|  |
| --- |
| An Analytical Approach to Tackling Income Inequality |
| Data Scientist Track  Project Milestone 1 |

|  |
| --- |
| **Team 17**  Ramesht Shukla  Supratik Sekhar Bhattacharya  Arvind Subramanian |
| 26/02/2021 |

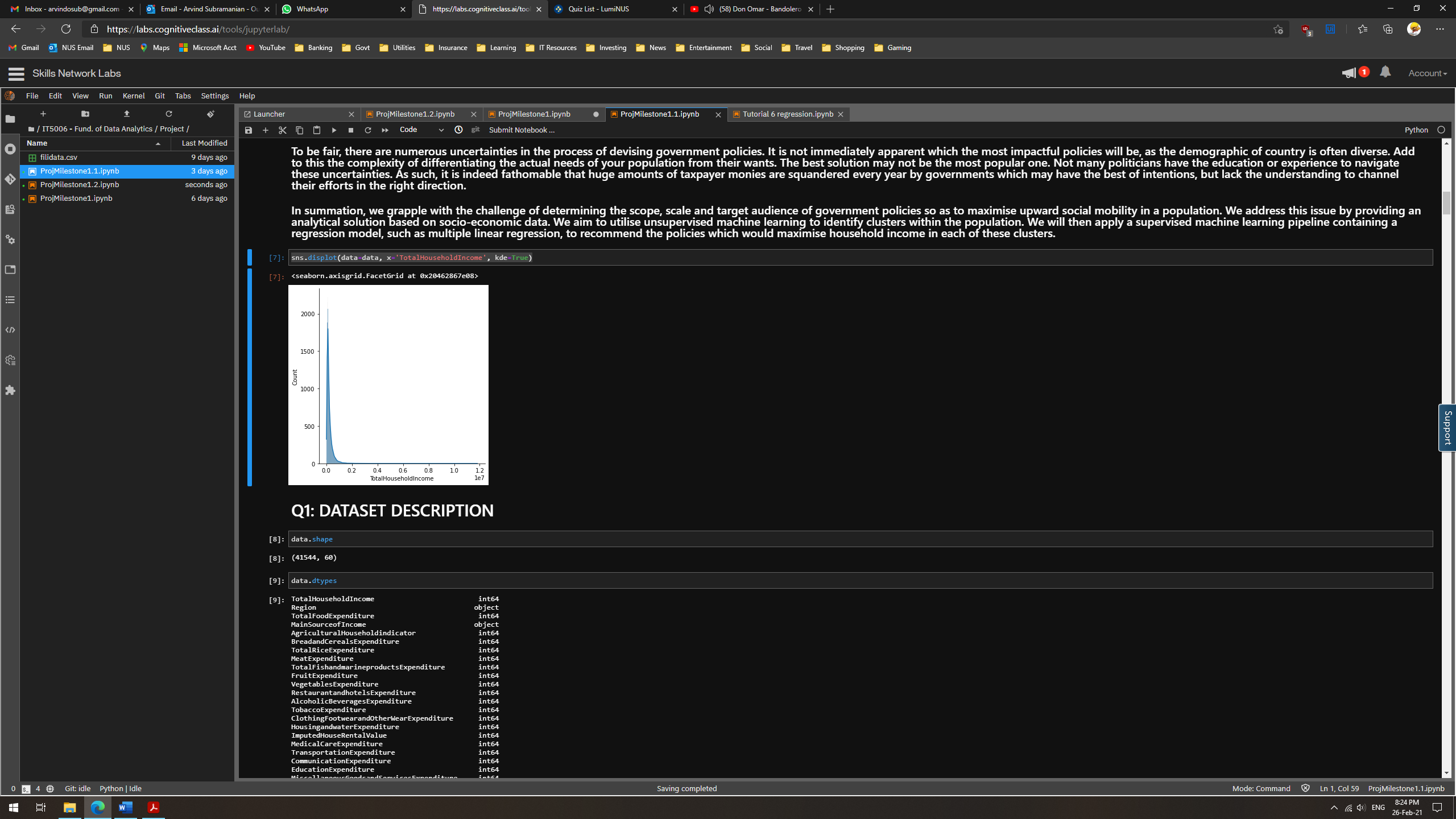
**Problem Statement**

‘Curating Government Policies to Maximize Upward Social Mobility in a Country.’

**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 (Fig 1) reveals that a huge chunk of the population is still at the lowest rungs of the economic ladder.

Fig 1: Plot of Income Distribution in the Philippines



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. There are numerous factors affecting income, both in the long-term and short-term. 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. We plan to provide a data analytics-based solution which would address this issue on two levels.

Group Level: In our analysis, we need to segregate the population into multiple socio-economic groups. Different segments of a country’s demographic would need different policies to maximise their household income. For example, we cannot expect all the policies that work for a white-collar worker to be effective for a farmer. We also cannot ‘convert’ the employment types of people to maximise their income. As such, we are limited by the structure of the country’s economy and need to be cognizant of which factors we can meaningfully influence. We should not attempt a ‘one size fits all’ approach.

National Level: This level would involve comparing across the previously mentioned social groups to identify which factors would enable upward social mobility or which groups need more assistance. For example, we may find a strong correlation between income and education expenditure. Households with higher income may spend more on education, which would enable them to maintain higher incomes in the future. Therefore, subsidising education for lower-income families could be a possible solution to close the income gap.

**Plan of Action**

We aim to determine the scope, scale and target of government policies so as to maximize household income and upward social mobility in a population. This will be achieved as follows:

1. Segmenting our dataset for Group Level analysis. This can be done by dividing the dataset along delineations such as employment type.
2. Utilizing unsupervised machine learning to perform clustering. This could pinpoint hidden trends in the data.
3. Applying a supervised machine learning pipeline containing a regression model, such as multiple linear regression. This would enable us to identify the combination of features most likely to impact income, and thereby recommend the appropriate policies.
4. Repeat steps 2 and 3 for National Level analysis.

**Dataset Description**

Source of Data and Background

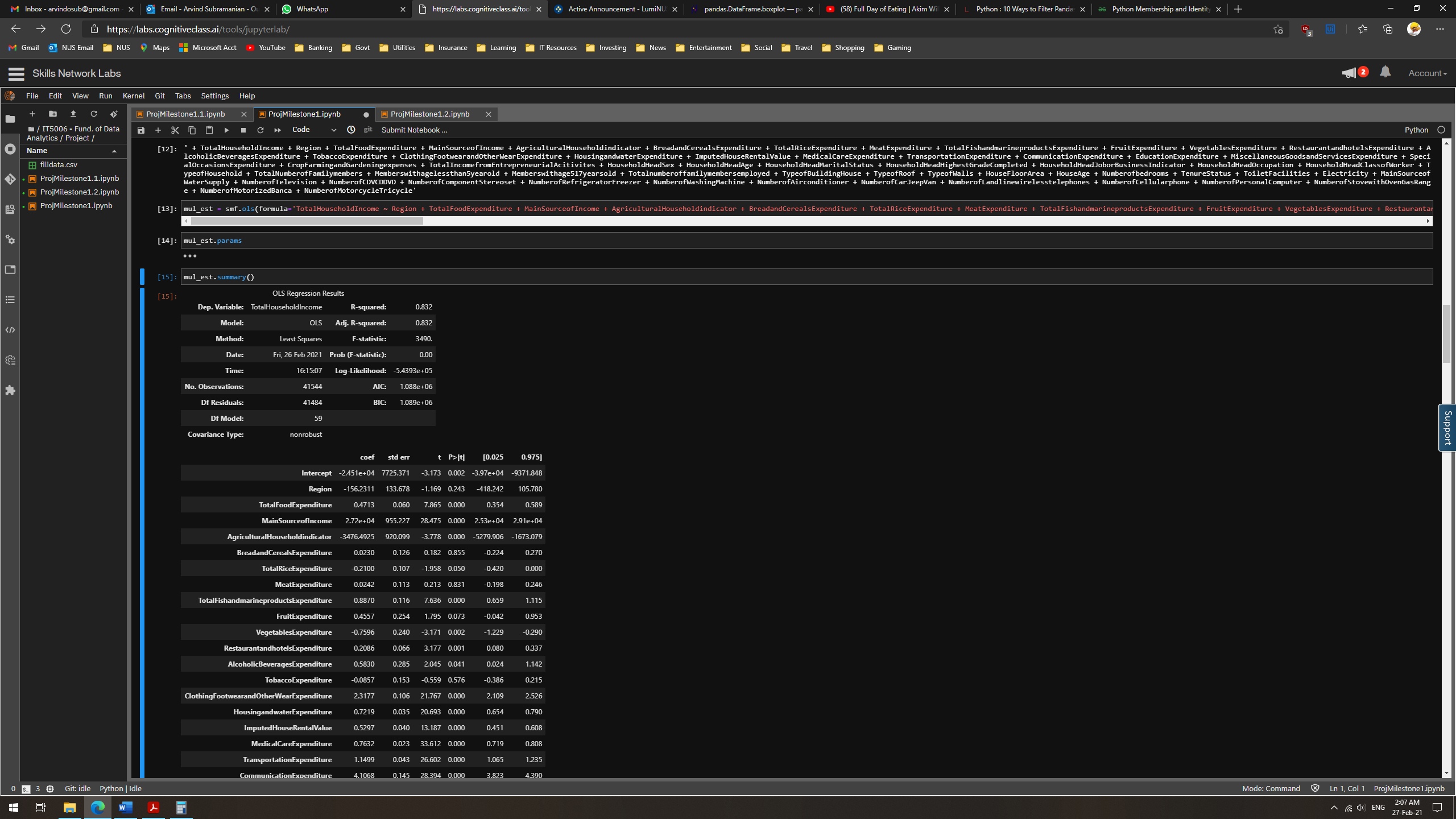
The data was obtained from the records of the Philippine government. It is the census data gathered from the Family Income and Expenditure Survey, which was carried out by the government. It contains various indicators and metrics which illustrate the socio-economic background, income and spending habits of the demographic in Philippines. The survey cuts across several sectors such as education, employment, healthcare and housing. As such, this dataset will enable us to perform a holistic analysis of how variables across multiple sectors correlate to influence household income.

Features and Data Points

The dataset contains 60 features and 41544 datapoints.

As our primary analysis objective is addressing income inequality, the target feature has been designated to be ‘Total Household Income’. Our goal is to discover the combination of features that would maximize this value for each segment of the population. Preliminary testing of our data using multiple linear regression has yielded a strong R-Score.

Fig 2: OLS Regression Results



Description of Features

There are several features that caught our attention, such as:

Total Household Income: Our target variable. Provides the combined income across all members in the household.

Region: Provides the area of residence of each family and can be used to estimate the overall wealth distribution across the Philippines.

Household Head Occupation: Describes the job or class of work of the household head. It can have a huge impact on the Total Household Income as well.

Total Number of Family Members: This would significantly influence the expenditure of the family in all sectors as more family members = more mouths to feed.

Total Food Expenditure: Illustrates the amount spent by each family on the basic essential of food. Can be used to calculate the food expenditure per person in a family to further analyse their spending habits.

Education Expenditure: Illustrates the amount spent by each family on education. Can be compared alongside household income and household education level to assess the accessibility of education.

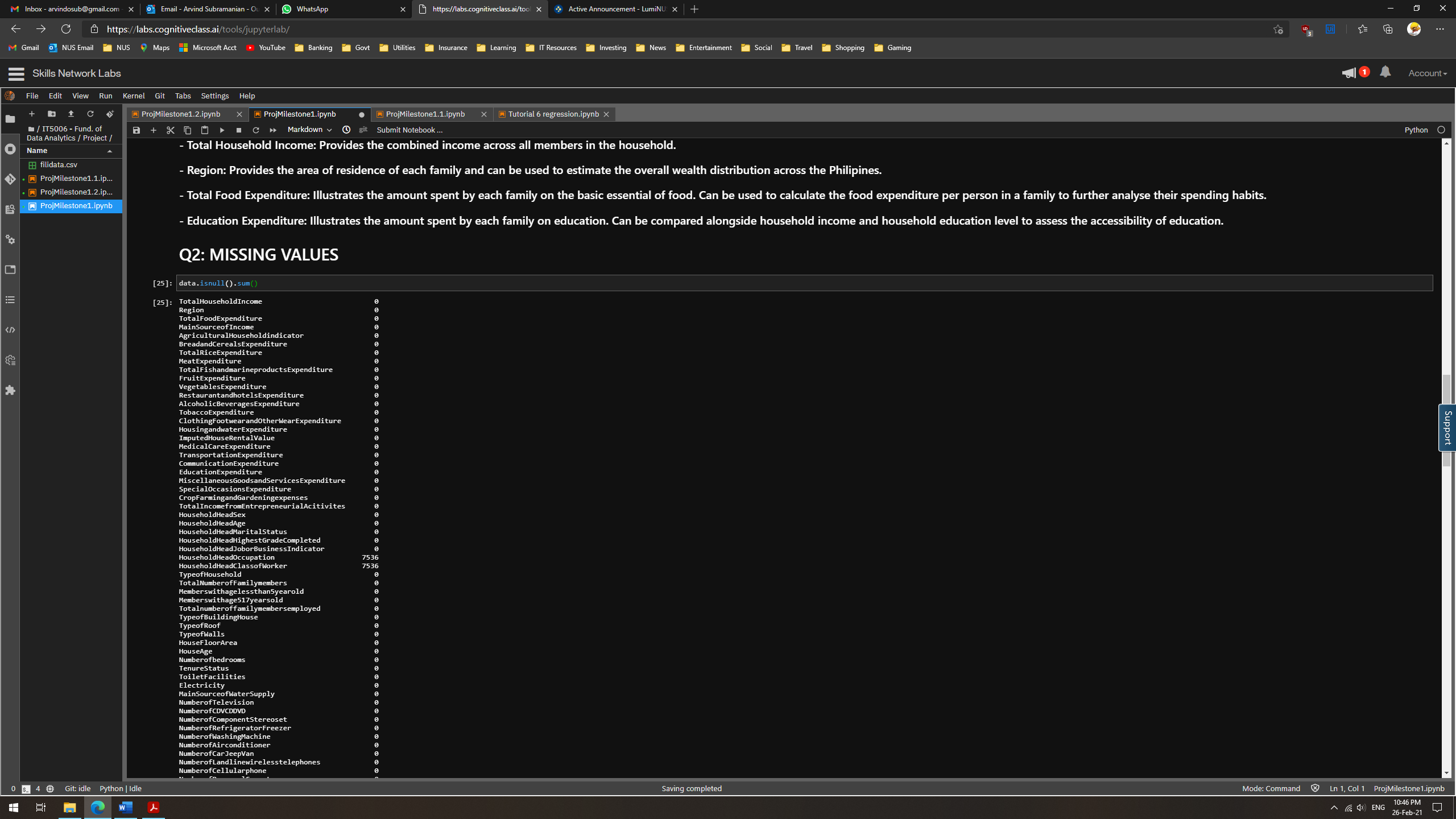
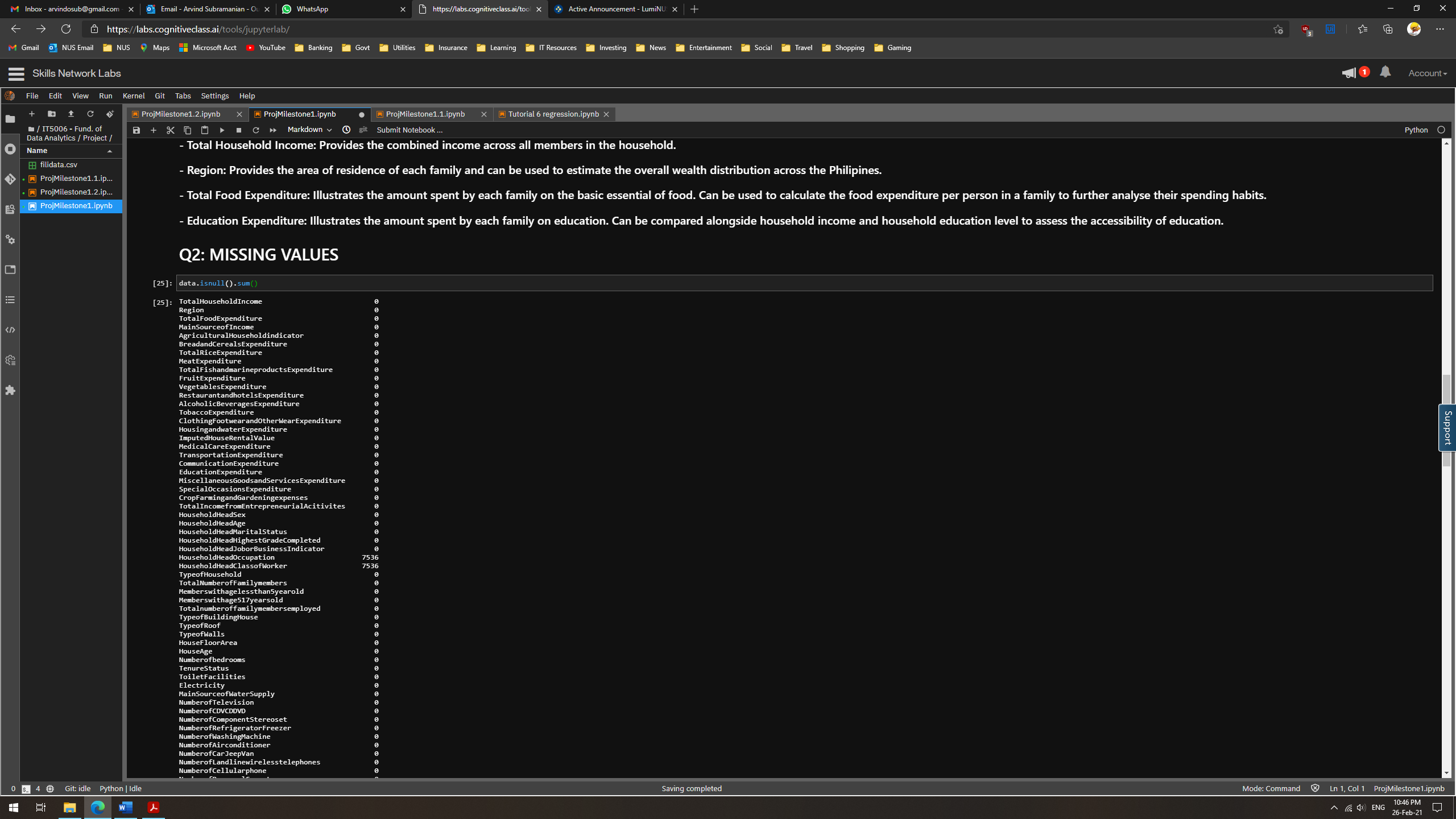
Medical Care Expenditure: Medical costs can put a severe dent on the income or even affect the earning capacity of individuals. This feature could help to identify if certain segments of the population incur higher medical costs and if so, why.

**Missing Values**

Details of Missing Values

Only 2 features have missing data, each with 7536 null values (Household Head Occupation and Household Head Class of Worker). These values were object values.

Fig 3: Missing values in data.



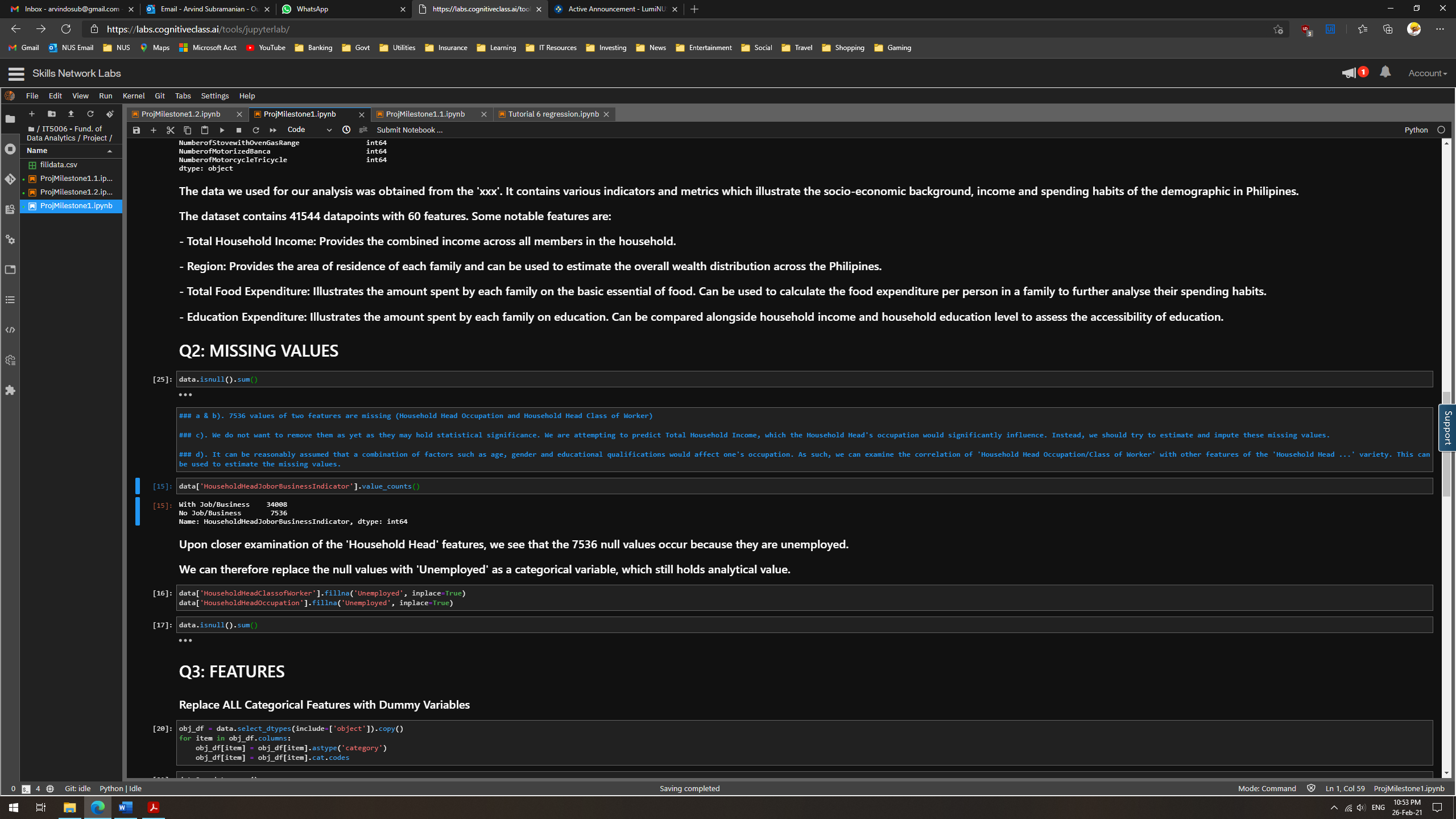
Managing Missing Data

Removal: We do not want to remove them yet as they may hold statistical significance. We are attempting to predict Total Household Income, which the Household Head's occupation would significantly influence. Instead, we should try to estimate and impute these missing values.

Imputation: This is our preferred strategy. 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.

However, upon closer examination of the 'Household Head' features, we see that the 7536 null values occur because they are unemployed.

Fig 4: Number of Unemployed in Data.



We can therefore replace the null values with 'Unemployed' as a categorical variable, which still holds analytical value. Thus, the number of missing values in the dataset was ascertained to be zero.

**Data Preparation**

Though the missing values were filled, there was still some data preparation left to do. The dataset needed some modifications to support future analytics work.

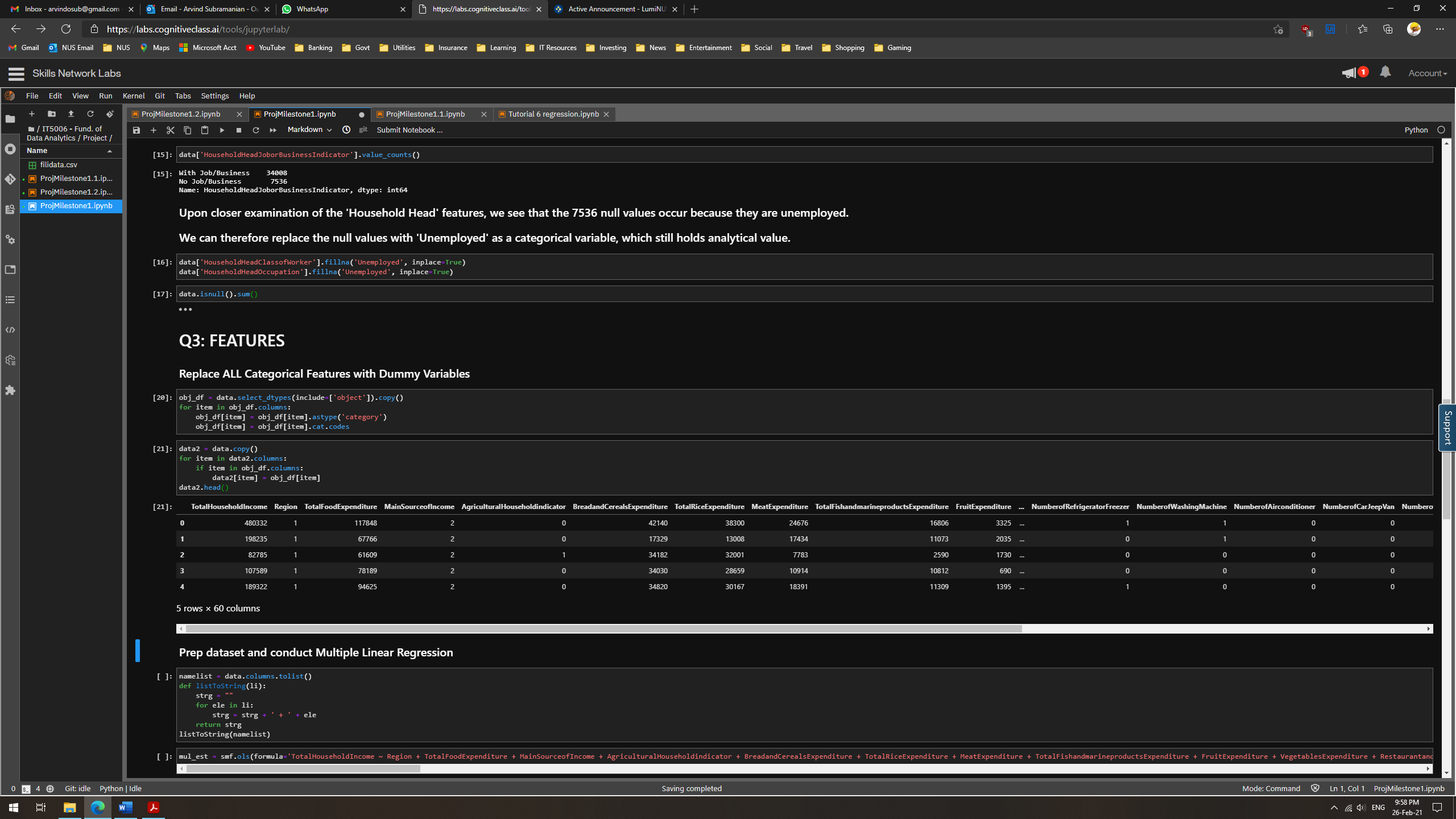
Firstly, we noticed that the feature names had spaces and other special characters. This would cause errors when using python libraries such as statsmodel. As such, we opted to remove all special characters in the feature names.

Fig 5: Removal of Special Characters in Data



Secondly, some features where categorical variables in nature and contained ‘object’ data types. They would be difficult to analyse against integer type variables. As such, we performed categorical encoding on these variables to insert dummy integer values.

Fig 6: Replacing Categorical Features with Dummy Integer Variables.



This concluded our Data Preparation for now and we can proceed to perform visualizations.

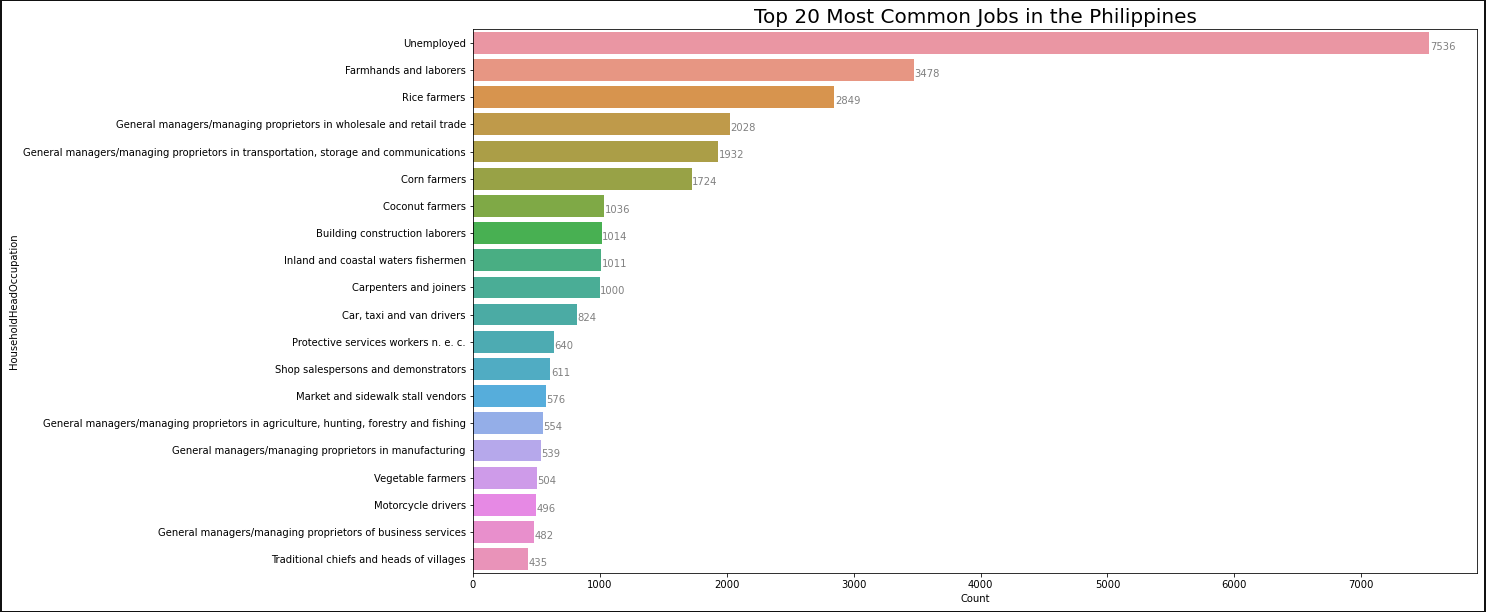
**Features**

There were some important patterns that we wished to observe using visualizations.

Common Occupations

This would allow us to identify the composition of Philippines’ economy and the distribution of its workforce. 7536 unemployed equates to approximately 18% of the families surveyed. This significant portion of the population is surviving on little to no income.

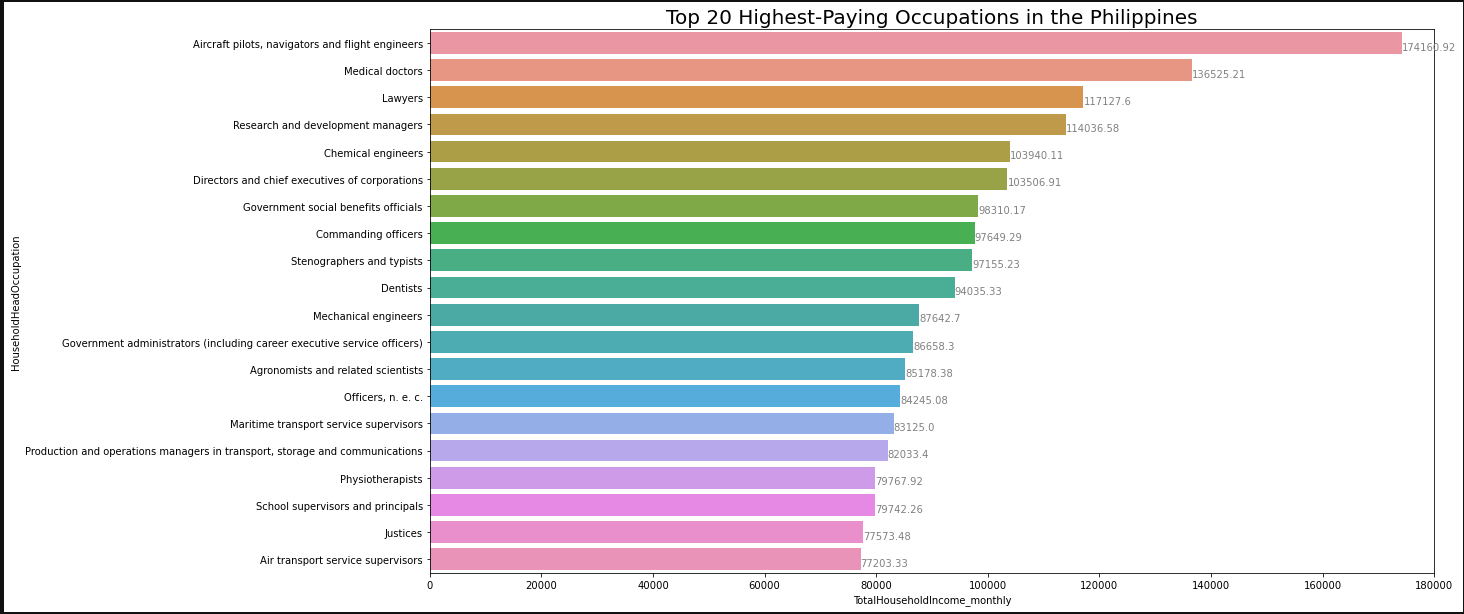
Fig 7: Most Common Jobs in the Philippines

****

Most/Least Lucrative Occupations

This could allow us to identify ‘classes’ of workers and thereby provides a good starting point to segment the population. It is possible that workers within these classes also share other traits.

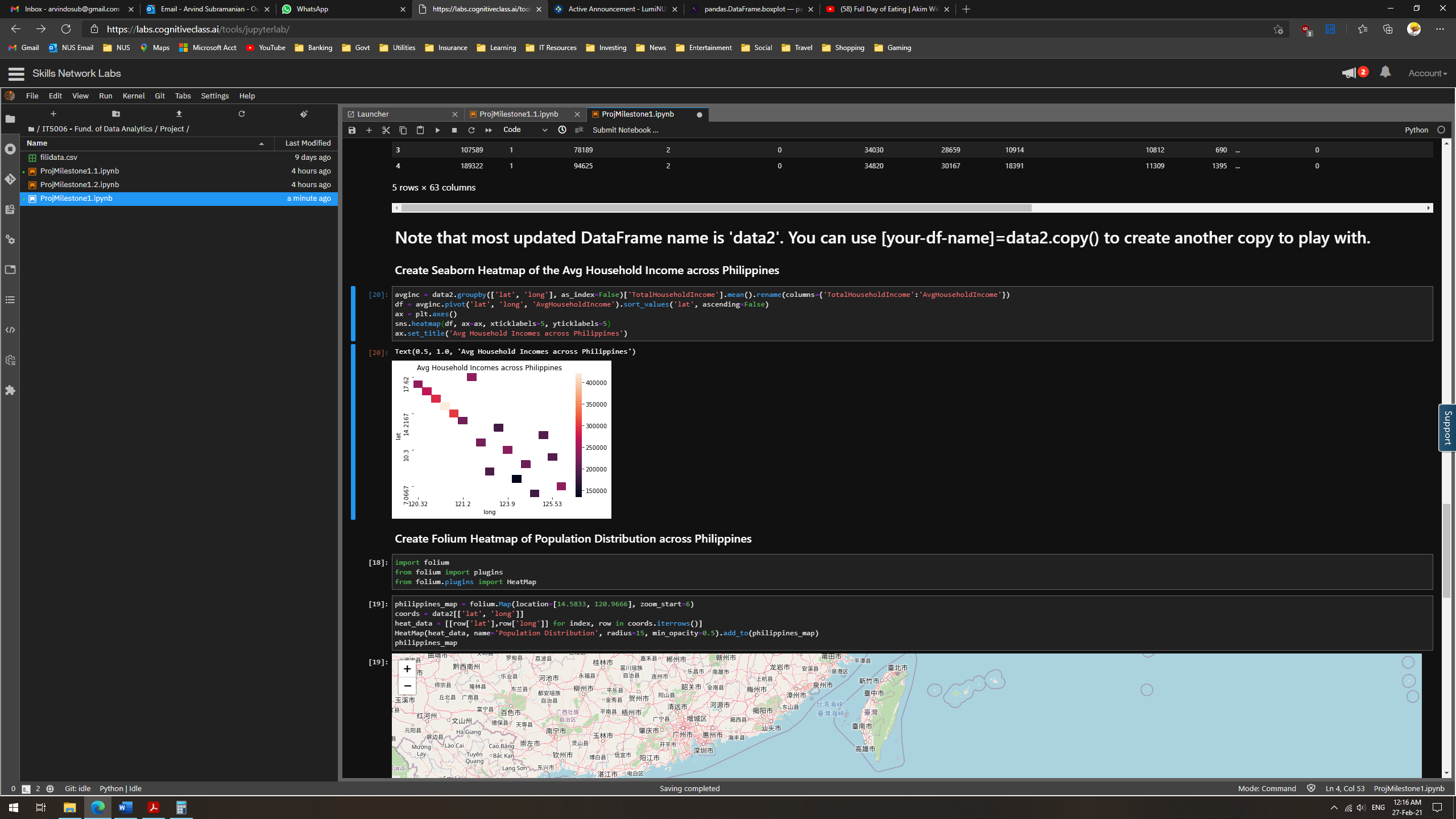
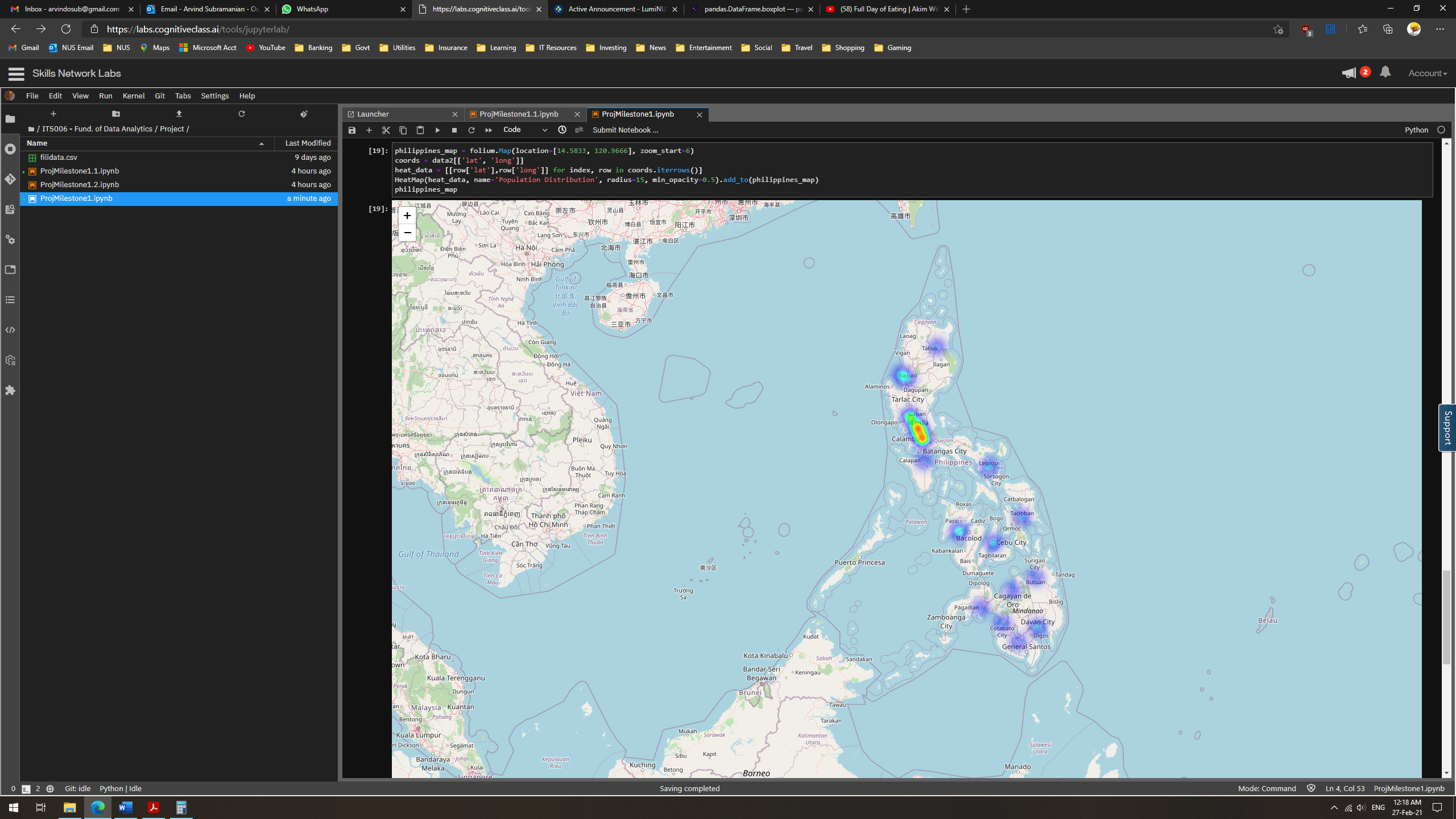
Fig 8: Most Lucrative Jobs in the Philippines

****

Geographic Distribution of Population and Income

Lat/Long data was sourced and appended to the dataset. This allows us to visualize the spread of the population and could lead to further analysis on the most prevalent occupations in each region as well.

Fig 9: Geographical Population and Income Distribution



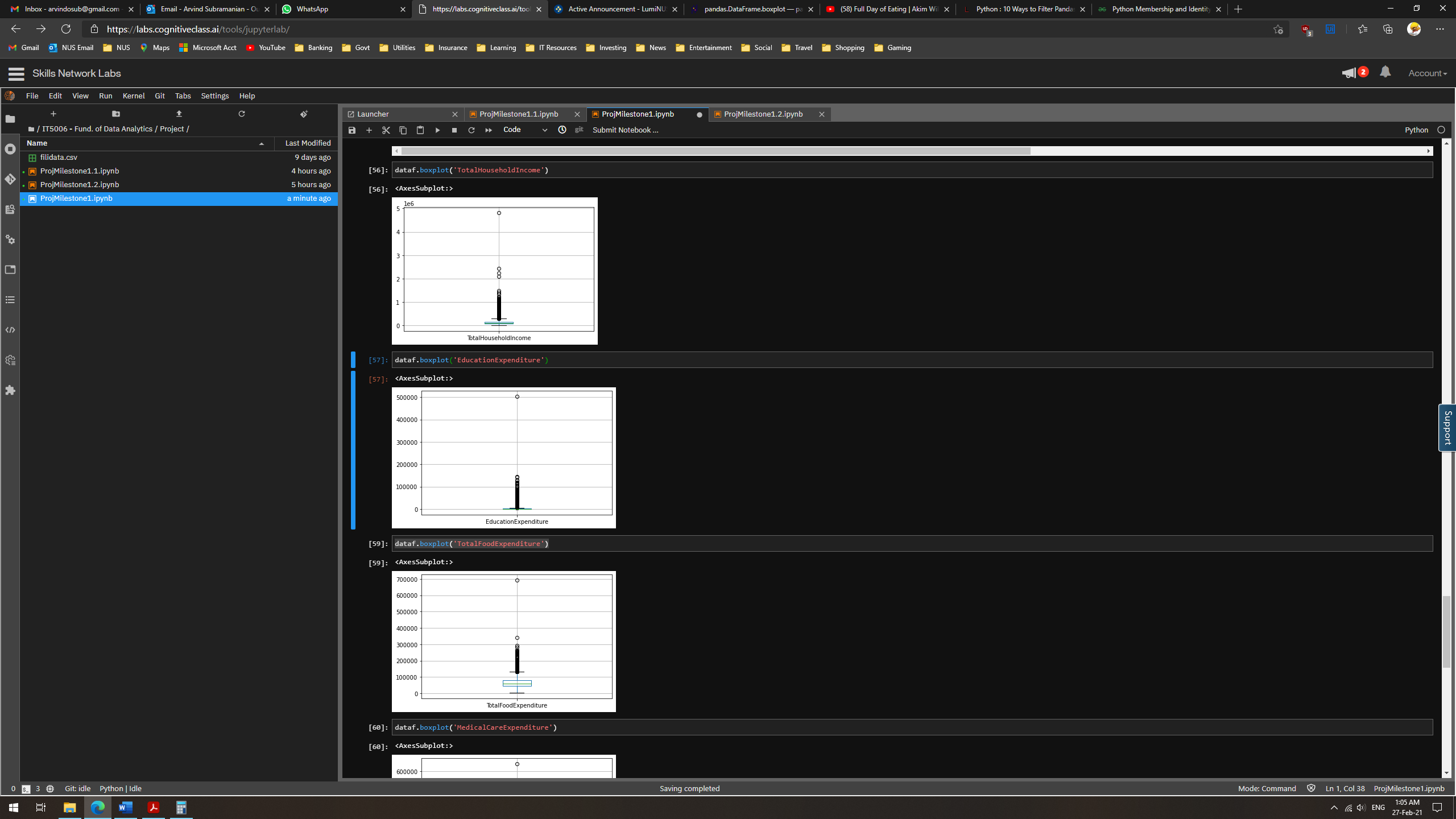
This shows that the bulk of the population resides in the large cities. The average incomes in these cities also seems to be correspondingly higher.

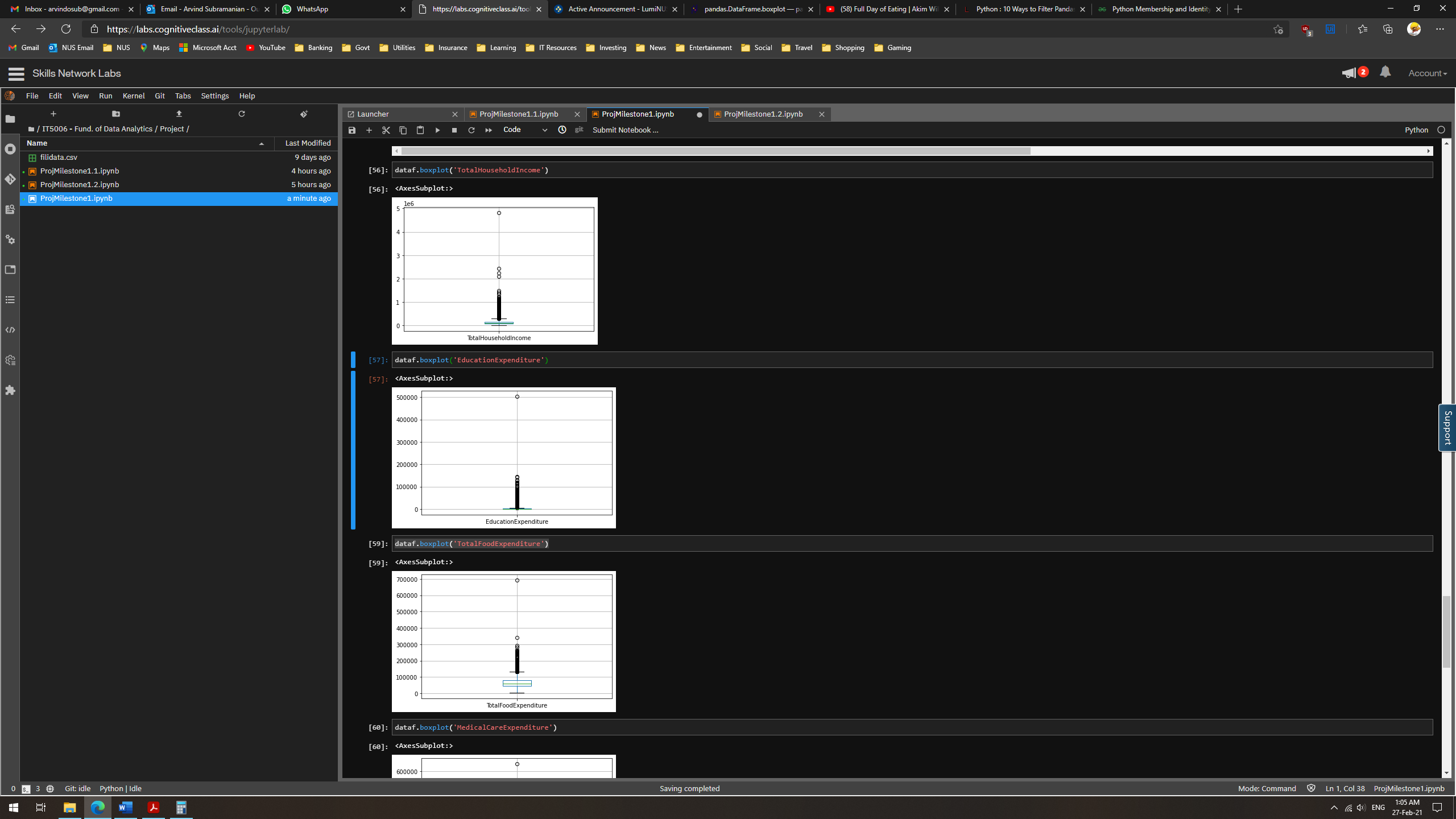
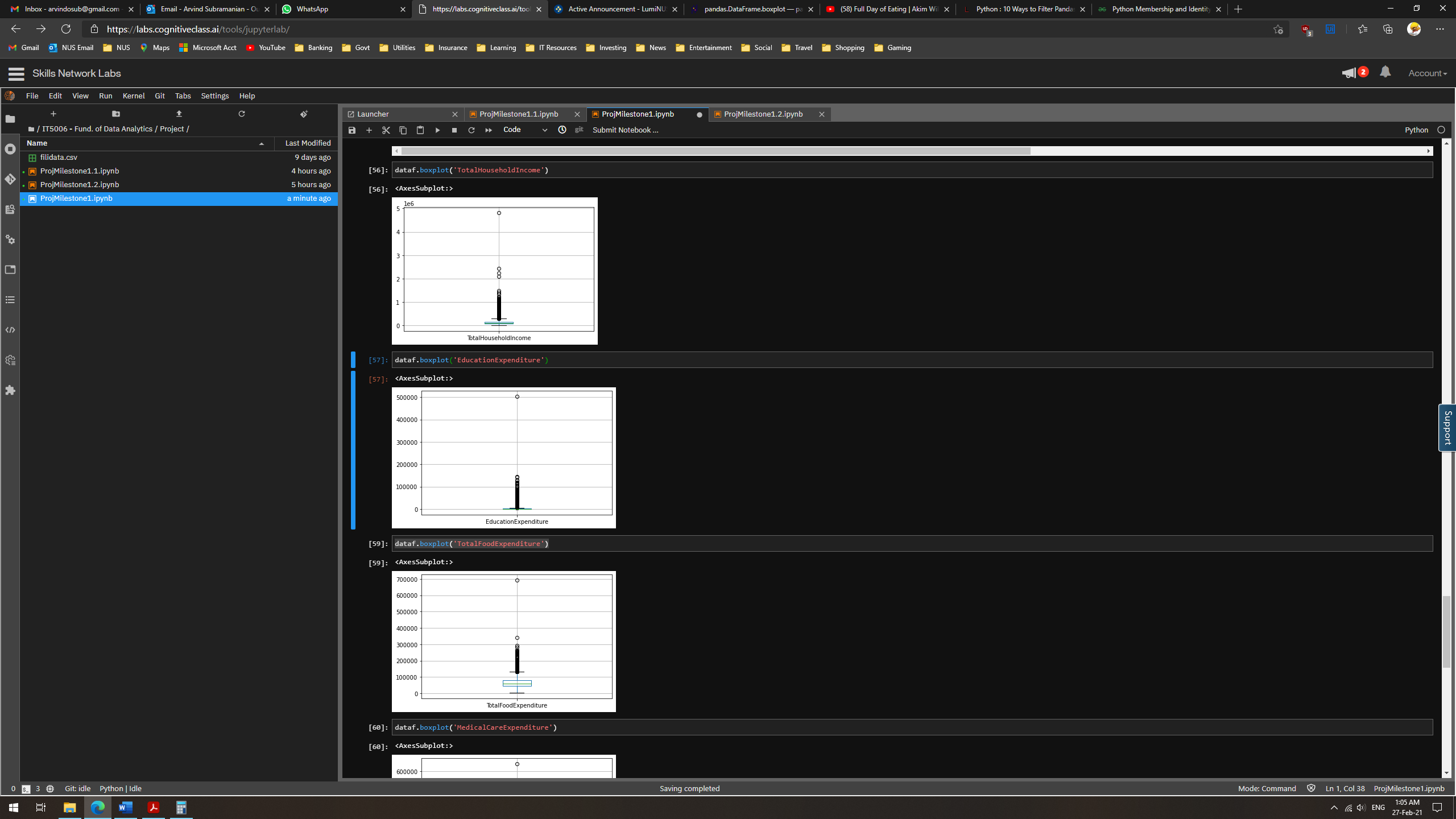
**Outliers**

As explained previously, we need to analyse our data in distinct groups. Only then can we obtain meaningful outliers, as we expect data within each group to be comparable. For example, if we were to compare Total Household Income across all Occupations, the Lawyers’ incomes would all seem like outliers next to that of the Farmers.

Therefore, we performed some preliminary segregation of the occupations based on the trends observed in our visualizations. Below, we present some outliers observed when analysing data for the ‘Farmer/Farmhand’ occupation group. These outliers were found by plotting boxplots of certain features of interest in the data subset.

Fig 10: Boxplots of Variables for the ‘Farmers’ Data Subset





As seen in these boxplots, there are some outliers in each of selected features. We plan to remove them from the data subsets before running the data through the machine learning models. This is because the outliers are significantly out of the interquartile range and the maximum point as well. They would significantly skew the results and predictions.

**Conclusion**

The preliminary analysis of the data has offered many angles to approach this complicated issue. In addition to maximizing income, we could also explore minimizing unwanted expenditure. We feel that our approach of tackling the issue on two levels (Group and National) would yield a holistic outcome. With this strategy, we hope to deliver both short-term and long-term solutions to minimize income inequality.