|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dublin&#39;s €10m Liberties distillery opens | Scotch Whisky | | | | |
|  |  |  | |  |
| Iowa Whiskey Sales  PRoject Deliverable III  GitHub: https://github.com/StevenEJordan/Network\_Science | | |
| Steven Jordan | | sjorda41@uncc.edu |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Empty Bourbon Barrel | Kings County Distillery | | | | |
|  |  |  | |  | |
|  | 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. | |  | |

|  |  |
| --- | --- |
|  |  |
| iowa sales DataData 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 it’s upload to Neo4J. | |  |

|  |  |
| --- | --- |
|  |  |
| Graph Data ModelsInitial 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 relationship between locations, stores, and purchases.   Cypher Query //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 Now that we have our data in our database, we can begin exploratory queries and algorithms. | |

|  |  |
| --- | --- |
|  |  |
| Graph Queries 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    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.  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    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    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    **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    **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 | |
| |  |  | | --- | --- | |  |  | | 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.      **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.      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 so they are familiar with the internal laws of transporting whiskey and Irish whiskies so they know about 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 center by both total sales and diversity of brands purchased. In order 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.    Des Moines | | | |
|  | |