Customer Personality Analysis

Presented By

Timileyin Anthony

Table of Content

Introduction

A company called "Fresh & Fine" sells a variety of food and beverage products online. Fresh & Fine has been experiencing inconsistent sales and struggling to understand its customers' buying habits. The company spends a lot of money on marketing campaigns, but they don't always see a good return on investment. Fresh & Fine's management team wants to understand their customers better so they can offer the right products to the right people and improve their sales.

Fresh & Fine does not have a clear understanding of who its customers are and what they want. The company needs to analyze customer data to identify different types of customers and their preferences. By doing this, Fresh & Fine can create targeted marketing campaigns and offer products that match their customers' needs. This will help the company increase sales, reduce wasted marketing spend, and improve customer satisfaction.

Data Source

The data used in this project was sourced from the Kaggle website, a popular platform for data science and machine learning datasets. The dataset contains attributes of people and products.

Dataset Details

Source: Kaggle

Link to Dataset: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis

People

- ID: Customer's unique identifier
- Year_Birth: Customer's birth year
- Education: Customer's education level
- Marital_Status: Customer's marital status
- Income: Customer's yearly household income
- Kidhome: Number of children in the customer's household
- Teenhome: Number of teenagers in customer's household
- Dt_Customer: Date of customer's enrollment with the company
- Recency: Number of days since customer's last purchase

Products

- MntWines: Amount spent on wine
- MntFruits: Amount spent on fruits
- MntMeatProducts: Amount spent on meat
- MntFishProducts: Amount spent on fish
- MntSweetProducts: Amount spent on sweets
- MntGoldProds: Amount spent on gold

Data Preprocessing

Data Wrangling

- Cleaned the dataset by removing duplicates, handling missing values, and correcting any inconsistencies.

Label Encoding

- Converted categorical data (e.g., marital status, education level) into numerical values using Label Encoder to make it suitable for machine learning algorithms.

Standard Scaling

- Applied StandardScaler to normalize the data, so all features had a similar scale.

Principal Component Analysis (PCA)

- Performed PCA to reduce the number of features in the dataset.

K-means Clustering

- Implemented the K-means algorithm to group customers into different segments based on their characteristics and behaviors.
- Determined the optimal number of clusters by analyzing the inertia and using the elbow method.

Cluster Analysis

- Analyzed the characteristics of each customer segment to understand their unique preferences and behaviors.

Data Modeling

```
# Modelling the data

scale = StandardScaler()
scale_df = scale.fit_transform(dff)
dff = pd.DataFrame(data = scale_df, columns= dff.columns)
dff.head(4)
```

```
pca = PCA(n_components=3)
columns = ['col1','col2','col3']
pca_df = pca.fit_transform(dff)
pca_df = pd.DataFrame(data = pca_df, columns=columns)
```

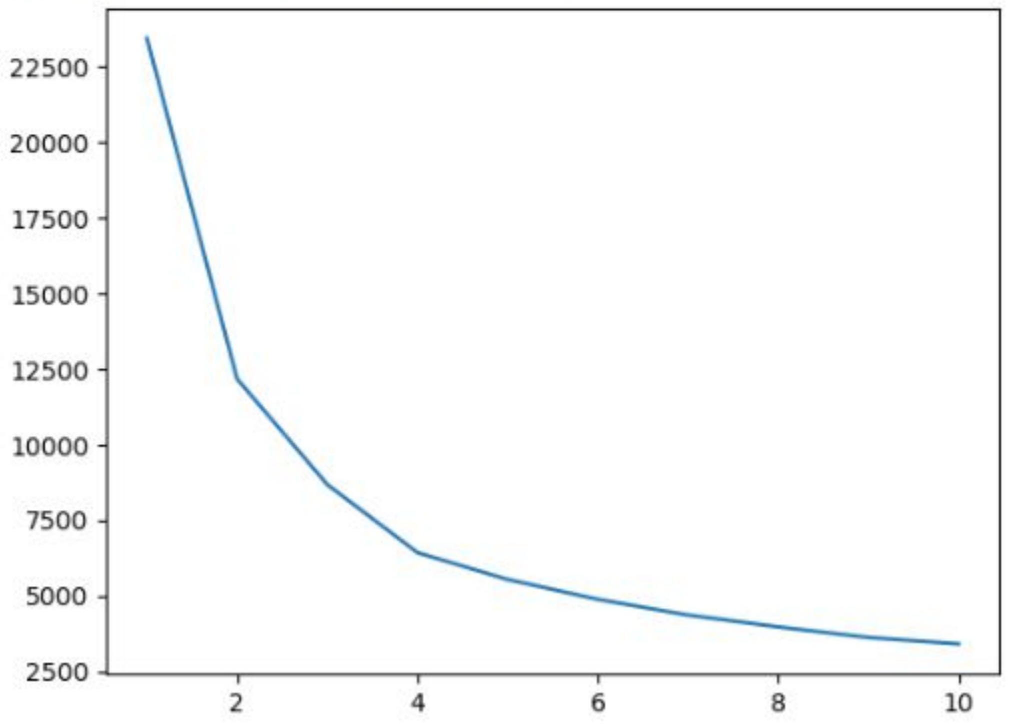
```
px.scatter_3d(data_frame=pca_df, x = 'col1', y = 'col2', z = 'col3', color='col1')
```

Eblow Graph

```
# Calculate WCSS for different number of clusters
# WCSS = Within-Cluster Sum of Square

WCSS_bank = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters=i, random_state=42)
    kmeans.fit(pca_df)
    WCSS_bank.append(kmeans.inertia_)
```

```
# chossing the number of clusters
# plot an elbow graph

plt.plot(range(1, 11), WCSS_bank)
```

```
[<matplotlib.lines.Line2D at 0x1e944cb80d0>]
```



Interpretation of the eblow graph and findings

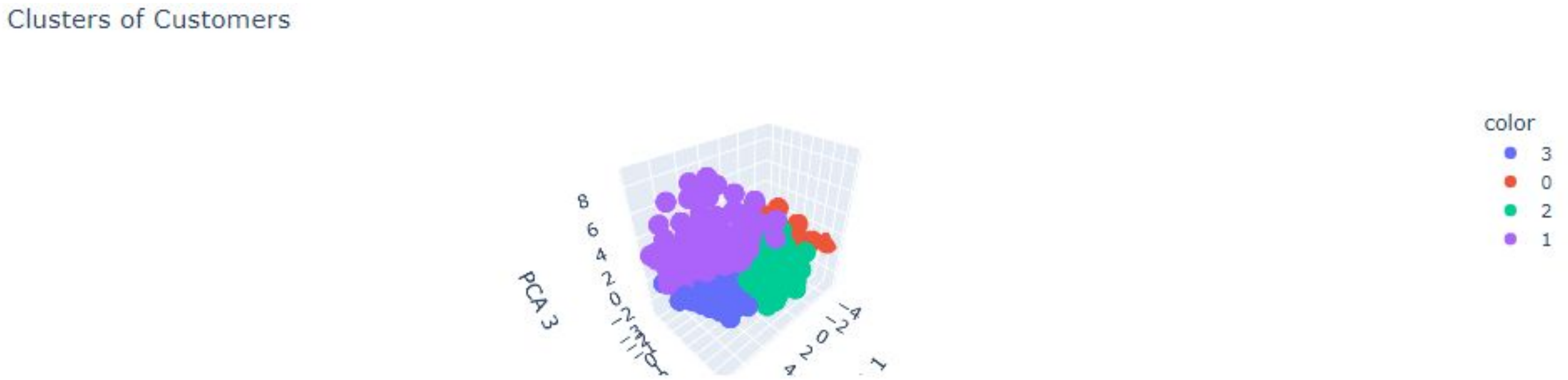
- The data points are best grouped into 4 distinct clusters.
- Four clusters provide a good balance between simplicity and the ability to capture the structure of the data.

Clusters Analysis Model

```
# Traning the KMeans Clustering Model

kmeans= KMeans(n_clusters=4, init='k-means++', random_state=42)
customer = kmeans.fit_predict(pca_df)
print(customer)
```

```
[3 0 3 ... 2 2 0]
```

```
# Add the cluster labels to the PCA DataFrame
pca_df['Cluster'] = customer

# Create scatter plot for each cluster using Plotly

px.scatter_3d(pca_df, x='col1', y='col2', z='col3', color=pca_df['Cluster'].astype(str),
              title='Clusters of Customers', labels={'col1': 'PCA 1', 'col2': 'PCA 2', 'col3': 'PCA 3'})
```



```python
# Create scatter plot for each cluster using plt

plt.figure(figsize=(10, 6))
plt.scatter(pca_df[pca_df['Cluster'] == 0]['col1'], pca_df[pca_df['Cluster'] == 0]['col2'], s=20, label='Cluster 1')
plt.scatter(pca_df[pca_df['Cluster'] == 1]['col1'], pca_df[pca_df['Cluster'] == 1]['col2'], s=20, label='Cluster 2')
plt.scatter(pca_df[pca_df['Cluster'] == 2]['col1'], pca_df[pca_df['Cluster'] == 2]['col2'], s=20, label='Cluster 3')
plt.scatter(pca_df[pca_df['Cluster'] == 3]['col1'], pca_df[pca_df['Cluster'] == 3]['col2'], s=20, label='Cluster 4')

# Plotting the centroids
centroids = kmeans.cluster_centers_
plt.scatter(centroids[:, 0], centroids[:, 1], s=100, c='black', marker='*', label='Centroids')

# Adding labels and title
plt.title('Clusters of Customers')
plt.legend()
```

```python
# Assign cluster labels

df['clusters'] = ['cl1' if x == 1 else 'cl2' if x == 2 else
                  'cl3' if x == 3 else 'cl4' for x in customer]
```

Clusters on Income by Education

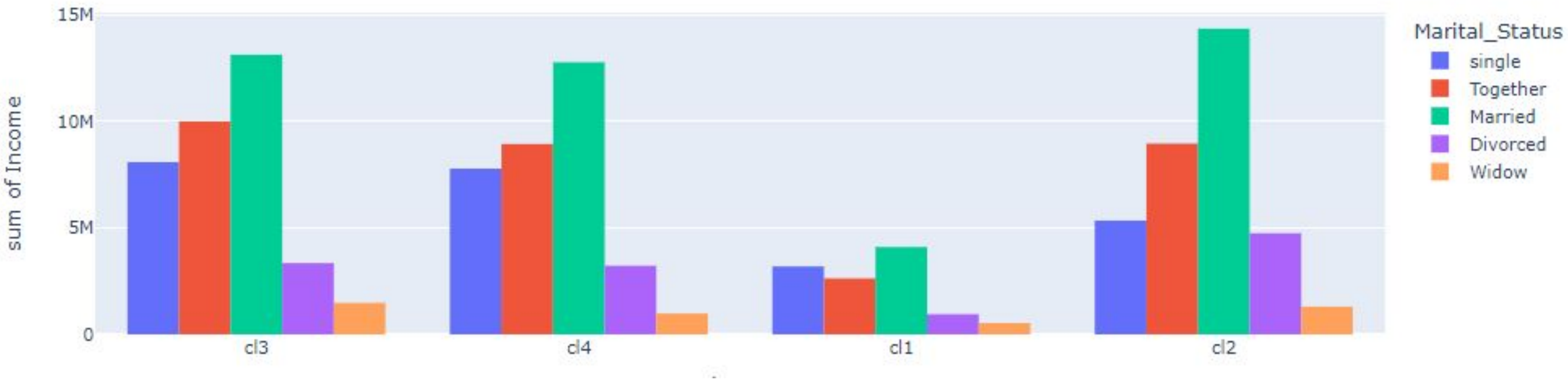
```
px.histogram(data_frame=df, x = 'clusters', y = 'Income',
             color = 'Education', barmode='group')
```

Insights

- Customers with a 'Graduation' education level consistently contribute the most income across all clusters.
- PhD and Master education levels also contribute, but significantly less than Graduation.
- Clusters cl3 and cl4 are the most profitable segments in terms of income.

Clusters on Income by Marital Status

```
px.histogram(data_frame=df, x = 'clusters', y = 'Income',
             color = 'Marital_Status', barmode='group')
```

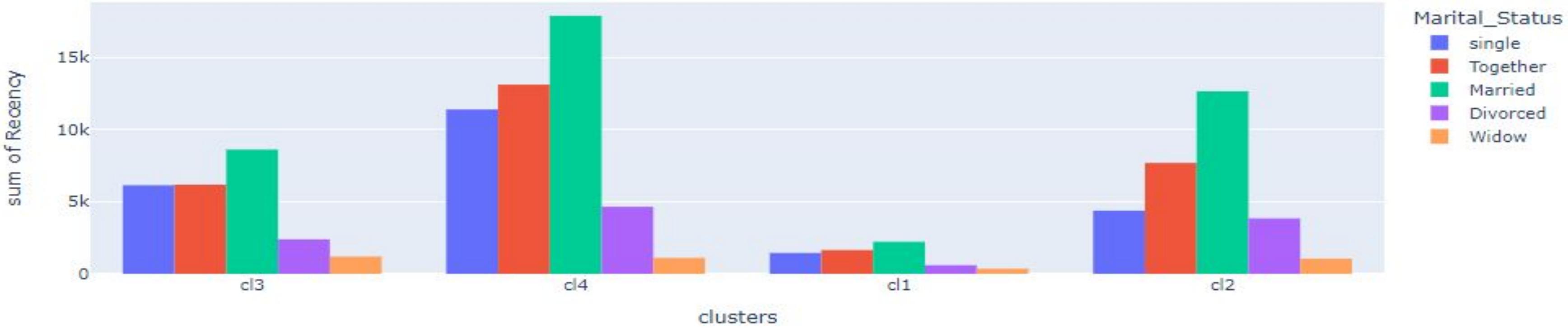


Insights

- 'Married' customers are the next significant contributors.
- 'Single' customers also contribute notably but less than 'Together' and 'Married' customers.
- 'Divorced' and 'Widow' customers generally contribute the least across all clusters.
- Cluster cl2 is the most profitable segment, primarily driven by 'Together' customers.

Clusters on Recency by Marital Status

```
px.histogram(data_frame=df, x = 'clusters', y = 'Recency',
             color = 'Marital_Status', barmode='group')
```



Insights

- 'Married' customers have the highest recency values in clusters cl4 and cl2, indicating that they are likely to have made recent interactions.
- 'Together' customers also show high recency values, especially in cluster cl4.
- 'Single' customers have moderate recency values across the clusters.
- 'Divorced' and 'Widow' customers generally have lower recency values, indicating less recent activity.
- Cluster cl4 shows the highest overall recency, driven mainly by 'Married' customers.
- Cluster cl3 shows the lowest overall recency.

Clusters on Mntwines by Marital Status

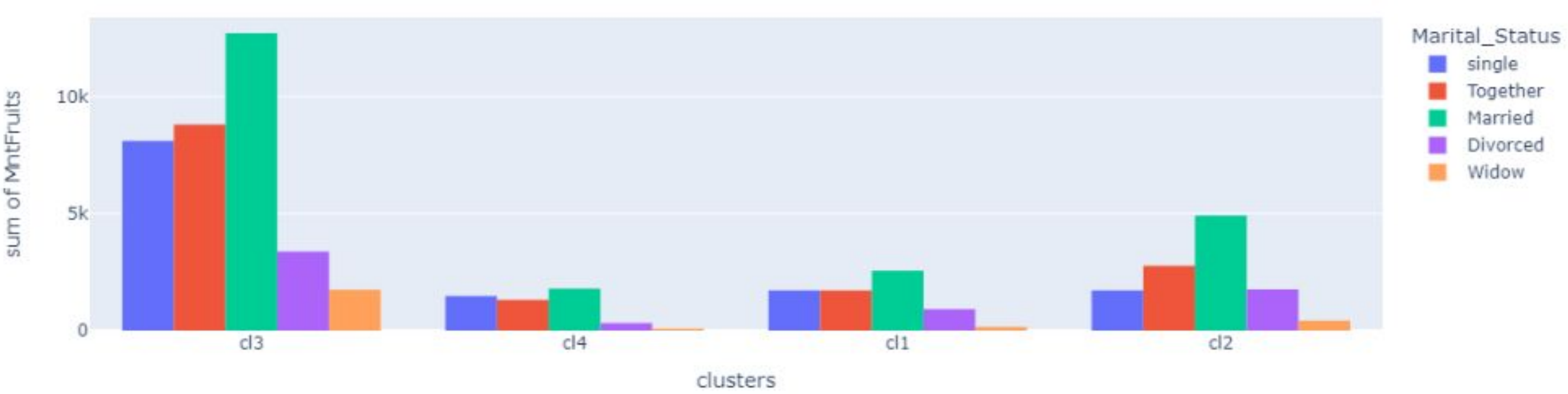
```
px.histogram(data_frame=df, x = 'clusters', y = 'MntWines',
             color = 'Marital_Status', barmode='group')
```

Insights

- Across all clusters, married customers are the highest spenders on wines.
- 'Together' customers contribute significantly to wine spending, second only to married customers.
- Single customers contribute a moderate amount to wine spending, less than married and together customers but still noteworthy.
- Divorced and Widow Customers, These groups contribute the least to wine spending across all clusters.

Clusters on MntFruits by Marital Status

```
px.histogram(data_frame=df, x = 'clusters', y = 'MntFruits',
             color = 'Marital_Status', barmode='group')
```



Insights

- 'Married' customers are the ones who spend the most on fruits.
- 'Together' customers also spend significantly on fruits, second only to married customers.
- 'Single' customers contribute a moderate amount to fruit spending, similar to their wine spending patterns.
- 'Divorced' and Widow Customers, These groups spend the least on fruits across all clusters.

Clusters on Recency by Education

```
px.histogram(data_frame=df, x = 'clusters', y = 'Recency',
             color = 'Education', barmode='group')
```

Insights

- Graduation holders consistently show high recency values across all clusters
- PhD holders also show high recency values
- Master holders show moderate recency values
- 2n Cycle and Basic education levels generally show lower recency values

Clusters on Mntwines by Education

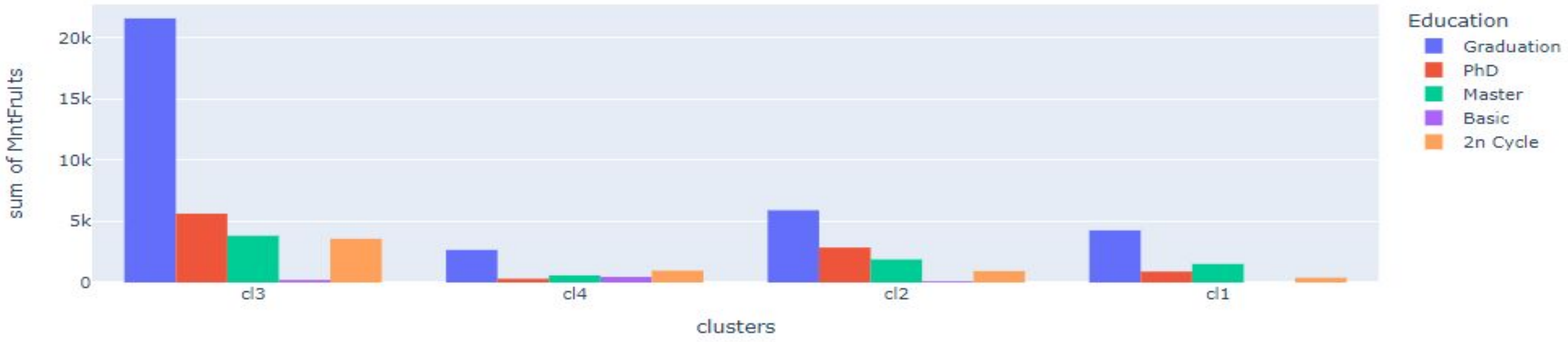
```
px.histogram(data_frame=df, x = 'clusters', y = 'MntWines',
             color = 'Education', barmode='group')
```

Insights

- Customers with a graduation level of education are the highest spenders on wines across clusters, particularly in clusters cl1 and cl2.
- PhD customers are significant wine spenders, especially in clusters cl1 and cl2.
- Customers with a Master’s degree also spend significantly on wines, especially in clusters cl1 and cl2.
- Basic and 2n Cycle These education levels have lower wine spending across all clusters.

Clusters on MntFruits by Education

```
px.histogram(data_frame=df, x = 'clusters', y = 'MntFruits',
             color = 'Education', barmode='group')
```



Insights

- Customers with a graduation-level education exhibit the highest spending on fruits, particularly noticeable in cluster cl1.
- Customers with PhD-level education show moderate spending on fruits across various clusters.
- Similar to the PhD level, customers with a Master’s level education also show moderate spending on fruits.
- Customers with 2n Cycle education show lower spending on fruits compared to higher education levels.
- Basic education level customers show the lowest spending on fruits across all clusters.

Recommendations

- Focus marketing efforts on segments with higher spending potential, such as married customers and those with higher education levels. Develop targeted campaigns highlighting premium and exclusive product offerings.
- Tailor product offerings to meet the preferences of high-spending segments, particularly for wines and fruits. Highlight the quality, exclusivity, and health benefits of these products in marketing campaigns.
- Leverage seasonal and festive occasions to offer targeted promotions, especially for high-value segments.
- Implement loyalty programs that reward frequent purchases and high spending.

THANK YOU