A Restaurant in Venice

Jacopo Trabona

22 March 2021

1 Introduction

1.1 Background

It's is often said that **Venice, Italy**, is a one-of-a-kind city. Far from being a futile cliché this sentence is – indeed – grounded in reality, and that is not only due to Venice's exceptional topography, but also to its very unique economy – which, notoriously, relies entirely on tourism.

Hence, it goes without saying that this peculiar economy has recently suffered a dramatic contraction as a result of the current pandemic, which **drove down the island's yearly tourism influx by 71.7%.**

However, times of crisis are often ones of **business opportunities**, and 2020 in Venice has been no exception, with historical Hotels selling for half their values – some of these sells made the international newspapers' headlines – and a wide plethora of properties back on the market.

1.2 Problem

In this context, we would like to propose an analysis that will help interested investors in starting a commercial activity – specifically a **restaurant** – in the historical island of Venice, or in the adjacent (and commercially viable) islands of Murano and Giudecca.

Specifically, we want help stakeholders identify the **most suitable neighbourhood for such investment**, considering different factors such as population, tourism influx, and typologies of the existing activities in each neighbourhood. Considering the peculiar season the city is living, as well as the still uncertain estimates regarding the future of tourism industry, we aim at choosing the neighbourhood with a **balanced proportion of travellers and current residents**.

Eventually, once the most suitable candidate is selected, we will provide stakeholders with **details regarding the composition of the existing food services** in that neighbourhood, so that to enable them to choose the best type of food service to offer.

In order to do so, we will leverage the API provided by *Foursquare*, an online platform and application that allows users to register, categorise and rate venues across the world.

1.3 Interest

The analysis that follows is aimed at both **Italian and international investors**, with a particular focus on those who lack a certain domain knowledge about the very peculiarity of Venice. Venetians and regular visitors might also be interested in such data driven description of the city and of its neighbourhoods, and will likely find useful insights, which might either confirm of invalid their rule-of-the-thumb suppositions.

2 Data

2.1 Data Sources

In order to develop our analysis, we will be leveraging **location data** via the free access API provided by *Foursquare*. This will enable us to cluster Venice's neighbourhoods - Venice's main islands and the areas historically called "Sestieri" - on the basis of the typologies of venues users have reviewed over the years. Such an homogenous source of data will be integrated with a more diverse series of datasets, which will be functional to the **geocoding** of the neighbourhoods themselves, as well as to providing useful insights into features such as **population density** and **tourism demand**.

Considering the peculiarity of the analysis, some datasets were manually assembled into tabular form, while other were web-scraped or directly downloaded by open access platforms such as *Inside Airbnb*.

The sourced data sets – in order of appearance within the analysis – are the ones that follow:

- User-defined dataset listing Venice's neighbourhoods and respective **geographical coordinates**;

- User-defined dataset organising Venice's neighbourhoods on the basis of their **population density**;

- Dataset directly downloaded from the platform *Inside Aribnb*, which catalogues all ***Airbnb* offerings** in the city, providing useful insights such as price per night and location information;

- Data fetched via the *Foursquare* API, containing information about the categories of the **food services** offered in each Venetian neighbourhood;

- Publicly available, web-scraped data on **rental cost** for commercial actives in each Venetian neighbourhood;

Finally, it is worth to mention that, when closely examined, this collection of data is flawed in some respect.

Foursquare data itself, for instance, is sometimes too granular to be throughly functional, as it seldom results in a series of noisy, redundant entries - e.g. "Veneto Restaurant", "Italian Restaurant", "Mediterranean", and "Local Restaurant" all labelled under different categories.

Nevertheless, it cannot be underestimated how the aforementioned collection allows non-enterprise users to build an accurate and commercial viable analysis as the one that follows.

2.2 Data Cleaning

The amount of Data cleaning operations across the aforementioned data sources was minimal. This was primarily due to the fact that two of the main datasets used were **manually assembled in tabular form** by the author, since needed figures were only sparsely available, namely scattered across several individual pages in the web.

The data regarding Venetians venues were also clean by definition, since we were able to **define a function that fetched and directly assembled them** in tabular form as interacting with the *Foursquare* API. Then we limited our operations to regrouping several redundant entries under the same label "Italian Restaurant".

Data from *Inside Airbnb* was directly available in tidy, tabular form. We simply red the aforementioned .csv file into a data frame, and then selected the feature we needed for our analysis. Specifically, we decided to **derive the new features** combining data from *Inside Airbnb* and figures concerning Venetian population density in each neighbourhood.

Most of the cleaning operations were focused on the data frame containing figures about the **average rental costs in each neighbourhood**. That entailed, translating some of the columns' names from Italian to English, as well as removing unnecessary characters that prevented the use of the needed numerical data.

3 Methodology

3.1 General Approach

As briefly stated in the introduction, we set ourselves the goal of both **identifying the optimal neighbourhood** where to start a restaurant and providing stakeholders with **insights** as to the compositions of the existing food services in that neighbourhood.

Necessarily, most of our efforts concentrated into selecting the optimal neighbourhood, which we did by **progressively excluding the suboptimal ones** in the two clusters that were able to obtain.

Schematically, we proceeded as follows:

- Cluster Venetian neighbourhoods on the basis of their **popularity across tourists**;
- Select the **most popular cluster**;
- Cluster the remaining neighbourhoods on the basis of the **categories of existing food services**;
- Select the cluster with the **least diverse set of categories** - so that to have the widest plethora of investment options;
- **Select the most populated neighbourhood** - where current rent rates are below the average;
- Provide a **visual description** of the existing categories in the candidate neighbourhood;

3.2 Cluster Analysis

3.2.1 First Clustering

After excluding the island of Burano (which, due to its distance from the city centre, could be spatially considered an outlier in our dataset), we decided to cluster the Venetians neighbourhoods and main islands on the basis of their **popularity among visitors to the city**.

In this instance, limited by the open-access resources at our disposal, we were not able to obtain any specific data about **tourism influx per neighbourhood**. Accordingly, within a certain range of accuracy, we decided to infer them from **the number of Airbnb listings** in each neighbourhood, which is publicly available on tabular form via the platform *Inside Airbnb*.

Additionally, since neighbourhoods widely differs in size, we derived more proportional data by creating a variable that describes the **ratio between the number of residents in each neighbourhood and the number of corresponding listings.**

We also extracted the **mean price per night** for each neighbourhood and store these value in a data frame. Presumably, this will be a rather descriptive feature, our assumption being that the higher the price guests are keen to pay for their accommodation the more central and exclusive the neighbourhood – and, as a consequence, the higher the price costumers of a restaurant would be also keen to pay.

Neighborhood	Population	Number of listings	Mean price	Residents per listing
Cannaregio	15662	1552	142.097938	10.091495
Castello	11642	1484	140.082884	7.845013
Dorsoduro	6429	603	157.510779	10.661692
Giudecca	4481	139	157.719424	32.237410

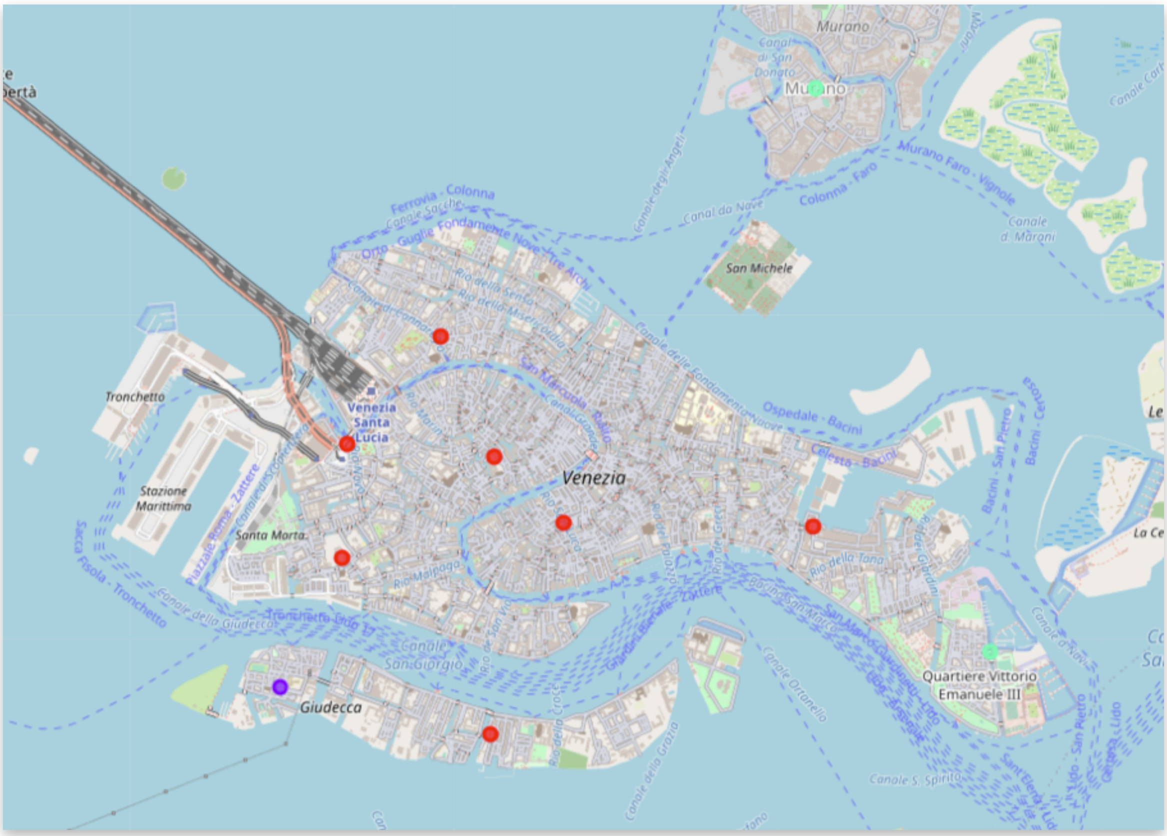
Neighborhood	Population	Number of listings	Mean price	Residents per listing
Murano	4338	82	95.609756	52.902439
Sacca Fisola	1452	3	95.000000	484.000000
San Marco	3788	961	183.387097	3.941727
San Polo	4628	651	148.671275	7.109063
Sant'Elena	1861	51	105.549020	36.490196
Santa Croce	4996	596	139.203020	8.382550

From the table above, we selected only the **Mean price** and **Residents per listing** columns in order to implement our cluster analysis via the popular algorithm **KMeans**. As it is often the case, we decided not to refer to a predetermined number of clusters. Rather, we will choose the optimal number of clusters using the widely adopted **Elbow Method** and **Silhouette Coefficients Method** in a complementary manner.

The implementation of these methods resulted in **3 optimal clusters**, each including their corresponding neighbourhoods as follows.

Neighborhood	Cluster
Cannaregio	0
Castello	0
Dorsoduro	0
Giudecca	0
Murano	2
Sacca Fisola	1
San Marco	0
San Polo	0
Sant'Elena	2
Santa Croce	0

We then decided to visualise each cluster in a **map**.



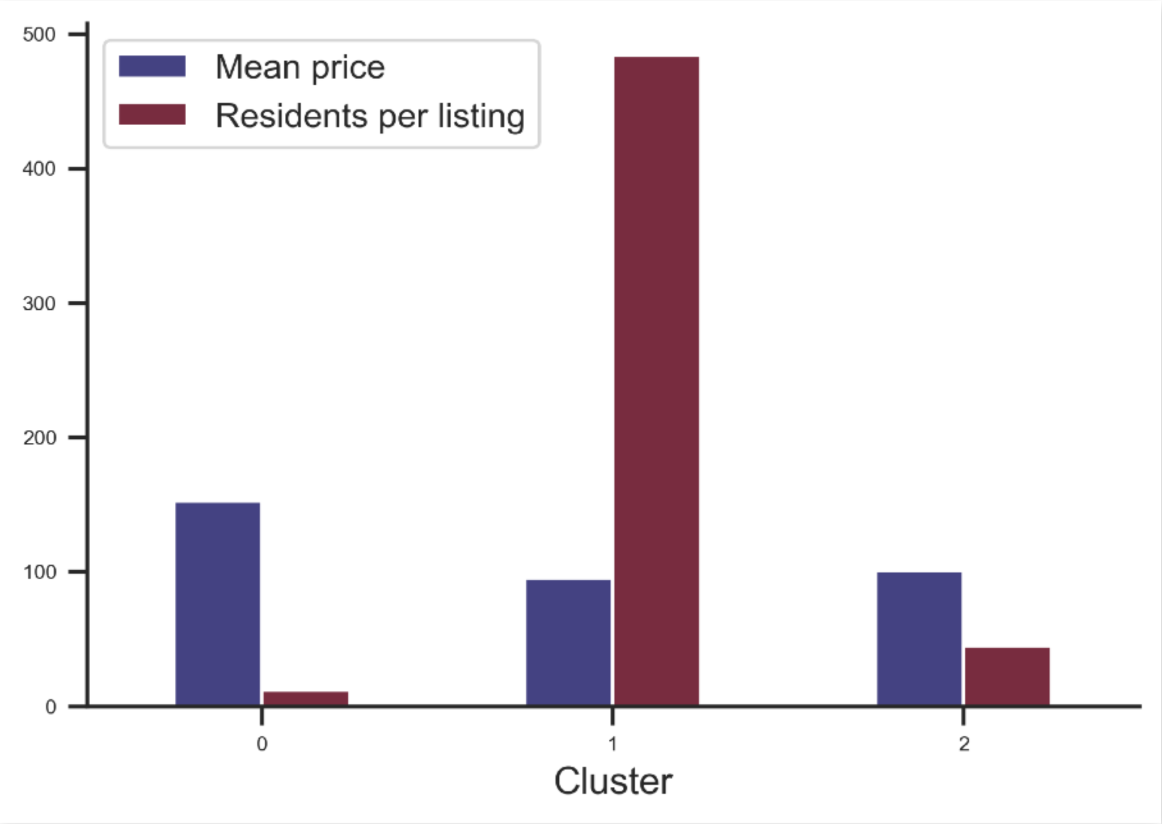
Map of Venice’s main island with the resulting cluster superimposed on top.

We noticed that, not surprisingly, **different clusters correspond to different distances from the city centre.**

This is likely due to the fact that central ares and neighbourhoods are the most popular choices among visitors who decide to spend the night in Venice.

We confirmed that by **exploring the characteristics which define each cluster** – which we obtain by taking the mean of both the mean listing price and the residents per listing ratio in each neighbourhood.

This resulted in the following **bar chart**, which clearly visualises the composition of each cluster.



Bar chart displaying the Mean price and mean Residents per listing ratio for each cluster.

Cluster 0

We can assume that the first cluster, which includes the neighbourhoods of Cannaregio, Castello, Dorsoduro, Giudecca, San Marco, San Polo, and Santa Croce, is **the one for which tourism demand is the highest**, since is the one whose neighbourhoods has the highest listing prices, and, by far, the one with the highest concentration of *Aribnb* listings (namely the lowest residents per listing ratio).

Cluster 1

Cluster 1, including the island of Sacca Fisola only, is **by far the least popular among tourists**, since only 3 accommodations (with rather low prices on average) are listed.

Cluster 2

Cluster 2, including the island of Murano and the neighbourhood of Sant'Elena, is not the most popular among visitors to the city. However, as compared to Cluster 1, it still hosts a rather high amount of *Airbnb listings* and it is probably targeted by those who are looking for **cheaper and quieter accommodations**, as compared to the congested city centre.

From these findings, it is clear that opening a restaurant in each given cluster would correspond to a different choice, in turn implying a different targeted audience, and accordingly to a different business model.

Especially considering the peculiar season the city is leaving due to the pandemic, we assumed that **stakeholders might be interested in Cluster 0**, namely the one which is usually by far the most targeted by tourists, and where – because of the aforementioned current drop in tourism figures – the most profitable business opportunities are likely to be found.

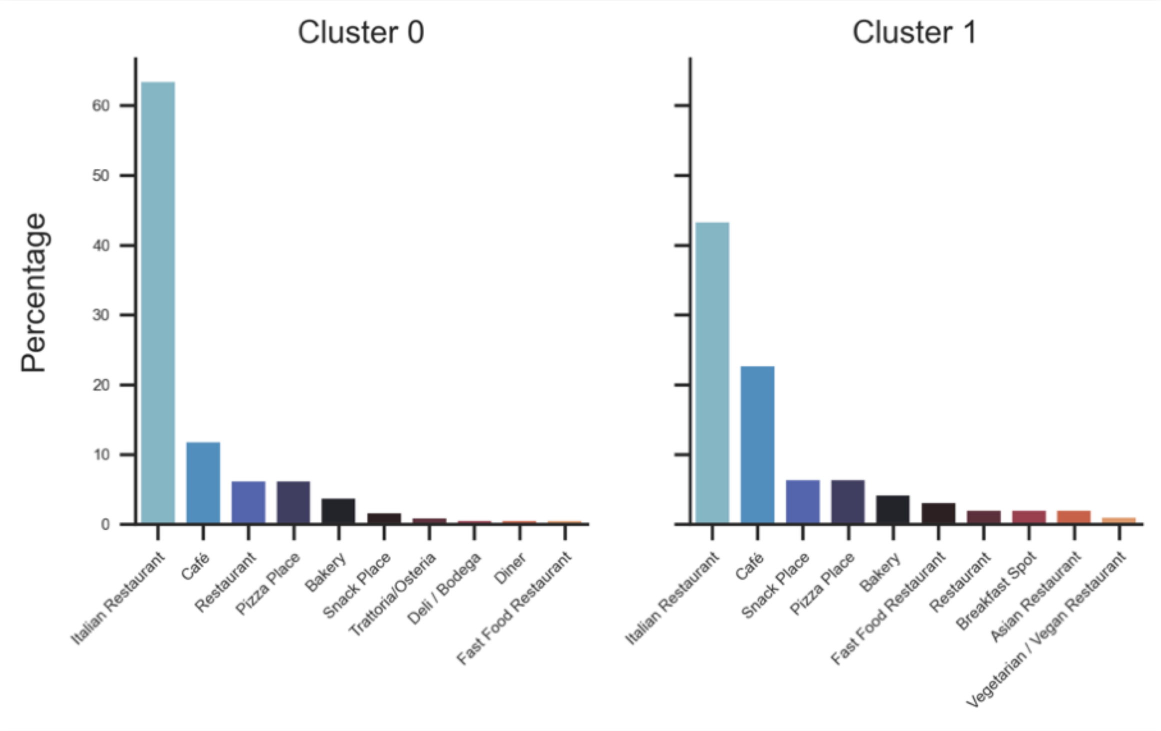
3.2.2 Second Clustering

In our second clustering process we used data we obtained by interacting with the *Foursquare* API. We thereby extracted the most popular food services in each neighbourhood and implemented the same algorithm and approaches we used in the first clustering.

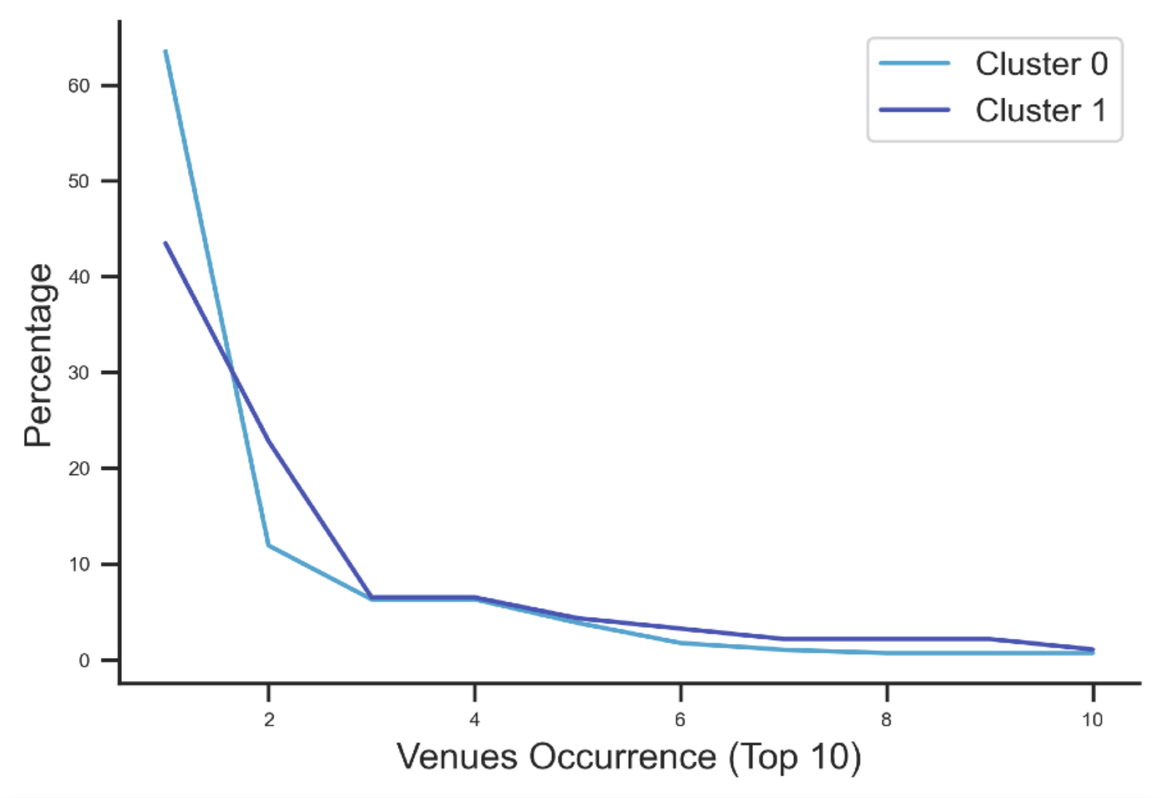
This resulted in **2 optimal clusters**, Cluster 0 including the neighbourhoods of Cannaregio, Dorsoduro, San Marco, San Polo, and Santa Croce, and Cluster 1 second including the neighbourhoods of Castello and Giudecca

From exploring the data we understood that the main factor leading our **KMeans** algorithm to produce these clusters was a dramatic preponderance of Italian restaurants in **Cluster 0**.

We can clearly see that in the next two graphs, which **plot the percentage of occurrence of the 10 most frequent categories in each cluster**.



Bar charts plotting the 10 most frequent categories in each cluster.



Line plot displaying the percentage of occurrence of 10 most frequent categories in each neighbourhood.

Both clusters share the first and second most frequent categories, namely Italian Restaurant and Café.

However, **Cluster 0 has a much more defined identity**, with more than 60% of the venues belonging to the category Italian Restaurant, while in Cluster 1's categories seem to be more proportionally distributed.

3.3 Further Analysis

The **lack of diversity in Cluster 0 makes it a good candidate for our Restaurant**, since it will be easier to choose a Restaurant category which is currently under-represented in the neighbourhoods belonging to Cluster 0.

In other words, especially considering that Cluster 0 is both bigger and contains more central neighbourhoods than Cluster 1, our assumption is that **any category of restaurant which is not "Italian Restaurant" will be meet customers diversified demand more easily** than it would do in the more diverse Cluster 1 – where Restaurants belonging to other categories are much more numerous.

Thusly, we sharpened the focus on 5 neighbourhoods – Cannaregio, Castello, Giudecca, San Marco, San Polo – and finally decided to refer to our initial conditions, namely that of choosing an area with a **balanced proportion of travellers and residents**, and, considering that the remaining neighbourhoods were originally clustered together on the basis of their popularity among tourists, we limited ourselves to select the **most populated one**.

This will allow our restaurant to be profitable even in the light of the current crisis vexing the travel industry.

In order to guide our choice we use a table as the one shown below.

Neighborhood	Population
Cannaregio	15662
Castello	11642
San Polo	4628
Giudecca	4481
San Marco	3788

The **most populated neighbourhood in our cluster is Cannaregio**, with a population density way above the average – and 4020 above the second most populated neighbourhood Castello**.**

However, since Castello's population is also more than twice as bigger as than that of the other remaining neighbourhoods, we decided to confirm our selection by **cross-referencing population figures with data regarding the rental cost within each neighbourhood**. The results are displayed below.

Neighborhood	Monthly_price_m2	For_rent
Giudecca	29.500000	30
Cannaregio	30.700001	100
San Polo	34.700001	100
San Marco	35.400002	100
Castello	40.500000	50

We were thereby able to conclude that **Cannaregio** is the optimal neighbourhood where to start a food service commercial activity, since not only is the most populated one, but it also cheaper then the majority of the other neighbourhoods, and – investors might be also interested in that – one where approximately 100 venues are available for rent.

4 Results

Our analysis unfolded through 3 main steps. The first two focused on clustering different Venetian neighbourhoods. The third one focused on elaborating on the findings we obtained through the first steps.

Clustering via **KMeans**, a popular unsupervised machine learning algorithm, is a technique whose most popular adoption is generally associated with approaches such as customer segmentation – usually by applying it to very large datasets. However, as part of the Capstone project of the IBM Professional Certificate in Data Science, we were tasked to apply this clustering technique to **location data**, which we were to obtain by making use of the aforementioned *Forsquare* API.

However, considering the goal that we set ourselves to – namely that of indicating an optimal neighbourhood where to start a restaurant in the city of Venice – our initial assumption was that such location data would not be very useful, unless integrated with a differently descriptive type of information.

By combining data regarding neighbourhoods population density, as well as the number and characteristics of the *Airbnb* listings in each neighbourhood, we were able to obtain two different sets of insights, which, correspondingly, implied **two different sets of clustered results**.

The first set allowed us to clearly identify **Cluster 0** as the **most popular among tourists**, since, not only it is the one where the ratio between the number of residents and the number of *Airbnb* listed is the lowest by far, but only the mean price per night guests of the listings are willing to pay is the highest.

For the neighbourhoods included in **Cluster 0** – namely Cannaregio, Castello, Dorsoduro, Giudecca, San Marco, San Polo, and Santa Croce – we then applied **KMeans** a second time, using the frequency of occurrence of each unique food service category fetched via the *Foursquare* API.

This allowed us to discover two more clusters, which, although presenting similar sets of categories, did so in rather different proportions. Specifically **Cluster 0** is much more defined, with more 60% of its venues being Italian restaurants. **Cluster 1**, on the other hand, is more diverse and thereby might present some criticality if stakeholders were to choose among which typology of restaurant to cover.

Accordingly, we finally analysed the neighbourhoods in **Cluster 0**, and find our that **Cannaregio** is the most populated one, and where, accordingly, profits are not necessarily bounded to the currently precarious tourism industry.

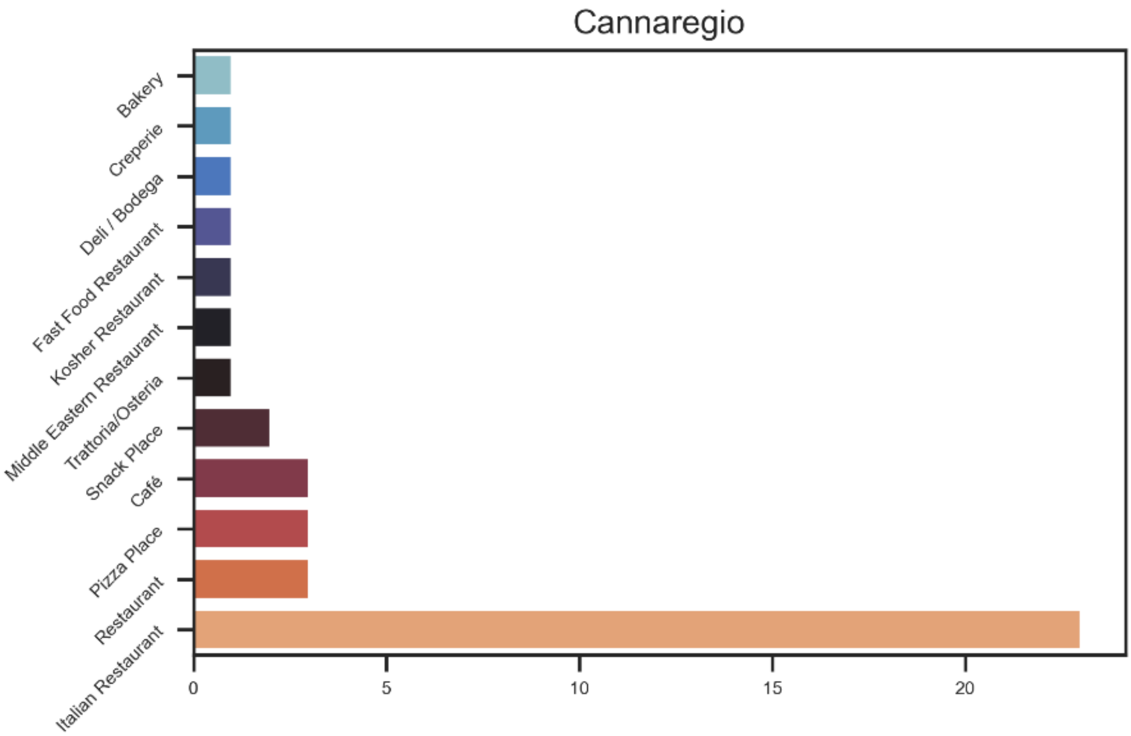
5 Discussion

Thanks to the analysis conducted in this report, stakeholders are presented with a defined choice of the **optimal neighbourhood** – **Cannaregio** – where to invest if the wanted to start a restaurant in the city of Venice.

Of course, throughout our analysis, with no other informative factor at our disposal, we assumed **profitability** and **risk aversion** as the main factors leading to the final choice.

Specifically, **risk aversion** was the reason why we identified the optimal neighbourhood as the one which is both the most populated and, in circumstances different from the ones currently ushered in by the pandemic, one of the most popular among tourists and visitors to the city.

Additionally, we can present stakeholders with the following visual description of the categories of the existing food services in the neighbourhood of **Cannaregio**.



From the plot displayed above, stakeholders would ideally be able to draw their own conclusions and, accordingly, chose the most pertinent typology of restaurant to invest in.

Our own suggestion, which directed our analytical approach, would be that of **choosing a category which is highly under-represented** – if not throughly absent – in the graph.

Considering the very diverse demand of our potential customers – both residents and visitors from all over the world – this might reveal to be the optimal strategy.

6 Conclusion

The combination of Machine Learning (ML) and traditional analytical techniques hereby employed led us to methodologically accurate definition of the optimal neighbourhood where, **as of March 2021**, to locate a restaurant in the city of Venice. However, in order for this analysis to be thoroughly useful, we recommend stake holders to bear in mind the following facts.

This analysis was extremely **time-sensitive**, and the reasoning that led to most of its finding was thoroughly grounded in the exceptionality of the current health emergency. Under different circumstances, a different course of action would have been probably taken into consideration.

By definition, this analysis was based on non-enterprise and **free access resources.** Many of the figures and datasets implied, despite being both accurate and reliable, represent only a fraction of the data that, were stakeholders to invest in enterprise based data collection, would be available for gaining more complete insights.
Above all, data regarding the aforementioned venue categories was obtained using the free access *Foursquare* Sandbox Tier Accounts, which limits the number of entries displayed. This implies that our description of the neighbourhood of Cannaregio should be taken only as reference, which, despite providing a realistic account of the distribution of its top food services, necessarily left out an indefinite number of restaurants. Stakeholders should be thereby be cautious in drawing conclusions such as "there is no Asian Restaurant in the neighbourhood of Cannaregio". The data we visualised simply state that no Asian Restaurant is in the best 50 rated food services in the area of Cannareggio as reported in *Foursquare*.