Project Report

US Supply Chain Risk Analysis

Submitted by

Saran Chandrasekharan Unnithan

Contents

Abstract

In this paper, we present a full-scale analysis on the risks of disrupting U.S. supply chains by an operational hazard like the extremely cold weather event in January 2014 provided in a Kaggle dataset that was composed of 1,000 transactions with 24 qualitative and quantitative attributes.

After the dataset was validated for completeness and consistency, exploratory data analysis and statistical inference methods, such as Z-tests, T-tests, One-Way and Two-Way ANOVA, Chi-Square tests, and regression modeling, were conducted using Python, Excel, and R. It is shown that most perturbations are non-existent or the result of under supply and inclement weather, and there are no statistically significant differences in energy utilization, supplier dependability, or order values between shipping modes, risk groups, or disruption seriousness.

The regression also shows that the influential variables of quantity ordered, communication cost, supplier reliability, and disruption history cannot significantly explain variability in order of value. In sum, these results indicate that the supply chain system modeled in this data set has a steady operating character with little statistical difference across the groups, suggesting that additional unobserved or external influences may be driving shipment risk and value results. These findings also may help inform the U.S. defense logistics community in enhancing risk management, in focusing data-sharing efforts, and in advancing resilience planning.

Introduction

Stable U.S. supply chains are essential for everything, including healthcare and manufacturing, technology, and food distribution. These interlinked systems have been subjected to increasing strains in recent years, with disruptions to logistics, suppliers, and external shocks originating from geopolitical, environmental, and pandemic incidents. Successfully navigating such complex risks is critical, not only to ensure that operations continue to run smoothly, but also that the entire economy can withstand shocks.

In this paper, we provide a full-scale data driven analysis on a real-world U.S. supply chain data set from Kaggle, which includes 1,000 transaction records and 24 features both qualitative and quantitative in nature. The main purposes are: (i) to detect the influencing factors with the greatest correlation with the shipment risk and analyze supplier and buyer behavior under diverse operational situations, and (ii) to identify trends in the frequency and severity of interruptions.

Using Python, Excel, and R, descriptive statistics and exploratory data visualization techniques are used to analyze delivery delays, supplier reliability, disruption types, energy consumption, and mode of shipment. In so doing, the report seeks to provide insight that the U.S. defense logistics community can use to improve risk management and supply chain performance in a highly complex U.S. logistics environment.

Dataset Definition and Description

This study is based on "U.S. Supply Chain Risk Analysis," a dataset that includes transactional data that simulates real supply chain activities. The data set contains 1,000 unique observations on 24 variables, giving a rich mixture of qualitative and quantitative measures. These variables represent the entire process of an order from ordering to delivery, including risk and performance measures.

The dataset comprises both qualitative (15) and quantitative (9) variables. It represents supplier-buyer interactions, orders, disruptions, and model optimization of metrics.

Quantitative Variables:

Quantity Ordered, Order Value (USD), Communication Cost (MB), Energy Consumption (Joules), Historical Disruption Count, Supplier Reliability Score, Available Historical Records, Federated Round, Parameter Change Magnitude.

Qualitative Variables:

Order ID, Buyer ID, Supplier ID, Product Category, Shipping Mode, Delay Days, Disruption Type, Disruption Severity, Organization ID, Data Sharing Consent, Dominant Buyer Flag, Supply Risk Flag, Order/Dispatch/Delivery Dates.

Pre-Cleaning Dataset Variable Summarization

The dataset was observed to be tidy and non-redundant with no missing records before data preprocessing.

Categorical Variable Summary

Some relevant operational insights can be observed from the frequency tables of essential categorical variables.

Product Category: The number of orders is balanced between categories; there are slightly more orders in Electronics (210) and Textile (206).

Shipping Mode: The shipments were uniformly distributed over the modes of shipment, Air (263) being the predominant one.

Disruption Type: A large share of shipments (486) was undisrupted. When they did, the leading causes were short supply (135) and weather (133).

Supply Risk Flag: Most shipments (514) are tagged as risk, emphasizing that risk is a salient feature in this supply chain ecosystem.

Data Sharing Consent: Most suppliers (805) opt to share data, which bodes well for collaborative analytics.

Quantitative Variable Summary

Central Tendency of Quantitative Variables: We utilized Excel to obtain the descriptive statistics for the US Supply Chain Risk Analysis dataset. R and Python were used to further explore and confirm the descriptive statistics. The descriptive statistics of our numerical variables (Quantity Ordered, Order Value, Communication Cost, Energy Consumption, Historical Disruption Count, Supplier Reliability Score, Available Historical Records, Federated Round, and Parameter Change Magnitude) are represented in Table 1 highlighting Mean, Median, Range (Minimum and Maximum values), IQR, Lower and Upper Bounds, Standard Deviation, and Q1 and Q3. A complete view of descriptive statistics is displayed in Table 1.

Dispersion of Quantitative: When reviewing our variables, we can see that Quantity Ordered, Energy Consumption, Supplier Reliability Score, and Parameter Change Magnitude are slightly skewed right. Variables of Order Value, Communication Cost, Historical Disruption Count, Available Historical Records, and Federated Round are slightly left skewed. All nine numerical variables in the dataset have skew values very close to zero ranging from –0.070 to 0.0415, suggesting the dataset is nearly symmetric with no strong pull from extreme low or high values. The kurtosis values of all nine variables lie between 1.76 and 1.90, making our data set platykurtic, which means having light tails and very few extreme outliers. For a full, detailed restatement of the descriptive statistics, please see Table 1.

Table 1: Data Summary of Numerical Variable

Measures	Quantity ordered	Order value (USD)	Communication cost (mb)	Energy consumption (Joules)	Historical disruption count	Supplier reliability score	Available historical records	Federated round	Parameter change magnitude
Mean	504.288	25290.073	2.768	270.855	9.841	0.745	510.317	5.679	0.251
Median	499	25596.005	2.76	268.37	10	0.74	521.5	6	0.255
Min - Max	10- 989	500.58- 49444.6	0.5- 5	50.49- 449.35	0- 19	0.5- 1	10- 988	1- 10	0.001- 0.488
Standard deviation	295.545	14380.329	1.275	130.218	5.656	0.1430	281.700	2.864	0.139
Skewness	0.026	-0.041	-0.043	0.0415	-0.070	0.012	-0.020	-0.031	0.002
Kurtosis	1.763	1.788	1.906	1.8	1.845	1.799	1.791	1.785	1.825
IQR	512.5	24682.592	2.082	220.237	10	0.25	489.25	5	0.237
Q1	248.00	12891.838	1.74	163.59	5.00	0.62	262.25	3.00	0.01
Q3	761.50	37607.41	3.83	384.01	15.00	0.87	752.00	8.00	0.37
Lower Bound	-522.25	- 24132.051	-1.39	-167.03	-10	0.25	-472.38	-4.50	-0.224
Upper Bound	1531.75	74598.319	6.96	714.62	30	1.25	1486.23	15.50	0.726

Table 2. Data Summary of Categorical Variables

Variable Name & Mode							
Product Category	Textile- 206	Electronics- 210	Machinery- 204	Food- 190	Pharma- 190		
Shipping Mode	Air- 263	Rail- 255	Road- 256	Sea- 226			
Delay Days	0 days- 486	1 day-133	2 days- 105	3 days- 99	5 days- 70	7 days- 39	10 days- 68
Disruption Type	None- 486	Shortage- 135	Weather- 133	Customs- 124	Strike- 122		
Disruption Severity	None- 486	Low- 248	Medium- 152	High- 114			
Dominant Buyer Flag	0's- 714	1's- 286					
Data Sharing Consent	0's- 195	1's- 805					
Supplier Risk Flag	0's- 486	1's- 514					

Cleaning Steps Details

Prior to analysis, the dataset was thoroughly checked for quality-related issues, including completeness and format consistency. The review revealed that the dataset was exceptionally clean and uniform.

Missing values: we checked all 24 columns for any null or missing values. The data was complete; no entry was missing.

Duplicated rows: Check if there are any duplicated rows in the data set. This testifies to the nonexistence of duplicated records.

Format and special characters: All the categorical and string-based columns were checked for inconsistencies, e.g. leading or trailing spaces or special characters. None was found. None were encountered.

Outlier Detection: The lower and upper limit for the quantitative data variables were calculated by using the IQR method for identifying the statistical outliers. A few points did lie beyond these bounds, but as contextualized review, they were not seen as erroneous but indeed seemed like operational plausible extremes rather than simple data entry error. Hence, we keep all points. All data points were retained.

As the original data was already of high quality, no clean-up or transformation of the data was necessary. The analysis was carried out on raw data.

Post-Cleaning Dataset Variable Summarization

Since no changes were introduced into the dataset in the quality of data verification procedure, the summarization of the dataset variables after cleaning is the same as the pre-cleaning summaries given in Pre-cleaning.

Tests

Z-Test

One-Tailed Z-test for order value

Alternate Hypothesis (H_1): The average order value for the Electronics product category is significantly higher than the overall average order value.

Null Hypothesis (H_0): The average order value for the Electronics product category is not significantly higher than the overall average order value

H0: $\mu_0 \leq \mu_1$ H1: $\mu_0 > \mu_1$

Assumptions Testing:

Normality Testing: The average order value was determined to be 25,290 (USD). A Shapiro Wilk normality test1 was conducted on R Studio version 2025.09.1+401 for the order value of electronic products. The normality test for the same returned W = 0.94774, and the p-value was $6.666*10^{-07}$ (as indicated in Figure 1 below), indicating non-normal data.

```
> Electro=filter(US_Supply_Chain,Product_Category=="Electronics")
> shapiro.test(Electro$Order_Value_USD)

        Shapiro-Wilk normality test

data:  Electro$Order_Value_USD
W = 0.94774, p-value = 6.666e-07
```

Figure 1. Normality test result for Z Test variable from R- Studio

Transformation of the Order Value: We then conducted logarithmic transformation using the Excel command (LOG10*)*. The normality of order value post transformation returned W = 0.81657; the corresponding p- value was $5.338*10^{-15}$ indicating non-normal data.

We conducted the Z test assuming that the data was normal. The Z score of Order Value resulted in 1.8144. *(Please refer to appendix K, Figure 12 for logarithmic transformation)* The p-value corresponding to z-statistic of the Order value of the electronics is 0.03481.

In a nutshell, we understand that **at the 1% level**, results are **not** statistically significant. We do not have enough evidence to conclude the true mean is greater than 25,290. **At the 5% level** and **10% level**, results **are** statistically significant. We conclude that the average order value for the Electronics product category is **not significantly higher** than the overall average order value. (*Please refer to appendix K, Figure 14 for Z- test results computed using R).*

T-Test

Two-sample T-test for Supply Risk Flag and Supplier Reliability Score.

Alternate Hypothesis (H_1): Suppliers with supplier risk flag 0 have significantly higher reliability scores.

Null Hypothesis (H_0): Suppliers with supplier risk flag 0 do not have significantly higher reliability scores.

$H0: \mu_0=\mu_1$ $H1: \mu_0>\mu_1$

Assumptions Testing:

Normality Testing: The low and high-risk reliability scores were considered to conduct the right-tailed t-test. This test for reliability scores resulted in W= 0.96024 and W= 0.9482, p- value $3.524*10^{-10}$ and $1.982*10^{-12}$ for 0 and 1 reliability scores respectively, indicating non-normal data (*Please refer to Figure 2 below*). Due to non-normal data, the logarithmic transformation was necessary to proceed further.

```
        Shapiro-Wilk normality test

data:  Low_risk$Supplier_Reliability_Score
W = 0.96024, p-value = 3.524e-10


        Shapiro-Wilk normality test

data:  High_risk$Supplier_Reliability_Score
W = 0.9482, p-value = 1.982e-12
```

Figure 2. Normality test result for T-Test variables from R- Studio

Transformation of the Supplier Reliability Score variable: We conducted the logarithmic transformation on Supplier Reliability Score using the Microsoft Excel function and command LOG 10 *(Please refer to Figure 15 in Appendix L)*. As seen in Figure 16, the Shapiro Wilk normality test post logarithmic transformation resulted in W=0.9566 and W=0.94442 where p value is $9.14*10^{-11}$ which is less than 0.05, indicating non-normal distributions. However, we move forward assuming that data are normal.

Homogeneity of Variance Test: Levene's test for homogeneity of variances was conducted, using R version 2025.05.1+513. We find that at 5% significance level, the assumptions of equal variances between groups are met. *(Please refer to Figure 17 in Appendix L)*

As seen from *Appendix L,* Suppliers with low-risk flags have significantly higher reliability scores. The p-value is 0.9324, which is very high (much greater than the typical significance level of 0.05). Therefore, we conclude at all conventional significance levels (1%, 5%, and 10%), the p-value of 0.9324 is substantially larger than the threshold.

Therefore, there is no statistically significant evidence to conclude that the true difference in means between supplier risk flag 0 and supplier risk flag 1 is greater than 0 and Suppliers with supplier risk flag 0 do not have significantly higher reliability scores. The slight observed difference in sample means (0.7388 vs. 0.7523) is likely due to random chance. (*Please refer to appendix L, Figure 18 for T- test results computed using R)*

One-Way ANOVA Test

Alternate Hypothesis (H_1): Air, Road, and Rail shipping mode have similar energy consumption.

Null Hypothesis (H_0): Air, Road, and Rail shipping mode do not have similar energy consumption.

Assumption Testing

Normality Testing: The Shapiro-Wilk tests were run to check if the Energy_Consumption_Joules data is normally distributed within each Shipping_Mode group from Figure 3, we get the p-value for Air is $1.821*10^{-07}$, Rail is $2.004*10^{-06}$ and Road is $1.289*10^{-07}$ which are all significantly less than 0.05. This shows that the data for all three shipping modes is not normally distributed.

```
> Air=filter(US_Supply_Chain,Shipping_Mode=="Air")
> Rail=filter(US_Supply_Chain,Shipping_Mode=="Rail")
> Road=filter(US_Supply_Chain,Shipping_Mode=="Road")
> shapiro.test(Air$Energy_Consumption_Joules)

        Shapiro-Wilk normality test

data:  Air$Energy_Consumption_Joules
W = 0.95331, p-value = 1.821e-07

> shapiro.test(Rail$Energy_Consumption_Joules)

        Shapiro-Wilk normality test

data:  Rail$Energy_Consumption_Joules
W = 0.96076, p-value = 2.004e-06

> shapiro.test(Road$Energy_Consumption_Joules)

        Shapiro-Wilk normality test

data:  Road$Energy_Consumption_Joules
W = 0.95069, p-value = 1.289e-07
```

Figure 3. Normality test result for One- way ANOVA variables from R- Studio

Transformation of the Supplier Reliability Score variable: After finding the original Energy Consumption Joules was not normally distributed, applied a log transformation to see if that would normalize the data. However, the new Shapiro-Wilk test results on the log-transformed data from figure 6 still show p-values for Air is $1.334*10^{-07}$, Rail is $1.903*10^{-09}$ and Road is $1.121*10^{-09}$ that are all significantly less than 0.05. This means the logarithmic transformation failed in making the data normal. *(Please refer to Figure 19 in Appendix M)*

Homogeneity Of Variance: The Levene's Test was performed to check if the variance of Energy Consumption Joules is equal across the different Shipping Mode groups. From *Figure 20, Appendix M*, the resulting p-value is 0.7233. Since this p-value is greater than 0.05, the variances are homogeneous, thus satisfying this assumption for testing.

The Standard One-Way ANOVA test is performed using **aov () command in** R (version 4.5.1), from Figure 8 we get the f value as 0.543, p-value is 0.582 is much greater than 0.05 show that there is no statistically significant difference in the mean Energy Consumption Joules across the different Shipping Mode groups. *(Please refer to Appendix M, figure 21)*.

Post-hoc Tuckey HSD Test:

This output is the Tukey HSD post-hoc test, which was run after the ANOVA. The test compares all possible pairs of Shipping Mode groups *(please refer to Appendix M, figure 22)*. Crucially, all p-adjusted values are greater than 0.05 (the smallest is 0.609), confirming that no two individual shipping modes are statistically different from each other in terms of their mean Energy Consumption Joules.

Two-Way ANOVA

Two- Way Anova for Shipping Mode, Disruption Severity and Energy Consumption: Shipping Mode and Disruption Severity are independent variables and Energy Consumption is the dependent variable

Alternative Hypotheses:

H_{11}: At least one mean energy consumption differs across shipping modes

H_{12}: At least one mean energy consumption differs across disruption severity levels

H_{13}: There is a significant interaction between Shipping Mode and Disruption Severity on Energy Consumption

Null Hypotheses:

H_{01}: The energy consumption does not impact air, rail, road shipping modes.

H_{02}: The energy consumption does not impact high and low disruption severities.

H_{03}: There is no interaction effect between shipping mode and disruption severity on the energy consumption.

Assumptions Testing:

Normality Testing: The means of energy consumption for air, rail and road shipping modes are 258.85, 269.6, 268.37, respectively. The means of energy consumption for high and low types of disruption severity is 290.19, 277.49, respectively. The Shapiro Wilk normality test was conducted on R Studio for shipping modes, disruption severity, and energy consumption. As seen in *figure 4* below, we find that p-value is less than 0.05. Based on p-value, we conclude that samples do not belong to a normally distributed population.

```
> Air=filter(df,Shipping_Mode=="Air")
> Rail=filter(df,Shipping_Mode=="Rail")
> Road=filter(df,Shipping_Mode=="Road")
> shapiro.test(Air$Energy_Consumption_Joules)

        Shapiro-Wilk normality test

data:  Air$Energy_Consumption_Joules
W = 0.94615, p-value = 0.001319

> shapiro.test(Rail$Energy_Consumption_Joules)

        Shapiro-Wilk normality test

data:  Rail$Energy_Consumption_Joules
W = 0.9511, p-value = 0.001589

> shapiro.test(Road$Energy_Consumption_Joules)

        Shapiro-Wilk normality test

data:  Road$Energy_Consumption_Joules
W = 0.94124, p-value = 0.0003953
```

Figure 4. Normality test result for Two- way ANOVA variables from R- Studio

Transformation of Energy Consumption Variable: After applying logarithmic transformation for energy consumption and we repeated the Shapiro Wilk test. Even then p value is less than 0.05 *(please refer figure 23 in Appendix N)*, meaning that the distribution is not normal. However, given the large sample size, we proceeded with the ANOVA under the assumption of approximate normality based on the Central Limit Theorem.

Homogeneity of Variance Test: We conducted Levene's test for equality of variances using R Studio version 4.5.1, to find if the variance of energy consumption joules is same between the combined groups of shipping mode and disruption type. After the test we found that the p-value is 0.232 *(please refer figure 24 in Appendix N)*, which is more than the significance level of 0.05 and the assumption of homogeneity of variances is satisfied. (F value is 1.3801)

As assumptions are met, a standard two-way ANOVA test is performed. The two-way ANOVA produced an F-statistic of 0.032 ($p = 0.968$) *(please refer figure 25 in Appendix N)*, for shipping mode, 0.341 ($p = 0.559$) for Disruption Severity, and 1.221 ($p = 0.297$) for their interaction. Based on these results, we conclude that at the 1%, 5%, and 10% significance levels, the energy consumption joules do not differ significantly across shipping modes, disruption severities or their interaction in our dataset.

Post-hoc Tukey HSD Test

Post-hoc Tukey HSD comparisons were conducted to identify significant differences in mean energy consumption joules across shipping modes and disruption severity. For Shipping mode, based on the mean differences, the p values are 0.97 between rail and air, 0.973 between road and air, 0.99 between road and rail. For disruption severity, the p value based on mean difference is 0.559. For all the mean differences of pairwise comparisons of disruption severity and shipping modes and in their interaction, there is a significant difference in the mean values but the p values for all the combinations exceed 0.05 and 0.1. *(Please refer figure 26 in Appendix N)*,

Hence as the p values are not significant at 1%, 5%, 10% level of confidence we fail to reject the null hypothesis meaning that there is no statistical significance of energy consumption joules on shipping modes, disruption severity and on their interaction.

Chi- Square test:

Two Way Chi- Square Test for Disruption severity and Product Category

Alternative Hypothesis (H_1): Disruption severity depends on product category.

Null hypothesis (H_0): Disruption severity is not dependent on product category.

The Chi-square test of independence was conducted using R Studio version 4.5.1 to determine the relationship between Product Category and Disruption Severity. The Chi-square test results showed the values as χ^2= 7.3044, degrees of freedom=12 and p value of 0.8368 *(please refer figure 27, in Appendix O)*, Based on these results we conclude that at 1%, 5% and 10% level of significance there is no statistically significant dependency between Product Category and Disruption severity meaning that the distribution of disruption severity does not depend on type of product. The disruptions seem to occur similarly across all product categories.

Linear Regression

Alternate Hypothesis(H_1): Quantity Ordered does have a linear effect on Order Value USD

Null Hypothesis(H_0): Quantity Ordered does not have a linear effect on Order Value USD

Simple linear regression was performed using R Studio version 4.5.1 on Quantity ordered and Order value to know the effect of Quantity ordered on the Order value. From figure 15, The predictor Quantity Ordered is not statistically significant, as evidenced by its p-value of 0.492, which exceeds all conventional significance thresholds (1%, 5%, and 10%). Because the predictor is not significant at any of these levels, it provides no meaningful explanatory power for the dependent variable, accounting for only 0.047% of the variation in Order Value USD. As a result, the overall regression model is also not statistically significant, indicating that the model does not reliably explain variation in the outcome. This outcome reflects a very weak linear association, demonstrating that Quantity Ordered does not exert a meaningful linear influence on Order Value USD in this dataset. *(Please refer figure 28, in Appendix P)*,

Multi Linear Regression

Multi Linear Regression for Order Value, Quantity Ordered, Historical Disruption Count, Supplier Reliability Score and Communication Cost:

Alternative Hypothesis (H_1): Quantity Ordered, Historical Disruption Count, Supplier Reliability Score, Communication Cost has significant linear effect on Order Value

Null Hypothesis (H_0): Quantity Ordered, Historical Disruption Count, Supplier Reliability Score, Communication Cost has no significant linear effect on Order Value

Multi Linear Regression Equation:

$$Order\ Value = \beta 0 + \beta 1(Quantity\ Ordered) + \beta 2(Historical\ Disruption\ Count) + \beta 3(Supplier\ Reliability\ Score) + \beta 4(Communication\ Cost) + \epsilon$$

These coefficients quantify the relationship between each predictor and the target variable, which is the **Order Value**. The tests were performed using R Studio version 4.5.1. The multi linear regression test returned p-value for **Quantity Ordered** is 0.429, which is not significant at 5%,10% and at 1%. This provides enough evidence to support quantity ordered does not have a linear relationship with Order Value at all significant levels.

Secondly, the **Historical Disruption Count** is 0.784, which is not significant at 1%, 5%,10%. This proves there is not enough evidence to support that historical disruption count has a linear relationship with Order value. Likewise, the p-value for **Supplier Reliability Score** is 0.385, which is not significant at 1%, 5%,10%. This proves there is not enough evidence to support the Supplier Reliability Score having a linear relationship with Order Value. Finally, the p-value for **Communication Cost** is 0.171, which is not significant at 1%, 5%, and 10%. This proves there is not enough evidence to support the Communication Cost having a linear relationship with Order Value.

From the above testing we conclude, Order Value is driven by factors NOT included in this model. Selected predictors do not statistically explain variation in order value.

From the adjusted R-squared value given in *Figure 30 in Appendix P* shows that only 0.07% of the variation in Order value is explained by the predictors variables such as Quantity Ordered, Historical Disruption Count, Supplier Reliability Score and Communication Cost, proposing that the model has very less predictive power.

Linear Regression Coefficients:

Intercept (28313.534): This is the expected value of the Order Value when all the predictor variables are zero. This serves as the baseline from which the effect of each predictor variable is measured. **Quantity Ordered (1.222).** For each additional unit of quantity ordered, the order value is expected to increase by 1.222, assuming all other variables remain constant. **Historical**

Disruption Count (-22.058) This negative coefficient suggests that for every unit increase of the historical disruption count, there is a decrease in the order value by 22.058, while all other variables remain constant. **Supplier Reliability Score (-2767.892)** The negative coefficient in the given scenario shows that for every unit increase in the supplier reliability score, there is an expected decrease of **2767.892** in order value, keeping other variables constant. **Communication Cost (-490.660)** From this negative coefficient here, we understand that for every unit increase in the communication cost, there is an expected decrease of **490.660** in order value, while assuming other variables constant.

The variance inflation factor (VIF) values for all predictors are approximately 1 (*please refer to Figure 31, in appendix P)*, indicating no multicollinearity exists among all the predictor variables such as Quantity Ordered, Historical Disruption Count, Supplier Reliability Score and Communication Cost. This suggests that the predictors are independent of each other and do not misshape the linear regression estimates.

Conclusion

This study presents a detailed analysis of U.S. supply chain performance serving 1,000 real-world–simulated transactional records. Results across a number of hypothesis tests and models consistently indicate that major operational variables (e.g., shipping mode, disruption severity, supplier reliability, order size) have no statistically significant impact on order value, energy usage, and shipment reliability. In some cases, there were differences in means but these differences were due to chance rather than any meaningful structural difference. The regression models had very low R-squared values, suggesting that the variables in this dataset do not capture what primarily determines the value and risk of a shipment.

Yet, the descriptive results portray a system in which risk abounds, empirically observed disruptions are not so frequent, and these disruptions are fairly evenly spread across product categories. The high amount of supplier data sharing also implies that there is significant scope for collaborative analytics and federated learning solutions. Taken together, these findings provide further support for the investigation of additional levels of data granularity, including other operational, environmental, or contractual considerations that may enhance predictiveness in subsequent supply chain risk analyses. For practitioners and defence logistics organizations, the research demonstrates the necessity of prioritizing systemic drivers of performance over isolated operational measures in constructing resilient and adaptive supply chain strategies.

Upon further analysis of data through hypothesis testing and carrying out tests taught in the class, it is evident that the dataset does not establish and significant relationships between variable. Only the Z test proved a weak significance between the population and sample mean of the order valued and the Electronic product at the 5% and 10% thresholds. However, the other tests have proved to be consistently insignificant which implied that this a lack or correlation or causation present between the numerical and categorical variable present in the dataset. Although logically disruption severity could be dependent on the shipping mode at time, our dataset lacked the statistical evidence to prove the hypothesis.

Regression models on the other hand, we were unable to provide any meaningful predictors for the Order Value variable in both simple and multiple linearity models. The absence of multicollinearity between the variables could best explain the lack of statistical evidence to prove all or any relationships at all. Overall, we have observed that our dataset signals were weak thereby limiting statistical evidence to prove any of our rational and logical hypothesis.

Recommendations: There is a scope for further analysis of the US supply chain risk management, provided there is an immense increase in the variability of the dataset and the integrate a product category partitioned using the help of a stock keeping unit (SKU) variable. Thus concluding, although the data pertains to real world, the data set could be improved with many more variables and rows to analyze existing correlations through hypothesis testing and conducting unbiased predictive analysis.

Appendices

Appendix A. Description of all the variables in the dataset

Order_ID: 1000 Unique identifier for each shipment/transaction.

Buyer_ID: 50 Unique identifier for the buyer making the order.

Supplier_ID: 30 Unique identifier for the supplier fulfilling the order.

Product_Category: Category of product (eg. Electronics, Food, Textiles).

Quantity_Ordered: Number of units ordered in the transaction.

Order_Date: Date when the order was placed.

Dispatch_Date: Date when the order was shipped.

Delivery_Date: Date when the order was delivered.

Shipping_Mode: Method of shipment (Air, Sea, Road, Rail).

Order_Value_USD: Value of the order in U.S. dollars.

Delay_Days: Number of days late compared to planned delivery.

Disruption_Type: Type of disruption (eg. Weather, Strike, Equipment).

Disruption_Severity: Severity level of disruption (Low, Medium, High).

Historical_Disruption_Count: Number of past disruptions for the supplier.

Supplier_Reliability_Score: Score (0–1) showing how reliable the supplier is.

Organization_ID : 20 Unique identifier for the organization involved.

Dominant_Buyer_Flag: Binary indicator (0/1) if a buyer is dominant for supplier.

Available_Historical_Records: previous historical records between supplier & buyer.

Data_Sharing_Consent: Binary indicator (0/1) if supplier consents to share data.

Federated Round: Iteration or training round number in a federated learning process for Supply chain optimization.

Parameter_Change_Magnitude: Amount of change in model parameters.

Communication_Cost_MB: Amount of communication data used (megabytes).

Energy_Consumption_Joules: Energy used to process/transport the shipment (joules).

Supply_Risk_Flag – Binary indicator determining if a shipment is risky or not.

Appendix B. Shipping Mode Positioning based on Order Value and Quantity ordered

Row Labels	Sum of Order_Value_USD	Sum of Quantity_Ordered
Air	**3619116.23**	**70332**
Pharma	817155.29	20198
Food	654716.72	15542
Machinery	874388.39	14345
Textiles	585118.06	10204
Electronics	687737.77	10043
Road	**3209890.64**	**63936**
Textiles	714659.83	15984
Machinery	737464.8	14814
Food	575493.16	11742
Electronics	690503.27	11299
Pharma	491769.58	10097
Rail	**2881662.86**	**59443**
Electronics	811702.74	17986
Machinery	803393.94	16981
Pharma	425138.09	9648
Textiles	446247.32	7865
Food	395180.77	6963
Sea	**2620264.4**	**48580**
Electronics	548246.9	12143
Textiles	639180.05	10023
Food	491172.95	9810
Pharma	560921.35	9296
Machinery	380743.15	7308
Grand Total	**12330934.13**	**242291**

Appendix C. Suppliers based on Quantity Ordered and Order Value

Row Labels	Sum of Quantity_Ordered	Sum of Order_Value_USD
S24	25656	1246160.45
S4	22814	1224751.12
S25	22568	1118566.8
S7	19413	1050956.12
S9	17551	1043628.11
S16	20672	1027280.62
S29	16387	1000010.58
S23	20257	975418.98
S26	19952	935512.51
S27	12237	916059.17
S1	20418	889887.87
S10	18580	863460.71
S3	16396	847631.91
S12	17599	824810.99
S21	12527	821360.29
S22	17345	805828.62
S19	13548	787063.64
S17	17348	778023.5
S11	17254	771866.33
S14	17358	745168.58
S6	12132	742148.67
S28	15106	739820.19
S5	14492	719954.62
S20	12366	694908.44
S30	12608	652712.31
S13	14138	644504.97
S8	15664	644285.99
S18	14292	640575.46
S2	16268	637418.04
S15	11342	500298.08
Grand Total	**504288**	**25290073.67**

Appendix D. Supplier Positioning

Supplier ID	Count of Supplier_ID
S24	48
S25	45
S16	44
S4	42
S7	42
S26	42
S10	39
S23	39
S1	37
S9	35
S28	35
S29	35

	Order_Value_USD	On-time delivery
Textiles	594469.34	
S9	214878.72	20.58958722
S25	197516.74	17.65801917
S19	182073.88	23.13331105
Electronics	560920.66	
S30	208365.7	31.92305351
S7	179086.66	17.0403556
S4	173468.3	14.16355512
Pharma	515953.04	
S25	181685.3	16.24268662
S4	170500.27	13.92121772
S20	163767.47	23.56676946
Machinery	453410.39	
S16	153758.65	14.9675412
S22	152987.54	18.98512118
S7	146664.2	13.95531147
Food	387982.07	
S16	141449.52	13.76931651
S1	127781.01	14.35922596
S24	118751.54	9.529394068

Figure 5. Suppliers Based on Efficiency

Row Labels	On-time delivery	Relative Volume
Textiles		
S9	20.58958722	83.74749094
S25	17.65801917	89.76105766
S19	23.13331105	63.1590932
Electronics		
S30	31.92305351	52.37787076
S7	17.0403556	84.33553801
S4	14.16355512	98.28197645
Pharma		
S25	16.24268662	89.76105766
S4	13.92121772	98.28197645
S20	23.56676946	55.76396202
Machinery		
S16	14.9675412	82.43566228
S22	18.9851211	64.66491695
S7	13.95531147	84.33553801
Food		
S16	13.76931651	82.43566228
S1	14.35922596	71.41037657
S24	9.529394068	100

OTD% = OTD/Total Quantity Ordered

Relative Vol = Quantity ordered/Max (Quantity Ordered)

Figure 6. Supplier Positioning - On-Time Delivery vs Relative Volume

Appendix E. Buyer Positioning

Buyer ID	Count of Buyer_ID
B12	31
B40	30
B29	29
B34	27
B47	26
B4	25
B23	25
B22	25
B6	24
B30	24

Dominant_Buyer_Flag	1		
Row Labels	**Count of Supplier_ID**	**Supplier Base**	**Dominance Ratio**
B16	12	22	54.54
B12	9	31	29.03
B34	9	27	33.33
B45	9	23	39.13
B39	8	20	40
B41	8	18	44.44
B1	8	20	40
B29	8	29	27.58

Dominance Ratio = Buyer dominance on Suppliers/total supplier base

Figure 7. Dominant Buyer Flag

Appendix F. Supplier Optimization

Row Labels	Average of Parameter_Change_Magnitude	Sum of Federated_Round	Average of Supplier_Reliability_Score
S24	0.266	267	0.72
S25	0.244	264	0.725
S26	0.256	257	0.717
S16	0.272	256	0.752
S7	0.274	247	0.772

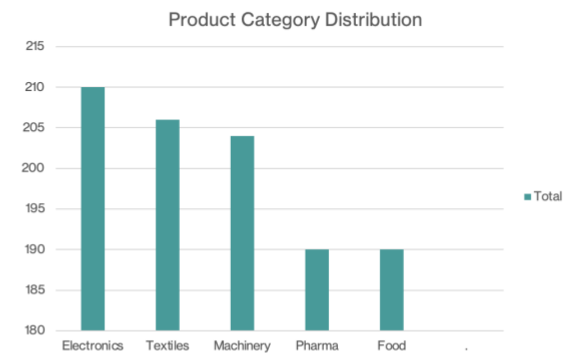
Delay_Days	0			
Supplier ID	Average of Parameter_Change_Magnitude	Sum of Federated_Round	Sum of Order_Value_USD	Average of Supplier_Reliability_Score
S25	0.227	136	617127.63	0.725
S4	0.271	103	562706.25	0.714
S7	0.302	149	535689.97	0.755
S9	0.261	98	526246.36	0.664
S11	0.2	113	499096.01	0.689

Figure 8. Best Supplier based on various parameters

Delay_Days	0			
Row Labels	Average of Communication_Cost_MB	Sum of Order_Value_USD	Average of Energy_Consumption_Joules	Count of Shipping_Mode
Air	2.632	3619116.23	250.638	138
Rail	2.743	2881662.86	263.042	118
Road	2.890	3209890.64	265.644	125
Sea	2.985	2620264.4	286.429	105

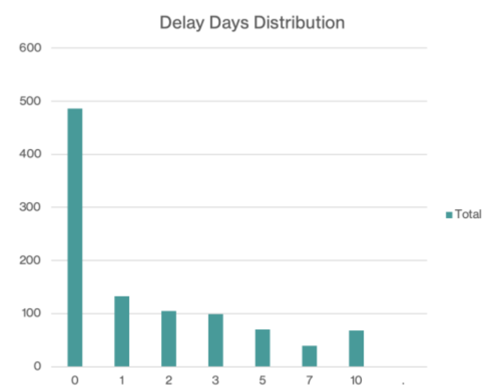
Figure 9. Best Shipping Mode

Appendix G. Histograms



Histogram showing distribution by Shipping Mode

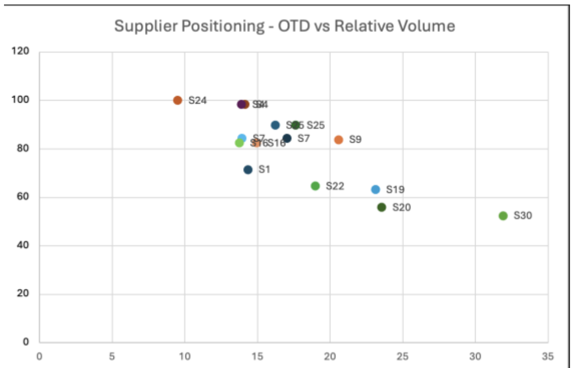
Histogram showing distribution by Product Category



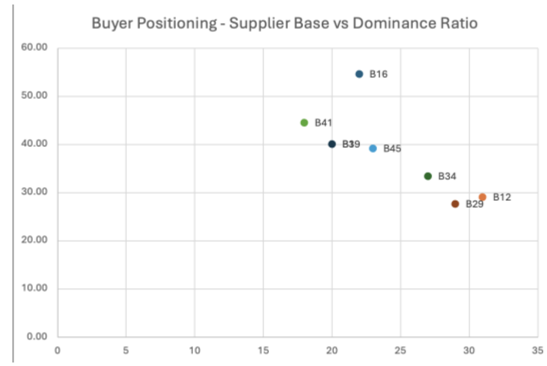
Histogram showing distribution by Delay Days Category

Figure 10. Histograms for categorical variables

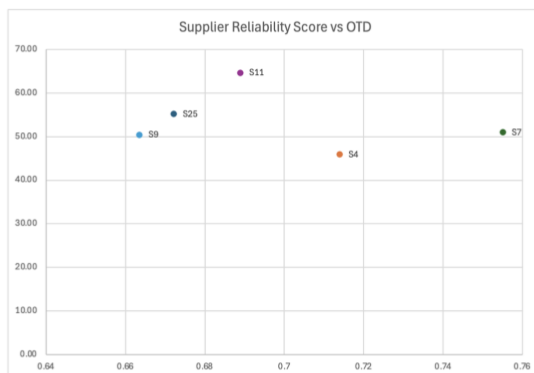
Appendix H. Scatter Plots



Scatter plot for supplier positioning



Scatter plot for buyer positioning



Supplier Positioning – On-Time Delivery vs Supplier Reliability Score

Figure 11. Scatter Plots

Appendix I. Python Code for Descriptive Statistics

```
from google.colab import files
import pandas as pd

# Upload file
uploaded = files.upload()
df = pd.read_csv("Group7_USSupplyChainRiskAnalysis.csv")
display(df.head())

# Python code descriptive statistics for each numerical variable

import numpy as np

# List of numerical columns to analyze
numerical_cols = [
    'Quantity_Ordered',
    'Order_Value_USD',
    'Communication_Cost_MB',
    'Energy_Consumption_Joules',
    'Historical_Disruption_Count',
    'Supplier_Reliability_Score',
    'Available_Historical_Records',
    'Federated_Round',
    'Parameter_Change_Magnitude'
]

# Calculating the descriptive statistics for each numerical column
for col in numerical_cols:
    print(f"--- Statistics for {col} ---")

    # Central Tendencies
    mean_val = df[col].mean()
    median_val = df[col].median()
    mode_val = df[col].mode().tolist() # Get all modes
```

```
print(f"Mean: {mean_val:.2f}")
print(f"Median: {median_val:.2f}")
print(f"Mode: {mode_val}")

# Measures of Dispersion
std_dev = df[col].std()
skewness = df[col].skew()
kurtosis = df[col].kurtosis()

print(f"Standard Deviation: {std_dev:.2f}")
print(f"Skewness: {skewness:.2f}")
print(f"Kurtosis: {kurtosis:.2f}")

# Quartiles, IQR, Bounds, Min, Max
quartiles = df[col].quantile([0.25, 0.5, 0.75])
Q1 = quartiles[0.25]
Q3 = quartiles[0.75]
IQR = Q3 - Q1

lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

min_val = df[col].min()
max_val = df[col].max()

print(f"Q1 (25th percentile): {Q1:.2f}")
print(f"Q3 (75th percentile): {Q3:.2f}")
print(f"IQR (Interquartile Range): {IQR:.2f}")
print(f"Lower Bound (for outliers): {lower_bound:.2f}")
print(f"Upper Bound (for outliers): {upper_bound:.2f}")
print(f"Minimum: {min_val:.2f}")
print(f"Maximum: {max_val:.2f}")
print("-" * 30)
```

Appendix J. Descriptive Statistics before Data Cleaning

```
--- Statistics for Order_Value_USD ---
Mean: 25290.07
Median: 25596.01
Mode: [500.58, 506.67, 567.0, 572.95, 598.61, 689.09, 715.98, 763.02, 888.16, 911.84, 914.0, 938.73
Standard Deviation: 14380.33
Skewness: -0.04
Kurtosis: -1.21
Q1 (25th percentile): 12891.84
Q3 (75th percentile): 37574.43
IQR (Interquartile Range): 24682.59
Lower Bound (for outliers): -24132.05
Upper Bound (for outliers): 74598.32
Minimum: 500.58
Maximum: 49945.18
```

```
--- Statistics for Quantity_Ordered ---
Mean: 504.29
Median: 499.00
Mode: [327]
Standard Deviation: 295.55
Skewness: 0.03
Kurtosis: -1.24
Q1 (25th percentile): 248.00
Q3 (75th percentile): 760.50
IQR (Interquartile Range): 512.50
Lower Bound (for outliers): -520.75
Upper Bound (for outliers): 1529.25
Minimum: 10.00
Maximum: 999.00
```

```
--- Statistics for Communication_Cost_MB ---
Mean: 2.77
Median: 2.76
Mode: [1.87]
Standard Deviation: 1.28
Skewness: -0.04
Kurtosis: -1.09
Q1 (25th percentile): 1.74
Q3 (75th percentile): 3.82
IQR (Interquartile Range): 2.08
Lower Bound (for outliers): -1.38
Upper Bound (for outliers): 6.95
Minimum: 0.50
Maximum: 5.00
```

Appendix K: Screen grabs pertaining to Z Test

Product_Category	Quantity_Ordered	Order_Date	Dispatch_Date	Delivery_Date	Shipping_Mode	Order_Value_USD	log_10_Order_Valu
Electronics	735	05/11/23	07/11/23	14/11/23	Sea	7571.09	3.879158409
Electronics	622	23/03/23	24/03/23	03/04/23	Air	42373.11	4.62709034
Electronics	414	30/07/23	03/08/23	12/08/23	Rail	30873.4	4.48958446
Electronics	959	31/08/23	04/09/23	09/09/23	Air	40507.64	4.607536942
Electronics	670	28/11/23	29/11/23	30/11/23	Rail	15328.5	4.185499658
Electronics	318	01/08/23	05/08/23	10/08/23	Air	39828.52	4.600194169
Electronics	322	04/01/23	08/01/23	16/01/23	Air	20442.23	4.31052827
Electronics	815	27/04/23	28/04/23	16/05/23	Road	11688.44	4.067756552
Electronics	673	09/02/23	13/02/23	25/02/23	Rail	35840.66	4.554375999
Electronics	865	10/04/23	12/04/23	23/04/23	Rail	38879.63	4.589722123
Electronics	671	30/11/23	01/12/23	05/12/23	Rail	31663.11	4.50055357
Electronics	891	21/11/23	23/11/23	28/11/23	Rail	16220.55	4.210065576
Electronics	478	10/02/23	14/02/23	18/02/23	Sea	506.67	2.70472519
Electronics	58	14/11/23	15/11/23	18/11/23	Rail	25800.89	4.411634687
Electronics	976	12/11/23	13/11/23	14/11/23	Road	2819.17	3.450121265
Electronics	698	25/03/23	28/03/23	31/03/23	Road	42047.26	4.6237377
Electronics	351	22/11/23	23/11/23	03/12/23	Rail	37174.75	4.570248057
Electronics	570	12/06/23	15/06/23	23/06/23	Sea	42214.24	4.625458975
Electronics	778	18/09/23	19/09/23	24/09/23	Rail	39552.16	4.597170206

Figure 12. Screen grab of logarithmically transformed variable for Z- Test

```
> Electro$log10_Order_Value <- log(Electro$Order_Value_USD, base = 10)
> shapiro.test(Electro$log10_Order_Value)

	Shapiro-Wilk normality test

data:  Electro$log10_Order_Value
W = 0.81657, p-value = 5.338e-15
```

Figure 13. Screen grab of logarithmically transformed variable for Z- Test

```
> library(BSDA)
> z.test(x = Electro$Order_Value_USD,mu = 25290,sigma.x = 14380,alternative = "greater")

	One-sample z-Test

data:  Electro$Order_Value_USD
z = 1.8144, p-value = 0.03481
alternative hypothesis: true mean is greater than 25290
95 percent confidence interval:
 25458.21       NA
sample estimates:
mean of x
 27090.42
```

Figure 14. Z- Test result from R Studio

Appendix L: Screen grabs pertaining to T- Test

Log_10_Supplier_Reliability_Scor	Supplier_Reliability_Score
-0.13667714	0.73
-0.055517328	0.88
-0.022276395	0.95
-0.036212173	0.92
-0.22184875	0.6
-0.142667504	0.72
-0.22184875	0.6
-0.200659451	0.63
-0.124938737	0.75
-0.27572413	0.53
-0.214670165	0.61
-0.013228266	0.97
-0.13667714	0.73
-0.142667504	0.72
-0.292429824	0.51
-0.207608311	0.62
-0.031517051	0.93

Figure 15. Screen grab of logarithmically transformed variable for T- Test

```
        Shapiro-Wilk normality test

data:  Low_risk$log10_Supplier_Reliability_Score
W = 0.9566, p-value = 9.14e-11


        Shapiro-Wilk normality test

data:  High_risk$log10_Supplier_Reliability_Score
W = 0.94442, p-value = 5.882e-13
```

Figure 16. Shapiro Wilk Test result post logarithmic transformation

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1   2.616 0.1061
      998
```

Figure 17. Screen grab of Levene's Test for Homogeneity

```
	Two Sample t-test

data:  Supplier_Reliability_Score by as.factor(Supply_Risk_Flag)
t = -1.4955, df = 998, p-value = 0.9324
alternative hypothesis: true difference in means between group 0 and group 1 is greater than 0
95 percent confidence interval:
 -0.02842133         Inf
sample estimates:
mean in group 0 mean in group 1
      0.7388066       0.7523346
```

Figure 18. Screen grab of T- Test results using R

Appendix M: Screen grabs pertaining to One Way ANOVA Test

```
> Air$log10_Energy_Consumption_Joules<- log(Air$Energy_Consumption_Joules, base = 10)
> Rail$log10_Energy_Consumption_Joules<- log(Rail$Energy_Consumption_Joules, base = 10)
> Road$log10_Energy_Consumption_Joules<- log(Road$Energy_Consumption_Joules, base = 10)
> shapiro.test(Air$log10_Energy_Consumption_Joules)

        Shapiro-Wilk normality test

data:  Air$log10_Energy_Consumption_Joules
W = 0.93238, p-value = 1.334e-09

> shapiro.test(Rail$log10_Energy_Consumption_Joules)

        Shapiro-Wilk normality test

data:  Rail$log10_Energy_Consumption_Joules
W = 0.92027, p-value = 1.903e-10

> shapiro.test(Road$log10_Energy_Consumption_Joules)

        Shapiro-Wilk normality test

data:  Road$log10_Energy_Consumption_Joules
W = 0.92966, p-value = 1.121e-09
```

Figure 19. Shapiro Wilk Test result post logarithmic transformation

```
> leveneTest(Energy_Consumption_Joules ~ as.factor(Shipping_Mode), data = USSupplyChain)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   2  0.3241 0.7233
      771
```

Figure 20. Screen grab of Levene's Test for Homogeneity

```
> ShippingMode_Anova<-aov(Energy_Consumption_Joules ~ as.factor(Shipping_Mode), data = USSupplyC
ain)
> summary(ShippingMode_Anova)
                          Df   Sum Sq Mean Sq F value Pr(>F)
as.factor(Shipping_Mode)   2    18063    9031   0.543  0.582
Residuals                771 12835105   16647
```

Figure 21. Screen grab of One-Way ANOVA results using R

```
> TukeyHSD(ShippingMode_Anova)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Energy_Consumption_Joules ~ as.factor(Shipping_Mode), data = USSupplyChain)

$`as.factor(Shipping_Mode)`
               diff       lwr      upr     p adj
Rail-Air  10.767377 -15.86026 37.39501 0.6090516
Road-Air   9.521406 -17.07981 36.12262 0.6779264
Road-Rail -1.245972 -28.05222 25.56028 0.9934536
```

Figure 22. Tukey HSD post-hoc test comparing all possible pairs of Shipping Mode

Appendix N: Screen grabs pertaining to Two- Way ANOVA Test

```
> Air$log10_Energy_Consumption_Joules<- log(Air$Energy_Consumption_Joules, base = 10)
> Rail$log10_Energy_Consumption_Joules<- log(Rail$Energy_Consumption_Joules, base = 10)
> Road$log10_Energy_Consumption_Joules<- log(Road$Energy_Consumption_Joules, base = 10)
> shapiro.test(Air$log10_Energy_Consumption_Joules)

        Shapiro-Wilk normality test

data:  Air$log10_Energy_Consumption_Joules
W = 0.88937, p-value = 2.205e-06

> shapiro.test(Rail$log10_Energy_Consumption_Joules)

        Shapiro-Wilk normality test

data:  Rail$log10_Energy_Consumption_Joules
W = 0.90877, p-value = 7.521e-06

> shapiro.test(Road$log10_Energy_Consumption_Joules)

        Shapiro-Wilk normality test

data:  Road$log10_Energy_Consumption_Joules
W = 0.91057, p-value = 9.182e-06
```

Figure 23. Shapiro test for shipping mode and disruption severity after logarithmic transformation

```
> leveneTest(Energy_Consumption_Joules ~ Shipping_Mode * Disruption_Severity, data = df)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   5  1.3801  0.232
      266
```

Figure 24. Homogeneity of Variance test – Levene's Test for Two-way Anova

```
> model <- aov(Energy_Consumption_Joules ~ Shipping_Mode * Disruption_Severity, data = df)
> summary(model)
                                   Df  Sum Sq Mean Sq F value Pr(>F)
Shipping_Mode                       2    1092     546   0.032  0.968
Disruption_Severity                 1    5783    5783   0.341  0.559
Shipping_Mode:Disruption_Severity   2   41361   20681   1.221  0.297
Residuals                         266 4505069   16936
```

Figure 25. Two Way Anova test

```
> tukey_results <- TukeyHSD(model)
> tukey_results
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Energy_Consumption_Joules ~ Shipping_Mode * Disruption_Severity, data = df)

$Shipping_Mode
                 diff       lwr      upr     p adj
Rail-Air   4.33622656 -41.55066 50.22311 0.9730270
Road-Air   4.28128032 -41.60561 50.16817 0.9736971
Road-Rail -0.05494624 -45.03566 44.92576 0.9999954

$Disruption_Severity
              diff       lwr      upr     p adj
Low-High -9.878595 -43.18881 23.43162 0.5597754
```

Figure 26. Post Hoc Tukey- test for Two Way Anova

Appendix O: Screen grabs pertaining to Chi- Square

Actual Values:
27 30 50 103
18 25 58 89
20 33 47 104
21 31 43 95
28 33 50 95

Expected Values:
23.94 31.92 52.08 102.06
21.66 28.88 47.12 92.34
23.256 31.008 50.592 99.144
21.66 28.88 47.12 92.34
23.484 31.312 51.088 100.116

Chi-Squared Values:
0.391128 0.115489 0.0830722 0.00865765
0.618449 0.521274 2.51219 0.12081
0.455862 0.127969 0.25503 0.237843
0.0201108 0.155623 0.360238 0.0766255
0.868432 0.0909985 0.0231707 0.261431

Chi-Square = 7.3044

Degrees of Freedom = 12

p = 0.836859

Figure 27. Results of Chi-Square Test

Appendix P: Screen grabs pertaining to Linear Regression Model

```
Call:
lm(formula = Order_Value_USD ~ Quantity_Ordered, data = US_Supply_Chain)

Residuals:
     Min       1Q   Median       3Q      Max 
-25055.4 -12204.9    445.8  12207.7  25008.4 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      24756.429    899.941  27.509   <2e-16 ***
Quantity_Ordered     1.058      1.540   0.687    0.492    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14380 on 998 degrees of freedom
Multiple R-squared:  0.000473,	Adjusted R-squared:  -0.0005285 
F-statistic: 0.4723 on 1 and 998 DF,  p-value: 0.4921
```

Figure 28. Summary of L-reg for Quantity Ordered and Order Value USD

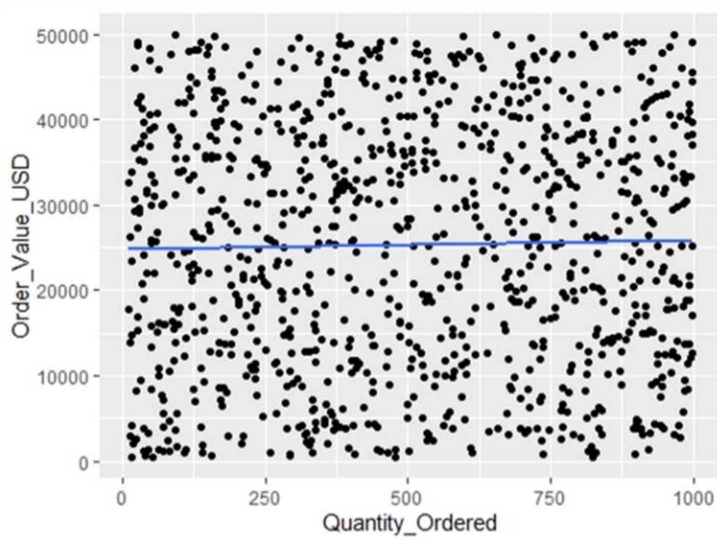


Figure 29. Plot graph for linearity between Quantity Ordered and Order Value USD

```
Call:
lm(formula = Order_Value_USD ~ Quantity_Ordered + Historical_Disruption_Count +
    Supplier_Reliability_Score + Communication_Cost_MB, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-26391.2 -12221.7    308.9  12216.5  25364.0

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 28313.534   2802.830  10.102   <2e-16 ***
Quantity_Ordered                1.222      1.546   0.791    0.429
Historical_Disruption_Count   -22.058     80.504  -0.274    0.784
Supplier_Reliability_Score  -2767.892   3182.590  -0.870    0.385
Communication_Cost_MB        -490.660    358.139  -1.370    0.171
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14390 on 995 degrees of freedom
Multiple R-squared:  0.003247,	Adjusted R-squared:  -0.0007599
F-statistic: 0.8104 on 4 and 995 DF,  p-value: 0.5186
```

Figure 30. Multi Linear Regression Output from R

```
vif(model)
       Quantity_Ordered Historical_Disruption_Count  Supplier_Reliability_Score       Communication_Cost_MB
               1.007262                    1.001034                    1.000726                    1.006873
```

Figure 31. VIF