Allen Jackson

Supply Chain Data Project

January 2022

Final Visualization: <https://public.tableau.com/app/profile/allen.jackson7251/viz/Supply_Chain_Data_Project/Story1?publish=yes>

Intro

I found a dataset with supply chain data meant for an optimization project. My goal is to query and explore the data and present what I learn. The data can be found at the following link:

<https://brunel.figshare.com/articles/dataset/Supply_Chain_Logistics_Problem_Dataset/7558679>

My first goal is to understand the data as presented before asking any questions. The dataset is in the form of an excel document with 7 sheets. I will list the sheets here and fill out the descriptions over time:

* OrderList
* FreightRates
* WhCosts
* WhCapacities
* ProductsPerPlant
* VmiCustomers
* PlantPorts

I’m executing this process using google bigquery. I have added each sheet as a table in a database so I can query it using SQL syntax.

Exploring Relationships

I’ve run a couple queries to get information on the commonalities between the tables:

SELECT column\_name FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.INFORMATION\_SCHEMA.COLUMNS` GROUP BY column\_name HAVING COUNT(\*) > 1

SELECT \* FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.INFORMATION\_SCHEMA.COLUMNS` WHERE column\_name

IN(SELECT column\_name FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.INFORMATION\_SCHEMA.COLUMNS` GROUP BY column\_name HAVING COUNT(\*) > 1)

ORDER BY column\_name

Exploring by Table

Each table seems to only have a relationship with one other table. I’m now going to look at the largest tables and refer to the data dictionary to learn more.

SELECT \* FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.INFORMATION\_SCHEMA.COLUMNS`

ORDER BY table\_name

FreightRates, OrderList, and PlantPorts have the most columns which indicates they likely have the most information. OrderList and FreightRates are the first two sheets on the excel document so I’ll start with those.

Having read back over the data description I understand the data is meant for an optimization problem focused on sending several products on the optimal routes given several ports varying capacities. I’m beginning to query the first table (freights rates) and will record my insights on each table.

Freight Rates

SELECT \* FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.FreightRates`  LIMIT 100

Initial look at the table

SELECT DISTINCT dest\_port\_cd FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.FreightRates`  LIMIT 100

SELECT DISTINCT orig\_port\_cd FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.FreightRates`  LIMIT 100

There are 10 origin ports and one target destination port.

Just from looking at the freight rates data I’m starting to think of some questions I can answer:

**Which origin ports tend to be the most costly?**

**Which mode of transportations are used most often by each origin port?**

**What is the average minimum cost by origin port?**

**How many day on average does it take shipments from each port?**

I should look into value counts for Carrier, svc\_cd (don’t yet know what that means), and carrier type

OrderList

SELECT  \* FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.OrderList`  LIMIT 100

Orders only came from PORT04, PORT09 and PORT05

There are 3 service levels: CRF, DTP, and DTD

There are 46 different customers

All of the orders in the database are on a single day: 2013-05-26

Continuing my exploration of the OrderList table. I wanted to generate some questions to answer on certain aspects of it:

**Which Service level is the most common?**

**What are the cumulative weights of each order?**

**Average cumulative weights by Port, Carrier, service level, or customer**

PlantPorts

Although a large table, plant ports just describes the actual allowed links between the warehouses. There are several empty columns. A visualization might better represent these relationships.

ProductsPerPlant

Products per plant list every product and which plants they’re available in.

**Which plants have the most products?**

**Which products are in the most plants?**

VMI\_customers

Lists warehouses and customers only those warehouses are allowed to deliver to. If the warehouse isn’t on the list then it has no description.

**Which warehouses have restrictions?**

**Which warehouses have the most restrictions?**

WhCosts

Cost of storing unit in warehouse by unit.

**Which Plants are the cheapest?**

**Which plants have the largest expense based on current inventory?**

WhCapacities

Lists the total amount of orders coming from the warehouses.

**Which customers make up the largest share of order by each warehouse?**

Answering Questions

I’ve explored each table and have questions to answer for each. I’m now going to attempt to answer these questions using SQL queries and visualizations.

**Which origin ports tend to be the most costly?**

**What is the average minimum cost by origin port?**

SELECT orig\_port\_cd, AVG(minimum\_cost) avg\_min\_cost, AVG(rate) avg\_rate FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.FreightRates`

GROUP BY orig\_port\_cd

There was an issue because minimum cost is actually a string with dollar sign values:

SELECT orig\_port\_cd, AVG(CAST(REPLACE(minimum\_cost, '$', '') AS DECIMAL)) avg\_min\_cost, AVG(CAST(REPLACE(rate, '$', '') AS DECIMAL))

 FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.FreightRates`

 GROUP BY orig\_port\_cd

**Which mode of transportations are used most often by each origin port?**

SELECT orig\_port\_cd, mode\_dsc, COUNT(\*)

 FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.FreightRates`

 GROUP BY orig\_port\_cd, mode\_dsc

**How many day on average does it take shipments from each port?**

WITH Averages AS

(SELECT orig\_port\_cd, (SELECT AVG(tpt\_day\_cnt) FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.FreightRates`) avg\_general,

AVG(tpt\_day\_cnt) avg\_port

FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.FreightRates`

GROUP BY orig\_port\_cd)

SELECT orig\_port\_cd, avg\_general, avg\_port, (avg\_general - avg\_port) avg\_diff FROM Averages

**Which Service level is the most common?**

SELECT Service\_Level, COUNT(\*) FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.OrderList`

GROUP BY Service\_Level

**What are the cumulative weights of each order?**

SELECT Order\_ID, Unit\_quantity \* Weight FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.OrderList`

**Average cumulative weights by Port, Carrier, service level, or customer**

SELECT Origin\_Port, AVG(Unit\_quantity \* Weight) avg\_cumulative\_weight

FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.OrderList`

GROUP BY Origin\_Port

SELECT Carrier, AVG(Unit\_quantity \* Weight) avg\_cumulative\_weight

FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.OrderList`

GROUP BY Carrier

SELECT Customer, AVG(Unit\_quantity \* Weight) avg\_cumulative\_weight

FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.OrderList`

GROUP BY Customer

Continuing to answer these questions. The queries are saved here and the results are saved in a google sheets page.

<https://docs.google.com/spreadsheets/d/13APdSa3LkTc1uAnh5tCG1jPyUMxopbnM6KyL_Vu47Jc/edit#gid=0>

Also should be noted I didn’t make a question for the plant ports table but I should count that visualization as a question itself.

**Which plants have the most products?**

SELECT Plant\_Code, COUNT(\*) Count FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.ProductsPerPlant`

GROUP BY Plant\_Code

ORDER BY Count DESC

**Which products are in the most plants?**

SELECT Product\_ID, COUNT(\*) Count FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.ProductsPerPlant`

GROUP BY Product\_ID

ORDER BY Count DESC

Additional Question (Jan 1st):

**I got exactly 100 rows as a result. Are there exactly 100 distinct products?**

SELECT Count(DISTINCT Product\_ID) Count FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.ProductsPerPlant`

The following query returned a count of 1540. Upon further investigation I realized big query only presents the first 100 rows and that all 1540 were properly grouped.

This won’t keep me from getting the answer but it is worth noting that the first row in the VMI customer table is set as the column names. I should adjust this later.

**Which warehouses have restrictions?**

**Which warehouses have the most restrictions?**

SELECT string\_field\_0, COUNT(\*) count FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.VmiCustomers`

WHERE string\_field\_0 != 'Plant Code'

GROUP BY string\_field\_0

ORDER BY count DESC

**Which Plants are the cheapest?**

SELECT \* FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.WhCosts`

ORDER BY Cost\_unit DESC

**Which plants have the largest expense based on current inventory?**

WITH total\_units AS (SELECT Plant\_Code, SUM(Unit\_quantity) Unit\_quantity FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.OrderList`

GROUP BY Plant\_Code)

SELECT t.Plant\_Code, t.Unit\_quantity, c.Cost\_unit, (t.Unit\_quantity \* c.Cost\_unit) total\_cost

FROM total\_units t

LEFT JOIN (SELECT \* FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.WhCosts`) c

ON c.WH = t.Plant\_Code

ORDER BY total\_cost DESC

**Which customers make up the largest share of orders by each warehouse?**

SELECT Plant\_ID, Daily\_Capacity\_,

ROUND(Daily\_Capacity\_ / (SELECT SUM(Daily\_Capacity\_) FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.WhCapacities`)\*100,2) Percent\_of\_total

FROM `platinum-voice-334221.supply\_chain\_logistics\_problem.WhCapacities`

ORDER BY Percent\_of\_total DESC

Visualizing Insights

Having answered all of the questions I’m going to visualize my insights. After trying a few methods of organizing the data I realized it makes the most sense to organize the graphs based on the problem. If I were to begin solving this problem the visualizations should give me an overlay of the Warehouses and Plants so I can have a general understanding of their utility. Given I made 3 dashboards:

1. Limitations (PLantPorts, ProductsPerPlant, VmiCustomer, WhCapcities)
2. Costs (Freight Rates, WhCosts)
3. Visualizing the problem (Order List)

Explanation

Having completed all three dashboards and combined them into a storyboard. The last part of this project will be a total write up explaining the problem, the variables and each chart.

Going to walk through the storyboard slide by slide:

Slide 1 – Explanation of the problem

The following visualizations are derived from a dataset for an optimization problem. The dataset is made up of 7 tables. The first table Order list is the list of orders that need to be delegated. The following six tables describe a plethora of limitations based on conditions in both warehouses (plants) from which products are stored as well as the ports and shipping lanes (ports) from which they are moved.

This first slide depicts the distribution of the orders in the order list table. There are a total of 9,215 orders of 772 products made by 46 customers. Most of these orders originate in PORT09 with virtually the rest originating from PORT04. The second pie chart indicates the type of service requested. Door to Port (DTP) was the most requested followed by Door to Door (DTD) and lastly Customer Referred Freight (CRF). The final pie chart is the distribution of different carrier types. Most of order being requested from carrier V444\_0, followed by V444\_1 and lastly V44\_3. It is also notable that V44\_3 is the carrier that handles Customer referred Freights which explains them both having exactly 854 instances.

Slide 2 – Plant Information

This slide depicts the options we have based on 11 ports from which the items are shipped. This information is generally relevant when optimizing for things such as cost and time. The total amount of shipping lanes that exist from each port. Port 06 clearly has the most options with 479. There are clearly more air options compared to ground options with only PORT03 and PORT09 having Land options available. The average lane cost table averages the costs of shipping objects along each lane. PORT03 and PORT10 have significantly large average costs compared to the rest. Avg Days by Port indicate which port lanes tend to take longer for delivery.

Slide 3 – Warehouse information

The warehouse information describes limitations as well as average storage costs. In terms of optimizing the orders this most describes model limitations. Not all ports ship to each Warehouse as can be seen in the compatibility chart. PORT04 is compatible with the most PLANTS while each PLANT is compatible with two warehouses at most. The bar graph at the bottom depicts unique products held at each warehouse. PLANT03 clearly holds the most diverse array of products. Finally, the tables in the top right corner indicate Capacity and Average cost. Plants 01 and 03 hold can handle over 1,000 different orders.

Following me applying the text information into a storyboard and finalizing the design, the project is complete. I’m now going to organize the relevant documents and upload them to github.

Conclusion and Next Steps

With the storyboard complete this concludes my total exploration of the dataset and problem. There are documents explaining Ant Colony Optimization which is meant to be applied to this situation. I think it’d be interesting to explore this methodology and later visualize the results against the visualization of the problem.