Assignment 3

- Manvi Goel

Question 1.

1. Preprocessing

   1.1 Replacing the values using *pd.replace()*. We need to note that '*?*' is present in form of " ?" and needs to be replaced appropriately.

```
pop_path.replace(" ?", np.nan, inplace = True)
pop_path.head()
```

S	DIVVAL	FILESTAT	GRINREG	GRINST	HHDFMX	HHDREL	MIGMTR1	MIGMTR3	MIGMTR4	MIGSAME	MIGSUN	NO
0	0	Nonfiler	Not in universe	Not in universe	Other Rel 18+ ever marr not in subfamily	Other relative of householder	NaN	NaN	NaN	Not in universe under 1 year old	NaN	

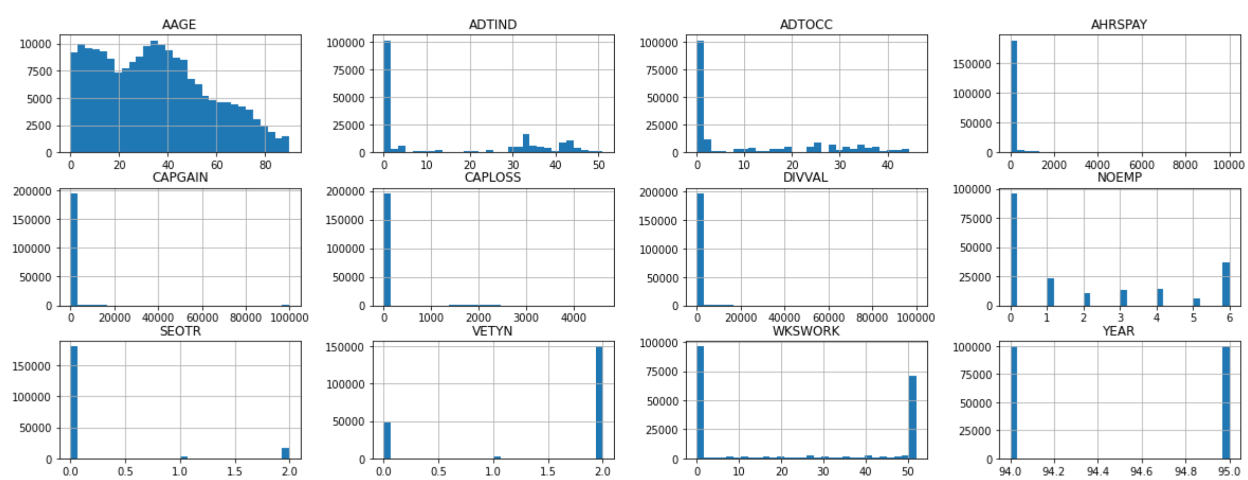
   1.2 Checking the percentage of missing data.

	Number of missing values	Percentage of Missing Values
MIGMTR1	99696	49.967172
MIGSUN	99696	49.967172
MIGMTR4	99696	49.967172
MIGMTR3	99696	49.967172
PEFNTVTY	6713	3.364524
PEMNTVTY	6119	3.066814

   Thus the top four columns namely "MIGMTR1, MIGSUN, MIGMTR4, MIGMTR3" are dropped from the dataset.

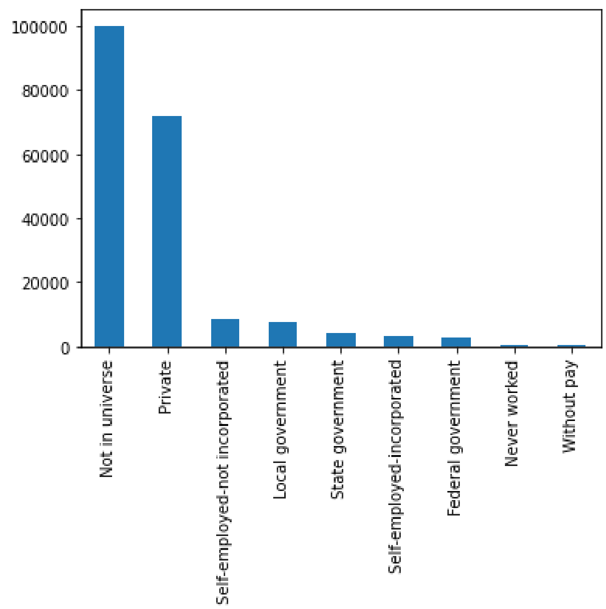
2. <u>Feature Analysis</u>

   2.1 Numerical Columns

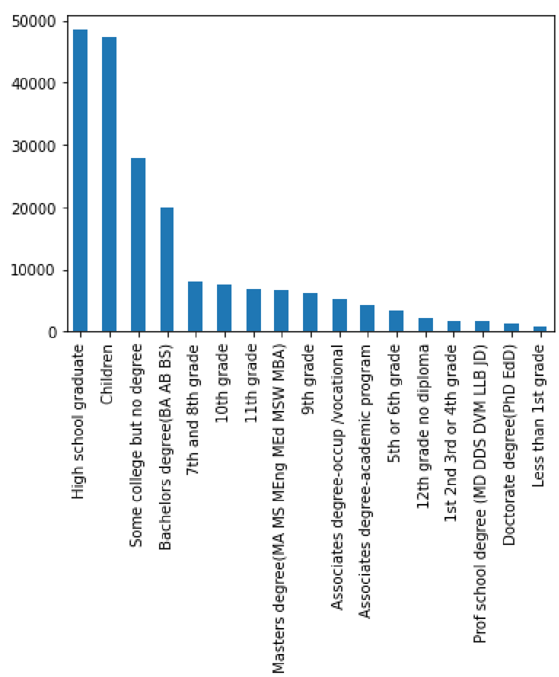


   Categorical Columns

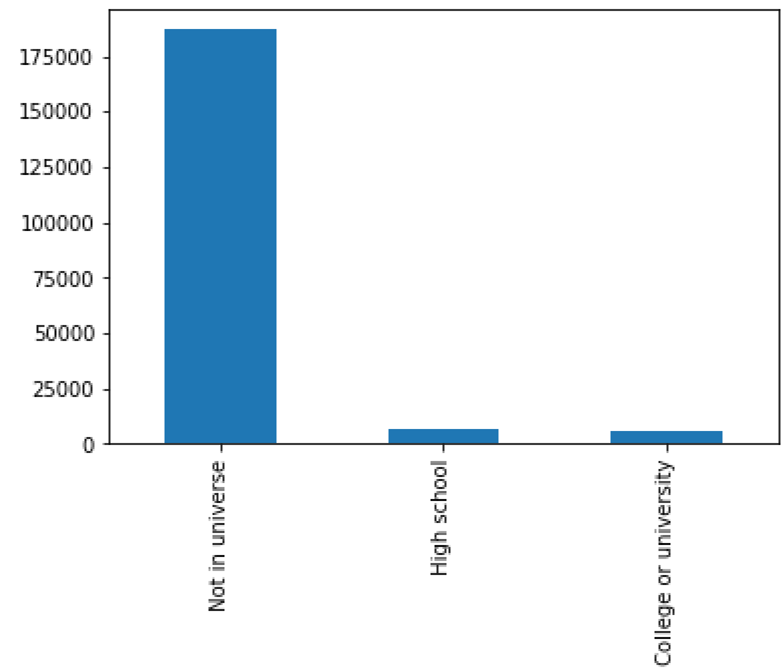
The histogram for column ACLSWKR



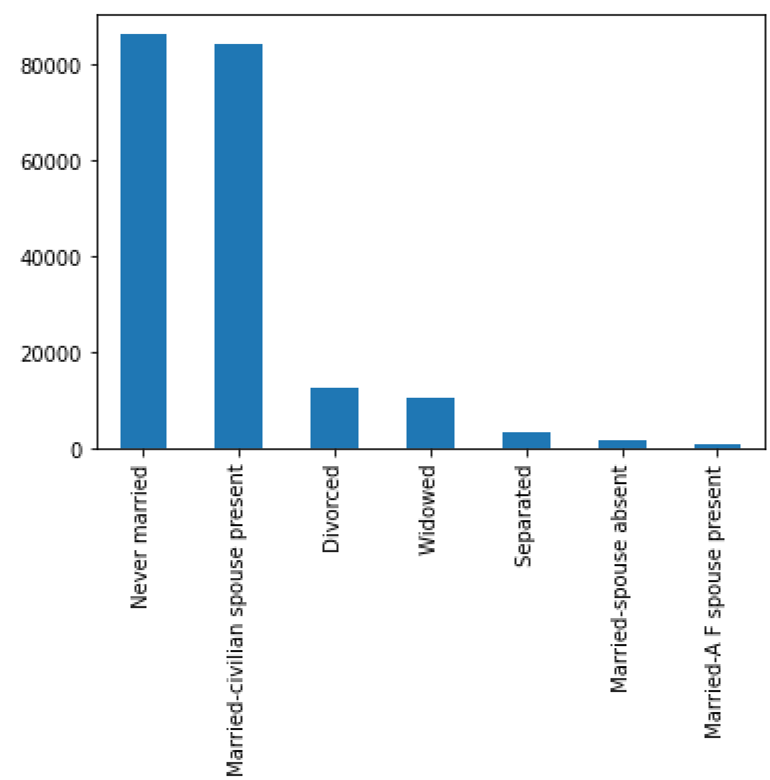
The histogram for column AHGA



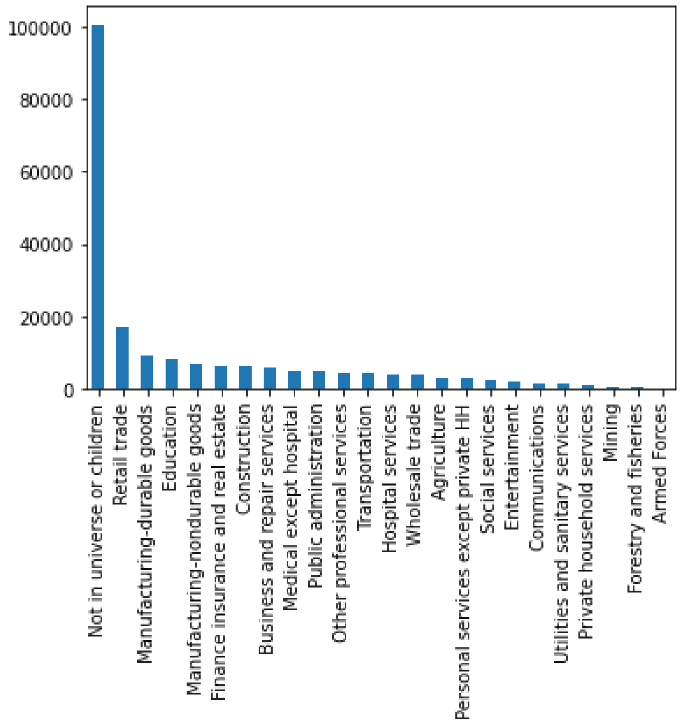
The histogram for column AHSCOL



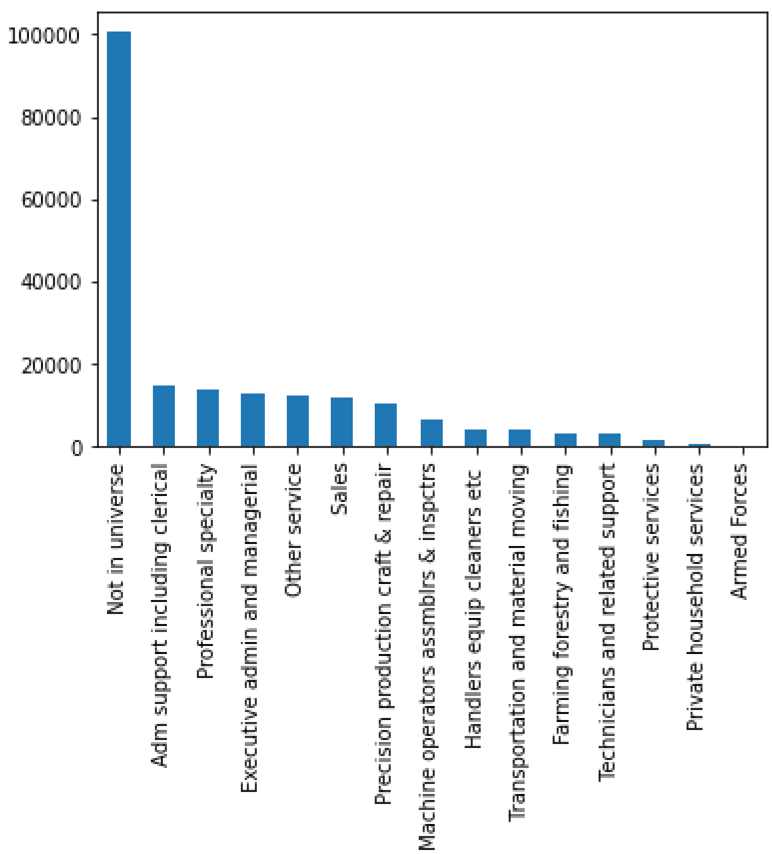
The histogram for column AMARITL



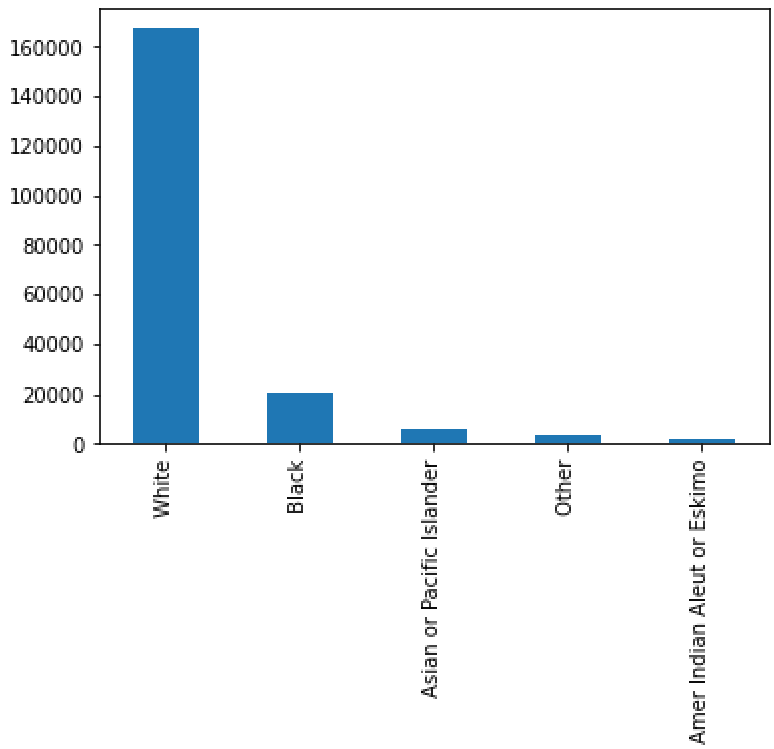
The histogram for column AMJIND



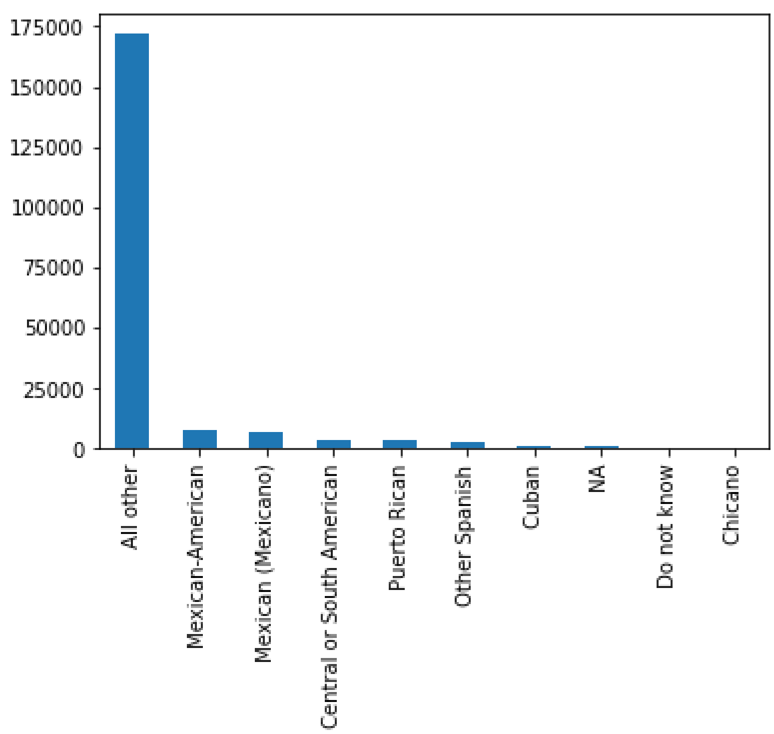
The histogram for column AMJOCC



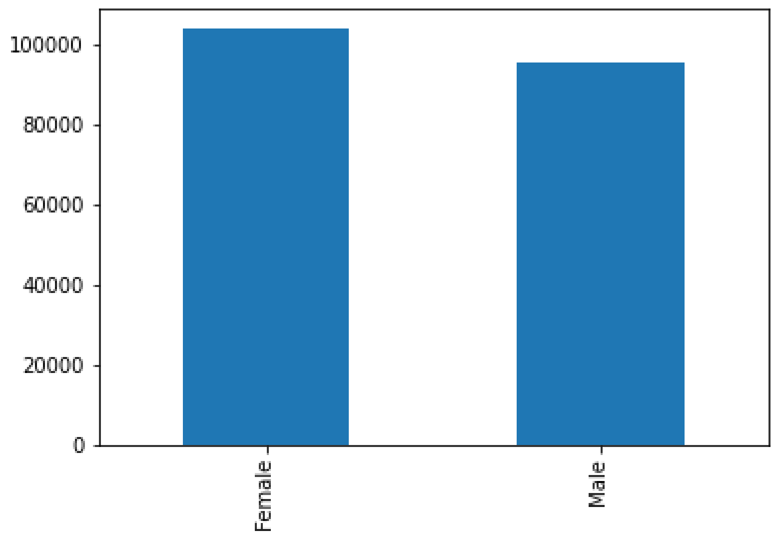
The histogram for column ARACE



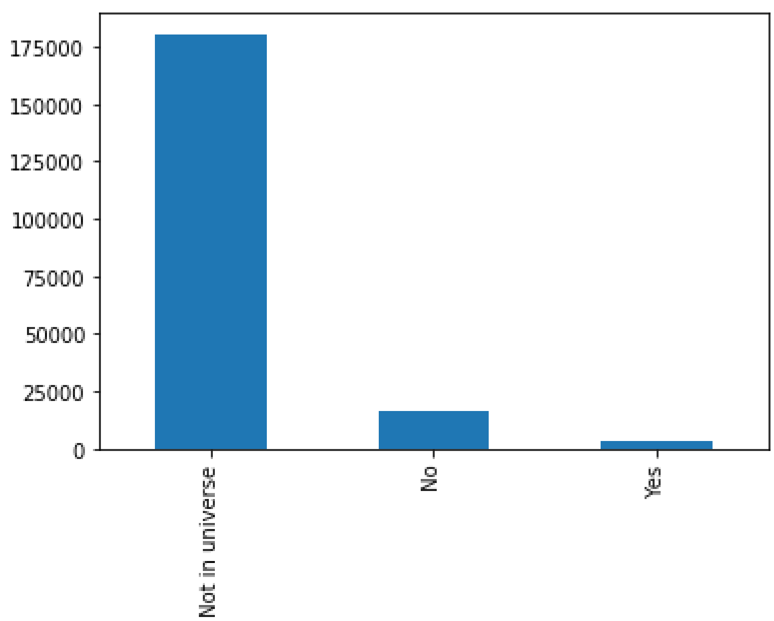
The histogram for column AREORGN



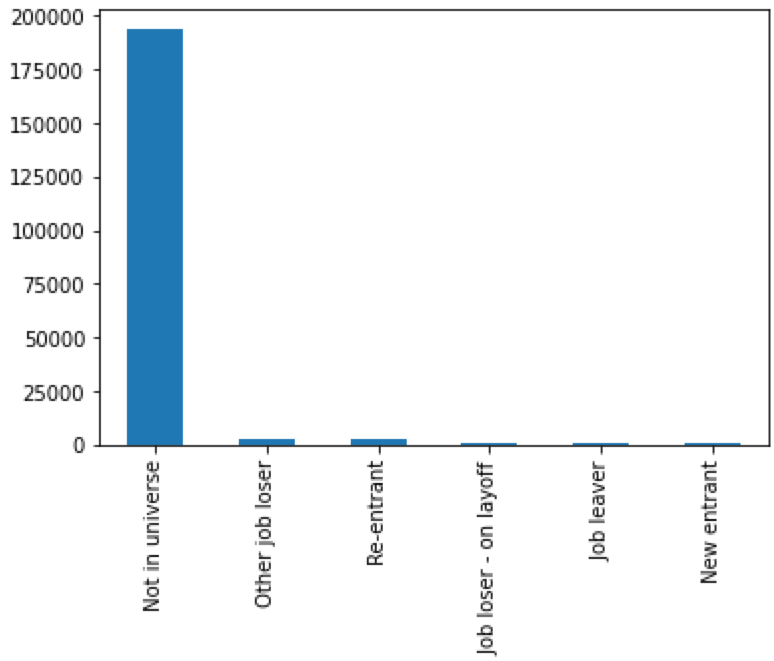
The histogram for column ASEX



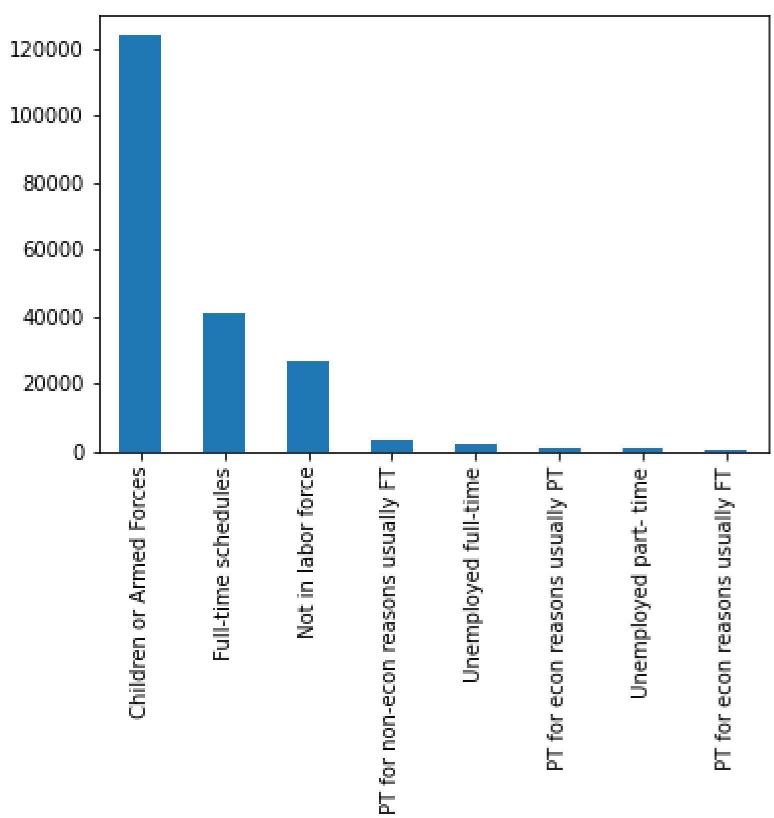
The histogram for column AUNMEM



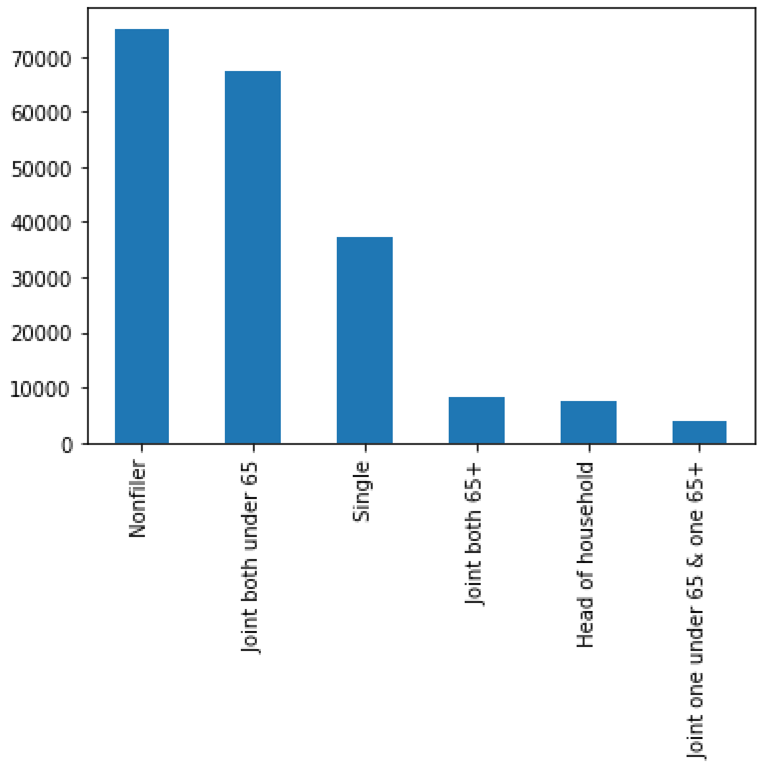
The histogram for column AUNTYPE



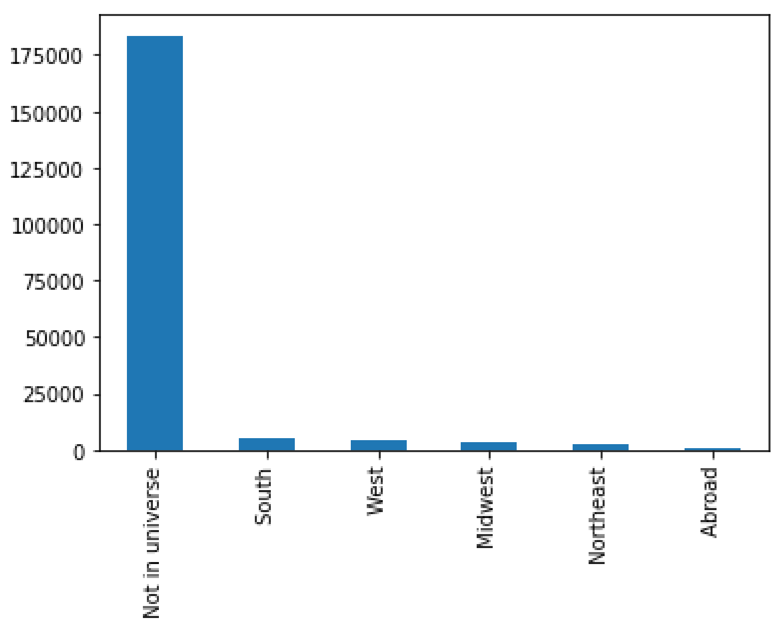
The histogram for column AWKSTAT



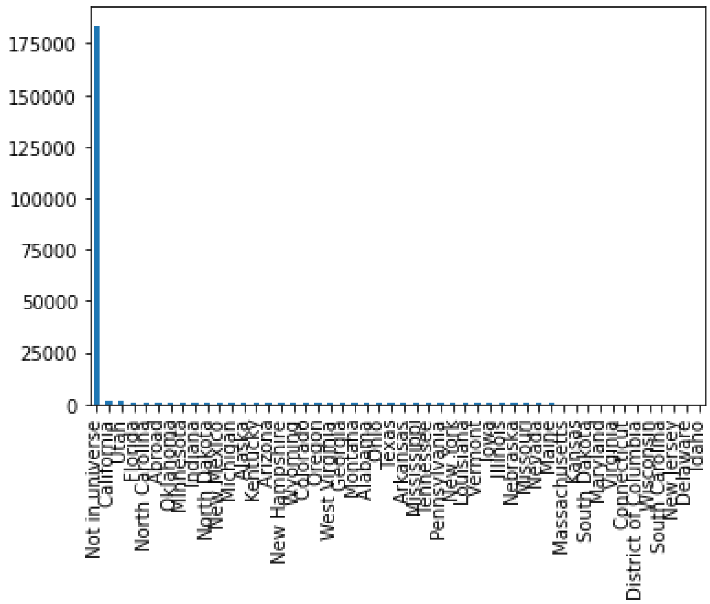
The histogram for column FILESTAT



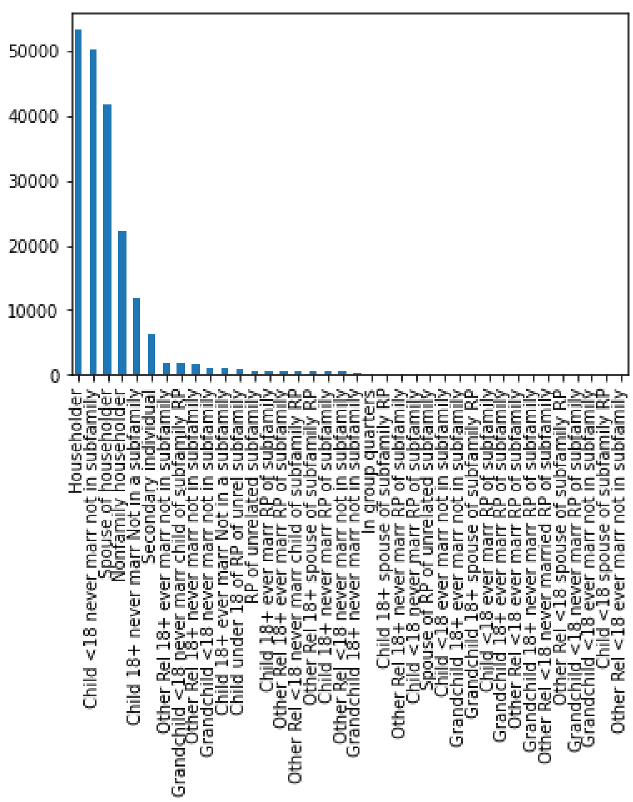
The histogram for column GRINREG



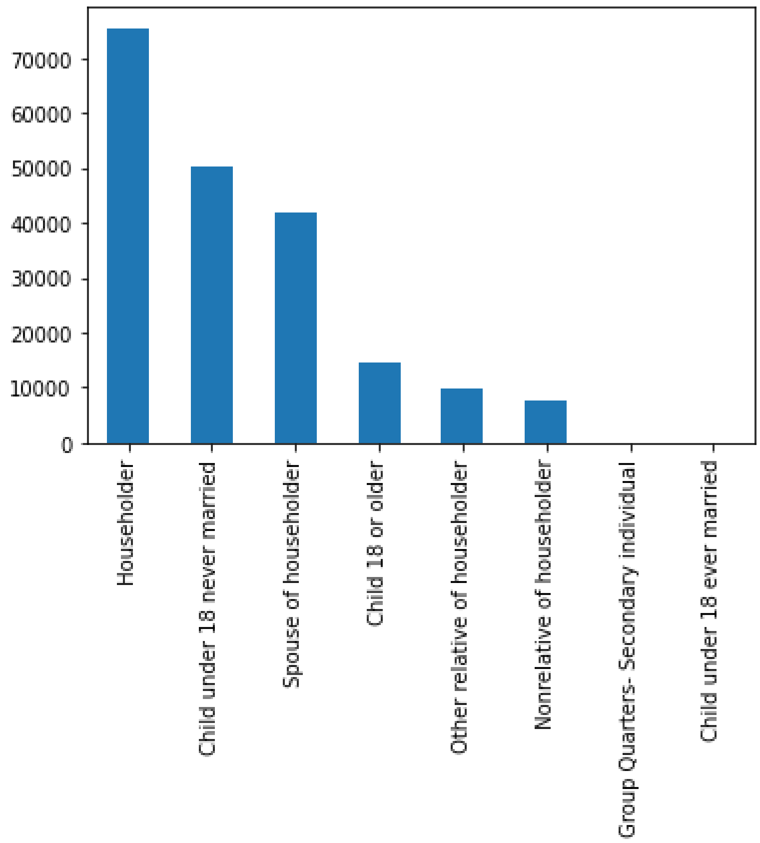
The histogram for column GRINST



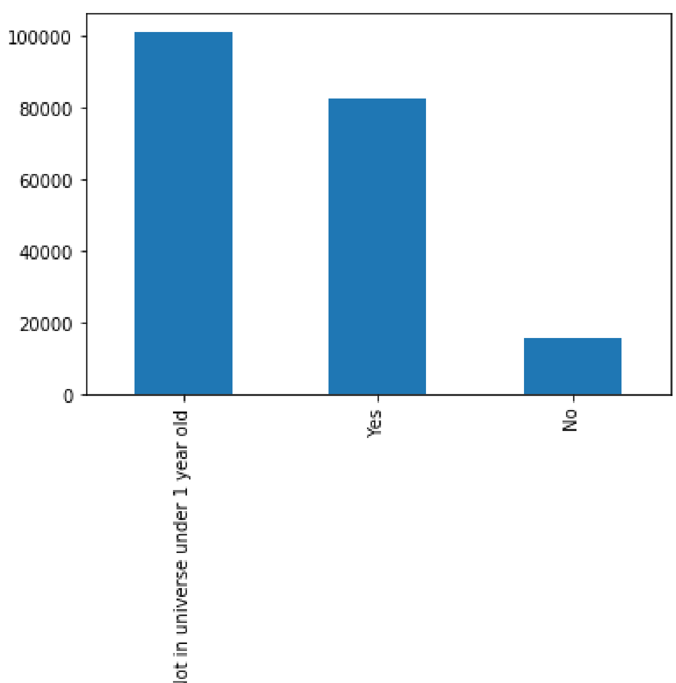
The histogram for column HHDFMX



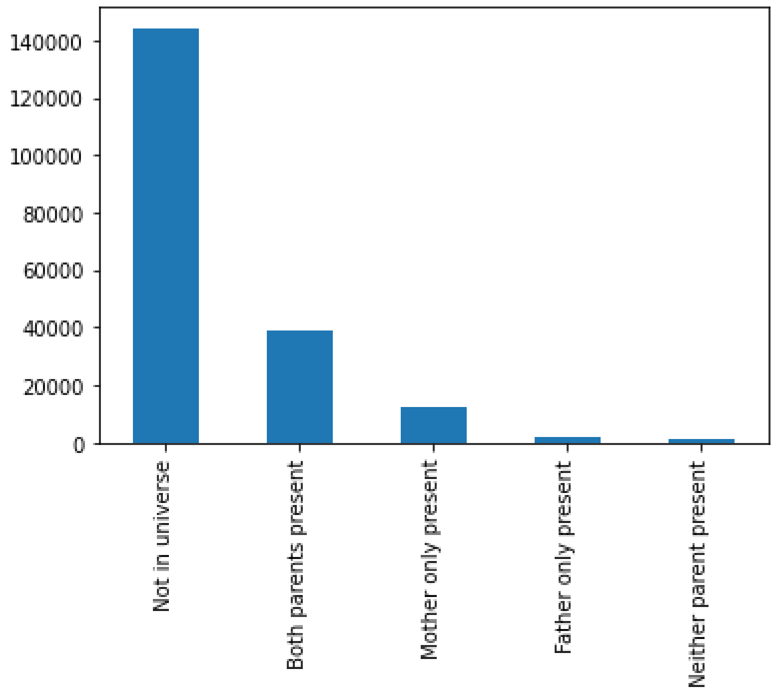
The histogram for column HHDREL



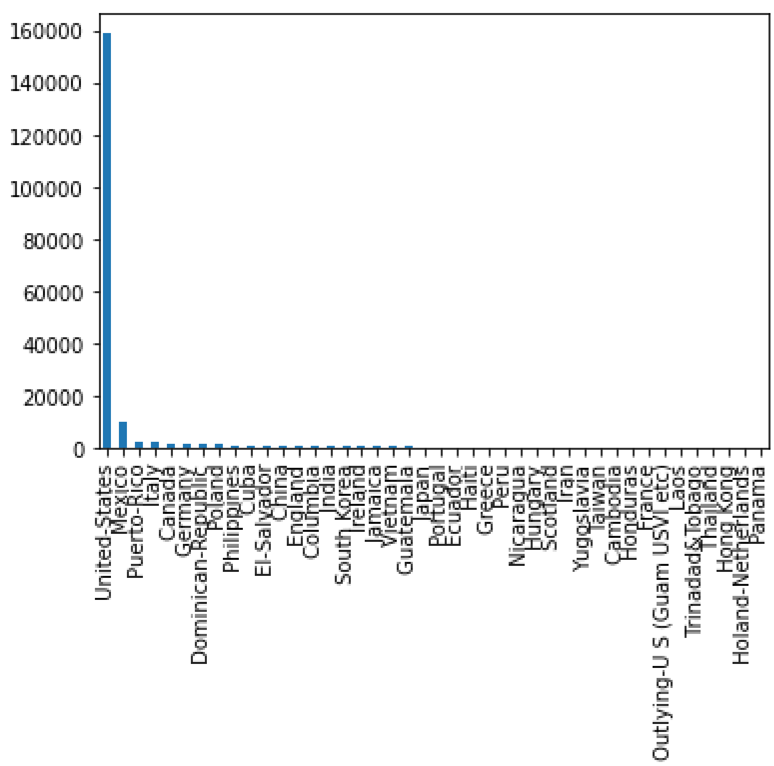
The histogram for column MIGSAME



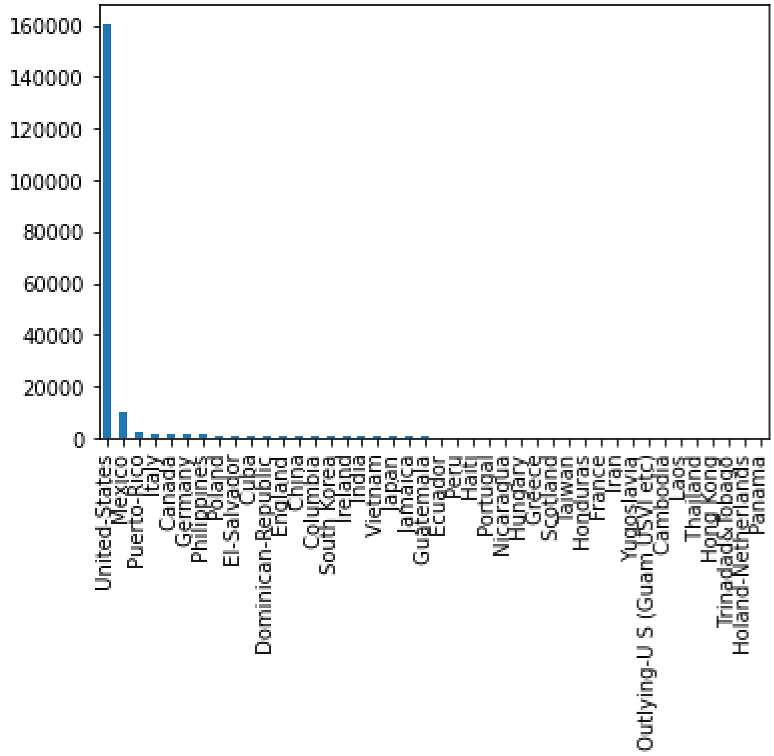
The histogram for column PARENT



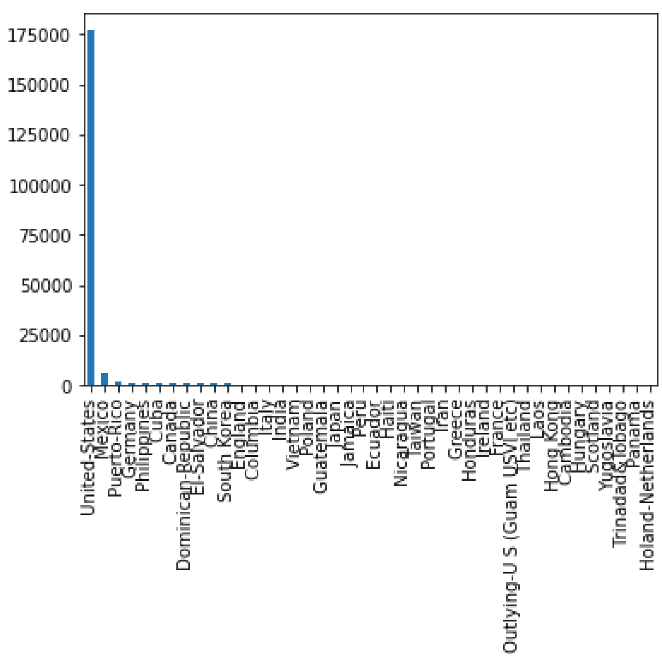
The histogram for column PEFNTVTY



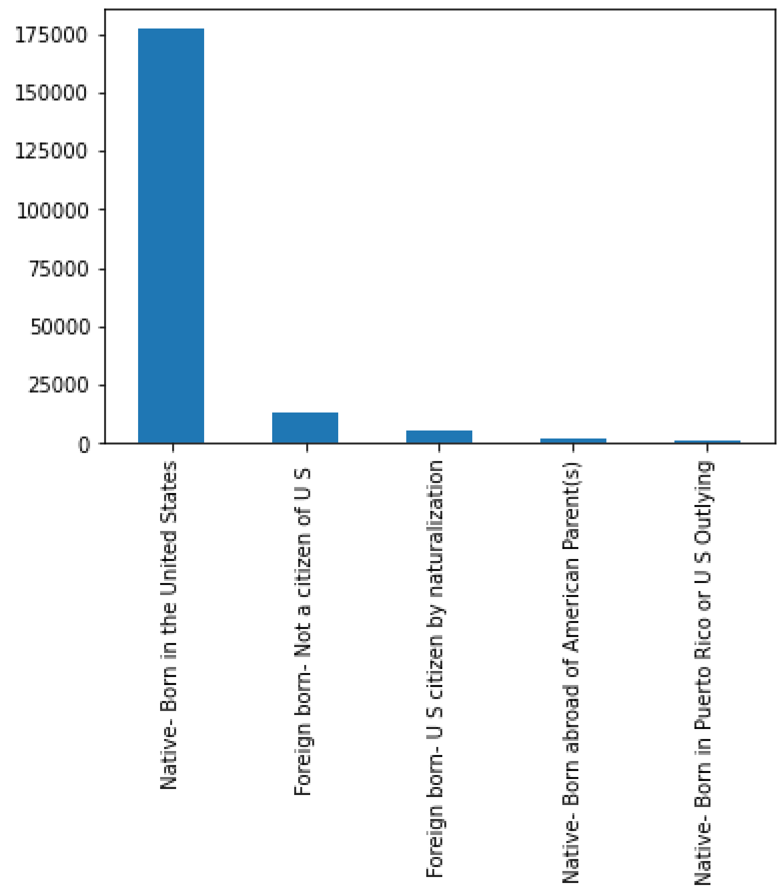
The histogram for column PEMNTVTY



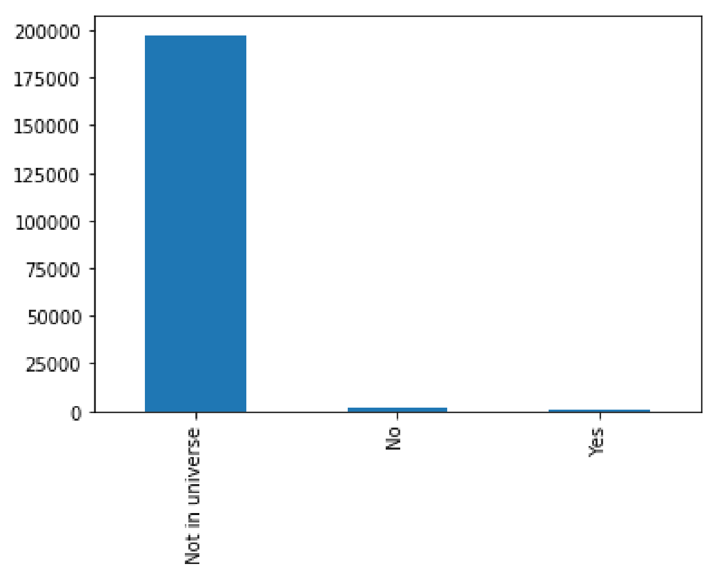
The histogram for column PENATVTY



The histogram for column PRCITSHP



The histogram for column VETQVA



2.2 The columns to drop: VETQVA, PRCITSHP, PENATVTY, PEMNTVTY, PEFNTVTY, GRINST, GRINREG, AUNTYPE, AUNMEM, AREORGN, ARACE, AMJOC, AMJIND, AHSCOL, ADTIND, ADTOCC, AHRSPAY, CAPGAIN, CAPLOSS, DIVYAL, NOEMP, SEOTR, YEAR

3. Imputation, Bucketization, One-Hot Encoding

3.1 Making a dictionary to store the mode value and replacing the nan values with fillna().

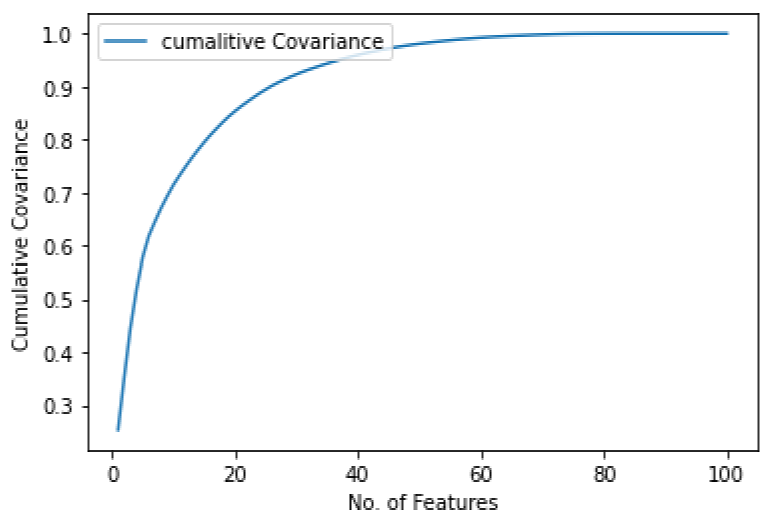
` 3.2 Making bins and categorizing the only left numerical feature using searchsorted().

3.3 One hot encoding by using the column name as suffix and pandas.

	0	1	2	3	4	5	6	7	8	9	Federal government	Local government	Never worked	Not in universe_PARENT	Private	Self-employed-incorporated	Self-employed-not incorporated	State government	Without pay	10th grade	11th grade	12th grade no diploma	gr
0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	
2	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	
3	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
4	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	

5 rows × 113 columns

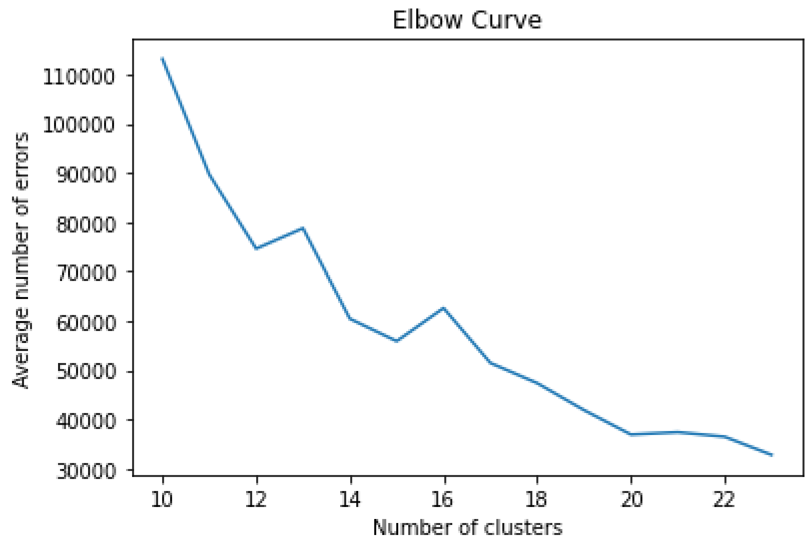
3.2 PCA



Plotting against the number of features the cumulative covariance explained. Taking threshold = 0.85. We go with 20 features.

4. <u>Clustering</u>

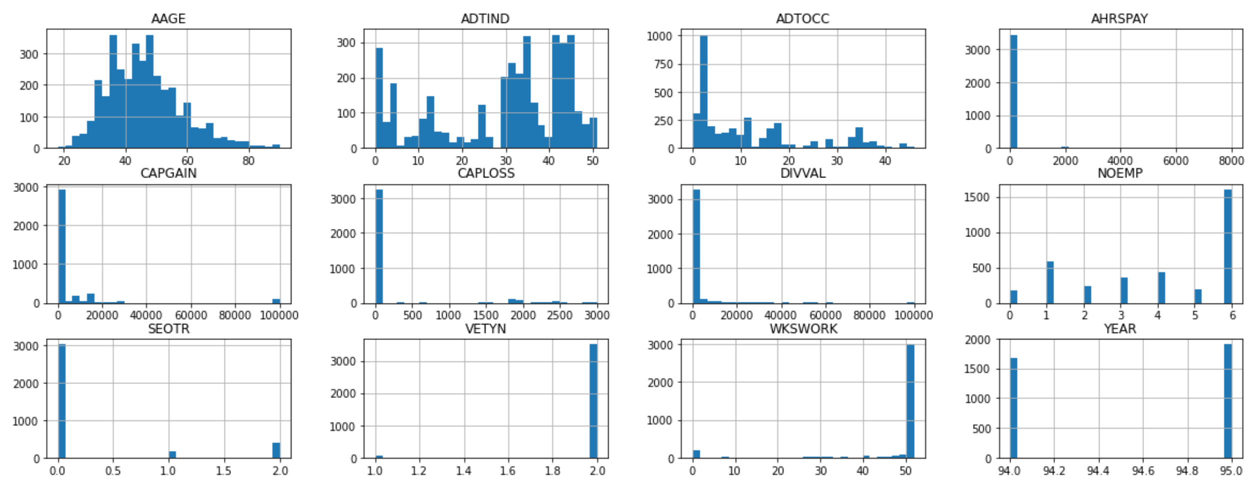
4. 1 Making the elbow curve to choose the value of K.



4.2 The value is chosen as K = 20 as we can see that the average distance becomes nearly constant after it reaches 20. Before 20, it is dropping at a very fast rate.

4.3 Applied K-median clustering for the value.

5. Handling more_than_50K data
5.1 Removing the columns with more than 40% data that are "MIGMTR1, MIGSUN, MIGMTR4, MIGMTR3".
5.2 Plotting the column histogram



The rest of the features can be seen in the notebook for brevity.

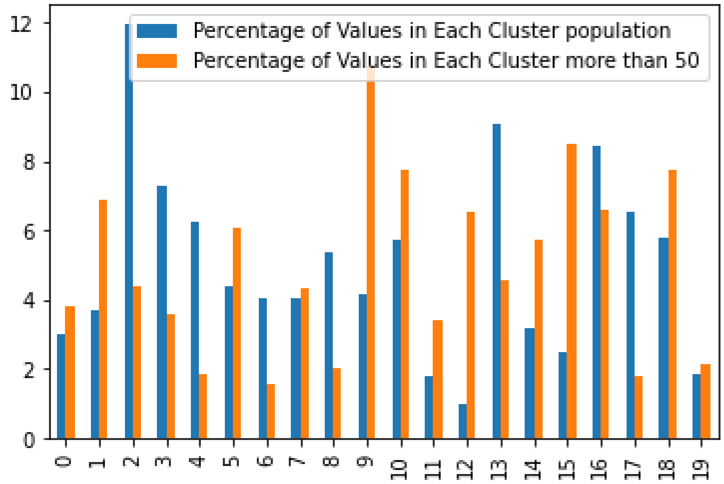
We do not delete any new column but retain *ADTIND and ADTOCC* as they show differences from previous graphs.

5.3 All the other steps are repeated. The same number of columns for PCA and clusters is retained for uniformity.

6. <u>Analysis</u>

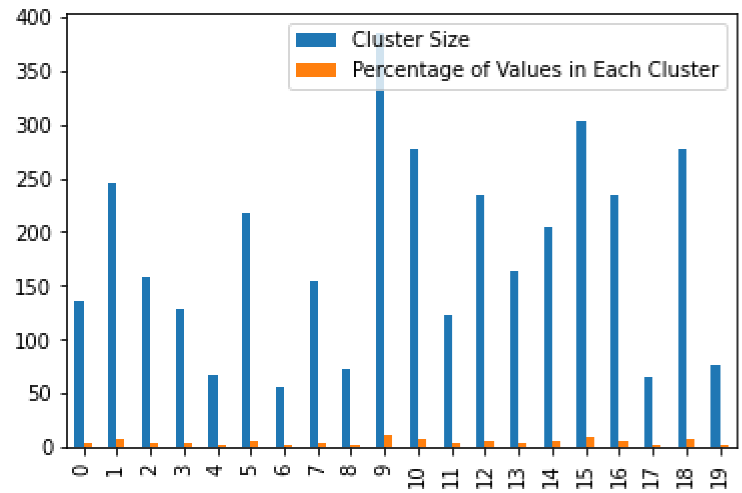
6.1 We first find out the percentage of data in each cluster for both datasets.
The clusters 15, 13, 12, 9, 4, and 2 show the maximum difference between the size of clusters. We can deduce that these clusters should include population dynamics that do not classify for the more than 50K income and hence the difference in the cluster adaption.

Plot the percentage difference between them.

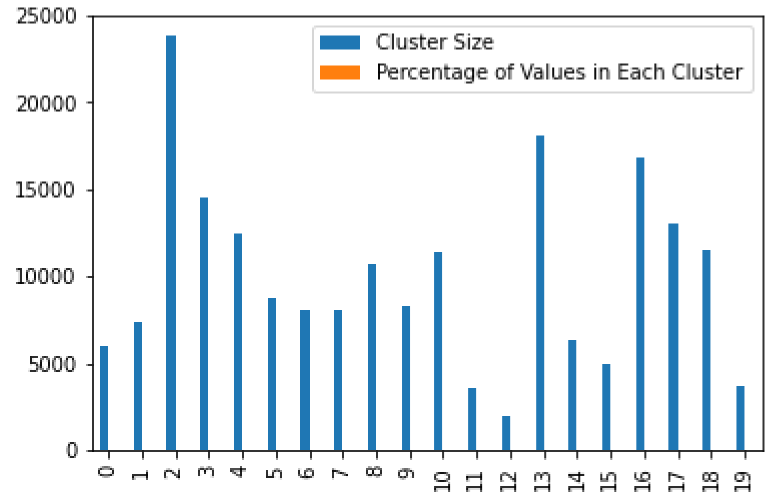


Here the differences are apparent.

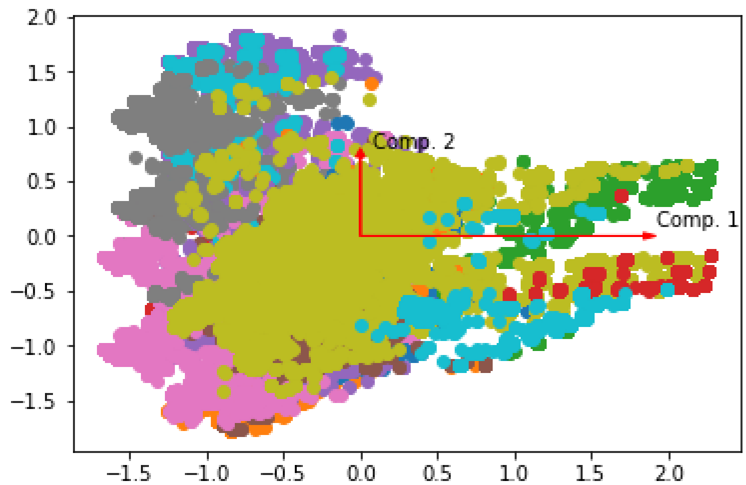
6.2 The clusters overrepresented in population data are 2 and 13. The clusters overrepresented in the more than 50K data are 9 and 15. The reason can be understood by the differences in the features. That is by analyzing the problem as a classification problem, we can understand the underlying difference between the classes. The same is being captured by the clusters.
For more than 50,



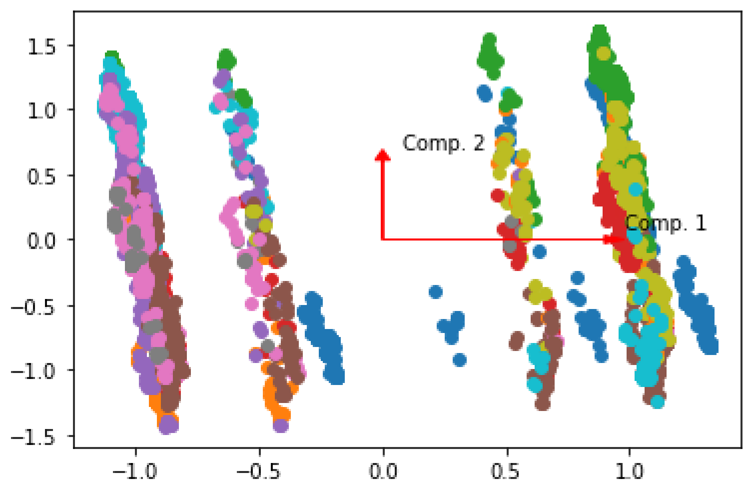
For population,



6.3 and 6.4 Plotting the clusters.
For population data.

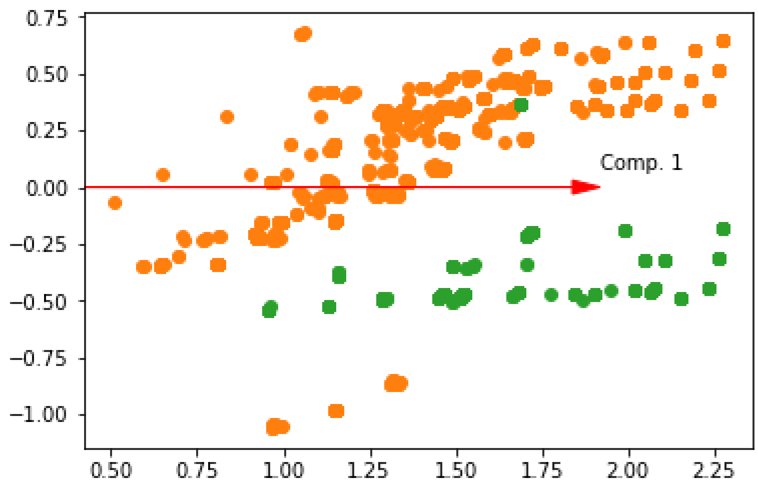


For more than 50K data,

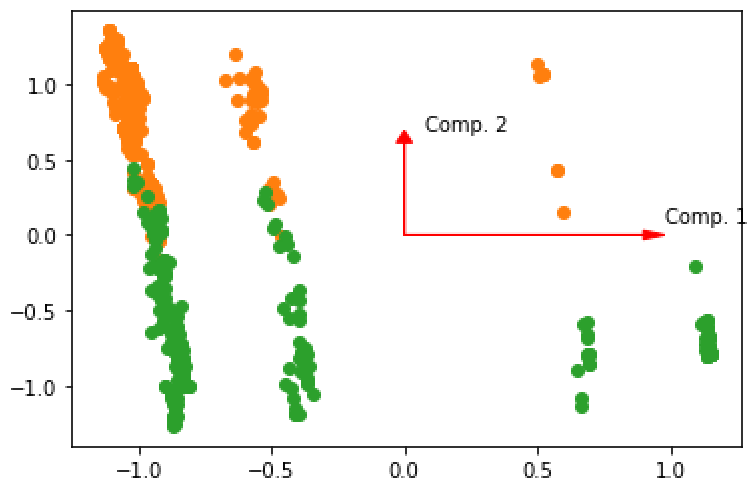


Though the difference is not as apparent in the population data, it comes the way because of reducing the dimensions. In higher dimensions, the data should be divided. More than 50K shows a better picture which can be attributed to the difference in classes and better feature distribution.
I also plot the principal components to notice the maximum covariance.

Next, I plot the overrepresented clusters from both datasets separately.
The populdation data.

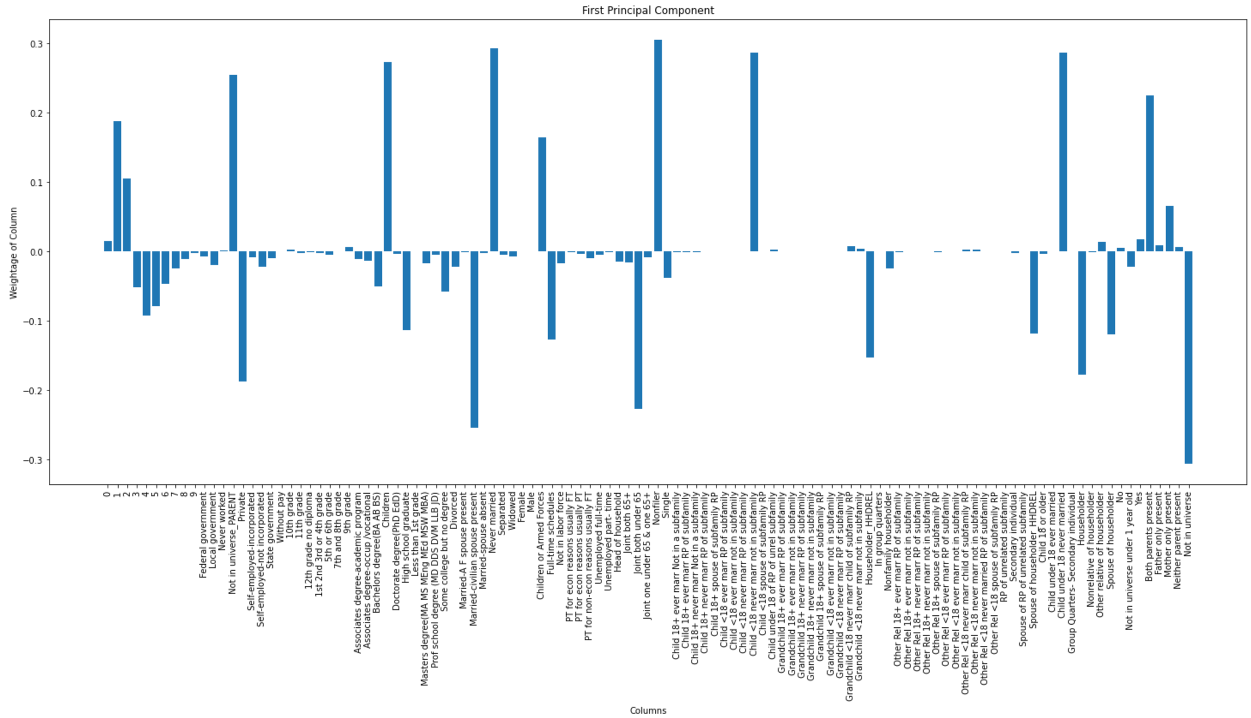


The more than 50K data.

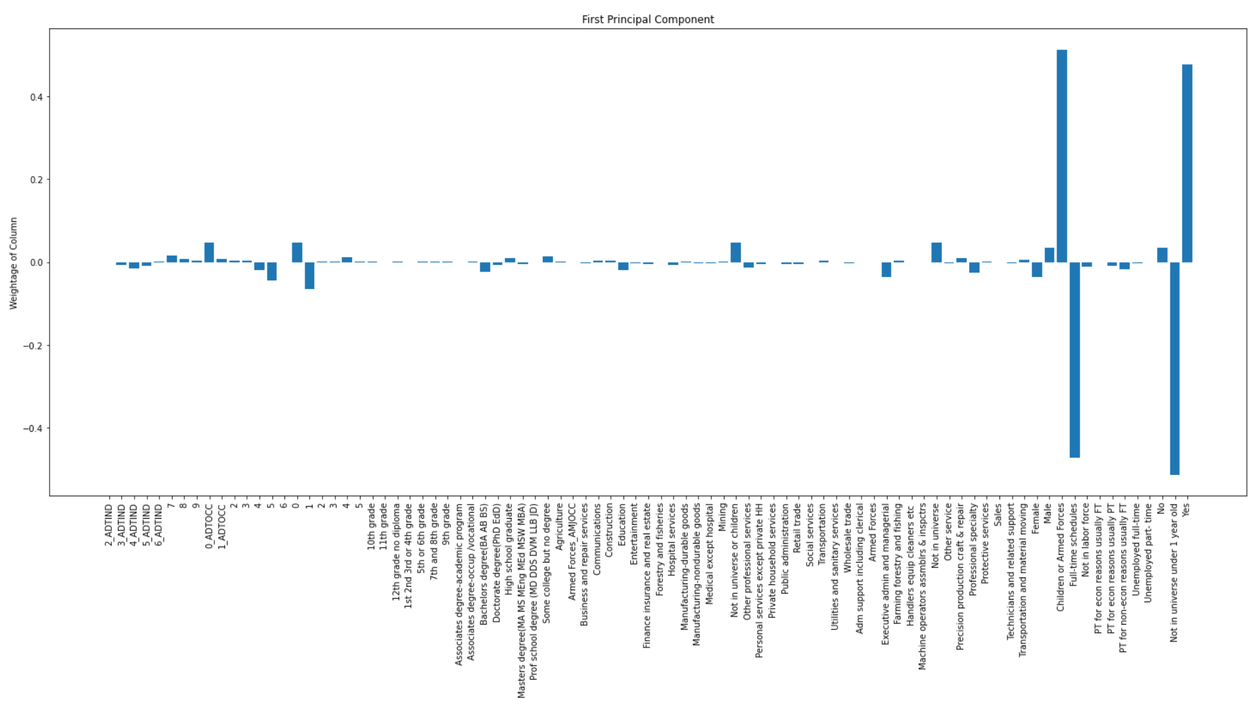


Here we can see the overrepresented classes clearly.

I also plot the first principal component with columns for both datasets.
Population Data.



More than 50K data.



Now, take the values of the overrepresented cluster for the top three columns.
The columns for the population data are: ' Nonfiler'), ' Never married'), ' Child under 18 never married')

The value of the medians for these three columns are
For cluster 2.
1.0077403067690933
1.0365323285011432
1.0500213369286593

For cluster 13
1.0069728875220072
0.9982745913398192
1.0225425792756258

We notice a high value for three columns. This signifies that the population in the clusters should either be in their mid 30-50. Thus the children but not married. Or unmarried youths which form the majority of the population.

The columns for the more than 50K data are: ' Children or Armed Forces, ' Yes', '0_ADTOCC')]
The columns being the job code being 0, migration prev res in sunbelt, and full or part time employment stat.

The values for cluster
For cluster 9,
-0.014120579118281 31
-0.025812822865021845
-0.019722690940121468

For cluster 15
-0.011825265225512499
-0.04225052747874741
0.015147865638000968

The negative value states that they work full time and have not migrated. Thus the presence in the more than 50K data. It also classifies their job as service, negative for 9 being that they are not in service while in 15 positive means that they work in service.

Hence the clusters are making the patterns correctly.

Assignment 3
Manvi Goel
2019472

Ques 2.

$$\min_{w,b,\epsilon} \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^{m} \epsilon^2$$

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

a). let us assume that there exists a possible solution with $\epsilon < 0$.

We have $\epsilon < 0$
$\Rightarrow -\epsilon > 0$.
$\Rightarrow 1 - \epsilon > 1$

This would imply that
$y^{(i)}(w^T x^{(i)} + b) \geq 1 - \epsilon_i$
will be valid for $\epsilon_i = 0$.
This means that the objective would be lower.
Thus this would not be an optimal solution.

Hence, we can remove the constraint without affecting the solution.

(b.) The formula is $L(x, y, \alpha) = f(x,y) - \alpha\, g(x,y)$.

Lagrangian for the problem is

$$L(w, b, \epsilon, \alpha) = \frac{1}{2} w^T w + \frac{C}{2}\sum_{i=1}^{m} \epsilon_i^2 - \sum_{i=1}^{m} \alpha_i [y^{(i)} (w^T x^{(i)} + b) - 1 + \epsilon_i]$$

where $\alpha_i \geq 0$ for $i = 1, 2 \dots m$.

and primal is $J(w, b, \alpha) = \frac{1}{2} w^T w + \frac{C}{2}\sum_{i=1} \epsilon^2 - \alpha [y^{(i)} (w^T x^{(i)} + b$

c). The objective function for the dual is

$$W(\alpha) = \min_{w, b, \epsilon} L(w, b, \epsilon, \alpha)$$

$$= \frac{1}{2} \sum_{i=1}^{m}\sum_{j=1}^{m} (\alpha_i y^{(i)} x^{(i)})^T (\alpha_j y^{(j)} x^{(j)}) + \frac{1}{2}\sum_{i=1}^{m} \frac{\alpha_i}{\epsilon_i} \epsilon_i^2$$
$$- \sum_{i=1}^{m} \alpha_i \left[y^{(i)} \left(\left(\sum_{j=1}^{m} \alpha_j y^{(j)} x^{(j)} \right)^T x^{(i)} + b \right) - 1 + \epsilon_i \right]$$

$$= -\frac{1}{2} \sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)} + \frac{1}{2}\sum_{i=1}^{m} \alpha_i \epsilon_i$$
$$- \left(\sum_{i=1}^{m} \alpha_i y^{(i)}\right) b + \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \alpha_i \epsilon_i.$$

$$= \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)} - \frac{1}{2}\sum_{i=1}^{m} \alpha_i \epsilon_i$$

$$= \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)} - \frac{1}{2}\sum_{i=1}^{m} \frac{\alpha_i^2}{C}$$

So, the dual formation is

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)} - \frac{1}{2} \sum_{i=1}^{m} \frac{\alpha_i^2}{C}$$

such that $\alpha_i \geq 0, \quad i = 1, \ldots, m$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0.$$

Ans 3:

$$K(x,z) = \exp(-|x-z|^2/\tau^2)$$

a). let us set $\alpha_i = 1$ for all $i = 1, \ldots, m$ and $b = 0$.
let us take an example from training data $\{x^{(i)}, y^{(i)}\}$

then
$$|f(x^{(i)}) - y^{(i)}| = \left| \sum_{j=1}^{m} y^{(j)} K(x^{(j)}, x^{(i)}) - y^{(i)} \right|$$
$$= \left| \sum_{j=1}^{m} y^{(j)} \exp(-\|x^{(j)} - x^{(i)}\|^2/\tau^2) - y^{(i)} \right|$$
$$= \left| y^{(i)} + \sum_{j \neq i} y^{(j)} \exp(\|x^{(j)} - x^{(i)}\|^2/\tau^2) - y^{(i)} \right|$$
$$= \left| \sum_{j \neq i} y^{(j)} \exp(-\|x^{(j)} - x^{(i)}\|^2/\tau^2) \right|$$
$$\leq \sum_{j \neq i} \left| y^{(j)} \exp(-\|x^{(j)} - x^{(i)}\|^2/\tau^2) \right|$$
$$= \sum_{j \neq i} |y^{(j)}| \exp(\|x^{(j)} - x^{(i)}\|^2/\tau^2)$$
$$= \sum_{j \neq i} \exp(-\|x^{(j)} - x^{(i)}\|^2/\tau^2).$$
$$\leq \exp(-\epsilon^2/\tau^2)$$
$$= (m-1) \exp(-\epsilon^2/\tau^2).$$

We are considering the first inequlity using the triangle inequality and second from assuming $||x^{(j)} - x^{(i)}|| \geq \epsilon$ for all $i \neq j$.

If we choose γ such that

$$(m-1)\exp(-\epsilon^2/z^2) < 1$$

or

$$z < \frac{\epsilon}{\sqrt{\log(m-1)}}$$

One value can be $z = \frac{\epsilon}{\sqrt{\log m}}$.

Thus we can use these values to train the kernel correctly.

b).

The classifier will obtain zero training error. The SVM that does not use slack variables will get zero training error if it finds a solution.

We need to show the existance of atleast one such feasible point.

let us take the constraint $y^{(i)}(w^T x^{(i)} + b)$ for some i.

let $b = 0$, then

$$y^{(i)}(w^T x^{(i)} + b) = y^{(i)} \cdot f(x^{(i)}) > 0$$

As ~~Now~~, $f(x^{(i)})$ and $y^{(i)}$ have the same sign.

Thus, by choosing any α_i's large enough,

$$y^{(i)}(w^T x^{(i)} + b) > 1.$$

This will make optimization feasible.

Scanned with CamScanner