

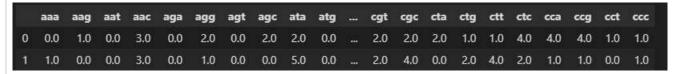


← Go Back to Data Analysis & Visualization

≔ Course Content

Clarification Note - Genomic Data Clustering

The frequency table displayed is the following:

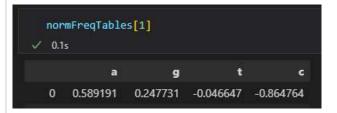


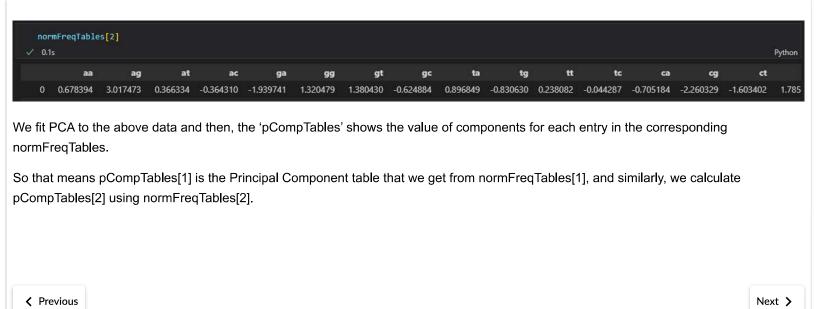
In the video, at 3:20, the instructor says "the PCA table is going to have one row for every column of the frequency table".

```
pca = PCA(n components = 2)
   pCompTables = {}
   for i in range(1,5):
       pca.fit(normFreqTables[i])
       pComponents = pca.transform(normFreqTables[i])
       pCompTables[i] = pd.DataFrame(pComponents[:, [0,1]], columns = ['pc1', 'pc2'])
       print('Explained variance for ' + str(i) + ' letters: ' + str(pca.explained_variance_ratio_.sum()))
   print(pCompTables[2].head())
Explained variance for 1 letters: 0.7489363490534278
Explained variance for 2 letters: 0.22774966356188592
Explained variance for 3 letters: 0.31670201938180154
Explained variance for 4 letters: 0.029334525098403036
        pc1
0 -0.949207 -0.396447
1 -0.100967 -0.873554
  1.198262 0.366900
  0.579445 1.174832
  0.102275 1.788382
```

Here, this refers to the fact that for every entry/row in the frequency table we calculate it. But, the other part of the explanation may be a little confusing.

At first, we scale the 'freqTables' into 'normFreqTables'. If you print 'normFreqTables[1] or normFreqTables[2]', you will be able to see the kind of data we are fitting the PCA algorithm into:





Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

© 2023 All rights reserved.

Help