Brief analysis of local Foursquare recommendation and US Presidential Elections results in Florida

Alejandro Ciriano

April 2020

1 Introduction

1.1 Background

The US presidential election is one of the most important events in the country. They happen every four years, the last one in 2016, where the main candidates were Donald Trump and Hillary Clinton. Results are obtained after deciding which candidate has won in each of the states, which are divided into counties. Many specialists try to forecast what the trend is in each of the states, in order to use that information and convey an adapted message to each of them in the electoral campaign. It is not an easy task, many parameters and factors are very important and need to include them to develop a correct model.
The purpose of this project consists of analyzing if is possible to link the venues and recommendations of a local zone to the vote intention of the population in the zone. It is an ambitious purpose, we cannot make reliable and accurate predictions about political trendings only with data about venues, but it is interesting seeing how different counties are clustered depending on the places of their zones.

1.2 Problem

So, as the previous section describes, the objective is try to predict the behaviour of the population based on the venues which are recommended in their zone. To do that, some data is collected using the Foursquare API, where everybody can make queries about location data.
Note that US is very large, analyzing every state is out of the scoped of this project and a free account in Foursquare. For this reason, I have chosen Florida because it's large enough, there are many counties and cities and the results of the elections show that both, Trump and Clinton, could won in this state. Final results are shown bellow and you can interact with the map in the link. [1]

[1] https://www.nytimes.com/elections/2016/results/president

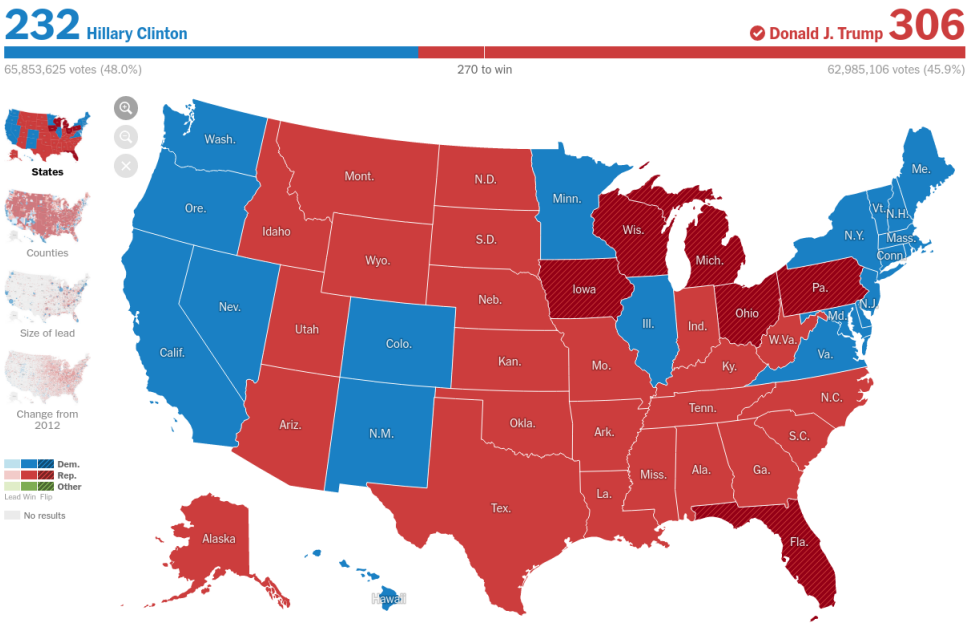

Figure 1: Presidential Elections 2016 results

1.3 Interest

Obviously, a lot of people work and make a huge efforts to develop these kind of ML (*Machine Learning*) models. Political parties or even venues are some examples of entities that are interested in this kind of research, but there are much more examples.

1.4 Some things to keep in mind

- I will use the venue recommendations in the cities of Florida to characterize each county.
- I will cluster the counties in two groups trying to simulate the principal options, which are Donald Trump and Hillary Clinton. If the clusters are similar to the results map of Florida (shown in the next figure), we can conclude that the local business of a region is partially connected with the vote intention or with political ideology in the zone.
- I don't have any expectations about the result, I won't use some parameters that could be important. The real world is much more complicated and forecasting about the results of US Elections is a huge challenge. This is just for fun.

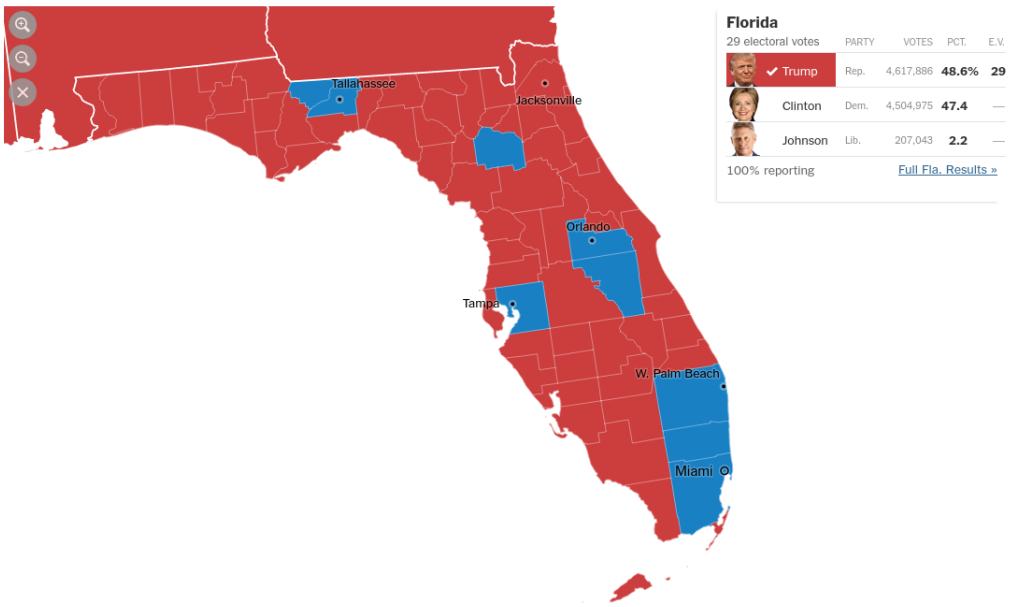

Figure 2: Results in Florida's counties

2 Data acquisition and cleaning

2.1 US Presidential Elections 2016

[2] This dataset contains the results of the past elections. It stores a lot of information which can be used to cluster the population and forecasting, like population description, population incomes, medium age, etc. Each row in this table represents the result that was obtained by both candidates in all of the counties of US. Of all the available columns, only those associated with income, mean age, population size, and poverty levels have been chosen.

	county	state	population	65plus	income	poverty	total_votes	clinton	trump
0	Autauga County	AL	55395	13.8	24571	12.1	24661	0.239569	0.734358
1	Baldwin County	AL	200111	18.7	26766	13.9	94090	0.195653	0.773515
2	Barbour County	AL	26887	16.5	16829	26.7	10390	0.466603	0.522714
3	Bibb County	AL	22506	14.8	17427	18.1	8748	0.214220	0.769662
4	Blount County	AL	57719	17.0	20730	15.8	25384	0.084699	0.898519

Figure 3: Results in US counties

[2] https://www.kaggle.com/prashant111/us-presidential-election-data

2.2 Counties Locations

[3] The second dataset contains the coordinates of the cities of every county. It is necessary for making queries in Foursquare, setting the latitude and the longitude of the point where search for venues.

	zip_code	latitude	longitude	city	state	county
0	501	40.922326	-72.637078	Holtsville	NY	Suffolk
1	544	40.922326	-72.637078	Holtsville	NY	Suffolk
2	601	18.165273	-66.722583	Adjuntas	PR	Adjuntas
3	602	18.393103	-67.180953	Aguada	PR	Aguada
4	603	18.455913	-67.145780	Aguadilla	PR	Aguadilla

Figure 4: Sample of US locations

2.3 Florida

- **Votes in Florida:** the data with results is really useful because it contains the list of the counties that are part of the state of Florida. After filtering by State, the final dataframe contains 67 rows, one for each county of Florida.

- **Locations in Florida:** the dataset with locations is much larger than that of counties. This is because it stores both latitude and longitude of postal codes within cities.
  After selecting the Florida rows, duplicate records need to be removed because different ZIP codes share coordinates. This makes a total of 1,447 geographic points within the state, which is too much for the scope of this project. To solve it, only one point per city is considered. The point that is chosen in the cities is the closest to the mean of all of the points in the city. In this way, we can suppose that this is the closest point to the center of the city (it doesn't have to be true, but in most of the cases yes). Finally, there are 525 locations in Florida, which is much more viable to analyze with a free account.

Before starting collecting data from this locations, we must be sure about the name of the counties in both datasets are the same. First, the final word 'county' in the column 'county' of votes dataset is deleted. Second, the counties 'Saint Johns', 'Saint Lucie' and 'De Soto' in locations dataset are exchanged for 'St. Johns', 'St. Lucie' and 'DeSoto'.

[3] https://docs.gaslamp.media/download-zip-code-latitude-longitude-city-state-county-csv/

```
(67, 10)
```

	county	state	population	65plus	income	poverty	total_votes	clinton	trump	winner
0	Alachua	FL	256380	12.5	24857	24.9	127827	0.589625	0.364430	1
1	Baker	FL	27093	12.9	19852	17.3	12634	0.167168	0.814785	0
2	Bay	FL	178985	16.1	24498	14.7	87151	0.248867	0.711524	0
3	Bradford	FL	26702	17.6	17749	18.2	12098	0.241693	0.736733	0
4	Brevard	FL	556885	22.6	27009	13.5	314337	0.380245	0.577788	0

Figure 5: Extracted data of votes in Florida and its shape.

```
(525, 6)
```

	zip_code	latitude	longitude	city	state	county
0	32082	30.102212	-81.382302	Ponte Vedra Beach	FL	Saint Johns
1	32033	29.813208	-81.468724	Elkton	FL	Saint Johns
2	32085	29.937673	-81.420603	Saint Augustine	FL	Saint Johns
3	32145	29.688750	-81.406081	Hastings	FL	Saint Johns
4	32259	29.877289	-81.561245	Jacksonville	FL	Saint Johns

Figure 6: Extracted data of locations in Florida and its shape.

3 Exploratory data analysis

3.1 Exploring US data

Many relevant aspects of a country can be analyzed with its demographic, economic and social data. The electoral results dataset contains all this information, so some checks and investigations can be carried out. The correlation of the main parameters of US counties is shown bellow.

	population	65plus	income	poverty
population	1.000000	-0.222524	0.260816	-0.063957
65plus	-0.222524	1.000000	-0.040189	-0.101606
income	0.260816	-0.040189	1.000000	-0.726398
poverty	-0.063957	-0.101606	-0.726398	1.000000

Figure 7: Correlation between parameters in US

A scatter plot is shown in the figure 8. In this plot, we can see the relation between poverty and income in the counties of US. The behaviour that everybody expect to see in this curve is a decrease of income levels when poverty increases, and that is what it shows. But there is an isolated point with the highest income level and a much more higher level of poverty than the mean. This point represents the county of New York, which is one of the most populated and multicultural counties in the country.

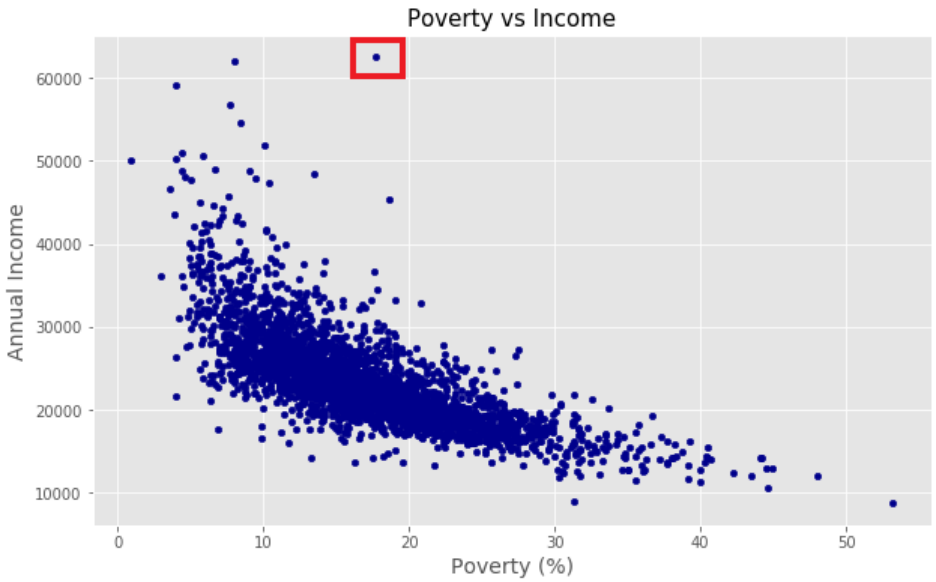

Figure 8: Poverty vs Income in US

And we can also obtain the average of the parameters in the counties where each candidate won. The next figure shows that Clinton was more popular in populated counties with young people. She was also more popular in counties where the wealth is not equally distributed (higher income and poverty).

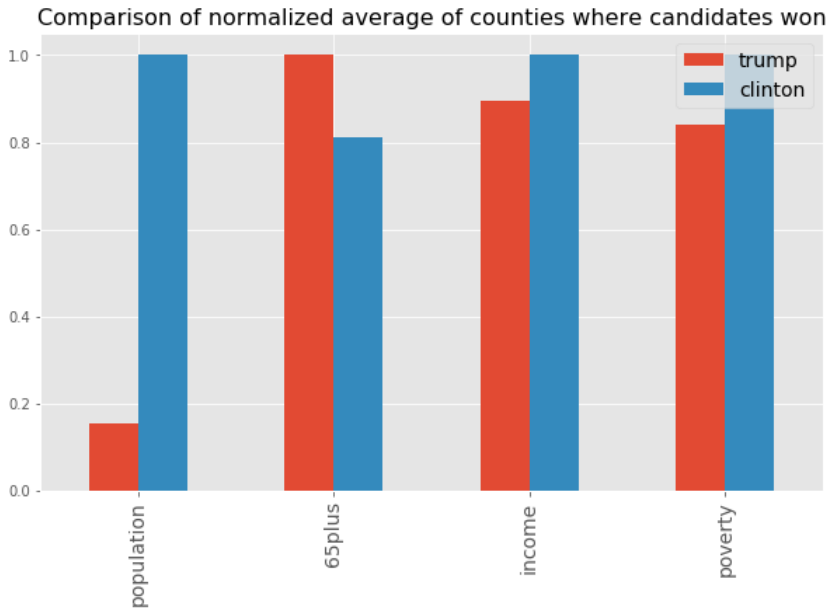

Figure 9: Comparison of normalized average of counties where candidates won

3.2 Exploring Florida data

If we focus on the state of Florida, we can do similar analysis as we have done with all the country.

	population	65plus	income	poverty
population	1.000000	-0.110823	0.372807	-0.244812
65plus	-0.110823	1.000000	0.305341	-0.310828
income	0.372807	0.305341	1.000000	-0.751438
poverty	-0.244812	-0.310828	-0.751438	1.000000

Figure 10: Correlation between parameters in Florida

Again, as the correlation in the country, the unique parameters which correlation's value is close to 1 are income and poverty. This value is -0.75 and previously was -0.726, so nothing new or remarkable to analyze.
Finally, the bar chart with the average of population, mean age, income and poverty of the counties for each candidate is similar as before.

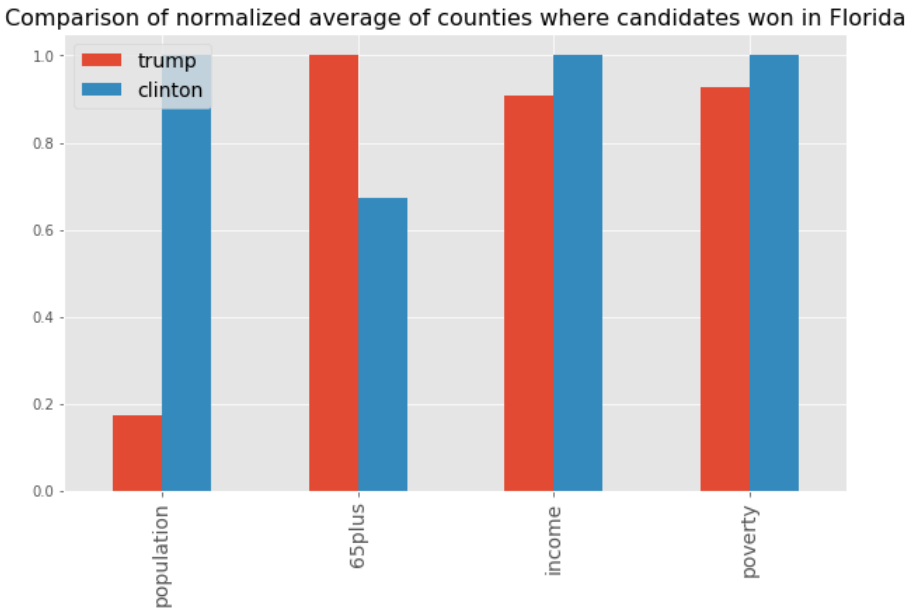

Figure 11: Comparison of normalized average of counties where candidates won in Florida.

4 Counties clustering with KMeans

In this chapter results that has been obtained are shown. First, the process of querying information to Foursquare is described and after some maps represents the clusters of counties in Florida depending on the data that is used to implement KMeans algorithm. Geo data about Florida has been downloaded from this link.[4]

4.1 Collecting venue recommendations from Foursquare

1. Obtain a list of categories and subcategories of Foursquare in the date of the last elections.

2. For each location register in Florida, send a request to the Foursquare API about recommendation in the zone. The limit is 20 and the radius 2 Km.

3. Extract the parent categories of the venues that appear in the response and add 1 in a dataframe with [index]=Counties and [columns]=Categories. Note that each county has multiple cities, and there is one register per city.

[4]https://github.com/johan/world.geo.json

4.2 Clustering with venue recommendations

So, if we apply KMeans algorithm to the counties using normalized data from recommendations and 2 clusters, the map result is:

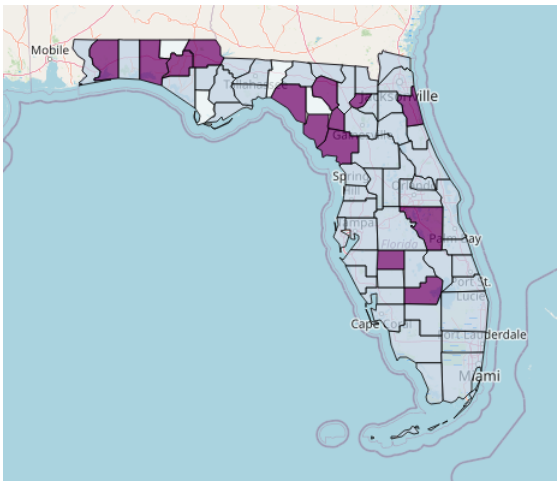

Figure 12: Clusters of counties based on recommendations.

4.3 Clustering with venue recommendations + counties data

Using population, income, poverty and percentage of people which is 65 or older, and we add the data from recommendations, results improve a lot:

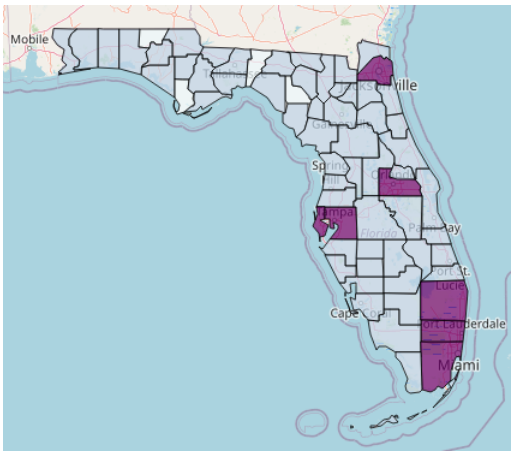

Figure 13: Clusters of counties based on recommendations.

5 Conclusions

- If we compare the maps with the result after clustering the counties and the map with electoral results, we can see that is more relevant and describes better the people behaviour the data from population, medium age, income or poverty levels than the data from venues in local zones.
- It does not mean that the venues and business are not connected with the political ideology or the vote intention. In this project only a few data have been used, and we cannot draw conclusions from this.
- With much more information of trendings and venues we could deploy a better model which integrates this massive data from each city and then achieve better results after clustering counties.
- Despite of the fact that the results of the elections in Florida show a deep correlation with population volume and income, final results and real results are close enough to be validated, and the resources of this project have been really limited.