Problem/Overiew:

Individual retail stores based in population dense areas are always out-of-stock or running low on many of their more popular products. The stem of this issue occurs are three different levels of the retail experience: the micro-level and macro-level of the consumers (individual retail stores and corporate), and the supplier (production company). Often when customers are faced with the "out-of-stock" reply for their desired products, they become deterred to try again at the same store later, or, even worse, to completely give up on physical stores to make their purchases. This is a large cause for the movement of physical shopping to online shopping where they can easily get the status of all products at a moment's notice, and the purchases are made and delivered at the customers' convenience.

The loss of physical customers usually results in the loss of loyal customers as most online customers will purchase their desired item from the cheapest, reliable source available, which is usually the larger, more well-known corporations that can afford to cut their sale prices. This causes many local or smaller corporations to have increased churn rates, and eventually go out of business.

My objective for this project is to experiment with a sample of retail data, specifically from 1C Company, a Russian software firm. I will analyze the data and generate a predictive model that can reliably forecast the future sales of their products for up to a month.

Potential Clients:

The potential clients for this project include the individual retail stores, the larger corporations, and the production companies. Individual retail stores would utilize this data for two main purposes. The first is to determine which products are still desired and which ones should be replaced to bring in potentially larger profit margins. The second is to help managers keep track of the influx and efflux of each product to more efficiently maintain stocked inventory, thus preventing reduced profits if the customer chooses to buy the product elsewhere. The short-term loss of customers usually has long-term results as poor customer experience can affect customer-retention rates.

Corporate retail stores would use this data for similar reasons as individual retail stores, but at a larger magnitude. However, poor customer experience here has a larger affect as it affects customer loyalty and the reception of corporate reliability.

Production companies could apply this data to better prioritize their manufacturing process to focus their efforts on the production of items that have higher sale value. If expanded with geographical information, production companies can focus their shipments to areas where their products are selling well.

Data:

This dataset was provided by the 1C Company to a Kaggle Competition several months ago. The time-series dataset consists of daily sales data ranging from Jan. 2013 to Oct. 2015. It consists of three tables containing information about over twenty thousand individual products, multiple product categories, and sixty individual stores. The time series data contains approximately a dozen attributes, consisting of a mixture of both categorical and quantitative information.

Approaching the Problem:

The first focus will be to explore how each variable affects the items sold per day, then aggregate this data over each week and month to better approximate the associate. This initial association study will be limited to individual retail stores, then expanded to the macro-level by applying it to larger samples of stores. Another approach could focus on individual products then expanded to more products.

Next, I will focus on cleaning the data. This involves thoroughly checking the data for any wrongly inputted data as well as any missing values as null values. If there are any incorrect data types, such as floats casted as strings or, more commonly, datetime objects casted as strings, the cleaning will address that.

The last part after getting a better understanding of the variable associations and domain is to dive into the data and begin an in-depth visual and statistical analysis.

Deliverables:

With the project, I hope to have solid code with visuals and statistical analysis that allow me to provide a solid presentation of both my code and my results.