Cluster classification of safe places in the US to reopen Starbucks*

Balachandar Guduri
4/20/2020
(*This project is chosen voluntarily for a course to earn IBM Data Science Professional Certificate)

---

1. Introduction

The pandemic due to COVID-19 has been significantly affecting not only human lives but also the economy of government, industries, and restaurants globally. By 18th April, 2020, a few countries have reached a plateau in terms of COVID-19 cases and the governments are started investigations to opening up their countries to run the economy with an assumption that the worst situation has been just handled. Recently, the president of the US, Mr. Trump, has announced that the Whitehouse committee is looking forward to "Opening up America Again" by lifting the lockdown and relaxing social distancing measures in a few states in the first week of May [1].

1.1 Business problem:

As data analytic enthusiast and certified Barista at Starbucks, it motivates me to inspect the safer locations in the US which were not affected by COVID-19 for reopening a particle business. The largest coffee business giant, **Starbucks Inc.**, sees 47% drops in second-quarter earnings from Coronavirus hit [2]. Like many restaurants, Starbucks closed many stores in the US for two weeks and has been operating drive-thru and delivery only in some places in aid to stop the spread of the virus [3]. Now the US government is focusing on relaxing social distancing in some places, which required further investigation by Starbucks to inspect which are safe places to operate so that it ensures the safety to its customers. Thus the objective of this project is to find clusters of safe locations in the US to reopen **Starbucks** stores.

1.2 Point of interest:

The safety and the satisfaction of customers should be of utmost importance to any business or industry.

1.3 Deliverables:

1. Safe counties in the US to operate or reopen the business
2. Region based clusters of Starbucks venues located in those safe counties for reopening

2. Data

The work focused on the counties of the United States. In this work, three datasets have been used.

Dataset 1: The county-wise COVID-19 cases in the US

Source: https://github.com/nytimes/covid-19-data/blob/master/us-counties.csv

For the USA, New York Times has released and has been daily updating state-wise and county-wise cases of COVID-19 in a github repository [4]. The NY Times is compelling the record of registered cases due to this ongoing pandemic from local government, health department, and hospitals. The dataset is released in the public interest to better understand the outbreak. The county-wise dataset is considered for this analysis, which includes COVID-19 cases from 21st Jan to 18th April 2020. As listed in Table 1, this dataset consists of 6 features - date, county, state, FIPS id, COVID 19 cases, and number of deaths.

Table 1: County-wise COVID-19 cases in the US

	date	county	state	fips	cases	deaths
0	2020-01-21	Snohomish	Washington	53061.0	1	0
1	2020-01-22	Snohomish	Washington	53061.0	1	0
2	2020-01-23	Snohomish	Washington	53061.0	1	0
3	2020-01-24	Cook	Illinois	17031.0	1	0
4	2020-01-24	Snohomish	Washington	53061.0	1	0
...	...	...	...	...	...	...
70213	2020-04-18	Sublette	Wyoming	56035.0	1	0
70214	2020-04-18	Sweetwater	Wyoming	56037.0	10	0
70215	2020-04-18	Teton	Wyoming	56039.0	62	0
70216	2020-04-18	Uinta	Wyoming	56041.0	6	0
70217	2020-04-18	Washakie	Wyoming	56043.0	4	0

The total number of counties collected for COVID-19 cases is 2704. As per 2016, however, there are 3007 counties in the US. It seems COVID-19 data is not available for nearly 300 counties and these counties have been excluded from this study. This work only focuses on county-wise cases registered by April 18th and a total of 590123 cases are reported in the US.

Dataset 2: County-wise geo JSON data file for visualization

Source: https://raw.githubusercontent.com/python-visualization/folium/master/examples/data/us_counties_20m_topo.json

The visualization of these counties is carried out using **folium** library of Python with 'county_geo JSON' data file. For analysis and visualization, these counties are characterized into the following 5 categories as shown in Figure 1:

1. *Black county:* No COVID-19 data is available
2. *Safe county:* COVID-19 cases are less than or equal to 10
3. *Yellow county:* Number of cases greater than 10 and less than or equal to 1000
4. *Orange county:* Number of cases greater than 1000 and less than or equal to 10000
5. *Red county:* Number of cases greater than 10000

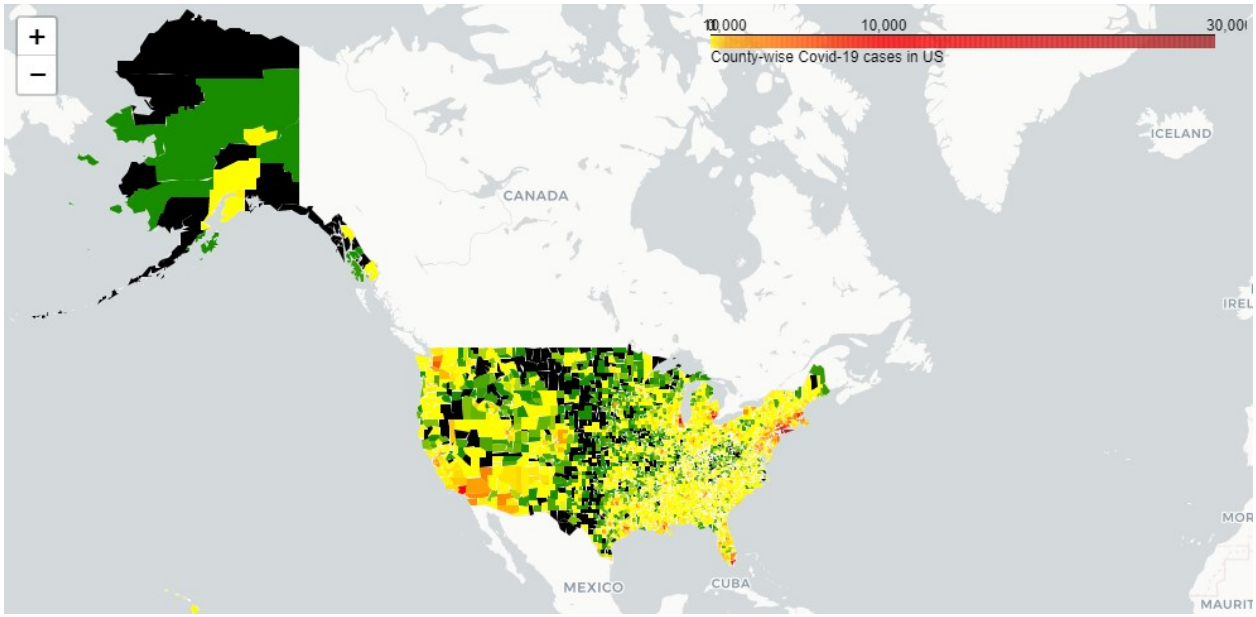


Figure 1: County-wise distribution of COVID-19 cases in the US.

Dataset 3: Venue locations of Starbucks using Foursquare API

Source: https://developer.foursquare.com/places

Foursquare is a very popular location-based networking website. It has been used by millions of users to explore nearby places and also been actively using for location-based data analyses by many researchers and analysts. Before using the Foursquare for my analysis, latitude and longitude of each safe county is obtained using 'geocoder' from GeoPy library. Then the Starbuck's venue locations (latitude and longitude) within the radius of 15 km of each safe county are obtained using Foursquare APIs. Since the free account of Foursquare is used for this work, the number of venues for a specific query (i.e. Starbuck) for each county is limited (max

50). The number of calls to be made in a day is also limited. Therefore, I had to run the code multiple days and the corresponding results are merged into a new data frame.

The ambiguity and misinformation of the venue results from Foursquare are inspected thoroughly. For example, using 'unique()' function the wrong data entries 'Languedoc-Rousillon' in the column of US states as shown in Fig. 2 are identified and corresponding data entries are discarded.

```
starbucks_venues['State'].unique()
```

```
array(['AL', 'AZ', 'AR', 'CA', 'CO', 'FL', 'GA', 'Georgia', 'NC', 'ID',
       'OR', 'Illinois', 'IL', 'KY', 'IN', 'Iowa', 'IA', 'KS', 'LA', 'MI',
       'MN', 'MS', 'MO', 'Missouri', 'NE', 'NM', 'VA', 'North Carolina',
       'SC', 'OH', 'Languedoc-Roussillon', 'Ohio', 'OK', 'PA', 'SD', 'TN',
       'TX', 'Texas', 'UT', 'VT', 'NY', 'Virginia', 'MD', 'D.C.', 'WA',
       'WV', 'WI', 'Wisconsin'], dtype=object)
```

Figure 2: Array of states obtained from Foursquare API using Starbucks as query.

For some counties, a few of Starbucks coordinates obtained from the Foursquare are located in the neighboring counties, for example, 598, 599 and 600 cases are shown in Table 2. This problem is solved by obtaining and verifying the FIPS codes from Starbucks's latitude and longitude coordinates with the FIPS codes in the COVID-19 data set. The results have been verified to make sure that each one of Starbuck venues is located into the corresponding county used for Foursquare API.

Table 2: Sample data of Starbucks venue location from Foursquare

	Requested county	Requested FIPS	Category ID	Venue name	Venue latitude	Venue longitude	FIPS	State
0	Bullock	1011	5793e354498e922ae570bdf3	Starbucks	32.014877	-85.746056	1011	Alabama
1	Graham	4009	4fc9457fd4f24895b4467c83	Starbucks	32.835233	-109.734477	4009	Arizona
2	Graham	4009	4c8fdf1c590ab1f7ac93e27d	Starbucks In Safeway	32.835672	-109.734042	4009	Arizona
3	Baxter	5005	5637f89dcd104b1868f1d572	Starbucks	36.349273	-92.371507	5005	Arkansas
4	Baxter	5005	5637f82ecd1099bc2507a5e5	Starbucks	36.347948	-92.371883	5005	Arkansas
...	...	...	...	...	...	...	...	...
598	Vilas	55125	549db741498ea7de5014f63e	Starbucks	43.075622	-89.386640	55025	Wisconsin
599	Vilas	55125	5a4a7a59bcbf7a68d9d7e169	Starbucks	43.025453	-89.417466	55025	Wisconsin
600	Vilas	55125	4cc96d214650a35ddb358e1e	Starbucks (inside The Sheraton)	43.047460	-89.373318	55025	Wisconsin
601	Washburn	55129	4b8187f2f964a520deac30e3	starbucks	45.896496	-91.827999	55129	Wisconsin
602	Wood	55141	57fd457b498e08976645bd16	Starbucks	44.017366	-90.508795	55081	Wisconsin

Python libraries used:

1. *Numpy*: For array operations
2. *Pandas*: For data manipulation and analysis
3. *Folium:* For geo map visualization
4. *Matplotlib*: For plotting data

5. *Sklearn*: For k-means clustering
6. *Geopy and geocoder:* For converting address to coordinates
7. *JSON:* For handling JSON files

3. Methodology

Using data analytics and machine learning tools, this problem is solved in two stages. In the first stage, the investigation for the safe counties which have a very low number of COVID-19 cases is carried out using Exploratory Data Analysis. In the second stage, using Foursquare API and machine learning tools, the clusters of Starbucks located in the safe counties in the US are classified.

3.1 Exploratory data analysis

Exploratory data analysis is carried out on county-wise COVID-19 cases by April 18th. In this analysis, a summary of the statistical parameters such as count, mean, standard deviation, max and min, and percentage of quartiles are estimated for safe, yellow, orange and red counties using python libraries. These parameters have provided strong insights into data patterns of COVID-19 cases. Next data visualization is carried out using **folium** and **matplotlib** libraries to inspect the data with the help of choropleth maps and bar charts. The corresponding results of this analysis are presented in the Results section.

3.2 Modeling

An unsupervised machine-learning algorithm such as *k*-mean clustering is used to form clusters of Starbucks locating the safe counties of the US. The *k*-means clustering is simple to use and a very popular algorithm for the classification of clusters. With this clustering, the Starbucks' locations are portioned into *k* clusters in such a way each Starbucks belongs to the cluster with the nearest mean. Following steps are carried out to classify the clusters:

1. Standardization of venue location coordinates of Starbucks is carried out using 'StandardScalar()' function from 'sklearn' library.
2. Elbow criterion is used to find the optimum number of clusters using distortion. The Euclidean distance metric is used.
3. With the optimal value of k, the cluster classification of Starbucks is done.
4. Classified clusters are inspected and corresponding results are presented in the Result section.

4. Results

4.1 Exploratory data analysis

A total of 7 counties from California, Illinois, Michigan, New Jersey, and New York have cases more than 10000. As we know, New York had recorded the highest number of COVID-19 cases. The number of counties and other statistical parameters of red, orange, yellow and safe counties are listed in Table 3 and the corresponding distribution of these counties are presented in the US map as shown in Figure 3.

Table 3: Statistical parameters of red, orange, yellow and safe counties in the US

Parameter	Red county	Orange county	Yellow county	Safe county
Count	7	97	1473	1153
Mean	19507.4	3009.9	106.5	4.1
Std	7050.81	2389.1	155.8	2.7
Min	12021	1003	11	1
25%	12817	1408	20	2
50%	20395	2073	42	3
75%	24661	3659	114	6
Max	29180	9956	987	10

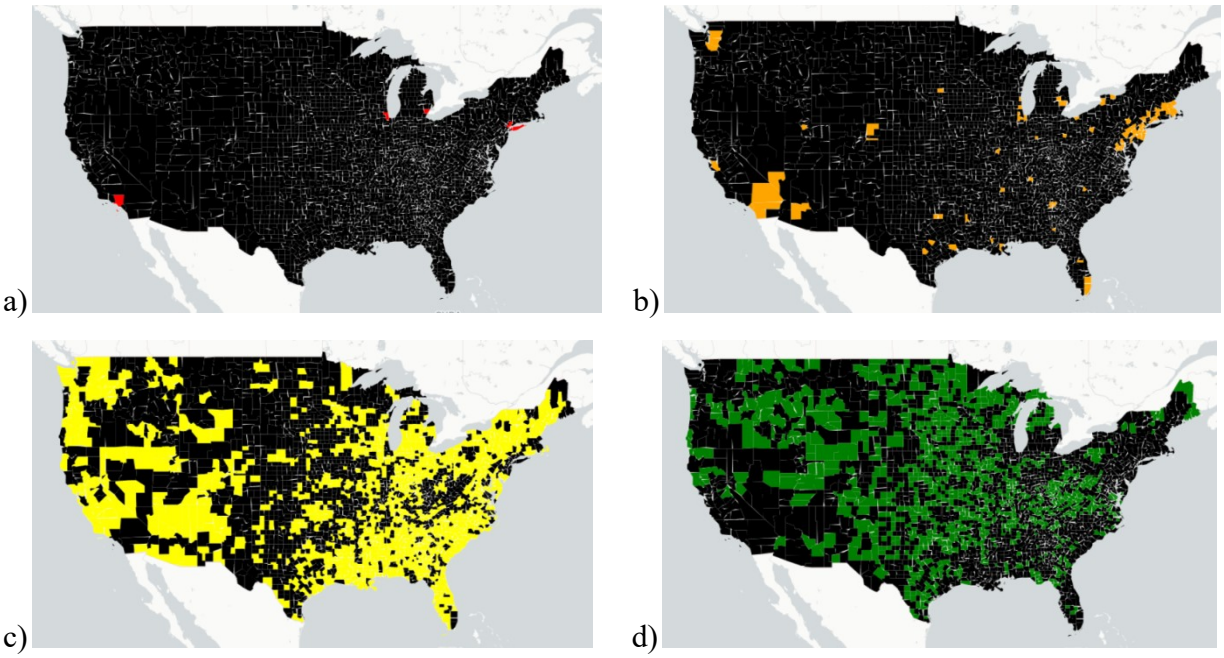


Figure: Choropleth map of red, orange, yellow and orange counties based on COVID-19 cases

There are 1153 counties, which are represented as safe counties since the number of cases ≤ 10. One can observe that most of the safe counties are located in the middle part of the US. The number-distribution of safe counties in each state is depicted in Figure 4. The Texas state has the highest number of safe counties i.e. 107.

4.2 k-mean clustering

In this work, '*k*-means++' approach is used for the initialization of parameters with a value of 12 to speed up the convergence. *k*-means clustering model divided the Starbucks venues into mutually exclusive region-based clusters. Figure 5 shows the variation of distortion for each value of k. With close inspection, there is a slight rate of change of variation that occurred at $k = 4$, which is considered as an optimal value.

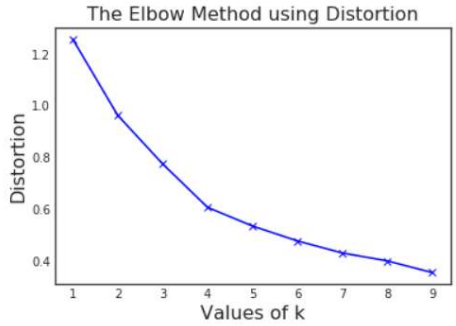


Figure 5: Variation of distortion with k

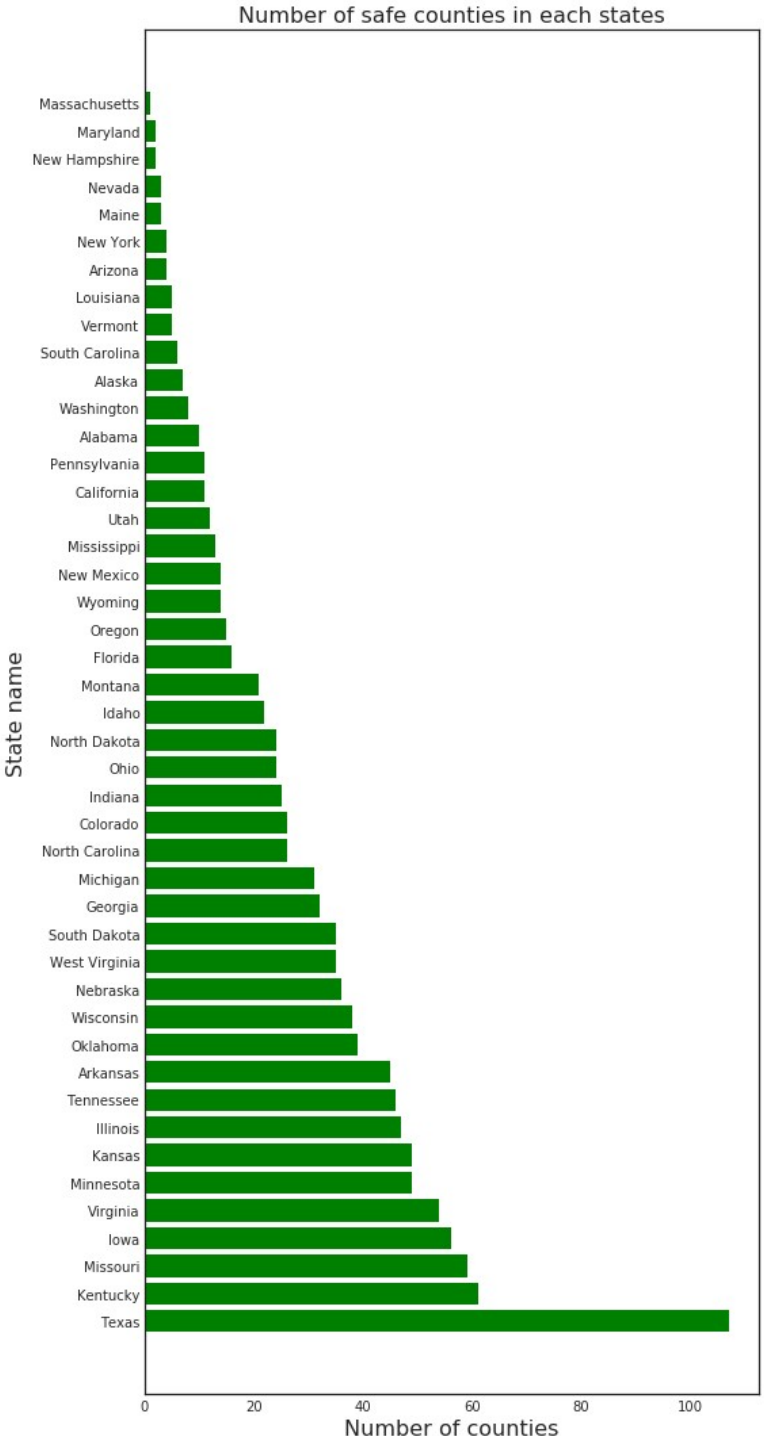


Figure 4: Bar chart distribution of safe counties in each state

Folium map in Figure 6 represents the color labeled markers of classified Starbucks on a Choropleth map of safe counties. Visually, we can clearly distinguish the clusters by region-wise.

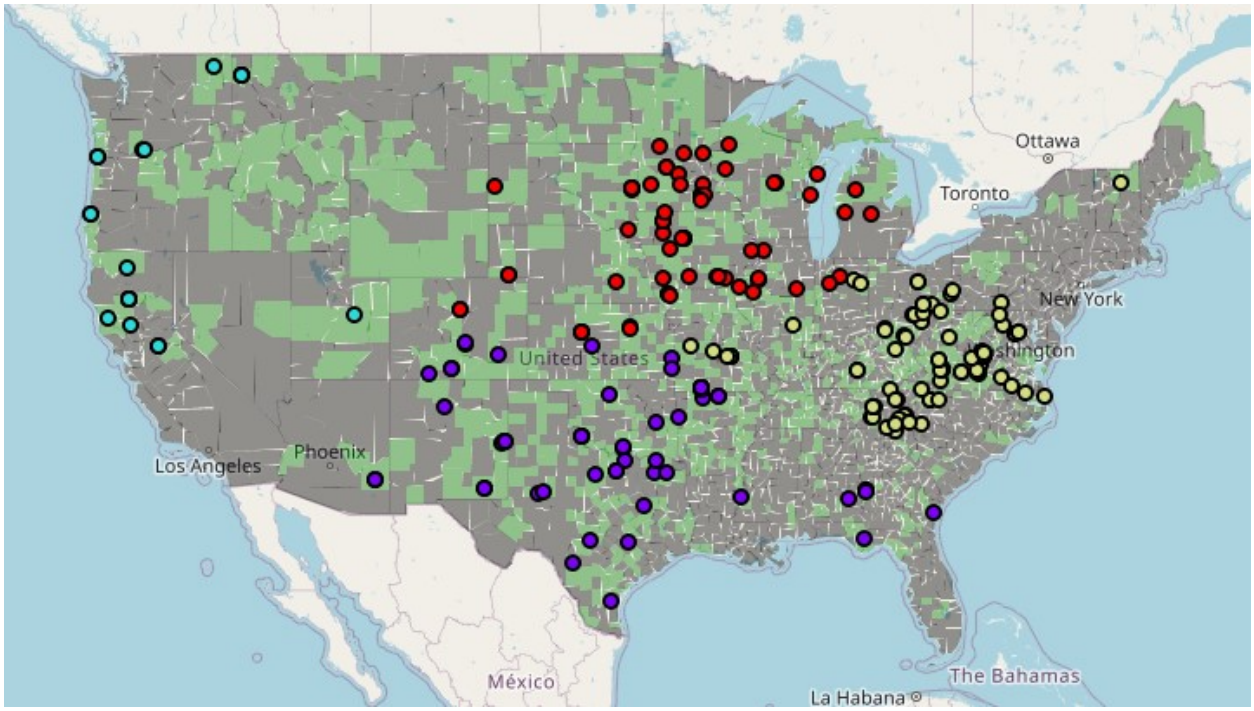
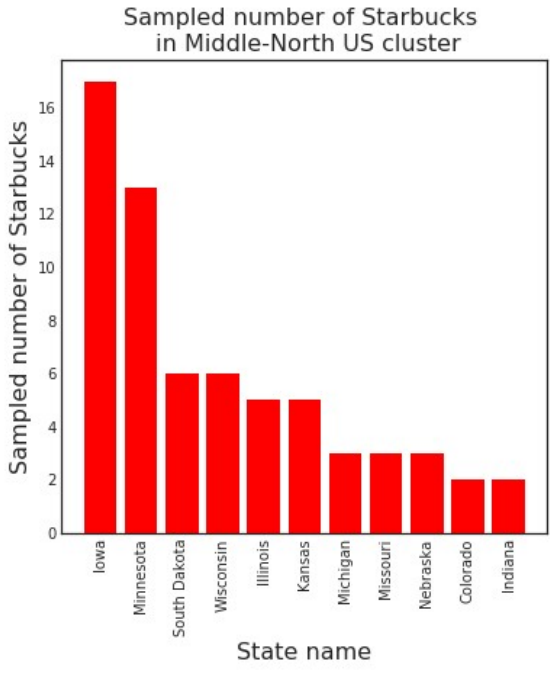
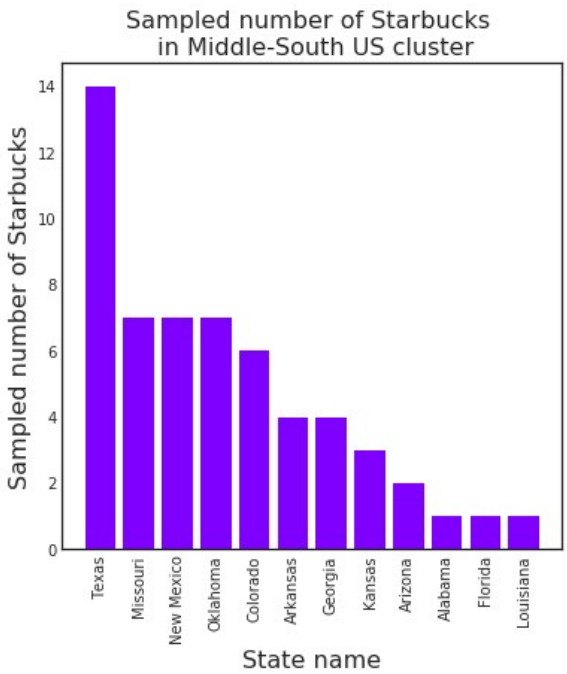


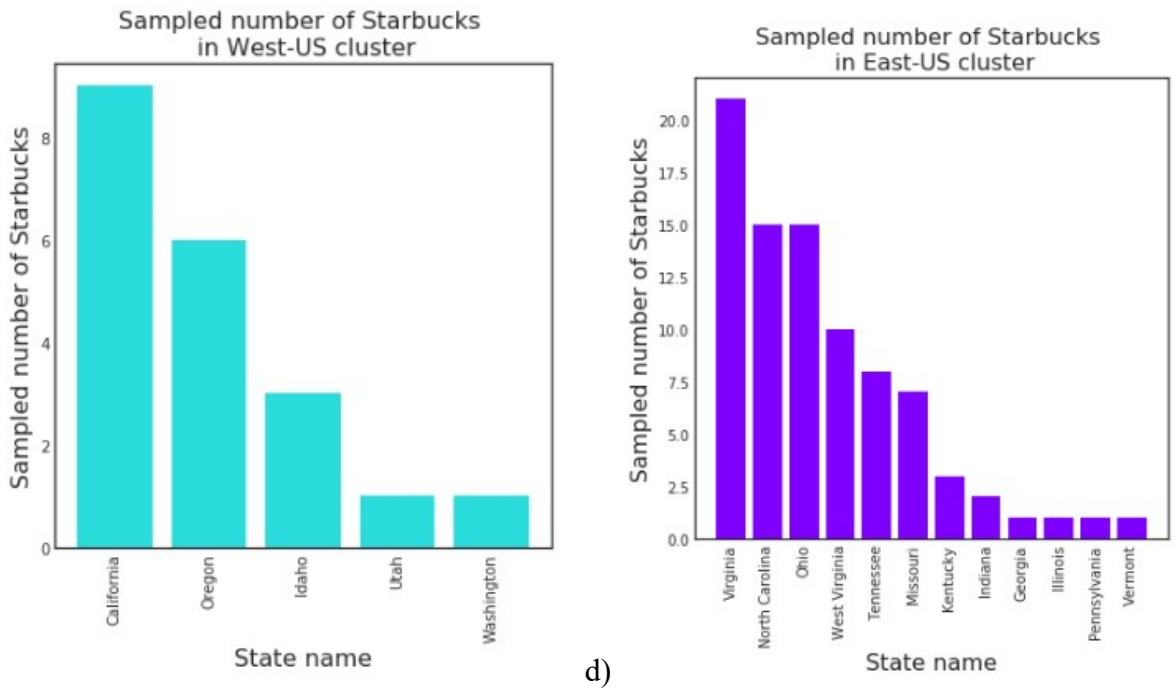
Figure 6: Classified clusters of Starbucks located in safe counties



a)



b)



c) d)

Figure 7: Bar charts of sampled number distribution of Starbucks in states in four clusters shown in Figure 6

These clustered data are inspected further and are presented as bar charts of sampled number distribution of Starbucks located safe counties for each state as shown in Figure 7.

1. For 0^{th} cluster (red) - **Middle-North US**, important states to focus on reopening Starbucks are **Iowa**, **Minnesota**, **South Dakota,** and **Wisconsin**.
2. For 1^{st} cluster (blue) - **Middle-South US**, important states to focus on reopening are **Texas**, **Missouri**, **New Mexico,** and **Colorado**.
3. For 2^{nd} cluster (cyan) – **West-Coast US,** important states to focus on reopening are **California,** and **Oregon**.
4. For 3^{rd} cluster (purple) – **East-Coast US,** important states to focus on reopening **Virginia**, **North Carolina**, **Ohio**, **West Virginia**, **Tennessee,** and **Missouri**.

5. Discussions

1. By April 18, 2020, nearly 1/3 of the US counties are recorded less than or equal 10 COVID-19 cases as shown in Figure 3(d).
2. From Figure 6, most of the Starbucks in the safe counties are located east-side half of the US.
3. The highest sampled numbers of Starbuck to reopen are located in Virginia, Iowa, North Carolina, Ohio, and Texas.

4. It is important to note that only one location request with a 15 km radius is made for each county despite the size of the county. In reality, therefore, the number of Starbucks stores in each county may be higher than the numbers represented in this work. The accuracy of the numbers in this work can be improved by using a grid-based location search approach for each county; however, this requires a significantly higher number of calls to make in Foursquare API. This can be done by using a premium membership account with Foursquare. Using a free account of Foursquare API, a sampled group of Starbucks venue points in each county are sufficient to provide meaningful insights about which counties would be allowed to reopen for the businesses.
5. In this analysis, the effects of other venues located near to Starbucks are not considered. However, one can understand that the Starbucks located near schools and universities should be extra cautious to re-open for business compared to those located in less crowded places or near high-ways.
6. The statement defined for Safe County in this work is subjective. One can also define on the basis if there is no increase in the number of cases for a certain number of days.

6. Conclusions

In this notebook, I carried out a rudimentary analysis to investigate safe counties to reopen Starbucks during an ongoing pandemic. First, safe counties in the US have been located using a dataset from the New York Times. Next, the venue locations of the Starbucks in the safe counties are obtained using data from Foursquare API. Finally, the cluster classification of Starbucks to reopen is done using *k*-means clustering. Four clusters of Starbucks have been located in the US, which are reported 10 or less COVID-19 cases, thus assumed as the safer locations to reopen for business.

On a personal note, I feel that Starbucks needs to put great emphasis on social distance measures to curb the spread of coronavirus in the near future. A few of them I can think of:

1. All the baristas and managers need to be tested daily or maybe twice in a week.
2. There should be strict enforcement on wearing a mask.
3. Limit the number of customers to enter.
4. Rearrange the seating and dining places to enable safe distance measures.
5. Just to accept the new normal lifestyle!

Thanks for visiting!

Stay home and stay safe!!

References:

1. https://www.yahoo.com/lifestyle/trump-announces-opening-america-again-010134209.html
2. https://www.reuters.com/article/us-health-coronavirus-starbucks/starbucks-sees-47-drop-in-second-quarter-earnings-on-coronavirus-hit-idUSKCN21Q3BG
3. https://edition.cnn.com/2020/03/21/business/starbucks-closing-covid-19/index.html
4. https://github.com/nytimes/covid-19-data