HULT
INTERNATIONAL
BUSINESS SCHOOL

Retail Product Sales Data Analysis and Reporting

Visualizing and Analyzing Data with R: Methods & Tools

Angel E. Lanto

Master's in Business Analytics

Table of Contents

I. Introduction

This report aims to demonstrate the application of data analysis methods learned throughout the course, focusing on cleaning, exploring, manipulating, visualizing, and analyzing a real-world dataset. The dataset used for this analysis is the Retail Product Sales Data, sourced from Kaggle. It contains detailed information on product sales and customer purchases across multiple stores, offering a comprehensive view of sales transactions and customer behavior.

The dataset includes several key attributes such as Order ID, Order Date, Ship Date, Customer ID, Product ID, Sales, and geographical information, among others. It also provides details about the products, including categories and sub-categories. By analyzing this data, the goal is to derive valuable insights that can address specific business questions, such as understanding sales trends, identifying high-performing product categories, and exploring customer segmentation.

Through this report, I will apply various data analysis techniques using R, including data cleaning, transformation, exploratory analysis, and visualization, to uncover patterns and insights that can aid in business decision-making and forecasting.

Please note that there are no data labels on the visualizations below. However, it is recommended to explore the values interactively within the notebook, as the visualizations were created using the Plotly library, which offers interactive features such as zooming, hovering for details, and dynamic updates. This allows for a more in-depth and flexible analysis of the data.

II. Data Loading and Exploration

train.csv

The dataset consists of 9,800 entries and 18 columns, each representing different aspects of product sales and customer information. The columns include identifiers for each row (Row ID), order (Order ID), and customer (Customer ID), as well as order and shipment details like Order Date, Ship Date, and Ship Mode. Additional information such as customer name, segment, and geographical data (Country, City, State, Postal Code, Region) is provided. The dataset also includes product-related details, such as Product ID, Category, Sub-Category, Product Name, and Sales, with the Sales column representing the total sales amount for each order.

Upon initial inspection, it was found that the dataset has no major inconsistencies in terms of data types or structure. The data is a mix of numeric and categorical values, with the Postal Code and Sales columns being numeric. The remaining columns are categorical, providing information about the customer, product, and order. One issue that needs to be addressed is the presence of missing values in the Postal Code column, with 11 missing entries out of 9,800. Additionally, the Order Date and Ship Date columns are currently stored as character data types, but they should be converted to date format to facilitate proper analysis of time-related trends and patterns.

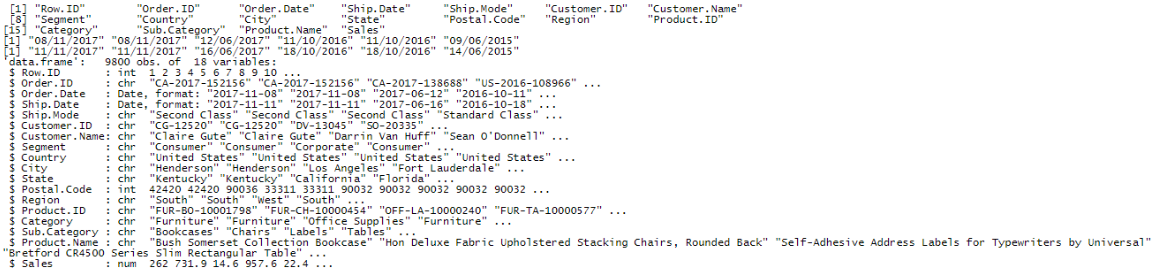
Overall, the dataset is well-structured for analysis, offering clear identifiers for each order, product, and customer, along with useful categorical variables for segmentation and deeper insights. The issues mentioned will be addressed during the data cleaning process, ensuring that the data is both complete and correctly formatted for further analysis.

The provided code first loads the necessary libraries required for data manipulation, visualization, and analysis. These libraries include readr for reading CSV files, dplyr for data manipulation, zoo for handling time series data, plotly for interactive visualizations, tidyr for reshaping data, ggplot2 for static data visualization, and lubridate for working with date and time.

Next, the code loads the dataset from a specified path using the read.csv() function, ensuring that string variables are not converted into factors by setting stringsAsFactors = FALSE. The head() function is then used to inspect the first few rows of the dataset, providing a quick view of the data's content.

The str() function is applied to display the structure of the dataset, including the data types of each column. This helps in understanding the overall structure of the data and identifying any inconsistencies, such as columns that may need to be converted to a different data type for analysis (e.g., dates or numeric variables). This initial exploration is crucial for preparing the data for further cleaning and analysis.

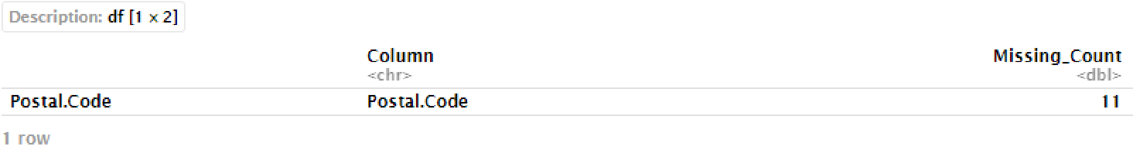
III. Data Cleaning and Handling Missing Values



```
 [1] "Row.ID"        "Order.ID"      "Order.Date"    "Ship.Date"     "Ship.Mode"     "Customer.ID"   "Customer.Name"
 [8] "Segment"       "Country"       "City"          "State"         "Postal.Code"   "Region"        "Product.ID"   
[15] "Category"      "Sub.Category"  "Product.Name"  "Sales"        
[1] "08/11/2017" "08/11/2017" "12/06/2017" "11/10/2016" "11/10/2016" "09/06/2015"
[1] "11/11/2017" "11/11/2017" "16/06/2017" "18/10/2016" "18/10/2016" "14/06/2015"
'data.frame':	9800 obs. of  18 variables:
 $ Row.ID       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Order.ID     : chr  "CA-2017-152156" "CA-2017-152156" "CA-2017-138688" "US-2016-108966" ...
 $ Order.Date   : Date, format: "2017-11-08" "2017-11-08" "2017-06-12" "2016-10-11" ...
 $ Ship.Date    : Date, format: "2017-11-11" "2017-11-11" "2017-06-16" "2016-10-18" ...
 $ Ship.Mode    : chr  "Second Class" "Second Class" "Second Class" "Standard Class" ...
 $ Customer.ID  : chr  "CG-12520" "CG-12520" "DV-13045" "SO-20335" ...
 $ Customer.Name: chr  "Claire Gute" "Claire Gute" "Darrin Van Huff" "Sean O'Donnell" ...
 $ Segment      : chr  "Consumer" "Consumer" "Corporate" "Consumer" ...
 $ Country      : chr  "United States" "United States" "United States" "United States" ...
 $ City         : chr  "Henderson" "Henderson" "Los Angeles" "Fort Lauderdale" ...
 $ State        : chr  "Kentucky" "Kentucky" "California" "Florida" ...
 $ Postal.Code  : int  42420 42420 90036 33311 33311 90032 90032 90032 90032 90032 ...
 $ Region       : chr  "South" "South" "West" "South" ...
 $ Product.ID   : chr  "FUR-BO-10001798" "FUR-CH-10000454" "OFF-LA-10000240" "FUR-TA-10000577" ...
 $ Category     : chr  "Furniture" "Furniture" "Office Supplies" "Furniture" ...
 $ Sub.Category : chr  "Bookcases" "Chairs" "Labels" "Tables" ...
 $ Product.Name : chr  "Bush Somerset Collection Bookcase" "Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back" "Self-Adhesive Address Labels for Typewriters by Universal" 
"Bretford CR4500 Series Slim Rectangular Table" ...
 $ Sales        : num  262 731.9 14.6 957.6 22.4 ...
```

Figure 1. Preview of the Dataset

In the data cleaning process, several steps were taken to ensure the dataset was properly prepared for analysis. First, the column names were checked to ensure no extra spaces or issues. The Order Date and Ship Date columns were initially in character format, so they were converted into date format using the as.Date() function. The format "%d/%m/%Y" was applied to both columns to ensure proper conversion. After this transformation, the structure of the dataset was verified to ensure the dates were correctly formatted.



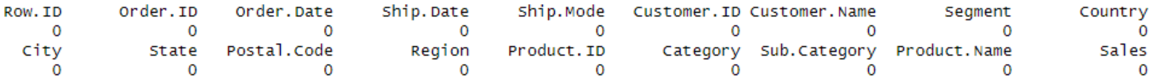
Description: df [1 × 2]

	Column <chr>	Missing_Count <dbl>
Postal.Code	Postal.Code	11

1 row

Figure 2. Missing Values

Next, the dataset was checked for missing values across all columns using the colSums(is.na()) function. This revealed that the Postal Code column had 11 missing values. Since the Postal Code is not critical for the analysis but still needed to be represented, these missing values were filled with a placeholder value of "00000" because the proponent was unsure which of the 7 possible postal codes pertaining to Burlington, Vermont should be applied to these records.



```
       Row.ID      Order.ID    Order.Date     Ship.Date     Ship.Mode   Customer.ID Customer.Name       Segment       Country 
            0             0             0             0             0             0             0             0             0 
         City         State   Postal.Code        Region    Product.ID      Category  Sub.Category  Product.Name         Sales 
            0             0             0             0             0             0             0             0             0 
```

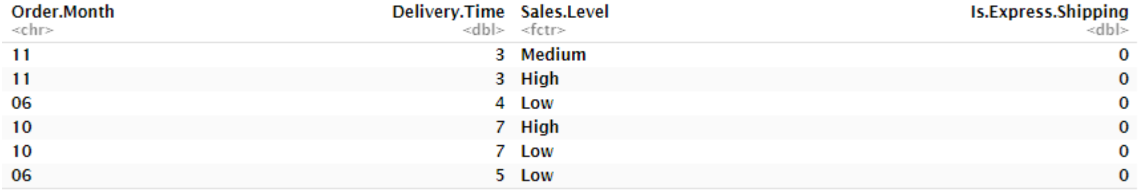
Figure 3. Result of Filling the Missing Values

After filling in the missing values, the dataset was checked again to ensure there were no remaining missing values.

updated_train_data.csv

Finally, the cleaned dataset was saved to a new CSV file for further use. These steps ensured that the data was free from missing values in key columns and that the date columns were appropriately formatted for time-based analysis.

IV. Data Transformation and Feature Engineering



Order.Month	Delivery.Time	Sales.Level	Is.Express.Shipping
<chr>	<dbl>	<fctr>	<dbl>
11	3	Medium	0
11	3	High	0
06	4	Low	0
10	7	High	0
10	7	Low	0
06	5	Low	0

Figure 4. Adding New Features

The provided code performs several data transformation and feature engineering steps on the dataset. First, the Order Date and Ship Date columns are converted to Date type, ensuring that the data is properly formatted for time-related analysis. A new column, Order Month, is created by extracting the month from the Order Date, allowing for easier analysis of monthly sales patterns.

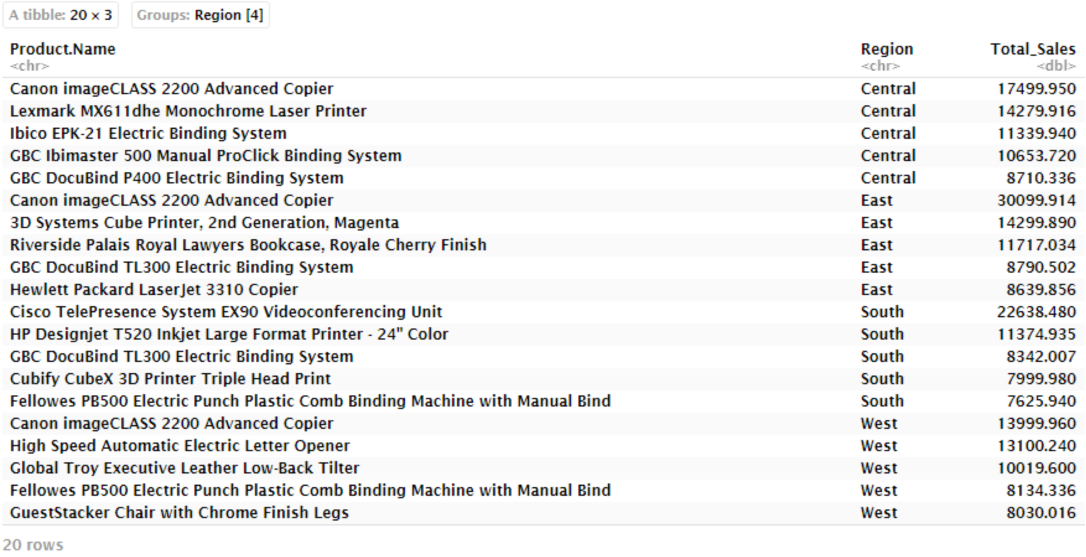
Next, a new feature, Delivery Time, is added by calculating the difference between the Ship Date and Order Date in days, which helps analyze shipping efficiency and potential delays. A Sales Level feature is also introduced, categorizing sales into three levels such as Low, Medium, and High which are based on predefined sales thresholds. This categorization can help identify the performance of products or orders at different sales levels.

Additionally, a binary feature, Is Express Shipping, is created, which assigns a value of 1 if the shipping mode is First Class or Same Day (indicating express shipping) and 0 for Second Class or Standard Class (indicating non-express shipping). This feature helps to analyze the impact of express shipping on sales or delivery times.

The table function is used to verify that the Is.Express.Shipping feature correctly contains values of 0 and 1, confirming that the transformation was successful. Finally, the first few rows of the dataset are displayed to inspect the newly created features.

The resulting dataset is enriched with new, meaningful features such as Order Month, Delivery Time, Sales Level, and Is Express Shipping, which can be used for further analysis or predictive modeling. These transformations enhance the ability to gain insights into sales trends, delivery performance, and customer behavior based on shipping methods.

V. Grouping and Aggregation



A tibble: 20 x 3 Groups: Region [4]

Product.Name <chr>	Region <chr>	Total_Sales <dbl>
Canon imageCLASS 2200 Advanced Copier	Central	17499.950
Lexmark MX611dhe Monochrome Laser Printer	Central	14279.916
Ibico EPK-21 Electric Binding System	Central	11339.940
GBC Ibimaster 500 Manual ProClick Binding System	Central	10653.720
GBC DocuBind P400 Electric Binding System	Central	8710.336
Canon imageCLASS 2200 Advanced Copier	East	30099.914
3D Systems Cube Printer, 2nd Generation, Magenta	East	14299.890
Riverside Palais Royal Lawyers Bookcase, Royale Cherry Finish	East	11717.034
GBC DocuBind TL300 Electric Binding System	East	8790.502
Hewlett Packard LaserJet 3310 Copier	East	8639.856
Cisco TelePresence System EX90 Videoconferencing Unit	South	22638.480
HP Designjet T520 Inkjet Large Format Printer - 24" Color	South	11374.935
GBC DocuBind TL300 Electric Binding System	South	8342.007
Cubify CubeX 3D Printer Triple Head Print	South	7999.980
Fellowes PB500 Electric Punch Plastic Comb Binding Machine with Manual Bind	South	7625.940
Canon imageCLASS 2200 Advanced Copier	West	13999.960
High Speed Automatic Electric Letter Opener	West	13100.240
Global Troy Executive Leather Low-Back Tilter	West	10019.600
Fellowes PB500 Electric Punch Plastic Comb Binding Machine with Manual Bind	West	8134.336
GuestStacker Chair with Chrome Finish Legs	West	8030.016

20 rows

Figure 5. Summarizing Data by Product and Region

The code provided groups the dataset by Product Name and Region, then calculates the total sales for each combination. Afterward, the data is sorted by region and descending total sales, and the top 5 products in each region are identified based on total sales. The resulting table provides a detailed view of the highest-selling products across different regions.

The results reveal that the Canon imageCLASS 2200 Advanced Copier is the top-selling product in multiple regions, including Central, East, and West, showcasing its widespread popularity. In the Central Region, it leads the sales, followed by other office equipment such as Lexmark MX611dhe Monochrome Laser Printer and Ibico EPK-21 Electric Binding System. Similarly, in the East Region, office-related products dominate, but we also see a variety of products like the 3D Systems Cube Printer and Riverside Palais Royal Lawyers Bookcase, which indicates a more diverse market. In the South Region, technology products such as the Cisco TelePresence System EX90 Videoconferencing Unit and HP Designjet T520 Inkjet Large Format Printer are top performers, highlighting a stronger demand for tech and printing equipment. In the West Region, the Canon imageCLASS 2200 Advanced Copier remains a leader in sales, with office equipment like the High Speed Automatic Electric Letter Opener and Fellowes PB500 Electric Punch Plastic Comb Binding Machine also performing well.

These findings suggest that office-related products are consistently popular across regions, but there is some regional diversity, with the East showcasing a broader product mix and the South leaning more towards technology-driven products. This information can inform inventory, marketing strategies, and product targeting, ensuring businesses focus on high-demand products based on regional trends.

VI. Data Visualization

The provided code aims to visualize total sales by product category using an interactive bar chart. Initially, new columns for Year and Month were created by extracting the respective values from the Order Date. The dataset was then grouped by Category, Year, and Month, with the total sales calculated for each combination. The data was sorted by total sales in descending order. An interactive bar chart was created using plotly, allowing users to filter the data by Year and Month. The bar chart displays the total sales for each product category, with the option to select a specific year or month for further analysis.

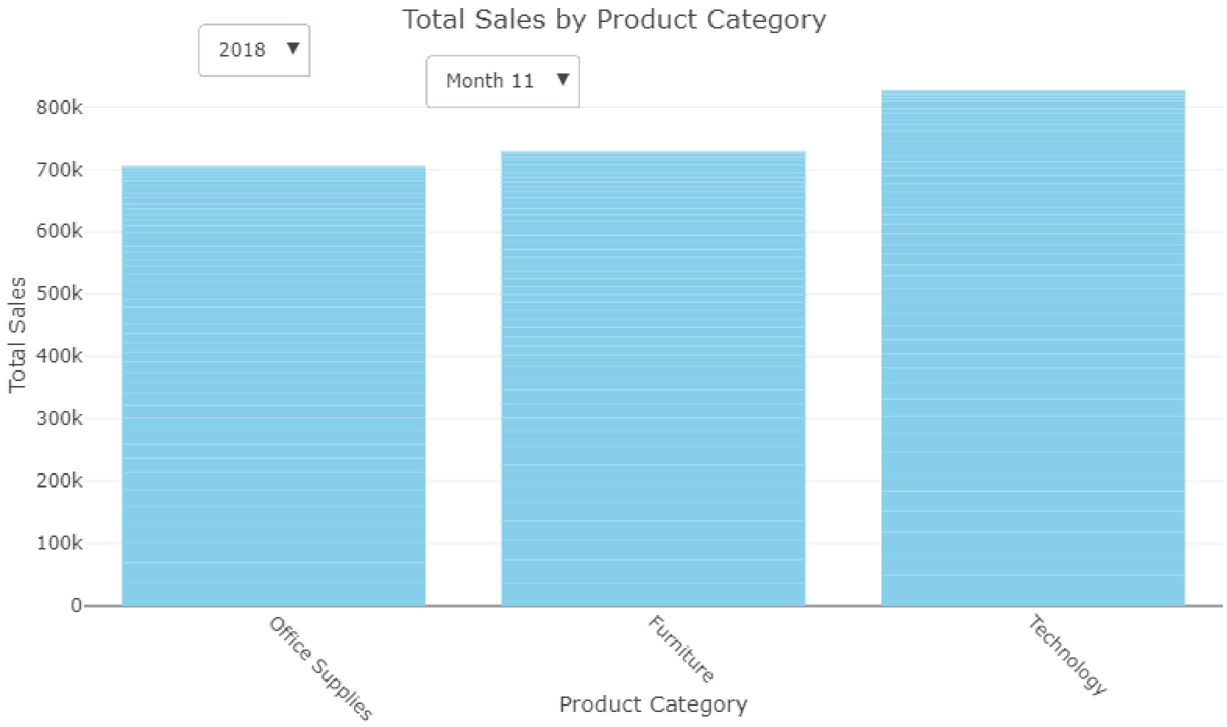


Figure 6. Visualizing Total Sales by Product Category

The resulting bar chart shows that the total sales for Office Supplies, Furniture, and Technology are fairly close, with Technology and Office Supplies showing slightly higher sales than Furniture. This suggests that while all three product categories contribute significantly to sales, Technology and Office Supplies may have a stronger market presence or higher demand. The interactive nature of the chart allows for a more detailed exploration of sales trends by filtering data by year or month, making it easier to identify patterns and trends over time.

The provided code creates an interactive line chart that visualizes the trend of total sales over time, grouped by Year and Month. First, new columns for Year and Month were extracted from the Order Date to enable time-based aggregation. The dataset was then grouped by these time dimensions, and the total sales were calculated for each combination. A new Date column was created, using the first day of each month, to plot the data on a timeline. The code uses plotly to create an interactive line chart, which displays total sales over time, with the ability to filter the data by Year and Month via dropdown menus.

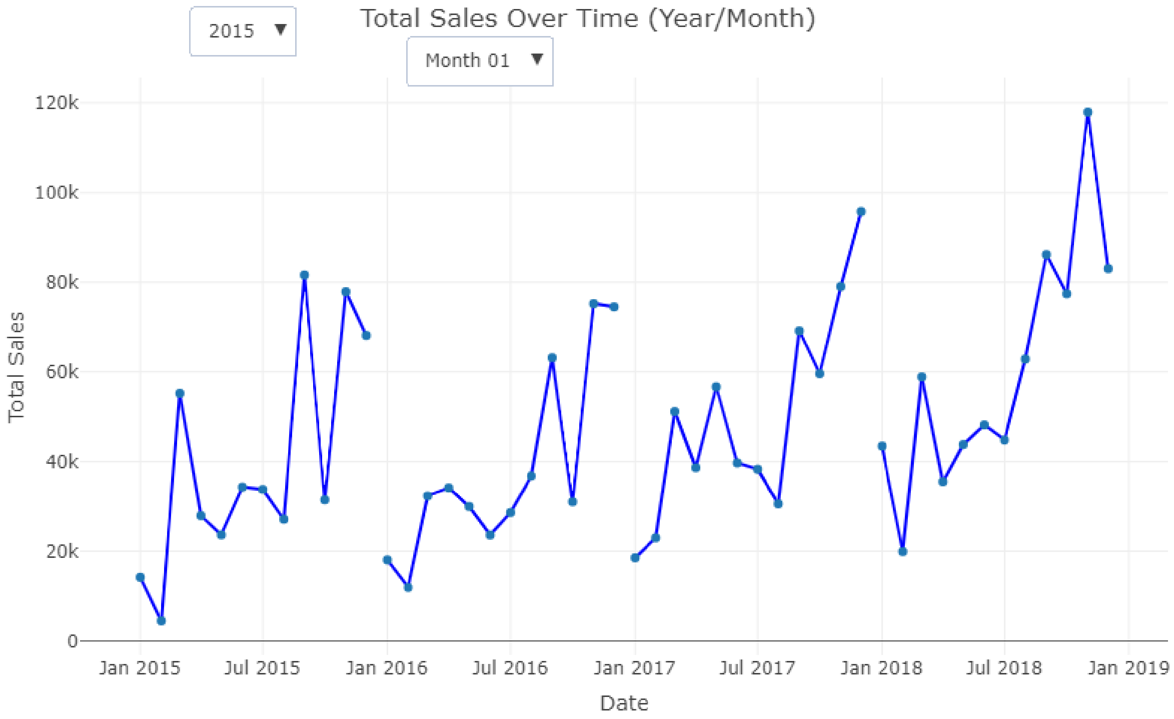
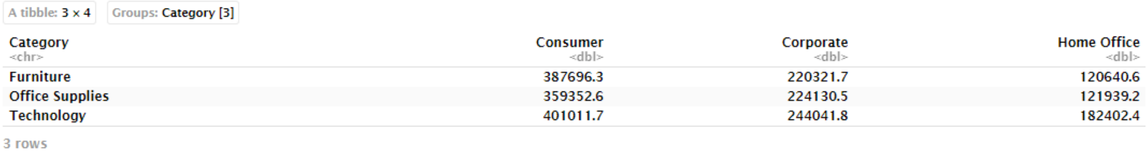


Figure 7. Visualizing Total Sales Over Time (Year/Month)

The resulting line chart shows fluctuations in total sales over time, with noticeable peaks in specific months, especially towards the end of each year. The interactive nature of the chart allows users to filter data by Year and Month, offering an in-depth exploration of sales patterns. The analysis reveals that sales tend to be higher in later years, indicating potential growth or seasonal demand increases, such as end-of-year sales or promotions. This trend can provide valuable insights for forecasting and planning future sales strategies, highlighting seasonal variations and growth patterns over time.

VII. Data Manipulation and Reshaping

The provided code aggregates the total sales by product category and customer segment. It groups the data by Category (such as Furniture, Office Supplies, and Technology) and Segment (Consumer, Corporate, and Home Office), then calculates the total sales for each combination. The data is then reshaped using pivot_wider to create a table where each customer segment is represented as a separate column, with Total Sales as the values. If there are any missing values, they are replaced with 0, ensuring a complete table for analysis.



A tibble: 3 × 4 Groups: Category [3]

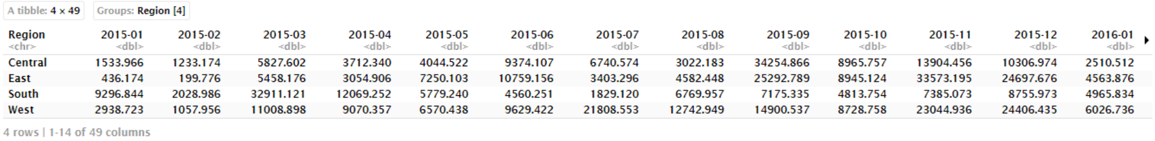
Category	Consumer	Corporate	Home Office
<chr>	<dbl>	<dbl>	<dbl>
Furniture	387696.3	220321.7	120640.6
Office Supplies	359352.6	224130.5	121939.2
Technology	401011.7	244041.8	182402.4

3 rows

Figure 8. Pivot Table for Total Sales by Product Category and Segment

The resulting pivot table shows the total sales for each product category across different customer segments. For instance, Technology leads in sales across all segments, with Consumer and Corporate segments contributing significantly, while Furniture and Office Supplies also show strong performance in the Consumer segment. These insights provide a clearer view of which product categories are most popular within each segment, helping businesses tailor their strategies for different customer groups. The analysis reveals that Technology has the highest sales overall, particularly in the Corporate and Home Office segments, while Furniture performs well in the Consumer segment.

On the other hand, another code provided aggregates total sales by region and month. It first creates a new column for the Month by extracting the year and month from the Order Date. The data is then grouped by Region and Month, and the total sales for each combination are calculated. The pivot_wider function is used to reshape the data so that each month becomes a separate column, with Total Sales as the values, filling any missing months with 0 to ensure complete data for analysis.



A tibble: 4 × 49 Groups: Region [4]

Region	2015-01	2015-02	2015-03	2015-04	2015-05	2015-06	2015-07	2015-08	2015-09	2015-10	2015-11	2015-12	2016-01
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Central	1533.966	1233.174	5827.602	3712.340	4044.522	9374.107	6740.574	3022.183	34254.866	8965.757	13904.456	10306.974	2510.512
East	436.174	199.776	5458.176	3054.906	7250.103	10759.156	3403.296	4582.448	25292.789	8945.124	33573.195	24697.676	4563.876
South	9296.844	2028.986	32911.121	12069.252	5779.240	4560.251	1829.120	6769.957	7175.335	4813.754	7385.073	8755.973	4965.834
West	2938.723	1057.956	11008.898	9070.357	6570.438	9629.422	21808.553	12742.949	14900.537	8728.758	23044.936	24406.435	6026.736

4 rows | 1-14 of 49 columns

Figure 9. Pivot Table for Sales by Region and Month

The resulting pivot table displays the total sales for each Region (Central, East, South, and West) across different months. This table provides insights into how sales fluctuate across regions over time, allowing for a better understanding of regional trends. For instance, some regions, such as the South, show higher sales in certain months, while others like the West may display more consistent sales or notable peaks in specific months. These insights help businesses analyze regional performance and plan accordingly, tailoring strategies to optimize sales in each region based on seasonal or time-specific trends.

VIII. Predicting Future Sales with Regression

The provided code aggregates total sales by order month and visualizes the monthly sales trend using an interactive Plotly plot. First, a new Order Month column is created by extracting the month from the Order Date. The data is then grouped by Order Month, and the total sales for each month are calculated. This aggregated data is used to create a scatter plot with lines and markers, where the x-axis represents the Order Month and the y-axis shows the corresponding Total Sales. The plot uses blue lines to represent the sales trend and red markers to highlight each month's total sales.

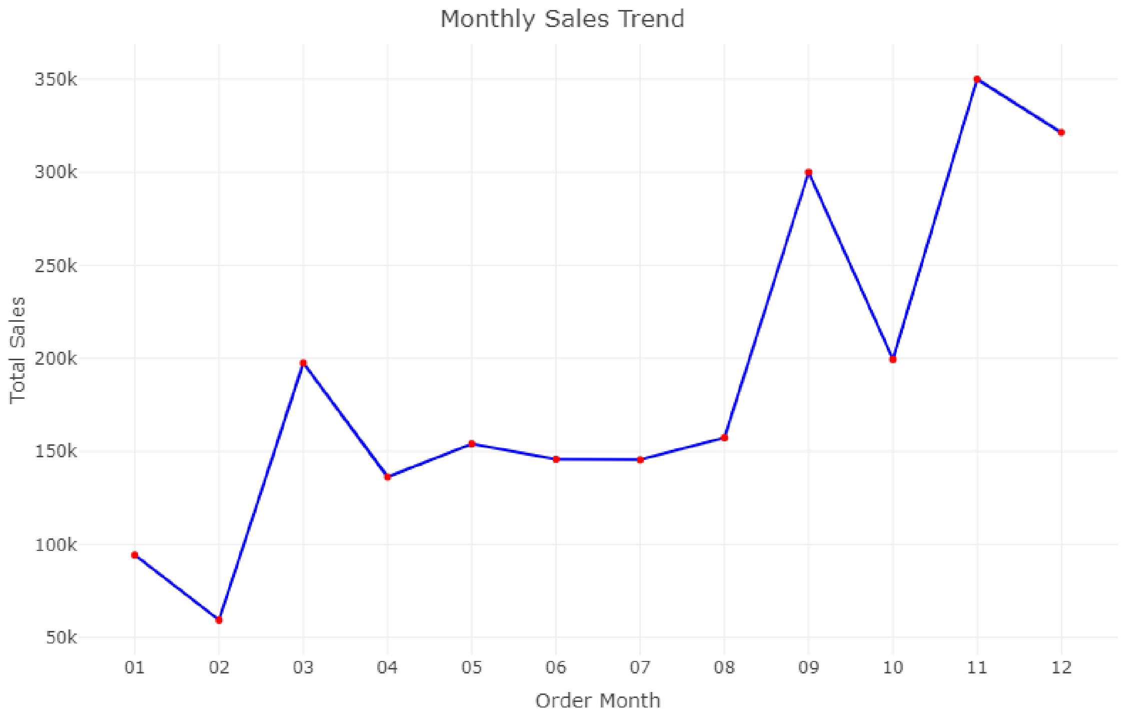


Figure 10. Monthly Sales Trend

The resulting plot shows the sales trend over the months, with visible peaks and valleys, indicating fluctuations in sales. Specifically, there is a sharp increase in sales towards the end of the year, particularly in November and December, which suggests higher sales during this period. This plot helps to identify seasonal trends, enabling businesses to recognize high-demand months and plan accordingly. The interactive nature of the plot allows users to explore the sales data for each month, offering insights into sales performance over time and aiding in future sales forecasting.

For the next set of code, it uses linear regression to predict future monthly sales based on historical sales data. First, the Order Month and Order Year are extracted from the Order Date column. The sales data is aggregated by Order Month, and a linear regression model is created to predict total sales based on the month number.

Next, the code predicts sales for the next three months by extending the Order Month values. The predicted sales are combined with the existing data, and an interactive Plotly plot is generated. The plot displays the historical sales trend with blue lines and red markers for each month's total sales, while green markers represent the predicted sales for the future months.

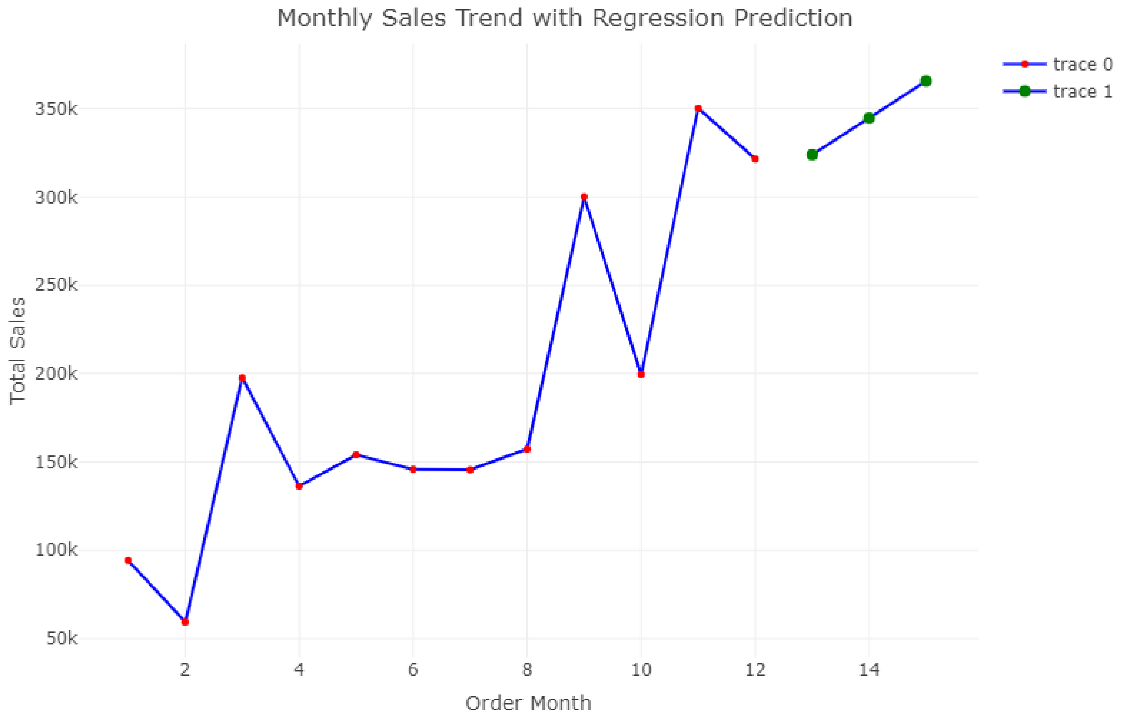
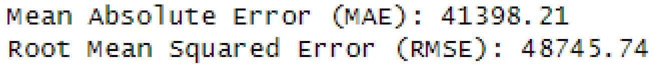


Figure 11. Monthly Sales with Regression Prediction for the Next 3 Months

The plot reveals the historical trend in sales, with noticeable fluctuations over time. The regression model is used to forecast sales for the upcoming months, providing a prediction that shows a potential increase in sales based on the historical pattern. This analysis can be useful for forecasting sales trends and planning future business strategies based on expected sales growth.

The next code evaluates the performance of the linear regression model by calculating two key metrics: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). After predicting the sales using the regression model, the code calculates the MAE, which represents the average absolute difference between the predicted and actual sales, and the RMSE, which measures the square root of the average squared differences. These metrics give insights into the accuracy of the model’s predictions.



```
Mean Absolute Error (MAE): 41398.21
Root Mean Squared Error (RMSE): 48745.74
```

Figure 12. Result of the Regression Model using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)

The results indicate that the MAE is 41,398.21, meaning the average difference between predicted and actual sales is approximately 41,398 units. The RMSE value is 48,745.74, indicating a higher error when considering the squared differences. These values suggest that while the model can capture the overall trend in sales, there is still a considerable amount of error in its predictions, particularly for more significant sales values.

These evaluation metrics are essential for assessing the reliability of the regression model and understanding its predictive capabilities. The relatively high MAE and RMSE indicate that improvements in the model, such as adding more features or using a more complex algorithm, may be necessary to improve prediction accuracy.

IX. Conclusion

In conclusion, the Retail Product Sales Data Analysis has provided valuable insights into the dynamics of product sales, regional performance, inventory management, and delivery efficiency. The analysis revealed distinct patterns across regions, with Technology products performing strongly in the South and West, while Office Supplies and Furniture dominated in the Central and East regions. Additionally, the examination of seasonal sales trends highlighted peak demand during the holiday months, providing crucial information for inventory planning and sales forecasting.

The introduction of new features, such as Delivery Time and Is Express Shipping, has enriched the dataset, offering deeper insights into shipping efficiency and its potential impact on customer satisfaction and sales. The regression model used to forecast future sales has shown potential, though the high Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) suggest that further refinement of the model could improve its predictive accuracy.

Based on these findings, it is recommended that the business focuses on region-specific inventory management, ensuring sufficient stock of high-demand products like Technology and Office Supplies in the appropriate regions. Additionally, targeted marketing campaigns should be employed to cater to the demand trends in each region. Improving delivery efficiency by offering express shipping options, especially during peak seasons, could further enhance customer satisfaction and drive higher sales.

Overall, by leveraging these insights, the business can optimize its operations, improve forecasting accuracy, and tailor its strategies to boost sales performance across regions, while also enhancing customer satisfaction through more efficient delivery systems. The data-driven approach demonstrated in this analysis sets a strong foundation for informed decision-making and strategic planning in the future.

X. Final Analysis and Reporting

The analysis of the Retail Product Sales Data revealed valuable insights into sales performance across regions, inventory management, and delivery efficiency. The sales trends across different regions showed significant variability, with the South region demonstrating stronger demand for technology-related products, such as the Cisco TelePresence System and HP Designjet T520. The West region, on the other hand, showed more consistent sales, while the Central and East regions exhibited a more diverse mix of products, with office equipment and furniture performing well in the Central region, and more varied products like 3D printers and bookcases in the East. The findings also highlighted seasonal trends, with peaks in November and December, suggesting increased demand during the holiday season.

In terms of inventory management, it is recommended that businesses focus on maintaining a high stock of Technology products, particularly in the South and West regions, where demand is higher. Additionally, ensuring a diverse range of products in the East and maintaining seasonal stock for peak periods like November and December will help meet customer demand and maximize sales. The regression model used to predict future sales indicated potential growth, but the Mean Absolute Error (MAE) of 41,398.21 and Root Mean Squared Error (RMSE) of 48,745.74 suggest that further refinement of the model may be needed for more accurate predictions.

For sales performance monitoring, the business should regularly track monthly sales trends to identify shifts in customer preferences and make proactive inventory adjustments. Targeted marketing campaigns should be tailored to capitalize on the demand in specific regions, especially in the South for technology products and the East and Central regions for office supplies and furniture. In terms of delivery efficiency, the creation of the Is Express Shipping feature suggests that offering faster delivery options, particularly during peak seasons, can improve customer satisfaction and potentially boost sales. Investing in real-time tracking and optimizing logistics for faster and more reliable delivery would help improve delivery times and increase customer retention.

Therefore, by strategically managing inventory based on regional demand and product categories, refining sales forecasts, and improving delivery efficiency, businesses can optimize operations and increase overall sales performance. The analysis also suggests that enhancing the regression model for more accurate predictions could help refine future sales forecasting, ensuring better decision-making and resource allocation.

XI. References

ChatGPT. (2025). Chatgpt.com. https://chatgpt.com/g/g-p-67b89f5a4de88191b004808e58af027e-visualizing-analyzing-data-with-r/c/67d1fd85-b6a0-8002-9d68-5656114a33a3