The Wine Alliance

Applied Data Science - Coursera
Capstone Project Course
by Leopoldo Sprandel

On the web – January 2018

Capstone Report

Business Problem

In May 25th we celebrate the National Wine Day in Brazil. All enthusiasts participates on a two weeks of ppreciation, learning sections and winery visits around the country. This event promotes the interchange of producers, wine shops and consumers moving all economy around the wine. The tourism in the city of São Paulo is in the rout of the event.

Now imagine we want to choose some Wine Bars in the city to organize a large alliance between them to promote the consumers interchange over the Wine shops. The Wine shops need to be well known and we are looking for somehow connections between them.

Data

The data we choose to select the wine bars that can participate on this alliance are the stores with high recommendations and are part of a network based on the public.

So, we need to answer two questions:

Which wine bars in São Paulo are best evaluated?

Is the wine bars connected? Which winery shares the same public?

To answer these questions we can analyze the wine shops listed in the Foursquare API in the city of São Paulo, regarding the following points:

Check the rank of wine bars (the first 100th)

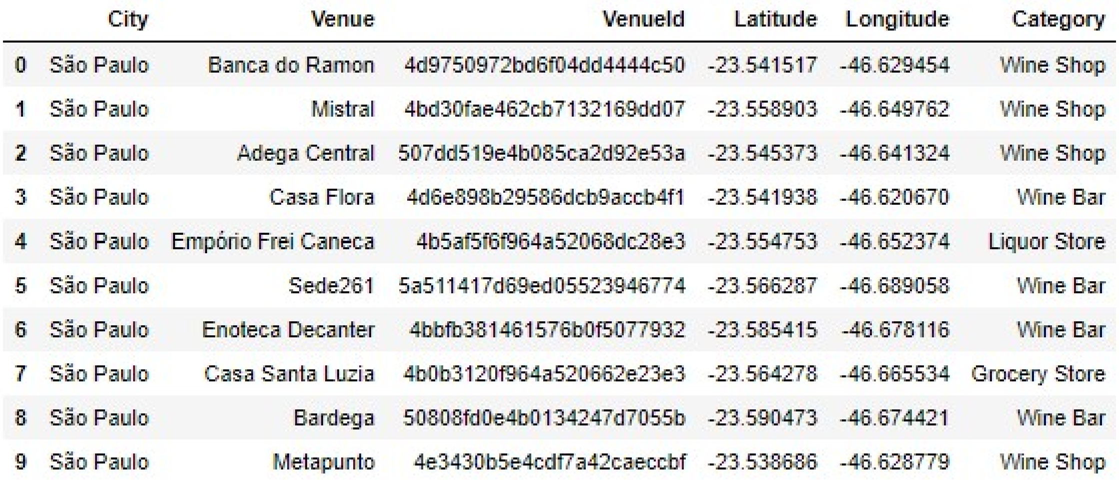
Wine bar	Ranking
WineBarA	1
WineBarB	2
WineBarC	3
WineBarC	4
WineBar...	...
WineShopX	100

A network representation can be done connecting stores who shares the same consumer (signalised by the likes or tips).

Wine Store	Conections
WineBarA	WineBarB, WineBarC
WineBarB	WineBarA, WineBarC, WineBarD, WineBarX
WineBarC	...

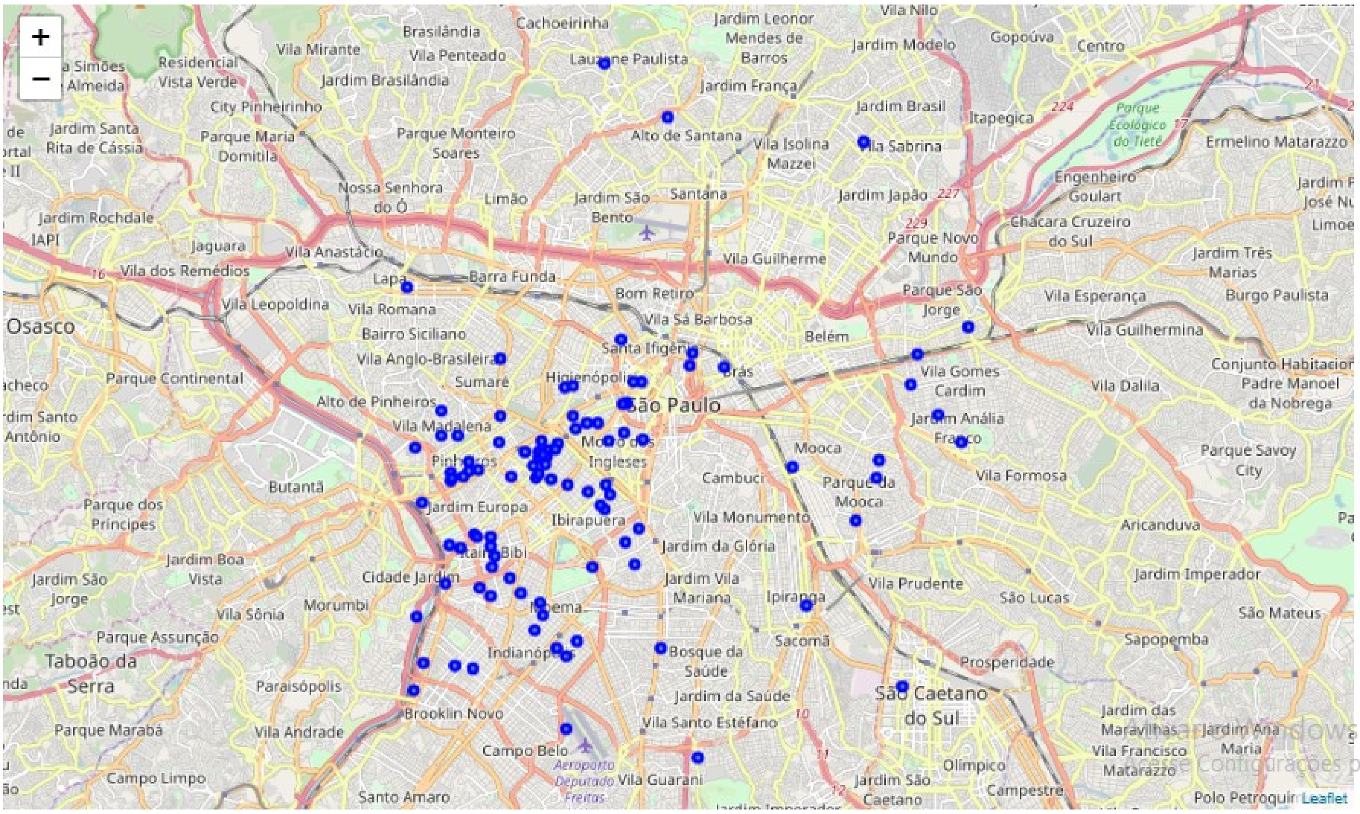
Based on the localization of the Center city of São Paulo, I've get all venues related with WINE in the region with 10km radius.

Here, the first 10 venues:



	City	Venue	VenueId	Latitude	Longitude	Category
0	São Paulo	Banca do Ramon	4d9750972bd6f04dd4444c50	-23.541517	-46.629454	Wine Shop
1	São Paulo	Mistral	4bd30fae462cb7132169dd07	-23.558903	-46.649762	Wine Shop
2	São Paulo	Adega Central	507dd519e4b085ca2d92e53a	-23.545373	-46.641324	Wine Shop
3	São Paulo	Casa Flora	4d6e898b29586dcb9accb4f1	-23.541938	-46.620670	Wine Bar
4	São Paulo	Empório Frei Caneca	4b5af5f6f964a52068dc28e3	-23.554753	-46.652374	Liquor Store
5	São Paulo	Sede261	5a511417d69ed05523946774	-23.566287	-46.689058	Wine Bar
6	São Paulo	Enoteca Decanter	4bbfb381461576b0f5077932	-23.585415	-46.678116	Wine Bar
7	São Paulo	Casa Santa Luzia	4b0b3120f964a520662e23e3	-23.564278	-46.665534	Grocery Store
8	São Paulo	Bardega	50808fd0e4b0134247d7055b	-23.590473	-46.674421	Wine Bar
9	São Paulo	Metapunto	4e3430b5e4cdf7a42caeccbf	-23.538686	-46.628779	Wine Shop

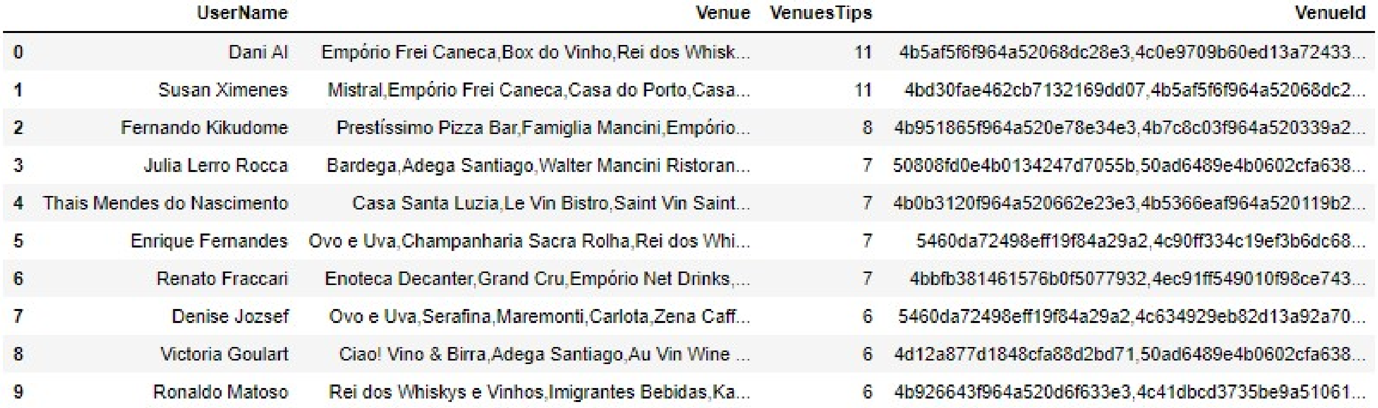
The representation of all first 100 venues from the Foursquare API. We can see the majoritie is located in the region between Higienópolis, Pinheiros and Ibirapuera. Another group inthe south city (Itain Bibi and Moema). All these regions are well known as noble neghborhoods.



To find a connection between the venues I try to take the common likes from the Foursquare API, but the free account limits the number of user I can see. They limits to 3 or 4.

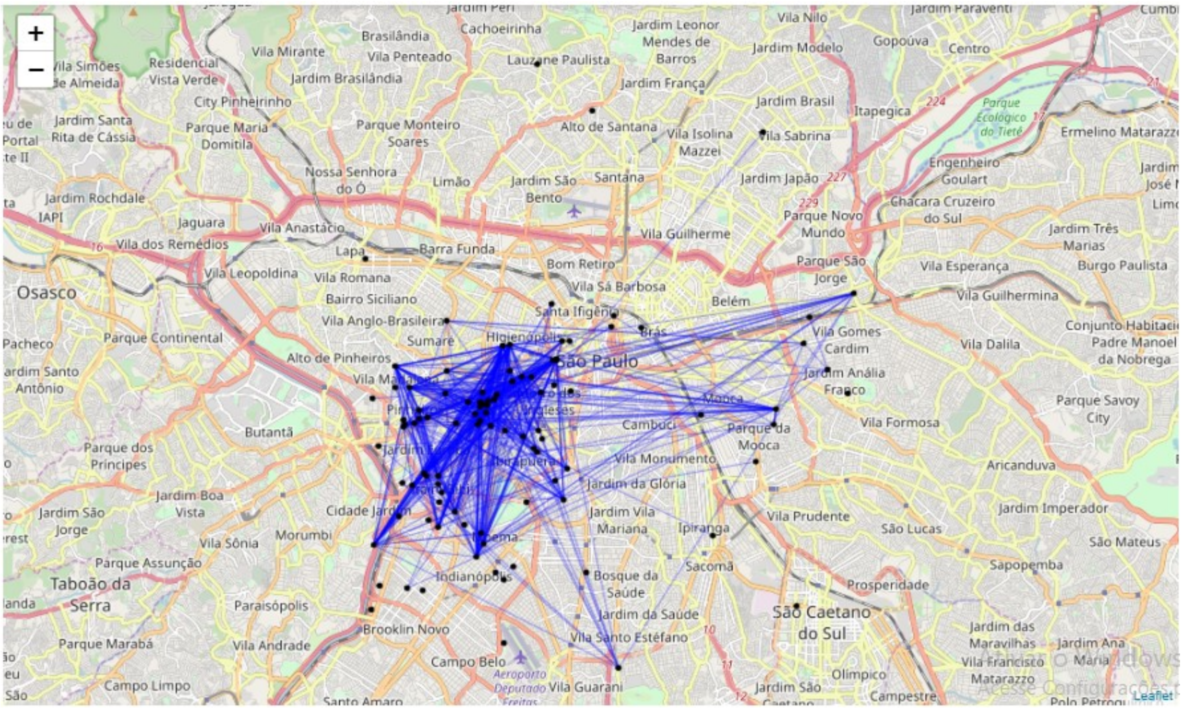
Due to this constrain, I took all the tips (not likes) directly from the web site looking inside the HTML code for each venue. And we got the following result:

A list of the users who did a tip for each venue (we limited to 50 users per venue). In the column 'Venue' is presented all venues that the user have did a tip.



	UserName	Venue	VenuesTips	VenueId
0	Dani Al	Empório Frei Caneca,Box do Vinho,Rei dos Whisk...	11	4b5af5f6f964a52068dc28e3,4c0e9709b60ed13a72433...
1	Susan Ximenes	Mistral,Empório Frei Caneca,Casa do Porto,Casa...	11	4bd30fae462cb7132169dd07,4b5af5f6f964a52068dc2...
2	Fernando Kikudome	Prestíssimo Pizza Bar,Famiglia Mancini,Empório...	8	4b951865f964a520e78e34e3,4b7c8c03f964a520339a2...
3	Julia Lerro Rocca	Bardega,Adega Santiago,Walter Mancini Ristoran...	7	50808fd0e4b0134247d7055b,50ad6489e4b0602cfa638...
4	Thais Mendes do Nascimento	Casa Santa Luzia,Le Vin Bistro,Saint Vin Saint...	7	4b0b3120f964a520662e23e3,4b5366eaf964a520119b2...
5	Enrique Fernandes	Ovo e Uva,Champanharia Sacra Rolha,Rei dos Whi...	7	5460da72498eff19f84a29a2,4c90ff334c19ef3b6dc68...
6	Renato Fraccari	Enoteca Decanter,Grand Cru,Empório Net Drinks,...	7	4bbfb381461576b0f5077932,4ec91ff549010f98ce743...
7	Denise Jozsef	Ovo e Uva,Serafina,Maremonti,Carlota,Zena Caff...	6	5460da72498eff19f84a29a2,4c634929eb82d13a92a70...
8	Victoria Goulart	Ciao! Vino & Birra,Adega Santiago,Au Vin Wine ...	6	4d12a877d1848cfa88d2bd71,50ad6489e4b0602cfa638...
9	Ronaldo Matoso	Rei dos Whiskys e Vinhos,Imigrantes Bebidas,Ka...	6	4b926643f964a520d6f633e3,4c41dbcd3735be9a51061...

So, we consider all venues that have a tip from the same person are connected. And the resultant network is shown below:



Based on this network we can do some interesting analysis about the venues and their connections:

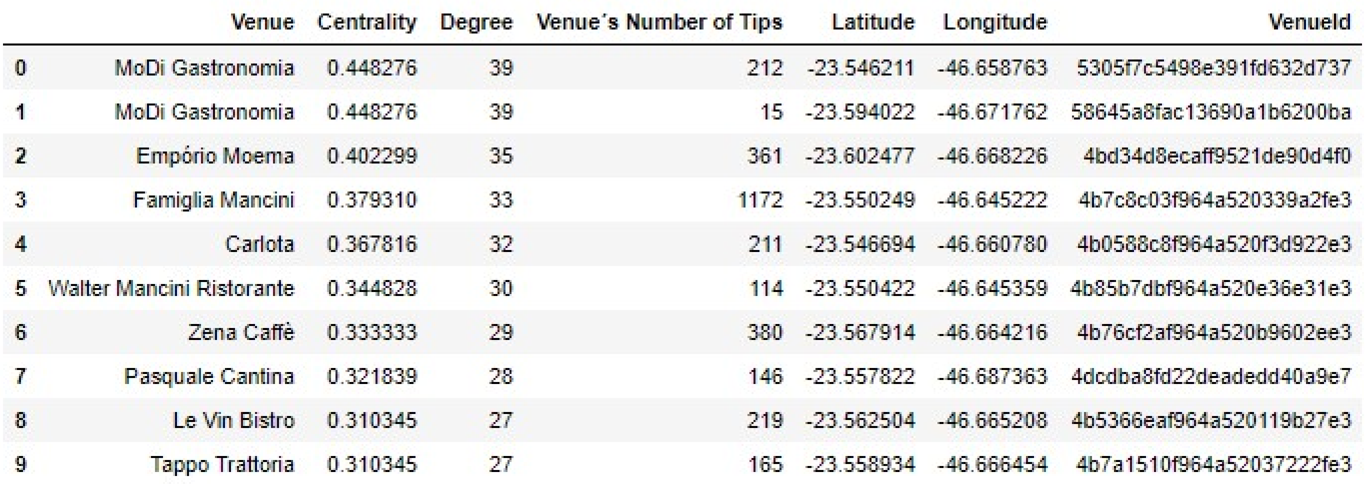
Network Analysis

Degree

With the Networkx, we can calculate the number of connections for all nodes (It's the number os edges in a venue).

Degree Centrality

One of the most widely used and important conceptual tools for analysing networks. **Centrality aims to find the most important nodes in a network** . Centrality measures themselves have a form of classification. In the next table, is presented the 10^{th} venues with highest centrality.



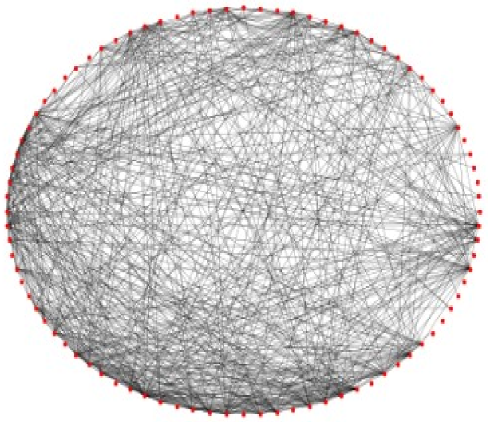
	Venue	Centrality	Degree	Venue's Number of Tips	Latitude	Longitude	VenueId
0	MoDi Gastronomia	0.448276	39	212	-23.546211	-46.658763	5305f7c5498e391fd632d737
1	MoDi Gastronomia	0.448276	39	15	-23.594022	-46.671762	58645a8fac13690a1b6200ba
2	Empório Moema	0.402299	35	361	-23.602477	-46.668226	4bd34d8ecaff9521de90d4f0
3	Famiglia Mancini	0.379310	33	1172	-23.550249	-46.645222	4b7c8c03f964a520339a2fe3
4	Carlota	0.367816	32	211	-23.546694	-46.660780	4b0588c8f964a520f3d922e3
5	Walter Mancini Ristorante	0.344828	30	114	-23.550422	-46.645359	4b85b7dbf964a520e36e31e3
6	Zena Caffè	0.333333	29	380	-23.567914	-46.664216	4b76cf2af964a520b9602ee3
7	Pasquale Cantina	0.321839	28	146	-23.557822	-46.687363	4dcdba8fd22deadedd40a9e7
8	Le Vin Bistro	0.310345	27	219	-23.562504	-46.665208	4b5366eaf964a520119b27e3
9	Tappo Trattoria	0.310345	27	165	-23.558934	-46.666454	4b7a1510f964a52037222fe3

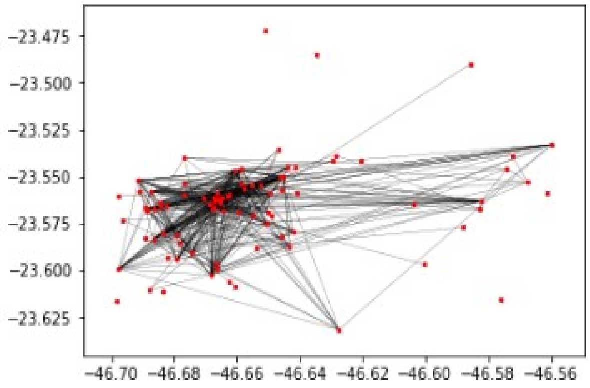
Network Density

A measure of how many edges a Graph has.

The actual definition will vary depending on type of Graph and the context in which the question is asked. For a complete undirected Graph the Density is 1, while it is 0 for an empty Graph. Graph Density can be greater than 1 in some situations (involving loops).

Our network density is 0.146. Seems to be small, but when we see the network representation in a circle it looks to bee well connected:



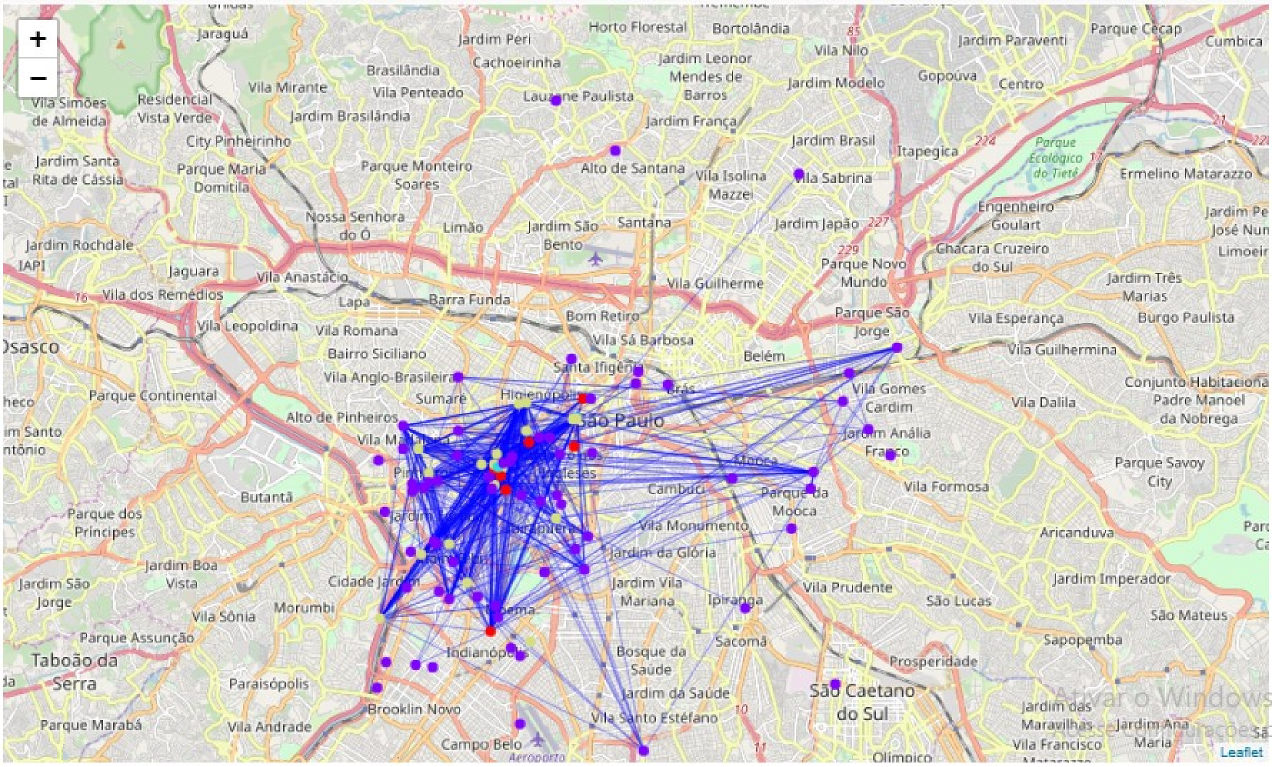


Clustering

To cluster the venues we can use the k-means technique besed on the number of tips, centrality, degree and location (number of clusters:4)

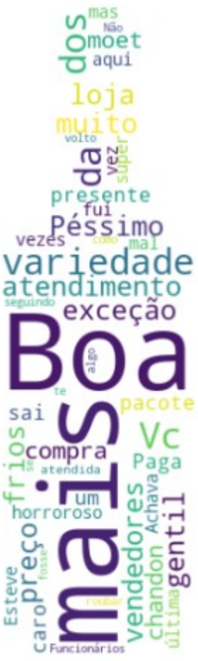
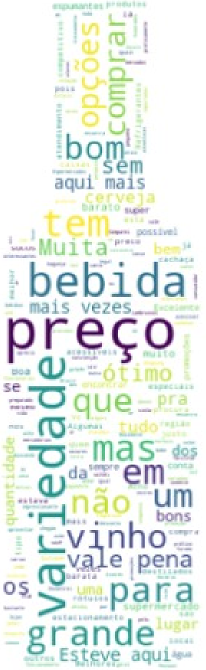
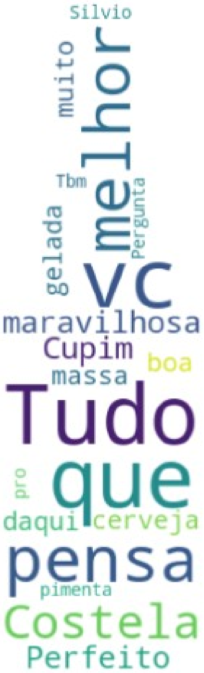
The result is:

- Cluster 0:(67 venues) is the group of Venues with low degree and low number of tips around all city
- Cluster 1:(7 venues) is a group with a big number of tips and located in the center city (perhaps the main wine shops)
- Cluster 2:(2 venues) a high number of tips and majoritary located in the center city
- Cluster 3:(18 venues) High number of tips and located and high centrality



Miscellaneous

A work cloud can be genereted based on the tips that all user did for a venue. It is a interesting print of the main words used by all user over each venue.
As example we took 3 interesting venues:

Paris 6	Famiglia Mancini	Cantina C... Que Sabe!
		
This is a very controversial venue. It is new (less than 10 years) it is in the top of tips. Has good word and bad words. (Ex.: good, variety, lousy, gentle, care, appalling)	This is a good one. Very traditional. Many tips, good words. (Ex.: price, variety, drink, great, wine, worth, more times)	With small centrality and with a high number of tips only good words on the word cloud. (Ex.: everything, think, costa, best, wonderful, pasta, beer, cold, perfect)

Conclusion

With the data free available on the internet, we are able to find any kind of venue and make interesting researches. The only thing we need is curiosity and creativity.

On that case of study, we took the name and location of a hundred of venues that are labelled with *WINE.* That means a place where we can drink a good glass of wine.
We found a connection between all that venues based on the shared clients or users on the foursquare. We can infer two venues have a same client when this client does a tip for these venues.

Analysing the constructed network, we can see the most central venue in that network (maybe the most influent). Based on the degrees of each node, or venue, we can see how they are connected and if we would like separate in groups, perhaps to organise different fronts of work, we can use machine learn to cluster the dataframe.

Clustering in 4 groups, we see hall the Wine shops can be grouped based on network properties and geographical position.

To finalise, looking the words people used to make all tips, we can do a word cloud and have an idea of how the user see each venue (and in a nice picture).