MINI HACKATHON 2023
PRELIMINARY ROUND

TEAM DOOS

UNIVERSITY OF JAFFNA

DECEMBER 3, 2023

a. What synthetic features are you able to generate from the provided data sets? Compile a table with the name, a brief description, and the data type of the synthetic features. (4 marks)

Name	Description	Data type
week_of_month	Generated From week_start_date and shows which week of the month	int64
day_of_week	Generated From week_start_date and shows which day of the week (assumed the week starts on Monday, which is denoted by 0 and ends on Sunday which is denoted by 6)	int64
month	Generated From week_start_date and shows which month of the year	int64
quarter	Generated From week_start_date and shows which quarter of the year	int64
year	Generated From week_start_date and shows which year	int64

b. Conduct EDA on the datasets. You need to generate the summary statistics for all the synthetic feature columns as well. Include 4 key observations from the EDA. (8 marks)

Summary statistics for all synthetic feature columns

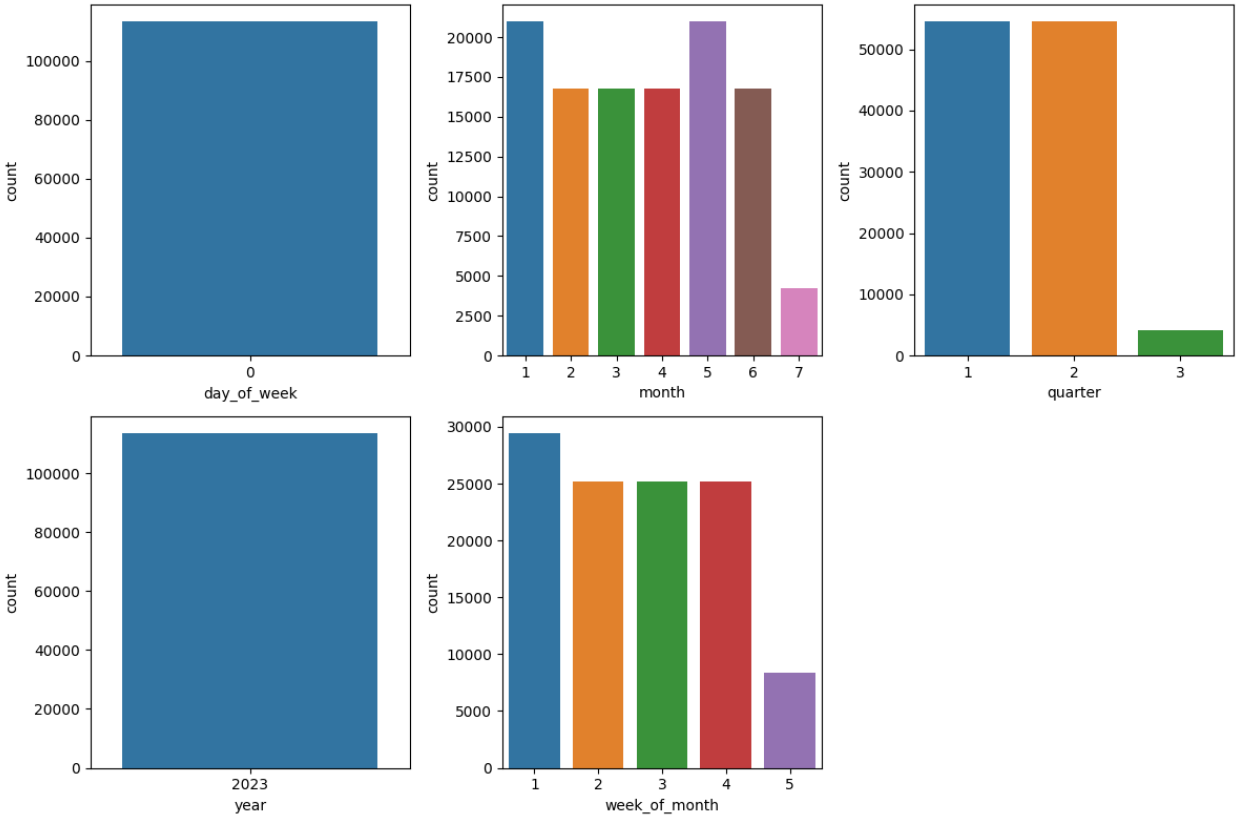
	day_of_week	month	quarter	year	week_of_month
Count	113400.0	113400.000000	113400.000000	113400.0	113400.000000
Mean	0.0	3.592593	1.555556	2023.0	2.629630
Std	0.0	1.831002	0.566560	0.0	1.280866
Min	0.0	1.000000	1.000000	2023.0	1.000000
25%	0.0	2.000000	1.000000	2023.0	1.000000
50%	0.0	4.000000	2.000000	2023.0	3.000000
75%	0.0	5.000000	2.000000	2023.0	4.000000
max	0.0	7.000000	3.000000	2023.0	5.000000

Observations

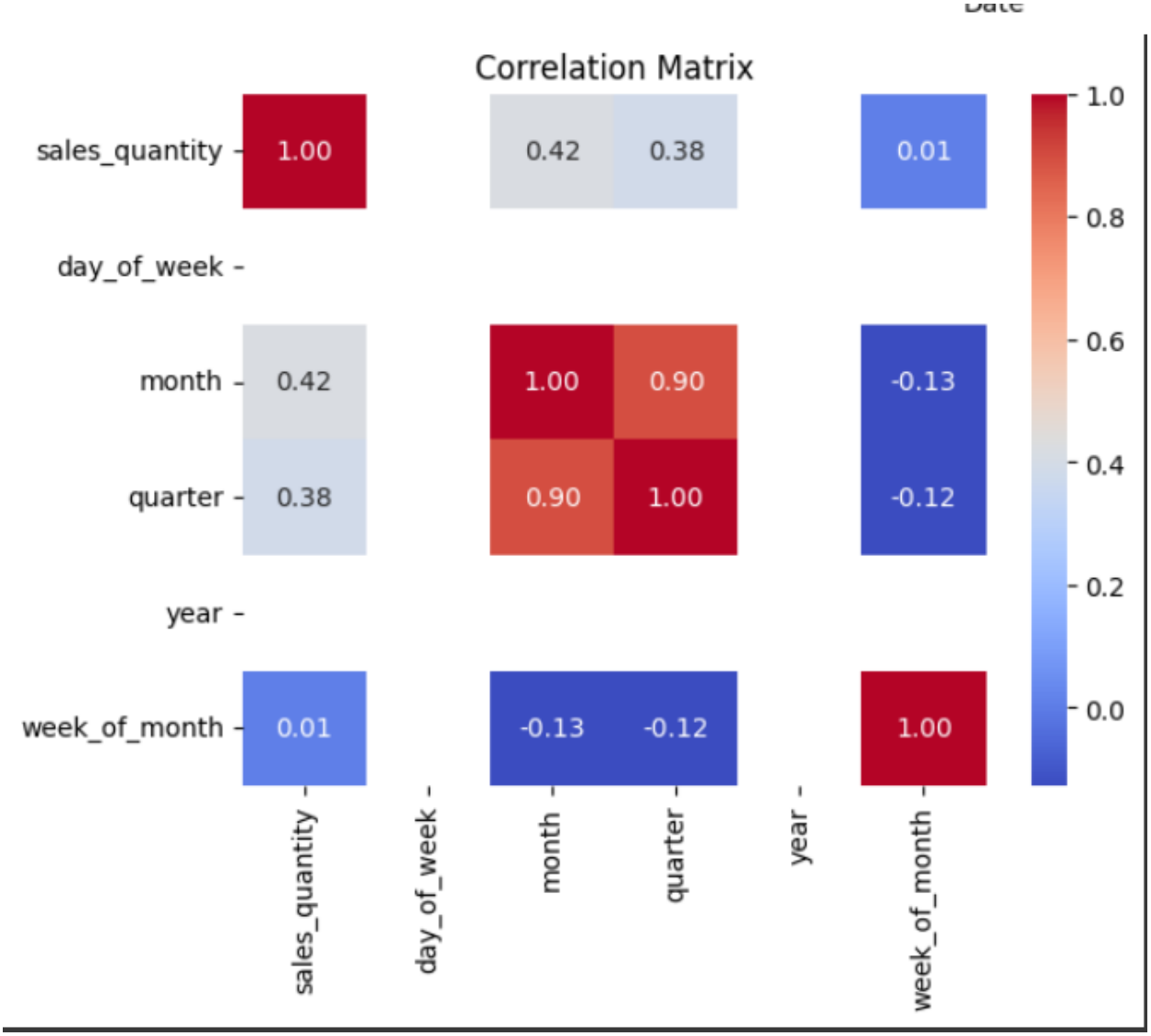
Outlier Detection on ‘sales_quantity’

```
Outliers: (2076, 18)
```

Examine the distribution of sales across synthetic features.



Correlation Matrix over synthetic features.



c. What is the selected target variable? How did you engineer it? (4 marks)

Target variable - sales_quantity

d. What was your approach for forecasting the sales? Justify your methodology. (4 marks)

For forecasting the sales quantity, we adopted a multi-method approach leveraging a combination of linear regression, decision tree regression, logistic regression, and random forest regression models. The choice of these models was driven by the nature of the data and the need to capture different aspects of the underlying patterns. Additionally, mutual information was employed for feature engineering to enhance the model's predictive capabilities.

e. Confectionery Manufacturer Company X has instructed your team that they are looking for data insights and they have specifically asked for the following questions. (10 marks)

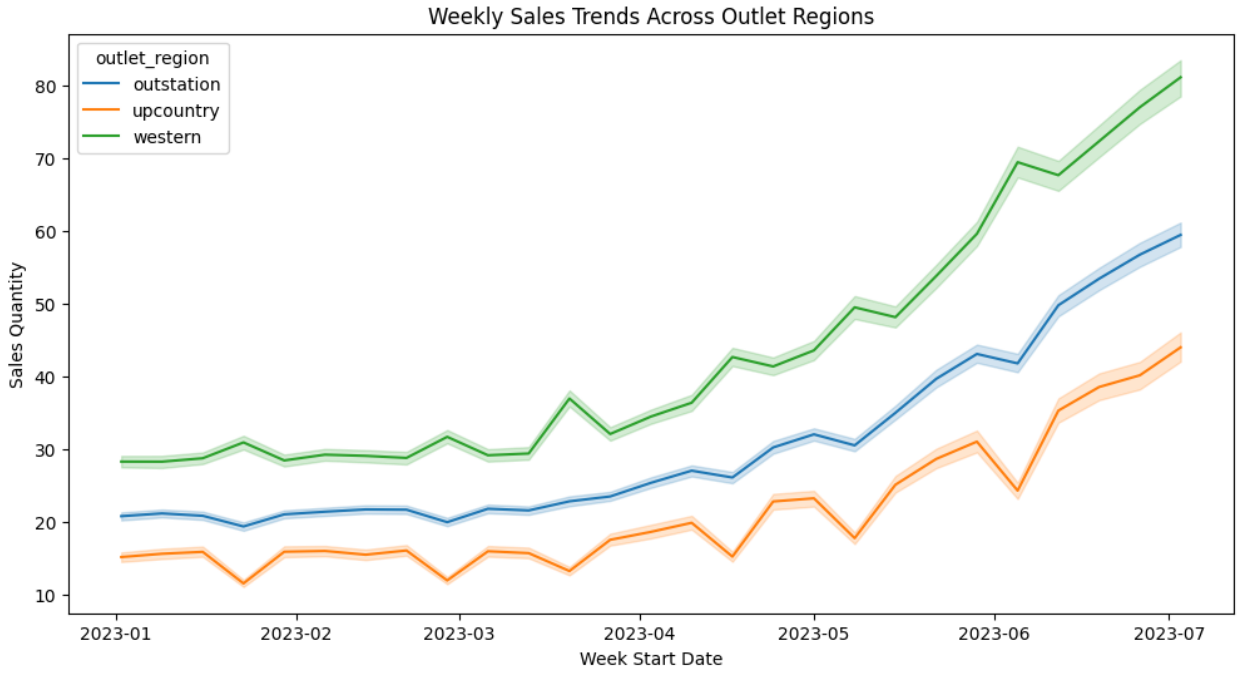
- **What are the average weekly sales volumes for each outlet region?**

```
Average Weekly Sales Volumes by Outlet Region:
outlet_region
outstation    30.669778
upcountry     21.515503
western       43.271235
Name: sales_quantity, dtype: float64
```

- **How do you assess the impact of rainfall on the weekly sales volumes? Is there any correlation between rainfall and the total weekly sales?**

```
Correlation between Rainfall and Weekly Sales Volumes: -0.0935959187399799
```

- **Visualize the weekly sales trends across different outlet regions. Are there any noticeable patterns in the sales data?**



Yes, considering upcountry and outstation both patterns behave like same. But Western pattern, at some points, there are some opposite movements.