H-Farm Project

Bottani Sophie
Chiarini Emanuele
Gatteschi Giulia
Giannelli Enrico

20595 – Business Analytics

Academic Year 2021/2022

Table of Contents

Data preparation

To define our new business strategy we used Capital Bikeshare's dataset which contains information regarding rides made by users in Washington DC. We also used the Visual Crossing dataset which collects all the information regarding the weather in Washington DC by day and hour. We considered the data for the months of February, March and April 2021 to have some variability with respect to the weather. The Visual Crossing dataset contains 2,112 observations, one for each hour of each day, and 17 columns:

Name	Type	Description
`Name`	STRING	City of weather condition. In this case only Washington DC
`Date time`	STRING	Date and time of weather condition
`Maximum Temperature`	FLOAT	Maximum temperature measured at that time in °C
`Minimum Temperature`	FLOAT	Minimum temperature measured at that time in °C
`Temperature`	FLOAT	Average temperature measured at that time in °C
`Wind Chill`	FLOAT	Wind's cooling effect in °C
`Heat Index`	FLOAT	Index that combines air temperature and relative humidity to give a measure of how hot it really is. Scale starts at 26°C
`Precipitation`	FLOAT	Amount of precipitation of rain in mm
`Snow`	FLOAT	Amount of precipitation of snow in cm
`Snow Depth`	FLOAT	Amount of snow on the ground in cm
`Wind Speed`	FLOAT	Speed of wind in km/h
`Wind Gust`	FLOAT	Sudden brief increase in speed of wind measured in km/h
`Wind Direction`	FLOAT	Direction of the wind
`Visibility`	FLOAT	Measure of the distance at which an object can be clearly discerned in km
`Cloud Cover`	FLOAT	Fraction of sky obscured by clouds
`Relative Humidity`	FLOAT	Percentage of water vapor present in the air

Conditions	STRING	Summary of weather conditions. Can be: overcast, snow and overcast, rain and overcast, partially cloudy, clear, snow, rain, partially cloudy

We first analyzed and prepared the Visual Crossing dataset. Since the `Date time` variable is of string type we created two new variables `date` and `time` which represent respectively the date and time of the weather condition. These two variables are of Datetime type which makes the manipulation of time dependent variables easier and more efficient. By inspecting the data we notice that there are many missing values for the `Wind Chill` (1047 NaNs) and `Wind Gust` (1566 NaNs) variables, which however in this case represents a phenomenon rather than an error. For example, a NaN for `Wind Gust` means that at that time there weren't any sudden increase in the wind speed. Therefore, we replaced the missing values with 0s. `Heat Index` has 2086 missing values out of 2112 observations. This is reasonable since we are considering the months of February, March and April and the heat index has a scale that starts at 26°C. In this case we decided to simply drop the column since the heat index can be inferred from the `Temperature` variable. Finally we also drop the `Minimum Temperature` and `Maximum Temperature` columns since for the same hour they are equal to the `Temperature` variable.

The Capital Bikeshare's dataset contains 2,598,241 observations and 13 columns, and these are:

Name	Type	Description
ride_id	STRING	Unique string identifying each ride
rideable_type	STRING	There are three possible types of bike: classic, electric and docked
started_at	STRING	Starting day and time of the bike ride
ended_at	STRING	Ending day and time of the bike ride
start_station_name	STRING	Name of the station where the ride begins, if any
start_station_id	INT	ID of the station where the ride begins, if any
end_station_name	STRING	Name of the station where the ride ends, if any
end_station_id	INT	ID of the station where the ride ends, if

		any
start_lat	FLOAT	Latitude coordinate of where the ride begins
start_lng	FLOAT	Longitude coordinate of where the ride begins
end_lat	FLOAT	Latitude coordinate of where the ride ends
end_lng	FLOAT	Longitude coordinate of where the ride ends
member_casual	STRING	Specifies whether the user is a member (Annual/30-Day pass member) or a casual rider (single trip, 24h pass, 3/5 Day pass)

By analyzing the Capital BikeShare dataset we notice that there are some missing values. In particular, there are 4809 observations that are missing a coordinate of the end of the ride (i.e., they are missing either the `end_lat` or `end_lng`) and 2 observations that are missing a coordinate of the starting point (i.e., they are missing either the `start_at` or `start_lng`). Since they represent a small portion of the dataset we decided to simply remove them. It can be seen that also `start_station_id` and `end_station_id` have some missing values, which in this case are probably due to the fact that the bike was not parked at a Capital Bikeshare station. Therefore, we replaced the missing start or end station with “undocked”.

Then since the variables `started_at` and `ended_at` are strings representing both the date and time at which the ride started/ended, we created four new variables: `start_date`, `end_date`, `start_time` and `end_time`. They represent the start/end date and the start/end time of the ride. These features are of Datetime type. We also added a variable `start_hour` which is equal to the hour at which the ride started. This was done so we could merge the data with the weather dataset on the hour the ride started so that we could view the weather conditions of the ride. Moreover, we created a `duration` variable which indicates the duration of the ride. By analyzing this new variable it can be seen that the average duration of a ride is 18:23 minutes and that there are some outliers in the dataset. In fact the minimum duration of a ride is of -21 days, which is clearly an error, while the maximum value is of 58 days, which is probably again an error. Therefore, we removed the outliers by looking at the different quantiles and kept the rides with a duration between the 0.05 and 0.999 quantiles which have reasonable values of respectively 2:56 minutes and 17

hours and 16 minutes. In fact, we reckon one could use a bike either to get quickly somewhere close, especially if a member, or to go sightseeing around the city for up to the whole day (likely for no more than 18 hours).

Then, we created the variable `distance` that computes the air distance between the starting and ending point. Since rides that have the same starting and ending station had values very close to zero, we set the distance of those observations equal to zero by approximation. By further analyzing the data we notice how many rides have a distance of zero but a long duration, meaning that these aren't outliers but that many ride a bike for a while, maybe to run an errand or visit the city, and then park the bike at the station where they started.

To get better and more informative clustering results in the following steps we applied a label encoder on the variables `rideable_type` and `casual_member` to have a different integer value for each category. Then we created a variable `duration_sec` which represents the duration of the ride in seconds to make the following operations easier. Finally we created the variable `weekday` which indicates on which day of the week the ride started.

Clustering

There are many techniques that we could use to cluster our data. For our clustering we decide to use a k-means algorithm. This type of clustering partitions the data into k clusters so as to minimize the inertia or within-cluster-sum-of-squares. This process is based on a measure of distance between the observations that, in our case, will be the Euclidean distance.

Formally, defining X our dataset, C_j the cluster j and having n_j as the number of data points in C_j we have that to minimize the inertia for C_j we do as follows:

$$\underset{\mu_j \in C_j}{\operatorname{argmin}} \sum_{i=0}^{n_j} (||x_i - \mu_j||^2)$$

Then, the optimization problem the k-means algorithm solves, calling C the set of all clusters (i.e. $C = \{C_1, C_2, ..., C_K\}$), is the following:

$$\underset{C}{\operatorname{argmin}} \sum_{j=1}^{K} \sum_{x \in C_j} (||x - \mu_j||^2)$$

Interestingly this optimization problem is not solved directly because of how computationally expensive doing so would be. Instead, the Scikit-learn implementation of this, the one we will use, follows a process equal to the one of a Expectation Maximization algorithm with a small, all-equal, diagonal covariance matrix.

In practice this translates to three steps:

- Initial step: creation of the initial centroids using, in our case, the k-means++ initialization scheme
- Assignment step: assignment of observations to the clusters of which the centroids are closest to them
- Updating step: creation of new centroids as the means of the data points assigned to each cluster

The last two steps are repeated until the difference between the new centroids and the old ones is lower than a threshold or a stopping criteria is reached.

An important aspect to consider before running the algorithm is the scaling of the data. Indeed, being this process based on distance, if we don't scale the data properly the features that have bigger values in absolute terms will be the ones that drive the choice of the clusters more. As we want to avoid this, we perform a min-max scaler on all the features but the dummies (as they already are bounded in $[0, 1]$). A min-max transformation formally entails:

$$X_{i,j}^{(Scaled)} = \frac{X_{i,j} - \min(X_j)}{\max(X_j) - \min(X_j)}$$

Where j denotes a feature in the data and i an observation. This scaling techniques ensures that all the features j have their min at 0 and their max at 1. Additionally, we will divide by 2 the longitudes and latitudes so as to give the starting point and the ending point (as a whole and not to their single coordinates) the same weight of the other features.

As k-means requires us to specify the number of clusters the data will be partitioned into it is important to note that we chose them using the elbow method based on distortion. Distortion is the average of the squared distances from the observations to the cluster centers of the clusters they're assigned to. To find k we run the k-means algorithm repeatedly, each time increasing the number of clusters from 1 to 14. Then, we choose the number of clusters from

which the distortion trend becomes linear or almost so (see Figure 1 for the results in our clusterings).

Having discussed all these theoretical points, we're ready to jump into the implementation of the algorithm. We select a few features from our dataset on which to carry out the clustering process. In particular, we include `start_hour`, `weekday`, `member_casual_enc`, `Precipitation`, `Snow`, `start_lat`, `start_lng`, `end_lat` and `end_lng`. Running this first k-means we see how varied the different 8 clusters tend to be in terms of `start_hour` and `weekday` (see Table 1). Indeed, `weekday` is possibly the variable that changes the most, while those related to weather are very similar across clusters.

cluster	start_hour	weekday	member_casual_enc	precipitation	snow
0	14.810727	0.978842	0	0.023104	0.000779
1	9.379898	1.098204	1	0.045631	0.004504
2	14.250042	5.423992	0	0.012805	0.000469
3	14.866767	5.491965	1	0.032369	0.001730
4	14.953224	3.535300	0	0.017098	0.007454
5	17.282812	3.021910	1	0.041091	0.000645
6	17.197825	0.537447	1	0.008417	0.000570
7	8.859675	3.913729	1	0.065571	0.036877

Table 1: means of the variables of interest across clusters

We, therefore, decide to explore this variation and set up another k-means model taking into account only `start_hour`, `weekday`, `start_lat`, `start_lng`, `end_lat` and `end_lng`. Note that in this second clustering we use data for the whole year from November 2020 to October 2021 included (previously we only considered February, March and April 2021) to ensure our results are valid across seasons.

In this second clustering, cluster 0 immediately drew our attention. It is by far the smallest of the 8 clusters, spans rides in the middle of the night, almost only on Saturdays and Sundays, and it contains more casual rides than all other clusters (see Table 2). Given this small size, we might be able to increase this type of rides by introducing an offer that incentivizes them. The fact that these rides are mostly done by non-members also suggests this a potentially

interesting business opportunity as it means an increase in them would lead to an increase in revenues (or, potentially, in subscribers). To design an effective offer we first have to analyze why these rides are being made. Given the time of the day and the days of the week such rides occur at, it is likely that people are using bikes to move around nightlife spots.

		start_hour		weekday		member_casual_enc	
Cluster	**count**	**mean**	**std**	**mean**	**std**	**mean**	**std**
0	64381.0	1.801168	1.960109	5.127771	0.888774	0.470123	0.499110
1	290390.0	10.53412	2.521691	3.508891	0.499922	0.628486	0.483210
2	397084.0	17.32061	2.425293	0.503629	0.499987	0.605978	0.488640
3	401115.0	11.57136	1.903248	5.460227	0.498416	0.489426	0.499889
4	369600.0	18.10133	2.230353	3.519916	0.499604	0.586991	0.492375
5	392404.0	17.75938	2.315284	5.449646	0.497459	0.473280	0.499286
6	220382.0	16.90673	2.841537	2.000000	0.000000	0.623068	0.484619
7	325139.0	8.893722	2.651249	0.960408	0.775859	0.664611	0.472127

Table 2: summary of `start_hour`, `weekday`, and `member_casual_enc` across clusters

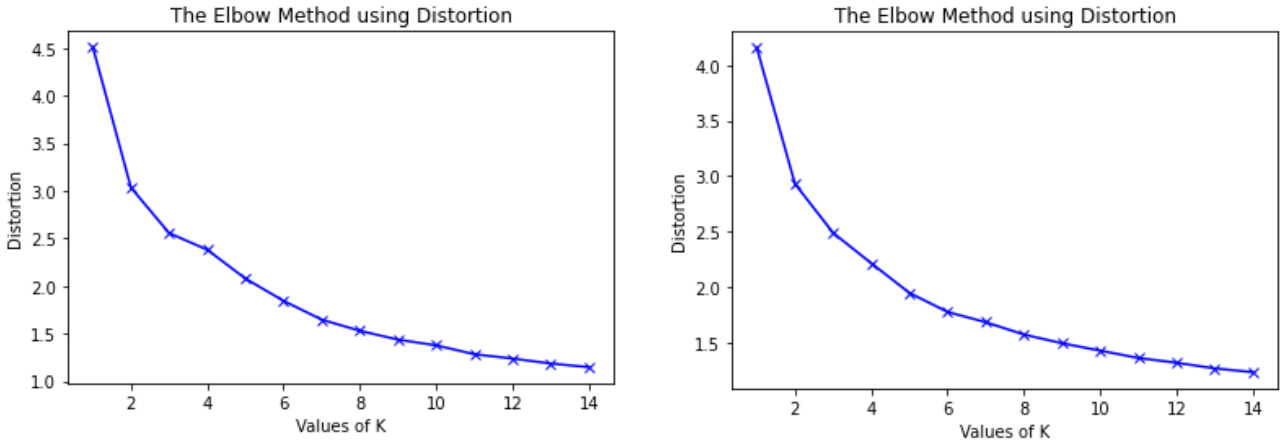


Figure 1: elbow method plot of the first (left) and second clustering (right)

Data visualization

In order to assess whether our intuition is correct, we visualize which areas are more popular in terms of starting points and routes using some heatmaps made with Plotly. We first plotted the starting points for all rides in general, and then we plotted the starting points for our cluster of interest, i.e. Cluster 0. In both cases we used a sample of 50,000 observations for computational reasons. As it can be seen in the figures below, the starting points differ when considering the whole dataset or only a cluster of it. In general the starting stations tend to be concentrated near the National Mall and Memorial Parks, location of the main monuments in Washington DC, such as the Lincoln Memorial or the Washington Monument. Many rides also start in the northern area of the city. However, when we look at the starting stations for rides made during the night, we notice how there are fewer rides and they are much more concentrated in the northern part of the city, with very few rides near the National Mall.

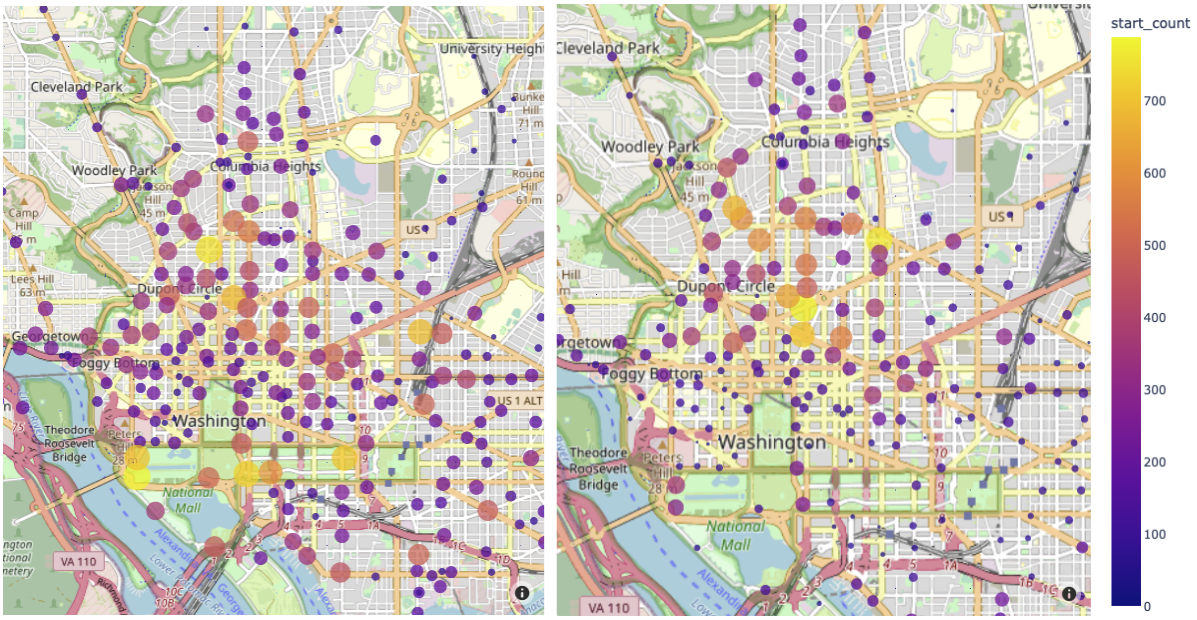


Figure 2: starting stations for the whole dataset (left) and for rides made during the night (right)

We also plotted the most popular routes as straight lines between the starting and ending stations. Even in this case, we first plotted the routes for the whole dataset and then for our cluster of interest. The figures below show how when considering the whole dataset, the most popular routes are along the National Mall. This could be because as it's a park it is more bike friendly, or because many ride a bike to see the most interesting monuments in Washington DC. When we look at the most popular routes during the night, the thick horizontal lines along the National Mall are, instead, not present anymore. It is clear that

rides are concentrated in the northern area of the city, and that there is a big difference compared to the rest of the data. If this area north of the National Mall is where most restaurants, bars and clubs are located we would confirm our intuition that rides at night are used to move around from/in the nightlife area.

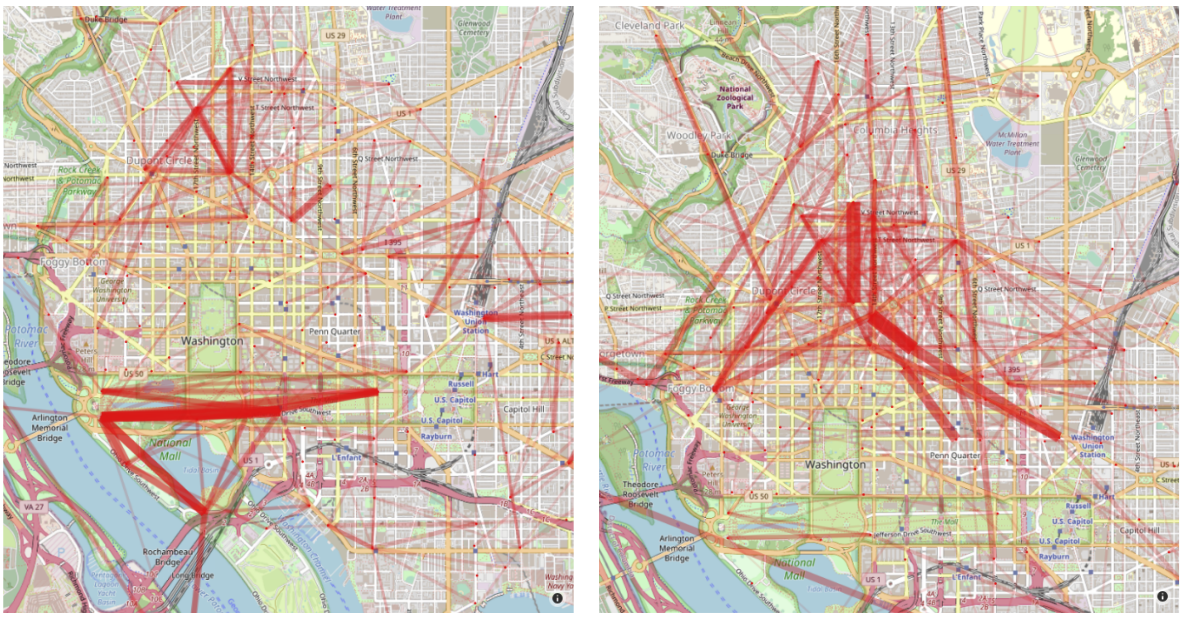


Figure 3: the most popular routes in the whole dataset (left) and during the night (right)

Nightlife locations in Washington DC

We, thus, looked for a data-driven way of confirming this hypothesis.

The Alcoholic Beverage Regulation Administration in Washington publishes online a list of all venues in Washington that own a license to sell alcoholic beverages, ranging from restaurants to clubs. We considered this publication a good proxy for the areas that are more populated during the night.

In the pdf file, besides the name of the venue and the status of the license, we can find its address and type (e.g.. Hotel, Tavern, Restaurant).

From the pdf file, we extracted the data to a pandas dataframe to perform better data manipulation. We then converted the address of each venue to its latitude and longitude so that we can better visualize the locations and perform additional analyses.

We then filtered this dataset keeping just the venues that were operating and dropped the one whose license was suspended. After plotting all the venues listed on a Washington map, establishment type by establishment type, we noticed that the majority of clubs (and hotels even if their role in the nightlife is really idiosyncratic) was indeed located in the zone that saw the vast majority of rides starting there during the night (between the National Mall and

Columbia Heights). Other types of establishment, like bars and restaurants, were instead more dispersed across the city. However, the bulk of them were again located in the center together with the clubs (see Figure 5).

Considering that the bulk of restaurants was located near clubs and taverns and that the role played by hotels in DC's may not be central, we decided to keep in the dataset just taverns and clubs. Keeping restaurants in the outskirts would have risked considering a consistent number of venues that likely have nothing to do with Washington's nightlife.

To separate all of the venues in identifiable areas, we run a K-means clustering on the data containing latitude and longitude of the taverns and clubs in the capital. The elbow method indicated an optimal number of clusters equal to 4, and we also found this number to be the most informative. The first, and most interesting cluster, was located in the center of the capital, north of the National Mall, in the area where most nightly rides originated from. The other three clusters were located in a peripheral position, one in the north, another in east and the last one in the south-east. All the results can be viewed in figures 4 and 5 below. Indeed, the amount of clubs situated in the central cluster was significantly higher, in both an absolute and a relative sense, compared to the other three. This confirms our initial intuition that indeed, the bulk of Washington nightlife is concentrated in the center and that, thus, the rides we witnessed are connected to it.

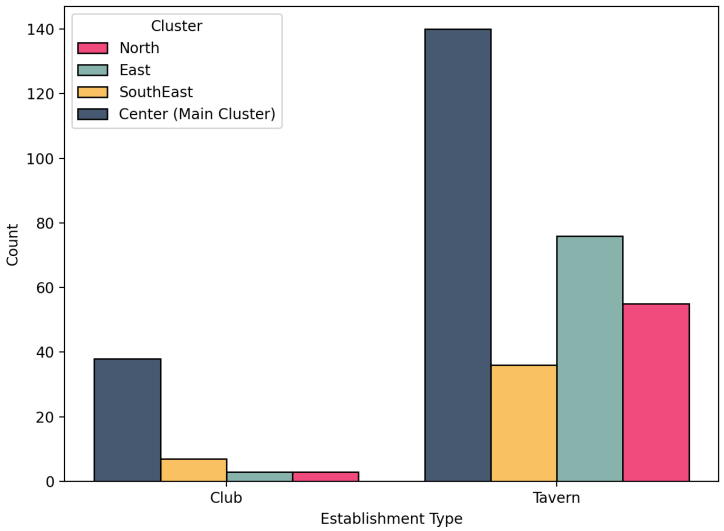


Figure 4: cluster composition for clubs and taverns

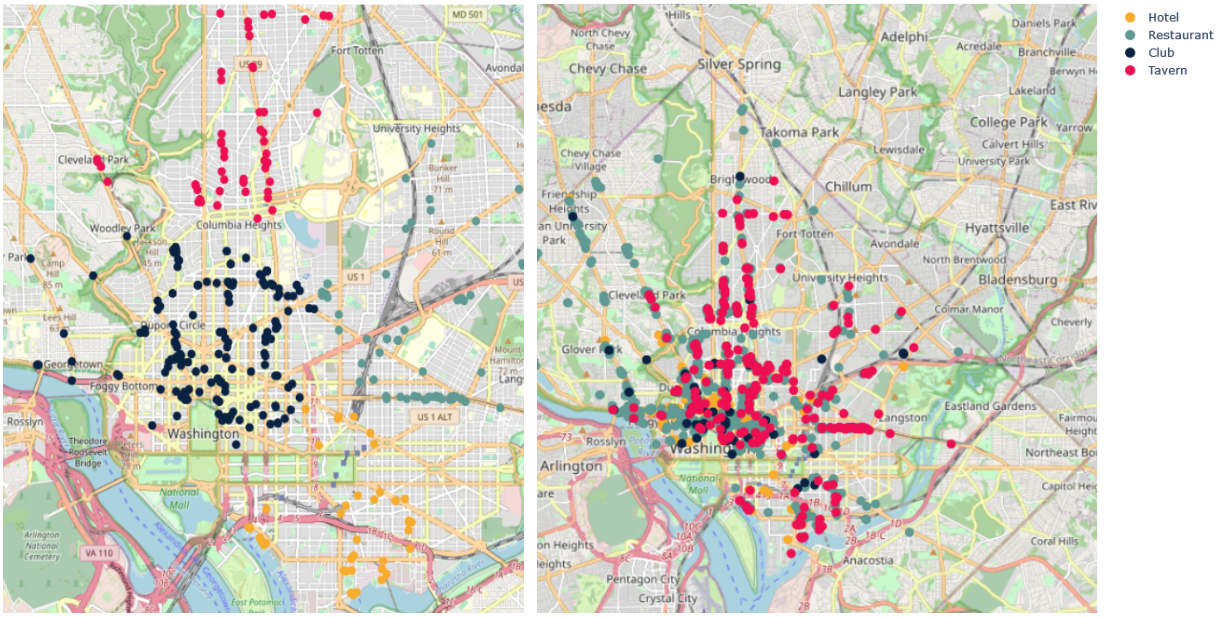


Figure 5: visualization of the 4 different clusters (left), location of hotel, restaurants, clubs and taverns in DC (right)

Business Theory

Having corroborated our idea that the few, mostly not taken by subscribers, night rides we saw in our clustering tend to happen around nightlife hotspots we can finally develop our theory.

In addition to the empirical results we obtained, we looked at some newspaper articles to further confirm our intuition that it is possible to stimulate night rides in the city. We found out the city is among the safest in the US (US News, 2020)[1], the majority of its citizens are not scared of roaming around during the night (dcist, 2013)[2], and that it ranks really high in bikeability among American cities (Yahoo! finance, 2021)[3] all aspects that make it very much possible to successfully push night rides.

Lastly, looking at the costs an increase of these rides would bring, we reckon variable costs for Capital Bikeshare's bikes are negligible and utilization of assets would not really impact profits. Managing to increase the utilization rate of Capital Bikeshare's bikes during the night is, therefore, likely to lead to an increase in profits also because, as we saw before, most of them are casual rides and, thus, are pay-per-use (not part of a subscription).

[1] US News. (2020). Washington, district of columbia; crime rate & safety. Retrieved from https://realestate.usnews.com/places/district-of-columbia/washington/crime

[2] dcist (2013). 72 percent of D.C. area residents say they feel safe at night. Retrieved from https://dcist.com/story/13/04/05/72-percent-of-dc-area-residents-say/

[3] Yahoo! finance. (2021). Most bike-friendly cities in America – 2021 edition. Retrieved from https://finance.yahoo.com/news/most-bike-friendly-cities-america-110029615.html

All this leads us to believe that there is an opportunity for Capital Bikeshare to better compete in the transportation market during the night.

As such, we tried to devise and test different ideas aimed at increasing the number of nightly rides. We came up with three different business proposals. Our first idea concerned a fixed-price bundle that would allow users to “book” a bike for the whole night and use it to move between clubs or bars.

Second, we thought of offering a free return ride during the night to encourage people to use bikes for their outward journeys. Importantly, this option would not include any restriction on the destination of the ride.

Third, we devised a discount on the first ride, for those that will use it to reach the main cluster of DC’s nightlife and leave the bike there. This option could potentially lead to both an increase in the number of rides and to an increased number of bikes available in the areas of the city where nightlife is concentrated. This, in turn, would make it easier to find a bike in these locations and potentially increase the utilization rates of Capital Bikeshare’s bikes even further. Indeed, instead of being scattered around the city, the bikes would be located where potential users aggregate during the night, increasing the likelihood that they will use them to return home.

We can summarize our theory in the following three scenario-action maps.

Prior: people would use bikes to come home from and go to nightlife hotspots if they were cheap and easy to take

	People are willing to use bikes for moving at night	**People are not**
Develop Idea	+++	-
Do not	0	0

In this scenario action map, we are assuming that the costs of developing the idea are almost negligible and close to 0 as, besides sending a discount notification to the Capital BikeShare app’s user and potentially a limited marketing campaign, this idea does not involve any variable costs. We will test in which scenario we are using 5 hypotheses and we are going to consider people as willing to bike at night if at least 4 out of our 5 hypotheses turn out to be true.

Prior: college students are the perfect target customers of the offer

	Service is more relevant for college students	Service is equally relevant for college students and the rest of the population	Service is more useful for rest of population
Focus marketing on students	++	+	-
Run unfocused marketing	+	+	+
Focus on non-students clients	-	+	++

The minus represents wasted additional ad spending on a non-responsive demographic.

Prior: one of the three offers (say X) is preferred to the others

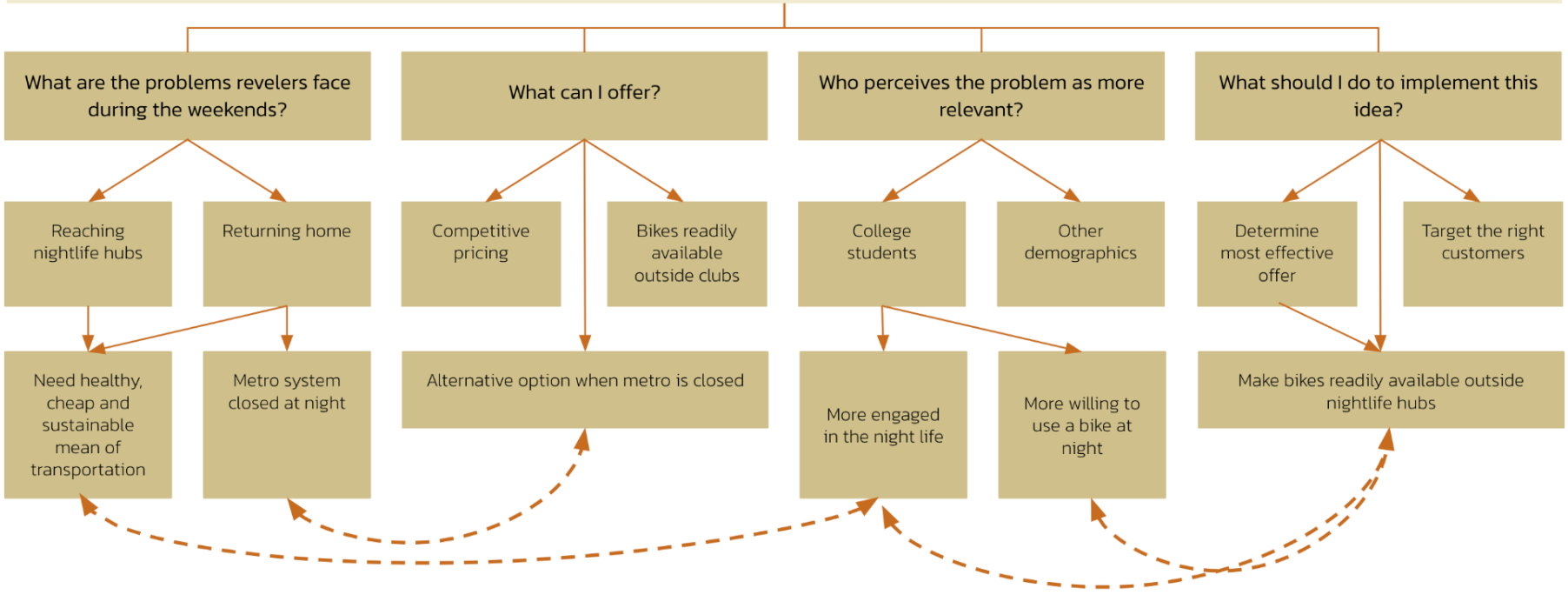
	Offer X is more attractive than the other	The offers are equally attractive
Implement offer X	+++	++
Implement another offer	+	++

In the scenario-action map we are assuming that all offers, even in the case that an alternative is more attractive, are still more profitable than the baseline scenario. In the following page a story tree can be viewed which summarizes all of our questions and possible theories.

STORY TREE

CAPITAL BIKESHARE'S STORY

- People during the night need cheap and versatile means of transportation.
- The Metro system closes early, using your own car may be difficult in big cities and services like Taxis or Ubers are expensive and may have long waiting times.
- Using discounts or offers to encourage revelers to use bikes as a mean of transportation can solve the problems mentioned before for them, while also increasing utilization rates for Capital Bikeshare bikes during the night.



Data collection and hypothesis testing

To better understand if our theory could be successful we started by building a survey.

Survey

The survey was aimed at assessing the sentiment towards biking in Washington DC in general, and specifically to the idea of riding a bike to go out at night. Many factors affecting the decision on whether to ride a bike at night or not are highly dependent on the city in which the respondent lives. For example, a person living in a city with many cycle paths or in a city with low crime rates is probably going to be more willing to ride a bike at night than someone who lives in a less bike-friendly city. For this reason and given the very peculiar characteristics of Washington DC we chose to survey its residents only. To achieve this goal we published our survey on many Facebook groups of the city and were able to obtain, after removing some bots, 228 valid and complete responses.

With the answers from this survey we wish to test six hypotheses (of which 1 is related to the target) and understand which of the three offers (bike all night, free return, 50% discount if parked near DC's main nightlife hubs) is the most popular. To test Hypotheses 1 and 2 we collected answer to the question:

"Do you currently use a bike to go out at night?". Which are shown in Table 3:

Answer	Freq.	Perc.[%]	Cum.[%]
Yes, my own bike	72	31.58	31.58
Yes, a bike-sharing service bike	87	38.16	69.74
No	69	30.26	100.00
	228	100	

Table 3: answers to question "Do you currently use a bike to go out at night?"

To confirm our first hypothesis we require that at least 60% of the people surveyed use a bike at night, a value which is exceeded by our data as 69.74% of the respondents use either their own bike or a bike sharing service at night. Our threshold for hypothesis 2 is that

at least 30% but not more than 50% of the survey-takers use a bike-sharing service bike to go out at night. In fact, we are also interested in assessing whether the market is already saturated. This hypothesis is confirmed too as 38.16% of the respondents use a bike-sharing service bike at night, suggesting that with our offer we could convince more people to use a bike at night as a sizeable but not huge portion of people is already doing it.

To test Hypothesis 3 we used the answers collected for the question:

“Out of the times you go out at night how many times do you use a bike?”

Which was displayed only to those who answered “Yes, my own bike” or “Yes, a bike-sharing service bike” to the question “Do you currently use a bike when you go out at night?”.

The answers collected are shown in Table 4. Our threshold for this hypothesis to be confirmed is that more than 35% of the respondents chose “Always” and at least 25% of the respondents chose “Sometimes”.

From these results we can see that a large percentage of people who use a bike at night use it either always or sometimes, suggesting that those who are used to biking to go out at night do it frequently. Thus, if we convince some other users to use a bike at night we can hypothesise that the offer could succeed in making them repeated customers.

Answer	Freq.	Perc.[%]	Cum.[%]
Always	66	41.51	41.51
Sometimes	80	50.31	91.82
Rarely	13	8.18	100.00
	159	100	

Table 4: answers to question “Out of the times you go out at night how many times do you use a bike?”

In order to test Hypothesis 4 we analyzed the answers to the question:

“Would you be interested in a bike-sharing offer for renting a bike at night?”

which are shown in Table 5. To confirm our hypothesis that people are interested in our offer we would like more than 50% of the respondents to choose either “Definitely yes” or “Probably yes”.

Answer	Freq.	Perc.[%]	Cum.[%]
Definitely yes	72	31.58	31.58
Probably yes	47	20.61	52.19
Might or might not	53	23.25	75.44
Probably not	38	16.67	92.11
Definitely not	18	7.89	100.00
	228	100	

Table 5: answers to question “Would you be interested in a bike-sharing offer for renting a bike at night?”

As we can see from the results displayed in the two tables the two hypotheses can be confirmed. To obtain further insights we have also asked people who responded “No” to the question on whether they ride a bike at night the reason behind this choice. We would like to understand how many of these respondents do not ride a bike at night because they are scared as that is hardly a feeling we can change with an offer.

Answer	Freq.	Perc.[%]	Cum.[%]
I use other means of transportation	33	47.83	47.83
There are no bikes available when I need them	2	2.90	50.73
I do not feel safe riding a bike at night	27	39.13	89.86
I do not know how to ride a bike	2	2.90	92.76
Other	5	7.25	100.00
	69	100	

Table 6: answers to question on why not bike at night

To do so we asked a follow-up question that, among the answers, included the option “I do not feel safe riding a bike at night”. We report in Table 6 the answers to this question.
From the results we see that 39.13% of the people who never use a bike at night do so because they do not feel safe.
From table 1 we can see that 69 people, 30.26% of the respondents, do not use a bike at night. We can thus conclude that out of the surveyed people, 39.13% * 30.26% = 11.84% do not use a bike at night because they do not feel safe. Our hypothesis that required < 20% of the respondents are scared of using a bike at night is thus not rejected.
Following the results from these three analyses, we can conclude that it is worth it to develop the offer as all five hypotheses regarding the Develop/Do not Develop actions are confirmed.
We then verified if the hypothesis regarding the target could be confirmed. In particular we would like to understand whether college students are more interested in the offer than non-college students.
To test our hypothesis that college students would be more interested in the offer we ran an ordered logit regression. In particular we encoded the answers regarding the interest in an offer for a bike sharing service at night (the one from Hypothesis 4). We performed this operation so as to have numbers ranging from 1 to 5 such that 1= “Definitely yes” and 5= ”Definitely not”. We called this variable `interest`.
We regressed `interest` on the following variable created from the answers of the survey using an ordered logistic regression:

variable	question	possible values
`age`	How old are you?	• (14,17) • (18,24) • (25,30) • (30,40) • (40+) • Prefer not to say
`gender`	What gender do you identify with?	• Male • Female • Non-binary / Third-gender • Prefer not to say
`occupation`	Are you a college student?	• Yes • No • Prefer not to say
`bikesatnight`	Out of the times you go out at night how many times do you use a bike?	• Always • Sometimes • Rarely
`bikesharingmember`	Are you a subscriber of an	• Yes

	annual/monthly pass for a bike-sharing service?	• No

We have summarized the obtained results in the Appendix in Table 9. We notice a positive coefficient for occupation = 2 which corresponds to non-college students. This positive coefficient implies that non-college students are more likely to choose a higher value for the variable interest which implies a lower interest (as “Definitely not” = 5); This is in-line with our prior. However the p-value of the coefficient is 0.771 and, thus, insignificant. We, therefore, have to reject our hypothesis that college students are more likely to be interested in the service.

We have summarized the hypotheses explored so far and the corresponding outcomes in the table below.

		Prior	Signal	Decision
Develop idea	*Hypothesis 1*	> 60% use bike at night	**69.74%** use bike at night	✅
	Hypothesis 2	> 30% and <50% use bike-sharing bike at night	**38.16** use bike-sharing at night	✅
	Hypothesis 3	> 35% always bikes at night and > 25% sometimes bikes at night	**31.6%** always **38.2%** sometimes	✅
	Hypothesis 4	> 50% definitely yes + probably yes interested in offer	**52.19%** definitely yes + probably yes	✅
	Hypothesis 5	< 20% not scared of riding a bike at night	**11.84%** do not ride a bike because they are scared	✅
Target	*Hypothesis T-1*	(interest in offer | college student) > (interest in offer | non-college student)	Prior confirmed but not stat significant	❌

Table 7: scenarios, corresponding hypothesis and outcomes

As we had three potential offers we decided to include in our survey three questions, one per offer, in order to assess which was the preferred one. In particular we asked:

- *"Would you be interested in an offer that would give you access to a bike for the whole night?"*
- *"Would you be interested in getting a free return trip when you ride a bike at night?"*
- *"Would you be interested in getting a discount on your bike trip at night if you park near a club/bar?"*

All the above questions had 5 possible answers ranging from "Definitely yes" to "Definitely no". We started by tabulating the answers to the three different questions in Table 8:

	Whole night			Free return			Park discount		
Answer	**Freq.**	**Perc**	**Cum.**	**Freq.**	**Perc**	**Cum.**	**Freq.**	**Perc**	**Cum.**
Definitely yes	59	25.88	25.88	85	37.28	37.28	95	41.67	41.67
Probably yes	30	13.16	39.04	54	23.68	60.96	43	18.86	60.53
Might or might not	55	24.12	63.16	50	21.93	82.89	50	21.93	82.46
Probably not	56	24.56	87.72	22	9.65	92.54	24	10.53	92.98
Definitely not	28	12.28	100.00	17	7.46	100.00	16	7.02	100.00
	228	100		228	100		228	100	

Table 8: answers to questions on different offers

From the frequency table above we can already notice how the offer that grants access to a bike for the whole night is the least popular. In fact, comparing it with the two other offers we can see how only ~ 40% of the respondents would be "Definitely" or "Probably" interested in the offer, while the same metric is ~ 60% for each of the other two offers.

We then ran t-tests on the mean of the difference between couples of offers. Each test compared the null hypothesis: H_0 : mean(offer1-offer2) = 0 vs. H_1 : mean(offer1-offer2) ≠ 0. To perform the difference between the two offers we encoded the answers of the questions so as to have numbers ranging from 1 to 5 such that 1= "Definitely yes" and 5="Definitely not".

We have reported below the results obtained. As we were expecting, when comparing the answers regarding the offer for the whole night with each of the other two offers we obtain a

mean of the difference statistically different from 0. On the other hand, when we compare the two most popular offers the difference of their means is not statistically different from zero.

ttest wholenight == parkdiscount

Variable	Obs	Mean	Std.err.	Std. dev.	[95% conf.	interval]
wholenight	228	2.842105	.0909707	1.373628	2.66285	3.02136
parkdiscount	228	2.223684	.0849608	1.28288	2.056272	2.391097
diff	228	.6184211	.070577	1.06569	.4793512	.7574909
	mean(diff) = mean(wholenight) -mean(parkdiscount)				t =	8.7624
H0 :	mean(diff) = 0			Degrees of freedom =		227

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

ttest wholenight == freereturn

Variable	Obs	Mean	Std.err.	Std. dev.	[95% conf.	interval]
wholenight	228	2.842105	.0909707	1.373628	2.66285	3.02136
freereturn	228	2.263158	.0834253	1.259695	2.098771	2.427545
diff	228	.5789474	.0645391	.9745189	.4517751	.7061197
	mean(diff) =mean(wholenight)-mean(freereturn)				t =	8.9705
H0 :	mean(diff) = 0			Degrees of freedom =		227

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

ttest freereturn == parkdiscount

Variable	Obs	Mean	Std.err.	Std. dev.	[95% conf.	interval]
freereturn	228	2.263158	.0834253	1.259695	2.098771	2.427545
parkdiscount	228	2.223684	.0849608	1.28288	2.056272	2.391097
diff	228	.0394737	.0521292	.7871343	-.0632454	.1421928
	mean(diff) =mean(freereturn)-mean(parkdiscount)				t =	8.7624
H0 :	mean(diff) = 0			Degrees of freedom =		227

Ha : mean(diff) < 0	Ha : mean(diff) != 0	Ha : mean(diff) > 0
Pr(T < t) = 0.7752	Pr(|T| > |t|) = 0.4497	Pr(T > t) = 0.2248

Updated business theory

After the results of the first survey, we weren't able to conclusively select which offer is poised to be better received by users. Therefore, we decided to do an A/B test to determine the preferred offer among the two most popular ones: the one that offers a discount (which we will fix at 50% for the experiment) when parking a bike near clubs/restaurants and the one that offers a free return ride at night. The scenario-action map in this case can be summarized as follows.

Prior: one offer is preferred to the other

	50% discount for trips that end near club/bar more attractive	**Equally attractive**	**One-way gets you return trip free more attractive**
Implement 50% discount	+++	++	+
Implement free return trip	+	++	+++

A/B Testing

To perform this A/B test we have created a second survey asking respondents, randomly, either their interest in the first offer or the second one. In particular each respondent was asked one of the following questions:

A) "It's Saturday evening and you decided to go out. You're looking for means of transportation and a bike sharing app offers you a free return trip if you ride a bike to get to your destination. Would you take the offer?"

B) "It's Saturday evening and you decided to go out. You're looking for means of transportation and a bike sharing app offers you a 50% discount if you ride a bike to popular spots. Would you take the offer?"

Both questions had the same 5 possible answers ranging from "Definitely yes" to "Definitely no".

Following the same reasoning from Survey 1 we collected answers from 150 Washington DC residents, out of which 73 saw question A and 77 saw question B.

We first performed some balance tests on the age and gender of the respondents which were both passed as we never reject the hypothesis that they are equal in mean (see Stata code for details).
We then encoded the answers to questions A) and B) with the usual scale: 1 = “Definitely yes” and 5= “Definitely no” into a variable called `likelihood`. Since our balance tests achieved good results we can assume the randomization was well performed and we, thus, proceeded with a mean linear regression of `likelihood` on `offer_type` (1 if offer A, 0 if offer B).
The regression yielded a significant coefficient for `offer_type` of -0.883, suggesting that people who have been presented with offer A have a higher likelihood of taking advantage of the offer (lower in value as per our encoding). The full results of this regression are shown in the Appendix in table 10.

To gather more insights In our survey we also asked an additional question: “How important was the offer for your choice?”, which allowed for 5 possible answers ranging from “Extremely important” to “Not at all important”. With this question we are interested in exploring further if the offer had an impact on the decision of taking the bike to go out at night.
We performed two ordered logistic regressions on `likelihood` by `offer_impact` for each of the two offers. The results of these regression, whose full results are shown in the appendix (Table 11 and 12), show that not only the offer is significant but that it also has a positive impact in the decision taken by the respondent.

Conclusions

After analyzing the results of the A/B test, we found a statistically significant preference for the option that offers users a free return trip for any outward journey at night. The next step in our analysis would be to implement this idea by testing it on the real users of the Capital Bikeshare’s ride-sharing app. To be sure of not cannibalizing any already existent business, e.g. offering a free return trip to someone already using the bike during the night for both the outward journey and the return, we would use the data gathered by the app. We would send this offer only to clients that do not usually use Capital Bikeshare’s rides during the night. We would then send out a notification on the app to the targeted clients, detailing the offer on Friday and Saturday afternoon. This would ensure that they would consider the offer for their transportation needs during the evening.

It is important to note that the potential downsides of doing this live experiment are quite limited as rides during the night are not particularly popular at the moment. Succeeding in Increasing the utilization rates of Capital Bikeshare's bikes during weekend nights, instead, will lead to a rise in profits as the majority of costs incurred by the company are fixed costs.

In conclusion, we, thus, believe that Capital Bikeshare should proceed with a live experiment, especially considering that the downsides are limited and the potential upsides are much greater.

Appendix

Table 9: Ordered logit regression output

VARIABLES	(1) interest
2.gender	0.148
	(0.275)
3.gender	0.842
	(1.400)
4.gender	2.337
	(1.459)
2.age	-0.285
	(2.047)
3.age	0.247
	(2.058)
4.age	0.543
	(2.073)
5.age	0.828
	(2.087)
6.age	-0.0254
	(2.358)
2.occupation	0.111
	(0.382)
3.occupation	-0.995
	(1.852)
2.bikesatnight	-0.526*
	(0.319)

3.bikesatnight	1.502***
	(0.436)
2.bikesharingmember	1.112***
	(0.366)
/cut1	-0.272
	(2.062)
/cut2	0.906
	(2.061)
/cut3	2.574
	(2.072)
/cut4	4.457**
	(2.093)
Observations	228

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 10: A/B testing offer type

	(1)
VARIABLES	likelihood
offer_type	-0.883***
	(0.208)
Constant	3.143***
	(0.145)

Observations	150
R-squared	0.109

*** p<0.01, ** p<0.05, * p<0.1

Table 11: ologit regression on offer A impact on decision of taking bike at night

	(1)
VARIABLES	abikelikelihood
aofferimpact	0.826***
	(0.213)
/cut1	1.311**
	(0.566)
/cut2	2.634***
	(0.624)
/cut3	3.775***
	(0.705)
/cut4	6.079***
	(1.009)
Observations	73

*** p<0.01, ** p<0.05, * p<0.1

Table 12: ologit regression on offer B impact on decision of taking bike at night

	(1)
VARIABLES	bbikelikelihood
bofferimpact2	0.611***
	(0.187)
/cut1	-0.247
	(0.565)
/cut2	1.065*
	(0.567)
/cut3	2.109***
	(0.615)
/cut4	3.251***
	(0.674)
Observations	77

*** p<0.01, ** p<0.05, * p<0.1