Datathon: Big Supply Co.

Presented by Correlation One

Problem Statement

Welcome to the 2022 Datathon! This document explains the topic of the Datathon, important details about the datasets you'll be using, and guidance on how to submit your results.

Background (adapted from Mendeley Data[1])

A supply chain is a network established by a company and its suppliers with the intent purpose of producing and distributing products. While the supply chain is typically afterthoughts to the common consumer, lingering impacts of the COVID-19 pandemic have thrust supply chain management into the forefront of consumers' minds. Many executives are using these mounting supply chain pressures as a catalyst to open the hood, and evaluate the overall health of their companies supply chains.

The process of supply chain management is characterized by controlling the flow of goods and services, and oversight of all processes that transform raw materials into final products. When supply chain management is optimized, a company is able to maximize customer value and cultivate a competitive advantage in their market place.

Supply chain management has continued to evolve as businesses find more strategic and efficient modes of operation. For example, an increasing degree of integration of separate tasks underpinned significant advancement in the 1960s. The most prevalent current trend for supply chain management is **digitization** – the ability to harness the power to technology and data to optimize the supply chain.

For this datathon, you will be analyzing BigSupply Co's supply chain. BigSupply Co. is a multinational retailer that specializes in clothing, fitness, and electronics. They offer over 100 distinct products which belong to 32 different categories. BigSupply Co. also sells their products to a large customer base that includes individual consumers, home office consumers, and corporate clients.

[1] Constante, Fabian; Silva, Fernando; Pereira, Antonio (2019), "DataCo Smart Supply Chain for Big Data Analysis", Mendeley Data, V5, doi: 10.17632/8gx2fvg2k6.5

Your Task

Your goal is to analyze the provided datasets, potentially in combination with supplementary datasets, in order to better understand patterns in BigSupply Co.'s supply chain that can be practically applied to improve company performance.

You are asked to pose your own question and answer it using the available datasets as well as any supplementary datasets you may find. What is important is both the creativity of your question and the quality of your data analysis. **You need not be comprehensive; depth of insight is more important over breadth of question posed.**

Submissions may be predictive, using machine learning to classify or predict patterns. Submissions may also be illuminating through use of data visualizations or through sound statistical tests.

Consider exploring one of the sample questions below, or creating your own variation. Creativity in formulating your own question is encouraged; **however, it should not be at the expense of analytical depth, precision, and rigor, which are far more important**.

Sample Question 1: What is BigSupply Co.'s overall shipping effectiveness? Does the effectiveness change based on customer type, customer location, or product ordered?

Sample Question 2: BigSupply Co. offers a wide array of products spanning many categories. Are there any products that have an outsized impact on BigSupply Co.'s bottom line? Are shipping difficulties localized to certain products?

Sample Question 3: Where is BigSupply Co. most susceptible to fraudulent consumer activity?

Datasets

The provided datasets are stored in the "Datathon Materials" folder on Google Drive. Your team should only use the datasets that are relevant to your choses question / topic.

The dataset consists of 180,519 orders placed between January 1st, 2015 and January 31st, 2018. For each order, the dataset included details relating to financial impact and execution of the order. The other data tables provide additional detail on each order by providing details on the product, origination store, and customer.

Orders

This table contains records for all placed orders between January 1st, 2015 and January 31st, 2018. Each row represents an order.
180,519 rows & 24 columns. Size: 33.1MB

Categories

This is a reference table that includes all of BigSupply Co.'s product categories.
51 rows & 2 columns. Size: 883 bytes

Customers
This table contains records for all of BigSupply Co.'s customers over the same timeframe.
20,653 rows & 11 columns. Size: 2MB

Departments
This is a reference table that includes all of BigSupply Co.'s stores.
11 rows & 4 columns. Size: 1.5MB

Products
This is a reference table with all of BigSupply Co.'s current products with details on each product.
118 rows & 7 columns. Size: 15KB

Additional Datasets

Participants are welcome to scour the Web for their own custom datasets to supplement their analysis. All additional data used should be public and should not exceed 2GB unzipped (consult Correlation One's R&D team if you believe your idea is a worthy exception).

Other Materials

We will provide you with the schema for each of the data tables in another packet.

Submissions: Content

Submissions should have two components:

1. Report – this should have two main components:
    a. Non-Technical Executive Summary – What is the question that your team set out to answer? What were your key findings, and what are their significance? You must communicate your insights clearly – summary statistics and visualizations are encouraged to help explain your thought process.
    b. Technical Exposition – What was your methodology/approach towards answering the questions? Describe your data manipulation and exploration process, as well as your analytical and modeling steps. Again, the use of visualizations is highly encouraged when appropriate.
2. Code – please include all relevant code that was used to generate your results. **Although your code will not be graded, you MUST include it or your entire submission will be discarded.**

Additional information (e.g. roadblocks encountered, caveats, future research areas, and unsuccessful analysis pathways) may be placed in an appendix.

Judges will be evaluating your technical report without your team there to explain it; therefore, **your submission must "speak for itself"**. Please ensure that your main findings are clear and that any visualizations are functionally labeled.

Submissions: Evaluation

The competition will have multiple rounds of evaluation. Your Report will be judged as follows:

- **Non-Technical Executive Summary**
  - *Insightfulness of Conclusions.* What is the question that your team set out to answer, and how did you choose that question? Are your conclusions precise and nuanced, as opposed to over-generalizations?
- **Technical Exposition**
  - *Wrangling & Cleaning Process.* Did you conduct proper quality control and handle common error types? How did you transform the datasets to better use them together? What sorts of feature engineering did you perform? Please describe your process in detail within your Report.
  - *Investigative Depth.* How did you conduct your exploratory data analysis (EDA) process? What other hypothesis tests and ad-hoc studies did you perform, and how did you interpret the results of those tests and analyses? What patterns did you notice, and how did you use these to make subsequent decisions?
  - *Analytics & Modeling Rigor.* What assumptions and choices did you make, and how did you justify them? How did you perform feature selection? If you built models, how did you analyze their performance, and what shortcomings do they exhibit? If you constructed visualizations and/or conducted statistical tests, what was the motivation behind the particular models you built, and what did they tell you?

Submissions: Format

Reports can be produced using any tool you prefer (Python Notebook, Shiny Application, Microsoft Office, etc.); however, **your report MUST be in a universally accessible and readable format (HTML, PDF, PPT, Web link)**. It must not require dedicated software to open. For example, if your report is a Python Notebook, it should be exported to HTML. If you create a Shiny App, it should be published at an accessible Web link.

Please include the source file used to generate your report. For example, if you submit a PDF with math-type, equations, or symbols, please include your LaTeX source file.

Code should be submitted in a single zipped collection of files separate from your report.

Your team will be sent a Google Form at the beginning of the competition; you will use this form to upload and send in your submitted content. **Submissions MUST be received by 10:00 PM**

GMT on SUNDAY, March 13th. Any submissions received after that time will NOT be evaluated by the judges.

Tips and Recommendations

For data engineering, exploration, and modeling, we highly recommend that you install Jupyter Notebook: http://jupyter.org/install.html. Jupyter Notebook is an interactive, real-time development environment that eliminates many pain points of the standard "terminal + text editor" environment, and is compatible with both Python and R.

We also recommend that your team not try to learn new tools if it can be avoided; instead, leverage your existing skills to extract as much insight from the data as you can.

We've compiled 3 additional commonalities of successful teams and 3 pitfalls of unsuccessful teams. Of course, these may not apply to every team, so we recommend that you and your team apply any tips accordingly.

Tips for Success	Try to Avoid
1. Focus on hypothesis testing when brainstorming your research question	**1.** Do not try to exhaust all different models you know just to yield an ideal cross validation accuracy
2. Spend at least 3 hours on your report to ensure strong communications through both visualizations and writing	**2.** Do not violate assumptions of statistical models. Sometimes, specific models require specific features so it is best to make sure those conditions are sufficiently met
3. Engage in proper causal analysis. Just because your model passes standard cross-validation checks it does not demonstrate (or even suggest) causality	**3.** Do not pick research statements and blindly stick to it trying to get it to work. Often times, further data exploration will show that it's not true or worthwhile

Ask for Help

Correlation One's R&D team is here to help. Let us know about your struggles as early on as you can and we may be able to offer advice on how to best move your analysis forward.