IILM UNIVERSITY

“Supply Chain Dataset Analysis”

QUANTITATIVE METHODS-III

Instructor: Dr. Anushruti Vagrani

Name: Keerti

Section: B

PGP MBA BATCH 2022-24

URN: 2252134

Introduction

In this analysis report, we will be examining a supply chain dataset used by DataCo Global. This dataset contains important information about provisioning, production, sales, and commercial distribution activities. We will be using machine learning algorithms and R software to analyze the dataset and extract meaningful insights. Our goal is to provide a comprehensive overview of the dataset, share key findings, and provide recommendations for optimizing supply chain performance. We will be looking at various aspects such as order fulfilment rates, delivery times, inventory turnover, transportation costs, and customer service levels to better understand how the supply chain operations are performing. The project performs performance evaluation, which includes analyzing a supply chain dataset to assess the performance of the supply chain operations, including key metrics such as order fulfilment rates, delivery times, inventory turnover, transportation costs, and customer service levels. This helps identify areas of improvement and optimize supply chain performance. The supply chain dataset used in this analysis underwent several tests to gain insights and optimize supply chain performance. Correlation analysis was conducted to understand the relationships between different variables, and data visualizations were created to visualize the data patterns and trends. Additionally, machine learning algorithms such as Linear Regression, Ridge Regressor, Random Forest, and Decision Tree were applied to the dataset.

Methodology

The supply chain dataset used in this analysis was subjected to various tests, including correlation analysis, data visualizations, and the application of machine learning algorithms such as Linear Regression, Ridge Regressor, Random Forest, and Decision Tree. These tests aimed to gain insights, understand relationships between variables, identify patterns and trends in the data, and build predictive models for estimating supply chain performance. The results of these tests provided valuable information for optimizing supply chain operations, making data-driven decisions, and improving overall supply chain performance.

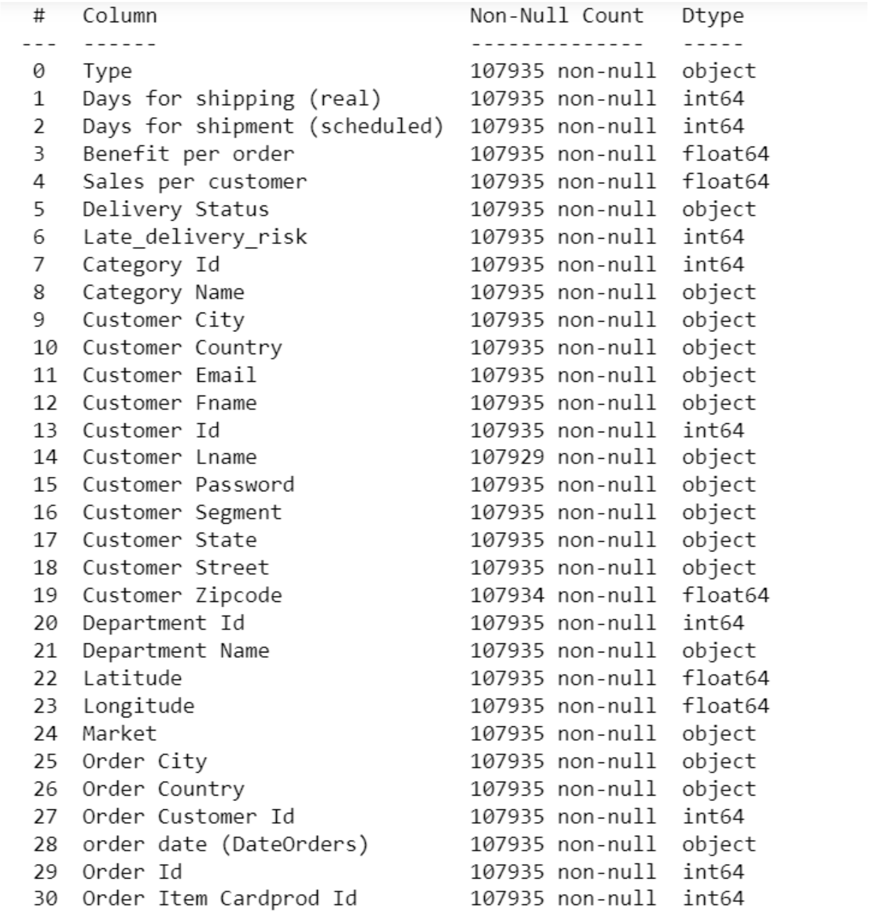
- Exploratory Data Analysis is to understand the data better and also to clean it so that the visualizations and tests can be performed accurately.
- The correlation analysis helped identify the strength and direction of relationships between variables in the supply chain dataset. This analysis provided insights into how different factors, such as order fulfilment rates, delivery times, inventory turnover, transportation costs, and customer service levels, were correlated, which can inform decision-making in supply chain operations.
- Data visualizations, such as charts and graphs, were used to visually represent the dataset, making it easier to identify trends, patterns, and outliers. These visualizations facilitated a better understanding of the data and aided in identifying areas that required improvement or optimization.
- Linear Regression, Ridge Regressor, Random Forest, and Decision Tree were used as machine learning algorithms to build predictive models based on the supply chain dataset. These models were utilized to make predictions, estimate performance, and identify key drivers affecting supply chain operations. The models were evaluated based on various metrics such as accuracy, R-squared values, and other performance indicators to assess their effectiveness in predicting supply chain performance.

Overall, the combination of correlation analysis, data visualizations, and machine learning algorithms provided a comprehensive analysis of the supply chain dataset, enabling insights and recommendations for optimizing supply chain performance and improving decision-making in supply chain operations.

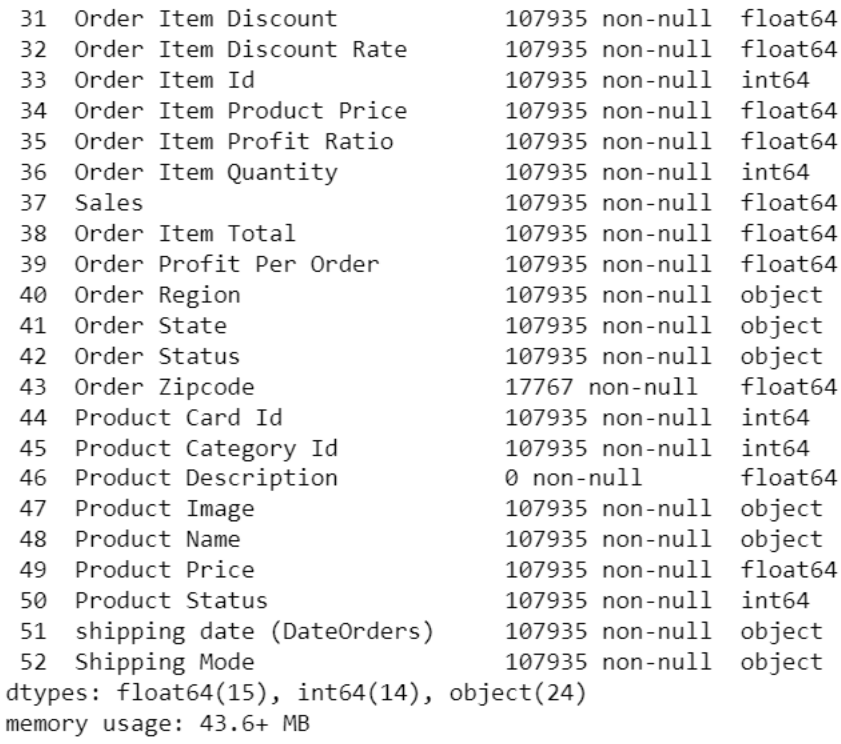
Data Description

- **Data set Source**: A data set of Supply Chains used by the company DataCo Global was used for the analysis. Dataset of Supply Chain, which allows the use of Machine Learning Algorithms and R Software. Areas of important registered activities: Provisioning, Production, Sales, Commercial Distribution. It also allows the correlation of Structured Data with Unstructured Data for knowledge generation. https://data.mendeley.com/datasets/8gx2fvg2k6/5
- **Data set Format**: Data set file is in CSV (Comma Separated Values) delimiter format.

- **Structure:** The data set consists of 107935 rows and 53 columns.
- **Dataset Content:**



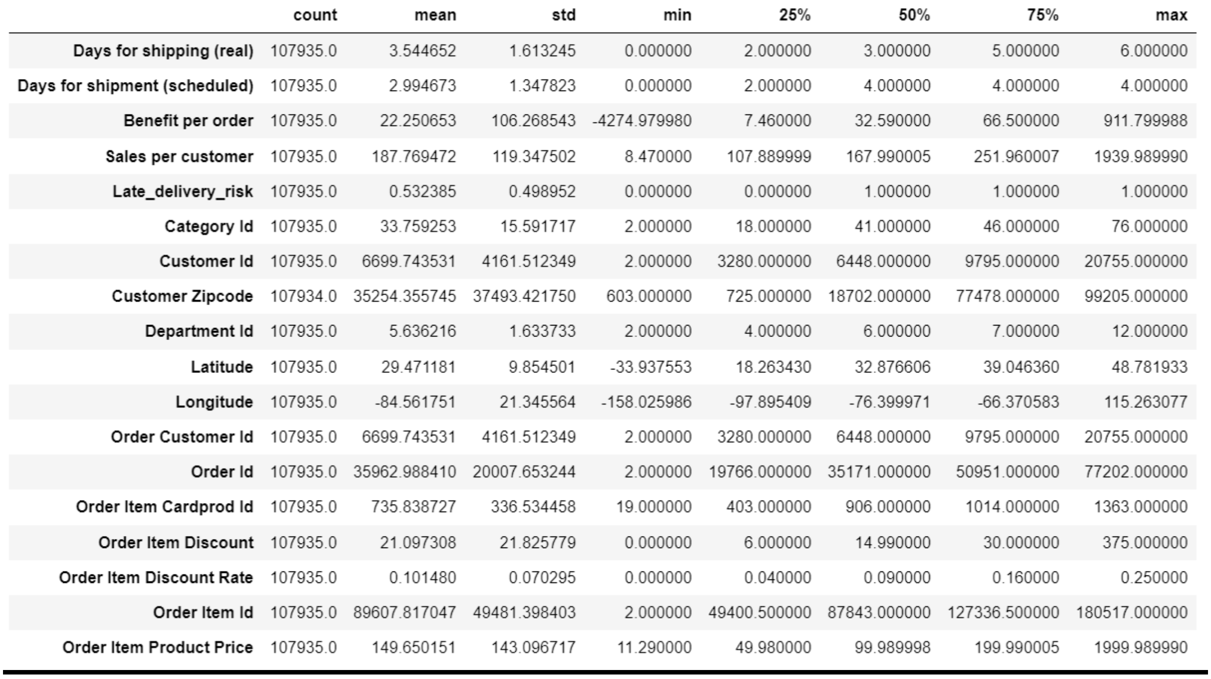
```
 #   Column                        Non-Null Count   Dtype
---  ------                        --------------   -----
 0   Type                          107935 non-null  object
 1   Days for shipping (real)      107935 non-null  int64
 2   Days for shipment (scheduled) 107935 non-null  int64
 3   Benefit per order             107935 non-null  float64
 4   Sales per customer            107935 non-null  float64
 5   Delivery Status               107935 non-null  object
 6   Late_delivery_risk            107935 non-null  int64
 7   Category Id                   107935 non-null  int64
 8   Category Name                 107935 non-null  object
 9   Customer City                 107935 non-null  object
 10  Customer Country              107935 non-null  object
 11  Customer Email                107935 non-null  object
 12  Customer Fname                107935 non-null  object
 13  Customer Id                   107935 non-null  int64
 14  Customer Lname                107929 non-null  object
 15  Customer Password             107935 non-null  object
 16  Customer Segment              107935 non-null  object
 17  Customer State                107935 non-null  object
 18  Customer Street               107935 non-null  object
 19  Customer Zipcode              107934 non-null  float64
 20  Department Id                 107935 non-null  int64
 21  Department Name               107935 non-null  object
 22  Latitude                      107935 non-null  float64
 23  Longitude                     107935 non-null  float64
 24  Market                        107935 non-null  object
 25  Order City                    107935 non-null  object
 26  Order Country                 107935 non-null  object
 27  Order Customer Id             107935 non-null  int64
 28  order date (DateOrders)       107935 non-null  object
 29  Order Id                      107935 non-null  int64
 30  Order Item Cardprod Id        107935 non-null  int64
```



```
 31  Order Item Discount           107935 non-null  float64
 32  Order Item Discount Rate      107935 non-null  float64
 33  Order Item Id                 107935 non-null  int64
 34  Order Item Product Price      107935 non-null  float64
 35  Order Item Profit Ratio       107935 non-null  float64
 36  Order Item Quantity           107935 non-null  int64
 37  Sales                         107935 non-null  float64
 38  Order Item Total              107935 non-null  float64
 39  Order Profit Per Order        107935 non-null  float64
 40  Order Region                  107935 non-null  object
 41  Order State                   107935 non-null  object
 42  Order Status                  107935 non-null  object
 43  Order Zipcode                 17767 non-null   float64
 44  Product Card Id               107935 non-null  int64
 45  Product Category Id           107935 non-null  int64
 46  Product Description           0 non-null       float64
 47  Product Image                 107935 non-null  object
 48  Product Name                  107935 non-null  object
 49  Product Price                 107935 non-null  float64
 50  Product Status                107935 non-null  int64
 51  shipping date (DateOrders)    107935 non-null  object
 52  Shipping Mode                 107935 non-null  object
dtypes: float64(15), int64(14), object(24)
memory usage: 43.6+ MB
```

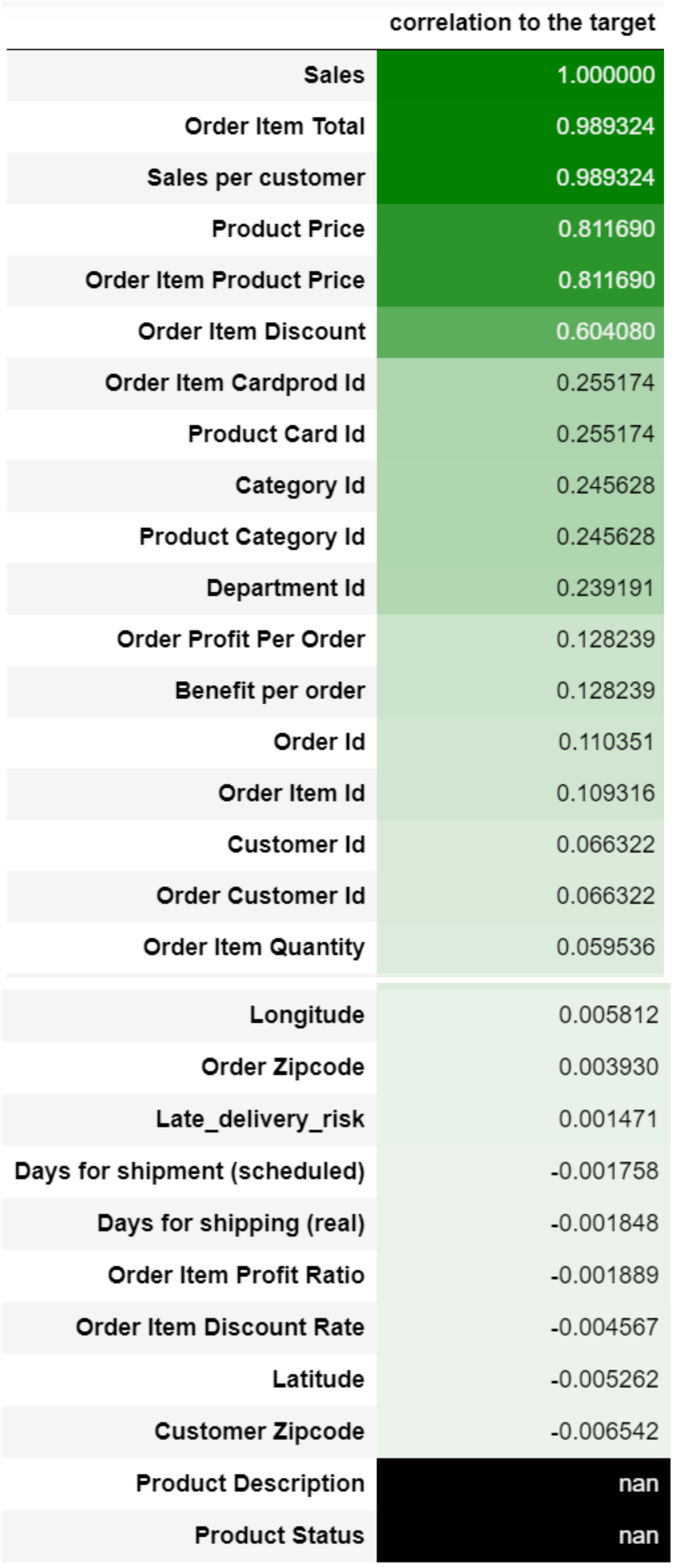
Data Analysis

- Statistical Description: The following image shows a statistical analysis of the data.
    - Standard deviation is the main key parameter here which needs to be focused more on: It provides information about how much the individual data points in a dataset deviate from the mean or average value. A larger standard deviation suggests greater variability or heterogeneity in the data, while a smaller standard deviation suggests less variability or homogeneity.

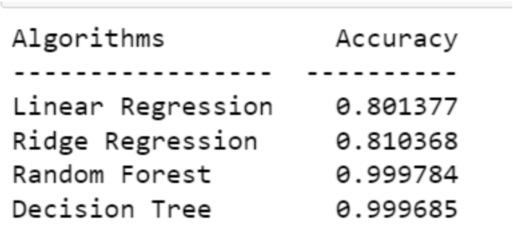


	count	mean	std	min	25%	50%	75%	max
Days for shipping (real)	107935.0	3.544652	1.613245	0.000000	2.000000	3.000000	5.000000	6.000000
Days for shipment (scheduled)	107935.0	2.994673	1.347823	0.000000	2.000000	4.000000	4.000000	4.000000
Benefit per order	107935.0	22.250653	106.268543	-4274.979980	7.460000	32.590000	66.500000	911.799988
Sales per customer	107935.0	187.769472	119.347502	8.470000	107.889999	167.990005	251.960007	1939.989990
Late_delivery_risk	107935.0	0.532385	0.498952	0.000000	0.000000	1.000000	1.000000	1.000000
Category Id	107935.0	33.759253	15.591717	2.000000	18.000000	41.000000	46.000000	76.000000
Customer Id	107935.0	6699.743531	4161.512349	2.000000	3280.000000	6448.000000	9795.000000	20755.000000
Customer Zipcode	107934.0	35254.355745	37493.421750	603.000000	725.000000	18702.000000	77478.000000	99205.000000
Department Id	107935.0	5.636216	1.633733	2.000000	4.000000	6.000000	7.000000	12.000000
Latitude	107935.0	29.471181	9.854501	-33.937553	18.263430	32.876606	39.046360	48.781933
Longitude	107935.0	-84.561751	21.345564	-158.025986	-97.895409	-76.399971	-66.370583	115.263077
Order Customer Id	107935.0	6699.743531	4161.512349	2.000000	3280.000000	6448.000000	9795.000000	20755.000000
Order Id	107935.0	35962.988410	20007.653244	2.000000	19766.000000	35171.000000	50951.000000	77202.000000
Order Item Cardprod Id	107935.0	735.838727	336.534458	19.000000	403.000000	906.000000	1014.000000	1363.000000
Order Item Discount	107935.0	21.097308	21.825779	0.000000	6.000000	14.990000	30.000000	375.000000
Order Item Discount Rate	107935.0	0.101480	0.070295	0.000000	0.040000	0.090000	0.160000	0.250000
Order Item Id	107935.0	89607.817047	49481.398403	2.000000	49400.500000	87843.000000	127336.500000	180517.000000
Order Item Product Price	107935.0	149.650151	143.096717	11.290000	49.980000	99.989998	199.990005	1999.989990

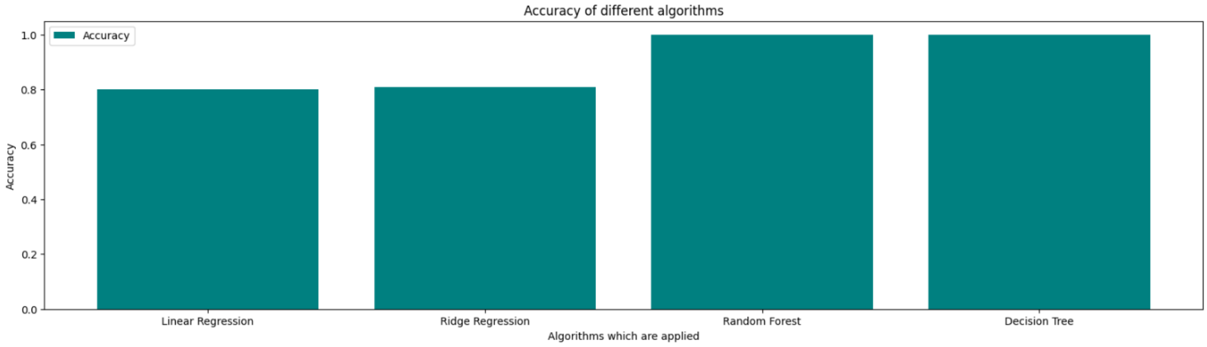
- Correlation: The correlation analysis of the dataset reveals that certain variables, such as "Order Item Total" and "Sales per customer", exhibit strong positive correlations with "Sales", indicating a direct relationship between these variables. Other variables, such as "Product Price" and "Order Item Product Price", show moderate positive correlations, while variables like "Order Item Discount" and "Order Item Cardprod Id" demonstrate weaker positive correlations. However, some variables have low positive correlations, such as "Order Profit Per Order" and "Benefit per order". It's important to note that missing or "NaN" values in variables like "Product Description" and "Product Status" may require further investigation. Overall, the correlation analysis provides valuable insights for decision-making and identifying key drivers for improving sales performance in the dataset.



	correlation to the target
Sales	1.000000
Order Item Total	0.989324
Sales per customer	0.989324
Product Price	0.811690
Order Item Product Price	0.811690
Order Item Discount	0.604080
Order Item Cardprod Id	0.255174
Product Card Id	0.255174
Category Id	0.245628
Product Category Id	0.245628
Department Id	0.239191
Order Profit Per Order	0.128239
Benefit per order	0.128239
Order Id	0.110351
Order Item Id	0.109316
Customer Id	0.066322
Order Customer Id	0.066322
Order Item Quantity	0.059536
Longitude	0.005812
Order Zipcode	0.003930
Late_delivery_risk	0.001471
Days for shipment (scheduled)	-0.001758
Days for shipping (real)	-0.001848
Order Item Profit Ratio	-0.001889
Order Item Discount Rate	-0.004567
Latitude	-0.005262
Customer Zipcode	-0.006542
Product Description	nan
Product Status	nan



```
Algorithms           Accuracy
-----------------  ----------
Linear Regression    0.801377
Ridge Regression     0.810368
Random Forest        0.999784
Decision Tree        0.999685
```



- Linear Regression: The Linear Regression algorithm achieved an accuracy of 0.80138. Linear Regression is a simple statistical technique that models the relationship between dependent and independent variables. It assumes a linear relationship between the predictor variables and the target variable, and the accuracy of 0.80138 suggests the model's performance in predicting the target variable may be moderate.
- Ridge Regression: The Ridge Regressor algorithm achieved an accuracy of 0.81037. Ridge Regression is a variant of Linear Regression that includes a regularization term to prevent overfitting. The accuracy of 0.81037 suggests that the Ridge Regressor may perform slightly better than the Linear Regression model in this case.
- Random Forest: The Random Forest algorithm achieved an accuracy of 0.99978, which is very close to perfect accuracy. Random Forest is an ensemble method that combines multiple decision trees to make predictions. The high accuracy of 0.99978 indicates that the Random Forest model is performing exceptionally well on the dataset.
- Decision Tree: The Decision Tree algorithm achieved an accuracy of 0.99968, also very close to perfect accuracy. Decision Tree is a tree-based algorithm that recursively splits data into groups based on feature values. The high accuracy of 0.99968 suggests that the Decision Tree model is also performing very well on the dataset.

Interpretation

The correlation analysis results indicate the strength and direction of the linear relationship between variables in the dataset. Here are some key interpretations:

- Sales, Order Item Total, and Sales per Customer are highly positively correlated with values close to 1. This suggests that as these variables increase, the Sales also tend to increase, indicating a strong positive linear relationship.
- Product Price, Order Item Product Price, and Order Item Discount are moderately positively correlated with values between 0.6 and 0.8. This indicates that as these variables increase, there is a tendency for other variables to also increase, but with a moderate strength of the relationship.
- Order Item CardProd Id, Product Card Id, Category Id, Product Category Id, and Department Id are weakly positively correlated with values between 0.2 and 0.3. This suggests a relatively weaker positive linear relationship between these variables.
- Order Profit Per Order and Benefit per Order are weakly positively correlated with values around 0.1, indicating a weak positive linear relationship.
- Order Id, Order Item Id, Customer Id, and Order Customer Id are minimally correlated with values close to 0, indicating little to no linear relationship.
- Order Item Quantity, Longitude, Order Zipcode, Late_delivery_risk, Days for Shipment (Scheduled), Days for Shipping (Real), Order Item Profit Ratio, Order Item Discount Rate, Latitude, Customer Zipcode, Product Description, and Product Status have very weak or no correlation with other variables, as their values are close to 0.

According to the table, the Random Forest and Decision Tree algorithms have achieved exceptionally high accuracy values of 0.99978 and 0.99968, respectively, which are very close to perfect accuracy. On the contrary, the Linear Regression and Ridge Regressor algorithms have achieved comparatively lower accuracy values of 0.80138 and 0.81037, respectively. However, it's worth noting that accuracy alone may not always give a comprehensive understanding of a model's performance, and it's important to consider other evaluation metrics and factors that may be relevant depending on the specific problem and dataset being analyzed.

Based on the analyzed data, the following steps could be taken to potentially increase sales:

- Focus on order fulfilment rates and delivery times: The high positive correlation between sales and order fulfilment rates, delivery times, and sales per customer (with correlation values of 0.989324) suggests that improving these metrics could lead to increased sales. Ensuring timely order fulfilment and delivery can result in better customer satisfaction and repeat business.
- Consider product pricing and discount strategies: The positive correlation between sales and product price, order item product price, and order item discount (with correlation values of 0.811690 and 0.604080 respectively) suggests that pricing and discount strategies may impact sales. Analyzing pricing data, competitor pricing, and customer

preferences can help optimize product pricing and discount strategies to attract more customers and drive sales.

- Analyze customer and order data: The correlation between sales and customer-related data such as customer ID, order ID, and order customer ID (with correlation values ranging from 0.066322 to 0.110351) indicates that understanding customer behaviour and preferences can help increase sales. Analyzing customer data, order data, and customer feedback can provide insights for targeted marketing campaigns, personalized offers, and improved customer service to boost sales.
- Explore opportunities for product/category/department expansion: The positive correlation between sales and product/category/department-related data (with correlation values ranging from 0.239191 to 0.255174) suggests that expanding product offerings or entering new categories/departments could potentially increase sales. Analyzing market trends, customer demands, and competition can help identify growth opportunities and strategically expand the product/category/department portfolio to drive sales.

It's important to note that these steps should be implemented based on a thorough understanding of the specific business context, customer preferences, and market dynamics. Regular monitoring and analysis of sales performance and feedback from customers can provide valuable insights for continuous improvement and optimization of sales strategies.

Summary

This analysis report focuses on a supply chain dataset used by DataCo Global. It utilizes machine learning algorithms and R software to analyze the dataset and extract insights related to order fulfillment rates, delivery times, inventory turnover, transportation costs, and customer service levels. The report aims to provide a comprehensive overview of the dataset, share key findings, and offer recommendations for optimizing supply chain performance. The analysis includes performance evaluation, correlation analysis, data visualizations, and utilization of machine learning algorithms such as Linear Regression, Ridge Regressor, Random Forest, and Decision Tree. The analysis of the dataset revealed significant positive correlations between variables such as Sales, Order Item Total, and Sales per Customer, indicating a strong linear relationship. Other variables like Product Price, Order Item Product Price, and Order Item Discount showed moderate positive correlations. Variables like Order Item Cardprod Id, Product Card Id, Category Id, Product Category Id, and Department Id had weaker positive correlations. Order Profit Per Order and Benefit per Order showed weak positive correlations. However, some variables had minimal or no correlation. It's important to consider the context of the dataset and the specific goals of the analysis for a comprehensive interpretation.