MACQUARIE
University

BUSA8000: TECHNIQUES IN BUSINESS ANALYTICS

Assignment-2 (Session 1- 2024)

Report for Dibs Retail

BY

	Group #	Student ID	Student Name	Responsible Task #
1	**54**	48023787	Deepali Raj	Task 1
2	**54**	47937629	Abhay Sachdeva	Task 2
3	**54**	47998105	Sarah Jane Daniel	Task 3
4	**54**	48078638	Varun Wadhwani	Task 4

Table of Contents

Executive Summary

This report analyzes Dibs Retail's sales data to enhance sales and customer loyalty. Key findings include 2019 as the best sales year, with peak sales in December and October. California led in sales, and AAA Batteries were top-sellers.

Predictive models, particularly the random forest, accurately forecast sales. These models are instrumental for Dibs in strategic planning and decision-making.

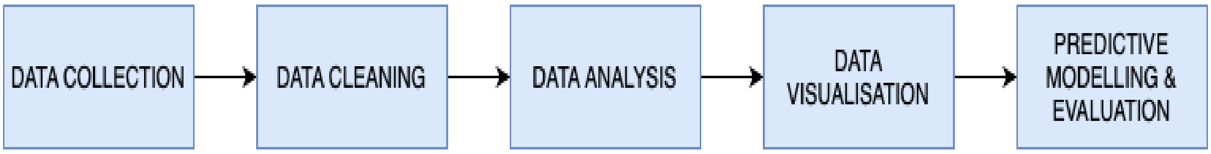
Recommendations are to collect data closely correlated with sales and gather product reviews focusing on price, quality, satisfaction, and usefulness to identify loyal customers. These insights aim to optimize Dibs' sales strategies, ensuring sustained growth and improved customer loyalty.

Introduction

Dibs, a rapidly growing online retailer specializing in accessories, home goods, and electronics, faces challenges in boosting sales and customer loyalty. Despite substantial customer purchase data, Dibs lacks expertise in leveraging it.

Our team will analyze their data using data visualization and statistical techniques to uncover customer behavior patterns and trends, aiding in targeted marketing campaigns. Employing machine learning, we'll develop a predictive model for forecasting sales trends. Our goal is to provide actionable recommendations to enhance Dibs' sales and customer loyalty, ensuring sustained growth in the competitive retail market.

Our Workflow



Data Cleaning

Overview

The data cleaning process involved several steps to ensure the dataset was ready for analysis. The steps included:

- **Data Importation**: Imported data from multiple sources.
- **Data Type Conversion**: Converted data types for consistency.
- **Handling Missing Values**: Detected and corrected non-numeric values in Quantity_Ordered and Price_Each columns.
- **Handling Duplicate Values**: Detected and corrected duplicate values in Product_Name and Purchase_Address columns.
- **Outlier Detection**: Identified and corrected errors such as outliers or inconsistent entries.
- **Date Conversion**: Converted date columns to appropriate date formats.
- **Data Merging**: Combined cleaned datasets into a master dataset for analysis.

Data Quality Insights and Recommendations

We identified data discrepancies and errors in Dibs Retail's sales records and recommend standardizing entry, enhancing validation, and ensuring consistency to improve data quality and strategic decision-making.

Data Types

- **Issue:** Columns such as Quantity_Ordered and Price_Each are stored as characters instead of numeric types.
- **Recommendation:** Ensure that the data types are correctly defined at the time of data collection. Numeric fields should be stored as integers or floats, and dates should be stored in an appropriate date/time format.

Missing and Invalid Values:

- **Issue**: Presence of non-numeric values in numeric fields (e.g., "$11.95" in Price_Each).
- **Recommendation**: Implement validation checks during data entry to prevent invalid data. For example, numeric fields should not accept alphabetic characters or special symbols (except for decimal points in prices).

Error Handling

- **Issue**: Erroneous data points such as incorrect product names and missing product IDs have led to numerous errors.
- **Recommendation**: At data entry, handle errors and validate data. Use drop-down menus or prepared lists for product names to reduce typos and provide each product a unique product ID to protect data integrity. Product IDs simplify inventory tracking, order processing, and sales data analysis. This improves operating efficiency and reduces mistakes.

Data Analysis and Insights

Sales Analysis

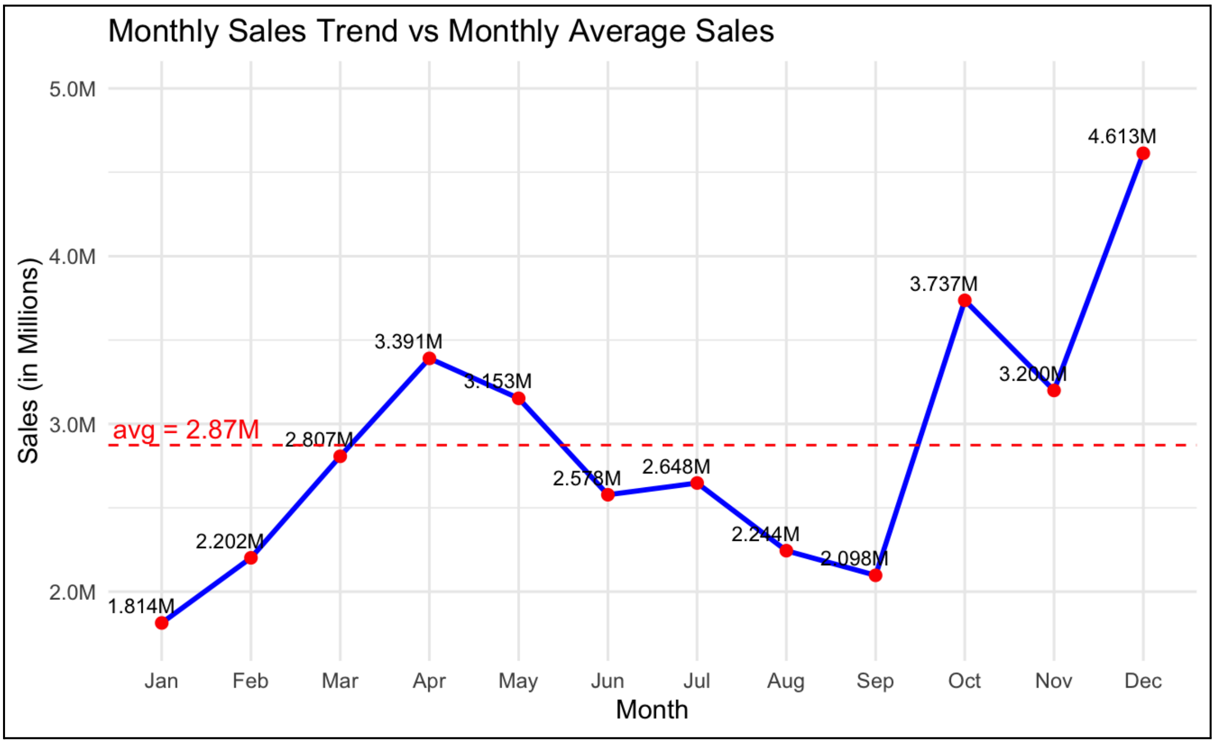
Key Findings:

1. **Worst Year of Sales**: 2021 with $3,927 in sales.
2. **Best Year of Sales**: 2019 with $34,483,366 in sales.
3. **Best Month of Sales in Best Year**: December 2019
4. **Revenue generated in Best Month of best Year of Sales:** $4,613,443
5. **Top Sales City in Best Year**: San Francisco, with a total order value of $8,259,719
6. **Optimal Advertising Time**: 7 PM - 8 PM (according to best year of sales i.e. 2019)
7. Most Sold Products Together: “iPhone” & “Lightning Charging Cable” are the most sold products as a bundle.
8. **Top Selling Product**: Product “AAA Batteries (4-pack)” sold the most(31,020 units sold), due to their universal utility, frequent need, and low price point ($2.99), making them a convenient and cost-effective choice for customers.
9. **Least Sold Product**: Due to its greater price, lesser demand than other gadgets, and intense competition from other companies offering identical features at lower costs, the LG Dryer was the least sold product during the best sales year.

Visualizations

NOTE: The data used for these was only from the best year of sales i.e. 2019.

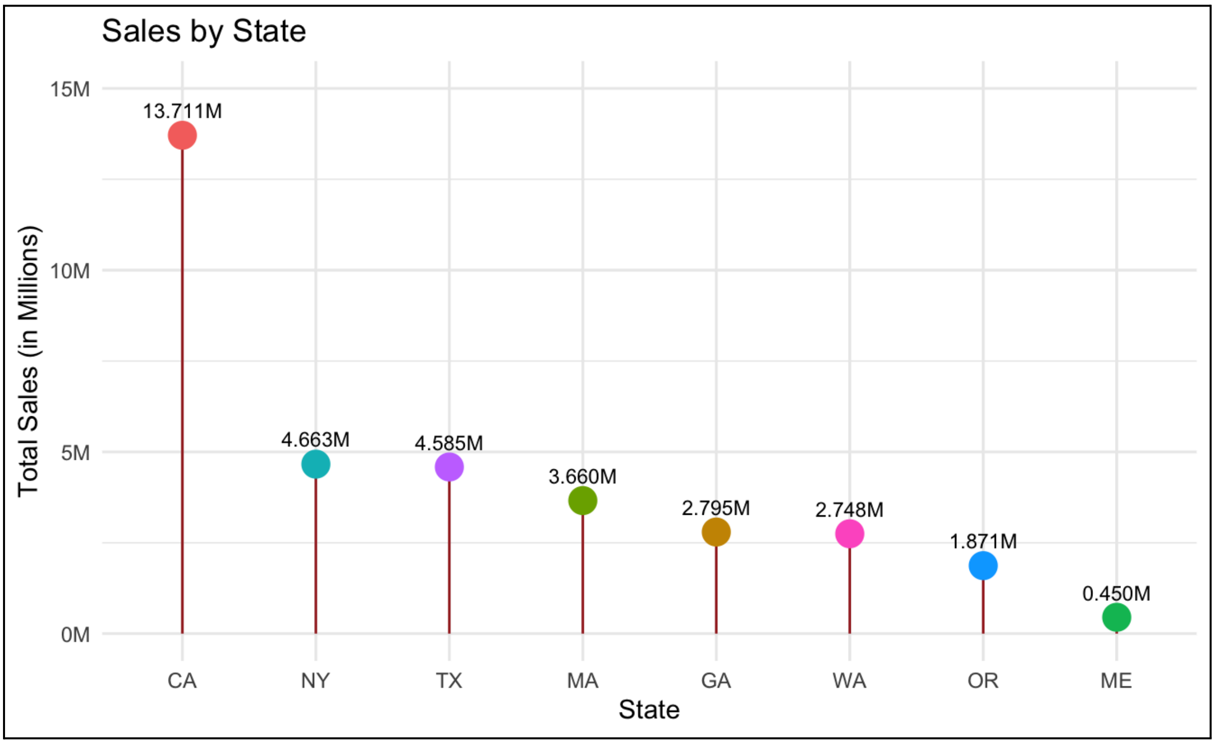
1. Monthly Sales Trend vs Monthly Average Sales



Sales show significant variability, with peaks in December (4.613 million) and October (3.737 million). The lowest sales occurred in January (1.814 million) and September (2.098 million). The average monthly sales for the year is approximately 2.87 million. Sales exceeded the average in April, May, October, November, and December, indicating strong performance in these months. In contrast, January, February, March, June, July, August, and September were below average, marking weaker periods.

Recommendation: Strategic marketing efforts should focus on peak periods to maximize revenue and targeted campaigns during low sales months to balance sales distribution

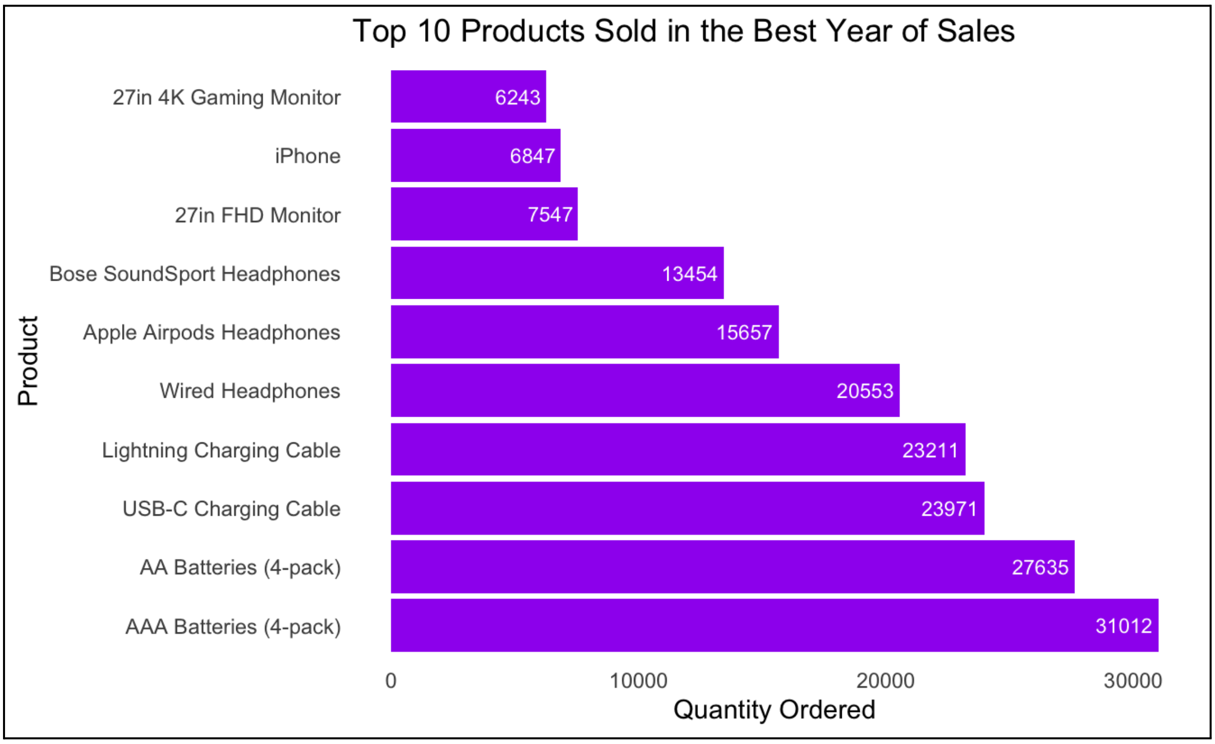
2. Sales by State



California (CA) leads with 13.711 million in sales, much outpacing other states. California's bigger population and economic activities may explain its supremacy. NY and TX follow with 4.663 million and 4.585 million sales, respectively, suggesting considerable market presence. Massachusetts (MA) and Georgia (GA) also had substantial sales of 3.660 million and 2.795 million. Washington (WA), Oregon (OR), and Maine (ME) had lower sales, with Maine having the lowest at 0.450 million.

Recommendation: This research advises targeting marketing in lower-performing states to enhance sales and capitalize on undiscovered areas. The need for strong operations and promotional efforts in top-performing states to maintain high sales volumes is also highlighted.

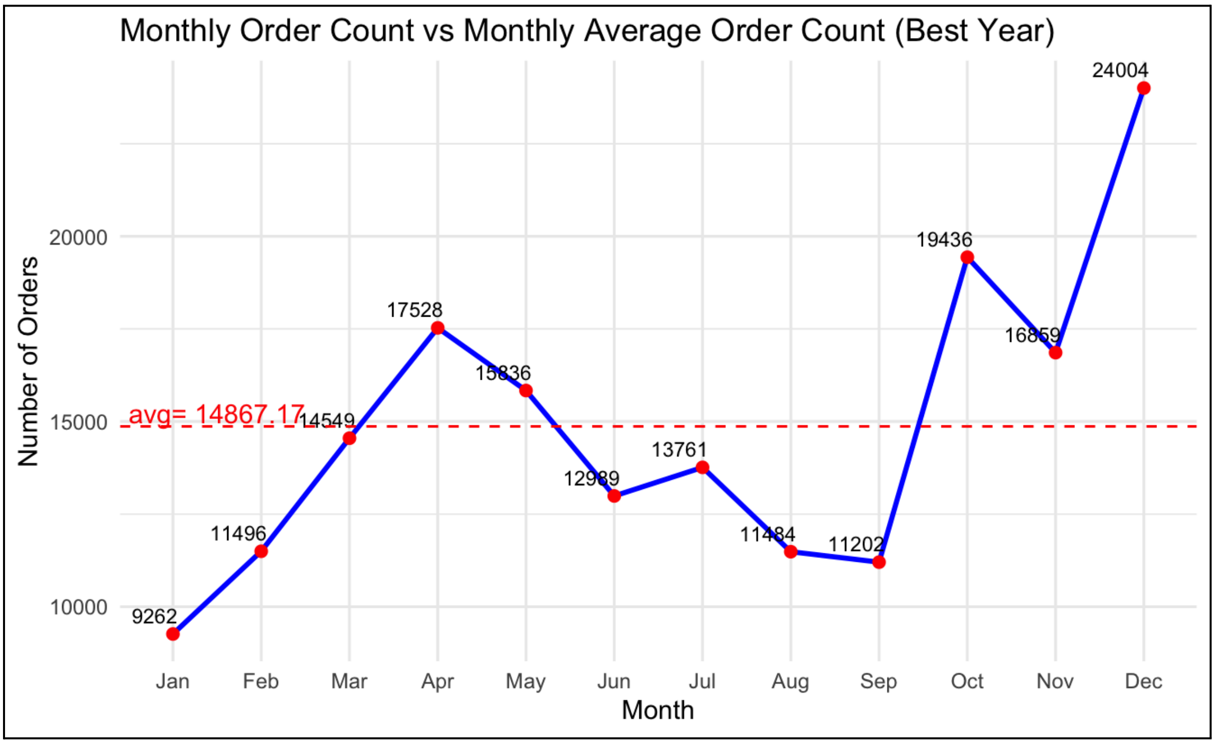
3. Top 10 products sold in the best year of sales



Essentials like AAA Batteries (4-pack) sold 31,012 units, indicating great demand. AA Batteries (4-pack) sold 27,635 units. At 23,971 and 23,211 units, USB-C and Lightning charging cables were popular. Wired Headphones (20,553), Apple Airpods (15,657 units), and Bose SoundSport Headphones (13,454) were top audio accessories. 27in FHD (7,547 units) and 27in 4K Gaming (6,243 units) monitors were popular for gaming and general use. The iPhone sold 6,847 units, demonstrating a robust high-end phone market.

Recommendation: These sales suggest a requirement for inventory diversity. The popularity of batteries and cables suggests opportunities for bundling or promotions. The strong performance of headphones and monitors indicates a growing demand for tech accessories, suggesting an expansion of these product lines.

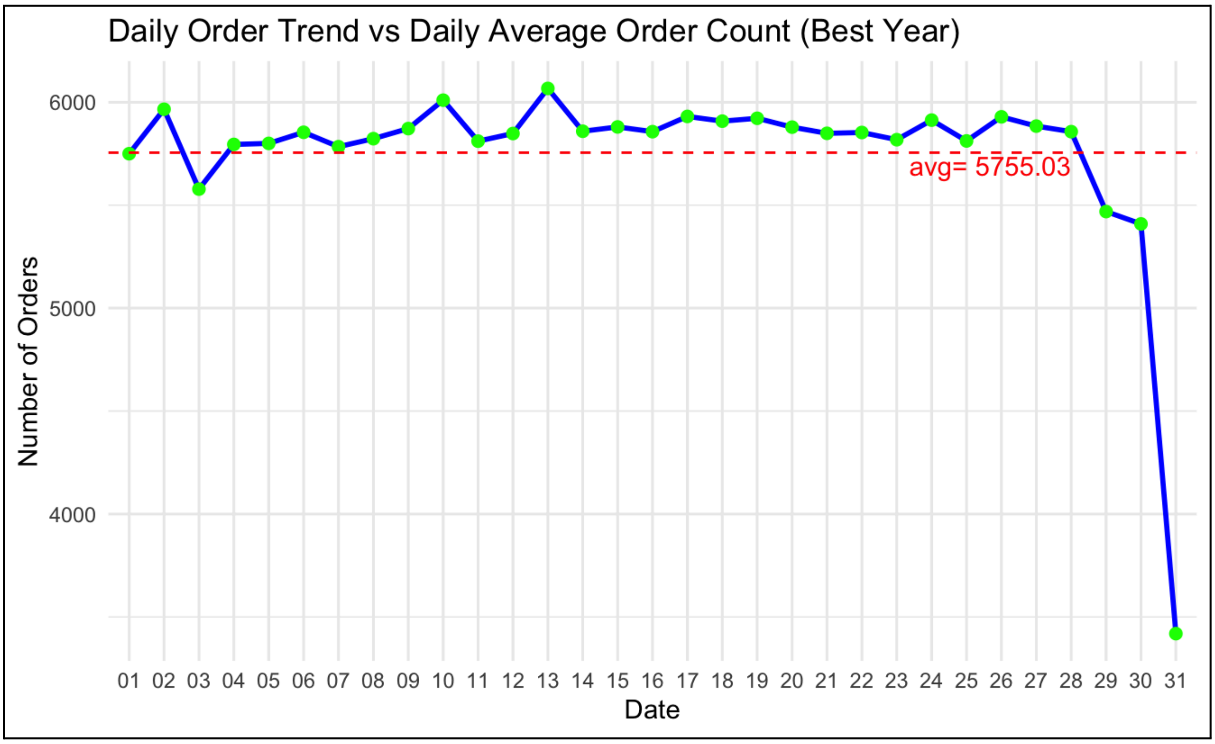
4. Monthly order trend vs monthly average order



December marks the highest (24,004), possibly due to holiday buying. October peaks with 19,436 orders, presumably from early Christmas specials. The lowest orders are in January (9,262) and September (11,120), reflecting post-holiday and late-summer slowdowns. The average monthly order is 14,867. Orders over average in April, May, October, November, and December indicate significant performance. January, February, March, June, July, August, and September had lower orders, suggesting weaker times.

Recommendation: Marketing during high-order months and focused ads during low-order months might balance order volume.

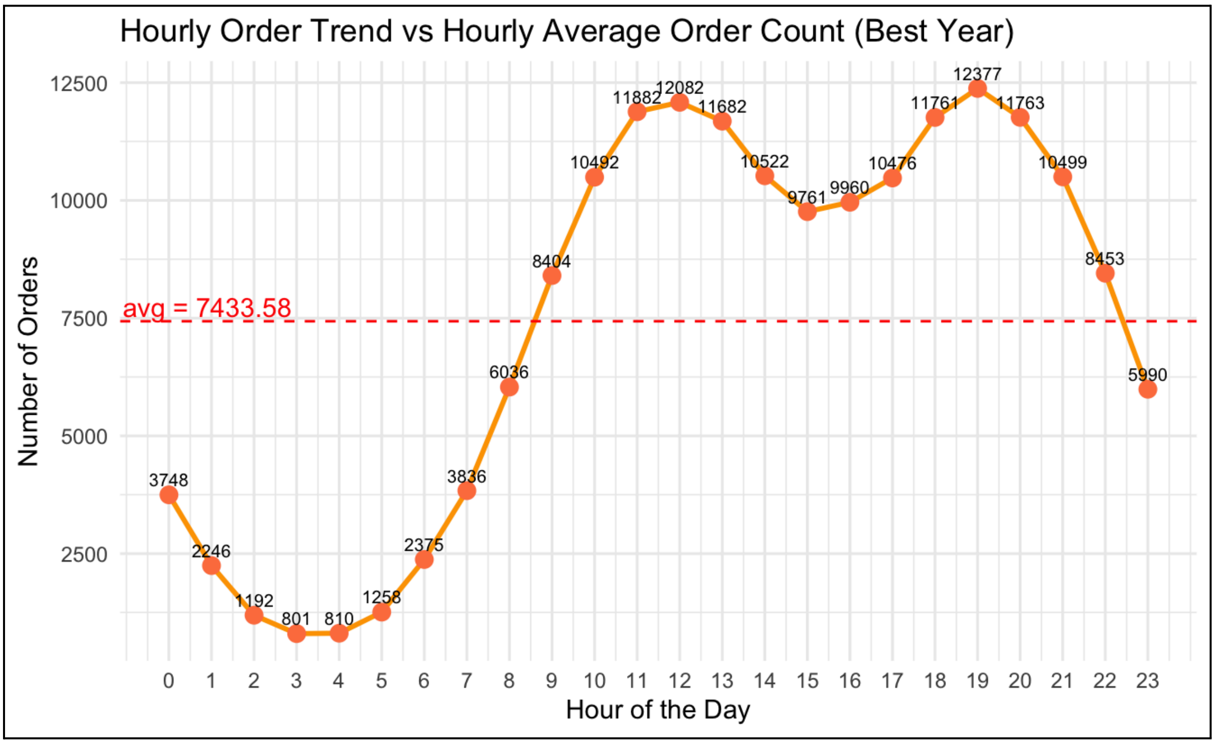
5. Daily order trend vs daily average



Daily orders are relatively stable, hovering around the average of 5,755.03 orders. There is a noticeable drop in orders towards the end of the month, particularly after the 26th, which could be due to month-end buying patterns or stock limitations. The consistency in daily orders suggests a steady demand, but the significant drop at month-end indicates potential issues with inventory or customer purchasing behavior.

Recommendation: Focusing on inventory management and end-of-month promotions could help mitigate this decline and maintain consistent order volumes throughout the month.

6. Hourly order trend vs hourly average order



Order activity peaked at 12,082 between 10 and 11 AM. Another surge totalling 12,377 orders appears between 6 and 7 PM, suggesting popular purchasing periods. The lowest order activity occurs between 2 AM and 4 AM, indicating poor customer activity. The average hourly order is 7,433.58. Orders are typically above this average from 8 AM to 9 PM, the busiest shopping hours.

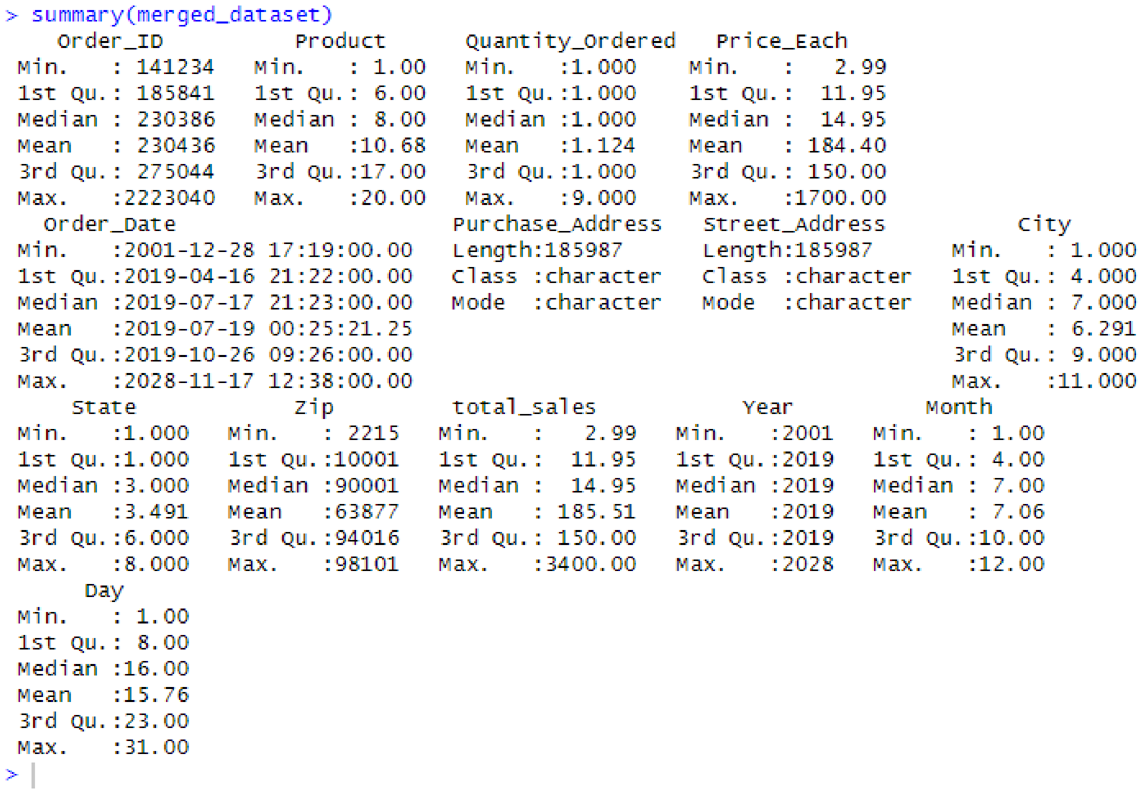
Recommendation: To boost revenue during peak times, focus on marketing and consumer involvement. Offer targeted specials during off-peak hours to boost daily purchase volume.

Building Predictive Model To Predict Future Sales

OBJECTIVE: Build a prediction model for Dibs which is useful for marketing and sales strategies. Predict Total Sales (A prediction problem)

Exploratory Data analysis and Data Preparation.

We added a total_sales (Price_Each * Quantity_Ordered) column for Dibs sales. Summary statistics and graphs explained data dispersion. We created categorical factors for better analysis. Years, months, and days were separated in Order_date. Splitting Address into City, State, and Zip columns generated factors. This examined if these factors impact total_sales projection. These characteristics' summary statistics follow.



```
> summary(merged_dataset)
    Order_ID           Product       Quantity_Ordered   Price_Each     
 Min.   : 141234   Min.   : 1.00   Min.   :1.000    Min.   :   2.99  
 1st Qu.: 185841   1st Qu.: 6.00   1st Qu.:1.000    1st Qu.:  11.95  
 Median : 230386   Median : 8.00   Median :1.000    Median :  14.95  
 Mean   : 230436   Mean   :10.68   Mean   :1.124    Mean   : 184.40  
 3rd Qu.: 275044   3rd Qu.:17.00   3rd Qu.:1.000    3rd Qu.: 150.00  
 Max.   :2223040   Max.   :20.00   Max.   :9.000    Max.   :1700.00  
   Order_Date                     Purchase_Address   Street_Address          City       
 Min.   :2001-12-28 17:19:00.00   Length:185987      Length:185987      Min.   : 1.000  
 1st Qu.:2019-04-16 21:22:00.00   Class :character   Class :character   1st Qu.: 4.000  
 Median :2019-07-17 21:23:00.00   Mode  :character   Mode  :character   Median : 7.000  
 Mean   :2019-07-19 00:25:21.25                                         Mean   : 6.291  
 3rd Qu.:2019-10-26 09:26:00.00                                         3rd Qu.: 9.000  
 Max.   :2028-11-17 12:38:00.00                                         Max.   :11.000  
     State            Zip          total_sales          Year          Month      
 Min.   :1.000   Min.   : 2215   Min.   :   2.99   Min.   :2001   Min.   : 1.00  
 1st Qu.:1.000   1st Qu.:10001   1st Qu.:  11.95   1st Qu.:2019   1st Qu.: 4.00  
 Median :3.000   Median :90001   Median :  14.95   Median :2019   Median : 7.00  
 Mean   :3.491   Mean   :63877   Mean   : 185.51   Mean   :2019   Mean   : 7.06  
 3rd Qu.:6.000   3rd Qu.:94016   3rd Qu.: 150.00   3rd Qu.:2019   3rd Qu.:10.00  
 Max.   :8.000   Max.   :98101   Max.   :3400.00   Max.   :2028   Max.   :12.00  
      Day       
 Min.   : 1.00  
 1st Qu.: 8.00  
 Median :16.00  
 Mean   :15.76  
 3rd Qu.:23.00  
 Max.   :31.00  
> 
```

Figure 1: Showing descriptive statistics of the dataset.

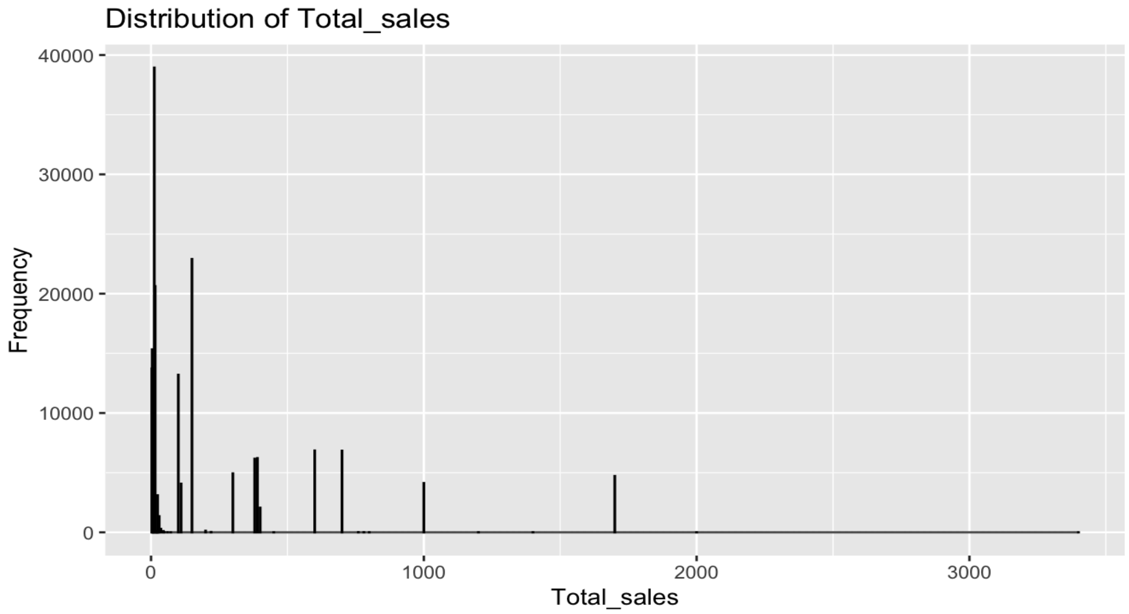


Figure 2: Showing distribution of sales data- Right Skewed.

As expected with time-series sales data, Figure 2 reveals right-skewed data. This skewness is predicted since some things sell better or cost more. Normalizing the data would hide price and demand effects and dispersion.

Identifying significant explanatory features.

Logically, certain dataset variables are useless for the prediction model and degrade its accuracy. We chose these columns as explanatory variables: Product, Quantity_Ordered, Year, Month, Day, City, State, Zip, Price_Each. Correlation matrix and Random Forest for variable significance examined these factors' relevance. The outcomes are:

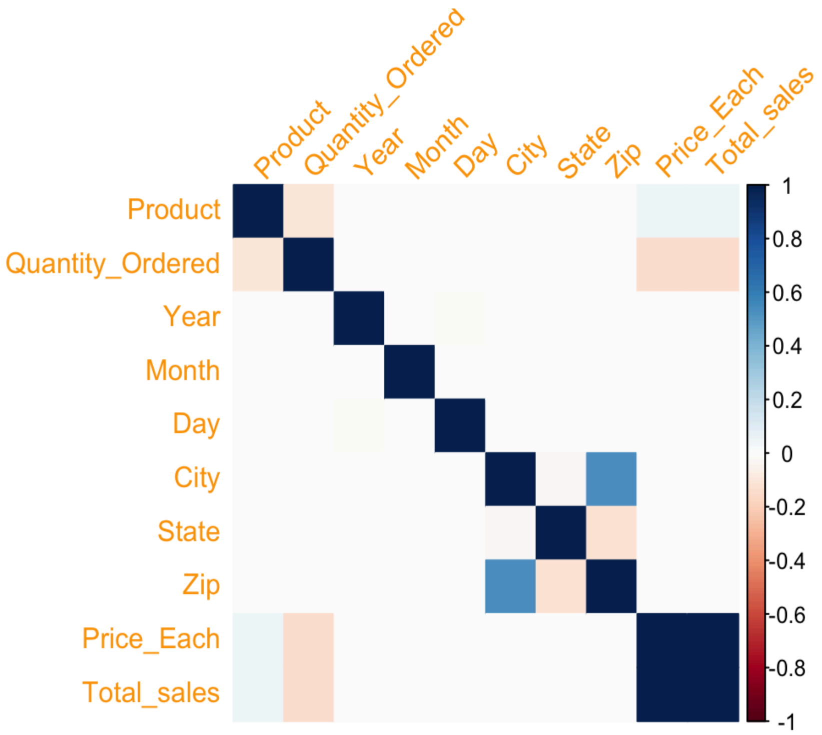
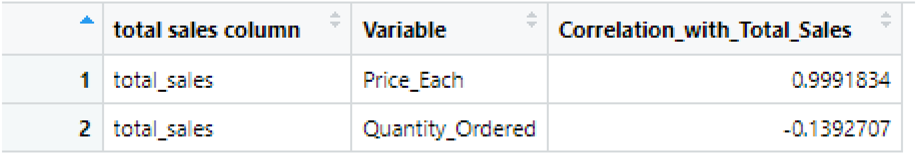


Figure 3: Correlation Heatmap between features

Figure 3's correlation heat map reveals that most variables in the dataset are unrelated to total sales. Figure 4 shows variables with correlations above 0.1. Quantity_Ordered correlates -13.92% with total_sales, but Price_Each correlates 99%.



	total sales column	Variable	Correlation_with_Total_Sales
1	total_sales	Price_Each	0.9991834
2	total_sales	Quantity_Ordered	-0.1392707

Figure 4: Correlation above abs(0.1) with total_sales

Thus according to this method, significant explanatory variables can be Price_each and Quantity_Ordered.

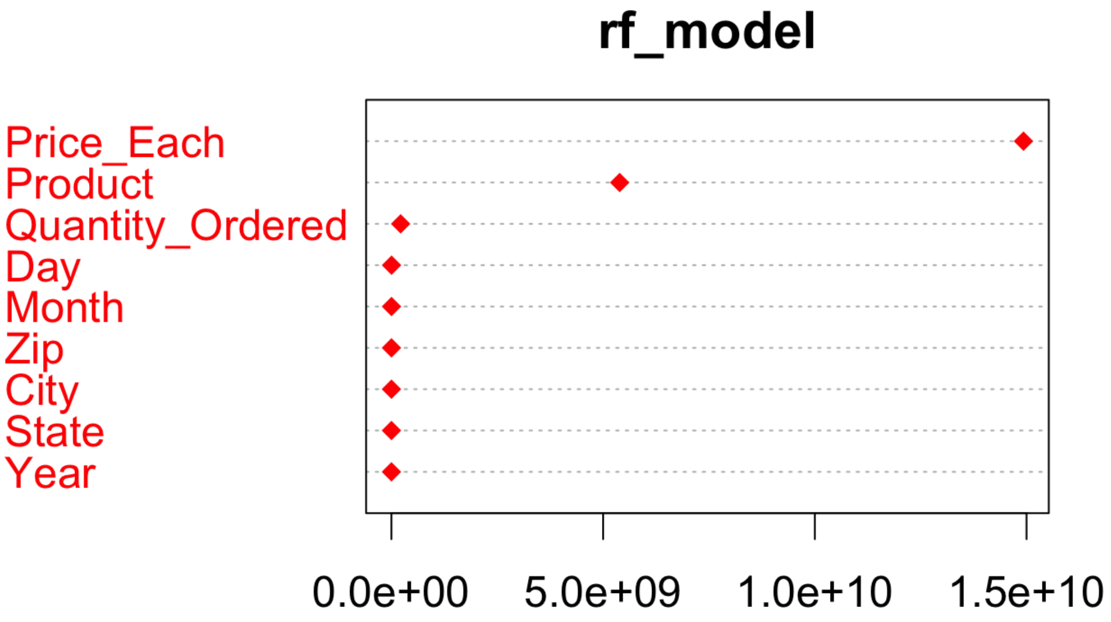


Figure 5: Variable Importance Plot using Random Forest

Since correlation alone can't indicate explanatory variable importance, we employed a random forest model, which is resilient to varied data structures and yields deep insights. Figure 5 displays the variable order of significance for forecasting total sales using the feature importance model. The regression model just needs Price_Each, Quantity_Ordered, and Products.

Choosing the Predictive Models

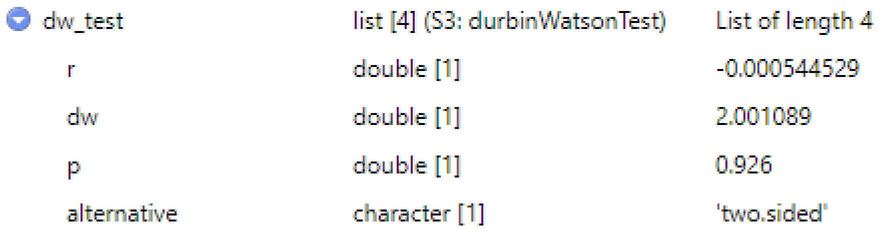
1) Linear Regression model

Total_sales = F(Product, Quantity_Ordered, Price_Each)

The time-series data and substantial explanatory factors with unambiguous correlations led us to use a linear regression model. This baseline model compares to more complicated ones. Checking linear regression model assumptions yielded these results:

- Check for Autocorrelation-

Autocorrelation shows whether explanatory variables are affected by their lagged values. A 2 autocorrelation means zero.



```
dw_test         list [4] (S3: durbinWatsonTest)   List of length 4
  r             double [1]                        -0.000544529
  dw            double [1]                        2.001089
  p             double [1]                        0.926
  alternative   character [1]                     'two.sided'
```

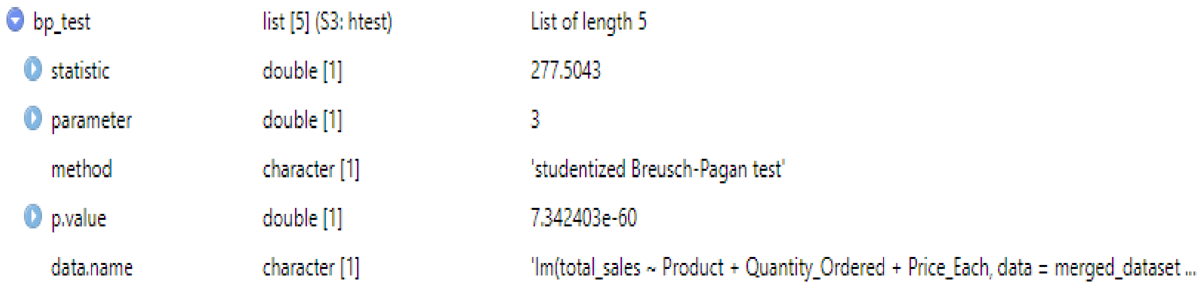
Figure 6: Durbin-Watson test statistic for Autocorrelation.

Since the DW test value is approximately 2, Figure 6 demonstrates that our model with total_sales as the prediction variable and Price_Each, Quantity_Ordered, and Products as explanatory variables has no autocorrelation.

- Check for Heteroskedasticity

To detect heteroskedasticity, check for higher values of explanatory factors that result in more mistakes or residuals.

Breusch Pagan Test:



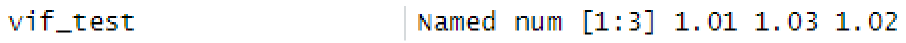
```
bp_test         list [5] (S3: htest)   List of length 5
  statistic     double [1]             277.5043
  parameter     double [1]             3
  method        character [1]          'studentized Breusch-Pagan test'
  p.value       double [1]             7.342403e-60
  data.name     character [1]          'lm(total_sales ~ Product + Quantity_Ordered + Price_Each, data = merged_dataset ...
```

Figure 7: Breusch Pagan test statistic for heteroskedasticity.

When p-value is below significance threshold, heteroskedasticity exists. The p-value is below 0.05 in Figure 7, showing heteroskedasticity.

- Test for Multicollinearity.

The Multicollinearity Test measures the correlation between explanatory factors using VIF values.



```
vif_test                      Named num [1:3] 1.01 1.03 1.02
```

High VIF scores suggest more association with other factors. Our variables have low VIF values near 1, indicating non collinearity. Our model is resilient because it meets autocorrelation and multicollinearity assumptions. The heteroscedastic data is typical of time-series data.

2) Random Forest.

Our second model is Random Forest. It can detect nonlinear predictor-response correlations that linear models overlook. This helps when few variables directly affect sales. Random Forest resists noise and handles unnecessary or duplicate features without degrading performance. Random Forests decrease overfitting and improve generalization to new data by averaging numerous decision trees.

Total_sales = F(*Product, Quantity_Ordered, Price_Each, City, State, Zip, Year, Month, Day)*

Splitting the dataset into Train and Test data

We split the dataset into 80% training and 20% testing data to prevent overfitting and evaluate our models.

Cross validation and Hyperparameter tuning

We utilized K-fold cross-validation to test the linear regression model, splitting the dataset into k equal sections. We train the model k times, using (k-1) folds as training and the remaining fold as testing.

To determine the number of decision trees the Random Forest will generate, we choose the hyperparameter “ntree” or “n_estimators.” Adding trees reduces variation and improves performance, but it increases computational complexity and training time.

Evaluation of the model-

We assess the model using Root Mean Squared Error (RMSE). RMSE indicates the model's data fit by measuring the average size of predicted-to-actual errors. Model performance improves with lower RMSE.

Comparison between models

Figure 8 demonstrates that the Random Forest model outperforms linear regression with a reduced RMSE (6.22) compared to 14.21. Since non-significant explanatory variables were eliminated previously, cross-validation had no effect on the linear regression model's initial high R-square value (0.99). Adjusting the number of decision trees enhanced the Random Forest model's performance, lowering its RMSE to 0.2.

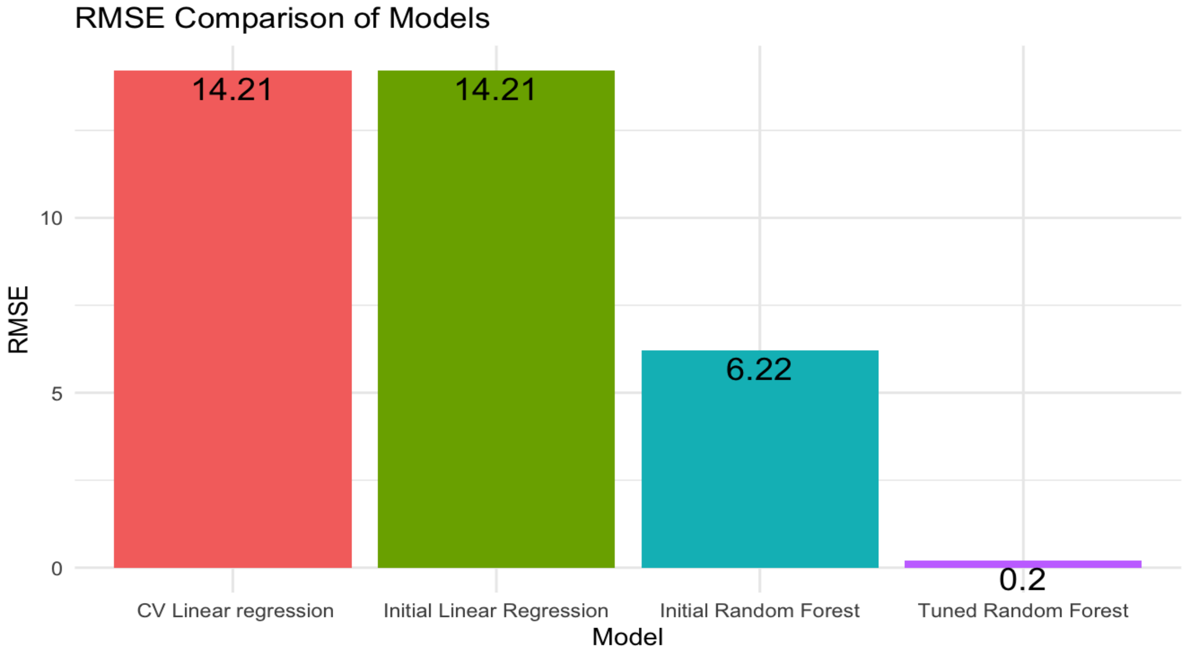


Figure 8: Comparison of both models based on RMSE

Recommendations

Since present data lacks sales-related elements, the organization needs to gather target variable-correlated data. We advocate gathering product feedback on pricing, quality, satisfaction, and usefulness for consumer loyalty. This data helps identify loyal clients and understand their needs.

Conclusion

This research analyzed Dibs Retail's sales data to suggest ways to boost sales and customer loyalty. Accurate data cleansing revealed the best sales year (2019), ideal advertising hours, and top-performing goods.

Sales trends, state-wise performance, and product appeal were visualized for marketing and operations strategy. Our predictive algorithms, notably the random forest model, accurately predicted sales, assisting strategic planning.

Optimizing techniques involve using peak sales periods and top items. This data-driven strategy helps Dibs expand and retain customers.