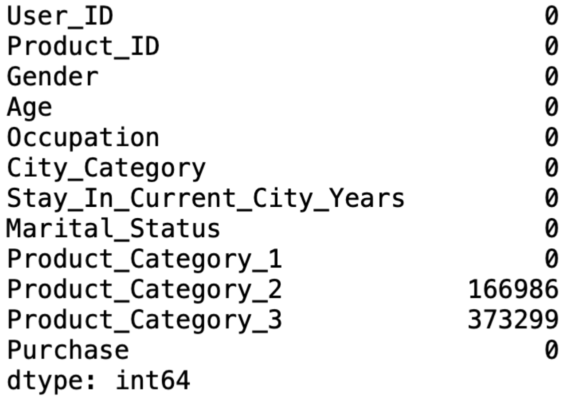
The goal of this work is to analyze the data from Black Friday sales and find insights that can be applied to increase earnings. Through early exploration it was revealed that the product category information could be used to determine most popular product categories to use for advertising, and the demographic data could be used to find the highest paying consumer segments. Together, this information can be used to advertise specific product categories to who would be their most likely consumers.

Data Clean-up

First, it was necessary to check for null values.



```
User_ID                            0
Product_ID                         0
Gender                             0
Age                                0
Occupation                         0
City_Category                      0
Stay_In_Current_City_Years         0
Marital_Status                     0
Product_Category_1                 0
Product_Category_2            166986
Product_Category_3            373299
Purchase                           0
dtype: int64
```

Product Category Null Values

There are only null values for columns “Product_Category_1” and “Product_Category_2”. This suggests that those values aren’t missing, but rather indicate that the system was designed so that there will not always be a value included for those features because these features represent a level of specificity of the product categories. Every entry belongs to at least 1 category, “Product_Category_1,” and therefore, any item with a value for “Product_Category_2” also has a value for “Product_Category_1.” Further, every item with a value for “Product_Category_3” has a value for “Prouct_Category_2”.

This is proven when this code is run:
has_category_3_without_category_2 = any(df['Product_Category_3'].notnull() & df['Product_Category_2'].isnull())

Which checks if there are any items with Product_Category_3 values but without Product_Category_2 values.

All of this is to say that this information represents categories, subcategories, and sub-subcategories (for example, Apparel > Shoes > Sandals), rather than missing data, and to make use of this information, the data will be reformatted—though this will be addressed later.

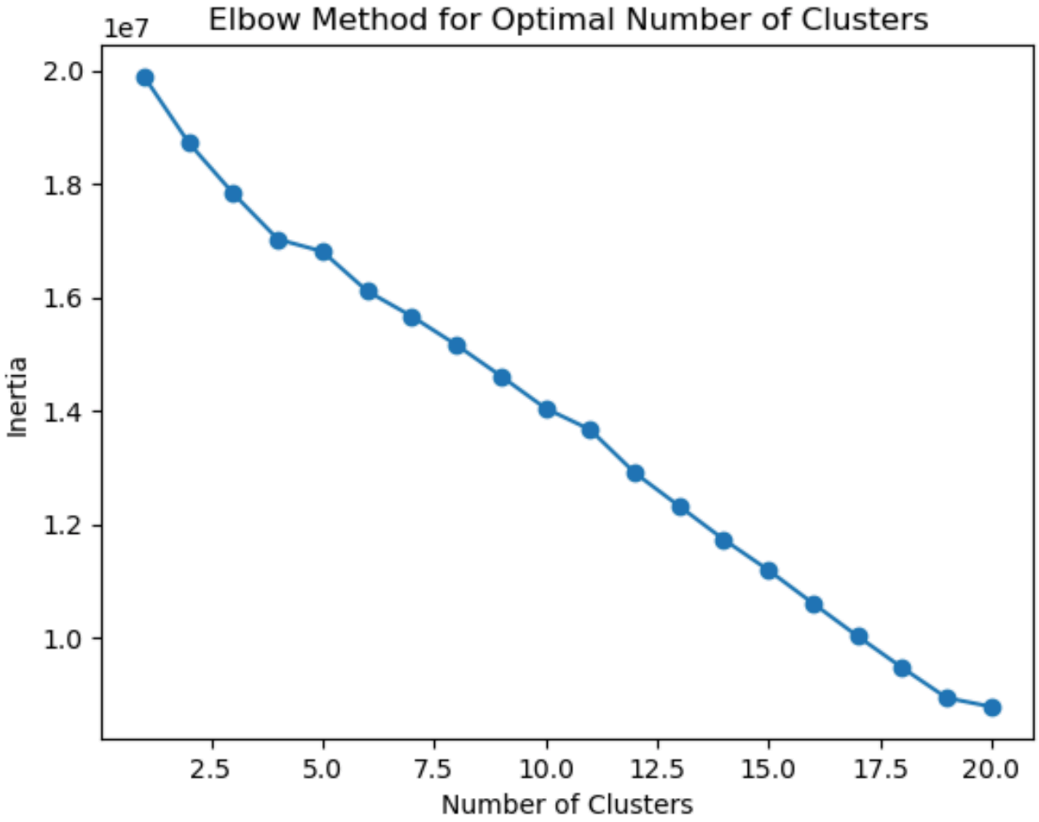
Categorical features

Next, the categorical columns needed to be one-hot encoded, to be in the correct format for getting clusters. This included Gender, Age, Occupation, City_Category, and Marital_Status. By replacing '4+' with '4' and converting the data type to integer, Stay_In_Current_City_Years was changed from categorical to numerical so that the average could be calculated.

Getting clusters

The clusters needed to be predictive of purchase amount. In order to do this, it was necessary to aggregate the rows by User_ID and Purchase, so that there would not be duplicate rows per user, which would affect the results of the clustering. Product categories and product number would be collinear with purchase so they were removed from the dataframe that would be used for clustering, this way, only demographic information would be used for predicting purchase amount.

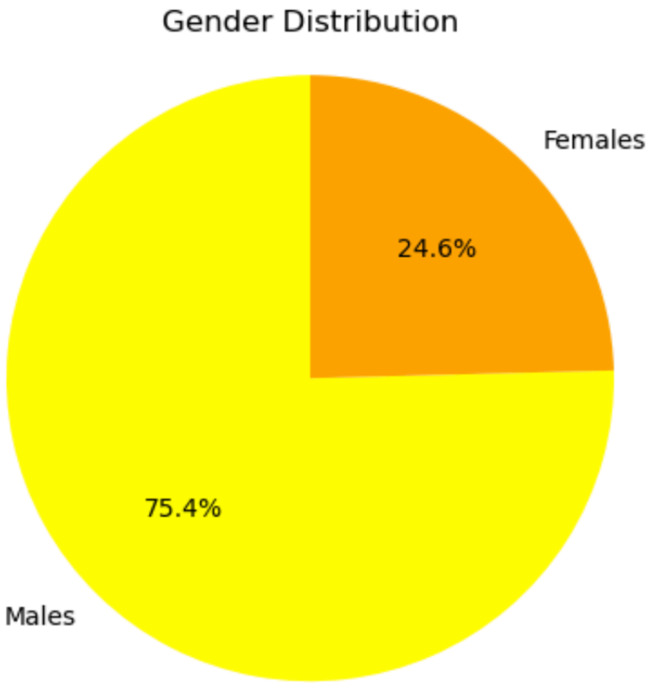
The elbow method was used to choose the optimal number of clusters, which it was decided is 18.

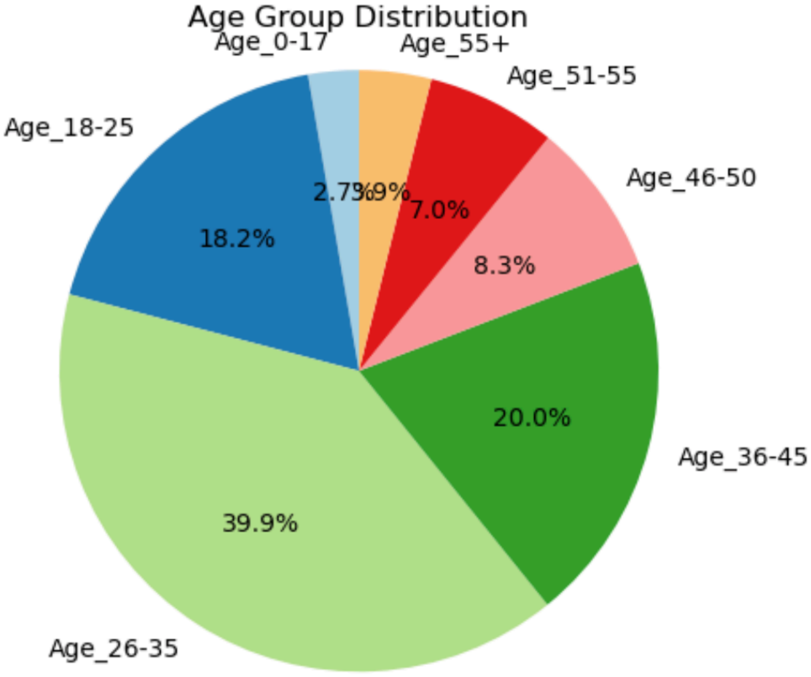


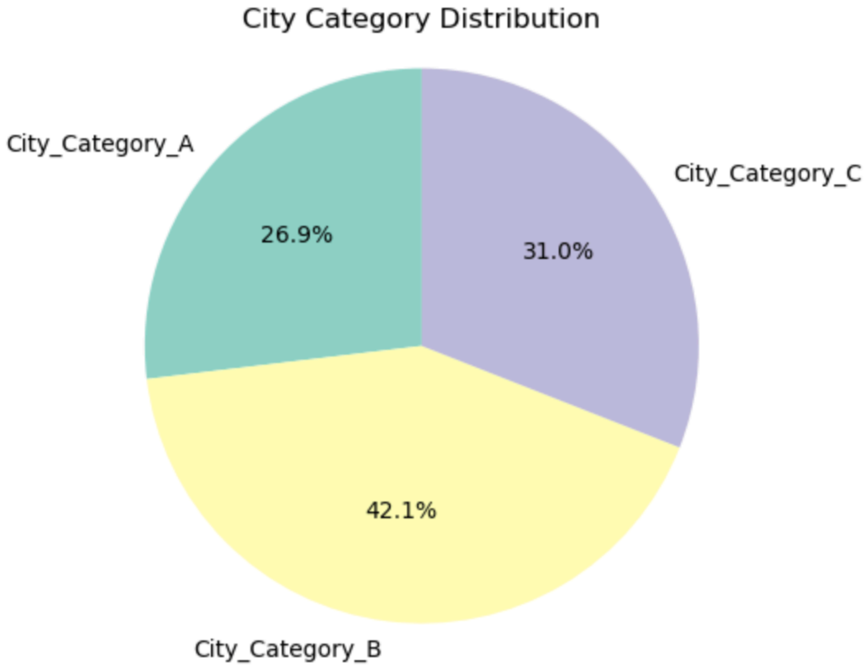
Demographics of Entire Dataset

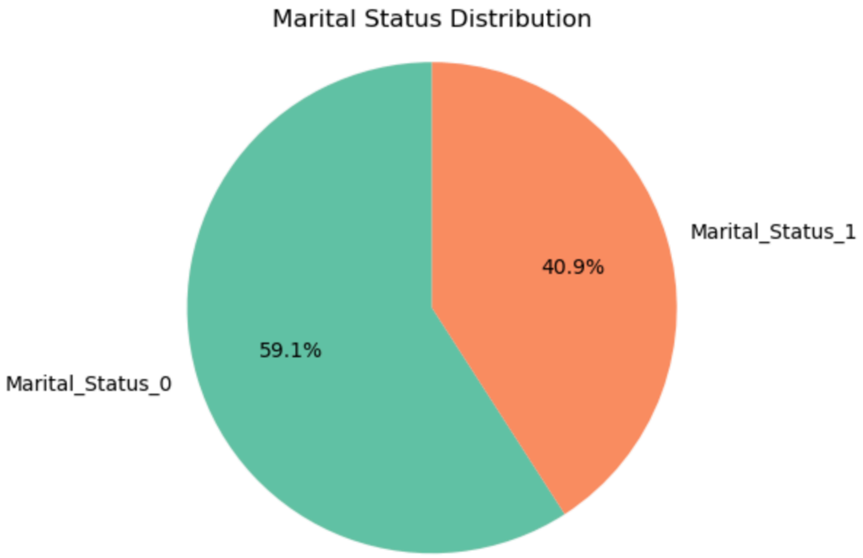
Before analyzing the demographics of the clusters extracted from the dataset, it was necessary to examine the general demographics of the dataset.

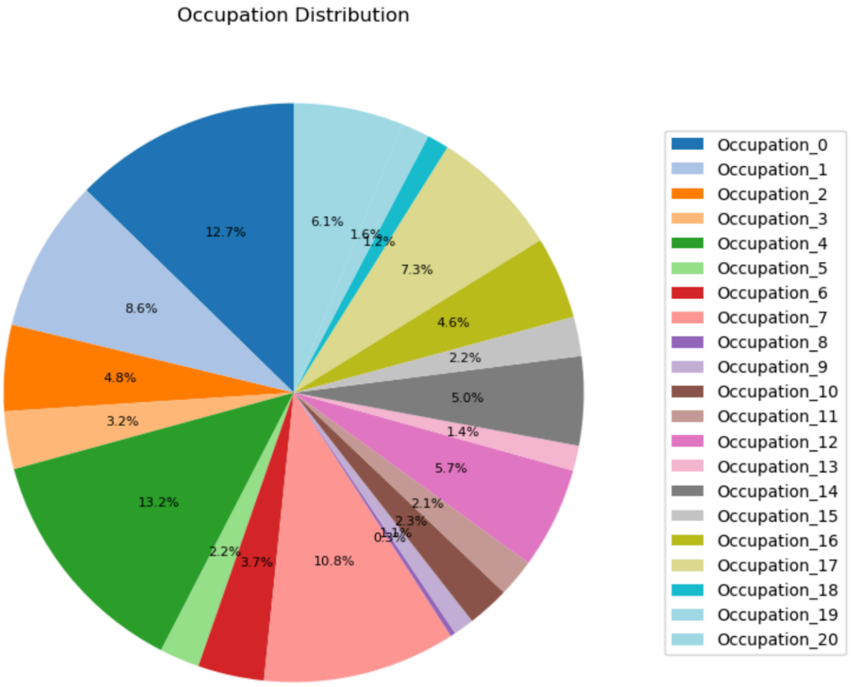
We can see that most of the customers are male, ages 18 - 35.


Gender Distribution

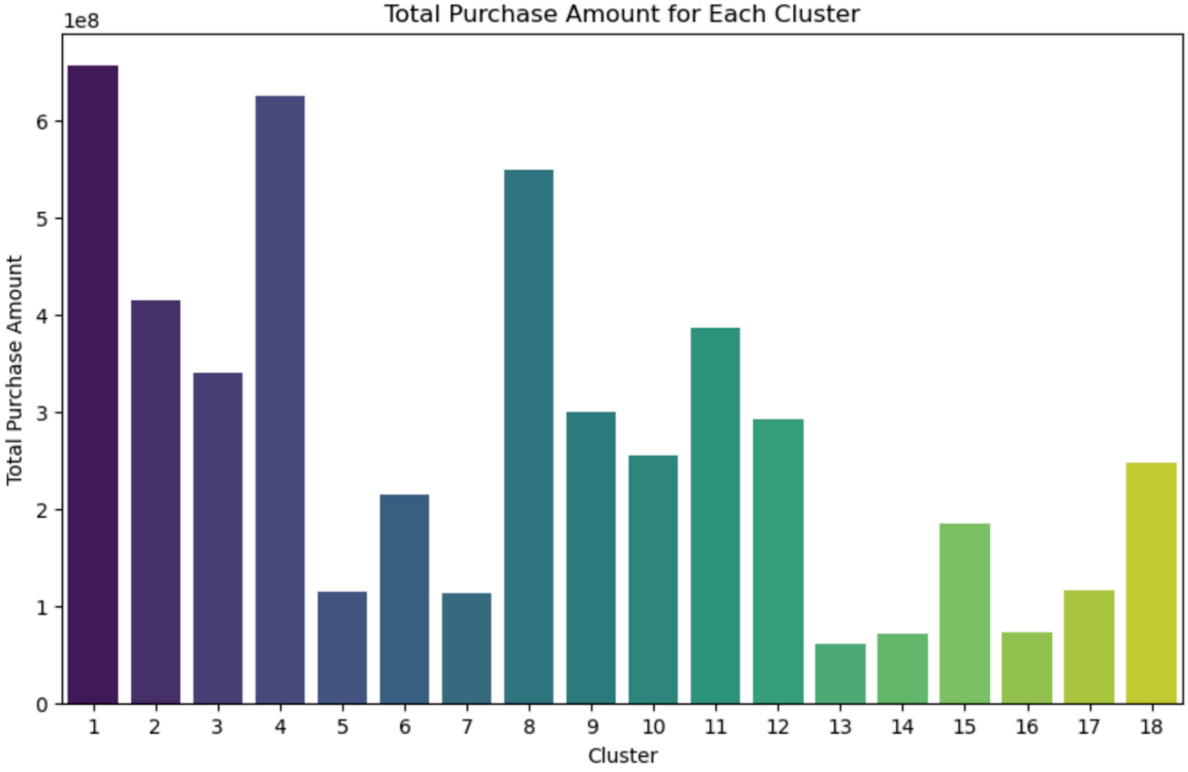

Age Group Distribution


City Category Distribution
City_Category_A
26.9%
City_Category_C
31.0%
42.1%
City_Category_B


Marital Status Distribution
Marital_Status_1
40.9%
59.1%
Marital_Status_0

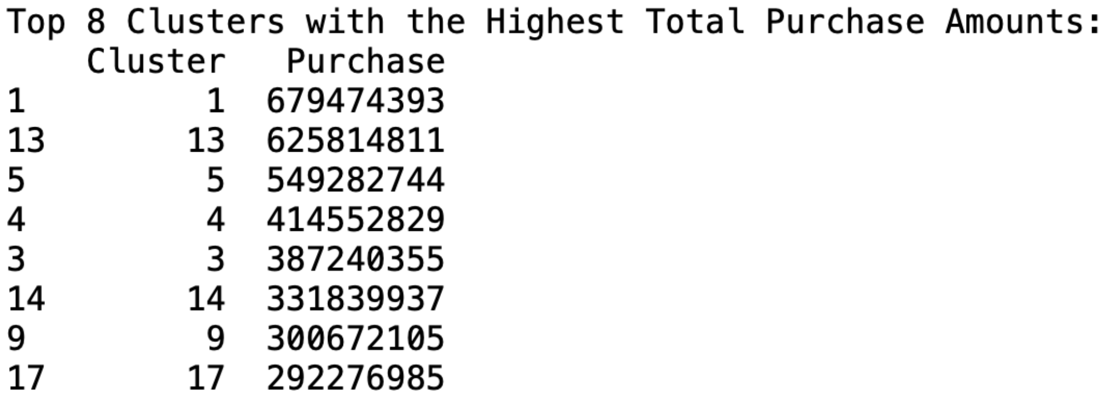


Before examining the demographic characteristics of the clusters, it is important to determine which clusters are of interest, by finding the highest-purchasing segments.



We can see that there's a wide range of purchase amounts per cluster, so there are some that would likely be much more effective for targeting with advertisements. In order to do so, the business should use the demographic information for each of the highest spending clusters and target them with advertisements for the products that they have shown to be interested in through their purchasing behavior captured in this dataset.

The top 8 clusters are who should be targeted with advertisements.



```
Top 8 Clusters with the Highest Total Purchase Amounts:
    Cluster   Purchase
1         1  679474393
13       13  625814811
5         5  549282744
4         4  414552829
3         3  387240355
14       14  331839937
9         9  300672105
17       17  292276985
```

Below are the demographics for each of the top 8 clusters, and a major distinguishing factor is occupation (bolded for each segment below). For each of the 8 highest-paying segments, the segment number is listed with the average years in their current city, and most occurring gender, age bracket, occupation, city category, and marital status with the corresponding percentages of occurrence for each.

Segment Number: 1
Average Stay_In_Current_City_Years: Avg=1.89
Gender with higher average (Gender_M): Avg=0.70
Average of the one with the highest value for Age (Age_18-25): Avg=0.71
Average of the one with the highest value for Occupation (Occupation_4): Avg=0.96
The one with the highest value for City_Category (City_Category_C): Avg=0.46
Average of the one with the highest value for Marital_Status (Marital_Status_0): Avg=0.74

Segment Number: 13
Average Stay_In_Current_City_Years: Avg=1.85
Gender with higher average (Gender_M): Avg=0.67
Average of the one with the highest value for Age (Age_26-35): Avg=0.42
Average of the one with the highest value for Occupation (Occupation_0): Avg=1.00
The one with the highest value for City_Category (City_Category_C): Avg=0.52
Average of the one with the highest value for Marital_Status (Marital_Status_0): Avg=0.58

Segment Number: 5
Average Stay_In_Current_City_Years: Avg=1.87
Gender with higher average (Gender_M): Avg=0.80
Average of the one with the highest value for Age (Age_26-35): Avg=0.38

Average of the one with the highest value for Occupation (Occupation_7): Avg=1.00
The one with the highest value for City_Category (City_Category_C): Avg=0.58
Average of the one with the highest value for Marital_Status (Marital_Status_0): Avg=0.53

Segment Number: 4
Average Stay_In_Current_City_Years: Avg=1.87
Gender with higher average (Gender_M): Avg=0.61
Average of the one with the highest value for Age (Age_26-35): Avg=0.29
Average of the one with the highest value for Occupation (Occupation_1): Avg=1.00
The one with the highest value for City_Category (City_Category_C): Avg=0.55
Average of the one with the highest value for Marital_Status (Marital_Status_0): Avg=0.54

Segment Number: 3
Average Stay_In_Current_City_Years: Avg=1.92
Gender with higher average (Gender_M): Avg=0.90
Average of the one with the highest value for Age (Age_26-35): Avg=0.40
Average of the one with the highest value for Occupation (Occupation_17): Avg=1.00
The one with the highest value for City_Category (City_Category_C): Avg=0.60
Average of the one with the highest value for Marital_Status (Marital_Status_0): Avg=0.59

Segment Number: 14
Average Stay_In_Current_City_Years: Avg=1.71
Gender with higher average (Gender_M): Avg=0.58
Average of the one with the highest value for Age (Age_26-35): Avg=0.43
Average of the one with the highest value for Occupation (Occupation_3): Avg=0.52
The one with the highest value for City_Category (City_Category_C): Avg=0.52
Average of the one with the highest value for Marital_Status (Marital_Status_0): Avg=0.57

Segment Number: 9
Average Stay_In_Current_City_Years: Avg=1.86
Gender with higher average (Gender_M): Avg=0.88
Average of the one with the highest value for Age (Age_26-35): Avg=0.47
Average of the one with the highest value for Occupation (Occupation_12): Avg=1.00
The one with the highest value for City_Category (City_Category_C): Avg=0.49
Average of the one with the highest value for Marital_Status (Marital_Status_0): Avg=0.58

Segment Number: 17
Average Stay_In_Current_City_Years: Avg=1.85
Gender with higher average (Gender_M): Avg=0.72
Average of the one with the highest value for Age (Age_26-35): Avg=0.39
Average of the one with the highest value for Occupation (Occupation_20): Avg=1.00
The one with the highest value for City_Category (City_Category_C): Avg=0.40
Average of the one with the highest value for Marital_Status (Marital_Status_0): Avg=0.53

For each of these clusters, many of the demographic details are consistent. For example, the most popular City_Category for the highest paying clusters is B, most are men, there are not significantly more married customers than unmarried, and the most popular age segment is 26-35. This is consistent with the demographics of the dataset in general. Using the occupation will be how the business targets the prospective customers most acutely.

Product category popularity by segment

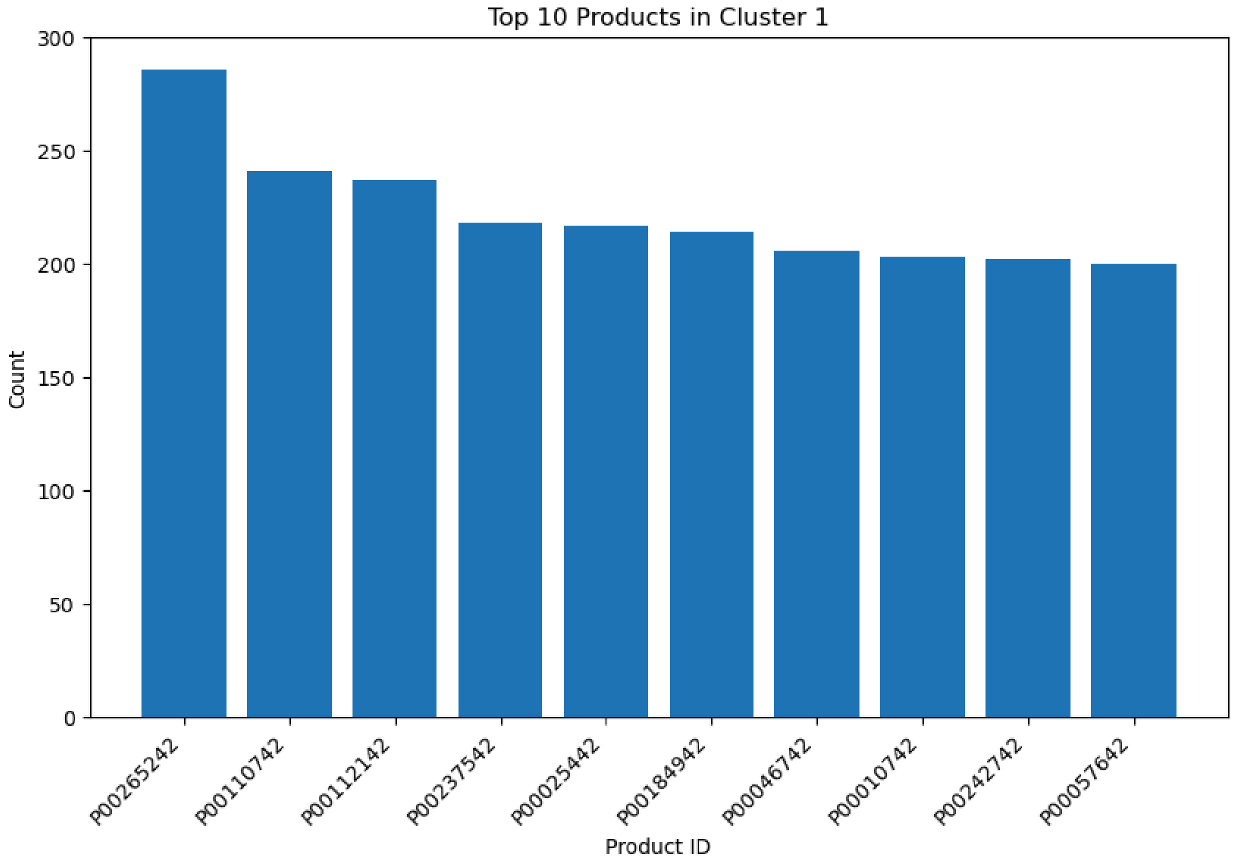
The next step was to see which products should be advertised to these top segments.

Product category data transformation

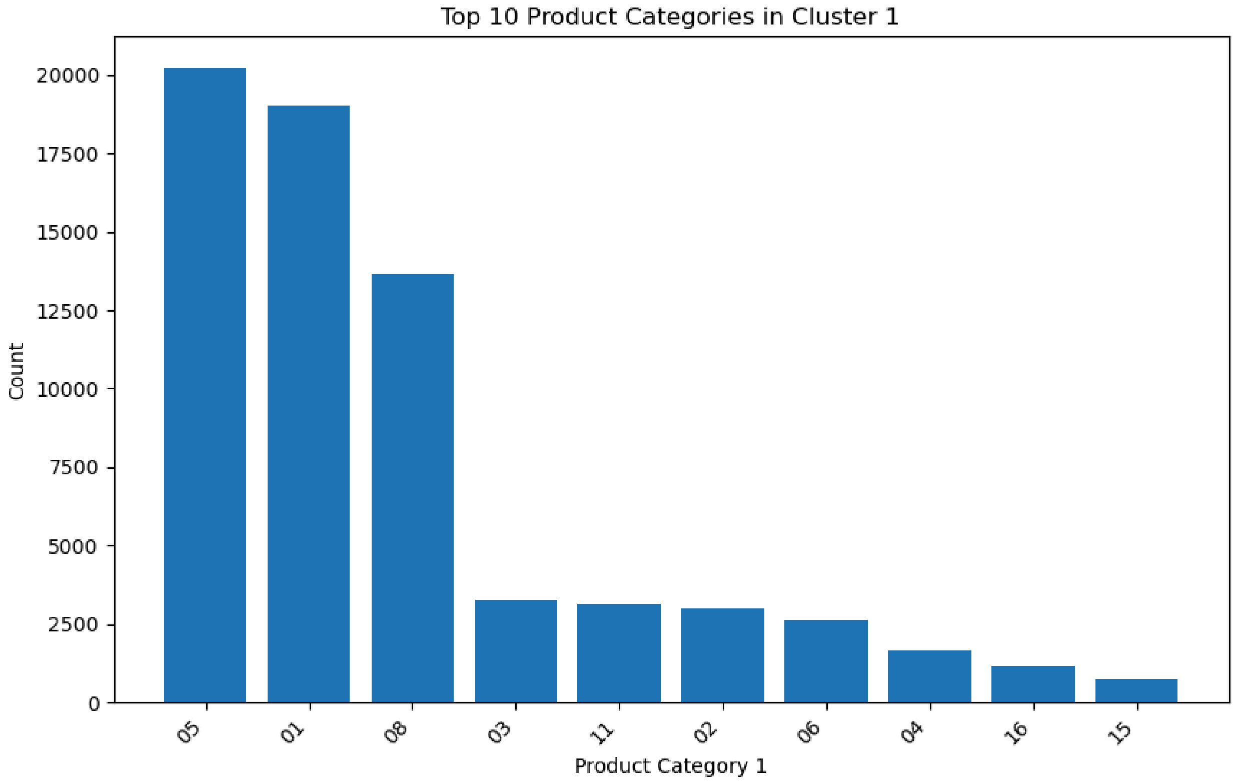
To explore the product categories, data cleanup needed to be done for the product category columns. To use the numbers as labels, they were converted to strings and single digit numbers were converted to 2 by prepending them with a 0, and the null values were converted to “00”. A new column was created to represent subcategories by combining the Product_Category_1 column with the Product_Category_2 column, and a new column was created to represent sub-subcategories by combining Product_Category_1, Product_Category_2, and Product_Category_3. With these new columns and with the clusters, it was now possible to examine the most popular product subcategories and sub-subcategories for each cluster.

Purchasing behavior

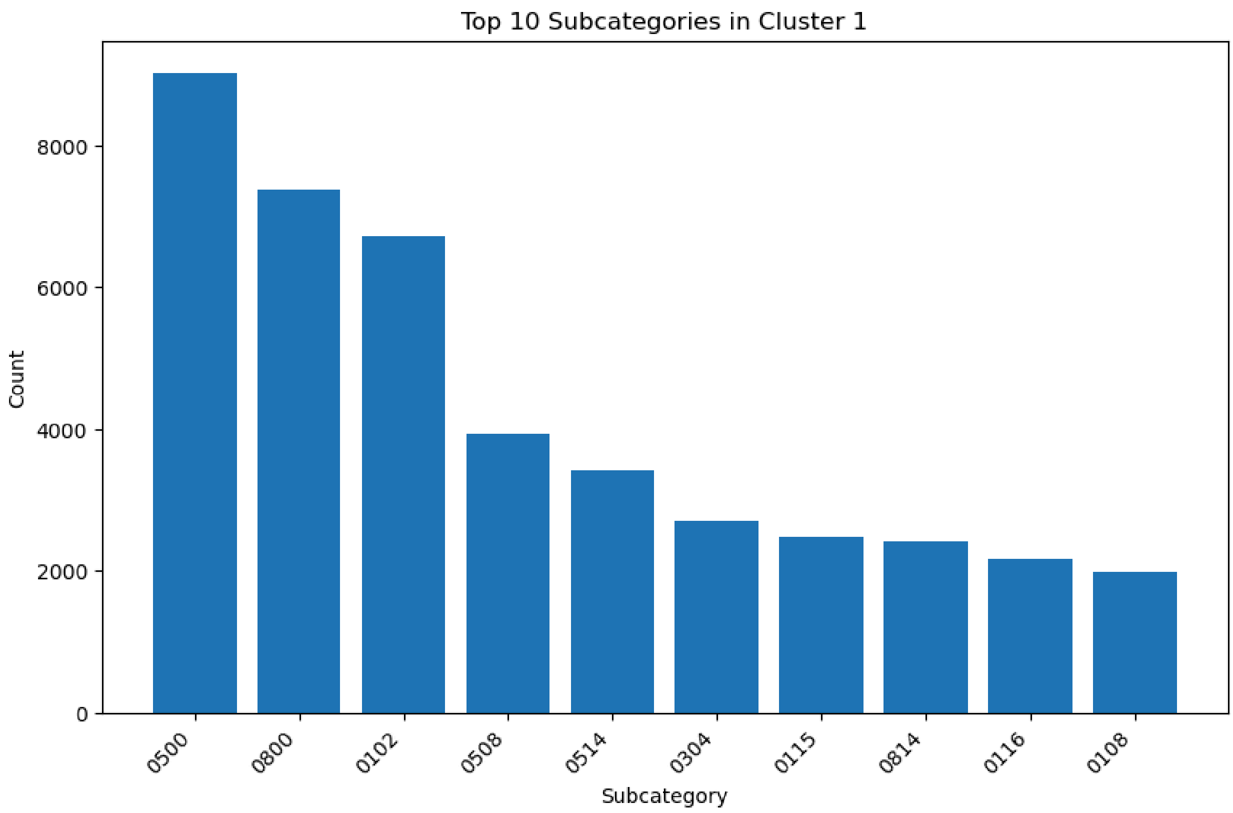
Looking at the purchasing habits of the most high-purchasing clusters– in the following example, 1– is necessary for deciding which items and item categories to include in advertising targeting this group.



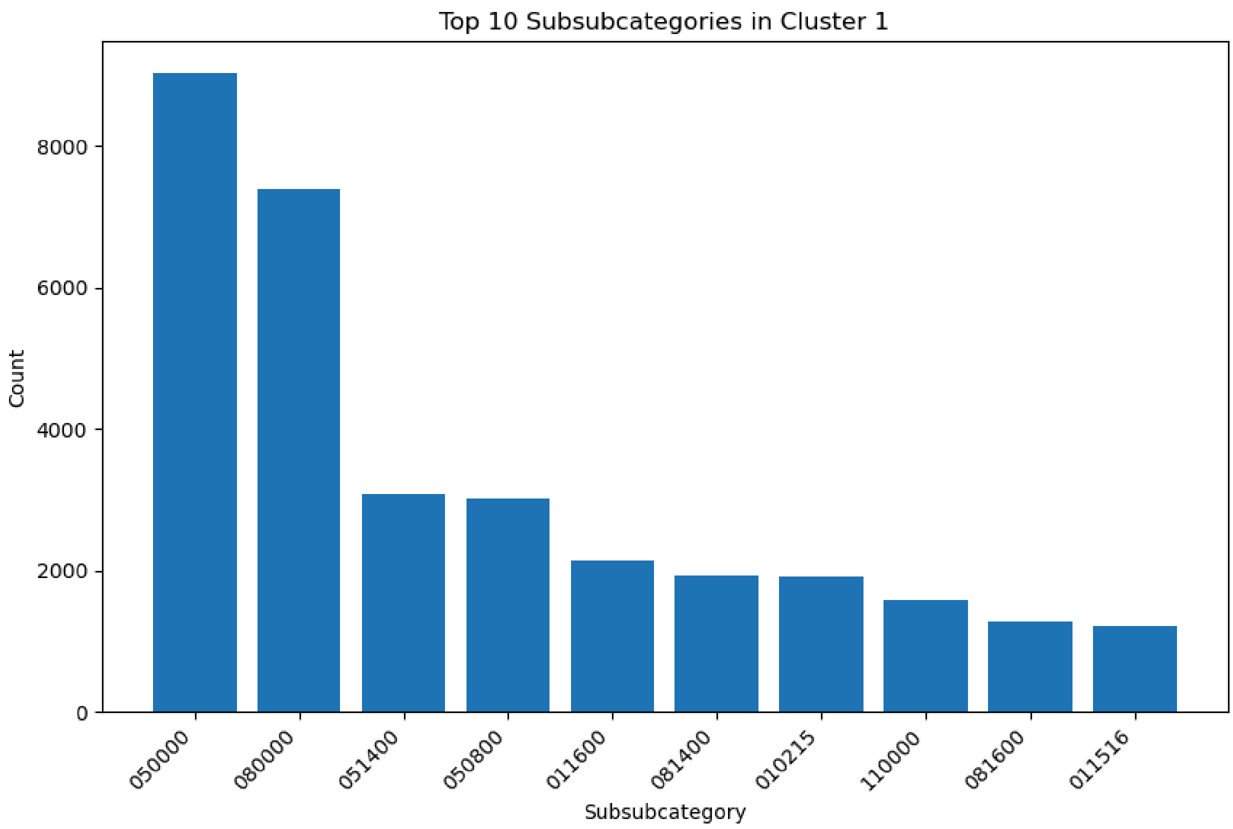
Within this cluster, certain products are more popular than others. For example, the top-selling individual product sold 250–300 units, while the broader category it belongs to accounted for roughly 20,000 units in total. This suggests that both category-level and product-level targeting could be valuable for advertising strategies.



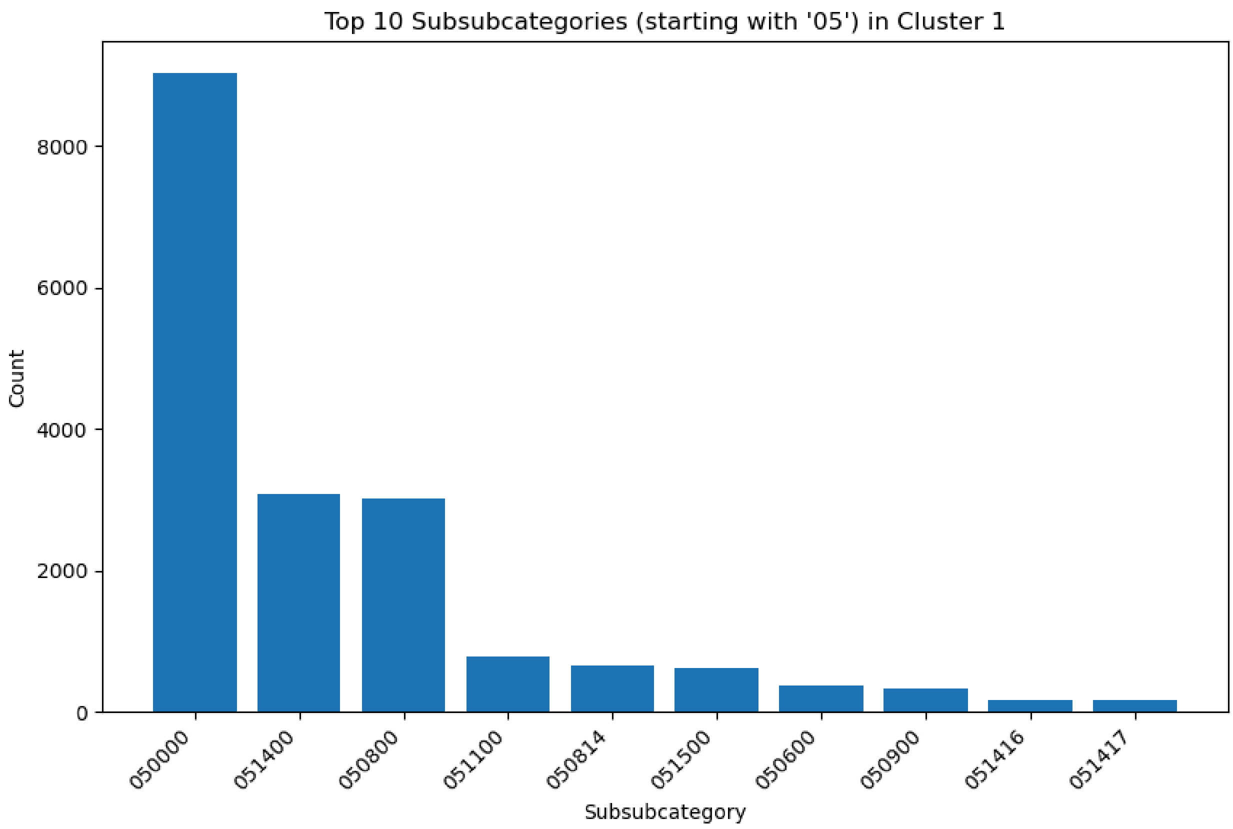
This shows that it would be more effective to advertise a category to this cluster. But we can be more specific by examining which subcategories within this category are most popular. Below, we can see that the subcategories that were most purchased by this cluster were in the 05, 08, 01 categories.



Since those categories are still broad, we can look more specifically at each to be more effective in advertising. Below, we see that there are some sub-subcategories that are significantly more popular than the rest.



If the business is going to advertise a specific category to this segment, like category 05 for example, they could look at the sub-subcategories in the “05” category to determine which would be most effective to advertise. Below we see that “050000”, “051400”, and “050800” are significantly more popular than the rest, so these would be among the most effective categories to advertise to segment 1. The same analysis could be done to determine the most effective categories to advertise to the rest of the highest purchasing segments.



Conclusion

In summary, this analysis provides actionable insights to fine-tune the business's advertising strategies. By examining the preferences and demographics of high-spending clusters, it's clear which segments should be targeted and which product categories should be advertised to them.