Georgia Counties and the 2020 Election: Lifestyle Factors

11.2020

Preeti Putcha

Capstone

Pipeline and Summary

SUMMARY
The rationale behind this project was to: a)investigate voter lifestyles in light of the extremely close liberal win in Georgia this 2020 presidential election. b)construct a pipeline to analyze counties and Foursquare check-ins in tandem with other demographic factors of interest.

159 counties in Georgia were analyzed for political skew and characterization of venue check-ins as an indicator of lifestyles associated with conservative and liberal voters.

latitude, longitude, voting data, total voting populations, definition and calculation of political skew ratios, and county binning were merged with kmeans clustering of Foursquare check-ins gathered with 500 and 5000 meter queries.

kmeans Clustering by venue preference showed significant differences in political skew via ttest, with Yoga Studios and Cuisine (Cupcake Shops and Mexican Restaurants) as unique venues representing liberal skew and Fast Food and Baseball Fields as unique venues associated with conservative skew. These findings are supported by literature and studies analyzing fan bases for these activities.

While interesting, this work is preliminary and comes with many statistical caveats and need for further analysis. These will be discussed in the report below.

This work nevertheless provides a baseline and potential pipeline for studying county distribution across lifestyle preferences and political preferences.

Code is available here
https://github.com/PPutchaML/Coursera_Capstone/blob/main/PPcapstone.ipynb

Part II:Data Analysis

Notes

- The following languages and libraries were used: -Folium(mapping), Python, (Pandas,Numpy,Scikitlearn, Matplotlib,Seaborn,Datacompy, StatAnno), GEOJson,Git.

CLEANUP

- **null or absent values were dropped**: there were actually none, since all 159 counties appeared to be present with no missing values after this command.This makes sense, a)since voting tallies are up to the minute b)lat/long files are for a very well known region not an unknown set of values c)The smaller scale: At 159 counties with the total voting population of 4.9 million, this is not such a high throughput project that errors accumulate by probability alone.

- **Defining voting population**:The total voting population per county was calculated, and votes for the Libertarian third party candidate were removed:while less than 1% of the total votes and thus significant in a narrow race, they did not count towards the main total of blue over red.

- **Defining conservative and liberal votes**:Then, the ratio of blue to red voters was calculated. If the ratio was greater than 1, the county was marked blue. If the ratio was less than 1, the county was marked red. The actual ratio, however, provided more precision for analysis, especially in "swing counties" where very few votes might result in a ratio of 0.99, or 1.01, for instance. An example of this process is shown below.

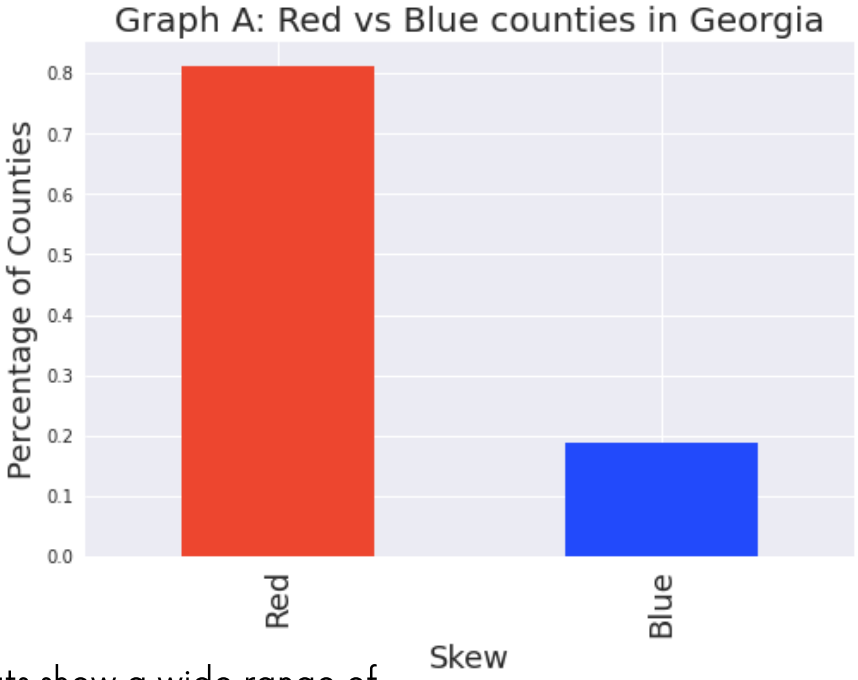
- Finally the dataframes were merged to proceed:see example below.

[9]:

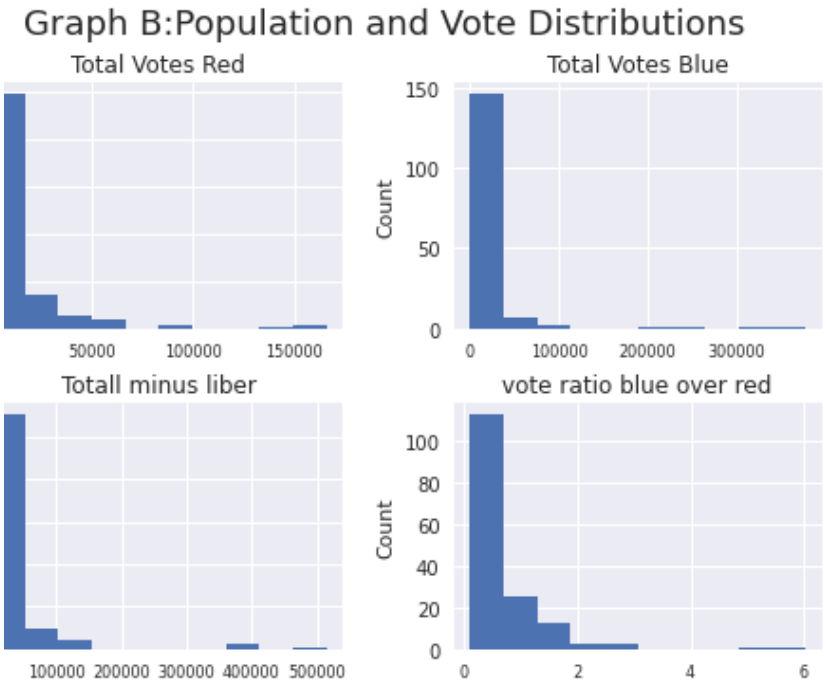
	County	Total Votes Red	Total Votes Blue	Totall minus liber	vote ratio blue over red	Skew	Lat	Long
0	Baker	897	652	1549	0.726867	Red	31.326183927	-84.4446694741
1	Calhoun	923	1259	2182	1.364030	Blue	31.5291972743	-84.6245076946
2	Chattahoochee	880	667	1547	0.757955	Red	32.346971275	-84.7870462059
3	Clay	637	790	1427	1.240188	Blue	31.6262755156	-84.9801029119
4	Echols	1256	167	1423	0.132962	Red	30.7100896074	-82.8939351883
5	Glascock	1403	155	1558	0.110478	Red	33.2292799721	-82.6107022318
6	Quitman	604	497	1101	0.822848	Red	31.8673305599	-85.0187841576
7	Schley	1800	462	2262	0.256667	Red	32.2616858077	-84.3147208166
8	Stewart	801	1182	1983	1.475655	Blue	32.0784621843	-84.8352023744
9	Taliaferro	360	561	921	1.558333	Blue	33.5660908408	-82.8787644687
10	Webster	748	639	1387	0.854278	Red	32.0466488616	-84.5510526557
11	Wheeler	1583	688	2271	0.434618	Red	32.1170652349	-82.7245932678

Part II:Data Analysis

1.Initially, voting population and the actual percentage red vs blue was examined. Georgia votes comprise ~5 million. Across counties, Georgia skews red 81%, leaving ~19% blue(Graph A). But this does not take into account population -the popular vote-and actual skew number, so it bolsters our analysis on the popular vote moving forward.



2.Descriptive Stats show a wide range of county populations, corroborated by a histogram showing clear bins of county populations (**Graph B,Totall minus liber**). Histogram bin optimization was calculated using **Sturge's Law**[].Sturge's Law optimizes the number of bins for a histogram with between 30 and 200 observations, resulting in 7 bins here.
k = 1 + 3.322 log n, Where:
k = the number of bins
n = the number of observations in the data set.
The diverse population ranges of these counties are revealed in clearly separated bins(red marker),and presents a variable to be considered later: How will venues be sorted when thinking of populations under 2000 vs populations around 20000? However, we will first proceed with the baseline analysis.

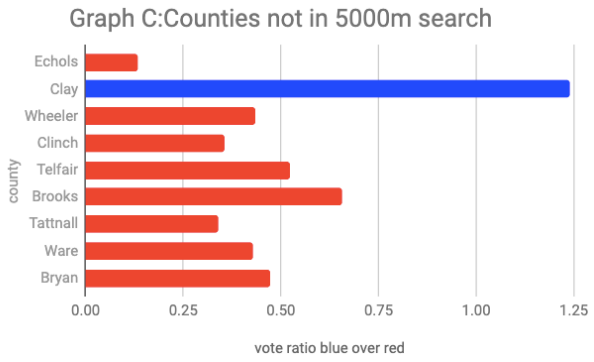


Foursquare:5000 meters

API calls were made for 159 counties,looking for the top 100 venues within 5000 meters(~3 miles) and 500 meters per county. The radius of 5000 was picked due to extrapolation: Manhattan and other densely populated cities can use a radius of 500 meters, but Georgia counties have a more diverse population size, as demonstrated in the histogram.

5000m Observations:

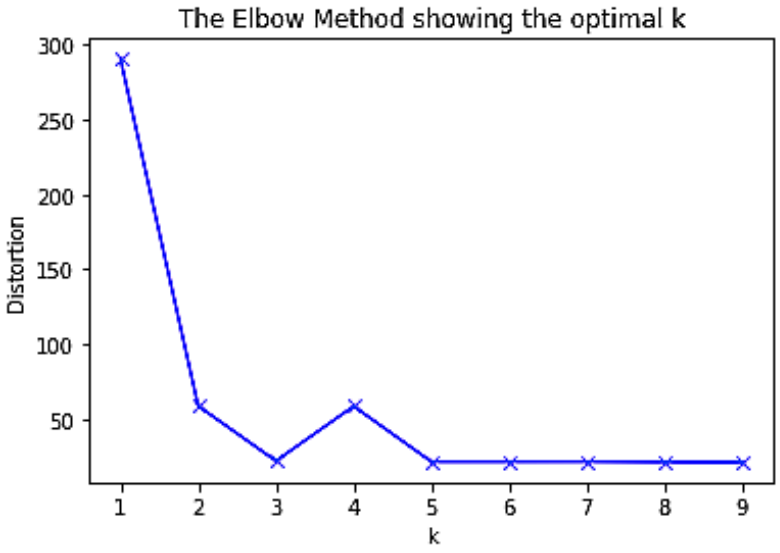
- 300 unique venue categories
- 150 counties had data
- 4802 total entries
- 9 counties did not return venues at all--of these 8 were red(**Graph C**).



Given the higher proportion of liberals more sympathetic to science and tech vs conservatives, it's possible red counties who are also rural don't use Foursquare. The notable exception here is Clay County (blue), which is strongly blue with a ratio of 1.24 and the county of the late John Lewis, an African-American and prominent civil rights senator.

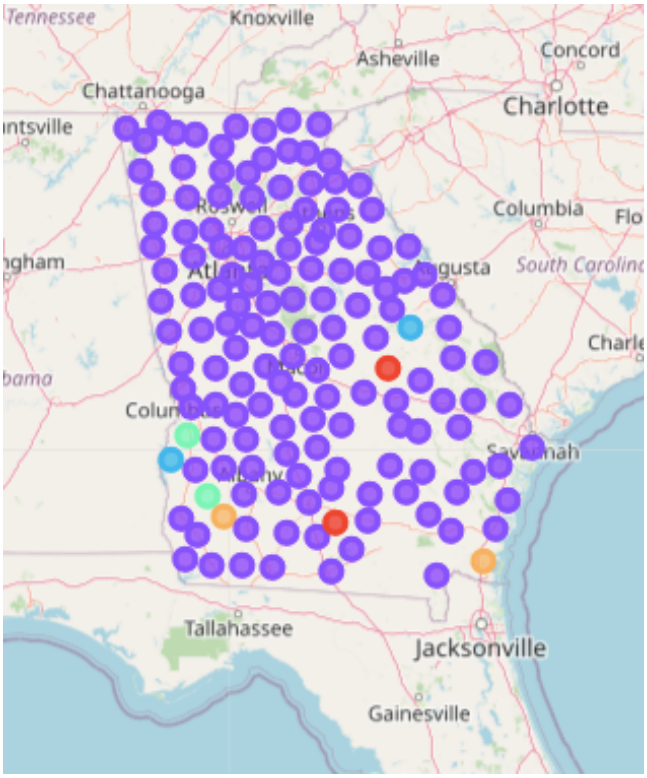
Cluster Label optimization(Graph D):
The elbow method of optimal kmeans was inconclusive, with peaks beyond the inflection point of 2. A **k of 5** was selected for initial trials, the idea being to run clustering iteratively over k =4 and 3 to compare cluster accuracy.

Graph D



KMEANS CLUSTER: RESULTS FOR 5000 METERS

Graph E

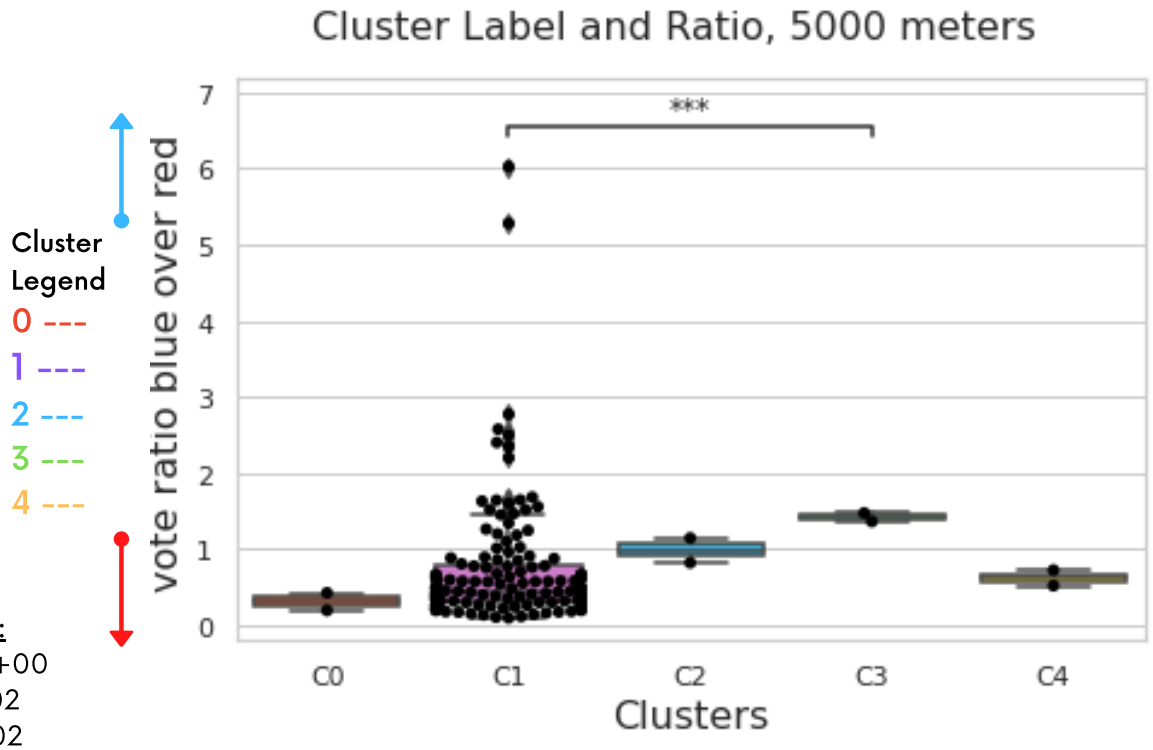


With a **kmeans cluster number of 5,** the counties sorted very unevenly, with most in **cluster 1(**purple **on Graph E left).**

Welsh's t test:
showed significant variance only between Cluster **0 and 3(p<.05**),The boxplot and swarmplot overlaid atop it show that there are a disproportionate amount of samples in cluster 1 as well as many outliers, complicating accurate statistical analysis.

The one-way ANOVA/F_oneway score was not significant, as is expected from this more stringent measure and the above sample distribution issues. You will especially note outliers and a wide range of scores within Cluster 1. Summary table and legend are below the boxplot(Graph F).

Graph F



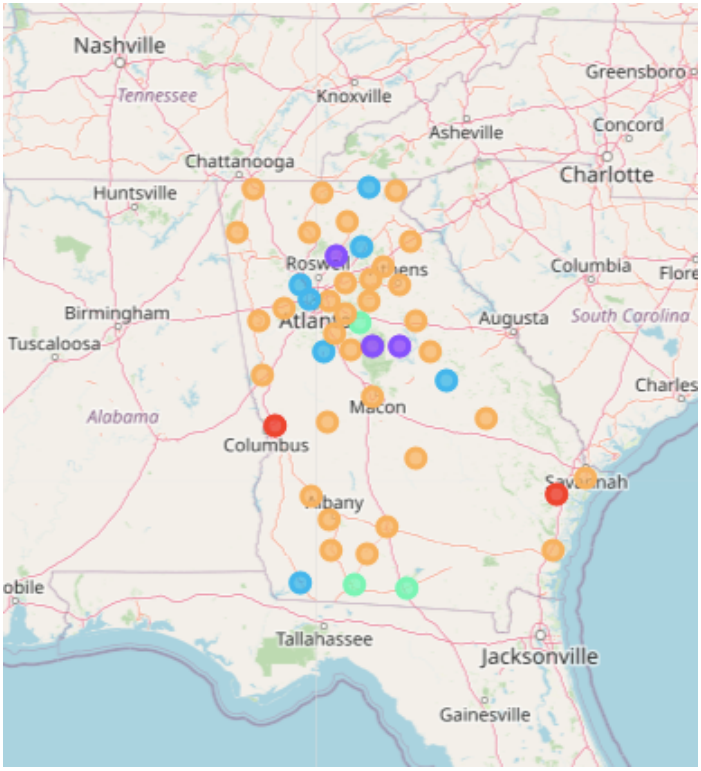
p-value annotation legend:
ns: 5.00e-02 < p <= 1.00e+00
*: 1.00e-02 < p <= 5.00e-02
**: 1.00e-03 < p <= 1.00e-02
***: 1.00e-04 < p <= 1.00e-03
****: p <= 1.00e-04
C1 v.s. C3: Welch's t-test independent samples with Bonferroni correction, P_val=2.096e-04 stat=-8.043e+00
F_oneway
F_oneway result(statistic=0.587, pvalue=0.672

KMEANS CLUSTER RESULTS FOR 500 METERS

With a **kmeans cluster number of 5 and a 500 meter query,** the goals were to see if:

- This query would result in more data points per county (assuming the venues in immediate vicinity of this narrow radius were bound to be in more populous areas and therefore better candidates for many check-ins,) compared to the 5000 meters query.
- This query would result in a more even distribution of venues and counties than the 5000 meter query.
- The resultant statistical analysis would have less error and more robust sampling, leading to a more accurate analysis of variance.

Graph G



Results

This set of venues had a more even distribution, albeit one still weighted to one cluster(orange dots in the kclustering figure/Graph G).

Welchs T-test with Boneferroni correction:

More, but not all, clusters showed statistically significant variance(Graph H), indicating that clustering by venue could be accurately reflecting political ideologies.

ANOVA

F_oneway ANOVA scores were still not significant. Some factors influencing this result: the samples here are still not optimally distributed enough to be robust with an ANOVA. However, the distribution is much improved compared to the 5000 meter query, and this is reflected in the t tests. more cluster comparisons here are significant and with lower p values.

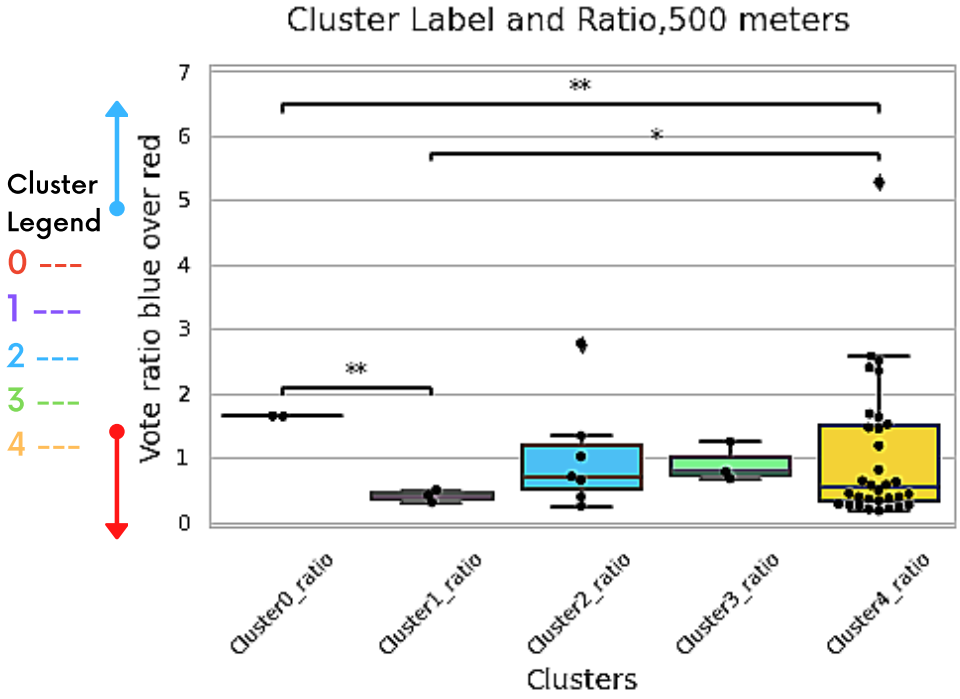
p-value annotation legend:

ns: 5.00e-02 < p <= 1.00e+00
*: 1.00e-02 < p <= 5.00e-02
**: 1.00e-03 < p <= 1.00e-02
***: 1.00e-04 < p <= 1.00e-03
****: p <= 1.00e-04

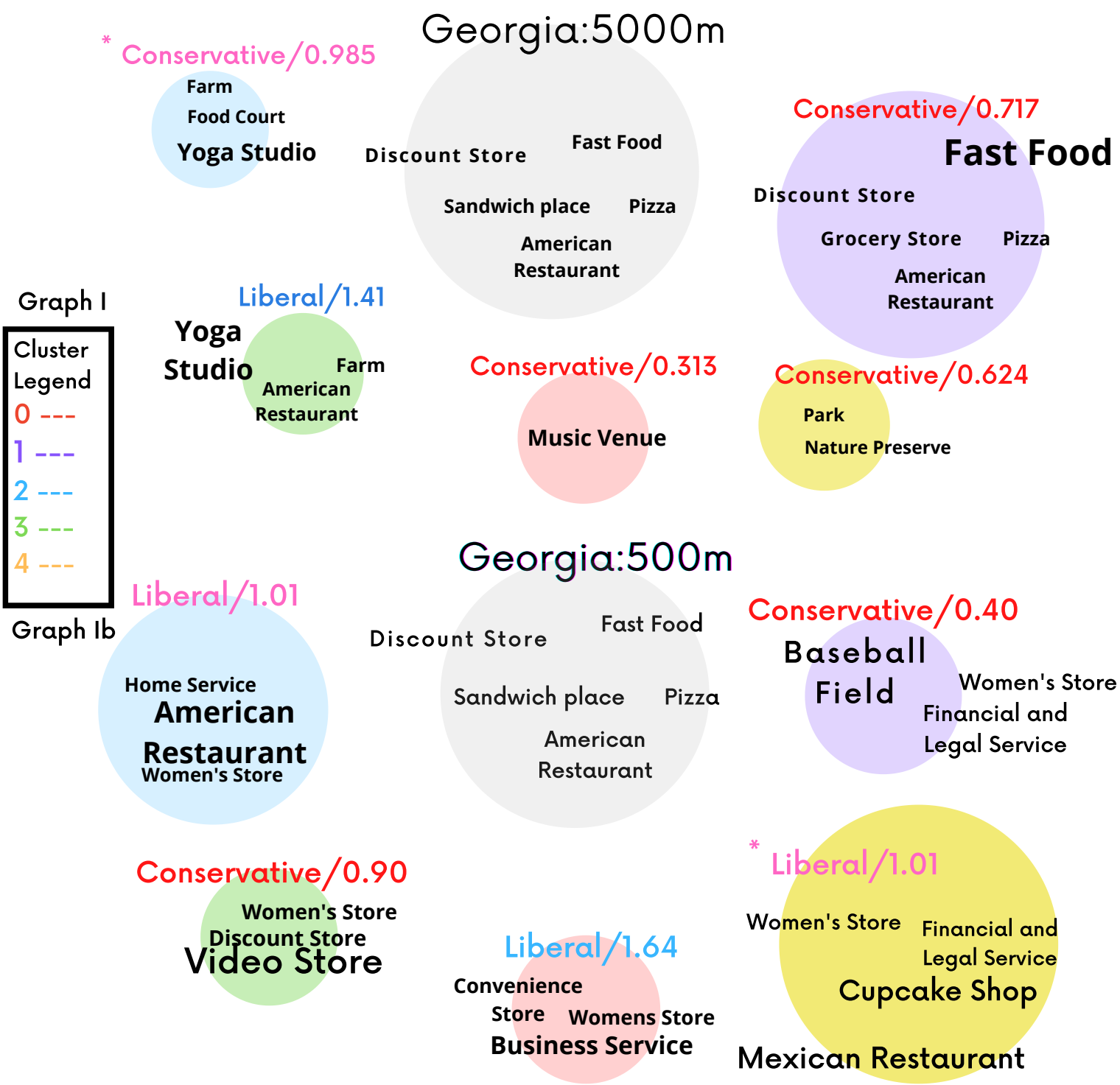
Cluster0_ratio v.s. Cluster1_ratio:
P_val=6.058e-03 stat=2.210e+01
Cluster1_ratio v.s. Cluster4_ratio:
P_val=1.434e-02 stat=-3.027e+00
Cluster0_ratio v.s. Cluster4_ratio:
P_val=7.352e-03 stat=3.298e+00

F_oneway
F_onewayResult(statistic=0.490, pvalue=0.742

Graph H



VENUES UNIQUE TO POLITICAL SKEW



- Venues uniquely associated with cluster skew are representatively selected here in larger font. Unique venues in the 500m analysis also align with ttest significance between corresponding clusters.
- For instance Yoga Studios, Cuisine, and cross-cultural food seem associated with a liberal political skew, while Baseball Fields and Fast Food seem associated with a conservative political skew.

*Ratios close to one are considered politically to have "purple", or "swing" voting populations. In this analysis, 0.98 to 1.02 range ratios are considered "purple."

OBSERVATIONS(Graph I and Ib):

- The top 3-5 venues were selected from each cluster through sorting and grouping per set of cluster venues quantitatively. Unique differences were selected.
- The cluster circles roughly indicate the relative sizes of each cluster to the others within the 500 or 5000m query. (NOTE:This approach was done qualitatively and not via a specific computational technique--the clustering approach and stats account for quantitative and standardized representation.
- Venues uniquely associated with cluster skew are representatively selected here in larger font, a more quantitative analysis of all unique venues remains to be done and is mentioned in Future Directions.

CONCLUSION AND FUTURE DIRECTIONS

CONCLUSION

Venue analysis showed many things in common: discount and convenience stores, American restaurants, and Women's Stores were ubiquitous across clusters. Since retail and convenience stores are generic institutions and most likely to have foot traffic, it makes sense that people would use them across political persuasions.

Venue analysis also yielded a few differentiating factors and provided direction for Future Directions:

MOST importantly, there were some significant differences in political view by cluster, as shown in the boxplots. This is valuable in being able to consider and compare the venues in those clusters.

Next, the unique venues in those clusters almost all occurred in clusters between which the political ratio means were significantly varied. This also lends some credence to what the different venues might indicate about the voters, although a further quantitative analysis needs to be done to confirm these preliminary conclusions. This would involve precise %s of unique venues in all clusters and cluster ratios with statistically significant variance and venue differences vs nonsignificant.

Another vital factor to be examined in particular detail is the **purple/swing counties.**
This is the most pressing Future Direction. If Georgia was won by only 12,000 votes, then some of those crucial swing vote populations may be active in counties where very few votes decided the liberal or conservative overall skew. Much more indepth location and venue data neds to be soruced from these crucial counties, and trends studied.

It would also be helpful to contrast results with those from another supervised or unsupervised learning method. What about hierarchical clustering? do we see the same results? Is there more or less balance in how these 159 counties will fall into categories? Historical voting data from the last few Georgia elections could also be merged with 2020 data in order to form a more robust data pool for analyzing voting trends. The MIT Election Lab has several testing techniques to contrast[].

It is tempting to note that yoga studios and cuisines from other cultures are generally linked to a more liberal lifestyle[13-15], while sports, especially the classic American institution of baseball, have more conservative fans[11]. Fast food chains also generally fund Republican candidates, in part due to lighter conservative regulation of corporation labor[12] and the lack of red state interest in healthier eating options and physiology[15].

However, there are factors that need to be further examined.
Sociopolitically, how do these counties sort by:
Poverty index(which ranges of people below the poverty index, people slightly above, people solidly above, middle class, and upper middle class vote Republican and Liberal, and how does this interact with available venues in each cluster? And how accurately does it reflect red state attitudes on lifestyle choices? Is poor eating a function of cheap available food?

Ethnicity,gender, and voting practices: more African-Americans tend to vote liberal, a factor complicated by the history of voter suppression in Georgia. This practice has historically made poll access, voting requirements, and political information harder to access for people of color, women, and disadvantaged groups.
Gerrymandering, the practice of drawing county and district lines to abet a political advantage, has also had a similar effect.How do counties with a notable history of questionable voting practices, and African-American, people of color,--gender balance is also a factor--sort per cluster in venue content and political skew?
These and other factors when analyzed can provide a more comprehensive portrait of what happens in swing states.
———————

Hopefully this has been an interesting establishment of project parameter and baseline. Now to see how the Senate Georgia races play out!

Sources

1.https://www.cnn.com/2020/11/21/politics/georgia-presidential-election-recount/index.html
2.https://apnews.com/article/georgia-certify-election-joe-biden-ea8f867d740f3d7d42d0a55c1aef9e69
3.https://fivethirtyeight.com/features/how-georgia-turned-blue/
4.https://www.timesfreepress.com/news/local/story/2020/nov/14/how-georgia-blue/536081/
5.https://www.nielsen.com/us/en/solutions/capabilities/nielsen-political-solutions/
6.https://www.nytimes.com/2020/11/23/us/politics/ossoff-perdue-loeffler-warnock.html
7.https://results.enr.clarityelections.com/GA/105369/web.264614/#/summary
8.https://public.opendatasoft.com/explore/dataset/us-county-boundaries/table/?disjunctive.statefp&disjunctive.countyfp&disjunctive.name&disjunctive.namelsad&disjunctive.stusab&disjunctive.state_name&refine.statefp=13
9.https://state.1keydata.com/state-population-density.php
10.https://foursquare.com/p/explore-georgia/40090318/list/georgias-best-bbq
11.https://www.xminstitute.com/blog/baseball-fans-lean-right/
12.https://www.eater.com/2020/10/21/21505080/mcdonalds-wendys-political-donations-trump-biden
13.https://theconversation.com/partisan-divide-creates-different-americas-separate-lives-122925
14.https://www.usatoday.com/story/news/politics/2017/05/29/donadl-trump-may-be-driving-a-yoga-transformation-liberal/102159616/
15.https://mashable.com/2011/05/25/political-eating/