Short Analysis of Homework 4

Hang Xu (U51199140)

First, I took a look at the raw data and have a sense of what we have. We have two major categories that one is countries along with regions which in a scale of states in the world. Another one is ingredients which each country in different states could have different ingredients used by them. Therefore, it is naturally for me to dig a bit more about the difference of ingredients used in America and China, given my background that I am an international student from China studying in America.

Hypothesis:

1.**China and America would be the major ingredients providers in East Asian and North American.**

2.More ingredients used from American than that from China, since YY is an American based company.

3.YY uses most ingredients from America since it is an American based company.

Parsing the data and store them into a matrix where it is a 'different regions' x 'different ingredients' matrix for me to use. For example, the first entry would be the number of that specific ingredient appears in the recipe which belongs to that specific country.

Next, I do a analysis of what would the most popular ingredient in North America and East Asian. I found in North America, it is eggs and in East Asia, it is soy sauce. It is exactly like what I personally expect since in Boston, I eat eggs almost for every meal. There are so many different ways to cook eggs here. And back home, almost every dish would need soy sauce to be a ingredient and it is really important in Chinese cuisine. Frankly, I am not surprised by the result.

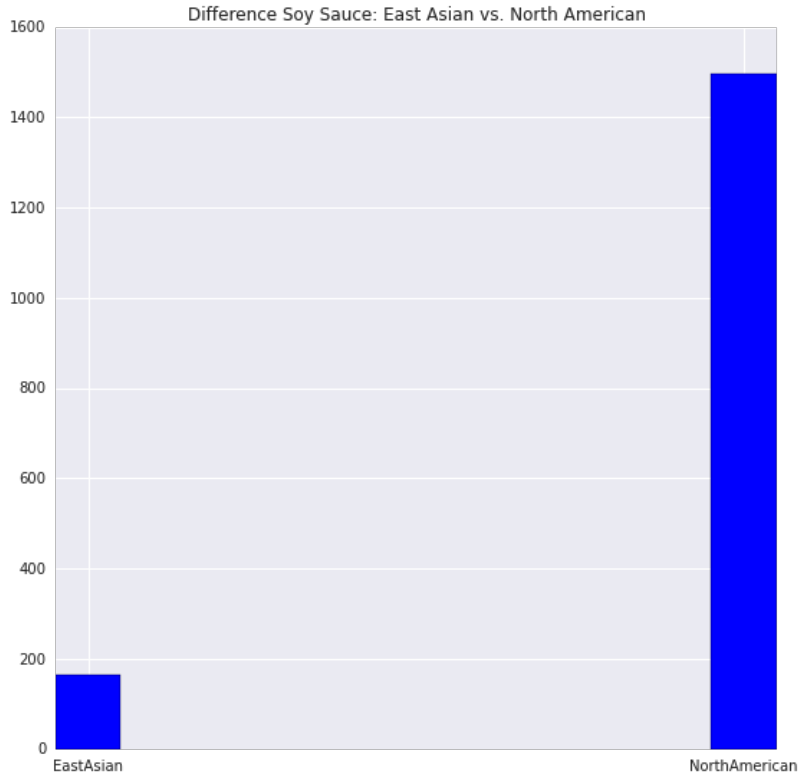
	North American	East Asian
Eggs	14802	164
Soy_Sauce	1498	84

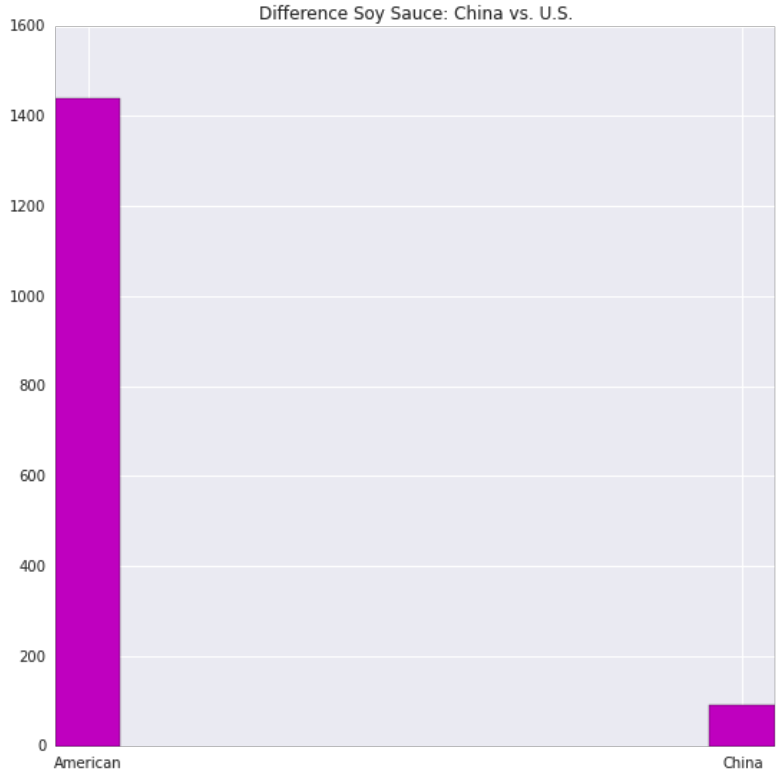
	American	China
Eggs	14528	92
Soy_Sauce	1441	51
Percentage_Eggs	14528/14802 = **98%**	92/164 = **56%**
Percentage_SS	1441/1498= **96%**	51/84 = **60%**

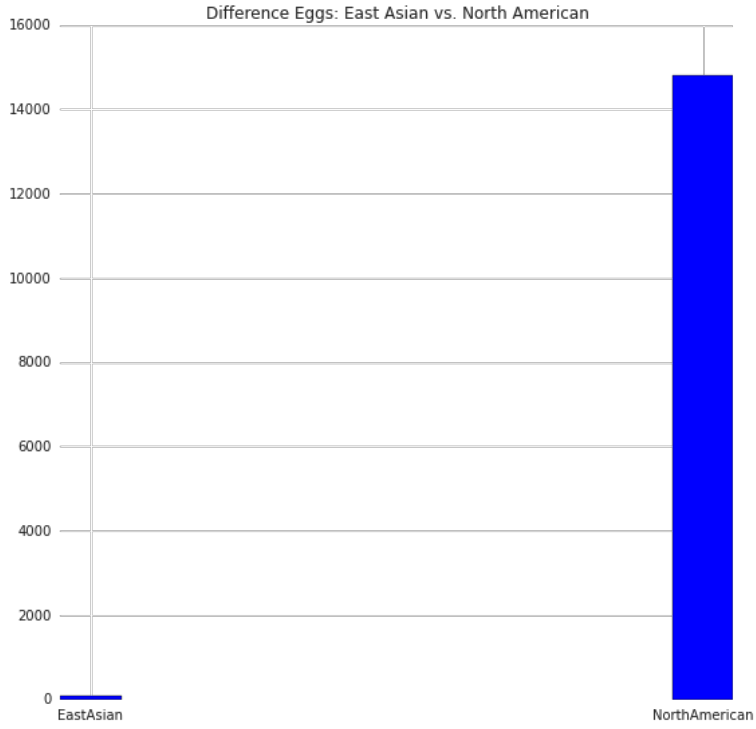
From the result above, we can see that America is the major supplier for eggs in North American. However, China supply a bit more than half of eggs among other countries in East Asia. So, for eggs, we could say American in the major supplier and China is an important but not major supplier from East Asia

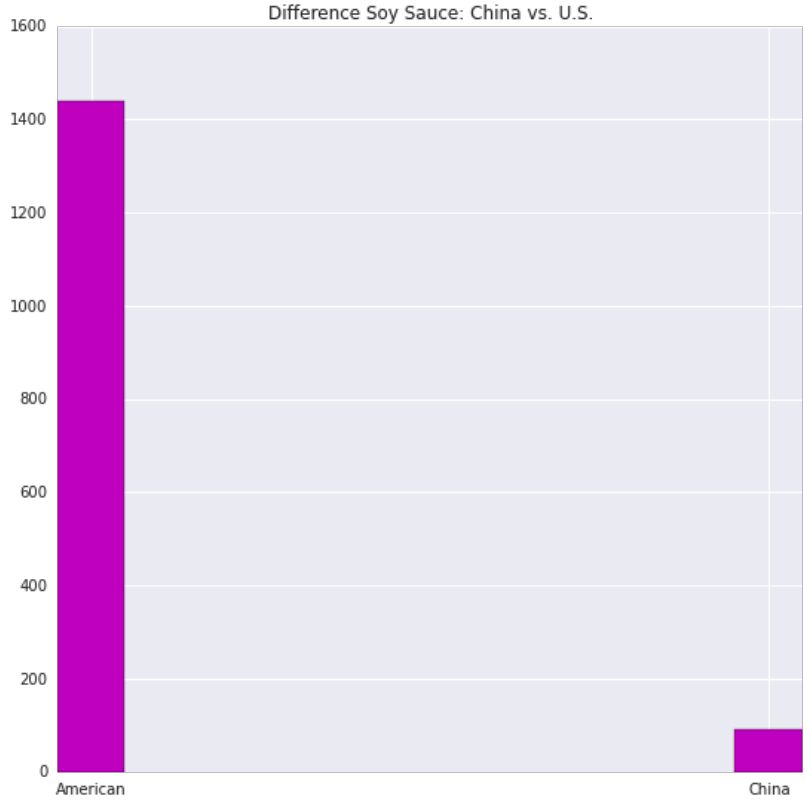
For soy sauce, China supply 60% among all other countries in East Asia and America provides 96%. Again, even for the ingredients that China uses most, America is a major supplier for YY and China is an important supplier.

Charts

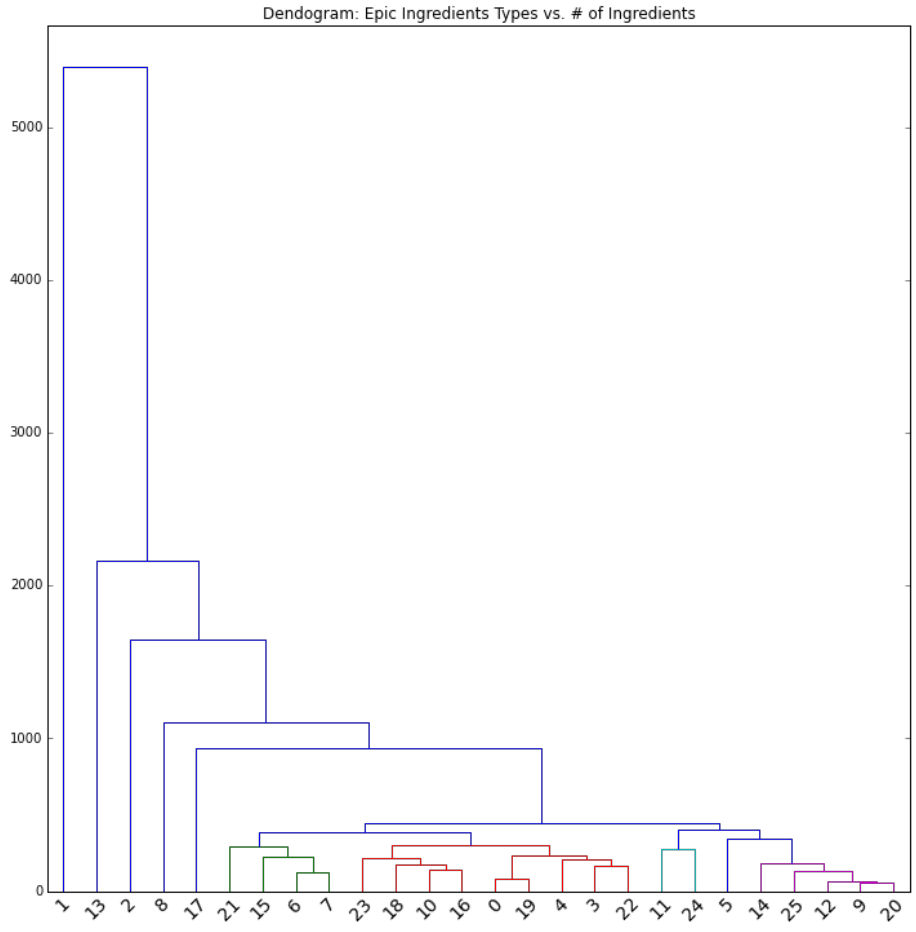

Difference Soy Sauce: East Asian vs. North American
1600
1400
1200
1000
800
600
400
200
0
EastAsian
NorthAmerican

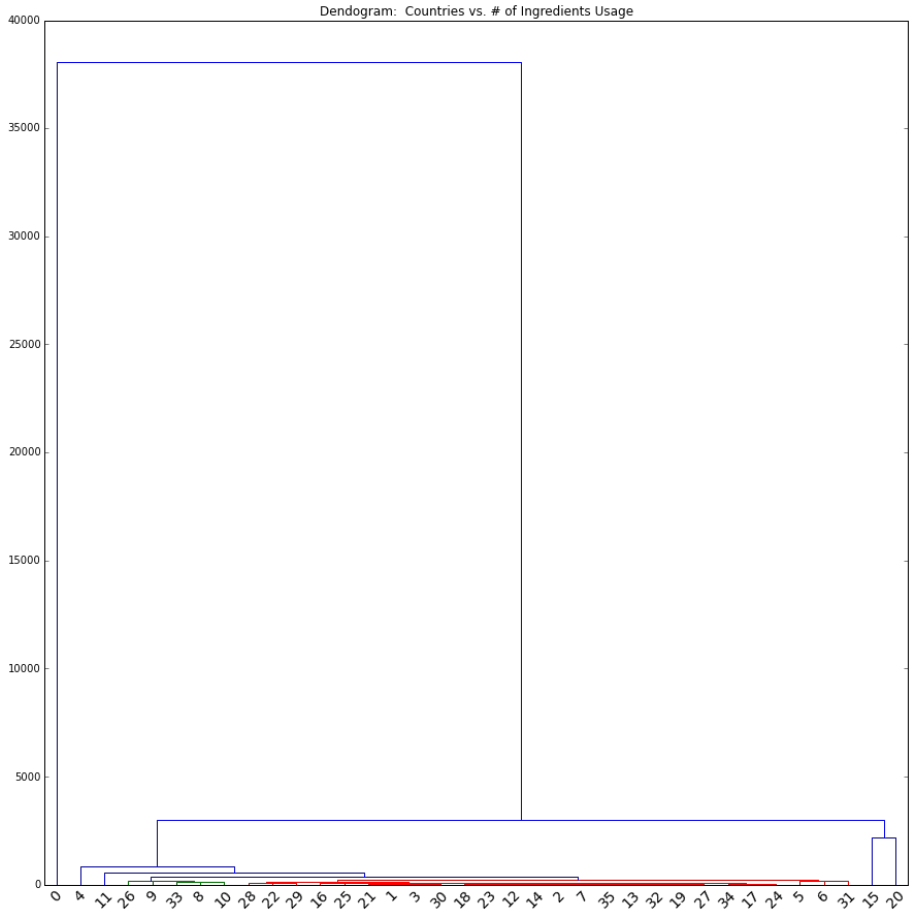

Difference Soy Sauce: China vs. U.S.
1600
1400
1200
1000
800
600
400
200
0
American
China


Difference Eggs: East Asian vs. North American
16000
14000
12000
10000
8000
6000
4000
2000
0
EastAsian
NorthAmerican


Difference Soy Sauce: China vs. U.S.
1600
1400
1200
1000
800
600
400
200
0
American
China

Then, I perform a hierarchical clustering to see the number of times that the used ingredients are from which specific region or country. And the usage of the epic recipes from the epic-recipes.txt file.

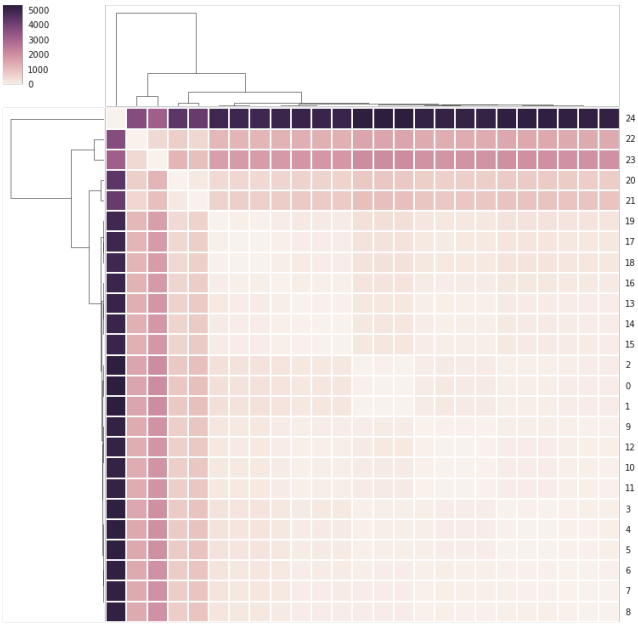


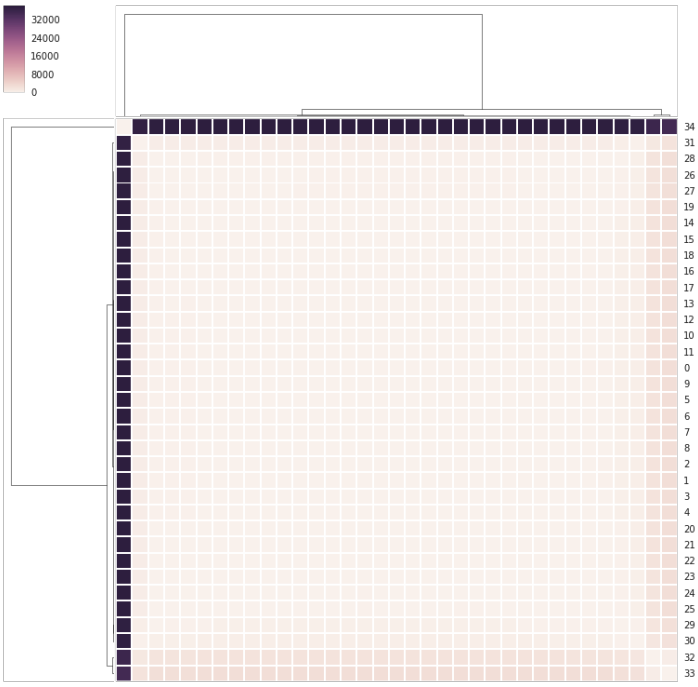


From the graph, you can tell that America as a country has most ingredients used by YY and American as a receipt also used most by YY. It is reasonable since YY is an American restaurant and it targets mostly to American customers.

In conclusion, we have three hypothesis and by doing clustering and visualization, we have proved that all of our hypotheses are correct.

Additional Charts:


5000
4000
3000
2000
1000
0
24
22
23
20
21
19
17
18
16
13
14
15
2
0
1
9
12
10
11
3
4
5
6
7
8


32000
24000
16000
8000
0
34
31
28
26
27
19
14
15
18
16
17
13
12
10
11
0
9
5
6
7
8
2
1
3
4
20
21
22
23
24
25
29
30
32
33