1. Introduction

1.1 Description of the Problem and Discussion of the Background

With the entry and spread of Covid-19 into the United States, a vast majority of the population has refrained from going outdoors, in a process we call quarantine. With the data collected by the New York Times [1], from the dates 1/21/2020 to 5/13/2020, there has been an estimated total of 1,194,663 cases and 63,999 deaths. With a population of around 328,239,523 estimated by the government census [2], which amounts to about a 3.6% Cases/Population and 0.019% Deaths/Population in the USA within this timeframe.

With the country and most of its states in complete lockdown towards the end of March, the rate at which Covid-19 was spreading had begun to drastically reduce. As a result, states have slowly started to open varying types of businesses, which inherently brings an influx of human interaction back into society. This brings an increase in the possible spread of Covid-19. Since I live in Washington state, this begged the question to me on how different counties might be dealing with the spread of Covid-19. More so, I wanted to find out what businesses would be safe/unsafe options for people to visit, as we go through the recovery process of quarantine.

What businesses within counties in Washington State are the most popular in correlation to their relative recovery process rates for Covid-19?

1.2 Data Description

USA Covid-19 Stats [1] – (5/13/2020) https://github.com/nytimes/covid-19-data

USA Population 2019 Stats [2] - https://www2.census.gov/programs-surveys/popest/datasets/2010-2019/counties/totals/

WA Covid-19 Stats [3] – (5/13/2020) https://www.doh.wa.gov/emergencies/coronavirus

Foursquare API Documentation [4] – (5/10/2020) https://developer.foursquare.com/docs/api-reference/venues/explore/

JSON for County Borders (FIPS geo-borders) [5] - https://raw.githubusercontent.com/plotly/datasets/master/geojson-counties-fips.json

[1] - New York Times has gone through government / state public datasets to compile a country-wide spreadsheet, in which we use the updated version from 1/21/2020 to 5/13/2020. The dataset comes with a daily update on cases and deaths within counties with

a spread of Covid-19. This data will be used to create a perspective on the spread of Covid-19 in the USA within the specified dates.

[2] - Contains Population Estimates in 2019 for all counties within the USA. This data will be used to aid in various percentage-based calculations for counties within the USA and Washington state.

[3] - Data from the WA state Department of Health. It contains data on a week by week account for relevant counties, regarding the count of cases, deaths, and correlating age buckets they belong to. This is used to depict a danger-ranking system for the counties within Washington state based off the week by week change in cases.

[4] - Foursquare will be used to highlight venues within the boundaries of each county that are most popular in the general week / time spectrum (not any specific time of day or day of week). This will showcase the most used / popular venues that are seen as safe or unsafe dependent on how the county was ranked in terms of weekly Covid-19 change rates for cases.

[5] – This is the set of boundary points for each of the counties within the USA. This is used to help graph varying labels and variables on a map of the USA.

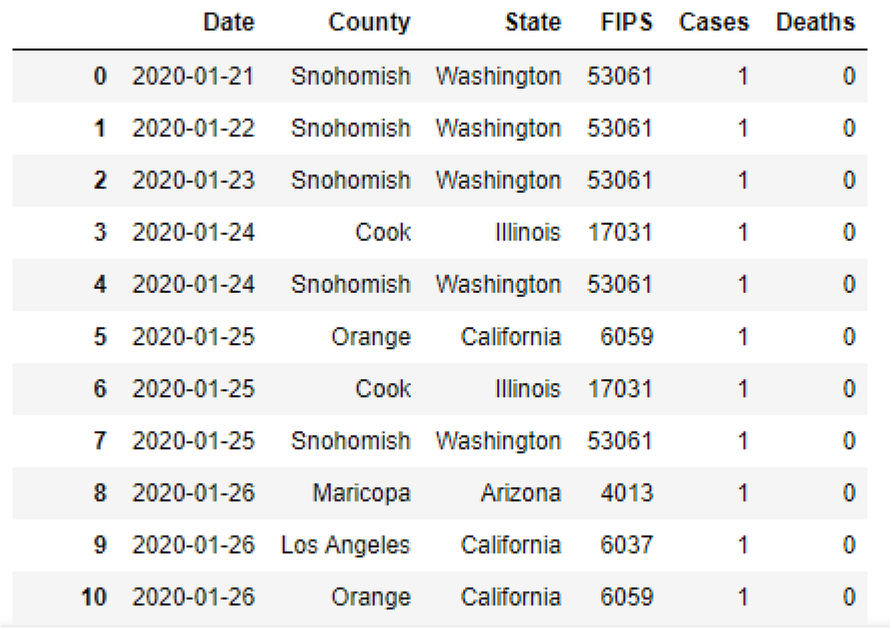
2. Methodology

2.1 USA Data

1. Setting up USA Covid-19 per County Data

The data provided by New York Times [1] comes in the format of containing the total counts of cases and deaths for relating counties on a day-by-day basis.

(Table 1)



	Date	County	State	FIPS	Cases	Deaths
0	2020-01-21	Snohomish	Washington	53061	1	0
1	2020-01-22	Snohomish	Washington	53061	1	0
2	2020-01-23	Snohomish	Washington	53061	1	0
3	2020-01-24	Cook	Illinois	17031	1	0
4	2020-01-24	Snohomish	Washington	53061	1	0
5	2020-01-25	Orange	California	6059	1	0
6	2020-01-25	Cook	Illinois	17031	1	0
7	2020-01-25	Snohomish	Washington	53061	1	0
8	2020-01-26	Maricopa	Arizona	4013	1	0
9	2020-01-26	Los Angeles	California	6037	1	0
10	2020-01-26	Orange	California	6059	1	0

In essence, this gives us a day by day look at how counties have been doing in terms of cases and deaths, from 1/21/2020 to 5/13/2020. We can use this to setup our data

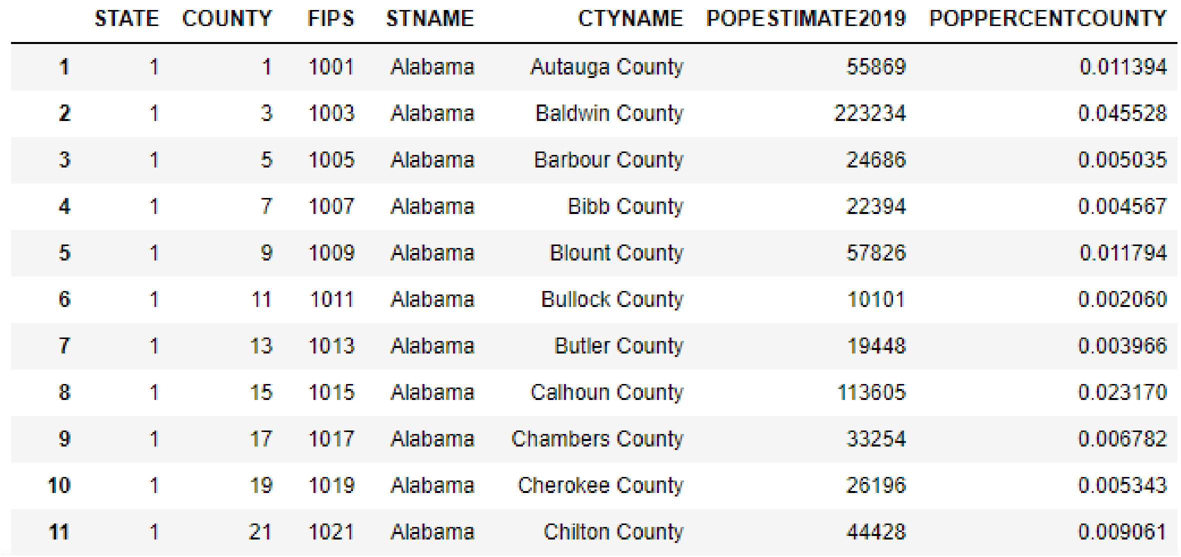
2. Setting up USA Population 2019 per County Data

These populations metrics are taken from the census statistics provided by the government [2]. The variables given are information concerning the county / location, along with the estimated population of that region in the year 2019. To help merge datasets, I created a FIPS value for the population data using some of its existing variables.

Next, I had to separate the population data between states and counties so that I did not count any of the population data twice. In addition, I added a County Population per State Population Percentage metric.

County:

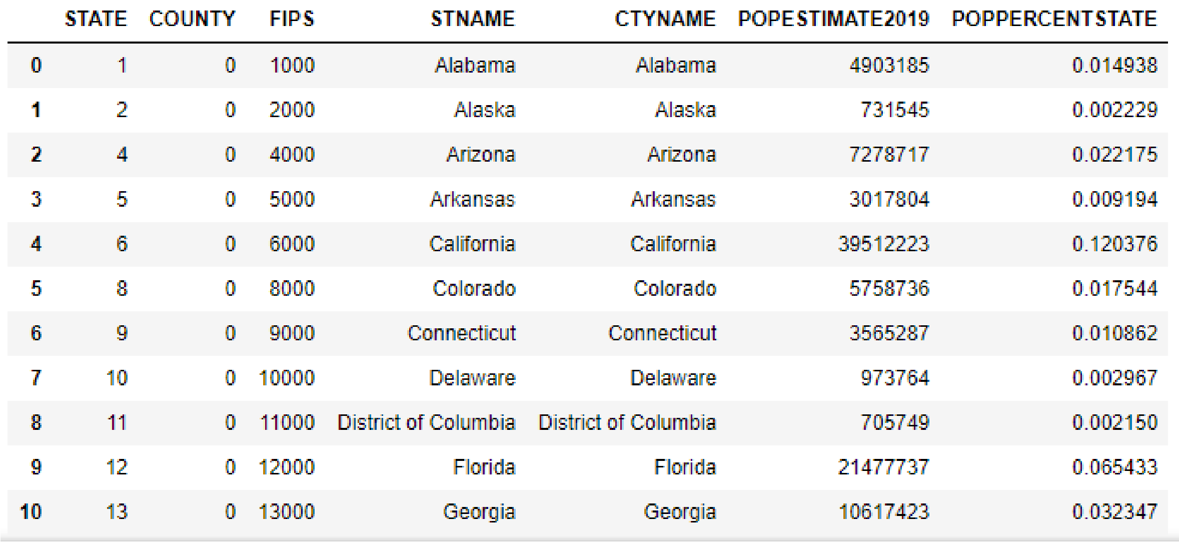
(Table 2)



	STATE	COUNTY	FIPS	STNAME	CTYNAME	POPESTIMATE2019	POPPERCENTCOUNTY
1	1	1	1001	Alabama	Autauga County	55869	0.011394
2	1	3	1003	Alabama	Baldwin County	223234	0.045528
3	1	5	1005	Alabama	Barbour County	24686	0.005035
4	1	7	1007	Alabama	Bibb County	22394	0.004567
5	1	9	1009	Alabama	Blount County	57826	0.011794
6	1	11	1011	Alabama	Bullock County	10101	0.002060
7	1	13	1013	Alabama	Butler County	19448	0.003966
8	1	15	1015	Alabama	Calhoun County	113605	0.023170
9	1	17	1017	Alabama	Chambers County	33254	0.006782
10	1	19	1019	Alabama	Cherokee County	26196	0.005343
11	1	21	1021	Alabama	Chilton County	44428	0.009061

State:

(Table 3)



	STATE	COUNTY	FIPS	STNAME	CTYNAME	POPESTIMATE2019	POPPERCENTSTATE
0	1	0	1000	Alabama	Alabama	4903185	0.014938
1	2	0	2000	Alaska	Alaska	731545	0.002229
2	4	0	4000	Arizona	Arizona	7278717	0.022175
3	5	0	5000	Arkansas	Arkansas	3017804	0.009194
4	6	0	6000	California	California	39512223	0.120376
5	8	0	8000	Colorado	Colorado	5758736	0.017544
6	9	0	9000	Connecticut	Connecticut	3565287	0.010862
7	10	0	10000	Delaware	Delaware	973764	0.002967
8	11	0	11000	District of Columbia	District of Columbia	705749	0.002150
9	12	0	12000	Florida	Florida	21477737	0.065433
10	13	0	13000	Georgia	Georgia	10617423	0.032347

These depictions help us create a baseline for how I can equally evaluate counties when it comes to their population size and how cases and deaths correlate to that.

3. Merging Population with USA Covid-19 Data

Now that both datasets are prepared, we can merge them so that we can create more varying data metrics for the counties in the USA.

For our County population data, I separated the Cases and Deaths values and merged them with their matching FIPS value from the population dataset. The same method was applied for the State population data, except it was done by merging the Cases and Deaths based off the State Number.

<u>Looking at our data</u>:

Total C19 Cases in USA (1/21 to 5/13) – 1,194,663
Total C19 Deaths in USA (1/21 to 5/13) – 63,999
USA 2019 Population - 328,239,523

<u>Along with these statistics, I bring in 3 different data metrics</u>:

Cases of C19 per Population Percentage
Deaths of C19 per Population Percentage
Deaths by Cases Ratio

These metrics will help give us variables to rank states and their counties in terms of the rate of cases for Covid-19.

4. Graphing General Higher Risk Counties / States

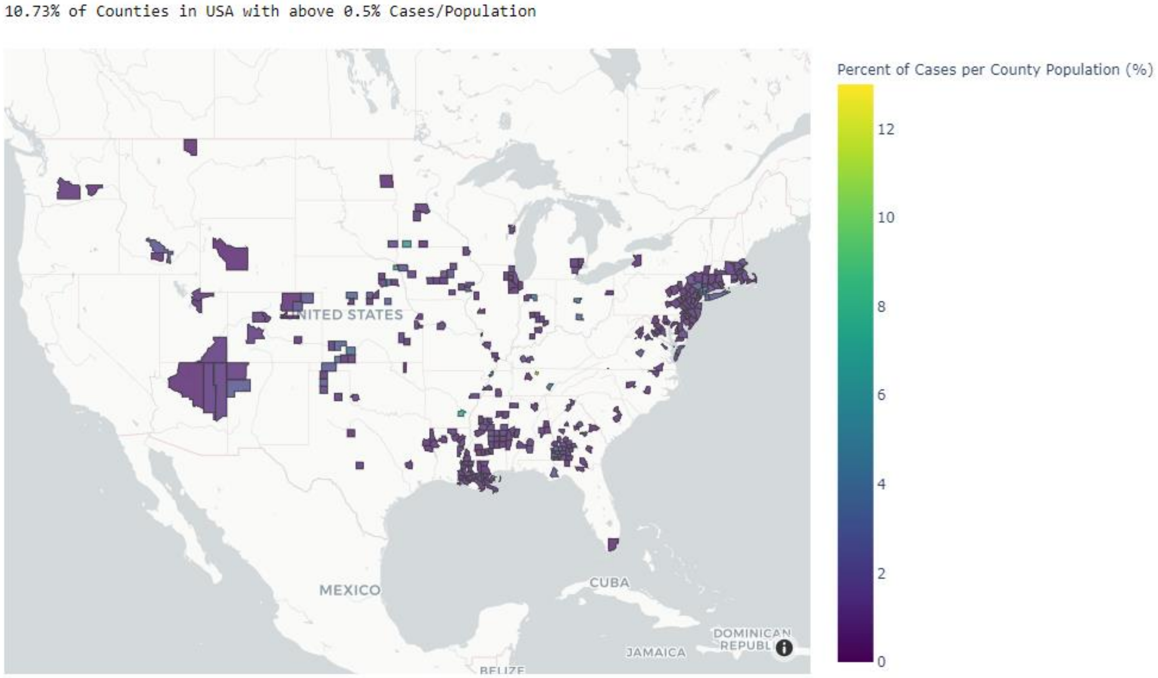
I used a geo-json file [5] with relating FIPS values to help graph the boundaries of the counties.

To graph variables within the merged dataset, I needed to make sure all FIPS values were 5 characters in length. I fixed this issue by adding a 0 to the front of any 4-character FIPS values.

Another step I took, was to remove any counties that had lower than 5 cases, since when dealing with percentages, could produce a skewed spectrum of values.

Cases/Population:
Counties in USA with above 0.1% Cases/Population: 49.48%
Counties in USA with above 0.5% Cases/Population: 10.73% (Graph 1)
Counties in USA with above 1.0% Cases/Population: 4.15%

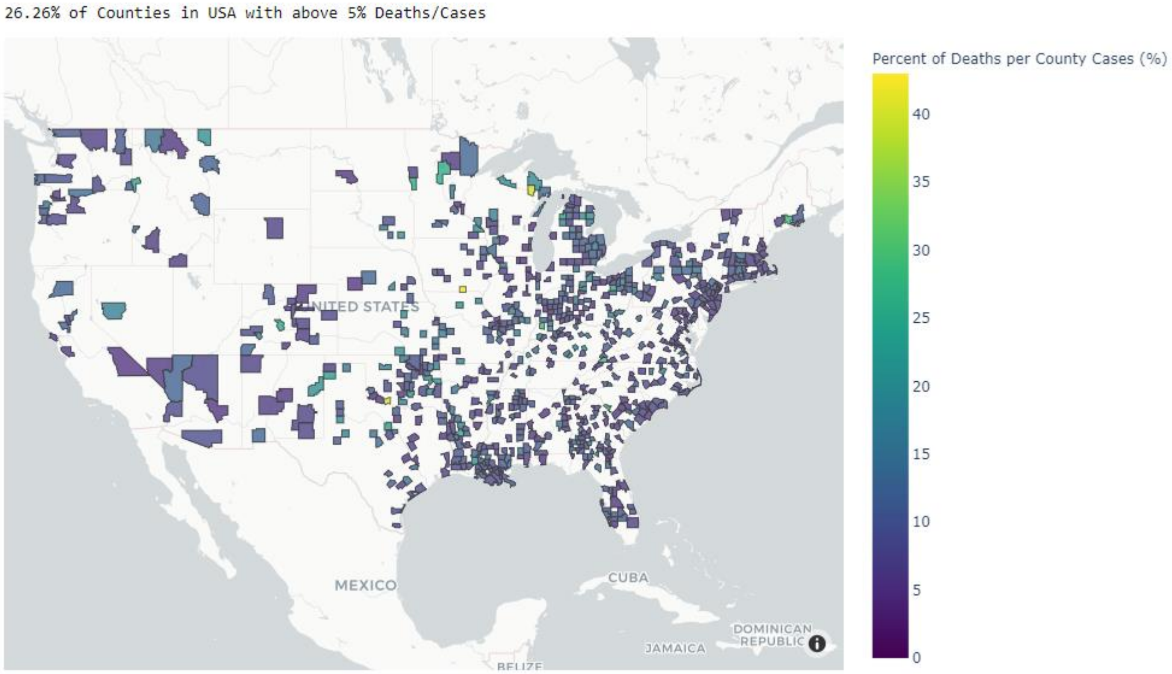
(Graph 1)


10.73% of Counties in USA with above 0.5% Cases/Population
Percent of Cases per County Population (%)
12
10
8
6
4
2
0
NITED STATES
MEXICO
CUBA
DOMINICAN REPUBL
JAMAICA
BELIZE

Graph 1 represents the counties and their corresponding percentages for counties where .5% of the county population tested positive for Covid-19.

Deaths/Cases:
Counties in USA with above 1% Deaths/Cases: 52.56
Counties in USA with above 5% Deaths/Cases: 26.26% (Graph 2)
Counties in USA with above 10% Deaths/Cases: 9.62

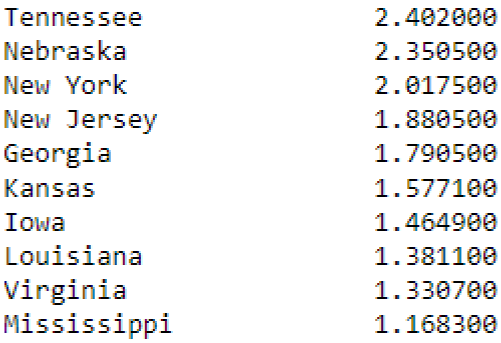
(Graph 2)


26.26% of Counties in USA with above 5% Deaths/Cases
Percent of Deaths per County Cases (%)
40
35
30
25
20
15
10
5
0
UNITED STATES
MEXICO
CUBA
JAMAICA
DOMINICAN REPUBLIC
BELIZE

Graph 2 represents the counties and their corresponding percentages for counties where 5% of the cases for Covid-19 resulted in death. This is an indication of areas that may contain people who are more vulnerable to Covid-19.

<u>Ranking of top 10 States by CCP percentages</u>:

(Table 4)

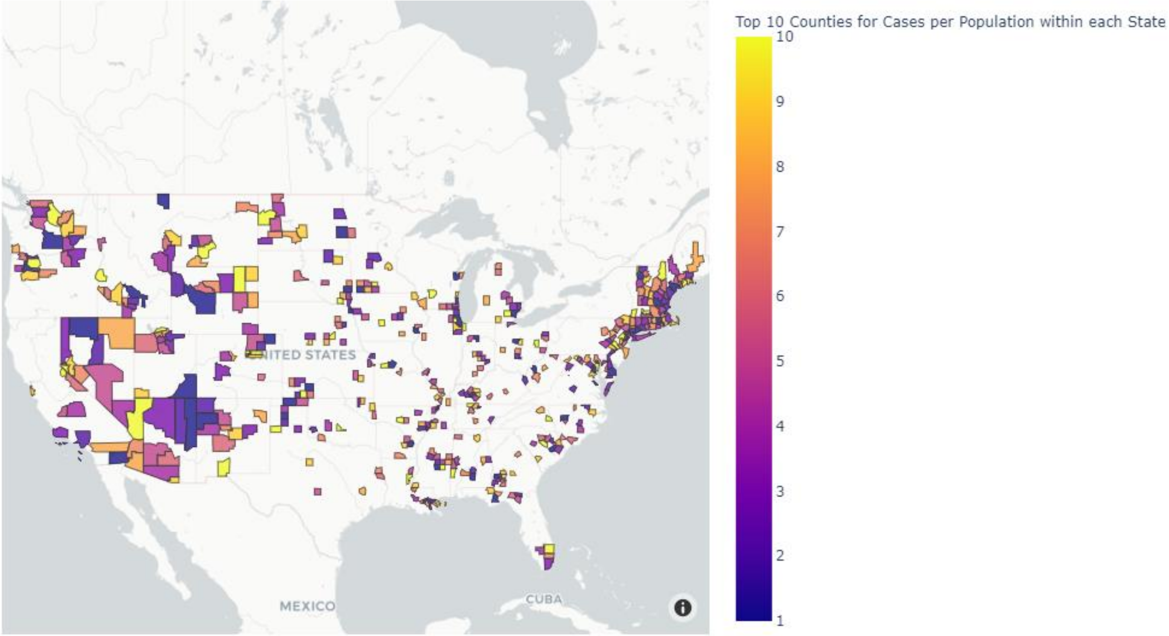


```
Tennessee           2.402000
Nebraska            2.350500
New York            2.017500
New Jersey          1.880500
Georgia             1.790500
Kansas              1.577100
Iowa                1.464900
Louisiana           1.381100
Virginia            1.330700
Mississippi         1.168300
```

Table 4 shows an ordered list of the top 10 States with their corresponding Cases per Population percentages. This is further illustrated by delving deeper into their relating counties.

Top 10 Counties per State:

(Graph 3)



Graph 3 represents the counties on a top 10 basis, relevant to how they rank versus other counties in their state, when it comes to Cases per Population. A lower rank signifies a lower percentage, while a higher rank signifies a higher percentage.

2.2 WA Data

5. Cleaning and Creating Metrics for WA data

Different from our USA data, the Washington state data [3] is based off a weekly timeseries, including two varying tables for cases and deaths.

To start off, I noticed that each of counties did not contain the same count of weeks. So to help create a unified structure that I can use later on to merge, I decided to populate the missing weeks for every county and include 0 as the entry value, whether for cases or deaths.

Now that I have both tables for WA Cases and Deaths, we can merge them and create metrics. As a baseline I chose to use the Cumulative Sum of cases and deaths.

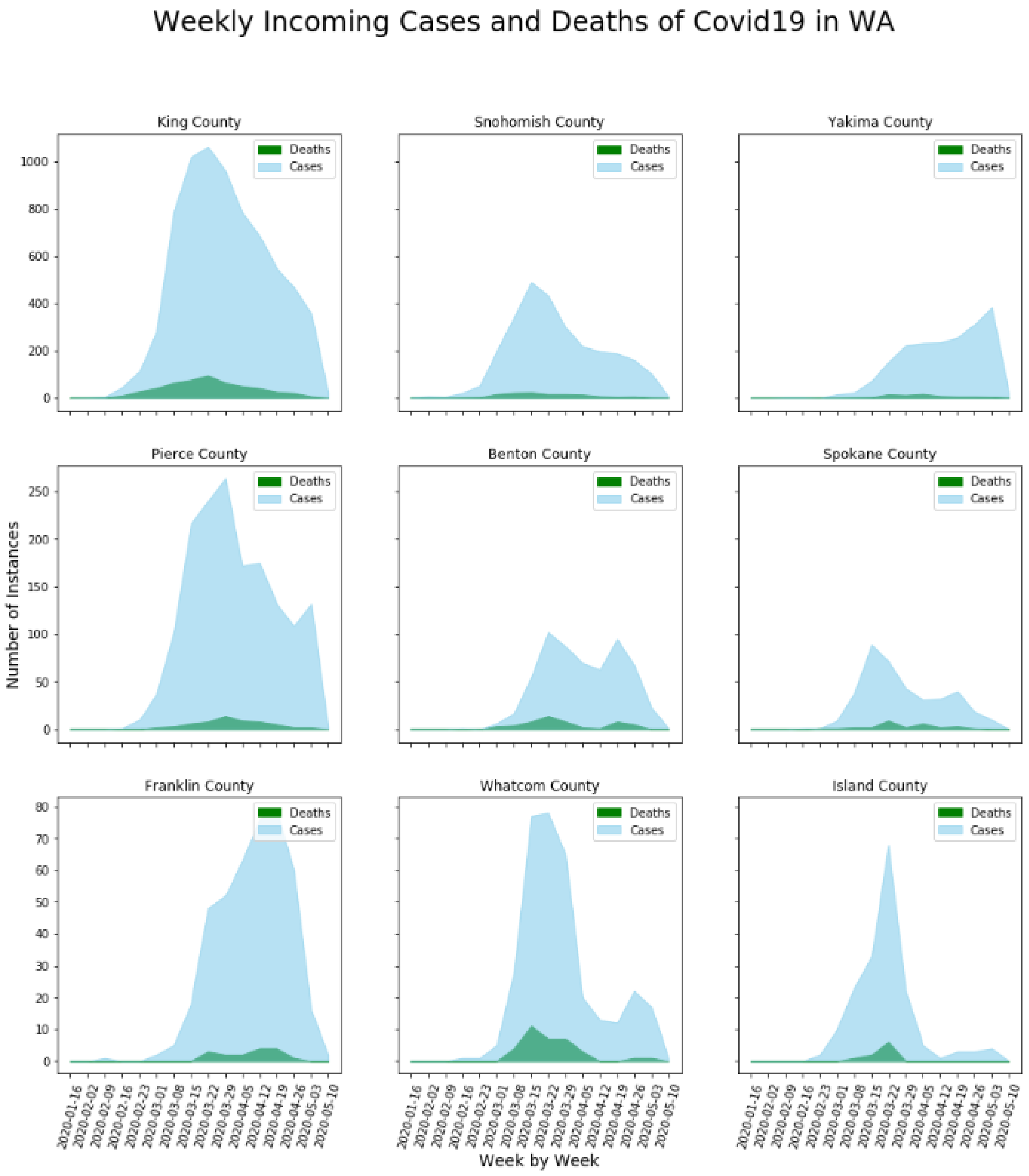
6. Graphing Covid-19 WA data

Since there 33 unique counties within this data, I decided graph the top 9 relative to the variable being looked at.

The current ranking of these countries (for the following graphs) are: King, Snohomish, Yakima, Pierce, Benton, Spokane, Franklin, Whatcom, and Island.

Weekly Incoming Cases and Deaths of Covid-19 in WA:

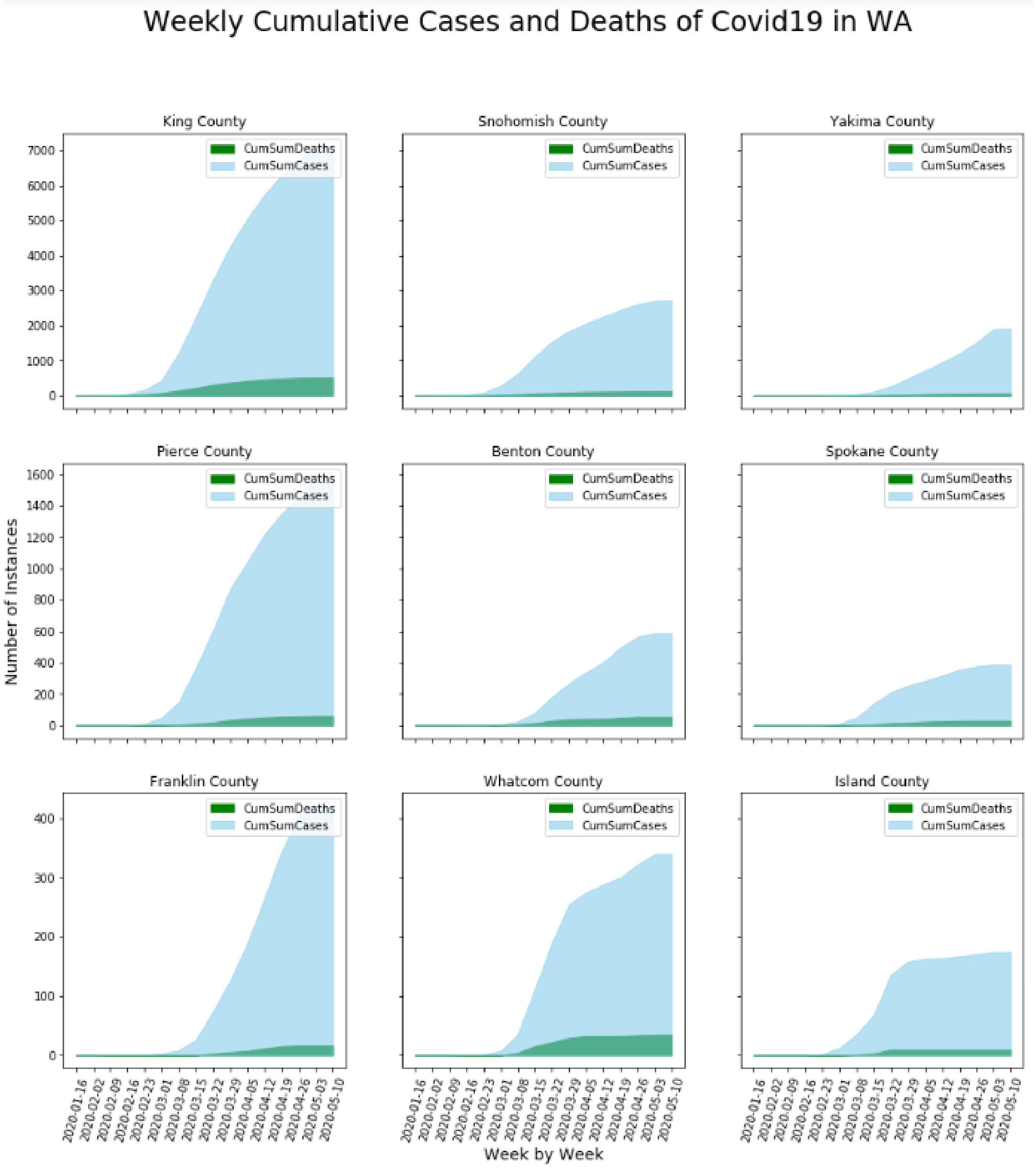
(Graph 4)



When it comes to sheer magnitude of incoming cases, Graph 4 shows King County is at the lead with a peak of 1,063 cases. This makes sense considering its population is the highest in Washington State at about 2,252,782. For the most part, many of these counties seem to have a significant drop off in cases over latter half of the time set, which we can call the set of recovery weeks.

<u>Weekly Cumulative Cases and Deaths of Covid-19 in WA</u>:

(Graph 5)



In graph 5, King County is at the top when it comes to cumulative cases with 7137 cases. Also illustrated in Graph 5, is the gradual slowdown in rate of cumulative cases towards the last few weeks, which we will take into consideration in the clustering process.

7. Clustering (K-Means) WA Counties

For this project, I am primarily dealing with metrics related to Cases. It is independent to Deaths, especially when it comes to our problem statement and the rate of spread for Covid-19.

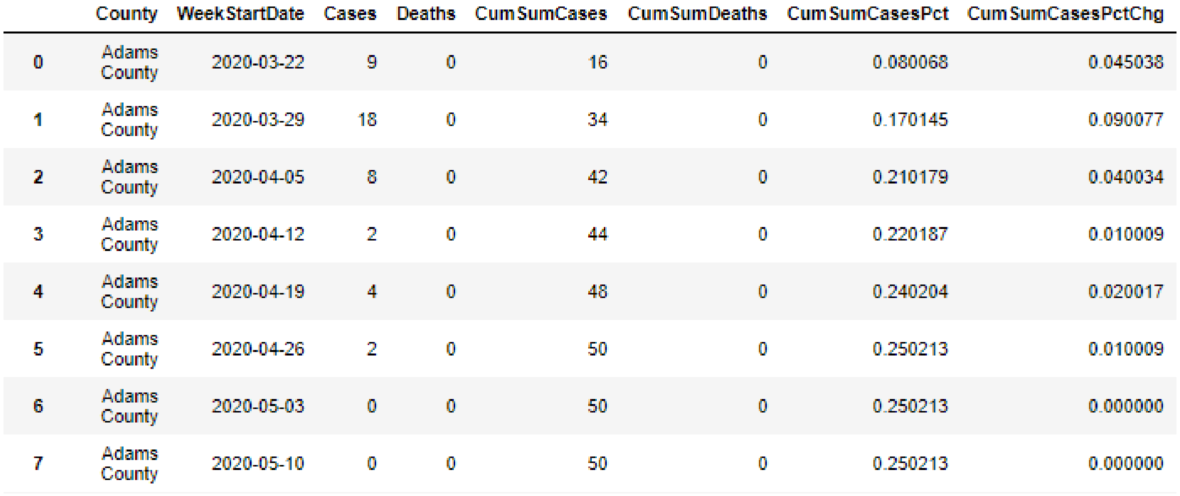
In terms of clustering these counties, quantifying the data in terms of the quantity of cases would not suffice, since there is such a wide spectrum of populations within these counties. As a result, I went along with creating a Cumulative Sum Percentage metric based off the counties Cases and Population. This helps better illustrate the relative rate of increase in cases per county.

However, when it comes to predicting the relative danger of a county for Covid-19 on a weekly basis, it would be better to analyze the percent change in the cumulative sum of cases. This way, there is a heavier weight on how varying counties might relate in terms of similar percent change rates every week.

Normally, I would cluster on the entire set of weeks within the dataset. However, since I am testing for how counties are dealing with the recovery process, I wanted to narrow the scope to the set of weeks that are within that realm. I did this by checking for the week in which counties had the highest incoming count of cases recorded. The data revealed the peak week to be 3/22/2020.

Now continuing with the clustering process, I set up a table consisting of the Percent Change in the cumulative sum of cases per population, for each county. The weeks were set up from 3/22/2020 to 5/10/2020.

(Table 5)

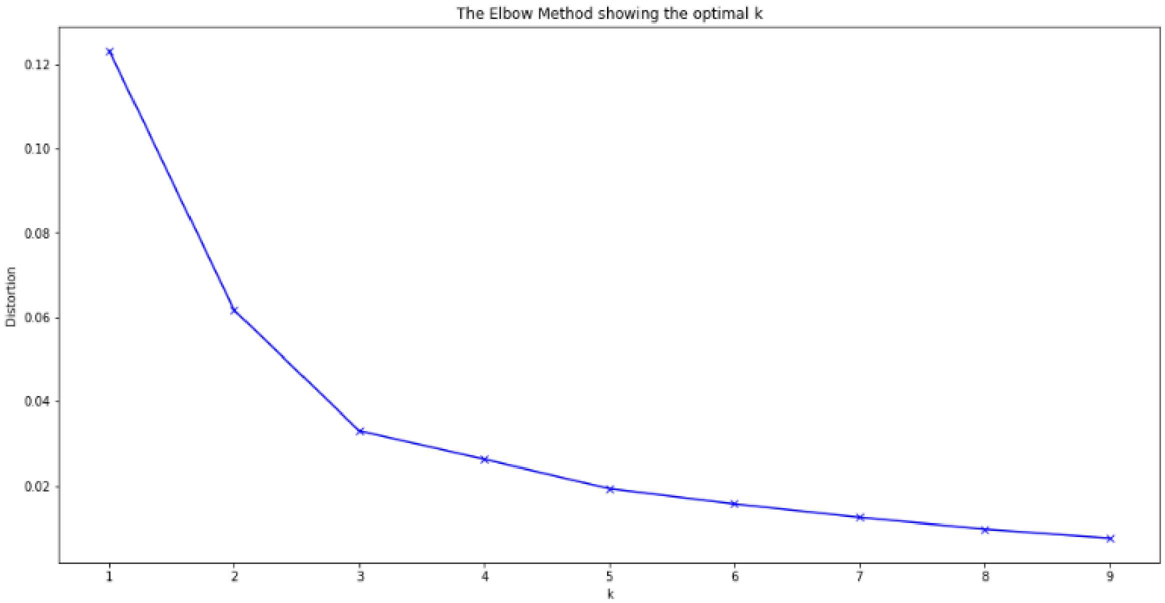


	County	WeekStartDate	Cases	Deaths	CumSumCases	CumSumDeaths	CumSumCasesPct	CumSumCasesPctChg
0	Adams County	2020-03-22	9	0	16	0	0.080068	0.045038
1	Adams County	2020-03-29	18	0	34	0	0.170145	0.090077
2	Adams County	2020-04-05	8	0	42	0	0.210179	0.040034
3	Adams County	2020-04-12	2	0	44	0	0.220187	0.010009
4	Adams County	2020-04-19	4	0	48	0	0.240204	0.020017
5	Adams County	2020-04-26	2	0	50	0	0.250213	0.010009
6	Adams County	2020-05-03	0	0	50	0	0.250213	0.000000
7	Adams County	2020-05-10	0	0	50	0	0.250213	0.000000

Clustering Counties based from (3/22/2020 to 5/10/2020):

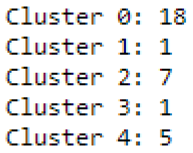
Having set up the input for our k-means clustering, I needed to figure out how many clusters I needed to use. I used the elbow method ranging from 1 to 9 clusters, which is shown in Graph 6. The result showed a k value of 5 to be the most optimal.

(Graph 6)



Here’s a count of the counties per Cluster:

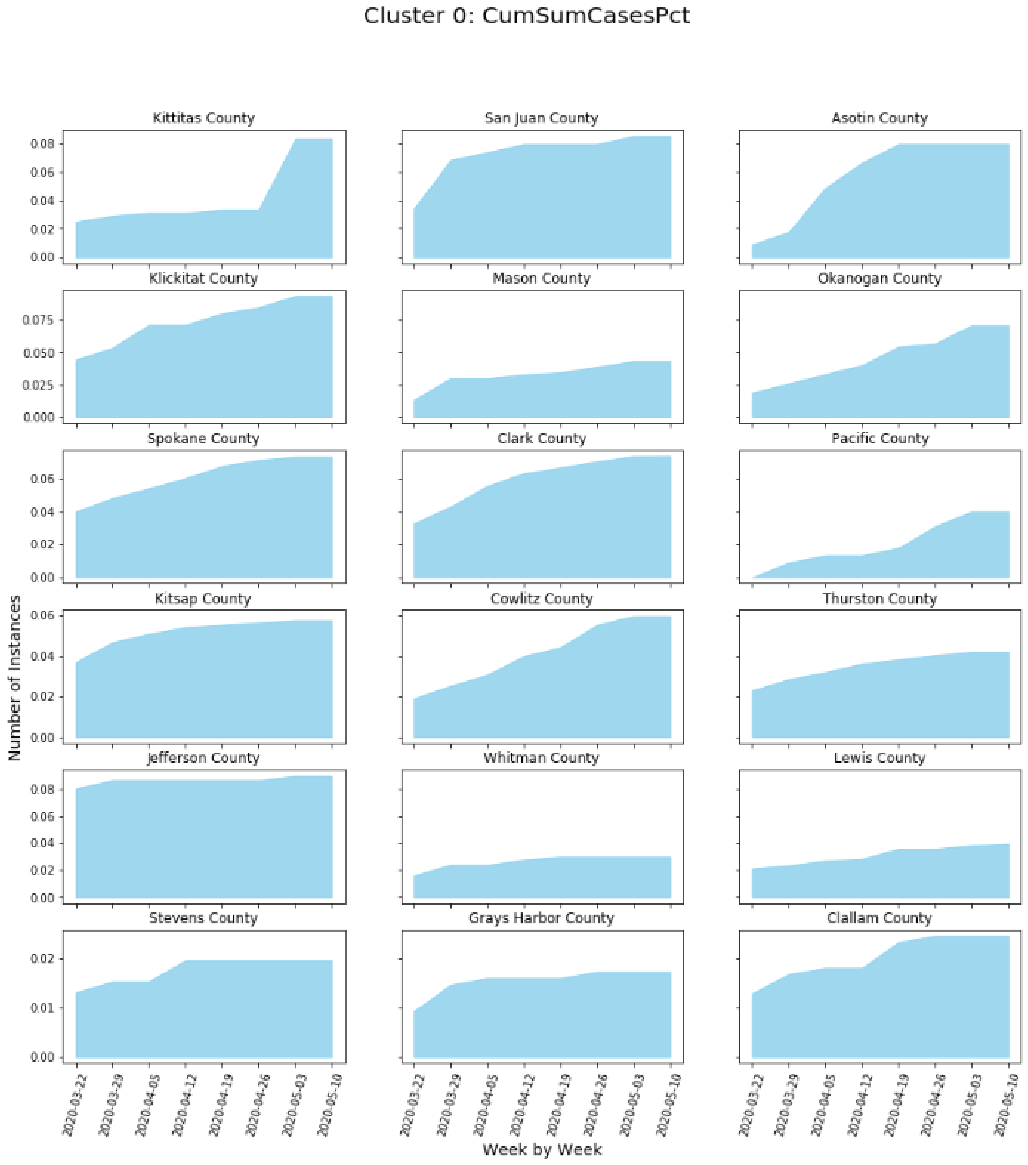
(Table 6)



```
Cluster 0: 18
Cluster 1: 1
Cluster 2: 7
Cluster 3: 1
Cluster 4: 5
```

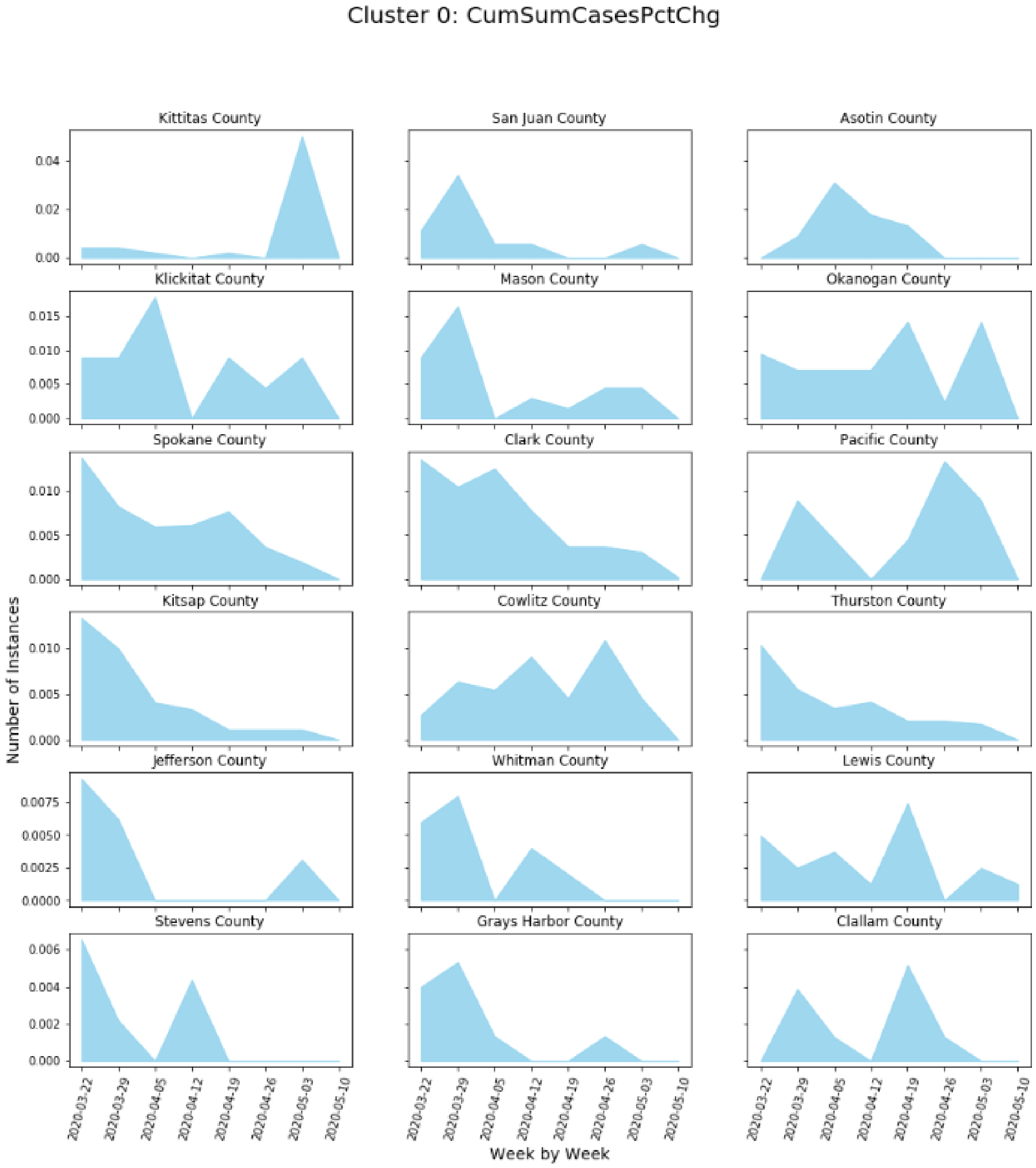
Example of Cluster 0: CumSumCasesPct

(Graph 7)



Example of Cluster 0: CumSumCasesPctChg

(Graph 8)



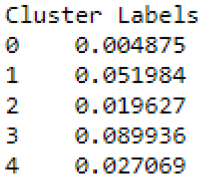
As you can see, Graph 7 and 8 illustrates counties (in Cluster 0) with similar changing rates when it comes to percentage changes in the cumulative sum of cases.

8. Geo-graphing of WA Counties based off Clusters

Now that we have grouped the counties of WA in 5 separate clusters, we can rank these clusters using two different statistical metrics.

a. Average Cumulative Sum of Cases – Percent Change

(Table 7)

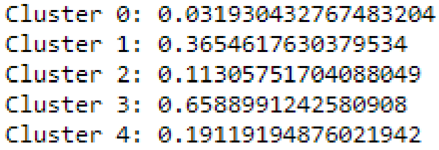


```
Cluster Labels
0    0.004875
1    0.051984
2    0.019627
3    0.089936
4    0.027069
```

This is the same metric we used to cluster the counties in WA state. In this result, each value is the average of all the county values for each separated cluster.

b. Average Max-Min Delta for Cumulative Sum of Cases – Percent

(Table 8)



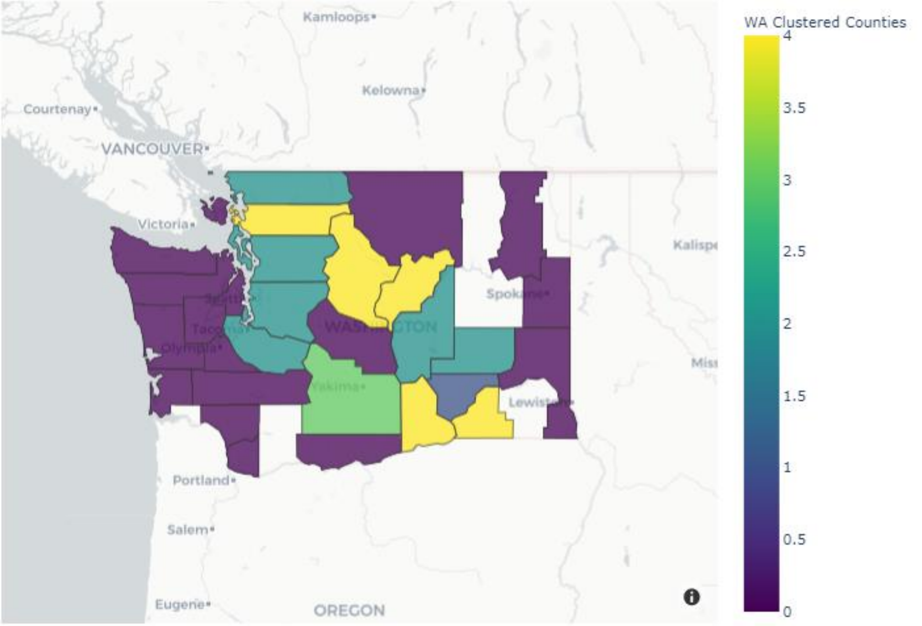
```
Cluster 0: 0.031930432767483204
Cluster 1: 0.3654617630379534
Cluster 2: 0.11305751704088049
Cluster 3: 0.6588991242580908
Cluster 4: 0.19119194876021942
```

Calculating the Max-Min Difference for each of our countries averaged out between our clusters helps illustrate the relative range for how cases have escalated within our narrowed time frame starting on 3/22/2020.

Using these statistical metrics, we can rank our clusters from lowest risk to highest risk (Cluster Number): 0, 2, 4, 1, 3

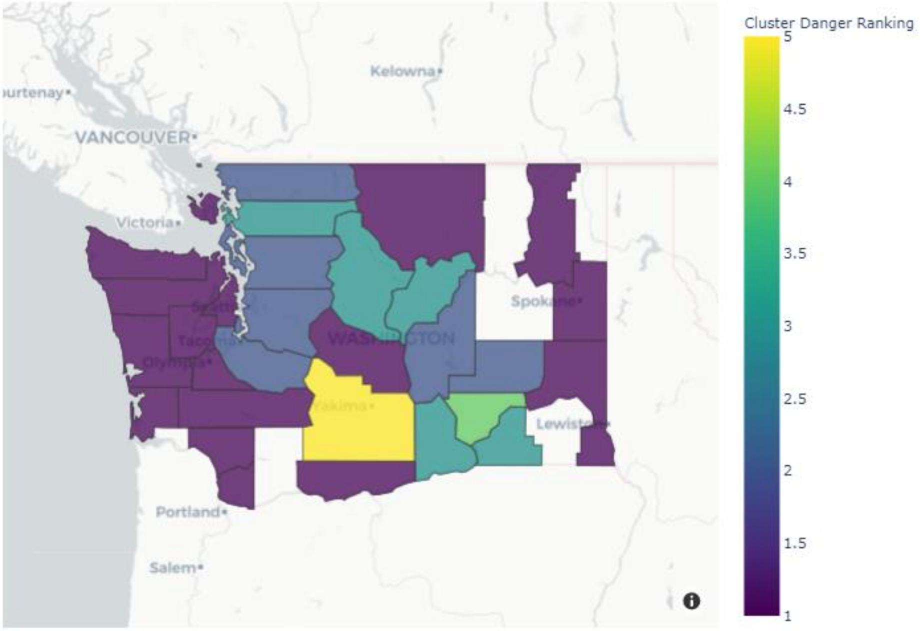
Graph of Counties based off Cluster Number:

(Graph 9)



Graph of County-Clusters based off Recovery Process/Rate of Covid-19 (3/22/2020 to 5/10/2020):

(Graph 10)



Graph 10 illustrates the county-clusters relative to their danger ranking within Washington State.

We will take a closer look at the specific counties within these clusters in the next section.

2.3 Foursquare Venue Selection

9. Receiving Venue data from Foursquare for WA Counties

Successfully having ranked the counties by their associated clusters, I could utilize the Foursquare API [4] to explore counties for popular venues, which could be deemed as safe or unsafe depending on the danger ranking of the related county.

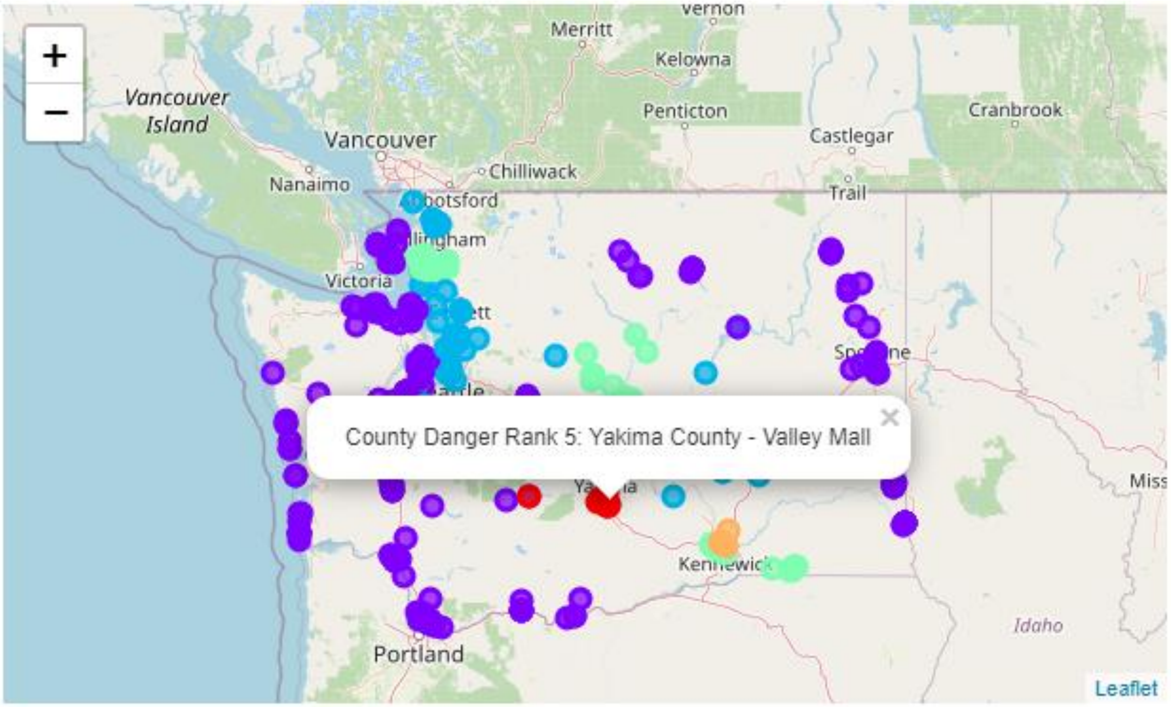
Within the Foursquare API, I used several variables. I started with using the explore action for my URL. Additional variables I used for the query includes the county name which the API can create a geo-boundary for, and a limit on the amount of responses returned. The version (date) being used is the last week in our dataset: 5/10/2020. This venue search was performed for all unique counties within our data.

Once I received the response of venues for each county, I copied the latitude and longitude of the venue along with their name. Having this, helps us graph and study the varying venues within the surrounding counties of Washington State.

10. Graphing and exploring most popular venues within clustered counties

Graph of Venues around WA Counties:

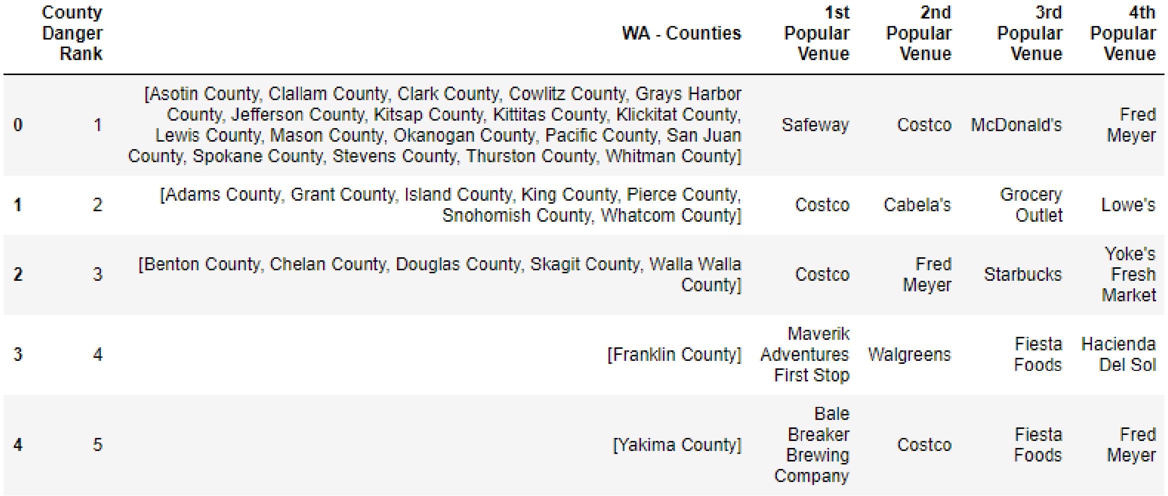
(Graph 11)



Now that we visualized the venues in Graph 11, we can check for which venues are most popular within the entirety of each cluster. Doing this will help us rank whether a certain cluster (set of counties) would be safe / unsafe visiting the set of relating popular venues.

County Danger Rank – WA Counties – Popular Venues:

(Table 9)



	County Danger Rank	WA - Counties	1st Popular Venue	2nd Popular Venue	3rd Popular Venue	4th Popular Venue
0	1	[Asotin County, Clallam County, Clark County, Cowlitz County, Grays Harbor County, Jefferson County, Kitsap County, Kittitas County, Klickitat County, Lewis County, Mason County, Okanogan County, Pacific County, San Juan County, Spokane County, Stevens County, Thurston County, Whitman County]	Safeway	Costco	McDonald's	Fred Meyer
1	2	[Adams County, Grant County, Island County, King County, Pierce County, Snohomish County, Whatcom County]	Costco	Cabela's	Grocery Outlet	Lowe's
2	3	[Benton County, Chelan County, Douglas County, Skagit County, Walla Walla County]	Costco	Fred Meyer	Starbucks	Yoke's Fresh Market
3	4	[Franklin County]	Maverik Adventures First Stop	Walgreens	Fiesta Foods	Hacienda Del Sol
4	5	[Yakima County]	Bale Breaker Brewing Company	Costco	Fiesta Foods	Fred Meyer

This is the final list of popular venues within the clustered counties based off their Danger Rank.

3. Results

To sum up the County Danger Rank's we have set up, along with the results displayed in Table 9, we can measure these varying clusters using two values. Their most popular venues as well as their correlating metrics when it comes to the percent change in the cumulative sum of cases (same metric we used to cluster the counties).

County Danger Rank - (County-Averaged) Percent Change in the Cumulative Sum of Cases:
(3/22/2020 – 5/10/2020) (Week-by-Week values):

1. .004875%
2. .019627%
3. .027069%
4. .051984%
5. .089936%

In terms of answering our problem statement of finding popular businesses/venues, we can now better determine which venues might be safe/unsafe to visit for counties in the recovery process of Covid-19. Counties within Danger Rank 1 seem to have the lowest rate of cases on a weekly basis at .004875%, which illustrates that their popular venues: Safeway, Costco, McDonald's and Fred Meyer are safer to visit during this quarantine period.

Counties within Danger Rank 2 and 3 also have a relatively low percentage at .019627% and .027069%, respectively. Like Danger Rank 1, Costco seems to be a very popular choice within their associated counties.

Now looking at Danger Rank 4 (Franklin County) and Danger Rank 5 (Yakima County), they have relatively higher percentages at .051984% and .089936%, respectively. Both these counties include Fiesta Foods as a popular venue, as well as Walgreens in Franklin County, and Costco and Fred Meyer in Yakima County. People should be more cautious at venues within these locations since they contain unsafe areas that are not performing as well in the recovery process of Covid-19. As a result, anyone visiting these areas/venues should put increased efforts in preventative measures.

4. Discussion

When initially figuring out the metrics that the clustering model should be based on, I realized that I could use my analysis of the USA data as a baseline for how I might cluster the WA data. I realized early on that the FIPS values would be important for graphing, so I made sure to keep those in most of the data tables. When dealing with the USA data, I noticed that using percentages when relating cases/deaths with population metrics, proved best at keeping any future statistics and analysis on counties on the same scale.

Now looking at the WA data, since I was testing for variables that meshed with the recovery process, I created a metric for the cumulative sum of cases. From there, I used each county's population [2] to create a percentage for cumulative sum of cases for Covid-19. However, since I wanted to test on how well counties in WA state were recovering in quarantine, I created a metric for the change in the percentage of the cumulative sum of cases, on a weekly basis. I decided to not utilize data on deaths, since it is unrelated to the analysis of the contraction/spread rate of Covid-19 and our problem statement. Maybe in future studies when dealing with mortality rates, we can use and create more metrics associated with data on deaths.

The USA data was structured on a daily basis; however, the WA data was based on a weekly basis. Considering this, I decided to signify the "recovery process" as the weeks from the peak week to the last week listed. I calculated the peak week to be the most common week in which counties had the highest count of incoming cases. This ended up being 3/22/2020, which resulted in a timeframe for the K-Means Clustering process from 3/22/2020 to 5/10/2020. In future studies, the timeframe can be altered to review different phases of the recovery process.

Using the Elbow method, I tested a range of clusters from 1 to 9 on the confined recovery weeks data. Observing Graph 6, I decided to use 5 clusters for my K-Means Clustering. Other options are 3 or 7 clusters which will give a narrow or wider spread on the set of counties in Washington State.

The Foursquare API has a plethora of parameters that can be used to explore geographical regions for venues. I used the "near" parameter which automatically finds the bounds of the geocode for the location, which in this case is the counties of WA state. An alternative is to use the latitude and longitude of the counties and set a radius to explore. I limited each of the queries to 10 responses per county. I also set "time" and "day" to retrieve results for any time of the day or day of the week, respectively. This allows the results to not be biased towards the specific time I query for a response. The last parameter I utilized is "sortByPopularity", which helps return responses more suited towards our problem statement of finding popular venues within counties.

As a note after analyzing the WA data, similar studies can be done on any of the other states in the USA. FIPS values must be available along with metrics for cases and deaths in relation to Covid-19.

5. Conclusion

WA state is ranked #36 in the USA when it comes to Cases per Population of Covid-19 with a total rate of .371%, from 1/21/2020 to 5/13/2020. When observing the County Danger Ranks and their associated counties, we concluded that people should be more cautious and put extra effort into preventative methods if they are visiting venues in Franklin County and Yakima County. This is due to their relatively slow decrease in cases from week to week. These counties must do better in reducing the cases in order to increase the safety of anyone visiting venues in these areas.

People within the counties associated with County Danger Ranks 1, 2, and 3 are doing much better when it comes to preventing the spread of Covid-19. Popular venues in these counties include Costco, Fred Meyer, and Safeway. These venues are deemed to be safer for people to visit considering the relatively lower weekly percent change for the cumulative sum of cases. With that being said, anyone venturing outside within these counties should make sure to keep up the preventative methods in the weeks to come.