Select a borough to set up a productive business in Colombia

APPLIED DATA SCIENCE

Capstone Project

The Battle of Neighborhoods

Sept. 2019

Sandra Lucía Rodríguez Ochoa

TABLE OF CONTENTS

1 INTRODUCTION

1.1 Background

Colombia, my country, is growing in terms of production, industry, agriculture and business; so, more and more foreign companies are betting on investing capital in these lands. Globalization has allowed shortening borders and establishing a diverse trading.

The vast mountainous territory that Colombia has, along with the climatic variations of the different vertical zones, allows for the production of an unusually represented the backbone of the Colombian economy, bringing premium prices on the world market and constituting about half of all exports.

But the big question for a foreign investor would be, in which part of the Colombian territory can set up his business, and invests his capital?, what kind of business could be more productive and could let more profit?. In which type of products are interested the developed countries to require exports from Colombia?

1.2 Business Problem

For a foreign investor that want to set up business in Colombia what kind of product commercialize with the objective of export to other countries, keeping in mind the Colombian economy and the wide range of climatic variations?

Where the investor could set up his office? In addition, how similar could result the selected neighborhood in Colombia in comparison with the neighborhood he was used to living.

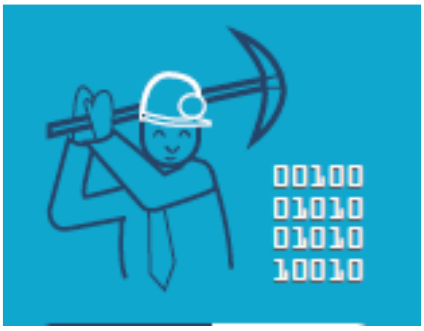
1.3 Interest

- ✓ Investors looking for projects to fund, and interested in Colombian diversity.
- ✓ Someone who wants to identify opportunities in Colombian lands and want to know a little about Colombian production, business and exports before making any investment decision.
- ✓ New foreign investors who want to establish their headquarters in Colombia and are looking for a place to live similar to their place of origin

2 DATA

2.1 Data sources

MARO (https://www.maro.com.co/) is a consultation platform with updated information on the dynamics and situation of the economic sectors in Colombia. Trough MARO a data base in a XLMSX format could be downloaded, this data base contains powerful information about exports. For each department in Colombia it is discriminated the kind of product exported, the amount exported in weigh (Kg) and in dollars and the country destiny of exportation. This is a huge data set to explore, clean, organize and extract the information to present to our investor.



As Data Scientist, these data will be used to analyze the Colombian exports, identify the most productive regions, recognize the start products and branches of the production that are sought as exports; and finally, know the countries with more demand of products from Colombia.

When the region with the best business opportunity is identified, a database of the DANE (National Administrative Department of Statistics) will be used to identify all the boroughs in the selected region, and with support of Argis API, the coordinate system will be found for each borough.

Then the Foursquare location data will be used to explore the boroughs, review the popular venues and identify similarities with neighborhoods in other countries.

2.2 Data wrangling and cleaning

In order to make data easy to use, the first step is understanding our data bases, sort the relevant data from the least necessary data, establish keys for each tables and identify the interrelation between them.

Data wrangling brings out the best features of data; so for our purpose, we identify de “Department” column as the key. We are only interested in exports values in dollars for the last three years, so we can drop the columns with the exports values in weigh and it is also possible to drop columns from the 2010 trough 2015 years.

For each row in the data set we verify that values were complete and meaningful, it was corroborated that all the values were consistent and that in the data set there were not NaN nor special characters replacing values.

It was also created a new column in the data set called “Total” that consist in the sum of all the exports values for the 2016, 2017, 2018 and the months elapsed in this year.

It was also necessary to fix some character strings and convert to capital letters and remove accents to allow the match between the “Department” in Exports data base and the name in the Json file to get the choropleth map.

2.3 Feature selection:

The following table describes the content of each table

3 METHODOLOGY

The main database of this study was downloaded from MARO (Opportunities Regional Map) web site, where there is available information about the economics sector in Colombia, my country.

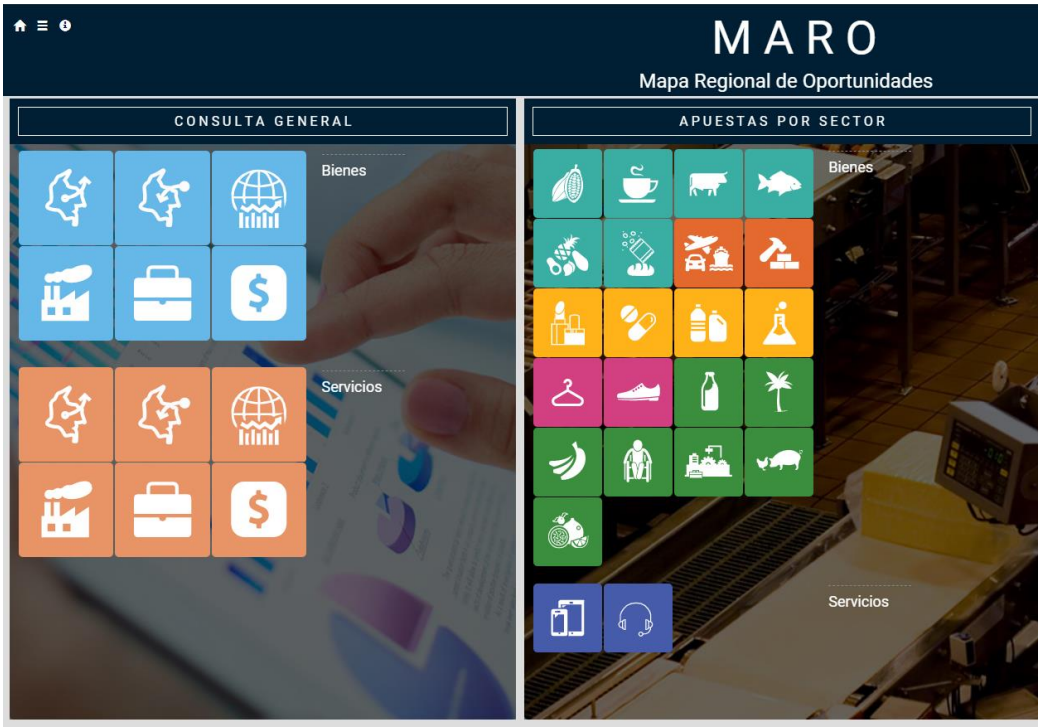


Figure 1: MARO, consultation platform

General Colombian exports data was the information considered appropriate for the analysis. It is a XLSX file with 168700 rows and 28 Columns. The following table shows the name of each column in the table and its content

INDEX	Column Name	Description
0	NANDINA	ID Product subcategory
1	Descripción NANDINA	Product subcategory description
2	CIIU Rev.4	ID Product Category
3	Descripción CIIU Rev.4	Product Category Description
4	Destino	Country Export destination Country
5	Departamento	Department in colombia of export origin
6	FOBDO 2010	Value exported in dollars for each year
7	FOBDO 2011	
8	FOBDO 2012	
9	FOBDO 2013	
10	FOBDO 2014	
11	FOBDO 2015	
12	FOBDO 2016	
13	FOBDO 2017	
14	FOBDO 2018	
15	FOBDO ENE-JUN 2018	
16	FOBDO ENE-JUN 2019	
17	KNETO 2010	Value exported in net kilogram
18	KNETO 2011	
19	KNETO 2012	
20	KNETO 2013	
21	KNETO 2014	
22	KNETO 2015	
23	KNETO 2016	
24	KNETO 2017	
25	KNETO 2018	
26	KNETO ENE-JUN 2018	
27	KNETO ENE-JUN 2019	

Table 1: Colombian exports Database, column description

[65]:

	NANDINA	Descripción NANDINA	CIIU Rev.4	Descripción CIIU Rev.4	Destino	Departamento	FOBDO 2010	FOBDO 2011	FOBDO 2012	FOBDO 2013	...	KNETO 2011	KNETO 2012	KNETO 2013	KNETO 2014	KNETO 2015	KNETO 2016	KNETO 2017	KNETO 2018	KNETO ENE-JUN 2018	KNETO ENE-JUN 2019
0	1905100000	Pan crujiente llamado "knackebrot".	1081	Elaboración de productos de panadería	Estados Unidos	Antioquia	0.0	961.64	4562.42	324.42	...	242.7	1411.37	71.0	0.00	0.00	0.00	0.00	0.00	0.00	0.0
1	1905100000	Pan crujiente llamado "knackebrot".	1081	Elaboración de productos de panadería	Panamá	Antioquia	0.0	0.00	0.00	0.00	...	0.0	0.00	0.0	0.00	0.00	0.00	0.00	20011.62	5867.67	20291.0
2	1905100000	Pan crujiente llamado "knackebrot".	1081	Elaboración de productos de panadería	Chile	Valle del Cauca	0.0	0.00	0.00	0.00	...	0.0	0.00	0.0	1351.97	5957.11	0.00	0.00	0.00	0.00	0.0
3	1905100000	Pan crujiente llamado "knackebrot".	1081	Elaboración de productos de panadería	Reino Unido	Valle del Cauca	0.0	0.00	0.00	10863.60	...	0.0	0.00	1440.0	555.60	0.00	0.00	0.00	0.00	0.00	0.0
4	1905100000	Pan crujiente llamado "knackebrot".	1081	Elaboración de productos de panadería	Aruba	Bogotá, D.C.	0.0	0.00	0.00	0.00	...	0.0	0.00	0.0	0.00	0.00	94.19	8.18	0.00	0.00	0.0

Figure 2: Colombian exports Database into Pandas Data Frame

With this large volume of data, it is possible to perform different types of analysis trying to focus in solve our problem: identify the better option for a foreign capital investment. The following sections show data statistics and analysis that can support making the right decision.

3.1 Exports by department.

The first analysis of the exports data, was oriented around regions; the data base was grouped by region and all export value in dollars were added, regardless the category of exported product or the destination of the export. Two columns were added the first one with the total export value in the last 6 years (since 2014 up to date) and the second one with the average of exports in the same period. This group resulted in a new data frame with 31 rows and 9 columns.

Exports by region has 31 Rows and, 9 columns

	Region	2014	2015	2016	2017	2018	2019	Total	Average
0	Antioquia	447.059325	576.554927	570.775513	613.578035	598.856402	336.743726	3143.567928	523.927988
1	Caldas	514.785876	486.371843	470.123975	641.742743	666.141690	304.714387	3083.880515	513.980086
2	Huila	423.530769	427.367289	404.387033	420.240562	406.720305	190.544706	2272.790665	378.798444
3	Risaralda	441.926184	419.326413	371.881648	318.681931	235.384163	143.385564	1930.585901	321.764317
4	Quindío	281.722253	309.061594	258.519142	241.316915	254.569640	127.360940	1472.550484	245.425081

Figure 3: Colombian exports Database by region Top 5 (Millions of dollars)

After sorting by total exports, the Figure 4 shows the top 12 Regions in Colombia with the highest export value, having each year as a series.

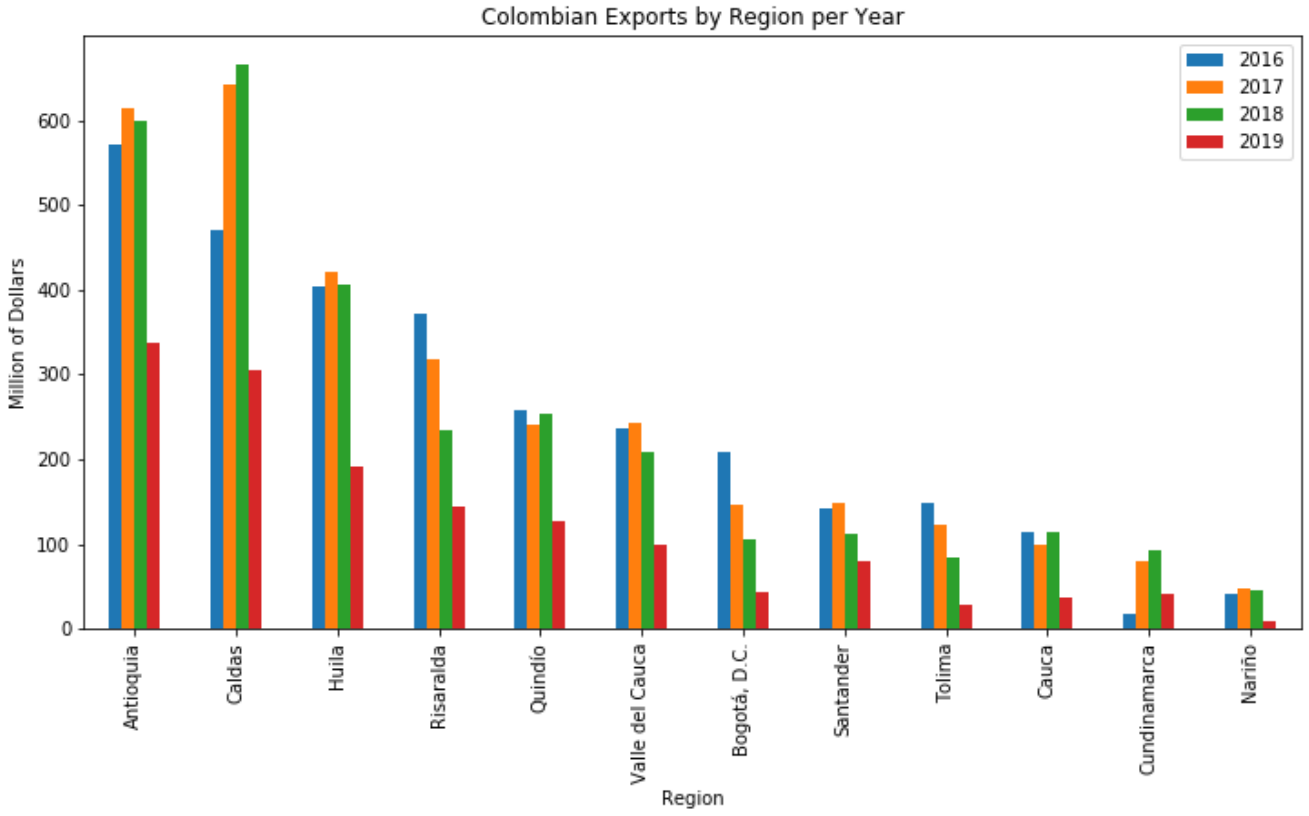


Figure 4: Top 12 Colombian exports by region per year (Millions of dollars)

As is evident, the departments of Antioquia and Caldas are the regions leaders in exports. Let´s see what happen in terms of growth, which departments show the largest percentage growth between years?. To identify that, a new data frame was created calculating the percentage increase or decrease in exports between years: 2014-2015, 2015-2016, 2016-2017, 2017-2018 and then the average growth.

	Region	2014-2015	2015-2016	2016-2017	2017-2018	Average
0	Cundinamarca	31.337458	-13.331624	138.710173	29.809296	46.631326
1	Boyacá	0.000000	2.143277	433.493868	-273.417566	40.554895
2	Meta	0.000000	40.135241	4.976893	110.458957	38.892773
3	Córdoba	27.321792	14.460533	98.195294	13.819262	38.449220
4	Bolívar	70.761086	-127.789570	102.224966	78.521597	30.929520

Figure 5: Colombian exports percentage growth between years Top 5 Data Frame

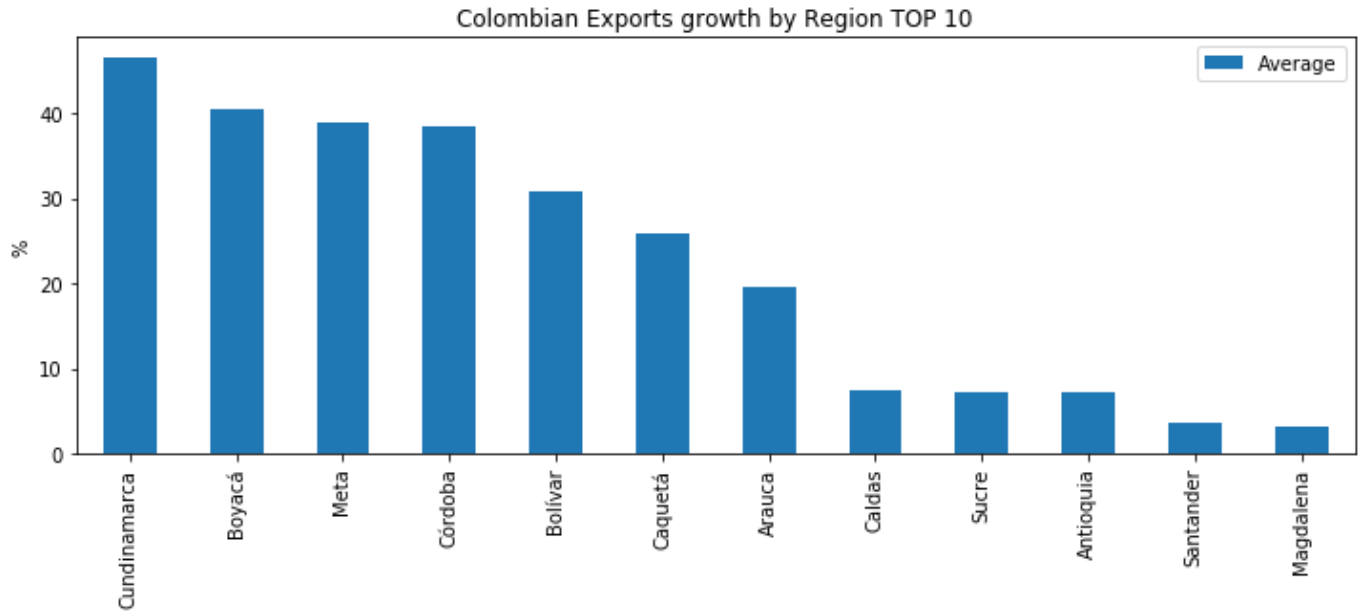


Figure 6: TOP 10 Colombian exports percentage growth between years

Let´s check regions with a negative growth, the departments with the largest decrease.

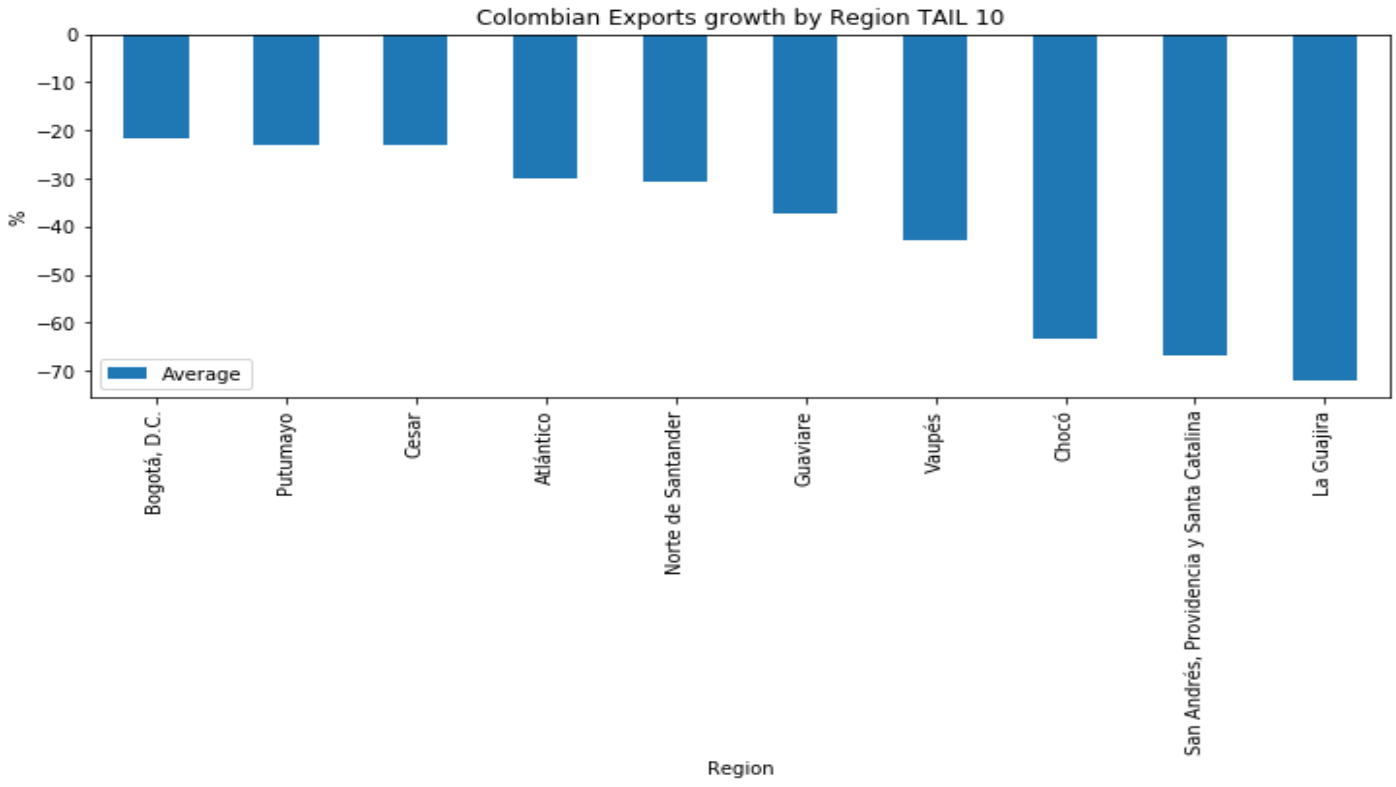


Figure 6: Tail 10 Colombian exports percentage growth between years

3.2 Exports by product category.

In the same way, the second analysis of the exports data, was oriented around product category, to identify the kind of products with the best opportunities to invest in Colombia.

The data base was grouped by product category and all export value in dollars were added, regardless the region of origin or the destination of the export. As in the previous section, two columns were added Total export value (since 2014 up to date) and the average of exports in the same period. This group resulted in a new data frame with 12 rows and 9 columns.

Exports by Product Category has 12 Rows and, 9 columns

	Category	2014	2015	2016	2017	2018	2019	Total	Average
0	Trilla de café	2473.260563	2526.470898	2379.267592	2513.090734	2267.459840	1081.603544	13241.153172	2206.858862
1	Otros derivados del café	221.299644	233.621727	220.016536	224.865705	233.153149	116.476565	1249.433326	208.238888
2	Elaboración de productos de panadería	133.272579	97.685515	92.212480	100.973334	95.496111	42.503701	562.143720	93.690620
3	Elaboración de cacao, chocolate y prod de conf...	121.515210	87.200311	105.070720	84.871409	91.606604	39.058075	529.322330	88.220388
4	Descafeinado, tostión y molienda del café	43.433770	49.981457	44.801176	68.495834	67.767740	51.836802	326.316779	54.386130

Figure 7: Colombian exports Database by Product Category Top 5 (Millions of dollars)

After sorting by total exports, the Figure 8 shows the top 5 product categories in Colombia with the highest export value, having each year as a series.

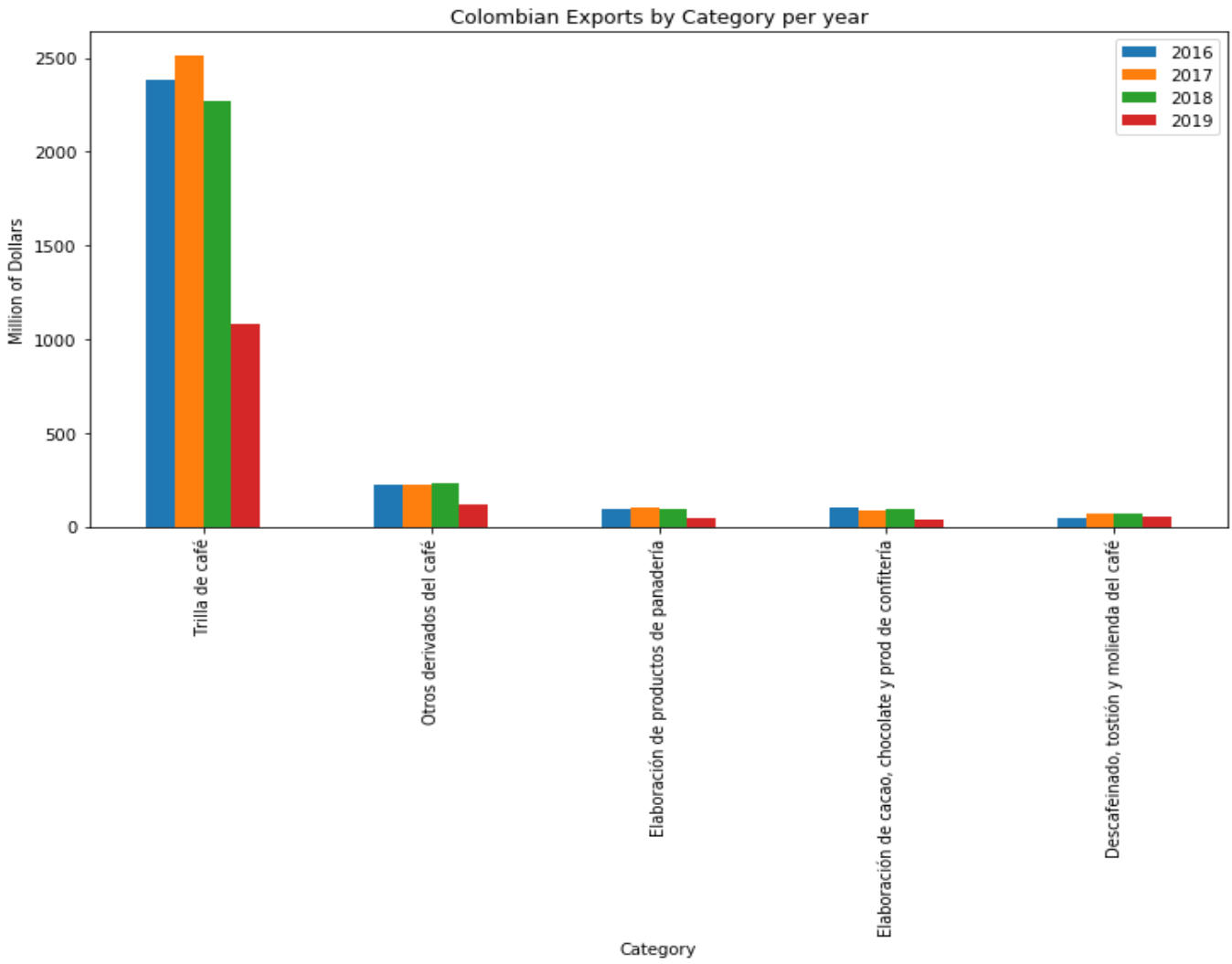


Figure 8: Top 5 Colombian exports by category per year (Millions of dollars)

By a wide margin, coffee and its derivatives are the star export product in Colombia. Let´s see what happen in terms of growth, which product categories shows the largest percentage growth between years?. To identify that, a new data frame was created calculating the percentage increase or decrease in exports between years: 2014-2015, 2015-2016, 2016-2017, 2017-2018 and then the average growth.

Category

	Category	2014-2015	2015-2016	2016-2017	2017-2018	Average
0	Cultivo de frutas tropicales y subtropicales	18.620080	70.259075	51.366242	12.688275	38.233418
1	Procesamiento y conservación de frutas, legumb...	15.632954	13.455669	20.246361	14.825145	16.040032
2	Procesamiento y conservación de carne y produc...	-23.442304	2.871297	67.222330	15.159912	15.452809
3	Descafeinado, tostión y molienda del café	12.039259	-9.525003	43.567465	-1.338751	11.185743
4	Otros derivados del café	5.917282	-6.533453	2.328657	3.979777	1.423066

Figure 9: Colombian exports percentage growth between years by category – Top 5 Data Frame

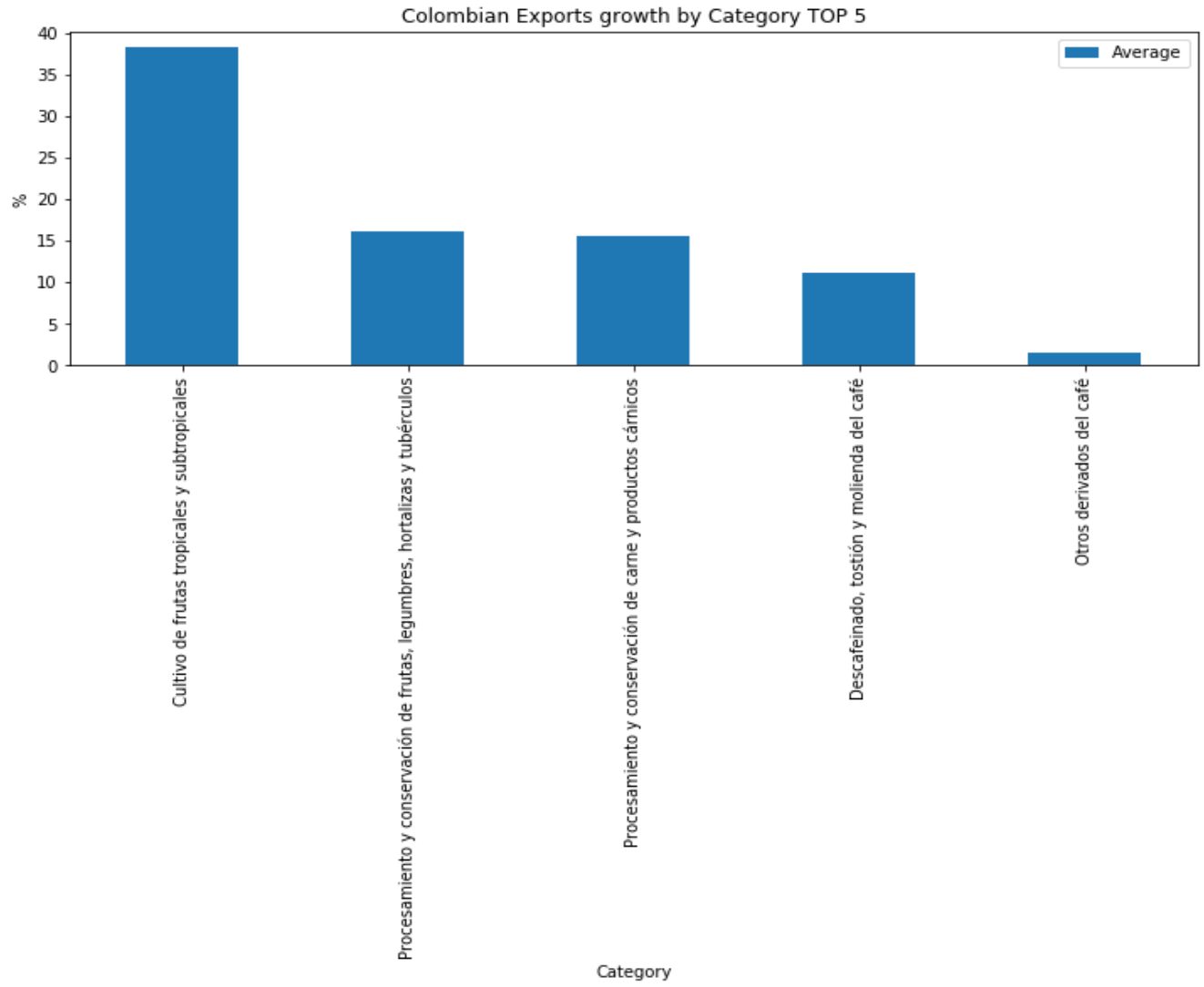


Figure 10: TOP 5 Colombian exports percentage growth between years by category

Let´s check products with a negative % of growth, the categories with the largest decrease.

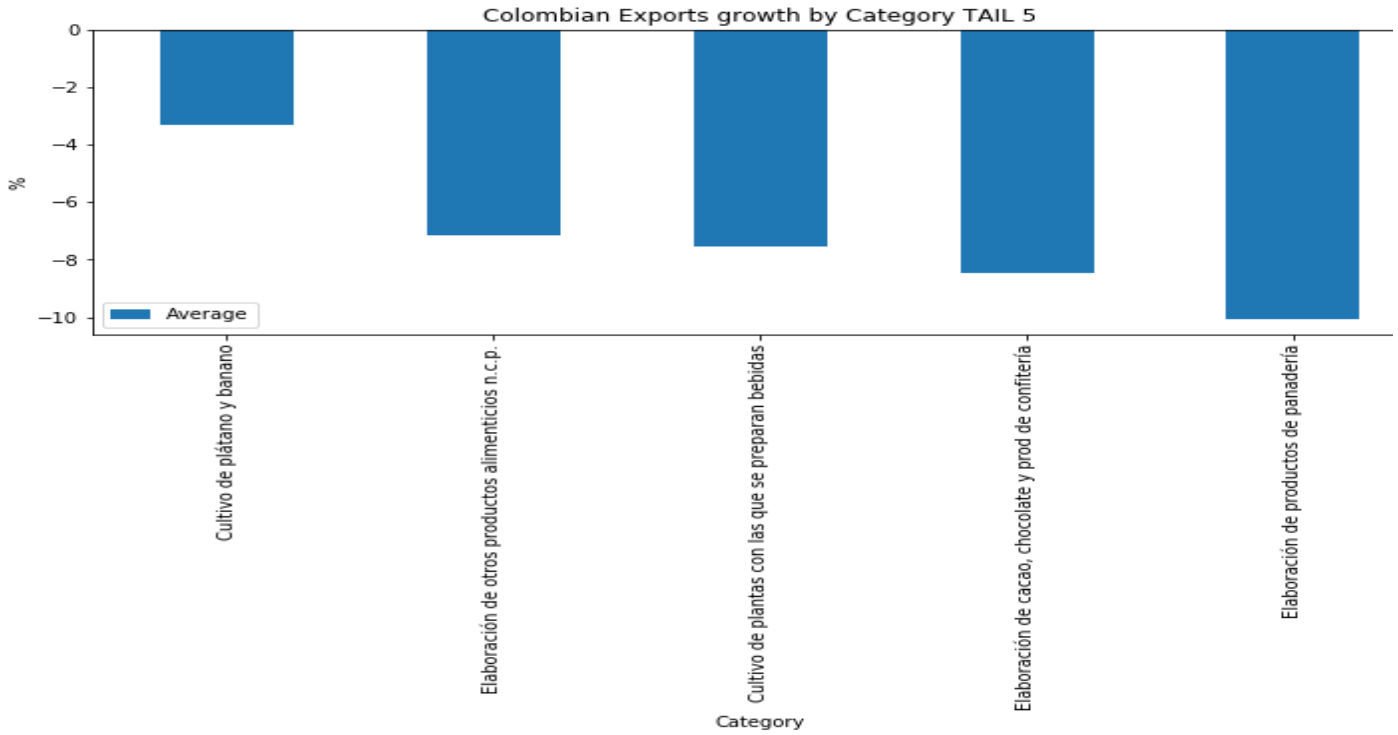


Figure 11: Tail 5 Colombian exports percentage growth between years by product category

3.3 Exports by destination.

The last analysis of the exports data, was oriented around destination, to identify the countries with the greatest interest in acquiring Colombian products.

The data base was grouped by destination and all export value in dollars were added, regardless the region of origin or the product category. As in the previous section, two columns were added Total export value (since 2014 up to date) and the average of exports in the same period. This group resulted in a new data frame with 12 rows and 9 columns.

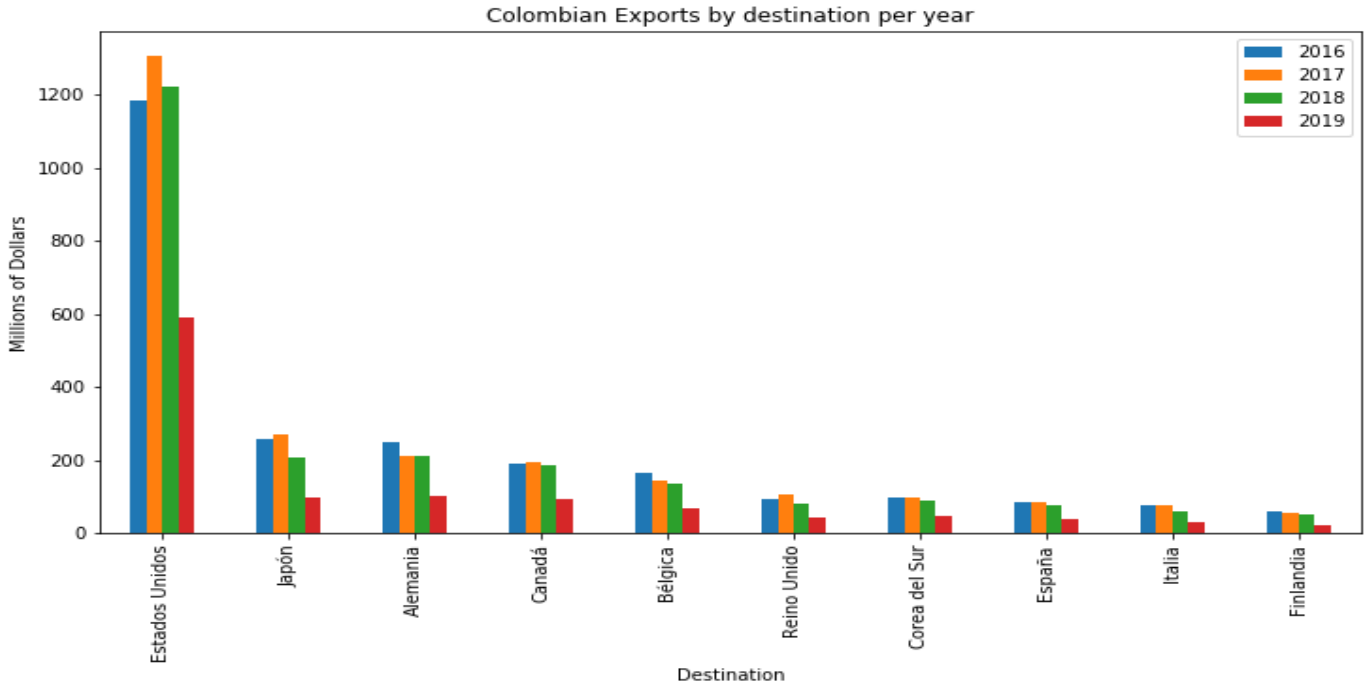


Figure 12: Top 10 destinations of Colombian exports per year (Millions of dollars)

due to the proximity and current trade agreements, United States is the top export destination of Colombia with about $11.1B of dollars. Japon, Germany and Canada follow the list.

3.4 Explore Region with best opportunities.

Once the region with the best opportunities to invest in Colombia has been identified, another data base will be included in our analysis. This new Database will contain the lists of boroughs located in this department and they will be passed as argument to a function that returns the coordinates of each borough using argi API.

Cluster boroughs according with popular venues

Then, by means of Python folium library, it could be visualized geographical details of the department and its boroughs.

As final step, Foursquare API will be used to explore the popular venues 750 m around the borough and with a limit of maximum 100 venues per borough. With those results we can analyze which borough has the grater quantity of popular venues and which kind of venues exist per borough.

Using unsupervised learning K-means algorithm, boroughs will be grouped into 5 clusters according with the venues category that exist in each borough.

3.5 Select the borough with more similarities

Assuming our investor used to live in Newhan, London, using the recommender matrix, we will calculate the Euclidean distance between Newham and each borough in our database, and then, select the borough with the shortest distance to set up the business there.

4 RESULTS

4.1 Find the region with better opportunities

The first objective of this analysis consist in identify the department in Colombia with the best opportunities to invest and set a business to export products to other countries. Using the Colombian exports database downloaded from MARO and matching it with the Colombia JSON file; this region was identified, through a choropleth map:

[34]:

	Region	2017	2018	2019	Total
0	ANTIOQUIA	613.578035	598.856402	336.743726	1549.178164
1	ARAUCA	0.140239	0.128348	0.164041	0.432628
2	ATLANTICO	6.056956	2.230487	1.542750	9.830194
3	SANTAFE DE BOGOTA D.C	146.757452	106.787289	43.160371	296.705112
4	BOLIVAR	2.631158	4.348195	0.912072	7.891425
5	BOYACA	0.053314	0.019853	0.000000	0.073167
6	CALDAS	641.742743	666.141690	304.714387	1612.598820
7	CAQUETA	0.228826	0.048947	0.000000	0.277773
8	CASANARE	0.001371	0.000000	0.002100	0.003471
9	CAUCA	98.766098	113.592569	37.102860	249.461527

Figure 13: Colombian regions and they exports in millions of dollars (First ten rows of Data frame)

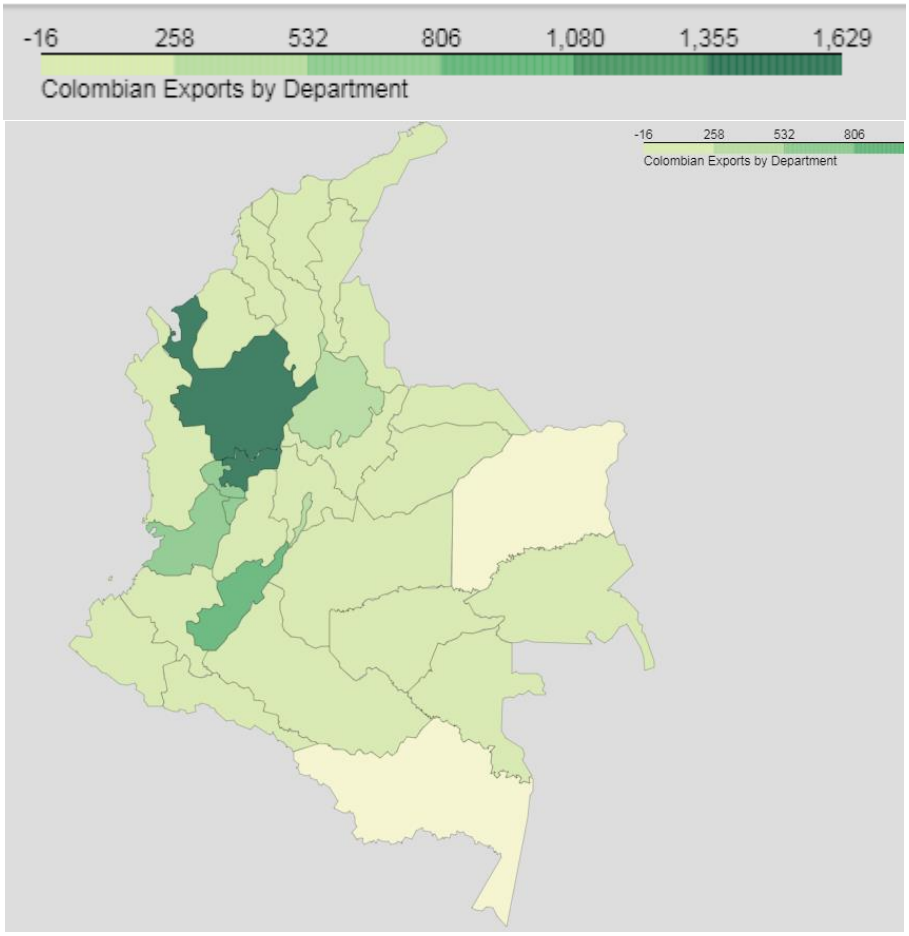


Figure 14: Choropleth map export value in millions of dollars per region (2017-2019)

As we can see the darkest regions are those that in the period 2017 – 2019 had a greater value of the exports in millions of dollars:

1. Antioquia
2. Caldas
3. Huila
4. Risaralda
5. Quindio

Those regions are located in the Andean lands and are listed as the coffee axis of Colombia

4.2 Antioquia's Market

Having Antioquia as the best opportunity for an investor, with the Colombian exports database we could explore more from Antioquia to select a proper market.

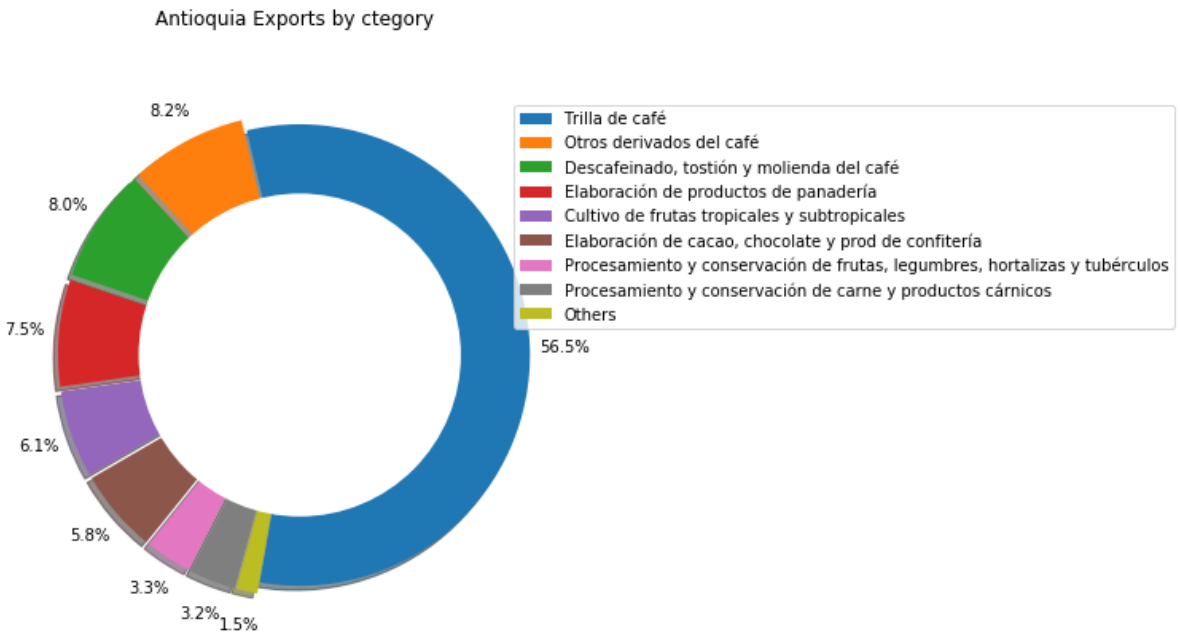


Figure 15: Product category exported from Antioquia

As we can see, the raw coffee market is covers more than half of exports.

In the same way, as figure 16 shows, United States is the favorite destination for Antioquia´s exports followed by Canada with a 5.3% and Japon with 5.1%

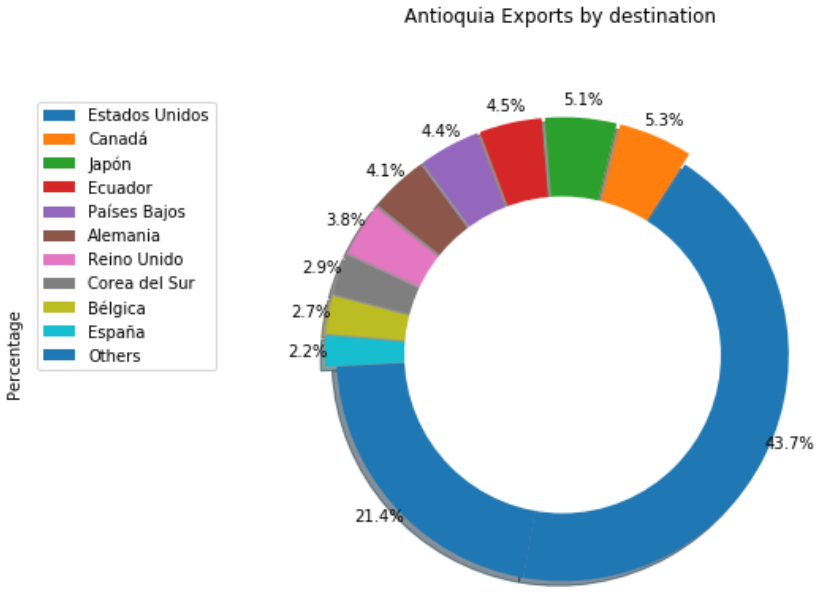


Figure 16: Antioquia's exports destinations

4.3 Cluster venues in Antioquia's boroughs

With a database that contains the list of boroughs in Antioquia, we can build a Data Frame that include the coordinate system of each borough.

	Código región	Código zona	Código de municipio	Municipios	Código municipal (1)	Zonas	Código zonal (3)	Latitude	Longitude
0	SR01	Z01	5001.0	Medellín	5001.0	Centro	1.0	6.24589	-75.57457
1	SR01	Z02	5079.0	Barbosa	5079.0	Norte	2.0	6.43624	-75.33157
2	SR01	Z02	5088.0	Girardota	5308.0	Norte	2.0	6.37735	-75.44398
3	SR01	Z02	5212.0	Bello	5088.0	Norte	2.0	6.34034	-75.55916
4	SR01	Z02	5308.0	Copacabana	5212.0	Norte	2.0	6.34641	-75.50800

Figure 17: List of Antioquia's boroughs including coordinate (first 5 rows in Data Frame)

Using the Foursquare location data, a new Data Frame was build including 402 veneus located in Antiquia's boroughs. The figure 17 shows the amount of popular venues that were found by borough

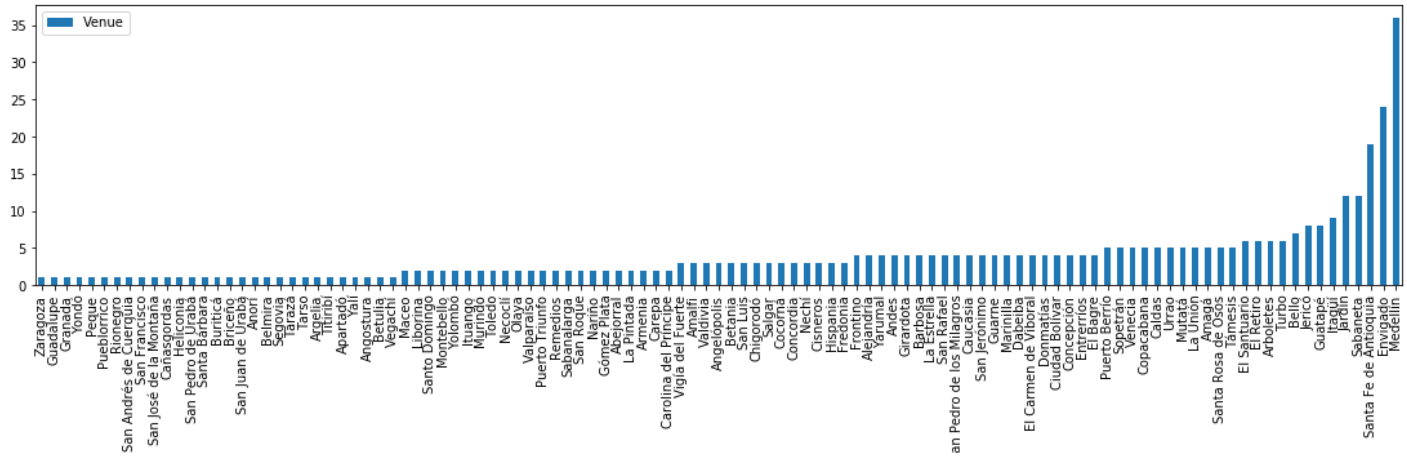


Figure 17: Number of venues by borough

To analyze each borough in Antioquia a matrix was built having as columns the venue category and as rows the boroughs. The matrix was grouped by borough to identify the amount of venues in each category. Then, with the normalized matrix, the unsupervised k-means was applied to cluster the boroughs into 5 clusters as Follows:

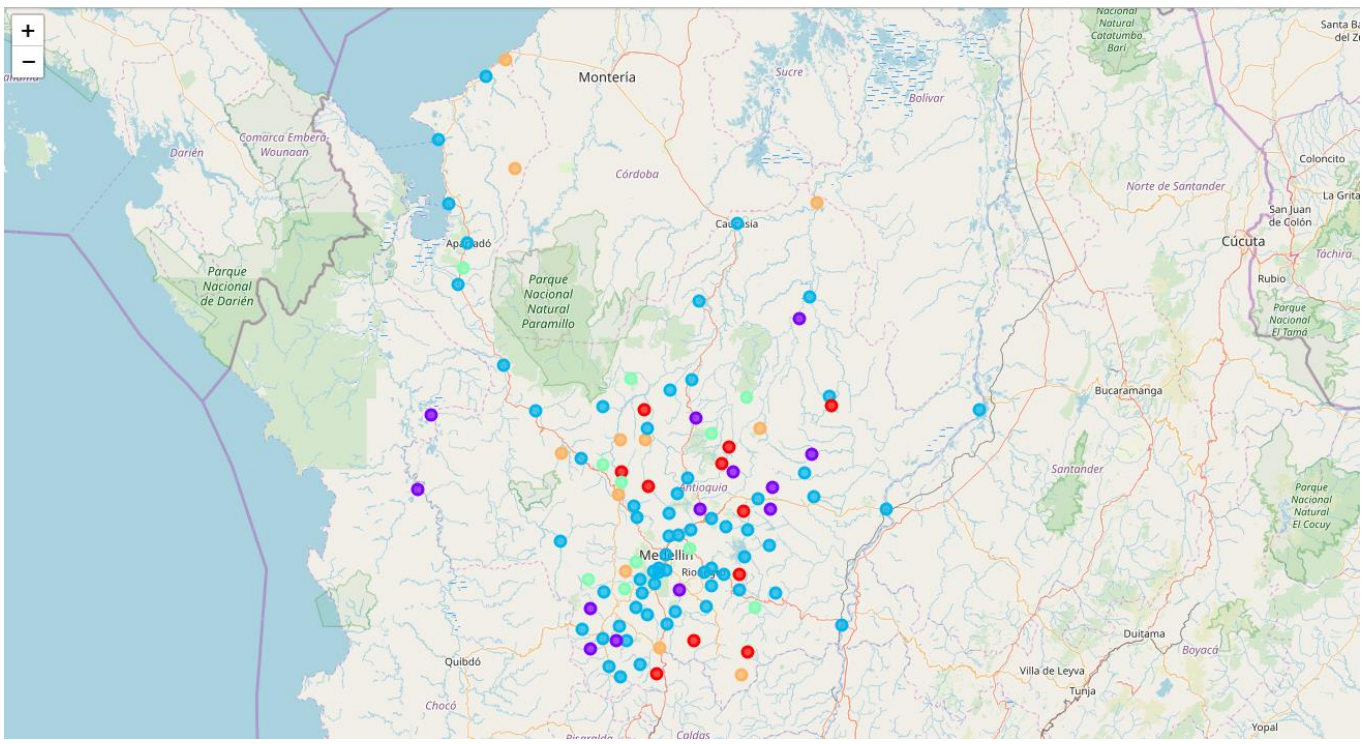


Figure 18: Boroughs into 5 Clusters according with venues categories

Cluster 1: Boroughs with a Plaza as a first or second most common venue

Cluster 2: Boroughs with Hotels as a first most common venue

Cluster 3: Boroughs with a Park as a first or second most common venue

Cluster 4: Boroughs with Coffees, bars, farms into their 10 first common venues

Cluster 5: Boroughs with Restaurants, Fast food restaurants, farmers Market, electronic stores into their 10 first common venues

4.4 Find a Borough in Antioquia similar to Southwark, London.

Assuming our investor used to live in Newhan, London, and with the Foursquare location data we calculate the same matrix with the categories of common venues in Southwark

	Borough	Airport	Aquarium	Arepa Restaurant	Asian Restaurant	Australian Restaurant	BBQ Joint	Bakery	Bar	Basketball Court	...	Theater	Tour Provider	Town	Toy / Game Store	Track	Trail	Water Park	Wine Bar	Wine Shop	Wings Joint
92	Southwark	0.0	0.0	0.0	0.02	0.01	0.0	0.02	0.02	0.0	...	0.0	0.0	0.0	0.0	0.0	0.01	0.0	0.03	0.01	0.0

Then we calculate the Euclidian distance between the Southwark vector and each borough in Antioquia with the following results

	Borough	Distance
59	Medellín	0.253296
38	Envigado	0.295701
51	Jardín	0.313139
45	Guatapé	0.335857
49	Itagüí	0.338260
...	...	...
78	San Francisco	1.014298
16	Betulia	1.014298
43	Guadalupe	1.014298
47	Heliconia	1.014298
18	Buriticá	1.014298

109 rows × 2 columns

Figure 18: Euclidian Distancce between Antioquia Boroughs and Southwark according with popular venues

As result, we find Medellin as Borough in Antioquia that result more similar to Southwark, London in terms of popular venues around (The shortest Euclidian distance)

5 DISCUSSION

Colombia is the 55th largest export economy in the world and the 53rd most complex economy according to the Economic Complexity Index (ECI).

Being Colombia a country with a great diversity of products and having a unique variety of crops, it has not sense that the exports of Colombia have decreased during the last five years. I found in Data science an opportunity to check the numbers with a realistic view and try to motivate foreign investors to support Colombian economy and gain benefits from part and part.

Data is an objective way to find the best alternatives and just it is what I try to recommend to the foreign investors. Antioquia is a city with a state of the art development, and characterized by its large coffee crops.

I used the Kmeans algorithm as part of this clustering study to group Antioquia Boroughs according with their popular venues, I also compare those boroughs with a London city, trying to advice a foreign investor the best place to set up his business.

6 CONCLUSION

- Colombian exports data estimated that the department with the greatest export value is Antioquia, where there are variety of products to process and export like coffee byproducts, fruits and vegetables and bakery products.
- The preferred product to export in Colombia is the raw coffee with which the greatest amount of export value has been realized in at least the last 7 years.
- Medellín the capital city of Antioquia is the borough with the greater number of popular venues. Using the Euclidian distance as comparative factor, Medellin is the borough with more similarities to Southwark, London in terms of popular venues around (The shortest Euclidian distance)

7 REFERENCES

- https://www.maro.com.co/
- https://oec.world/en/profile/country/col/
- https://gist.githubusercontent.com/john-guerra/43c7656821069d00dcbc/raw/be6a6e239cd5b5b803c6e7c2ec405b793a9064dd/Colombia.geo.json
- http://www.antioquiadatos.gov.co/index.php/1-3-1-division-del-departamento-de-antioquia-por-subregion-zonas-y-municipios