Most efficient cultural holiday in Florence

by Francesca Scalesse

data: 03/06/2021

1. Introduction

1.1 Background

Florence is an italian city rich of history and cultural heritage sites like museums, churches... They are both in the city's centre or in peripheral zones. You can easily access to a lot of them because Florence is not a huge city so you can easily move even by foot. The possibility of cultural holiday are very large.

1.2 Problem

My client wishes spent a cultural holiday (a week) in Florence visiting as many cultural sites as possible. In particular the client is interested in museums and churches. The most part of them are in the city's centre. The problem is that public transport cannot go in this zone because it is a pedestrian area. The best way is go by foot. In order to visit as many churches and museums as possible, my client wishes to find an hotel in a zone with the highest concentration of these sites. Price is not a problem.

1.3 Interest

This work can be useful to anyone wants to spend an holiday in Florence visiting museums or churches for personal interest, job, study or research.

2. Data acquisition and cleaning

2.1 Data source

The source of data about the location of museums, churches and hotels is Foursquare[1].

2.2 Importing, cleaning data and feature selection

Data about churches and museums were taken using the website Foursquare to extract them using the keywords *Museo* and *Chiesa* (respectively museum and church in italian). The data were

acquired separately and transformed in data frames using pandas. Then these data frames were combined in a unique one, containing all the interest sites for the client. Informations about hotels' localization were taken from Foursquare and transformed in a second data frame using the previous procedure.

3. Methodology

The model used was a descriptive model. The first step consisted in applying a clustering model on the data about churches and museums, in order to find the zone with the highest concentration of this cultural sites. Subsequently, the position of the hotels was overlapped on the formed clusters. The cluster with the highest amount of churches or museums and containing at least one hotel was selected and the seven nearest hotels to the cluster's centroid were considered the best choice for the client.
The clustering algorythm used is K-Nearest Neighbours (KNN).This method wad used because of the following advantages:

- Quick calculation time
- Simple to interpret
- Versatility (useful both for regression and classification)
- High accuracy

When using KNN it is important to know the number of clusters. A good method to validate the optimal number of clusters is the elbow method[2]. The idea of the elbow method is to run k-means clustering on the dataset for a range of values of k (say, k from 1 to 10), and for each value of k calculate the Sum of Squared Errors (SSE). When k increases, the assigned centroids are closer to the clusters centroids. At some point there would not be sensitive improvement and a elbow shape would appear. This point would be is the optimal value for k.
After the clusters' creation, it was chosen the cluster with the highest amount of data. Then the distances between hotels and this cluster's center was calculate, in order to find some hotels which permit to easily reach the highest number of these cultural sites. It was chosen the first seven hotels in order to give the client a good choice, avoiding too far hotels. The method used is **haversine formula**[3], easily imported by scikit library. It determines the great-circle distance two points on a sphere, given their latitude and longitude, which are the data about churches museums and hotels.

This assumes the earth is a true sphere making the computation fast. The computation assumes the radius of the sphere is 1, so to get the distance in kilometers we multiply the output of the sklearn computation by 6371 km, the average radius of the earth (to get the distance in miles this number would be 3959 miles).

4. Results

4.1 Clustering with KNN

In this paragraph the data frame containing data about museums and churches were analyzed the clustering model KNN, in order to find the zone with the highest concentration of these sites. Before applying the KNN method the optimal number of clusters was found applying the elbow method. The k-means was applied using values of k (number of cluster) from 1 to 10 (Figure 1).

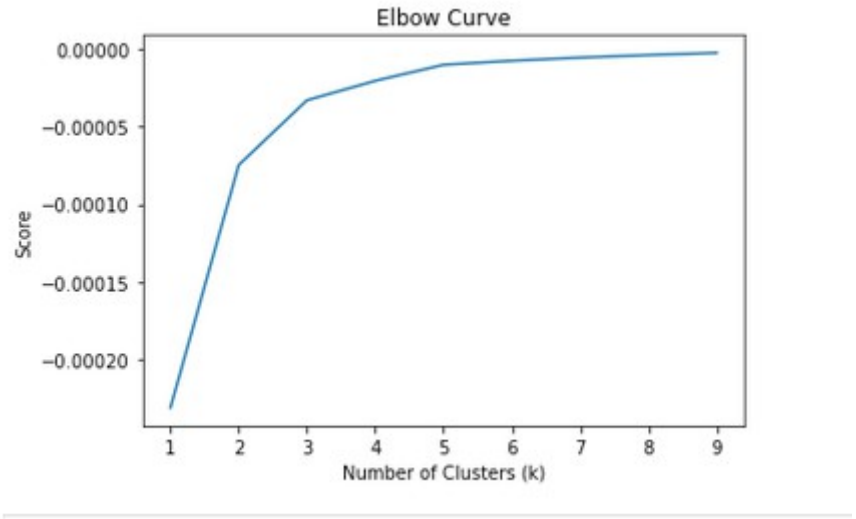

Figure 1. Elbow method on the data frame containing coordinates of Florence's churches and museums. The k-means was applied with the value of k from 1 to 10.

The improvement is very high until k = 3 and slow until k = 5. After that point there was not any significant changing. Therefore, five clusters was considered the best value. Subsequently, KNN algorithm was applied on the same data frame (Figure 2). The k-Means parameters used was these three:

- **init**: Initialization method of the centroids. Value was "k-means++", which selected initial cluster centers for k-means clustering in a smart way to speed up convergence.

- **n_clusters**: the number of clusters to form as well as the number of centroids to generate. Value was 5
- **n_init**: number of times the k-means algorithm will be run with different centroid seeds. The final results will be the best output of n_init consecutive runs in terms of inertia. The value was 10

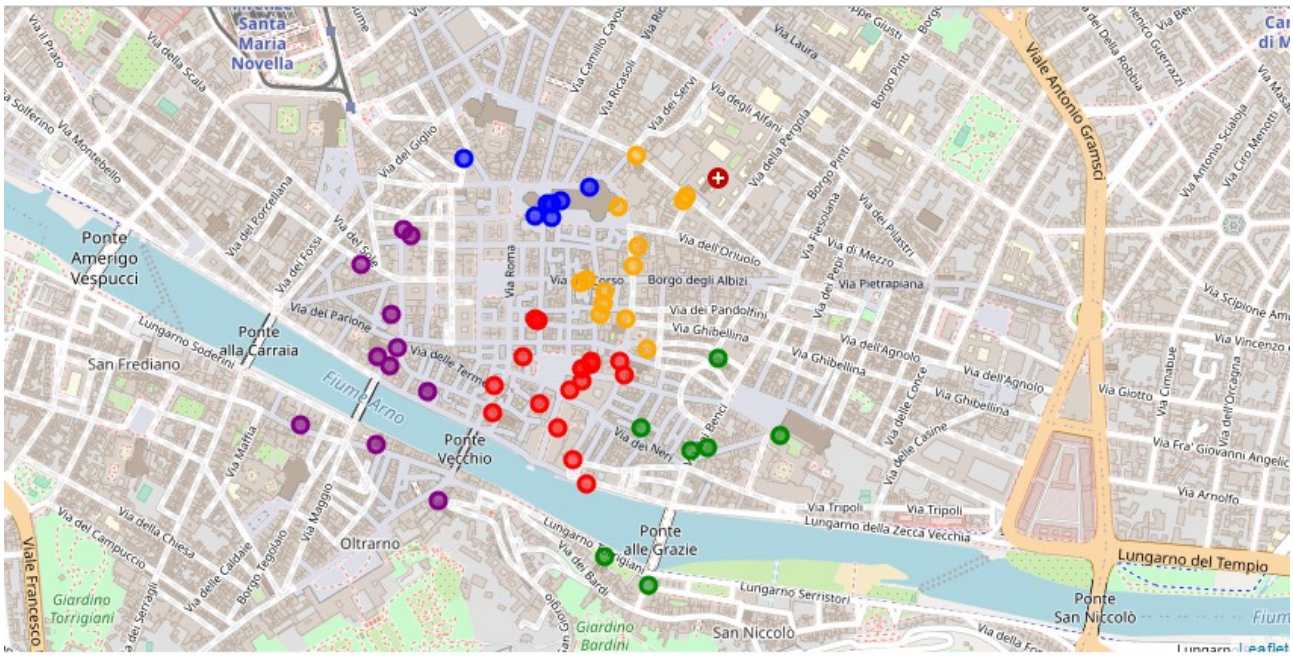

Figure 2. Visualization of clusters on the data frame containing Florence's museums and churches. Cluster 0 = blue, cluster 1 = red, cluster 2 = green, cluster 3 = purple, cluster 4 = orange.

The number of data points in every cluster is summarize in the **Table 1**. According these data the cluster with the highest amount of museums or churches is the n. 1 (18 data points, red circles in Figure 2).

Cluster	Data point in a cluster
0	7
1	18
2	7
3	11
4	13

Table 1. Number of data for every clusters. The cluster with the highest amount of cultural sites is the number 1 (red circles in the Figure 2).

4.2 Calculate the haversine distance between hotels and centroid of cluster 1

After data containing Florence's churches and museums were clustered (Figure 2) and discovered the most densest cluster (number 1, red circles), the data frame containing the hotels (black circles) was overlapped on the map containing the clusters (Figure 3).

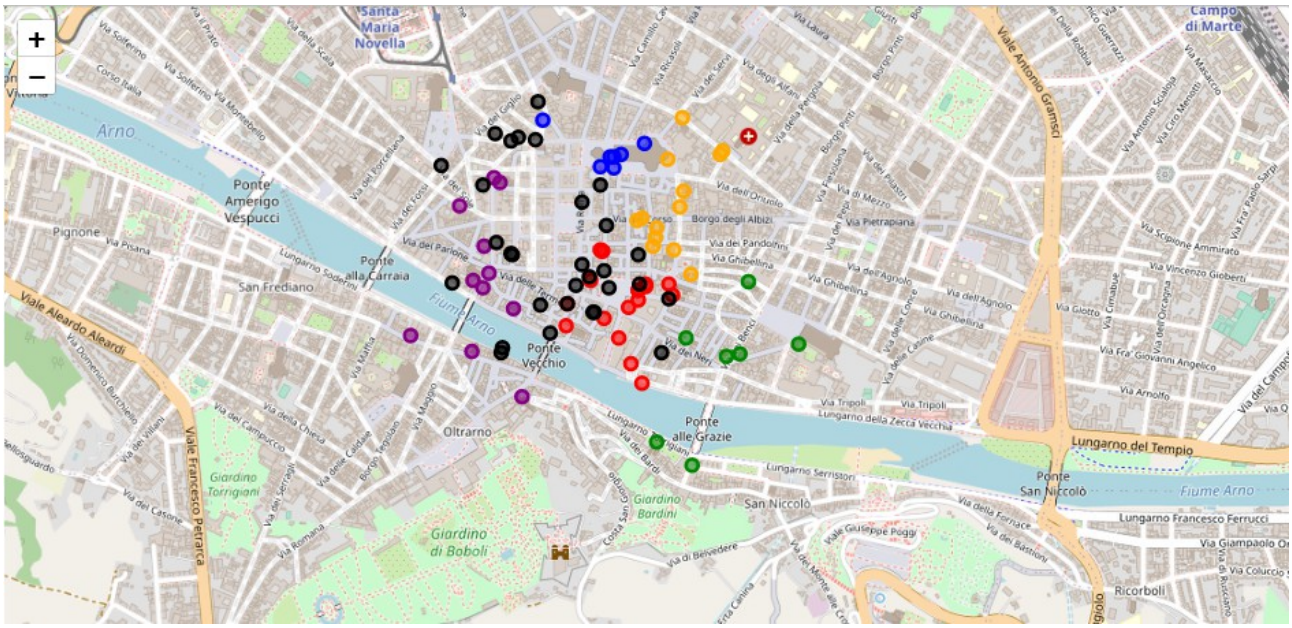

Figure 3. Data points of churches and museums (colored circles) and hotels (black circles). Cluster 0 = blue, cluster 1 = red, cluster 2 = green, cluster 3 = purple, cluster 4 = orange, hotels = black.

In the red clusters there are several hotels. The distance between hotels and cluster's centroid was calculate in order to find the ones which are the most adequate to reach as many cultural sites as possible. The method used was the haversine distance[3]. Before the application of this method the value of the coordinated were converted in radians, as requested from the method. The resulting distance was expressed in km and only the first 7 hotels was considered (Table 2). According the obtained data, the nearest hotel to cluster's centroids is "*Hotel Torre Guelfa*".

Hotel's name	Distance from the centroid (km)
Hotel Torre Guelfa (Palazzo Acciaiuoli)	0.050652
La Casa Del garbo Hotel	0.066505
Relais Hotel Uffizi	0.087016
Hotel Relais Uffizi	0.088538
Fh Hotel Calzaiuoli	0.096272
Olga's House Hotel Florence	0.114010
Hotel Bernini Palace	0.129837

Table 2. Haversine distance of Florence's Hotels from densest cluster's centroid.

5. Discussion

The simplicity and rapidity of the method KNN made it a good choice for this work. Once applied on the data frame containing Florence's museums and churches, the method divided the city in 5 clusters (Figure 2). The number 1 (red circles) is the most densest so the best suitable to find hotels for a cultural holiday. Then the haversine distance between hotels and this cluster's center was calculate, in order to find hotels which permit to easily reach the most part of these cultural sites. It was chosen the first seven hotels in order to give the client a good choice avoiding too far hotels (Table 4). The nearest hotel to cluster's centroids is "*Hotel Torre Guelfa*" which made it the best choice for an efficient cultural holyday in Florence (Figure 4). Moreover, another good choice can be the "*La Casa Del garbo Hotel*" because in case of longer holiday, it is not far from the cluster 4 (orange circles), the second densest cluster.

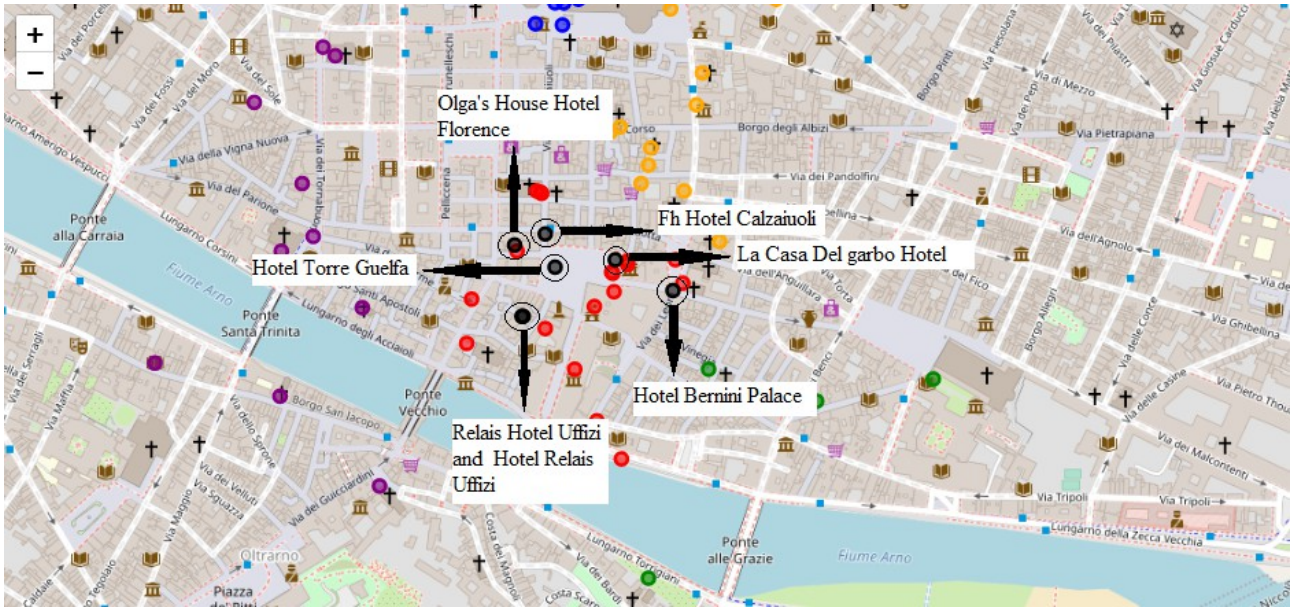

Figure 4. Data points of churches and museums (colored circles) and hotels in the cluster 1(black circles). Cluster 0 = blue, cluster 1 = red, cluster 2 = green, cluster 3 = purple, cluster 4 = orange, hotels = black.

6. Conclusion

In conclusion:

1- Data about churches museum and hotels were obtained from the website Foursquare[1].

2- Churches and museums were clustered using the method k- Nearest Neighbours, finding the cluster containing the highest amount of them.

3- Hotels easily permitting to reach the most part of these cultural sites was found calculating the haversine distance between hotels and the centroid of the densest cluster. The best choices resulted to be "*Hotel Torre Guelfa*", the one nearest to cluster's centroid, and "*La Casa Del garbo Hotel*" because in case of longer holiday, it is not far from the cluster 4 (orange circles), the second densest cluster.

7. References

1- https://foursquare.com/

2- https://levelup.gitconnected.com/clustering-gps-co-ordinates-forming-regions-4f50caa7e4a1

3- https://medium.com/@danalindquist/finding-the-distance-between-two-lists-of-geographic-coordinates-9ace7e43bb2f