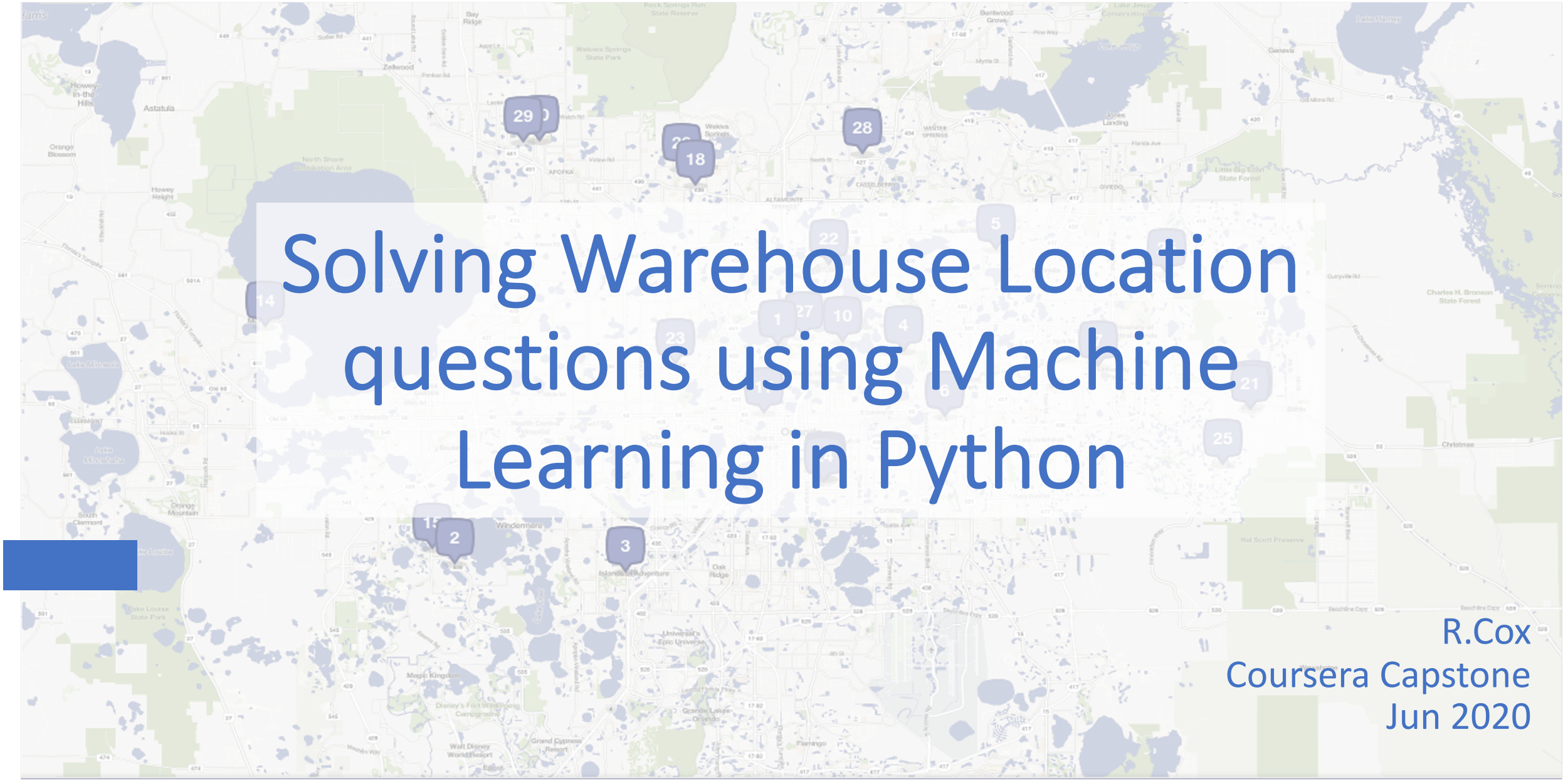


# Solving Warehouse Location questions using Machine Learning in Python

R.Cox

Coursera Capstone

Jun 2020

# Problem Description

As a supplier of daily meals to High Schools in Central Florida, you have been awarded the contract to supply all schools within a 20mile radius of Orange County city center. To meet this increased demand you will need to expand your preparation, storage and distribution facilities.

The questions to be answered then become:

***How many distribution facilities do you need?***

***Where should these distribution facilities be located?***

# Problem Approach

The approach to this problem is broken down into the three broad steps below;


i. **Explore & Extract**: Find & extract location data for schools to create Feature Set

ii. **Group Observations**: Group these schools into clusters using Machine Learning

iii. **Analyze Results**: Analyze groups to uncover insights and answer questions

# Data Workflow

Explore & Extract

Data for all schools in this assignment was gathered through a single API GET request to Foursquare's *Search* Endpoint within the *Venues* Group using the parameters shown:

```
search_query='High School'
        LIMIT=200
        RADIUS=30000
        URL =
'https://api.foursquare.com/v2/venues/search?client_id={}&client_secret={}&ll={},{}&v={}&query=
                {}&radius{}&limit={}'.format(CLIENT_ID, CLIENT_SECRET, latitude, longitude, VERSION,
search_query,
                RADIUS, LIMIT)
        results = requests.get(URL).json()['response']['venues']
```

The results of which are shown here with target features emphasized: (*sample only*)


[{'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', 'lat': 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 Schools', 'shortName': 'High School', 'icon': {'prefix

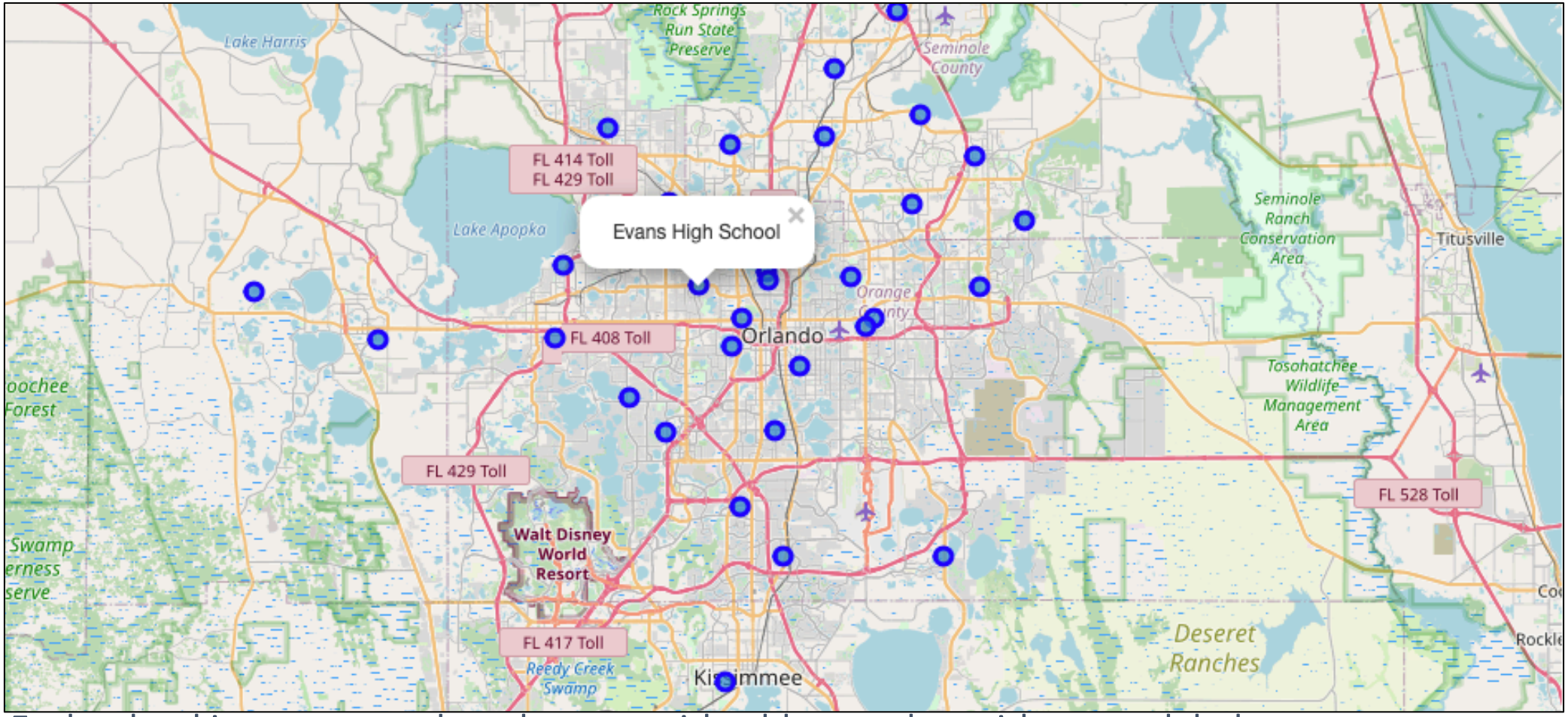
Explore & Extract

Data wrangling, cleaning and formatting resulted in the DataFrame below, reducing the data set from 50 samples to 32 target schools. The intended *Feature Set* is highlighted below. (*sample only shown*)

| | 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 | Oviedo High School | 601 King St | 32765 | Oviedo | 28.672018 | -81.219145 | High School |
| 4 | Lake Howell High School | 4200 Dike Rd | 32792 | Winter Park | 28.637595 | -81.271969 | High School |
| 5 | Dr. Phillips High School | 6500 Turkey Lake Rd | 32819 | Orlando | 28.470967 | -81.476313 | High School |
| 6 | Bishop Moore Catholic High School | 3901 Edgewater Dr | 32804 | Orlando | 28.587962 | -81.392106 | High School |
| 7 | Lake Nona High School | 12500 Narcoossee Rd | 32832 | Orlando | 28.380683 | -81.245804 | High School |
| 8 | Olympia High School | 4301 S Apopka Vineland Rd | 32835 | Orlando | 28.496692 | -81.505520 | High School |
| 9 | Oak Ridge High School | 700 W Oak Ridge Rd | 32809 | Orlando | 28.472290 | -81.385922 | High School |
| 10 | St Cloud High School | 19th St | 34769 | Saint Cloud | 28.240517 | -81.279506 | High School |

Explore & Extract

Data exploration through visualizations revealed the true scope of the problem showing the spatial distribution of schools.




Each school is represented on the map with a blue marker with popup label

Group Locations

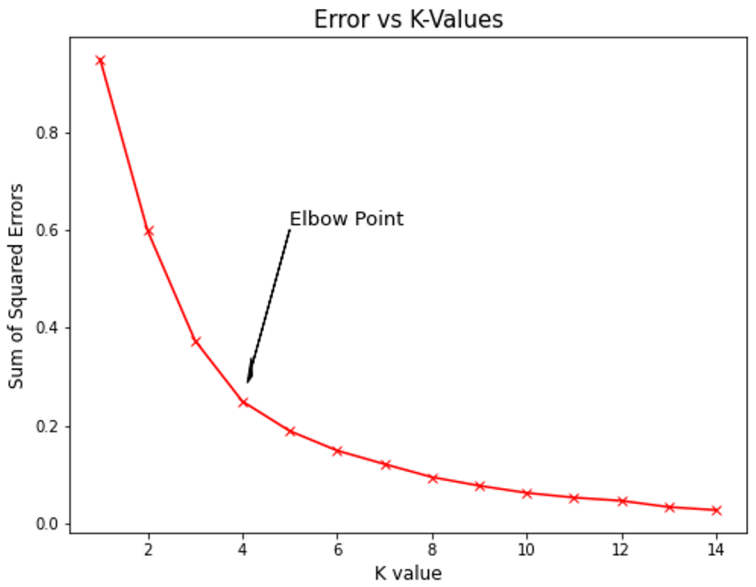
**K-Means** is a partition based unsupervised clustering algorithm that produces distinct non-overlapping clusters. The algorithm groups observations in a way that minimizes the error within a cluster while maximizing the error between different clusters where error is measured as the distance between observations and their centroids. The process converges on a solution by iteratively reducing the error by adjusting the centroid locations.

## Determining Number of Clusters

The optimal number of clusters is determined through analysis of the plot shown below. The elbow point is chosen at **K= 4** as the point where the further increasing the value of K yields increasing smaller reductions in error.
***This represents the optimal number of distribution centers***.

The Machine Learning K-Means model is then built and fit to the *Feature Set* with the following lines of code in Python:

```
kmeans = KMeans(n_clusters=4, init='k-means++', n_init=12)
kmeans.fit(schools_df.iloc[:,4:6].values)
```



Analyze Results

## K-means produces the following results which are then incorporated into the final DataFrame shown

**kmeans.labels_**
array([0, 1, 1, 1, 1, 0, 0, 3, 0, 0, 3, 3, 1, 1, 3, 0, 1, 0, 3, 0, 1, 1, 1, 1, 0, 2, 0, 0, 0, 0, 2, 0], dtype=int32)

**kmeans.cluster_centers_**
array([[ 28.56736717, -81.45334576],
[ 28.64583597, -81.277227 ],
[ 28.5565066 , -81.76637662],
[ 28.34134914, -81.34884609]])

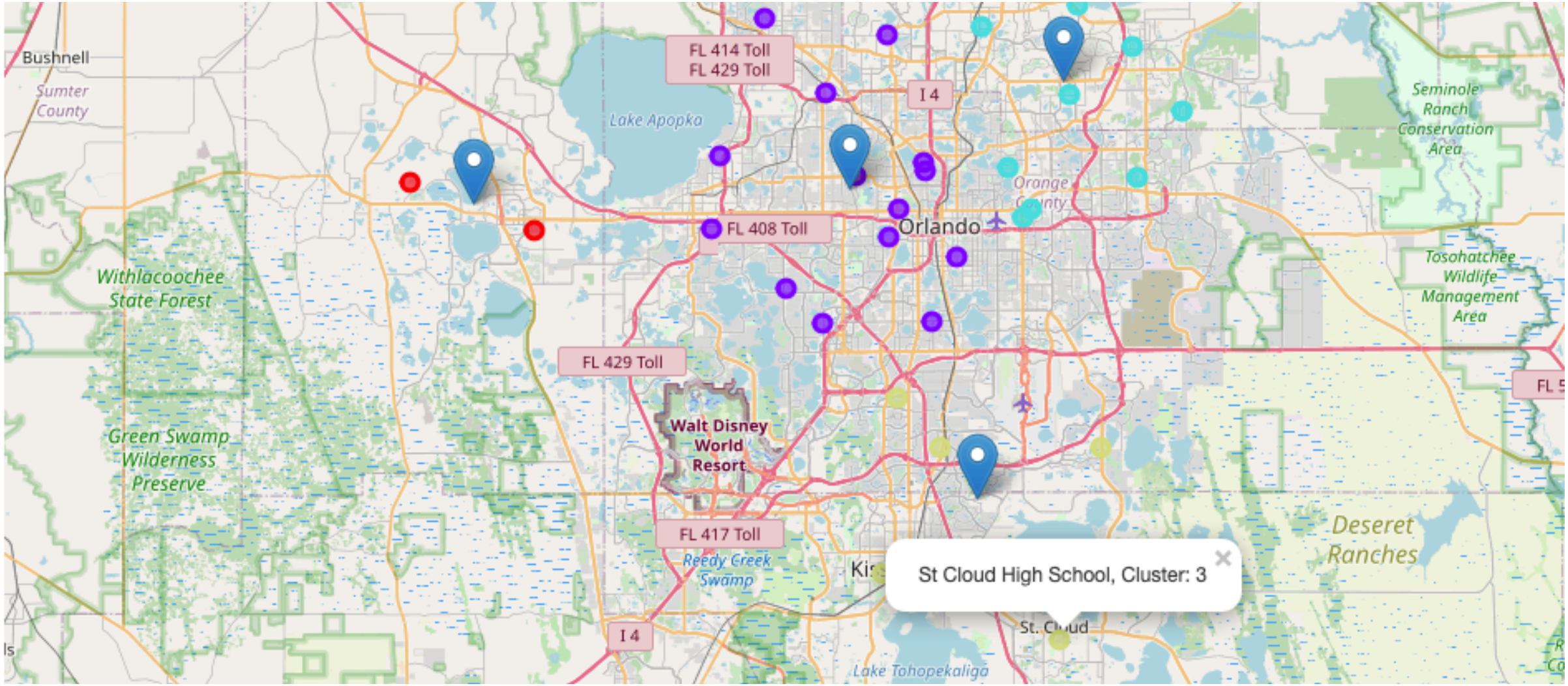
*where,*
**labels_** represents the assigned cluster for each school

**cluster_centers_** represents the centroid locations for each cluster (latitude & longitude)

| | Name | Address | Zip | City | School Latitude | School Longitude | Category | Cluster | Central Address |
|---|---|---|---|---|---|---|---|---|---|
| 0 | William R. Boone High School | 2000 S Mills Ave | 32806 | Orlando | 28.519488 | -81.365116 | High School | 0 | 5340, Ferdinand Drive, Pine Hills. |
| 1 | Colonial High School | 6100 Oleander Dr | 32807 | Orlando | 28.554220 | -81.302382 | High School | 1 | 1009, Chatham Pines Circle, Lost Creek. |
| 2 | Winter Park High School | 2100 Summerfield Rd | 32792 | Winter Park | 28.585042 | -81.322753 | High School | 1 | 1009, Chatham Pines Circle, Lost Creek. |
| 3 | Oviedo High School | 601 King St | 32765 | Oviedo | 28.672018 | -81.219145 | High School | 1 | 1009, Chatham Pines Circle, Lost Creek. |
| 4 | Lake Howell High School | 4200 Dike Rd | 32792 | Winter Park | 28.637595 | -81.271969 | High School | 1 | 1009, Chatham Pines Circle, Lost Creek. |
| 5 | Dr. Phillips High School | 6500 Turkey Lake Rd | 32819 | Orlando | 28.470967 | -81.476313 | High School | 0 | 5340, Ferdinand Drive, Pine Hills. |
| 6 | Bishop Moore Catholic High School | 3901 Edgewater Dr | 32804 | Orlando | 28.587962 | -81.392106 | High School | 0 | 5340, Ferdinand Drive, Pine Hills. |

Analyze Results

# Visualizing Results: Each school is now color coded according to cluster with cluster centers marked



# Results

A summary of results is presented in the following format for each school within Python. (*sample only*)

**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 |

# Results are also exported to spreadsheet format as shown here (sorted by cluster & distance)

| Name | Address | Zip | City | School Latitude | School Longitude | Category | Cluster | Central Address | Point to Point Distance |
|---|---|---|---|---|---|---|---|---|---|
| Cypress Creek High School | 1101 Bear Crossing Dr | 32824 | Orlando | 28.38079293 | -81.37867936 | High School | 0 | 118, Green Cove Court, Osceola County. | 5.268585694 |
| Osceola High School | 420 S Thacker Ave | 34741 | Kissimmee | 28.28817213 | -81.42668309 | High School | 0 | 118, Green Cove Court, Osceola County. | 9.644762076 |
| Freedom High School | 2500 W Taft Vineland Rd | 32837 | Orlando | 28.41658052 | -81.41355807 | High School | 0 | 118, Green Cove Court, Osceola County. | 10.49092738 |
| Lake Nona High School | 12500 Narcoossee Rd | 32832 | Orlando | 28.38068325 | -81.24580353 | High School | 0 | 118, Green Cove Court, Osceola County. | 10.99035377 |
| St Cloud High School | 19th St | 34769 | Saint Cloud | 28.24051685 | -81.27950637 | High School | 0 | 118, Green Cove Court, Osceola County. | 13.10741213 |
| Evans High School | 4949 Silver Star Rd | 32808 | Orlando | 28.57867691 | -81.44882843 | High School | 1 | 5177, Cortez Drive, Pine Hills. | 1.887938852 |
| Orlando School High School | West Colonial, Ferrand Drive | 32804 | Orlando | 28.554918 | -81.41259 | High School | 1 | 5177, Cortez Drive, Pine Hills. | 3.90422388 |
| Jones High School | Lake Mann Estates, Cottage Hill Road | 32805 | Orlando | 28.53444854 | -81.42083453 | High School | 1 | 5177, Cortez Drive, Pine Hills. | 4.29461363 |
| Edgewater High School | 3100 Edgewater Dr | 32804 | Orlando | 28.58180132 | -81.39074755 | High School | 1 | 5177, Cortez Drive, Pine Hills. | 6.35665971 |
| Bishop Moore Catholic High School | 3901 Edgewater Dr | 32804 | Orlando | 28.58796177 | -81.39210564 | High School | 1 | 5177, Cortez Drive, Pine Hills. | 6.507214104 |
| Olympia High School | 4301 S Apopka Vineland Rd | 32835 | Orlando | 28.49669249 | -81.50551951 | High School | 1 | 5177, Cortez Drive, Pine Hills. | 8.952156798 |
| William R. Boone High School | 2000 S Mills Ave | 32806 | Orlando | 28.51948796 | -81.36511648 | High School | 1 | 5177, Cortez Drive, Pine Hills. | 9.689221737 |
| Dr. Phillips High School | 6500 Turkey Lake Rd | 32819 | Orlando | 28.47096658 | -81.47631334 | High School | 1 | 5177, Cortez Drive, Pine Hills. | 10.39139604 |
| Legacy High School | 1550 E Crown Point Rd | 34761 | Ocoee | 28.59256288 | -81.56070263 | High School | 1 | 5177, Cortez Drive, Pine Hills. | 11.17053305 |
| West Orange High School | 1625 Beulah Rd | 34787 | Winter Garden | 28.54043293 | -81.56816625 | High School | 1 | 5177, Cortez Drive, Pine Hills. | 11.61673129 |
| Oak Ridge High School | 700 W Oak Ridge Rd | 32809 | Orlando | 28.47229023 | -81.38592158 | High School | 1 | 5177, Cortez Drive, Pine Hills. | 11.86021111 |
| Lake Brantley High School | 991 Sand Lake Rd | 32714 | Altamonte Springs | 28.68109058 | -81.42262134 | High School | 1 | 5177, Cortez Drive, Pine Hills. | 13.55577514 |
| Apopka High School | 555 Martin St | 32712 | Apopka | 28.69331389 | -81.52351577 | High School | 1 | 5177, Cortez Drive, Pine Hills. | 16.20430073 |
| Lake Howell High School | 4200 Dike Rd | 32792 | Winter Park | 28.63759472 | -81.27196914 | High School | 2 | 1009, Chatham Pines Circle, Lost Creek. | 1.050258184 |
| Oviedo High School | 601 King St | 32765 | Oviedo | 28.67201812 | -81.21914485 | High School | 2 | 1009, Chatham Pines Circle, Lost Creek. | 6.371283964 |
| Winter Springs High School | 130 Tuskawilla Rd | 32708 | Winter Springs | 28.70308251 | -81.26473766 | High School | 2 | 1009, Chatham Pines Circle, Lost Creek. | 6.481095454 |
| Lyman High School | 865 S Ronald Reagan Blvd | 32750 | Longwood | 28.68718329 | -81.34441313 | High School | 2 | 1009, Chatham Pines Circle, Lost Creek. | 8.006671228 |
| Winter Park High School | 2100 Summerfield Rd | 32792 | Winter Park | 28.58504228 | -81.3227532 | High School | 2 | 1009, Chatham Pines Circle, Lost Creek. | 8.08985507 |
| University High School | 11501 Eastwood Dr | 32817 | Orlando | 28.57726325 | -81.21547561 | High School | 2 | 1009, Chatham Pines Circle, Lost Creek. | 9.71988635 |
| Hagerty High School | 3225 Lockwood Blvd | 32765 | Oviedo | 28.62519383 | -81.17831604 | High School | 2 | 1009, Chatham Pines Circle, Lost Creek. | 9.922287254 |
| Colonial High School | 6100 Oleander Dr | 32807 | Orlando | 28.55422021 | -81.30238234 | High School | 2 | 1009, Chatham Pines Circle, Lost Creek. | 10.47906003 |
| Aloma Charter High School | 495 N Semoran Blvd | 32807 | Orlando | 28.547944 | -81.310158 | High School | 2 | 1009, Chatham Pines Circle, Lost Creek. | 11.34998471 |
| Lake Mary High School | 655 Longwood Lake Mary Rd | 32746 | Lake Mary | 28.73646169 | -81.33599592 | High School | 2 | 1009, Chatham Pines Circle, Lost Creek. | 11.59352669 |
| Seminole High School | 2701 Ridgewood Ave | 32773 | Sanford | 28.7781918 | -81.28415108 | High School | 2 | 1009, Chatham Pines Circle, Lost Creek. | 14.73279976 |
| South Lake High School | 15600 Silver Eagle Rd | 34736 | Groveland | 28.57425461 | -81.81772076 | High School | 3 | 674, W Minneola Ave, Clermont. | 5.388628351 |
| East Ridge High School | 13322 Excalibur Rd | 34711 | Clermont | 28.53875859 | -81.71503248 | High School | 3 | 674, W Minneola Ave, Clermont. | 5.38941501 |

# Conclusion

The application of data science and machine learning, in particular ***K-Means Clustering,*** has shown in this example to be an effective method for determining the optimal number of distribution centers and their locations. It was successfully implemented to answer the questions presented in the problem description.

The simplicity of this approach and its scalability will be apparent when applied to larger datasets, for example a distributor expanding across an entire state.

This implementation is based on a few assumptions and is not without limitations. Please refer to the full report for more in-depth discussions on these issues.

*Report:*
https://github.com/Richardhaydn/Coursera_Capstone/blob/master/Capstone%20Project_Full%20Report.pdf

*Notebook:*
https://github.com/Richardhaydn/Coursera_Capstone/blob/master/Coursera%20Capstone_Applied%20Data%20Science_Project.ipynb

*Clustered Map:*
http://capstone-project.droppages.com