# A. Introduction

## A.1. Business Problem

Florida is the southeasternmost U.S. state with population of around 21.48 Million, with the Atlantic on one side and the Gulf of Mexico on the other. It has hundreds of miles of beaches. The city of Miami is known for its Latin-American cultural influences and notable arts scene, as well as its nightlife, especially in upscale South Beach. Orlando is famed for theme parks, including Walt Disney World.

Florida is a very rich place with its multi culture, tourist destination and lots of beaches and theme park. People across the USA are very keen to invest or get settle down here due to its warm weather and rich culture and other attractions.

This project is to find out a best county to invest in real estate in Florida USA. The investment might be buying a house for staying or renting or could be an investment in some local business. The idea of this project to give insight of the neighborhood so that it is easy to narrow down the right place suited to the stakeholders need.

In order get the details we will be looking for information like population of a county of people from different ages, and race. We will also check crime rate in different county to find out a comparatively safe place for investment. We will check house price and its trend and what are different venues available to decide what type of business might be suitable in a county.

This project is not intended to narrow down to find a place for a specific investment in real estate but to provide different parameters or opinion to decide which place might be suitable for what type of investment.

We will use Data Science to collect all the details and consolidate and present it such a way so that it can give us an idea about the neighborhood and help us to take decision on a place and investment. Based on the final conclusion we can narrow down it further to get more details insight but that is not in the scope of this project.

## A.2. Data Description

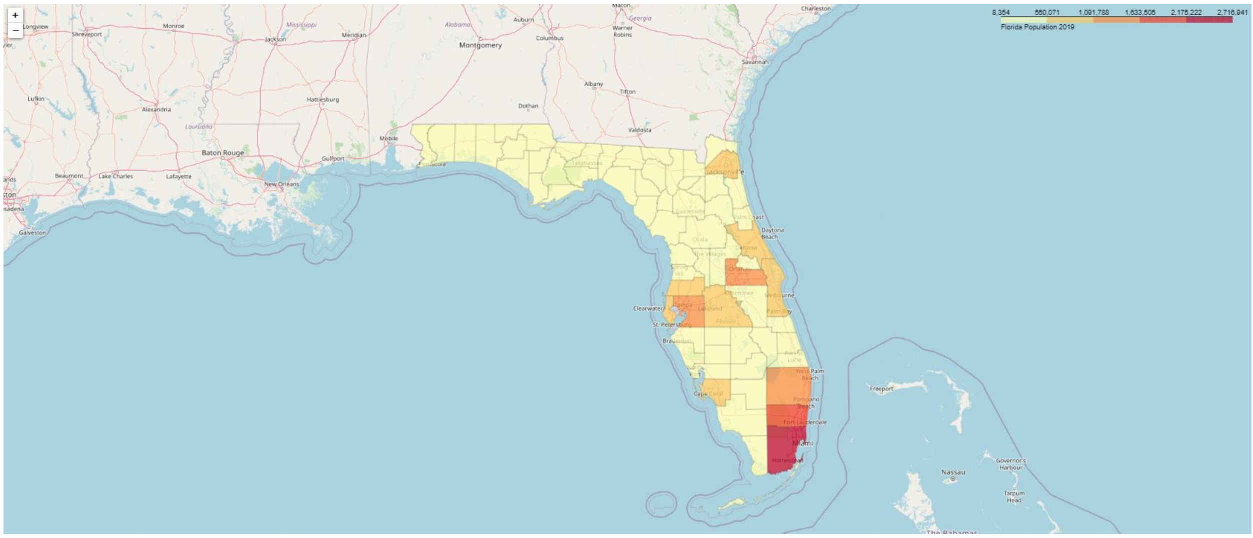
To get a solution of the stated business problem we would need the data as given below.

- Census data of Florida USA with the breakdown into age, sex and race. I found this data census.gov site. The data from census has more granular level details then we need, so I have removed the unnecessary details and cleaned the dataset to use for my purpose.
- I have also used crime data from Florida Law Enforcement website to get an understanding of what type of crime are more common or frequent by County and how we can relate it with other details.
- From the Federal Housing Financial Agency website I have taken data related housing price and index and its growth over the past years which I have used to find out which county might be more preferable to invest in housing or real estate with respect to cost.
- To put all the analysis into a more understandable map I have downloaded the co-ordinate details of Florida county from census website.
- I have used Google Map api to get the data related to venues in each county.

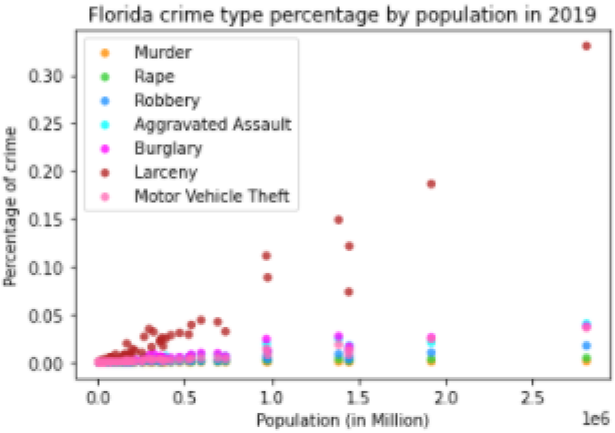
# B. Methodology

Most of the data used for this project are being used directly from the website, some data I have downloaded from website and saved in Azure Blob public container and accessed the data directly in my notebook.

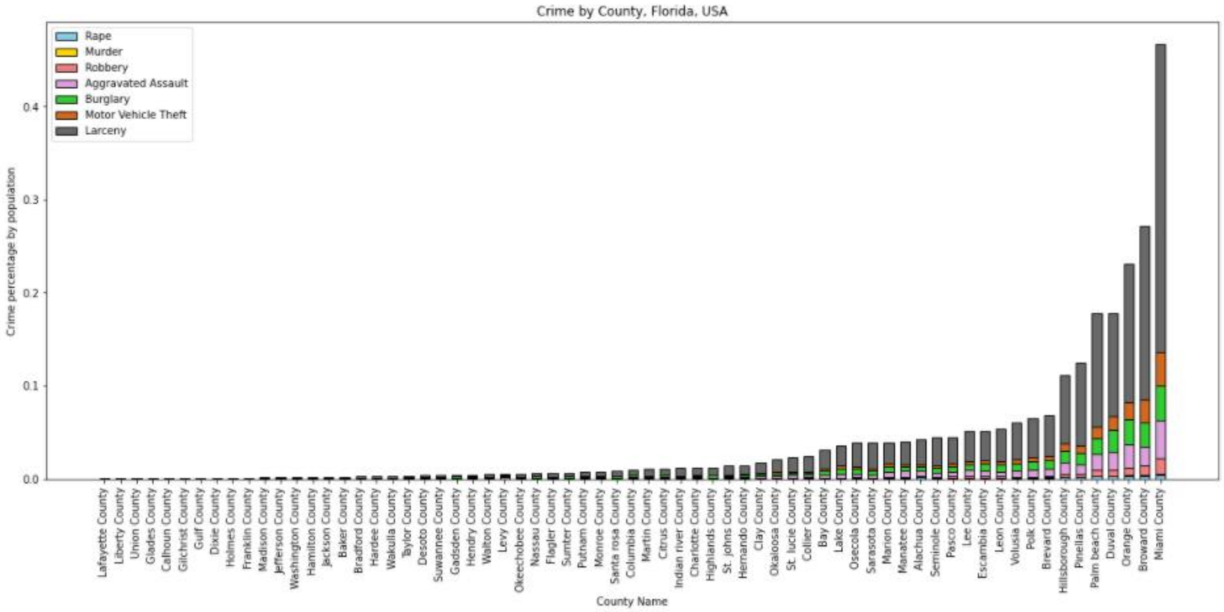
I have used python Folium library to visualize the data in a map. Below map showing population density in each Florida County.



I have analyzed the crime data and tried to get a relation with population and as shown in the scatter plot below though other crime majorly robbery, assault, burglary and vehicle theft increased a bit with population, but Larceny has increased exponentially with the population.



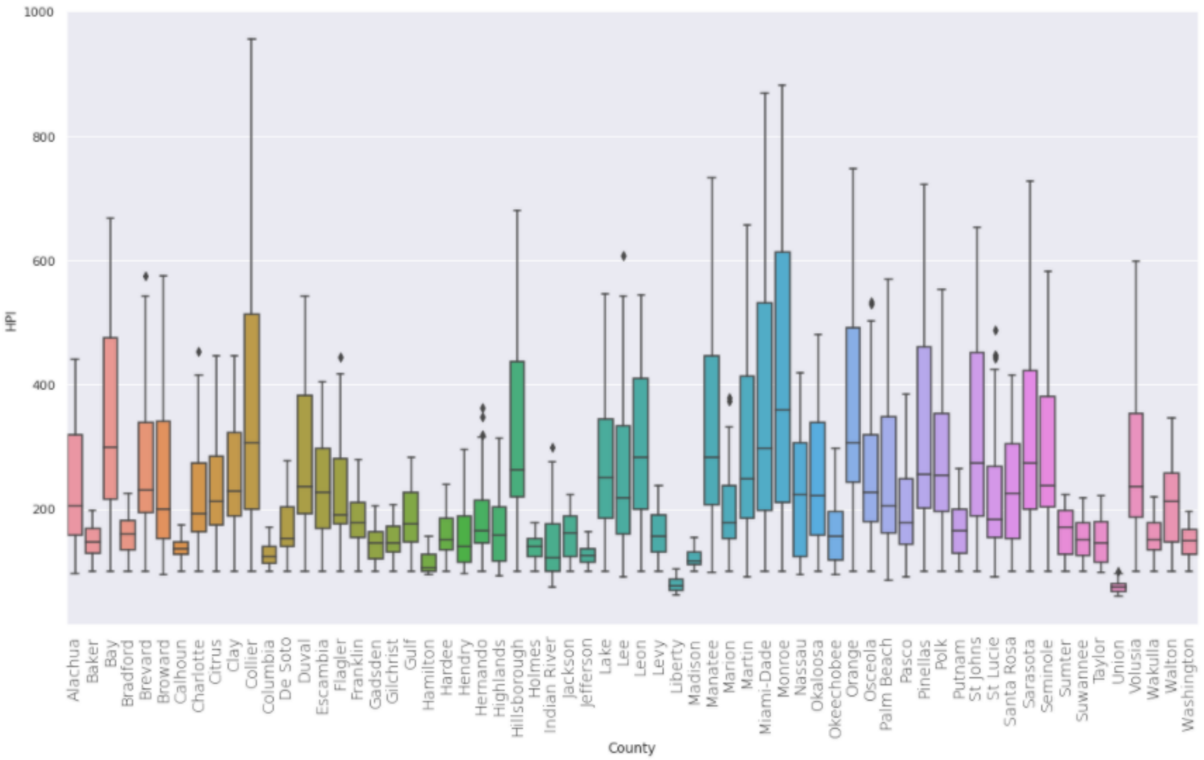
Now let me put it in a bar chart to get a better visual of the crime data by county.



## Housing price Index

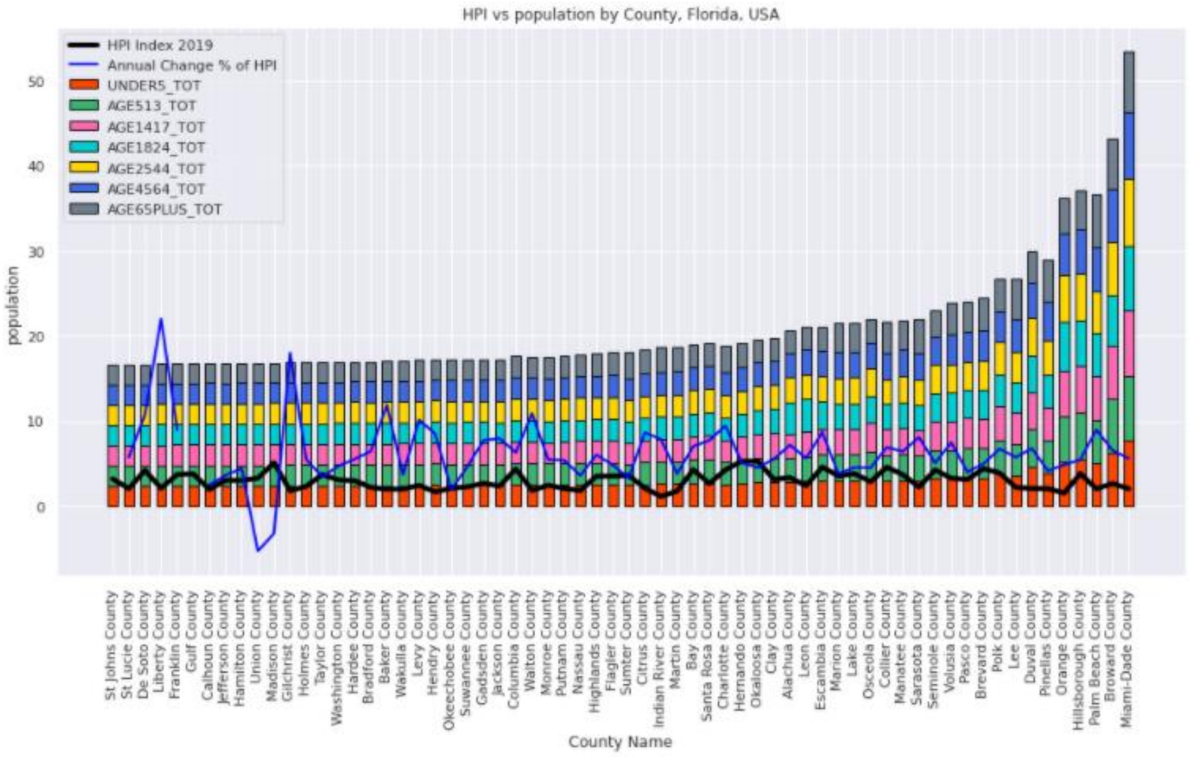
Now let us have a look at the housing price index by County. Let us first have a look how the HPI has changed by County over the last 10 years. To get a better visual we will put it in a boxplot.

As we can see below Union and Liberty county has the lowest HPI and along with Madison, Jefferson and Calhoun counties house market flourished much in the past few years where as Monroe, Miami-Dade, Clay and Bay county has good increase in HPI over the past few years means the house market boomed in past several years in these counties and we can safely assume that it will continue booming in coming years.

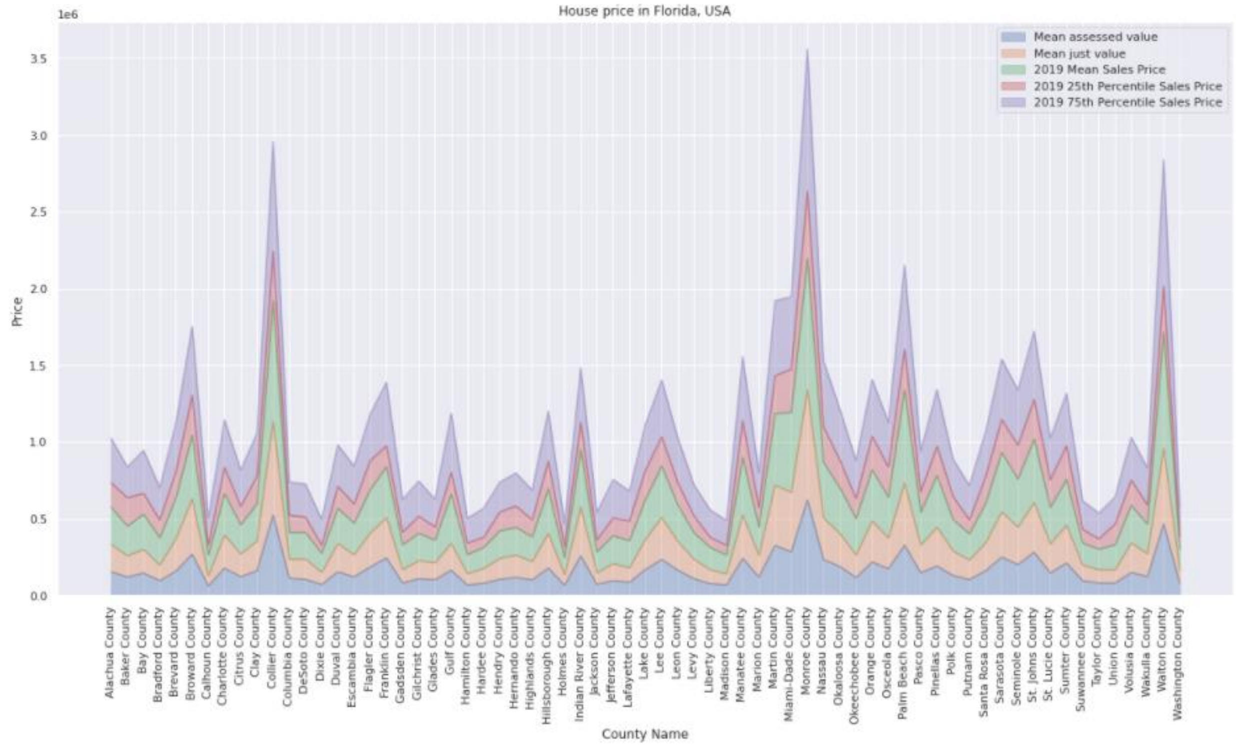


Next, I have checked HPI for the year 2019 only and put it in a plot along with population by age. In the plot the blue line represents the HPI change percentage from the year 2018 to 2019. We can see Liberty and Gilchrist County has a good increase of HPI from 2018 to 2019 where as HPI has dropped in Union and Madison County.

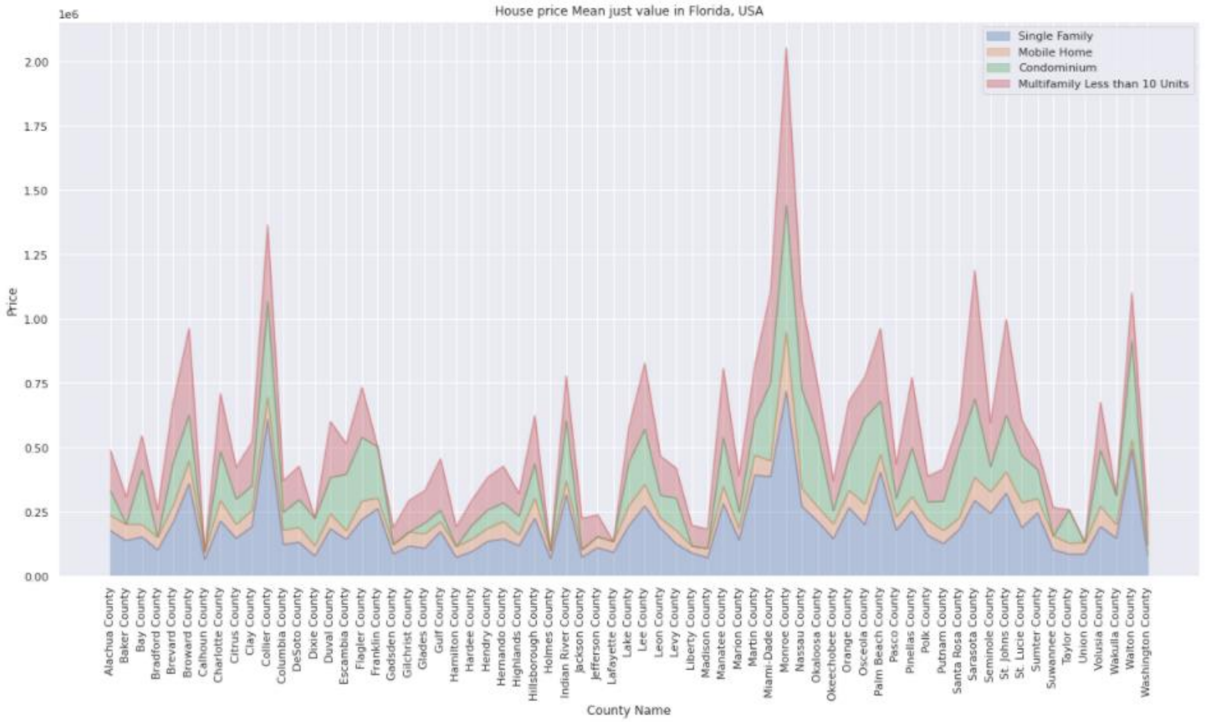
When I put the different age group also in the same chart, it looks like different age group people has almost equal percentage by population, so age group details may not help to draw any conclusion in this analysis. So, I have excluded these details later part from my analysis.



Next, I have checked the House price in Florida, and put it in an area chart for a better visualization. We can see that Collier, Monroe and Walton County has highest average house price and Calhoun, Holmes, Hamilton and Madison County has lowest average house price.



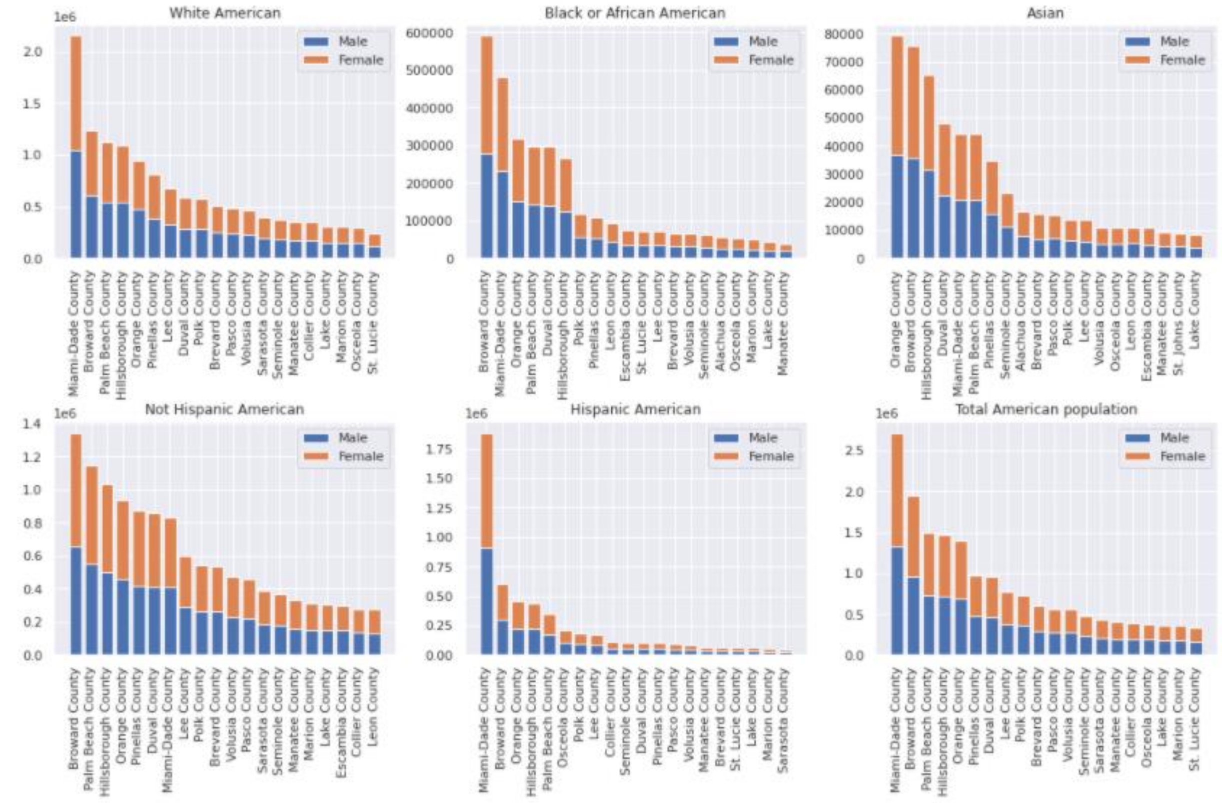
I have also looked at the price of different house type in Florida counties, and below is an area chart to visualize the numbers and we can see that it is almost aligned to the mean house price. One observation I have here is that Calhoun and Holmes county doesn’t have huge difference in house price between different house type whereas other counties has significant difference and some counties like Collier, Monroe and Walton County where house price is high has notable difference in different house type.



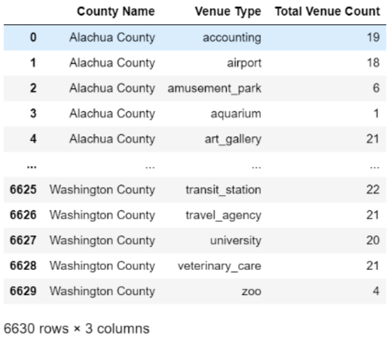
For investment I think a major parameter to consider is the race. For opening a business, we should consider type of the customers their culture, likes and dislikes and considering these in business decision will help for better growth as we can make it customer centric, for example if someone have a plan to open a Restaurant they can decide on cuisine based on the type of the people in that area, or someone would like to buy house they may want to consider a locality where most people belongs to similar community as him/her.

So, I have categorized the Counties based on the population of people from different race. In Florida there are few major races as given below, and I will see top 20 counties in Florida in each category.

- White American
- Black or African American
- Asian
- Hispanic
- Not Hispanic

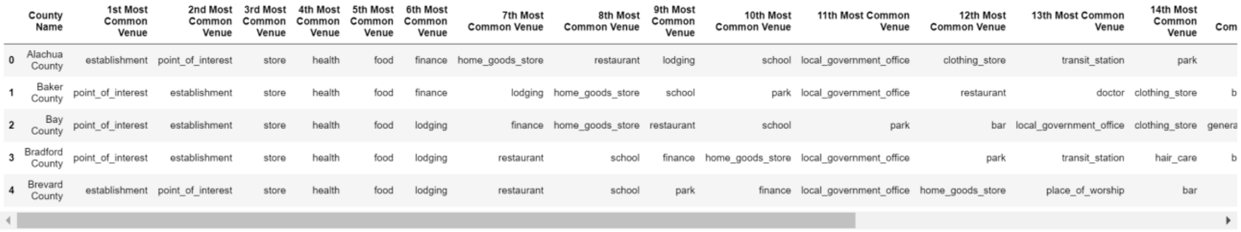


I have also collected the venue details and I have used Google nearby paces api to gather this information. I have used googlemaps python package to connect to google api and finally put all venue details in panda dataframe. There are 96 venue categories defined in google nearby places api and I have collected details for all the category type. Below is the snapshot of the data after cleaning. Here as you can see, I put total number of venue by each category and county. From this dataframe I have collected 20 most popular venue in each county.



| | County Name | Venue Type | Total Venue Count |
|---|---|---|---|
| 0 | Alachua County | accounting | 19 |
| 1 | Alachua County | airport | 18 |
| 2 | Alachua County | amusement_park | 6 |
| 3 | Alachua County | aquarium | 1 |
| 4 | Alachua County | art_gallery | 21 |
| ... | ... | ... | ... |
| 6625 | Washington County | transit_station | 22 |
| 6626 | Washington County | travel_agency | 21 |
| 6627 | Washington County | university | 20 |
| 6628 | Washington County | veterinary_care | 21 |
| 6629 | Washington County | zoo | 4 |

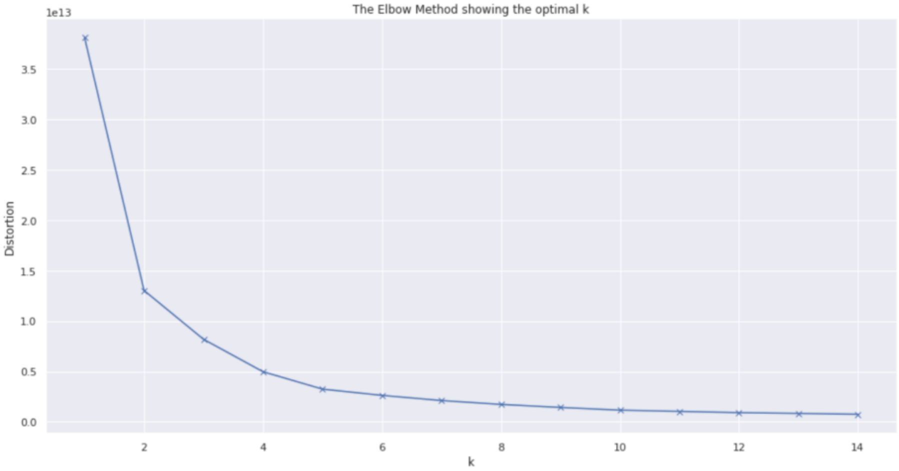
6630 rows × 3 columns



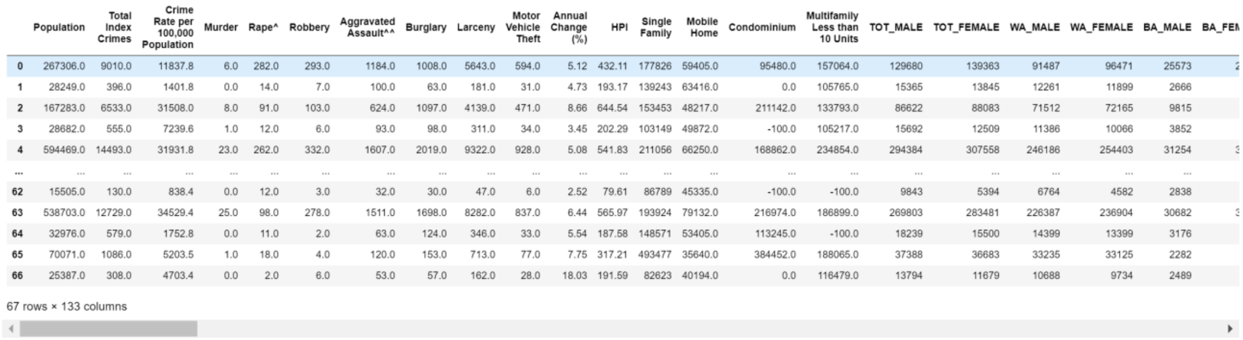
| | County Name | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue | 11th Most Common Venue | 12th Most Common Venue | 13th Most Common Venue | 14th Most Common Venue | Com |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Alachua County | establishment | point_of_interest | store | health | food | finance | home_goods_store | restaurant | lodging | school | local_government_office | clothing_store | transit_station | park | |
| 1 | Baker County | point_of_interest | establishment | store | health | food | finance | lodging | home_goods_store | school | park | local_government_office | restaurant | doctor | clothing_store | b |
| 2 | Bay County | point_of_interest | establishment | store | health | food | lodging | finance | home_goods_store | restaurant | school | park | bar | local_government_office | clothing_store | genera |
| 3 | Bradford County | point_of_interest | establishment | store | health | food | lodging | restaurant | school | finance | home_goods_store | local_government_office | park | transit_station | hair_care | b |
| 4 | Brevard County | establishment | point_of_interest | store | health | food | lodging | restaurant | school | park | finance | local_government_office | home_goods_store | place_of_worship | bar | |

Finally, I have merged all the data I have discussed so far in a single dataframe and used unsupervised learning **K-means algorithm** to cluster the counties. k-means clustering is a popular method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

First, I have used Elbow method to find optimum number of clusters and as shown in the graph below 5 degree would be the optimum k for this dataset to run K-means algorithm.



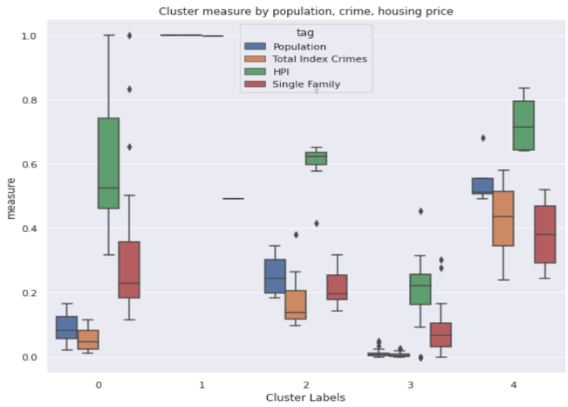
Below is the snapshot of my merged dataset.

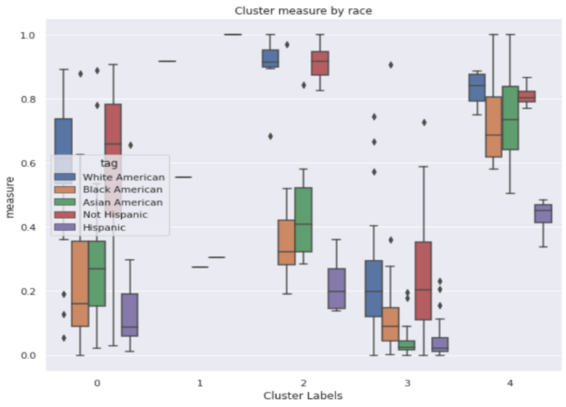


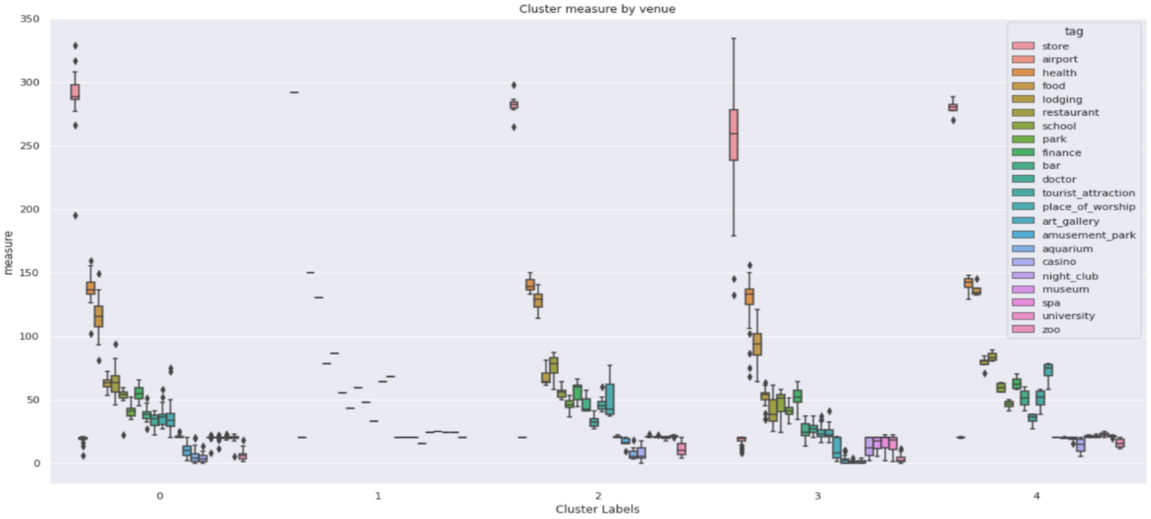
| | Population | Total Index Crimes | Crime Rate per 100,000 Population | Murder | Rape^ | Robbery | Aggravated Assault^^ | Burglary | Larceny | Motor Vehicle Theft | Annual Change (%) | HPI | Single Family | Mobile Home | Condominium | Multifamily Less than 10 Units | TOT_MALE | TOT_FEMALE | WA_MALE | WA_FEMALE | BA_MALE | BA_FEM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 267306.0 | 9010.0 | 11837.8 | 6.0 | 282.0 | 293.0 | 1184.0 | 1008.0 | 5643.0 | 594.0 | 5.12 | 432.11 | 177826 | 59405.0 | 95480.0 | 157064.0 | 129680 | 139363 | 91487 | 96471 | 25573 | 2 |
| **1** | 28249.0 | 396.0 | 1401.8 | 0.0 | 14.0 | 7.0 | 100.0 | 63.0 | 181.0 | 31.0 | 4.73 | 193.17 | 139243 | 63416.0 | 0.0 | 105765.0 | 15365 | 13845 | 12261 | 11899 | 2666 | |
| **2** | 167283.0 | 6533.0 | 31508.0 | 8.0 | 91.0 | 103.0 | 624.0 | 1097.0 | 4139.0 | 471.0 | 8.66 | 644.54 | 153453 | 48217.0 | 211142.0 | 133793.0 | 86622 | 88083 | 71512 | 72165 | 9815 | |
| **3** | 28682.0 | 555.0 | 7239.6 | 1.0 | 12.0 | 6.0 | 93.0 | 98.0 | 311.0 | 34.0 | 3.45 | 202.29 | 103149 | 49872.0 | -100.0 | 105217.0 | 15692 | 12509 | 11386 | 10066 | 3852 | |
| **4** | 594469.0 | 14493.0 | 31931.8 | 23.0 | 262.0 | 332.0 | 1607.0 | 2019.0 | 9322.0 | 928.0 | 5.08 | 541.83 | 211056 | 66250.0 | 168862.0 | 234854.0 | 294384 | 307558 | 246186 | 254403 | 31254 | 3 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **62** | 15505.0 | 130.0 | 838.4 | 0.0 | 12.0 | 3.0 | 32.0 | 30.0 | 47.0 | 6.0 | 2.52 | 79.61 | 86789 | 45335.0 | -100.0 | -100.0 | 9843 | 5394 | 6764 | 4582 | 2838 | |
| **63** | 538703.0 | 12729.0 | 34529.4 | 25.0 | 98.0 | 278.0 | 1511.0 | 1698.0 | 8282.0 | 837.0 | 6.44 | 565.97 | 193924 | 79132.0 | 216974.0 | 186899.0 | 269803 | 283481 | 226387 | 236904 | 30682 | 3 |
| **64** | 32976.0 | 579.0 | 1752.8 | 0.0 | 11.0 | 2.0 | 63.0 | 124.0 | 346.0 | 33.0 | 5.54 | 187.58 | 148571 | 53405.0 | 113245.0 | -100.0 | 18239 | 15500 | 14399 | 13399 | 3176 | |
| **65** | 70071.0 | 1086.0 | 5203.5 | 1.0 | 18.0 | 4.0 | 120.0 | 153.0 | 713.0 | 77.0 | 7.75 | 317.21 | 493477 | 35640.0 | 384452.0 | 188065.0 | 37388 | 36683 | 33235 | 33125 | 2282 | |
| **66** | 25387.0 | 308.0 | 4703.4 | 0.0 | 2.0 | 6.0 | 53.0 | 57.0 | 162.0 | 28.0 | 18.03 | 191.59 | 82623 | 40194.0 | 0.0 | 116479.0 | 13794 | 11679 | 10688 | 9734 | 2489 | |

67 rows × 133 columns

After I have clustered the dataset I analyzed each cluster and in order to do that I have used boxplot to put the data of each cluster by each category, for venues I have used the most common venues.







After analyzing the above plots we can categorize the clusters as below ..

## B.1. Categorize the clusters

### Cluster 0:

- Average Population
- Below Average Crime rate
- HPI below average
- Average White American Population
- Below Average Black American population
- Below Average Asian American Population
- Above Average Not Hispanic American Population

- Below Average Hispanic American Population
- More Tourist attraction

### Cluster 1:

- Very High in Population
- Highest in Crime
- High HPI and High House Price
- High in White American Population
- Average Black American population
- Below Average Asian American Population
- Below Average Not Hispanic American Population
- Below Average Hispanic American Population
- More number of Universities, Night club, Museum and more amusement park

### Cluster 2:

- Above Average Population
- Average in Crime
- HPI is low
- High in White American Population
- Below Average Black American population
- Below Average Asian American Population
- High Not Hispanic American Population
- Below Average Hispanic American Population

### Cluster 3:

- Low in Population
- Low in Crime
- Average HPI but low House Price

- Below Average in White American Population
- Low Average Black American population
- Low Asian American Population
- Below Average Not Hispanic American Population
- Low Hispanic American Population
- Few Tourist attraction, low in cafe, bar, restaurant, store, casino, amusement park and place of worship

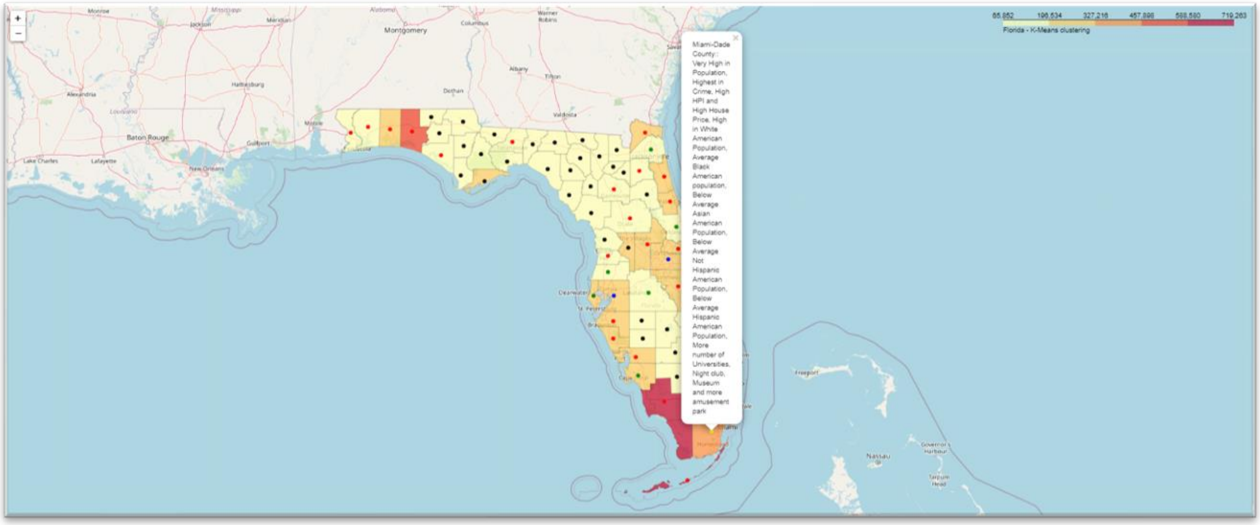
Cluster 4:

- High in Population
- Above average in Crime
- lowest HPI
- Above Average in White American Population
- Above Average Black American population
- Above Average Asian American Population
- Above Average Not Hispanic American Population
- Average Hispanic American Population
- More number of lodging, amusement park and place of worship compare to other clusters

# C. Result

Now let me combine all the data together and create a visualization of the summary. To do that I have downloaded the Florida county boundary json file from www.igismap.com and used that to create folium choropleth path.

In the final section I have put the cluster analysis and average housing price in folium map where mean average price showed in choropleth map and the cluster with dotted color and cluster analysis as a popup.


65,852
196,534
327,216
457,898
588,580
719,263
Florida - K-Means clustering
Miami-Dade County:: Very High in Population, Highest in Crime, High HPI and High House Price, High in White American Population, Average Black American population, Below Average Asian American Population, Below Average Not Hispanic American Population, Below Average Hispanic American Population, More number of Universities, Night club, Museum and more amusement park
Montgomery
Baton Rouge
New Orleans
Lafayette
Jackson
Hattiesburg
Freeport
Nassau

# D. Discussion

As I have mentioned earlier Florida is a very rich state with respect to culture along with manu beautiful beaches, amusement, tourist attraction etc.. Exploring Florida need complex analysis and for that many different approach and data can be used.

In this project my analysis was limited to population, race, crime, HPI, Housing price and venue and I have used K-means algorithm to make a final conclusion, but as the area I choose is very vast, to get more detail analysis by each county, more granular data and other datasets can also be included to pin point the conclusion and obviously we can use other approaches like different machine learning technique to come to final decision.

I have perofirmed different analysis with all the different datasets and provided different visualization to get better understanding of the data.

Finally I have merged all the conclusion from my analysis and put the visualzion in a map.

# E. Conclusion

As Florida is a rich state with different culture and activity, it is always lucrative for the investiors to invest in such a place, and to get the best outcome from the investment its always necessary to dig through the different data to get a good understanding of the neighborhoods before investing.

# F. References

[1] Florida - Wikipedia

[2] Census

[3] Florida Department of Law Enforcement

[4] Florida Housing Finance Agency

[5] Google MAP API