Applied Statistical Modelling (CS7DS3)

Main Assignment: Option C

Catalin Gheorghiu (22305257)

1 Data & Methodology

Per the assignment specification, we focus on two series of 17430 observations from the 2007 US Survey of Consumer Finances. The first is `lsam` $= log(x + 50)$, where x is the average amount of savings as of the previous month average. The second is `linc` $= log(y + 50)$, where y is the total income over the previous year.

Figure 1 plots the two series against each other, with an overlapped contour plot of point density which reveals that most observations fall in the middle of the log-income range and either at 0 savings or close to just before the middle of the log-savings range. There is also a visible positive correlation between log-income and log-savings, with the added observation that savings often (and understandably) fall below the level of income. Finally, a noteworthy anomaly is how much log-income can vary while savings remain 0. The upcoming analysis will reveal whether these intuitions about the data are in any way captured or exploited by the investigated clustering algorithms.

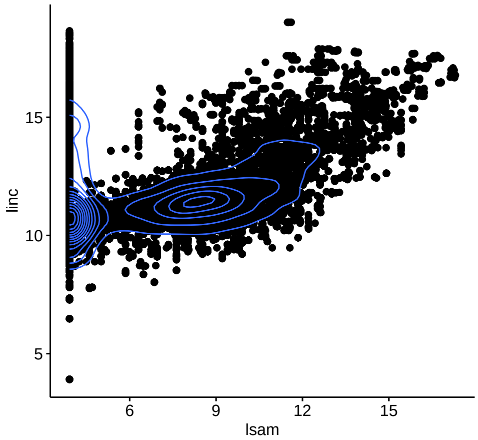

Figure 1: Scatter plot of log-income against log-savings with an overlapped density plot.

I will compare the performance of a model-based clustering algorithm based on expectation-maximisation (EM) to a more simple partitioning algorithm around K medoids. Exact algorithm choice and hyperparameter optimisation for each algorithm are discussed in the algorithms' respective sections.

2 Model-based clustering: EM

As we are working with an assumed multivariate normal model, the 14 possible covariance structures given by the different combinations of volume, shape, and orientation must be compared for each candidate number of clusters. I have chosen the Bayesian Information Criterion (BIC) as the approximation for model evidence to be maximised, and figure 2 shows the comparison of the BIC resulting from the different covariance structures when up to 12 clusters are considered.

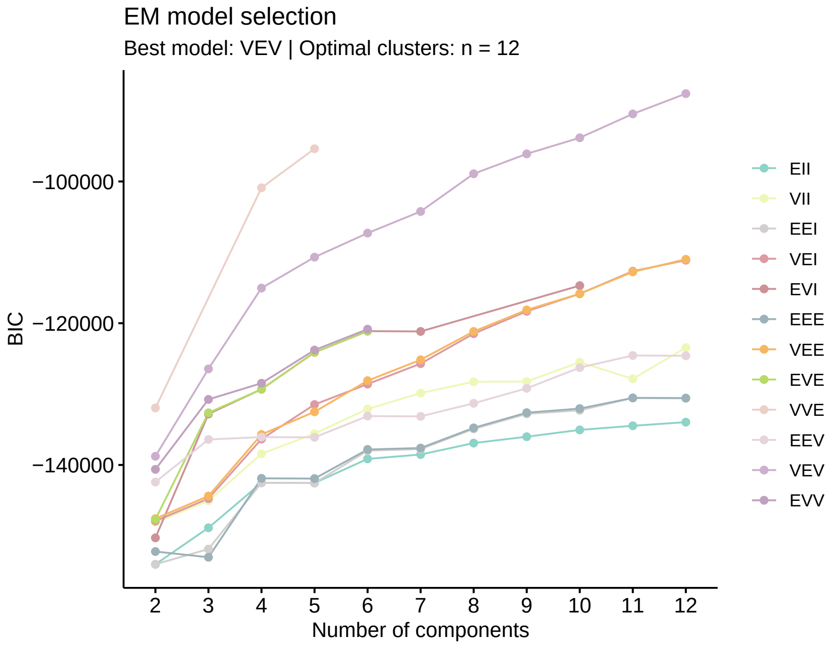

Figure 2: BIC comparison for the 14 covariance matrix structures of the assumed Gaussian distributions that generated up to 12 possible clusters.

Varying volume, equal shape and varying orientation (VEV) is the winning Gaussian distribution model, with the BIC maximised for 12 clusters. As the background literature for this exercise concurs, this result falls in line with the tendency of BIC maximisation to return the highest number of clusters available, as the BIC itself continues to increase – albeit less-than-linearly – with the number of components (Hennig and Liao, 2013). This tendency to overfit is natural if the underlying distributions are not Gaussian, which is likely to be the case.

The resulting clusters are visualised on the scatterplot in figure 3; larger individual points have been placed in their respective clusters with higher uncertainty. Clusters 10, 9, 3, 2, and 5 all capture different groups of income for the individuals with no savings, while clusters 1, 12, 8, 7, and 4 capture similar but smaller income brackets of individuals that also save. Clusters 11 and 6 are much wider in surface, also capturing the outliers with the highest and smallest incomes respectively; note that cluster 6 is the only one that does not separate individuals with no savings from the others, likely due to the high variance along the savings axis incurred by the inclusion of the outliers. It is interesting to see log-income lead to a much finer group separation (between 5 and 7 brackets) than log-savings (2 brackets).

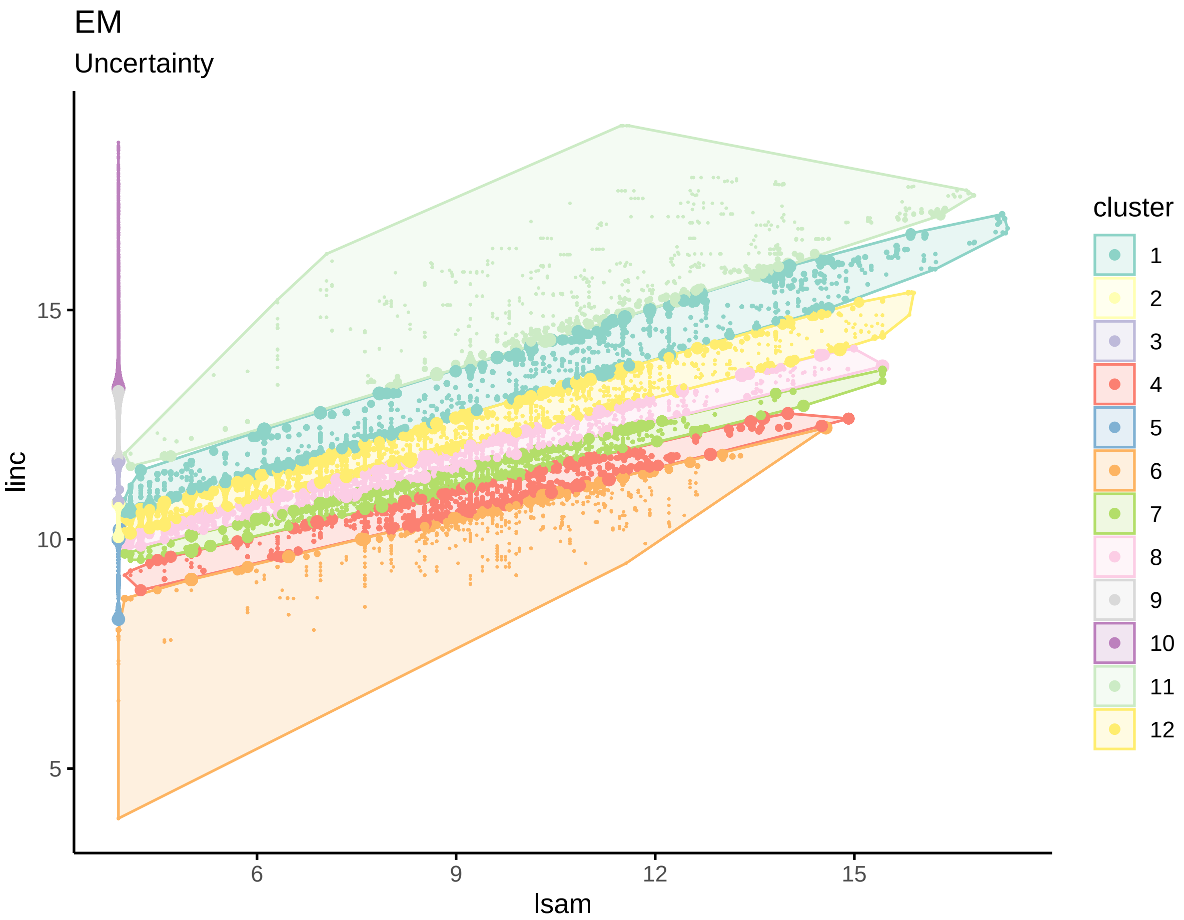

Figure 3: VEV Gaussian EM clusters coded in colour and bounded by convex borders, with uncertainty coded in point size.

The image of uncertainty from figure 3 is complemented by the histogram in figure 4. About 70% of the points (12000) are assigned with certainty, with few uncertainties above 0.1 and rather evenly distributed over the remaining range. The highest uncertainties can are intuitively found at the borders of the various clusters, and the phenomenon is exacerbated for smaller clusters like 4, 7, and 8 that cover small ranges of income which may overlap with the confidence intervals of neighbouring higher-variance Gaussians. Nevertheless, the high proportion of certainly assigned points is encouraging, attesting to relatively clean divisions in spite of the highly concentrated points and high number of clusters.

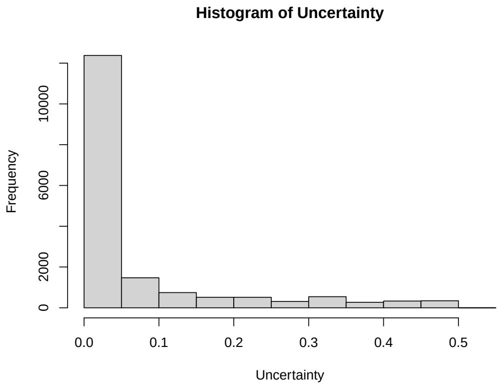

Figure 4: Histogram of uncertainty levels for the classification of every point in EM-generated clusters.

The bar plot in figure 5 supports the following interpretation of the cluster structure. Clusters 12, 8, 2, and 3 are the most populous, shortly followed by 1 and 7. These can be understood to form the canonical middle class in the economy, both due to their relative size and the mid-range level of log-income. Much like in the real-world middle class, there is a noticeable difference between what may be construed as upper-middle class (cluster 1, with an average log-income of 13.2) and lower-middle class (cluster 2, with an average log-income of 10.4). In terms of monetary units, this amounts to a difference of 507,000: the upper-middle class would earn on average 1500% more!

Clusters 4, 5, and 6 form the lower end of the distribution in terms of log-income and show markedly lower frequencies compared to the previously discussed clusters, making them interpretable as the working class. Clusters 9, 10, and 11, on the other hand, show the highest average log-incomes while maintaining low relative frequencies, making them interpretable as the upper class. It is noteworthy that this is the only group where one third of the clusters have non-zero savings, compared to two thirds in the others; since we are working with such small numbers, these proportions are not necessarily significant, but they contradict the positive correlation between class and savings that is otherwise fairly commonly supported in the real world. On the other hand, in support of the correlation between savings and income identified earlier, the lowest average log-income is found in cluster 5, where savings are not significant, while the highest average log-income is found in cluster 11, where savings are indeed common.

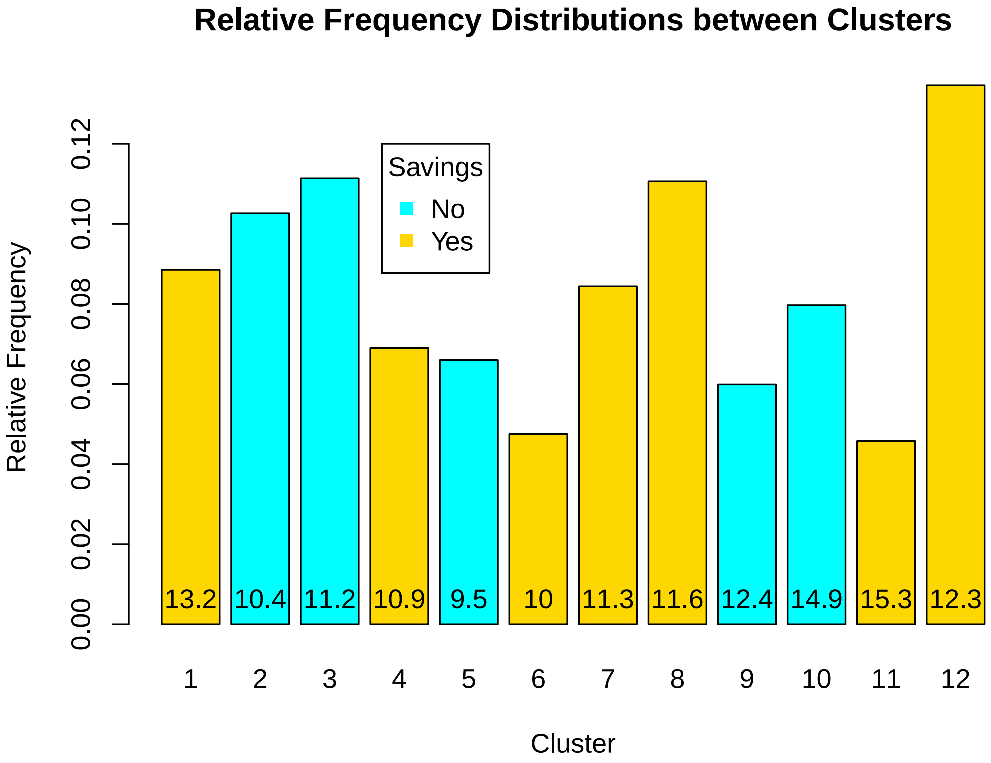

Figure 5: Relative frequency of EM cluster classifications, coloured by the existence of savings in the cluster and showing the average level of log-income.

3 Partition-based clustering: K-Medoids

The entire dataset is too large for K-medoids to finish in a reasonable time frame, especially for higher values of K. Consequently, I sampled 20% of the data points without replacement 9 times and plotted the average silhouette width (ASW) for K ranging from 1 to 12. Figure 6 shows scatter plots of 3 of these samples, as well as their corresponding ASW plots as K is varied. This method consistently returns K=2 as the optimal number of clusters for all 9 samples, but K=4 is also consistently very close behind. As a result, both will be considered in the upcoming analysis.

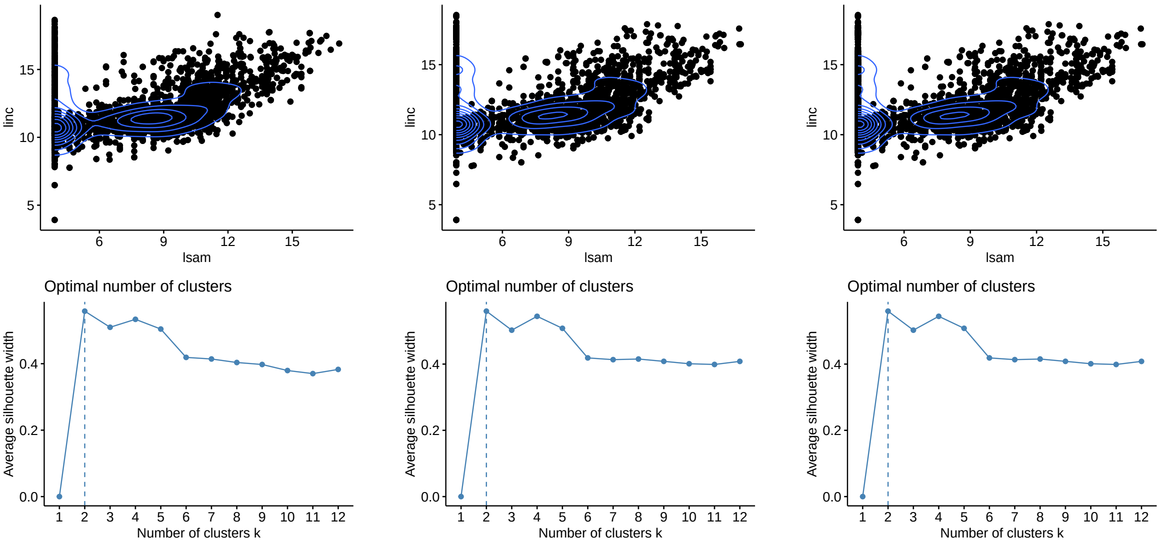

Figure 6: Scatter plots and ASW against number of clusters plots for 3 different samples of 3400 individuals.

Applying the 2-medoid and 4-medoid clustering algorithms to the entire dataset yields the structures seen in figure 7. The 2-medoid clustering separates on the basis of log-savings first and foremost, and the shapes of the fitted normal-distribution ellipses indicate that different features control most of the variance in different clusters. In the low-savings cluster 1, log-income intuitively drives the variance, especially as there are so many observations with no savings whatsoever that lie on a very large spectrum of income. In the high-savings cluster 2, however, there is significantly more variance in savings than there is in income.

The 4-medoid clustering further splits the previous two clusters, creating what may be interpreted as an upper and working class both within the group with savings (clusters 2 and 3, respectively) and within the group with no savings (clusters 4 and 1, respectively). The old border between savings-based groups is mostly unchanged, but it is noteworthy that the sparse points around log-savings of 7 and log-income of 15 mostly move to the low-savings group where they were part of the high-savings group beforehand. It is also noteworthy how for both the 2-medoid and 4-medoid settings, the high-income, high-savings outliers significantly bloat the variance of cluster 2, made obvious by the ellipses including lots of points from other clusters. All things considered, it seems more informative to continue with the 4-cluster structure, as the loss in ASW does not outweigh the additional insight given into the data.

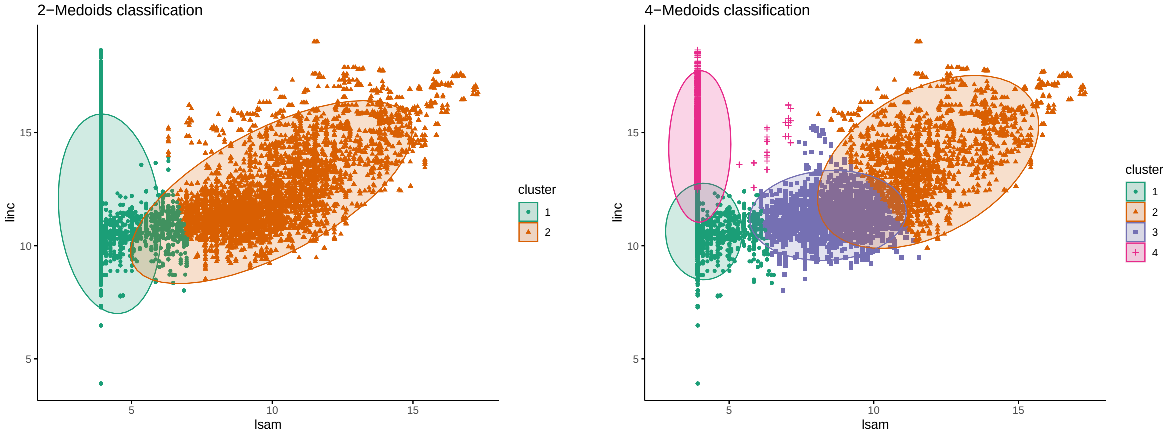

Figure 7: Scatter plots and ASW against number of clusters plots for 3 different samples of 3400 individuals.

Figure 8 gives additional insight into the structure of the 4-medoid clustering. Clusters 1 is the most populous by far, followed by cluster 3. These cannot necessarily be separated into a working and middle class as the difference in average log-income is rather small, but the average savings differ significantly. Nevertheless, we still observe the group that saves edging out the advantage in log-income. This trend is reversed when looking at the upper-class clusters 2 and 4; it is worth noting, however, that the group who saves is the more populous in the upper class (in spite of lower average log-income) while the group who does not save is more populous in the middle and working class. The hypothesis that income and savings are positively correlated is thus further enforced, although it is still impossible to talk about causality between the two. At most, the small frequency of the upper-class individuals who do not save can be interpreted to imply that making a high-income without any savings is usually the result of good fortune, making savings a more reliable predictor of financial success and therefore suggesting that savings do indeed increase the probability of socio-economic climbing.

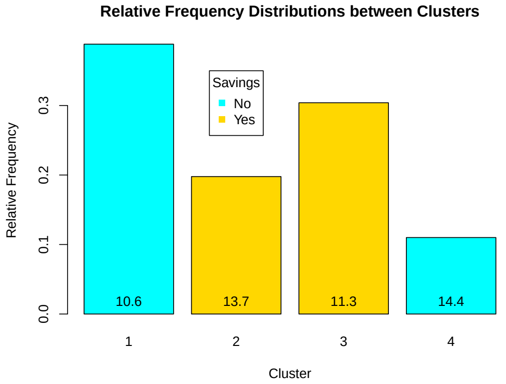

Figure 8: Relative frequency of 4-medoid cluster classifications, coloured by the existence of savings in the cluster and showing the average level of log-income.

4 Concluding Objective Statement

Two main algorithms were considered: a model-based EM approach assuming Gaussian underlying distributions, and a partition-based K-medoids approach to compare against the former and act as a sanity check. The optimal EM algorithm, 12-cluster VEV, was chosen from among the $14 \cdot 12 = 168$ attempted models as it maximised the BIC and thus the model evidence. The optimal K-medoid algorithm, with $K = 4$, was chosen after inspecting the ASW measures up to $K = 12$ and discarding the slightly-higher ASW $K = 2$ configuration as I did not believe the difference in ASW offset the additional intuition that could be gained.

While it is likely that BIC-based model selection would have suggested many more clusters for the EM algorithm if given the chance, the analysis in the assignment suggests the differences would likely be rather small, further splitting pre-existing clusters and over-fitting the data. Some symptoms of over-fitting are already apparent in figure 3, as the difference between some clusters (e.g. 7 and 8) is not very quantitatively significant (11.3 against 11.6 average log-income for instance) and interpreting the difference between them is difficult as well. On the other hand, the fine-grain clustering generally allowed for deeper insights when trying to interpret the data from a standpoint of socio-economic class and how income and savings are tied to them compared to 4-medoids, leading me to prefer the model-based algorithm overall.

To make a conclusive interpretation of the results, both algorithms primarily separate those with high savings and high incomes from those that score low on both features, usually leaving a very populous middle class between. High incomes and low savings, as well as high savings and low incomes, are relatively rare, but the former is still markedly more prevalent, which would suggest that a high income with bad saving habits may be possible with a stroke of luck, but high savings are likely to keep an individual from having to work low-income jobs to get by.

References

Hennig, C. and Liao, T. F. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(3):309–369.

5 Appendix - R Code

```
#Installing packages
install.packages('ggpubr')
install.packages('mclust')
install.packages('factoextra')
install.packages('fpc')

#Enable libraries
library('ggpubr')
library('mclust')
library('factoextra')
library('fpc')
library('cluster')

#Read data
data = read.csv('SCF07.csv')
data = data[,2:3]
#Plot data
ggscatter(data, x = 'lsam', y = 'linc') +
  geom_density2d()

#Model-based clustering (EM)
cl = Mclust(data, G = 1:12)
summary(cl)
fviz_mclust_bic(cl, palette = 'Set3', main = 'EM model selection',
                legend = 'right')
fviz_mclust(cl, 'classification', geom = 'point', stand = FALSE,
            ellipse.type = 'convex', pointsize = 1.3, palette = 'Set3',
            main = 'EM')
fviz_mclust(cl, 'uncertainty', stand = FALSE,
             ellipse.type = 'convex', palette = 'Set3',
            main = 'EM')

#EM analysis
hist(cl$uncertainty, main = 'Histogram of Uncertainty', xlab = 'Uncertainty')
cluster_table <- table(cl$classification)
cluster_prop <- prop.table(cluster_table)
cluster_means <- aggregate(data, by = list(cl$classification), mean)
bar_colors <- ifelse(cluster_means[, 'lsam'] < 4, 'cyan', 'gold')
bp = barplot(cluster_prop, main = 'Relative Frequency Distributions between Clusters',
        xlab = 'Cluster', ylab = 'Relative Frequency', col = bar_colors)
text(bp, 0, labels = round(cluster_means$linc, 1), pos = 3)
legend(4, 0.12, title = 'Savings', legend = c('No', 'Yes'),
       col = c('cyan', 'gold'), pch = 15)
```

```
#K-Medoids clustering - find best k
best_k = vector(mode = 'integer', length = 3)
for (i in 1:3) {
  sm = data[sample(nrow(data), 3400), ]
  pdftitle = paste('sample_sca', i, '.pdf', sep='')
  ggscatter(sm, x = 'lsam', y = 'linc') + geom_density2d()
  ggsave(pdftitle)
  pdftitle = paste('sample_sil', i, '.pdf', sep = '')
  plot = fviz_nbclust(sm, cluster::pam, method = 'silhouette', k.max = 12)
  ggsave(pdftitle)
  info = plot$data
  best_k[i] = as.numeric(info$clusters[which.max(info$y)])
}
best_k

#K-Medoids clustering - apply to dataset
plot = pam(data, k = 2, metric = 'euclidean', stand = FALSE)
fviz_cluster(plot, palette = 'Dark2', ellipse.type = 'norm',
             stand = FALSE, geom = 'point', pointsize = 1.3,
             main = '2-Medoids classification'
             ) + theme(axis.line = element_line(),
                       panel.background = element_blank())

#KM analysis
km_cluster_table <- table(plot$clustering)
km_cluster_prop <- prop.table(km_cluster_table)
km_cluster_means <- aggregate(data, by = list(plot$clustering), mean)
bar_colors <- ifelse(km_cluster_means[, 'lsam'] < 6, 'cyan', 'gold')
bp = barplot(km_cluster_prop, main = 'Relative Frequency Distributions between Clusters',
             xlab = 'Cluster', ylab = 'Relative Frequency', col = bar_colors)
text(bp, 0, labels = round(km_cluster_means$linc, 1), pos = 3)
legend(0.55, 0.35, title = 'Savings', legend = c('No', 'Yes'),
       col = c('cyan', 'gold'), pch = 15)
```