How are you, really?

Measuring mental health and **identifying** people in critical need to support the **British Red Cross** in **providing help** to those who need it the most

QTEM Data Challenge - 2022
Q-Team (8)

Demarets Mathieu
Galletta Erika
Moreira Laura
Rösch Florian

Executive Summary | Defining the problematic of mental health support and elaborating the LSI and MHI to measure mental health

Understanding the mental health issue

When **basic needs** are met but **inequalities** are at their worst, mental health issues arise

Strong stigma (Public, self and institutional) around mental health prevent individuals from **reaching out**

Covid-19 **pandemic** drove the mental health down through the combined effects of **loneliness** and **lack** of physical **activities**

The British Red Cross mission and actions

The British Red Cross (BRC) aims at **providing help** to those who **need** it most

Besides their numerous operations, they opened a **support line** in 2020 to support people during the pandemic

We can help the BRC in their resource allocation by:
1) Developing a **measure** of mental health
2) **Identifying** people in need

Tools to measure mental health

BRC **Understanding Vulnerability Survey** 2022 (Wave 2) will be used to achieve our goals after some **cleaning** and **feature engineering** for the most relevant variables

Literature on mental health promotes the use of **indexes** based on **survey** answers (scores on point scales) to assess life satisfaction and mental health

Life Satisfaction and Mental Health index

Solution to the first objective

Life Satisfaction Index will be created based on 'life worth, happiness and life satisfaction' indicators
& **Mental Health Index** will also be created based on 'anxiousness , lack of interest and loneliness' indicators

Linear regressions will help us define which **variables significantly** impact each index

Executive Summary | Assigning individuals to help groups and predicting the help group based on easily observable variables

Clustering individuals in need groups

Partial solution to the second objective (1)

Naïve clustering based on LSI and MHI helps us to identify critical **need groups** and show potential for refined analysis (clustering and logit regression)

Advanced **machine learning** leads us to cluster the sample in four need groups with k-means and assign a **priority** to each group

Sociodemographics of the need group

We look at the **socio-demographic** variables which characterize the individuals in the highest **priority** help group to understand who is in that group

The **findings** are **aligned** with the results of previous linear regressions, which enables us to confirm our choice of **socio-demographic** variables to use in the predicting model

Predicting needs with observable variables

Partial solution to the second objective (2)

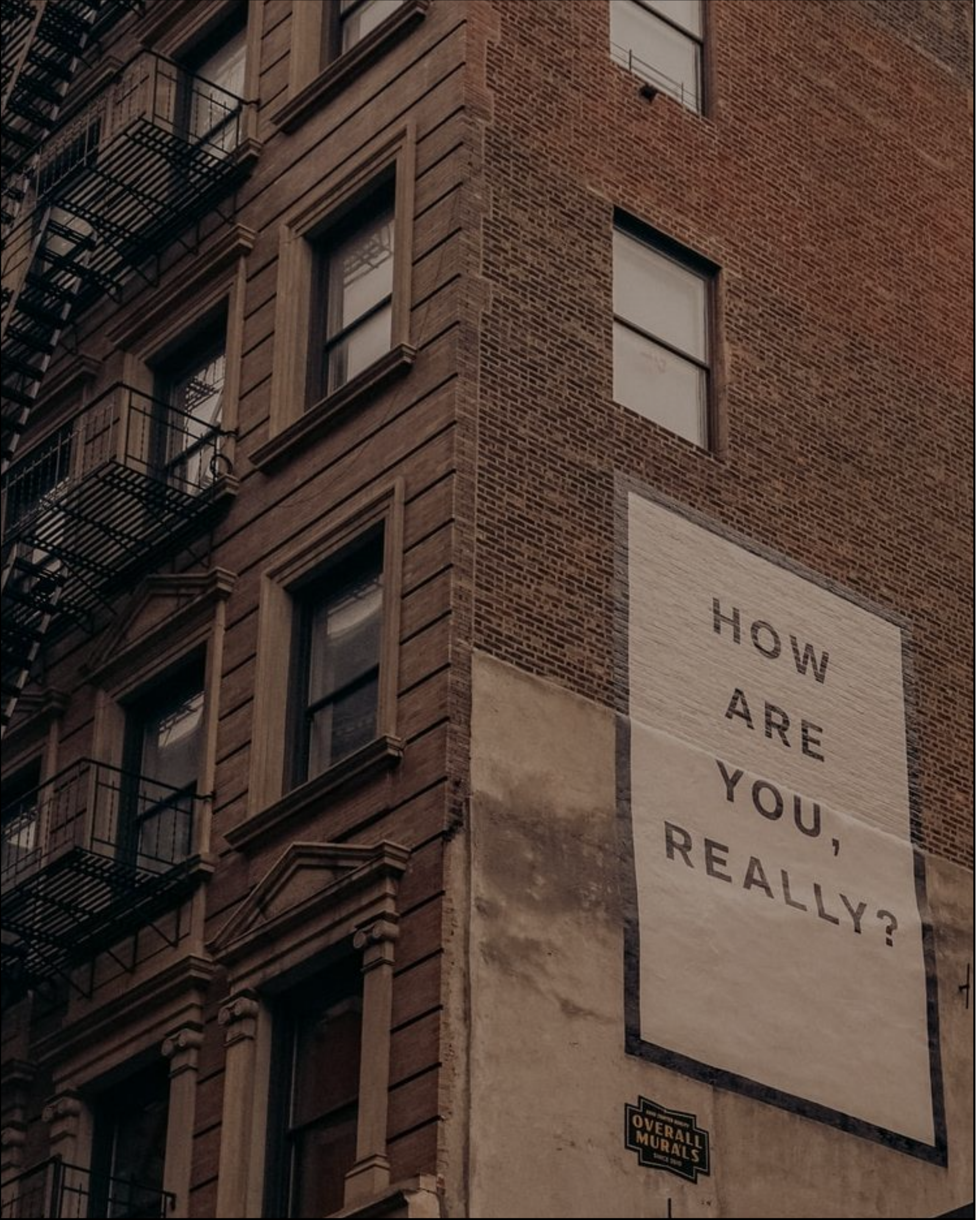
Instead of having to survey the mental health and life satisfaction to **allocate resources**, we develop a **predictive model** (probit) based on the significant socio-demographic variables

Based on the **coefficients** of the model we offer a few paths for **interpretation** and **action**

Limits of our research and next steps

To achieve greater impact:

- **Enrich** the **survey** with more complete data and variables drawn from other surveys
- Develop an interactive **dashboard** to update and **visualize** results
- Research the **efficiency** of the **support** provided given the support **cost**



PART 1 | Mental health trends

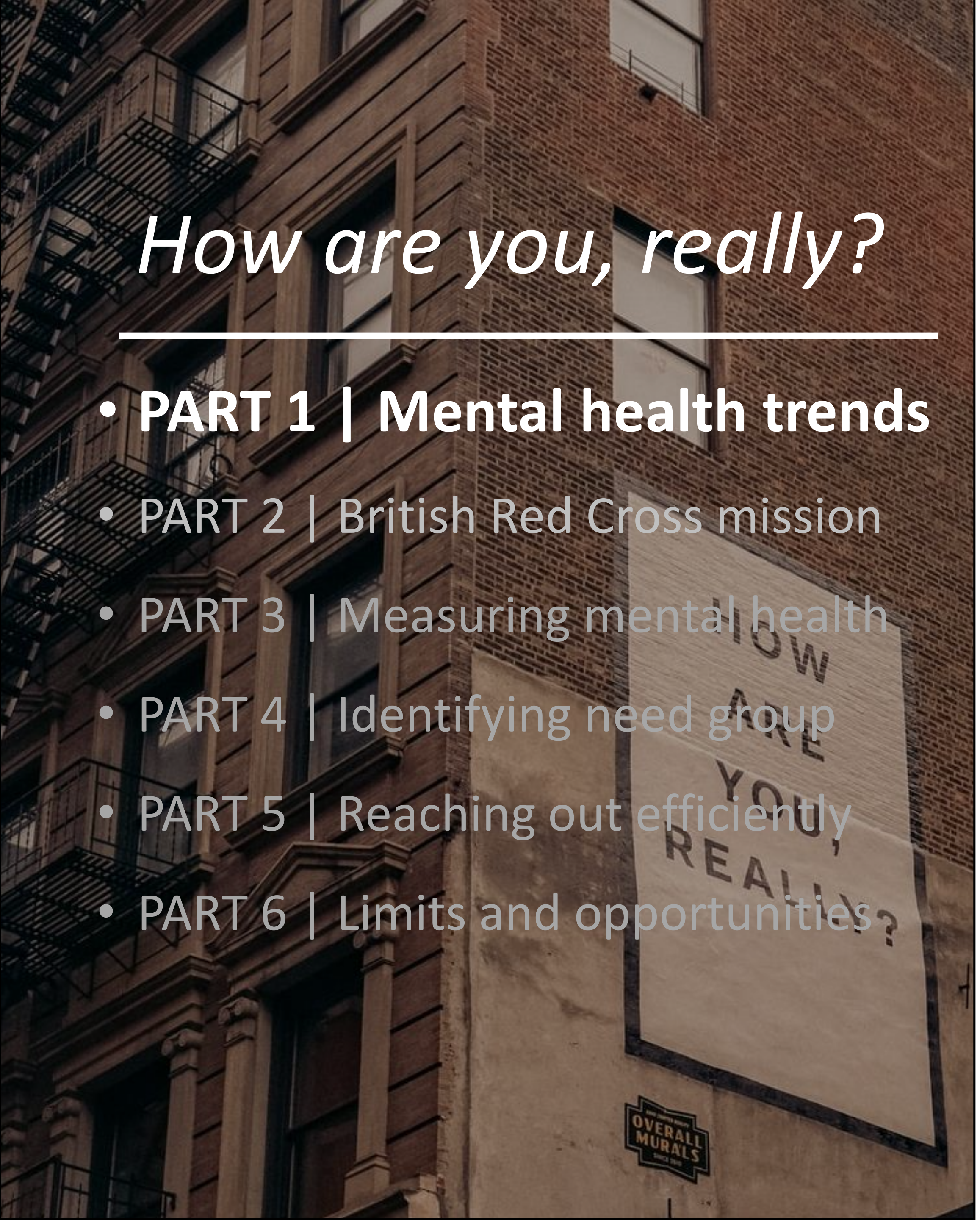
PART 2 | British Red Cross mission

PART 3 | Measuring mental health

PART 4 | Identifying need group

PART 5 | Reaching out efficiently

PART 6 | Limits and opportunities



How are you, really?

- **PART 1 | Mental health trends**
- PART 2 | British Red Cross mission
- PART 3 | Measuring mental health
- PART 4 | Identifying need group
- PART 5 | Reaching out efficiently
- PART 6 | Limits and opportunities

British Red Cross

Mental health trends

a. Mental health importance (psychological needs & inequalities)

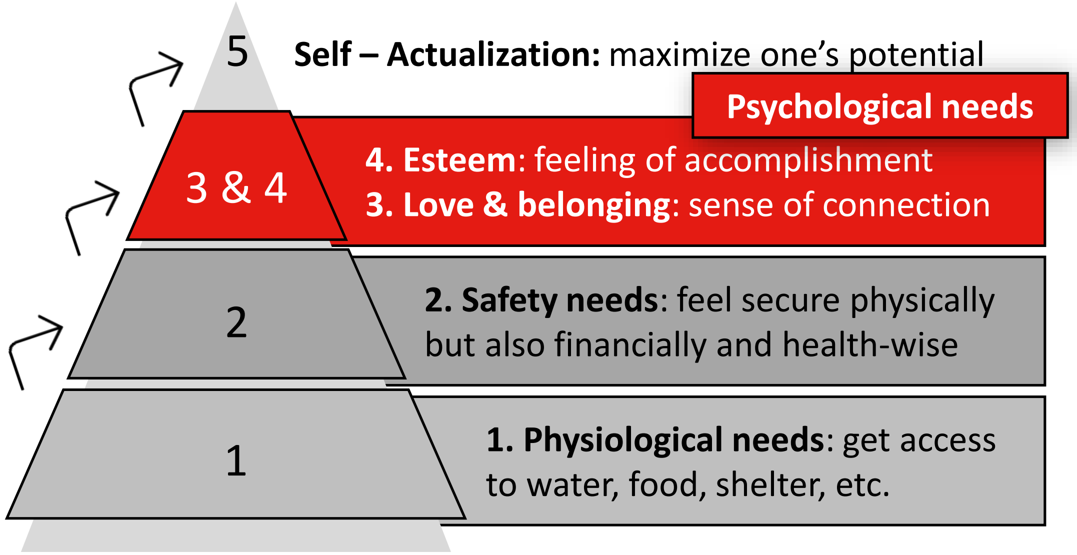
b. Stigmatization around mental health issues obstructs treatment

c. Covid-19 pandemic reinforced the mental health issues

QTEM DATA CHALLENGE

Mental health trends | When basic needs are met but inequalities are at their worst, mental health issues become a top focus

Address psychological needs after physiological and safety needs are fulfilled



In richer countries such as the United Kingdom, most of the population can meet their **physiological needs** and their **safety needs**. According to the Maslow pyramid (Maslow, 1942), individuals will then start adressing their **psychological needs** of belonging and esteem. **Mental health support** then becomes critical as individuals will be confronted to loneliness, doubt, and depression in their search for love, belonging and estime.

Mental health issues should be addressed early because they reinforce pre-existing inequalities

Uncovering the hidden impacts of inequality on mental health: a global study (Yu, 2018)

In the study, it was shown that not everyone is equal in the face of mental health issues. In fact, they seem to affect individuals in disadvantaged groups more strongly:

1. Twice as many **women** are affected by mental illnesses as men, which may be related to societal **gender inequalities**
2. The **GINI** index (measure of distribution of income) is connected to the degree of **depression disorder** for males
3. Gender **differences** in depressive illnesses are **related** to the **richness** of a country

→ ***Importance of prevention: Tackle mental health issues to avoid making the pre-existing inequalities worst***

Mental health trends | Effective treatment of mental disorders is obstructed by public, self, and institutional stigmas

Issue in providing treatment

Over half of people with mental disorders do not obtain treatment. People frequently refuse or postpone therapy due to fears of being treated unfairly or of losing their employment and livelihood. (American Psychiatric Association, 2020)

Underlying Mechanisms

Public stigma

The negative or discriminating attitudes that others hold about mental illness are referred to as public stigma

Self stigma

Self-stigma refers to the unfavourable sentiments that people with mental illnesses hold regarding their own situation

Institutional stigma

Institutional stigma is more structural; government and private-sector who purposefully or not limit opportunities for people with mental illnesses

To help reduce these issues, it is critical to educate individuals and work to eliminate the stigma associated with mental health

Mental health trends | The COVID-19 pandemic makes loneliness and isolation worst for individuals

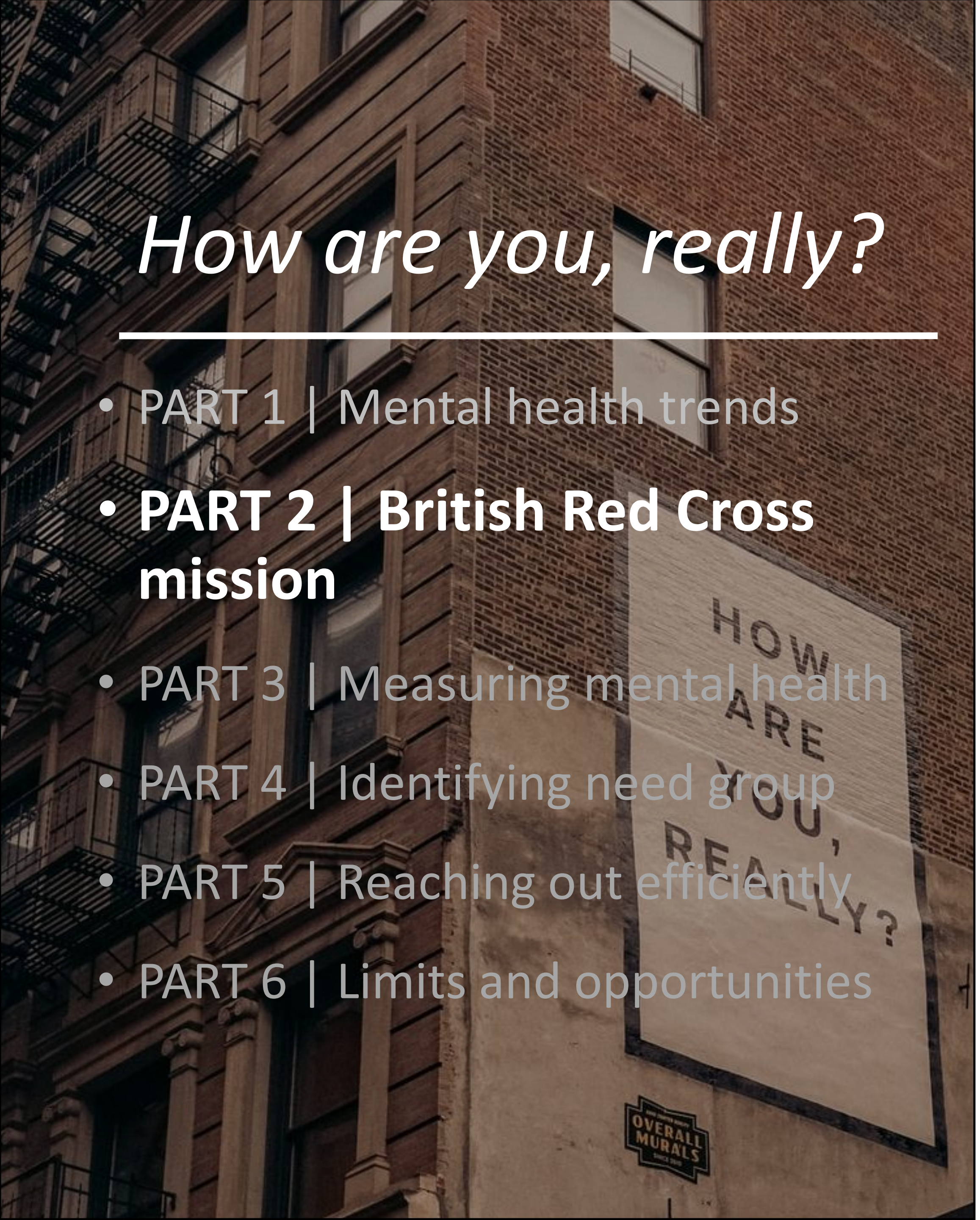
Attachment style and mental health during the later stages of COVID-19 pandemic: the mediation role of loneliness and COVID-19 anxiety (Vismara et al. 2022)

Loneliness has a direct impact on mental health problems and serves as a bridge between insecure attachment patterns and COVID-19-related anxiety symptoms, which causes mental health difficulties

The relationship between physical activity and mental health in a sample of the UK public: A cross-sectional study during the implementation of COVID-19 social distancing measures (Jacob et al. 2022)

Some groups were more severely affected by lockdown-caused isolation, such as: women, young individuals, impoverished people, smokers and those with physical problems

Combined with the reinforcement of **pre-existing inequalities** and the **stigmatization** of mental disorders, the Covid-19 **pandemic effects** push Mental Health Issues in the **top priority** issues to tackle



How are you, really?

British Red Cross

British Red Cross mission

a. Mission to provide help to people who need it most

b. New support line in 2020 on top of numerous previous actions

d. ‘Qteam’ can help the BRC in their resource allocation by measuring and identifying needs

QTEM DATA CHALLENGE

British Red Cross' Mission is to assist **disadvantaged people** around the world in **preparing** for, **surviving**, and **recovering** from disasters in their own communities

British Red Cross' focus is on **providing help** to those who **need it the most**, thus contributing to the **dignity of** the **life** of those who face vulnerability

British Red Cross Mission | The BRC opened a support line to relieve mental distress on top of their numerous other actions

Refugees support
Help refugees and asylum seekers to get back on their feet after a traumatic time

Emergency Response
Assist people in case of fires, floods or transport accidents

Research and speak up
Carry out research on a range of issues to help the case for change and speak up to bring that change to UK policies

British Red Cross

Support Line (2020)
+
Help people cope with loneliness and lockdowns and provide food and medicine to those who need it most (+ actions of psychosocial team)

People in armed conflicts
Protect people in countries torn by war and conflict, save lives

International Work
Partner with other Red Cross and Red Crescent societies around the world to prevent or ease human suffering

https://www.redcross.org.uk/about-us/what-we-do

British Red Cross Mission | Our goal is twofold, first measure mental health problems and then better identify people in need

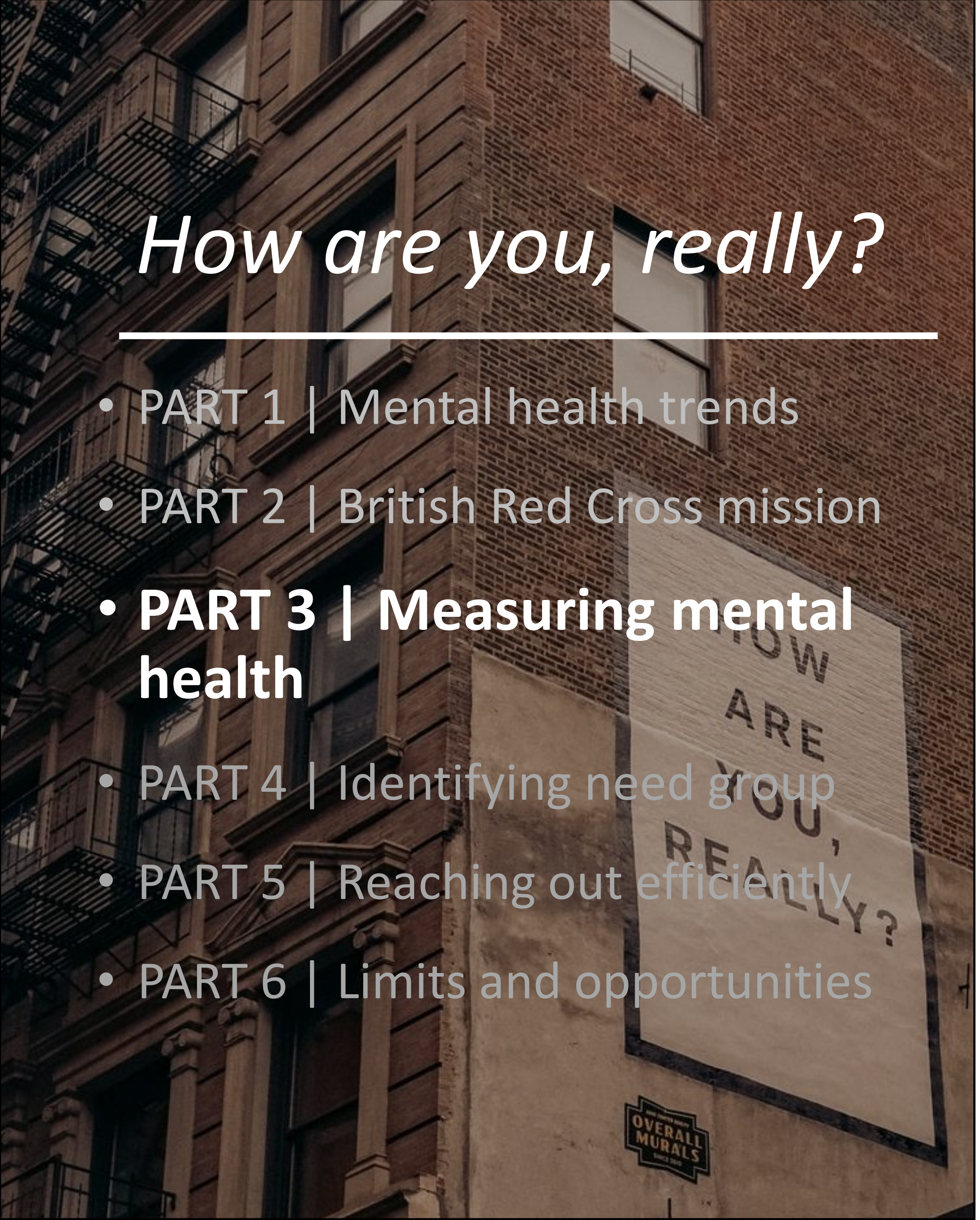
The British Red Cross can only rely on a limited budget to support its numerous actions and therefore needs to know where its resources will be best spent to alleviate mental health disorders

Our Goal

Clearly identify **where the British Red Cross efforts should be focused** and design a tool to **consistently identify this target in future surveys**

1) How can we **measure** mental health? (PART 3)

2) How can we better **identify** people in need? (PART 4 & 5)



How are you, really?

- PART 1 | Mental health trends
- PART 2 | British Red Cross mission
- **PART 3 | Measuring mental health**
- PART 4 | Identifying need group
- PART 5 | Reaching out efficiently
- PART 6 | Limits and opportunities

BritishRedCross

Measuring mental health

a. Exploring the raw survey data and engineering features

b. Literature review on mental health measures (surveys and indexes)

c. Computation of our LSI and MHI based on the vulnerability survey

d. LSI and MHI exploration through linear regressions

QTEM DATA CHALLENGE

Measuring mental health | Working on the raw survey data to create valuable and complete information through feature engineering

From raw data → *feature engineering* → To valuable information

Raw survey data exploration

Understanding Vulnerabilities Survey 2022 (Wave 2)
→ BRC wide scope survey to identify people in need

Facts and Figures:

Longitudinal online survey of 11 questions collected from July 2021 to January 2022

- ✓ 2 **multiple** choice questions (dummy parametrization for each: 0 or 1)
- ✓ 9 **single** choice questions

4000 participants all over the UK from age 18 and above

- **Demographic data** including gender, age, ethnicity, employment, social grade, age of children and living area
- **Geographic data** including the region, the nearest city and the related UK two-digit postcode

Feature engineering examples

Region:
Data: Nearest City, Region, 2-digit UK Postcode
Problem: Region data contains **47% missing** data

FROM

Solution: matching postcodes to UK regions and create new region data

TO

Youngest Child:
Data: info about having a child (=1) or not (=0)
Problem: No information about number of children per participant. We cannot control if a certain children age influences the participants.

FROM

Solution: Calculating youngest child of each participant to assess if a certain age of children has an impact

TO

Measuring mental health | Feature engineering enables a better understanding of the previously disperse and incomplete variables

From raw data — *feature engineering* → **To valuable information**

Post_EM	Post_East	Post_Lon	Post_NE	Post_NW	Post_Scot	Post_SE	Post_SW	Post_Wal	Post_WM	Post_York
NG										
										LS
					G					
NG										
		E								
		UB								
S										
LE										

Children1_expecting	Children2_0_3	Children3_4_6	Children_4_7_10	Children5_11_13	Children6_14_15
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	1
0	0	0	0	0	0
0	0	1	1	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	1	1	1
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	1	0	0

youngest	postcode	postcode_area	region
teenager	NG	Nottingham	East Midlands
adult	LS	Leeds	North East
No children	G	Glasgow	Scotland
teenager	KY	Kirkaldy	Scotland
adult	NG	Nottingham	East Midlands
kid	E	London	Greater London
adult	BL	Bolton	North West
adult	M	Manchester	North West
adult	B	Birmingham	West Midlands
adult	BH	Bournemouth	South West
kid	S	Sheffield	East Midlands
No children	DA	Dartford	Greater London
No children	UB	Southall	Greater London
adult	ST	Stoke on Trent	West Midlands
No children	S	Sheffield	East Midlands
No children	LD	Llandrindod	Wales
No children	LE	Leicester	East Midlands
kid	E	London	Greater London

For a detailed description of the regional distribution, see Appendix.

Region: Multiple postcode columns + Incomplete region data are difficult to use (47% NA's)

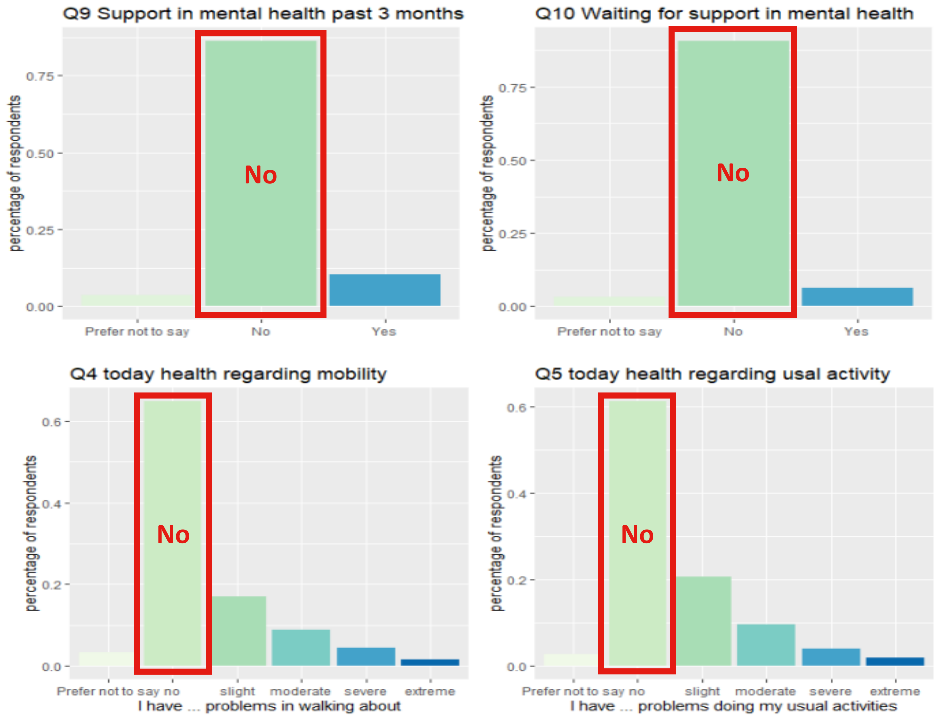
Children: Binary variable for each age category prevents proper readability

feature engineering →

Engineered Features:

The new variables summarize the most important information in a few clear and compact variables

Measuring mental health | Imbalances in the answers reduce the information richness and make interpretation more difficult



For a similiar information on the other questions, see Appendix

Q1_a13: "None of the above" → **61** %
Q2_a15: "None of the above" → **89** %
Q2_a16: "Not applicable" → **73** %
Q2_a17: "Prefer not to say" → **99** %

A very **imbalanced** dataset can lead to problems in classification and regression analysis (*kmeans and logit in further sections*)

The model trained on imbalanced data has a big risk of **overfitting** because of insufficient different values to be trained with

The explanatory power of our analysis could be affected and this limit should be considered when examining our results

Measuring mental health | Indexes defined on the basis of survey questions are the most common tool to measure mental health

Common methodology to measure mental health

Many **surveys**, such as the BRC Vulnerability Survey, are employed to have a large and representative **sample** of individuals answer a few questions, and enable researchers to **create** mental health **indexes**

"Does Life Satisfaction Change in Old Age: Results From an 8-Year Longitudinal Study"

(Gana et al. 2012)

Survey (0 to 7 scale)

1. "In most ways, my life is close to my ideal"
2. "The conditions of my life are excellent"
3. "I am satisfied with my life"
4. "So far I have gotten the important things I want in life"
5. "If I could live my life over, I would change almost nothing"

Life Satisfaction Index

"The Mental Health Index in the Italian Regions"

(Leogrande 2022)

Questions extracted from short form survey related with:

1. Anxiety
2. Depression
3. Emotional control
4. Psychological well being

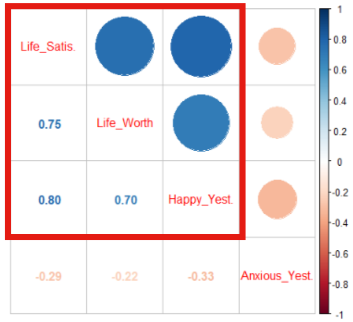
Mental Health Index

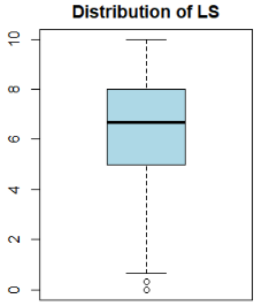
Measuring mental health | The LSI and MHI will facilitate the analysis by summarizing the information of 7 correlated variables

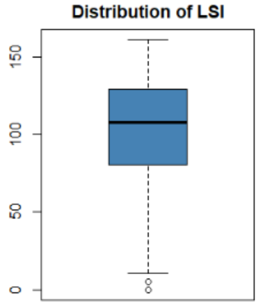
Life satisfaction index

Interesting questions in the survey to assess **life satisfaction** were:

Q6_a1: How satisfied are you with your life nowadays? (Life satisfaction)
Q6_a2: To what extent do you feel the things you do in your life are worthwhile? (Life worth)
Q6_a3: How happy did you feel yesterday? (Happy yesterday)







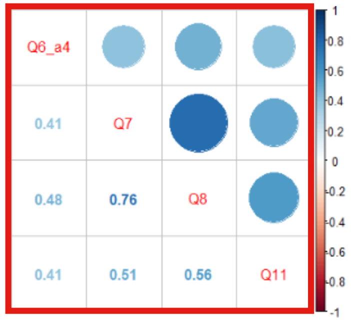
$\boldsymbol{LS} = (Q6_{a1} + Q6_{a2} + Q6_{a3})/3$ $\boldsymbol{LSI} = 100 * LS/mean(LS)$

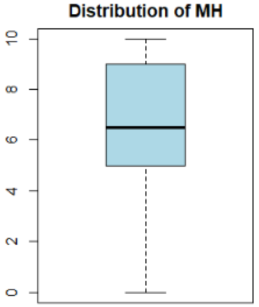
*The three questions were answered on a scale from **0 to 10** and given **equal weight** in the index.*

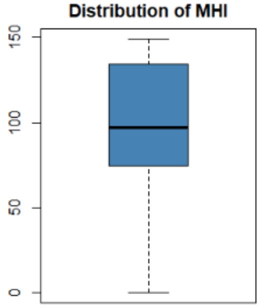
Mental health index

Interesting questions in the survey to assess **mental health** were:

Q6_a4: How anxious did you feel yesterday? (not used in LSI)
Q7: Over the last two weeks, how often, if at all, have you been bothered by not having interest or pleasure in doing things?
Q8: Over the last two weeks, how often have you been bothered by feeling down, depressed, or hopeless?
Q11: How often, if at all, do you feel lonely?







$\boldsymbol{MH} = (Q6_{a4} + Q7 + Q8 + Q11)/2$ $\boldsymbol{MHI} = 100 * MH/mean(MH)$

*The answers to **Q7, 8 and 11** address a similar aspect of mental health issue and are recoded on a **cumulated** and **inverted** scale of **0 to 10** (7: 0 to 3 + 8: 0 to 3 + 11: 0 to 4) whereas **Q6** is answered on a scale of **0 to 10** and is given a **50% weight** for MH.*

Reaching out efficiently | Linear regressions show that the MHI and LSI are correlated with the 6 socio-demographic variables

MHI & LSI are difficult to observe

LSI

Q6_a1: How satisfied are you with your life nowadays?
Q6_a2: To what extent do you feel the things you do in your life are worthwhile?
Q6_a3: How happy did you feel yesterday?

MHI

Q6_a4: How anxious did you feel yesterday?
Q7: Over the last two weeks, how often, if at all, have you been bothered by not having interest or pleasure in doing things?
Q8: Over the last two weeks, how often have you been bothered by feeling down, depressed, or hopeless?
Q11: How often, if at all, do you feel lonely?

Problem: These questions are often hard to answer and require an in-depth survey to be collected

Objective: Identify proxy variables that can be used to approximate the value of the indexes quickly and easily

Methodology: Identify the socio-demographic variables that are significantly correlated with the indexes MHI and LSI through linear regressions

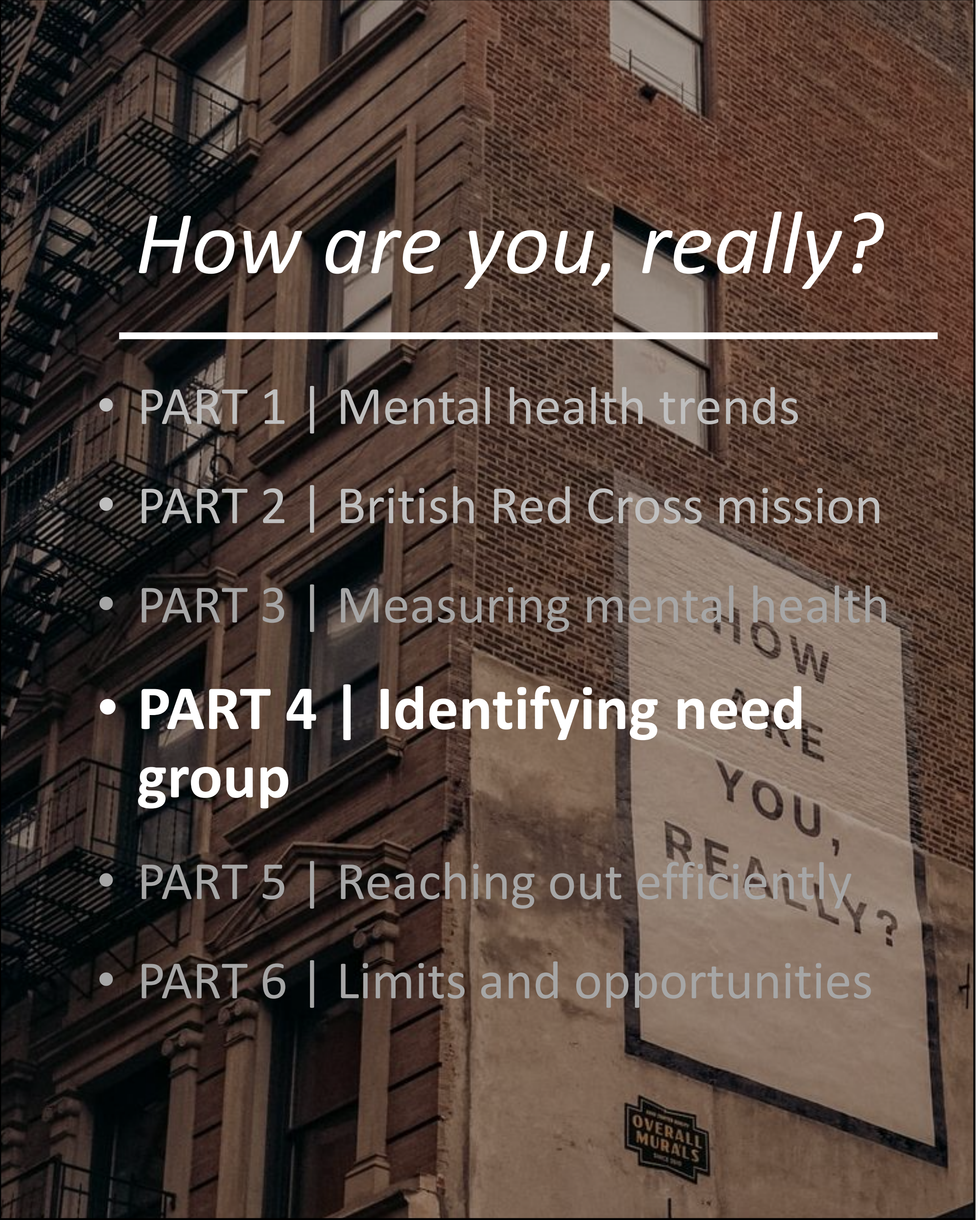
Significant variables from linear regressions

Regress indexes on demographic and geographic variables (c)

$$MHI = \beta_0 + \beta_c Demographic_c$$
$$LSI = \beta_0 + \beta_c Demographic_c$$

LSI			MHI	
↑		Age		↑
↑	Male	Gender		→
↓	Unemployed	Employment		→
↑	Higher Managerial	Working situation	Skilled Manual	↑
→		Region	Greater London North. Ireland	↓
↓	No Children	Age of Youngest Child	Kid & teenager	↓

For complete regression results, see Appendix



How are you, really?

- PART 1 | Mental health trends
- PART 2 | British Red Cross mission
- PART 3 | Measuring mental health
- **PART 4 | Identifying need group**
- PART 5 | Reaching out efficiently
- PART 6 | Limits and opportunities

BritishRedCross

Identifying need group

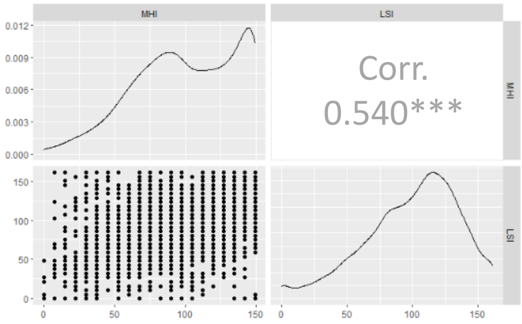
a. Naïve approach to clustering enables to identify help groups

b. Refined clustering confirms the naïve approach with a similar priority help group (low LSI & MHI)

c. The priority need group differs from other cluster in terms of sociodemographic variables

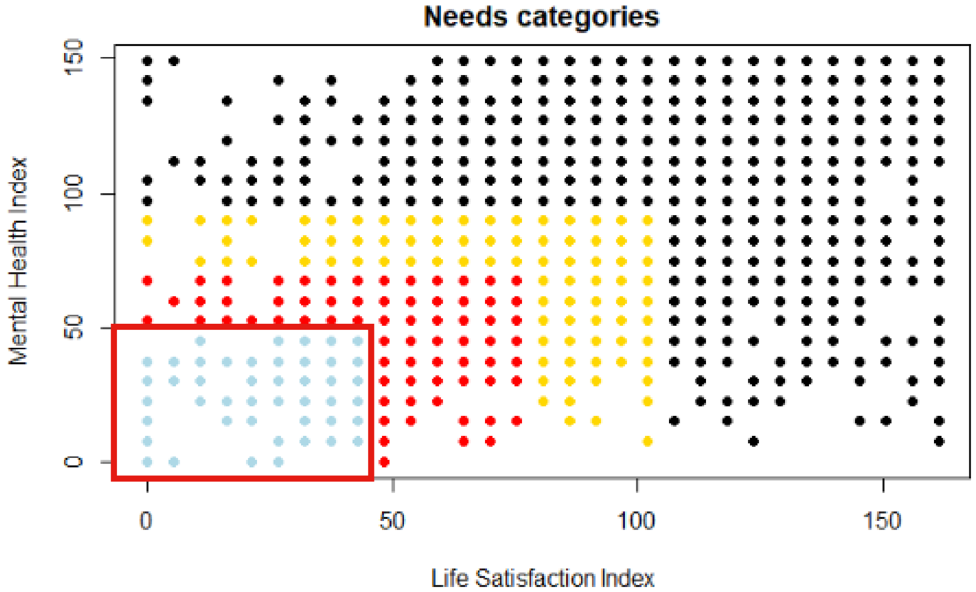
QTEM DATA CHALLENGE

Identifying need group | Initial naïve clustering helps identifying priority help group and shows potential for refined analysis



MHI and LSI are significantly **correlated**, thus we can summarize their information by assigning individuals to **need categories** based on a **naïve quantile clustering**

- MHI & LSI < q(10%) → n = 121 (3.2%)
- MHI & LSI < q(25%) → n = 321 (8.4%)
- MHI & LSI < q(50%) → n = 711 (18.6%)



1. Measure mental health

MHI and LSI can help us to identify specific need groups with higher needs through a subjective naïve clustering.
→ ***Refine the clustering with advanced machine learning techniques (see next slide)***

2. Identify the need group easily

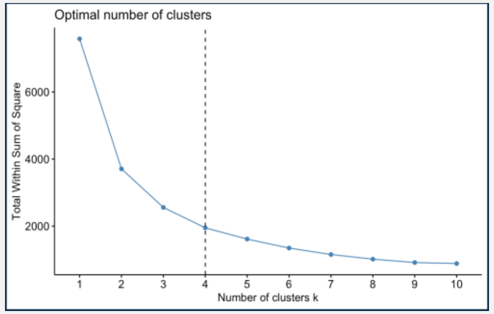
The group with most critical needs can be identified based on the indexes but these are difficult to observe in real life
→ ***Observe difference in socio-demographic variable distribution between the focus group and the rest***

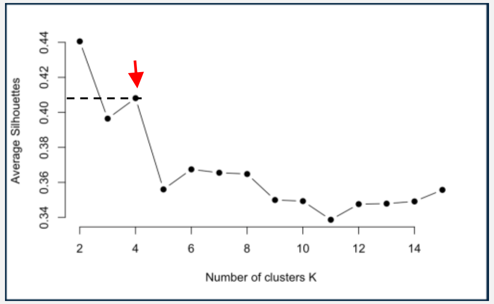
	Rest <dbl>	Focus_Group <dbl>	Gap <dbl>	p-value <dbl>	Significant <dbl>
Female_prop	0.23	0.31	0.08	5.586946e-02	0
Age_mean	52.32	49.11	-3.21	1.813268e-02	1
Minority_prop	0.10	0.13	0.03	3.803345e-01	0
Parent_prop	0.67	0.55	-0.12	8.319006e-03	1
Unemployed_or_Sick_prop	0.04	0.21	0.17	1.754125e-05	1
Urban_Living_Area_prop	0.31	0.39	0.08	8.812769e-02	0
Support_Received_prop	0.10	0.23	0.13	7.941422e-04	1
Waiting_Support_prop	0.06	0.20	0.14	2.271001e-04	1

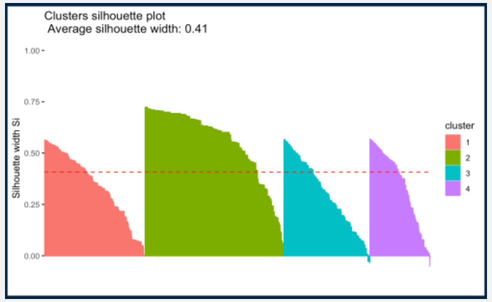
The significant socio-demographic variables can later be used to predict whether the individual belongs to the critical need group (defined by the k-means) based on a probit model (cf. part 5)

Identifying need group | Refined analysis shows that clustering in four groups enables us to best identify the priority help group

1. Optimal order: Silhouette & Elbow







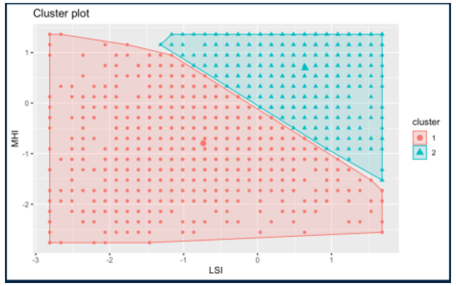
The **optimal** order according to the methods implemented would be **2** clusters, but the clustering would then only focus on "mentally healthy" individuals (*see plot*)

The second-best order appeared to be **4** clusters. This clustering still shows to be **partially optimal** by both methods

Choosing four clusters instead of two enable us to add **depth** to our analysis and better **segment** based on the MHI & LSI

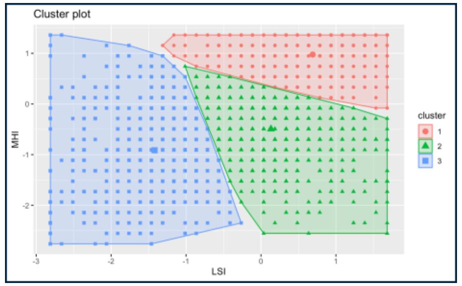
2. Clustering: K-mean of order 4 (sum of square)

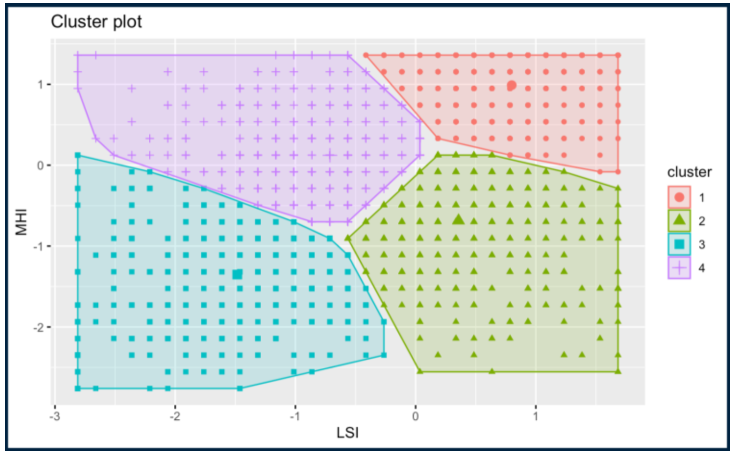
While 2 clusters is analytically optimal, it only enables us to identify individuals with no health issue (*in blue*)



Similarly, if we use 3 clusters, the priority help group will be defined merely based on the LSI as the clustering doesn't allow for differentiation along the MHI axis

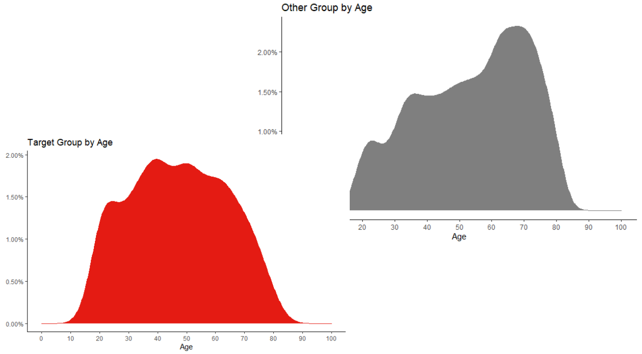
We thus to split the data in 4 clusters as we are able to better explain each cluster meaning and identify the priority help group (*cluster 3*) who are **unsatisfied** with **their life** and **suffering** from **mental disorder**

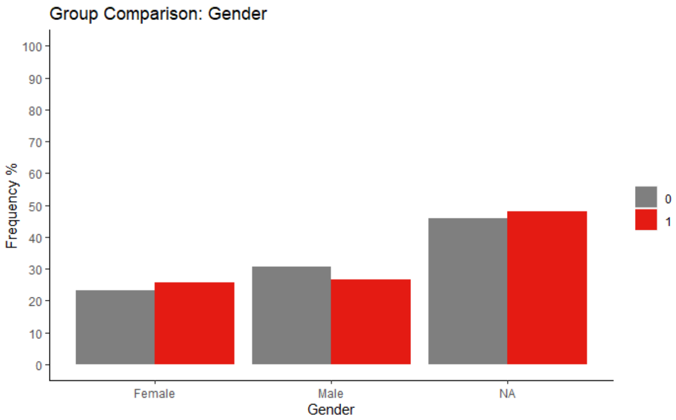


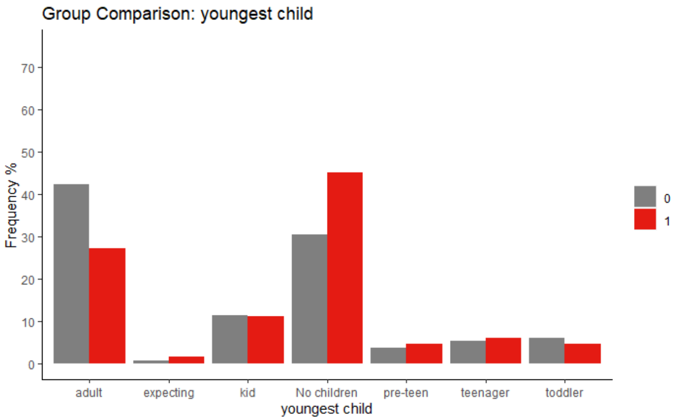


Limits: *While the clustering enables easy interpretation, the* ***high within cluster variance*** *and* ***low between clusters variance*** *make it hard to differentiate individuals between clusters*

Identifying need group | The priority cluster has a higher proportion of female, child-less individuals and younger individuals







1. Age

Interpretation of graphs: while the target group age distribution looks quite normal, the other group is clearly composed for the majority of older individuals. This doesn't allow us to conclude much about the target group, but it can give us insight on the remaining population

Priority group | Rest of population

See appendix for a better view of the graphs

2. Gender

Interpretation of graphs: There are more women and less men in the target group. This is aligned with the theory on gender differences in mental health. Since the difference between group is small, we don't expect this variable to be significant.

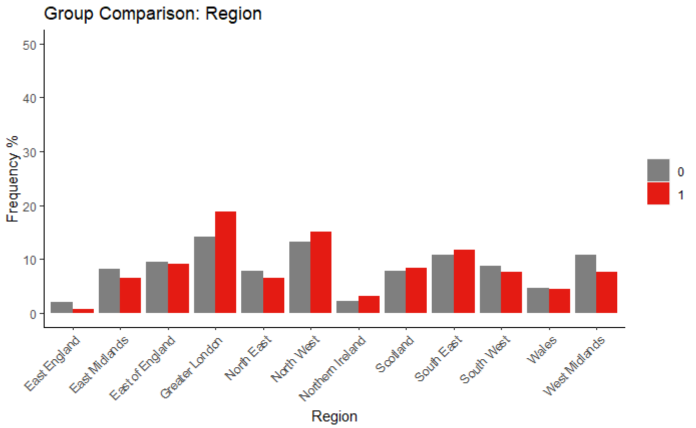
Issues with the data: majority of answers were NAs which makes generalisation less reliable and interpretation more difficult

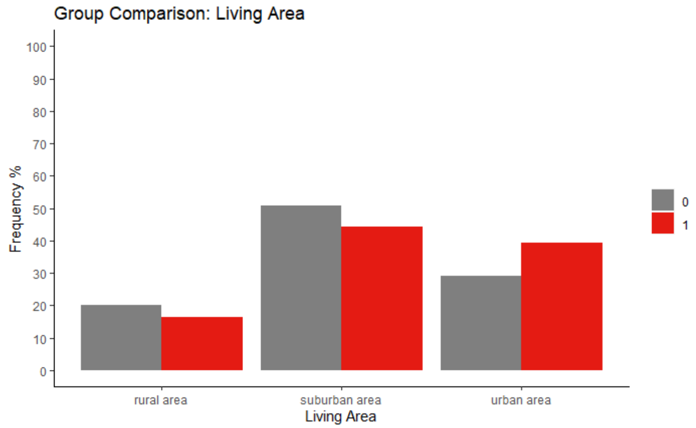
3. Youngest child

Interpretation of graphs: high percentage of child-less individuals in the target group. It could be interesting to explore this further: is being childless the possible cause (e.i. people without children feel lonelier) or consequence of the low mental health (e.i. the person doesn't have children because of mental distress)

Issues with the data: most of the sample had adult kids so the difference in this variable between groups might not be significant

Identifying need group | The priority cluster has a higher proportion of individuals coming from the South and from urban areas





4. Region

Interpretation of graphs: Target group has higher percentage of individuals from London, South east and West. Other regions show similar levels

Issues with the data: the difference in Londoners in the population might not be significant given that the majority of the sample was from London. Furthermore, we observe a high number of NAs

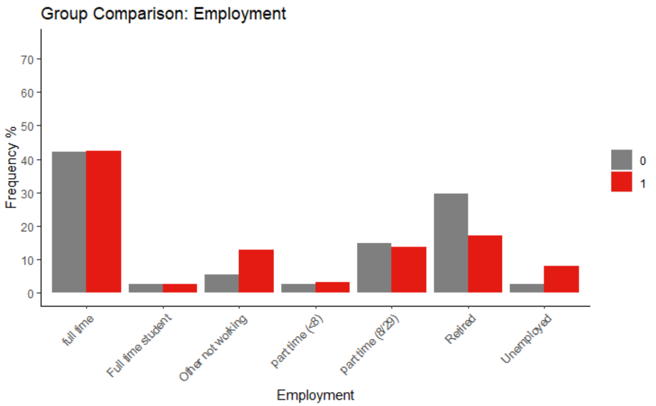
Priority group | Rest of population

See appendix for a better view of the graphs

5. Living Area

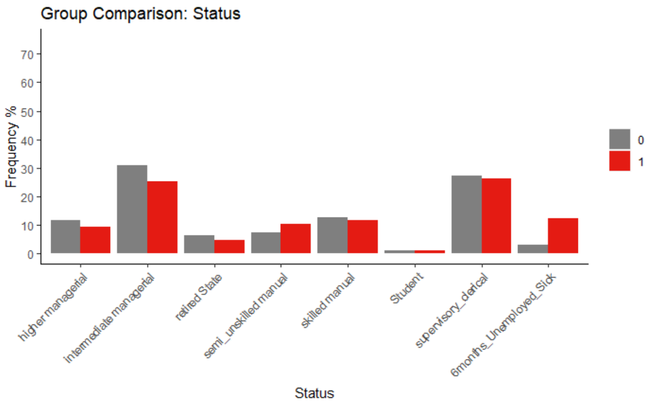
Interpretation of graphs: a bigger proportion of the target group living in urban areas: this could reflect the fact that city life is usually more frenetic and stressful. It might be a consequence of longer commuting time, traffic, transportation and frenzyness of living in an urban area. We believe it could be a topic worth to explore more deeply since it is a growing phenomenon experienced by the majority of the population

Identifying need group | The priority cluster has a higher proportion of unemployed, not working, sick, or low-skilled individuals



6. Employment

Interpretation of graphs: higher percentage of not working and unemployed individuals in the target group. This is not surprising, and it is definitely an area where the British red cross could intervene. Aligned with the other findings, we also see a lower percentage of retired people in the target groups

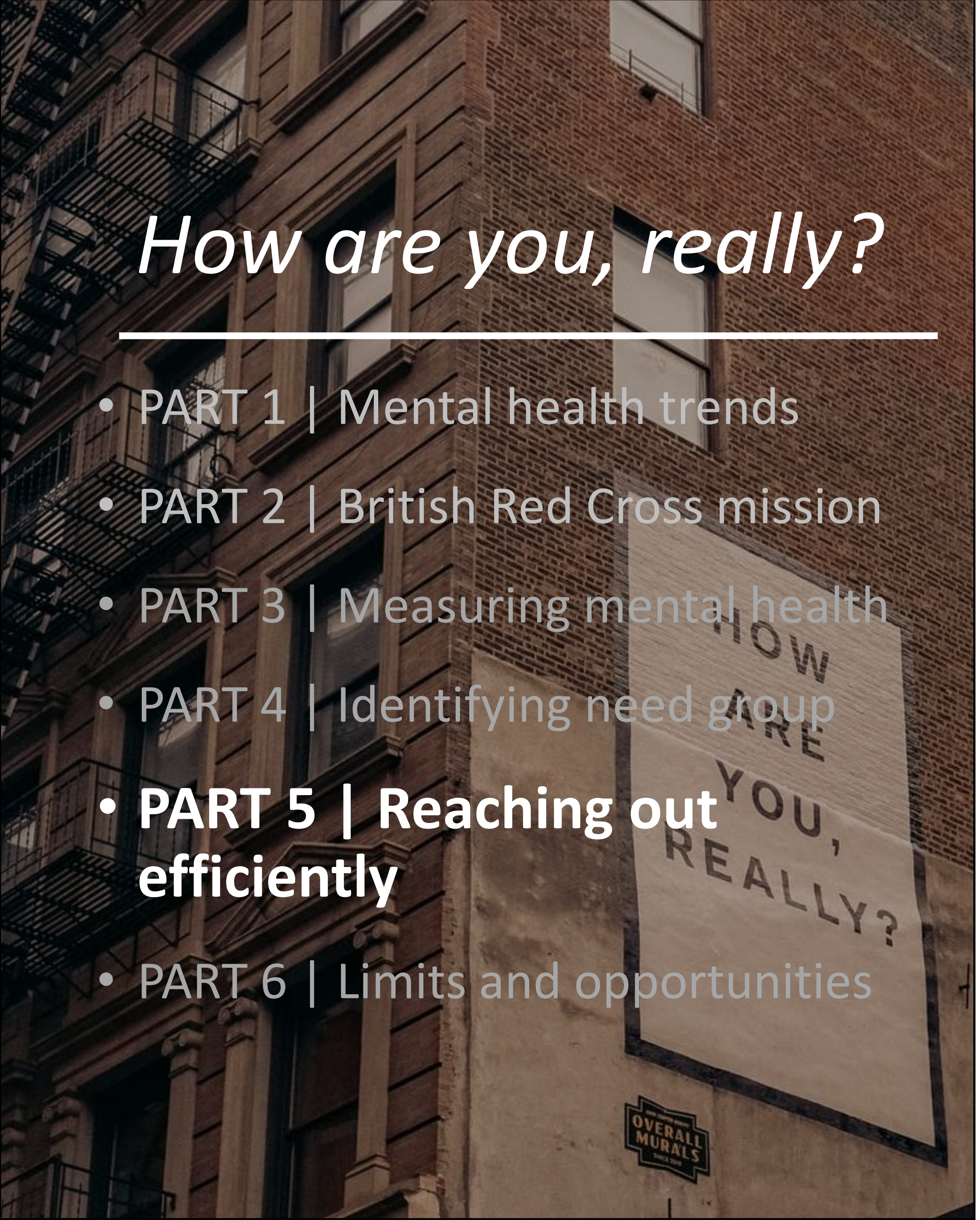


7. Working Status

Interpretation of graphs: Confirming previous findings, we have again higher percentage of target individual une mployed and/or sick.
We can also observe a higher percentage, although not as high, of semi/unskilled worker.
We note that high and intermediate level managers are highly represented in "others" group

Priority group | Rest of population

See appendix for a better view of the graphs



How are you, really?

- PART 1 | Mental health trends
- PART 2 | British Red Cross mission
- PART 3 | Measuring mental health
- PART 4 | Identifying need group
- **PART 5 | Reaching out efficiently**
- PART 6 | Limits and opportunities

British Red Cross

Reaching out efficiently

a. Choice of sociodemographic variables that can be used as proxy to the clustering

b. Probit regression model highlights significant variables

c. The interpretation of the model outcome leads to various propositions to allocate resources

QTEM DATA CHALLENGE

Reaching out efficiently | The explanatory variables for the binary regression models are chosen based on the differences identified

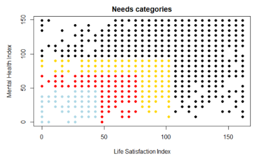
$$LSI = \beta_0 + \beta_i \times Sociodemo_i + \varepsilon$$
$$MHI = \beta_0 + \beta_i \times Sociodemo_i + \varepsilon$$

LSI		MHI
	Age	
Male	Gender	
Unemployed	Working (vs. Full Time)	
Higher Managerial	Social Grade (vs. Unemployed/Sick)	Skilled Manual
	Region	Greater London North. Ireland
No Children	Youngest Child (vs. adult)	Kid & teenager

1. Linear Regression

Differences identified:
The linear regression gives mixed results. Whereas 'Gender', 'Age' and 'Working status' affects the Indexes positively, 'Employment', 'Region' and 'youngest child' have the opposite effect

Variables: Age + Gender + Employment + Working situation + Region and youngest child

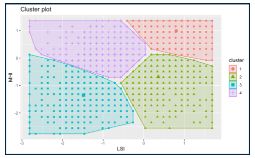


	Rest <dbl>	Focus_Group <dbl>	Gap <dbl>	p-value <dbl>
Female_prop	0.23	0.31	0.08	5.586946e-02
Age_mean	52.32	49.11	-3.21	1.813268e-02
Minority_prop	0.10	0.13	0.03	3.803345e-01
Parent_prop	0.67	0.55	-0.12	8.319006e-03
Unemployed_or_Sick_prop	0.04	0.21	0.17	1.754125e-05
Urban_Living_Area_prop	0.31	0.39	0.08	8.812769e-02
Support_Received_prop	0.10	0.23	0.13	7.941422e-04
Waiting_Support_prop	0.06	0.20	0.14	2.271001e-04

2. Naïve Clustering

Differences identified:
The t test shows a significant difference in means only for the 'Age', the 'Proportion of parents' and the 'Proportion of unemployed or sick' individuals

Variables: Age + Youngest child + Unemployed or sick



Variable
Age
Full time student
Other/Not working
Unemployed
Semi unskilled worker
Unemployed/Sick for more than 6 months
No Children
Toddler

3. K-Means Clustering

Differences identified:
The graphs show interesting differences in 'employment' and 'working status', as well as geographical distribution and age of youngest children, calling for further analysis.

Variables: Age + Ethnicity + Gender + Employment + youngest + Working situation + Region

Reaching out efficiently | Sociodemographic variables can predict the probability of being in the priority group through a probit model

Linear regression $P(Priority = 1) = \beta_0 + \beta_i \times Sociodemo_i + \varepsilon$

Linear regressions reach their limit when we try to predict a binary variable (0: no priority, 1: priority) because the outcomes of the model are unbounded and thus can go from '$-\infty$' to '$+\infty$' which doesn't make sense for a probability

Use a non-linear function to bound the outcome from 0 to 1

Probit regression $P(Priority = 1) = \Phi(\beta_0 + \beta_i \times Sociodemo_i) + \varepsilon$

Thanks to the normal cumulative distribution function $\Phi(Z)$, the outcome of the probit model will be bound between 0 ($\Phi(-\infty)$) and 1 ($\Phi(+\infty)$) which is in line with the binary variable that we try to predict

Apply the logit regression to predict whether the individual belongs to the priority cluster or not

Our probit regression

We will regress the binary variable 'Priority Cluster' based on the sociodemographic variables: *"Age + Ethnicity + Gender + Employment + youngest + Working situation + Region"* as identified in the previous slide

```
Wald test:
----------

Chi-squared test:
X2 = 12.3, df = 1, P(> X2) = 0.00045
```

V

```
fitting null model for pseudo-r2
          llh       llhNull           G2      McFadden
-1.516994e+03 -1.630409e+03 2.268304e+02 6.956242e-02
         r2CU
1.006834e-01
```

?

1. Significance of the model

The Wald test enables us to reject the null hypothesis that no sociodemographic variable is significant in predicting whether the individual belongs to the priority cluster ***pvalue (0.00045) < α(0.05)***

2. Predictive power benchmark

The predictive power of our model can be assessed through the **pseudo-R^2 of 0.101.** Our logistic regressions showed a predictive power of 0.078 and 0.148 for the two indexes.

The **average R^2** in social studies and psychology models is **0.19,** our model falls just below the 25th percentile (Gignac & Szodorai, 2016). Therefore, while our model predictive power is below average, it is still relevant

→ We believe that by including some other variables and improving the current survey the BRC would be able to achieve a higher predictive power (cf. part 6)

Reaching out efficiently| 14 sociodemographic variables are identified as significant for predicting the cluster of the individual

Goal of the regression: Predict the probability of belonging to the priority cluster based on the sociodemographic variables

Mathematical Expression: $P(Priority = 1) = \Phi(\beta_0 + \beta_i \times X_i) + \varepsilon$ which is bound between 0 (at $\Phi(-\infty)$) and 1 (at $\Phi(+\infty)$)

Interpretation: Based on all the sociodemographic variables, we can compute a score Z for the individual ($= \beta_0 + \beta_i \times X_i$) which gives us the probability of belonging to the priority cluster when fed into the normal cumulative distribution function ($\Phi(Z)$). Thus a positive coefficient will increase the Z score and the probability to belong to the priority group whereas a negative coefficient will decrease the Z score and the probability to belong to the priority group .

Variable	Level of significance (p value)	Estimate
Age	5.89e-10	-0.016
Full time student	0.022	-0.423
Other/Not working	0.001	0.409
Unemployed	0.025	0.304
Semi unskilled worker	0.031	0.259
Unemployed/Sick for more than 6 months	2.48e-05	0.643
No Children	0.021	0.175
Toddler	0.029	-0.296

Intercept modalities: Woman, Minority, Full-time worker, Adult children, High manager and East Midland (as reference point – significance: 0.001, estimate: -1.067)

Reaching out efficiently | 14 sociodemographic variables are identified as significant for predicting the cluster of the individual

Variable	Level of significance (p value)	Estimate
East of England	0.030	0.542
Greater London	0.011	0.622
North East	0.067	0.464
North West	0.013	0.611
Northern Ireland	0.025	0.632
Scotland	0.025	0.562
South East	0.016	0.594
South West	0.054	0.484
Wales	0.071	0.0478

Intercept modalities: Woman, Minority, Full-time worker, Adult children, High manager and East Midland (as reference point – significance: 0.001, estimate: -1.067)

The **logit binary regression** uses another non-linear transformation of the score: $\frac{exp(Z)}{1+exp(Z)}$ instead of $\Phi(Z)$.

→ The results of the regression are similar to those displayed above and thus **confirm** the results of the probit model.

Reaching out efficiently| Allocate resources to young and low skilled workers to alleviate stress and financial unsecurity

1. Non-student young people

Interpretation: while the **age** variable predicts **worst mental health** for younger individuals, the variable **Employment / full time student** has a **negative** coefficient which means it decreases likelihood for the individual to be in the target group. Our suggestion is that young workers might suffer from **stress** and **employment uncertainty**, especially if they are **low-skilled**.

Proposal: Career development support to guide young generations through job application and interviews, economical help to allow for independence and affordable psychological help to deal with uncertainty and stress

2. Low skilled workers

Interpretation: semi and **unskilled workers**, as well as **not working** individuals are shown to have higher probability to be in the target group. The reasons could be multiple: jobs might be **less rewarding**, both economically and personally. They could be at higher risk to be **automated** and workers might feel anxious about their unstable working status.

Proposal: Professional training and further formation courses to allow people to develop new skills and possibly improve their working situation

Reaching out efficiently| The impact of other significant variables should be further studied to implement relevant policies

3. Unemployed or Sick

Interpretation: **Biggest impact** on the probability of falling into the need group. The inability to work can lower **self-esteem** and the sense of **belonging** whereas sickness can put one's **work security** at risk

Proposal: Improve employability for a better access to jobs and ensure health care support and job stability for workers on sick leave

4. Toddlers vs no children

Interpretation: While having a **toddler decreases** likelihood of experiencing **negative emotions** such a as loneliness and anxiety, having **no children increases it**. Our interpretation is that many people might be unable, but willing, to raise children

Proposal: Offer economical support to new parents, create affordable childcare facilities, further investigate childless individuals

5. South region

Interpretation: Multiple **specific regions** inhabitants appear more likely to need help. This could be due to differences in **cultural** and **institutional** factors but it might be partially due to **sample bias** (majority from London)

Proposal: This initial observation could spark a wider research on the impact of the geographical location on mental health needs

How are you, really?

- PART 1 | Mental health trends
- PART 2 | British Red Cross mission
- PART 3 | Measuring mental health
- PART 4 | Identifying need group
- PART 5 | Reaching out efficiently
- **PART 6 | Limits and opportunities**

British Red Cross

Limits and opportunities

a. Understand how the unbalanced dataset limits our analysis and interpretations

b. Offer solutions to enrich the survey and gain more insights

c. Extend the research agenda with further research paths and internal communication ideas

QTEM DATA CHALLENGE

Limits and opportunities | Imbalance dataset proves to be a limit for our analysis and interpretations in multiple ways

Imbalanced dataset
High number of NAs and "prefer not to say" answers, especially in personal questions, gender identification and region

Solution: Increase sample selection and diversity

Solution: Include more variables, based on literature best practice in mental health measures

Logistic regression's predictive power
The logit and probit models show small sized pseudo-R. Similarly, both linear models have below-average effect size

Clustering relevance
High within cluster variance and low between cluster variance. This is also due to complexity of mental health phenomenon

Usage of predictive model
The model is not yet able to predict well on a test sample – more data is needed, as the target group is quite small and sample is incomplete

Result interpretability
For both 'youngest kid' and 'region' variables, our interpretation need further analysis with more complete data to carefully evaluate significance of findings

Solution: improve question formulation + communicate results internally with interactive Dashboards

Limits and opportunities | Enriching and extending the survey would enable to make our analysis more relevant

Imbalanced dataset
High number of NAs and "prefer not to say" answers, especially in personal questions, gender identification and region

Solving the lack of information to improve the regression power and clustering quality

Our clustering and probit model show promising results but are not usable on the field yet because the imbalanced dataset makes it hard to discriminate between people who need or not → Improve the **completeness** of the answers + Add more **information** in the survey

How to get there?

1. Improvement to the survey (completeness)

Tackling the problem of numerous NAs/Prefer not to say:

→ Socio-demographic items: questions could be made mandatory and offer more modalities to avoid NA's (especially Gender, Region and Ethnicity)

→ Personal questions: splitting question in two (general and detailed), giving the possibility to answer in more details to those who feel more comfortable

→ Partial data regarding children: rephrasing questions so that the respondent can clarify the number of kids and the age of each child

2. Inclusion of new variables (information)

Based on other studies and similar surveys, we suggest to include the following variables:

→ Risk status of profession (Nikčević et al., 2020)

→ Income level (Jacob et al., 2020)

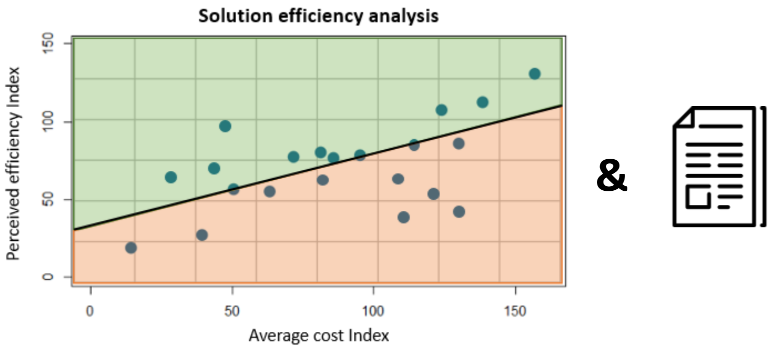
→ Physical issues – detail current survey (Jacob et al., 2020)

→ Smoking and alcohol consumption habits (Jacob et al., 2020)

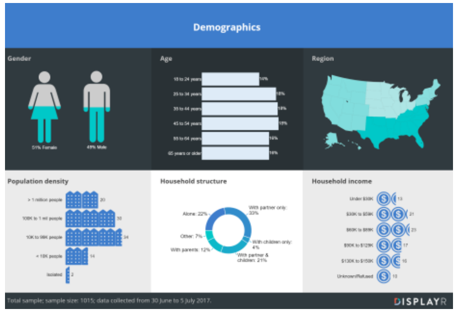
→ Family situation & Marital status (Jacob et al., 2020)

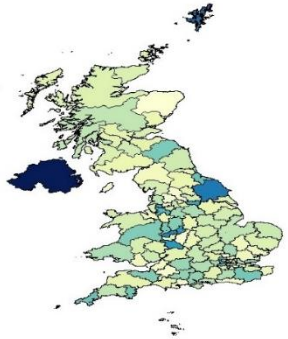
Limits and opportunities | Analyse the efficiency of resource used after allocation and better communicate the needs internally

Better usage of the predictive models



Better result interpretability





Additional research on the efficiency of the different help means for optimal allocation

Why? The regression analysis has only a limited power of interpretation because it focuses on correlation
→ Further studies will help understand if there is a causal link
What? Confirm the causal link behind the correlation we observe with our linear regressions (*see regression results in Appendix*).
How? More sophisticated methods supported by qualitative insights from theoretical studies could confirm our results and make them more robust

Internal communication of the priorities through a dynamic dashboard

Why? A Dashboard is an interactive way to convey information, it is an easy and appealing way to communicate the results to a broader audience
What? If someone wants to know the information of his region, they can hover over a map and the information pops up
How? Different kinds of software like R give the opportunity to create an interactive dashboard, for example based on a UK Map to show differences in regional trends

Take away | Clustering and predictive model show great initial insights but the input and models should be refined to perform well

Clustering individuals in need groups

Models: naïve clustering, elbow and silhouette method, kmeans

Key results: creation of mental health measures (MHI & LSI), clustering of sample and identification of "target" group (15,5%)

Socio-demographics of the need group

Models: t-test, distribution graphs, linear regressions

Key results: analysis of 7 socio-demographic variables in target group and of factors influencing the MHI & LSI created

Predicting needs with observable variables

Models: logit and probit regressions

Key results: creation of a predictive model based on significant socio-demographic variables, identification of 5 possible areas of actions

Limits of our research and next steps

Key results: limits of our analysis due to imbalanced dataset, suggestions for fut ure surveys: inclusions of 7 new variables, discussion of next steps- efficiency of resource allocation and dashboard creation

Thank you for your interest in our work!

Demarets Mathieu
Student at the Solvay Brussels School (ULB – BE), I have a passion to vulgarize statistical insights to make them understandable for the greater number

Galletta Erika
Student at Politecnico di Milano, I want to create value through my work, not only for potential customers but for people first and foremost

Moreira Laura
School of Economics and Management of the University of Porto (FEP), one of the reasons I wanted to join QTEM was to explore data analytics and understand what's behind the numbers.

Rösch Florian
Student at Johann Wolfgang Goethe-University Frankfurt am Main, I am curios how big data and new statistical models can further improve our understanding of the world

ANNEXES

a. **Sources** for the literature review and the model benchmarking

b. **Description** of the demographic variables and simple linear regressions

c. **Plots** of the imbalances, of the demographics of the cluster and of the logit and probit cumulative probabilities

Sources

A) Literature Review:

Leogrande, A. (2022). The Mental Health Index in the Italian Regions. *Munich Personal RePEc Archive*.

Gana, K., Bailly, N., Saada, Y., Joulain, M., & Alaphilippe, D. (2012). Does Life Satisfaction Change in Old Age: Results From an 8-Year Longitudinal Study. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *68*(4), 540–552.

Das, R., Hasan, M. R., Daria, S., & Islam, M. R. (2021). Impact of COVID-19 pandemic on mental health among general Bangladeshi population: a cross-sectional study. *BMJ Open*, *11*(4), e045727.

Savage, M. J., James, R., Magistro, D., Donaldson, J., Healy, L. C., Nevill, M., & Hennis, P. J. (2020). Mental health and movement behaviour during the COVID-19 pandemic in UK university students: Prospective cohort study. Mental Health and Physical Activity, 19, 100357.

Jacob, L., Tully, M. A., Barnett, Y., Lopez-Sanchez, G. F., Butler, L., Schuch, F., López-Bueno, R., McDermott, D., Firth, J., Grabovac, I., Yakkundi, A., Armstrong, N., Young, T., & Smith, L. (2020). The relationship between physical activity and mental health in a sample of the UK public: A cross-sectional study during the implementation of COVID-19 social distancing measures. Mental Health and Physical Activity, 19, 100345.

Sources

Nikčević, A. V., Marino, C., Kolubinski, D. C., Leach, D., & Spada, M. M. (2021). Modelling the contribution of the Big Five personality traits, health anxiety, and COVID-19 psychological distress to generalised anxiety and depressive symptoms during the COVID-19 pandemic. Journal of Affective Disorders, 279, 578–584.

Healy, K. (2016). A theory of human motivation by Abraham H. Maslow (1942). The British Journal of Psychiatry, 208(4), 313-313.

B) About R square significance:

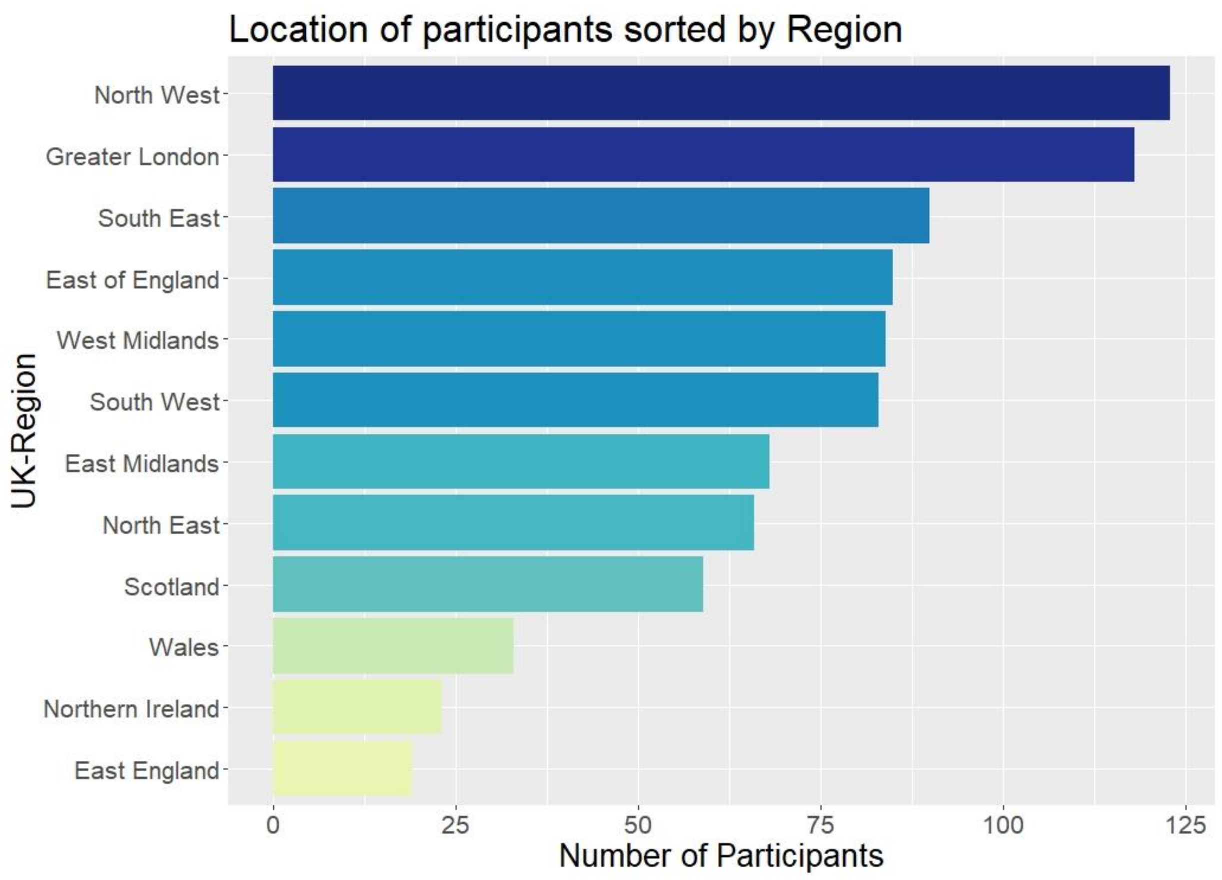
Richard et al., 2003 - One hundred years of social psychology quantitatively described.

Fraley & Marks, 2007 - The null hypothesis significance testing debate and its implications for personality research.

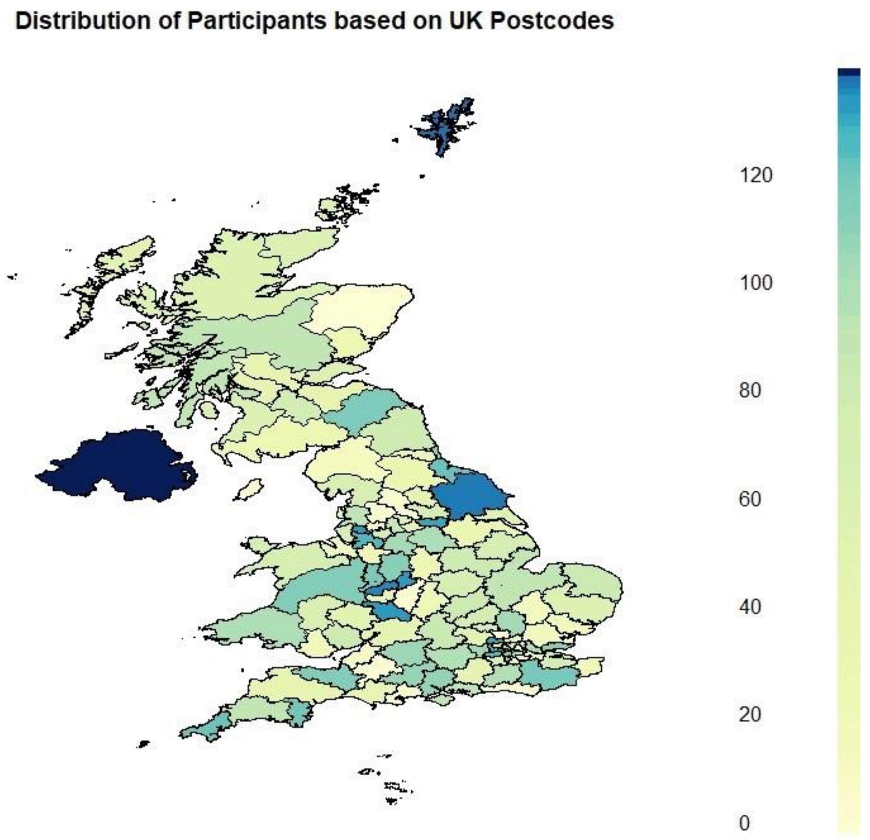
Gignac & Szodorai, 2016 - Effect size guidelines for individual differences researchers.

Description| Variable description – Regional Distribution

- Distribution of all participants sorted by region



- Distribution of all participants sorted by 2-digit UK Postcode



Description| Variable description – demographic data

VARIABLE	FEATURES
Gender (Categorical/Dummy)	Male, Female, not_available (NA)
Age (Numerical)	1, …, 80
Employment (Categorical/Dummy)	**Full time**: working full time (30 or more hours per week) **Full time student**: full time student **Part time (< 8):** working part time (less than 8 hours A week) **Part time (8/29):** working part time (8 - 29 hours per week) **Retired**: retired **Unemployed**: unemployed **Other not working**: other not working

VARIABLE	FEATURES
Social Grade (Categorical/Dummy)	**Higher managerial**: higher managerial/ professional/ administrative (e.G. Established doctor, solicitor, board director in large organisation (200+ employees), top level civil servant/ public service employee, head teacher etc.) **Intermediate managerial**: intermediate managerial/ professional/ administrative (e.G. Newly qualified (under 3 years) doctor, solicitor, board director of small organisation, middle manager in large organization, principal officer in civil service/ local government etc."), **Retired state**: retired and living on state pension only **Semi_unskilled** manual: semi-skilled or unskilled manual worker (e.G. Manual jobs that require no special training or qualifications, apprentices to be skilled trades, caretaker, cleaner, nursery school assistant, park keeper, non-hgv driver, shop assistant etc.) **Skilled manual**: skilled manual worker (e.G. Skilled bricklayer, carpenter, plumber, painter, bus/ ambulance driver, hgv driver, unqualified teaching assistant, pub/ bar worker etc.) **Supervisory_clerical**: supervisory or clerical/ junior managerial/ professional/ administrator (e.G. Office worker, student doctor, foreman with 25+ employees, sales person, student teacher etc.) **6months_unemployed**_sick: unemployed for over 6 months or not working due to long term sickness **Student**: student

Description | Variable description – demographic data

VARIABLE	FEATURES
Living area type **(Categorical/Dummy)**	**Suburban area**: suburban area - residential areas on the outskirts of cities and towns **Urban area**: urban area - cities or towns **Rural area**: rural area - villages or hamlets
Region **(Categorical/Dummy)**	**East midlands, east of england, greater london, north east, north west, northern ireland, scotland, south east, south west, wales, west midlands, wales** Based on: https://ideal-postcodes.Co.Uk/guides/postcode-areas
Ethnicity **(Catgorical/Dummy)**	**Minority**: includes: asian or asian british, black, african, caribbean or black british, mixed or multiple ethnic groups, other ethnic group **White**: **Don't think of myself as any of these:** **Prefer not to say:**

Description| Simple regression results – demographic data

Regression of MHI and LSI only on demographic variables:

$$LSI_i = \beta_0 + \beta_{ic}Control_{ic} + \epsilon_i$$

$$MHI_i = \beta_0 + \beta_{ic}Control_{ic} + \epsilon_i$$

where i = participant, c = {Control Variable}

Only significant coefficients are stated in the following

Description| Simple regression results – demographic data

c	Question_Answer	Feature	α < 5% - LSI	α < 5% - MHI
c.1	Male	Male		6.529
I	Female	Femle	Intercept	
c.3	Not_available	NA		
c.4	Age	1, …, 80	0.298	0.609
I	full time	Working full time (30 or more hours per week)	Intercept	
c.5	full time student	Full time student		
c.6	part time (<8)	Working part time (Less than 8 hours a week)		
c.7	part time (8/29)	Working part time (8 - 29 hours per week)		
c.8	Retired	Retired		
c.9	Unemployed	Unemployed		
c.10	Other not working	Other not working	-13.424	

Description| Simple regression results – demographic data

c.11	higher managerial	Higher managerial/ professional/ administrative (e.g. established doctor, solicitor, board director in large organisation (200+ employees), top level civil servant/ public service employee, head teacher etc.)	25.039	21.297
c.12	intermediate managerial	Intermediate managerial/ professional/ administrative (e.g. newly qualified (under 3 years) doctor, solicitor, board director of small organisation, middle manager in large organization, principal officer in civil service/ local government etc."),	18.872	23.703
c.13	retired state	Retired and living on state pension only	13.236	15.679
c.14	semi_unskilled manual	Semi-skilled or unskilled manual worker (e.g. manual jobs that require no special training or qualifications, apprentices to be skilled trades, caretaker, cleaner, nursery school assistant, park keeper, non-HGV driver, shop assistant etc.)	15.087	20.665
c.15	skilled manual	Skilled manual worker (e.g. skilled bricklayer, carpenter, plumber, painter, bus/ ambulance driver, HGV driver, unqualified teaching assistant, pub/ bar worker etc.)	19.592	23.038
c.16	supervisory_clerical	Supervisory or clerical/ junior managerial/ professional/ administrator (e.g. office worker, student doctor, foreman with 25+ employees, sales person, student teacher etc.)	17.153	21.592
I	6months_unemployed_sick	Unemployed for over 6 months or not working due to long term sickness	Intercept	
c.17	Student	Student		

Description| Simple regression results – demographic data

c.18	suburban area	suburban area - residential areas on the outskirts of cities and towns		
c.19	urban area	Urban area - cities or towns		
I	rural area	Rural area - villages or hamlets		
c.20	East Midlands	East Midlands		
c.21	East of England	East of England		
c.22	Greater London	Greater London		-11.629
c.23	North East	North East		
c.24	Northern Ireland	Northern Ireland		-16.023
c.25	Scotland	Scotland		
c.26	South East	South East		
c.27	South West	South West		
c.28	Wales	Wales		
c.29	West Midlands	West Midlands		
I	North West	North West		
c.30	youngest_expecting	Kind of youngest child: expecting		
c.13	youngest_kid	Kind of youngest child: kid		-8.303
c.14	youngest_no children	Kind of youngest child: no children	-6.443	
c.15	youngest_pre-teen	Kind of youngest child: pree-teen		
c.16	youngest_teenager	Kind of youngest child: teenager		-7.670
c.17	youngest_toddler	Kind of youngest child: toddler		
I	youngest_adult	Kind of youngest child: adult	Intercept	

Description | Variable description and simple regression results – question data – Q1 – Multiple Choice

Regression of LSI and MHI on all variables controlling for demographic data:

$$LSI_i = \beta_0 + \beta_{ic}Control_{ic} + \beta_{ia}Answer_{ia} + \epsilon_i$$

$$MHI_i = \beta_0 + \beta_{ic}Control_{ic} + \beta_{ia}Answer_{ia} + \epsilon_i$$

where i = participant, c = {Control Variable}, a = {Question Answer}

Only significant coefficients are stated in the following

Description | A first data analysis – Simple linear regression – What leads to a high LSI or MHI?

Experiences

LSI

Good Experience (Q1):
- None of the options/ Absence of problems

Bad Experience:
- Serious debt

MHI

Good Experience (Q1):
- None of the options could mean that background information is missing and relevant

Bad Experience:
- Getting behind on bills
- Serious physical health problems
- Domestic abuse or violence
- Alcohol or drug problems

Help

LSI

Where should help come from (Q2):
- Local welfare fund
- Organisations supporting migrants
- None
- Not applicable – nothing received

Where help should not come from:
- Prefer not to say -> missing background information?

MHI

Where should help come from (Q2):
- Not applicable – nothing received/ no help was neccessary

Where help should not come from:
- Other relatives
- Food banks
- Day centre and drop in centre
- Organisations supporting migrants

For complete regression results, see Appendix

Description| A first data analysis – Simple linear regression – What leads to a high LSI or MHI?

Daily Life, Mobility and Usual Acitvities

Daily Activity, Mobility and Usual activities (Q3, Q4, Q5):

- Limitations in daily activities strongly decreases the indices
- Mobility problems are only relevant if they are extreme
- If problems in usual activities become more severe, they become more influential
- strong positive significant value for “Prefer not to say” -> missing background information?

Receiving and Waiting for Mental Health Support

Received and waiting for mental health (Q9, Q10):

- Received mental health in the past three month has no influence on the LSI (no significance)
- In contrast received mental health strongly decreases the MHI

  “Prefer not to say” is decreasing the MHI
- Waiting for mental health is decreasing both indices, but especially the MHI decreases very strong
- “Prefer not to say” is decreasing both indices by a similar amount -> missing background information?

For complete regression results, see Appendix

Description| Variable description and simple regression results – question data – Q1 – Multiple Choice

Q1: In the last three months, have you experienced any of the following? Please tick all that apply.

Only significant coefficients are stated

a	Question_Answer	Feature	α < 5% - LSI	α < 5% - MHI
1.1	Q1_a1	Benefit sanctions (e.g. benefits being stopped or reduced)		
1.2	Q1_a2	Delays to benefit payments		
1.3	Q1_a3	Getting behind on bills		-11.882
1.4	Q1_a4	Getting behind on rent or mortgage payments		
1.5	Q1_a5	Serious debt	-18.724	
1.6	Q1_a6	Being evicted from your home		
1.7	Q1_a7	Applying to the council or Northern Ireland Executive as homeless or being threatened with homelessness		
1.8	Q1_a8	Serious physical health problems		-14.107
1.9	Q1_a9	Domestic abuse or violence		-16.805
1.10	Q1_a10	Alcohol or drug problems		-12.705
1.11	Q1_a11	Coming to the UK to live		
1.12	Q1_a12	Problems with your right to live or work in the UK		
1.13	Q1_a13	None of the above	11.048	6.720

Description | Variable description and simple regression results – question data – Q2 – Multiple Choice

Q1: In the last three months, have you received money and/or any non-cash items (such as food, clothing, toiletries, prepaid cards for utilities such as energy, or other items) from any of the following? Tick all that apply.

a	Question_Answer	Feature	α < 5% - LSI	α < 5% - MHI
2.1	Q2_a1	Universal Credit/benefits/social security		
2.2	Q2_a2	Parents		
2.3	Q2_a3	Other relatives		-11.103
2.4	Q2_a4	Friends		
2.5	Q2_a5	Colleagues		
2.6	Q2_a6	Charities/churches		
2.7	Q2_a7	Food banks		-13.794
2.8	Q2_a8	Local welfare fund, if it exists (run by the council)	23.697	
2.9	Q2_a9	Paid work (including cash-in-hand work)		
2.10	Q2_a10	Begging		
2.11	Q2_a11	Advice service (e.g. Citizens Advice, money advice, welfare advice, etc.)		
2.12	Q2_a12	Day centre or drop-in centre		-27.403
2.13	Q2_a13	Organisations supporting migrants	24.745	-22.807
2.14	Q2_a14	Other (please specify)		
2.15	Q2_a15	None of the above	7.801	
2.16	Q2_a16	Not applicable – I have no received money or support from additional sources	7.508	8.697
2.17	Q2_a17	Prefer not to say	-21.228	

Description| Variable description and simple regression results – question data – Q3 and Q4 – Single Choice

Q3: Are your day-to-day activities limited because of a health problem or disability which has lasted, or is expected to last, at least 12 months? Include problems related to old age.

a	Question_Answer	Feature	α < 5% - LSI	α < 5% - MHI
3.1	lot	Yes, limited a lot	-5.465	-15.251
3.2	little	Yes, limited a little	Intercept	
3.3	no	No	12.514	15.183

Q4: Please tick the response that best describes your health today regarding mobility.

a	Question_Answer	Feature	α < 5% - LSI	α < 5% - MHI
4.1	no	I have no problems in walking about	22.101	18.733
4.2	slight	I have slight problems in walking about		
4.3	moderate	I have moderate problems in walking about		
4.4	severe	I have severe problems in walking about		
4.5	extreme	I have extreme problems in walking about	Intercept	
4.6	prefer not to say	Prefer not to say	22.101	

Description| Variable description and simple regression results – question data – Q5 and Q9 – Single Choice

Q5: Please tick the response that best describes your health today regarding usual activities (e.g. work, study, housework, family or leisure activities).

a	Question_Answer	Feature	α < 5% - LSI	α < 5% - MHI
5.1	no	I have no problems doing my usual activities	32.851	30.039
5.2	slight	I have slight problems doing my usual activities	16.841	14.181
5.3	moderate	I have moderate problems doing my usual activities	11.998	
5.4	severe	I have severe problems doing my usual activities		
5.5	extreme	I have extreme problems doing my usual activities	Intercept	
5.6	prefer not to say	Prefer not to say	21.489	17.589

Q9: Have you received support for your mental health in the past three months?

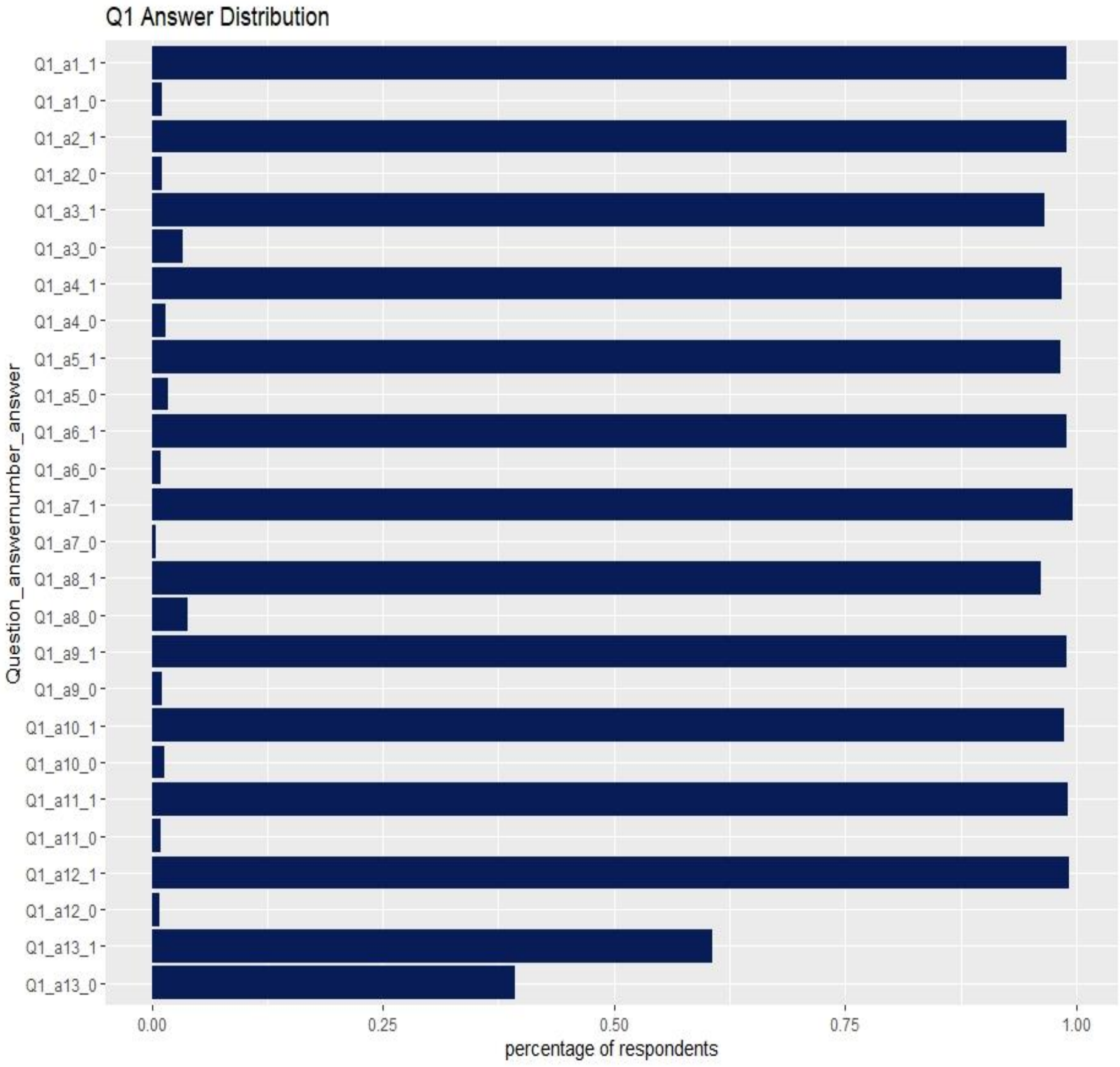
a	Question_Answer	Feature	α < 5% - LSI	α < 5% - MHI
9.1	Yes	Yes		-23.760
9.2	No	No	Intercept	
9.3	Prefer not to say	Prefer not to say		-16.339

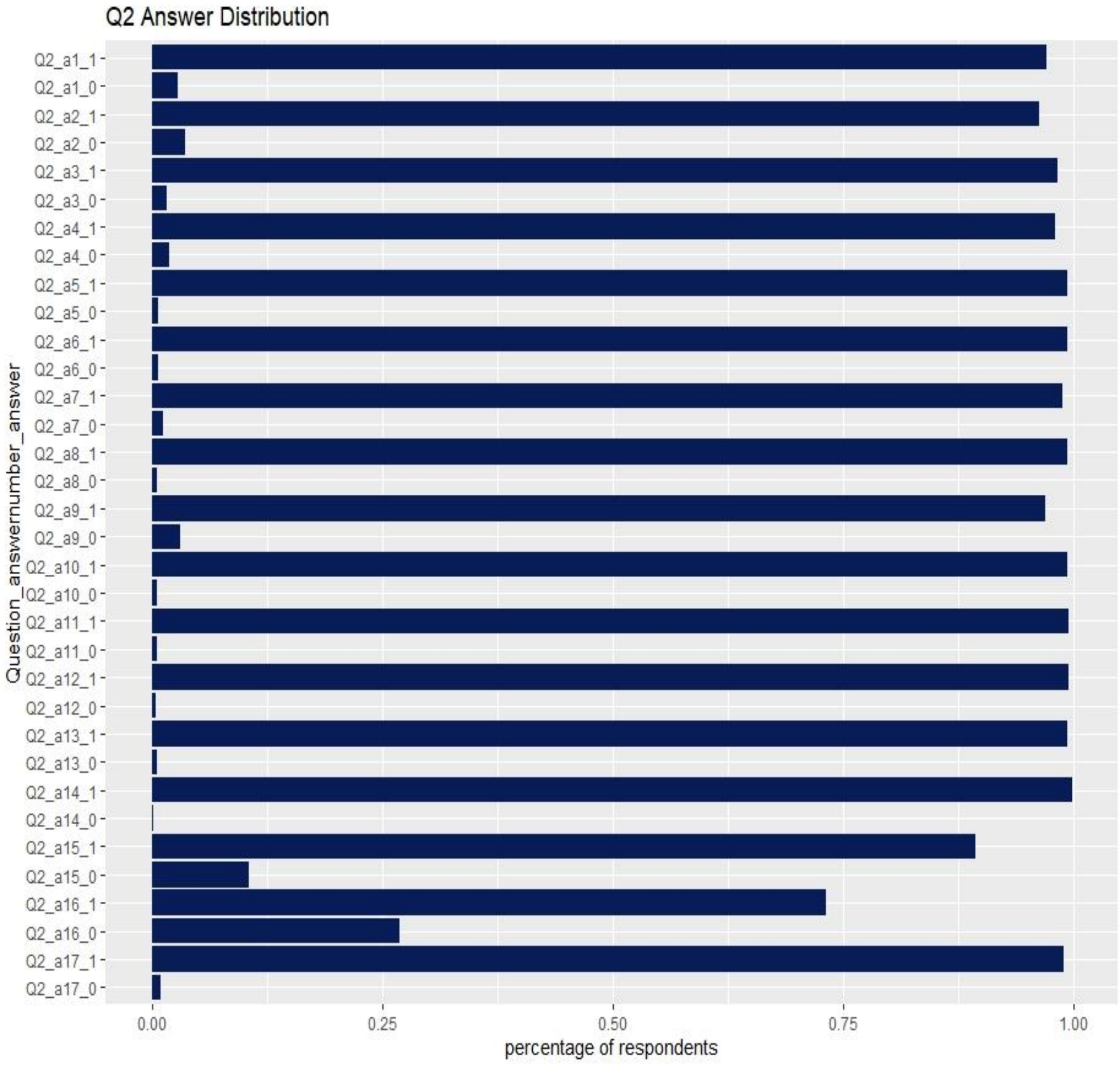
Description| Variable description and simple regression results – question data – Q10 – Single Choice

Q10: Are you currently waiting for support with your mental health?

a	Question_Answer	Feature	α < 5% - LSI	α < 5% - MHI
10.1	Yes	Yes	-7.951	-33.003
10.2	No	No	Intercept	
10.3	Prefer not to say	Prefer not to say	-22.445	-25.729

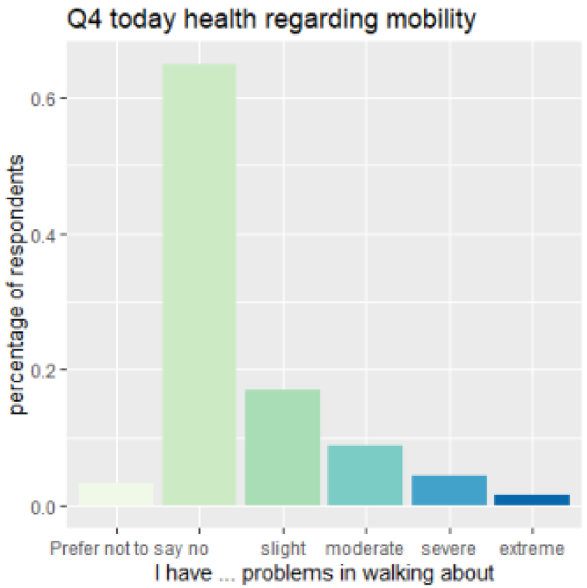
Description| Answer frequency of Q1 and Q2

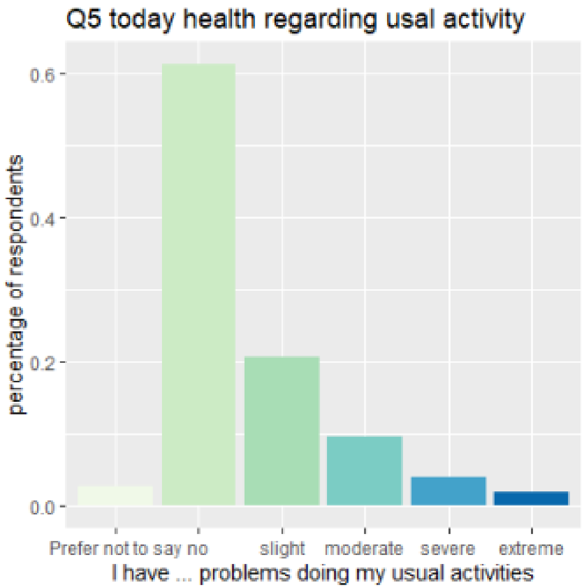


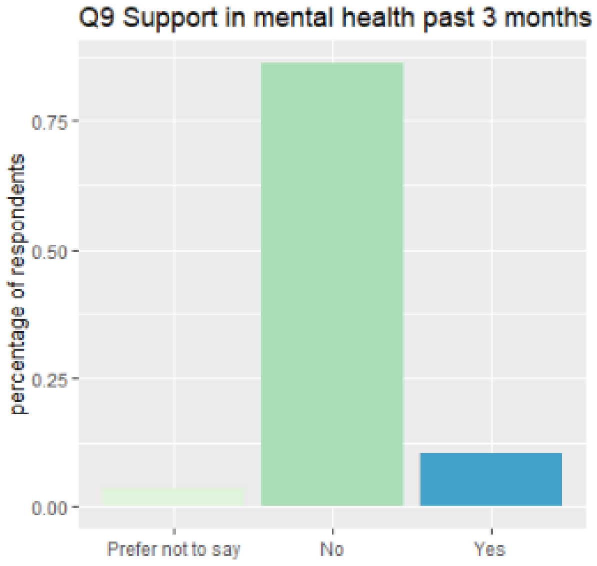


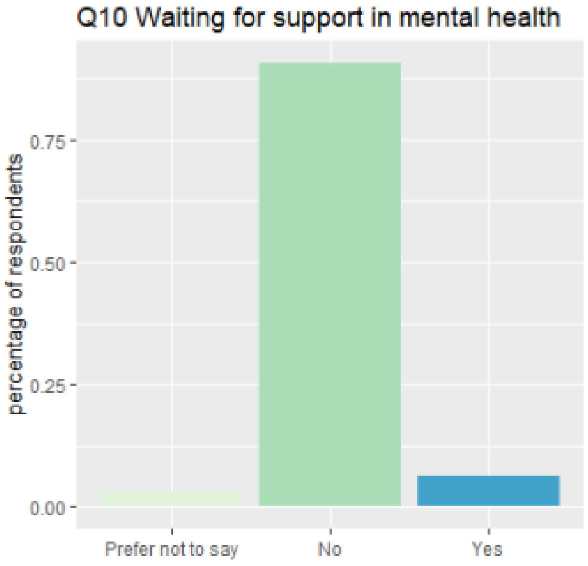
Plots| Variable description – Frequency of Question Answers

Frequency of answers for questions 4,5,9 and 10:





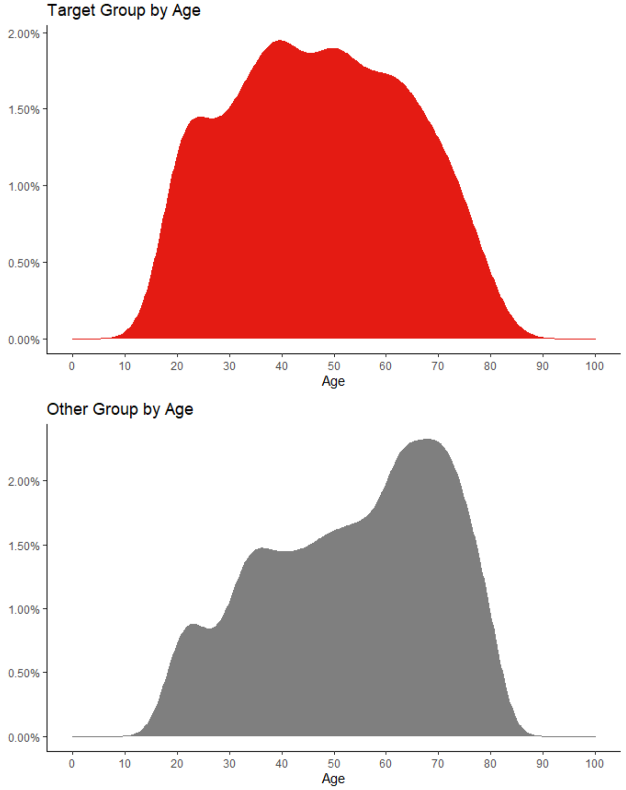




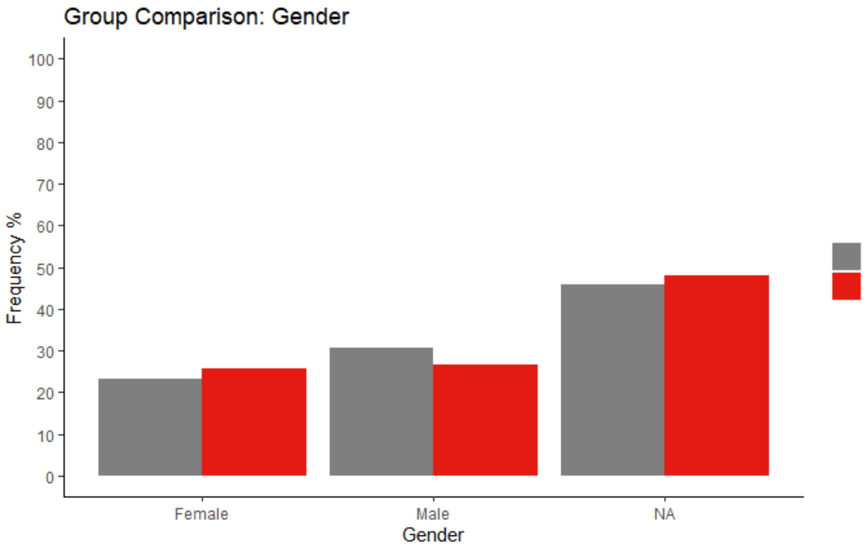
Plots| Variables distribution in clusters

We decoded the target group as 1 and the other clusters as 0

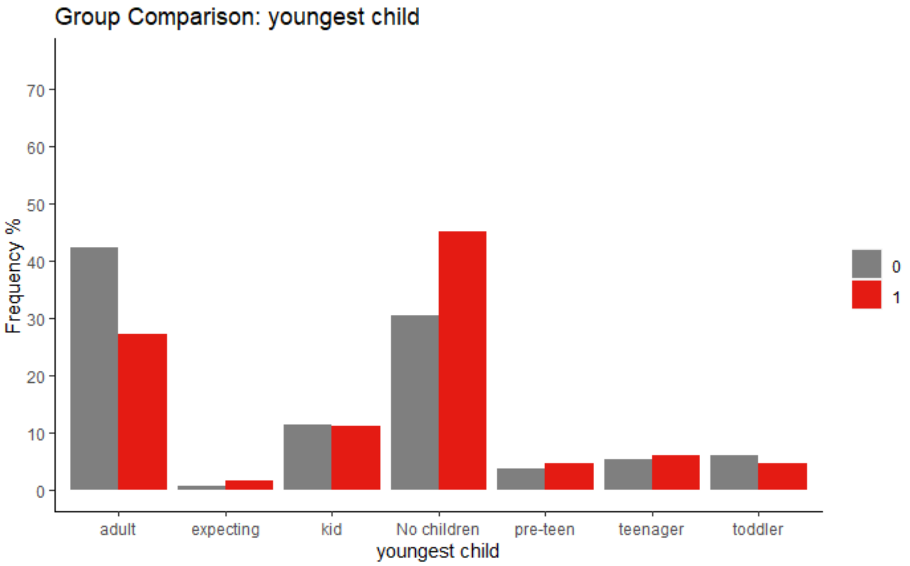
1. Age



2. Gender



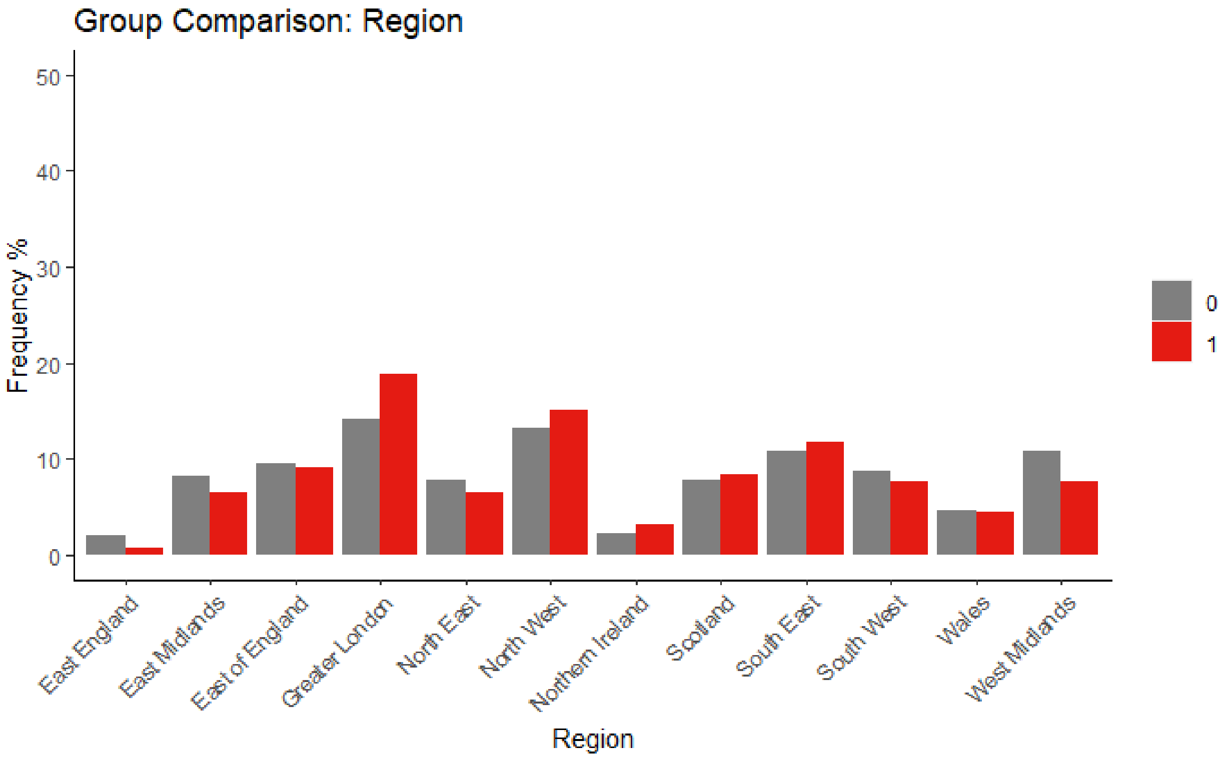
3. Youngest child



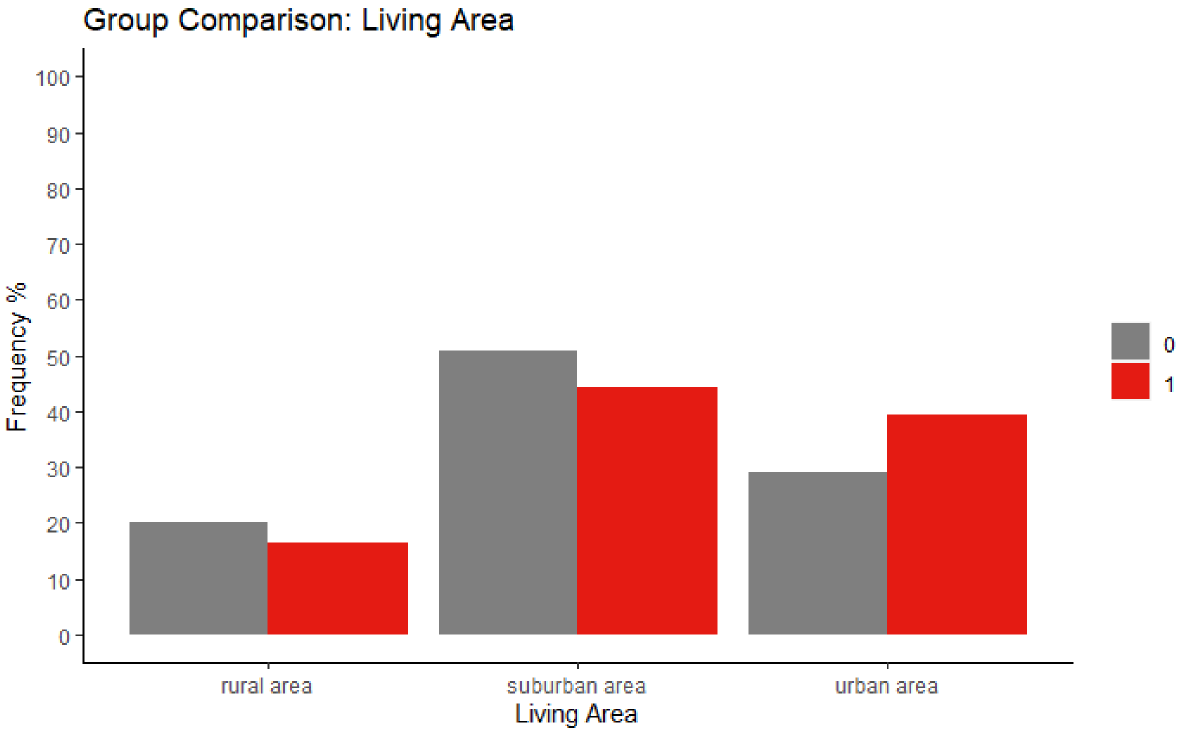
Plots | Variables distribution in clusters

We decoded the target group as 1 and the other clusters as 0

4. Region



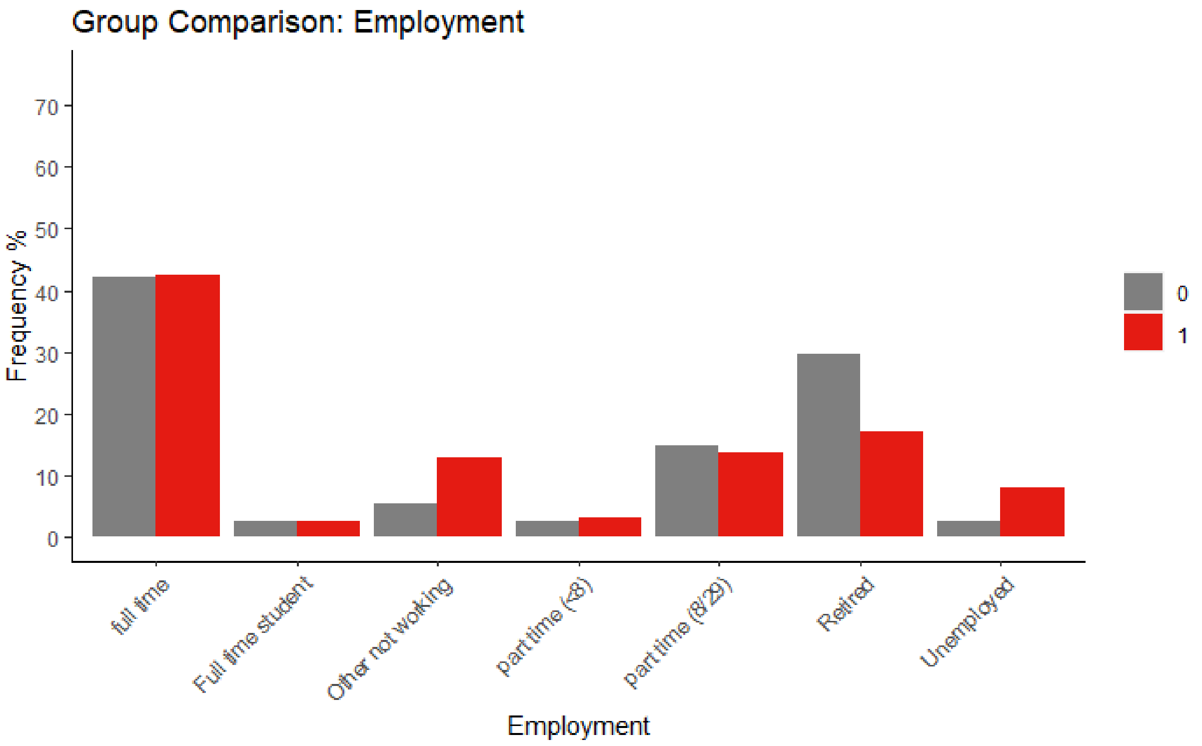
5. Living Area



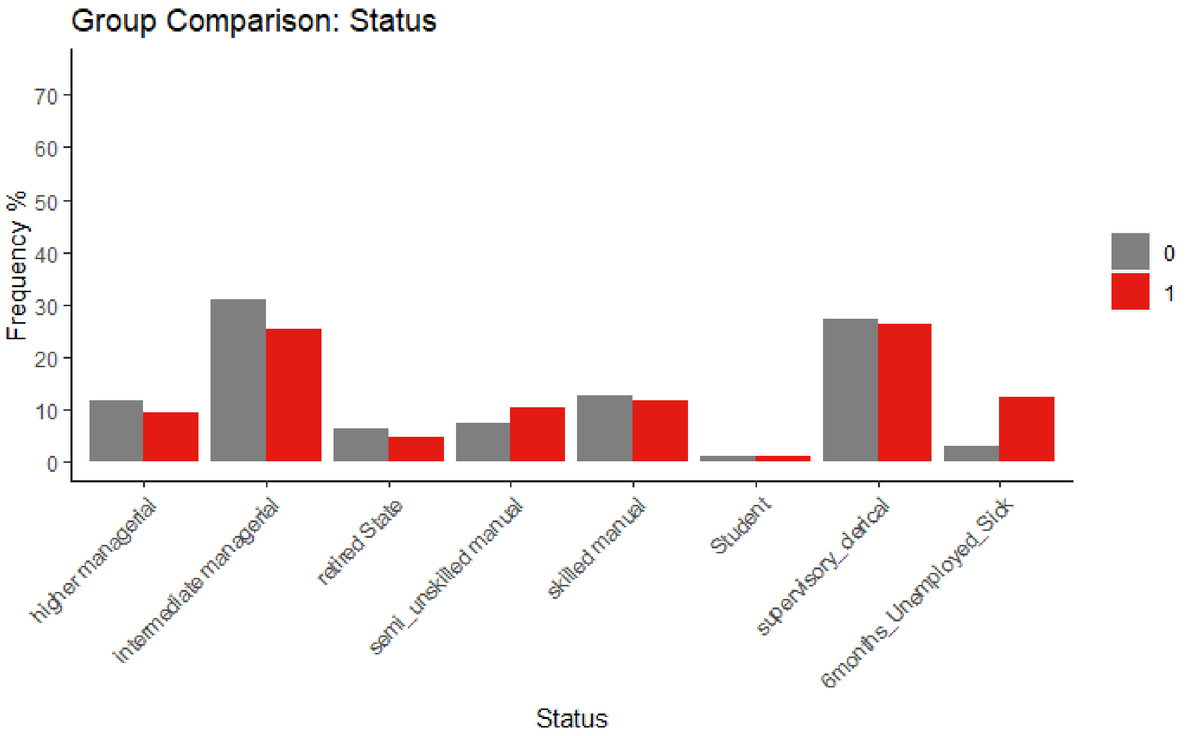
Plots| Variables distribution in clusters

We decoded the target group as 1 and the other clusters as 0

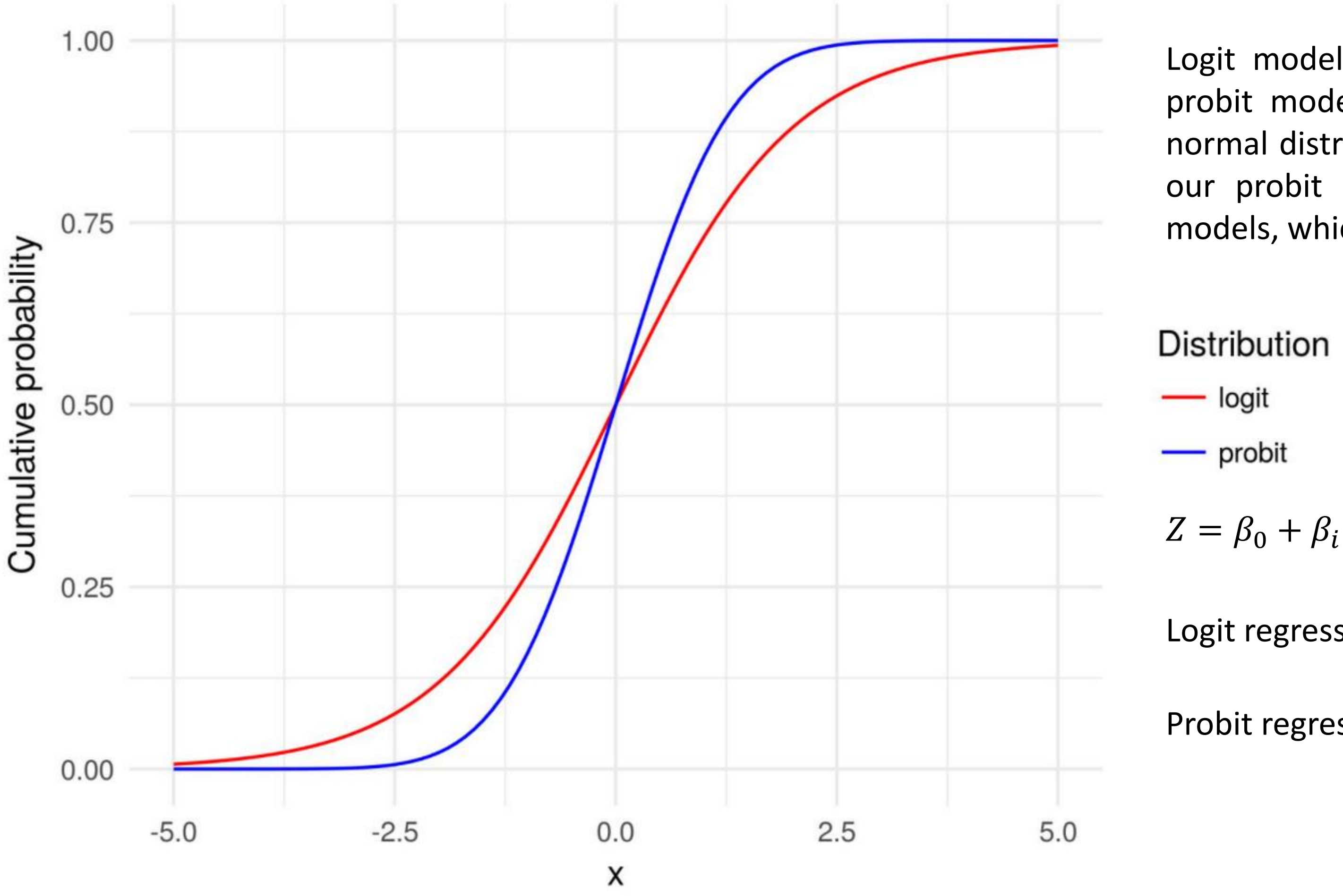
6. Employment



7. Working Status



Plots| Difference in logit and probit function



Logit models are used to model Logistic distribution while probit models are used to model the cumulative standard normal distribution. We wanted to confirm the predictions of our probit regression and therefore decided to use both models, which ended up giving us the same results.

$Z = \beta_0 + \beta_i \times X_i$

Logit regression: $Y_i = \frac{\mathrm{e}^{\mathrm{Z}}}{1+\mathrm{e}^{\mathrm{Z}}} + \varepsilon$

Probit regression: $Y_i = \Phi(Z) + \varepsilon$