The echo data consists of 96 rows and 64 columns. Exploratory analysis on vorticity was done using PCA and Hierarchical Clustering on 8 vortex variables viz., VortexArea, VortexIntensity, VortexDepth, VortexLength, EnergyDissipation, VorticityFluctuation, KineticEnergyFluctuation, ShearStressFluctuation.

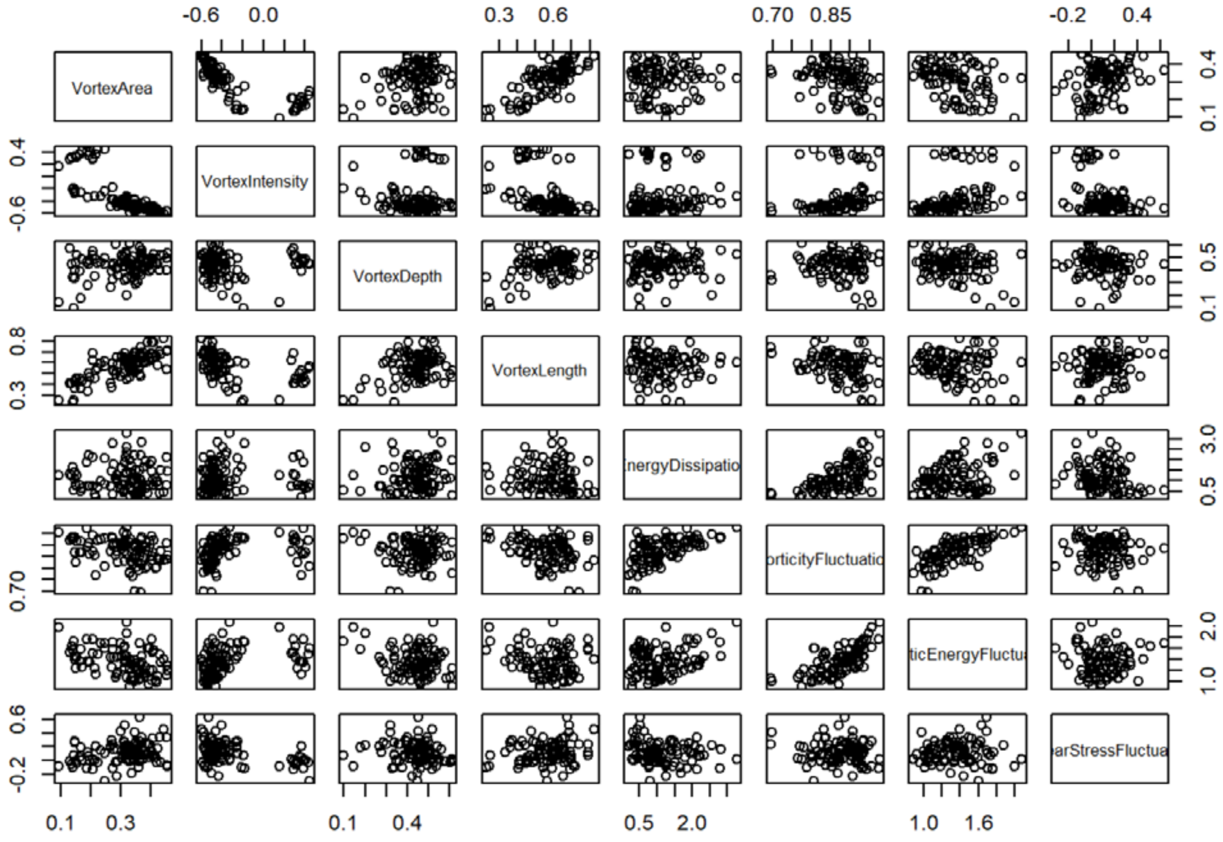


Figure 1: Scatterplot Matrix of vorticity variables

From the scatterplot matrix (Figure 1), there are evidence that some variables are correlated. Interestingly, there's an almost perfect linear relationship VoticityFluctuation and KineticEnergyFluctuation. Also, there's a distinct pattern in VorticityIntensity and all other variables; it seems to form two separate clusters. To better visualize the high-dimensional data, a dimensionality reduction method called Principal Component Analysis (PCA) is used.

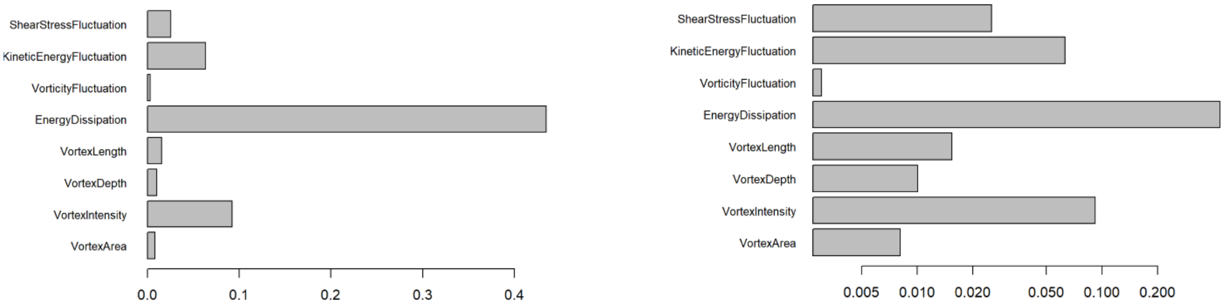


Figure 2: Variance of columns. a) unscaled columns, b) scaled columns

The first principal component of the unscaled data has a standard deviation of around 0.66 and accounts for about 68% of the variance in the data. The first column of loadings shows that the first principal component is just the Energy Dissipation. This is due to the variability in scale of the different variables in the data set. Figure 2 shows the variance of all columns used in the analysis where a is unscaled columns and b is scaled columns.

Thus, after scaling the data set, PCA was applied which shows that the first component is no longer dependent on one variable. Figure 3 shows the plot of components based on the variance they explain.

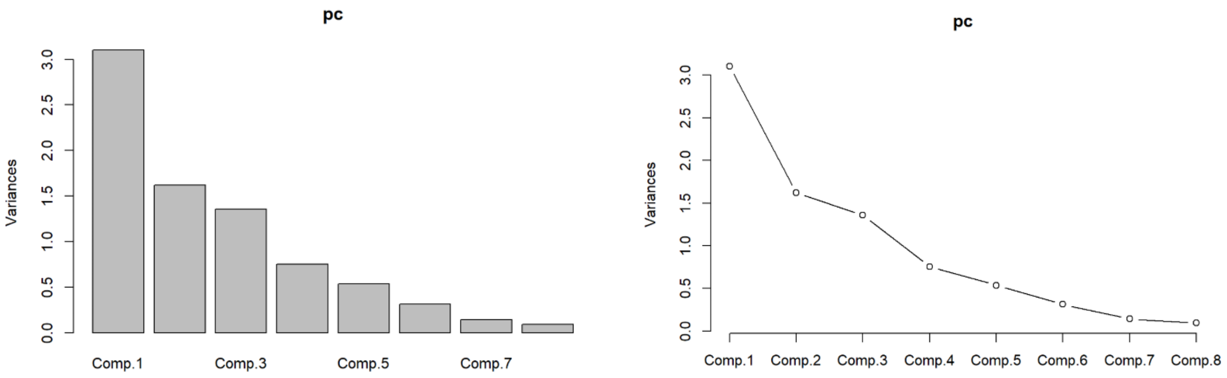
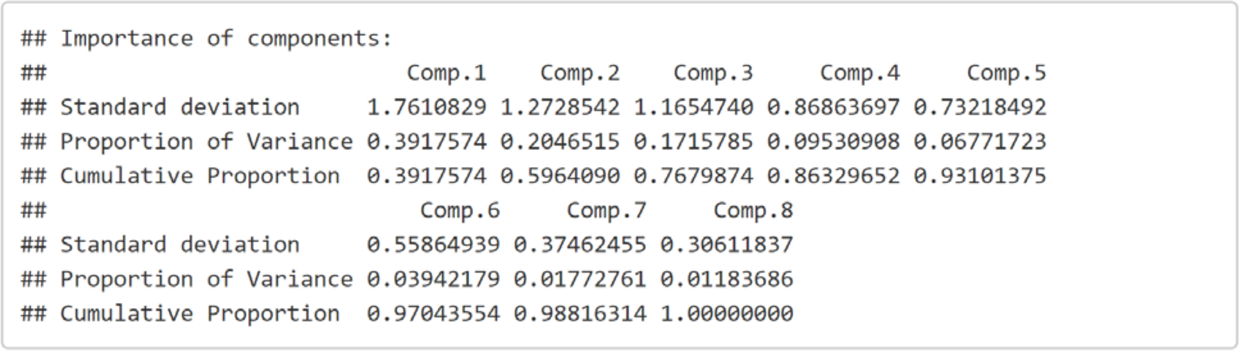


Figure 3: Variance explained by principal components



```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     1.7610829 1.2728542 1.1654740 0.86863697 0.73218492
## Proportion of Variance 0.3917574 0.2046515 0.1715785 0.09530908 0.06771723
## Cumulative Proportion  0.3917574 0.5964090 0.7679874 0.86329652 0.93101375
##                            Comp.6     Comp.7     Comp.8
## Standard deviation     0.55864939 0.37462455 0.30611837
## Proportion of Variance 0.03942179 0.01772761 0.01183686
## Cumulative Proportion  0.97043554 0.98816314 1.00000000
```

Table 1: Summary of principal components

Table 1 gives us the summary of the principal components. We choose four components as it explains 86% of the variance of the data.

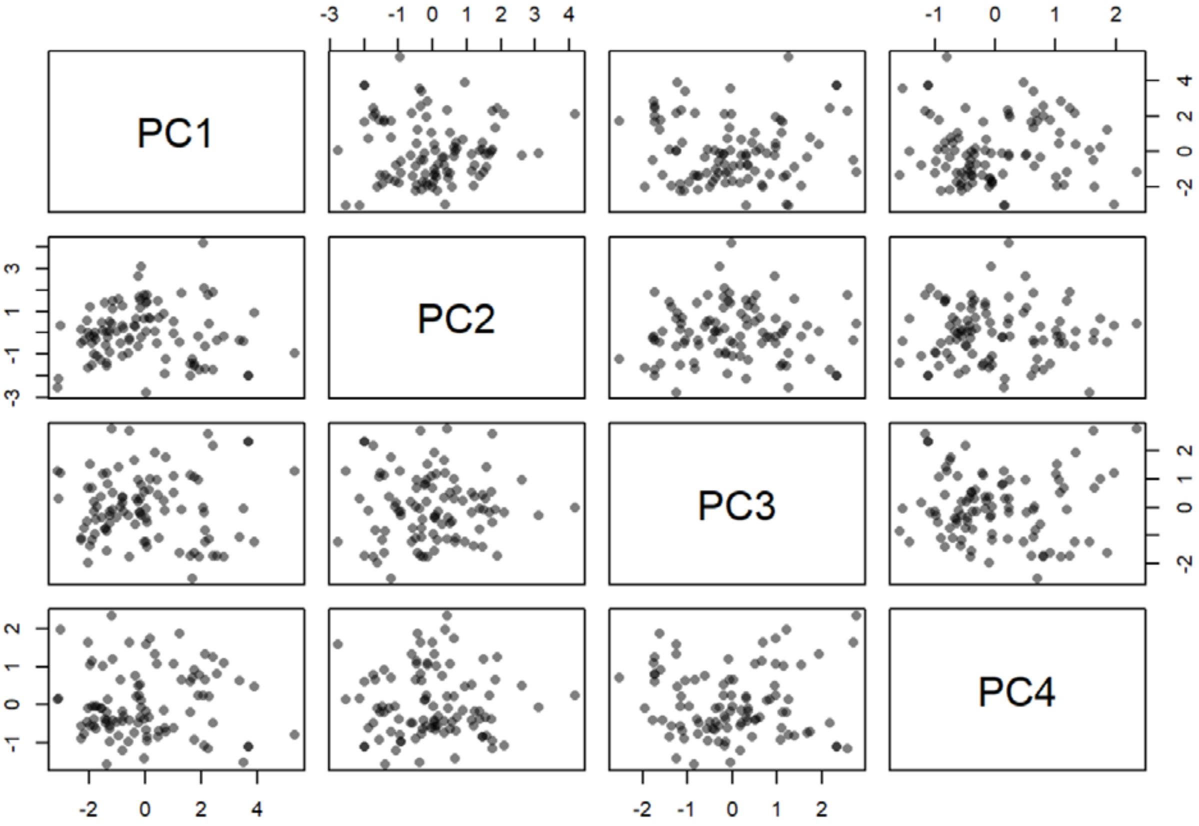


Figure 4: Scatterplot Matrix of Four Principal Components

The scatterplot matrix in Figure 4 shows 12 2-D projections of data which are in 4-D space. The plot shows that they seem to be have some pattern such that some clusters are formed in different projections. To visualize this and find patterns, as in which patients are most similar, we used an unsupervised machine learning technique Hierarchical Clustering.

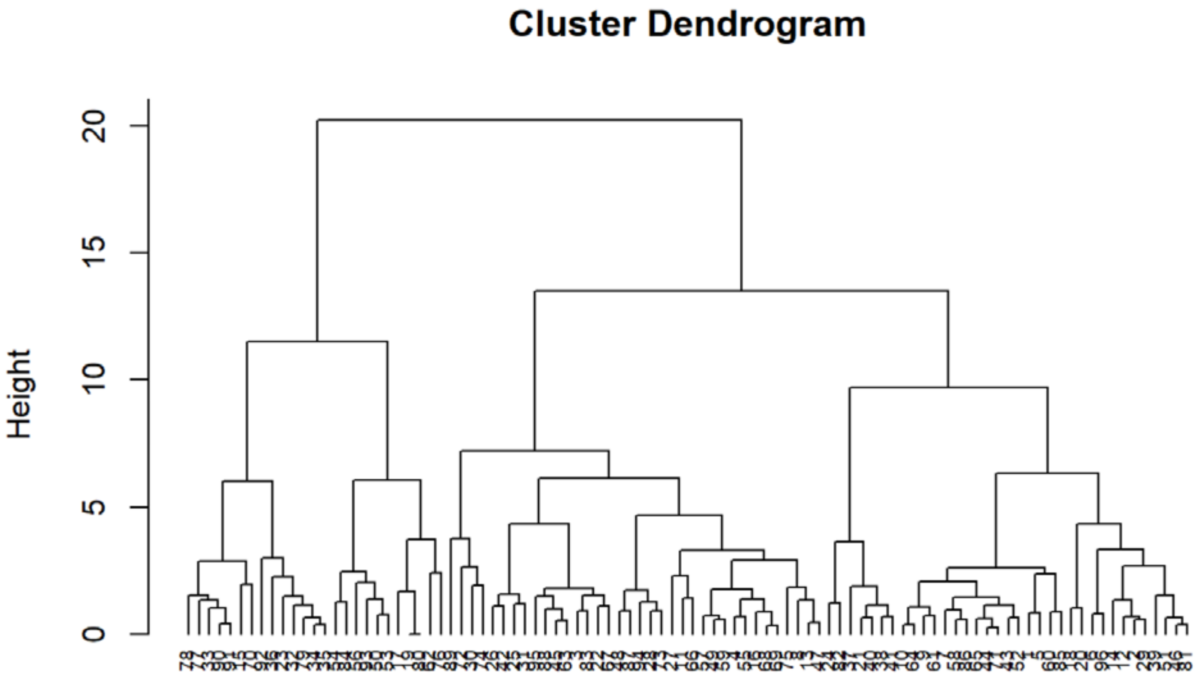


Figure 5:Cluster Dendrogram using Ward's Method

The hierarchical clustering was performed on the principal components determined in the previous steps. The dendrogram shown in Figure 5 shows the structure of the cluster using Ward's Method. Agnes package in r allows us to identify the structure strength of the cluster using various algorithms. It shows that Ward's method identifies the strongest clustering structure of the four methods assessed. From the dendrogram shown above, we can identify the sub-groups or clusters. It seems that we can use four clusters. The identification of the clusters in the dendrogram and in the data set is shown below in Figure 6 and Figure 7 respectively.

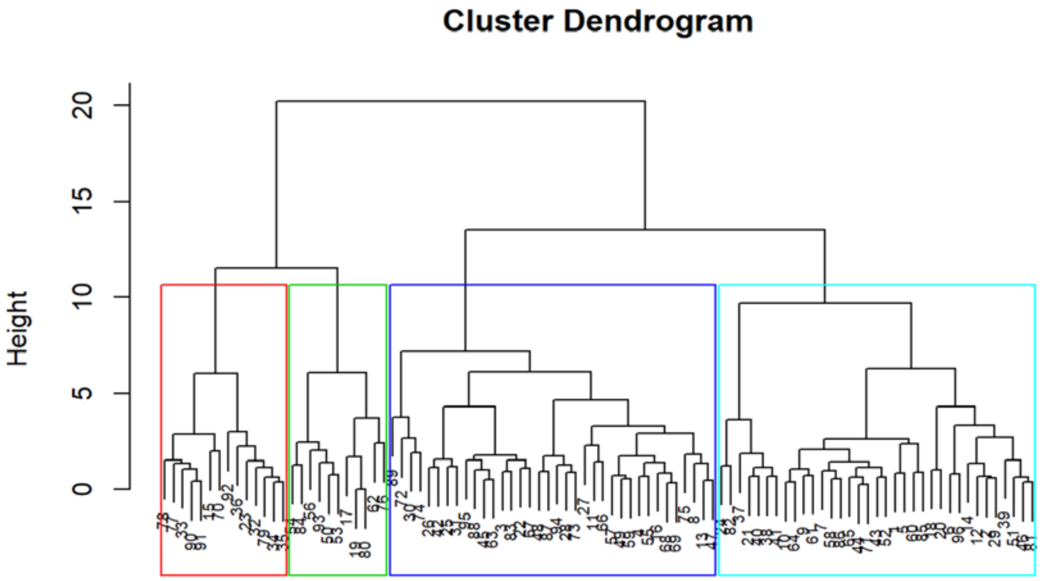


Figure 6: Dendrogram with clusters

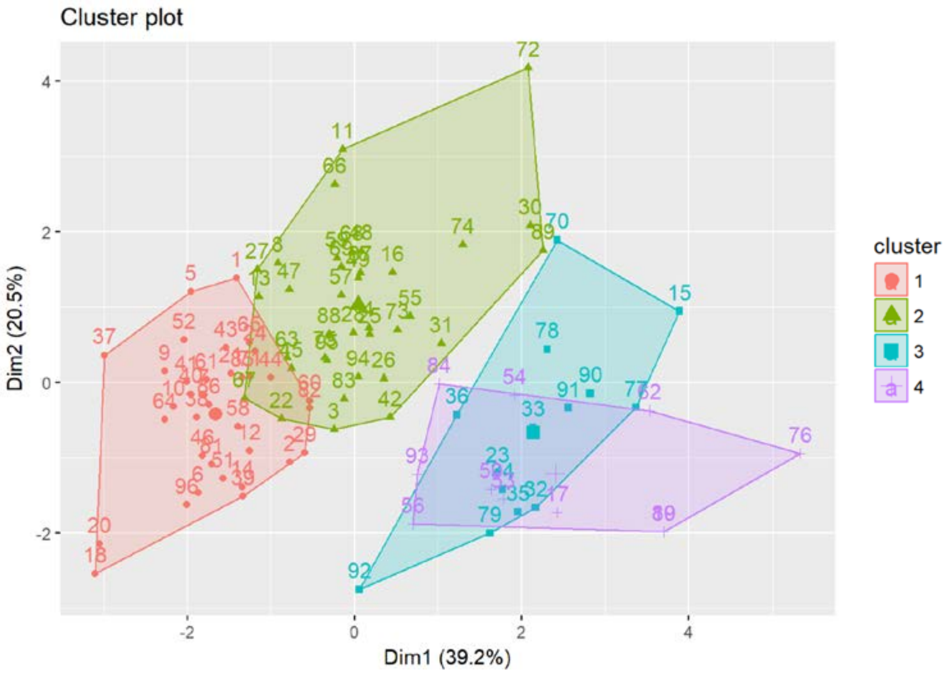


Figure 7: Clusters

Using the clusters identified, we can visualize the principal components to see if any pattern emerges. The scatter plot matrix in Figure 8 shows clusters in the principal components.

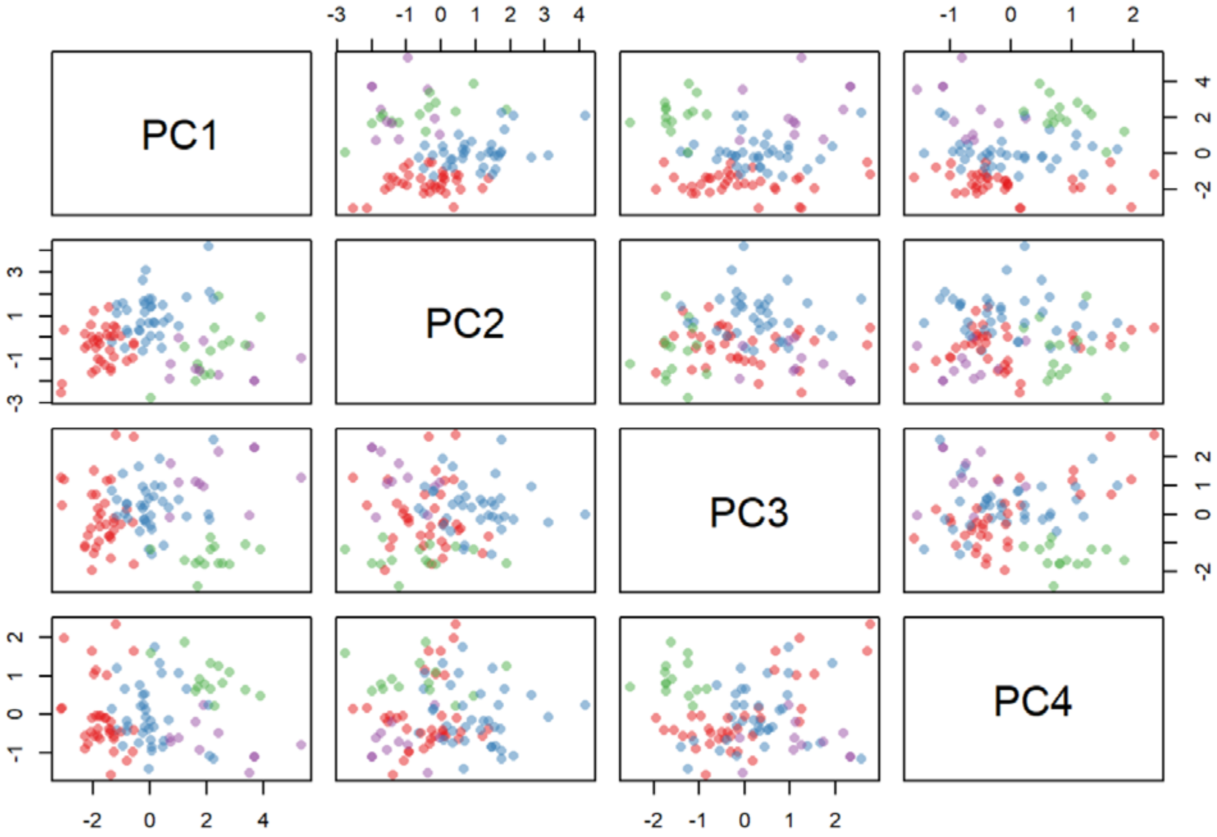


Figure 8: Scatter plot matrix of cluster in principal components

The following table shows number of patients in each cluster.

Cluster sizes			
4	**3**	**2**	**1**
11	14	35	36

Table 2: Table showing size of each cluster

Cluster 1	17 19 50 53 54 56 62 76 80 84 93
Cluster 2	15 23 32 33 34 35 36 70 77 78 79 90 91 92
Cluster 3	1 2 5 6 7 9 10 12 14 18 20 21 24 29 37 38 39 40 41 43 44 46 51 52 58 60 61 64 65 71 81 82 85 86 96
Cluster 4	3 4 8 11 13 16 22 25 26 27 28 30 31 42 45 47 48 49 55 57 59 63 66 67 68 69 72 73 74 75 83 87 88 89 94 95

Table 3: List of patients in each cluster

The following table shows that cluster 3 seems to have patients with higher NYHA class and Abnormal EF. Cluster 1, 2, or 4 doesn't show such high NYHA patients.

Cluster 3		
Patient ID	NYHA	AbnormalEF50
15	NA	0
23	1	0
32	3	1
33	3	1
34	3	1
35	3	1
36	1	0
70	NA	0
77	3	1
78	3	1
79	3	1
90	4	1
91	3	1
92	3	1

Cluster 4		
Patient ID	NYHA	AbnormalEF50
17	1	NA
19	1	0
50	1	0
53	1	0
54	1	0
56	1	0
62	3	1
76	2	1
80	1	0
84	1	0
93	1	0
17	1	NA
19	1	0
50	1	0

To identify if any of the categorical values have any relationships with the cluster, χ^2 test is performed. The table below show the χ^2 test along with the table with number of patients in each cluster.