Coursera Capstone Project
Applied Data Science Specialization

R.Cox
rigdog@me.com
June 2020

Table of Contents

1. Introduction 3

1.1. Problem Description 3
1.2. Background 4
1.3. Target Audience 4

2. Data Section 5

3. Methodology 6

3.1. Machine Learning Model Selection 7

4. Results 9

5. Discussion 12

6. Conclusion 13

1. INTRODUCTION

This project seeks to showcase the knowledge and skills gained on my path to a Data Science Specialization with Coursera. I will attempt to solve a simulated real-world problem entirely with Python, employing various web scraping, data wrangling, data visualization and machine learning techniques. Within the code, I will utilize data returned through API calls to Foursquare to meet the requirements of this assignment.

All code will be written in Python, tested and published from a Jupyter Notebook within Watson Studio on IBM Cloud. The code will be made available on Github through the links provided at the end of this exercise.

1.1 PROBLEM DESCRIPTION

This project will attempt to solve to the problem of selecting the ideal number and locations for Warehouses or Distribution Centers. This question often emerges at some point in the growth of a business whether it is expending across a county, state or nation.

Picking the best locations for a distribution center can, among other things, reduce cost, improve efficiency, reduce lead time on fulfillment of orders, reduce length of delivery routes and therefore lower driver risk & insurance costs.

Conversely, too few or poorly located distribution centers can overload workers, increase overtime, reduce productivity, lengthen delivery routes, increase cost of delivery, increase delivery times and risking spoilage if the product is of a perishable nature.

PROBLEM SCENARIO:
A supplier of daily meals to High School Cafeterias across Orange County, Florida has just been awarded the contract to supply all schools within a specified area, defined within a 20mile radius of county center. To fulfill this contract, he will need to expand his food preparation and storage facilities as well as strategically position them to permit dependable and timely delivery considering the perishable nature of the product (food).

QUESTION**:**
How many preparation/distribution facilities should he rent and what are the best locations for these facilities?

1.2 BACKGROUND

A frequently encountered problem within the business world among medium to large businesses seeking to expand their distribution networks. The question of how many and where to setup distribution centers can greatly impact the bottom line of the business.

Typically, the question answers itself through trial and error with companies selecting locations within the general vicinity of its most frequent clients. Unfortunately, as they expand they could experience the ill effects of too few or poorly located distribution centers before a decision is made to take corrective action. The hope is that action is taken before customer are lost.

The simple answer to such a problem is one of picking central locations among groups of customers and can be broken down into the following broad steps;

i. Identify the delivery locations of all or most frequent customers
ii. Gather these customers into groups
iii. Finding the central location among these groups

With the variety of libraries and modules available within Python, this problem is apt for a programmatic solution through data science.

There are many approaches and solutions to this question. Many involved suggestions and discussions can be found through a simple web search. Some links are provided at the bottom of this page for reference. My approach focuses only on geographical data for the relatively small scale of the problem scenario.

1.3 TARGET AUDIENCE

This project and approach should be of interest to businesses seeking to expand storage, warehousing or distribution facilities to meet customer demands across large areas; cross-county, state or country.

This implementation, while simple, explores the use of unsupervised machine learning techniques to provide an answer to a common business question. This key limitation here is available data. The larger and more varied the collection of related factors, the greater the potential insights from leveraging data science techniques.

Link: Warehouse locations with k-means (Marton Trencseni)
Link: Selection of warehouse location for a global supply chain: A case study
(Rajesh Kr Singh Nikhil Chaudhary, [b]Nikhil Saxena)
Link: Choosing a Warehouse Location: 7 Critical Criteria to Consider
(CYZERG Warehouse Technonogy)

2. DATA SECTION

DATA WORKFLOW: Data Source – Data Wrangling - Feature Set for Analysis
This implementation, as described above was approached in three simple steps, the first of which involves determining the location of each school. This data constitutes the *Feature Set* for this implementation and is acquired through API calls within the Python code.

Foursquare popularized the concept of location sharing back in 2009 and today leverages the location data collected to serve approximately 150,000 partners through their developer tools. Due to its extensive database and accessibility, Foursquare is the location provider of choice for locating the geographical coordinates of the target schools. *Figure1* below shows a sample of the web based search results.

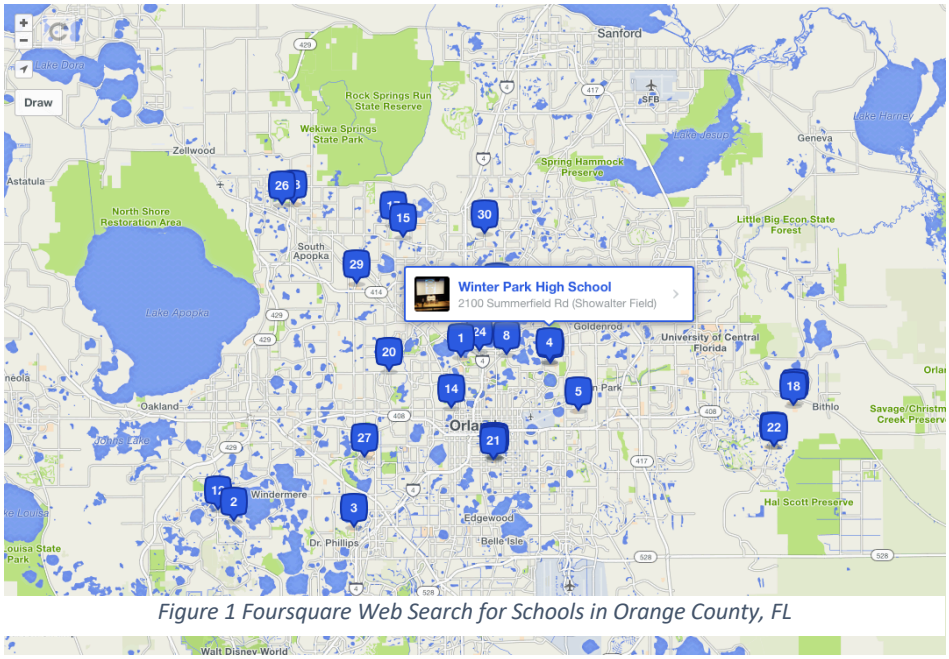

Figure 1 Foursquare Web Search for Schools in Orange County, FL

The API call returns the requested data in JavaScript Object Notation (JSON), from which the required data (e.g., names, addresses and geographical coordinates) needs to be extracted, filtered, corrected and formatted into a pandas DataFrame for analysis within Python. *Figure2* shows an example of JSON results.


```
{"meta":{"code":200,"requestId":"5edff55a7762772fe1d5295c
"},"response":{"venues":[{"id":"4b58cc3ff964a520b56b28e3"
,"name":"William R. Boone High
School","location":{"address":"2000 S Mills
Ave","crossStreet":"Kaley
St","lat":28.51948796222489,"lng":-
81.36511647712614,"labeledLatLngs":[{"label":"display","l
at":28.51948796222489,"lng":-
81.36511647712614}],"distance":2862,"postalCode":"32806",
"cc":"US","city":"Orlando","state":"FL","country":"United
States","formattedAddress":["2000 S Mills Ave (Kaley
St)","Orlando, FL 32806","United
States"]},"categories":[{"id":"4bf58dd8d48988d13d941735",
"name":"High School","pluralName":"High
```

Figure 2 Sample JSON results from Foursquare API Call showing target data

Following the data extraction and wrangling into an easily manipulated pandas DataFrame, the data will be initially explored through map visualizations due to the spatial nature of the problem. F*igure 3* below shows an example of the cleaned DataFrame from which the highlighted features will be selected for analysis.

	Name	Address	Zip	City	School Latitude	School Longitude	Category
0	William R. Boone High School	2000 S Mills Ave	32806	Orlando	28.519488	-81.365116	High School
1	Colonial High School	6100 Oleander Dr	32807	Orlando	28.554220	-81.302382	High School
2	Winter Park High School	2100 Summerfield Rd	32792	Winter Park	28.585042	-81.322753	High School
3	Lake Howell High School	4200 Dike Rd	32792	Winter Park	28.637595	-81.271969	High School
4	Oviedo High School	601 King St	32765	Oviedo	28.672018	-81.219145	High School
5	Dr. Phillips High School	6500 Turkey Lake Rd	32819	Orlando	28.470967	-81.476313	High School
6	Lake Nona High School	12500 Narcoossee Rd	32832	Orlando	28.380683	-81.245804	High School
7	Bishop Moore Catholic High School	3901 Edgewater Dr	32804	Orlando	28.587962	-81.392106	High School
8	Olympia High School	4301 S Apopka Vineland Rd	32835	Orlando	28.496692	-81.505520	High School
9	Osceola High School	420 S Thacker Ave	34741	Kissimmee	28.288172	-81.426683	High School

Figure 3 Cleaned data showing primary Feature Set of geographical coordinates

For initial visual exploration of the geographical area in question, a geojson file for the state of Florida was downloaded from the Official State of Florida Geographic Portal (http://geodata.myflorida.com).

From this point, the coordinates for each school will be fed into a machine learning model to perform unsupervised grouping of these schools into clusters. The choice of machine learning algorithm is discussed in the following Methodology section.

The data provided and insights gained from the results of this algorithm will again be analyzed through data visualizations to extract the ideal number of distribution centers and their precise locations.

The free online encyclopedia, Wikipedia (https://en.wikipedia.org) was also consulted during this project but only as a reference for general information on the county selected.

3. METHODOLOGY

The approach chosen to solve the presented problem required only geographical coordinates of each target school as the *Feature Set* for analysis and model input.

To retrieve the raw data set, an API GET request was made to Foursquare's search Endpoint of the Group venues. Data extraction and wrangling methods within Python were then carried out on the results prior to the initial analysis. *Figure 3* above shows a sample of the cleaned and formatted data.

Due to the spatial nature of the question presented, data visualization through geographical maps was utilized to explore the scope of the problem. All target schools within the specified area were superimposed on a map centered around Orange County, Florida as shows in *figure 4* below.

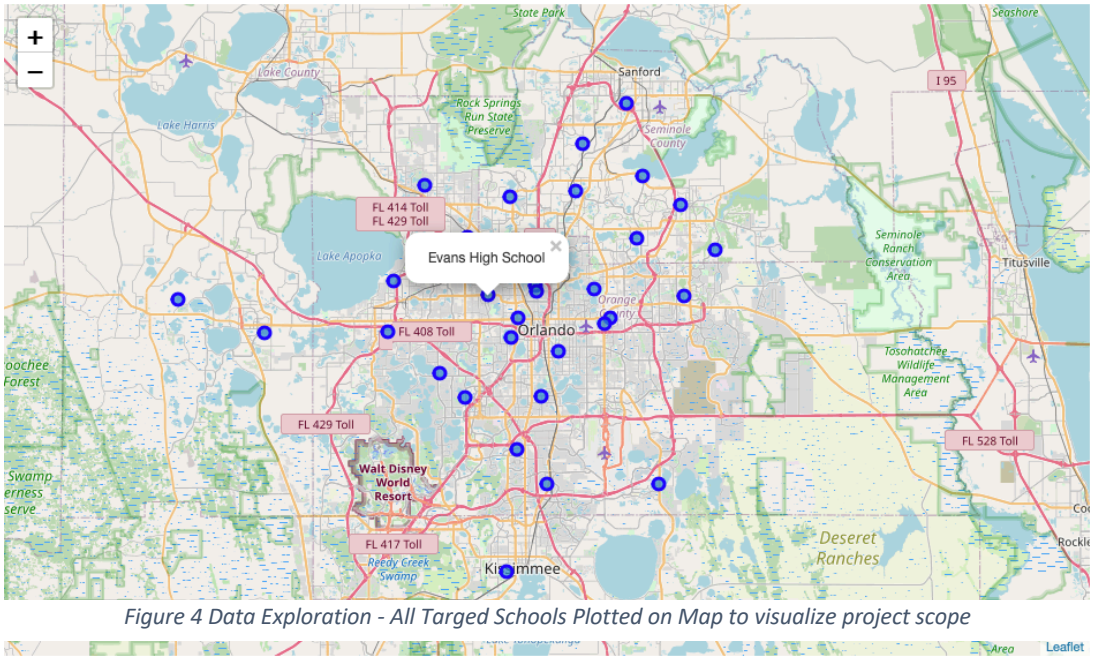

Figure 4 Data Exploration - All Targed Schools Plotted on Map to visualize project scope

3.1 MACHINE LEARNING MODEL SELECTION

Unsupervised Clustering was chosen as the most suitable approach to solving the problem of how many and where to position distribution/storage facilities. By specifying location as the *feature set*, the clustering algorithm will automatically group schools into distinct clusters based on location.

While several clustering algorithms are available, including DBSCAN which is especially suited to spatial data, the wide separation between locations, feature comparison between algorithms and code testing suggested that it was not the best candidate for this application.

K-Means is a partition based clustering algorithm that iteratively steps towards a solution by reducing its error at each iteration. It is designed to cluster observations in a way that minimizes the separation within clusters and maximizes the separation between clusters. This feature makes K-Means suitable for this application as the goal is to minimize distances from centrally located distribution centers.

Distance Metric

The separation or distance metric used internally by k-means is based on *Euclidian distance* calculations (straight line measurement between two points). While not equivalent to the actual distance along the earth's curved surface, for the 'relatively' small surface area concerned (20mile radius of Orange County's city center) this calculation method will suffice. It

will not provide actual distances between locations in this case but can be used as the separation metric with success.

The *Haversine formula* (implemented through the haversine package in Python) is used for calculating the actual distance between two points on the earth's curved surface using latitude and longitude and is included in this implementation for reporting purposes only.

Cluster Centers

In addition to the labels assigned to each observation (school) indicating the assigned cluster, the k-means algorithm also returns the locations of the cluster centers as the mean value of all observations within a cluster. By feeding the algorithm with latitudes and longitudes for each school, and not scaled or normalized values, the cluster centers returned by k-means will represent the geographical coordinates.

K-Value

Central to the application of the k-means algorithm is the required input of how many clusters to create, *K*. This directly translates to one of the questions presented by in the problem statement; "*How many preparation/distribution facilities should he rent?*" and is answered within the Python code through testing and analyses.

Different values for *K* are tested in the k-means model and the associated errors recorded. The error in this case is again a measure of distance or separation, where *Inertia* = Sum of Squared Distances from an observation to the cluster center. This data is then plotted and analyzed to determine the elbow point or point on the graph were the changes to *K* result in smaller changes to the error (point of diminishing returns). The *figure5* below shows an example of such a plot.

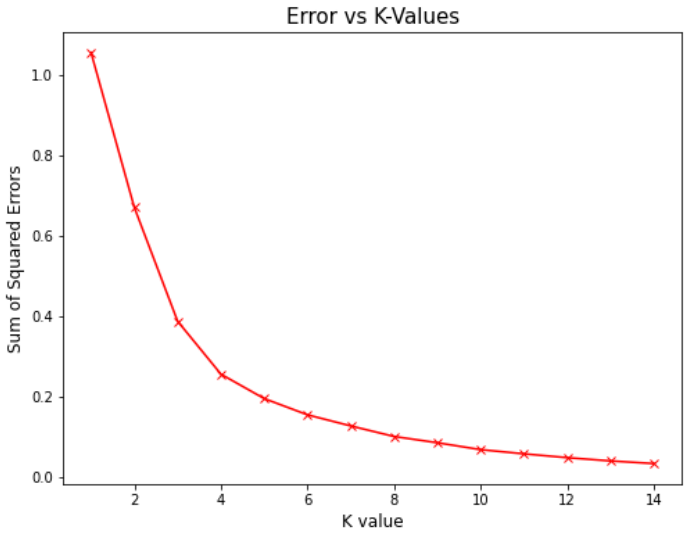

Figure 5 Plotting Errors vs K-Value to determine Optimal value for K

Armed with the optimal value for the number of clusters, and therefore the number of distribution centers, the definitive machine learning model is then fed the location data or *Feature Set* for final clustering and analysis. These results of which are discussed in the following section.

4. RESULTS

Testing the k-means algorithm on values of K ranging from 1 through 14 yielded the graph shown in *figure6* below

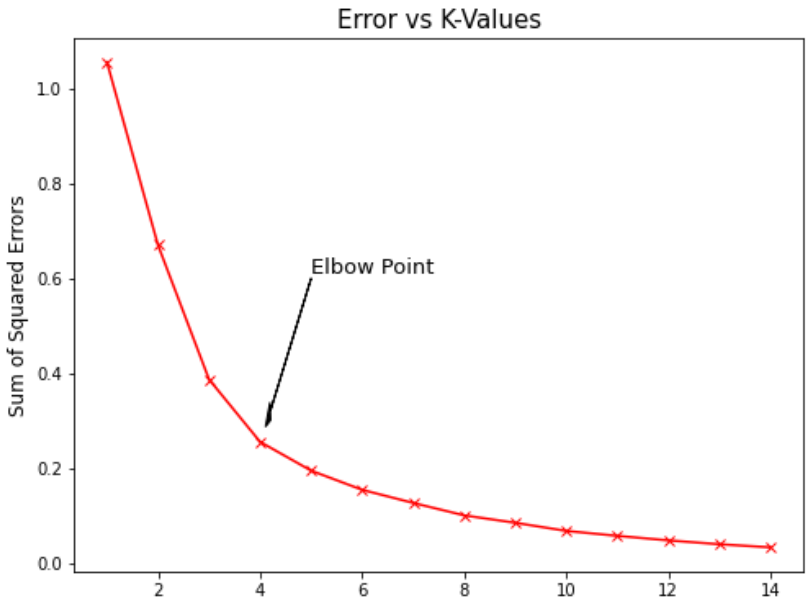

Figure 6 Errors vs K-Value with Elbow Point Indicated

The elbow point was chosen for an optimal value of K=4 (or 4 distribution centers). This is chosen from the above graph as the point beyond which the increase in K, the number of distribution centers, would yield increasingly smaller advantages.

A k-means model was then created in Python with the parameters below to group the schools into *4 distinct clusters*.

Kmeans = Kmeans(n_clusters=4, init='k-means++', n_init=12)

where;

n_clusters = 4 : *number of clusters to form and number of centroids to generate*
init = k-means++ : *selects initial cluster centers in order to speed up convergence*
n_init = 12 : *number of time the algorithm will be run with different centroid seeds*

Results in the form of an array of labels, *Kmeans.labels_*, one for each observation (school) with a value from 0-3 based on the assigned cluster is then added to the original DataFrame as column '*Cluster*'. See *Table 1* below.

The centroids locations, *Kmeans.cluter_centers_*, are returned as an array of four pairs of coordinates (latitude & longitude) which are then reverse geocoded to extract an address and again added to the original DataFrame as column '*Central Address'.* See *Table 1* below.

Physical distances from each school to the address of its cluster center are calculated using the haversine formula to report actual point-to-point, or radial distances in kilometers. This data is added to the final DataFrame as column '*Point to Point Distance'.* See *Table 1* below.

Name	Address	Zip	City	School Latitude	School Longitude	Cluster	Central Address	Point to Point Distance
William R. Boone High School	2000 S Mills Ave	32806	Orlando	28.51948796	-81.36511648	1	5177, Cortez Drive, Pine Hills.	9.689221737
Colonial High School	6100 Oleander Dr	32807	Orlando	28.55422021	-81.30238234	2	1009, Chatham Pines Circle, Lost Creek.	10.47906003
Winter Park High School	2100 Summerfield Rd	32792	Winter Park	28.58504228	-81.3227532	2	1009, Chatham Pines Circle, Lost Creek.	8.08985507
Lake Howell High School	4200 Dike Rd	32792	Winter Park	28.63759472	-81.27196914	2	1009, Chatham Pines Circle, Lost Creek.	1.050258184
Oviedo High School	601 King St	32765	Oviedo	28.67201812	-81.21914485	2	1009, Chatham Pines Circle, Lost Creek.	6.371283964
Dr. Phillips High School	6500 Turkey Lake Rd	32819	Orlando	28.47096658	-81.47631334	1	5177, Cortez Drive, Pine Hills.	10.39139604
Bishop Moore Catholic High School	3901 Edgewater Dr	32804	Orlando	28.58796177	-81.39210564	1	5177, Cortez Drive, Pine Hills.	6.507214104
Lake Nona High School	12500 Narcoossee Rd	32832	Orlando	28.38068325	-81.24580353	0	118, Green Cove Court, Osceola County.	10.99035377
Olympia High School	4301 S Apopka Vineland Rd	32835	Orlando	28.49669249	-81.50551951	1	5177, Cortez Drive, Pine Hills.	8.952156798
Osceola High School	420 S Thacker Ave	34741	Kissimmee	28.28817213	-81.42668309	0	118, Green Cove Court, Osceola County.	9.644762076
St Cloud High School	19th St	34769	Saint Cloud	28.24051685	-81.27950637	0	118, Green Cove Court, Osceola County.	13.10741213
Oak Ridge High School	700 W Oak Ridge Rd	32809	Orlando	28.47229023	-81.38592158	1	5177, Cortez Drive, Pine Hills.	11.86021111
Aloma Charter High School	495 N Semoran Blvd	32807	Orlando	28.547944	-81.310158	2	1009, Chatham Pines Circle, Lost Creek.	11.34998471
Seminole High School	2701 Ridgewood Ave	32773	Sanford	28.7781918	-81.28415108	2	1009, Chatham Pines Circle, Lost Creek.	14.73279976
Freedom High School	2500 W Taft Vineland Rd	32837	Orlando	28.41658052	-81.41355807	0	118, Green Cove Court, Osceola County.	10.49092738
University High School	11501 Eastwood Dr	32817	Orlando	28.57726325	-81.21547561	2	1009, Chatham Pines Circle, Lost Creek.	9.71988635
Edgewater High School	3100 Edgewater Dr	32804	Orlando	28.58180132	-81.39074755	1	5177, Cortez Drive, Pine Hills.	6.35665971
West Orange High School	1625 Beulah Rd	34787	Winter Garden	28.54043293	-81.56816625	1	5177, Cortez Drive, Pine Hills.	11.61673129
Lyman High School	865 S Ronald Reagan Blvd	32750	Longwood	28.68718329	-81.34441313	2	1009, Chatham Pines Circle, Lost Creek.	8.006671228
Lake Brantley High School	991 Sand Lake Rd	32714	Altamonte Springs	28.68109058	-81.42262134	1	5177, Cortez Drive, Pine Hills.	13.55577514
Cypress Creek High School	1101 Bear Crossing Dr	32824	Orlando	28.38079293	-81.37867936	0	118, Green Cove Court, Osceola County.	5.268585694
Hagerty High School	3225 Lockwood Blvd	32765	Oviedo	28.62519383	-81.17831604	2	1009, Chatham Pines Circle, Lost Creek.	9.922287254
Winter Springs High School	130 Tuskawilla Rd	32708	Winter Springs	28.70308251	-81.26473766	2	1009, Chatham Pines Circle, Lost Creek.	6.481095454

Lake Mary High School	655 Longwood Lake Mary Rd	32746	Lake Mary	28.73646169	-81.33599592	2	1009, Chatham Pines Circle, Lost Creek.	11.59352669
Orlando School High School	West Colonial, Ferrand Drive	32804	Orlando	28.554918	-81.41259	1	5177, Cortez Drive, Pine Hills.	3.90422388
Apopka High School	555 Martin St	32712	Apopka	28.69331389	-81.52351577	1	5177, Cortez Drive, Pine Hills.	16.20430073
Jones High School	Lake Mann Estates, Cottage Hill Road	32805	Orlando	28.53444854	-81.42083453	1	5177, Cortez Drive, Pine Hills.	4.29461363
South Lake High School	15600 Silver Eagle Rd	34736	Groveland	28.57425461	-81.81772076	3	674, W Minneola Ave, Clermont.	5.388628351
Evans High School	4949 Silver Star Rd	32808	Orlando	28.57867691	-81.44882843	1	5177, Cortez Drive, Pine Hills.	1.887938852
Legacy High School	1550 E Crown Point Rd	34761	Ocoee	28.59256288	-81.56070263	1	5177, Cortez Drive, Pine Hills.	11.17053305
East Ridge High School	13322 Excalibur Rd	34711	Clermont	28.53875859	-81.71503248	3	674, W Minneola Ave, Clermont.	5.38941501

Table 1 Final Python DatFrame showing Columns appended to original data including Cluster, Central Address & Distance

A visual representation of the results can be seen in *Figure 7* below. With the four clusters represented by different colors and the distribution centers identified by popup markers. Interactive Map link:
http://capstone-project.droppages.com (**Note**: linked map my vary due to time of execution and results returned. Please allow time to load)

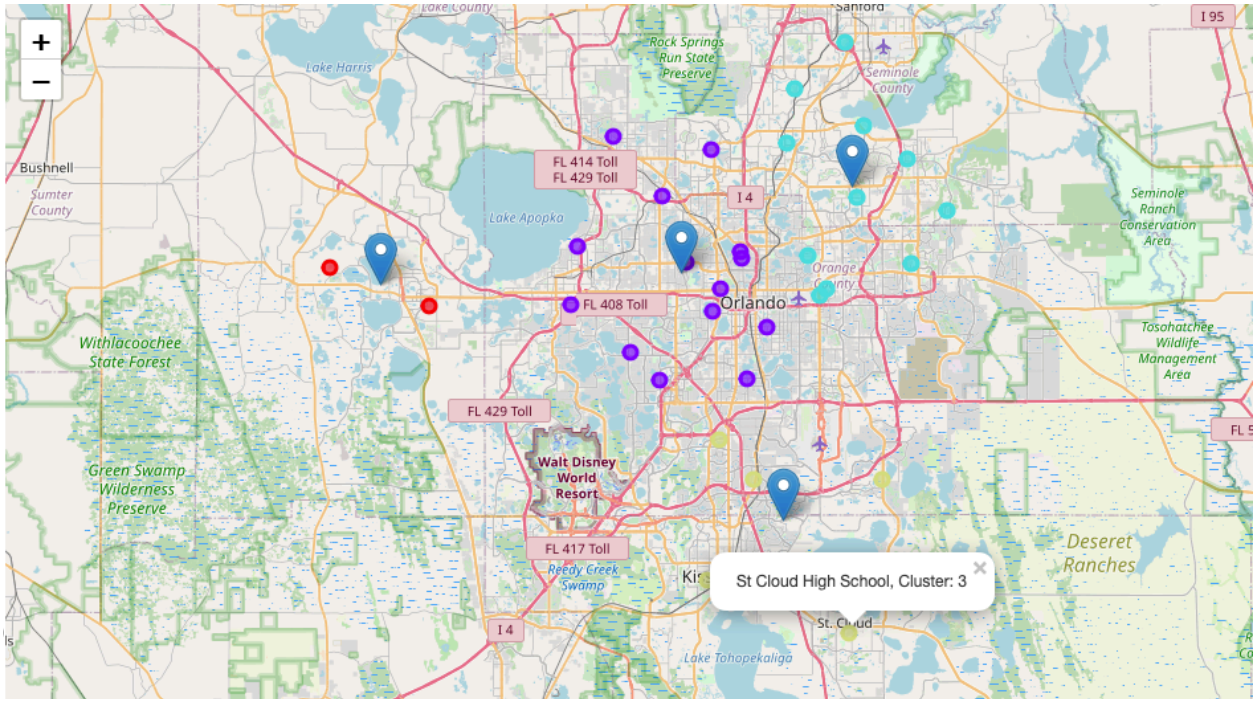

Figure 7 Results of K-Means clustering. Map showing School Locations and Clusters identified by different colors

Results are also summarized in the following format within Jupyter Notebook as well as sorted and exported as an excel workbook. Link: https://github.com/Richardhaydn/Coursera_Capstone/blob/master/Clustered%20School.xlsx (**Note**. Linked file may vary due to time executed and results returned)

Distribution Center located near: 118, Green Cove Court, Osceola County.
should serve the following **5 schools,**
at a maximum radial distance of **13.11 km.**

Name	Address	City
Lake Nona High School	12500 Narcoossee Rd	Orlando
Osceola High School	420 S Thacker Ave	Kissimmee
St Cloud High School	19th St	Saint Cloud
Freedom High School	2500 W Taft Vineland Rd	Orlando
Cypress Creek High School	1101 Bear Crossing Dr	Orlando

Distribution Center located near: 5177, Cortez Drive, Pine Hills.
should serve the following **13 schools,**
at a maximum radial distance of **16.20 km.**

Name	Address	City
William R. Boone High School	2000 S Mills Ave	Orlando
Dr. Phillips High School	6500 Turkey Lake Rd	Orlando
Bishop Moore Catholic High School	3901 Edgewater Dr	Orlando
Olympia High School	4301 S Apopka Vineland Rd	Orlando
Oak Ridge High School	700 W Oak Ridge Rd	Orlando
Edgewater High School	3100 Edgewater Dr	Orlando
West Orange High School	1625 Beulah Rd	Winter Garden
Lake Brantley High School	991 Sand Lake Rd	Altamonte Springs
Orlando School High School	West Colonial, Ferrand Drive	Orlando
Apopka High School	555 Martin St	Apopka
Jones High School	Lake Mann Estates, Cottage Hill Road	Orlando
Evans High School	4949 Silver Star Rd	Orlando
Legacy High School	1550 E Crown Point Rd	Ocoee

Distribution Center located near: 1009, Chatham Pines Circle, Lost Creek.
should serve the following **11 schools,**
at a maximum radial distance of **14.73 km.**

Name	Address	City
Colonial High School	6100 Oleander Dr	Orlando
Winter Park High School	2100 Summerfield Rd	Winter Park
Lake Howell High School	4200 Dike Rd	Winter Park
Oviedo High School	601 King St	Oviedo
Aloma Charter High School	495 N Semoran Blvd	Orlando
Seminole High School	2701 Ridgewood Ave	Sanford
University High School	11501 Eastwood Dr	Orlando
Lyman High School	865 S Ronald Reagan Blvd	Longwood
Hagerty High School	3225 Lockwood Blvd	Oviedo
Winter Springs High School	130 Tuskawilla Rd	Winter Springs
Lake Mary High School	655 Longwood Lake Mary Rd	Lake Mary

Distribution Center located near: 674, W Minneola Ave, Clermont.
should serve the following **2 schools,**
at a maximum radial distance of **5.39 km.**

Name	Address	City
South Lake High School	15600 Silver Eagle Rd	Groveland
East Ridge High School	13322 Excalibur Rd	Clermont

5. DISCUSSION

The application of k-means to this problem yields the desired results and answers the questions presented by the problem scenario; *'How many preparation/distribution facilities should he rent and what are the best locations for these facilities?'*

The simple two-dimensional feature set of latitude and longitude fits well with the algorithm to provide clustering and deliver central locations based on minimizing intra-cluster separation

while maximizing inter-cluster separation. This methodology translates well to this problem but it is not without limitations.

This implementation is based on the follow main assumptions;

I. Foursquare is able to locate all High Schools within the given area, i.e., all High Schools are listed in their database
II. The best location for distribution centers is within the geographical center of each cluster.
III. k-means will always converge to a local optimum which may not be the same as the global optimum.
IV. Euclidian distance computed on latitude & longitude coordinates can adequately serve as a separation metric used for minimizing errors. This is discussed in a previous section.

It also does not take into account factors that may also influence the location of distribution centers such as;

I. Volume to be served or number of schools served per distribution center
II. Traffic conditions in locations around proposed distribution centers
III. Access to roads, highways, routes in the immediate vicinity of the proposed centers

Working with the assumptions above, the machine learning algorithm can be used to provide answers to the business questions presented. In the event that the proposed location is either inaccessible or needs to be reconsidered due to one of the above considerations, the original proposal can still serve a region of interest for finding available options nearby.

The ability to limit the number of distribution centers, *K* in k-means, due to possible financial constraints for example, can be considered an advantage of this algorithm.

This clustering based on geographical coordinates can serve as an initial step and be further developed by the examination of individual clusters to find further similarities or apply weightings to locations based on the above mentioned factors. This approach may be useful for managing deliveries or workload.

6. CONCLUSION

Machine Learning has shown in this example to provide answers to the business question presented in the Introduction. In particular, *K-Means Clustering* has proven effective in the case of determining the optimal number of distribution centers and their locations given the coordinates of all targets.

The small-scale implementation in this example can quickly and easily be scaled to state-wide or nation-wide distributions provided that the required location of all targets is available. Such a large distribution, if considered with additional features would undoubtedly benefit from the speed and application of machine learning and a data science approach.