The cost of living crisis in Scotland*: predicting the impact on life satisfaction and wellbeing*

Student ID: 202258581

October 2023

Contents

1 Introduction

Inequalities in health are related to inequalities in wealth, income and social capital (*Bambra 2021*). The COVID-19 pandemic exacerbated preexisting health and wealth inequalities. Living in a deprived or rural area, being on a low income or minority ethnic significantly increased the chances of death (*Bambra 2021*). In the UK, some of the most alarming inequalities were seen in Scotland, with the death rate being over double for those in the most deprived areas of Scotland, compared to the least (*Bambra 2021*). Figure 1 shows how these structural inequalities can be evidenced by how financially secure people feel, and how this, in turn, affects the resources they have available to adopt healthy lifestyles and choose behaviors positively impacting their health and well-being (*Shatz 2019*).

Figure 1: Structural Inequalities: *social and economic factors affecting people's ability to cope with daily living, leading to anxiety, stress, and a reduction in overall life satisfaction.*

In 2022, JRF published two waves of data relating to the impact of the COVID-19 pandemic (first) and the Cost of Living Crisis (second). This paper focuses on the second wave of data to consider the impact the Cost of Living Crisis (COLC) has had on people living in Scotland. The focus will be on the social, economic and behavioural variables and how these relate to life satisfaction. In particular, we will look to assess to what extent an individual's circumstances and characteristics may be associated with their life satisfaction having been negatively impacted (NI) by the COLC. For example, is somebody with a disability more likely to be NI by the COLC? Are those with no savings, on a low income, more likely to be NI than those with a high income and savings? Being able to predict life satisfaction as a result of the COLC can enable Government to develop evidence based policies as an effective preventative method.

2 Methods

2.1 Data Set

The JRF data consists of 4203 rows, representing individuals, and 186 columns, representing variables, which have been re-coded from the original questionnaire. I have sub-setted the data to include the following variables shown in Table 1. Integer and one-hot encoding has been performed on all variables already. Throughout my code, I have dealt with answers such as *do not know*, blank values and *NaNs*, by removing these from the results, as this was the best way to deal with categorical missing values that were only covering only a small section of the data set.

Variable Name	**Variable topic**	**Questionnaire answers**
Negmental	Mental health NA	Recoded 1 OR 2
Negphysical	Physical health NA	Recoded 1 OR 2
Negsocial	Social life NA	Recoded 1 OR 2
Negpersonal	Personal life NA	Recoded 1 OR 2
Q141	Mental health NA	Scale 1-4
Q142	Physical health NA	Scale 1-4
Q143	Social life NA	Scale 1-4
Q144	Personal life NA	Scale 1-4
Disability	Disabled or not	1 or 2
D3D	Rural or city	1 or 2
Employment	Employment type	Sliding scale
Famtype	Family type	Sliding scale
Q2Benefits	Benefits or not	1 or 2
HIDD10	Tenure	Sliding scale
Essentials	Cut back on essentials	1 or 2
Q1510	Used a warm bank	1 or 2
Q12	Seek help with bills	Sliding scale
Q8W22	Belief in Government	Sliding scale
Income	Income level OECD	1 -3
Savings	Savings level	Sliding scale
Debtcat	Debt level	Sliding scale
Financial Security	How financially secure you feel	Sliding scale

Table 1: Variables used from original JRF dataset

3 Analyses

3.1 Life Satisfaction Variables

I have called the four outcome variables *life satisfaction* variables, and they include mental health, physical health, social life and personal life. Firstly, I considered whether the four life satisfaction variables are linear, and whether they are normally distributed, as this will impact the type of analysis used. I have created histograms and density plots to visualise the distributions. As Figure 2 shows, the variables are either right skewed, or equally distributed across 3 data points. Normality was also checked using *Shapiro-Wilk test* from SciPy. This was also looked at for the other financial variables (more detail on these below) as these are scale variables. However, as you can see in Figure 3, these were also not normally distributed.

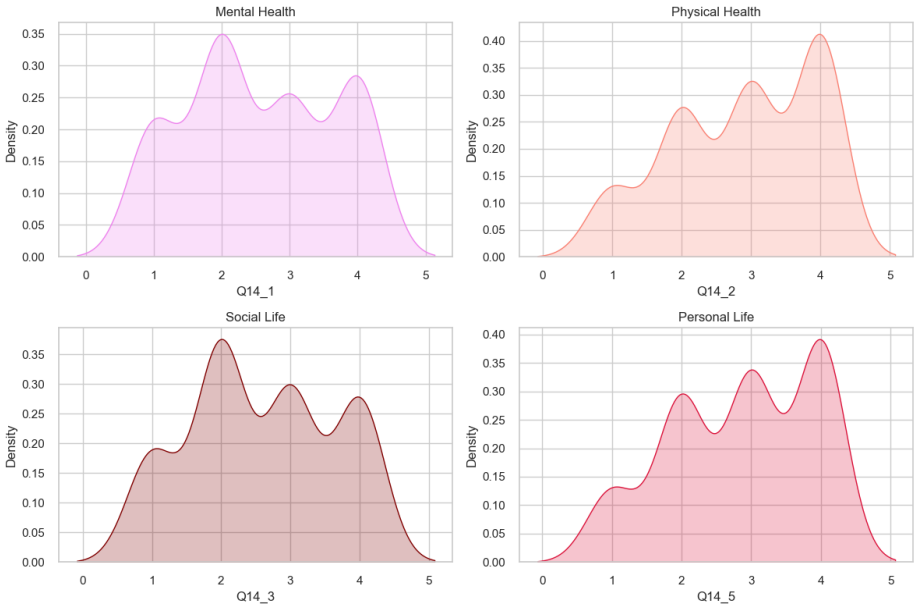

Figure 2: *Density plot showing life satisfaction variables distribution*

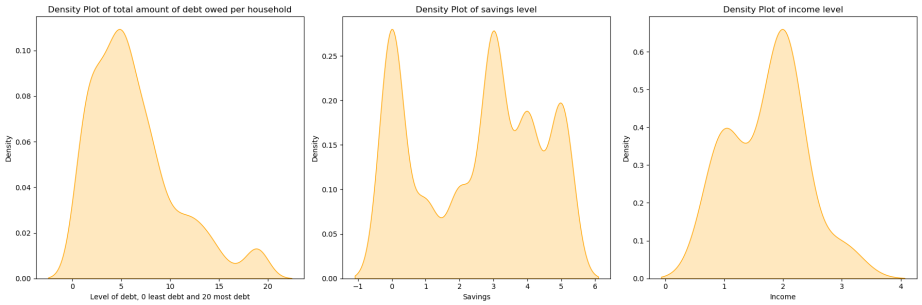

Figure 3: *Density plot showing financial variables distribution*

3.2 Sociodemographic Variables

Here I considered the impact the COLC has had on people depending on sociodemographics. In particular, I assessed whether individuals with the demographics on the left hand side of Table 2, were more likely to be NI by the COLC, in comparison to the reference group, on the right of Table 2. This table is just looking at the mental health variable shown in Table 1, where the answers have been re-coded to a binary outcome. This is due to the variable not being normally distributed, therefore a categorical analysis is preferred to linear.

If we consider those who are disabled, 60 percent had their mental health NI by the COL crisis in comparison to only 33 percent of those who are not disabled. For unemployed people, 79 percent have been affected, in comparison to high professionals, where it is 40 percent. The biggest difference can be seen by tenure type, with 70 percent of those socially renting having been NI and only 28 percent who own outright. It should be noted that some of these variables increased on a sliding scale; for example a higher proportion of private renters or unskilled workers were NI compared to mortgage owners or skilled professionals.

Consideration was next given to each variable and how these relate to all life satisfaction variables. Figure 4 shows these comparisons for the variables with the largest differences. Famtype was removed, as although there is a difference between the groups, the proportion of both groups being NI is high, therefore the graph was not so clear.

Variable	Category of interest	Percent NI	Reference Category	Percent NI
Disability	Disabled	60	Not Disabled	33
D3D	Remote	40	Accessible	50
Employment	Unemployed	79	High Professional	40
Famtype	Single Kids	79	Couple Kids	61
Q2Benefits	Benefits	59	No Benefits	38
HIDD10	Social Renter	70	Own Outright	28
D4DUM	Ethnic Minority	54	White	50

Table 2: *Percentage having had their mental health affected, by demographic group.*

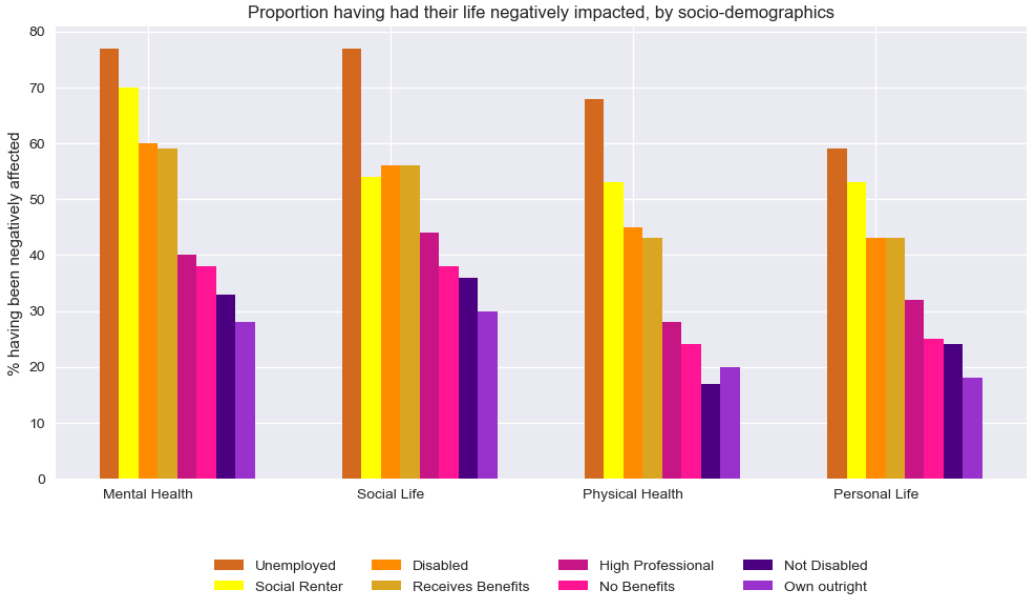

Figure 4: *Percentage having had their life satisfaction negatively impacted across the four areas (mental health, social life, physical health and personal life) by demographic.*

As Table 2 and Figure 5 show, a higher proportion of those who have the characteristics on the left, are more likely to be NI by the COLC (warm colors) than the comparison categories (cool colors).

3.3 Behavioural and Belief Variables

Next, we will consider any associations between the impact the COL has had on people depending on their behaviours and beliefs. We are looking to see whether cutting back on items, seeking help from warm banks and charities alongside believing the Government is not providing enough support, are associated with being NI by the cost of living crisis. Table 3 shows the variables considered, and

the percentages of those having been NI.

Variable	Category of interest	Percent NI	Reference Category	Percent NI
Essentials	Cut back on essentials	61	Did not cut back on essentials	24
Q1510	Used warm bank	76	Did not use warm bank	42
Q12	Seek help with bills	80	Pay bills with own money	23
Q8W22	Economy is unfair	59	Economy is fair	35

Table 3: *Percentage having had their mental health affected, by behaviour or belief.*

As Figure 5 shows from the warmer pastel colors, a higher percentage of those from the categories on the left of Table 3, have had their life satisfaction negatively impacted compared to a smaller percentage of those in the cold pastel colors, from the right side of Table 3.

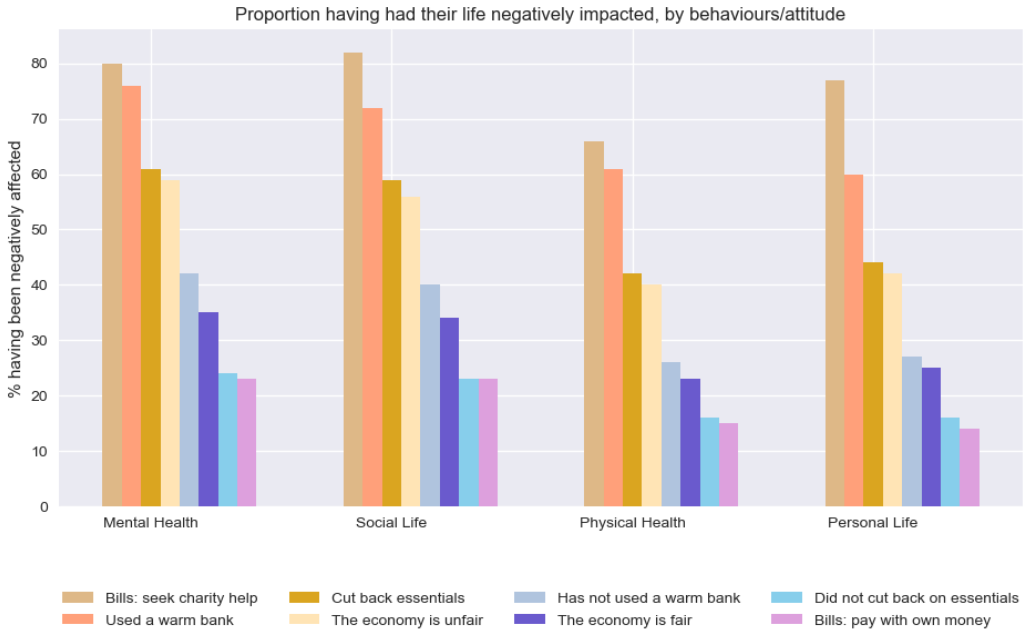

Figure 5: *Percentage having had their life satisfaction negatively impacted across the four areas (mental health, social life, physical health and personal life) by behaviors and beliefs.*

3.4 Socioeconomic Variables

Here we will consider the impact the cost of living has had on people depending on their financial circumstances. Table 4 shows the four variables and the percentages of those whose mental health has been NI, by socioeconomic factor.

Figure 6 shows the same financial variables, but by each life satisfaction variable; mental health, physical health, social life and personal life. As we can see, a high proportion of those on a low income, those with no savings, those with 3 or more items of debt, and those who feel financially insecure have been negatively impacted (warm colors). This can be compared to those categories on the right of Table 4 where we can see a much lower percentage of people have been negatively affected in terms of their life satisfaction. Some extra re-coding was completed here, so the scale switched into the opposite direction, as this affected the correlation matrix explored next.

Variable	**Category of interest**	**Percent NI**	**Reference Category**	**Percent NI**
Income	Low income	69	High income	25
Savings	No savings	78	Savings above £50,000	18
Debt	Debt 3+ items	82	No debt	37
Financial Security	Not finally secure	85	Financially secure	19

Table 4: *Percentage having had their mental health affected, by socioeconomic factor.*

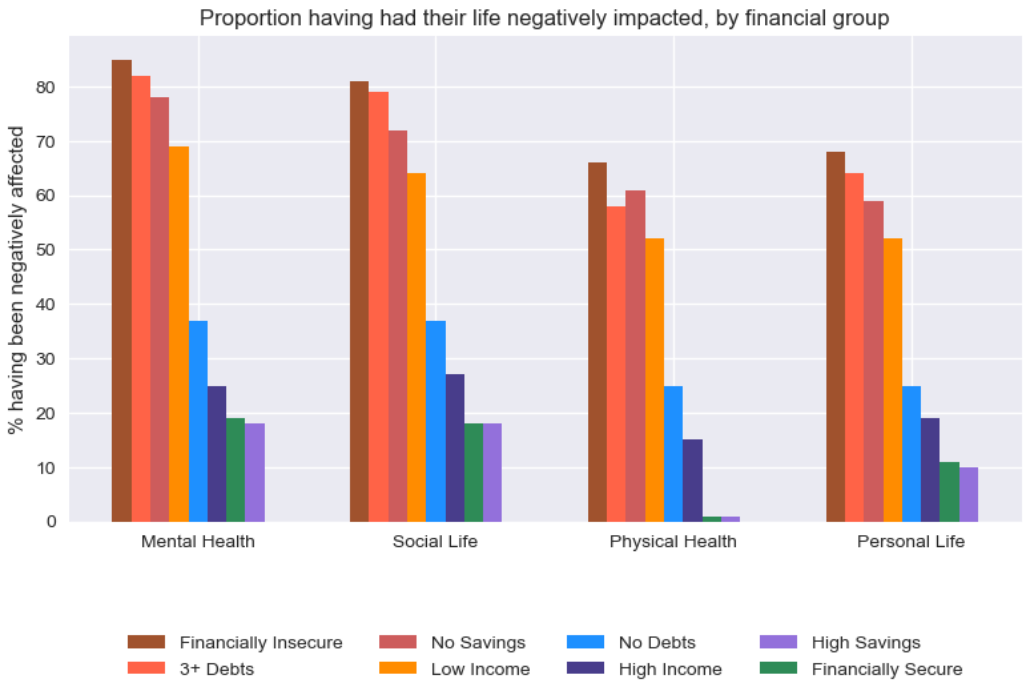

Figure 6: *Percentage having had their life satisfaction negatively impacted across the four areas (mental health, social life, physical health and personal life) by socioeconomic factor.*

3.5 Correlation Matrix of All Variables

Figure 7 below displays the Spearman's rank correlation coefficients for variables relating to the impact of the cost of living across the four life satisfaction variables.

The heat-map shows the strongest positive relationships are between the life satisfaction variables, as these show an association between your mental health, social life, personal life and physical health not being impacted by the cost of living. These are to be expected however, as these are quite similar variables. Consideration was given to creating a new variable showing a mean score across the four satisfaction variables, for a more general 'life satisfaction' score. However, as these variables are quite similar we may start by running further analysis looking at just one of these variables.

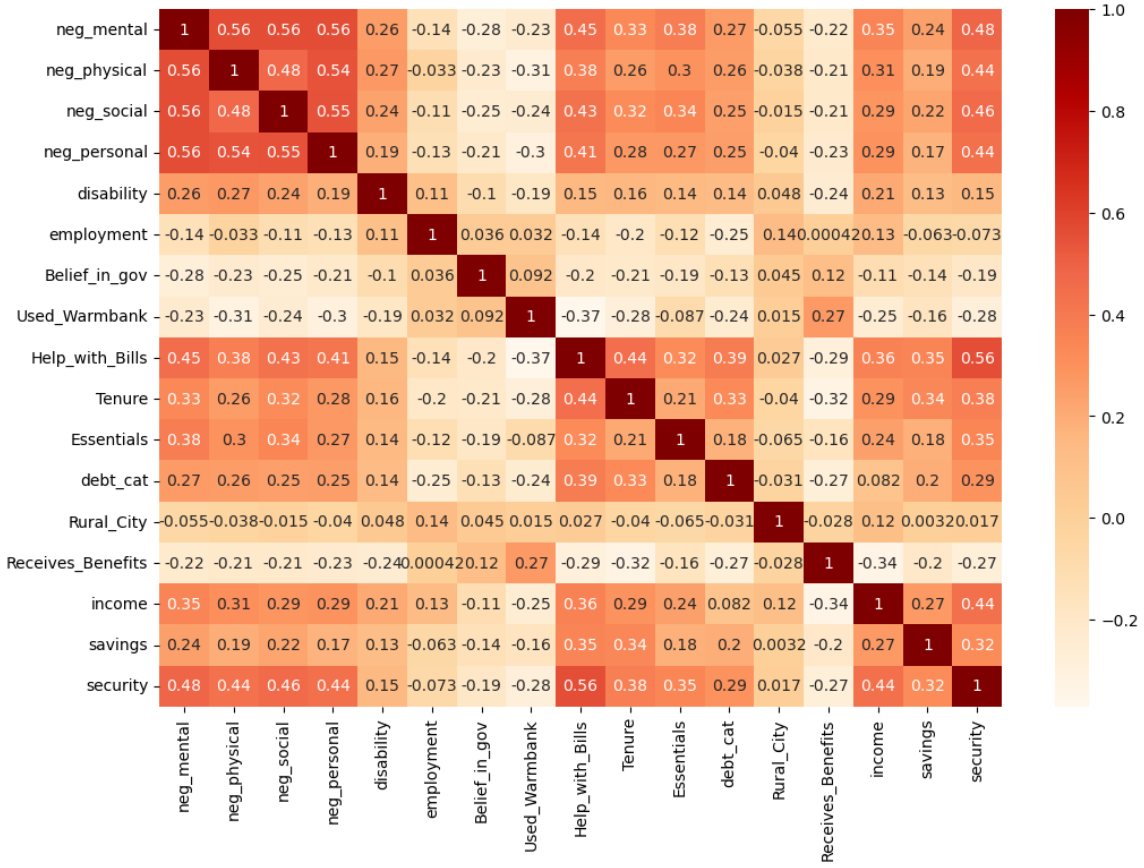

Figure 7: *Heatmap showing all variables considered so far. The variables with the strongest correlations will then be used in further analysis.*

We can see the strongest associations are between life satisfaction variable scores and help with bills, tenure type, cutting back on essentials, savings, income and financial security. We can also note that how financially secure a person feels, or perceives themselves to be, has a stronger association with how the COLC has impacted them, when compared to income or savings, despite the fact that actual income and savings are potentially a truer and more objective representation of financial insecurity than how financially secure somebody feels.

4 Machine Learning Approaches

4.1 Unsupervised learning

Clustering is an unsupervised machine learning method used on unlabelled data, allowing us to group data points together, based on similarities between features. Two frequently used methods are Hierarchical Clustering and K-Means.

K-means looks at the instances around a particular point (its centroid) to try to assign instances to the closest blob, and uses Euclidean distance, which means the data should be on a continuous scale. Here, we will use K-means, as we are not looking for hierarchical clusters, but rather groups of clusters based on similar features. Some of these variables will not suit k-means as it is categorical. The new, subsetted data, containing just continuous data can be seen in Table 5.

Variable Name	**Variable topic**	**Questionnaire answers**
Q141	Mental health NA	Scale 1-4
HIDD10	Tenure	Sliding scale
Income	Income level OECD	1 -3
Savings	Savings level	Sliding scale
Debtcat	Debt level	Sliding scale
Financial Security	How financially secure you feel	Sliding scale

Table 5: *The selection of continuous only variables used for k-means.*

K-means can be used as an exploratory technique to assess underlying patterns in the data. In this case, we have the labels for our data, and some theory and consideration of what we may expect to see from clustering. If this were to work well, we would hope to see different clusters with similar groupings for factors such as mental health, income and financial security. Our outcome variable is Q141, which means we also know the number of clusters to specify will be four.

An assumption of k-means is that the data can be divided into similar clusters. At this point, it may help to visualise the data, to see whether we already have obvious clusters. The scatter matrix in Figure 8, which has been jittered with an adjusted alpha score, can aid us. We can see there are no obvious clusters for the mental health scores, when considered alongside each feature. However, its hard to tell this just visually without performing k-means.

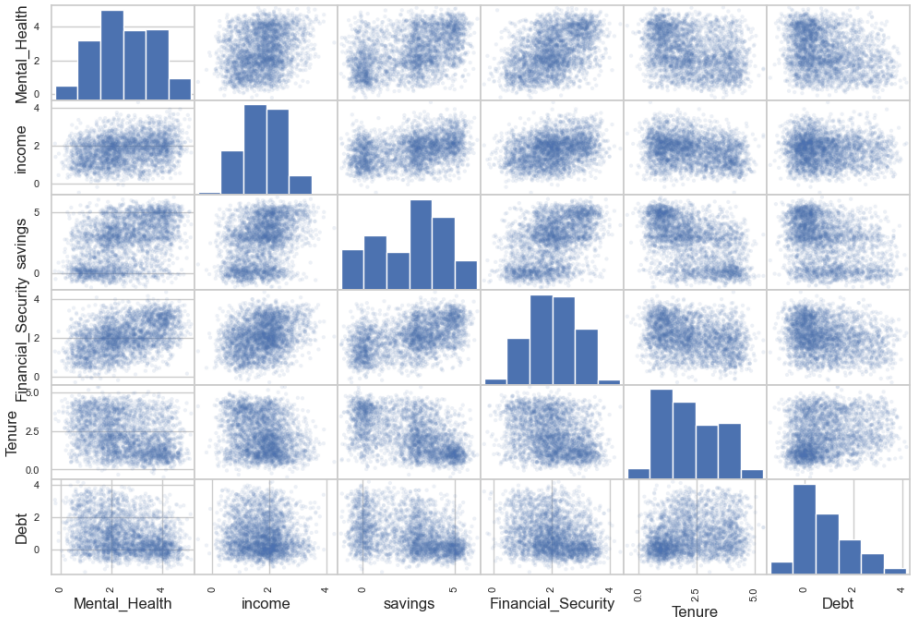

Figure 8: *Scatter matrix for each variable, showing the density and sparseness of data points for each variable included in the k-means, alongside any existing clusters.*

Table 6 shows results of the first k-means. For silhouette scores, a coefficient close to +1 means the instance is well inside its own cluster, and far away from other clusters, whilst close to 0 means it is close to the boundary between clusters. A completeness score close to 1 shows the extent to which all data points with the same label belong to the same cluster, whilst a homogeneity score close to 1 shows the extent to which all data points within one cluster, belong to the same true class. Considering the overall model, With a score of 0.23, the first k-means did not perform very well.

Silhouette Coefficient	0.23
Calinski harabasz Coefficient	1199.68
Completeness score	0.11
Homogeneity score	0.11

Table 6: *First K-means scores: low silhouette, completeness and homogeneity scores can be seen and should be improved.*

I next performed a Principal Component Analysis (PCA) and compared results. By performing the analysis on a lower dimensional data set, the silhouette score was improved (Table 7) and can be visualised in Figure 9.

Silhouette Coefficient	0.37
Calinski harabasz Coefficient	3154.425
Completeness score	0.12
Homogeneity score	0.121

Table 7: *K-means scores after peforming a PCA: a reasonable silhouette score can be seen, however the completeness and homogeneity scores should be improved.*

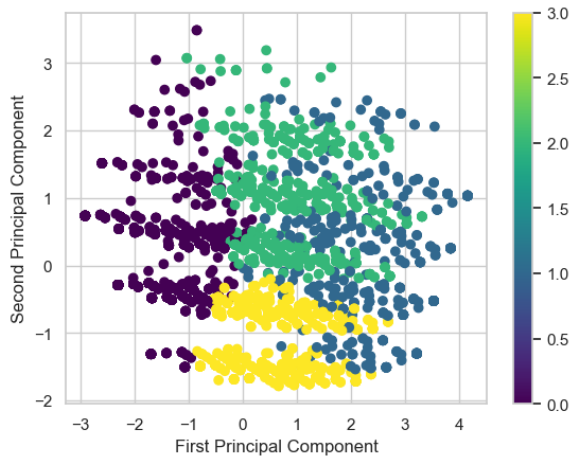

Figure 9: *Results of k-means after performing a PCA on the data. We can see the four clusters, and how some of these clusters are not spherical, and are overlapping one another, which is not ideal for k-means algorithms.*

However, we can see clusters are overlapping, and shape-wise, are not spherical. Figure 10 shows this in more detail; ideally each cluster should hold majority one colour, showing one target and a clear separation of true labels. We can see to some extent a color dominates each cluster, but not as much as would be expected, had the algorithm performed well.

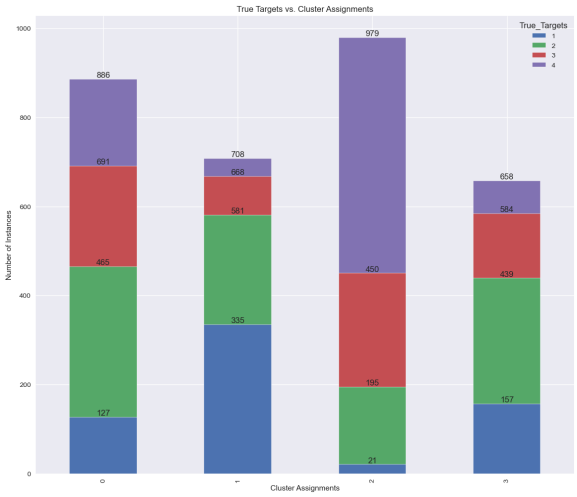

Figure 10: *True labels assigned to each clusters, where we hope to see results showing on color dominating each cluster box.*

For improving the model, we may consider feature selection, or a different number of clusters. We can see in Figure 11, that 2 clusters may yield the best results, before the model starts to loose inertia. As the target variable for k-means does not need to be continuous, we may change our target variable to a binary outcome, and consider results for two clusters.

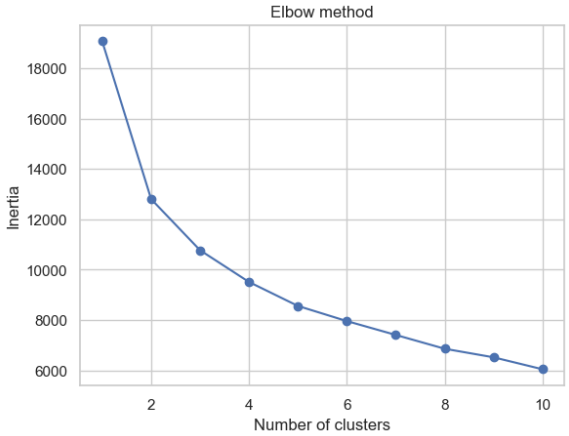

Figure 11: *Elbow method showing the optimal number of clusters before the model looses inertia. A decrease in the within cluster number of squares can be first seen at 2, indicating more clusters beyond this point may not improve the quality.*

I performed a new analysis looking for 2 clusters, and as Table 8 and Figure 12 show, the silhouette and completeness score improved, providing us with a reasonably good model. The completeness score has increased as the clusters will have merged smaller clusters into one; more data points in each cluster will therefore belong to the same class. However, this is at the expense of homogeneity. Whilst the silhouette score shows a degree of similarity of data points within each cluster, the true classes are overlapping within clusters.

Model	**Silhouette Score**	**Completeness Score**	**Homogeneity Score**
Model One	0.23	0.11	0.11
Model Two	0.37	0.12	0.12
Model Three	0.45	0.19	0.09

Table 8: *Clustering scores for three different models, showing improvements each time.*

Improvements include further feature selection. However, we may also consider that k-means may not be the best algorithm to use on this data. It is unfortunate that k-means can not also look at categorical variables, as I have had to exclude some of the binary categorical variables from the analysis, where we can see on the heat-map they have a medium correlation, which may have led to better results. Perhaps combining categorical and continuous data in a different algorithm will be best, as we can then consider some of those stronger differences in categorical groups considered

in the exploratory analysis, which may capture better results. For the continuous variables alone, perhaps there is not a clear enough pattern in the data to show strong enough clusters for the different levels of mental health scores.

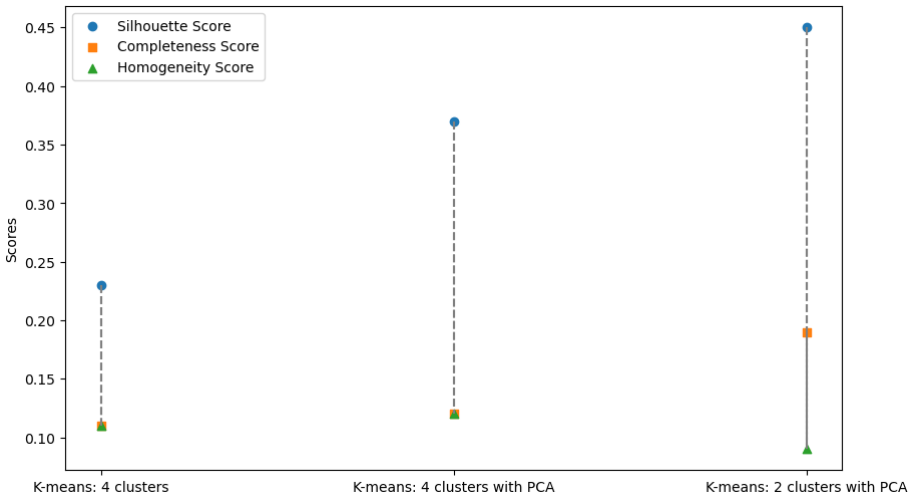

Figure 12: *Results of the three different k-means models, with adjustments made to each model to improve the final scoring. The best model is model three, where I performed the PCA reduction technique and specified only two clusters.*

4.2 Supervised learning

As the data includes categorical variables, there are two types of Supervised learning methods that may be appropriate; Logistic Regression or Decision Trees. Here, we will consider a binary decision tree, as not only is this a useful algorithm for classification tasks, but considering the split nodes can help us to visualise how decisions have been made, and perform feature selection.

The first decision tree provided reasonably accurate results, as seen in Table 10. This was with all of the variables in table 9 added into the decision tree, and these were based on the correlation matrix.

Variable Name	Variable topic	Questionnaire answers
Disability	Disabled or not	1 or 2
Employment	Employment type	Sliding scale
Q2Benefits	Benefits or not	1 or 2
HIDD10	Tenure	Sliding scale
Essentials	Cut back on essentials	1 or 2
Q1510	Used a warm bank	1 or 2
Q12	Seek help with bills	Sliding scale
Q8W22	Belief in Government	Sliding scale
Income	Income level OECD	1 -3
Savings	Savings level	Sliding scale
Debtcat	Debt level	Sliding scale
security010	How financially secure you feel	Sliding scale

Table 9: Variables included in the first decision tree

Class	Precision	Recall	F1-Score	Support
1.0	0.70	0.68	0.69	466
2.0	0.72	0.74	0.73	523
Accuracy			0.71	989

Table 10: *Decision Tree classification report showing an accuracy of 71 percent. This is a reasonable model, but this can be improved.*

Decision Trees are considered *white box models*, as they allow us to check the calculations preformed in making predictions. Figure 15 shows the features considered by the decision tree at each node, which helps us to understand the most important factors for predicting mental health scores, and feature select accordingly. These 5 features were used in the final model; I performed the analysis a few times, starting with a greater selection of features, and reducing these until I had optimal results. The results did not change a huge amount, but the best results, after feature selection and nodes, can be seen in Table 11, an improvement from 71 to 76. The max depth of the tree was 8.

Class	Precision	Recall	F1-Score	Support
1.0	0.76	0.71	0.74	479
2.0	0.76	0.80	0.78	532
Accuracy			0.76	1011

Table 11: *Decision Tree classification report showing an accuracy of 76 percent. This is a reasonable model, and is the result of feature selection of the five most important features.*

The confusion matrix in Figure 13 shows the true positives, false positives, true negatives, and false negatives for the decision tree.

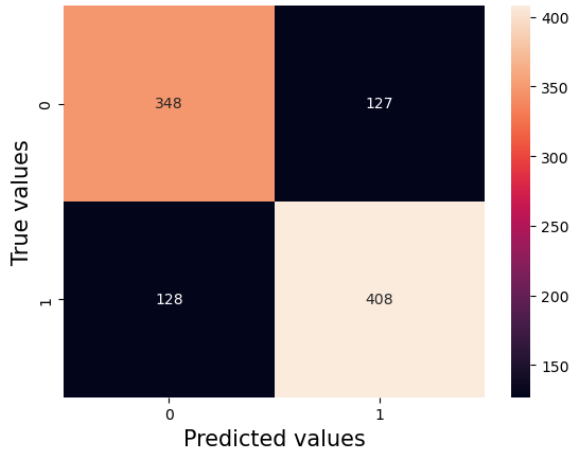

Figure 13: *Decision Tree Confusion Matrix: 348 and 408 represent the true positives and true negatives, and 127 and 128 represent the false positives and false negatives.*

Figure 14 shows the results of a binary decision tree, with 5 nodes and with two children; mental health has been affected, or has not been affected. The first tree was overly complex, so I limited the number of nodes to give a clearer representation (this tree shows 5, but the best result was with 8). As we can see the majority of those who have been negatively impacted are on the left, and the purer nodes are in dark blue, whilst purer nodes on the right are mainly orange, and includes those who have not been negatively impacted. The gini coefficients give us clearer understanding of where the training has performed well (gini = 0 if it is pure) and where the nodes are less pure, they contain instances from different classes.

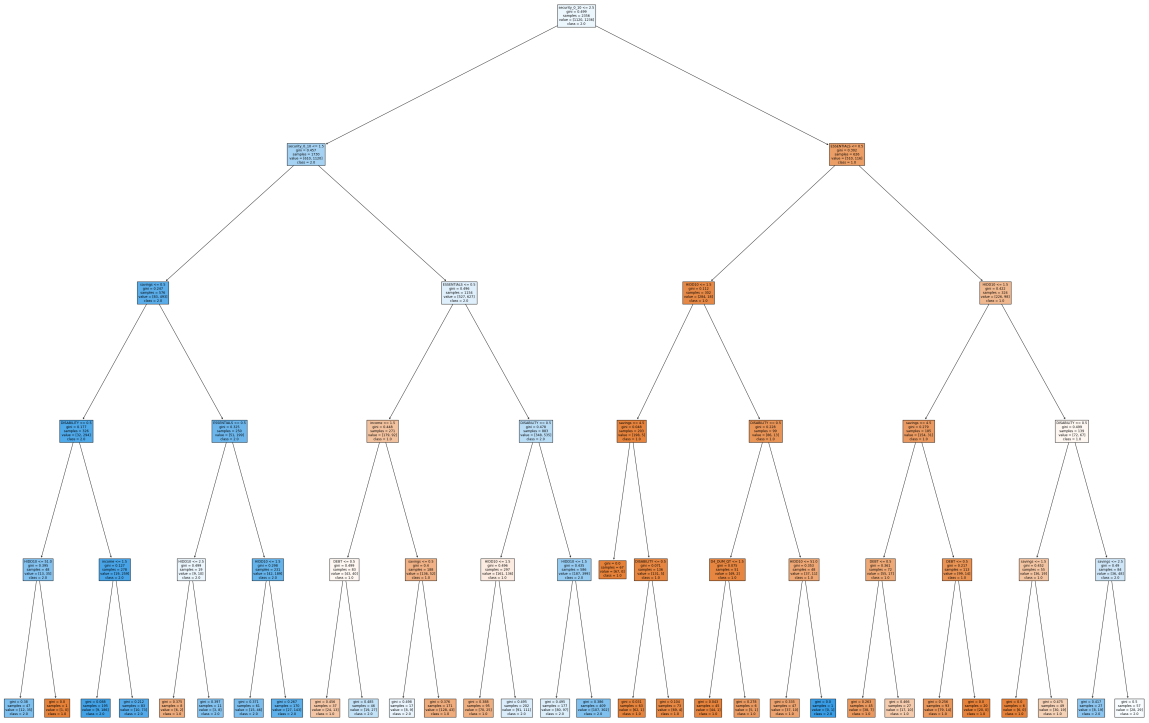

Figure 14: *Decision tree with two children; mental health has been affected or has not been affected by the cost of living crisis - this shows max depth cut at 5, so we can visualise the detail of the tree. The best model had a max depth of 8.*

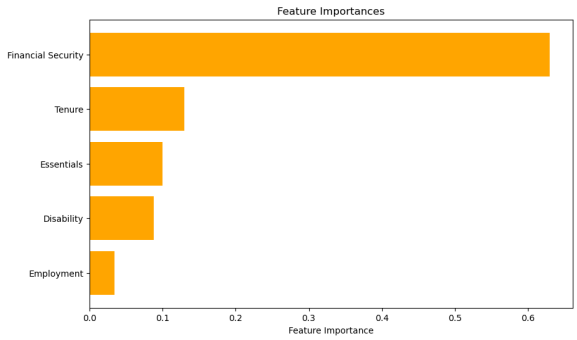

Figure 15: *Features selected for the final model with the highest accuracy scoring. The best features for predicting mental health scores are financial security, tenure type, cutting back on essentials, disability and employment type.*

Finally, I checked the model for over-fitting by performing a cross-validation technique. Over-fitting occurs when the model performs well on the training set, but poorly on the test set. If the model picks up small nuances in training data, it will then not be able to make more generalised assumptions with the test data, leasing to a poor result for the test data. The test set accuracy here is 75, which is the same score for the training set. Furthermore, having performed a cross validation, with a mean score of 74 across the 5 folds, we can see the model performed well on each test set and therefore, over-fitting is probably not a concern here.

5 Conclusion

As we can see from the initial exploratory analysis, there appears clear associations between a persons life satisfaction scores and their socioeconomic background, beliefs, behaviours and demographics. Effectively clustering groups can be beneficial to understand the underlying patterns within the data. However, overall, k-means as an unsupervised learning method did not perform particularly well here, perhaps due to an absence of a clear underlying structure within the data set that I expected to see - an alternative mixed model with categorical data, or better feature selection of continuous data, may yield different results. However, as a supervised method, a decision tree provided reasonably accurate results and we can see that how financially secure somebody feels may be the biggest predictor of mental health post COLC. This analysis may be considered in a

policy context, namely, should policy considerations be based not just on raising income levels, but on creating a financial safety net, so concerns of financial insecurity will be reduced. However, the decision tree should ideally be improved in accuracy to reduce its current type two errors; incorrectly classifying somebody as not being NI, when they in fact will be NI, will have consequences on who will receive the right support.

6 Bibliography

Bambra, C., et al. (2021). Pale Rider: Pandemic Inequalities. In The Unequal Pandemic: COVID-19 and Health Inequalities (1st ed., pp. 13–34). Bristol University Press.

Bambra, C., et al. (2021). Introduction: Perfect Storm. In The Unequal Pandemic: COVID-19 and Health Inequalities (1st ed., pp. 1–12). Bristol University Press.

Princeton University. (2009). Financial Security: More Money Alone May Be Key to Happiness, Princeton Study Says. Retrieved from https://www.princeton.edu/news/2009/03/17/financial-security-more-money-alone-may-be-key-happiness-princeton-study-says

Miron-Shatz, T. (2009). "Am I going to be happy and financially stable?": How American women feel when they think about financial security. Judgment and Decision Making, 4(1), 102-112. doi:10.1017/S1930297500000747

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.