Predicting mental health issues in individuals using unrelated lifestyle attributes

BY: TEAM NSMR

INTRODUCTION

- A significant portion of the population suffers from mental health issues and the topic is not being openly discussed.
- The ongoing pandemic only exacerbates these problems and hence even more reason to understand these problems better.

Introduction(cont.)

- We want to provide actionable insights into this ubiquitous problem by analysing lifestyle factors not generally associated with mental health issues.
- Our dataset spanned over 20 years and over 6000 unique rows with over a 100 column for each row. The sheer size of the dataset necessitated the use of big data framework and analytics.

Background

Machine-learning boosted regression analysis found some biomarkers related to depression in the National Health and Nutrition Examination Survey dataset.[1, 7]

The effectiveness of predicting a set of common mental conditions based on textual description of patient's history of present illness was discussed by Tung, 2014. [2]

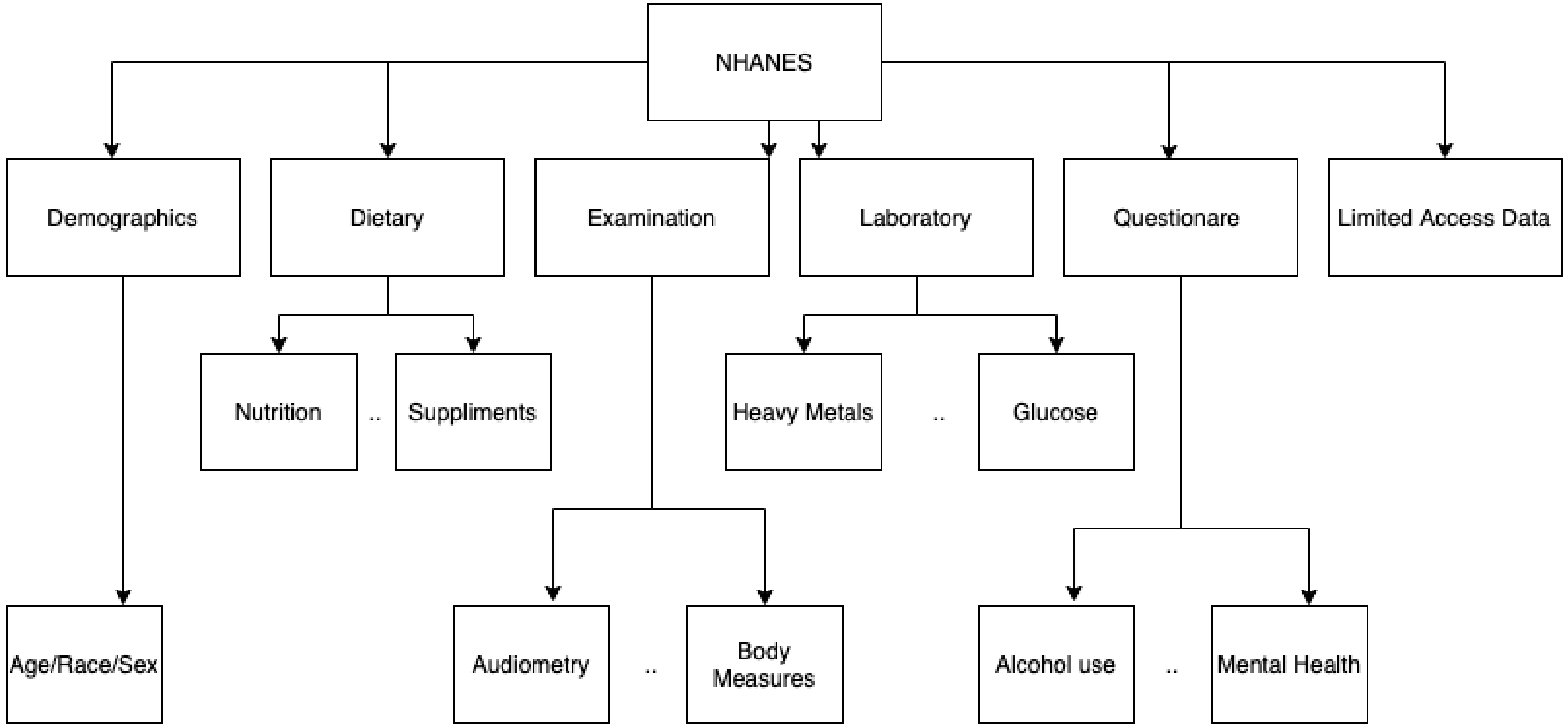
There are works identifying depression from social media that use natural language processing and quantify activities on social media platforms, identifying positive or negative emotions, social isolation and other platform specific statistics.[3, 4]

There is also growing interest in using sensing data from wearables to infer human mobility and mental health. [5,6].

Data

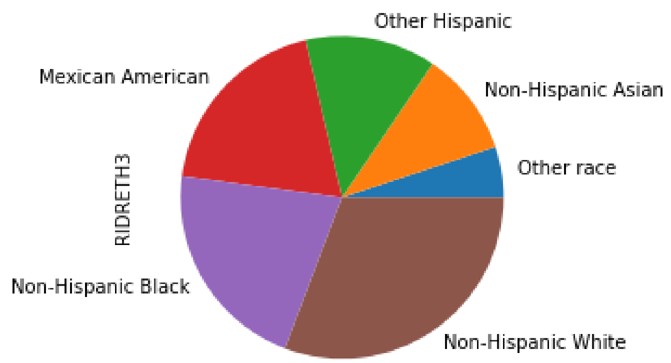
- We used the National Health and Nutrition Examination Survey(NHANES) dataset which examines a nationally representative sample of about 5,000 persons each year.
- The NHANES interview includes demographic, socioeconomic, dietary, and health-related questions in addition to the laboratory tests administered by highly trained medical personnel.

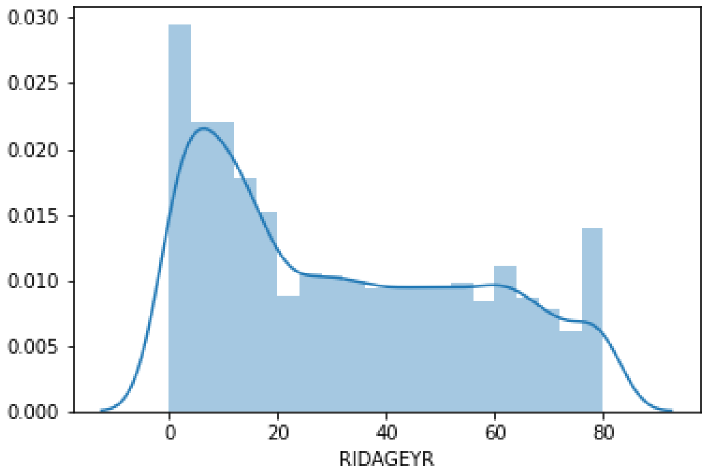
Data(Cont.)


NHANES
Demographics
Dietary
Examination
Laboratory
Questionare
Limited Access Data
Nutrition
..
Suppliments
Heavy Metals
..
Glucose
Age/Race/Sex
Audiometry
..
Body Measures
Alcohol use
..
Mental Health

Data(Cont.)

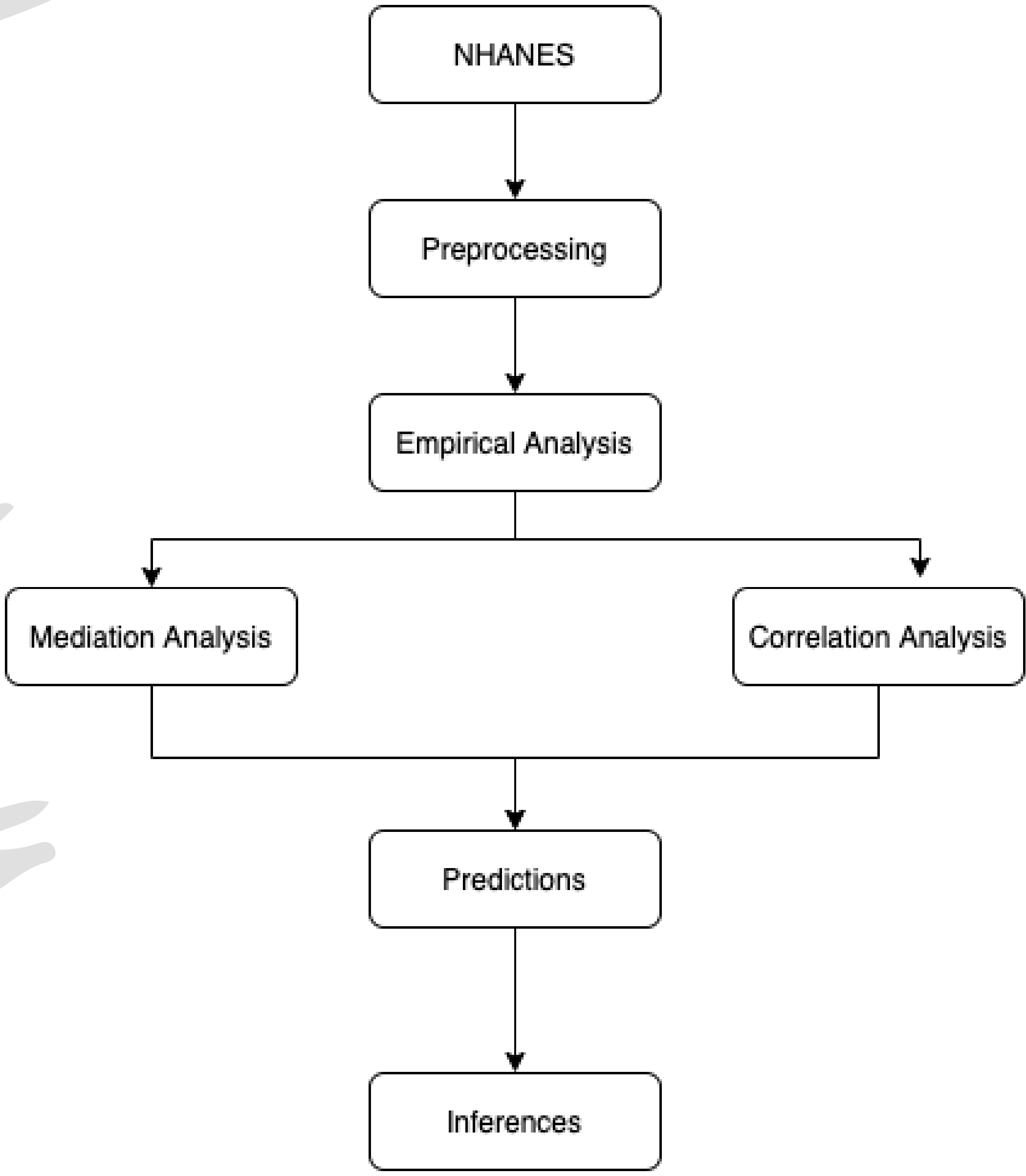
- The data consisted of individuals from all age group and ethnicity from different areas of the united states.
- The data part of the questionnaire subset was categorical while the data in other subsets were mostly continuous.





Method

- The complete project breaks down into the following subsections
    1. Data Preprocessing
    2. Empirical Data Analytics
    3. Correlation Analysis
    4. Mediation Analysis
    5. Predictions
    6. Inferences



Data Preprocessing

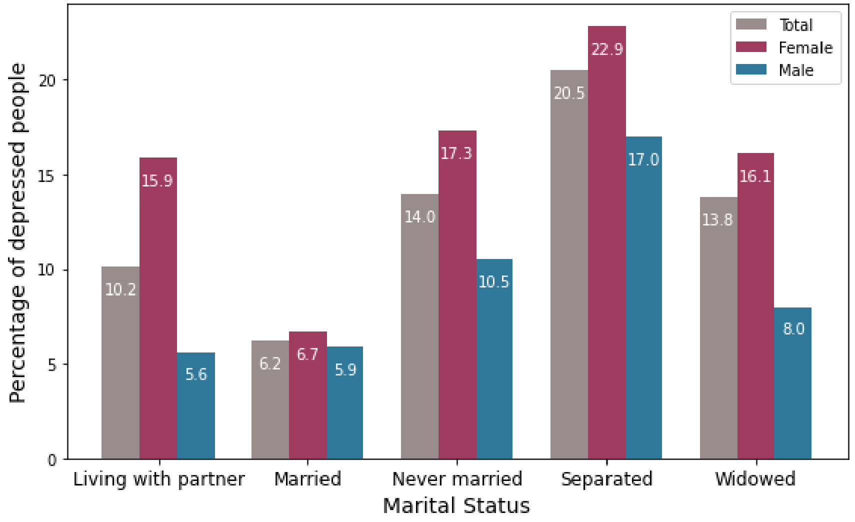
FEATURE SELECTION: WE SELECTED A SUBSET OF THE DATASET WHICH WE THOUGHT WOULD PRODUCE INTERESTING INSIGHTS.

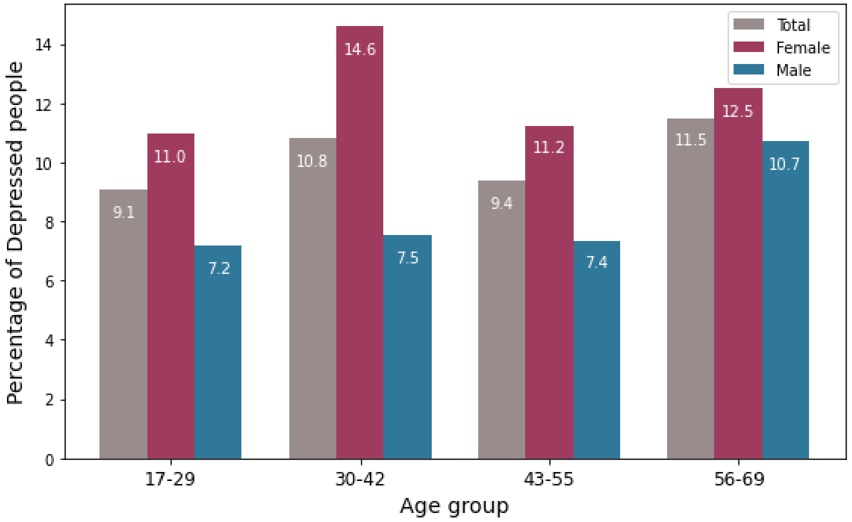
VALUE MAPPINGS: UNDERSTAND AND CONVERT THE VALUES GIVEN IN DATASETS TO APPROPRIATE SCALES BY USING THE DOCUMENTS PROVIDED TO UNDERSTAND DATA BETTER.

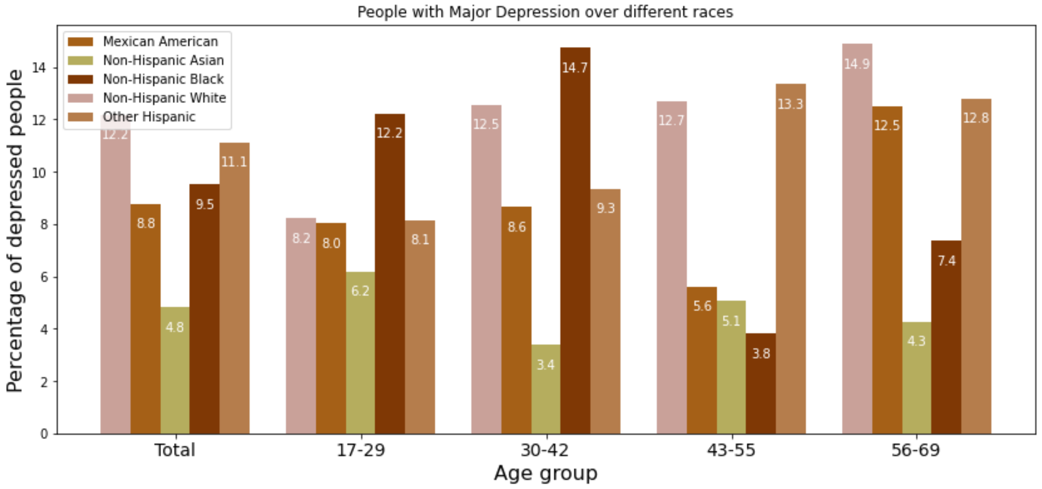
HANDLE MISSING VALUES: REMOVE/IMPUTE MISSING VALUES FROM THE DATA BY REASONING WHY THE VALUES MIGHT BE MISSING.

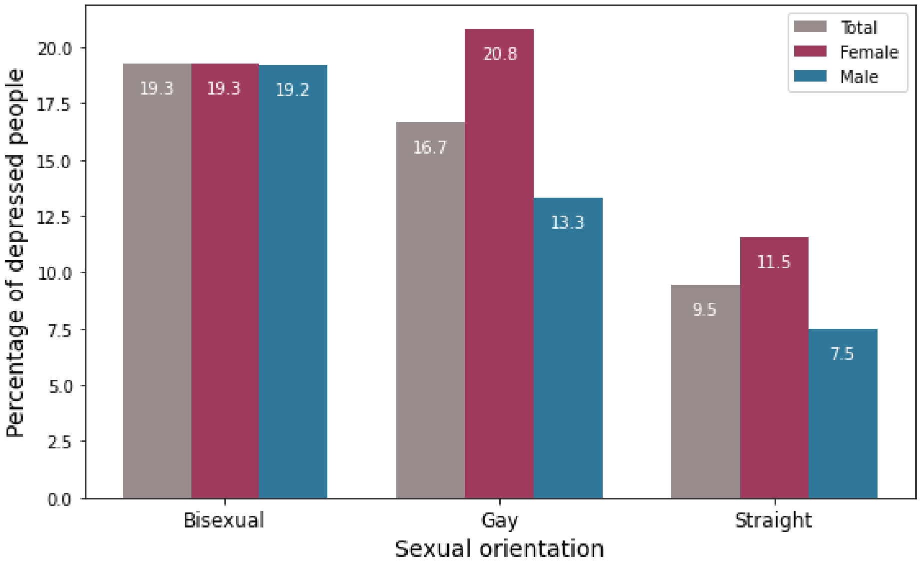
PRODUCE OUTPUT CSV FOR USE IN LATER PART.

Empirical Data Analysis

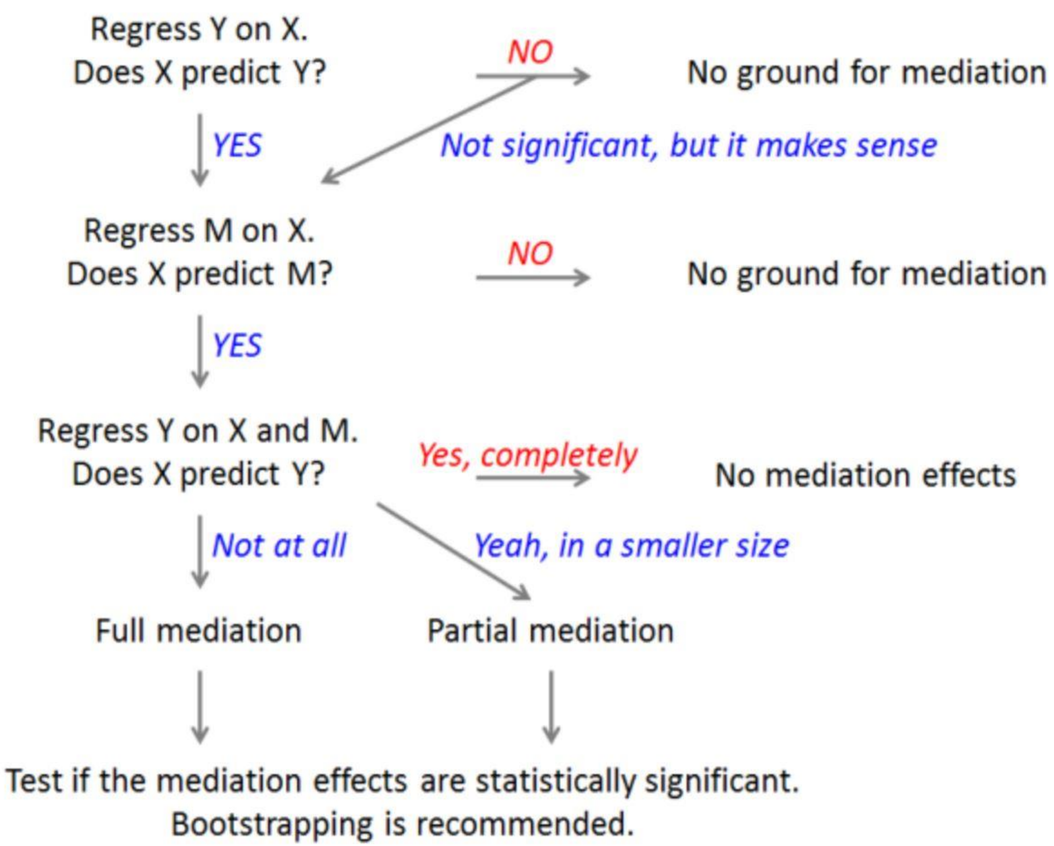






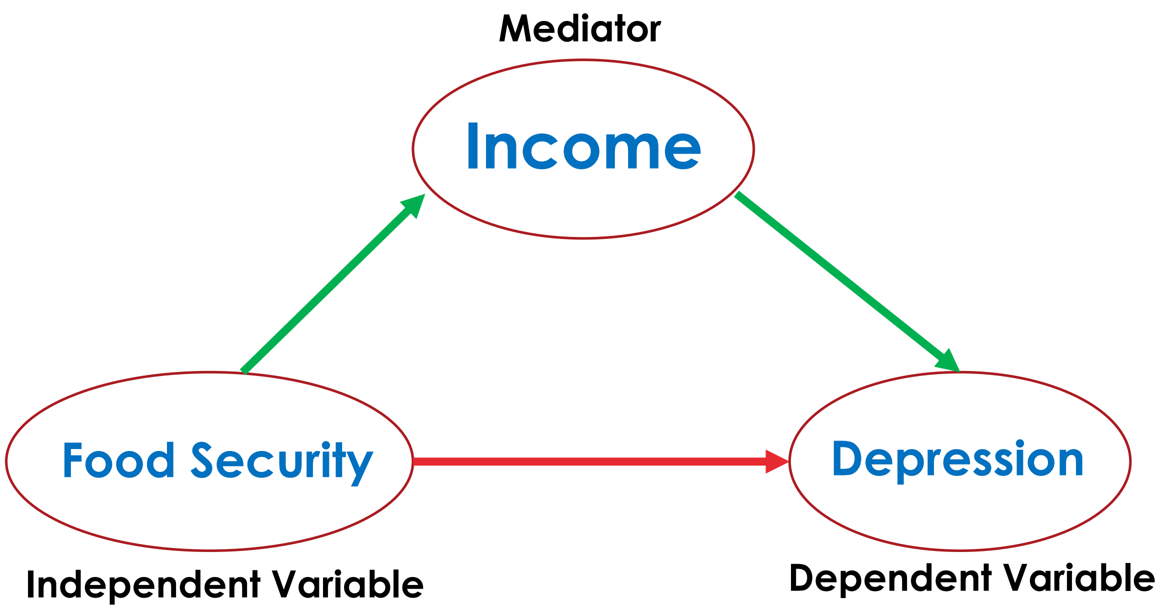


Mediation Analysis



University of Virginia Library Research Data Services

Terms	Meaning
Average causal mediation effects (ACME)	Indirect effect of the IV on the DV
Average direct effects (ADE)	Direct effect of the IV on the DV
Total Effect	Total effect (direct + indirect) of the IV on the DV
Prop. Mediated	Proportion of the effect of the IV on the DV that goes through the mediator



Correlation Analysis

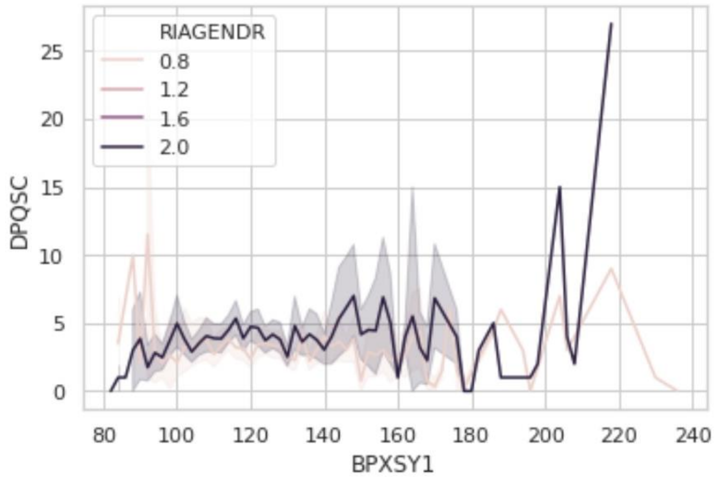
- To make population inferences, all sample data was weighted.
- NHANES recommendations were followed to calculate weighted depression scores and unfold complex survey data.
- Statistical significance was assigned based on 95 % confidence intervals.
- Hypothesis testing involved calculation of P-values and performing the T-Test on the dataset. The betas values enabled selection of relevance features for our prediction model.
- Multivariate regression was performed using tensorflow to observe the weights of every feature of each of the dataset.
- The betas values enabled selection of relevance features for our prediction model.

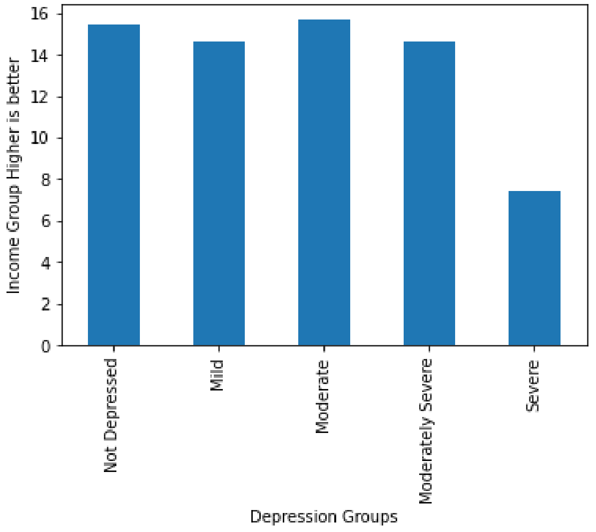
Predictions

Selected features from Analysis → Random Sample → Split into X,Y → Find Neighbors → Compute Values → Calculate Error

Evaluation/Result

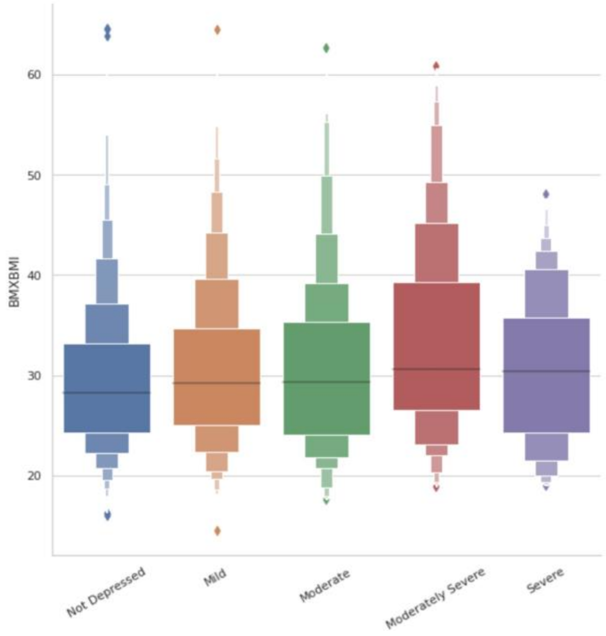
- Some of the insights that we discovered were as expected:
    1. People with higher income showed lower depression rates. But people who are moderately depressed have similar income on average as compared to people who are not depressed.
    2. Higher Blood Pressure (Systole Readings) is positively correlated with depression scores.

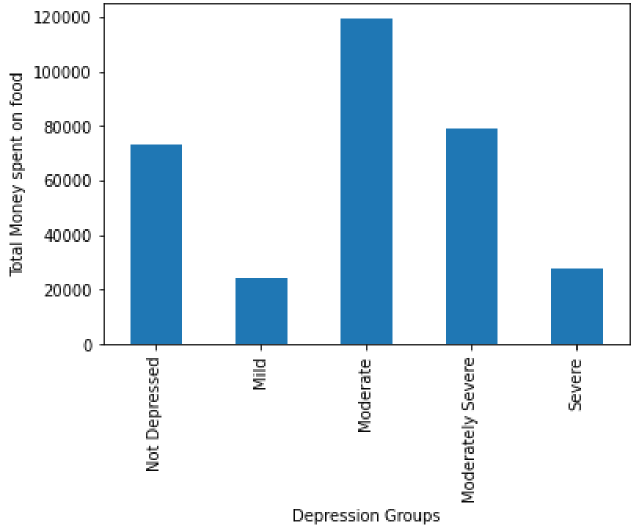




Evaluation/Result

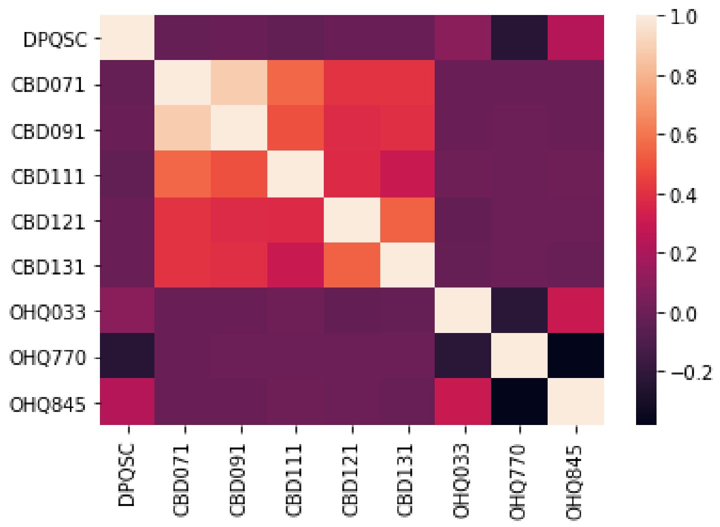
- On the other hand some of the factors that are thought to have strong correlation with mental health show no real correlation for ex: BMI.
- The total food expenditure also varies by the level of depressive symptoms in an individual with moderately depressed being the highest spenders.

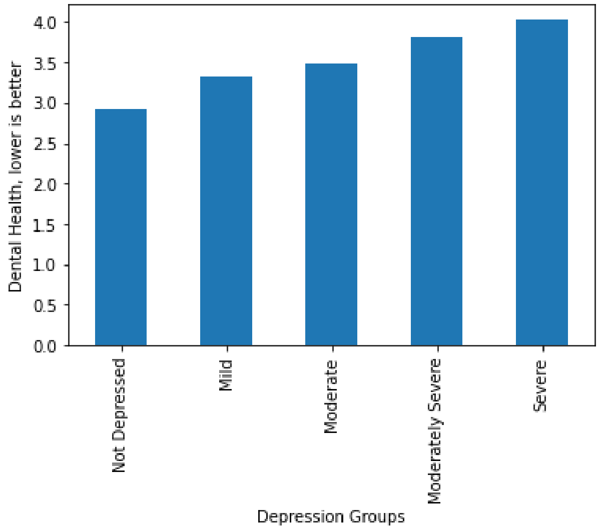




Evaluation/Result

- Factors which are generally not correlated with mental health like dental health show a correlation of 0.2
- Consumer behavior on the other hand show very little correlation(~0.0055).
- For the Prediction, the MSE value predicting the column most directly related to depression was 0.83.





Conclusion

- Our study combines simultaneously assessed self-reported questionnaires and physical measurements to generate guidelines for predicting and tackling the mental health problems.
- We were able to determine statistically significant correlation b/w mental health and lifestyle factors such as dental hygiene and income using the continuous NHANES dataset.
- Most of the features that we selected didn’t show statistically significant relationship with the target variable which was to be expected.

References

1. NHANES Questionnaires, Datasets, and Related Documentation
2. Spark Python API Docs
3. Seaborn – Statistical Data Visualization Documentation
4. https://github.com/fedorgrab/nhanes-analysis/blob/master/NHANES_analysis.ipynb
5. Introduction to Mediation Analysis
6. Mediation Statsmodel