COMP20008 – ELEMENTS OF DATA PROCESSING

PHASE 3-B

RESEARCH QUESTION

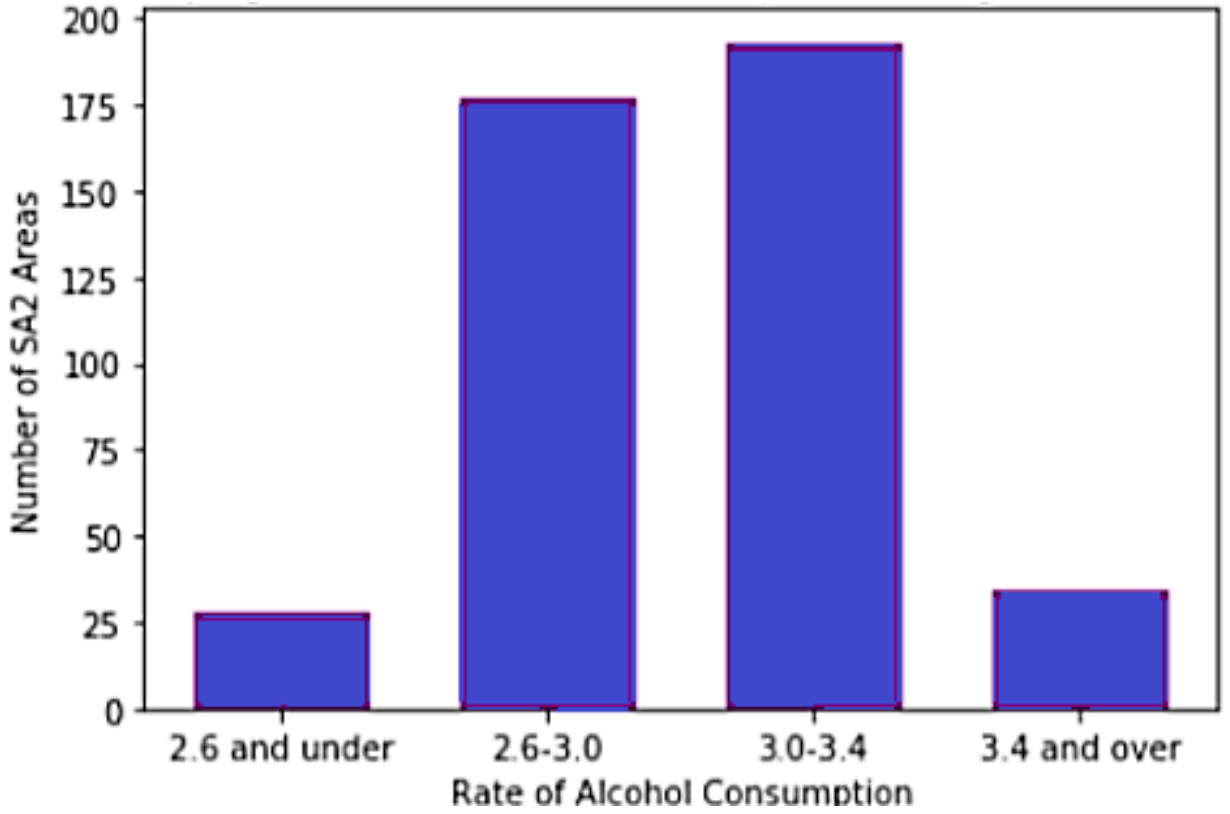
- Do people who are socioeconomically disadvantaged have a high health risk factor?

MOTIVATION

- Smoking has been associated with cancers and lung disease; obesity has been associated with mature onset diabetes and heart disease; and risky drinking has been linked to liver disease and acute short term effects, for example, dangerous driving and violence.
- Alcohol and Tobacco (5.1% and 9.0% respectively) are amongst the highest contributing risk factors to burden of disease in Australia (AIHW).
- Being obese poses major health risks by increasing the risk of chronic illnesses such as diabetes, cardiovascular disease and some cancers (WHO).

MOTIVATION

Changing rates of Alcohol Consumption across Victoria
Worth investigating why the rates differ across different areas



DATASETS

- The following datasets from AURIN were used:
- SA2 Health Risk Factors – Modelled Estimate 2011-2013
- SA2 Chronic Disease – Modelled Estimate 2011-2013
- SA2 National Regional Profile (NRP) – Economy2009-2013
- SA2 OECD Indicators: Income, Inequality and Financial Stress 2011
- SA2 SEIFA 2011 – The Index of Relative Socio-Economic Disadvantage (IRSD)

REASONS

- All datasets were from the same region (SA2)
- Helpful to show relations between attributes

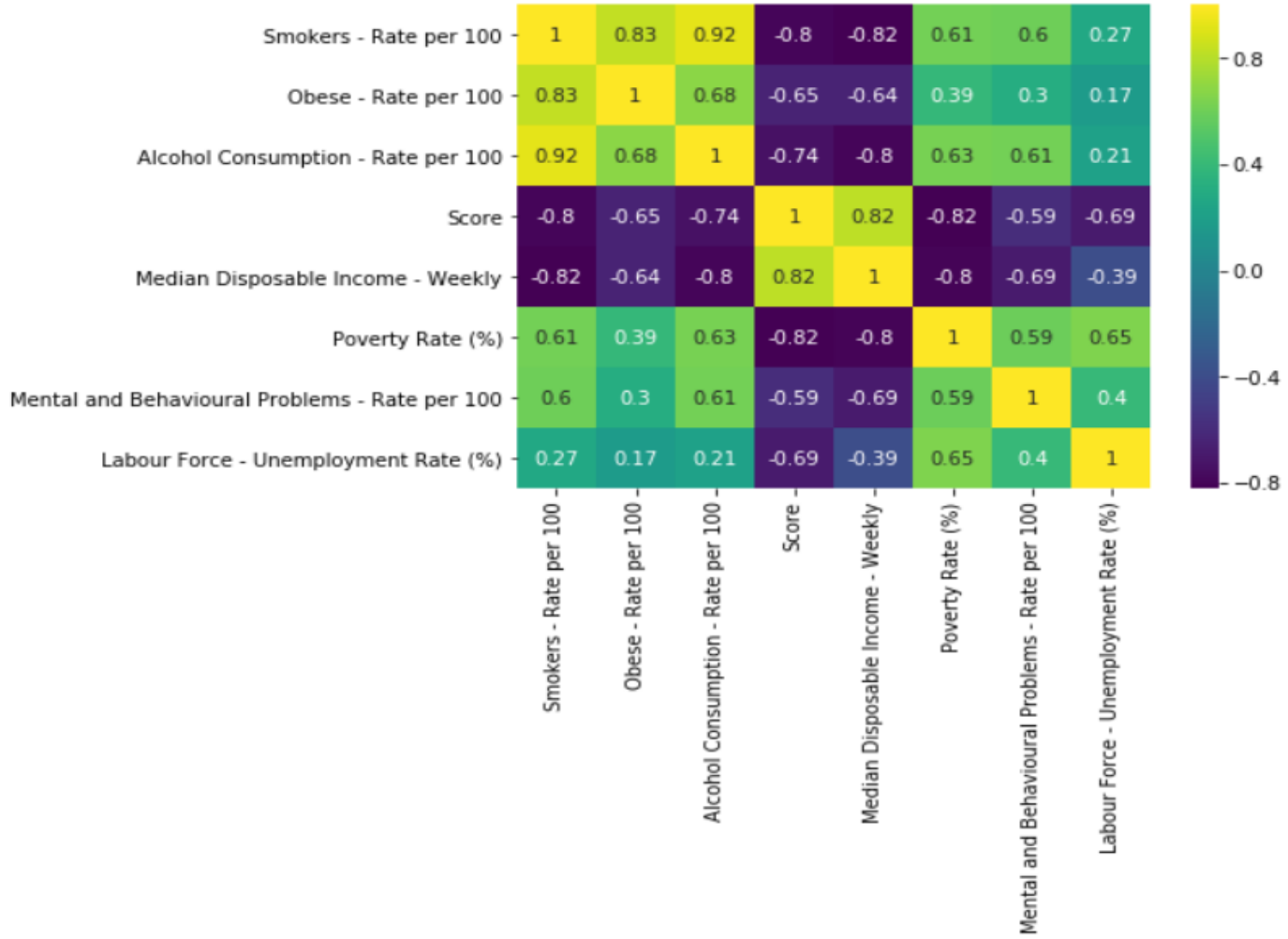
DATA WRANGLING

- Data Preprocessing – Renaming columns, removing missing values to get consistent data
- Data Integration – Merging the datasets by the Statistical Area Level 2 Code
- Data Transformation – Normalizing data
- Data Visualization – Pearson Correlation and Scatter plots were used to find the relations between the attributes
- Clustering – Dendrograms were made but did not help too much with the visualizations

PEARSON CORRELATION

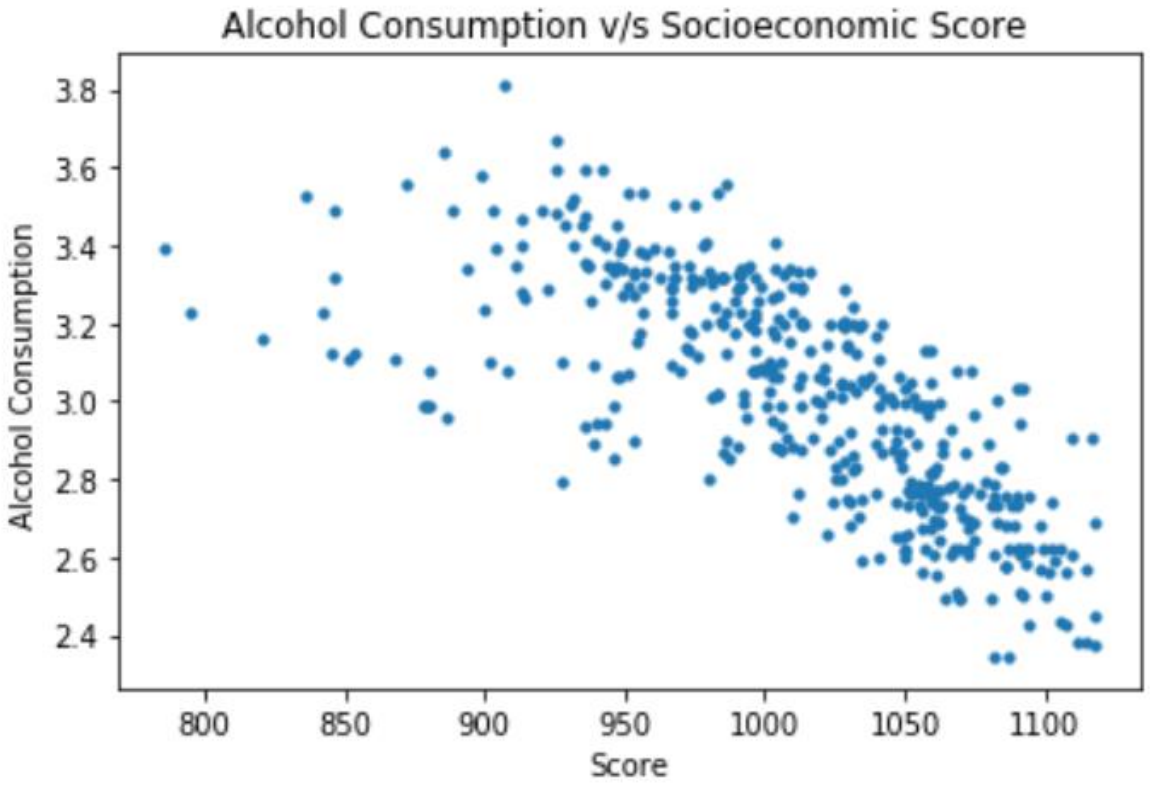
- Deductions:

- Strong negative correlations between Income and Smokers (-0.82) and Income and Alcohol Consumption (0.80)
- Negative correlation between Obesity and Score (-0.65)
- Very strong positive correlations between Smoking, Alcohol Consumption, Obesity and Mental problems
- Generally strong negative correlations between Score and the health risk factors (ranges from -0.59 to -0.82)

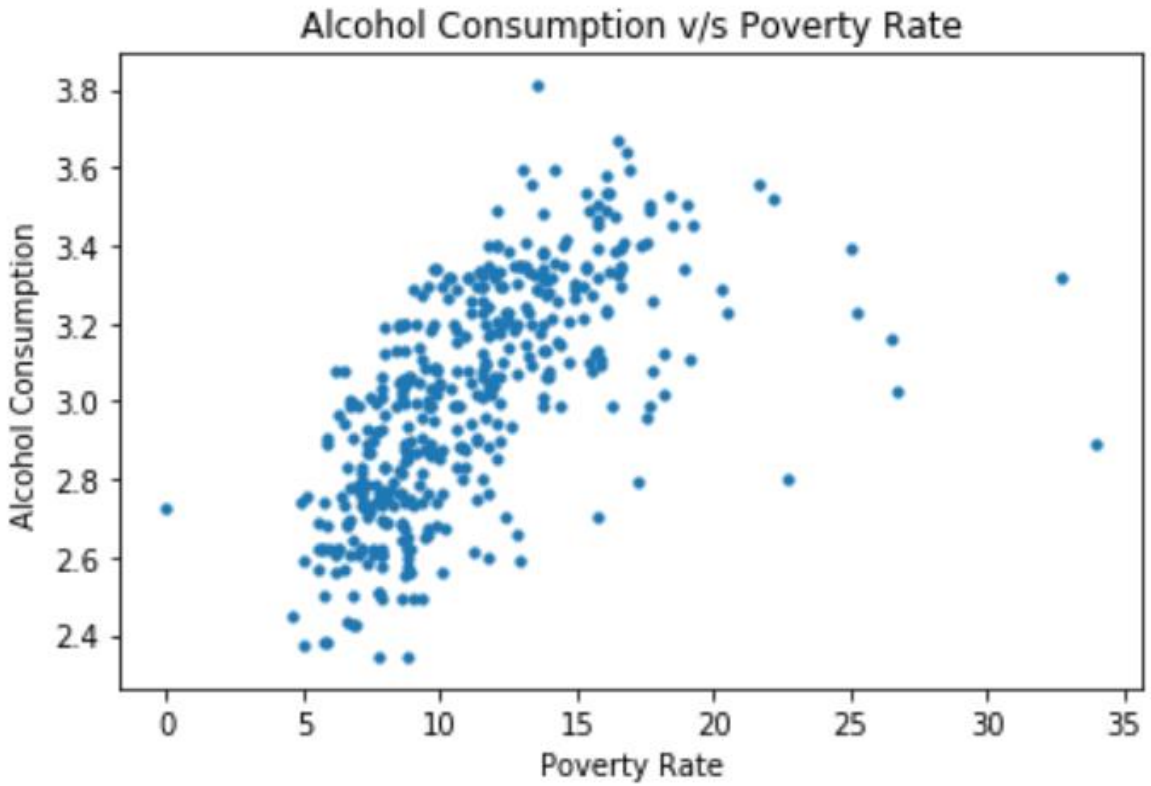


ALCOHOL CONSUMPTION V/S POVERTY RATE AND SOCIOECONOMIC SCORE

- Positive Relation

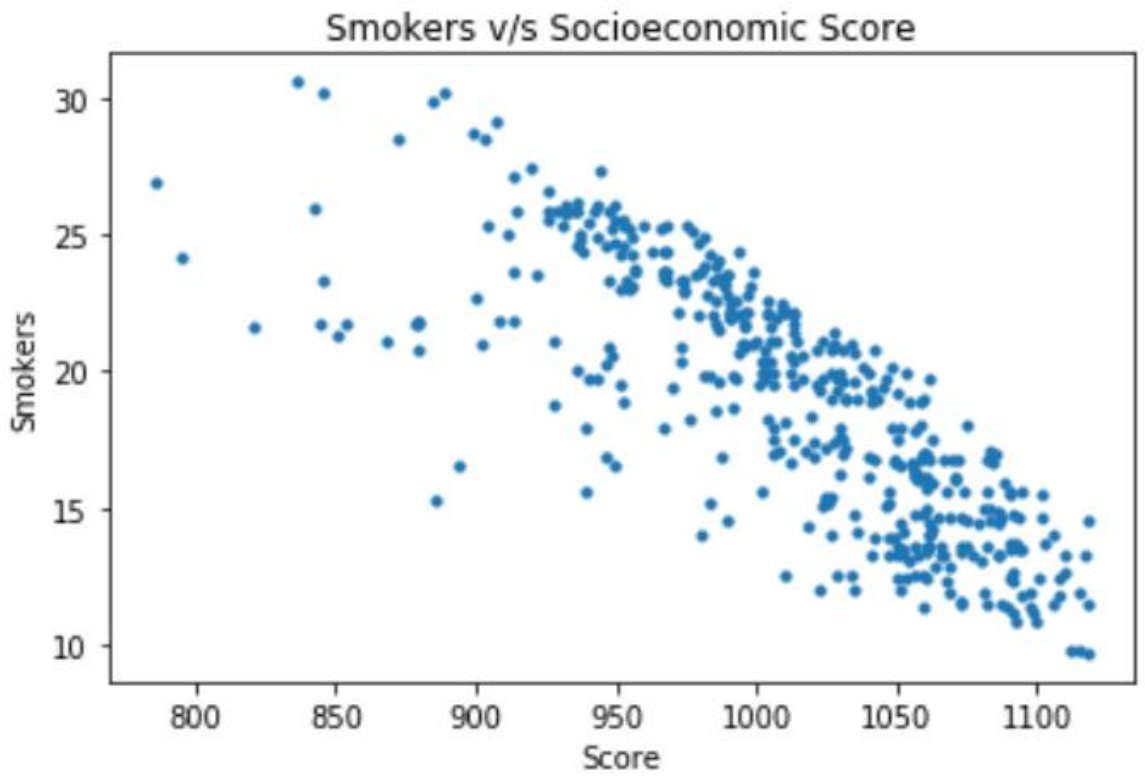


- Positive Relation

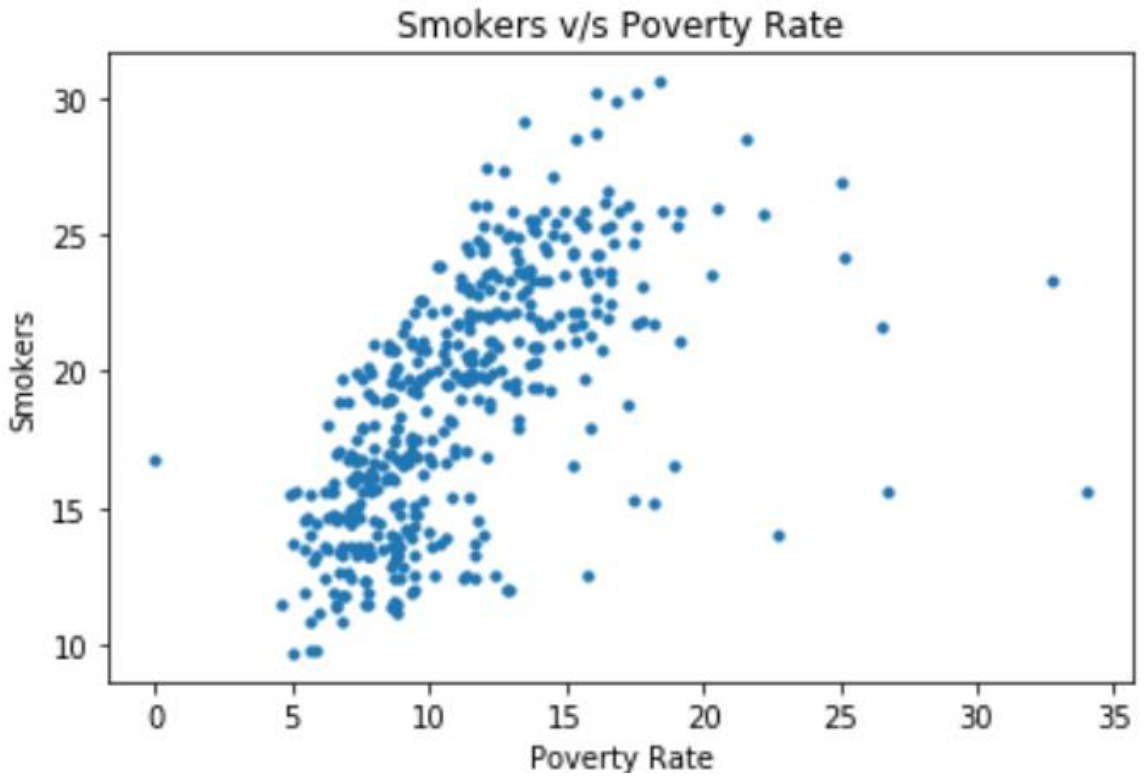


SMOKERS V/S SOCIOECONOMIC SCORE AND POVERTY RATE

- Positive Relation

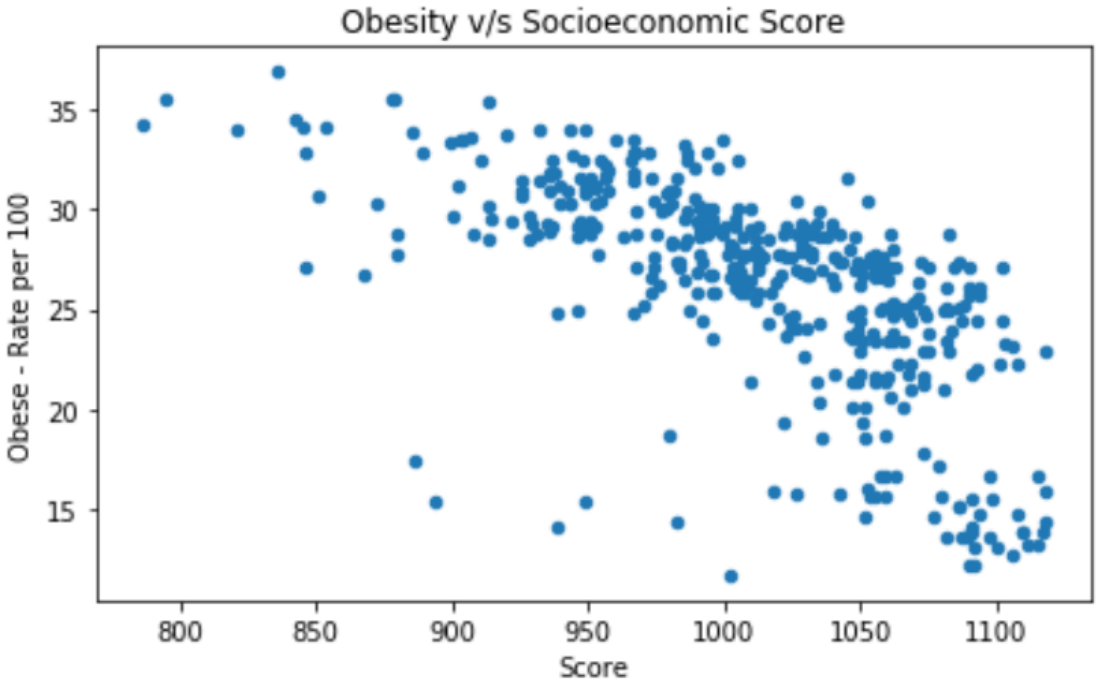


- Positive Relation

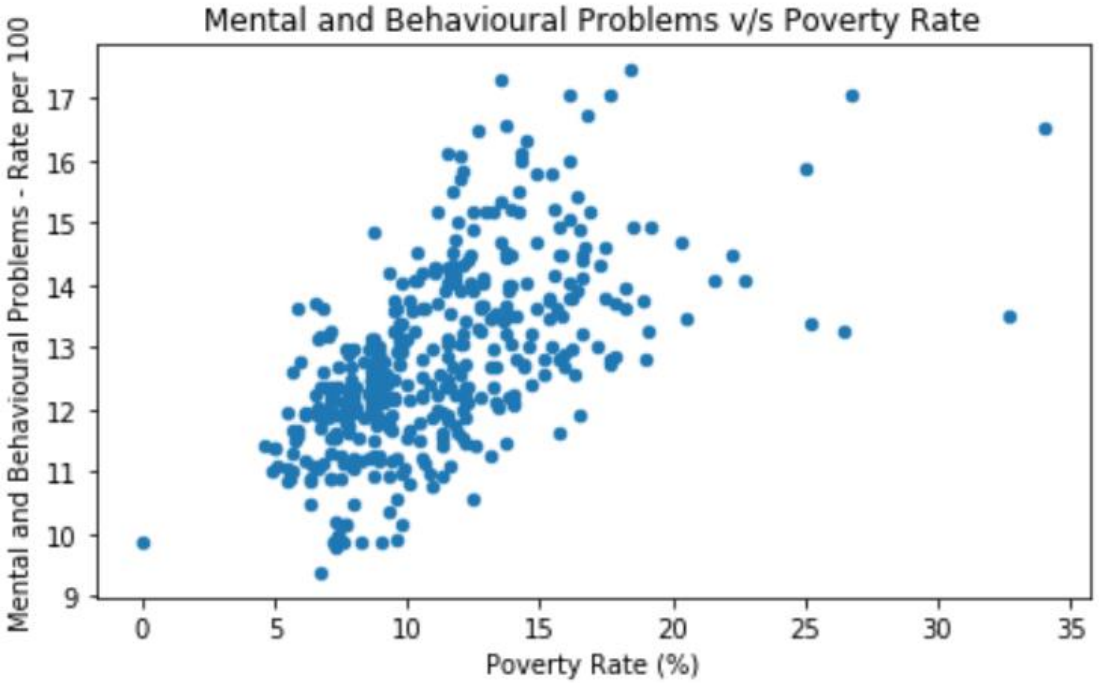


OBESITY V/S SOCIOECONOMIC SCORE
MENTAL AND BEHAVIORAL PROBLEMS V/S POVERTY RATE

- Negative Relation



- Positive Relation



CHALLENGES

- Finding data from the same geographical region (SA2) to compare data attribute-wise or to merge data
- Since most attributes were affected by more than one attribute, it was difficult to find the exact correlation between two attributes
- Data preprocessing - Removing missing values and renaming columns

IMPROVEMENTS

- Account for confounding factors to get greater accuracy for the results
- Consider more health risk factors from different datasets