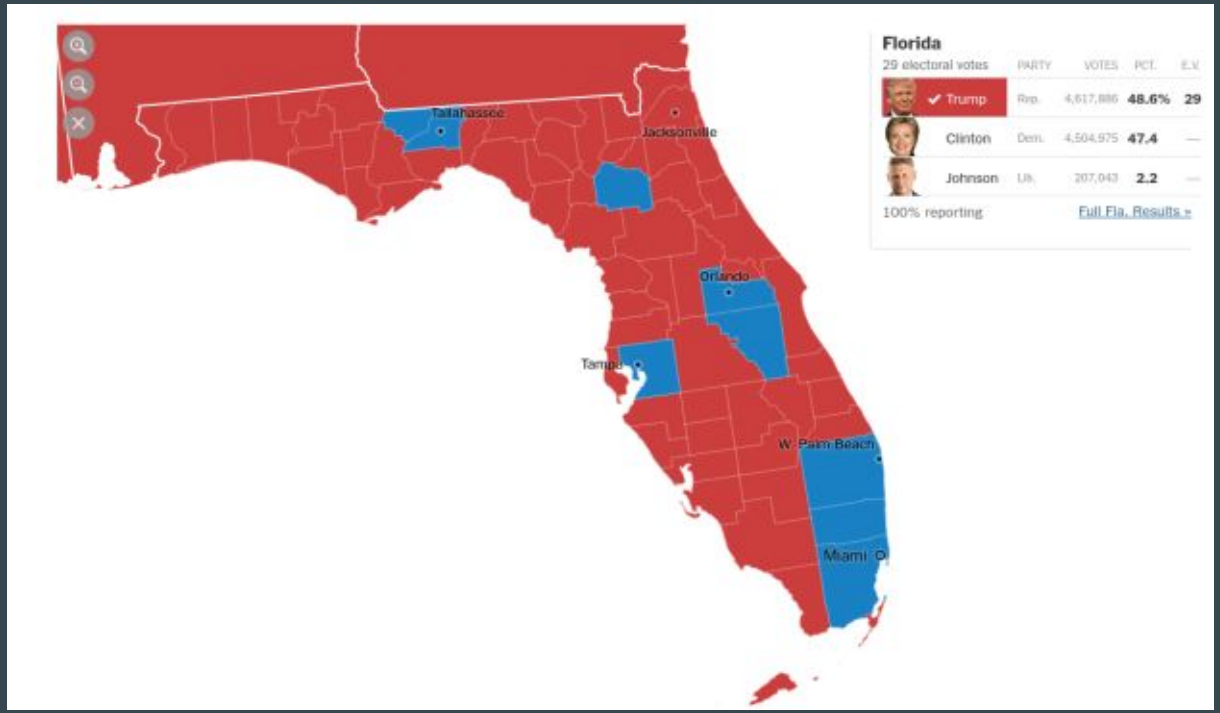
Brief analysis of local Foursquare recommendation and US Presidential Elections results in Florida

Alejandro Ciriano
April, 2020

Problem

- Predicting the vote intention of the population is really difficult nowadays despite of we are a developed society.
- 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.
- Find out connections between politics and other facts of the real life can decide who will be the winner of the next Elections.

Florida results in US Presidential Elections 2016



Question

Is there any connection between trending venues and the vote intention of the population?

Answer

Let’s collect some data from Foursquare to cluster Florida counties and compare the results with the Elections results.

Data acquisition

1. **US Presidential Elections 2016:** the results of the Presidential Elections. Contains one register per county with the state, population, income, poverty and much more interesting parameters. Only some features are chosen. *(https://www.kaggle.com/prashant111/us-presidential-election-data)*

	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

Data acquisition

1. **Counties locations:** 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.
(*https://docs.gaslamp.media/download-zip-code-latitude-longitude-city-state-county-csv/*)

	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

Selecting and Cleaning data

1. **Votes in Florida:** we can filter and select the counties where state = 'FL' to obtain the counties of Florida. It is also needed to modify the column 'county', each county name ends with the word 'County'. We should delete this word.

```
(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

Selecting and Cleaning data

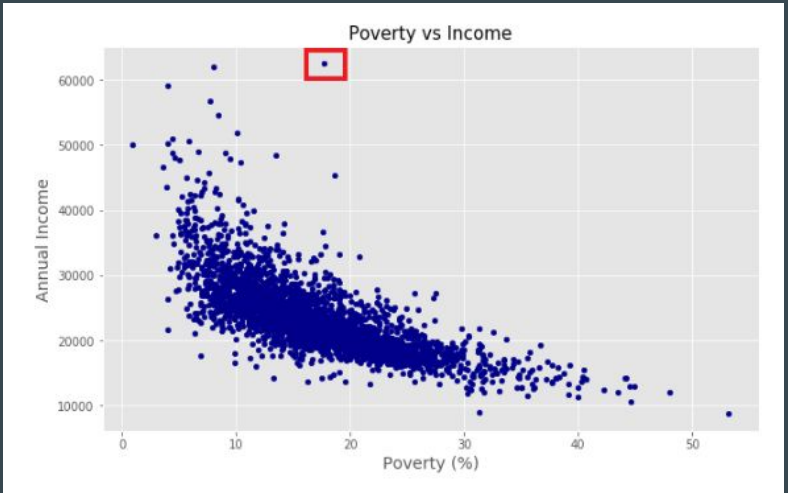
1. **Locations in Florida:** Next step is filter the locations data. Again, we can filter and select locations from Florida by state = ‘FL’. To reduce the total number of locations, only the closest point to the center of each city is selected. Finally, to avoid incompatibilities with the other dataset, some counties names are exchanged (’De Soto’ -> ’DeSoto’, etc.)

```
(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

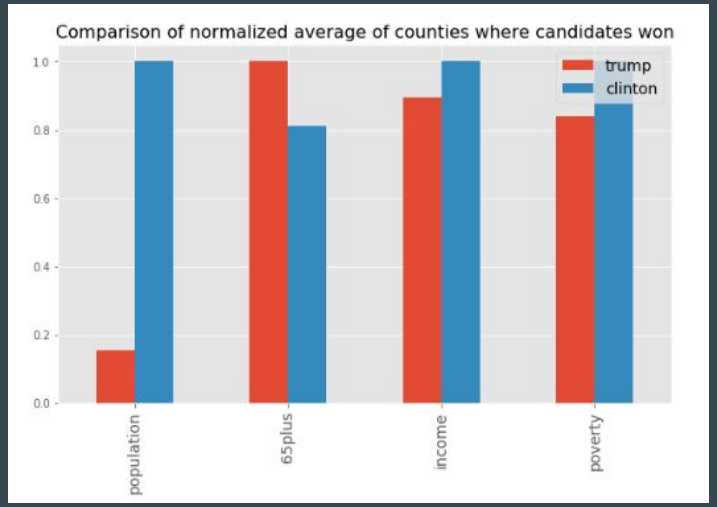
Exploratory Data Analysis: correlation

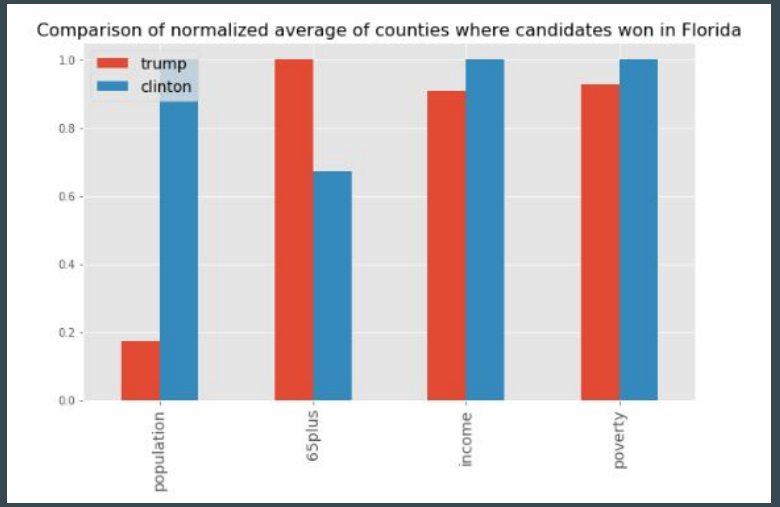
- Correlation factors between parameters show that the unique facts that are linked are poverty and income. It's normal that if poverty goes up, wealth goes down, but we can check it:



New York County(red square) is the most polarized region in poverty and income. The rest of the counties behave as we expected.

Exploratory Data Analysis: candidates and voters





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).

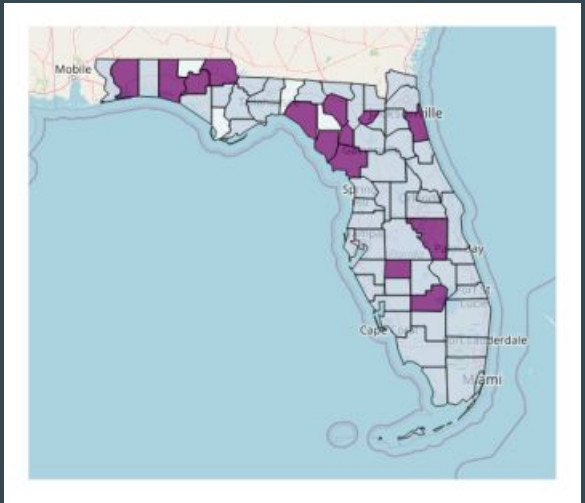
Collecting from Foursquare

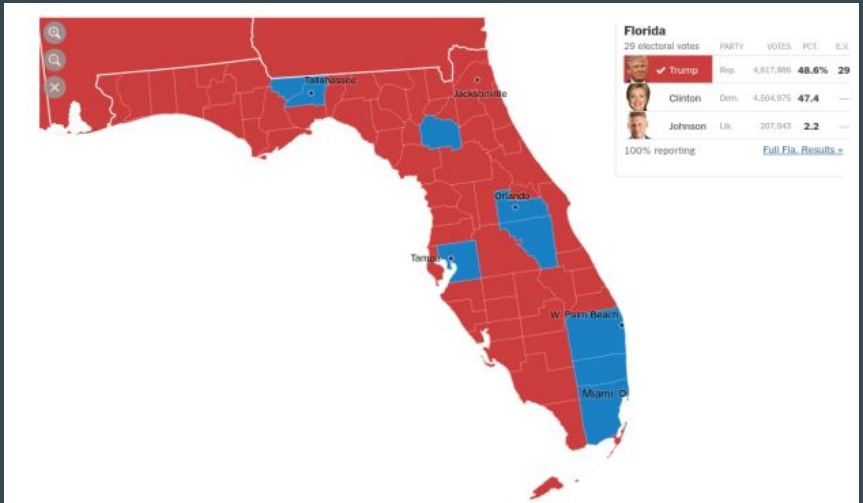
1. For each location in Florida, get its latitude and longitude.
2. Send a request to the Foursquare API about recommendation in the zone with ***explore***. The limit is 20 and the radius 2 Km.
3. Extract the parent categories of the venues that appear in the response and store them in the county register.
4. Normalize the result.

	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
Alachua	0.032787	0.0	0.0	0.622951	0.016393	0.032787	0.016393	0.065574	0.213115	0.000000
Baker	0.000000	0.0	0.0	0.666667	0.000000	0.166667	0.000000	0.000000	0.166667	0.000000
Bay	0.000000	0.0	0.0	0.600000	0.000000	0.200000	0.000000	0.000000	0.133333	0.066667
Bradford	0.000000	0.0	0.0	0.181818	0.090909	0.181818	0.000000	0.000000	0.363636	0.181818
Brevard	0.028037	0.0	0.0	0.514019	0.037383	0.102804	0.000000	0.000000	0.224299	0.093458

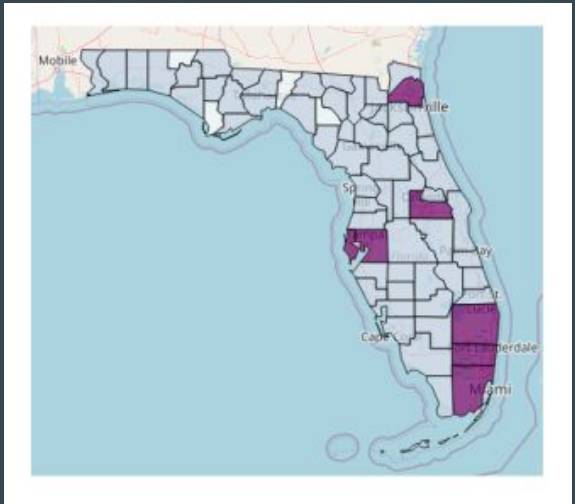
Counties clustering

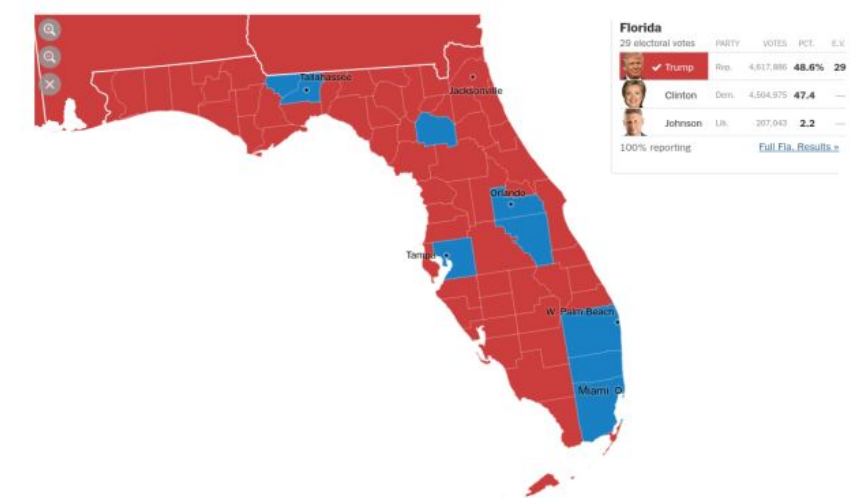
Applying KMeans algorithm to the counties based on the information that was collected from Foursquare, the result is not as good as we want:





... but, if we include demographic data, the accurate improves a lot:





- Error rate with Foursquare venue recommendations: 19.40 %
- Error rate with recommendations + poverty, income, population and age: 7.42 %

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.