**IST 707 Final Report: Refugee Asylum Seekers**

**Introduction**

The complex global refugee crisis presents challenges for developing and developed countries alike. As persecution, climate change, conflict and poverty become more acute, millions of refugees are being forced to flee their homes and seek extended residence in peripheral countries. Some of these countries, facing their own economic and developmental issues, do not have the capacity or public will to support these refugees for significant periods of time, increasing instability and threatening a disjointed and potentially xenophobic response.

Natural disasters and conflict in countries such as South Sudan, Myanmar, Venezuela and Syria have produced millions of refugees that are left to live in austere conditions. With the advent of the Coronavirus, ravaged medical infrastructure and a decreasing availability of basic necessities, these refugees face a grim outlook as countries look inward to ameliorate their own public health crises. Aid organizations, which have historically stood up to provide humanitarian assistance, must navigate restricted transportation networks and a lack of funding, hamstringing their life-saving capacity.

International organizations, such as the United Nation’s High Commissioner for Refugees (UNHCR), were created for the sole purpose of responding to this refugee crisis. Through private donations and UN member nation burden-sharing, UNHCR seeks to: ease pressure on host countries, enhance refugee self-reliance, expand access to third country solutions and support setting conditions conducive for a safe return to host countries. As funding and involvement from member states decreases, UNHCR must effectively prioritize flashpoint locations to maximize their response options amidst increasing uncertainty.

**Analysis and Models**

***About the Data***

**Initially, data was pulled from the UNHCR’s website. The original data set was comprised of 15 tables (sheets) in Excel each with similar attributes. For the purposes of this project, only 5 tables were used: 1 and 2, which covered Destination country (where applications for asylum are submitted)) totals, and 3-5, which covered Origin country (where applicants come from) totals. All tables include a base of similar attributes with raw aggregate totals per year for application numbers from 2010-2014, with countries in rows. The tables also had an Aggregate Total and Scaled Total columns already included. The scaled attributes included applications per 1,000 inhabitants and applications per 1 USD/GDP per capita. Each total was then in another preset ordinal rank column for all countries in the table. The most cleaning that had to occur with the initial data was converting Rank attributes (labeled clearly in 6 different attributes) from numeric to categorial.**

|  |  |
| --- | --- |
| **Original Attributes (UNHCR)** | |
| **Attribute** | **Description** |
| **Country** | **List of country name (same in both Origin and Destination tables)** |
| **2010** | **Total aggregate count of applications either to or from the country (depending on table) in the year 2010** |
| **2011** | **Total aggregate count of applications either to or from the country (depending on table) in the year 2011** |
| **2012** | **Total aggregate count of applications either to or from the country (depending on table) in the year 2012** |
| **2013** | **Total aggregate count of applications either to or from the country (depending on table) in the year 2013** |
| **2014** | **Total aggregate count of applications either to or from the country (depending on table) in the year 2014** |
| **Total** | **Total numeric count of applications from 2010-2014** |
| **Annual change ’14-‘13** | **Percent change in applications from 2013 to 2014** |
| **Rank2014** | **Categorical ranking of high number of applications for just 2014** |
| **Rank2010-2014** | **Categorical ranking of high number of applications for 2010-2014** |
| **Per1000Inhabitants2014** | **Scaled number of applications per 1000 inhabitants in the country (AKA a control for population) for 2014** |
| **Per1000Inhabitants2010-2014** | **Scaled number of applications per 1000 inhabitants in the country (AKA a control for population) for 2010-2014** |
| **Rank1000\_2014** | **Categorical ranking of 1000 inhabitant scaled applications for just 2014** |
| **Rank1000\_10-14** | **Categorical ranking of 1000 inhabitant scaled applications for 2010-2014** |
| **Per1USDGSP2014** | **Scaled number of applications per 1 USD/GSP per capita in the country (AKA control for GDP) for 2014** |
| **Per1USDGSP2010-2014** | **Scaled number of applications per 1 USD/GSP per capita in the country (AKA control for GDP) for 2010-2014** |
| **RankGDP\_2014** | **Categorical ranking of applications per 1 USD/GDP per capita for just 2014** |
| **RankGDP\_2010-2014** | **Categorical ranking of applications per 1 USD/GDP per capita for 2010-2014** |

**In addition to relabeling the Rank attributes as categorical, the data had to be cleaned to remove the top and bottom chunk of rows for each table. The initial format of the sheets included supplementary information and instructions on how to analyze the results. While useful, it resulted in several NAs when read into R and were removed when imported as data frames for the purpose of this project.**

**After initial analysis, which is to be discussed below, it was determined that additional data would need to be pulled in order to better make sense of the differences in scaling for different countries. The addition of the data from the Freedom House aimed to provide greater insight onto why different countries would have a high number of applications. The next pull of data from Freedom House includes the following indices for each country:**

|  |  |
| --- | --- |
| **Appended Attributes (Freedom House)** | |
| **Attribute** | **Description** |
| **Region** | **Geographical region, as a factor (Europe or Asia)** |
| **Free** | **Indication of freedom status as Free, Not Free, or Partially Free (F, NF, PF)** |
| **PoliticalRights** | **Numeric rating from 1-7 (1 most free, 7 least free)** |
| **CivilLiberties** | **Numeric rating from 1-7 (1 most free, 7 least free)** |
| **Governance** | **Function of government, score from 1-12 (1 poor, 12 best)** |
| **Expression** | **Freedom of expression and belief rating 1-16 (1 poor, 16 best)** |
| **Electoral** | **Electoral process rating 1-12 (1 poor, 12 best)** |
| **PolPluralism** | **Political pluralism and participation rating 1-12 (1 poor, 12 best)** |
| **IndividualRights** | **Personal authority and individual rights rating 1-16 (1 poor, 16 best)** |
| **RuleofLaw** | **Rule of law rating 1-16 (1 poor, 16 best)** |
| **Organization** | **Association and organizational rights 1-12 (1 poor, 12 best)** |
| **PolRights** | **Score from 1-40 (total of Electoral, PolPluralism, and Government)** |
| **Liberties** | **Score from 1-60 (total of Expression, Organization, IndividualRights, and RuleofLaw)** |

Once the data from the UNHCR and Freedom House were located, they needed to be consolidated. In order to properly do this, the data was matched on a key. The key was country in this case and the data was merged in Excel rather than in R. The Freedom House Index values were appended to the original data of the UNHCR. Once the tables were properly merged in Excel it needed to be cleaned in order to analyze for future modeling.

The rows that were either totals or did not have a country related were removed from the tables. Then the data was given proper headers to then be able to easily fix column types. All data was read in as “character” and needed to be changed to either factor or numerical. Country, Region, Free, Political Rights, and Civil Liberties were converted to factors while all other variables were transformed to numerical.

Once all columns were associated to their proper type, another data frame was created to make cuts within the data either to be binned. Binning was done with three to four bins and was based on the range of the data in each of the data. These separate data frames allowed for flexibility for modeling that will be shown in the below analysis.

***EDA***

**Initial evaluative analysis of the data revealed a lot of what would be expected. Since the UNHCR data was pulled from a site conducting a report in 2014, part of their goal was to compare across years how the total number of applications was affected. The below first set of EDA includes histograms for 2010 and 2014 to compare the change in years. Each has an obvious right skew with the majority of applications residing in the lower range per country, however, as years go on the bin widths widen. The comparison of the bin sizes and max and mins of each histogram per year demonstrates a near tripling in the total number of applications from 2010 to 2014.**

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

**Initial analysis was also conducted on the aggregate totals and scaled total values. The below are depictions of the top 10 ranked countries for each category for origin countries, with the top two in royal blue:**

A picture containing drawing

Description automatically generated

Figure 1: Total Aggregate Applications

A picture containing drawing

Description automatically generated

Figure 2: Total per 100 Inhabitants

A screenshot of a cell phone

Description automatically generated

Figure 3: Total per1 USD/GDP

**From the creation of the above visualizations, it was clear the Germany and the U.S. were quite high-volume destination countries, but more peculiarly, Turkey was listed as a country with a high-volume of applications in both the aggregate and GDP scaled models.**

Once application totals were inspected, further analysis using the Freedom House indices was conducted to better understand the behavior of the country scores for both the origin and destination countries. Destination countries tended to have most labeled as F (red) and very few as PF (blue) or NF (green). The only NF was, as expected, Turkey. The below are visual representations of the total scores for PoliticalRights and CivilLiberties for destination countries.

A picture containing people

Description automatically generated

Figure 4: Destination Political Rights Scores

A screenshot of a cell phone

Description automatically generated

Figure 5: Destination Civil Liberties Scores

The same visualizations were created for the origin countries – however, the results are a bit more scattered across the board – there are two countries labeled “Free”, Ghana and India, that are useful to note in modeling.

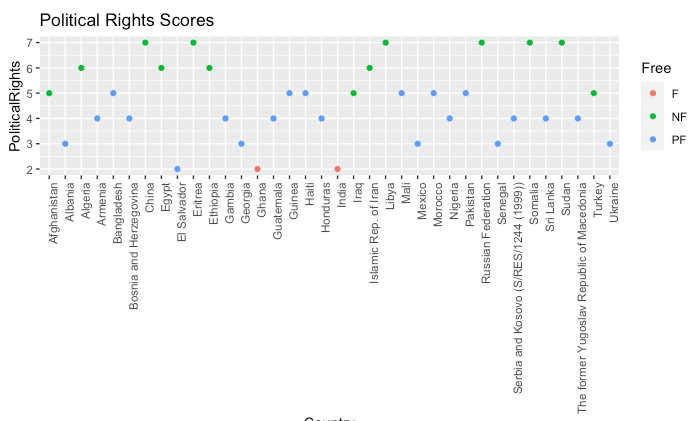
****

Figure 6: Origin Political Rights Scores

A picture containing clock

Description automatically generated

Figure 7: Origin Civil Liberties Scores

**Models**

Association Rule Mining

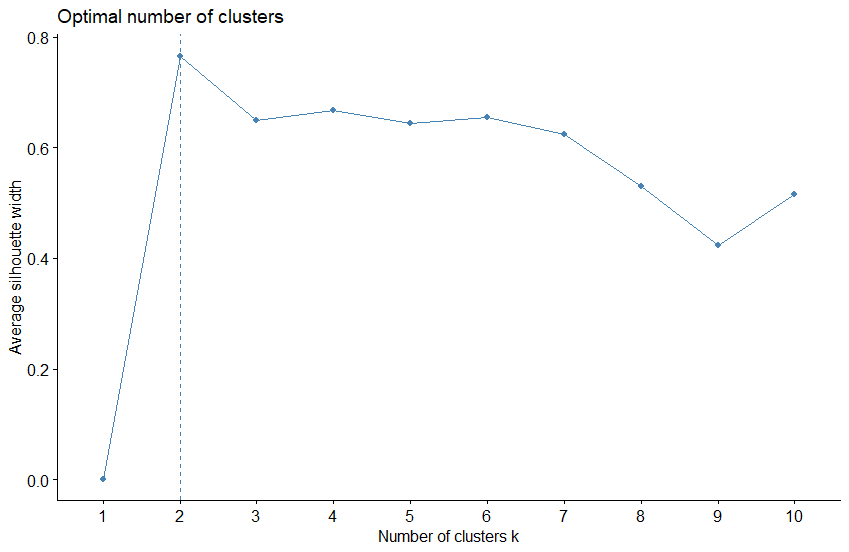
The purpose of Association Rule Mining (ARM) is to determine interesting and viable relations in variables to see if the data can be explained by certain variables. In the case of this analysis, the team was interested in looking at where refugees were going to and originating from. With this the team was interested if the variables could define a trend that could be pointed towards with a strong and reliable correlation towards predicting trends.

The trends that come from the rules can be identified with reason that if something occurs then it can be explained by certain factors. The reliability of these rules can be deemed from the confidence, support, and lift. With these values the rules can be identified to be trusted for future analysis when collecting further data on more countries.

K-Means

Refugee destination and refugee origination countries were grouped according to K-means clustering, given the extensive amount of numerical data in the dataset. The intent was to determine if freedom indices versus refugee numbers and other economic considerations were a driving factor in country similarity. K means is an unsupervised model that defines a user-specified number of centroids and clusters groups of observations to these centroids based on variable values of both the centroid and observations. For each newly established cluster, the algorithm then updates the mean of these points, and selects a new centroid. This process repeats until the centroids remain the same.

In order to determine the number of centroids, the “silhouette method” was implemented, which seeks to minimize the average distance within each cluster compared the average distance to the next nearest cluster. The optimal number of clusters is the point where the average silhouette width is closest to 1, in our case that was true for 2 clusters in every case. Here is a representation:



K-means was run at 2, 3 and 5 clusters to understand how a larger number of centroids would group the data. These different choices for k had a significant impact on the results and due to the relatively small number of countries in each table, it was decided that K>2 did not portray an accurate representation. Analysis was continued utilizing a K of 2. After clusters were assigned, Euclidean distance was measured between the destination countries, and origin countries in two distinct ways: including all variables or selecting the freedom indices only.

Hierarchical Agglomerative Clustering

Hierarchical Agglomerative Clustering (HAC) was run to assess the comparison between this approach and k-means. Given that K- means chooses centroids at random and updates those centroids based on their relationship with the data, a “bottom-up” driven approach was desired. Hierarchical agglomerative clustering achieves this goal by illustrating the distance between observations and allows us to determine the true similarity between them. This is done by successively merging the two closest observations into small clusters until all clusters are merged and only one cluster remains.

The HAC model first calculated the Euclidean distance between each observation to create a dissimilarity object. From there, using the “complete” method, which calculates the complete linkages of the dataset and finds similar clusters, the object was transformed into a dendrogram with countries on the y axis and distance on the x axis. This was completed using two methods: all variables and the freedom indices.

Decision Trees

While HAC and K means focus on the similarity of particular observations, the overarching goal is to assess how economic and political factors and refugee numbers would determine if a country was “free”, “partially free” and “not free” as compared to Freedom House’s freedom index standards. While Freedom House assigns a rating and overall classification to each country, identifying their “freeness” this classification did not take refugee numbers into account.

A decision tree is a simple classification technique that can make predictions on a specific problem when offered a series of “questions” or information about the data and a training dataset from which it has learned the “answers.” A decision tree is a hierarchical structure that has a root node, internal nodes and terminal nodes. A root node has no higher node in the hierarchy and often greatly divides the data for further “questions”. An internal node has exactly one incoming edge and one or more outgoing edges. A terminal node has exactly 1 incoming edge and no outgoing edges - it is the classification decision. Each node may be connected to other nodes via edges, which can be considered the answer to the question posed by that node. For example, a node might question “Region = Middle East North Africa?” and the edges might be “yes” or “no,” indicating a path to take to further internal nodes or terminal nodes.

The decision trees implemented in this analysis include all available variables, including freedom indices and refugee numbers. The complexity parameter (CP) is set to 0 which gives keeps the tree from pruning splits that are less worthwhile, and all variables will be able to form splits – given the small dataset, the saved computational time is negligible. Minimum splits are set to 2, stating that there must be at least two observations in a node for the split to be attempted. This model risks overfitting due to the low CP value.

Random Forest

With the decision to run Random Forest models the intent was to effectively identify whether a country can be defined as Free, Not Free, or Partially Free. This modeling is again a supervised model that generates decision trees and then outputs the most predicted values from those trees. With this modeling on such a small data set it was found that models were easily predictable, but important to show that the data on hand was valuable in future findings.

**Results**

*ARM Assessment*

For the Association Rule Mining the team initially looked to run rules on the tables that involved the information of where the asylum seekers were sending their applications.

lhs rhs support confidence lift

{2010=small, Rule of Law=mid-low} => {2014=small}    0.2 0.8 1.10

This rule shows that if the country had a small 2010 of applications submitted and low rule of law then the outcome would usually result in a small number of applications in 2014. This rule shows decent support and confidence. The lift is promising at 1.10 and shows that it can be generally dependent.

lhs rhs support confidence lift

{Free=F,Governance=mid-high}     => {2013=small}       0.3    0.8 1.07

The above rule shows that if the Freedom Index has indicated that the country is Free and their Governance is mid-high then they would have a small number of applications in 2013. The confidence is fairly high at 80 percent, but the lift again is showing that this is a rule that can be trusted.

lhs rhs support confidence lift

{2013=small-high,Liberties Score=F}   => {2012=small-high} 0.1    1    5

This rule shows that if the country had a small-high number of applications in 2013 and a Liberties score of Free then their 2012 was likely small-high. The small support and confidence of 1 is somewhat questionable in this case. The high lift at 5 too also can point that this rule is important, but should be taken with a grain of salt as there may have only be a handful of countries that follow this.

Next, the team wanted to see if when looking at where the asylum seekers were coming from, results would differ from the above rules that were generated from where the applications were being put in.

lhs rhs support confidence lift

{2013=highest,PolPluralism=mid} =>  {PolRights=mid-low}    0.36  0.8    1.9

This rule is showing that if the PolRights is mid-low then it is likely that the country had a highest 2013 and mid PolPluralism. This is interesting to show what can make up some of the Freedom Index values from both Freedom Index and UNHCR variables. The support at 36 percent and confidence of 80 percent are good. The lift of 1.9 is also a good indicator that this rule can be trusted.

lhs rhs support confidence lift

{2014=highest,Expression=low}    =>    {Free=NF}        0.36  0.86    2.2

This rule is showing that if 2014 had a highest number of seekers from the country and a low Expression then the country would be considered Free. The support and confidence show strong reason to believe in this rule. The lift also supports both the support and confidence and show that the rule can be valid.

lhs rhs support confidence lift

{Expression=low,Liberties Score=NF}    =>             {Governance=low}    0.36  0.86    1.9

This final rule is saying that if the Governance of the country is considered low then they have a strong chance of having a low Expression and a Not Free Liberties value. The support and confidence match the rule above and as previously discussed show good value to the rule. Then the lift is a bit lower than the previous but is still valid at 1.9 and shows strong promise to mark this rule as reliable.

From here the top rules with confidence, support, and then lift were run. With these rules the team took the top 5 from the where the seeker was submitting their applications and where they were from. With this, it gave a good indicator into the various effects on the data.

*Confidence Asylum Applications Submitted*

lhs rhs support confidence lift

{Governance=high,Rule of Law=mid-high} => {PolRights=high} 0.2 1 1.8

{Governance=high,Rule of Law=mid-high} => {PolPluralism=high} 0.2 1 1.4

{Governance=high,Rule of Law=mid-high} => {Political Rights=1} 0.2 1 1.4

{Governance=high,Rule of Law=mid-high} => {Liberties Score=F} 0.2 1 1.3

{Governance=high,Rule of Law=mid-high} => {Organization=high} 0.2 1 1.2

*Support Asylum Applications Submitted*

lhs rhs support confidence lift

{Organization=high,Liberties Score=F} => {Free=F} 0.78 1.00 1.1

{Free=F,Liberties Score=F} => {Organization=high} 0.78 1.00 1.2

{Free=F,Organization=high} => {Liberties Score=F} 0.78 0.91 1.2

{Political Rights=1,Organization=high} => {Free=F} 0.72 1.00 1.1

{Free=F,Political Rights=1} => {Organization=high} 0.72 1.00 1.2

*Lift Asylum Applications Submitted*

lhs rhs support confidence lift

{Expression=mid,Rule of Law=mid-low} => {Civil Liberties=2} 0.2 1 4.4 8

{2012=small,Expression=mid,Rule of Law=mid-low} => {Civil Liberties=2} 0.2 1 4.4 8

{2014=small,Expression=mid,Rule of Law=mid-low} => {Civil Liberties=2} 0.2 1 4.4 8

{2013=small,Expression=mid,Rule of Law=mid-low} => {Civil Liberties=2} 0.2 1 4.4 8

{2011=small,Expression=mid,Rule of Law=mid-low} => {Civil Liberties=2} 0.2 1 4.4

*Confidence Origin of Asylum Applications Submitted*

lhs rhs support confidence lift

{PolRights=mid-high,IndRights=mid-high} => {Civil Liberties=low} 0.21 1 2.5

{PolRights=mid-high,IndRights=mid-high} => {2014=highest} 0.21 1 1.0

{Organization=mid,IndRights=mid-high} => {Civil Liberties=low} 0.21 1 2.5

{Civil Liberties=low,IndRights=mid-high} => {2014=highest} 0.24 1 1.0

{2014=highest,IndRights=mid-high} => {Civil Liberties=low} 0.24 1 2.5

*Support Origin of Asylum Applications Submitted*

lhs rhs support confidence lift

{2013=highest,Liberties Score=NF} => {2014=highest} 0.64 1.00 1.0

{2014=highest,Liberties Score=NF} => {2013=highest} 0.64 1.00 1.1

{Civil Liberties=mid,Liberties Score=NF} => {2013=highest} 0.61 1.00 1.1

[{2013=highest,Civil Liberties=mid} => {Liberties Score=NF} 0.61 1.00 1.6

{2013=highest,Liberties Score=NF} => {Civil Liberties=mid} 0.61 0.95 1.6

*Support Origin of Asylum Applications Submitted*

lhs rhs support confidence lift

{PolPluralism=low,Rule of Law=low} => {PolRights=low} 0.27 1 3.3

{Free=NF,PolPluralism=low} => {PolRights=low} 0.33 1 3.3

{Civil Liberties=low,Expression=mid,Organization=mid} =>{Liberties Score=PF} 0.24 1 3.3

{Civil Liberties=low,Organization=mid,Rule of Law=mid-low =>{Liberties Score=PF} 0.24 1 3.3

{Free=NF,PolPluralism=low,Rule of Law=low} => {PolRights=low} 0.27 1 3.3

With a smaller data set these rules show some variation in that they look to be viable, but they can be deceiving with such good confidence numbers they were not overwhelmed by the outcomes. They also wanted to see the rules on the tables when the left-handed side was set to Free on the table of the Asylum applications submitted. The for the origin of the seekers they set the left-handed side to be Not Free. The rules are straightforward and common sense but still important to note when making further inspections into the countries.

*LHS Free Asylum Applications Submitted*

lhs rhs support confidence lift

{Free=F} => {Organization=high} 0.85 0.97 1.1

{Free=F} => {Liberties Score=F} 0.78 0.89 1.1

*LHS Not Free Origin of Asylum Applications Submitted*

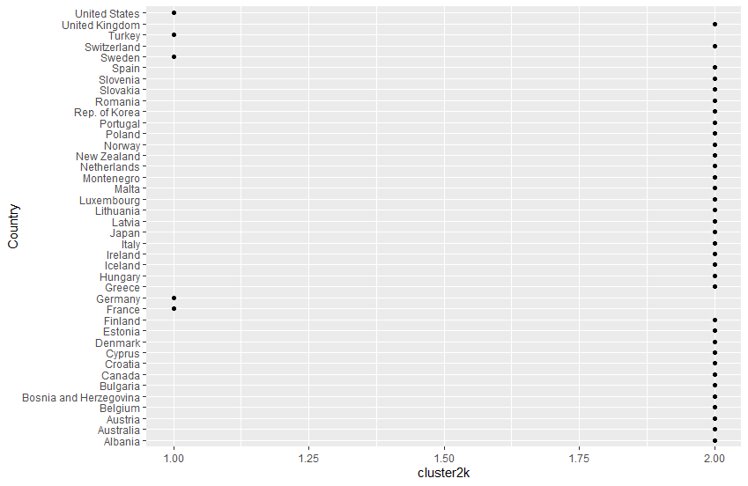
lhs rhs support confidence lift

{Free=NF} => {Governance=low} 0.39 1 2.2

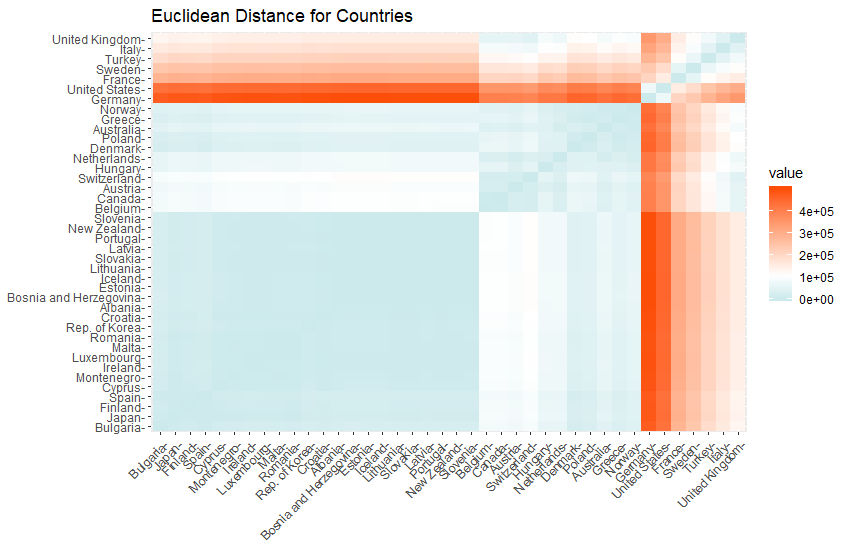
{Free=NF} => {Organization=low} 0.39 1 1.8

*Clustering Assessment*

Below are the results and visualizations of each data subset after performing K-means with 2 centroids, Distance Matrix Visualization and Hierarchical Agglomerative Clustering. The order of presentation illuminates the thought process and seeks to ascertain both root cause of refugee asylum applications and conditions present in countries with significant refugees.

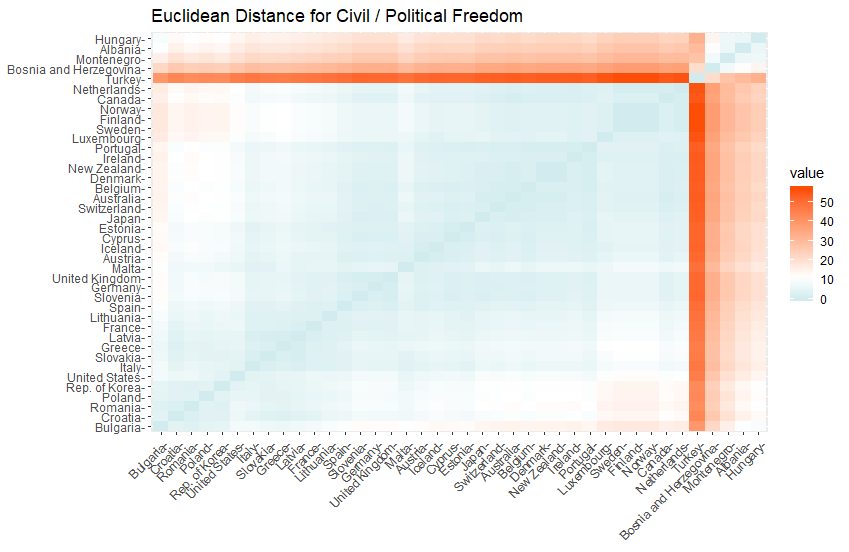


The first table includes countries in Europe and selected others where refugees sought asylum. K-means assigned two clusters with the United States, Turkey, Sweden, Germany and France in one and all other European countries and Australia, Japan, New Zealand in the other. Intuitively, one with knowledge of the subject matter would understand the similarity between countries such as the United States, Germany and France given their similar liberal worldviews, democratic nature and expected refugee intake. For these countries to be clustered with Turkey with its greatly differing views on democracy, was unexpected. To assess why Turkey was included in the first cluster, a distance matrix was performed using two methods: all variables included and freedom indices only.

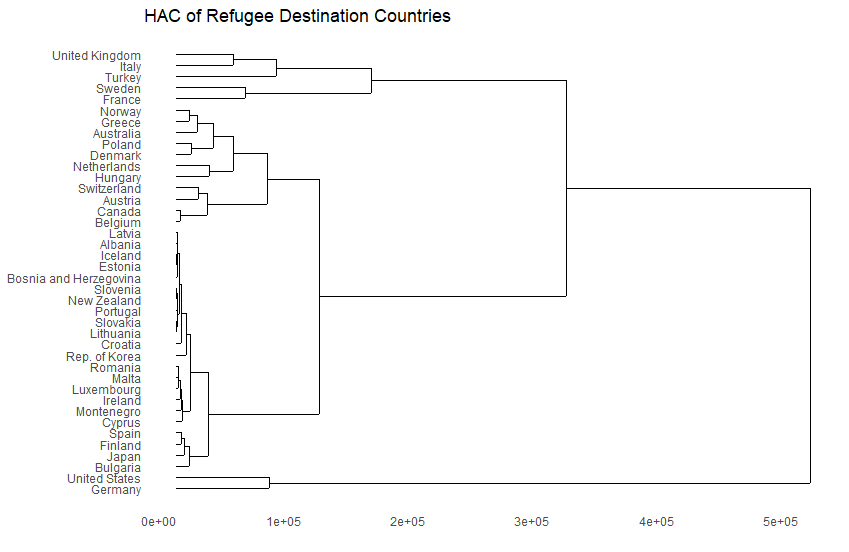


This visualization is a distance matrix with countries on both the x and y axes. This visualization describes the “distance” or difference between a selected country and all other countries. A large difference is colored in red, while small difference is colored in blue, with white in between. For example, when selecting the United States on the Y axis, one will notice that each corresponding intersection with an x axis country will result in a dark red color, except for where the United States intersects with itself, which is predictably blue (they are the same) and Germany which is a smaller difference.

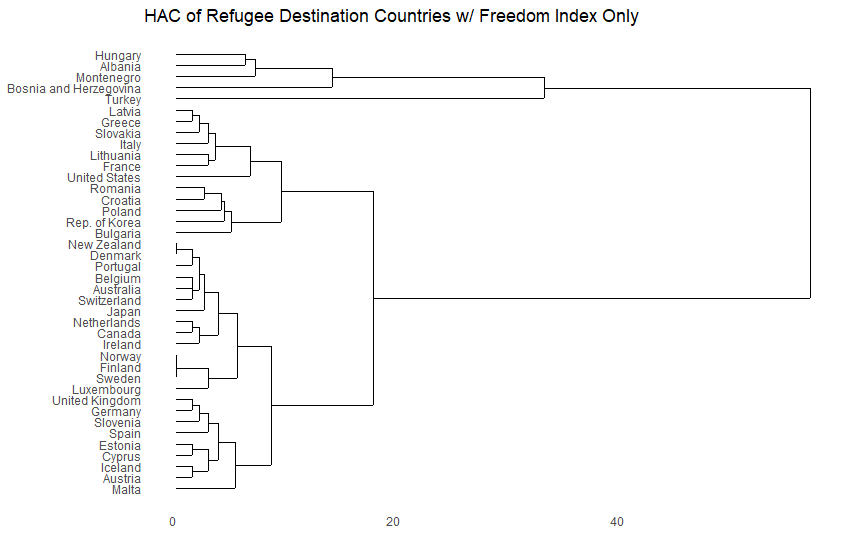
The first visualization shows two, arguably 3 distinct clusters: One including the US, Germany, Turkey, France, Sweden, Italy and the United Kingdom; and the other including All other European countries, New Zealand, Japan and Australia. This was expected given the results from the simple cluster assignments from K-means. While informative, it is important to understand that large values in certain variables have a propensity to skew the results and minimize the impact of other variables. Such a value includes the refugee application columns. To understand the similarity between these countries in terms of ideals, political freedom and civil liberties, a different visualization was completed using the freedom indices, shown below.



This visualization is more telling; Turkey, Bosnia and Herzegovina, Montenegro, Albania and Hungary are shown in one cluster with all other countries located in the other cluster. All these countries in the first cluster were rated as “not free” or “partially free” by Freedom House. There is a significant difference between Turkey and the other European countries from the prior figure, refugee numbers, GDP, population or other economic indicators may have been the leading cause for Turkey to be assigned in cluster 1 originally. This makes sense due to Turkey’s proximity to conflict-afflicted countries such as Syria, Iraq and Afghanistan as well as its gateway to Eastern Europe.

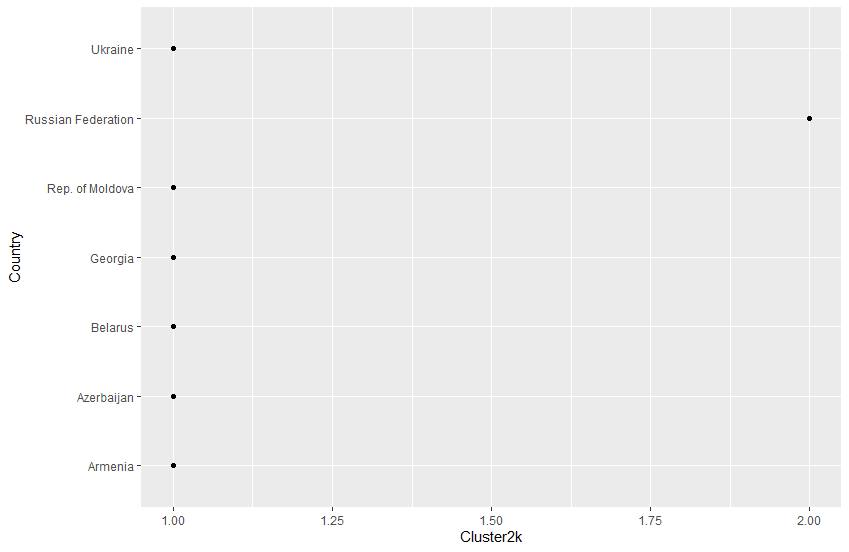


The above Hierarchical clustering dendrogram provides another interesting perspective: many of the smaller countries appear negligible and illustrate the previous point regarding refugee numbers. None of the countries, such as Estonia, Latvia or New Zealand, received significant refugee numbers and it was believed that this dendrogram provides a false assessment of what these countries represent. Turkey is once again clustered with Italy and the United Kingdom at a relatively small distance. All three of these countries accept large numbers of refugees. What is most interesting about this dendrogram is that the United States and Germany are still much “different” than all other countries, either by virtue of their population, political freedoms, liberties or simply refugee numbers.

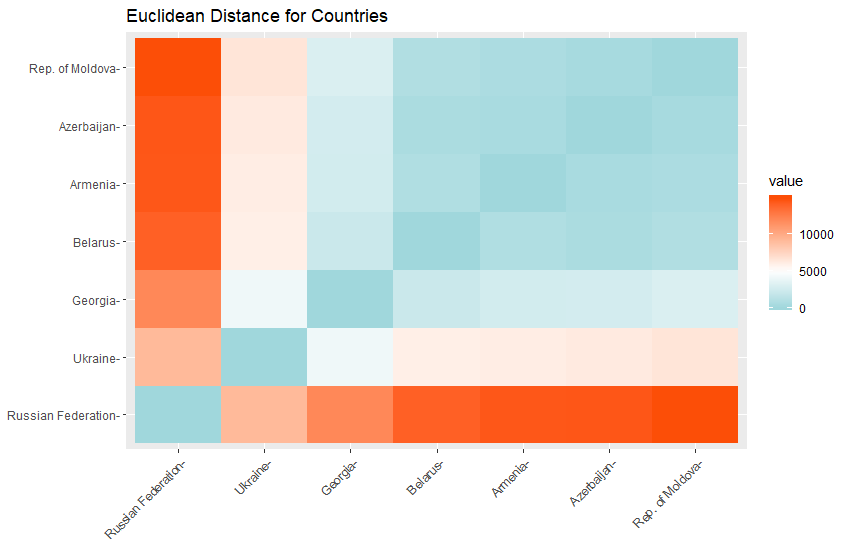


A HAC dendrogram with freedom indices only is more congruent with the observation shown in the freedom index distance matrix: Turkey, Bosnia and Herzegovina, Albania, and Hungary are vastly different from all other countries when accounting for ideals alone. Proximity, combined with a shared culture and religion with the origin countries might be accounting for the majority of Turkey’s refugee application numbers, not their freedom index.

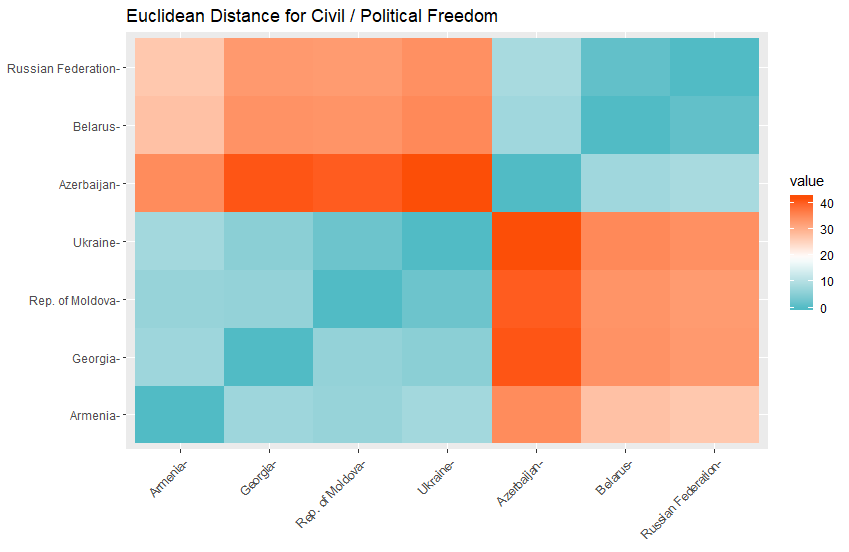
The second data subset included refugee asylum destination countries in Eastern Europe. 7 countries were included in this subset. A subset of this size is not ideal for K-means because there are so few datapoints. Russia was “clustered” with itself, alone which makes it difficult for one to make a prediction of its freedom index, population, GDP or refugee intake. Even accounting for the countries in the other cluster, it is difficult to reach a meaningful conclusion



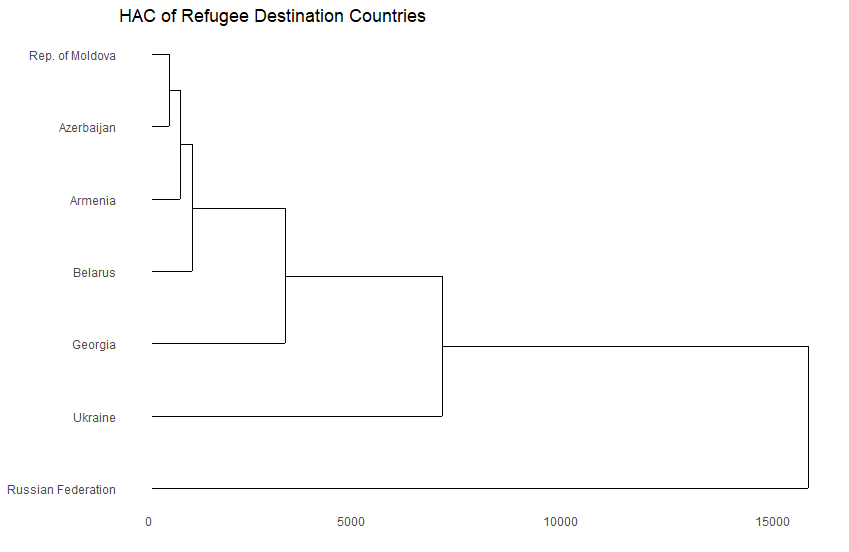
Including all variables, Russia is shown to be vastly different than the other countries. Its next closest neighbors, Ukraine and Georgia, are significantly different. It was believed that refugee numbers may have been the reason given the large “value” notated on the key. When accounting for freedom indices only, the “value” key is much smaller.

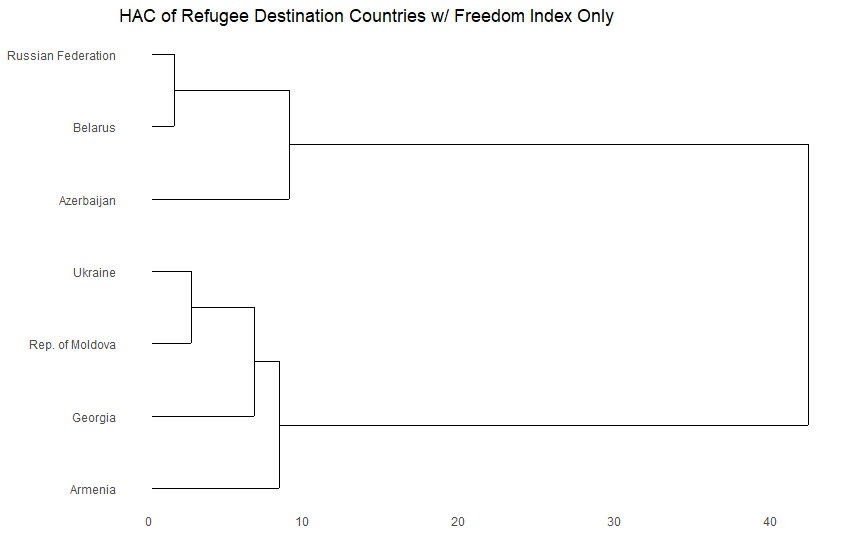


When accounting for freedom indices, there are two distinct groups: Russia, Belarus and Azerbaijan; and Ukraine, Rep. Of Moldova, Georgia and Armenia. All the countries included in this dataset are were members of the former Union of Soviet Socialist Republics. Azerbaijan and Belarus are more like Russia than all other former Republics, when not accounting for GDP, refugees, population and other economic factors. Russia has either intervened in militarily or outright invaded the countries in the other cluster. Russia has invaded both Georgia in 2008, and Ukraine in 2014, perhaps in response to these countries adopting a more western-influenced worldview.

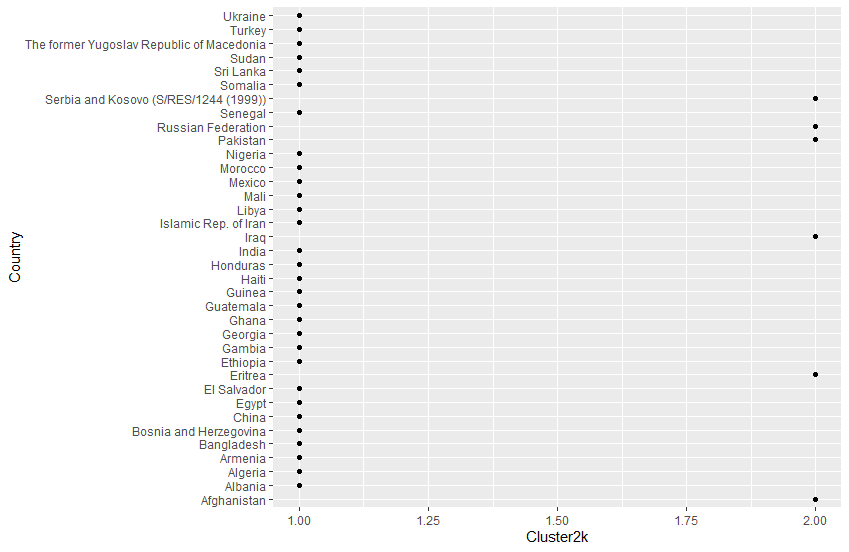


A HAC dendrogram of all variables shows how applications for Russia are influencing the clustering method. When the freedom index dendrogram is placed below, it is clearer. Ukraine, Moldova, Georgia and Armenia are approximately as similar to each other as Russia, Belarus and Armenia are to each other. It appears that each of these countries have similarly modeled their civil liberties and political freedoms.

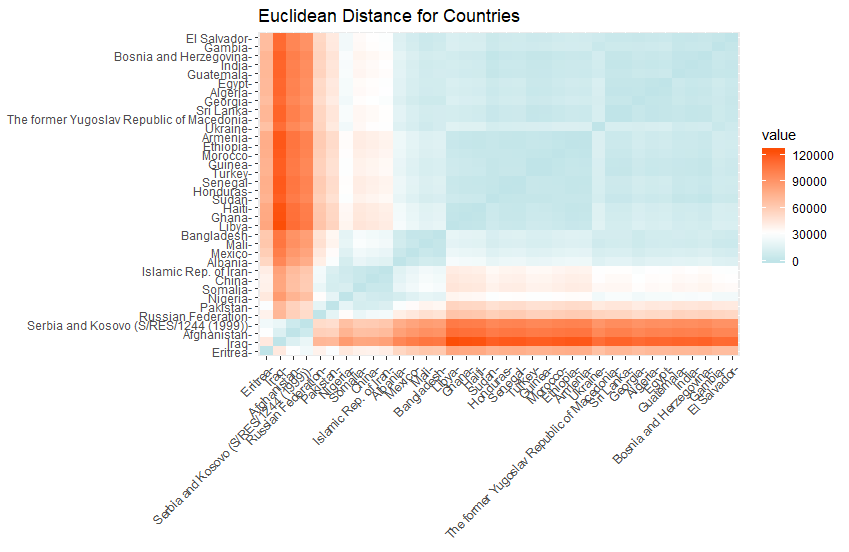




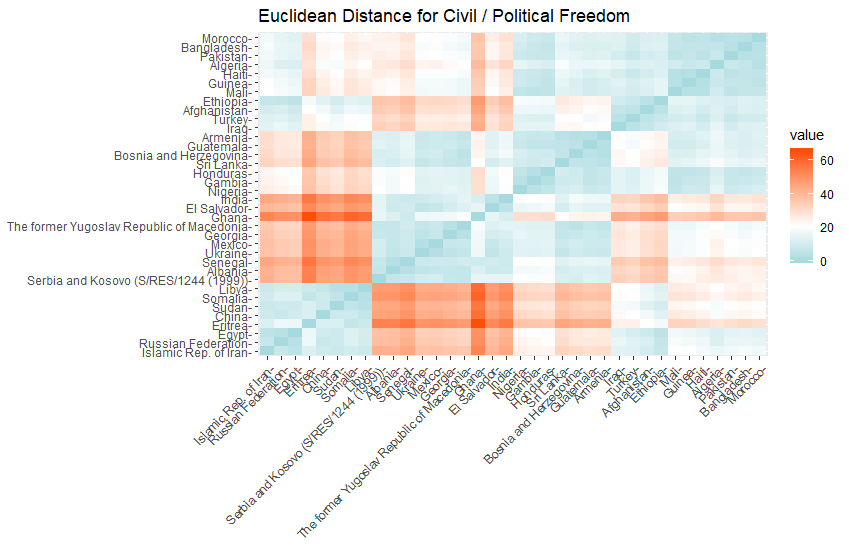
The following table is comprised of refugee origination countries. These countries’ refugee numbers are reflected by applications submitted to industrialized countries. There are two clusters: one including Serbia and Kosovo, Russia, Iraq, Afghanistan Eritrea and Pakistan; and the other including several African countries, Middle Eastern countries, South American and Caribbean countries. While all these countries have refugees, they have varying levels of freedom. The list includes countries from “free”, “not free” and “partially free” categories of the Freedom House index. The team sought to understand if these countries had a common unifying theme with which we could predict potential refugee crises.



In this distance matrix visualization, there are two clear clusters, one including Eritrea, Iraq, Afghanistan and Serbia Kosovo: and one including all other countries. It was obvious from this visualization and the previously assigned clusters that these countries were clustered because they had significant refugee numbers. All four of these countries have been recently involved in or are currently involved in extensive conflict.



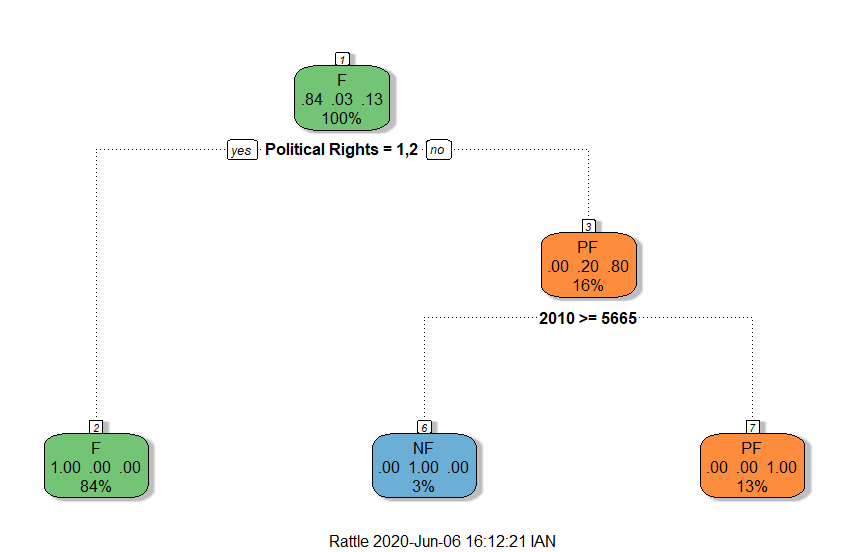
When adjusting for political freedom and civil liberties, a completely different picture is presented. Countries such as Iraq and Afghanistan are more like Russia, Egypt, Libya, Somalia and China than they are Eritrea and Serbia. There are free, not free and partially free countries that have experienced a significant refugee situation and that seems to be independent of that countries particular civil liberties and political freedom index. The biggest driving factor for refugees is understandably, conflict, natural disaster and ethnic tensions.



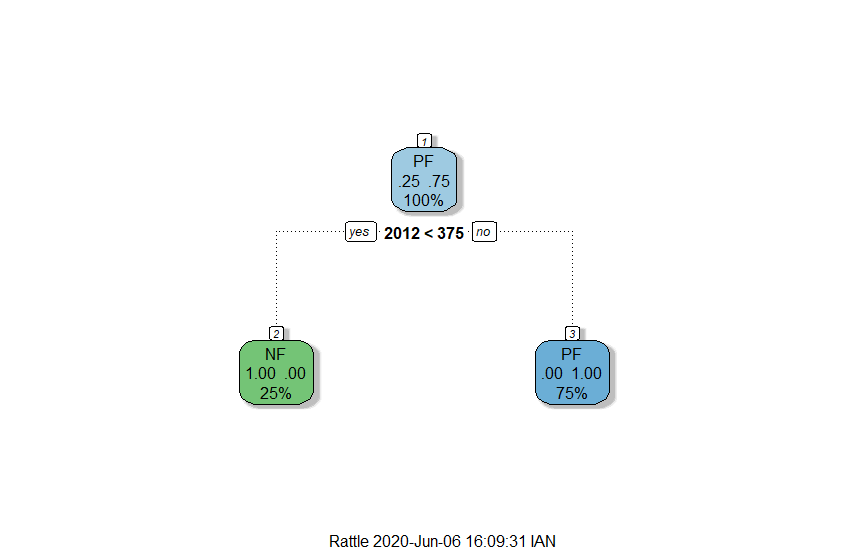
The clustering assessment, a collection of K-means, HAC and their corresponding visual displays is a simple and quick way to assess numeric data. The largest limitation is that in order to answer questions regarding refugee origination, destination, routes and the drivers of instability, a vast amount of data must be collected. This data will likely span both numeric qualitative and categorical variables which may make this method of assessment very difficult. This method would have been more effective if the data was previous scaled as it would have reduced the influence of refugee numbers, which disproportionately affected clustering and forced the team to calculate similarity using two distinct methods: all variables and freedom index alone. A scaled refugee score would have allowed an analyst to spot nuance between countries when accounting for all variables instead of refugee numbers greatly skewing the data.

*Decision Trees*

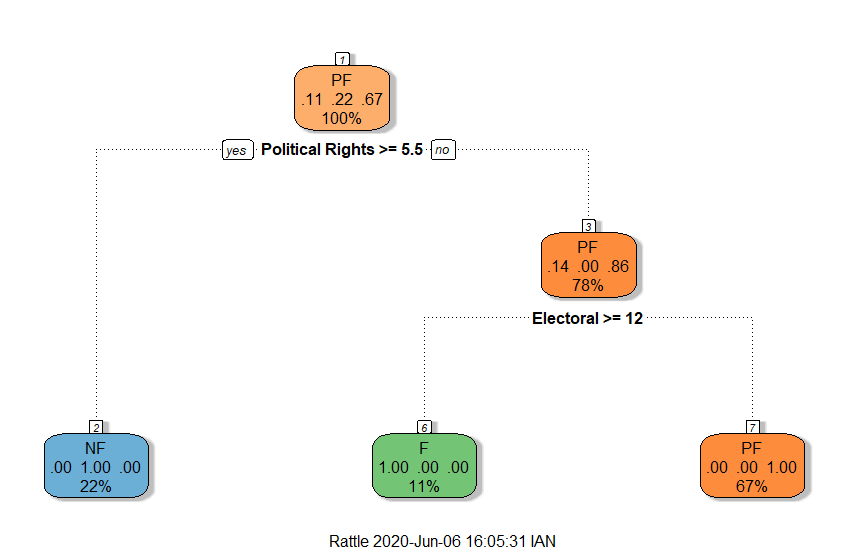
After clustering was completed, the it was desired to determine if the variables present in the UNHCR refugee data, combined with Freedom House’s freedom index data would be indicative of a country’s Freedom House “freedom” classification. Freedom House's model does not include refugee data, but if that data were added, could a decision tree model be constructed which achieved accurate results? The team desired to understand if a country’s freedom rating might suggest the presence of refugees or not. Additionally, for destination countries, would a “free” country receive more refugee applications from those who were desperate to escape a conflict-afflicted or “not free” country?



For the destination countries including Europe and selected others (USA, AUS, NZ, Japan), the model achieved a 95% accuracy rating at 39/42 accurate predictions. The majority of the countries in this subset are “free” countries, with only one “not free” country, Turkey. Turkey’s vastly different freedom indices, but similar refugee numbers may have been a confounding factor in the model. The root node for this subset was Political Rights, where a value of 1 or 2 guaranteed a “free” rating. When political rights rating fell above that threshold, the “year 2010 total refugee applications” became the next internal node. When applications for a specified country were >= 5665, that country was classified as "not free”, otherwise it was classified as “partially free.” The ”2010” split is counter-intuitive because one might think that a large number of refugee applications would occur in a free country, however, when assessed further, one finds that Turkey is a country with a higher “political rights” rating (1 is the best) that has a significant number of refugees. This model would not generalize well to another dataset because of confounding factors like this. If a country had a moderately high political rights rating with more than 5665 refugee applications in one 2010, this model would classify that country as not free, which is an overfit generalization.



For the data subset including Russia and the former Soviet Socialist Republics, a 71% accuracy level was achieved. This model is simply overfit to the training data and does not offer substantive value to the analysis. This model was trained on 7 datapoints and should have instead used the first decision tree model on those seven observations.

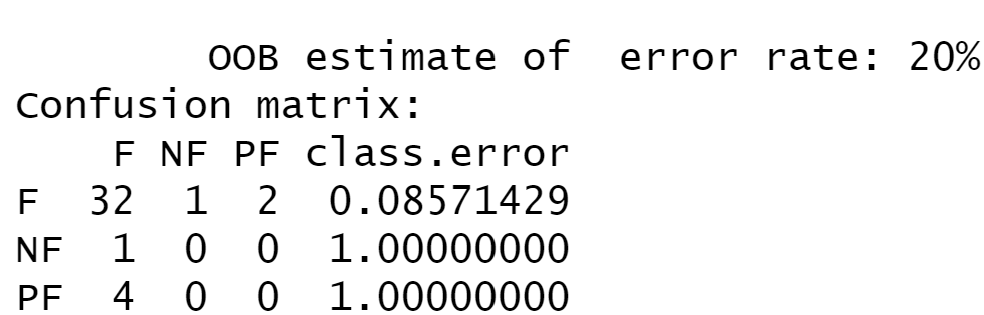


This data subset included all refugee origin countries that had refugees apply for asylum in industrialized countries. This model achieved 77% accuracy. The root node is “Political Rights >= 5.5,” considering that this is a rating from 1 to 7 (1 is best), this is a very low score. Any country that received this score was classified as “not free.” Next, the internal node of “Electoral >= 12” was considered. For values >= 12, “free” was selected, otherwise “partially free” was selected. \

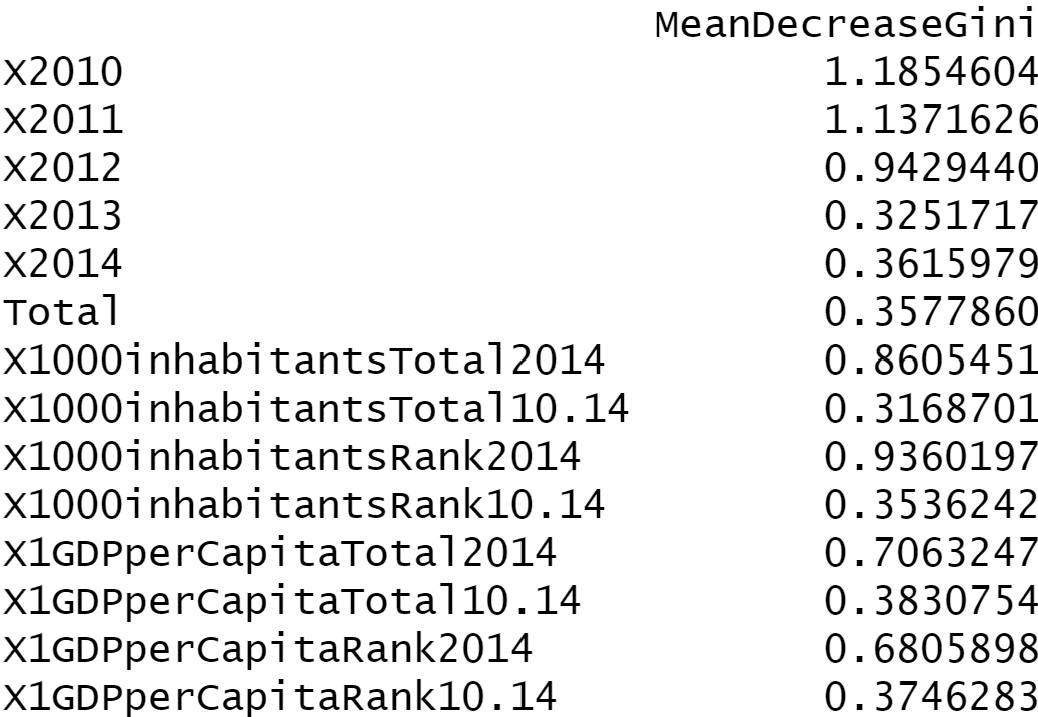
While these decision trees achieved accuracy ratings between 71% and 95%, it would have been more effective to further explore parameters as to not overfit the model to the training data. While the actual classifications were assigned by Freedom House, the attempt to achieve a correct classification by adding refugee numbers resulted in a mediocre response. This is likely due to Freedom House’s model already accounting for refugee numbers or other significant factors in their surveys which led to incorrect classifications when the team added unnecessary complexity with refugee numbers.

*Random Forest Assessment*

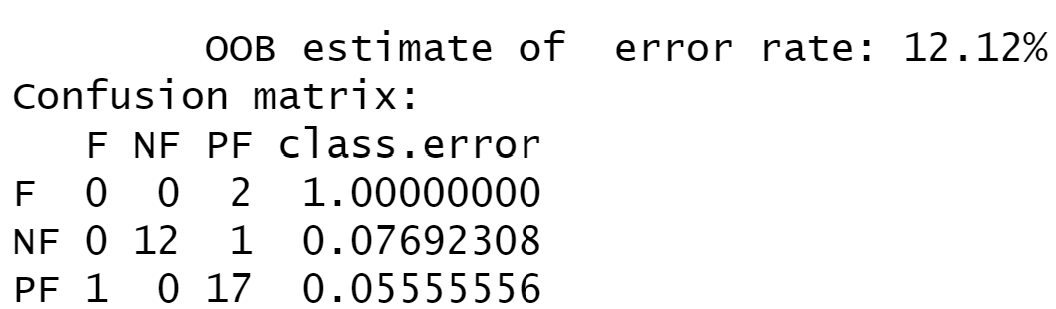
The final models that were run were Random Forest on both tables for where asylum seekers were seeking that asylum and where thy originated from. The first model below is showing the Random Forest run on the first table of where asylum seekers were seeking the asylum. It was run to predict whether a country was labeled as Free. For the Freedom Index values not to show a perfect prediction those values were removed for the model.



The error rate of 20% is promising and can be taken that defining a Free country off the first table is somewhat reliable. The importance of the variables below shows that the asylum numbers from the years showed strong importance in the model. They would recommend looking into collecting more data from these countries by territories in them.



The next model was run on the table of where the applicants were coming from. The team was more interested in running the model on this table as it could produce a better path forward when determining if a country is Free. As with the previous the Freedom Index values were taken out except for Political Rights and Civil Liberties. It was determined that these values important and good values for the UNCHR to populate.



This model shows a better prediction model than the model built on the previous data and used even less variables. The importance of the variables shows that the variables that come from the Freedom House are showing strong importance. This makes sense but the other variables are also encouraging as these are values that the UNHCR already have.



**Conclusion**

UNHCR, nongovernmental and other intergovernmental organizations have an ever-increasing challenge in responding to refugee crises. It is necessary that these organizations maximize their forecasting capabilities to identify areas of interest for further investigation and proactively respond to major humanitarian disasters. It’s important to note that as an ongoing crisis, data be continuously collected and compared across different methodologies. Descriptive analysis, while an important basis for further exploration, does not sufficiently provide cause for these organizations to mobilize resources and personnel for an expedient and efficient response. Models such as Association Rules Mining, Clustering and Decision Trees may be able to inform decision making by identifying unexpected trends in data or predicting changes in a country's likelihood for a crisis, given changes to several factors. For example, predicting changes in refugees given changes in political rights and GDP after an economic catastrophe.

Clustering and decision trees grant insights that simple summary statistics cannot. When isolating freedom indices from refugee numbers, GDP or other economic indicators, one can identify countries that share similar civil liberties and political freedom ratings. When two countries mirror each other politically, but differ geographically, or in refugee numbers, there is an opportunity to explore potential flashpoints that might lead to a refugee crisis. We identified this in democratic countries that are surrounded by relatively “not free” countries, such as Ukraine, Armenia and Moldova. Simply classifying surrounding countries as “not free” gives reason to explore the possibility of future strife.

Conversely, countries that differ from “not free” countries, such as Eritrea compared to Iraq or Afghanistan, present a false sense of security with their freedom indices. When accounting for conflict afflicted countries in their periphery, these two countries were much more similar and were even clustered together when refugee numbers were included. Another interpretation could be taken from the rules defined through the ARM model, in which the connection among the countries was more clear based on low Governance and low Organization scores more so than the other indices.

Decision trees can provide similar insights; identifying patterns in countries that experienced refugee crises and training a model to predict that outcome in other countries is an invaluable forecasting tool. Just as the team trained a model to predict a freedom classification based on available data, so too can you train a model to predict refugee application numbers given user-adjusted parameters. This model can be seen also when multiplied within Random Forest to show the power of decision tree modeling and the importance of the variables within in the data set.

Moving forward, this type of analysis may be better suited for country-specific deep dives. Within a country, freedom indices tend to vary from province to province or city to city, especially if there is ethnic violence, disenfranchised populations, limited or full-scale war. One province could be “partially free” and unaffected while another is devastated by horrendous circumstances that drive massive instability, violence and displaced persons. Countries where this is true include South Sudan, Myanmar and the Rohingya and The People’s Republic of China with the Uighurs. Additionally, while we would like to think that refugees could apply and be granted asylum in Europe or the US, it isn’t feasible for them to actually reach those destinations, which is why we see excessive refugees application numbers in countries like Turkey, Russia, Italy – all coming from conflict afflicted areas.