IBM Data Science Professional Certificate

Applied Data Science Capstone Project

The Battle of the Neighborhoods

15th February 2021

Table of Contents

Executive Summary

The state of Florida is third largest state by population in the US and fourth largest by GDP. Due to the 134 million tourist visits annually and the highest domestic arrivals in the US, the state of Florida was chosen to start a business.

The challenge for most startups is to determine what type of business to open and to identify the most convenient location. This project uses data science to develop a model which can be used by entrepreneurs to answer the business problem of 'how to identify a business category and determine the geographic location of a startup which generates profit and maximizes revenue'.

The data used was sourced online and used to find the latitude and longitude of the counties. Foursquare was used to identify the various types of businesses in a geographic location. Exploratory data analysis was performed to identify the top ten counties for population using population density and average household income. Checks were conducted using pairplots and heatmaps to explore relationships between attributes.

The latitude and longitude of the trending venue and categories for the top 10 counties were identified using Foursquare. The venues were then grouped by counties and the top ten venues were listed for each county. One hot encoding was performed on the table to modify the categorical data. The k-means algorithm was deployed to create clusters of venue categories in each county.

The results of the cluster identified 19 food establishments, 7 home essentials stores, 1 clothing store and 10 other type establishments which ranged from dentists, cosmetic shops, gas stations and hotels. The four counties identified for a startup as Broward, Miami Dade, Orange and Pinellas. The most trending category of business in the four counties identified is the food/restaurant business.

The results indicate the counties of Miami Dade and Broward has the greatest potential for a successful business opportunity in the food restaurant industry. Additional analysis is required to determine location, type of restaurant, menu and pricing.

1. Introduction

Market research is one of the key elements for all successful startups. The study consists of analyzing the market to determine the viability of a new product or service. One of the challenges experienced by new entrants is the identification of a product and a convenient location for the business to maximize revenue. One solution is to geographically segment areas of interest by state city or neighborhood and the use of market orientation to identity the various trends in the market. The data obtained then is used to create a product or service to satisfy the market. This project uses a top down approach of data science and machine learning to develop a model which can be used by entrepreneurs to identify the most suitable location and business type for their startup. An unsupervised learning algorithm was deployed (k-means) using data generated from Foursquare. The foursquare application identifies trending venues and their types given the geographic coordinates of a city or neighborhood

The state of Florida is located in the south eastern region of the United States. Bordered by state of Georgia, the Gulf of Mexico and the Atlantic Ocean, the sunshine state has the longest coastline (1197 miles) in the US with 825 miles of beaches. Being the fourth largest economy in the US, with a population of 21.48 million and 131.4 million tourist visits spending USD40 billion annually, Florida was chosen as the state to start a business. Due its large land mass the challenge was the choice of location and the type of business. This project aims to identify the most appropriate county and the type of business for a startup.

2. Business Problem

How to identify a business category and determine the geographic location of a startup which generates profit and maximizes revenue.

3. Data Sources and Description

The data used for the analysis was sourced through www.simplemaps.com. This website provided updated data (Jan 2021) on latitude and longitude for over 108,000 cities and towns in the US for all 50 states. The geodata was used in conjunction with the folium library to visualize the geographic locations of the counties. The data was read into the IDE via a CSV file.

Another website used was https://data.census.gov/cedsci/. This is the official website of the US Census Bureau which provided the population, population density and average income per household for the counties in Florida. The data was used to analyze each county using their attributes to identify the most suitable county for a startup. The data was read into IDE via a CSV file.

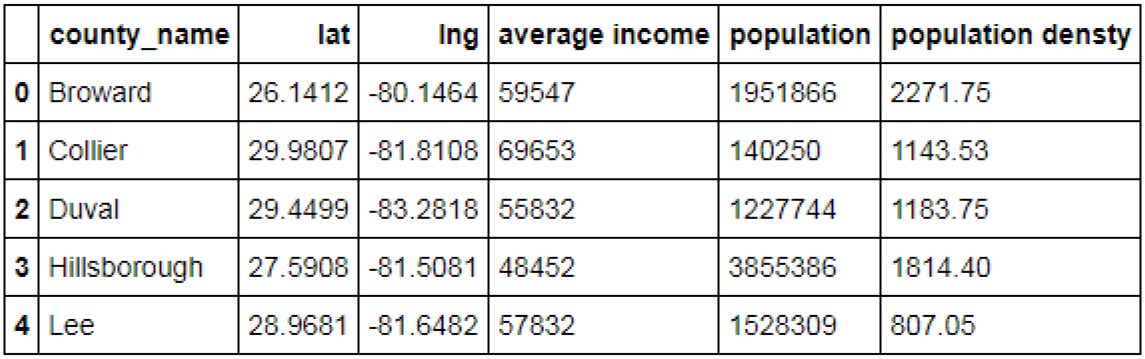
Foursquare is a social media mobile application used to discover various types of businesses in a geographic location. The application provided the trending venues by category for the cluster analysis model.

All of the datasets used in this project were checked for null/NAN values and duplicates. Columns with data not relevant to the scope of this project were removed (State Id, time zone, zip code, ranking, military, etc) The Foursquare category data was grouped by county and one hot encoded to convert the categorical data. Table 3.1 is a data dictionary for all of the datasets used in this project.

Table 3.1 – Data Dictionary

Term	Data Type	Description
County _name	Object	Name of the county
lat	Float64	Latitude of the object
lng	Float64	Longitude of the object
average income	Int64	Average income for each household in the county
population	Int64	Population of each county
Population density	Float64	Population density of each county
County	Object	Name of the county
County Latitude	Float64	Latitude of the county
County Longitude	Float64	Longitude of the county
Venue	String	Name of the venue
Venue latitude	Float64	Latitude of the venue
Venue Longitude	Float64	Longitude of the venue
Venue Category	String	Classification of the venue
Most Common Venue	String	10 most trending venues
Cluster Labels	Int64	The number of cluster in each county

Table 4.1 – County Data

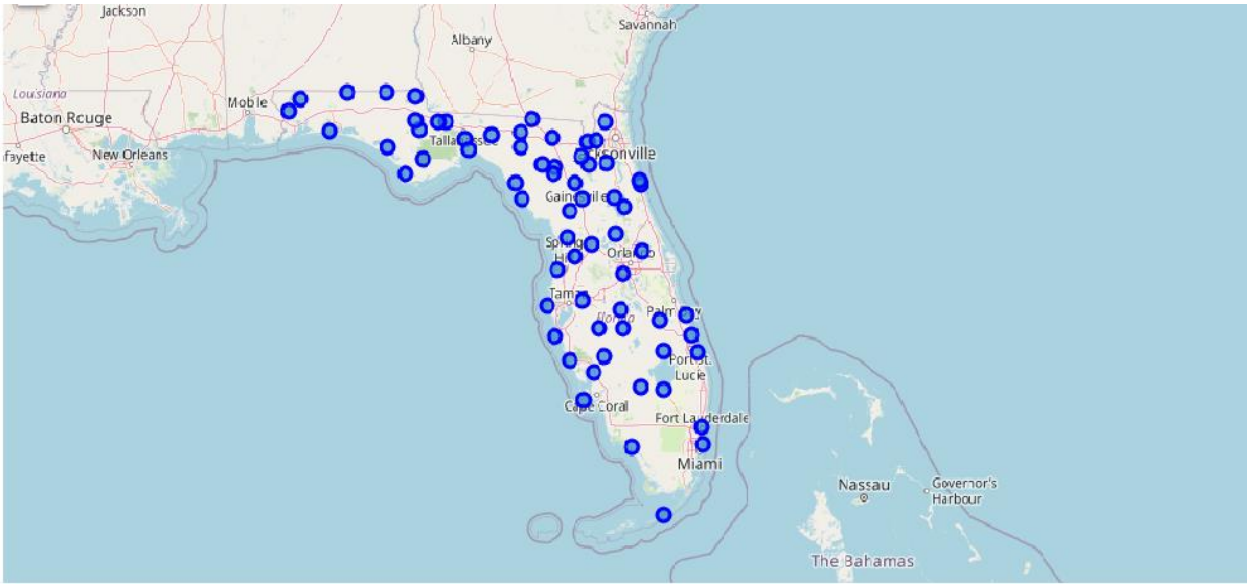


	county_name	lat	lng	average income	population	population densty
0	Broward	26.1412	-80.1464	59547	1951866	2271.75
1	Collier	29.9807	-81.8108	69653	140250	1143.53
2	Duval	29.4499	-83.2818	55832	1227744	1183.75
3	Hillsborough	27.5908	-81.5081	48452	3855386	1814.40
4	Lee	28.9681	-81.6482	57832	1528309	807.05

4. Analysis and Methodology

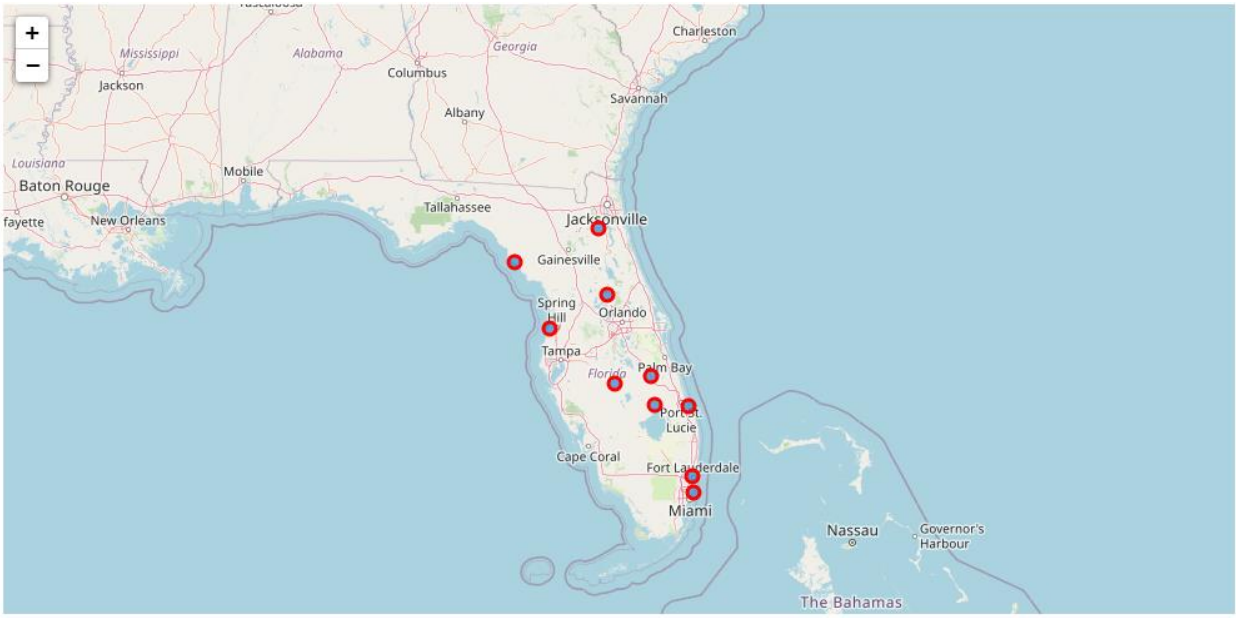
There are approximately 912 cities/towns listed in the 66 counties of Florida.

Fig 1.1 - County Locations in Florida



In order to choose the appropriate location for the startup, the top ten counties with the highest average income, population density and/or population were selected. The counties identified were Duval, Orange, Hillsborough, Miami Dade, Palm Beach, Broward, Pinellas, Monroe, Lee, and Collier (fig4.2). Their geographic locations are given below.

Fig4.2 – Top County Locations in Florida



4.1 Exploratory Data Analysis

The dataset used for preliminary analysis was shown in table 4.1 . This table provided data on population, population density and household income for each household per county. The following relationships were explored:

Population and Population Density per County

The graph in figure 4.3 shows the top three counties by population as Miami Dade, Hillsborough and Orange with Miami Dade having more than twice as much residents as Hillsborough however the top three counties by population density is Miami Dade, Broward and Hillsborough (fig 4.4). Orange County placed eight in population density. For this exercise population density will be used to analysis the locations due to potential for greater traffic and result increased revenue.

Average Household income per County

The graph in figure 4.5 showed the top three counties by household income as Monroe, Collier and Palm Beach. The counties of Broward Duval Miami Dade and orange came in fourth, sixth, eight and tenth respectively. Average household income is also a lead attribute to determine a successful location for a startup and will be a factor in the analysis however more weight will be given to population density for this exercise.

Fig4.3

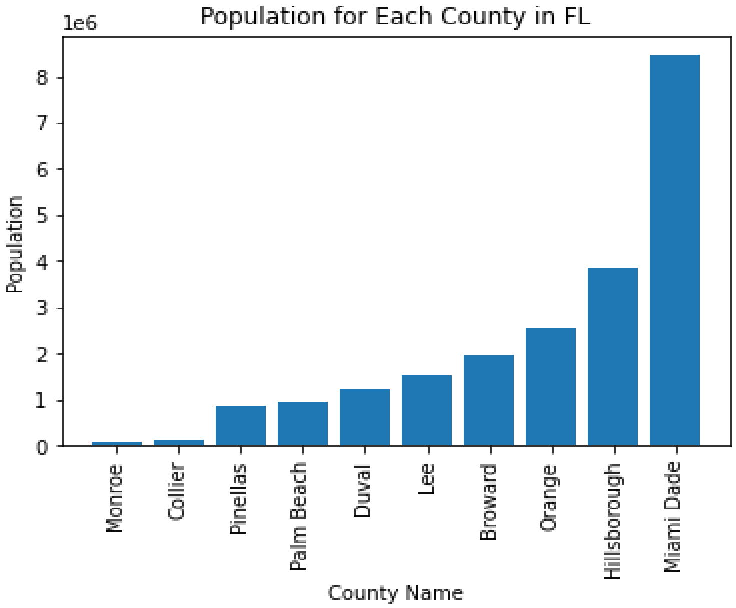


Fig4.4

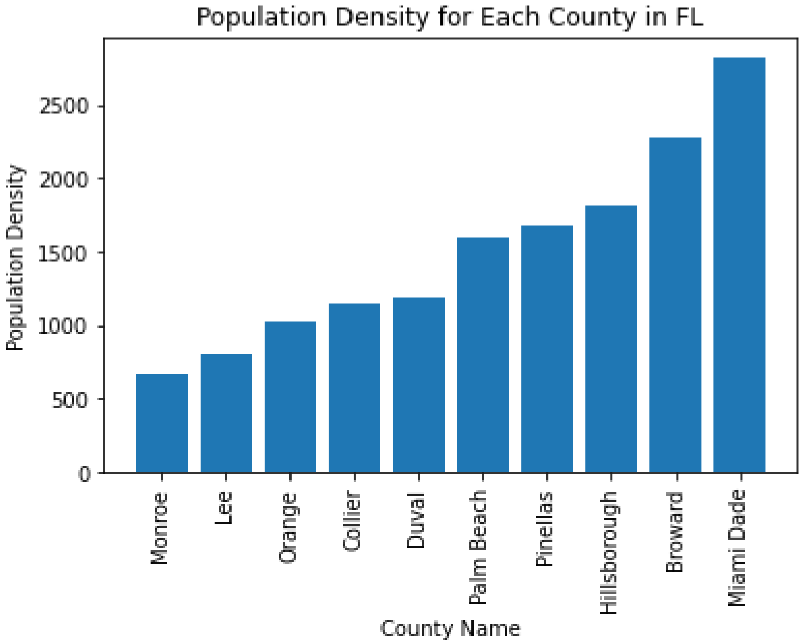
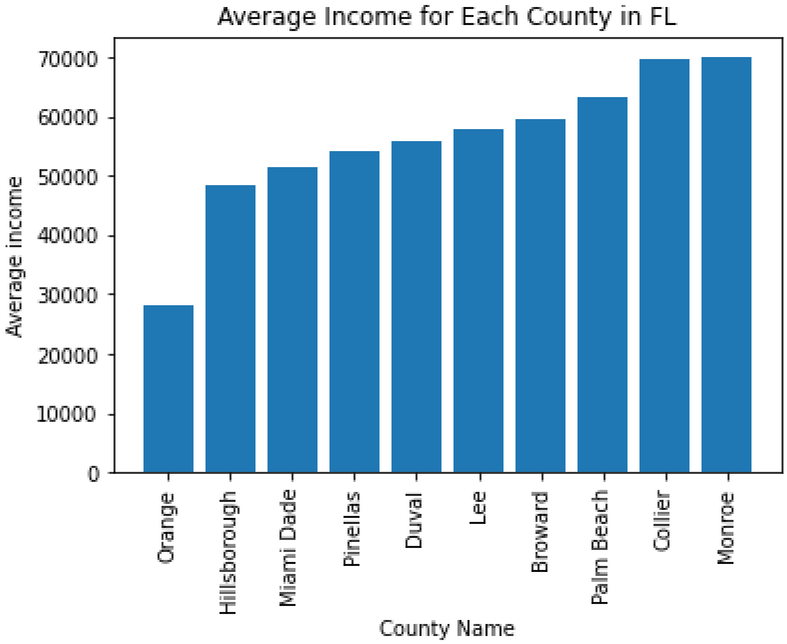

Population Density for Each County in FL
Population Density
0
500
1000
1500
2000
2500
Monroe
Lee
Orange
Collier
Duval
Palm Beach
Pinellas
Hillsborough
Broward
Miami Dade
County Name

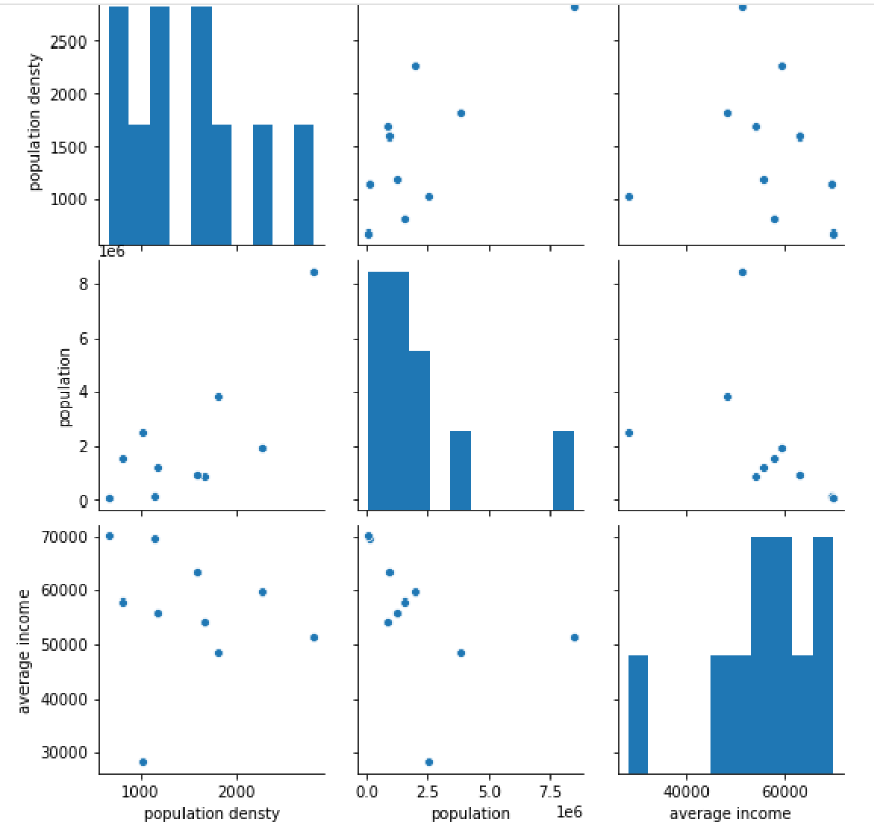
Fig4.5


Average Income for Each County in FL
Average income
0
10000
20000
30000
40000
50000
60000
70000
Orange
Hillsborough
Miami Dade
Pinellas
Duval
Lee
Broward
Palm Beach
Collier
Monroe
County Name

4.2 Correlation Analysis

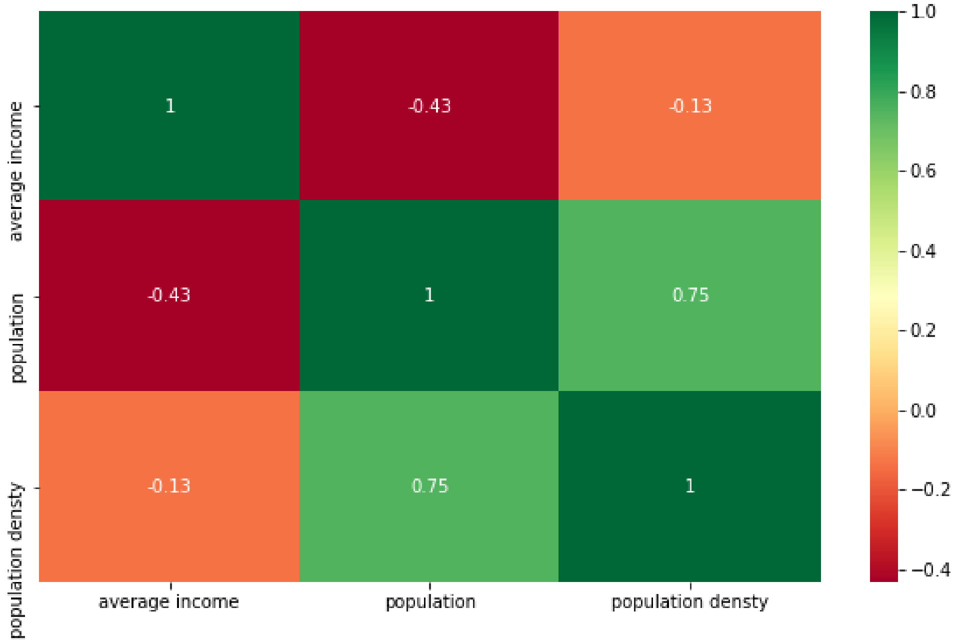
Additional analysis was undertaken to identify dependent relationships. This was achieved using pairplots and heatmaps.

Fig.4.6



The pairplots shown (fig.4.6.) indicates there is some relationship between population density and population however there was little relationship between average income and population or population density.

Fig 4.7 - Heatmap

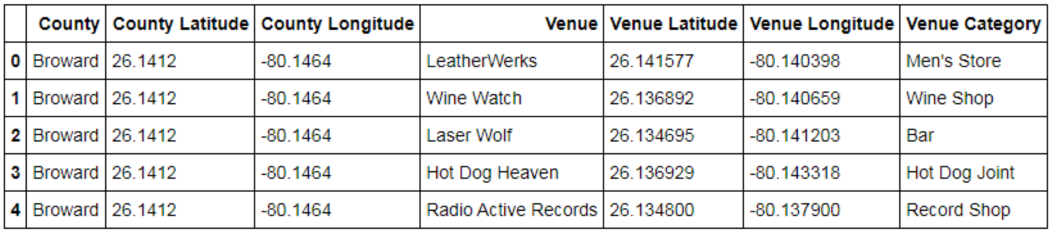


Further analysis using a heatmap shows the relationship between population density and population and possible relationship between population and income. There was no relationship average identified between income and population density

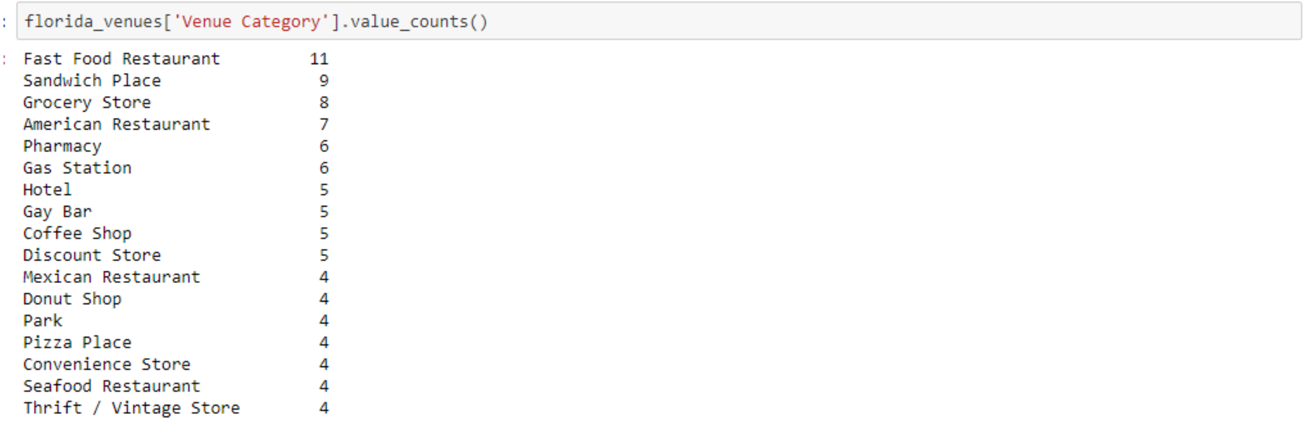
4.3 Foursquare Data

Foursquare is a local app which provides recommendations given a user's current location and interests. The Foursquare application was used to locate and identify trending venues in the 10 counties (Table 4.2) Foursquare provided the latitude and longitude, venue name and category for the top 196 trending venues for each state. The venues were grouped by counties and the top ten venues were listed for each county. Table (4.3)

Table 4.2 - Foursquare Results by County



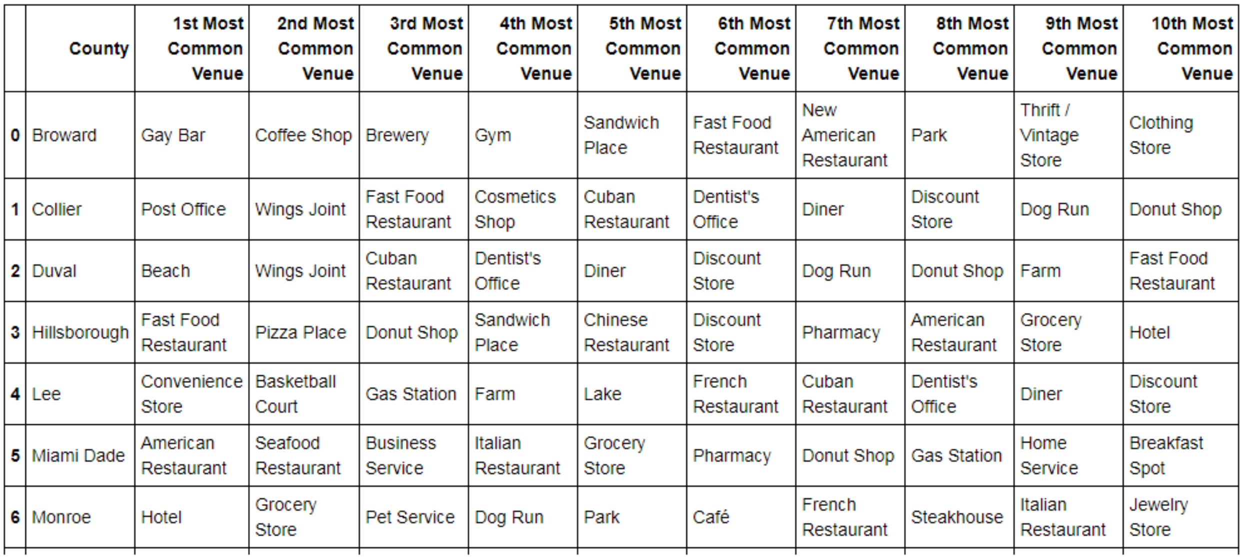
	County	County Latitude	County Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Broward	26.1412	-80.1464	LeatherWerks	26.141577	-80.140398	Men's Store
1	Broward	26.1412	-80.1464	Wine Watch	26.136892	-80.140659	Wine Shop
2	Broward	26.1412	-80.1464	Laser Wolf	26.134695	-80.141203	Bar
3	Broward	26.1412	-80.1464	Hot Dog Heaven	26.136929	-80.143318	Hot Dog Joint
4	Broward	26.1412	-80.1464	Radio Active Records	26.134800	-80.137900	Record Shop



```
: florida_venues['Venue Category'].value_counts()

: Fast Food Restaurant         11
  Sandwich Place                9
  Grocery Store                 8
  American Restaurant           7
  Pharmacy                      6
  Gas Station                   6
  Hotel                         5
  Gay Bar                       5
  Coffee Shop                   5
  Discount Store                5
  Mexican Restaurant            4
  Donut Shop                    4
  Park                          4
  Pizza Place                   4
  Convenience Store             4
  Seafood Restaurant            4
  Thrift / Vintage Store        4
```

Table 4.3 County Top 10 Venue by County

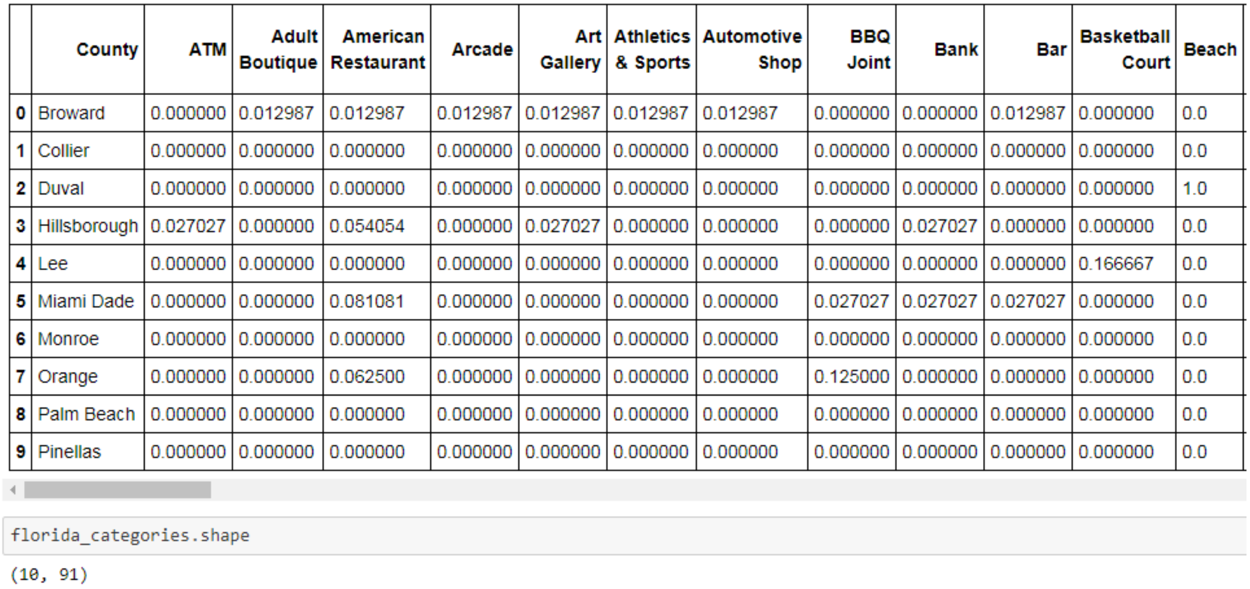


	County	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Broward	Gay Bar	Coffee Shop	Brewery	Gym	Sandwich Place	Fast Food Restaurant	New American Restaurant	Park	Thrift / Vintage Store	Clothing Store
1	Collier	Post Office	Wings Joint	Fast Food Restaurant	Cosmetics Shop	Cuban Restaurant	Dentist's Office	Diner	Discount Store	Dog Run	Donut Shop
2	Duval	Beach	Wings Joint	Cuban Restaurant	Dentist's Office	Diner	Discount Store	Dog Run	Donut Shop	Farm	Fast Food Restaurant
3	Hillsborough	Fast Food Restaurant	Pizza Place	Donut Shop	Sandwich Place	Chinese Restaurant	Discount Store	Pharmacy	American Restaurant	Grocery Store	Hotel
4	Lee	Convenience Store	Basketball Court	Gas Station	Farm	Lake	French Restaurant	Cuban Restaurant	Dentist's Office	Diner	Discount Store
5	Miami Dade	American Restaurant	Seafood Restaurant	Business Service	Italian Restaurant	Grocery Store	Pharmacy	Donut Shop	Gas Station	Home Service	Breakfast Spot
6	Monroe	Hotel	Grocery Store	Pet Service	Dog Run	Park	Café	French Restaurant	Steakhouse	Italian Restaurant	Jewelry Store

4.4 Cluster Analysis

In order to analyze the categorical data of the venues identified, one hot coding was performed to the Foursquare results (Table 4.4). The result was a table consisting of 10 rows and 91 columns

Table 4.4 – One Hot Encoding



	County	ATM	Adult Boutique	American Restaurant	Arcade	Art Gallery	Athletics & Sports	Automotive Shop	BBQ Joint	Bank	Bar	Basketball Court	Beach
0	Broward	0.000000	0.012987	0.012987	0.012987	0.012987	0.012987	0.012987	0.000000	0.000000	0.012987	0.000000	0.0
1	Collier	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0
2	Duval	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.0
3	Hillsborough	0.027027	0.000000	0.054054	0.000000	0.027027	0.000000	0.000000	0.000000	0.027027	0.000000	0.000000	0.0
4	Lee	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.166667	0.0
5	Miami Dade	0.000000	0.000000	0.081081	0.000000	0.000000	0.000000	0.000000	0.027027	0.027027	0.027027	0.000000	0.0
6	Monroe	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0
7	Orange	0.000000	0.000000	0.062500	0.000000	0.000000	0.000000	0.000000	0.125000	0.000000	0.000000	0.000000	0.0
8	Palm Beach	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0
9	Pinellas	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0

```
florida_categories.shape
```

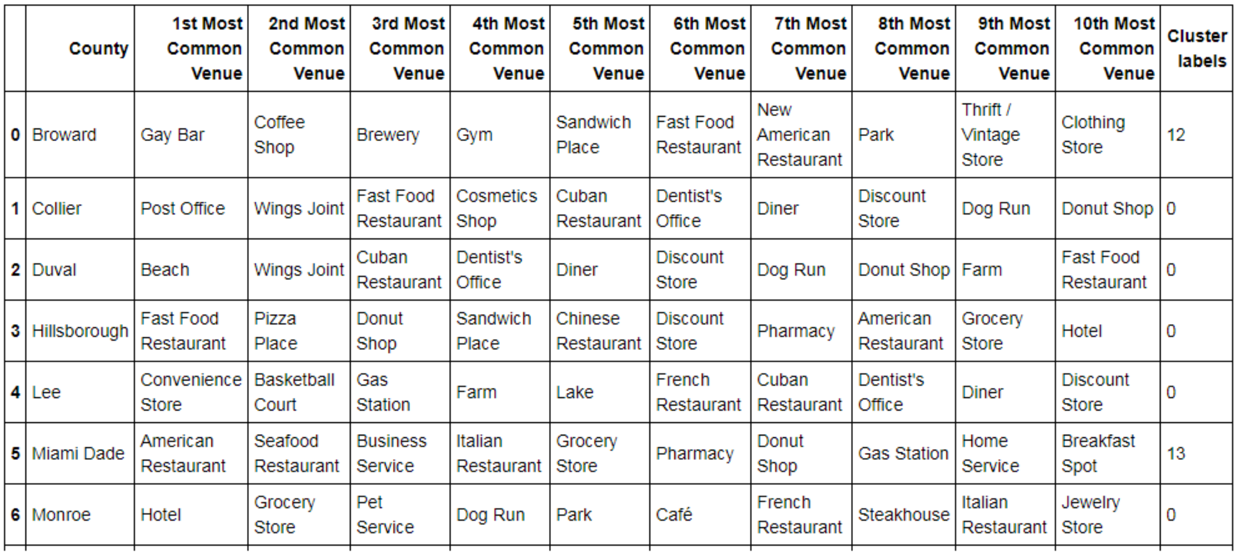
```
(10, 91)
```

The county column was removed and a cluster analysis was performed on the new table using k-means. This type of analysis will return the various trending categories of business types in each county to be considered for a startup.

5. Results and Discussion

The results of the cluster analysis were a group of clusters in the counties of Broward, Miami Dade, Orange and Pinellas (Table 5.1). Broward County results returned four food establishments, two bars, one clothing store and two other type business categories. Miami Dade returned five food establishments, three home essentials and two other type businesses. Orange results returned six food establishments, one home essentials and two other type businesses. Pinellas returned four food establishments, two home essentials and four other type establishments. The coordinates of the clusters for each county were plotted as shown in the map in figure 5.1 .

Table 5.1 County Clusters

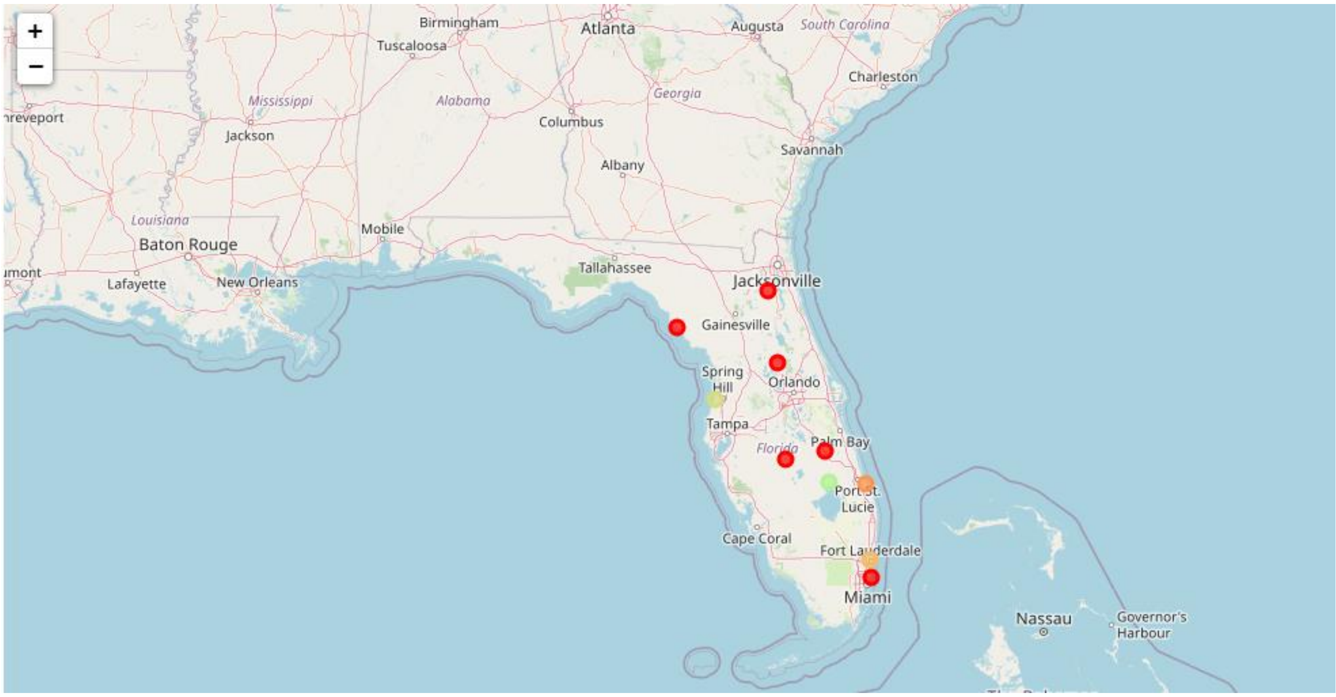


	County	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Cluster labels
0	Broward	Gay Bar	Coffee Shop	Brewery	Gym	Sandwich Place	Fast Food Restaurant	New American Restaurant	Park	Thrift / Vintage Store	Clothing Store	12
1	Collier	Post Office	Wings Joint	Fast Food Restaurant	Cosmetics Shop	Cuban Restaurant	Dentist's Office	Diner	Discount Store	Dog Run	Donut Shop	0
2	Duval	Beach	Wings Joint	Cuban Restaurant	Dentist's Office	Diner	Discount Store	Dog Run	Donut Shop	Farm	Fast Food Restaurant	0
3	Hillsborough	Fast Food Restaurant	Pizza Place	Donut Shop	Sandwich Place	Chinese Restaurant	Discount Store	Pharmacy	American Restaurant	Grocery Store	Hotel	0
4	Lee	Convenience Store	Basketball Court	Gas Station	Farm	Lake	French Restaurant	Cuban Restaurant	Dentist's Office	Diner	Discount Store	0
5	Miami Dade	American Restaurant	Seafood Restaurant	Business Service	Italian Restaurant	Grocery Store	Pharmacy	Donut Shop	Gas Station	Home Service	Breakfast Spot	13
6	Monroe	Hotel	Grocery Store	Pet Service	Dog Run	Park	Café	French Restaurant	Steakhouse	Italian Restaurant	Jewelry Store	0

A summary of the results for the four counties is as follows:

- 19 Food establishments – This includes restaurants, donut shops, diners and fast food restaurants
- 7 Home essentials –This includes home services, discount stores and pharmacies
- 1 Clothing store
- 10 Other establishments. This ranged from dentists, post offices, cosmetic shops gas stations and hotels

Fig 5.1 County Cluster locations



6. Conclusion

The model results identified four potential counties for a startup as Broward, Miami Dade, Orange and Pinellas. The most trending category of business in the four counties identified was the food/restaurant business. This is followed by other types which ranged from gas stations to hotels and home essentials. The counties of Miami Dade and Broward were identified as the first and second most densely populated counties while Pinellas was the fourth most densely populated and Orange having the lowest average income per family.

This project lays the groundwork of the market research necessary for a successful startup in the state of Florida. The results show the counties of Miami Dade and Broward has the greatest potential for a successful business opportunity in the food restaurant industry. This analysis should be supported with additional market research to determine location, type of restaurant, menu, pricing in addition to a rigorous business plan.

References

https://data.census.gov/cedsci/

https://simplemaps.com/data/us-cities