# Milestone 1 – Project Proposal

The purpose of this document is to define the first project of the DSC 680 Applied Data Science course. The data source will be identified and preliminarily reviewed for the project. My current role is a Material Quality Engineer (MQE) within the Global Procurement Group. My main objective as an MQE is to be a subject matter expert on incoming materials. This entails solving critical problems with suppliers and addressing quality concerns within the supply chain. Another crucial aspect within the supply chain entails on-time delivery of shipments. Essentially, suppliers need to be committed to delivering material in the agreed upon timeframe with their customers. With this background in mind, I have chosen to focus on sales forecasting and late delivery prediction for the first project. The Milestone 1 template will be followed for the remainder of this document and can be referenced in the Appendix.

## Topic

Project Name: Sales Forecasting and Late Delivery Prediction for Supply Chain

This project focuses on the prediction of future sales and late deliveries to improve supply chain performance.

## Business Problem

In today’s world, a supplier that understands their customers’ needs and can provide the material (or service) in a timely fashion becomes invaluable. Customers tend to retain suppliers with high performance and this project will focus on two main aspects. The first objective is to develop a model for sales forecasting based on historical data. A foundational understanding of customer demand can help suppliers better manage their own internal supply chain and processes. The second focus is to determine a method for predicting late deliveries. This allows the supplier to correct ongoing issues with late deliveries or communicate with customers early in the process. The relationships between customers and suppliers within a system often get compared to a stream. One supplier may report to a particular customer, then that customer serves as a supplier for an alternate customer downstream and so on... Organizations can relay these strategies to their suppliers “upstream” to continue to meet customer expectations. The research questions associated with this project are outlined below:

Objective 1 Prediction of Future Sales

* Which categories had the highest sales historically?
* Which customers bought the most historically?
* Are there any variables strongly correlated with sales?
* Which model provides the best accuracy for forecasting sales?

Objective 2 Prediction of Late Deliveries

* Which categories had the highest number of late deliveries historically?
* Does a particular product tend to be late historically?
* Do customers in a particular geographic area tend to receive late shipments?
* Are there any variables strongly correlated with late deliveries?
* Which model provides the best accuracy for predicting late deliveries?

## Datasets

The dataset for this analysis is from Mendeley Data and represents supply chain data from the company DataCo Global. Only the structured data from the DataCoSupplyChainDataset.csv will be utilized. The link to the data can be found below:

[DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS - Mendeley Data](https://data.mendeley.com/datasets/8gx2fvg2k6/5)

The shape of the dataset includes 53 variables and over 180,000 records. The source for the data includes a quick description of each variable as shown below:

|  |  |
| --- | --- |
| FIELDS | DESCRIPTION |
| Type | Type of transaction made |
| Days for shipping (real) | Actual shipping days of the purchased product |
| Days for shipment (scheduled) | Days of scheduled delivery of the purchased product |
| Benefit per order | Earnings per order placed |
| Sales per customer | Total sales per customer made per customer |
| Delivery Status | Delivery status of orders Advance shipping, Late delivery, Shipping canceled, Shipping on time |
| Late\_delivery\_risk | Categorical variable that indicates if sending is late (1), it is not late (0). |
| Category Id | Product category code |
| Category Name | Description of the product category |
| Customer City | City where the customer made the purchase |
| Customer Country | Country where the customer made the purchase |
| Customer Email | Customer's email |
| Customer Fname | Customer name |
| Customer Id | Customer ID |
| Customer Lname | Customer lastname |
| Customer Password | Masked customer key |
| Customer Segment | Types of Customers Consumer, Corporate, Home Office |
| Customer State | State to which the store where the purchase is registered belongs |
| Customer Street | Street to which the store where the purchase is registered belongs |
| Customer Zipcode | Customer Zipcode |
| Department Id | Department code of store |
| Department Name | Department name of store |
| Latitude | Latitude corresponding to location of store |
| Longitude | Longitude corresponding to location of store |
| Market | Market to where the order is delivered Africa, Europe, LATAM, Pacific Asia, USCA |
| Order City | Destination city of the order |
| Order Country | Destination country of the order |
| Order Customer Id | Customer order code |
| order date (DateOrders) | Date on which the order is made |
| Order Id | Order code |
| Order Item Cardprod Id | Product code generated through the RFID reader |
| Order Item Discount | Order item discount value |
| Order Item Discount Rate | Order item discount percentage |
| Order Item Id | Order item code |
| Order Item Product Price | Price of products without discount |
| Order Item Profit Ratio | Order Item Profit Ratio |
| Order Item Quantity | Number of products per order |
| Sales | Value in sales |
| Order Item Total | Total amount per order |
| Order Profit Per Order | Order Profit Per Order |
| Order Region | Region of the world where the order is delivered Southeast Asia ,South Asia ,Oceania ,Eastern Asia, West Asia , West of USA , US Center , West Africa, Central Africa ,North Africa ,Western Europe ,Northern , Caribbean , South America ,East Africa ,Southern Europe , East of USA ,Canada ,Southern Africa , Central Asia , Europe , Central America, Eastern Europe , South of USA |
| Order State | State of the region where the order is delivered |
| Order Status | Order Status COMPLETE, PENDING, CLOSED, PENDING\_PAYMENT, CANCELED, PROCESSING , SUSPECTED\_FRAUD , ON\_HOLD , PAYMENT\_REVIEW |
| Product Card Id | Product code |
| Product Category Id | Product category code |
| Product Description | Product Description |
| Product Image | Link of visit and purchase of the product |
| Product Name | Product Name |
| Product Price | Product Price |
| Product Status | Status of the product stock If it is 1 not available, 0 the product is available |
| Shipping date (DateOrders) | Exact date and time of shipment |
| Shipping Mode | The following shipping modes are presented Standard Class, First Class, Second Class, Same Day |

After a quick review, not all variables will be retained in the analysis. For example, Customer Email and Product Image will be removed initially. Other variables will be evaluated for importance related to forecasting sales or predicting late shipments.

## Methods

The Cross Industry Standard Process for Data Mining (CRISP-DM) will be followed for this project. The high-level phases for this process include Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The Business Understanding phase was considered above in the Business Problem section. To recap, this project aims to improve supply chain performance through sales forecasting and prediction of late delivery. There will hopefully be actionable insights that surface from this analysis. The Data Understanding step will involve Exploratory Data Analysis. This will include univariate and bivariate analysis of the variables. Next, the Data Preparation step will be performed to prepare the data for each respective model. The forecasting sales model will be regression whereas the late delivery model will be classification. Supervised learning algorithms will be utilized for each. For forecasting sales, I am considering Linear, Lasso, and Ridge Regression models. For predicting late deliveries, I am considering Logistic Regression, Decision Trees, and potentially Ridge or Lasso Regression. Since there are two types of problems being addressed, the evaluation metrics will be different for the models predicting future sales (R2 or RMSE) compared to prediction of a late delivery (accuracy, precision, recall, F1 score). Cross-validation will be considered to better understand model performance. Lastly, a recommendation will be made whether to deploy the models. This will be an iterative process and steps may be revisited often throughout the analysis.

## Ethical Considerations

Ethics is essential when working on any type of data science task/project. Since this project focuses on supply chain, it is important to remember privacy for customers and/or vendors. This project will consider ethics throughout each step in the analysis. Honesty and transparency for all major decisions will be communicated clearly to the audience. The story within the data will be unfolded without any deceitful intent.

## Challenges/Issues

This is not a one-size fit all solution for all organizations. Supply chains are structured differently depending on industries and products (or services) sold. Several challenges may surface during the analysis. First, there may be key variables not included in the dataset for predicting sales or late deliveries. It will be important to understand the most influential variables within the dataset. This challenge can also be mitigated by considering an alternate dataset. Second, the selected models may not have strong enough performance to be deployed. The evaluation metrics will serve useful for selecting the best model and determining if it is ready for deployment. Third, the frequency for retraining the model after deployment (if necessary) will need to be chosen carefully. Supply chain data may quickly evolve as customers’ needs change and markets fluctuate.

## References

AI Monks (n.d.). Top 10 Analytics Projects in Operations and Supply Chain. Resources. Retrieved June 10, 2023, from <https://aimonks.com/supply-chain-analytics-top-10-analytics-projects/>

Ariwala, P. (2022, September 30). 9 Ways Machine Learning Can Transform Supply Chain Management. Artificial Intelligence and Machine Learning. Retrieved June 10, 2023, from <https://marutitech.com/machine-learning-in-supply-chain/>

Constante, F., Silva, F., & Pereira, A. (2019, March 12). Datasets. DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS. Retrieved June 10, 2023, from <https://data.mendeley.com/datasets/8gx2fvg2k6/5>

IBM (2021, August 17). CRISP-DM Help Overview. Documentation. Retrieved June 10, 2023, from <https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview>

Tiwari, S. (2019, March 12). DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS. About Dataset. Retrieved June 10, 2023, from <https://www.kaggle.com/datasets/shashwatwork/dataco-smart-supply-chain-for-big-data-analysis?resource=download&select=DataCoSupplyChainDataset.csv>

Ying, K. (2021, September 16). Data Science Ethics – What Could Go Wrong and How to Avoid It. Data Science. Retrieved June 10, 2023, from <https://www.freecodecamp.org/news/the-ethics-of-data-science/>

# Appendix – Milestone Template Criteria

Each Milestone must follow the following template

**Milestone 1 - Proposal**

Submit a proposal with the following topics covered as a PDF

* Topic - Describe and name your project in 1-2 sentences max
* Business Problem - Describe the business problem your project is trying to solve and/or the research questions you will explore
* Datasets - where are you getting your data? Describe the data that you will use to solve the problem
* Methods - What analysis methods will you use to complete this project? Note this is just a proposal, your project can adapt as you work on it
* Ethical Considerations - What are some potential ethical concerns of this topic or analyzing the data?
* Challenges/Issues - What are some issues and challenges do you think you might face?
* References - What sources will you use to validate your results and support your project topic?

**Milestone 2 - Draft White Paper**

Submit a draft of your white paper with the following topics covered as a PDF

* Business Problem
* Background/History
* Data Explanation (Data Prep/Data Dictionary/etc)
* Methods
* Analysis
* Conclusion
* Assumptions
* Limitations
* Challenges
* Future Uses/Additional Applications
* Recommendations
* Implementation Plan
* Ethical Assessment