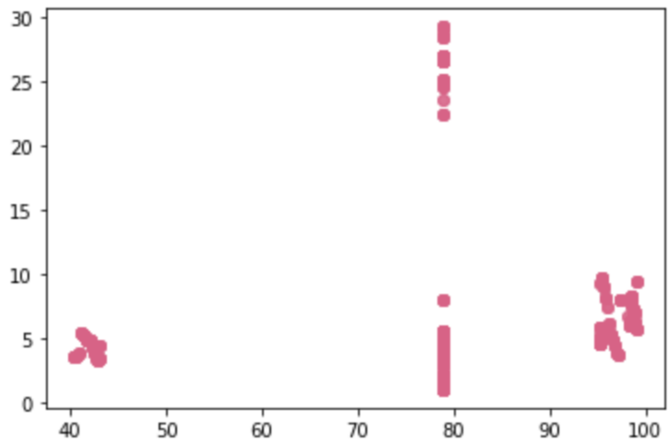
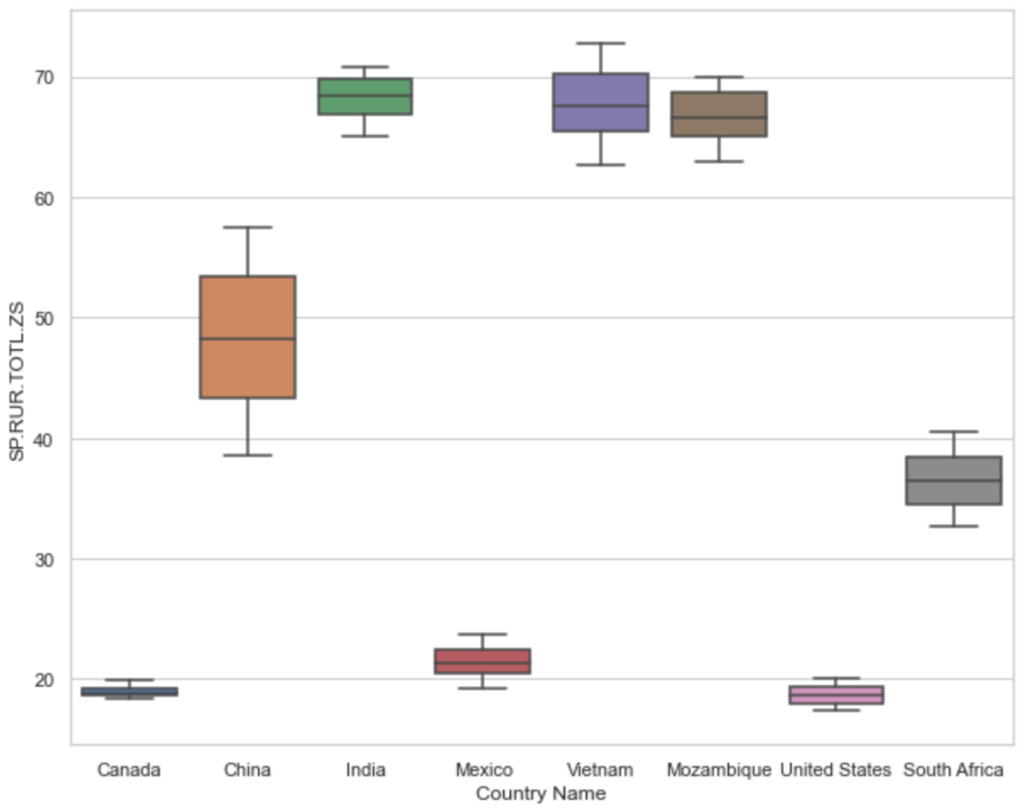
1. Data summarization, data preprocessing and feature selections

1.1 Data summarization

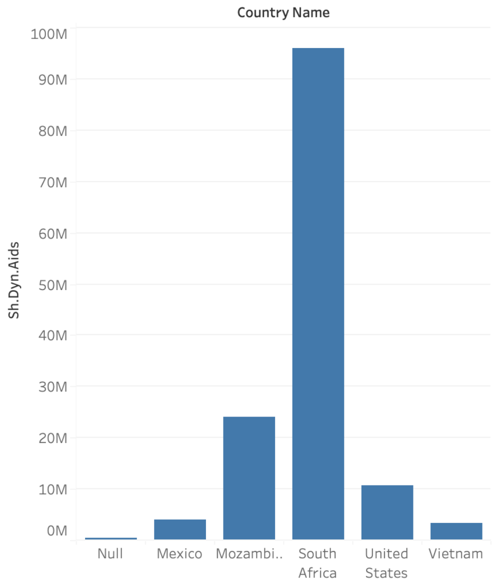
1.1.1 Scatter plot

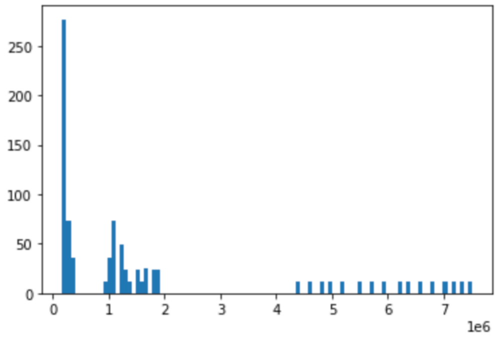


1.1.2 Boxplots



1.1.3 Histograms





According to these three types of graphs, we know that the overall data is unbalanced, and there is an unequal distribution of class in it. Also, the dataset does not follow a Gaussian distribution, so the normalization step following is important.

1.2 Data processing

1.2.1 Missing values

- Missing data for consecutive years
  - When the data are missing for consecutive years, for example, data are missing from 2005 to 2010, we consider it not helpful to replace missing data with a median value, so we replace them with 0.
- Missing data for a couple of years
  - When the data are missing for a couple of years, for example, data are missing in 2008 and 2015, we decided to replace them with a median value in a specific range.

1.2.2 Categorical attributes

We use one-hot encoding to deal with categorical attributes.

When a category has several levels, assigning numbers to each level implies an order of the levels. This means that one level of the category has a lower rank than another level. So we use 1 to 5 for levels of adults living with HIV. 1 stands for the smallest population, and 5 stands for the largest population. While this makes sense for ordinal variables, it is a wrong assumption for nominal variables such as nationality. Therefore, we decided not to use one-hot encoding in countries.

1.2.3 Normalization

From the histogram above, we know that our data does not follow a Gaussian distribution. Therefore, we use normalization to shift and rescale the dataset so that they end up ranging between 0 and 1 via MinMaxScalar from the sklearn library.

There is a subset of the original dataset:

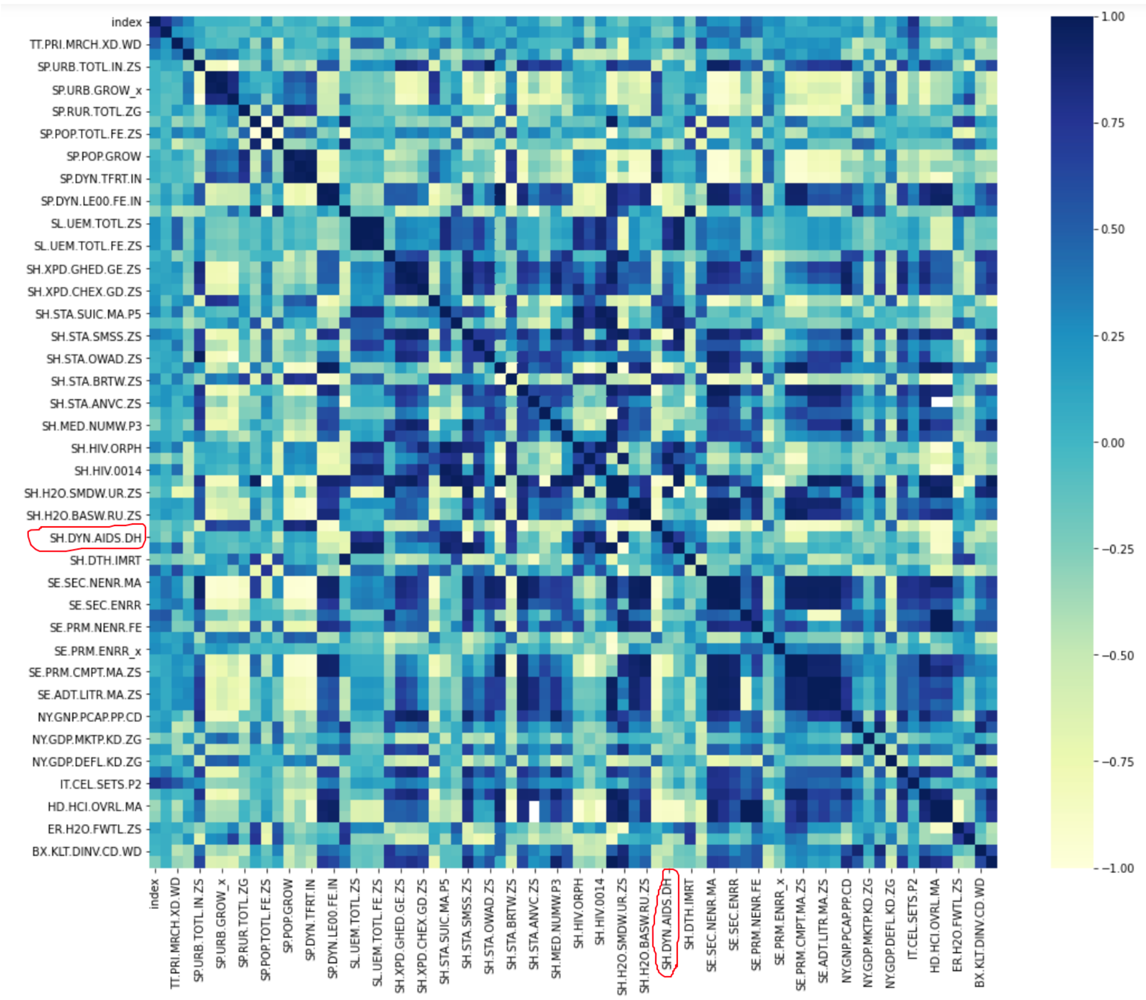
	index	Year	TT.PRI.MRCH.XD.WD	TG.VAL.TOTL.GD.ZS	SP.URB.TOTL.IN.ZS
0	0	2005	100	70	80.122
1	1	2005	100	70	80.122
2	2	2005	100	70	80.122
3	3	2005	100	70	80.122
4	4	2005	100	70	80.122
...	...	...	...	...	...
1540	1540	2020	165	51	67.354
1541	1541	2020	165	51	67.354
1542	1542	2020	165	51	67.354
1543	1543	2020	165	51	67.354
1544	1544	2020	165	51	67.354

	SP.URB.GROW_y	SP.URB.GROW_x	SP.RUR.TOTL.ZS	SP.RUR.TOTL.ZG	\
0	1.040619	1.4	19.878	0.557844	
1	1.040619	1.4	19.878	0.557844	
2	1.040619	1.4	19.878	0.557844	
3	1.040619	1.4	19.878	0.557844	
4	1.040619	1.4	19.878	0.557844	
...	...	...	...	...	
1540	2.015480	2.0	32.646	-0.240580	
1541	2.015480	2.0	32.646	-0.240580	
1542	2.015480	2.0	32.646	-0.240580	
1543	2.015480	2.0	32.646	-0.240580	
1544	2.015480	2.0	32.646	-0.240580	

There is a subset after normalization:

```
[[0.54992658 0.8        0.28       ... 0.18576143 0.         0.        ]
 [0.89427313 0.73333333 0.68       ... 0.01896006 0.         1.        ]
 [0.02936858 0.2        0.13333333 ... 0.1906109  0.         0.        ]
 ...
 [0.58443465 0.4        0.18666667 ... 0.09156732 0.         0.        ]
 [0.60058737 0.53333333 0.18666667 ... 0.09156732 0.         0.        ]
 [0.84581498 0.33333333 0.68       ... 0.01896006 0.         0.75      ]]
```

1.3 Feature selection



	index	Year	TT.PRI.MRCH.XD.WD	TG.VAL.TOTL.GD.ZS	SP.URB.TOTL.IN.ZS	SP.URB.GROW_y	SP.URB.GROW_x	SP.RUR.TOTL.ZS	SP.RUR.TOTL.ZG
index	1.000000	0.760606	0.487843	0.293436	0.007641	0.012231	0.013680	-0.007641	-0.01456
Year	0.760606	1.000000	0.211410	0.163910	0.110473	-0.090001	-0.047249	-0.110473	-0.10542
TT.PRI.MRCH.XD.WD	0.487843	0.211410	1.000000	0.398295	0.074240	-0.027345	-0.078836	-0.074240	0.18668
TG.VAL.TOTL.GD.ZS	0.293436	0.163910	0.398295	1.000000	-0.333070	0.347214	0.384973	0.333070	0.01852
SP.URB.TOTL.IN.ZS	0.007641	0.110473	0.074240	-0.333070	1.000000	-0.839113	-0.828789	-1.000000	-0.28205
SP.URB.GROW_y	0.012231	-0.090001	-0.027345	0.347214	-0.839113	1.000000	0.974099	0.839113	0.25575
SP.URB.GROW_x	0.013680	-0.047249	-0.078836	0.384973	-0.828789	0.974099	1.000000	0.828789	0.14305
SP.RUR.TOTL.ZS	-0.007641	-0.110473	-0.074240	0.333070	-1.000000	0.839113	0.828789	1.000000	0.28205
SP.RUR.TOTL.ZG	-0.014565	-0.105423	0.186689	0.018520	-0.282051	0.255757	0.143051	0.282051	1.00000
SP.POP.TOTL.MA.ZS	-0.185271	0.027242	-0.350460	-0.178211	-0.320328	-0.001605	0.065398	0.320328	-0.47140
SP.POP.TOTL.FE.ZS	0.185271	-0.027242	0.350460	0.178211	0.320328	0.001605	-0.065398	-0.320328	0.47140
SP.POP.TOTL	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

We mapped the population of adults living with HIV into {Very small = 1, Small= 2, Medium = 3, Large = 4, Very large= 5}. We want to use the following three models to classify the level of the population of adults living with HIV.

Our dataset contains features of Education, Health, Quality of life, Population, Country etc. Because the dataset contains lots of features, picking some of them as relevant features for classification is critical. First, we see the whole picture of correlation in general via the heatmap. We know that around 10 features are relatively dark, which means the correlation with SH.DYN.AIDS is higher than others. Then, in order to dig into the exact correlation among features, we use the correlation matrix that shows the values of correlation so that we can pick the higher ones.

```
SH.H2O.SMDW.ZS        0.991027
SH.STA.SMSS.ZS        0.975533
SH.HIV.0014           0.962035
SL.UEM.TOTL.MA.ZS     0.956075
SL.UEM.TOTL.ZS        0.948709
SH.HIV.ORPH           0.935999
SL.UEM.TOTL.FE.ZS     0.933612
SH.STA.SUIC.MA.P5     0.919786
SH.STA.OWGH.ME.ZS     0.884812
SH.STA.SUIC.FE.P5     0.853521
SH.DYN.AIDS.DH        0.734165
```

Here are the top 11 correlations with SH.DYN.AIDS. We decided to choose the top 10 values since the correlation of the last one dropped from 0.85 to 0.73, and the difference between the 10th and the 11th is larger than the difference of others.

The ten features are

- People using safely managed to drink water services (% of the population)
- People using safely managed sanitation services (% of the population)
- Children (0-14) living with HIV
- Unemployment, male (% of the male labor force)
- Unemployment, total
- Children orphaned by HIV/AIDS
- Unemployment, female (% of the male labor force)
- Suicide mortality rate, male (per 100,000 male population)
- Prevalence of overweight (modeled estimate, % of children under 5)
- Suicide mortality rate, female (per 100,000 female population)

2. Classification (Supervised Learning)

2.1 Decision Tree

Round	criterion	max_depth	Accuracy	Recall	Precision
1	/	2	0.8490196078431372	0.38493723849372385	0.7695085425983258

2	entropy	3	0.9655058523529412	0.901058823529912	0.954701882348395
3	entropy	2	0.8647058823529412	0.735678391959799	0.8818861522676595

criterion="entropy", max_depth=2

```
--------Decision Tree--------
Time to construct the model:  0.004157066345214844
Accuracy: 0.8647058823529412
Recall:  0.735678391959799
Precision:  0.8818861522676595
```

2.2 Gradient Boosting

Round	iterations	learning_rate	max_depth	Accurancy	Recall	Precision
1	100	0.1	1	0.9654736981342076	0.9768816811323645	0.9911698113200765
2	200	0.005	1	0.9490196078431372	0.9754716981132076	0.9540214576396597
3	100	0.005	1	0.8196078431372549	0.7169283919597994	0.81925708699902244

```
--------Gradient Boosting--------
Time to construct the model:  0.7605419158935547
Accuracy: 0.8196078431372549
Recall:  0.716928391959799
Precision:  0.8192570869990224
```

2.3 Random Forest

Round	max_depth	Accurancy	Recall	Precision
1	1	0.7588235294117647	0.339622641509434	0.7250839562344275
2	3	0.9921568627450981	0.8	0.9856209150326798
3	2	0.9	0.7170489949748744	0.9047319075723589

```
--------Random Forest--------
Time to construct the model:  0.14213919639587402
Accuracy: 0.9
Recall:  0.7170489949748744
Precision:  0.9047319075723589
```

2.4 Comparison

Model	Accuracy	Recall	Precision	Time
Decision tree	0.8666666666666667	0.739622641509434	0.8861341337753684	0.004441976547241211
Gradient Boosting	0.8176470588235294	0.6881940700808625	0.8295121663123081	0.7514901161193848
Random Forest	0.8509803921568627	0.5650943396226416	0.8299535268652916	0.12217378616333008

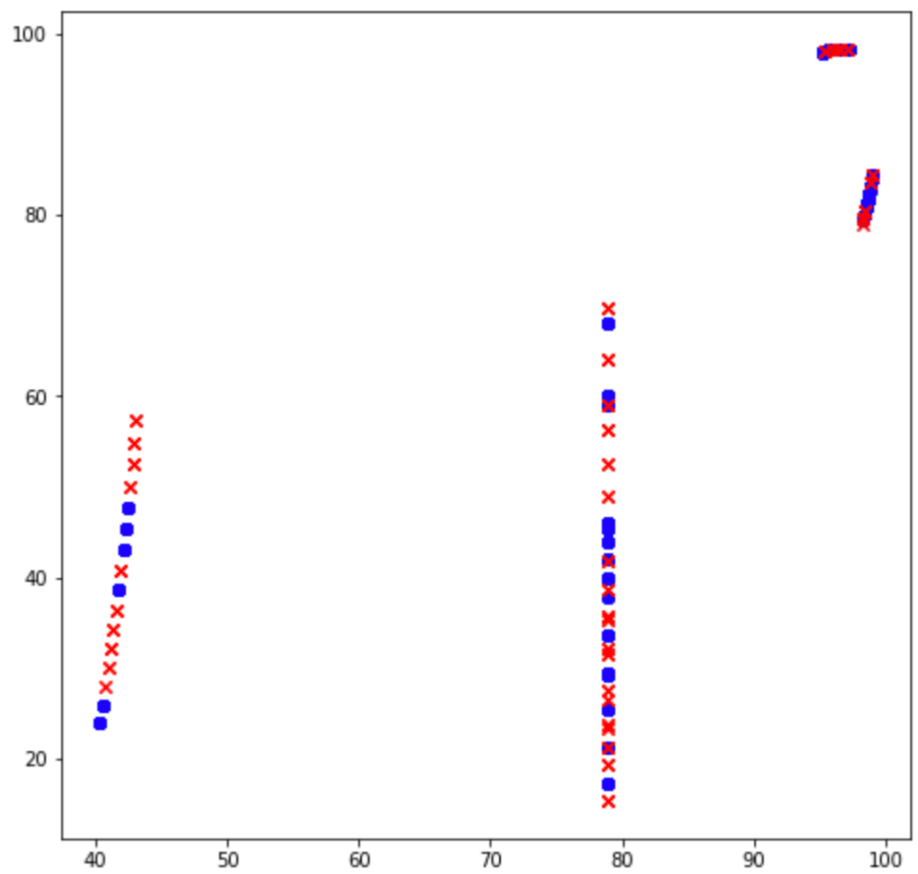
Overall, the decision tree model did the best!

2.5 Insights we had got

- Feature selection helps solve two problems: having too much data that is of little value or having too little data that is of high value. The decision tree algorithm can naturally select which features are most important. Based on our test, the accuracy, precision and recall are higher on the whole dataset instead of the subset of feature selection. In other words, If we somehow know which features are the most important, then DT should be able to acquire accuracy while saving computing power.

- A decision tree can handle both numerical and categorical variables at the same time as features. Because our dataset contains a large amount of numerical and categorical data, DT is an excellent choice.
- Random Forest is suitable for situations when we have a large dataset.
- Random forests typically outperform gradient boosting in high noise settings (especially with small data). But in our dataset, we have small data and low noise, so gradient boosting outperforms random forests in this case.

3. Detecting Outliers



We identify the outliers based on the level of population of adults living with HIV (Very small, Small, Medium, …). The one-class SVM assumes all data belong to the normal class(Very small, Small, Medium, …), and detect novelties outside the boundaries.