IBM-Coursera
Applied Data Sciences
Capstone Project
08/2020

Abstract

This is a presentation for IBM-Coursera - “Applied Data Sciences” Capstone project. The study focusses on finding the most liveable neighbourhoods by clustering of Foursquare venue data. It limits the scope to exploring various venue categories, along with other key influencers, like Population, Home-prices, and COVID, before moving into a neighbourhood. I introduce a methodology – “idY”, a wholistic guidance to transform an idea into a productive app. This analysis by itself is not exhaustive or complete by any stretch of imagination. It simply lays ground to an idea and approach for building Data Sciences enabled intelligent tools. A tool to make decisions based on open data sources. But one that can take you beyond common intuitive, incremental insights, to make better life choices. I have built this analysis as a reusable Jupyter Notebook Template. Please feel free to copy the template and explore away. Don’t forget to give a ThumbsUp to “idY”.

Data Sciences to Explore Liveable Neighbourhoods

“idY” METHODOLOGY

RAMESH YELISETTY, CONSULTANT ALL THINGS BIG AND SMALL

Jupyter Notebook Reference:
Coursera-IBM CapStone Liveable Neighbourhoods DC.ipynb

NYUG, INC. | ramesh.yelisetty@nyuginc.com

Contents

1. Overview

It is often a challenge to find the most appropriate place to live. We often go by intuitive takes, or concerns that linger at the top of our attention span. As we go live at that given neighbourhood, we soon ponder, "what if", I had just known better. Hence, I make my research/ project statement:

"Explore liveable neighbourhoods"

This body of work presents a simple analysis to enable identification of neighbourhoods (group of zip codes) that are closer to what may be suitable to one's living requirements. It could be converted to a simple app. An app that could evolve to be a common window to explore neighbourhoods.

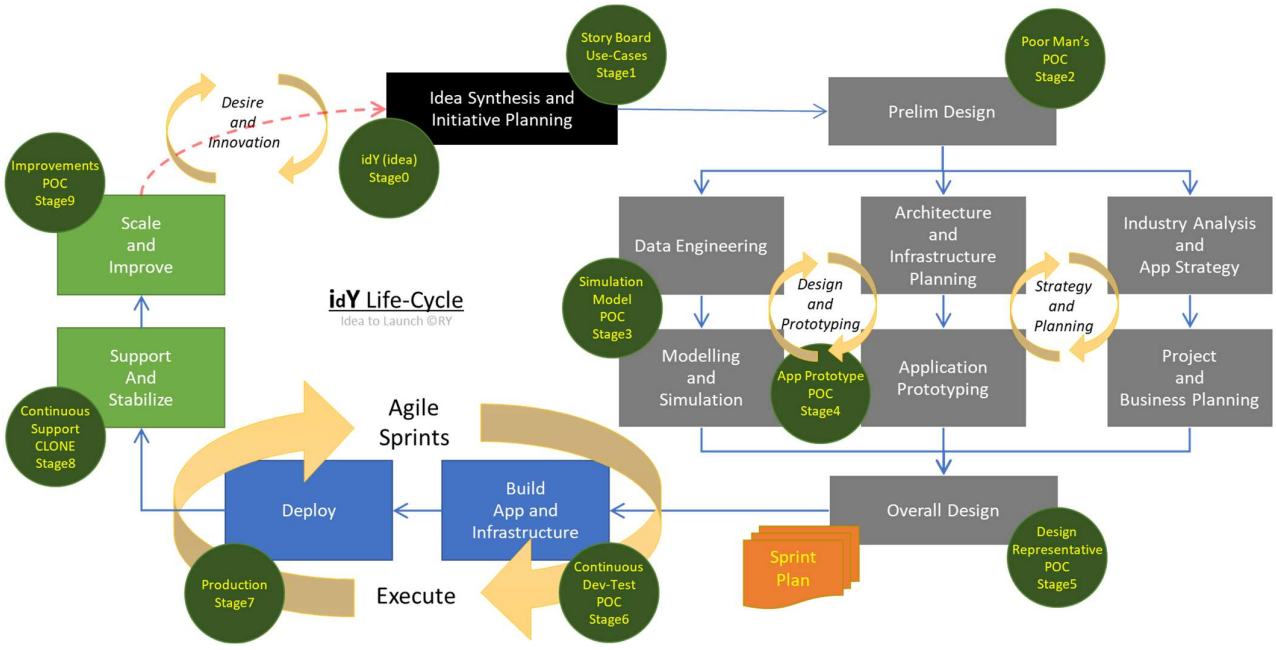
I make no claims to the completeness or superior efficacy of this body of work. I would only expect this to be a starting point to build a more significant application or utility for the future. Additionally, my intention is to limit myself to the broad requirements of IBM-Coursera "Applied Data Sciences" Capstone project requirements.

In the next section we will delve into the methodology "idY" before we float into various aspects of data analysis, and final conclusions.

Throughout this project, I will try to stay off mathematics. I am seriously not doing anything too complex here. So, no sigmas, epsilons, or any other such exotic variables or constants. Let's just chill with simple python code.

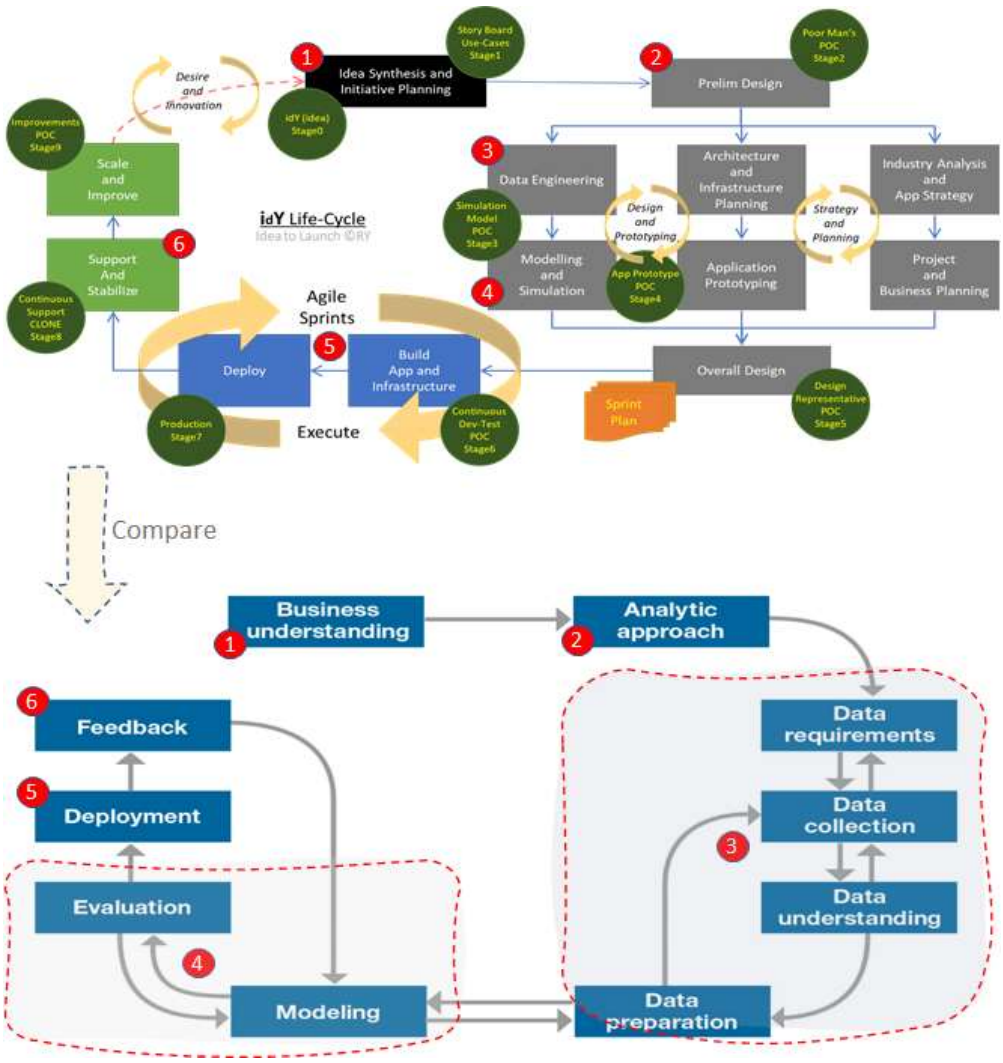
2. "idY" Methodology

idY is a wholistic methodology that provides guidance from idea to launch. It is not limited to Data Science tasks. "idY" can be of considerable advantage to deliver an initiative, in a way, that maximizes the overall outcome, while reducing waste/cost.



"idY" is consistent with "Foundational Data Sciences Methodology", as illustrated here:
https://www.ibmbigdatahub.com/blog/why-we-need-methodology-data-science

"idY" roughly translates to Foundational Data Sciences Methodology, as illustrated below.



The purpose of this document is not to explore "idY" methodology. It is to present an approach to identify liveable neighbourhoods, as a demonstration of beginner-level Data Sciences capabilities. We will focus on only those Data Sciences oriented aspects of idY here. Other idY steps can be explored separately, to convert this basic idea into a viable product for market.

3. Playing by "idY"

We will now walkthrough relevant "idY" methodology steps rapidly to build upon our focus.

1. Idea Synthesis
    - Business or need understanding/assessment
    - Idea to use cases
2. Prelim Design
    - Analytic Approach
    - Define a High-Level approach to the overall problem
3. Data Engineering
    - Data Requirements->Data Collection->Data Understanding->Data preparation
    - Define Data Chain – Sources, Clean-up and Other processing requirements
4. Modelling and Simulation

- Data modelling and Evaluation
- Use Data Sciences platform to simulate and model core processing requirements

5. Build and Deploy
    - Not relevant to this project
6. Support Stabilize and Improve
    - Not relevant to this project

4. Idea Synthesis

Liveability can mean different things to different people. We could build a complex scoring/ranking metrics, but often complex evaluations become either incomprehensible or emotionally inaccessible from a common decision maker's perspective. Hence, we will take a less precise approach to decision making, with a potential to building a coherent model, if the patterns are consistent.

Often life decisions are not binary or discrete choices. They are complex amalgamation of preferences and compromises. Data or information alone cannot lead to decision end-points, as much as, unstructured emotional preferences also cannot. A good mix of the two could lead to a decision with a relatively more sustainable satisfaction level, or so I would like to believe.

Story boards and use cases can be a good way to evolve ideas and setting realizable goals.

4.1 Use-Cases

As a step forward, I would like to encapsulate my analysis objective into a few use-cases as listed below,

1. Families looking for residential neighbourhoods, with schools, shopping, and other such family friendly services. This middle-income demographic typically wouldn't hesitate to drive.

2. Young individuals or couples who prefer to have a more outward lifestyle with nightlife, bars, good food around them; have a fun life. Some of them might prefer to live closer to their areas of work.

3. Older folks or empty nesters, who value a quieter lifestyle with easy of access to medicare and other life services.

4. A bird's-eye view for policy makers and agencies/ businesses, so that they can identify improvement potential by areas to attract any of the above groups.

5. Preliminary Design

As we go about building design outline, I would like to adhere to some of the guidelines laid out for the Capstone. Especially, the usage of Foursquare API. Outside of this Capstone, we could envisage more open and exploratory research, nevertheless adopting a similar approach.

I would like to look at this project as largely driven by a central driver, which would be influenced by multiple supporting or influencing factors to arrive at a final decision.

Central idea for this analysis:

> Foursquare provides data on various venues by neighbourhood. Use this venue data to classify neighbourhood into distinct clusters.

Supporting ideas to influence the central driver:

> Liveability is by no means a mere aggregation of schools, shops or nightlife around you. There are other aspects that need to be taken into consideration to make a decision. The clusters derived from the above venue analysis can be further evaluated for other aspects, such as,

- Population – Population and density
- Home Prices – affordable homes
- COVID Spread – sensitivity to diseases - how did COVID play; if there aren't drastic structural adjustments, maybe, maybe, the next pandemic will chase the same fault-lines
- Indices indicative of affordability – cost of living, besides home prices
- Crime indices – how safe is the place
- School ratings – How do schools here compare ?
- Incident history – major events in the past, like earthquakes, accidents, weather events, or such notable calamities
- Existence of large environmentally sensitive sites – Nuclear facilities, Chemical Plants, Alien visitations, and such
- And more of such factors that may impact your primary decision approach.

This project will analyze the first three factors listed above – Population, Home Prices, and COVID.

5.1 Analytic Outline

1. Prepare a ZIP code master list for the locale of interest
    - Say in Washington DC
2. Get Population Data by ZIP code.
3. Get Home Price data by ZIP code.
4. Collect venues data by ZIP code, Latitude-Longitude from Foursquare.
5. Build a feature matrix of venue and non-venue data/characteristics
6. Use a clustering algorithm to classify ZIPs into Clusters.
    - As a starting point to our analysis use kMeans method for clustering.
    - Other clustering models such as Decision Tree and DBSCAN can be assessed separately, they are outside the scope of this analysis/project
7. Explore the key characteristics of these clusters to ascertain suitability for each of the above listed use-cases.
8. Assess/Compare population and density by clusters.
9. Assess/Compare home affordability by clusters.
10. Assess/Compare sensitivity to disease spread by clusters.

5.2 Key resource requirements

Identify hardware, software and other resource requirements to execute upon the above analytical outline. For this purpose of this Capstone project, I would limit my imagination to identifying python packages, relevant to this analysis and simulation. We will use Jupyter notebook as a platform for analysis, on a local machine(laptop). All data sources used in this analysis are freely available over internet.

Refer: Jupyter Notebook, Section 5.2

5.3 Key Parameters

Parameterize some of the key data input variable that will help generalize the data model and application simulation. For example, by simply changing the P_select_state to say “NY”, the program could be run for NY state dataset. This will ensure minimization to changes to core code for analysing additional or alternate datasets.

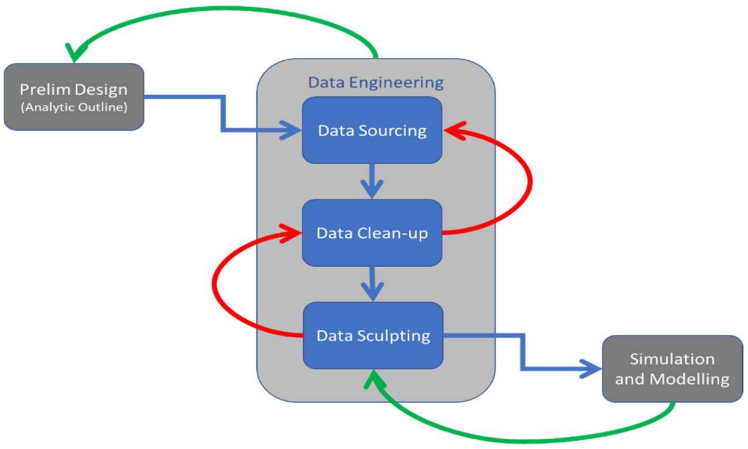
The main sections of this document present a simulation model with DC as the reference state, P_select_state = “DC”. Analysis of Additional states is presented in Annexure 10.1.

Refer: Jupyter Notebook, Section 5.3

```
#key parameters
P_select_state = "DC" # state in USA for analysis
P_POP_data_source = "Data/Population_uszips.csv" # data source population data by ZIP
P_ZHVI_data_source = "Data/ZHVI_uszips.csv" # data source Home-price data by ZIP
P_ZHVI_date = "06/30/2020" # effective date of Home-price data by ZIP
P_COVID_cases_data_source = "Data/covid_confirmed_usafacts.csv" # COVID confirmed cases by County
P_COVID_deaths_data_source = "Data/covid_deaths_usafacts.csv" # COVID deaths by County
P_COVID_date = "07/08/2020" # effective date of Home-price data by ZIP
P_path = r'D:\GIT\IBM_Coursera\Coursera_Capstone\Data' # file path to store venue counts extracts from
Foursquare
P_4sq_CLIENT_ID = 'N233FNIITUBDDXZ4OR5AWM0YEY1XXVKWP2L2IAZOWISCPLWC' # your Foursquare ID
P_4sq_CLIENT_SECRET = 'YHREMIYEIW5TLJQYFDZGVHZVTW13AFPT3GLBJRS11KWRC2OP' # your Foursquare Secret
P_4sq_VERSION = '20180605' # Foursquare API version
P_rainbow = ['Red', 'Orange', 'Green', 'Blue', 'Yellow', 'Brown', 'Black', 'Purple'] # color pallet for
clusters
P_default_zoom = 8.0 # default zoom for visualization of clusters on a map, set 12 for small states
like DC, 6 for large states like CA
```

6. Data Engineering

Data engineering as per "idY" follows a 3-step sequence:



6.1 Data Sourcing

Identify and evaluate data sources. The evaluation process must consider – availability, goodness of fit, quality/consistency, cost, and other strategic/business considerations. Check the scope of analytics that is feasible with the identified data sources. There should be an obvious preference for data sources that have the capability to scale or support a wider scope of analysis. This would ensure that we can add more functionality to our app as we evolve. For the scope of this project, we have limited our appetite to data availability at a ZIP code level, and free of charge.

Ref: Jupyter Notebook, Section 6.1

Data sources used, include:

a. Population Data
   - Source: https://simplemaps.com/data/us-zips
   - Type: Download CSV file
   - Data freely available by ZIP, State, and County
   - Other demographic details can also be fetched by paid subscription

b. Home Price Data
   - Source: https://www.zillow.com/research/zhvi-methodology/
   - Type: Download CSV file
   - Data freely available by ZIP, State, and County
   - Data is available in a time-series format by month-end; for this analysis choose the latest published data

c. COVID Data
   - Data available in 2 parts, separately for Confirmed Cases and Deaths
   - Source: https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/
   - Type: Download CSV files
   - Data freely available by State, and County
   - Data is available in a time-series format by month-end; for this analysis choose the latest published data

d. Venue Data
   - Source: https://api.foursquare.com/v2/venues/
   - Type: API call
   - Call limit of 1000/day with unregistered account; 100000/day with a registered account

6.2 Data Extraction and Clean-up

All Data sources do not follow the same access and clean-up pattern. Each of the data sources need specific processing steps to make them consistent and useful. No matter how good the data source is, there will be a need to check for gaps in the data. Assess impact of these gaps on the overall analytical goals. Define a strategy to address these gaps. If the data does not meet the requirements as defined, do we have the flexibility to redefine the analytic strategy or the requirement itself. Nice try! Don't touch the requirements. Could we go back to sourcing ? Explore additional data sources to enrich existing data sources or extrapolate the given data to fill for gaps.

Ref: Jupyter Notebook

Population Data: Section 6.2.1

Home-price Data: Section 6.2.2

COVID Data: Section 6.2.3

Venue Data: Section 6.2.4

Each of the csv data sources (as listed in the previous section) can be read using standard pandas function(read_csv) to read CSV files. Clean-up the dataframes for blank, Null, or invalid values. Some of these steps specific to individual data sources can be seen in Section 6.2 of Jupyter Notebook. Some of the clean-up tasks include,

- Dropping NAs
- Clean-up County names to be consistent
- Eliminate Unallocated COVID data records
- Drop blank Home-price records
- … and more, based on specific datasets

Some of the clean-up tasks could lead to loss of data, which has been ignored in this analysis, with an expectation that these incomplete lost records will not have a noticeable impact on the analysis. This would require a more concerted view when a production grade application is built.

Extracting venue data from Foursquare has been quite a challenge. It takes over 6-8Hrs for larger states like California, or New York. This can be a serious constraint while developing the program. Hence, make a set of backup venue counts database (CSV files in this case) to store the Foursquare data relevant to this analysis. These backup files can then be used to recreate venue data, when required.

Here is a code snippet to define credentials for accessing Foursquare API,

```
#Set Credentials
CLIENT_ID = 'XXXXXXXXXXXXXXXXX' # your Foursquare ID
CLIENT_SECRET = 'XXXXXXXXXXXXXXXXX' # your Foursquare Secret
VERSION = '20180605' # Foursquare API version
print('Your Credentials:')
print('CLIENT_ID: ' + CLIENT_ID)
print('CLIENT_SECRET:' + CLIENT_SECRET)
```

Define a reusable function to extract data from Foursquare,

```
def get_venues_count_ll(lat,lon, radius, categoryId):
    explore_url = '
https://api.Foursquare.com/v2/venues/explore?client_id={}&client_secret={}&v={}&ll={},{}&radius={}&c
ategoryId={}'.format(
                CLIENT_ID,
                CLIENT_SECRET,
                VERSION,
                lat,
                lon,
                radius,
                categoryId)

    # make the GET request
    return requests.get(explore_url).json()['response']['totalResults']
```

Foursquare calls can sometimes time-out or give network error. Write the extract program, in a way that can give some flexibility to start mid-way, in case of failures, as illustrate in Jupyter Notebook, Section 6.3.4.

```
import time
RADIUS = 2000 # 1000
processing_start_index = 0 # 0
file_start_index = 0 # 0
block = 200 # 200
message_block = 25 # 50
df_size= len(EXTRACT_venues_df)

processing_blocks = [(x, min(x+block,df_size)) for x in range(processing_start_index,
df_size, block)]

main_df = pd.DataFrame()

for n, p in enumerate(processing_blocks):
    start = p[0]
    end = p[1]

    print(" Start Index: {}, Start Time: {}".format(start, time.ctime(time.time())))
    process_df = EXTRACT_venues_df.iloc[start:end,:].copy(deep = True)
# anaconda warning, dont write to a slice, chain indexing risk

     for i, row in process_df.iterrows():

        if i%message_block == 0:
            print("     Start Processing Index: {}, ZIP: {}, at: {}".format(i,
row['ZIP'], time.ctime(time.time())))

        for c in venue_cats.items():
            try:
                state = process_df.State.loc[i]
#                 zip = process_df.ZIP.loc[i]
                lat = process_df.Latitude.loc[i]
                lon = process_df.Longitude.loc[i]
                process_df.loc[i, c[0]] = get_venues_count_ll(lat, lon, radius=RADIUS,
categoryId=c[1])
```

Continued…code snippet

```
            except:
                process_df.loc[i, c[0]] = 0

    pd.concat([main_df, process_df])

    #save to file
    filename = P_select_state+'_bkup_venues_counts_{}.csv'.format(str(n+file_start_index))
    file_location = os.path.join(P_path,filename)
    process_df.to_csv(file_location)

    print(" End Index: {}, End Time: {}\n          File: {}".format(end-1,
time.ctime(time.time()), filename))
```

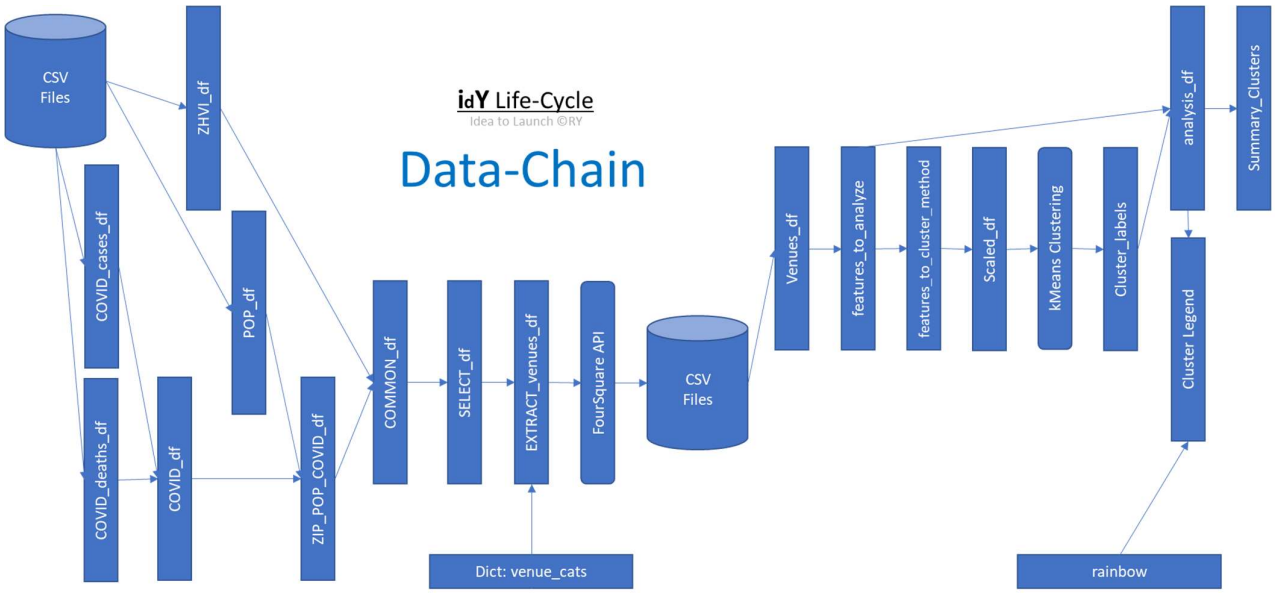
6.3 Data Sculpting

Sculpting implies taking cleaned data and preparing it for the required analytics goals. Some of the key aspects to keep in mind while sculpting your data,

1. Performance, how long does it take to run ?
2. Flexibility of data structure to expand the scope of analysis
3. Is there a need to change data extract and clean-up strategy ?

"idY" runs with a strong belief in doing things right, in-line with an evolving (not static/rigid) expectation, hence, all the feedback loops that take you back – "a step forward and another back is better than rushing 20 steps forward to a crash landing." Define data-chain and evaluate integration touchpoints to the ongoing app design. It maybe necessary to redo your data sculpting strategy based on the needs of Simulation and Modelling exercise.

Data Chain to meet the analytical demands of this project is illustrated below.



Ref: Jupyter Notebook, Section 6.3

Some of the key sculpting tasks include,

- Derive COVID Cases and Deaths by ZIP – 6.3.1

    - COVID data is available by county. Calculate population data by County and use it as the denominator to distribute the COVID data by ZIP.
        - COVID-ZIP = (COVID-CASES(or DEATHS) * POPULATION-ZIP)/POPULATION-COUNTY
- Build a common data frame by joining all data components
    - COMMON = Join by ZIP(Population, Home-prices, COVID)
- Final Data Filter
    - Create a more focussed/reduced dataset "SELECT" from the COMMON all-inclusive dataset
    - If there is scope for reducing the data size based on certain criteria in the interest of performance; this analysis will include all the ZIP records from Washington DC dataset
    - Some of the examples of this filtering could be,
        - Include only those records that belong in the Population inter-quartile range
        - Include only those records that have home prices within a required range
- Venue Data
    - Pull data from Foursquare using the functions and credentials defined in the previous section (Jupyter Notebook: 6.2.4)
    - Use SELECT dataframe as the ZIP reference to call the Foursquare API
    - The given call for CA state ZIPs created 9 files, for the selected call parameters,
        - Venue categories limited to venue_cats dictionary,

```
venue_cats = {
'Arts & Entertainment': '4d4b7104d754a06370d81259',
'Event': '4d4b7105d754a06373d81259',
'Nightlife Spot': '4d4b7105d754a06376d81259',
'Outdoors & Recreation': '4d4b7105d754a06377d81259',
'College & University': '4d4b7105d754a06372d81259',
'Travel & Transport': '4d4b7105d754a06379d81259',
'Professional & Other Places': '4d4b7105d754a06375d81259',
'Food': '4d4b7105d754a06374d81259',
'Shop & Service': '4d4b7105d754a06378d81259',
'Residence': '4e67e38e036454776db1fb3a',
'School': '4bf58dd8d48988d13b941735',
'Medical Center': '4bf58dd8d48988d104941735',
'Hospital': '4bf58dd8d48988d196941735',
'Spiritual Center': '4bf58dd8d48988d131941735',
}
```

      The above list of categories includes the top most hierarchy published by Foursquare, plus some more at the lower levels that may be of interest to enhance or impact liveability, such as, "Schools", "Medical Center", "Hospital", "Spiritual Center", which were part of "Professional & Other Places".
        - Radius of 2 KMs
        - this radius may not be appropriate for suburbs or rural areas, but we will proceed for now
    - Build venue count dataframe and store the data to local CSV file database
    - Extract venue data from the stored CSV file database
    - Remove null or NA data
    - Remove empty venue data
        - Remove records with zero venues data across all venues

7. Simulation and Modelling

It's time to execute our analytics strategy. For this analysis/project, lets focus on KMeans algorithm for clustering. We will execute upon the following key steps to fulfil this analysis,

1. Build a Feature Matrix
2. Evaluate for best k-value
3. Build Cluster: Perform clustering with the selected value of k
4. Evaluate individual clusters

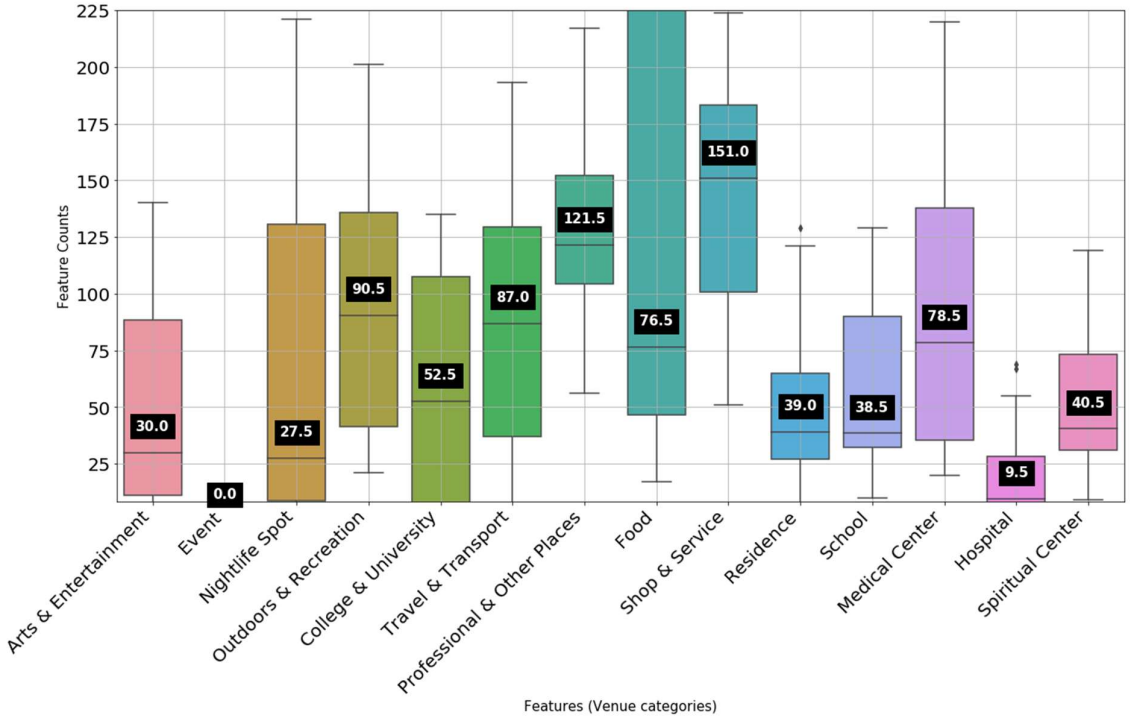
7.1 Build a Feature Matrix

Features are key characteristics or attributes to be used as a basis for identifying clusters. Our dataset includes 2 types of features:

1. Foursquare venue data which tells us about various venues available at a given neighbourhood, to enrich liveability
2. Non-venue data like population, population density, COVID cases/deaths, and home affordability, which will be further influence choice of a neighbourhood. We will not use non-venue data for clustering but will analyze the same for clusters derived using venue data, to make the final recommendations/decisions on liveability.

Ref: Jupyter Notebook, Section 7.1.2

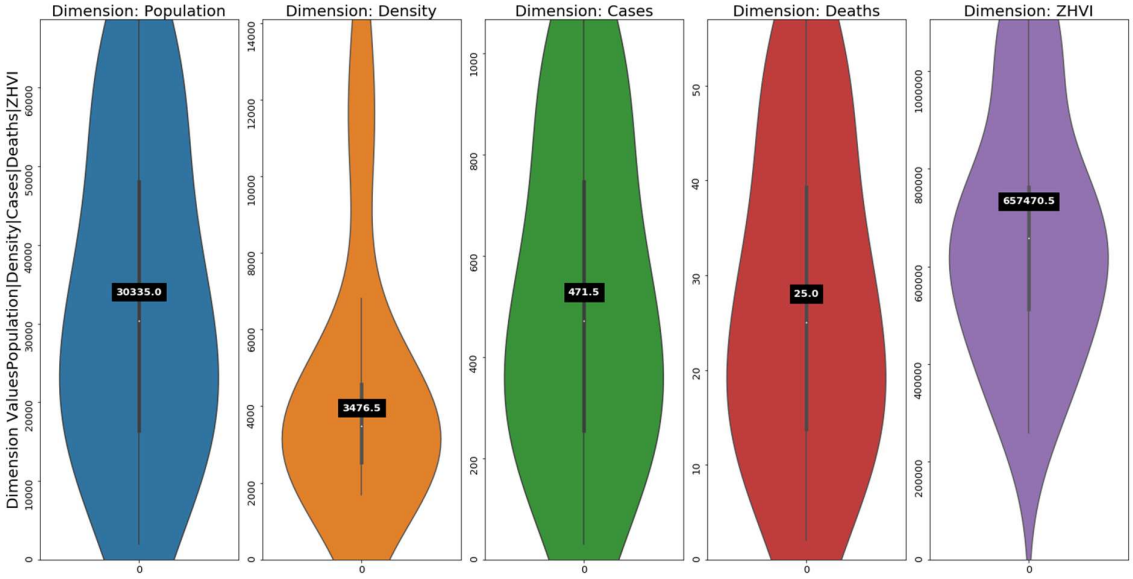
Box-plot of Venue Feature Counts



Analysis:

- *Most venue categories seem to have a wide distribution across records, in Washington DC area.*
- *"Event" venue category seems to have a negligible presence, in DC area. We will retain it in the model to let the model be sufficiently generalized across states. A more intelligent or dynamic filter can be applied while designing the main app*

Violin-plot of non-venue features



Analysis:

- *Population, COVID Cases, COVID Deaths, all seem to have similar distributions, in DC area. This could be due to proportional distribution of one county data across all ZIPs.*
- *ZHVI plot seems to imply shortage of lower cost homes, in DC area. The violin curve seems to indicate a skew towards homes upwards of $500,000.*
- *Density plot doesn't seem tell us a whole lot besides the fact that most records are concentrated around the median.*

7.2 Evaluate for best k-value

In kMeans method of clustering, k-value implies number of clusters to be derived. We should choose a k-value that would make the most practical sense. One of the ways to derive such a value is by evaluating an error metric – SSE.

SSE, Sum of Squared Errors, is the sum of squared difference of all points in analysis domain to their respective cluster centers (centroid).

Let's stretch our imagination a little more. If we had just one cluster SSE will be huge. If we had as many clusters as points/records in our data set, SSE would be zero. None of the two extremes would make practical sense. Hence, we would like to find a minimal k-value with the maximum impact on reducing SSE. Incremental k after this k-value only reduces SSE marginally. This is often referred to as elbow-method to determine k-value. SSE is also referred to as inertia. Elbow is a point on inertia curve where the slope drastically reduces to become linear on the way down.

7.2.1 Initialize Feature Matrix

Select/filter the features to be clustered

7.2.2 Normalize Feature Matrix

Before evaluating the k-value, the feature matrix is normalized.

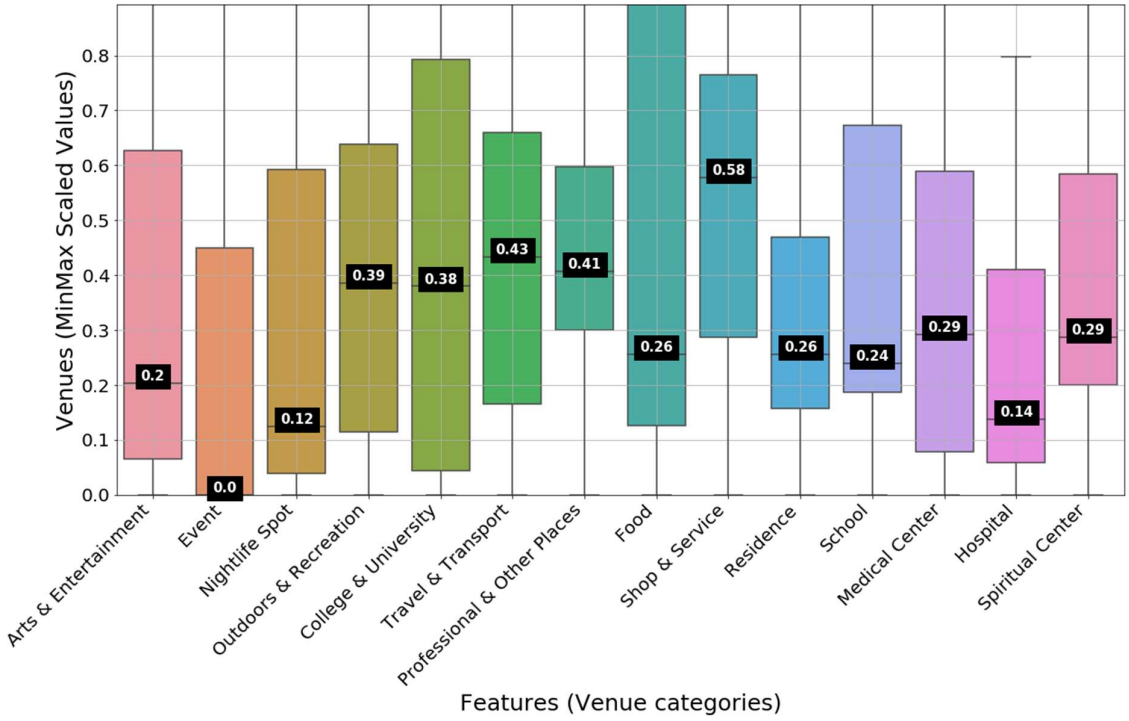
Ref: Jupyter Notebook, Section 7.2.2

```
scaled = MinMaxScaler().fit_transform(features_to_cluster_method)
scaled_df = pd.DataFrame(scaled)
scaled_df.columns = venue_cats_list
```

7.2.3 Visualize Feature Matrix

Box-plot of Normalized Feature Matrix using MinMax Scaler

Ref: Jupyter Notebook, Section 7.2.3



Analysis:

- *Values show a similar pattern to Venue counts discussed above, except for "Events" category shows a wider distribution.*

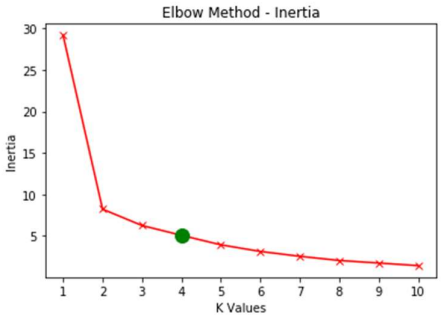
7.2.4 Inertia for k-Values

Calculate inertia for various values of k ranging from 1 to 10.

```
inertia = []
X = scaled_df
K = range(1,10)
for k in K:
    #kmean model
    Model = KMeans(n_clusters=k).fit(X)
    Model.fit(X)
    inertia.append(Model.inertia_)
```

Line-Plot of Elbow Curve

Ref: Jupyter Notebook, Section 7.2.4



Analysis:

- *k-value 4 seems like the most appropriate "Elbow" value for clustering this dataset.*

7.3 Build Clusters

Execute kMeans clustering algorithm on the above normalized dataset to derive k(4) clusters. Output of this algorithm is in the form of a data series of cluster labels.

Ref: Jupyter Notebook, Section 7.3.1

```
# set number of clusters from the above elbow curve k = 6
num_clusters = 4
# run k-means clustering
kmeans = KMeans(n_clusters=num_clusters, random_state=0).fit(X)
# check cluster labels generated for each row in the dataframe
cluster_labels = kmeans.labels_[:]
```

Combine the clusters with original feature matrix, to create a common dataset for further analysis. Add a color code to clusters. This data set has cluster reference to ZIP and non-venue data.

7.4 Analyze Clusters

7.4.1 Cluster Counts

kMeans has distributed 20 data records across 4 clusters. While RED, and ORANGE are larger clusters with 7 members each, BLUE and Green are smaller clusters with 3 members each.

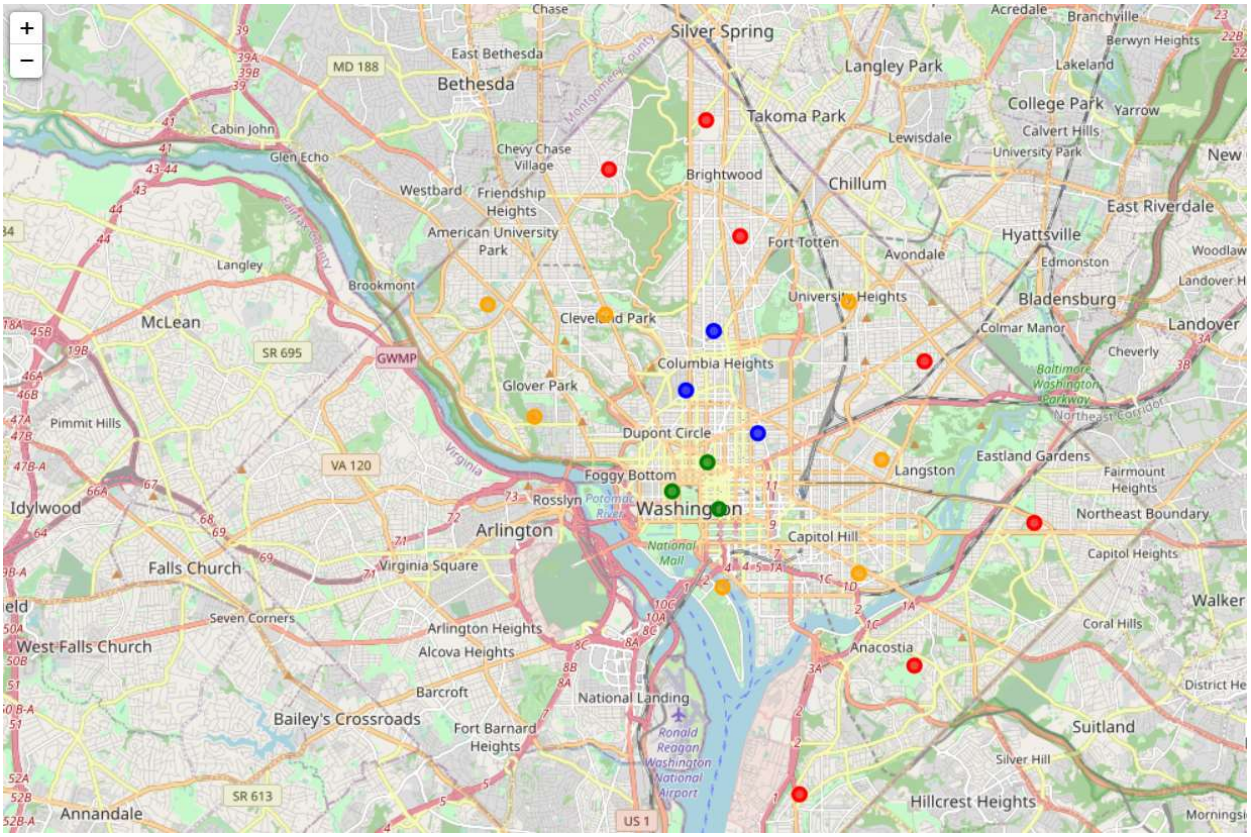
Ref: Jupyter Notebook, Section 7.4.1

Clusters	Color	ZIP
0	Red	7
1	Green	3
2	Orange	7
3	Blue	3

7.4.2 Visualize Clusters – Geographic

Ref: Jupyter Notebook, Section 7.4.2

Plot Clusters on Geo-Map



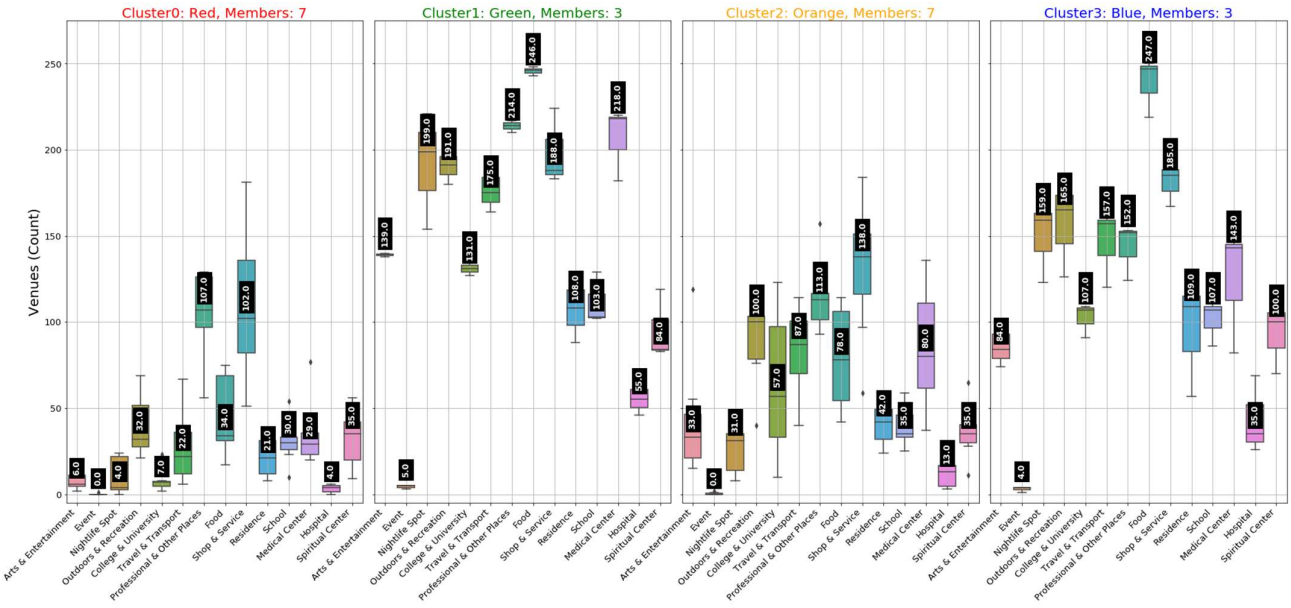
Analysis:

- *The clusters seem to have a radial pattern. GREEN -> BLUE -> ORANGE -> RED.*
- *While GREEN and BLUE represent inner areas of the city, ORANGE and RED are outskirts or closer to suburbs.*
- *GREEN and BLUE seem like thick city areas – National Mall, Capitol Hill, and DuPont Circle*
- *ORANGE and RED seems to have a touch of green - parks, gardens, and open areas*

7.4.3 Compare Clusters by Venues

Ref: Jupyter Notebook, Section 7.4.3

Box-Plots to Compare Venue Distribution Across Clusters



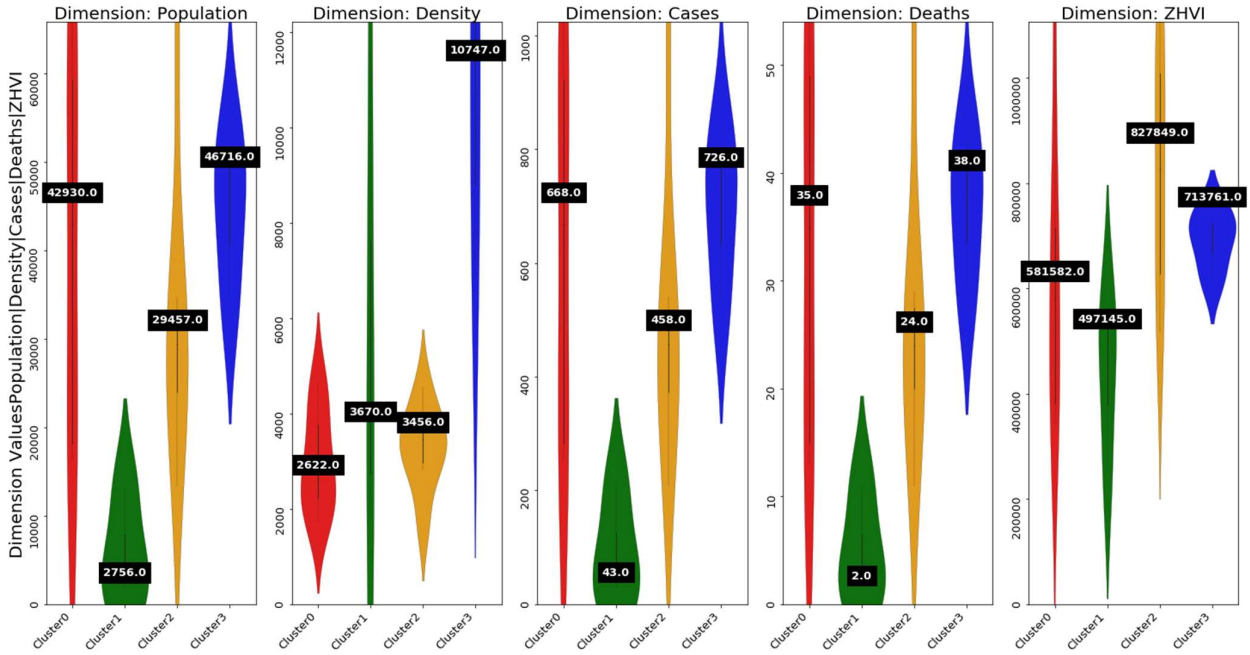
Spider-Plot to Compare Clusters by Venue Medians

<u>Evil Note:</u> Most references that I came across discouraged me from using Spider plots, especially the mixed one like you see above. I couldn't resist the temptation of using it. It looks kinda cool, don't' you think ?

Analysis:

- *All of the clusters seem to have a good distribution of various venues.*
- *Going by distribution of various venues*
  - *Cluster0-RED is comparable to Cluster2-ORANGE*
  - *Cluster1-GREEN is comparable to Cluster3-BLUE*
- *RED-ORANGE*
  - *comparable on - Spiritual Center, Professional&Others*
  - *ORANGE has superior - Art&Entertainment, Nightlife, Outdoor&Recreation, College&University, Travel&Transport, and Food*
  - *Is ORANGE more expensive(Home-prices) than RED ?*
- *GREEN-BLUE*
  - *Both GREEN and BLUE are way better than RED-ORANGE*
  - *GREEN is marginally better than BLUE on all venues, except for maybe Spiritual Centers*
  - *Is GREEN more expensive(Home-prices) than BLUE ?*

Violin-plots to Compare non-Venue Feature Distribution Across Clusters



Analysis:

- *Population of GREEN is much less than other clusters, even if Density is comparable to ORANGE.*
- *BLUE has the highest Population and Density*
- *COVID data (Cases and Deaths) seems to be of exactly the same distribution as Population, not so much by density; but that could be because we distributed County COVID data as a proportion of Population.*
- *ORANGE has the most expensive(Home-prices) neighbourhoods.*
- *In the thick of the city BLUE is more expensive(Home-prices) than GREEN. Isn't that surprising, why is BLUE more expensive ? It has a higher population density. What else ?*

8. Discussion

This section will build upon "Analysis" listed in above sections. Here is our opportunity to revisit use-cases we created in section 4.1. We will later talk about limitations and improvement opportunities to the above model simulation.

8.1 Use-Case1: Young Families

Families looking for residential neighbourhoods, with schools, shopping, and other such family friendly services. This middle-income demographic typically wouldn't hesitate to drive.

- *Overall DC is a great area to raise a family, with good distribution on venues like Residence, Schools, Medical Centers, Shopping, and Recreation.*
- *GREEN or BLUE seem like good places if you prefer a City life, and close to workplace (Professional&Others).*
- *One could evaluate RED or ORANGE if they prefer to drive around.*
- *Since the whole of Washington DC is just one county, deriving COVID data by proportional distribution across the ZIPs by population, leaves us with little scope for much more analysis. Hence, we will have to ignore the COVID data for DC. COVID analysis could however be significant for other larger states like CA or NY.*
- *Radius to explore venues was assumed as 2km. Population and venue count in a 2km radius is not a fair comparison between the thick of a city and suburb. 2km in a city could be comparable to say 5km of a certain suburb. Hence, RED or ORANGE may not be as low on venue counts after all, if you are habituated to a lifestyle that involves driving fair distances. We could evolve a method to calculate a normalization factor to radius by location/ZIP to make a more appropriate comparison.*

8.2 Use-Case2: Young Individuals or Couples

Young individuals or couples who prefer to have a more outward lifestyle with nightlife, bars, good food around them; have a fun life. Some of them might prefer to live closer to their areas of work.

- *BLUE and GREEN seem like areas that are just right for this group. These areas are in the thick of a city buzzing with nightlife, food/bars, and close to workplaces.*
- *If some of them prefer to stay closer to university, may need to explore ORANGE. For most of them cost may also be a key factor.*
- *Having home rental data could be very useful. This group would consist mostly of renters and not so much of home buyers.*

8.3 Use-Case3: Elderly-Retired Individuals or Couples

Older folks or empty nesters, who value a quieter lifestyle with easy of access to medicare and other life services.

- *GREEN or ORANGE may be the right areas for this demographic.*
- *GREEN seems to have a good distribution of venues such as Medical Centers, Spiritual Centers, Shopping, and others.*
- *ORANGE Could be another choice based on their mobility and affordability, if they do not prefer the city.*

- *GREEN has a lower population compared to BLUE. We could expect it to be less susceptible to COVID like diseases, in comparison to BLUE.*

8.4 Use-Case4: Govt and Agencies

A bird's-eye view for policy makers and agencies/ businesses, so that they can identify improvement potential by areas to attract any of the above groups.

The dataset for DC does not have significant patterns to analyze improvement areas. Datasets from other larger states can be explored to identify such insights. Some of the aspects that are not quite obvious from the given dataset are:

- *Why are Home-prices in GREEN and BLUE so low when they are so well supported by venues ? Do they have crime and other issues ? Is there too much traffic ?*

8.5 Limitations

This study has many limitations. It primarily goes with a pre-meditated assumption of using Foursquare API. A more detailed research is required to define liveability, before we build an analytical model that simulates exploration process.

- *This Analysis is significantly overpowered by availability of data from Foursquare. As you can see from Annexure 10.1, FL cannot have fewer venues than smaller states like WA. OR, Can it ? Most likely Foursquare is not suitable for a state-wide analysis, more suitable for analyzing large cities ?*
- *Some of the key aspects that most people consider while exploring a place to live include quality of schools, Crime Statistics, New or Old infrastructure, etc, which are not part of this study.*
- *Many a time, ratings on specific services of interest – elder homes, hospice, hospitals, daycares, speciality hospitals (like say a specific cancer or Alzheimer care), etc. Such ratings have not been considered in this study.*
- *Some of the analysis showcased above can be more insightful with a larger dataset like CA or FL state.*
- *A more detailed healthcare/disease data could be useful, COVID alone may not be a key decision influencer.*
- *Everybody doesn't buy a home. Having rental data can be very useful. But initial preliminary analysis with Zillow ZORI data was not very encouraging. The dataset was missing data points for most ZIP codes. May require more research.*
- *Other limitations:*
    - *Should non-venue data also be part of clustering algorithm ?*
    - *Did we use the right technique to Normalize/Standardize the data ?*
    - *What is the impact of data loss from CSV files to COMMON data frame ?*
    - *Did we use the right clustering algorithm ?*
    - *Should we have built a more detailed venue exploration model ?*
    - *Should we have included detailed demographic data ?*
    - *Would multi-level clustering be more appropriate for this kind of analysis ? Clusters in a Cluster ?*

9. Conclusion

We can use Data Sciences to empower our decisions, as demonstrated above. However, to accomplish our objectives adequate research on the subject matter and a well-defined methodology/roadmap like "idY" are critical. As a part of IBM-Coursera Capstone Project requirement (for Applied Data Sciences Course), I have tried to present a simple model with limitations. This model can be replicated as a template to execute upon small Data Sciences projects. Simply copy the Jupyter Notebook, Change State to, say "IL", execute and there you go. Add additional data references. Add better visualizations. Take the sections as guide and fill your own code. Do whatever.

Annexure 10.1 presents the model on additional states – CA, WA, FL, NY and NJ. Now that you have come this far,

- How do you compare the 5 states ? Rural Vs Urban ? COVID in NY, NJ Vs other states ?
- Where would you choose to live ?
- Did I miss a state in MidWest or South ? Why don't you play with the Notebook template - do it yourself ?

I am grateful to IBM, Coursera, and all those references available online to make my task so effortless. I am no Data Scientist or Programmer (yet), but they made it easy for me. ***Thank You All***

10. Annexures

10.1 Analyze Additional States

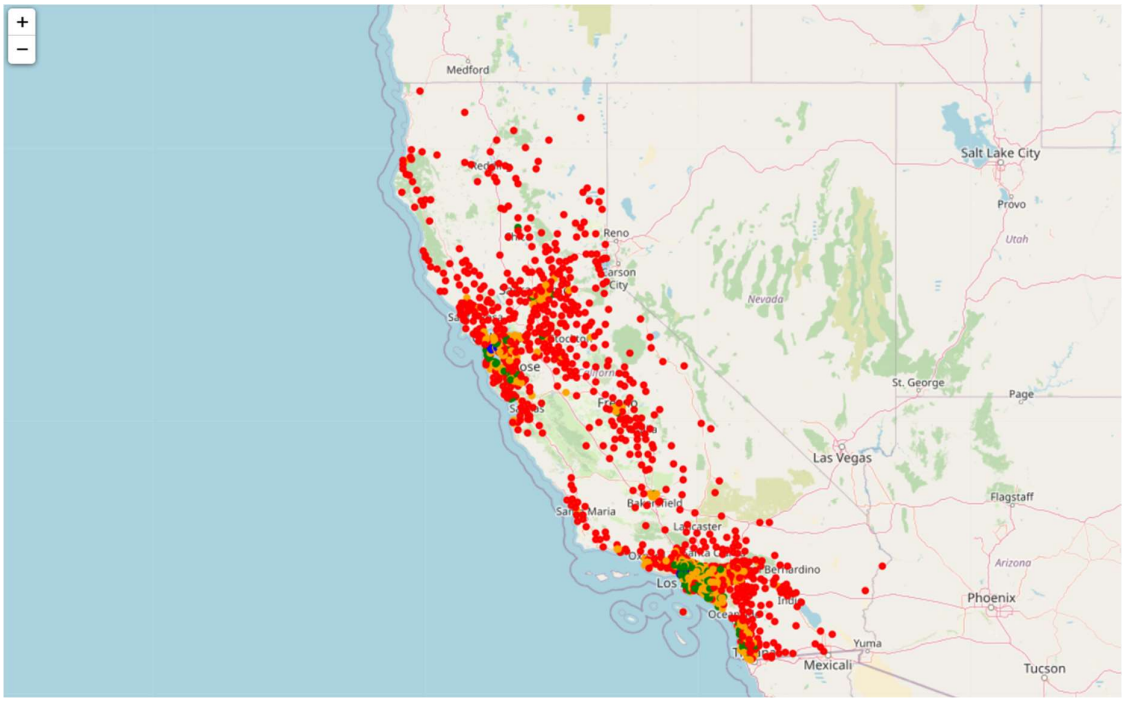
States Analyzed in this section:

State	Ref: Jupyter Notebook
California(CA)	Coursera-IBM CapStone Liveable Neighbourhoods CA v17.ipynb
Washington(WA)	Coursera-IBM CapStone Liveable Neighbourhoods WA v17.ipynb
Florida(FL)	Coursera-IBM CapStone Liveable Neighbourhoods FL v17.ipynb
New York(NY)	Coursera-IBM CapStone Liveable Neighbourhoods NY v17.ipynb
New Jersey(NJ)	Coursera-IBM CapStone Liveable Neighbourhoods NJ v17.ipynb

10.1.1 Analyze – California (CA)

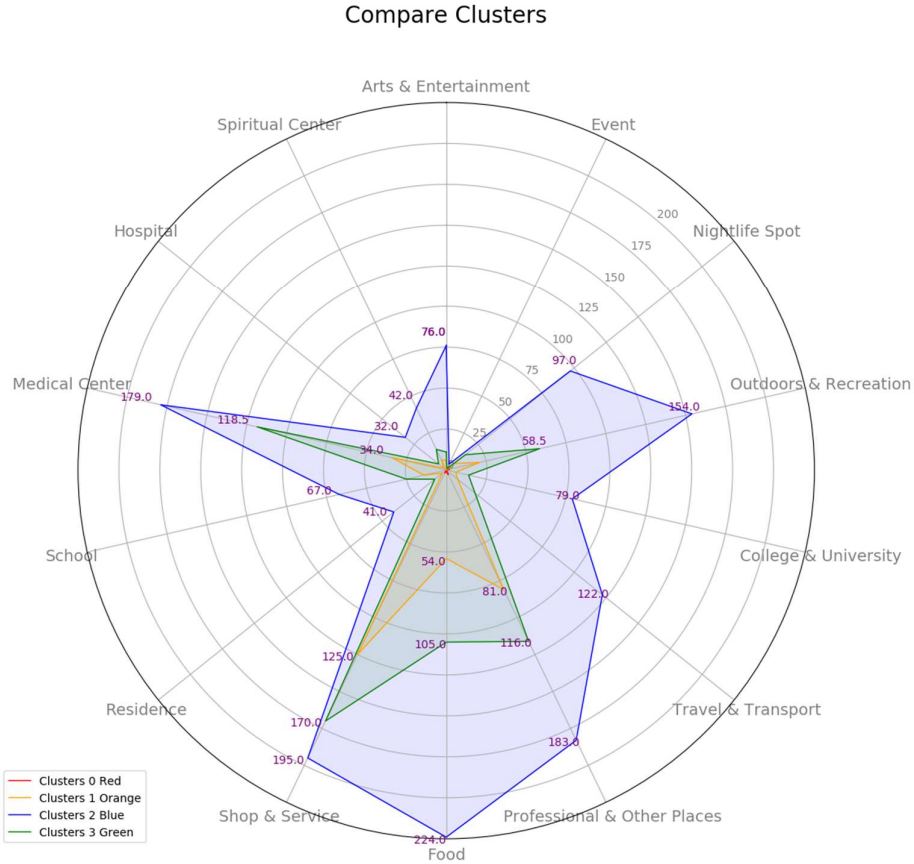
10.1.1.1 Visualize Clusters – Geographic

Plot Clusters on Geo-Map

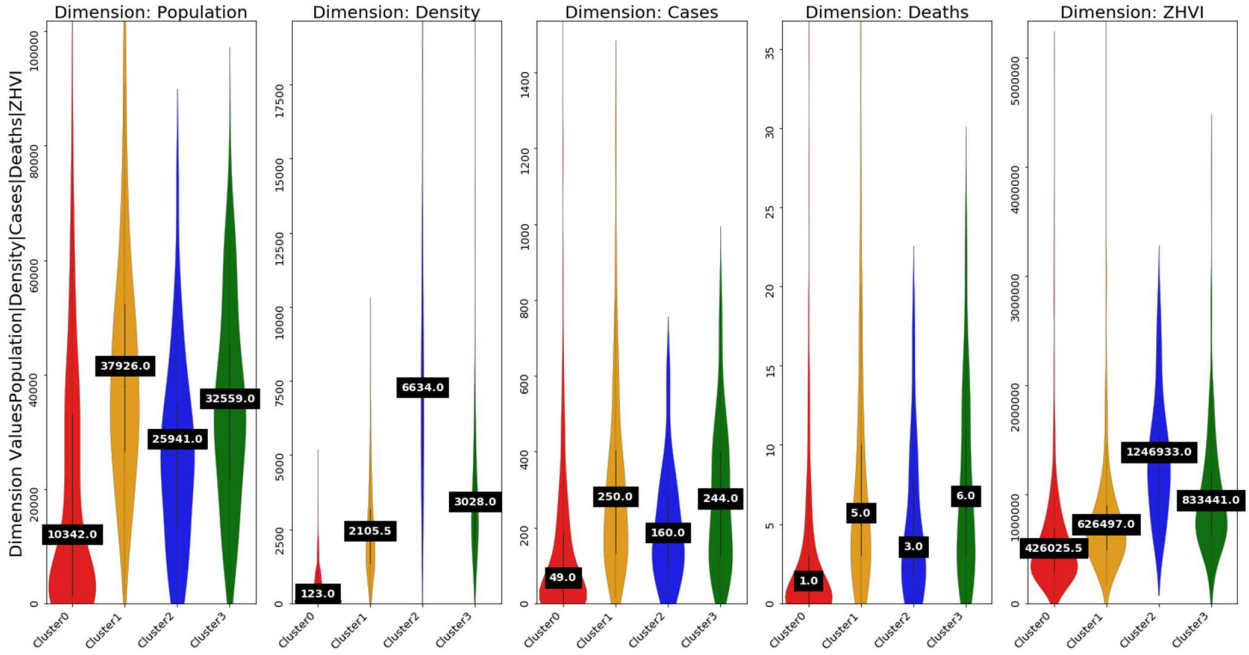


10.1.1.2 Compare Clusters by Venues

Spider-Plot to Compare Clusters by Venue Medians



Violin-plots to Compare non-Venue Feature Distribution Across Clusters



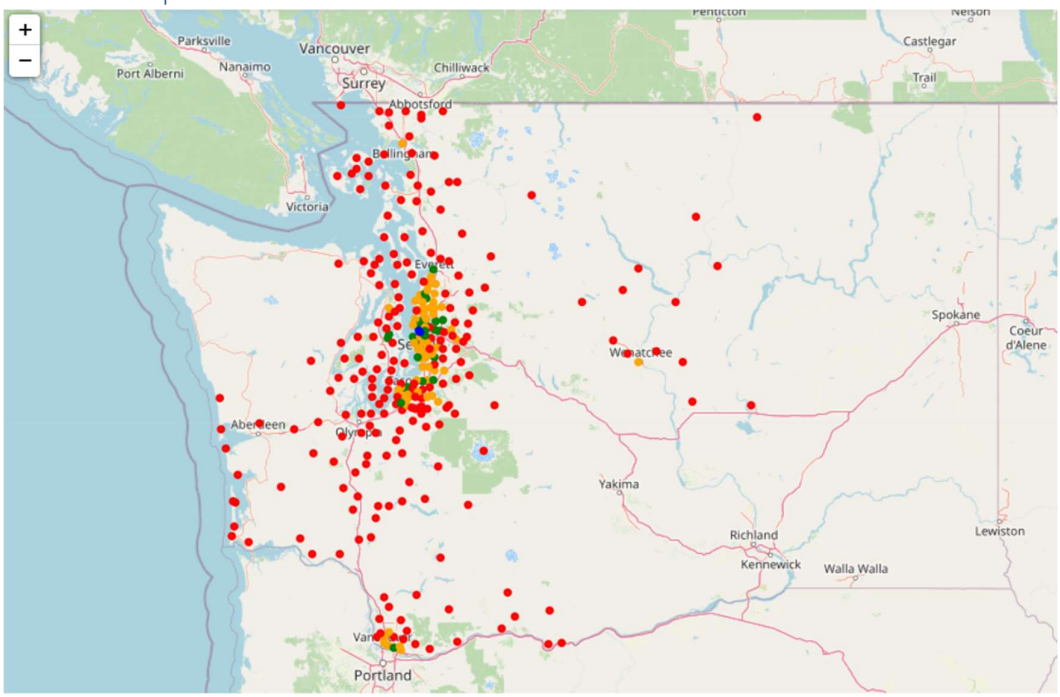
10.1.1.3 Discussion

Use Case		What do you think ?
1. Families looking for residential neighbourhoods, with schools, shopping, and other such family friendly services. This middle-income demographic typically wouldn't hesitate to drive.		
2. Young individuals or couples who prefer to have a more outward lifestyle with nightlife, bars, good food around them; have a fun life. Some of them might prefer to live closer to their areas of work.		
3. Older folks or empty nesters, who value a quitter lifestyle with easy of access to medicare and other life services.		
4. A bird's-eye view for policy makers and agencies/ businesses, so that they can identify improvement potential by areas to attract any of the above groups.		

10.1.2 Analyze – Washington (WA)

10.1.2.1 Visualize Clusters – Geographic

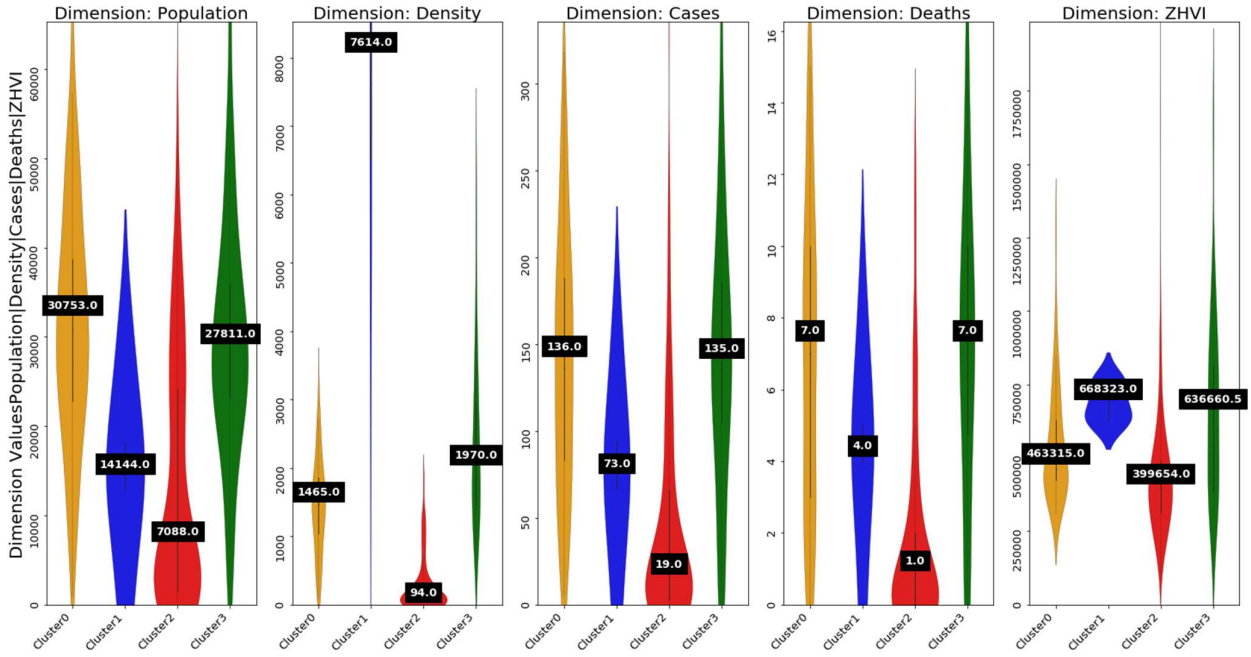
Plot Clusters on Geo-Map



10.1.2.2 Compare Clusters by Venues

Spider-Plot to Compare Clusters by Venue Medians

Violin-plots to Compare non-Venue Feature Distribution Across Clusters



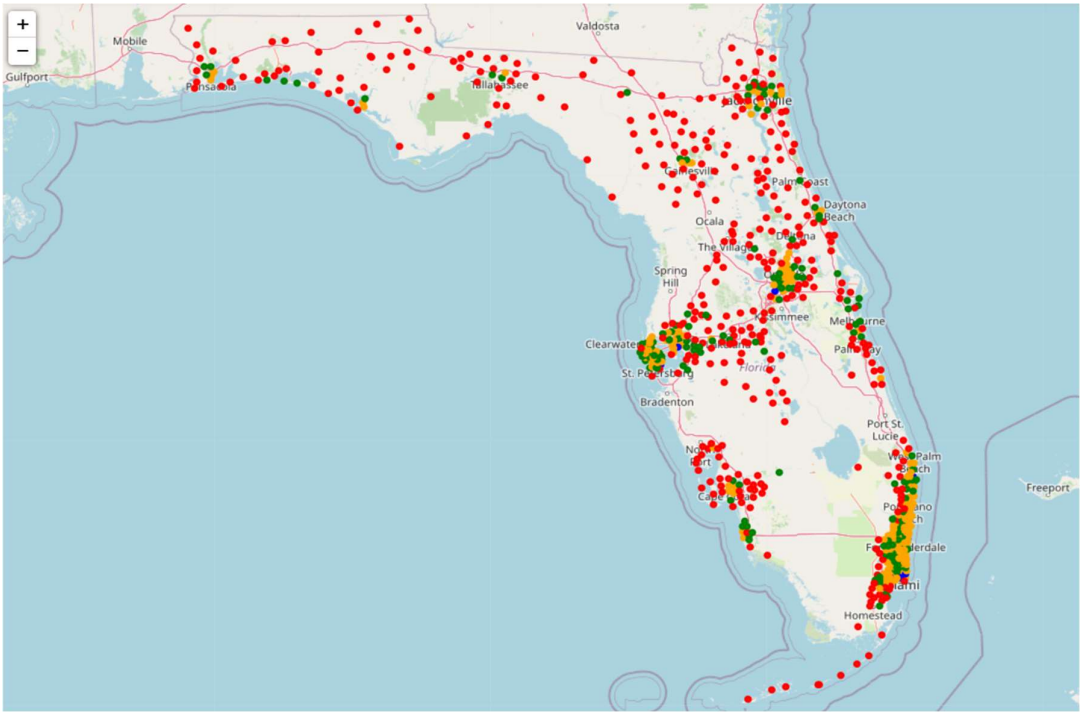
10.1.2.3 Discussion

Use Case		What do you think ?
5. Families looking for residential neighbourhoods, with schools, shopping, and other such family friendly services. This middle-income demographic typically wouldn't hesitate to drive.		
6. Young individuals or couples who prefer to have a more outward lifestyle with nightlife, bars, good food around them; have a fun life. Some of them might prefer to live closer to their areas of work.		
7. Older folks or empty nesters, who value a quitter lifestyle with easy of access to medicare and other life services.		
8. A bird's-eye view for policy makers and agencies/ businesses, so that they can identify improvement potential by areas to attract any of the above groups.		

10.1.3 Analyze – Florida (FL)

10.1.3.1 Visualize Clusters – Geographic

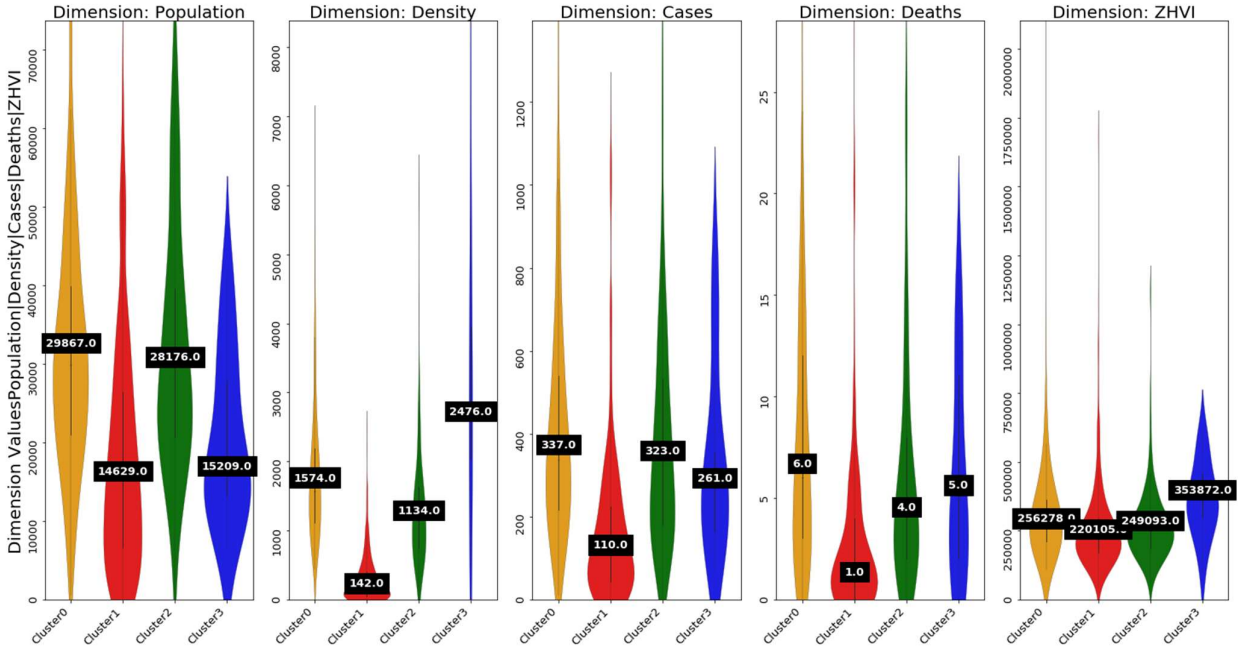
Plot Clusters on Geo-Map



10.1.3.2 Compare Clusters by Venues

Spider-Plot to Compare Clusters by Venue Medians

Violin-plots to Compare non-Venue Feature Distribution Across Clusters



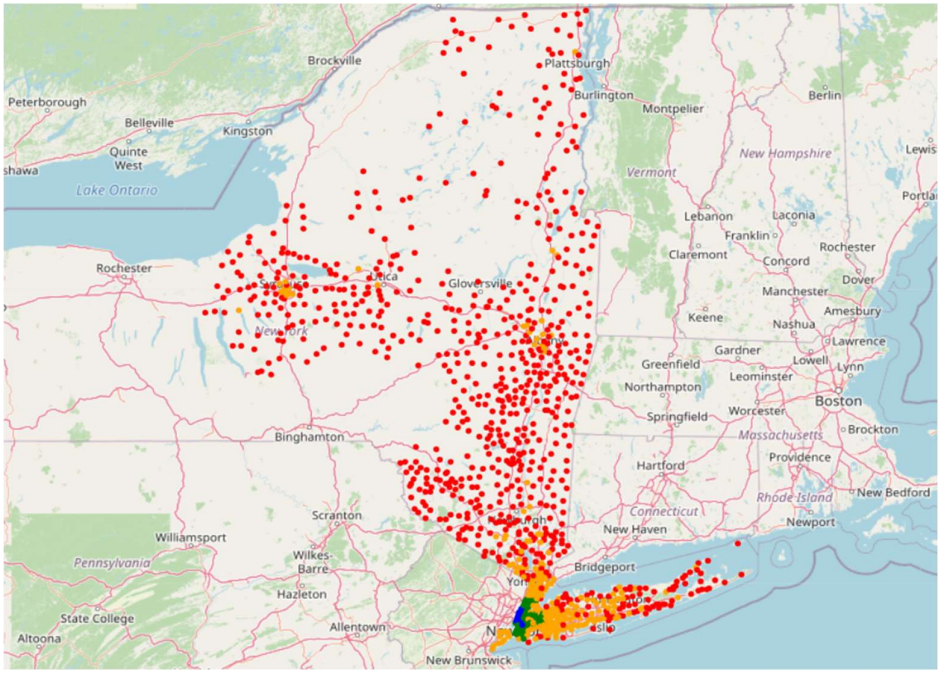
10.1.3.3 Discussion

Use Case		What do you think ?
9. Families looking for residential neighbourhoods, with schools, shopping, and other such family friendly services. This middle-income demographic typically wouldn't hesitate to drive.		
10. Young individuals or couples who prefer to have a more outward lifestyle with nightlife, bars, good food around them; have a fun life. Some of them might prefer to live closer to their areas of work.		
11. Older folks or empty nesters, who value a quitter lifestyle with easy of access to medicare and other life services.		
12. A bird's-eye view for policy makers and agencies/ businesses, so that they can identify improvement potential by areas to attract any of the above groups.		

10.1.4 Analyze – New York (NY)

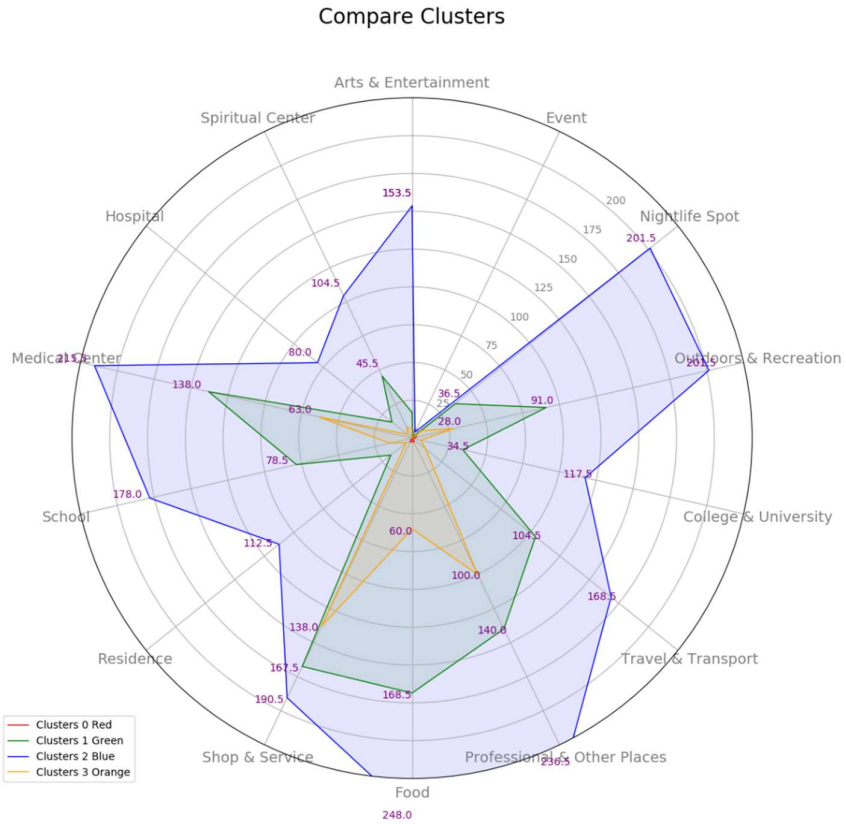
10.1.4.1 Visualize Clusters – Geographic

Plot Clusters on Geo-Map

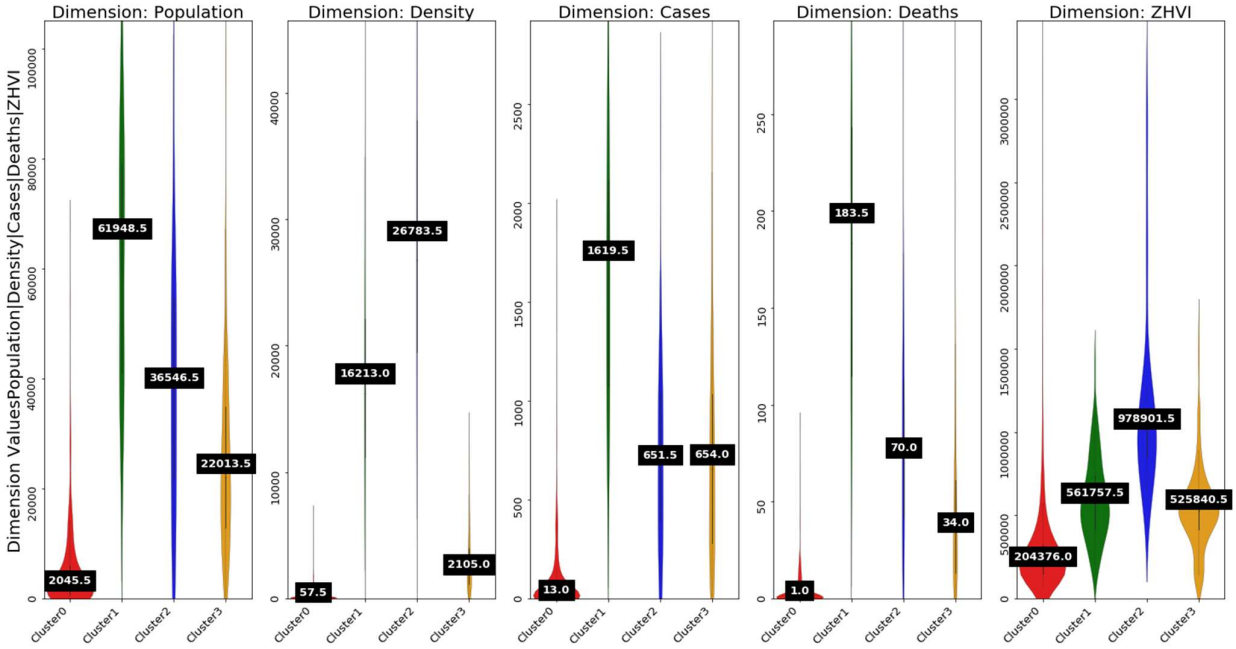


10.1.4.2 Compare Clusters by Venues

Spider-Plot to Compare Clusters by Venue Medians



Violin-plots to Compare non-Venue Feature Distribution Across Clusters



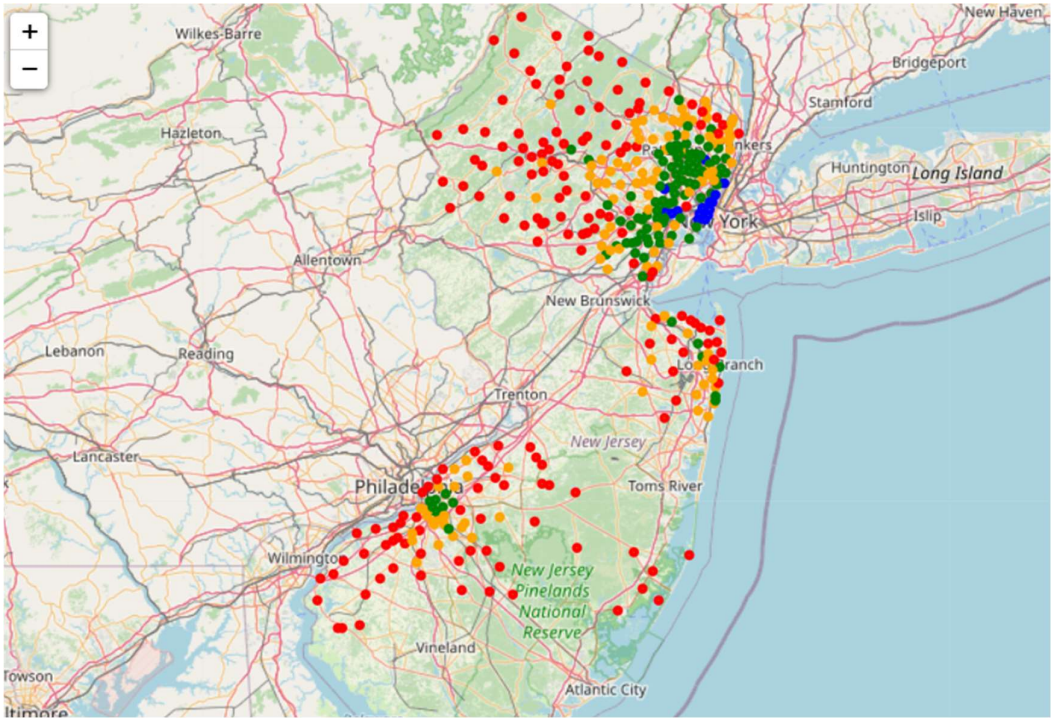
10.1.4.3 Discussion

Use Case		What do you think ?
13. Families looking for residential neighbourhoods, with schools, shopping, and other such family friendly services. This middle-income demographic typically wouldn't hesitate to drive.		
14. Young individuals or couples who prefer to have a more outward lifestyle with nightlife, bars, good food around them; have a fun life. Some of them might prefer to live closer to their areas of work.		
15. Older folks or empty nesters, who value a quitter lifestyle with easy of access to medicare and other life services.		
16. A bird's-eye view for policy makers and agencies/ businesses, so that they can identify improvement potential by areas to attract any of the above groups.		

10.1.5 Analyze – New Jersey (NJ)

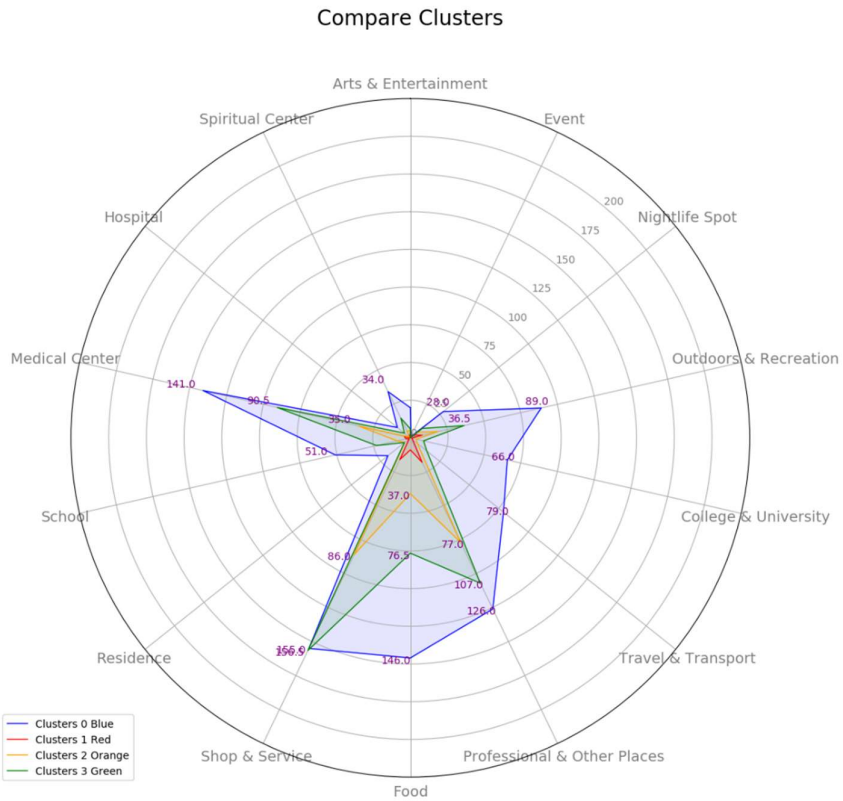
10.1.5.1 Visualize Clusters – Geographic

Plot Clusters on Geo-Map

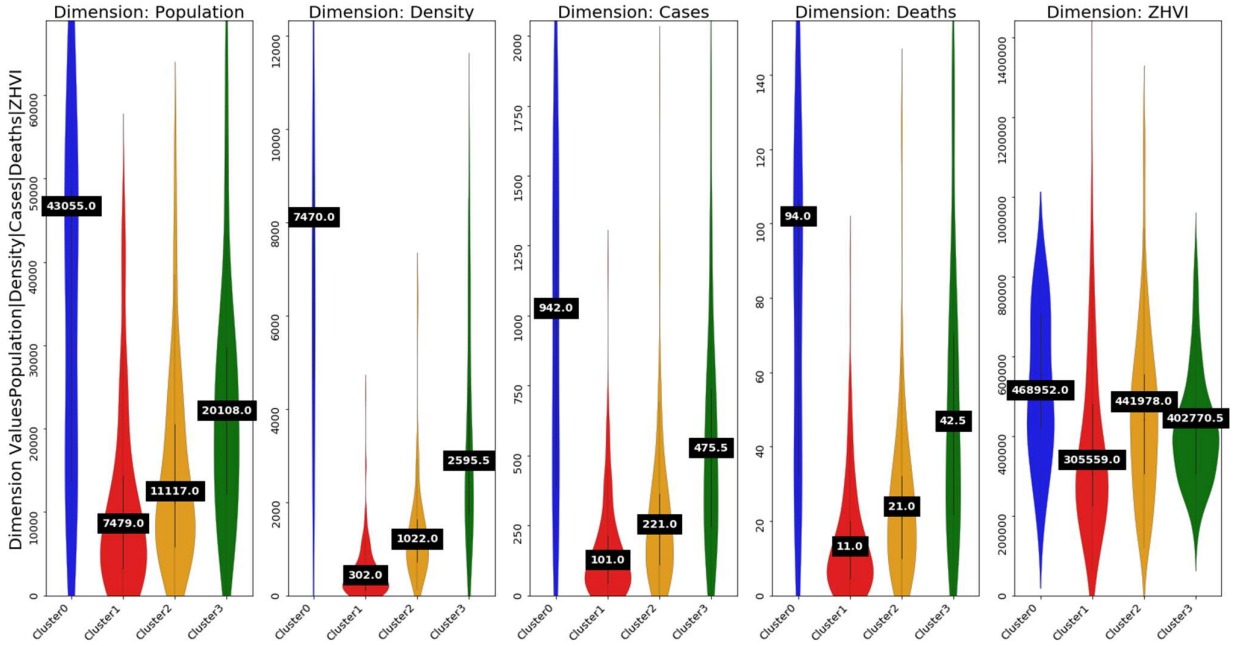


10.1.5.2 Compare Clusters by Venues

Spider-Plot to Compare Clusters by Venue Medians



Violin-plots to Compare non-Venue Feature Distribution Across Clusters



10.1.5.3 Discussion

Use Case		What do you think ?
17. Families looking for residential neighbourhoods, with schools, shopping, and other such family friendly services. This middle-income demographic typically wouldn't hesitate to drive.		
18. Young individuals or couples who prefer to have a more outward lifestyle with nightlife, bars, good food around them; have a fun life. Some of them might prefer to live closer to their areas of work.		
19. Older folks or empty nesters, who value a quitter lifestyle with easy of access to medicare and other life services.		
20. A bird's-eye view for policy makers and agencies/ businesses, so that they can identify improvement potential by areas to attract any of the above groups.		

10.2 Key References

Reference	Link
Coursera	https://Coursera.org
IBM Skills Network	https://labs.cognitiveclass.ai/login?next=https%3A%2F%2Flabs.cognitiveclass.ai%2F
IBM Data&Analytics Hub	https://www.ibmbigdatahub.com/blog/why-we-need-methodology-data-science
Python Graph Gallery	https://python-graph-gallery.com/
Learn By Marketing	http://www.learnbymarketing.com/tutorials/k-means-clustering-by-hand-excel/
Analytic Vidhya	https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/
Geek for Geeks	https://www.geeksforgeeks.org/
StackOveflow	https://stackoverflow.com/
Capstone Project Stanislav Rogozhin	https://towardsdatascience.com/classification-of-moscow-metro-stations-using-Foursquare-data-fb8aad3e0e4