CAPSTONE PROJECT - THE BATTLE OF NEIGHBORHOODS

Created with IBM Watson Studio

Muzammil Ahmed

April 1, 2020

TABLE OF CONTENTS

SUITABLE NEW STORE LOCATIONS IN PARIS FOR A FASHION RETAILER

This notebook contains multiple parts:

This report contains multiple parts (which are consistent with the Jupyter notebook

- ✓ A description of the problem and a discussion of the background
- ✓ A description of the data and how it will used to solve the problem
- ✓ Methodology and Exploratory Data Analysis
- ✓ Inferences, Discussion and Conclusions.

INTRODUCTION AND DISCUSSION OF THE BUSINESS OBJECTIVE AND PROBLEM

Locations for New Fashion Stores in High Traffic Areas in Paris France

The Task At Hand

A digitally native vertical fashion retailer, with a substantial e-commerce footprint, has begun the rollout of brick and mortar stores as part of their Omni channel retail strategy. After rolling out stores in a few select cities by guessing where the best locations were to open, as part of their store expansion for Paris they have decided to be more informed and selective, and take the time to do some research.

I have given the exciting task of assisting them to make data-driven decisions on the new locations that are most suitable for their new stores in Paris. This will be a major part of their decision-making process, the other being on the ground qualitative analysis of districts once this data reviewed and studied.

The fashion brand not considered as high end, there positioned in the upper end of the fast fashion market. As such, they do not seek stores in the premium upmarket strips like Avenue Montaigne, but rather, in high traffic areas where consumers go for shopping, restaurants and entertainment. Foursquare data will be very helpful in making data-driven decisions about the best of those areas.

Criteria

Qualitative data from another retailer that they know, suggests that the best locations to open new fashion retail stores may not only be where other clothing is located. This data strongly suggests that the best places are in fact areas that are near ***French Restaurants, Cafés and Wine Bars***. Parisians are very social people that frequent these place often, so opening new stores in these locations is becoming popular.

The analysis and recommendations for new store locations will focus on general districts with these establishments, not on specific store addresses. Narrowing down the best district options derived from analysis allows either further research to be conduct, advising agents of the chosen district, or on the ground searching for specific sites by the company's personnel.

Why Data?

Without leveraging data to make decisions about new store locations, the company could spend countless hours walking around districts, consulting many real estate agents with their own district biases, and end up opening in yet another location that is not ideal.

Data will provide better answers and better solutions to their task.

Outcomes

The goal is to identify the best districts - *Arrondissements* - to open new stores as part of the company's plan. The results will be translate to management in a simple form that will convey the data-driven analysis for the best locations to open stores.

THE DATA SCIENCE WORKFLOW

Data Requirements

The main districts in Paris are divided into 20 *Arrondissements Municipaux* (administrative districts), shortened to *arrondissements*.

The data regarding the districts in Paris needs to research and a suitable useable source identified. If it founds but not in a useable form, data wrangling and cleaning will have to be performed.

The cleaned data will used alongside Foursquare data, which is readily available. Foursquare location data will be leveraged to explore or compare districts around Paris, identifying the high traffic areas where consumers go for shopping, dining and entertainment - the areas where the fashion brand are most interested in opening new stores.

The Data Science Workflow for Part 1 & 2 includes the following:

- Outline the initial data that is required:
  - District data for Paris including names, location data if available, and any other details required.
- Obtain the Data:
  - Research and find suitable sources for the district data for Paris.
  - Access and explore the data to determine if it can manipulated for our purposes.
- Initial Data Wrangling and Cleaning:
  - Clean the data and convert to a useable form as a data frame.

The Data Science Workflow for the next section:

- Data Analysis:
  - Foursquare location data will leveraged, to explore or compare districts around Paris.
  - Identifying the high traffic areas using Data Visualization and Statistical Analysis.
- Machine Learning:
  - Analysis and Visualization with Clustering.
  - Data Visualization using Choropleth Mapping.

DATA RESEARCH AND PREPARATION

Import the Paris District Data

Arrondissements Municipaux for Paris CSV (administrative districts)

Paris is divided into 20 Arrondissements Municipaux (or administrative districts), shortened to just arrondissements. They and normally referenced by the arrondissement number rather than a name.

Data for the arrondissements is necessary to select the most suitable of these areas for new stores.

Initially looking to get this data by scraping the relevant Wikipedia page (https://en.wikipedia.org/wiki/Arrondissements_of_Paris), fortunately, after much research, this data is available on the web and can be manipulated and cleansed to provide a meaningful dataset to use.

Data from Open|DATA France:
https://opendata.paris.fr/explore/dataset/arrondissements/table/?dataChart

Also available from Opendatasoft:
https://data.opendatasoft.com/explore/dataset/arrondissements%40parisdata/export/

The data was imported from the source, but as can be seen, was not in the right format.

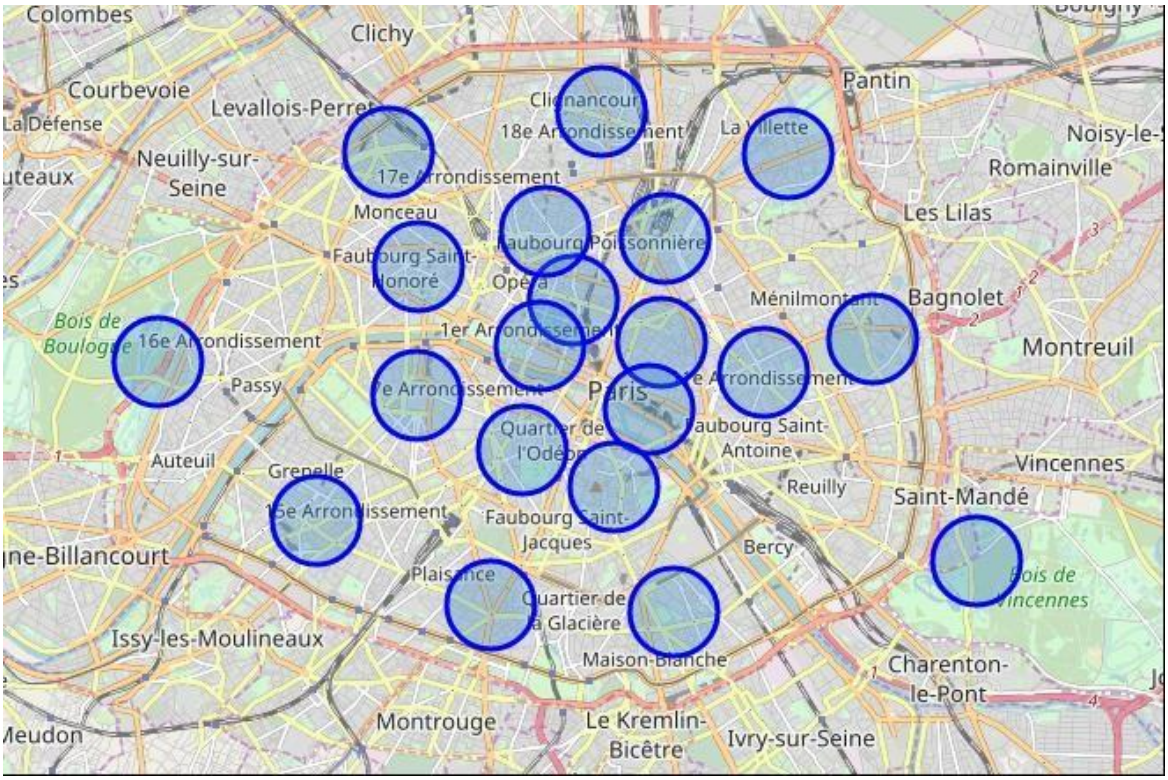
	CAR	NAME	NSQAR	CAR.1	CARINSEE	LAR	NSQCO	SURFACE	PERIMETRE	Geometry_X	Geometry_Y
0	3	Temple	750000003	3	3	3eme Ardt	750001537	1170882828	4519264	48.862872	2.360001
1	19	Buttes-Chaumont	750000019	19	19	19eme Ardt	750001537	6792651129	11253182	48.887076	2.384821
2	14	Observatoire	750000014	14	14	14eme Ardt	750001537	5614877309	10317483	48.829245	2.326542
3	10	Entrepot	750000010	10	10	10eme Ardt	750001537	2891739442	6739375	48.876130	2.360728
4	12	Reuilly	750000012	12	12	12eme Ardt	750001537	16314782637	24089666	48.834974	2.421325
5	16	Passy	750000016	16	16	16eme Ardt	750001537	16372542129	17416110	48.860392	2.261971

After some data wrangling and cleaning – renaming and dropping unnecessary columns - the data frame was in a structure that might be use.

	Arrondissement_Num	Neighborhood	French_Name	Latitude	Longitude
0	3	Temple	3eme Ardt	48.862872	2.360001
1	19	Buttes-Chaumont	19eme Ardt	48.887076	2.384821
2	14	Observatoire	14eme Ardt	48.829245	2.326542
3	10	Entrepot	10eme Ardt	48.876130	2.360728
4	12	Reuilly	12eme Ardt	48.834974	2.421325

METHODOLOGY AND EXPLORATORY DATA ANALYSIS

The GeoPy library is used to geocode location data and get location coordinates. It was import here to get the latitude and longitude values of Paris. From this using folium, I created a map of Paris with the location of the 20 districts superimposed.



As above, the image of the neighborhoods - *Arrondissements* - are located in a circle around central Paris. Although not apparent in the image, the circle of neighborhoods is quite large, making searching for new store locations a laborious task. Therefore, out task, using data is to reduce the amount and target only a few districts.

Foursquare

Use the Foursquare API to explore the Arrondissements of Paris (Neighborhoods)

After setting up the Foursquare API, we can explore the geolocation data. Exploratory data analysis allows us to look at what we are dealing with, and is this case, the first district in our data frame is explored to become familiar with the data (for the data we here use the French descriptive arrondissement name).

The first arrondissement is identified as *3eme Ardt*

```
# Get the Neighborhood's latitude and longitude values.

neighborhood_latitude = paris.loc[0, 'Latitude'] # Neighborhood latitude value
neighborhood_longitude = paris.loc[0, 'Longitude'] # Neighborhood longitude value

neighborhood_name = paris.loc[0, 'French_Name'] # Neighborhood name

print('Latitude and longitude values of the neighborhood {} are {}, {}.'.format(neighborhood_name,
                                                                                 neighborhood_latitude,
                                                                                 neighborhood_longitude))
```

```
Latitude and longitude values of the neighborhood 3eme Ardt are 48.86287238, 2.3600009859999997.
```

We determine that the first neighborhood is 3eme Ardt and we were able to easily its latitude and longitude.

Passing the search parameters and location details for this district into the API, we're able to get the top 100 venues that are in the neighborhood 3eme Ardt within a radius of 500 meters. Not that 100 is the limit on the free Foursquare accounts, so this is a limitation we need to recognize throughout the entire analysis.

	name	categories	lat	lng
0	Mmmozza	Sandwich Place	48.863910	2.360591
1	Square du Temple	Park	48.864475	2.360816
2	Marché des Enfants Rouges	Farmers Market	48.862806	2.361996
3	Chez Alain Miam Miam	Sandwich Place	48.862781	2.362064
4	Chez Alain Miam Miam	Sandwich Place	48.862369	2.361950
5	Fromagerie Jouannault	Cheese Shop	48.862947	2.362530
6	Les Enfants Rouges	Wine Bar	48.863013	2.361260
7	Okomusu	Okonomiyaki Restaurant	48.861453	2.360879
8	Hôtel Jules & Jim	Hotel	48.863496	2.357395
9	Musée de la Chasse et de la Nature	Museum	48.861507	2.358624
10	Bontemps	Dessert Shop	48.863956	2.360725

OK, looks good. So now, we did the same for all of the areas - creating a nearby venues function for all the neighborhoods in Paris.

Just a few changes to the code to include all of the locations:

```
def getNearbyVenues(names, latitudes, longitudes, radius=500):

    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        print(name)

        # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            LIMIT)

        # make the GET request
        results = requests.get(url).json()["response"]['groups'][0]['items']

        # return only relevant information for each nearby venue
        venues_list.append([(
            name,
            lat,
            lng,
            v['venue']['name'],
            v['venue']['location']['lat'],
            v['venue']['location']['lng'],
            v['venue']['categories'][0]['name']) for v in results])

    nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in venue_list])
    nearby_venues.columns = ['French_Name',
                  'Latitude',
                  'Longitude',
                  'Venue',
                  'Venue Latitude',
                  'Venue Longitude',
                  'Venue Category']
```

In addition, we're able to generate a new data frame with all of the nearby venues for all of the neighborhoods.

	French_Name	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	3eme Ardt	48.862872	2.360001	Mmmozza	48.863910	2.360591	Sandwich Place
1	3eme Ardt	48.862872	2.360001	Square du Temple	48.864475	2.360816	Park
2	3eme Ardt	48.862872	2.360001	Marché des Enfants Rouges	48.862806	2.361996	Farmers Market
3	3eme Ardt	48.862872	2.360001	Chez Alain Miam Miam	48.862781	2.362064	Sandwich Place
4	3eme Ardt	48.862872	2.360001	Chez Alain Miam Miam	48.862369	2.361950	Sandwich Place
5	3eme Ardt	48.862872	2.360001	Fromagerie Jouannault	48.862947	2.362530	Cheese Shop
6	3eme Ardt	48.862872	2.360001	Les Enfants Rouges	48.863013	2.361260	Wine Bar
7	3eme Ardt	48.862872	2.360001	Okomusu	48.861453	2.360879	Okonomiyaki Restaurant
8	3eme Ardt	48.862872	2.360001	Hôtel Jules & Jim	48.863496	2.357395	Hotel
9	3eme Ardt	48.862872	2.360001	Musée de la Chasse et de la Nature	48.861507	2.358624	Museum

With new data frame from the data, it was then possible to check how many venues were return for each neighborhood, and it was possible to calculate how many unique venue categories there are. This is a useful statistic in itself.

This all very useful to accomplish our task of finding the best location areas for new stores, but also is great data to have access to as a resource in future planning of new stores down the track.

It was then possible to analyses each of the neighborhoods from the results, displaying how many venues of each category were in each neighborhood.

```
paris_onehot
```

]:

	Neighborhood	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	...	Udon Restaurant	Vegetarian / Vegan Restaurant	Venezue Restaur
0	3eme Ardt	0	0	0	0	0	0	0	0	0	...	0	0	
1	3eme Ardt	0	0	0	0	0	0	0	0	0	...	0	0	
2	3eme Ardt	0	0	0	0	0	0	0	0	0	...	0	0	
3	3eme Ardt	0	0	0	0	0	0	0	0	0	...	0	0	
4	3eme Ardt	0	0	0	0	0	0	0	0	0	...	0	0	
5	3eme Ardt	0	0	0	0	0	0	0	0	0	...	0	0	
6	3eme Ardt	0	0	0	0	0	0	0	0	0	...	0	0	
7	3eme Ardt	0	0	0	0	0	0	0	0	0	...	0	0	

An important part of the process was to group the rows by neighborhood, and take the mean of the frequency of occurrence of each category. This will be use in narrowing down the suitable neighborhoods for new stores.

Group rows by neighborhood and take the mean of the frequency of occurrence of each category

```
paris_grouped = paris_onehot.groupby('Neighborhood').mean().reset_index()
paris_grouped
```

:

	Neighborhood	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	...	Udon Restaurant
0	10eme Ardt	0.000000	0.020000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	0.020000	...	0.00
1	11eme Ardt	0.015152	0.015152	0.000000	0.00	0.00	0.000000	0.015152	0.000000	0.015152	...	0.00
2	12eme Ardt	0.000000	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	0.000000	...	0.00
3	13eme Ardt	0.000000	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	0.206897	...	0.00
4	14eme Ardt	0.000000	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	0.000000	...	0.00

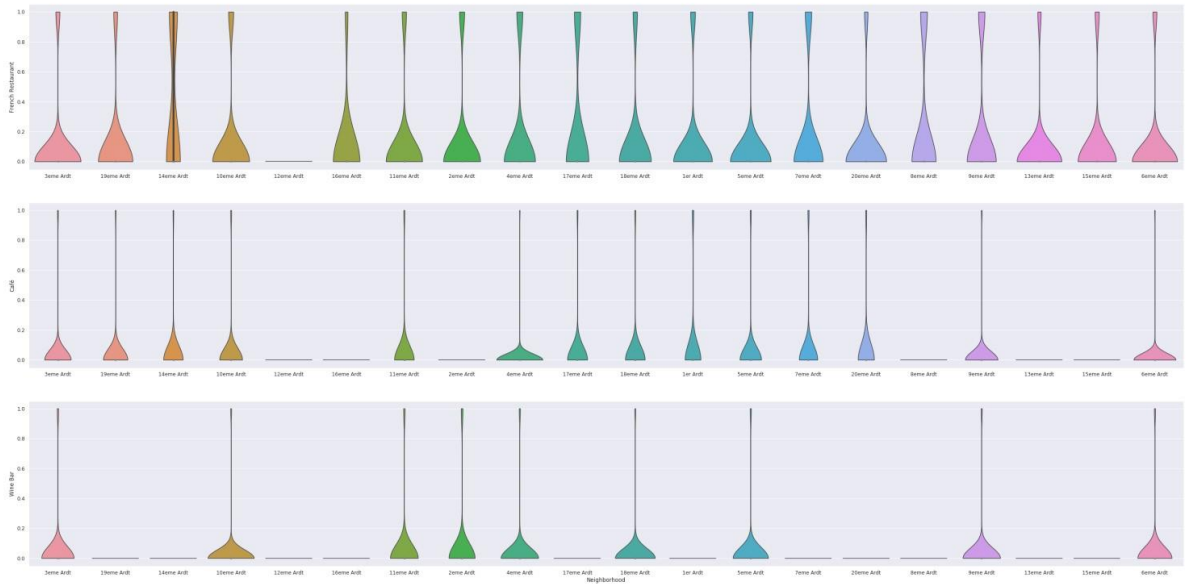
Referring back to the original task - the business types criteria specified by the client! *'French Restaurants', 'Cafés'* and *'Wine Bars'.*

Let's look at their frequency of occurrence for all the Paris neighborhoods, isolating the categorical venues.

These are various venue type that the client wants to have an abundant density of in the ideal store locations. I've used a violin plot from the seaborn library - it is a great way to visualize frequency distribution datasets, they display a density estimation of the underlying distribution.

Let's see the results.

Frequency distribution for the top 3 venue categories for each neighborhood (click to enlage)



The Neighborhoods

Therefore, as we can see from the analysis there are eight neighborhoods to open new stores - according to the criteria that they have the three specified venues in a great frequency (*French Restaurants, Cafés and Wine Bars).* They are as follows:

- 3eme Ardt
- 10eme Ardt
- 11eme Ardt
- 4eme Ardt
- 18eme Ardt
- 18eme Ardt
- 5eme Ardt
- 9eme Ardt
- 6eme Ardt

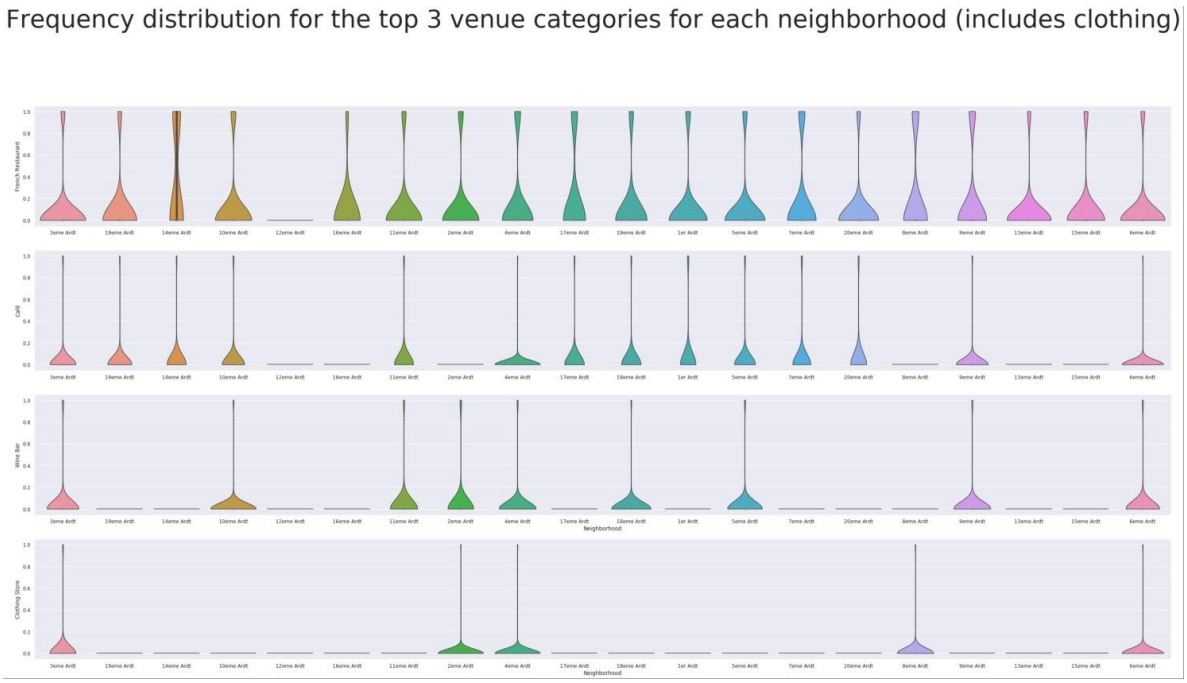
Let's take this further with some exploration and inferential analysis.

We have the eight neighborhoods that all include the venue category criteria.

But if we included the ***Clothing Store'*** venue category into the analysis, then we might be able to make some inferences based on the data, and domain knowledge of marketing and the industry, to focus the list.

Therefore, I looked at the venue category ***Clothing Store*** and plotted the result.

Therefore, there are five neighborhoods that have a significant frequency density of clothing stores. Let's add this to the analysis with the other 3 specified categories as below, then we get:

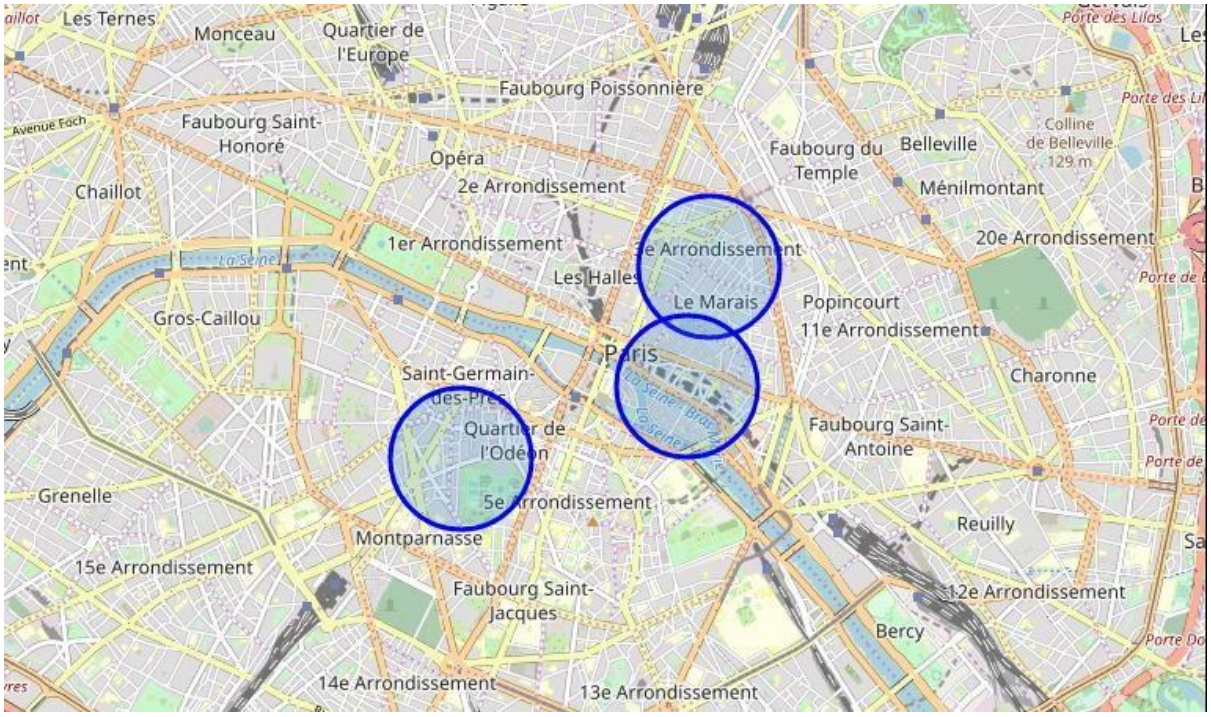


This narrows it down to just 3 neighborhoods as below:

	Arrondissement_Num	Neighborhood	French_Name	Latitude	Longitude
0	3	Temple	3eme Ardt	48.862872	2.360001
1	4	Hotel-de-Ville	4eme Ardt	48.854341	2.357630
2	6	Luxembourg	6eme Ardt	48.849130	2.332898

So where are our chosen districts?

Visualized on a folium map of Paris shows that they are all in fact quite close to the center of the city.



The **3rd arrondissement of Paris** is one of the 20 arrondissements of the capital city of France. In spoken French, this arrondissement is colloquially referred to as *troisième*.

The **4th arrondissement of Paris** is one of the 20 arrondissements of the capital city of France. In spoken French, this arrondissement is referred to as *quatrième*

The **6th arrondissement of Paris** is one of the 20 arrondissements of the capital city of France. In spoken French, this arrondissement is referred to as *sixième*

INFERENCES, DISCUSSION, CONCLUSION

Results - Chosen Neighborhoods

Inferential analysis using the data, as well as domain knowledge of retail and marketing, allow the list to be focused to just 3 neighborhoods from the previous 8.

The reasoning being that if the 3 criteria have been met - identifying neighborhoods that are lively with Restaurants, Cafés and Wine Bars - adding Clothing Stores into the mix of stores in the area is a significant bonus. Having some of the same category of stores in the same area - especially in fashion retail - is very desirable as a retailer.

So we can increase the criteria to include *Restaurants, Cafés, Wine Bars and Clothing Stores* - which narrows down and focuses the suggested districts for new stores to be located, and at the same time provides better locations for the brand.

So the final 3 prospective neighborhoods for new store locations are where 4 criteria are met:

- 3eme Ardt : Arrondissement 3, Temple
- 4eme Ardt : Arrondissement 4, Hotel-de-Ville
- 6eme Ardt : Arrondissement 6, Luxembourg

OBSERVATIONS AND DISCUSSION

I guess it's not a surprise that these districts are all very centrally located in the circular arrangement of Paris's arrondissements. Locations fitting the criteria for popular venues would normally be in central locations in many cities of the world.

From this visualization, it is clear that on a practical level, with no data to base decisions on, the circle of the 20 districts is very large, and researching and then visiting them all would be a daunting and time-consuming task. We have narrowed the search area down significantly from 20 potential districts to 3 that should suit the client's retail business.

Inferences

We have made inferences from the data in making the location recommendations, but that is exactly the point. There is no right or wrong answer or conclusion for the task at hand. The job of data analysis here is to steer a course for the location selection of new stores (i) to meet the criteria of being in neighborhoods that are lively with abundant leisure venues, and (ii) to narrow the search down to just a few of the main areas that are best suited to match the criteria.

Conclusions

There are many ways; this analysis could performed on different methodology and perhaps different data sources. I chose the method I selected as it was a straightforward way to narrow down the options, not complicating what is actually simple in many ways – meeting the criteria for the surrounding venues, and in my case, domain knowledge I have on the subject. I originally intended to use the clustering algorithms to cluster the data, but as it progressed, it became obvious that this only complicated the task.

The analysis and results are not an end-point, but rather a starting point that will guide the next part of the process to find specific store locations. The next part will involve domain knowledge of the industry, and perhaps, of the city itself. Nevertheless, the data analysis and resulting recommendations have greatly narrowed down the best district options based on data.

Without leveraging data to make focused decisions, the process could draw out and resulted in new stores opening in sub-standard areas for this retailer. Data has helped to provide a better strategy and way forward; these data-driven decisions will lead to a better solution in the end.

THANKS FOR TAKING PART IN MY DATA SCIENCE JOURNEY!

REFERENCES

Open|DATA France: https://opendata.paris.fr

Opendatasoft: https://data.opendatasoft.com

Quarters of Paris: https://en.wikipedia.org/wiki/Quarters_of_Paris

Arrondissements: https://www.data.gouv.fr/en/datasets/arrondissements-1/

Arrondissements of Paris: https://en.wikipedia.org/wiki/Arrondissements_of_Paris

Atelier Parisien d'Urbanisme: http://opendata.apur.org/

Paris – MapIt: https://global.mapit.mysociety.org/area/29746.html