Cleaning And Exploring a Data Set To Identify Correlations

By: Sai Ananthula

Table of Contents

[Project Overview 3](#_Toc131964205)

[Data Exploration 3](#_Toc131964206)

[Data Cleaning 3](#_Toc131964207)

[Data Transformation 4](#_Toc131964208)

[Data Analysis 5](#_Toc131964209)

[Conclusion 8](#_Toc131964210)

# Project Overview

This project was initiated to process a data set regarding ABC company Lead and Shipping times during the Coronavirus Pandemic. The question that was asked was that was there a relationship between Logistics lead time and manufacturer/shipping modality. If there is a relationship it would allow you to cover your weaknesses and optimize shipping times.

# Data Exploration

Upon extracting the data, the first step was to explore the data and search for both inconsistencies and potential predictor variables that might have an impact on Lead Time. 2 problems were initially found: missing values and non-standardized column names. Solutions were built for both problems and are covered in the Data Cleaning section.

# Data Cleaning

As mentioned in the last section, the initial data exploration showed that there was missing data. The n\_miss () function was used from the Naniar package, and it reported 290 missing values. These missing values were divided equally into the Ship\_Date and Receipt\_Date columns. As these were date-time objects within R there was no way to impute the values. Impute refers to the process of attempting to calculate values that are missing from other data. Multiple hours of work were put in, but the code would return nonsensical values such as the 75th of April, which of course does not exist. Due to this the rows with the missing values were dropped since Ship\_Date and Receipt\_Date would be needed later to calculate both Lead Time and Manufacturing Time. This resulted in a decrease in observations from 9124 to 8979.

During the cleaning process certain column names were problematic and not standardized so the second part of cleaning the data was to insert an underscore (“\_”) in any column names with spaces. This made manipulating the data easier down the line. For Example: “Ship Date” became “Ship\_Date”. This did not remove any data or change any values but allowed for consistency across the data table.

# Data Transformation

The data also had to undergo transformations before it could be analyzed. The first transformation was to append year and quarter to the initial data set from a different data set to track shipping times over a longer time span. A function named date fixer was built to take in the Receipt Date and used the date bound of each quarter/year combination to figure which quarter/year each observation was in. This quarter/year was then appended to the data. This resulted in a data set with the same number of rows but two additional columns: year and row. These columns could potentially have some correlation with lead time, so it was vital that they existed in the data set before further analysis was conducted.

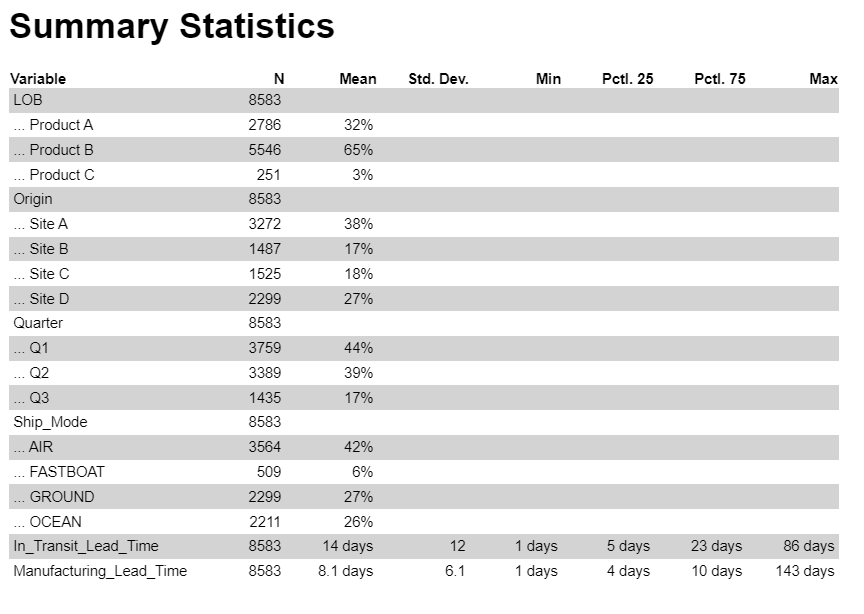
The other transformation was adding 2 more columns that were calculated data. The first was In\_Transit\_Lead\_Time which was Receipt\_Date – Ship\_Date. The second was Manufacturing\_Lead\_Time which was Ship\_Date – PO\_Download\_Date. These 2 newly generated columns are vital to future analysis. After that any rows in the newly generated columns were scanned for values below zero. A lead time cannot be less than 1 day so 396 rows were removed. This reduced the total number of rows from 8979 to 8583.

The year column also had to be removed since it was all in 2020 so it had to impact on the correlation.

The last transformation was converting all the data to numeric values since correlation function doesn’t work on non-numeric data. The categorical values were converted to factors first and then to numbers so correlation calculations would be possible. The dates were converted to numbers as well so correlation analysis could be carried out.

# Data Analysis

After the cleaning and transformations were completed, the summary statistics were generated for the columns that made sense. The summary stats (on next page) show that transit lead time on average is longer than manufacturing lead time.



After summary stats were compiled, the next step was to calculate possible correlation between variables in the data set. The numbers are on a scale of -1/solid red (strong negative correlation) to 1/solid blue (strong positive correlation). Graphical user interface

Description automatically generated with low confidence

Chart, histogram

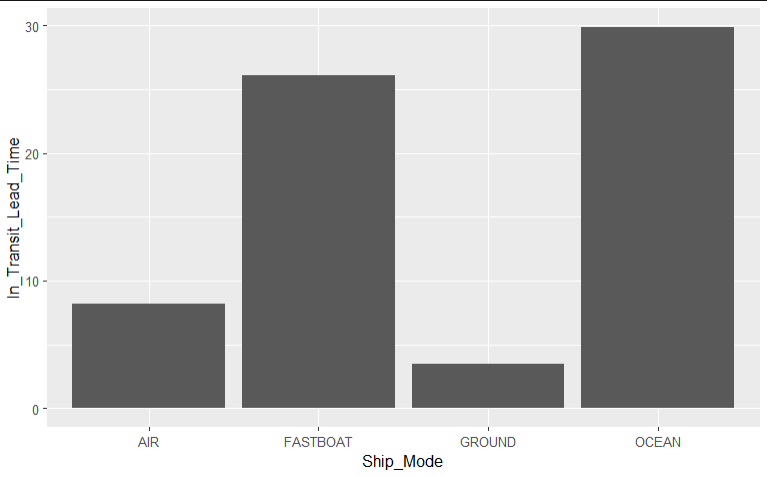
Description automatically generatedIt seems Origin has a significant negative correlation with In\_Transit\_Lead\_Time while Ship\_Mode has a significant positive correlation with In\_Transit\_Lead\_Time. When analyzing the correlation plots the bound for significant is plus/minus -.5 since there is a lot of correlation but not all of it relevant to the questions we are asking.

Further analyzing the impacts of Origin and Ship\_Mode generated the graphs below.

Chart, bar chart

Description automatically generated

This graph above plots In\_Transit\_Lead\_Time vs Origin shows that the Origin point impacts lead time since each site has rather different average lead times. This is further compounded by the graph below which is In\_Transit\_Lead\_Time vs Ship\_Mode. This shows that the average transit time varies by shipping method.



However, the real insights come from when the looking at In\_Transit\_Lead\_Time vs Origin colored by Ship\_Mode.

Chart

Description automatically generated

Here you can see that the sites don’t equally employ the same shipping method which could mean that is why some sites have lower average transit times.

# Conclusion

For example, weight or size of a package might impact transit time or complexity of a product might impact manufacturing time. There also could be an interaction effect between predictor variables. Further analysis is needed to determine which predictors matter and which predictor variables are useless. As can be observed in the last graph there is a high possibility of an interaction effect. But since correlation does not imply causation, no conclusions should be drawn since there could be another variable since no data set has all variables. Also, more variables describing the package itself might help draw better conclusions. After looking through and analyzing the correlation plots it seems Origin and Ship\_Mode both correlate with In\_Transit\_Lead\_Time but a definite conclusion cannot be drawn.