**Procedure**

1. Get the long format dataset for one household (regarded session\_id as new id)
2. Transform long format into wide format. (web\_id/category)
3. PCA (normalization)
4. K-means & silhouette method

**Two-people Household**

**Website**

Raw dataset (201×3)

|  |  |  |
| --- | --- | --- |
| session\_id | web\_id | count |

Train dataset (72×42)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| session\_id | web\_id1 | web\_id2 | … | web\_idn |

Y (72×41)

|  |  |  |  |
| --- | --- | --- | --- |
| web\_id1 | web\_id2 | … | web\_idn |

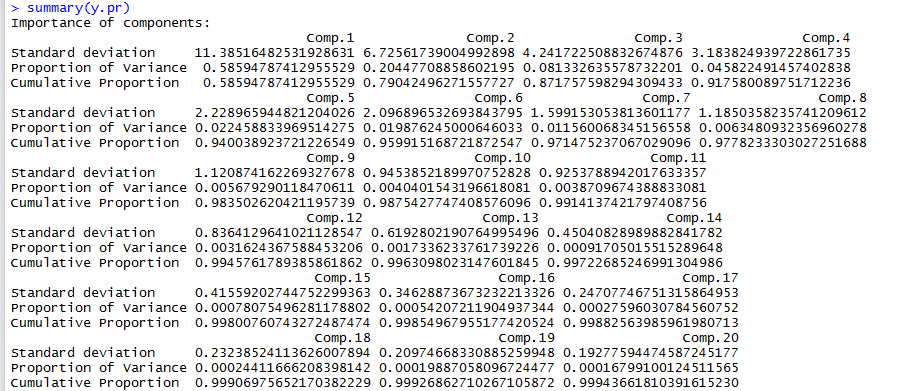
**1.**

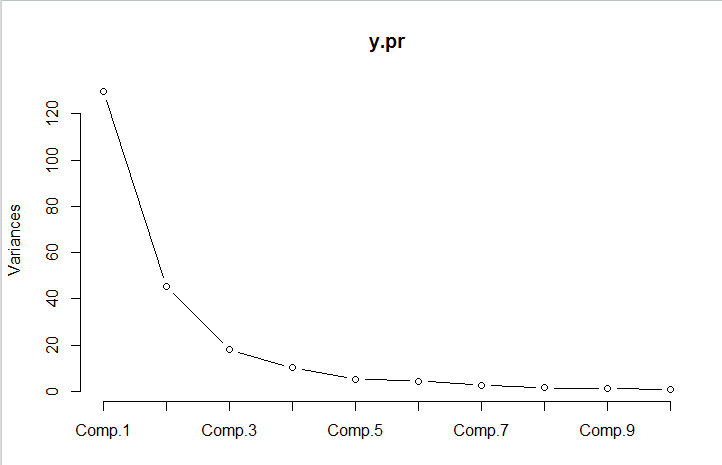
Do not normalize data

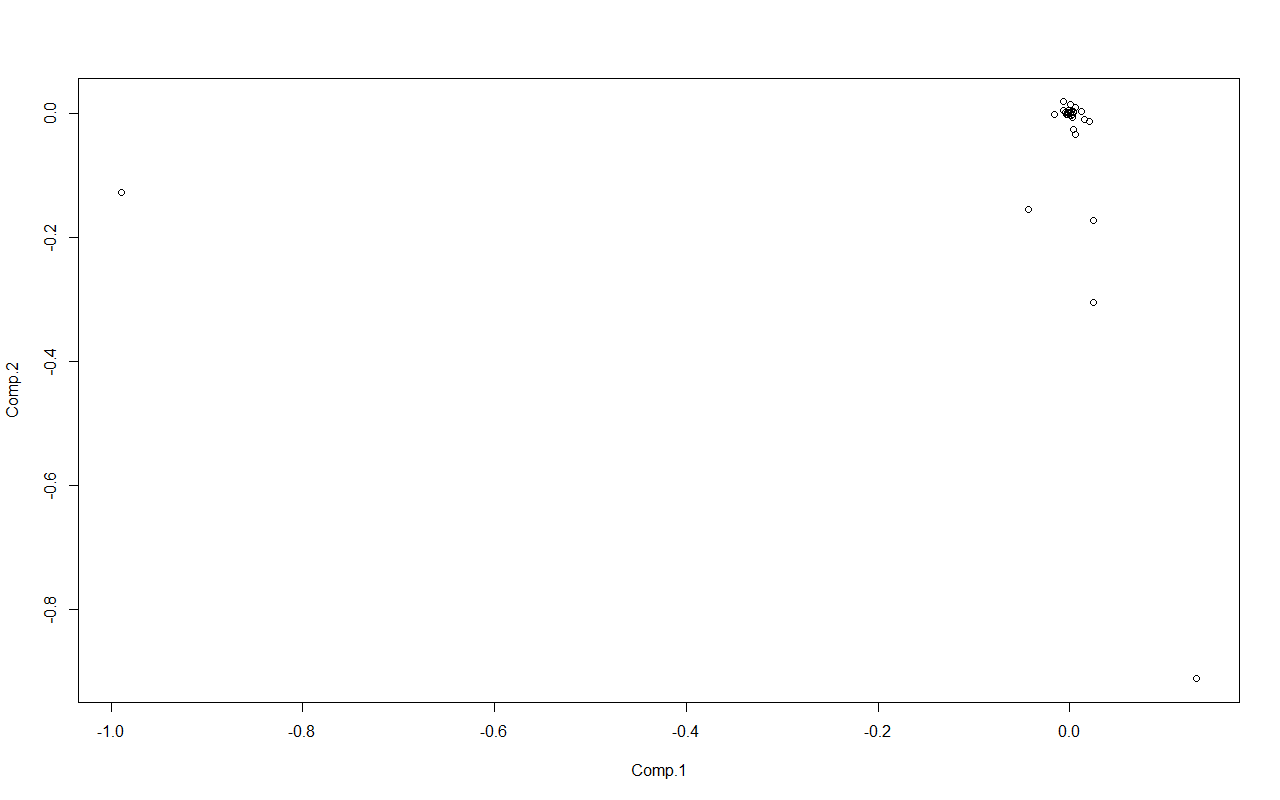
Variance of first principal component is large.

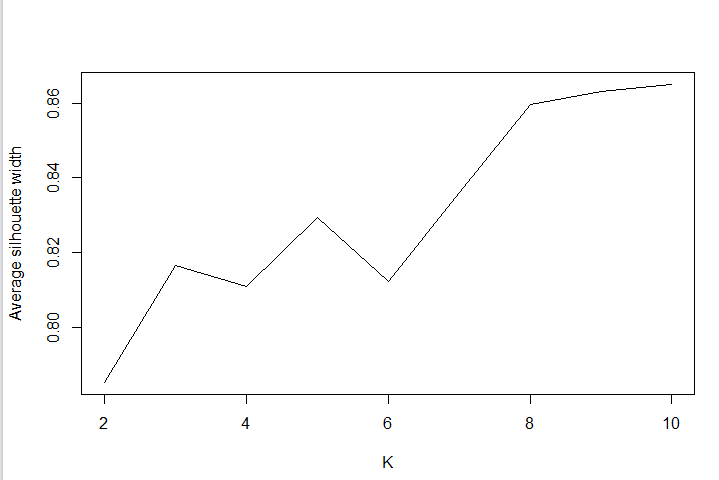
Select the first four principal components as features to do K-means clustering

K-means clustering doesn’t perform well









First four principal components

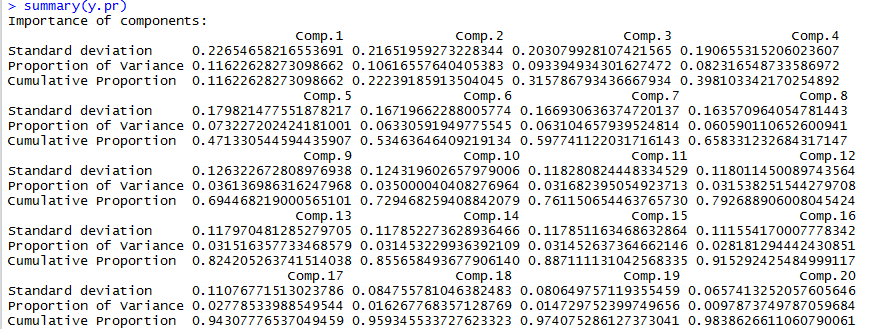
**2.**

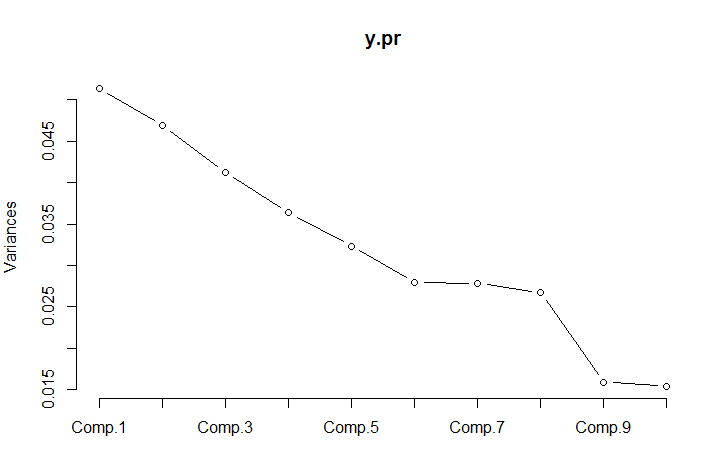
normalize data by column (count/sum(count))

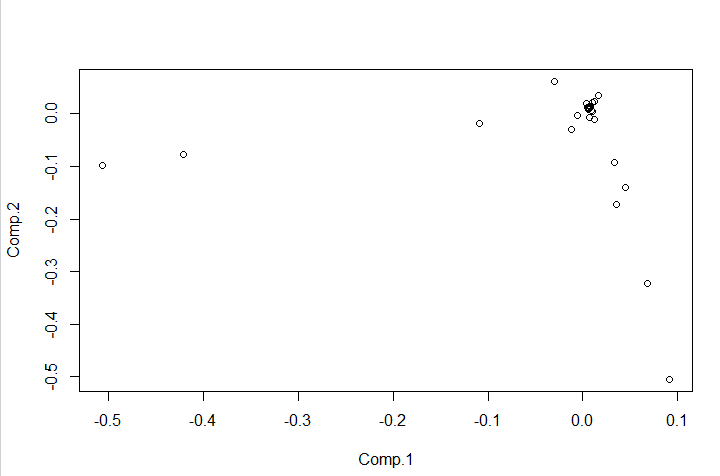
the proportion of variance is small.

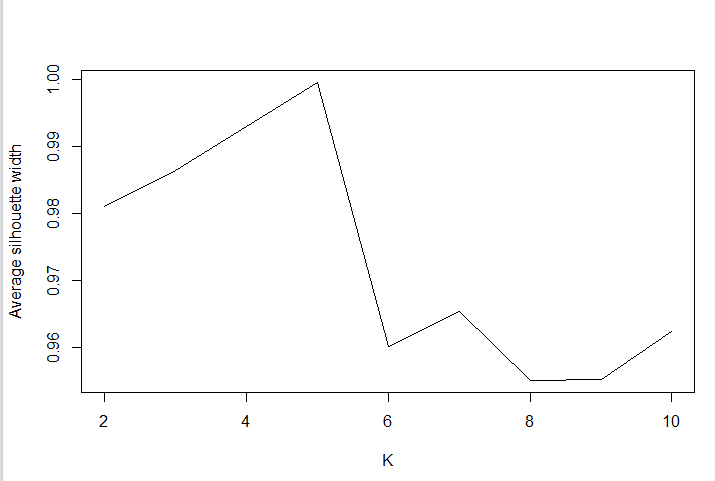
Select the first four principal components as features to do K-means clustering, and k = 5

Select the first ten principal components as features, and k = 9

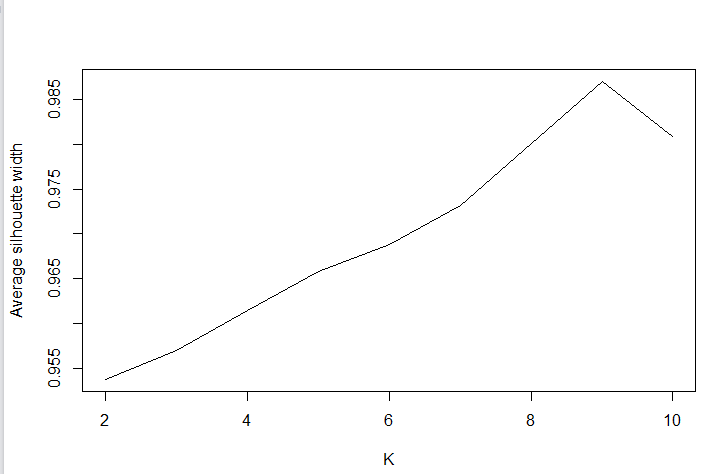








First four principal components

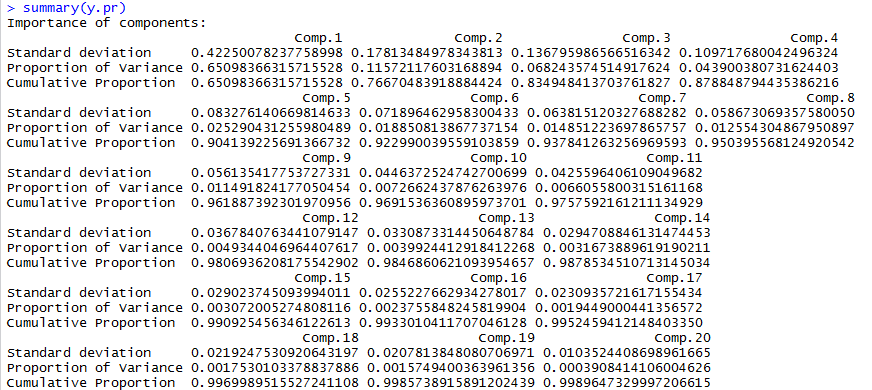


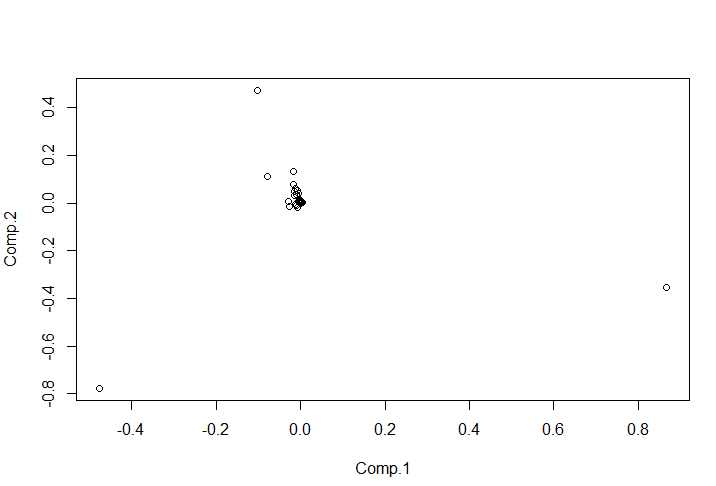
First ten principal components

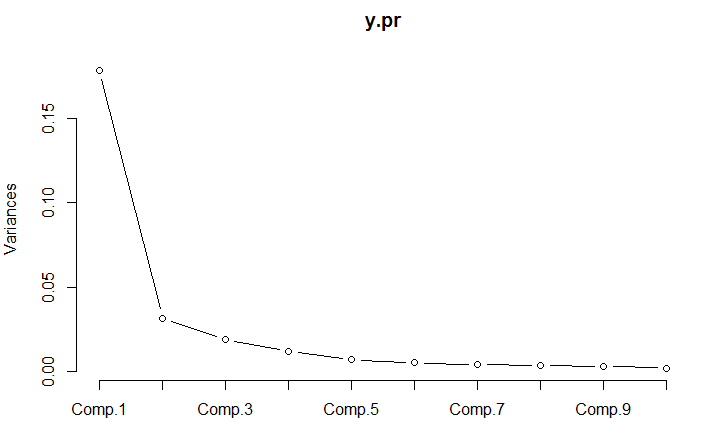
**3.**

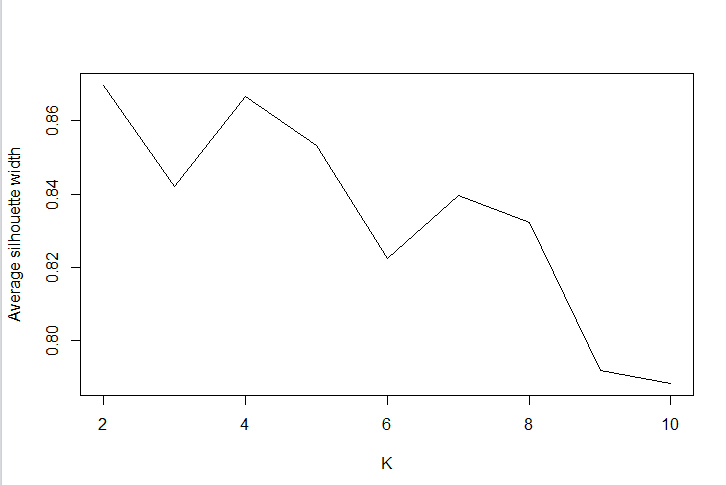
normalize data by rows (count/sum(count))

Select the first four principal components as features to do K-means clustering, and it sometimes works.



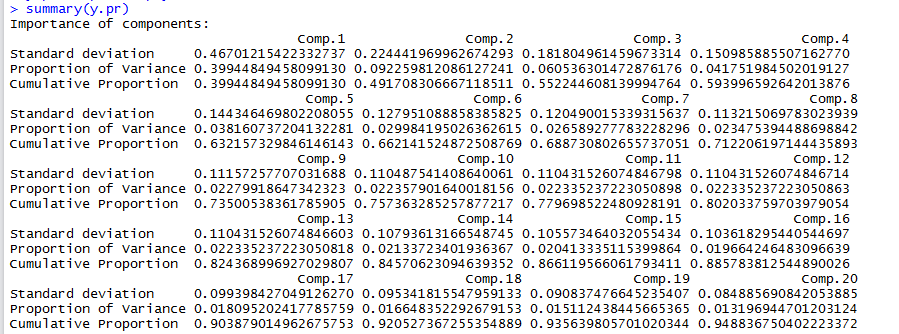


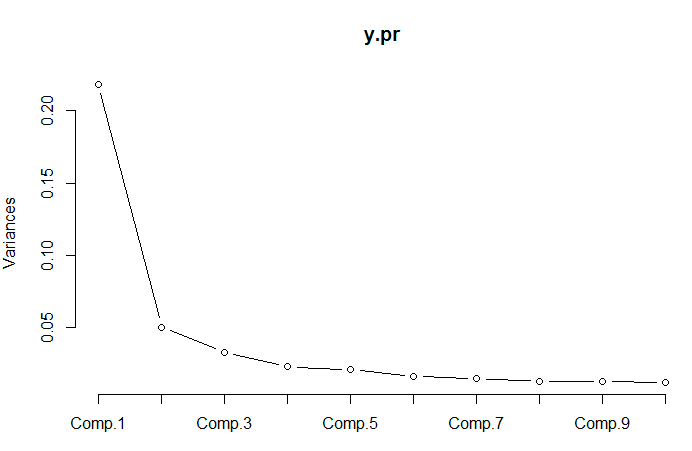


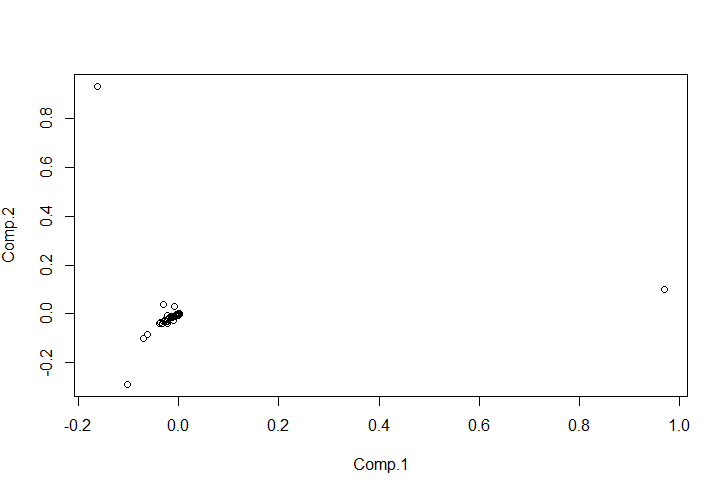


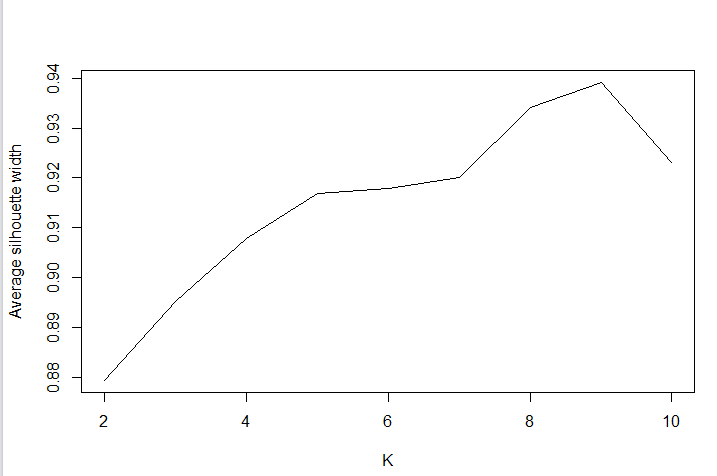
First four principal components

Try it on the other household, which household size = 5









First four principal components

**Category**

Raw dataset (66×3)

|  |  |  |
| --- | --- | --- |
| session\_id | category | count |

Train dataset (37×15)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| session\_id | Category1 | Category2 | … | categoryn |

Y (37×14)

|  |  |  |  |
| --- | --- | --- | --- |
| Category1 | Category2 | … | categoryn |

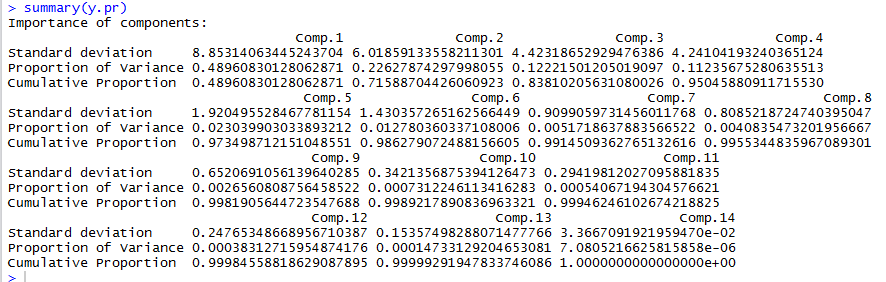
**1.**

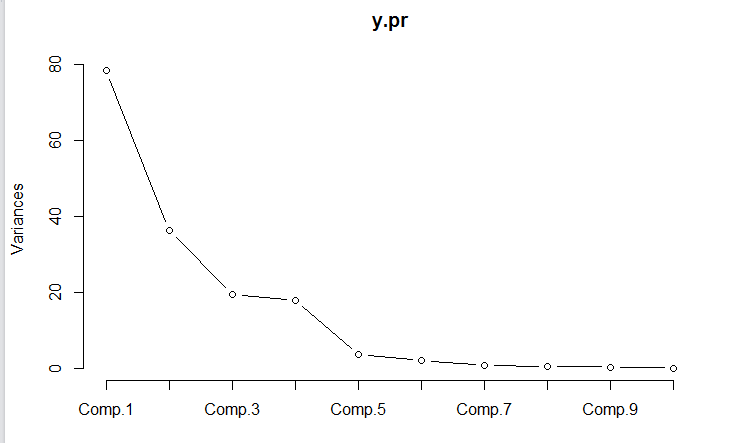
Do not normalize data

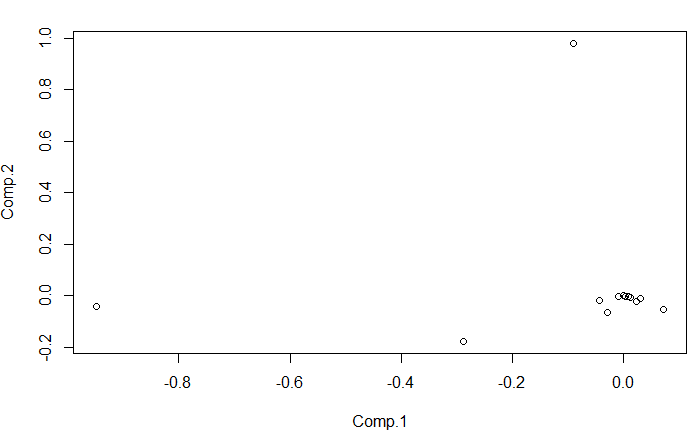
Variance of first principal component is large.

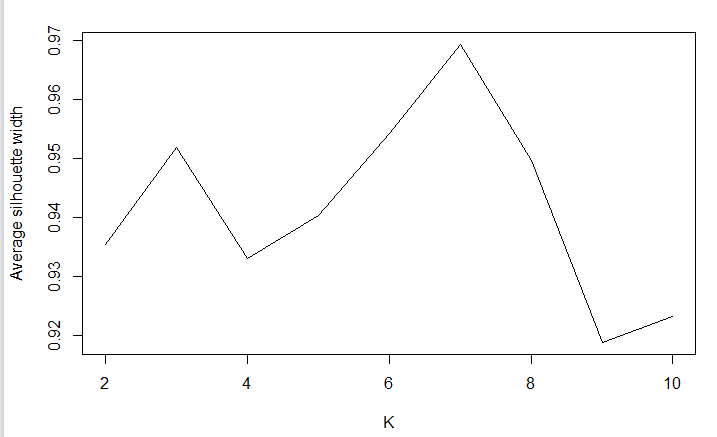
Select the first four principal components as features to do K-means clustering

K-means clustering doesn’t perform well, and K=7









First four principal components.

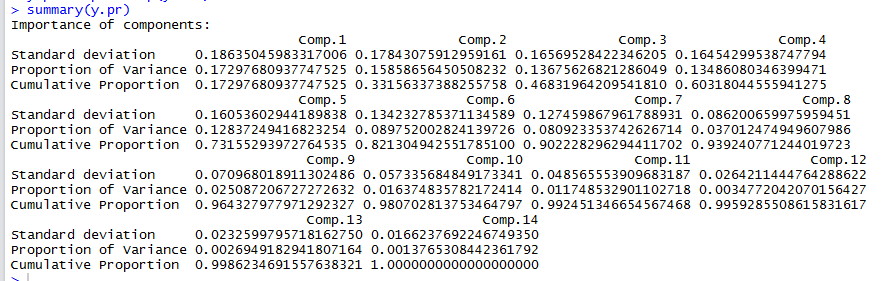
**2.**

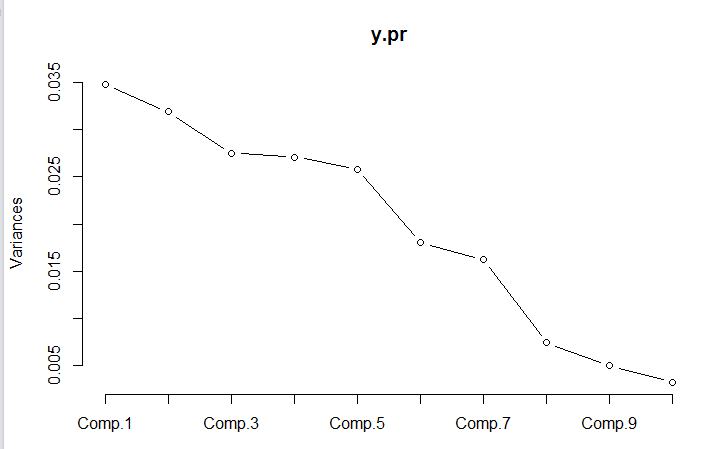
normalize data by column (count/sum(count))

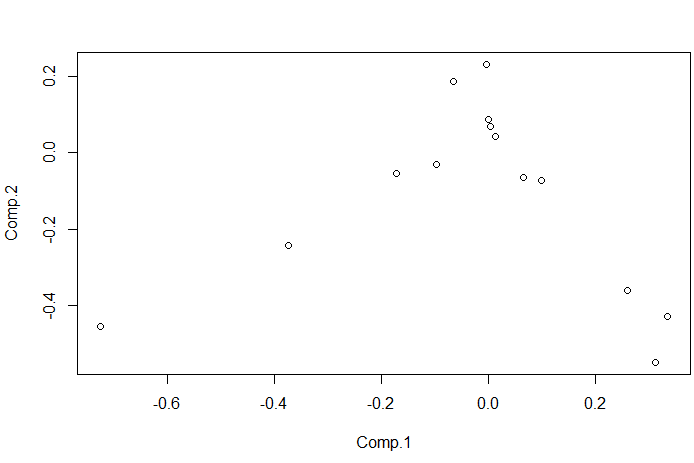
the proportion of variance is small.

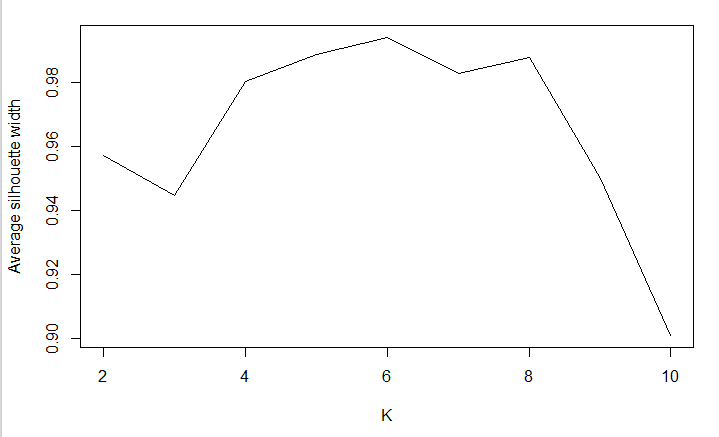
Select the first four principal components as features to do K-means clustering, and k = 6

Select the first ten principal components as features, and k = 6

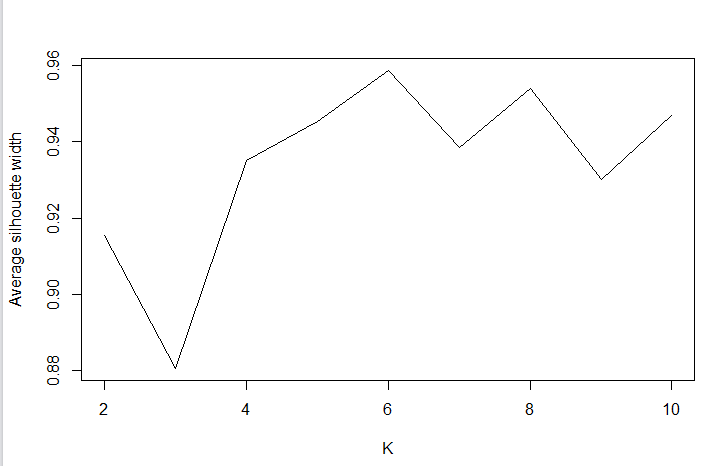








First four principal components

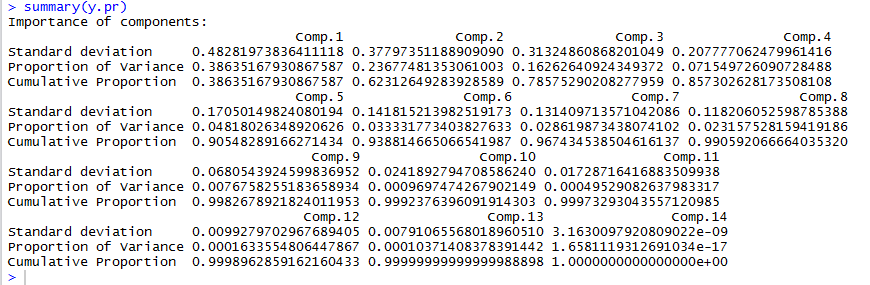


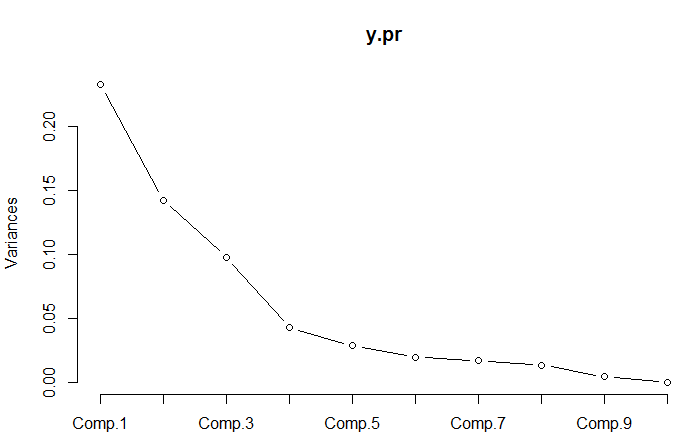
First ten principal components

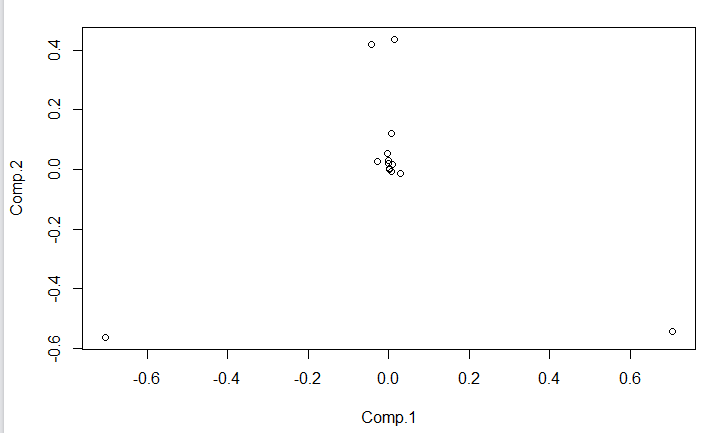
**3.**

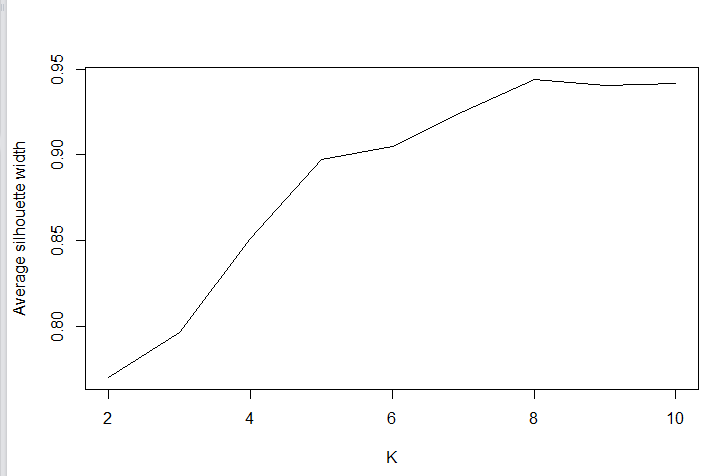
normalize data by rows (count/sum(count))

Select the first four principal components as features to do K-means clustering, and it doesn’t perform well.









First four principal components

**Improvement**

1. Increase sample size. Used several months’ data rather than one.
2. Try other feature selection methods. Because if we want to compute n PCs, we need n2 observations.
3. Get some new features, such as the time one session last.
4. Delete outliers, such as count = 1.