Chocolate Bar 2020 Project

Yifei Dong

DSBA 6520 – Network Science

Jun 11th 2021

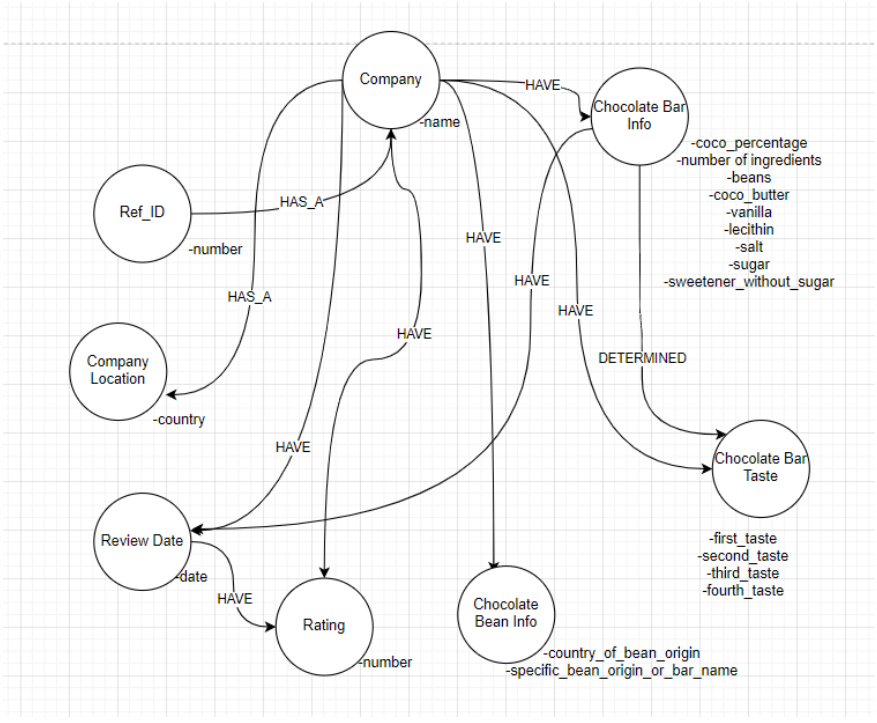
Professor Robinson

As a Chocolate shop manager at the South Park Mall, we import and sell chocolates from all over the world from hundreds of different companies. However, after Covid-19 has started, the sales have been declining. To provide a better shopping experience and great quality for our customers, I will use graph analytics to identify few questions below using the Chocolate Bar dataset:

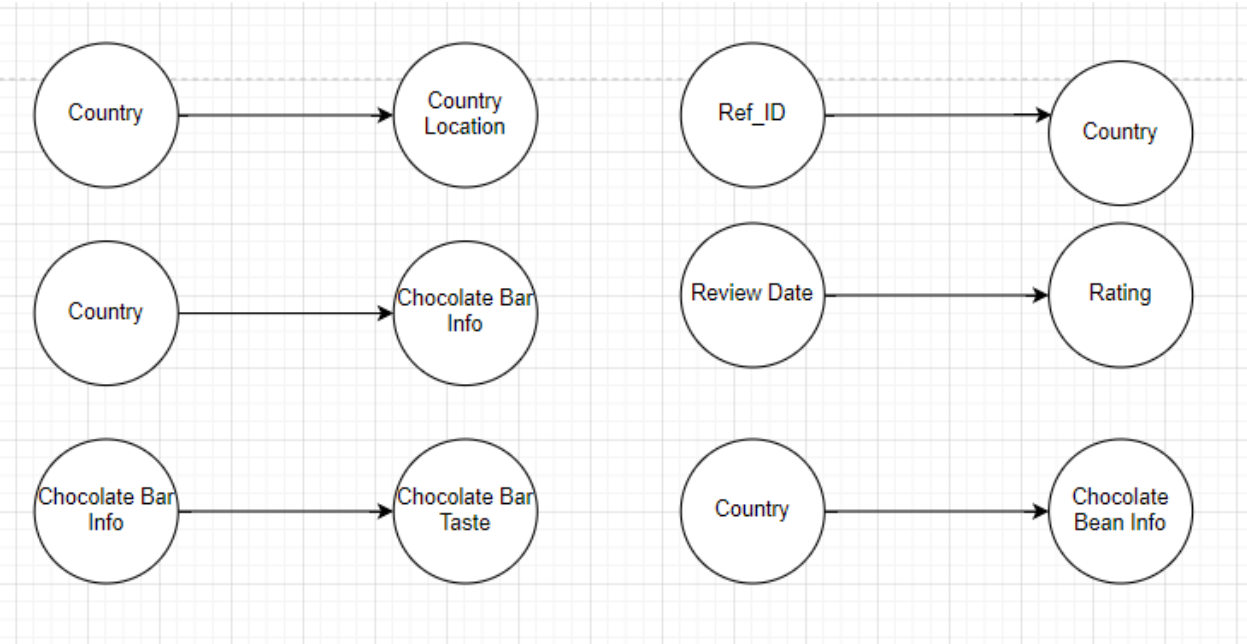
- Where are the best cocoa beans grown?
- Which countries produce the highest-rated chocolate bars?
- Which company has the highest rate?
- What is the most popular taste?

This dataset contains 21 columns and 2225 rows. The first question can help me find out where to purchase the best quality cocoa beans for my company to produce a high-quality chocolate bar. The second and third questions can help me identify which country and company I should import my product from that are most populated and received the highest rating from customers. The last question can help me determine what kind of variety my chocolate bar shop should have to target a wider range of customers. I will use the dataset to answer these questions and determine what chocolate will be purchased from which company in the next year.

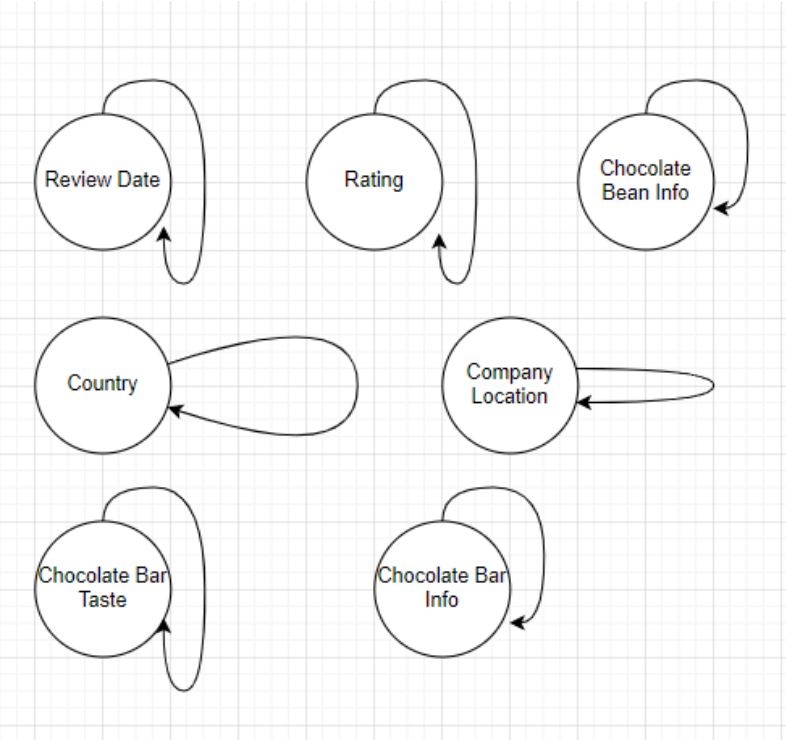
<u>Graph data model</u>



Bi-partite graph

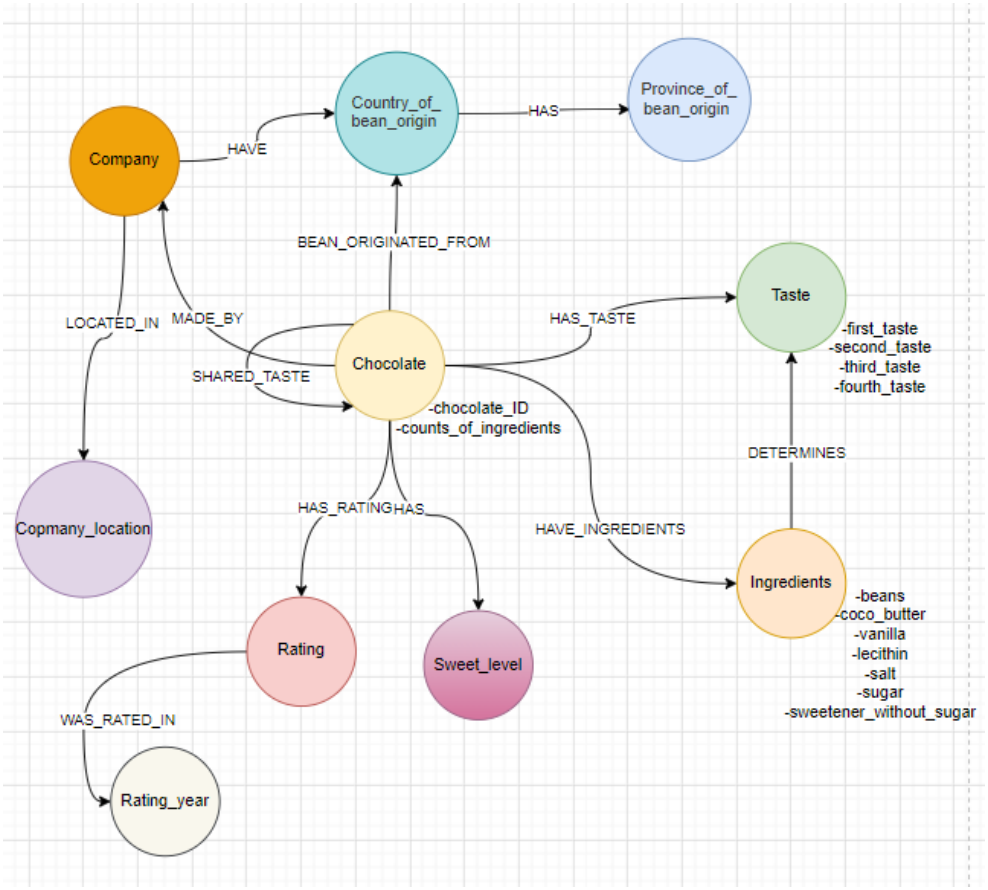



Mono-partite graph



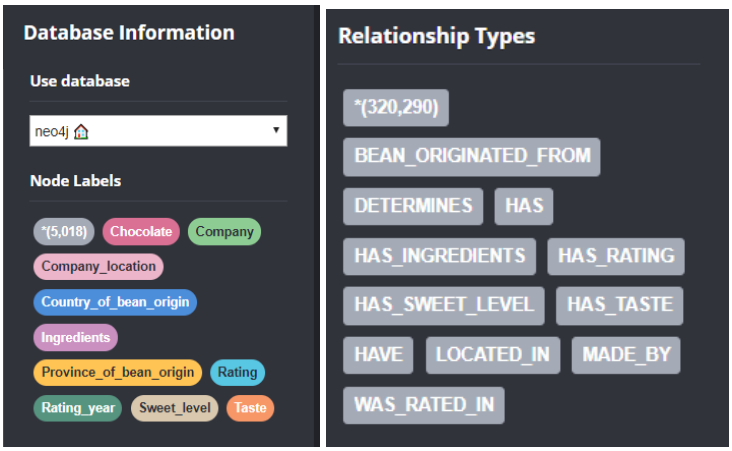
***Part II***

The updated version of the Data graph model:




The reason why I updated my graph model this way is because I grouped several ranges of cocoa percentage into sweet levels for better understanding. Ingredients and Taste nodes contain unique values that are connected with the chocolate node through chocolate ID. I have also connected company with country of bean origin for algorithm purpose.

Neo4j Database Setup and Screenshots



Cypher Queries

1. This cypher query looks for companies' name which are in the U.S.A and has a chocolate rating of 4. This information can help me to identify which company has the highest rating that is in the United State, so the shipping cost will be minimized when I purchase their products.

```
MATCH (l:Company_location{name:'U.S.A'})-[r3:LOCATED_IN]-(c:Company)-
[r1:MADE_BY]-(a:Chocolate)-[R2:HAS_RATING]-(r:Rating{rating:'4'})
RETURN DISTINCT c.name AS Copmany,l.name AS Location, r.rating AS Rating
LIMIT 10
```

Result table:

| Copmany | Location | Rating |
|---|---|---|
| Ruket | U.S.A | 4 |
| Sjolinds | U.S.A | 4 |
| Meadowlands | U.S.A | 4 |
| Mutari | U.S.A | 4 |
| Public Chocolatory | U.S.A | 4 |
| Exquisito | U.S.A | 4 |
| Friis Holm | U.S.A | 4 |
| Escazu | U.S.A | 4 |
| Argencove | U.S.A | 4 |
| Urzi | U.S.A | 4 |

2. This cypher query can determine which company makes the highest rating chocolate flavor and its sweet level.

```
MATCH (c:Company)-[r1:MADE_BY]-(a:Chocolate)-[:HAS_TASTE]-(t:Taste)
MATCH (s:Sweet_level)-[:HAS_SWEET_LEVEL]- (a:Chocolate)-[R2:HAS_RATING]-
(r:Rating{rating:'4'})
RETURN DISTINCT c.name AS Company,t.Taste AS Taste,s.Sweet_level AS Sweet_Level,r.
rating AS Rating LIMIT 10
```

Result table

| Company | Taste | Sweet_Level | Rating |
|---|---|---|---|
| Pralus | fourth_taste-lemon | Extra-Bittersweet | 4 |
| Pralus | third_taste-ashey | Extra-Bittersweet | 4 |
| Pralus | second_taste-burnt | Extra-Bittersweet | 4 |
| Pralus | first_taste-creamy | Extra-Bittersweet | 4 |
| Frederic Blondeel | first_taste-banana | Bittersweet | 4 |
| Shattell | second_taste-perfume | Extra-Bittersweet | 4 |
| Shattell | first_taste-strong spice | Extra-Bittersweet | 4 |
| Shattell | third_taste-roasty | Extra-Bittersweet | 4 |
| Danta | second_taste-floral | Bittersweet | 4 |
| Danta | first_taste-sweet spice | Bittersweet | 4 |

3. This cypher query answers my first use case question, and it displayed cocoa beans from these countries have the highest rating and have the best quality cocoa beans.

```
MATCH (r:Rating{rating:'4'})-[r2:HAS_RATING] -(a:Chocolate)-
[:BEAN_ORIGINATED_FROM]-(b:Country_of_bean_origin)
RETURN DISTINCT b.name AS Bean_origin,r.rating AS Rating
LIMIT 10
```

Result table

| Bean_origin | Rating |
|---|---|
| Principe | 4 |
| Bolivia | 4 |
| St.vincent-grenadines | 4 |
| Blend | 4 |
| U.s.a. | 4 |
| Samoa | 4 |
| Puerto rico | 4 |
| Belize | 4 |
| Haiti | 4 |
| Vietnam | 4 |

Graph Algorithms

1. Louvain Community Algorithm- These three queries return each country of bean origin name and the ID of the community to which it belongs. In the result tables, it showed 512 countries belong to the same community. This could be that these counties are located near each other.

```
CALL gds.graph.create('origin-related-
entities', ['Company', 'Company_location', 'Country_of_bean_origin', 'Province_of_bean
_origin'], '*')

CALL gds.louvain.stream('origin-related-entities')
YIELD nodeId, communityId
RETURN gds.util.asNode(nodeId).name AS Country, communityId
ORDER BY communityId DESC
```

Partial result table

| Country | communityId |
|---|---|
| Nicaragua | 1341 |
| Brazil | 1340 |
| Italy | 1339 |
| Canada | 1338 |
| U.S.A | 1337 |
| France | 1337 |
| Fiji | 1337 |
| Vanuatu | 1337 |

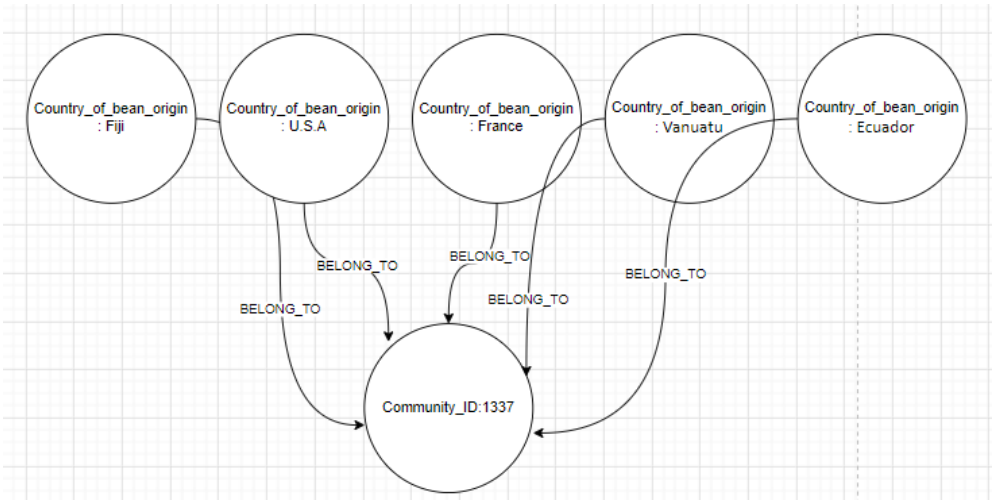
| | |
|---|---|
| Ecuador | 1337 |

```
CALL gds.louvain.stream('origin-related-entities')
YIELD nodeId, communityId
RETURN communityId, COUNT(DISTINCT nodeId) AS members
ORDER BY members DESC
```

Partial result table

| communityId | members |
|---|---|
| 1337 | 512 |
| 573 | 72 |
| 574 | 1 |
| 575 | 1 |
| 576 | 1 |

Projection Graph




2. *Updated*: PageRank-This query showed the top 10 Chocolate PageRank, and which chocolate most important one. From the result table we can see chocolate ID 101 have the highest PageRank score, which means chocolate ID 101 have the highest number and quality of links to other nodes and is the most important node. Chocolate 101 also has the most common taste and share taste with 676 other chocolates.

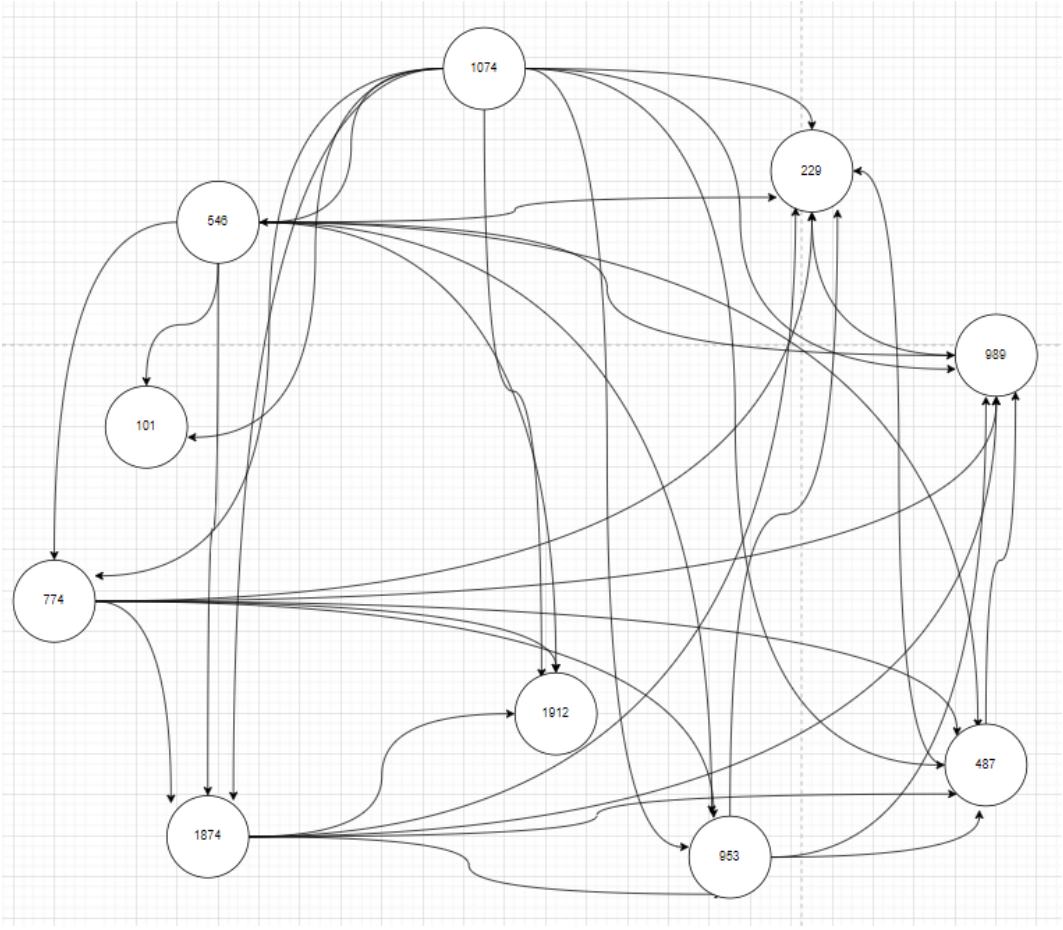
```
CALL gds.graph.create('Taste-graph','Chocolate','SHARE_TASTE')

CALL gds.pageRank.stream('Taste-graph') YIELD nodeId, score AS pageRank
WITH gds.util.asNode(nodeId) AS n, pageRank
MATCH (n)-[i:SHARE_TASTE]-()
RETURN n.chocolate_ID AS c_ID, pageRank, count(i) AS interactions
```

```
ORDER BY interactions DESC LIMIT 10
```

Result table

| c_ID | pageRank | interactions |
|---|---|---|
| 101 | 2.662793 | 676 |
| 546 | 2.399912 | 592 |
| 1074 | 2.399912 | 592 |
| 1874 | 2.399912 | 592 |
| 229 | 2.399912 | 592 |
| 774 | 2.417375 | 590 |
| 953 | 2.332458 | 584 |
| 1912 | 2.415265 | 578 |
| 487 | 2.351179 | 572 |
| 989 | 2.351179 | 572 |

Projection Graph




3. Betweenness Centrality- this query finds the node that serves as a bridge from one part of a graph to another. As the result tables show, 'have_bean' has the highest score, which tells me

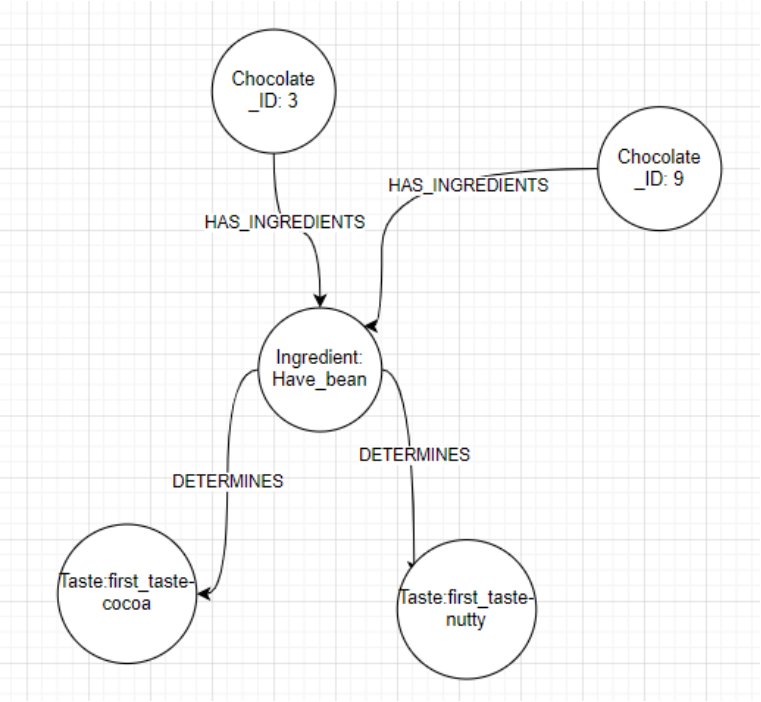
that the cocoa bean is the most important ingredient when making a chocolate bar, and all other tastes including the ingredient of the cocoa bean.

```
CALL gds.graph.create('Chocoate-taste-related-
entities', ['Chocolate', 'Ingredients', 'Taste', 'Sweet_level'], '*')
CALL gds.betweenness.stream('Chocoate-taste-related-entities') YIELD nodeId, score
RETURN gds.util.asNode(nodeId).Ingredients AS Ingredients, score
ORDER BY score DESC LIMIT 10
```

Result table

| Ingredients | score |
|---|---|
| have_bean | 429614.7 |
| have_not_salt | 422467.6 |
| have_not_sweetener_without_sugar | 414933.3 |
| have_sugar | 413774.4 |
| have_not_vanila | 362391.7 |
| have_not_lecithin | 337088.6 |
| have_cocoa_butter | 296520.3 |
| have_not_cocoa_butter | 133094.4 |
| have_lecithin | 92526.14 |
| have_vanila | 67223 |

Projection Graph



***Part III***

Cypher Action:

1. This cypher query action can help the user quickly find companies who have a rating of 4 in a specific country. This can help minimize the shipping cost when import chocolate products.

   Search Phrase: Find companies with a rating of 4 located at $company_location
   Cypher query:

```
MATCH (l:Company_location{name:$company_location})-[r3:LOCATED_IN]-
(c:Company)-[r2:COMPANY_RATING]-(r:Rating{rating:'4'})
WHERE r.rating='4'
MATCH (l)-[r3:LOCATED_IN]-(c)-[r2:COMPANY_RATING]-(r)
RETURN l,r3,c,r2,r
LIMIT 10
```

2. This cypher query action can help users to find which company makes a certain sweet level of chocolate, this can help easily navigate when purchasing chocolate by sweet level.

   Search Phrase: Companies makes $sweet_level chocolate
   Cypher query:

```
MATCH (s:Sweet_level{Sweet_level:$sweet_level})-[r:HAS_SWEET_LEVEL]-
(c:Chocolate)-[r1:MADE_BY]-(m:Company)
RETURN s,r,c,r1,m
limit 10
```

Graph Visualization:

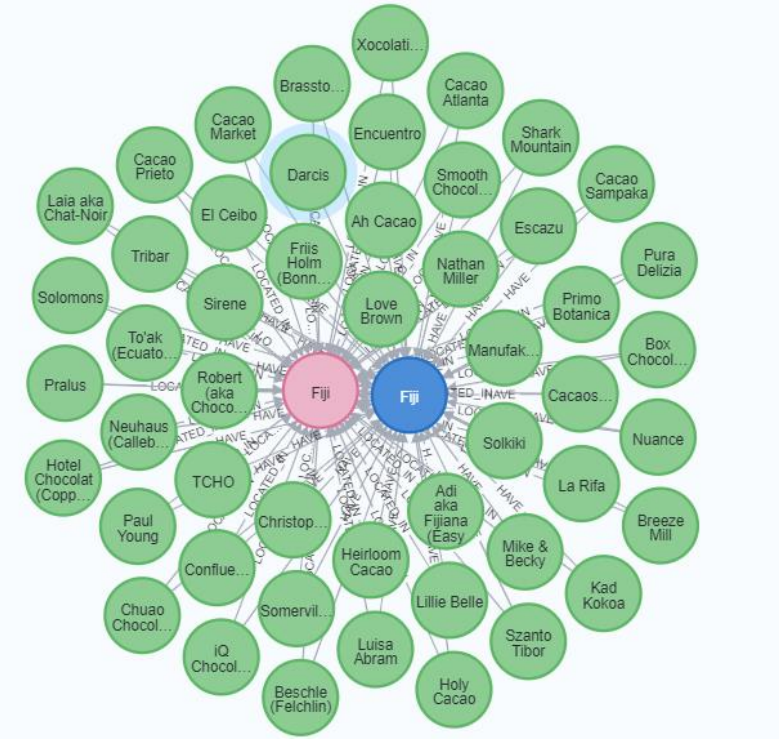
1. Find chocolate by chocolate ID with all its tastes and have a rating of 4 and all other chocolate which share similar tastes. This can help users to navigate specific tastes.

   Search Phrase: Find chocolate that shares some similar tastes by ID $chocolate_ID
   Cypher query:

```
MATCH (t:Taste)-[r:HAS_TASTE]-(c:Chocolate{chocolate_ID:$chocolate_ID})-
[r2:SHARE_TASTE]-(d:Chocolate)-[r1:HAS_TASTE]-(t2:Taste)
RETURN *
```

2. This visualization showed some companies where the company location and which country they get their cocoa bean from are the same country. This indicates that this specific country must be one of the popular places that grow cocoa beans.

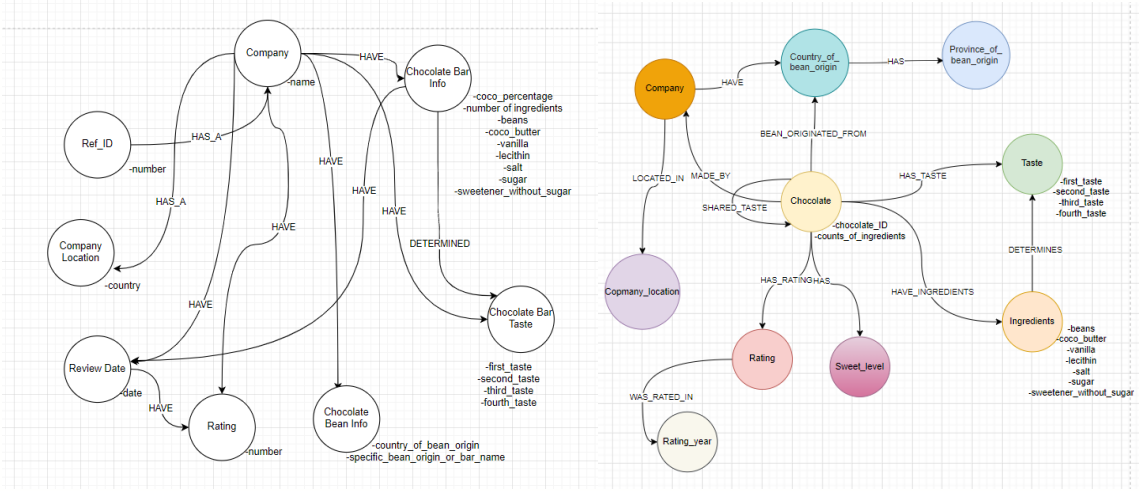
```
Match (c:Country_of_bean_origin)-[:HAVE]-(a:Company)-[:LOCATED_IN]-
(b:Company_location)
WHERE c.name=b.name
RETURN * LIMIT 50
```


Xocolati...
Brassto...
Cacao Atlanta
Cacao Market
Encuentro
Shark Mountain
Cacao Prieto
Darcis
Smooth Chocol...
Cacao Sampaka
Laia aka Chat-Noir
El Ceibo
Ah Cacao
Escazu
Tribar
Friis Holm (Bonn...
Nathan Miller
Pura Delizia
Solomons
Sirene
Love Brown
Primo Botanica
To'ak (Ecuato...
Manufak...
Box Chocol...
Pralus
Robert (aka Choco...
Fiji
Fiji
Cacaos...
Neuhaus (Calleb...
Solkiki
Nuance
Hotel Chocolat (Copp...
TCHO
La Rifa
Paul Young
Christop...
Adi aka Fijiana (Easy
Breeze Mill
Conflue...
Heirloom Cacao
Mike & Becky
Kad Kokoa
Chuao Chocol...
Somervil...
Lillie Belle
Szanto Tibor
iQ Chocol...
Luisa Abram
Holy Cacao
Beschle (Felchlin)

**Final Report**

In this project, I have made several changes throughout the way. From the first deliverable to the second deliverable, I updated my data graph model from the first to the second deliverable because, when I began to import data into Neo4j, I learned that the first data graph model did not support how I wanted my data to be connected in Neo4j.

The data graph model on the right side, as shown in the graph below, is clearer and can readily retrieve data. Importing data is the most difficult component of this project, but if the data graph model is established correctly, the process of importing data becomes much easier.




To accommodate my use case, I made a few simple changes to my queries and algorithms to present more meaningful and full information. I include each chocolate ID's taste, as well as the ones they share, in the first graph visualization to provide users a complete picture.

My use case stayed the same this entire project. From the analysis of the dataset, I found that countries that grow high-quality cocoa beans, many of them are located in South America, for instance, Peru, Ecuador, Venezuela, and Mexico. Many companies that produce high rating chocolate bars are also located in the United States. Which can tell me that the United States consumes a very large amount of chocolates each year. From the result of this analysis, I can consider partnerships with some of the domestically chocolate bar companies.

From the analysis, I also found out there are few extremely popular flavors that people love, which are ‘Creamy’, ’Nutty’, ‘Lemon’ and ‘Roasty’. However, there are hundreds of different flavor combinations in this dataset, some flavors are very unique, for instance, spicy, vegetal, and smoky. I will consider having a corner in my chocolate shop that displays some of the unique flavor chocolate bars to give my customers some different chocolate tastes experiences.

Nowadays, healthy food and a healthier lifestyle are very popular, and because we can buy high-quality raw materials in South America at lower shipping costs, this could be a great opportunity for my shop to make healthier chocolate bars with simple ingredients and an honest and transparent baking process.

Overall, this project provided me with strong core abilities in Neo4j, the basic cypher queries can be learned in a short amount of time and it is very straightforward. Exploring Bloom was a fun experience. I found being about to pre-set cypher actions is a remarkably useful feature for users who are not familiar with cypher queries. The three algorithms I used in the project brought me views beyond the surface, for example, chocolate 101 shares tastes with 676 other chocolate bars. Because of the dataset's limitations, this information does not convey many tales, but it can be highly valuable when applied to other social network datasets. I learned a new approach to explore and visualize data, which I will use in my future job and studies.