Applied Data Science Capstone Project Report:
The Battle of Neighborhoods: College Campuses

Submitted By: Shiv Vyas

Goal

The goal of this project is to help students/parents to select a university/campus for college/schools especially international students who are totally new to the country and do not have any major idea about the locality of the campus.
Students can select the College/School/ Campus and find out the top & places nearest places to the campus area such as shops, restaurants, hospitals which can help to decide the college/campus.

This Problem & Use Cases

The goal of this project is to help students/parents to select a university/campus for college/schools especially international students who are totally new to the country and do not have any major idea about the locality of the campus.

I am an international student, and it was really difficult for me to select the campus without having any contact in the United States.

Although there is some information available about neighborhoods on the college website, this information is limited and not exactly in the way we want it.

University, Students, etc anyone can use this program to find out everything about the neighborhood not only around the campus but also any other residential places too.

Data

The main data is obtained from the website: https://ope.ed.gov/dapip/#/home
The data has 32833 rows and 14 columns:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32833 entries, 0 to 32832
Data columns (total 14 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   DapipId         32833 non-null  int64  
 1   OpeId           10092 non-null  object 
 2   IpedsUnitIds    9388 non-null   object 
 3   LocationName    32833 non-null  object 
 4   ParentName      32833 non-null  object 
 5   ParentDapipId   32833 non-null  object 
 6   LocationType    32833 non-null  object 
 7   Address         32833 non-null  object 
 8   GeneralPhone    7273 non-null   object 
 9   AdminName       6109 non-null   object 
 10  AdminPhone      5940 non-null   object 
 11  AdminEmail      4382 non-null   object 
 12  Fax             2316 non-null   float64
 13  UpdateDate      6737 non-null   object 
dtypes: float64(1), int64(1), object(12)
memory usage: 3.5+ MB
```

The data is shown below:

	DapipId	OpeId	IpedsUnitIds	LocationName	ParentName	ParentDapipId	LocationType	Address	GeneralPhone	AdminName	AdminPhone	AdminEmail	Fax	UpdateDa
0	100016	01230800	100636	Community College of the Air Force	-	-	Institution	130 W Maxwell Blvd, Montgomery, AL 36112-6613	3349536436	ERIC A. ASH	3349536436	eric.ash@maxwell.af.mil	3.349538e+09	N
1	100016002	NaN	NaN	Community College of the Air Force	Community College of the Air Force	100016	Additional Location	2250 Stanley Road, Unit 161, Fort Sam Houston,...	NaN	NaN	NaN	NaN	NaN	N
2	100025	00100200	100654	Alabama A & M University	-	-	Institution	4900 Meridian Street, Normal, AL 35762	2563725000	Andrew Hugine	2563725230	andrew.hugine@aamu.edu	NaN	9/14/2(12:00:00
3	100034	00105200	100663	University of Alabama at Birmingham	-	-	Institution	1720 2nd Avenue South, Birmingham, AL 35233	2059344011	Dr. Ray L. Watts	2059344636	rlawtts@uab.edu	NaN	3/12/2(12:00:00
4	100034002	NaN	NaN	UAB School of Optometry	University of Alabama at Birmingham	100034	Additional Location	1716 University Boulevard, Birmingham, AL 35294	NaN	NaN	NaN	NaN	NaN	N

EDA | Data Pre-Processing | Data Visualization

The main focus in this section is to obtain the Postal Code from the Address Column and Find the latitude and longitude of each institute.
Select the required features only, remove the unwanted features.

	LocationType	LocationName	ParentName	Address
0	Institution	Community College of the Air Force	-	130 W Maxwell Blvd, Montgomery, AL 36112-6613
1	Additional Location	Community College of the Air Force	Community College of the Air Force	2250 Stanley Road, Unit 161, Fort Sam Houston,...
2	Institution	Alabama A & M University	-	4900 Meridian Street, Normal, AL 35762
3	Institution	University of Alabama at Birmingham	-	1720 2nd Avenue South, Birmingham, AL 35233
4	Additional Location	UAB School of Optometry	University of Alabama at Birmingham	1716 University Boulevard, Birmingham, AL 35294

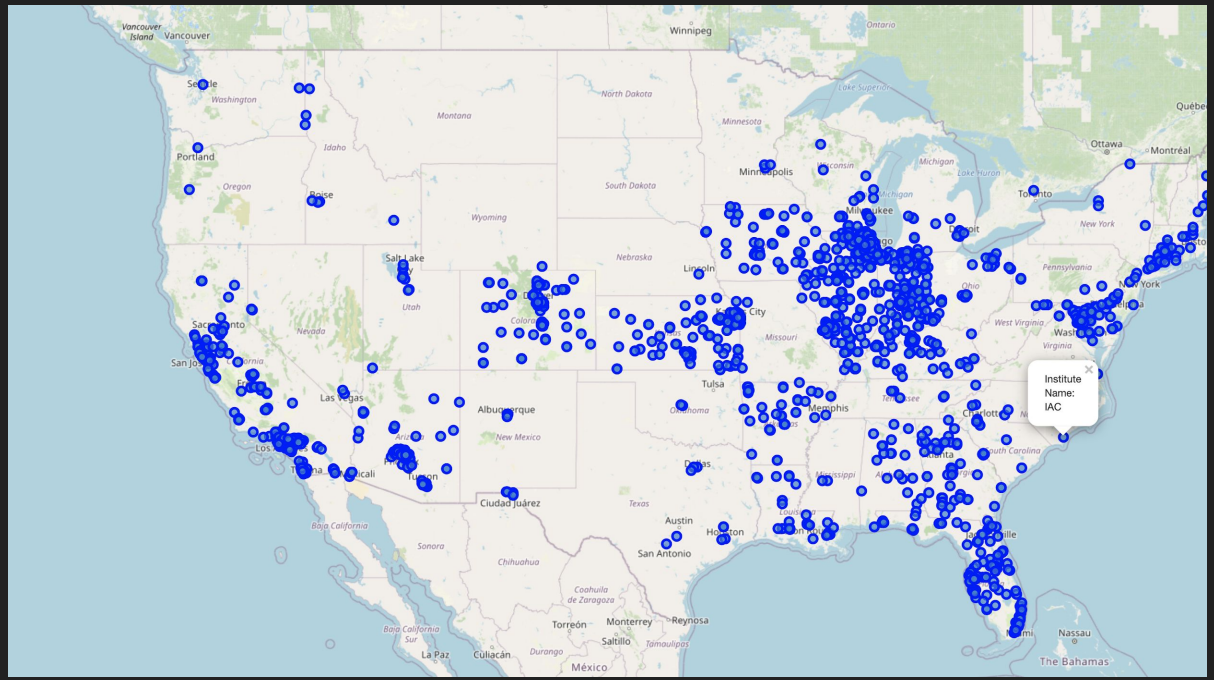
Now, the remaining Postal codes were obtained from the Address column using the regex as shown below. Although some addresses were not proper and were forced to be removed.

```
post_code_2 = r'\d{5}'

PC_List = []
to_remove = []
for x in range(len(df_new)):
        PC = re.search(post_code_2,str(df_new["Address"][x]))
        if ( PC != None):
            PC_List.append(PC.group())
        else:
            to_remove.append(int(df_new['Seq'][x]))
```

Visualization

Geopy and Folium were used to find the latitude and the longitude of the institution and plot them on the map of the USA. The map is shown below.



FourSquare

Foursquare api was used to find the neighbours of the university.
Foursquare returned a json file which was processed to find the top “N” nearest venues.

```
results

[{'reasons': {'count': 0,
    'items': [{'reasonName': 'globalInteractionReason',
      'summary': 'This spot is popular',
      'type': 'general'}]},
  'referralId': 'e-0-4c597cd47b049521a4e9881f-0',
  'venue': {'categories': [{'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/parks_outdoor
      'suffix': '.png'},
     'id': '4bf58dd8d48988d163941735',
     'name': 'Park',
     'pluralName': 'Parks',
     'primary': True,
     'shortName': 'Park'}],
   'id': '4c597cd47b049521a4e9881f',
   'location': {'address': '1600 1st Ave S',
    'cc': 'US',
    'city': 'Birmingham',
    'country': 'United States',
    'crossStreet': 'btw 14th & 18th St S',
    'distance': 100,
    'formattedAddress': ['1600 1st Ave S (btw 14th & 18th St S)',
     'Birmingham, AL 35233',
     'United States'],
    'labeledLatLngs': [{'label': 'display',
      'lat': 33.509764495688025,
      'lng': -86.80793066227511}],
    'lat': 33.509764495688025,
```

After taking important features into account from results, the following data frame was created.

	Postal Code	Location Name	Location Type	Loc_Latitude	Loc_Longitude	Venue Name	V_Latitude	V_Longitude	V_Category
0	20912	Washington Adventist Hospital	Site	33.509110	-86.807192	Railroad Park	33.509764	-86.807931	Park
1	20912	Washington Adventist Hospital	Site	33.509110	-86.807192	Red Cat	33.509741	-86.807950	Café
2	20912	Washington Adventist Hospital	Site	33.509110	-86.807192	Hilton Garden Inn	33.508873	-86.806958	Hotel
3	20912	Washington Adventist Hospital	Site	33.509110	-86.807192	Planet Smoothie	33.509726	-86.807465	Smoothie Shop
4	94708	Dominican School of Philosophy and Theology	Institution	34.072432	-86.778004	Warehouse Discount Groceries	34.072573	-86.777053	Grocery Store

Analysis : Top Nearby Venues

The data frame obtained in (5) was processed with foursquare api to obtain the top venues.
The result can be seen below:

	Postal Code	Location Name	Venue Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	2135	Clinical Pastoral Education Program at Caritas...	Corner Coffee at ACC	Coffee Shop	Theater	Gym	Grocery Store	Football Stadium
1	2148	New England Hair Academy	Sagan's Space	Playground	Yoga Studio	Coffee Shop	Grocery Store	Football Stadium
2	8534	College of New Jersey, The - Capital Health Me...	Pence Theatre	Theater	Yoga Studio	Gym	Grocery Store	Football Stadium
3	8629	Gwynedd Mercy College at St. Francis Medical C...	Tree House Studios	Park	Yoga Studio	Health & Beauty Service	Grocery Store	Football Stadium
4	10946	ICR Graduate School	Ramiro's Mexican Food	Mexican Restaurant	Yoga Studio	Health & Beauty Service	Grocery Store	Football Stadium
5	14200	Trinity International University at Christ Com...	The Grille	American Restaurant	Yoga Studio	College Stadium	Gym	Grocery Store
6	14200	Trinity International University at Christ Com...	Tine Davis Gym	Gym	Yoga Studio	Theater	Grocery Store	Football Stadium

Clustering

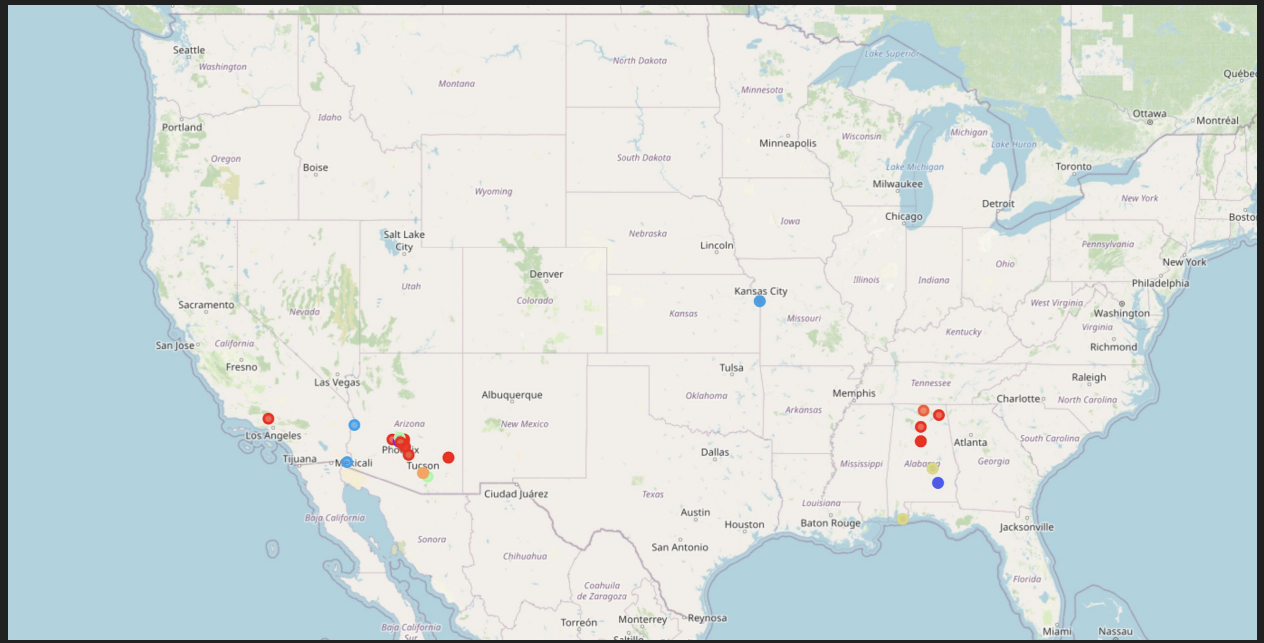
Top 5 Venues near each institute were searched, so that students can visualize the locality and places where they can find regular stuff.
K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.
AndreyBu, who has more than 5 years of machine learning experience and currently teaches people his skills, says that “the objective of K-means is simple: group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (*k*) of clusters in a dataset.”
As, per the category the venues were clustered into a total group of 10 clusters.

```
0    107
3     19
2     18
4     14
7     12
1     11
8      9
5      9
9      2
6      2
Name: Cluster Labels, dtype: int64
```

Results & Conclusion

Visualizing Result Clusters on US Map

(Please note due to limit in request calls for foursquare api only 100 rows were used)



Thank you