

# Data Processing

## Contents

|  |    |
|--|----|
| Overview . . . . .                               | 1  |
| Redefine Factor Variables . . . . .              | 1  |
| Remove Missing Observations . . . . .            | 3  |
| Generate Factor Blocs . . . . .                  | 8  |
| Extract Numeric & Categorical Features . . . . . | 8  |
| Dummy Encoding Categorical Variables . . . . .   | 9  |
| Standardise Numeric Data . . . . .               | 9  |
| Data Reduction . . . . .                         | 10 |
| Data Output . . . . .                            | 11 |

## Overview

This Rmarkdown report processes and cleans the Eurovision Song Contest (ESC) data for modeling.

This includes:

1. Redefine variables as factor or numeric
2. Dividing the variables into the three predefined groups;
  - Performance
  - External
  - Competition
3. Normalize the Numeric Data to have mean 0 and standard deviation 1
4. Dummy Encoding all Categorical Factor levels.
5. Data Reduction
  - Redundant Variables
  - Variables of Linear Combinations
  - Categorical Variables via Chi-Squared Tests of Association

## Redefine Factor Variables

Some of the numeric music features need to be redefined as nominal variables, the variables are key, mode and time signature.

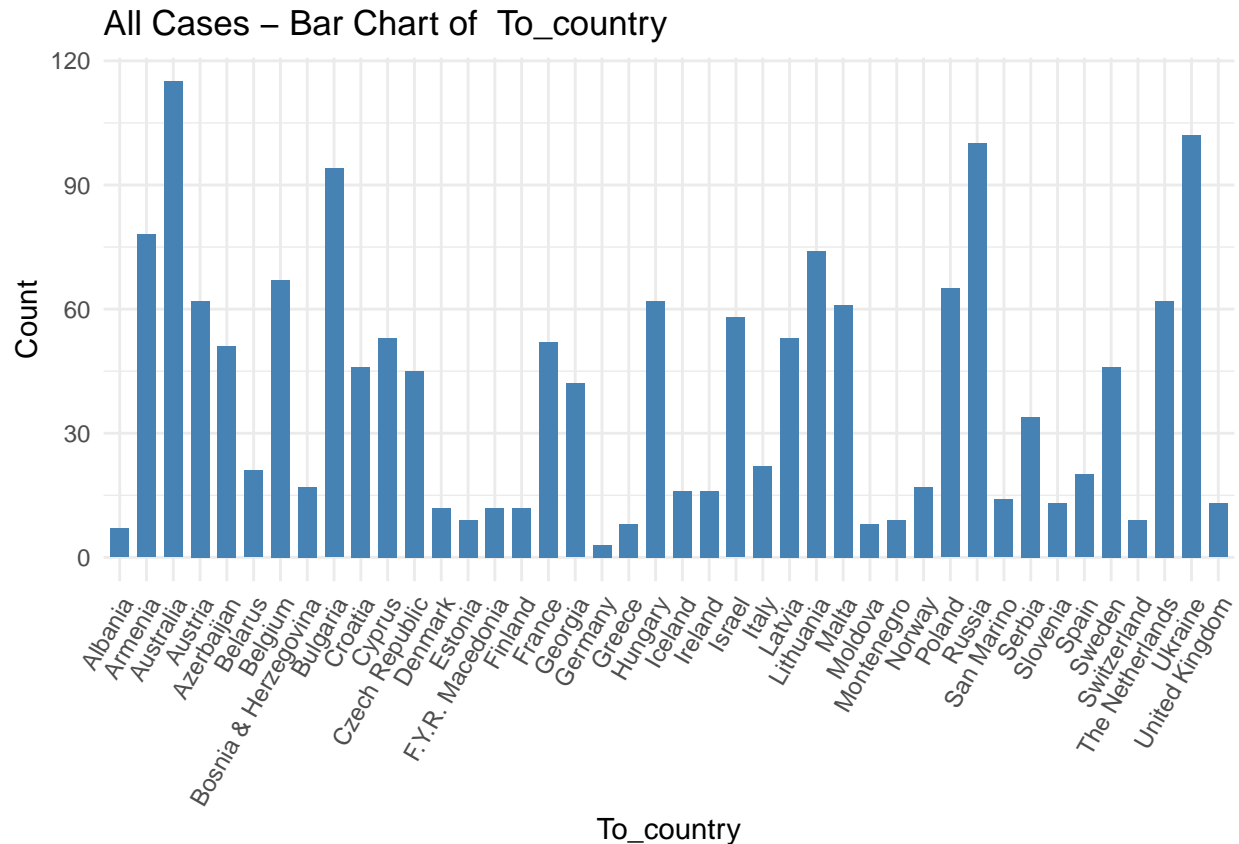
|              | Data Type |
|--------------|-----------|
| From_country | character |

|                  | Data Type |
|------------------|-----------|
| To_country       | character |
| Points           | integer   |
| Round            | character |
| Voting_Method    | character |
| Host_Nation      | character |
| OOA              | integer   |
| Average_Points   | numeric   |
| VBlocs1_FC       | factor    |
| VBlocs2_FC       | factor    |
| VBlocs1_TC       | factor    |
| VBlocs2_TC       | factor    |
| ComVBlocs1       | character |
| ComVBlocs2       | character |
| FC_LANGFAM       | character |
| TC_LANGFAM       | character |
| ComLANGFAM       | character |
| Neighbours       | character |
| TC_NumNeigh      | integer   |
| FC_NonCOB        | integer   |
| FC_NonCitizens   | integer   |
| FC_COB           | integer   |
| FC_Citizens      | integer   |
| FC_Population    | integer   |
| METRIC_COB       | numeric   |
| METRIC_Citizens  | numeric   |
| METRIC_COBCit    | numeric   |
| FC_GDP_mil       | numeric   |
| TC_GDP_mil       | numeric   |
| GDP_PROP         | numeric   |
| FC_CAP_LAT       | numeric   |
| FC_CAP_LON       | numeric   |
| TC_CAP_LAT       | numeric   |
| TC_CAP_LON       | numeric   |
| CAP_DIST_km      | numeric   |
| TC_PerfType      | character |
| TC_SingerGender  | character |
| FC_SONGLANG      | character |
| TC_SONGLANG      | character |
| ComSONGLAN       | integer   |
| danceability     | numeric   |
| energy           | numeric   |
| key              | factor    |
| loudness         | numeric   |
| mode             | factor    |
| speechiness      | numeric   |
| acousticness     | numeric   |
| instrumentalness | numeric   |
| liveness         | numeric   |
| valence          | numeric   |
| tempo            | numeric   |
| duration_ms      | integer   |
| time_signature   | factor    |

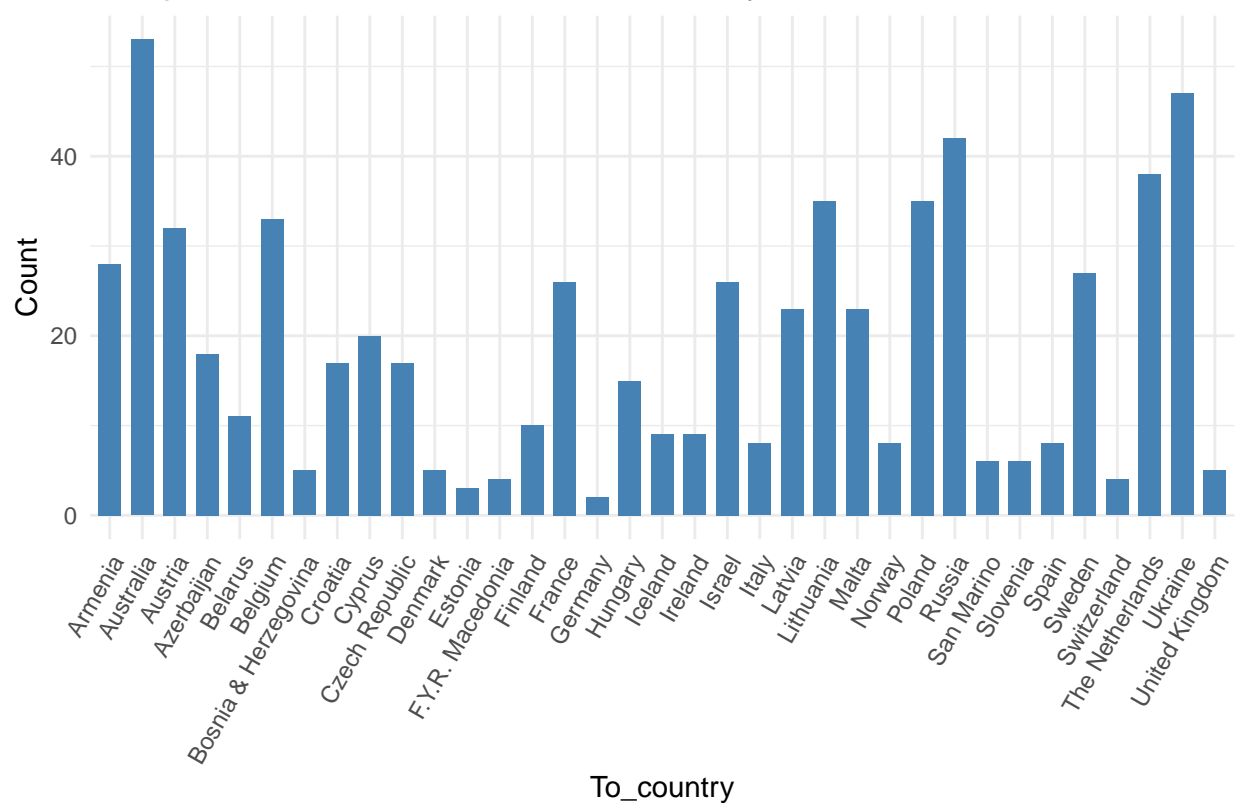
## Remove Missing Observations

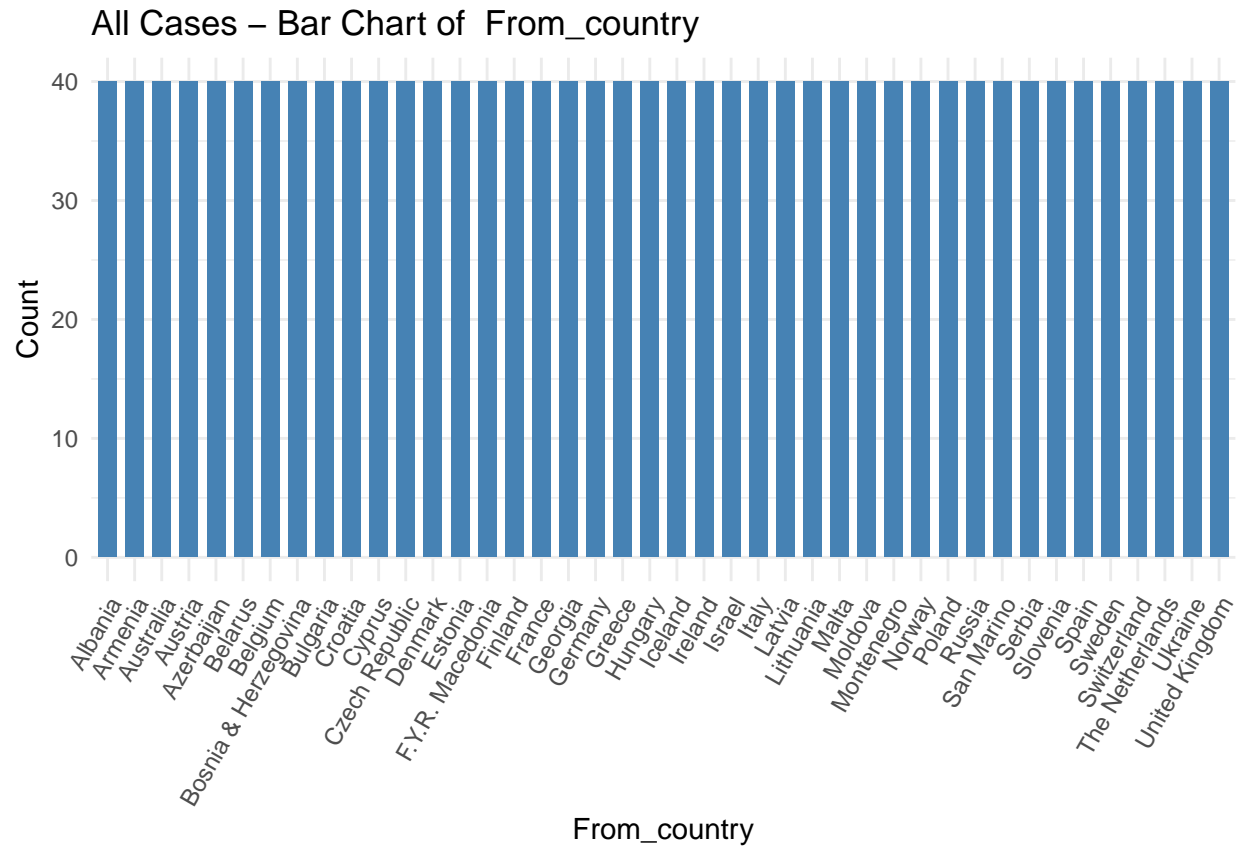
There is a total of 1022 rows missing from the data. This corresponds to approximately 60.83% of the data. Removing all of these rows with missing values leaves a total of 658 rows remaining. This is a substantial loss of data, a notable limitation to the research, and an area of improvement for future research iterations. An alternative solution could be to use different sources such as the world bank for the migration / diaspora based data. However at time of research, these data repositories did not store migration / diaspora information by country of birth / citizenship.

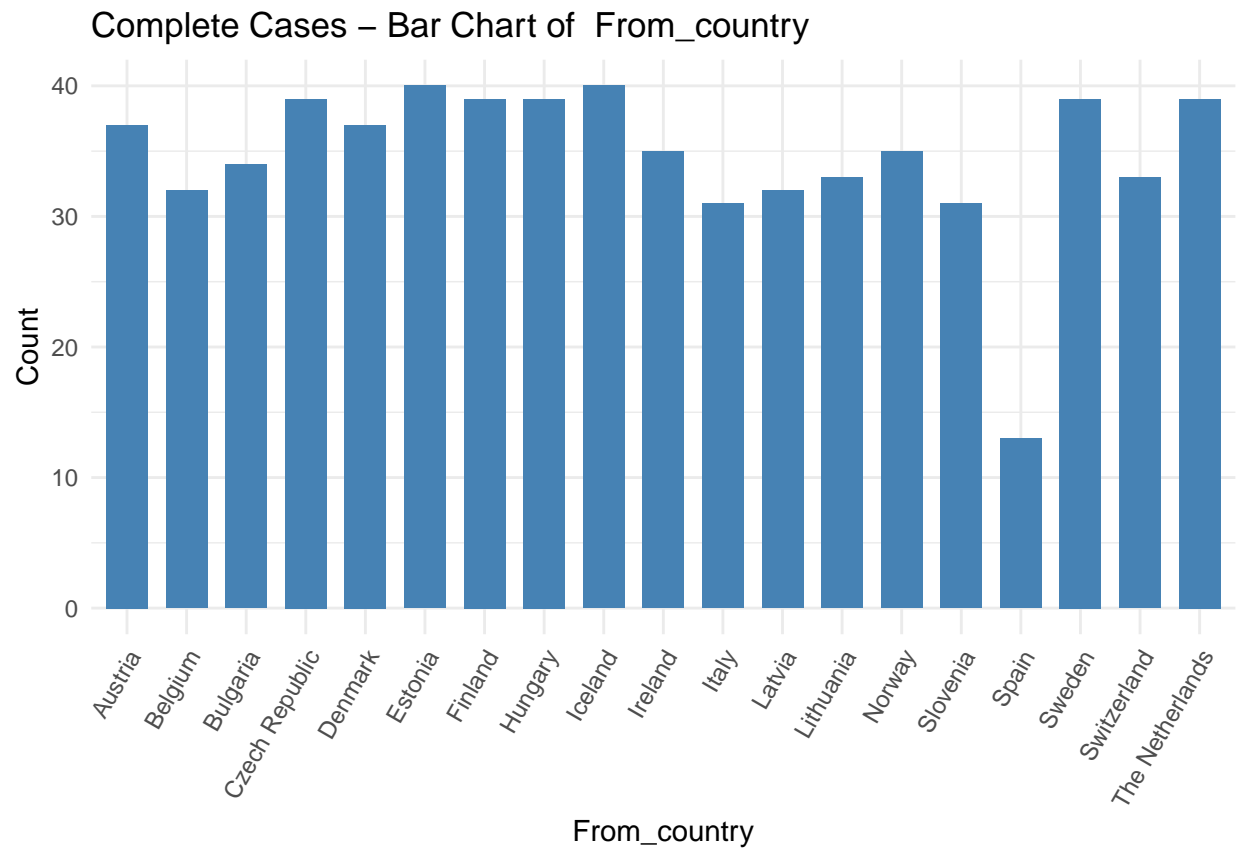
<http://www.worldbank.org/en/topic/migrationremittancesdiasporaissues/brief/migration-remittances-data>

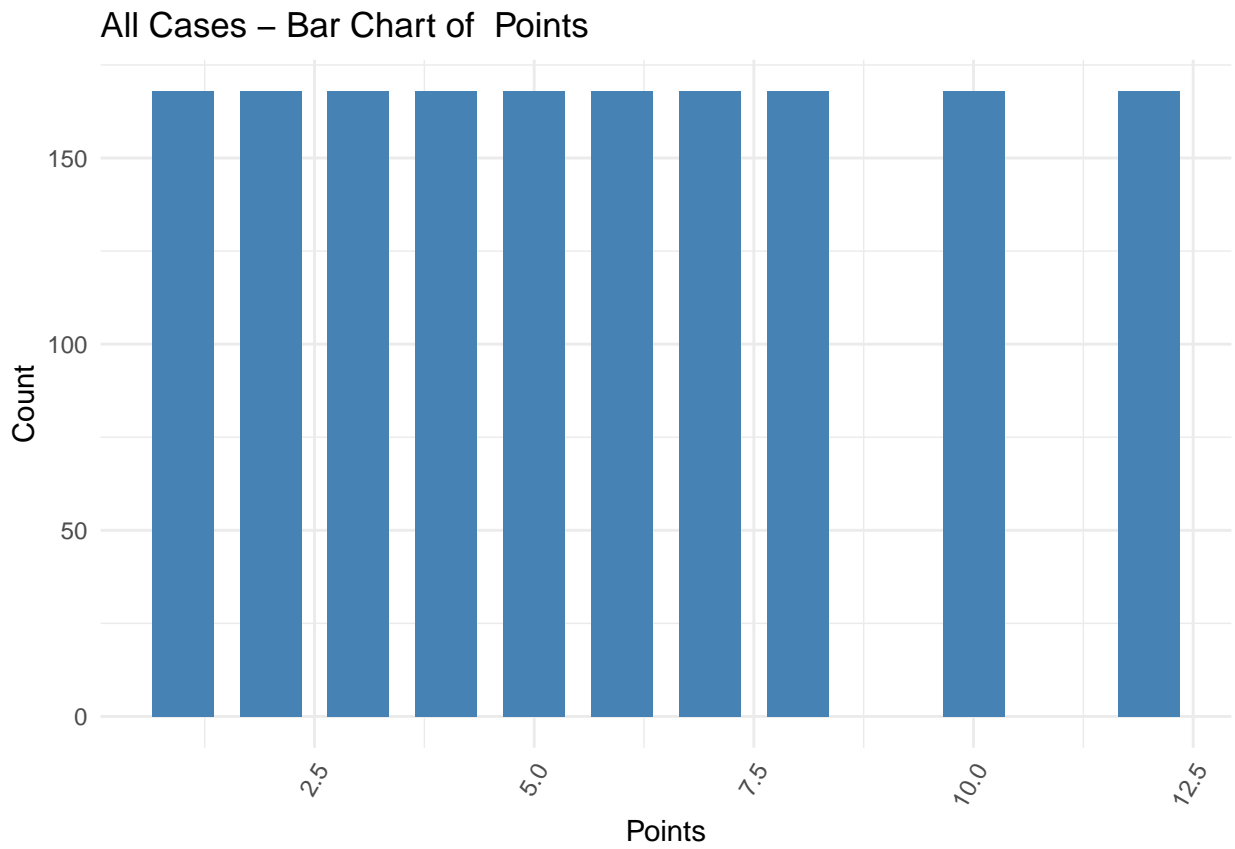


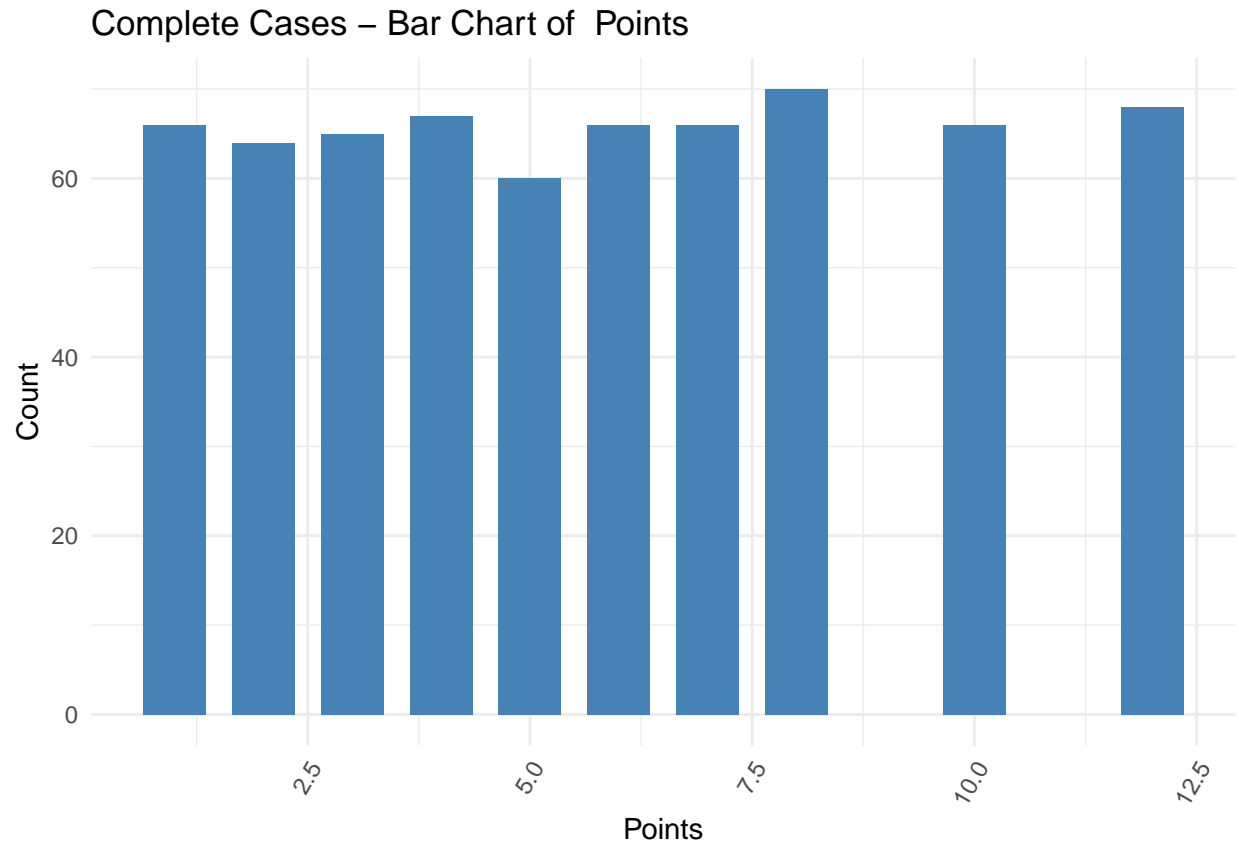
Complete Cases – Bar Chart of To\_country











## Generate Factor Blocs

In total, there are:

- 3 Voting Factors
- 5 Competition Factors
- 27 External Factors
- 18 Performance Factors

## Extract Numeric & Categorical Features

In total, there are:

- 2 Numeric Voting Factors
- 3 Categorical Voting Factors
- 17 Numeric Competition Factors
- 10 Categorical Competition Factors
- 11 Numeric External Factors
- 7 Categorical External Factors
- 1 Numeric Performance Factors
- 2 Categorical Performance Factors



## Dummy Encoding Categorical Variables

All categorical variables for each variable bloc are dummy encoded. It is not necessary to dummy encode the voting factors To\_country and From\_country as we will not be incorporated as model predictor variables

## Standardise Numeric Data

This section normalizes or range standardizes the numeric variables in each block. There is also a special case for OOA in the competition Bloc. OOA will need to be standardized in relation to each round. This shall be done after the numeric variables from each bloc have been standardized

|                | mean | sdev |
|----------------|------|------|
| Average_Points | 0    | 1    |
| OOA            | 0    | 1    |

|                 | mean | sdev |
|-----------------|------|------|
| TC_NumNeigh     | 0    | 1    |
| FC_NonCOB       | 0    | 1    |
| FC_NonCitizens  | 0    | 1    |
| FC_COB          | 0    | 1    |
| FC_Citizens     | 0    | 1    |
| FC_Population   | 0    | 1    |
| METRIC_COB      | 0    | 1    |
| METRIC_Citizens | 0    | 1    |
| METRIC_COBCit   | 0    | 1    |
| FC_GDP_mil      | 0    | 1    |
| TC_GDP_mil      | 0    | 1    |
| GDP_PROP        | 0    | 1    |
| FC_CAP_LAT      | 0    | 1    |
| FC_CAP_LON      | 0    | 1    |
| TC_CAP_LAT      | 0    | 1    |
| TC_CAP_LON      | 0    | 1    |
| CAP_DIST_km     | 0    | 1    |

|                  | mean | sdev |
|------------------|------|------|
| ComSONGLAN       | 0    | 1    |
| danceability     | 0    | 1    |
| energy           | 0    | 1    |
| loudness         | 0    | 1    |
| speechiness      | 0    | 1    |
| acousticness     | 0    | 1    |
| instrumentalness | 0    | 1    |
| liveness         | 0    | 1    |
| valence          | 0    | 1    |
| tempo            | 0    | 1    |
| duration_ms      | 0    | 1    |

| From_country   | To_country | Round | OOA |
|----------------|------------|-------|-----|
| Austria        | Belgium    | f     | 1   |
| Bulgaria       | Belgium    | f     | 1   |
| Czech Republic | Belgium    | f     | 1   |
| Denmark        | Belgium    | f     | 1   |
| Denmark        | Belgium    | f     | 1   |
| Iceland        | Belgium    | f     | 1   |

| From_country   | To_country | Round | OOA |
|----------------|------------|-------|-----|
| Austria        | Belgium    | f     | 0   |
| Bulgaria       | Belgium    | f     | 0   |
| Czech Republic | Belgium    | f     | 0   |
| Denmark        | Belgium    | f     | 0   |
| Denmark        | Belgium    | f     | 0   |
| Iceland        | Belgium    | f     | 0   |

| From_country | To_country | Round | OOA       |
|--------------|------------|-------|-----------|
| Austria      | Armenia    | f     | 1.0000000 |
| Austria      | Armenia    | sf1   | 0.3529412 |
| Austria      | Armenia    | sf1   | 0.3529412 |
| Belgium      | Armenia    | f     | 1.0000000 |
| Bulgaria     | Armenia    | f     | 1.0000000 |
| Bulgaria     | Armenia    | f     | 1.0000000 |

|    | From_country | To_country | Round | OOA |
|----|--------------|------------|-------|-----|
| 9  | Austria      | Armenia    | f     | 26  |
| 10 | Austria      | Armenia    | sf1   | 7   |
| 11 | Austria      | Armenia    | sf1   | 7   |
| 14 | Belgium      | Armenia    | f     | 26  |
| 18 | Bulgaria     | Armenia    | f     | 26  |
| 19 | Bulgaria     | Armenia    | f     | 26  |

|    | Average_Points | OOA        |
|----|----------------|------------|
| 9  | 1.339881       | 2.2137693  |
| 10 | 1.339881       | -0.6879442 |
| 11 | 1.339881       | -0.6879442 |
| 14 | 1.470934       | 2.2137693  |
| 18 | 1.034092       | 2.2137693  |
| 19 | 1.034092       | 2.2137693  |

## Data Reduction

This section performs data reduction whereby Categorical / Numeric variables are removed if: 1. They have only a single type of observation, 0 or 1 2. They form part of a linear combination with other variables 3. They are strongly associated / correlated with other variables

In total, there are:

- 13 Redundant External Categorical Factors
- 0 Redundant Competition Factors
- 0 Redundant Performance Factors

If two variables are feature a lot of 0s or 1s then there will be a strong association between the two variables as they share a lot of common observations. In such cases, one of the variables can be removed as they both measure the same entity. This lowers the chance of collinearity and reduces the number of dimensions

If two numeric variables are very highly correlated and represent the same entity then it is unnecessary to include them in the data modeling stage. This is particular the case for the migration data. The correlation test function implement is the same as the one that was used during the exploratory analysis section.

## **Data Output**