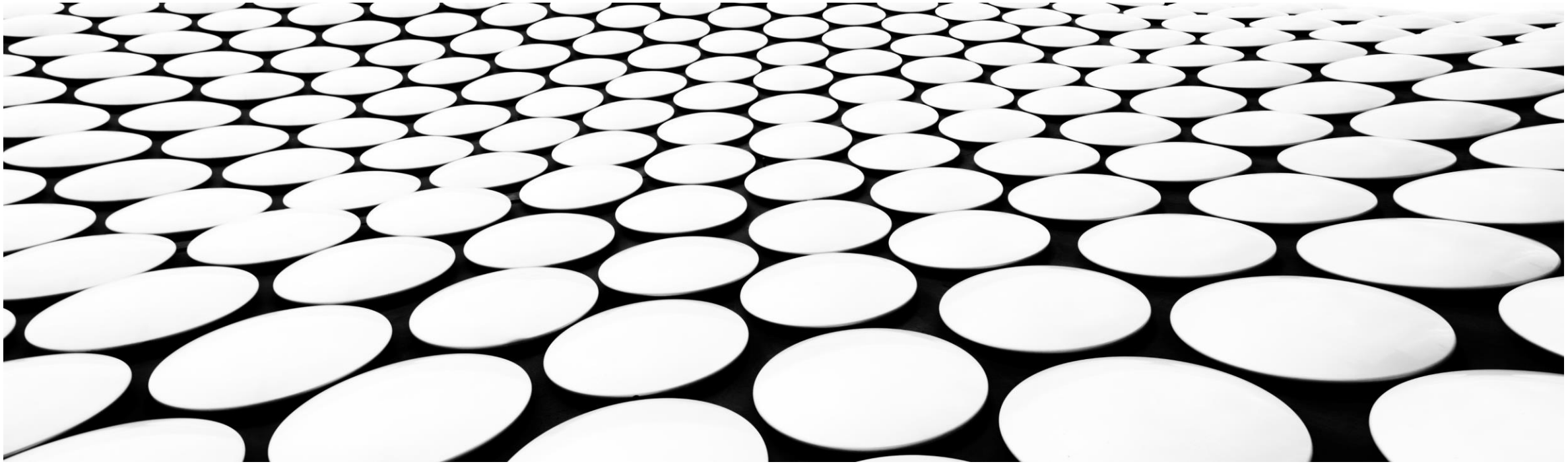
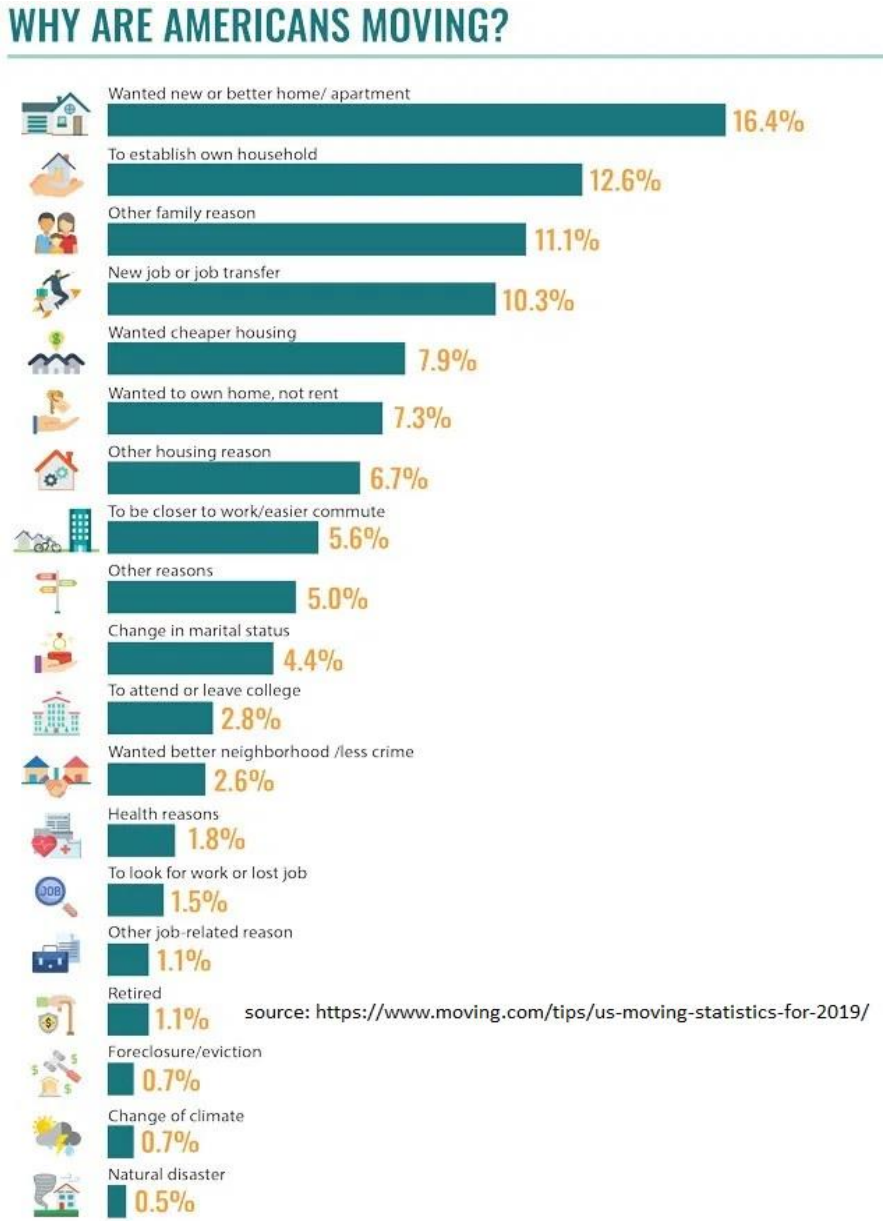
APPLIED DATA SCIENCE CAPSTONE – RELOCATION ASSISTANT

SUBMITTED FOR COURSERA STUDY



BUSINESS PROBLEM

- Americans relocate from one place to another for different reasons (as shown in the graphic on the adjacent side)
- People could use a tool / application that takes the profile of the person along with the requirements, and suggest a set of zip codes / neighbourhood for the person to consider exploring
- For the purpose of this implementation., Los Angeles and its suburbs are considered
- Data is crowd sourced, and scraped from various public web based sources covering various aspects of a geography that affects living preferences
- The goal is to create a segmentation of various zipcodes based on various living preferences for the user to select exploring

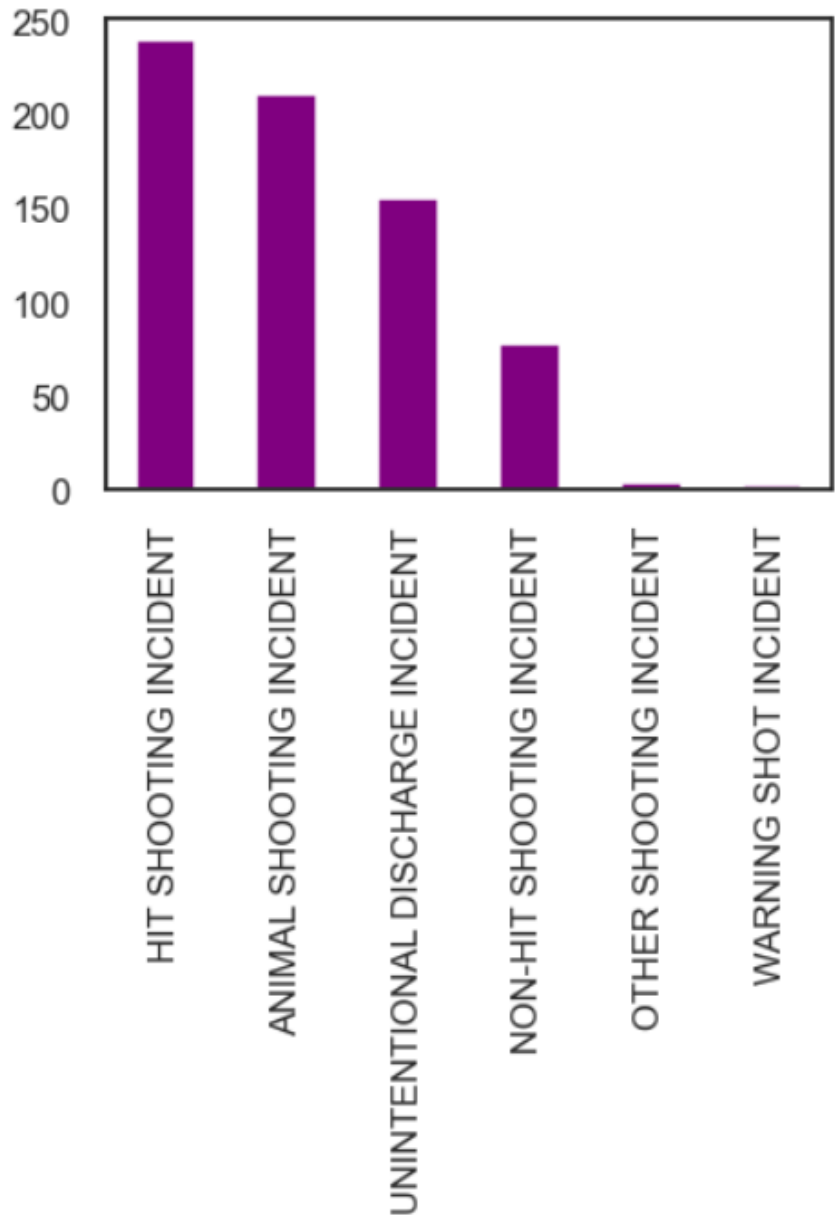


DATA SOURCES

- The data for this study includes factors that affect one's choice of living in a certain area (the area being Los Angeles for this study)
- The variables used for clustering are population, number of household, size of the household, median age, shooting incidents involving human, shooting incidents involving animals, top 4 most common venues of that neighbourhood – all by zip codes
- Zip code information:
  - https://namecensus.com/igapo/zip_codes/counties/alpha/Los Angeles County-California1.html
  - https://namecensus.com/igapo/zip_codes/counties/alpha/Los%20Angeles%20County-California1.html
  - http://download.geonames.org/export/zip/ (downloaded the USZipcodes.zip)
- Census: https://catalog.data.gov/dataset/2010-census-populations-by-zip-code
- Population by Zip codes: https://www.zip-codes.com/city/ca-los-angeles.asp
- Latitude / Longitude: "https://gist.githubusercontent.com/erichurst/7882666/raw/5bdc46db47d9515269ab12ed6fb2850377fd869e/US%2520Zip%2520Codes%2520from%25202013%2520Government%2520Data"
- Shooting Incidents: https://data.lacounty.gov/Criminal/All-Shooting-Incidents-for-Deputy-Involved-Shootin/d5zc-33fr

DATA PREPARATION AND CLEANSING

- Data is cleansed to remove duplicates, missing values etc.
- Missing values are sourced / substituted from other data sources from across the Internet to construct a complete dataset
- Venue information was added (using Foursquare APIs)
- Shooting incidents are added using a crime incident database. Since not all shootings are equal –
  - Animal related shootings are used as a single feature / variable: Those who see risk in living in a neighbourhood with lot of wild animals may prefer to avoid those neighborhoods that involve high animal sightings and shootings
  - Non Animal related, but shooting incidents of concern: These include any kind of shooting, including warning shots etc
  - Shootings causing hit / human injury
- As shown in the chart, there have been significant number of shooting incidents in a given year



OVERALL APPROACH TO FEATURE ENGINEERING & CLUSTERING

Zipcode City

Zipcode
Latitude
Longitude

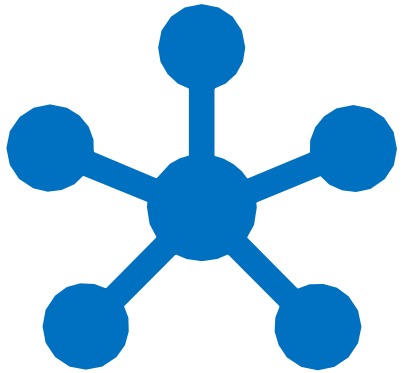
Zipcode
Population
Median Age Total Males
Total Females
Total Households
Average Household Size

Shooting Incident Type
Shooting Incident Date
Shooting Incident Location City
State
Zip
Reporting District
Geo Location
Handling Unit ID
Handling Unit Name
Approximate Latitude
Approximate Longitude

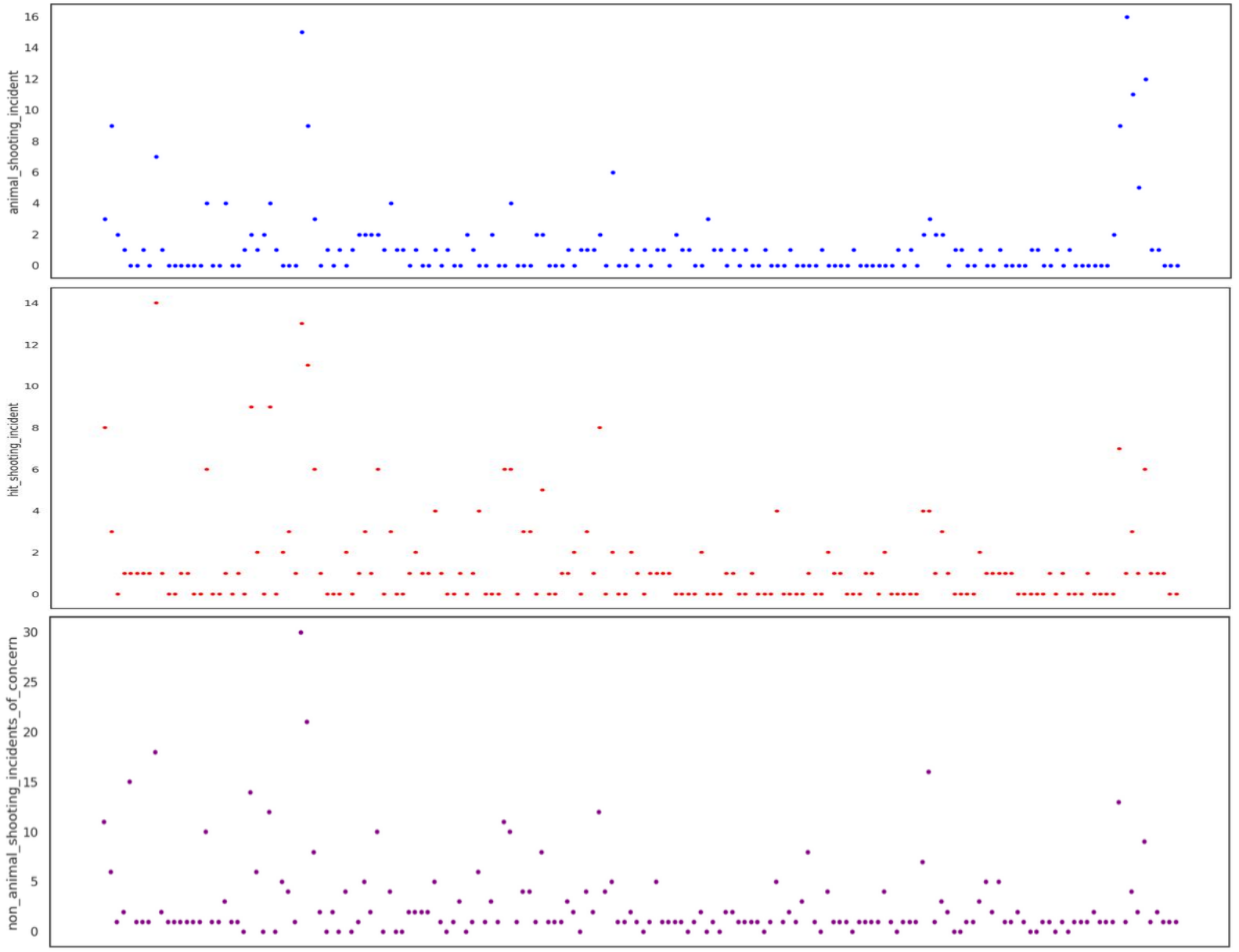
Foursquare APIs

Zipcode
Latitude & Longitude
Venue & Categories
Population
Median Age
Total Males
Total Females
Average Household Size
Shooting incidents
1st Common Avenue
2nd Common Avenue
3rd Common Avenue

clustering

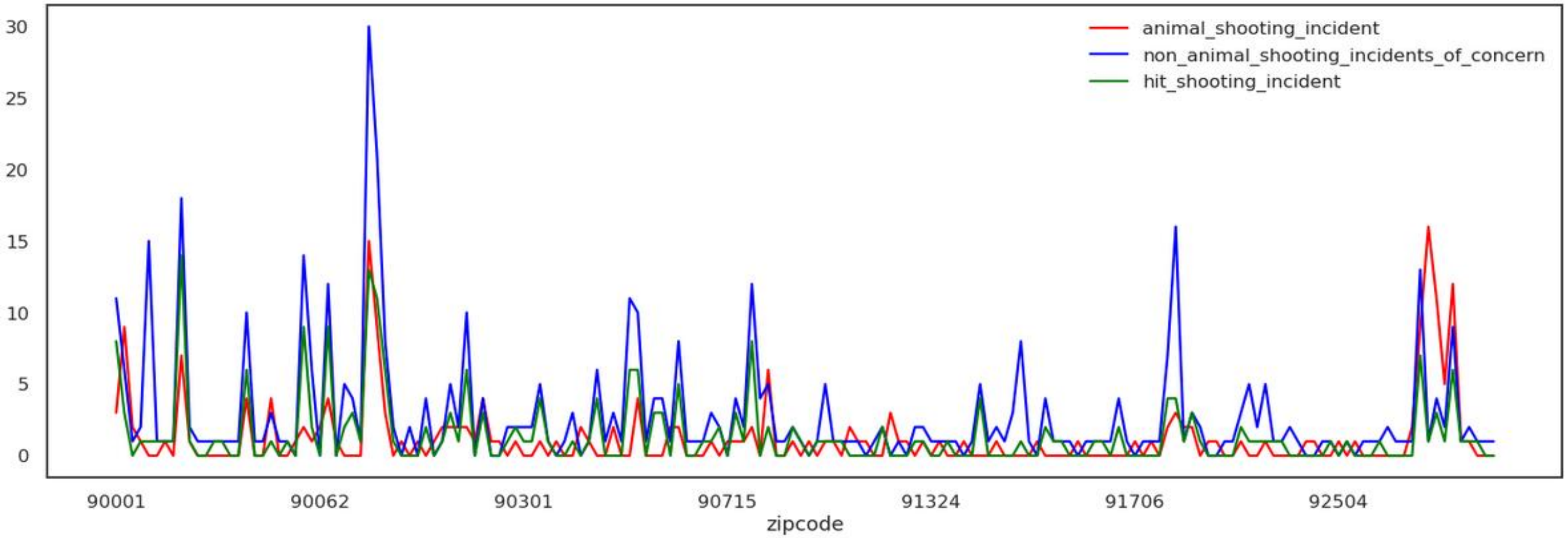


SHOOTING INCIDENT SCATTER



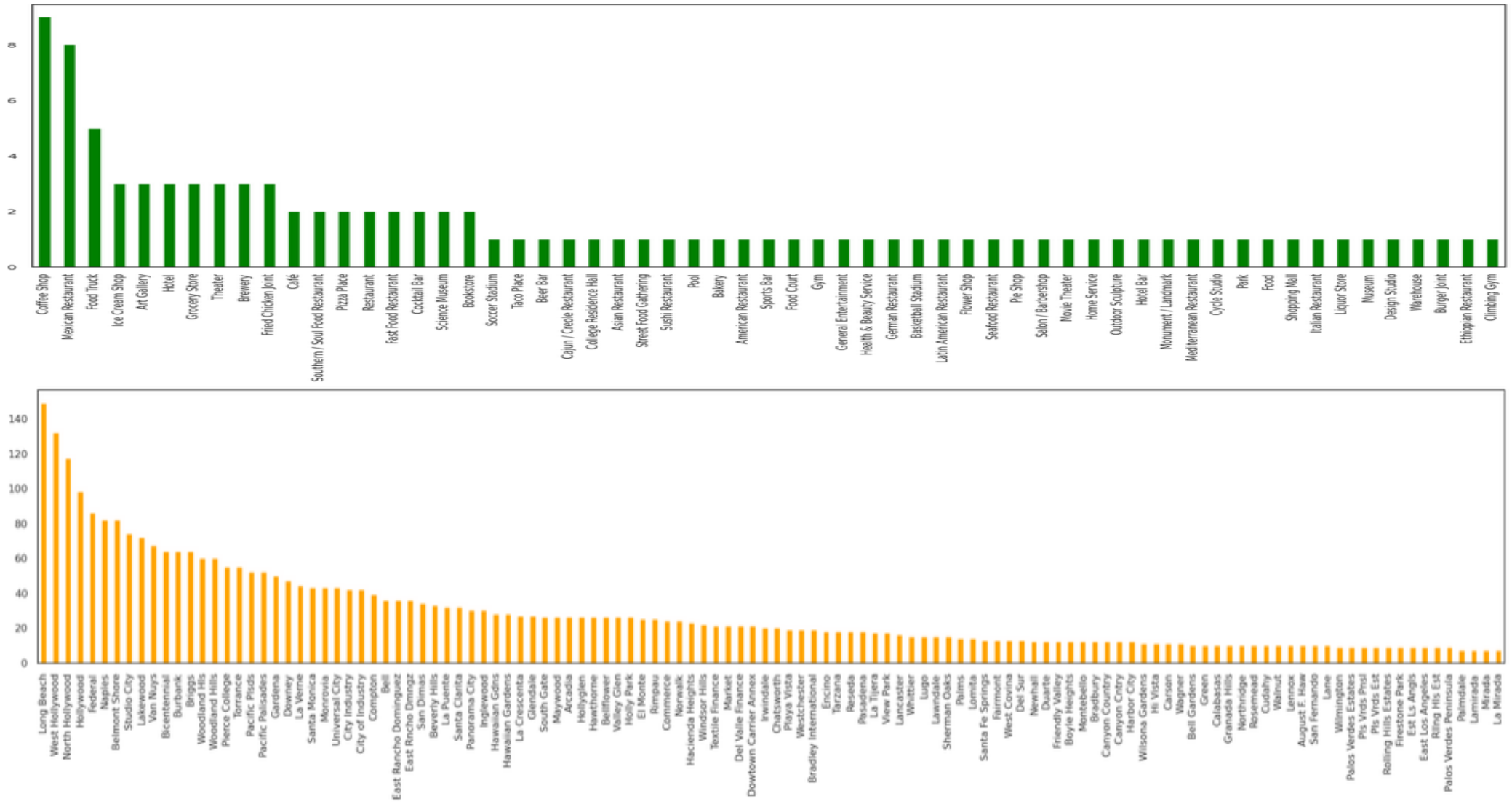
- As shown in the charts, shooting incidents are elevated in some areas / zip codes
- Animal shooting could involve wild animals and other ones that could spread disease
- Shootings involving human injury (hit) could be particularly prone to high crime rates
- The yearly statistics of all these shooing incidents could vary by zip code and regions

ALL THREE TYPES OF SHOOTING COMPARED AGAINST ZIP CODES



As shown in the plot above, there are areas with significant episodes of all kinds of shootings, and they seem to be concentrated to certain localities more than the other

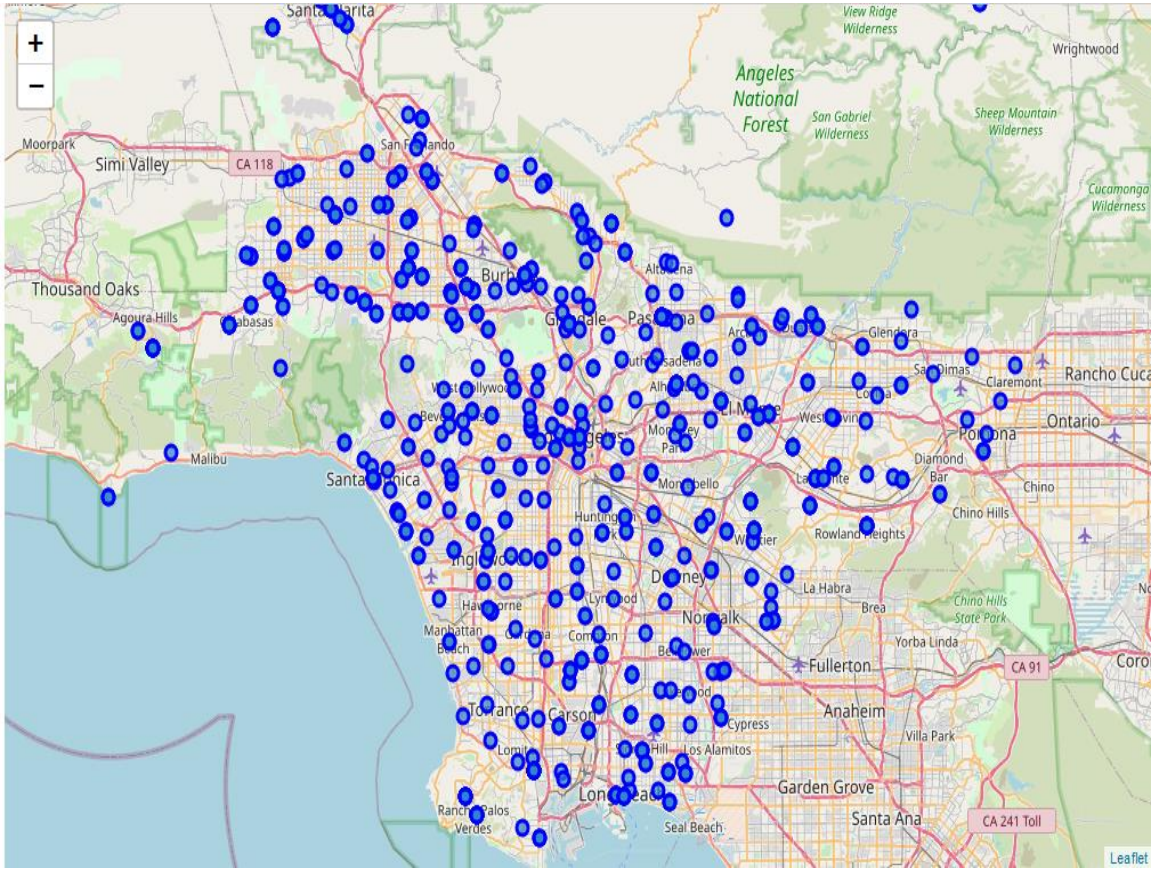
CATEGORIES ADDED USING FOURSQUARE APIS

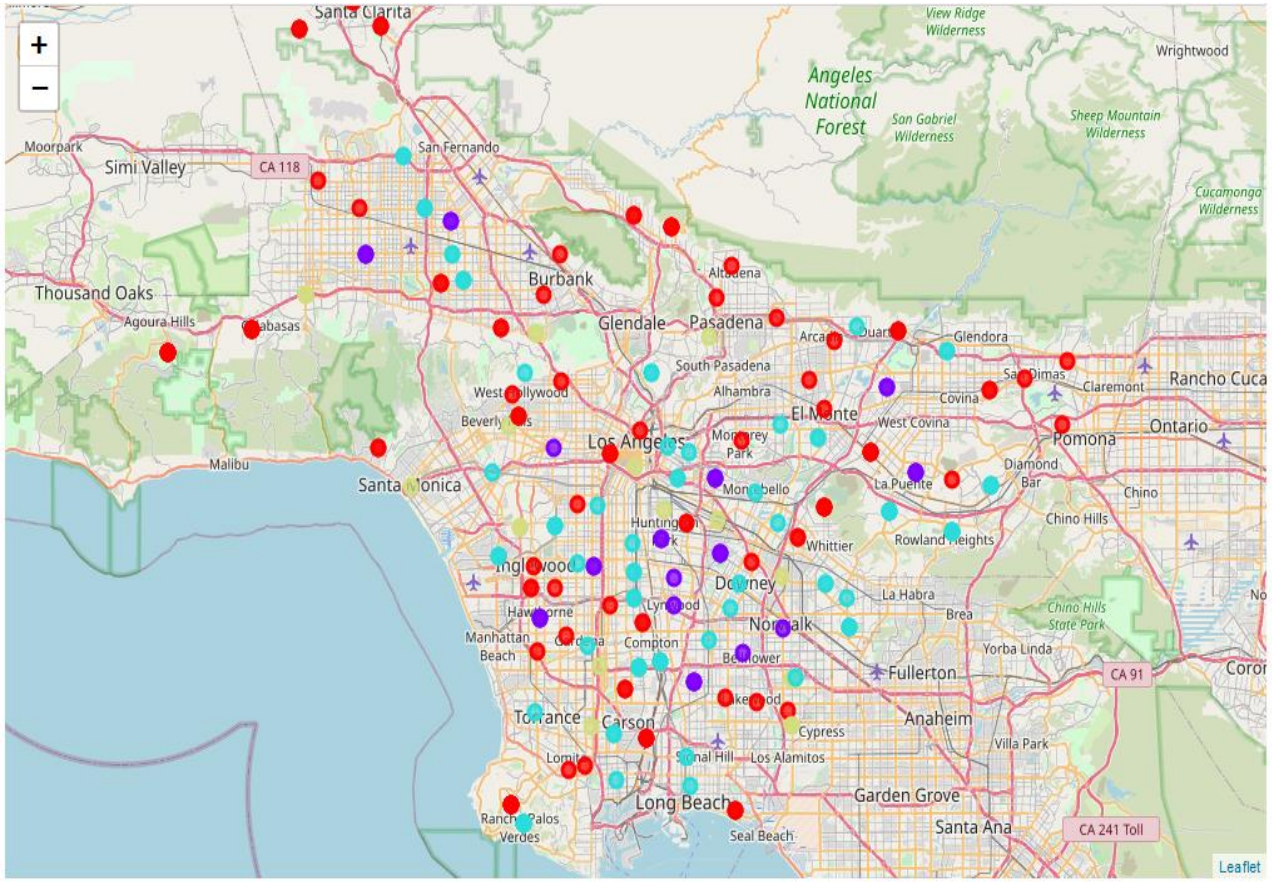


The plot shows the number of venues added using Foursquare APIs across the zip codes to enrich the feature set

DISCUSSION OF RESULTS

COVERAGE AREA OF LOS ANGELS WITH AND WITHOUT CLUSTERING

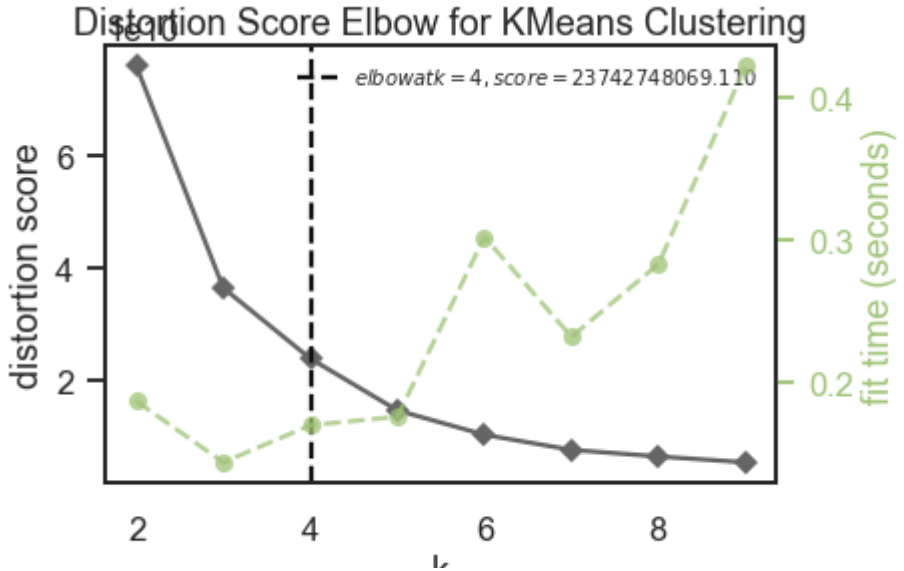


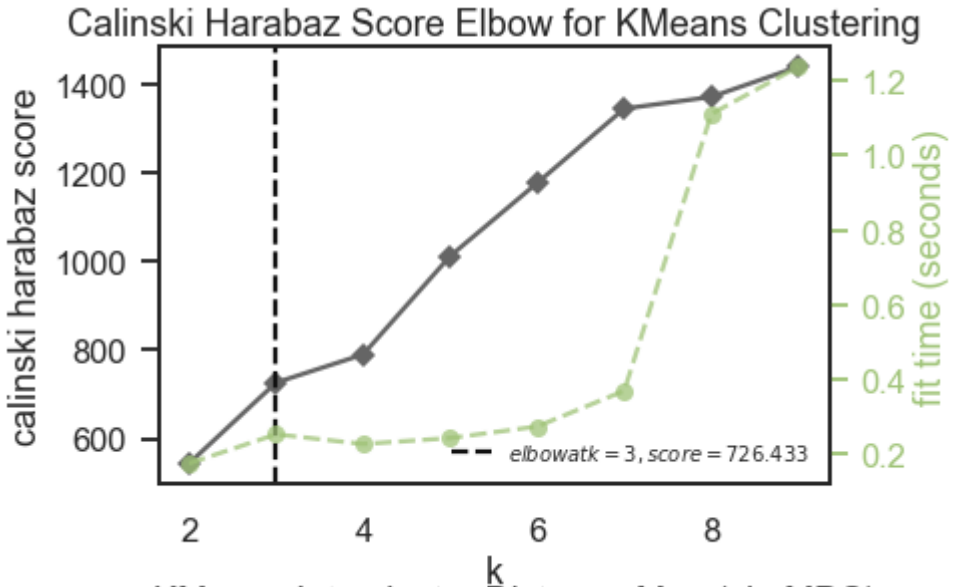


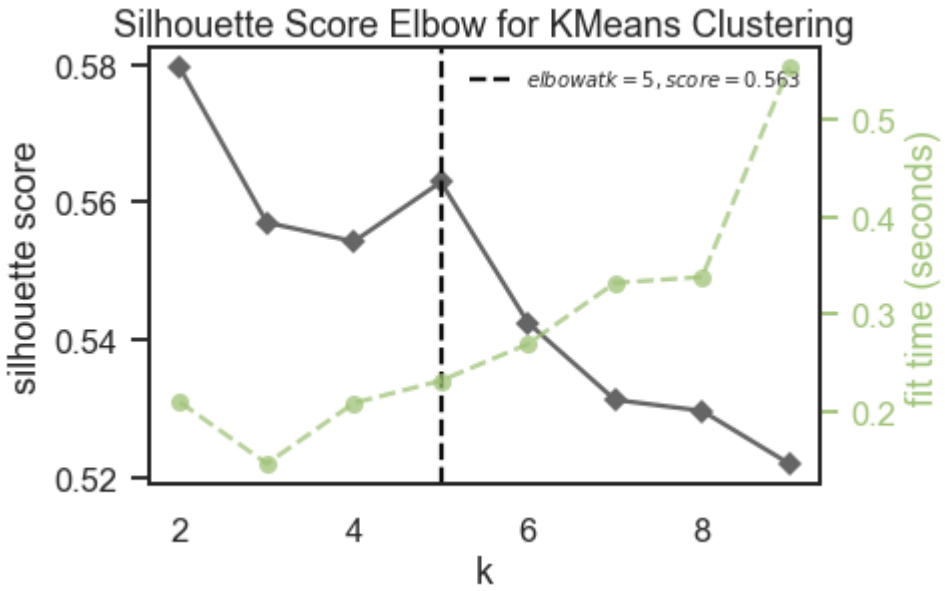
Cluster 1 Cluster 2 Cluster 3 Cluster 4

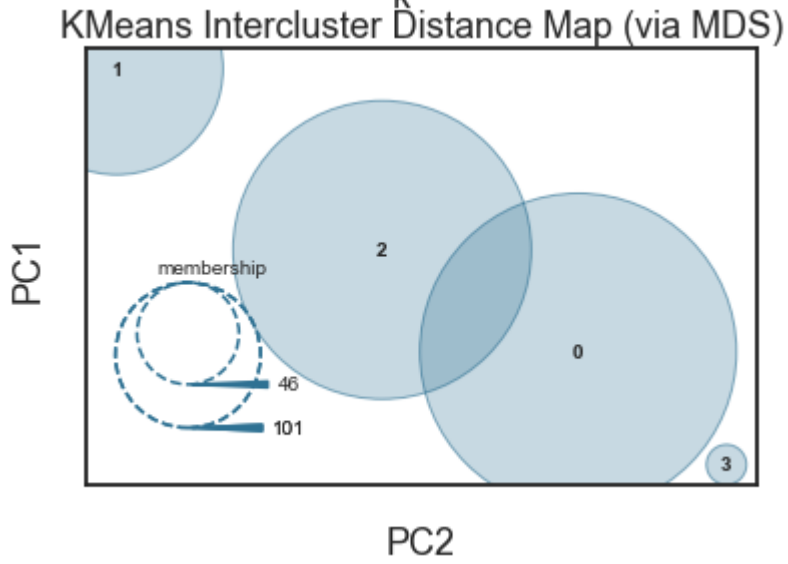
As maps show above, the area represented on the left with venues, is clustered on the right in terms of population, number of households, avg median household age, shooting incidents, and the top four most common venues that are found in the neighborhood. Expanding the area can result in more meaningful clusters

CLUSTER METRICS AND DETERMINATION OF NUMBER OF CLUSTERS



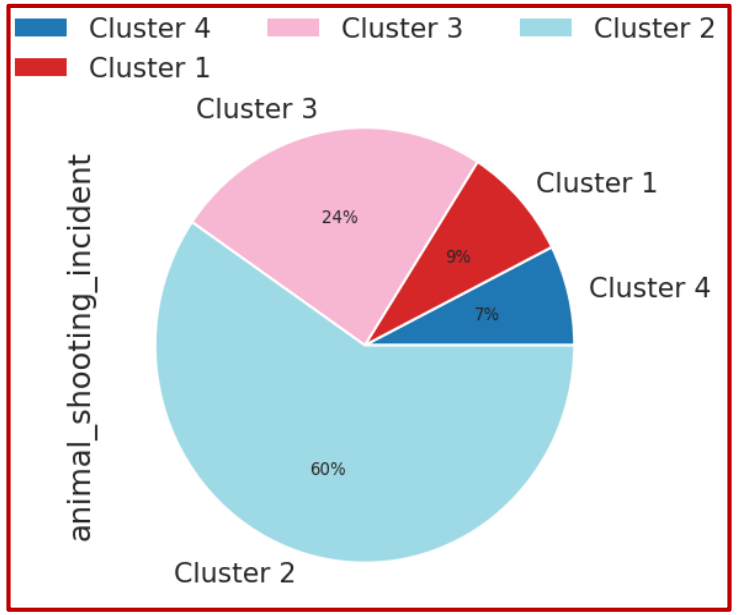


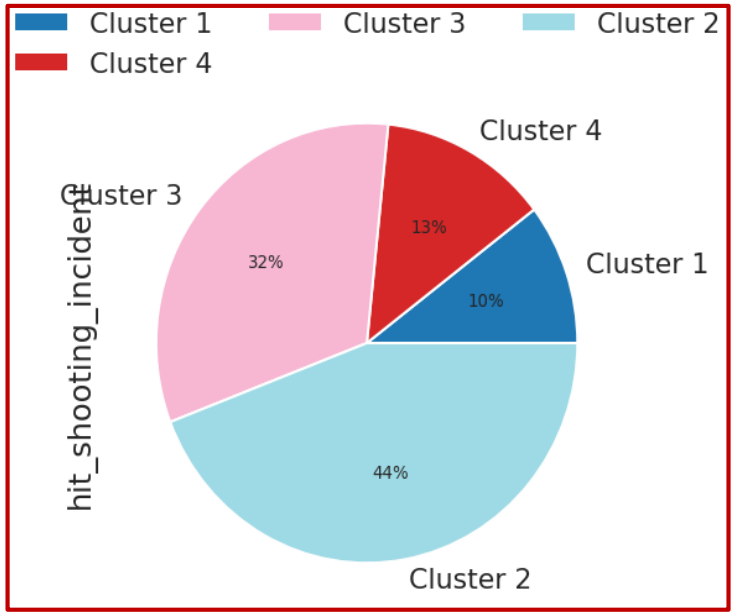


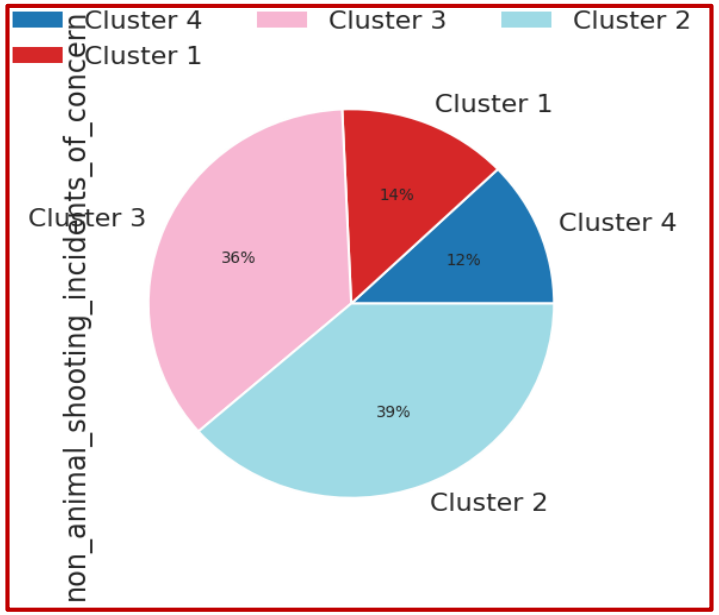


As shown here multiple metrics are tried to find the optimum number of clusters. 4 clusters seem to result in the best outcome

CLUSTER COMPARISONS – BY SHOOTING INCIDENTS



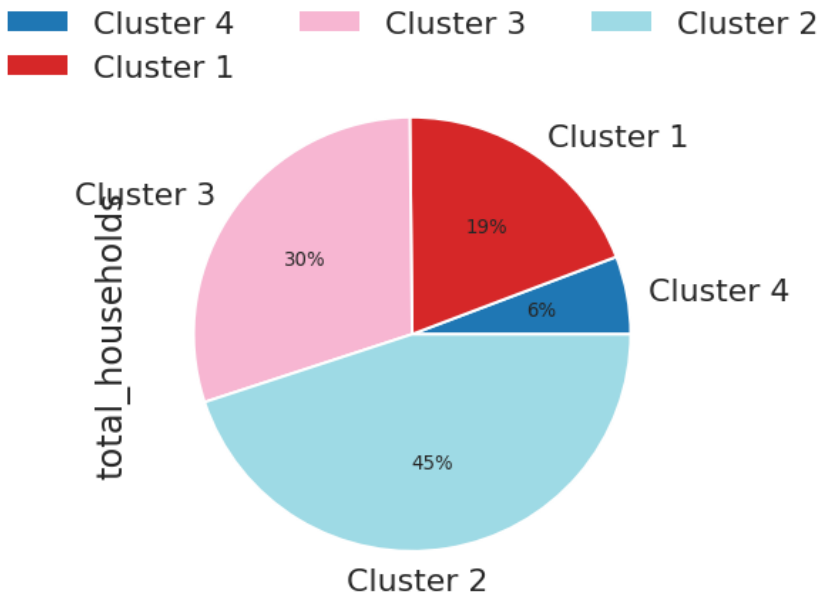


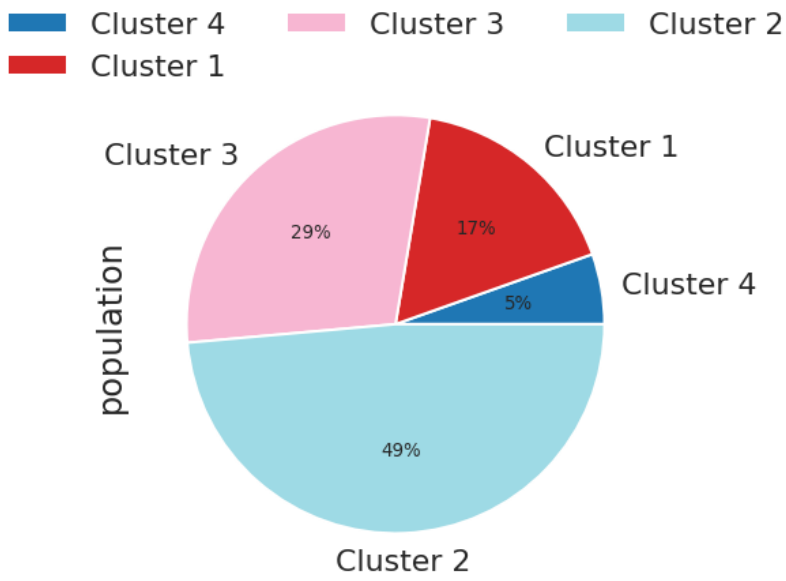


The charts show how the clusters compare with respect to various types of shooting incidents – Animal shooting, Shooting of a human being on another causing a hit, and all types of human shooting including warning shots. It is clear from the analysis that

- Cluster 2 is the worst when it comes to high rate of shooting incidents of various kinds, but has many shopping and mall areas
- Cluster 3 follows cluster 2 closely
- Cluster 4 is arguably the safest from the standpoint of shooting and Gun related crime
- Cluster 1 is the next to the cluster 4 when it comes to safety from Gun related crimes

CLUSTER COMPARISONS – BY POPULATION AND HOUSEHOLD

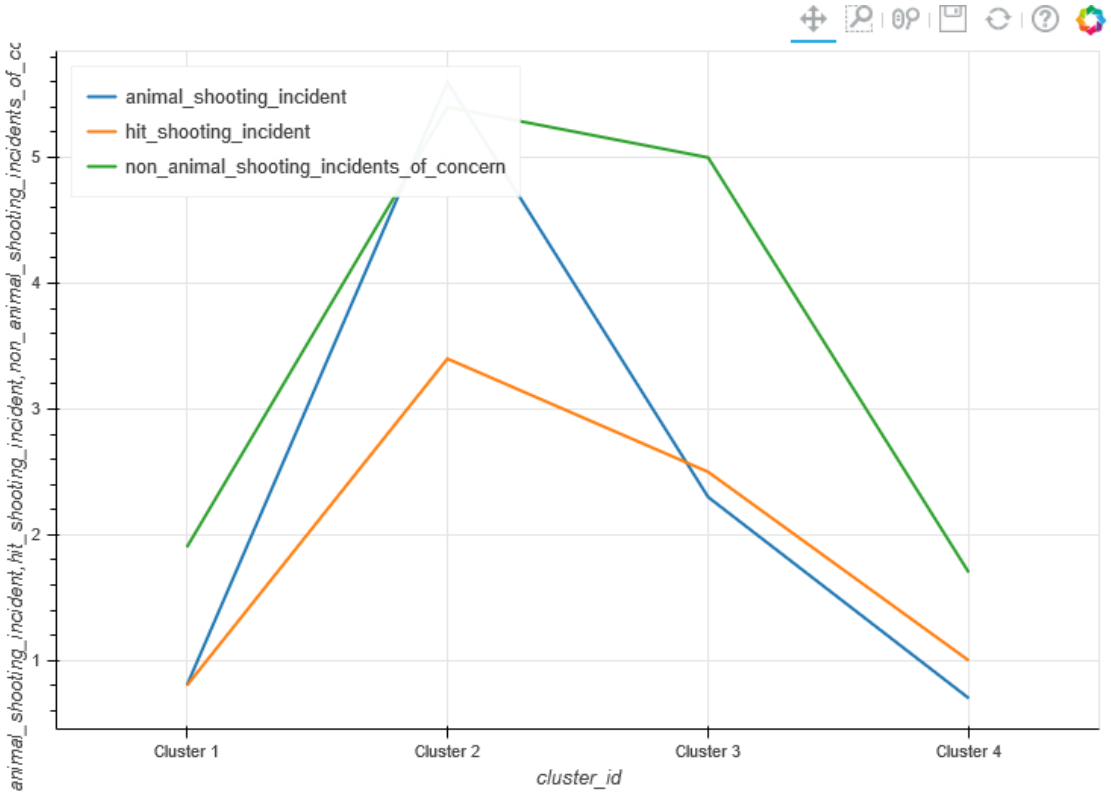


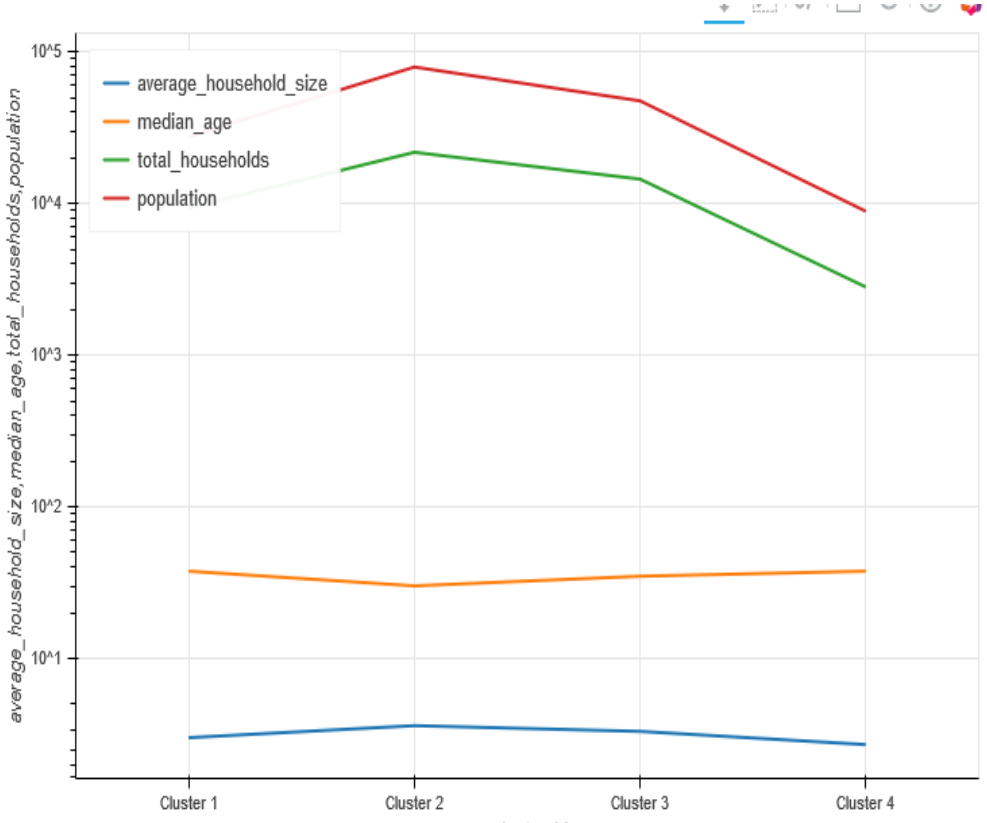


The charts show how the clusters compare with respect to total number of household, and population. It is clear from the analysis:

- Cluster 2 ranks high when it comes to high population and number of households. This has a positive and negative implications for people who relocate. Likely to include many communities, restaurants, and shops, while less of natural landscapes
- Cluster 3 is a close cousin of cluster 2, like in the case of shooting incidents
- Cluster 4 is arguably ideal one for sparsely populated area
- Cluster 1 is one with less population with good blend of natural outdoor areas

CLUSTER COMPARISON IN TERMS OF FEATURES



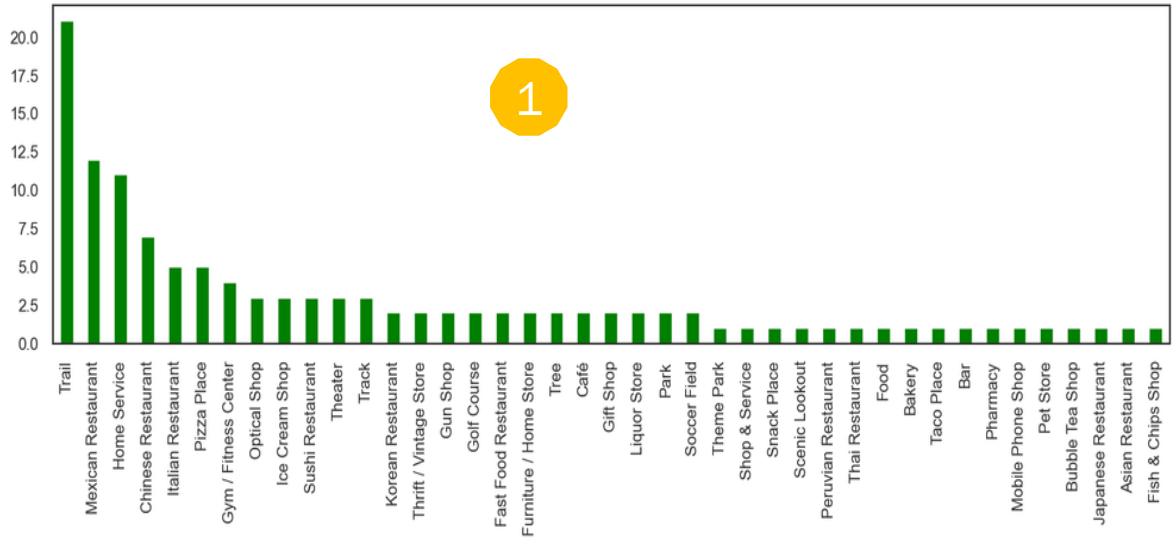


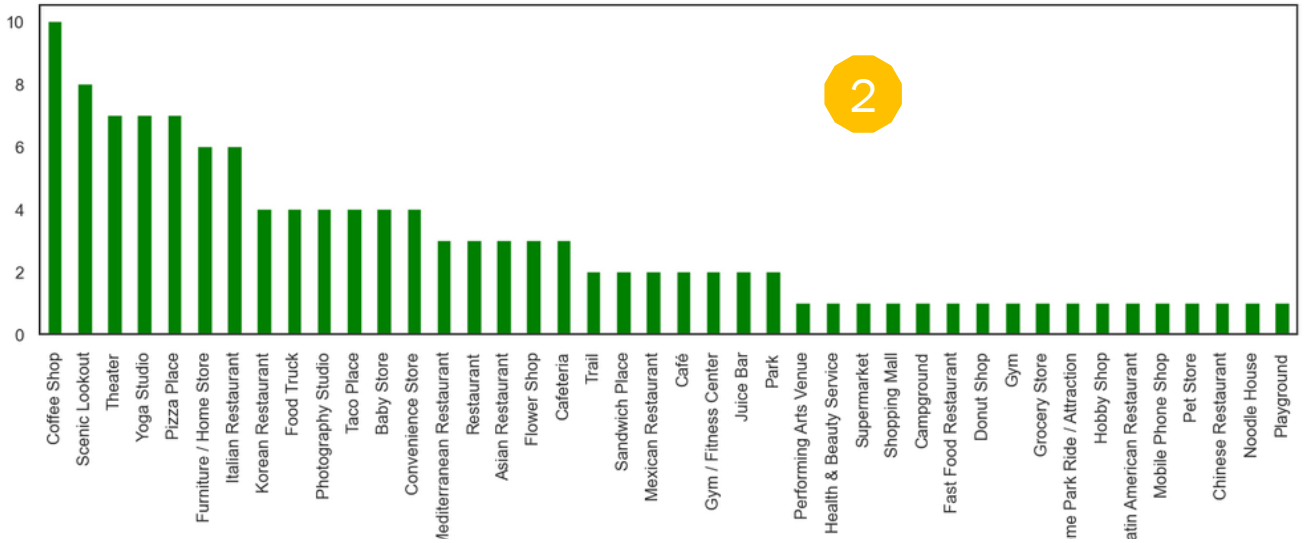
The line plots show the feature wise comparison of the clusters for easier illustration and comparison

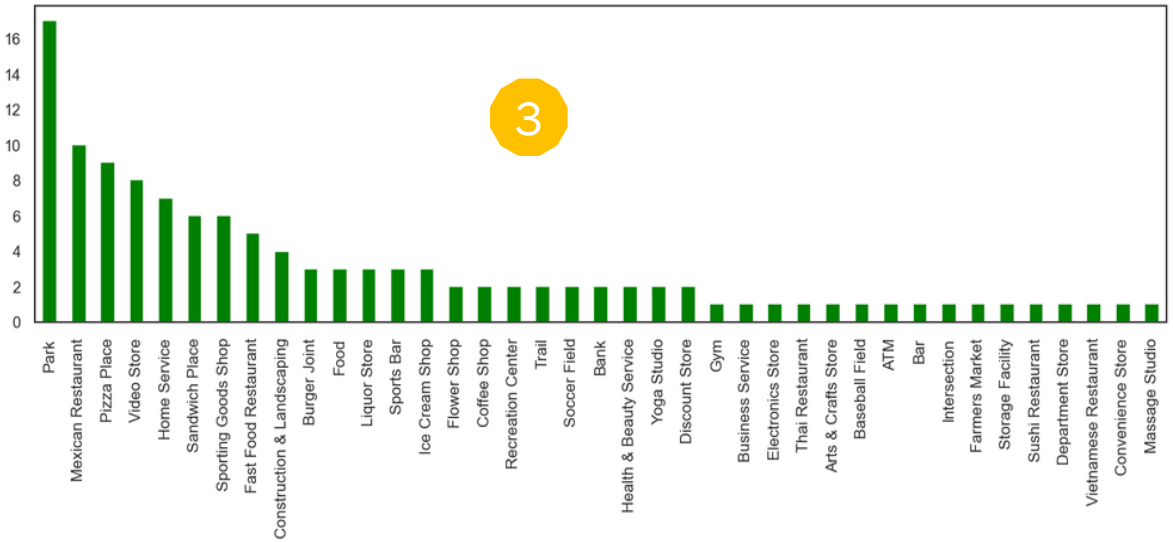
CLUSTER DEFINITIONS AND COMPARISON ACROSS THE FEATURES

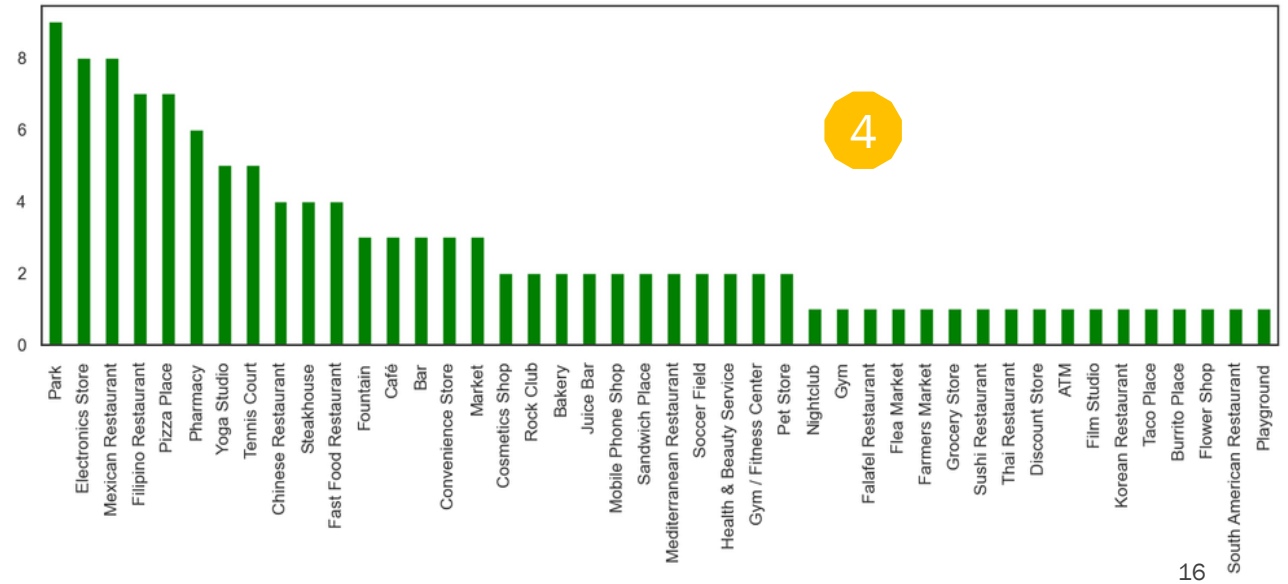
Cluster ID	Mean Population	Mean Age	Number of Households	Avg Household Size	Shooting involving Human injury	Non-Animal Shooting Incidents Of Concern	Animal Shooting Incident	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
1	27705	37.60	9390.6	3.0	0.8	1.9	0.8	Trail	Coffee Shop	Park	Park
2	79555	30.09	21789.6	3.6	3.4	5.4	5.6	Mexican Restaurant	Mexican Restaurant	Shopping Mall	Pizza Place
3	47588	34.83	14497.2	3.3	2.5	5.0	2.3	Mexican Restaurant	Mexican Restaurant	Home Service	Pharmacy
4	8896	37.61	28145	2.7	1.0	1.7	0.7	Mexican Restaurant	Mediterranean Restaurant	Ice Cream Shop	Yoga Studio

TOP 4 MOST COMMON AVENUES FOR CLUSTER 1

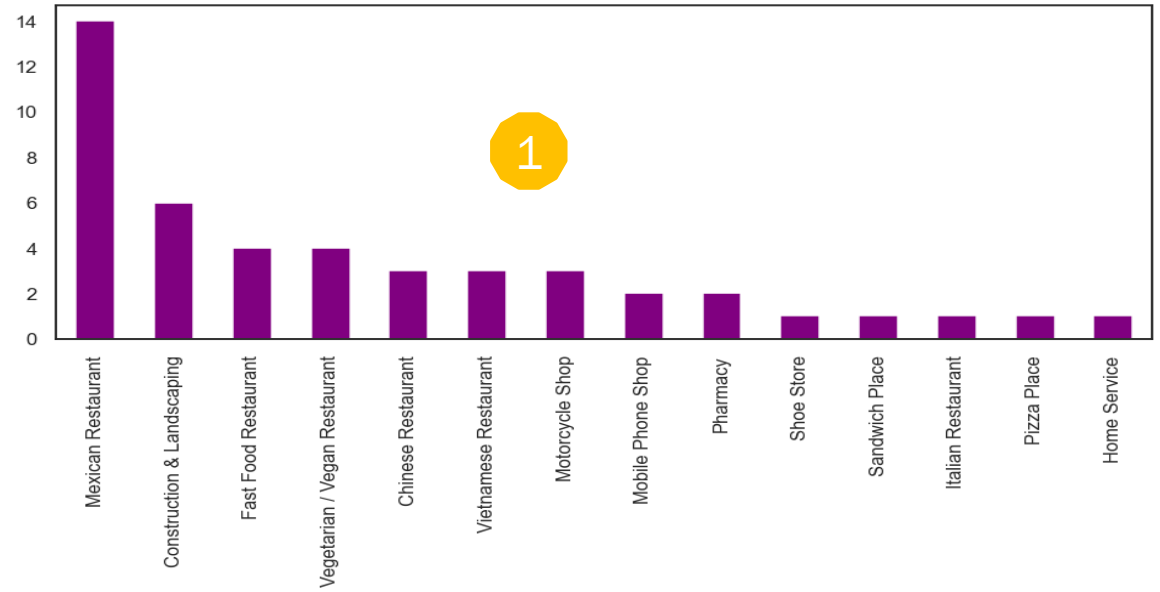


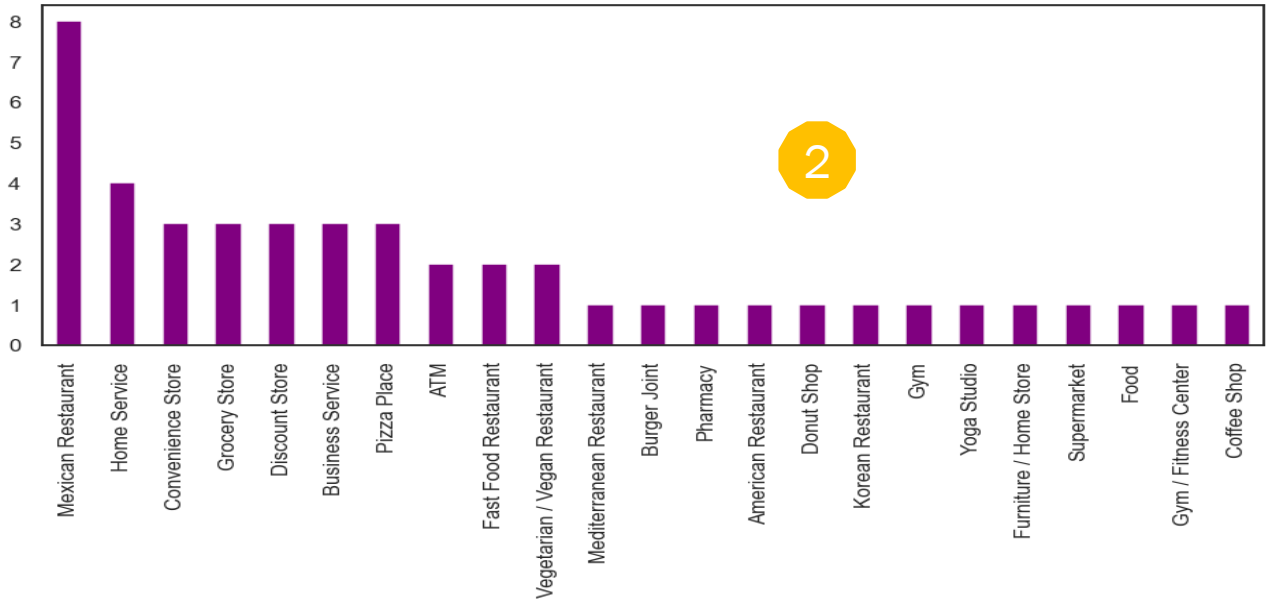


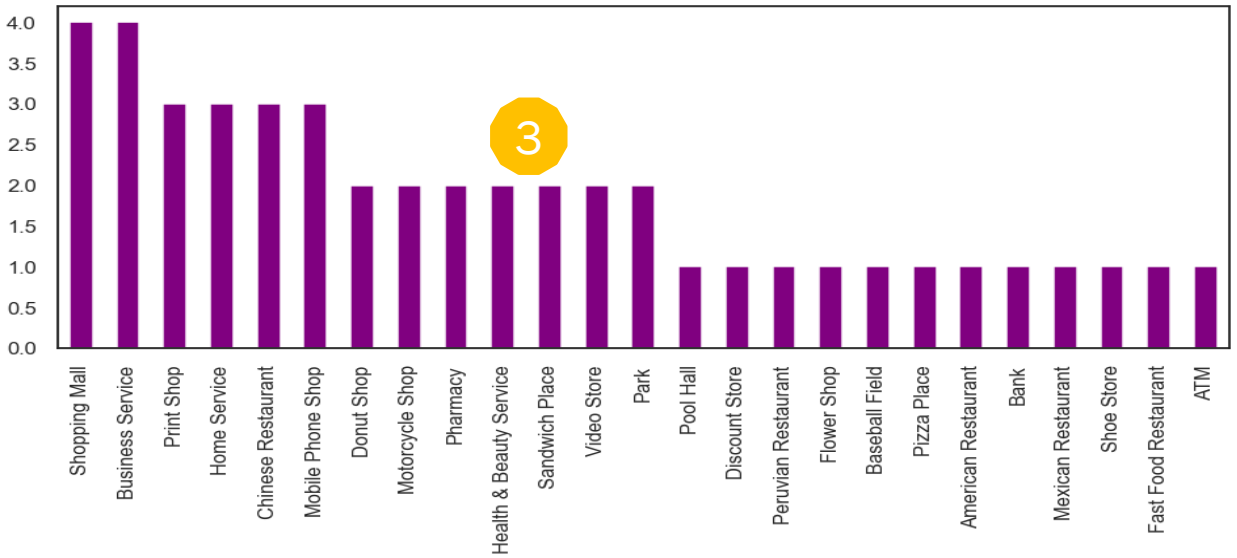


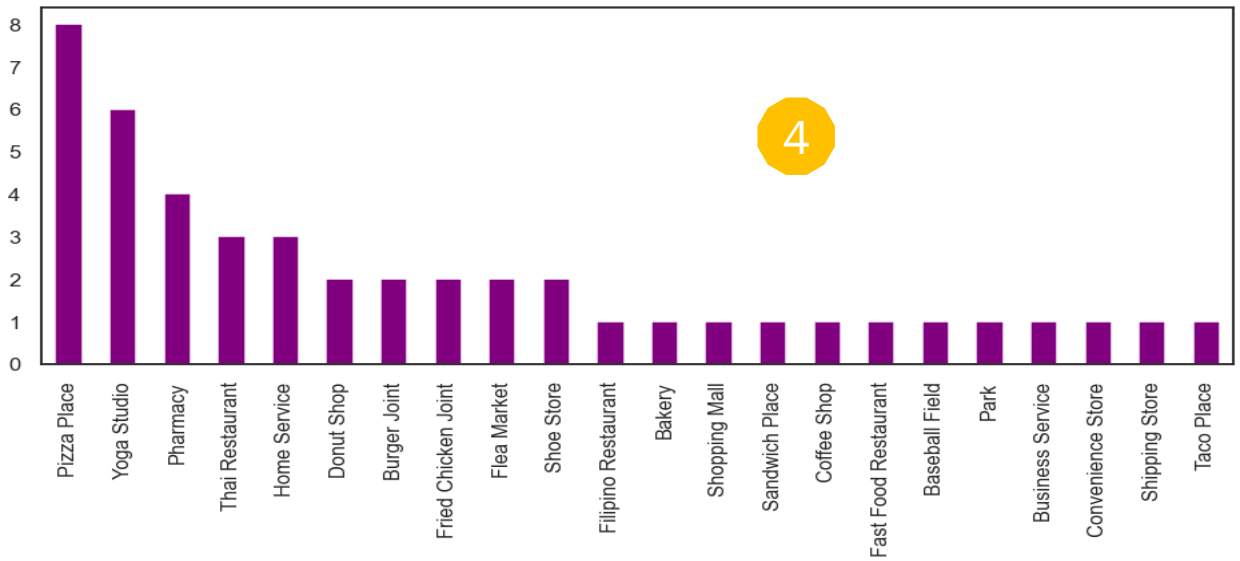


TOP 4 MOST COMMON AVENUES FOR CLUSTER 2

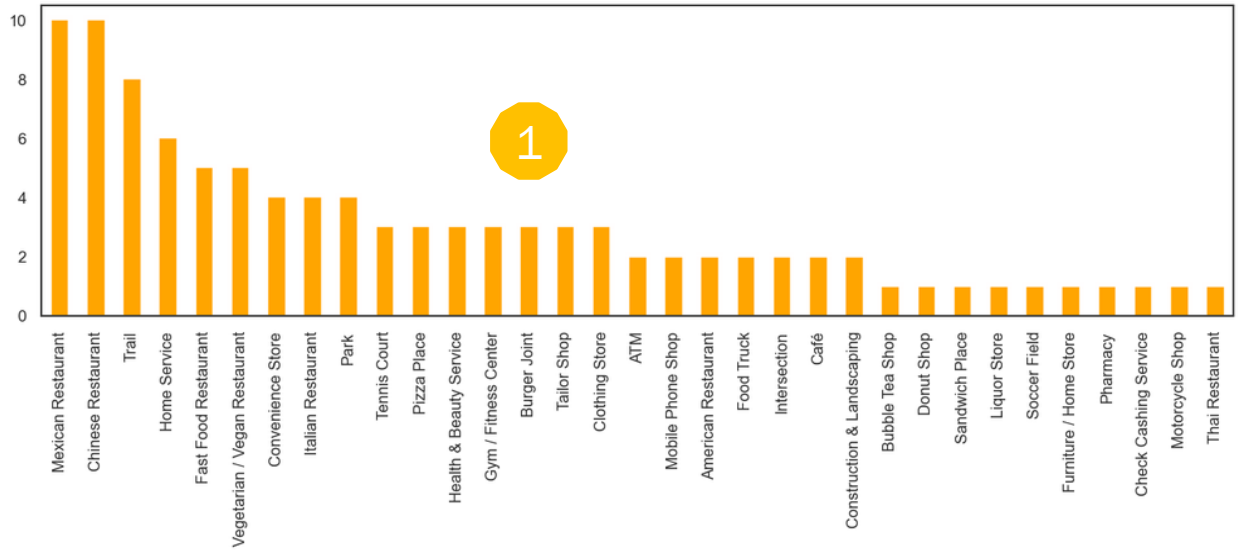


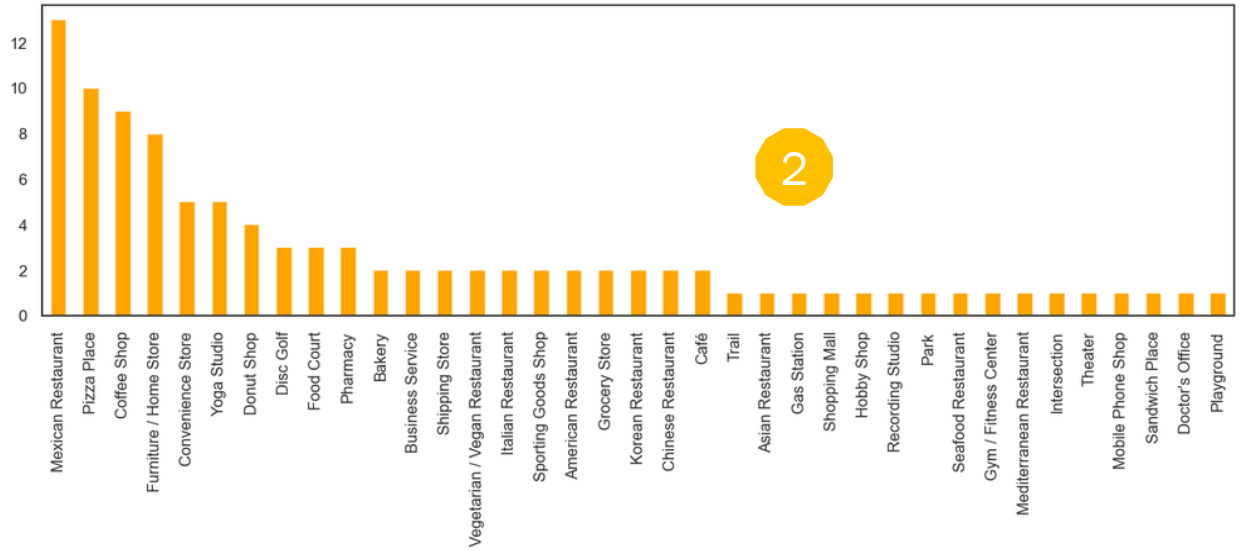


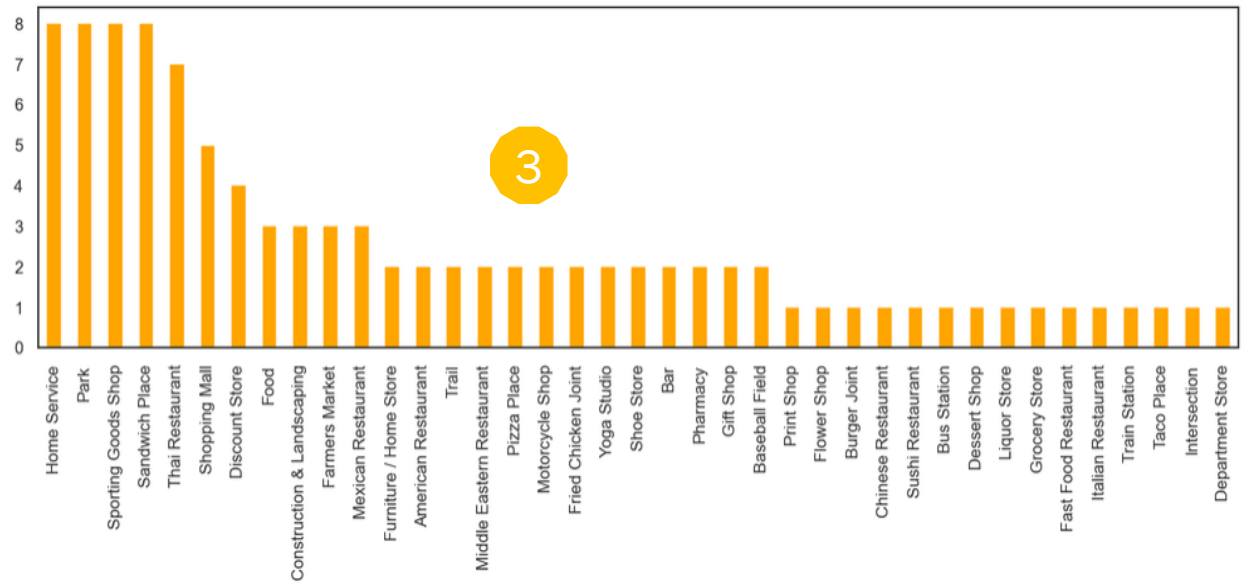


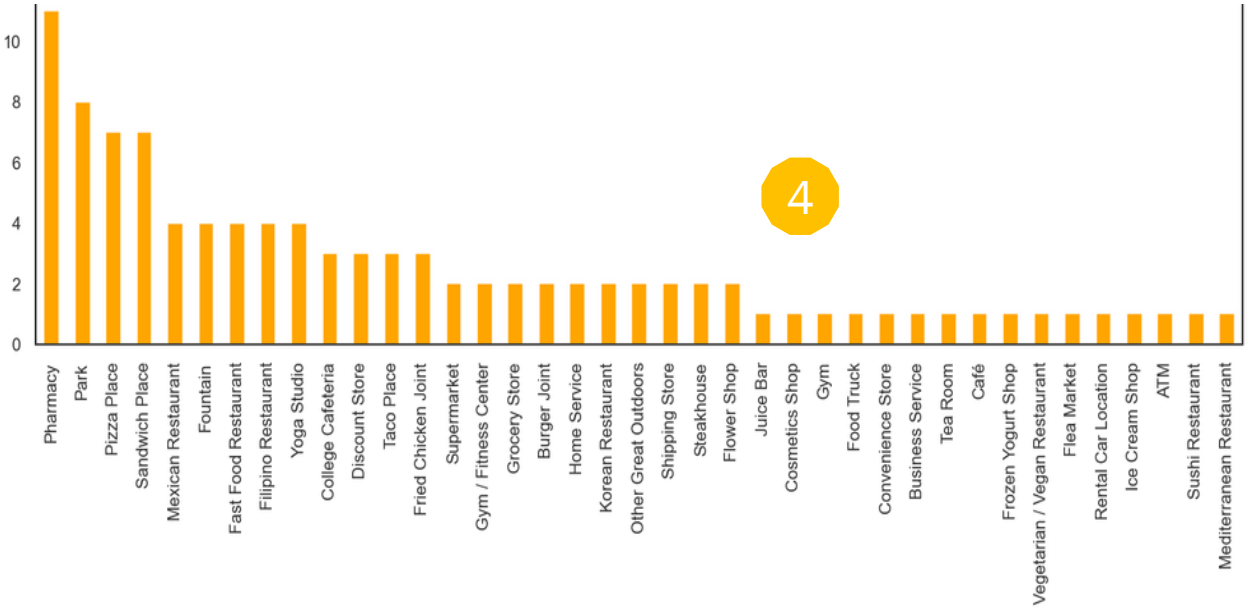


TOP 4 MOST COMMON AVENUES FOR CLUSTER 3

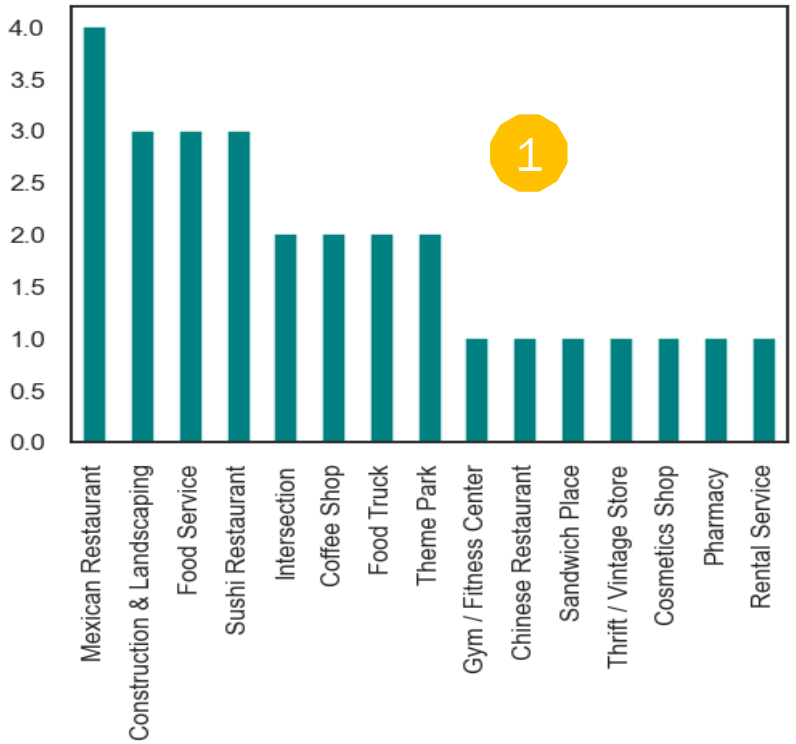


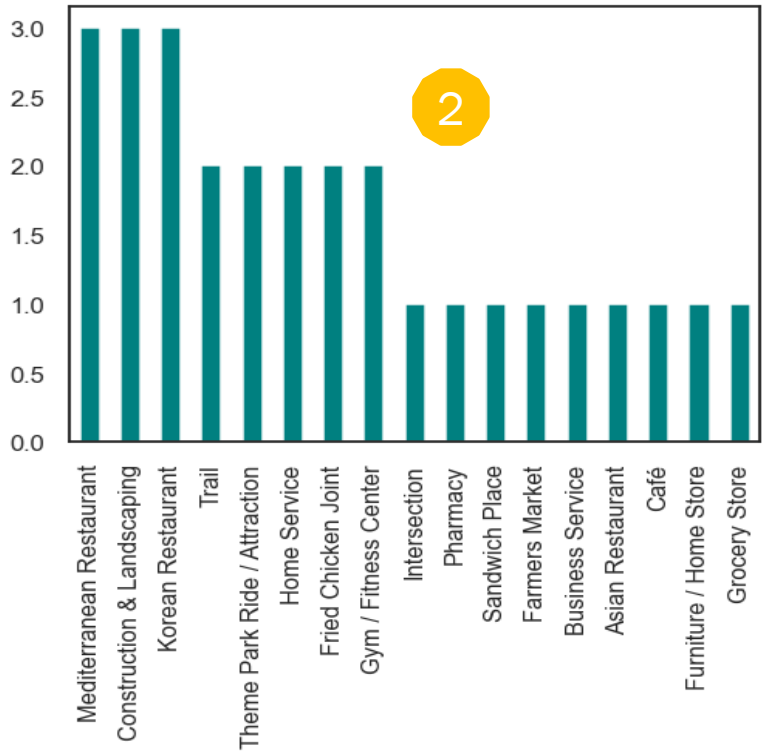


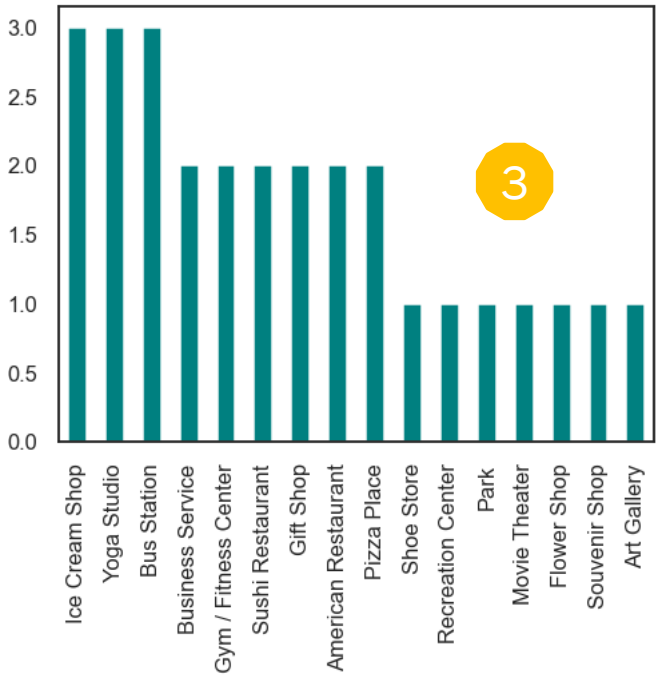


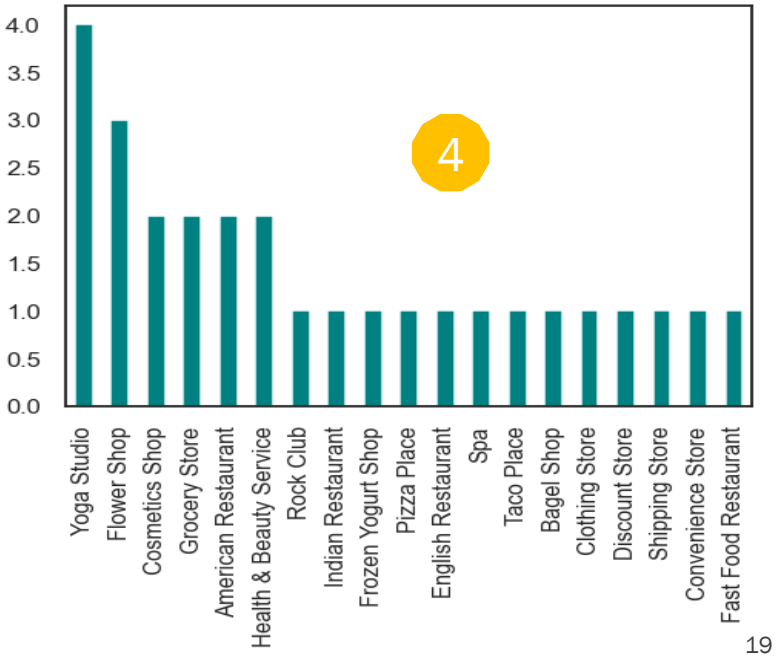


TOP 4 MOST COMMON AVENUES FOR CLUSTER 4









SUMMARY

- Out of the four clusters identified,
  - **Cluster 1:** The neighborhood in this cluster are suitable for those who like a scenic surroundings with park, and natural landscapes with low crime rate (shooting incidents), and moderate population with slightly higher age group.
  - **Cluster 2:** The neighborhood in this cluster involves high shooting incidents, lower age group individuals, high population, with lot of restaurants and shopping complexes.
  - **Cluster 3:** The neighborhood in this cluster involves moderately high shooting incidents, second lower age group individuals in the cluster, high population, with lot of restaurants and other businesses.
  - **Cluster 4:** The neighborhood in this cluster are suitable for those who like restaurants involving different cuisines, health clubs, seeking safe, and low crime rate (shooting incidents), and very low population density with slightly higher age group individuals
- As illustrated in this short exercise, with additional data one can create more meaningful clusters to place relocating people in suitable clusters to help narrow down their search. In that sense, this is a capstone project and a small proof of concept demonstrating the feasibility of this kind of idea.

RECOMMENDATIONS

- Adding more datasets to the study will help bring out additional features and criteria
  - As an example, adding more crime related datasets like burglary etc. in addition to just shooting incidents
  - Adding more dataset related to people of special needs
  - Adding features related to senior / adult assisted living
- Adding more features will also help form better, and more meaningful clusters
- Expanding the region to include larger area, adding more zip codes could help form larger, and possibly better clusters
- The same concept can be extended to include epidemic / outbreak related incidents which would be particularly helpful during a pandemic such as Covid-19
- The same concept can be extended to help people who travel for vacation by taking the interests of people and combining transportation related data