Breakout activity Machine learning

Exploratory Data Analysis (EDA)

1. What is the total revenue generated by each salesperson (Kelci Walkden, Brien Boise, and Others)? How do their performances compare?
2. Which country (Australia, India, or Others) contributes the most to the total sales revenue? What percentage of the total revenue does it account for?
3. How many unique products are sold in the dataset? Which product (e.g., 50% Dark Bites, Eclairs, or Others) has the highest number of transactions?
4. What is the trend of sales over time based on the Date column? Identify the month or period with the highest transaction count.
5. What is the distribution of the **Amount** column? Are there any outliers in the revenue generated per transaction (e.g., values above 638.20 or below 1.00)?
6. How does the number of **Boxes Shipped** correlate with the **Amount**? Is there a strong relationship between these two variables?
7. Which salesperson shipped the most boxes of chocolate? Does this align with their total revenue contribution?

Data Cleaning

8. Are there any missing values in the dataset (e.g., in Sales Person, Country, Product, Date, Amount, or Boxes Shipped)? How would you handle them?
9. The dataset shows date ranges (e.g., 07/14/2022 - 08/07/2022). Should these be split into individual dates or kept as ranges? Justify your choice and suggest a cleaning approach.
10. The **Amount** column has a range from 1.00 to 709.00. Are there any negative or unrealistic values that need to be removed? How would you identify and clean them?
11. The **Country** and **Product** columns have "Other" categories (64% and 89%, respectively). Should these be treated as a single category, or should you attempt to break them down further? Why?
12. Are there duplicate transactions in the dataset (e.g., same Sales Person, Date, Product, Amount, and Boxes Shipped)? How would you detect and resolve them?

Business Expectations

13. Based on the dataset, which salesperson should the chocolate company recognize as the top performer? Justify your answer using revenue and boxes shipped.
14. Which country should the company focus on for expanding its chocolate sales? Support your recommendation with data.
15. What insights can you provide about the popularity of chocolate products (e.g., 50% Dark Bites vs. Eclairs)? How can the company use this to adjust its production or marketing strategy?
16. Are there seasonal patterns in chocolate sales (e.g., higher sales in certain months)? How can the business use this information for inventory planning?

17. If the company wants to increase revenue by 20%, how many additional boxes of chocolate would need to be shipped, assuming the current average revenue per box remains constant?

Machine Learning Models (KNN, Decision Tree Classifier, Logistic Regression, Random Forest)

18. Suppose you want to predict whether a transaction will generate "High" revenue (e.g., Amount > 355.00) or "Low" revenue (e.g., Amount ≤ 355.00). How would you preprocess the dataset (e.g., encoding categorical variables like Sales Person, Country, Product) for use in KNN, Decision Tree, Logistic Regression, and Random Forest models?
19. Train a KNN classifier to predict whether a transaction is "High" or "Low" revenue based on features like Boxes Shipped, Country, and Product. What value of K would you choose, and why?
20. Build a Decision Tree Classifier to predict the same "High" vs. "Low" revenue outcome. Which feature (e.g., Boxes Shipped, Country, Sales Person) is the most important in splitting the data? How do you interpret this?
21. Use Logistic Regression to predict the probability of a transaction being "High" revenue. How do the coefficients of the model help you understand the impact of Boxes Shipped on revenue?
22. Implement a Random Forest Classifier to predict "High" vs. "Low" revenue. How does the ensemble approach improve performance compared to a single Decision Tree? What is the importance of each feature (e.g., Boxes Shipped, Country) in the Random Forest model?
23. Compare the performance of KNN, Decision Tree, Logistic Regression, and Random Forest models using accuracy, precision, and recall. Which model performs best for this classification task, and why?
24. Tune the Random Forest model by adjusting hyperparameters (e.g., number of trees, maximum depth). How do these changes affect the model's performance?
25. Can you use the dataset to cluster sales transactions into groups (e.g., using K-Means) instead of classification? What features would you use, and how might the clusters help the business?

Notes for Students

- **EDA**: Use visualizations (e.g., bar charts for sales by country, line plots for sales over time, histograms for Amount) to explore the data.
- **Data Cleaning**: Handle missing data, outliers, and categorical variables appropriately before modeling.
- **Business Insights**: Tie your findings back to actionable recommendations for the chocolate company.

- **Machine Learning**: Split the data into training and testing sets, evaluate models, and justify your choices. For Random Forest, explore how the ensemble nature reduces overfitting compared to a single Decision Tree.