Exploratory Data Analysis of Online Retail business using Python

1.Introduction:

In data analytics, raw datasets provide multiple opportunities to explore further and provide valuable insights to help businesses grow and improve efficiency. Using the online retail dataset from Kaggle, we perform further analysis using Python programming to identify important trends and patterns. This report builds on our previous work to identify additional insights that can aid businesses' overall growth. This report investigates sales performance day of the week and hour of the day to find peak periods and provide recommendations. We also conduct product analysis, which reveals the most returned or cancelled items, which will allow better inventory management. Additionally, we segment our customers into loyalty tiers so that they can receive targeted marketing messages and customized discount offers. Finally, sales forecasting is carried out in order to forecast total sales for the coming year. All in all, this report gives actionable strategies for the overall growth of the business by insights from data analytics.

2.Exploratory Data Analysis:

2.1 Sales by hour of the day

To identify the peak hour of sales, we analyse the dataset for the United Kingdom, where our sales are high, to identify when customers are online for further engagement.

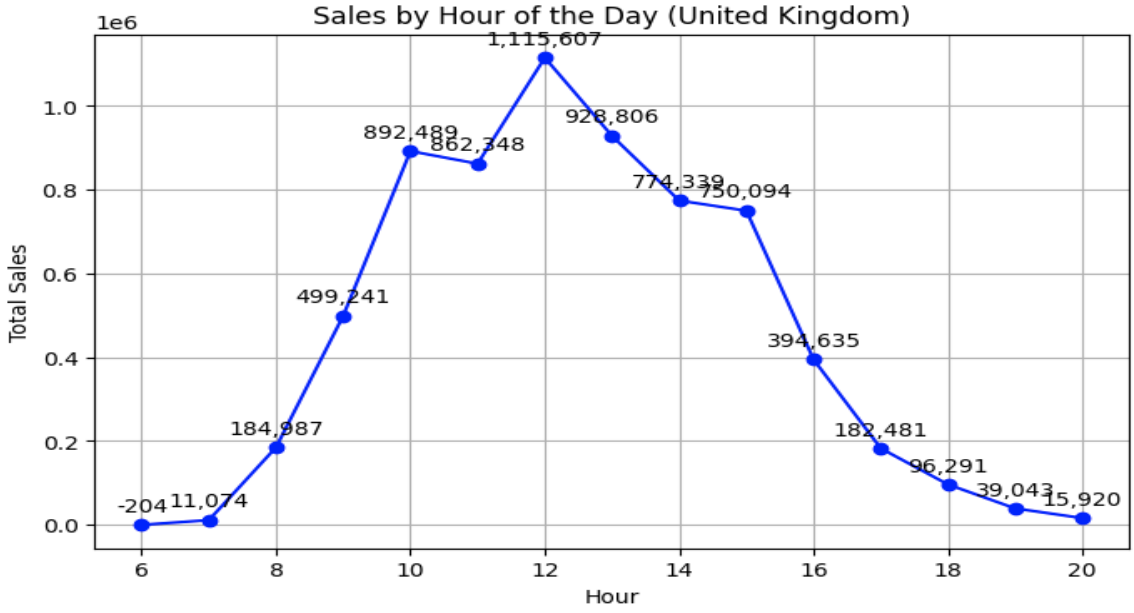

Figure 1: Line graph of sales by hour of the day in United Kingdom

As shown in Figure 1, from 7 AM to 12 PM, customers purchasing products online steadily increase, and sales in the United Kingdom steadily decrease from noon to 8PM.

Using this analysis, we can further improve our sales by engaging with customers online to suggest products and cross-sell products during 7AM to 12 PM. We can also assist customers during this period in finding the products that they are looking for to improve customer satisfaction.

2.2 Sales by the day of the week

We identify the total sales by the day of the week by analysing when and how much the customer purchased the respective product.

```
Monday       1271078.601
Tuesday      1562715.681
Wednesday    1526440.000
Thursday     1902316.050
Friday       1238556.741
Saturday             NaN
Sunday        777412.351
Name: TotalSales, dtype: float64
```

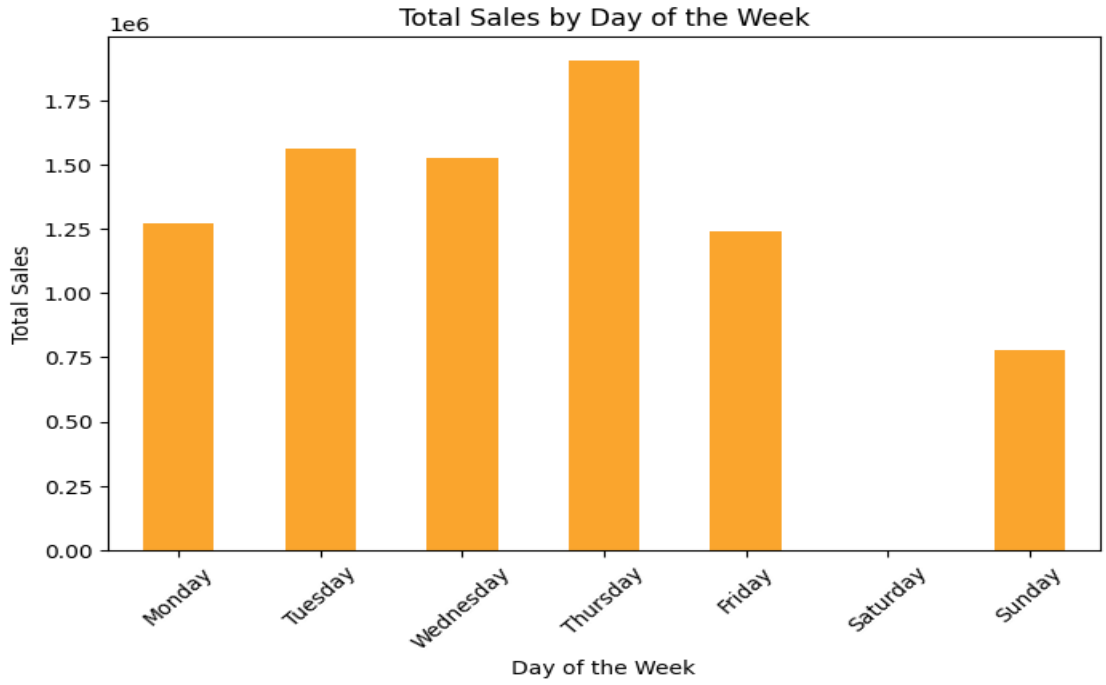

Figure 2: Bar graph of sales by the day of the week

We have the highest total sales of 1,902,316 on Thursday and the lowest on Sunday of 777,412. We see similar total sales throughout Monday, Tuesday, and Wednesday.

We can create discount plans and offers on Friday and Sunday to improve the total sales. Additionally, we can create promotional schemes and host competitions on weekends, which

could attract more customers to visit the website. This increased traffic may encourage purchases. Additionally, we can introduce new products or product bundles on Saturday to create exclusivity and demand, which will improve total sales on the weekends.

2.3 Sales forecast

To find the next year performance of the business we run a machine learning model 'Prophet" on total sales by date. This analysis shows the sales forecast for next year based on the previous sales data.

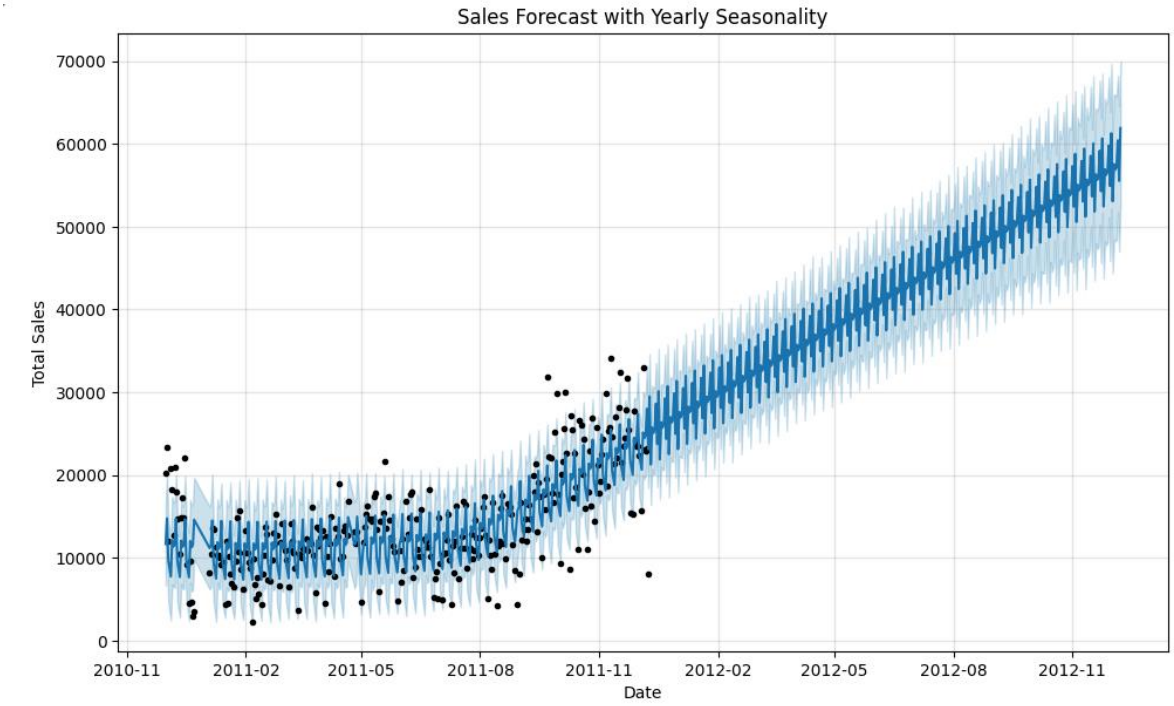

Figure 5: Sales Forecast for the next year

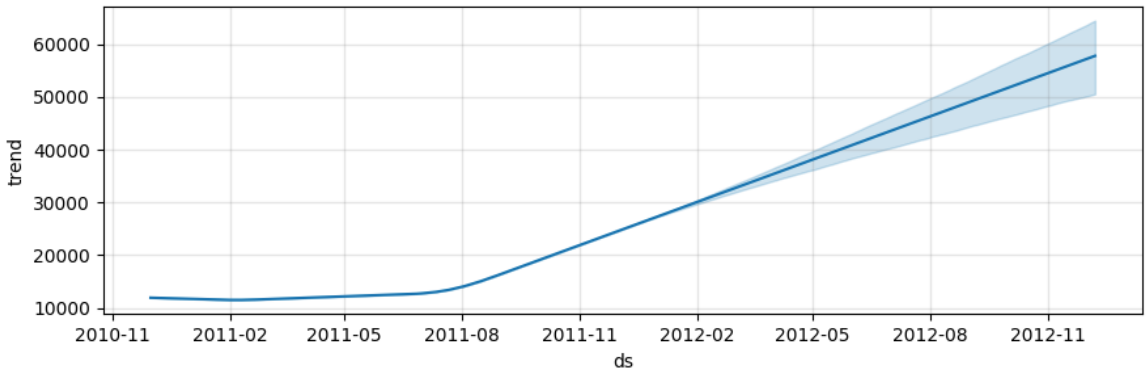

Figure 6: Trendline for total sales

As shown in Figures 5 and 6, it indicates steady growth in total sales for the following year. Total sales would reach around 70000 in December 2012. The black dot indicates the current year's sales, and the blue line shows the maximum and minimum sales during the period.

The growth in the total sales forecast only relied on the previous year's sales data; it does not consider any other variables. Other variables, such as inflation and changes in demand, can impact the model's output, so we recommended that businesses do frequent data updates and add variables that impact the model's forecast to identify the trend and make decisions accordingly precisely.

2.4 Top Cancelled or returned products

We Identify products that are cancelled or returned, since they are important to understand the demand and manage inventory.

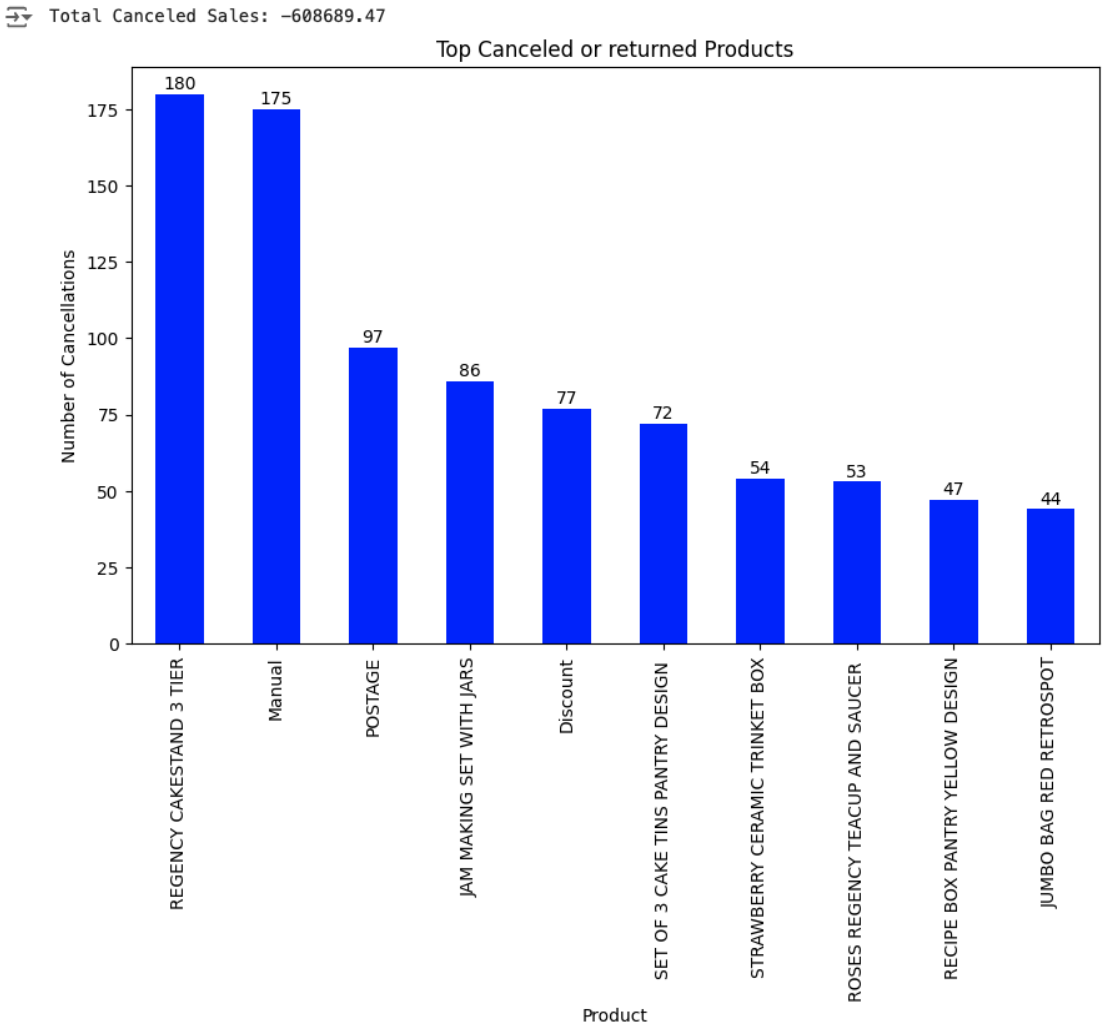

Figure 3: Bar graph of top cancelled or returned products

As shown in the figure above, “Regency cake stand 3 tier” is the most returned item, with 180 returns, followed by “Manual” with 175 returns. The business has lost 608,689.47 in sales from returns.

We can ask for feedback from customers to identify the reason for their return. We can then decide to fix the issues faced by the customers for the respective product. If we identify returns without satisfactory reasons, then we amend our return policy to curb returns, which can reduce the total loss from returns.

2.5 Customer loyalty analysis

We analyse the dataset to identify loyal customers based on their purchase frequency, which can be identified based on the number of times purchased.

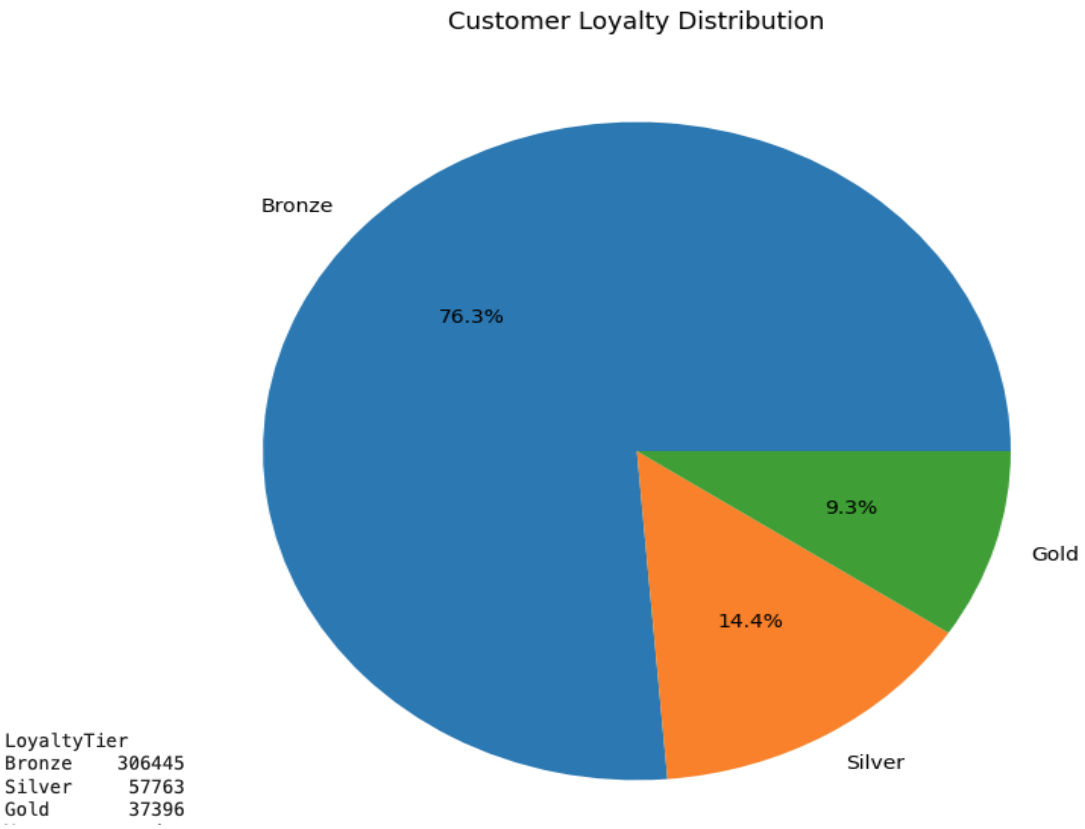

Figure 4: Pie chart of customer loyalty distribution

We can see that 9.3% of the total customers, 37,396, are loyal and purchase more than 50 times. It is followed by 14.4% of customers, 57,763, who purchased more than 20 times. The rest,

76.3%, purchase less than 20 times. The tier segmentation is based on purchase frequency, which should be altered accordingly.

Exclusive rewards like early bird passes or exclusive discounts can be offered by the organization to keep gold-tier customers. It may also provide discounts on bundles to silver-tiers and point out the benefits of gold-tier membership so that they are persuaded to make purchases quite frequently. Strategies such as gamification and loyalty points will increase their purchase frequency.

3. Conclusion

In conclusion, we have gathered further insights on the same dataset which can improve business decisions. We found that 8 AM to 12 PM is the busiest time for purchases, which presents a chance to interact with customers and increase sales. Thursday is the most profitable day, with maximum sales of 1,902,316, while Sunday is the least profitable. 9.3% of the customer base are loyal and repeat customers engaging with them can improve the total sales and growth. The sales forecast indicates an upward trend, projecting the total sales to reach approximately 70,000 in the upcoming year. Along with these insights, implementing the recommended strategies can improve inventory management, engagement and retention, ultimately contributing to the sustained growth of the business.

References

1. Panda-monium (2024). *Online Retail Dataset*. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/divanshu22/online-retail-dataset/data.

2. Žunić, E., Korjenić, K., Hodžić, K. and Đonko, D. (2020). Application of Facebook's Prophet Algorithm for Successful Sales Forecasting Based on Real-world Data. *International Journal of Computer Science and Information Technology*, 12(2), pp.23–36. doi:https://doi.org/10.5121/ijcsit.2020.12203

3. Neale, H. (2020). *Key Analytics to Boost Retail Marketing Strategies*. [online] Alytix Marketing. Available at: https://www.alytixmarketing.com/post/analytics-optimise-retail-marketing-strategy.

4. Stone, M., Bearman, D., Butscher, S.A., Gilbert, D., Crick, P. and Moffett, T. (2003). The effect of retail customer loyalty schemes — Detailed measurement or transforming marketing? *Journal of Targeting, Measurement and Analysis for Marketing*, 12(3), pp.305–318. doi:https://doi.org/10.1057/palgrave.jt.5740117

5. Liu, L. (2024). *Python: Effective Techniques for Managing Dates in DataFrame*. [online] Hackernoon.com. Available at: https://hackernoon.com/python-effective-techniques-for-managing-dates-in-dataframe [Accessed 4 Dec. 2024].

Appendix:

Programming code file:

https://colab.research.google.com/drive/1OSrhoCRRJTXXh5cd8JQg45j_0hEbKEAT?usp=sharing