Exploring Burger Joint Clusters in São Paulo

GUILHERME LEITE ALVES DA COSTA

Exploring Burger Joint Clusters in São Paulo

IBM Data Science Capstone Project

São Paulo-SP-Brazil
2020

Table of Contents

Introduction

São Paulo, is one of the largest cities in the world, it has more than 12 million people and is Brazil's economy powerhouse. It has a plethora of venues and restaurants with specialties from all around the world. In special, there is one type of restaurant category that is remarkably popular in São Paulo: burger joints. São Paulo has more than 600 burger joints in total and some neighborhoods have more than one per block.

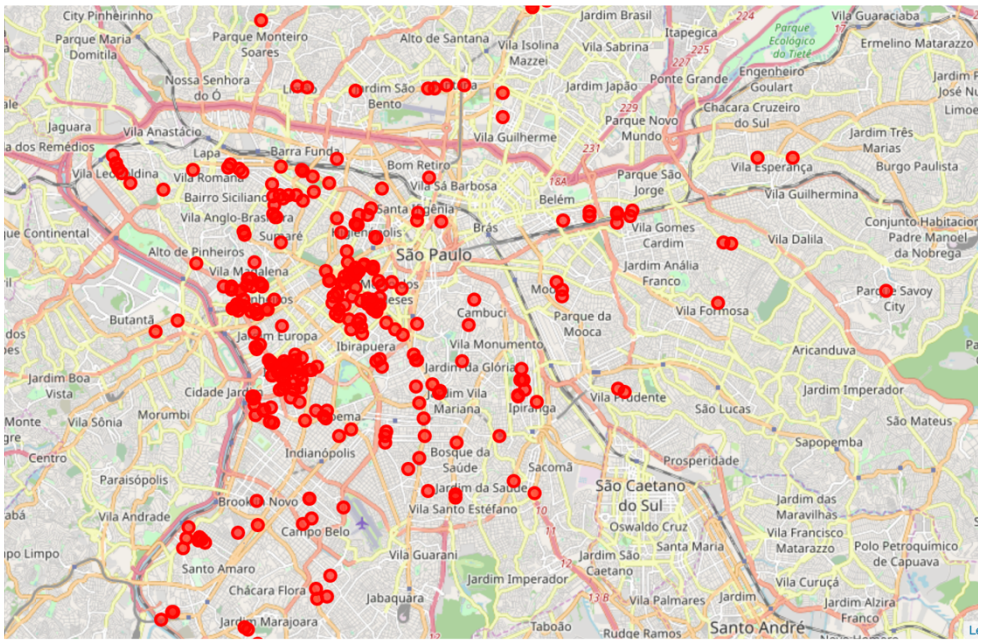

Figure 1- Map that shows in red all the burguer joints in São Paulo registered in the Foursquare API

With this in mind, since it is a very competitive market, the purpose of this study is to explore and provide insights for someone wanting to know more about the types of burger joints in São Paulo and to serve as a market research for someone that wants to open up a burger joint or to someone who already owns a venue and wants to understand a bit more about the market they are inserted on. To do this, some tools and data are going to be used such as the Foursquare API to obtain data about the venues and basic clustering algorithms such as k-means to try to segment these venues based on their characteristics.

Data Acquisition

This project had 2 steps of data acquisition on different kinds of data to get the results on the burger venues:

1. Getting location data
2. Getting venue data

Step 1 consisted in acquiring data about the location of the venues and the location of the neighborhoods in São Paulo. The first requirement was to get the neighborhood latitude and longitude information using the *geopy* library. After the neighborhood locations were retrieved, it was possible to get all the burger venues within 900 meter radius from each neighborhood using the Foursquare API. This radius was selected because larger radius would only cause overlap of venues and create more duplicates in the data set. After retrieving the venues, the API also returns the latitude and longitude of each venue, making it possible to map the neighborhood of each venue. Once there is information about the venue's name, id, neighborhood and location, it is possible to go to step 2 of the data acquisition phase.

Step 2 consists of retrieving specific information about each venue that will bring insights for future analysis. In the case of this study, 4 types of information were chosen that will help cluster the venues and also give insights about the types of customers that attend these venues:

- Venue Rating: A straightforward way to differentiate between the venues. We will retrieve the best venues by requesting the venue rating from the Foursqaure API. This data is a rating that varies from 0-10.

- Venue Price: Comma separated list of price points. Currently the valid range of price points are [1,2,3,4], 1 being the least expensive, 4 being the most expensive. 1 is < 10 USD an entree, 2 is 10 to 20 USD an entree, 3 is 20 to 30 USD an entree, 4 is > 30 USD an entree. This data is also within the json from the Foursquare API response.

- Venue Review: A text review of a customer that attended to a specific venue registered in the Foursquare platform. This will be used to give us more qualitative insights about what the customers value the most.

- Neighborhood meter squared price: This is the real estate price per meter squared of a specific neighborhood. This will help retrieve insights about not only the cost of maintaining a venue but also the type of customer that attends that venue if we assume that the higher the real estate price of a certain area, the higher the purchase power of a person that lives in that area. This feature is measured in Reais (R$).

Below there is a snippet of the dataset acquired:

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Id	Price_M2	Rating	Review	Price
Moema	-23.594585	-46.661801	Chicohamburger	-23.600686	-46.659306	4b30180ef964a520f6f524e3	10160	8.3	['Chico Buarque. Chico Anysio. Chico Mendes. C...	2
Moema	-23.594585	-46.661801	Tradi	-23.595413	-46.666676	56633f71498ec93dd3483489	10160	8.5	['Tão bom que comi dois! Oráculo e Ipiranga. Ó...	2
Moema	-23.594585	-46.661801	Bullguer	-23.597058	-46.667724	5506eeac498e2883360e7b7d	10160	8.3	['Já comeu no Shake Shack? Gostou? Tá com saud...	2
Moema	-23.594585	-46.661801	Stop Dog	-23.595320	-46.670190	4b4fb4e6f964a520d41127e3	10160	7.5	['Cheese Burguer Especial = Cheese Burger + Ma...	2
Moema	-23.594585	-46.661801	America	-23.597604	-46.667176	4b2b99e8f964a52013b824e3	10160	6.6	['Frozen LAJOTINHA!!! Sensacional!!!!!!', 'O s...	2
...	...	...	...	...	...	...	...	...	...	...
Itaim Bibi	-23.583656	-46.677918	Classic Burger Haüs	-23.583059	-46.676177	564bd2e3498e1cba92476299	11270	7.0	['Otima hamburger!!!', 'Burguer incrível! Pão ...	2
Itaim Bibi	-23.583656	-46.677918	Varanda Burguer	-23.582869	-46.669926	56bb905c498e886d33450790	11270	8.2	['Burguer fenomenal! Preço justo, atendimento ...	1
Itaim Bibi	-23.583656	-46.677918	Achapa	-23.580422	-46.674754	51c88f78498edb10ae45797d	11270	7.4	['senha do wifi: itaim198198', 'problema: vc v...	2
Itaim Bibi	-23.583656	-46.677918	TheDog Haüs	-23.583016	-46.676225	5058b9cae4b0700b20010a3b	11270	7.6	['Hot dog bom preparado e servido por gatos: g...	2
Itaim Bibi	-23.583656	-46.677918	Dona Deôla	-23.586195	-46.680645	5384815d498e59fb623fcc3d	11270	7.3	['Inauguração hj (15/09). A partir do dia 18/0...	2

Figure 2- São Paulo burger venues and their information dataset

Methodology

After acquiring the necessary data, 3 features were chosen to be inserted in the clustering algorithm: Venue Price, Venue Rating and Neighborhood meter squared price. The problem with 2 of those features are that they are ordinal data. This makes it very difficult for an unsupervised algorithm such as k-means to cluster the data because it clusters data based on euclidian distance and there is no distance between categorical data. But the data is also not purely categorical so it would not make sense to run a k-modes algorithm on the ordinal features. Running k-prototypes algorithm would be a solution for this problem with mixed data but assuming rating or price as purely categorical data would mean that all ordered information about the data would be lost. The final decision was to treat all the data as numeric and use the k-means clustering algorithm.

To use this clustering algorithm, it is important that the data is scaled and normalized so no feature prevails over the other while clustering. After preprocessing the data with sklearn's *StandardScaler*, the next step is to discover the optimal number of clusters. This was done by using the elbow method and the sum of squares as the function to minimize. The optimal number of clusters were found to be 4.

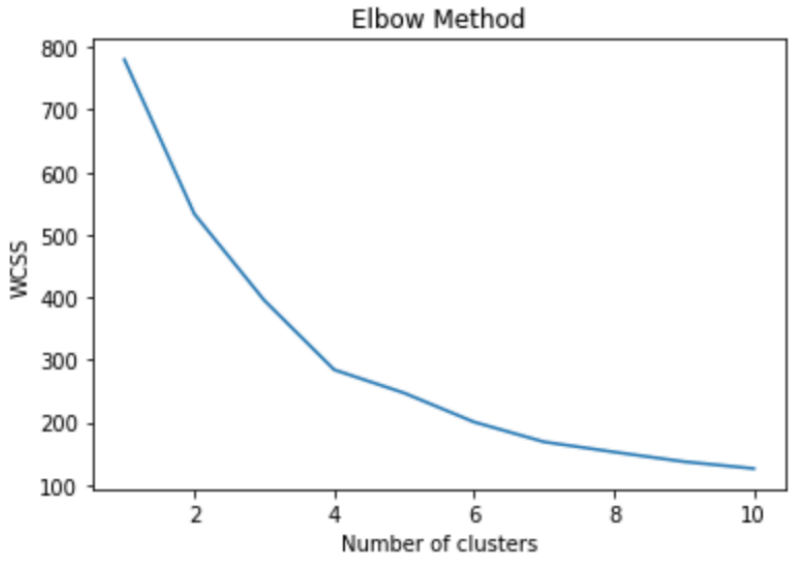

Figure 3- Elbow method graph shows an elbow at 4 clusters

Once the venues are clustered, there is room for analysis of each feature within a cluster and categorize the clusters. The other analysis will be made using the reviews obtained by each venue. The method will be to extract all the reviews of the venues of each cluster and build a word cloud of each cluster and then analyze what the customers of the venues of these clusters are paying more attention to based on the frequency of the words in the reviews.

Resulsts

After choosing 4 as the number of clusters, the k-means algorithm returned the clusters for each venue. A map of the burger joints in São Paulo color coded by cluster is shown below. The red cluster is cluster 0, the purple cluster is cluster 1, cluster 2 is color coded in blue and cluster 3 is show in beige:

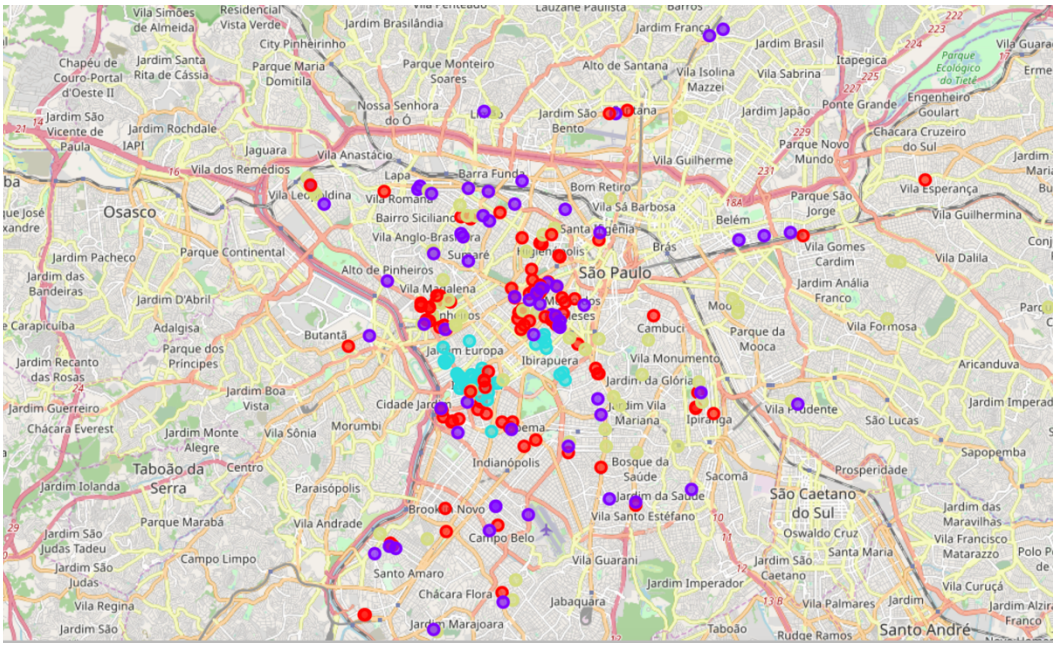

Figure 4- Clustered map of São Paulo burger joints

To give a more general vision how the algorithm clustered the venues based on each feature, there is a graph below that portrays that view:

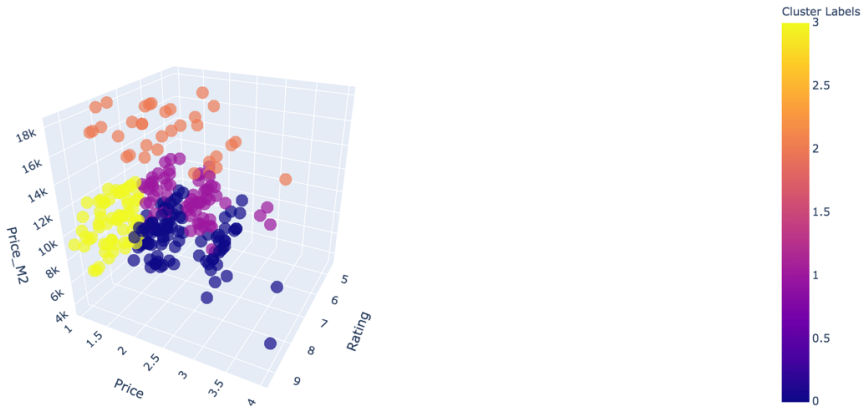

Figure 5- 3D view of the clustering based on the chosen features

The 3D graph may bring some insights on the division of the clusters but to get a clear detail between the clusters there are 3 graphs that portrays the mean of each feature of each cluster shown below. The first one is a graph showing the mean price of each cluster, the second portrays the mean real estate price of each cluster and the third graph shows the mean rating of each cluster:

Figure 6- Mean price of each cluster

The graph above shows cluster 0 and 2 being the most expensive, having entrees ranging from 10-20 USD. If we assume a linear relationship between price rating and real price, Cluster 0 is on average, approximately 17% more expensive than cluster 2 which are the second most expensive venues.

This draws a clear divide between cluster 0 and 2 and cluster 1 and 3. Cluster 0 and 2 are the most expensive venues and clusters 1 and 3 are the cheapest venues, with special attention to cluster 3, which is less than half of the average price of cluster 0. This means that the average price of the entrees of the venues of cluster 3 are less than 10 dollars.

Figure 7- Mean real estate price of each cluster

The graph above shows that the one of the main clustering feature that separated cluster 2 from the rest was real estate price. Its average price is 75% higher than the second most expensive neighborhood cluster, which is cluster 0.

With this in mind, it is expected that the people that attend cluster 2 venues have more purchasing power when compared to other clusters.

Figure 8- Mean rating of each cluster

Almost all of the clusters had no relevant percentage difference on rating, all of them ranging from 7.6 to 8.1 on average. It's assumable that the venues of these clusters are pretty satisfied customers.

On the other hand, we observe another clear division between clusters on cluster 1. It's average rating was 5.86, 23% less than the second lowest rated cluster.

After the quantitative analysis of each cluster, the reviews of each cluster were used to create a word cloud of each cluster. This will not bring insights about what is particularly best in each cluster but rather will show what aspects the customers from each cluster talk about the most. Since this project is dealing with a country that speaks Portuguese, it will be hard for someone that doesn't understand the language to retain information from the word cloud as a whole. For this reason, the key points generated in this word cloud will be explained and properly translated. Below are the four word clouds of each cluster:

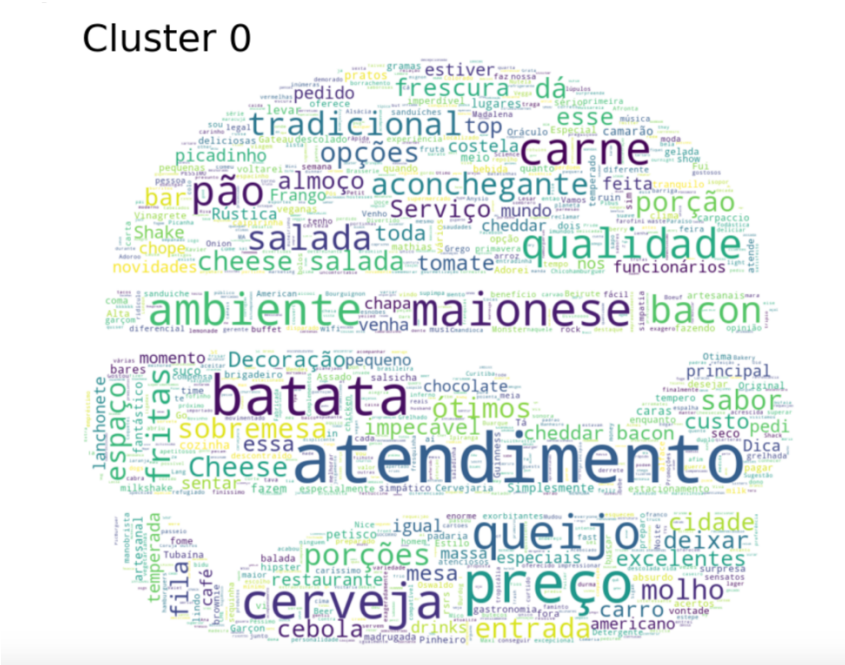

Figure 9- Word cloud of cluster 0

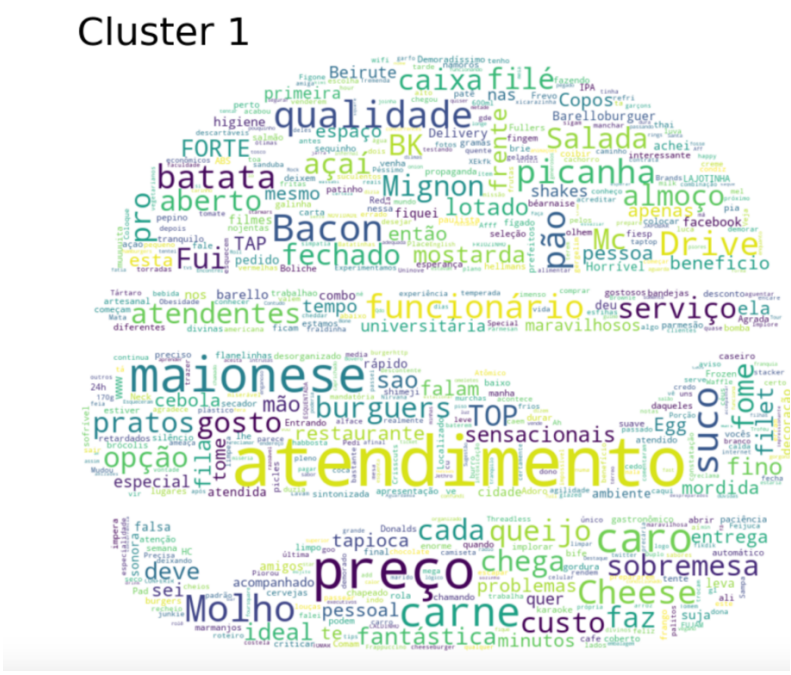

Figure 10- Word cloud of cluster 1

Figure 11- Word cloud of cluster 2

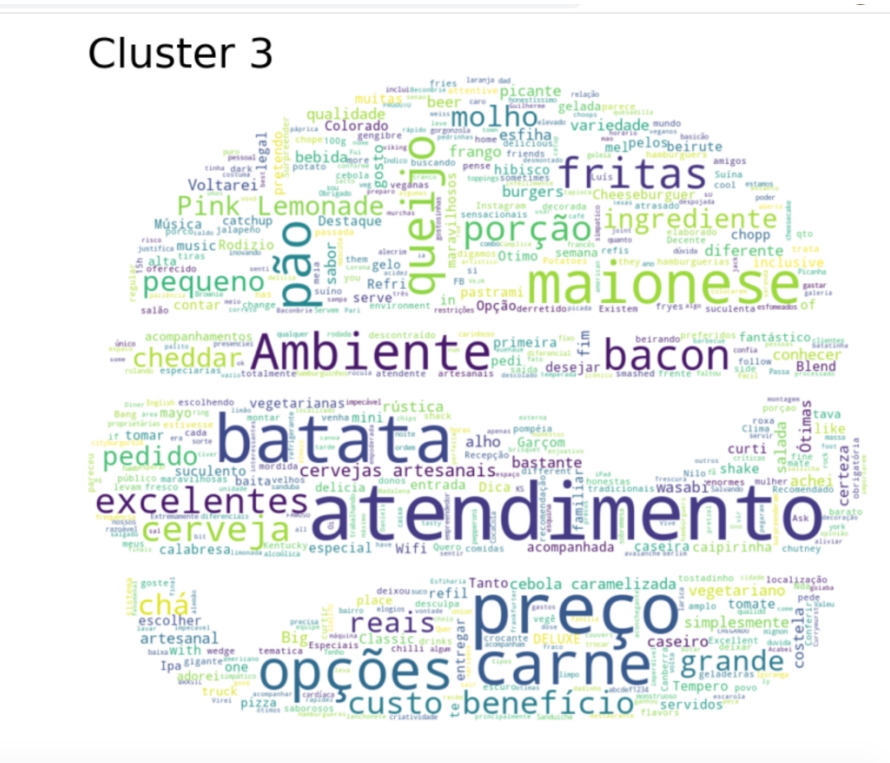

Figure 12- Word cloud of cluster 3

Taking a look at the top 10 words ordered by frequency of each cluster we have:

Cluster 0		Cluster 1		Cluster 2		Cluster 3	
1	service	1	service	1	mayo	1	service
2	price	2	price	2	fries	2	price
3	fries	3	mayo	3	service	3	fries
4	beer	4	meat	4	wifi	4	meat
5	cheese	5	expensive	5	coffee	5	mayo
6	mayo	6	sauce	6	waiting	6	options
7	meat	7	bacon	7	vegetarian	7	environment
8	environment	8	juice	8	cheese	8	bread
9	bread	9	quality	9	brie	9	cheese
10	quality	10	picanha	10	environment	10	fries

Table 1- Top 10 words from the 4 clusters

Taking a look at everything they have in common within their top ranked words, these could be the must-haves for someone looking to open a burguer joint. All of them have the word 'atendimento' which means 'service' in portuguese. So first key takeway from this

information is that no matter the customer segment, price, or location, the venue has to have a good service, this word is within the top 3 words in all clusters. Another word that appeared in all clusters is 'maionese' which is equivalent to mayonnese in English. This shows that it is another must-have in all burger joints.

Now, taking a look at the differences between clusters, we can see that all of them but cluster 2 have the word 'preço', which means 'price' in portuguese. This means that cluster 2 probably doesn't show too much interest in evaluating price as a defining category when visiting a burger joint. This makes sense based on the price per meter squared because cluster 2 was categorized as having the richest customers. Another interesting fact is that cluster 2 refers to the word vegetarian ('vegetariano' in Portuguese) and brie (as for the cheese), showing that they are also very demanding.

Discussion

All of this information gathered about the venues and analysis on them gave us a plethora of insights. Although unsupervised algorithms are not always reliable, they proved to reveal some insights and help not only segment the venues but also take a deep look into the customers of these venues. From the clusters per se, the key take-aways are:

- Cluster 0: High priced venues, on neighborhoods that are average priced and have high ratings. The customers talk about the price, service and another feature that was not present in any other category which is beer. This shows us that the people that attend

these high priced venues don't have as much purchasing power as cluster 2 but when they choose to spend, money, it is probably on leisure time (hence beer on top words) and they value this type of experience (hence higher ratings).

- Cluster 1: Average priced (compared to others), low priced neighborhoods and low ratings. This cluster is highlighted by dissatisfaction not only in the ratings but in the top words where price and expensive are among the top 10. This could be explained by the fact that people within this cluster don't have much purchasing power and when they try to spend a little bit more on average priced venues, they usually don't reach their expectations. Points that a cluster 1 venue should look out for are probably cost benefit related, giving people expectations that are not really met.

- Cluster 2: High priced venues and neighborhood locations and high ratings. This cluster is defined by high standards and and differentiation. One explanation for the rating being lower than clusters 0 and 3 are that people that have higher purschasing power and attend high priced venues elevate their expectations, hence, they will have less mercy when submitting a rating. Based on the top words, we can see that the customers of these venues pay attention to differentiated foods and condiments and place mayo as the top word. A clear point that cluster 2 venues should look out for are clearly to place a priority on uniqueness within their venue's menu, with less regards when it comes to pricing.

- Cluster 3: Low priced venues in low priced neighborhoods with high ratings. This cluster is about having lower prices and getting what you paid for. One aspect that was not discussed but is a word that appears on the 15th place on the word cloud is cost-benefit.

This already shows us that the customers are getting what they are looking for: good price for a good meal. The price frequency on the cluster 3 word cloud was greater than all the others and from there the top words highlight the basic aspects of a burger joint which is fries, meat, mayo and lots of options. The key take-away for cluster 3 venues is: keep it simple and cheap.

Conclusion

This closes up the study on burger joints in São Paulo. There are tons of data that could make this study more reliable statistically and probably a supervised algorithm would be a good choice if we could label the venues. With that said, it is a good high level analysis of the burger joint market in São Paulo for people that are thinking of opening up a burger joint or for people that are curious about how data science and machine learning can help retrieve insights and learn new things about almost every topic.