Sales Forecasting on DataCo Smart Supply Chain

Project Report Group No: 12

Group Members: Abilash Sivakumar, Soham Palnitkar, Venkata Gattamaneni, Yash Jaigude

Executive Summary:

The DataCo Smart Supply Chain represents a complex network integrating a multitude of data streams, including product categories, orders, inventory levels, customer demographics, and sales forecasting metrics. This project aims to address critical inefficiencies observed within the system, which have led to increased costs, operational disruptions, and a decline in customer satisfaction. Through comprehensive data analysis employing multivariate regression techniques, our objective is to identify key factors impacting the efficiency of the DataCo Smart Supply Chain. By examining the interplay between historical sales data, procurement lead times, order quantities, customer region, and product category, we intend to uncover actionable insights that can streamline operations. The proposed methodology includes rigorous data preprocessing, exploratory data analysis (EDA), and regression modeling to ensure the robustness of our findings. Our report concludes with strategic recommendations intended to enhance the supply chain's efficiency, thus fostering improved business outcomes for DataCo.

Introduction:

In today's global economy, the efficacy of a supply chain is a significant determinant of a company's success. The DataCo Smart Supply Chain is an embodiment of modern supply chain complexities, encompassing an extensive array of interconnected data points and processes essential for timely and effective product delivery. However, recent trends have indicated that the DataCo Smart Supply Chain is grappling with suboptimal sales forecasting accuracy, leading to a cascade of negative repercussions ranging from stock shortages to customer dissatisfaction.

Understanding and optimizing the underlying factors contributing to these inefficiencies is more than an operational necessity; it is a strategic imperative. This project is borne out of the necessity to dissect and reconstruct the forecasting mechanisms of the DataCo Smart Supply Chain. By embarking on this analysis, we seek to not only rectify the current inefficiencies but also to lay down a scalable and sustainable framework for future operations. The strategic significance of this endeavor cannot be overstated, as it holds the potential to transform data-driven insights into a competitive advantage for DataCo.

Problem Description and Objectives:

The DataCo Smart Supply Chain is a complex system that encompasses various interconnected data, including product categories, orders, inventory, customers, and Sales forecasting. Inefficiencies like incorrect demand forecasting within this system can lead to increased costs, operational disruptions, and customer dissatisfaction. This project aims to analyze the DataCo Smart Supply Chain data to identify keys. Factors affecting its performance and propose improvements to enhance its overall efficiency.

Why It Is of Interest:

The efficiency and effectiveness of supply chain operations have a significant impact on an organization’s profitability and customer satisfaction. As a data analyst, we are passionate about optimizing processes and leveraging data to improve business outcomes in Supply Chain.

Conjectures to Investigate:

The impact of procurement lead times, order quantities, and Customer region, Product Category on Categorical Sale, and Categorical Demand. Investigating the relationship between historical sales data, seasonal trends, and demand forecasting accuracy.

Methods of Analysis:

The proposed analysis will involve multivariate regression techniques, including linear and potentially nonlinear models, to investigate the relationships between various supply chain variables and performance metrics. The analysis will encompass the following steps:

Data Source and Data Dictionary:

Gathered historical data related to procurement, inventory, demand, and logistics within the DataCo Smart Supply Chain.

- We used Origin – Kaggle for data sets and Company - DataCo Supply Chain for project and we have raw data - 180519 rows & 53 columns and this data period is Covered - May 2016 to Jan 2018
- We have Categorical Data, Demographic Data of Customers, Geographical Data of Customers, Sales Orders, Product Categories, Customer Segment, Purchase Order and Shipping Data.
  The following figure shows the data dictionary and sample of few data columns and rows.

FIELDS	DESCRIPTION
Type	: Type of transaction made
Days for shipping (real)	: Actual shipping days of the purchased product
Days for shipment (scheduled)	: Days of scheduled delivery of the purchased product
Benefit per order	: Earnings per order placed
Sales per customer	: Total sales per customer made per customer
Delivery Status	: Delivery status of orders: Advance shipping , Late delivery , Shipping canceled , Shipping on time
Late_delivery_risk	: Categorical variable that indicates if sending is late (1), it is not late (0).
Category Id	: Product category code
Category Name	: Description of the product category
Customer City	: City where the customer made the purchase
Customer Country	: Country where the customer made the purchase
Customer Email	: Customer's email
Customer Fname	: Customer name
Customer Id	: Customer ID
Customer Lname	: Customer lastname
Customer Password	: Masked customer key
Customer Segment	: Types of Customers: Consumer , Corporate , Home Office
Customer State	: State to which the store where the purchase is registered belongs
Customer Street	: Street to which the store where the purchase is registered belongs
Customer Zipcode	: Customer Zipcode
Department Id	: Department code of store
Department Name	: Department name of store
Latitude	: Latitude corresponding to location of store
Longitude	: Longitude corresponding to location of store

Order City	: Destination city of the order
Order Country	: Destination country of the order
Order Customer Id	: Customer order code
order date (DateOrders)	: Date on which the order is made
Order Id	: Order code
Order Item Cardprod Id	: Product code generated through the RFID reader
Order Item Discount	: Order item discount value
Order Item Discount Rate	: Order item discount percentage
Order Item Id	: Order item code
Order Item Product Price	: Price of products without discount
Order Item Profit Ratio	: Order Item Profit Ratio
Order Item Quantity	: Number of products per order
Sales	: Value in sales
Order Item Total	: Total amount per order
Order Profit Per Order	: Order Profit Per Order
Order Region	: Region of the world where the order is delivered : Southeast Asia ,South Asia ,Oceania ,Eastern Asia, West Asia , West of USA , US Center , West Africa, Central Africa ,North Africa ,Western Europe ,Northern , Caribbea
Order State	: State of the region where the order is delivered
Order Status	: Order Status : COMPLETE , PENDING , CLOSED , PENDING_PAYMENT ,CANCELED , PROCESSING ,SUSPECTED_FRAUD ,ON_HOLD ,PAYMENT_REVIEW
Product Card Id	: Product code
Product Category Id	: Product category code
Product Description	: Product Description
Product Image	: Link of visit and purchase of the product
Product Name	: Product Name
Product Price	: Product Price
Product Status	: Status of the product stock :If it is 1 not available , 0 the product is available
Shipping date (DateOrders)	: Exact date and time of shipment
Shipping Mode	: The following shipping modes are presented : Standard Class , First Class , Second Class , Same Day

Type	Days for shipping (real)	Days for shipment (scheduled)	Benefit per order	Sales per customer	Delivery Status	Late_delivery_risk	Category Id	Category Name	Customer City	Customer Country	Customer En
DEBIT	3	4	91.25	314.6400146	Advance shipping	0	73	Sporting Goods	Caguas	Puerto Rico	XXXXXXXXX
TRANSFER	5	4	-249.0899963	311.3599854	Late delivery	1	73	Sporting Goods	Caguas	Puerto Rico	XXXXXXXXX
CASH	4	4	-247.7799988	309.7200012	Shipping on time	0	73	Sporting Goods	San Jose	EE. UU.	XXXXXXXXX
DEBIT	3	4	22.86000061	304.8099976	Advance shipping	0	73	Sporting Goods	Los Angeles	EE. UU.	XXXXXXXXX
PAYMENT	2	4	134.2100067	298.25	Advance shipping	0	73	Sporting Goods	Caguas	Puerto Rico	XXXXXXXXX
TRANSFER	6	4	18.57999992	294.980011	Shipping canceled	0	73	Sporting Goods	Tonawanda	EE. UU.	XXXXXXXXX
DEBIT	2	1	95.18000031	288.4200134	Late delivery	1	73	Sporting Goods	Caguas	Puerto Rico	XXXXXXXXX
TRANSFER	2	1	68.43000031	285.1400146	Late delivery	1	73	Sporting Goods	Miami	EE. UU.	XXXXXXXXX
CASH	3	2	133.7200012	278.5899963	Late delivery	1	73	Sporting Goods	Caguas	Puerto Rico	XXXXXXXXX
CASH	2	1	132.1499939	275.3099976	Late delivery	1	73	Sporting Goods	San Ramon	EE. UU.	XXXXXXXXX
TRANSFER	6	2	130.5800018	272.0299988	Shipping canceled	0	73	Sporting Goods	Caguas	Puerto Rico	XXXXXXXXX
TRANSFER	5	2	45.68999863	268.7600098	Late delivery	1	73	Sporting Goods	Freeport	EE. UU.	XXXXXXXXX
TRANSFER	4	2	21.76000023	262.2000122	Late delivery	1	73	Sporting Goods	Salinas	EE. UU.	XXXXXXXXX
DEBIT	2	1	24.57999992	245.8099976	Late delivery	1	73	Sporting Goods	Caguas	Puerto Rico	XXXXXXXXX

Data Preprocessing:

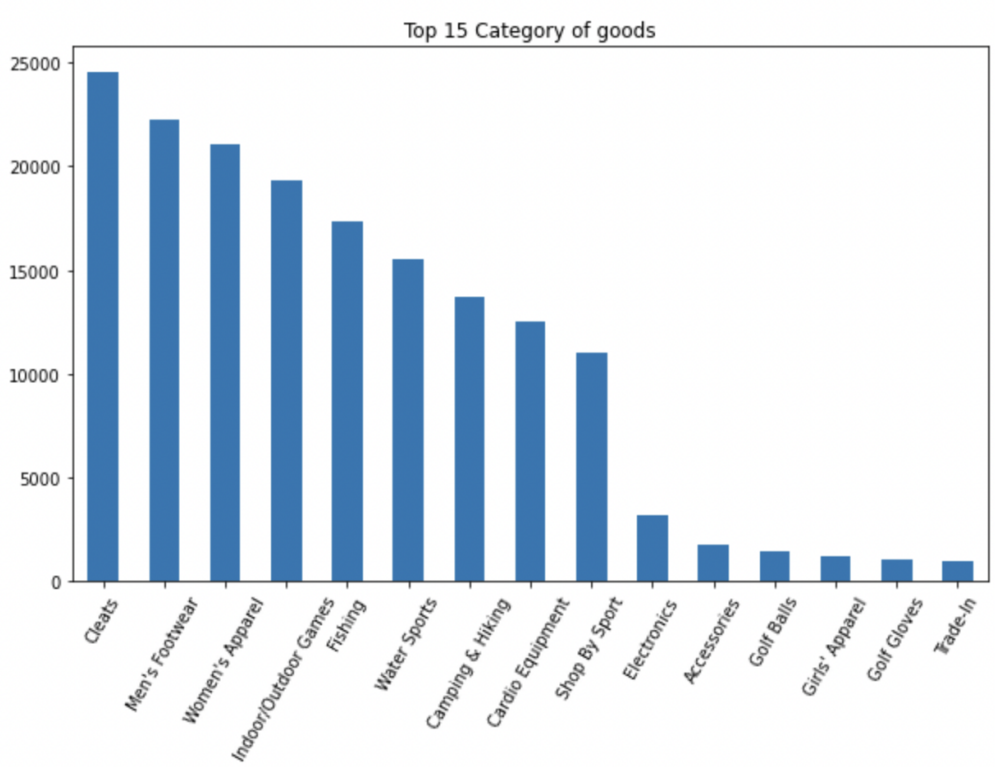
Clean and prepare the data, addressing missing values and outliers.
Exploratory Data Analysis (EDA): Before performing EDA, We cleaned and prepared the data, addressing missing values and outliers. Explore the data to identify patterns, trends, and potential relationships between variables.

- Identified and addressed missing or null values within the dataset to ensure data integrity and completeness for analysis.
- Conducted univariate analysis, examining individual variables' distributions and characteristics to understand their behavior and outliers.

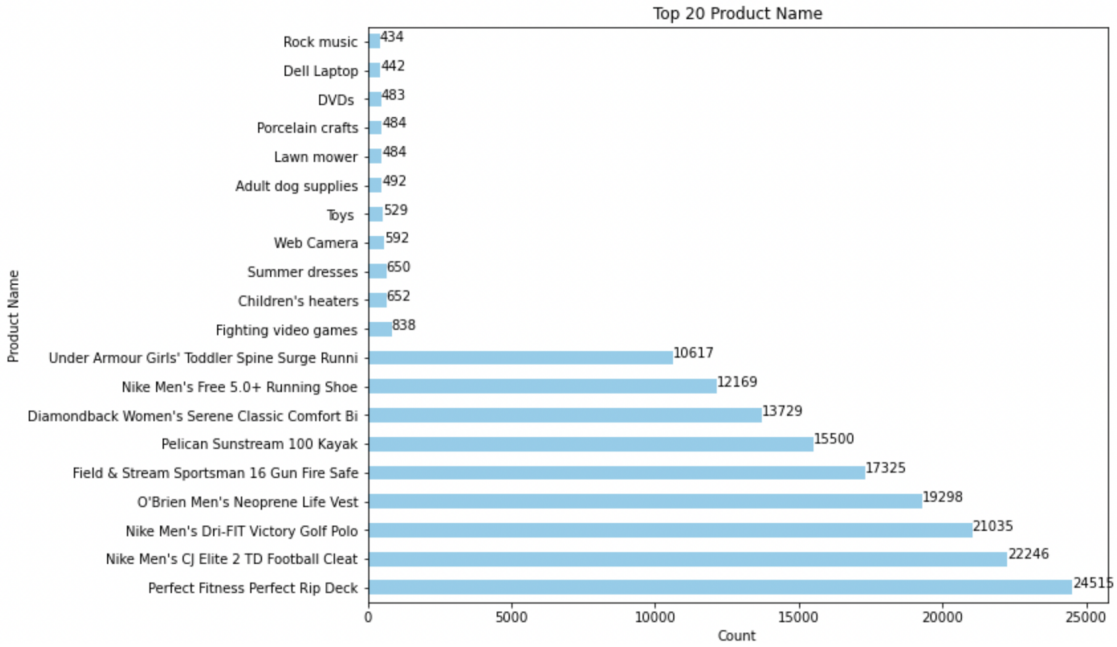
Multivariate Regression Modeling: Develop multiple regression models to examine the impact of

different factors on supply chain demand forecast and performance metrics.

Model Validation: Assess the model's accuracy and validity using appropriate statistical tests and measures.
Analysis and results:

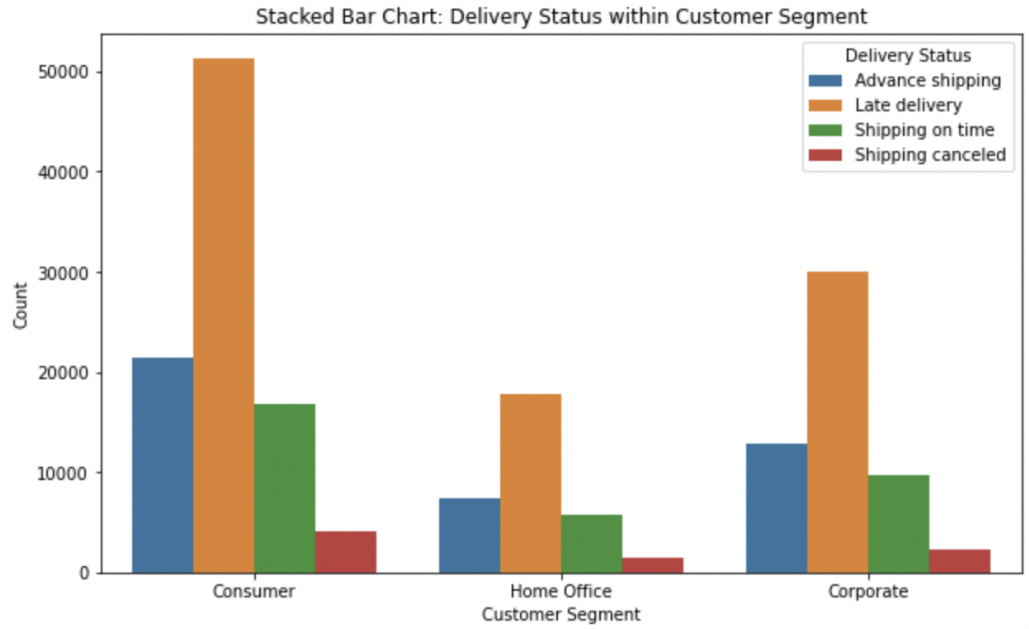


Graph Insights: The bar chart indicates that cleats and men's footwear are the most popular goods categories, with a steep decline to niche items like golf gloves and trade-ins. For data preparation, one would focus on cleaning and normalizing the data, while feature engineering might involve creating new features such as seasonality indicators or combining related categories for analysis.

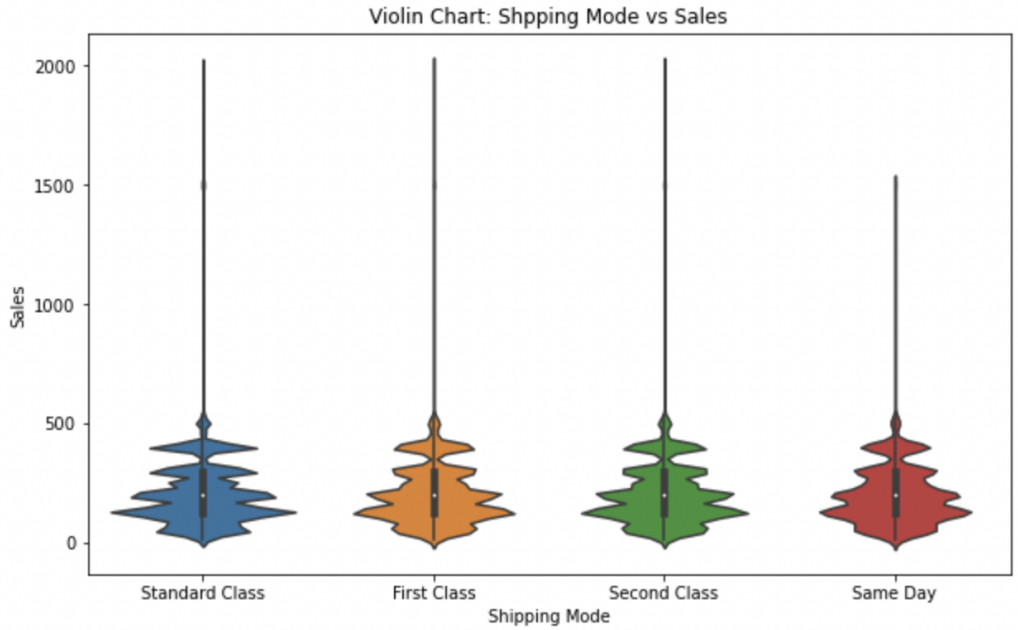


Graph Insights: The horizontal bar chart displays the top 20 products by count. The product variety is broad, ranging from fitness equipment like the "Perfect Fitness Perfect Rip Deck" to electronics such as "Dell Laptop." The counts vary significantly, with the most popular item being the "Perfect Fitness Perfect Rip Deck" with 24,515 units, suggesting a high demand in fitness or a successful marketing campaign for this product. Conversely, items like "Rock music" and "Dell Laptop" have much lower counts, indicating less frequency in sales or stock.

For data preparation, one would ensure the product names are uniformly formatted and check for duplicates. Feature engineering could involve categorizing these products into broader groups to identify trends across categories or creating features that capture the seasonality of purchases.



Graph Insights: The chart shows that Corporate has the highest delivery counts, predominantly on time, while Consumer and Home Office have more varied delivery statuses. Data preparation would include segmenting delivery status, and feature engineering might involve creating time-to-delivery metrics.



Graph Insights: The violin chart suggests Standard Class is the most common shipping mode, with a wide distribution of sales, while Same Day shipping shows higher sales but less frequency. Data preparation would involve sales normalization, and feature engineering could create shipping mode efficiency ratios.

Graph Insights: The scatter plot indicates a positive correlation between sales and product price discount, with dense clustering at lower discounts. Data preparation would involve outlier detection, and feature engineering could create a new variable representing discount tiers.

DATA PREPARATION AND FEATURE ENGINEERING:

Data Preparation:

- **Column Removal:** Eliminated unnecessary columns that aren't pertinent for modeling, focusing on features independent of the target variable and dropping highly correlated columns to prevent multicollinearity.
- **Handling Missing Values:** Addressed columns with a significant number of missing values by either imputing missing data or removing columns with excessive missing values for improved model performance.

Feature Engineering:

- **Label Encoding:** Transformed categorical variables into numerical format using label encoding to represent categories as integer values, allowing algorithms to interpret and process categorical data.
- **One-Hot Encoding:** Converted categorical variables into a binary format using one-hot encoding, creating binary columns to represent different categories within a feature.
- **Standardization:** Standardized numerical features to a common scale to ensure uniformity and avoid bias towards variables with larger magnitudes, enabling a fair comparison between different features.

```
Normalized Data:
```

	Days for shipping (real)	Days for shipment (scheduled)	Benefit per order	Delivery Status	Late_delivery_risk	Order Item Discount	Order Item Discount Rate	Order Item Profit Ratio	Order Item Quantity	Sales	...	Product Name_Under Armour Women's Ignite Slide	Product Name_Under Armour Women's Micro G Skulpt Running S	Product Name_Web Camera
0	0.500000	1.0	0.841800	0	0.0	0.02622	0.16	0.935385	0.0	0.159678	...	0	0	0
1	0.833333	1.0	0.776183	1	1.0	0.03278	0.20	0.600000	0.0	0.159678	...	0	0	0
2	0.666667	1.0	0.776435	3	0.0	0.03606	0.24	0.600000	0.0	0.159678	...	0	0	0
3	0.500000	1.0	0.828614	0	0.0	0.04588	0.28	0.870769	0.0	0.159678	...	0	0	0
4	0.333333	1.0	0.850082	0	0.0	0.05900	0.36	0.984615	0.0	0.159678	...	0	0	0

5 rows × 220 columns

MODEL BUILDING AND PERFORMANCE EVALUATION:

Model Building:

- Implemented various regression models including Linear Regression, Ridge Regression, Decision Tree, and Random Forest Regressor to explore diverse modeling approaches.
- Employed Linear Regression for establishing linear relationships, Ridge Regression for regularization, and Decision Tree & Random Forest for non-linear modeling, expanding model diversity.

Data Analysis and Interpretation:

ANOVA Results:

```
349  anova(model1_LRM)
350
351  # Extracting feature weights
352  feature_weights_LRM <- coef(model1_LRM)
353
354  # Printing the feature weights
349:18  (Top Level)
```

```
Console   Terminal   Background Jobs
R 4.2.2 · ~/Documents/MS NEU/IE 7280 Stat Methods Xie/Project I/Dataset/
Response: y_train
                          Df Sum Sq Mean Sq    F value  Pr(>F)
Days.for.shipping..real.   1   0.00   0.003 5.5352e+09 < 2e-16 ***
Category.Name             49 440.30   8.986 1.5776e+13 < 2e-16 ***
Customer.Segment           2   0.00   0.002 3.0365e+09 < 2e-16 ***
Order.Item.Discount        1  20.79  20.789 3.6500e+13 < 2e-16 ***
Order.Item.Discount.Rate   1  31.93  31.933 5.6066e+13 < 2e-16 ***
Order.Item.Profit.Ratio    1   0.00   0.002 3.4942e+09 < 2e-16 ***
Order.Item.Quantity        1  66.80  66.799 1.1728e+14 < 2e-16 ***
Order.Item.Total           1  31.48  31.477 5.5264e+13 < 2e-16 ***
Order.Region              22   0.00   0.000 4.2840e-01 0.99081
Order.Status               8   0.00   0.000 2.0562e+00 0.03639 *
Product.Price              1   0.00   0.000 2.5292e+00 0.11176
Shipping.Mode              3   0.00   0.000 4.4590e-01 0.72017
Residuals             135299   0.00   0.000
```

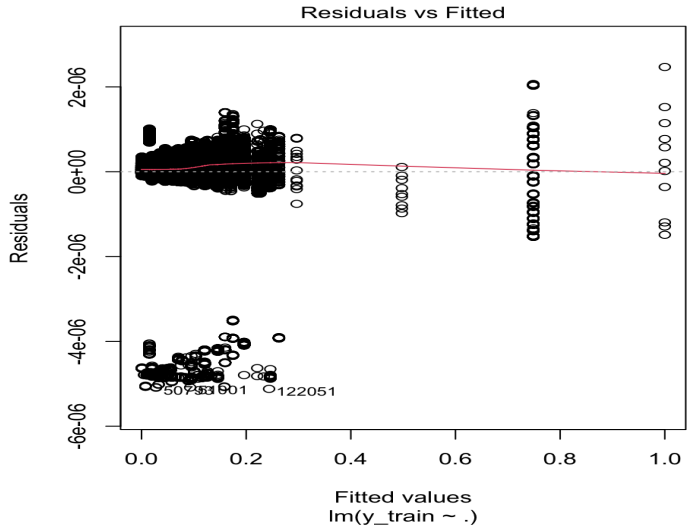
Customer.Category, Order.Item.Discount, Order.Item.Discount.Rate, Order.Item.Profit.Ratio, Order.Item.Quantity, Order.Item.Total:
All these variables have extremely high F-values and p-values less than 2e-16, indicating their strong statistical significance in predicting Sales.

Order.Region, Order.Status, Product.Price, Shipping.Mode:
These variables have higher p-values (0.99081, 0.03639, 0.11176, 0.72017, respectively), suggesting that they are not statistically significant predictors of Sales. The F-values are low for Order.Region and Shipping.Mode, further supporting this observation.

Residuals: The Residuals section represents the unexplained variation in the response variable after accounting for the predictors. The fact that the sum of squares for residuals is zero indicates a perfect fit in the model.

Variation in Residuals:

The spread or variation in residuals is crucial to assess model performance. If the residuals are spread relatively evenly around the zero line, it indicates that the model's predictions are consistent across different levels of the predictor variables. In your case, with symmetric residuals, the variation appears to be relatively constant across the range of fitted values. In summary, the described characteristics of your residual plot suggest a reasonably well-behaved model with symmetric and evenly spread residuals.



Hypothesis Testing:

1. Is the Product Price a significant factor in determining sales?

- Null Hypothesis (H0): Beta = 0 (There is no effect of Product Price on Sales)
- Alternative Hypothesis (H1): Beta ≠ 0 (There is an effect of Product Price on Sales)

Given the p-value of 0.116, and assuming a significance level (alpha) of 0.05, you make a decision based on the p-value.

- Decision Rule: If p-value < alpha, reject the null hypothesis.
- Interpretation: Since p-value (0.116) is greater than alpha (0.05), you do not have enough evidence to reject the null hypothesis.
- Conclusion: Fail to reject the null hypothesis.

Therefore, based on your analysis, you don't have sufficient evidence to claim that Product Price has a significant effect on Sales. However, be cautious about interpreting this result. If the p-value is close to the significance level, you may want to consider the practical significance and conduct a more detailed analysis.

2. Does the discount rate have any significant effect on Sales?

   Null Hypothesis (H0): Beta = 0 (There is no effect of Discount Rate on Sales)

   Alternative Hypothesis (H1): Beta ≠ 0 (There is an effect of Discount Rate on Sales)

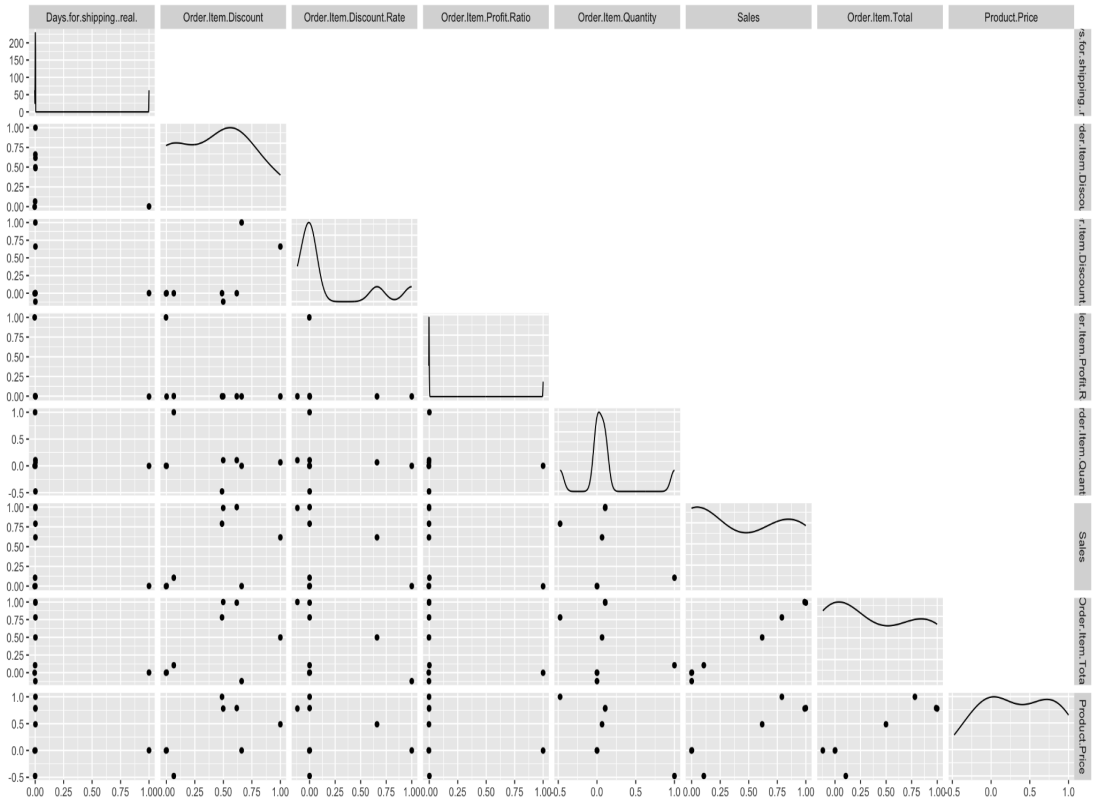
Given the p-value of 2e-16 (a very small value), and assuming a significance level (alpha) of 0.05, you make a decision based on the p-value.

- Decision Rule: If p-value < alpha, reject the null hypothesis.
- Interpretation: Since the p-value (2e-16) is much smaller than alpha (0.05), you have enough evidence to reject the null hypothesis.

- Conclusion: Reject the null hypothesis.

Therefore, based on your analysis, you conclude that Discount Rate has a significant effect on Sales. The p-value being extremely small suggests a very high level of statistical significance, supporting the idea that the effect of Discount Rate on Sales is not likely due to random chance.

3. Is there any correlation between Order Discount Rate and Profit Ratio?



From the Scatter Plot Matrix, there is no correlation between Order Discount Rate and Profit Ratio.

4. Are Days for Shipping Real significant in predicting Sales ?

- Null Hypothesis (Ho): Beta =0 (Days for Shipping is not significant in predicting Sales)
- Alternative Hypothesis (Ha): Beta ≠0 (Days for Shipping is significant in predicting Sales)

If the p-value from the ANOVA table is less than the chosen significance level (alpha), which is 0.05 in this case, you reject the null hypothesis. In your case, you mentioned that p=2×10−16 which is significantly smaller than 0.05.

So, you correctly reject the null hypothesis. Therefore, based on the given p-value, you have evidence to conclude that Days for Shipping is significant in predicting Sales.

Performance Evaluation:

- Utilized regression evaluation metrics like RMSE, R2, and MAE to assess model accuracy and fit.

- Conducted a comparative analysis of multiple models, leveraging evaluation metrics to select the best-performing model for predicting the target variable.
- The Linear Model demonstrates exceptional performance with considerably lower error metrics (MAE, MSE, RMSE) compared to the Decision Tree Model.
- The Linear Model achieves near-perfect accuracy with an R-squared value of 1.0, indicating an exact fit to the data, while the Decision Tree Model shows high accuracy (R2 = 0.999) but with slightly larger errors. Also, RMSE score is less for Linear Regression.
- Overall, the Linear Model appears to outperform the Decision Tree Model in terms of accuracy and precision in predicting the target variable based on the provided scores.

Model	R-squared (R2)	Adjusted R2	RMSE	MAE
Multiple Linear Regression Model	0.999	0.969	0.0115	0.0051
Decision Tree	0.966	0.92	0.012	0.009
Ridge Regression	0.895	0.86	0.054	0.039

INTERPRETATION AND SIGNIFICANCE OF RESULTS:

- **Make Informed Decisions:** Understanding influential factors behind sales forecasts aids in strategic planning, resource allocation, and inventory management.
- **Optimize Strategies:** Insights derived from the model allow stakeholders to tailor marketing campaigns, adjust pricing strategies, and enhance product availability to meet forecasted demand.

Hence, stakeholders should use these findings as guidance rather than absolute predictions, complementing them with market intelligence and domain expertise for nuanced decision-making.

Shortcomings of the Study:

Limited Data Granularity:

If the data collection procedure lacks detailed information on certain relevant factors such as customer behavior, market trends, or external economic factors, the forecasting model may not fully capture the complexity of the sales environment. This limitation could lead to oversimplified predictions and reduced accuracy in forecasting sales.

Sampling Bias in Data Collection:

Another potential shortcoming could arise from sampling bias in the data collection process. If the data collected primarily represents a specific subset of customers, regions, or product categories, it may not be reflective of the entire customer market dynamics. This bias could result in a skewed model that fails to generalize well to broader scenarios, leading to inaccurate sales forecasts for segments not adequately represented in the dataset.

Assumption of Stationarity:

Sales forecasting models often assume that the underlying patterns in the data remain stable over time. If the data collection procedure does not account for changes in market conditions, consumer preferences, or other external factors, the model's assumptions may be violated. This lack of consideration for non-stationarity can introduce errors into the forecasts, especially in dynamic environments where trends and patterns evolve over time.

NEXT STEPS FOR IMPACT:

- **Fine-Tuning Models:** Continuously refine and optimize the supervised learning models based on performance evaluations and feedback loops.
- **Incorporate External Factors:** Expand the models to include external factors (e.g., market trends, economic indicators)
- **Real-Time Integration:** Implement real-time data integration for immediate adaptation to changing market conditions.
- **Continuous Monitoring:** Set up a system for continuous monitoring of forecasting accuracy and model performance.
- **Explore Advanced Models:** Investigate the applicability of advanced machine learning models, such as ensemble methods or neural networks.

Interpretation and Recommendations: Provide insights and recommendations based on the analysis results to enhance the DataCo Smart Supply Chain's efficiency.