IBM Applied Data Science Capstone

Recommending a Data Science Business in South Florida

Alfredo Castañeda

Table of Contents

1. Introduction 3
2. Background of Study (Literature Review) 3
   2.1. Preliminary Analysis of Business Formation Statistics 3
   Total Business Applications from for top ten Counties in Florida 3
   Annual Total of Business Applications by Top Five Counties 4
3. Business Problem 5
4. Methodology – Clustering through use of k-means 5
   4.1. Collecting the Data 5
   4.2. Building the Data Set 6
   4.3. Description of Data Set 6
   4.4. Mapping Industry categories 7
   4.5. Preprocessing / Wrangling 7
   4.6. Mining the Data (Data Visualization) 9
   Which Cities have the most number for businesses 9
   Relationship between DB Hoover industry and total businesses 10
   Relationship between Owner type and number of businesses 12
   Relationship between Parent Country and total businesses 12
5. Analysis of k-means 12
6. Results / Discussion 13
7. Conclusion / Recommendations 15
8. References / Acknowledgement 16

1. Introduction

Practicing data science provides me an exciting and interesting career opportunity to leverage my mathematics, engineering background and systems thinking approach to solving practical business problems. Moreover, it allows me to choose from several locations to reside since I need not be on site to provide this service to local organization. One form of employment is to provide consulting or contract employment. If successful I would consider opening a firm.

This project will explore setting up a firm that provides data science services as a business in Florida. The project will attempt to use the data science approach to develop a strategy to identify potential opportunities to provide data science tools to local organizations to solve their business challenges.

2. Background of Study (Literature Review)

Enterprise Florida, Inc. (EFI) is a public-private partnership between Florida's business and government leaders and is the principal economic development organization for Florida.

EFI promotes various advantages that Florida provides including:

- 4th largest economy in the U.S. and 18th largest economy in the world (if Florida were a nation). (BEA & IMF).
- Ranked 4th in the nation for high-tech employment by CompTIA, Florida boasts nearly 237,000 high-tech workers.
- Number 4 in high-tech employment in the US (TechAmercia).
- Number 2 business birthrate (U.S. Chamber of Commerce).
- Number 3 for high-tech establishments (TechAmerica).

2.1. Preliminary Analysis of Business Formation Statistics

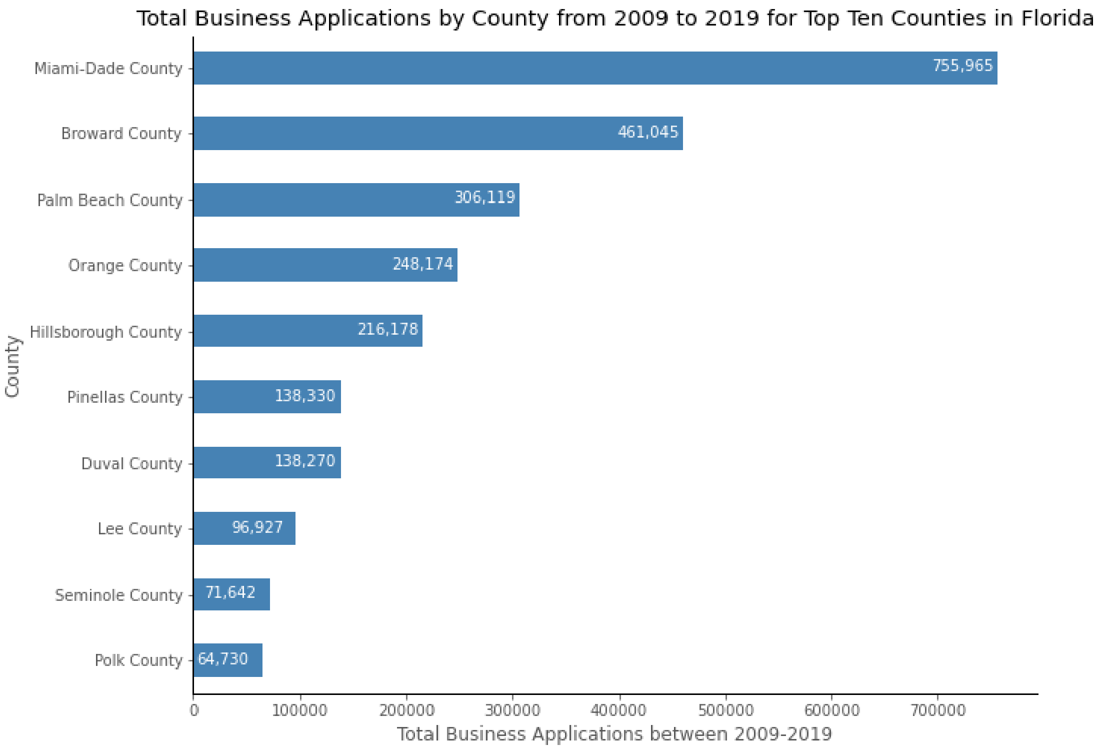
The US Census Bureau Let collects business startup statistics these are the starting point to identify what counties indicate economic growth in Florida. We will limit the review to the last ten years ending in 2019.

This United States Census Bureau offers a downloadable csv file that contains data for all 50 states and corresponding counties. This file can be located at website Business Formation Statistics (census.gov).

Total Business Applications from for top ten Counties in Florida

Data wrangling steps performed on csv file.

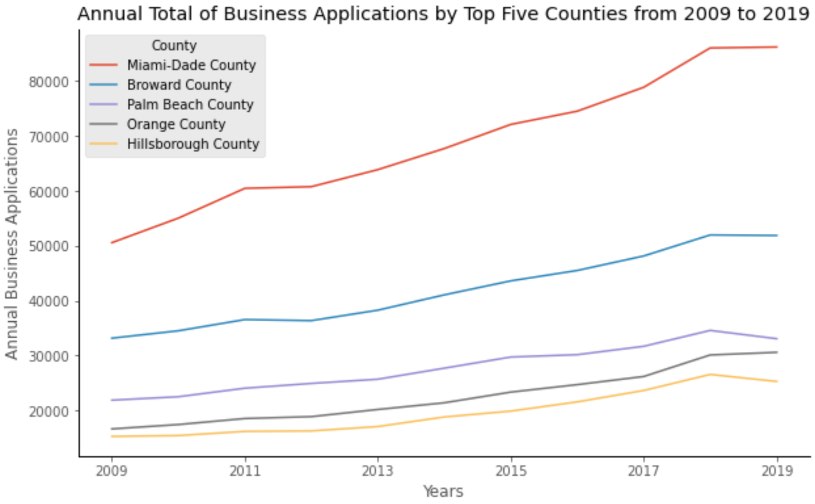
- The file was downloaded as an Excel file.
- To limit the analysis to ten years (2009-2019) columns with earlier years were deleted.
- Data was sliced to evaluate only state of Florida for each county.
- Data over ten years was summed then sorted for top ten counties in Florida.
- See Total Business Application chart.



Annual Total of Business Applications by Top Five Counties

To look at trending over ten years additional steps were performed on data frame.

- Data was sliced to evaluate the top five counties base on totals from prior chart.
- Dropped the totals column then transposed the columns to view data in chart.



3. Business Problem

The objective of this project is to analyze the businesses located within the cities of both Broward and Palm Beach counties. These represent the second and third counties with the highest new business applications over a ten-year period across the 67 counties in Florida. These two counties are also within the Miami Metropolitan area. The Miami Metropolitan also referred to as the Miami-Fort Lauderdale-West Palm Beach metropolitan area is the seventh largest metropolitan area in the United States.

A key *assumption* here is that data science services can be done effectively with occasional in-person interactions reinforced through virtual meetings. Therefore, we can expand the clientele across larger distance (across both counties) to increase potential for economic benefit.

The data search will focus on businesses that fall within the four key industry groups that are well represented in South Florida based on the Industry Key Areas. Both these counties exhibited strong representation of businesses as illustrated by corresponding industry specific choropleth maps for each key industry see links below.

- Information Technology
- Life Sciences
- Financial
- Clean Tech

The *benefit* of this study is that it produces a strategy that can be tailored to service the four key industries from either a location in Broward county or Palm Beach County. In addition, this strategy can also be used to focus on specific industries.

The *target audience* for this project includes people who are interested in practicing data science or providing technology services / products that are applicable to the selected industries mentioned within and across both counties.

4. Methodology – Clustering through use of k-means

The methodology used for this project includes several steps: Collecting the data, Building the data set, Mapping categories, Preprocessing, Transforming, Mining, and Analysis.

4.1. Collecting the Data

D & B Hoovers is among the world's largest commercial database of 120 million business records. Fortunately, I have limited access during the time of this project to research businesses based on select criteria. I have used the following parameters to work within the limited access and obtain records for businesses based on the following criteria:

- Located within each county Broward or Palm Beach
- Employees (Across All Sites) - more than 25 employees
- Annual Revenue - More than $5M annual revenue
- D & B Hoovers categories (later they will be matched to Enterprise Florida Industries)

The extracted D&B Excel file contains records for 495 businesses along with the business attributes, address information.

4.2. Building the Data Set

To build the set I started with the DB Hoovers downloaded Excel file then performed some activities to seed the data set with geo coordinates for each respective business.

Google Sheets provides an add on (Geocode by Awesome table) that can seed a spreadsheet with geo coordinates provided adequate/accurate address information is provided. After uploading the Excel file onto google sheets I then used Geocode to seed the file for each business location.

Lastly after downloading the seeded file I used a VBA macro in Excel to calculate distances to two central locations (at the heart of each county) using the Haversine equations.

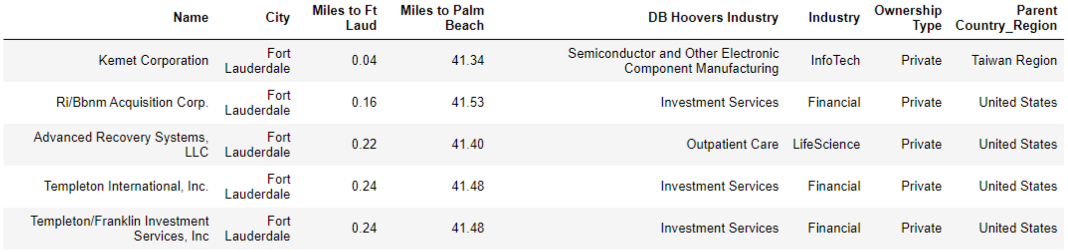
Operations in Excel & Google Sheets on initial Excel file from DB Hoovers Database included:

- Removed unnecessary Columns - Deleted rows
- Seeded geo coordinates in google sheets with geo cords app then verified sample values
- Added VBA Macro to calculate distances for Fort Lauderdale and Palm Beach locations
- Created two columns to stored miles from each business to Fort Lauderdale and Palm Beach Locations

Note: After calculating distances the formulas were verified then to simplify transfer a csv file was exported from Excel with the values, no macros/formulas.

4.3. Description of Data Set

The collected dataset consists of a single data frame with 495 rows and 12 columns containing Company Name, City, Postal Code, Latitude, Longitude, Distance to FL, Distance to PB, Industry, Revenue, Employees, Ownership Type, and Parent Country/Region.



Name	City	Miles to Ft Laud	Miles to Palm Beach	DB Hoovers Industry	Industry	Ownership Type	Parent Country_Region
Kemet Corporation	Fort Lauderdale	0.04	41.34	Semiconductor and Other Electronic Component Manufacturing	InfoTech	Private	Taiwan Region
Ri/Bbnm Acquisition Corp.	Fort Lauderdale	0.16	41.53	Investment Services	Financial	Private	United States
Advanced Recovery Systems, LLC	Fort Lauderdale	0.22	41.40	Outpatient Care	LifeScience	Private	United States
Templeton International, Inc.	Fort Lauderdale	0.24	41.48	Investment Services	Financial	Private	United States
Templeton/Franklin Investment Services, Inc	Fort Lauderdale	0.24	41.48	Investment Services	Financial	Private	United States

4.4. Mapping Industry categories

The B&D Hoovers database categories were mapped to the Enterprise Florida key Industries after reviewing the corresponding Industry briefs and maps from Enterprise Florida. This was achieved by a creating an Industry Column that matched each key industry to the corresponding DB Hoover Industry category as illustrated in following table.

Table 1 – Mapping Industry Categories

Key Industry Category	Mapped Values to Category
InfoTech	Business Support Services, Market Research and Opinion Polling, Computer and Peripheral Equipment Manufacturing, Computer Programming, Computer System Design Services, Data Processing, Internet and Web Services, Software, Magnetic and Optical Media Manufacturing, Semiconductor and Other Electronic Component Manufacturing, Communications Equipment Manufacturing, Miscellaneous Telecommunication Services, Broadcasting and Media, Miscellaneous Information Services
LifeScience	Health and Personal Care Wholesale, Pharmaceutical Manufacturing, Ambulatory Services, Dentists, Diagnostic Laboratories, Hospitals, Medical Equipment and Supplies, Outpatient Care, Physicians and Health Practitioners.
Financial	Banking, Investment Banking, Investment Services, Mortgage and Credit
Clean Tech	Electricity Generation and Distribution.

4.5. Preprocessing / Wrangling

Preprocessing activities summarized in table below.

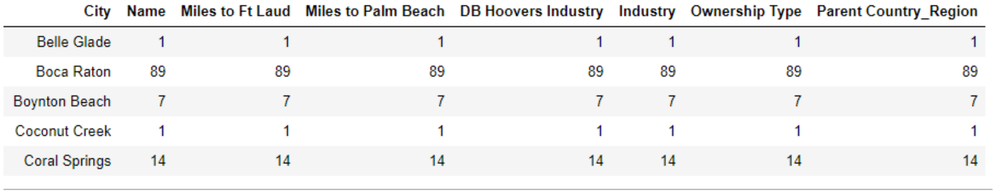
Table 2 - Preprocessing steps

Data / Activity	Columns / Rationale
Industry classification codes / Removed	ANZSIC 2006, NACE Rev 2, ISIC Rev 4, UK SIC 2007, NAICS 2017, US SIC 1987, US 8-Digit SIC Data was redundant with DB Hoover categories.
Redundant information / Removed	Parent company information: Ultimate Parent Country/Region, Parent Company/Region, State or Province. Address Line, Postal Code
Incomplete or missing information for entire data set / Removed	Assets, PreTax Profit
Information not relevant to study / Removed	Direct Marketing Status, Source, Key ID, TPS Flag, Ticker, Is Headquarters, URL, Fax, Phone, Entity Type
Dropped redundant and unneeded columns	Postal Code, Revenue & Employees – ***see note**** Latitude & Longitude – N/A using distance from one location Palm Beach.
Replace missing values for Names and Employees at site	Employees column was updated with missing values from Employees_All.

****Note:*** *Data is not accurate for every business since many are subsidiaries or business units within larger corporations. Therefore information in P&L statements is consolidated differently for each making it difficult to assign contribution for each site.*

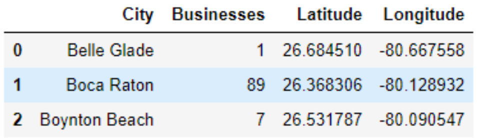
The Data wrangling included an additional 7 steps.

1. First aggregating the data by city.



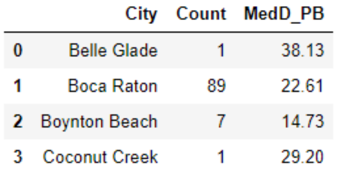
City	Name	Miles to Ft Laud	Miles to Palm Beach	DB Hoovers Industry	Industry	Ownership Type	Parent Country_Region
Belle Glade	1	1	1	1	1	1	1
Boca Raton	89	89	89	89	89	89	89
Boynton Beach	7	7	7	7	7	7	7
Coconut Creek	1	1	1	1	1	1	1
Coral Springs	14	14	14	14	14	14	14

2. Then creating a small data frame for seeding with Geocode in sheets for the 38 cities. This file would be used latter with map illustrations.



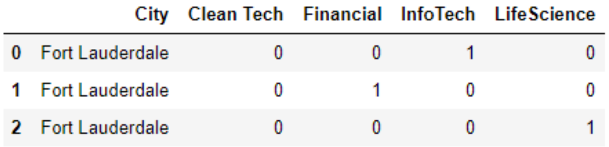
	City	Businesses	Latitude	Longitude
0	Belle Glade	1	26.684510	-80.667558
1	Boca Raton	89	26.368306	-80.128932
2	Boynton Beach	7	26.531787	-80.090547

3. Next calculated the median distance of the location in miles of businesses within a city to Palm Beach. This provides a reference distance to use for clustering algorithm.



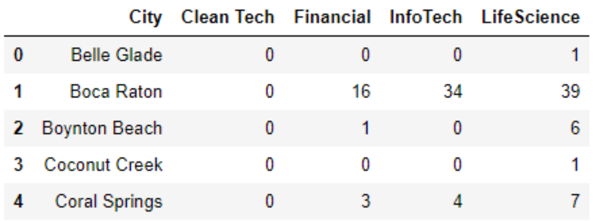
	City	Count	MedD_PB
0	Belle Glade	1	38.13
1	Boca Raton	89	22.61
2	Boynton Beach	7	14.73
3	Coconut Creek	1	29.20

4. Followed by creating dummy variables to transform the key industry categories into variables.



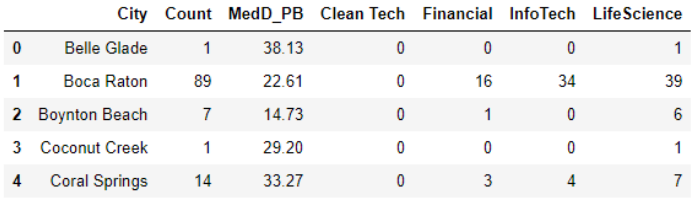
	City	Clean Tech	Financial	InfoTech	LifeScience
0	Fort Lauderdale	0	0	1	0
1	Fort Lauderdale	0	1	0	0
2	Fort Lauderdale	0	0	0	1

5. After that aggregated the results by city to get the counts by industry.



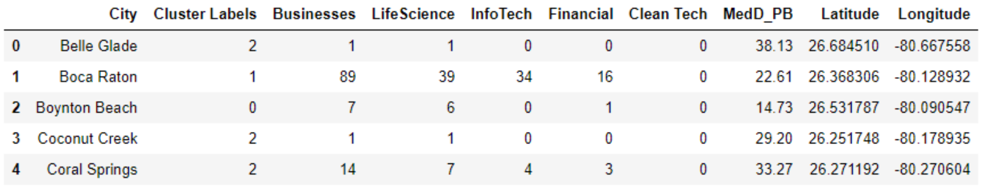
	City	Clean Tech	Financial	InfoTech	LifeScience
0	Belle Glade	0	0	0	1
1	Boca Raton	0	16	34	39
2	Boynton Beach	0	1	0	6
3	Coconut Creek	0	0	0	1
4	Coral Springs	0	3	4	7

6. Now the data frame is merged to add median distance, total count and industry category counts to creating the file for the k-means cluster algorithm



	City	Count	MedD_PB	Clean Tech	Financial	InfoTech	LifeScience
0	Belle Glade	1	38.13	0	0	0	1
1	Boca Raton	89	22.61	0	16	34	39
2	Boynton Beach	7	14.73	0	1	0	6
3	Coconut Creek	1	29.20	0	0	0	1
4	Coral Springs	14	33.27	0	3	4	7

7. After the model ran, I then merged the city geo-location data and added the cluster labels to segregate the clusters for final examination. This is final data frame.



	City	Cluster Labels	Businesses	LifeScience	InfoTech	Financial	Clean Tech	MedD_PB	Latitude	Longitude
0	Belle Glade	2	1	1	0	0	0	38.13	26.684510	-80.667558
1	Boca Raton	1	89	39	34	16	0	22.61	26.368306	-80.128932
2	Boynton Beach	0	7	6	0	1	0	14.73	26.531787	-80.090547
3	Coconut Creek	2	1	1	0	0	0	29.20	26.251748	-80.178935
4	Coral Springs	2	14	7	4	3	0	33.27	26.271192	-80.270604

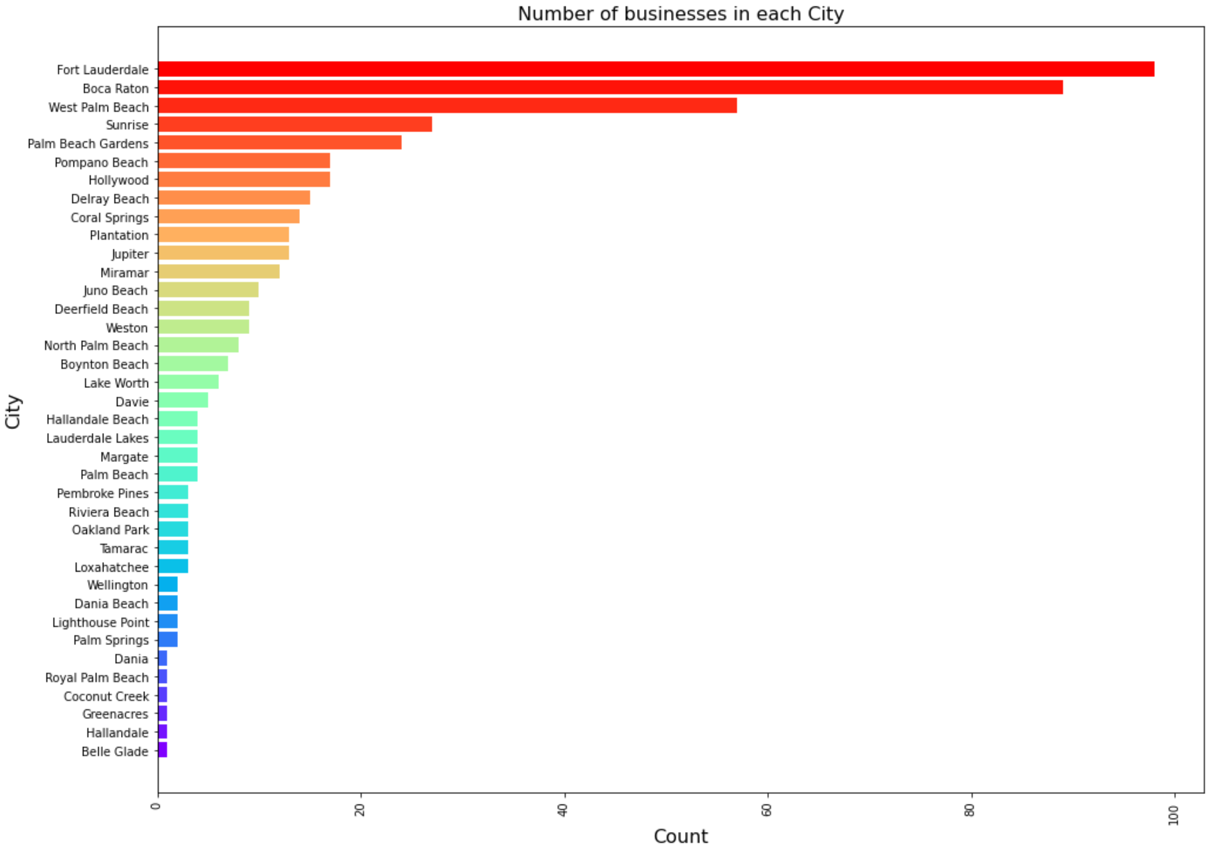
4.6. Mining the Data (Data Visualization)

After the data set was preprocessed. The data was mined through exploratory analysis. This involved characterization of the following attributes. This activity guided the selection of attributes for k-means algorithm. The exploration included evaluating:

- Number for Businesses per city as defined per the criteria and frequency across data set
- DB Hoovers Industry category counts and frequency across data set
- Ownership Type category counts and frequency across data set
- Parent Country counts and frequency across data set
- Omission of Revenue and Employee data from further analysis

Which Cities have the most number for businesses

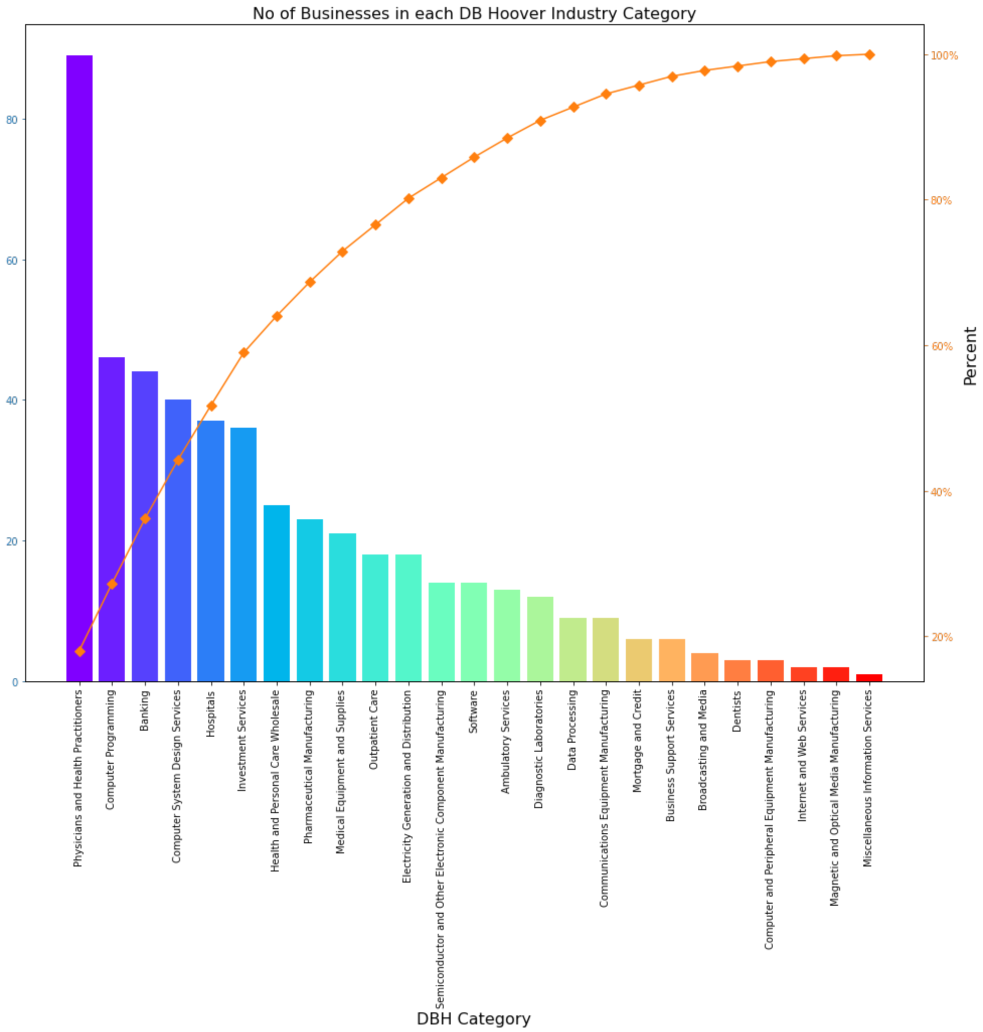
The bar chart below illustrates the number of businesses per city in descending order. The chart illustrates a higher concentration of economic activity with the top three cities: Fort Lauderdale, Boca Raton and West Palm Beach.



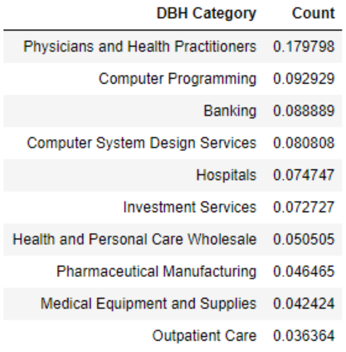
Beyond the top five cities the number of businesses declines significantly. This can later be explored through the clustering algorithm.

Relationship between DB Hoover industry and total businesses

The next attribute to explore is the relationship between the DB Hoover Industry category and the businesses in total. We start here to explore if we can mine the data set with a higher resolution such as using the DB Hoover categories or continue to use the Enterprise Florida Key Industries. The bar and pareto chart below illustrate the counts and distribution by DB Hoover categories.



The above chart indicates that Physicians and Health Practitioners make up about 18% of the total businesses.

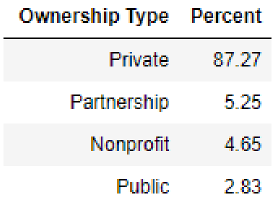


DBH Category	Count
Physicians and Health Practitioners	0.179798
Computer Programming	0.092929
Banking	0.088889
Computer System Design Services	0.080808
Hospitals	0.074747
Investment Services	0.072727
Health and Personal Care Wholesale	0.050505
Pharmaceutical Manufacturing	0.046465
Medical Equipment and Supplies	0.042424
Outpatient Care	0.036364

A closer review of the data as illustrated in the table at the left shows only the top ten categories. Nevertheless, except for the Physicians category the distribution across categories seems to be a broad.

Based on this information we will continue with the intention of using the key industry categories versus the DB Hoover categories for further analysis.

Relationship between Owner type and number of businesses

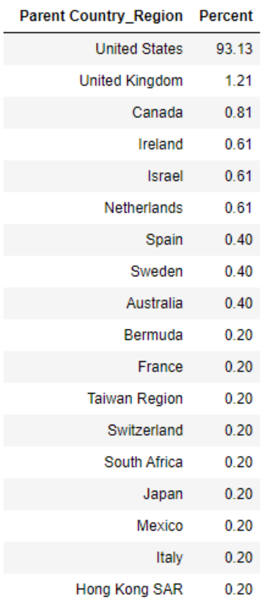


Ownership Type	Percent
Private	87.27
Partnership	5.25
Nonprofit	4.65
Public	2.83

The distribution of ownership type is heavily weighted towards private industry at 87.3 %. Adding partnerships, the total increases to 92.5%. Refer to adjacent table for summary of results.

This factor will not influence further analysis and will not be included in clustering.

Relationship between Parent Country and total businesses



Parent Country_Region	Percent
United States	93.13
United Kingdom	1.21
Canada	0.81
Ireland	0.61
Israel	0.61
Netherlands	0.61
Spain	0.40
Sweden	0.40
Australia	0.40
Bermuda	0.20
France	0.20
Taiwan Region	0.20
Switzerland	0.20
South Africa	0.20
Japan	0.20
Mexico	0.20
Italy	0.20
Hong Kong SAR	0.20

The distribution of parent country is heavily weighted towards the United States at 93.13 %. South Florida is considered an area with international economic activity, but the remaining countries are diversity distributed making up less than 7% of the total businesses withing the key industries.

Refer to adjacent table for summary of results.

This factor will not influence further analysis and will not be included in clustering.

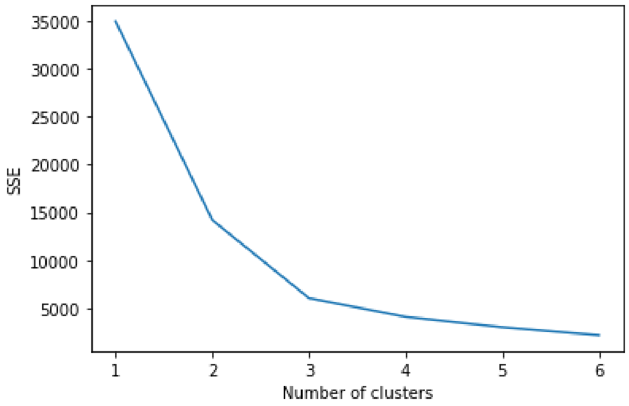
5. Analysis of k-means

This project will apply clustering *k-means* to explore a data set of businesses then partition the data into clusters. These clusters can then be analyzed and characterized.

The k-means clustering algorithm will group the 38 cities into clusters aggregating them into ‘n’ number of clusters in which each observation (city) belongs to a given cluster. It will use the selected attributes to determine how to aggregate / assign each city based on select criteria.

The elbow method will be used to determine the optimum value of k to perform k-means clustering.

The resulting graph is below illustrating the values for the number of clusters.

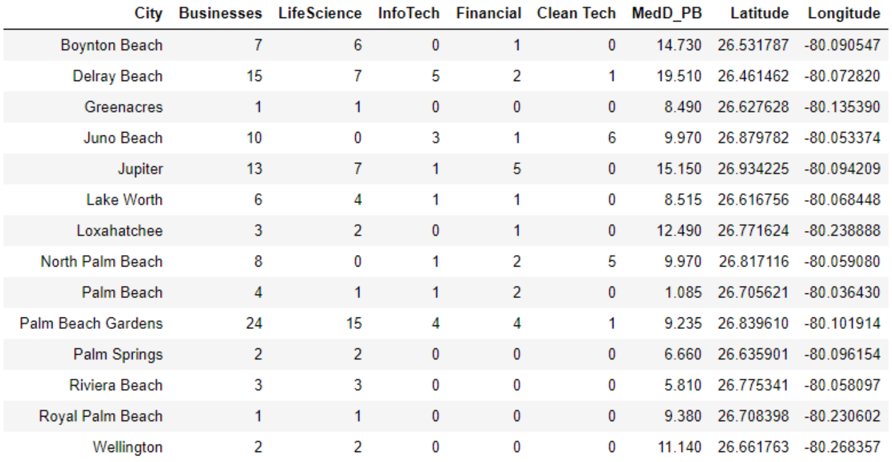


The graph indicates that the elbow is at 3 clusters. Notice at this point the slope begins to decrease less dramatically compared to the prior value by a significant amount. This appears as an elbow. Hence, we will run the algorithm with k=3.

6. Results / Discussion

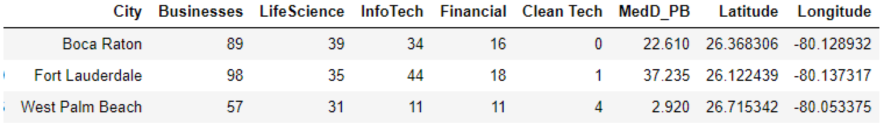
Since we identified three clusters, we will get three corresponding groupings to segregate the 38 cities. We can then examine the characteristics for each cluster.

The first cluster below will be referred to as the North Cluster. It contains mostly cities in Palm Beach county.



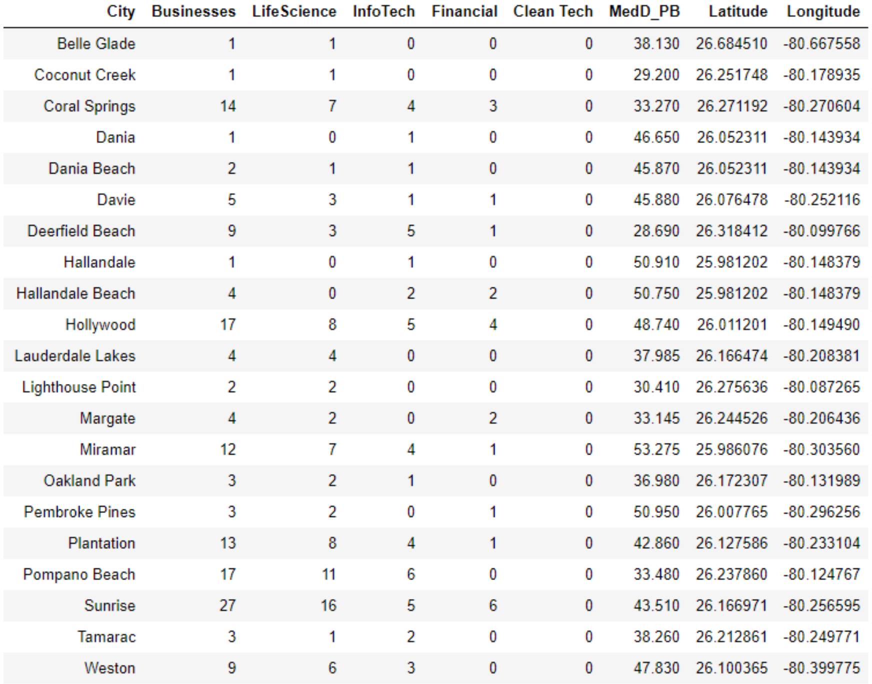
City	Businesses	LifeScience	InfoTech	Financial	Clean Tech	MedD_PB	Latitude	Longitude
Boynton Beach	7	6	0	1	0	14.730	26.531787	-80.090547
Delray Beach	15	7	5	2	1	19.510	26.461462	-80.072820
Greenacres	1	1	0	0	0	8.490	26.627628	-80.135390
Juno Beach	10	0	3	1	6	9.970	26.879782	-80.053374
Jupiter	13	7	1	5	0	15.150	26.934225	-80.094209
Lake Worth	6	4	1	1	0	8.515	26.616756	-80.068448
Loxahatchee	3	2	0	1	0	12.490	26.771624	-80.238888
North Palm Beach	8	0	1	2	5	9.970	26.817116	-80.059080
Palm Beach	4	1	1	2	0	1.085	26.705621	-80.036430
Palm Beach Gardens	24	15	4	4	1	9.235	26.839610	-80.101914
Palm Springs	2	2	0	0	0	6.660	26.635901	-80.096154
Riviera Beach	3	3	0	0	0	5.810	26.775341	-80.058097
Royal Palm Beach	1	1	0	0	0	9.380	26.708398	-80.230602
Wellington	2	2	0	0	0	11.140	26.661763	-80.268357

The next cluster is only three cities. This cluster represents cities with largest economic activity (businesses). We will call this one the highest density cluster.



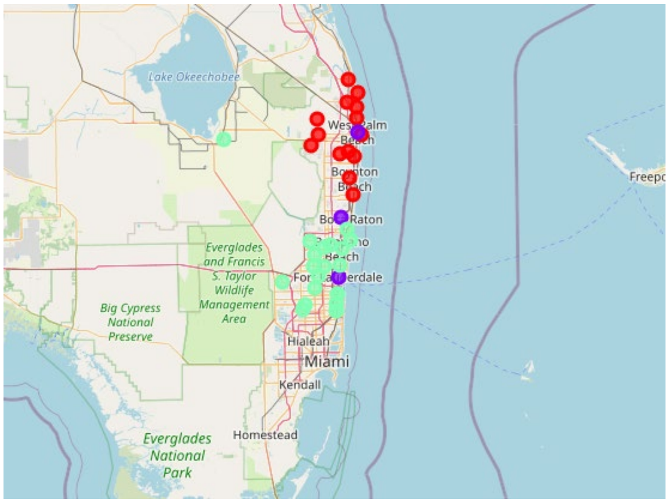
City	Businesses	LifeScience	InfoTech	Financial	Clean Tech	MedD_PB	Latitude	Longitude
Boca Raton	89	39	34	16	0	22.610	26.368306	-80.128932
Fort Lauderdale	98	35	44	18	1	37.235	26.122439	-80.137317
West Palm Beach	57	31	11	11	4	2.920	26.715342	-80.053375

The third cluster below will be referred to as the South Cluster. It contains mostly cities in Broward county.

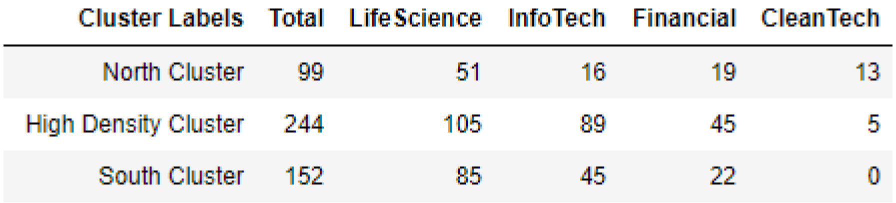


City	Businesses	LifeScience	InfoTech	Financial	Clean Tech	MedD_PB	Latitude	Longitude
Belle Glade	1	1	0	0	0	38.130	26.684510	-80.667558
Coconut Creek	1	1	0	0	0	29.200	26.251748	-80.178935
Coral Springs	14	7	4	3	0	33.270	26.271192	-80.270604
Dania	1	0	1	0	0	46.650	26.052311	-80.143934
Dania Beach	2	1	1	0	0	45.870	26.052311	-80.143934
Davie	5	3	1	1	0	45.880	26.076478	-80.252116
Deerfield Beach	9	3	5	1	0	28.690	26.318412	-80.099766
Hallandale	1	0	1	0	0	50.910	25.981202	-80.148379
Hallandale Beach	4	0	2	2	0	50.750	25.981202	-80.148379
Hollywood	17	8	5	4	0	48.740	26.011201	-80.149490
Lauderdale Lakes	4	4	0	0	0	37.985	26.166474	-80.208381
Lighthouse Point	2	2	0	0	0	30.410	26.275636	-80.087265
Margate	4	2	0	2	0	33.145	26.244526	-80.206436
Miramar	12	7	4	1	0	53.275	25.986076	-80.303560
Oakland Park	3	2	1	0	0	36.980	26.172307	-80.131989
Pembroke Pines	3	2	0	1	0	50.950	26.007765	-80.296256
Plantation	13	8	4	1	0	42.860	26.127586	-80.233104
Pompano Beach	17	11	6	0	0	33.480	26.237860	-80.124767
Sunrise	27	16	5	6	0	43.510	26.166971	-80.256595
Tamarac	3	1	2	0	0	38.260	26.212861	-80.249771
Weston	9	6	3	0	0	47.830	26.100365	-80.399775

The results show that cities are segregated by those within each county Palm Beach in Red and Broward in green. Purple illustrates the third cluster, the cities with highest number of businesses.

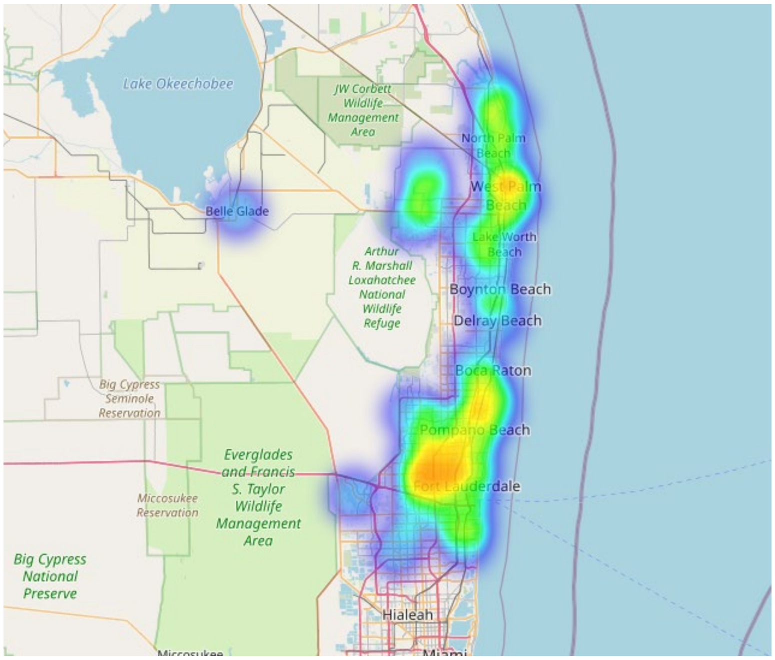


Below is a summary of the three clusters based on totals by city and by specific key industry. The High Density cluster accounts for largest share (49.3%) of total businesses with significant activity across three industries. The least active is the Clean Tech Sector.



Cluster Labels	Total	LifeScience	InfoTech	Financial	CleanTech
North Cluster	99	51	16	19	13
High Density Cluster	244	105	89	45	5
South Cluster	152	85	45	22	0

Another useful visualization is the heat map illustrating number of businesses by city. See map below. The closer color approaches red the more business in that area.


Lake Okeechobee
JW Corbett Wildlife Management Area
North Palm Beach
West Palm Beach
Belle Glade
Lake Worth Beach
Arthur R. Marshall Loxahatchee National Wildlife Refuge
Boynton Beach
Delray Beach
Boca Raton
Big Cypress Seminole Reservation
Pompano Beach
Everglades and Francis S. Taylor Wildlife Management Area
Fort Lauderdale
Miccosukee Reservation
Big Cypress National Preserve
Hialeah
Miccosukee
Miami

7. Conclusion / Recommendations

This study was an effort to understand the current economic climate of the Northern two-thirds of the Miami-Fort Lauderdale-West Palm Beach metropolitan area. With a focus on four key Industries: Information Technology, Life Sciences, Financial, and Clean Tech.

These industries have been identified as growth sectors for South Florida based on the selection criteria. 495 businesses were identified with a potential for data science services either through potential contracting or consulting. These 495 businesses span 38 cities and can be segregated into three groups (clusters). The groups are divided both geographically by North and South areas (Broward vs Palm Beach County respectively). In addition, a separate group was identified to address the higher concentration of economic activity. This group spans across North and South it includes three cities (Boca Raton, Ft Lauderdale, and West Palm Beach).

The results of this study indicate potential economic opportunities for a prospective entrepreneur to target these cities and business based on location and/or key industry.

Moreover, the data can further be analyzed if one desires to target one of specific industries.

8. References / Acknowledgement

Source US Census Bureau, 2020 November 24, Business Formation Statistics, Retrieved from https://www.census.gov/econ/bfs/index.html

Source Wikipedia, 2020 November 20, Miami metropolitan area, Retrieved from https://en.wikipedia.org/wiki/Miami_metropolitan_area

Source Enterprise Florida, 2020 November 11, Industries, Retrieved from https://www.enterpriseflorida.com/industries/

I would like to acknowledge the folks a LHH for access to DB Hoovers.

Note: report has hyperlinks to all cited material pointing to direct sources.