Cluster Analysis of Ghanaian Households using Household Expenditure Data

Ameyaw Emmanuel

I. Introduction

For policymakers and practitioners to design policies and services that affect household expenditure patterns for particular segments of the population (for example poorer households), there need to be a good understanding of how households are segmented based on their expenditure patterns. For example, are expenditure patterns among different segments of the population the same but differ on expenditure levels? Or is there a huge difference in regard to how different segments of the population spend their money on the various expenditure categories in Ghana. The findings of this project reveals that while lower and middle class Ghanaian households have similar expenditure patterns, upper class Ghanaian households, however, have a completely different spending pattern.

In this report, we present the results of two machine learning clustering algorithms/models (auto-encoder with kmeans and auto-encoder with DBSCAN). The results of these two models are very similar including model performance and the number of clusters found in the dataset although they are completely different from each other in how they work. Also, these two models are the best models among several other clustering algorithms used to cluster the dataset. The silhouette coefficient for the auto-encoder with kmeans model is 0.986 whiles the silhouette coefficient for the auto-encoder with DBSCAN is 0.982.

II. The Dataset and Features

This project uses household expenditure data made available by the Ghana Living Standards Survey 7 (GLSS 7). The 12 household expenditure categories used in this project includes the following:

1. expenditure on alcoholic beverages, tobacco and narcotics (TOTALCH)
2. expenditure on clothing and footwear (TOTCLTH)
3. expenditure on communication (TOTCMNQ)
4. expenditure on education (TOTEDUC)
5. expenditure on food & non-alcoholic drinks (TOTFOOD)
6. expenditure on health (TOTHLTH)

7. expenditure on housing, water, electricity, gas and other Fuels (TOTHOUS)

8. expenditure on furnishing, household equipment and routine maintenance (TOTFURN)

9. expenditure on transport (TOTTRSP)

10. expenditure on recreation and culture (TOTRCRE)

11.expenditure on miscellaneous goods and services (TOTMISC)

12. Hotel Cafes and Restaurants (TOTHOTL)

A normalized version of the features above were used as inputs to the machine learning model. In the next section of this report, we will describe the data in much detail and how it was gathered and analyzed. We will then present the results of the cluster analysis. Finally, we will present the conclusion of this study.

III. Descriptive statistics and cluster analysis

The final dataset used for the project is a merger of 13 different dataset including demographic data of the households and 12 household expenditure categories. All the 13 dataset are available here: http://www2.statsghana.gov.gh/nada/index.php/catalog/97/study-description. The data is filtered for households that reported all values for the 12 household expenditure categories. The final dataset contains 8922 households fairly spread across the 10 regions of Ghana. Urban and rural households are also fairly distributed in the data. This helps to avoid any biases that might result from an imbalanced dataset, that is, a dataset that is not a true representation of the country. Food expenditure is the dominant expenditure among the 12 expenditure categories followed by expenditure on education. Expenditure on hotels and restaurants is the lowest among the 12 categories. Furthermore, there is low correlation between the attributes with the highest correlation coefficient less than 0.5. Also, the distribution of all attributes in the dataset are heavily skewed to the right suggesting that there are outliers in the dataset. As a result, we compare the results of two clustering algorithms – kmeans and DBSCAN. While DBSCAN is robust to outliers, kmeans is not. But as we will see in later sections of this project, both models give similar results in a deep learning framework.

For all the 12 categories, urban households had higher expenditures than rural households except expenditure on drugs and narcotics. Also, not surprisingly, household expenditures in the southern part of the country is generally higher than those in the northern part of the country due to the presence of larger cities in the southern part of the country.

A. Cluster Analysis: Auto-encoder with kmeans

The auto-encoder with kmeans is one of the several clustering algorithms used for this project. This is one of the baseline models due to the presence of outliers in the dataset. While the kmeans algorithm without an auto-encoder performed poorly with a silhouette coefficient of 0.378, the performance of the kmeans model significantly improved under a deep learning framework. The silhouette value for the auto-encoder with kmeans model was 0.986. The auto-encoder which is a deep learning model first extract relevant information from the input data and as a result reduces its dimensionality. This information is then passed to a kmeans algorithm which serves as a clustering layer to the auto-encoder. The result is a cluster of the data points into various groups. To visualize the results, we use the PCA technique to reduce the dimension of the dataset to 2 components. The result is shown below:

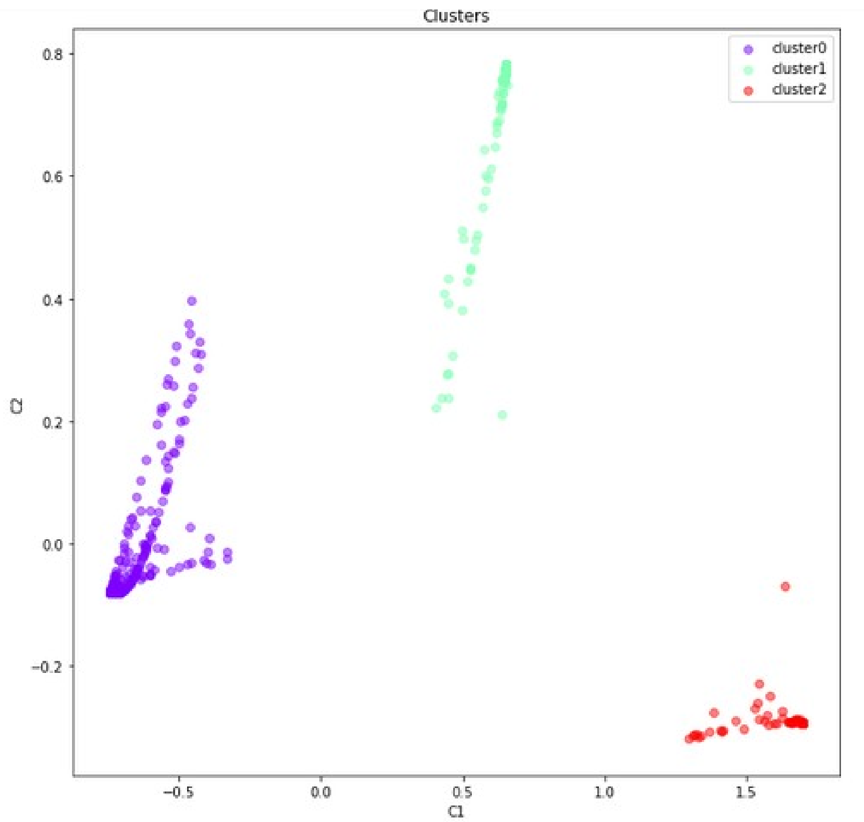


Fig 1: Auto-encoder with kmeans and PCA

From the figure above, we see that the clusters are clearly not overlapping. Thus there are three distinct clusters in the dataset. The number of optimal clusters for the dataset is three as revealed by the elbow method for optimal number of clusters. The elbow method plot is shown below where it can be seen that the sum of squared distances is constant after k = 3.

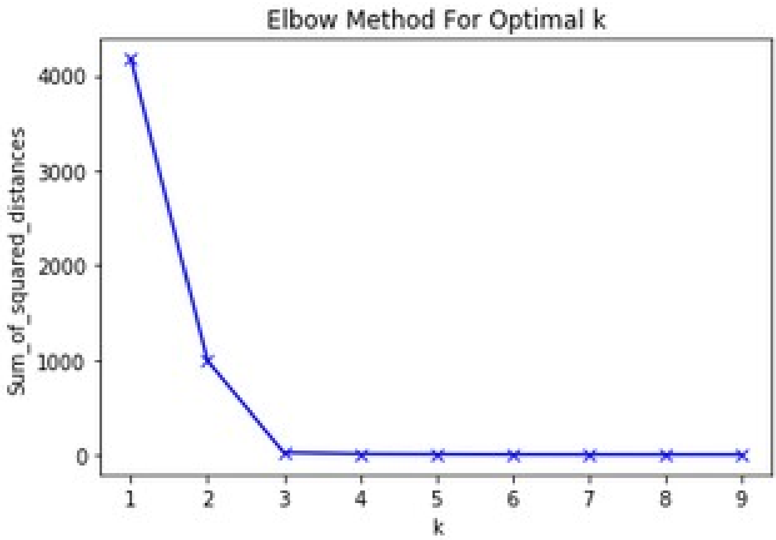


Fig 2: Elbow method for optimal k

To analyze which attributes are important for each cluster, we build a parallel plot using the results of the auto-encoder. In the figure below, we see the three clusters with the vertical axis measuring the importance of each feature/attribute in a given cluster. For cluster 0 and 1, expenditure on alcohol, tobacco and narcotics (totalch) is the most important feature among the 12 expenditure categories followed by food and health expenditures, however, the level of expenditure differs in these two clusters. Generally, mean household expenditure in cluster 1 is more than mean household expenditure in cluster 0.

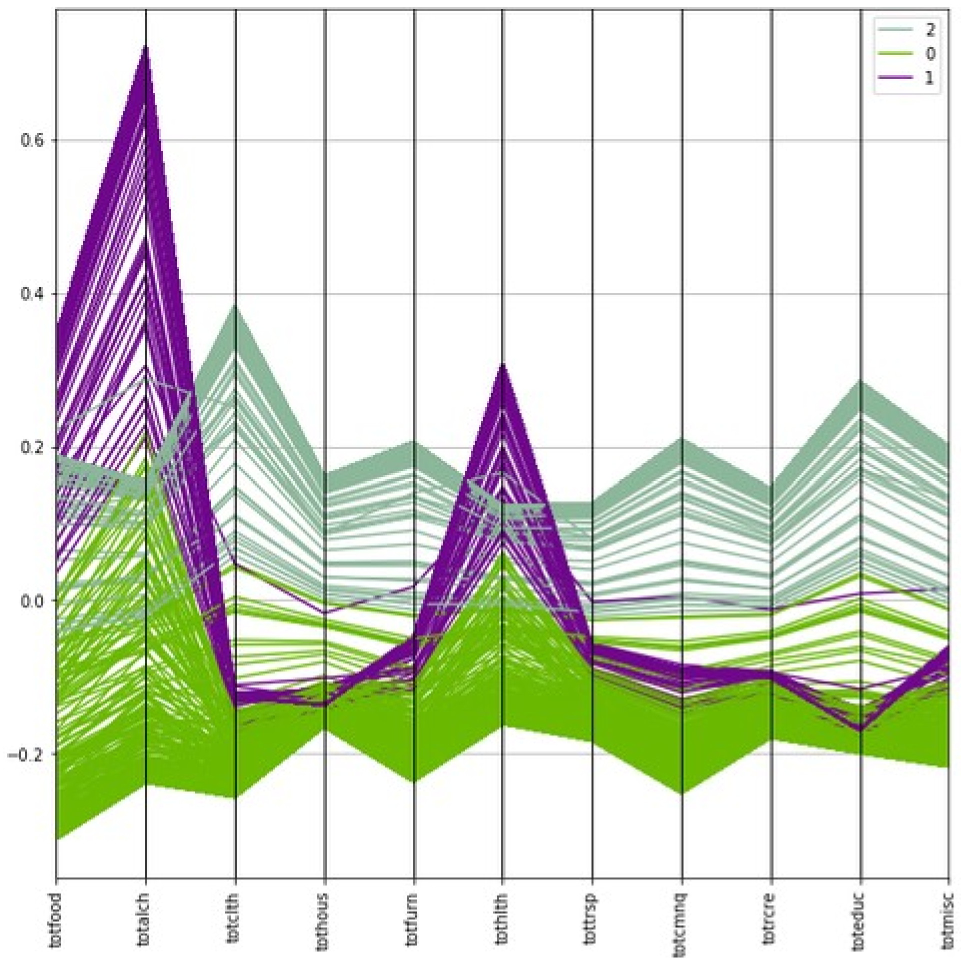
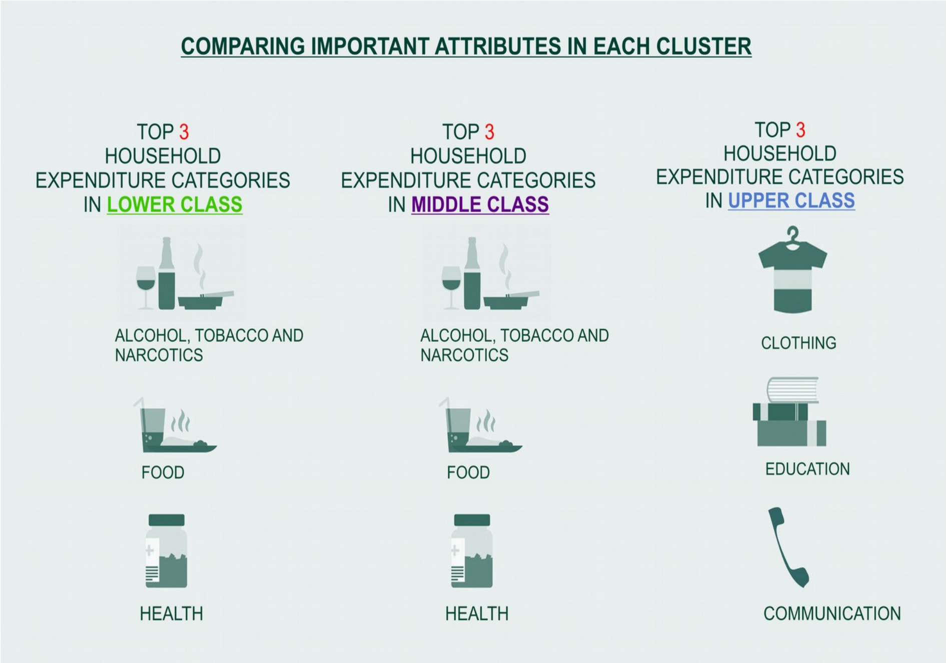


Fig 3: Parallel plot

For cluster 2, the top two important features or attributes that identifies this cluster are clothing expenditure and expenditure on education. This expenditure pattern is completely different from the household expenditure patterns exhibited in cluster 0 and 1. Generally, we also see that mean household expenditures are higher in cluster 2 compared to cluster 0 and 1 for majority of the 12 household expenditure categories.

Given the discussion above, we can view the clusters as lower class households (cluster 0), middle class households (cluster 1) and upper class households (cluster 2) based on the mean value of household expenditures in each cluster. In other words, the clusters resemble these divisions (lower, middle and upper class). In this regard, we see that majority of middle class households spend more on alcohol and health than upper class households.

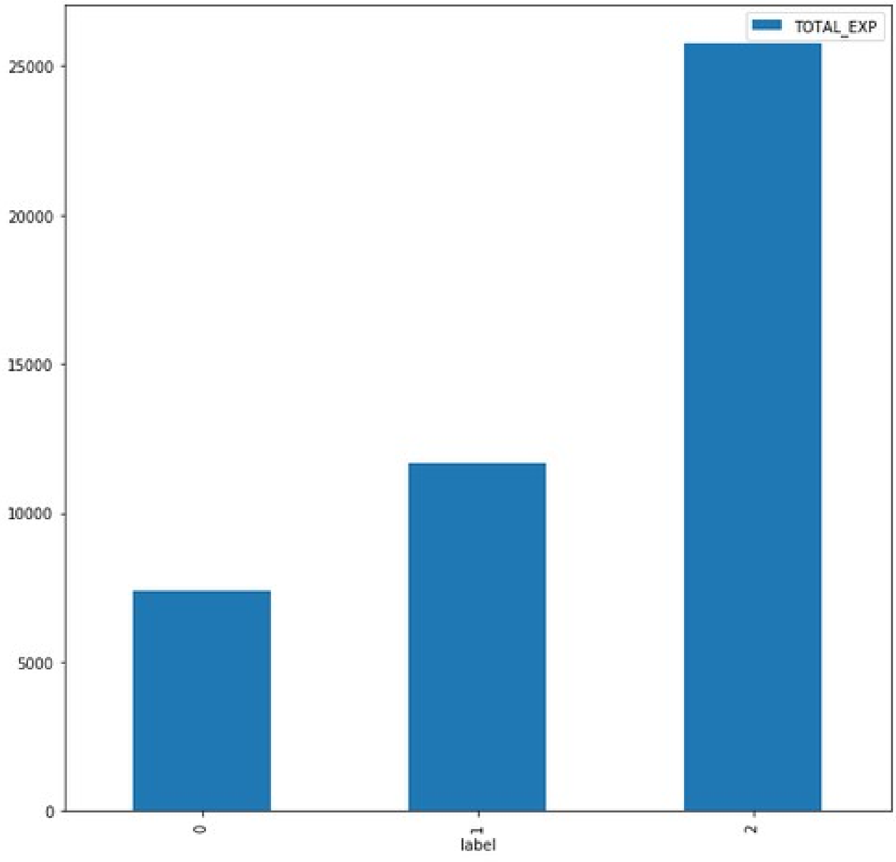
In the figure below, we show the top 3 expenditure categories for each of the clusters.



Important features in each cluster

In the graph below, we show the mean value of total household expenditure for each cluster. Clearly, we see that mean household expenditure in cluster 2 (Upper class) is almost twice the mean household expenditure in cluster 1 (middle class). Also, mean household expenditure in cluster 1 (middle class) is significantly higher than mean household expenditure in cluster 0 (lower class).

Fig 4: Mean household expenditure by cluster



Also, by counting the number of households in each cluster, we see that majority of the households in our sample belongs to cluster 0 (lower class) which is typical of household expenditure distribution.

Table 1: Households per cluster

Cluster	**Number of households**
Cluster 0 (lower class)	5657
Cluster 1 (Middle class)	1301
Cluster 2 (Upper class)	1964

Lastly, we also show the mean expenditure for the 12 expenditure categories for each cluster.

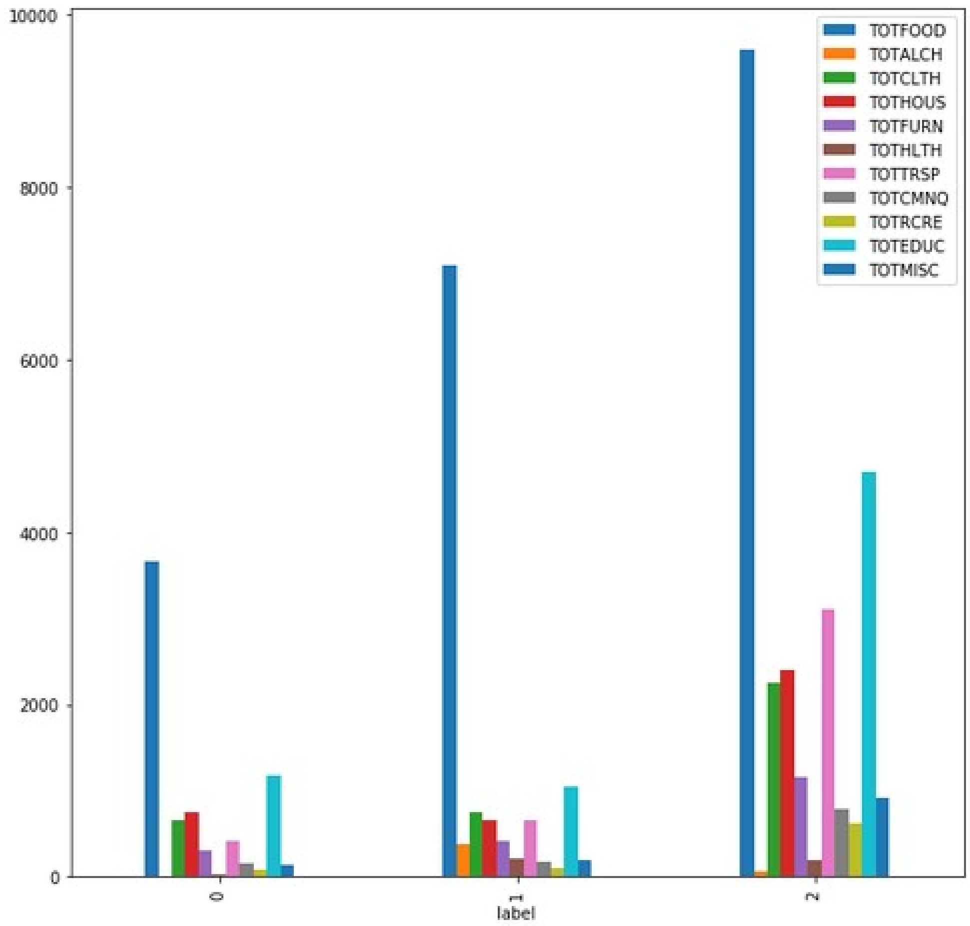
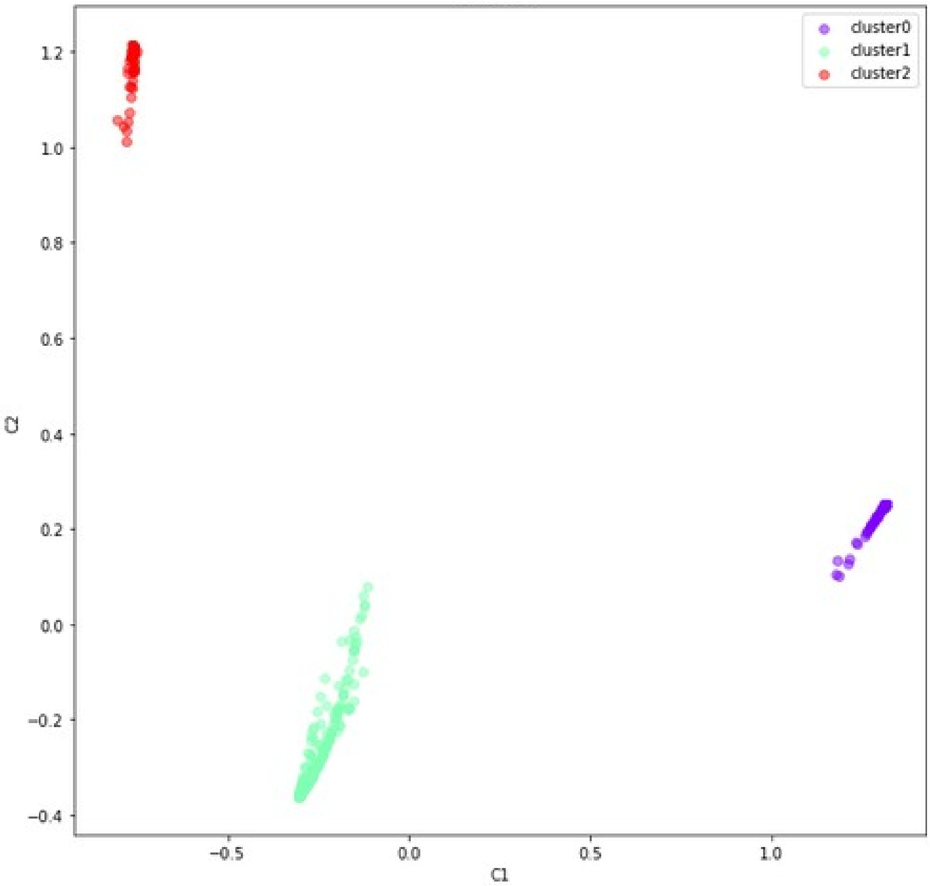


Fig 5: Mean expenditure for the 12 expenditure categories for each cluster

B. Cluster Analysis: Auto-encoder with DBSCAN

The auto-encoder with DBSCAN produces similar results as the auto-encoder with kmeans. One difference between these two models is that DBSCAN is robust to outliers while kmeans is not. Moreover, the optimal number of clusters is not predefined in the DBSCAN algorithm. Rather, it is optimally chosen and the optimal number of clusters simultaneously reported with the clustering results. Using this dataset, three distinct clusters were found. Using PCA, we can visualize these clusters.

Fig 6: Cluster visualization using PCA



Similar to the auto-encoder with kmeans, these clusters resembles lower class, middle class and upper class as suggested by the parallel plot below. In the table below, we also count the number of households in each cluster. Take note of the change in clusters labels. The cluster with a label of -1 contains noisy data points and should be ignored.

Table 2: Households per cluster

Cluster	Number of households
Cluster 1 (lower class)	5626
Cluster 2 (Middle class)	1281
Cluster 0 (Upper class)	1945

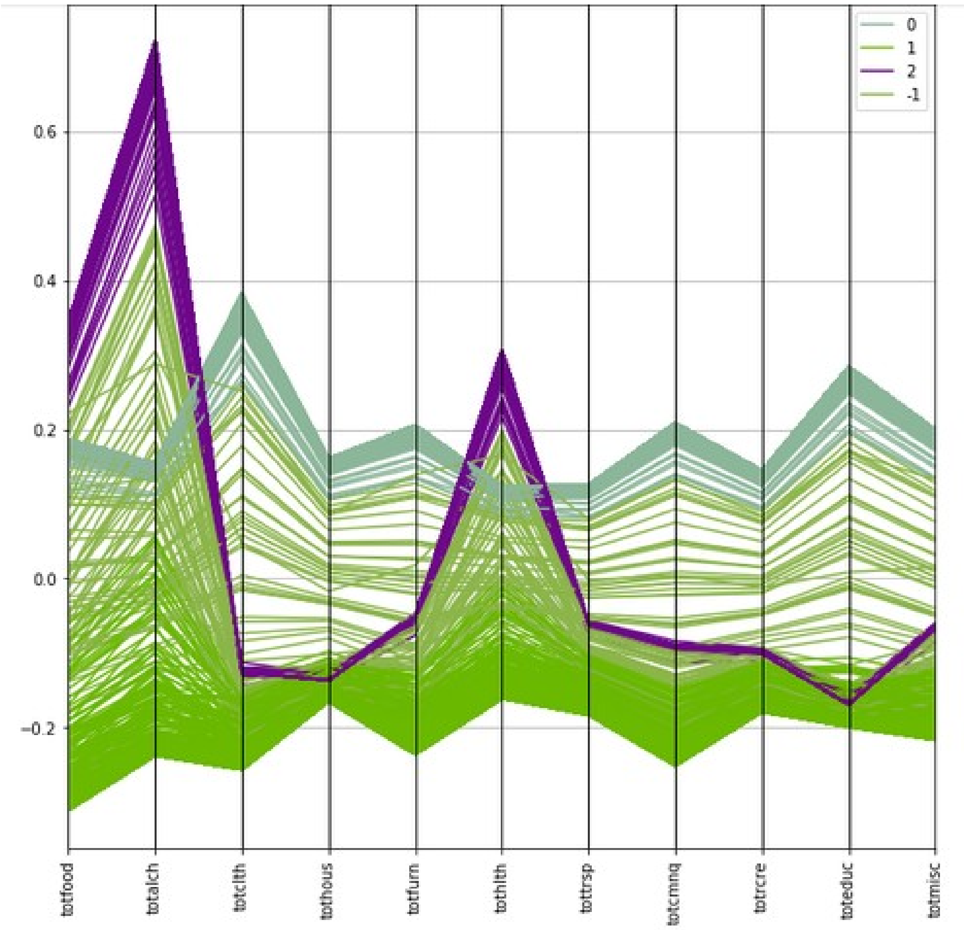


Fig 7: Parallel Plot

In the parallel plot above, we see that while lower (cluster1) and middle class (cluster2) households have similar spending patterns, upper class households (cluster0) have a completely different spending pattern. The top two expenditure categories in the upper class are expenditure on clothing & footwear and expenditure on education while the top three expenditure categories in the both the lower and middle class households are expenditures on ‘alcohol, tobacco and narcotics’, ‘food’ and ‘health’.

We also show mean household expenditure by cluster which is similar to the results under auto-encoder with kmeans. Mean household expenditure in cluster 0 (Upper class) is almost twice the mean household expenditure in cluster 2 (middle class). Also, mean household expenditure in cluster 2 (middle class) is significantly higher than mean household expenditure in cluster 1 (lower class).

Finally, we also present mean household expenditure by cluster and by household expenditure category.

Fig 9: Mean household expenditure by cluster

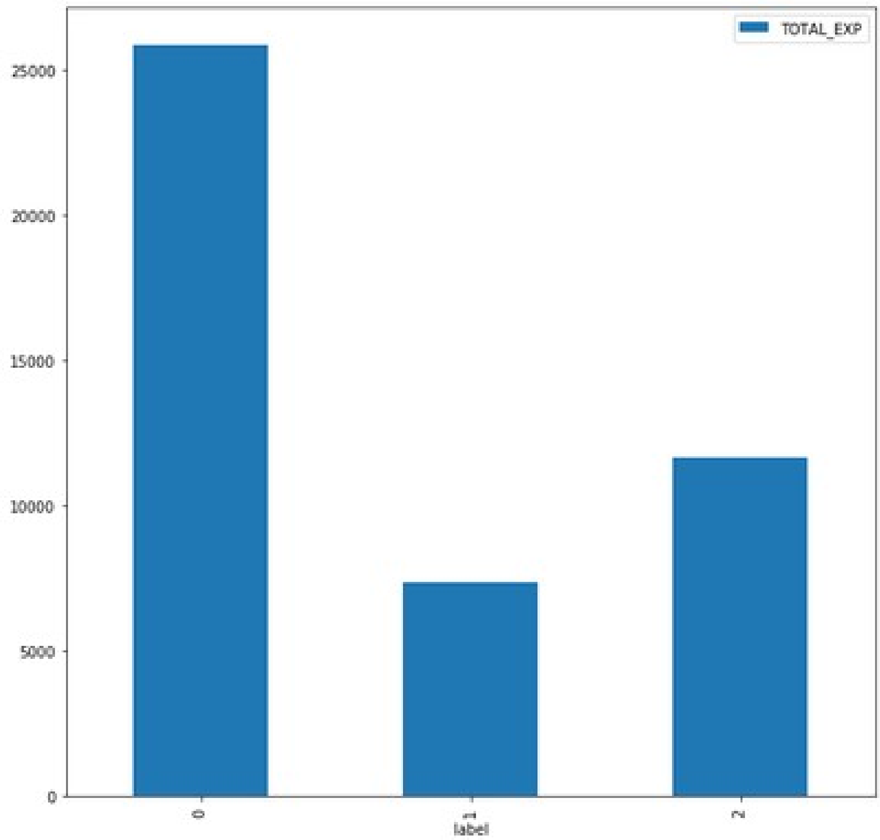
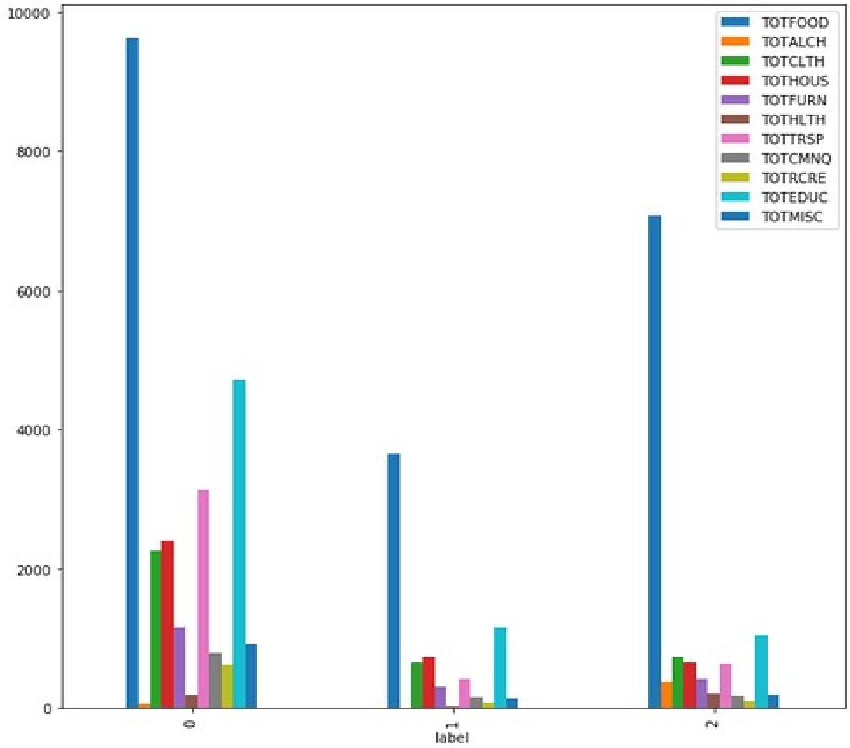


Fig 10: Mean household expenditure by cluster and household expenditure category



IV. Conclusion

We have clustered a representative sample of Ghanaian households to determine expenditure patterns among various segments of the population. Using a normalized data and two deep learning models, we found 3 optimal clusters in the dataset for both models. These two models (auto-encoder with kmeans and auto-encoder with DBSCAN) reported high silhouette coefficient – 0.986 and 0.982 respectively. Also, we found that Ghanaian households are generally segmented based on their level of household expenditure. The three optimal clusters resembles lower, middle and upper class households. Lower and middle class households have similar expenditure patterns which is completely different from expenditure patterns of upper class households. For example, the majority of middle class households spend more on alcohol and narcotics compared to upper class households. Similarly, majority of middle class households spend more on food than upper class households. We need further research to explain this discrepancy in household spending especially between middle and upper class households for expenditure categories such food, alcohol, tobacco and health.

NB: All notebooks from data exploration to model deployment could be found here: https://github.com/EmmanuelAmeyaw/IBM-capstone-prject