

Intro

The following is the summation of work performed to complete the Google Data Analytics Capstone Project. This write-up consists of 4 parts. First, I will present the scenario and data sources, all of which have been provided by Google's program. Next, I will detail and elaborate on the data cleaning and transformation process before moving on to the analysis and presentation of the data. The end of this section will ultimately result in addressing the stakeholder's request. This section will include charts and screenshots of data and will also be interpreted as a Powerpoint and video presentation.

This capstone was done using Excel. At the end of this write-up will be a supplemental, in an educational and meta-analytical-sense, detailing my experience going through the project with Excel and my expectations going forward doing the same capstone with R and subsequently SQL thru Google's Big Query.

PROMPT

What follows is the case study prompt provided by Google. Essentially the prompt has been copied here, excluding student instruction and links.



You are a junior data analyst working in the marketing analyst team at Cyclistic, a bike-share company in Chicago. The director of marketing believes the company's future

success depends on maximizing the number of annual memberships. Therefore, your team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, your team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations.

Characters and teams

- **Cyclistic:** A bike-share program that features more than 5,800 bicycles and 600 docking stations. Cyclistic sets itself apart by also offering reclining bikes, hand tricycles, and cargo bikes, making bike-share more inclusive to people with disabilities and riders who can't use a standard two-wheeled bike. The majority of riders opt for traditional bikes; about 8% of riders use the assistive options. Cyclistic users are more likely to ride for leisure, but about 30% use them to commute to work each day.
- **Lily Moreno:** The director of marketing and your manager. Moreno is responsible for the development of campaigns and initiatives to promote the bike-share program. These may include email, social media, and other channels.
- **Cyclistic marketing analytics team:** A team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy. You joined this team six months ago and have been busy learning about Cyclistic's mission and business goals — as well as how you, as a junior data analyst, can help Cyclistic achieve them.
- **Cyclistic executive team:** The notoriously detail-oriented executive team will decide whether to approve the recommended marketing program.

About the company

In 2016, Cyclistic launched a successful bike-share offering. Since then, the program has grown to a fleet of 5,824 bicycles that are geotracked and locked into a network of 692 stations across Chicago. The bikes can be unlocked from one station and returned to any other station in the system anytime.

Until now, Cyclistic's marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the flexibility of its pricing plans: single-ride passes, full-day passes, and annual memberships. Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members. Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps Cyclistic attract more customers, Moreno believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, Moreno believes there is a very good chance to convert casual riders into members. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs.

Moreno has set a clear goal: Design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the marketing analyst team needs to better understand how annual members and casual riders differ, why casual riders would buy a membership, and how digital media could affect their marketing tactics. Moreno and her team are interested in analyzing the Cyclistic historical bike trip data to identify trends.

Data Cleaning and Transformation

Data for a company similar to the prompt “Cyclistic” was provided. I downloaded the last 12 months of data plus an additional month. My decision to do so is a result of taking a quick glance at the data and noticing a key change in Cyclistic’s business: during the year of 2020, 2 more bike types were introduced. I thought it important to know how much the introduction of bikes affected ride numbers and diversity. So, I added an additional month which included only docked bikes. Each month is in its own separate CSV file.

Before delving into the data cleaning, I’d like to present some basis of understanding for your reference. Again, the tool I chose was Excel, the data’s integrity is considered ensured due to it coming from the Google program and is public data. Sources will be at the end.

The terminology used will be as follows:

<u>Term</u>	<u>Refers to</u>
Member(s)	Cyclistic users who pay for the annual subscription
Casual(s)	Cyclistic users who pay for single-ride or full-day passes
Ride(s)/Usage	The usage of a Cyclistic bike. One single user can have multiple rides per day or month. Each ride has its own “RIDE ID”
Ride Length	The length of any particular ride
Day	The day of the week for which any particular ride occurred
Month	The Month for which any particular ride occurred
Docked (bikes)	Cyclistic bikes that have docking stations. Each station has an ID and is registered with

	the RIDE ID when the bike is engaged or returned.
Electrical (bikes)	Cyclistic bikes that are electrical. These bikes SEEMINGLY don't need to be docked/ Hence the distinction
Classic (bikes)	Cyclistic bikes that are in a classic style. These bikes SEEMINGLY don't need to be docked, Hence the distinction

The business task is to identify differences between Members and Casuals and determine why a casual would become a member and how to convert them using the data we have. The data is organized into a number of columns detailing the following: Ride ID's, start and end date and time of the rides, start and end longitude/latitude of rides, start and end station ID's, start and end station names, member/casual, ride type (docked, electrical, or classic).

I converted all CSV files to XLSX. For this analysis I removed the longitude/latitude data. This was done primarily because there were a growing number of blanks month to month and the long/lat data was becoming increasingly unreliable or non-existent. The same would become true for the start/end station ID's and names. This is partly because of recording errors but also due to the increase in bikes that don't require docking, leading to rides being started and ended in locations that did not correspond with stations. Also, Excel does not easily calculate or map locations with lat/long data. However, I did calculate the most popular start/end stations by finding the MODE of each, using the station ID's. In later cases when no ID is available, I substituted blank start end ID's with 998 and 999 respectively. I will revisit this action later in the analysis section.

With the long/lat data removed and the station ID blanks accounted for I added a few columns to reflect ride length, day, and month. After doing so I needed to clean the data more.

I removed any rows wherein the ride date (start) did not match the month. Also, any ride lengths that returned negative were deleted. These would become negative because the end time recorded would be BEFORE the start time. While I could have switched these times, I didn't. Firstly, there is no way to actually know why the times were out of order. Was there a desync issue between the bike and app or however the times are recorded? Were the times mistakenly switched when recorded? Have users found a way to circumvent payment through some hack or glitch? There is no way to actually know. Second, most of these negative times, if switched in order to produce positives, would have registered as times averaging around 3 seconds and maxing at nearly 3 minutes. As a result, I removed this data to not interfere with my analysis.

I also removed any rows that had blanks in the member/casual column as this distinction is paramount to the analysis.

I did my cleaning in chronological order. As the months went by more errors would arise. For example, station ID's began as 3 digit numbers but became inconsistent by June 2020, and were no longer reliably managed and as a result were not used much in my analysis. Also as more bike types were introduced, as mentioned earlier, all location data became less reliable.

After cleaning the data, I performed a few extra calculations on a separate sheet. I calculated the most popular start/end stations, the most popular day, and the max ride length. This was done for the overall numbers and also for members and casuals separately. Of note is that it was often common that ride lengths would be recorded as day and even weeks long in some cases.

Next was the creation of pivot tables to better analyze the data. For each month I created the following : Total Rides & Average Ride Length by user type and bike type, Percent of Total Rides & Ride Length per user type and bike type, Daily Percentage of Total Rides per

user type and bike type, Weekly Percentage of Total Rides per user type and bike type, Daily Total Rides per user type and bike type, Daily Percentage of Ride length per user type and bike type, Weekly Percentage of Total Rides per user type and bike type, Average ride length per day per user type per bike type.

The cleaning and transformation is complete. Next is the analysis of the data. This section will be in a more presentational style.

Analysis and Presentation

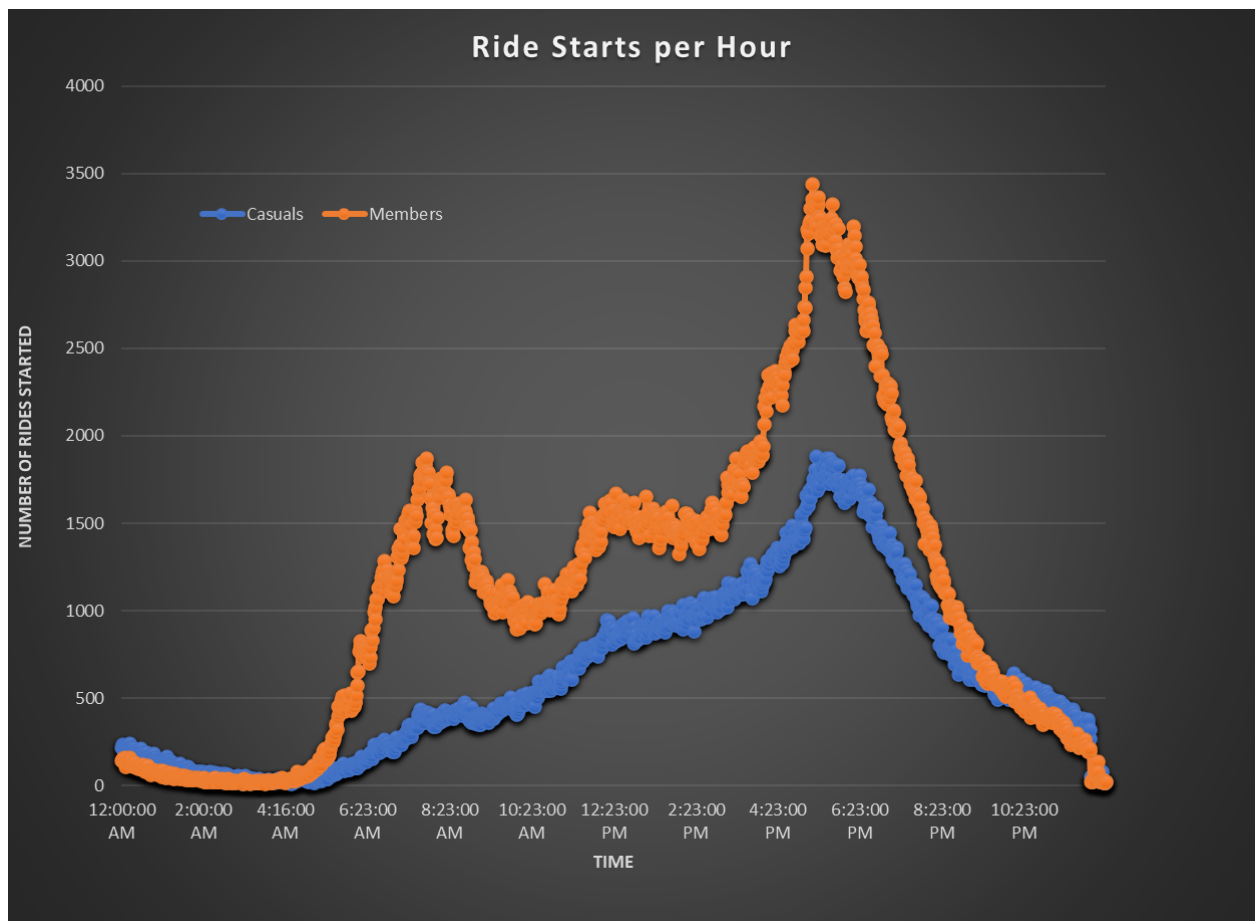
Purpose

The purpose of this project is to find actionable differences between Members and Casuals using the data provided and ultimately facilitate effective marketing driven attempts to convert Casuals into Members. Members are more profitable and converting Casuals into Members is seen as the way forward for growth. After analyzing the data provided, I have found a number of differences between the two rider types. I will explain and show these differences and also address how well we may be able to use these differences effectively.

Observation 1 - Why We Ride

It's been previously determined that about 30% of riders use Cyclistic bikes to commute to work each day. However, this fact only details a sliver of the narrative while also providing Cyclistic a window for growth. While the data provided shows each ride without distinguishing actual users, and we don't know how many Members we have, there is still lots that we can do with the data and the "70%leisure/30%commute" fact when compared to our data.

Regarding the “30% Commute” fact as true, there may be potential in convincing Casuals to bike to work, and as a result become Members. As shown below the most active times at which bikes are engaged are during rush hour. In the morning, Member usage spikes between 6AM and 8AM, a time when most people would be heading to work. Then we see two more spikes in bike starts at 12PM (lunch) and 5PM (leaving work). While Casuals follow the same trend line, the spikes are most prominent with Members.

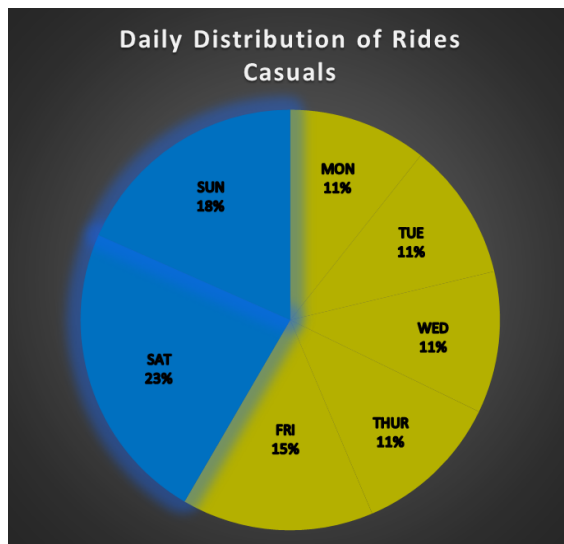


More data is needed to confirm, however, we can infer that even though only 30% of users ride for commutes, a larger portion of individual rides are actually used to get to and from work than. From here we can take a step towards answering why Members use their bikes: commuting.

Our purpose is finding differences between Members and Casuals. So if Members use rides for commuting, why do Casuals use rides? We can't for sure say exactly why Casuals get on a bike (and we are still inferring with Members) but we can make a short list that we can take action on. While not as strong as Members, Casuals' trend also follows rush hour expectations. Generally, Casuals ride steadily more and more throughout the day before peaking at around 5PM and decreasing thereafter. It's likely that many Casuals also use their bikes to commute. Also, even though we have no data pertaining to income, location, or occupation of users, we do have over 3 million points of data. We will revisit this later.

Observation 2 - When We Ride

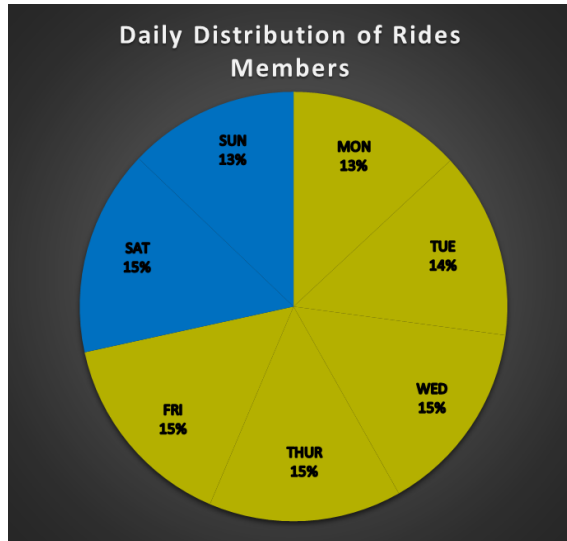
When we look at the week and days on which Members and Casuals ride, we can find more differences between the two. Casuals prefer weekends with 41% of rides happening on Saturday and Sunday (2 days) as opposed to 43% on Monday through Thursday combined (4 days), with Friday sitting just under 15%. We can also see that the number of rides on Saturday or Sunday reach nearly double the number of rides on any other day, besides Friday.



This observation also contributes to the last; Casuals likely ride for leisure while Members ride to commute. However, we can refine that perspective with a look at Members' daily ride distribution. Member ride distribution is almost spread evenly throughout the week, with the margin between the most and least popular days (Friday and Sunday) being 2%. Members use their rides as often on weekdays as the weekend.

Member riders are habitual riders and routine riders. Cyclistic bikes, for Member riders, are

ingrained into their daily lives while Casuals are more inclined to ride on the weekend. Members



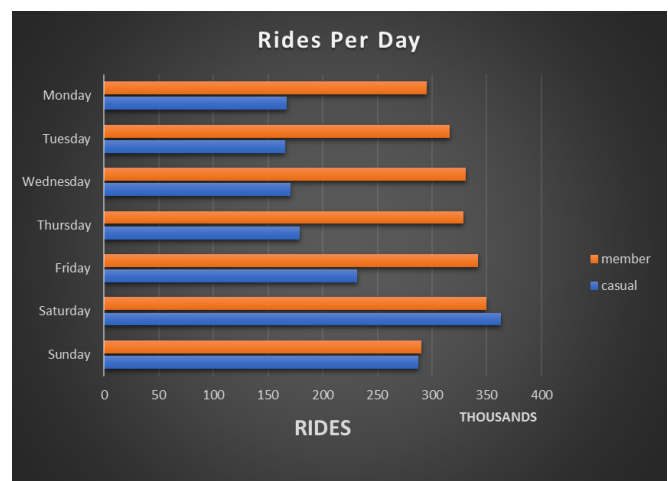
are consistent riders who ride everyday and we'll soon see that consistency in riding is a defining characteristic of Members, and a key difference from Casuals. We can also see this when looking at the bar chart to the right. We see that Members make up a significant amount of the total rides. Every weekday, Member rides outnumber Casuals until the weekend where Casual usage spikes significantly. This supports the previous

thought that Casuals tend to ride for leisure.

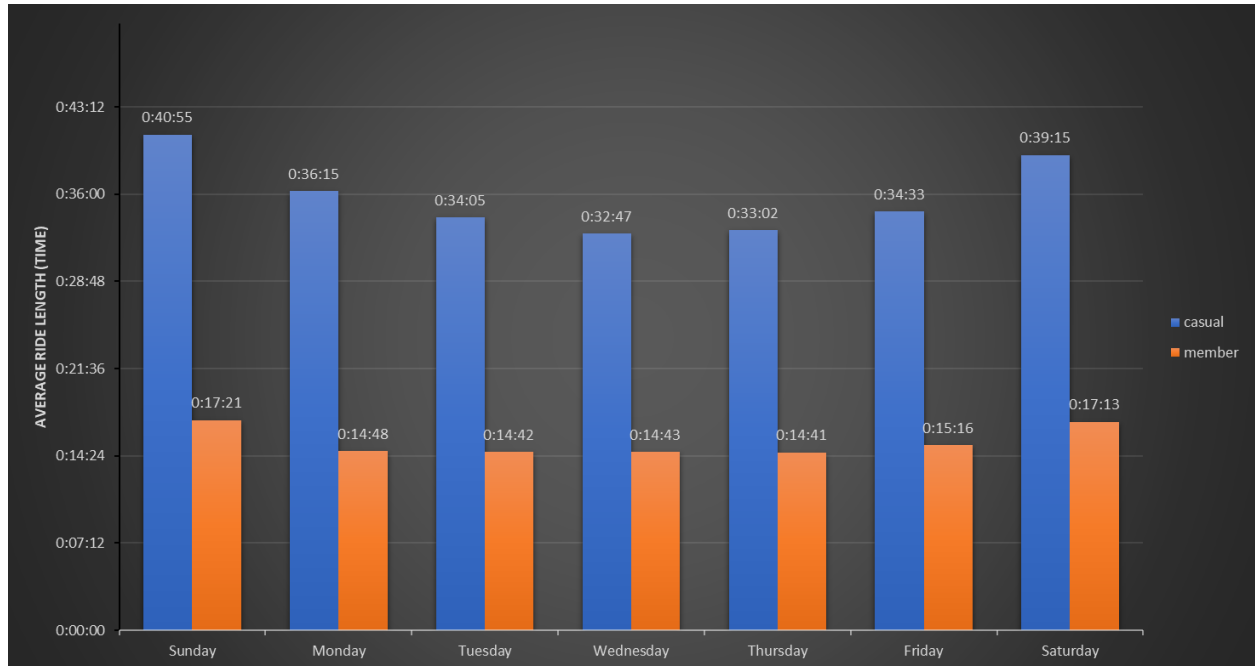
We may be able to use this difference to flip Casuals to Members.

Members and Casuals tendencies toward daily rides also extend to their average ride time. Casuals prefer to ride on the weekend and also ride for longer average times on the weekend. The time margin

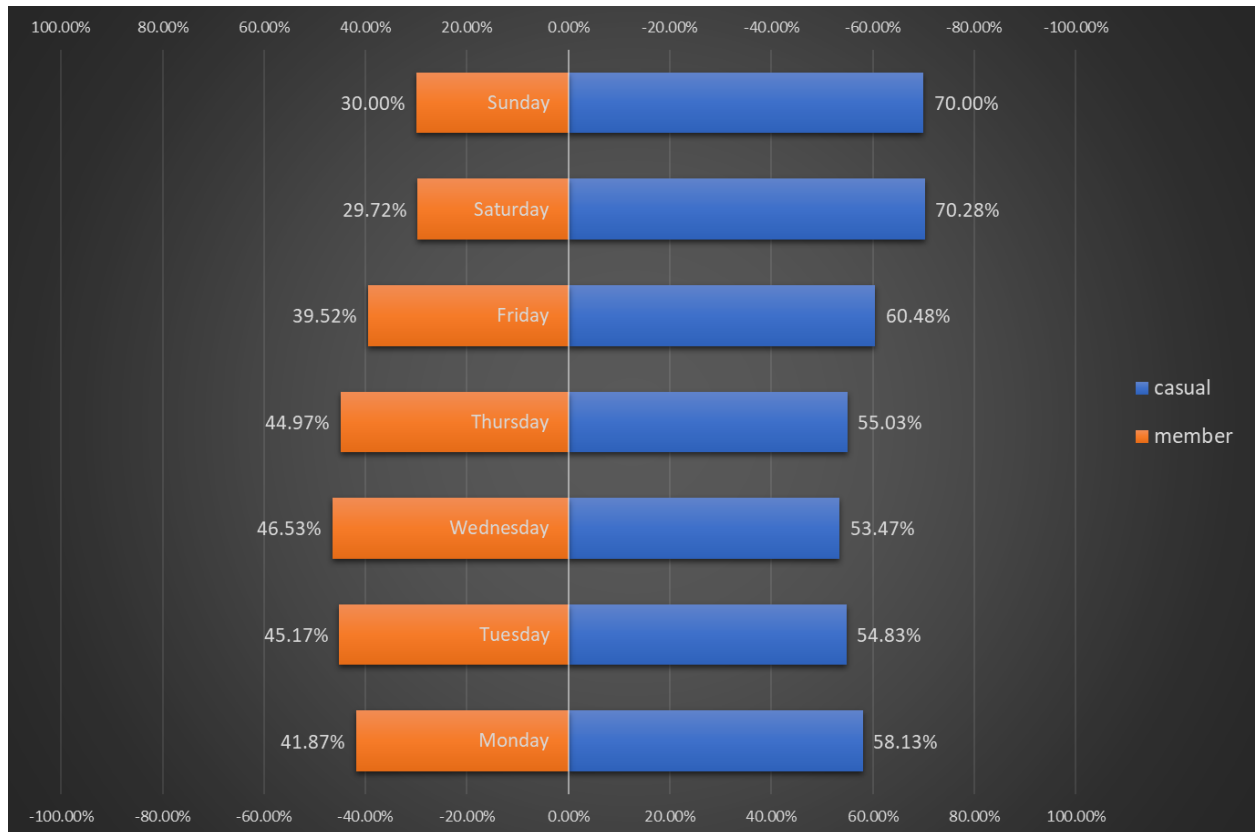
between the days is relatively small and the difference between the weekend and weekdays is between 4-8 minutes. However, with Members the average ride time seems more consistent throughout the week with the largest margin of difference being under 3 minutes. More



interestingly is the difference in average and overall times between the two riders.



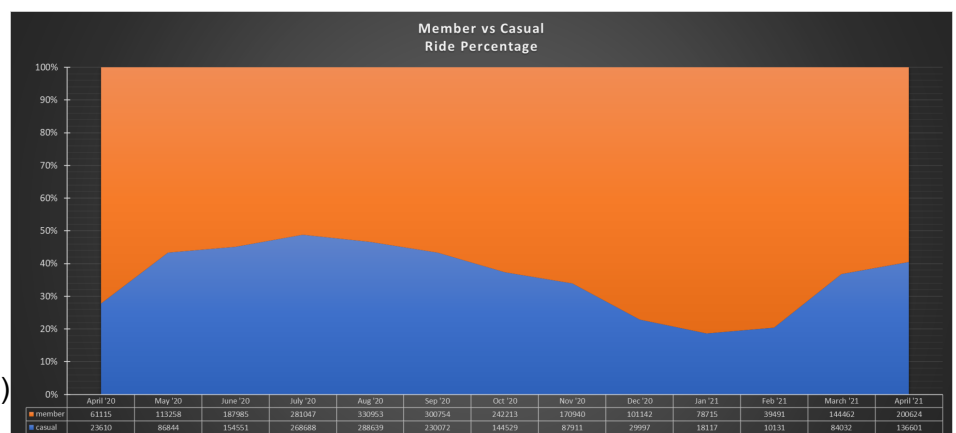
Casuals make up around 40% of bike rides but account for 62% of overall ride time and average over double Member average ride length. They also account for around 70% of all time spent with the bikes on weekends and 56% of the ride time on weekdays. Casuals are using Cyclistic bikes to the extent that Members, who ride consistently, account for nearly 60% of all rides and ride all year long, are dwarfed when it comes to ride length and average ride length. Casuals and their use of bikes on the weekend and for large periods of time is something we will review later in order to create our marketing strategy.



So far, we've seen differences in the consistency and length of rides through the 7-day week. Now, we'll look at differences across the seasons. As expected, ride count fluctuates depending on the weather.

The fluctuations are drastic with the most popular season (Summer-40%) outperforming the least popular season (Winter-7%)

by over fivefold. Chicago



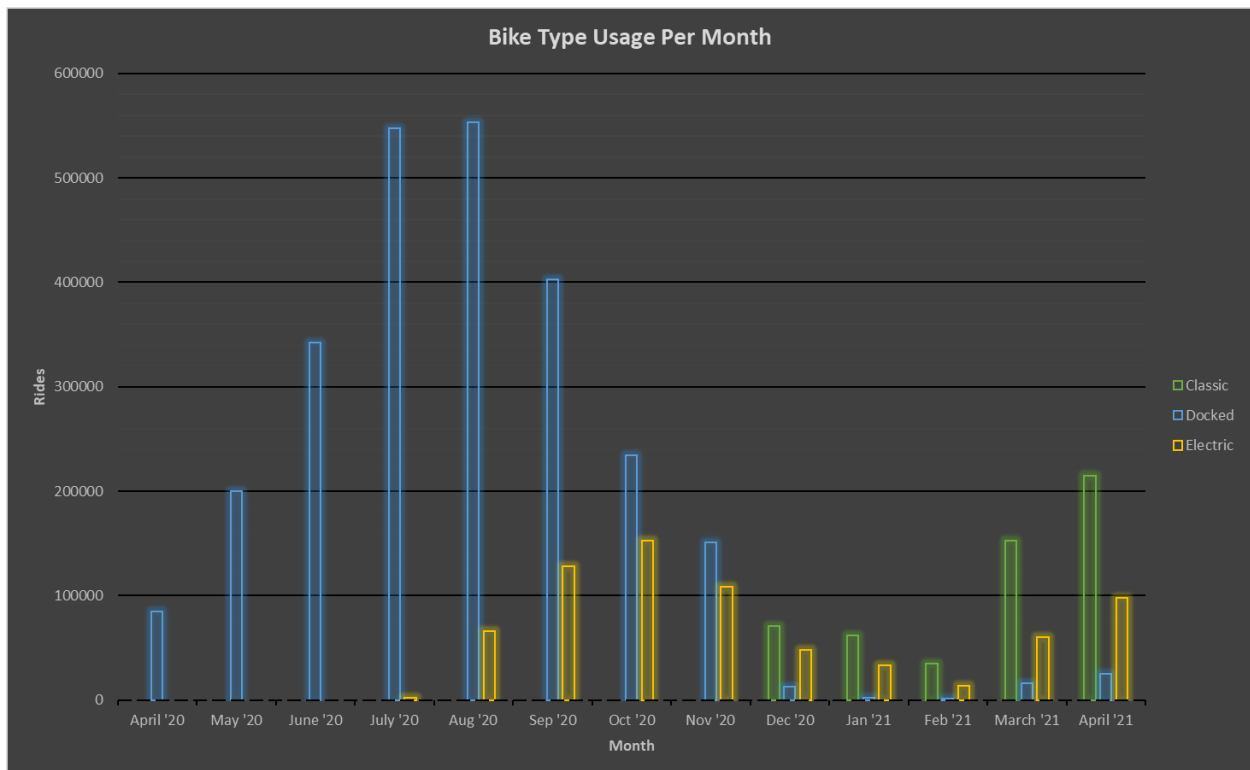
Winters can average low temperatures in the 20's and both types of riders are less likely to ride in the Winter. It must be noted that 2020 is also the year of COVID and rider numbers likely dropped faster and earlier than normal. However, while both riders avoid Winter, Members are

much more likely to ride in Winter and account for upwards of 75% of all rides in the Winter. If our previous theory holds true, this supports the thought that Members ride for work and habit. Other than that, it is difficult to differentiate between the users. This is due to 2 variables that are hard to account for without more data. In general, Cyclistic has become much more popular over time. April '21 has nearly 4x as many rides as April '20. With such a large jump in usage and only 13 months of data, it may obscure some differences present. Also, there is a large spike in Summer, which is presumably due to better weather AND tourist season, which would account for more Casual riders in the city. But we can't know currently. We can see that near the end of Summer, all ride usage steadily drops until Spring surges numbers again. Without knowing how much users have grown, which users are tourists, and why the number of rides has grown so much, we can't draw many conclusions from the seasonal data in regards to differences between riders.

What They Ride

Another possible avenue for converting Casuals may be present in the data for the bike types. In April '20, docked bikes were the only bikes available. Electric bikes were introduced in July '20 and Classic bikes in December '20. While the sum of the data shows that docked bikes were ridden the most, it is clear that their popularity is trending down. Docked bikes were very popular in early 2020, however, this is likely due to there being no other choice. Electric bikes were introduced at the peak of Cyclistic ride popularity. Over the next 4 months, the electric bike gained popularity while the docked bike declined. When the slow season starts, bike usage declines across the board but electric and classic bikes hold stronger than the docked, which becomes and stays the least popular throughout the rest of our observation. When comparing April '20 to April '21, we can see a few important things: The total ride count in '21, with 3 bike types available, is nearly 4x higher than '20 which only had docked bikes; Docked bike usage in

'21 is less than 30% of that in '20; And both electric and classic are more sought-after than docked by 4x and 8.6x, respectively.



Looking at the differences and similarities in bike usage by user type could lead us to some marketing solutions. Generally, between both Members and Casuals, bike type usage for each type follows the same trend as seasonality usage. In the summer, bikes are popular and when it gets cold fewer people like to ride. When the weather heats up, usage goes up and so does bike popularity. This is true for Members and Casuals, except in the case of docked bikes. After the introduction of the other two bike types, docked bikes continued to lose popularity even when the weather improved. So much so that Members stopped using them completely in January of '21. Casuals continued to use the docked bikes but they only gained a fraction of the popularity of electric and classic bikes when the trend rose.

It is clear that docked bikes are the least popular bike amongst both users. But, why? Without more data, like surveys or interviews, we cannot say definitively for sure, but we can

make some leaps. The following claims are made using what information has been provided in the case study and general knowledge. I'm assuming, after researching other similar companies, that electric and classic bikes do not have this limitation. I make this assumption because if all the bikes needed to be docked, there would be no reason in distinguishing a classic bike from a "docked" bike, and once the classic and electric bikes are introduced, the data begins showing more and more blanks with location data, possibly because bikes are not being left at the designated docking stations. While this could be for various other reasons, it trends heavily with the rise in popularity of the new bikes and fits easily into the theory of docked bikes needing to be docked while the other bikes don't. Without being given detail of what differentiates the bikes, I will be moving forward with this assumption.

Classic and electric bikes might have an advantage because of the inherent nature of docked bikes. Docked bikes are meant to be ridden to and from docking stations. Using a docked bike means that users will be required to: start and end your ride from a designated station, meaning users may have to travel to some extent to start and finish their ride; those who commute may feel the need to dock and then find another bike when arriving and leaving work; plan their rides. Electric and classic bikes seemingly do not have these limitations. These users don't need to walk from a dock station to their actual destination and are able to use the bikes as a "last mile" solution because the bikes can be left and found anywhere and users can use the bikes to begin, continue, or end whatever travel they are doing all around Chicago. The new bikes are simply more convenient than docked bikes and this appeal was strong enough for many users, particularly Members, to stop using docked bikes completely. Knowing this popularity and speculating on why could lead us to some marketing decisions.

Solutions

Our goal is to design marketing strategies to convert Casuals into Members. Below I will provide a few ideas for the marketing strategies using the insights we've gained from the data. Then I will briefly go over other options that may be available to us and how we can improve our efforts to answer the stakeholder question.

Marketing Opportunity 1

We established that Members are consistent riders, seemingly riding to commute and riding more consistently through the week than Casuals. But we also see that on weekdays, Casuals' trends in rides closely matches Members'. This likely means that some Casuals use the bikes to commute, at least sometimes. We can target these Casuals by highlighting to them that Cyclistic bikes are regularly and reliably used to commute to work. Through social media we can highlight easy and popular it is to commute with a bike. Assuming Cyclistic has an app, we can implement the use of location services to notify users of nearby stations when we detect they are traveling but not with our bikes. This will increase users' awareness of Cyclistic availability and convenience. Also, following the hypothesis that Casuals largely ride for leisure, we can utilize our social media and email platforms to alert users to events, new shops, and entertainment near Cyclistic docking stations, enticing Casuals to use Cyclistic bikes more often. The more often a single casual user uses the bike the more likely they are to see the value in an annual membership.

Marketing Opportunity 2

Making Casuals aware of their use of Cyclistic bikes could push them to become Members. On average, Casuals ride twice as long as Members and account for over 60% of all

ride time while only making up 40% of total rides. Making Casuals aware of how often and long they use Cyclistic bikes will likely push them towards becoming members in order to save money. The marketing should focus on prospective cost, the money and time saved by switching to an annual membership. This is also an opportunity to use the app and emails by informing Casuals of how much time they are spending with their bikes. Outside of social media and similar channels, we could market a promotion for weekend passes or an annual membership that only includes weekends. This would allow us to entice the Casuals who regularly ride but only on the weekends.

We could also consider rewarding users depending on how often or how much they ride. For Members this could offer benefits like bike reservations or redeemable points for events. For Casuals, this could be a special offer on starting an annual membership once they reach a certain distance or use milestone. We already know that Casuals use our bikes for a long average time so we know that this reward notification will also act as an ad for our targeted riders. Offering more of these reward benefits to Members would make a membership seem more valuable and entice Casual.

Other solutions