THE XINOMAVRO PROJECT

WINE BAR ANALYSIS IN GREECE

ILIAS KAPETANAKIS 2021

INTRODUCTION

Xinomavro:" (Greek: Ξινόμαυρο [ksiˈnomavro], *lit.* 'sour black') is the principal red wine grape of the uplands of Naousa in the regional unit of Imathia, and around Amyntaio, in Macedonia, Greece.".

This project is inspired from the love for the good Wine and the places you can find it, drink it and have amazing time with friends, the Wine Bars.

The main purpose of this analysis is to help people who are interested on opening a new wine bar in Greece to get an idea which is the current situation in the biggest Greek Cities and to make easier for them to understand where and why it's better to invest their money and open their Wine Bar.

We will involve in our analysis data like i) the population of the cities to have an indication of the size of each one of them, ii) the number of the Wine Bars, iii) the number of the Bistros as they are both wine related places and iv) metrics that characterize each Wine Bar like Foursquare Rating, Price Range and Number of Likes.

DATA

The following data will be used in the analysis:

1. GREEK CITIES POPULATION AND COORDINATES

DATA EXTRACTION

Data Source: *GREECE City & Town Population Geography Population Map cities coordinates location - Tageo.com (http://www.tageo.com/index-e-gr-cities-GR.htm)*

Extraction Method: Scrap HTML to get the table containing the Greek Cities, population and coordinates from the relevant table of the Tageo.com page

Number of rows: 60

DATA DESCRIPTION

The data Source contains the 60 biggest Greek cities in terms of population , the population of each one of them and the Latitude/Longitude.

DATA QUALITY, CLEANSING AND TRANSFORMATION

- Remove unnecessary columns like Rank
- Fix Header of the Dataset (Header is appearing as 1st row)
- Convert numeric fields - Population
- Handle Duplicate City Names: Iraklion Crete duplicate with Iraklion Attica

DATA SAMPLE

	City	Population	Latitude	Longitude
1	Athinai	762100	37.980	23.730
2	Thessaloniki	372100	40.640	22.940
3	Piraieus	179600	37.960	23.640
4	Patrai	164000	38.240	21.730
5	Peristerion	141000	38.020	23.700

2. GREEK CITIES VENUES (WINE BARS + BISTROT)

DATA EXTRACTION

Data Source: FOURSQUARE API : search Endpoint (*GET https://api.foursquare.com/v2/venues/search*)

Extraction Method: Call the API for each city found in dataset [1. GREEK CITIES POPULATION AND COORDINATES] by passing the latitude and longitude of each city and the specific category id of the "Wine bars" (4bf58dd8d48988d123941735) and "Bistro" (52e81612bcbc57f1066b79f1) as found in Foursquare documentation (https://developer.foursquare.com/docs/build-with-foursquare/categories/)

Number of rows for Wine Bars: 462

Number of rows for Bistro: 240

DATA DESCRIPTION

The data set contains all the venues belonging to the WineBar category for each city in a radius of 1500m from the City Coordinates. The information retrieved includes the Venue ID, Venue Name, Venue Latitude, Venue Longitude, Venue Distance from the City Coordinates and Venue Category. In addition by using exactly the same method the venues of the category 'Bistro' were retrieved.

Having the number of Wine Bars per City will be able to create a very important KPI by combining the Number of Winebars and the Population of each city: the Number of Winebars per Person. The same KPI was created for Bistrot showing the Number of Bistro per Person.

DATA QUALITY, CLEANSING AND TRANSFORMATION

- The major problem in this data source is that the same venue can be returned for more than 1 cities. This is possible to metropolitan areas like Athens and Thessaloniki because many cities consist the metropolitan area. The result is that there are venues having distance less than 1500m from multiple city centers. In cases like this the minimum distance will be used to keep each venue only one time and assign it to the city that is closer to the venue.
- Venues belonging both to Wine Bars and Bistro categories were kept only as Wine Bars, as this is our main interest, and were removed from the Bistro data set.
- It should be noted that not all cities have wine bars.

DATA SAMPLE

	City	Population	City_Latitude	City_Longitude	Venue_Id	Venue_Name	Venue_Latitude	Venue_Longitude	Venue_Distance	Venue_Category
0	Athinai	762100	37.980	23.730	50c37cd7e4b04a2d9cd2a324	Harvest	37.979581	23.728421	146	Café
1	Athinai	762100	37.980	23.730	53880882498e450b0a1fbd04	Delight	37.979810	23.732515	221	Café
2	Athinai	762100	37.980	23.730	5b0af271f193c0002c3c53d2	Kalimeres	37.978510	23.723880	562	Wine Bar
3	Athinai	762100	37.980	23.730	56a2a79b498ef61d4b2b036a	Wine O'Clock	37.967770	23.729855	1361	Wine Bar
4	Athinai	762100	37.980	23.730	56733fe7498e0010a67a750b	Acropaul's	37.966752	23.728304	1482	Wine Bar

3. WINE BAR DETAILS

DATA EXTRACTION

Data Source: FOURSQUARE API : details Endpoint (*GET https://api.foursquare.com/v2/venues/VENUE_ID*)

Extraction Method: Call the API for each venue found after processing of venues returned from data set [2. GREEK CITIES VENUS (WINE BARS + BISTRO)]

Number of rows: 388

DATA DESCRIPTION

The data source contains details about each Winebar venue. We will use it to get important data for each venue like:

- Rating (score from 1 to 10 with 1 decimal point)
- Price Range (from 1 (least pricey) - 4 (most pricey))
- Number of Likes

We will use the Average Rating, Price Range and Number of Likes for the Winebars of each city, to the analysis and the clustering of the cities.

DATA QUALITY, CLEANSING AND TRANSFORMATION

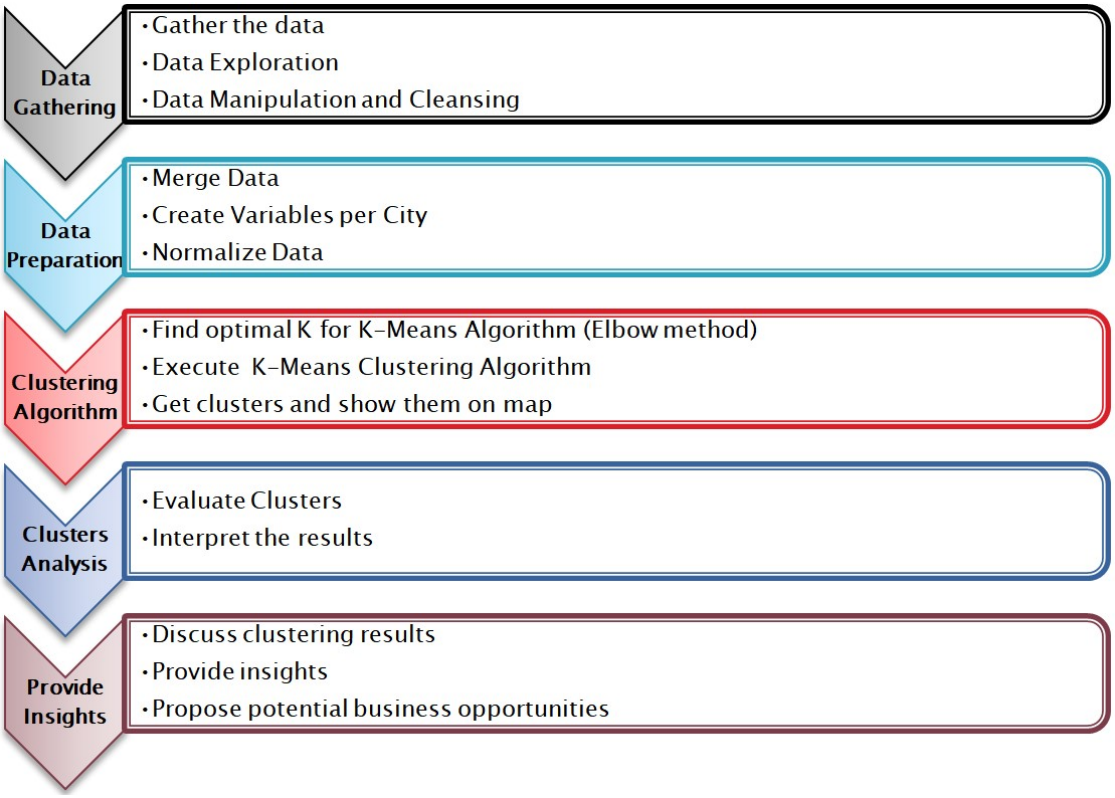
- In many venues price and rating info were missing from the data
- It was decided to replace the missing values as follows:
    - Missing Price = Average City Price Value rounded in the closer integer.
    - Missing Rating = Average City Rating Value rounded in 1 decimal
    - Missing Likes = The Median of total Likes

DATA SAMPLE

	Venue_ID	Venue_Name	Price	Likes	Rating
0	4adcdadef964a520f15721e3	Peacock Roof Garden Restaurant	2	4	6.7
1	4adcdadff964a520215821e3	CAFÉ & BISTROT VIENNA	1	15	7.5
2	4b5b2f7ff964a52087e928e3	Gala 1985	2	59	7.7
3	4b684c8bf964a52057702be3	Chocolat Royal	2	355	6.9
4	4b8589b7f964a5201f6431e3	Scala Vinoteca	3	229	9.2
5	4ba296a8f964a520f00638e3	Franco's	1	30	6.7
6	4bb25ff3eb3e9521a8f0c90a	Thema Coffees and Drinks	2	281	7.9

METHODOLOGY

We will gather the data described in the previous section, we will explore them, manipulate them and finally prepare them to be included in a Cluster Analysis. Our goal is to execute a K-Means Clustering Algorithm in order to get clusters that will help us better understand and segment the greek cities based on certain wine bar related characteristics. We will follow a bottom up approach by gathering raw wine bar data from the Foursquare API and gradually convert them to clusters containing group of cities with multiple wine bars each. After analyzing the clusters we will follow the opposite way and we will deep down the cities of each cluster to find details that will help us find results and insights.

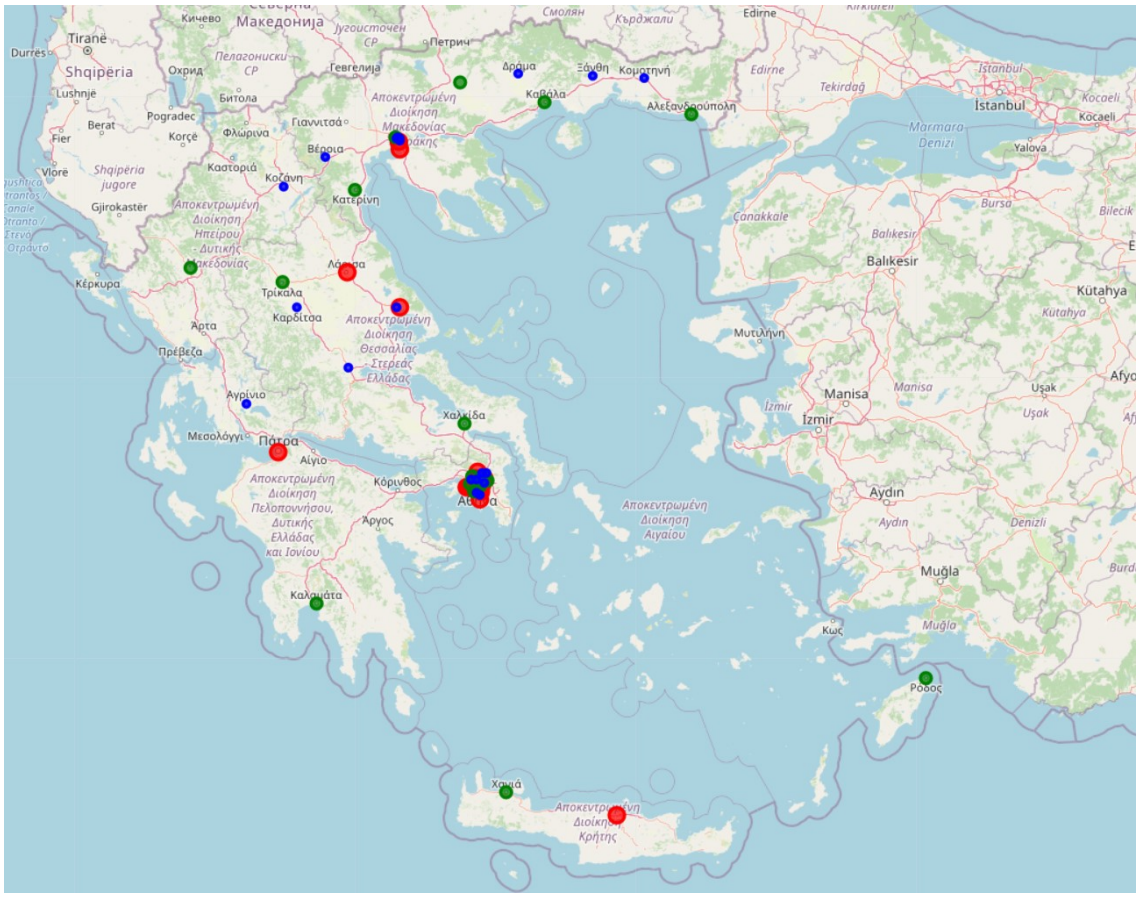


DATA GATHERING AND EXPLORATION

CITIES POPULATION AND LOCATION

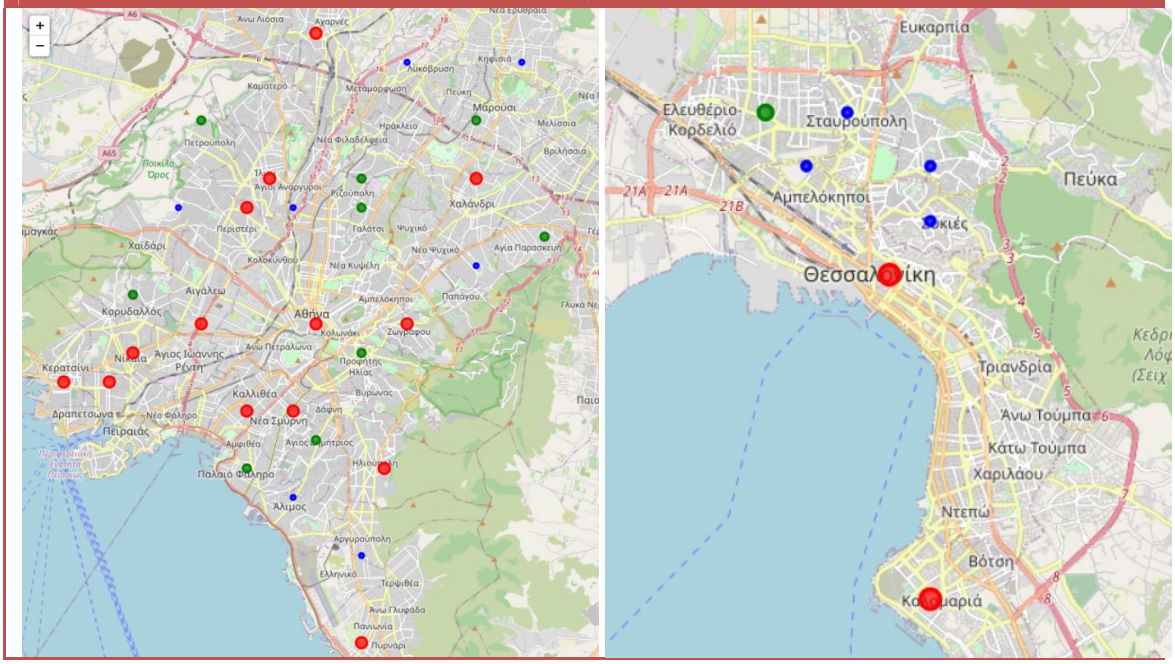
A first view of the 60 Greek Cities participating in the analysis is shown in the map below. The color and size of the circles are related to each city population. A big circle means high population. The size and color were based on 3 bins that were created based on population.

Color	Population Range
Blue	Greater than 73300
Green	Between 49400 and 71000
Red	Lower than 47400



By zooming in to the metropolitan areas of Athens and Thessaloniki, we notice that there are many cities the one next to the other. **This fact creates an alert to be careful regarding the venues that the Foursquare API will return because there may be cases where a venue is returned for more than 1 city.**

Athens	Thessaloniki



WINE BAR AND BISTRO VENUES

Explore Venue data. Checking for duplicate venues:

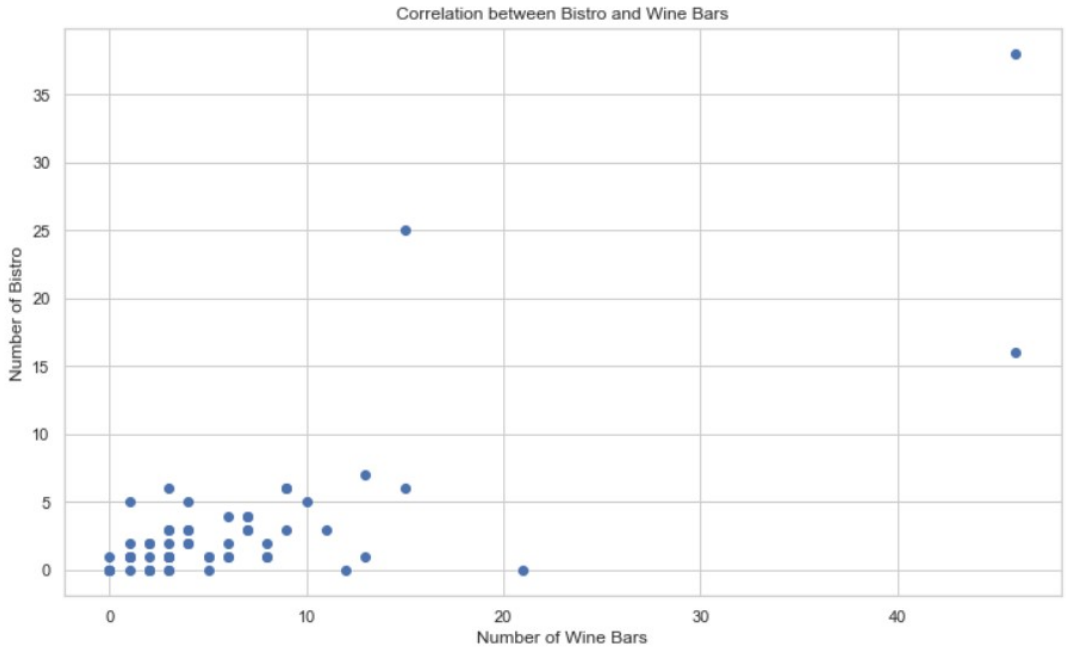
Venue_Id	Venue_Name	count(*)
4adcdadef964a520f15721e3	Peacock Roof Garden Restaurant	2
4c24c439a852c928549de36c	Coq Au Zen	2
4d41d00789c3a143abd2f183	Oinoscent	2
4d598cb524466ea852ac789f	Lithinoi (Ληθηνόη)	2
4d63e70b1a83f04d209d782b	Το Δώμα	2

It seems that we should clean the data in order for each venue to be assigned in only one city (the minimum distance from city center location). Take as example “Oinoscent”. It can be found in both Athinai and Viron. We will keep only the Athinai row (distance 689 <1631).

City	Population	City_Latitude	City_Longitude	Venue_Id	Venue_Name	Venue_Latitude	Venue_Longitude	Venue_Distance	Venue_Category
Athinai	762100	37.980	23.730	4d41d00789c3a143abd2f183	Oinoscent	37.974037	23.732125	689	Wine Bar
Viron	62500	37.970	23.750	4d41d00789c3a143abd2f183	Oinoscent	37.974037	23.732125	1631	Wine Bar

As shown in the scatter plot and after calculating Pearson’s Correlation Coefficient, the correlation between Number of Wine Bars and Bistro is relatively strong. This is normal based on the fact that both are related to the same product (wine) thus in general in a big city the more wine bars exist, the more bistro we will find.

Pearson's correlation coefficient： 0.7709239068926617
p-value. 5.80417973210664e-13



Although there is a strong correlation we can see some outliers. Lets examine the three ouliers having Number of Bistro>15 more carefully.

City	Population	Latitude	Longitude	wine	bistro
Athinai	762100	37.980	23.730	46	38
Thessaloniki	372100	40.640	22.940	46	16
Viron	62500	37.970	23.750	15	25

As expected Athinai the capital and the biggest city of Greece has a big number of wine bars and bistro. Thessaloniki, the second biggest city in Greece has a big number of wine bars but an unexpected much smaller number of bistros. Finally Viron with a small population, it has an unexpected big number of bistros (bigger than Thessaloniki that is 6 times bigger than Viron) and wine bars. In addition the second surprise is that the Wine bars are less than the bistros and this is something we should keep it in mind. The big number of Venues in Viron is partially explained because the Pagrati neighborhood, a vibrant and crowd place, is near (part of Athens city) and its closer to Viron than Athens city center.

WINE BAR DETAILS - PRICE, LIKES, RATING

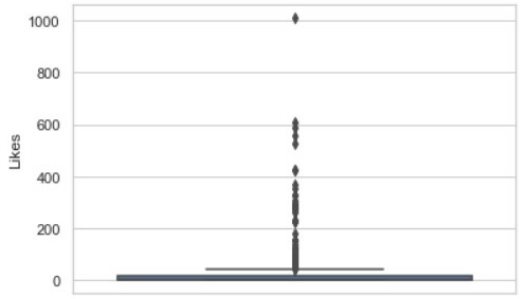
After retrieving Wine Bar Details data, we checked them for missing values. The API definition was clear that Price and Rating might be missing from the data. Checking the 388 Wine Bars returned by the API we found a big number of missing ratings, a not so big Price missing values and many Likes with value equals to zero (meaning either that no one liked the place or no one reviewed it).

missing_price	missing_rating	missing_likes
14	231	109

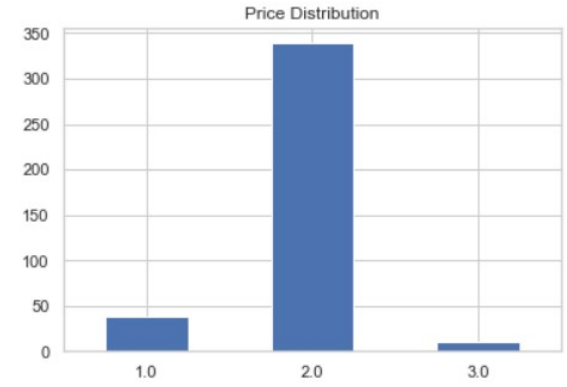
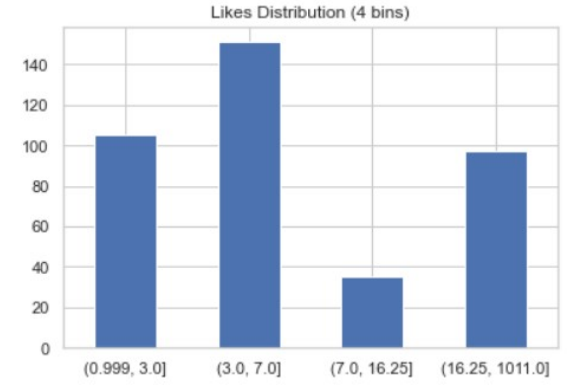
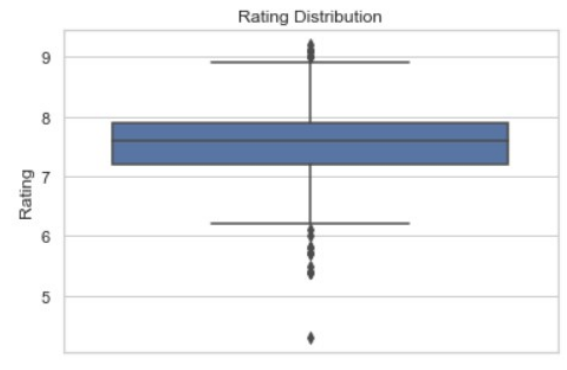
In order to fix missing rating, likes and prices, we proceeded in three steps:

1. Find the Average price and rating per city and then replace the missing values with the average of each city
2. If following step 1 there are still missing values, because there are no data to calculate average for a city, then we replace the missing value with the total average price or rating.

3. For Likes we will replace the 0 values with the median of all likes. As shown in the boxplot below there are some extreme values and all the others are in the same small range it is much preferable to use the median instead of average.



After missing values replacement the big picture regarding the Price, Likes and Rating is the following:

	Most Wine Bars have moderate prices (price band 2). There are no very expensive Wine Bars (Price band 4).
	Most Wine Bars are inside the bin 3-7 likes. It seems normal as the media value we replaced missing values was 5. Almost the 75% of the Wine Bars have less than 16 likes.
	Wine Bar Rating except some outliers with rating greater than 9 and less than 6 are in the range 6-9.

DATA PREPARATION

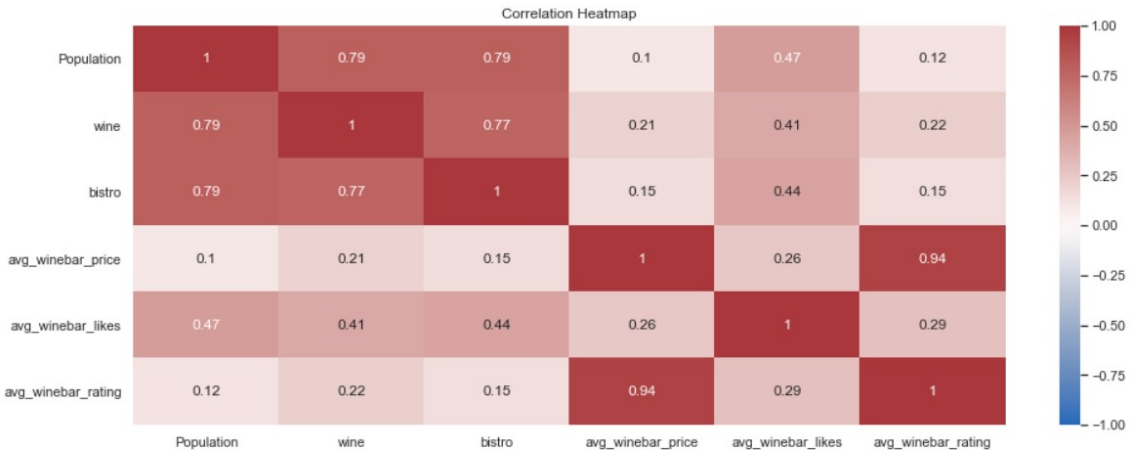
Following the exploration and manipulation of each data set, its time to bring everything together. Our point of reference and the dimension that our analysis will focus is City. After the data aggregation to City level, our data set looks like this:

	City	Population	Latitude	Longitude	wine	bistro	avg_winebar_price	avg_winebar_likes	avg_winebar_rating
0	Athinai	762100	37.980	23.730	46	38	2.000000	132.413043	7.695652
1	Thessaloniki	372100	40.640	22.940	46	16	1.869565	41.500000	7.602174
2	Piraieus	179600	37.960	23.640	4	2	2.000000	26.500000	9.100000
3	Patrai	164000	38.240	21.730	13	7	2.076923	26.384615	7.500000
4	Peristerion	141000	38.020	23.700	2	2	2.000000	5.500000	7.800000

We now have the Population, total wine bars, total bistro, average wine bars price band, average wine bars likes and average wine bars rating per City in order to continue with our analysis.

CORRELATION MATRIX FOR ALL VARIABLES

Let's examine the correlation between the variables:



We have already discussed the correlation between the Number of Wine bars and the Number of Bistro. The second strong correlation is the one between Wine Bar Average Price and Average Rating. Although we have a strong correlation we will keep the Price the variable in order to use it to Cities clustering and to help us interpret the results. In addition the correlation in reality it doesn't seem logic (the rating of place isn't straight forward connected with the prices of the place) and most probably it is because most Wine bars (almost 80%) has the same value (2) in Price.

FINALIZE THE VARIABLES

We saw that the Population, total wine bars, total bistro, average wine bars price band, average wine bars likes and average wine bars rating per City are the variables based on which, we will execute our clustering algorithm. But are all the variables meaningful and comparable? Can we transform some variables to have more meaning in our analysis? The answer is yes. Number of Wine Bars and Number of Bistro don't say anything when compared from city to city. That's why it is a good idea to divide them with the population of

the city to extract a new KPI called Number of Wine Bars per person. After creating our two new variables (one for wine bar and one for bistro), our data set looks like this:

City	population	wine_var	bistro_var	avg_winebar_price	avg_winebar_likes	avg_winebar_rating
Athinai	762100	0.000060	0.000050	2.000000	132.413043	7.695652
Thessaloniki	372100	0.000124	0.000043	1.869565	41.500000	7.602174
Piraieus	179600	0.000022	0.000011	2.000000	26.500000	9.100000
Patrai	164000	0.000079	0.000043	2.076923	26.384615	7.500000
Peristerion	141000	0.000014	0.000014	2.000000	5.500000	7.800000

Finally our variables are the following:

Variable Name	Definition	Business Meaning
population	Population of each city.	It is an indication of the size of each city.
wine_var	Number of Wine Bars per Person.	How a city competes in terms of the number of wine bars compared to other cities.
bistro_var	Number of Bistro per Person.	How a similar business to the wine bars, is doing in the same city.
avg_winebar_price	Average Wine Bar Price Band.	How expensive the Wine Bars of a city are.
avg_winbar_likes	Average Wine Bar Likes.	How much the customers like the Wine Bars of a city on average.
avg_winbar_rating	Average Wine Bar Rating.	The quality of the Wine Bars of a city on average.

NORMALIZE VALUES

Before proceeding with executing the clustering algorithm, we need to bring the data in a form where a certain variable not to influence the result more than the others variables just because the scaling is different. By making the ranges consistent between variables, normalization enables a fair comparison between the different features, making sure they have the same impact.

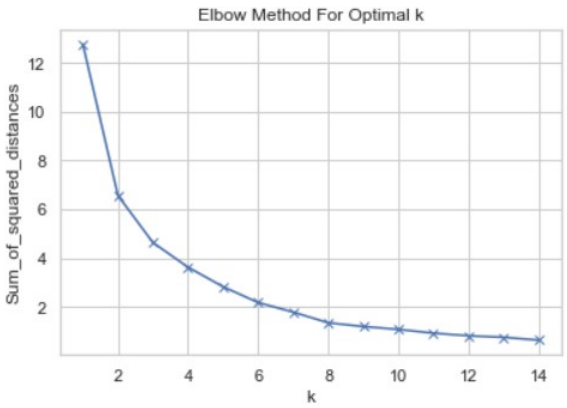
Our data looks like this after applying a MinMax Scaler to normalize the values:

	0	1	2	3	4	5
0	1.000000	0.103473	0.124656	0.800000	0.903843	0.845676
1	0.466192	0.211925	0.107498	0.747826	0.283276	0.835404
2	0.202710	0.038180	0.027840	0.800000	0.180887	1.000000
3	0.181358	0.135889	0.106707	0.830769	0.180100	0.824176
4	0.149877	0.024316	0.035461	0.800000	0.037543	0.857143

EXECUTING THE K-MEANS CLUSTERING ALGORITHM

FIND THE OPTIMAL K

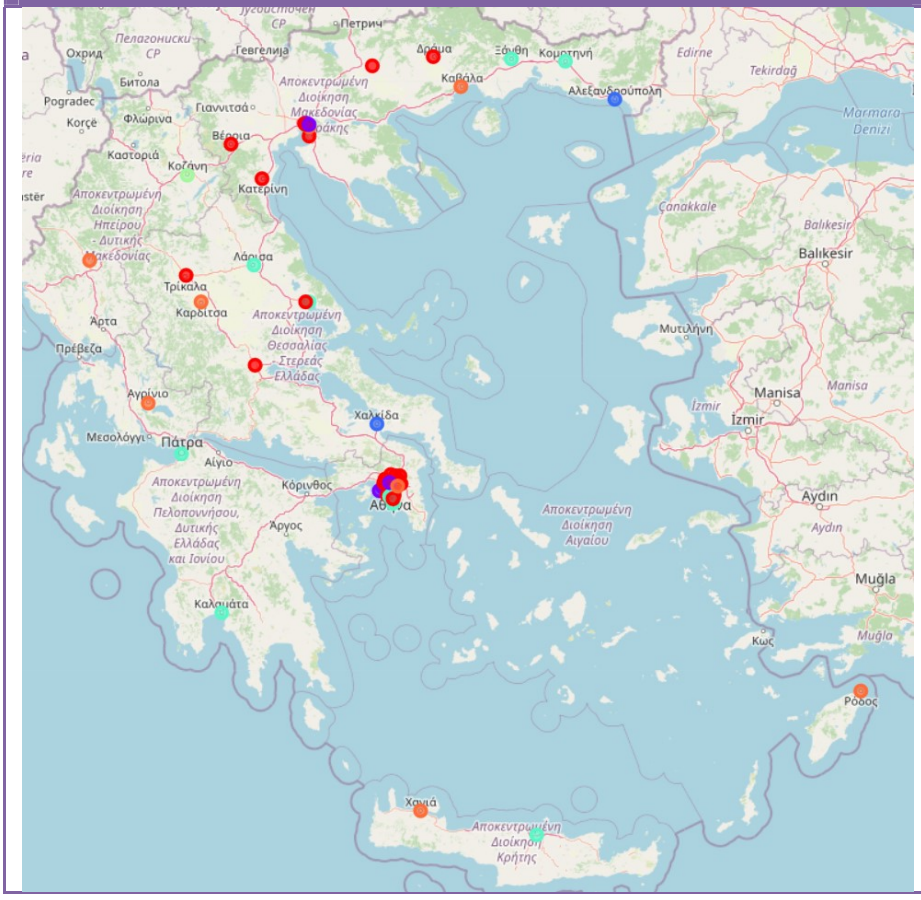
We will use K-Means to cluster the cities and one of the major questions is how many clusters to create or what is the optimal K parameter? The answer is given by the elbow method. Below there is the elbow method for our case showing that the optimal K is 8.



THE K-MEANS CLUSTERING

Using as parameter the k=8 and after executing the algorithm each city is assigned to one of the clusters created based on the characteristics of the wine bars of the city. By creating city clusters we will be able to identify common patterns and extract useful insights regarding the wine bar business in different cities. Let's see the clusters created on the map:

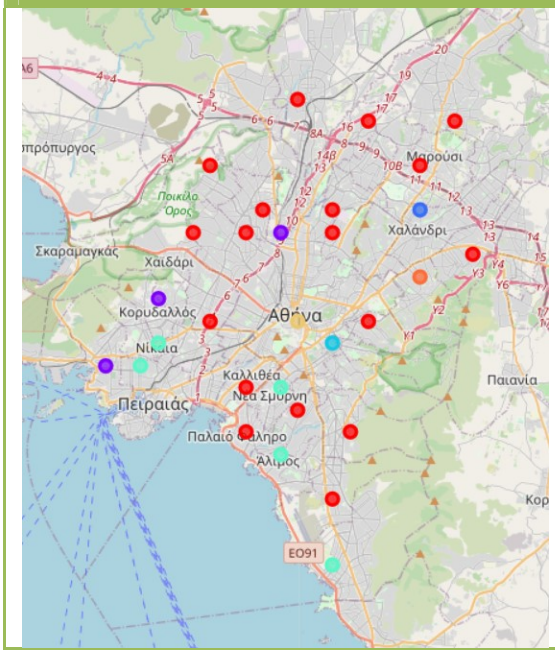
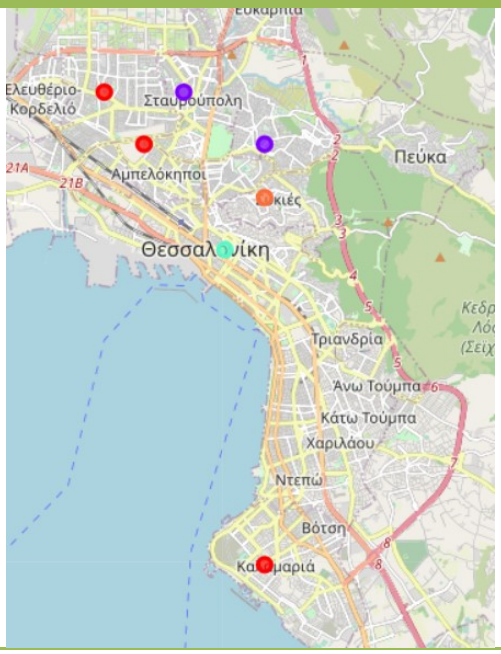
Greek Cities Clusters based on Wine Bar characteristics



Clusters Legend

Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7

Let’s zoom in to the metropolitan areas:

Athens	Thessaloniki
	

K-MEANS RESULTS EVALUATION

CLUSTERS EVALUATION WITH RAW DATA

In order to evaluate the clusters created, we will aggregate each cluster’s cities metrics to cluster level:

Cluster	cities	avg_population	avg_winebar	avg_bistro	avg(avg_winebar_price)	avg(avg_winebar_likes)	avg(avg_winebar_rating)
0	28	62921.428571	5.446856	2.548391	1.939838	8.307653	7.453784
1	5	52020.000000	0.000000	0.289855	0.000000	0.000000	0.000000
2	3	59366.666667	8.399810	5.093948	2.023810	112.103175	8.127778
3	1	62500.000000	24.000000	40.000000	1.866667	25.200000	7.213333
4	13	114892.307692	8.951285	2.855173	1.941875	40.294551	7.795584
5	1	36000.000000	58.333333	0.000000	1.904762	5.523810	6.804762
6	1	762100.000000	6.035953	4.986222	2.000000	132.413043	7.695652
7	8	47837.500000	17.596208	8.980689	1.927778	11.463095	7.222153

By comparing the metrics of each cluster with the total averages, we can interpret each cluster:

total_avg_population	total_avg_winebar	total_avg_bistro	total_avg_price	total_avg_likes	total_avg_rating
85032.727273	9.513037	4.374287	1.942271	24.50172	7.52587

We can take a first look at clustering results by checking how many cities consists each cluster, the average population of the cluster, the average number of wine bars/100000 persons, the average number of bistro/100000 persons and the average price band, likes and ratings of the wine bars inside each cluster and compare them to the total averages. As you may noticed we transformed the number of venues per person to number of venues per 100000 persons in order to help us with the interpretation and comparison of the results.

ADVANCED CLUSTERS EVALUATION

In order to easier identify the main characteristics of each cluster, interpret the clustering results and proceed with the results of our analysis, we can visualize the evaluation of the clusters as a Heatmap.

In the y-axis of the Heatmap we see the cluster number and x-axis has the six different variables In order for the evaluation to be easier, we will use three main values to depict the results of the comparison between cluster average metrics and total averages:

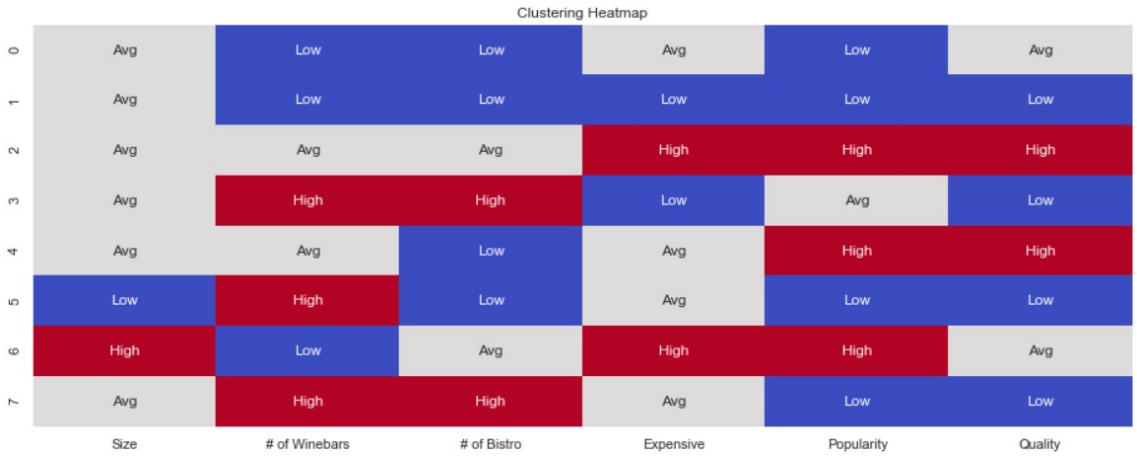
- Low: blue color, indicating that the average variable value of the cluster is below a certain threshold of the total average
- Avg: grey color, indicating that the average variable value of the cluster is around the total average
- High: red color, indicating that the average variable value of the cluster is above a certain threshold of the total average

Different threshold were used depending the variable used for the comparison, because of the different scale and distribution of each variable (i.e population vs rating):

- very_small_threshold=2% diff from the total average, used for price comparison
- small_threshold=3% diff from the total average, used for rating comparison
- medium_threshold=20% diff from the total average, used for number of wine bars and bistro comparison
- big_threshold=50% diff from the total average, used for population and likes comparison

Finally we will give more descriptive names to our variables. Population shoes how big is a city and that's why we will name it "Size", Price is showing how expensive is a wine bar so we will name it "Expensive", Likes shows if the customers love the place so we will name it "Popularity" and Rating shows how good a wine bar is so we will name it "Quality".

This is the final cluster evaluation heatmap, showing the main attributes of each cluster:



RESULTS

By examining the data, checking the metrics and the picture the Heatmap gives us for each cluster, we can give each cluster a descriptive name:

Cluster	Cities	Name	Description
0	28	The Not so Popular Average	Small number of wine bars, average quality and price, low popularity
1	5	Dry Cities	No wine bars
2	3	High Performance	Consists of cities with an average number of wine bars and exceptional quality
3	1	Champion with low quality	Many wine bars and bistro with below average quality
4	13	The Popular Average	Small number of wine bars, average price, high in quality and popularity
5	1	Quantity not Quality	Many wine bars but low quality and likes
6	1	The Capital	The biggest in size, not so many wine bars with average quality, popular places and high prices
7	8	Trying to get Improved	Many wine bars with below the average popularity and quality

Going into deep to each Cluster we can have a detailed analysis on each one:

Cluster 0

Name: The Not so Popular Average

Summary: High: - | Low: Winebars, Bistro, Popularity | Average: Size, Expensive, Quality

Cluster's Cities

Cluster	City	Population	Latitude	Longitude	wine	bistro	avg_winebar_price	avg_winebar_likes	avg_winebar_rating
0	Peristerion	141000	38.020	23.700	2	2	2.000000	5.500000	7.800000
0	Kallithea	112100	37.950	23.700	6	2	2.000000	13.333333	7.500000
0	Kalamaria	89200	40.580	22.950	7	3	1.857143	15.714286	7.814286
0	Ilion	82700	38.030	23.710	1	0	2.000000	7.000000	7.500000
0	Zografos	77800	37.980	23.770	4	5	1.750000	19.250000	8.200000
0	Ilioupoli	77600	37.930	23.760	2	0	2.000000	11.500000	8.500000
0	Akharnai	77000	38.080	23.730	2	0	2.000000	5.000000	7.500000
0	Aigaleo	75700	37.980	23.680	1	1	2.000000	7.000000	7.500000
0	Amarousion	71000	38.050	23.800	3	3	2.000000	5.000000	7.500000
0	Nea ionia	67500	38.030	23.750	1	1	2.000000	3.000000	7.500000
0	Ayios dimitrios	66600	37.940	23.730	5	0	2.000000	8.000000	8.700000
0	Palaion faliron	66200	37.930	23.700	6	1	1.833333	15.833333	6.816667
0	Galatsion	59300	38.020	23.750	3	0	1.666667	9.000000	6.200000
0	Ayia paraskevi	58100	38.010	23.830	1	5	2.000000	1.000000	7.500000
0	Serrai	55500	41.090	23.550	8	1	1.875000	11.000000	7.700000
0	Euosmon	53800	40.670	22.910	3	3	2.000000	19.000000	7.600000
0	Katerini	51600	40.270	22.500	5	1	2.000000	4.400000	7.500000
0	Trikala	49800	39.560	21.770	3	1	2.000000	3.666667	5.400000
0	Petroupoli	49400	38.050	23.680	2	2	2.000000	9.000000	7.650000
0	Lamia	47400	38.900	22.430	6	1	1.833333	2.166667	7.500000
0	Iraklion	47000	38.070	23.770	2	0	1.500000	3.000000	7.500000
0	Khaidarion	46200	38.020	23.670	1	1	2.000000	2.000000	7.500000
0	Kifisia	44900	38.070	23.820	3	2	2.000000	10.666667	6.200000
0	Veroia	43700	40.520	22.200	4	3	2.000000	19.250000	7.625000
0	Drama	43400	41.150	24.140	3	1	2.000000	11.000000	7.600000
0	Ambelokipoi	41900	40.660	22.920	1	2	2.000000	4.000000	7.500000
0	Aryiroupoli	33900	37.900	23.750	3	0	2.000000	4.333333	7.400000
0	Nea ionia	31500	39.370	22.920	1	1	2.000000	3.000000	7.500000

Comment: **The biggest cluster consists of cities with no strong points. We can assume that this is the average having at the same time less number of wine bars per person than the average, that at the same time are not so popular. To be more precise the average Likes of cluster's wine bars are the second worst from the seven clusters that have wine bars. The characteristics of this cluster don't show an opportunity or something special. It's an average after all.**

Cluster 1

Name

Dry Cities

Summary

No Wine Bars

Cluster's Cities

Cluster	City	Population	Latitude	Longitude	wine	bistro	avg_winebar_price	avg_winebar_likes	avg_winebar_rating
1	Keratsinion	77800	37.960	23.620	0	0	0.0	0.0	0.0
1	Koridallos	69000	37.990	23.650	0	1	0.0	0.0	0.0
1	Stauroupolis	42600	40.670	22.930	0	0	0.0	0.0	0.0
1	Polikhni	37000	40.660	22.950	0	0	0.0	0.0	0.0
1	Ayioi anaryiroi	33700	38.020	23.720	0	0	0.0	0.0	0.0

Comment

This is an expected cluster with cities that don't have any wine bars. This fact can be probably explained form the fact that all 5 cities are part of metropolitan areas, so there maybe another center next to them where restaurant and bars exists and social life is more developed. There is no reason to invest in such an area.

Cluster 2

Name

High Performance

Summary

High: Popularity, Expensive, Quality | Low: - | Average: Size, Winebars, Bistro

Cluster's Cities

Cluster	City	Population	Latitude	Longitude	wine	bistro	avg_winebar_price	avg_winebar_likes	avg_winebar_rating
2	Khalandrion	73300	38.030	23.800	7	4	1.571429	110.142857	8.200000
2	Khalkis	54800	38.460	23.600	2	1	2.500000	146.500000	8.100000
2	Alexandroupoli	50000	40.850	25.870	6	4	2.000000	79.666667	8.083333

Comment

These 3 cities are top in terms of Wine bars quality and Popularity. It is a bit risky to invest in such an area where people are happy with the existing wine bars.

Cluster 3

Name

Champion with low quality

Summary

High: Winebars, Bistro | Low: Expensive, Quality | Average: Size, Popularity

Cluster's Cities

Cluster	City	Population	Latitude	Longitude	wine	bistro	avg_winebar_price	avg_winebar_likes	avg_winebar_rating
3	Viron	62500	37.970	23.750	15	25	1.866667	25.2	7.213333

Comment

This one city cluster is the only one that has at the same time a high number of wine bars and bistro but the number of bistro (40 per 100k persons) is greater (almost double) than the number of wine bars (24 per 100k persons). It is worth mention that the Quality of the existing wine bars are below average. This city needs more analysis and it is a good candidate to host one more wine bar.

Cluster 4

Name: The Popular Average

Summary: High: Popularity, Quality | Low: Bistro | Average: Winebars, Expensive, Size

Cluster's Cities:

Cluster	City	Population	Latitude	Longitude	wine	bistro	avg_winebar_price	avg_winebar_likes	avg_winebar_rating
4	Thessaloniki	372100	40.640	22.940	46	16	1.869565	41.500000	7.602174
4	Piraieus	179600	37.960	23.640	4	2	2.000000	26.500000	9.100000
4	Patrai	164000	38.240	21.730	13	7	2.076923	26.384615	7.500000
4	Iraklion_Crete	133800	35.330	25.130	9	6	2.111111	55.111111	7.233333
4	Larisa	127200	39.640	22.420	13	1	2.076923	31.692308	7.592308
4	Nikaia	95200	37.970	23.650	5	1	2.000000	26.600000	7.900000
4	Volos	84300	39.370	22.950	7	3	2.000000	33.571429	7.600000
4	Glifada	82200	37.870	23.750	8	2	1.875000	38.875000	7.812500
4	Nea smirni	75600	37.950	23.720	11	3	1.818182	29.636364	7.827273
4	Kalamata	50300	37.040	22.110	8	1	2.000000	54.875000	7.575000
4	Xanthi	46100	41.140	24.880	4	2	1.750000	36.750000	8.600000
4	Komotini	44300	41.120	25.400	3	1	2.000000	61.333333	7.833333
4	Kalamakion	38900	37.920	23.720	3	1	1.666667	61.000000	7.166667

Comment: The second biggest cluster with 13 cities consists of cities with nothing special in terms of number of wine bars but it performs high in terms of quality and popularity. We could say that this is the alter ego of cluster 0 having similar characteristics but the exact opposite behavior regarding popularity and quality. In addition it is similar to "High Performance" cluster 2 both having an average number of wine bars and high popularity and quality. Their difference is that cluster 2 is much better in popularity with an average of 112 likes vs 40 likes and an average rating of 8.13 vs 7.79. The same comment as in cluster 2 is valid here. Too risky to invest to a high performance city with already popular places with good quality.

Cluster 5

Name: Quantity not Quality

Summary: High: Winebars | Low: Size, Bistro, Quality | Average: Expensive

Cluster's Cities:

Cluster	City	Population	Latitude	Longitude	wine	bistro	avg_winebar_price	avg_winebar_likes	avg_winebar_rating
5	Kozani	36000	40.300	21.790	21	0	1.904762	5.52381	6.804762

Comment: One of a kind city and the only one that compose this cluster. Kozani is a small city having an impressive number of wine bars compared to the population with 58 wine bars per 100k persons. On the other hand is very poor when it comes to popularity and quality. This fact makes Kozani a city worth thinking it as the city of your investment.

Cluster 6

Name: The Capital

Summary: High: Size, Expensive, Popularity | Low: Winebars | Average: Bistro, Quality

Cluster's Cities:

Cluster	City	Population	Latitude	Longitude	wine	bistro	avg_winebar_price	avg_winebar_likes	avg_winebar_rating
6	Athinai	762100	37.980	23.730	46	38	2.0	132.413043	7.695652

Comment: The Greek capital is the only city of cluster 6. This is normal as its characteristics are unique. The major difference here is the size (762000 population) that is double than the second biggest city in Greece (Thessaloniki - 372100). At the same time it is an expensive city with wine bars of average quality. With the results we have in our hands now, it is not so attractive to invest there. In order to check the possibility of opening a new wine bar in Athens a different analysis should take place on each neighborhoods or hot areas of the city.

Cluster 7

Name	**Trying to get Improved**
Summary	**High: Winebars, Bistro | Low: Popularity, Quality | Average: Size, Expensive**

Cluster's Cities

Cluster	City	Population	Latitude	Longitude	wine	bistro	avg_winebar_price	avg_winebar_likes	avg_winebar_rating
7	Ioannina	63000	39.670	20.850	10	5	1.700000	14.700000	6.810000
7	Kavala	60000	40.940	24.400	9	3	1.888889	6.444444	7.300000
7	Khania	54600	35.510	24.020	15	6	2.000000	13.600000	7.986667
7	Rodos	53500	36.440	28.220	9	6	2.000000	4.888889	6.788889
7	Agrinion	43300	38.620	21.410	7	4	2.000000	8.571429	6.900000
7	Sikeai	42700	40.650	22.950	12	0	1.833333	22.583333	7.591667
7	Kholargos	32900	38.000	23.800	4	3	2.000000	17.250000	6.900000
7	Karditsa	32700	39.370	21.920	3	6	2.000000	3.666667	7.500000

Comment

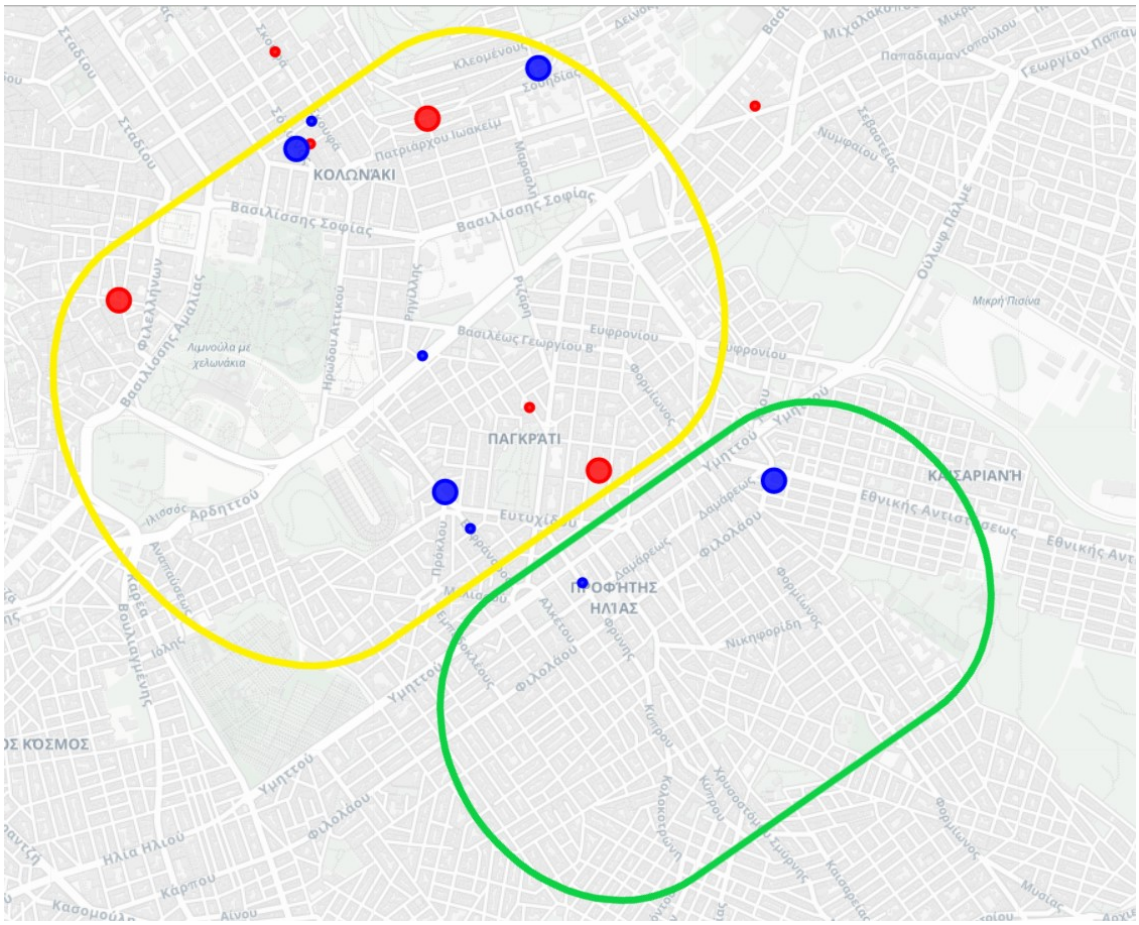
The third cluster in size (8 cities) that compares to cluster 0 "The Not so Popular Average" and cluster 4 "The Popular Average", is a cluster with high number of wine bars and bistros, almost the double of the average. This is exactly the difference with cluster 0 and the low popularity and quality is its difference with cluster 4. The fact that the winebars of the cities of this cluster are not so popular and of low quality makes this cluster interesting to examine one by one the 8 cities in order to identify potential business opportunity.

DISCUSSION

Following the interpretation of clustering results, we came up with three clusters that have interest in investigate the possibility to invest to a new wine bar. The three clusters were Cluster 3 – **"Champion with low quality"**, Cluster 5 – **"Quantity not Quality"** and Cluster 7 – **"Trying to get Improved".** From the clusters we selected, it seems that the number of wine bars per person and the quality are the features played the most important role in the selection. Features like Popularity played role on discouraging us of selecting cluster. Price showing how expensive a wine bar is didn't played significant role in this phase but it is an important feature to categorize the wine bars into clusters and it will definitely be used when it come to open a new wine bar in order to understand how the competition works. Let's examine each one of the selected cluster in depth.

CLUSTER 3 – **"CHAMPION WITH LOW QUALITY"**,

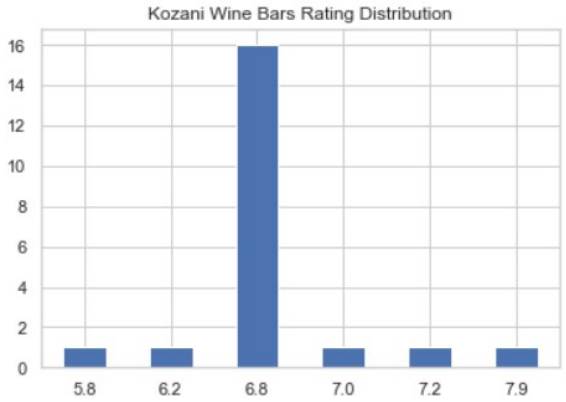
Champion because of the very high number of wine bars and bistro but low quality because of the below the average ratings. Looking better into the data we see that the number of bistro are almost double than the number of wine bars and that is a strange thing because this is the only city where the bistros are more than wine bars. In addition, Bistro are a wine related restaurant very close to wine bars (where in general they serve small plates to accompany the wine) but not so popular. For sure this is a wine friendly city and more wine bars are welcome. Now let's see where Viron is and where the wine bars are located. In the map below the blue/red color shows low/high likes and the size of the bullets shows low/high rating.



Viron borders city of Athens, especially Pagrati area where many bars, café and restaurants exist and is a part of Athens metropolitan area. The campus of University of Athens is near and many students are living in Pagrati and Viron area. The anthropogeography of the area is very important and it must be combined with the numeric data in order to decide how to proceed. Inside the yellow shape is the wine bars that belong to Athens city but are closer to Viron. Green area is mainly the Viron area with a part of Pagrati. The big red wine circled (High rating with high popularity) are the top performance wine bars of the area while the small blue circles are places that are low performing in terms of popularity and rating. All these must be taken into account in order to evaluate specific places before opening a wine bar in Viron, this crowded city (part of a metropolitan area) with the low quality wine bars.

CLUSTER 5 – **"QUANTITY NOT QUALITY"**

Cluster 5 includes only Kozani, a small Greek city located in northern Greece. Kozani has an impressive number of wine bars compared to the population with an average of 58 wine bars per 100k persons and at the same time it is one of the worst city (bottom 5) in the average rating of the area's wine bars with 6.8/10. Also an impressive fact for Kozani is that there are no bistros in the area. All these facts show a city that people like hanging out to wine bars but they are not very happy with them. Only 3 out of 21 wine bars have rating more than 7 and only 1 is above the Greek total average of 7.53.

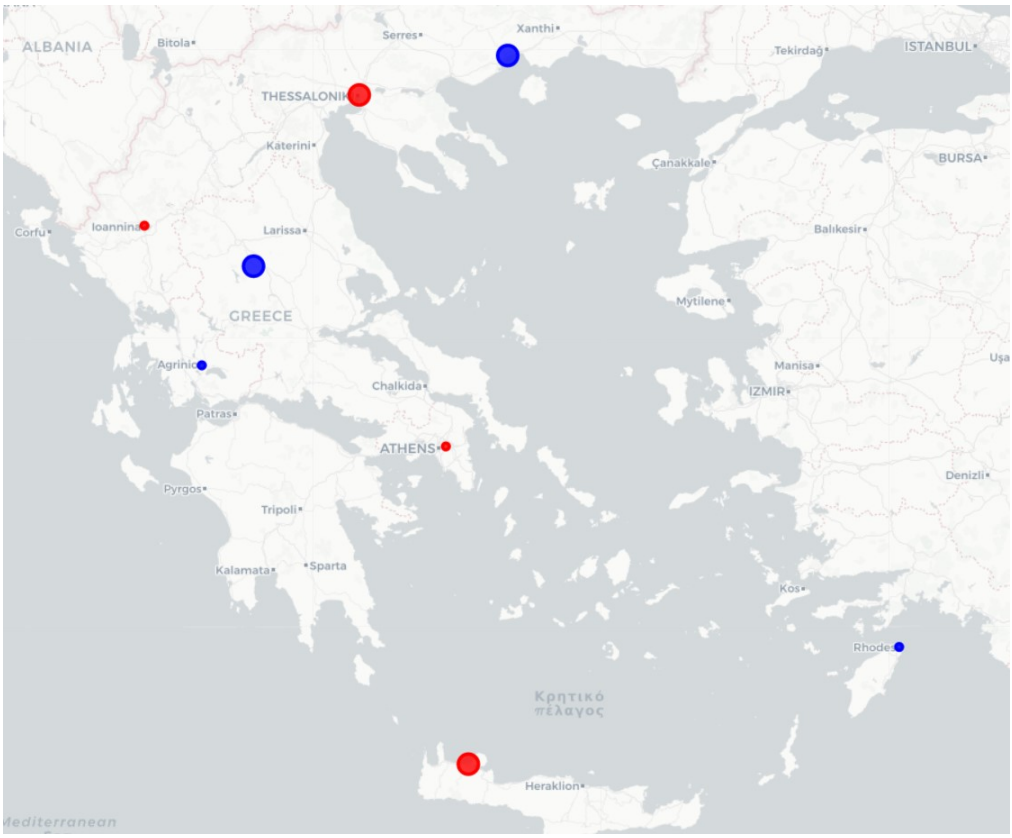


To complete the picture of the city some basic economy information taken from Wikipedia (https://en.wikipedia.org/wiki/Kozani) is that Kozani is well known for its important contribution to the Greek electricity supply, and a large part **of the population works in the Public Power Corporation's Agios Dimitrios Power Plant**, the largest power plant in Greece. The Ptolemaida Basin hosts the Western Macedonia Lignite Center, which is accountable for the production of 40% of the electric energy of the country. Other famous products are marble, **saffron (Krokos, Kozanis), fruits, local wines** and specialized arts and crafts industry. The Commercial Exhibition of Kozani takes part in the Exhibition Centre of West Macedonia in Koila Kozanis every September.

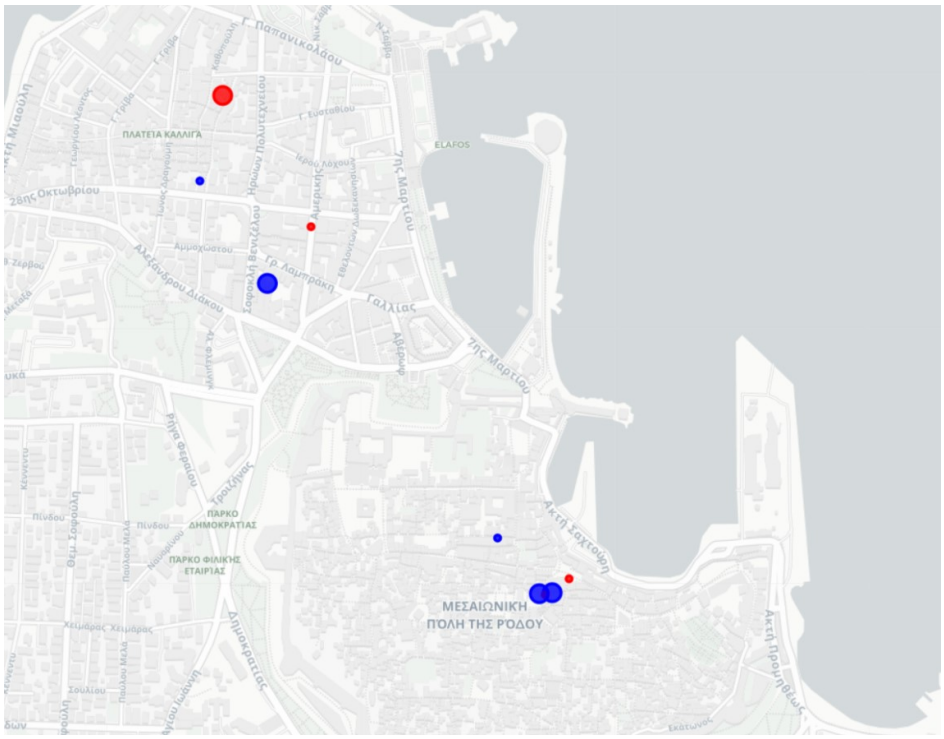
A good quality wine bar serving local wines and products with below the average prices has good possibilities to win the locals and become the new hot spot of the area.

CLUSTER 7 – **"TRYING TO GET IMPROVED"**

This cluster has 8 cities placed all over Greece. Main characteristics of the cluster are the high number of wine bars and the poor popularity and quality. Let's remember the cities of the cluster. The blue/red color shows low/high likes and the size of the bullets shows low/high rating. As a reminder the big red circles (High rating with high popularity) are the top performance cities of the cluster while the small blue circles are places that are low performing in terms of popularity and rating.



From the cities on the map, the two small blue circles, cities Agrinio and Rhodes maybe interesting cases for investment. Agrinio is an agriculture area mostly known for the tobacco industry and the tobacco crops. On the other hand Rhodes is a famous very touristic island of the Aegean sea. From both areas let's focus on Rhodes which because of the tourism seems a very good choice. The city of Rhodes has 9 wine bars with an average rating below 7. The location of the wine bars on the map is this.



Making a quick search on Rhodes we find the following on Wikipedia (https://en.wikipedia.org/wiki/Rhodes): “Rhodes is the largest of the Dodecanese islands of Greece and is also the island group's historical capital. The economy is tourist-oriented, and the most developed sector is service. Tourism has elevated Rhodes economically, compared to the rest of Greece”. It is clear even you have not visited Rhodes, that tourism is the big chance here. With almost 2 millions of tourists every year (in 2015 the arrival number of tourists was about 1,901,000) a new wine bar with good quality (missing from the island) around or inside the historic center serving mainly Greek wines from all over Greece it is the right thing to do. With the power of social networks and travel guides/place recommendation sites like Instagram, Foursquare, Trip Advisor etc the new wine bar will quickly travel all over the world.

CONCLUSION

In this report we analyzed the wine bar status of the 60 biggest cities in Greece via clustering, resulting proposal on where and why someone could open a new wine bar. We used Greek Population and location info combined with Wine bars data taken from Foursquare API. 6 main variables for each city were extracted in order to be used to our analysis: Population, Number of Wine bars / Person, Number of Bistro / Person, Wine Bar average Price Range, average Number of Likes and Rating average. We explored the data, transform them and prepare them for K-Means Clustering. We executed the clustering algorithm producing 8 clusters and we analyzed each one of them taking into account the numerical data. Finally we proposed 3 clusters on which potential for new business opportunity was identified. What we need to have in mind is that numeric data and analysis are a very strong tool in our hands, but understanding the culture of the people of a city, talk to them, walk to the streets and feel the vibes and the atmosphere is equally important before making your next business move. Stay calm and drink wine!