Table of content

Executive Summary

With the objective of supporting the Dibs retailer to increase sales and customer loyalty, this report aims to provide the company with extensive data insights and predictive models, then recommend some solutions to finish the objectives. The report is divided into 4 key tasks, including: Cleaning data, Descriptive Analytics, Visualisation, and Building Predictive Models. Particularly, the first task focuses on exploring and cleaning data to have a general view and reveal hidden patterns, trends in the data analysis. The second and third tasks will show the descriptive analysis and visualisation of the data, aiming to answer all the crucial questions given, such as month of best sale during the trend, etc. Finally, task 4 will focus on building predictive models for Dibs organisation to be able to predict future sales, specifically using K nearest neighbours and Linear regression models. The recommendations laid out in this report are designed to equip Dibs with the tools necessary for optimising their marketing efforts, improving customer engagement, and ultimately driving significant growth in sales and customer loyalty.

1. Introduction

Background

Dibs, an expanding online retailer, specialising in selling accessories, home goods and electronics, is facing challenges in boosting sales and customer loyalty. To address these challenges, our team has been commissioned to conduct a thorough analysis of the available data with the goal of uncovering underlying patterns and trends in customer behaviour.

Methodology

To perform data pre-processing and subsequent analysis, this project will utilise RStudio, employing both descriptive and inferential statistics to uncover underlying trends and preferences. Initially, the project utilises several R libraries, like tidyverse, dplyr, ggplot2, etc. All files CSV of data sales monthly will be combined for the whole team to work on. For the predictive models, due to high variation in daily sales, the K-nearest neighbour prediction model is used with 3 groups. For the sales prediction, we utilised linear regression and regression tree in this analysis. Year, Month, Date and Total Quantity Ordered are independent variables, in which Total Quantity Ordered is the only non-factor variable. Limitations of this study include its focus on urban populations which may not represent rural consumer behaviour.

2. Data Cleaning and Preprocessing

The raw data had 186,894 observations and 6 columns, containing several issues that need resolution before effective analysis. There are also inconsistencies in formatting, especially with order_date' and 'purchase_address'. Addressing these challenges is vital to ensure the dependability of our analytical results and recommendations.

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
1	141234	iPhone	1	700	01/22/19 21:25	944 Walnut St, Boston, MA 02215
2	141235	Lightning Charging Cable	1	14.95	01/28/19 14:15	185 Maple St, Portland, OR 97035
3	141236	Wired Headphones	2	11.99	01/17/19 13:33	538 Adams St, San Francisco, CA 94016
4	141237	27in FHD Monitor	1	149.99	1/05/2019 20:33	738 10th St, Los Angeles, CA 90001
5	141238	Wired Headphones	1	11.99	01/25/19 11:59	387 10th St, Austin, TX 73301
6	141239	AAA Batteries (4-pack)	1	2.99	01/29/19 20:22	775 Willow St, San Francisco, CA 94016
7	141240	27in 4K Gaming Monitor	1	389.99	01/26/19 12:16	979 Park St, Los Angeles, CA 90001
8	141241	USB-C Charging Cable	1	11.95	1/05/2019 12:04	181 6th St, San Francisco, CA 94016
9	141242	Bose SoundSport Headphones	1	99.99	1/01/2019 10:30	867 Willow St, Los Angeles, CA 90001
10	141243	Apple Airpods Headphones	1	150	01/22/19 21:20	657 Johnson St, San Francisco, CA 94016
11	141244	Apple Airpods Headphones	1	150	1/07/2019 11:29	492 Walnut St, San Francisco, CA 94016
12	141245	Macbook Pro Laptop	1	1700	01/31/19 10:12	322 6th St, San Francisco, CA 94016
13	141246	AAA Batteries (4-pack)	3	2.99	1/09/2019 18:57	618 7th St, Los Angeles, CA 90001
14	141247	27in FHD Monitor	1	149.99	01/25/19 19:19	512 Wilson St, San Francisco, CA 94016
15	141248	Flatscreen TV	1	300	1/03/2019 21:54	363 Spruce St, Austin, TX 73301
16	141249	27in FHD Monitor	1	149.99	1/05/2019 17:20	440 Cedar St, Portland, OR 97035
17	141250	Vareebadd Phone	1	400	1/10/2019 11:20	471 Center St, Los Angeles, CA 90001
18	141251	Apple Airpods Headphones	1	150	01/24/19 08:13	414 Walnut St, Boston, MA 02215
19	141252	USB-C Charging Cable	1	11.95	01/30/19 09:28	220 9th St, Los Angeles, CA 90001
20	141253	AA Batteries (4-pack)	1	3.84	01/17/19 00:09	385 11th St, Atlanta, GA 30301

Figure 2.1 Data before cleaning

Following this, we start cleaning the data by loading and combining the data sources provided and then work on each column, the task will go through the steps of handling missing values, converting data type format, splitting values ('order_date' and 'purchase_address'). As it initially appeared there were no missing values, closer inspection revealed empty strings in critical columns like 'order_id', 'product', 'price_each', 'order_date', and 'purchase_address', which we identified as missing data. To rectify this, we removed rows with incomplete data, resulting in a refined dataset of 186,311 rows and 6 columns. We also convert those columns into the correct datatype.

Particularly, for column 'order_date', which contains the missing values from years 2001, 2028 and January of 2020. Additionally, this column and 'purchase_address' are splitted into different columns. Before creating the new cleaned dataset, we have cleaned all the missing values,

reformatted the data and column name. To sum up, the original dataset had 186,894 observations of 6 variables, when after cleaning, it had 185,951 observations and 13 variables, with additional columns of street, city, state, time, year, month and day, and the postal codes are removed for the correctness and completeness of the data set.

	order_id	product	quantity_ordered	price_each	order_date	street	city	state	time	year	month	day
1	141254	AAA Batteries (4-pack)	1	2.99	2019-01-08	238 Sunset St	Seattle	WA	11:51	2019	1	8
2	141283	Flatscreen TV	1	300.00	2019-01-02	68 Hickory St	Seattle	WA	16:16	2019	1	2
3	141285	AAA Batteries (4-pack)	3	2.99	2019-01-14	447 Cedar St	Seattle	WA	14:13	2019	1	14
4	141303	AA Batteries (4-pack)	1	3.84	2019-01-19	313 14th St	Seattle	WA	09:23	2019	1	19
5	141315	USB-C Charging Cable	1	11.95	2019-01-10	842 8th St	Seattle	WA	01:32	2019	1	10
6	141316	AAA Batteries (4-pack)	3	2.99	2019-01-01	235 South St	Seattle	WA	07:26	2019	1	1
7	141328	ThinkPad Laptop	1	999.99	2019-01-06	736 5th St	Seattle	WA	23:18	2019	1	6
8	141342	Wired Headphones	1	11.99	2019-01-26	119 9th St	Seattle	WA	21:13	2019	1	26
9	141359	AAA Batteries (4-pack)	2	2.99	2019-01-09	383 11th St	Seattle	WA	22:21	2019	1	9
10	141366	Flatscreen TV	1	300.00	2019-01-17	803 Church St	Seattle	WA	22:34	2019	1	17
11	141370	AAA Batteries (4-pack)	2	2.99	2019-01-24	375 Center St	Seattle	WA	20:53	2019	1	24
12	141392	Apple Airpods Headphones	1	150.00	2019-01-25	755 Cherry St	Seattle	WA	13:51	2019	1	25
13	141406	USB-C Charging Cable	1	11.95	2019-01-28	110 13th St	Seattle	WA	12:02	2019	1	28
14	141438	Wired Headphones	2	11.99	2019-01-29	331 Center St	Seattle	WA	17:36	2019	1	29
15	141446	Bose SoundSport Headphones	1	99.99	2019-01-16	150 14th St	Seattle	WA	20:03	2019	1	16
16	141457	iPhone	1	700.00	2019-01-09	820 Jackson St	Seattle	WA	22:11	2019	1	9
17	141457	Apple Airpods Headphones	1	150.00	2019-01-09	820 Jackson St	Seattle	WA	22:11	2019	1	9
18	141466	AAA Batteries (4-pack)	2	2.99	2019-01-24	269 4th St	Seattle	WA	16:09	2019	1	24
19	141504	iPhone	1	700.00	2019-01-13	855 11th St	Seattle	WA	13:37	2019	1	13
20	141515	34in Ultrawide Monitor	1	379.99	2019-01-12	179 Highland St	Seattle	WA	16:34	2019	1	12

Figure 2.2 Data after cleaning

3. Performance Analysis

3.1 Descriptive Analysis

a. What is the worst year of sales and how much sales was earned?

2021 was the worst year of sales among 3 years at $3,927.

b. How much was earned in the best Year of sales?

The data revealed that the highest sales were recorded in 2019, amounting to $34,483,366.

c. In the best year of sales, which was the best month for sales?

In 2019, December emerged as the peak month for sales.

d. In the best year of sales how much was earned in the best month?

In 2019, December contributed $4,613,443, or 13.38% of the total sales.

e. Which City had the most sales in the best year of sales?

San Francisco was identified as the city with the highest sales, totaling $8,259,719.

f. To maximise the likelihood of customers buying a product, what time should Dibs business displaying advertisements in the best year of sales?

A closer look at the hourly sales data indicated 19:00 being the hour with the highest sales at $2,412,939. This suggests that Dibs could strategically place advertisements during this time to maximise the likelihood of customer purchasing decisions.

g. Which products are most often sold together?

The most common combinations were 'iPhone and Lightning Charging Cable' with 1014 times sold together.

h. Overall which product sold the most and why do you think it has sold the most?

- In terms of **quantity**, AAA Batteries being the most sold item.
- In terms of **total sales**, the Macbook Pro Laptop being the most sold item.

Regarding why certain products sold the most, either by quantity or total sales, it can generally be attributed to the nuance of the products. For the items like AAA batteries, which are not high the individual values, are frequently purchased due to their broad uses and affordability, leading to high volume sales. On the other hand, premium items like the Macbook Pro Laptop, despite their higher price point, generate significant total sales due to their high individual value and strong brand recognition.

i. What is the least sold product in the best year of sales?

- The 'AAA Batteries' was the least sold product with regards to **the total sales**.
- The LG Dryer was the least sold product with regards to **the quantity.**

3.2 Visualisation

a. Monthly sales trend vs monthly average sales

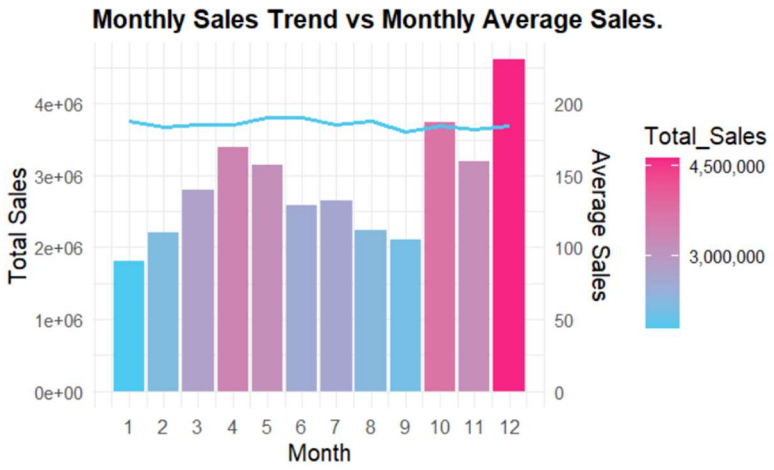

Figure 3.1 Monthly sales trend vs monthly average sales

Figure 3.1 showed a significant increase in sales during the holiday season (April, October and December), indicating more sales were made compared to other months. However, each individual sale is not very large (180 to 190), showing a sign that customers bought a high volume of low-priced items or a small amount each time. After these months, there has been a noticeable dip in sales (January and February), possibly due to consumers cutting back on spending after the holidays.

b. Sales by state

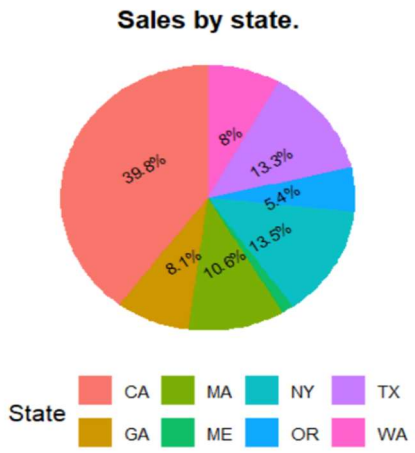

Figure 3.2 Sales by state

The figure 3.2 shows the distribution of sales across different states. California (CA) leads in total sales, followed by New York (NY) and Texas (TX) while Oregon (OR) and Maine (ME) hold the minor market. This could be due to a difference in customer base in these states as different in disposable income, or demand for Dibs' products.

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

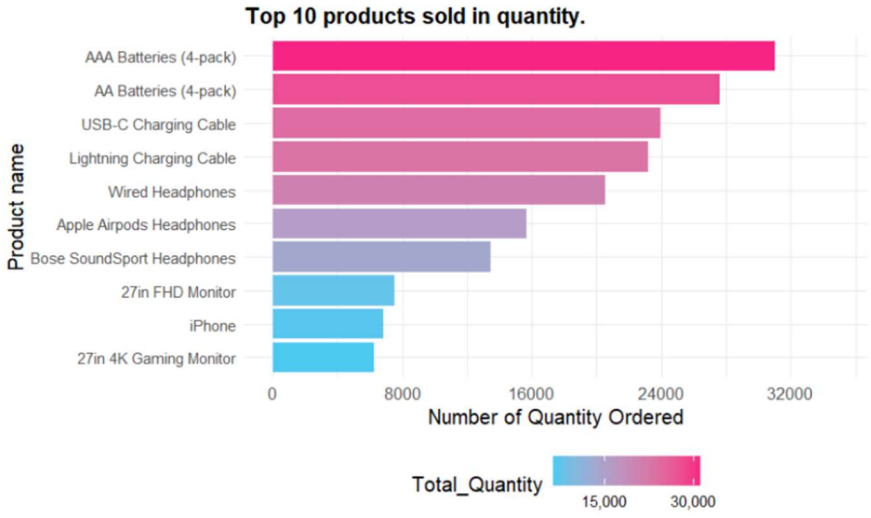

Figure 3.3 Top 10 products sold in quantity

Figure 3.3 shows that the batteries A-type designated represented the most sold products, followed by the USB-C/Lightning Charging Cable and Wired Headphones, suggesting a strong customer demand for electronic products, particularly mobile phone accessories and audio devices.

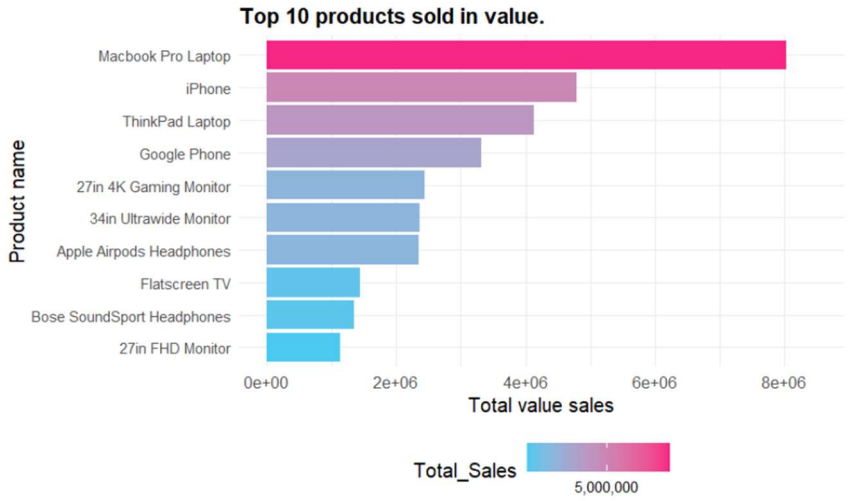

Figure 3.4 Top 10 products sold in value

For the value of sales for products like phones or premium monitors, their values in sales are in the top 3 while their quantity is not *(figure 3.4)*. Therefore, customers are willing to pay for high-quality products or relevant products combined with the main one to enhance their experience.

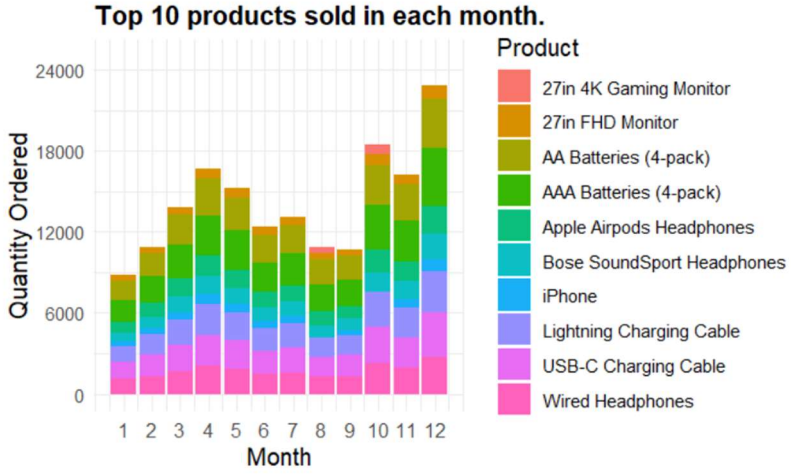

Figure 3.5 Top 10 products sold in each month

	<chr>	<int>
1	27in 4K Gaming Monitor	1
2	27in 4K Gaming Monitor	2
3	27in 4K Gaming Monitor	3
4	27in 4K Gaming Monitor	4
5	27in 4K Gaming Monitor	5
6	27in 4K Gaming Monitor	6
7	27in 4K Gaming Monitor	7
8	27in 4K Gaming Monitor	9
9	27in 4K Gaming Monitor	11
10	27in 4K Gaming Monitor	12
11	iPhone	8
12	iPhone	10

Figure 3.6 Appearance of "iPhone" and "27 in 4K Gaming Monitor"

When we deep dive into the top product sales *(figure 3.5*), "iPhone" and "27 in 4K Gaming Monitor" do not appear in some months *(figure 3.6),* indicating that for the specific period, customers are paying attention to the special products and tend to wait until the new model is released, which meets their interest more, or buying at a discounted price.

d. Monthly order trend vs monthly average order

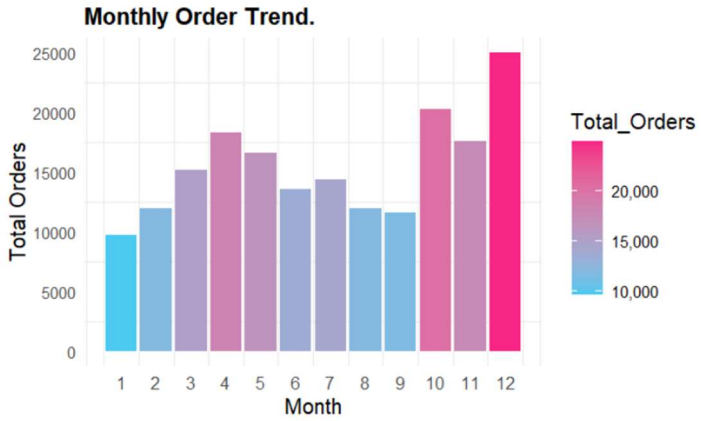

Figure 3.7 Monthly Order Trend.

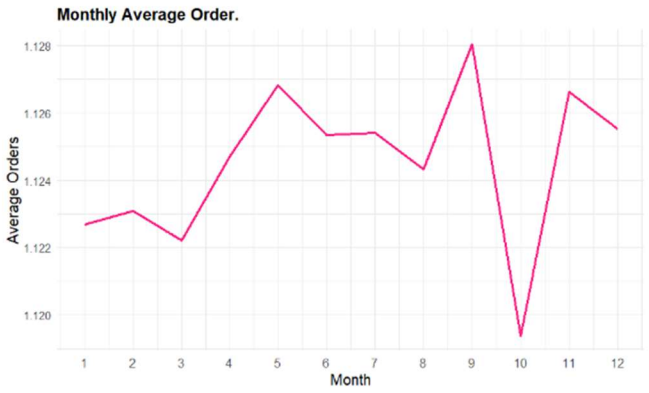

Figure 3.8: Monthly Average Order.

There was a stable increase starting from January and reaching its peak in April as more people made more purchases to prepare for spring/summer activities. After that, the number of orders decreased, explained by people cutting back on spending and preparing for Christmas's occasion or end-year sale as Black Friday/Cyber Monday.

For the average order, it appeared to have a decline in March and September and increased dramatically in April, November, and December. Hence, during special occasions, consumers willing to buy in greater quantities or larger numbers of consumers make purchasing decisions.

e. Daily order trend vs daily average

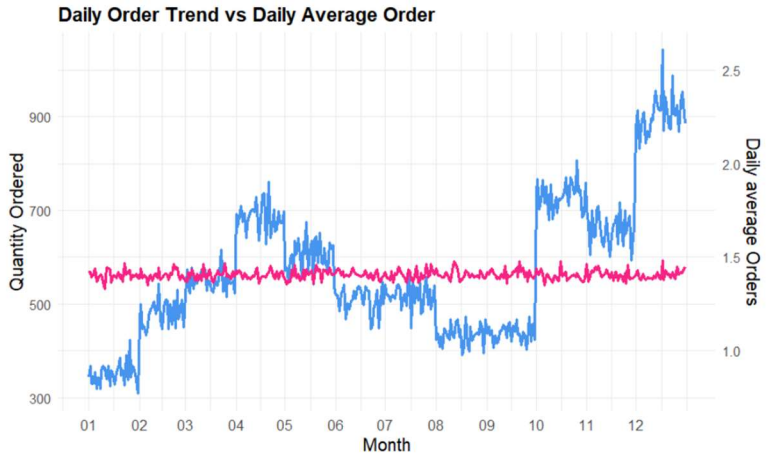

Figure 3.9 Daily Order Trend vs Daily Average Order

There was a dramatic increase in the total number of daily orders (blue line) from the start of October onwards. Moreover, the number of orders seems to be relatively consistent across different

days of the month, with the peaks around the 1st, 15th, and last days of the month *(figure 3.10)* while slight dips in orders around the 5th and 20th, explained by consumers tend to make purchases as their pay cycles renewed.

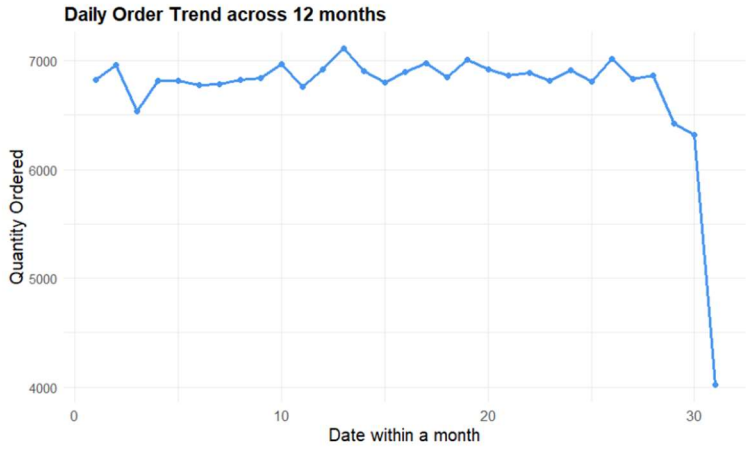

Figure 3.10 Daily Order Trend across 12 months

f. Hourly order trend vs hourly average order

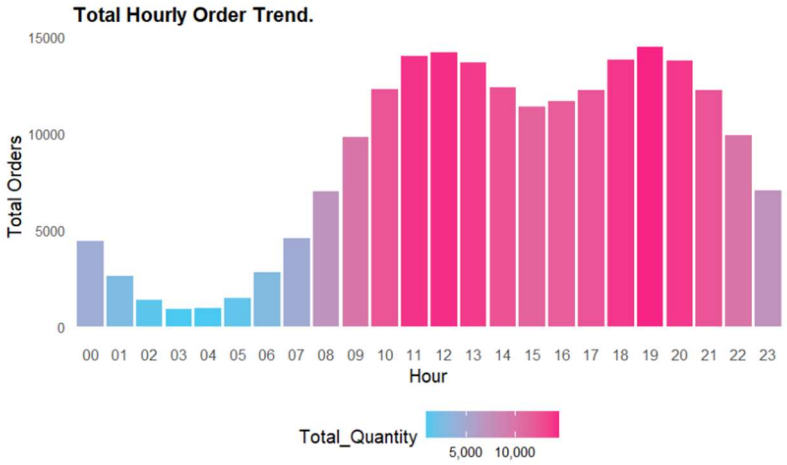

Figure 3.11 Total hourly order trend

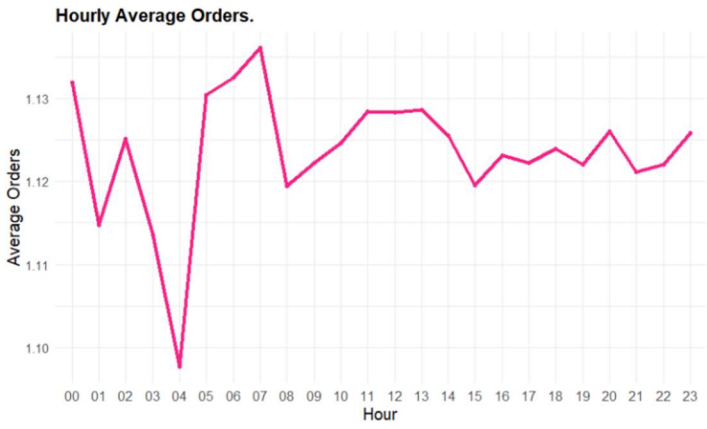

Figure 3.12 Hourly Average Orders

The number of orders was relatively low in the early morning hours (from 0 to 6) as most people are likely asleep during this time. There was a significant increase in orders starting at 7a.m, peaking at 2 points at around 12p.m and 7p.m, illustrating that people make purchases during the day, especially before lunch/finishing work. However, when looking at the average orders, the period from 5a.m to 7a.m showed the highest value, indicating that fewer consumers are making purchases during this period but buying more products at this time.

4. Predictive analysis

a. Data manipulation

The "sales" dimension was created from the product of "price_each" and "quantity_ordered". The data frame is then grouped by distinct year, month, and date for the calculation of "total_sales". Then randomised the rows so that all the values are included in the training and testing data set for model prediction, also set seed for reproductivity. "Sales_evaluation" column is created based on the daily sales range with labels poor, average and good assigned to values under $60,000; $60,000-$100,000; and above $100,000 respectively.

b. Model selection

Regarding the non-parametric model, the K-nearest neighbour is applied to predict sales range. The other 2 chosen parametric models are Linear Regression and Regression Tree to predict the exact total sales.

c. Data split for training and testings

The data frame is splitted into training and testing dataset with the ratio 60/40. The "sales_evaluation" column is used for the diagnosis of KNN models.

d. Model testing and recommendation

- **Non-parametric model**

	actuals	predicted
actuals	1.0000000	0.7348868
predicted	0.7348868	1.0000000

Figure 4.1 Correlation between the prediction and actual

	predictions		
testOutcomes	average	good	poor
average	71	9	5
good	6	46	0
poor	2	0	18

Figure 4.2 Most frequency of discrepancy

The table of differences between actual and predicted sales shows that the K-nearest neighbour algorithm performs well, with a correlation of 0.73 (Figure 4.1). The main discrepancies are between average and good performance, with 9 over-predictions and 6 under-predictions (Figure 4.2). Other differences are minimal. From that result, despite not predicting the correct sales, the model can be used as a reference to predict the sales range and help organisations to adjust the staff needed for specific days and improve efficiency.

- **Parametric models (Linear Regression and Regression Tree)**

For Linear Regression and Regression tree models, rooted mean square errors (RMSE) are calculated from the test data set to infer the average errors. The higher the RMSE, the less fit the model.

- RMSE for Linear Regression Model: 8647.08922845259
- RMSE for Regression Tree Model: 9845.81557440559

According to the RMSE of the 2 models, the number indicates a high variation for both, which can be understood as a high dispersion of errors. However, judging the number alone, the linear regression model is a more optimal choice.

e. Model visualisation and clarification

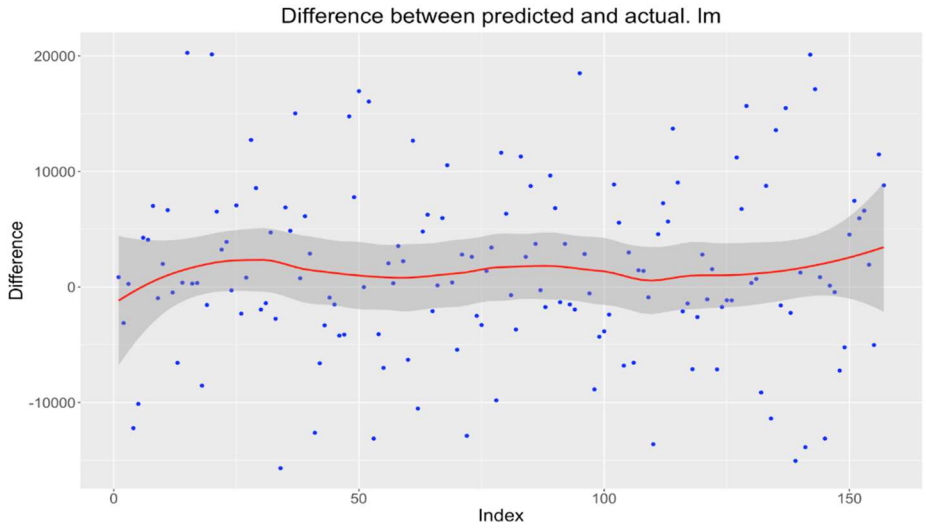

Figure 4.3 Difference between predicted and actual

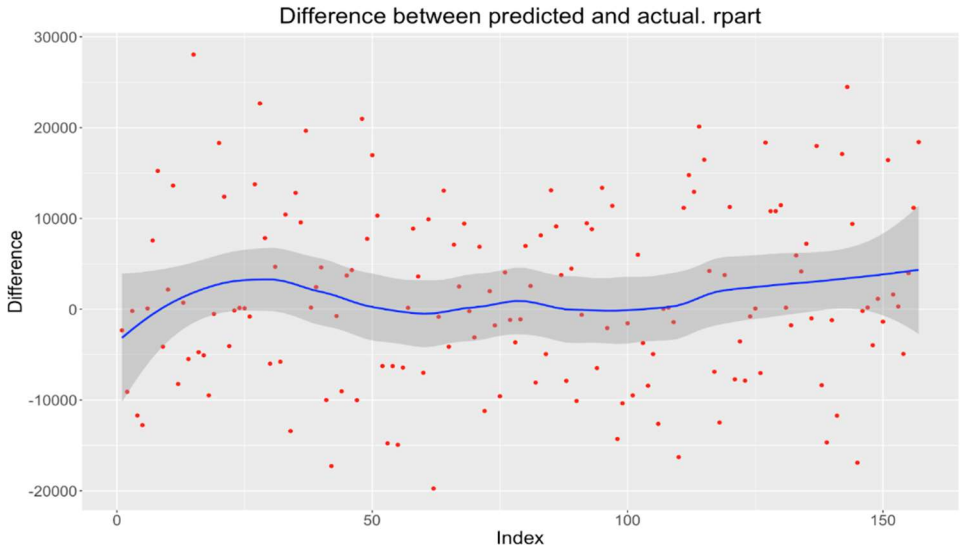

Figure 4.4 Difference between predicted and actual

Figure 4.5 Regression model

Both regression models perform well in predicting price based on quantity and order date. The regression tree predicts better for the trend of the data with the deviation around 20,000 dollars *(figure 4.4)*.

However, it can be seen that the linear model's difference between predicted and actual data has the tendency to gather more around 0 deviation *(figure 4.3)* while the regression tree is more scattered. This is because the actual data clearly showed linearity *(figure 4.5)*.

f. Conclusion

Hence, It can be concluded that the linear model is better to predict the majority of the daily sales.

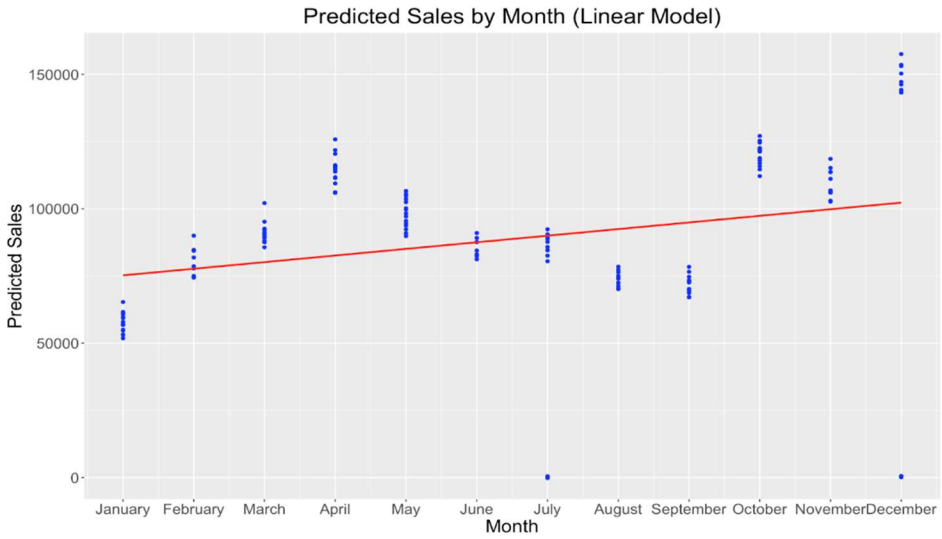

Figure 4.6 Predicted Sales by Month

Figure 4 visualises the prediction of selected model (Linear Model) overtime, which will be useful for seasonality prediction for future years. However, due to the huge lack of 2020 and 2021 data, the model can not be sure to predict the increasing trend well.

5. Future scenarios & Recommendations

Staffing

Dibs can increase staff during peak hours to ensure quick service and a good shopping experience. This can help increase sales and customer loyalty by keeping the supply chain smoothly.

Optimise the advertising times & selling strategies

Regarding the optimal time slot in hours or days, Dibs should concentrate their efforts during these periods by initialising marketing campaigns to capture the attention of potential consumers.

Additionally, during off-peak hours or days, Dibs can implement special promotions and introduce flexible payment options such as Buy Now Pay Later to incentivize customers to make purchasing decisions even when they are waiting for their next paycheck.

Geographical Focus

For Dibs, the company can tailor their selling strategies for each state/city, as different states/cities will have different disposable income and demand. For high performing ones, Dibs can develop expansion and marketing strategies to gather consumers attention while localised marketing campaigns can be tailored to suit underperforming cities to enhance the company's presence and attract new customers.

Product Bundling

By understanding the frequent combination of electronic products like phones with accessories (headphones/charging cables), Dibs can develop bundle offerings with promotions or special deals, ultimately increasing the average quantity sold and enhance the consumers' experience at a discounted price, thereby increasing sales.

Inventory Strategy Adjustments

Dibs can increase the inventory levels of high-value products like the Apple products, monitors and explore the introduction of higher-margin items in this segment while ensuring high-demand products are well-stocked before peak hours. This can prevent lost sales due to out-of-stock items while maintaining the customer's experience when making purchases.

Utilise the Seasonal Tendency

Given the highest total sales in December, attributed by the impact of holiday seasons, Dibs can utilise the holiday-targeted campaigns along with limited seasonal products to boost sales further.

Additionally, the linear model *(figure 4.4)* can also be used to check the seasonality effects, in which the sales tend to peak in April, October and December while they are notably lower in January and September. Despite its difficulty to project future trends due to insufficient data, it is anticipated that equivalent patterns will recur in the following years.

6. Conclusion

In conclusion, our comprehensive analysis of Dibs' sales data has provided valuable insights into consumer behaviour and sales trends. Through tasks including Cleaning, Descriptive Analytics, Visualisation, and Predictive Modelling, we have identified key areas where Dibs can optimise their marketing strategies and operations to enhance customer engagement and sales performance. Our predictive models suggest effective strategies for targeting key sales periods and customer segments. We recommend continued monitoring of sales trends and regular updates to the predictive models to ensure they remain relevant as market conditions change. Ultimately, by implementing our recommendations, Dibs is well-positioned to increase sales and strengthen customer loyalty, thereby achieving sustained business growth.