# Customer Segmentation and Analysis

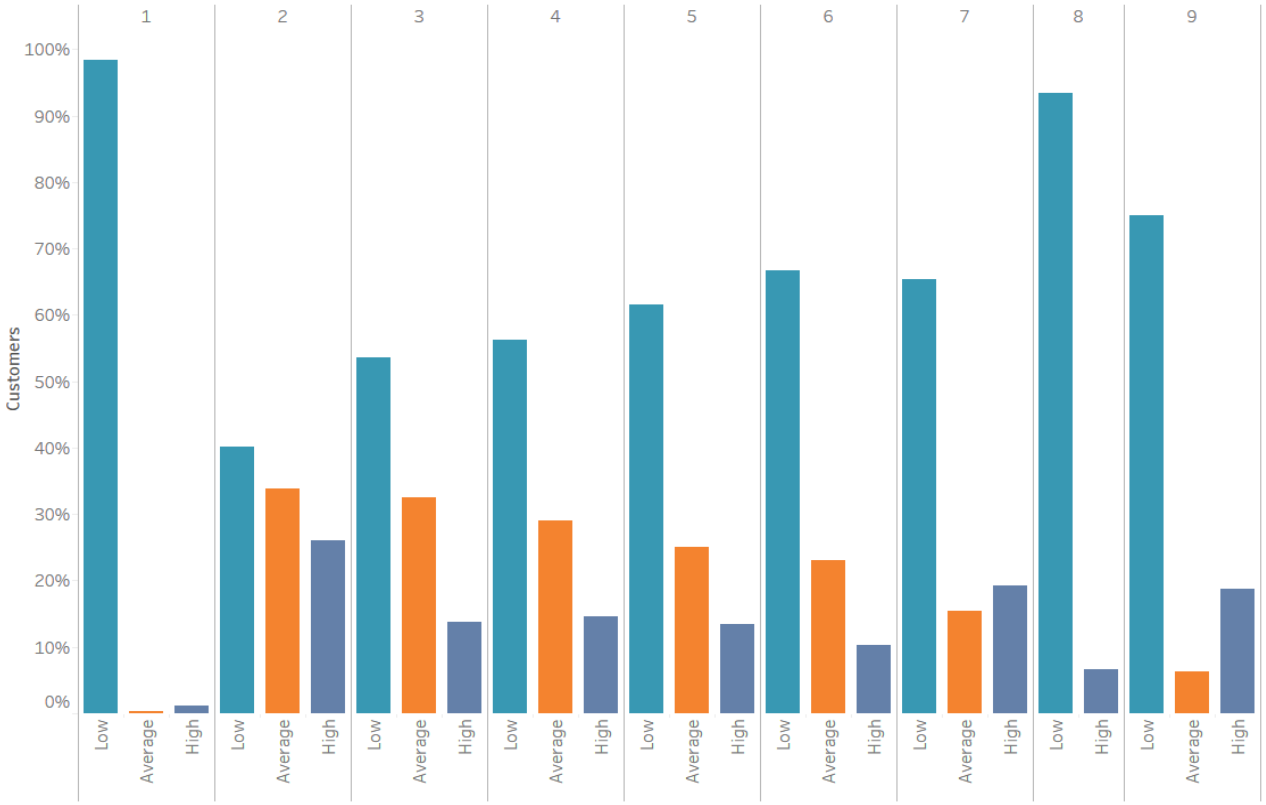
Garvit Kashyap
BITS Pilani, Hyderabad Campus

# Table of Contents

# 1. Exploratory Data Analysis

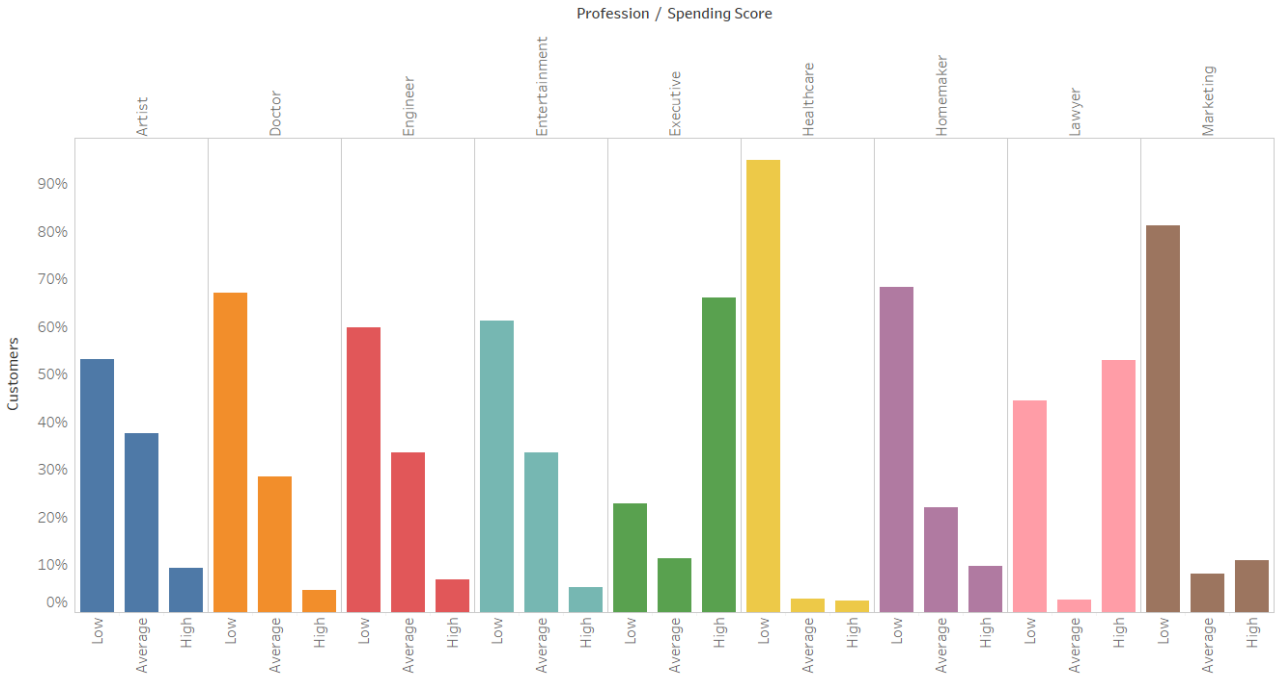
The data contains customer data, where a customer has specific properties like Age, Education, Marital Status, Work Ex,  Family size, etc. A spending score can take three values, *Low*,  *Average*, and *High*.

The following shows some preliminary plots based on the data -




A plot of Spending Scores with the **size of the family** of the customer. In all of the different types of families, the highest count is of *Low* spending score. Apart from family size 2, the difference of *Low* with *High* and *Average* doesn't seem comparable, with almost everyone single being a *Low* spender.

Now we will look at how the spending score seems to vary with the **profession** of the customer.

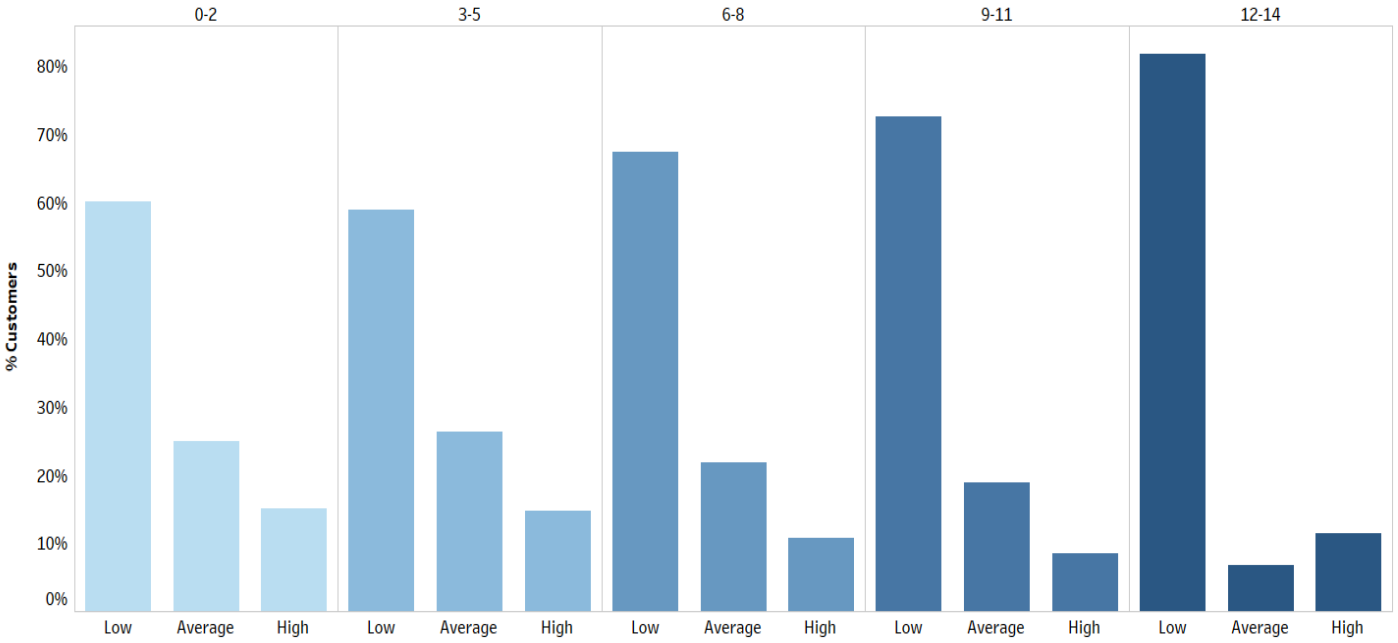



Given that the count for *Low* spending customers is higher than other categories, in most professions, *Low* is the highest, except for Lawyers, who have similar counts of *Low* and *High*, and for Executive, which has a majority of *High* spenders. As all professions don't have an equal number of records, the percentage of the total is taken as a metric. Most healthcare personnel seem to be *Low* spenders.

*Low* spenders again are in the majority as far as the **educational background** of the customer is concerned. Graduate customers have a much higher count of being *Average* spenders than those who are not graduates.

| Graduated | Low | Average | High |
|---|---|---|---|
| No | 694 | 167 | 140 |
| Yes | 908 | 452 | 242 |

Based on **work experience**, similar to previous plots, the majority of customers are *Low* spending. The fraction of *Low* spenders seem to be increasing as work experience increases.




Now we will see how spending scores appear to vary with **marital status**. All customers who have never been married are *Low* spenders, whereas most married customers are low/*Average* spenders, with the remaining being *High* spenders.

| Ever Married | Spending Score | |
|---|---|---|
| No | Low | 1,057 |
| Yes | Low | 535 |
| | Average | 610 |
| | High | 375 |

## 2. Dataset

The customers dataset has 10 (used 9) columns, which contain the features/properties.

***Gender***: *Male or Female*

***Ever Married***: *Yes or No*

***Age***: *Integer (Range - 18 to 89)*

***Graduated***: *Yes or No*

***Profession***: *[Engineer, Artist, Lawyer, Executive, Healthcare, Executive, Doctor]*

***Work Ex***: *Integer*

***Spending Score***: *Average, Low or High*

***Family Size***: *Integer (Range- 1 to 9)*

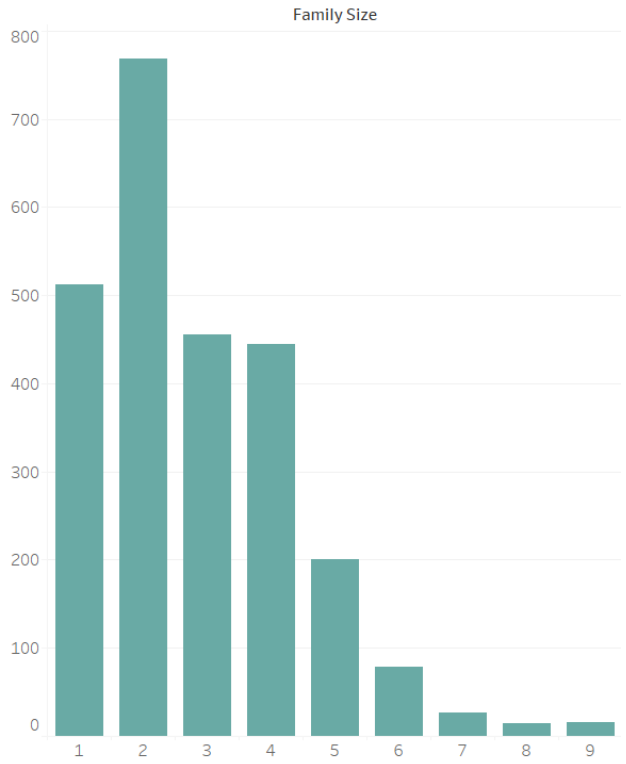
*Dataset View -*

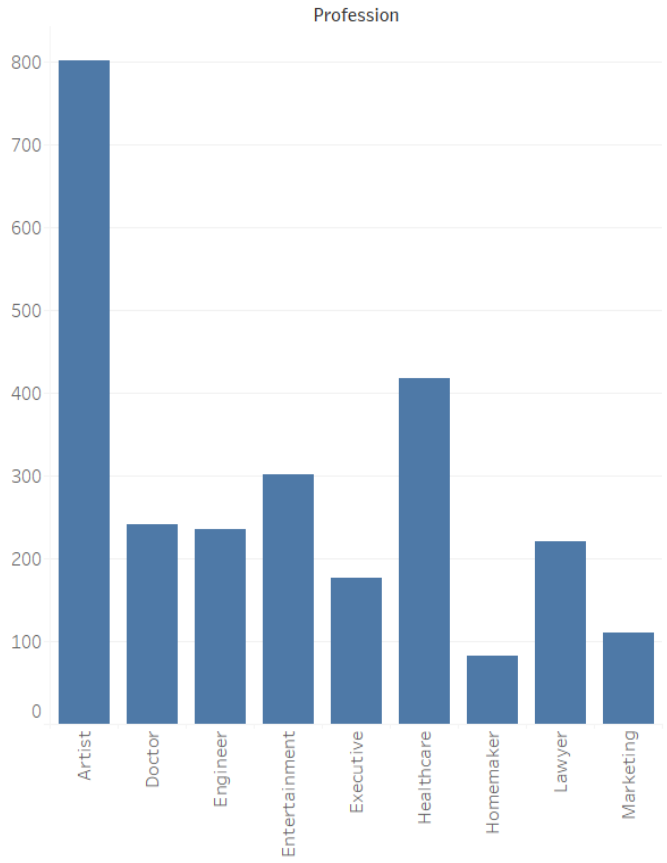
| ID | Gender | Ever_Married | Age | Graduated | Profession | Work_Experience | Spending_Score | Family_Size |
|---|---|---|---|---|---|---|---|---|
| 458989 | Female | Yes | 36 | Yes | Engineer | 0.0 | Low | 1.0 |
| 458994 | Male | Yes | 37 | Yes | Healthcare | 8.0 | Average | 4.0 |
| 459000 | Male | Yes | 59 | No | Executive | 11.0 | High | 2.0 |
| 459003 | Male | Yes | 47 | Yes | Doctor | 0.0 | High | 5.0 |
| 459005 | Male | Yes | 61 | Yes | Doctor | 5.0 | Low | 3.0 |

### a. Dataset Visualization

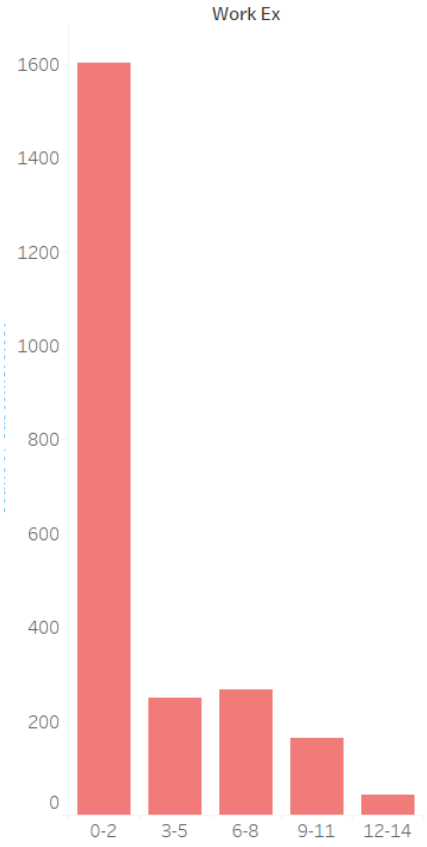
Now we will see the distribution of features in the dataset, starting with the numerical fields:

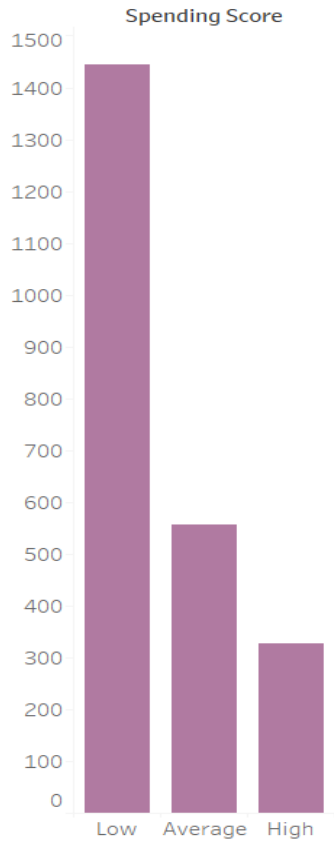
| | Age | Work_Experience | Family_Size |
|---|---|---|---|
| **count** | 2154.000000 | 2154.000000 | 2154.000000 |
| **mean** | 43.461467 | 2.551532 | 2.837047 |
| **std** | 16.761895 | 3.344917 | 1.566872 |
| **min** | 18.000000 | 0.000000 | 1.000000 |
| **25%** | 30.000000 | 0.000000 | 2.000000 |
| **50%** | 41.000000 | 1.000000 | 2.000000 |
| **75%** | 52.000000 | 4.000000 | 4.000000 |
| **max** | 89.000000 | 14.000000 | 9.000000 |




Apart from most customers being *artists* or *healthcare workers*, it is clear that many customers in the dataset are *Low* spenders. The majority of people have *work experience* between 0-2 years.



### b. Dataset Preparation

In the dataset, we have binary features, categorical features and discrete features. We will have to transform such features into numerical form to prepare them for further steps.

#### 1. One-Hot Encoding

One-Hot encoding is a popular way to convert categorical data into numerical form.
It splits one feature into a matrix of 1s and 0s.
*Profession*, *Spending* has been transformed the same way.

| Artist | Doctor | Engineer | Entertainment | Executive | Healthcare | Homemaker | Lawyer | Marketing | Average | High | Low |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |

#### 2. Binary Features

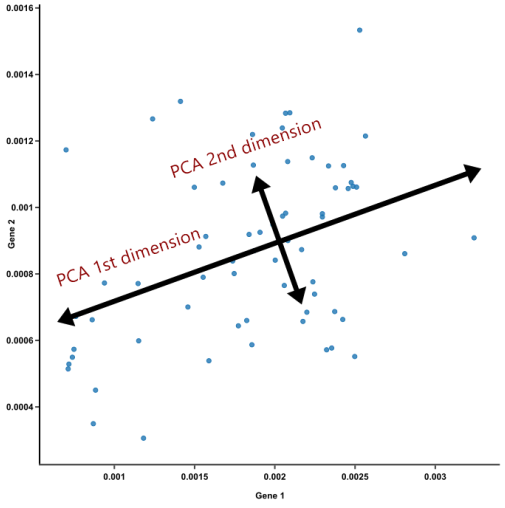
***Graduated***: *1 if Yes else 0*
***Gender***: *1 if Male else 0*
***Ever_Married***: *1 if Yes else 0*

After performing these transformations, and removing any records with NaN, the size of the dataset is (2154,18). Now the next step will be dimensionality reduction.

## 3. PCA (Principal Component Analysis)

PCA is a dimensionality reduction method. It transforms high-dimensional datasets into smaller ones and tries to contain the critical information in the dataset. PCA attempts to reduce the dimensionality and push most of the information in the dataset's first few components (dimensions). The first principal component is the one with the maximum variance.

*Figure* 1 and *Figure* 2 show the plots obtained after applying K-Means Clustering to the reduced data for 2 and 3 dimensions, respectively.



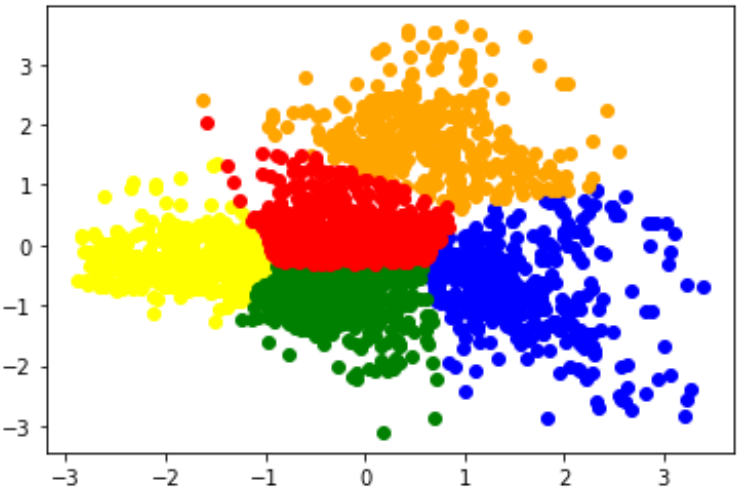




*Figure 1*

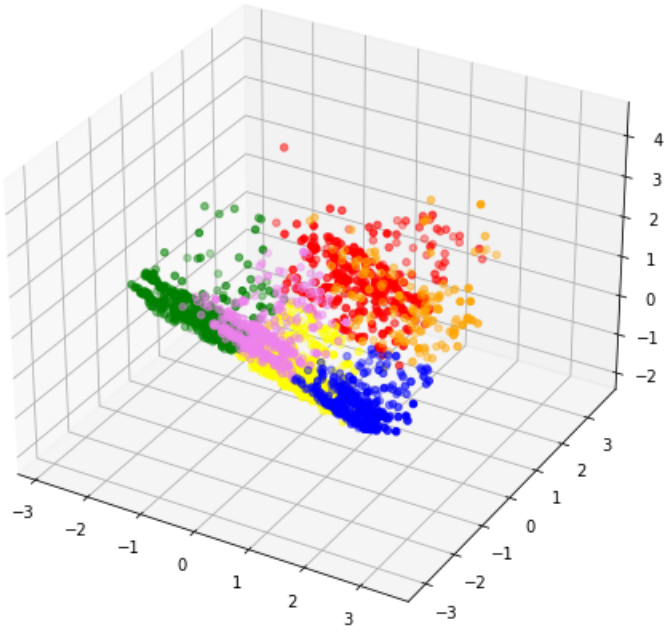
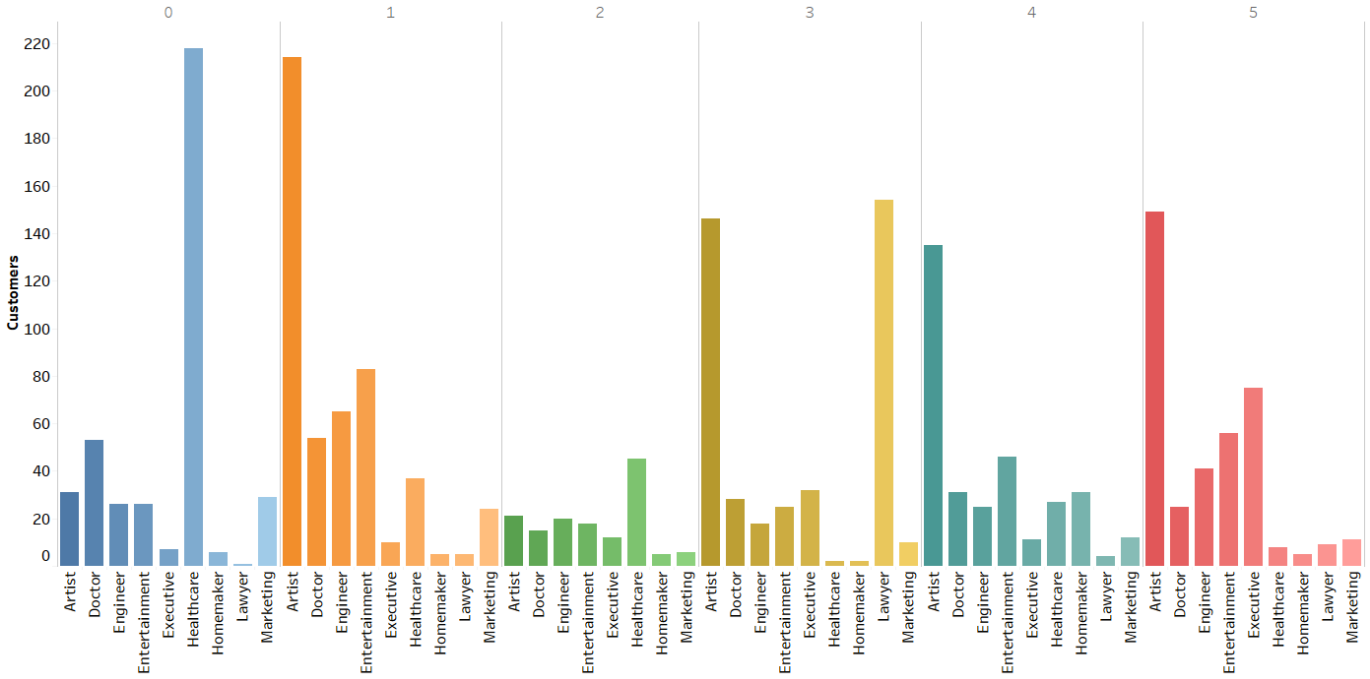



*Figure 2*

# 4. Visualising The New Segmented Data

Now we will visualise the newly segmented clusters and see if we can get any insights into the data. It shows that *Segment* 0 has most healthcare workers, most of whom happen to be *Low* spenders.




Distribution of spenders across all segments -

| Segment | Spending Score Average | Spending Score High | Spending Score Low |
|---|---|---|---|
| 0 | 5 | 2 | 390 |
| 1 | 77 | 11 | 409 |
| 2 | 19 | 18 | 105 |
| 3 | 112 | 155 | 150 |
| 4 | 70 | 20 | 232 |
| 5 | 238 | 101 | 40 |

# Analysing Generated Segments

| Spending Score | Profession | Segment 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|
| Average | Artist | | 37 | 3 | 66 | 33 | 121 |
| | Doctor | 1 | 9 | 2 | 18 | 11 | 22 |
| | Engineer | 1 | 10 | 7 | 7 | 6 | 31 |
| | Entertainment | | 16 | 4 | 12 | 9 | 47 |
| | Executive | | 1 | 1 | 2 | 1 | 11 |
| | Healthcare | 2 | 2 | 1 | 1 | 3 | 2 |
| | Homemaker | | 1 | 1 | | 6 | 2 |
| | Lawyer | 1 | | | 4 | | 1 |
| | Marketing | | 1 | | 2 | 1 | 1 |
| High | Artist | | 3 | 3 | 33 | 5 | 18 |
| | Doctor | | 2 | 1 | 3 | | 1 |
| | Engineer | | 3 | 1 | 2 | | 6 |
| | Entertainment | | 1 | | 2 | 2 | 4 |
| | Executive | 2 | 2 | 8 | 22 | 8 | 58 |
| | Healthcare | | | 2 | | 3 | 3 |
| | Homemaker | | | 1 | 1 | 1 | 1 |
| | Lawyer | | | | 87 | 1 | 5 |
| | Marketing | | | 2 | 5 | | 5 |
| Low | Artist | 31 | 174 | 15 | 47 | 97 | 10 |
| | Doctor | 52 | 43 | 12 | 7 | 20 | 2 |
| | Engineer | 25 | 52 | 12 | 9 | 19 | 4 |
| | Entertainment | 26 | 66 | 14 | 11 | 35 | 5 |
| | Executive | 5 | 7 | 3 | 8 | 2 | 6 |
| | Healthcare | 216 | 35 | 42 | 1 | 21 | 3 |
| | Homemaker | 6 | 4 | 3 | 1 | 24 | 2 |
| | Lawyer | | 5 | | 63 | 3 | 3 |
| | Marketing | 29 | 23 | 4 | 3 | 11 | 5 |

*Segment* 0 contains a lot of Low spenders and barely any spenders from the *High* or *Average* category.

*Segment* 1, along with *Segment* 0 and *Segment* 4, seems to contain most *Low* spending customers.

*Segment* 3 contains many high spending *Artists*, *Lawyers*, *Executives* who were *High* spenders in the initial analysis.

*Segment* 5 contains many *Average* Spenders and a few *High* Spenders from the *Executive* profession.

## 5. Training ML Models

The dataset with the new segments was trained on multiple algorithms, *RandomForest*, *DecisionTrees*, *Multi-Layer Perceptron* and *Logistic Regression*.

### a. Logistic Regression

It is a viral machine learning algorithm, and it uses the logistic function to model the probability distribution of a dataset.

### b. RandomForest

A random forest is an ensemble classifier that fits multiple Decision Trees on the dataset and takes the result by voting.

### c. DecisionTree

A decision tree splits the dataset into subsets based on attribute values. It sorts data down from the root node to the leaf node, which gives the classification result.

### d. Multi-Layer Perceptron

A multi-layer perceptron is a neural network with multiple hidden layers.

### e. Results

All of the models attained an accuracy of 95% and above, with the highest accuracy of 97% achieved by Multi-Layer Perceptron and RandomForestClassifier.

# Truncated SVD (Singular Value Decomposition)

Truncated SVD is another dimensionality reduction technique that uses Singular Value Decomposition to transform data to a lower dimension and only returns the specified number of columns.

*Figure* 3 and *Figure* 4 show the data distribution in the 2D and 3D space, where each colour represents a segment allotted after K Means.

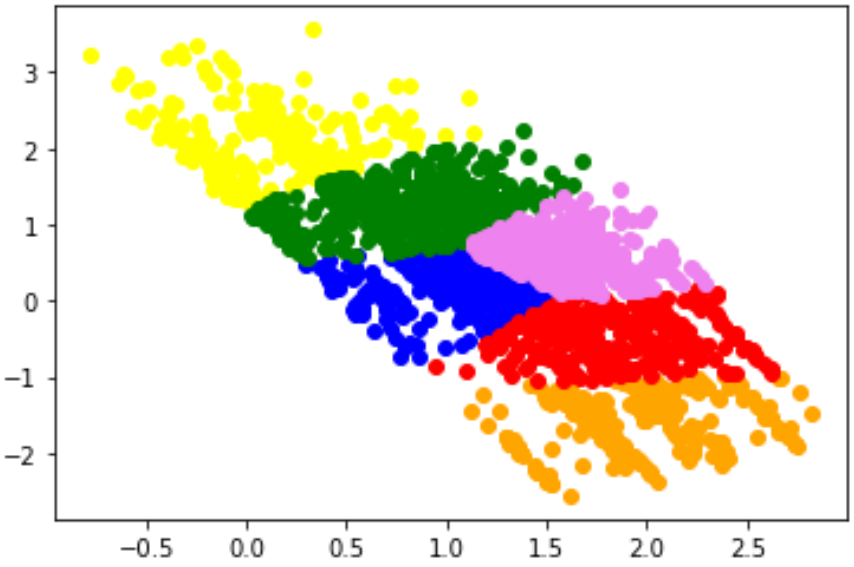



*Figure 3*

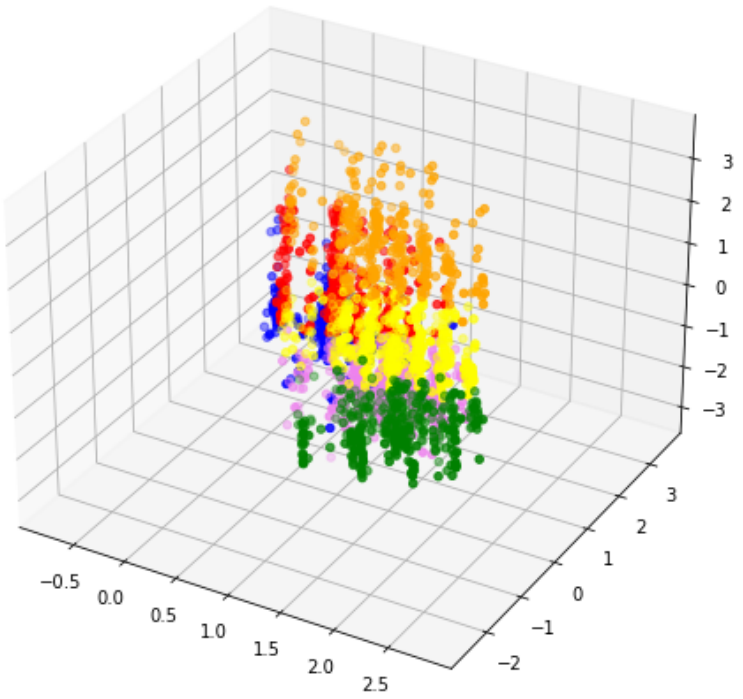



*Figure 4*

Now we'll be plotting similar graphs for SVD as we did for PCA.

| | | Segment | | | | | |
|---|---|---|---|---|---|---|---|
| Spending Score | Profession | 0 | 1 | 2 | 3 | 4 | 5 |
| Average | Artist | 2 | 68 | 20 | 44 | 1 | 125 |
| | Doctor | 3 | 17 | 8 | 14 | 2 | 19 |
| | Engineer | 2 | 12 | 7 | 3 | 7 | 31 |
| | Entertainment | | 19 | 9 | 11 | 4 | 45 |
| | Executive | 2 | 2 | | 2 | 1 | 9 |
| | Healthcare | 1 | 2 | 2 | 1 | 1 | 4 |
| | Homemaker | | 3 | 4 | | 1 | 2 |
| | Lawyer | | | | 4 | 1 | 1 |
| | Marketing | | 2 | | 1 | | 2 |
| High | Artist | | 10 | 4 | 31 | 2 | 15 |
| | Doctor | | 3 | | 3 | | 1 |
| | Engineer | 1 | 2 | | 2 | 1 | 6 |
| | Entertainment | | 1 | 2 | 2 | | 4 |
| | Executive | 6 | 20 | 2 | 22 | 10 | 40 |
| | Healthcare | | | 4 | | 1 | 3 |
| | Homemaker | | 1 | 2 | 1 | | |
| | Lawyer | 1 | 4 | 1 | 84 | | 3 |
| | Marketing | 1 | | 1 | 4 | 1 | 5 |
| Low | Artist | 16 | 151 | 98 | 45 | 37 | 27 |
| | Doctor | 33 | 32 | 22 | 8 | 38 | 3 |
| | Engineer | 16 | 33 | 22 | 11 | 26 | 13 |
| | Entertainment | 14 | 57 | 35 | 9 | 26 | 16 |
| | Executive | 5 | 4 | 2 | 8 | 3 | 9 |
| | Healthcare | 128 | 22 | 29 | 1 | 130 | 8 |
| | Homemaker | 6 | 4 | 25 | | 1 | 4 |
| | Lawyer | | 5 | 2 | 62 | | 5 |
| | Marketing | 22 | 16 | 11 | 4 | 15 | 7 |

Similarly, like PCA, using SVD, it's become possible to group customers into different segments, *Segment* 3 containing many *High* spenders and *Segments* 0 *and* 1 containing a majority of *Low* Spenders. The same ML models were trained for this dataset as well, and the minimum accuracy achieved was 94% on the test data set, which contained 154 of 2154 examples.

# Conclusion

Using dimensionality reductions techniques, such as *Principal Component Analysis and Singular Value Decomposition*, we have managed to cluster the dataset into segments with somewhat similar properties. Using Machine Learning algorithms, it has also become possible to make models that can predict customers based on these segments. *Segment 0* contains a lot of *Healthcare* workers, the majority of whom are *Low Spenders*. *Segment 3* in both the analyses includes many *High-Spending Lawyers*, *Executives and Artists*, and similarly many *Low* and *Average* spending customers of the same profession, who can be future loyal customers.