ML Assignment -3
Report

Pankaj Kumar
2019262

Q1.

Preprocessing

For PreProcessing part, I have first loaded the dataset.

- Replacing the ' ?' with the NaN value
- Then I have calculated how much null value is there in each column
- After calculating the percentage, I drop those columns which contain null values > 40
- 4 Column were dropped from the data frame which is as follow MIGMTR1, MIGMTR3, MIGMTR4, and MIGSUN containing the null values 99696.

Feature Analysis

For Feature Analysis, I create the list of the numerical and categorical columns as per the Dataset Description CSV, and then I have plotted the histogram values for each feature belonging from numerical as well as categorical.

And using the plots I have dropped those features in which most of the data is in one column and there is no data in the remaining column.

Numerical Columns

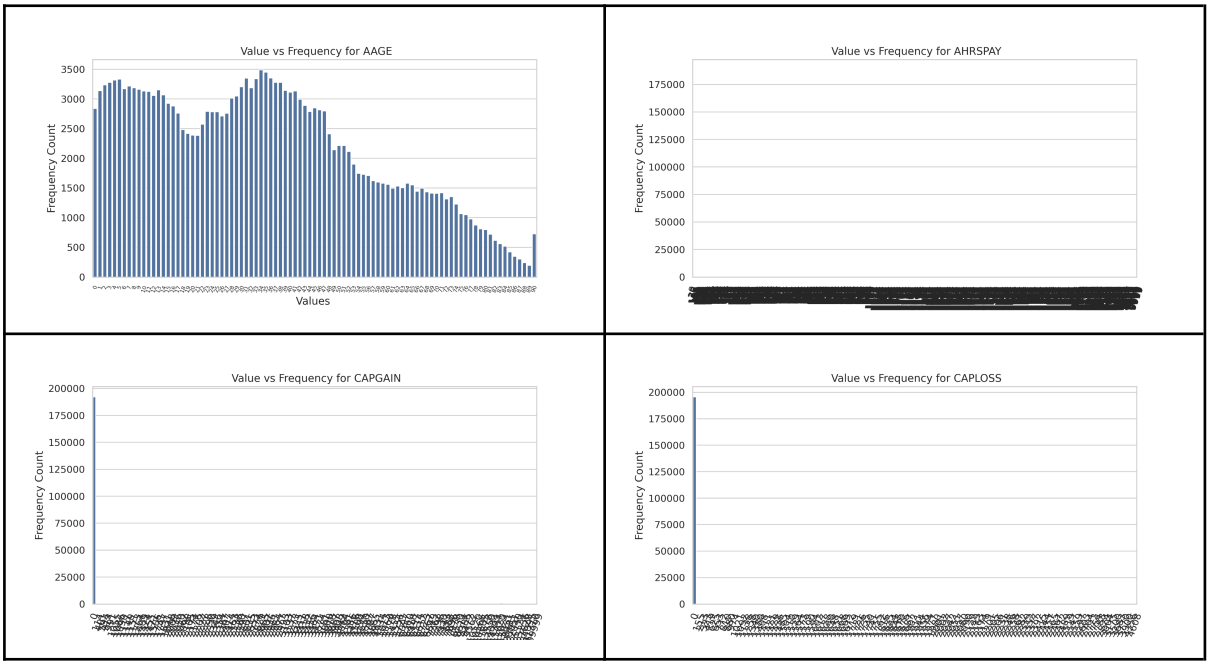
```
['ACLSWKR', 'ADTIND', 'ADTOCC', 'AHGA', 'AHSCOL', 'AMARITL', 'AMJIND',
'AMJOCC', 'ARACE', 'AREORGN', 'ASEX', 'AUNMEM', 'AUNTYPE', 'AWKSTAT',
'FILESTAT', 'GRINREG', 'GRINST', 'HHDFMX', 'HHDREL', 'MIGSAME',
'NOEMP', 'PARENT', 'PEFNTVTY', 'PEMNTVTY', 'PENATVTY', 'PRCITSHP',
'SEOTR', 'VETQVA', 'VETYN', 'YEAR']
```

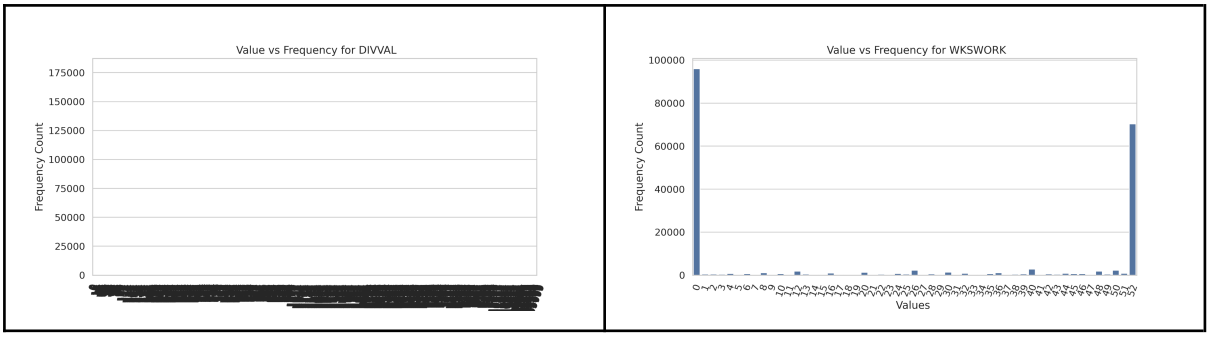
Categorical Columns

```
['AAGE', 'AHRSPAY', 'CAPGAIN', 'CAPLOSS', 'DIVVAL', 'WKSWORK']
```

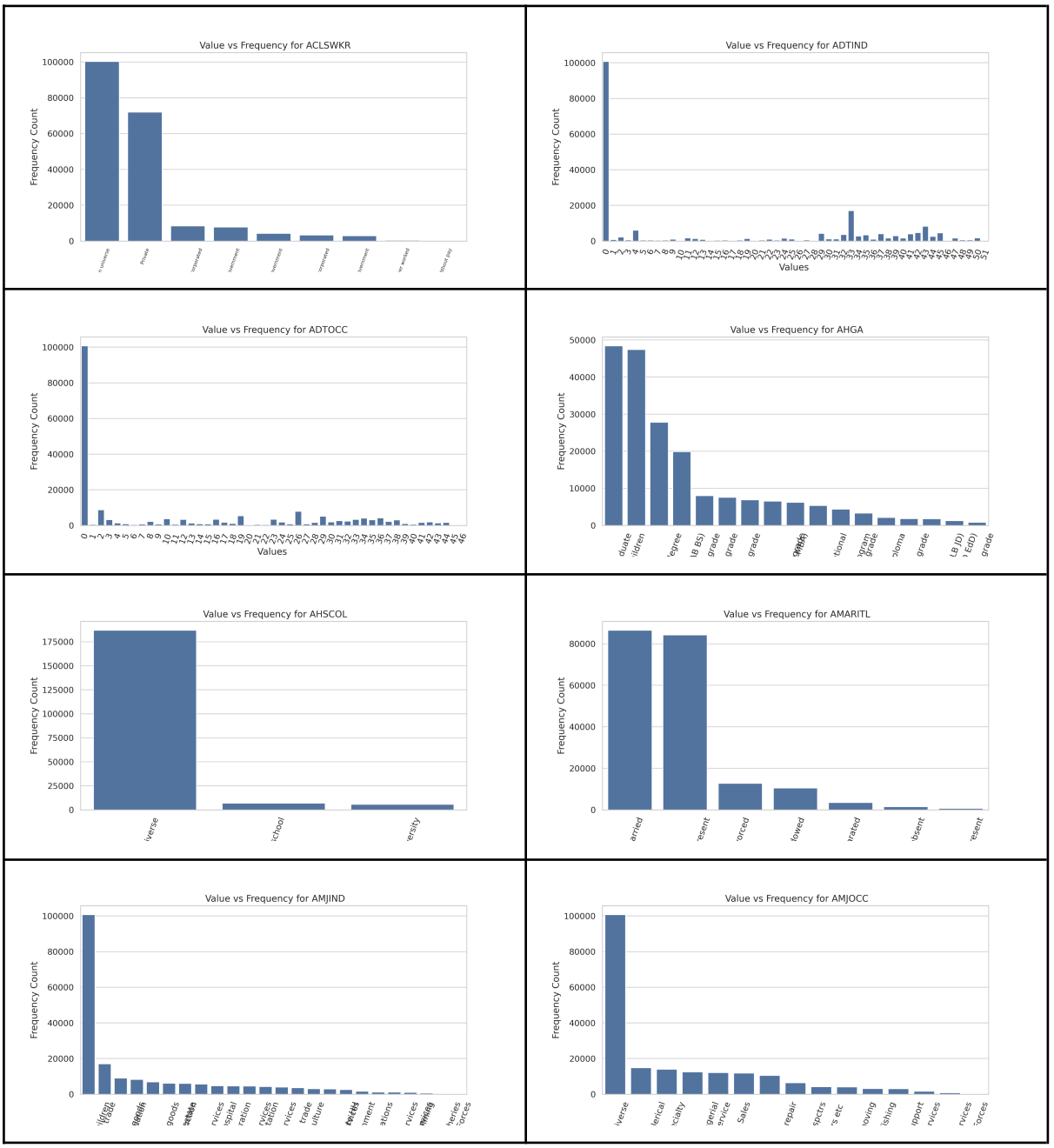
Now here the plots of the each columns

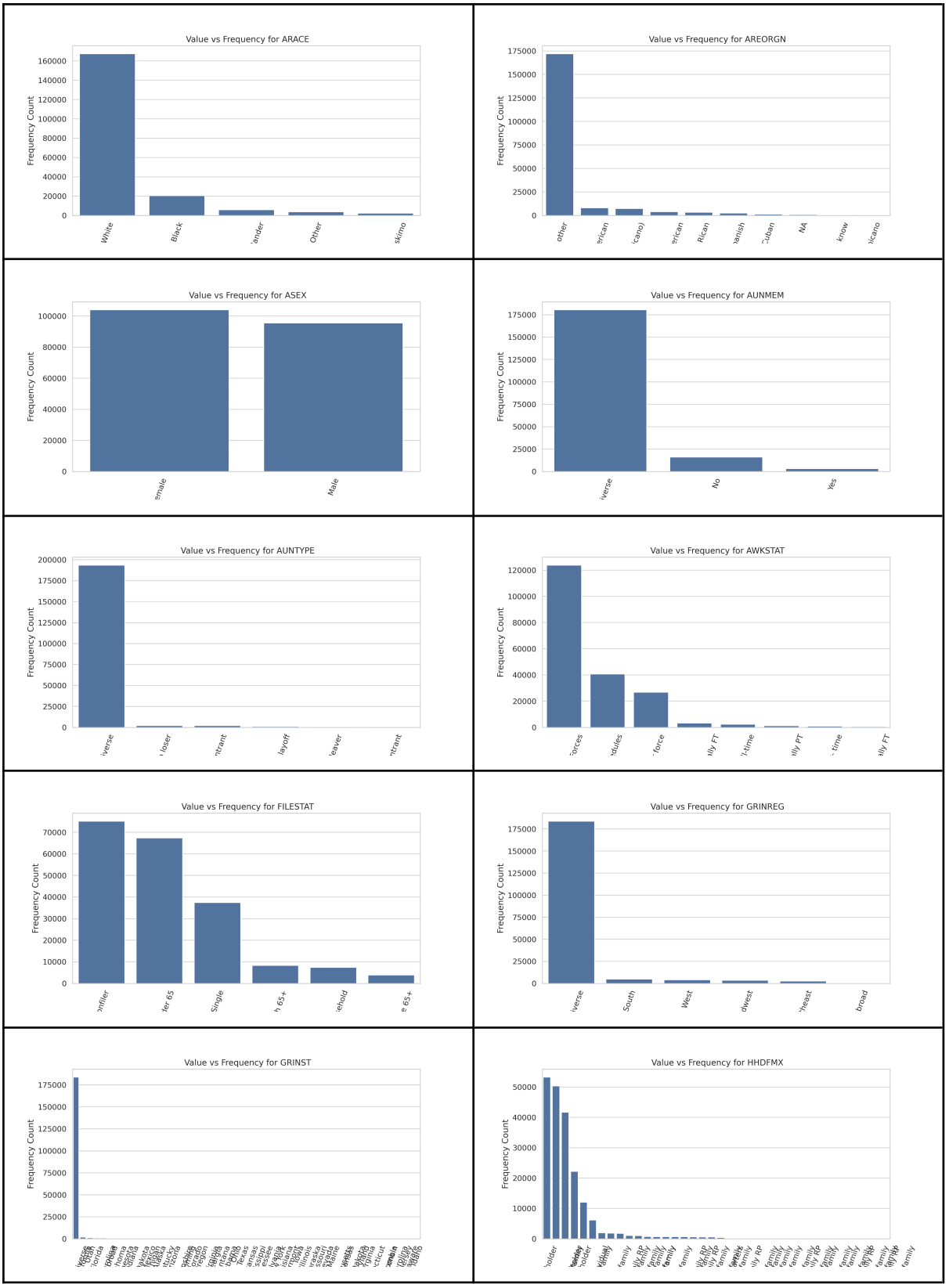
Numerical

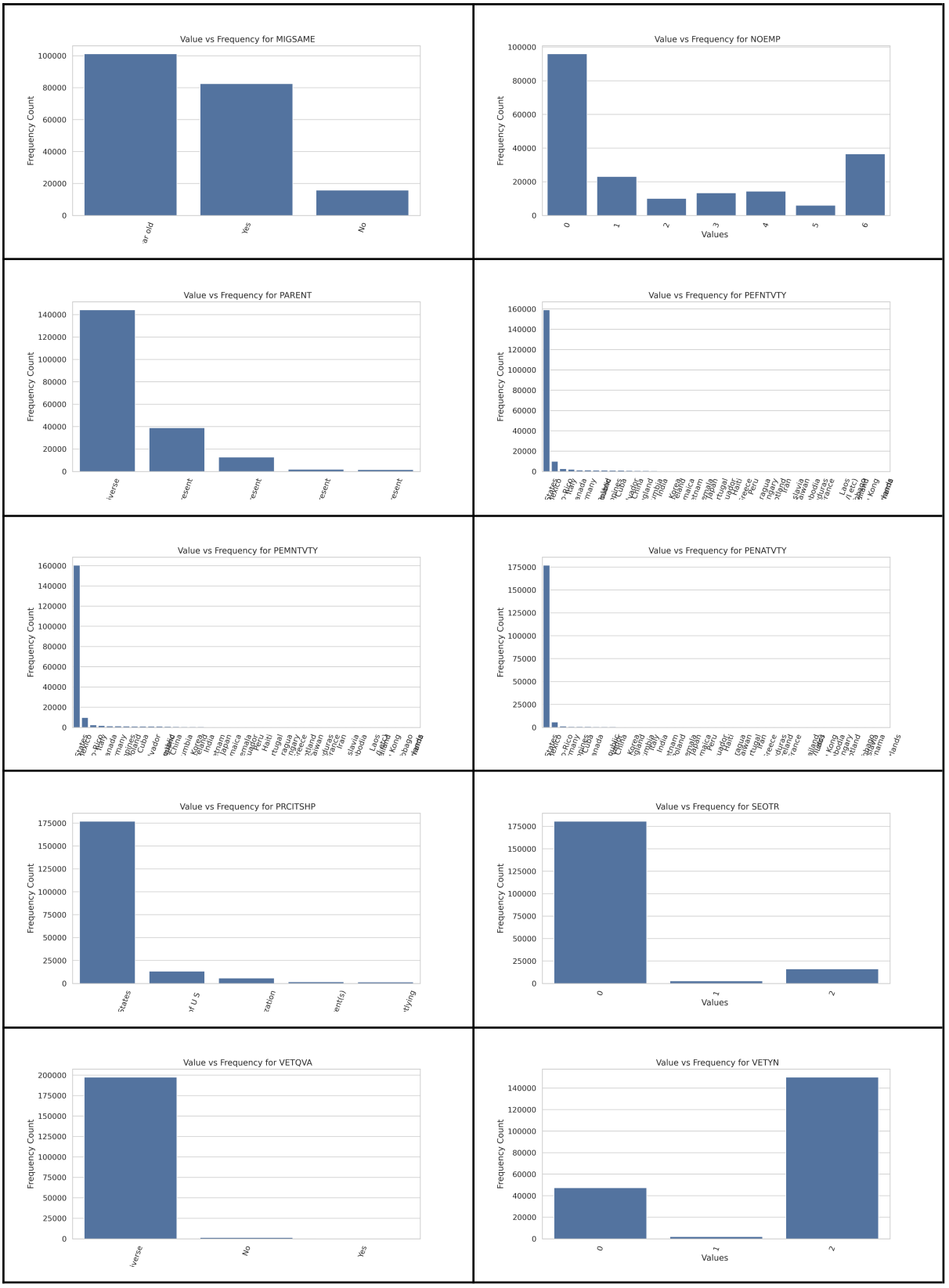


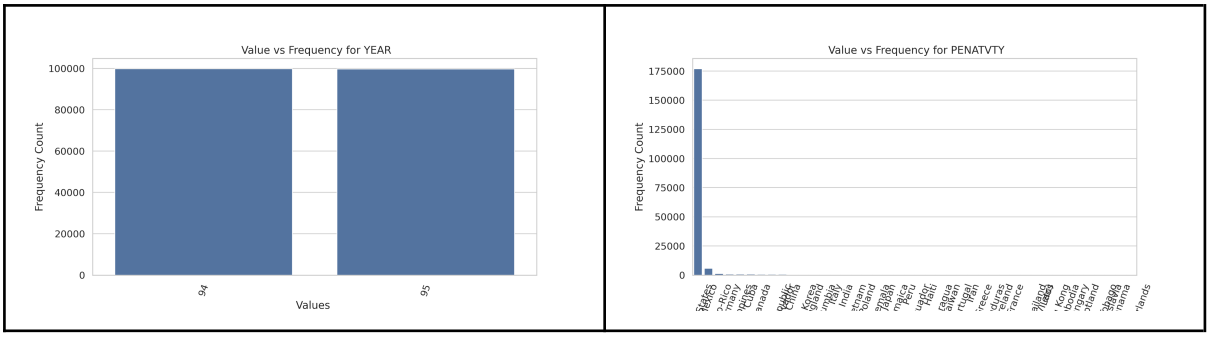


Categorical




Value vs Frequency for ARACE
Frequency Count
White
Black
Other
Value vs Frequency for AREORGN
Frequency Count
NA
Value vs Frequency for ASEX
Frequency Count
Male
Value vs Frequency for AUNMEM
Frequency Count
No
Yes
Value vs Frequency for AUNTYPE
Frequency Count
layoff
Value vs Frequency for AWKSTAT
Frequency Count
Value vs Frequency for FILESTAT
Frequency Count
Single
Value vs Frequency for GRINREG
Frequency Count
South
West
broad
Value vs Frequency for GRINST
Frequency Count
Value vs Frequency for HHDFMX
Frequency Count

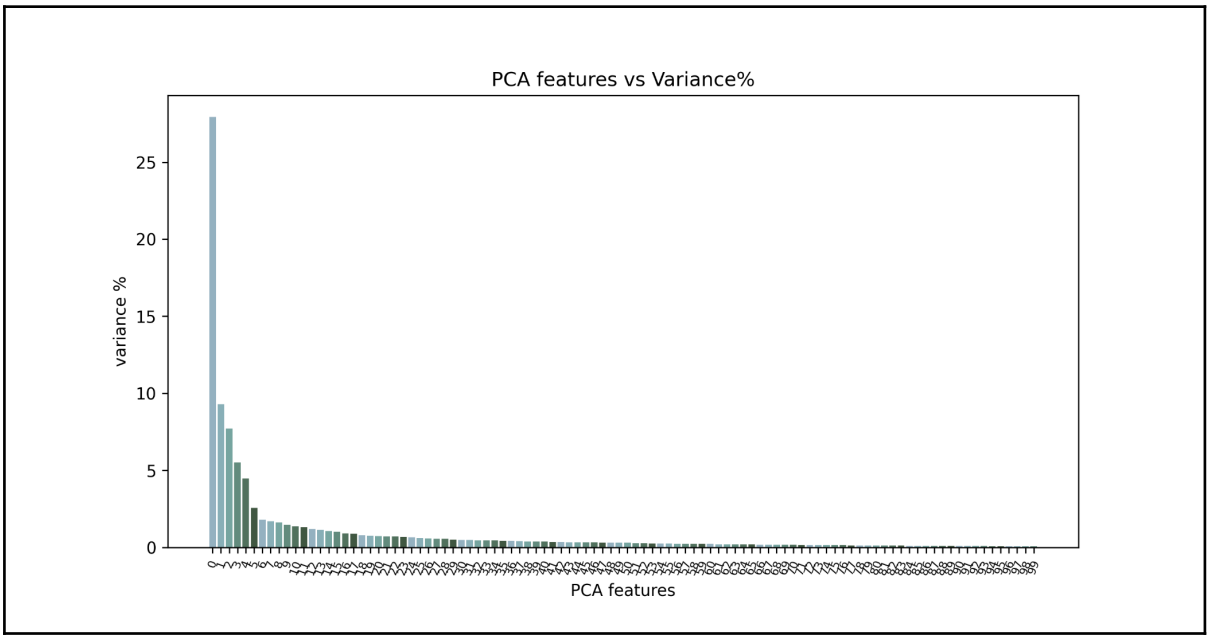

Value vs Frequency for MIGSAME
Frequency Count
Yes
No
Value vs Frequency for NOEMP
Frequency Count
Values
Value vs Frequency for PARENT
Frequency Count
Value vs Frequency for PEFNTVTY
Frequency Count
Value vs Frequency for PEMNTVTY
Frequency Count
Value vs Frequency for PENATVTY
Frequency Count
Value vs Frequency for PRCITSHP
Frequency Count
Value vs Frequency for SEOTR
Frequency Count
Values
Value vs Frequency for VETQVA
Frequency Count
No
Yes
Value vs Frequency for VETYN
Frequency Count
Values



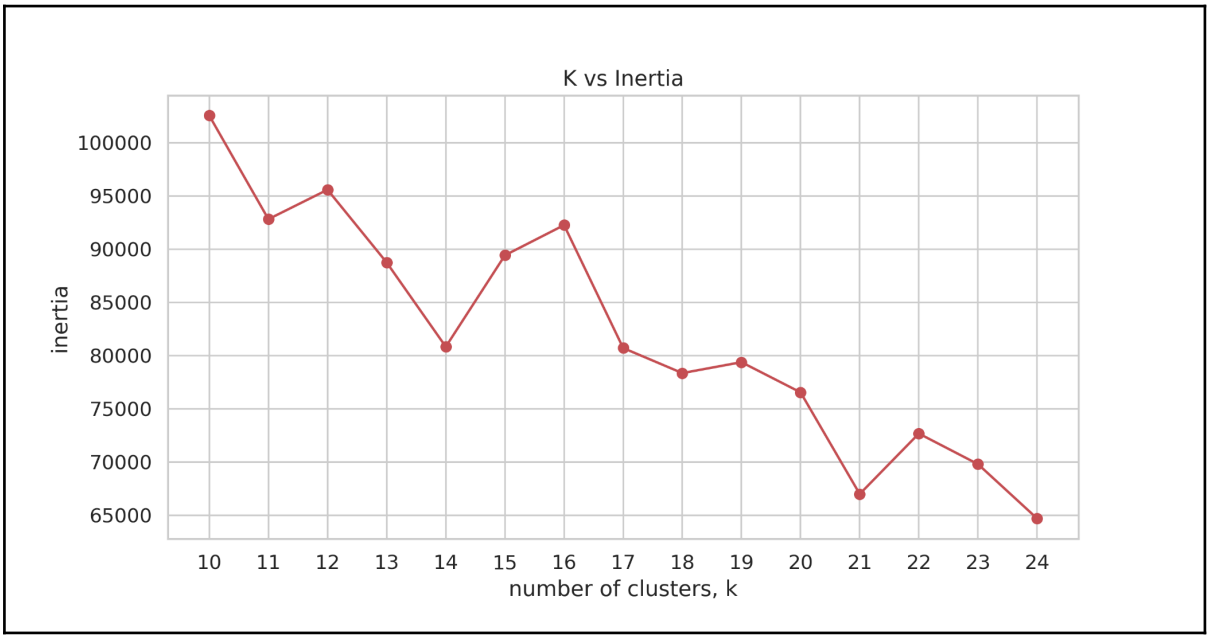
Imputation, Bucketization, and One-Hot Encoding

- For this part, I have replaced the NaN value with mode for each column.
- Now for the bucketization, I find the final numerical and categorical column present in the Dataframe.
- Then I have bucketed (Column “AAGE” and “WKSWORK”) them with the proper label.
- For AAGE ,
  `labels=['Child','Youth', 'Adult','Senior ']`
- For WKSWORK,
  `labels=['Fresher','Experienced','Highly-Experienced','Senior ']`
- Now For One hot encoding, I set the categorical column type to ‘object’.
- Then using the get_dummies(), I have done the One Hot Encoding.

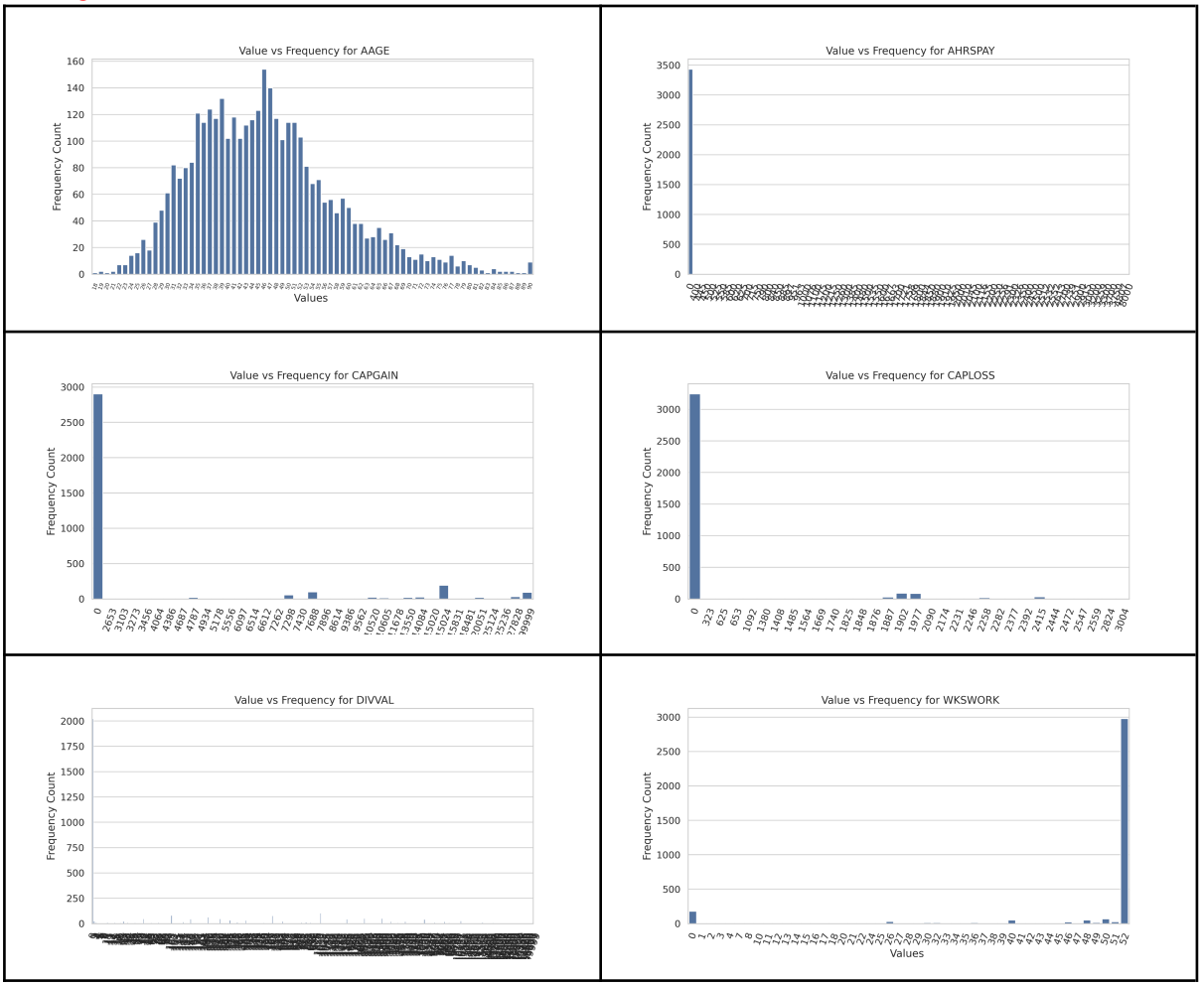
PCA



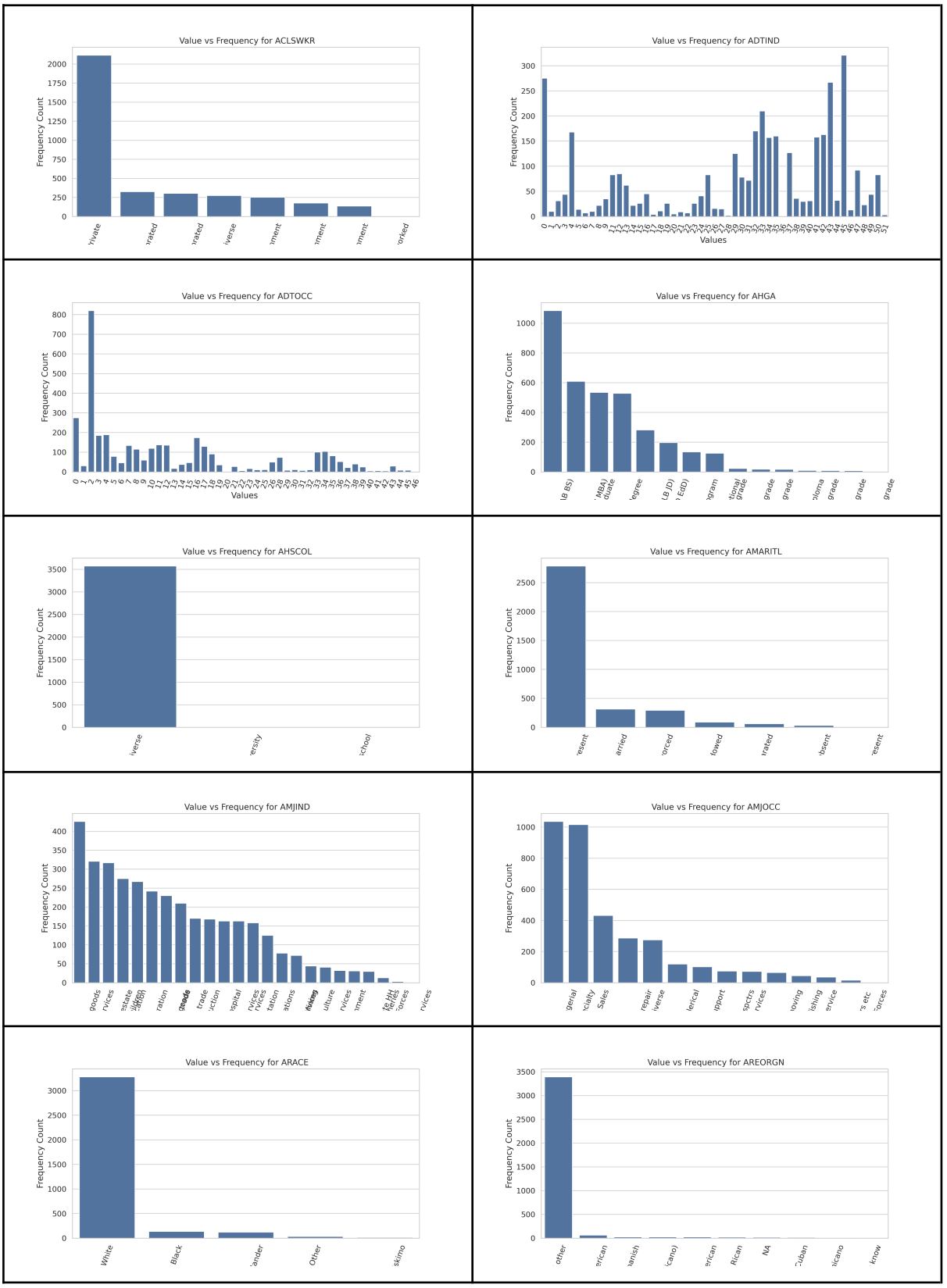
Elbow Plot

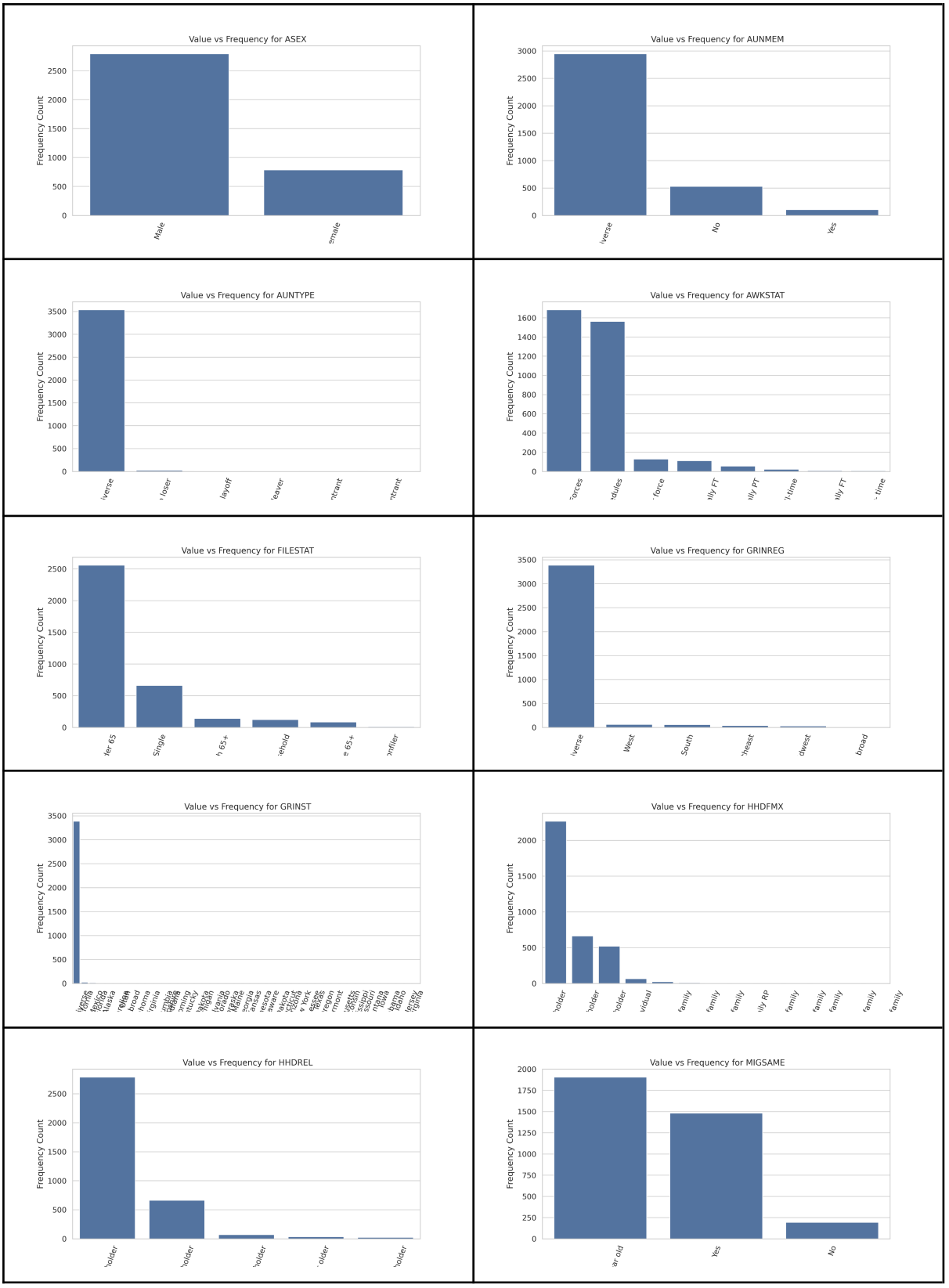


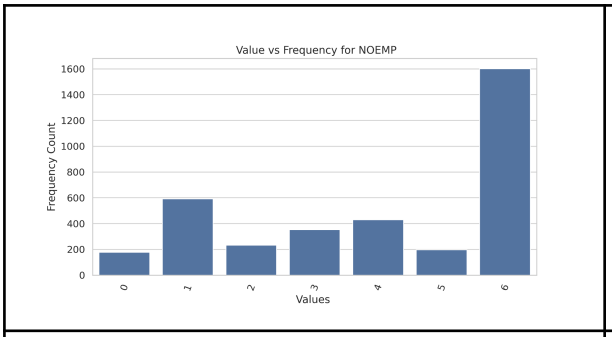
Categorical

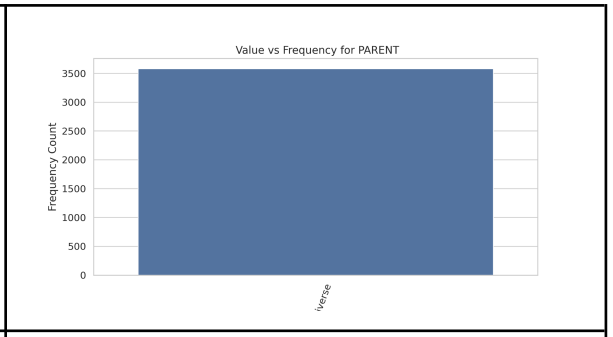


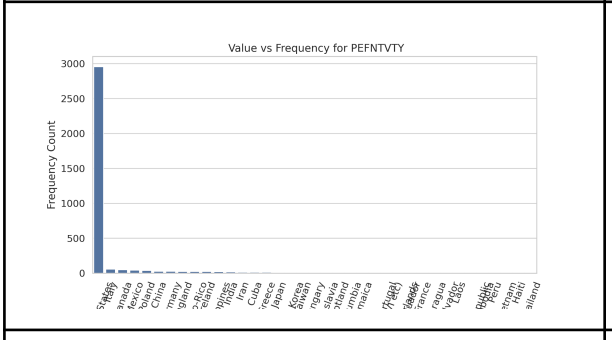
Categorical

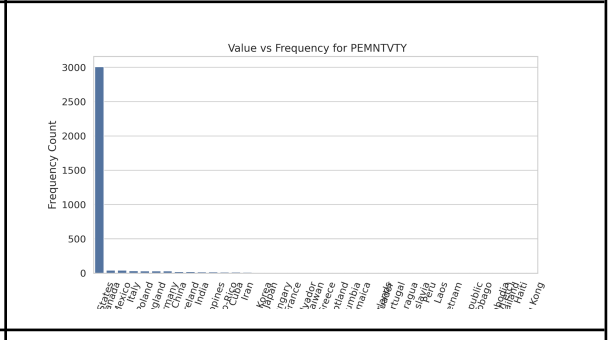


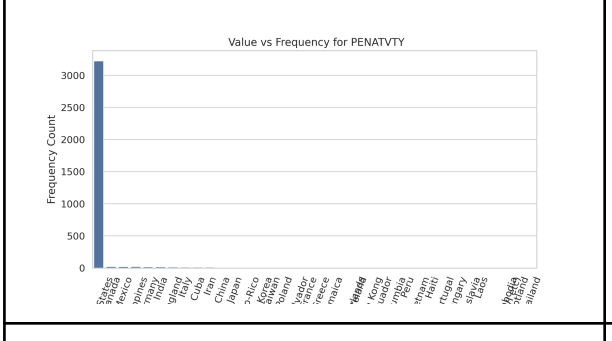

Value vs Frequency for ASEX
Frequency Count
Male
Value vs Frequency for AUNMEM
Frequency Count
No
Yes
Value vs Frequency for AUNTYPE
Frequency Count
Value vs Frequency for AWKSTAT
Frequency Count
Value vs Frequency for FILESTAT
Frequency Count
Single
Value vs Frequency for GRINREG
Frequency Count
West
South
broad
Value vs Frequency for GRINST
Frequency Count
Value vs Frequency for HHDFMX
Frequency Count
Value vs Frequency for HHDREL
Frequency Count
Value vs Frequency for MIGSAME
Frequency Count
Yes
No

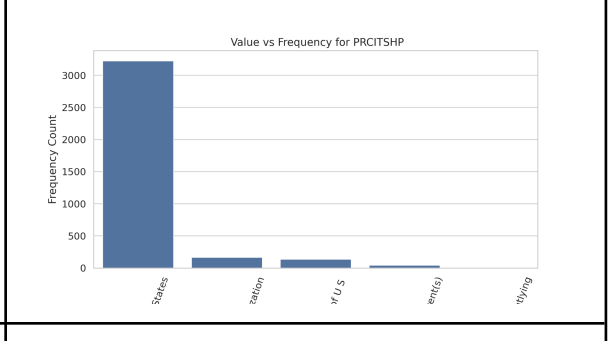

Value vs Frequency for NOEMP
Frequency Count
Values

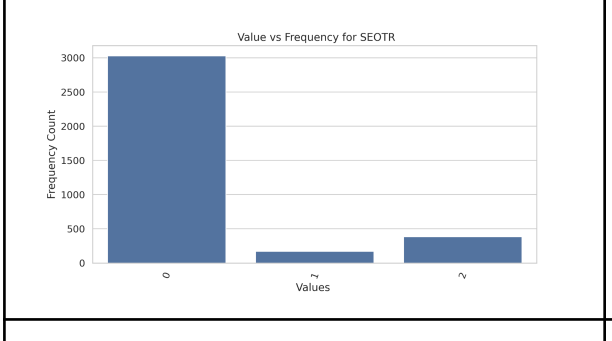

Value vs Frequency for PARENT
Frequency Count

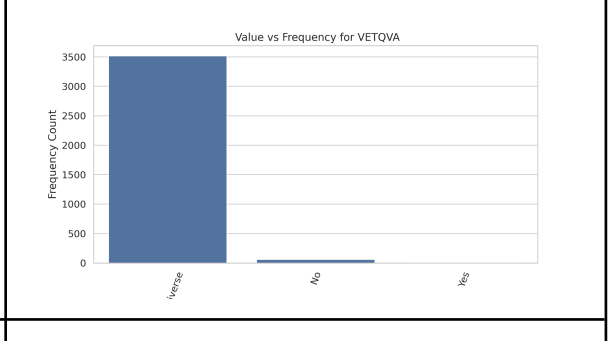

Value vs Frequency for PEFNTVTY
Frequency Count

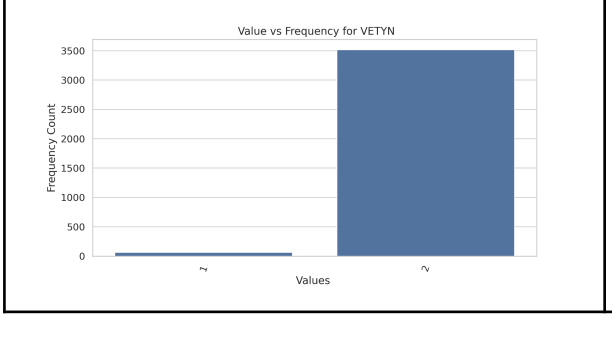

Value vs Frequency for PEMNTVTY
Frequency Count

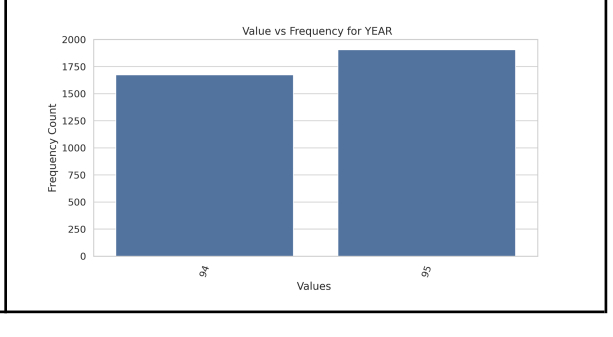

Value vs Frequency for PENATVTY
Frequency Count


Value vs Frequency for PRCITSHP
Frequency Count

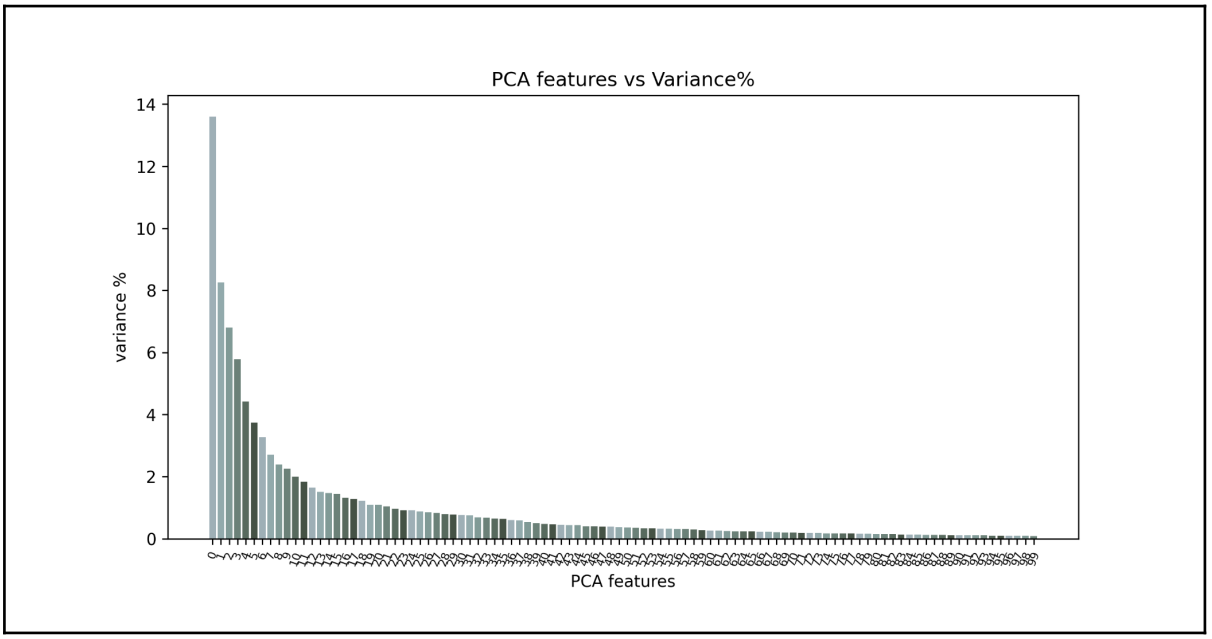

Value vs Frequency for SEOTR
Frequency Count
Values


Value vs Frequency for VETQVA
Frequency Count

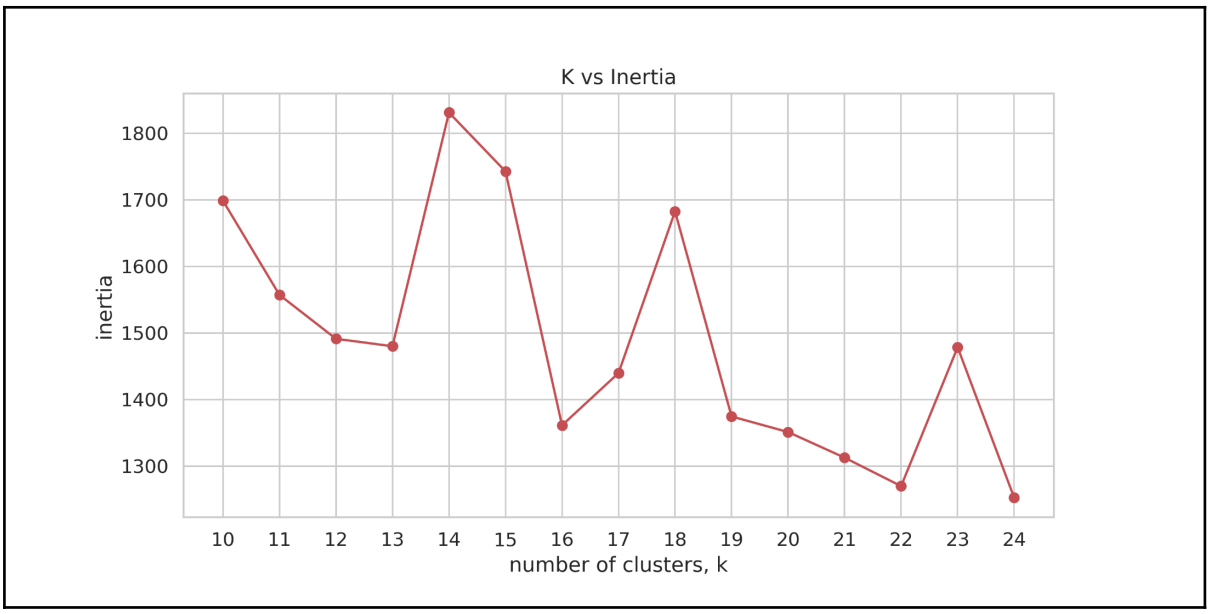

Value vs Frequency for VETYN
Frequency Count
Values


Value vs Frequency for YEAR
Frequency Count
Values

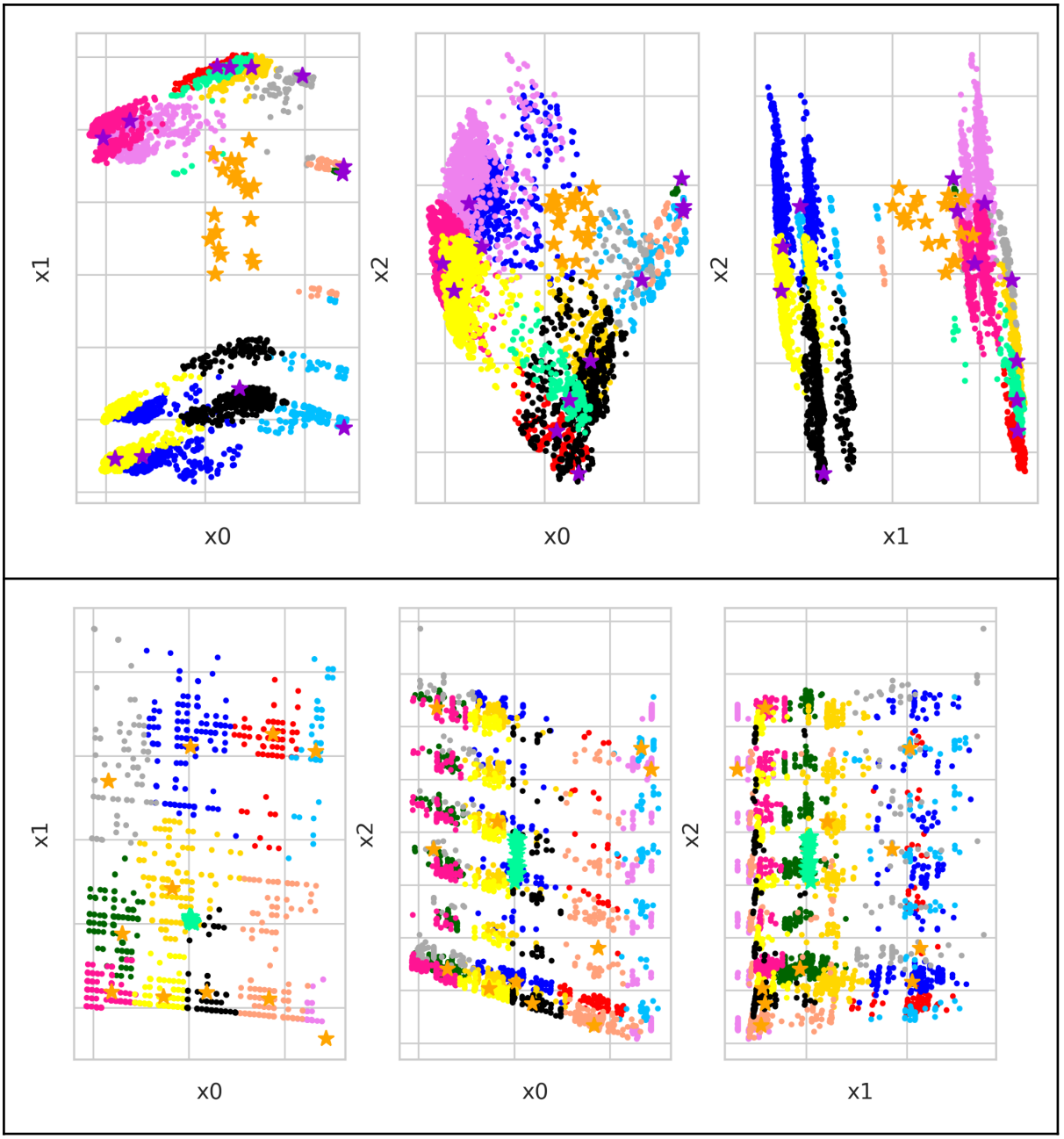
PCA



Elbow Plot



Scatter Plot at K=22



Scatter Plot at K=22

Q6.2 & 6.4
I can interpret over-represented clusters as follows based on the preceding math and visualisation:
In General Population data ,1 3 4 8 9 10 13 14 15 16 17 (Clusters)
All clusters are overrepresented in more than 50k data

.Q6.3

```
array([[-8.98849346e-03, -1.12084034e-02, -5.01675266e-04, ...,
        -1.38262873e-02, -7.56169498e-04,  1.00075617e+00],
       [ 2.83216183e-03, -2.80299457e-03,  9.94440036e-01, ...,
         9.99348032e-01,  1.00515486e+00, -5.15485549e-03],
       [-1.29423599e-03,  1.01464731e+00,  1.41661991e-02, ...,
```

```
        -1.19879640e-02, -1.04302002e-02,  1.01043020e+00],
       ...,
       [-1.32989875e-02, -1.43796726e-02,  9.89258106e-01, ...,
         9.89597222e-01, -7.25865113e-03,  1.00725865e+00],
       [ 1.04434945e-02,  9.57984504e-01, -4.79371344e-02, ...,
        -1.09178372e-02,  3.30735655e-02,  9.66926435e-01],
       [-1.14341337e-02, -6.67487743e-03,  9.94230489e-01, ...,
         9.93228242e-01,  1.00836499e+00, -8.36499263e-03]])
```

```
array([[-7.15702002e-17, -2.78127892e-03,  1.00067974e+00, ...,
         9.98534848e-01,  7.67751430e-04,  9.99232249e-01],
       [-3.30015669e-17, -2.43093251e-03,  1.00452947e+00, ...,
         1.00411942e+00,  9.97525760e-01,  2.47423973e-03],
       [-1.54799055e-17, -1.88983770e-03,  1.00413863e+00, ...,
         1.01414120e+00,  9.98505745e-01,  1.49425495e-03],
       ...,
       [ 1.87567913e-17, -1.00174857e-03,  9.98182565e-01, ...,
         1.00304301e+00,  1.94502562e-03,  9.98054974e-01],
       [-8.61412911e-17, -1.43578070e-03,  1.00048372e+00, ...,
         9.99425354e-01,  1.00140303e+00, -1.40302721e-03],
       [-8.56608970e-17, -1.07261507e-03, -1.78250029e-03, ...,
        -1.03195305e-03,  1.00136671e+00, -1.36670817e-03]])
```

```
PCA_components_inverse_test=pca.inverse_transform(PCA_components_test)
PCA_components_inverse_test
```

Q2.

Q2 SVM

3 Given : we have linearly separable data

optimization problem : $\min_{w,b,\epsilon} \; 1/2\,\|w\|^2 + C/2 \sum_{i=1}^{m} \epsilon^2$

such that $y^{(i)}(w^T x^i + b) \geq 1 - \epsilon_i$ where $i = 1 \dots m$

a) No, because if we look at the ϵ, then at particular ϵ_i, if ϵ_i becomes negative, then objective function will reduced that to 0.

i.e $\epsilon_i = 0$.

Therefore there is no need of $\epsilon \geq 0$ for L2 norm soft margin optimization problem

b) Lagrangian of the L2 norm soft margin optimization problem is given by

$$L(w, b, \epsilon, \alpha) = \frac{1}{2}\|w\|^2 + \frac{C}{2}\left(\epsilon_i^2 + \sum \alpha_i \left(1 - (y_i(w^T x^i + b) - \epsilon_i)\right)\right)$$

$$= \frac{1}{2}\|w\|^2 + \frac{C}{2}\left(\epsilon_i^2 + \sum \alpha_i \left(1 - (y_i(w^T x^i + b) - \epsilon_i)\right)\right)$$

(c) Now we have, $L(w, b, \varepsilon, \alpha)$

$\frac{\partial L}{\partial w} = 0$, $w - \sum \alpha_i y_i x^i$

$\therefore w = \sum_{i=1}^{m} \alpha_i y_i x^i$ — (i)

$\frac{\partial L}{\partial b} = 0$, $\sum \alpha_i y_i = 0$ — (ii)

$\frac{\partial L}{\partial \varepsilon} = 0$, $C\varepsilon - \alpha_i = 0$, $\varepsilon = \frac{\alpha_i}{C}$ — (iii)

we found the above result, using KKT Conditions

Therefore

Dual = $\max \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_j y_i y_j x^{iT} x^j$

provided that $\sum \alpha_i y_i = 0$.

$\therefore 0 \leq \alpha_i \leq C$

Q3

Given: decision function by the sum.

$$f(x) = \sum_{i=1}^{m} \alpha_i y^{(i)} k(x^i, x) + b$$

Now,

here $\alpha_i = 1$ and $b = 0$ for all $i \in [1, m]$.

Therefore,

$$|f(x^i) - y_i| = \left| \sum_{j=1}^{m} y^j k(x^j, x) - y^j \right|$$

$$= \left| \sum_{j=1}^{m} y^j e^{-\|x^j - x^i\|^2/\tau^2} - y^i \right|$$

$$= \left| y^i + \sum_{j=1} y^j e^{-\|x^j - x^i\|^2/\tau^2} - y^i \right|$$

Now, here, we can see that

$$\leq \sum_{j \neq i} |y_i e^{-\|x^j - x^i\|/|\tau^2|}$$

because we know that $|\Sigma a| \leq \Sigma |a|$

$$= \sum_{j \neq i} |y^j| e^{-\|x^j - x^i\|/\tau^2}$$

$$\therefore \leq \sum_{j \neq i} e^{-\varepsilon^2/\tau^2}$$

Now, $(m-1) e^{-\varepsilon^2/\tau^2}$

we can also say that $(m-1) e^{-\varepsilon^2/\tau^2} < 1$ (taking log)

$$\therefore \boxed{\tau < \frac{\varepsilon}{\log(m-1)}}$$

(b) No, because we know that parameter 'c' will reset the relative weight of $c\sum_{i=1}^{m} \xi_i$ and $1/2 \|w\|^2$ and we can say that it will effectively control the both terms alone.

On other hand, if c is small then weight will have small effect or small room.

Therefore we can say that resultant weight will be approximately zero, but not zero.