**Introduction**

In an ideal world, we would be able to eradicate poverty and create upward social mobility for all segments of our population. However, even the best governments in the world find it difficult to formulate socio-economic initiatives and budgets to effectively meet these noble ends. Instead, what we are left with is a world where the rich get richer and the poor get poorer. This effect is particularly exacerbated in developing countries. For example, a quick distribution plot of household income in the Philippines reveals that a huge chunk of the population is still at the lowest rungs of the economic ladder.

To be fair, there are numerous uncertainties in the process of devising government policies. It is not immediately apparent which the most impactful policies will be, as the demographic of country is often diverse. Add to this the complexity of differentiating the actual needs of your population from their wants. The best solution may not be the most popular one. Not many politicians have the education or experience to navigate these uncertainties. As such, it is indeed fathomable that huge amounts of taxpayer monies are squandered every year by governments which may have the best of intentions but lack the understanding to channel their efforts in the right direction.

In summation, we grapple with the challenge of determining the scope, scale and target audience of government policies so as to maximize upward social mobility in a population. We address this issue by providing an analytical solution based on socio-economic data. We aim to utilize unsupervised machine learning to identify clusters within the population. We will then apply a supervised machine learning pipeline containing a regression model, such as multiple linear regression, to recommend the policies which would maximize household income in each of these clusters.

**Dataset Description**

Q.1 What is the source of your data? What is it about?

Q.2 How many features and datapoints does contain?

Q.3 List a few (at most 10) features and describe them.

**Missing Values**

**Q.1** **Are there missing values in the dataset?**

**Ans.** Yes

**Q.2 If yes, please provide the statistics (such has how many, which features, etc.) of the missing values.**

**Ans.** 7536 values were found missing from two features (Household Head Occupation and Household Head Class of Worker)**.** These values were object values.

**Q.3 Do you want to remove such data-points from the dataset? Why?**

**Ans.** We do not want to remove them as they may hold statistical significance. We are attempting to predict Total Household Income, which could significantly be influenced by the Household Head's occupation class. Instead, we should try to estimate and impute these missing values.

It can be reasonably assumed that a combination of factors such as age, gender and educational qualifications would affect one's occupation. As such, we can examine the correlation of 'Household Head Occupation/Class of Worker' with other features of the 'Household Head ...' variety. This can be used to estimate the missing values.

**Q.4 Do you perform imputation to fill in missing values? What technique would you use?**

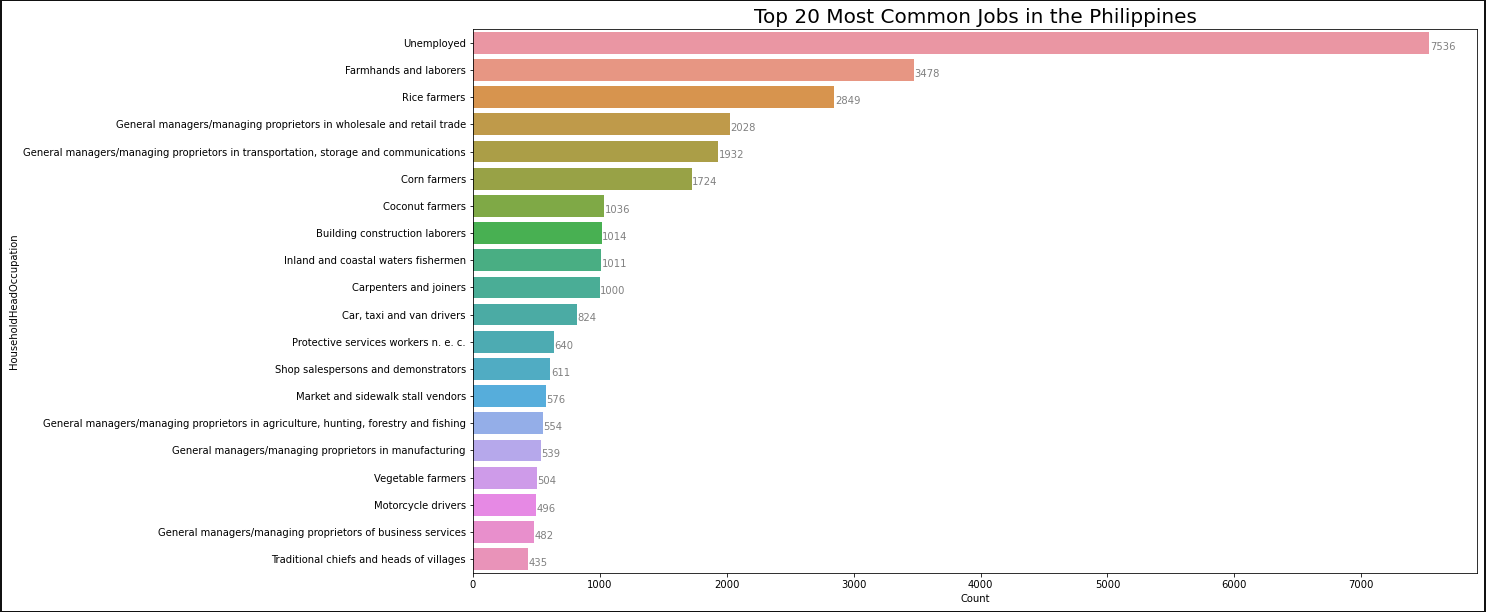
**Ans.** Upon closer examination of the 'Household Head' features, we found that the 7536 null values occur because they are unemployed.

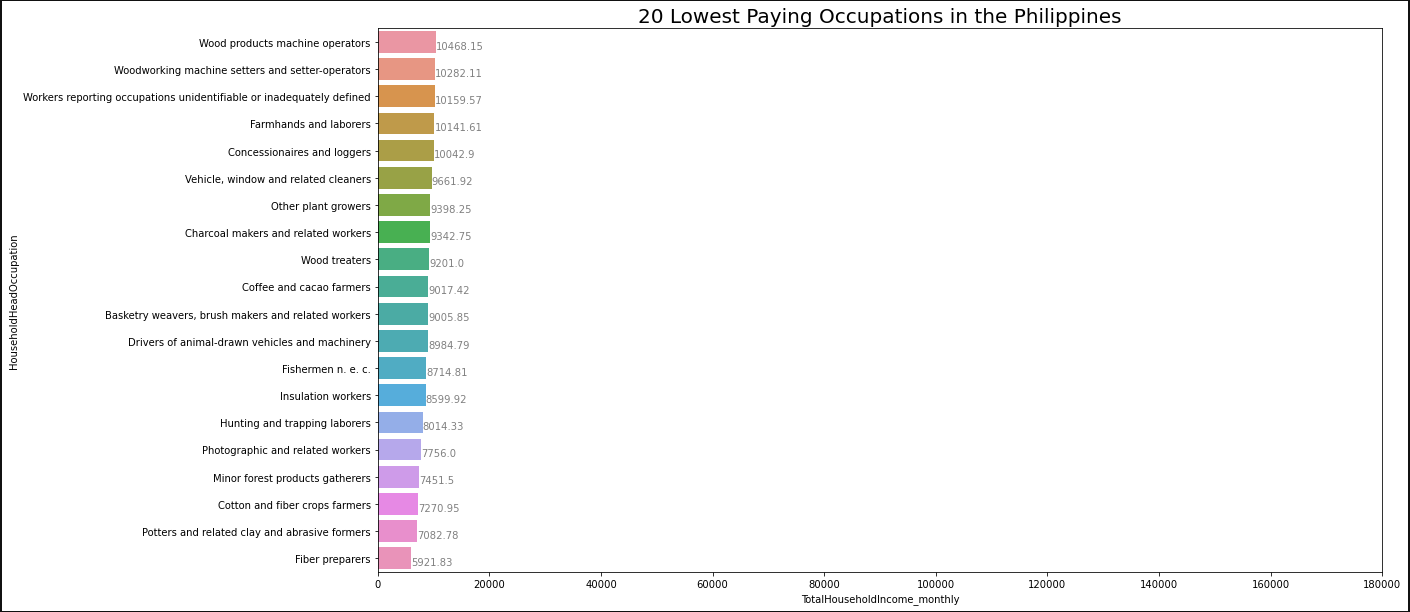
We can therefore replace the null values with 'Unemployed' as a categorical variable, which still holds analytical value.

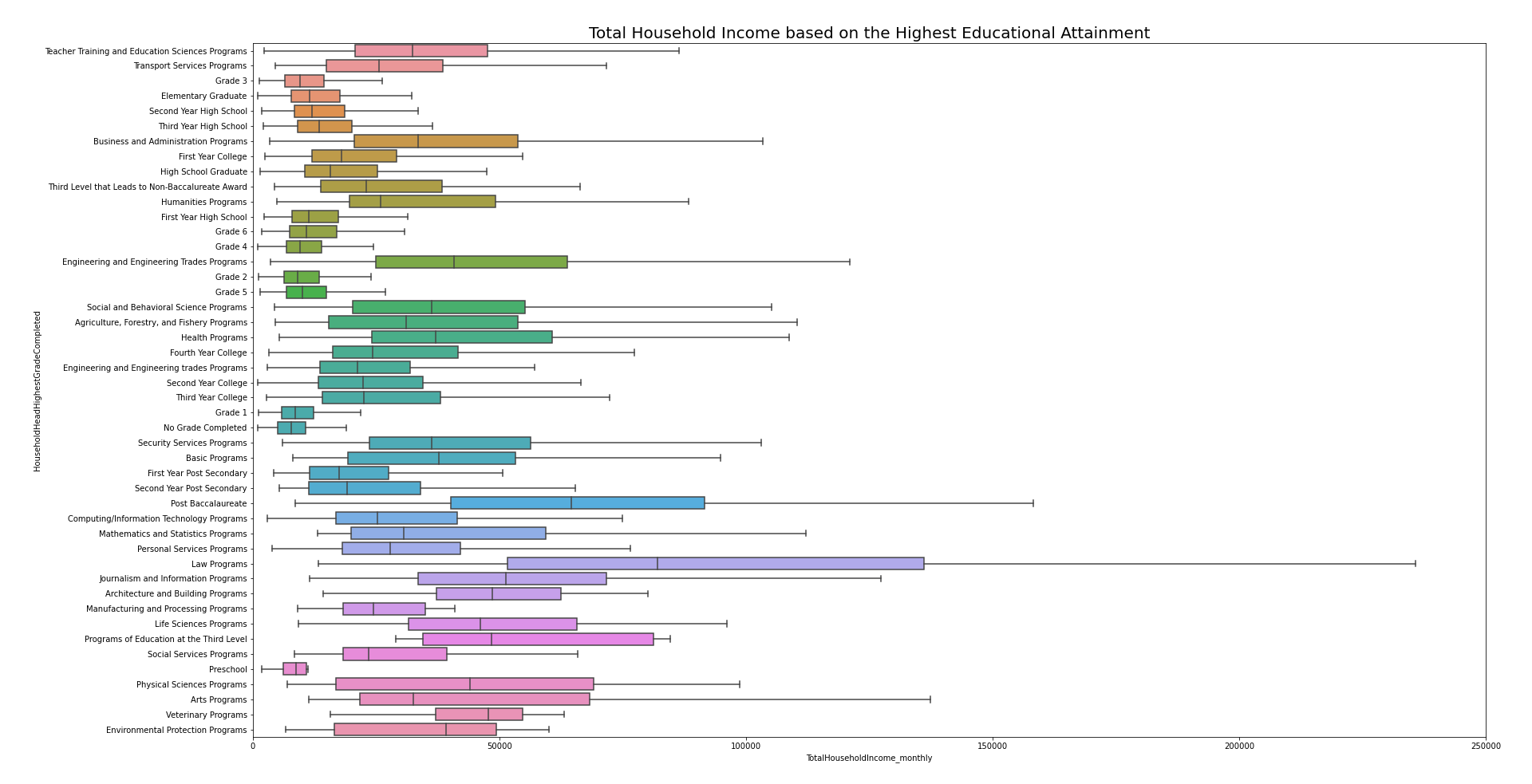
**Features**

Q.1 Please show some interesting patterns (with the help of visualizations) that you would have discovered by visualizing certain features in the data.

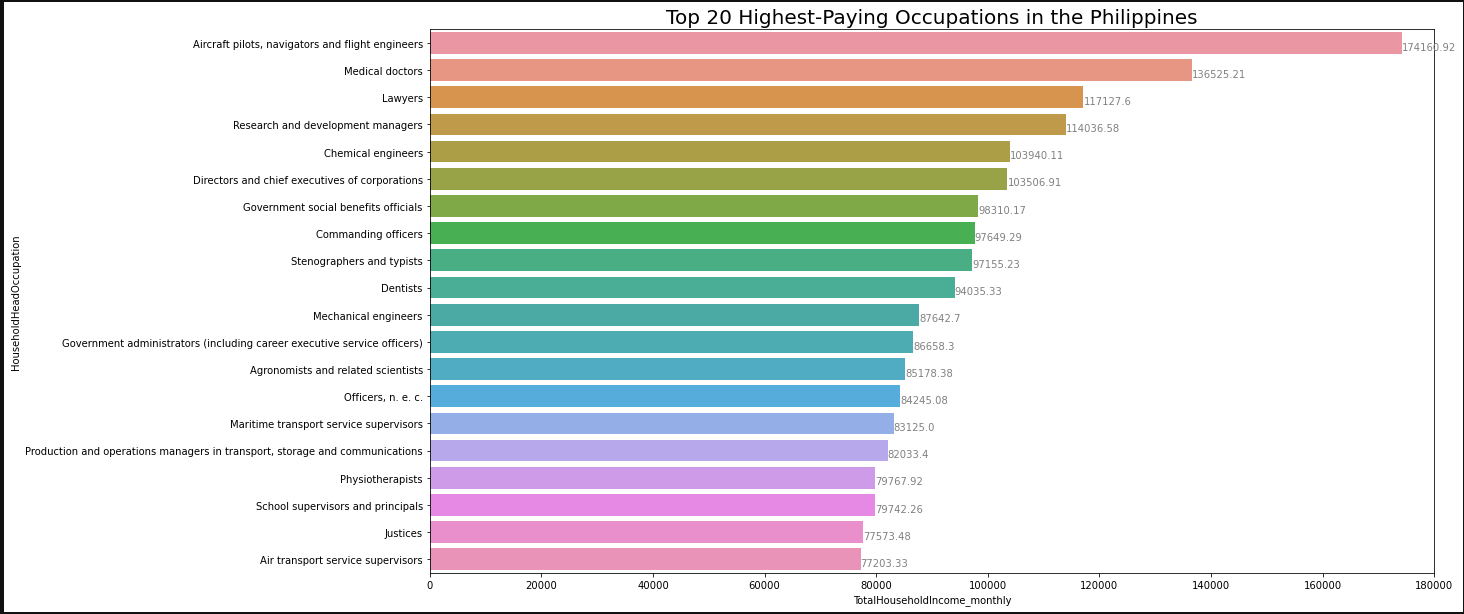
**PLOT-1**

****

**PLOT-2**

**PLOT-3**

**PLOT-4**

****

Q.2 Does exploratory data analysis help you to remove a few features? Which and how?

**Outliers**

Q.1 Have you found any outliers in your data? How have you found them?

Q.2 Do you plan to remove them to keep them in your data? Why?

**Data Scientist**

Q.1 State your problem statement along with the motivation.

Q.2 How does your exploratory data analysis support the problem statement?

Q.3 Propose a plan of action to address the problem. (Which model do you plan to use?)