Contents

1. Introduction

A flag is an emblem consisting of a piece of cloth with distinctive design. It is rectangular in shape and depicts a symbol. Flags bear motley connotations encompassing strong loyalty to a faith or belief. The national flags interpret some kind of religious intentions tinged with patriotic potent in a kind of veneration and reverence of the flag[1]. The purpose of the following assignment is twofold. The first one is that based on the flag dataset, which contains details of various nations and their flags, to try to predict the religion of a country. The second is to identify clusters of flags with respect to their characteristics.

Regarding the 1st aim, the problem we try to solve is classification. The response variable under study that we focus on predicting is religion which is a categorical variable with 8 levels (Catholic, Other Christian, Muslim, Buddhist, Hindu, Ethnic, Marxist, Others). We will use three methods: Classification Tree, Random Forest and Bagging. Based on their accuracy we will select the best classification method out of those three.

As far as the 2nd aim is concerned, we will use PAM clustering method to identify the different clusters of flags in our dataset and then describe each one of them.

1 Scientific Study of Religion in Vexillology URL , Muhammad Naeem1* and Sohail Asghar2, May 2013

2. Data Transformation and Descriptive Analytics

Below are the transformation steps taken in order to prepare the data for analysis:

1. We eliminated six features *"name of the countries", "landmass", "area in tsqm", "geographic quadrant", "population" and "language"* as these features have no direct or indirect relation towards religion of a nation in a sense of cause and its effect in probabilistic models.
2. Regarding the *"main hue predominant color", "top left color", "bottom right color"* characteristics of the flag which were the only categorical predictors with more than 2 levels, we created for each level a new variable and transformed them to binary (1 for those flags the have the respective level and 0 for those that don't have it).
3. We converted every categorical variable to factor and every discrete to numeric.

After performing the aforementioned actions, we got a dataset with 194 observations and 43 variables. Let's proceed with some descriptive analysis in order to get familiar with the dataset. After this, in the next section we will perform variable selection in order to select the "best" subset of predictors to avoid overfitting.

The Image 1 illustrates the number of flags of countries whose population believe in a certain religion. We can say that around 50% of the dataset are Catholics or other Christians, approximately 20% are Muslims and the remaining percentage is spread to the rest of the levels Ethnic, Marxist, Buddhist, Hindu and others. The Image 2 depicts a bar plot with the number of flags per each main hue. We notice that around 41% of the flags in the dataset have as predominant color the red, followed by 20% of the blue and 15% the green. Regarding the Image 3, we can see that blue color in flags might indicate that the religion of the respective populations are non-catholic Christians. Also, we might say that the countries where people are Muslims have flags with dominant colors red and green.

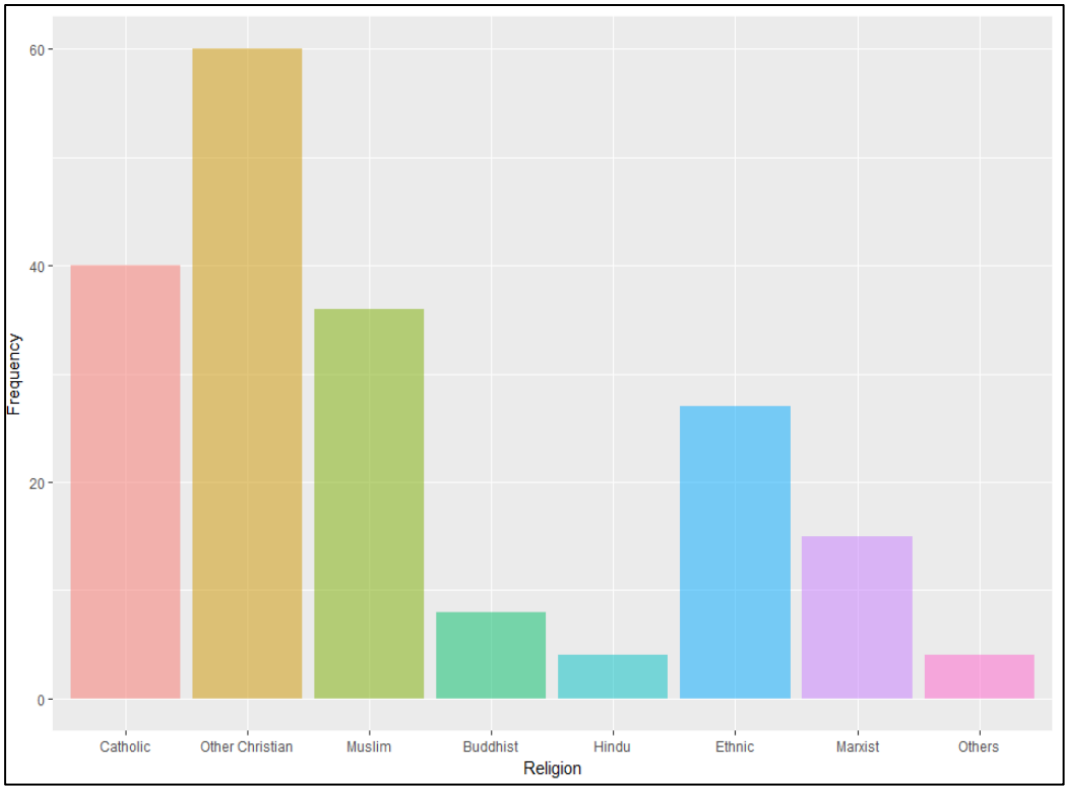


Image 1: Bar plot: No of flags per Religion

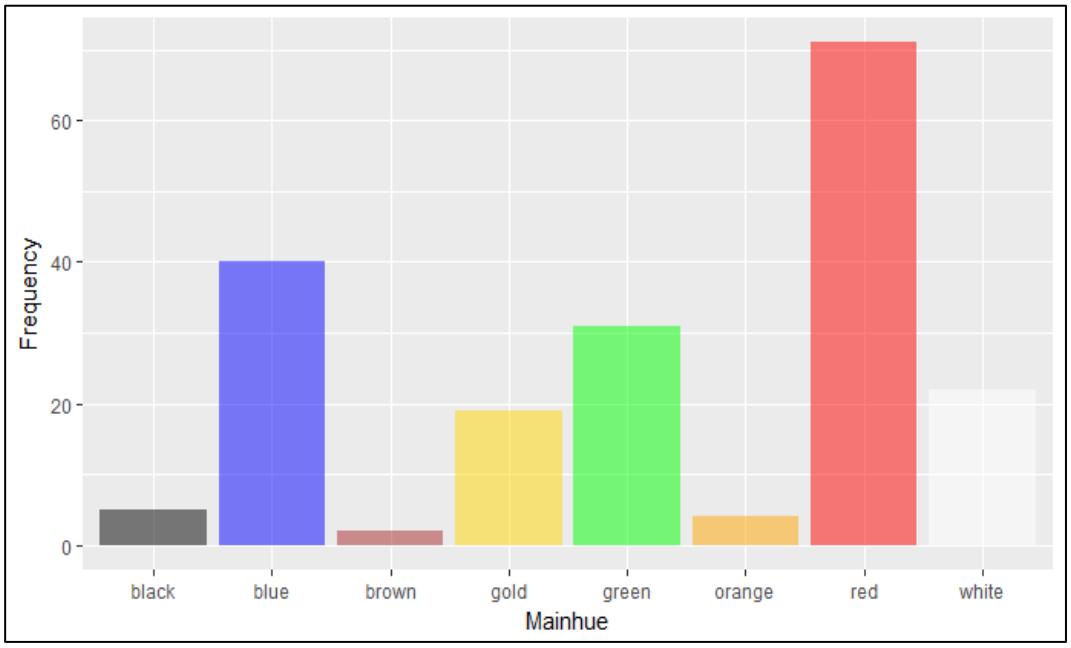


Image 2: Bar plot: Main hue

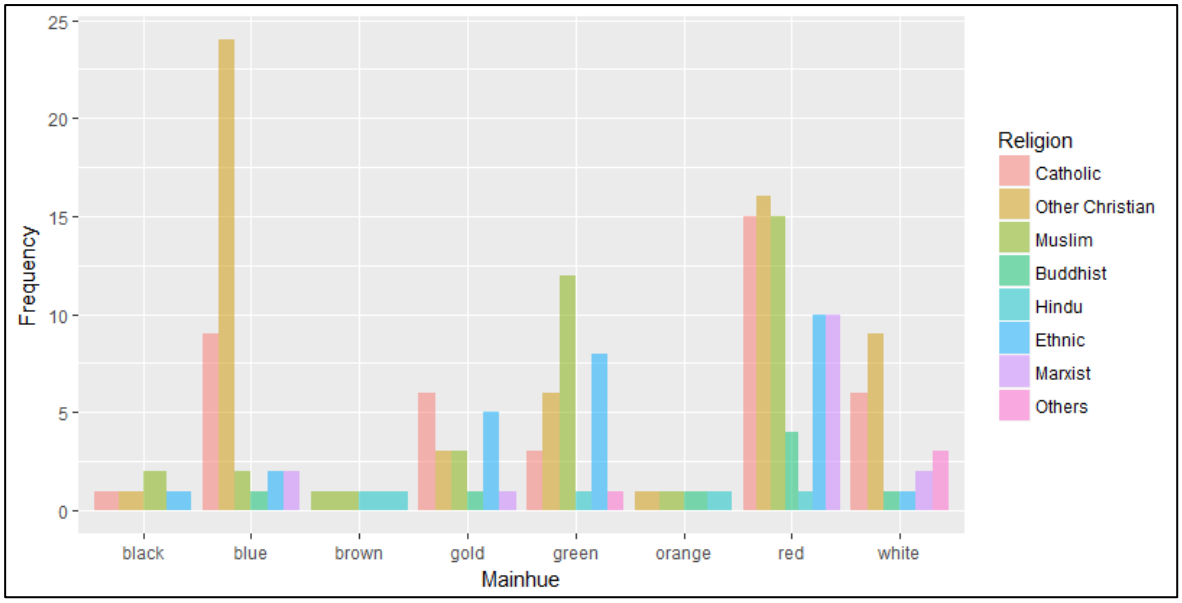


Image 3: Bar plot: Number of flags per Religion per dominant color

3. Variable Selection

In order to remove the redundant predictors from the dataset, avoid the possible noise that might be added to the estimation of the religion response variable and eliminate the collinearity issue, we will procced by implementing variable selection. To perform variable selection for the case that the response variable is categorical with more than 2 levels we will use a method called Grouped Lasso, which is an extension of the common Lasso[2]. Lasso is a penalized linear regression method that involves the penalization of the absolute size of the coefficients. By penalizing (or equivalently constraining the sum of the absolute values of the estimates) the objective is to be in a situation where some of the parameter estimates might be exactly zero. The larger the penalty applied, the further estimates are shrunk towards zero. Group Lasso performs similarly to Lasso except that it selects a group of variables rather than a single variable at each step of selection. The groups were pre-assigned on covariates. Therefore, in 1 categorical case, we group the stack of dummy variables originated from a factor variable and each group represents a corresponding factor variable[3].

After performing this method and selecting as value for the penalty the one that gives minimum mean cross-validated error of the group lasso, we end up in the following model and reduce the number of variables from 43 variables to the following 26 variables.

bars, stripes, red, green, blue, gold, white, black, circles, crosses, saltires, sunstars, crescent, triangle, animate, mainhue_blue, mainhue_brown, mainhue_orange, mainhue_white, topleft_green, topleft_gold, topleft_white, botright_black, botright_gold, botright_orange, botright_red

[2] Glmnet. Vignette Trevor Hastie and Junyang Qian Stanford. June 26, 2014 URL

[3] Lasso on Categorical Data. Yunjin Choi, Rina Park, Michael Seo. December 14, 2012 URL

4. Classification

4.1. Classification Tree[4]

A classification tree is used to predict a qualitative response. For a classification tree, we predict that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs. The method uses recursive binary splitting to grow a classification tree. The classification error rate is the fraction of the training observations in that region that do not belong to the most common class (classification error rate).

To proceed with the classification, we split the observations into a training set (70%) and a test set (30%), build the tree using the training set, and evaluate its performance on the test data.

The tree library is used to construct classification and regression trees. We use this library to cultivate the classification tree for the flag dataset and we get the tree of Image 4. Regarding the interpretation of the tree we can say for example that if a flag has no crosses, is green, has no gold color, 0 or 1 bars and the bottom right color is black, then the religion that is predicted for the population of the country that this flag belongs is Muslim, etc.

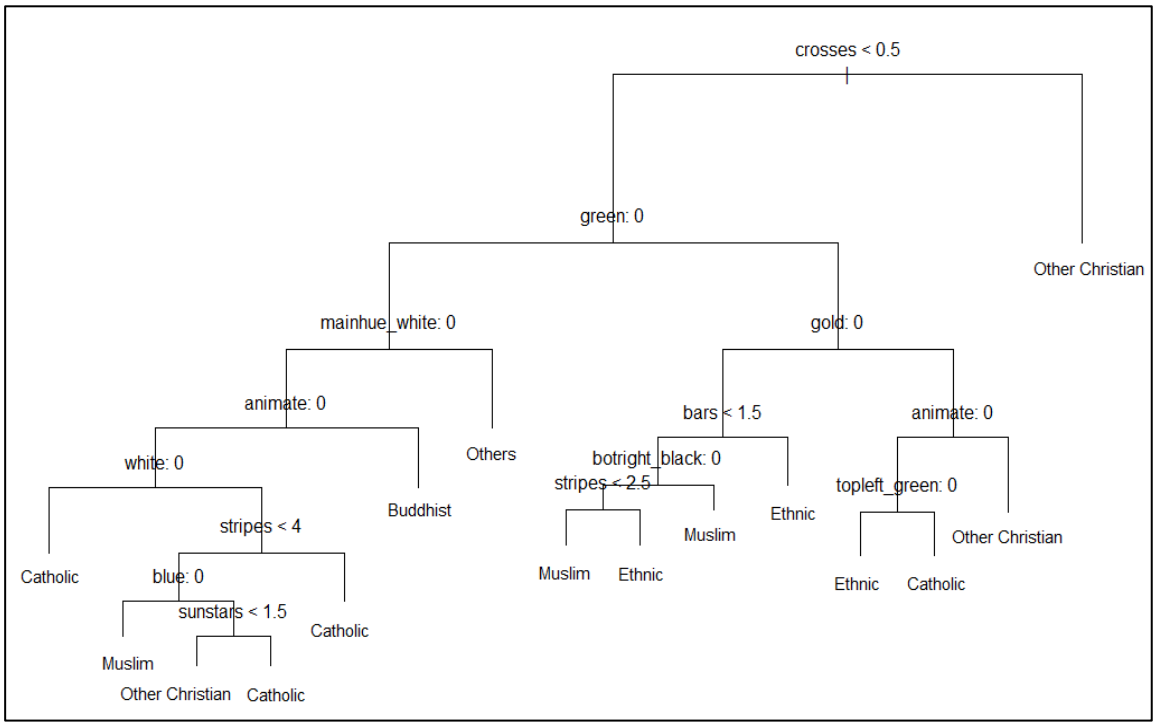


Image 4: Classification Tree on Training Dataset

Afterwards, we test the accuracy of the tree we build on the test dataset with the predict() function. We get an accuracy of 24.19%. Next, we consider whether pruning the tree might lead to improved results. The function cv.tree() performs cross-validation in order to determine the

4 An Introduction to Statistical Learning. Gareth James. Daniela Witten. Trevor Hastie. Robert Tibshirani. Springer Texts in Statistics.

optimal level of tree complexity. The cv.tree() function reports the number of terminal nodes of each tree considered *("size") and "dev"* corresponds to the cross-validation error rate. The tree with 4 terminal nodes results in the lowest cross-validation error rate, with 84 cross-validation errors. We then apply the prune.misclass() function to prune the tree and get the 4-node tree.

Command:
cv.tree(tree_model, FUN=prune.misclass)

Results:

- size: 15 11 9 7 4 1
- dev: 87 86 86 86 84 94

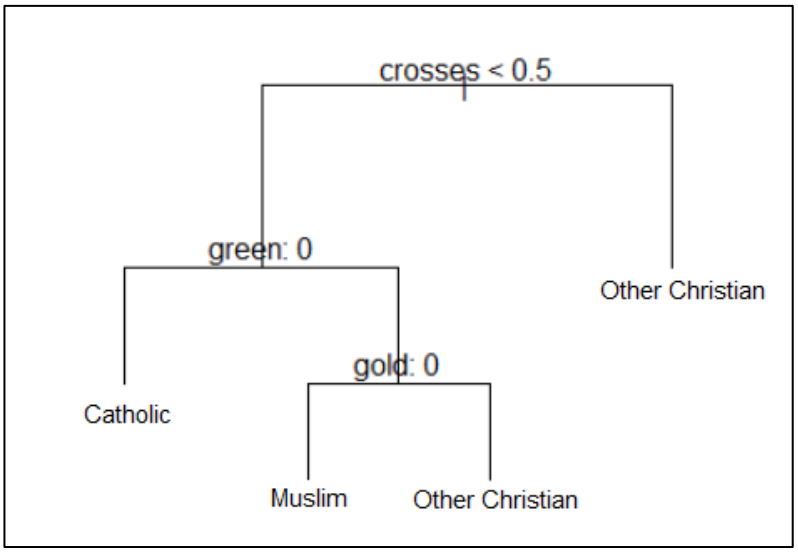


Image 5: Pruned tree

We apply again the predict() function on the test data and now 41.93% of the test observations are correctly classified. Thus, the pruning process not only produced a more interpretable tree, but also improved the classification accuracy.

4.2. Bagging[5]

The classification trees suffer from high variance. This means that if we split the training data into two parts at random, and fit a decision tree to both halves, the results that we get are quite different. Indeed, that was the case in the previous paragraph. In contrast, a procedure like bagging with low variance will yield similar results if applied repeatedly to distinct data sets. Bagging, is a general-purpose procedure for reducing the bagging variance of a statistical learning method. To reduce the variance and hence increase the prediction accuracy of a statistical learning method we have to take many training sets from the population, build a separate prediction model using each training set, and average the resulting predictions. Because there are not multiple training sets, bugging uses bootstrap method by taking repeated samples from the training dataset. To apply bagging to regression trees, we simply construct B regression trees using B bootstrapped training sets, and average the resulting predictions. These trees are grown deep, and are not pruned. Hence each individual tree has high variance, but low bias. Averaging these B trees reduces the variance. In Bagging there is a very straightforward way to estimate the test error of a bagged model, without the need to perform cross-validation or the validation set approach. This is done by computing an out-of-bag estimate of the misclassification error. On average, each bagged tree makes use of around two-thirds of the observations. The remaining one-third of the observations not used to fit a given bagged tree are referred to as the out-of-bag (OOB) observations. To apply bugging method on the flag dataset, we use the bagging() function from the “ipred” library and we set the coob parameter to TRUE. The method cultivated 25 trees with out-of-bag estimate of misclassification error 56%. The accuracy of the method in the flag dataset by using the predict() function is 43.3%.

[5] An Introduction to Statistical Learning. Gareth James. Daniela Witten. Trevor Hastie. Robert Tibshirani. Springer Texts in Statistics.

4.3. Random Forest[6]

Random forests provide an improvement over bagged trees by "decorrelating" the trees. In other words, in building a random forest, at each split in the tree, the algorithm is not allowed to consider a majority of the available predictors. Supposing that there is one very strong predictor in the data set, along with a number of other moderately strong predictors, then in the collection of bagged trees, most or all of the trees will use this strong predictor in the top split. Consequently, all of the bagged trees will look quite similar to each other. Hence the predictions from the bagged trees will be highly correlated. Averaging many highly correlated quantities does not lead to as large of a reduction in variance as averaging many uncorrelated quantities. This means that bagging will not lead to a substantial reduction in variance over a single tree in this setting. Random forests overcome this problem by forcing each split to consider only a subset of the predictors.

Instead of using a 70%-30% split in our dataset for training and testing, we will use a 10-fold cross validation and run the random forest 10 times (library randomForest). This approach involves randomly dividing the set of observations into 10 folds, of approximately equal size (sample() method used). The first fold is treated as a validation set, and the method is fit on the remaining 9 folds. The miss-classification error, is then computed on the observations in the held-out fold(predict() function). This procedure is repeated 10 times and each time, a different group of observations is treated as a validation set. The accuracy after applying this method on the flag dataset is 51.5%.

Comparing the accuracy found from the previous methods we end up with the following table. We can see that Random forests performed better than any other model in predicting successfully the religion based on the flags characteristics. However, the accuracy percentage is still rather low in general. Further analysis might be performed in two directions. The first one is to check whether our methods still suffer from a high bias or high variance problem and then fit other classification methods to check if this problem is fixed. The second one is the case that there might be no strong connection between the characteristics of the flags and the religion of the country for some cases. Domain expertise from scientists who study Vexillology would have been helpful in this case in order to evaluate if such a percentage of accuracy is acceptable, or whether they expected it to be larger.

Method	Accuracy Percentage
Classification Tree	24.2%
Pruned Tree	41.93%
Bagging	43.3%
Random Forest	51.5%

[6] An Introduction to Statistical Learning. Gareth James. Daniela Witten. Trevor Hastie. Robert Tibshirani. Springer Texts in Statistics.

5. Clustering[7]

For clustering we will use the PAM (partition around medoids) Clustering Algorithm. The algorithm is intended to find a sequence of objects called medoids which are always restricted to be members of the data set and are centrally located in clusters. The algorithm has the following phases:

1. In the first phase, choose k random objects to become the medoids.
2. Assign every other observation to its closest medoid.
3. For each cluster, identify the observation that would yield the lowest average distance if it were to be re-assigned as the medoid. If so, make this observation the new medoid. If at least one medoid has changed, return to phase 2. Otherwise, end the algorithm.

As we see this algorithm requires the calculation of distances. A careful selection of the dissimilarity measure must be used. Due to the fact that the variables of the dataset are both numeric discrete (Number of something) and binary categorical we will use the Gower distance[8](daisy function of cluster package). In the Gower distance, the dissimilarity between two rows is the weighted mean of the contributions of each variable (quantitative (interval): range-normalized Manhattan distance is used, qualitative: Jaccard's distance is used[9]).

To select the appropriate number of clusters for the PAM method, we use the silhouette plot. The silhouette plot measure is calculated for each observation to see how well it fits into the cluster that it's been assigned to. This is done by comparing how close the object is to other objects in its own cluster with how close it is to objects in other clusters. Values near one mean that the observation is well placed in its cluster; values near 0 mean that it's likely that an observation might really belong in some other cluster[10].

We calculate the silhouette width in the flag dataset 2 to 15 for the PAM algorithm and we see that 13 clusters yield the highest value, as shown in Image 6. Additionally, the silhouette plot, for the 13-cluster pam solution is illustrated in Image 7. The plot indicates that there in is a good structure with most observations seeming to belong to the cluster that they're in. This is not true for cluster 5 though where no structure has been found. However, this is the most appropriate number of clusters based on the overall average silhouette width among clusters 2-15. In the next § the will try to interpret the different clusters.

[7] PAM: Partitioning Around Medoids, UMB. URL
[8] Gower Distance. URL.
[9] Gower Similarity Coefficient. URL.
[10] Cluster Analysis. Berkeley URL.

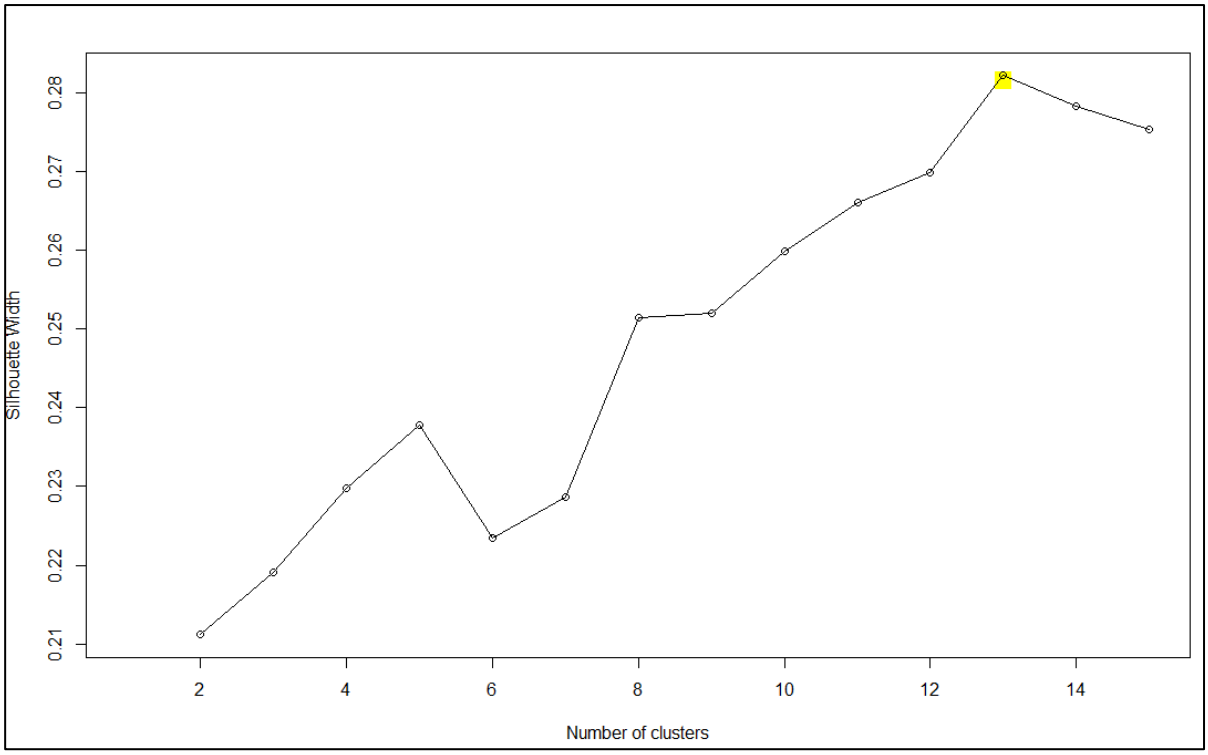


Image 6: Silhouette width of 2-15 clusters using the PAM algorithm

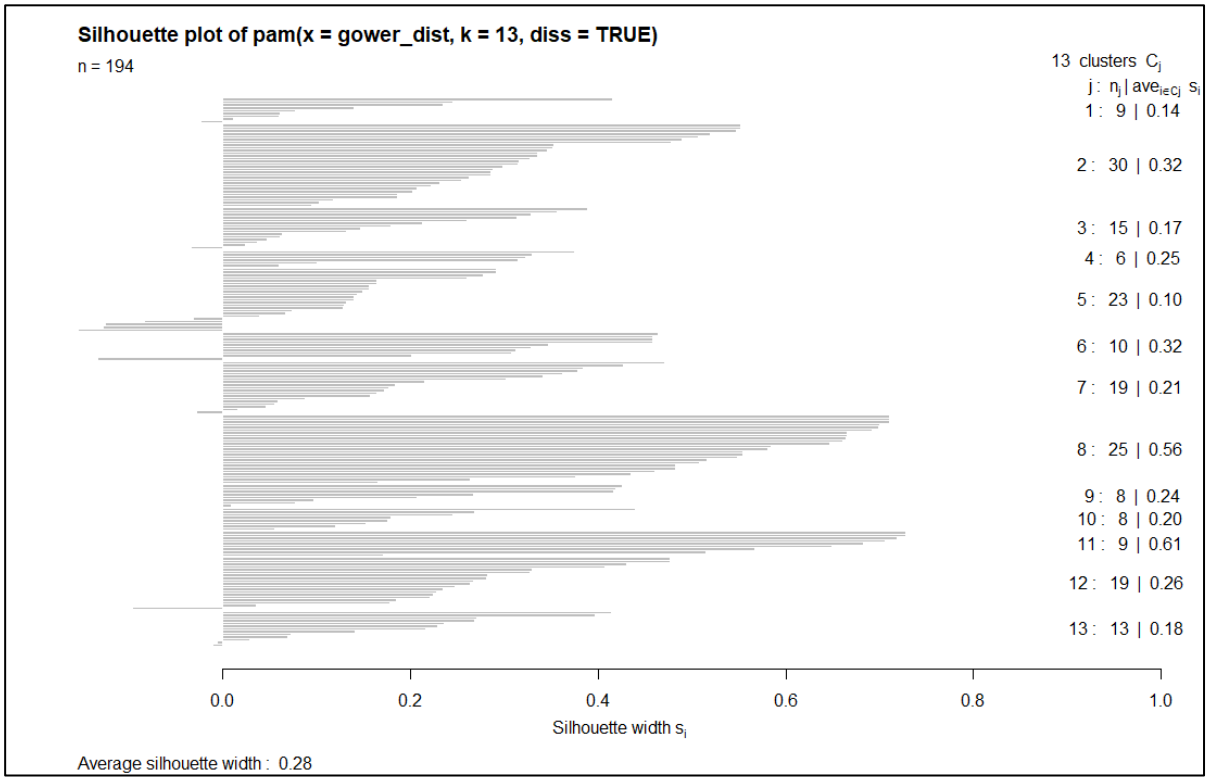


Image 7: Silhouette Plot

5.1. Cluster Interpretation

In order to visualize many variables in a lower dimensional space a dimension reduction technique that tries to preserve local structure so as to make clusters visible in a 2D or 3D visualization can be used. We will use the t-SNE[11]. The t-SNE algorithm focuses to preserve the local distances of the high-dimensional data in some mapping to low-dimensional data. We can see that clusters 2, 7 and 9, 11, 12 are well separated.

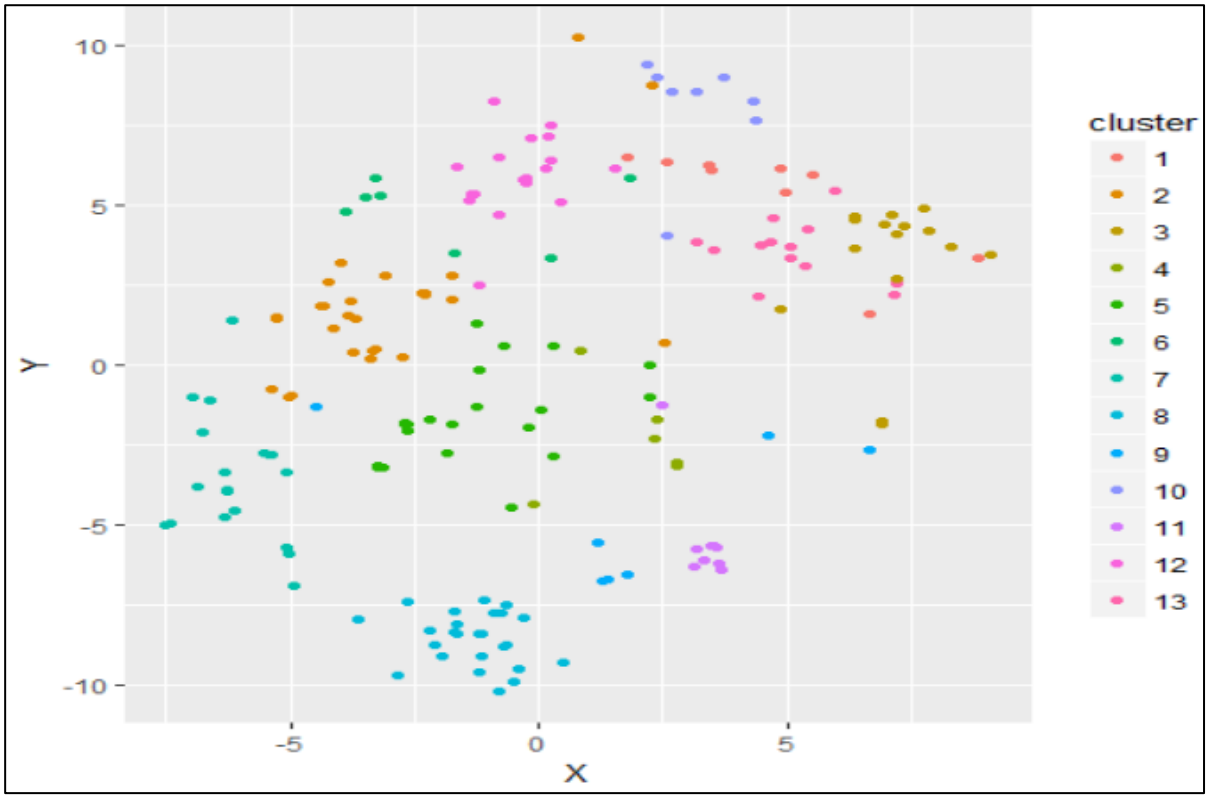


Image 8: Visualization of Clusters

In order better to understand the clusters found we combine the information from the summary statistics of all features grouped by each cluster with the information given by the bar plots for each feature per cluster. We indicatively provide a bar plot of the feature *main hue red.* We will not include all the bar plots of features in this report. Instead, we will provide a table that summarizes the main characteristics of each cluster as derived from summary and bar plots. For example, what we can understand from the bar plot of Image 9 is that 93% of the observation of cluster two have as dominant color the red. Also red is prevalent in clusters 5,6,12.

[11]t-Distributed Stochastic Neighbor Embedding (t-SNE). URL.

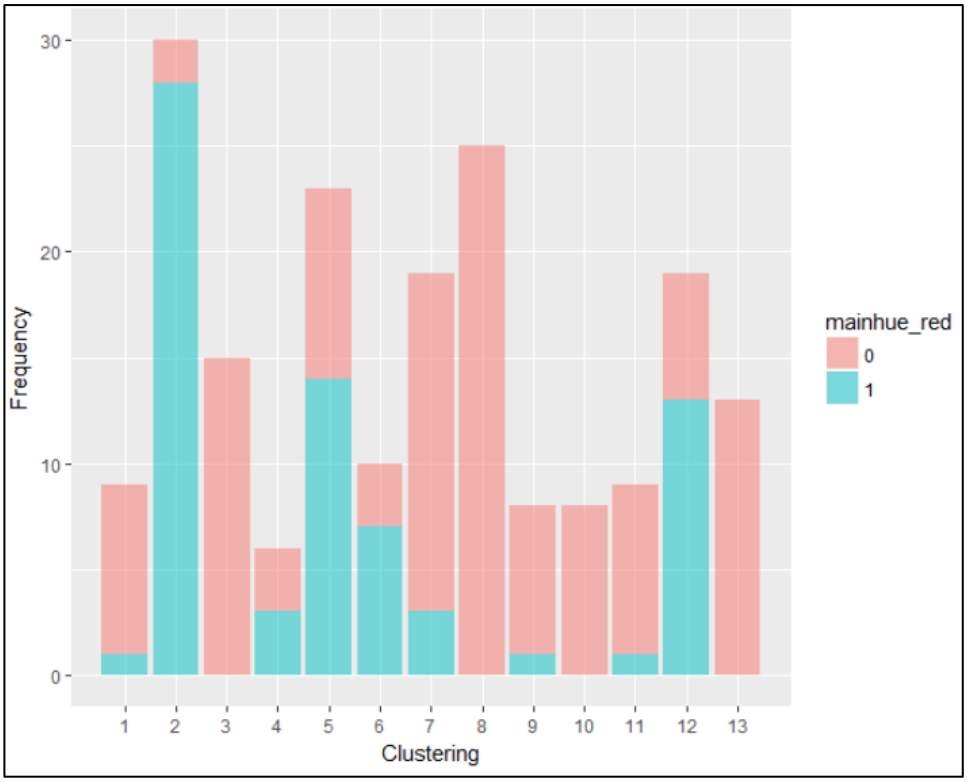
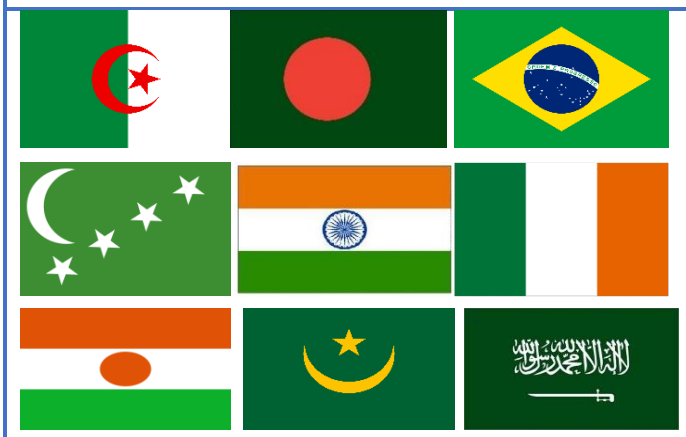
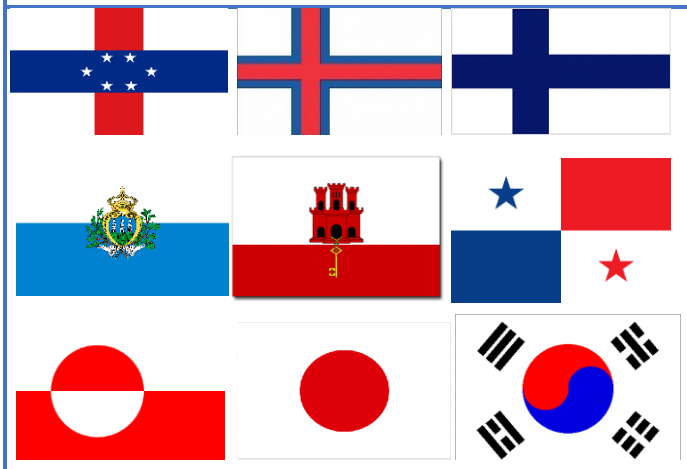
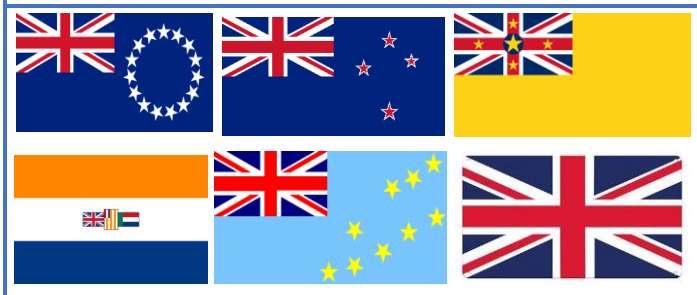
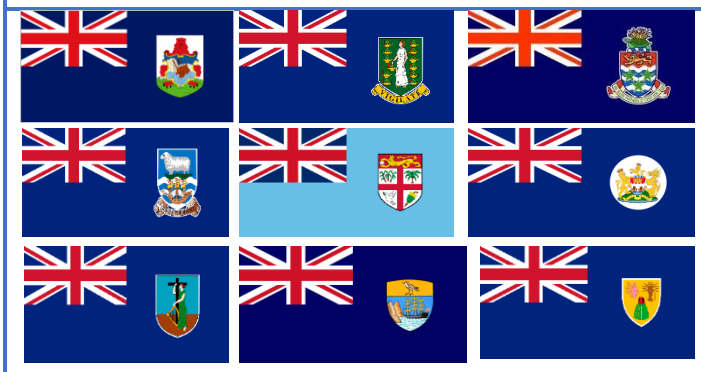


Image 9: Main hue red per cluster

Indicative Flags	Description of cluster and countries
	**1st cluster:** • Main hue: Green • Bottom-right: Green • 4-6 colors: Red, Gold, Black, white • triangles **Countries:** Dominica, Guyana, Jamaica, Jordan, Sao-Tome, Soloman-Islands, Vanuatu, Zimbabwe

Indicative Flags	Description of cluster and countries
	**2nd cluster:** • Main hue: Red • Bottom-right: Red • Top-left: Red • Other colors: White, Green, Gold • Mainly 2 colors **Countries:** Albania, Austria, Bahrain, Bhutan, Canada, China, Denmark, Haiti, Indonesia, Iran, Kampuchea, Laos, Lebanon, Luxembourg, Maldive-Islands, Malta, Monaco, Morocco, Netherlands, Norway, Peru, Puerto Rico, Spain, Switzerland, Thailand, Tonga, Tunisia, Turkey, USSR, Vietnam
	**3rd cluster:** • Main hue: Green • Bottom-right: Green • Top-left: Green • 2-3 colors: Green, Red, White • Crescent, Circles **Countries:** Algeria, Bangladesh, Benin, Brazil, Comorro-Islands, India, Ireland, Libya, Mauritania, Niger, Nigeria, Pakistan, Saudi-Arabia, Sierra-Leone, St-Vincent
	**4th cluster:** • Main hue: Blue/Red • Bottom-right: Blue/Red • Top-left: Blue/Red • Inanimate **Countries:** American-Samoa, Belize, Guam, Parguay, Romania, Venezuela
	**5th cluster:** • Main hue: Red • Bottom-right: Red • Top-left: Blue • Other colors: White, Gold • Sun/stars **Countries:** Andorra, Antigua-Barbuda, Bulgaria, Burma, Central-African-Republic, Chad, Chile, Colombia, Ecuador, France, French-Guiana, French Polynesia, Kiribati, Liberia, Liechtenstein, Malaysia, Mongolia, Philippines, Portugal, Taiwan, USA, Western-Samoa, Yugoslavia

Indicative Flags	Description of cluster and countries
	**6th cluster:** • Main hue: Red • Bottom-right: Black • Top-left: Red • Other colors: White, Green, Gold • Mainly three stripes **Countries:** Angola, Egypt, Iraq, Kenya, North-Yemen, Papua-New-Guinea, South-Yemen, Sudan, Syria, UAE
	**7th cluster:** • Main hue: White • Bottom-right: White • Top-left: White • Other colors: Blue, Red, Gold • 2-3 colors **Countries:** Anguilla, Burundi, Cyprus, Czechoslovakia, Faeroes, Finland, Gibraltar, Greenland, Japan, Netherlands-Antilles, Panama, Poland, Qatar, San-Marino, Singapore, South-Korea, Trinidad-Tobago, Uruguay, US-Virgin-Isles
	**8th cluster:** • Main hue: Blue • Bottom-right: Blue • Top-left: Blue • Other colors: White, Red, Gold, Black • 2-3 colors **Countries:** Argentina, Argentine, Bahamas, Barbados, Botswana, Costa-Rica, Cuba, Dominican-Republic, El-Salvador, Greece, Guatemala, Honduras, Iceland, Israel, Lesotho, Marianas, Micronesia, Nauru, Nepal, Nicaragua, North-Korea, Somalia, St-Lucia, Swaziland, Sweden
	**9th cluster:** • British Commonwealth flags • 3-5 colors • red, blue, white, crosses, saltires **Countries:** Australia, Cook-Islands, Djibouti, New-Zealand Niue, South-Africa, Tuvalu, UK

Indicative Flags	Description of cluster and countries
	**10th cluster:** • Main hue: Gold • Bottom-right: Gold • Other colors: Red, Black, White, • 2-3 colors **Countries:** Belgium, Brunei, Germany-DDR, Germany-FRG Mozambique, Sri-Lanka, Uganda, Vatican-City
	**11th cluster:** • Blue British Commonwealth flags with animate and inanimate images • Main hue: Blue • Bottom-right: Blue, • 6-7 colors, • crosses, saltires, quarters **Countries:** Bermuda, British-Virgin-Isles, Cayman-Islands Falklands-Malvinas, Fiji, Hong-Kong, Montserrat, St-Helena, Turks-Cocos-Islands
	**12th cluster:** • Main hue: Red • Bottom-right: Green • Top-left: Red • Other colors: Gold, White, Black • Sun-stars **Countries:** Bolivia, Burkina, Cape-Verde-Islands, Congo, Gambia, Ghana, Grenada, Guinea, Guinea-Bissau, Hungary, Ivory-Coast, Malagasy, Malawi, Mauritius, Oman, Rwanda, Seychelles, Surinam, Togo
	**13th cluster:** • Main hue: Green • Bottom-right: Red • Top-left: Green • Other colors: Gold, White, Black **Countries:** Cameroon, Equatorial-Guinea, Ethiopia, Gabon, Italy, Kuwait, Mali, Mexico, Senegal, St-Kitts-Nevis, Tanzania, Zaire, Zambia