Case Study: How Does a Bike-Share Navigate Speedy Success?

Table of Contents

Introduction

Cyclistic, a bike-sharing organisation initially founded in 2016, has seen a noticeable rise in popularity throughout recent years, growing from a small local company to providing its services throughout the whole of Chicago with its current fleet of 5,824 geo-tracked bicycles and 692 docking stations. One of the pillars of its success is its differentiated brand that focuses both on a wide product range as well as inclusivity, having expanded its offerings to include adult tricycles, reclining bikes and cargo bikes, targeted specifically at those with disabilities, all alongside traditional bicycles that are the preferred option of 92% of customers. Whilst their primary use case is for leisure, approximately 30% of bicycles are rented to commute daily to work.

Cyclistic pricing plans consist of an annual subscription as well as full-day and single-ride passes, with customers using the former being referred to as members and those purchasing the latter as casual riders. Thanks to the work of the company's finance analysts, it was identified that the profitability of the first pricing plan greatly exceeds that of the second, making the firm adopt a new marketing strategy, namely targeting casual riders in

an attempt to convert them into annual members due to the already existing awareness of this group when contrasted with new customers.

Consequently, the executive team's as well as this report's objective is to study the habits and preferences of these two target groups, members and casual riders, in order to derive actionable insights that will allow Cyclistic to understand the motivations behind their customers' chosen pricing plan and ultimately implement measures to convince the latter group of taking up an annual subscription. Therefore, the business task that this analysis sets out to solve is the following: **How do annual members and casual riders use Cyclistic bikes differently?**

Used Data Sources

In order to answer the aforementioned question, two data sets containing information on bicycle trips were used: Divvy 2019 Q1 and Divvy 2020 Q1[1]. Both of these are two Google Docs files belonging to Google and can be accessed via the following links:

- Divvy 2019 Q1: https://docs.google.com/spreadsheets/d/1uCTsHlZLm4L7-ueaSLwDg0ut3BP_V4mKDo2IMpaXrk4/template/preview?resourcekey=0-dQAUjAu2UUCsLEQQt20PDA#gid=1797029090
- Divvy 2020 Q1: https://docs.google.com/spreadsheets/d/179QVLO_yu5BJEKFVZShsKag74ZaUYIF6FevLYzs3hRc/template/preview#gid=640449855

Divvy 2019 Q1 has the following structure:

Field name	Data type	Description
trip_id	Integer	The ID of the given trip
start_time	Timestamp	The date and time at which the customer began using the bike
end_time	Timestamp	The date and time at which the customer returned the bike to a docking station
bikeid	Integer	The ID of the given bike
tripduration	Float	The duration of the trip in minutes
from_station_id	Integer	The ID of the docking station the bike was taken from
from_station_name	String	The name of the docking station the bike was taken from
to_station_id	Integer	The ID of the docking station the bike was returned to

[1] It should be noted that the latter two are not named Cyclistic 2019 Q1 and Cyclistic 2020 Q1 simply due to the fact that Cyclistic is a fictional company, but these data sets will allow us to solve the business task at hand nevertheless. Motivate International Inc. has made them publicly available via the following licence: https://www.divvybikes.com/data-license-agreement

to_station_name	String	The name of the docking station the bike was returned to
usertype	String	The type of user: either a member or a casual rider
gender	String	The customer's gender
birthyear	Integer	The customer's year of birth

Divvy 2020 Q1, whilst quite similar to the latter dataset, still has some differences and is organised as follows:

Field name	**Data type**	**Description**
ride_id	String	The ID of the given trip (Note: it is not identical to tripid and is instead expressed as a hexadecimal)
rideable_type	String	The type of bike used
started_at	Timestamp	The date and time at which the customer began using the bike
ended_at	Timestamp	The date and time at which the customer returned the bike to a docking station
start_station_name	String	The name of the docking station the bike was taken from
start_station_id	Integer	The ID of the docking station the bike was taken from
end_station_name	String	The name of the docking station the bike was returned to
end_station_id	Integer	The ID of the docking station the bike was returned to
start_lat	Float	The latitude of the docking station the bike was taken from
start_lng	Float	The longitude of the docking station the bike was taken from
end_lat	Float	The latitude of the docking station the bike was returned to
end_lng	Float	The longitude of the docking station the bike was returned to
member_casual	String	The type of user: either a member or a casual rider

Even from a high-level description of the data, we can already identify some issues. Whilst both data sets contain the same information such as the type of user, the date and time a bicycle was taken from and returned to a docking station or the name and ID of the latter, they nevertheless do not apply the same standards or naming conventions. For example, after comparing the two, from_station_id from Divvy 2019 Q1 and start_station-id from Divvy 2020 Q1 do in fact represent the same information, namely the date and time when

the given bicycle was taken from a docking station. This inconsistency shall be addressed in the next section through standardisation.

Moreover, one data set contains information that the other one lacks and vice versa, meaning that some data for a particular year is not available. For instance, the Divvy 2019 Q1 data set also contains additional information on the gender and birth year of the customer, something which Divvy 2020 Q1 lacks, but which instead has data that is absent in the former such as the latitude and longitude of the docking stations.

The reason why these two particular data sets were used as basis for the analysis is because they give us access to the trips undertaken by both members and casual riders in 2019 and 2020. Comparing the behaviour of these two groups (such as their average trip durations) will allow us to derive valuable insights about the differences between them.

Data Cleaning and Manipulation

Please note that all the exact details of how the data processing and, in the subsequent section, analysis was performed can be viewed in the attached R file titled "Data Processing.R".

The entire process, all from cleaning and manipulating the data to analysing it and creating visualisations, was performed entirely using R through RStudio after having loaded both data sets. Whilst using spreadsheets was technically potentially feasible as well, the relatively large size of the data sets would have made it cumbersome to work with and hence suboptimal. RStudio in particular was chosen because it offers all the tools and functionalities allowing to efficiently resolve this business task as well as granting the flexibility that programming languages provide. Furthermore, it also has the advantage of making the analysis fully reproducible as opposed to being forced to document all spreadsheet steps used, SQL queries written or other operations undertaken. Hence, R appeared to be a natural choice for this particular case study.

Turning to the actual data cleaning and manipulation, a copy was created of both of the data sets, which were named rides_2019 and rides_2020 respectively. The very first step of transforming the data consisted of standardising the column names, as already briefly hinted at in a previous section. For more details, please refer to the "Data Processing.R" file. Following that, in order to facilitate future data manipulation, the *start_time* and *end_time* fields in both data sets were converted from character to datetimes types. From the description of Divvy 2019 Q1 and Divvy 2020 Q1, as outlined in the section "Used Data Sources", one can see that whilst the former does contain a field called *tripduration* (which was renamed to *ride_length*), it was absent in the latter. Thus, a new column *ride_length* was added to rides_2020 by subtracting *start_time* from *end_time*.

After having scrutinised the rideable_type field, it was found that the only possible value it had in Divvy 2020 Q1 was “rideable_type”, thus being useless and not bringing any information. For this reason, that column was removed entirely from rides_2020. Next, a new column called *week_day* was derived by extracting from *start_time* using a function as this field would prove to be useful for analysis later on. Once this was completed, the columns in rides_2019 and rides_2020 were reordered for more uniformity and to provide more clarity.

The next step was to verify the data types present in each column, before confirming that all trip and ride IDs were unique. It should be noted both datasets do not include duplicates because the trip and ride IDs for each record are distinct, which is why this important topic is not discussed in more detail. In rides_2019, it was ensured that *start_time* was strictly less than *end_time*, whereas for rides_2020 precisely 117 instances of inconsistencies were detected. Without further information, it appears self-evident that when the times were inputted into the data set, the start and end times were simply switched by accident, which is why this was corrected at this stage. Naturally, this modification required the *ride_length* and *week_day* fields to be updated. Also, just to be certain, it was verified that only non-negative ride lengths existed in both data sets.

Further data validation performed included looking at all possible values in the *week_day* column to guarantee that the data stated in rides_2019 and rides_2020 were actual week days. Upon comparing the two data sets, one could see that for the field *user_type*, the one used the values “Subscriber” and “Customer”, whereas for the other these were called “member” and “casual”. The latter were chosen in this case.

The next step in the data cleaning process consisted of verifying data constraints. More precisely all values of the *gender*, *birth_year*, *start_lat*, *end_lat*, *start_lng* and *end_lng* were inspected to ensure that they conformed to expected or real values. For example, any birth years before 1900 were filtered out and removed from the data sets, given that the likelihood of finding a person not only aged over 120, but also riding a bike is extremely low.

Another crucial aspect to address were NA values. Upon further investigation it was revealed that there were 19,711 rows in rides_2019 that had missing values for the gender field and 18,023 records with NA values for the birth year. As there is no method of finding out what the actual value was for each of these, these will not be replaced with something else. In addition, it was decided to not delete all those records with NAs because they could still provide valuable insights in cases when either the gender or birth year were of less importance.

Doing the same inspection for rides_2020 showed that there was only one particular row that contained missing entries for the end station ID, the end station name and the end latitude and longitude. In this case, merely analysing the record allows us to deduce the missing values. Comparing the start and end times, we determine that the bike was used for

a mere 12 seconds, which allows us to conclude with a probability bordering on certainty that this bicycle was returned to the docking station it was taken from.

Afterwards, the data sets were inspected for outliers. In rides_2019, 192 records were found having ride lengths over 24 hours, with there being 290 of these in rides_2020, thus representing merely 0.053% and 0.068% respectively of the entire data set. Considering that for casual riders the longest pass they can purchase is a full-day one, it is extremely unlikely that a ride should last longer than 24 hours. Therefore, as these data points almost certainly constitute errors rather than actual bike rentals and as they only occupy a negligible fraction of the whole data set, these were excluded.

Once this was done, all the records that were removed for quality control (i.e. those having the start station "HQ QR") were filtered out, resulting in a further 3,767 records being deleted from rides_2020. Finally, the two data sets, rides_2019 and rides_2020 were merged, creating a new data set labelled all_rides, which will be the subject of the analysis in the subsequent section.

Note: Although it could be preferable to remove the fields gender, birth year, start_lat, start_lng, end_lat, and end_lng in order to ensure that the resulting data set is as uniform and complete as possible (as well as to limit the occurrences of NAs), these variables may nevertheless provide useful insights. For this reason, they have been retained.

Analysis

In order to provide a satisfactory answer to the question of how the bike usage behaviour of annual members and casual riders differs, we shall first begin by analysing general statistics about both categories. This shall be followed by exploring how, if at all, mean ride length varies based on the group using the bicycle, before analysing if the frequency of rides per day of the week is correlated with the user type. Finally, it shall be verified whether gender is a differentiating factor between casual riders and subscribers.

General Overview and Comparison

	General	Casual riders	Members
Mean	13 min 43 s	38 min 23 s	11 min 25 s
Median	8 min 59 s	23 min 8 s	8 min 28 s
Maximum	23 h 55 min 55 s	23 h 55 min 55 s	23 h 53 min 4 s
Minimum	1 s	2 s	1 s

Table 1

Before delving deeper into analysis, first the summary statistics of the two groups will be scrutinised, whose results are displayed in *Table 1*. Based on the table above, casual riders do indeed have quite different usage patterns compared to members, at least based on the ride length parameter. We can observe that casual customers, on average, use bicycles over

three times longer than subscribers per use with their means of 11 minutes 25 seconds and 38 minutes 23 seconds respectively. This finding makes sense since one-time pass purchases are generally more expensive and hence casual users appear to seek to maximise their use of the bicycle. A possible alternative explanation is that members, thanks to their annual subscription, do not hesitate to use bicycles for even short distances, which leads to a significantly lower average ride duration.

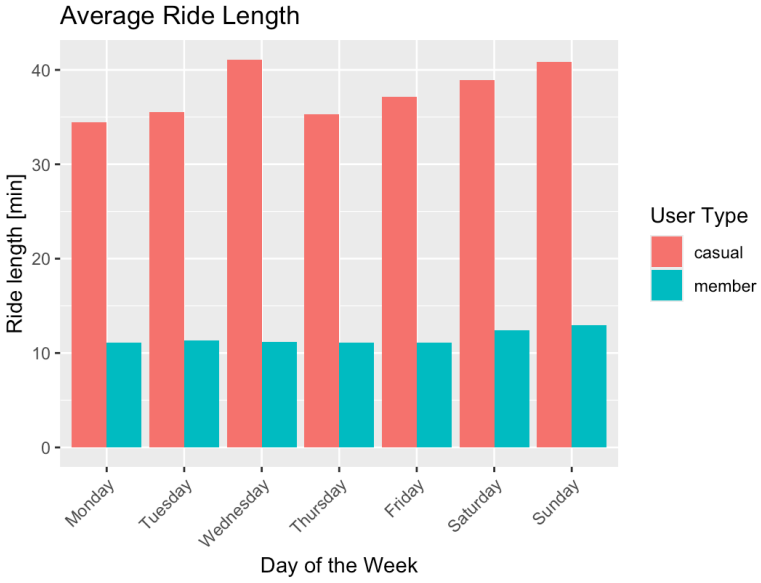
The median reveals a similar picture: The median ride length of casual riders is approximately 2.7 longer than that of members. The durations obtained are visibly lower than that for the mean, as one would expect, given that this data set still contains values which one could argue are outliers such as ride lengths of 20+ hours. Speaking of which, the maximums ride lengths for both categories naturally converge towards the same value, i.e. close to 24 hours, simply because all larger durations were filtered out.

As to the minimum bike use times, these are of 1s for subscribers and 2s for customers. Although these are merely pure speculations, the minimum bike usage of 2s for casual riders could be due to an accidental purchase or where someone merely experimented on how to rent a such a bicycle without having the immediate need to use it. In contrast, for members the minimum bike usage of 1s is logical, considering that they do not have to pay per use due to their annual subscription and can hence take and return bikes as often as they desire or even change their mind after having taken a bicycle from a docking station and immediately return it.

To recap, the main discovery we learned from the summary statistics is that casual riders have a tendency to rent bicycles for significantly longer periods of time compared to annual members.

Further Investigating the Mean Ride Length

As the previously mentioned observed phenomenon could reveal additional patterns, we shall now look at it in more detail by verifying whether the ride length differs by day of week per user group. To illustrate this more clearly, *Figure 1* displays two types of columns: orange ones for casual riders and blue ones for members. The x-axis lists the days of the week and the y-axis the mean ride duration. On it we can see what we already discovered previously: that average ride lengths for casual customers are three times longer than for members.



We notice that there is a

Figure 1

small general increase in the mean use time during the weekend for both user types, particularly on Sunday. This can be easily explained by the fact that the primary use of Cyclistic's bikes is for leisure and Saturdays and Sundays are the days when people have the most free time. What is interesting, however, is that casual riders specifically tend to use bikes for slightly longer periods of time on Wednesdays, namely 41 minutes 7 seconds in contrast to their mean of 38 minutes 23 seconds. Regrettably, to discover the cause behind this would require further investigations that are beyond the scope of this analysis. That aside, generally speaking, there are no perceptible differences trend-wise between subscribers and casual riders in their bicycle usage per day of the week.

Number of Rides and User Type

Having scrutinised the mean ride length in depth, we shall now turn to enquiring whether the number of trips can uncover any valuable insights into the two user types' bicycle use patterns. The natural expectation is for the number of members to be considerably higher than that of casual riders, which is confirmed in *Table 2* below. There, when comparing the mean number of rides for the two categories, we observe that of the daily bike use 96.6% is due to annual members. This can be seen even clearer in *Figure 2*.

	Mean number rides per day	Mean number rides per week day	Mean number rides during weekend
Casual riders	9,654.57	7,117.6	15,997
Members	102,875	120,113.2	59,779.5

Table 2

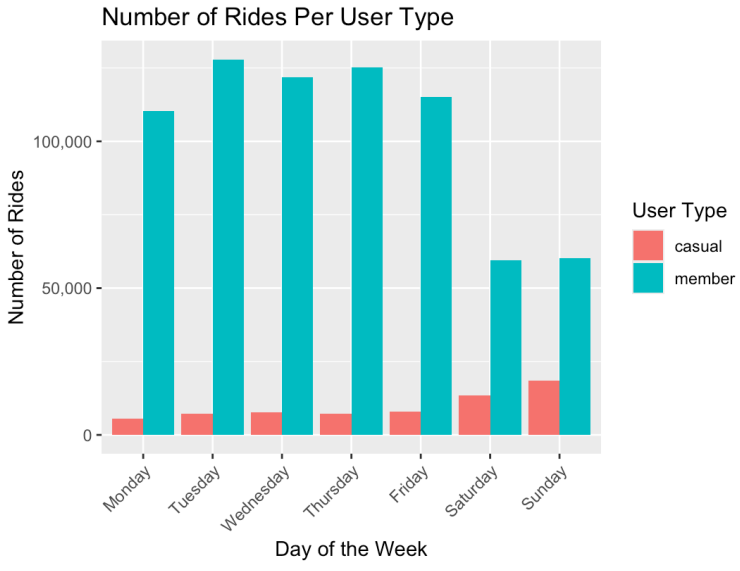


Figure 2

The graph's x-axis once again displays the day of the week, whilst the y-axis shows the number of rides, where the same colour coding is being used to represent both user types. The start contrast in size between the two is immediately apparent.

Looking at that graph, we can detect that there appears to be something that can almost be described as an inverse relationship: During the week, the mean number of rides for members constitutes 120,113.2/day (as stated in *Table 2*), but we witness an abrupt decline during the weekend, where the average number of rides for this group halves, dropping to a mere 59,779.5/day.

Conversely, for casual riders we observe the opposite trend, with 7,117.6 bicycle rides throughout the week and then suddenly rising by 225% during the weekend, reaching almost 16,000 trips per day on average.

A hypothesis that could explain this behaviour at least in part is that the majority of those aforementioned 30% of Cyclistic customers who use the bike-share program to commute to work are probably annual members, which would explain the sharp drop-off in number of rides for subscribers on Saturdays and Sundays. Should this be true, this would reveal that for this subgroup, the major justification behind becoming an annual member is to use bicycles on their daily work commute during the week. Following a similar line of reasoning, although further investigation would be required before one could state this with absolute certainty, it appears that one of the major motives behind casual riders renting a bicycle is to use it for leisure. These two findings provide valuable insights that will be further discussed in a later section.

Gender and User Type

With the average number ride length and number of trips per day having been examined, in this section we will consider if the user's gender may have an impact on membership or not. However, before beginning, it should be noted that the caveat for this part is that, as explained in section "Data Cleaning and Manipulation", our used data sets only contain information for the year 2019 and of those records a comparatively small number, namely 19,711, contained NAs rather than the true value. It is therefore possible that usage patterns by gender may have changed in 2020, but this would not be detected in this analysis due to those missing values. This limitation should be taken into account by the senior management before taking any courses of actions based on the conclusions reached in this section.

We began by generating summary statistics based on the user type and gender, the results of which are shown in *Table 3* below.

User Type	Gender	Count	Percentage
Casual rider	Female	1,872	0.542%
Casual rider	Male	4,052	1.17%
Member	Female	65,009	18.8%
Member	Male	274,290	79.5%

Table 3

We are immediately confronted by the fact that most of Cyclistic's user base are male as seen when looking at the percentage column. This discovery is visualised more easily in the graph labelled *Figure 3*, displaying the gender on the x-axis and the percentage on the y-axis, where the size of the columns for men is significantly higher than for women.

Using *Table 3* we can compute the proportion of male casual riders, giving us 14.5%, meaning that, in other words, 14.5% of all male Cyclistic bike users are not members. This is noticeably lower when compared to the ratio of 28.0% for women, which tells us that the chances of a male customer converting to a subscriber seem to be approximately double of that of a female customer becoming an annual member. It thus appears that there could potentially be a worse awareness amongst female customers of the benefits of annual memberships, which is a crucial aspect to consider.

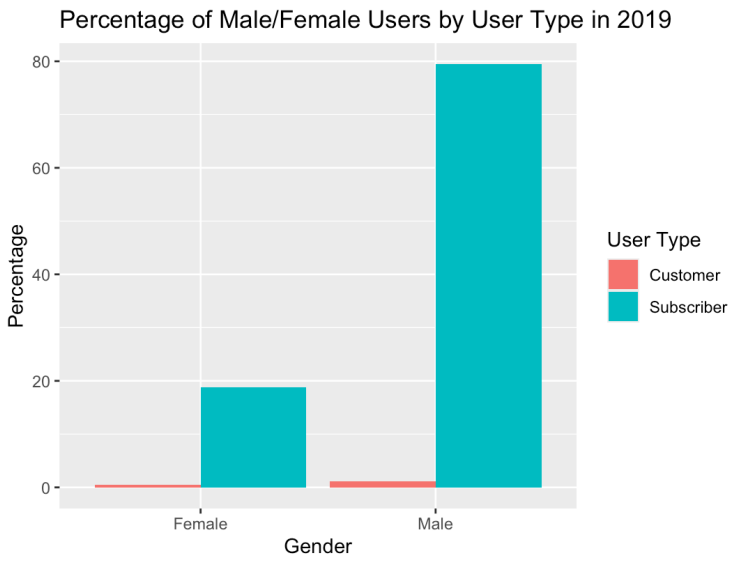


Figure 3

As a side note, due to the evident size difference between the two populations as seen in *Figure 3*, this inference should be taken with a grain of salt. Nevertheless, this graph does highlight that, for a future analysis, it would be worth digging into the causes of why the overwhelming majority of Cyclistic bicycle users, namely 80.67%, are male and potentially start a marketing campaign targeting women in particular.

Key Findings

For the sake of clarity, all the major discoveries will now be reiterated:

- **Casual riders use bicycles for periods of time over three times longer** than annual members on average (11 mins 25 s vs 38 mins 23 s respectively)
- **On Wednesdays, casual riders see a slight increase in mean bicycle use time** of 2 minutes 44 seconds
- The trip frequency follows an inverse relationship amongst the two categories: Whilst **members use bicycles two times more throughout the week** than on Saturdays and Sundays, the **number of rides for casual** riders is low during the week, but **doubles during the weekend**
- **Men are two times more likely to purchase an annual membership** than women.

Top Three Recommendations

With the most valuable insights fresh in mind, we shall now turn to the actionable takeaways from this analysis. First of all, with the finding that female customers are two

times less likely to purchase an annual membership, it is suggested to launch an advertising campaign that specifically attracts the attention of female customers and increases their awareness of the benefits of Cyclistic annual memberships. However, it should be noted that due to female clients only representing a comparatively small fraction of the company's customer base, namely 19.23%, the costs and resources required to perform this may outweigh the benefits achieved. Thus, if possible, it would be preferable to combine this with a marketing campaign targeted at women in order to gain more female customers in the process and maximise the effects of this initiative.

Turning to the next suggestion, it was discovered that one of the major differentiating factors between casual riders and subscribers is the fact that the latter mainly rent bicycles to use during the weekend. For that reason, purchasing a complete annual membership does not seem to be advantageous for them as they will only be using it during, at most, 104 days out of 365 (i.e. all Saturdays and Sundays), but paying for the entirety of the whole year.

Therefore, the second recommendation is to create a new, fourth offering, namely a so-called weekend membership targeted specifically at weekend bike trip leisure seekers – in other words, an annual subscription that gives the customer unlimited access to Cyclistic bicycles for free exclusively during the weekend. This is anticipated to convince many of the casual riders to opt for an annual weekend membership, thus gaining new customers in the process and increasing revenues.

Needless to say, in the event that this recommendation should indeed be deemed of interest, extensive market research as well as financial analysis ought to be conducted in order to ensure that this offering does not cannibalise on the already existing profitable annual membership. Particular care and planning are thus needed to implement it in such a way to ensure that more casual riders become weekend members, whilst maintaining customer loyalty to the yearly subscription currently in place. This strategy targets approximately half of the casual riders' population, i.e. the "weekend leisure seekers", however the other half still rents bicycles throughout the week, which leads us to our third and final recommendation.

Without further data on the casual riders renting bicycles between Mondays to Fridays, it would be difficult to determine what barriers and obstacles they face that discourage them from becoming annual members. Therefore, it is advised to launch a survey that should be sent to all casual riders asking them such a similar question, amongst others such as, to target specifically those who use it throughout the week, enquiring if they use Cyclistic bicycles to commute to work to better understand their use cases. The findings from this survey could prove to be the deciding factor, which, if correctly addressed, could encourage casual riders to make the switch. More precisely, it might very well be that certain casual riders might prefer not to purchase an annual membership not due to its price but for other reasons. For instance, perhaps there are specific areas in Chicago that do

not have sufficient bicycles or whose docking stations are too separated and far away from another to be conveniently used, thus making the local population only rent Cyclistic bicycles occasionally rather than on a regular basis.

To sum up, the recommendations are the following:

1. Launch a marketing campaign targeting women and increasing their awareness of the advantages provided by annual membership.
2. Introduce a new offering, i.e. an annual weekend subscription to win casual riders using bicycles for leisure on Saturdays and Sundays.
3. Conduct a survey to determine what other obstacles hinder casual riders, especially those renting bicycles throughout the week, from purchasing annual memberships.