Segmentation of Croatian cities

Andrea Pirsa
8-20-2019

Introduction

Croatia is a Central European and Mediterranean country with a long maritime border with Italy in the Adriatic Sea. Croatia has an amazing 5,835km of coastline, 4,057km of which belongs to islands, cliffs and reefs. The climate is Mediterranean along the Adriatic coast, meaning warm dry summers and mild winters, with 2,600 hours of sunlight on average yearly – it is one of the sunniest coastlines in Europe. The main industry in Croatia is tourism. Croatia has a big challange to make strong offer for tourists from different cultures. One of the challenges is to show tourists what the country has and what are some of the specialties of each country region. Another challenge have Croatian caterers when they need to create offers for tourists and open the right type of restaurant in right town. This project will try to solve both problems: show segmentation of Croatian cities based on the collected venues data and predict which city has some specific type of restaurant based on some other venues data. This solutions could help tourists to find out how similar some cities are and show investmens where to open some type of restaurant.

Data

To collect all the cities of Croatia and details like to which county city belongs and how many people live there, we have loaded data from Wikipedia: https://en.wikipedia.org/wiki/List_of_cities_and_towns_in_Croatia . Figure 1 Wikipedia data shows dataframe after some data preparation part where the name of cities was cleaned from special characters.

	City	County	Population
1	Zagreb	Zagreb	688163
2	Split	Split-Dalmatia	167121
3	Rijeka	Primorje-Gorski Kotar	128384
4	Osijek	Osijek-Baranja	84104
5	Zadar	Zadar County	71471
6	Velika Gorica	Zagreb County	31553
7	Slavonski Brod	Brod-Posavina	53531
8	Pula/Pola	Istria County	57460
9	Karlovac	Karlovac County	46833
10	Sisak	Sisak-Moslavina	33322
11	Varaždin	Varaždin County	38839
12	Šibenik	Šibenik-Knin	34302

Figure 1 Wikipedia data

For each city in dataframe was found right geolocation data with longitude and latitude values. For thos purpose was used *geopy* library. Figure 2 Map Croatia shows all collected cities in Croatia and the bubble size shows the population value.

Figure 2 Map Croatia

From https://api.foursquare.com was collected data about the venues which are in the radius of 5000 meters of each city in Croatia. Values collected for venues are:

- 'Venue name',
- 'Venue Id',
- 'Venue Distance',
- 'Venue Latitude',
- 'Venue Longitude',
- 'Venue Category',
- 'Venue Catgory Id'

API has limit of 100 venues per call, so it was defined process of collecting data per city and per venue category. Figure 3 Venues shows how collected data for venues looks like.

	City	Latitude	Longitude	Venue	Venue Id	Venue Distance	Venue Latitude	Venue Longitude	Venue Category	Venue Catgory Id
200	Rijeka	45.326936	14.440984	King's Caffe pub	5320499d498e106a8736c44f	194	45.328686	14.440913	Bar	4bf58dd8d48988d116941735
201	Rijeka	45.326936	14.440984	Cacao	56e6f24e498e82b22c027edc	135	45.326876	14.439252	Dessert Shop	4bf58dd8d48988d1d0941735
202	Rijeka	45.326936	14.440984	Conca D'oro	54a2f0e5498ef7f11af2f84d	117	45.327811	14.440149	Bistro	52e81612bcbc57f1066b79f1
203	Rijeka	45.326936	14.440984	CukariKafe	4d18f42c25cda14329f782d6	183	45.327594	14.443135	Lounge	4bf58dd8d48988d121941735
204	Rijeka	45.326936	14.440984	King's Caffe Food pub	55cb3065498e0fce8864c96a	381	45.324290	14.444085	Pub	4bf58dd8d48988d11b941735

Figure 3 Venues

Final dataframe with venues data had 14130 venues.

Methodology

Exploratory data analysis

Venues date has 14130 values and the first step in data analysis was to explore how distribution of number of venues per city looks like. Figure 4 Number of venues shows that in our dataset there are few cities with a lot of venues and then there is strong decrease in number of venues. Zagreb is the capital of Croatia and the correlation between the size of city and number of venues is obvious.

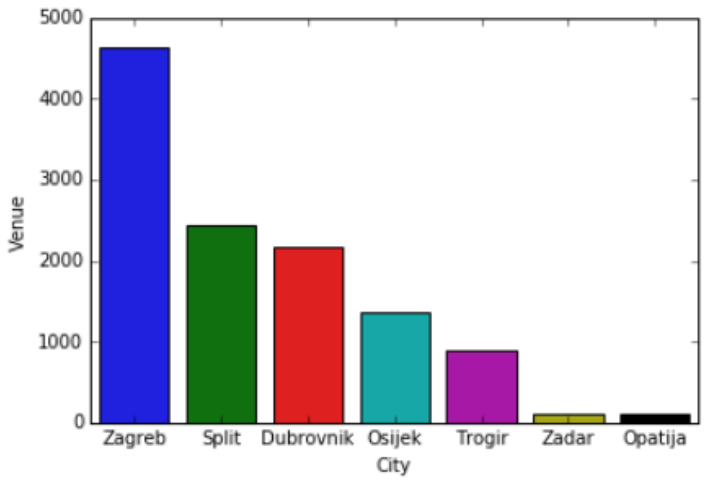

Figure 4 Number of venues

There are some cities with less than 10 venues and decision was to remove that cities from dataset. After removal, dataset had 13599 venues for 65 cities.

Next step was exploring correlation between number of venues and population of city. Figure 5 Population and number of venues shows that number of venues in city increases with the number of people living in that city.

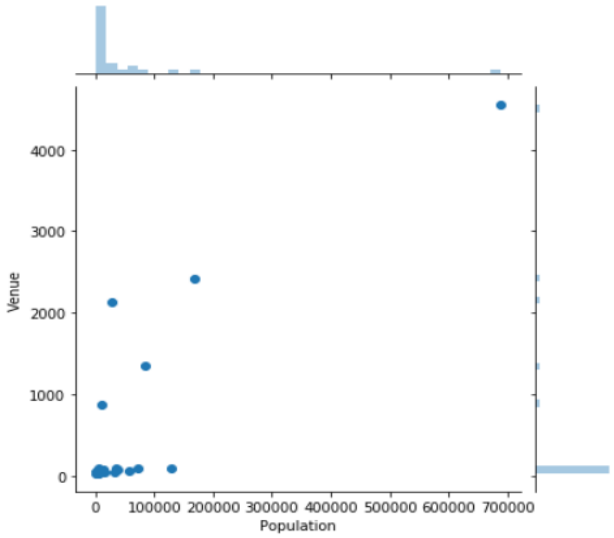

Figure 5 Population and number of venues

Dataset has pretty much different categories of venues and the next analysis was to see how many venues are in each category. Figure 6 Top categories shows number of venues in categories with the largest amount of venues. Figure 7 Bottom categories shows categories with the smallest amount of venues.

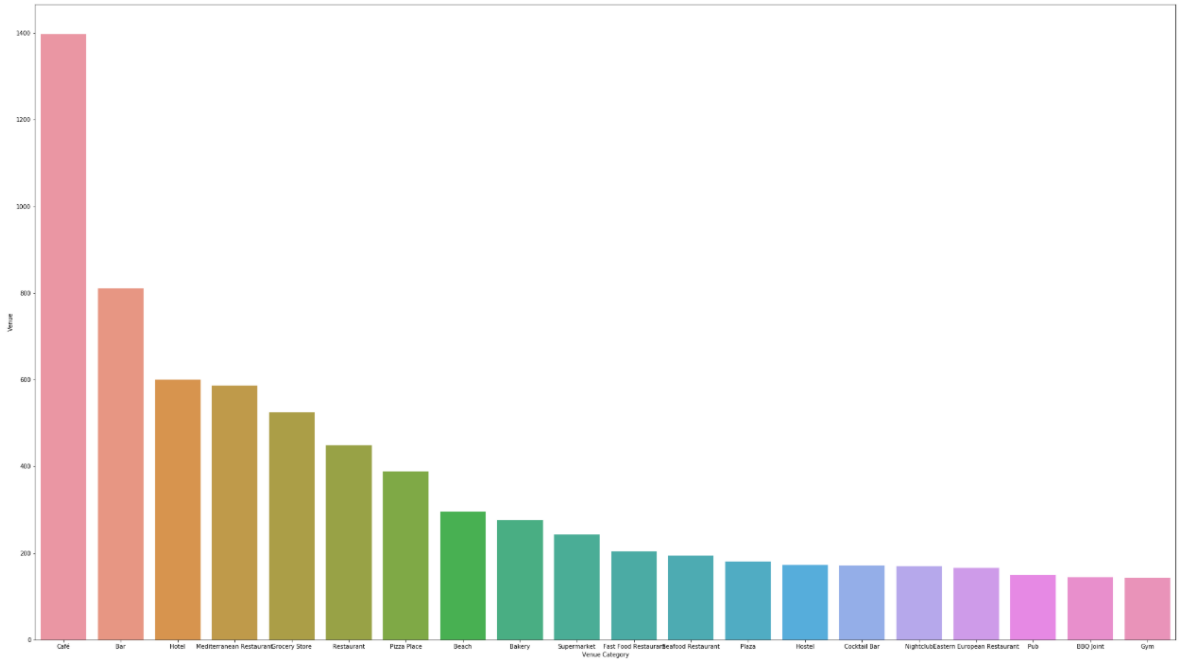

Figure 6 Top categories

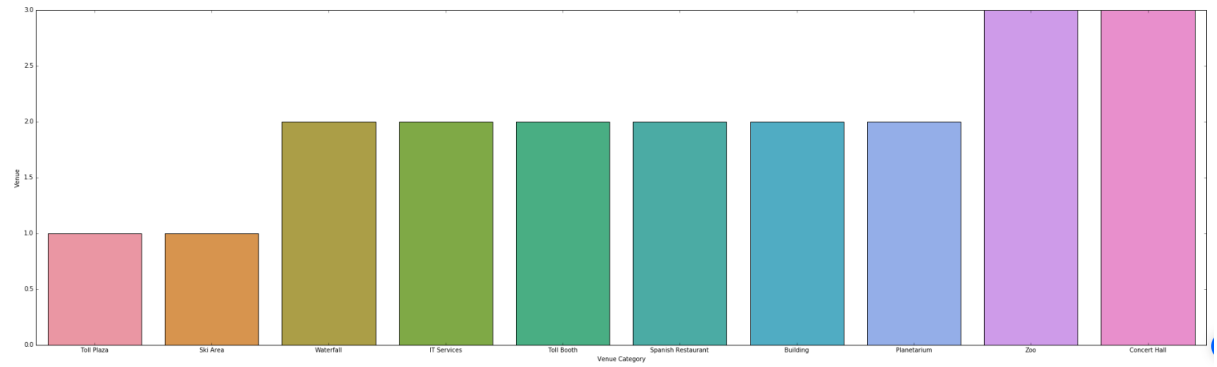

Figure 7 Bottom categories

The largest categories have few thousands venues and the smallest only few examples. This analysis showed that there could be a need for joining some of venues categories in one category.

To show number of venues per some part of category name, it was created wordcloud. Figure 8 Wordcloud categories names shows how many times each word was part of some example of venue in the city. Large words like 'Store', 'Bar', 'Park', 'Restaurant' are very often and word slike 'Korean', 'Donut', 'Indie' are very small. Based on the analysis the conclusion was that small words are some types of large words, for example 'Donut' is type of store. To reduce number of categories, some large words like 'Bar', 'Store', 'Park' were new categories in which were renamed all categories that have large words as part of name and less than 50 venues in total. „Restaurant“ wasn't part of this feature elimination because for the prediction of restaurant types there should be all restaurants part of dataset.

Figure 8 Wordcloud categories names

Figure 9 Wordcloud after categories reduction shows that some of small words are now even smaller our disappeared and big words become bigger. In this process was removed 14 categories and they bacame part of some other category.

Figure 9 Wordcloud after categories reduction

After data exploration and reducing of some features, categorical values in the dataset are one hot encoded and all features are aggregated on the level of city. Population was removed from dataset bacause of big correlation with the number of venues and counties.

Machine learning modeling

Segmentation of cities

For the problem of segmentation of the cities the machine learning model which was used is Kmeans implementation from sklearn. This was problem of clustering data in chosen number of clusters and for that problem KMeans was proven choice. Clustering was done with few number of clusters and results will be shown on the map. For the purpose of segmentation of the cities clustering was done on 2 datasets: first was dataset with the whole prepared data and the second was dataset with only venues which are considered as attractive for tourists (bars, restaurants, pubs, hotels, motels…).

Restaurant predictions

The idea of restaurant predictions was to develop model which could predict if some city has Mediterranean Restaurant, based on the dataset with only touristic attractive venues and county data. This is a classification problem for which were tested 2 models: Random Forest and SVM. Results and comparison of metrics are described in results section.

Results

Segmentation of cities

Figure 10 Kmeans 4 clusters shows results of cities segmentation with Kmeans model trained on the first dataset with all prepared variables. After joining of cluster labels to cities and their coordinates, data was shown on the map. Results of clusters are overlapping with some geographical and traditional regions in Croatia. Light blue bubbles are cluster which overlapps with North Croatia region, purple bubles are cluster which overlapps with Lika region, yellow bubbles are cluster which overlapps with Istria and Dalmatia and red bubbles are cluster which is a south Dalmatia with a lot of islands and huge amount of tourists throught the whole year. Results of this model are pretty good and they can show to the tourists what are some specific regions in Croatia. Based on the venues tourists can visit only few cities in that region and they can meet the lifestyle of the whole region.

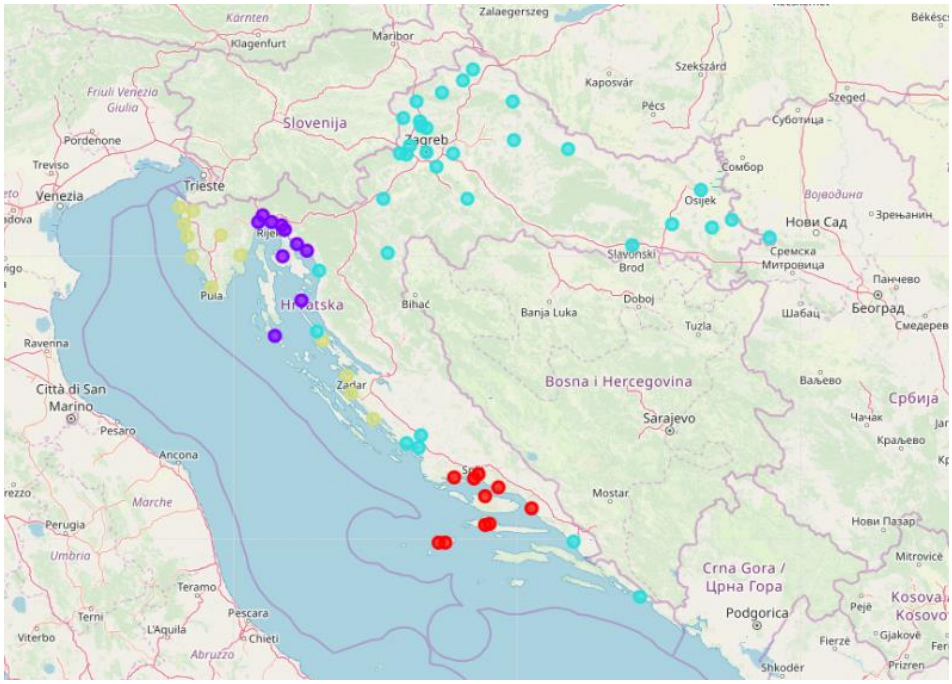

Figure 10 Kmeans 4 clusters

Figure 11 Kmeans 5 clusters shows result of Kmeans model with 5 clusters trained on the first dataset with all prepared variables. Results of this segmentation are isolating some specific parts of Croatia, like light green is one small region called Hrvatsko Zagorje which has a lot of green wineyards and the culture of spending a lot of time in pubs. Segmentation done by this model is different from the previous because now there is not very big overlapping with some geographical regions, but there are some new small and very specific clusters.

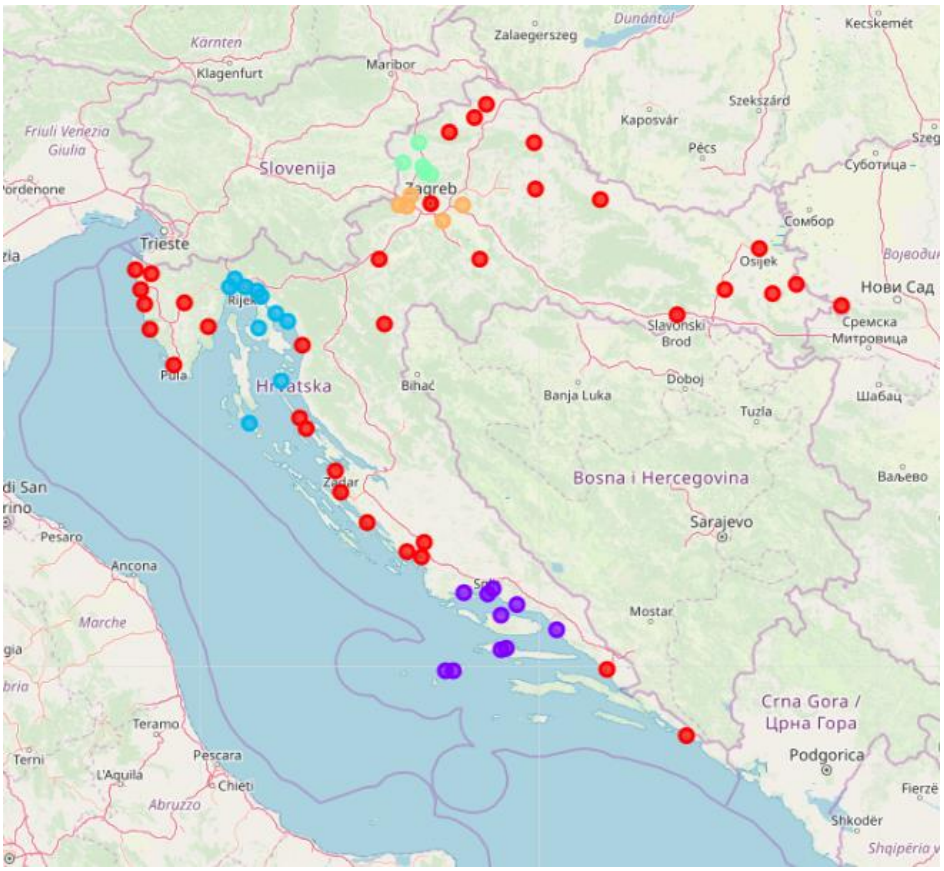

Figure 11 Kmeans 5 clusters

Kmeans on touristic venues

Figure 12 Kmeans touristic data shows results of K Means model with 4 clusters trained only on dataset with some venues which are considered as touristic attractive. Red cluster is the part of Croatia with beaches, a lot of coffee shops and some specific restaurants. Yellow cluster are cities with a lot of bistros, shops, hotels and different types of restaurants. Light blue cluster are one small part of Croatia with pubs and bistros. In purple cluster are only 2 cities Osijek and Đakovo, that are cities with some specific culture and traditional Croatian restaurants. Clustering based on the „touristic“ data is showing were to try some specific food and where tourists can go if they want to spend time in some specific way.

Figure 12 Kmeans touristic data

Figure 13 Beaches by clusters shows where are beaches located by clusters. Based on the boxplot, cluster 0 is region of Croatia located on the Adriatic sea with many beaches. Other clusters are continental, but in cluster 3 are some cities with beaches which are on the lakes or rivers.

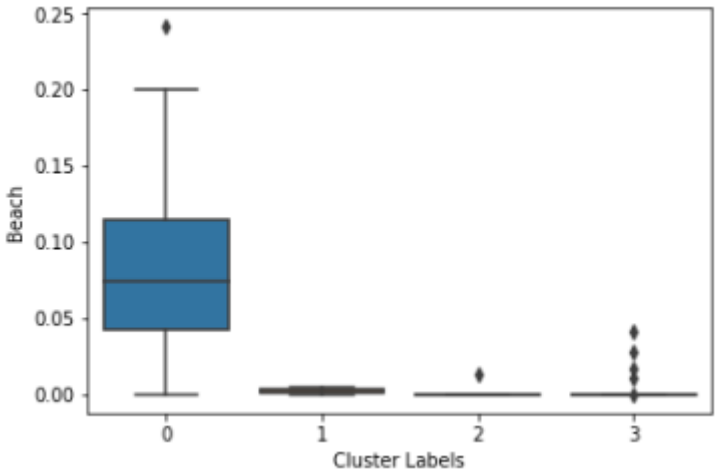

Figure 13 Beaches by clusters

Figure 14 Bars by clusters shows how are bars distributed by bars. In all clusters are located some bars, cluster 2 is above all other clusters by number of bars. Cluster 2 covers some regions in Croatia with a culture of wine yards and drinking, so bars are common places to go in this parts of Croatia.

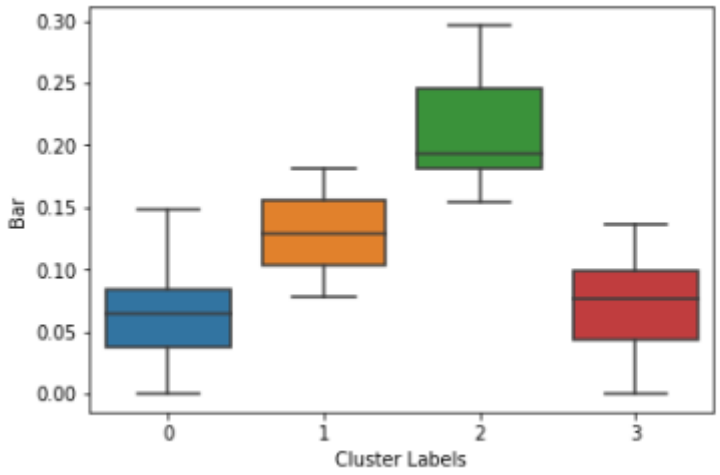

Figure 14 Bars by clusters

Figure 15 Pubs by clusters shows that cluster 1 has a lot of pubs. This is small region with only 2 cities in east part of Croatia. Clusters 1,2 and 3 are in parts of Croatia where are some beer industries and pubs are popular places. In region on the adriatic sea people drinks a lot of coffee and their bars and pubs are places which are manly open through the summer season.

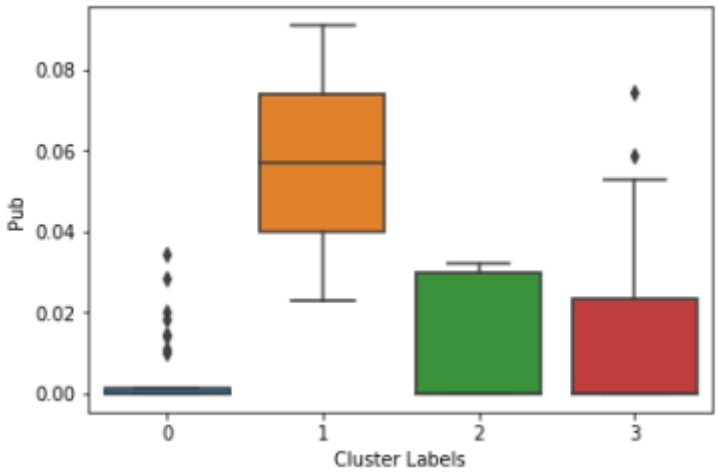

Figure 15 Pubs by clusters

Figure 16 Mediterranean restaurants by clusters shows that the most of Mediterranean restaurants are in cluster 0 and there are some in cluster 3. This confirms that cluster 0 is region of Croatia with a lot of sea, beaches and sea food. Cluster 3 has some large cities with many different restaurants. For mediterranean restaurant lovers great place could be any big city in Croatia or any city on the seaside.

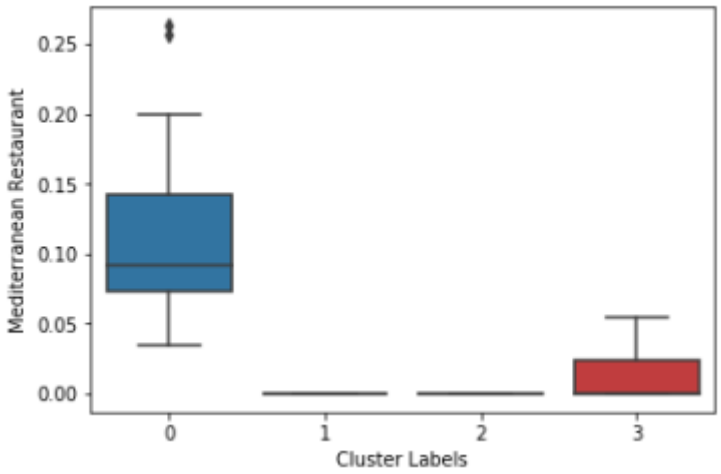

Figure 16 Mediterranean restaurants by clusters

Restaurant predictions

Based on the prepared dataset with variables which are „tourist attractive“ models have try to predict if there is any Mediterranean restaurant in cities of Croatia. Developed models were Random Forest Classifier with 250 estimators and default SVM model.

Figure 17 Results shows comparison of F1 score and Jaccard similarity score for Random Fores and SVM models trained and tested on the same datasets. Random Forest has higher F1 score and Jaccard score. 0.9375 F1 score and 0.9207 Jaccard score are showing that Random Forest is pretty good choice for predicting if some city has this specific type of restaurant.

	F1 score	Jaccard similarity score
RandomForest	0.9375	0.9207
SVM	0.7907	0.6538

Figure 17 Results

Figure 18 Random Forest confusion matrix and Figure 19 SVM confusion matrix are showing comparison of 2 models. Random Forest is good in predicting both classes and SVM for all cases chose the same class. Results are shown on test set which is 40% of total data, which means that Random Forest has good predictions even it has seen only 60% of very small dataset.

	0	1
0	9	0
1	2	15

Figure 18 Random Forest confusion matrix

	0	1
0	0	9
1	0	17

Figure 19 SVM confusion matrix

Discussion

Venues data has very big potential for many different purposes and analysis. Segmentation of cities in Croatia has shown that with correct data and a lot of preparation we can get some good insights from data. Based on Kmeans clustering cities in Croatia are grouped in clusters which have overlapped with some geographical and cultural regions in Croatia. Number of clusters influenced on insights that we get from data. Segmentation depends on the input data and based on the purpose of segments input data should be selected. Results have shown that Croatia has few interesting regions and for tourist this model could be some kind of trip advisor. Based on the some type of venues they are interested in, this model could gave them group of cities that they can visit. This model could be improved with some extra data about venues like price or comments. This was considered as second phase of this project, but the problem is API limit for details about venues. With venues details there is spece for making not only segmentation of cities , but also recommendation of some special parts of each segment.

Random Forest classifier has shown that it can make good predictions on very small dataset. This model can predict in which city is some type of restaurants based on some other venues in that city. The purpose of this model could be to help investors to find the right place for opening some type of restaurant, bar, shop… On the other hand, this model could be advisor for tourist if they want to find cities with some type of venue based on their preferences.

Conclusion

In this project was analysed data about cities and venues in Croatia. It has been done segmentation of Croatian cities based on different input data about venues in all the cities. Analysis of results has shown that clustering models can find some specific regions of Croatia. As a second part of this project it has been done prediction if some type of restaurant is located in the city or not, based on data about some other restaurants, bars, hotels and other touristic venues. This classification model has very good results and there is a space to work on similar recommendation models. In the future, this project could be extended with more data about venues, more venues details and development of some customized model which could help tourists with recommendations based on some real-time inputs.