TIME SERIES ANALYSIS ON PRODUCE TRUCK LOAD SHIPMENTS

By

Ernest Barzaga, BS Mathematics

A thesis submitted to the Graduate Committee of

Ramapo College of New Jersey in partial fulfillment

of the requirements for the degree of

Master of Science

in Data Science

August 2024

Committee Members:

Debbie Yuster, Advisor

Osei Tweneboah, Reader

Ernest Santos, Reader

Table of Contents

Abstract 1

Chapter 1: Introduction & Industry Background 2

Section 1.2: Trucking Industry Overview 3

Section 1.3: Pricing Mechanisms 4

Section 1.4: Produce Transportation 5

Chapter 2: Data Preparation 8

Section 2.1: Data Overview 8

Section 2.2: Lane Deductions 9

Section 2.3: Miles Logic 14

Section 2.4: Data Density Requirement 15

Section 2.5: Interquartile Range (IQR) 16

Section 2.6: Data Cleaning Conclusion 16

Chapter 3: Exploratory Data Analysis 18

Section 3.3: Introduction to Stationarity 22

Section 3.4: Augmented Dickey Fuller Test 22

Section 3.5: Engle-Granger Causality Test 25

Section 3.6: Cointegration Test 27

Section 3.7: Decomposition 29

Chapter 4: Modeling Trucking Costs 33

Section 4.1: Introduction to Multiple Linear Regression 33

Section 4.2: Correlation Map 34

Section 4.3: Results from MLR model 35

Section 4.5: ACF and PACF 38

Section 4.6: Results from ARIMA (9,1,3) 39

Section 4.7: Introduction to SARIMA 42

Section 4.8: ACF and PACF 43

Section 4.9: Results from SARIMA (9, 1, 3) (3, 1, 3, 12) 44

Section 4.10: Closing thoughts on SARIMA (9,1,3) (3, 1, 3) 45

Chapter 5: Closing Thoughts 47

Section 5.1: Closing thoughts on modeling 47

Section 5.2: Data Limitations 47

Section 5.3: Multivariate factors 48
Section 5.4: Modeling Limitations 48
Section 5.5: Conclusion 50
Appendix 52
References 58

Abstract

The dataset for this project is sourced from a major freight broker based in New Jersey, with an annual revenue of approximately $200 million. The primary objectives are to implement techniques for handling missing data across the client's highest-volume lanes, to prepare the dataset for modeling and analysis, and to predict truck costs. Additionally, the project explores the impact of external macroeconomic factors on the trucking industry and their relationship to truck costs. In the modeling phase, analysis is focused on a single lane—Salinas, California, to the Bronx, New York—due to its high shipment volume for produce. Various machine learning models were evaluated on this lane, with ARIMA performing best when the year 2022 was excluded from the training set, resulting in a root mean squared error (RMSE) of $436. SARIMA performed best when 2021 was excluded, yielding an RMSE of $834. Based on this initial iteration of modeling, recommendations for future modeling techniques were made, including the use of a vector autoregression model. This suggestion arose from hypothesis tests (Engle-Granger) that indicated the collected macroeconomic factors may have predictive power regarding truck costs.

Chapter 1: Introduction & Industry Background

In this study, we utilize a dataset sourced from a freight broker based in New Jersey, with annual revenue totaling approximately $200 million. The primary aim of the research is to develop and implement methodologies for managing missing data and to prepare the dataset for thorough analysis. While initial modeling techniques are introduced, the focus is primarily on the foundational work necessary to understand the data and the broader impact of external macroeconomic factors on the trucking industry. The study aims to establish a solid groundwork for future modeling efforts by exploring how these factors correlate with truck costs and by providing insights into potential directions for more advanced analyses in subsequent iterations.

Over the past 35 years, our client has developed a robust network of thousands of carriers. They have strived to stay on the cutting edge of logistics technology. This has been achieved through significant investments in advanced transportation management systems (TMS), which serve as end-to-end tools for data entry and capture for all shipments in the dataset. The company also utilizes pricing benchmark technologies such as DAT and Sonar, alongside a well-organized database structure that provides access to all historical company data. These technological advances have enabled the undertaking of a project like this.

The data used in this study has never been analyzed using machine learning techniques. The data mining conducted here will lay the foundation for future AI initiatives within the company. By studying and addressing business logic, we have also developed strategies to manage data gaps commonly found in company systems (as discussed in Chapter 2). This foundational work not only supports the current project but also positions the company to leverage advanced analytics and AI for future growth and efficiency improvements.

This study provides a comprehensive data analysis of all major produce lanes within the company to establish a foundation for future research. The actual modeling, utilizing ARIMA, SARIMA, Multiple Linear Regression, and Vector Autoregression (VAR), is focused specifically on the lane from Northern California (represented by Salinas) to the Bronx, NY, due to its high shipment volume. Modeling for other regions will be reserved for future work.

Section 1.2: Trucking Industry Overview

If it ends up on your dining room table, it was likely hauled by a carrier from a shipper, possibly with the help of a broker. There are roughly 13 million registered trucks in the U.S. These trucks generated approximately $940.8 billion in revenue in 2022, with about 1.86 million companies operating semi, straight, and hazardous materials trucks [1].

The trucking industry is dominated by the relationship between shippers, who own the goods to be transported, and carriers, who own the assets used to haul these goods [2]. There is a third player, the broker, who is responsible for supplying carriers when a shipper cannot find sufficient carriers to move their product. The need for brokers arises when capacity in the trucking industry becomes tight, meaning there are more products to be shipped than there are available trucks to move them. In such situations, the network and reach that a broker possesses becomes highly valuable for shippers. Brokers leverage their connections with carriers to secure the necessary trucks, and they make a margin by charging shippers a slightly higher rate than the cost they pay to the carriers. This premium compensates brokers for their expertise and the service they provide in connecting shippers with available carriers. Brokers not only secure carriers for shippers but also handle many additional tasks that simplify the shipping process. They negotiate for the best rates, provide digital services to track the location of trucks, facilitate communication between shippers and carriers, and plan routes. This comprehensive service reduces the operational burden on shippers and ensures efficient and cost-effective transportation of goods [3].

Section 1.3: Pricing Mechanisms

There are two types of carrier rates discussed: spot rates and contract rates. Spot rates apply to one-time shipments that are transactional and fluctuate based on current market conditions. Contract rates are pre-negotiated and remain fixed regardless of market changes. This paper focuses on spot rates, as most carriers acquired by the client operate on the spot market. Carriers that use contract rates were excluded from the dataset during analysis.

Spot rates are negotiated on the spot, with carriers often increasing prices as their capacity decreases and the demand for trucks surges in a region. The negotiation process also involves shippers, who may need to renegotiate the selling price as market conditions change.

Carriers, shippers, and brokers all use major load boards which are online lists of available trucks and loads based on origin city/state and destination city/state (what are called “lanes”). If a shipper has freight, they can post their load on these boards from these lanes. The same goes for carriers that have available assets (trucks) in the area. Some load boards, such as DAT Solutions, offer benchmarking on lanes by collecting and normalizing data from contributors, including clients. This benchmarking provides a reference point for lane pricing and is a factor in rate negotiations, as both carriers and brokers have access to this information. However, DAT uses historical data, which may not always predict changing market conditions. DAT provides does not provide benchmarks for produce, lumping produce and non-produce (assuming the product is temperature controlled) into the refrigerated equipment category.

Section 1.4: Produce Transportation

The key difference between produce shipments and non-produce refrigerated shipments lies in the urgency of the timeline. Produce begins to perish the moment it is harvested, creating a time-sensitive demand that strains the trucking market during certain months, leading to price increases. This seasonality is a characteristic of produce shipments, unlike non-produce shipments, which do not exhibit the same patterns. Most produce markets begin in March and extend through July, sometimes into August. Figure 1.4.1 illustrates a monthly aggregation of truck costs for a non-produce lane for the client, and as shown, while there is a general downward trend, no clear pattern is evident

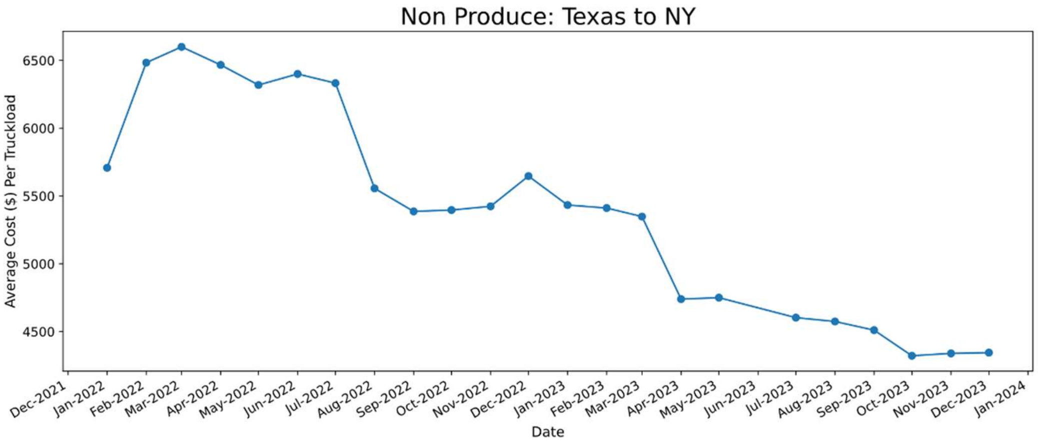


Figure 1.4.1 Graph of Non-Produce Lane: TX to NY: Cost vs Pick up Date

Figure 1.4.2 illustrates the time series of truck costs for a produce lane, highlighting the seasonality observed in the years 2022 and 2023. In both years, there is a noticeable rise in costs beginning in March, peaking in June before declining. This recurring pattern reflects the produce harvest season, which typically starts around March, leading to increased demand for trucks. As a result, truck rates rise, reaching their highest point in June when the harvest is at its peak. After

June, as the harvest season winds down, the demand for trucks decreases, leading to a subsequent decline in costs.

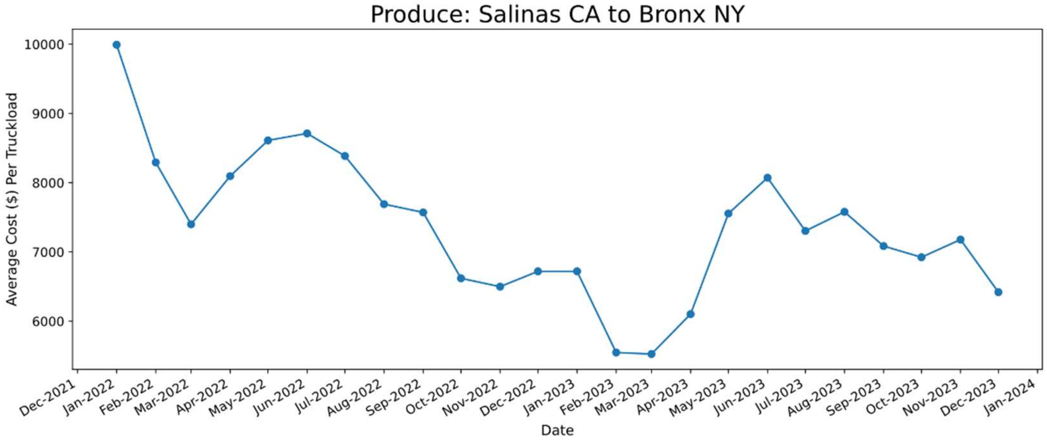


Figure 1.4.2 Graph of Produce Lane: Salinas CA to Bronx NY: Cost vs Pick up Date

Section 1.5: Literature Review

David Sokoloff's paper titled *Predicting and Planning for the Future: North American Truckload Transportation* laid the groundwork for getting us to think about machine learning possibilities in predicting truck pricing. Sokoloff developed a machine learning model to predict the US truckload dry van spot rate. The model achieved impressive accuracy, with an average Mean Absolute Percentage Error (MAPE) below 7% and a mean error below 0.05 for 12-month forecasts. This enabled companies to mitigate risks and unplanned costs from market volatility. While Sokoloff's study focuses on predicting dry van spot rates in freight, my study will focus on produce refrigerated rates [4].

Sokoloff's study employed vector autoregression (VAR) to capture the linkages between multiple variables throughout a time series. VAR predicts endogenous variables by calculating their own lagged values, the lagged values of other endogenous variables, and the error values. We used similar variables to analyze their relationships with our produce data. However, we ultimately left VAR for further work as we did not meet the data requirements for VAR. While Sokoloff used VAR for long-term forecasting, we recommend a short-term forecast for VAR in future work to incorporate more variables into our multivariate modeling.

To address the need for short-term forecasting, Sokoloff utilized the autoregressive integrated moving average (ARIMA) method. Similar to VAR, ARIMA uses time series data and lagged observations to inform future predictions. However, ARIMA is suited for non-stationary series, employing differencing to achieve stationarity. We employed ARIMA and SARIMA models, which account for seasonality, for our long-term forecast.

The ARIMA model in Sokoloff's study focused on univariate analysis of the monthly national DAT spot rate, utilizing a dataset from January 2009 to March 2020. The model parameters included a lag order (p) of 4, a degree of differencing (d) of 1, and a moving average order (q) of 2, defined as ARIMA (4,1,2). Diagnostic checks for autocorrelation, stationarity, and heteroscedasticity confirmed the model's fitness.

Chapter 2: Data Preparation

Section 2.1: Data Overview

The client's shipment data was initially divided into two periods: 2019-2021 and 2021-2023. Each row in these datasets corresponds to a unique shipment and includes information on the customer, carrier, number of stops (except the 2019-2021 data, which lacks stop information), produce or non-produce designation, and details about the margin the client received from each shipment. There is a Lane ID column which designates a particular origin and destination region which are defined internally by the client. These Lane IDs are Arizona to NJ, Desert to NJ, Florida to NJ, Georgia to NJ, Cali Oxnard & Santa Maria to the Bronx NY, NoCal Salinas to the Bronx NY, Pacific Northwest to NJ, and Texas to the Bronx NY. These are not all the lanes our client operates in but represent the highest volume for produce specific shipments.

The dataset for 2019-2021 includes a mix of single-stop shipments (one pickup and one delivery) and multi-stop shipments (more than one pickup or delivery). However, it does not provide specific columns to indicate the number of pickups and stops, which leads to inflated costs and total mileage. In contrast, the dataset for 2021-2023 includes columns that specify the number of pickups and stops, but neither dataset separately breaks out the charges for each additional stop. To address this, sections 2.2 and 2.3 outline the steps taken to normalize the data from 2019 through 2021, minimizing the skew that multi-stop shipments have on the cost.

Section 2.2: Lane Deductions

In this section, we aim to calculate a lane-specific deduction to truck costs to account for the inflated expenses associated with multi-stop shipments, which include additional pickups or drop-offs.

The original 2021-2023 dataset, prior to filtering, included columns that specified the number of pickups and drop-offs for each shipment. Each additional pickup or drop-off represents an extra stop, which typically increases the cost of the shipment due to higher labor and operational expenses, such as fuel for the carrier. We conducted an exploratory analysis comparing the cost distribution between single-stop and multi-stop shipments from each of the client's regions to the East Coast. Approximately 30% of all shipments across the lanes were multi-stop.

To quantify the cost difference, we calculated the average price of multi-stop shipments and subtracted the average price of single-stop shipments. This difference will be used as a deduction for shipments in the 2019-2021 dataset that we suspect are multi-stop based on the analysis detailed in section 2.3. The 2019-2021 dataset does not include columns that indicate whether a shipment is single or multi-stop, so this adjustment is necessary.

Figures 2.2.1 through 2.2.8 below show the distribution of single stop versus multiple stop shipment for each region to a representative destination on the East Coast for 2021-2023. The X-axis, labeled "Cost Bucket – Line Haul," represents the cost paid by the client for the truck. We selected the most popular destinations by shipment volume for these displays. Additionally, we compared the average cost of a random sample (with replacement) of both single-stop and multi-stop shipments to the average cost of the original dataset.

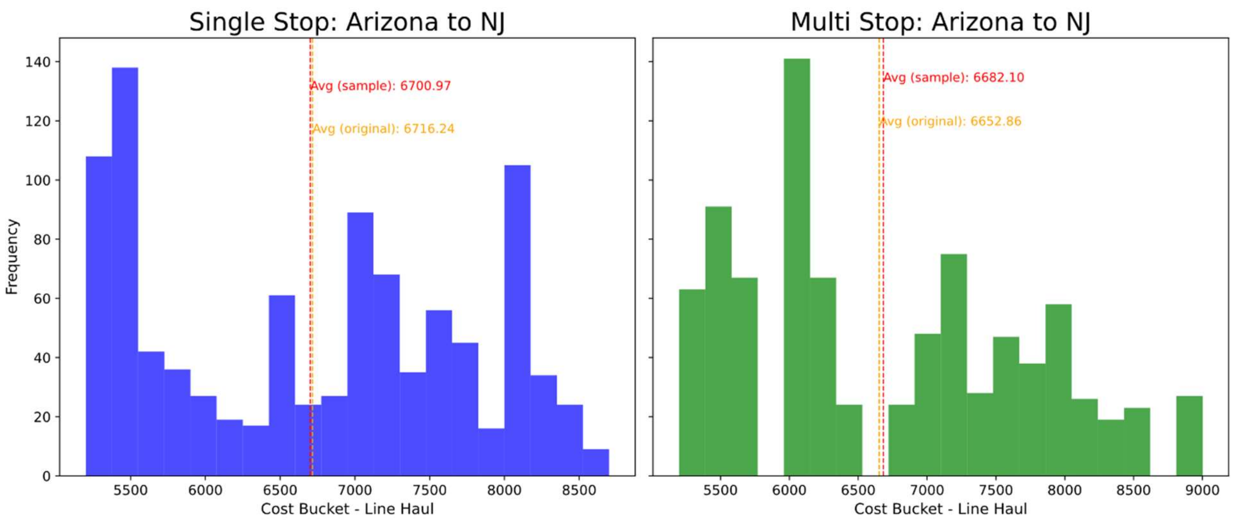


Figure 2.2.1 Distribution of Cost for single shipment (left) to multi stop (right) for AZ to NJ

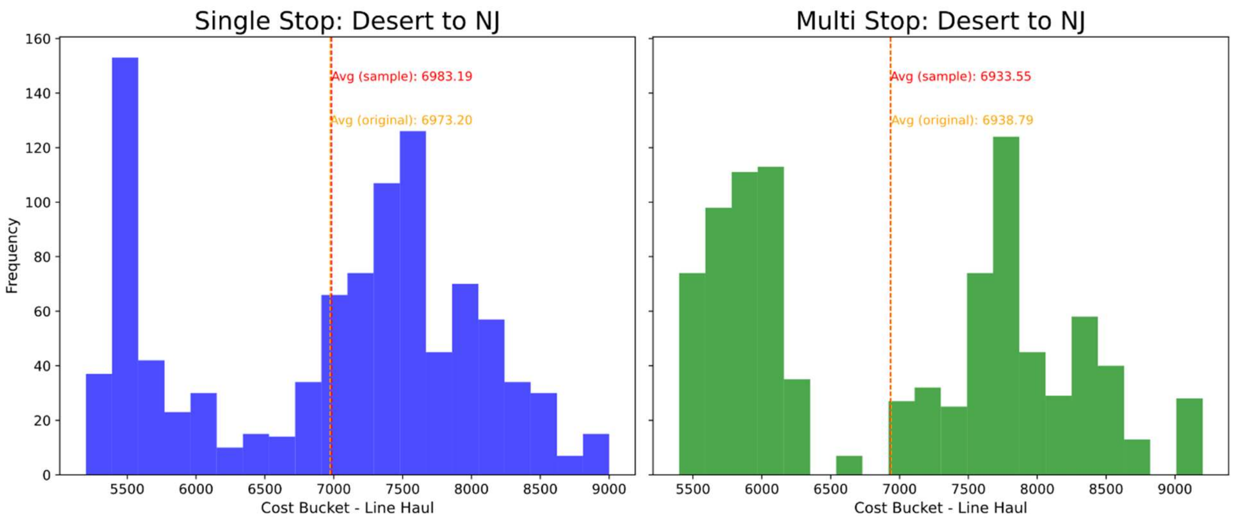


Figure 2.2.2 Distribution of Cost for single shipment (left) to multi stop (right) for Desert to NJ

The difference in averages between Arizona to NJ and Desert to NJ for single-stop and multi-stop shipments is small. This indicates that there is a minimal premium for additional stops. Therefore, we will not be deducting from the cost in the 2019-2021 data for these lanes.

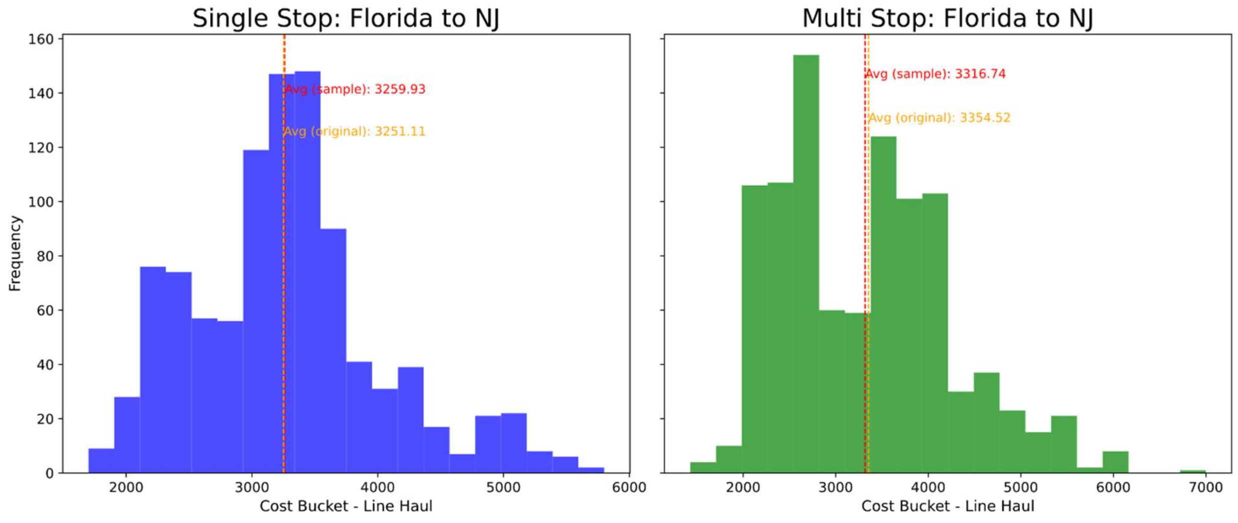


Figure 2.2.3 Distribution of Cost for single shipment (left) to multi stop (right) for FL to NJ

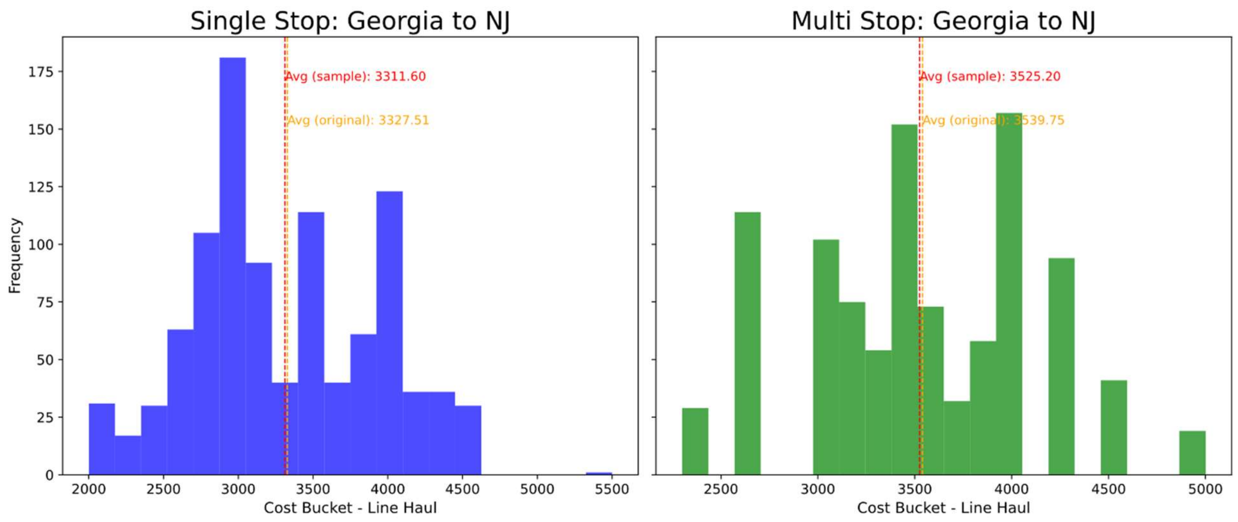


Figure 2.2.4 Distribution of Cost for single shipment (left) to multi stop (right) for GA to NJ

Florida and Georgia to NJ show differences of $57.74 and $214, respectively. Since these two regions are very similar at certain times of the year, I will take the average of these differences to deduct from the cost. This amounts to a lane deduction of $135.

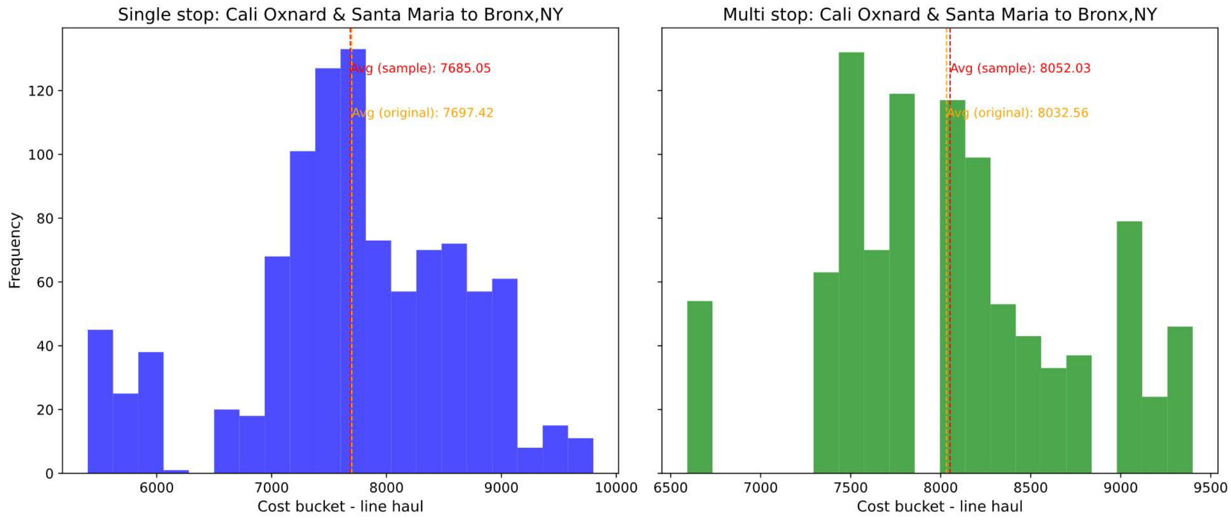


Figure 2.2.5 Distribution of Cost for single shipment (left) to multi stop (right) for Santa Maria & Oxnard CA to Bronx, NY

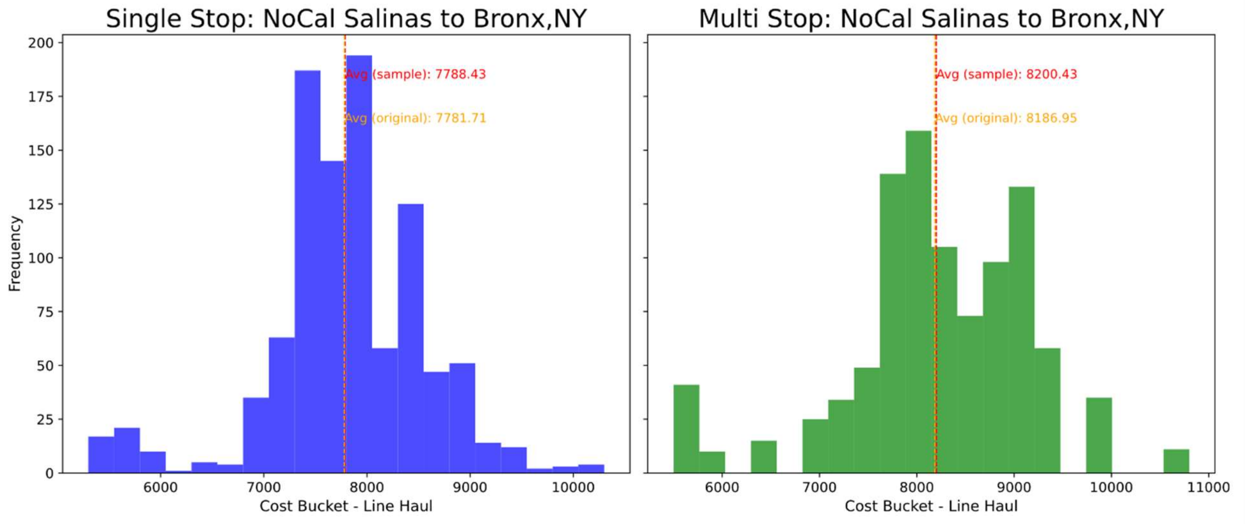


Figure 2.2.6 Distribution of Cost for single shipment (left) to multi stop (right) for NoCal Salinas to Bronx, NY

Cali Oxnard & Santa Maria (Southern California) and NoCal Salinas (Northern California) have differences of $367 and $412, respectively. These differences will be used to calculate lane deductions after mileage analysis. The difference is substantial, so further exploratory data analysis (EDA) is needed to ensure we do not deduct more than necessary.

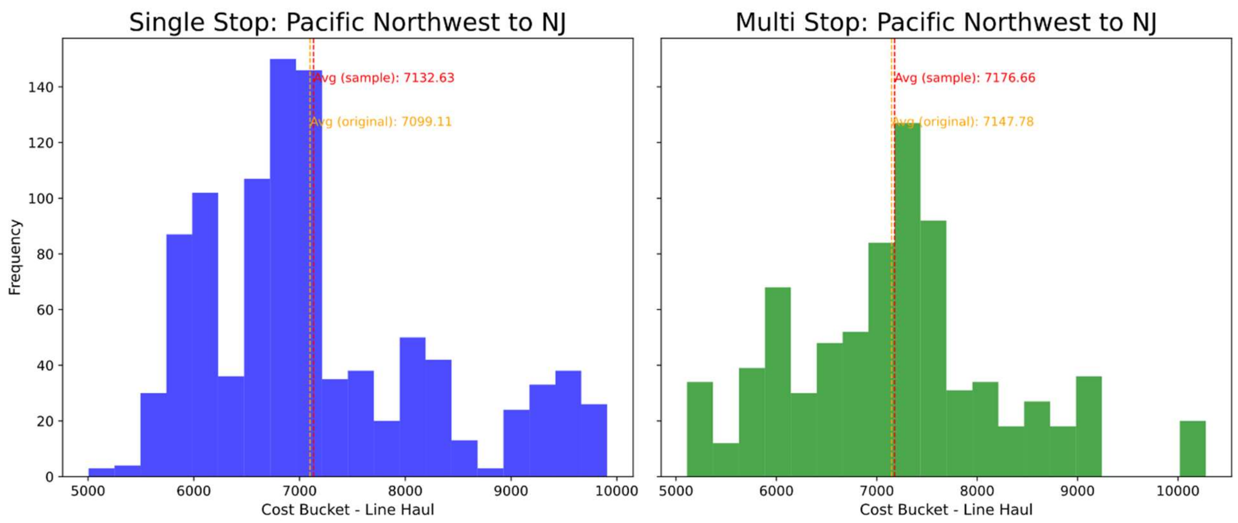


Figure 2.2.7 Distribution of Cost for single shipment(left) to multi stop (right) for Pacific Northwest to NJ

The Pacific Northwest (which includes Washington) shows no significant differences between single and multi-stops. Therefore, no lane deduction will be applied to the 2019-2021 lanes out of the Pacific Northwest.

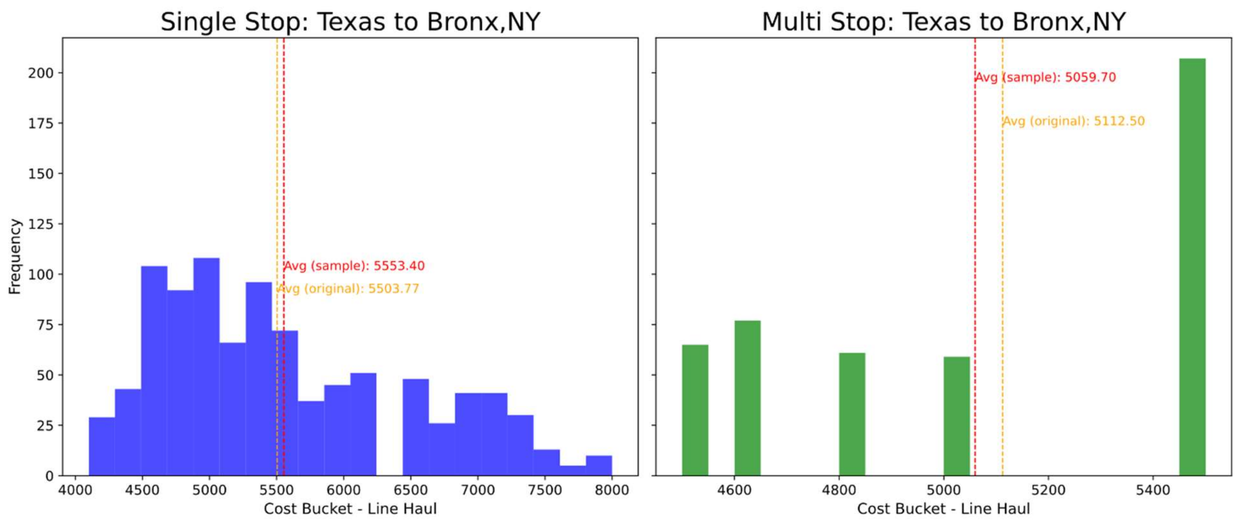


Figure 2.2.8 Distribution of Cost for single shipment(left) to multi stop (right) for TX to NJ

There are several gaps in the distribution on Figure 2.2.8 for Texas to Bronx NY that suggest a lack of variety in the rates as well as not much data density. Therefore, these shipments will retain the cost listed in the 2019-2021 data.

Section 2.3: Miles Logic

In this section, we explain the methodology used to apply the lane deductions discussed in section 2.2. Since we do not know which shipments in the 2019-2021 dataset are multi-stop, we used information from the 2021-2023 dataset to identify which shipments in the 2019-2021 dataset should have the deduction applied. This process involves examining the 2021-2023 dataset, where single stop shipments are identified. The total miles variable for these single stop shipments reflects the approximate direct mileage, unaffected by additional stops.

We grouped the single stop shipments by 3-digit ZIP code pairs, indicating the origin and destination ZIPs for each shipment. For example, the pair 089-104 indicates that the shipment picks up from the 089 ZIP code and delivers to the 104 ZIP code. We then calculated the mileage range by subtracting the lowest total miles in each 3-digit ZIP group from the highest, establishing a minimum and maximum for the single stop shipments.

Shipments in the 2019-2021 dataset with total miles exceeding the maximum mileage were classified as potential multi-stop shipments and subjected to the lane deductions outlined in Section 2.2. These deductions were applied to every origin region shipping to any East Coast state. For California regions (NoCal Salinas, Santa Maria, and Oxnard), the lane deduction of $416 was only applied to shipments that were at least 100 miles above the maximum direct mileage. Based on domain knowledge, we know that a deduction of $416 is substantial, so our threshold for applying this deduction was higher than for other lanes.

Section 2.4: Data Density Requirement

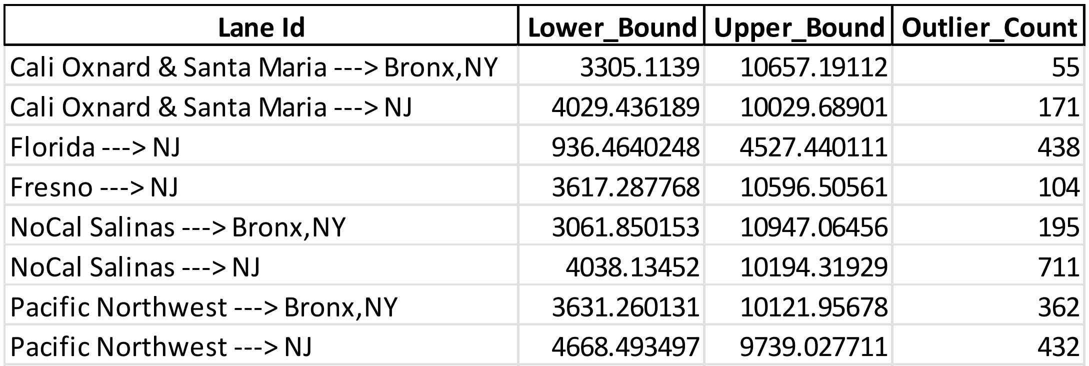
To address the issue of lanes with low shipment volume and limited carrier diversity, we implemented a filtering process to ensure that each lane has sufficient unique carriers and shipment volume. This prevents any single carrier from monopolistically setting market conditions. If only one carrier is used for a particular lane in a given month, it reflects that carrier's pricing rather than providing a reliable indicator of what the client spent on freight in that market.

The filtering process involves grouping data by Lane ID (e.g., "NoCal Salinas → Bronx, NY"), month, and year of shipment pickup. We then applied the criteria that each lane must have at least 5 different carriers and 5 loads during the months of April to August for each year from 2019 to 2024. If a lane fails to meet these criteria in any month within this range, it is excluded from future modeling as it does not meet our data requirements.

This filtering reduced our dataset from 54,474 shipments to 34,743, thereby enhancing the reliability of our modeling. The methodology was inspired by DAT's approach to benchmarking carrier rates, which requires at least 3 different carriers and a combined total of 7 loads to meet the criteria for reporting a rate for a lane [5].

Section 2.5: Interquartile Range (IQR)

To ensure that outliers do not skew our data, we employed the interquartile range (IQR) statistical technique. This method identifies outliers in shipment carrier costs that fall below or above 1.5 times the IQR, where the IQR is defined as the difference between the 3rd quartile (Q3) and the 1st quartile (Q1). For each lane, we calculated the IQR and used it to establish lower and upper bounds for identifying potential outliers. As outlined in Table 2.5.1, outliers were detected and removed for lanes such as Cali Oxnard & Santa Maria to the Bronx NY, NoCal Salinas to NJ, and Pacific Northwest to NJ. Lanes like Arizona to NJ and Texas to the Bronx NY did not exhibit any outliers in the data, ensuring a cleaner dataset for analysis.



Lane Id	Lower_Bound	Upper_Bound	Outlier_Count
Cali Oxnard & Santa Maria ---> Bronx,NY	3305.1139	10657.19112	55
Cali Oxnard & Santa Maria ---> NJ	4029.436189	10029.68901	171
Florida ---> NJ	936.4640248	4527.440111	438
Fresno ---> NJ	3617.287768	10596.50561	104
NoCal Salinas ---> Bronx,NY	3061.850153	10947.06456	195
NoCal Salinas ---> NJ	4038.13452	10194.31929	711
Pacific Northwest ---> Bronx,NY	3631.260131	10121.95678	362
Pacific Northwest ---> NJ	4668.493497	9739.027711	432

Table 2.5.1: Outlier bounds, and the number of outliers removed for each lane

Section 2.6: Data Cleaning Conclusion

For the 2021-2023 period, there were initially 85,100 shipment records. This dataset included columns for the number of pick-ups and drops. After filtering for shipments with one pick-up and one drop, as well as those with only one additional stop, applying a $200 lane deduction on the truck cost for these additional stop shipments, and focusing on refrigerated produce shipments from the client's region origins to East Coast destinations, the dataset was reduced to 24,108 records.

For the 2019-2021 period, there were initially 97,711 records, each representing a full truckload shipment. After filtering for produce and refrigerated shipments and focusing on the client's region origins to East Coast destinations, the dataset was reduced to 37,757 rows.

Following the initial filtering, we further refined the dataset by excluding shipments from a customer for whom we have contracted carrier rates. This adjustment resulted in a total of 21,336 records for the years 2022-2023, and 33,138 records for the period 2019-2021.

After combining the two datasets the total number of records was 54,474.

Chapter 3: Exploratory Data Analysis

In this section we explore a select group of macroeconomic factors that may impact a broker's cost for a truck. We examine similar endogenous variables as those cited in David Sokoloff's paper to explore the correlations between these variables and our data [4]. From this point forward, our focus narrows to the lane from Salinas, California, to the Bronx, New York. This lane has the highest volume, meaning the client uses it more than any other lane to for produce product. In fact, this lane is often used as a benchmark for pricing other lanes where data is sparse. Consequently, the multivariate analysis that follows will compare the time series of other variables to those of the Salinas-to-Bronx Lane.

Section 3.1: Variable Explanation

EIA Fuel Data: This dataset contains monthly diesel prices per gallon by region, provided by the Energy Information Administration (EIA). Each record is assigned a fuel price based on its region, e.g., California (CA) is in a region by itself, while Washington and Oregon are included in the West Coast region [6].

OTRI Origin and Destination: The Outbound Tender Rejection Index (OTRI) reflects the balance of supply and demand in the freight and logistics industry. A high OTRI indicates high demand and low truckload capacity, leading to higher outbound shipping rates. Conversely, a low OTRI suggests lower demand and higher capacity. Both origin and destination are assigned their corresponding OTRI score. These values were aggregated monthly [7].

CASS Datasets:

Cass is a freight payment management company that publishes datasets on operational costs for trucking companies. All of the datasets listed below are aggregated monthly.

- Expenditures Value: Total amount spent in the US on freight shipping, including all costs.
- Shipments Value: Volume of shipments processed, indicating shipping activity and demand.
- Inferred Rates: Average rate per shipment, derived from dividing total expenditures by total shipments.
- TL LH Index: Changes in linehaul rates, excluding additional fees. Line haul refers to the truck cost [8].

FRED Datasets:

The Federal Reserve Economic Data is a database that contains more macroeconomic datasets

- All Employees, Truck Transportation: Index representing the number of employees in the trucking sector [9].
- Export Price Index (NAICS): Crop Production: Measures average month by month change in prices received by domestic producers for crop exports [10].
- Producer Price Index: USDA PPI for food measures the average change in selling prices received by domestic producers for their output [11].

Section 3.2: Time Series vs. Macroeconomic Factors

Figure 3.2.1 illustrates how the target variable which is the cost for the truck (Adjust LH) in the graph below and the macroeconomic trucking variables changed with respect to the carrier pickup date for Salinas CA to the Bronx NY. Both the OTRI for destination and origin show a spike at the beginning of 2021 that sustained until early 2022. This trend is also evident in the Adjusted LH, which exhibits a similar spike in 2021 and a subsequent decline around the beginning of 2022. Cass Expenditures, Inferred Rates, and the TL LH Index show a gradual increasing trend throughout this period, followed by a slower decline. The FRED number of employees exhibits a steady increase during the same period, which could indicate that the demand for trucks was met by an influx of new carriers entering the market. Gas prices also increased during this period, likely due to rising demand. Crop production started increasing before Adjusted LH and the other macroeconomic factors. This might help explain the cause in demand and should be studied further by examining other factors associated with supply chain supply and demand such as import and exports, all of which need trucks to haul goods to and from ports of entry.

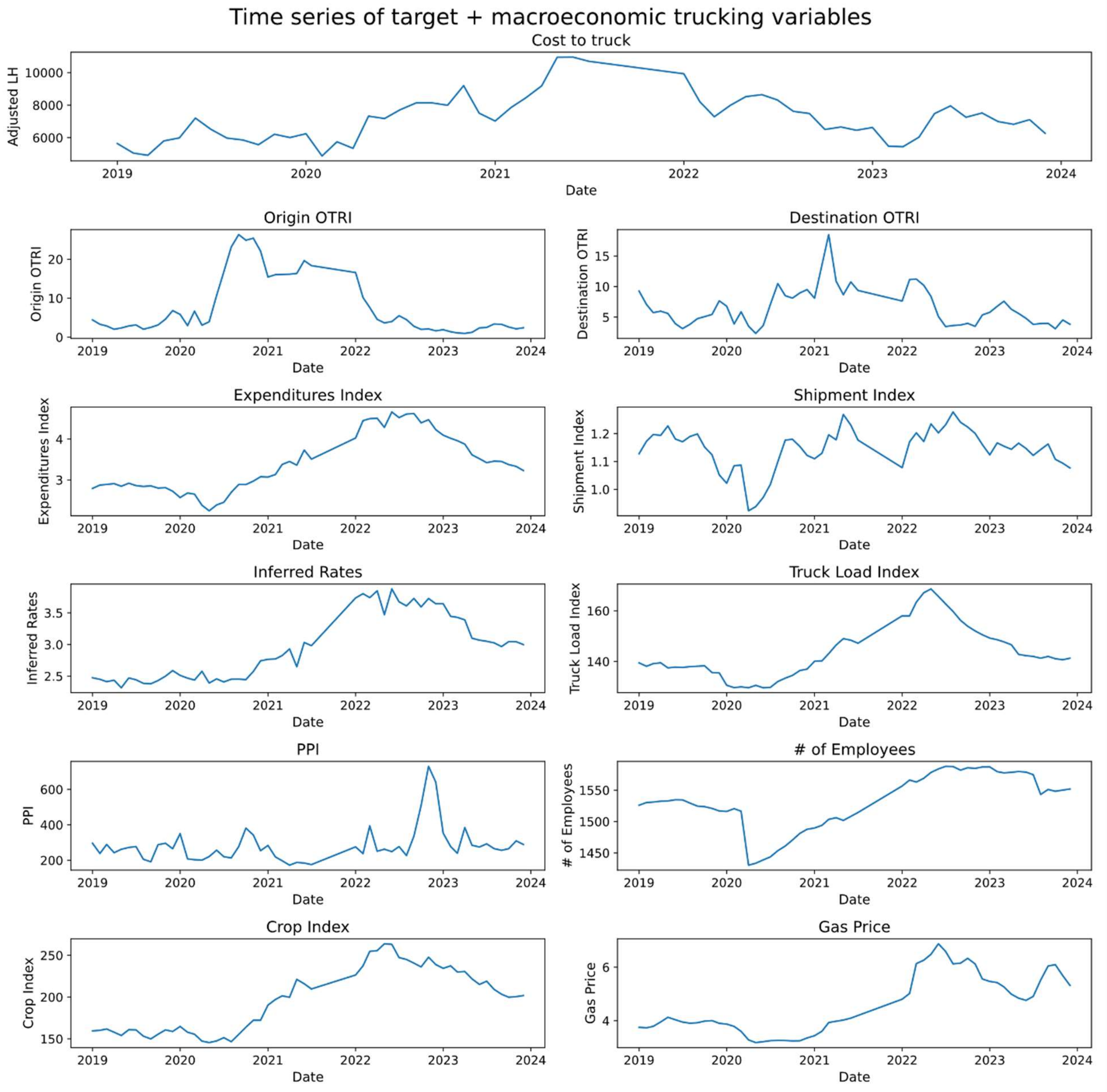


Figure 3.2.1 Cost to truck (target) and the exogeneous variables vs pickup date.

Section 3.3: Introduction to Stationarity

In this section, we introduce the concept of stationarity, an important assumption for the time series analysis that follows in Chapter 4.

A time series is considered stationary if it satisfies the following conditions:

i. Constant Mean: The mean value function of the series remains constant over time and does not depend on the specific time at which it is measured.
ii. Time-Invariant Autocovariance: The autocovariance of the series depends only on the time difference between two observations, not on the actual time at which the observations are taken. This implies that the statistical properties of the series, such as variance and covariance, do not change over time [12].

To test whether a time series is stationary or not we will employ the Augmented Dickey Fuller test below. If needed, we will difference the time series which involves taking the difference between adjacent observations [13].

Section 3.4: Augmented Dickey Fuller Test

The Augmented Dickey-Fuller (ADF) test is a statistical test used to determine if a time series is stationary or if it contains a unit root, which indicates non-stationarity [12].

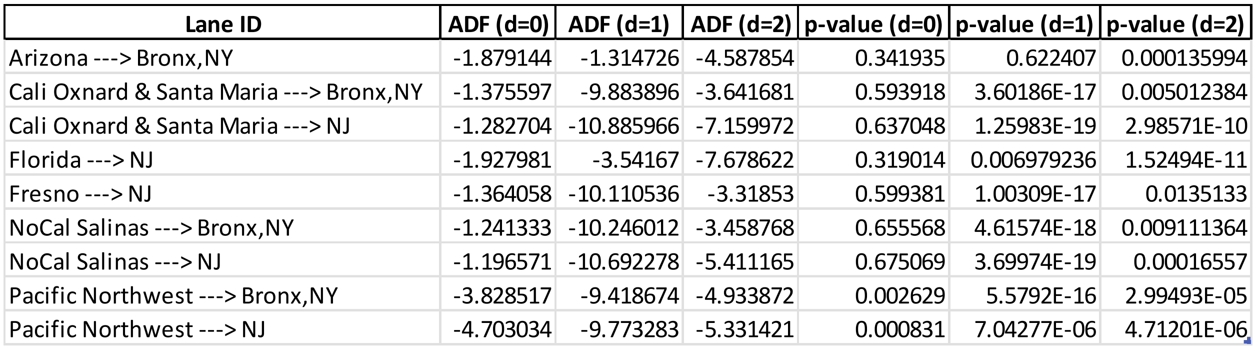
ADF test explanation

- Null Hypothesis: The time series has a unit root (i.e., it is non-stationary).
- Alternative Hypothesis: The time series does not have a unit root (i.e., it is stationary).

ADF Test Statistic (Score)

- The ADF test statistic is a negative number. The more negative it is, the stronger the evidence against the null hypothesis of a unit root.
- Interpretation:
    - Less Negative / Closer to Zero: Indicates weaker evidence against the null hypothesis (more likely the series is non-stationary).
    - More Negative / Further from Zero: Indicates stronger evidence against the null hypothesis (more likely the series is stationary) [14].

The significance level indicates whether my ADF value is sufficiently negative enough to reject the null hypothesis. The significance level of p = 0.05 was chosen. The results of the ADF test for each Lane ID, specifically on the Cost variable, are shown below in Table 5.2.1. These differencing components represent the number of times a time series must be differenced to make the time series stationary. Later in the paper, we will use these components to run ARIMA and SARIMA models.

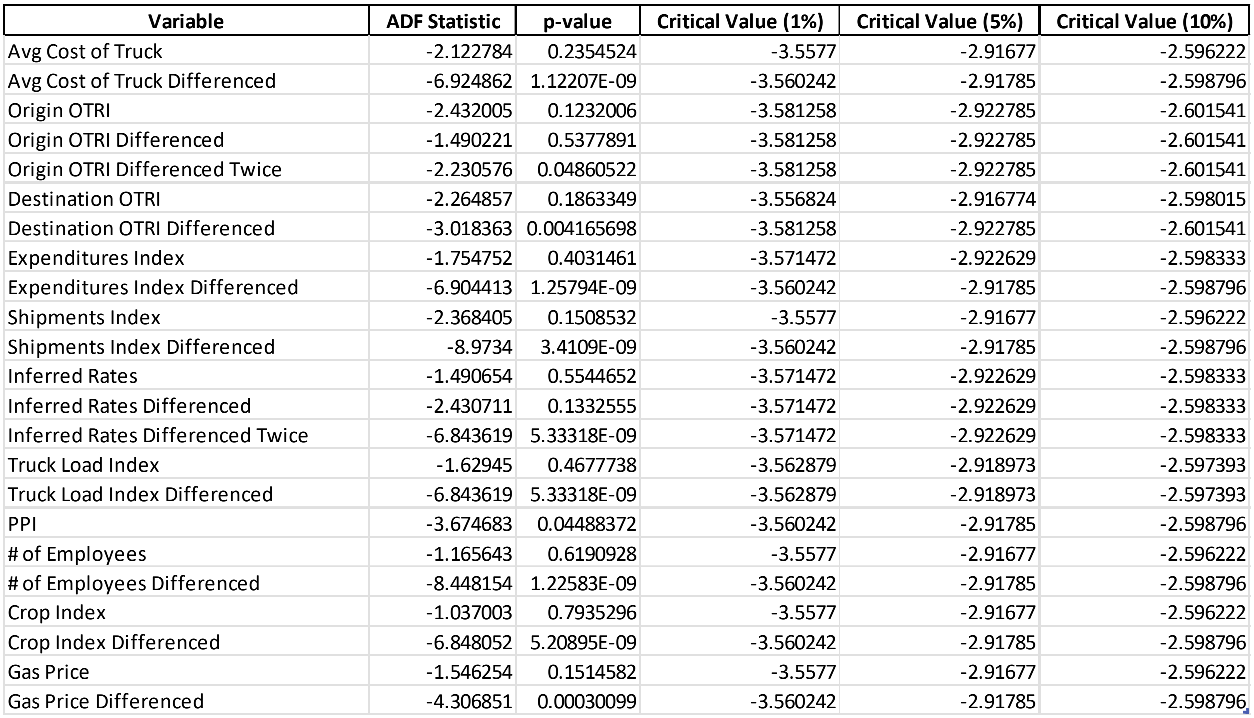


Lane ID	ADF (d=0)	ADF (d=1)	ADF (d=2)	p-value (d=0)	p-value (d=1)	p-value (d=2)
Arizona ---> Bronx,NY	-1.879144	-1.314726	-4.587854	0.341935	0.622407	0.000135994
Cali Oxnard & Santa Maria ---> Bronx,NY	-1.375597	-9.883896	-3.641681	0.593918	3.60186E-17	0.005012384
Cali Oxnard & Santa Maria ---> NJ	-1.282704	-10.885966	-7.159972	0.637048	1.25983E-19	2.98571E-10
Florida ---> NJ	-1.927981	-3.54167	-7.678622	0.319014	0.006979236	1.52494E-11
Fresno ---> NJ	-1.364058	-10.110536	-3.31853	0.599381	1.00309E-17	0.0135133
NoCal Salinas ---> Bronx,NY	-1.241333	-10.246012	-3.458768	0.655568	4.61574E-18	0.009111364
NoCal Salinas ---> NJ	-1.196571	-10.692278	-5.411165	0.675069	3.69974E-19	0.00016557
Pacific Northwest ---> Bronx,NY	-3.828517	-9.418674	-4.933872	0.002629	5.5792E-16	2.99493E-05
Pacific Northwest ---> NJ	-4.703034	-9.773283	-5.331421	0.000831	7.04277E-06	4.71201E-06

Table 3.4.1 ADF p-value and critical value results for each lane

We focus on “NoCal Salinas → Bronx, NY,” to run Multiple Linear Regression (MLR) and Vector Autoregression (VAR). This will involve performing another ADF test on each variable within this lane to ensure stationarity.

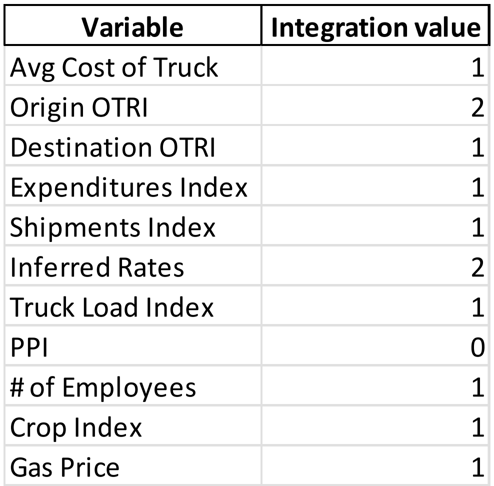
The test results show that ‘Adjusted LH’, ‘Destination OTRI’, ‘Cass Expenditures Value’, ‘Cass Shipments Value’, ‘Cass TL LH Index’, ‘FRED Num Employees’, ‘FRED Crop Production’, and ‘Gas Price’ required first differencing to become stationary. ‘Origin OTRI’ and ‘Cass Inferred Rates’ required second differencing. The ‘PPI’ variable was already stationary. These differencing components are essential for accurate ARIMA and SARIMA modeling.



Variable	ADF Statistic	p-value	Critical Value (1%)	Critical Value (5%)	Critical Value (10%)
Avg Cost of Truck	-2.122784	0.2354524	-3.5577	-2.91677	-2.596222
Avg Cost of Truck Differenced	-6.924862	1.12207E-09	-3.560242	-2.91785	-2.598796
Origin OTRI	-2.432005	0.1232006	-3.581258	-2.922785	-2.601541
Origin OTRI Differenced	-1.490221	0.5377891	-3.581258	-2.922785	-2.601541
Origin OTRI Differenced Twice	-2.230576	0.04860522	-3.581258	-2.922785	-2.601541
Destination OTRI	-2.264857	0.1863349	-3.556824	-2.916774	-2.598015
Destination OTRI Differenced	-3.018363	0.004165698	-3.581258	-2.922785	-2.601541
Expenditures Index	-1.754752	0.4031461	-3.571472	-2.922629	-2.598333
Expenditures Index Differenced	-6.904413	1.25794E-09	-3.560242	-2.91785	-2.598796
Shipments Index	-2.368405	0.1508532	-3.5577	-2.91677	-2.596222
Shipments Index Differenced	-8.9734	3.4109E-09	-3.560242	-2.91785	-2.598796
Inferred Rates	-1.490654	0.5544652	-3.571472	-2.922629	-2.598333
Inferred Rates Differenced	-2.430711	0.1332555	-3.571472	-2.922629	-2.598333
Inferred Rates Differenced Twice	-6.843619	5.33318E-09	-3.571472	-2.922629	-2.598333
Truck Load Index	-1.62945	0.4677738	-3.562879	-2.918973	-2.597393
Truck Load Index Differenced	-6.843619	5.33318E-09	-3.562879	-2.918973	-2.597393
PPI	-3.674683	0.04488372	-3.560242	-2.91785	-2.598796
# of Employees	-1.165643	0.6190928	-3.5577	-2.91677	-2.596222
# of Employees Differenced	-8.448154	1.22583E-09	-3.560242	-2.91785	-2.598796
Crop Index	-1.037003	0.7935296	-3.5577	-2.91677	-2.596222
Crop Index Differenced	-6.848052	5.20895E-09	-3.560242	-2.91785	-2.598796
Gas Price	-1.546254	0.1514582	-3.5577	-2.91677	-2.596222
Gas Price Differenced	-4.306851	0.00030099	-3.560242	-2.91785	-2.598796

Table 3.4.2 ADF p-value and critical value results of each variable for Lane ID NoCal Salinas to Bronx NY

Table 3.4.3 below shows the appropriate integration values used to difference NoCal Salinas -→ Bronx, NY



Variable	Integration value
Avg Cost of Truck	1
Origin OTRI	2
Destination OTRI	1
Expenditures Index	1
Shipments Index	1
Inferred Rates	2
Truck Load Index	1
PPI	0
# of Employees	1
Crop Index	1
Gas Price	1

Table 3.4.3 The differencing value (integration value) established by ADF for each variable

For example, Avg Cost of truck needs to be differenced once. Therefore, the integration value, d, is equal to one.

Section 3.5: Engle-Granger Causality Test

This section focuses on running hypotheses tests which will be used to understand how the macroeconomic factors we've been discussing have can be used to predict the cost of a truck. The following test is used for the purpose of running a vector autoregression model (VAR) which we will leave for future work. This test helps us to better understand the relationship between our target variable (cost of truck) and the macroeconomic factors.

The Engle-Granger causality test is an econometric method used to determine if one time series can be useful in forecasting another. Specifically, it assesses whether past values of one variable provide significant information about the future values of another variable.

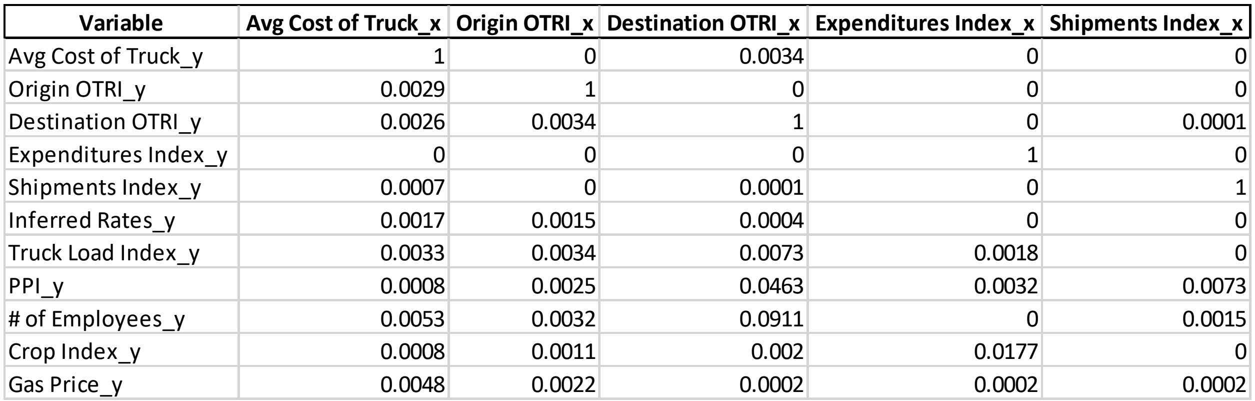
i. Granger-Causation: A variable X is said to Granger-cause a variable Y if the inclusion of past values of X in the forecasting model for Y significantly reduces the prediction error of Y.

ii. Failure to Granger-Cause: Conversely, X fails to Granger-cause Y if its inclusion does not significantly improve the forecasting accuracy of Y.

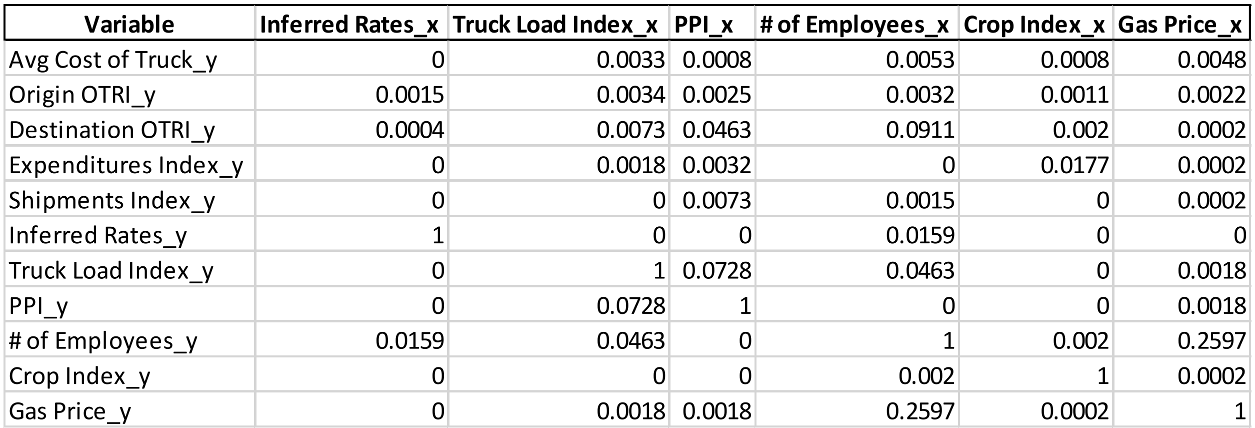
Formal Criteria:

1. In vector autoregressive models, X fails to Granger-cause Y if the lags of X are not statistically significant in the predictive equation for Y.
2. This indicates that past values of X do not provide significant information for predicting future values of Y.

We conducted the Granger causality test on our dataset to identify potential causal relationships. Each variable was systematically tested against all others in a pairwise manner. Specifically, each variable was treated as the independent (X) variable and tested against every other variable as the dependent (Y) variable, ensuring that every possible combination was examined.



Variable	Avg Cost of Truck_x	Origin OTRI_x	Destination OTRI_x	Expenditures Index_x	Shipments Index_x
Avg Cost of Truck_y	1	0	0.0034	0	0
Origin OTRI_y	0.0029	1	0	0	0
Destination OTRI_y	0.0026	0.0034	1	0	0.0001
Expenditures Index_y	0	0	0	1	0
Shipments Index_y	0.0007	0	0.0001	0	1
Inferred Rates_y	0.0017	0.0015	0.0004	0	0
Truck Load Index_y	0.0033	0.0034	0.0073	0.0018	0
PPI_y	0.0008	0.0025	0.0463	0.0032	0.0073
# of Employees_y	0.0053	0.0032	0.0911	0	0.0015
Crop Index_y	0.0008	0.0011	0.002	0.0177	0
Gas Price_y	0.0048	0.0022	0.0002	0.0002	0.0002



Variable	Inferred Rates_x	Truck Load Index_x	PPI_x	# of Employees_x	Crop Index_x	Gas Price_x
Avg Cost of Truck_y	0	0.0033	0.0008	0.0053	0.0008	0.0048
Origin OTRI_y	0.0015	0.0034	0.0025	0.0032	0.0011	0.0022
Destination OTRI_y	0.0004	0.0073	0.0463	0.0911	0.002	0.0002
Expenditures Index_y	0	0.0018	0.0032	0	0.0177	0.0002
Shipments Index_y	0	0	0.0073	0.0015	0	0.0002
Inferred Rates_y	1	0	0	0.0159	0	0
Truck Load Index_y	0	1	0.0728	0.0463	0	0.0018
PPI_y	0	0.0728	1	0	0	0.0018
# of Employees_y	0.0159	0.0463	0	1	0.002	0.2597
Crop Index_y	0	0	0	0.002	1	0.0002
Gas Price_y	0	0.0018	0.0018	0.2597	0.0002	1

Table 3.5.1 Engle Granger Causality

To interpret these results for the purpose of drawing insights on Adjusted LH, we observe that every predictor variable (with annotation y) has a p-value less than 0.05 at some lag when running the test against independent variable (with annotation x). The specific lag at which the minimum p-value occurs is presented in Table 5.4.1, particularly in the context of discussing future work and the VAR model.

These low p values indicate Granger causality, as the lagged values of all variables show strong evidence of being useful for predicting Adjusted LH (cost for the truck). We took the minimum p-value across 12 lags to determine these results. The values indicate whether the inclusion of one variable's lagged values significantly improves the forecasting accuracy of another variable. This helps in identifying potential causal relationships in the data and motivates the use of a Vector Autoregression (VAR) model for further analysis [15].

Section 3.6: Cointegration Test

In this section, we run a co-integration test to examine whether there is a long-term equilibrium relationship between truck costs and key macroeconomic trucking factors. The Engle cointegration test is particularly useful here, as it allows us to determine if these variables move together over time, despite short-term fluctuations. Understanding this relationship is crucial for the future implementation of a VAR model (left for future work), as it ensures that the model captures both the short-term dynamics and the underlying long-term trends in truck costs driven by macroeconomic conditions.

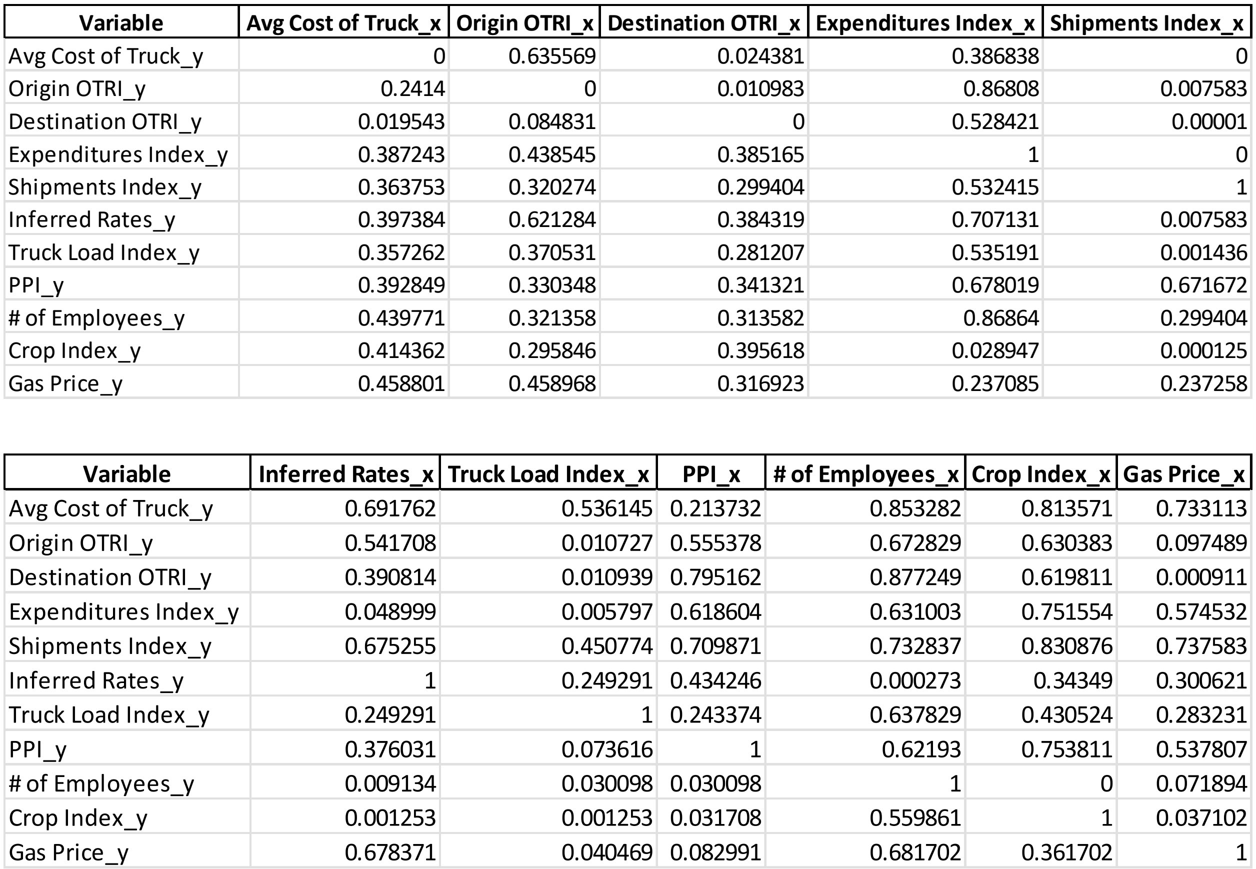
The Engle-Granger two-step method is commonly used for this test:

i. Estimate the long-run equilibrium relationship using ordinary least squares (OLS) regression. This involves regressing one time series on the other(s) and obtaining the residuals.

ii. Perform a unit root test (e.g., Augmented Dickey-Fuller test) on the residuals from the OLS regression. If the residuals are stationary, it indicates that the time series are cointegrated.

In essence, the cointegration test identifies whether the non-stationary time series move together over the long term, suggesting a stable relationship despite individual trends [15].

In Figure 3.6.1 we see that Adjusted LH (annotated with y) and Destination OTRI (annotated with) show strong evidence of cointegration, with a p-value of 0.024. This means that despite any short-term fluctuations, there exists a stable, long-term equilibrium relationship between these two variables. A linear combination of Adjusted LH and Destination OTRI is stationary, implying that the residuals from their linear regression do not have a unit root, confirming cointegration. We used the pair of Adjusted LH and Destination OTRI as it is the only significant p value between Adjusted LH (cost for the truck) which is the variable of interest for forecasting and a macroeconomic variable which is Destination OTRI.



Variable	Avg Cost of Truck_x	Origin OTRI_x	Destination OTRI_x	Expenditures Index_x	Shipments Index_x
Avg Cost of Truck_y	0	0.635569	0.024381	0.386838	0
Origin OTRI_y	0.2414	0	0.010983	0.86808	0.007583
Destination OTRI_y	0.019543	0.084831	0	0.528421	0.00001
Expenditures Index_y	0.387243	0.438545	0.385165	1	0
Shipments Index_y	0.363753	0.320274	0.299404	0.532415	1
Inferred Rates_y	0.397384	0.621284	0.384319	0.707131	0.007583
Truck Load Index_y	0.357262	0.370531	0.281207	0.535191	0.001436
PPI_y	0.392849	0.330348	0.341321	0.678019	0.671672
# of Employees_y	0.439771	0.321358	0.313582	0.86864	0.299404
Crop Index_y	0.414362	0.295846	0.395618	0.028947	0.000125
Gas Price_y	0.458801	0.458968	0.316923	0.237085	0.237258

Variable	Inferred Rates_x	Truck Load Index_x	PPI_x	# of Employees_x	Crop Index_x	Gas Price_x
Avg Cost of Truck_y	0.691762	0.536145	0.213732	0.853282	0.813571	0.733113
Origin OTRI_y	0.541708	0.010727	0.555378	0.672829	0.630383	0.097489
Destination OTRI_y	0.390814	0.010939	0.795162	0.877249	0.619811	0.000911
Expenditures Index_y	0.048999	0.005797	0.618604	0.631003	0.751554	0.574532
Shipments Index_y	0.675255	0.450774	0.709871	0.732837	0.830876	0.737583
Inferred Rates_y	1	0.249291	0.434246	0.000273	0.34349	0.300621
Truck Load Index_y	0.249291	1	0.243374	0.637829	0.430524	0.283231
PPI_y	0.376031	0.073616	1	0.62193	0.753811	0.537807
# of Employees_y	0.009134	0.030098	0.030098	1	0	0.071894
Crop Index_y	0.001253	0.001253	0.031708	0.559861	1	0.037102
Gas Price_y	0.678371	0.040469	0.082991	0.681702	0.361702	1

Table 3.6.1 Engle Granger Cointegration test

Section 3.7: Decomposition

To inform our forecast of truck cost for the transit route from Salinas, CA to the Bronx, NY, we decomposed the time series data into its trend and seasonal components. Trend refers to the direction of the time series. Seasonality refers to any recurring patterns in the time series. This decomposition can be done using either an additive or a multiplicative approach:

Additive approach: $Y(t) = Trend + Seasonality$, where $Y(t)$ is the time series.

Multiplicative approach: $Y(t) = Trend \times Seasonality$

After decomposing the time series into its trend and seasonal components, we overlay these components onto the original series to assess how well they fit the overall behavior of the data. Where the original series closely matches the sum or product of these components, it indicates that the trend and seasonal elements effectively capture the main patterns in the data. Conversely, noticeable discrepancies suggest the presence of noise or unexplained variation. This discrepancy highlights the influence of random fluctuations or other factors not accounted for by the decomposition, emphasizing the importance of considering noise in time series analysis [13].

Additive decomposition

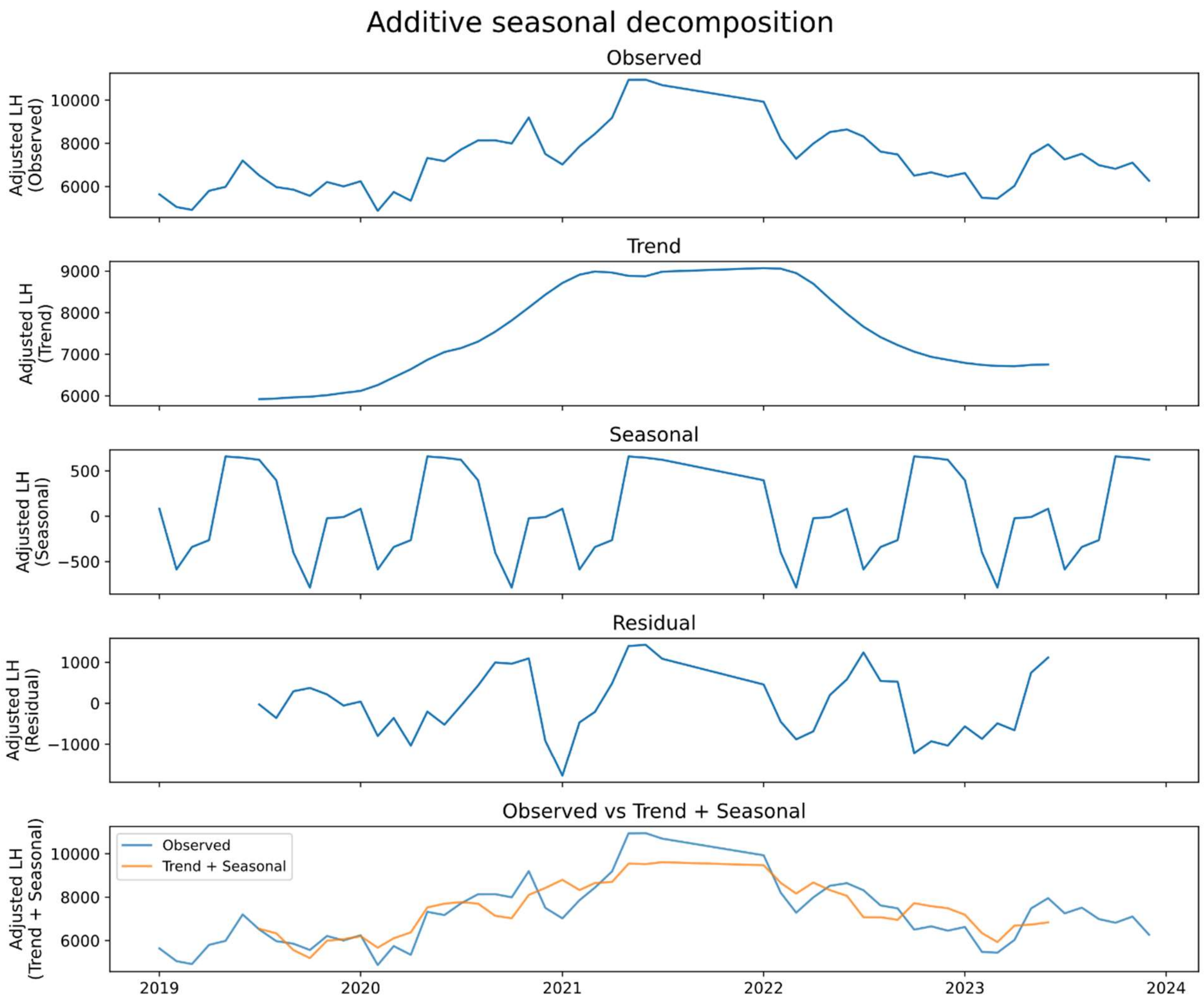


Figure 3.7.1: Original time series, trend, seasonality, residuals, and comparison of original time series vs. seasonality + trend.

Multiplicative decomposition

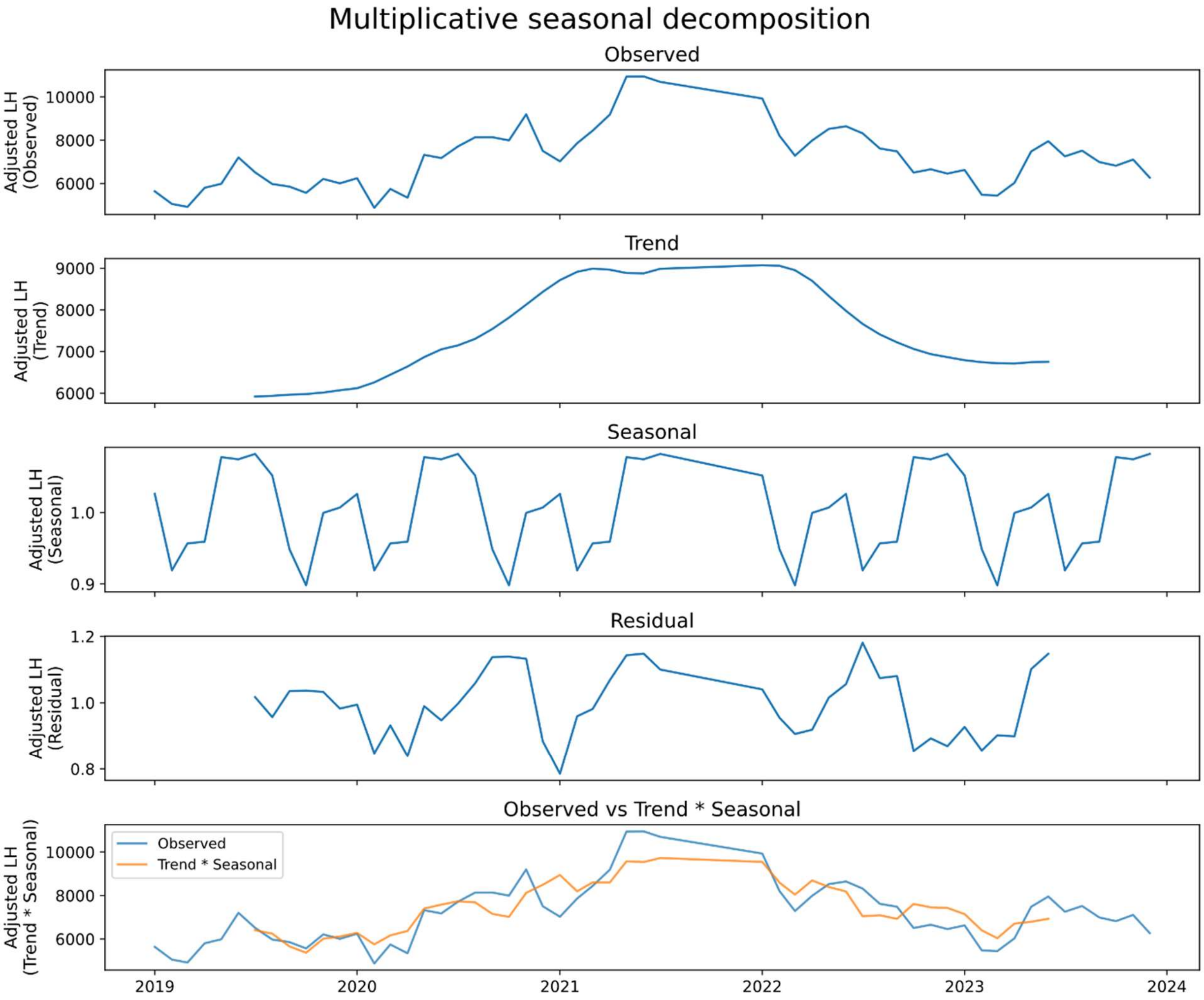


Figure 3.7.2: Original time series, trend, seasonality, residuals, and comparison of original time series vs. seasonality * trend.

We observe that the combined trend and seasonal components explain most of the patterns in the time series. The differences between the observed data (blue line) and the combined trend and seasonal components (orange line) can be attributed to other factors. These may include macroeconomic factors discussed in Sections 4.1 and 4.2.

The choice between choosing additive and multiplicative decomposition depends on how the data fluctuates around the trend. If fluctuations are constant, then additive decomposition might be more appropriate. Should the magnitude of the fluctuations vary greatly, then multiplicative decomposition would be more appropriate.

This analysis is left for further study as it would involve analyzing other time series lanes to confirm the behavior of the data for truck prices [13].

Chapter 4: Modeling Trucking Costs

Section 4.1: Introduction to Multiple Linear Regression

Multiple linear regression (MLR) is a modeling method used to predict one dependent variable using multiple independent variables. The equation below is the general set up of a multiple linear regression model.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

- Y is the dependent variable. In our case the cost for the truck (i.e. Adjusted LH)
- $X_1, X_2, \dots, X_n$ are the independent variables.
- $\beta_1, \beta_2, \dots, \beta_n$ are the coefficients
- β_0 is the intercept
- ϵ is the error term

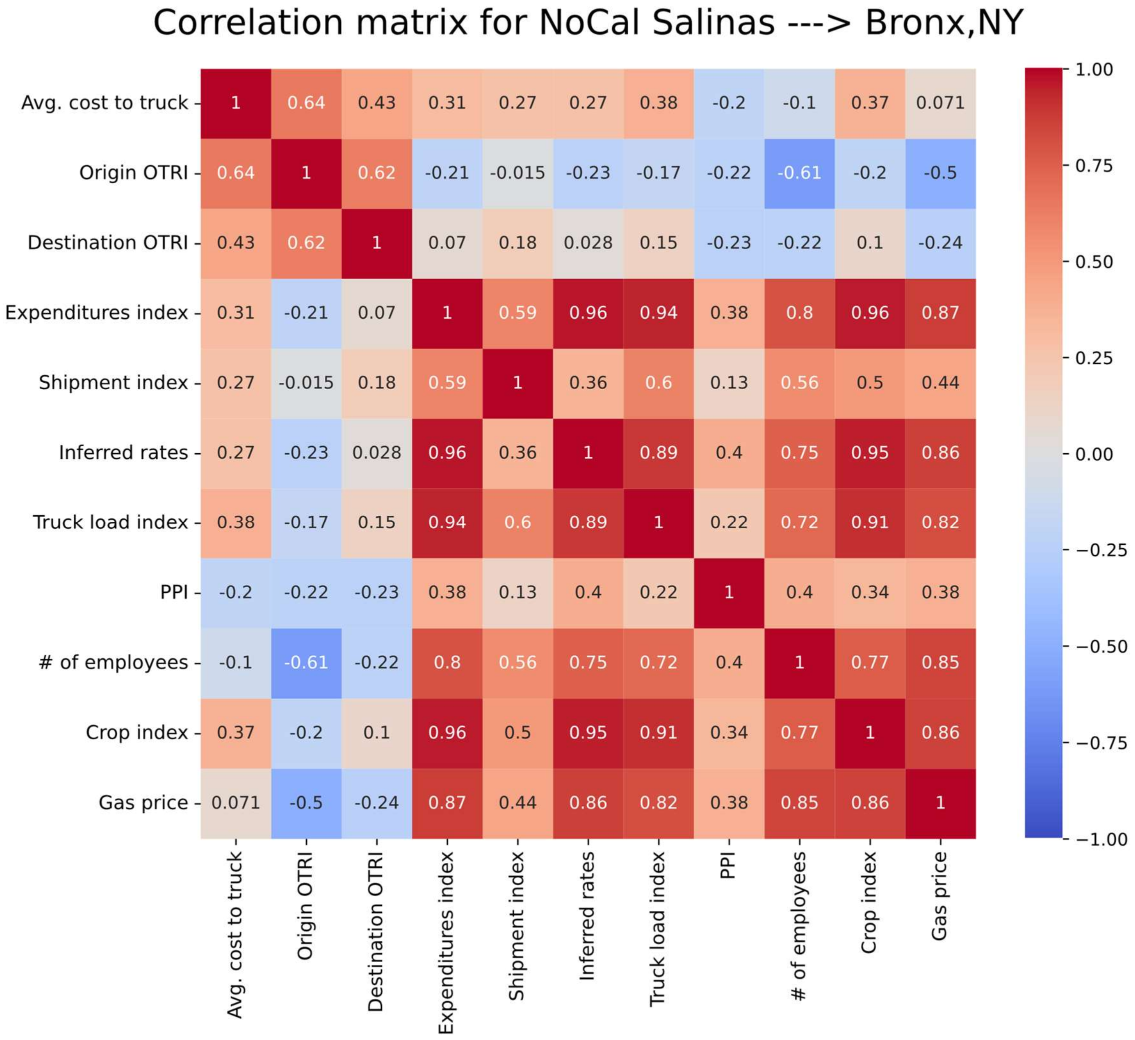
MLR minimizes the sum of squared residuals to estimate the coefficients $\beta_1, \beta_2, \dots, \beta_n$[16].

Substituting the selected independent variables into the MLR equation we arrive at the equation for our model.

$$\begin{aligned} AdjustedLH = {} & \beta_0 + \beta_1 OriginOTRI + \beta_2 DestinationOTRI + \beta_3 CASSExpenditures \\ & + \beta_4 CASSShipmentValue + \beta_5 CASSInferredRates + \beta_6 CASSTLLHIndex \\ & + \beta_7 ussappi + \beta_8 FREDNumofEmployees + \beta_9 FREDCropProduction + \beta_{10} Gas \\ & + \epsilon \end{aligned}$$

Section 4.2: Correlation Map

Before fitting the model, we analyze the correlations between the variables to identify useful relationships and possibly identify any multicollinearity in the model



	Avg. cost to truck	Origin OTRI	Destination OTRI	Expenditures index	Shipment index	Inferred rates	Truck load index	PPI	# of employees	Crop index	Gas price
Avg. cost to truck	1	0.64	0.43	0.31	0.27	0.27	0.38	-0.2	-0.1	0.37	0.071
Origin OTRI	0.64	1	0.62	-0.21	-0.015	-0.23	-0.17	-0.22	-0.61	-0.2	-0.5
Destination OTRI	0.43	0.62	1	0.07	0.18	0.028	0.15	-0.23	-0.22	0.1	-0.24
Expenditures index	0.31	-0.21	0.07	1	0.59	0.96	0.94	0.38	0.8	0.96	0.87
Shipment index	0.27	-0.015	0.18	0.59	1	0.36	0.6	0.13	0.56	0.5	0.44
Inferred rates	0.27	-0.23	0.028	0.96	0.36	1	0.89	0.4	0.75	0.95	0.86
Truck load index	0.38	-0.17	0.15	0.94	0.6	0.89	1	0.22	0.72	0.91	0.82
PPI	-0.2	-0.22	-0.23	0.38	0.13	0.4	0.22	1	0.4	0.34	0.38
# of employees	-0.1	-0.61	-0.22	0.8	0.56	0.75	0.72	0.4	1	0.77	0.85
Crop index	0.37	-0.2	0.1	0.96	0.5	0.95	0.91	0.34	0.77	1	0.86
Gas price	0.071	-0.5	-0.24	0.87	0.44	0.86	0.82	0.38	0.85	0.86	1

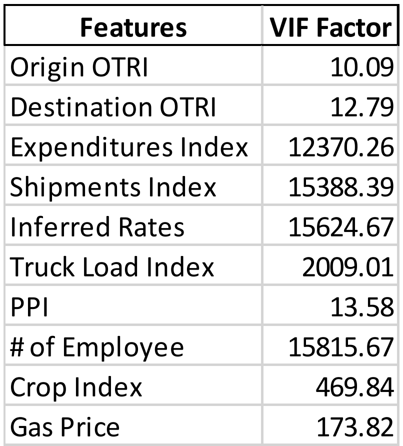
Figure 4.2.1 Pairwise Pearson Correlation Coefficients

The map above shows significant multicollinearities particularly with the CASS datasets. We use the Variance Inflation Factor (VIF) to confirm this visual representation [17].

Interpretation of VIF:

- **VIF = 1**: No multicollinearity.
- **1 < VIF < 5**: Moderate multicollinearity.
- **VIF > 5**: High multicollinearity. Consider corrective measures.
- **VIF > 10**: Very high multicollinearity. Indicates a significant issue that needs to be addressed [17].

The results, shown in Table 4.2.2, indicate significant multicollinearity among the variables in our dataset. This necessitates continued feature engineering to ensure that the model's coefficients are interpretable. We will proceed with running the model in the next section, leaving feature selection for future work in subsequent projects.



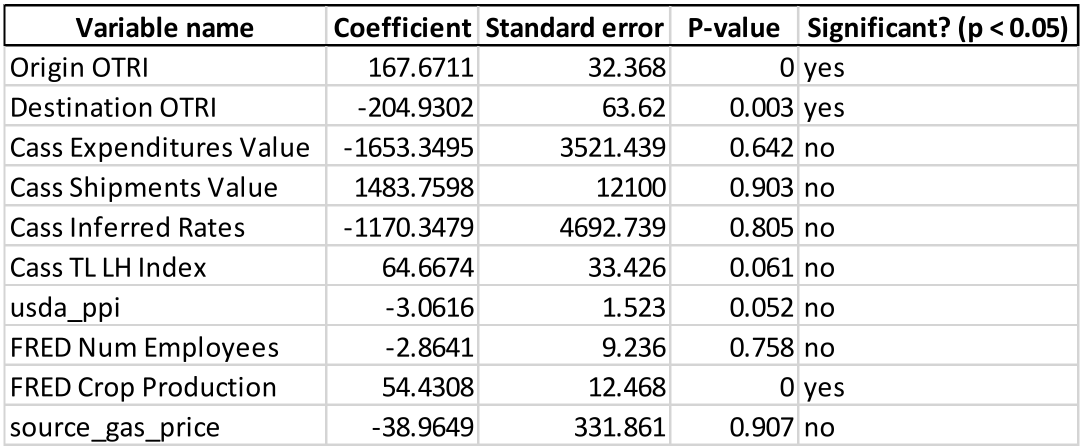
Features	VIF Factor
Origin OTRI	10.09
Destination OTRI	12.79
Expenditures Index	12370.26
Shipments Index	15388.39
Inferred Rates	15624.67
Truck Load Index	2009.01
PPI	13.58
# of Employee	15815.67
Crop Index	469.84
Gas Price	173.82

Table 4.2.2 VIF scores for independent variables

Section 4.3: Results from MLR model

In this section we run the multiple linear regression model to show how well the macroeconomic factors perform when predicting the cost for a truck.

We ran the model by randomly selecting our test data, regardless of date, in an 80% test and 20% train split. In Table 4.3.1, we list the p-values for each variable used to predict the cost of a truck. The variables Origin OTRI, Destination OTRI, and Crop Index showed significant p-values (below 0.05).



Variable name	Coefficient	Standard error	P-value	Significant? (p < 0.05)
Origin OTRI	167.6711	32.368	0	yes
Destination OTRI	-204.9302	63.62	0.003	yes
Cass Expenditures Value	-1653.3495	3521.439	0.642	no
Cass Shipments Value	1483.7598	12100	0.903	no
Cass Inferred Rates	-1170.3479	4692.739	0.805	no
Cass TL LH Index	64.6674	33.426	0.061	no
usda_ppi	-3.0616	1.523	0.052	no
FRED Num Employees	-2.8641	9.236	0.758	no
FRED Crop Production	54.4308	12.468	0	yes
source_gas_price	-38.9649	331.861	0.907	no

Table 4.3.1 Coefficients of MLR model with significance values

Figure 4.3.2 shows the results as a scatter plot, where the predictions (red dots) are closely clustered with the train (black dots) and test set (green dots). This indicates that the predicted values cover similar regions as the actual data, suggesting a strong fit.

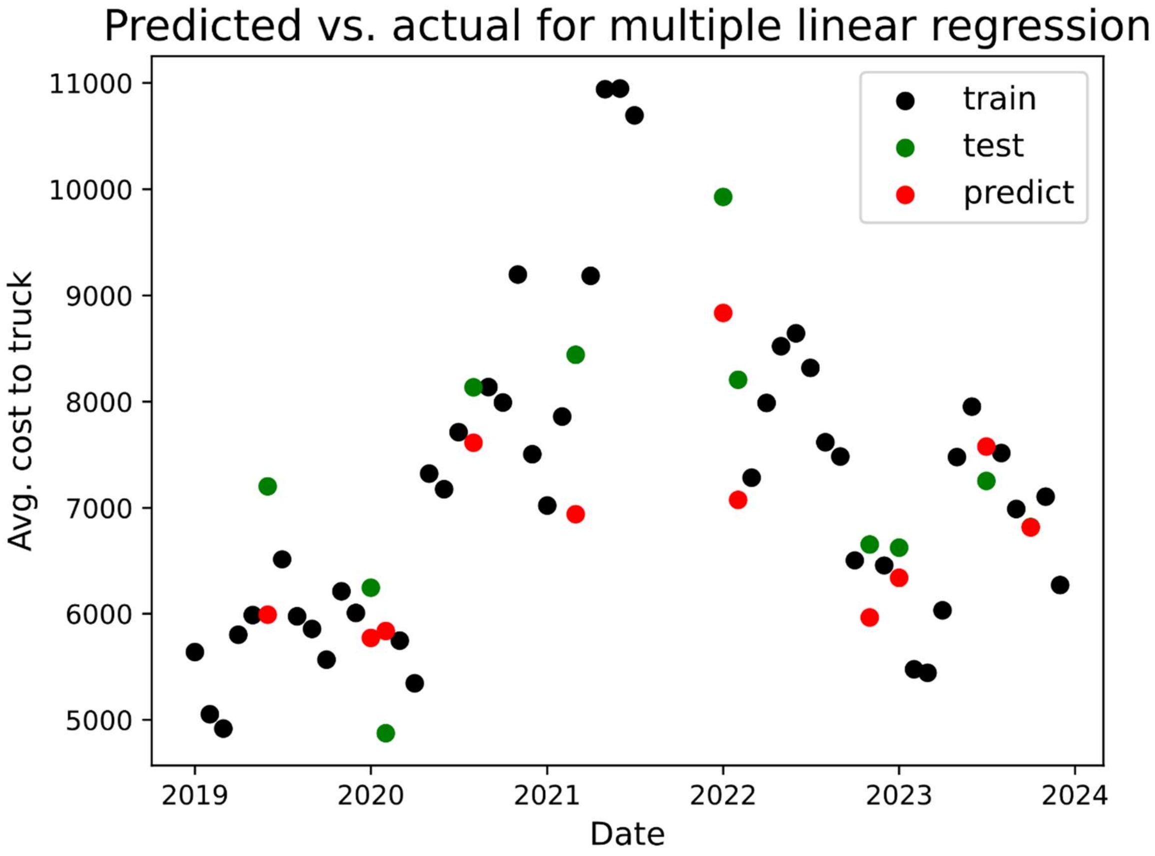


Figure 4.3.2 Results of MLR on randomly selected of shipments

The model produced a root mean squared error (RMSE) of $700. The results are promising enough to warrant continued feature engineering in the future specifically with Origin OTRI, Destination OTRI, and Crop Index due to their significant relationship with trucking cost. However, multiple linear regression (MLR) is not ideal for this time series data. As shown in Figure 4.3.1 we can visually assess that the residual errors increase with respect to time. This is particularly evident between 2021 and 2022. By visual inspection we suspect that the distribution of residuals is not normal.

This exercise was useful for understanding the predictive power of the chosen variables and capturing some noise in the freight market. Future iterations might benefit from forward selection of variables and researching more predictive variables with less multicollinearity.

Section 4.4: Introduction to ARIMA

In this section, we aim to predict truck costs using a univariate model. As observed in Figure 4.3.1, the residual errors show increasing variance over time, particularly between 2021 and 2022. This indicates that the gap between predicted and actual costs widened during these years, which were disruptive for the trucking industry due to the fallout from COVID-19. To address this, we shift our focus to the ARIMA model, which makes weaker assumptions about the normality of residual errors.

A forecasting model that combines autoregression, differencing, and moving average is known as an ARIMA model. This model is particularly effective for non-stationary time series data, as it allows for adjustments to be made to achieve stationarity.

- **AR (Autoregression)**: This component uses the dependency between an observation and a number of lagged observations (past values). It captures the relationship between the current value and its past values.

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

ϕ_0 is the mean of the time series

$\phi_1, \phi_2, \ldots, \phi_p$ are the parameters of the auto regressive terms

$y_{t-1}, y_{t-2}, .., y_{t-p}$ are the past lagged values of the time series

ϵ_t is the white noise (error term)

- **I (Integration/Differencing)**: This involves differencing the series to remove trends and make it stationary. By taking the difference of adjacent values, this component helps to stabilize the mean of the time series.
- **MA (Moving Average)**: This component represents the dependency between an observation and a residual error from a moving average model applied to lagged observations. It captures the relationship between an observation and past forecast errors.

$$y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

μ is the mean of the series

ϵ_t is the error term at time t

$$\theta_1, \theta_{t-2}, \ldots, \theta_{t-q} \epsilon_{t-1}, \epsilon_{t-2}, \ldots, \epsilon_{t-q}$$

ARIMA combined equation

$$y_t = \phi_0 + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$$

Together, these components make the ARIMA model.

The ARIMA model will take 3 parameters AR order p, differenced order d, and MA order q. Notation for the specific model is denoted ARIMA (p, d, q) [13].

Section 4.5: ACF and PACF

We used Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots to determine the appropriate $p$and q values for our $ARIMA(p, d, q)$model. The differencing component, d, was previously established in Section 5.2 using the Augmented Dickey-Fuller (ADF) test. We will be using d = 1.

Using Figure 4.5.1, we determined the q component by analyzing the ACF plot and the p component using the PACF plot. Both plots show a significant spike at lag 0, followed by oscillations between negative and positive values. We selected the number of lags for our p value based on the PACF plot, taking the point where the correlation drops to zero. Therefore, we chose p=9 for our AR order. The ACF plot shows that at lag 1 the correlation dramatically drops and spikes at lag 3. Therefore, we chose an MA component of 3 for our model.

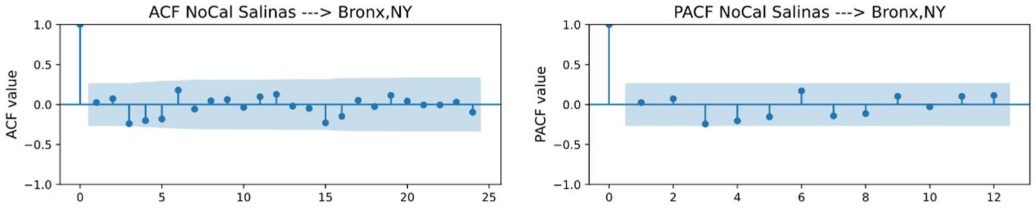


Figure 4.5.1 ACF and PACF plots for the transit route from Salinas CA to the Bronx NY

Section 4.6: Results from ARIMA (9,1,3)

For our holdout set, we used the last six months of data (March 2023 to December 2023). The data before this period was used to train our model.

We created six different models. The first model included all the data for each year. For the subsequent models, we progressively removed years from the dataset to see how the removal of certain years affected the model's performance. We do this because recent years have been disruptive in the supply chain. For example, in the aftermath of Covid–19 shutdowns and the closing of the economy many carriers left the market and shippers were forced to pause shipments. In 2021 there was a boom in trucking that was caused partially by the backup of product again caused by Covid-19. Checking which periods, when excluded, improve the results will help to validate, from a modeling perspective, these disruptions.

As a performance metric, we used Root Mean Square Error (RMSE). We provided a graph for each model showing the predicted versus actual cost for truck values. Here the red shaded region represents the confidence interval of 95%.

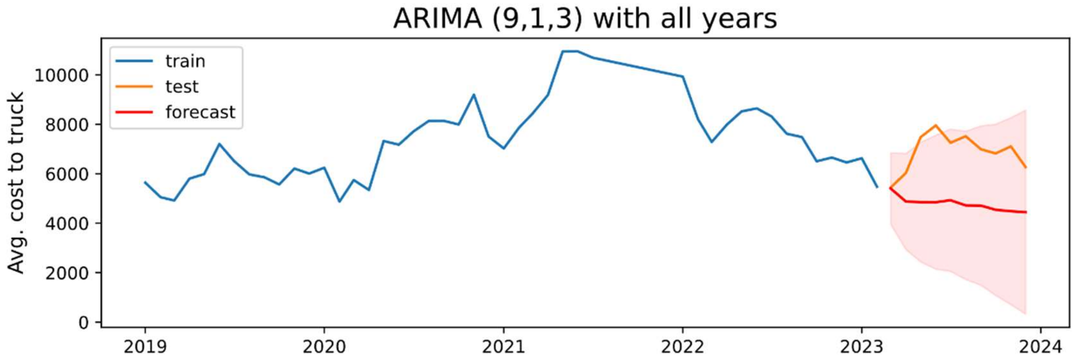


Figure 4.6.1 ARIMA (9,1,3) model with all years included in training. RMSE $2,273

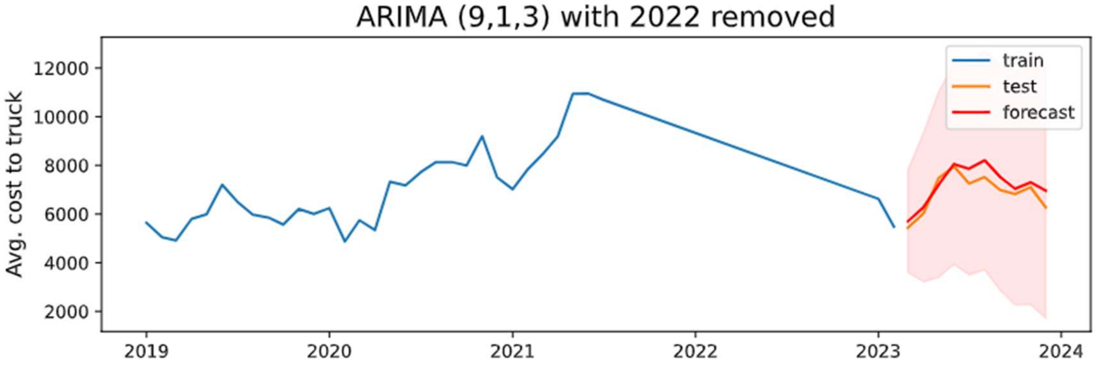


Figure 4.6.2 ARIMA (9,1,3) model with 2022 removed from training set. RMSE $436

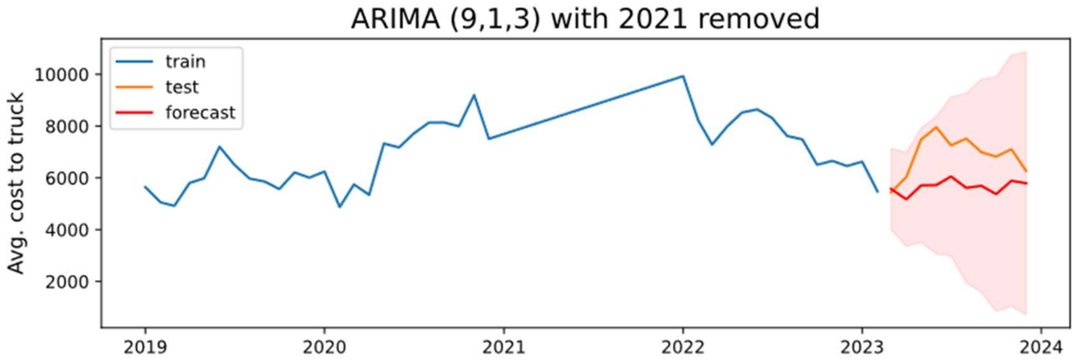


Figure 4.6.3 ARIMA (9,1,3) model with 2021 removed from training set. RMSE $1,394

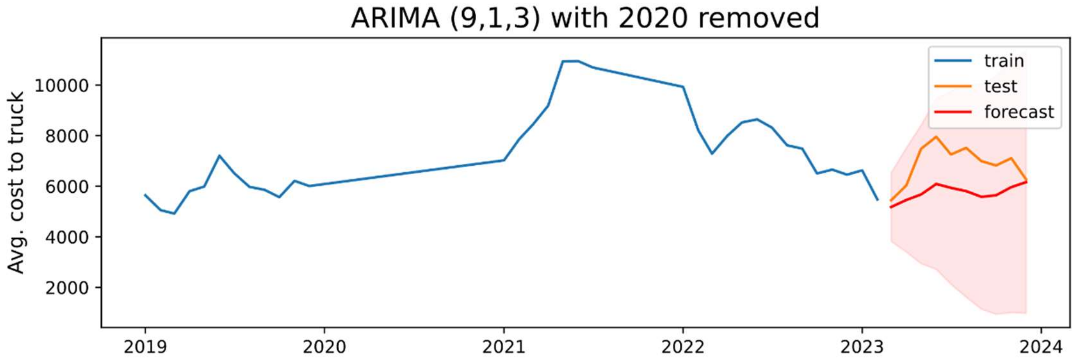


Figure 4.6.4 ARIMA (9,1,3) model with 2020 removed from training set. RMSE $1,285

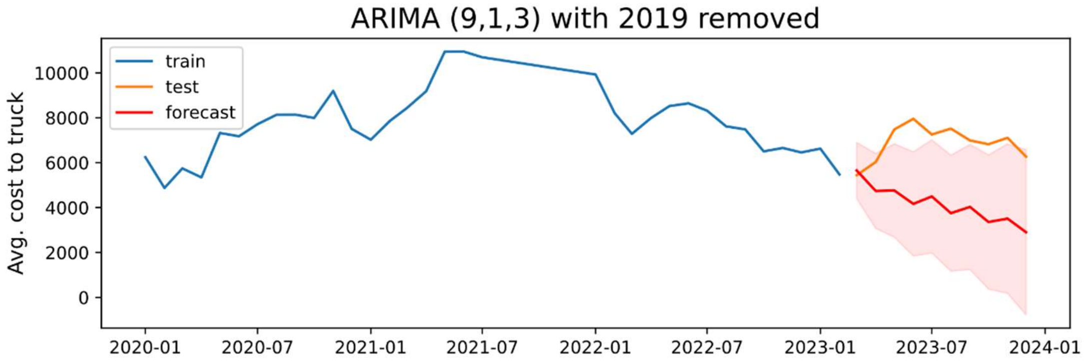
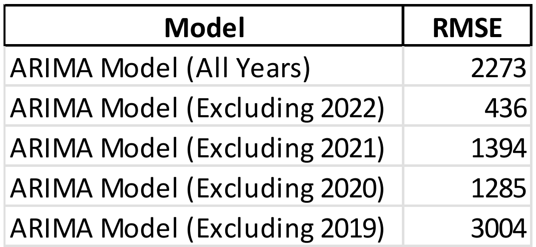


Figure 4.6.5 ARIMA (9,1,3) model with 2019 removed from training set . RMSE $3,004

The standard deviation of the full training dataset is $1558.75, which provides a reference point for evaluating the model's performance. With the lowest RMSE of $426, the model's error is significantly smaller than the overall variation in the data. This suggests that the model is able to capture much of the underlying pattern in the data, as the RMSE is well below the standard deviation, indicating that the model's predictions are reasonably accurate relative to the natural variability of the dataset.

Table 4.6.6 shows the resulting RMSE for each of the models we ran.



Model	RMSE
ARIMA Model (All Years)	2273
ARIMA Model (Excluding 2022)	436
ARIMA Model (Excluding 2021)	1394
ARIMA Model (Excluding 2020)	1285
ARIMA Model (Excluding 2019)	3004

Table 4.6.6 RMSE of each ARIMA model iteration.

Section 4.7: Introduction to SARIMA

Seasonal Autoregressive Integrated Moving Average (SARIMA) modeling is an extension of the ARIMA model with a seasonal component. SARIMA incorporates both non-seasonal and seasonal aspects of the data, providing a framework for time series analysis and forecasting. [13]

The SARIMA model is denoted as $SARIMA(p,d,q)(P,D,Q)_s$, where:

- $p$is the order of the non-seasonal autoregressive (AR) terms,
- $d$is the order of non-seasonal differencing,
- q is the order of the non-seasonal moving average (MA) terms,
- $P$is the order of the seasonal autoregressive (SAR) terms,
- $D$is the order of seasonal differencing,
- $Q$is the order of the seasonal moving average (SMA) terms,

- $s$is the length of the seasonal cycle.

The general form of the SARIMA model is expressed as:

$$\Phi_P(B^s)\phi_p(B)(1-B)^d(1-B^s)^D y_t = \Theta_Q(B^s)\theta_q(B)\epsilon_t$$

- y_t is the time series,
- B is the backshift operator, $By_t = y_{t-1}$
- $\Phi_P(B^s)$ and $\Theta_Q(B^s)$ are the seasonal AR and MA polynomials, respectively
- $\phi_p(B)$ and $\theta_q(B)$ are the non-seasonal AR and MA polynomials, respectively
- ϵ_t is the white noise error term

Section 4.8: ACF and PACF

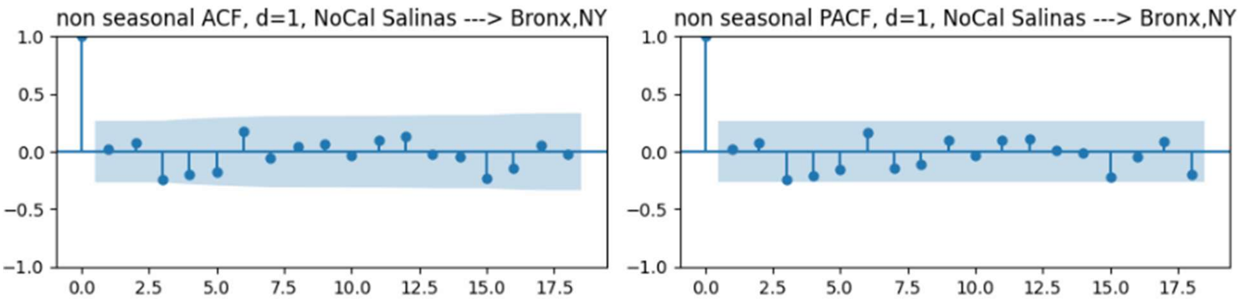


Figure 4.8.1 ACF and PACF plots of the transit route from NoCal Salinas to Bronx NY with differencing of 1

The ACF and PACF plots suggest a possible p and q value of 3, with no apparent seasonality. However, from domain knowledge, we know that produce is typically seasonal on a yearly basis. Recent supply chain disruptions may have obscured the seasonality in these plots. For modeling purposes, we will use s=12 for the seasonal component.

Section 4.9: Results from SARIMA (9, 1, 3) (3, 1, 3, 12)

We will use the same testing set for SARIMA of March 2023 – June 2023.

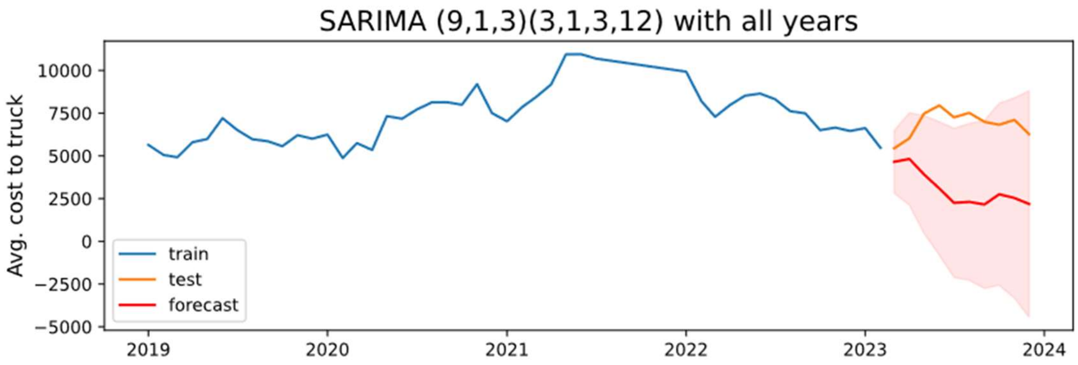


Figure 4.9.1 SARIMA (9,1,3) (3,1,3,12) model all years included in training. RMSE $4,100

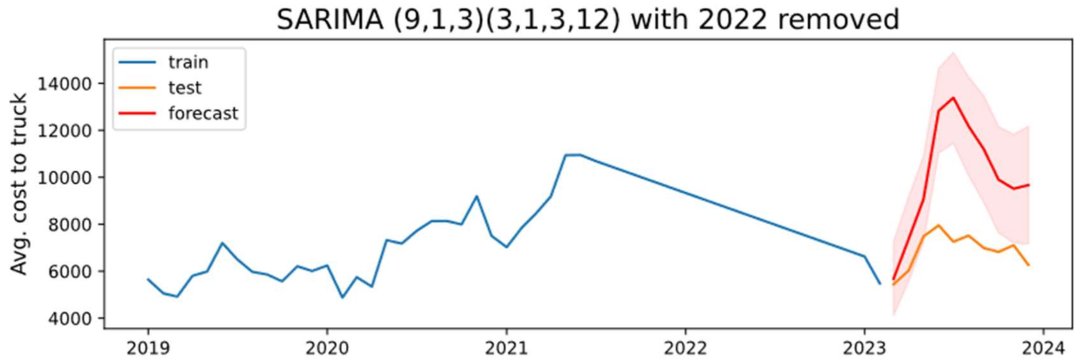


Figure 4.9.2 SARIMA (9,1,3) (3,1,3,12) model with 2022 removed from the training set. RMSE $3,633

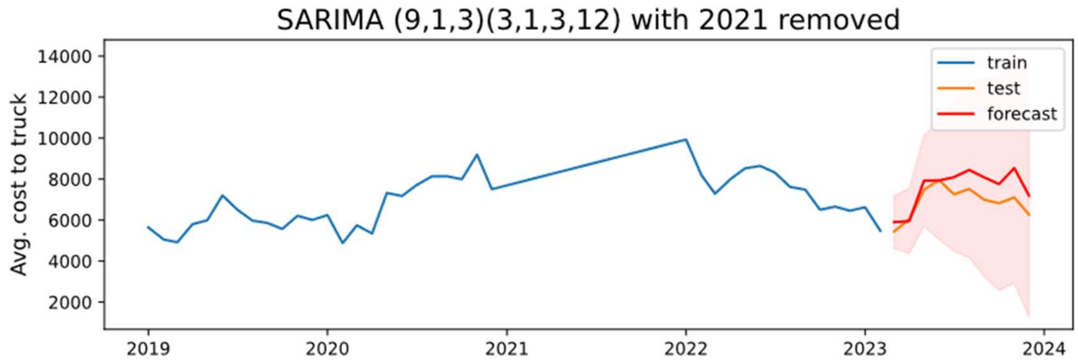


Figure 4.9.3 SARIMA (9,1,3) (3,1,3,12) model with 2021 removed from the training set. RMSE $834

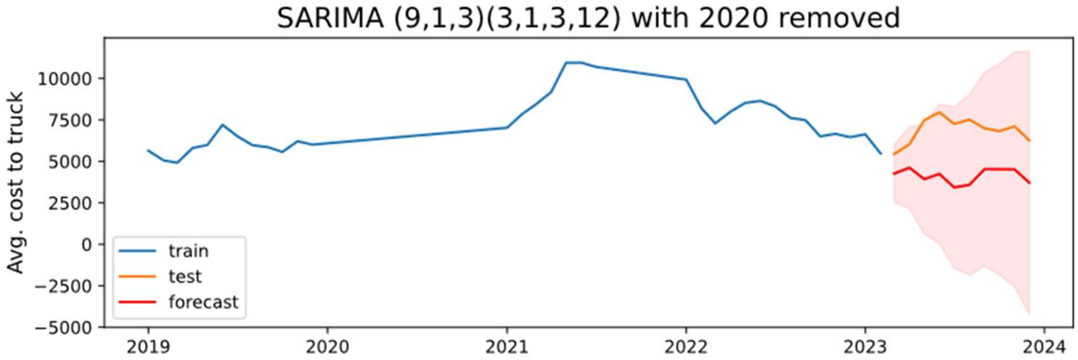


Figure 4.9.4 SARIMA (9,1,3) (3,1,3,12) model with 2020 removed from the training set. RMSE $2,906

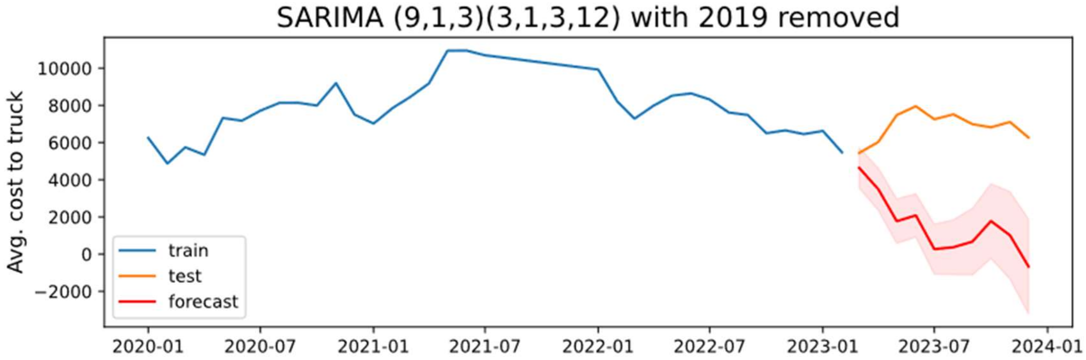
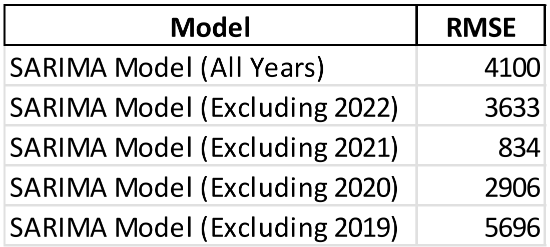


Figure 4.9.5 SARIMA (9,1,3) (3,1,3,12) model with 2019 removed from the training set. RMSE $5,696

Section 4.10: Closing thoughts on SARIMA (9,1,3) (3, 1, 3)

Figure 4.10.1 shows the resulting RMSE for each of the models we ran.

As mentioned earlier, the standard deviation of the full training dataset is $1558.75. With the lowest RMSE of $834 when the year 2021 was removed, this indicates that the model performs well, with the error being significantly lower than the dataset's variability, similar to previous results.



Model	RMSE
SARIMA Model (All Years)	4100
SARIMA Model (Excluding 2022)	3633
SARIMA Model (Excluding 2021)	834
SARIMA Model (Excluding 2020)	2906
SARIMA Model (Excluding 2019)	5696

Table 4.10.1 RMSE of each SARIMA model iteration.

Chapter 5: Closing Thoughts

Section 5.1: Closing thoughts on modeling

ARIMA performed better when 2022 was removed, and SARIMA performed better when 2021 was removed. During 2022 prices were coming down from the year prior that saw a disruption in the market due to labor shortages and high freight volume that came in the aftermath of Covid-19 pandemic. However, it remains unclear as to why the other disruptive years did not lead to a negative effect on the model's performance.

For SARIMA we can conjecture as to why the exclusion of 2021 led to better results. In Figure 4.2.1 we noticed that for most of 2021 costs for trucks remained constant and at very high prices. Seasonality during this time wasn't as evident as it was in other years. By excluding 2021, we are giving the model more data with seasonal patterns to train on. When removing any other year, the model results deviated significantly from the actuals as shown in Table 10.4.1.

Section 5.2: Data Limitations

The data cleaning required us to attempt to extrapolate limited information particularly on multi stop and single stop shipments. These are approximations based on domain knowledge. The 2021 – 2023 data which has fields to indicate which shipments are single and which are multiple stops is cleaner but alone does not suffice for our modeling purposes. As we saw in sections 10 and 11 both models needed to exclude one of the years, 2021 and 2022 to arrive at the best model with the possible data. However, approximations are not without margin of error. The 2019- 2021 data gave no indication of which shipments were single stop or multi stop. We extrapolated on mileage bands and approximated a lane deduction based on the average of the single and multiple stop - population for the 2021-2023 dataset. This data limitation encourages collection of more data for future years that have a broken-out system to display charges.

Another limitation is that there aren’t enough “normal” years in trucking to model within the dataset. The supply chain market has seen significant peaks and valleys in the years that followed the COVID 19 pandemic. This further encourages the collection of data for this project.

Section 5.3: Multivariate factors

There needs to be more feature engineering since the selected variables exhibit significant multicollinearity. Multicollinearity can distort the results of our models and reduce their predictive power. To address this, we will apply techniques such as variable transformation, interaction terms, and principal component analysis (PCA) to reduce multicollinearity and improve the robustness of our models.

Additionally, we must perform a thorough analysis of heteroscedasticity, which refers to the presence of non-constant variance in the error terms of a regression model. Heteroscedasticity can lead to inefficient estimations and affect the validity of hypothesis tests. By detecting and addressing heteroscedasticity, we can ensure that our model's assumptions are met and improve its accuracy and reliability.

These steps are crucial to refining our models and enhancing their predictive capabilities, ultimately leading to more accurate forecasts and better decision-making for the company.

Section 5.4: Modeling Limitations

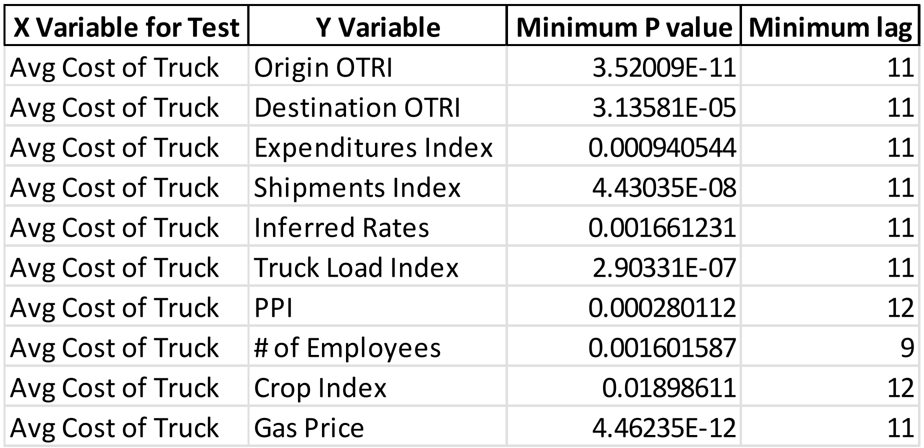
The modeling limitations go hand in hand with our data limitation and the additional multivariate work to continue to understand the influences on truck price. One of the models of interest during this project was the Vector autoregression. As we saw from Chapter 6 on Engle Granger causality and cointegration there are lags in which a significant forecasting power exists in our chosen group of variables. However, our number of parameters together with their lags greatly exceeds the amount of observations which are aggregated as a monthly average. The number of lags required to satisfy the Engle-Granger causality test restricts our ability to use certain variables.

For instance, Table 5.4.1 shows the minimum p-value from our Engle-Granger test. Ideally, the lag for a variable like Origin OTRI would be 11, but we only have 53 observations.

This issue arises because VAR models require enough observations to estimate the parameters accurately. The number of parameters in a VAR model can be calculated using the formula $K + pK^2$, where K is the number of variables and p is the number of lags. If the number of lags (p) and the number of variables (K) are too high relative to the number of observations, the model can become over-parameterized, leading to overfitting and unreliable estimates [13].

This issue can be resolved by using VAR for weekly, rather than monthly, predictions. By increasing the frequency of the data, we will have more observations, allowing us to include the necessary number of lags without exceeding the number of observations. This approach will ensure that the number of parameters in the VAR model does not exceed the number of observations, thereby improving the model's validity and reliability.

For future work, when a VAR will be implemented, it could be used as a weekly forecast to expand the number of observations and allows for a linear combination of the variables and lags listed in Table 5.4.1



X Variable for Test	Y Variable	Minimum P value	Minimum lag
Avg Cost of Truck	Origin OTRI	3.52009E-11	11
Avg Cost of Truck	Destination OTRI	3.13581E-05	11
Avg Cost of Truck	Expenditures Index	0.000940544	11
Avg Cost of Truck	Shipments Index	4.43035E-08	11
Avg Cost of Truck	Inferred Rates	0.001661231	11
Avg Cost of Truck	Truck Load Index	2.90331E-07	11
Avg Cost of Truck	PPI	0.000280112	12
Avg Cost of Truck	# of Employees	0.001601587	9
Avg Cost of Truck	Crop Index	0.01898611	12
Avg Cost of Truck	Gas Price	4.46235E-12	11

Table 5.4.1 Shows the minimum p-values from Engle Granger Causality

Section 5.5: Conclusion

In closing we accomplished three main objectives with this first iteration of machine learning on this data.

1. Formalized company logic for normalizing missing or misleading data: We achieved this by developing a strategy to identify multi-stop shipments, which can greatly skew our results, and apply an approximate lane deduction to each such shipment.

2. Collected and analyzed multivariate variables that impact the supply chain: Predicting carrier costs with a reliable degree of accuracy is a multivariate issue. Trucking can be impacted by changes in interest rates, gas prices, truck asset prices, insurance prices, and even weather conditions. We analyzed some of these variables to check their viability in model usage. Some promising results came from Engle Granger tests and Multiple Linear Regression which suggest that these variables can be used to predict some of the fluctuation in the trucking market.

3. We modeled the data with reasonably good results. Both the ARIMA and SARIMA model forecasts aligned well with the test set which forecasted below $1000 RMSE and captured the

directional trend in the two best performing models. The models gave us an understanding of which datasets to exclude for future models.

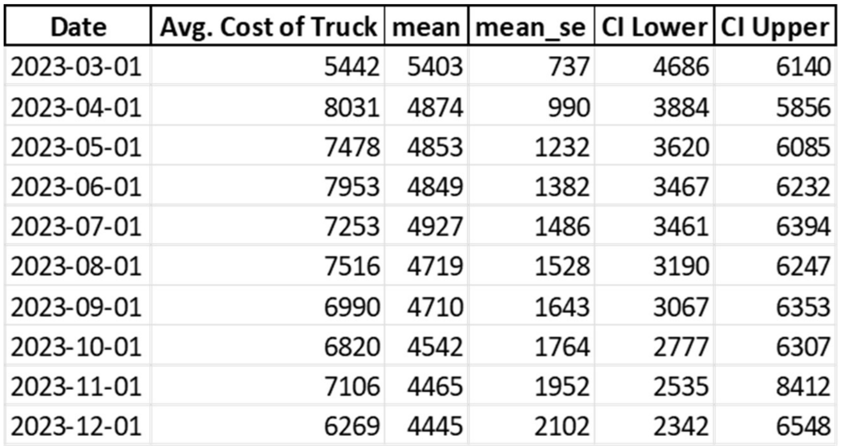
In future iterations of this project, ARIMA and SARIMA will be evaluated in different parts of the time series. More data will be added to include 2024 and future years.

Appendix

Below are tables associated with the ARIMA and SARIMA models that include lower and upper confidence intervals and average forecasts for each month.

ARIMA

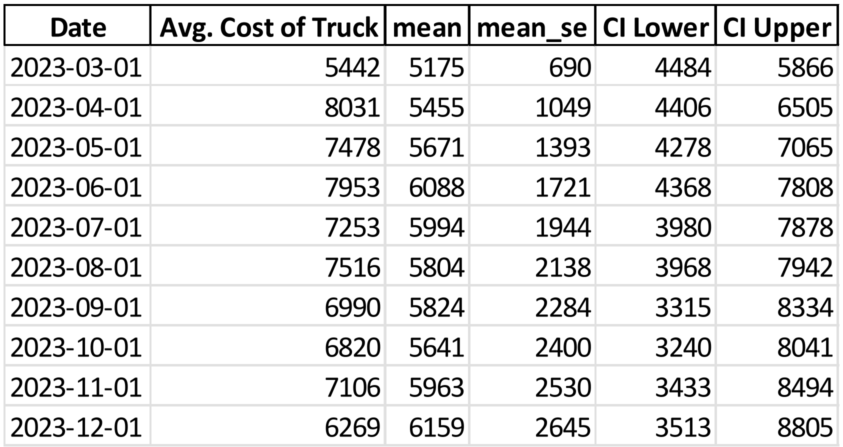
All Years included



Date	Avg. Cost of Truck	mean	mean_se	CI Lower	CI Upper
2023-03-01	5442	5403	737	4686	6140
2023-04-01	8031	4874	990	3884	5856
2023-05-01	7478	4853	1232	3620	6085
2023-06-01	7953	4849	1382	3467	6232
2023-07-01	7253	4927	1486	3461	6394
2023-08-01	7516	4719	1528	3190	6247
2023-09-01	6990	4710	1643	3067	6353
2023-10-01	6820	4542	1764	2777	6307
2023-11-01	7106	4465	1952	2535	8412
2023-12-01	6269	4445	2102	2342	6548

Appendix Table 1: ARIMA model results with all years included

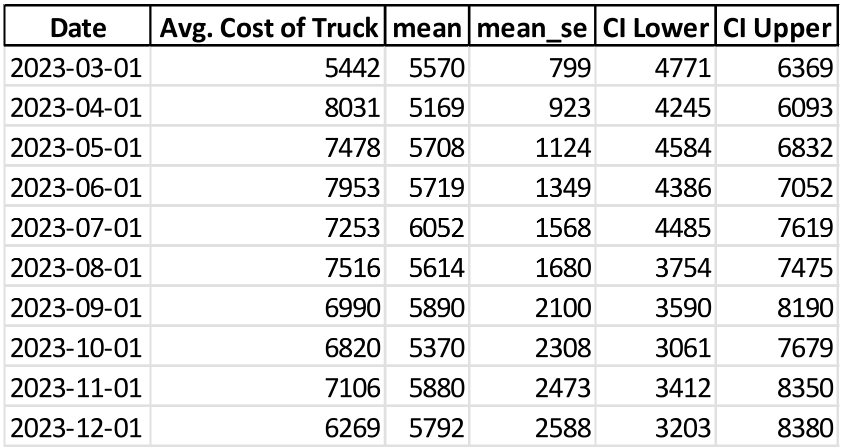
2022



Date	Avg. Cost of Truck	mean	mean_se	CI Lower	CI Upper
2023-03-01	5442	5175	690	4484	5866
2023-04-01	8031	5455	1049	4406	6505
2023-05-01	7478	5671	1393	4278	7065
2023-06-01	7953	6088	1721	4368	7808
2023-07-01	7253	5994	1944	3980	7878
2023-08-01	7516	5804	2138	3968	7942
2023-09-01	6990	5824	2284	3315	8334
2023-10-01	6820	5641	2400	3240	8041
2023-11-01	7106	5963	2530	3433	8494
2023-12-01	6269	6159	2645	3513	8805

Appendix Table 2: ARIMA model results with 2022 excluded from the training set

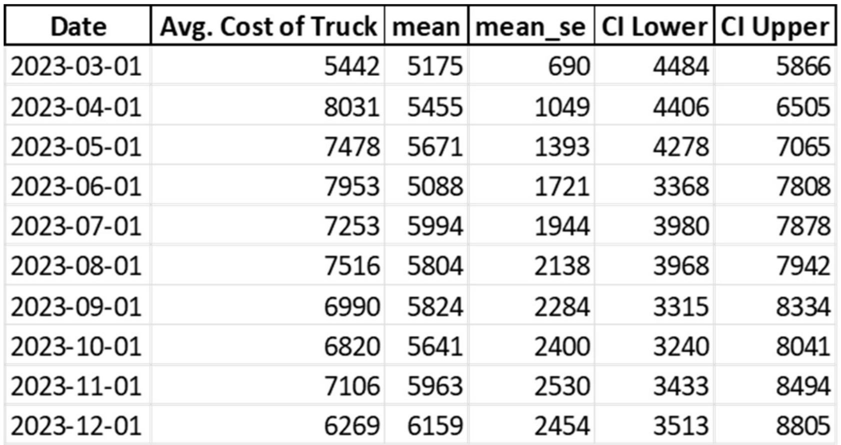
2021



Date	Avg. Cost of Truck	mean	mean_se	CI Lower	CI Upper
2023-03-01	5442	5570	799	4771	6369
2023-04-01	8031	5169	923	4245	6093
2023-05-01	7478	5708	1124	4584	6832
2023-06-01	7953	5719	1349	4386	7052
2023-07-01	7253	6052	1568	4485	7619
2023-08-01	7516	5614	1680	3754	7475
2023-09-01	6990	5890	2100	3590	8190
2023-10-01	6820	5370	2308	3061	7679
2023-11-01	7106	5880	2473	3412	8350
2023-12-01	6269	5792	2588	3203	8380

Appendix Table 3: ARIMA model results with 2021 excluded from the training set

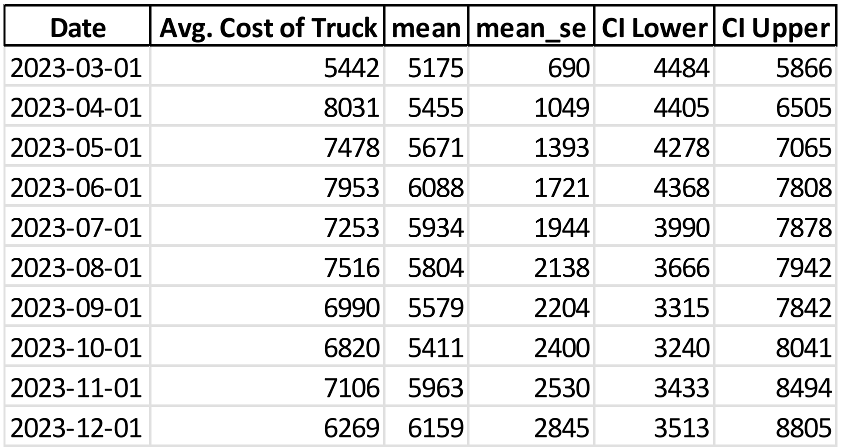
2020



Date	Avg. Cost of Truck	mean	mean_se	CI Lower	CI Upper
2023-03-01	5442	5175	690	4484	5866
2023-04-01	8031	5455	1049	4406	6505
2023-05-01	7478	5671	1393	4278	7065
2023-06-01	7953	5088	1721	3368	7808
2023-07-01	7253	5994	1944	3980	7878
2023-08-01	7516	5804	2138	3968	7942
2023-09-01	6990	5824	2284	3315	8334
2023-10-01	6820	5641	2400	3240	8041
2023-11-01	7106	5963	2530	3433	8494
2023-12-01	6269	6159	2454	3513	8805

Appendix Table 4: ARIMA model results with 2020 excluded from the training set

2019

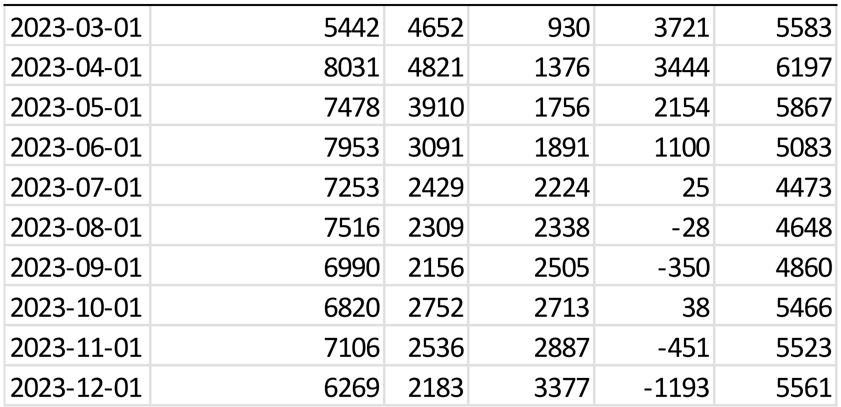


Date	Avg. Cost of Truck	mean	mean_se	CI Lower	CI Upper
2023-03-01	5442	5175	690	4484	5866
2023-04-01	8031	5455	1049	4405	6505
2023-05-01	7478	5671	1393	4278	7065
2023-06-01	7953	6088	1721	4368	7808
2023-07-01	7253	5934	1944	3990	7878
2023-08-01	7516	5804	2138	3666	7942
2023-09-01	6990	5579	2204	3315	7842
2023-10-01	6820	5411	2400	3240	8041
2023-11-01	7106	5963	2530	3433	8494
2023-12-01	6269	6159	2845	3513	8805

Appendix Table 5: ARIMA model results with 2020 excluded from the training set

SARIMA

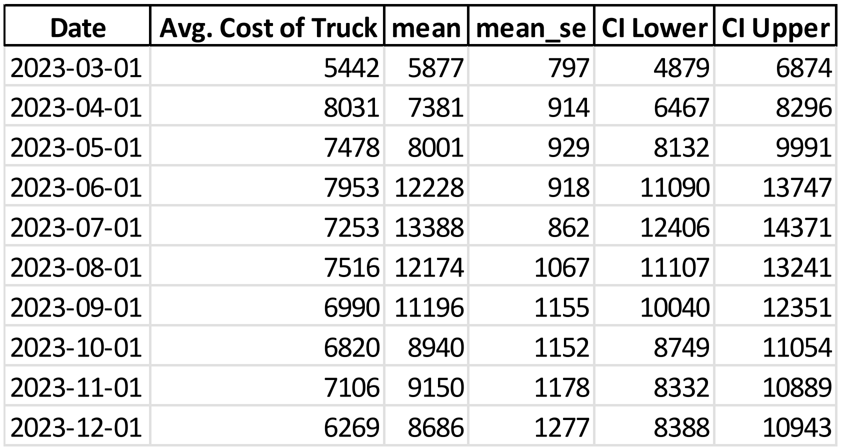
All Years



2023-03-01	5442	4652	930	3721	5583
2023-04-01	8031	4821	1376	3444	6197
2023-05-01	7478	3910	1756	2154	5867
2023-06-01	7953	3091	1891	1100	5083
2023-07-01	7253	2429	2224	25	4473
2023-08-01	7516	2309	2338	-28	4648
2023-09-01	6990	2156	2505	-350	4860
2023-10-01	6820	2752	2713	38	5466
2023-11-01	7106	2536	2887	-451	5523
2023-12-01	6269	2183	3377	-1193	5561

Appendix Table 5: SARIMA model results with all years included

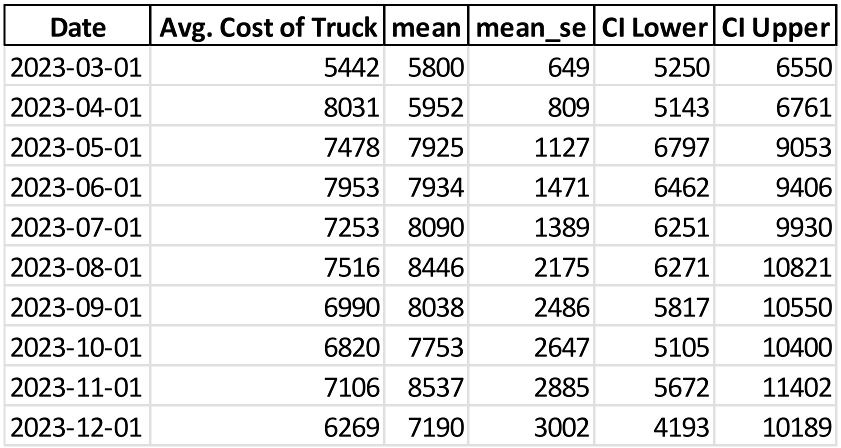
2022



Date	Avg. Cost of Truck	mean	mean_se	CI Lower	CI Upper
2023-03-01	5442	5877	797	4879	6874
2023-04-01	8031	7381	914	6467	8296
2023-05-01	7478	8001	929	8132	9991
2023-06-01	7953	12228	918	11090	13747
2023-07-01	7253	13388	862	12406	14371
2023-08-01	7516	12174	1067	11107	13241
2023-09-01	6990	11196	1155	10040	12351
2023-10-01	6820	8940	1152	8749	11054
2023-11-01	7106	9150	1178	8332	10889
2023-12-01	6269	8686	1277	8388	10943

Appendix Table 6: SARIMA model results with 2022 excluded from the training set

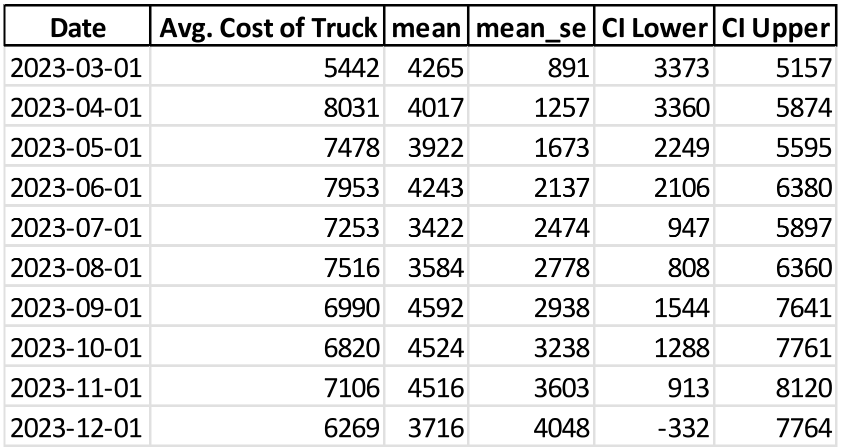
2021



Date	Avg. Cost of Truck	mean	mean_se	CI Lower	CI Upper
2023-03-01	5442	5800	649	5250	6550
2023-04-01	8031	5952	809	5143	6761
2023-05-01	7478	7925	1127	6797	9053
2023-06-01	7953	7934	1471	6462	9406
2023-07-01	7253	8090	1389	6251	9930
2023-08-01	7516	8446	2175	6271	10821
2023-09-01	6990	8038	2486	5817	10550
2023-10-01	6820	7753	2647	5105	10400
2023-11-01	7106	8537	2885	5672	11402
2023-12-01	6269	7190	3002	4193	10189

Appendix Table 7: SARIMA model results with 2021 excluded from the training set

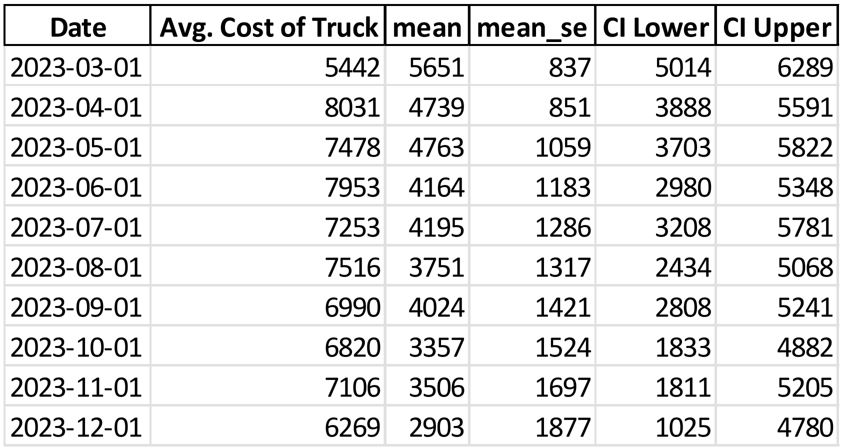
2020



Date	Avg. Cost of Truck	mean	mean_se	CI Lower	CI Upper
2023-03-01	5442	4265	891	3373	5157
2023-04-01	8031	4017	1257	3360	5874
2023-05-01	7478	3922	1673	2249	5595
2023-06-01	7953	4243	2137	2106	6380
2023-07-01	7253	3422	2474	947	5897
2023-08-01	7516	3584	2778	808	6360
2023-09-01	6990	4592	2938	1544	7641
2023-10-01	6820	4524	3238	1288	7761
2023-11-01	7106	4516	3603	913	8120
2023-12-01	6269	3716	4048	-332	7764

Appendix Table 8: SARIMA model results with 2020 excluded from the training set

2019



Date	Avg. Cost of Truck	mean	mean_se	CI Lower	CI Upper
2023-03-01	5442	5651	837	5014	6289
2023-04-01	8031	4739	851	3888	5591
2023-05-01	7478	4763	1059	3703	5822
2023-06-01	7953	4164	1183	2980	5348
2023-07-01	7253	4195	1286	3208	5781
2023-08-01	7516	3751	1317	2434	5068
2023-09-01	6990	4024	1421	2808	5241
2023-10-01	6820	3357	1524	1833	4882
2023-11-01	7106	3506	1697	1811	5205
2023-12-01	6269	2903	1877	1025	4780

Appendix Table 9: SARIMA model results with 2019 excluded from the training set

References

[1] S. Elgin, "Trucking Industry Trends, Statistics, & Forecast – 2024 Edition," 22 05 2024. [Online]. Available: https://www.truckinfo.net/research/trucking-statistics. [Accessed 12 08 2024].

[2] A. Bignell, "Characteristics of Spot-market Rate Indexes for Truckload Transportation," M.Eng. thesis, Eng. Syst. Div., Massachusetts Inst. Technol., Cambridge, MA, USA, 2013.

[3] A. Todd, "What is a freight broker?" Flock Freight, 03 01 2020. [Online]. Available: https://www.flockfreight.com/blog/what-is-a-freight-broker#:~:text=Freight%20brokers%20use%20their%20expertise,pass%20on%20to%20their%20customers. [Accessed 12 08 2024].

[4] D. Sokoloff, "Predicting and Planning for the Future: North American Truckload Transportation," M.S. thesis, Dept. Civil Eng., Massachusetts Inst. Technol., Cambridge, MA, USA, 2020.

[5] DAT, "What Goes into a Rate View Rate," DAT Freight & Analytics, Portland, OR, USA, Oct. 2019.

[6] U. E. I. Administration, "Petroleum & Other Liquids," EIA, [Online]. Available: https://www.eia.gov/petroleum/gasdiesel/. [Accessed 12 08 2024].

[7] FreightWaves, "FreightWaves SONAR Outbound Tender Rejection Index (OTRI)," FreightWaves, [Online]. Available: https://sonar.freightwaves.com/features/outbound-tender-rejection-index. [Accessed 12 08 2024].

[8] CASS, "Cass Freight Index," CASS, [Online]. Available: https://www.cassinfo.com/freight-audit-payment/cass-transportation-indexes/may-2024. [Accessed 12 08 2024].

[9] F. R. Bank, "All Employees, Truck Transportation," FRED, [Online]. Available: https://fred.stlouisfed.org/series/CES4348400001. [Accessed 12 08 2024].

[10] F. R. Bank, "Export Price Index (NAICS): Crop Production," FRED, [Online]. Available: https://fred.stlouisfed.org/series/IY111. [Accessed 12 08 2024].

[11] U. D. o. Agriculture, "Food Markets & Prices," USDA, [Online]. Available: https://www.ers.usda.gov/topics/food-markets-prices/consumer-and-producer-price-indexes/. [Accessed 12 08 2024].

[12] R. H. Shumway and D. S. Stoffer, Time Series Analysis and Its Applications, New York: Springer, 2011.

[13] G. Athanasopoulos and R. J. Hyndman, *Forecasting: Principles and Practice*, 1st ed., OTexts, Melbourne, Australia, 2013.

[14] S. Prabhakaran, "Augmented Dickey Fuller Test (ADF Test) – Must Read Guide," Machine Learning Plus, [Online]. Available: https://www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/. [Accessed 12 08 2024].

[15] R. F. Engle and C. Granger, "Co-Integration and Error Correction: Representation, Estimation, and Testing," *Econometrica,* vol. 55, no. 2, pp. 251-276, 1987.

[16] Z. Bobbitt, "Multiple Linear Regression by Hand (Step-by-Step)," Statology, 18 11 2020. [Online]. Available: https://www.statology.org/multiple-linear-regression-by-hand/. [Accessed 12 08 2024].

[17] Z. Bobbitt, "A Guide to Multicollinearity & VIF in Regression," Statology, 10 03 2019. [Online]. Available: https://www.statology.org/multicollinearity-regression/. [Accessed 12 08 2024].

[18] P. J. Brockwell and R. A. Davis, introduction to Time Series and Forecasting, New York: Springer, 2016.