Project Coversheet

June 6, 2025

Full Name	Jude Spellacy
Email	judespellacy@gmail.com
Contact Number	07891187181
Project Title	week 2

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: Travel Ease Case Study.)
- File Format: Submit as PDF or Word file.
- Page Limit: 4–5 pages, including the title and references.

2. Answer Requirements

- Word Count: Each answer should be within 100–150 words; Maximum 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

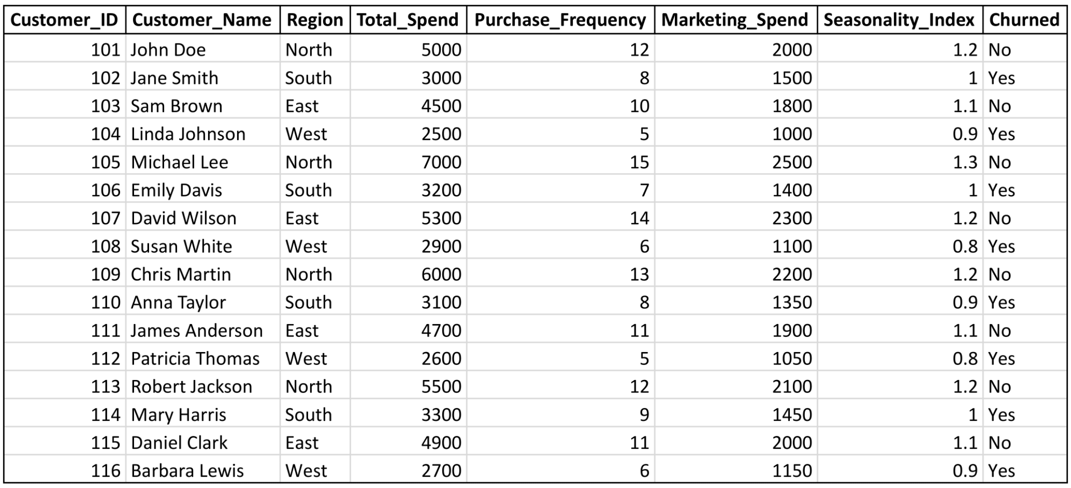
- Deadline: Submit on time; trainees who fail to submit the project will miss the “Certificate of Excellence”.

7. Additional Resources

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

Project 2: Data Cleaning, Analysis, and Business Insights

1 Data Wrangling and Summary



Customer_ID	Customer_Name	Region	Total_Spend	Purchase_Frequency	Marketing_Spend	Seasonality_Index	Churned
101	John Doe	North	5000	12	2000	1.2	No
102	Jane Smith	South	3000	8	1500	1	Yes
103	Sam Brown	East	4500	10	1800	1.1	No
104	Linda Johnson	West	2500	5	1000	0.9	Yes
105	Michael Lee	North	7000	15	2500	1.3	No
106	Emily Davis	South	3200	7	1400	1	Yes
107	David Wilson	East	5300	14	2300	1.2	No
108	Susan White	West	2900	6	1100	0.8	Yes
109	Chris Martin	North	6000	13	2200	1.2	No
110	Anna Taylor	South	3100	8	1350	0.9	Yes
111	James Anderson	East	4700	11	1900	1.1	No
112	Patricia Thomas	West	2600	5	1050	0.8	Yes
113	Robert Jackson	North	5500	12	2100	1.2	No
114	Mary Harris	South	3300	9	1450	1	Yes
115	Daniel Clark	East	4900	11	2000	1.1	No
116	Barbara Lewis	West	2700	6	1150	0.9	Yes

Figure 1: Dataset already clean

There was no need to alter the contents of the data as no clear outliers were identified and there are no duplicate or missing values, therefore the given values are assumed accurate.

1.1 Data Summary

Summarising the data the following was obtained:

Feature	Total Spend	Purchase Frequency	Marketing Spend	Seasonality Index
Mean	4137.50	9.50	1675.00	1.04
Median	3900.00	9.50	1650.00	1.05
Mode	All values are unique.	Multiple values appear twice.	2000.00	1.20
Range	4500.00	10.00	1500.00	0.50
Standard Deviation	1396.13	3.22	484.42	0.15
Total	66200.00	152.00	26800.0	N/A

These give indications on the average spending and frequency as well as the seasonality, the standard deviations also indicate how the sales are distributed with 68% of values falling within the ± of this around the mean.

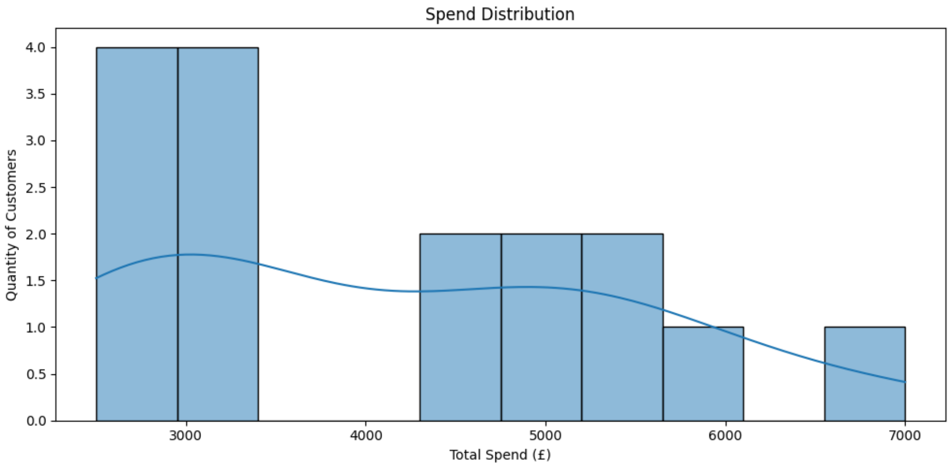


Figure 2: Total Spending distribution over customers

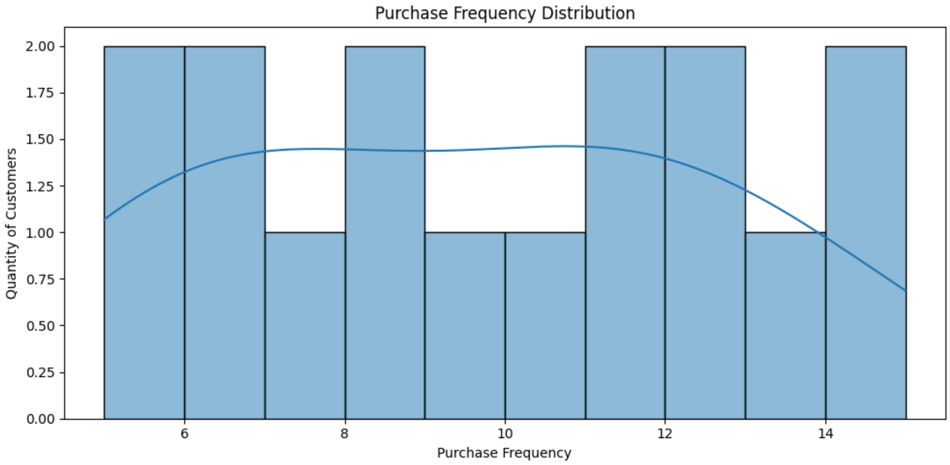
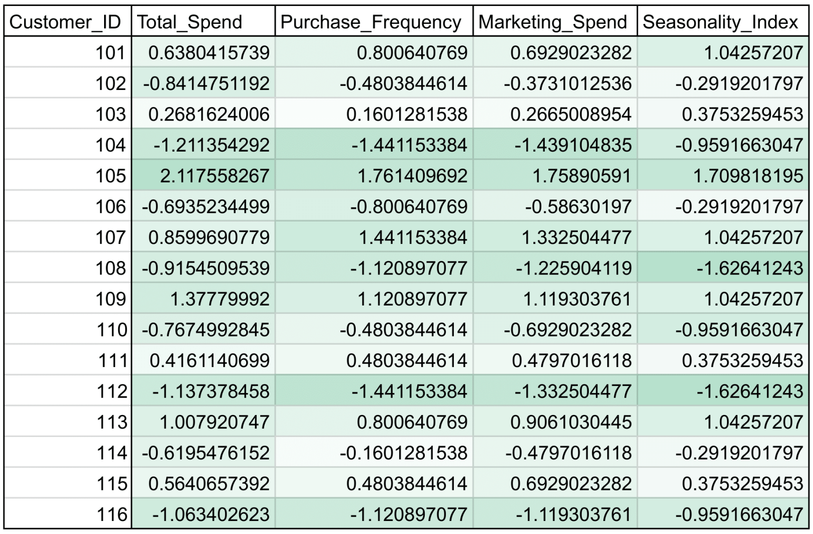


Figure 3: Purchase Frequency distribution over customers

This measure assumes a normal distribution over all features, Figures 2 and 3 demonstrates that this may not be the case for every feature over the data sampled.

The standard deviation allows the calculation of Z-scores to test for outliers, by scoring how many standard deviations out a data point is:



Customer_ID	Total_Spend	Purchase_Frequency	Marketing_Spend	Seasonality_Index
101	0.6380415739	0.800640769	0.6929023282	1.04257207
102	-0.8414751192	-0.4803844614	-0.3731012536	-0.2919201797
103	0.2681624006	0.1601281538	0.2665008954	0.3753259453
104	-1.211354292	-1.441153384	-1.439104835	-0.9591663047
105	2.117558267	1.761409692	1.75890591	1.709818195
106	-0.6935234499	-0.800640769	-0.58630197	-0.2919201797
107	0.8599690779	1.441153384	1.332504477	1.04257207
108	-0.9154509539	-1.120897077	-1.225904119	-1.62641243
109	1.37779992	1.120897077	1.119303761	1.04257207
110	-0.7674992845	-0.4803844614	-0.6929023282	-0.9591663047
111	0.4161140699	0.4803844614	0.4797016118	0.3753259453
112	-1.137378458	-1.441153384	-1.332504477	-1.62641243
113	1.007920747	0.800640769	0.9061030445	1.04257207
114	-0.6195476152	-0.1601281538	-0.4797016118	-0.2919201797
115	0.5640657392	0.4803844614	0.6929023282	0.3753259453
116	-1.063402623	-1.120897077	-1.119303761	-0.9591663047

Figure 4: Z-scores of data points per column

No data point exceeds a reasonable threshold of 3 standard deviations which confirms the initial assesment that the data does not need modifying.

The mean total sales and purchase frequency respectively per region is given in:

North	East	South	West
5875.0	4850.0	3150.0	2675.0
13.0	11.5	8.0	5.5

Showing that the North and East make purchases more frequently and offer a larger total spend each.

2 Forecasting

2.1 Sales Regression

In order to predict the total spend of a customer, a linear model was trained on a subset (80%) of the data, it was then verified on an unseen test set performing as follows:

Real Data	Prediction
5000	5171.14
3000	3748.00
3200	3494.92
4900	5092.27

The parameters for the model given as $\hat{y} = 2.53 \cdot x_1 + 788.75 \cdot x_2 - 836.92$ with x_1 and x_2 representing Marketing Spend and Seasonal Index respectively. This can then be used on new or other existing customers not included in the data when their total spend is unknown, with a similar expected accuracy.

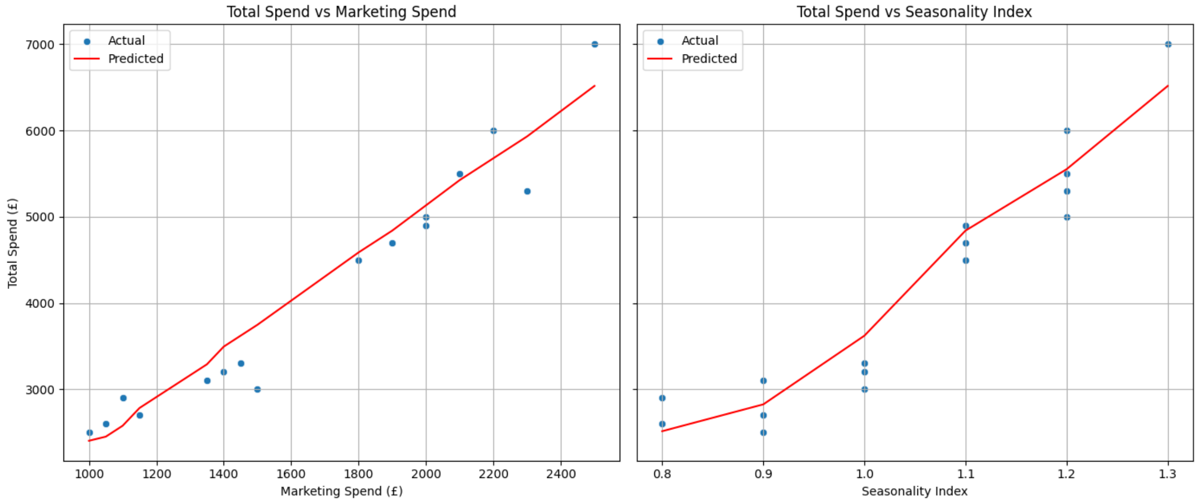


Figure 5: Linear regression in order to forecast total spending based upon Marketing spend and seasonal index

It is clear from both that the model interprets a strong positive correlation with both features.

2.2 Churn Probability

In order to produce a similar regression model to determine whether a customer will stay, a logistic regression model is used which returns a probability based upon the inputs to the model.

The logistic model has a similar structure to the linear regression model and over a similar (80%) training set the following equation was obtained: $\hat{y} = -0.057 \cdot x_1 - 0.000008 \cdot x_2 + 94.66$ with x_1 and x_2 representing Marketing Spend and Seasonal Index respectively. This indicates that the model interprets a much stronger negative correlation between churn probability and marketing spending than with seasonal variation. It is probable that this is due to the fact that these parameters are not normalised so the larger quantities in the marketing spending are out weighing the influence of the seasonal index.

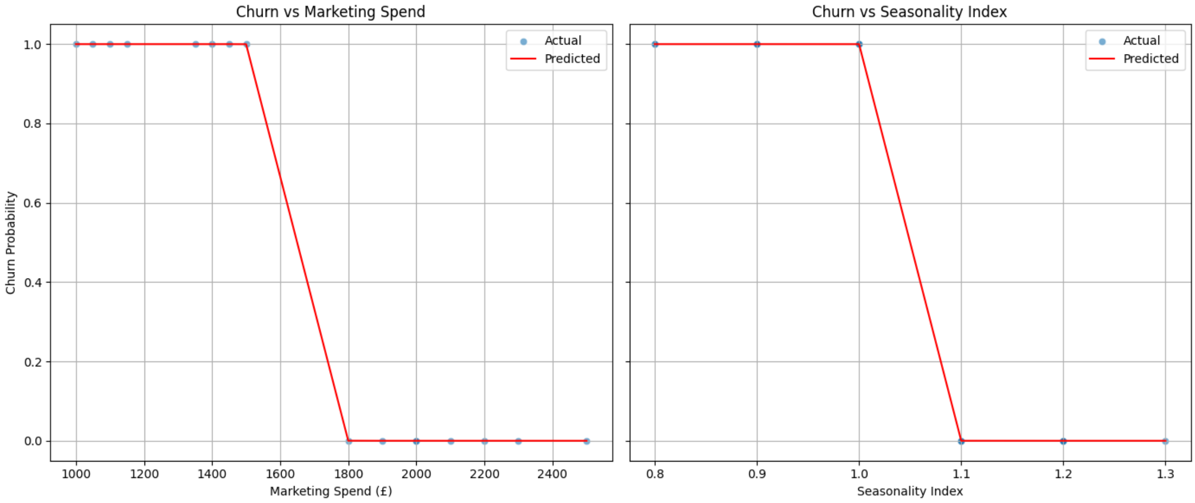


Figure 6: Logistic regression in order to predict churn based upon Marketing spend and seasonal index

3 Analysis

3.1 Regional Behaviour

Across Regions in order to test if they have the same distribution mean, the ANOVA p-value is obtained, over all 4 regions the P-value was in the order of 10^{-6} and so it is extremely unlikely that sales in all of the regions shows the same behaviour.

Performing the same on each Region to each other resulted in:

Region 1	Region 2	P-value
North	East	0.067
North	South	0.0007
North	West	0.0003
East	South	8×10^{-5}
East	West	2×10^{-5}
West	South	0.004

This would indicate that there is less significant evidence against the hypothesis that the distribution means are the same between the North and East but that there is relatively high evidence in all other regions.

3.2 Factor Analysis

Investigating the influence of each feature in the data set on total spend is done through decomposition to find a small number of coefficients like in the earlier linear regression that links the feature to total sales. The decomposition of total sales into three underlying factors was arbitrary and based on the result could have been reduced to two.

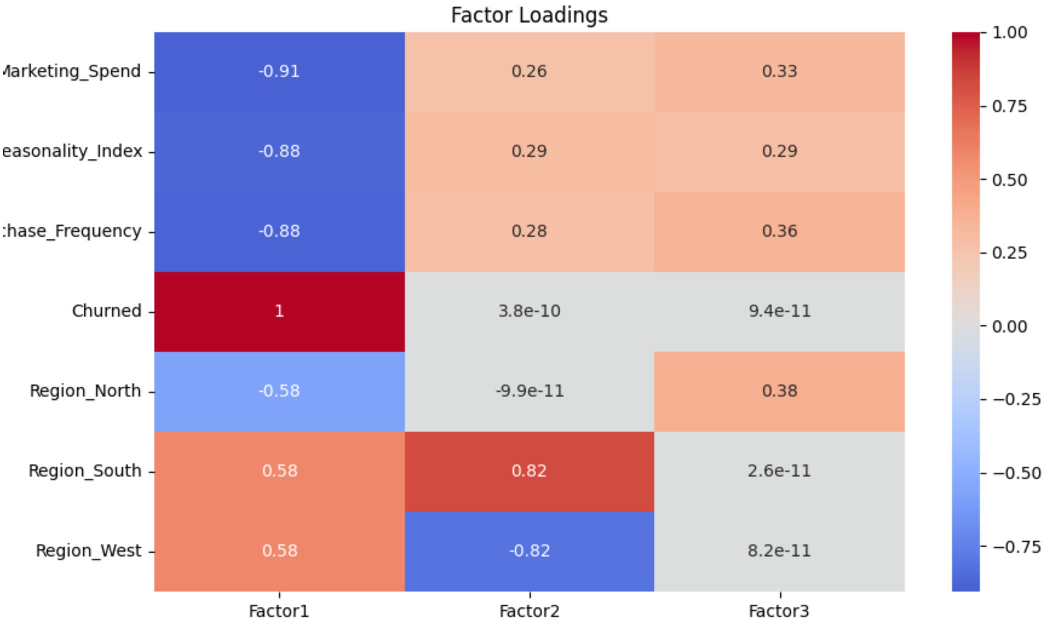


Figure 7: Factor analyis on total spend using decomposition.

It is clear from this that total sales is determined by all the factors to significant extents which is to be expected in real data that is likely related. The most significant factors in the first underlying 'behaviour' of total spend being; marketing spend, seasonality, purchase frequency, and churn. The secondary component of total spend seems to be more correlated with regional behaviour, specifically in the south and west.

4 Customer Classification

K-means clustering was applied over Total Spend, Marketing Spend, Seasonality Index, Purchase Frequency, and whether the customer is Churned. This was devised into three clusters which indicated distinct behaviour shown in Figure 8.

5 Key Findings and Report

Total Spend, highest spenders

- Based on the result of the clustering these high value demographic should be offered greater rewards to avoid churn.
- This group coincides with high marketing spend and this should indicate future actions to be taken.
- This group makes frequent purchases and this rather than low frequency large purchases is the main lure of income, discounts and lower value ranges should be encouraged.

Total Spend, middle range spenders

- There is a secondary group which have potential to draw more income.
- Churn in this and the highest spending groups is significantly absent which means that customer retention is not an issue for the most valuable groups.

Influential Factors

- Marketing spend and region are both actionable factors as shown above, heavier marketing does not correlate with the same underlying factors as region in total spend, so increasing marketing spend in the lower revenue regions is recommended.
- In the limited sample seasonality plays a large part in total spending, it would be advisable to utilise this by increasing marketing in the lead up to high sales seasons, to compensate the others.

Recommended Actions

- Introducing a membership or loyalty policy that focuses on the identified cluster could result in future retention, most importantly marketing this towards the middle of the range customer cluster in order to prevent further loss is recommended.
- Seasonal advertising to all audiences and capitalising on the prime season would be beneficial, depending on what season that is, it may be possible to market with a long lead up especially if the season is also a specific seasonal festival.
- Potentially advertising at lower frequency large spending could bring a larger share of income from the lowest spending cluster, as this group has a low frequency of purchases.

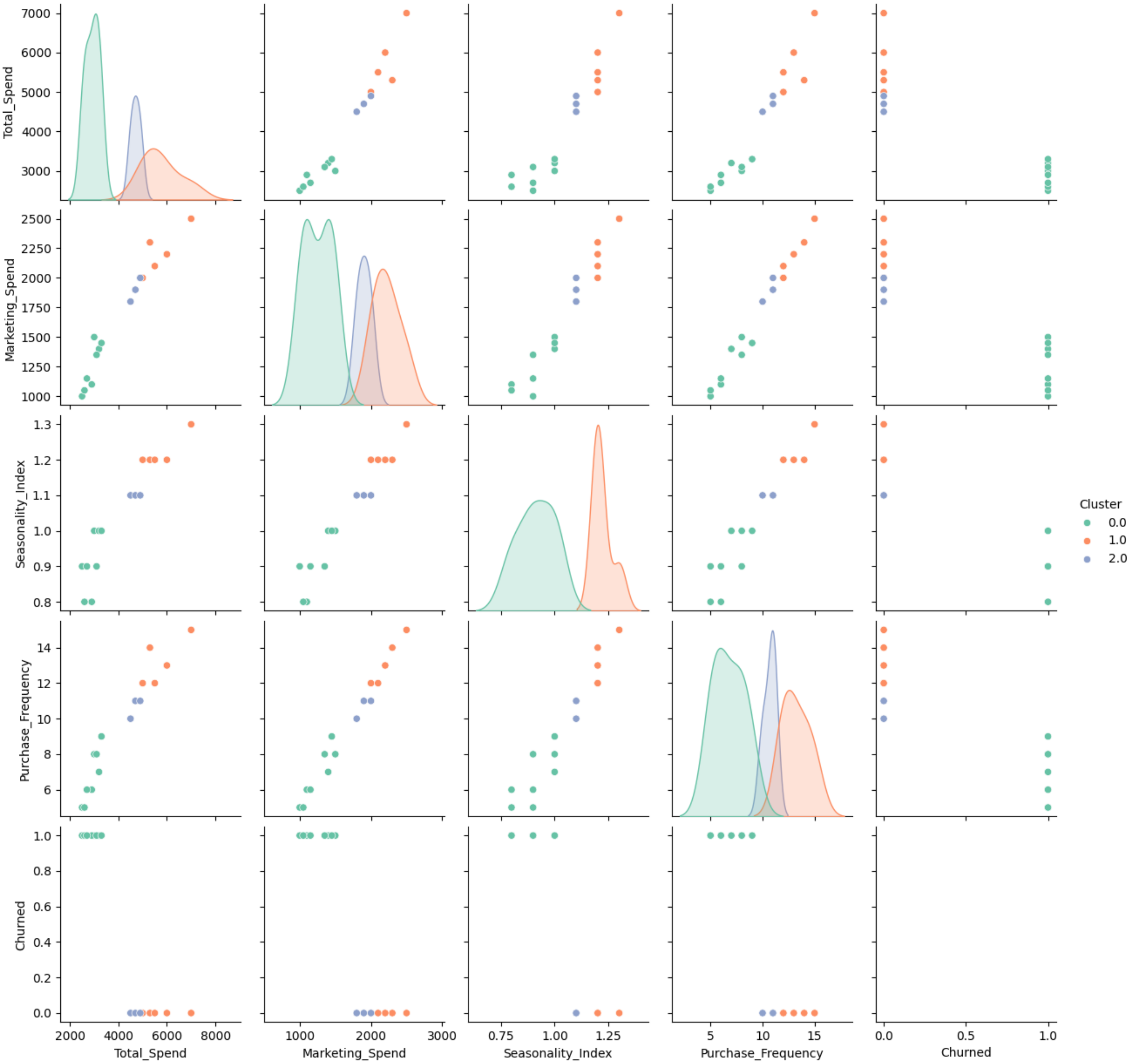


Figure 8: K-means clustering over the data producing three distinct groups of consumers.