Review of Differences in Ridership

By

Alexander McAndrew

The last year and a half has been a troubling and difficult time for everyone. The citywide quarantine, and the devastating economic loss companies have endured. All while struggling to provide services while helping colleagues ensure the health and financial safety of their families. We understand that the bike services Cyclistic has provided has been invaluable to casual riders and commuters alike.

Our goal is to grow our rider membership family by maximizing the conversion of leisure riders to annual membership holders. This analysis is to help identify how annual membership holders and casual riders ride differently. This insight will help our marketing analyst team pinpoint their next campaign.

Dataset, Cleaning, and Manipulation

I have dataset I have used was from the last 12 months which can be found on our website for download. ([Index of bucket "divvy-tripdata"](https://divvy-tripdata.s3.amazonaws.com/index.html)) These datasets contain a generated ride ID, bike type, start and end times, start and end stations, the corresponding ID’s, and geographic data(latitude, and longitude). There is no personal or financial identifying data within these logs.

I first wanted to identify the length of each ride which was done by formatting the date and time, then finding the difference between start and end times. I then took the date of each start time and found which day of the week it fell on. Afterwards I began to clean up the start and end stations. All rides are logged regardless of if it is done by internal or external use, so I removed all entries that were involved with our testing facilities. This was done with the assumption that these were not normal rides and often involved unrealistic ride times. There had been rides performed by casual riders that had uncharacteristically high ride lengths however as they had been recorded by casual riders I have left these in.

Throughout the months of datasets there had been multiple issues regarding how the times had been logged. Typically, the difference in time would create the appropriate ride length however there had been many entries that had the start and end times reversed. This caused an issue where the average ride length of both casual and member riders had been incorrect and had been shown as “####”. These rides have also been removed to ensure the accuracy of average ride times.

Along with this issue the new year found that we had changed the way we had been categorizing these test stations. Initially test stations had been labeled as such under the station name columns. Come the new year this had been changed to note the test status under the station id instead along with how the test stations were labeled as well.

There had been initial concern regarding station locations that had been completely blank or geographic data that was less accurate. However, these have been left in as there had been what appeared to be a realistic use case. This was decided based on the length of ride time which typically fell within reason. With this in mind geographic data was formatted down to the millionth as I was aiming to also record where these bikes were being used from.

With the data now cleaned and in order I began to import to my SQL database and created a view containing all twelve months so I would be able to begin answering my questions.

Analysis

I began with looking at the number of riders and the length of their rides for the entire year. Finding which riders rode longer, and which riders rode more frequently. While this gave a very clear-cut answer, I was looking to track down any other pieces to the puzzle that might give insight. I continued to review which days of the week were the busiest by rider type, as well as which starting stations were the most popular by month.

My findings show that membership holders are prone to riding far more frequently, however these rides tend to only last roughly half as long as casual riders. Where the opposite is true for casual riders. They will ride less frequently but for longer periods of time. Membership holders tend to consistently ride throughout the week where as casual riders will ride primarily on the weekends. Generally, there did not seem to be an overwhelming number of stations used by members but the areas closer to downtown did show the most activity. Casual riders had a few very high traffic stations which I found were closer to the lakefront and navy pier area.

As it stands, I have concluded that casual riders are not only taking more leisurely rides they are primarily leisure rides themselves, and membership holders will use the bikes to commute to work.

My recommendation to our marketing team would be to consider a campaign that focuses on leisure and exploration, and weekends. Using either of these two options will allow us to tap into additional casual riders who might not be fully aware of the benefits of the membership program. As another approach I would also request our teams consider altering the membership program to be more inclusive towards casual riders. This would have the membership work for the number of ride counts within a month/ calendar year or to appeal to those who will ride for longer periods of time. This alteration can be made to work for whichever set of criteria has been met first.

Another option would be to have the memberships set by the travel of relative distance from the starting station. If the membership were to include access to areas outside of a certain radius of the starting station, I feel that casual riders might be more inclined to join with the appeal of extended distance or unrestricted distances (within a reasonable limit).

Moving forward I would hope to work more with our other teams to focus on accuracy and data collection. There had been an issue regarding how the start and end ride times had been collected and recorded as this later caused an issue while analyzing. While ultimately this data was possible to format there were still issues in which attempting to work it in SQL. This combined with another recording issue regarding the formatting of member type would be useful. Luckily this issue was extremely minute but did impact when looking at the top 1 million longest rides. I had reviewed each month individually but was unable to track down this issue even after formatting. I would also hope to see, if possible, about recording the active distance. Geographic data is wonderful however many of the entries were lacking accuracy by various degrees. While I would later attempt to review and see if there was more of a station impact on our rider-base I feel that distance recording of each ride may make future analysis easier.

Conclusion

I highly recommend that our marketing team consider altering the membership program and focus on advertisements that appeal to those seeking leisure during the weekends. Here will be a majority of our membership expansion waiting to join so they can continue their enjoyment without any impediment.