Final Report

12 june 2002

THE PROBLEM

When people have to choose an university, they have lots of criterias : quality of studies, city of the university, costs,...

But people also like to eat. Everyone like to have a good meal at the restaurant. What will be analyzed in this report will be the quality of the restaurants around each university. And people don't have all the same tastes, so a distinction between each types of restaurants will be made.

We will answer the two following questions :

- What university in US is in the best area for good cuisine ?
- Which is the best place to stay for each type of cuisine ?

DATA

We will need to find a list of universities in the US. And we also need the coordinates (latitude /longitude) of each one. Those informations can be found on the following website that will be parsed :

- https://www.latlong.net/category/universities-236-47.html

To have a list of the restaurants for each area, we will use the Foursquare API. It will give us the list of venues for each area :

- https://foursquare.com

METHODOLOGY

Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why

First, we have to analyze and understand the problem. That part has been done in the first chapter of this report. After that, we did the following steps.

Acquiring data

Universities

First, the list of the universities has been collected. In order to do that, a web page has been parsed with BeautifulSoup. We got a list of 50 universities in the USA.

```
# Parse list universities
url = "https://www.latlong.net/category/universities-236-47.html"
soup = BeautifulSoup(requests.get(url).text, "html.parser")
table = soup.find("table")
# Create dataframe universities
df_uni = pd.read_html(str(table))[0]
df_uni.columns = ['name', 'lat', 'lng']
df_uni['name'] = df_uni['name'].str.split(',').str[0]
print(df_uni)
```

```
                                          name        lat         lng
 0                       Johns Hopkins University  39.328888  -76.620277
 1                         Utah Valley University  40.277779 -111.713890
 2                Indiana University of Pennsylvania  40.617001  -79.160004
 3                         University of Illinois  40.110558  -88.228333
 4              Massachusetts Institute of Technology  42.360001  -71.092003
```

Restaurants

The next step is to get the list of restaurants around each university. With the Foursquare API, it's easy to get that list. Because of the limits of the free account (500 premium call per day), we limit the radius to 300m and the amount of universities to the 15 first.

	id	university	name	category	price_category	rating	lat	lng
0	5c55cc12b9a5a8002ccbfafa	Johns Hopkins University	Sakoon Indian Fusion Restaurant		1	2	39.326053	-76.615654
1	4ad4c017f964a52041f020e3	Johns Hopkins University	Niwana Restaurant	Sushi Restaurant	1	2	39.328017	-76.617214
2	4b719093f964a520d54d2de3	Johns Hopkins University	Tamber's Restaurant	Diner	1	2	39.329069	-76.615864
3	4adf518bf964a520677921e3	Johns Hopkins University	Gertrude's	American Restaurant	1	2	39.326216	-76.618815
4	4f3f07bfe4b00d12eec35c41	Utah Valley University	Restaurant Forte	Restaurant	1	2	40.279333	-111.716264

Data cleaning

So now we have a nice table of restaurants. The next task to do is to clean the data.

For the exercice, we don't want to use restaurants with no category, price_category or rating Since we can't easily predict those variables, all rows without those data will be removed.

KMeans

All rows from the same universities were grouped. And after that, we made a new dataframe that will be used to KMeans :

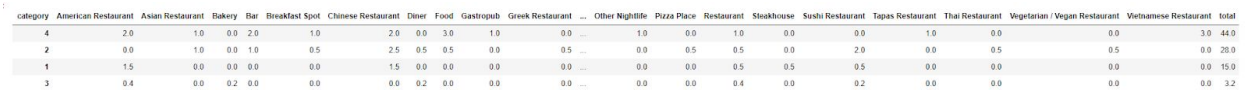
- Rows : university
- Columns : categories of restaurants

category university	American Restaurant	Asian Restaurant	Bakery	Bar	Breakfast Spot	Chinese Restaurant	Diner	Food	Gastropub	Greek Restaurant	...	Other Nightlife	Pizza Place	Restaurant	Steakhouse	Sushi Restaurant	Tapas Restaurant	Thai Restaurant	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	total
Harvard University	2.0	1.0	0.0	2.0	1.0	2.0	0.0	3.0	1.0	0.0	...	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	3.0	22.0
Indiana University of Pennsylvania	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Johns Hopkins University	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	3.0
Lawrence University	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Massachusetts Institute of Technology	1.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	...	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	7.0

KMeans will now be used, with 4 differents clusters

RESULTS

We see that group 4 is the best. So all universities in the group 4 have the most restaurants



category	American Restaurant	Asian Restaurant	Bakery	Bar	Breakfast Spot	Chinese Restaurant	Diner	Food	Gastropub	Greek Restaurant	...	Other Nightlife	Pizza Place	Restaurant	Steakhouse	Sushi Restaurant	Tapas Restaurant	Thai Restaurant	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	total
4	2.0	1.0	0.0	2.0	1.0	2.0	0.0	3.0	1.0	0.0	...	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	3.0	44.0
2	0.0	1.0	0.0	1.0	0.5	2.5	0.5	0.5	0.0	0.5	...	0.0	0.5	0.5	0.0	2.0	0.0	0.5	0.5	0.0	28.0
1	1.5	0.0	0.0	0.0	0.0	1.5	0.0	0.0	0.0	0.0	...	0.0	0.0	0.5	0.5	0.5	0.0	0.0	0.0	0.0	15.0
3	0.4	0.0	0.2	0.0	0.0	0.0	0.2	0.0	0.0	0.0	...	0.0	0.0	0.4	0.0	0.2	0.0	0.0	0.0	0.0	3.2

	university	group
0	Harvard University	4
1	Indiana University of Pennsylvania	3
2	Johns Hopkins University	3
3	Lawrence University	3
4	Massachusetts Institute of Technology	1
5	Penn State University	1
6	University of Chicago	3
7	University of Illinois	2
8	University of Michigan	2
9	Utah Valley University	3

DISCUSSION and CONCLUSION

We were able to group universities into 4 groups, and see which one is the best. To go further, we have to take all the ratings and use them to have more data to improve our system.

Because having lots of restaurants doesn't mean they are perfect.