**Question 1.** What are the best predictors of mental health (perhaps depression in particular / is it possible to differentiate between those with minor v major depression?)? (What features about a person have the highest correlation)

Are these features colinear or correlated to one another in another way, for example BMI and weight may both be positively correlated to poor mental health, but BMI is a function of weight and height, so weight and BMI are strongly correlated (though not linearly) – there is also likely a correlation between other personal features, such as smoking and cancer etc…

(including overarching categories: lifestyle choices, educational results – background too if available, zip code someone lives in, sex, age, race, presence of physical health conditions, weight / BMI, financial wellbeing, job) - other factors I see as possible.

What is a pen profile of someone that is at high risk / likely to have or get mental illness?

Is is possible to create an accurate model that can predict whether a person has, or will have a mental health issue based on these features?

How should this model be evaluated?

“Except for the ‘POORHLTH’ attribute, due to trivial and not useful link in the case of prediction. Which attributes in the dataset given are the most predictive of poor mental health? Here we will count an attribute as predictive if it has high ‘feature importance’ and ‘predictive power’. That is to say that there is a high correlation between the attribute and the response variable and therefore would be useful in a predictive model. In this case, the response variable by which we are measuring ‘poor mental health’ is an individual answer to the question ‘Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?’”

Based off the model we have created, is it likely that those who refused to answer the question, and that those who were unsure, in fact did have poor mental health? This may indicate they are worried about a result of poor mental health being recorded.

Tasks:

First create the relevant dataset

Use exploratory data analysis.

Plot graphs, use pandas describe and head methods.

Find the data type and unique values of each column.

Use the data dictionary to understand the data.

Separate the response variable.

Consider string variables to be categorical variables.

Discard the missing data points from the response variable.

Plot count plots with seaborn to visualise the correlation of each data point to the response variable.

**Question 2.** Are Zip codes that are generally associated with greater wealth (look up median incomes, wealth or house prices and site what sites gave me this information), healthier, mentally, and physical, have less – or more acute conditions) than others.

If so, does this correlate with better lifestyle conditions or features the person is born with like race / sex (also grades? – which could be attributed to among other things: intelligence, family support, wealth, education, or educational opportunities).

Which of these is the better predictor of a healthier life (less or more acute conditions)

Are there certain conditions that richer or poorer people get more often? (Rich more interestingly as my assumption would be that poor people due to less access to better and healthier decisions, would get more diseases, but perhaps a wealthy person’s lifestyle also brings on different diseases.)

**Question 3.** Does a person’s sex have a statistically significant effect on the number of, and type of health conditions someone experiences? If so, what are the most highly correlated conditions with each sex, and which conditions show the biggest and smallest divergence.

If so, what effect does it have?

Are these differences, if they exist, down to the persons sex, or is this more likely described by other factors like zip codes, race, lifestyle factors (are these linked to sex anyway?) – is there enough, or the right data / evidence to show this cause rather than correlation either way?

**Similar question to the last two:** What about whether someone has Medicare, or another form of health benefit, or not.

**Question 4.** Does a screening for a stroke (or other screenings for conditions) show a correlation with a decrease in the existence of other conditions besides that being screened for?

If so, this could give rise to a further research project into why? – Is it because the presence of another illness, for instance cancer, or menta health is discovered during the screening? (spill over diagnoses) – Could this encourage more people to go to screenings, or get check-ups, could this instead encourage doctors to take more samples at screenings, or perform more procedures to check for other things at the same time – could the research guide the correct other things to check for due to correlations, or predictive features found?

**Question 5.** Does the indication of any conditions early on in life have a strong (statistically significant) correlation with a condition later in life (is an initial condition at a particular age – or withing a given range of ages – a good predictor of another condition at a later age – or range of ages)

What if we take out the timeline factor and just look for correlations / predictors of one disease / condition from another.

What is we take out the condition factor and include personal features, or lifestyle choices?

Could we then add timelines back in and link lifestyle choices with a timeline as to when, proportionally, or nominally, based on the age of a particular lifestyle choice or ages that lifestyle choices were made, when a person will get a particular illness or condition. Or perhaps when we can be convinced, they have avoided it through good decisions, for example, if someone eats fruit and has not had diabetes or cancer by 55, they are unlikely to d so, or if someone smokes before they are 32, they are likely to have a stroke by 55, or between the ages of 60 and 75.