Cyrus Kwan

ID: 45200165  Date: 15/05/2021

REPORT: A COMPARISON BETWEEN K-means & DBSCAN Clustering Algorithms on Young People Survey

Contents

[Data Preparation: 2](#_Toc71841688)

[Structure of Survey Data: 2](#_Toc71841689)

[Data Dictionary: 2](#_Toc71841690)

[Summary of Survey Data: 2](#_Toc71841691)

[Data Treatment: 3](#_Toc71841692)

[Imputation 3](#_Toc71841693)

[Data Manipulation 3](#_Toc71841694)

[Standardization & Normalization 3](#_Toc71841695)

[Data Reduction/Principal Components Analysis: 3](#_Toc71841696)

[Clustering: 5](#_Toc71841697)

[K-Means Algorithm: 5](#_Toc71841698)

[DBSCAN Algorithm: 11](#_Toc71841699)

[Profiling: 13](#_Toc71841700)

[Demographic Analysis of K-means Clusters 13](#_Toc71841701)

[Conclusion: 17](#_Toc71841702)

[Comparative Analysis: 17](#_Toc71841703)

[Critique: 17](#_Toc71841704)

[Position: 17](#_Toc71841705)

[Appendix: 18](#_Toc71841706)

# Data Preparation:

## Structure of Survey Data:

### Data Dictionary:

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute Type | Definition | Variable Type | Number of Items |
| Music Preferences | The respondent’s overall opinion on music as well as specific genres of music. | Likert scale (1-5): continuous | 19 |
| Movie Preferences | The respondent’s overall opinion on movies as well as specific types of movies. | Likert scale (1-5): continuous | 12 |
| Hobbies & Interests | Respondent’s opinion on specific academics, arts, activities, etc. | Likert scale (1-5): continuous | 32 |
| Phobias | Opinion on common phobias e.g. flying, spiders, public speaking. | Likert scale (1-5): continuous | 10 |
| Health habits | Personal rating of respondent health habits derived from list of common habits. | Likert scale (1-5): continuous, categorical | 3 |
| Personality traits, Views on life, & opinions | Rating of how much the respondent agrees or disagrees with a view. | Likert scale (1-5): continuous, categorical | 57 |
| Spending habits | Overall opinion of spending and rating of how much the respondent spends on different sectors. | Likert scale (1-5): continuous | 7 |
| Demographics | General demographics information e.g. age, gender. | Likert scale (1-5): continuous, categorical | 10 |

## Summary of Survey Data:

> ncol(survey\_raw)

[1] 150

> nrow(survey\_raw)

[1] 1010

*Output 1: The dimensions of raw survey data are of 150 attributes and 1010 observations.*

> length(list\_NA(survey\_raw))

[1] 144

*Output 2: The number of incomplete columns in that dataset. Hence, a lot of the dataset requires imputation and treatment.*

## Data Treatment:

### Imputation

The dataset is composed of a small sample of 1010 observations. Hence, rows containing missing values were replaced as opposed to being removed. k-nearest neighbours (kNN) imputation was performed on the dataset to minimize potential skewing or distortions in distribution that are present in other imputation methods (replacing missing values with means or medians). Although this report will only require the use of music, movies, and demographics data, imputation was performed on the entire raw dataset such that the kNN algorithm could make better predictions.

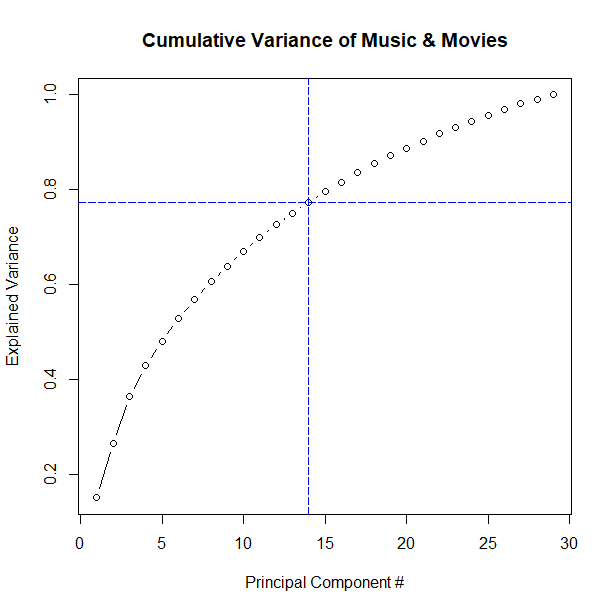
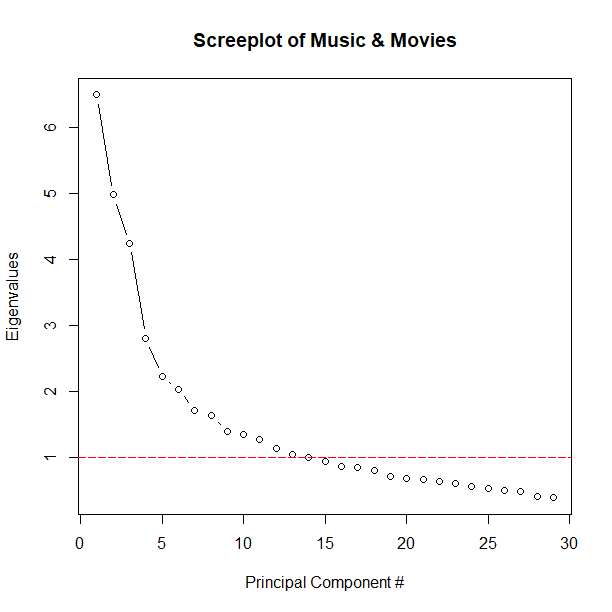
### Data Manipulation

Principal components analysis (PCA) and clustering algorithms were performed on solely on music and movie preference data. Hence, music and movies data were sliced from the imputed dataset. Additionally, overall ratings of music and movies were dropped; generally, most people enjoy music and movies making their overall ratings redundant.

### Standardization & Normalization

Standardization of the data was not performed to retain the patterns in observations. Additionally, standardization could be considered redundant since all values in music and movies are set on a 1-5 Likert scale.

# Data Reduction/Principal Components Analysis:



*Figure 1: Screeplot of music and movies show that the point at which principal components only account for one attribute is where eigen values are greater than one (approximately 13 or 14 principal components. The cumulative variance plot shows that 14 components account for about 75% of the variance of the total dataset.*

> MM\_eigen

[1] 6.4871365 4.9752053 4.2425997 2.8064139 2.2273615 2.0396663 1.7225502

[8] 1.6371801 1.4036245 1.3555821 1.2721656 1.1383384 1.0549512 1.0115392

[15] 0.9462665 0.8743575 0.8530362 0.8114711 0.7141094 0.6826805 0.6782830

[22] 0.6388835 0.6132091 0.5661814 0.5459523 0.5154966 0.4986427 0.4232753

[29] 0.3961534

*Output 3: A closer look at eigen values shows that the 14th principal component accounts for only 1% worth of data. Researcher judgement could be applied to further reduce the amount of principal components e.g. a principal component must provide and extra 30% worth of data. However, researcher judgement is usually arbitrary. Hence, this study will use principal components up until they no longer provide any comparative value to the original attributes.*

> MM\_cumprob

[1] 0.1504008 0.2657484 0.3641108 0.4291760 0.4808162 0.5281048 0.5680413

[8] 0.6059984 0.6385407 0.6699692 0.6994637 0.7258554 0.7503139 0.7737659

[15] 0.7957046 0.8159761 0.8357533 0.8545669 0.8711231 0.8869507 0.9026763

[22] 0.9174885 0.9317055 0.9448321 0.9574897 0.9694412 0.9810020 0.9908154

[29] 1.0000000

*Output 4: Raw values of the cumulative probability output show that 14 principal components account for 77.38% of the dataset. Researcher judgement can also be made to select principal components up to an arbitrary value e.g. where the amount of principal components accounts for 80% of the dataset. However, this study will use principal components for the reasons explained in Output 1.*



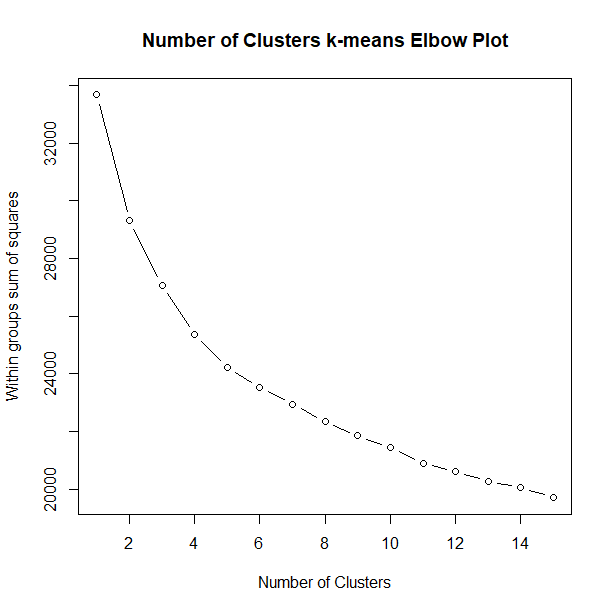
*Figure 2: Varimax rotation of principal components show the correlation between each principal component to the original attributes. Each component is shown to have correlations with distinct attributes e.g. PC4 shows correlations with “Rock”, “Metal or Hardrock”, and “Punk” which could be generalized to some overarching genre.*

*NOTE: Reproduced in excel for better readability compared to R output (see figure 17 in appendix)*

# Clustering:

## K-Means Algorithm:

K-means was initially used for music and movies data because it is one of the most used and reliable flat algorithms. One of its drawbacks is that it only clusters based on ovular partitioning. Additionally, it requires the calculation of optimal clusters.



*Figure 3: Elbow plot of within sum of squares used to determine the optimal number of clusters. The plot shows no clear elbow such that it is difficult to visually determine the number of clusters required for a k-means algorithm.*

> PC\_nClust = NbClust(opt\_MM\_PC, method = "kmeans")

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Among all indices:

\* 7 proposed 2 as the best number of clusters

\* 3 proposed 3 as the best number of clusters

\* 4 proposed 4 as the best number of clusters

\* 2 proposed 5 as the best number of clusters

\* 1 proposed 7 as the best number of clusters

\* 2 proposed 9 as the best number of clusters

\* 1 proposed 13 as the best number of clusters

\* 2 proposed 14 as the best number of clusters

\* 1 proposed 15 as the best number of clusters

\*\*\*\*\* Conclusion \*\*\*\*\*

\* According to the majority rule, the best number of clusters is 2

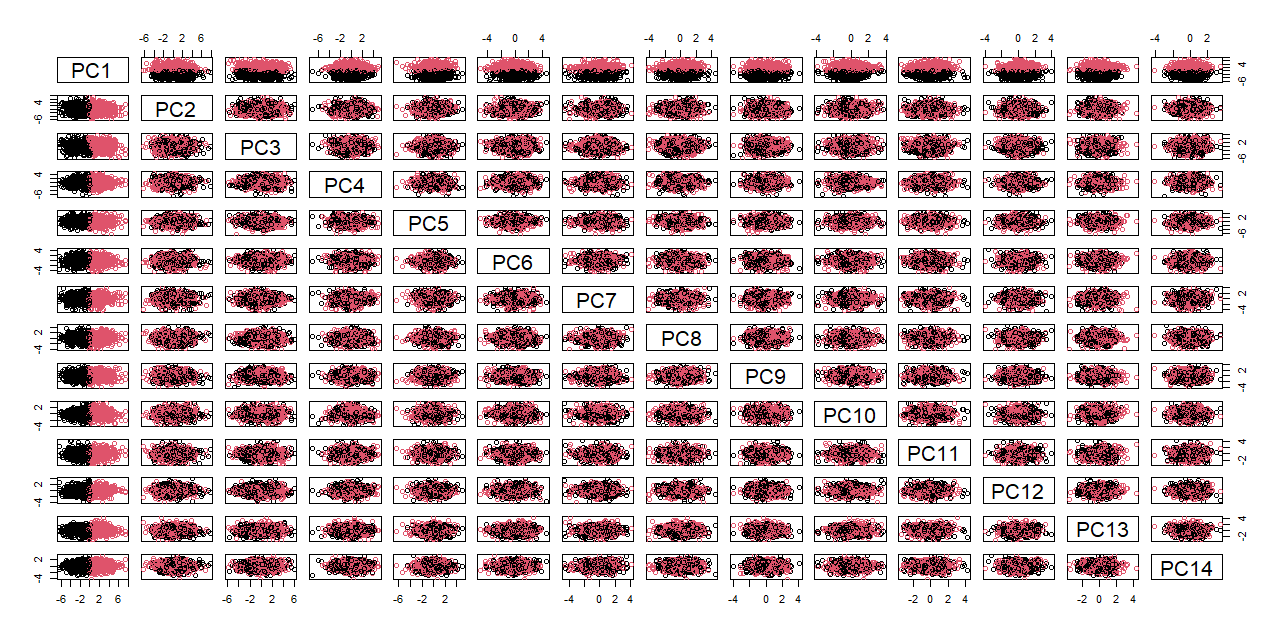
*Output 5: Using R package NbClust to determine the optimal number of clusters. Note that the suggested number of clusters is likely always greater than 1 since clustering implies separating the data into groups of more than 1.*

> table(PC\_KMClust$clust)

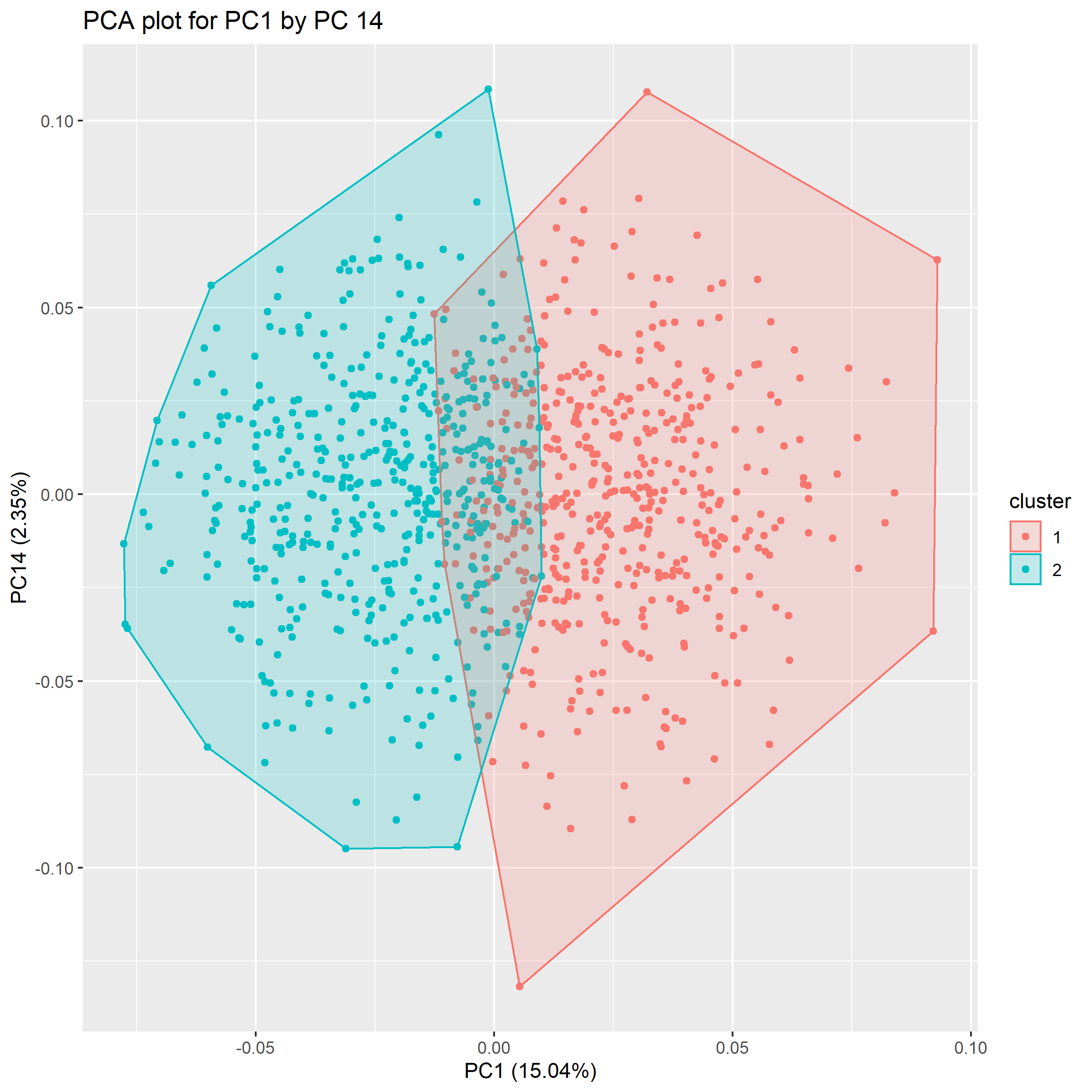
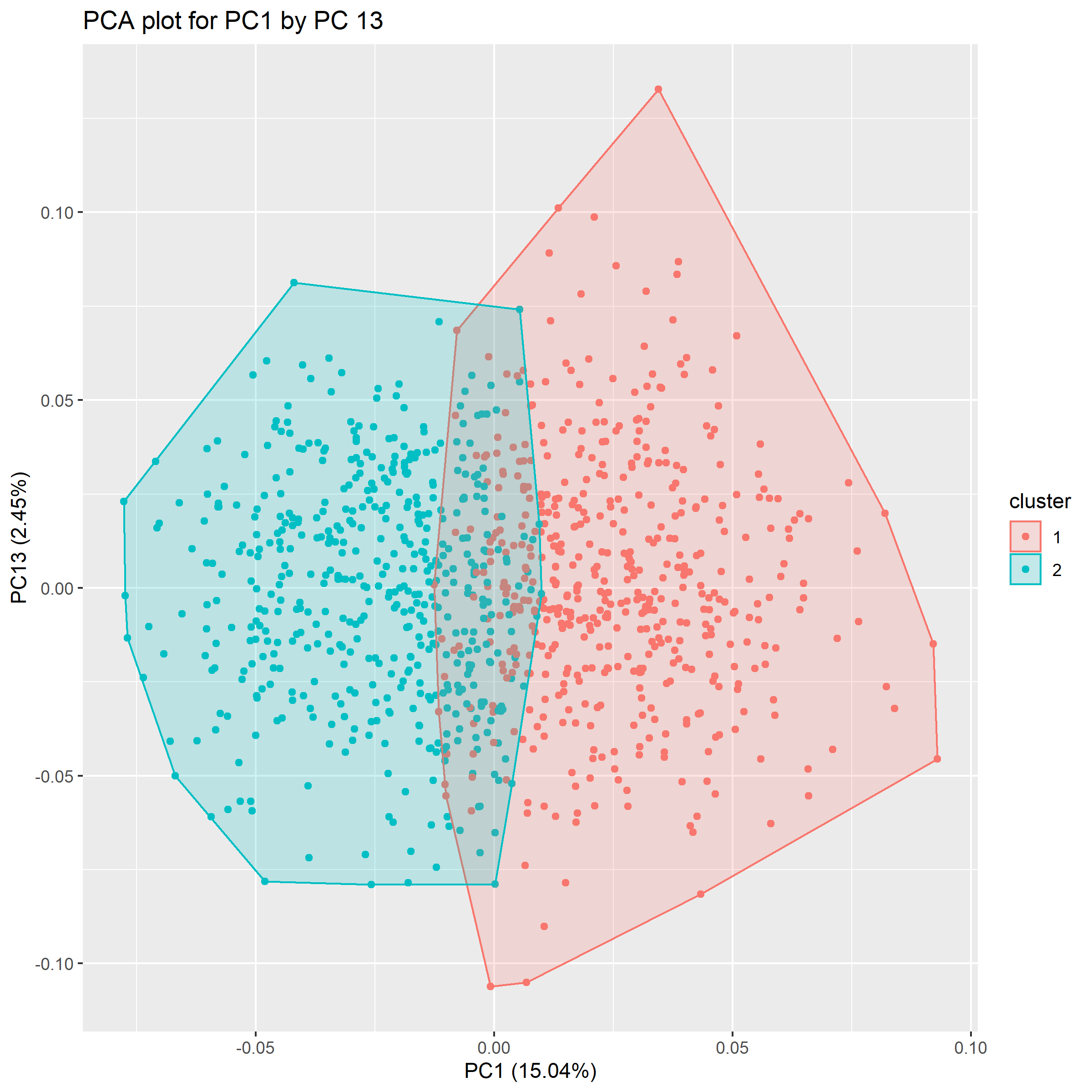
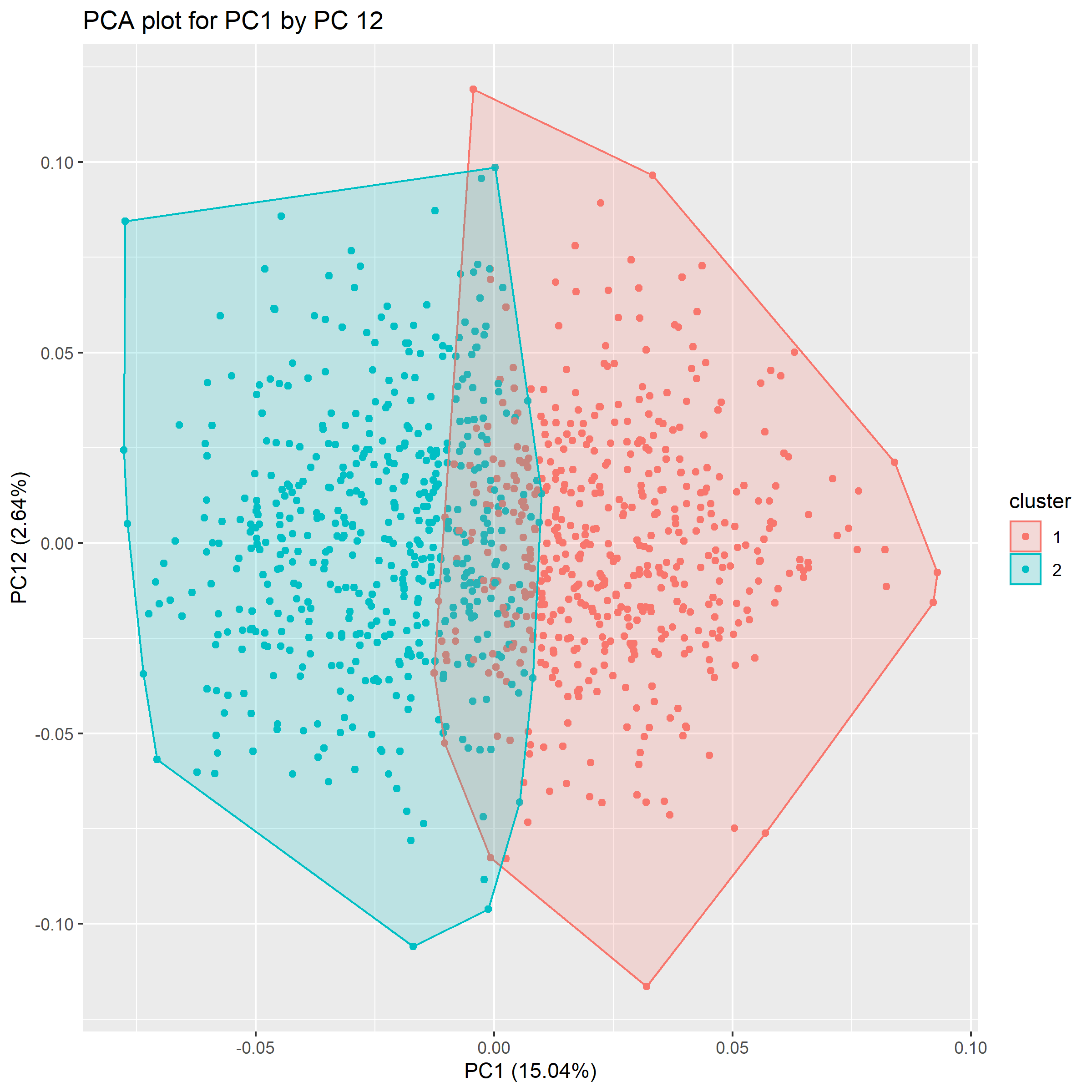
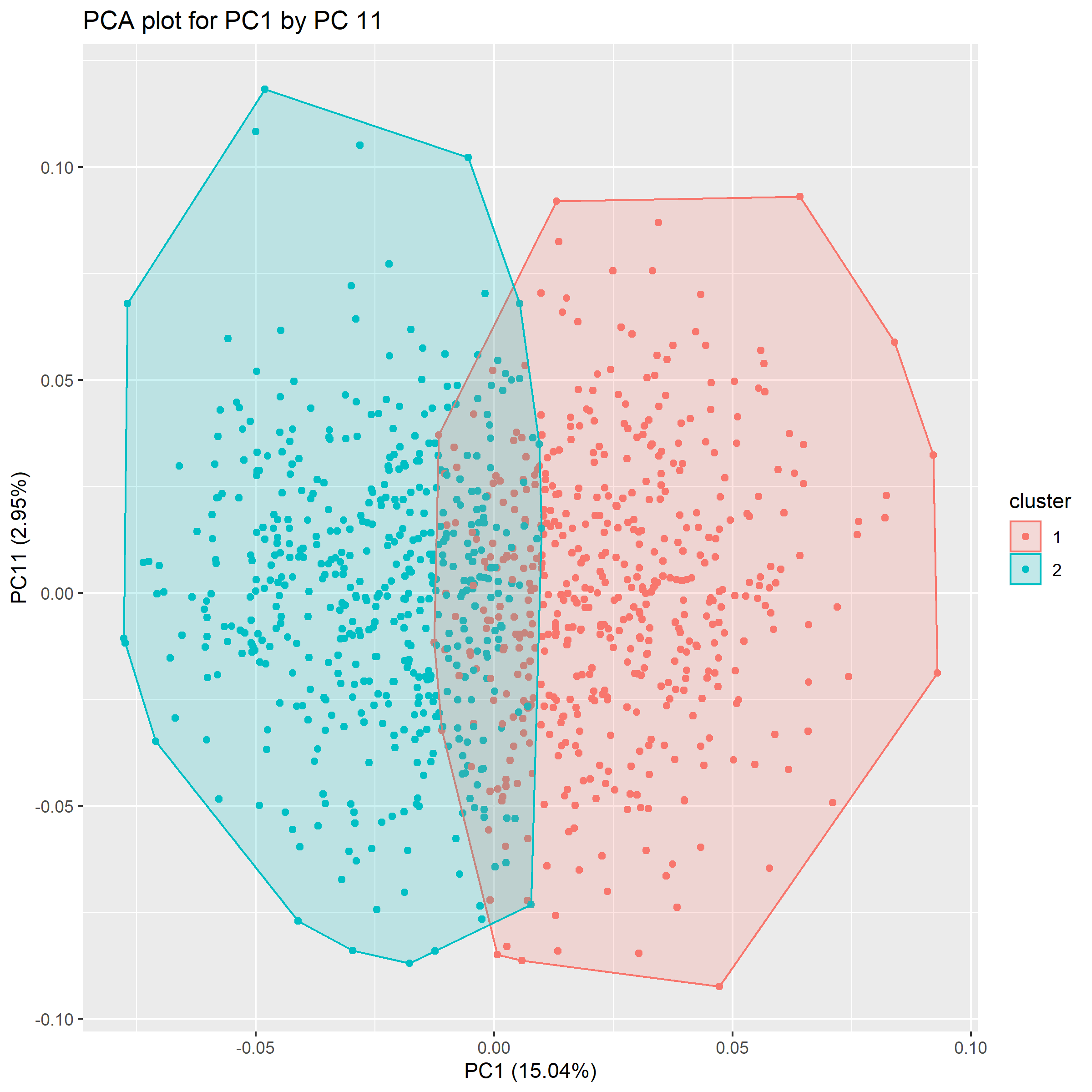
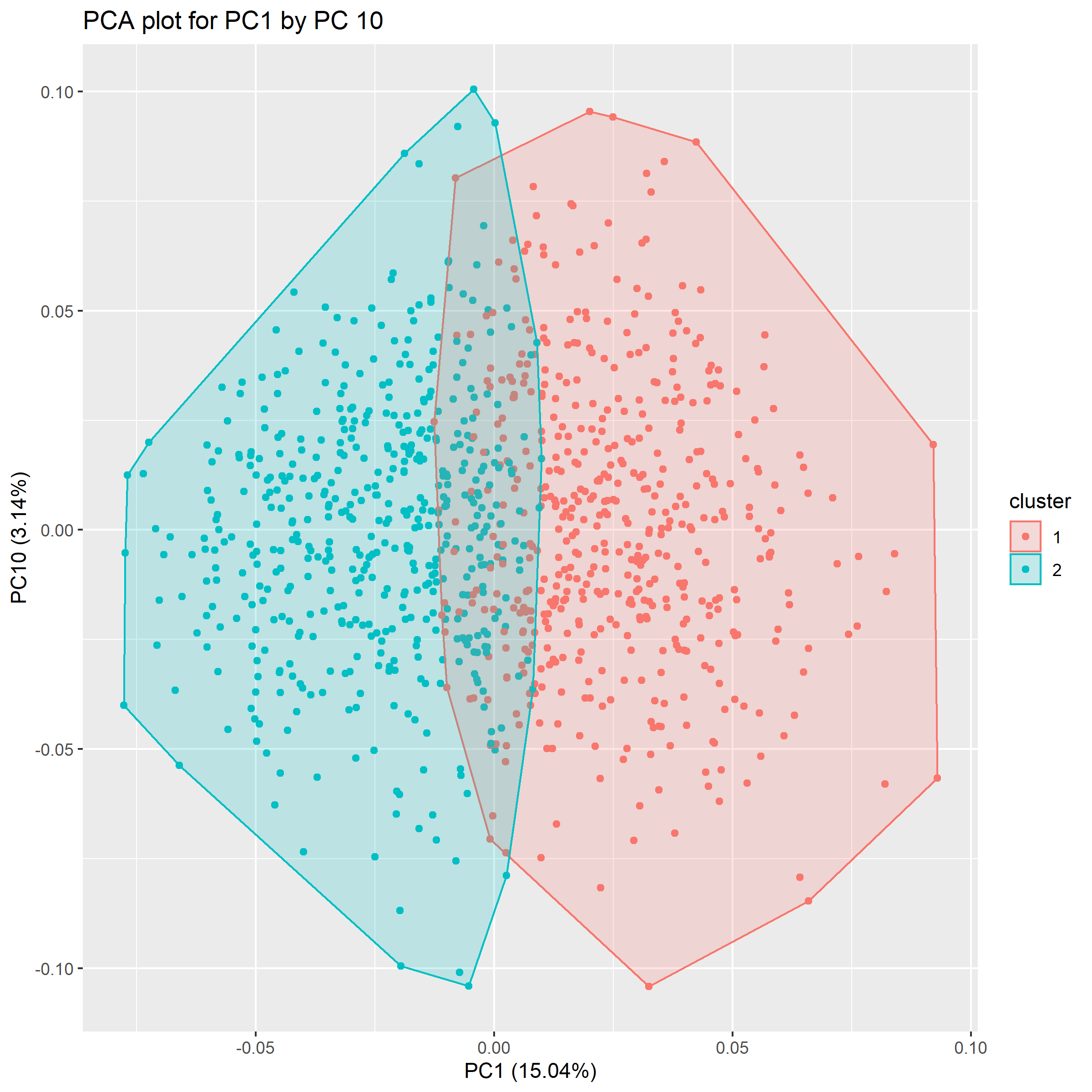
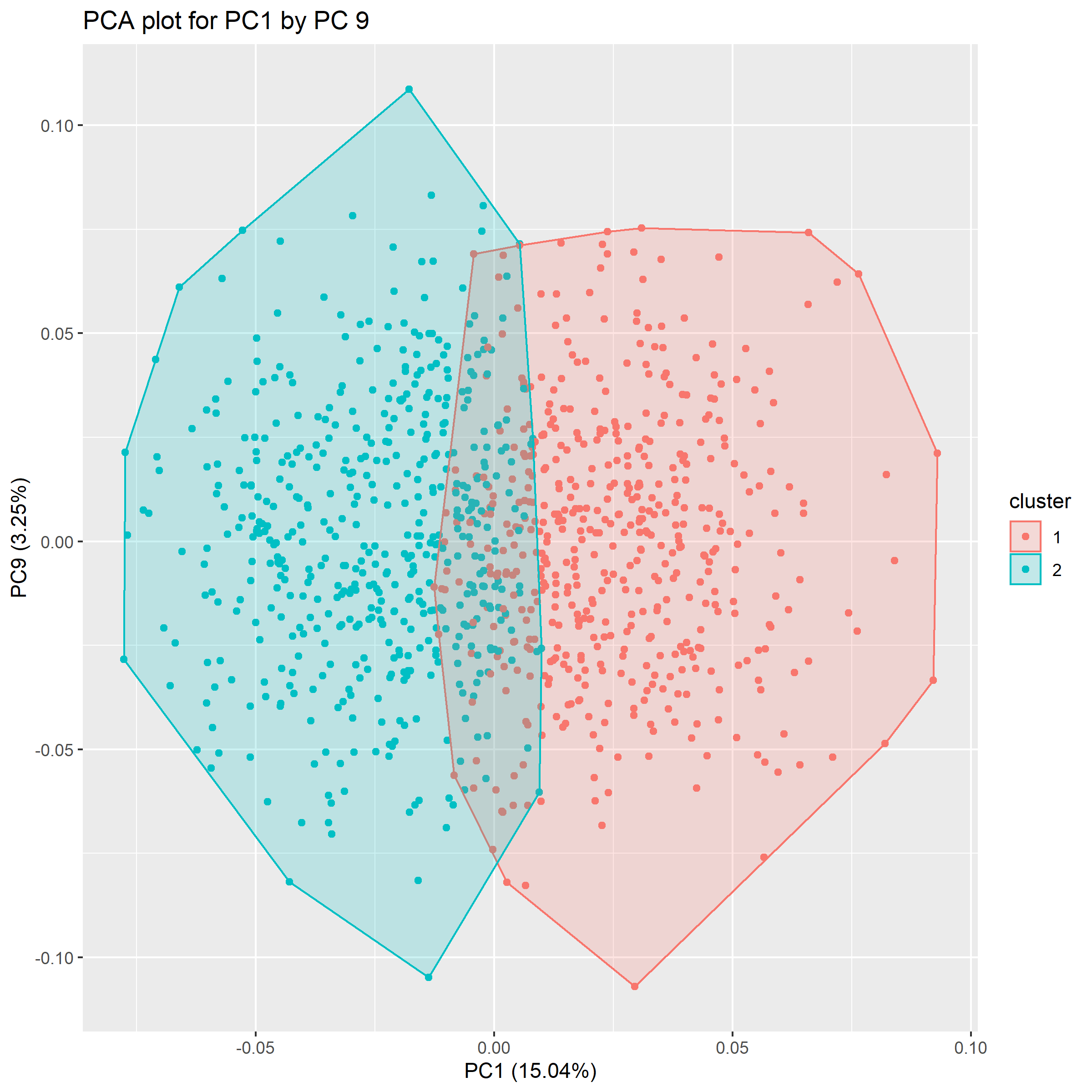
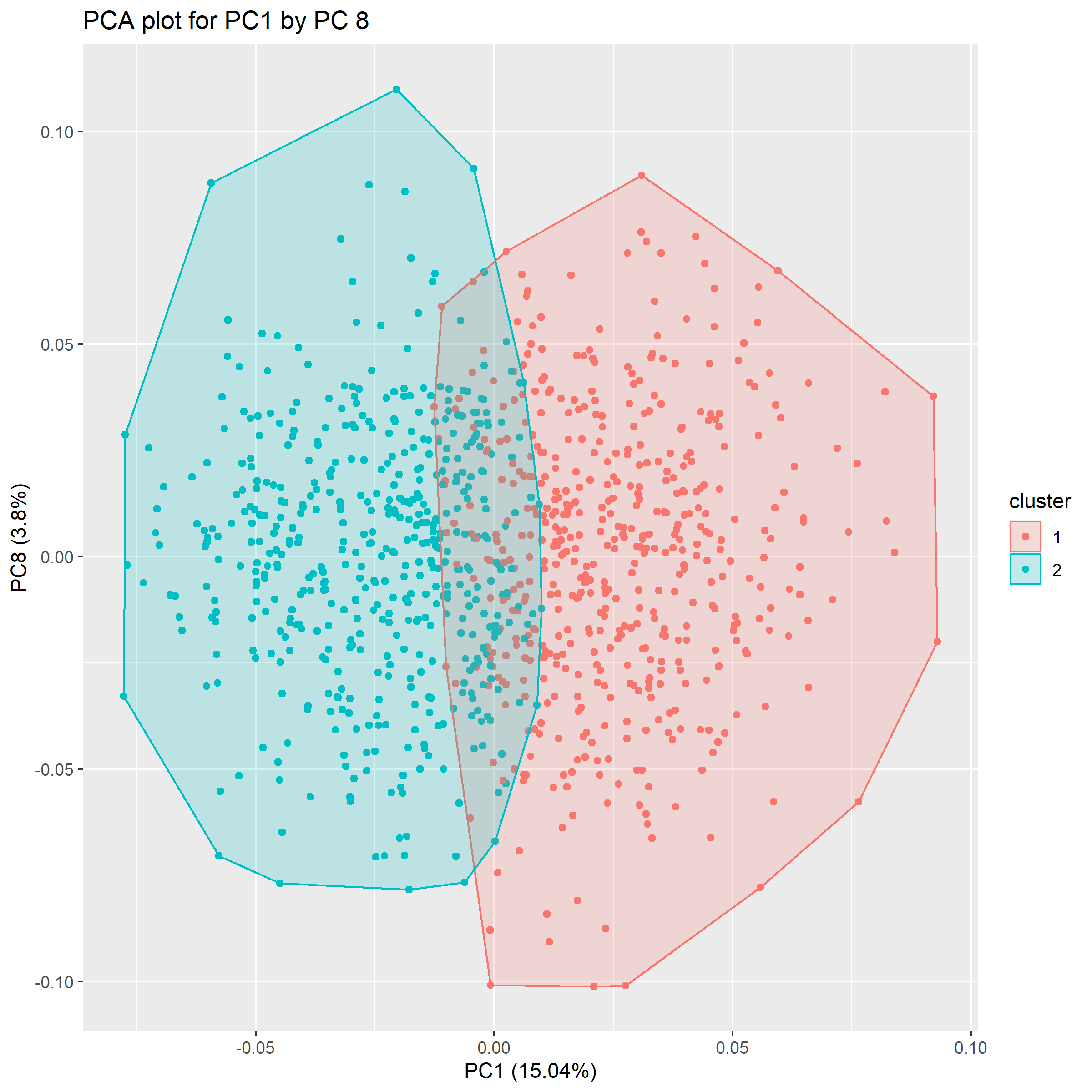
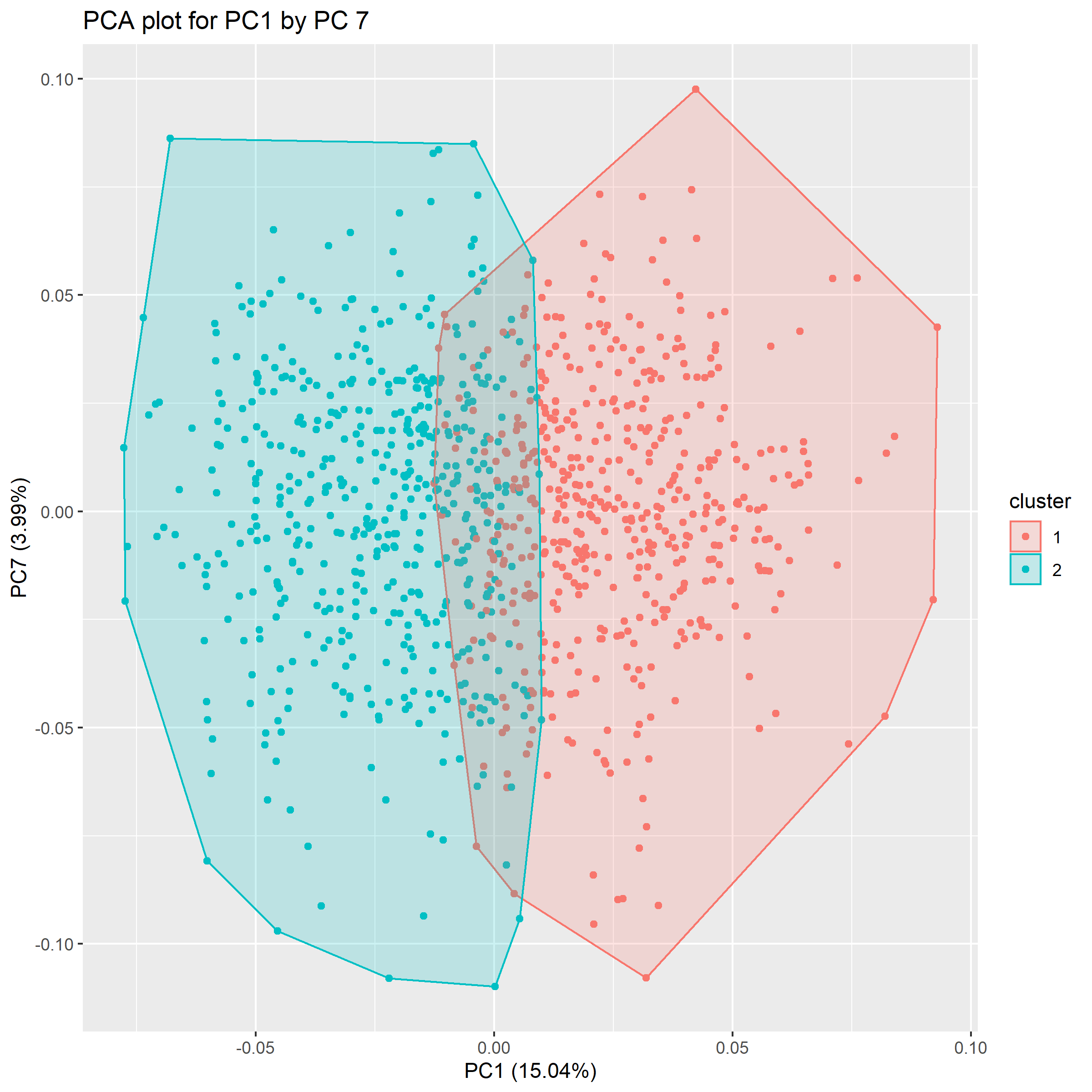
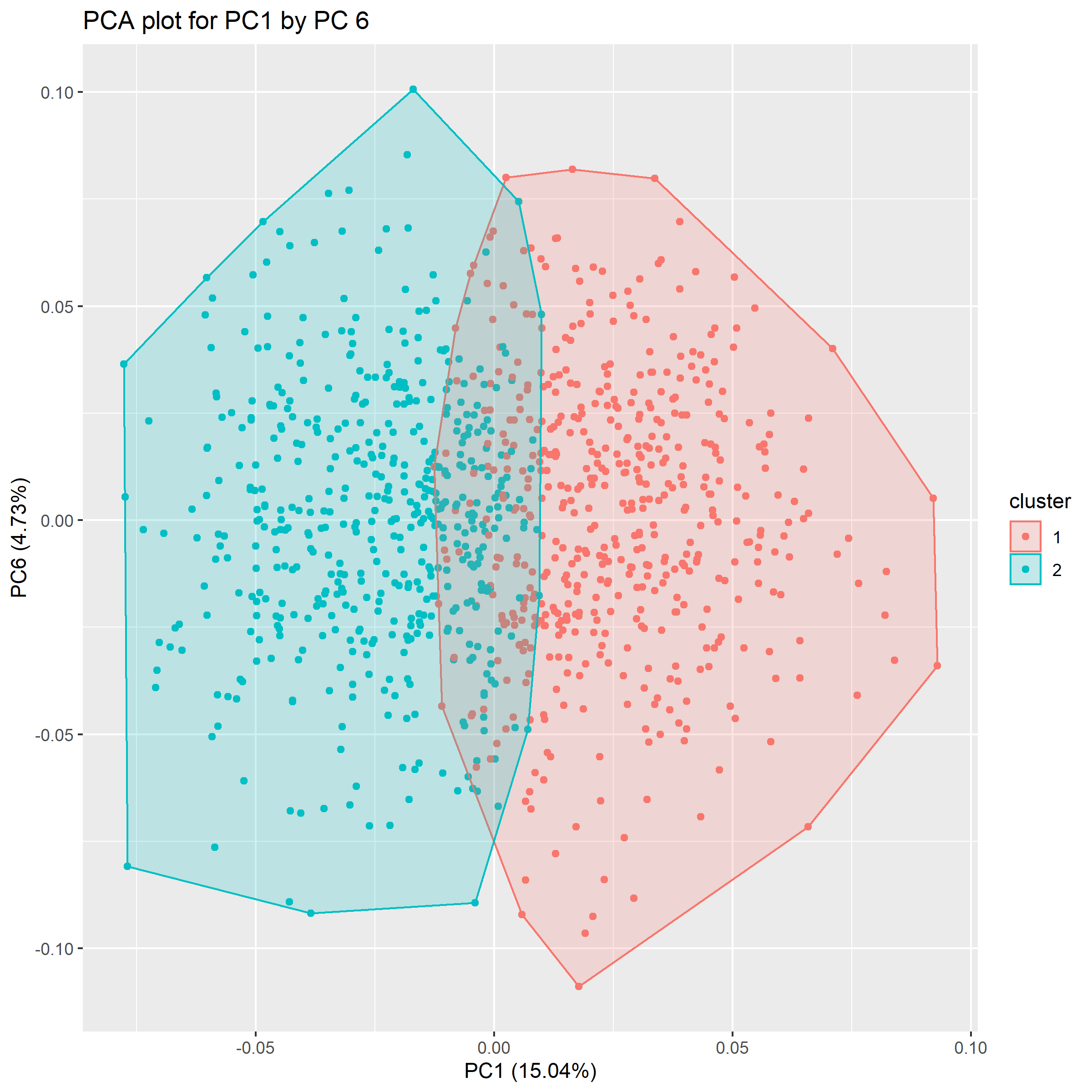
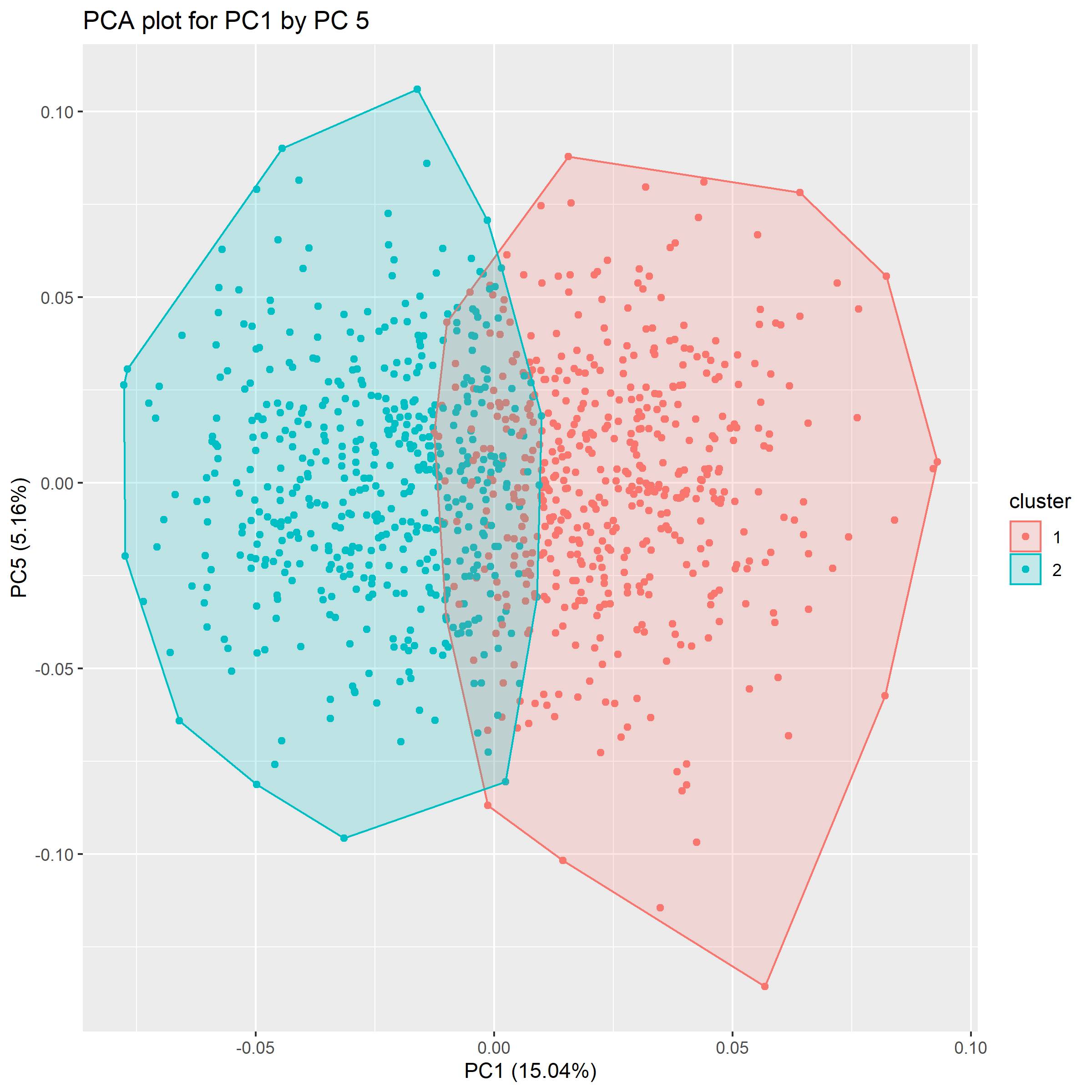
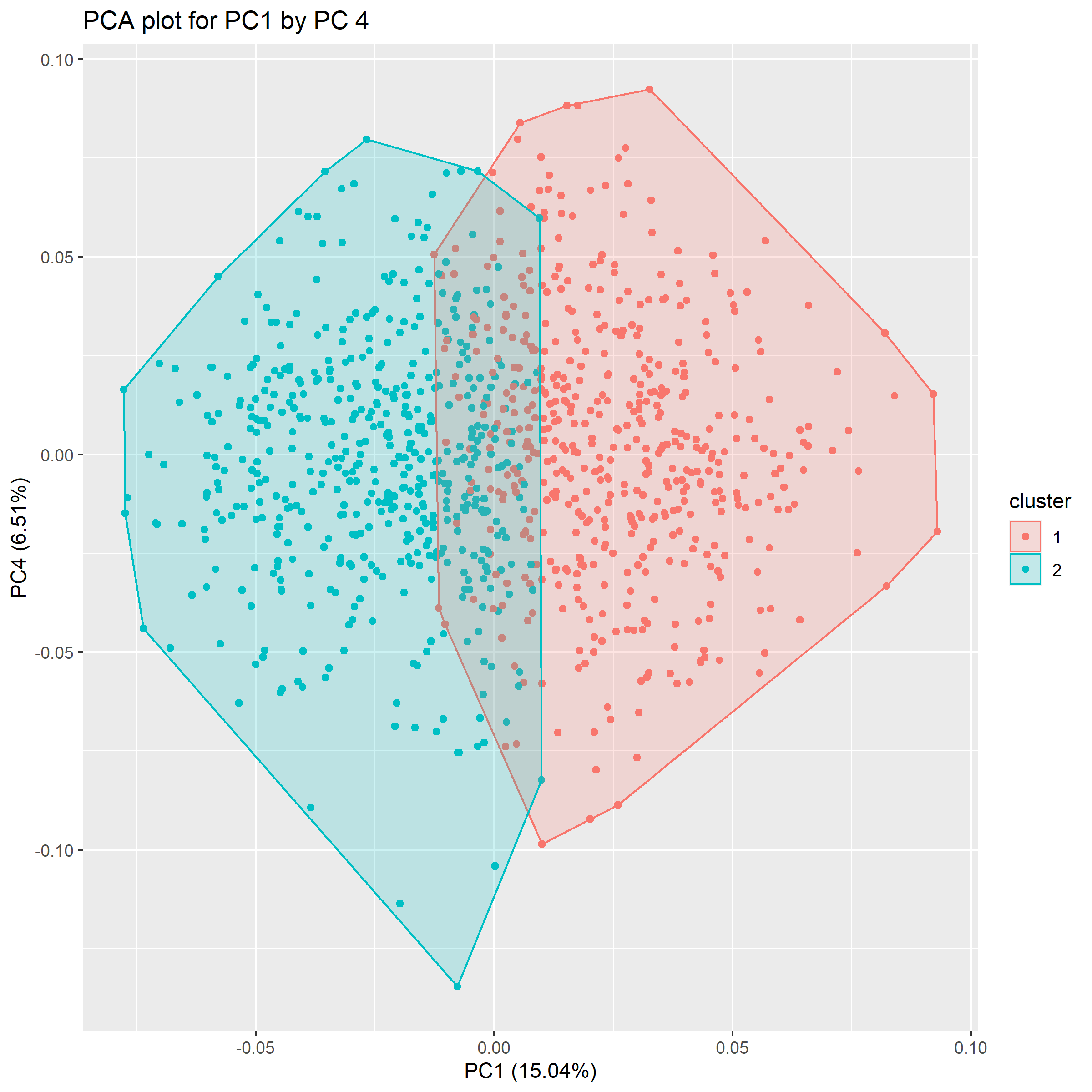
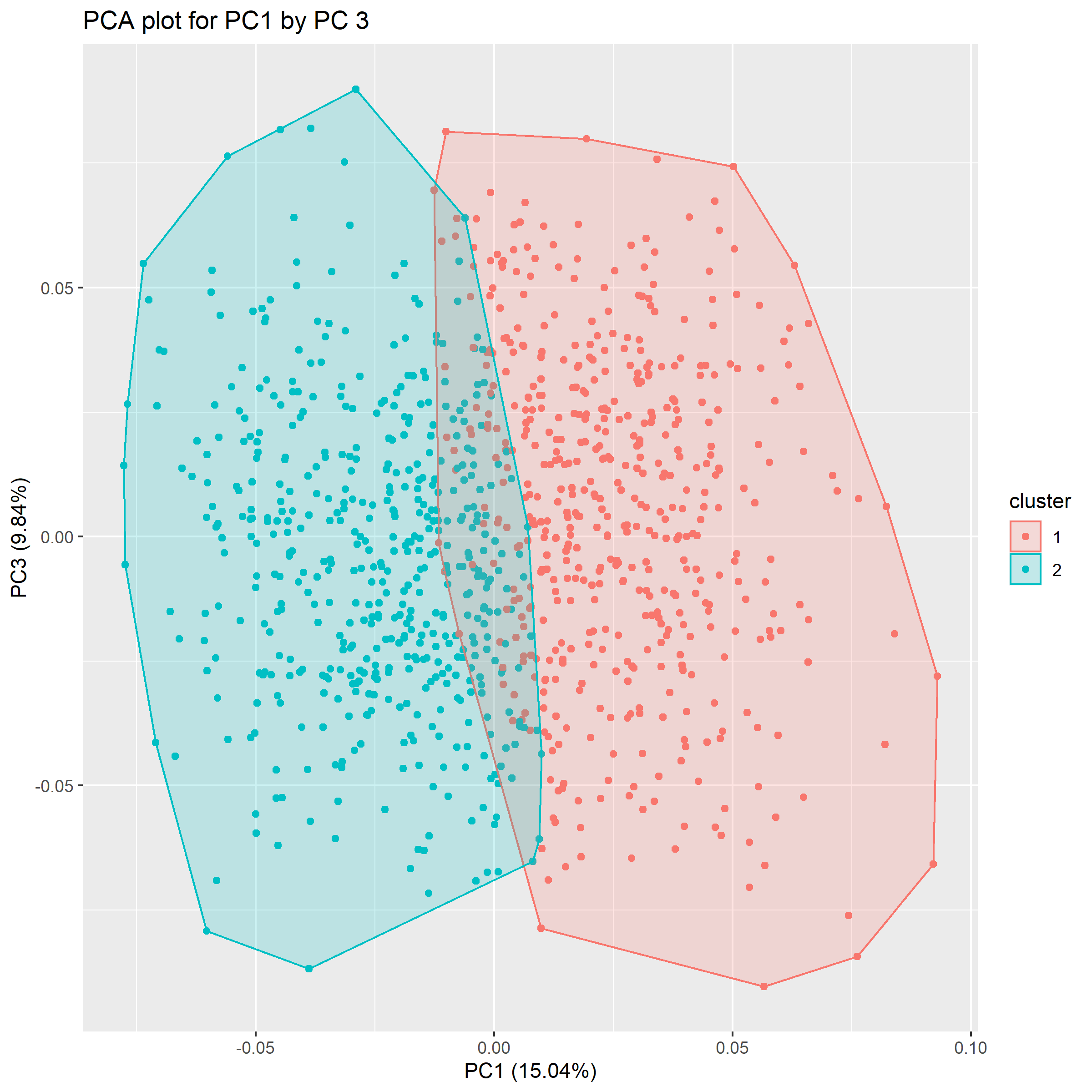
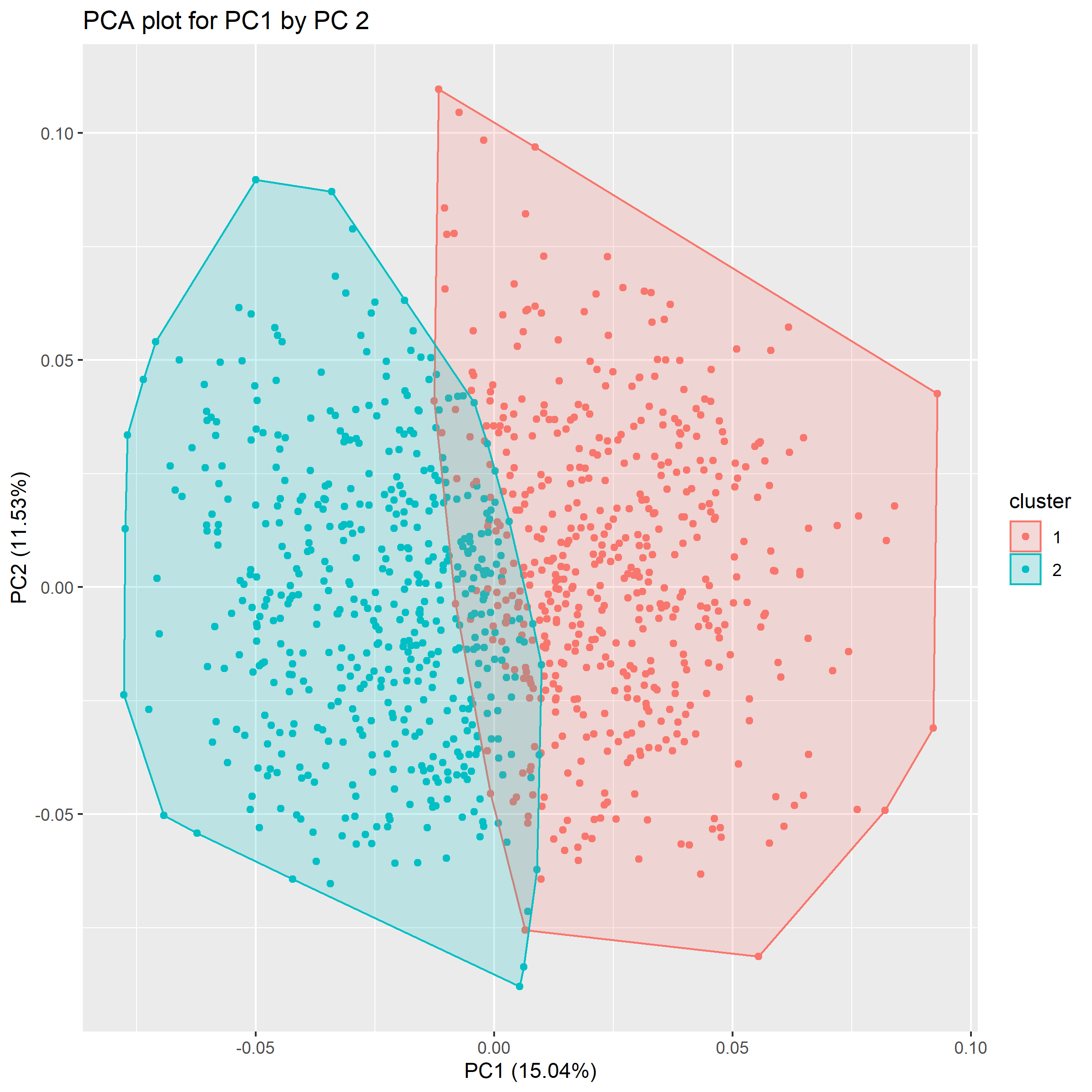
1 2

507 503

*Output 6: Shows the dimensions of the 2 clusters of which are approximately equal.*



*Figure 4: Pair plot of principal components show that PC1 is distinctly related to a single cluster while all other components are delegated to the secondary cluster. The lack of observable patterns in all other principal components suggest that half (output 4) of all young people enjoy or have a neutral opinion of pop and dance music (figure 2).*



*Figure 5: The overlap between clusters in the detailed PCA plots show that clusters are not uniquely different from each other. For all plots, overlap occurs along a central vertical showing that all components excluding PC1 are not determinant of the observations assigned cluster.*

> t(PC\_KMClust$centers)

1 2

PC1 2.053147881 -2.069475101

PC2 0.081534929 -0.082183318

PC3 0.204620766 -0.206247969

PC4 0.075777887 -0.076380495

PC5 -0.032965656 0.033227809

PC6 -0.003174637 0.003199883

PC7 -0.047174858 0.047550006

PC8 -0.036215176 0.036503170

PC9 -0.015621922 0.015746152

PC10 0.015270954 -0.015392393

PC11 -0.006510167 0.006561938

PC12 -0.044447876 0.044801339

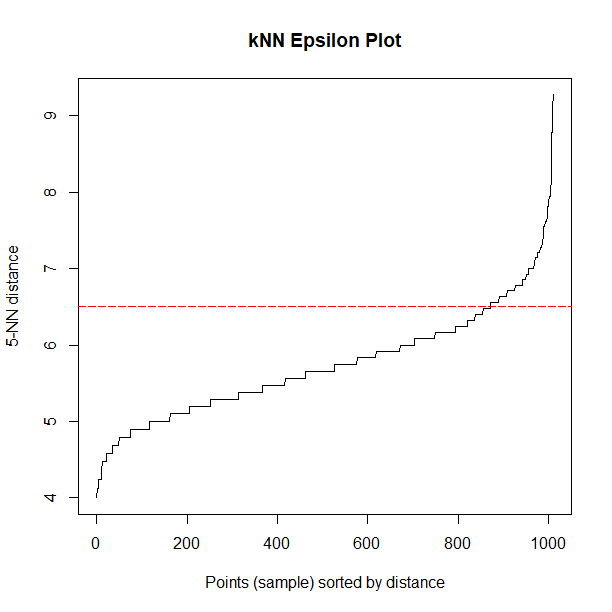
PC13 0.060687969 -0.061170577

PC14 -0.037901862 0.038203268

*Output 7: The table of principal component centres confirm the observations in figure 5 where k-means clusters all start from approximately 0 for all components excluding PC1 (centres at 2 and -2).*

## DBSCAN Algorithm:

DBSCAN is a hierarchical clustering algorithm that was chosen because its partitions can be more dynamically shaped in contrast to k-means. The optimal number of clusters for hierarchical algorithms are produced on their own.



*Figure 6: The epsilon plot has no clear bend point maximizing distance. Thus, its tangent point was approximated to 6.5. This epsilon value is required to determine whether observations within that distance can be clustered together.*

> MM\_db

DBSCAN clustering for 1010 objects.

Parameters: eps = 6.5, minPts = 2

The clustering contains 2 cluster(s) and 37 noise points.

0 1 2

37 971 2

*Output 8: DBSCAN output shows a large discrepancy between sizes of clusters. A secondary cluster is only formed when the minimum points required is reduced to 2 of which is too small to appropriately be considered as a cluster. Hence, DBSCAN produces only a single cloud-shaped cluster excluding noise/outlier observations.*

*Table

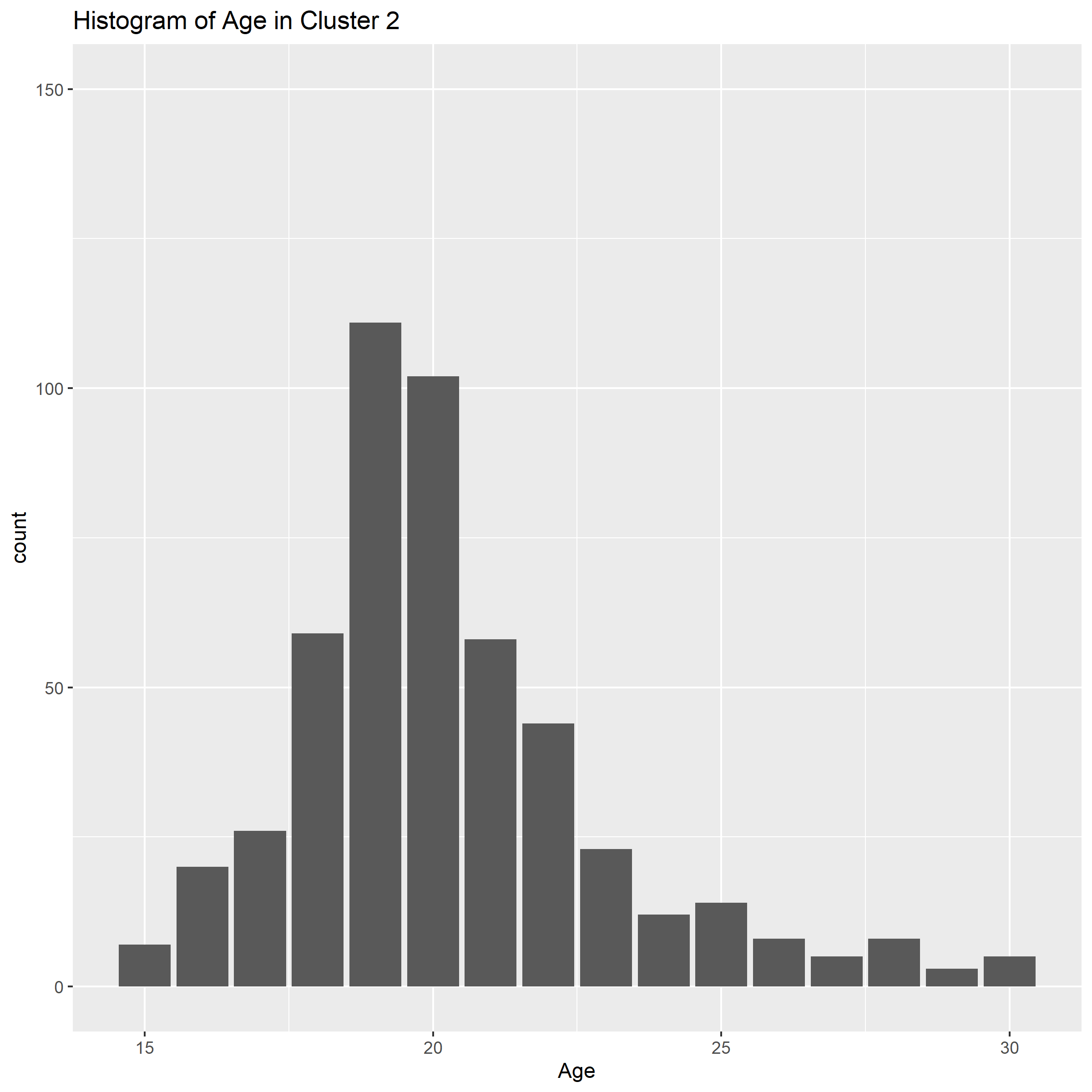
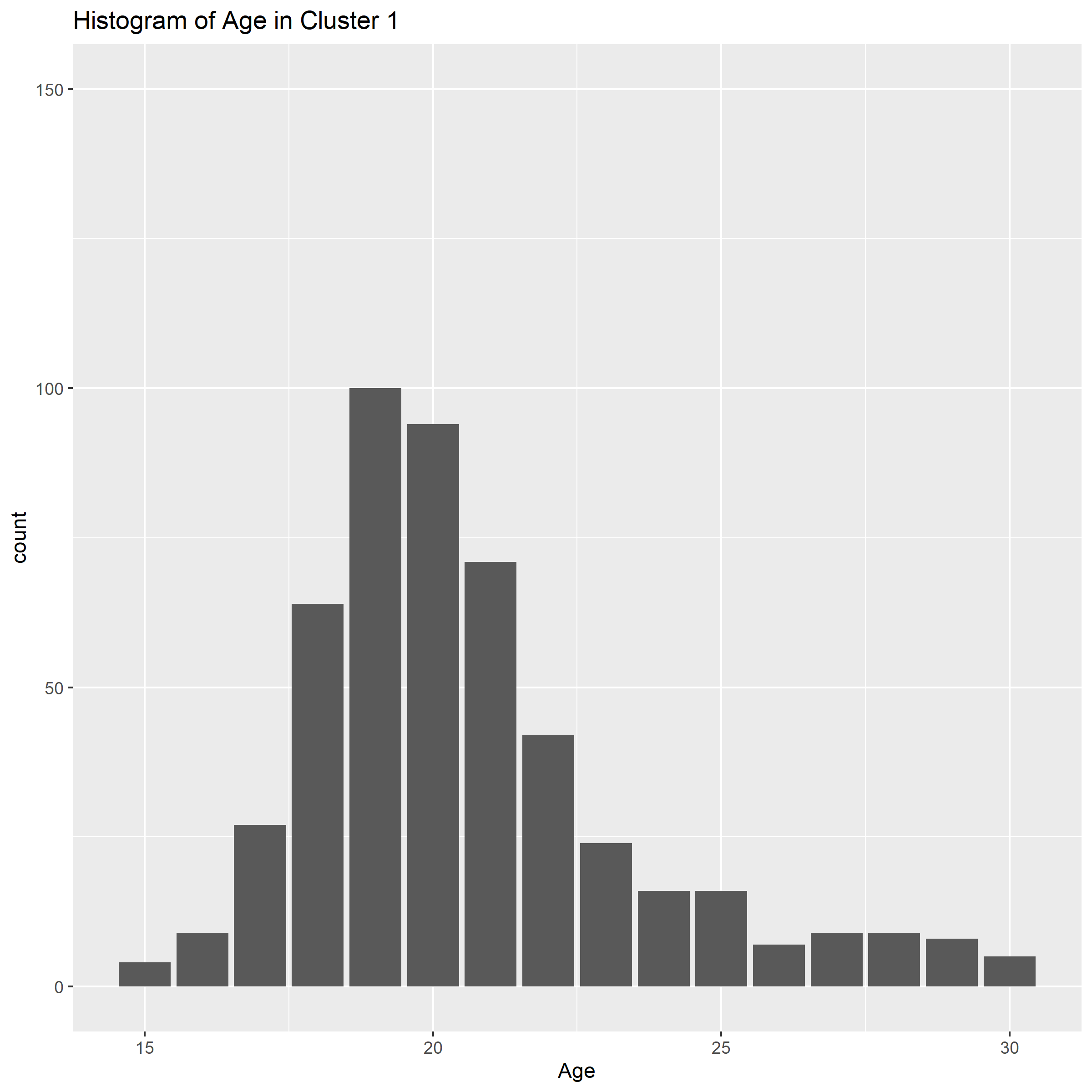
Description automatically generated with medium confidence*

*Figure 7: Pair plot confirms output 6 where for all principal components, DBSCAN produces only a single cluster. There are no observable patterns nor insights that can be determined.*

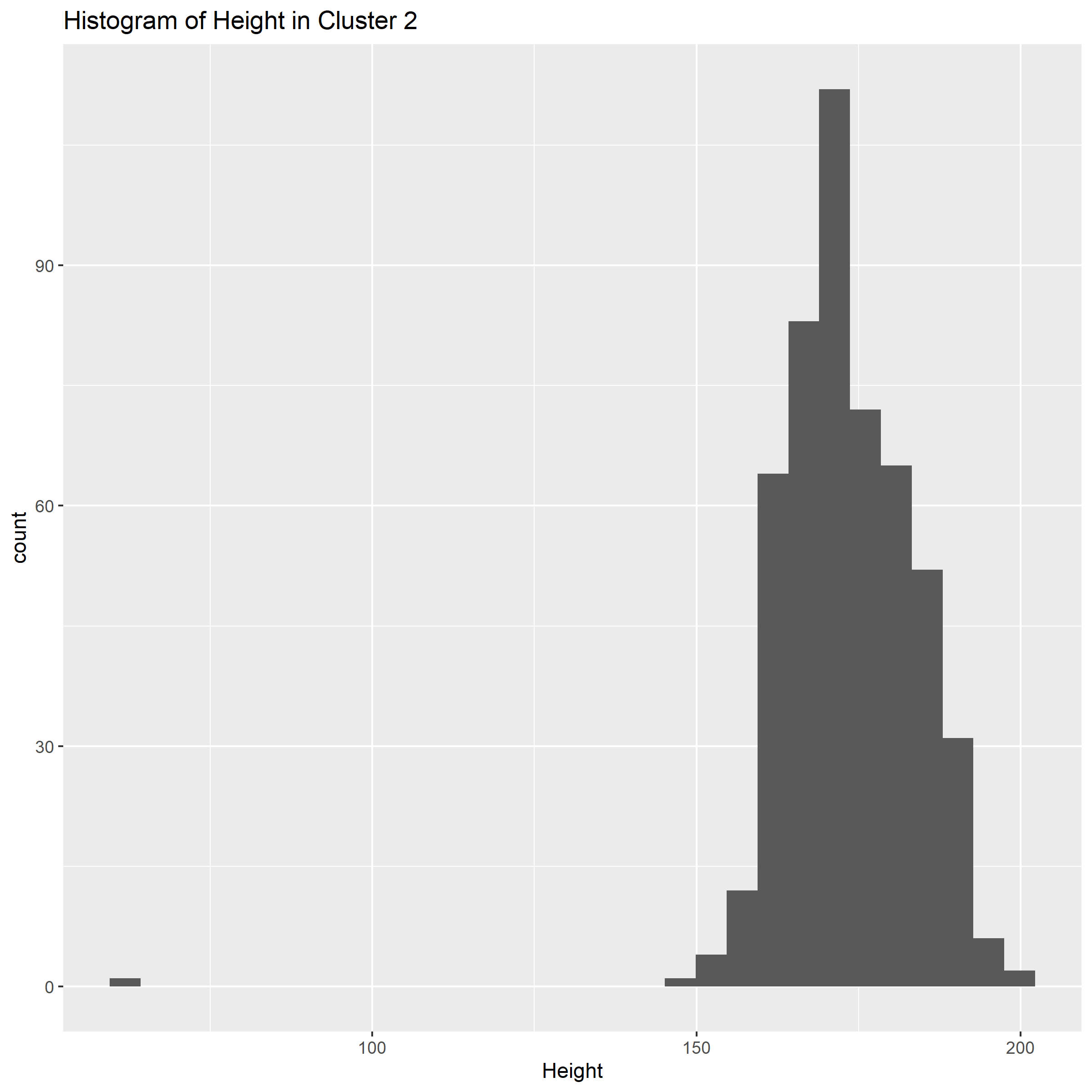
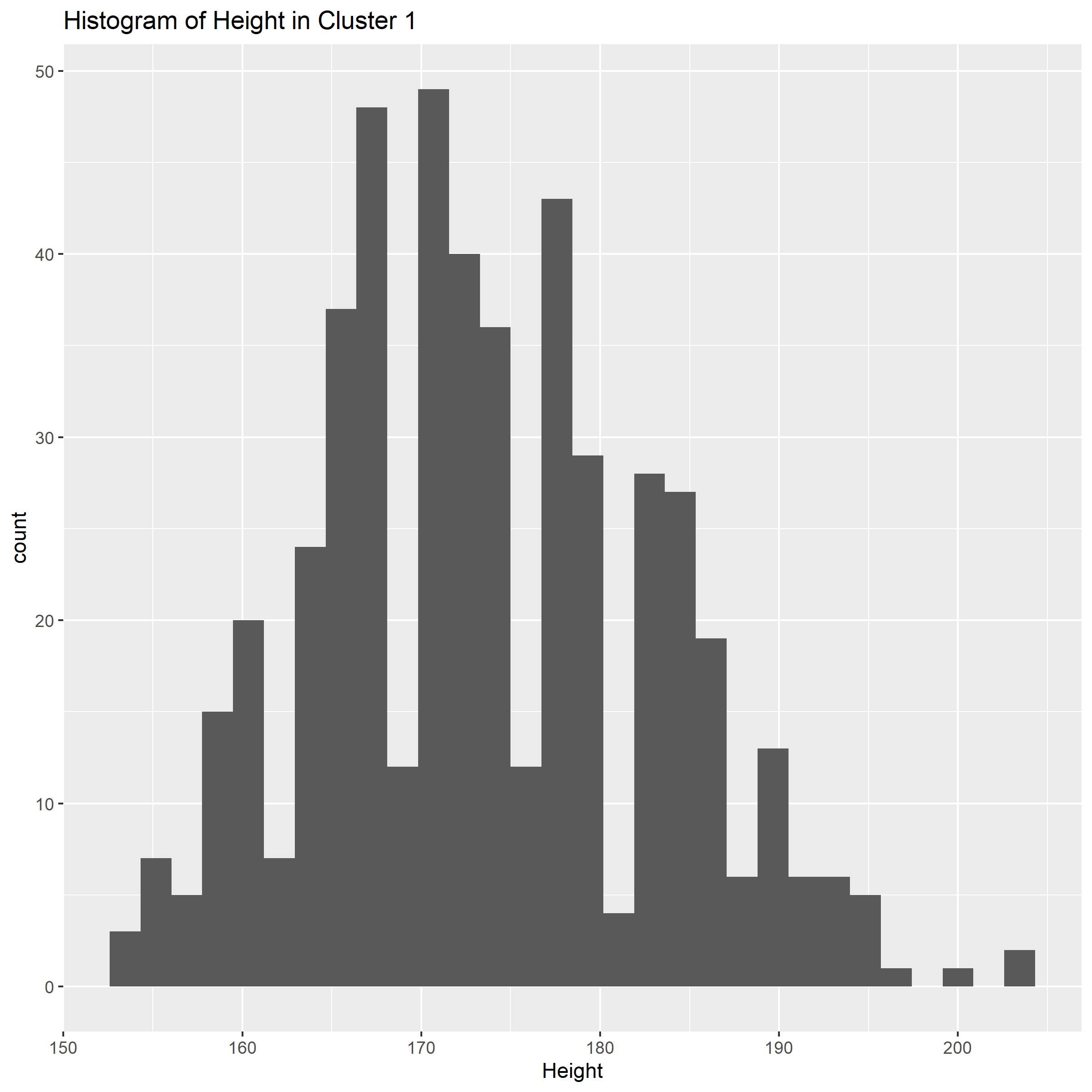
# Profiling:

Further data manipulation was made to combine clusters with music and moves preferences along with demographics information. Only k-means clusters were analysed against each other because DBSCAN produced a single cluster (nothing to compare). The purpose of this section was to identify if certain demographics of young aged individuals were related to any of the identified k-means clusters.

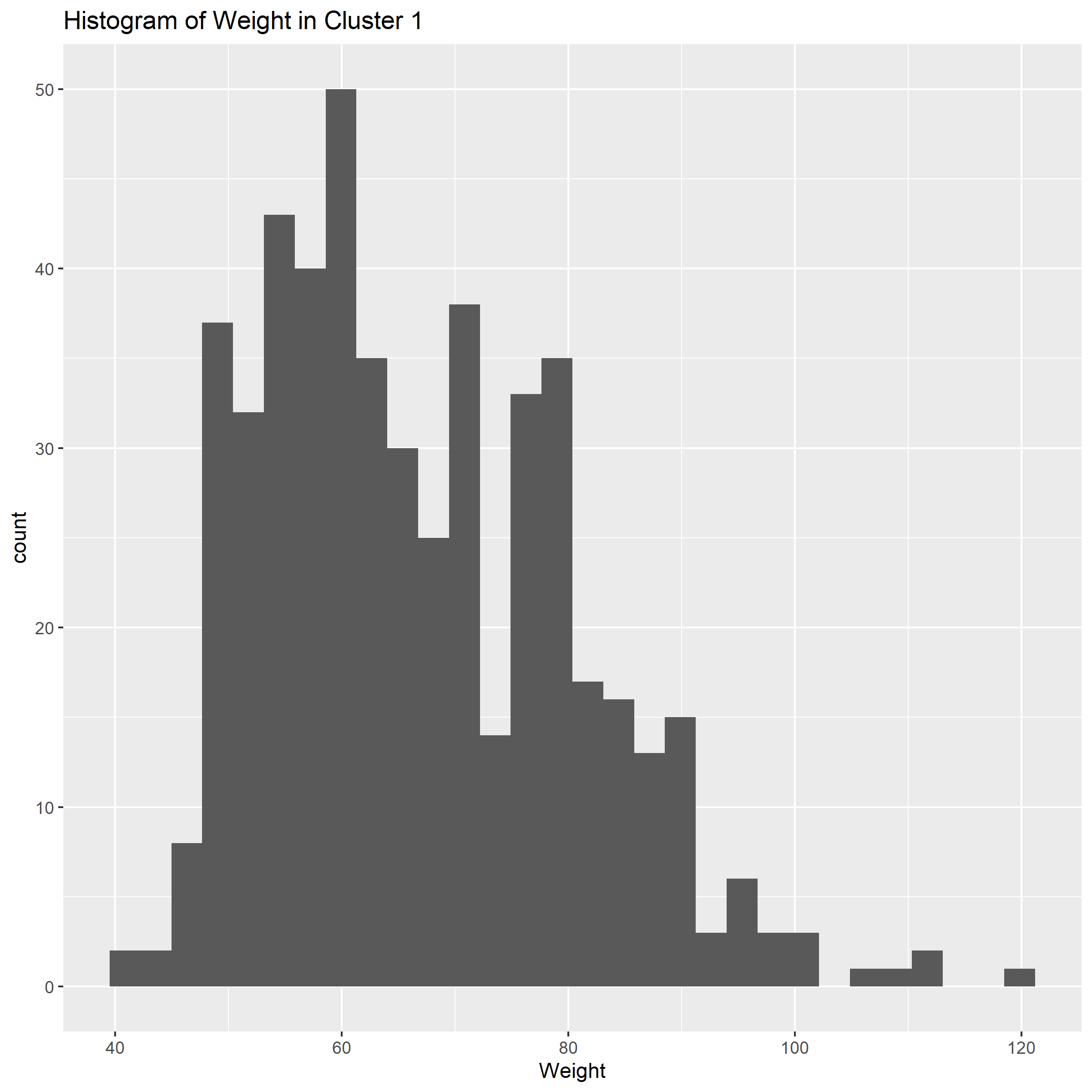
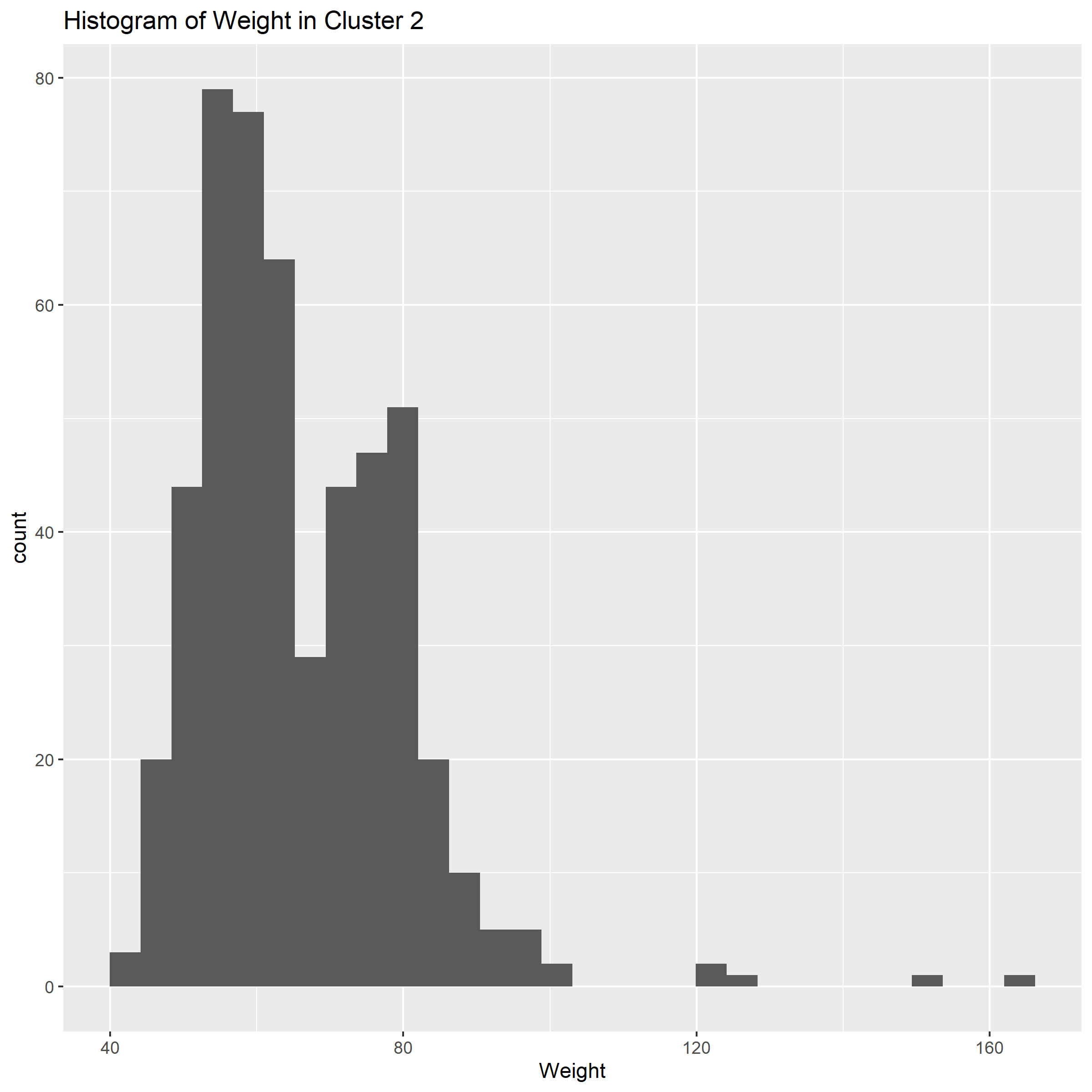
## Demographic Analysis of K-means Clusters



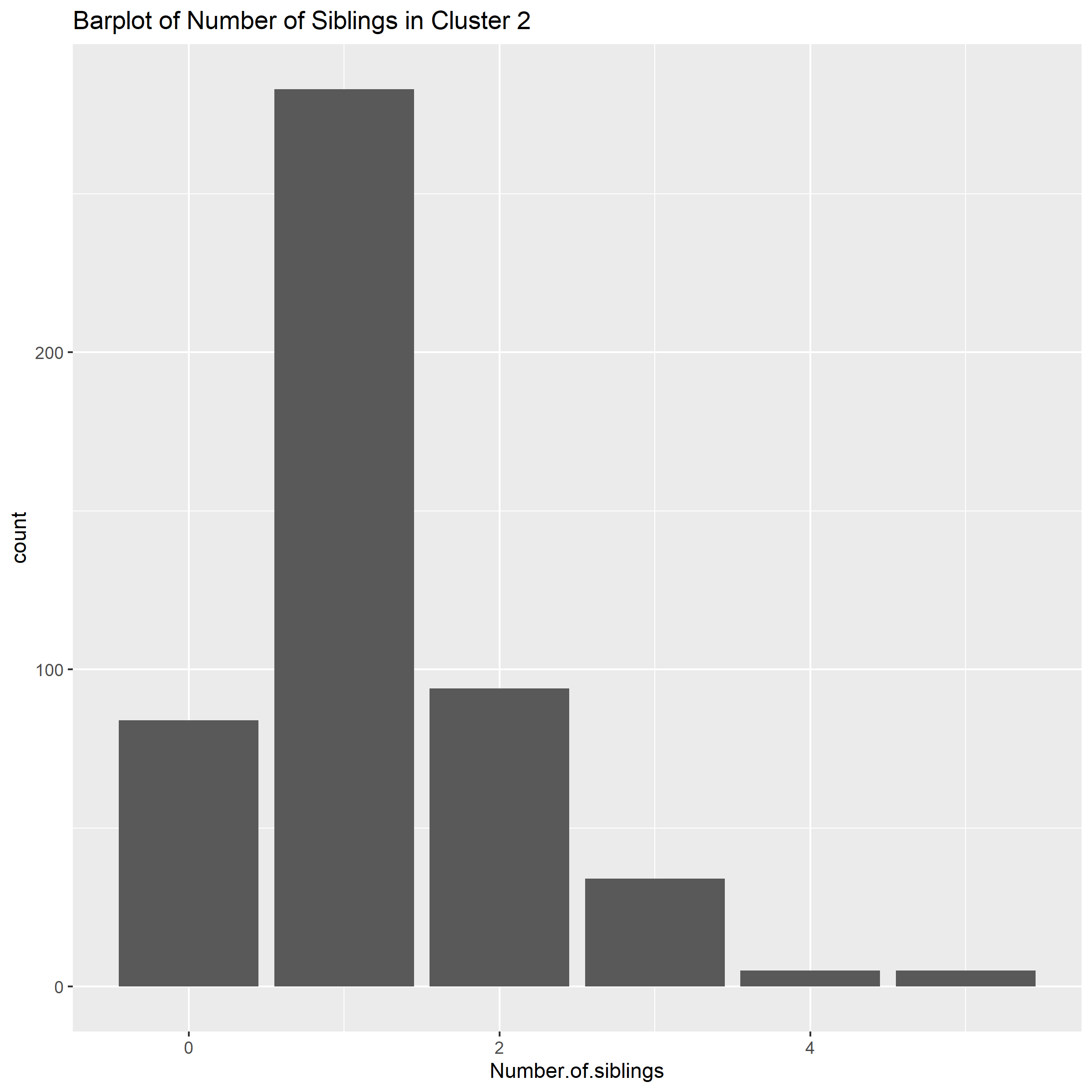
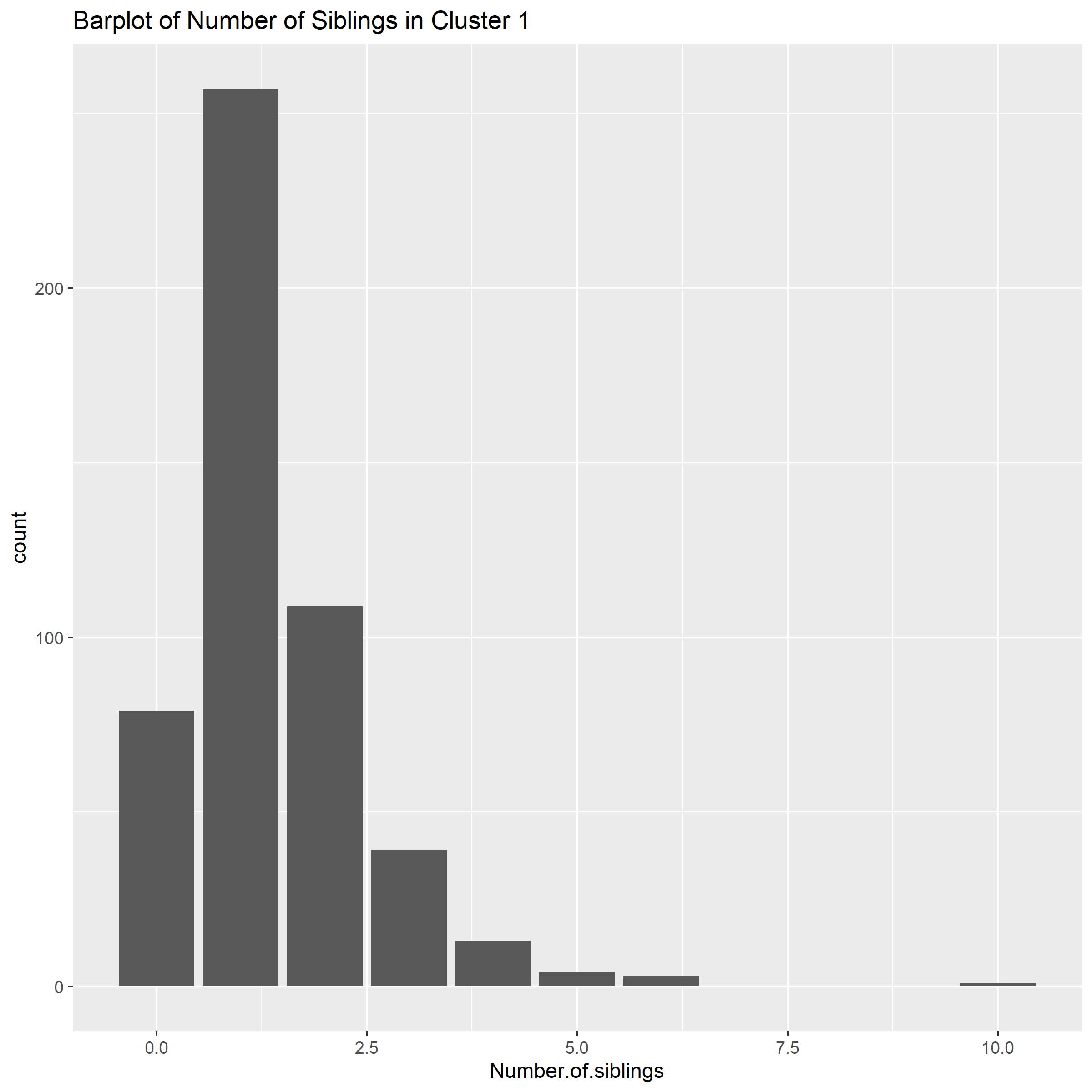
*Figure 8: Histogram of differences in age between clusters shows that both clusters are normally distributed with medians at 20 years old.*

**

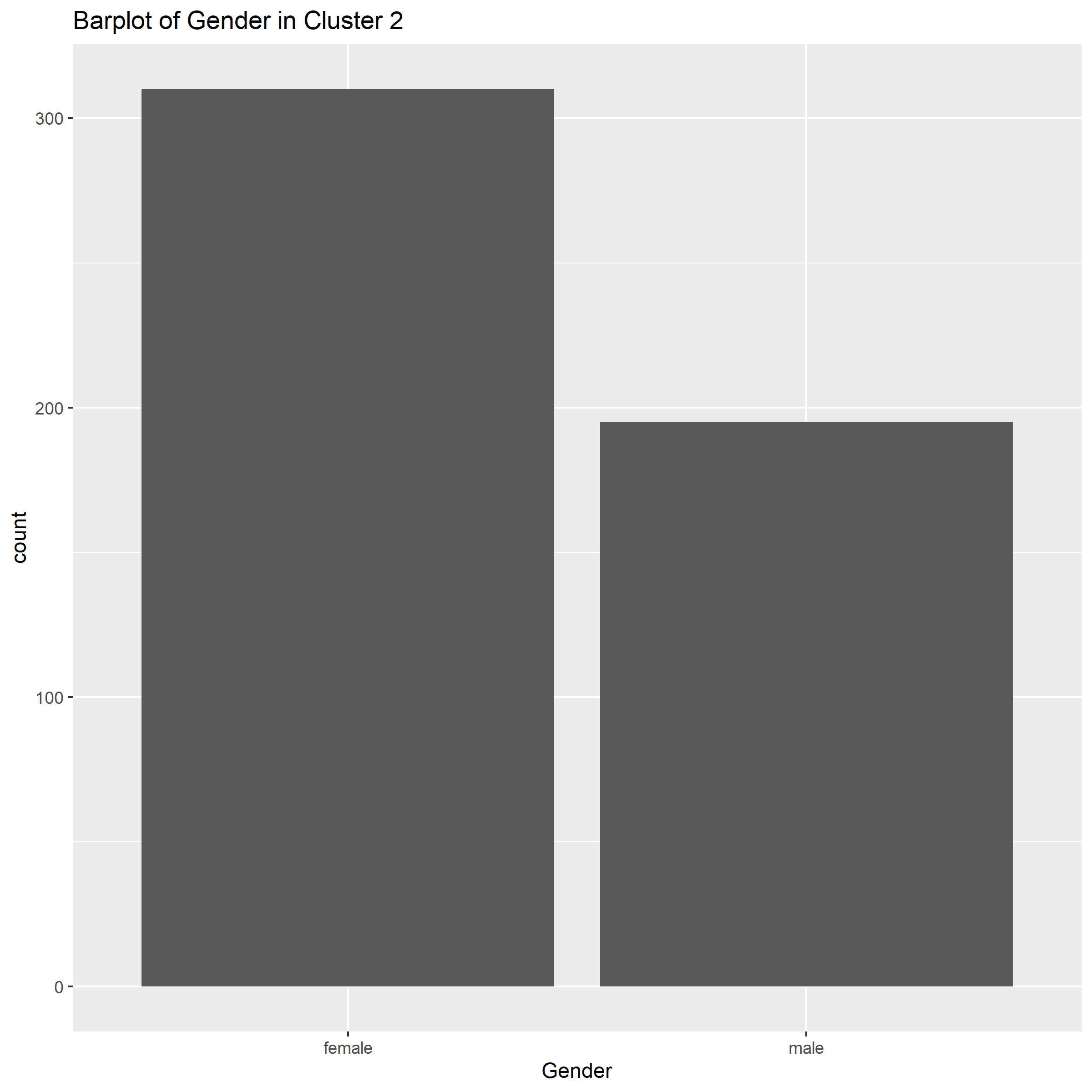
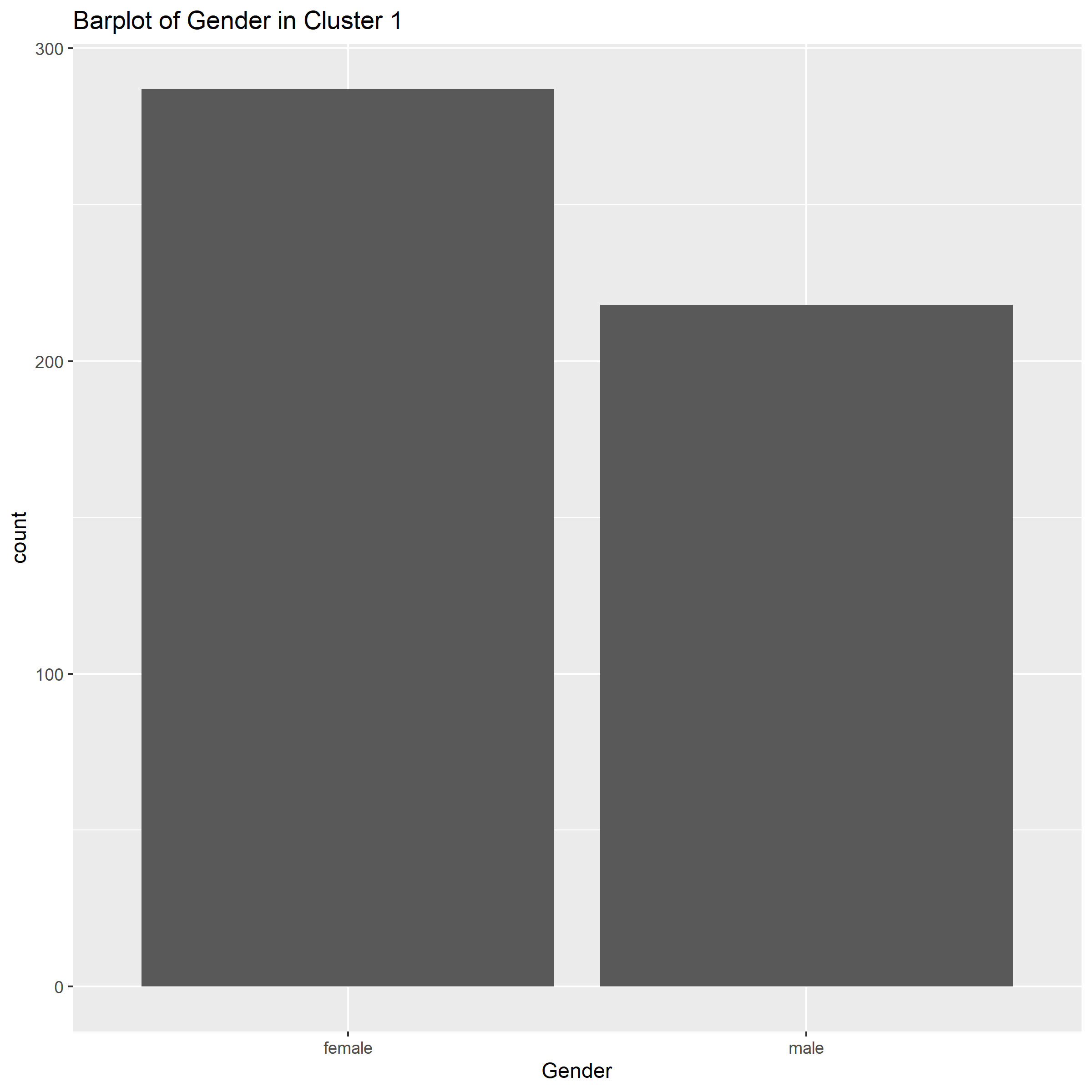
*Figure 9: Histogram of heights shows a single outlier in cluster 2 with a height of 65cm. Otherwise, both clusters are normally distributed at around 175cm.*

*­­­*

*Figure 10: Two peaks of majority weight groups distributed about 60kg and 80kg. General distribution of weight between clusters appears similar.*

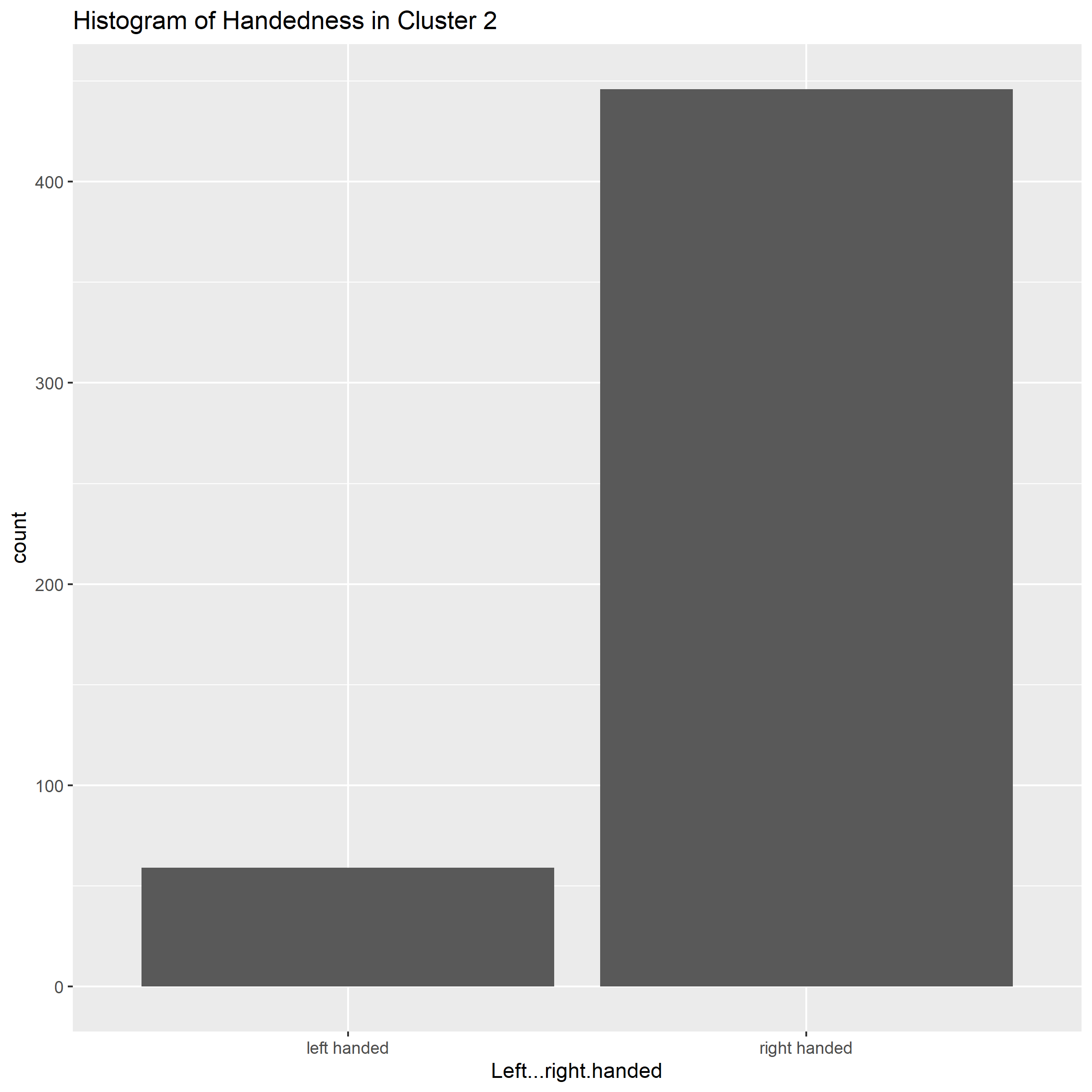
**

*Figure 11: Similar distribution of siblings between clusters with most respondents having only a single sibling.*

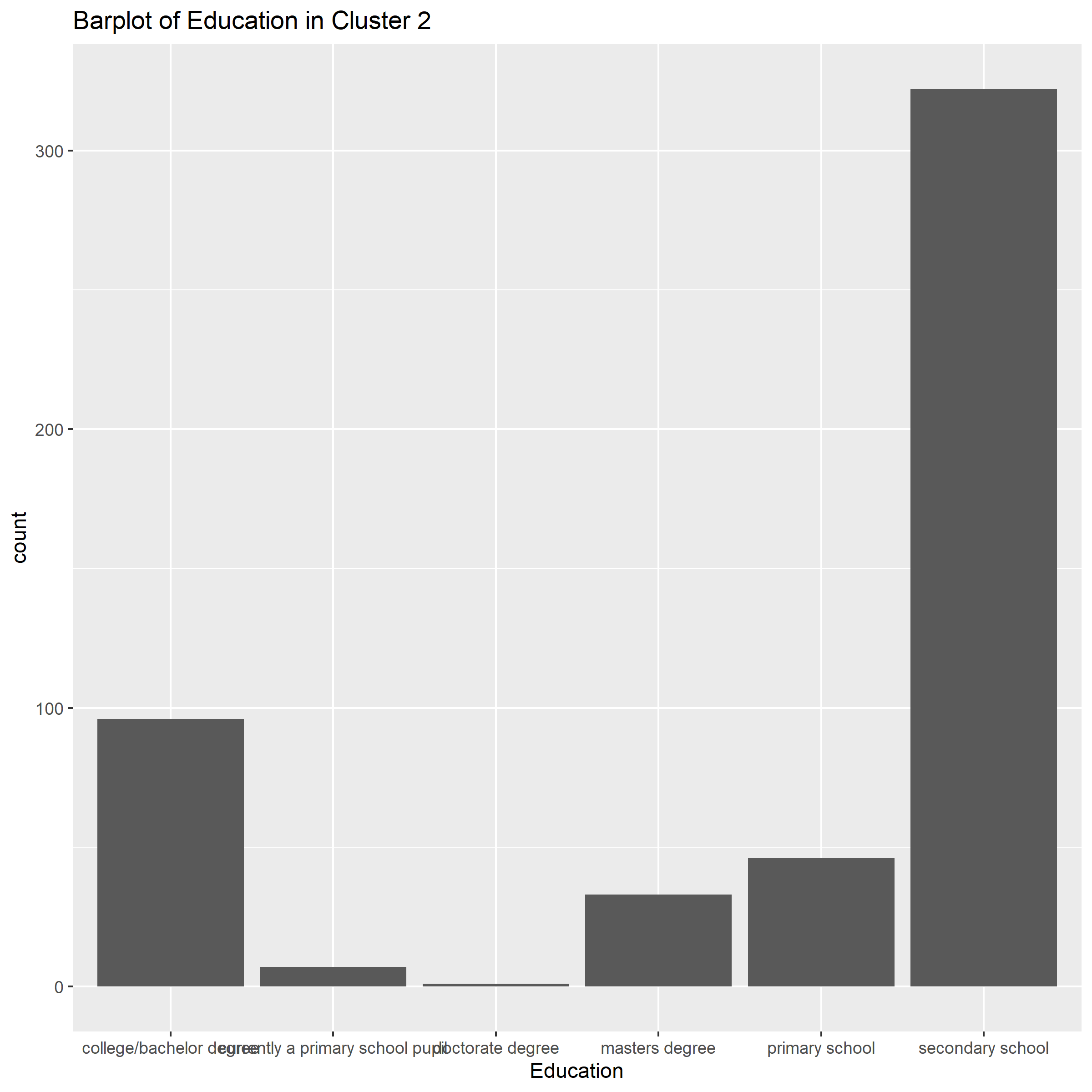
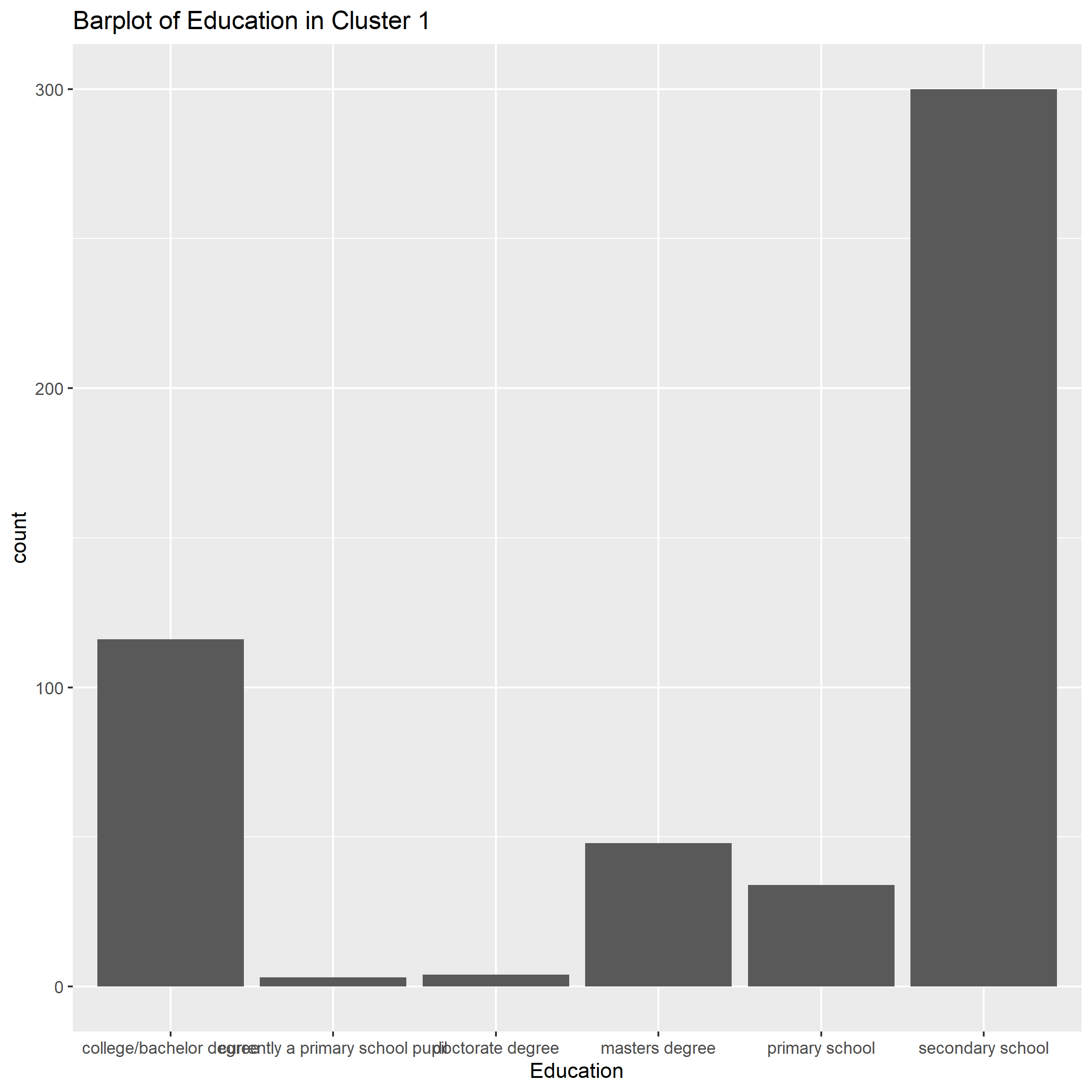
**

*Figure 12: It can be observed that there is a slightly smaller difference between frequency of males and females in cluster 1 compared to cluster 2. Assuming the plot is not influenced by sample bias, it can be inferred that the popularity of pop and dance music (figure 2) is more evenly distributed among genders according to k-means clustering.*

*Chart

Description automatically generated*

*Figure 13: Distribution of handedness is generally similar between clusters.*

**

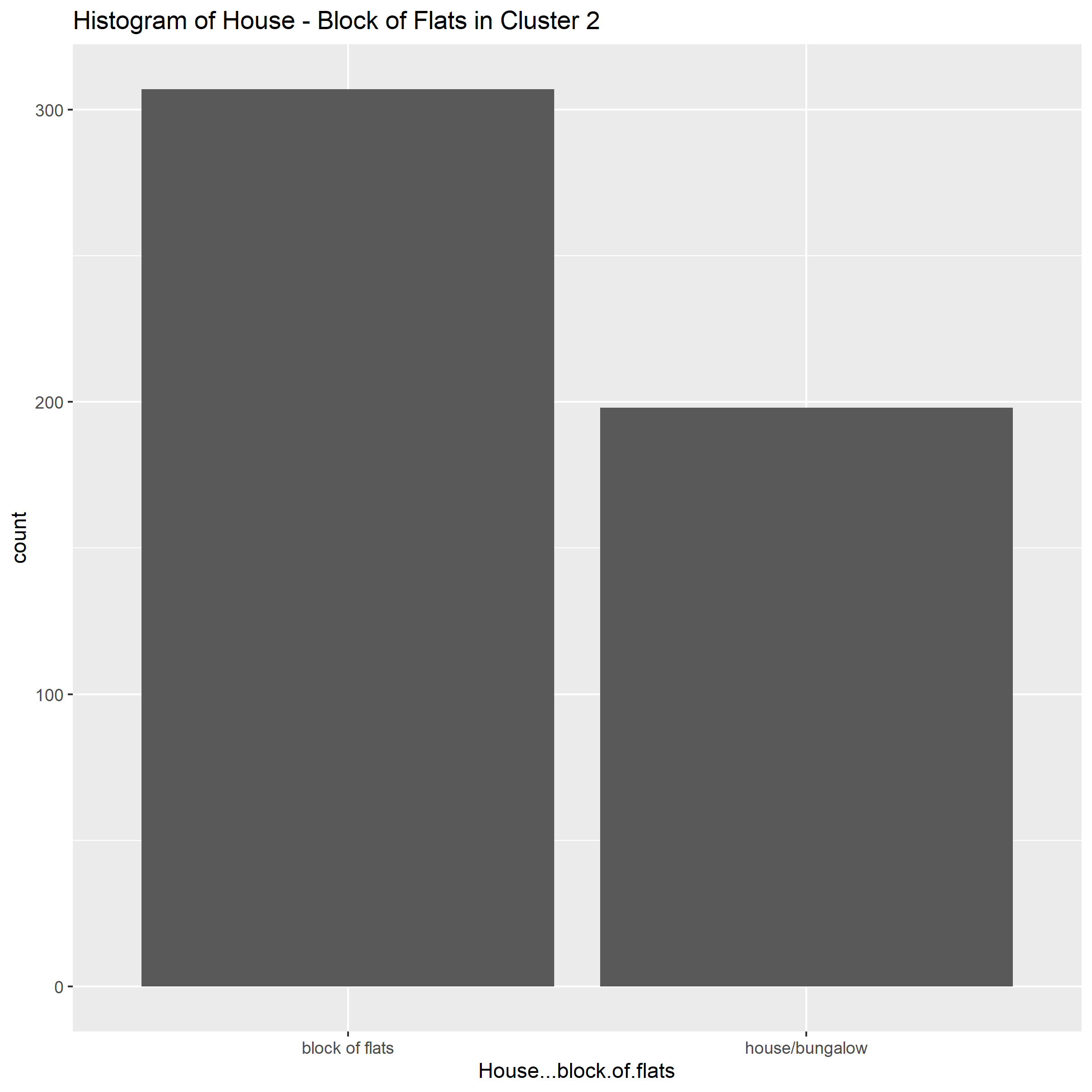
*Figure 14: Cluster 1 has fewer people whose highest level of education was primary and primary school pupils compared to cluster 2. However, the sample size of people aged 15-30 who have only achieved this level of education makes up about 10% of the data (approximately 100 observations) and may be subject to bias.*

*Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated*

*Figure 15: Environment of city or village does not appear to change between clusters.*

*Chart, bar chart

Description automatically generated*

*Figure 16: Slightly more people in cluster 1 who lived in a block of flats. However, distribution appears mostly similar.*

# Conclusion:

## Comparative Analysis:

The results produced from k-means and DBSCAN algorithms produced greatly differing results. DBSCAN results determined that there were no distinct partitions in the data where the whole dataset (excluding some outliers/noise) produced a single cluster. Cross-checking the k-means results shows that aside from some minor differences, the distribution of demographics remains generally the same across clusters. Hence, results from both k-means and DBSCAN suggest that no valuable insights of music and movie preferences of young individuals can be produced using these 2 algorithms.

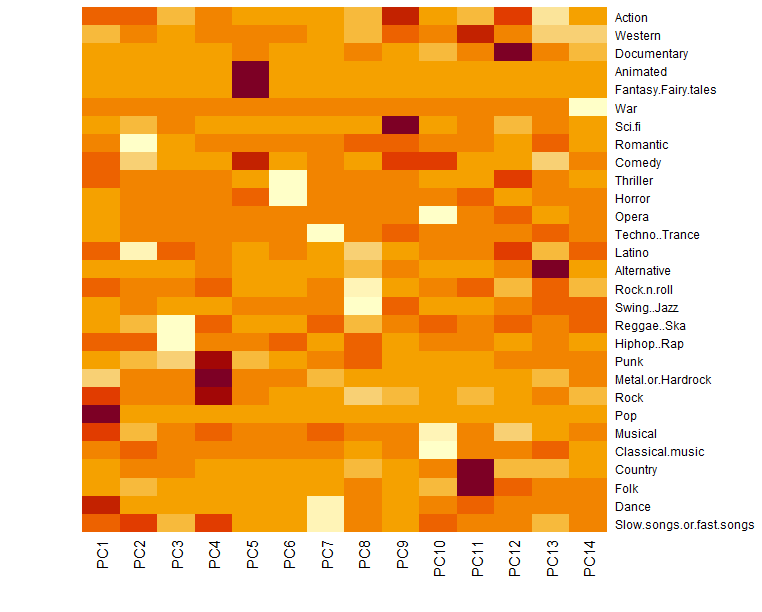
## Critique:

The research process of this report may have been overly methodical, further experimentation with cluster numbers (k-means) and epsilon values (DBSCAN) may have been more appropriate in contrast to the performed systematic method. Additionally, alternative clustering algorithms could have been further researched before being implemented.

## Position:

It can be concluded that k-means and DBSCAN clustering algorithms were not appropriate for young people survey data. Alternatively, the dataset may not have any observable clusters.

# Appendix:



*Figure 17: PC-Attribute correlation heatmap*