hottest month of the year, especially in August.

On the other side, data confirms the intuitions above: members do more rides, while casuals have a higher average ride length. Interestingly, for the sum of time spent riding, there is not a user type prevailing on the other. Indeed, during the hottest month, casuals spend more time riding, while during the coolest months is the opposite.

As evident, the relation between the number of rides during the day and the total or average time spent riding a bike is quite fundamental to understanding riding habits. Let's consider a scatterplot displaying them.

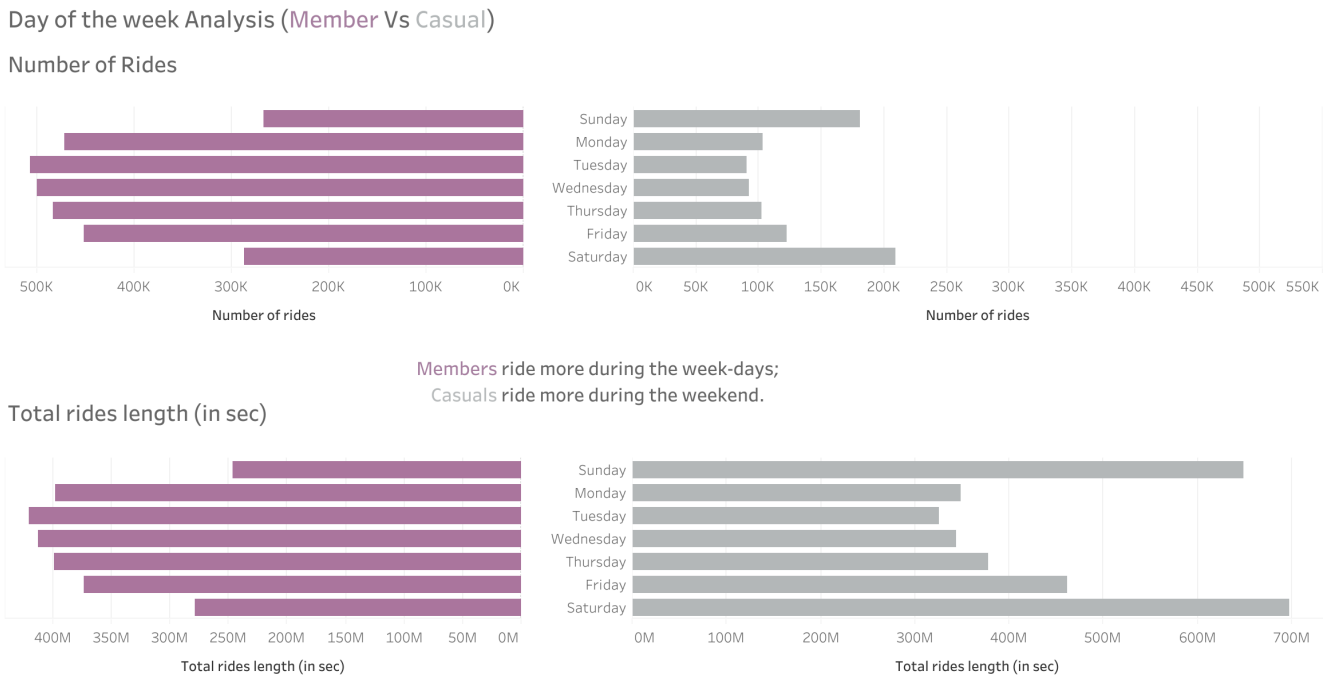


Picture 5. Sum of ride length against number of rides (above) and Average ride length against number of rides (below)

The first chart displays how the relation between total ride times/ride count is steeper for casuals. This means that for each additional ride, the total ride length will increase more than the one of the members. This is due to casuals' greater average ride time. On the other hand, what contributes more to increase members' time spent on a bike is how often they undock a bike.
The second chart confirms the difference. Casuals cluster at the beginning of the x-axis (i.e. lowest number of rides) rising on the vertical one (average ride time); while members distribute evenly on the x-axis without significantly increasing their average ride time (y-axis).

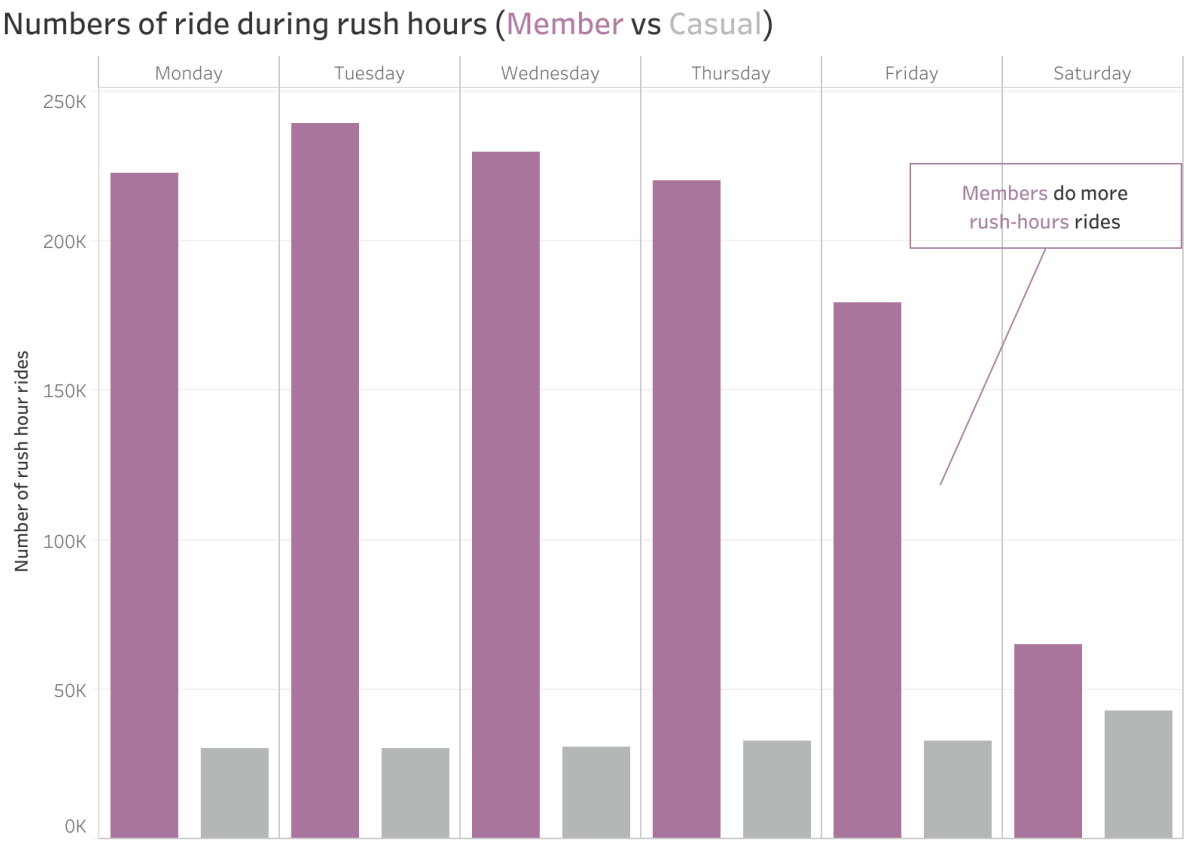
Day of the Week and Rush Hours Analysis

To delve more into profundity, a day-of-the-week analysis is essential. Indeed, knowing on which days members and casuals ride more is useful to understand their rides' purposes.



Picture 6. Count (above) and sum (below) of rides based on day of the week

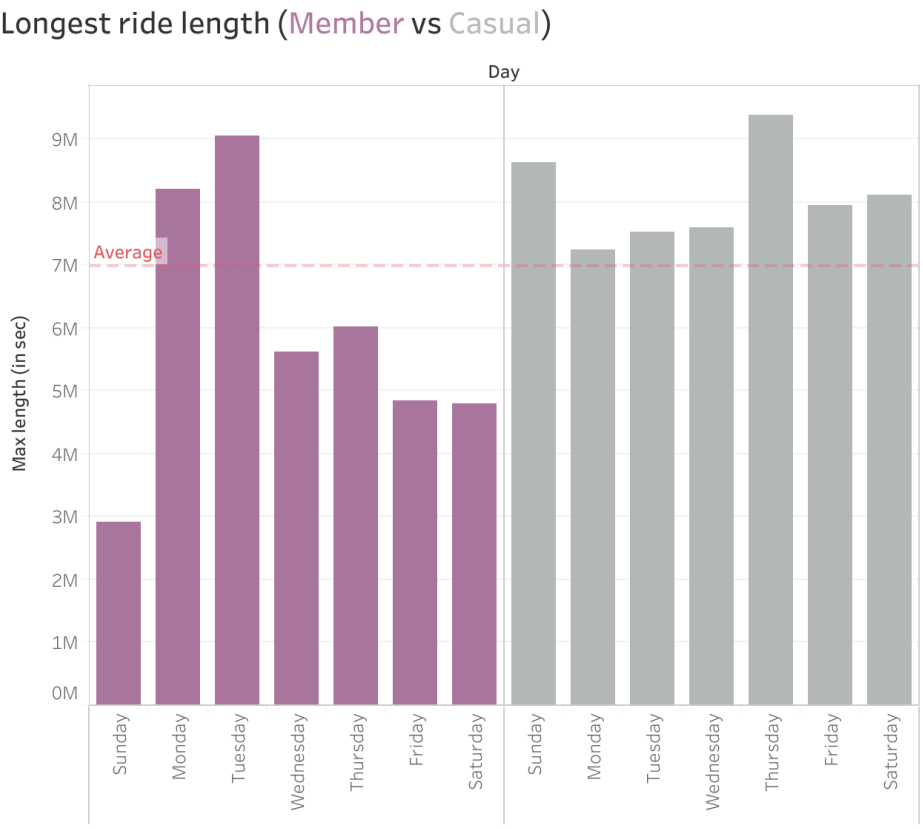
The charts show how annual members use more cyclistics' bikes during the weekdays, while casuals prefer to ride during the weekend. This may suggest users have different purposes: casuals use cyclistic services during leisure time and for leisure purposes; while members ride the docked bike mainly to commute and/or move during working days.
To double-check if this is the case, let's consider rides during rush hours. Again, "rush hours" are those between 7.30 and 9.30 am and 5.00 and 8.00 pm Monday to Saturday. These are the time ranges when people go to work or go back home. Note that Saturday has been included since there is a non-negligible amount of people that work on Saturday too.



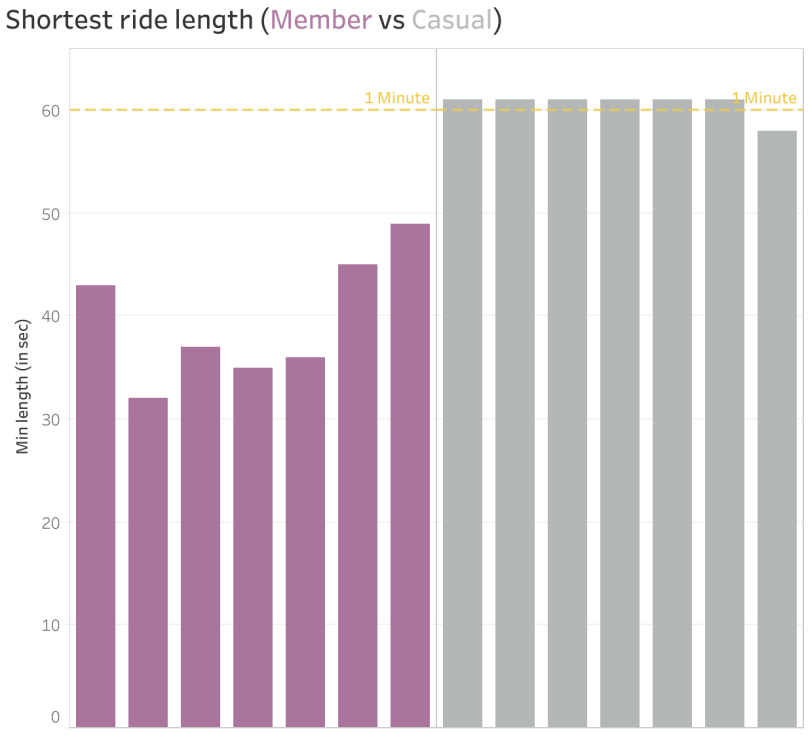
Picture 7. Count of rides during rush hours.

As evident, annual members use the cyclistics' bikes to commute to and from work more than casuals. Indeed, during the week members' rush hour rides count is more than 4-time the one of casuals. Interestingly, the day that sees the lowest margin between the counts is Saturday, the only weekend day.
Altogether, these considerations confirm that members prefer using the bikes for commuting purposes or to perform day-to-day tasks, whilst casuals prefer riding in their leisure time, maybe to explore the borough of Chicago.

Minimum and maximum ride lengths confirm all of these insights. Let's take a look.



Picture 8. Maximum ride length.

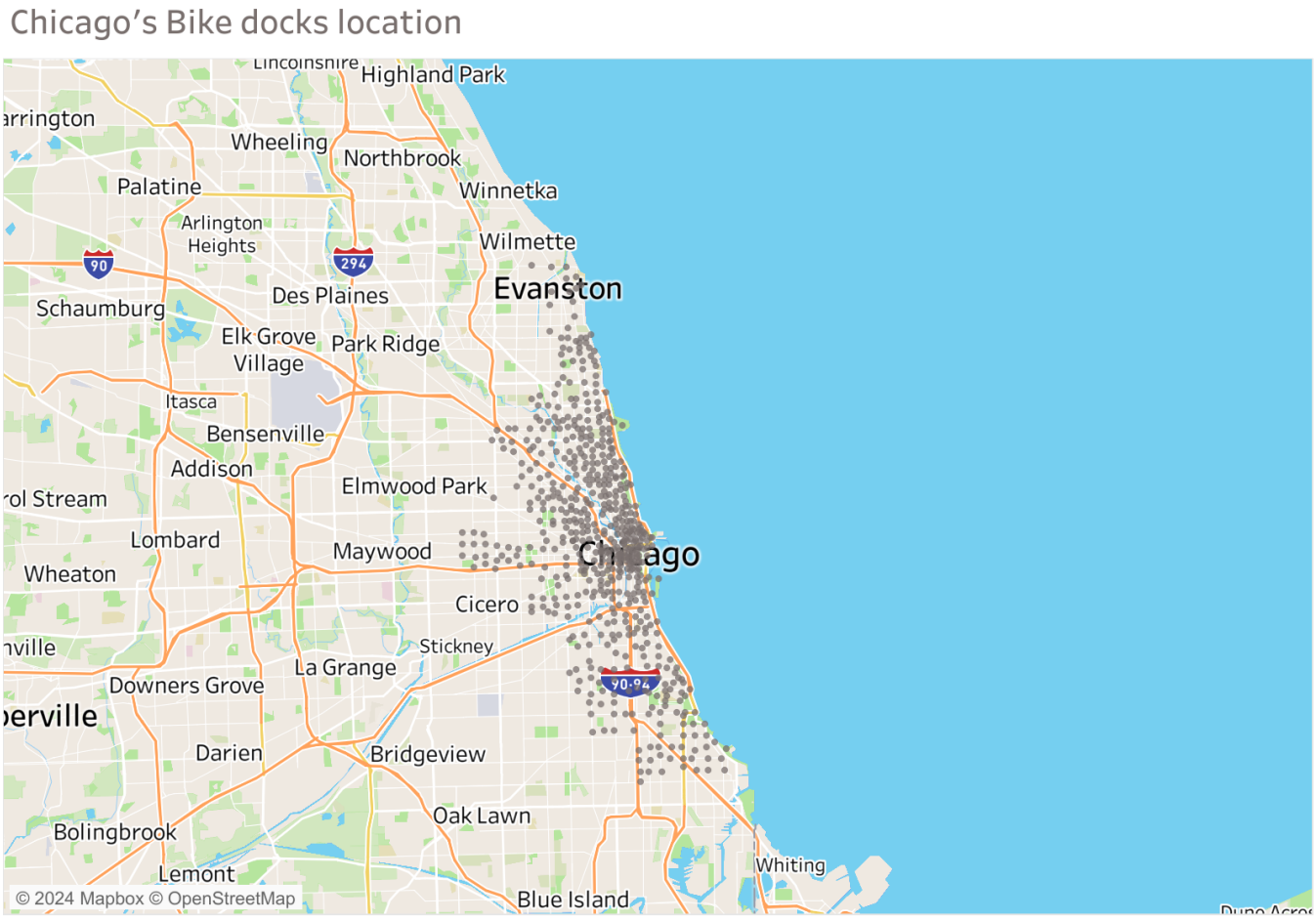


Picture 9. Minimum ride length.

Two things are worth noting. Primarily, casuals have a maximum ride length that is always above the average. Secondly, casuals do not undock bikes for rides shorter than 60 seconds (except on Saturdays). This suggests that the price for undocking the bike is a barrier for casuals: they will not bear that cost for short rides. Instead, Cyclistics membership cancels the undocking cost, as such members do not mind using the bike even for brief rides.

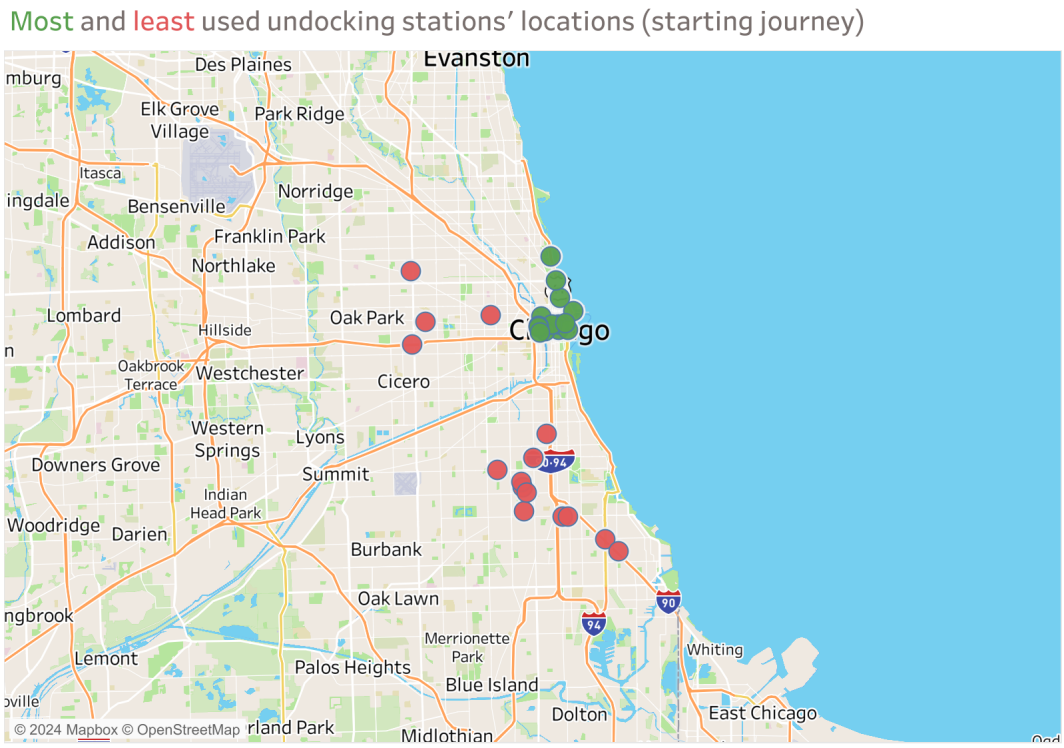
Docking and undocking stations analysis

A station-centered analysis helps both to understand user behaviors and to be prepared for possible future unexpected downturns.
Cyclistic did a good job in placing its stations all around the city. They cover not only the Chicago city center but also the suburbs, as the stations' map displays.

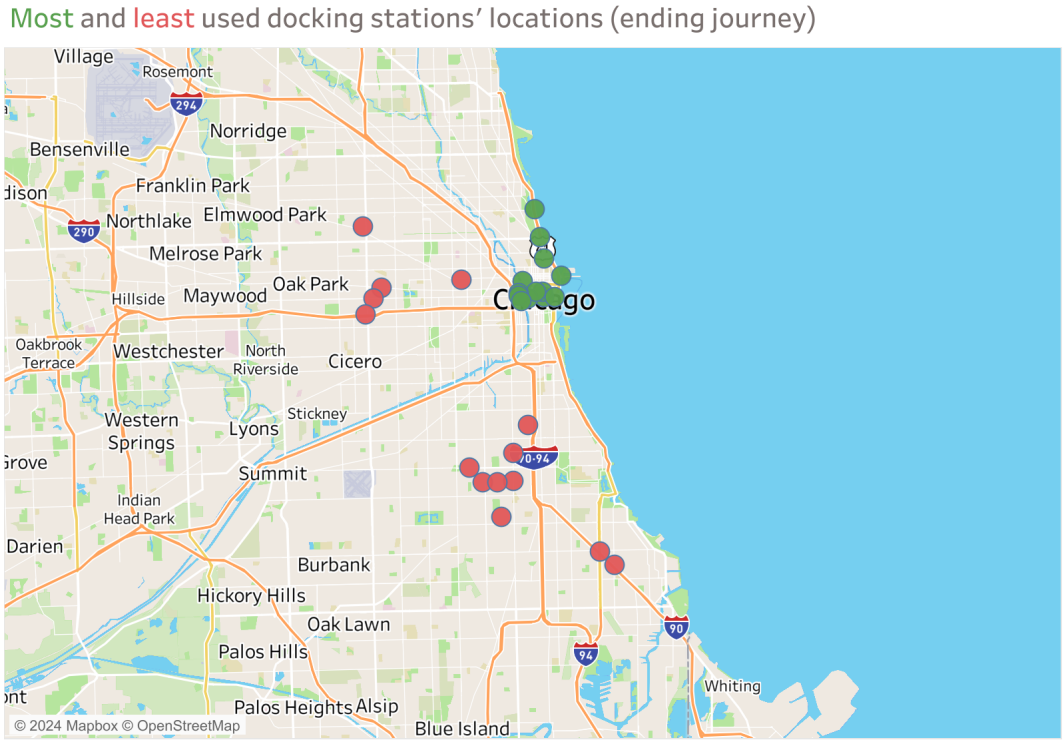


Picture 10. Cyclistics' stations locations in Chicago.

The map shows this service covers a wide geographical area. At this point, it may be useful to understand if some stations are used more/less than others. Let's consider that.



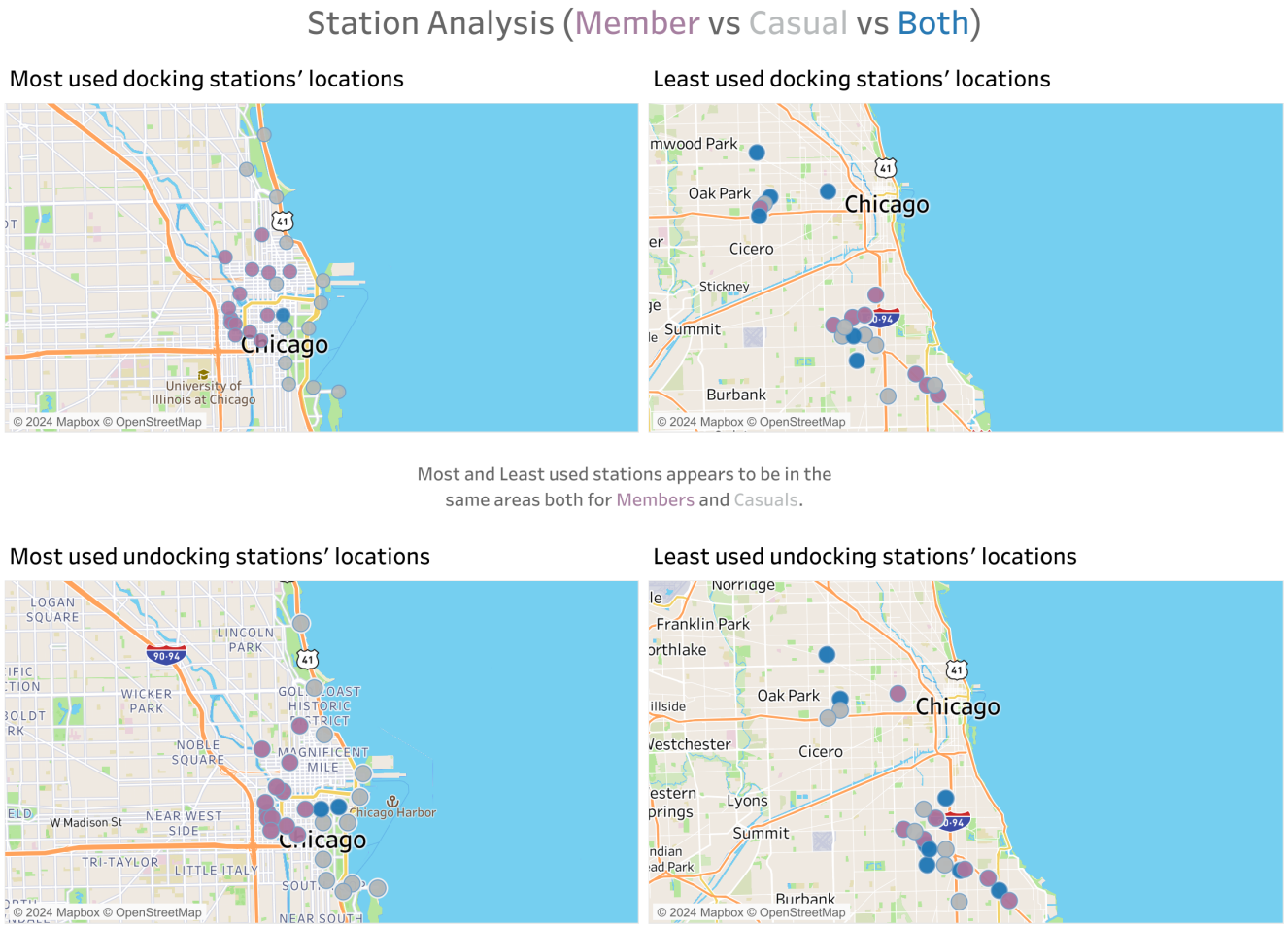
Picture 10. Most and least used undocking (i.e starting) stations



Picture 11. Most and least used docking (i.e ending) stations.

Displaying these two maps together allows to inquire about two different trends.
Considering docking altogether with undocking stations helps in comprehending if users have a preferred riding direction. If, for example, the most used docking stations are in the center while the most used undocking are in the periphery, then it is evident that users ride to go toward the city center. This is not the case. Most and least used stations are the same both for docking and undocking. This may be due to the higher concentration of people, commercial and working places and residential buildings in the city center.
However, what these maps are useful for is for managing purposes, especially during periods of financial distress. When, and if, Cyclistic experiences periods of economic downturn, a way to cut costs may be to shut down some stations. It would be wrong to close the most used station. Instead, closing the least used stations, just at least one of two really close stations, will decrease costs and avoid disrupting the service too much. The charts above suggest there are at least 3 stations that could be closed without harming the service.

If most and least used stations are examined based on user type, then it is possible to analyze even users' behaviors.



Picture 12. Most and least used stations based on user type.

Stations used with the highest and lowest frequency are in the same areas for members and casuals. This confirms the trend shown above. Members and causals also share more stations with low usage than stations with high usage. This is good news for Cyclistics. On one hand, there are a couple of

stations that can be closed without harming either annuals or casuals. This entails cutting expenditures used to sustain almost never-used stations. On the other hand, casuals' most used stations could have a higher undocking price to capitalize on the major number of rides and make the membership more attractive.

Summary

To sum up, the major trends that have been discovered during the analysis are:

1. Members do more rides (i.e have a higher count of rides)
2. Casuals have a higher average ride time.
3. Total ride time is higher for casuals during the hottest month. It is the opposite for the coldest months.
4. The number of rides contributes the most to members' total ride time; ride length, instead, contributes the most to casuals' total ride time.
5. Members use cyclist services more during the week for commuting or performing day-to-day tasks.
6. Casuals ride Cyclcistic bikes more during the weekend for leisure and visiting the city.
7. Members ride quite a lot during rush hours, suggesting that they may use Cyclcistic bikes to commute.
8. Casuals avoid undocking bikes for short (in terms of time) rides. This may be due to the cost that comes with undocking.
9. Most used stations are in the city center while the least used stations are in the suburbs.
10. Most and least used stations are located in the same areas for members and casuals.

ACT - DATA-DRIVEN MARKETING STRATEGY

Overall, many aspects distinguish annuals from casuals. The perfect marketing strategy should leverage them. Some strategies that may work better in increasing the user conversion rate (based on the data analysis) are:

1. **1,2,3… member**: Since the membership price is fairly competitive, strongly advertise that by paying as causal the price for just a few rides the users overcome the cost of an annual membership. The campaign could be something like: "*You know that you are just n-rides away from the price of your annual membership? Save money, be a member*". This will make the membership more attractive by highlighting its low cost.
2. **From Summer… to Summer**: Make the most of the campaign for casuals during the summer months. This is the period when they use the most Cyclistic's services, as such they may be more willing to subscribe to a membership. The catch is that the user is locked in up to the next year's hottest months.The perfect time for a casual user to re-new it, pushed by a tailor-made advertisement. It is a cyclical win-win campaign.
3. **From weekend riders to weekday riders**: Casuals ride more during the weekends. As such, the best shot could be to heavily advertise during those two days. A catchy sentence like: "*You know that you can ride on weekdays too?*" could work for this purpose.
4. **The magic of rush hour rides**: Casuals do not ride during rush hour. A good advertising strategy would be to show the benefits coming with it: no traffic, avoiding the risk of a stolen bike, faster than walking, staying fit and so on. Something like: "*Afraid of your bike being*

stolen? Just dock it" or "*If you have time to ride this, you are probably going too slow. Undock a bike and save time.*" should do the work.

5. **Place them where they see them**: This is an adv placing suggestion. The idea is to place advs in those areas that see the most used stations by casuals, to let them see advs clearly and on a daily basis. On the other way round, a good strategy is to avoid advertising the shutting down of a station if it is not constantly being used by casuals.

Aside from marketing, the top three management strategies that can tempt casuals to become members are:

1. **Infinite ride length? No thanks**: casuals have an average ride length far greater than members. Limiting the ride duration time for casuals while allowing infinite time for members could push casuals to subscribe to a membership.
2. **Cyclical prices**: slightly increasing the prices in the hottest months will make even more appealing a membership in those months. Instead, slightly decreasing the prices during the coolest month can induce casuals to do more rides, making, again, the membership more appealing. Aside from the conversion rate, this strategy may be good for increasing earnings (earning more in a single ride in summer; earning less but increasing the number of rides in winter).
3. **Higher price on the weekends**: Casuals ride on the weekends. Increasing those rides' prices may succeed in making casuals consider switching to an annual membership.

CONCLUSIONS

To sum up, annual and members do have different habits when it comes to use Cyclistics' service. This leaves room for many paths, in terms of marketing strategies, that Cyclistic can exploit to transform casuals into members. It is up to the marketing department to decide which strategy to pursue. Interestingly, the various options can be mixed to obtain even better results.

Note that by collecting more data, the analysis could drastically improve. Undocking price, bikes per station, and knowing how many rides are done by the same user are extremely useful pieces of information to tailor even better future strategies.
Besides that, to have up-to-date insights the analysis must be repeated at least every year, since users' habits may change quickly.

Finally, analysis results could be a proper starting point even for a marketing/management strategy aiming at increasing Cyclistics' customers, independent by their user type.