Data Analysis Report

Problem Statement

The cafe chain seeks to delve into sales data to uncover trends, improve customer experience, and optimize product offerings. The analysis aims to guide strategic decisions that drive growth and customer satisfaction**.**

Datasets Used

1) **Cafe Sales Data:** The dataset includes records of individual sales transactions, capturing various details about each sale, including customer information, product details, transaction specifics, and external factors that may impact sales. It includes following columns:
   - Transaction ID: A unique identifier for each sales transaction.
   - Date: The date when the transaction occurred.
   - Customer ID: A unique identifier for each customer.
   - Item Category: The category of the item sold (e.g., Coffee, Sandwiches).
   - Product ID: A unique identifier for each product.
   - Product Description: A description of the product sold.
   - Quantity Sold: The number of units of the product sold.
   - Sale Amount: The total amount of money received from the sale.
   - Payment Type: The method of payment used for the transaction (e.g., Credit Card, Cash).
   - Discount Applied: Indicates whether a discount was applied to the transaction (e.g., Yes/No).
   - Discount Amount: The amount of discount applied to the sale.
   - Employee ID: A unique identifier for the employee who processed the transaction.
   - Location: The location or branch where the transaction took place.
   - Temperature (°F): The temperature in Fahrenheit at the time of the transaction.
   - Branch: The specific branch of the cafe where the transaction occurred.

2) **Events data:** This xlsx file contains the event names and the respective dates each event took place on. Columns in this dataset are:
    - Date: Date on which an event took place
    - Event Description: Name of the event

Solution Design Structure

Tools used:

- Jupyter Notebook
- Tableau Desktop

a) Data Analysis and Model Building

1) Assessing Data Sets and Identifying Anomalies

- Employed basic pandas functions such as describe() and info() to summarize and understand the data structure.
- Detected missing values, outliers, and inconsistencies.
- Utilized techniques such as imputation for missing values, removal or correction of outliers, and data normalization.

2) Correlating External Factors with Sales Performance

- Merged events data with café sales data to analyze sales trends during events.
- Calculated correlation using corr() function and created scatterplot to visualize the relationship between temperature and sales.

3) Analyzing Sales Performance Across Categories

- Created a bar chart to visualize sales across different item categories.
- Developed a clustered column chart to analyze sales trends by month and item category.

4) Forecasting Sales for next quarter

- Built multiple predictive models including Linear Regression and Random Forest Regressor.
- Selected the model with the best accuracy for predicting Q1 2024 sales.
- Build a line chart to show the predicted Sale Amount by both models

5) **Analyzing the Effects of Discounts on Sale Amount Volume**

- Calculated a "Discount Rate" variable using Discount Amount and Sale Amount.
- Build a scatterplot of Discount Rate vs Sale Volume to understand the relationship between Discount and Sales.
- Used a heatmap to confirm the negative correlation between discount rate and sales amount.
- Proposed strategies based on the analysis to optimize discount offerings.

6) **Different Analysis on Dataset:**

- Tableau Dashboard: Conducted further analysis in Tableau, created interactive dashboard, and published it on Tableau Public.
- Applied filters and created interactive KPI's in Tableau Desktop.
- Included the URL to the Tableau dashboard in the report for detailed insights.

7) **Forecasting Sales for the Next 6 Months**

- Autoregressive Model: Utilized AR (Autoregressive) models to forecast sales for the next six months, considering event timings.

b) Strategic Recommendations and Documentation

8) **Formulate Strategic Recommendations**

- Addressed the issue of negative values in Sale Amount and Quantity Sold detected during the descriptive analysis. Due to insufficient information on why these negative values exist, recommended strategies include:

i) Data Verification: Implementing a verification process to ensure data accuracy before analysis.

ii) Data Cleaning Protocols: Establishing protocols to handle and correct negative values in sales data.

- Provided detailed insights for each task in Part I, ensuring actionable recommendations:

9) **Discussing Areas for Further Investigation**

Highlighted the need for improving data quality, especially addressing anomalies like negative values in Sale Amount and Quantity Sold.

Identified potential areas for further investigation, such as:

i) Customer Demographics: Exploring the impact of customer demographics on sales performance.

ii) Seasonal Trends: Conducting a more detailed analysis of seasonal sales trends to refine forecasting models.

10) Ensuring Comprehensive Documentation

- Ensured all data manipulations, model selections, and recommendations are thoroughly documented and justified.
- Maintained clear and concise comments and bullet points within the Jupyter notebook to enhance end-user understanding.
- Included appropriate visualizations in Tableau to support and present findings effectively.

Methodology

Summary Statistics and Data Anomalies: To begin the analysis, summary statistics were generated using the pandas functions describe() and info(). These functions provided summary of the dataframe, revealing anomalies such as negative values in the Sale Amount and Quantity Sold columns. The dataset contained no missing values. Replaced these negative values with zero to maintain data integrity and avoid losing valuable data. This strategy was chosen after engaging with stakeholders to discuss possible reasons for the anomalies.

Explore Relationship with External Factors: To explore the relationship between events and sales, the Events data was merged with the Cafe Sale data using a left join. This allowed for an analysis of the correlation between event days and Sale Amount. For examining the relationship between temperature and sales, the correlation coefficient was calculated using the corr() function, and a scatter plot was created to visualize the relationship between Temperature and Sale Amount.

Sales Performance Across Categories: The sales performance across various categories was analyzed by creating a bar chart to visualize the Sale Amount for each item category. Further insights were obtained by creating a clustered column chart, with the x-axis representing the month, the y-axis representing Sale Amount, and the item category color-coded using the hue parameter.

Sales Forecasting for Q1 2024: Data preprocessing involved creating new columns such as Month, Day of Year, and Is Weekend to ensure the predicted Sale Amount for Q1 2024 followed

existing trends. The GridSearchCV function was used with the Random Forest Regressor to identify the best model. The models were evaluated based on Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) values to determine their accuracy. A line chart was built at the end of this task to visualize the predicted Sale Amount by both the models.

Analysis of Discount Effects on Sales: A new feature named "Discount Rate" was created using the formula: Sum(Discount Amount) / Sum(Sale Amount). A scatter plot of Discount Rate versus Sale Amount was created. A correlation heatmap was generated to confirm the correlation value between Sale Amount and Discount Rate. This analysis helped to identify the optimal discount range to maximize sales.

Forecasting Sales for the Next 6 Months: An Autoregressive (AR) model was implemented to forecast sales for the next six months. This model used historical sales data to predict future sales trends.

Different type of Analysis: Build a dashboard in Tableau Desktop using both the datasets. To apply filters on the dashboard including both the datasets, a relationship was created.

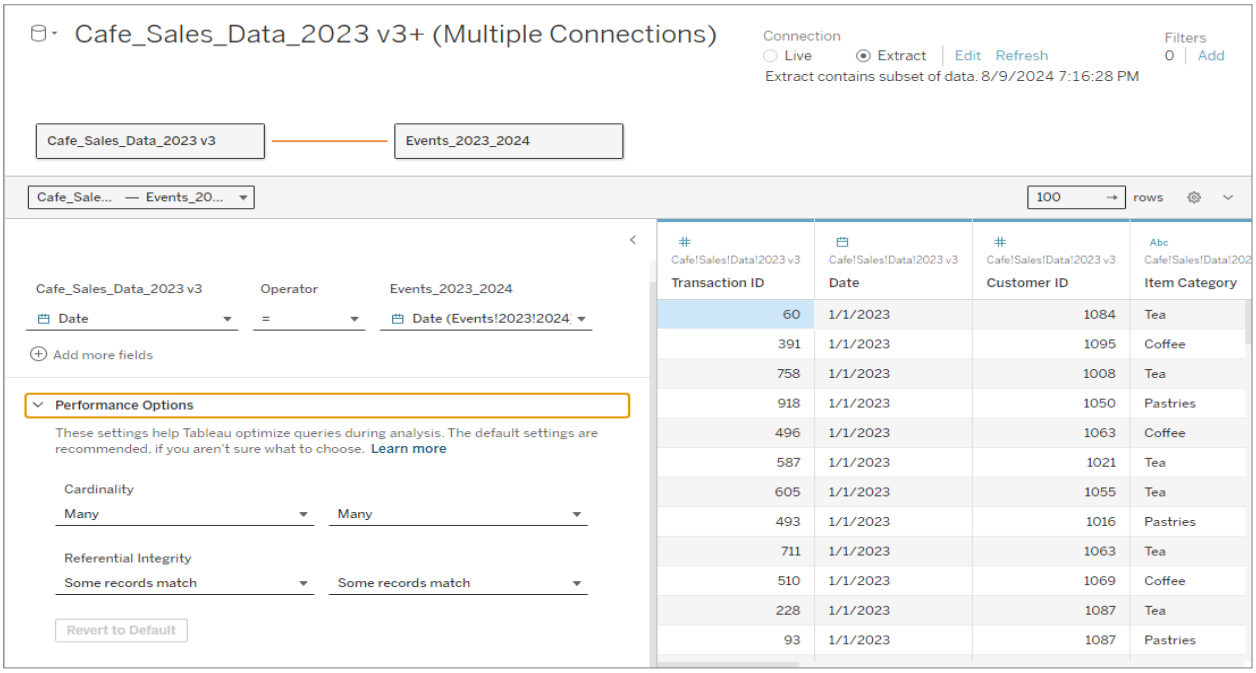

Figure 1: Relationship between Cafe Sale dataset and Events dataset

Calculated fields were used to create KPI cards. Tableau functionalities such as parameters and filters were used in dashboard to ensure different type of analysis and deeper insights.

Results and Discussion

The summary statistics revealed **1,035 rows with negative values for Sale Amount and Quantity Sold**. There were no missing values across both numeric and categorical columns. The **average Sale Amount per transaction was $26.6**, with temperature ranging from **30°F to 100°F**. **Coffee** was identified as the top-selling category, and **Credit Card** was the most frequently used payment method.

Analysis showed **no relationship between event days and Sale Amount**, as the dataset included only four events in 2023. The events number are too low to indicate any potential relationship. More data is required.

```
Date
2023-01-01    4260.54
2023-01-02    3620.97
2023-01-03    4470.87
2023-01-04    4129.23
2023-01-05    4215.27
Name: Sale Amount, dtype: float64
```

	Event Description	Sale Amount
0	Independence Day	3847.92
1	Labor Day	3414.36
2	Long Weekend	4631.49
3	Middle Beast Festival	4060.80

Figure 2: Output shows the average Sale amount value per day vs Events

The correlation coefficient between Temperature and Sale Amount was -0.0196, indicating no significant relationship. The temperature range, with a mean of 65.5°F and a maximum of 100°F, suggested that **temperature had negligible impact on sales**.

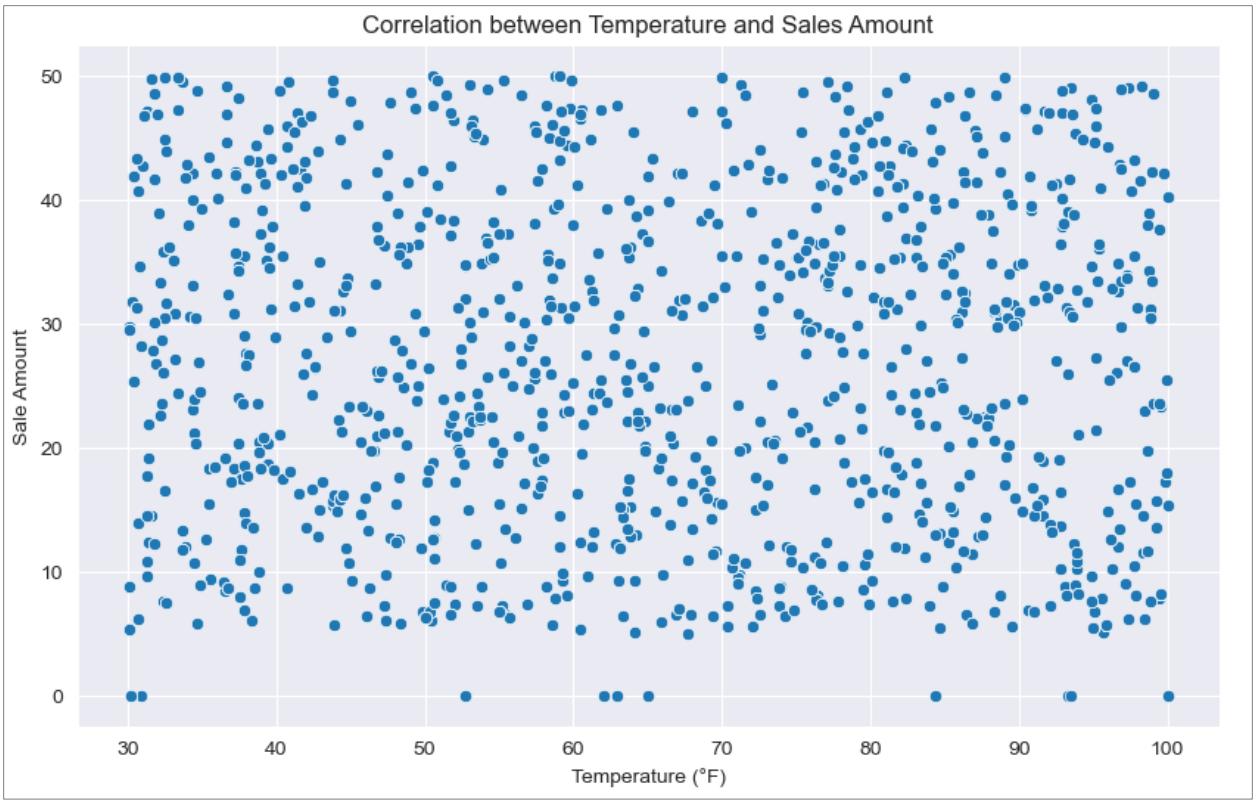

Figure 3: Scatter plot indicates no relationship between Temperature and Sale Amount

The sales performance analysis indicated that coffee generated the highest Sale Amount, followed by sandwiches and pastries. Tea had the lowest sales among the categories analyzed.

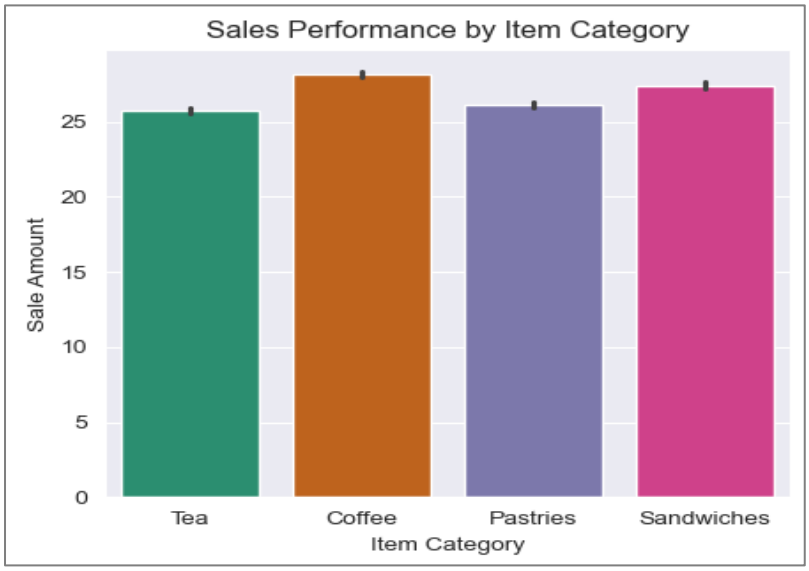

Figure 4: Bar chart indicating Sale Amount per Item Category

This provided valuable insights into product popularity and sales trends. For deeper insights, a clustered column chart was made.

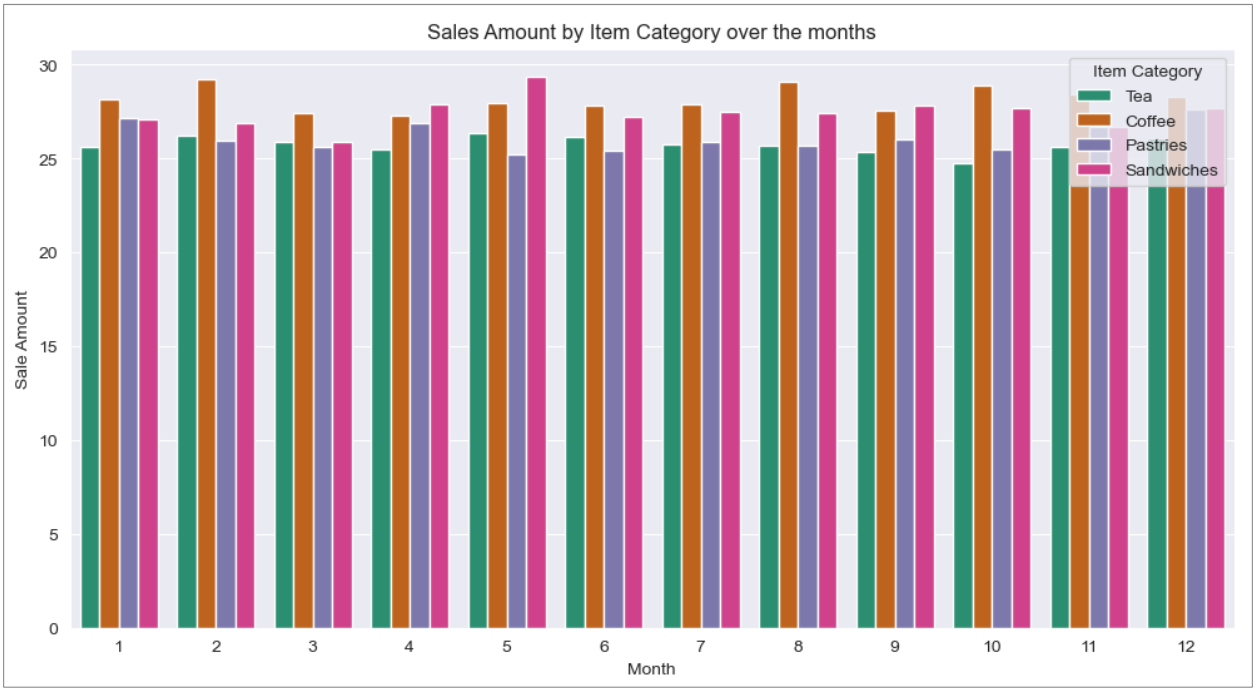

Figure 5: Clustered column chart showing Sale Amount volume overtime by Item Category

In forecasting sales for Q1 2024, Linear Regression and Random Forest Regressor models were implemented. The Linear Regression model had MSE and RMSE values **of 103,336.36 and 321.46**, respectively, while the Random Forest Regressor had MSE and RMSE values of **113,375.12 and 336.71**, respectively. Both models predicted daily Sale Amounts with similar accuracy, but the Random Forest Regressor, enhanced by GridSearchCV, showed greater precision in handling fluctuating sales prices.

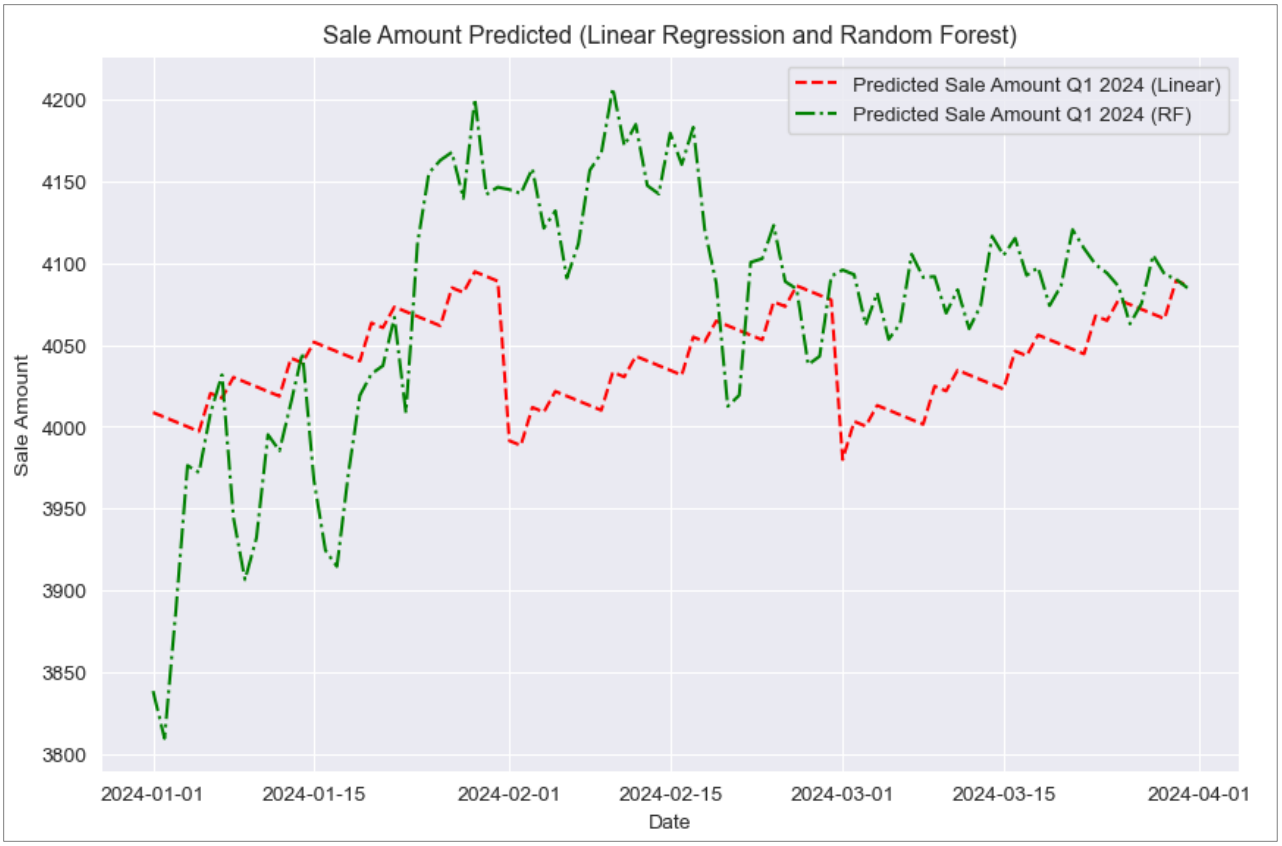

Figure 6: Predicted Sale Amount by Linear regression and Random Forest regressor

Scatterplot was built to show the relationship between Discount rate and Sale Amount. As the discount rate is greater than 20%, the Sale amount is declining sharply. The analysis of discount effects revealed that the optimal discount range was between 10% and 20%. Beyond a 20% discount rate, the Sale Amount tended to decrease.

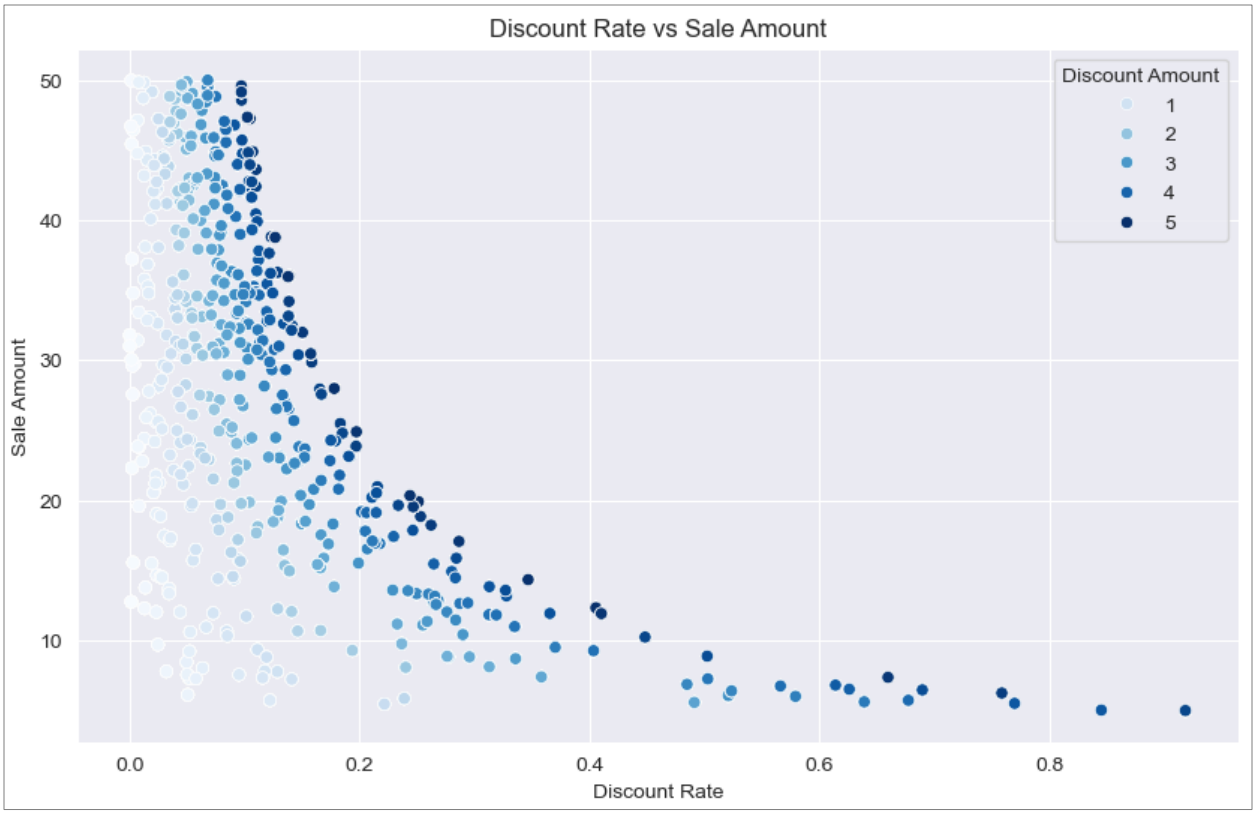

Figure 7: Scatterplot shows negative relationship of Discount Rate vs Sale Amount

The negative correlation between Discount Rate and Sale Amount highlighted the need for careful management of discount strategies to avoid less returns.

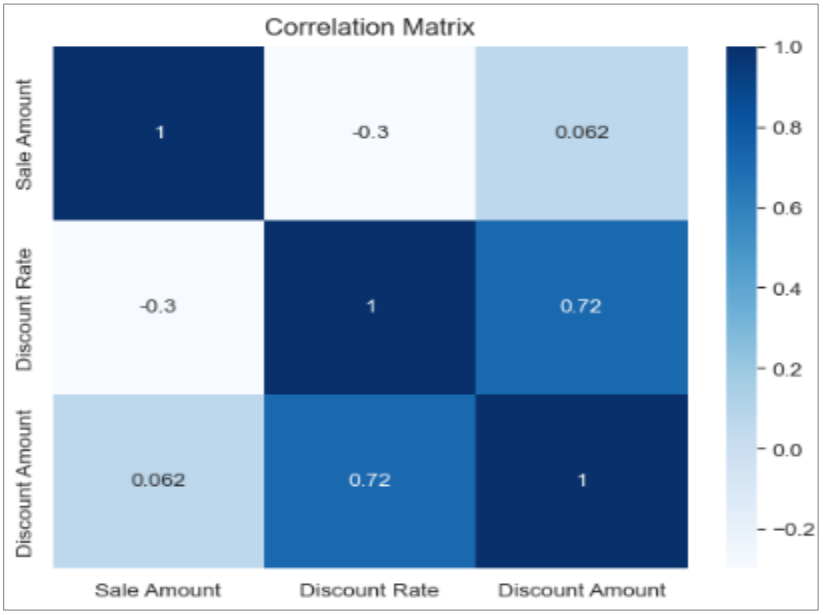

Figure 8: Heatmap visualizing Discount Amount correlation with Sale Amount

The AR model forecast for the next six months suggested that sales would be low for the first month of 2024, with the **highest Sale Amount predicted for the second month**. This forecast provided valuable insights into sales trends, although incorporating seasonal parameters could improve accuracy. Although, the model gave better results, yet the data is not enough to predict seasonal forecast. No seasonal parameters were included in this forecast.

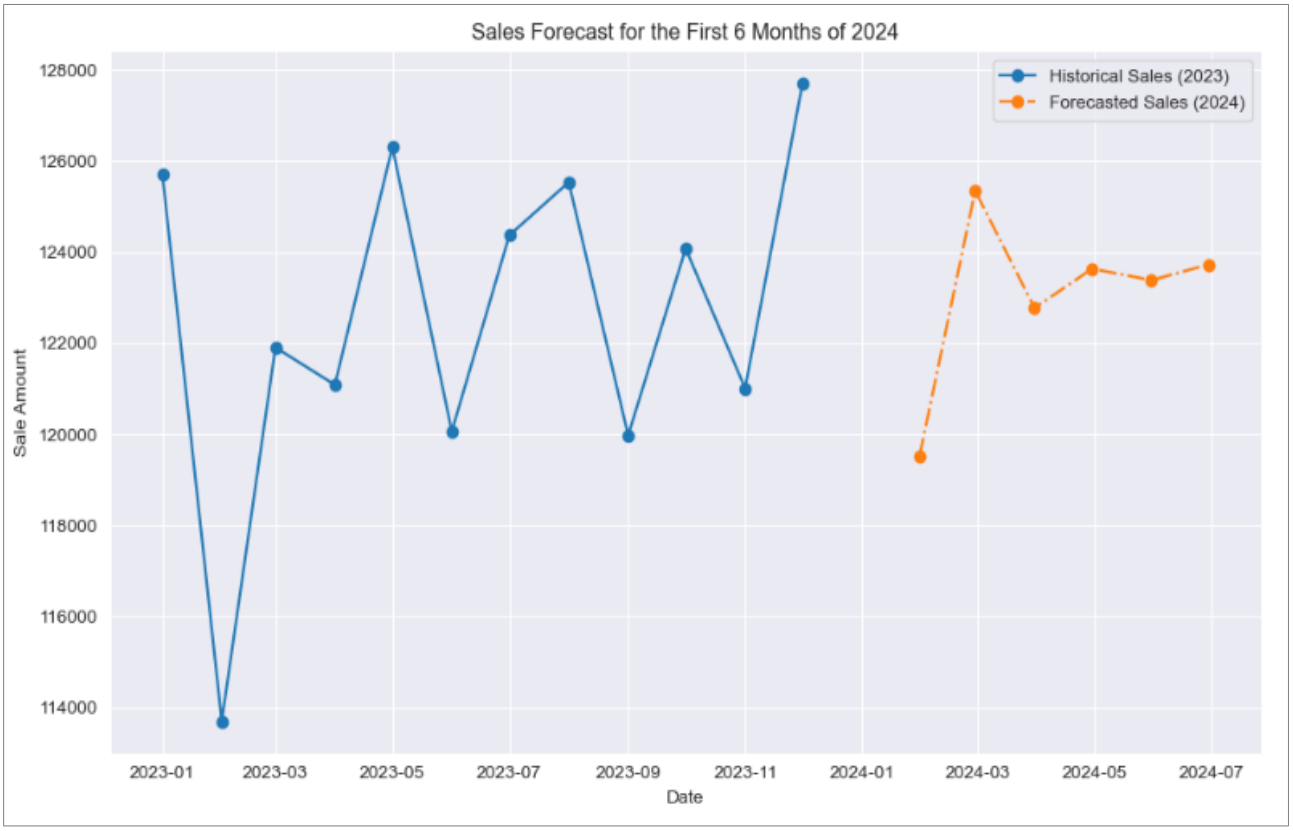

Figure 9: Sale Amount forecast using Auto Regressive model

To explore the data, a dashboard was made in Tableau Desktop. Tableau functionalities such as parameters, filters and calculated fields were used to make the dashboard more interactive. Filters for Employee ID, Item Category, Location, Quarter of Date and Payment Type are applied to whole dashboard to allow the end-user to get detailed insights.

Link to the dashboard:

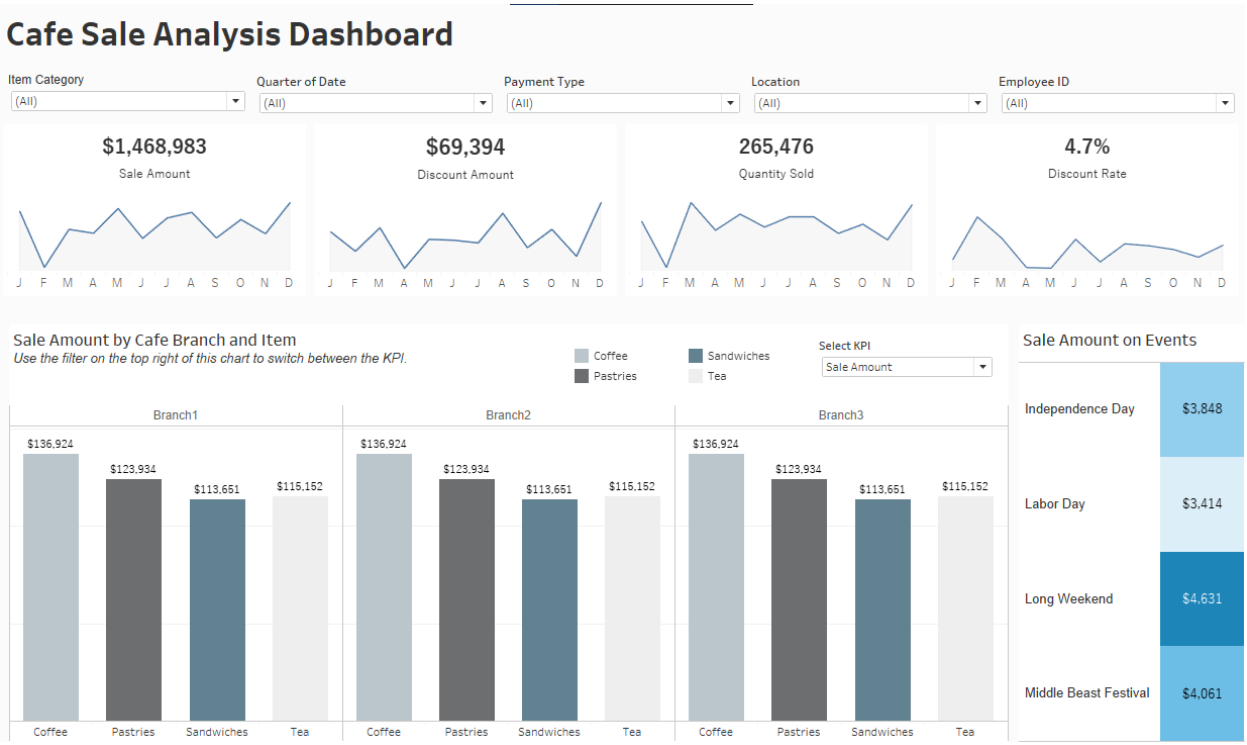

Figure 10: Cafe Sale Analysis Dashboard

Recommendations

- Engage with stakeholders to understand the cause of negative Sale Amount and Quantity Sold values.
- Prioritize and promote coffee, as it is the top-selling category.
- Apply discounts within the optimal range of 10% to 20% to maximize sales.
- Investigate the limited impact of events on sales and consider increasing promotional activities during events to boost sales.
- Focus marketing efforts on product quality and customer experience instead of weather conditions. Weather is not a factor in Sales growth.
- Explore the potential for seasonal factors in future sales by generating more data and to make more accurate forecasts.

Feedback

- Investigate the source of negative Sale Amount and Quantity Sold values and improve data validation processes.

- Expand event data and analyze its broader impact on sales, considering different types and scales of events.
- Refine discount strategy by conducting more detailed regression analysis to find the optimal discount rate.
- Explore advanced forecasting models and include additional features for improved prediction accuracy.
- Investigate other external factors beyond temperature that might influence sales.
- Incorporate seasonal parameters for more accurate sales forecasts.
- Analyze customer behavior and preferences to tailor marketing strategies and product offerings.