Python Assignment Report

Kushagra Srivastava
Department of Computer Science and Engineering
Indian Institute of Technology, Kanpur

April 2024

1 Methodology

The files for this project are contained in the following repository.

1.1 Data Pre-Processing

- The `Candidate`, `Constituency` columns are dropped from the dataframe after extracting the features.
- Since, the `Total Assets` and `Liabilities` have text-numeric data, that is entires of the form Crore or Lacs, they have been converted to their respective values in Hundreds.
- `Party`, `State` and `Education` have both been converted into their respective mapped integer values, using a dictionary.

1.2 Feature Engineering

- From the `Candidate` column, look out for 'Dr.' or 'Adv.' and create two new indicator features, `isDoctor` and `isAdvocate`.
- From the `Constituency` column, look out for '(SC)' or '(ST)' and create two new indicator features, `isSC` and `isST`. This feature was later dropped as it did not yeild any fruit-ful results, which indicates that there is little to no co-relation to these two factors.

1.3 Data-Set Analysis

We first analyse the correlation matrix as shown in 1 on the processed data, and observe. The conclusion that we arrive at from this matrix, is that Education does not have a strong correlation with a single factor over the entire dataset.

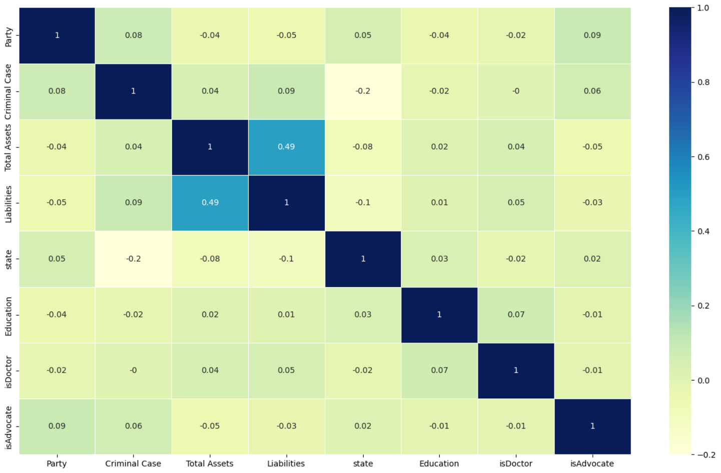

Figure 1: Correlation matrix b/w features in the dataset

Now, we take a look at the correlation over different state 2 and education now is strongly co-related to some factor which is different over different states.

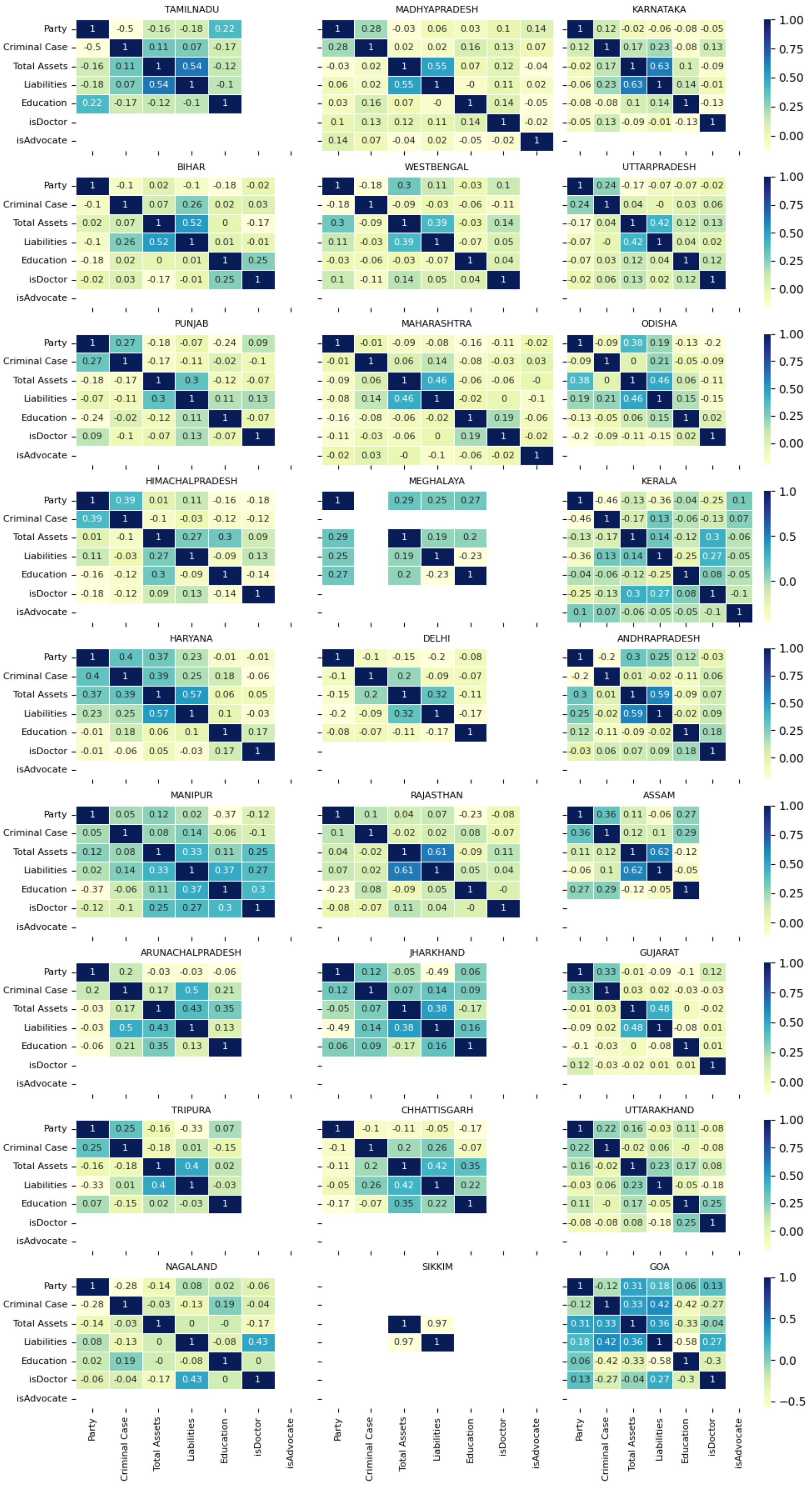

Figure 2: Corealtion matrix b/w features in the dataset

2 Experiment Details

The main idea here, is to use different model for different states, so, we try two models, `kNN`, `Decision Tree`, and choose the one with the better F1-Score, for each state.

Table 1: kNN for each state.

State	n	F1-Score
TAMILNADU	14	0.4432
MADHYAPRADESH	8	0.6375
KARNATAKA	12	0.4464
BIHAR	7	0.3185
WESTBENGAL	8	0.4071
UTTARPRADESH	14	0.3271
PUNJAB	14	0.2014
MAHARASHTRA	12	0.3146
ODISHA	6	0.3436
HIMACHALPRADESH	2	0.5091
MEGHALAYA	14	0.2815
KERALA	7	0.3367
HARYANA	7	0.1111
DELHI	13	0.3636
ANDHRAPRADESH	13	0.1400
MANIPUR	1	0.2667
RAJASTHAN	2	0.2768
ASSAM	6	0.2476
ARUNACHALPRADESH	12	0.2500
JHARKHAND	13	0.3807
GUJARAT	10	0.1905
TRIPURA	2	0.5714
CHHATTISGARH	12	0.5101
UTTARAKHAND	14	0.2545
NAGALAND	5	0.2204
SIKKIM	1	1.0000
GOA	14	0.3333
PUDUCHERRY	1	0.3750

These parameters have been saved in their respective files, to avoid training time, as `paramters_DT.txt` or `paramters_KNN.txt`

Using these parameters we can judge, which kind of model is to be used for each state, and a final model is formed.

Table 2: Decision Tree for each state.

State	maxdepth	minsamplesleaf	minsamplessplit	F1-Score
TAMILNADU	10	4	2	0.5156
MADHYAPRADESH	1	1	2	0.6429
KARNATAKA	6	1	3	0.2924
BIHAR	1	1	2	0.2227
WESTBENGAL	25	1	3	0.2497
UTTARPRADESH	9	4	2	0.3523
PUNJAB	6	8	2	0.3100
MAHARASHTRA	4	10	2	0.2399
ODISHA	7	14	2	0.1604
HIMACHALPRADESH	6	11	2	0.5455
MEGHALAYA	4	13	2	0.2815
KERALA	7	2	2	0.4325
HARYANA	6	3	2	0.1905
DELHI	4	11	2	0.3636
ANDHRAPRADESH	11	9	2	0.3253
MANIPUR	6	2	2	0.1111
RAJASTHAN	4	8	2	0.2075
ASSAM	1	1	2	0.2286
ARUNACHALPRADESH	4	8	2	0.2000
JHARKHAND	2	1	2	0.5273
GUJARAT	3	1	2	0.2630
TRIPURA	4	10	2	0.6349
CHHATTISGARH	7	2	2	0.4072
UTTARAKHAND	1	1	2	0.1169
NAGALAND	2	1	2	0.2204
SIKKIM	1	1	2	1.0000
GOA	1	1	2	0.3333
PUDUCHERRY	4	1	2	0.3750

2.1 Dataset Analysis for Given Train Data

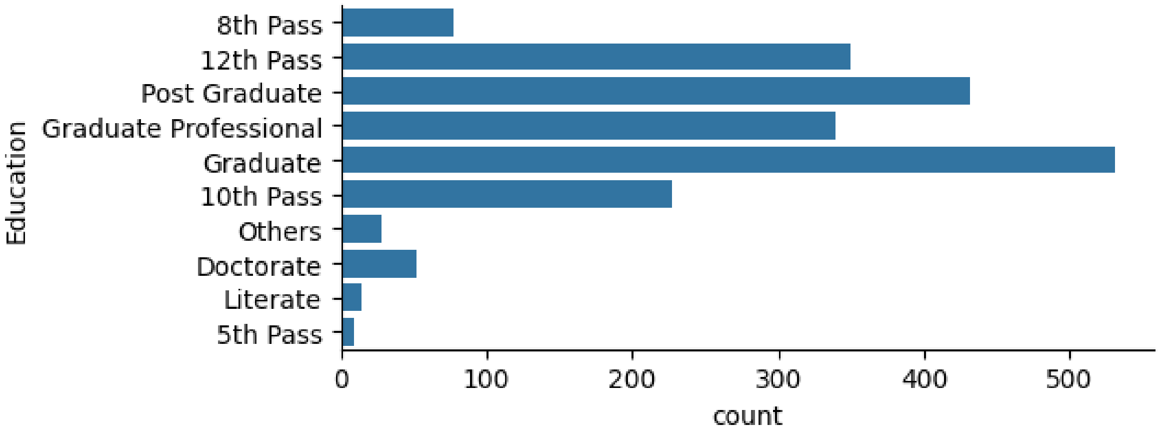

Figure 3: Distribution of Education

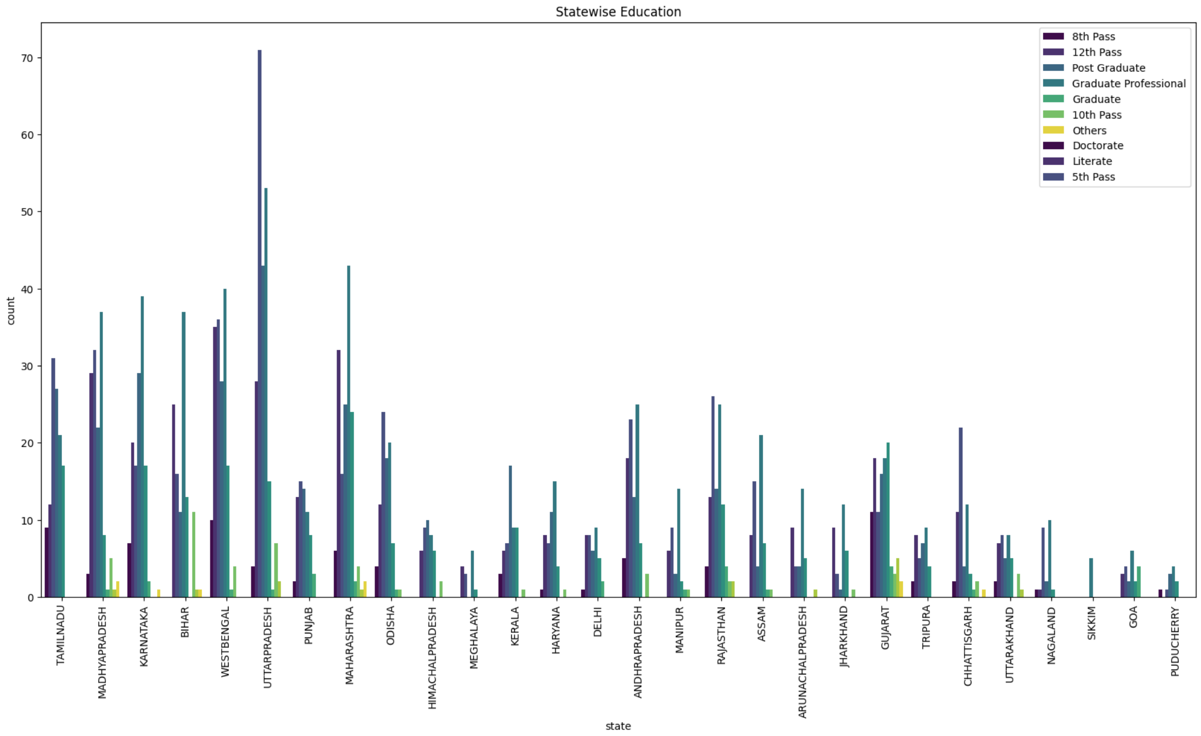

Figure 4: Distribution of Education with respect to state

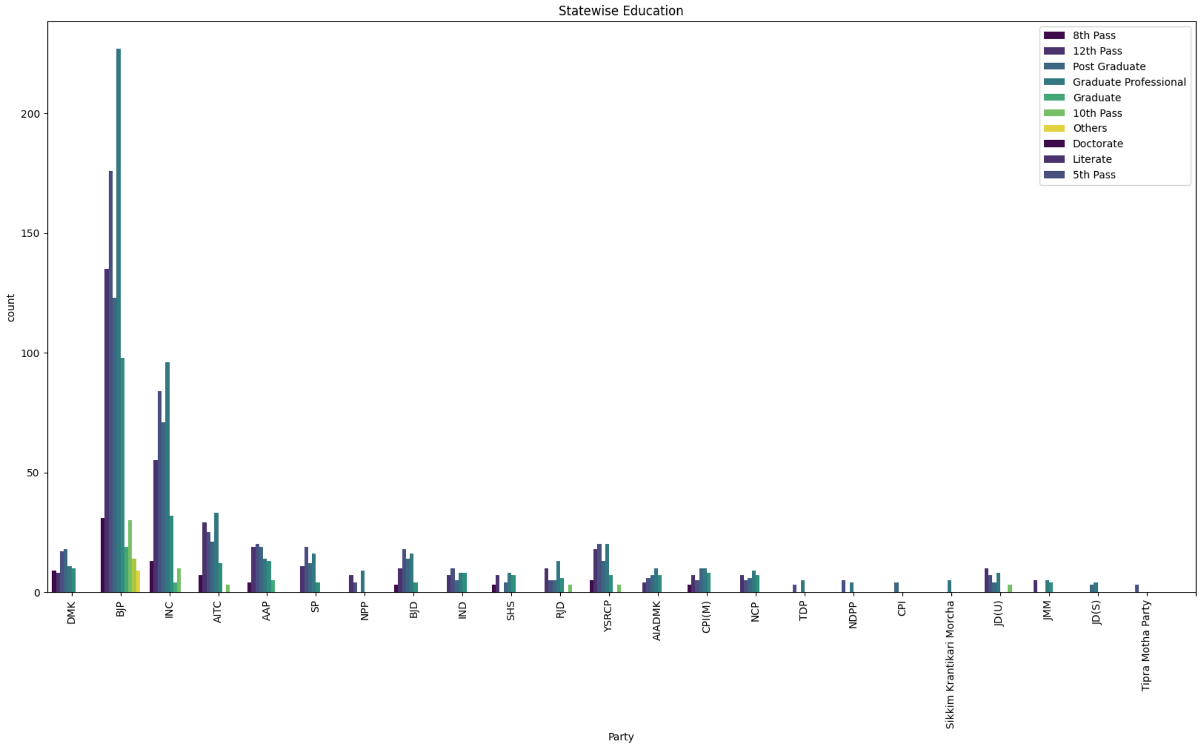

Figure 5: Distribution of Education with respect to Party

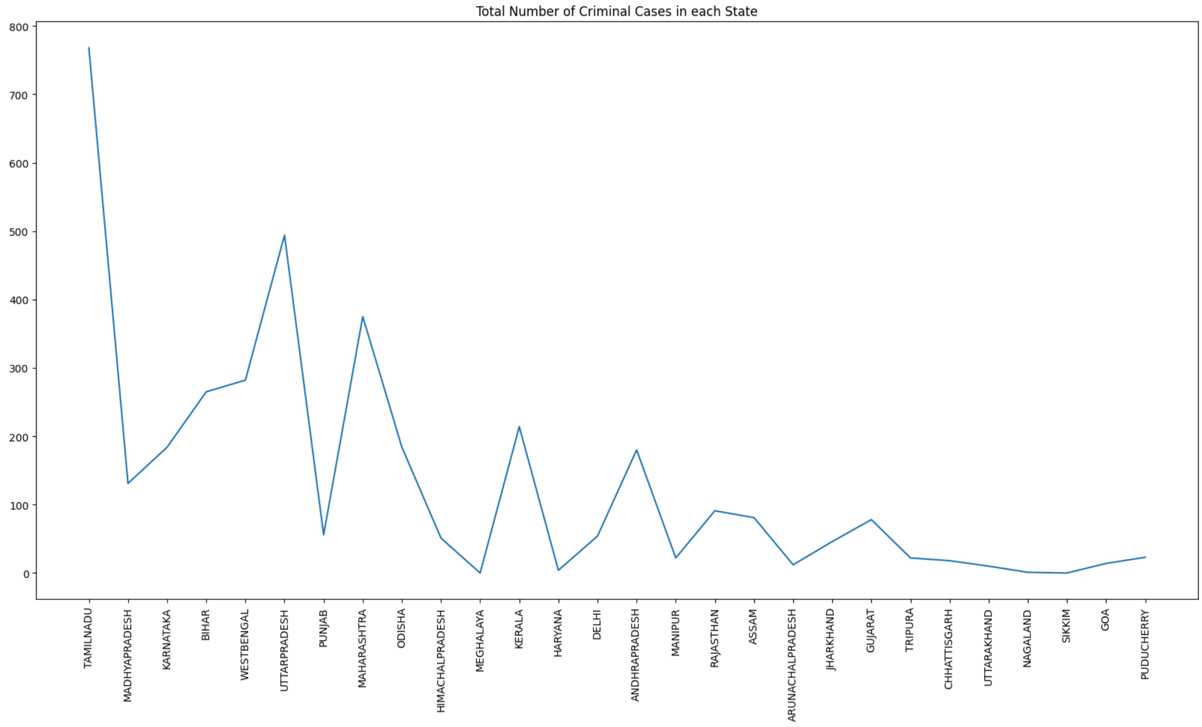

Figure 6: Criminal Cases vs State

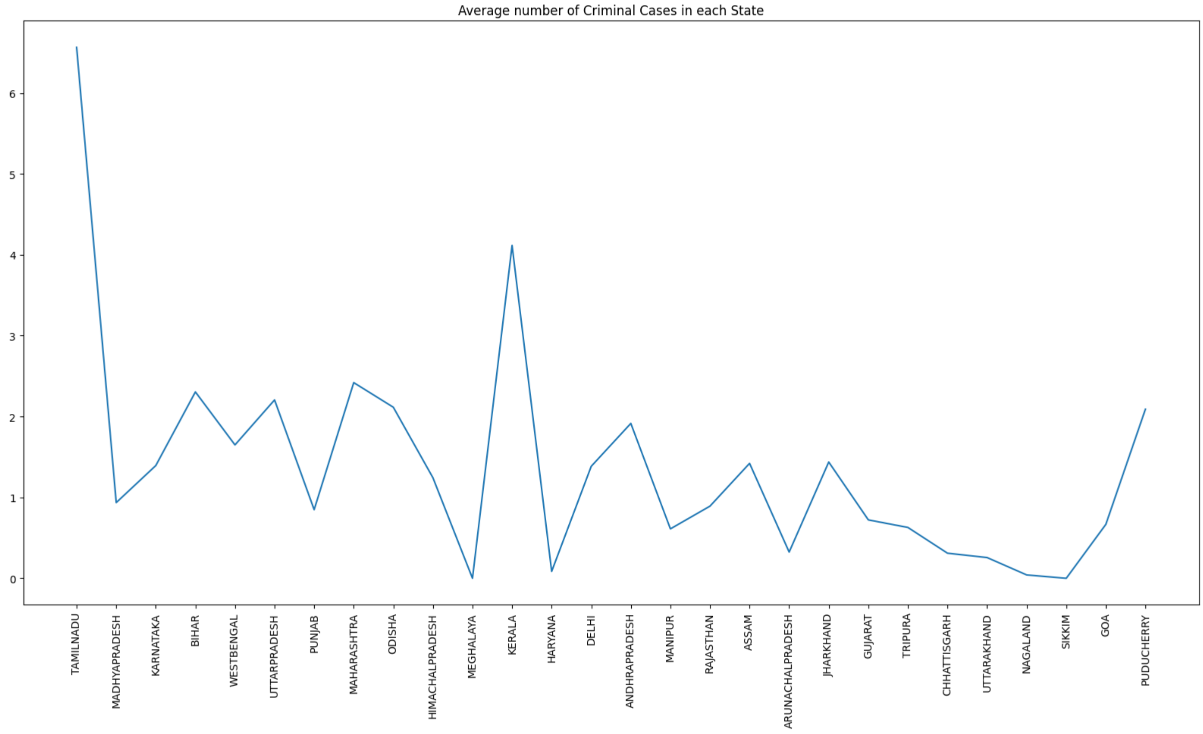

Figure 7: Average Criminal Cases vs State

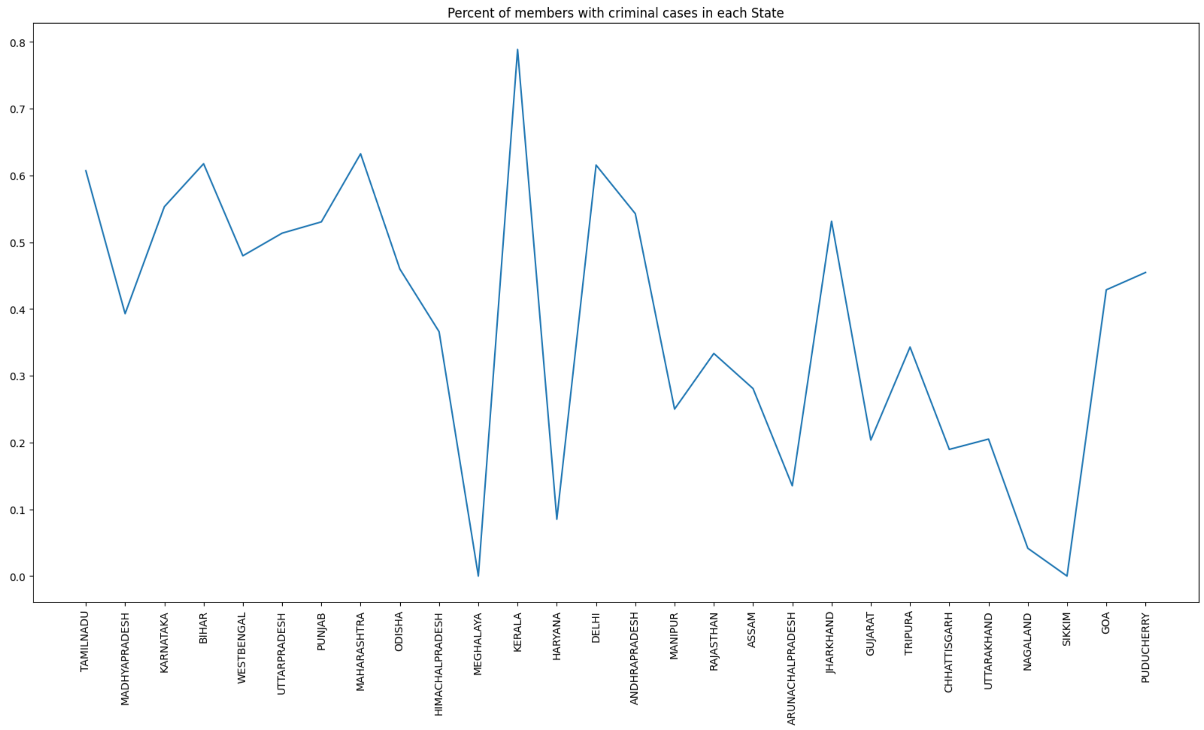

Figure 8: Percent of Assembly with Criminal Cases vs State

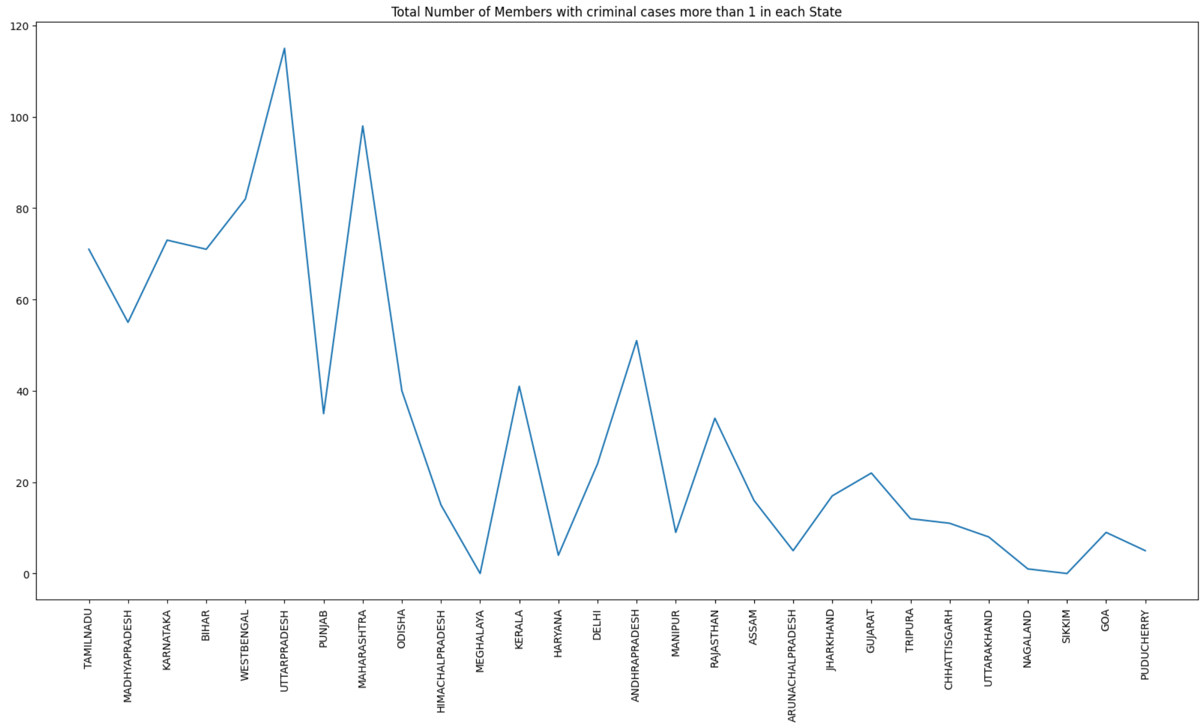

Figure 9: Members with more than one criminal cases vs State

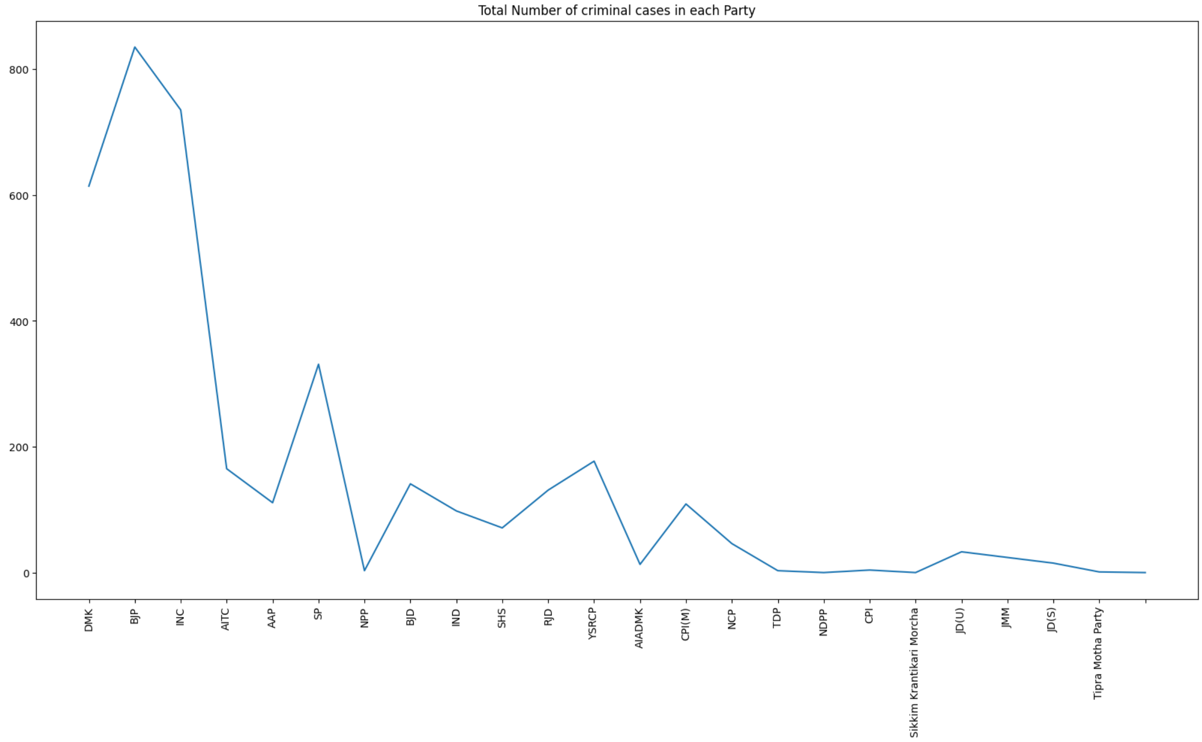

Figure 10: Total Criminal Cases vs Party

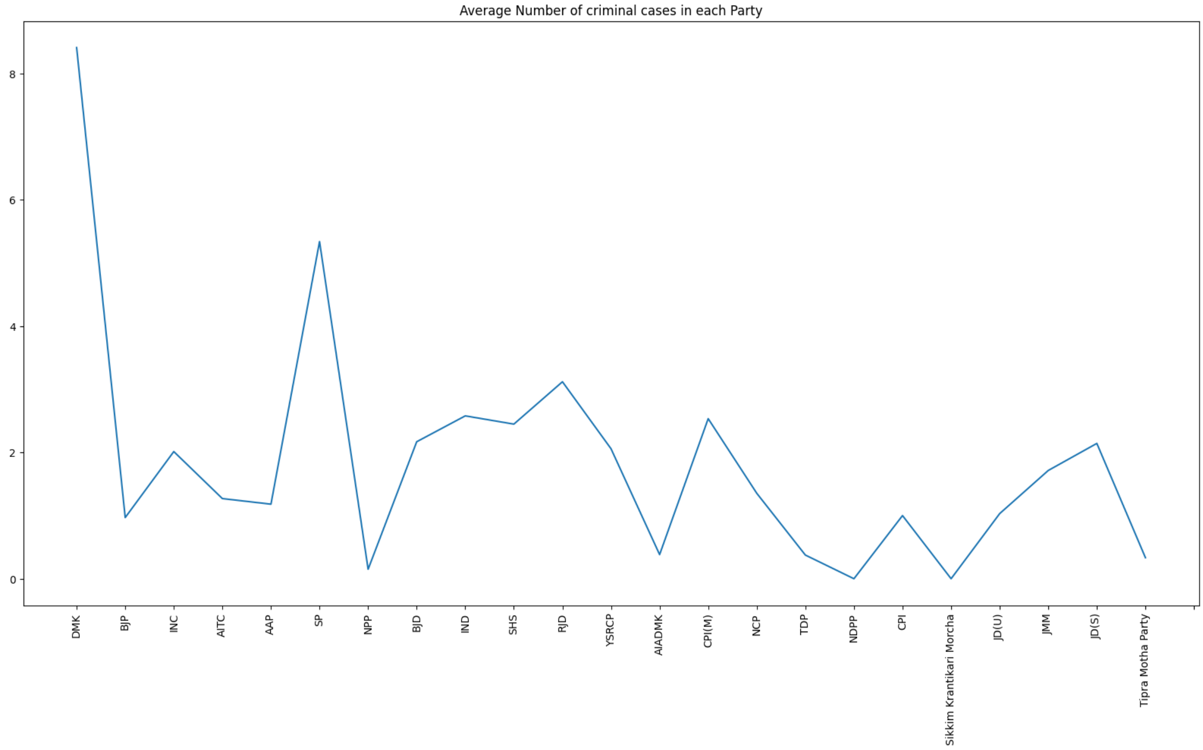

Figure 11: Average Criminal Cases vs Party

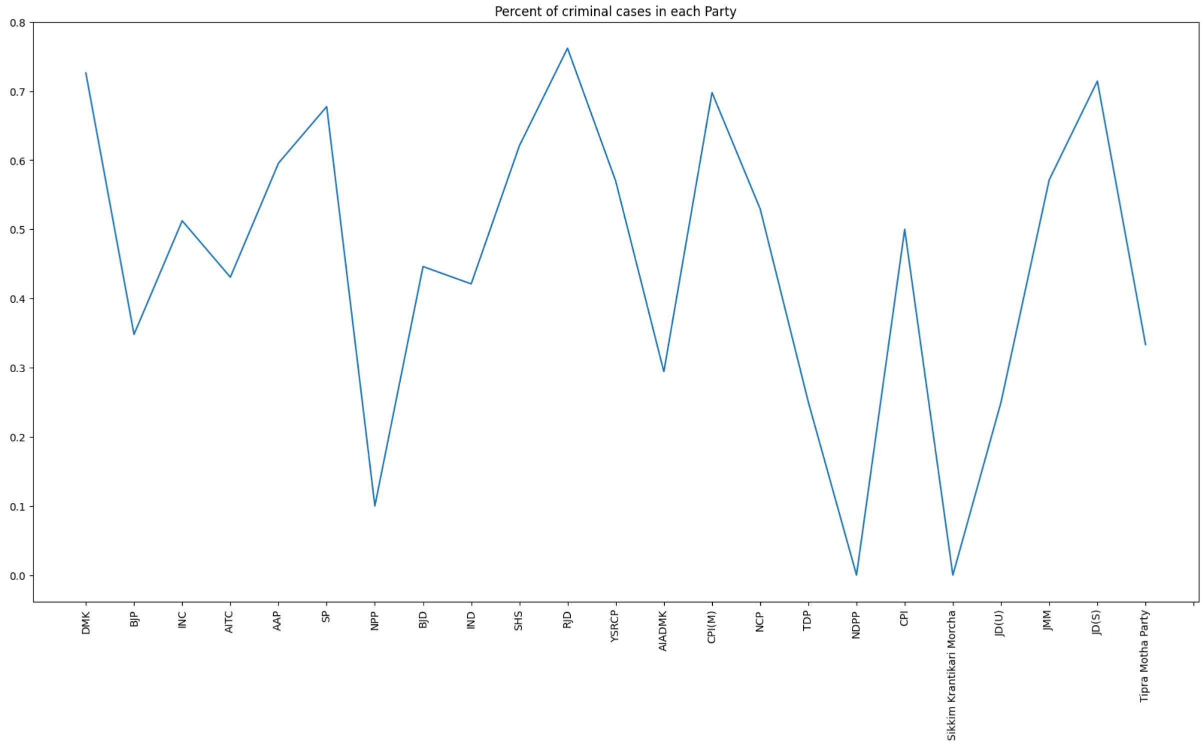

Figure 12: Percent of Party with Criminal Cases vs Party

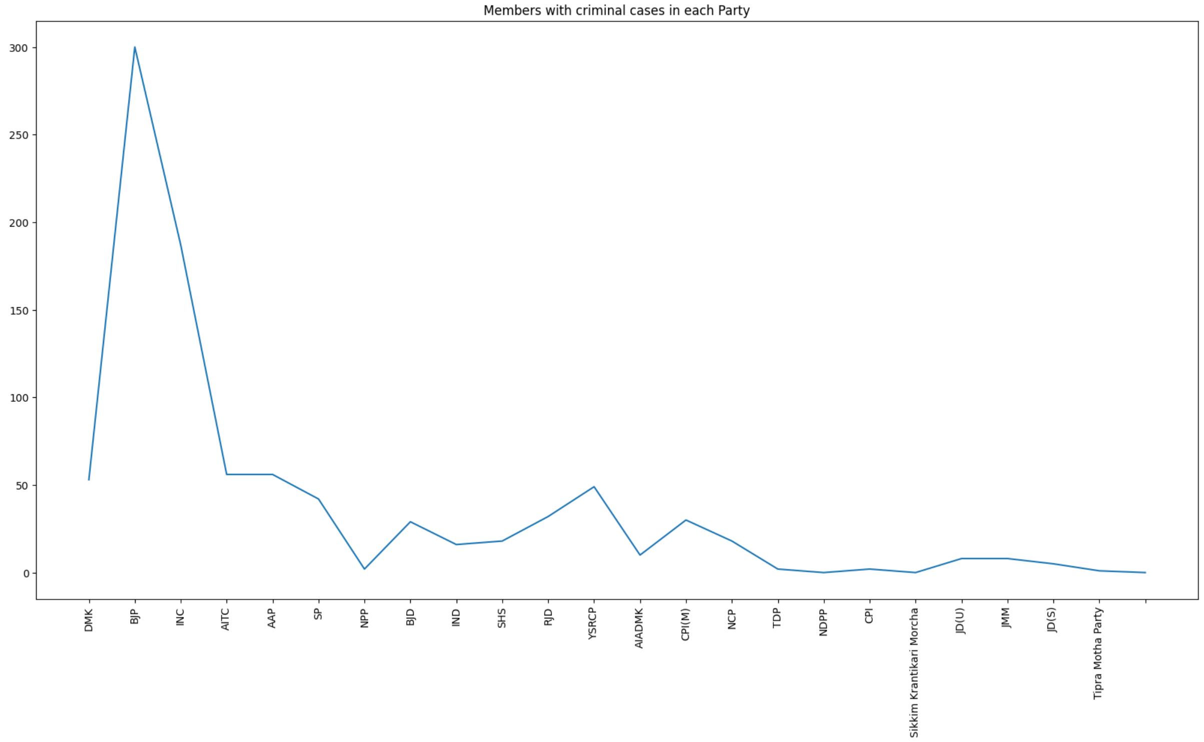

Figure 13: Members with more than one criminal cases vs Party

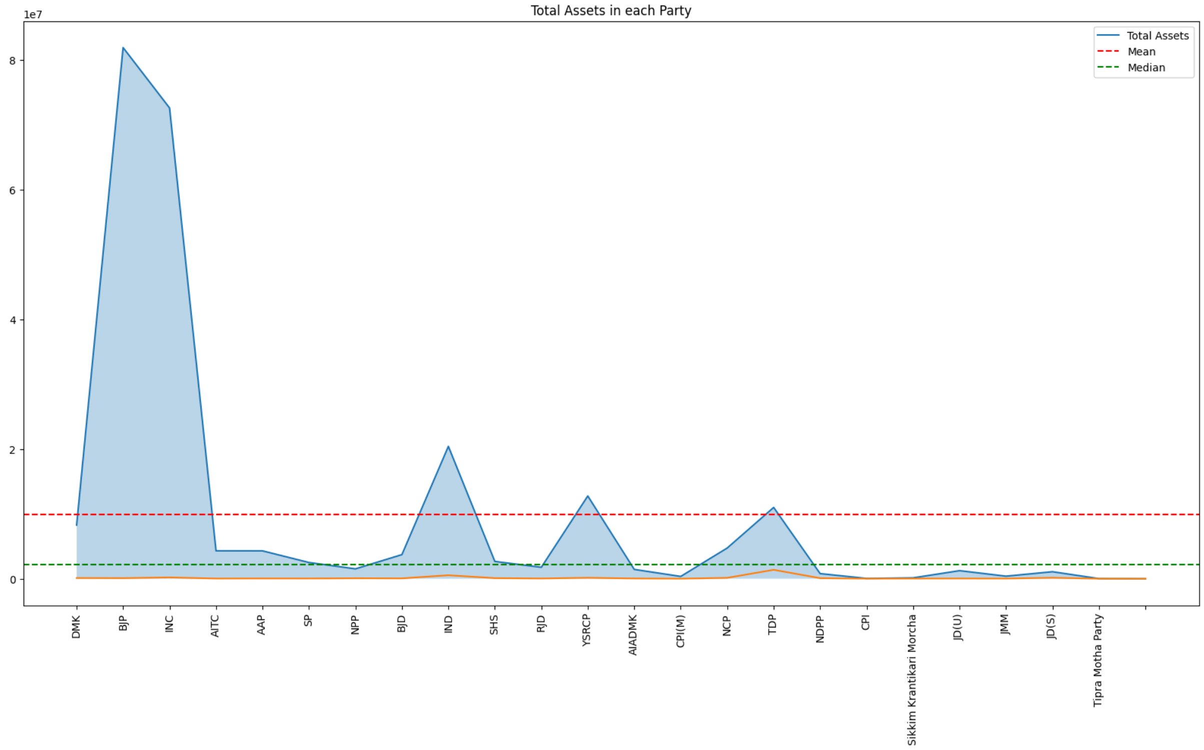

Figure 14: Assets for each party

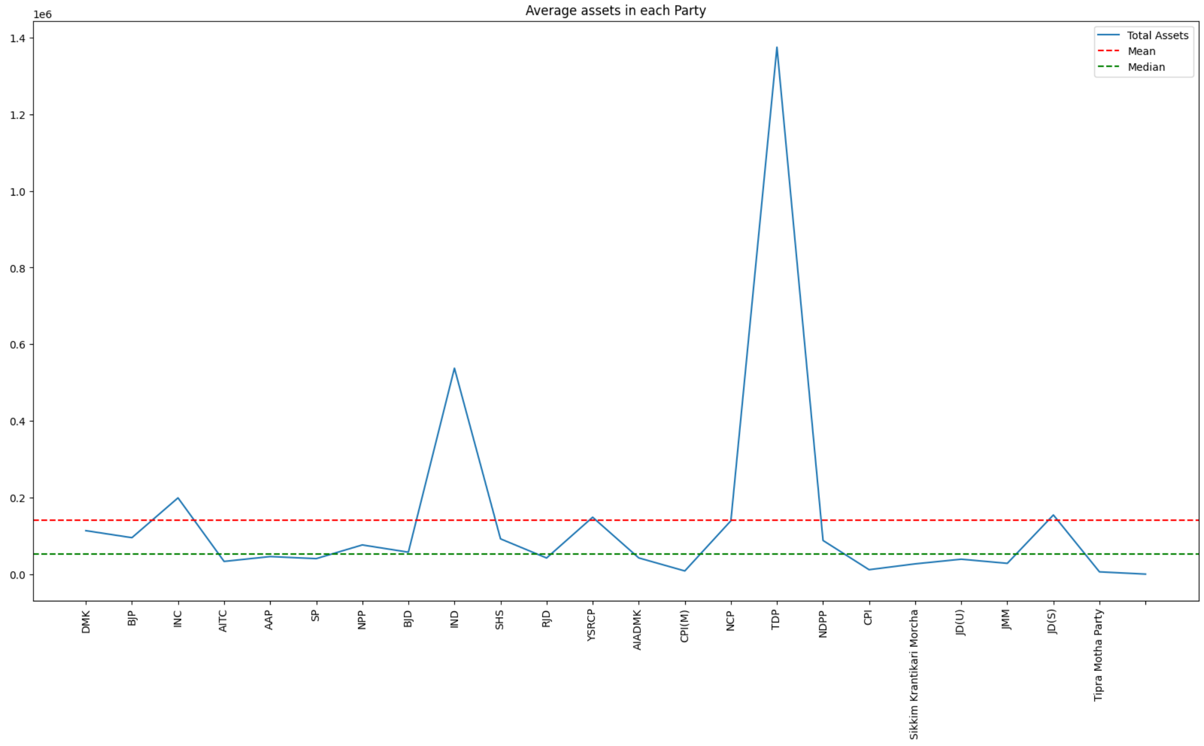

Figure 15: Average Assets in each party

2.2 Dataset Analysis for Enhanced Data

Data has been enhanced using `CTGAN`, by generating around 8000 Datapoints, and using 10000 epochs.

```python
ctgan = CTGAN(epochs=10000, log_frequency=True, verbose=True)
ctgan.fit(X, discrete_columns)

# Create synthetic data
synthetic_data = ctgan.sample(8000)
final_data = []

# Save the synthetic data
reverse_party = {value: key for key, value in party.items()}
reverse_edu = {value: key for key, value in education.items()}
reverse_state = {value: key for key, value in state.items()}

for row in synthetic_data.index:
    synthetic_data.loc[row, "Party"] = reverse_party[synthetic_data.loc[row, "Party"]]
    synthetic_data.loc[row, "state"] = reverse_state[synthetic_data.loc[row, "state"]]
    synthetic_data.loc[row, "Education"] = reverse_edu[synthetic_data.loc[row, "Education"]]

synthetic_data.to_csv('synthetic_data_5ke_8k.csv', index=False)
```

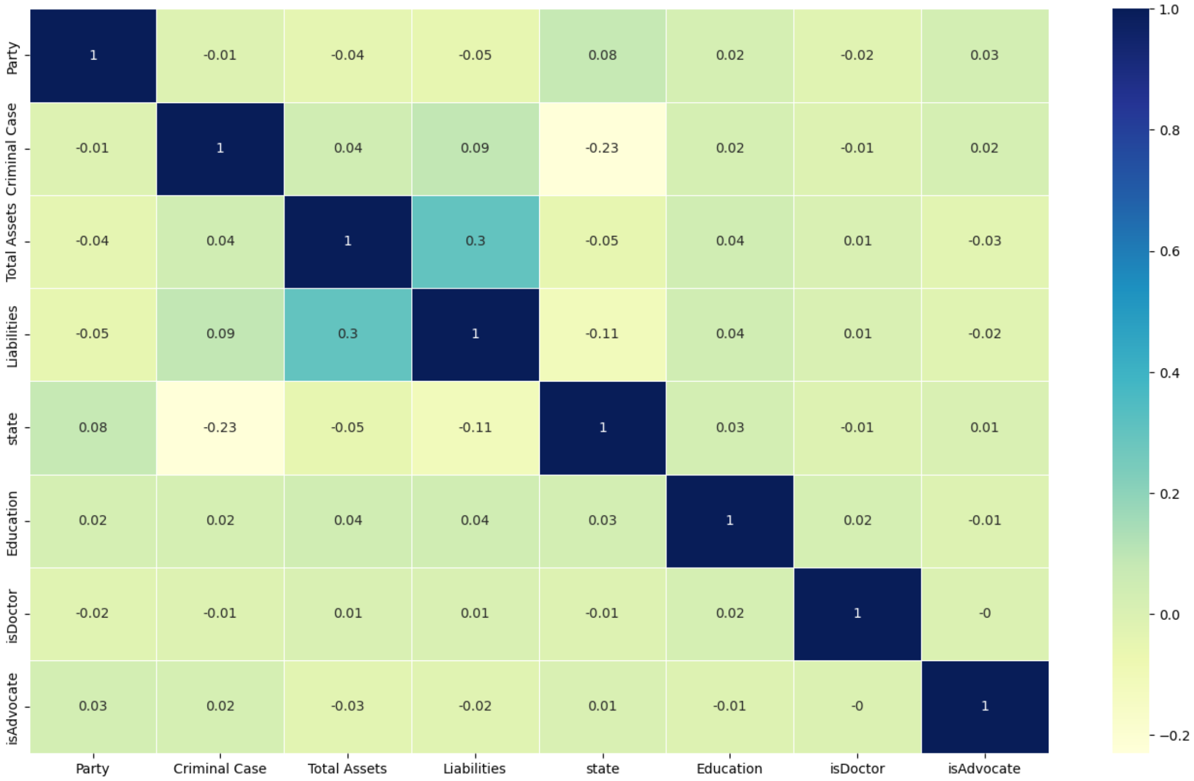

Figure 16: Co-relation

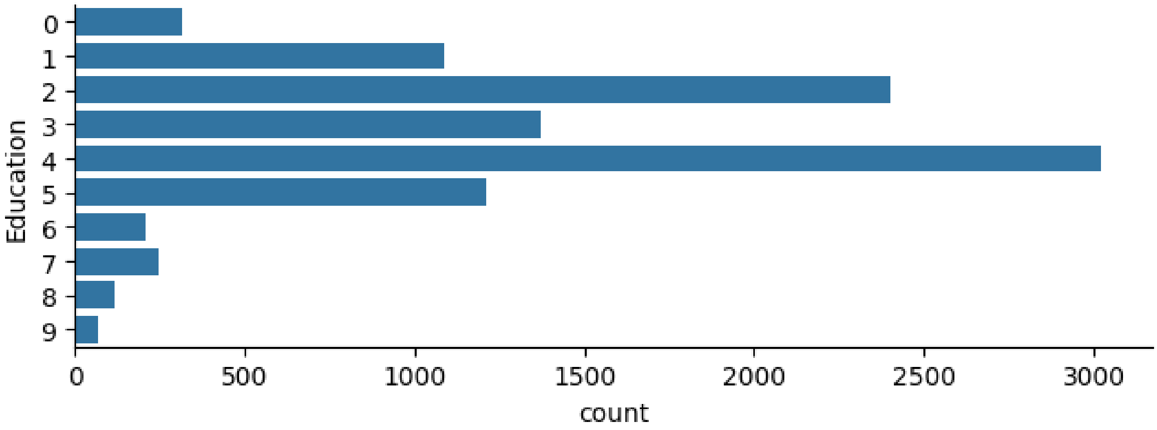

Figure 17: Education Distribution

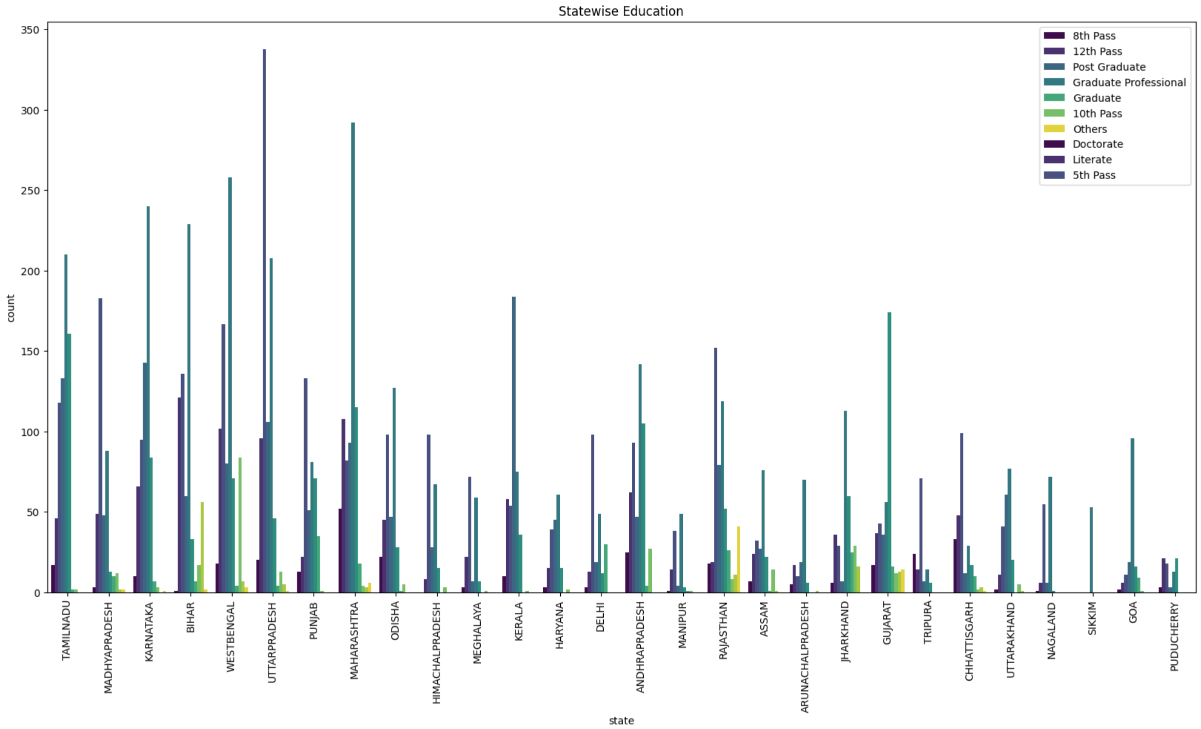

Figure 18: Education vs State

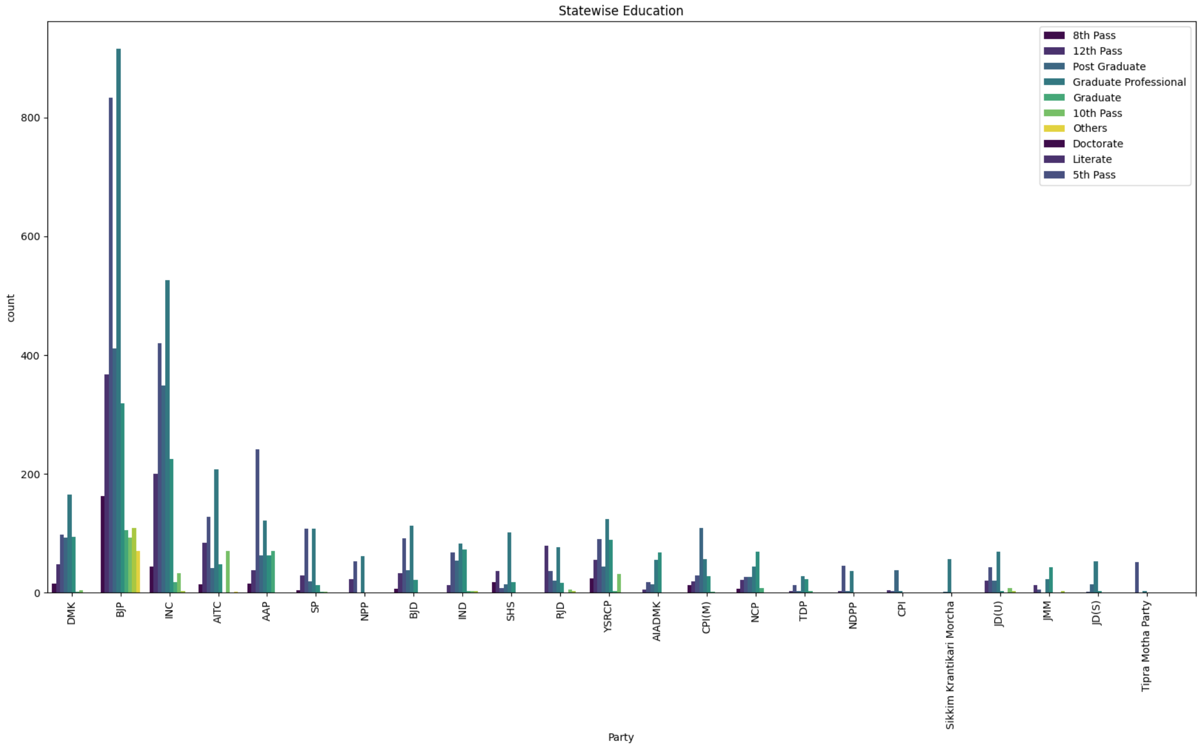

Figure 19: Education vs Party

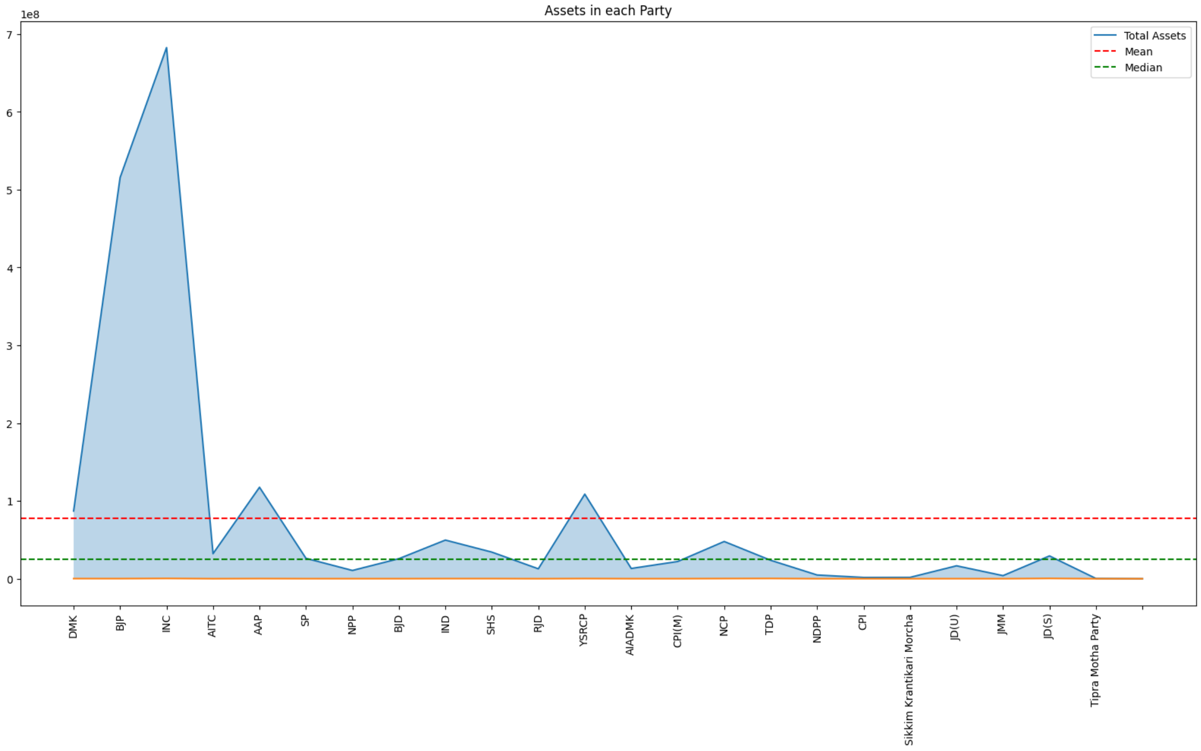

Figure 20: Assets vs Party

3 Results

The results generated using this methodology, generate the following Scores :

- `Public Score : 0.30737` [7/243]
- `Private Score : 0.32192` [6/243]

4 References

- Xu, Lei, et al. "Modeling Tabular data using Conditional GAN - NeurIPS Proceedings." [SNIPPET] Modeling Tabular data using Conditional GAN. (2019).
- Multi-Class Classification on Medium