ML Assignment 3 Report

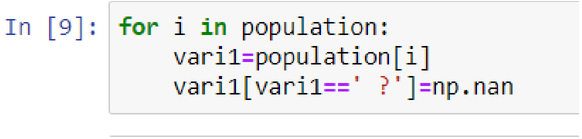
Roll No: 2019462

Name: Bhavya Narang

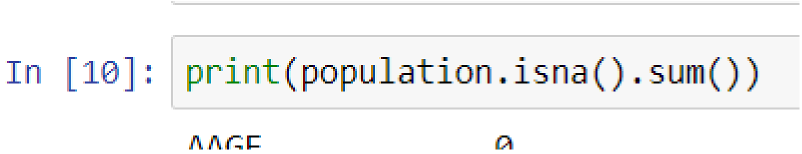
Ques 1)

Part 1)

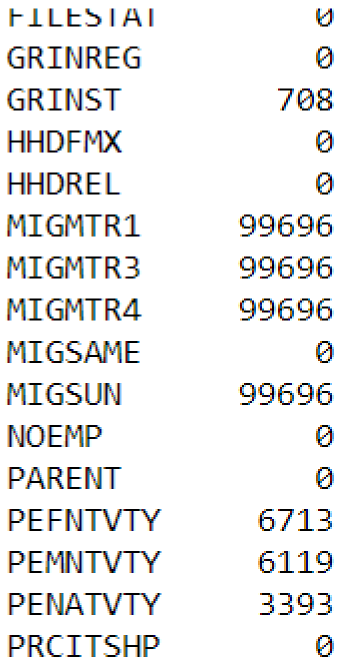
Missing data:



```
In [9]: for i in population:
            vari1=population[i]
            vari1[vari1==' ?']=np.nan
```

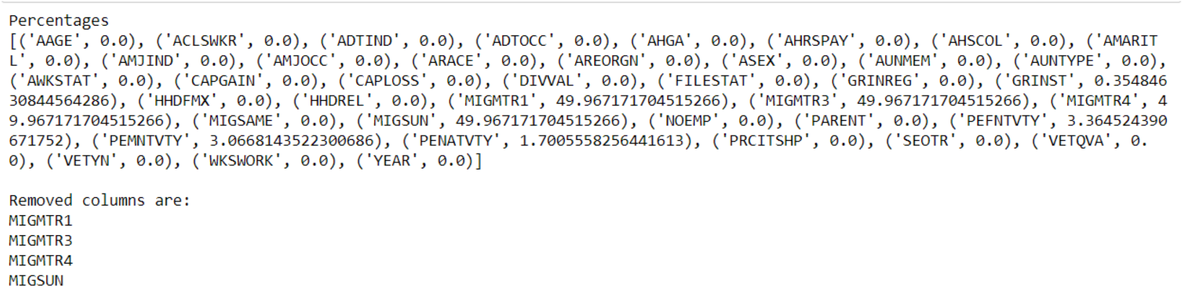


```
In [10]: print(population.isna().sum())

          AAGE              0
```



```
FILESTAT          0
GRINREG           0
GRINST          708
HHDFMX            0
HHDREL            0
MIGMTR1       99696
MIGMTR3       99696
MIGMTR4       99696
MIGSAME           0
MIGSUN        99696
NOEMP             0
PARENT            0
PEFNTVTY       6713
PEMNTVTY       6119
PENATVTY       3393
PRCITSHP          0
```

Columns with percentage of missing data and the removed columns:

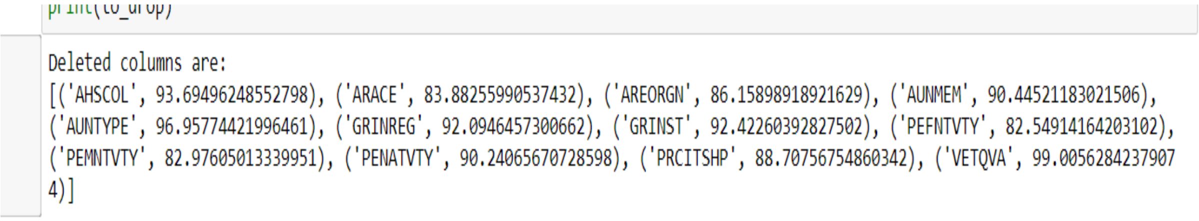


```
Percentages
[('AAGE', 0.0), ('ACLSWKR', 0.0), ('ADTIND', 0.0), ('ADTOCC', 0.0), ('AHGA', 0.0), ('AHRSPAY', 0.0), ('AHSCOL', 0.0), ('AMARIT
L', 0.0), ('AMJIND', 0.0), ('AMJOCC', 0.0), ('ARACE', 0.0), ('AREORGN', 0.0), ('ASEX', 0.0), ('AUNMEM', 0.0), ('AUNTYPE', 0.0),
('AWKSTAT', 0.0), ('CAPGAIN', 0.0), ('CAPLOSS', 0.0), ('DIVVAL', 0.0), ('FILESTAT', 0.0), ('GRINREG', 0.0), ('GRINST', 0.354846
30844564286), ('HHDFMX', 0.0), ('HHDREL', 0.0), ('MIGMTR1', 49.967171704515266), ('MIGMTR3', 49.967171704515266), ('MIGMTR4', 4
9.967171704515266), ('MIGSAME', 0.0), ('MIGSUN', 49.967171704515266), ('NOEMP', 0.0), ('PARENT', 0.0), ('PEFNTVTY', 3.364524390
671752), ('PEMNTVTY', 3.0668143522300686), ('PENATVTY', 1.7005558256441613), ('PRCITSHP', 0.0), ('SEOTR', 0.0), ('VETQVA', 0.
0), ('VETYN', 0.0), ('WKSWORK', 0.0), ('YEAR', 0.0)]

Removed columns are:
MIGMTR1
MIGMTR3
MIGMTR4
MIGSUN
```

Part 2)

Plots are in the ipynb notebook, not uploading screenshot here as there are 36 columns.

Deleted features with most data(criteria: greater than 80%) in one column with their percentages:

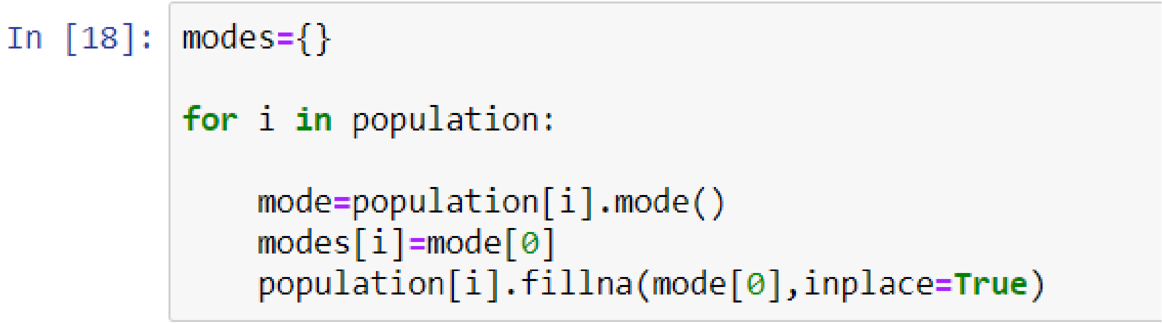


```
print(to_drop)
```

```
Deleted columns are:
[('AHSCOL', 93.69496248552798), ('ARACE', 83.88255990537432), ('AREORGN', 86.15898918921629), ('AUNMEM', 90.44521183021506),
('AUNTYPE', 96.95774421996461), ('GRINREG', 92.0946457300662), ('GRINST', 92.42260392827502), ('PEFNTVTY', 82.54914164203102),
('PEMNTVTY', 82.97605013339951), ('PENATVTY', 90.24065670728598), ('PRCITSHP', 88.70756754860342), ('VETQVA', 99.0056284237907
4)]
```

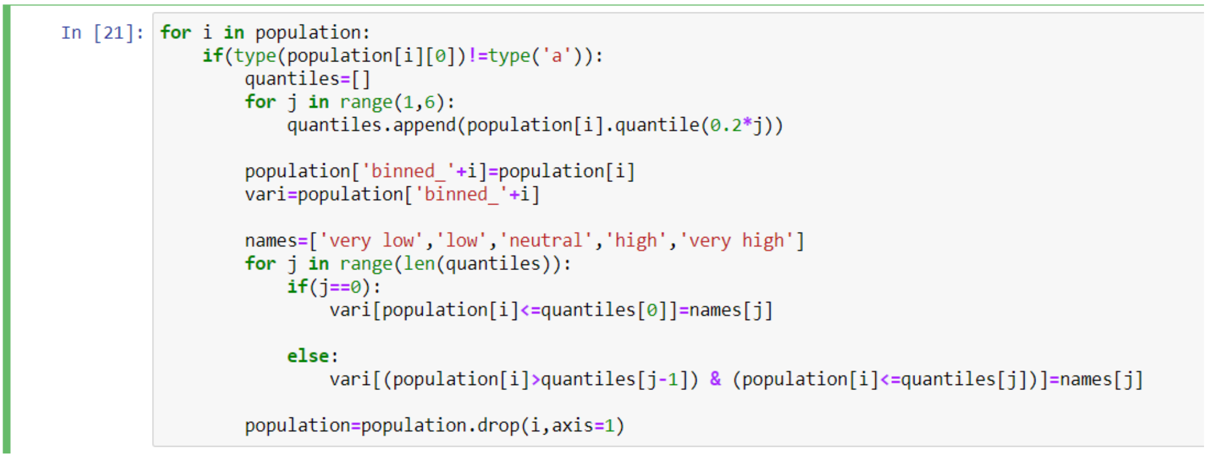
Part 3)

Replacing missing data with mode and storing the modes for later use:



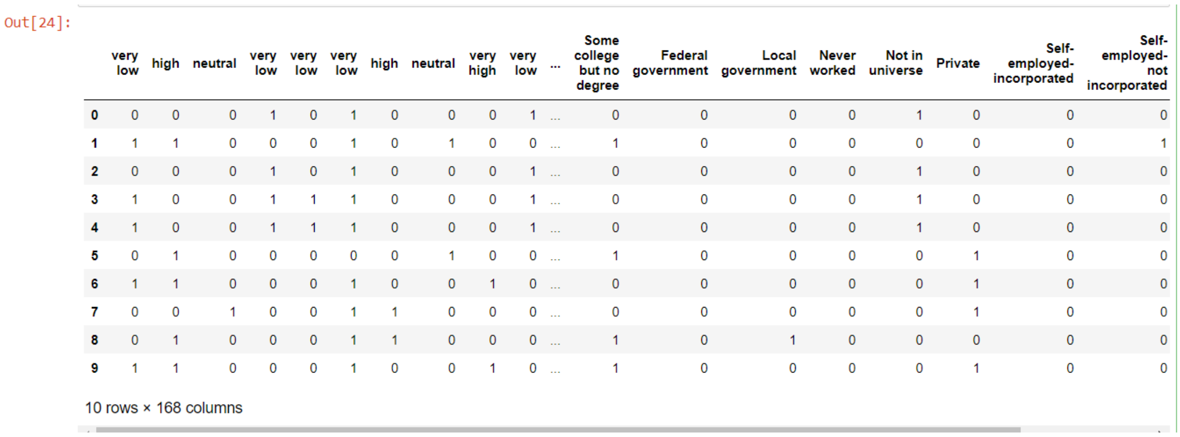
```
In [18]: modes={}

         for i in population:

             mode=population[i].mode()
             modes[i]=mode[0]
             population[i].fillna(mode[0],inplace=True)
```

Bucketing numerical features into 5 classes with percentile ranges as : 0-20,20-40,40-60,60-80,80-100 as very low, low, neutral, high, very high.



In [21]:
```
for i in population:
    if(type(population[i][0])!=type('a')):
        quantiles=[]
        for j in range(1,6):
            quantiles.append(population[i].quantile(0.2*j))

        population['binned_'+i]=population[i]
        vari=population['binned_'+i]

        names=['very low','low','neutral','high','very high']
        for j in range(len(quantiles)):
            if(j==0):
                vari[population[i]<=quantiles[0]]=names[j]

            else:
                vari[(population[i]>quantiles[j-1]) & (population[i]<=quantiles[j])]=names[j]

        population=population.drop(i,axis=1)
```

Data after one hot encoding contains 168 columns:



Out[24]:

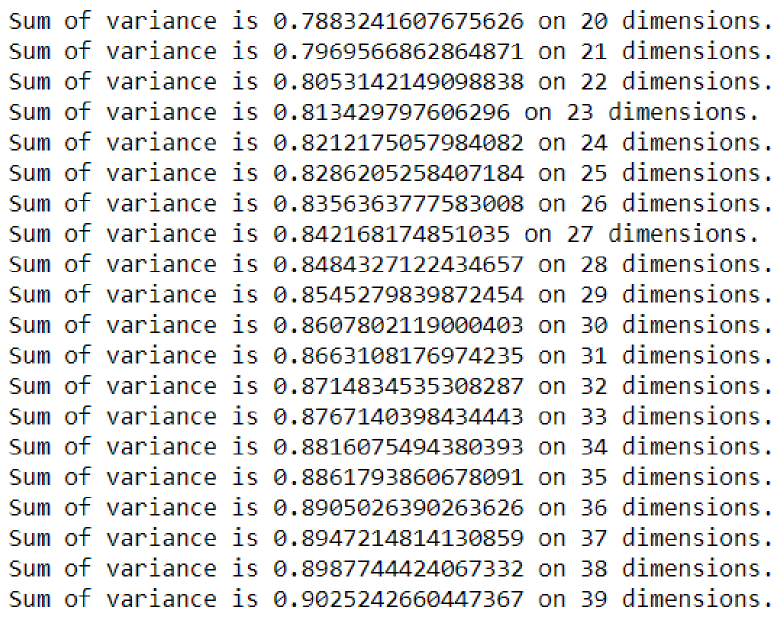
	very low	high	neutral	very low	very low	very low	high	neutral	very high	very low	...	Some college but no degree	Federal government	Local government	Never worked	Not in universe	Private	Self-employed-incorporated	Self-employed-not incorporated
0	0	0	0	1	0	1	0	0	0	1	...	0	0	0	0	1	0	0	0
1	1	1	0	0	0	1	0	1	0	0	...	1	0	0	0	0	0	0	1
2	0	0	0	1	0	1	0	0	0	1	...	0	0	0	0	1	0	0	0
3	1	0	0	1	1	1	0	0	0	1	...	0	0	0	0	1	0	0	0
4	1	0	0	1	1	1	0	0	0	1	...	0	0	0	0	1	0	0	0
5	0	1	0	0	0	0	0	1	0	0	...	1	0	0	0	0	1	0	0
6	1	1	0	0	0	1	0	0	1	0	...	0	0	0	0	0	1	0	0
7	0	0	1	0	0	1	1	0	0	0	...	0	0	0	0	0	1	0	0
8	0	1	0	0	0	1	1	0	0	0	...	1	0	1	0	0	0	0	0
9	1	1	0	0	0	1	0	0	1	0	...	1	0	0	0	0	1	0	0

10 rows × 168 columns

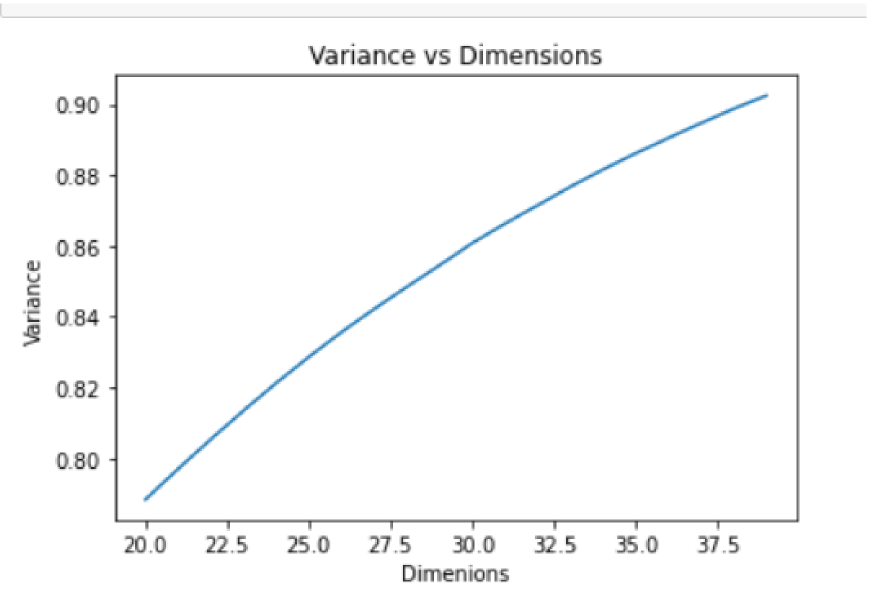
Dimensionality reduction:

As doing PCA is optional and marks won’t concern I am only doing the steps I feel important.

Calculating variance sum with dimensions after PCA:



```
Sum of variance is 0.7883241607675626 on 20 dimensions.
Sum of variance is 0.7969566862864871 on 21 dimensions.
Sum of variance is 0.8053142149098838 on 22 dimensions.
Sum of variance is 0.813429797606296 on 23 dimensions.
Sum of variance is 0.8212175057984082 on 24 dimensions.
Sum of variance is 0.8286205258407184 on 25 dimensions.
Sum of variance is 0.8356363777583008 on 26 dimensions.
Sum of variance is 0.84216817485103 5 on 27 dimensions.
Sum of variance is 0.8484327122434657 on 28 dimensions.
Sum of variance is 0.8545279839872454 on 29 dimensions.
Sum of variance is 0.8607802119000403 on 30 dimensions.
Sum of variance is 0.866310817697423 5 on 31 dimensions.
Sum of variance is 0.8714834535308287 on 32 dimensions.
Sum of variance is 0.8767140398434443 on 33 dimensions.
Sum of variance is 0.8816075494380393 on 34 dimensions.
Sum of variance is 0.8861793860678091 on 35 dimensions.
Sum of variance is 0.8905026390263626 on 36 dimensions.
Sum of variance is 0.8947214814130859 on 37 dimensions.
Sum of variance is 0.8987744424067332 on 38 dimensions.
Sum of variance is 0.9025242660447367 on 39 dimensions.
```

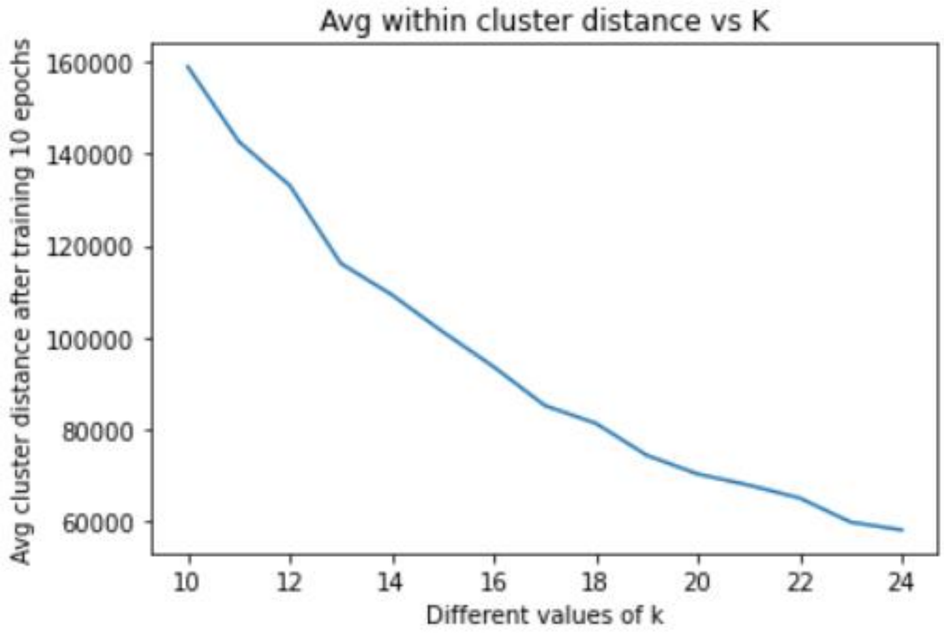


Using 29 dimensions out of 168 to keep variance above 85%.

Part 4) K Median Clustering:

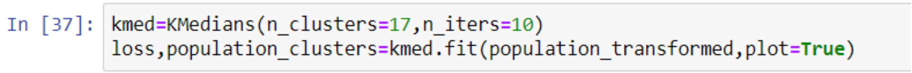
Self-written from scratch, comments in code to explain working. Distance used is Manhattan and not L2 (we can use any, both will give similar results).

Graph :

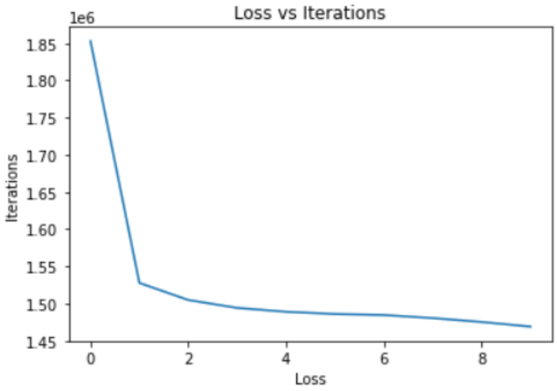


As there is no clear elbow observed in the figure hence using what seems visually to me, hence keeping 17 clusters.

Applying clustering and storing clusters for each data point.



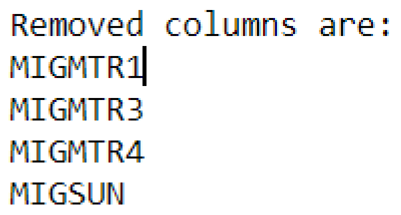
```
In [37]: kmed=KMedians(n_clusters=17,n_iters=10)
         loss,population_clusters=kmed.fit(population_transformed,plot=True)
```



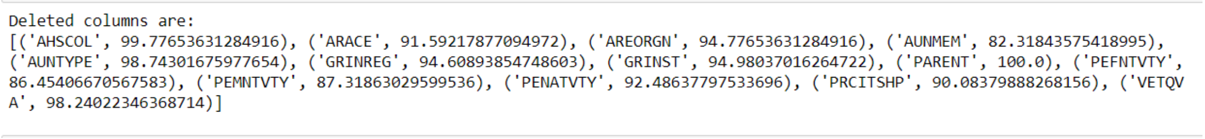
Part 5)

All steps done as above are done on the more than 50k data.

Some distinguished of them are:

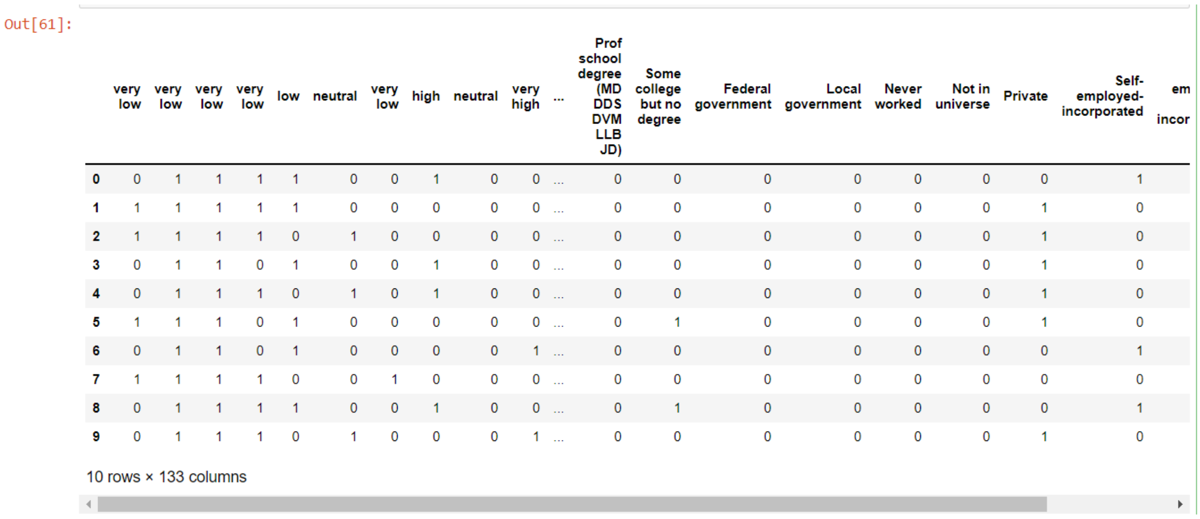


```
Removed columns are:
MIGMTR1
MIGMTR3
MIGMTR4
MIGSUN
```



```
Deleted columns are:
[('AHSCOL', 99.77653631284916), ('ARACE', 91.59217877094972), ('AREORGN', 94.77653631284916), ('AUNMEM', 82.31843575418995),
('AUNTYPE', 98.74301675977654), ('GRINREG', 94.60893854748603), ('GRINST', 94.98037016264722), ('PARENT', 100.0), ('PEFNTVTY',
86.45406670567583), ('PEMNTVTY', 87.31863029599536), ('PENATVTY', 92.48637797533696), ('PRCITSHP', 90.08379888268156), ('VETQV
A', 98.24022346368714)]
```

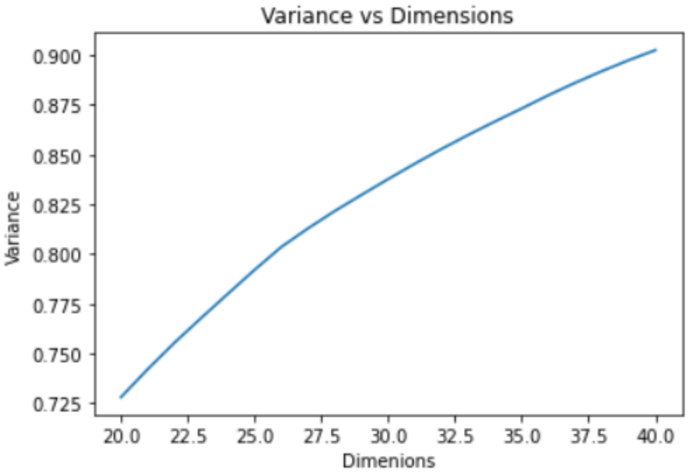
After one hot encoding we have 133 columns:



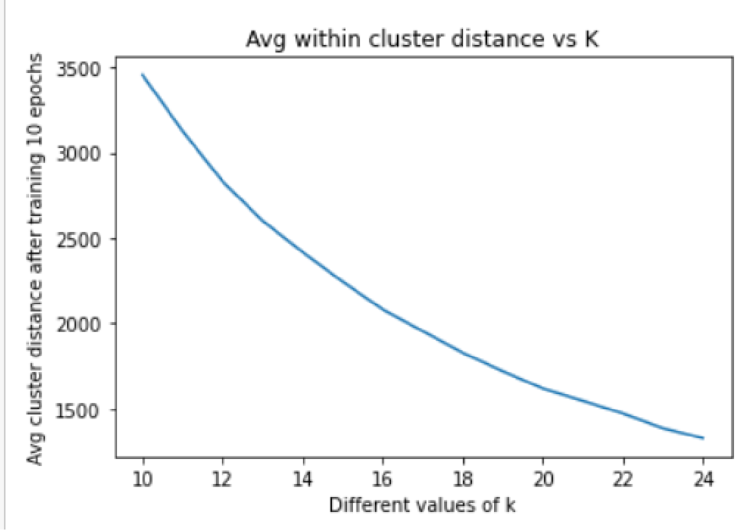
Out[61]:

	very low	very low	very low	very low	low	neutral	very low	high	neutral	very high	...	Prof school degree (MD DDS DVM LLB JD)	Some college but no degree	Federal government	Local government	Never worked	Not in universe	Private	Self-employed-incorporated	em incor
0	0	1	1	1	1	0	0	1	0	0	...	0	0	0	0	0	0	0	1	
1	1	1	1	1	1	0	0	0	0	0	...	0	0	0	0	0	0	1	0	
2	1	1	1	1	0	1	0	0	0	0	...	0	0	0	0	0	0	1	0	
3	0	1	1	0	1	0	0	1	0	0	...	0	0	0	0	0	0	1	0	
4	0	1	1	1	0	1	0	1	0	0	...	0	0	0	0	0	0	1	0	
5	1	1	1	0	1	0	0	0	0	0	...	0	1	0	0	0	0	1	0	
6	0	1	1	0	1	0	0	0	0	1	...	0	0	0	0	0	0	0	1	
7	1	1	1	1	0	0	1	0	0	0	...	0	0	0	0	0	0	0	0	
8	0	1	1	1	1	0	0	1	0	0	...	0	1	0	0	0	0	0	1	
9	0	1	1	1	0	1	0	0	0	1	...	0	0	0	0	0	0	1	0	

10 rows × 133 columns



Clustering graph:

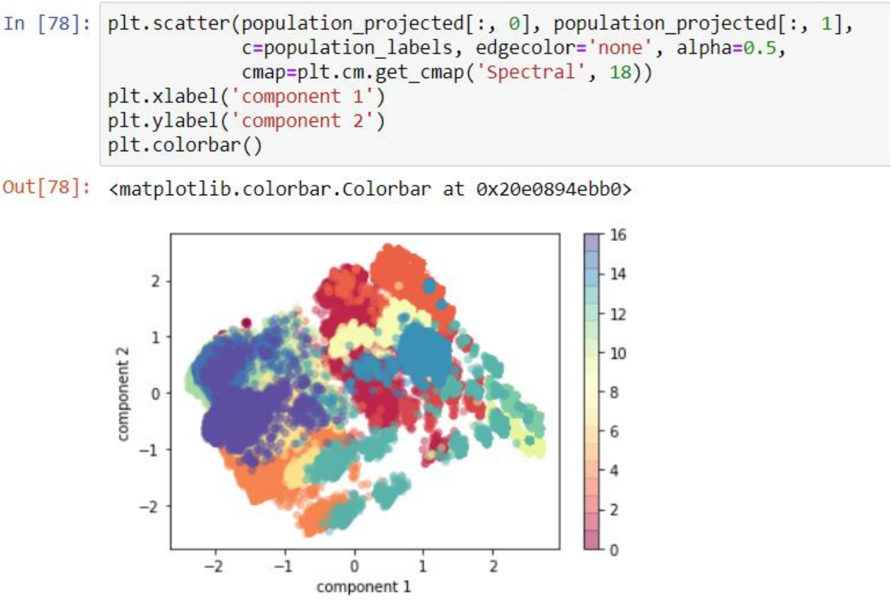


Again as we have no clear elbow taking clusters to be 17.

Part 6)

As we have to visualize the data again applying PCA to keep only 2 dimensions as we can only use 2 to plot irrespective of the variance.

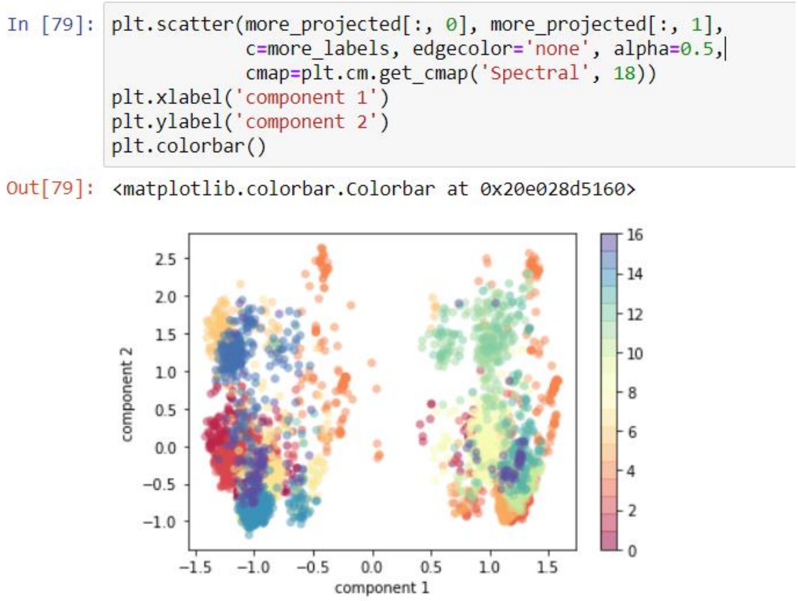
Cluster of general data:



```
In [78]: plt.scatter(population_projected[:, 0], population_projected[:, 1],
                     c=population_labels, edgecolor='none', alpha=0.5,
                     cmap=plt.cm.get_cmap('Spectral', 18))
         plt.xlabel('component 1')
         plt.ylabel('component 2')
         plt.colorbar()
```

```
Out[78]: <matplotlib.colorbar.Colorbar at 0x20e0894ebb0>
```

Cluster of more than 50k data:



```
In [79]: plt.scatter(more_projected[:, 0], more_projected[:, 1],
                     c=more_labels, edgecolor='none', alpha=0.5,
                     cmap=plt.cm.get_cmap('Spectral', 18))
         plt.xlabel('component 1')
         plt.ylabel('component 2')
         plt.colorbar()
```

```
Out[79]: <matplotlib.colorbar.Colorbar at 0x20e028d5160>
```

Observations from the clusters:

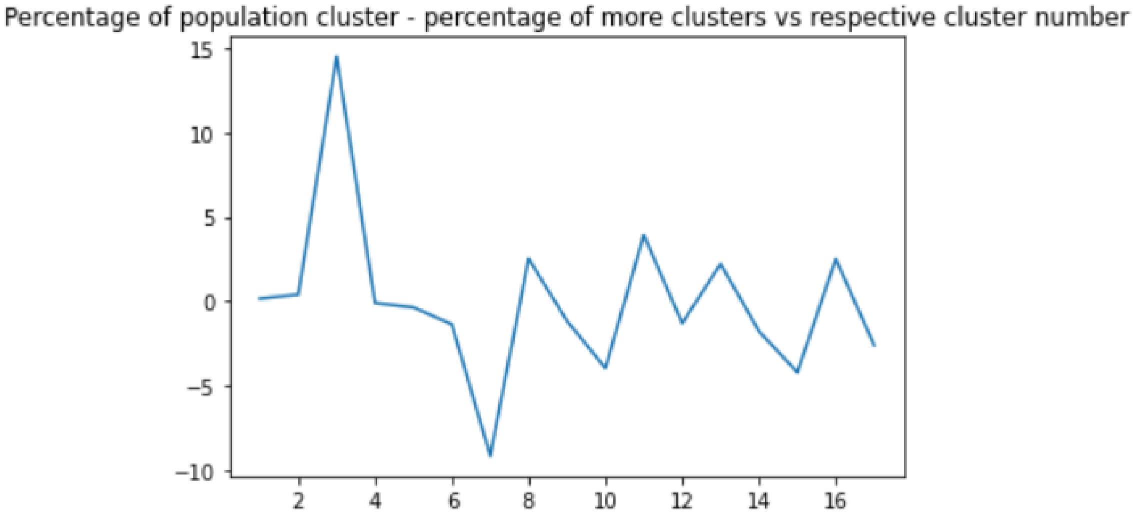
6.1 and 6.2)

By visual analysis:

Cluster 2 seems to be a subset of cluster 1 with data missing in the cental region (less than 50k population) and also 2 is tilted.

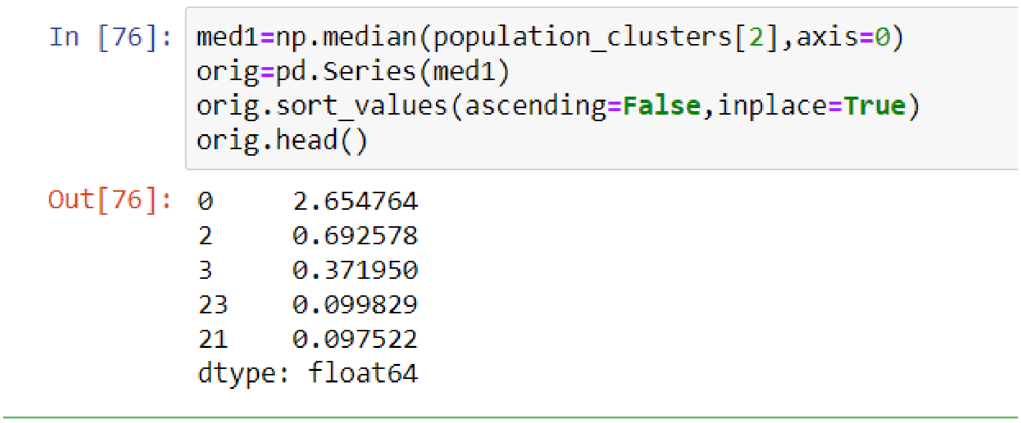
We can observe that cluster 1 has a very high density in centre which 2 does not, but the sides have similar clustering proportions with the data in more than 50k being a bit less.

By data percentages:

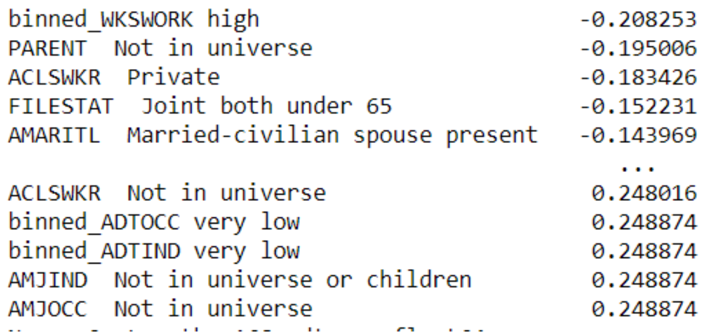


Clearly the 3rd and 7th component denote over represented, in 3rd population cluster is over represented whereas in 7th the more cluster is over represented.

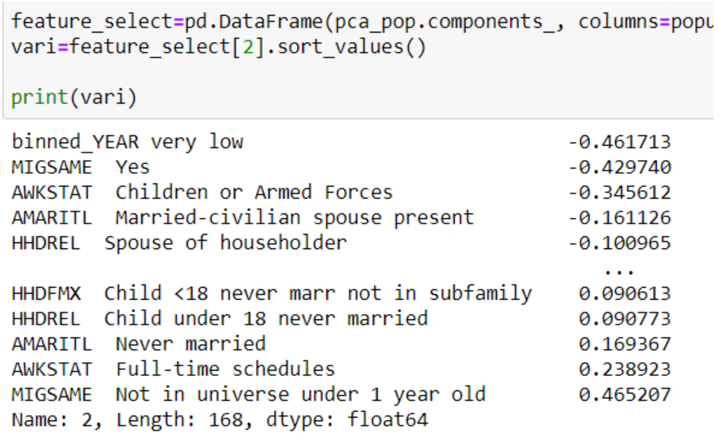
6.3)



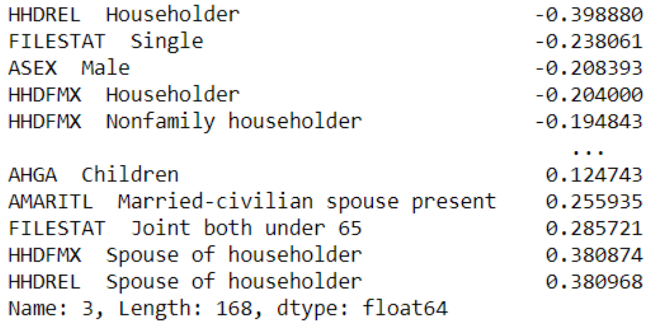
```
In [76]: med1=np.median(population_clusters[2],axis=0)
         orig=pd.Series(med1)
         orig.sort_values(ascending=False,inplace=True)
         orig.head()

Out[76]: 0     2.654764
         2     0.692578
         3     0.371950
         23    0.099829
         21    0.097522
         dtype: float64
```



```
binned_WKSWORK high                        -0.208253
PARENT  Not in universe                    -0.195006
ACLSWKR  Private                           -0.183426
FILESTAT  Joint both under 65              -0.152231
AMARITL  Married-civilian spouse present   -0.143969
                                              ...
ACLSWKR  Not in universe                    0.248016
binned_ADTOCC very low                      0.248874
binned_ADTIND very low                      0.248874
AMJIND  Not in universe or children         0.248874
AMJOCC  Not in universe                     0.248874
```



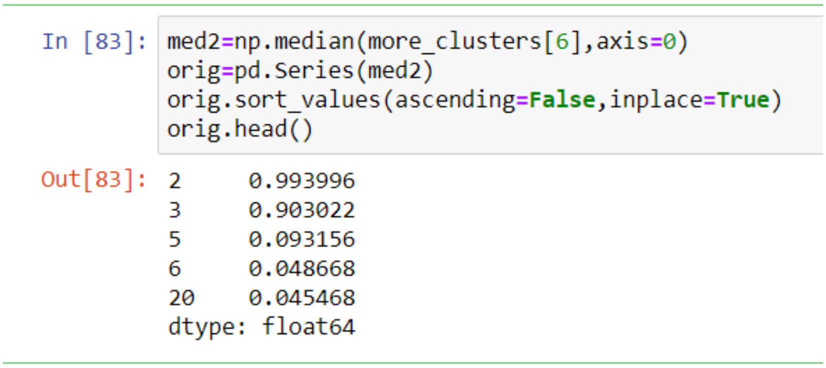
```
feature_select=pd.DataFrame(pca_pop.components_, columns=popu
vari=feature_select[2].sort_values()

print(vari)

binned_YEAR very low                         -0.461713
MIGSAME  Yes                                 -0.429740
AWKSTAT  Children or Armed Forces            -0.345612
AMARITL  Married-civilian spouse present     -0.161126
HHDREL  Spouse of householder                -0.100965
                                                ...
HHDFMX  Child <18 never marr not in subfamily  0.090613
HHDREL  Child under 18 never married          0.090773
AMARITL  Never married                        0.169367
AWKSTAT  Full-time schedules                  0.238923
MIGSAME  Not in universe under 1 year old     0.465207
Name: 2, Length: 168, dtype: float64
```



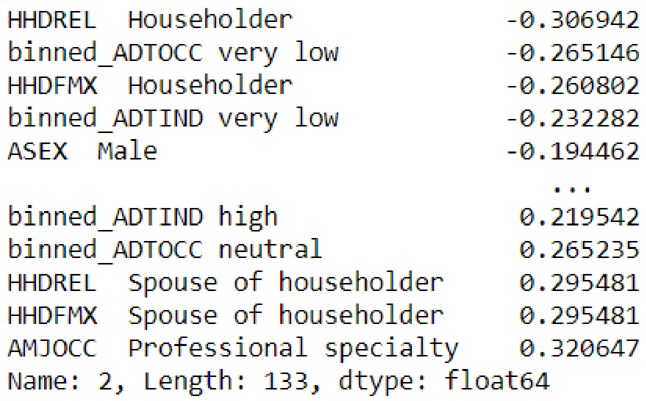
```
HHDREL  Householder                         -0.398880
FILESTAT  Single                            -0.238061
ASEX  Male                                  -0.208393
HHDFMX  Householder                         -0.204000
HHDFMX  Nonfamily householder               -0.194843
                                               ...
AHGA  Children                               0.124743
AMARITL  Married-civilian spouse present     0.255935
FILESTAT  Joint both under 65                0.285721
HHDFMX  Spouse of householder                0.380874
HHDREL  Spouse of householder                0.380968
Name: 3, Length: 168, dtype: float64
```

Columns over represented in top 3 principal features of population data is : WKSWORK -> high, Year-> low, HHDREL-> House holder

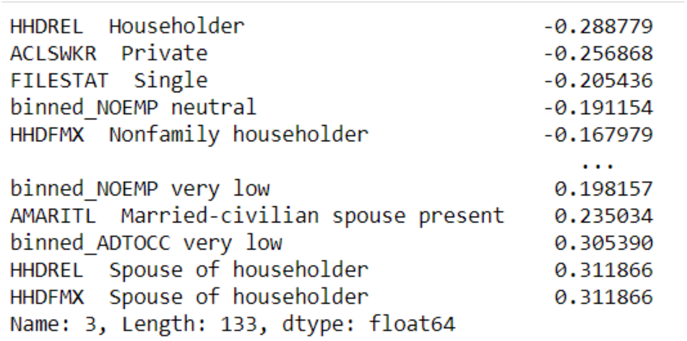
6.4)



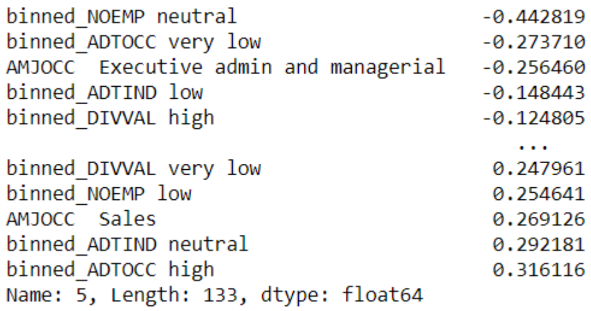
```
In [83]: med2=np.median(more_clusters[6],axis=0)
         orig=pd.Series(med2)
         orig.sort_values(ascending=False,inplace=True)
         orig.head()

Out[83]: 2     0.993996
         3     0.903022
         5     0.093156
         6     0.048668
         20    0.045468
         dtype: float64
```



```
HHDREL   Householder                 -0.306942
binned_ADTOCC very low               -0.265146
HHDFMX   Householder                 -0.260802
binned_ADTIND very low               -0.232282
ASEX   Male                          -0.194462
                                        ...
binned_ADTIND high                    0.219542
binned_ADTOCC neutral                 0.265235
HHDREL   Spouse of householder        0.295481
HHDFMX   Spouse of householder        0.295481
AMJOCC   Professional specialty       0.320647
Name: 2, Length: 133, dtype: float64
```



```
HHDREL   Householder                          -0.288779
ACLSWKR   Private                             -0.256868
FILESTAT   Single                             -0.205436
binned_NOEMP neutral                          -0.191154
HHDFMX   Nonfamily householder                -0.167979
                                                 ...
binned_NOEMP very low                          0.198157
AMARITL   Married-civilian spouse present      0.235034
binned_ADTOCC very low                         0.305390
HHDREL   Spouse of householder                 0.311866
HHDFMX   Spouse of householder                 0.311866
Name: 3, Length: 133, dtype: float64
```



```
binned_NOEMP neutral                      -0.442819
binned_ADTOCC very low                    -0.273710
AMJOCC   Executive admin and managerial   -0.256460
binned_ADTIND low                         -0.148443
binned_DIVVAL high                        -0.124805
                                             ...
binned_DIVVAL very low                     0.247961
binned_NOEMP low                           0.254641
AMJOCC   Sales                             0.269126
binned_ADTIND neutral                      0.292181
binned_ADTOCC high                         0.316116
Name: 5, Length: 133, dtype: float64
```

Over represented top 3 columns in principal component of cluster of more data is: HHDREL-> Householder, HHDREL-> Householder, NOEMP->neutral.