IOWA WHISKEY DISTRIBUTION

FINAL PROJECT DELIVERABLE

GitHub: https://github.com/StevenEJordan/Network_Science

STEVEN JORDAN sjorda41@uncc.edu

BUSINESS USE CASE

I am putting myself in the shoes of a Data Scientist working for a startup whiskey distillery tasked with understanding and optimizing the retail sales channels. Specifically, we would like to begin distribution in Iowa but don't know what vendor to use and which stores/areas to target with our initial launch. Our product is American made but in the Irish Whiskey tradition. This means we need a distributor who specializes in that type of product. Currently, all liquor sales go through the state who work as an intermediary between vendors and liquor stores. This means that the department can help us understand what items are selling where and what vendors are supplying them. This data will allow us to answer our business questions around distribution and marketing.

Business Questions:

1. What distributor should we work with to get our product into stores?

2. What geographic area and stores should we target first?

IOWA SALES DATA

Data Source

Our data is hosted on BigQuery in their public datasets, in the “bigquery-public-data.iowa_liquor_sales.sales” table. This table is updated monthly and has over 21.6 million records each of which is an individual product purchase. It has 24 columns that fall into the below three categories:

- Location of stores purchasing the alcohol
- Information about the product
- Metrics around quantity and price.

This dataset has many fields and records that aren’t useful for our analysis so we will need to filter and reform it a bit.

Exporting the Data - SQL Query

Since this is currently stored in a relational database, the easiest and most efficient way to transform the data is in a sql query. To accomplish this, I’ve written the below query.

```
SELECT
   upper(store_name) store_name,
   upper(city) city,
   zip_code,
   upper(county) county,
   upper(category_name) category_name,
   upper(vendor_name) vendor_name,
   upper(item description) item_description,
   sum(pack) pack,
   sum(bottle_volume_ml) bottle_volume_ml,
   round(sum(state_bottle_cost),2) state_bottle_cost,
   round(sum(state_bottle_retail),2) state_bottle_retail,
   round(sum(bottles_sold),2) bottles_sold,
   round(sum(sale_dollars),2) sale_dollars,
   round(sum(volume_sold_liters),2) volume_sold_liters,
   round(sum(volume_sold_gallons),2) volume_sold_gallons
FROM `bigquery-public-data.iowa_liquor_sales.sales`
where date >= "2021-01-01"
and upper(category_name) like "%WHISK%"
group by 1,2,3,4,5,6,7
```

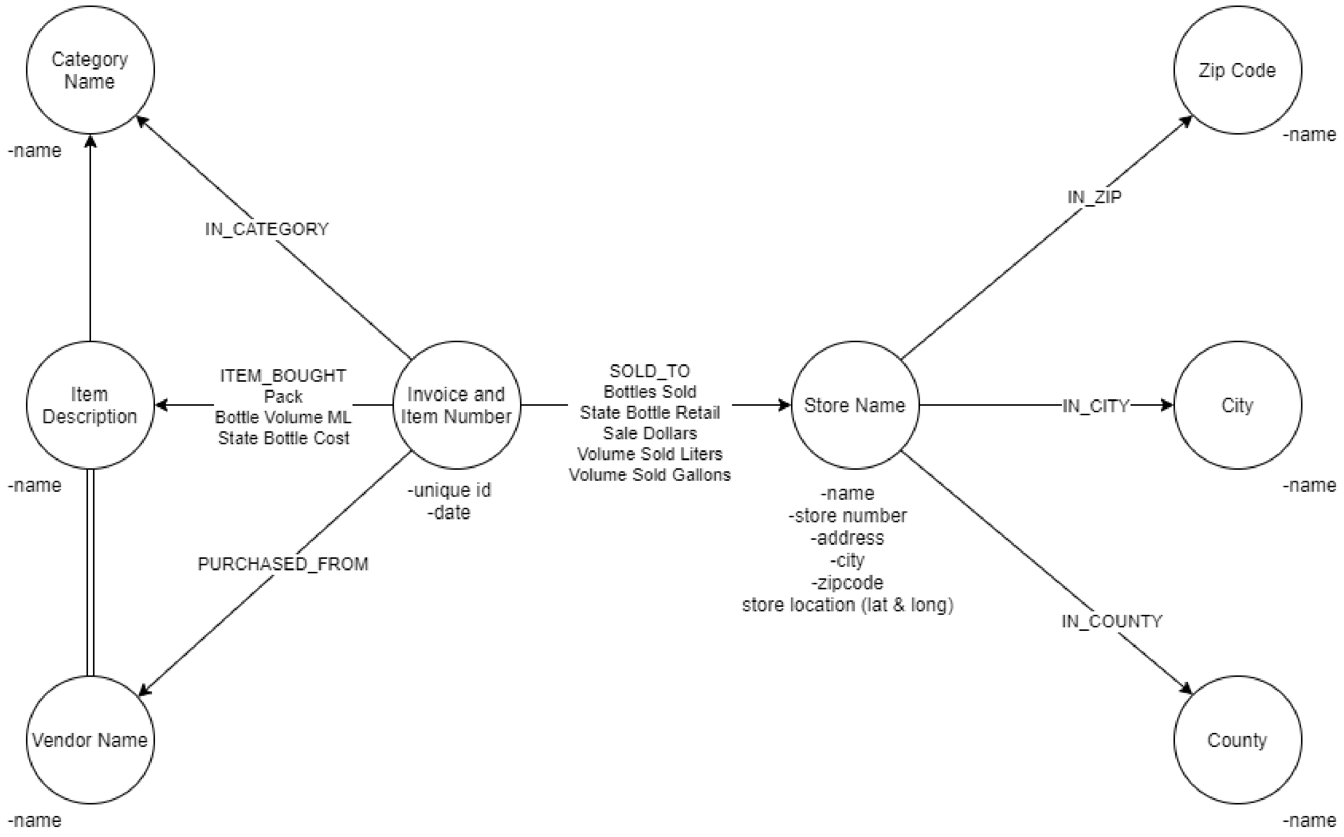
This query has limited the data to the most recent (this year) and only showing categories that are related to whiskey since this is our business. This now has ~99k records which is much more reasonable to work with. Once completed I exported the data to a csv file for its upload to Neo4J.

GRAPH DATA MODELS

Initial Data Model

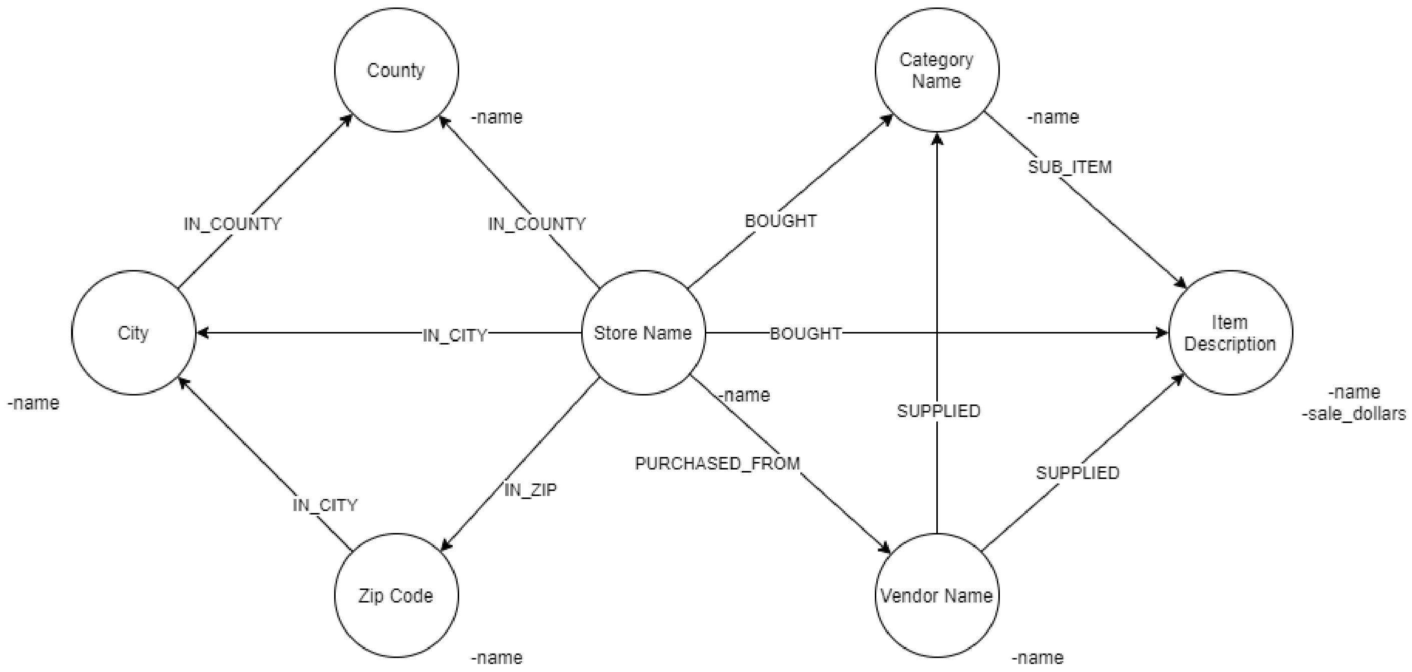
My first data model had a few issues that I discovered after creating the cypher query and inserting the data.

1. Storing the values on the relationships instead of nodes made querying more difficult.
2. The invoice and item number level of detail was not needed for my analysis. The extra nodes decreased the performance and made the visualizations more confusing.



New Data Model

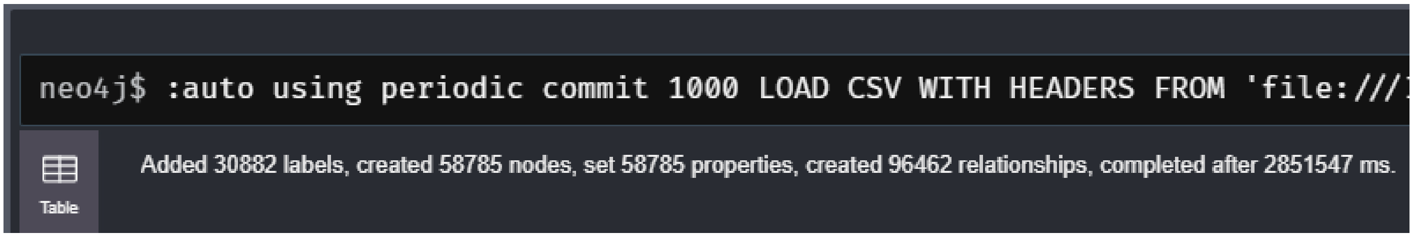
My new data model is much more agile and will allow me to focus on what specifically I'm trying to analyze. In this case, the relationships between locations, stores, vendors, and purchases.



Cypher Data Load Statement

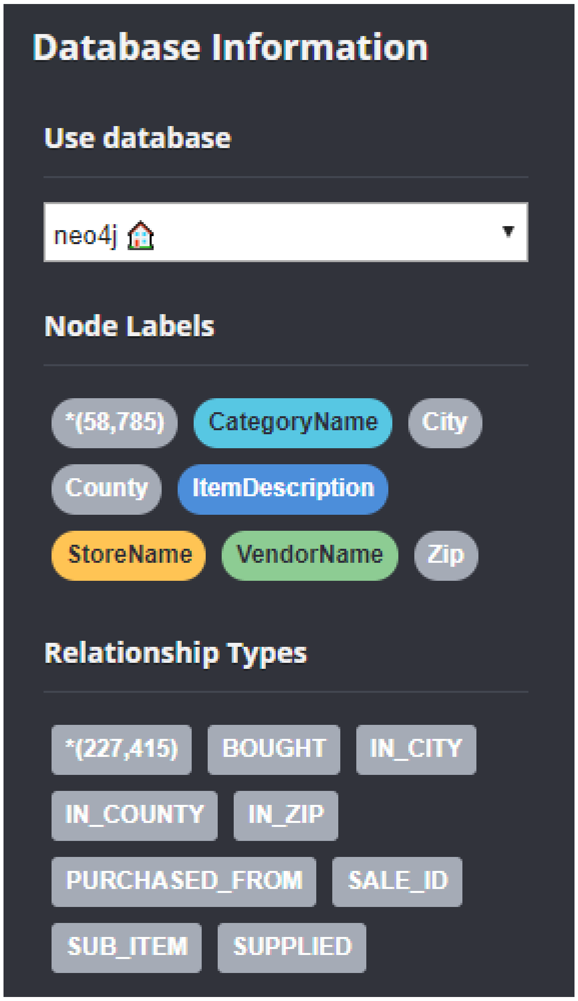
```
//CREATE CONSTRAINT ON (n:StoreName) ASSERT n.StoreName is UNIQUE;
//CREATE CONSTRAINT ON (n:City) ASSERT n.City is UNIQUE;
//CREATE CONSTRAINT ON (n:Zip) ASSERT n.Zip is UNIQUE;
//CREATE CONSTRAINT ON (n:County) ASSERT n.County is UNIQUE;
//CREATE CONSTRAINT ON (n:CategoryName) ASSERT n.CategoryName is UNIQUE;
//CREATE CONSTRAINT ON (n:VendorName) ASSERT n.VendorName is UNIQUE;
//CREATE CONSTRAINT ON (n:ItemDescription) ASSERT n.ItemDescription is UNIQUE;

:auto using periodic commit 1000
LOAD CSV WITH HEADERS FROM 'file:///Iowa_Whiskey_Sales_Clean.csv' AS line

MERGE (s:StoreName {StoreName: line.store_name})
MERGE (c:City {City: line.city})
MERGE (z:Zip {Zip: line.zip_code})
MERGE (co:County {County: line.county})
MERGE (cn:CategoryName {CategoryName: line.category_name})
MERGE (v:VendorName {VendorName: line.vendor_name})
MERGE (id:ItemDescription {ItemDescription: line.item_description, SaleDollars: line.sale_dollars})

//Sale Relationships
MERGE (s)-[:PURCHASED_FROM]->(v)
MERGE (s)-[:BOUGHT]->(cn)
MERGE (cn)-[:SUB_ITEM]->(id)
MERGE (id)-[:SALE_ID]->(i)
MERGE (v)-[:SUPPLIED]->(cn)
MERGE (v)-[:SUPPLIED]->(i)

//Location Relationships
MERGE (s)-[:IN_ZIP]->(z)
MERGE (z)-[:IN_CITY]->(c)
MERGE (c)-[:IN_COUNTY]->(co)
```

Data Load



```
neo4j$ :auto using periodic commit 1000 LOAD CSV WITH HEADERS FROM 'file:///
```

```
Added 30882 labels, created 58785 nodes, set 58785 properties, created 96462 relationships, completed after 2851547 ms.
```



Database Information

Use database

neo4j ▾

Node Labels

*(58,785) CategoryName City County ItemDescription StoreName VendorName Zip

Relationship Types

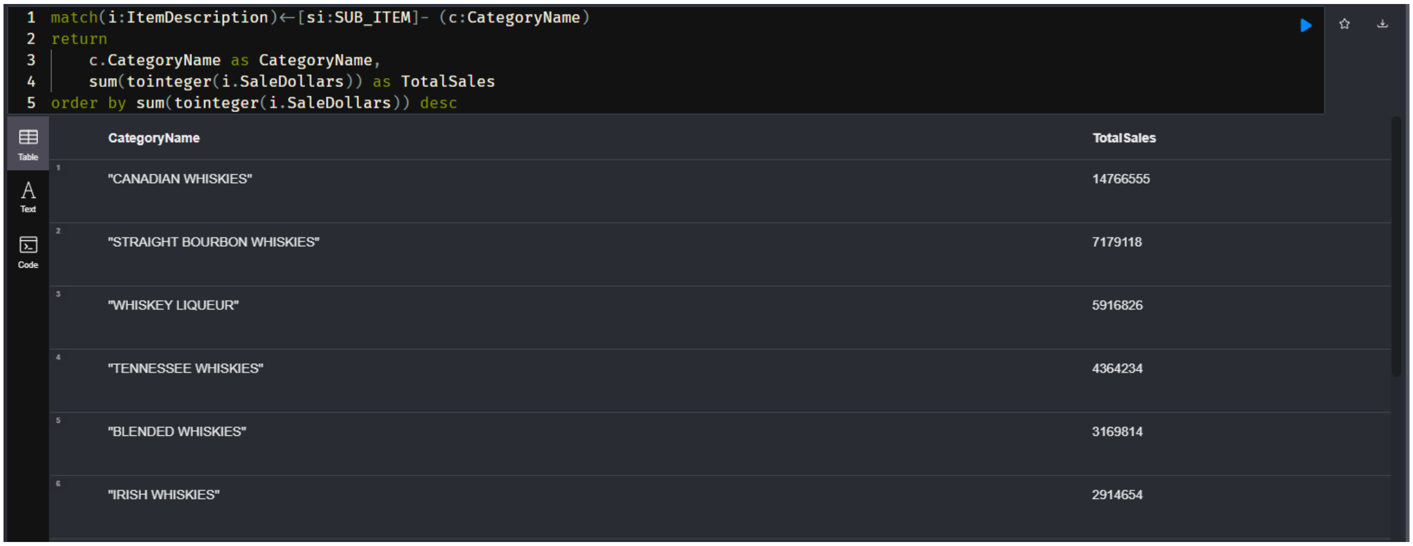
*(227,415) BOUGHT IN_CITY IN_COUNTY IN_ZIP PURCHASED_FROM SALE_ID SUB_ITEM SUPPLIED

GRAPH QUERIES

Now that we have our data in our database, we can begin exploratory queries and algorithms.

What Whiskey Categories Currently Sell the Most in Iowa?

```
match(i:ItemDescription)<-[si:SUB_ITEM]- (c:CategoryName)
return
   c.CategoryName as CategoryName,
   sum(tointeger(i.SaleDollars)) as TotalSales
order by sum(tointeger(i.SaleDollars)) desc
```



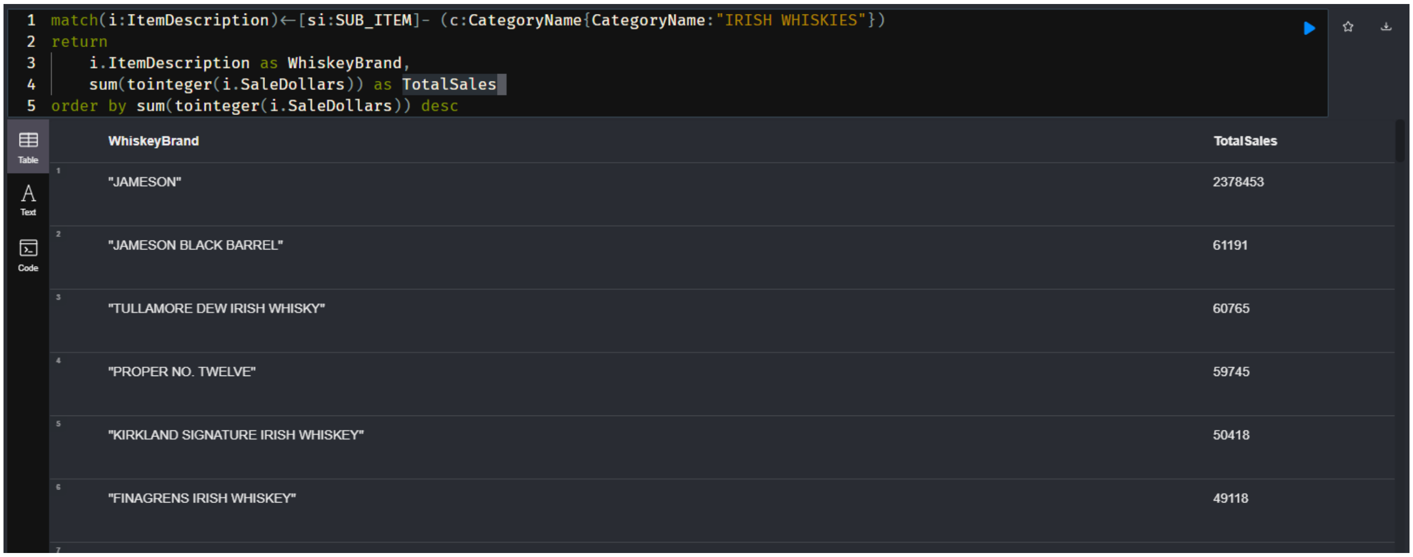
```
match(i:ItemDescription)←[si:SUB_ITEM]- (c:CategoryName)
return
    c.CategoryName as CategoryName,
    sum(tointeger(i.SaleDollars)) as TotalSales
order by sum(tointeger(i.SaleDollars)) desc
```

	CategoryName	TotalSales
1	"CANADIAN WHISKIES"	14766555
2	"STRAIGHT BOURBON WHISKIES"	7179118
3	"WHISKEY LIQUEUR"	5916826
4	"TENNESSEE WHISKIES"	4364234
5	"BLENDED WHISKIES"	3169814
6	"IRISH WHISKIES"	2914654

When we run the above, we see that Irish Whiskies (the style we make) is the sixth most popular category in Iowa with total sales of $2,914,654 YTD. This tells us there is a small market in Iowa right now so we will need to advertise how Irish Whiskies are different/similar to the more popular categories and try and educate potential customers. A different approach would be to target stores where Irish whiskies are currently selling to skip that step.

What Are the Best Selling Irish Whiskies?

```
match(i:ItemDescription)<-[si:SUB_ITEM]- (c:CategoryName{CategoryName:"IRISH WHISKIES"})
return
   i.ItemDescription as WhiskeyBrand,
   sum(tointeger(i.SaleDollars)) as TotalSales
order by sum(tointeger(i.SaleDollars)) desc
```



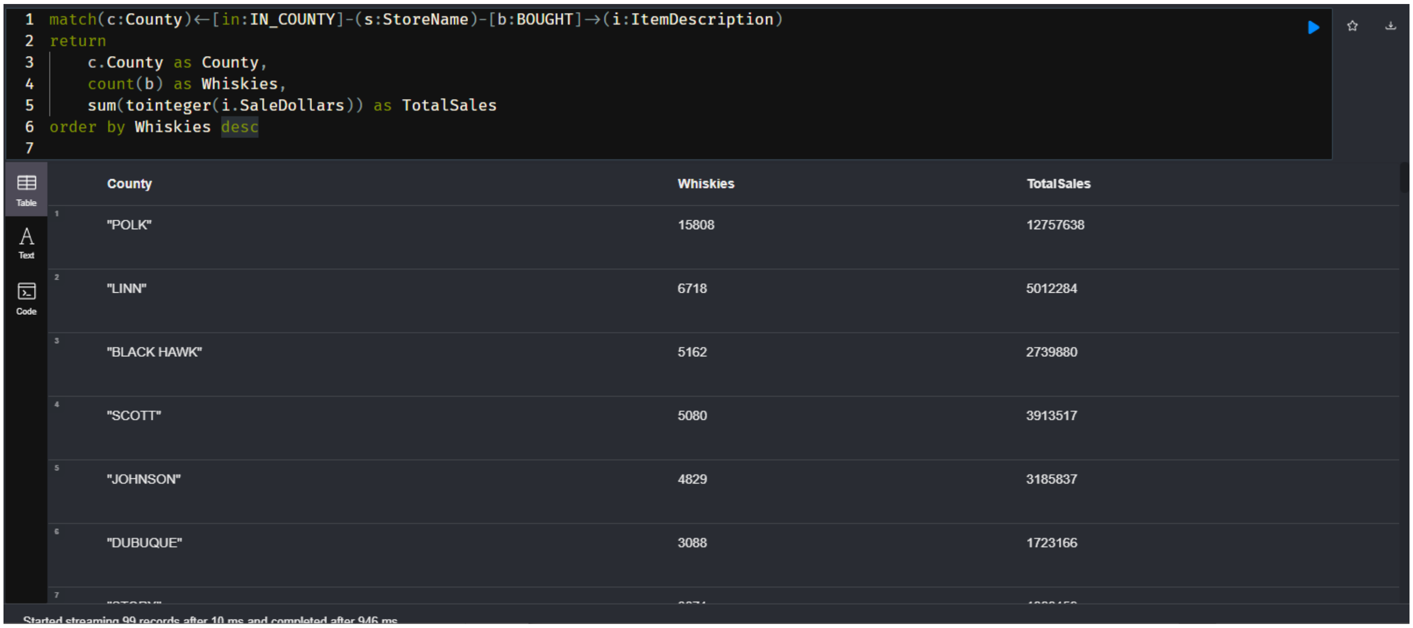
```
match(i:ItemDescription)←[si:SUB_ITEM]- (c:CategoryName{CategoryName:"IRISH WHISKIES"})
return
    i.ItemDescription as WhiskeyBrand,
    sum(tointeger(i.SaleDollars)) as TotalSales
order by sum(tointeger(i.SaleDollars)) desc
```

WhiskeyBrand	TotalSales
"JAMESON"	2378453
"JAMESON BLACK BARREL"	61191
"TULLAMORE DEW IRISH WHISKY"	60765
"PROPER NO. TWELVE"	59745
"KIRKLAND SIGNATURE IRISH WHISKEY"	50418
"FINAGRENS IRISH WHISKEY"	49118

No surprises here that Jameson is by far the biggest Irish Whiskey seller in Iowa. Accounting for over 71% of total Irish Whiskey sales. This tells us we should focus on converting this strong customer base by comparing our product to that one.

What County Buys the Most Variety of Whiskey?

```
match(c:County)<-[in:IN_COUNTY]-(s:StoreName)-[b:BOUGHT]->(i:ItemDescription)
return
    c.County as County,
    count(b) as Whiskies,
    sum(tointeger(i.SaleDollars)) as TotalSales
order by Whiskies desc
```



```
match(c:County)←[in:IN_COUNTY]-(s:StoreName)-[b:BOUGHT]→(i:ItemDescription)
return
    c.County as County,
    count(b) as Whiskies,
    sum(tointeger(i.SaleDollars)) as TotalSales
order by Whiskies desc
```

County	Whiskies	TotalSales
"POLK"	15808	12757638
"LINN"	6718	5012284
"BLACK HAWK"	5162	2739880
"SCOTT"	5080	3913517
"JOHNSON"	4829	3185837
"DUBUQUE"	3088	1723166

Started streaming 99 records after 10 ms and completed after 946 ms

This tells us that Polk county is by far the largest purchaser of whisky. This is where the city of Des Moines is, but it is interesting that they have such a wide margin as compared to the other counties in Iowa.

GRAPH ALGORITHMS

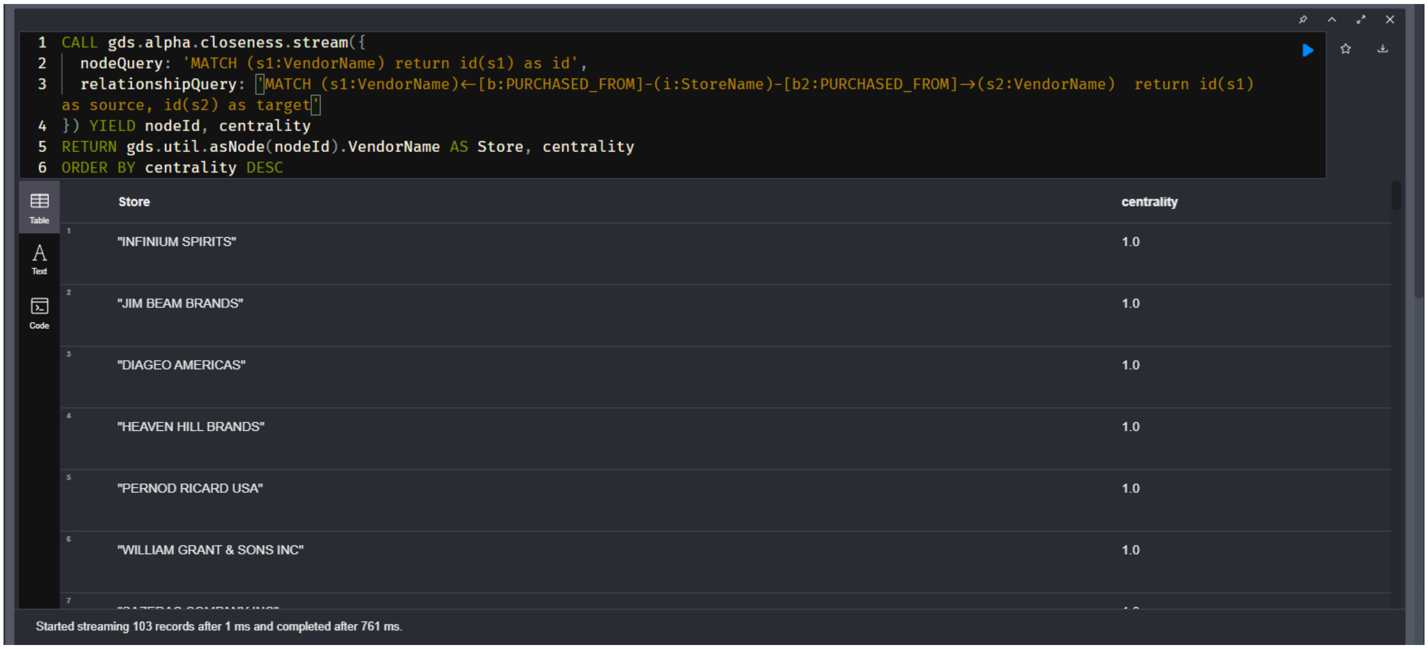
Now that we have a basic understanding of what types of whiskey are currently being sold in Iowa and where, we can move on to more advanced analysis.

Community Detection with Closeness Centrality & Label Propagation

Closeness Centrality

The purpose of this algorithm is to understand how closely connected each vendor is to the stores in the state. This will be useful as I begin to zero in on which vendors the distillery should approach about distributing our whiskey.

```
CALL gds.alpha.closeness.stream({
  nodeQuery: 'MATCH (s1:VendorName) return id(s1) as id',
  relationshipQuery: 'MATCH (s1:VendorName)<-[b:PURCHASED_FROM]-(i:StoreName)-
[b2:PURCHASED_FROM]->(s2:VendorName)  return id(s1) as source, id(s2) as target'
}) YIELD nodeId, centrality
RETURN gds.util.asNode(nodeId).VendorName AS Store, centrality
ORDER BY centrality DESC
```



```
CALL gds.alpha.closeness.stream({
  nodeQuery: 'MATCH (s1:VendorName) return id(s1) as id',
  relationshipQuery: 'MATCH (s1:VendorName)←[b:PURCHASED_FROM]-(i:StoreName)-[b2:PURCHASED_FROM]→(s2:VendorName)  return id(s1)
as source, id(s2) as target'
}) YIELD nodeId, centrality
RETURN gds.util.asNode(nodeId).VendorName AS Store, centrality
ORDER BY centrality DESC
```

	Store	centrality
1	"INFINIUM SPIRITS"	1.0
2	"JIM BEAM BRANDS"	1.0
3	"DIAGEO AMERICAS"	1.0
4	"HEAVEN HILL BRANDS"	1.0
5	"PERNOD RICARD USA"	1.0
6	"WILLIAM GRANT & SONS INC"	1.0
7	"SAZERAC COMPANY INC"	1.0

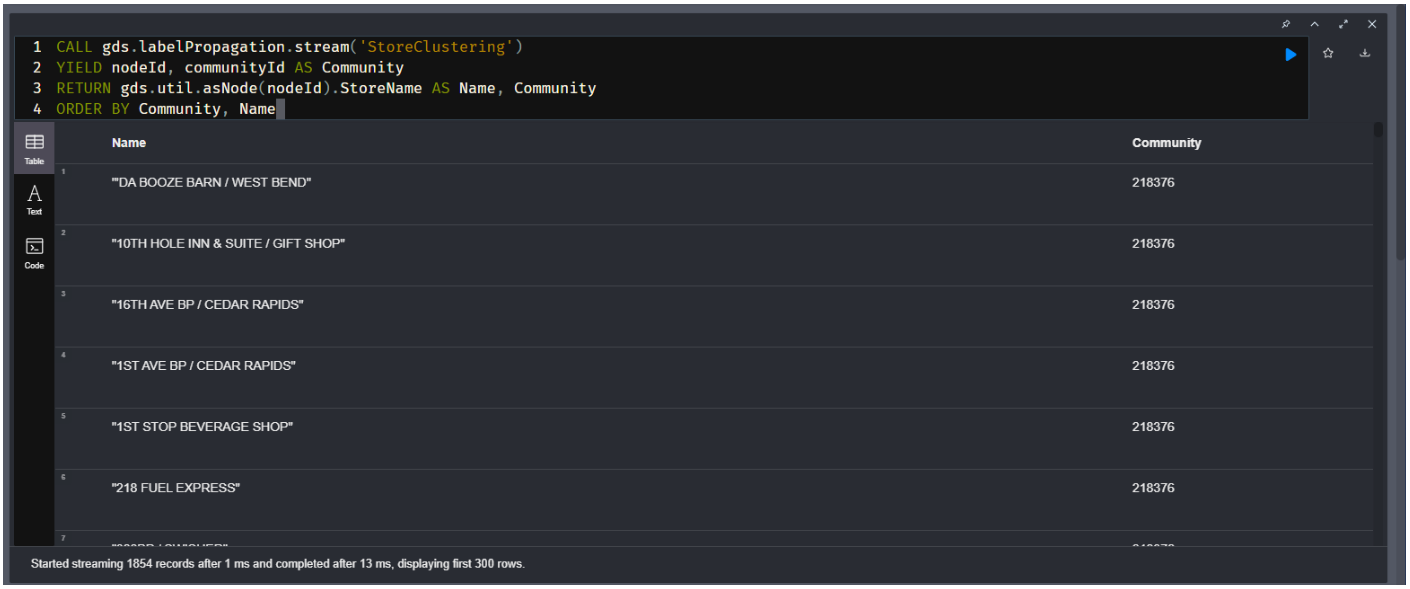
Started streaming 103 records after 1 ms and completed after 761 ms.

Label Propagation

This algorithm will help me understand what groups of stores exist based on what they purchase. This will help me understand which stores to target for distribution.

```
CALL gds.graph.create.cypher(
    'StoreClustering',
    'MATCH (s1:StoreName) return id(s1) as id',
    'MATCH (s1:StoreName)-[b:BOUGHT]->(i:ItemDescription)<-[b2:BOUGHT]-(s2:StoreName)
return id(s1) as source, id(s2) as target'
)
```

```
CALL gds.labelPropagation.stream('StoreClustering')
YIELD nodeId, communityId AS Community
RETURN gds.util.asNode(nodeId).StoreName AS Name, Community
ORDER BY Community, Name
```



```
CALL gds.labelPropagation.stream('StoreClustering')
YIELD nodeId, communityId AS Community
RETURN gds.util.asNode(nodeId).StoreName AS Name, Community
ORDER BY Community, Name
```

	Name	Community
1	"'DA BOOZE BARN / WEST BEND"	218376
2	"10TH HOLE INN & SUITE / GIFT SHOP"	218376
3	"16TH AVE BP / CEDAR RAPIDS"	218376
4	"1ST AVE BP / CEDAR RAPIDS"	218376
5	"1ST STOP BEVERAGE SHOP"	218376
6	"218 FUEL EXPRESS"	218376
7	[illegible]	[illegible]

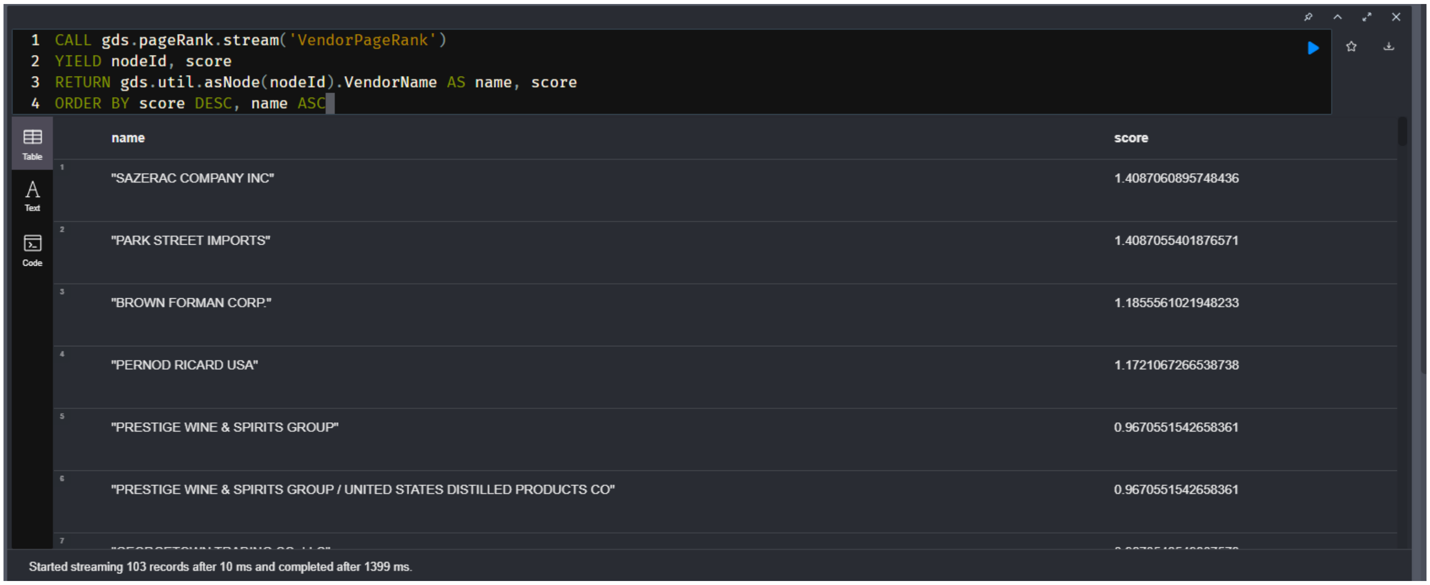
Started streaming 1854 records after 1 ms and completed after 13 ms, displaying first 300 rows.

Page Rank

The intention of this algorithm is to understand which vendors have the biggest impact on overall sales to help narrow in on the ones we should target to distribute our whiskey.

```
CALL gds.graph.create.cypher(
    'VendorPageRank',
    'MATCH (v1:VendorName) return id(v1) as id',
    'MATCH (v1:VendorName)-[s:SUPPLIED]->(i:ItemDescription)<-[s2:SUPPLIED]-
(v2:VendorName)  return id(v1) as source, id(v2) as target'
)
```

```
CALL gds.pageRank.stream('VendorPageRank')
YIELD nodeId, score
RETURN gds.util.asNode(nodeId).VendorName AS name, score
ORDER BY score DESC, name ASC
```



```
CALL gds.pageRank.stream('VendorPageRank')
YIELD nodeId, score
RETURN gds.util.asNode(nodeId).VendorName AS name, score
ORDER BY score DESC, name ASC
```

	name	score
1	"SAZERAC COMPANY INC"	1.4087060895748436
2	"PARK STREET IMPORTS"	1.4087055401876571
3	"BROWN FORMAN CORP."	1.1855561021948233
4	"PERNOD RICARD USA"	1.1721067266538738
5	"PRESTIGE WINE & SPIRITS GROUP"	0.9670551542658361
6	"PRESTIGE WINE & SPIRITS GROUP / UNITED STATES DISTILLED PRODUCTS CO"	0.9670551542658361

Started streaming 103 records after 10 ms and completed after 1399 ms.

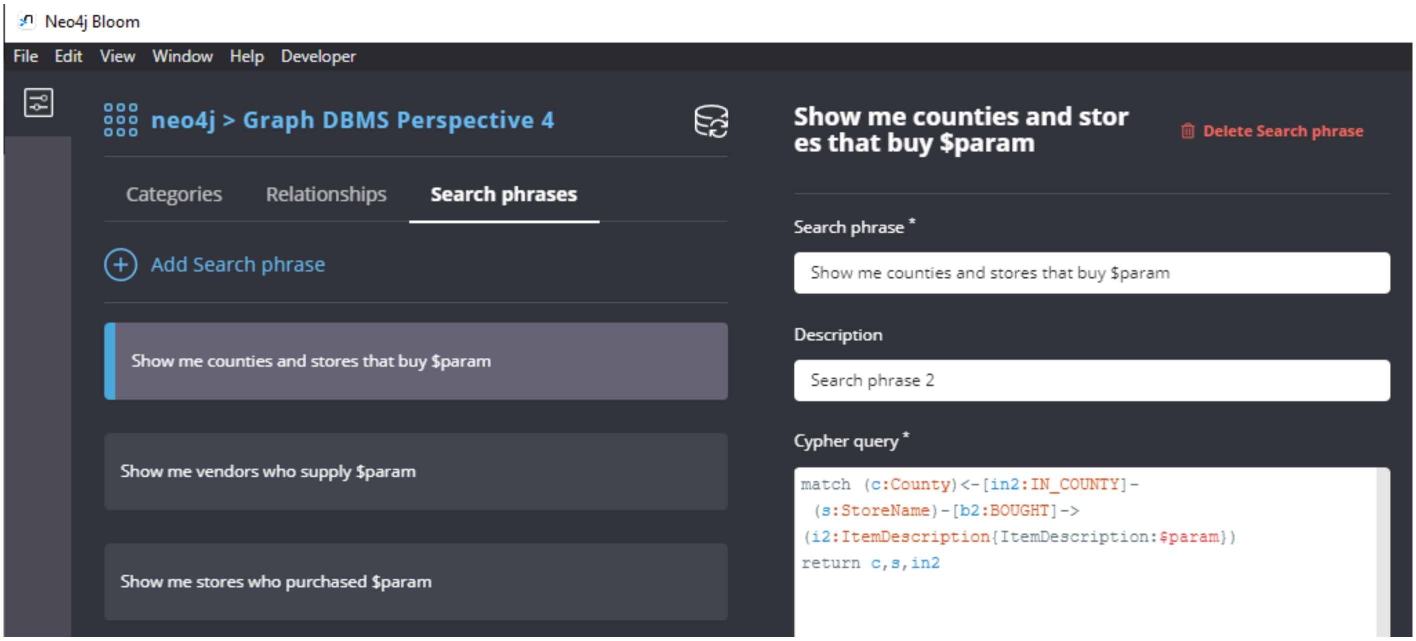
NETWORK VISUALIZATIONS

Search Phrases

For the next phase of my project, I've created two search phrases to make it easier for non-technical coworkers to perform self-serve analytics and quickly answer their own questions.

Show me counties and stores that buy $param

Since I know that we are experimenting with different types of whiskies at our distilleries, we will often need to know who is currently purchasing whiskies that are like our most recent creations. To see this, I've created the "Show me counties and stores that buy $param" search phrase. This is a dynamic search where you replace the $param with the whiskey of choice to see what stores have purchased that whiskey in the past and what county they are located in.


Neo4j Bloom
File Edit View Window Help Developer
neo4j > Graph DBMS Perspective 4
Categories
Relationships
Search phrases
Add Search phrase
Show me counties and stores that buy $param
Show me vendors who supply $param
Show me stores who purchased $param
Show me counties and stor
es that buy $param
Delete Search phrase

Search phrase *

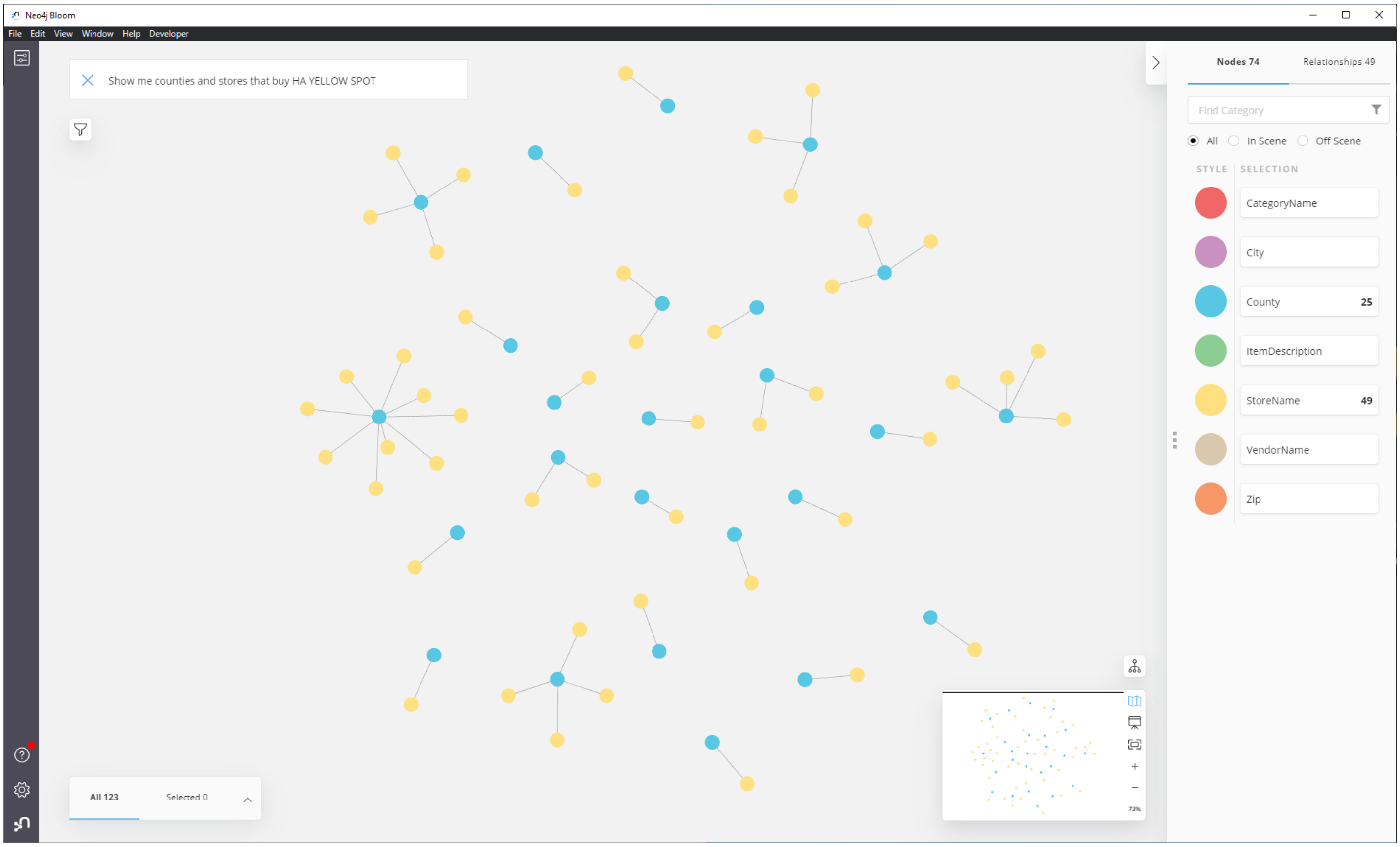
Show me counties and stores that buy $param

Description

Search phrase 2

Cypher query *

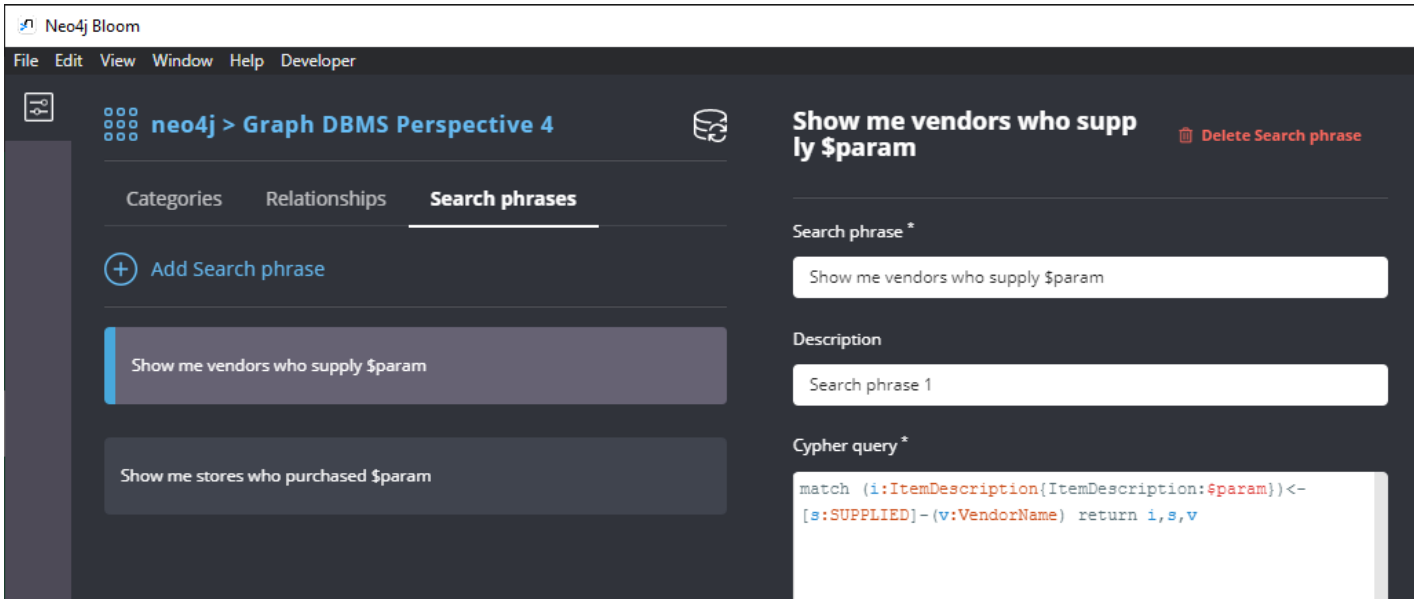
```
match (c:County)<-[in2:IN_COUNTY]-
 (s:StoreName)-[b2:BOUGHT]->
(i2:ItemDescription{ItemDescription:$param})
return c,s,in2
```



Show me counties and stores that buy HA YELLOW SPOT

Show me vendors who supply $param

The second part of that question will naturally be "how do we get our new product to those stores?" to answer that I've created a second search phrase "Show me vendors who supply $param". This allows you to replace the $param with a whiskey choice and see who has historically supplied that whiskey.


Neo4j Bloom
File Edit View Window Help Developer
neo4j > Graph DBMS Perspective 4
Categories
Relationships
Search phrases
Add Search phrase
Show me vendors who supply $param
Show me stores who purchased $param
Delete Search phrase

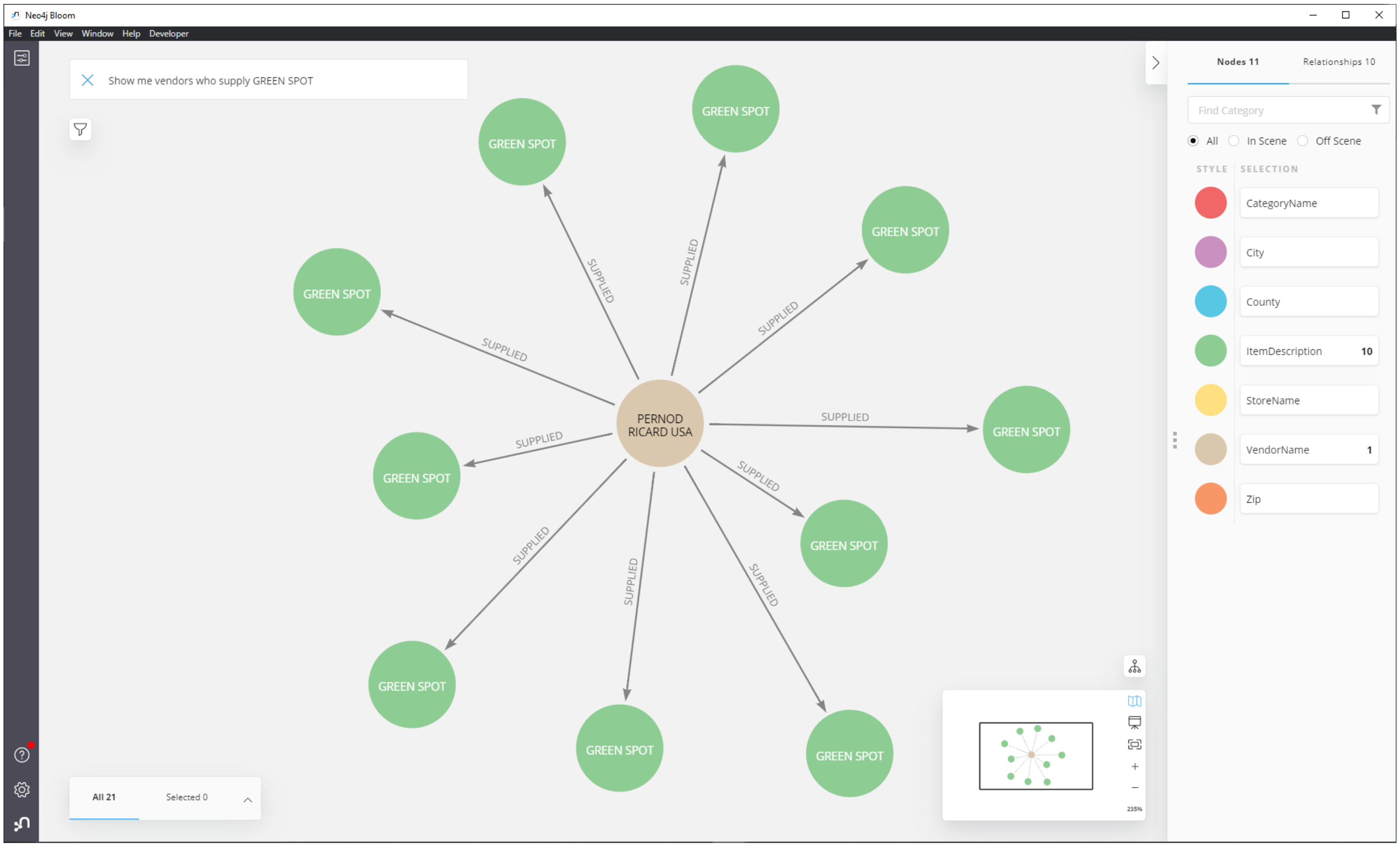
Show me vendors who supp ly $param

Search phrase *: Show me vendors who supply $param

Description: Search phrase 1

Cypher query *:

```
match (i:ItemDescription{ItemDescription:$param})<-
[s:SUPPLIED]-(v:VendorName) return i,s,v
```



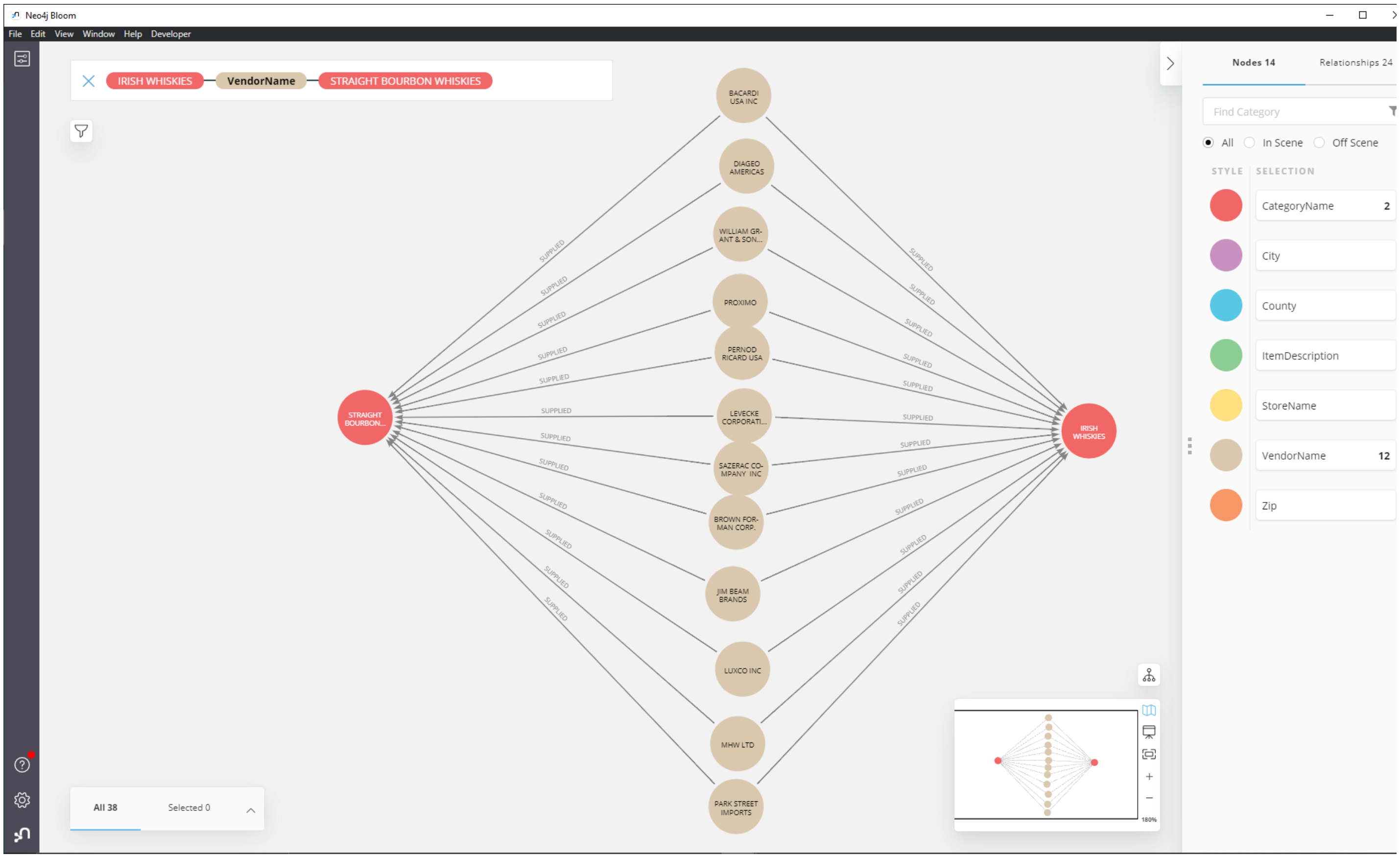
These two search phrases should make it very easy to see:

1. Who we should reach out to about distributing our new products.
2. What stores we would like to market to.

Graph Visualizations

Vendors to consider working with

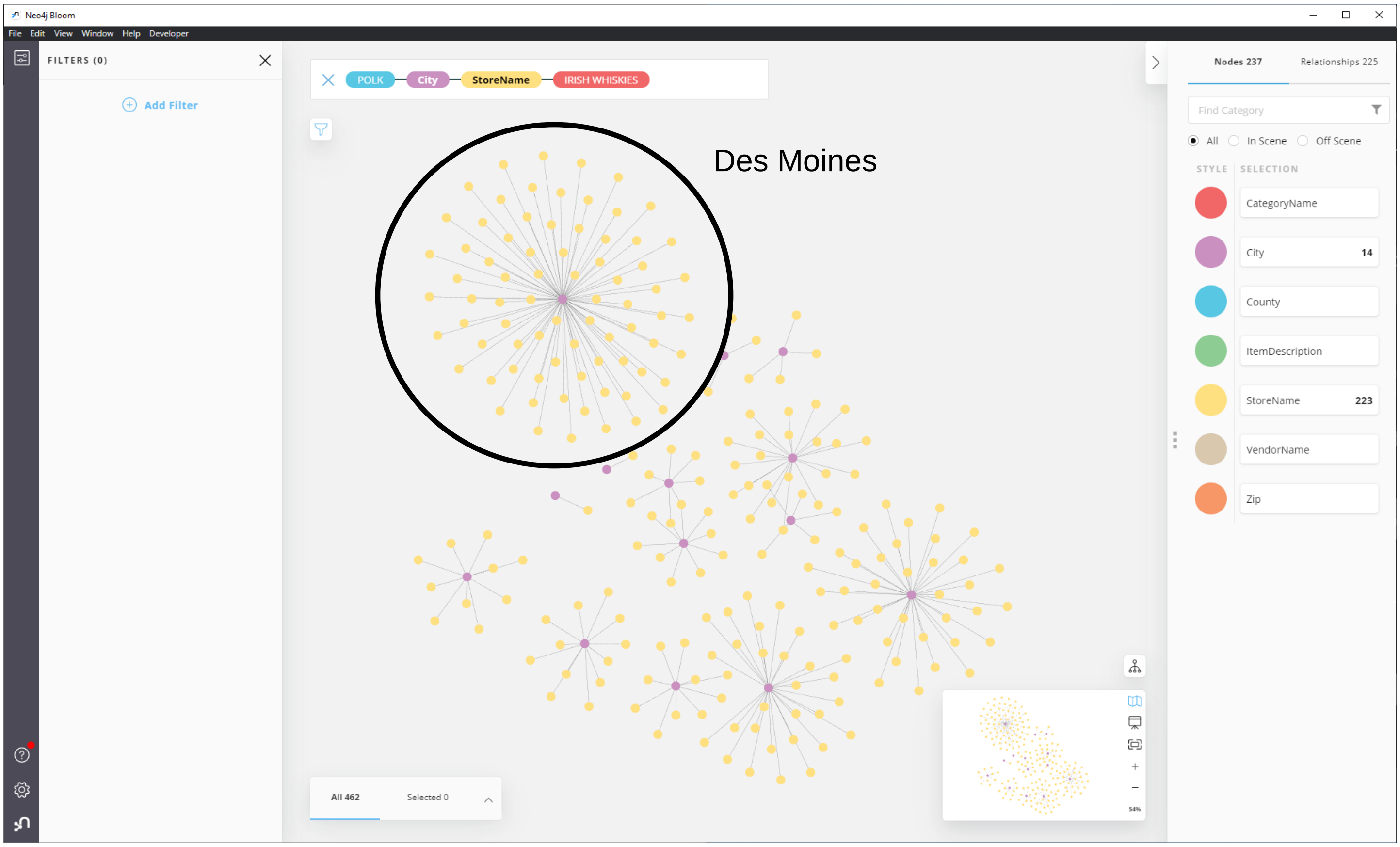
In my business use case, I note that a major problem we are trying to solve is which distributor we should use for our initial product launch. Since it is an American made whiskey in the Irish tradition, my thought is that we should look for a distributor who has experience in both American made bourbons and Irish Whiskies. Bourbons so they are familiar with the internal laws of transporting whiskey and Irish so they know what stores and counties we should focus on with our launch. To accomplish this, I created a visualization of the vendors who supply both to begin whittling down the list.



This shows us there are 12 vendors who currently supply both types to Iowa. This is a much more manageable list to explore than the 103 total we have in our dataset.

Cities to focus on first

As we discovered earlier when we looked at the counties who purchase the most whiskey, Polk is by far the largest by both total sales and diversity of brands purchased. To take this a step further, I've created a visualization to see which cities inside Polk have the most stores and by what margin. As we can see the largest cluster is in Des Moines. This means we could plausibly ship to only that city and distribute to the maximum number of stores from there.



RECOMMENDATIONS

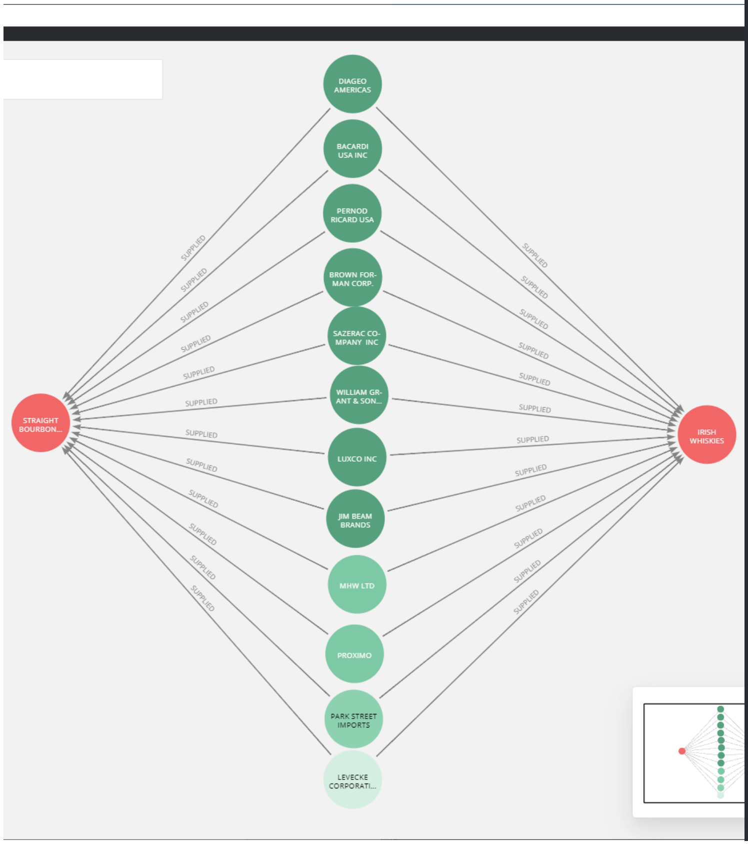
To make actionable recommendations on the business questions we are trying to answer. I'll take the analysis from the previous sections and combine them.

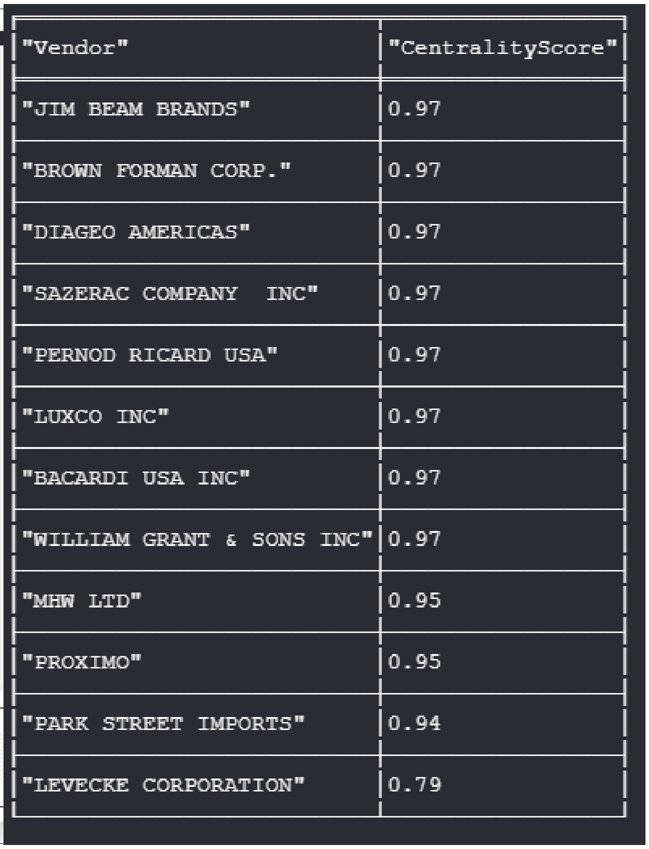
What distributor should we work with to get our product into stores?

For this first question I'll combine:

- The centrality algorithm I calculated earlier for how connected distributor are to stores in Iowa
- The visualization of distributors of both Irish and American style whiskies
- The search phrase "Show me vendors who supply $param"

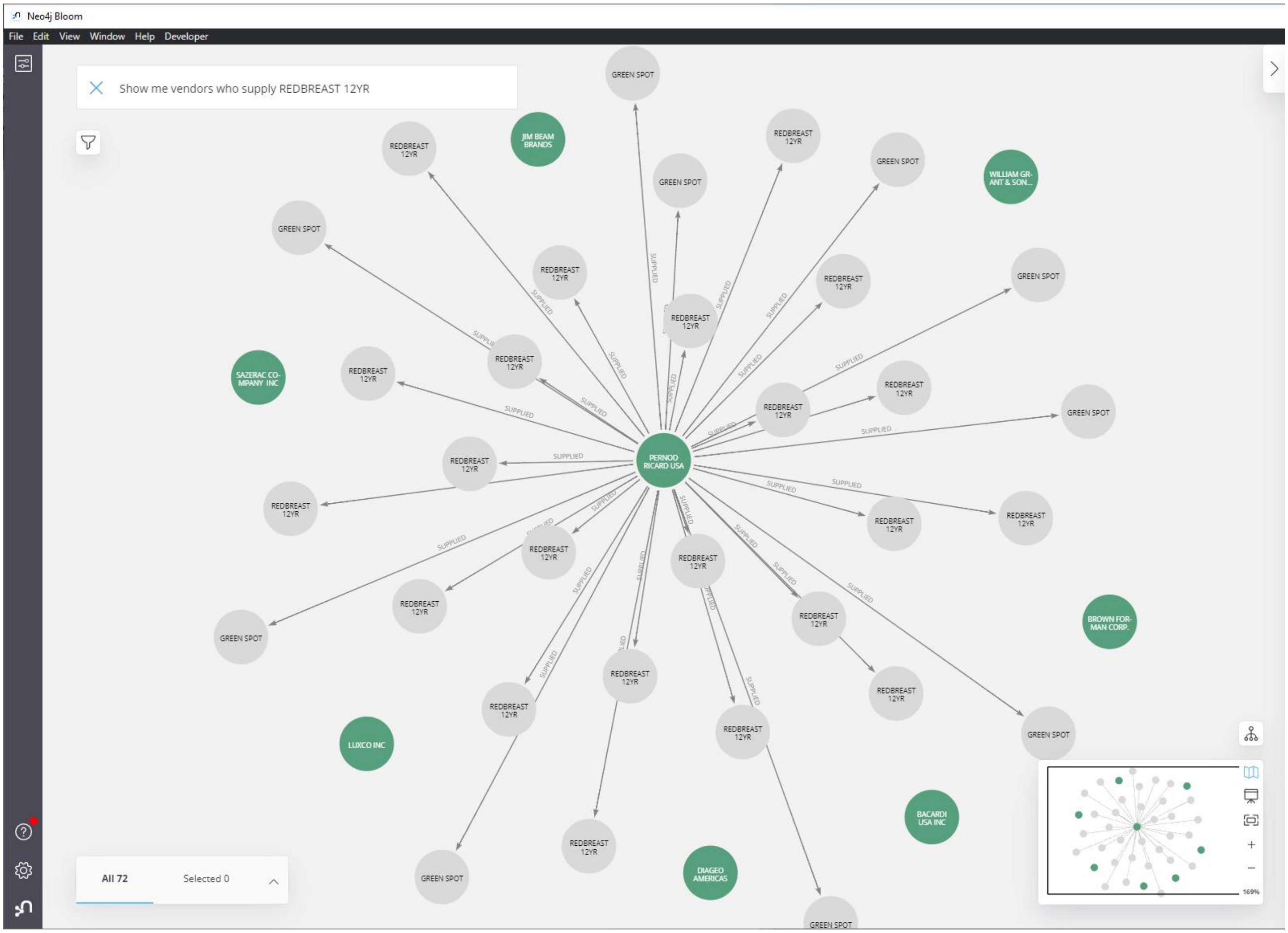
Combining the visualization and centrality score gives us a nice view of which vendors are most connected to stores in Iowa and carry both types of whiskies.





```
╒═══════════════════════════════╤════════════════╕
│"Vendor"                       │"CentralityScore"│
╞═══════════════════════════════╪════════════════╡
│"JIM BEAM BRANDS"              │0.97            │
├───────────────────────────────┼────────────────┤
│"BROWN FORMAN CORP."           │0.97            │
├───────────────────────────────┼────────────────┤
│"DIAGEO AMERICAS"              │0.97            │
├───────────────────────────────┼────────────────┤
│"SAZERAC COMPANY  INC"         │0.97            │
├───────────────────────────────┼────────────────┤
│"PERNOD RICARD USA"            │0.97            │
├───────────────────────────────┼────────────────┤
│"LUXCO INC"                    │0.97            │
├───────────────────────────────┼────────────────┤
│"BACARDI USA INC"              │0.97            │
├───────────────────────────────┼────────────────┤
│"WILLIAM GRANT & SONS INC"     │0.97            │
├───────────────────────────────┼────────────────┤
│"MHW LTD"                      │0.95            │
├───────────────────────────────┼────────────────┤
│"PROXIMO"                      │0.95            │
├───────────────────────────────┼────────────────┤
│"PARK STREET IMPORTS"          │0.94            │
├───────────────────────────────┼────────────────┤
│"LEVECKE CORPORATION"          │0.79            │
└───────────────────────────────┴────────────────┘
```

Most of these vendors are very well connected to stores in Iowa! There are a couple things we can do to make the list a bit smaller though. We can remove the ones that aren’t in the .97 group at the top to make sure we are working with the most connected distributors. Secondly, we can use the search phrase created earlier to see which of these vendors have experience working with high end Irish whiskies like “Green Spot” and “Red Breast 12”.



Show me vendors who supply REDBREAST 12YR

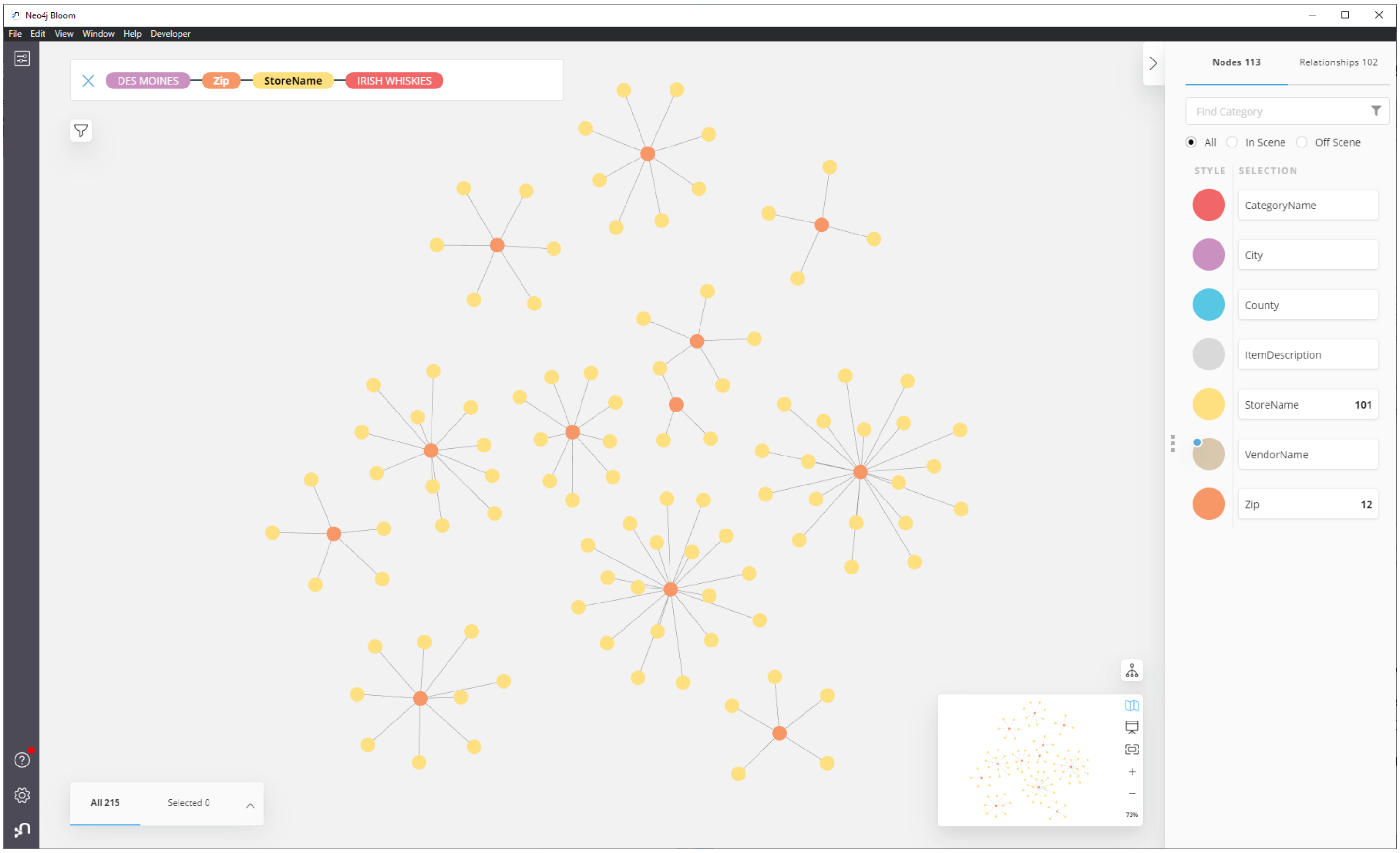
This leaves us with one distributor who supplies both Irish and American whiskies to the most stores in Iowa and have experience with high end whiskies!

Pernod Ricard USA

My recommendation is to reach out to them as our first choice for distribution.

What geographic area and stores should we target first?

As we discovered earlier, Polk county is by far the largest purchaser of Irish whiskies so is a logical place to start. However, with 247 stores in the county, it is a bit too big for an initial launch. Later, we saw that Des Moines is the largest city in the county as a market for Irish whiskies. That city alone has 76 stores that purchase Irish whiskies. To go one step farther, we will begin to break down the city to find a more manageable number for our initial launch. First up is to look for zip codes that purchase the most Irish whiskies.



These look like great test markets because of their existing Irish whisky sales along with being markets with huge potential as the distillery grows. Also, Pernod Ricard USA (our recommended distributor) currently distributes to all these stores.

Zip Code	Stores
50317	16
50313	11
50315	9
50321	8
50310	6

My recommendation is to pick one of the above zip codes as a test market and expand from there when appropriate.

Final Thoughts

There are many was to go about deciding on a distributor or geographic area for an initial product launch and these are by no means concrete recommendations. Depending on the specific business and marketing strategies this business would like to pursue, we could target in completely different ways. For example, if our whisky distillery would like to focus on exclusivity and rarity, we could change the targeted stores from time to time. Similarly, we could look for sites that do not currently have Irish whiskies but are similar to sites that do in demographics or sales. With that in mind, I have tried to build a database that could easily pivot as those decisions are made by the business and quickly answer questions about the best way to implement those strategies.