Analysis of How Correlations Between Per Capita Income and City Venues Explain Gun Violence Distributions

Lulu Ricketts

March 31, 2020

1. Introduction

1.1 Background

Without a doubt, gun violence in the US has become a huge problem in the past couple decades. Affecting hundreds of thousands nationwide, the number of people growing per year continues to increase. However, it is clear that some places are more at risk than others. It seems to be a persistent trend that lower-income areas are more susceptible to gun violence than higher-income areas. This project is seeking to explore the areas affected within 2019 and determining which areas tend to be more at risk.

1.2 Audience

Gun violence affects everyone that resides in the US, essentially putting lives at risk every day. My goal for this project is to provide clarity as to which places are more at risk so the residents of those areas can take better protective measures.

2. Data

2.1 Data Acquisition

For this project, I utilized two datasets. The first was a complete listing of the metadata of each gun violence incidence under the category of "mass shooting" in the year 2019, taken from https://www.gunviolencearchive.org. This data was directly downloaded from the website as a .csv file.

The second dataset I used for this project was the per capita incomes of each city or county that was listed within the shootings dataset. My initial plan of action to

go about obtaining this dataset was to scrape various Wikipedia pages with tables consisting of major cities and counties by per capita income. However, none of the Wiki pages I found contained an exhaustive account of all the cities I needed for my analysis. I found that this was because there was a mix of cities and counties in the first dataset, and some of those were small towns as well. Because those locations were so widespread over the entire country, I could not find a page to scrape that was satisfactory.

To get the dataset I used, I googled the per capita income of each location separately, aggregating the data into a new dataset which I then uploaded as a .csv file. The per capita income data within this dataset contains per capita incomes ranging from 2014-2018. Because of the semi-large range of years, this data may not be the most stable information accounting for fluctuations of income data across the years, but I went ahead with the analysis regardless.

2.2 Data Cleaning and Feature Selection

The first dataset consisting of all the shootings in 2019 was already extremely clean. I found no null values in the dataset or any values out of place. I dropped one row consisting of the location "Cascilla, Mississippi" because it is an unincorporated community, and thus I could not find a reliable per capita income to represent it.

After uploading both datasets into my notebook, the first step I took was to merge those datasets on the "City" and "State" features. Upon ensuring that there were no null values in the resulting dataset, I realized that many cities had multiple incidences (rows) within 2019. I aggregated equivalent cities into one row, summing up the metadata of the incidences to get a description of the total incidences for each location.

The descriptive features in the dataset for each location only consisted of the raw number of people killed and people injured. These numbers could be easily subject to skewness if a particular location has a higher population or more incidences, so thus are not a reliable measure of risk. To more accurately assess the risk of each location, I added the features of "number of incidences" and "killed/incidence". These are the features I relied heavily on in my analysis because they are more robust to fluctuations in population or otherwise.

2.3 Geolocations

A large portion of my upcoming analysis of the data involves visualizations in the form of maps. To obtain the geospatial coordinates of each location, I utilized the geopy package too locate those values and add them to my dataframe.

3. Methodology

3.1 Initial Distribution Maps

The first step in my exploratory data analysis was to map the distribution of all the shootings in my dataset using folium first without any clusters. For these maps and all following ones, the first map shows the distribution by number of incidences, and the second map shows the distribution by number of people killed per incidence. These maps together give a relatively exhaustive sense of the risk of each city.

Figure 1: Map of the distribution of mass shootings in 2019, the radius of each point representing the number of incidences in each location.

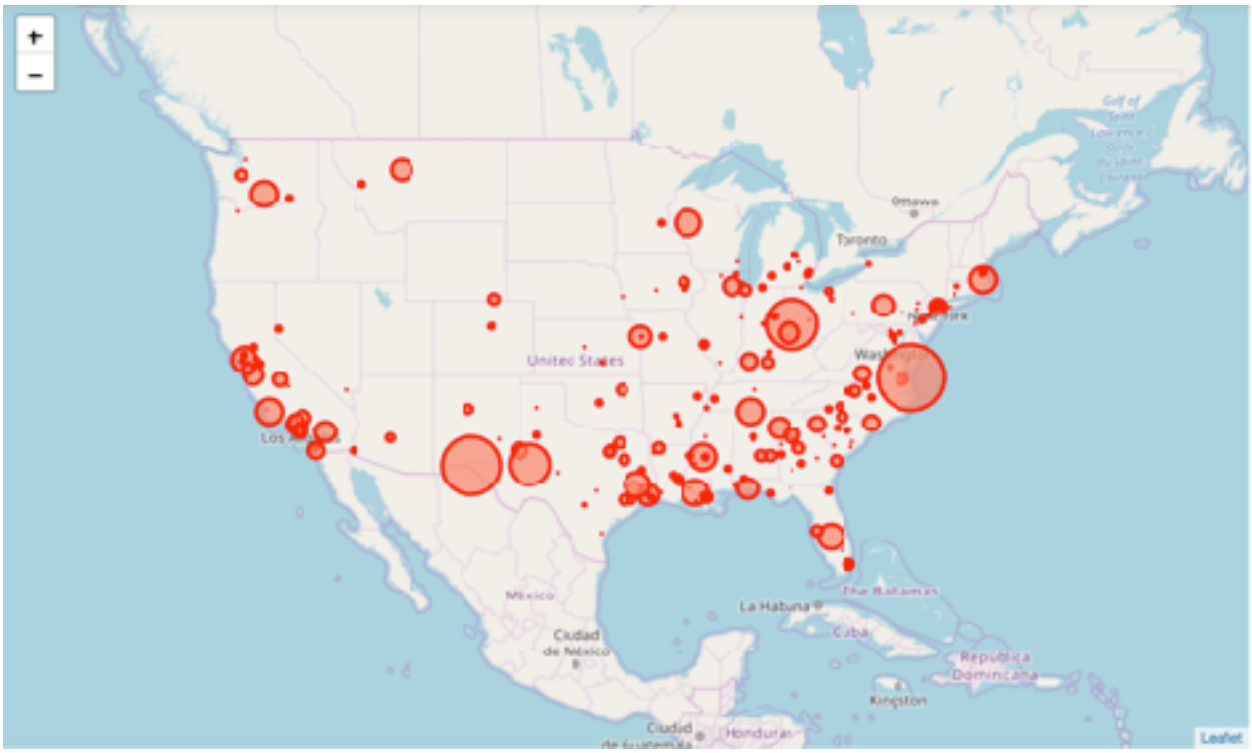

Figure 2: Map of the distribution of mass shootings in 2019, the radius of each point representing the number of people killed per incidence in each location.

3.2 Cluster via Per Capita Incomes

I continued my analysis by clustering in two different ways: via per capita income and via location venues. This portion contains the clustering via per capita income of each location. I observed a large distribution of the per capita incomes of my dataset, ranging from $9,507 to $84,985. I believe that clustering locations around similar incomes will give a sense of which social classes are more at risk of shootings.

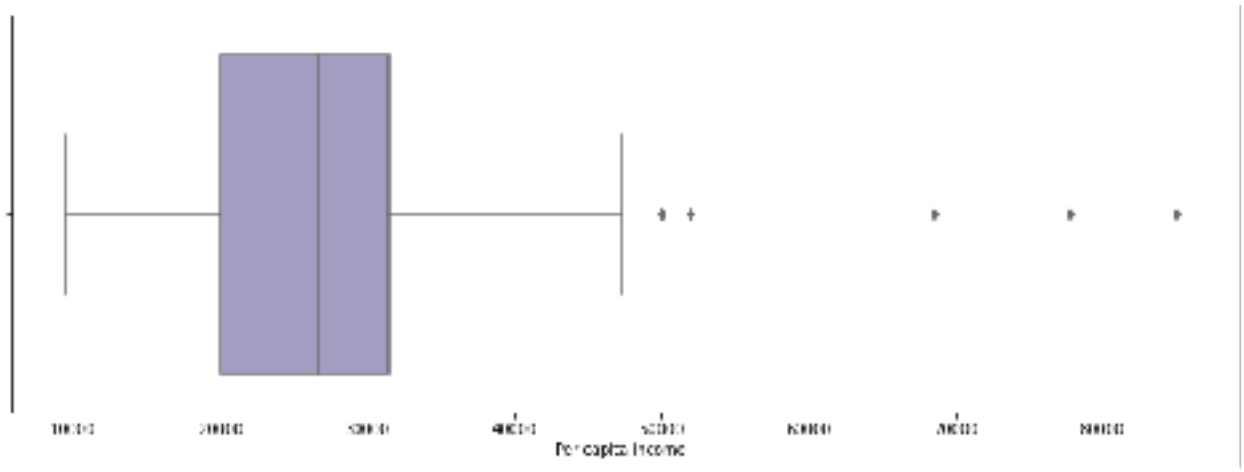

Figure 3: Boxplot showing the distribution of per capita incomes of the cities that have experienced shootings in 2019.

From the data in figure 3, I created five clusters encompassing all incomes:

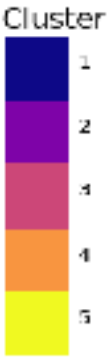

Figure 4: Legend for clusters by per capita income shown in following maps.

Cluster 1: the first quartile

Cluster 2: the second quartile

Cluster 3: the third quartile

Cluster 4: the fourth quartile

Cluster 5: all outliers to the right

Then the original distribution maps were recreated this time to show the per capita clusters, both in terms of number of incidences and number of people killed per incidence.

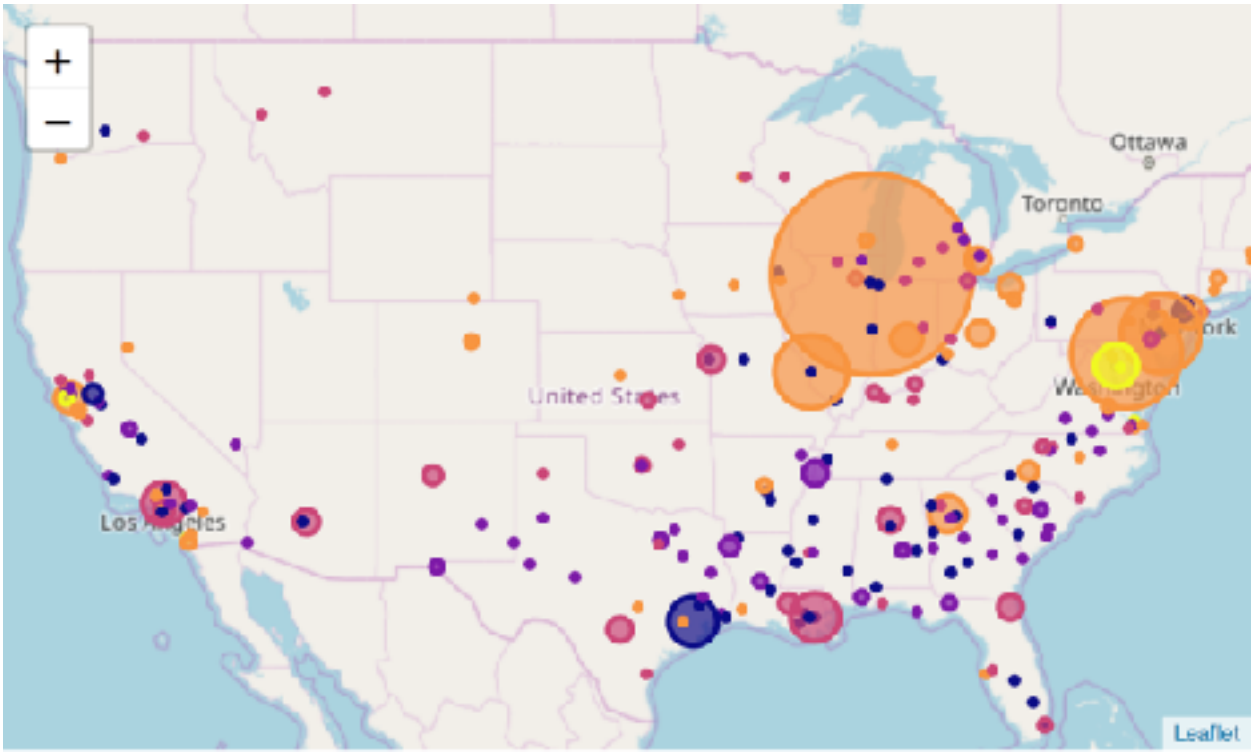

Figure 5: Map of the distribution of mass shootings in 2019 clustered by per capita income, the radius of each point representing the number of incidences in each location.

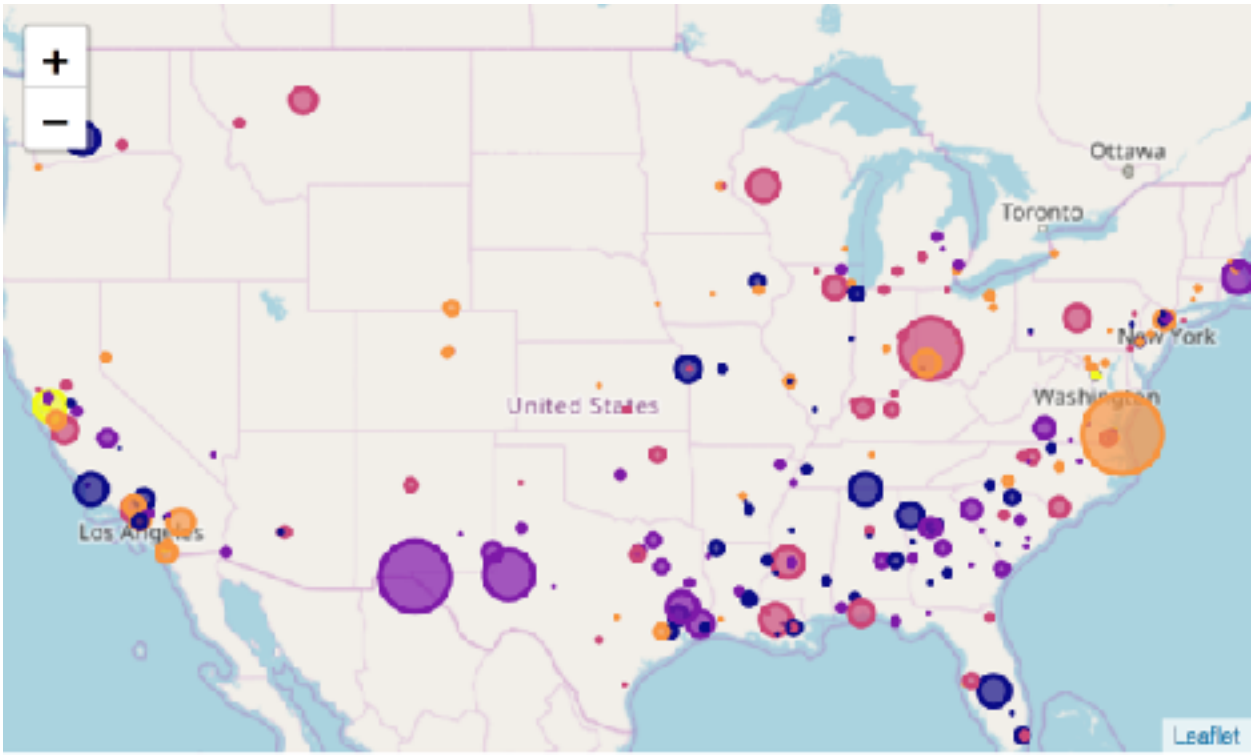

Figure 6: Map of the distribution of mass shootings in 2019 clustered by per capita income, the radius of each point representing the number of people killed per incidence in each location.

3.3 Cluster via Location Venues

The second factor that I chose to use for this analysis was to cluster the locations by the venues in that location using k-means clustering. This provides an idea of how the contents of each location could correlate with per capita income, subsequently correlating with the risk of shootings occurring in that area. For this clustering, I used the FourSquare Places API to search the top 20 venues around each area, then clustered them into groups based on those venues.

When obtaining the top 20 venues for each location, the resulting dataset required some minimal cleaning. I realized that a few of the smaller cities returned less than 20 venues. Determining that this discrepancy in amount of venues could skew the clusters, I dropped 11 locations that had less than 15 venues.

Once the data was cleaned, I began the process to cluster the locations by the venues that were returned by the Foursquare Places API. To do this, I one hot encoded each of the 308 unique venue categories (ie. restaurants, bars, museums, etc), which I used to create clusters.

The elbow method was semi-inconclusive at determining the optimal number of clusters to use was inconclusive. However, I had anticipated this issue early on because of the ambiguity and overlapping of many of the venue locations. However, the elbow method did show the slightest hint of a change in momentum at k=5 clusters, which I determined was best to use so that there was additionally some measure of symmetry with the clusters by per capita income.

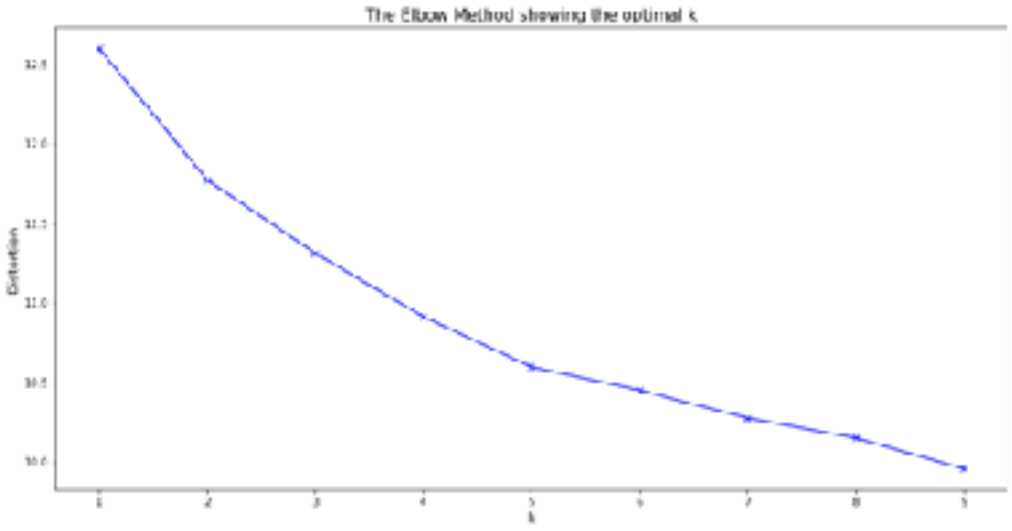

Figure 7: Semi-inconclusive elbow method showing the optimal k clusters.

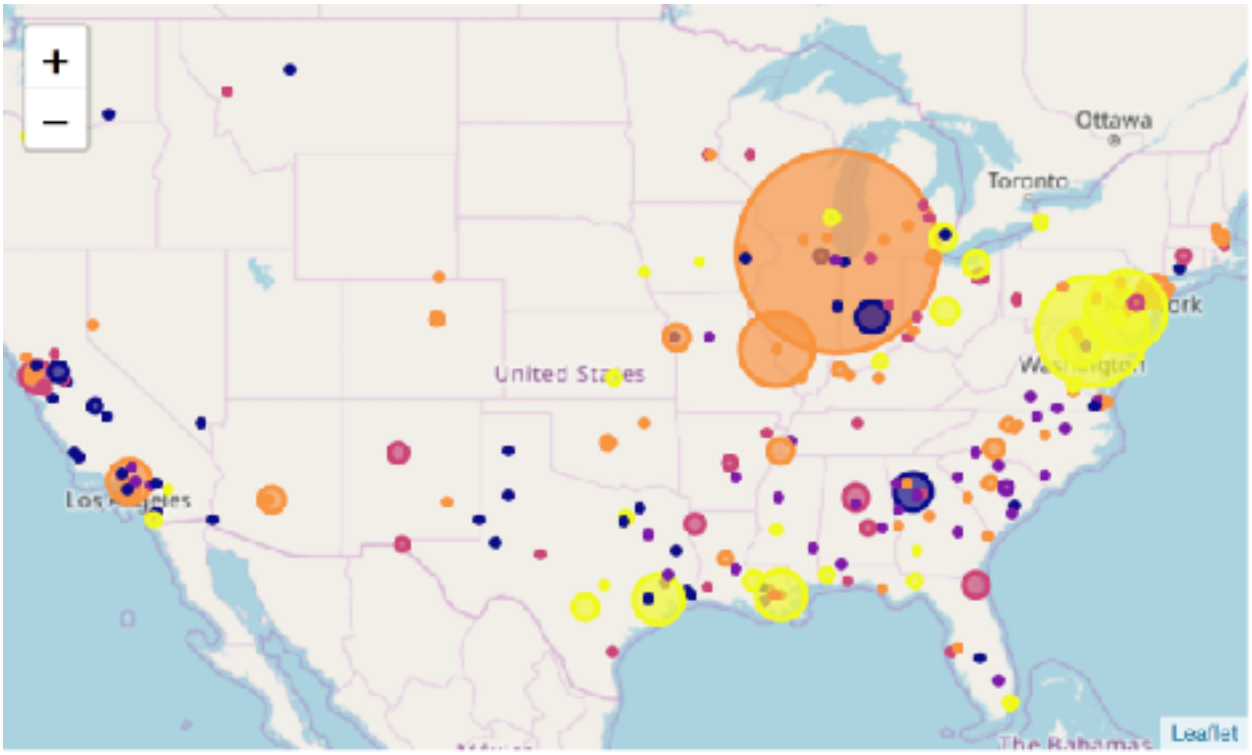

Figure 8: Map of the distribution of mass shootings in 2019 clustered by location venues, the radius of each point representing the number of incidences in each location (legend below).

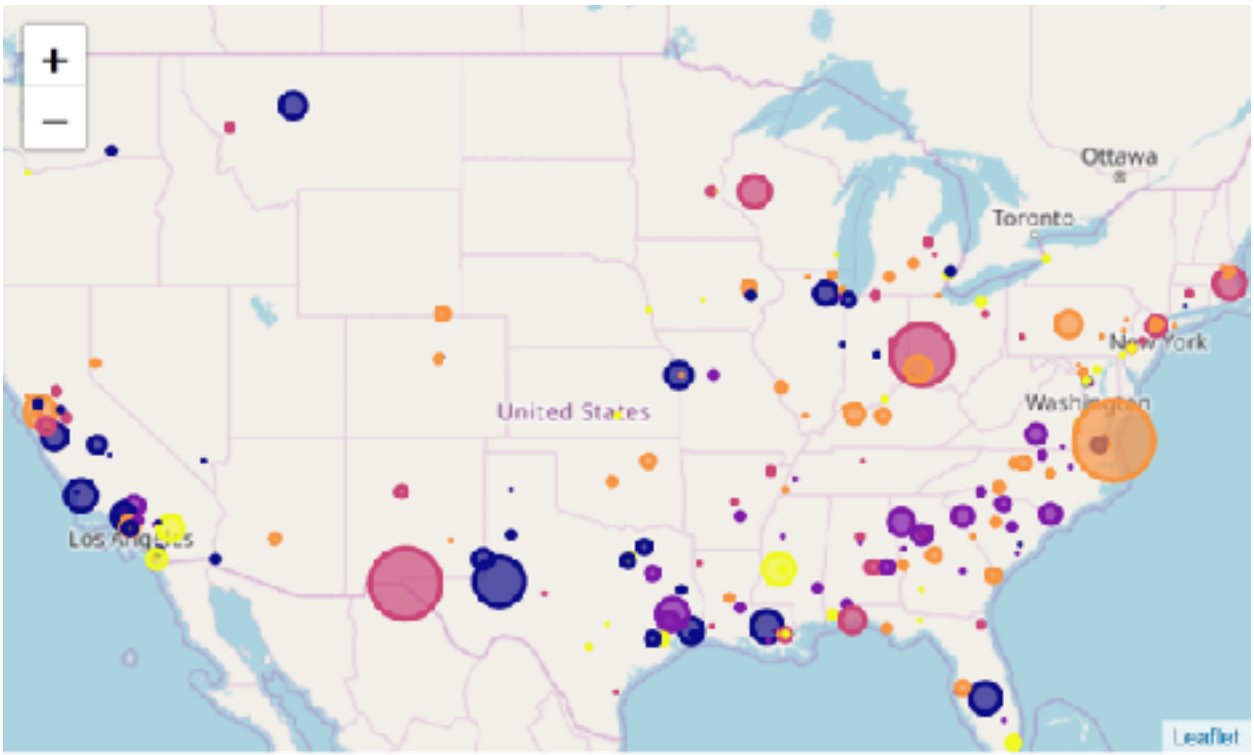

Figure 9: Map of the distribution of mass shootings in 2019 clustered by location venues, the radius of each point representing the number of people killed per incidence in each location.

Because the clusters formed by venue locations were unstructured, I was able to play around with the order of each cluster's color to get the maps to more closely resemble those shown in the per capita income clusters. The resulting legend for these maps is as follows:

Figure 10: Legend for clusters by location venues shown in previous maps.

3.4 Diving Further Into Location Clusters Meaning

Unlike the cluster by per capita income where each cluster represented a range of incomes, the clusters by location venues have little meaning without further analysis. To get a better idea of how the locations were clustered, I found the top three most common venues for each location, and then summing those results to find the top three venues per cluster, shown below

Cluster	Top 1 Venue	Top 2 Venue	Top 3 Venue
1	Fast Food Restaurant	Discount Store	Pizza Place
2	Hotel	Coffee Shop	Museum
3	Bar	Italian Restaurant	American Restaurant
4	Pizza Place	Coffee Shop	American Restaurant
5	Mexican Restaurant	American Restaurant	Coffee Shop

Figure 11: Top 3 venues within each cluster

As expected, there is a certain degree of overlap between each of the cluster venues. However, this gives us enough information about the most important contents of each of the clusters. Knowing this, I am able to assign the following descriptions to each of the clusters:

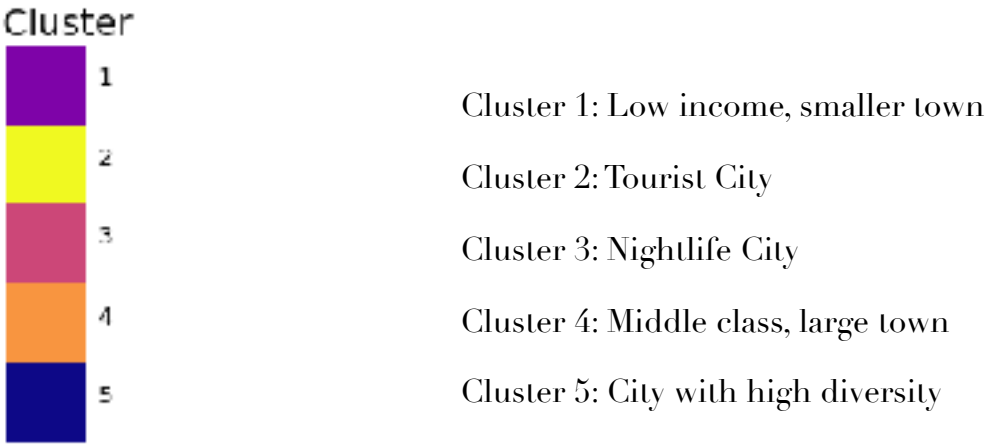

Figure 12: Legend for clusters by location venues, with descriptions.

3.5 Similarity of Per Capita and Location Venues Clusters

With these descriptions of the contents of each cluster by location's venues, the cluster colors should intuitively line up with one another.

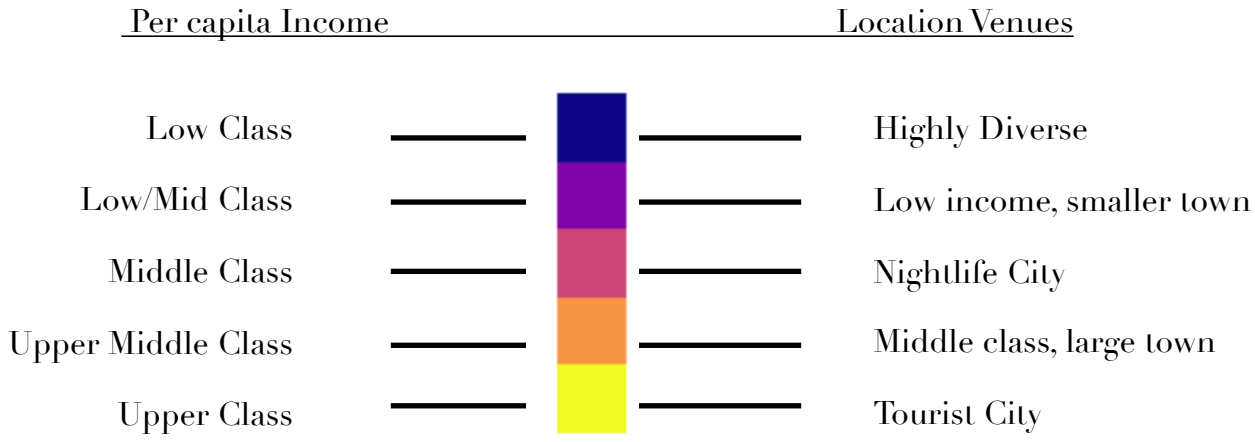

Figure 13: Alignment of the labels of clusters by per capita income and clusters by location venues.

For the most part, the clusters seem to match what we would intuitively pair together, suggesting that the two ways of clustering the data (by per capita income and by a location's venues) have some sort of correlation between them.

4. Results

4.1 Initial Distributions

I chose two ways to graph the geospatial data containing the locations of the shootings in 2019: one map showing the number of incidences for each location, while the other map shows the number of people that were killed per each incidence. These distributions, even before clustering, show a couple noteworthy properties.

The map representing the number of incidences each location had during the year 2019 (Figure 1) shows that Chicago, IL takes the lead in that category, surpassing the next highest city by almost double. The next couple cities with the highest number of incidences in 2019 are in the northeast, around New York and New Jersey.

However, when we look instead at the map representing the number of people killed per each incidence in 2019, the map is drastically different (Figure 2). This map shows a much more even distribution across all the locations mapped. Shifting away from Chicago and the tri-state area, the locations with the highest rates of people killed per incidence here include Virginia Beach, VA and El Paso, TX.

4.2 Per Capita Income Cluster Distributions

In the first map representing the clusters by per capita incomes and by number of incidences per location (Figure 5), it was surprising to note that the cities with higher per capita incomes tended to also have more incidences than those cities that have lower per capita incomes. These cities mostly reside in the northeast, spanning into the mid-west and the west slightly as well. The states that tended to have lower per capita incomes were heavily spread across the southern region of the US.

When shifting our view towards the second map showing the per capita income clusters by the number of people killed per incidence (Figure 6), the location with the highest rate shift away from higher per capita incomes towards those with lower per capita incomes. That is, while the radius of the dots in the northeast grow smaller, those spanning the south grow larger.

4.3 Location Venues Cluster Distributions

Clustering each location in my dataset by the top venues in each location provided a way to "look" inside each location. This provides the benefit of being able to draw speculative correlations about how the contents of each location contribute to its risk for shootings. After playing around with the legend's color order, the resulting clusters by location venues very similarly resembles those clusters by per capita income.

I was able to label each of the clusters by taking into account the top 3 venues in each of the clusters created by k-means. Looking at these labels (Figure 12), we see the same effect that we did for the per capita clusters. Intuitively, we should see a higher number of incidences for cities such as "tourist" or "middle class", and subsequently less incidences for cities and towns labeled "low income" or "nightlife". However, we see the opposite in the first map (Figure 8). For the second map however (Figure 9), we see that distribution shift towards what our intuition expects, very similarly to the behavior of the clusters by per capita income.

5. Discussion

5.1 Initial Distributions

The two initial maps (Figure 1 and Figure 2), while mapping the same locations, show a drastic difference when comparing the number of incidences that a location saw in 2019 and the number of people killed per incidence. This difference suggests that while a city may be more at risk of incidences, that doesn't necessarily predict the violence of each incidence. For example, Chicago by far took the lead in the number of mass shooting incidences seen in 2019, but was barely a speck on the map when mapping the number of people killed per incidence. With a lower fatality rate compared to other locations, it can be said that while cities such as Chicago may be more at risk of shootings, this does not necessarily go hand in hand with the notion that that city may be more at risk of fatalities. This idea is kept in mind while analyzing the clusters.

5.2 Per Capita Income Cluster Distributions

The two maps representing the clusters of each location by per capita income (Figure 5 and Figure 6) show a significant difference in which locations take the lead when comparing the number of incidences to the number of people killed per incidence.

When looking at just the map showing the number of incidences, the data showing that location with higher per capita income having more incidences seemingly goes against our intuition. Intuitively, we would most probably correlate lower incomes with more violence. However, as we established previously, being more at risk for incidences doesn't necessarily correlate with the violence of that incidence.

This is enforced when turning to the second map, showing that locations with lower per capita income actually have a higher rate of people killed for each incidence, even when the number of incidences itself is low. Thus, while locations with an overall higher number of incidences and per capita income may be more susceptible to shootings, but it is the locations with overall lower per capita incomes that are more at risk of incidences that are more violent. Analyzing the next set of clusters may give us better insight as to why that may be.

5.3 Relations Between Per Capita and Location Venues Clusters

The overall distribution of the clusters created by the venues in each location are very similar to that of clusters created by per capita incomes. Aligning with out intuitions, the labels for both of the clusters on the map seem to line up in a satisfactory way (Figure 13). Without having the relation between the two clusters, it is difficult to find a reason as to why higher income locations tend to have more incidences, while lower income locations tend to have a higher degree of violence for each of its incidences.

With the addition of the location venues clusters overlaying and describing the contents shown within the per capita income clusters, we are able to make a conjecture regarding the reason that locations with higher per capita income had more incidences in 2019. We can observe that the locations with higher income are labeled as "tourist" locations, or larger cities/towns. Because of the high population in these kinds of locations like Chicago, it may be possible that the number of incidences is simply an extension of the proportion of the population. For instance, I have labeled smaller towns are on the lower end of the per capita income scale, and by the same logic we could most likely assume that the low number of incidences is related to the lower population in those areas.

Because the distributions of the first map are more susceptible to fluctuations in population across different locations, it is imperative that the second map, showing the number of people that were killed per incidence, is taken fully into account. From my analysis, we can see that locations with lower per capita income are labeled as "highly diverse" or "small town", suggesting populations on the lower side. Because of

potentially low populations, the number of incidences is expected to be lower. However, the maps (Figure 6 and Figure 9) show that the violence of each incidence is significantly higher, as exemplified by the number of people killed per incidence, while the opposite is true for locations with higher per capita income. The relation between lower per capita income locations and higher levels of violence may be attributed to the usually low populations in those areas as well; the lack of bystanders in times of violence could potentially fail to stop the perpetrator, leading to longer more violent attacks. By the same logic, the presence of bystanders in highly populated locations with higher per capita income could halt the attacker from escalating to a high level of violence that results in more lives lost.

6. Conclusion

In this project, I analyzed the mass shootings that occurred in 2019 as reported by the Gun Violence Archive. I did so by splitting my analysis to look at two distinct features of the data: the number of incidences each unique location saw in that year and the average number of people killed per incidence. To further analyze why some areas may be more at risk of shootings and violence, I created two sets of clusters: by per capita income and by each location's venues, comparing them to find similarities. After resolving obstacles in data collection and analysis, I was able to conclude the relation that locations with higher per capita incomes may sometimes have more incidences, but with less violence. Conversely, locations with lower per capita incomes may have less incidences, but each incidence is of a greater violence. Both of these relations can additionally be tentatively attributed to the population size as well.

7. Future Direction

There are quite a few opportunities to refine the analysis or take it a step further. Refinement can be applied to the data collection of per capita incomes for each location. I ran into issues when scraping datasets, so many of the incomes were from a large range of years. Perhaps using median incomes of each location would have shown a different distribution. Furthermore, some locations were large cities, such as Chicago, which has drastically different incomes for different areas within the city, which is a discrepancy that this analysis doesn't account for. Lastly, I concluded that population size could play a factor in the violence of a location's shootings, and in the future that idea could be further analyzed alongside the current analysis.