Individual Final Report

Abhishek Nimmakayala

**Background Information:**

The portion of work I did was the exploratory data analysis , the data pre-processing and a decision tree model with raw data.Exploratory data analysis was required to get valuable insights from the data. Also I had to pre-process the data as our data had lots of missing values which needed to be dealt with before fitting it to a decision tree model.

**Brief explanation of my work:**

**Exploratory data analysis:**

Created visualizations to gain insights of the data which includes the below:

1. Scatter plot of family income in poverty VS Scatter plot of family income in non-poverty
2. Bar graph to demonstrate the percentage of population in poverty
3. Poverty population Vs non poverty population in different towns of New York
4. Bar graph illustrating missing data in columns

**Data pre-processing:**

The pre-processing I did was to **handle missing data** in columns and **feature selection**. I used the f-fill (forward fill) method to impute missing data in each column. The f-fill imputes the value of the observation above to the value of the missing data. As the data in our dataset was inputted sequentially the f-fill is the best method to replace missing values in our case.

For feature selection I dropped columns which were highly correlated with each other (for example

NYCgov\_FICAtax, NYCgov\_IncomeTax, NYCgov\_Income were all highly correlated. etc). Also dropped unwanted columns which would have no effect on the model (Example: Household Unit ID, Age Category, Poverty Gap, Tax Unit etc)

**Decision Tree using raw Data:**

Ran a decision tree model with un-processed data. Used sklearn to split my data into train and test and defined my x\_train, y\_train, x\_test, y\_test and fit to a decision tree model. Used confusion\_matrix from sklearn.metrics to view the accuracy of the model on raw data.