Assignment-4 DMS-672

Rohan Nimesh

Dataset Summary:

Shape: Rows-33212, Columns-17

Data types:

```
age                int64
workclass         object
fnlwgt             int64
education         object
education-num    float64
marital-status    object
occupation        object
relationship      object
race              object
sex               object
capital-gain       int64
capital-loss       int64
hours-per-week     int64
native-country    object
income            object
random_string     object
hashed_id         object
```

Missing Values:

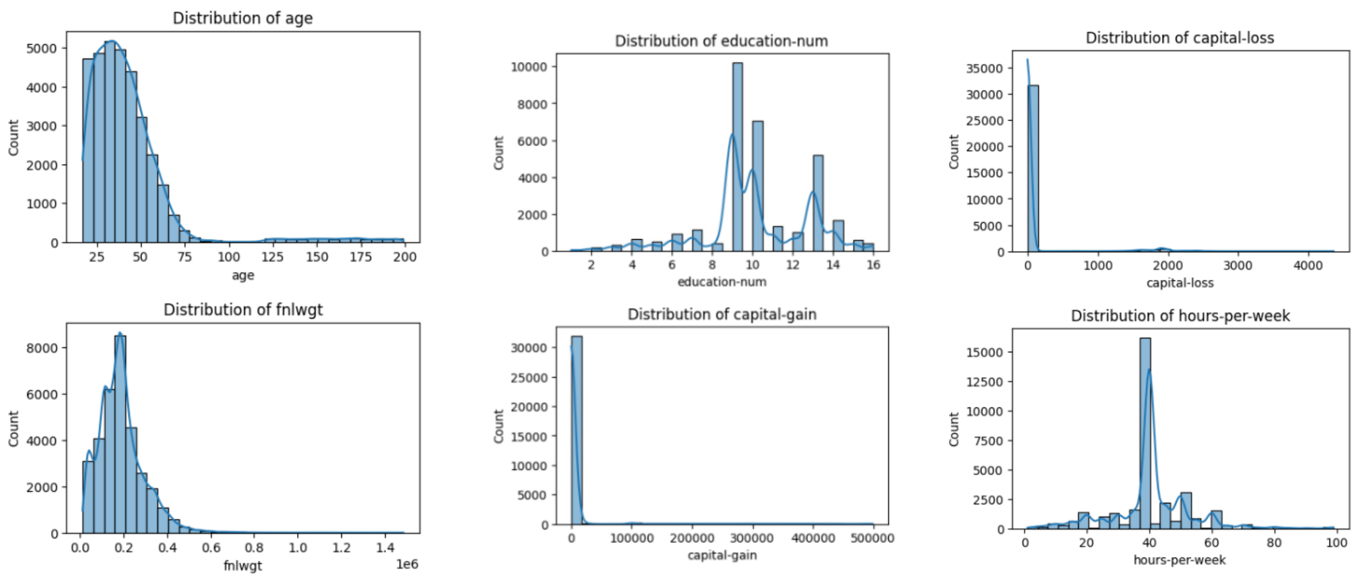
```
occupation        3910
workclass         1873
education-num     1638
native-country     591
age                  0
education            0
marital-status       0
relationship         0
fnlwgt               0
race                 0
sex                  0
capital-loss         0
capital-gain         0
hours-per-week       0
income               0
random_string        0
hashed_id            0
```

Numeric Summary

```
                  count           mean            std      min       25%  \
age             33212.0      42.215103      24.941444     17.0      28.0
fnlwgt          33212.0  189883.462815  105560.543548  12285.0  117849.0
education-num   31574.0      10.081903       2.576571      1.0       9.0
capital-gain    33212.0   10007.742262   54898.139383      0.0       0.0
capital-loss    33212.0      87.296911     402.795736      0.0       0.0
hours-per-week  33212.0      40.439269      12.341622      1.0      40.0

                     50%       75%        max
age                 38.0      49.0      199.0
fnlwgt          178430.0  237397.5  1484705.0
education-num       10.0      12.0       16.0
capital-gain         0.0       0.0   499096.0
capital-loss         0.0       0.0     4356.0
hours-per-week      40.0      45.0       99.0
```

Distribution of Numerical Column:



EDA:

Firstly, I looked for the number of duplicate rows in the dataset that came out to be 651. Also, in the dataset I looked at columns that had each entry as a unique entry an might not have any significant role in the modelling process those columns were 'random_string' and 'hashed_id'.

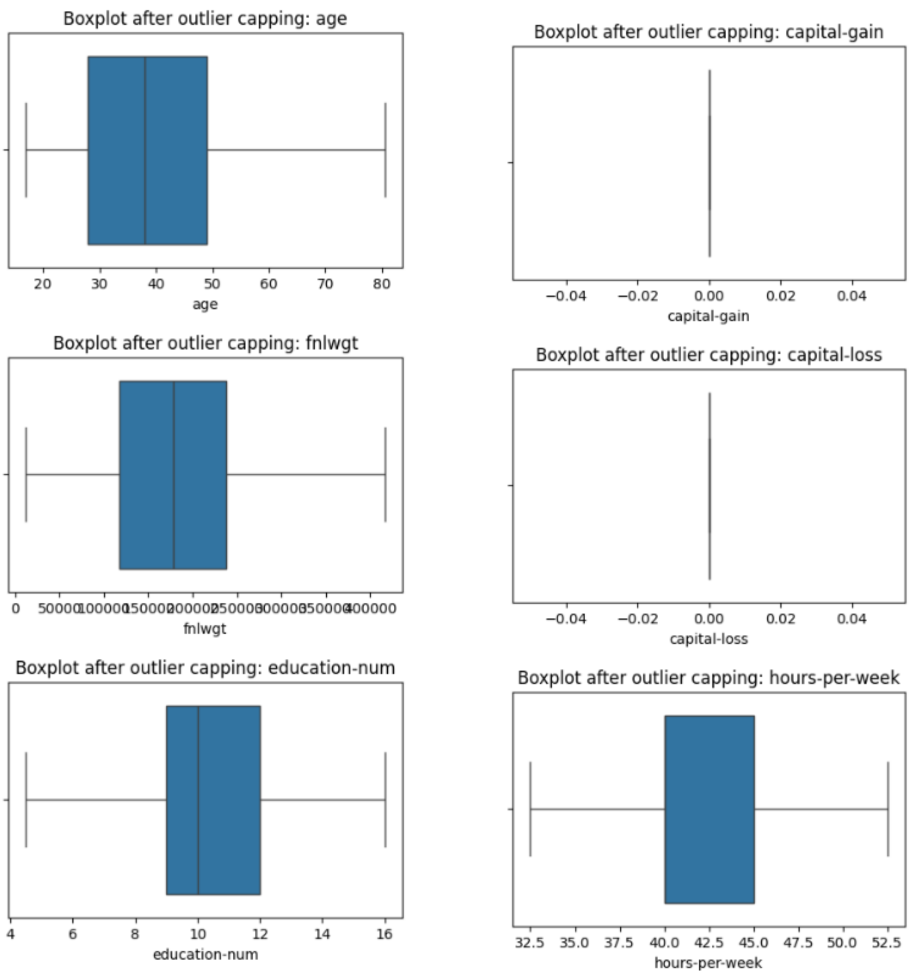
Next, I looked for any suspicious values in the dataset, here the age column had 995 suspicious entry (the age can't be greater than 120).

Next step was to handle the missing data for that we just updated the nan values with the median of the column for numerical values and for categorical values we updated the nan values with the most frequent observations.

Next step was to handle the outliers in the data set for that I first defined the potential outlier values for each numerical column. The potential outlier values obtained after performing this are shown below

```
Potential outliers per numeric column:
  age: 1088 outliers (range = [-3.50, 80.50])
  fnlwgt: 1000 outliers (range = [-61473.75, 416720.25])
  education-num: 1166 outliers (range = [4.50, 16.50])
  capital-gain: 3655 outliers (range = [0.00, 0.00])
  capital-loss: 1550 outliers (range = [0.00, 0.00])
  hours-per-week: 9206 outliers (range = [32.50, 52.50])
```

The boxplot for these numerical columns after outlier handling are shown below



From this we can clearly see that there are no more outliers in the dataset as we have capped the outlier values.

Next step was to handle the formatting inconsistencies in the dataset for that we converted all the categorial data in lower case. Some possible formatting issues are shown below

```
Possible formatting issues in 'workclass': ['Private', 'Private', 'Private', 'Private', 'Private', 'State-gov', 'Private', ' Private ', 'Self-emp-not-inc', 'Private']
Possible formatting issues in 'education': ['hs-grad', 'SOME-COLLEGE', 'Some-college', '9th', 'HS-grad', 'some-college', 'HS-GRAD', 'HS-grad', 'hs-grad', 'HS-grad']
Possible formatting issues in 'marital_status': ['Never-married', 'Divorced', 'NEVER-MARRIED', 'Never-married', ' Never-married ', 'Never-married', ' Married-civ-spouse ', 'Divorced', 'Married-civ-spouse', 'Married-civ-spouse']
Possible formatting issues in 'occupation': ['Other-service', 'Adm-clerical', 'Sales', 'Priv-house-serv', 'Machine-op-inspct', ' adm-clerical ', 'Machine-op-inspct', 'Craft-repair', 'Craft-repair', 'Farming-fishing']
Possible formatting issues in 'relationship': ['Other-relative', 'Unmarried', 'Not-in-family', 'Not-in-family', 'Not-in-family', 'Not-in-family', 'Husband', 'Unmarried', ' Husband ', 'Husband']
Possible formatting issues in 'race': ['Asian-Pac-Islander', ' White ', 'White', 'White', 'White', 'Black', 'white', 'White', 'White', 'White']
Possible formatting issues in 'sex': ['male', 'Female', ' Female ', ' Female ', 'Male', 'Female', 'Male', 'Female', 'Male', 'Male']
Possible formatting issues in 'native_country': ['United-States', 'United-States', 'United-States', 'El-Salvador', ' United-States ', ' United-States ', 'United-States', 'United-States', 'United-States', 'united-states']
Possible formatting issues in 'income': ['<=50K', ' <=50K ', '<=50K', ' <=50K ', '<=50K', '<=50K', '<=50K', '<=50K', '<=50K', '<=50K']
```

Our next step was to handle the noise in the dataset for that we removed the columns that had almost constant entries or which had unique entries. So, the final output of this process is as follows

```
Dropping constant columns: ['capital_gain', 'capital_loss']
Dropping ID-like columns: ['random_string', 'hashed_id']

 Remaining columns after cleaning: 13
['age', 'workclass', 'fnlwgt', 'education', 'education_num', 'marital_status', 'occupation', 'relationship', 'race', 'sex', 'hours_per_week', 'native_country', 'income']
```

We can clearly see that capital gain and loss did not provide any significant value in our analysis from the boxplot thus this explains the decision. Similarly for the string and hash id columns

<u>Feature Engineering:</u>

For the feature engineering part, we added 2 new features in our dataset

1. Age band- divides our dataset into 6 different groups that will better help us understand the trends with their net income the groups were <25, 26-35, 36-45, 46-55, 56-65 and 65+.
2. Weekly hour band- like the age band this feature will help us to understand the correlation between the total hours spent working with respect to the education qualification of the person.

<u>Normalization:</u>

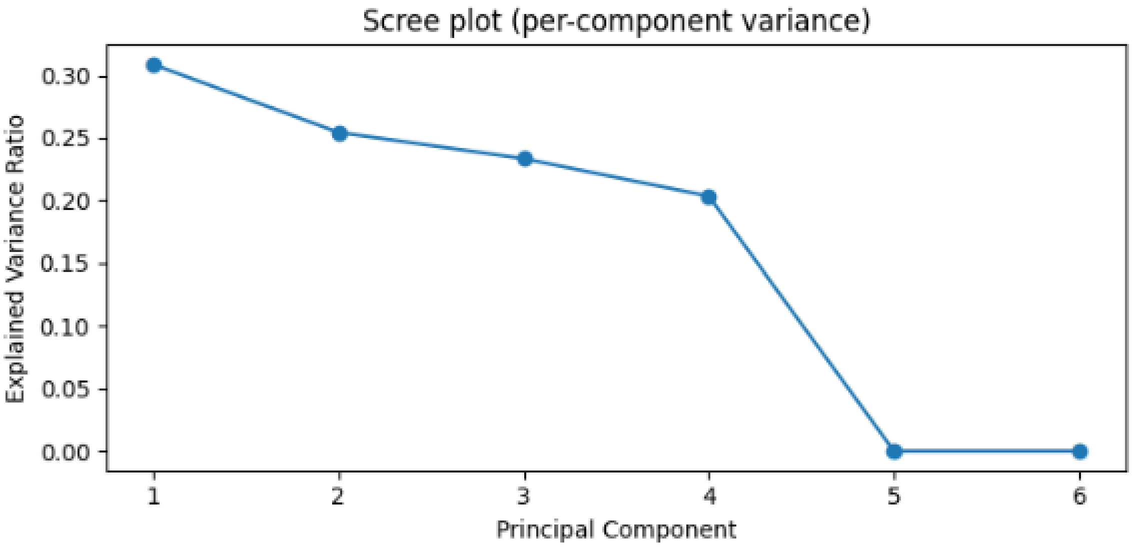
For normalization we used the standardscaler function to standardize the entries in numerical column as most of the data in our dataset follows normal distribution.

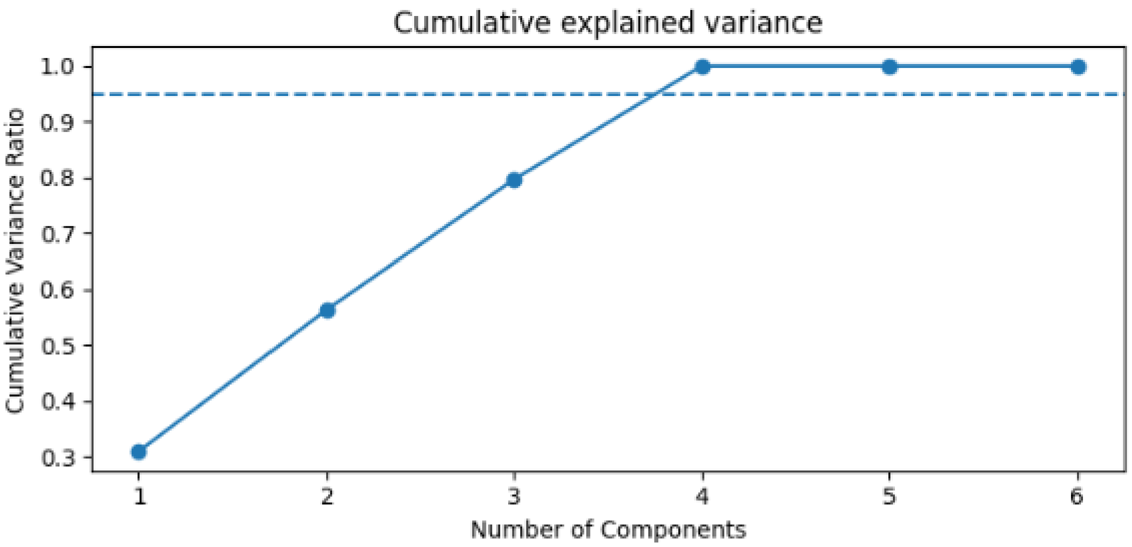
Given below is an snippet of the data after normalization

	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week
0	-0.977333	private	-0.388775	hs-grad	-0.465875	never-married	other-service	other-relative	asian-pac-islander	male	0.0	0.0	1.825037
1	0.143170	private	-0.533389	some-college	-0.049653	divorced	adm-clerical	unmarried	white	female	0.0	0.0	-0.517571
2	2.680778	private	0.933557	some-college	-0.049653	never-married	sales	not-in-family	white	female	0.0	0.0	-0.194453
3	-0.186390	private	0.536789	9th	-2.130764	never-married	priv-house-serv	not-in-family	white	female	0.0	0.0	-1.406147
4	-0.318214	private	0.007982	hs-grad	-0.465875	never-married	machine-op-inspct	not-in-family	white	male	0.0	0.0	-0.194453

Dimensionality reduction:

We performed PCA Analysis





- **Data used:** numeric-only matrix with **33,212 rows × 6 features**.
- **Dimensionality: 4 principal components (PCs)** are enough to reach **≥95% variance**.
  - EVR by PC ≈ PC1: 0.309, PC2: 0.254, PC3: 0.223, PC4: 0.203, PC5–PC6: ~0.
  - Clear “elbow” at **PC4** → after that, additional PCs add effectively no variance.

- **Rank/collinearity:** Two PCs having ~0 variance means the numeric block is effectively **rank-4** (strong multicollinearity or two variables carry negligible independent variance).

What each PC seems to capture (from loadings)

- PC1 (work/education intensity): dominated by hours_per_week (~0.63), education_num (~0.60), plus age (~0.41); fnlwgt contributes modestly; capital_gain/loss ≈ 0.
- PC2 (sampling weight + age contrast): strongest on fnlwgt (~0.72), with age (~−0.50), then education_num and hours_per_week.
- PC3 (age + weight): age (~0.77) and fnlwgt (~0.62) dominate; smaller role for education_num (~−0.26).
- PC4 (hours vs education): hours_per_week (~0.72) and education_num (~−0.67) drive this axis; age/fnlwgt minor.
- Across all PCs (max |loading|): top contributors are age, fnlwgt, hours_per_week, education_num.
  capital_gain and capital_loss show ~0 loadings in your table → they add virtually no variance in this scaled numeric space (likely due to being mostly zeros/sparse).

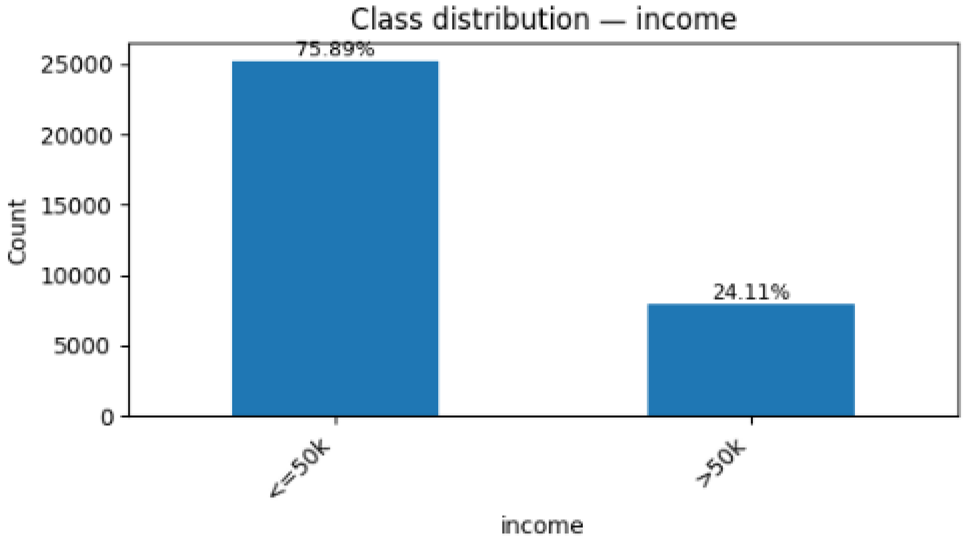
	PC	feature	abs_loading	signed_loading
0	PC1	hours_per_week	0.629443	0.629443
1	PC1	education_num	0.600225	0.600225
2	PC1	age	0.405336	0.405336
3	PC1	fnlwgt	0.281486	-0.281486
4	PC1	capital_gain	0.000000	-0.000000
5	PC1	capital_loss	0.000000	-0.000000
6	PC2	fnlwgt	0.722895	0.722895
7	PC2	age	0.504766	-0.504766
8	PC2	education_num	0.359899	0.359899
9	PC2	hours_per_week	0.305134	0.305134
10	PC2	capital_gain	0.000000	-0.000000
11	PC2	capital_loss	0.000000	-0.000000
12	PC3	age	0.737066	0.737066
13	PC3	fnlwgt	0.622235	0.622235
14	PC3	education_num	0.258865	-0.258865
15	PC3	hours_per_week	0.050470	0.050470
16	PC3	capital_gain	0.000000	-0.000000
17	PC3	capital_loss	0.000000	-0.000000
18	PC4	hours_per_week	0.712845	0.712845
19	PC4	education_num	0.665727	-0.665727
20	PC4	age	0.194031	-0.194031
21	PC4	fnlwgt	0.104939	-0.104939
22	PC4	capital_gain	0.000000	0.000000
23	PC4	capital_loss	0.000000	0.000000

```
Overall strongest contributors across ALL retained PCs (by max |loading| per feature):
```

	feature	max_abs_loading
0	age	0.737066
1	fnlwgt	0.722895
2	hours_per_week	0.712845
3	education_num	0.665727
4	capital_gain	0.000000
5	capital_loss	0.000000

<u>Distribution of Target Class:</u>

Our Target class is income the current distribution is as follows



Steps involved in the process to make the distribution equal
Loaded data; standardized names; summarized schema and distributions; flagged duplicates, constants, ID-like fields, formatting issues, and suspicious ranges. Imputed missing values (median/mode), capped outliers (IQR), normalized text, dropped noisy columns, engineered features (age bands, hours buckets, gain/marital flags), scaled numerics. Ran PCA on numeric block. Examined target balance, stratified split, frequency-encoded categoricals, applied SMOTE if imbalanced, and exported train/test CSVs.

Final outcome : Clean CSV file ready for modelling.