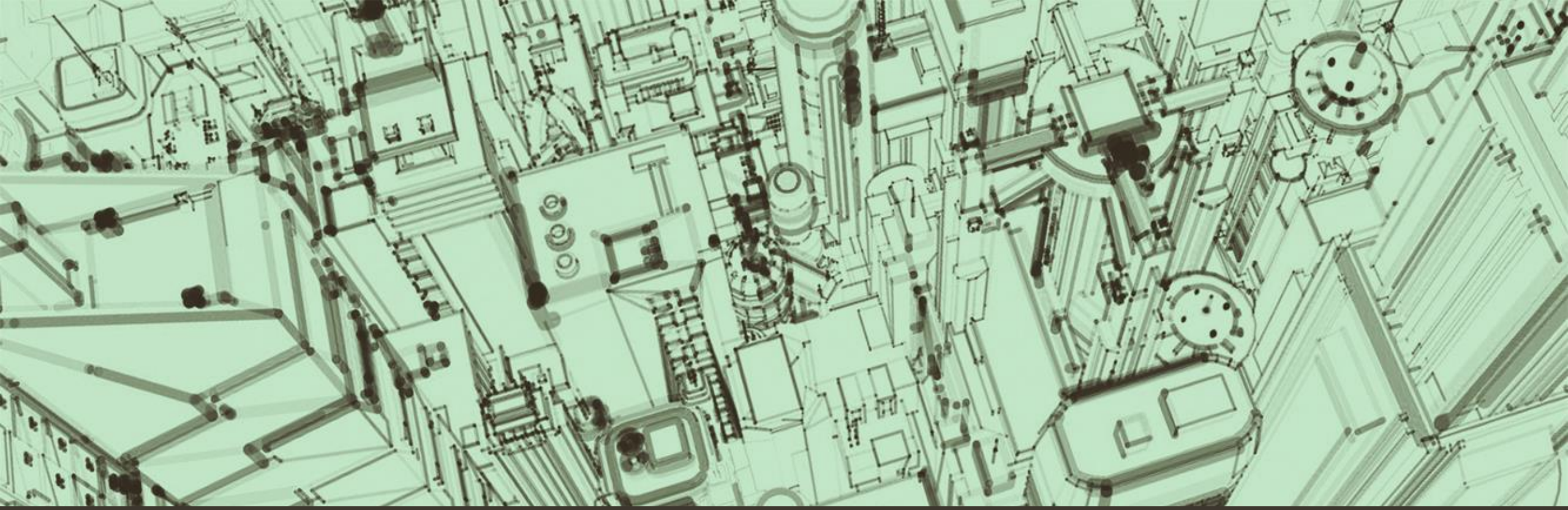


# Clustering cities by venues categories

Marzio Lanzoni

# PROBLEM AND TARGET

## Machine learning catches what conventional methods cannot

Big cities evolve very quickly, following economic perspectives, lifestyle trends and political scenarios.

Common thinking tends to assume similarities among cities based on geographical proximity or historical connections, but nowadays many factors affects cities evolution, and may turn common thinking and traditional assumptions outdate and very misleading.

The aim of the project is to build a machine learning approach able to catch similarity among big cities all around the world, based on actual and update characteristics like real-time data provided by Foursquare APIs.

Such an approach would allow a continuous update of how a city is changing, something that conventional approaches like surveys and inventories cannot accomplish: surveys need to be planned, executed and published, and they risk being outdate before becoming effective.

The deliverable will be a tool for people aiming to travel and looking for experiences and feelings they previously had visiting other cities

# DATA ACQUISITION AND CLEANING

Data source Machine learning catches what conventional methods cannot

Location and population data for cities all around the world were gathered from https://simplemaps.com/data/world-cities.

A subset of cities with medium-high population (**2.5M+ habitants**) are kept. China's main cities were dropped because too big. This returned a set of **36 cities**.

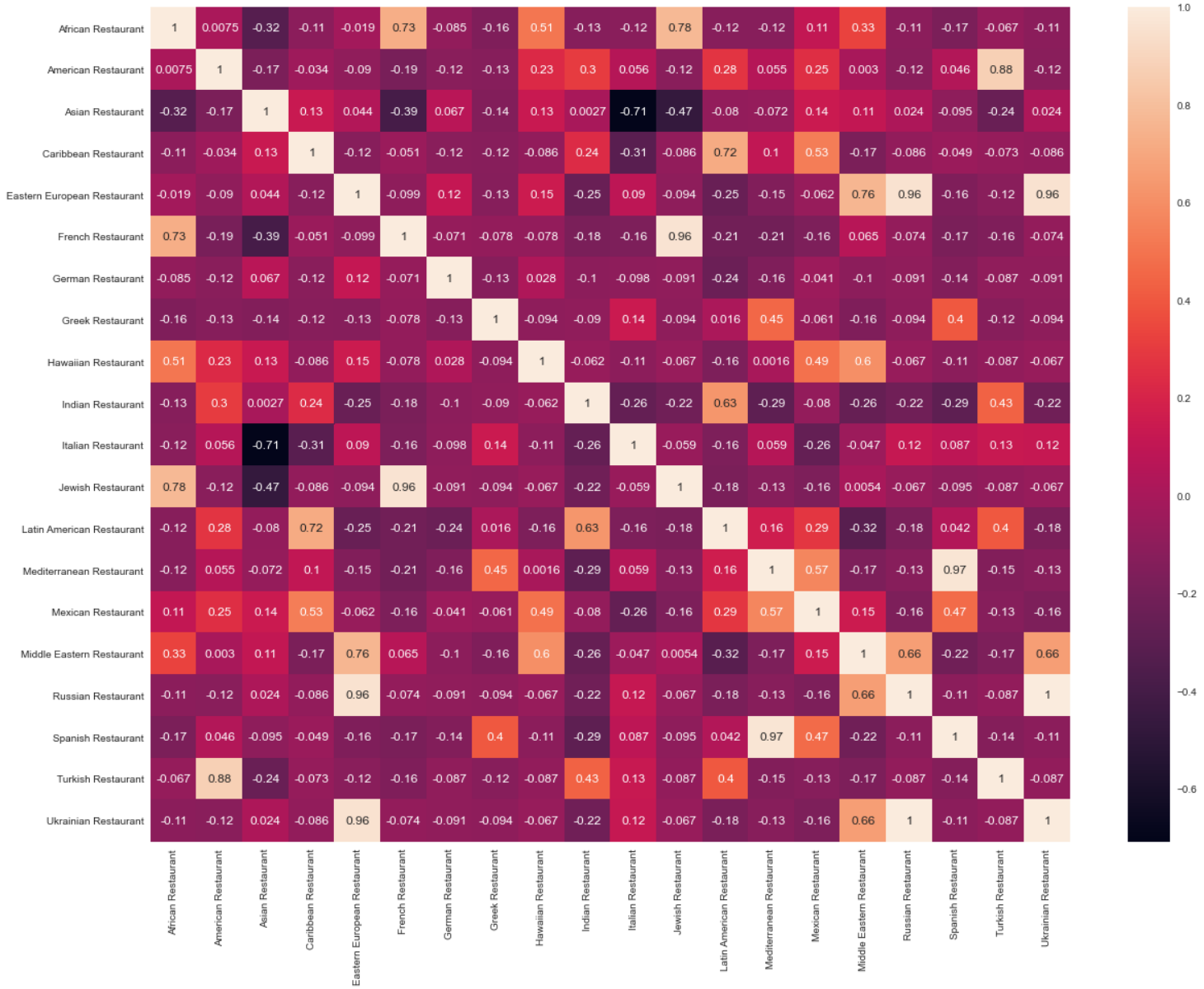
Foursquare APIs venues related to the main category "**Food**" were used, and **only venues located around the downtown** were kept.

Only cities well tagged on Foursquare were kept, namely cities with **at least 40 food-related venues in a 700 metres radius from the centre of the city**. This make the **final set made by 18 cities**.

**One-hot encoding** was performed to transform the original dataset (made by one row per venues) in a new dataset where each row represents a city and each column is the relative occurrence of a venues category.

# DATA ACQUISITION AND CLEANING

## Heatmap to search for correlations




Some features show correlations. This is not because of a logical dependence (each venue is per se), but rather because the Foursquare's hierarchical categorization is prone to ambiguities (categories too similar can be used one for another).

A different aggregative structure is suggested as **future development**.

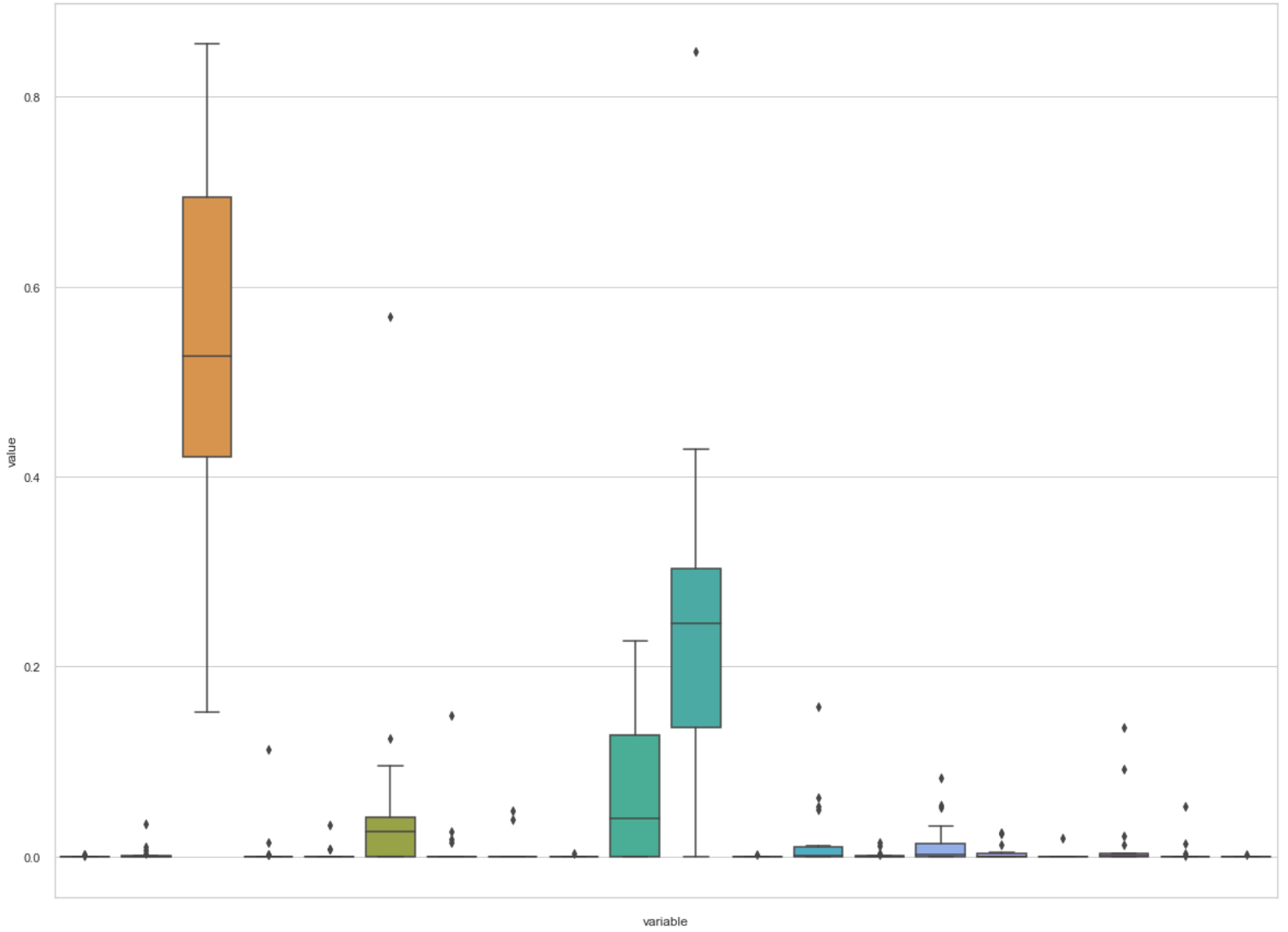
# DATA ACQUISITION AND CLEANING

## Variance analysis

**The features do not require any scaling**, since they all have the same meaning (relative occurrence) and range (from 0 to 1, by definition).

Nevertheless, a boxplot of the features shows that some of them are hardly significant because of a very low occurrence.

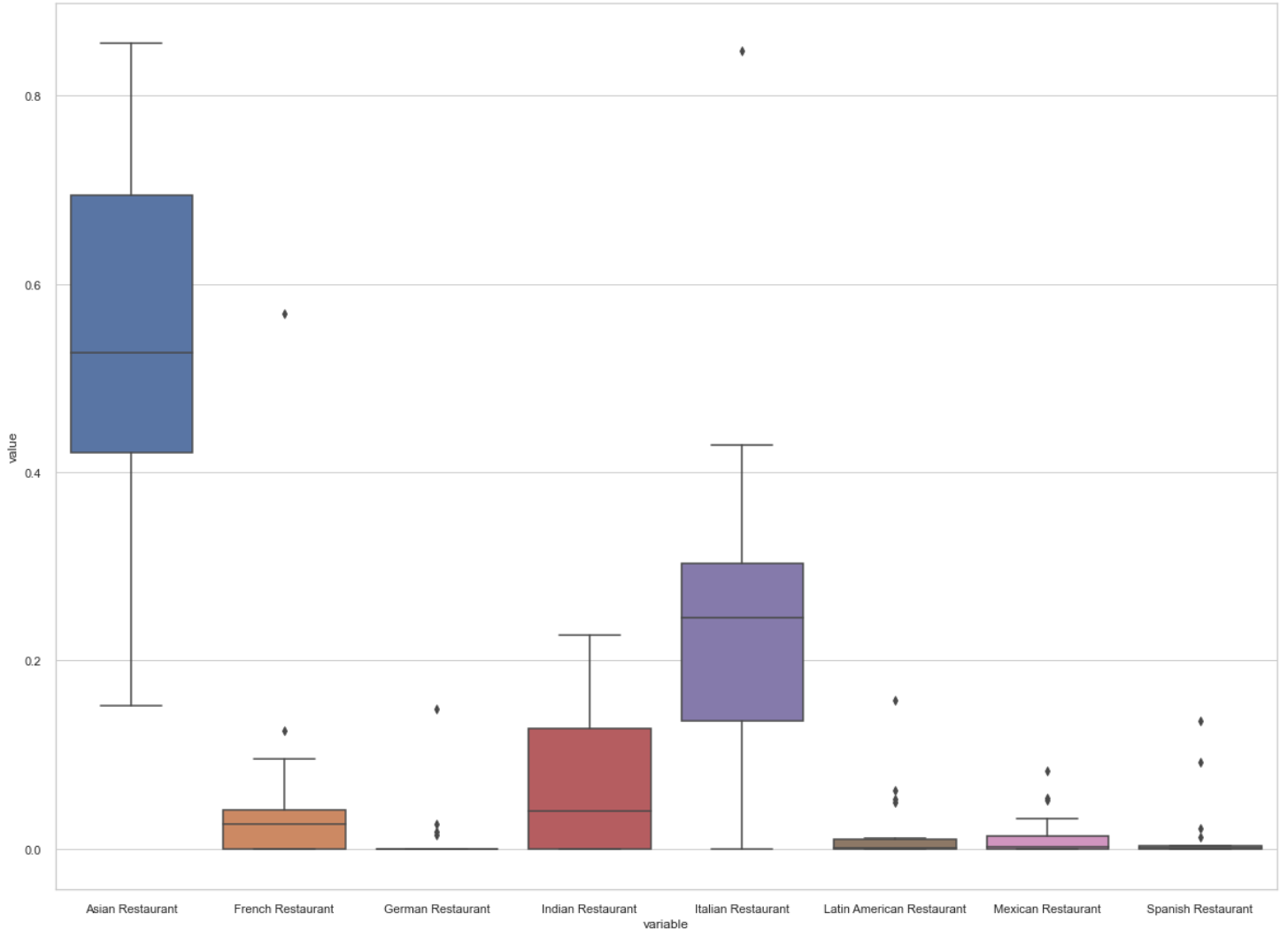
Using the complete set of features makes the modelling quite unstable (the elbow method using Silhouette metric does not converge).



# DATA ACQUISITION AND CLEANING

## Variance analysis and features selection

Keeping only **features with a mean occurrence grater than 0.01** stabilize the model.



# DATA ACQUISITION AND CLEANING

## Features selection

The reduced dataset, consisting of features with mean occurrence greater that 0.01, is made by **8 features.**

Their values have the same magnitude.
Being the mean occurrence not negligeable implies that all the remaining features correspond to well characterized type of cuisine (low noise).

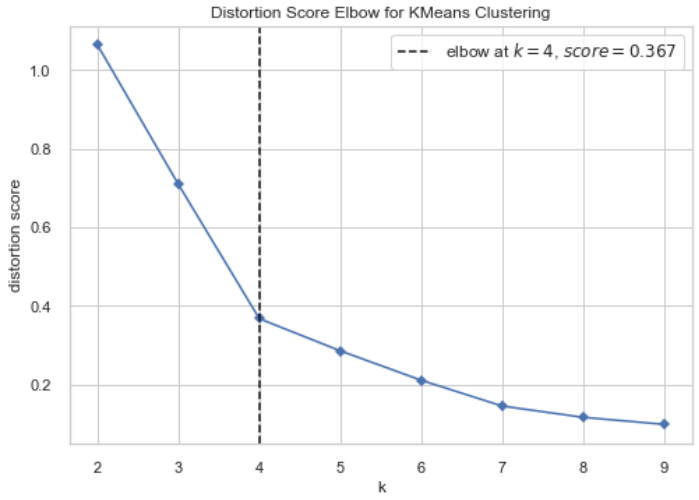
| | Asian | French | German | Indian | Italian | Latin American | Mexican | Spanish |
|---|---|---|---|---|---|---|---|---|
| count | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 |
| mean | 0.5399 | 0.0599 | 0.0115 | 0.0642 | 0.2463 | 0.0191 | 0.0147 | 0.0153 |
| std | 0.2022 | 0.1314 | 0.0351 | 0.0727 | 0.1902 | 0.0403 | 0.0242 | 0.0371 |
| min | 0.1524 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 25% | 0.4207 | 0.0000 | 0.0000 | 0.0000 | 0.1362 | 0.0000 | 0.0000 | 0.0000 |
| 50% | 0.5269 | 0.0260 | 0.0000 | 0.0403 | 0.2453 | 0.0008 | 0.0023 | 0.0012 |
| 75% | 0.6946 | 0.0417 | 0.0000 | 0.1275 | 0.3034 | 0.0096 | 0.0135 | 0.0033 |
| max | 0.8553 | 0.5680 | 0.1486 | 0.2266 | 0.8476 | 0.1583 | 0.0827 | 0.1362 |

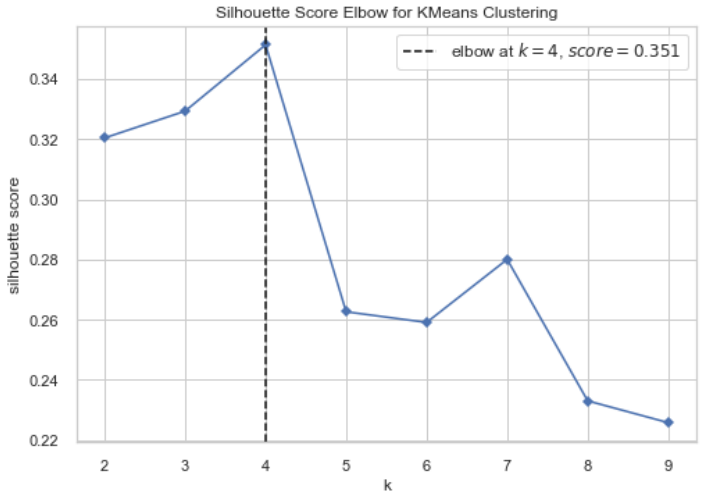
# MODELLING

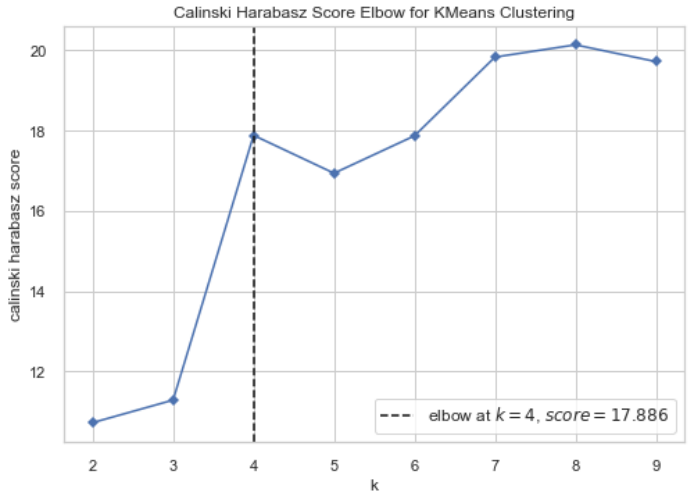
## Tuning the number of clusters

Both **K-means** and **Agglomerative** algorithm were used, in order to compare their clustering performances.

The only parameter to be tuned is the number of clusters, and the search for this parameter was done, for both algorithms, using the elbow method and three different metrics: **Distortion, Silhouette, and Calinski Harabasz**.




All the metrics returned the same value for both algorithm, suggesting to cluster the samples in **4 clusters**.

# CONCLUSIONS

How cities are clustered

| Red Cluster |
| --- |
| Barcelona |
| Berlin |
| Birmingham |
| London |
| Madrid |
| Manchester |
| Moscow |
| Washington |

| Blue Cluster |
| --- |
| Hong Kong |
| Montréal |
| New York |
| Seattle |
| Tokyo |
| Ōsaka |

| Cyan Cluster |
| --- |
| Paris |

| Yellow Cluster |
| --- |
| Rome |


K-means and Agglomerative algorithms returns exactly the same clustering with the same scores (K = 4):

- Silhouette Score: 0.3514
- Calinski Harabasz Score: 17.8863
- Davies Bouldin Score: 0.5610

# CONCLUSIONS

Possible intuitions

- All the European cities have been clustered altogether, exception made for Paris and Rome.
- Paris and Rome, despite European cities, have been clustered separately and constitute one cluster each. French and Italian cuisines are so renowned and valuable brands that the capitals of the respective countries bet on their national cuisine more than other capitals.
- USA and Japan big cities have been clustered together: they tend to have more a cosmopolitan characterization than a national or local one.

# FUTURE DIRECTIONS

Despite we can find a reasonable interpretation of the clustering results, still the scores obtained suggest that clusters are overlapping, with samples quite close to the decision boundary of the neighbouring clusters.
Here some future directions of the study that can possibly drive to more dense and well separated clusters.

## RADIUS

We set a 700 metres radius within which to search for venues, based on the idea that representative venues are closer to the downtown. Nevertheless, the definition of “close to the downtown” likely depends on the size and topology of the city, which suggest varying the radius based on the city physical properties.

## VENUES CATEGORIES AGGREGATION

Foursquare’s categories are hierarchically organized in such a way that categories located at the same level have very different meaningfulness. Trying a custom aggregation could result in a better representation of the culinary profile of the cities.
This approach is also suggested by the data exploration which shows that there are categories highly correlated because too similar and some also suffer of very low occurrence.
Remediation to both issues could be trying a different aggregation, where categories highly correlated are merged, and categories with very low occurrence are merged as well to the closest category.

# FUTURE DIRECTIONS

## MORE MAIN CATEGORIES

For sake of simplicity, the presented analysis is based only on “Food” main category, while other main categories (like Arts & Entertainment, Nightlife Spot, Outdoors & Recreation) could help profiling the cities.

## NUMBER OF VENUES

The dataset used is made by cities with very different number of venues, likely due to different population and different touristic appeal.
Also, for some city the number of venues reached the 100-limit imposed by the free Foursquare license, which may result in a bias because the overall number of venues in the downtown radius could be much larger and the returned venues are not necessarily the most representative.
Unlocking Foursquare license and trying to use cities with more similar number of venues is a possible direction of investigation.