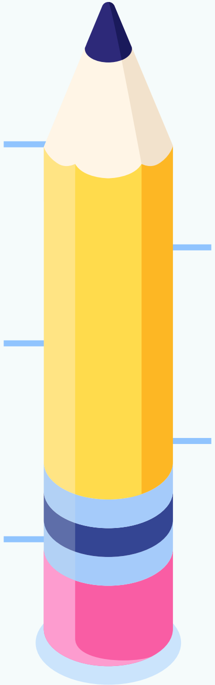
COLLEGE CLUSTERING

CAPSTONE DEVELOPED AS A PART OF THE IBM DATA SCIENCE PROFESSIONAL CERTIFICATION.

TABLE OF CONTENTS



01 INTRODUCTION

Brief overview of the project problem statement and proposed solution

02 TECH STACK

Technologies and platforms used to create project

03 THE DATA

Information about the data source

04 METHODOLOGY

A description of the data wrangling and cleaning techniques used as well as the Machine Learning algorithms applied

05 OUTRO

The results and conclusions of the project

01

INTRODUCTION

- PREFACE
- PROBLEM STATEMENT
- PROPOSED SOLUTION

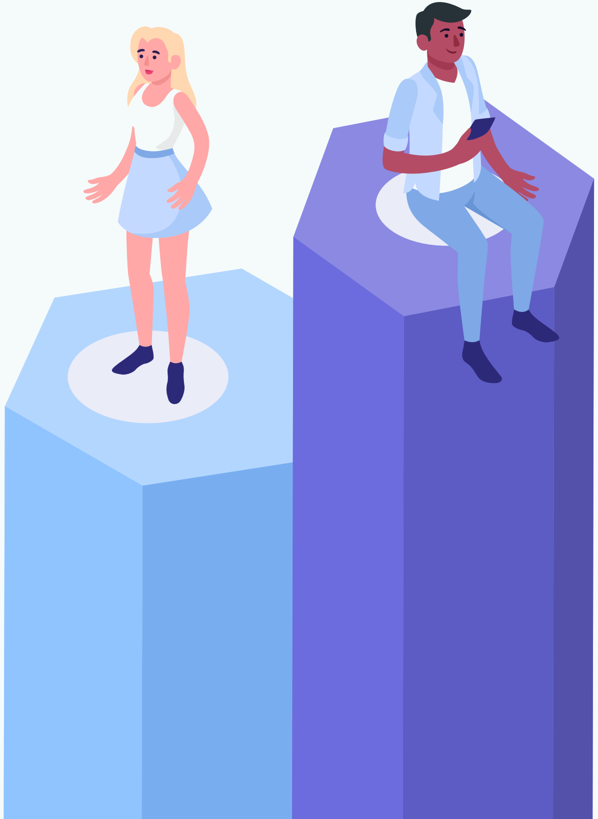
the PREFACE

- ➔ Every year, millions of students from around the world apply to undergraduate, graduate, and doctoral programs alike in the united states.
- ➔ While making decisions and shortlisting perspective schools, one of the factors that is considered is the location of the institution.
- ➔ The surrounding amenities and features pay a major role in college selection as different people prefer different institutional settings.
- ➔ The current project serves to provide a simple information repository to help prospective applicants make informed decisions by highlighting the kind of venues that are located around a given institution.
- ➔ Students can also discover the top 5 colleges similar to one they are already interested in based on location setting.

PROBLEM

Applying for college is hard and looking for universities that are in the kind of neighborhoods that one prefers can prove to be a tedious added burden.

To research the surroundings of the college, students have to scour the web for hours, looking for information on each institution



SOLUTION

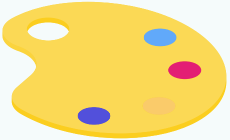
Creating a dynamic and cohesive dataset that has information on each college and the colleges are clustered into groups of like objects.

This dataset can thus be queried by the user to extract relevant information about venues surrounding a college. One can even find colleges with similar location settings.

02

TECH STACK

Technologies and platforms
used to create project



PLATFORMS USED

FOURSQUARE

This is the API service used to obtain the nearby venue data for each college given geographical coordinates.

WATSON STUDIO

An online integrated environment provided via IBM cloud that helps to centrally store all the data relevant to ones data science projects.

JUPYTER LABS

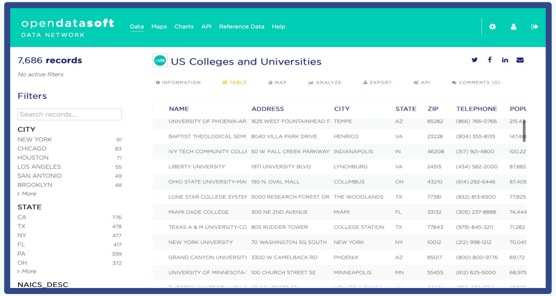
A web based integrated environment to created jupyter notebooks which are interactive documents used primarily for data analytics and research projects

THE DATA

Information about the data source

03

The data was sourced from a website called opendatasoft and was titled the US college and university data.The colleges and university dataset is composed of all Post Secondary Education facilities as defined by the Integrated Post Secondary Education System (IPEDS), National Center for Education Statistics, US Department of Education.



opendatasoft DATA NETWORK

7,686 **records**

No active filters

Filters

Search records...

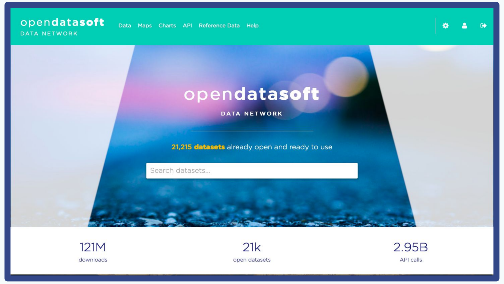
CITY	
NEW YORK	91
CHICAGO	83
HOUSTON	71
LOS ANGELES	55
SAN ANTONIO	49
BROOKLYN	48
› *More*	

STATE	
CA	776
TX	478
NY	477
FL	417
PA	399
OH	372
› *More*	

NAICS_DESC

US Colleges and Universities

NAME	ADDRESS	CITY	STATE	ZIP	TELEPHONE	POPU
UNIVERSITY OF PHOENIX-AR	1625 WEST FOUNTAINHEAD F	TEMPE	AZ	85282	(866) 766-0766	215,43
BAPTIST THEOLOGICAL SEMI	8040 VILLA PARK DRIVE	HENRICO	VA	23228	(804) 355-8135	147,48
IVY TECH COMMUNITY COLLI	50 W. FALL CREEK PARKWAY	INDIANAPOLIS	IN	46208	(317) 921-4800	100,22
LIBERTY UNIVERSITY	1971 UNIVERSITY BLVD	LYNCHBURG	VA	24515	(434) 582-2000	87,882
OHIO STATE UNIVERSITY-MAI	190 N. OVAL MALL	COLUMBUS	OH	43210	(614) 292-6446	87,405
LONE STAR COLLEGE SYSTEI	5000 RESEARCH FOREST DR	THE WOODLANDS	TX	77381	(832) 813-6500	77,825
MIAMI DADE COLLEGE	300 NE 2ND AVENUE	MIAMI	FL	33132	(305) 237-8888	74,444
TEXAS A & M UNIVERSITY-CO	805 RUDDER TOWER	COLLEGE STATION	TX	77843	(979) 845-3211	71,282
NEW YORK UNIVERSITY	70 WASHINGTON SQ SOUTH	NEW YORK	NY	10012	(212) 998-1212	70,04
GRAND CANYON UNIVERSITY	3300 W CAMELBACK RD	PHOENIX	AZ	85017	(800) 800-9776	69,172
UNIVERSITY OF MINNESOTA-	100 CHURCH STREET SE	MINNEAPOLIS	MN	55455	(612) 625-5000	68,975



opendatasoft DATA NETWORK

opendatasoft

DATA NETWORK

21,215 datasets already open and ready to use

Search datasets...

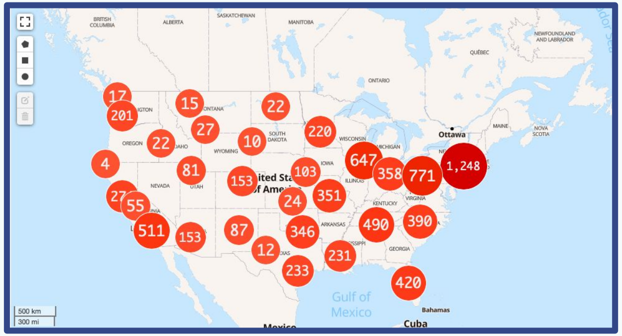
121M downloads

21k open datasets

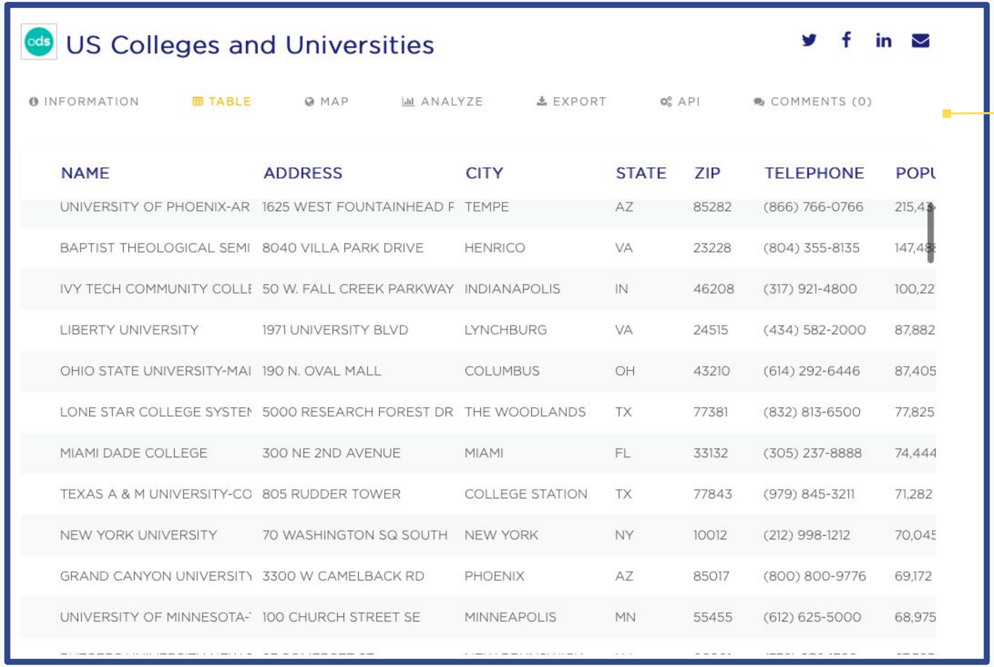
2.95B API calls

Included are Doctoral/Research Universities, Masters Colleges and Universities, Baccalaureate Colleges, Associates Colleges, Theological seminaries, Medical Schools and other health care professions, Schools of engineering and technology, business and management, art, music, design, Law schools, Teachers colleges, Tribal colleges, and other specialized institutions. Overall, this data layer covers all 50 states, as well as Puerto Rico and other assorted U.S. territories.

This feature class contains all MEDS/MEDS+ as approved by NGA. For each field the 'Not available' and 'NULL' designations are used to indicate that the data for the particular record and field is currently unavailable and will be populated when and if that data becomes available.



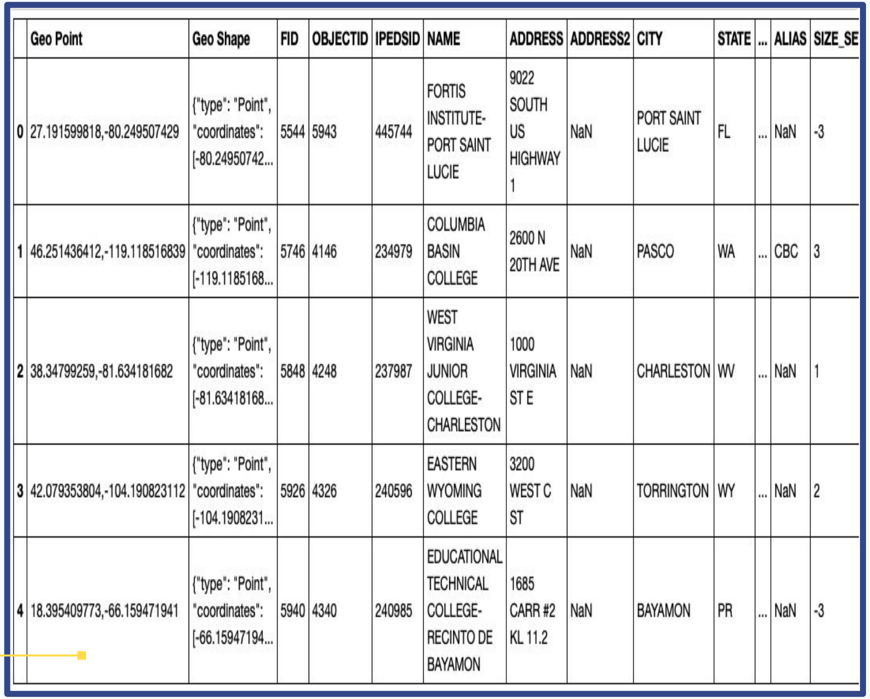
The dataset structure



US Colleges and Universities

NAME	ADDRESS	CITY	STATE	ZIP	TELEPHONE	POPU
UNIVERSITY OF PHOENIX-AR	1625 WEST FOUNTAINHEAD F	TEMPE	AZ	85282	(866) 766-0766	215,43
BAPTIST THEOLOGICAL SEMI	8040 VILLA PARK DRIVE	HENRICO	VA	23228	(804) 355-8135	147,48
IVY TECH COMMUNITY COLLE	50 W. FALL CREEK PARKWAY	INDIANAPOLIS	IN	46208	(317) 921-4800	100,22
LIBERTY UNIVERSITY	1971 UNIVERSITY BLVD	LYNCHBURG	VA	24515	(434) 582-2000	87,882
OHIO STATE UNIVERSITY-MAI	190 N. OVAL MALL	COLUMBUS	OH	43210	(614) 292-6446	87,405
LONE STAR COLLEGE SYSTEN	5000 RESEARCH FOREST DR	THE WOODLANDS	TX	77381	(832) 813-6500	77,825
MIAMI DADE COLLEGE	300 NE 2ND AVENUE	MIAMI	FL	33132	(305) 237-8888	74,444
TEXAS A & M UNIVERSITY-CO	805 RUDDER TOWER	COLLEGE STATION	TX	77843	(979) 845-3211	71,282
NEW YORK UNIVERSITY	70 WASHINGTON SQ SOUTH	NEW YORK	NY	10012	(212) 998-1212	70,045
GRAND CANYON UNIVERSITY	3300 W CAMELBACK RD	PHOENIX	AZ	85017	(800) 800-9776	69,172
UNIVERSITY OF MINNESOTA-	100 CHURCH STREET SE	MINNEAPOLIS	MN	55455	(612) 625-5000	68,975

The dataset as structured on the original website is shown adjacent. It has about 7500 entries and 45 attributes.



	Geo Point	Geo Shape	FID	OBJECTID	IPEDSID	NAME	ADDRESS	ADDRESS2	CITY	STATE	...	ALIAS	SIZE_SE
0	27.191599818,-80.249507429	{"type": "Point", "coordinates": [-80.24950742...	5544	5943	445744	FORTIS INSTITUTE-PORT SAINT LUCIE	9022 SOUTH US HIGHWAY 1	NaN	PORT SAINT LUCIE	FL	...	NaN	-3
1	46.251436412,-119.118516839	{"type": "Point", "coordinates": [-119.1185168...	5746	4146	234979	COLUMBIA BASIN COLLEGE	2600 N 20TH AVE	NaN	PASCO	WA	...	CBC	3
2	38.34799259,-81.634181682	{"type": "Point", "coordinates": [-81.63418168...	5848	4248	237987	WEST VIRGINIA JUNIOR COLLEGE-CHARLESTON	1000 VIRGINIA ST E	NaN	CHARLESTON	WV	...	NaN	1
3	42.079353804,-104.190823112	{"type": "Point", "coordinates": [-104.1908231...	5926	4326	240596	EASTERN WYOMING COLLEGE	3200 WEST C ST	NaN	TORRINGTON	WY	...	NaN	2
4	18.395409773,-66.159471941	{"type": "Point", "coordinates": [-66.15947194...	5940	4340	240985	EDUCATIONAL TECHNICAL COLLEGE-RECINTO DE BAYAMON	1685 CARR #2 KL 11.2	NaN	BAYAMON	PR	...	NaN	-3

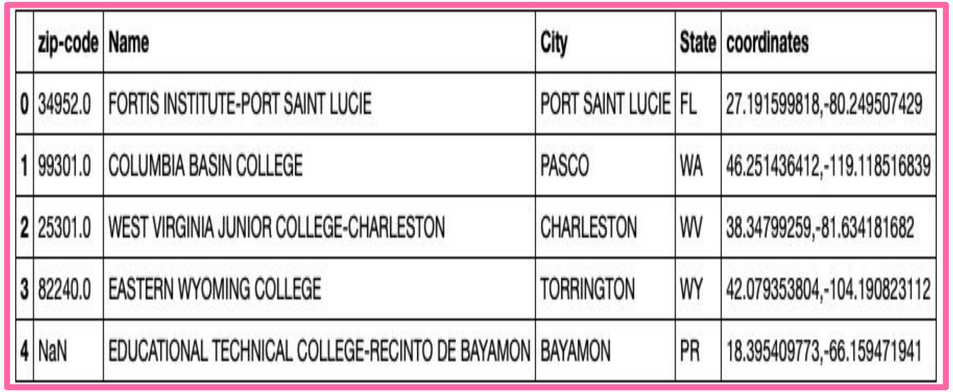
When imported into the jupyter labs environment, the dataset looks as shown adjacent.

METHODOLOGY

04

In depth description of data science and machine learning principles used

DATA CLEANING AND WRANGLING



	zip-code	Name	City	State	coordinates
0	34952.0	FORTIS INSTITUTE-PORT SAINT LUCIE	PORT SAINT LUCIE	FL	27.191599818,-80.249507429
1	99301.0	COLUMBIA BASIN COLLEGE	PASCO	WA	46.251436412,-119.118516839
2	25301.0	WEST VIRGINIA JUNIOR COLLEGE-CHARLESTON	CHARLESTON	WV	38.34799259,-81.634181682
3	82240.0	EASTERN WYOMING COLLEGE	TORRINGTON	WY	42.079353804,-104.190823112
4	NaN	EDUCATIONAL TECHNICAL COLLEGE-RECINTO DE BAYAMON	BAYAMON	PR	18.395409773,-66.159471941

01

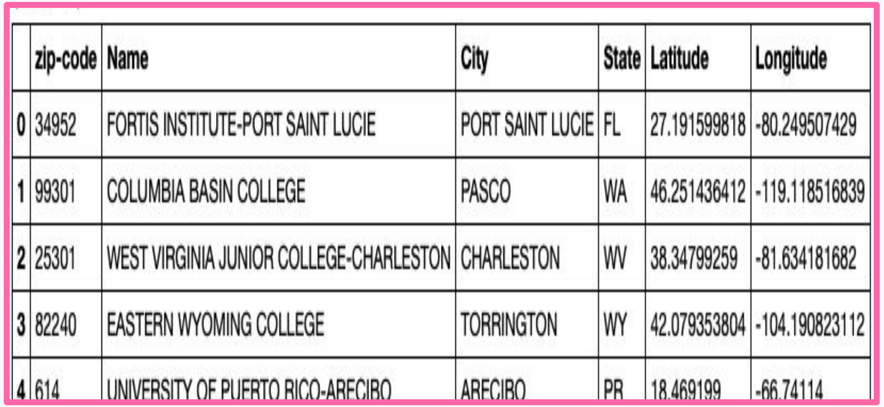
STREAMLINING

In the previous section, an image of the imported dataset was shown. From that image it was clear that a lot of extra information was available to us. Thus a new data frame was created holding information relevant to present context only.

02

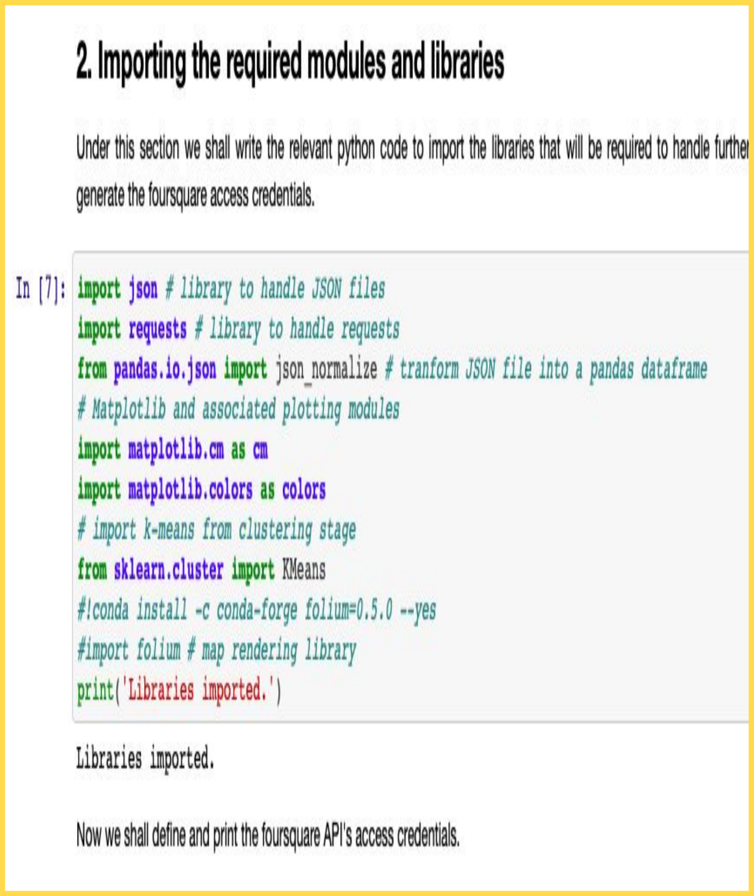
ENHANCING

After creation of a streamlined dataset, the coordinates column was converted into separate latitude and longitude attributes for ease of API request generation. Null value problem was also addressed.



	zip-code	Name	City	State	Latitude	Longitude
0	34952	FORTIS INSTITUTE-PORT SAINT LUCIE	PORT SAINT LUCIE	FL	27.191599818	-80.249507429
1	99301	COLUMBIA BASIN COLLEGE	PASCO	WA	46.251436412	-119.118516839
2	25301	WEST VIRGINIA JUNIOR COLLEGE-CHARLESTON	CHARLESTON	WV	38.34799259	-81.634181682
3	82240	EASTERN WYOMING COLLEGE	TORRINGTON	WY	42.079353804	-104.190823112
4	614	UNIVERSITY OF PUERTO RICO-ARECIBO	ARECIBO	PR	18.469199	-66.74114

IMPORTING LIBRARIES



2. Importing the required modules and libraries

Under this section we shall write the relevant python code to import the libraries that will be required to handle furthe
generate the foursquare access credentials.

```
In [7]: import json # library to handle JSON files
        import requests # library to handle requests
        from pandas.io.json import json_normalize # tranform JSON file into a pandas dataframe
        # Matplotlib and associated plotting modules
        import matplotlib.cm as cm
        import matplotlib.colors as colors
        # import k-means from clustering stage
        from sklearn.cluster import KMeans
        #!conda install -c conda-forge folium=0.5.0 --yes
        #import folium # map rendering library
        print('Libraries imported.')
```

```
Libraries imported.
```

Now we shall define and print the foursquare API's access credentials.

REQUESTS

To handle API request made to the foursquare API

PANDAS AND NUMPY

To handle and manipulate the data and associated dataframes

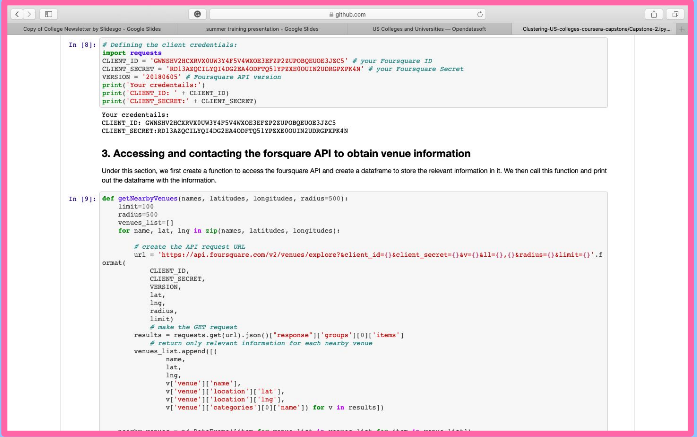
FOLIUM

To plot and visualize the results of the clustering

SCIKIT LEARN

For the K Means clustering algorithm

DEFINING CREDENTIALS AND CONTACTING THE API


github.com
Copy of College Newsletter by Slidesgo - Google Slides
summer training presentation - Google Slides
US Colleges and Universities — Opendatasoft
Clustering-US-colleges-coursera-capstone/Capstone-2.ipy...

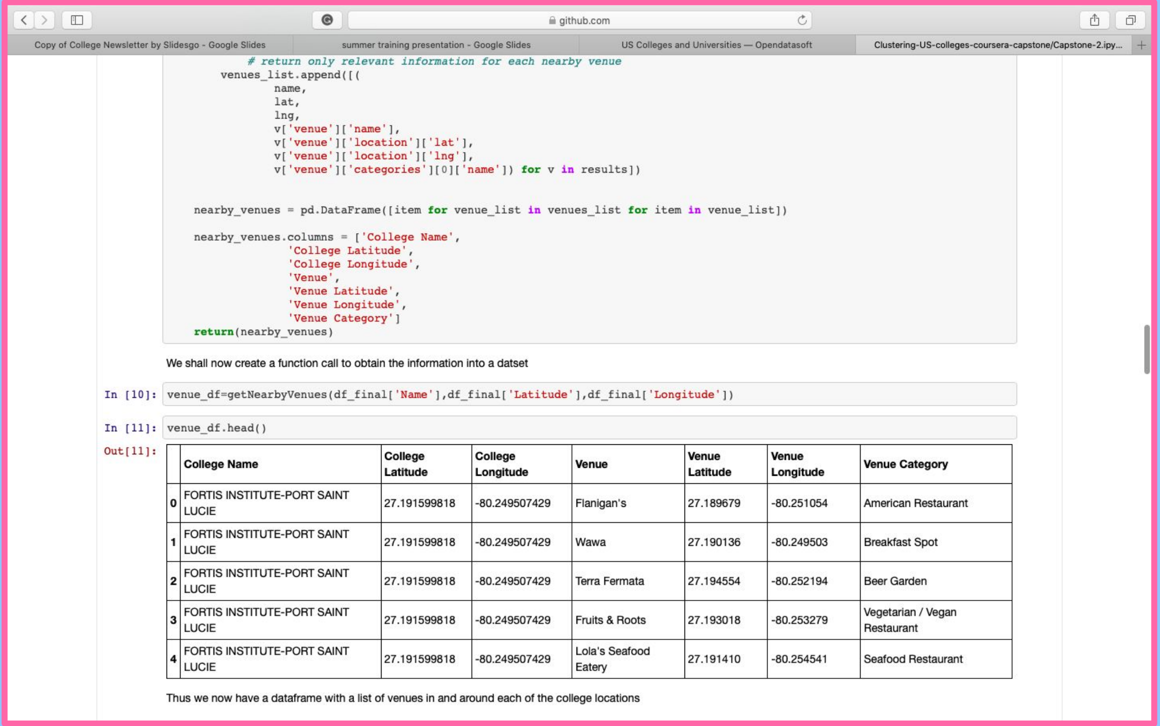
```
In [8]: # Defining the client credentials:
        import requests
        CLIENT_ID = 'GWNSHV2HCXRVX0UW3Y4F5V4WXOE3EFZP2ZUPOBQEUOE3JZC5' # your Foursquare ID
        CLIENT_SECRET = 'RD13AZQCILYQI4DG2EA4ODFTQ51YPZXE0OUIN2UDRGPXPK4N' # your Foursquare Secret
        VERSION = '20180605' # Foursquare API version
        print('Your credentails:')
        print('CLIENT_ID: ' + CLIENT_ID)
        print('CLIENT_SECRET:' + CLIENT_SECRET)

        Your credentails:
        CLIENT_ID: GWNSHV2HCXRVX0UW3Y4F5V4WXOE3EFZP2ZUPOBQEUOE3JZC5
        CLIENT_SECRET:RD13AZQCILYQI4DG2EA4ODFTQ51YPZXE0OUIN2UDRGPXPK4N
```

3. Accessing and contacting the forsquare API to obtain venue information

Under this section, we first create a function to access the foursquare API and create a dataframe to store the relevant information in it. We then call this function and print out the dataframe with the information.

```
In [9]: def getNearbyVenues(names, latitudes, longitudes, radius=500):
            limit=100
            radius=500
            venues_list=[]
            for name, lat, lng in zip(names, latitudes, longitudes):

                # create the API request URL
                url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.f
        ormat(
                    CLIENT_ID,
                    CLIENT_SECRET,
                    VERSION,
                    lat,
                    lng,
                    radius,
                    limit)
                    # make the GET request
                results = requests.get(url).json()["response"]['groups'][0]['items']
                    # return only relevant information for each nearby venue
                venues_list.append([(
                        name,
                        lat,
                        lng,
                        v['venue']['name'],
                        v['venue']['location']['lat'],
                        v['venue']['location']['lng'],
                        v['venue']['categories'][0]['name']) for v in results])
```

After Contacting the foursquare API, the data for each of the venue associated with a given college was stored in a data frame called the venues_df dataframe.


github.com
Copy of College Newsletter by Slidesgo - Google Slides
summer training presentation - Google Slides
US Colleges and Universities — Opendatasoft
Clustering-US-colleges-coursera-capstone/Capstone-2.ipy...

```
            # return only relevant information for each nearby venue
        venues_list.append([(
                name,
                lat,
                lng,
                v['venue']['name'],
                v['venue']['location']['lat'],
                v['venue']['location']['lng'],
                v['venue']['categories'][0]['name']) for v in results])
```

```
    nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in venue_list])
    
    nearby_venues.columns = ['College Name',
                  'College Latitude',
                  'College Longitude',
                  'Venue',
                  'Venue Latitude',
                  'Venue Longitude',
                  'Venue Category']
    return(nearby_venues)
```

We shall now create a function call to obtain the information into a datset

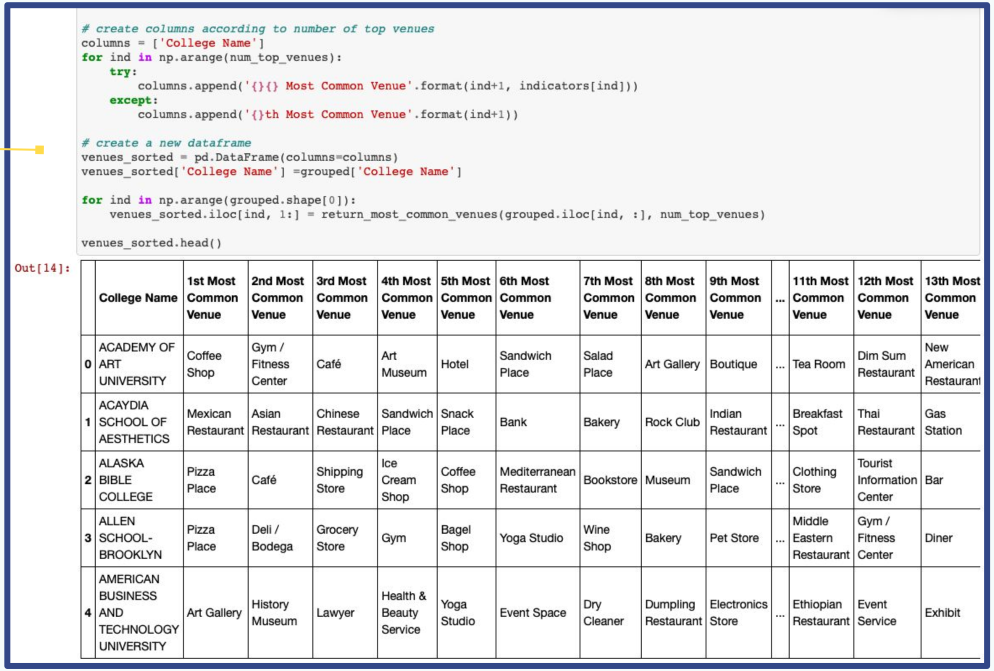
```
In [10]: venue_df=getNearbyVenues(df_final['Name'],df_final['Latitude'],df_final['Longitude'])

In [11]: venue_df.head()

Out[11]:
```

	College Name	College Latitude	College Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	FORTIS INSTITUTE-PORT SAINT LUCIE	27.191599818	-80.249507429	Flanigan's	27.189679	-80.251054	American Restaurant
1	FORTIS INSTITUTE-PORT SAINT LUCIE	27.191599818	-80.249507429	Wawa	27.190136	-80.249503	Breakfast Spot
2	FORTIS INSTITUTE-PORT SAINT LUCIE	27.191599818	-80.249507429	Terra Fermata	27.194554	-80.252194	Beer Garden
3	FORTIS INSTITUTE-PORT SAINT LUCIE	27.191599818	-80.249507429	Fruits & Roots	27.193018	-80.253279	Vegetarian / Vegan Restaurant
4	FORTIS INSTITUTE-PORT SAINT LUCIE	27.191599818	-80.249507429	Lola's Seafood Eatery	27.191410	-80.254541	Seafood Restaurant

Thus we now have a dataframe with a list of venues in and around each of the college locations

OPTIMIZING THE RESULT PRODUCTIVITY

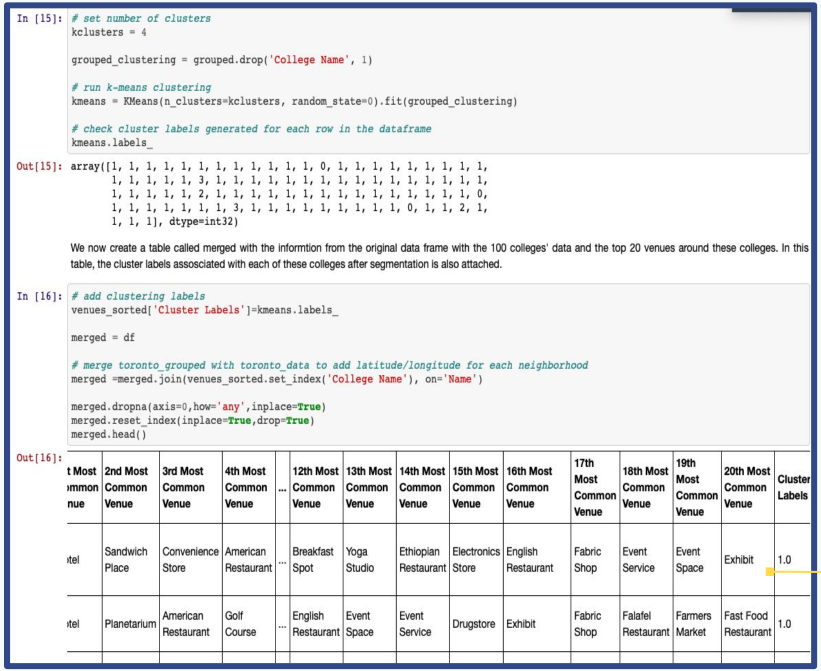
Grouping and one hot encoding the obtained data set and then obtaining a dataset with the top 20 most popular venues for each location.



```
# create columns according to number of top venues
columns = ['College Name']
for ind in np.arange(num_top_venues):
    try:
        columns.append('{}{} Most Common Venue'.format(ind+1, indicators[ind]))
    except:
        columns.append('{}th Most Common Venue'.format(ind+1))

# create a new dataframe
venues_sorted = pd.DataFrame(columns=columns)
venues_sorted['College Name'] =grouped['College Name']

for ind in np.arange(grouped.shape[0]):
    venues_sorted.iloc[ind, 1:] = return_most_common_venues(grouped.iloc[ind, :], num_top_venues)

venues_sorted.head()
```

Out[14]:

	College Name	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	...	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue
0	ACADEMY OF ART UNIVERSITY	Coffee Shop	Gym / Fitness Center	Café	Art Museum	Hotel	Sandwich Place	Salad Place	Art Gallery	Boutique	...	Tea Room	Dim Sum Restaurant	New American Restaurant
1	ACAYDIA SCHOOL OF AESTHETICS	Mexican Restaurant	Asian Restaurant	Chinese Restaurant	Sandwich Place	Snack Place	Bank	Bakery	Rock Club	Indian Restaurant	...	Breakfast Spot	Thai Restaurant	Gas Station
2	ALASKA BIBLE COLLEGE	Pizza Place	Café	Shipping Store	Ice Cream Shop	Coffee Shop	Mediterranean Restaurant	Bookstore	Museum	Sandwich Place	...	Clothing Store	Tourist Information Center	Bar
3	ALLEN SCHOOL-BROOKLYN	Pizza Place	Deli / Bodega	Grocery Store	Gym	Bagel Shop	Yoga Studio	Wine Shop	Bakery	Pet Store	...	Middle Eastern Restaurant	Gym / Fitness Center	Diner
4	AMERICAN BUSINESS AND TECHNOLOGY UNIVERSITY	Art Gallery	History Museum	Lawyer	Health & Beauty Service	Yoga Studio	Event Space	Dry Cleaner	Dumpling Restaurant	Electronics Store	...	Ethiopian Restaurant	Event Service	Exhibit



```
In [15]: # set number of clusters
         kclusters = 4

         grouped_clustering = grouped.drop('College Name', 1)

         # run k-means clustering
         kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(grouped_clustering)

         # check cluster labels generated for each row in the dataframe
         kmeans.labels_
```

```
Out[15]: array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 1, 1, 1, 1, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0,
                1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 2, 1,
                1, 1, 1], dtype=int32)
```

We now create a table called merged with the informtion from the original data frame with the 100 colleges' data and the top 20 venues around these colleges. In this table, the cluster labels assosciated with each of these colleges after segmentation is also attached.

```
In [16]: # add clustering labels
         venues_sorted['Cluster Labels']=kmeans.labels_

         merged = df

         # merge toronto_grouped with toronto_data to add latitude/longitude for each neighborhood
         merged =merged.join(venues_sorted.set_index('College Name'), on='Name')

         merged.dropna(axis=0,how='any',inplace=True)
         merged.reset_index(inplace=True,drop=True)
         merged.head()
```

Out[16]:

t Most ımmon nue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	...	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue	15th Most Common Venue	16th Most Common Venue	17th Most Common Venue	18th Most Common Venue	19th Most Common Venue	20th Most Common Venue	Cluster Labels
ıtel	Sandwich Place	Convenience Store	American Restaurant	...	Breakfast Spot	Yoga Studio	Ethiopian Restaurant	Electronics Store	English Restaurant	Fabric Shop	Event Service	Event Space	Exhibit	1.0
ıtel	Planetarium	American Restaurant	Golf Course	...	English Restaurant	Event Space	Event Service	Drugstore	Exhibit	Fabric Shop	Falafel Restaurant	Farmers Market	Fast Food Restaurant	1.0

Clustering the data in the above illustrated table and thus obtaining cluster labels and adding them back to the database to obtain penultimate data resource that is to be queried.

K MEANS ALGORITHM ADVANTAGES

IMPLEMENTATION

The algorithm is relatively simple to implement and comprehend.

CONVERGENCE

The algorithm guarantees convergence if it is run sufficient number of times.

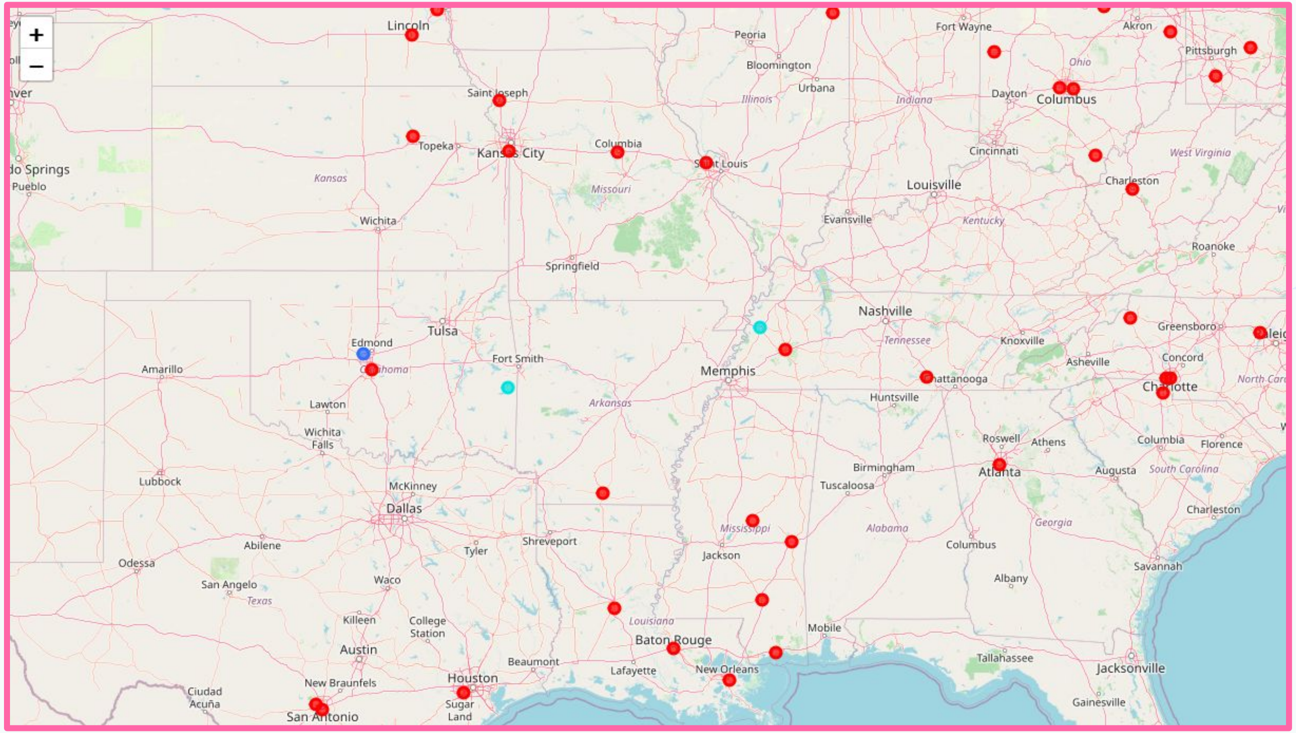
ADAPTIVE

Easily adapts to new examples and has low value of time complexity.

BIG DATASETS

The algorithm scales to large datasets quite easily.

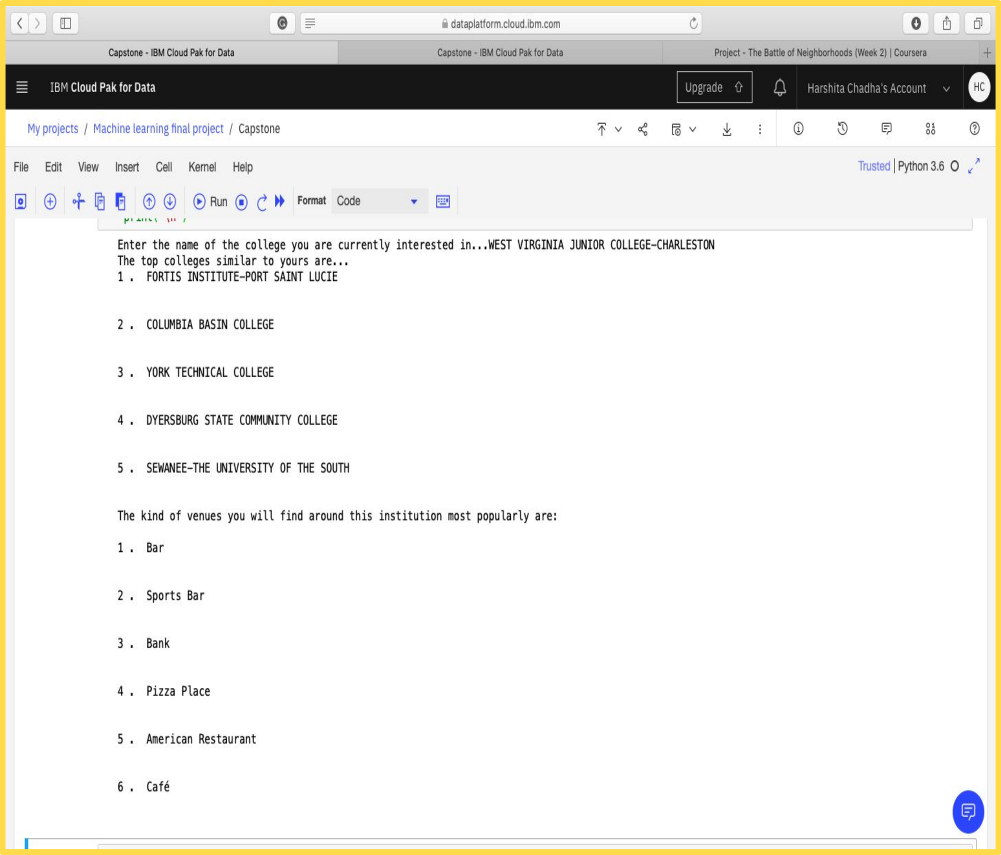
VISUALIZED RESULTS



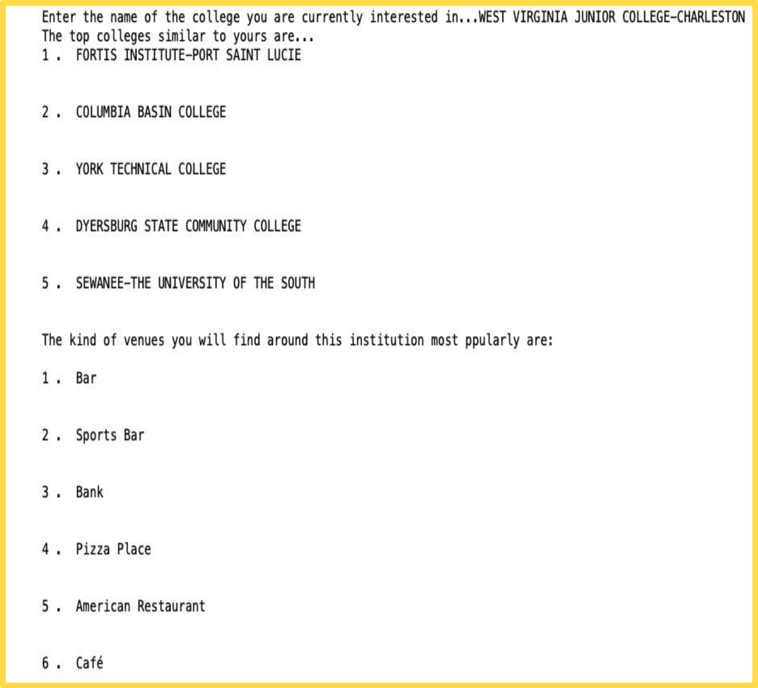
RESULTS OF CLUSTERING

USER INTERACTION

The images illustrate the rudimentary UI that was created for user interaction.


dataplatform.cloud.ibm.com
Capstone - IBM Cloud Pak for Data
Capstone - IBM Cloud Pak for Data
Project - The Battle of Neighborhoods (Week 2) | Coursera
IBM Cloud Pak for Data
Upgrade
Harshita Chadha's Account
HC
My projects / Machine learning final project / Capstone
File
Edit
View
Insert
Cell
Kernel
Help
Trusted
Python 3.6
Run
Format
Code

```
Enter the name of the college you are currently interested in...WEST VIRGINIA JUNIOR COLLEGE-CHARLESTON
The top colleges similar to yours are...
1 .  FORTIS INSTITUTE-PORT SAINT LUCIE


2 .  COLUMBIA BASIN COLLEGE


3 .  YORK TECHNICAL COLLEGE


4 .  DYERSBURG STATE COMMUNITY COLLEGE


5 .  SEWANEE-THE UNIVERSITY OF THE SOUTH


The kind of venues you will find around this institution most popularly are:

1 .  Bar


2 .  Sports Bar


3 .  Bank


4 .  Pizza Place


5 .  American Restaurant


6 .  Café
```



```
Enter the name of the college you are currently interested in...WEST VIRGINIA JUNIOR COLLEGE-CHARLESTON
The top colleges similar to yours are...
1 .  FORTIS INSTITUTE-PORT SAINT LUCIE


2 .  COLUMBIA BASIN COLLEGE


3 .  YORK TECHNICAL COLLEGE


4 .  DYERSBURG STATE COMMUNITY COLLEGE


5 .  SEWANEE-THE UNIVERSITY OF THE SOUTH


The kind of venues you will find around this institution most ppularly are:

1 .  Bar


2 .  Sports Bar


3 .  Bank


4 .  Pizza Place


5 .  American Restaurant


6 .  Café
```

OUTRO

A discussion on the results and the conclusions

05

RESULTS AND CONCLUSION

The project helped gain another new skill which is to interact with the foursquare API and manage and manipulate the results returned by it to draw meaningful conclusions. At the end of the development cycle, a highly informative database was created which was flexible enough to be updated to present real-time, up to date information and could be easily queried to extract useful information.

Thus, as a result of the training, the individual was able to master industry-relevant skills that have a high value in the tech market presently. What is more, is that practical hands-on experience in solving real-life problems using data science was also gained which helped build candidate portfolio and further career aspirations.

THANKS

Do you have any questions?

PLEASE FEEL FREE TO CONTACT ME ON MY LINKEDIN HANDLE:
https://www.linkedin.com/in/harshita-chadha-1b8576163/