

# H-Farm Project

Bottani Sophie  
Chiarini Emanuele  
Gatteschi Giulia  
Giannelli Enrico



# DATASETS

## 1) Visual Crossing

Collects all the information regarding the weather in Washington DC by day and hour between February and April 2021. There are 2,122 observations and 17 variables:

- Name
- Date time
- Maximum Temperature
- Minimum Temperature
- Temperature
- Wind Chill
- Heat Index
- Precipitation
- Snow
- Snow Depth
- Wind Speed
- Wind Gust
- Wind Direction
- Visibility
- Cloud Cover
- Relative Humidity
- Conditions



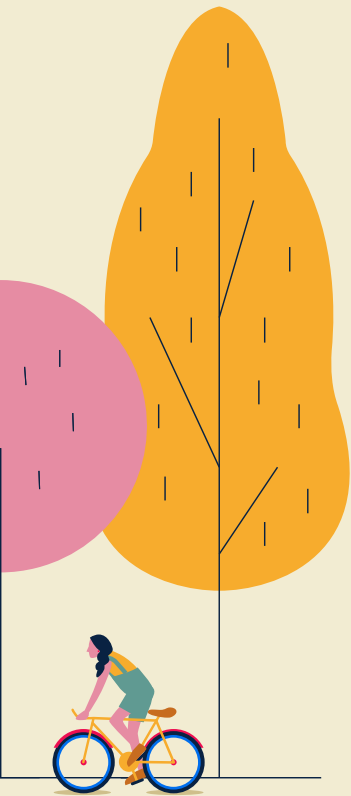
# DATASETS



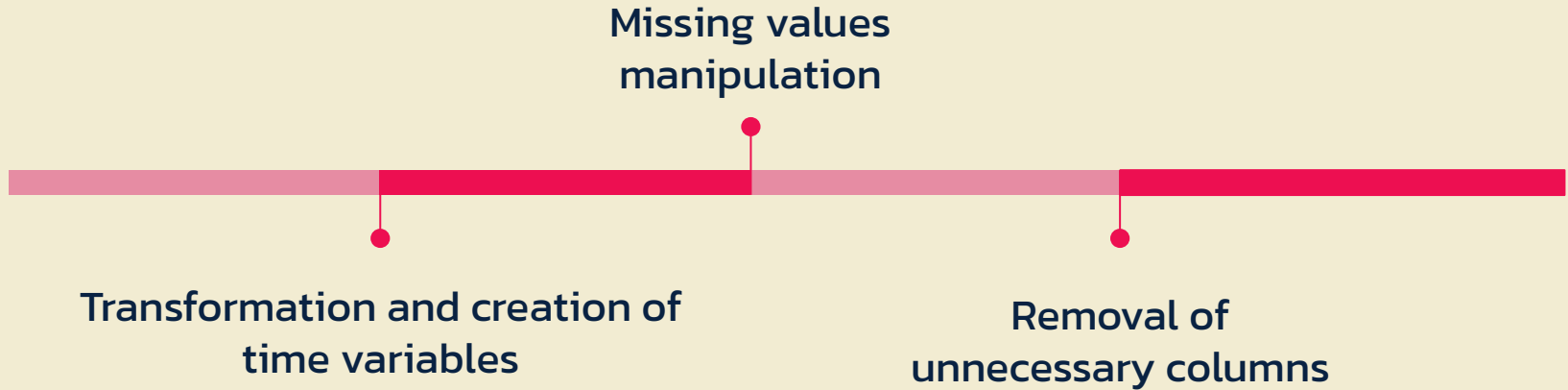
## 2) Capital Bikeshare

Contains information about each ride made by users in Washington DC and 13 variables:

- ride\_id
- rideable\_type
- started\_at
- ended\_at
- start\_station\_name
- start\_station\_id
- end\_station\_name
- end\_station\_id
- start\_lat
- start\_lng
- end\_lat
- end\_lng
- member\_casual



# DATA PREPARATION: Weather Data



# DATA PREPARATION: Capital Bikeshare



Missing values manipulation



Transformation and  
creation of time variables



Outliers manipulation



Label encoder for our  
categorical variables

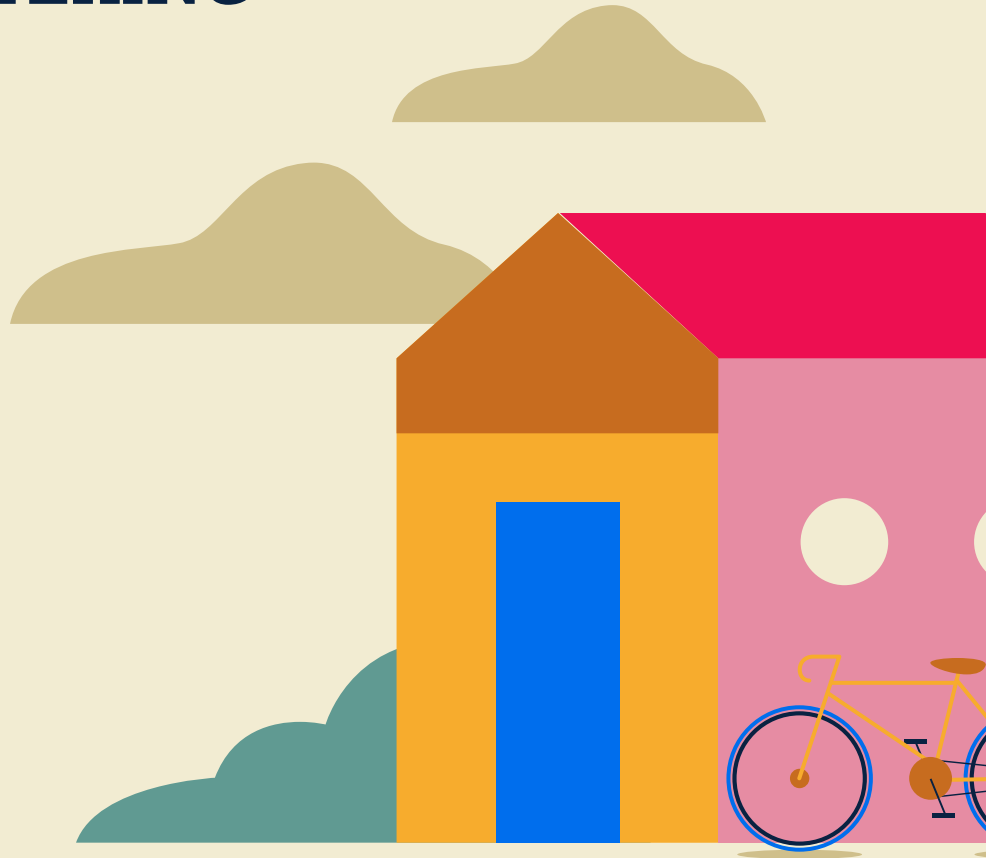


Creation of useful  
variables



# CLUSTERING

- **K-Means algorithm:**  
Sklearn implementation, K-Means++ initialization
- **Scaling of the data:**  
Min-Max scaler and division by 2 of longitude/latitude
- **Elbow Method:**  
based on distortion



# FIRST CLUSTERING

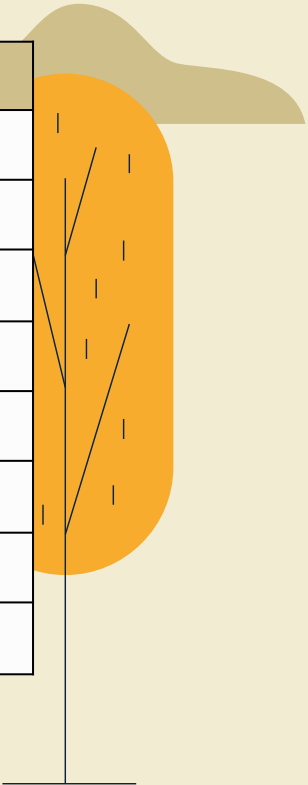
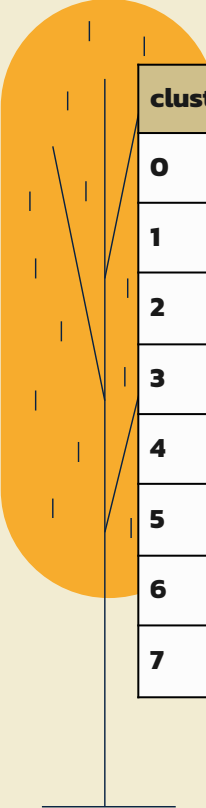
Performed a k-means algorithm on a dataset including some key variables from the rides and weather dataset:

- Starting and ending points of the ride
- Starting hour and day of the week
- User membership
- Precipitation and snow during ride starting hour

Results showed clusters differ a lot on the starting hour and day of the week



# FIRST CLUSTERING: RESULTS



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

Means of the variables of interest across clusters



# SECOND CLUSTERING



Decided to explore this pattern further!

Performed a second k-means algorithm on:

- Starting and ending points of the ride
- Starting hour and day of the week

Found a very interesting cluster that is:

- Small
- Concentrated on the weekend
- Concentrated in the first hours of the day
- Mostly made up by non-members rides



# SECOND CLUSTERING: RESULTS

Cluster	count	start_hour		weekday		member_casual_enc	
		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



# A POTENTIAL BUSINESS OPPORTUNITY



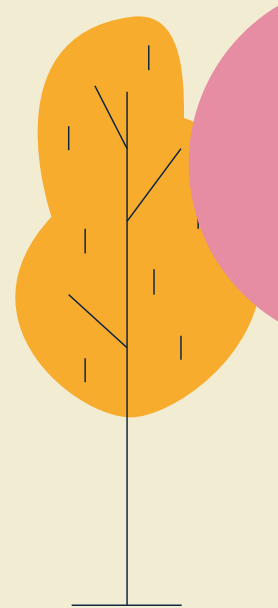
**Few** people use bikes in the middle of the night

Could design offer that increases this type of usage

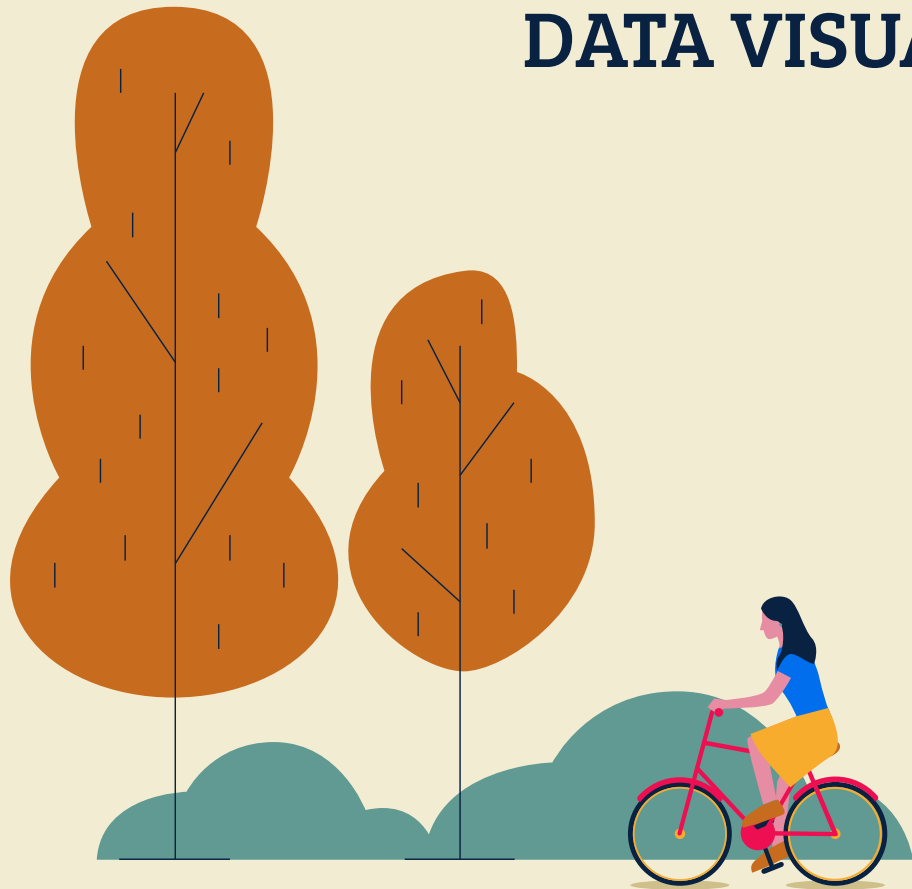
First need to understand why these rides are being done

Could be coming home from a night out

Let's visualize where these rides and the Washington's nightlife are!

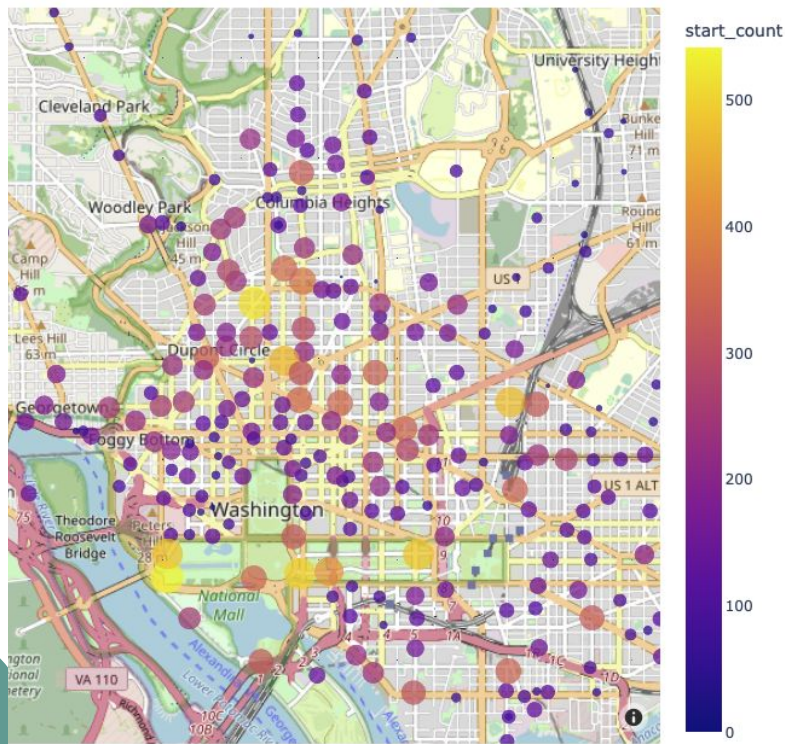


# DATA VISUALIZATION

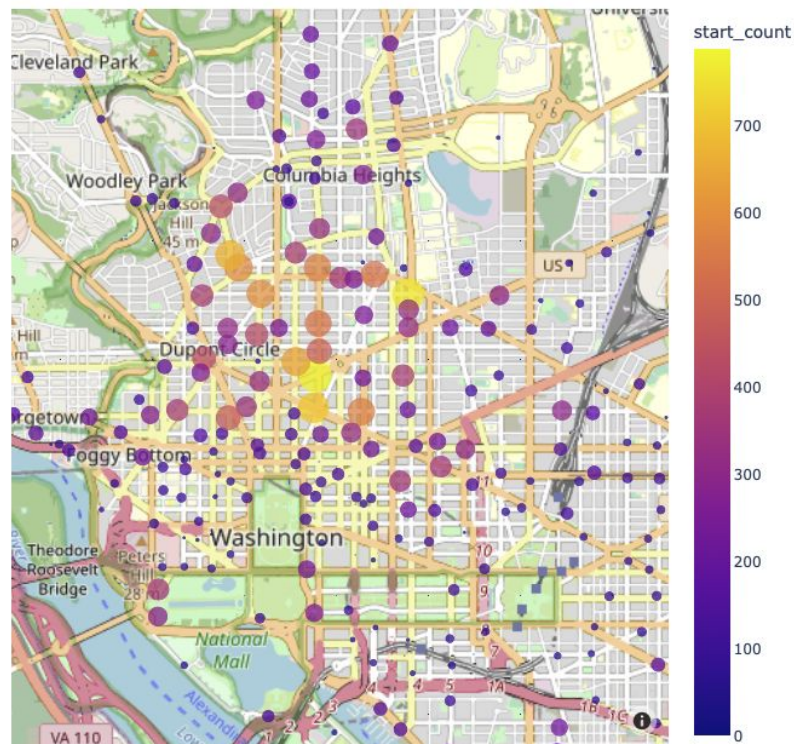


- We created some heat maps to see which areas are more popular in terms of starting points and routes
- We used the Plotly library
- We compared the starting station and routes between the whole dataset and our cluster of interest

# DATA VISUALIZATION: starting stations



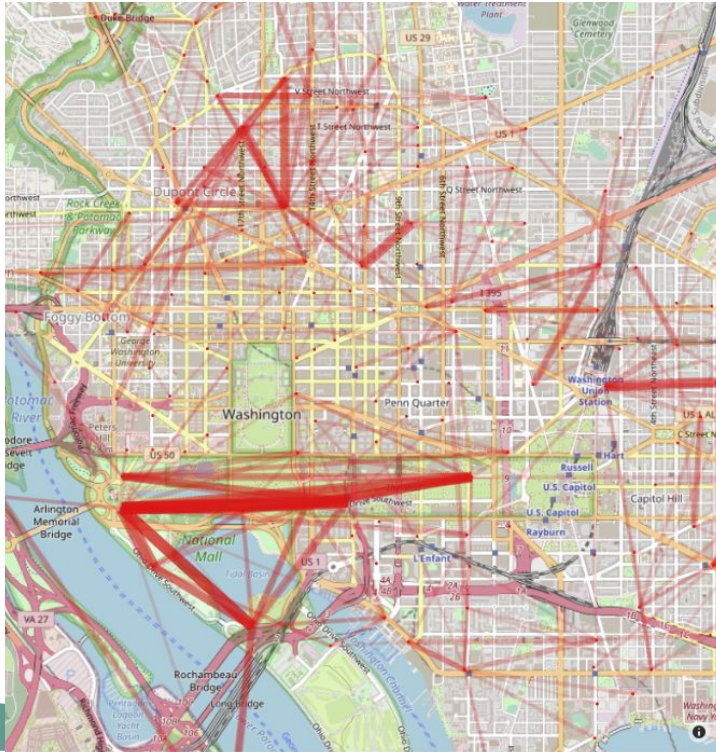
Starting stations for the whole dataset



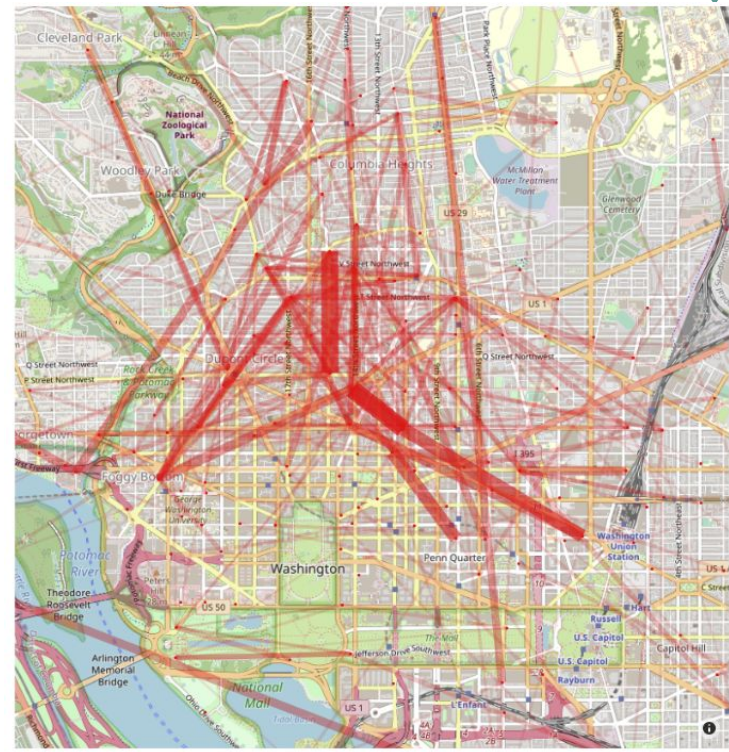
Starting stations for rides made during the night



# DATA VISUALIZATION: routes



Routes for the whole dataset



Routes for rides made during the night

# A PEAK AT DC'S NIGHTLIFE

- Rides are concentrated north of the National Mall
- Since they are done during a weekend night we expect the nightlife of Washington D.C. to be located in this area



# A PEEK AT DC'S NIGHTLIFE

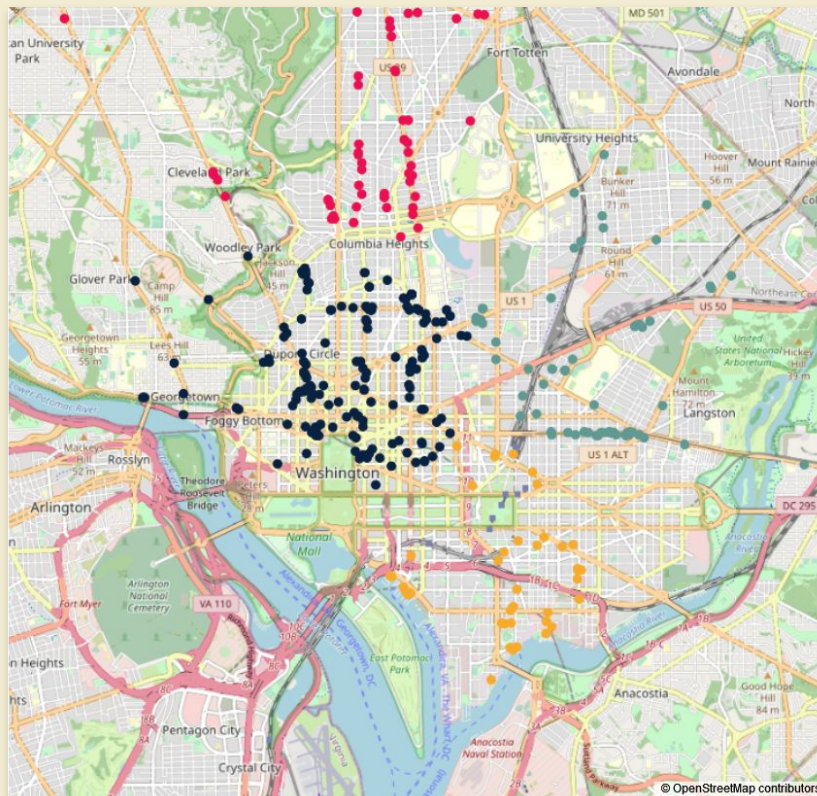
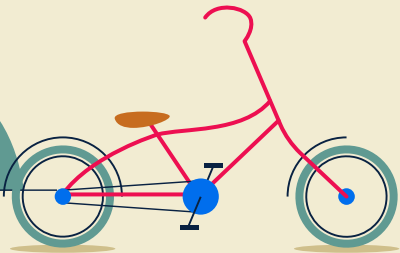


- We wanted to confirm this intuition
- To do so we used the list of all venues selling alcohol as a proxy
- Extracted the data from the pdf with records on DC's municipal government.
- Cleaned them and retrieved their latitude, longitude and category



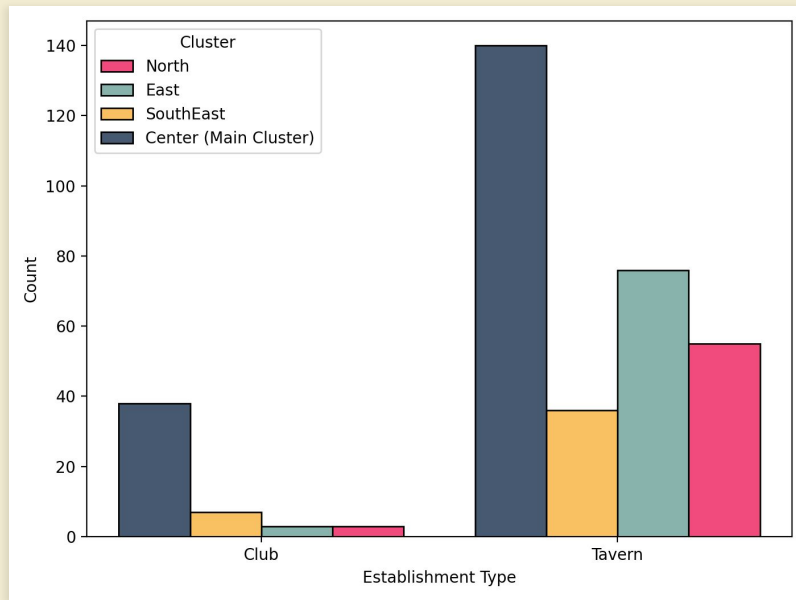
# A PEEK AT DC'S NIGHTLIFE

- Applying a K-means algorithm at this scraped data yields four main clusters of Washington's nightlife
- The main one is located in the center of the city, another in the North, one South East and the last in the east



# A PEEK AT DC'S NIGHTLIFE

- The composition of the clusters is really different
- The percentage of clubs compared to bars and taverns in the main cluster in the city center is significantly higher compared to the other three
- This confirms our theory that the region from where many of the rides during the night start is the main hub of the nightlife in Washington and that, thus, these rides are related to it



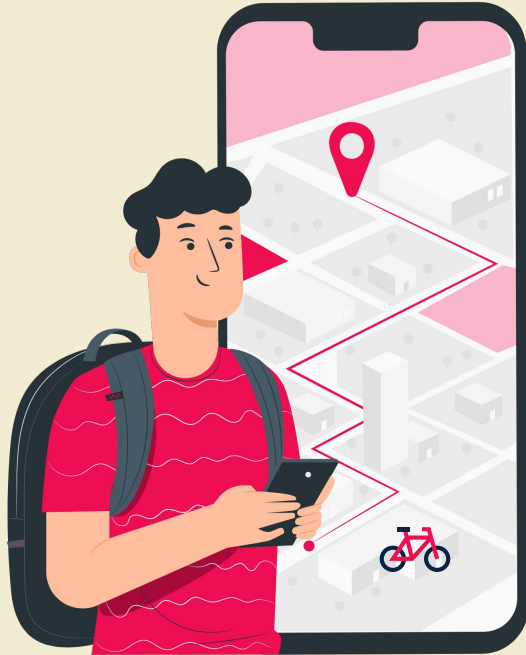
# INITIAL BUSINESS THEORY

Adding to what we've found up to now that:

- Washington is among the safest cities in the US
- Majority of citizens not scared of roaming around it at night
- One of the highest bikeability among US cities
- Most of these rides not part of a membership
- Increasing utilization rates may boost profitability as most costs are fixed



# INITIAL BUSINESS THEORY



**Opportunity to increase profits by increasing night rides!**

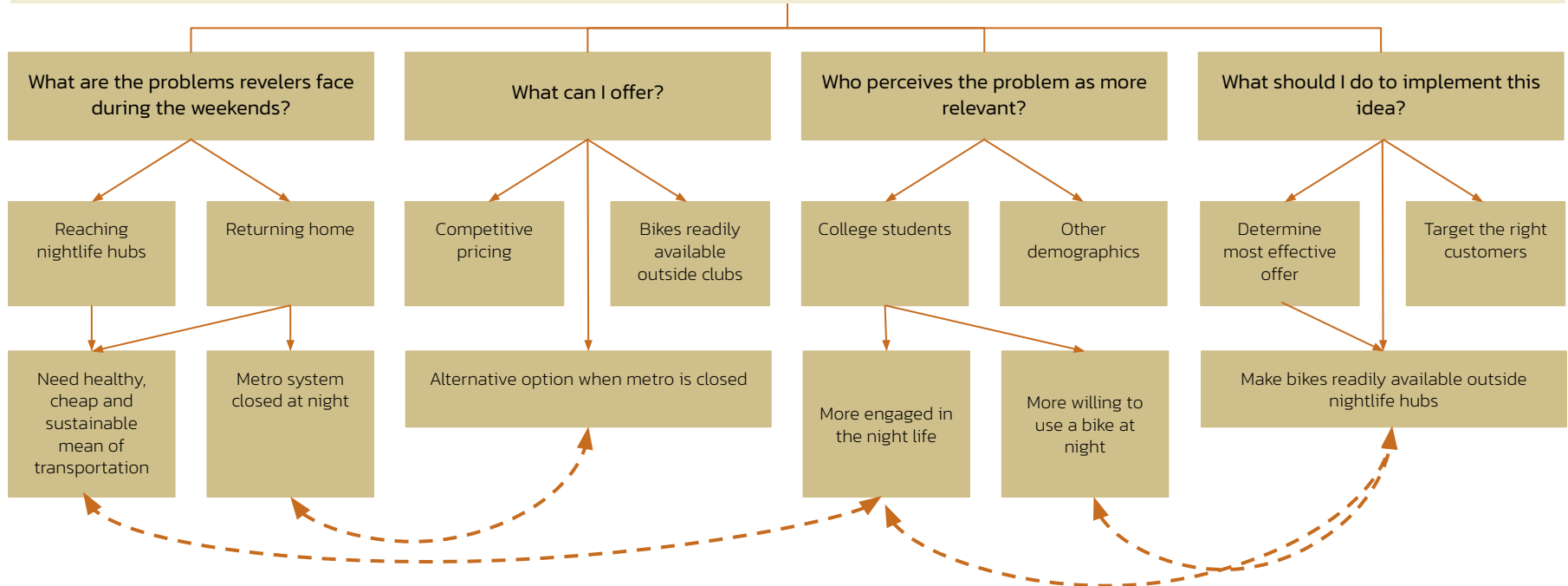
Our three business ideas are:

- A bundle that would allow users to “book” a bike for the whole night
- A free return ride for any ride during the night
- A discount on rides that end in DC’s nightlife main cluster

# 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



# SCENARIO-ACTION MAP 1

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

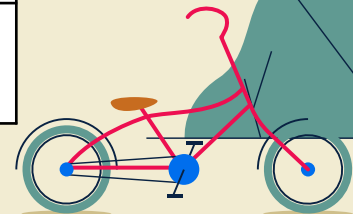
We will accept the first scenario if 4 out of the 5 hypotheses we use to test it will be verified



# SCENARIO-ACTION MAP 2

Prior: college students are the perfect target customers

	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	-	+	++

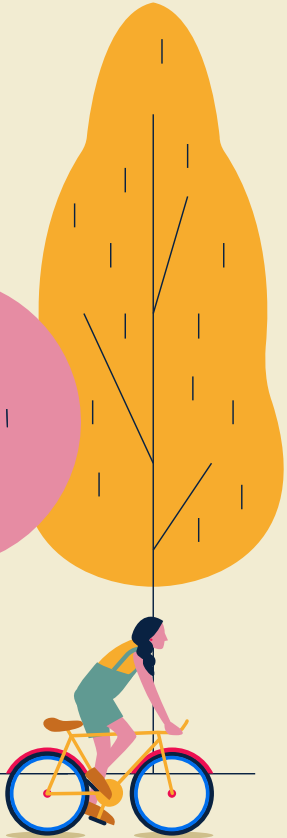


# SCENARIO-ACTION MAP 3



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	+	++





# SURVEY

## Objective:

- Assess general sentiment towards biking in Washington DC → **especially at night**
- Assess which of the offers is most popular
- Surveyed **Washington DC residents only** as willingness to ride bike highly depends on city's attributes (infrastructure, crime rate, traffic ...)

**Surveyed  
228 people**

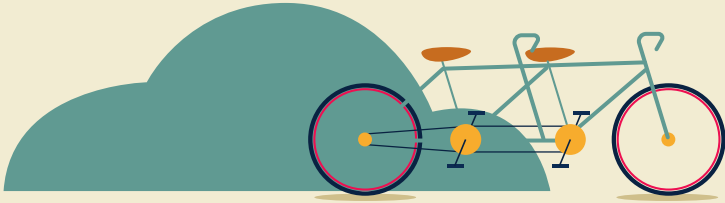
# HYPOTHESIS TESTING

Develop/Not develop

	Hypothesis	Prior
1	People use the bike to go out at night	> 60% use bike at night
2	People use bike-sharing at night but the market is not saturated	> 30% and <50% use bike-sharing bike at night

**"Do you currently use a bike to go out at night?"**

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	



# HYPOTHESIS TESTING

Develop/Not develop

Out of those who replied "Yes" to the previous question

	Hypothesis	Prior
3	People <b>often</b> use a bike to go out at night	> 35% always use bike at night > 25% sometimes use bike at night

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

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

if we convince some other users to use a bike at night the offer could succeed in making them **repeated** customers!



# HYPOTHESIS TESTING

Develop/Not develop

Hypothesis		Prior
4	People are <b>interested</b> in a bike sharing offer	> 50% "probably yes" & "definitely yes"

**"Would you be interested in a bike-sharing offer for renting a bike at night?"**

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
	159	100	





# HYPOTHESIS TESTING

Develop/Not develop



	Hypothesis	Prior
5	People are not scared of biking at night in Washington DC	< 20% scared of riding a bike at night

**"Why are you not using a bike to go out at night?"**

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	

$39.13\% \times 30.26\% = \mathbf{11.84\%}$  of the people do not feel safe riding a bike at night

**"Do you currently use bike to go out at night?"**

Answer	Freq.	Perc.[%]	Cum.[%]
No	69	30.26	100.00

# HYPOTHESIS TESTING

Target

Hypothesis		Prior
6	College students will be more interested in the service	$(\text{interest in offer} \mid \text{college student}) > (\text{interest in offer} \mid \text{non-college student})$

**Ordered logistic regression**, *interest* as dependent variable. Interested in the coefficient of *student*=2

Positive coefficient but **not statistically significant**! **Prior not confirmed**

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

Answer	interest
Definitely yes	1
Probably yes	2
Might or might not	3
Probably not	4
Definitely not	5

- Gender
- Age
- Bike sharing member

"Do you currently use a bike when you go out at night?"

Answer	bikenight
Yes, my own bike	1
Yes, bike-sharing bike <sup>a</sup>	2
No	3

"Are you a college student?"

Answer	student
Yes	1
No	2
Prefer not to say	3

# HYPOTHESIS TESTING

		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	✗

# HYPOTHESIS TESTING

Most popular offer

	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	

**Need A/B testing** to assess the best offer!

T-tests	$H_a : \text{mean}(\text{diff}) \neq 0$	Decision
$H_0 : \text{mean}(\text{whole night} - \text{park discount})$	$\Pr( T  >  t ) = 0.0000$	Reject $H_0$
$H_0 : \text{mean}(\text{whole night} - \text{free return})$	$\Pr( T  >  t ) = 0.0000$	Reject $H_0$
$H_0 : \text{mean}(\text{free return} - \text{park discount})$	$\Pr( T  >  t ) = 0.4497$	Cannot reject $H_0$

Offers on free return and park discount are best



# UPDATED BUSINESS MODEL

After the results of the first survey, we did another experiment to determine the best offer among the two most popular:

- 50% discount when parking a bike near clubs/restaurants
- Free return ride



# SCENARIO-ACTION MAP

Prior: one offer is preferred to the other

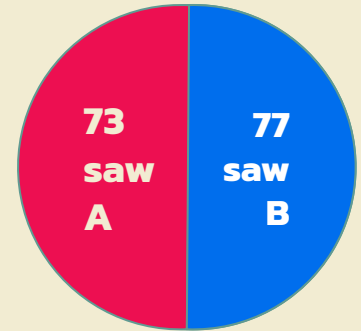
	50% discount for trips that end near club/bar more attractive	Equally attractive	One-way gets you return trip free more attractive
Implement 50% discount	+++	++	+
Implement free return trip	+	++	+++



# A/B Test



- We have thus created a second survey where 50% of the respondents are shown a question regarding their interest in one of the first offer and the other 50% on the second. The two offers are
  - A. Free return
  - B. Discount if parked in nightlife hubs
- Objective : find **most popular offer**
- Surveyed **Washington DC residents only**



# HYPOTHESIS TESTING

- **Balance checks** on age and gender passed
- Regression on **likelihood** and **offer\_type** ( offer A = 1, B = 0)
- The regression yielded a significant coefficient for offer\_type of **-0.883**, suggesting that people who have been asked question A have higher likelihood of taking advantage of the offer

"It's Saturday evening and you decided to go out. Offer A/B. Would you take the offer?"

Answer	likelihood
Definitely yes	1
Probably yes	2
Might or might not	3
Probably not	4
Definitely not	5

"How important was the offer for your choice?"

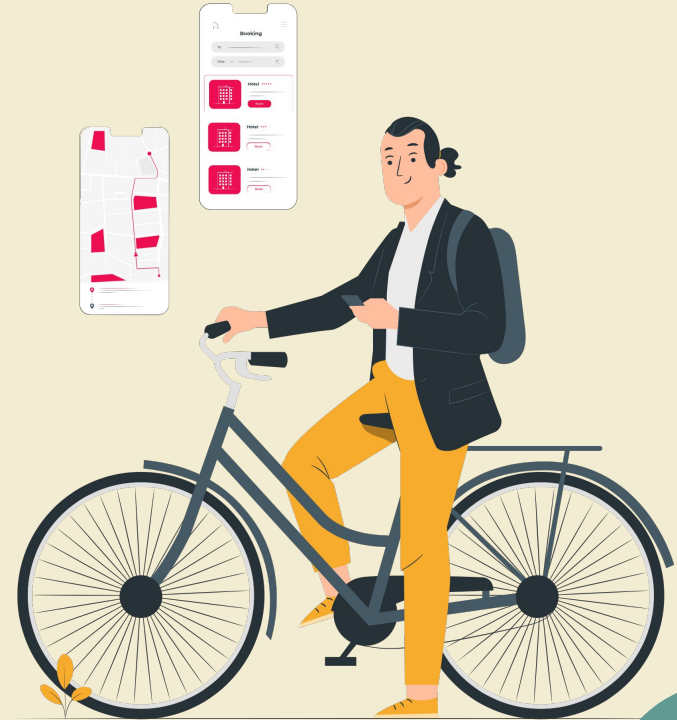
Answer	offer_impact
Extremely important	1
Very important	2
Moderately important	3
Slightly important	4
Not at all important	5

- Performed **ordered logistic regression** for offer A and B on likelihood by(offer\_impact)
- Offer has **significant impact** on decision!

**Offer A wins on offer B!**

# CONCLUSIONS

- Preference for the bundle offering a free return trip for a night outward journey
- Proceed with a live-test of this business ideas on the app's users
- Target users that usually do not use a bike at night
- Send notification detailing the offer during, Friday or Saturday afternoon
- Almost non-existent downside, potential significant higher profit



# Thanks!

Do you have any questions?



Bottani Sophie  
Chiarini Emanuele  
Gatteschi Giulia  
Giannelli Enrico