Brown Datathon 2019

MINDGAP

Act fast, save lives

Kyi Yeung Goh
Columbia University

THE WHAT, WHY AND WHERE

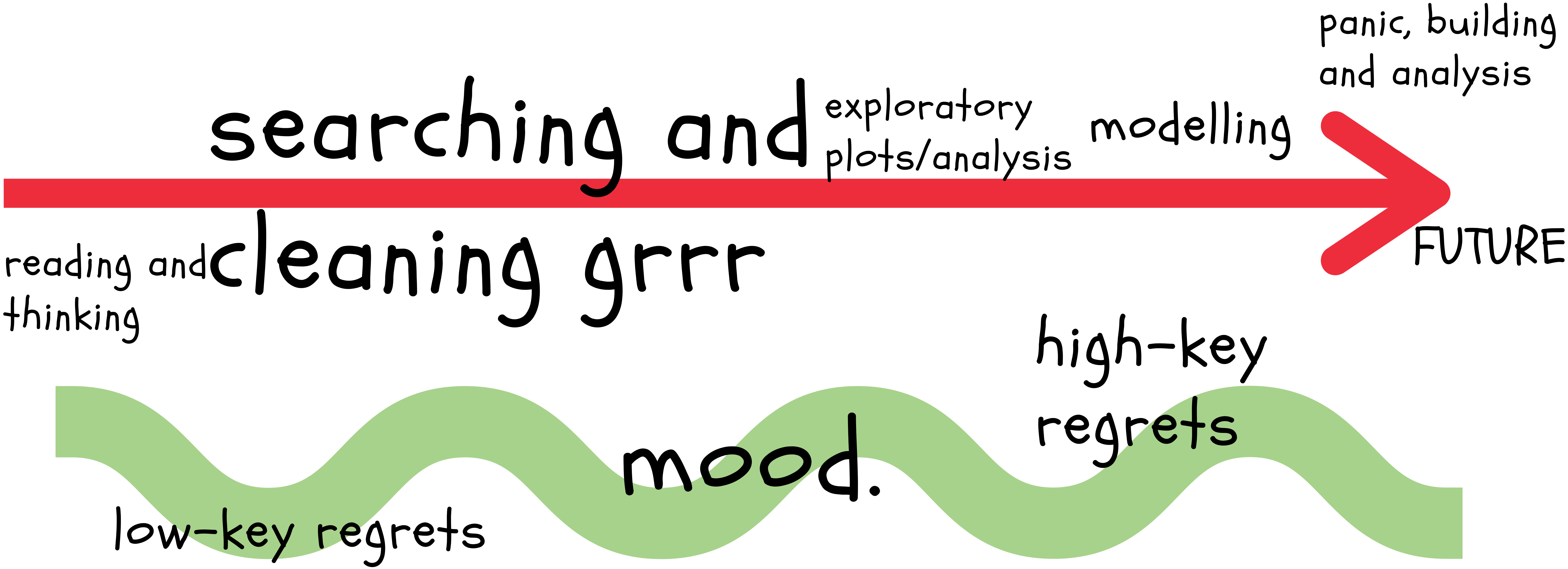
You will see a **very** rough sketch of the app at the end of the presentation predicting mental health outcomes and recommending interventions

Caveat: I had intended to spend more time learning rather than building (as part of a group) but decided to eventually give it a shot...

Why:

1. From a causal perspective, what drives mental distress?
2. Can we intervene early before things turn dire?

THE WHERE: WALKTHRU



MOTIVATION (ACADEMIC)

When exploring the dataset, I came across mental health issues in the codebook which was 195 pages long and on page 151: it is there

Label: Computed Mental Health Status
Section Name: Calculated Variables
Module Section Number: 2
Question Number: 2
Column: 1948
Type of Variable: Num
SAS Variable Name: _MENT14D
Question Prologue:
Question: 3 level not good mental health status: 0 days, 1-13 days, 14-30 days

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Zero days when mental health not good	300,134	66.69	63.56
2	1-13 days when mental health not good	92,902	20.64	22.68
3	14+ days when mental health not good	49,777	11.06	12.23
9	Don't know/Refused/Missing	7,203	1.60	1.53

MOTIVATION (PERSONAL)

For me I naturally gravitated towards the topic given my personal experiences. Over the past two years, a handful of my close friends have been left debilitated by mental health issues - a few grapple with crippling anxiety while others bravely confront depression.

Yet, as a friend, I did not notice warning signs. Indeed, many of them had not sought professional help prior to this, making it difficult to assess whether or not something is amiss.

DATASETS AND PROCESSING

Combined two CDC datasets:

1. CDC Behavioural Risk Factor Surveillance System (BRFSS) in **SAS?????** :(
2. CDC Chronic Disease Indicators 

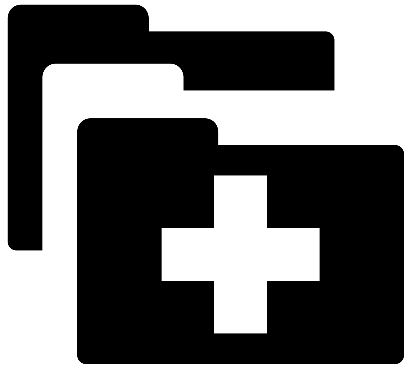
Target dependent variable, days that an individual had experienced mental distress in last 30 days. The dataset (after much cleaning) offered a total of 380+ variables, 17 of which were eventually included for testing - 2 were from the CDC Chronic Disease Indicators.

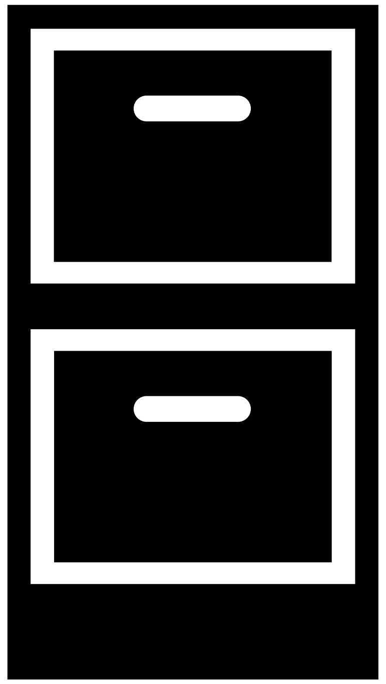
DATASET CLEANING

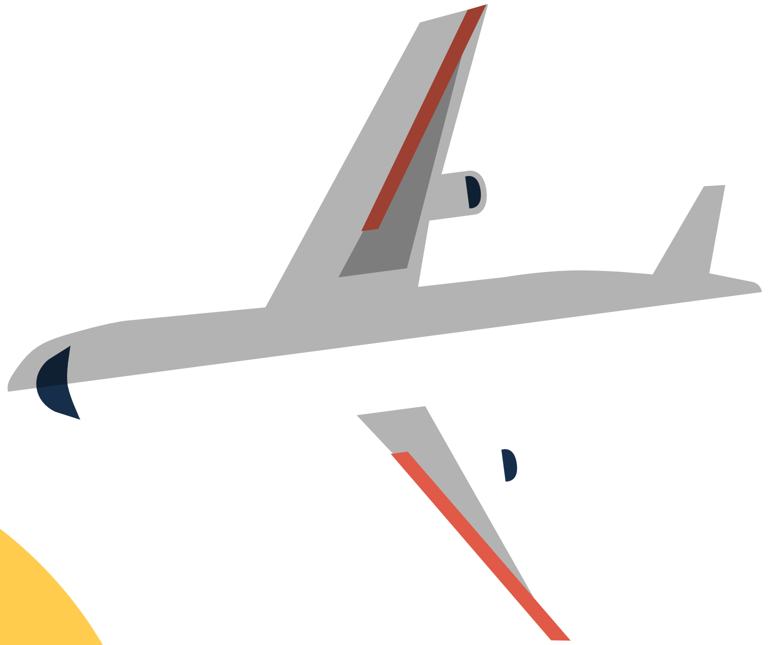
- Making initial scatterplots to look at NICEL assumptions - initially.... (its supposed to be weighted) - whether things needed to be squared, logged etc.
- Cleaning process included matching and partial matches
- Radically reshaping data
- Changing factor levels to ones that made more sense - iterating with continuous vs. categorical to see if it made sense - metropolitan areas
- Removing strange entries like 67000 minutes of exercise a week/ random imputations of nonsensical numbers etc.
- Matching based on states - one is in irregular numbers and the other one was in normal characters

SOME VARIABLES I SELECTED

My initial motivation surrounded the use of the data to answer a causal claim rather than a predictive question

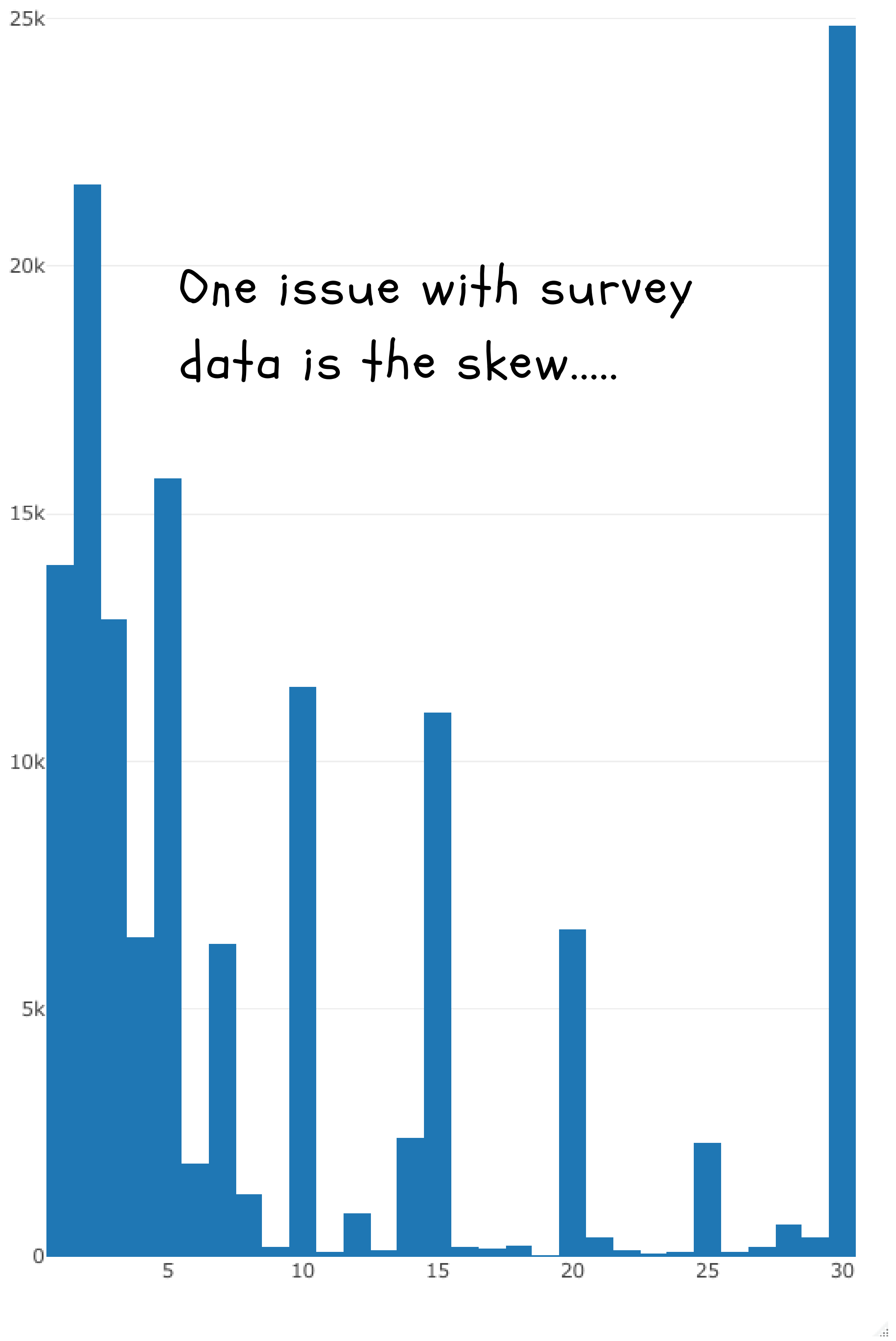


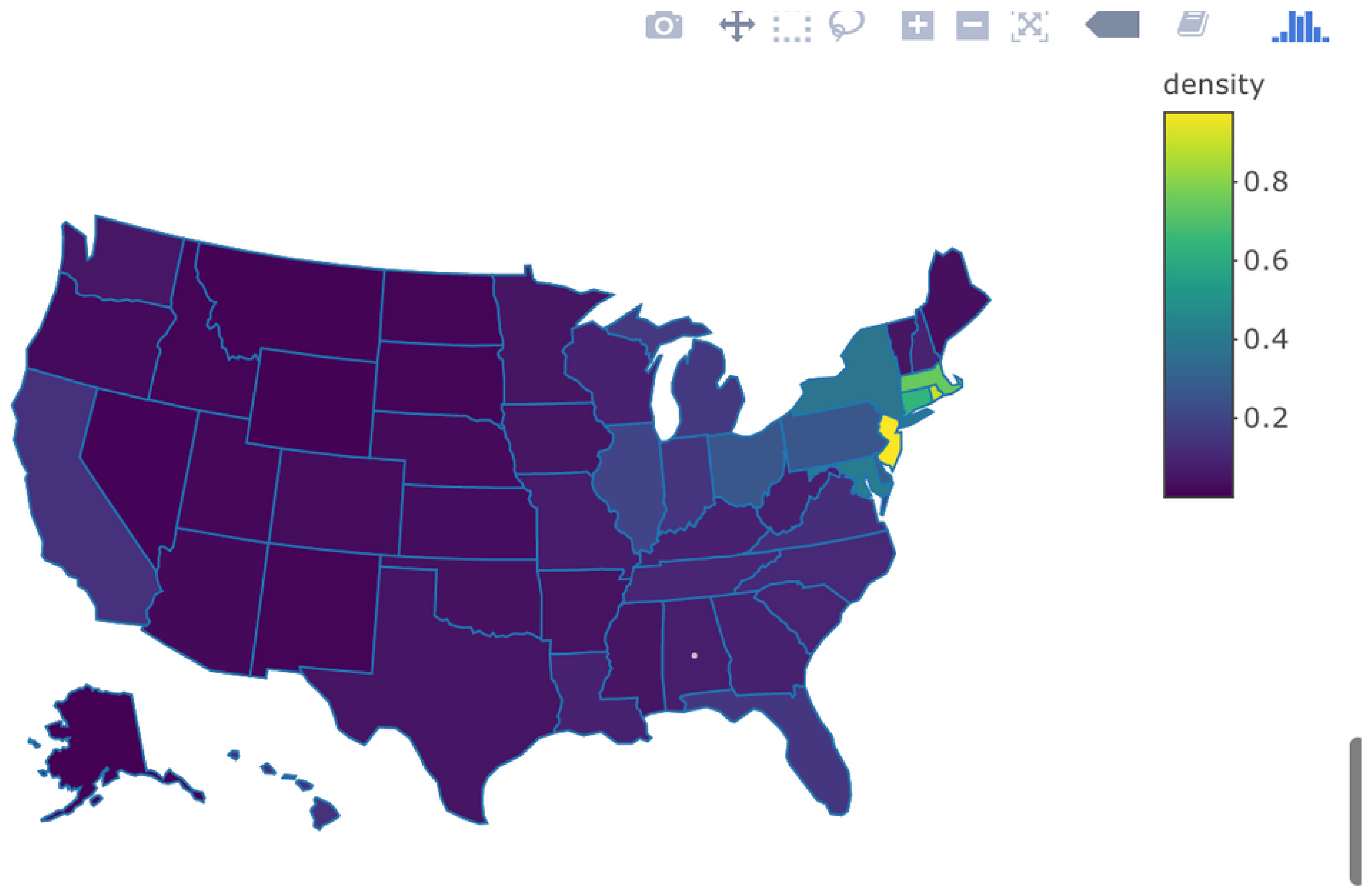




EXPLORATORY ANALYSIS

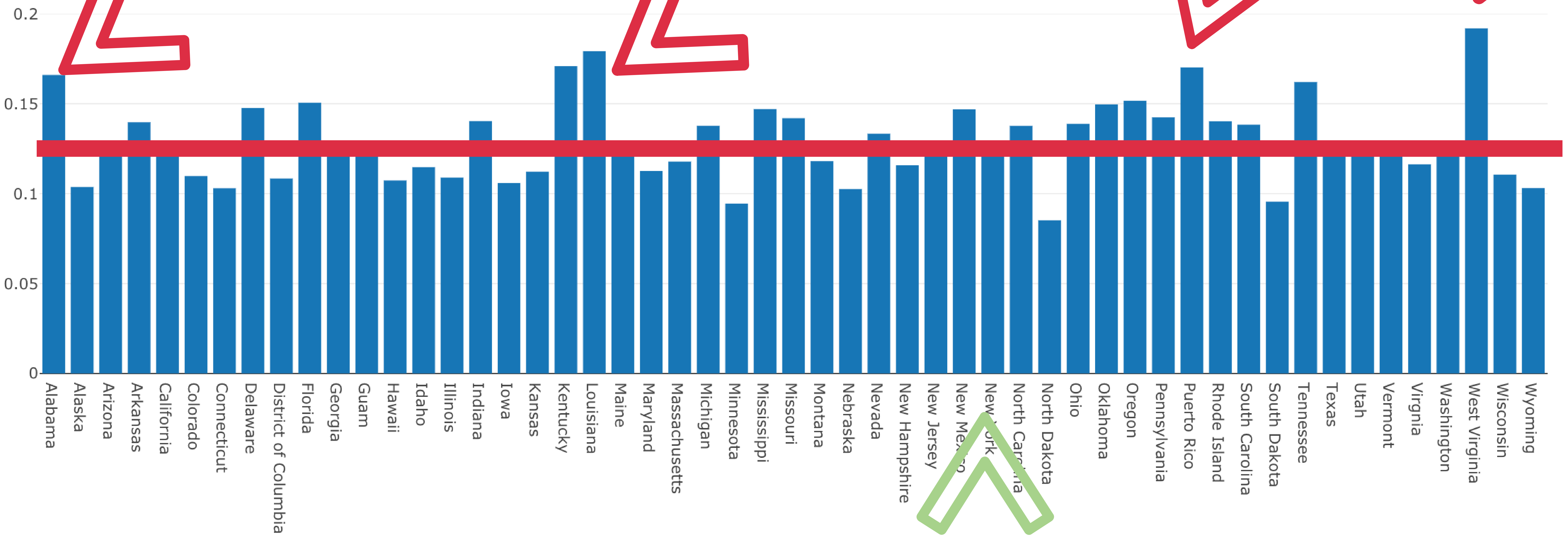
Interactive versions are included in a seperate html file





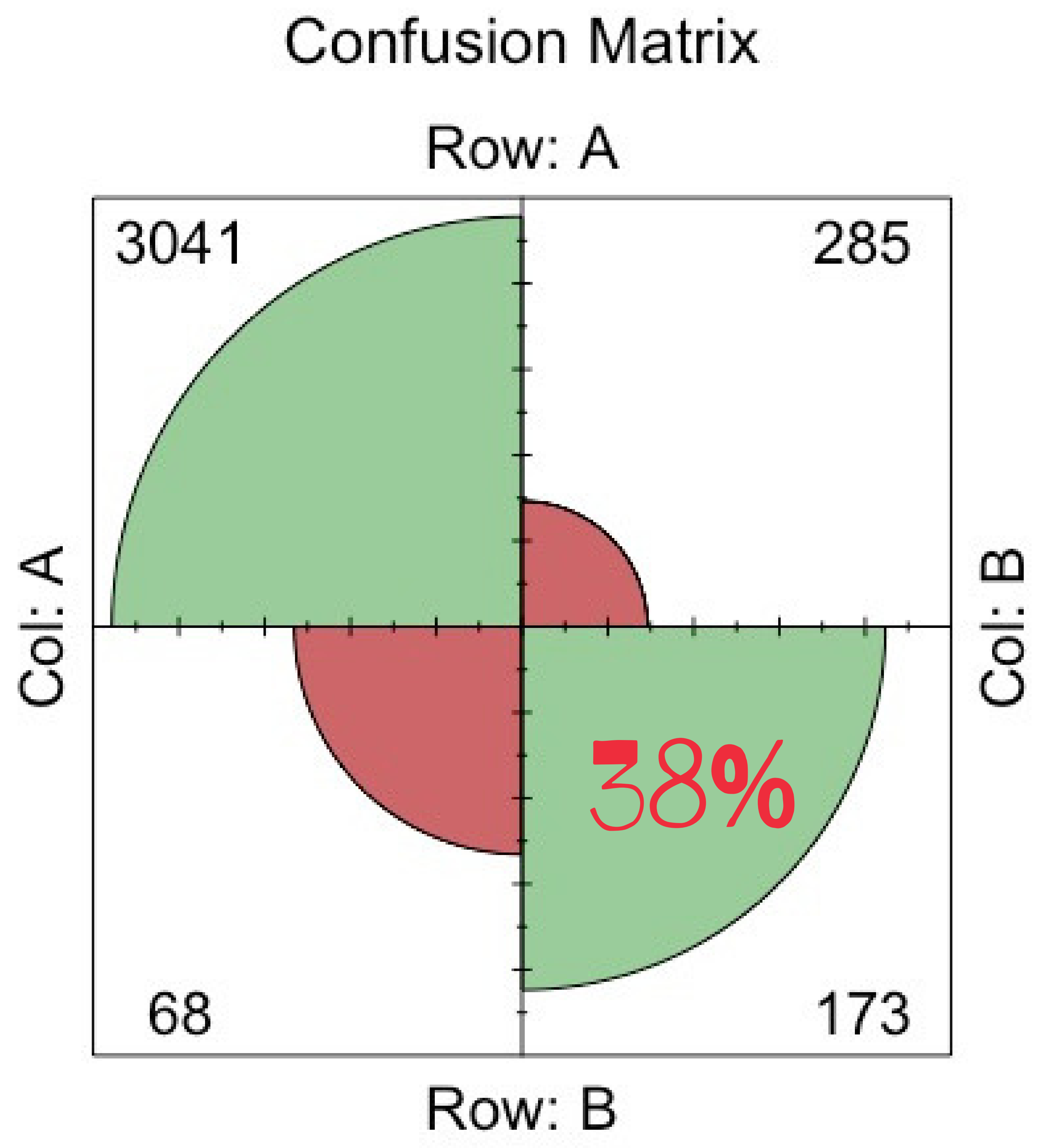
EXPLORATORY ANALYSIS

What we have here are the states (proportion) compared against one another - proportion of 'acute' mental health issues



Its supposed to be interactive on plot.ly but I am not a Pro member so you have to imagine that it is....

BINARY



FIRST MODEL - PART 1

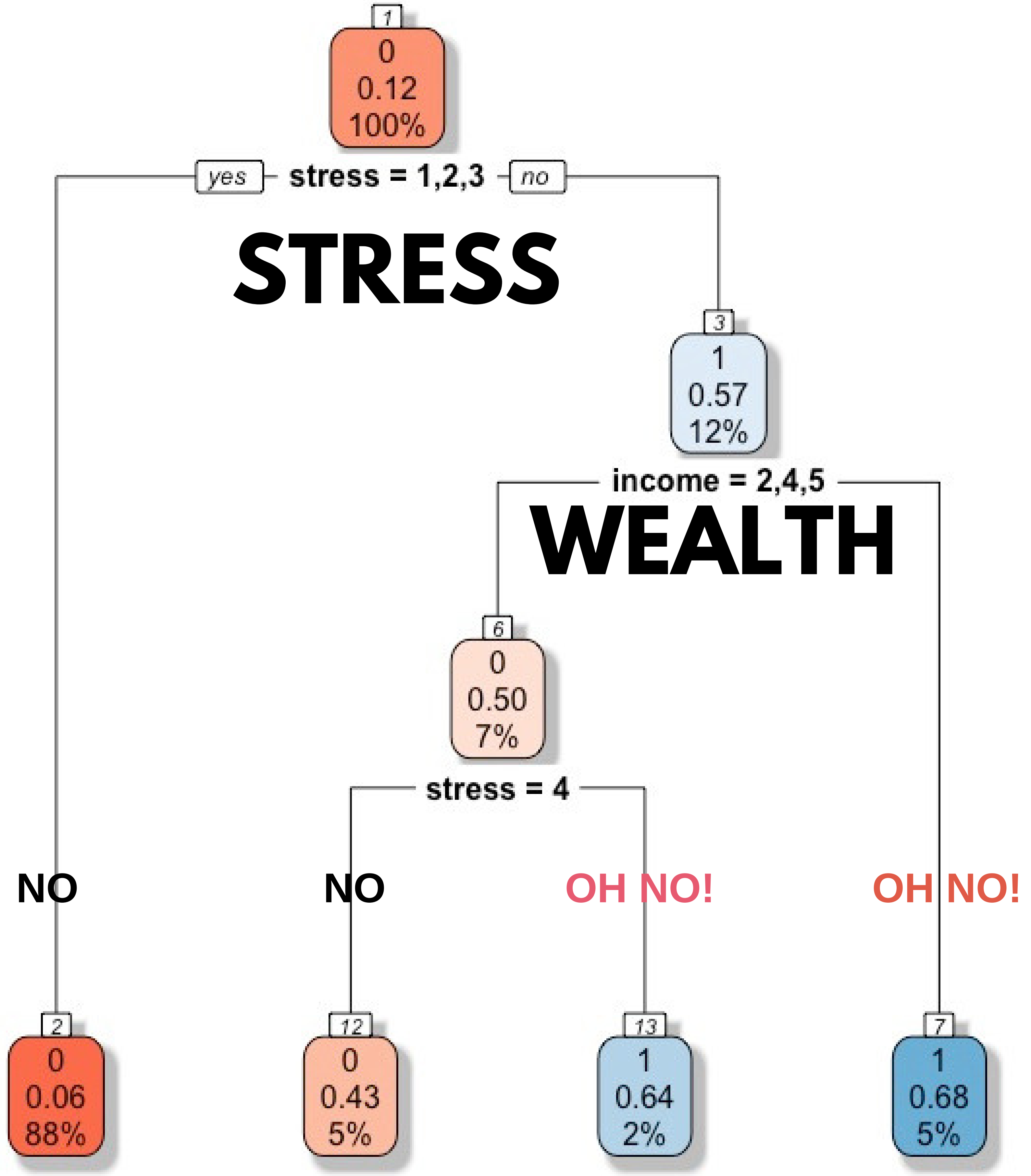
- Good ol' trusty logistic regression
- Why?

excellent at telling causal stories

collinearity can at least be addressed (L2-penalty for ridge*)

BUT......

bad for large feature space like my 17... categorical problems



FINAL MODEL* - PART 2

Decision tree ML model was ultimately the best performing, offering an overall prediction accuracy of around 92%**** and an accurate positive diagnosis rate of 50%!!!

An active decision was taken to remove previous medical records as a variable given that it would not uncover underreported cases. This would have given a much higher positive diagnosis score.

TRIAL MODEL - PART 1/2?

Stress is a very powerful predictor for how likely a person is undergoing soe kind of mental duress

Work appears to reduce stress, particularly amongst those with higher incomes

Interaction effects indicate that this wealth effect is offset as one ages around the 40 year old mark

Yet, we need to contextualise the results using a logistic regression model rather than an ML one (though thinking in more probabilistic terms)

I also gave **randomforest** (using R gridsearchcv) and **SVMs** a shot though the results weren't much different from the two, albeit with much larger false negatives.

VERY ROUGH PROJECT DEMO

A very rough guide of what this might look like from a user standpoint

THE FUTURE?

Opportunities:

1. Integrating with other data points

2. Life SOS-helplines thru data collection (*)

3. Eventually build an app with capabilities to serve these social-good purposes.