IBM Data Science –Capstone Report

Introduction

The United States is the 4th largest country by area [1] with a rich secondary school system (inaccurately but now referred to as a university). It has over 7,000 registered secondary schools [2]. For many young adults the environment around the university (referred to as university ambience) is just as important as the university education. The availability of cheap restaurants, foreign food, museums, bars and night life factors into many individual's choice of a university.

Problem Statement

With a large area and a rich option of universities it can be difficult for low income students to travel and assess the ambience of the universities the prospective students are interested in attending. Foursquare data can be used to obtain popular venues around a university and then to cluster universities based off the surrounding venues.

The university clusters could contain a list of other universities with similar near-campus venues or ambience. A prospective student could use this grouping to:

- Identify a nearby university to assess the ambiance for a university farther away
- Identify additional universities that share similar ambiance to a previously visited university
- Identify universities to not consider based off the similar ambiance to a previously visited university

Data

The National Center for Education Statistics maintains a robust data set for all U.S. secondary schools. The data is available from [2] and is summarized in the table below. For this analysis only the INSTNM, LAT and LON fields are used.

Field	Length	Type	Description
UNITID	8	String	School identification number
INSTNM	120	String	Name of institution
STREET	100	String	Reported street address
CITY	30	String	Reported city
STATE	2	String	Reported state
ZIP	10	String	Reported ZIP code
STFIP	2	String	State FIPS
CNTY	5	String	County FIPS
NMCNTY	40	String	County name
LOCALE	2	String	Locale code
LAT	10.6	Double	Latitude of school location
LON	11.6	Double	Longitude of school location
CBSA	5	String	Core Based Statistical Area
NMCBSA	100	String	Core Based Statistical Area name
CBSATYPE	1	String	Metropolitan or Micropolitan Statistical Area indicator
CSA	3	String	Combined Statistical Area
NMCSA	100	String	Combined Statistical Area name
NECTA	5	String	New England City and Town Area
NMNECTA	100	String	New England City and Town Area name
CD	4	String	115th Congressional District
SLDL	5	String	State Legislative District - Lower
SLDU	5	String	State Legislative District - Upper
SURVYEAR	4	String	Survey year

Foursquare [4] provides an API where a LAT and LONG can be provided and a recommended list of venues is returned in a JSON file. The details of the available return data is provided in the table below

Field	Description
warning	Presents an object with a `text` field that contains a warning message, if applicable (i.e. not enough results, try doing X).
groups	An array of objects representing groups of recommendations. Each group contains a `type` such as "recommended" a human-readable (eventually localized) `name` such as "Recommended Places," and an array `items` of recommendation objects.
suggestedRadius (optional)	If no radius was specified in the request, presents the radius that was used for the query (based upon the density of venues in the query area).
headerLocation	A text name for the location the user searched, e.g. "SoHo".
headerFullLocation	A full text name for the location the user searched, e.g. "SoHo, New York".
headerMessage	A message to the user based on their current context, e.g. "Suggestions for Tuesday afternoon".
id	A unique string identifier for this venue.
name	The best known name for this venue.
location	An object containing none, some, or all of `address` (street address), `crossStreet`, `city`, `state`, `postalCode`, `country`, `lat`, `lng`, and `distance`. All fields are strings, except for `lat`, `lng`, and `distance`. Distance is measured in meters. Some venues have their locations intentionally hidden for privacy reasons (such as private residences). If this is the case, the parameter `isFuzzed` will be set to true, and the `lat`/`lng` parameters will have reduced precision.

categories	An array, possibly empty, of categories that have been applied to this venue. One of the categories will have a `primary` field indicating that it is the primary category for the venue. For the complete category tree, see categories.

When this data is combined a ranking of the most popular venues can be created. An example for a few universities is provided below.

	School	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	ATA College - Cincinnati	Clothing Store	Gas Station	Mexican Restaurant	Supplement Shop	American Restaurant	Sandwich Place	Gym / Fitness Center	Pet Store	Restaurant	Discount Store
1	Abcott Institute	Fast Food Restaurant	Sandwich Place	Cosmetics Shop	Grocery Store	Coffee Shop	Gym / Fitness Center	BBQ Joint	Bank	Bar	Gas Station
2	Adler University	Hotel	Theater	Coffee Shop	Middle Eastern Restaurant	Gastropub	Salad Place	Donut Shop	Arts & Crafts Store	Sandwich Place	Fountain
3	Adrian College	Pizza Place	Café	Hotel	Discount Store	Optical Shop	Sandwich Place	Video Store	Fast Food Restaurant	Gym / Fitness Center	Gas Station
4	Adult and Community Education-Hudson	Discount Store	History Museum	Public Art	Food Court	Fast Food Restaurant	Soccer Stadium	Caribbean Restaurant	Gas Station	Chinese Restaurant	Park
...	...	...	...	...	...	...	...	...	...	...	...
886	Wright State University-Main Campus	Mexican Restaurant	Coffee Shop	Café	Sandwich Place	Pizza Place	Steakhouse	Kebab Restaurant	Burger Joint	Breakfast Spot	Smoke Shop
887	Xavier University	Pizza Place	Ice Cream Shop	Discount Store	Bar	High School	Home Service	Bowling Alley	College Basketball Court	Sandwich Place	Breakfast Spot

Methodology

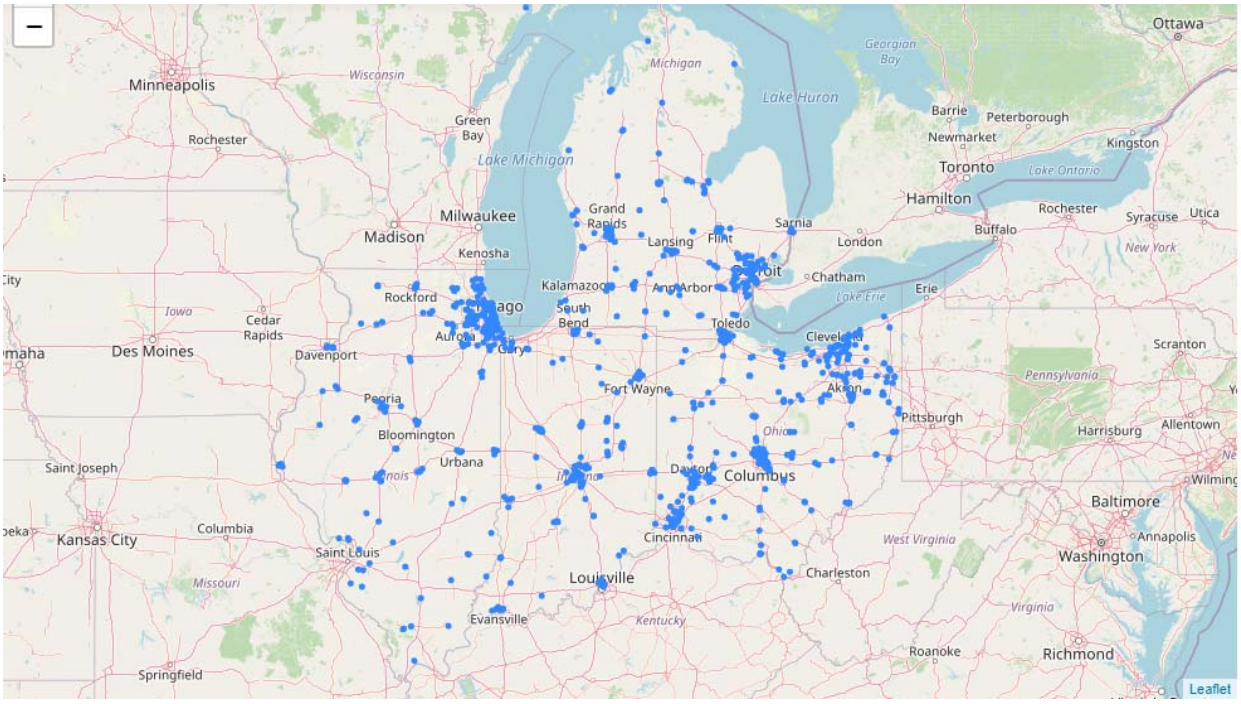
The machine learning algorithm used to group together universities is K-Means. The steps below provide a summary of the steps needed to move the data gathering to use.

1. Gather data from the National Center for Education Statistics
    a. Obtain CSV file
    b. Turn data into Pandas data frame
    c. Filter universities to the states of Indiana, Ohio, Michigan, and Illinois (done due to limitations on Foursquare API calls for anything larger)
2. Request venue data from the Foursquare API for each university in the data set
    a. Iterate through each university
    b. Store up to 30 recommended venues per university
3. For each university, sort the venues by most popular to least popular
4. Run a K-Means algorithm for all the universities, clustering based off the 10 ten venues
    a. K-Means was chosen due to its unsupervised nature and speed of computation
    b. Iterate through K values from 1 to 100
5. Use the elbow method to obtain a reasonable K
6. Map the different clusters and display the list of universities in each cluster

Methodology Features

A plot of all the U.S. secondary schools is provided below. The dataset contains 7,067 secondary schools around the world.

Due to limitations with processing and Foursquare API calls a smaller dataset was used to prove out the clustering methodology. The states of Indiana, Ohio, Michigan, and Illinois were used. The map of the associated secondary schools is below. This contain a smaller set of 924 secondary schools



Each university was passed through the Foursquare API to generate a list of nearby venues. Up to 30 venues where considered though not all places generated 30. The resulting dataset had 33,404 venues as seen below.

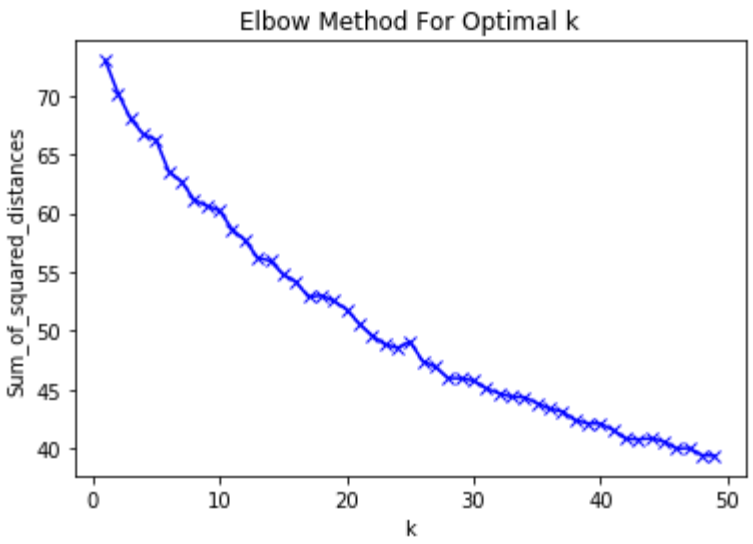
	School	School Latitude	School Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Flashpoint Chicago A Campus of Columbia Colleg...	41.882848	-87.631154	Cadillac Palace Theatre	41.884006	-87.633144	Theater
1	Flashpoint Chicago A Campus of Columbia Colleg...	41.882848	-87.631154	Kimpton Gray Hotel	41.880875	-87.631752	Hotel
2	Flashpoint Chicago A Campus of Columbia Colleg...	41.882848	-87.631154	Do-Rite Donuts & Coffee	41.884598	-87.629904	Donut Shop
3	Flashpoint Chicago A Campus of Columbia Colleg...	41.882848	-87.631154	James M. Nederlander Theatre	41.884416	-87.628861	Theater
4	Flashpoint Chicago A Campus of Columbia Colleg...	41.882848	-87.631154	Pret A Manger	41.883872	-87.628652	Sandwich Place
...	...	...	...	...	...	...	...
33400	Baker College - Flint	42.975177	-83.697246	SUBWAY	42.987833	-83.693476	Sandwich Place
33401	Baker College - Flint	42.975177	-83.697246	Family Dollar	42.974989	-83.686306	Discount Store
33402	Baker College - Flint	42.975177	-83.697246	Wendy's	42.973554	-83.684777	Fast Food Restaurant
33403	Baker College - Flint	42.975177	-83.697246	Capitol	42.973321	-83.712547	Diner
33404	Baker College - Flint	42.975177	-83.697246	Halo Burger	42.980844	-83.692392	Burger Joint

33405 rows × 7 columns

The venues where counted, ranked, and sorted based off of their popularity as seen below

	School	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	ATA College - Cincinnati	Clothing Store	Gas Station	Mexican Restaurant	Supplement Shop	American Restaurant	Sandwich Place	Gym / Fitness Center	Pet Store	Restaurant	Discount Store
1	Abcott Institute	Fast Food Restaurant	Sandwich Place	Cosmetics Shop	Grocery Store	Coffee Shop	Gym / Fitness Center	BBQ Joint	Bank	Bar	Gas Station
2	Adler University	Hotel	Theater	Coffee Shop	Middle Eastern Restaurant	Gastropub	Salad Place	Donut Shop	Arts & Crafts Store	Sandwich Place	Fountain
3	Adrian College	Pizza Place	Café	Hotel	Discount Store	Optical Shop	Sandwich Place	Video Store	Fast Food Restaurant	Gym / Fitness Center	Gas Station
4	Adult and Community Education-Hudson	Discount Store	History Museum	Public Art	Food Court	Fast Food Restaurant	Soccer Stadium	Caribbean Restaurant	Gas Station	Chinese Restaurant	Park
...	...	...	...	...	...	...	...	...	...	...	...
886	Wright State University-Main Campus	Mexican Restaurant	Coffee Shop	Café	Sandwich Place	Pizza Place	Steakhouse	Kebab Restaurant	Burger Joint	Breakfast Spot	Smoke Shop
887	Xavier University	Pizza Place	Ice Cream Shop	Discount Store	Bar	High School	Home Service	Bowling Alley	College Basketball Court	Sandwich Place	Breakfast Spot

The dataset was passed to a K-Means algorithm. The challenge of finding the optimal K was solved by using the elbow method. The elbow method visually looks for diminishing returns in the sum of squared distances for various Ks. Based off of the analysis below an ideal K of 30 was chosen.



Results

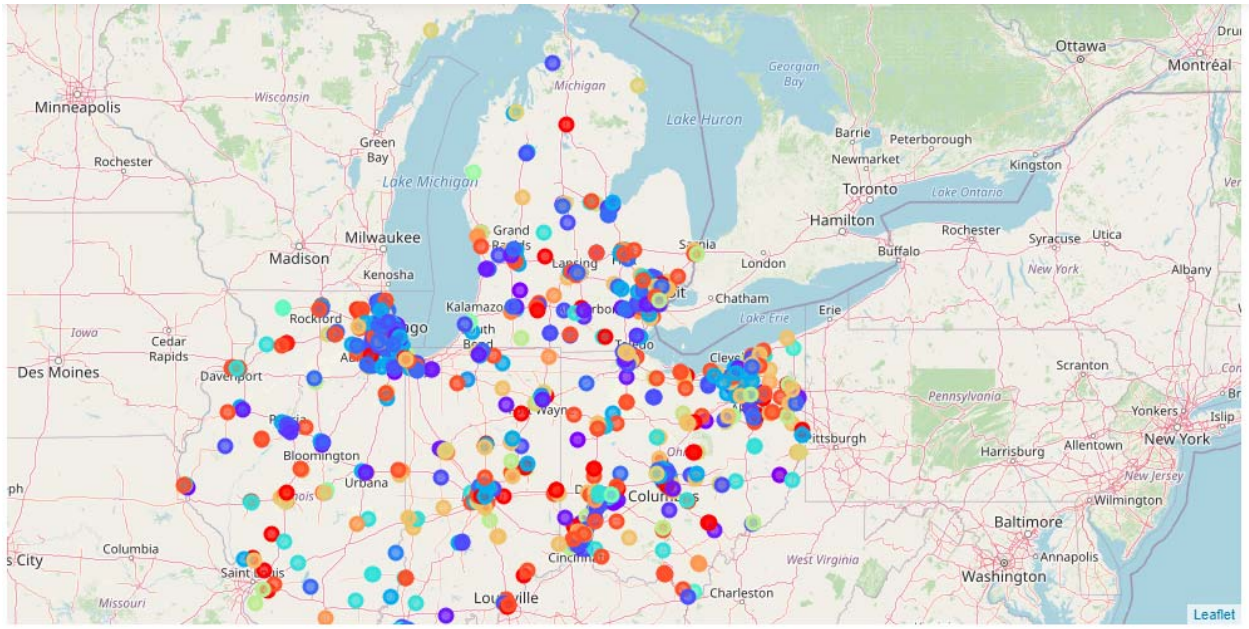
The available data set successfully allows an individual to search through similarly clustered universities to answer the questions original brought up in this report. An example of the school listing to answering these questions is below.

```
*****Universities in cluster 0 *****
                                                NAME
                      Southwestern Illinois College
                                  Blackburn College
                           Chicago State University
     City Colleges of Chicago-Kennedy-King College
                               John A Logan College
                 Lewis and Clark Community College
                               Prairie State College
        Southern Illinois University-Edwardsville
                                     Hanover College
       PJ's College of Cosmetology- Jeffersonville
                                 Huntington University
                  Ivy Tech Community College-Kokomo
               Ivy Tech Community College-Richmond
                           Indiana University-East
                        Ravenscroft Beauty College
           PJ's College of Cosmetology- Greenfield
               PJ's College of Cosmetology- Muncie
```

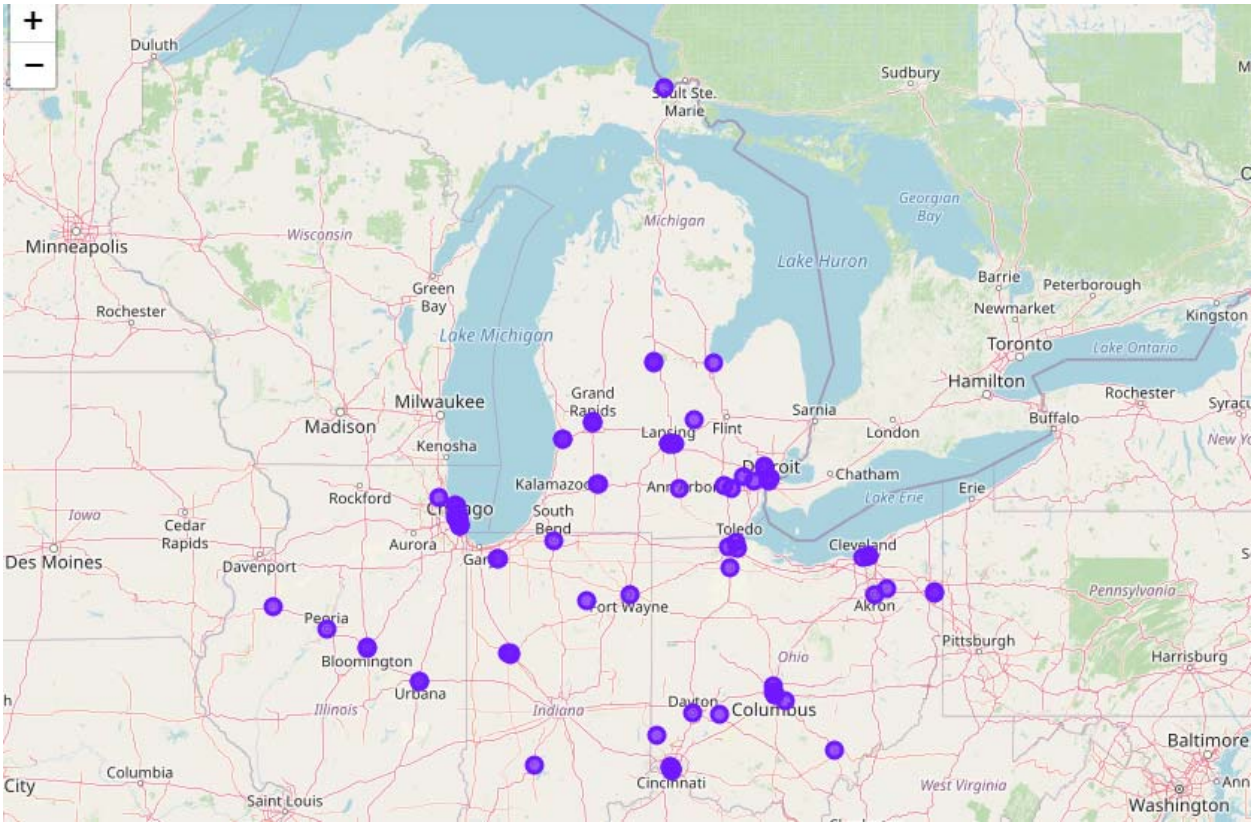
The result of using K-Means to cluster universities based off nearby popular venue data resulted in 30 different clusters of universities yielding a useful clustering. The largest cluster had 110 schools in it. A sample breakdown is below.

Cluster Labels	
23.0	110
12.0	94
6.0	84
1.0	79
2.0	73
3.0	64
26.0	52
29.0	51
39.0	50
20.0	47
15.0	46
31.0	38
28.0	25
10.0	23
24.0	11
13.0	6
0.0	5

The visual clustering of universities is presented below.



An example of a single cluster (cluster 2) is below and can be used to help prospective students strategically chose the closest school so that they can assess the ambience of a more distant school



Discussion

It is impossible to say if the clustering is completely accurate but qualitative checking provides some insights:

- Major universities such as Purdue, the University of Illinois, University of Michigan and Ohio state are all in the same cluster. As universities with other similar qualities it seems reasonable for them to be clustered together based off nearby venues.
- Schools in very rural areas tend to be clustered together.
- Schools in high population areas such as Chicago tend to be clustered together or share a small number of clusters. This qualitatively makes sense as they are relatively close to each other and share common venues

The data is potentially useful to future students to help assess the ambience of a future school. If more processing capabilities where available and Foursquare API access was increased, this analysis could be expanded to the entire U.S. using the same methods outlined here.

In addition to expanding to the entire U.S. it is suggested that additional data be considered outside of venue data. Crime statistics, population density and medium income would potentially create more insightful clusters for potential students.

Conclusion

This analysis provides a value to potential secondary school students who want to evaluate a universities ambience. The available groupings correlate well to other existing similarities in major universities and

highlight the commonality that rural universities share. The dataset can be helpful in answering the original questions presented by this report:

- Identify a nearby university to assess the ambiance for a university farther away
- Identify additional universities that share similar ambiance to a previously visited university
- Identify universities to not consider based off the similar ambiance to a previously visited university

Expanding the analysis to the entire U.S. and considering other data sources would potentially provide even more insights to potential students.

References

1. Largest Countries in the World - https://www.worldometers.info/geography/largest-countries-in-the-world/
2. National Center for Education Statistics - https://nces.ed.gov/ipeds/use-the-data/survey-components
3. National Center for Education Statistics, Data - https://nces.ed.gov/programs/edge/docs/EDGE_GEOCODE_POSTSEC_1617.pdf&sa=U&ved=2ahUKEwiaiLyfwYbqAhUSbq0KHeUyDhAQFjAAegQIAhAB&usg=AOvVaw3Bf6JohA_PCo7m4fWhY918
4. Four Square Developer API - https://developer.foursquare.com/