*Word count: 1,677*

**Consumer Personality and Purchasing Loyalty:**
**A Quantitative Analysis of Customer Spending Patterns**

## Introduction

In marketing science and consumer psychology, understanding why consumers remain loyal to certain brands is a central question. Customer retention is what businesses invest in when global markets become increasingly competitive, with research suggesting that retaining an existing customer costs substantially less than acquiring a new one (Reichheld and Sasser, 1990). Traditional marketing models have commonly focused on behavioural metrics such as purchase frequency; however, a growing amount of literature argues that consumer personality profiles and demographic characteristics are equally significant factors influencing purchasing loyalty (Fournier, 1998). This report investigates the extent to which consumer demographic and behavioural characteristics predict purchasing loyalty, operationalised as total spending, and whether distinct consumer segments can be identified. The study applies and extends quantitative methods of linear regression, k-means clustering, and logistic regression to a real-world consumer dataset.

## Literature Review

Jacoby and Chestnut (1978) distinguished between behavioural loyalty (repeat purchasing) and attitudinal loyalty (psychological commitment), but the concept of brand loyalty has evolved considerably ever since. Oliver (1999) proposed a four-stage loyalty model advancing from cognitive to action loyalty. However, such stage models have been critiqued by Uncles, Dowling, and Hammond (2003) for lacking empirical validation across diverse product categories, implying that loyalty may instead be context-dependent. The present investigation is driven by this tension between universal and situational models of loyalty.

The idea that consumers establish brand relationships shaped by demographic and lifestyle variables was introduced by Fournier (1998). The RFM (Recency, Frequency, Monetary) framework operationalises this by classifying consumers using quantitative behavioural variables (Fader et al., 2005). Nonetheless, RFM has been criticised for being descriptive rather than predictive despite being widely adopted in industry (Gupta and Zeithaml, 2006). This emphasises the need for methods that represent spending outcomes instead of merely categorising them.

K-means clustering serves as a data-driven alternative to a priori demographic groupings in consumer classification (Wedel and Kamakura, 2000). Meanwhile, logistic regression has been employed to segment consumers into loyalty categories based on demographic and transactional predictors (Neslin et al., 2006). Combining both approaches, this report uses clustering to identify consumer segments and logistic regression to predict loyalty classification.

## Research Question, Sub-questions, and Hypotheses

**Research Question:** To what extent do consumer demographic and behavioural characteristics predict purchasing loyalty, and can distinct consumer segments be identified based on spending patterns?
**Sub-question 1 (SQ1):** Is there a significant relationship between household income and total spending?
**Sub-question 2 (SQ2):** Can distinct consumer segments be identified through K-means clustering based on product category spending?
**Sub-question 3 (SQ3):** Can a logistic regression model accurately classify consumers as high or low spenders based on demographic variables?

*Table 1: Aims and hypotheses for this investigation*

| Aim | Null Hypothesis ($H_0$) | Alternative Hypothesis ($H_1$) |
|---|---|---|
| SQ1: Income - spending relationship | No significant correlation between income and total spending. | Significant positive correlation exists between income and total spending. |
| SQ2: Consumer segments | No meaningful clusters exist in consumer spending patterns. | Distinct consumer clusters with different spending patterns can be identified. |

| | | |
|---|---|---|
| SQ3: Loyalty classification | Demographic variables cannot predict high-spender classification above chance. | Demographic variables predict high-spender classification significantly above chance (AUC > 0.5). |

**Presentation of Data**

This study utilises the Customer Personality Analysis dataset (Romero-Hernandez, n.d.), comprising 2,240 customer records from a retail firm. After removing 24 missing income values and 4 extreme outliers (defined as incomes exceeding $200,000 as these likely represent data entry errors rather than genuine consumer profiles), 2,212 observations remain. The key variables and their data types are as follows: income ($) is continuous and unbounded; total spending ($, summed across six product categories) is continuous and bounded below by zero; age (years) is continuous; recency (days since last purchase) is discrete and bounded; number of children is discrete; and education and marital status are categorical. Total spending serves as the dependent variable and proxy for purchasing loyalty. For the logistic regression, a binary outcome variable (high spender = 1 if spending exceeds the median of $397, otherwise 0) is derived, making it a Boolean variable. Summary statistics are presented in Table 2.

*Table 2: Summary statistics and data types for key variables across 2,212 customers*

| Variable | Data Type | Mean | Median | Std Dev | Range |
|---|---|---|---|---|---|
| Income ($) | Continuous | 51,958.81 | 51,371 | 21,527.28 | 160,667 |
| Total Spending ($) | Continuous | 607.27 | 397 | 602.51 | 2,520 |
| Age (years) | Continuous | 55.09 | 54 | 11.70 | 56 |
| Recency (days) | Discrete | 49.02 | 49 | 28.94 | 99 |
| Total Children | Discrete | 0.95 | 1 | 0.75 | 3 |

Income and total spending display positive skew, where means exceed medians, likely determined by a subset of high-income, high-spending customers. Total spending demonstrates particularly high variability (standard deviation of $602.51 relative to a mean of $607.27), indicating substantial heterogeneity. Recency is roughly normally distributed, with mean and median both near 49 days. These distributions are visualised in Figure 1 and quantified in Table 3.

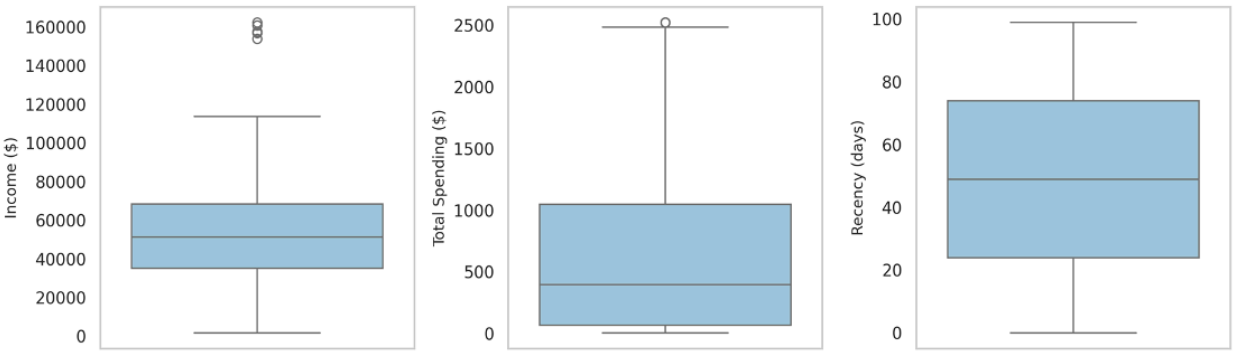



*Figure 1: Box plots displaying the distribution of Income, Total Spending, and Recency for 2,212 customers*

*Table 3: Box plot statistics for Income, Total Spending, and Recency across 2,212 customers*

| Box Plot Statistic | Income ($) | Total Spending ($) | Recency (days) |
|---|---|---|---|
| Minimum | 1,730 | 5 | 0 |
| Maximum | 162,397 | 2,525 | 99 |
| Lower Quartile (LQ) | 35,233.50 | 69 | 24 |
| Upper Quartile (UQ) | 68,487 | 1,048 | 74 |
| Interquartile Range (IQR) | 33,253.50 | 979 | 50 |
| Lo. Outlier Lim (LQ-1.5xIQR) | -14,646.75 | -1,399.50 | -51 |
| Hi. Outlier Lim (UQ+1.5xIQR) | 118,367.25 | 2,516.50 | 149 |

The box plot statistics reveal that income has several upper outliers beyond $118,367, while total spending has a narrow outlier window (upper limit of $2,516.50 is close to the maximum of $2,525), signaling that very high spenders are not statistical anomalies but rather a genuine subgroup. Recency does not demonstrate any outliers, with all values falling within the Tukey fences.

## Methodology

**Part 1:** A simple linear regression is conducted with Total Spending as the dependent variable and Income as the independent variable to assess the relationship between income and spending (SQ1), where higher income is associated with greater spending, addressing SQ1 directly.

**Part 2:** K-means clustering is applied to spending across six product categories to identify consumer segments (SQ2). Variables are standardised using z-scores to prevent scale dominance. Silhouette analysis has been utilised to determine the optimal number of clusters, comparing scores for k=2 through k=6. However, K-means assuming convex cluster shapes and being sensitive to initialisation, created a research problem. This is mitigated by using multiple random seeds (n_init=10). Moreover, K-means requires pre-specification of k, bringing in some subjectivity despite silhouette guidance.

**Part 3:** A logistic regression models the probability of being a high spender (spending above the median) as a function of income, age, recency, and number of children to predict consumer loyalty classification (SQ3). Unlike linear regression, logistic regression models a binary outcome using the logit function: $\log[P/(1-P)] = \beta_0 + \beta_1 x_1 + \ldots + \beta_4 x_4$, producing odds ratios that quantify how each predictor affects the odds of high spending (Hosmer et al., 2013). Model performance is evaluated using accuracy, a confusion matrix, and the AUC (Area Under the ROC Curve). The median-split threshold for defining high spenders being arbitrary brings about a limitation as alternative thresholds could yield different results.

All analyses are conducted using Python, with relevant code extracts provided in the Appendix.

## Results

**Part 1:**

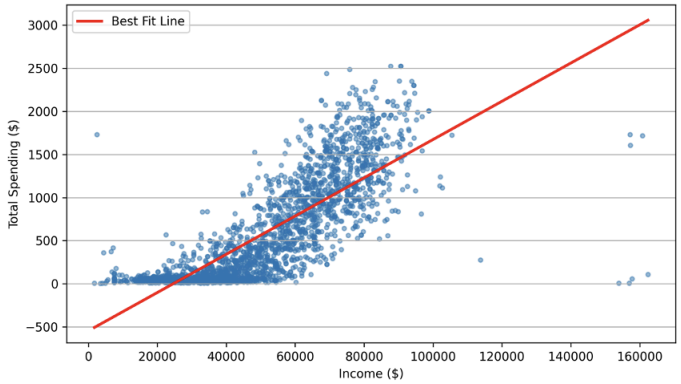



*Figure 2: Scatter plot of Income vs Total Spending with regression line, for 2,212 customers*

*Table 4: Simple linear regression statistics for Income and Total Spending*

| Regression Statistic | Value |
|---|---|
| Gradient (m) | 0.02219 |
| Intercept | -545.57 |
| R-squared ($R^2$) | 0.628 |
| Correlation coefficient (r) | 0.793 |
| P-value | < 0.001 |
| Regression Equation | $\hat{y} = 0.02219x - 545.57$ |

The scatter plot illustrates a strong positive linear correlation with moderate dispersion around the line of best fit. The R-squared value of 0.628 indicates that 62.8% of the variation in total spending is explained by income alone, while the remaining 37.2% is attributable to other factors. The correlation coefficient (r) of 0.793 confirms a strong positive association.

**Part 2:**




*Figure 3: Silhouette scores for different values of k (k=2 through k=6), K-means clustering on standardised spending variables*

Although k = 2 yields the highest silhouette score (0.514), k = 3 is selected for its superior interpretability, producing three meaningfully distinct segments rather than a simple binary split.

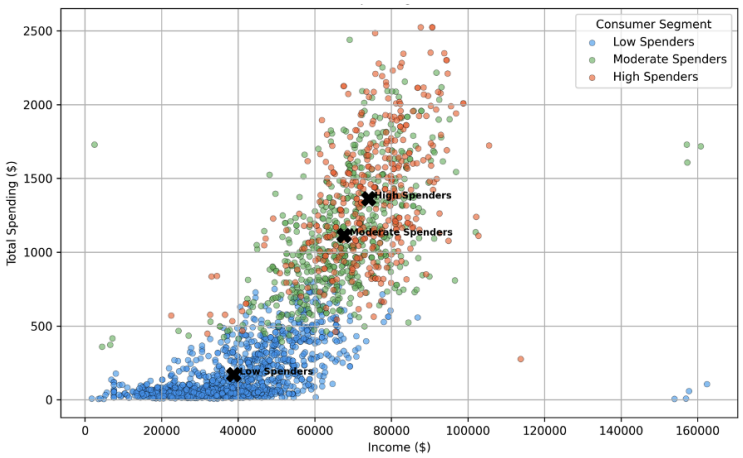



*Figure 4: K-means clustering of customers (k=3) based on spending patterns (Income vs Total Spending), for 2,212 customers. Silhouette score = 0.458.*

*Table 5: Centroid characteristics for each consumer segment*

| Consumer segment | Mean Spending ($) | Mean Income ($) | Mean Children | Count |
|---|---|---|---|---|
| High Spenders | 1,362.57 | 74,049.56 | 0.34 | 377 |
| Moderate Spenders | 1,114.15 | 67,602.10 | 0.67 | 548 |
| Low Spenders | 170.19 | 38,826.93 | 1.24 | 1,287 |

The clustering reveals that the majority of customers (1,287 out of 2,212) are Low Spenders, with a mean spending of just $170.19 and an average of 1.24 children. High Spenders, though the smallest group (377), have the highest mean income ($74,050) and notably fewer children (0.34 on average). This inverse relationship between children and spending is a key finding that is further examined in Part 3.

**Part 3:**

*Table 6: Logistic regression coefficients and odds ratios (standardised predictors)*

| Predictor | Coefficient (β) | Odds Ratio | Interpretation |
|---|---|---|---|
| Income | 3.200 | 24.52 | Strong positive effect on high-spending odds |
| Age | -0.036 | 0.96 | Negligible effect |
| Recency | -0.006 | 0.99 | Negligible effect |
| Total Children | -0.702 | 0.50 | Each child halves the odds of high spending |

*Table 7: Logistic regression classification performance (test set, n=443)*

| Model Performance Metric | Value |
|---|---|
| Accuracy | 90.7% |
| AUC (Area Under ROC Curve) | 0.961 |
| True Positives | 198 |
| True Negatives | 204 |
| False Positives | 21 |
| False Negatives | 20 |

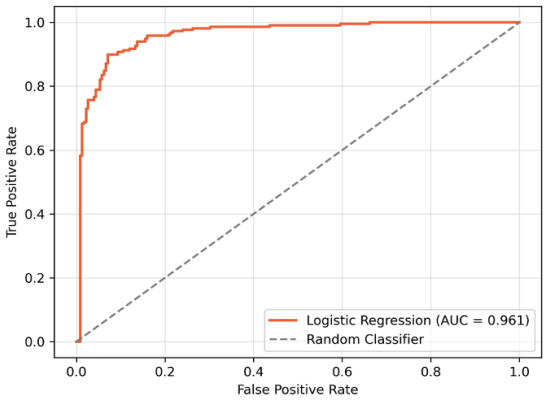



*Figure 5: ROC curve for the logistic regression model predicting high spender classification*

The logistic regression correctly classifies 90.7% of consumers in the test set. The confusion matrix shows that only 21 false positives and 20 false negatives occurred out of 443 test cases, demonstrating balanced performance across both classes.

## Discussion

**Part 1:** The regression reveals a strong positive relationship between income and total spending ($r = 0.793$, $p < 0.001$), with 62.8% of spending variation explained by income. The slope indicates that for each $1,000 increase in income, there is a $22.19 increase in spending. The null hypothesis for SQ1 is rejected. This aligns with consumer research linking economic resources to brand engagement (Gupta and Zeithaml, 2006). However, the unexplained 37.2% suggests that income alone is insufficient, motivating the multivariate approaches in Parts 2 and 3.

**Part 2:** Three distinct consumer segments were identified. High Spenders have substantially higher incomes ($74,050) and significantly fewer children (0.34) than Low Spenders ($38,827 income, 1.24 children). An independent samples t-test confirmed a statistically significant difference in spending between High and Low Spenders ($t = 78.17$, $p < 0.001$). The silhouette score of 0.458 indicates moderate cluster separation, with some overlap between Moderate and High Spenders, possibly reflecting transitional consumer profiles. The null hypothesis for SQ2 is also rejected. This supports Wedel and Kamakura's (2000) argument for data-driven segmentation over a priori groupings.

**Part 3:** The logistic regression achieved 90.7% accuracy and an AUC of 0.961, demonstrating excellent predictive power for classifying consumers as high or low spenders. Income is the dominant predictor (odds ratio = 24.52), while number of children halves the odds of high spending (odds ratio = 0.50). Age and recency show negligible effects, with odds ratios of 0.96 and 0.99 respectively, both very close to 1, indicating almost no change in the odds of being a high spender. The null hypothesis for SQ3 is also rejected. The high AUC indicates that demographic profiles can effectively predict purchasing loyalty, consistent with Neslin et al.'s (2006) findings on predictive consumer analytics.

**Limitations:** Lacking psychological variables (e.g., brand attachment, self-congruity), the dataset limits assessment of emotional drivers. The median-split threshold for defining high spenders is arbitrary. Omitted variable bias may affect estimates (Wilms et al., 2021). K-means assumes spherical clusters, potentially misclassifying boundary cases.

## Conclusion

This report has demonstrated that consumer income and household composition are significantly associated with purchasing loyalty, though these relationships are correlational rather than causal. Income alone accounts for 62.8% of spending variation, while a logistic regression model incorporating income and number of children classifies consumers with 90.7% accuracy (AUC = 0.961). Three distinct consumer segments were identified through K-means clustering, with High Spenders characterised by higher incomes and fewer children. These findings suggest that demographic profiles are meaningful predictors of brand spending behaviour; however, the direction of causality cannot be established from observational data alone, and unobserved confounders may explain part of these associations. Future research should employ experimental or longitudinal designs to confirm causal mechanisms, integrate attitudinal survey data alongside behavioural metrics, and explore alternative classification thresholds, supporting the broader transition from descriptive to predictive consumer analytics advocated in the literature.

**References and Appendix:**


Fader, P.S., Hardie, B.G.S. and Lee, K.L. (2005) 'RFM and CLV: Using iso-value curves for customer base analysis', *Journal of Marketing Research*, 42(4), pp. 415–430.

Fournier, S. (1998) 'Consumers and their brands: Developing relationship theory in consumer research', *Journal of Consumer Research*, 24(4), pp. 343–373.

Gupta, S. and Zeithaml, V. (2006) 'Customer metrics and their impact on financial performance', *Marketing Science*, 25(6), pp. 718–739.

Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) *Applied Logistic Regression*. 3rd edn. Hoboken: John Wiley & Sons.

Jacoby, J. and Chestnut, R.W. (1978) *Brand Loyalty: Measurement and Management.* New York: Wiley.

Neslin, S.A. et al. (2006) 'Defection detection: Measuring and understanding the predictive accuracy of customer churn models', *Journal of Marketing Research*, 43(2), pp. 204–211.

Oliver, R.L. (1999) 'Whence consumer loyalty?', *Journal of Marketing*, 63 (Special Issue), pp. 33–44.

Reichheld, F.F. and Sasser, W.E. (1990) 'Zero defections: Quality comes to services', *Harvard Business Review*, 68(5), pp. 105–111.

Romero-Hernandez, O. (n.d.) Customer Personality Analysis Dataset. Available at: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis (Accessed: 24 April 2026).

Uncles, M.D., Dowling, G.R. and Hammond, K. (2003) 'Customer loyalty and customer loyalty programs', *Journal of Consumer Marketing*, 20(4), pp. 294–316.

Wedel, M. and Kamakura, W.A. (2000) *Market Segmentation: Conceptual and Methodological Foundations.* 2nd edn. Boston: Kluwer Academic Publishers.

Wilms, R. et al. (2021) 'Omitted variable bias: A threat to estimating causal relationships', *Methods in Psychology*, 5, p. 100075.


I acknowledge that Google's Gemini 3 Flash was used for proofreading purposes, including grammar and syntax checking, as well as improving the readability and organisation of my code presented in this work. Available at: https://gemini.google.com/app (Accessed: 28 April 2026).

[2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv('marketing_campaign.csv', sep='\t')
print("Dataset shape:", df.shape)
print("\nFirst 5 rows:")
df.head()
```

```
Dataset shape: (2240, 29)

First 5 rows:
```

[2]:

| | ID | Year_Birth | Education | Marital_Status | Income | Kidhome | Teenhome | Dt_Customer | Recency | MntWines | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 5524 | 1957 | Graduation | Single | 58138.0 | 0 | 0 | 04-09-2012 | 58 | 635 | ... |
| **1** | 2174 | 1954 | Graduation | Single | 46344.0 | 1 | 1 | 08-03-2014 | 38 | 11 | ... |
| **2** | 4141 | 1965 | Graduation | Together | 71613.0 | 0 | 0 | 21-08-2013 | 26 | 426 | ... |
| **3** | 6182 | 1984 | Graduation | Together | 26646.0 | 1 | 0 | 10-02-2014 | 26 | 11 | ... |
| **4** | 5324 | 1981 | PhD | Married | 58293.0 | 1 | 0 | 19-01-2014 | 94 | 173 | ... |

5 rows × 29 columns

[3]:
```
df['Age'] = 2024 - df['Year_Birth']

df['TotalSpending'] = (df['MntWines'] + df['MntFruits'] +
                       df['MntMeatProducts'] + df['MntFishProducts'] +
                       df['MntSweetProducts'] + df['MntGoldProds'])

df['TotalChildren'] = df['Kidhome'] + df['Teenhome']

print("Missing values:")
print(df.isnull().sum()[df.isnull().sum() > 0])

df = df.dropna(subset=['Income'])

df = df[df['Income'] < 200000]
df = df[df['Age'] < 100]

print(f"\nCleaned dataset: {df.shape[0]} rows, {df.shape[1]} columns")
```

```
Missing values:
Income    24
dtype: int64

Cleaned dataset: 2212 rows, 32 columns
```

[4]:
```
summary_vars = ['Income', 'TotalSpending', 'Age', 'Recency', 'TotalChildren']

for var in summary_vars:
    print(f"\n{var}:")
    print(f"  Mean:    {df[var].mean():.2f}")
    print(f"  Median:  {df[var].median():.2f}")
    print(f"  Std Dev: {df[var].std():.2f}")
    print(f"  Range:   {df[var].max() - df[var].min():.2f}")
```

```
Income:
  Mean:    51958.81
  Median:  51371.00
  Std Dev: 21527.28
  Range:   160667.00

TotalSpending:
  Mean:    607.27
  Median:  397.00
  Std Dev: 602.51
  Range:   2520.00

Age:
  Mean:    55.09
  Median:  54.00
  Std Dev: 11.70
  Range:   56.00

Recency:
  Mean:    49.02
  Median:  49.00
  Std Dev: 28.94
  Range:   99.00

TotalChildren:
  Mean:    0.95
  Median:  1.00
  Std Dev: 0.75
  Range:   3.00
```

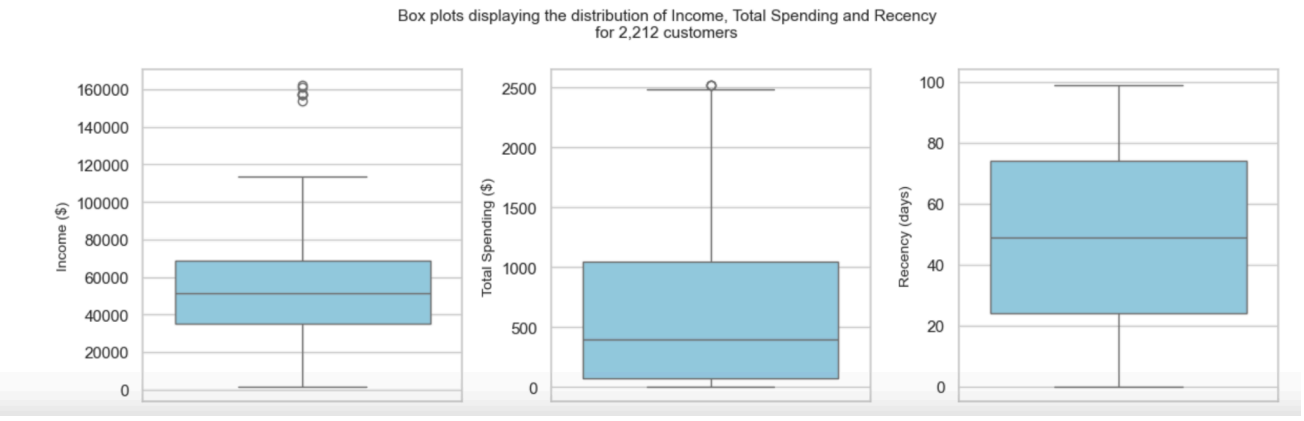
[5]:
```
sns.set(style='whitegrid')
fig, axes = plt.subplots(1, 3, figsize=(12, 4))

sns.boxplot(y=df['Income'], ax=axes[0], color='skyblue')
axes[0].set_ylabel('Income ($)', fontsize=10)

sns.boxplot(y=df['TotalSpending'], ax=axes[1], color='skyblue')
axes[1].set_ylabel('Total Spending ($)', fontsize=10)

sns.boxplot(y=df['Recency'], ax=axes[2], color='skyblue')
axes[2].set_ylabel('Recency (days)', fontsize=10)

plt.suptitle('Box plots displaying the distribution of Income, Total Spending and Recency\nfor 2,212 customers', fontsize=11)
plt.tight_layout()
plt.savefig('boxplots.png', dpi=300)
plt.show()
```



[17]:
```python
import statsmodels.api as sm
from scipy.stats import pearsonr

x = df['Income'].values
y = df['TotalSpending'].values

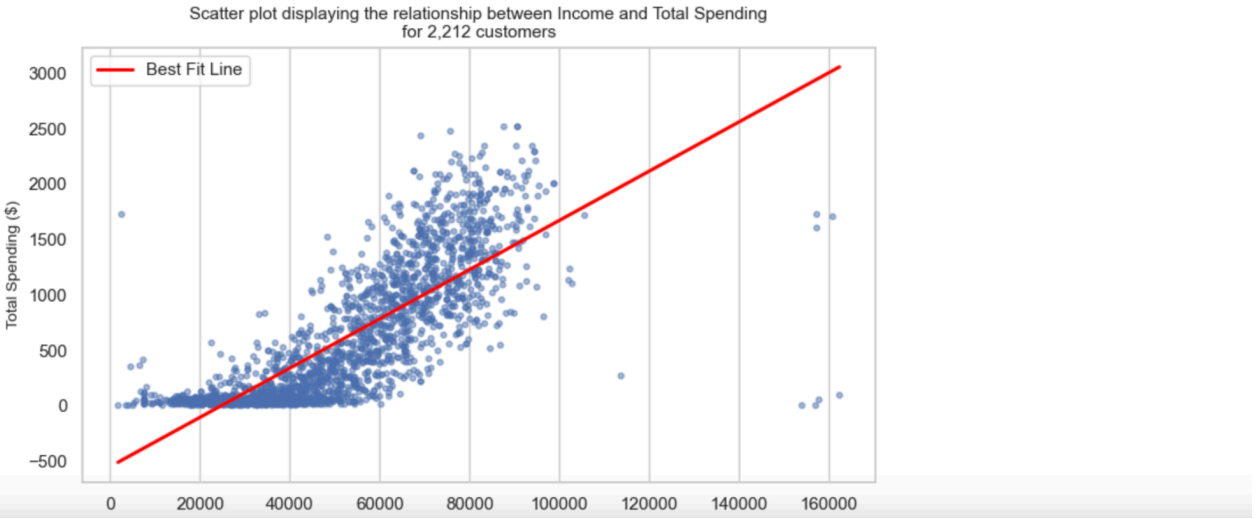
X = sm.add_constant(x)
model = sm.OLS(y, X).fit()

slope = model.params[1]
intercept = model.params[0]
r_squared = model.rsquared
correlation, _ = pearsonr(x, y)
p_value = model.pvalues[1]

print("=== Simple Linear Regression Results ===")
print(f"Slope:       {slope:.5f}")
print(f"Intercept:   {intercept:.2f}")
print(f"R-squared:   {r_squared:.4f}")
print(f"Correlation: {correlation:.4f}")
print(f"P-value:     {p_value:.2e}")
print(f"Equation:    y = {slope:.5f}x + ({intercept:.2f})")
```

```
=== Simple Linear Regression Results ===
Slope:       0.02219
Intercept:   -545.57
R-squared:   0.6284
Correlation: 0.7927
P-value:     0.00e+00
Equation:    y = 0.02219x + (-545.57)
```

[7]:
```python
fig, ax = plt.subplots(figsize=(8, 5))
sns.regplot(x='Income', y='TotalSpending', data=df, ci=None,
            scatter_kws={'s': 12, 'alpha': 0.5, 'label': 'Data Points'},
            line_kws={'color': 'red', 'label': 'Best Fit Line'})
plt.title('Scatter plot displaying the relationship between Income and Total Spending\nfor 2,212 customers', fontsize=11)
plt.xlabel('Income ($)', fontsize=10)
plt.ylabel('Total Spending ($)', fontsize=10)
plt.legend()
plt.grid(axis='y')
plt.tight_layout()
plt.savefig('scatter_regression.png', dpi=300)
plt.show()
```



[8]:
```
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import silhouette_score

spending_cols = ['MntWines', 'MntFruits', 'MntMeatProducts',
                 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds']

scaler = StandardScaler()
features_scaled = scaler.fit_transform(df[spending_cols])

print("=== Silhouette Analysis ===")
for k in range(2, 7):
    km = KMeans(n_clusters=k, random_state=42, n_init=10)
    labels = km.fit_predict(features_scaled)
    score = silhouette_score(features_scaled, labels)
    print(f"k = {k}: Silhouette Score = {score:.3f}")
```

```
=== Silhouette Analysis ===
k = 2: Silhouette Score = 0.514
k = 3: Silhouette Score = 0.458
k = 4: Silhouette Score = 0.469
k = 5: Silhouette Score = 0.473
k = 6: Silhouette Score = 0.445
```

[9]:
```
km_final = KMeans(n_clusters=3, random_state=42, n_init=10)
df['Cluster'] = km_final.fit_predict(features_scaled)

cluster_means = df.groupby('Cluster')['TotalSpending'].mean().sort_values()
cluster_names = {}
for i, (idx, val) in enumerate(cluster_means.items()):
    if i == 0:
        cluster_names[idx] = 'Low Spenders'
    elif i == 1:
        cluster_names[idx] = 'Moderate Spenders'
    else:
        cluster_names[idx] = 'High Spenders'

df['Segment'] = df['Cluster'].map(cluster_names)
```

[10]:
```
print("=== Cluster Centroids ===")
centroid_table = df.groupby('Segment')[['TotalSpending', 'Income', 'Age', 'TotalChildren']].mean().round(2)
print(centroid_table)

print(f"\nCluster sizes:")
print(df['Segment'].value_counts())

sil_score = silhouette_score(features_scaled, df['Cluster'])
print(f"\nSilhouette Score (k=3): {sil_score:.3f}")
```

```
=== Cluster Centroids ===
                  TotalSpending    Income    Age  TotalChildren
Segment
High Spenders           1362.57  74049.56  55.26           0.34
Low Spenders             170.19  38826.93  54.04           1.24
Moderate Spenders       1114.15  67602.10  57.43           0.67

Cluster sizes:
Segment
Low Spenders         1287
Moderate Spenders     548
High Spenders         377
Name: count, dtype: int64

Silhouette Score (k=3): 0.458
```

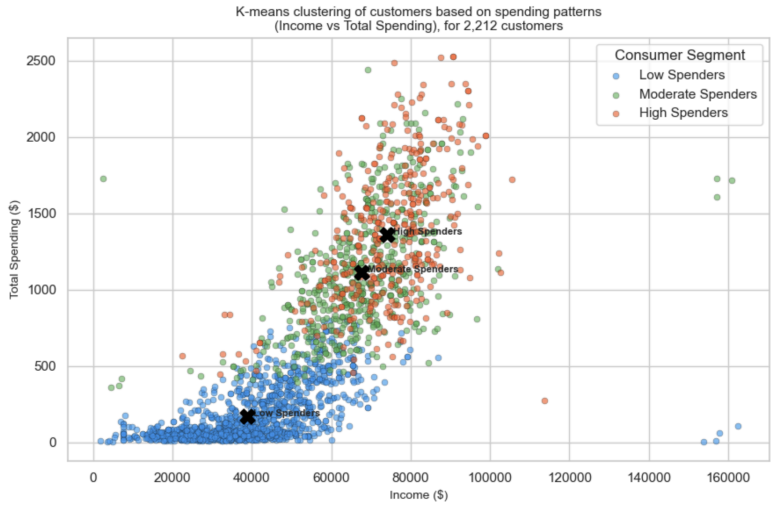
[11]:
```
fig, ax = plt.subplots(figsize=(9, 6))

colors = {'Low Spenders': '#2196F3', 'Moderate Spenders': '#4CAF50', 'High Spenders': '#FF5722'}

for seg, color in colors.items():
    mask = df['Segment'] == seg
    ax.scatter(df.loc[mask, 'Income'], df.loc[mask, 'TotalSpending'],
               c=color, label=seg, s=25, alpha=0.6, edgecolors='k', linewidth=0.3)

    cx = df.loc[mask, 'Income'].mean()
    cy = df.loc[mask, 'TotalSpending'].mean()
    ax.scatter(cx, cy, c='black', marker='X', s=150, zorder=5)
    ax.annotate(seg, (cx + 1500, cy), fontsize=8, fontweight='bold')

plt.title('K-means clustering of customers based on spending patterns\n(Income vs Total Spending), for 2,212 customers', fontsize=11)
plt.xlabel('Income ($)', fontsize=10)
plt.ylabel('Total Spending ($)', fontsize=10)
plt.legend(title='Consumer Segment')
plt.grid(True)
plt.tight_layout()
plt.savefig('cluster_plot.png', dpi=300)
plt.show()
```




[12]:
```
from scipy.stats import ttest_ind

high = df[df['Segment'] == 'High Spenders']['TotalSpending']
low = df[df['Segment'] == 'Low Spenders']['TotalSpending']

t_stat, t_pval = ttest_ind(high, low)
print("=== Independent Samples T-Test ===")
print(f"High Spenders mean: {high.mean():.2f}")
print(f"Low Spenders mean:  {low.mean():.2f}")
print(f"t-statistic: {t_stat:.2f}")
print(f"p-value:     {t_pval:.2e}")
```

```
=== Independent Samples T-Test ===
High Spenders mean: 1362.57
Low Spenders mean:  170.19
t-statistic: 78.17
p-value:     0.00e+00
```

[14]:
```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, roc_auc_score, roc_curve
from sklearn.model_selection import train_test_split


median_spending = df['TotalSpending'].median()
df['HighSpender'] = (df['TotalSpending'] > median_spending).astype(int)
print(f"Median spending threshold: ${median_spending}")
print(f"High Spenders: {df['HighSpender'].sum()}")
print(f"Not High Spenders: {(1 - df['HighSpender']).sum()}")

X = df[['Income', 'Age', 'Recency', 'TotalChildren']]
y = df['HighSpender']


X_scaled = StandardScaler().fit_transform(X)


X_train, X_test, y_train, y_test = train_test_split(
    X_scaled, y, test_size=0.2, random_state=42)

print(f"\nTraining set: {len(X_train)} rows")
print(f"Test set: {len(X_test)} rows")
```

```
Median spending threshold: $397.0
High Spenders: 1104
Not High Spenders: 1108

Training set: 1769 rows
Test set: 443 rows
```

[15]:
```
log_model = LogisticRegression(random_state=42, max_iter=1000)
log_model.fit(X_train, y_train)

y_pred = log_model.predict(X_test)
y_prob = log_model.predict_proba(X_test)[:, 1]


accuracy = (y_pred == y_test).mean()
auc = roc_auc_score(y_test, y_prob)
cm = confusion_matrix(y_test, y_pred)

print("=== Logistic Regression Results ===")
print(f"Accuracy: {accuracy:.3f} ({accuracy*100:.1f}%)")
print(f"AUC:      {auc:.3f}")
print(f"\nConfusion Matrix:")
print(cm)
print(f"\nTrue Negatives:  {cm[0,0]}")
print(f"False Positives: {cm[0,1]}")
print(f"False Negatives: {cm[1,0]}")
print(f"True Positives:  {cm[1,1]}")


print("\n=== Coefficients and Odds Ratios ===")
feature_names = ['Income', 'Age', 'Recency', 'TotalChildren']
for name, coef in zip(feature_names, log_model.coef_[0]):
    odds_ratio = np.exp(coef)
    print(f"{name}: coef = {coef:.3f}, odds ratio = {odds_ratio:.3f}")
```

```
=== Logistic Regression Results ===
Accuracy: 0.907 (90.7%)
AUC:      0.961

Confusion Matrix:
[[204  21]
 [ 20 198]]

True Negatives:  204
False Positives: 21
False Negatives: 20
True Positives:  198

=== Coefficients and Odds Ratios ===
Income: coef = 3.200, odds ratio = 24.524
Age: coef = -0.036, odds ratio = 0.964
Recency: coef = -0.006, odds ratio = 0.994
TotalChildren: coef = -0.702, odds ratio = 0.496
```

[16]:
```python
fpr, tpr, thresholds = roc_curve(y_test, y_prob)

fig, ax = plt.subplots(figsize=(6, 5))
ax.plot(fpr, tpr, color='#FF5722', linewidth=2,
        label=f'Logistic Regression (AUC = {auc:.3f})')
ax.plot([0, 1], [0, 1], color='gray', linestyle='--', label='Random Classifier')
ax.set_xlabel('False Positive Rate', fontsize=10)
ax.set_ylabel('True Positive Rate', fontsize=10)
ax.set_title('ROC Curve for Logistic Regression Model\npredicting High Spender classification', fontsize=11)
ax.legend(loc='lower right')
ax.grid(True, alpha=0.3)
plt.tight_layout()
plt.savefig('roc_curve.png', dpi=300)
plt.show()
```

