



BritishRedCross



UNDERSTANDING VULNERABILITY SURVEY

Help those in need through analytics

TEAM 17: pizza_beers_reindeer.py

Marco CORTESE
Hedda TORKELSEN
Hadrien VENANCE
Lorenzo VOTTA

TABLE OF CONTENT

- 1 Preface
- 2 Background Theory for Business Application
 - Key Domains of Vulnerability
 - Existing Literature on Vulnerability
 - Creation of Indexes
 - Disaster Risk Management
- 3 Strategy
 - Business Questions
 - Methodology to Answer

- 4 CRISP-DM
 - Data Understanding
 - Data Preparation
 - Modeling
 - Robustness of the Model
 - Evaluation
- 5 Business Advices
- 6 Appendices
- 7 References

PREFACE



VULNERABILITY INDEXES TO SHAPE BRITISH RED CROSS'S STRATEGY

Motivation for Vulnerability Indexes

Vulnerability indices are an important tool for **identifying the most vulnerable areas in specific regions**.

They provide a set of relative measures of vulnerability for small geographical areas across the UK, based on different domains of vulnerability.

Objective

Understanding vulnerabilities is essential for the purpose of this work, i.e., help the British Red Cross in **identifying geographical trends** to **help resource allocation**, and broader national trends to **help future planning and service delivery**.

There is reason to believe that the **pandemic** has further increased the **importance of inclusive social protection and health care system**, given that more and more people are experiencing mental health challenges as a result of the quarantine and the worsened economic conditions.

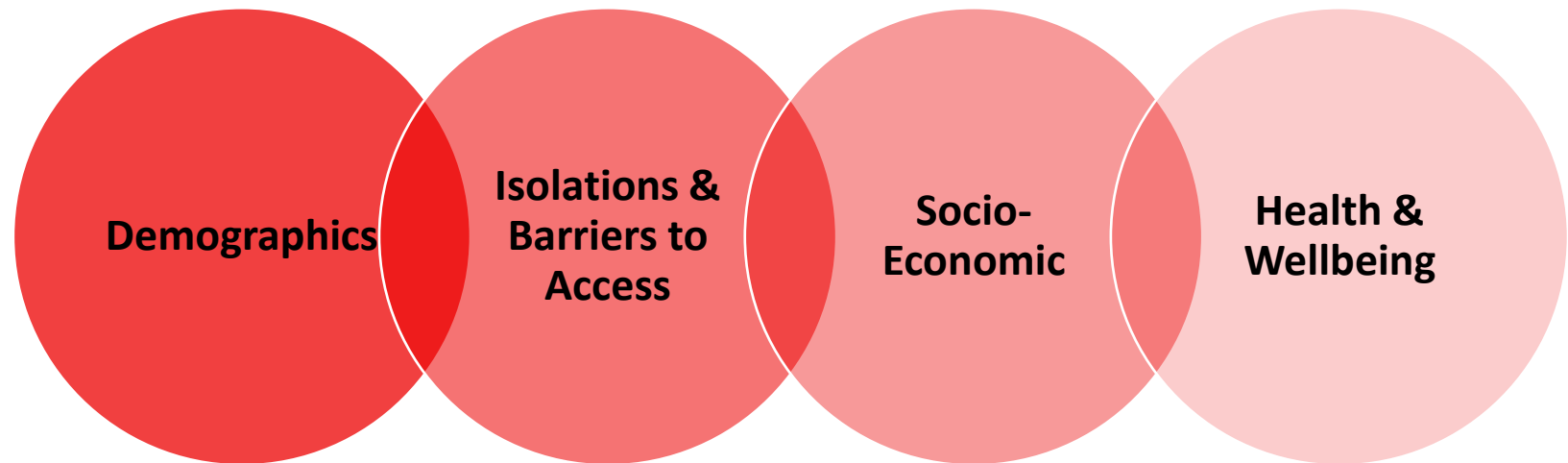
BACKGROUND THEORY FOR BUSINESS APPLICATION

- Key Domains of Vulnerability
- Existing Literature on Vulnerability
- Creation of Indexes
- Disaster Risk Management

THERE ARE DISTINCT DOMAINS WHERE VULNERABILITY ARISES

Each of the domains of vulnerability is based on a **basket of indicators**, each derived from **two waves** of data provided by the British Red Cross (BRC), dated July 2021 and January 2022, respectively.

It is important to understand that vulnerability is **multi-dimensional** and can manifest in relation to a number of distinct domains:



Multiple vulnerability is measured at an **area level** by combining these domains.

LITERATURE HIGHLIGHTS DIFFERENT DOMAINS FOR VULNERABILITY

Demographics

Age (over-70s), insecure immigration status, refugees, asylum seekers, victims of crime (including domestic abuse & trafficking), children and young people.

Isolation & Barriers to Access

Physical isolation, social isolation, digital isolation, geographical isolation, language barriers, barriers accessing services due to immigration status.

Socio-Economic

Income deprivation, economic vulnerability, rough sleepers, homeless, unsuitable accommodation or living environment (overcrowded, temporary, emergency or supported), working age benefits, pension.

Health and Wellbeing

Multiple health conditions, health inequalities, addiction, poor mobility, mental health and wellbeing, loneliness, pregnancy.

DISASTER RISK MANAGEMENT IS A TOOL THAT WILL HELP THE MOST AT-RISK PEOPLE



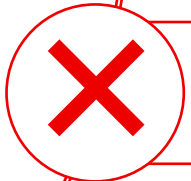
The **identification of vulnerable groups**, such as the elderly, children and the mentally and/or physically impaired, in the case of hazards or when a crisis unfolds, is an issue that any crisis and **disaster risk management** should address, since people are affected unequally.



A vulnerable group can be defined as a 'population within a country that has specific characteristics that make it at a higher risk of needing humanitarian assistance than others or being excluded from financial and social services.'



In a crisis such groups would need **extra assistance**, which appeals for additional measures.



In general, countries and international organizations define vulnerable groups according to pre-crisis social, economic and cultural factors, which usually perpetuate inequality, exclusion, and lack of access to resources.



These factors are very much ingrained in our societal systems and function **as multipliers of marginalisation**, which can become especially apparent in crises and disasters.

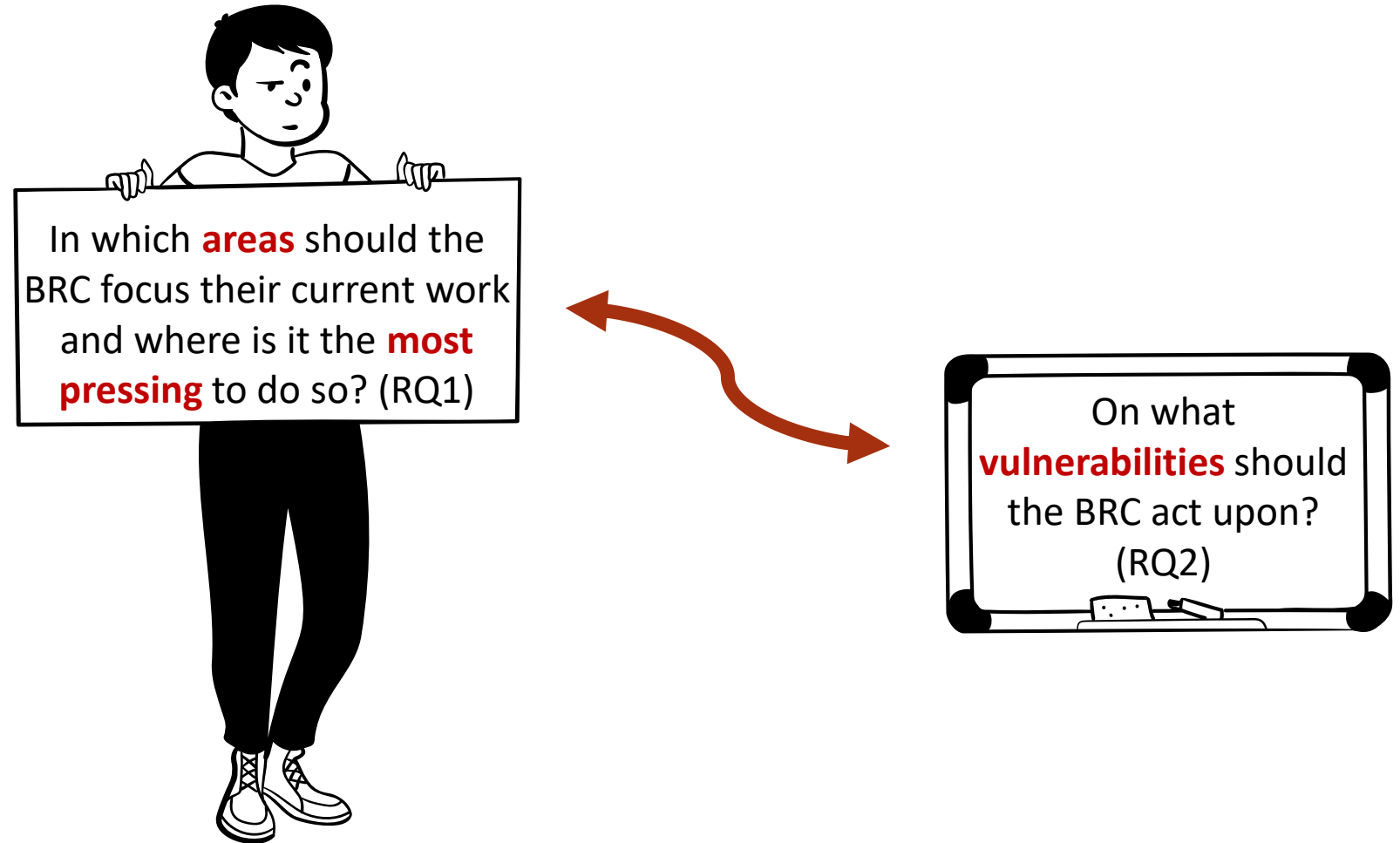
STRATEGY*

- Business Questions
- Methodology to Answer

*based on the strategy used in the article of M.Seyedan, et al. 2022

Spring 2022

TWO MAIN RESEARCH QUESTIONS HAVE TO BE ANSWERED



METHODOLOGY USED TO ANSWER THE 2 BUSINESS QUESTIONS



Our main focus was to conduct a **cross-sectional** analysis in order to identify the main **current geographical trends** among survey respondents.



This study aims at finding some **longitudinal** differences utilizing statistical techniques (such as t-tests) to ensure that the British Red Cross has targeted those areas with the most pressing demand.

1

Creation of 3 Vulnerability Indexes.

2

Clustering on these Indexes & Geographical Visualizations (by Nearest Cities).

3

Training of Decision Trees on each Index to allow the BRC to implement these ML Models on new unseen cases.

4

Based on T-tests, means of the input variables of the ML models are compared and Business Recommendations stem from the results of the mean comparison.

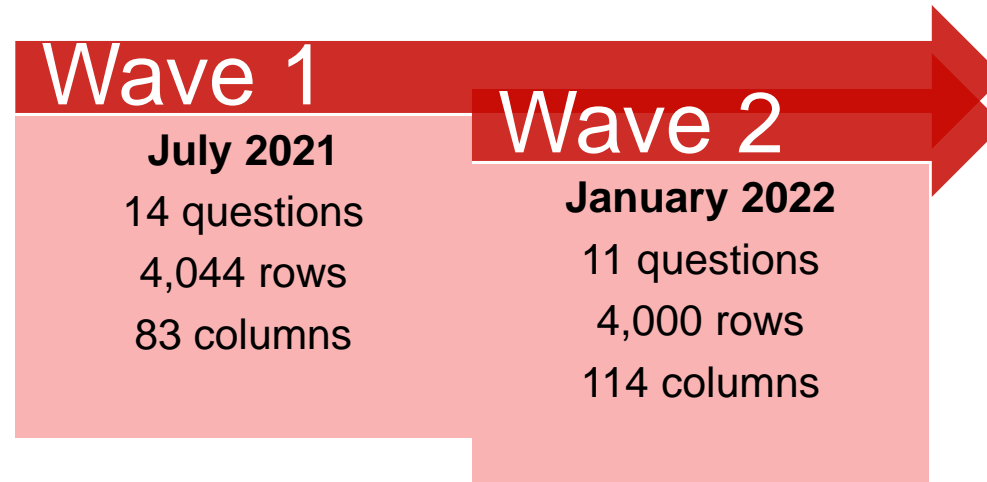
CRISP-DM*

- Data Understanding
- Data Preparation
- Modeling
- Robustness
- Evaluation

*Cross-Industry Standard Process
for Data Mining

Spring 2022

DATA UNDERSTANDING



Binary variables

- Values are either 1 or 0. As we have survey data, most of the records are either “Yes” or “No” answers to specific questions and we translated them to 1 or 0 for modelling purposes.

Discrete variables

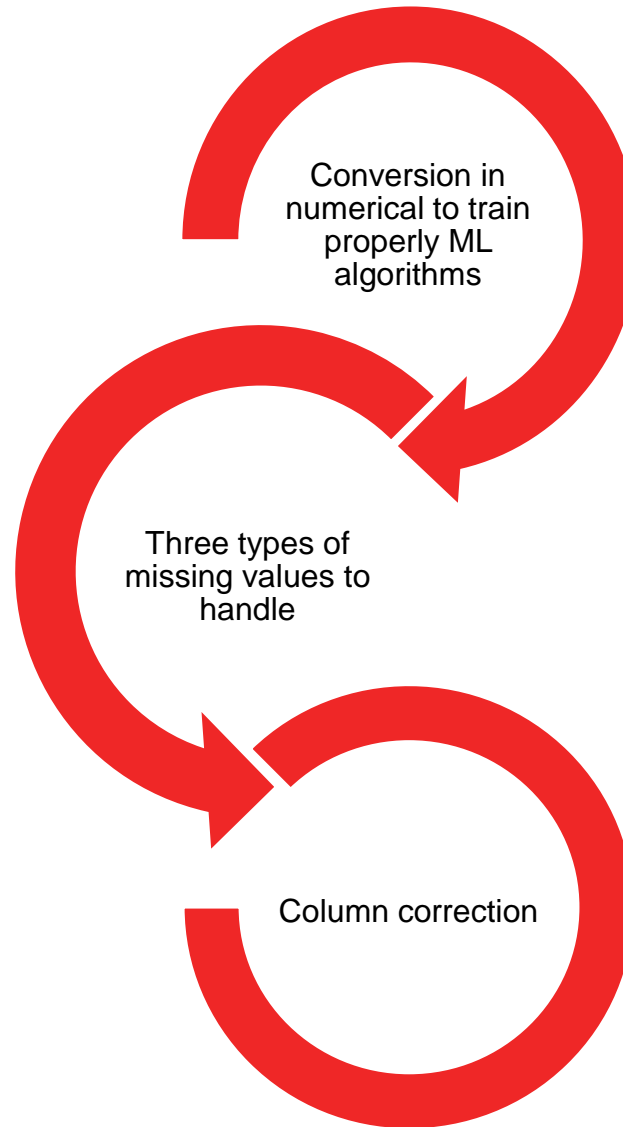
- The variable can take on a specific set of numeric variables. Mostly in the way of scaled answers.

Categorical variables

- Every respondent belongs to one specific category. It accounts for most of the demographic variables.

DATA PREPARATION TO ENABLE OUR MODEL TO BE PROPERLY TRAINED

1. **NA answers to multiple choice questions:** Important to keep it as it is informative (the respondent did not select this answer). For the case where respondent did not tick any of the possible choices, we used an own-built function PreferNotToSay that created an additional column.
2. **NA in 'Region' feature:** Value of the region could be inferred from the "nearest_city" feature.
3. **NA in 'Gender' feature:** If NAs are removed, the datasets were quite well balanced (i.e. 56% of Male and 44% of women in wave 1). So it was decided that this feature was not of the uppermost importance.



For questions with **Boolean answers** (yes/no or chosen/not-chosen), values were transformed in **0/1**.

For questions with scale of **non-numerical values**, those were translated into **ordinal ones** (i.e. from 1 to 5 or from 1 to 10).

To have a **clearer understanding** of each feature, it has been decided to give a defined **naming system for each feature**. E.g. Q3_multi_welfarefund.


VULNERABILITY WAS HANDLED BY CREATING INDEXES

We decided to create **three different indexes** related to the three different types of vulnerability:

- **Socioeconomic Index**
- **Physical Index**
- **Mental Index**

To do so, we proceeded as follows:

We created variables from the questions and/or answers from the Understanding Vulnerabilities Survey provided by the BRC.



Then, the different variables have been aggregated for each index according to what the survey question was trying to explain.

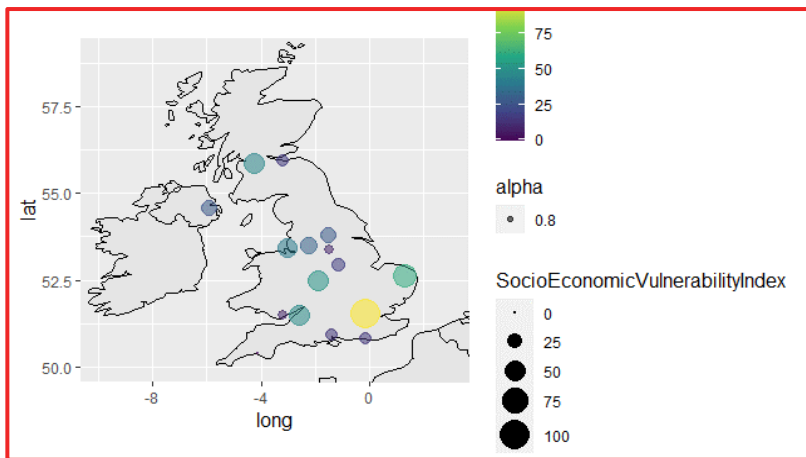
for further details click [here](#)

FEATURE ENGINEERING TO GET A HIGH-LEVEL VIEW OF VULNERABILTIES AMONG RESPONDENTS

Feature engineering

To have a more **high-level view** of the vulnerabilities faced by respondents and make the most out of the use of ML models, it was decided to create three different indexes related to the three different types of vulnerability the BRC is trying to handle: **socio-economic, mental and physical health vulnerabilities** (for further details click [here](#)). The 3 indexes were built on variables chosen according to health business knowledge and availability of information.

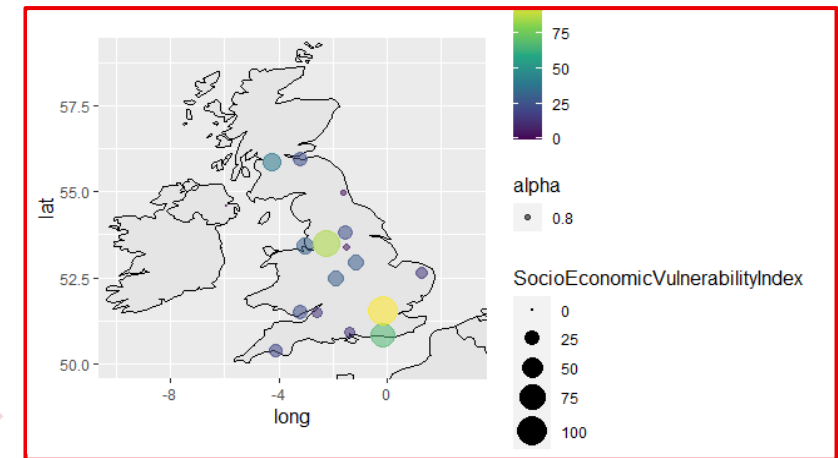
Socio-economic Index (Wave 1)



Insights:

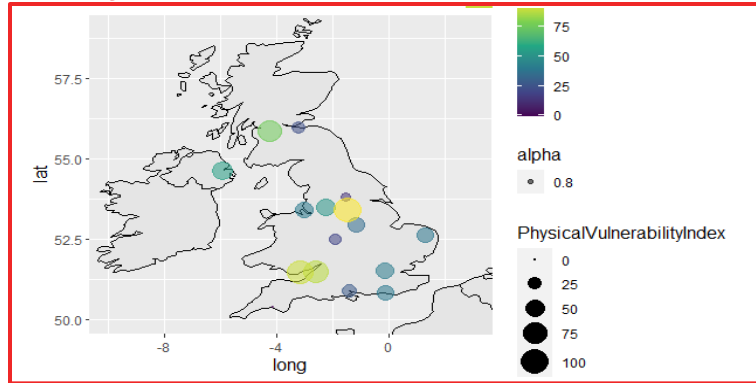
- **Big cities** seem to have a higher score than others. It could be explained by the fact that there is more unemployment.

Socio-economic Index (Wave 2)



TRENDS ARE DIFFERENT BETWEEN THE 2 WAVES FOR PHYSICAL AND MENTAL INDEXES

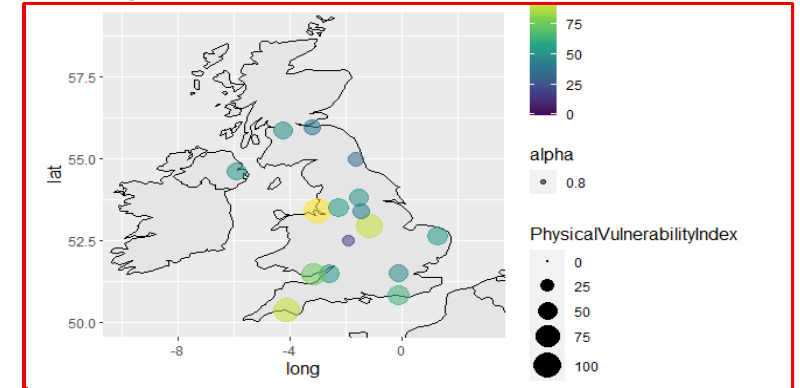
Physical Index (Wave 1)



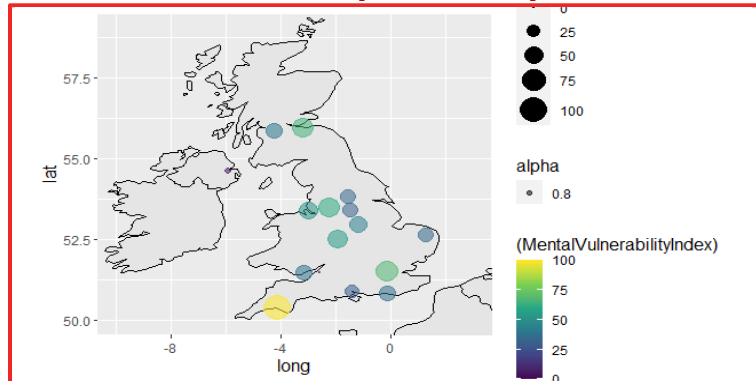
Insights:

- **Increase** in wave 2 for the more **southern cities**.
- **Norwich** stands out from the other cities.

Physical Index (Wave 2)



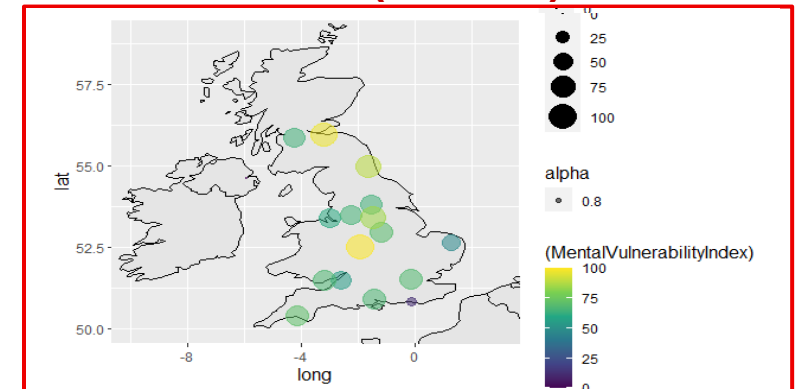
Mental Index (Wave 1)



Insights:

- Most cities have a **higher average** in wave 2.
- **Newcastle** is the most-at-risk in wave 1 and still scores bad after.

Mental Index (Wave 2)

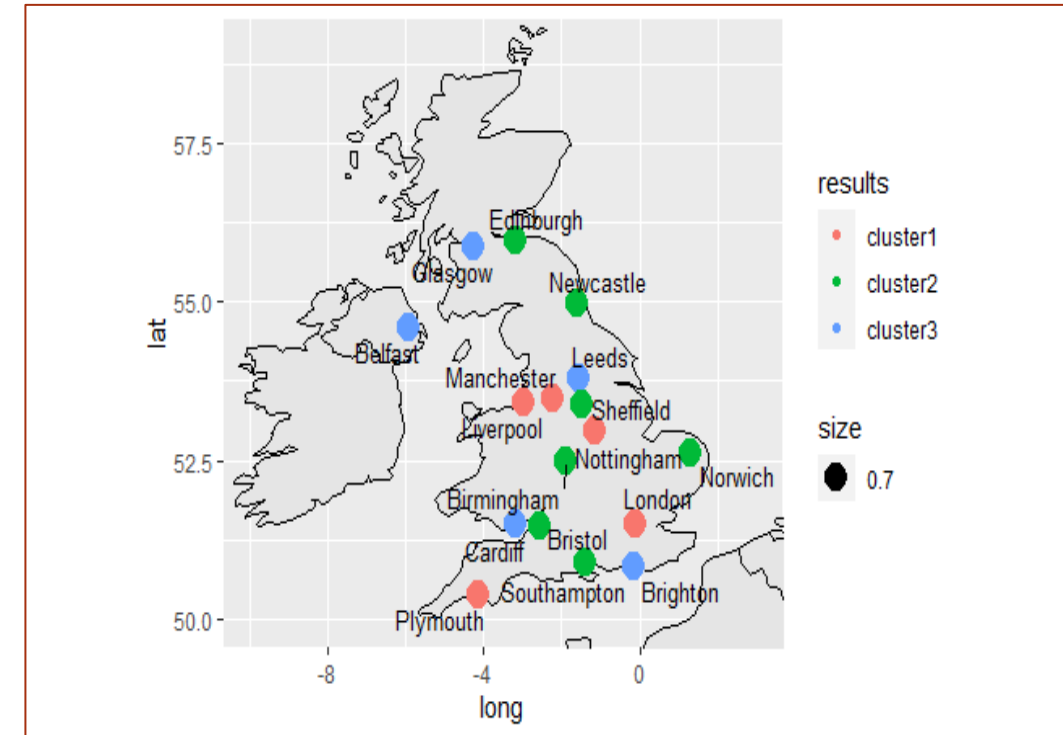


CLUSTERING ENABLES TO FIND THE MOST URGENT AREAS WHERE ACTION IS NEEDED

Unsupervised ML: Clustering

To answer our second question on **where it is more critical** to act for the BRC, it was decided to cluster the respondents of the 2nd wave. The clustering was motivated by regressions for Spatial Inference (see Appendix). The clustering method used is **Kmeans on the 3 previously build Vulnerability Indexes**.

1. Elbow method and gap statistics advise to go for 3 clusters (see Appendix).
2. After training the model, 3 clusters of interest. **Cluster 1** encompasses respondents with high socio-economic and physical issues; **Cluster 2** is made of people with only high mental vulnerability and **Cluster 3** includes people with relatively low vulnerability overall.
3. To make it useful for the BRC, it was decided to take the most important cluster for each city by majority voting (i.e. regarding the population in each city, what is the most important cluster).



for further details click [here](#).

FEATURE SELECTION AND REGULARISATION TO REDUCE MODEL COMPLEXITY AND PREVENT OVERFITTING

It was necessary to decide on which features would be fed the ML models' inputs. The first issue was to **avoid multicollinearity**. It was done **removing the variables directly correlated** with the indexes.

After taking care of the multicollinearity issue, a lot of variables were still available. Therefore, **feature selection** techniques were used:

- ✓ **Feature Selection**

- ✓ **High Correlation Filter:** High correlation among features means they are likely to carry similar information, which can bring down the performance of our models. We implemented a threshold of 0.6, i.e., if correlation coefficient crosses this number, we drop one of the correlated feature.

Then, two **regularisation** techniques were used to reduce model complexity and prevent over-fitting which may result from simple linear regression:

- ✓ **Lasso Regularisation:** In lasso regression, the cost function is altered by adding a penalty equivalent to the (absolute value) magnitude of the coefficients. It performs feature selection by moving some coefficients to 0. Therefore, it is a good choice when there are many features.


- ✓ **Ridge Regularisation:** In ridge regression, the cost function is altered by adding a penalty equivalent to the square of the magnitude of the coefficients. It shrinks coefficients and works well in presence of highly correlated features.

REGRESSIONS GRASP THE RELEVANT VARIABLES FOR A CLASSIFICATION MODEL

Supervised ML: Standard Regressions

The motivation for the use of regressions underlies **in choosing only the most relevant variables** and in making a **classification model as understandable as possible**.

According to GridSearchCV lasso regressions were the most efficient for all the indexes. Thus, **lasso regressions** were trained for each of the indexes. It allowed to **retrieve some variables** thanks to its embedded **shrinkage coefficient** method.



<u>Indexes</u>	Variable Names
Socio-Economic	'Q1_multi_DomesticAbuse', 'Q1_multi_AlcoholDrug', 'Q10_MentalHealthSupportWaitingPreferNotToSay', 'Q2_multi_MigrantOrga', 'Age', 'Q2_multi_AdviceService', 'Q11_LonelyPreferNotToSay', 'NumberChildren', 'Q7_NoPleasure'
Physical	'Q4_MobilityHealth', 'Q7_NoPleasure', 'Q1_multi_AlcoholDrug', 'EmploymentOther not working', 'Q2_multi_DayCentre', 'Q4_MobilityHealthPreferNotToSay', 'Q1_multi_Homeless', 'Q5_UsualActiHealthPreferNotToSay', 'Q6_Happy', 'SocialGradeUnemployed for over 6 months or not working due to long term sickness'
Mental Index	'Q6_Happy', 'Q6_LifePurpose', 'Q8_DepressedPreferNotToSay', 'Q7_NoPleasure', 'NoChildren', 'Q1_multi_DomesticAbuse', 'Q2_multi_MigrantOrga', 'Q9_MentalHealthSupportPreferNotToSay', 'SocialGradeHigher managerial/ professional/ administrative (e.g. established doctor, etc.)', 'Q10_MentalHealthSupportWaitingPreferNotToSay', 'Q2_singleBlack, African, Caribbean or Black British'

ONE DECISION TREE MODEL FOR EACH OF OUR CLUSTERS

Supervised ML: Decision Tree

Methodology

- The decision tree aims to form groups of respondents with high vulnerability for one index against respondents not part of this most vulnerable people, minimizing grouping errors.
- It was decided to implement 3 decision trees as they all rely upon a different set of explanatory variables.

Response Variable

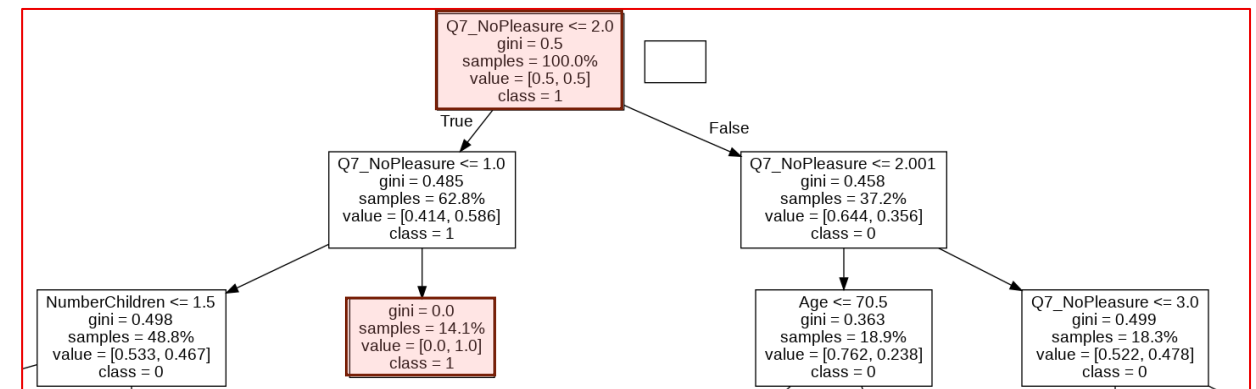
- 0 if not part of the group
- 1 if part of the vulnerable group

Insights:

- In this decision tree, the most important variable to discriminate between belonging to cluster 1 and not is the answer to the **question 'Having no interest in doing things'**.
- People answering that they did not experience it seems to belong to our cluster; i.e. **low mental 'breakdown'** for these people **with high economical and physical vulnerabilities**.
- Having a high **number of children (2 or more)** also seems to be the source of **economical and physical vulnerabilities**.

Decision Tree to classify the 1st cluster

(only the 3rd nodes, the rest of the decision tree can be found in our code)



Red box means significant sample (high % size, relatively low gini/entropy and majority class = 1)

ROBUSTNESS OF THE DECISION TREE MODEL

A rigorous testing protocol was used to make sure that the results are **reproducible** in the future and that the model will **hold on unseen cases**.

Overfitting is a big problem in data science, and it is of the uttermost importance to avoid it. Overfitting means that a model performs well on the training set but poorly on unseen cases.

Techniques



Test set: a portion of the data was set apart to test the model at the end (on unseen cases)



When **calibrating** the models (try to find the best hyper-parameters), **k-fold cross-validation** was used. This methods divide the training set into k different subsets. While training the model, it trains it on k-1 fold and test it on the remaining one. It is repeated until each fold is used on the test set.



An **analysis of the variance** of the predictions on every fold was conducted, making sure that it was as low as possible.

METRICS TO EVALUATE THE DECISION TREE MODEL

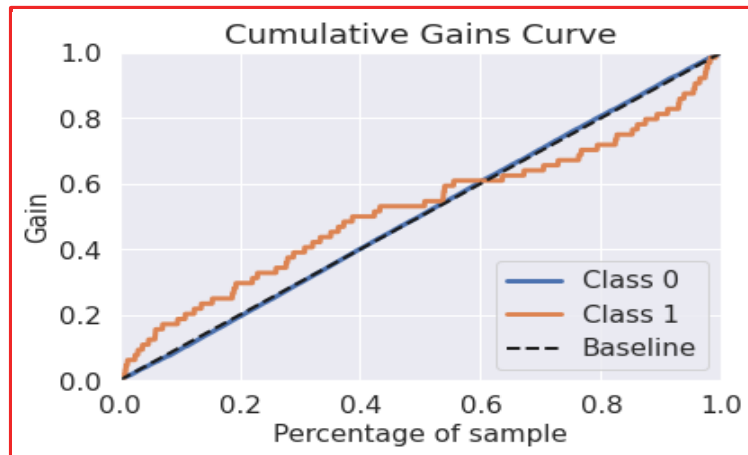
Measures	Cluster 1	Cluster 2	Cluster 3
Cross-validation variance	0.00104	0.000524	0.000629
Accuracy	84.41%	59.75%	82.08%
AUC (on the test set)	0.52	0.53	0.58
Recall	0.22	0.45	0.22
Precision	0.09	0.04	0.03

It is worth noting that all cross-validation variances are really small, and it shows that these models **prevent some overfitting**. The AUC score does not seem to be really high, but it can be explained as the dataset is **highly unbalanced** in terms of 1 cluster against the rest of data.

CUMULATIVE GAIN CHARTS TO EVALUATE THE DECISION TREE MODEL

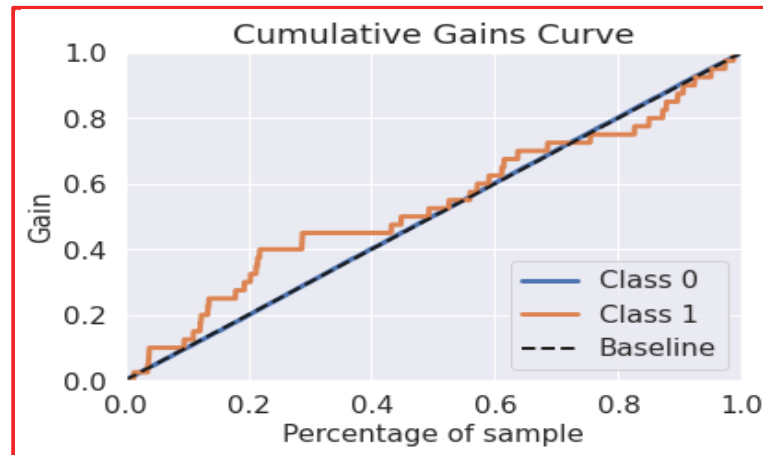
A cumulative gain chart shows the **percentage of cases in each category by targeting a percentage of the total number of cases**. It plots the *tp rate* (i.e., % of targets correctly classified) as a function of the % of the total population.

1st Cluster Decision Tree



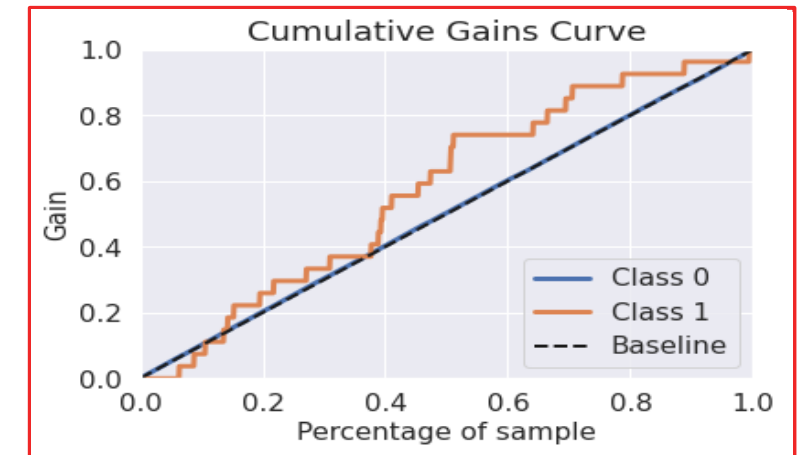
Above we see that by targeting 20% of the population, we are able to retrieve a bit more than 35% of positive targeted cities.

2nd Cluster Decision Tree



Above we see that by targeting 30% of the population, we are able to retrieve a bit more than 50% of positive targeted cities.

3rd Cluster Decision Tree



Above we see that by targeting 50% of the population, we are able to retrieve more 75% of positive targeted cities.

T-TESTING HAS BEEN USED TO ASSESS IF THERE ARE SIGNIFICANT INCREASES IN INDEXES AMONG THE TWO WAVES

To understand if there are **significant differences** in the three indexes among the two waves, it was used hypothesis testing. These tests have also been used to **forecast** what we can expect in the future. To understand which are the **drivers that have had the biggest impact on these changes** we t-tested also the most important variables for each index following the results of the regressions.

Analyzing the results, we focused on two different types of cities:



- **High risk cities:**
 - cities with high value of the index.



- **On trend cities:**
 - cities that are at risk and present a significant change in the value of the index between the two waves.

Hypothesis:

- $H_0: \text{mean}(\text{wave_1}) - \text{mean}(\text{wave_2}) = 0$
- $H_1: \text{mean}(\text{wave_1}) - \text{mean}(\text{wave_2}) \neq 0$

It has been decided to use two-sided t-test to be able to assess both increases and decreases. Significance level taken in all the tests is 0.05.

VARIABLES DRIVING THE CHANGES IN CLUSTER 1

Cities belonging to the cluster 1 are displayed in red on the map.

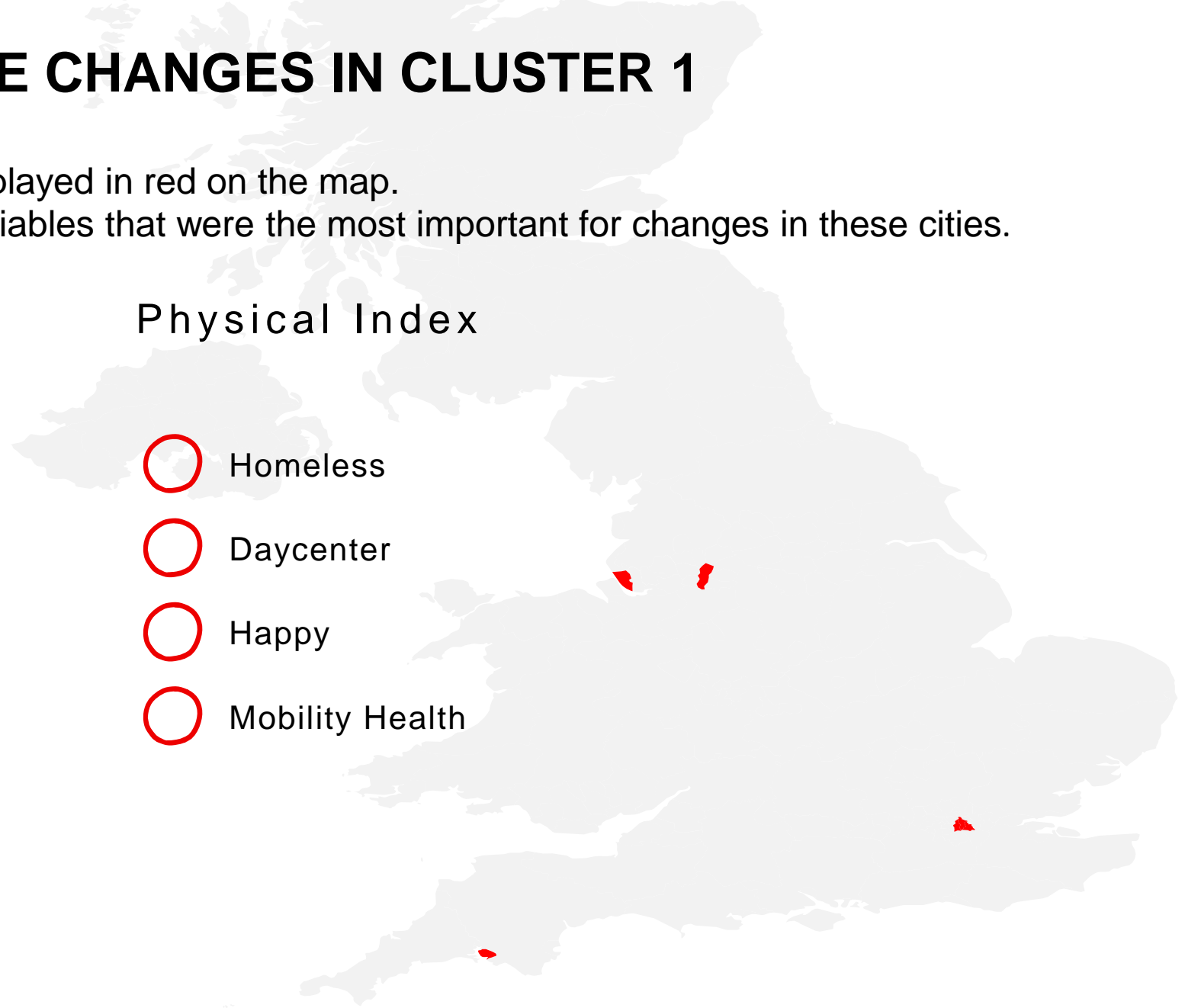
Between the 2 waves, these are the variables that were the most important for changes in these cities.

Socio Economic Index

- ☐ Coming UK
- ☐ None
- ☐ Migrant Organisation
- ☐ Life Purpose
- ☐ Alchohol Drug

Physical Index




- ☐ Homeless
- ☐ Daycenter
- ☐ Happy
- ☐ Mobility Health



MANCHESTER IS THE MOST AT RISK CITY FOR THE SOCIO ECONOMIC INDEX, AND IS WORSENING

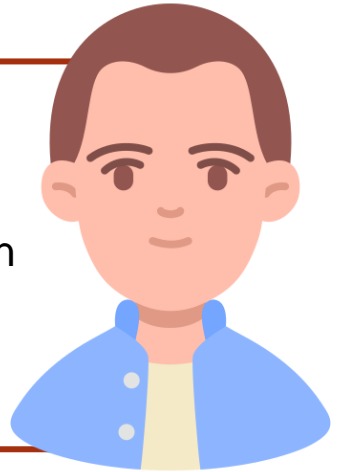
Manchester's socio economic index is worsening. We obtain a t-value of -2,14 when testing the change between the waves, which is statistically significant on a 96,7 significance level.

Drivers of these changes are:

-  Coming UK
-  None
-  Migrant Organisation

The typical person doing worse than the average:

- Male
- Tends to be young
- Works full time in a intermediate managerial position
- Has no kids
- Lives in a suburban area
- Is white



For the “Coming UK” variable we have a negative t-value (-1,50). So does the variable “Migrant Organisation” with t-value (-2,13), which implies that more individuals in wave 2 experience troubles with coming to the UK and with migrant organisations. These changes are statistically significant. The variable “None” has a t-value of 2,82, meaning that far less individuals experience none of issues BRC listed in question 1 of the two waves. This result shows that **overall the population is experiencing more socio economic issues in wave 2, compared to wave 1**. As we see, the individuals experiencing such issues, tend to be **young**.

LONDON IS ONE OF THE MOST AT RISK CITIES FOR THE PHYSICAL INDEX, BUT HAS A POSITIVE TREND

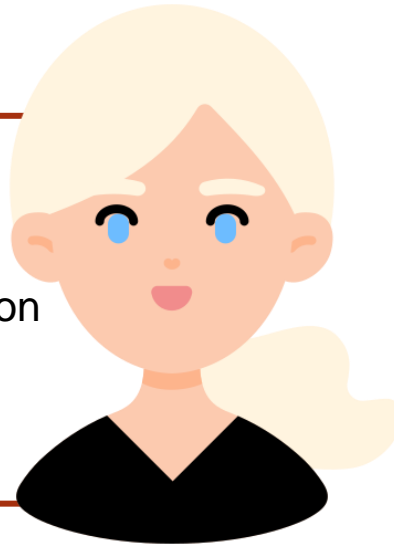
London's physical index is improving. Testing the change between the waves, gives us a t-value of 4,24, statistically significant on a 100 percentage significance level.

Drivers of these changes are:

-  Homeless
-  Daycenter
-  Happy

The typical person doing worse than the average:

- Unknown gender
- Is between 30 and 45 or above 65
- Works full time in a intermediate managerial position
- Has 1 kid
- Lives in an urban area
- Is white



For the “Homeless” variable we have a positive t-value of 3,01, which is significant on a 97,4 level. We see the same trend for the variables “Day Center” and “Happy”. Their t-values are 4,56 and 1,92 with respective significance levels of 100 and 94,5. Overall, we have **a highly positive trend for London when evaluating the physical index**. However, one should pay attention to the **typical person doing worse than the average**, as this part of the population could be the ones dragging the "Happy" variable down.

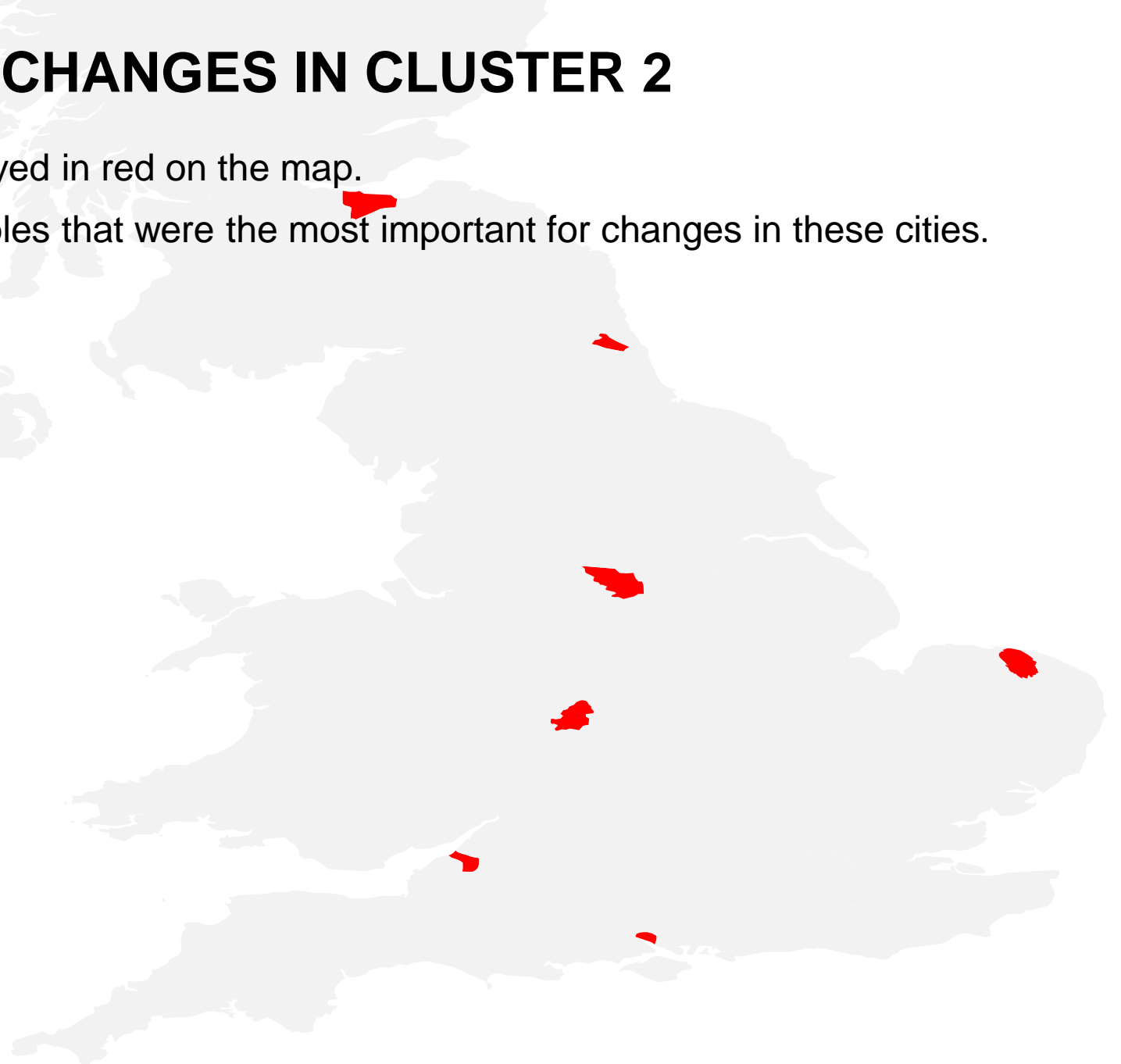
VARIABLES DRIVING THE CHANGES IN CLUSTER 2

Cities belonging to the cluster 2 are displayed in red on the map.

Between the 2 waves, these are the variables that were the most important for changes in these cities.

Mental Index

- Life Purpose
- No Pleasure
- Domestic Abuse



BIRMINGHAM IS THE MOST AT RISK CITY ON THE MENTAL INDEX, AND CONTINUES TO WORSEN

Birmingham's mental health index is worsening. Comparing the two waves gives us a t-value of -1,83, which is statistically significant on a 93,1 percentage level.

Drivers of this change are:

- ✓ Life Purpose
- ✓ No Pleasure

The typical person doing worse than the average:

- Male
- Between 50 – 75 years
- Works full time in a intermediate managerial position
- Has 1 kid
- Lives in a suburban area
- Is white



For the variable “Life Purpose” we have a negative t-value (-2,58), which implies **individuals in wave 2 feel like life has less purpose than before**. On the other hand, the “No Pleasure” variable has a t-value of 3,21, which is statistically significant on the 99,9 percentage level - implying that individuals feel like life is more pleasurable than before. These variables are the main drivers on **Birmingham's worsening mental index**. Reasons could be that a long pandemic has taken its toll on the population, but also the fact that **most of the individuals doing worse than the average are older people, who in fact has been greater affected by the pandemic**.

BUSINESS ADVICES

- Capitalize on our findings

1. Allocate more resources in the cities at high risk

There are cities that have a “high risk” profile given by the high value of one or more indexes. Our suggestion is to focus the resources on these cities in order to improve the vulnerability situation.

Socio-Economic Index & Physical Index	Mental Index	
Manchester	Newcastle	Norwich
Liverpool	Birmingham	Sheffield
Plymouth	Southampton	Edinburgh
London	Bristol	

2. Offer different services according to the different vulnerability

Different cities present different high vulnerability index values. Our suggestion is to take our findings and allocate the different resources and services that BRC offers according to the different issues that cities are facing in Great Britain. For this purpose, our suggestions are the following:



In cities with high **physical index** mobility aid services and support at home services can be reinforced in order to help people with high mobility issues in particular in Plymouth where respondents suggested that they faced high vulnerability given by MobilityHealth issues.



In cities with high **mental index** BRC could help providing more intense services of emotional support. In particular, the “Loneliness service” and “BRC support line” are services that could help improve the life purpose and pleasure, the two biggest issues respondents highlighted they have and worsened from wave 1 to wave 2.



In cities with high **socioeconomic index** BRC financial services like “Get help with money problems” or “UK victims of terrorism abroad” could help to improve the situation. Could be also useful to try to allocate more resources in migrant organizations since many respondents highlighted this as the main serviced used and increasing from wave 1 to wave 2.

BUSINESS ADVICES

- Capitalize on our findings

3. Allocate resources checking the cities that are on trend

Cities that are on negative trends (e.g., Manchester for the socioeconomical) are the ones that require more urgent aid. For these cities, it would be helpful to get in touch with the municipality to implement solutions as soon as possible.

There are also cities that are improving in wave 2 with respect to wave 1 (e.g., London for the physical Index and Birmingham for the mental index) but they still have some of the highest values of vulnerability in the country. We suggest to monitor the improvements in these cities and try to understand how those issues have been handled in order to be able to replicate the same solutions also in the other cities.

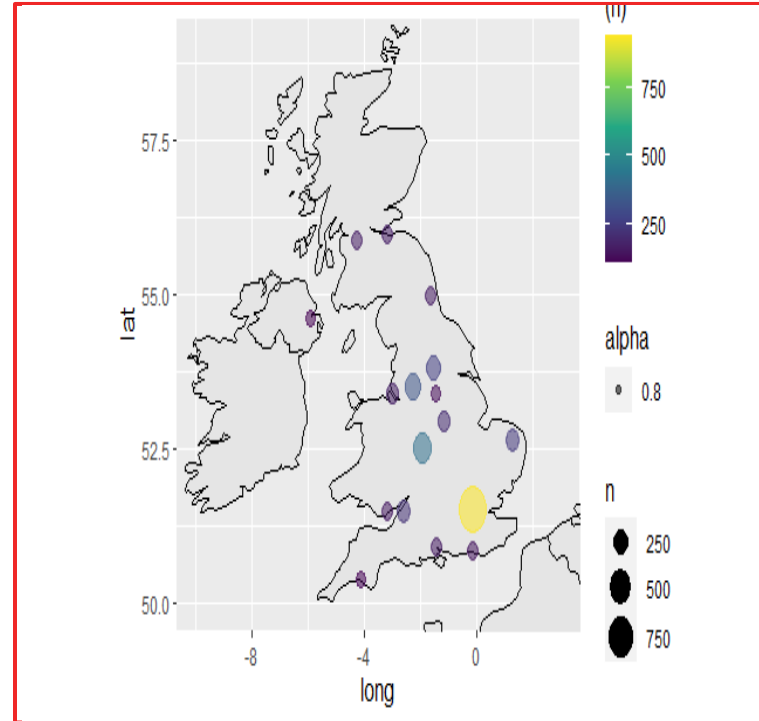
APPENDICES

APPENDICES

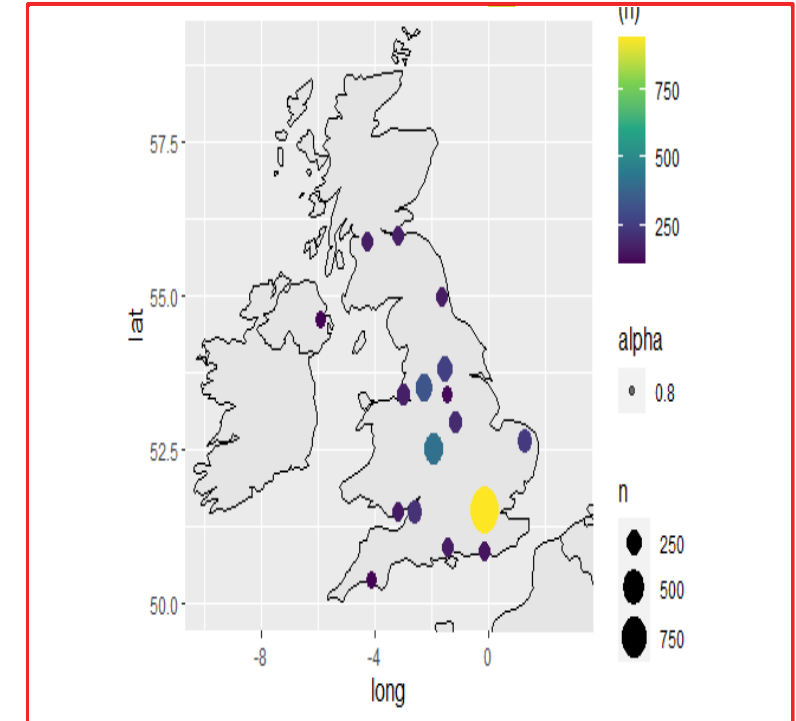
- Data Exploration
- Data Preparation
- Support for the Geographical Hypothesis
- Clustering
- Decision Tree
- Profiles of at risk cities

DATA EXPLORATION – MAP OF POPULATIONS FOR THE 2 WAVES

Population of Respondents
(Wave 1)



Population of Respondents
(Wave 2)



Insight: the distribution of population in cities across waves seem to be quite similar.

DATA PREPARATION – FEATURE ENGINEERING

For each index **different variables have been aggregated** according to what one answer is trying to explain.

Under the variables chosen for each indexes can be seen:

<u>Indexes</u>	<u>Variable Names</u>
Socio-Economic	"Q2_multi_UniversalCredit","Q2_multi_Parents","Q2_multi_OtherRelatives","Q2_multi_Friends","Q2_multi_Colleagues", "Q2_multi_Charities","Q2_multi_LocalWelfareFund","Q2_multi_PaidWork","Q2_multi_Begging","Q2_multi_Other", "Q2_multi_PreferNotSay","Q1_multi_BenefitSanctions","Q1_multi_BenefitDelays","Q1_multi_BehindOnBills","Q1_multi_BehindOnRent", "Q1_multi_SeriousDebt","Q1_multi_Evicted","Q1_multi_None","Q1_multi_RightLiveWork"
Physical	"Q1_multi_PhysicalProb","Q3_HealthLimitActivities", 'Q5_UsualActiHealth'
Mental Index	'Q6_LifeSatisfaction','Q6_Anxious','Q8_Depressed','Q9_MentalHealthSupport','Q10_MentalHealthSupportWaiting','Q11_Lonely'

DATA PREPARATION – FEATURE ENGINEERING

After choosing the relevant variables for each index, some operations have been processed. Overall, a **higher value of the index means a higher level of vulnerability**.



Where needed, the variables have been scaled to have the same range $[0,1]$ for all of them.



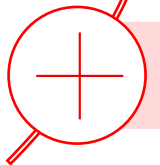
For ‘_none’ and “_Happy” answers, it has been taken the negative range since it is a measure of welfare not of illness.



For each variable has been computed the mean value (weight) on the total dataset.



The different variables of each instance have been divided by their weight.



At this point to get the index, we sum the different variables.

SUPPORT FOR THE GEOGRAPHICAL HYPOTHESIS

Before diving into the Machine Learning part, it seemed relevant to assess whether there was significant **evidence supporting geographical differences among the cities** of respondents. To answer this hypothesis, **Regression for Spatial Inference** was implemented on the cities of interest.

As it could be expected all different alphas are highly significant (**p-value < 0.01**) and the coefficient are quite high in line with the hypothesis of spatial fixed effects.

Test of median residuals within each cities:

- There seems to be a distinctive effect of intangible cities effect (see Appendix for the residual plots).

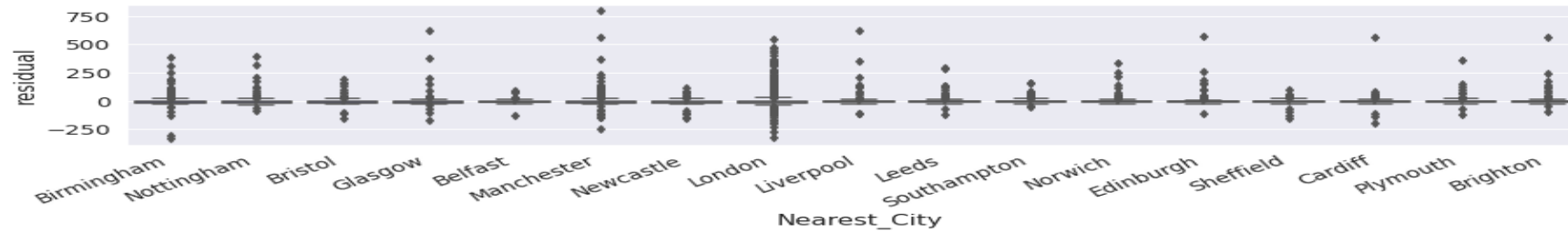
Introduction of Spatial Heterogeneity with Spatial Fixed Effects:

- Spatial Heterogeneity means that parts of the model may vary systematically with geography change. Spatial Fixed Effects are incorporated in the model by letting alpha vary in the regression according to the city (and after it was tested according to the **cluster** to confirm the hypothesis).

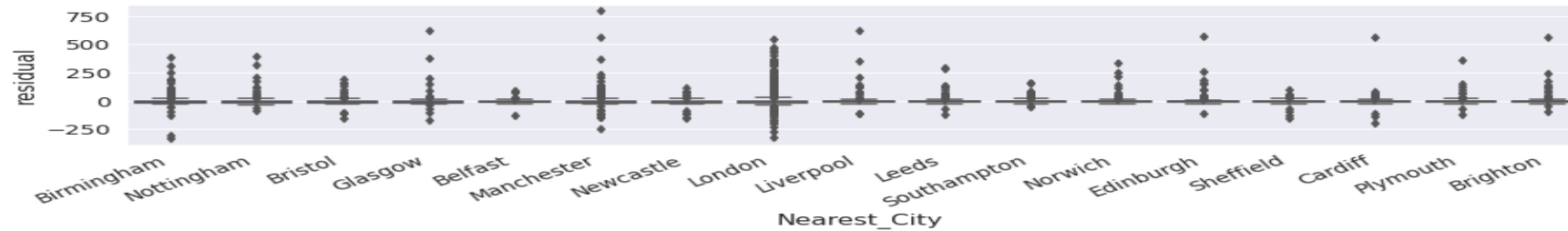
SUPPORT FOR THE GEOGRAPHICAL HYPOTHESIS

Median Residual within each city

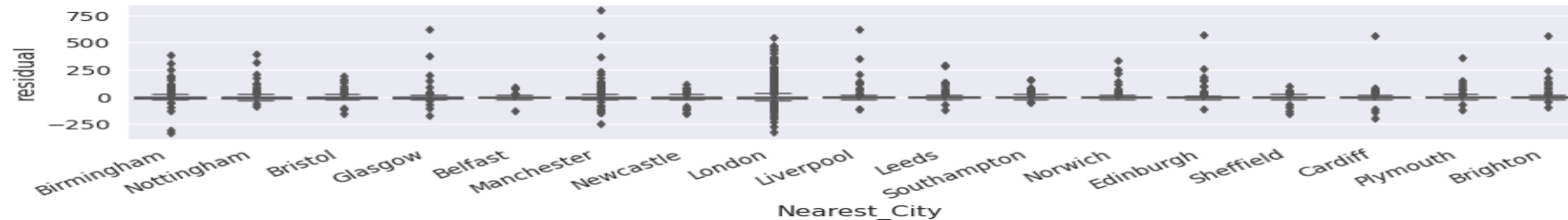
Socio-Economic



Physical



Mental

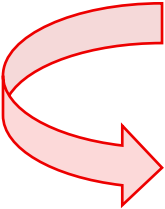


REGRESSIONS

The **motivation** for the use of regressions underlies **in choosing only the most relevant variables** and in making a **classification model as understandable as possible**.

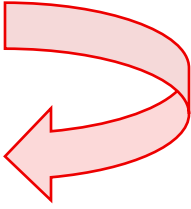
To create the **optimal** model and make sure the latter was as **robust** as possible, GridSearch Cross-Validation was used. The parameter alpha (i.e. responsible for the magnitude of the shrinkage of the coefficients) was evaluated at different values in order to **maximize the R-squared**.

The results are of the GridSearchCV for the 3 indexes with Lasso Regression (which is the best regression found):



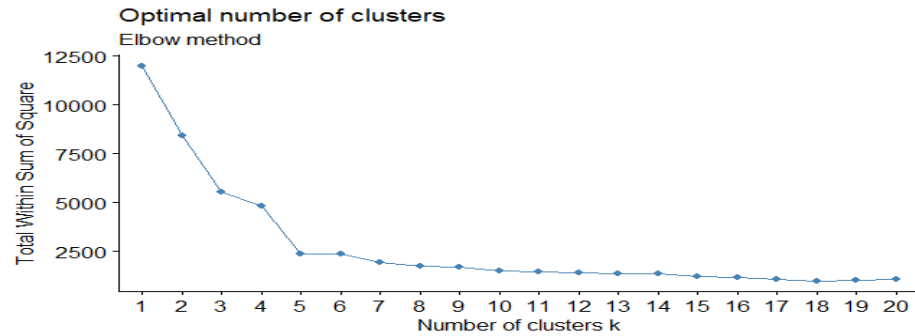
<u>Measures</u>	Cluster 1	Cluster 2	Cluster 3
Hyperparameters used	Alpha = 0.03	Alpha = 0.02	Alpha = 0.025
R-squared	31.65 %	32.7 %	52.2 %
Mean abs error	24.25	2.529	11.8
Mean squared error	3497.05	31.98	247.03

From these results, lasso regression seems to yield relevant results only for the Cluster 3. Thus, further analysis will be required to fully grasp what can explain the indexes and classify the respondents according to it.

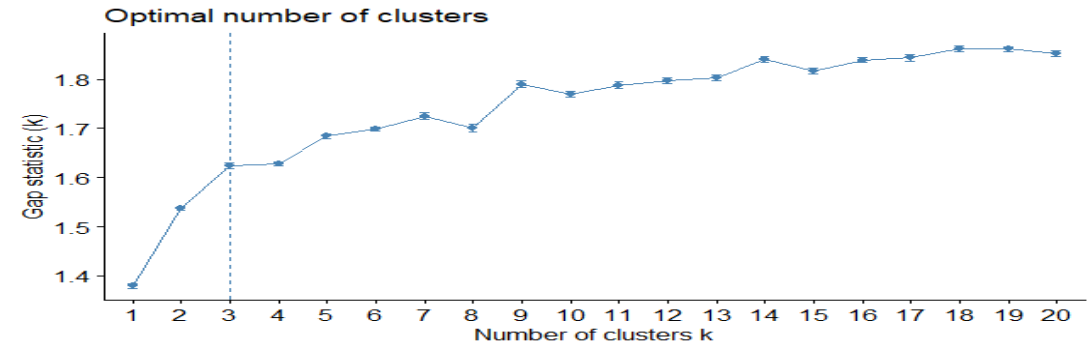


CLUSTERING

The choice of the number of clusters relies on 2 statistics: **Elbow method** and **Gap statistics**.



Elbow method: it consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve.



Gap Statistic: it standardizes the graph of the log within-cluster dispersion, by comparing it to its expectation under an appropriate null reference distribution.



Cluster plot on the Physical and Socio-Economic dimensions



Cluster plot on the Mental and dimensions

DECISION TREE MODELING

To implement the decision tree model, several steps have been processed (in a **modeling function**):

Separate the target variable 'y' (1 if the respondent belongs to one cluster, 0 otherwise) from the explanatory variables 'X'.



Split the dataset between a training (70%) and a test test (30%).



Use of an OverSampler method on the training set (and **not** on the test set as it is required to keep a natural density to test) to enable the model to learn more easily. This choice is motivated by the fact that decision trees are **very sensitive to unbalanced dataset**. The method is ADASYN. The essential idea of ADASYN is to use a weighted distribution for different minority class examples according to their level of difficulty in learning, where more synthetic data is generated for minority class examples that are harder to learn (H He et al., 2008).



Use of the built model on the test set and compute various metrics such as accuracy, AUC, Cohen kappa, etc.

LONDON IS THE MOST AT RISK CITY FOR THE SOCIO ECONOMIC INDEX WITH AN UNDECIDED TREND

London is identified as the most at risk city for the socio economic index, but we do not see any statistically significant changes between wave 1 and wave 2.

Drivers of the socio economic index for London are:

- ✓ Coming UK
- ✓ None
- ✓ Migrant Organisation
- ✓ Homeless
- ✓ Food Banks

The typical person doing worse than the average:

- Male
- Is between 30 and 50
- Works full time in a higher managerial position
- Has no kids
- Lives in an urban area
- Is white



All the values mentioned above has a positive t-value, and are statistically significant on a 99,7 significance level and above. This implies that individuals in wave 2 score lower on the variables Coming UK, Migrant Organisation, Homeless and Food Banks, which overall is a very good result. However, the variable «None» is also increasing, thus implying that the number of individuals experiencing none of the issues listed in question 1 of survey 2, is higher in wave 2 compared to wave 1. It could therefore be that the issues these individuals experience is now more spread out.

LIVERPOOL IS ONE OF THE MOST AT RISK CITY FOR THE SOCIO ECONOMIC INDEX WITH AN UNDECIDED TREND

Liverpool is identified as one of the most at risk city for the socio economic index, but we do not see any statistically significant changes between wave 1 and wave 2.

Drivers of the socio economic index for Liverpool are:



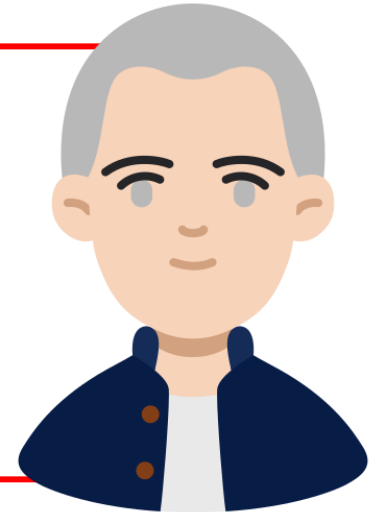
Food Banks



None

The typical person doing worse than the average:

- Male
- Is between 50 and 60
- Works full time as a skilled manual worker or supervisory
- Has 1 kid
- Lives in an urban area
- Is white



Both the variable “None” and “Food Banks” has a positive t-value. For “None”, the t-value is 3,54 and for “Food Banks” the t-value is 1,75. “None” is statistically significant on a 100 percentage significance level, while “Food Banks” is significant on a 91,9 significance level. This means that both variables are measures to be higher in wave 1, compared to wave 2. This means that less individuals depends on food banks in wave 2, but the overall population in Liverpool is also experiencing more issues overall in wave 2, compared to wave 1. The individuals doing the worst, are the elder.

PLYMOUTH IS ONE OF THE MOST AT RISK CITY FOR THE SOCIO ECONOMIC INDEX WITH AN UNDECIDED TREND

Plymouth is identified as one of the most at risk city for the socio economic index, but we do not see any statistically significant changes between wave 1 and wave 2.

The driver of the socio economic index for Plymouth are:



None

The typical person doing worse than the average:

- Male
- 23 years old
- Works full time as a supervisory
- Has no kids
- Lives in a suburban area
- Is white



The variable “None” has a positive t-value of 2,25 on a 97,5 percentage significance level. This means that the overall population in Plymouth is experiencing more issues overall in wave 2, compared to wave 1. The individuals doing the worst, were the youngest in this city.

MANCHESTER IS ONE OF THE MOST AT RISK CITIES FOR THE PHYSICAL INDEX

Manchester's physical index is positively increasing. We obtain a t-value of 1,83 when testing the change between the waves, which is statistically significant on a 93,2 significance level.

There are no statistically significant changes in any of the variables driving the physical index. Therefore it seems like the variables have stayed high and stable in both wave 1 and 2. Notice that it is both the elderly and younger population who is doing the worst in Manchester.

The typical person doing worse than the average:

- Male
- 20 to 40 or 60 to 70 years old
- Works full time in a supervisory position
- Has 1 kid
- Lives in a suburban area
- Is white



LIVERPOOL IS THE MOST AT RISK CITY FOR THE PHYSICAL INDEX WITH AN UNDECIDED TREND

Liverpool is identified as the most at risk city for the physical index, but we do not see any statistically significant changes between wave 1 and wave 2.

There are no statistically significant changes in any of the variables driving the physical index. Therefore it seems like the variables have stayed high and stable in both wave 1 and 2. Notice that it is the elderly population who is doing the worst in Liverpool.

The typical person doing worse than the average:

- Unknown sex
- Is 60 or above
- Is retired, but has worked as a skilled manual worker
- Has 1 kid
- Lives in a suburban area
- Is white



PLYMOUTH IS ONE OF THE MOST AT RISK CITY FOR THE PHYSICAL INDEX WITH AN UNDECIDED TREND

Plymouth is identified as one of the most at risk city for the physical index, but we do not see any statistically significant changes between wave 1 and wave 2.

The driver of the physical index for Plymouth are:



Mobility Health

The typical person doing worse than the average:

- Male
- Is between 50 and 60
- Works full time as a skilled manual worker or supervisory
- Has 1 kid
- Lives in an urban area
- Is white



The variable “Mobility Health” has a t-value of -2,65 with a significance level of 99,1. This means that more people are experiencing problems related to their mobility health in Plymouth in wave 2, compared to wave 1. This could be explained by that fact that the typical person doing worse than the average tends to be a older, retired individual.

EDINBURGH IS ONE OF THE MOST AT RISK CITY FOR THE MENTAL INDEX WITH AN UNDECIDED TREND

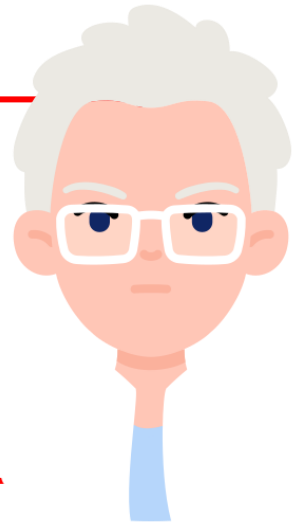
Edinburgh is identified as one of the most at risk city for the mental health index, but we do not see any statistically significant changes between wave 1 and wave 2.

Drivers of the mental health index for Edinburgh are:

- ✓ No Children
- ✓ Life Purpose
- ✓ No Pleasure

The typical person doing worse than the average:

- Unknown sex
- Is between 55 and 70
- Works full time as a supervisory or is retired
- Has 1 kid
- Lives in a suburban area
- Is white



“No Children” has a t-value of 2,09 with a significance level of 96,3, which implies that individuals are more likely to have kids in wave 2. The variable “Life Purpose” has a negative t-value of -2,05, which is statistically significant on a 95,9 percentage significance level. More individuals in Edinburg experience that life has a purpose in wave 2, compared to wave 1. The variable “No Pleasure” has a t-value of 1,84, which is statistically significant on a 93,3 percentage significance level. This implies that less people experience that life has no pleasure in wave 2, compared to wave 1.

NEWCASTLE IS ONE OF THE MOST AT RISK CITY FOR THE MENTAL INDEX WITH AN UNDECIDED TREND

Newcastle is identified as one of the most at risk city for the mental health index, but we do not see any statistically significant changes between wave 1 and wave 2.

The driver of the mental health index for Newcastle is:



Migrant Organisation

The typical person doing worse than the average:

- Unknown sex
- Is between 40 and 50, or above 65
- Works full time in an intermediate managerial position
- Has 1 kid
- Lives in a suburban area
- Is white



The variable “Migrant Organisation” has a positive t-value of 1,74, which is statistically significant on a 91,7 percentage significance level. This implies that less individuals in Newcastle experience troubles with migrant organisations in wave 2, compared to wave 1.

SHEFFIELD IS ONE OF THE MOST AT RISK CITY FOR THE MENTAL INDEX WITH AN UNDECIDED TREND

Sheffield is identified as one of the most at risk city for the mental health index, but we do not see any statistically significant changes between wave 1 and wave 2.

Drivers of the mental health index for Sheffield are:

- ✓ No Children
- ✓ Life Purpose

The typical person doing worse than the average:

- Unknown sex
- Is between 55 and 75
- Works full time as a supervisory
- Has 1 kid
- Lives in a suburban area
- Is white



The variable “No Children” has a t-value of 2,58, and is statistically significant on a 99,0 percentage significance level. This implies that the habitants of Sheffield are less likely to not have kids in wave 2, compared to wave 1. The variable “Life Purpose” has a negative t-value of -1,52, which is statistically significant on a 87,0 percentage significance level. More individuals in Edinburg experience that life has a purpose in wave 2, compared to wave 1.

NORWICH IS ONE OF THE MOST AT RISK CITY FOR THE MENTAL INDEX WITH AN UNDECIDED TREND

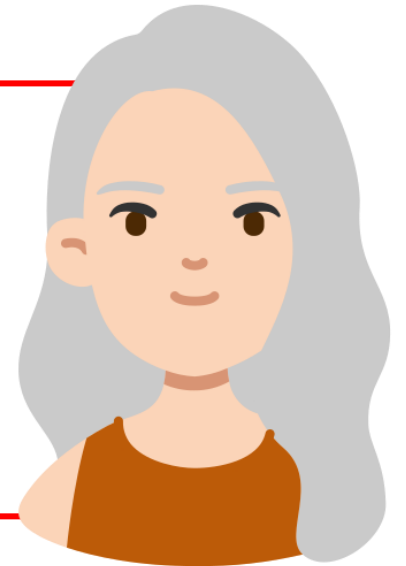
Norwich is identified as one of the most at risk city for the mental health index, but we do not see any statistically significant changes between wave 1 and wave 2.

Drivers of the mental health index for Norwich are:

- ✓ Migrant Organisation
- ✓ Domestic Abuse

The typical person doing worse than the average:

- Unknown sex
- Is above 60 years
- Works full time as a supervisory or is unemployed
- Has 1 kid
- Lives in a suburban area
- Is white



The variable “Migrant Organisation” has a positive t-value of 3,05, which is statistically significant on a 99,8 percentage significance level. This means that less individuals in Norwich is experiencing issues with migrant organisations in wave 2, compared to wave 1. The variable “Domestic Abuse” has a positive t-value of 3,15, and is statistically significant on a 99,8 percentage significance level. This implies that individuals now are experiencing less domestic abuse, compared to before in wave 1.

SOUTHAMPTON IS ONE OF THE MOST AT RISK CITY FOR THE MENTAL INDEX WITH AN UNDECIDED TREND

Southampton is identified as one of the most at risk city for the mental health index, but we do not see any statistically significant changes between wave 1 and wave 2.

There are no statistically significant changes in any of the variables driving the mental health index. Therefore it seems like the variables have stayed high and stable in both wave 1 and 2. Notice that it is the elderly population who is doing the worst in Southampton.

The typical person doing worse than the average:

- Unknown sex
- Is between 50 and 75 years old
- Works full time as an intermediate managerial position, or is retired
- Has 1 kid
- Lives in a suburban area
- Is white



BRISTOL IS ONE OF THE MOST AT RISK CITY FOR THE MENTAL INDEX WITH AN UNDECIDED TREND

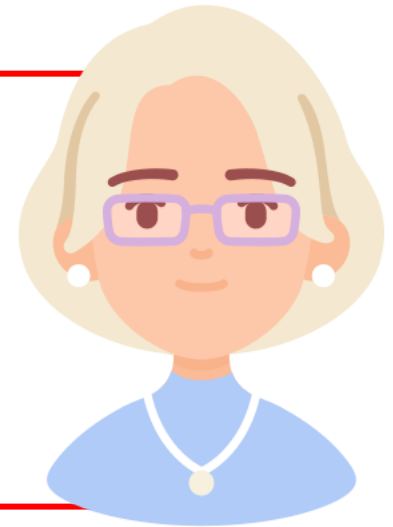
Bristol is identified as one of the most at risk city for the mental health index, but we do not see any statistically significant changes between wave 1 and wave 2.

Drivers of the mental health index for Bristol are:

- ✓ No Pleasure
- ✓ Life Purpose
- ✓ Migrant Organisation
- ✓ Domestic Abuse

The typical person doing worse than the average:

- Unknown sex
- Is between 40 and 60 years old
- Works full time in an intermediate managerial position
- Has 1 kid
- Lives in a suburban area
- Is white



All variables, except “Life Purpose” has a positive t-value, with significance levels above 90 percentage. “No Pleasure” has a positive t-value of 2,92, implying that more people experience life as more pleasurable in wave 2. “Life Purpose” has a negative t-value of -2,77, which is statistically significant on a 99,4 percentage significance level. This means that more individuals in wave 2 experience life as less meaningful. “Migrant Organisations” has a t-value of 1,74. This means that less individuals in Bristol is experiencing issues with migrant organisations in wave 2. “The variable “Domestic Abuse” has a positive t-value of 2,26. This implies that individuals now are experiencing less domestic abuse, compared to before in wave 1.

REFERENCES

- Analytics Vidhya. *A Complete Tutorial on Ridge and Lasso Regression in Python*. (n.d.). Retrieved May 21, 2022, from <https://www.analyticsvidhya.com/blog/2016/01/ridge-lasso-regression-python-complete-tutorial/>.
- He, H., Bai, Y., Garcia, E., & Li, S. (2008). ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. In *Proceedings of the International Joint Conference on Neural Networks* (p. 1328). <https://doi.org/10.1109/IJCNN.2008.4633969>.
- *Home—Geographic Data Science with Python*. (n.d.). S.Rey et al. Retrieved May 19, 2022, from <https://geographicdata.science/book/intro.html>.
- Seyedan, M., Mafakheri, F., & Wang, C. (2022). Cluster-based demand forecasting using Bayesian model averaging: An ensemble learning approach. *Decision Analytics Journal*, 3, 100033. <https://doi.org/10.1016/j.dajour.2022.100033>.
- Kuran, C.H.A, et al. (2020). Vulnerability and vulnerable groups from an intersectionality perspective, *International Journal of Disaster Risk Reduction*, 50. <https://doi.org/10.1016/j.ijdr.2020.101826>.
- Coding files: https://drive.google.com/drive/folders/1RXxB1AavjobXGDPI5TM1Bmh7eWhTMN_2?usp=sharing.