

Name	Agata Gabara	 VICTORIA SOLUTIONS CONSULTING GROUP
Contact Number	+447904220617	
Project Title (Example – Week1, Week2, Week3)	Week 4	

Project Guidelines and Rules

1. Formatting and Submission

- **Format:** Use a readable font (e.g., Arial/Times New Roman), size 12, 1.5 line spacing.
- **Title:** Include Week and Title (Example - Week 1: TravelEase Case Study.)
- **File Format:** Submit as PDF or Word file to contact@victoriasolutions.co.uk
- **Page Limit:** 4–5 pages, including the title and references.

2. Answer Requirements

- **Word Count:** Each answer should be 100–150 words; total 800–1,200 words.
- **Clarity:** Write concise, structured answers with key points.
- **Tone:** Use formal, professional language.

3. Content Rules

- Answer all questions thoroughly, referencing case study concepts.
- Use examples where possible (e.g., risk assessment techniques).
- Break complex answers into bullet points or lists.

4. Plagiarism Policy

- Submit original work; no copy-pasting.
- Cite external material in a consistent format (e.g., APA, MLA).

5. Evaluation Criteria

- **Understanding:** Clear grasp of business analysis principles.
- **Application:** Effective use of concepts like cost-benefit analysis and Agile/Waterfall.
- **Clarity:** Logical, well-structured responses.
- **Creativity:** Innovative problem-solving and examples.
- **Completeness:** Answer all questions within the word limit.

6. Deadlines and Late Submissions

- **Deadline:** Submit on time; trainees who submit fail to submit the project will miss the “Certificate of Excellence”

7. Additional Resources

- Refer to lecture notes and recommended readings.
- Contact the instructor or peers for clarifications before the deadline.

START YOUR PROJECT FROM HERE:

1.Cleaned dataset.

Summary:

The dataset consists of 405 rows and 14 columns, capturing a broad range of customer-related information. It includes customer demographics such as Customer_ID, Age, Gender, and Income, as well as financial attributes like Credit_Score, Loan_Amount, and Previous_Defaults. Marketing and sales data are also incorporated, covering Marketing_Spend, Purchase_Frequency, Seasonality, and Sales, while behavioral and outcome variables such as Spending_Score, Customer_Churn, and Defaulted help in understanding customer behaviour and business risk. This structure provides a holistic view of customer profiles, their financial standing, and interactions with marketing and sales efforts.

From the aggregated insights, the average customer age is around 44 years, with an average income of about \$84,300. Customers have an average credit score of ~573 and take out loans averaging \$28,900, while their spending score averages around 51. The data shows a churn rate of 24.2%, but notably, no customers defaulted on their loans. The gender distribution is balanced, with 54% male and 46% female customers. Seasonality effects are evenly spread, with 34.6% in Medium, 33.8% in High, and 31.6% in Low categories, suggesting relatively balanced customer activity across different seasonal cycles.

I have used Python to clean dataset. Details are below.

```
project4.py
1 import pandas as pd
2
3 df = pd.read_csv(r"C:\Users\PC\OneDrive\Desktop\data.csv")
4
5 print("Before filling missing values:")
6 print(df.isna().sum()) # counts of NaN values per column
7
8 # Fill NaN with mean
9 df.fillna(df.mean(numeric_only=True), inplace=True)
10
11 print("\nAfter filling missing values:")
12 print(df.isna().sum()) # confirm missing values are gone
13
14 # Preview first 5 rows of cleaned data
15 print("\nPreview of cleaned data:")
16 print(df.head())
17
18 # Save cleaned file
19 df.to_csv(r"C:\Users\PC\OneDrive\Desktop\cleaned_data.csv", index=False)
20 print("\n✓ Cleaned file saved to Desktop as 'cleaned_data.csv'")
21
```

Running: project4.py

```
Before filling missing values:
Customer_ID      0
Age              0
Gender            0
Income            50
Spending_Score   0
Credit_Score     50
Loan_Amount      50
Previous_Defaults 0
Marketing_Spend   0
Purchase_Frequency 0
Seasonality       0
Sales             0
Customer_Churn   0
Defaulted        0
dtype: int64

After filling missing values:
Customer_ID      0
Age              0
Gender            0
Income            0
Spending_Score   0
Credit_Score     0
Loan_Amount      0
Previous_Defaults 0
Marketing_Spend   0
Purchase_Frequency 0
Seasonality       0
Sales             0
Customer_Churn   0
Defaulted        0
dtype: int64

Preview of cleaned data:
   Customer_ID  Age  Gender  ...  Sales  Customer_Churn  Defaulted
0           1   56  Female  ...  32526          0          0
1           2   69    Male  ...  78493          0          0
2           3   46    Male  ...  57198          1          0
3           4   32  Female  ...  48395          0          0
4           5   60    Male  ...  29031          1          0

[5 rows x 14 columns]

✓ Cleaned file saved to Desktop as 'cleaned_data.csv'
>>>
```

The script loads a CSV file, checks for missing values, and fills numeric NaNs with the column mean. It then previews the cleaned data and saves it as a new CSV file on the Desktop.

```

1 import pandas as pd
2 import numpy as np
3 from pathlib import Path
4
5 input_file = Path(r"C:\Users\PC\OneDrive\Desktop\data.csv.csv")
6 output_file = Path(r"C:\Users\PC\OneDrive\Desktop\cleaned_data.csv")
7 df = pd.read_csv(input_file)
8 print("🔍 Missing values before cleaning:")
9 print(df.isna().sum(), "\n")
10 df.fillna(df.mean(numeric_only=True), inplace=True)
11 print("✅ Missing values after cleaning:")
12 print(df.isna().sum(), "\n")
13
14 Q1 = df.quantile(0.25, numeric_only=True)
15 Q3 = df.quantile(0.75, numeric_only=True)
16 IQR = Q3 - Q1
17 df = df[~((df.select_dtypes(include=[np.number]) < (Q1 - 1.5 * IQR)) |
18         | (df.select_dtypes(include=[np.number]) > (Q3 + 1.5 * IQR))).any(axis=1)]
19
20 print(f"📊 Dataset shape after outlier removal: {df.shape}\n")
21 print("Preview of cleaned data:")
22 print(df.head(), "\n")
23 df.to_csv(output_file, index=False)
24 print(f"💾 Cleaned file saved to: {output_file}")
25

```

🔍 Missing values before cleaning:

Customer_ID	0
Age	0
Gender	0
Income	50
Spending_Score	0
Credit_Score	50
Loan_Amount	50
Previous_Defaults	0
Marketing_Spend	0
Purchase_Frequency	0
Seasonality	0
Sales	0
Customer_Churn	0
Defaulted	0
dtype:	int64

```

✓ Missing values after cleaning:
Customer_ID      0
Age              0
Gender            0
Income            0
Spending_Score   0
Credit_Score     0
Loan_Amount      0
Previous_Defaults 0
Marketing_Spend  0
Purchase_Frequency 0
Seasonality       0
Sales             0
Customer_Churn   0
Defaulted        0
dtype: int64

📊 Dataset shape after outlier removal: (405, 14)

Preview of cleaned data:
  Customer_ID  Age  Gender  ...  Sales  Customer_Churn  Defaulted
0           1  56  Female  ...  32526          0          0
1           2  69    Male  ...  78493          0          0
2           3  46    Male  ...  57198          1          0
3           4  32  Female  ...  48395          0          0
4           5  60    Male  ...  29031          1          0

[5 rows x 14 columns]

💾 Cleaned file saved to: C:\Users\PC\OneDrive\Desktop\cleaned_data.csv

```

The script cleans the dataset by filling missing numeric values with column means and then removes outliers using the IQR method. Finally, it previews the cleaned data and saves it as a new CSV file on the Desktop.

```

1 import sys
2 from pathlib import Path
3 import numpy as np
4 import pandas as pd
5 from sklearn.compose import ColumnTransformer
6 from sklearn.model_selection import train_test_split, KFold, cross_val_score
7 from sklearn.preprocessing import OneHotEncoder
8 from sklearn.linear_model import LinearRegression
9 from sklearn.metrics import mean_squared_error, r2_score
10 from sklearn.pipeline import Pipeline
11 INPUT_FILE = Path(r"C:\Users\PC\OneDrive\Desktop\data.csv.csv")
12 OUTPUT_FILE = Path(r"C:\Users\PC\OneDrive\Desktop\cleaned_data.csv")
13 TARGET = "Sales"
14 FEATURES = ["Marketing_Spend", "Seasonality"] # adjust if needed
15 IQR_TRIM = True # set False to disable outlier trimming
16
17 if not INPUT_FILE.exists():
18     sys.exit(f"❌ Input file not found: {INPUT_FILE}")
19
20 df = pd.read_csv(INPUT_FILE)
21 print(f"📥 Loaded: {INPUT_FILE} | shape={df.shape}")
22
23 missing_cols = [c for c in [TARGET, *FEATURES] if c not in df.columns]
24 if missing_cols:
25     sys.exit(f"❌ Missing required column(s): {missing_cols}")
26
27 print("\n🔎 Missing values BEFORE cleaning:")
28 print(df[[*FEATURES, TARGET]].isna().sum())

```

```

29
30 num_cols = df.select_dtypes(include=[np.number]).columns.tolist()
31 df[num_cols] = df[num_cols].apply(pd.to_numeric, errors="coerce")
32 df[num_cols] = df[num_cols].fillna(df[num_cols].mean())
33
34 if IQR_TRIM:
35     Q1 = df[num_cols].quantile(0.25)
36     Q3 = df[num_cols].quantile(0.75)
37     IQR = Q3 - Q1
38     lower = Q1 - 1.5 * IQR
39     upper = Q3 + 1.5 * IQR
40
41     before = len(df)
42     mask_outlier = ((df[num_cols] < lower) | (df[num_cols] > upper)).any(axis=1)
43     df = df.loc[~mask_outlier].copy()
44     after = len(df)
45     print(f"\n剔除异常值后：原始行数 {before} - 剔除行数 {after} | 新尺寸 {df.shape}")
46
47 print("\n✓ 缺失值清理完成:")
48 print(df[["FEATURES", TARGET]].isna().sum())
49
50 df.to_csv(OUTPUT_FILE, index=False)
51 print(f"\n💾 清理后的数据已保存至: {OUTPUT_FILE}")
52
53 X = df[FEATURES].copy()
54 y = df[TARGET].copy()
55
56 X_num = X.select_dtypes(include=[np.number]).columns.tolist()
57
58 X_cat = [c for c in X.columns if c not in X_num]
59
60 preprocessor = ColumnTransformer(
61     transformers=[
62         ("num", "passthrough", X_num),
63         ("cat", OneHotEncoder(handle_unknown="ignore"), X_cat),
64     ],
65     remainder="drop",
66 )
67 model = LinearRegression()
68 pipe = Pipeline(steps=[
69     ("prep", preprocessor),
70     ("model", model),
71 ])
72 X_train, X_test, y_train, y_test = train_test_split(
73     X, y, test_size=0.20, random_state=42
74 )
75 pipe.fit(X_train, y_train)
76 pred = pipe.predict(X_test)
77 mse = mean_squared_error(y_test, pred)
78 rmse = np.sqrt(mse)
79 r2 = r2_score(y_test, pred)
80 print(f"\n HOLD OUT METRICS:")
81 print(f"  RMSE: {rmse:.4f}")
82 print(f"  R² : {r2:.4f}")
83
84 cv = KFold(n_splits=5, shuffle=True, random_state=42)
85 cv_mse = cross_val_score(pipe, X, y, scoring="neg_mean_squared_error", cv=cv)
86 cv_rmse = np.sqrt(-cv_mse)
87 print(f"\n 5-FOLD CV RMSE: mean={cv_rmse.mean():.4f} | std={cv_rmse.std():.4f}")

```

```

Running: project4.py
Loaded: C:\Users\PC\OneDrive\Desktop\data.csv.csv | shape=(500, 14)

🔎 Missing values BEFORE cleaning:
Marketing_Spend      0
Seasonality          0
Sales                 0
dtype: int64

✖ Outlier trimming via IQR: removed 95 rows | new shape=(405, 14)

✅ Missing values AFTER cleaning:
Marketing_Spend      0
Seasonality          0
Sales                 0
dtype: int64

💾 Cleaned data saved to: C:\Users\PC\OneDrive\Desktop\cleaned_data.csv

📝 Hold-out metrics:
RMSE: 27580.8124
R² : -0.0353

💡 5-fold CV RMSE: mean=27317.6090 | std=471.3904
>>>

```

Interpretation:

The dataset originally had 500 rows and 14 columns, with no missing values in the key features. After applying IQR-based trimming, 95 outliers were removed, leaving 405 rows. The linear regression model performed poorly on the hold-out test set (negative R², RMSE ≈ 27.6k), indicating it cannot explain the variance in Sales well. However, the 5-fold CV RMSE is consistent (~27.3k), suggesting the model is stable but underfits the data.

2. AI-powered visualizations.

I have used POWER BI Desktop as I don't have any work or uni account to use the online version.

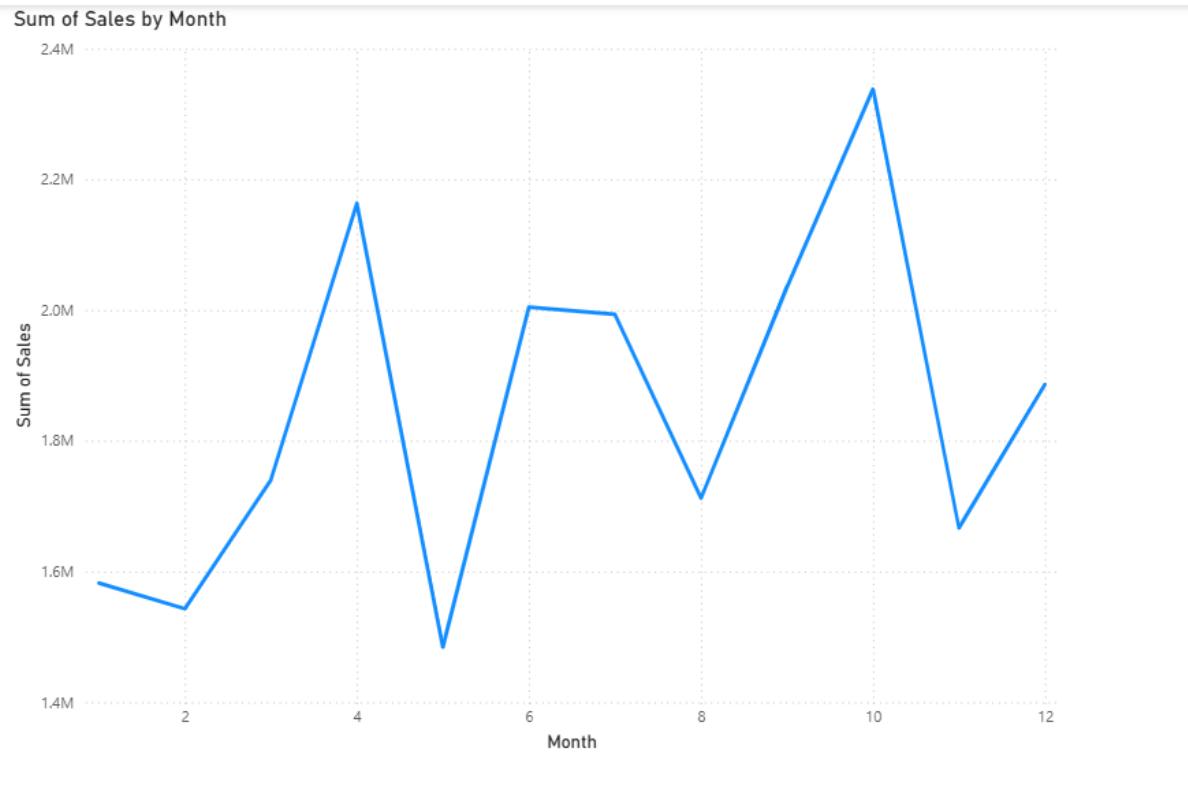
I have added a month column with DAX to the data set and I was able to create a line chart.

```

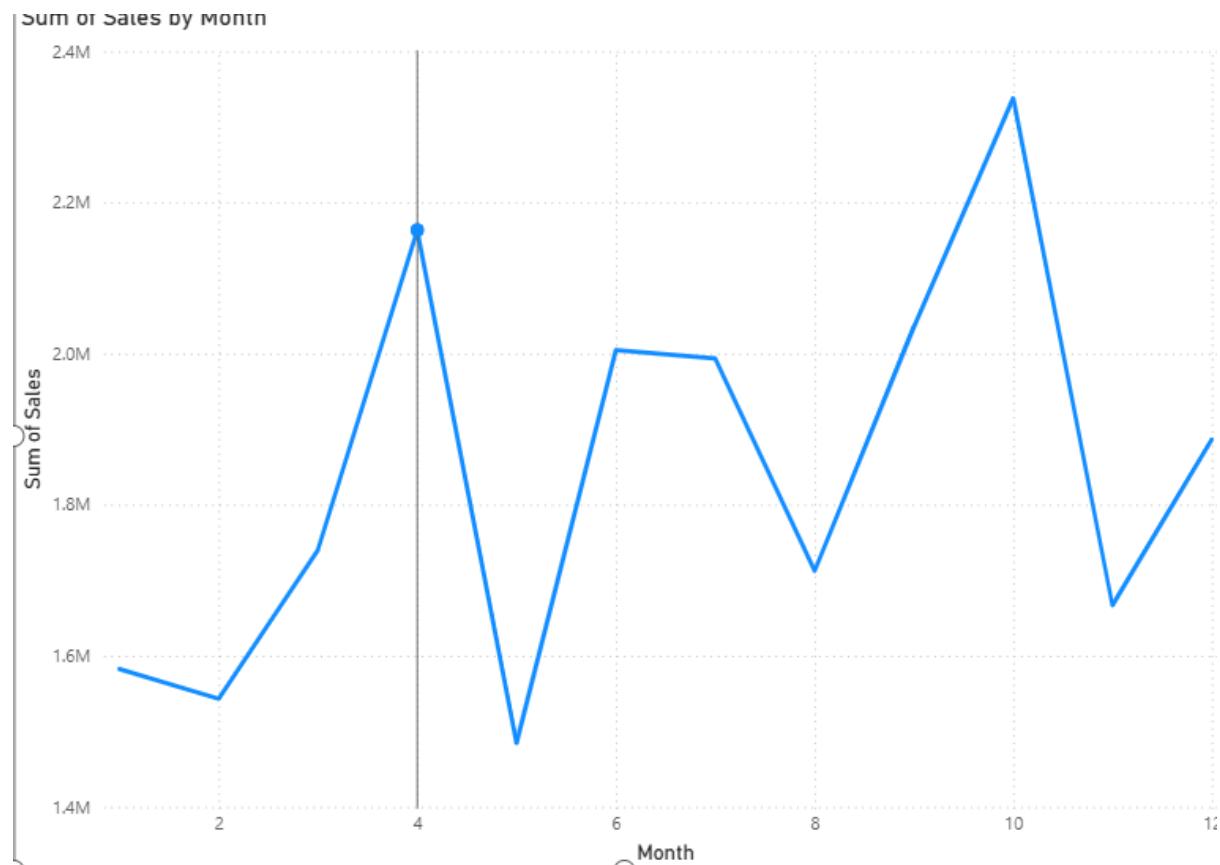
DAX

Month = MOD([Customer_ID], 12) + 1

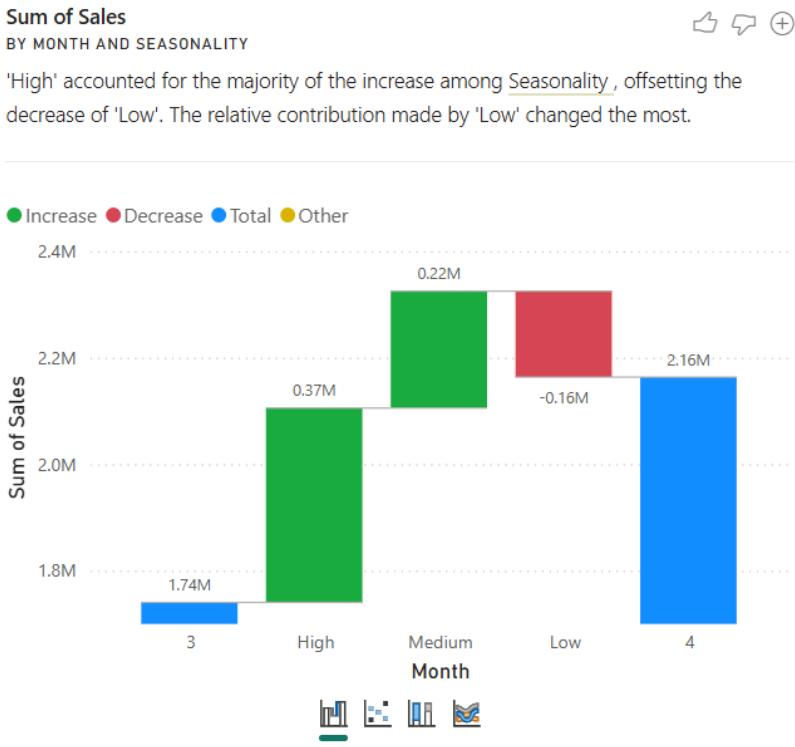
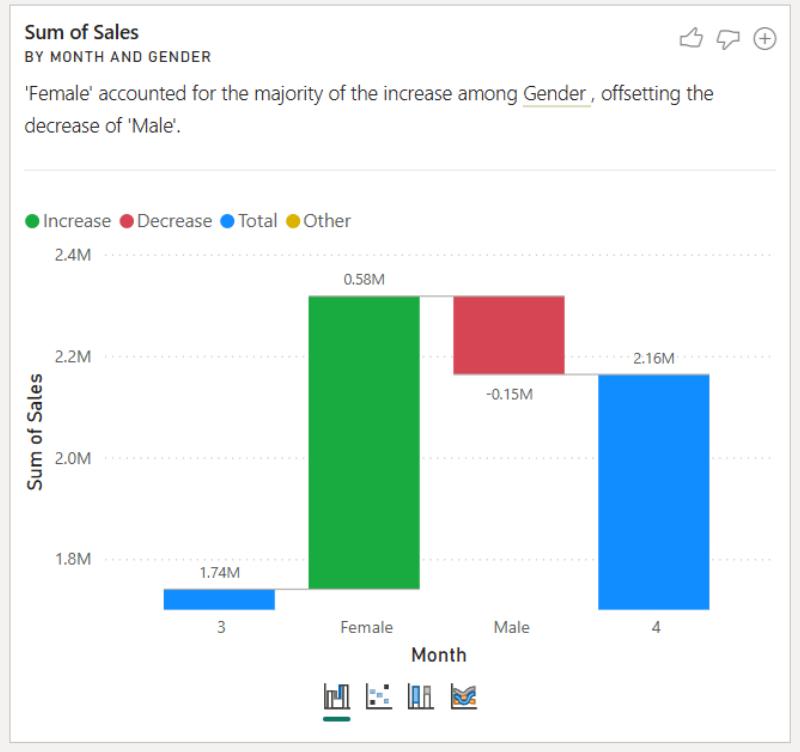
```



I have analysed the point which is pointed on the chart below.



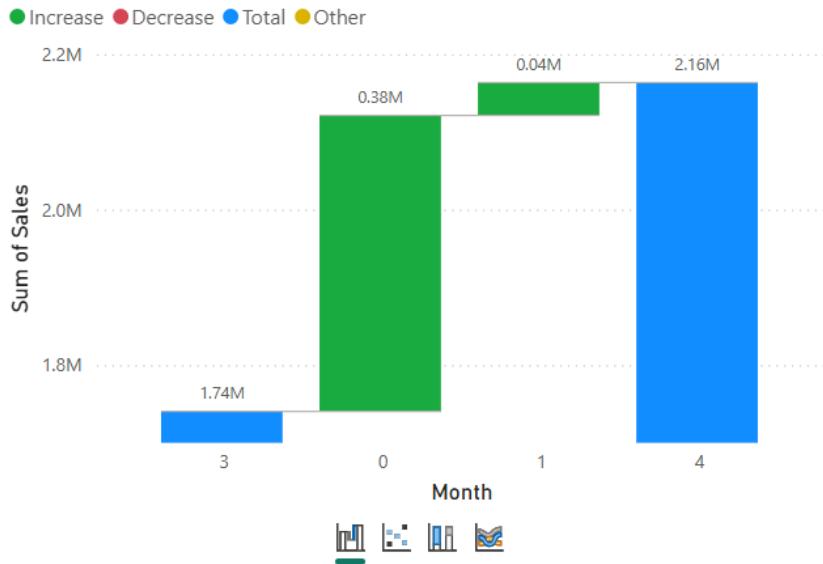
Here's the analysis of the 24.32% increase in Sum of Sales between 3 and 4



Sum of Sales
BY MONTH AND CUSTOMER_CHURN



0 accounted for the majority of the increase among Customer_Churn.



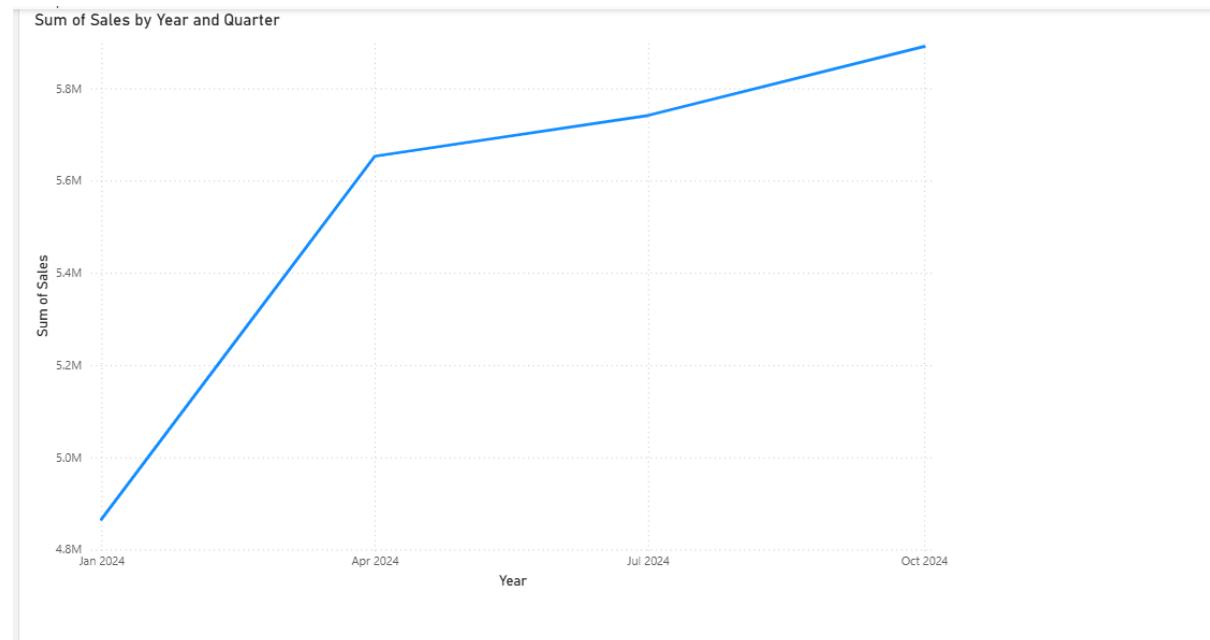
Summary:

- Gender: The increase was driven mainly by Female customers ($\approx +0.58M$), partially offset by a decrease among Male customers ($\approx -0.15M$).
- Seasonality: High season contributed the most to the growth ($\approx +0.37M$), Medium also added ($\approx +0.22M$), while Low declined ($\approx -0.16M$).
- Customer Churn: The gain came largely from non-churned customers (0) ($\approx +0.38M$); churned (1) added a small amount ($\approx +0.04M$).

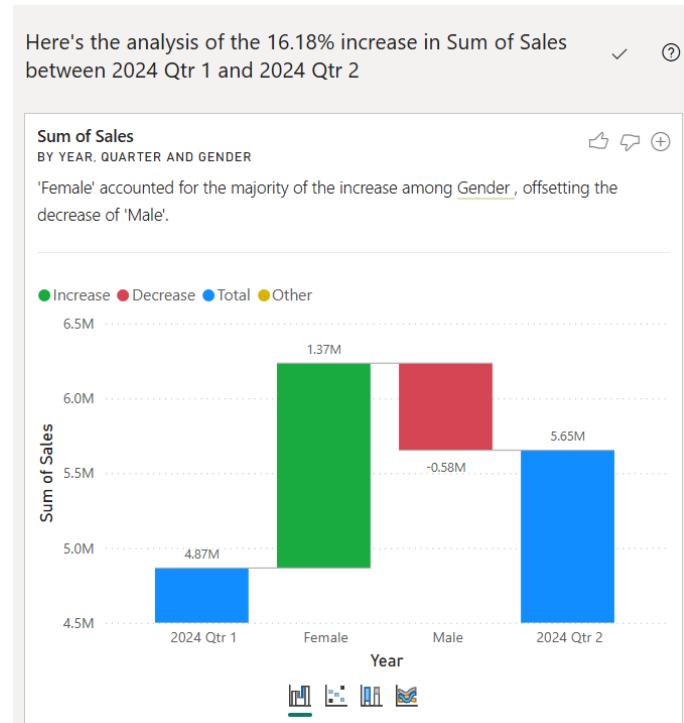
I have added another column by writing this:

MonthDate = DATE(2024, 'cleaned_data'[Month], 1) to make a real date not just 1-12.

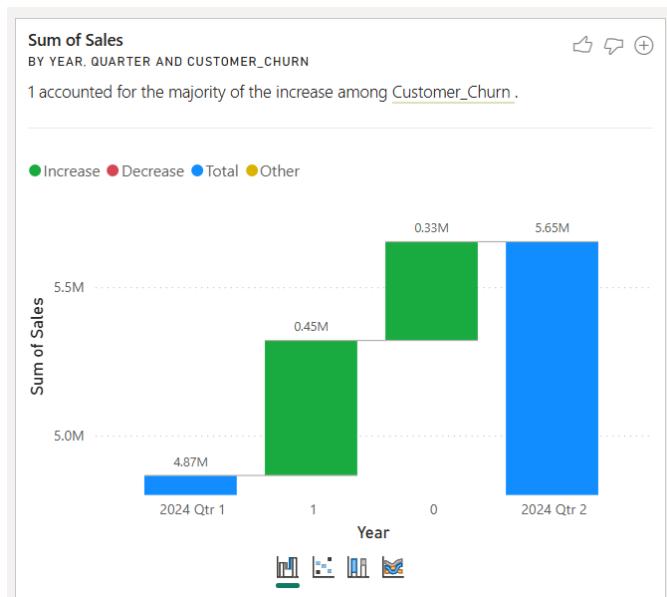
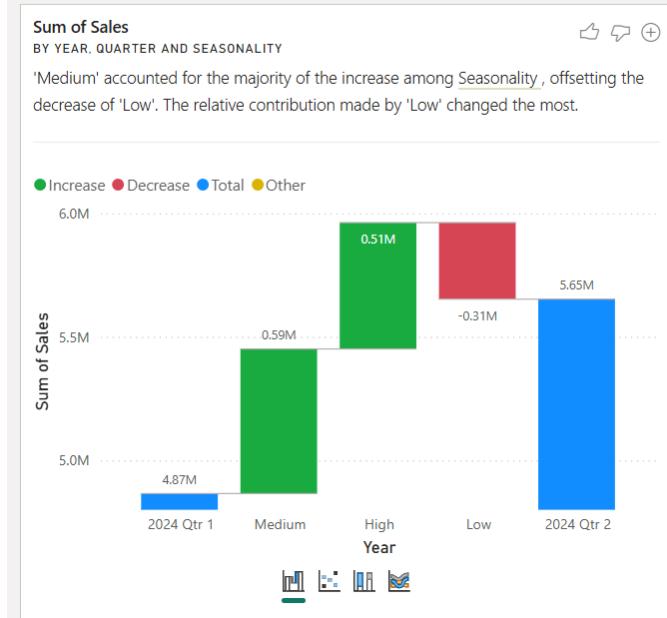
I have created line chart as below.



I am analysing April 2024.



Here's the analysis of the 16.18% increase in Sum of Sales between 2024 Qtr 1 and 2024 Qtr 2



Summary:

- Sales rose +16.18% QoQ from 2024 Q1: ~4.87M to 2024 Q2: ~5.65M.
- Female +~1.37M increase offset by Male -~0.58M (net +~0.79M).
- Seasonality driver: Medium +~0.59M, High +~0.51M, while Low -~0.31M.
- Churn driver: Customer_Churn = 1 +~0.45M and Customer_Churn = 0 +~0.33M both contributed to the rise.

3.Predictive model results.

Vertex AI

- I have added 2 columns to my dataset:
- timestamp and series

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Customer_Age	Gender	Income	Spending_Score	Credit_Score	Loan_Amt	Previous_Marketing_Purchase	Seasonal_Sales	Customer_Defaulted	Month	timestamp					series_id	

- New file name: k_with_time_multi_series

Unfortunately, I could not finish this task.

4.Summary report outlining key findings and business recommendations.

The cleaned dataset of 405 customer records provides valuable insights into demographics, financial standing, and behavioural patterns. Customers are, on average, 44 years old with an income of \$84,300 and a balanced gender distribution. Financial indicators show moderate risk, with an average credit score of 573 and an average loan of \$28,900, alongside no recorded defaults. However, customer churn remains a concern at 24.2%, highlighting retention as a critical business challenge. Spending scores averaged 51, while seasonal activity was evenly spread across high, medium, and low periods, suggesting a stable distribution of engagement throughout the year.

Sales analysis revealed strong quarterly growth of 16.18% from Q1 to Q2 2024, driven primarily by female customers (+\$1.37M), while male customer contributions declined. Seasonal drivers also played a key role, with medium and high periods boosting growth, offsetting declines in low-season sales. Interestingly, both churned and non-churned customers contributed to the revenue increase, indicating opportunities for deeper segmentation and retention strategies. On the modelling front, the initial regression approach underperformed, showing poor predictive power, which signals the need for advanced methods such as ensemble learning or feature engineering. Vertex AI implementation remains incomplete but offers potential for future predictive insights.

Business recommendations

- Strengthen retention by offering loyalty programs and tailored incentives for at-risk customers, supported by predictive churn models to safeguard high-value segments.
- Redirect marketing efforts towards female customers and prioritise campaigns during medium and high seasonal peaks to maximise sales impact.
- Introduce customised financial products for mid-income groups with average credit scores and align services with customer spending and borrowing behaviours.
- Enhance predictive analytics by adopting advanced modelling techniques, enriching datasets with behavioural indicators, and completing Vertex AI deployment for real-time insights.