Sales Data Analysis & Predictive Modeling

Zoljargal Batbaatar

Introduction

Dibs, an online retail company selling accessories and home goods, is facing challenges in increasing sales and ensuring customer loyalty. The company aims to gain a better understanding of their customers using data from purchase histories. In this report, we used R Studio and various business analytics techniques to analyze the existing data, identify trends, and provide insights and recommendations to enhance sales performance.

Task 1 - Data cleaning

Firstly, the data cleaning process involved column by column observation. After reading and binding all datasets, checked and cleaned duplicated column names, as the duplicated labels do not contain any value. When checking missing values, observed and cleaned 3,289 missing values. For further clarity and analysis convenience, changed all column names into lowercase and snake case:

Original column name	Changed column name
Order ID	order_id
Product	product
Quantity Ordered	quantity_ordered
Price Each	price
Order Date	order_date
Purchase Address	purchase_address

Order ID column observations:

An “order_id” is the number system used to keep track of orders. Each order receives its own order ID that will not be duplicated. The “order_id” column should not be duplicated, but found a total of 7,537 duplicates. However, after examining, found that those duplicated IDs have different products, thus assuming a single order may contain several different products, leaving the “order_id” column as it is.

Product column observations:

After checking “product” column unique values, we identified several misspelling errors. Therefore, mutated the misspelling and changed system/fault errors into NAs. Newly issued NAs had only 2 entries thus deleted them.

Original values	Mutated values
IPhone	iPhone
USBC Charging Cable	USB-C Charging Cable
LightCharging Cable	Lightning Charging Cable
AAA Batteries (4pack)	AAA Batteries (4-pack)
Goo0gle Phone	Google Phone
Wired Headphoness	Wired Headphones
##system error##	NA
Fault error	NA

Quantity ordered column observations:
Firstly, changed the character mode of the column into numerical value. When checking the unique value of the "quantity_ordered" column, found that there is 0 value. Since the already made order quantity cannot be equal to zero, I cleaned the rows where equals 0.

Price column observations:
Firstly, changed type from character into numerical value and checked for unique values. When checking unique values we discovered NAs. Thus, imputed NA values with the mean price of corresponding products.

Order date column observations:
Since the "order_date" column contains both data and time, divided it into 2 separate columns "date" and "time" with the date and time format and deleted the original column. Since these columns are in date and time format it is now possible to extract unique year values (2019, 2021, 2028, 2001, 2020). I counted 2001 and 2028 years entries, each of which had only 1 entry, thus, deleted rows with these entries.

Purchase address column observations:
Firstly, I splitted the column into 4 separate columns: "address", "city", "state", "postcode". Then checked unique values of city and state, fixed some spelling errors.

Original values	Mutated values
SanFrancisco	San Francisco
Las Angeles	Los Angeles

Task 2 - Data Investigation

In this part, we investigated Dibs Retail Company's customer purchase history data from 2019 to 2021.

A. Dibs earned $34280627 in 2019, $11905.08 in 2020 and $4365.56 in 2021. 2021 is marking the year with the lowest sales.

B. The best year for sales is 2019. $34280627 was earned.

C. December was the most successful month in 2019.

D. In December 2019, $4613443 was earned, marking the best year of best month sales.

E. San Francisco city had the highest sales of 8259719 in 2019.

F. The highest sale in 2019 occurred at 19:01:00, indicating that this time, or slightly earlier, could be strategically advantageous for Dibs to schedule their display advertisements. Displaying advertisements shortly before this peak buying period maximizes the likelihood of engaging customers when they are most receptive to making purchases. This timing could significantly enhance the effectiveness of their advertising strategy.

G. The iPhone and Lightning Charging Cable were the most frequently sold items together from 2019 to 2021, with a total of 891 bundles sold.

H. AAA Batteries (4-pack) are sold the most (31020) the time period.

AAA Batteries (4-pack) are essential for household devices, including TV remotes, electronic toys, flashlights, and other everyday gadgets. These devices often require frequent battery replacements due to their limited lifespan. Purchasing batteries in 4-packs provides convenience and ensures a ready supply, saving customers the hassle of frequent purchases.

Additionally, the bulk packaging of these batteries offers a cost-effective solution, reducing the cost per unit and offering greater value to consumers.

I. In 2019, 646 LG Dryers were sold, making it the least sold product in the best year for sales.

Task 3 - Data Analysis and Visualisation

Based on the data provided by Dibs, we determined key trends and patterns in the sales data, providing insights into the company's performance and identifying opportunities for improvement.

A. Monthly sales trend vs monthly average sales

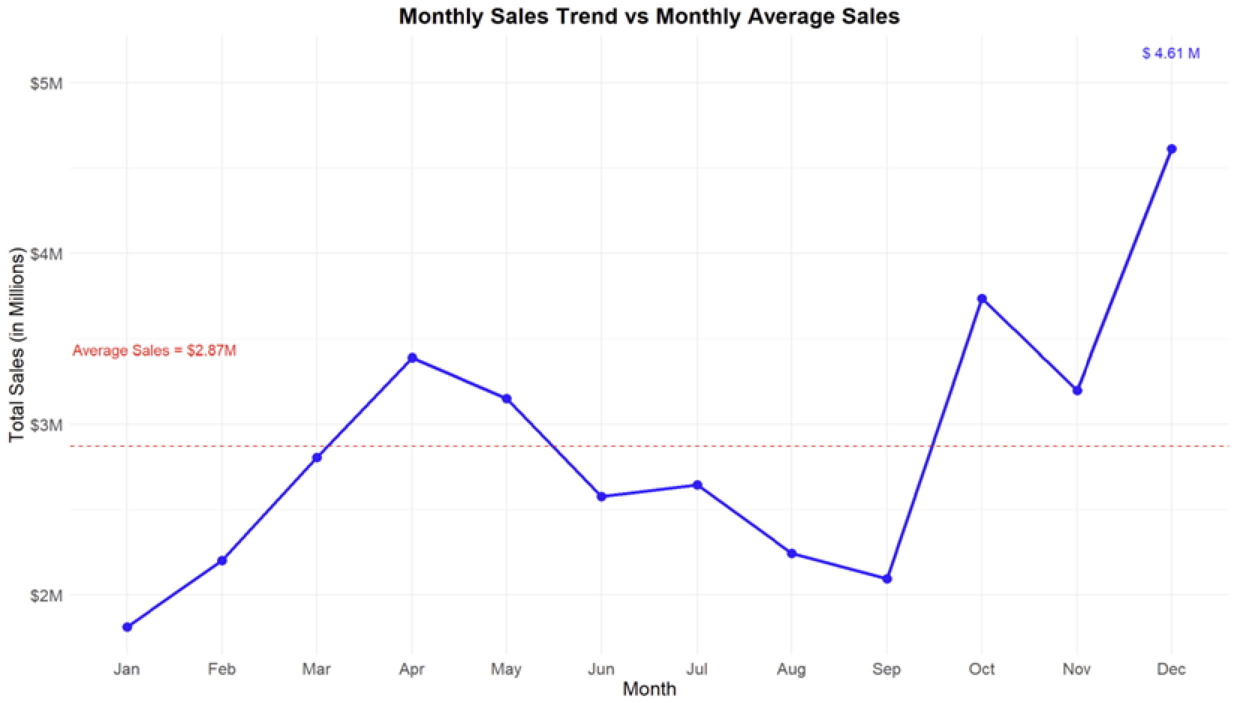

Figure 1: Dibs Monthly Sales for 2019

The sales for 2019 were highest in April, October, and December, suggesting these months had higher consumer activity, likely due to successful promotions or seasonal demand. The average monthly sales (presented in a red line) were $2.8 million while the highest monthly sales were $ 4.61 million for December. On the other hand, sales were lowest in January, February and September, indicating periods of reduced consumer spending or a need for improved marketing strategies during these months.

B. Sales by state

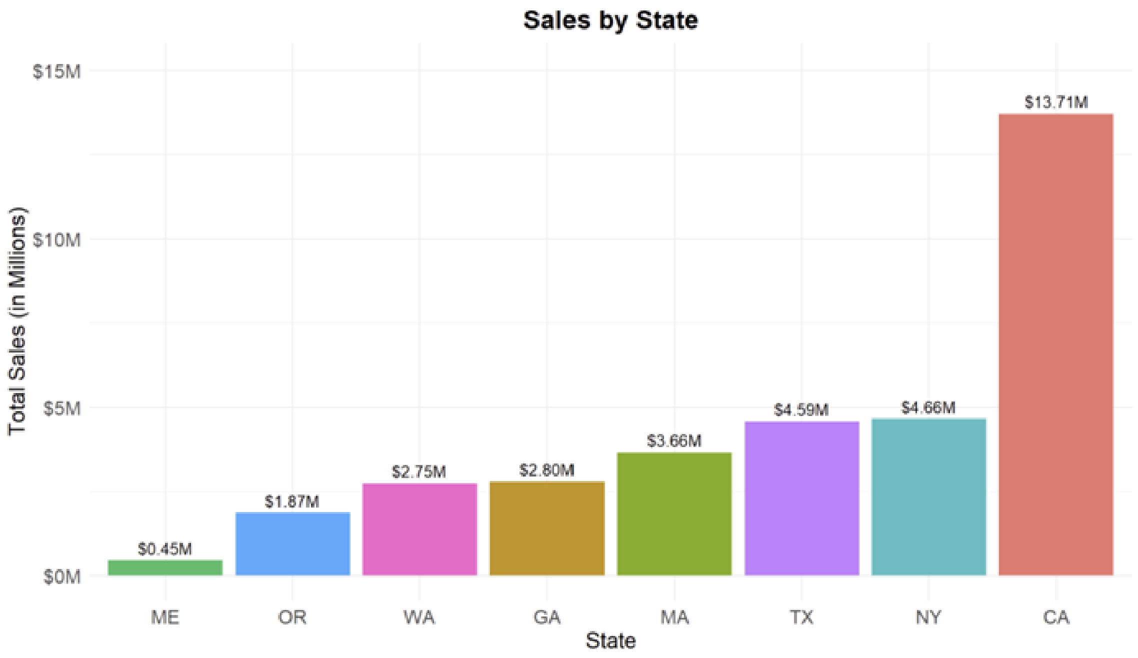

Figure 2: Sales per State for 2019

California stands out with the highest sales with $ 13.71 million, indicating a strong customer base and high demand. New York, Texas, and Massachusetts also show substantial sales, contributing significantly to overall revenue. Conversely, states like Maine and Oregon have relatively lower sales, highlighting potential areas for growth or the need for improved marketing strategies.

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

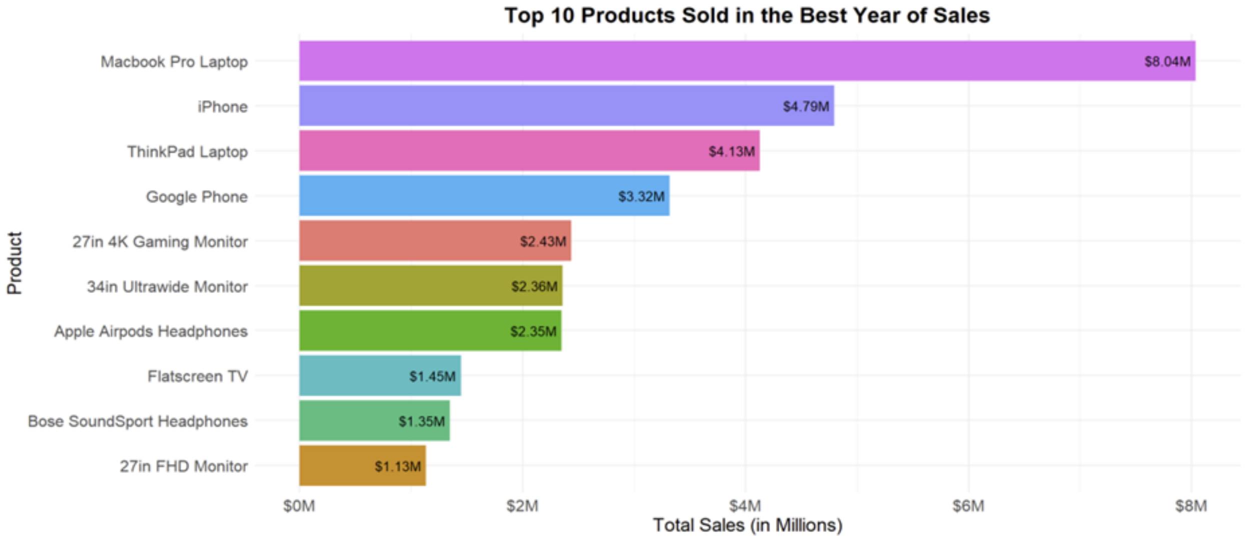

Figure 3: Top Ten Products according to total sales

The "Top 10 Products Sold" graph reveals the highest-selling products, which are key revenue contributors. The Macbook Pro Laptop leads as the top-selling product, followed by the iPhone and ThinkPad Laptop, all contributing significantly to sales. Other high-performing products include the Google Phone, 27in 4K Gaming Monitor, and 34in Ultrawide Monitor.

D. Monthly order trend vs monthly average order

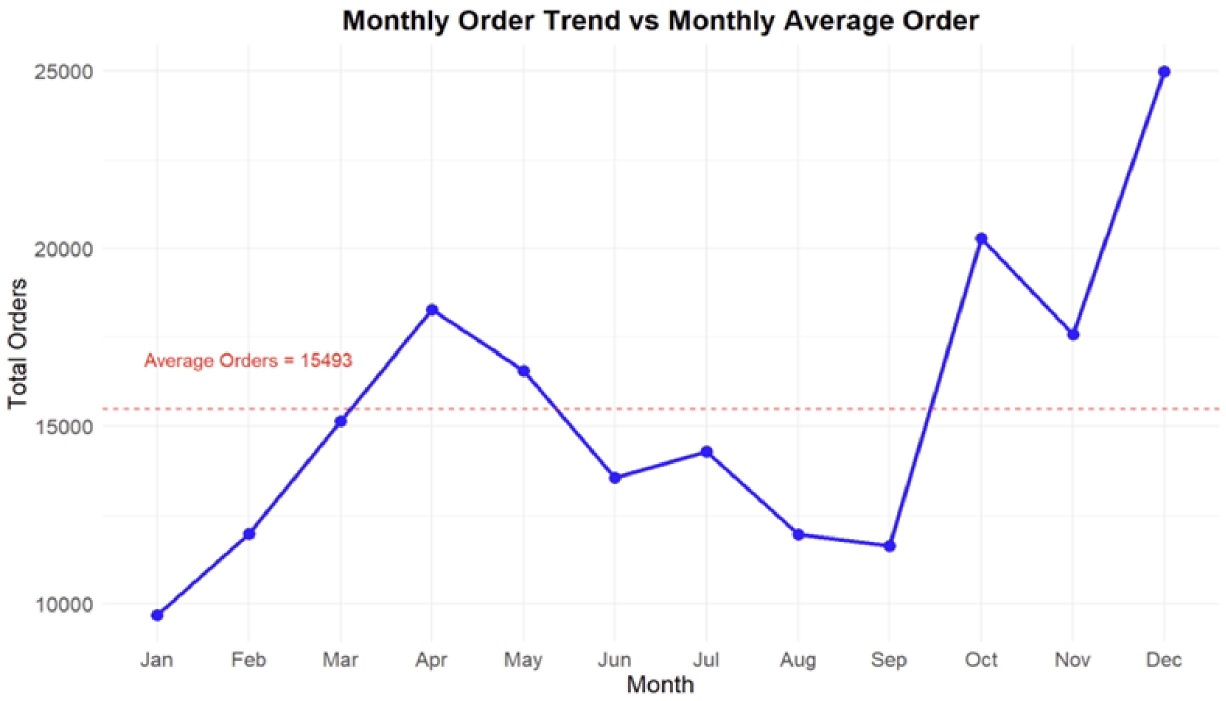

Figure 4: Dibs Monthly Order Trend for 2019

The average monthly orders, represented by the red dashed line, stand at 15,493. The highest monthly orders were recorded in December, with a total of over 25,000 orders. This trend mirrors the “Monthly sales graph” indicating consistent consumer behavior in order volume and sales value.

E. Daily order trend vs daily average

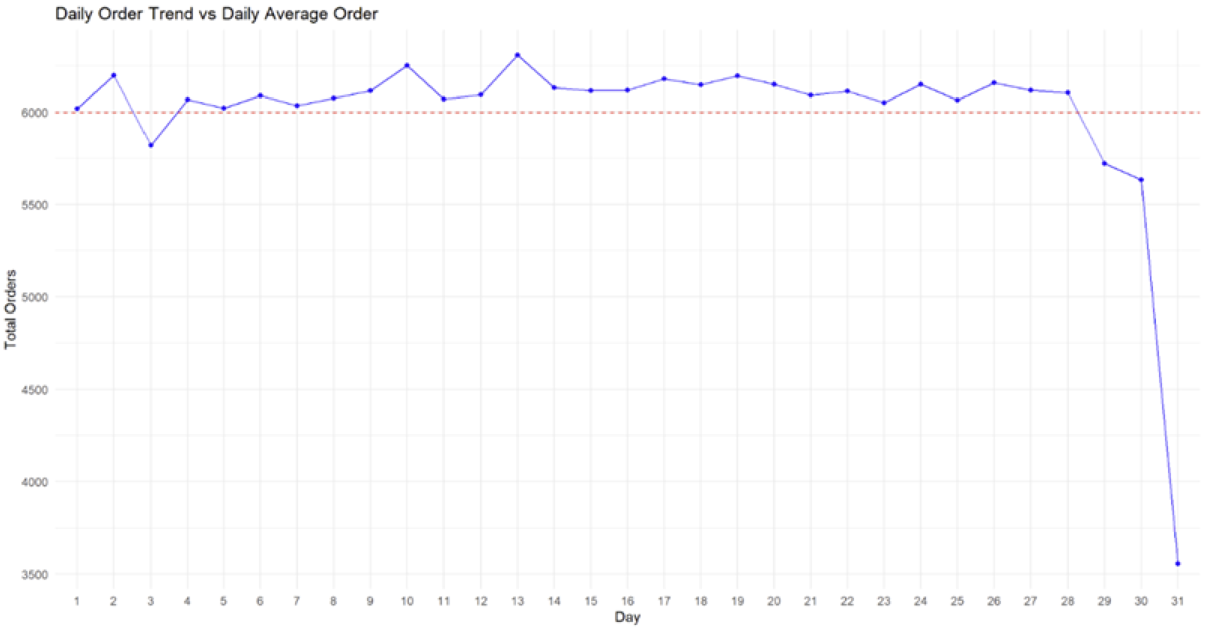

Figure 5: Dibs Daily Order Trend for 2019

The graph shows consistent daily orders throughout the month, generally around the average line. Minor peaks occur around the 4th and 18th, while a significant drop is noted on the last day of the month.

F. Hourly order trend vs hourly average order

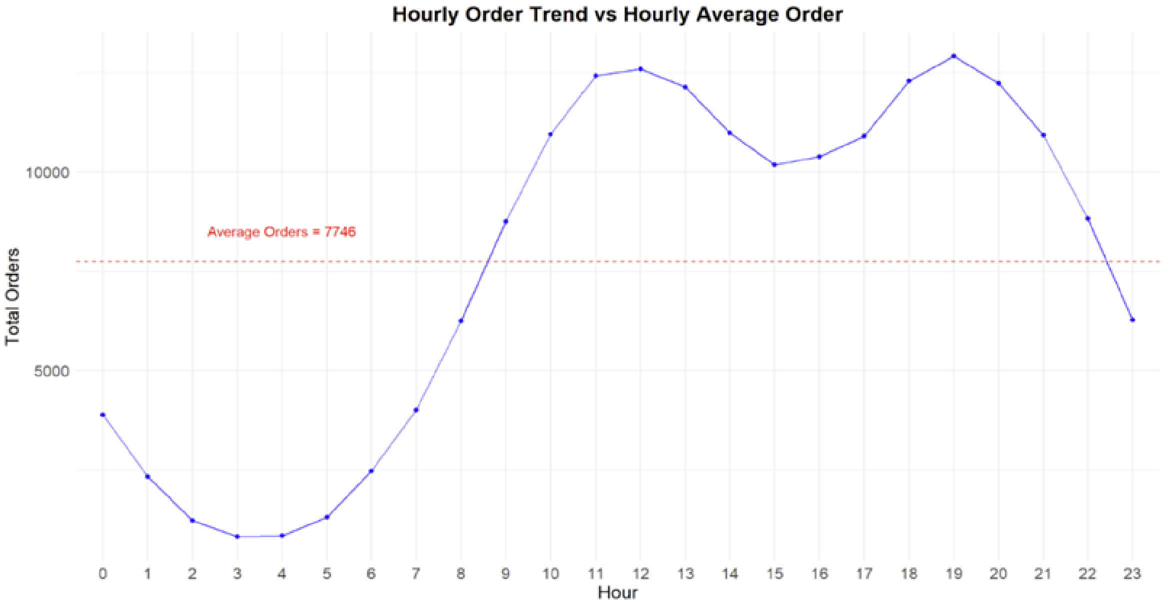

Figure 6: Dibs Hourly Order Trend for 2019

The graph shows clear trends in hourly order activity. Early morning hours (midnight to 6 AM) exhibit the lowest order numbers, reflecting typical consumer behavior. Orders increase around 8 AM, peaking between 10 AM and 3 PM, indicating the most active shopping hours. Another peak occurs between 5 PM and 9 PM, followed by a sharp drop after 9 PM.

Task 4 - Predictive Modelling

Objective:

Build a predictive model for Dibs organization to predict future sales.

Data Preparation:

As the objective is to predict future sales, the prediction variable used will be ‘sales’, created in Task 2 which is quantity_ordered *x* price. This added variable gives us the actual sales value for each order.

The data was split into two sets using a 90/10 split:

- Training Data: Trains our model based on known data
- Testing Data: Used to test the accuracy of our trained data

Model Selection:

For this analysis, linear regression and decision trees were used.

Linear Regression provides a solid foundation for predicting sales data such as this, although being a basic model, it is usually a good place to start and then build up to more advanced models. Using a decision tree as our second model builds upon the linear regression's accuracy with a more complex and robust prediction model which can capture more complex relationships if there are any.

Model Results:

Linear Regression:

- RMSE = 11.32

Decision Tree:

- RMSE = 24.59

Having an RMSE of 11.32 for our linear regression model suggests that this model was accurate in predicting sales to a +- 11.33 margin of error. In the context of sales ranging from 2.99 - 1700, this is an incredibly accurate model for predicting sales. Comparing these results to our decision tree resulted in an RMSE of 24.59, although still bearing a great result, it is over double the margin of error of the linear regression model.

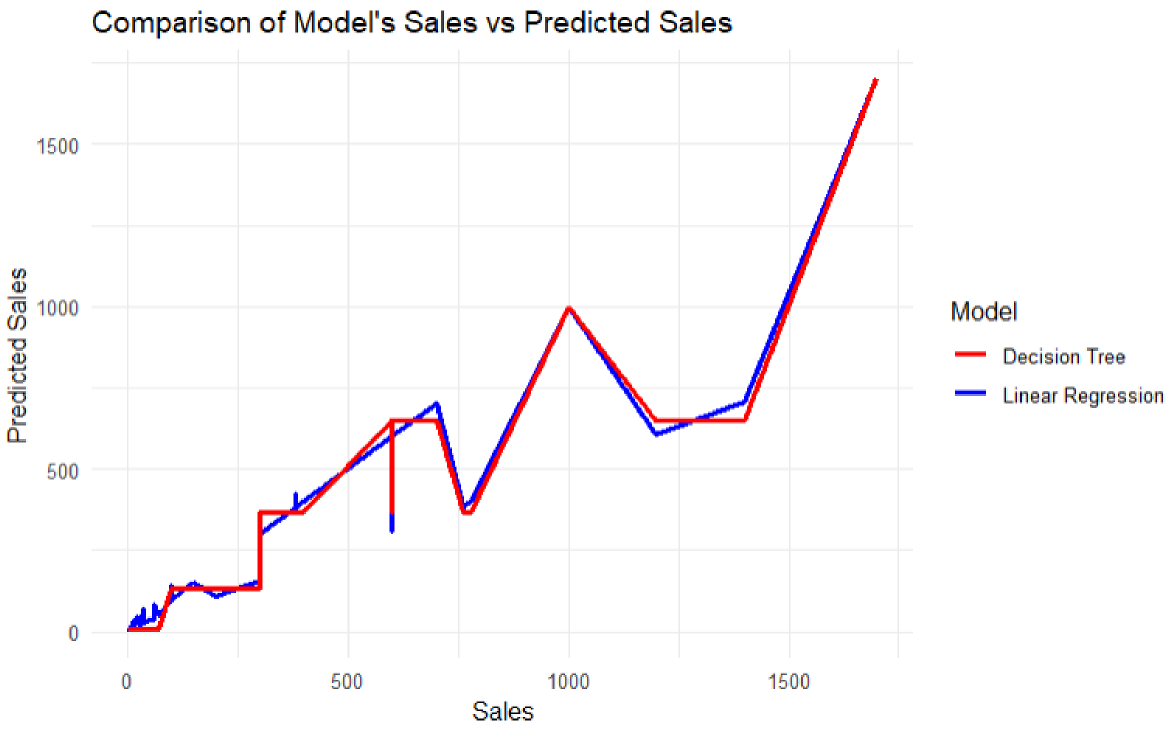

Figure 7: Comparison of models

Figure x shows the two models' prediction accuracies of sales against predicted sales, results show that they both perform similarly with a slight variation in the decision tree which supports the RMSE score.

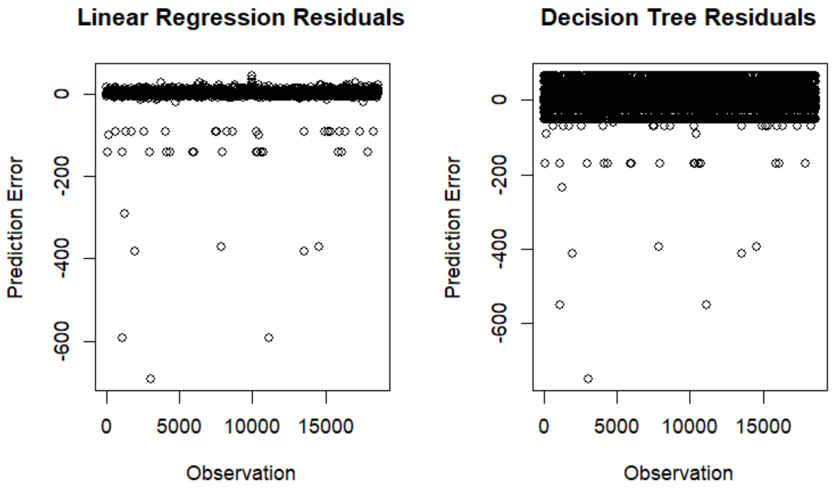

Figure 8: Model's Residuals

By using the above scatterplot, we can further analyze our models by visualizing each model's residuals:

- Linear Regression: Tightly space residuals, suggesting higher accuracy and lower prediction error.
- Decision Tree: More spread out residuals, suggesting lower accuracy and higher prediction error.

Although a simpler model, the linear regression model demonstrates higher accuracy and fewer margins of error, suggesting that the predictor variables have a linear relationship with the prediction variable. This supports the case for the decision tree performing worse, as the decision tree model predicts relationships in a more stepwise approach rather than forming linear relations.

Based on these findings, it is recommended that Dibs use the linear regression model when aiming to predict sales, as it is more accurate as shown in our RMSE and visual representations of the differences between actual sales and prediction sales.

Recommendation for Dibs:

To enhance sales and market presence, Dibs could implement a multi-faceted strategy based on the following insights:

1. **Targeted Marketing and Promotions**:
   - Invest in targeted marketing campaigns, special promotions, and customer loyalty programs in high-performing states such as California, New York, Texas, and Massachusetts.
   - Tailor marketing efforts and product offerings in states with lower sales, like Maine and Oregon, to attract more customers and increase sales.
   - Analyze successful strategies from high-performing states and replicate them in other regions to expand market presence.
2. **Product Strategy**:
   - Continue promoting and stocking top-selling products like the Macbook Pro Laptop and iPhone while exploring strategies to boost sales of lower-performing items.
   - Use insights from best sellers to guide future product development and inventory decisions.
3. **Order Activity and Trends**:
   - Investigate the significant drop in orders on the last day of the month to determine if it’s due to system issues, reporting cutoffs, or specific events.
   - Maintain consistent daily orders by leveraging the stable customer base to boost orders on below-average days.
   - Implement daily promotions to increase order volumes on traditionally lower activity days, smoothing out fluctuations and driving overall growth.
4. **Hourly Order Optimization**:
   - Focus marketing efforts, special promotions, and flash sales during peak hours (10 AM to 3 PM and 5 PM to 9 PM) to maximize order volumes.
   - Offer time-sensitive discounts during low activity hours (midnight to 6 AM) to drive orders and utilize downtime effectively.
   - Ensure adequate staffing and resources during peak hours to handle increased order volumes efficiently and maintain high customer satisfaction.

By adopting these recommendations, Dibs can capitalize on current insights, optimize sales strategies, and drive sustained growth.