Introduction

The App Happy Company wishes to understand what they think is the market for a new social application. In order to understand the customers in the market better, they hired the Consumer Spy Corporation (CSC) to survey consumers in the market. They collected data from a sample of consumers and provided App Happy with a data set of the responses.

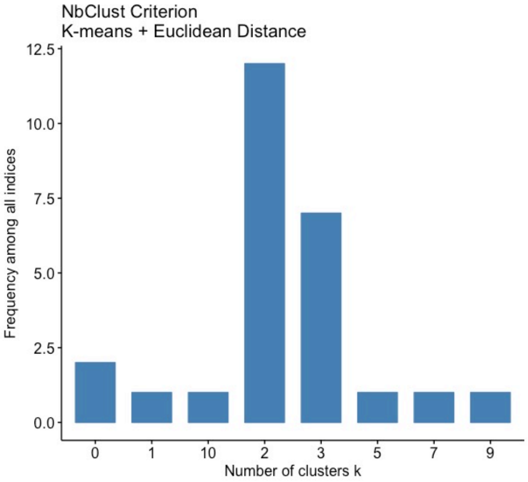
Market segmentation

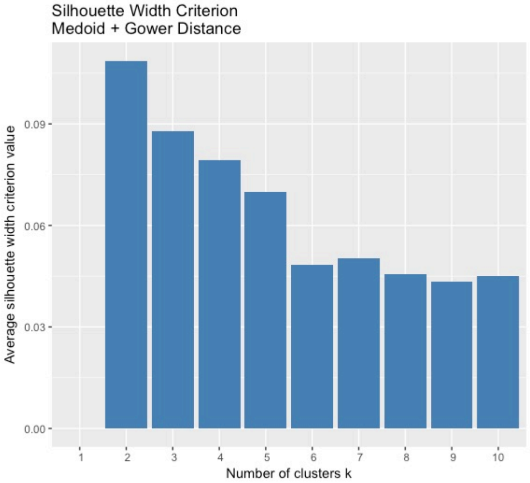
To conduct the general attitudinal post hoc segmentation analysis, we need to find attitudinal item variables to make up the basis variables to be used for the assessment. These variables would be able to characterize the items ang groups based on attitudes. Looking at the survey data, questions 24, 25, and 26 would make the best items for basis variables.

There were a couple of different methods that were used to conduct the post hoc segmentation analysis. The first method that was used was the clustering method. The desired number of clusters is determined in advance to set the number of cluster centers. The data points are assigned to its nearest cluster center by minimizing and maximizing a desired criterion. For this method, the relationship between the clusters stays fairly undetermined. Both non-hierarchical clustering and hierarchical clustering methods were used to segment by the set of attitudinal variables.

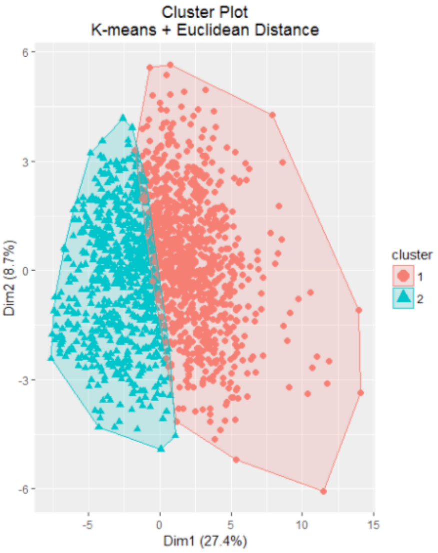
Each of the basis variables will employ a six level Likert scale ranging from Agree Strongly to Disagree Strongly. Each variable can be considered to be a categorical type. The daisy function in the cluster R package can compute distances using the Gower distance metric. Cluster analysis on the assumption of continuous data using the Euclidean distance metric and a comparative analysis assuming non-continuous data using the Gower distance metric will be conducted.

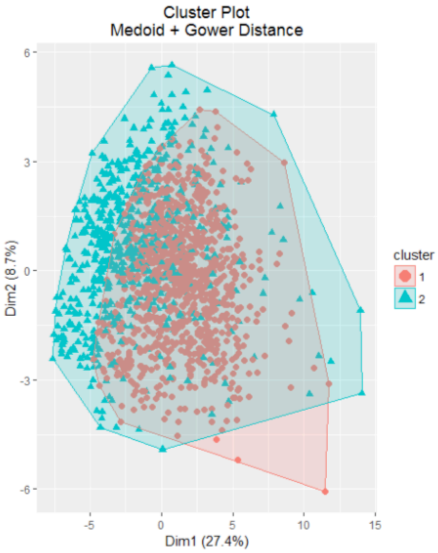
The NbClust R package will allow us to employ up to 30 cluster criterion measures as a form of index. The function will be applied to k-means clustering using the Euclidean distance metric and show index frequencies for a range of clusters below.



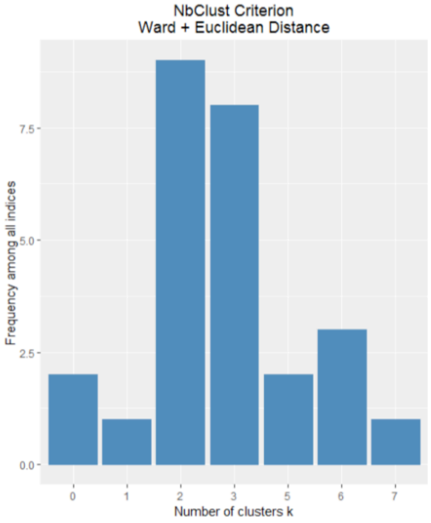


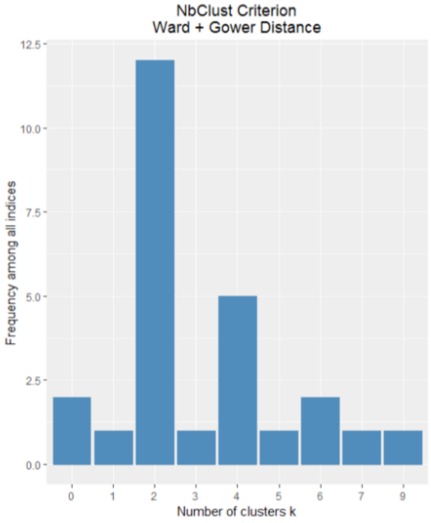
The use of the two clusters is reported to be the best for both clustering methods, while the use of three clusters is reported to be the next option. The number of clusters are pre-determined based on the criteria and we can do the non-hierarchical clustering method next. Below are the non-hierarchical cluster plots. In the plots, the cluster boundaries are determined by connecting the most extreme data points for the clusters. The medoid cluster technique has a much greater overlap of clusters than the k-means technique. The overlap of cluster points and inability to distinguish clusters is a disadvantage for the Medoid cluster.



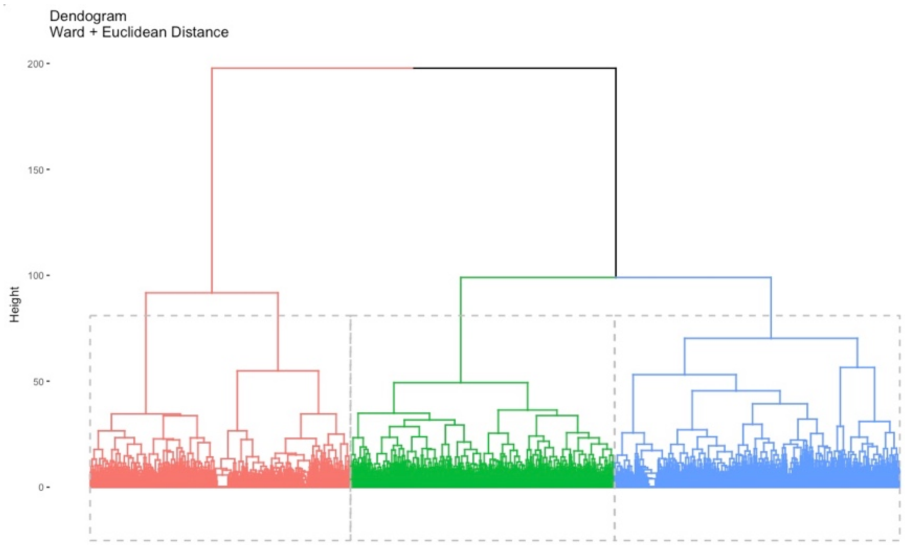


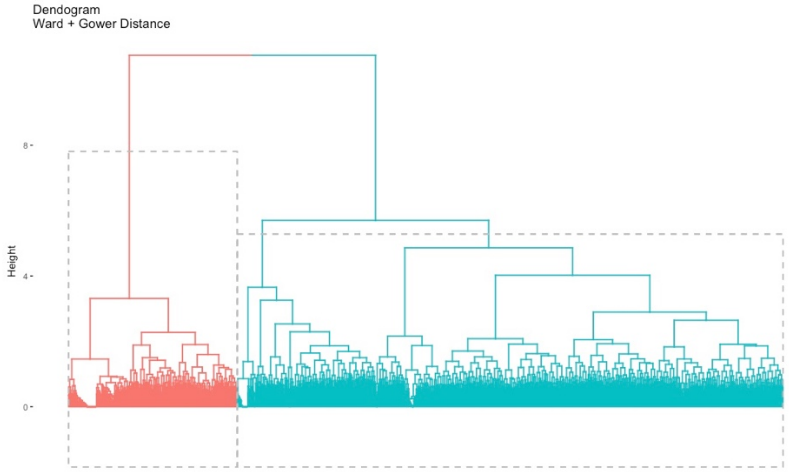
The next clustering is the hierarchical Clustering. For this cluster method, we followed a similar process as above to keep a standard in analysis. The pre-determined number of clusters will aid in interpretation of the plots below. The NbClust function was also used as well as the ward clustering technique using both Euclidean and Gower Distance Metric.



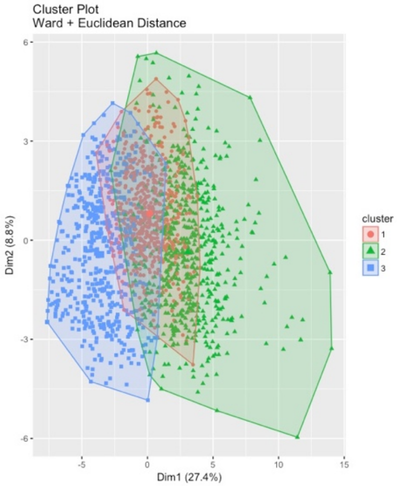


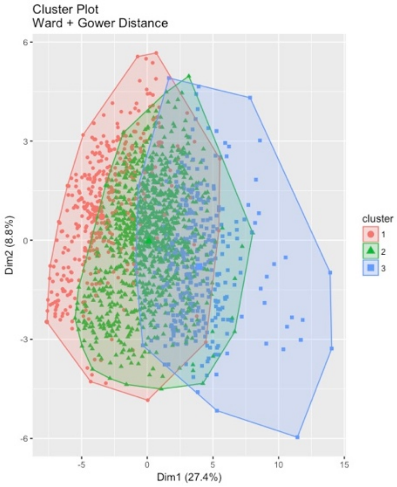
The use of three clusters is reported to be optimal when using the Euclidean distance metric on the left. When the Gower distance metric is used, the use of two clusters is reported as being optimal. Dendograms for both metrics are shown below. Deciding what is the optimal cut is a decision we will have to make. For the graphs below, the assessment of three clusters is appropriate under each distance metric. Making a cut after the second Clade when using either metric results in data points being attributed evenly between the clusters.





Below are the scatter plots highlighting the hierarchical cluster assignments based on two-dimensional reduction of our original set of attitudinal survey responses. These gras show a large amount of overlap between the clusters. Assigning the three clusters under the Euclidean distance metric makes it extremely difficult to distinguish the second cluster from the remaining two. According to the Gower distance metric, it is hard to distinguish the first cluster from the third.





Generalized linear models were also applied to each question based on each clustering method. Each model was measured on the AIC value. We are looking for the lowest AIC value to find the better clustering method for predicting each cluster. An assessment of the AIC values shows that each clustering method tends to demonstrate low AIC values for questions two and four and high AIC values for questions one, and 56. The average AIC value for each clustering method allows us to understand if a particular method is better for predictive accuracy. The summary table can be seen below. The Ward clustering technique with Euclidean distance metric has the lowest average AIC value with the k-means technique with the Euclidean distance metric having the second lowest AIC value.

AIC

```
k-means + Euclidean      Medoid + Gower   Ward + Eucildean         Ward + Gower
           4359.223            4452.868           4469.116             4390.677
```

Chi

```
k-means + Euclidean      Medoid + Gower   Ward + Eucildean         Ward + Gower
       0.0004997501        0.0004997501       0.0004997501         0.0004997501
```

Segment Profiles

Segment profiles have been able to be created from the previous analysis. We were able to use visual methods to evaluate both the demographic and consumer behavior characteristics of the clusters determined by the selected method. There were a number of similarities over demographic characteristics for both clusters. For example, we found that there were significantly more respondents who were under the age of 35, college graduates, who were white, married, had no children and had an annual income of $30,000 to $70,000 among both clusters. The clusters were distinguished by many demographic characteristics. The first cluster had a greater number of younger respondents who were more educated, were Caucasian, or had a higher income. This same cluster was more likely to have been married, divorced, or to have older children than those from cluster two who also had children.

There were also a lot of similarities over consumer preferences for both of the clusters. The majority of respondents owned an iPhone, used gaming or social networking applications, had over 10 applications on their chosen device, or they acquired at least half of their device applications for free. The differences in the clusters found that respondents from the first cluster had a preference for iPhone over Android devices. However, this cluster had less of a preference for iPods and tablet devices in general. The first cluster had less of a preference for entertainment or television applications. This same cluster had a larger bias towards gaming, music, and social applications. First cluster respondents generally had less applications, more free applications, and were less likely to visit the website designed by the survey.

Segment profiling shows that consumers in the first cluster have less of a preference for entertainment and television applications. Instead, they have more of a preference for social media applications. App Happy is trying to produce a social entertainment platform. The analysis would suggest that the social media portion of the product will appeal to the first cluster and the entertainment portion of the product will appeal to the second cluster. App Happy could target a product towards consumers designated in the first cluster and would most likely have success if it were available on the iPhone and Android platforms. Consumers in the second cluster would appeal to the same platforms because they prefer iPod and tablet devices.

Classification Models

The ultimate goal is to create a classification model that could be replicated and re-used for future survey data. There are many different classification models that they could use to predict customer loyalty. These models include Decision Trees, K-Nearest Neighbors and a Support Vector Machines. For the company App Happy, it is not cut and dry to determine which model to use in the future. Depending on what type of resources they are willing to use, and depending on the amount of data, App Happy would be satisfied in using either Decision Trees or a Support Vector Machine. If they may want to use K-Nearest Neighbors if they start collecting data with varying types, scale, and distributions.

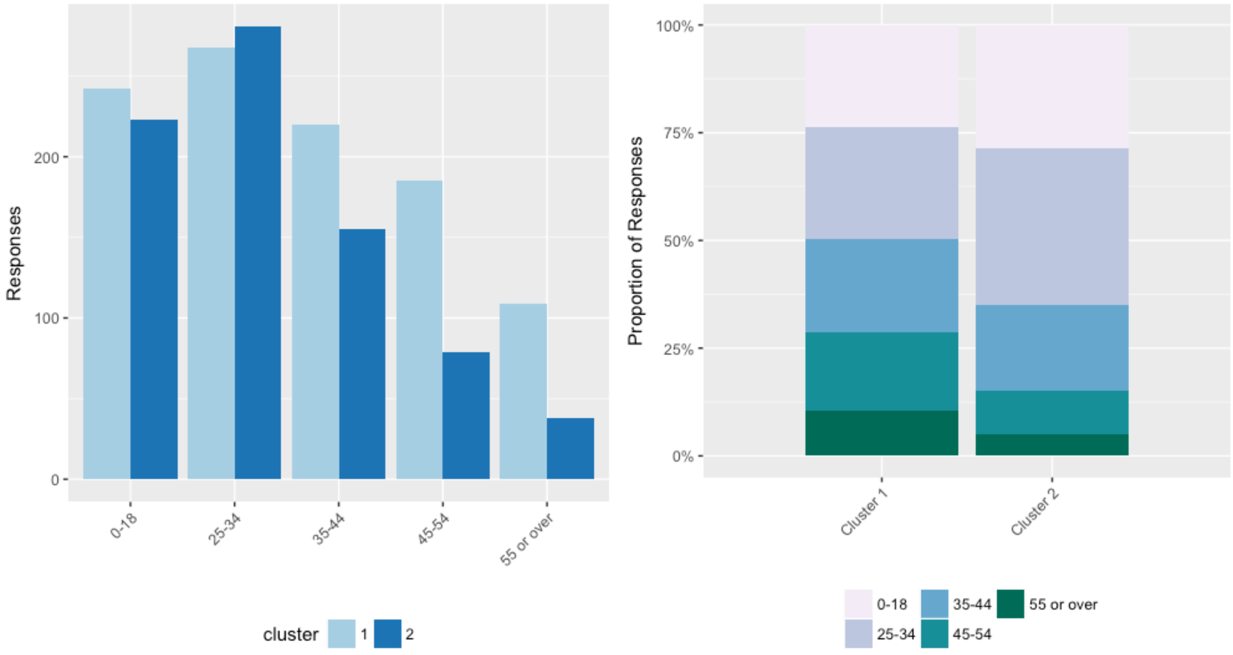
Conclusion

There were many different clustering combinations used and different distance metrics to segment the survey data. Even though the data may be too continuous, the K-Means clustering with Euclidean distance metric was the best performing model for the data. Consumers in both clusters were found to share demographic and behavioral characteristics. The clusters could distinguish themselves from one another according to respondent age, education, incomes, and ethnicity. In regard to consumer dispositions, clusters were able to be told apart according to their preferences for type of electronic device, use of applications, and their propensity to purchase applications. The first cluster had less of a preference for entertainment and television applications, a greater preference for using iPhone and Android devices, a greater preference for social media applications, and were less likely to browse websites such as Facebook or Twitter where the survey was being hosted.

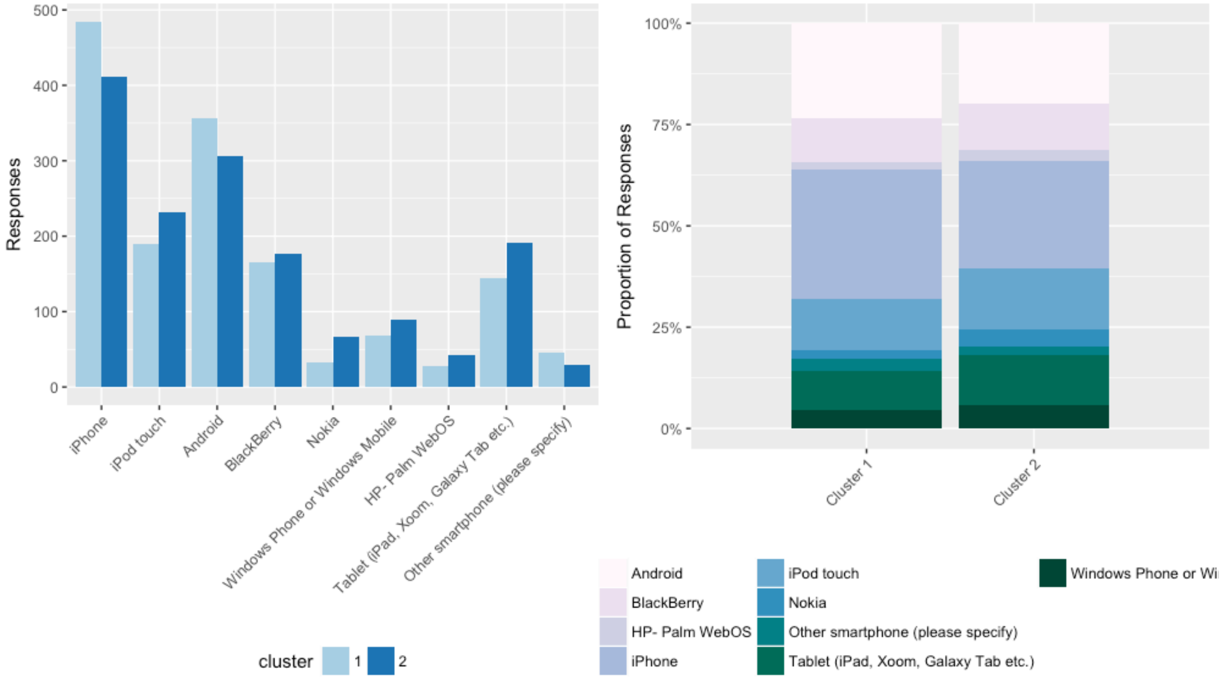
Different types of models could be used by App Happy to predict customer loyalty for their products. Each model had their own strengths and weaknesses. These models include Decision Trees, K-Nearest Neighbors, and a Support Vector Machine. When App Happy is thinking about moving into a new marketplace in the future, we recommend the use of a classification model.

Appendix

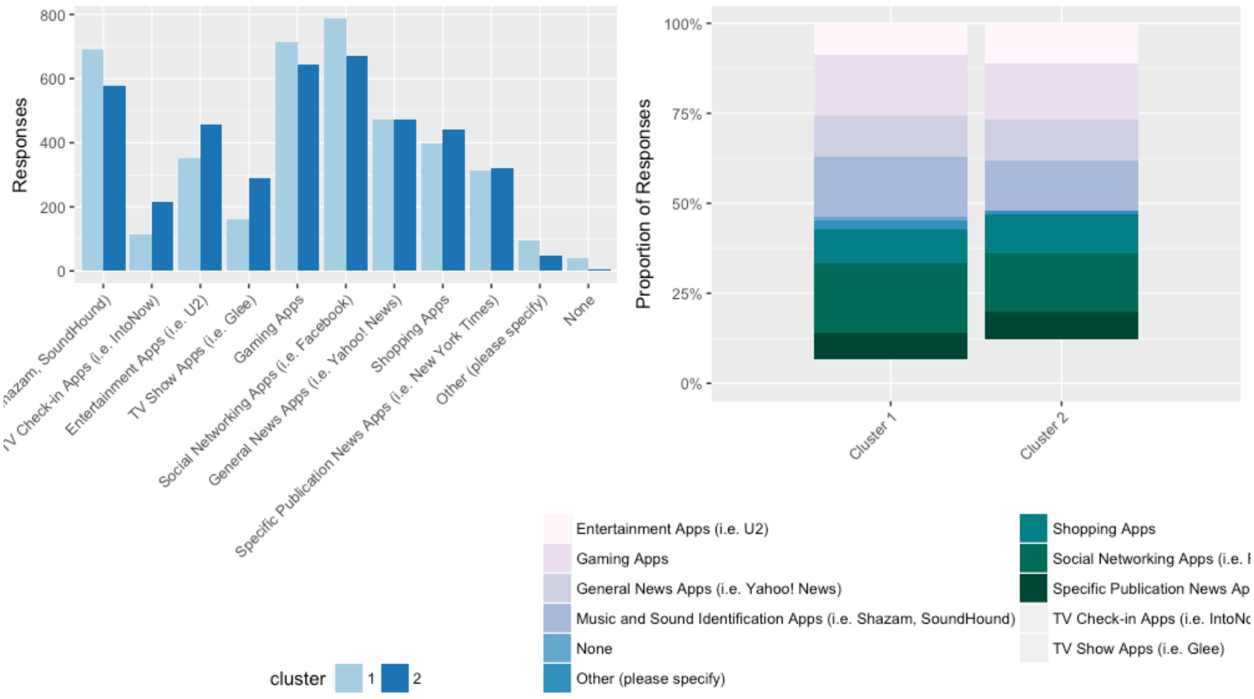
q1. Which of the following best describes your age?



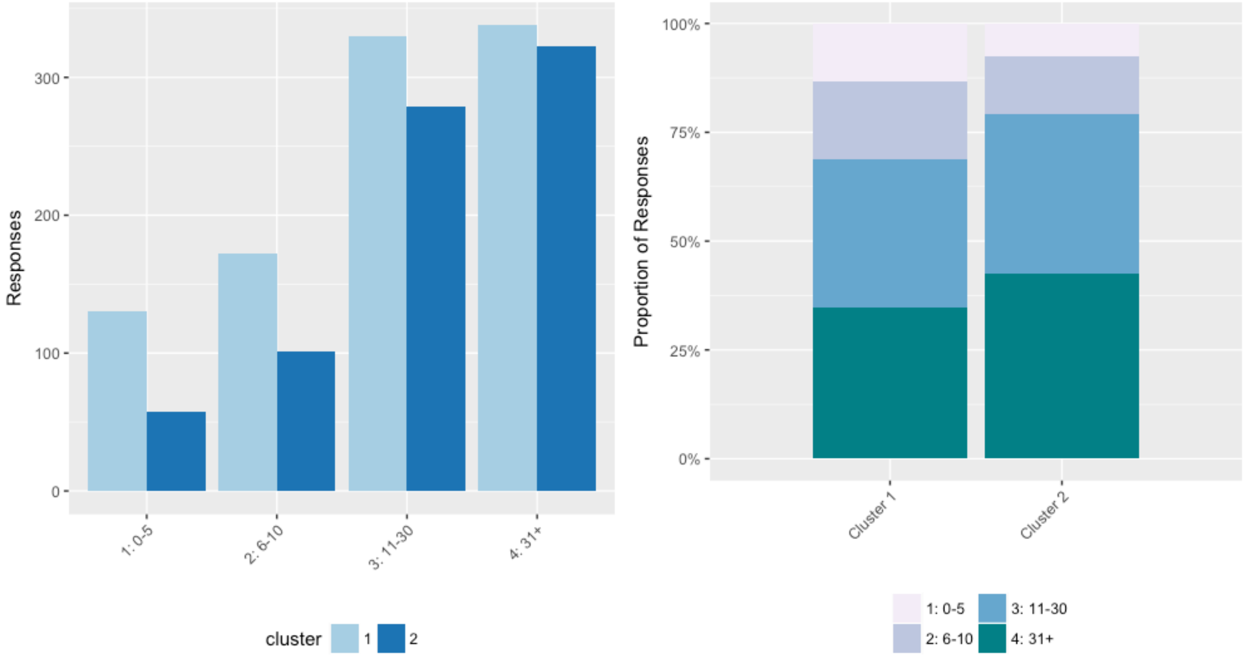
q2. Do you own any of the following smartphones or other web-enabled devices?



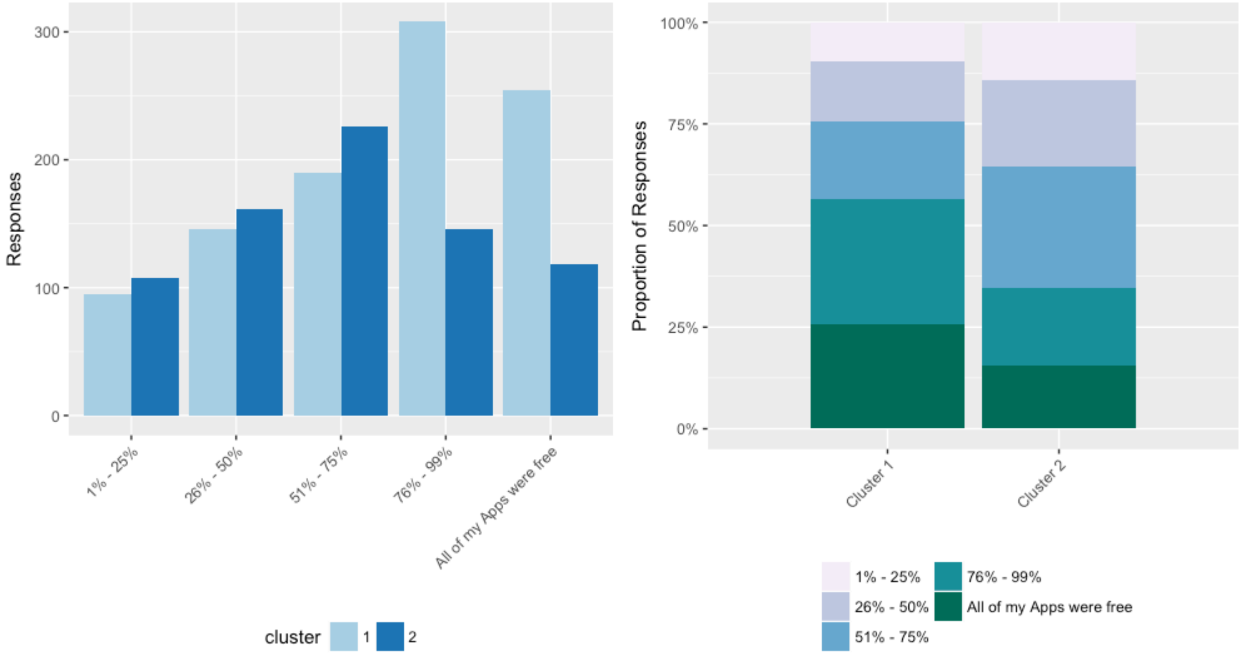
q4. Do you use any of the following kinds of Apps?



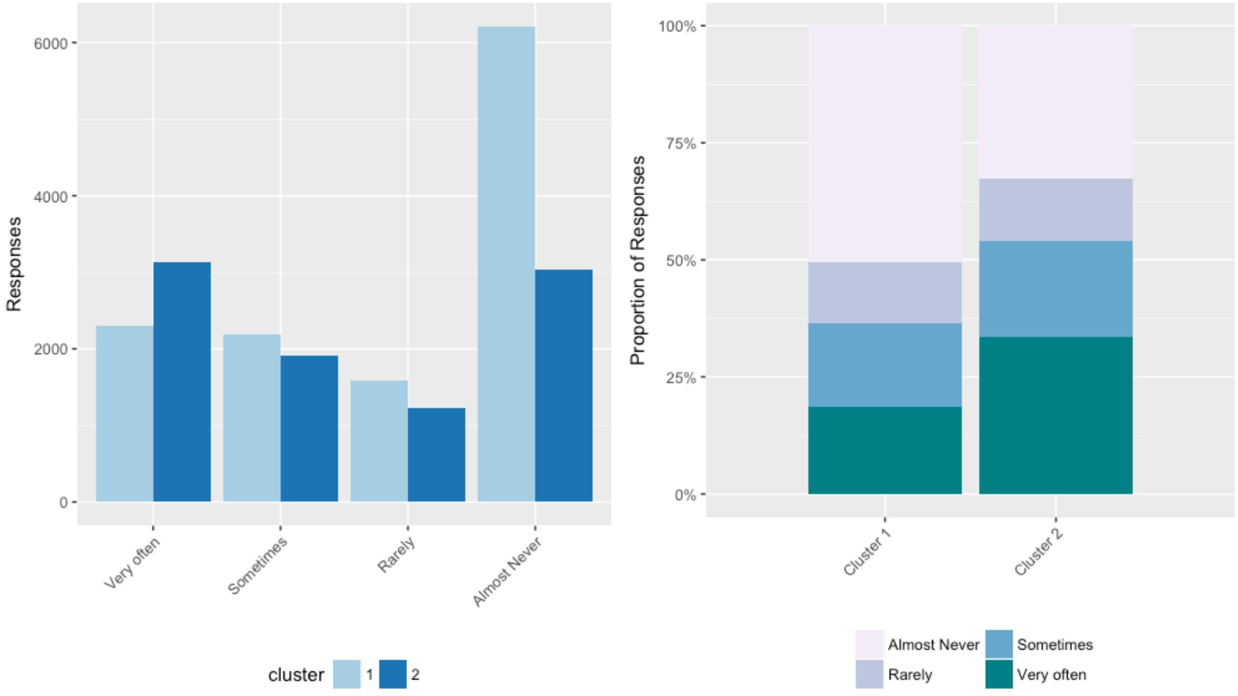
q11. How many Apps do you have on your smartphone/iPod Touch/Tablet?



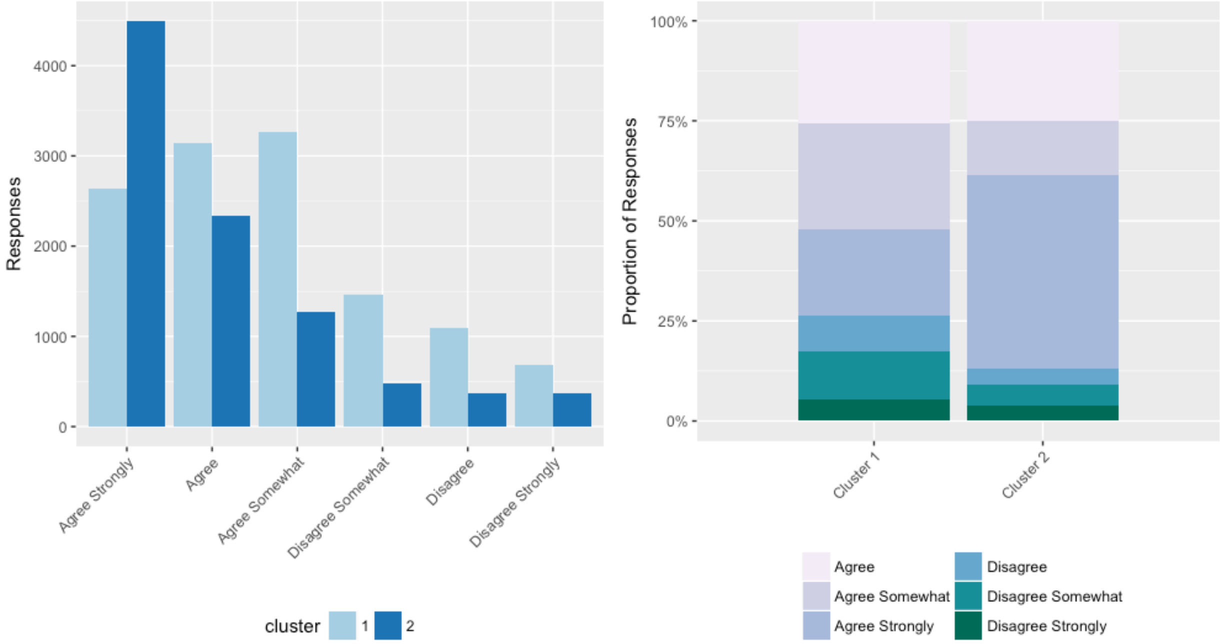
q12. Of your Apps, what percent were free to download?



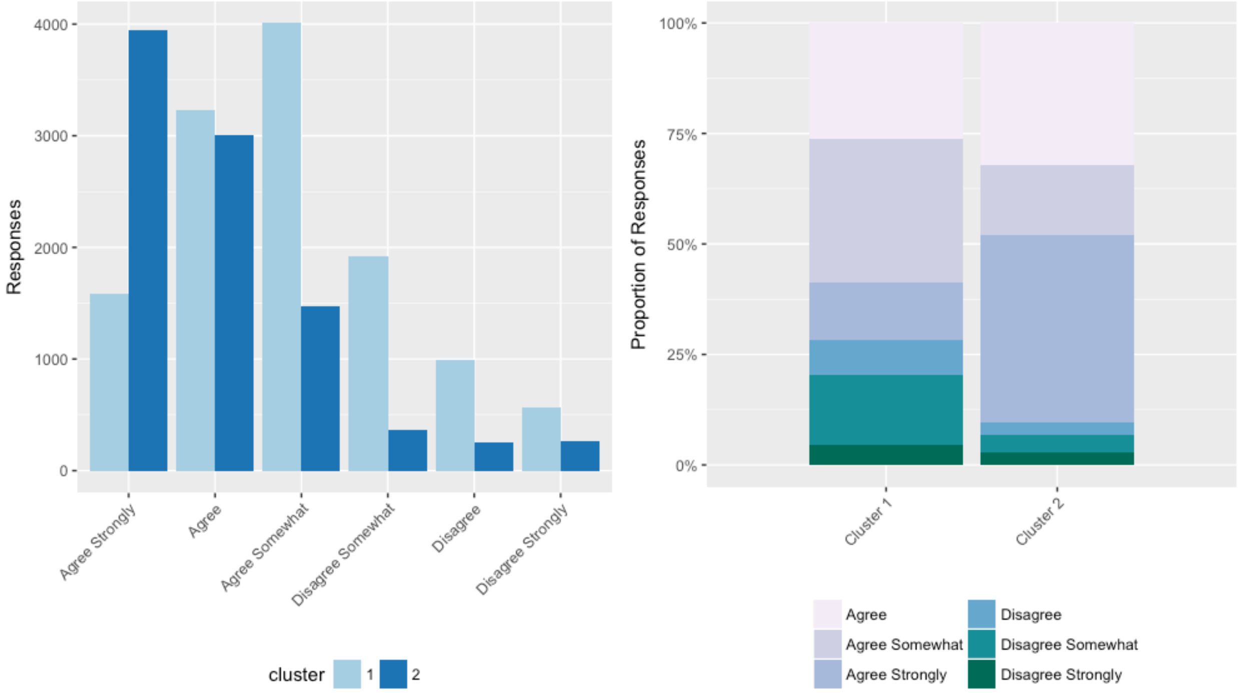
q13. How many times per week do you visit each of the following websites?



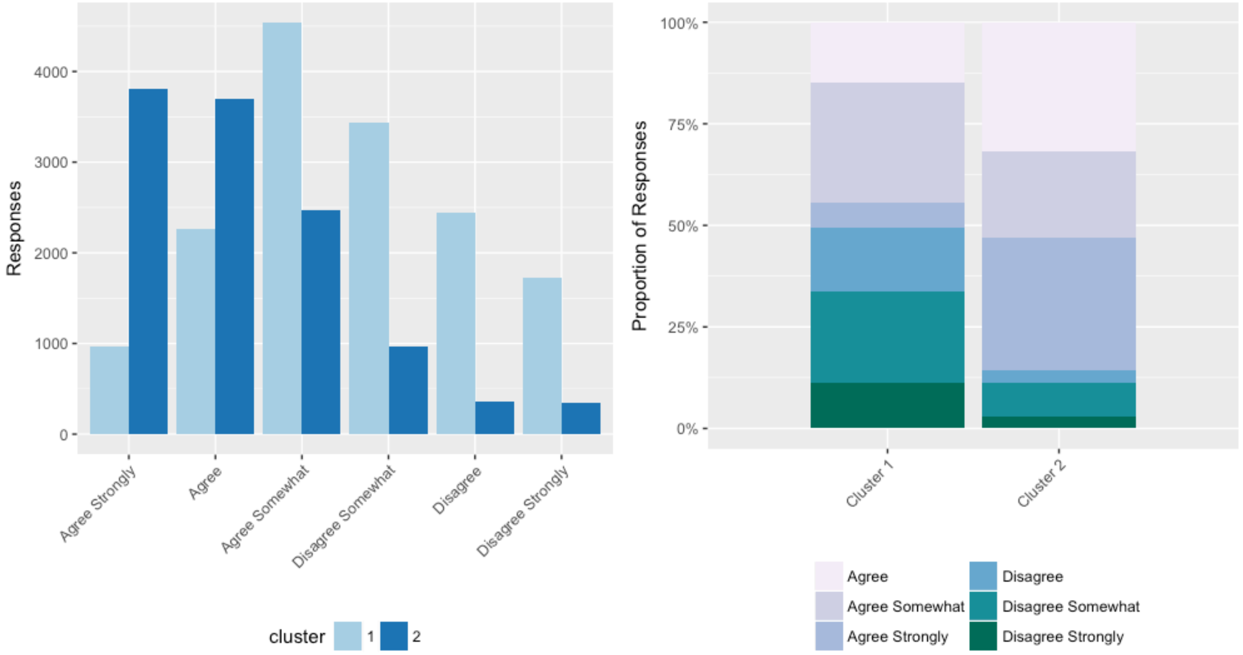
q24. Please tell us how much you agree or disagree with each of the follow statements



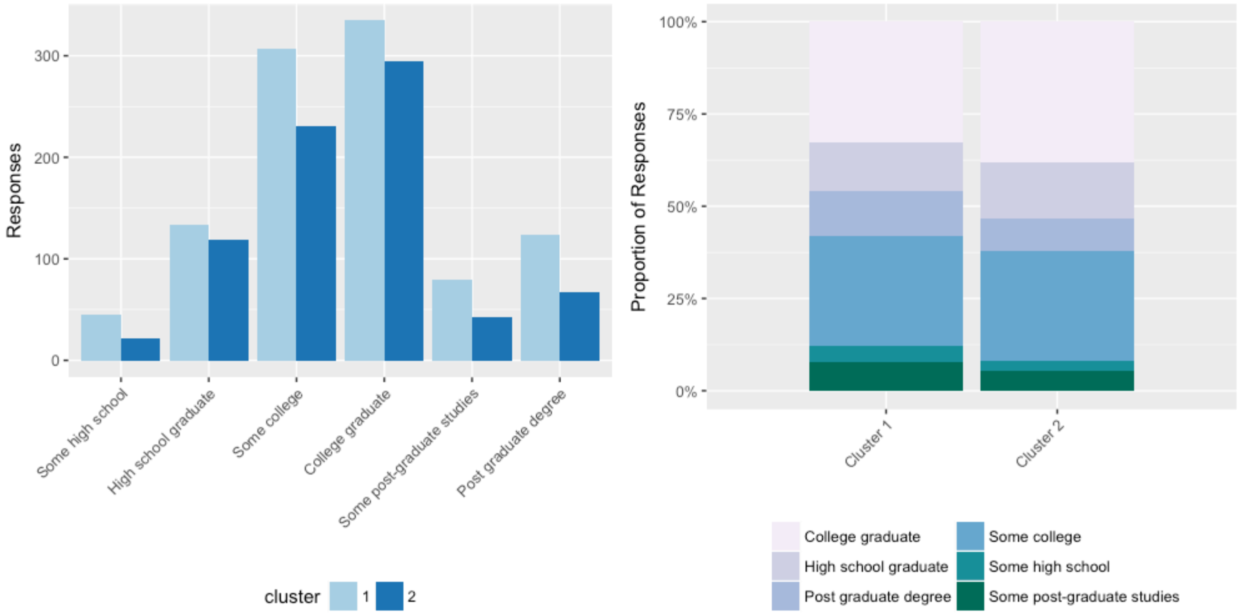
q25. And how much do you agree or disagree with each of the following?



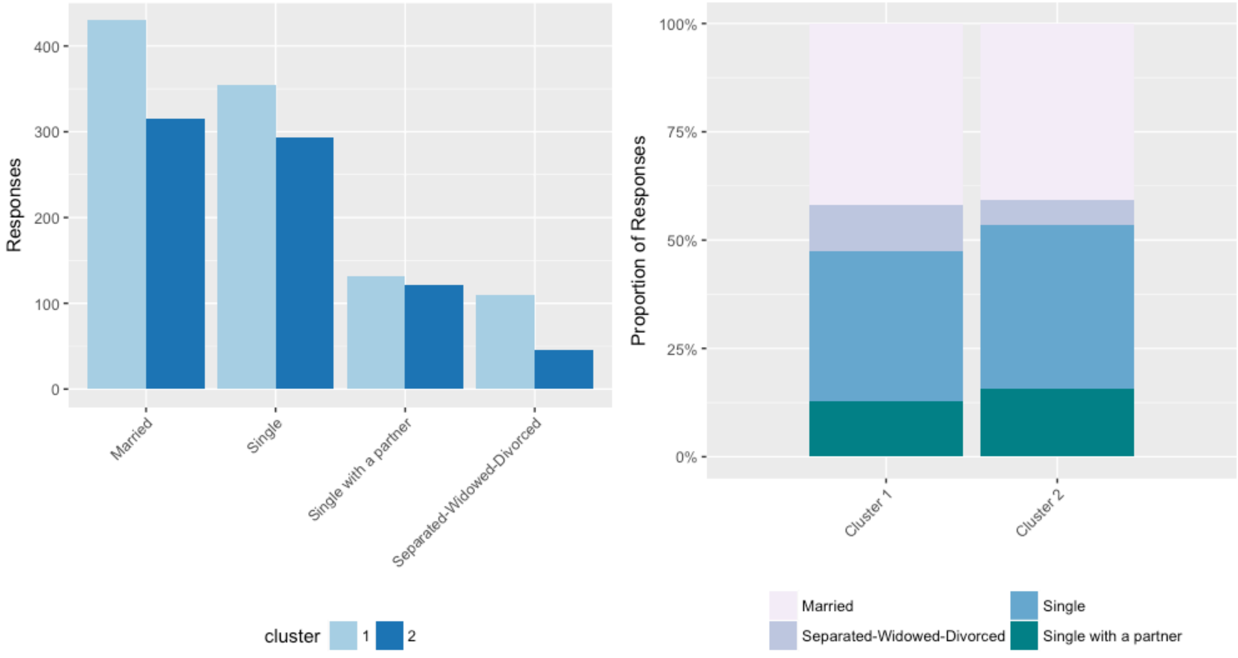
q26. And finally how much do you agree or disagree with each of these statements?



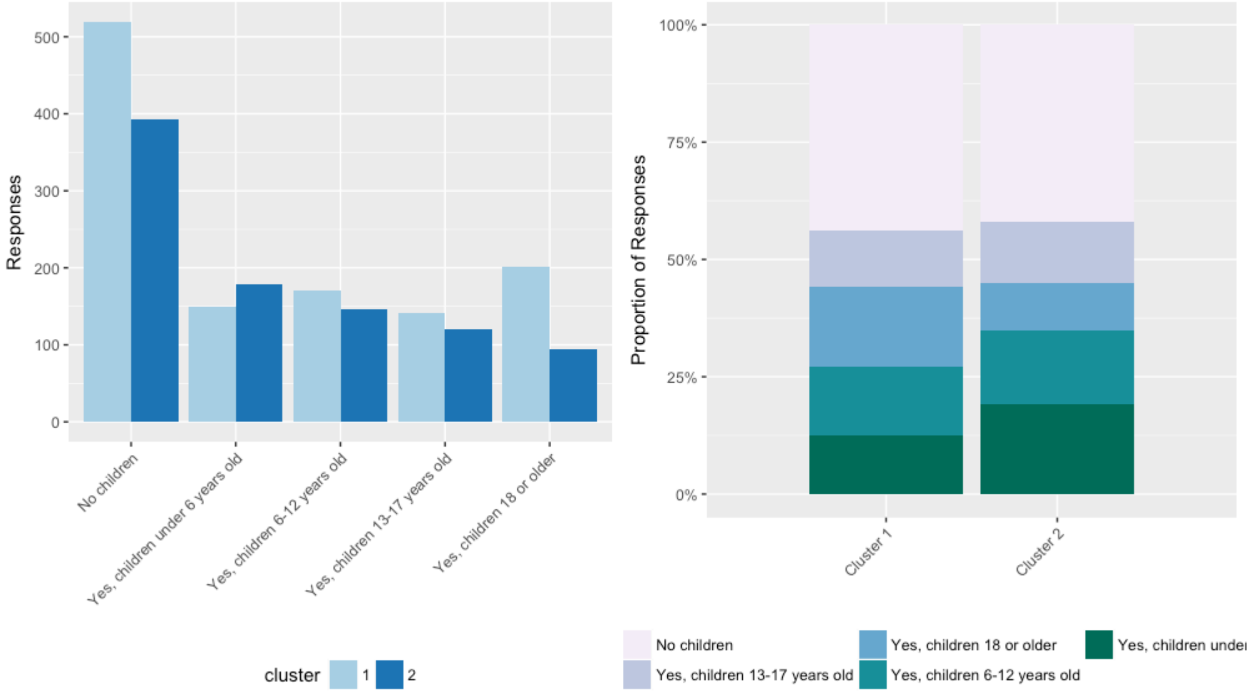
q48. Which of the following best describes the highest level of education you have attained?



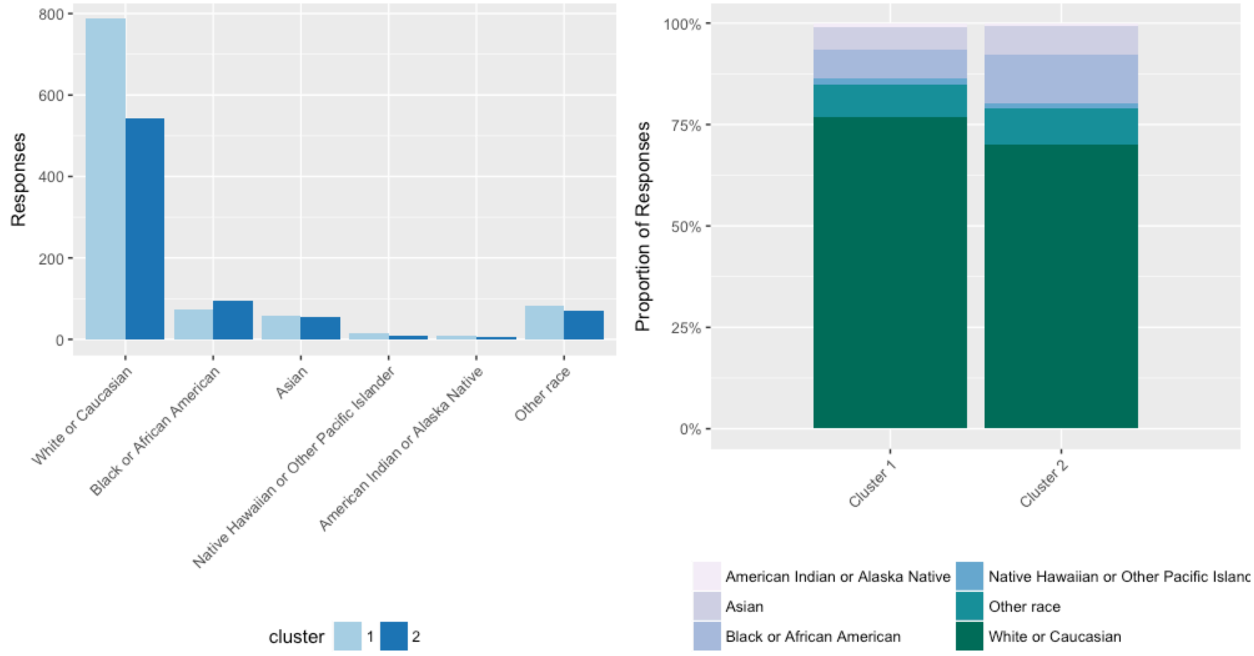
q49. Which of the following best describe your marital status?



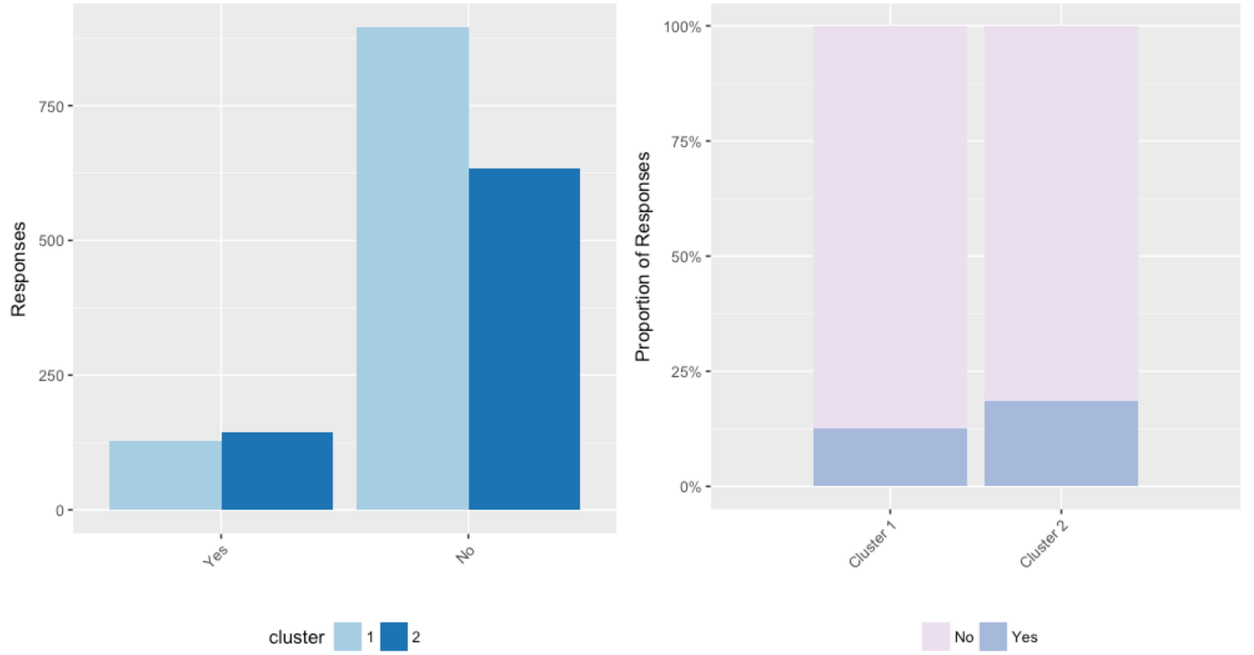
q50. Do you currently have any children in the following age groups?



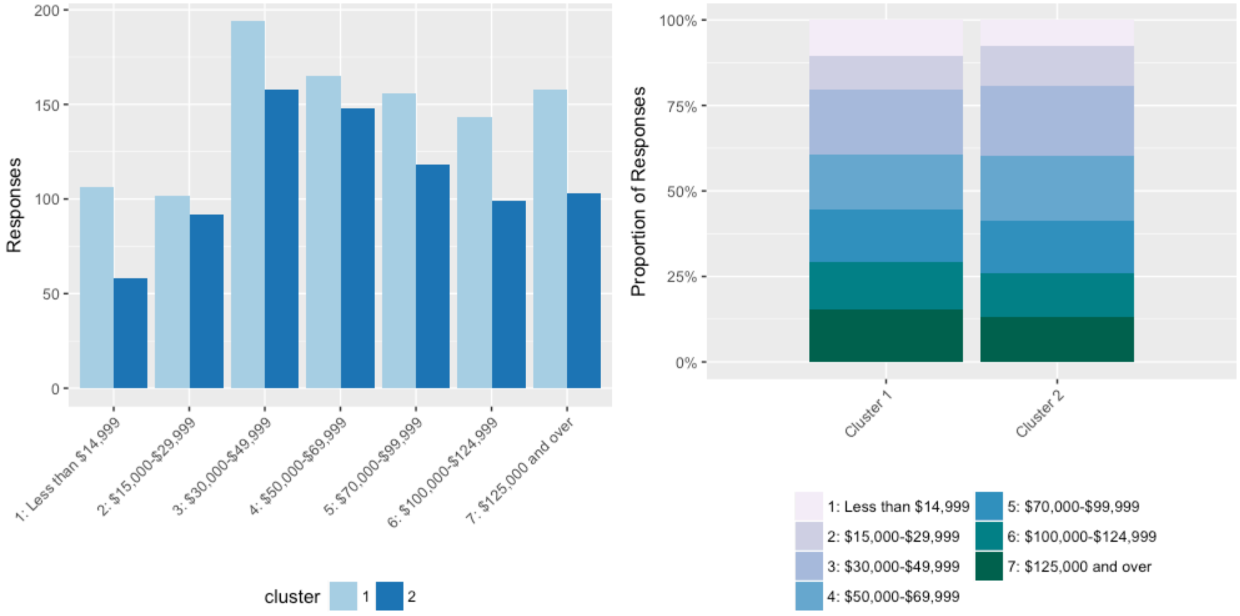
q54. Which of the following best describes your race?



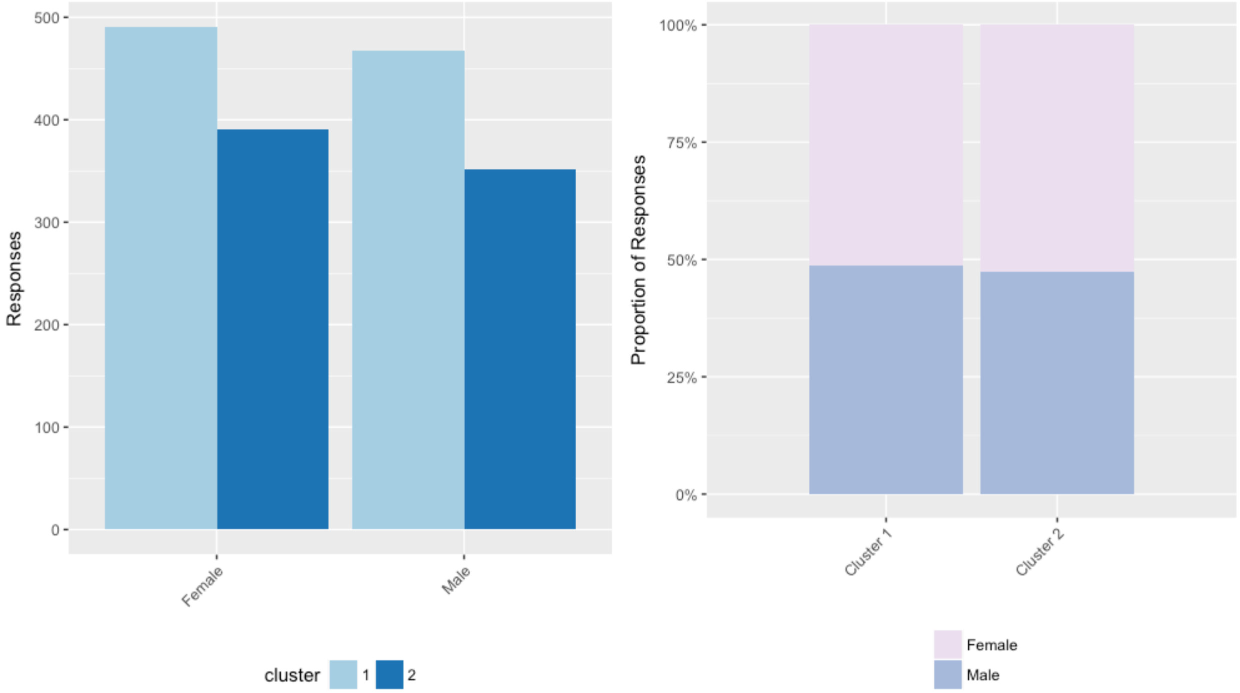
q55. Do you consider yourself to be of Hispanic or Latino ethnicity?



q56. Which of the following best describes your household annual income before taxes?



q57. Please indicate your gender


Responses
Female
Male
cluster
Proportion of Responses
Cluster 1
Cluster 2
Female
Male