Customer Stream Improvement Proposal

Action commitment analysis with Clustering and Applied Scientific approach for Capital Bikeshare's business

Group 3

- *Chiavarino Federico*
- *Giancola Simone*
- *Liscai Dario*
- *Marchetti Simone*

Bocconi University, MS in Data Science and Business Analytics, 2021 – 2022 a.y.
Business Analytics 20595

Table of Contents

List of Tables

List of Figures

Introduction

Capital Bikeshare is metro DC’s bike share system disposing of more than 5000 bikes and about 600 stations across seven jurisdictions: Washington, DC.; Arlington, VA; Alexandria, VA; Montgomery, MD; Prince George's County, MD; Fairfax County, VA; and the City of Falls Church, VA. It offers a simple service able to satisfy the citizens’ needs, guaranteeing affordability, sustainability, and the possibility of moving easily into the city.
It was born in September 2010 and since then it has kept growing, with new jurisdictions and stations becoming part of the system. Capital Bikeshare is operated by Motivate (see appendix, Introduction).

The service works in a very elementary way. The customer first joins Capital Bikeshare’s system by using a kiosk or the app. The choices are essentially between a long-term deal or short-term passes so that both occasional and frequent needs can be addressed. The app offers a map of the city and illustrates the location of available bikes. Once the bike of interest is spotted, the customer can scan the QR code in the app, get a ride code at the kiosk, or use a bike key to unlock it. At the end of the ride, the bike can be returned to any station making sure that a green light is displayed on the dock.
The offered packages and related prices are the following:

- Single trip: $1 + $0.05 per minute for a classic bike or $0.15 per minute for an e-bike;
- 24-hour pass: $8 to be paid at the beginning of the ride with no additional fee for the time spent;
- Annual membership: $7.92 per month (billed upfront for $95), with free usage for the first 45 minutes. (and additional benefits such as bike angle points, group rides, etc...).

The company also offers special memberships reserved for universities and corporations (Pricing @ Special Promotions). In any case, if a user does not return a bike by 24 hours, a fee corresponding to the value of the bike (same as a purchase) will be claimed.
Capital Bikeshare’s website and app suggest a set of popular and interesting rides to help customers enjoy a ride in nature or the city. As the single trip pricing suggests, there are two types of bicycles: classic bikes and e-bikes. The latter are more expensive than the former

as they are provided with an electric motor to assist propulsion, reducing riders' effort for longer trips.

If interested in the story about Capital Bikeshare and the competitors in the city, a more detailed description is attached in the appendix (Competitors).

The final scope of this work is to propose some initiatives, potentially having a positive impact on the company's revenues. To achieve this, we will apply the eight steps of the scientific approach, starting from a theory raised by some evidence caught from our data. We will then develop some tests that will reveal to us how to act. Through this paper, we will go over the eight steps in detail, after a comprehensive description of the data used and their processing into an analytical format.

Preliminary Data Analysis and Processing

In the firm's framework, our task is conducting some analysis, possibly using clustering techniques, to deduce insights for Capital Bikeshare. We are provided with two raw datasets (Rides Data, Weather Data). The former stores specific information about customers' rides during the three months of 2021, specifically from February to April. The latter is a general weather dataset with hourly measurements of key indices in the same period.

One important aspect to consider is that Capital Bikeshare's dataset does not include the IDs of riders, but rather proposes the IDs of rides. For this reason, any analysis will relate to the nature of rides and riders' habits rather than the types of customers that use the service. If, in addition to this, the identities of users could be identified, then the analysis could also include considerations of the subjects of interest and not only of their usage patterns.

After cleaning (appendix**,** Data cleaning) and merging the two datasets, we obtained a final dataset with more than 450000 rows and 21 columns, which lost around 5% of the initial observations.

Moreover, for the purpose of the analysis, some columns such as the duration of the ride, polar hours (to avoid the categorical form of hours), spatial distance (between the starting station and the ending one), and general data transformations are added. For the interested reader, a more detailed analysis is provided in the Jupyter Notebook attached to this document.

The aim of the features created is that of *summarizing* some of the original variables to make them clustering friendly (continuous) and efficient.

Plots & Basic Statistics

This section will provide an overview of the data we observed.

Let us begin by observing the information provided about Washington DC’s meteorology. First of all, we analyze the types of weather conditions encountered in the city between February and April: the weather is overcast around 46% of the time; clear in one out four days; partially cloudy 15% of the time; it rains around 1 every 10 days and it snows only 2% of the time (even though the actual frequencies vary across the months, without showing huge differences). Hence, we are provided with a majority of “cyclable days” samples, meaning those days in which it does not rain or snow, something we think might obstruct bicycle usage. As just said, conditions vary across different months. Without considering the temperature, they are comparable for the last two months, March and April, while they appear to be worse in the earlier one, with more rainy and overcast days (a detailed histogram is proposed in the appendix, Conditions by Month Histogram).

From here we can analyze how the rides are distributed across months. Rides’ frequency follows our expectations given the nature of the service. As shown in the following histogram, bicycle rides are less attractive during colder months.

Almost half of the total number of rides is performed in April, while just around 16% of the total rides are done in February. This huge gap makes us think that customers are heavily influenced by weather conditions and temperature when deciding to use Capital Bikeshare’s bicycles. Another evidence in support of this idea: half of the rides happen with a temperature higher than 14.3°C, even though there are just 21 days out of the 90 considered that have a mean temperature equal to or higher than 14.3°C. Hence, our belief that temperature plays an important role is concrete.

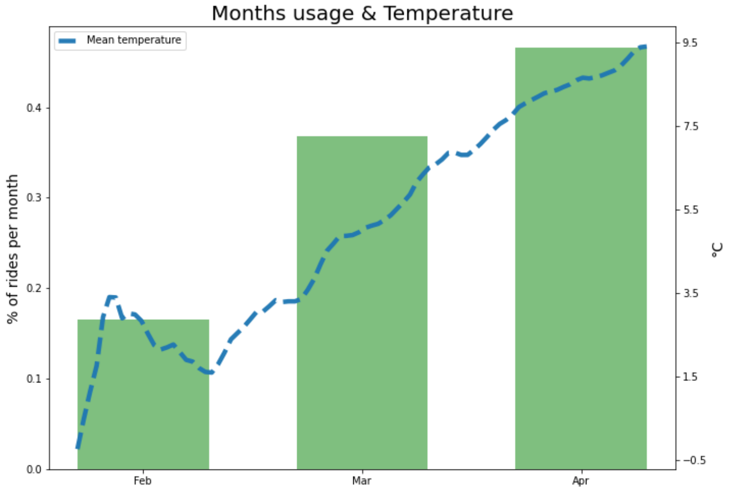


Figure 1: Mean temperature and proportion of rides per month combined

At this point, it is worth considering how the service is used across the day’s hours. Peak hours, as expected, are concentrated in the afternoon, when either leisure or general activities combine. The second part of the afternoon is a moment in which both workers come back home and people like to spend time engaging in social/open-air activities. A plot is reported in the appendix, Peak Hours Histogram.

Member vs Casual problem

One of the most important attributes regarding Capital Bikeshare's data is the one related to the "*type of ride*". There are two possible types of ride: members and casuals. The former are the ones who use an annual membership, while the latter use all the other types of offers, such as day-pass or single ride. Moreover, differently from other features, they present an *almost* balanced share of observations: members account for 58 % of the rides, casuals for 42 %. The binary nature of this characteristic, its innate tractable distribution, and its business meaning make it a well-posed reference point.

The distinction between casuals and members is underlined in almost all plots of density and histograms, once conditioned on it. Consequently, we can now show some plots to understand these fundamental differences across the two groups. Even aspects not straightforward in terms of intuition return a difference between what we will refer to as members and casual from here onwards.

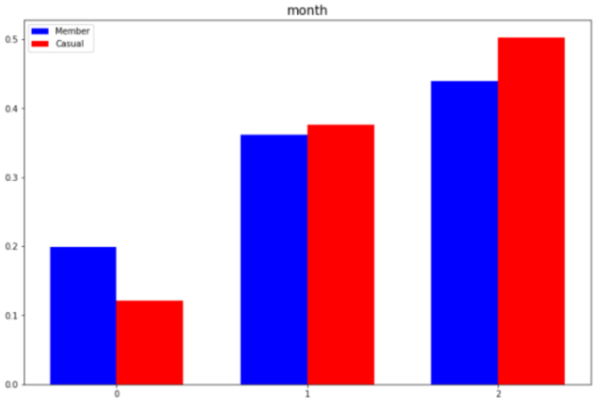


Figure 2: Type of customer share across Months

The first thing worth analyzing is the difference in usage of the service across the three months. Members appear to be less month dependent, with a steadier use from February to April. Contrarily, casual users present an evidently bigger portion of rentals only in the last month.

Moreover, we can consider the differences across days of the week. While it is almost constant for members, casuals tend to use it significantly more on weekends.

On the other hand, when considering weather conditions, it can be argued that the difference is less noticeable. Similar trends are recorded, with a lower eagerness to ride with bad conditions of casuals, being that they are more likely to ride for leisure rather than for impellent needs.

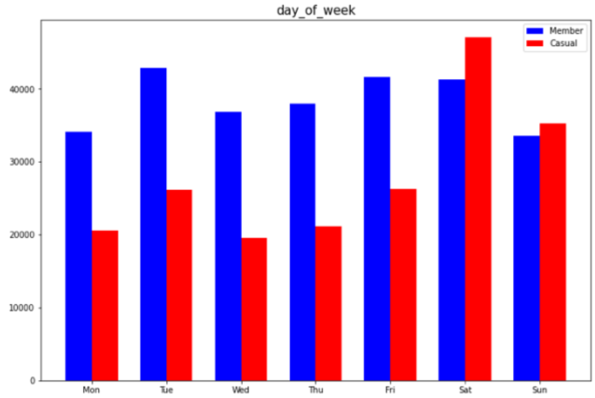


Figure 3: Type of customer share across days of the week

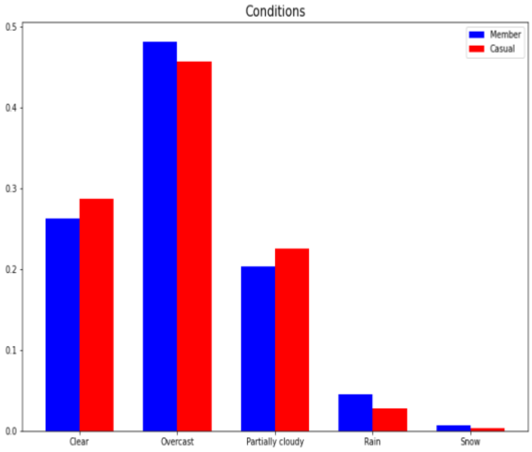


Figure 4: Type of customer share across weather conditions

Finally, we consider the difference in the duration of the rides. After having winsorized rides length distribution, it is possible to pivot on it. The result is as expected. Members tend to do shorter rides, which is probably due to the fact they use the service for commuting purposes and hence do shorter trips. Differently, casual riders use the bicycles for a greater amount of time, indicating that these rides are not likely related to commuting but rather to leisure. The estimated density plot is proposed in the appendix, Time of rental Density Plot.
To sum up, all these facts combined strengthen the idea that a regular customer uses the service not only during free time, presenting regularities that a casual user does not feature on average.

Clustering

Apart from the insights provided by the data, we want to dig deeper into the analysis of the rides. *How can we do that?* Given the nature of the data, we would like to discover some patterns in the rides observed. To do this, we run a cluster analysis to identify distinct groups of rides with similar characteristics. Once we identify group-related patterns, we analyze the most important traits for each cluster and make inferences about the type of ride we might be observing. Since we do not know the IDs of users from the dataset, we can identify only general trends and not individual ones. Nevertheless, from the type of ride, we can hypothesize the type of user using the group's most relevant characteristics. For example, if we have a group of rides with the peculiarity of doing longer rides, we can suppose those rides identify a particular type of user: a tourist or somebody who is doing a bike ride. Instead, if we observe a group of rides all having a shorter usage, we could associate them with a group of commuters. One thing to note is that we are creating an identikit for each cluster of rides, but this does not mean that a user belongs to one identikit only. In fact, a user may belong to different groups depending on the rides he or she has done.

There are many algorithms suitable to run a cluster analysis and, in our case, we decided to use a K-Means algorithm. A K-means algorithm has the objective of partitioning the observations in K different clusters so that each observation belongs to the cluster with the nearest mean. The main advantage of this type of algorithm is that it is very efficient when dealing with large amounts of data. Since we have more than 450k observations, K-Means is the only algorithm able to solve with the computing power we had available. The only drawback related to this type of algorithm is that we need to choose the number of clusters beforehand.

Given the statistics of the dataset seen before, together with our experience, we think we might observe two or three distinct types of rides: rides that can be classified as “routine”, rides that can be classified as “leisure time” and something in the middle of the two. After running some tests, we decide that three is the optimal number of clusters based on the characteristics we observe within them.

While with two clusters distinctions were not marked enough, as we increased the number of centroids, the algorithm started splitting the three previous clusters into smaller subsamples according to weather conditions only. There were groups composed of a small number of observations on single meteorological conditions.

Before looking at the results, it is important to discuss which features we decided to include in the clustering and why. Note that only numeric columns are accepted since we cannot define a “distance” between categories even if we encode them. The columns we picked are the following:

- temperature, precipitation, and cloud cover of the hour in which the ride started, since we think that different meteorological conditions might play a role in determining the type of ride we observe;
- distance of the ride and time spent as they are main features characterizing the ride; hour started and hour ended to capture different cycling habits;
- geographical information (latitude and longitude) about the starting and ending positions, since these might identify different groups of cyclists.

Now we discuss the results obtained and visualize their differences with the help of some plots. The first thing we notice is the size of each cluster. The three groups have 209396, 144098, and 110427 observations respectively (we will refer to them as *group 2*, *group 1,* and *group 0* to identify them from now on). *Group 2* is the largest, with almost double the observations of *group 0*, which is instead the smallest among the three. First of all, we compare the share of rides performed by members in each group. We remember that each group needs to be associated with a certain type of bicycle usage, rather than user identity.

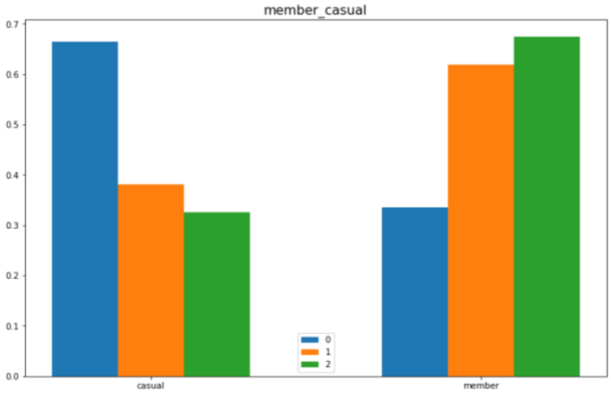


Figure 6: Clusters proportions wrt customer type

From the plot above we immediately see a difference between the groups: *groups 2* and *1* have a similar percentage of members, attaining around 65%, while the last group (*group 0)* has a majority of casual riders. Now we check if we can spot any difference between the groups for different weather conditions in the plot below.

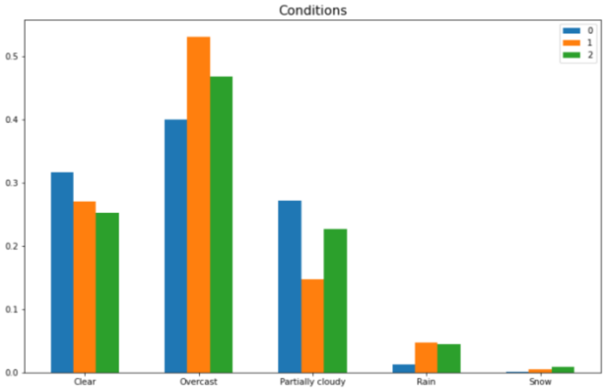


Figure 7: Cluster proportions wrt weather conditions

Group 0 seems to use the service mainly with favourable conditions. In fact, the usage on rainy or snowy days is close to 0. Furthermore, the usage is higher than the other two groups when we have a clear or partially cloudy day. Again, there are just slight differences between groups *2* and *1*. *When do we observe a difference between these two groups then?* Here is an example of when this happens:

The plot on the right shows the distribution of starting hours of the rides, divided by clusters. Here, *group 2* starts using the bicycles around 5 am and finishes using the service around 4 pm. This share of rides seems to capture all those clients who use the bike from home to work. *Group 1* is instead highly concentrated from the afternoon to the end of the day. We think this captures the share of rides of those who use it to go back home. *Group 0* rides are mainly present between 10 am and 8 pm. Let us look at what are the most preferred days for each group.

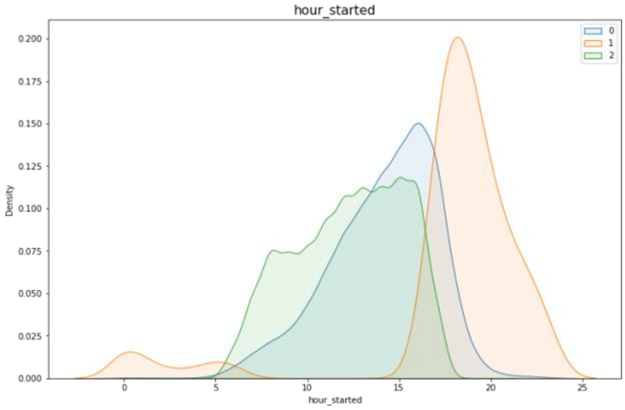


Figure 5: Clusters' densities wrt starting hour

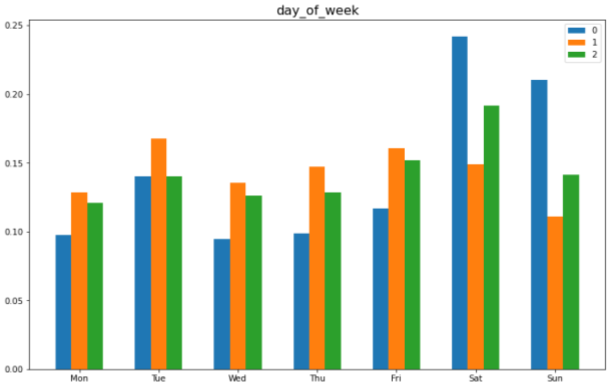


Figure 8: Cluster proportions wrt days of the week

The graph above seems to be in line with the difference we observed in the member vs casual problem: *groups 2* and *1* have more or less stationary usage across the weekdays, with a peak on Saturdays for the latter. Instead, *group 0* rides concentrate almost 50% of the cases during weekends. This can be linked with the fact that the majority of the rides are done by casual riders in this group. The next graph shows the difference in monthly usage.

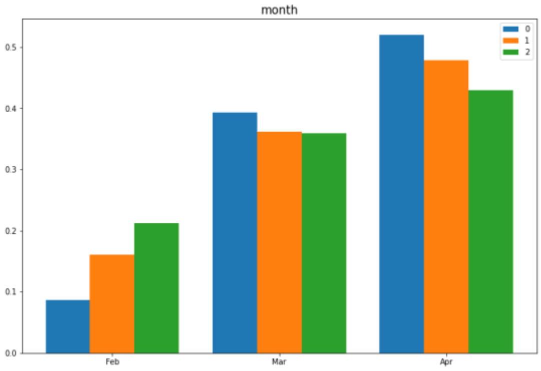


Figure 9: Cluster proportions wrt months

First of all, we notice there is not a uniform usage across months for all groups, as shown in the preliminary statistics. However, the difference is more pronounced for those rides belonging to *group 0*, with more than half of the rides happening in April only. Together with the features discussed above, *group 0* seems to be more associated with leisure time, meaning that these rides typically happen in favourable meteorological conditions (high temperature, no precipitations, sunny days…) which is easier to encounter during April rather than February or March. Nevertheless, all groups show an increasing trend in usage from February to April. An additional consideration must be done on the distribution of time usage for each cluster. Also, in this case, the main difference arises between *group 0* and the other two. The difference in distributions closely follows the one seen in the Member vs Casual problem, with *group 2* and *group 1* having a right-skewed distribution (indicating

shorter rides), while *group 0* rides appear to be longer.
To sum up, we observe two similar clusters of rides that differ only in the time of the day in which they happen. Apart from this trait, they follow similar characteristics when it comes to other features: similar time used per ride, similar share of members, and similar days and weather conditions in which they are done. All these traits make us think that these are rides used for commuting needs since they are typically shorter and they take place in hours of the day that can be associated with house-to-work travels. Moreover, their uniform usage across the week seems to confirm a routine-based usage. On the other hand, *group 0* differs on almost all the features observed: longer rides, usage concentrated at the weekends, and during days with more favourable weather conditions. These traits make us think these rides belong to a free time usage in which the aim is to relax with a bike trip or something similar. This is corroborated by the fact the majority of rides are performed by casual customers, who typically do not follow a routine-based usage. However, the clustering helped us understand that rides cannot be split only using the membership condition: *group 2* and *1*, associated with commuting purposes, are not performed by members only: this might indicate the willingness of some casuals to use the service they way members typically do, but for some reasons, they do not prefer to remain casuals.

Story Tree

Given what we observed in the previous analyses, we built a tool that should help us to derive new promotions logically: the story tree.

Step 1: Identify your story

What am I observing? As previously seen in the sections "*Member vs casual problem*" and "*Clustering*", there are different types of rides depending on different days of the week, hours of the day, time spent, weather conditions, and month.

Why is that? Customers use the service for different needs and purposes, which can be mostly identified in the distinction between commuting and leisure. Weather conditions, day of the week, and other traits discussed above play a major role in determining whether these kinds of needs can be fulfilled.

What can I do? We aim to derive promotions that can capture different needs which are not taken into account in Capital Bikeshare's current offer. Our starting point is that it is in the interest of the company to have customers that pay an amount of money *now* for some *future rides* (i.e. members) as opposed to people who sporadically pay and get the rides. Keeping this in mind, we try to craft new promotions which are favorable both for customers (satisfy their needs) and for the company (fixed income).

Step 2: Identify the most important elements

At this point, we define the most important elements we must consider. We propose a hierarchical structure (from most to least important):

- Casual vs Member: this is the most important element because it delineates two different kinds of customers both in the usage of the service and company cash flow. As just said, we assume it is in the interests of the company to have many members, because they represent a fixed income.
- Current offers provided by the company: to craft new promotions, it is necessary to know which are the current ones. We presented these in the "*Introduction*" section.
- The attributes that allow outlining different needs about the rides, such as *time spent during the ride*, *day of the week*, *conditions, month*.
- *What are the current promotions of bike-sharing companies in other cities of the US?*
- Given our insights, *what could we offer to exploit consumers' needs?*

Step 3: Identify the reasons

1. *Why casuals?* We think the main reason behind the choice of not becoming members is related to flexibility, both in terms of payment (they might not want to pay an annual membership if they are not sure to use it for so long) and usage (sporadic use of the service). This is also confirmed by the data analysis: it can be observed that on average casual rides are prevalent during weekends and with favorable weather conditions.

   We think casual riders can be divided into three groups: those who use bicycles for leisure purposes only; "potential members", users who might have an almost regular usage, but think the actual membership offer might not be fully exploited; those in the middle of the two groups, who might have sporadic commuting needs and also want to use the service to enjoy a ride.

   Why members? On the contrary, there is a share of people who regularly uses the service and exploits its routine-based usage to save money. Hence, we can identify these customers with people who have constant specific needs repeated over time and already know they will use Capital Bikeshare's bicycles to move around the city.

2. Capital Bikeshare's current offers show a huge gap in terms of duration: either an annual membership or a single-day pass/single ride. Moreover, there are also strong differences in the pricing: the annual membership costs $7.92/month ($95 billed upfront), while a 24-hour pass is $8/day. This imbalance of pricing shows how the company pushes its customers towards an annual membership as it is in their interests to anticipate profits.

3. *What are the reasons behind the difference in the attributes?*
   - Time spent: a high ride-time means a moment of leisure or relaxation, while a low ride-time is typically related to a specific objective, such as going to work or university.
   - Days of the week: the difference is the same as the previous point, usage on the weekends may underline times to have fun, while during the week it may be related to more specific needs. In addition, weekend rides can be attributed to tourists coming for a visit.
   - Conditions: combined with the attributes above, we expect that rides happening during any meteorological condition serve commuting needs, while those happening during favorable conditions only are related to leisure time.
   - Month: the different usage across months is strictly related to differences in meteorological conditions. People tend to use the service less during colder months, with this being especially true for sporadic users.
4. Let us consider the comparison of Bikeshare's offer with respect to the ones of some other bike-sharing services across the US:
   - Boston's Bluebikes: monthly membership, annual membership (both paid upfront or monthly).
   - Philadelphia Indego Bike Share: 24-hours access, 30-day pass, annual pass.
   - New York Citi Bike: Day Pass, Annual Membership, Annual Membership "Premium" (with more facilities, such as 3 free scooter rides).
5. Our aim is to craft new promotions that come closer to customers' needs. We will see how by making connections among the most important elements outlined so far.

Step 4: make connections

Here we try to make connections among them to get more insights. There is a link between the flexibility required by some casual customers and the current offer of Capital Bikeshare: the ones available are either too restrictive (annual membership) or too vague (pay-per-ride). Moreover, the relationship between Capital Bikeshare's current offers and the ones proposed by other bike-sharing companies can give us a hand to find out what is missing: monthly memberships, missing in Capital Bikeshare's current offer, may meet the needs of additional flexibility. Furthermore, other promotions suggest we should not focus exclusively on bicycle-related offers, but rather to expand our concept of "promotion" by adding facilities that would ease customers' decision to become members. The additional facilities should be linked to the issues preventing consumers from using the service to discover what we should address first in tackling our problems.

Scientific Approach

Step 1: Idea

The story tree guided us along the ideation process. We began by observing different usage of the service depending on several factors: membership to the company, weather conditions, day of the week, month. Consequently, we can think about promotions dealing with flexibility in terms of frequency of usage and partnerships to try to take into account the difference in usage depending on diverse weather conditions.

Step 2: Problems

We think the company is facing two main problems: casuals not becoming members (even if there might be an interest in doing so) and the huge gap in usage across months.
We believe the former can be addressed to the lack of flexibility in the current membership offer. It is an important problem because casual riders do not represent a constant income for the company, but rather just a *"one-shot income".* As stated several times, it is in the interest of the company to retain as many customers as possible. Moreover, there may be a possibility that these customers will not use the service again if attracted by competitors. Again, we could not measure exactly the magnitude (i.e. number of users involved) of this problem since we could not recognize individual customers in the dataset, but we imagine the company should tackle this as their priority.

Moving on to the second problem, we observed a consistent usage decrease during months typically associated with cold and bad weather conditions. In fact, among the three months considered, February saw the lowest number of rides. Even though we observed this fact for one month only, we do not consider it just a coincidence, and we believe our claim can be reasonably extended to the other winter months, maybe even with a stronger effect. Hence, the company's revenues might suffer more during colder months.

Step 3: Scenarios

At this point, we can identify different scenarios. These have been designed from the analysis of the dataset, in particular by looking at the different types of rides. The scenarios we outline refer to different ways the customer might be interested in using the service.

- S1: *seldom usage with exclusive interest for the weekend*.
  This is inspired by the regular proportion of rides performed by casuals, or *group 3* rides in the cluster analysis, who show a consistent usage mostly during weekends. This suggests there may be a possibility that some customers have an exclusive interest in using the service during weekends only. Tourists coming to visit the city and townspeople who have no time during the week are potential candidates.
- S2: *seldom usage without exclusive interest for the weekend*.
  The idea is that there could be customers who might have sporadic commuting needs across weekdays and also use the service to enjoy a ride. So, they do not have an exclusive interest in the weekend. We can think of these customers like those in the middle between casuals and members, but who do not have such a constant usage of the service to become members.
- S3: *regular usage with people influenced by bad weather (rain, snow, or terrible conditions)*.
  There are customers with constant needs that bring them to regularly use the service. In addition to this, given the nature of our service, they might be influenced by weather conditions in deciding whether to use it or not.
- S4: *regular usage with people not influenced by the weather*.
  In general terms, this is similar to the previous point with the exception that there may be people using the service independently from weather conditions.

Step 4: Actions

At this point, we can identify the actions we have thought about. For each one of these a possible price is proposed in the appendix (Pricing).

A1: **Weekend pass**. Get access to all members' facilities for a weekend. The idea behind this offer is to target both tourists who go to Washington during the weekends and townspeople who have free time only during the same period. Obviously, the weekend pass would be more beneficial than taking two one-day passes. In this case, we would offer a prepaid service.

A2: **Rides carnet**: package of 10 rides with access to members' facilities to use anytime. The idea behind this promotion is to offer something that is paid in advance and that can be used with no restrictions. This would be suitable for people who rarely use the bike, without a dominant preference of some days over others. For instance, a person who occasionally takes a bicycle for commuting purposes or simply to enjoy a ride.

Here it is important to underline that this promotion is about the number of rides and not the number of days. This implies that each unlock action accounts for a decrease in the remaining rides. In this case, we would offer a prepaid service.

A3: **Monthly membership with a partnership of the Washington Metropolitan Area Transit Authority (WMATA)**. The idea is to propose a partnership with WMATA, responsible for managing Washington's public transport services. In this way, people who do not feel good about riding a bicycle with bad or extreme weather conditions (extreme cold, snow, or rain) have free access to public services through this innovative bike-sharing membership. The partnership should be crafted to be beneficial for both agents, so the choice on whether to offer free access or tickets' discount should be decided with the other company. The choice would obviously also be reflected in the price of this type of membership. In this case, we would offer a subscription-based service that can be canceled at any time.

A4: **Monthly membership**. In this case, the idea is to offer a meeting point between an annual membership and a day-pass for the reasons outlined before. Some people may not consider the possibility of getting the membership only because they are not willing to pay for a year-long service without the certainty of actually exploiting it. The downside is that they do not frequently use the service due to the high cost of a single ride and because of all drawbacks of not being a member, such as not paying for the first 45 minutes. In this case, we would offer a subscription-based service that can be canceled at any time.

A5: **no action / pivot**

Step 5: Scenario-Action Map

SXA	S1: seldom usage + exclusive interest for the weekend	S2: seldom usage without exclusive interest for the weekend	S3: regular usage + bad weather influence	S4: regular usage + no bad weather influence
A1: weekend pass	++	-	-	-
A2: carnet	+	++	+	+
A3: month + WMATA	-	+	+++	++
A4: month	-	+	++	+++
A5: no action/pivot	0	0	0	0

Table 1: Scenario Action map

Once identified scenarios and actions, it is now the moment to combine them in the scenario-action map to assess every possible combination of them.

The scope of this matrix representation is to identify the most profitable action per scenario. Starting from S1, the most suitable action would be the weekend pass because it perfectly reflects what customers are looking for: they are provided with a "small membership" for the only two days of the week during which they are exclusively willing to ride. Another good option can be a carnet of rides as it still grants flexibility. However, in this case, the latter fits customer needs less than the former. Nevertheless, the expected value of its implementation would be positive. Monthly memberships – both with and without the partnership of WMATA- would cause a negative impact instead: under this scenario people prefer shorter contracts and cheaper payments.

Differently from S1, under S2 a monthly membership might positively affect Capital Bikeshare's revenues. Customers do not show an exclusive interest for the weekend, hence A3 and A4 partially accomplish them. *Why partially?* Because these actions may better fit a more regular user, but if the customer thinks that in a particular month he or she may use bike-sharing more than usual or for longer tracks, then interest for monthly memberships may increase. Under S2, a Carnet of rides would be the perfect option because people do not prefer the weekend over other days and can distribute their rides in the way they most prefer. Only a weekend pass would be a negative investment under S2 as we explicitly know that these customers do not have a particular preference on weekends.

The third and fourth scenarios capture an interest in using bike-sharing services regularly and differ on whether weather conditions influence a user or not.

Under S3, where this influence is relevant, A3 would be the best action to undertake. The regular need would be addressed by the monthly membership and the bother caused by the weather would be solved with the partnership with the public transportation services of the city. Again, monthly membership and carnet would partially satisfy customers, as there would be no benefits dealing with poor weather conditions. A carnet will be less effective than a monthly membership because it would be more costly. The weekend pass (both in S3 and S4) would cause negative revenues because it is different from what customers are looking for.

In the last scenario, S4, the monthly membership would be the best option. Users would prefer A3 to avoid bad weather conditions (because even if not influenced by meteorological conditions, they might still prefer to use a public service under these circumstances). However, from the perspective of the firm, it would be better to offer only a monthly membership since weather conditions are thought not to influence customers, making the partnership with WMATA an investment that might not be profitable. Obviously, A3 remains the second best because it still satisfies the needs of users by offering something more than they are asking. A carnet would also be a positive investment in this case, although less effective than the two actions just mentioned. Its proportional cost would be greater for a customer who makes regular use of the service. The only initiative that would be out of context and therefore of a negative return is the weekend pass.

To conclude, it is worth showing a possible caveat of our approach: some members (the ones with an annual membership) could decide not to renew their past subscription in favor of "*monthly membership*" or "*carnet*". *Is this a problem?* In our opinion, it is not. The reason is that in this case, they will pay more in proportion (annual members pay less than $8 per month, while our membership would charge more per month) and, given their commuting needs, they will probably buy the monthly membership more than two or three times. That is why we do not consider this as a "loss" for the company.

Step 6: Testable Hypotheses

In this section, we identify the hypotheses that will help us understand in which of the four scenarios we are located. In later sections, a survey will be used to test them. For each hypothesis, an explanation about how we collect data through the survey is provided. More details are outlined in the "*Data collection"* section and Step 7.

First, we want to test whether customers can be classified as regular users or not. Given the current links between actions and scenarios, we could summarize it as {S1, S2} vs {S3, S4}. In this setting, our null hypothesis is:

H_0*: the majority of customers uses the service in a regular way*

The idea is that to propose actions that are more limited in their possibilities (only for sporadic users), the proportion of sporadic users must be larger than regular ones. This is because the latter represents a constant income for the company, something preferable. On the other hand, promotions crafted for occasional users have a revenue stream of smaller magnitude.

Why do we set "regular use" as the null hypothesis? Because this is what was suggested by our analysis of the data frame where we found that there were more regular rides than casual ones.

In addition to this, there is a statistical reason behind this choice. We decided to design H_0 to identify a high frequency of use because of the rule of thumb that type I error (reject H_0| H_0 true) is worse than type II error (fail to reject H_0 | H_0 false). In light of this, we select the null hypothesis in order to have the "worst-case" as a type I error. In this setting rejecting the null (i.e. equal to saying we are in "*seldom use*") when regular users are predominant (H_0 true) is far worse than the other way around, since we would propose a weekend pass or a carnet when customers are looking for a membership different from the annual one. On the contrary, in the case of identifying regular (not reject H_0) when it is actually seldom (H_0 false), we would craft a monthly membership. This scenario would not perfectly suit customer needs, but it could be somehow considerable given the high number of rides members perform (as shown in "*Data analysis*"). The only customers not satisfied would be tourists, but we argue that this service must rely on Washington's citizens to be stable over time and a day-pass is already offered for them.

How did we gather data to test this hypothesis? We provided the following multiple-choice question: "*On average, during any week, how often do you use a bicycle?* ", with four possible answers: "*Never*", "*Between 1 and 2 days per week*", "*Between 3 and 6 days per week*", "*Every day*". Our idea is to consider regular users all respondents who claim to use the bicycle at least 3 days per week. This threshold (more than 2 days) is provided only by our theory since we could not measure how many times per week customers use the service (given the fact we were not provided with users' ID). Nonetheless, we think the threshold is strict enough to identify *true* regular users.

Once we identified the nature of the respondent (sporadic vs regular), we directed them to different questions according to their type: "seldom" users were asked about their preferences towards weekends, while "regular" ones about their concerns with weather conditions.

Let us suppose we are in the *"sporadic setting"*, that is we rejected the null of being in the "*regular"* usage in the previous hypothesis. Now we consider a split given by an *exclusive or not exclusive interest for the weekend*. Therefore, the hypothesis to test is:

H_0*: the majority of sporadic users does not have an exclusive interest in weekends*

We choose "not exclusive interest for the weekend" because offering the weekend pass when customers do not have a strong preference towards the weekend (that is, reject H_0 | H_0 true) is far worse than proposing a carnet in the opposite setting. The former would mean proposing a service that is not necessary for customers' demand. The latter is an offer that does not perfectly match the exclusive interest for weekends, but does not prevent those users from using the carnet during weekend days.

How did we gather data to test this hypothesis? We provide the following multiple-choice question: "*On which days of the week do you go cycling the most?*" with four possible answers: "*Weekend*", "*Weekday*", "*Both*" and "*I never go cycling*". Our idea would be to drop the observations regarding the respondents who choose the last option ("*I never go cycling"*) since they are not even potential customers. Regarding the others, the second choice (*"Weekend"*) is the one which highlights an exclusive interest for the weekend.

Finally, let us consider the *"regular setting",* that is we fail to reject the first hypothesis. Now the focus is on understanding whether meteorological conditions play a role in the choice of using a bicycle or not. Consequently, we test:

H_0*: people are not influenced by the weather*

For this hypothesis, we change our way of setting the null: H_0 is the worst scenario. In the previous hypotheses, this approach could not hold because we were not able to state what would have been the worst setting between regular and seldom or weekend preferences. As a consequence, in the previous hypotheses, we decided to state H_0 to have as type I error the worst case to minimize.

As said, in this case, we use the rule of thumb to state H_0 as the worst scenario, meaning that we would like to reject it. In our mind "Customers are not influenced by the weather" is the worst because it would mean that the decrease in the company's revenue during fall would not be related to the weather, meaning that during that period there are other reasons affecting it. This could mean that there is a structural problem in the company or factors that were not considered.

How do we gather data through the survey to test this hypothesis? We ask respondents to provide a ranking of different means of transport (public transports, bicycle, walk, taxi, car/moped, scooter) from the most used to the least used in two different settings: first, assuming any given day; then, assuming they would be facing a day with bad weather conditions. The idea is to capture the different ranks of "bicycle" in the two standings to check whether the weather has an impact on its use.

To summarize, the tests are executed in the following way:

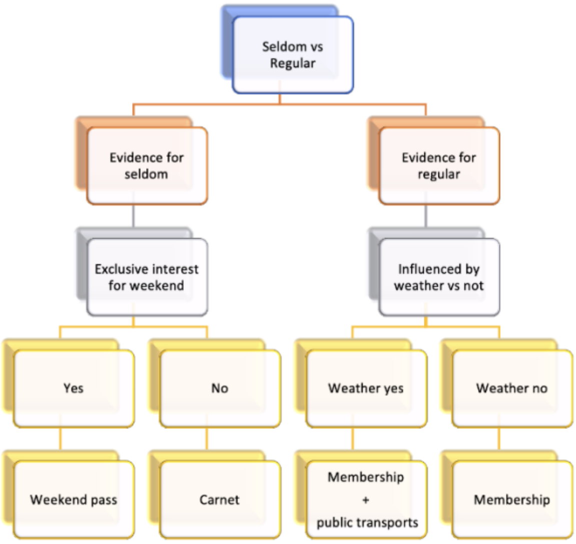


Figure 10: Hypothesis tree structure (top-down)

Data collection

Before moving on, it is important to show how we collected the data needed to test our hypothesis. We decided to create a survey[1] to capture the cycling habits of the respondents. *Who are the targets of our survey?* Ideally, if we were in Capital Bikeshare's position, we would like to collect answers from a pool of customers (i.e. both regular bicycle users and seldom ones) and potential customers. This could be done by running a survey on a random subset of customers and a random sample of Washington's population from those who were not already included in the first sample. Indeed, we do not restrict our target to just one type of user: our promotions span across different kinds of customers. Indeed, both members and casuals might potentially increase the revenues of the company.

Obviously, we were not able to reach the ideal sample, so we decided to share this survey on Reddit, especially in communities related to cycling in large cities (such as Washington

[1] Number of respondents > 1000, click here to view a pdf report

DC, Boston, Houston, Toronto, London, Milan, Singapore, etc…), trying to mimic the company's riders (both regular and sporadic ones). Furthermore, we decided to share the survey also with any possible individual with no specific traits, trying to capture hypothetical inhabitants of the city we are interested in.

Can we consider this as an appropriate sample of the people we would like to target?

Our survey cannot be considered a perfectly faithful representation of riders and non-riders living in Washington, but rather an attempt to obtain the most significant sample we could think of: we addressed people living in large cities like Washington is, especially looking for Northern American ones to consider possible common habits among American citizens.

Step 7: Test

As we saw in the previous part, the tests follow a hierarchical structure. First, we want to see whether we are in a *regular usage* or *sporadic usage* world. To verify this, we classified all those respondents who answered to use a bicycle at least 3 days per week as *regulars* and all the others as *seldom.* We built the following test: if the majority of respondents is classified as *seldom* (more than 50% is *seldom*), then we reject the null of being in a world of regular users. *Why do we use 50% as a threshold?* Because, as explained before, the survey was addressed both to a sample of people who showed a particular interest in cycling (members of cycling in large cities' subreddits) and a sample of people selected at random from Milan and other Italian cities. From the latter, we expect to observe a majority of sporadic users, but we expect the majority of the respondents to be regulars.

We used a one-tail t-test on the proportion of regular users, to test whether the proportion of regular users is below 50%, with a p-value to reject the null at 5%[2]. The result is shown in the appendix, First Test.

We fail to reject the null (which would be rejected if shown that the proportion of regular users is below 50%) since the p-value of the mean is lower than 50% is equal to 0.8257, significantly higher than the p-value we set before. So, the first test leads us into the branch of regular usage.

At this point, we want to test whether regular users are affected by bad weather conditions. Rain, snow, and low temperatures are among the bad weather conditions we imagined a citizen of Washington might encounter.

To test this, we verify if the average position of "bicycle" changes between the two rankings (*usage during any day* vs *usage in bad weather conditions*). We should observe a difference

[2] This percentage is the canonical alpha for tests. Depending on the context, one could argue that it is not appropriate. Nevertheless, being that bike riding is not a particular setting such as clinical trials / any risk assessment procedure, the usual cutoff is a way to standardize the procedure

between the two average positions equal to zero if bicycle users were not influenced by the weather (the measure to be tested is "*average position in the first ranking - average position in the second ranking*). As before, the results are in the appendix, Second Test.

We can immediately see that the difference is statistically different from zero, and indeed it is negative, meaning that in the second ranking "bicycle" 's position dropped with respect to the first one.

Had the first hypothesis been rejected, we should have tested whether sporadic users show an exclusive interest in weekends. Our null hypothesis is that there is no difference between weekends and weekdays, and this is rejected if we observe that more than 50% of seldom users express a preference for weekends only using a one-tail t-test. Also, in this case, we would have used the classical 5% p-value to test whether the null can be rejected or not.

To conclude, we showed through these tests that we are in the *"regular usage"* setting, where cyclists change their bicycle usage with bad weather conditions. Hence, we can say we are in scenario S3.

Step 8: Make a Decision

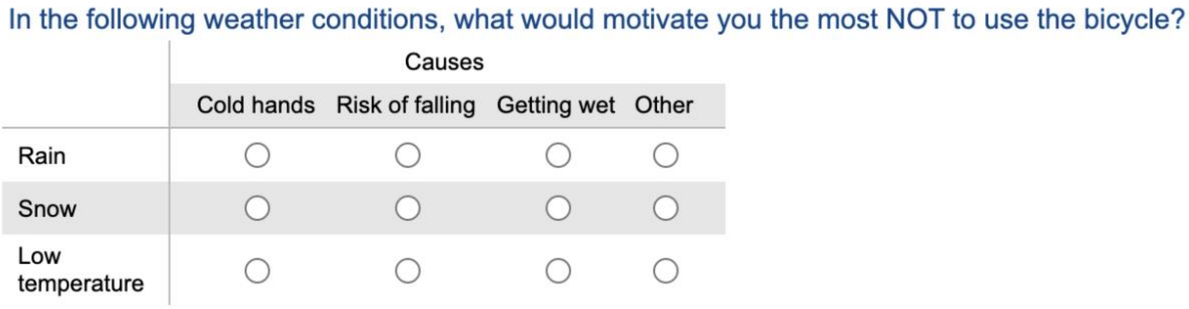
The results provided by the tests guide us to propose a monthly membership plus a partnership with the WMATA. However, the tests we ran did not make an explicit reference to the position of Public Transports in the two rankings. So, to strengthen the development of this promotion, we developed an additional test that shows that the true needs of regular cyclists are in line with our promotion. The test follows a similar structure to a Differences in Differences approach. We want to measure how the position of public transports changes with respect to the bicycle's one in the two rankings. So, the measure we want to test is the following:

DiD = (average position of the bicycle in the first ranking – average position of public transports in the first ranking) – (average position of the bicycle in the second ranking – average position of public transports in the second ranking)

If this difference is equal to 0, then it means that we do not observe any variation in public transport usage with respect to the difference in bicycle usage. If this difference is positive, public transports lose positions with respect to the bicycle in the second ranking. Instead, if this difference is negative, it means that public transports have gained positions with respect to the bicycle in the second ranking. In the appendix, Additional Test, we provide STATA's output.

The “DiD” measure is statistically different from zero at the 1% level, meaning that there has been a change in the relative rankings in the before-after comparison. Moreover, it has a value equal to -1.98, meaning that on average the two means of transport have swapped their positions in the two rankings (in fact, imagine having bicycles placed first and public transport second in the first ranking. This means a relative change equal to: *bicycle – public transports = -1*. Then imagine the two means of transport swap their positions, with public transports first and bicycle second in the second ranking. Then, the relative change in this ranking is: *bicycle – public transports = 1*. Computing the difference between the two relative changes we obtain: *-1 - 1 = -2*). This is a nice result which makes us think that the promotion appropriately suits riders’ habits when they do not have the possibility of using the bicycle. But there is more. We noticed a correlation between bike-sharing services’ usage in (the survey we also asked respondents to say whether they had ever used bike-sharing services) and the position of public transports in the ranking. So, we ran two ordered logit models (results in the appendix, Ordered Logit), one for each ranking, with the position of public transport as a dependent variable and a dummy variable indicating bike-sharing service usage as our variable of interest. We then added age, profession, and gender as control variables. Interestingly, those who claimed to have already used bike-sharing services had a higher probability of placing public transports in the first positions, and the marginal effect was even higher in the ranking of the bad weather conditions scenario. These two tests give us the confidence to say that a promotion of this type would be highly appreciated by Capital Bikeshare’s customers who use the service with regularity, prefer some more flexibility as opposed to the annual membership, and are concerned with bad meteorological conditions.

After this analysis confirmed our promotion “Monthly membership + public transports”, we would advise investigating more about the reasons that push people to use the service with less frequency during fall or on days with bad weather conditions. In this way, we might be able to understand the *mechanism* behind this phenomenon. *Why is this interesting?* Because it might reveal something useful for the company, such as considering the possibility to add different facilities for those users who do not have easy access to public transports or are not interested in using them. We tried to analyze the reasons behind the inability to use the bicycle during bad weather conditions with a quick question in our survey:



In the following weather conditions, what would motivate you the most NOT to use the bicycle?

	Causes			
	Cold hands	Risk of falling	Getting wet	Other
Rain	○	○	○	○
Snow	○	○	○	○
Low temperature	○	○	○	○

Figure 11: Survey question to explore additional factors

Our idea is that the main reason for not taking a bike during a rainy day is because of *"Getting wet"*, while during a snowfall would be the *"Risk of falling"*. Finally, considering that some bike-sharing systems already offer a system for protecting their legs, we thought about the same problem with hands. In this way, we identified "Cold hands" as the main reason behind not taking a ride during low-temperature periods.

For each weather condition, we also provide the "Other" option to cope with possible problems we are not considering.

For each hypothesis, we may think of an action as "*improve our promotion*" that suits these needs. As a consequence, a simple scenario-action map can be designed as follows:

	Getting wet	All the others
Improving our promotion with additional facilities related to the problem of getting wet	+++	-
Do nothing/pivot	0	0

Table 2: Bonus Scenario Action map

Here we propose as an example only the scenario-action map related to "Getting wet" during rainy days (the others can be created following the same logic). If the scenario "All the others" would be verified, we would suggest the company try to investigate more about the reasons behind it, i.e. "*pivot*".

We continue with the example of "getting wet" during rainy days. Now we identify the hypothesis we want to test:

H_0*: getting wet is the main problem during rainy days*

By testing this hypothesis with a simple t-test (more options in the appendix**,** Running t-tests), we find that there's statistical evidence for “getting wet is the main problem during rainy days”.

Consequently, what we would do in this setting is to add some facilities to the current promotion, that is the first action of the previous scenario action map. For instance, we might think about offering a free k-way. At this point, we would run an experiment to see whether it is worth introducing this new facility to the current promotion. Adding this benefit would represent additional costs for the company and it is important to understand if it is a “worthy cost”. In this setting, a simple A/B test could be enough. The idea is to create two different web pages where people are randomly directed while checking the company’s website: one page will show the “membership + Public Transports + free k-way”, another one all but the “free k-way option”. Given this setting, a comparison of the difference in the number of purchases of the “monthly membership + Public Transports” option between the two groups should show whether the “k-way” option has a positive impact. A possible way to test its impact would be running a probability model, such as a probit or logit, on the share of membership sold: our variable of interest would be a dummy indicating whether the website visitor saw the k-way offer or not. If the coefficient related to our variable of interest is positive and statistically significant, then we could think of integrating the additional facility in our promotion.

What we are suggesting to Capital Bikeshare is to create different customers’ profiles and consequently craft promotions that suit their needs. The idea is that through these kinds of analysis they would be able to add some facilities to current ones to make them more attractive. *How would the company proceed?* Assuming that the company has in its hands these data (in the appendix**,** Lack of data**,** we discuss options when this is not feasible), the procedure to follow is quite straightforward:

1. Think about different actions according to different scenarios (i.e., different customer profiles or different reasons for not taking a ride with bad weather conditions).
2. Run different tests (t-test, regressions, ...) in order to see whether there is evidence for a specific scenario, exactly in the way we have done previously.
3. If there is evidence for that scenario, run an experiment to check whether the additional facility has an impact on the company’s revenue.

   If there is no evidence, pivot the theory and think about other possible reasons (similar to the approach to the “cold hands” option).
4. If the experiment shows the promotion has a positive effect on the company’s revenue, launch it.

Limitations

In terms of limitations, we will briefly list what was not considered and what could have helped strengthen our beliefs:

- IDs of riders are missing
  - No information on individuals, rather only on their rides
- It is assumed throughout the process that a member is more profitable than a casual customer
  - This is not in general true, but nevertheless works as a reasonable assumption
  - There could be cases in which a casual customer ends up paying more than a member, once adjusting for the time component of payments. Any kind of payment is preferred in advance by definition, but *what if some customers end up paying way more with time?*
- The survey sample is not proved to be representative of Washington's population or Capital Bikeshare's customers
  - This is indeed not possible to prove unless the firm provided us with real information
  - Despite this fallacy, the survey was thought of in a way such that it would have applied to any sample
- Cost proposals and financial analysis were not viable
  - No information about these choice contributors was provided, nor asked
- The final experiment was only designed and not carried out
  - The feasibility of the experiment on the field was not an option
  - A potential experiment was proposed, choice makers would presumably decide whether to carry it out, controlling for financial constraints and opinions on its stability across time and states of nature

Conclusions

Across the document, we attempted to develop a sustainable growth option for Capital Bikeshare. The result was achieved thanks to a well-defined number of steps following the scientific approach for business decisions.

This method relied on rational operations and mindful considerations, which implied the necessity to explore further some aspects of customers' needs. Eventually, the result of such exploration was verified in a partition of the real world that identified pairs of scenarios and actions arranged in hypotheses. Strong evidence in favour of some or others in terms of data then helped isolate a potential offer for customers.

Given the structure of our reasoning, proposing a monthly membership and a partnership with Washington’s public transport service is not the only option, but rather is the only action featured with statistical evidence of benefit. Differently from the basic scenario or the other possible ones, information gathered with databases and a survey supports its validity.

While it is not ensured that such a choice would be profitable, a simple experiment, once designed correctly, would return such an answer.

In an ideal setting, once the company identifies a somewhat comparable city as Control (it could be exploited the fact that the company offers the service also in other cities), it would be possible to assess the effectiveness of the offer for a whole winter season in Washington. Controlling for potential heterogeneous factors across the two cities, the procedure would just require administering the offer to DC’s customers only and observing any differences across the two groups in terms of absolute numbers, profits, cash flow, or any variable of interest. How this process of causal inference is to be carried out is highly dependent on the features of the two groups and cannot be discussed thoroughly in a paragraph. From Naive estimation to very subtle refinements, there are plenty of methods to assess the potential of our proposal. However, given the setting and the type of data that could be gathered, a Difference-in-difference approach might be the most appropriate one.

We recognize that with a more heterogeneous set of information coming from different relevant parts of a business problem, the solution might have changed. At the same time, we claim that the whole reasoning, once constrained on scarce data, is robust and well structured. Moreover, limitations’ awareness is a pivotal indicator of further improvements, thus making the whole document a thoughtful starting point for a stronger analysis.

If this proposal was considered enough in terms of evidence, it would be ready to be enacted in real life, with a clear and powerful process, eventually requiring the gathering of additional evidence. This new information would then lead to the estimation of the real effect of the offer, returning knowledge about its effect.

Even in the worst-case scenario, the limitations listed above would help understanding how to pivot the project towards a more appropriate setting.

Sources

Throughout the document, we did not exploit any work / publication or reference worth mentioning. The methods used can be found in any introductory Machine Learning / Regression book. In terms of Programming, this applies as well.

For what concerns the line of thought, we followed guidelines provided across the course of Business Analytics at Bocconi University (course page), of which this work is part of the evaluation.

We would also like to mention Hfarm, the company that provided us with both the challenge and the datasets (h-farm.com).

The Jupyter Notebook is available at this link.

The survey we designed with the help of Qualtrics is available at the following link.

Any other information extracted can be directly recovered with a simple Google Search of Capital Bikeshare's website, its competitors in Washington, or other cities.

Appendix

Introduction

Click here to go back

- Capital Bikeshare is operated by Motivate (Motivate.com), a company that cooperates with brands to enhance sustainability in large cities currently managing the largest bike-share systems in the US and the world. Every Motivate system is guided by a general director, who is responsible for daily operations. The general manager is supported by business intelligence tools and by a network of other managers in the bike-sharing sector. Motivate takes care of rebalancing regulation[3].
- The company also offers special memberships reserved for universities and corporations. For the former, an annual fee of $25 is asked where users only pay the usage fees once over 30 minutes of rent. The latter group is offered a double option: either only the company (full subsidy: $50) or both the company and the employee (partial subsidy: $25 + $25) can contribute, with customers still paying usage fees starting from the expiry of the first 30 minutes. In both cases, a 41% discount is applied from the retail price of $85.

Competitors

Click here to go back

- The fact that this service takes place in the US capital city is peculiar for American habits. In 2018 the city won the first prize in the competition "League of American Bicyclists" for bicycle transport policies and became the second US city for the number of people who travel to work by bike. These facts illustrate a trend of the rising consideration bikes are gaining as an alternative to the classic transportation options across the city. However, it is not only a matter of needs. A huge project has been implemented for the realization of a Coast-to-Coast bicycle lane 6000 kilometers long passing through twelve States. The project became real in 2019 and is estimated to be completed by 2040.

[3] That is repositioning bikes so that their availability matches customers' needs, customer service, and maintenance and repair

- Capital Bikeshare succeeded Smartbike DC, created in 2008 and closed in January 2011 due to low memberships and a limited number of bike rental stations. The launch of the new service required $5 million for planning and implementation costs, followed by an additional $2.3 million for the operating costs of the first year. Then more money has been invested to support the expansion. This led to the number of stations, jurisdictions, and bikes that Capital Bikeshare claims right now.
- The company was the largest bike-share system in the US till May 2013 when Citi Bike began its business in New York City. Nevertheless, it remains the main actor in Washington DC beating the competition arisen, in particular, by four dockless bike-share systems: Mobike, LimeBike, Spin, and the new entry Jump.
- Their service offers the possibility to park not necessarily to the station, but in any public place, or almost. Private properties, the Mall, the White House, and Capital Complexes for example should be avoided. In case the bike is not left in an appropriate parking place, an additional fee can be awarded to the last bike user.

Preliminary Data Analysis and Processing

Tables

Below two tables briefly describe the variables included. It is worth pointing out that some descriptions are simply taken by general knowledge definitions as the dataset is not provided with a proper explanation. More information can be found on the source websites.

Weather Dataset	
Variable Name	Description
Name	Placeholder, it is always “Washington DC, United States”
Datetime	measurement, in one-hour intervals. Day-month-year hour:minute:second format
Maximum Temperature	Maximum daily temperature in Celsius
Minimum Temperature	Minimum daily temperature in Celsius
Temperature	The temperature at the hour of measurement
Wind Chill	The cooling effect of wind blowing on a surface

Heat Index	an index that combines air temperature and relative humidity, in shaded areas
Precipitation	Rain quantity index at the time of measurement
Snow	Snow quantity index at the time of measurement
Snow Depth	Snow depth index at the time of measurement
Wind Speed	Wind Speed
Wind Direction	Wind Direction (polar)
Wind Gust	Sudden, brief increase in speed of the wind indicator
Visibility	Visibility index
Relative Humidity	Humidity percentage
Conditions	Categorical, stores classes of general weather conditions such as: Snow, Rain, Overcast, Clear, and some combinations

Capital Bikeshare Dataset	
Variable Name	Description
ride_id	The ID of the single ride recorded
rideable_type	Type of bike used, can be docked, normal, electric
started_at	Starting rental time in day-month-year hour:minute:second format
ended_at	Ending rental time in day-month-year hour:minute:second format
start_station_id	The ID of the starting station
end_station_id	The ID of the ending station
start_station_name	Name of the starting station
end_station_name	Name of the ending station
start_lat	Latitude of start

start_lng	Longitude of start
end_lat	Latitude of the final position
end_lng	Longitude of the final position
member_casual	Categorical indicator distinguishing members from casual users.

Data cleaning

Click here to go back

In this subsection, we reported a summary of the procedure that made the clustering meaningful and machine-executable. For a detailed explanation, please refer to the Jupyter Notebook attached.

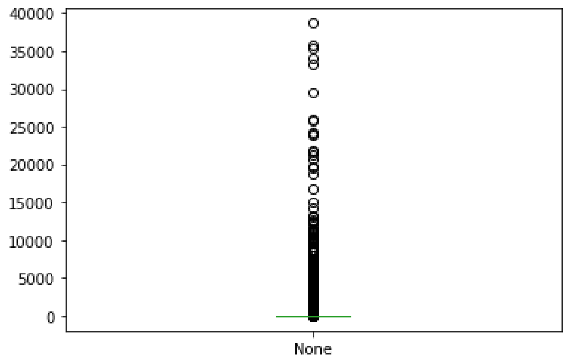
At first glance, the two distinct Dataframes present some mild problems related to Python’s interpretability, and columns with almost all missing values. These are quickly solved after changing some data types to machine-readable and dropping the columns, which would have been not very much informative anyway.

The merging operation is done in such a way that the hour in which the rental started is matched with its respective measurements in the weather datasets. Here we notice that it is often the case that rental durations go well over the Interquartile Range, presenting a right-skewed distribution that will be dealt with just before the algorithm. It is also worth noting that Capital Bikeshare considers any rental lasting more than 24 hours as “lost bike” or “stolen bike” so we could expect that those rentals should not be considered.

After merging the two, a general tidy_data function, that drops useless columns and removes or imputes missing values is applied. The result is a dataset with more than 450000 rows and 21 columns, which lost around 5% of the initial observations.

Additionally, after noticing that some columns were considered continuous while being categorical, we agreed to transform them to a categorical type for consistency.

As the last step, to avoid the outlier sensitivity problem of our algorithm, we combined drops and winsorization operations for the duration column "*Time_spent*". This proved indeed to be useful in terms of performance and shape of the clusters while losing again very few corner observations.



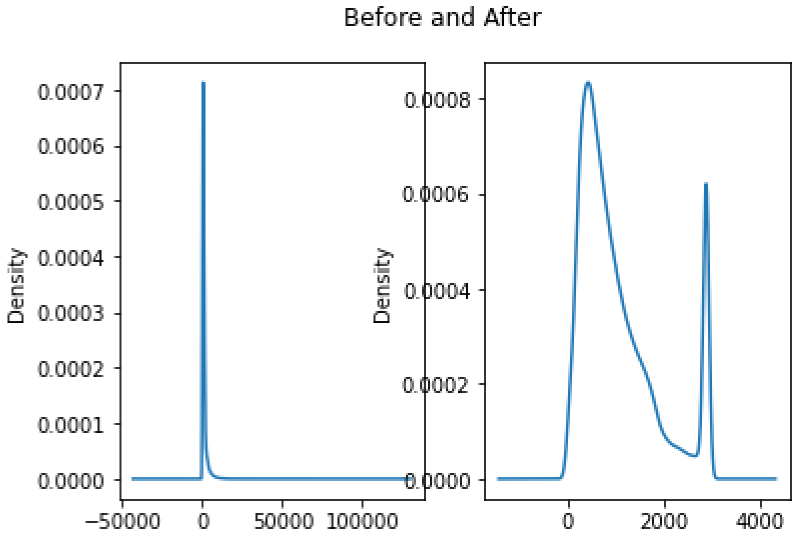


Figure 12: Above, Boxplot of the Duration of rides distribution. Circles are outliers

Figure 13: Density estimation of duration before and after dropping lost bikes and winsorizing

Plots & Basic Statistics

Conditions by Month Histogram

Click here to go back.

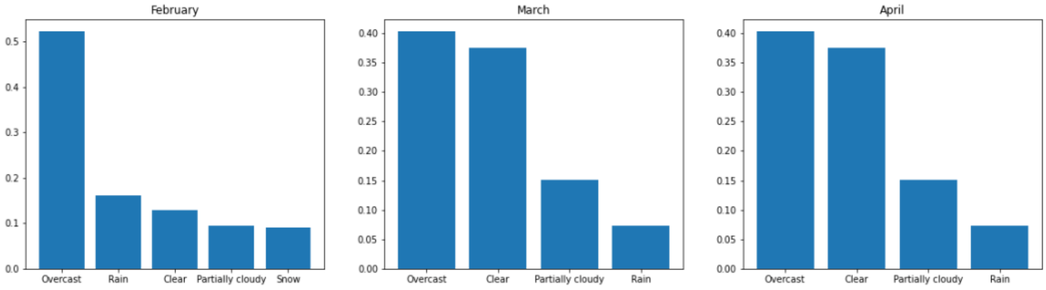


Figure 14: Rides conditions distribution grouped by month

Peak Hours Histogram

Click here to go back.

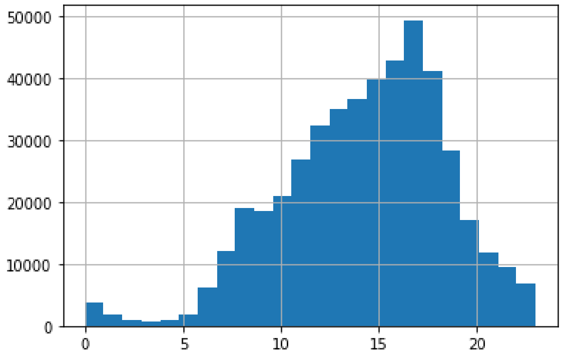


Figure 15: Starting hours distribution

Geographical Data

Click here to go back

Geographical data does not present unusual features. From single and aggregate plots of starting (right) and ending positions (left) there is a noticeable concentration around a center which is most likely matching with Washington’s. Having no particular flaw in its form, this piece of information could be stored for the algorithm and potential positional analysis. However, given that the joint distribution is mostly uniform, we expect little or no information from it.

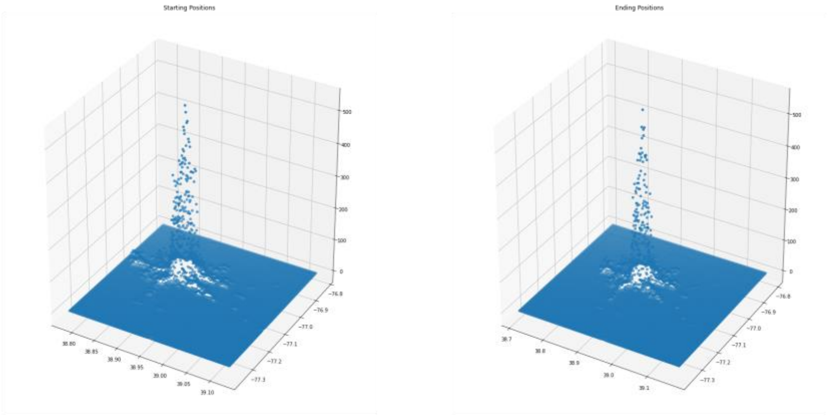


Figure 16: Joint distribution of starting and ending latitude and longitude

At first glance, it might seem that the distribution of starting and ending points is not uniform at all, as suggested by the marginal graphs proposed below. Latitudes and Longitudes appear to be very much concentrated on the same paired values, with a long-tailed distribution that probably spreads to the extremes of Washington. However, this reasoning is wrong in terms of what we wish to measure. Geographical positions are expressed in tuples, pairs of numbers that *together* convey a meaning in the imaginary grid, and thus their distribution is not 100% meaningful if analyzed coordinate by coordinate. In the corresponding document section, a joint plot is thus proposed.

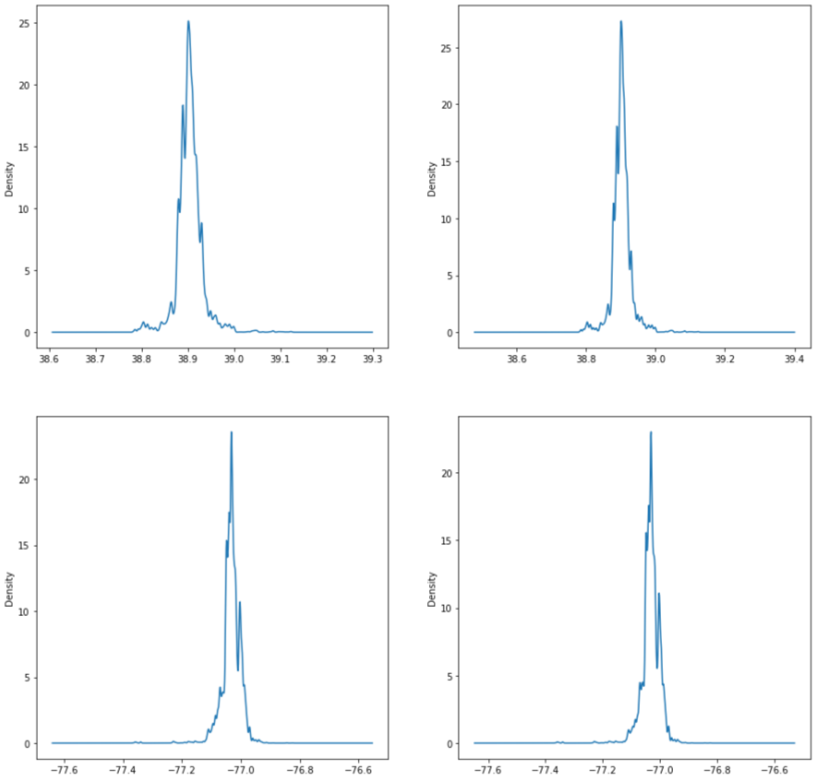


Figure 17: Disjoint distribution of starting and ending latitude and longitude

Time of rental Density Plot

Click here to go back.

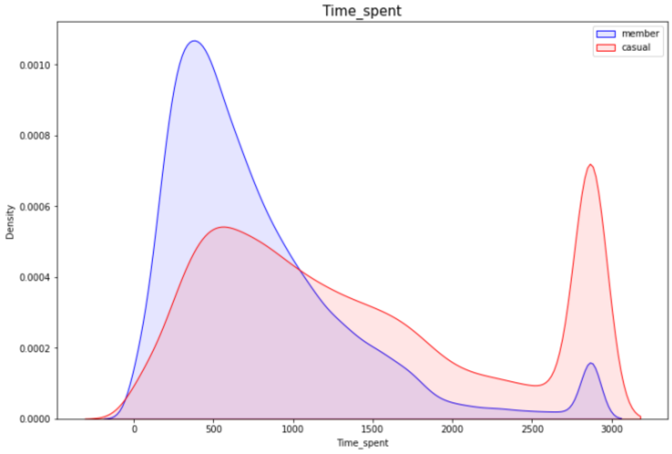


Figure 18: Type of customers' time of rental densities

Scientific approach

Pricing

Click here to go back

Determining these prices with the data in our hands is difficult. For instance, we cannot calculate the mean price spent by a casual member given that we have only data about single rides. Moreover, we do not know anything about the company's fixed costs. Therefore, what we can do is to compare the price of our company to the others in the US and make a simple proportion.

- ❖ *Weekend pass*:
  - ➢ Given that the daily pass costs $8/day, our weekend pass could have a price between 10$ and 12$ for the whole weekend.
- ❖ *Carnet*:
  - ➢ Given that the unlock for a bike would cost $1, casual rides last around 40 minutes and the price of a classic bike is $0.05/minute, on average casual riders would spend $1 + $0.05*40, that is 3$ per ride. Consequently, we think about a price of $20 for a carnet of 10 rides, which is around $2 for each ride: the mean price spent by casual without considering the unlock price.
- ❖ *Month Membership + WMATA*:

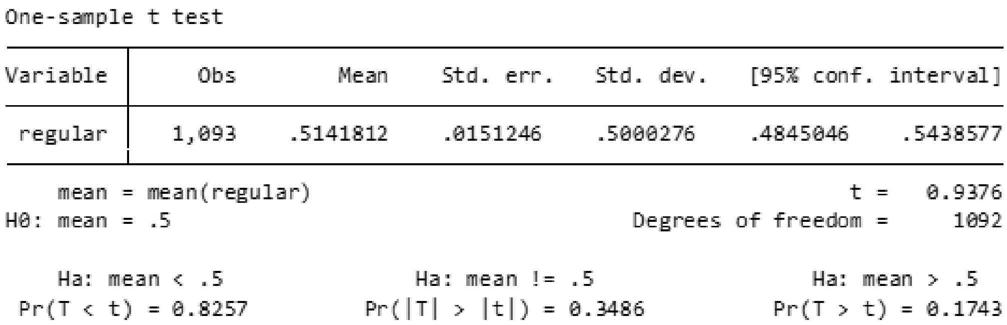
- Studying the relationship between monthly pass and annual one in the other US's bike-sharing companies, we have found that the first one costs 50% more with respect to the monthly price of the annual one (in many cases this membership is billed upfront). Capital Bikeshare proposes a price of $7.92/month for the annual membership, billed upfront. So, a monthly membership may have a price of around $12/month.
- Our promotion also has an additional facility: free use of WMATA's public transports during days with bad weather conditions or discount on the tickets. In the first case, in order to add the price of this to our promotion, we studied the mean price of a ride using these public transports: regular routes with metrobus costs $2, while peak fares (the price you pay for the service during a time when a lot of people travel) of metrorail is $2.25 - $6.00. We can think about a mean cost of $3.5-$4 for using public services. Moreover, we observed that the mean number of bad weather conditions during a month is on average 4-5 days (considering what we observed and that this may varies a lot across months), which would mean a cost of $20-$40 per month, considering also the return trip. As a consequence, we may think about adding $20 to our promotion. This would mean a final cost of around $32/month. This is just speculation because we should also consider WMATA's demands.

- *Month Membership*:
  - Based on what we said in the previous point, the price of this promotion would be around $12/month.

Step 7: Test

First Test

Click here to go back.



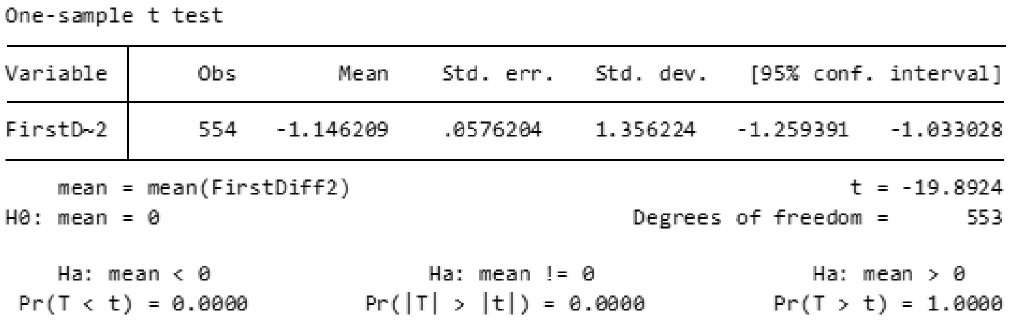
```
One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
 regular |   1,093    .5141812    .0151246    .5000276    .4845046    .5438577
------------------------------------------------------------------------------
    mean = mean(regular)                                          t =   0.9376
H0: mean = .5                                    Degrees of freedom =     1092

    Ha: mean < .5               Ha: mean != .5                 Ha: mean > .5
 Pr(T < t) = 0.8257         Pr(|T| > |t|) = 0.3486          Pr(T > t) = 0.1743
```

Figure 19: First test for use frequency

Second Test

Click here to go back.

This is what we obtain when we test the difference equal to zero with a two-tail t-test:



```
One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
FirstD~2 |     554   -1.146209    .0576204    1.356224   -1.259391   -1.033028
------------------------------------------------------------------------------
    mean = mean(FirstDiff2)                                       t = -19.8924
H0: mean = 0                                     Degrees of freedom =      553

    Ha: mean < 0                 Ha: mean != 0                 Ha: mean > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

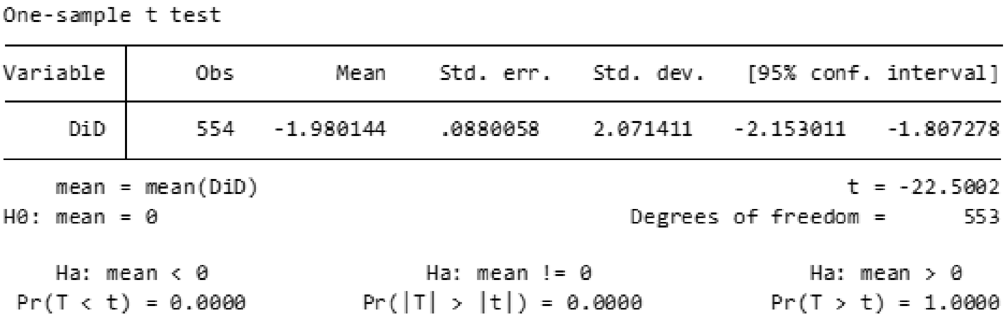
Figure 20: Second test for position change

Step 8: make decisions

Additional Test

Click here to go back

This is the test we performed:

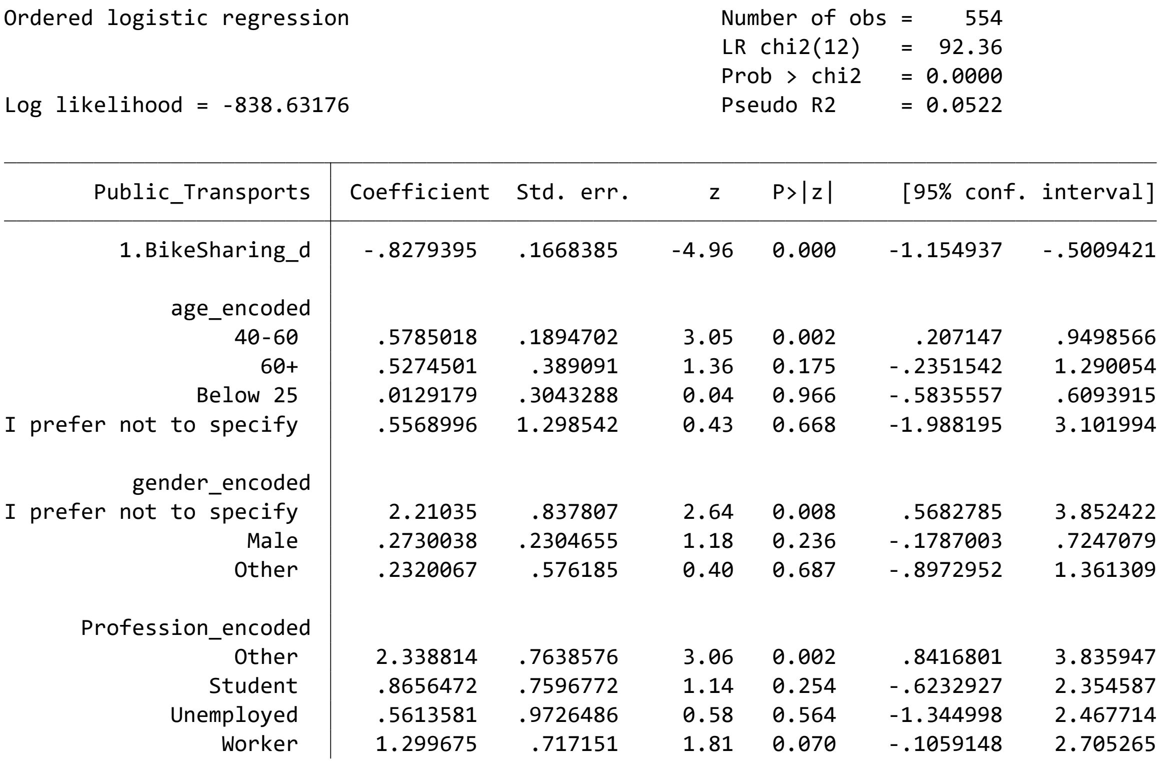


```
One-sample t test

Variable |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
     DiD |     554   -1.980144    .0880058    2.071411   -2.153011   -1.807278
---------+--------------------------------------------------------------------
    mean = mean(DiD)                                              t = -22.5002
H0: mean = 0                                     Degrees of freedom =      553

    Ha: mean < 0                 Ha: mean != 0                 Ha: mean > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Figure 21: Additional evidence DiD test

Ordered Logit

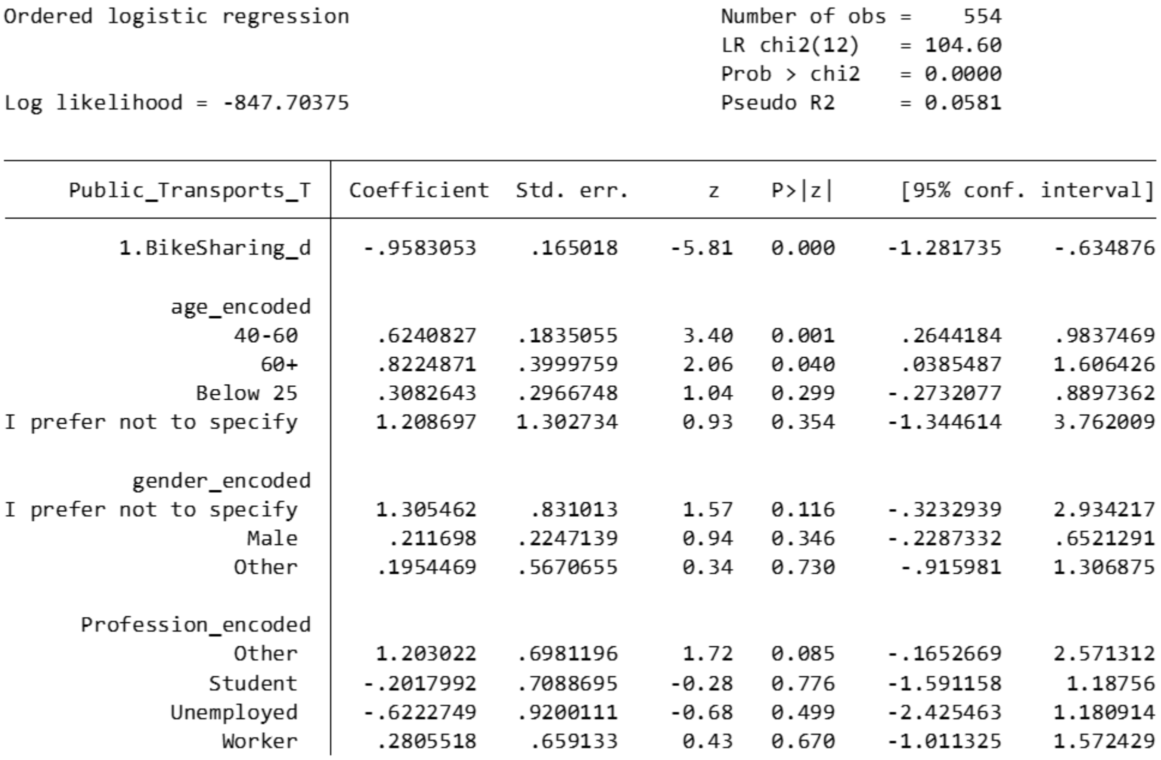
Click here to go back.



```
Ordered logistic regression                             Number of obs =    554
                                                        LR chi2(12)   =  92.36
                                                        Prob > chi2   = 0.0000
Log likelihood = -838.63176                             Pseudo R2     = 0.0522

--------------------------------------------------------------------------------------------
       Public_Transports | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------------------+------------------------------------------------------------------
         1.BikeSharing_d |  -.8279395   .1668385    -4.96   0.000    -1.154937   -.5009421
                         |
             age_encoded |
                  40-60  |   .5785018   .1894702     3.05   0.002      .207147    .9498566
                    60+  |   .5274501    .389091     1.36   0.175    -.2351542    1.290054
               Below 25  |   .0129179   .3043288     0.04   0.966    -.5835557    .6093915
I prefer not to specify  |   .5568996   1.298542     0.43   0.668    -1.988195    3.101994
                         |
          gender_encoded |
I prefer not to specify  |    2.21035    .837807     2.64   0.008     .5682785    3.852422
                   Male  |   .2730038   .2304655     1.18   0.236    -.1787003    .7247079
                  Other  |   .2320067    .576185     0.40   0.687    -.8972952    1.361309
                         |
      Profession_encoded |
                  Other  |   2.338814   .7638576     3.06   0.002     .8416801    3.835947
                Student  |   .8656472   .7596772     1.14   0.254    -.6232927    2.354587
             Unemployed  |   .5613581   .9726486     0.58   0.564    -1.344998    2.467714
                 Worker  |   1.299675    .717151     1.81   0.070    -.1059148    2.705265
```

Figure 22: Ordered Logit model A

The model above represents the coefficients of an ordered logit model, where the position of public transports is used as depended variable and the dummy “BikeSharing”, indicating whether the respondent has ever used bike-sharing services or not, is our variable of interest. We added other variables as controls. As we can see, the coefficient of the variable of interest is statistically significant and different from zero, with a negative value: this means that those who have used bike-sharing services above have also placed Public transports upper in their rankings.



```
Ordered logistic regression                             Number of obs =     554
                                                        LR chi2(12)   = 104.60
                                                        Prob > chi2   = 0.0000
Log likelihood = -847.70375                             Pseudo R2     = 0.0581

------------------------------------------------------------------------------------------
       Public_Transports_T | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
---------------------------+--------------------------------------------------------------
            1.BikeSharing_d |  -.9583053    .165018   -5.81   0.000    -1.281735    -.634876
                            |
                age_encoded |
                     40-60  |   .6240827   .1835055    3.40   0.001     .2644184    .9837469
                       60+  |   .8224871   .3999759    2.06   0.040     .0385487    1.606426
                  Below 25  |   .3082643   .2966748    1.04   0.299    -.2732077    .8897362
   I prefer not to specify  |   1.208697   1.302734    0.93   0.354    -1.344614    3.762009
                            |
             gender_encoded |
   I prefer not to specify  |   1.305462    .831013    1.57   0.116    -.3232939    2.934217
                      Male  |    .211698   .2247139    0.94   0.346    -.2287332    .6521291
                     Other  |   .1954469   .5670655    0.34   0.730     -.915981    1.306875
                            |
         Profession_encoded |
                     Other  |   1.203022   .6981196    1.72   0.085    -.1652669    2.571312
                   Student  |  -.2017992   .7088695   -0.28   0.776    -1.591158     1.18756
                Unemployed  |  -.6222749   .9200111   -0.68   0.499    -2.425463    1.180914
                    Worker  |   .2805518    .659133    0.43   0.670    -1.011325    1.572429
```

Figure 23: Ordered Logit Model B

The model above is equal to the ordered logit model seen before, with the only difference that our dependent variable in this case is the position in the ranking of public transport with bad weather conditions. Here the effect seems to be even stronger, meaning that with bad weather conditions those who claim to have already used bike-sharing services placed public transports at an even higher position than in the first ranking.

Lack of data

Click here to go back

In the case where the company has a lack of data, what we would suggest is to create a survey to gather this data and then to apply the previous four steps, as we have done previously. Another problem is to whom to address with this survey. There are three possibilities:

1. If the company believes that they need new customers, they should direct the survey towards people who do not use the service.
2. If the company believes that decreasing revenue is caused by a lack of frequency of casual customers, they should propose the survey to them. Actually, this is something we have been able to observe also from our analysis: the percentage of customers decreases a lot while considering February, the coldest month of the three considered by us.

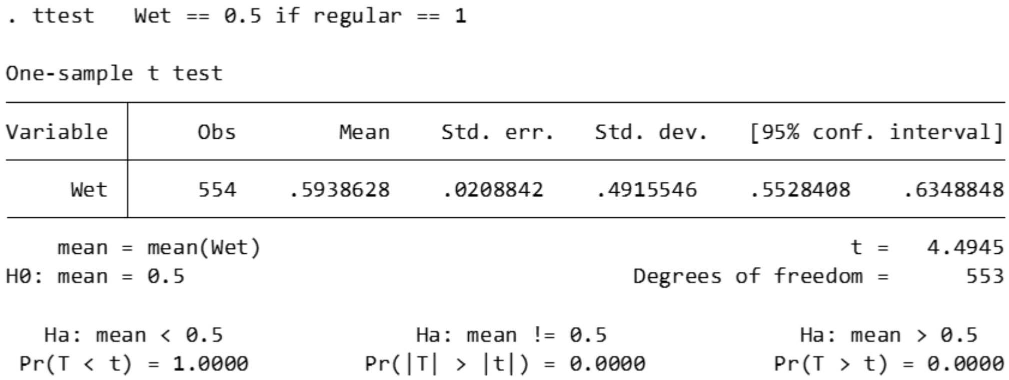
3. If the company believes this is a mixture of the previous points, that is both the need of new customers and the lack of reuse of the service by the casual, then they should provide the survey to both the groups. There should be an “if condition” in it to display questions according to the group they belong to. This is because questions regarding the service of the company can be provided to casual customers, while for the potential new ones there should be others.

After this, they would move on with the same four previous steps.

Running t-tests

Click here to go back

We ran the following hypothesis test:



```
. ttest   Wet == 0.5 if regular == 1

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
     Wet |     554    .5938628    .0208842    .4915546    .5528408    .6348848
------------------------------------------------------------------------------
    mean = mean(Wet)                                              t =   4.4945
H0: mean = 0.5                                   Degrees of freedom =      553

   Ha: mean < 0.5              Ha: mean != 0.5               Ha: mean > 0.5
 Pr(T < t) = 1.0000       Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Figure 24: t-test for getting wet

Furthermore, we can do the same also with the “snow” and “low temperature”. First of all, we define the two hypotheses we want to test:

1. H_0: falling during snowfalls
2. H_0: cold hands for low temperature

As before, we test these hypotheses through two different t-test, one for each weather condition. We find there is statistical evidence for the “risk of falling” regarding snowfalls. Consequently, what we would do in this setting is to add some facilities to match what we have found. For instance, we may think about proposing a discount for falls insurance or a partnership with insurance companies.

Regarding the third hypothesis, there is no evidence for “cold hands” during low-temperature days. Consequently, what we would do is to create a more specific survey to understand which are the reasons behind not moving during those days. Presumably, there is something hidden in the “Other” option that we are not considering and so it is important to get more insights about it.

An example would be: if the company finds that potential customers are sporty people, they could think about adding some coupons to spend in selected sports shops. *How could they find that a person is sporty?* Through an analysis of better-informed datasets (i.e. containing customer IDs that would allow creating consumers identikits).

The objective of these two additional pieces of analysis (investigation of the reasons behind not taking a ride with bad weather conditions and identification of customers) is to add something different to the promotions. In this way, customers will have to choose not only between different memberships or passes according to “time duration”, but also regarding other specific needs or additional facilities.