**CS253 – Assignment 3 Report**

**Chinmay Pillai**

**200298**

**Feature Engineering:**

A new feature called net\_worth = Total Assets – Liabilities has been generated. This better represents the individual’s financial capacity.

**Features Used:**

Since ID, Candidate Name and Constituency are unique features, they won’t be useful in training and hence are dropped. Additionally, we don’t expect to have any correlations between the party an individual belongs to and their education level, hence this feature is also dropped. The below plots also reflect the same. The feature net\_worth is used instead of Total Assets and Liabilities since it better represents the individual’s financial capacity.

The catagorical feature state and been one-hot encoded before being used for training.

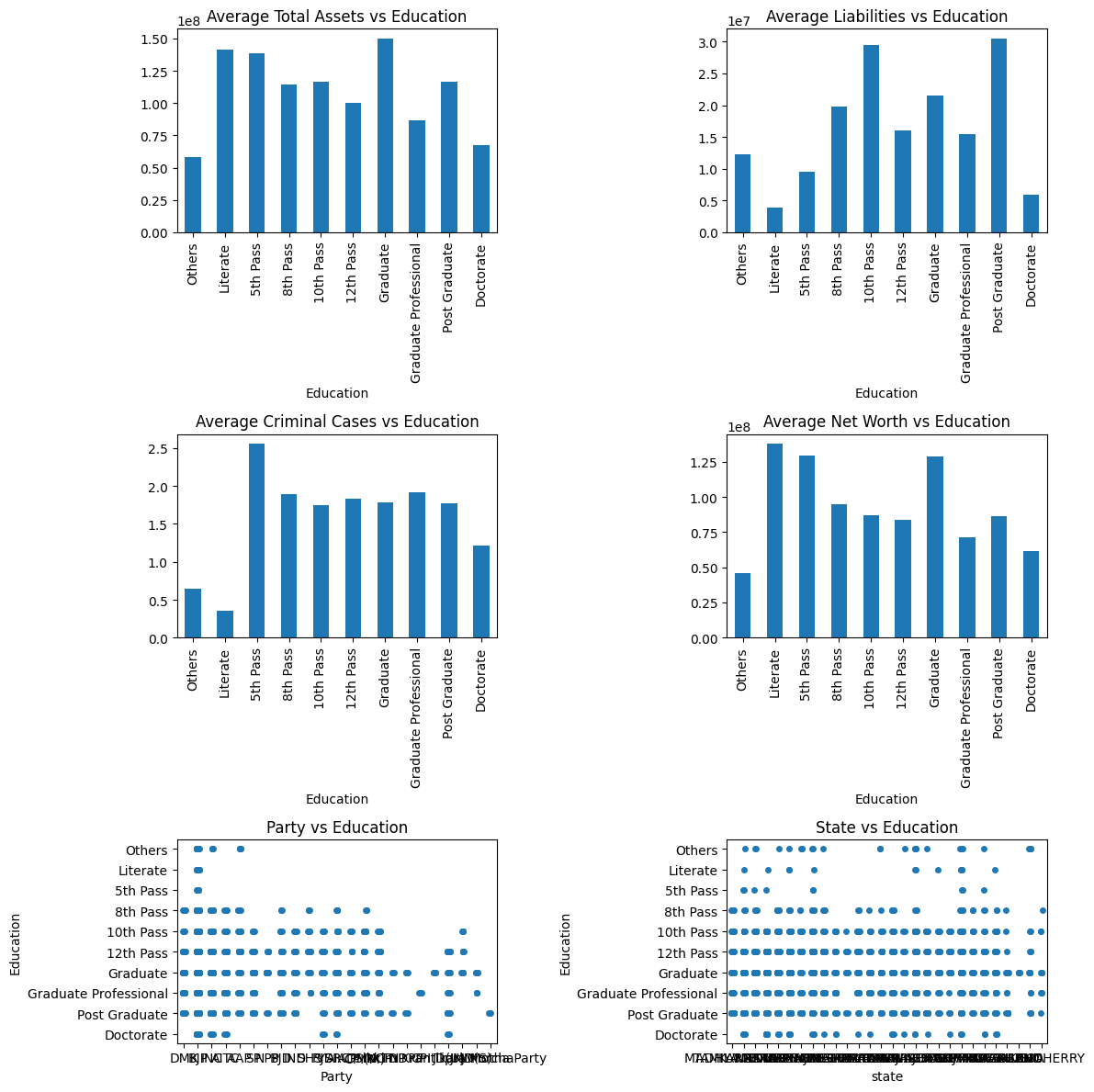
**Numerical Features Used**: net\_worth, Criminal Cases

**Catagorical Features Used**: State

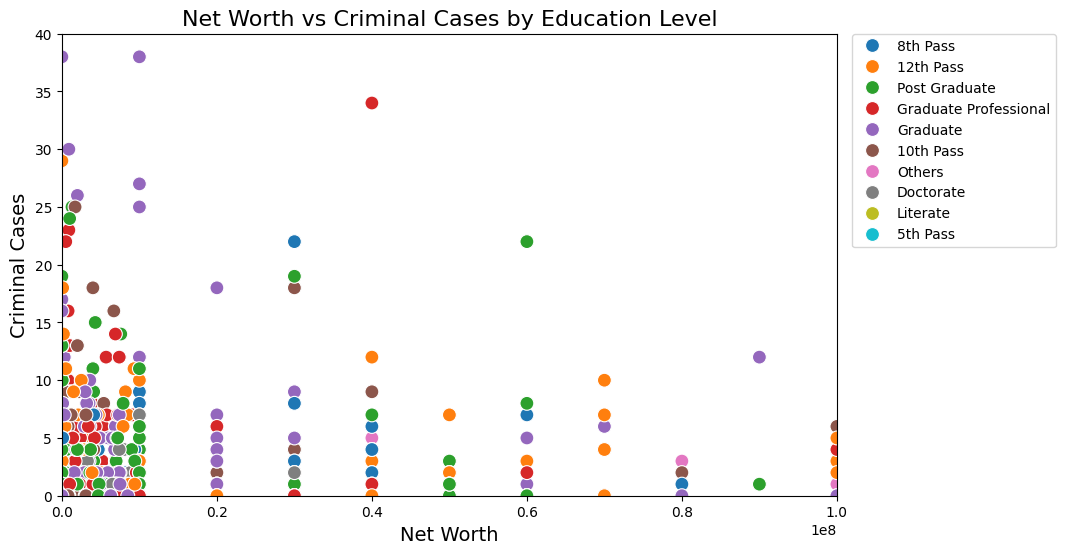
**Target Variable:**

Initially one-hot encoding was used on the catagorical target ‘Education’ as well. But, since different classes of education do lie on an scale, using Labe Encoding where the lower levels of education are given a lower integer value and the higher levels are given a higher value, better represents the target feature. The classes have been orders as - 'Others', 'Literate', '5th Pass', '8th Pass', '10th Pass', '12th Pass', 'Graduate', 'Graduate Professional', 'Post Graduate', 'Doctorate', where ‘Others’ maps to a label of 0, ‘Literate’ to 1 and so till ‘Doctorate’ to ‘9’.

**Plots of each feature vs Education Level:**



**Scatter Plot of net\_worth and Criminal Cases:**



We do not observe any clear cluster formation for each Education class and hence clustering methods like Gaussian Mixture Models are not being considered to model the data.