Busara Depression Dataset – Exploratory Analysis

Introduction:

The 2015 data, gathered by the Busara Center, Siaya County in Kenya has 1143 objects and explains the rural locals' economic situations, expenditures, and familial makeup. The modified data contains 24 features. Three variables; survey id and village id may be useful when examining clusters further, and the final variable is the classifier: depressed (0=not depressed, 1=depressed). The rest of the 21 variables are as follows:

age: in years; **married** (binary, 0 or 1); **children**: birthed; **edu**: education level; **hh_children**: living in the household; **asset_livestock**: value of livestock ($); **asset_durable**: value of durable goods ($); **asset_phone**: value of cell phone ($); **asset_savings**: value of savings ($); **asset_land_owned_total**: Land owned (acres); **cons_allfood**: food total ($); **cons_ownfood;** food own production ($); **cons_ed**: Education expenditure ($); **cons_socia**l: Social expenditure ($); **cons_other**: other expenditure ($); **ent_ownfarm:** own farm primary income (0 or 1); **ent_nonagbusiness**: Non-agricultural business owner (0 or 1); **ent_farmexpenses**: farm flow expenses, monthly ($); **ent_animalstockrev:** livestock sales and meat revenue, monthly ($); **fs_adskipm_often**: meals skipped by adults in last month; **labor_primary:** casual or wage labor primary source of income (0 or 1).

Note: *According to the original data source, 0 = no/false and 1 = yes/true. Additionally, $ refers to US dollars and the children and hh_children variables are coded as factors with levels: 1 [Below average], 2 [Average] and 3 [Above Average] which was completed during the data cleansing phase. Further details in R code.*

Data Cleansing:

The original dataset contained 75 columns and were reduced based on the following criteria:

- **Missing Values:** Columns which contained over 75% missing values were removed
- **Same Values:** Columns which contained values that were all the same (i.e. where the variance was 0). Femaleres was retained to see if depression rates varied by gender specifically.
- **Excess Values the Same:** Columns which contained over 80% 0s or 1s (implying that a majority of the participants answered similarly to yes/no questions e.g. they didn't have employees)

Further data cleansing involved only taking the 2 significant values given for age as some were formatted as decimals. Additionally, all monetary values were rounded to two decimal places and it was discovered that following this step, there were no missing values at all (by chance) in any of the rows as well as zero duplicates. The children, hh_children and hhsize variables were recoded. First it was determined what the average, minimum and maximum values were for each. Using these results the variables were sorted into 1 = Below Average, 2=Average and 3=Above Average so that it would be easier to determine whether, relative to other households, having more or less children/family members affects the depression prediction result?

Overall, this step brought forward 34 total variables to the exploratory analysis stage where further variables were removed based on insights uncovered due to analysis of both variation and correlation in the data.

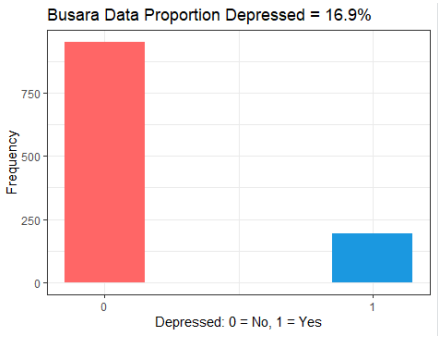
Exploratory Analysis:

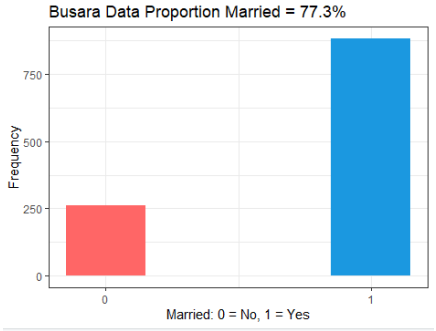
1). Spread and Variance:

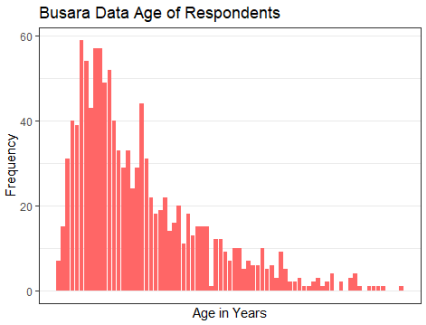
The spread and variance of the data was analysed with respect to, first the demographic variables including marriage status, age, household members and then secondly, to the economic variables including assets, land etc.

- **Demographic:** It was established that 16.9% of the people included in the dataset are depressed. There is relatively little difference between rates of depression by gender (17.9% for 0 vs 16.9% for 1). Since femaleres=1 made up 91.7% of the dataset, but had no significant difference to the femaleres=0, it was removed from the analysis.

Busara Depression Dataset – Exploratory Analysis



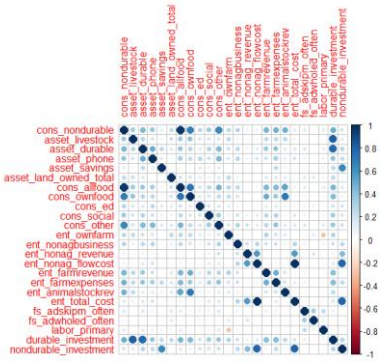


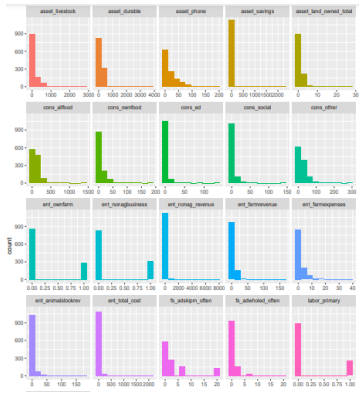


The majority of respondents were married which is expected for a rural area in Sub Saharan Africa, and notably, 22.3% of married people screened positive for depression versus 15.28% of those who weren't married. Most of those surveyed are between 20 and 50 years in age, the majority being 20-30 years old. Finally, when establishing whether the amount of people living in a household or the number of children birthed, had any relation to those who were depressed, it was found that those with an above average amount of children given birth to/raised or living in a household (5+ or 6+) had far higher levels of depression than those with an average or below average amount. Those respondents' rates of depression were around 20.6% vs 15.5%-16.7% for the rest. The number of children had more of an impact on depression than the size of a household, which saw no notable difference in depression rates at various levels and so this was also dropped from the analysis. The majority of people had factor levels of 1 or 2 for all three of those variables with around 8%-12.4% falling into above average ranges.

2). Correlation:

- **Economic:** Using a correlation matrix, the economic variables which were highly correlated to one another were found (the threshold being 0.75 for the correlation coefficient). Since these variables added no new information for predicting depression, they were also removed.





The darker the blue the circle, the higher the correlation with the following variables being highly correlated: **cons_nondurable** and **cons_allfood** (presumably non_durable expenditure is mostly for food); **durable_investment** and **asset_livestock**; **non_durableinvestment** with **ent_ nonag_flow_cost** and also **ent_total_cost;** and **ent_nonag_flowcost** with **ent_animalstockrev**. The first variable mentioned in each of the above groups was removed.

The plot_num function was used to plot all of the variables to uncover any more insights and it was found that **nonag_revenue** and **ent_total cost** still had a substantial amount of 0 values. It seems as if the other economic variables will suffice to explain depression regardless so those two were also taken out. Interesting relationships were found using correlation plots. For example, those more highly educated had lower rates of depression even if they had a more children than those with very little education and few children.

Conclusions: The dataset is quite sparse and has large ranges between a variety of different variables, particularly those with an economic focus where most values hover near or at 0. However, based on early exploratory analysis, there do seem to be a certain number of variables which look to be very clearly related to depression. Further analysis can help determine whether there are any clusters or groups and what characteristics they share which may put people at higher risk. Additionally, when making predictions, the regsubsets function might shed some light on which variables actually have an influence on depression as an outcome which will be very interesting to see, especially taking these variables into consideration all at once.

Summary Statistics: