**What problem are you attempting to solve?**

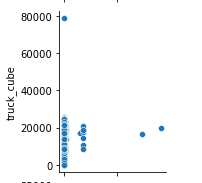
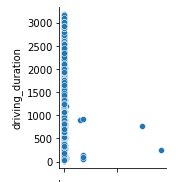
Was there any delay in shipping for certain transactions? If so, what factors might have contributed to delay in shipping and prevent delays in the future.

Predict number of delay days for a shipment

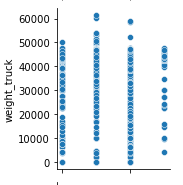
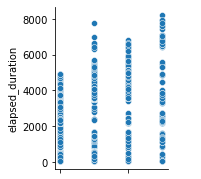
Predict cost of shipment based on average crude prices, season, and other metrics .

**What have you observed thus far in the data?**

In order to start exploring the data a bit further, we decided we wanted to create various scatterplots to try to get a sense of any existing relationships in the data, especially in regard to what could cause shipping delays since that was one of the problems we were attempting to solve. We started by creating a new column in the data to calculate the number of days a shipment arrives late by taking the dispatch date and subtracting from it the delivered date. We then created a new subset of the data containing only those shipments that arrived after the shipment date by setting all negative delays (early shipments) to zero, and then removing all rows where the days delayed equaled zero. Then, taking a closer look at the data through various scatterplots, using a pair plot, we were able to see that the number of shipments who were only delayed by one day was exponentially larger than shipments which were later than that.



To get a more accurate picture of the rest of the data, another subset was created by removing shipments that were only late by one day, and the same pairplot was created, allowing for a much clearer picture of the rest of the data. One interesting thing to take from the scatter plots is that the duration of the shipment seems positively correlated with the number of days a shipment is delayed. Specifically the elapsed duration seemed to be strongly correlated with shipping delays between two and four days.



Added additional information to the dataset

Extracted Inflation data from the website

<https://www.usinflationcalculator.com/inflation/current-inflation-rates/>

Merged inflation information to the shipping data based on dispatch date

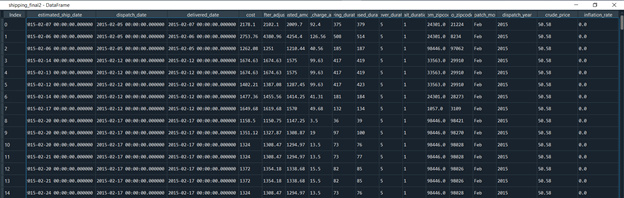
Extracted Crude oil prices data per month from the website

<https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=RWTC&f=M>

Added Season Data based on dispatch date

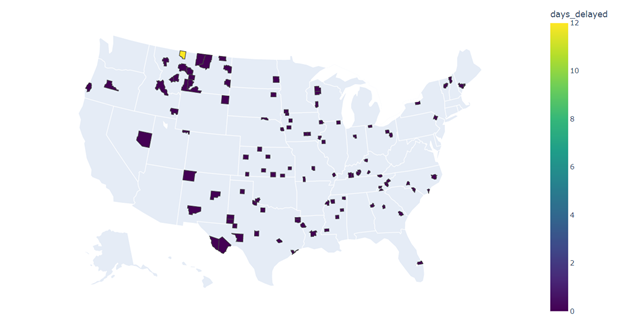
Created additional column shipping days to know the number of days between dispatch date and delivered date for a shipment

Combined all this external information to the shipping dataset

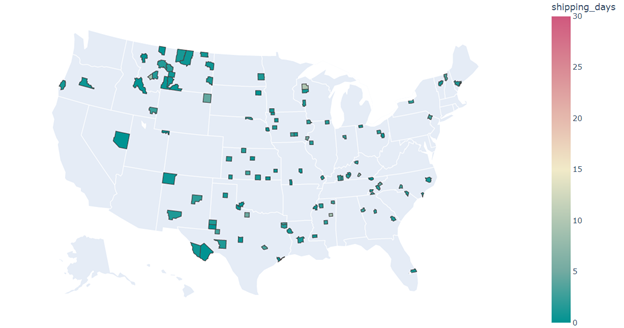


**Trends and observation:**

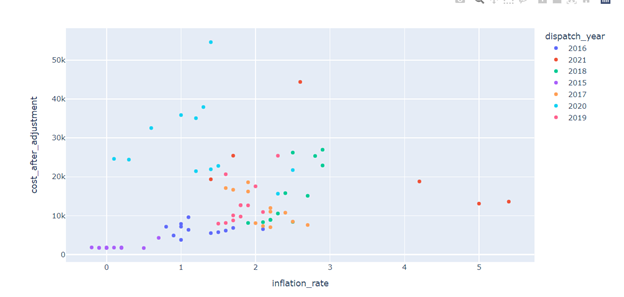
Delay Days by Zip Code



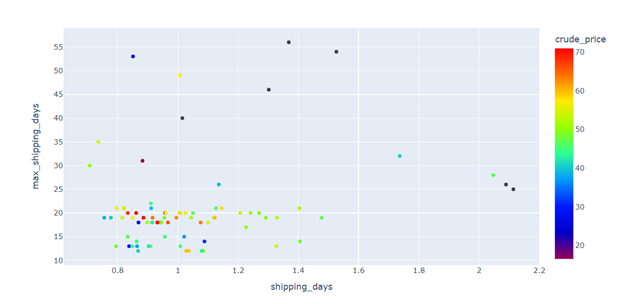
Average Shipping days by Zipcode



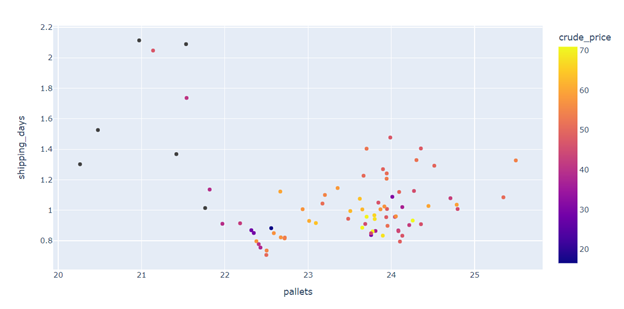
Inflation vs Cost



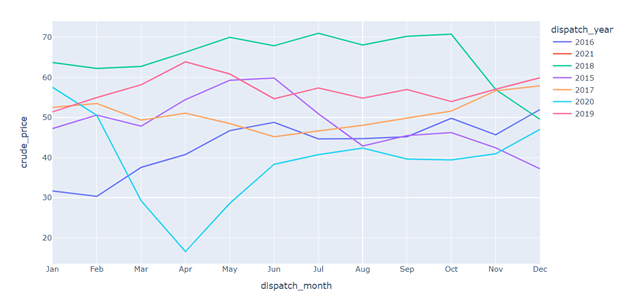
Shipping Days vs Maximum Shipping Days Vs Crude Price



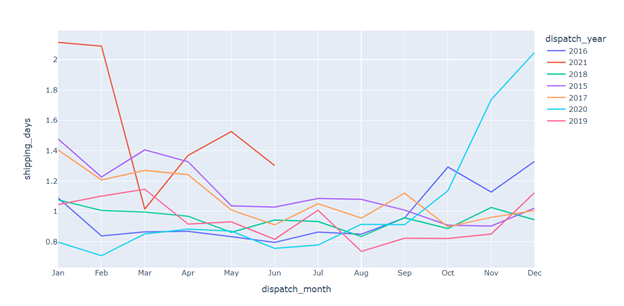
Number of Pallets vs Shipping days Vs Crude Price



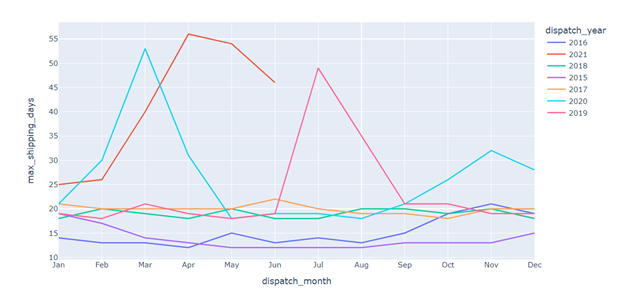
Crude Prices Trend



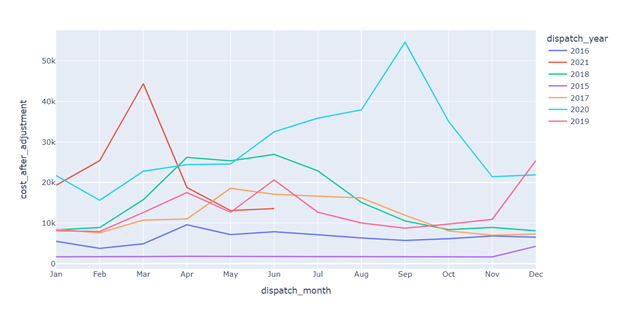
Shipping Days Trend



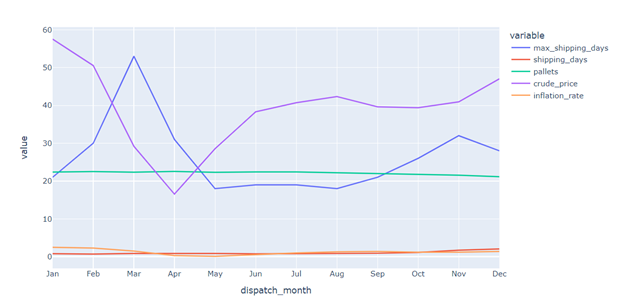
Maximum Shipping Days



Cost Trend



Comparing metrics for 2020



**What modeling techniques will you use?**

Logistic Regression for identifying delay in shipping

Linear regression for delay days or days taken to ship from one location to another

Time Series for predicting cost of shipping based on historical data

**What work do you still have to do?**

Now that our data manipulation, cleanup, validation, and analysis is complete, we need to apply our modeling techniques on the data to come up with answers to the problem we are solving