Optimization of Barcelona’s new subway line

Laia Alcaraz

Applied Data Science Capstone,

IBM Data Science Professional Certificate

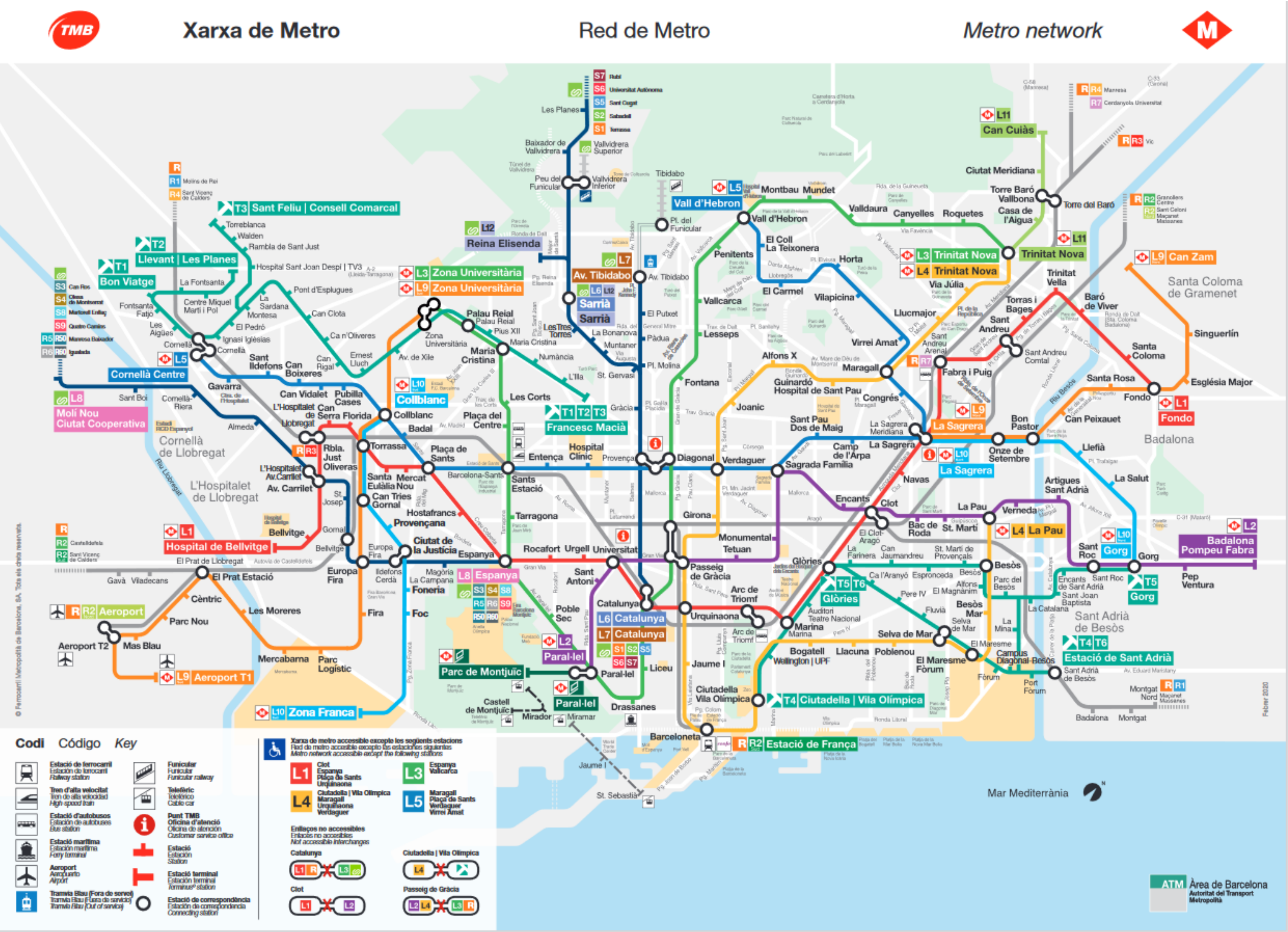
December 2020

Outline

- Introduction
- Data
- Methodology
- Results
- Conclusions and future work

Introduction

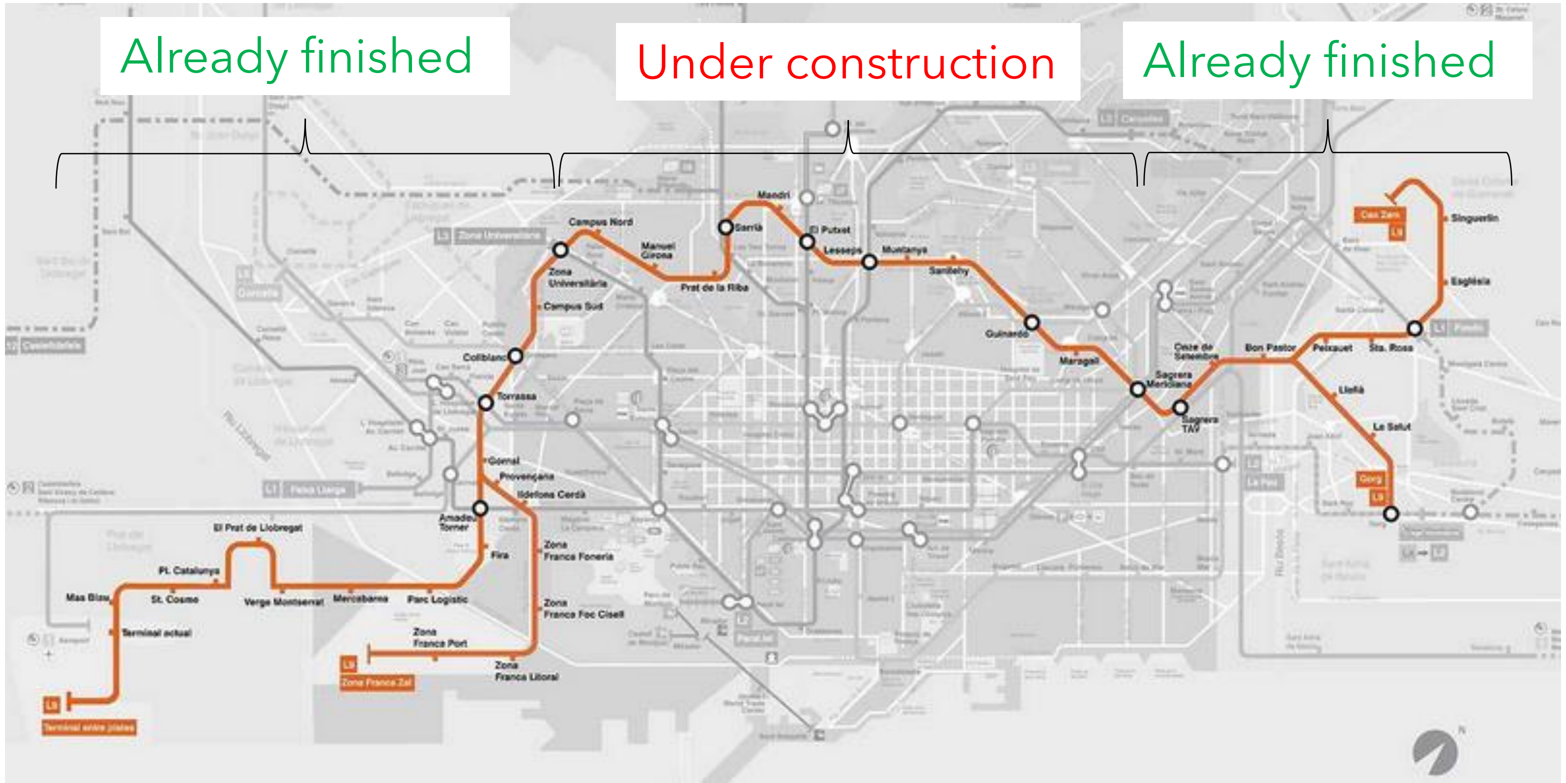
- Barcelona is a city of Spain with 1.6 milion of citizens.
- In 2019, the demand of its subway (called Metro) reached 411.95 milions.



Barcelona's Metro network (source: www.tmb.cat)

Introduction

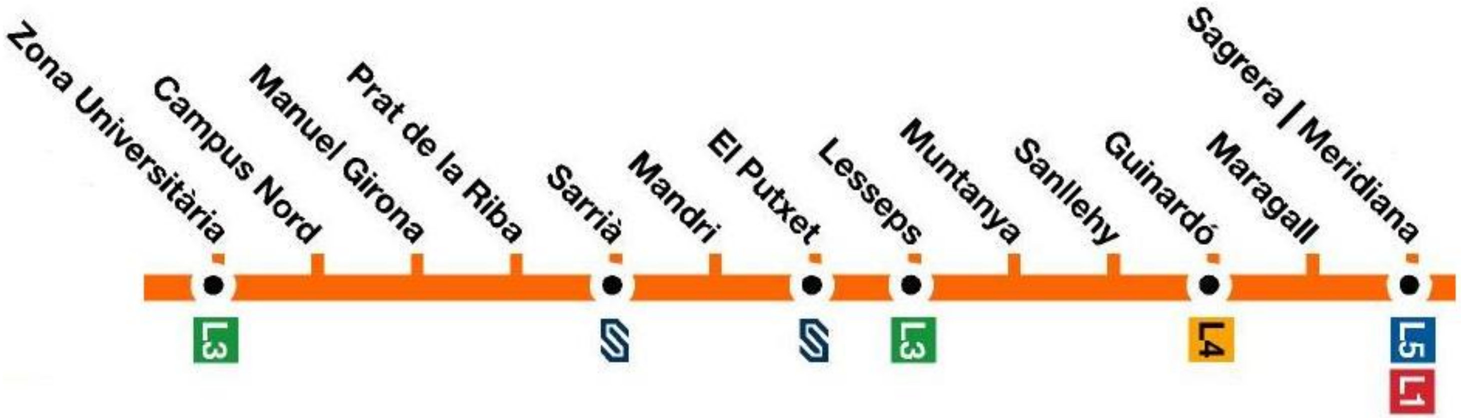
- Currently, the line 9 (orange) of the Metro is being built.
- Once it is finished, this length of this Metro line will be 27.7 km.



Planned stations of Barcelona's Metro line 9 (source: Pinterest)

Introduction

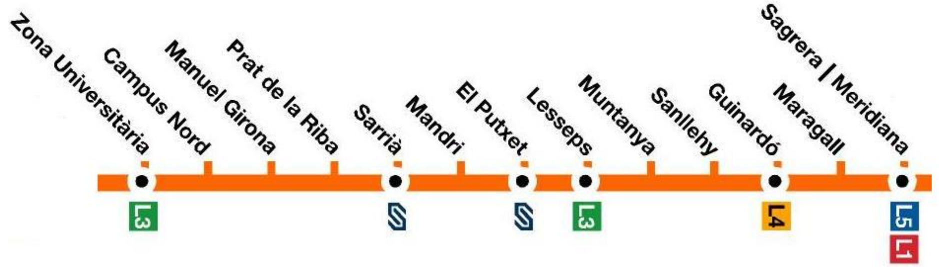
- This work aims to study if the location of the Metro stations that are still being built could be optimized taking into acount their proximity to shops, hospitals, universities…
- The stations that will be analysed are the ones that are still under construction:
    - Campus Nord
    - Manuel Girona
    - Prat de la Riba
    - Mandri
    - Muntanya
    - Sanllehy
    - Maragall

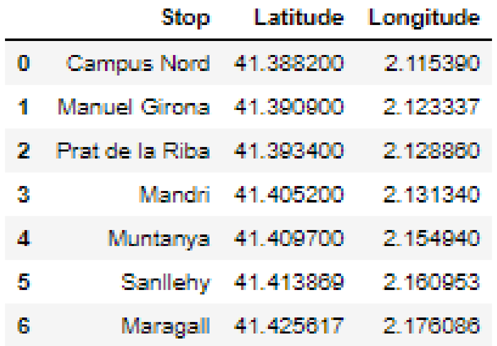


Section of the L9 line to be optimized (source: www.tmb.cat)

Data

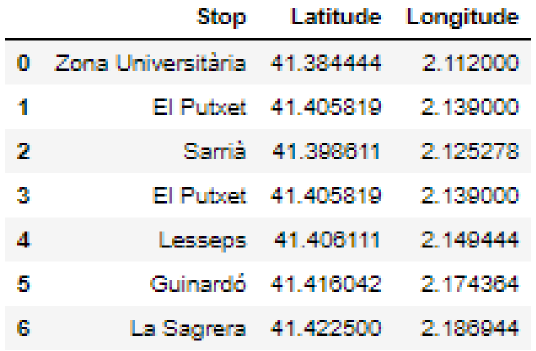
- The coordinates (latitude and longitude) of the Metro stops can be found in Wikipedia. In the code, they are stored in two different Pandas dataframes.
- For each Metro stop under construction, the 100 top venues within a radius of 500 m are obtained through the Foursquare API. The data is stored in a third dataframe.





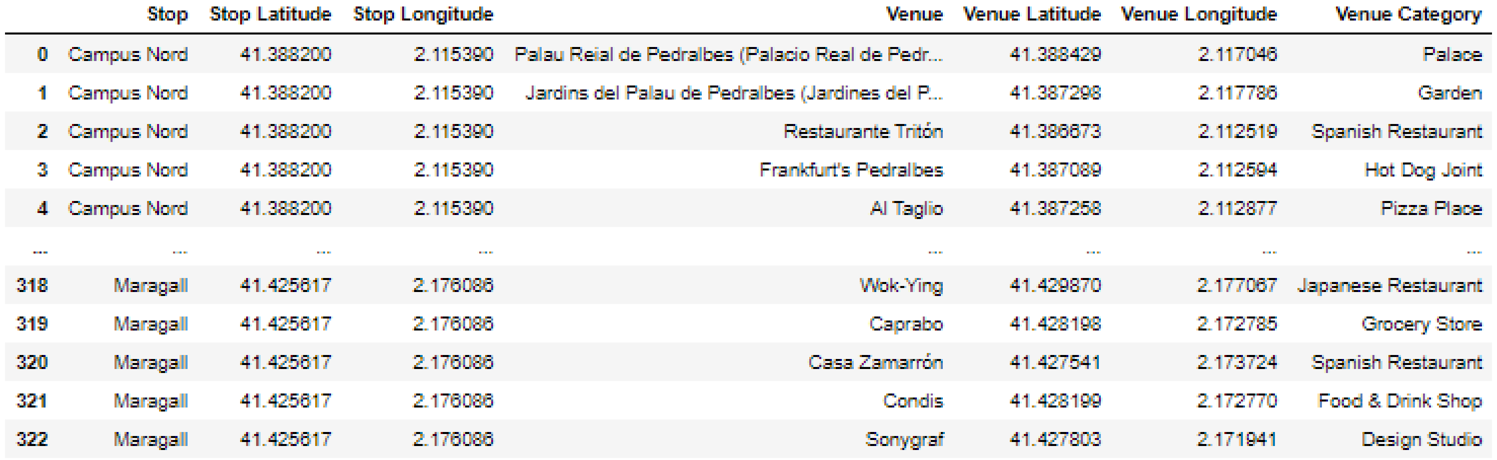
	Stop	Latitude	Longitude
0	Campus Nord	41.388200	2.115390
1	Manuel Girona	41.390900	2.123337
2	Prat de la Riba	41.393400	2.128860
3	Mandri	41.405200	2.131340
4	Muntanya	41.409700	2.154940
5	Sanllehy	41.413869	2.160953
6	Maragall	41.425617	2.176086

df_coordstops (stops to be optimized)



	Stop	Latitude	Longitude
0	Zona Universitària	41.384444	2.112000
1	El Putxet	41.405819	2.139000
2	Sarrià	41.398611	2.125278
3	El Putxet	41.405819	2.139000
4	Lesseps	41.408111	2.149444
5	Guinardó	41.416042	2.174364
6	La Sagrera	41.422500	2.186944

df_coordfixedstops (built stops)

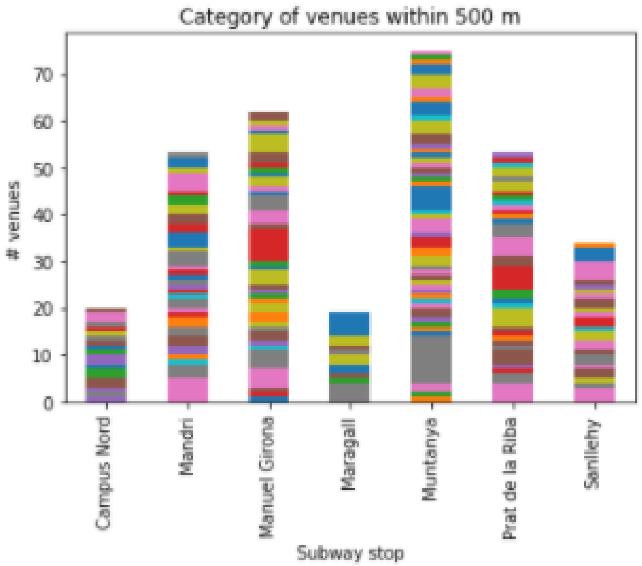


	Stop	Stop Latitude	Stop Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Campus Nord	41.388200	2.115390	Palau Reial de Pedralbes (Palacio Real de Pedr...	41.388429	2.117046	Palace
1	Campus Nord	41.388200	2.115390	Jardins del Palau de Pedralbes (Jardines del P...	41.387298	2.117786	Garden
2	Campus Nord	41.388200	2.115390	Restaurante Tritón	41.386673	2.112519	Spanish Restaurant
3	Campus Nord	41.388200	2.115390	Frankfurt's Pedralbes	41.387089	2.112594	Hot Dog Joint
4	Campus Nord	41.388200	2.115390	Al Taglio	41.387258	2.112877	Pizza Place
...	...	...	...	...	...	...	...
318	Maragall	41.425617	2.176086	Wok-Ying	41.429870	2.177067	Japanese Restaurant
319	Maragall	41.425617	2.176086	Caprabo	41.428198	2.172785	Grocery Store
320	Maragall	41.425617	2.176086	Casa Zamarrón	41.427541	2.173724	Spanish Restaurant
321	Maragall	41.425617	2.176086	Condis	41.428199	2.172770	Food & Drink Shop
322	Maragall	41.425617	2.176086	Sonygraf	41.427803	2.171941	Design Studio

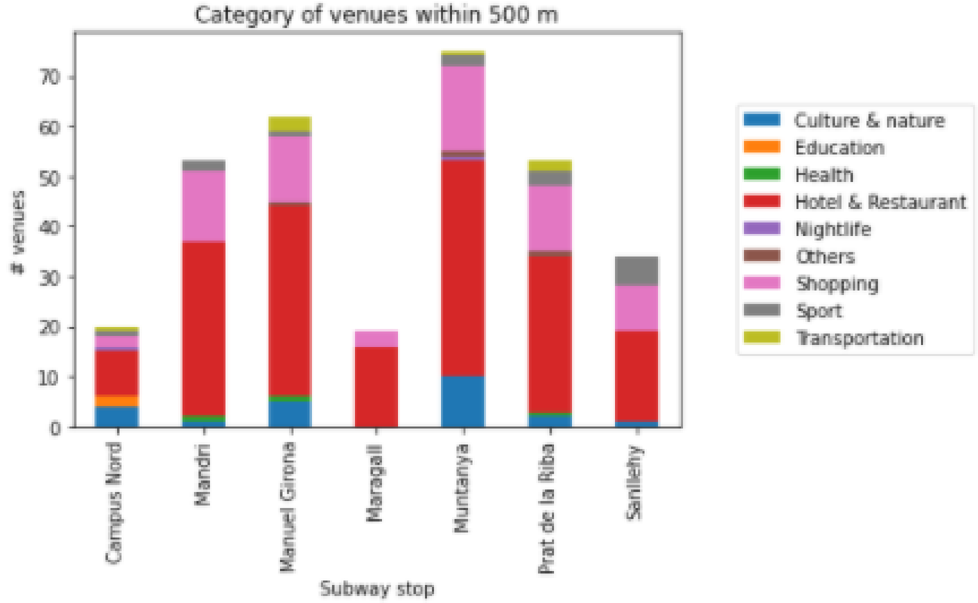
df_bcn_venues

Data

The df_venues_dataframe shows that the categories of the venues are too specific. Therefore, data is processed to group the venues in more general categories.



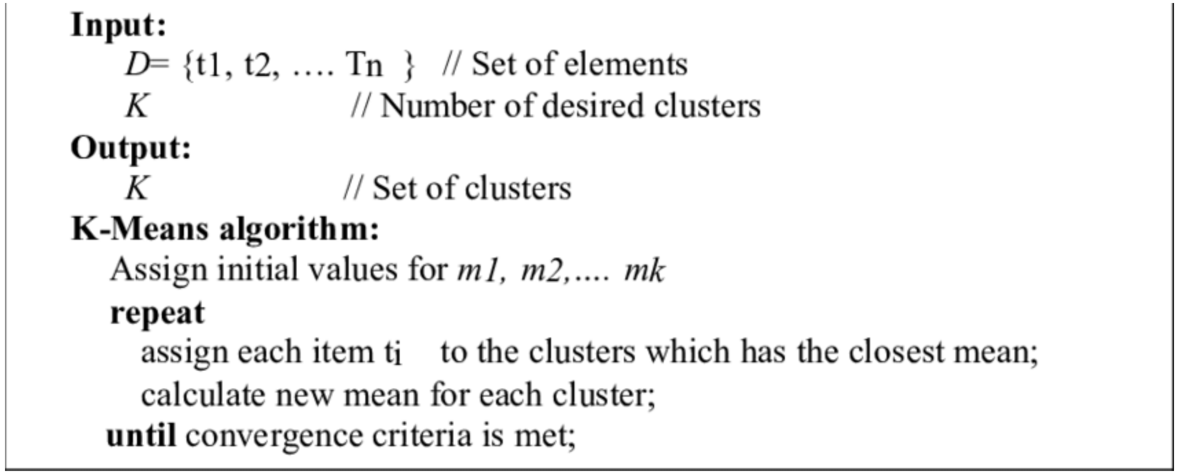
(for the sake of simplicity the legend has been deleted)



The plot on the right shows that despite having only 20 venues, the "Campus Nord" stop is the one where its variety is the highest.

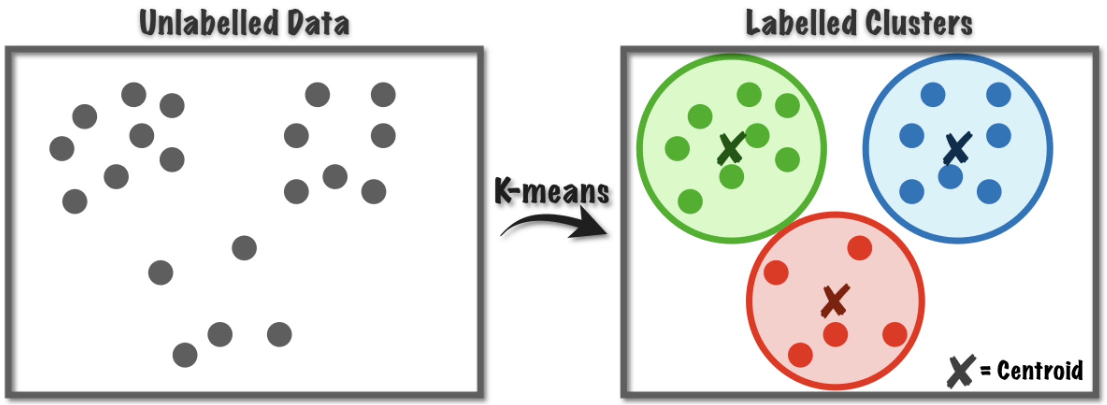
Methodology

The goal of this work is to find the optimal position of several Metro stops based on the venues around them → Usage of **K-means** (unsupervised clustering algorithm) where the centroids of each cluster will be the optimized Metro stop



```
Input:
    D= {t1, t2, .... Tn }   // Set of elements
    K                       // Number of desired clusters
Output:
    K                       // Set of clusters
K-Means algorithm:
  Assign initial values for m1, m2,.... mk
  repeat
    assign each item ti   to the clusters which has the closest mean;
    calculate new mean for each cluster;
  until convergence criteria is met;
```

K-Means pseudocode (source: Research Gate)

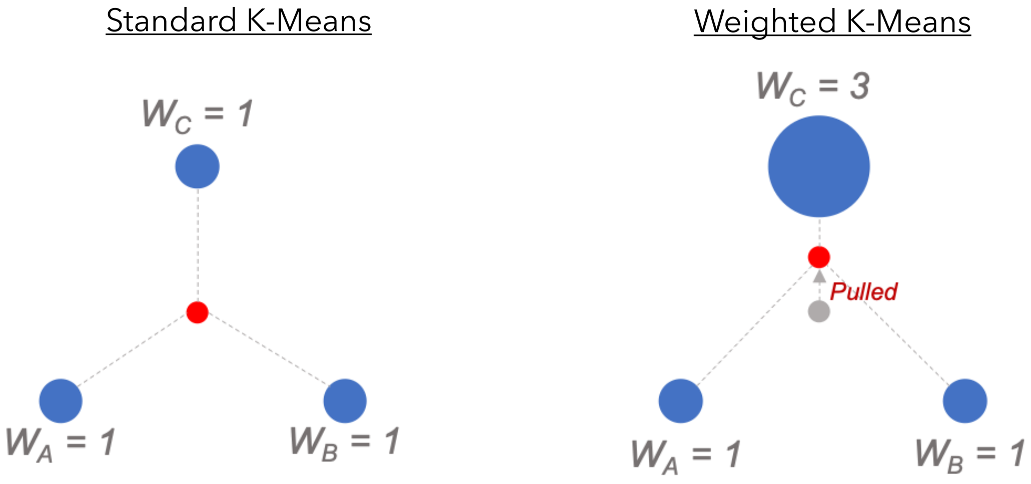


K-Means scheme (source: https://www.towardsdatascience.com)

Methodology

However, it has been seen that despite having only 20 venues, the “Campus Nord” stop is the one where its variety is the highest:

- This observation suggests that applying the standard K-means algorithm on our data may penalize such cases.
- As an alternative, the **weighted K-Means** algorithm is taken into account.



Standard K-Means vs weighted K-Means (source: https://www.towardsdatascience.com)

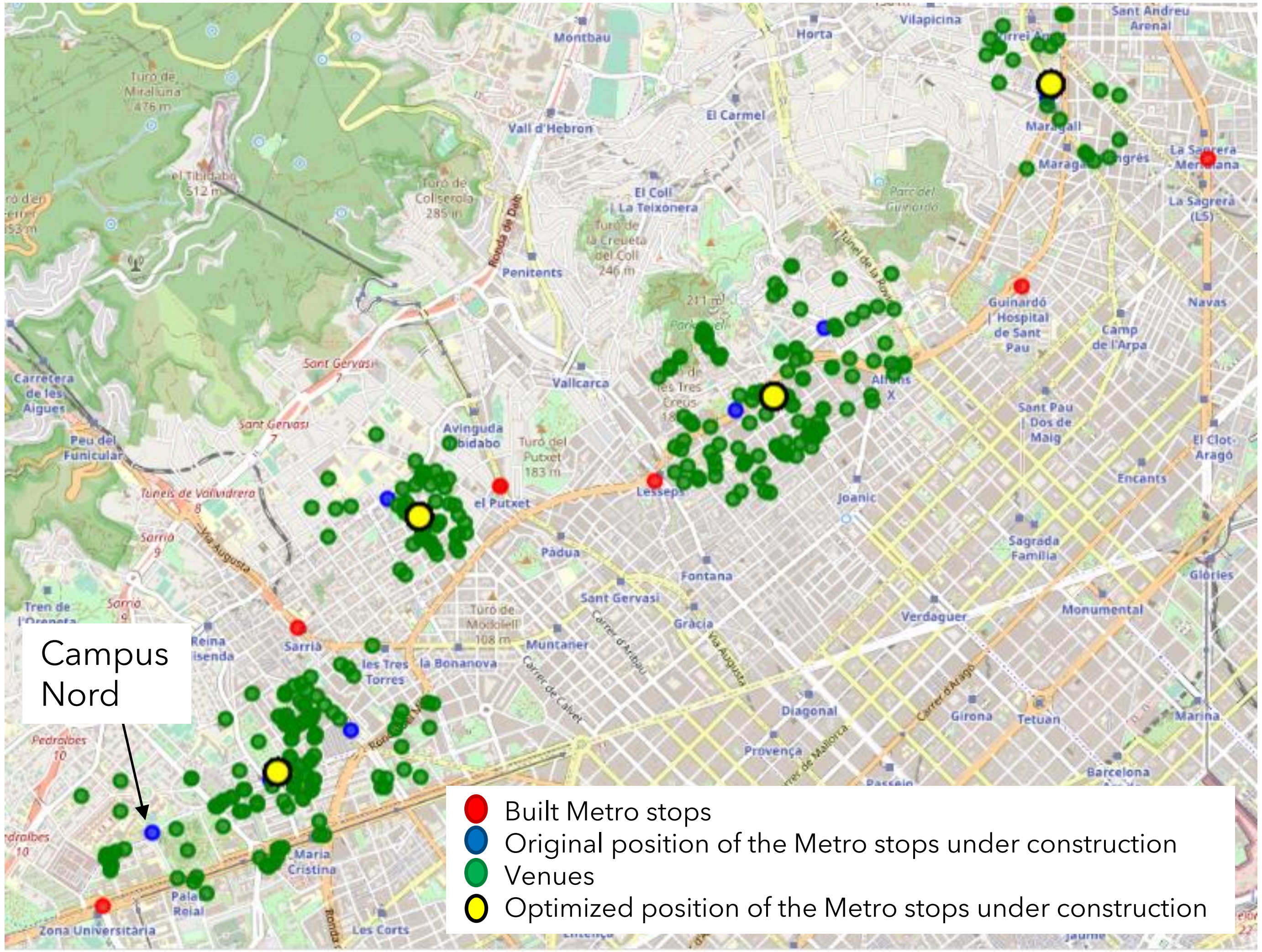
Venue Category	Weight (W_i)
Education, Health and transportation	10
Culture & nature	2.5
Hotel & Restaurant, nightlife, shopping, sport, others	1

Results

Standard K-means

According to the Elbow method, the optimal number of clusters is four

Metro stop under construction	Distance to the optimized stop (m)
Campus Nord	**789.53**
Manuel Girona	64.72
Prat de la Riba	475.16
Mandri	204.85
Muntanya	231.64
Sanllehy	475.69
Maragall	77.26



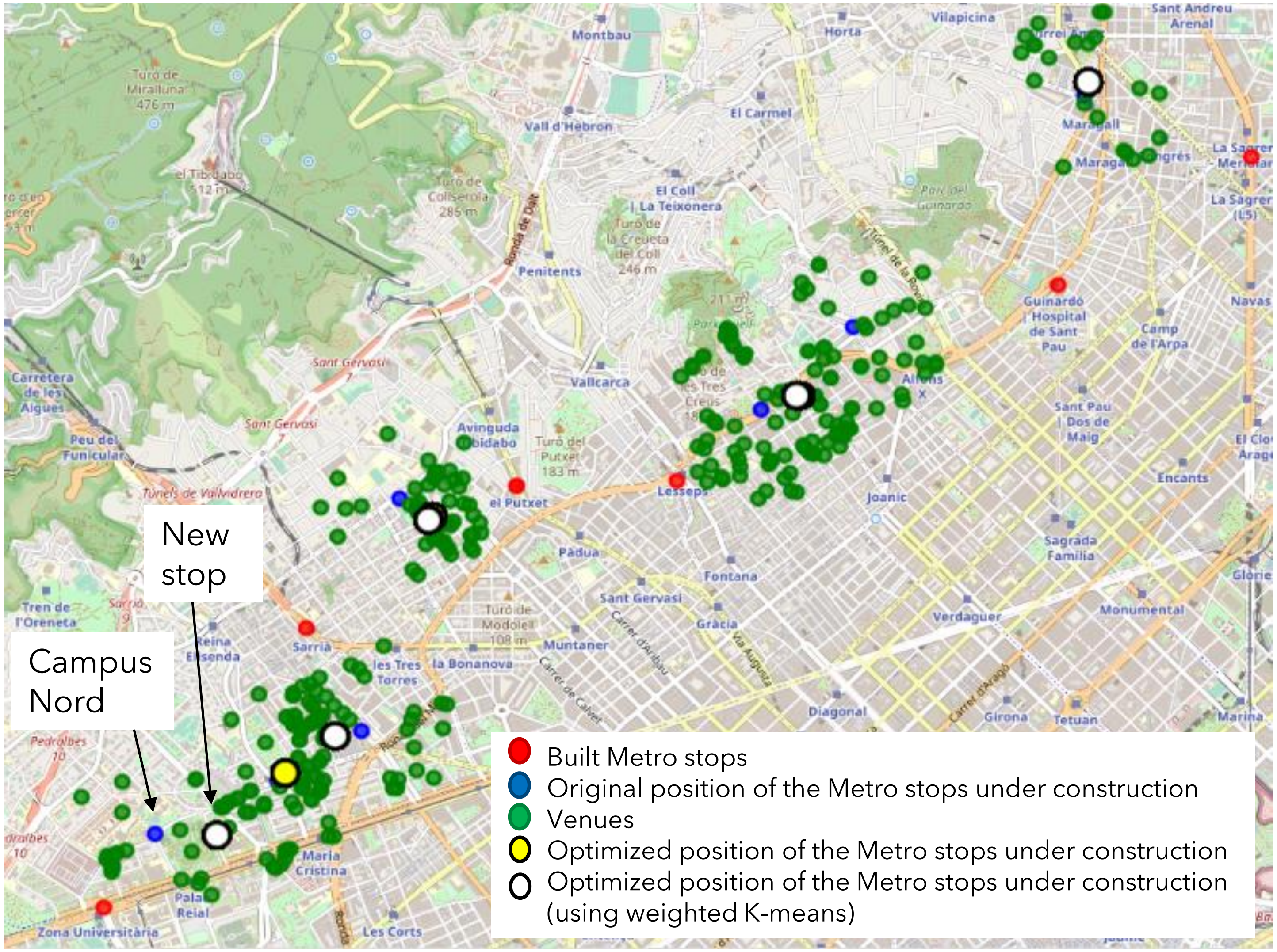
It can be seen that the optimal stop is too far from "Campus Nord"

Results

Weighted K-means

According to the Elbow method, the optimal number of clusters is five

Metro stop under construction	Distance to the optimized stop (m)
Campus Nord	**338.87**
Manuel Girona	408. 46
Prat de la Riba	143.58
Mandri	200.08
Muntanya	211.96
Sanllehy	489.84
Maragall	77.26



New stop

Campus Nord

🔴 Built Metro stops
🔵 Original position of the Metro stops under construction
🟢 Venues
🟡 Optimized position of the Metro stops under construction
○ Optimized position of the Metro stops under construction (using weighted K-means)

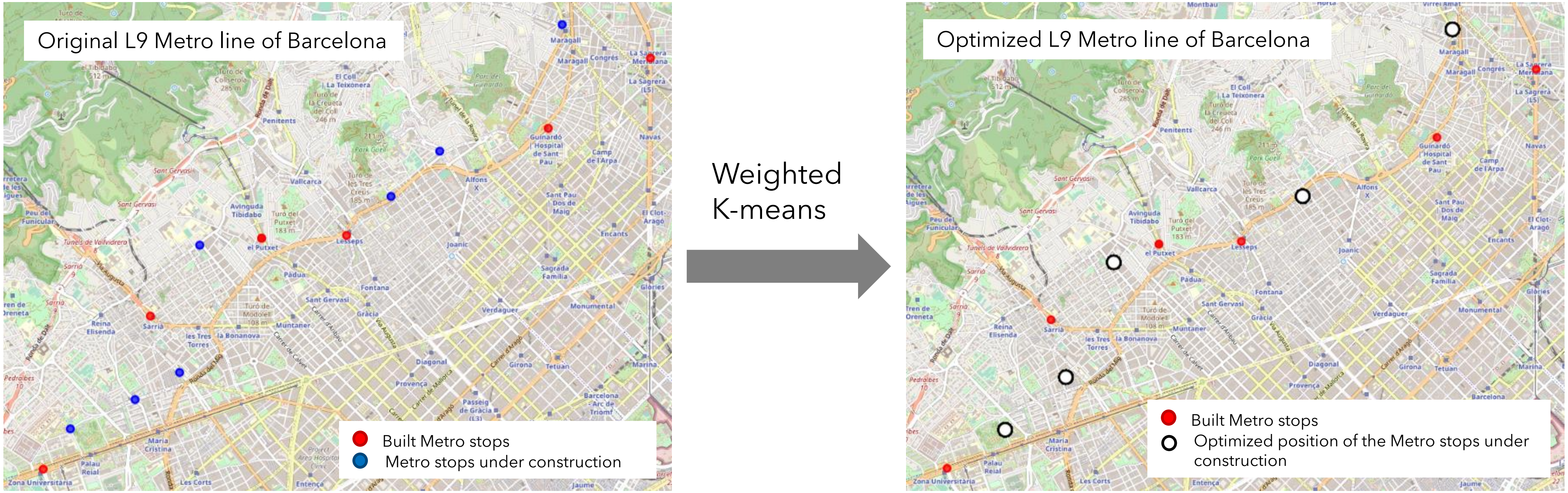
In this case, the agorithm adds a new Metro stop closer to "Campus Nord"

Conclusions and future work

- For the L9 of Barcelona's Metro, a study to find the optimal position of the stops under construction based on the venues around them has been done.
    - A Jupyter Notebook with the Python code is available under: https://github.com/laiaalcaraz/Coursera_Capstone
- Two different unsupervided clustering algorithms have been applied:
    - The weighted K-means algorithm provides better results because it takes the variety of venues around the" Campus Nord" stop into account.
    - The optimized "Maragall" stop is less than 80 m away than the original stop, so the original one can be kept.

Metro stop under construction	Distance to the optimized stop (m) using standard K-means	Distance to the optimized stop (m) using weighted K-means
Campus Nord	**789.53**	**338.87**
Manuel Girona	64.72	408. 46
Prat de la Riba	475.16	143.58
Mandri	204.85	200.08
Muntanya	231.64	211.96
Sanllehy	475.69	489.84
Maragall	77.26	77.26

Conclusions and future work



Original L9 Metro line of Barcelona

● Built Metro stops

● Metro stops under construction

Weighted K-means

Optimized L9 Metro line of Barcelona

● Built Metro stops

○ Optimized position of the Metro stops under construction

- In the future, this optimization study could be improved by taking into account the popultion density and also the proximity of bus stops, other Metro lines, train stops, etc.