# Predicting population growth

Mons Even Oppedal

# Abstract

This project looks at the population growth from the world indicators dataset. In the project I try to fit a regression model using some relevant (and irrelevant?) indicators.

# Motivation

The development of demographics are very relevant for most countries when deciding policy.

I would like to find out if certain development indicators can predict a country’s rate of population growth across several countries.

# Dataset(s)

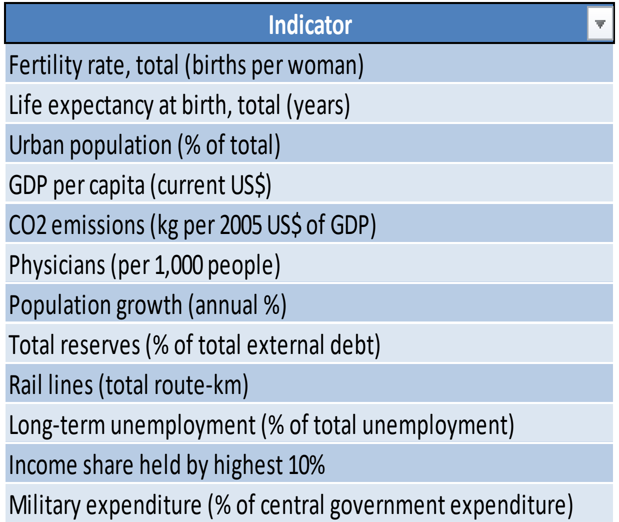
I used the world development indicators dataset that we already explored.

# Data Preparation and Cleaning

The first issue was to choose relevant indicators. The dataset contains more than 1000 indicators.

I needed an easy way to scan all the indicators to find some relevant and perhaps also some irrelevant indicators.

I began with these:



| Indicator |
|---|
| Fertility rate, total (births per woman) |
| Life expectancy at birth, total (years) |
| Urban population (% of total) |
| GDP per capita (current US$) |
| CO2 emissions (kg per 2005 US$ of GDP) |
| Physicians (per 1,000 people) |
| Population growth (annual %) |
| Total reserves (% of total external debt) |
| Rail lines (total route-km) |
| Long-term unemployment (% of total unemployment) |
| Income share held by highest 10% |
| Military expenditure (% of central government expenditure) |

I did not expect that they all would be relevant

# Research Question(s)

How well will a model predict the population growth of a country by using the indicators that I’ve chosen?

# Methods

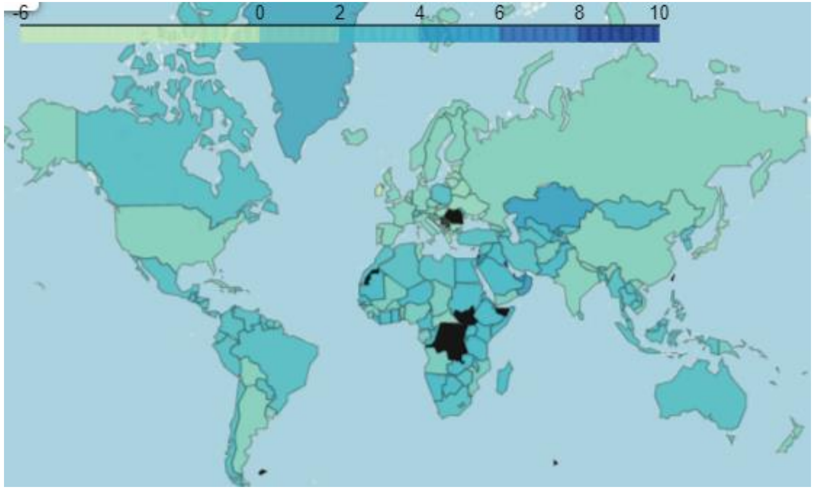
What methods did you use to analyze the data and why are they appropriate? Be sure to adequately, but briefly, describe your methods.

I used a visual display to show how population growth differs in the world.

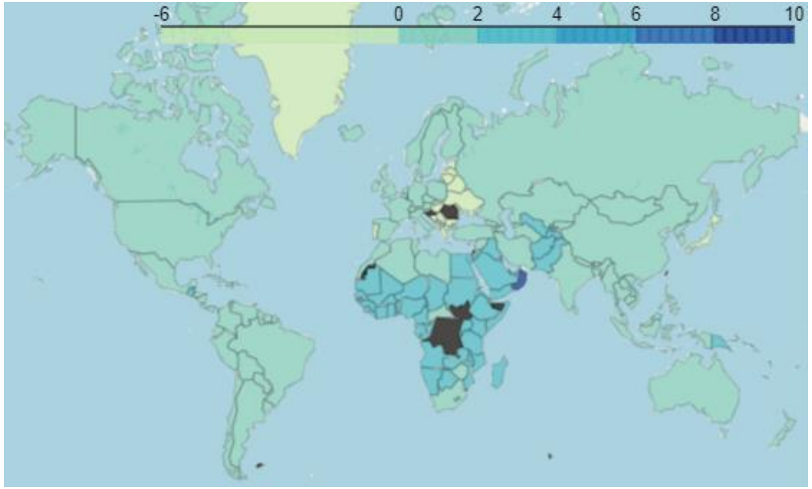
Then i used a linear regression model to train and test the models.

# Findings Visual display of growth:

1960 Scale annual growth %. Mean is 2,3 %



2011 Scale annual growth %. Mean is 1,3 %



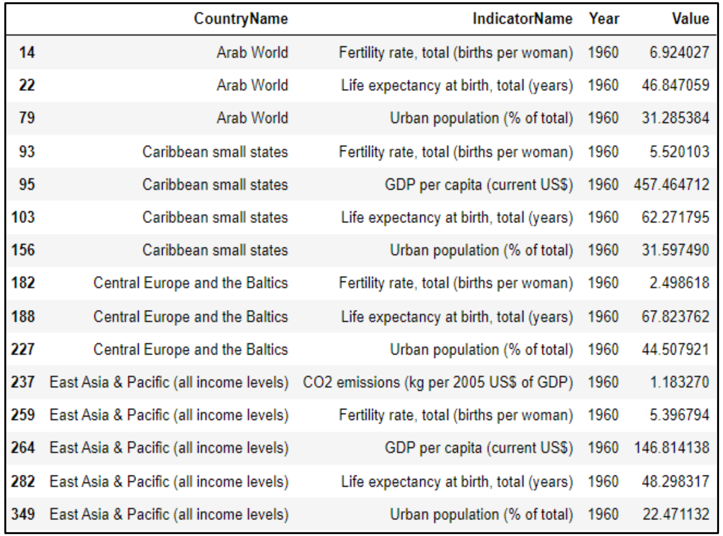
We can easily see that high growth is centered in Africa and the Middle-East, and that the population growth on average is lower in 2011 than in 1960

# Findings

I am trying to make a model that can predict population growth. Population growth is a number so I will use a regression model.

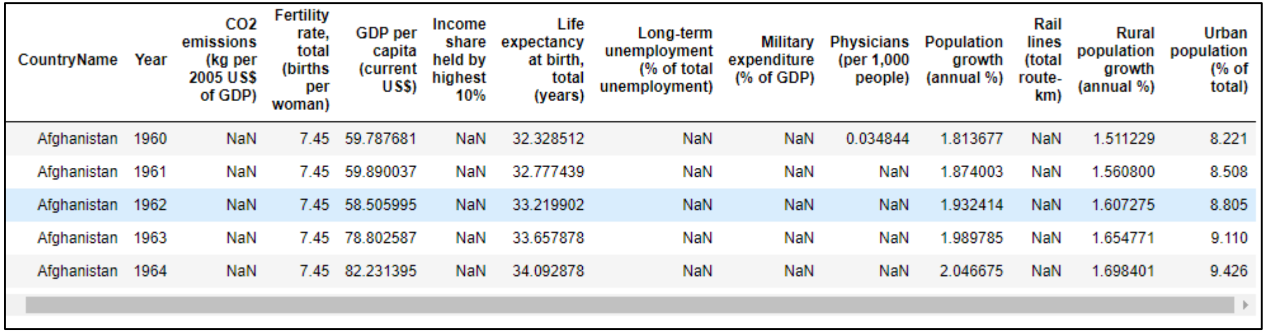
In order to use the model from the course I needed to transform the data from a «tall» to a «flat» dataset.

From a «tall» dataset where the indicators are stacked like this:



| | CountryName | IndicatorName | Year | Value |
|---|---|---|---|---|
| **14** | Arab World | Fertility rate, total (births per woman) | 1960 | 6.924027 |
| **22** | Arab World | Life expectancy at birth, total (years) | 1960 | 46.847059 |
| **79** | Arab World | Urban population (% of total) | 1960 | 31.285384 |
| **93** | Caribbean small states | Fertility rate, total (births per woman) | 1960 | 5.520103 |
| **95** | Caribbean small states | GDP per capita (current US$) | 1960 | 457.464712 |
| **103** | Caribbean small states | Life expectancy at birth, total (years) | 1960 | 62.271795 |
| **156** | Caribbean small states | Urban population (% of total) | 1960 | 31.597490 |
| **182** | Central Europe and the Baltics | Fertility rate, total (births per woman) | 1960 | 2.498618 |
| **188** | Central Europe and the Baltics | Life expectancy at birth, total (years) | 1960 | 67.823762 |
| **227** | Central Europe and the Baltics | Urban population (% of total) | 1960 | 44.507921 |
| **237** | East Asia & Pacific (all income levels) | CO2 emissions (kg per 2005 US$ of GDP) | 1960 | 1.183270 |
| **259** | East Asia & Pacific (all income levels) | Fertility rate, total (births per woman) | 1960 | 5.396794 |
| **264** | East Asia & Pacific (all income levels) | GDP per capita (current US$) | 1960 | 146.814138 |
| **282** | East Asia & Pacific (all income levels) | Life expectancy at birth, total (years) | 1960 | 48.298317 |
| **349** | East Asia & Pacific (all income levels) | Urban population (% of total) | 1960 | 22.471132 |

To a «flat» dataset like this:



| CountryName | Year | CO2 emissions (kg per 2005 US$ of GDP) | Fertility rate, total (births per woman) | GDP per capita (current US$) | Income share held by highest 10% | Life expectancy at birth, total (years) | Long-term unemployment (% of total unemployment) | Military expenditure (% of GDP) | Physicians (per 1,000 people) | Population growth (annual %) | Rail lines (total route-km) | Rural population growth (annual %) | Urban population (% of total) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Afghanistan | 1960 | NaN | 7.45 | 59.787681 | NaN | 32.328512 | NaN | NaN | 0.034844 | 1.813677 | NaN | 1.511229 | 8.221 |
| Afghanistan | 1961 | NaN | 7.45 | 59.890037 | NaN | 32.777439 | NaN | NaN | NaN | 1.874003 | NaN | 1.560800 | 8.508 |
| Afghanistan | 1962 | NaN | 7.45 | 58.505995 | NaN | 33.219902 | NaN | NaN | NaN | 1.932414 | NaN | 1.607275 | 8.805 |
| Afghanistan | 1963 | NaN | 7.45 | 78.802587 | NaN | 33.657878 | NaN | NaN | NaN | 1.989785 | NaN | 1.654771 | 9.110 |
| Afghanistan | 1964 | NaN | 7.45 | 82.231395 | NaN | 34.092878 | NaN | NaN | NaN | 2.046675 | NaN | 1.698401 | 9.426 |

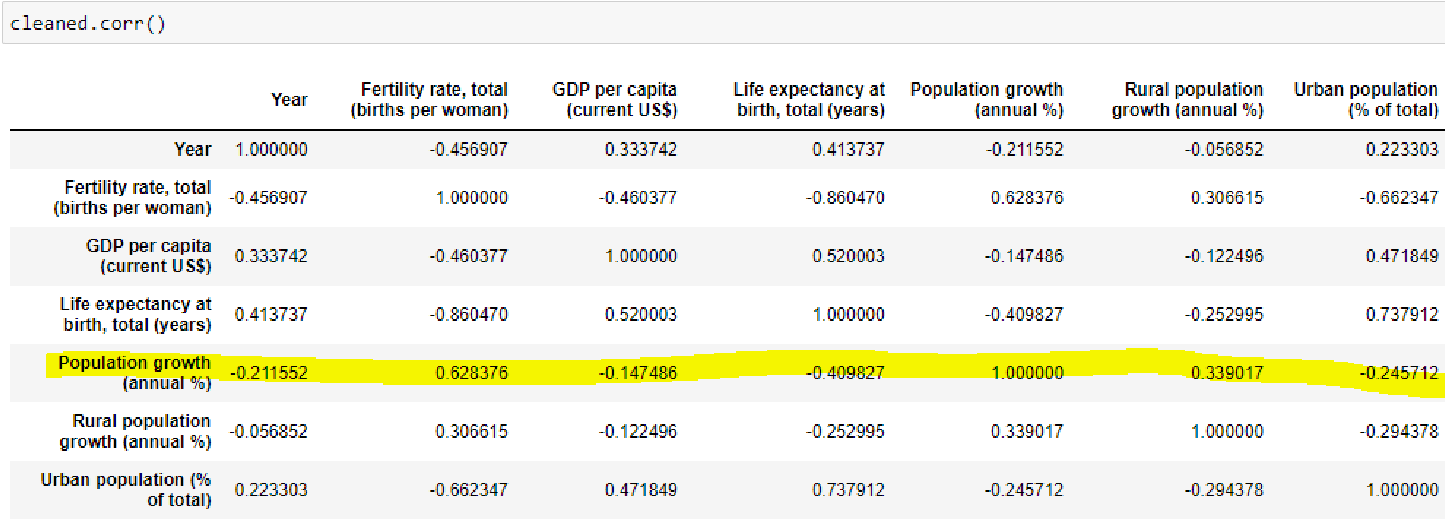
# Missing values.

I was not lucky with my chosen indicators.
There were 13442 observations of population growth, and I found that six of the indicators didn’t have sufficient observations to be of use. I there fore dropped theese.

I ended up with a dataframe with 9560 rows and 8 coulumns

# Correlation

None of the indicators showed much correlation with the population growth so you will not get a good prediction if you use only one of the chosen indicators.



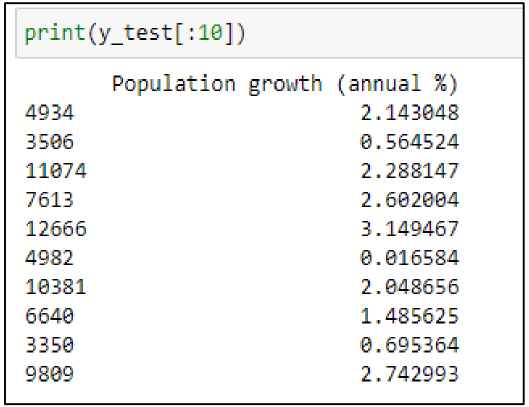
```
cleaned.corr()
```

| | Year | Fertility rate, total (births per woman) | GDP per capita (current US$) | Life expectancy at birth, total (years) | Population growth (annual %) | Rural population growth (annual %) | Urban population (% of total) |
|---|---|---|---|---|---|---|---|
| **Year** | 1.000000 | -0.456907 | 0.333742 | 0.413737 | -0.211552 | -0.056852 | 0.223303 |
| **Fertility rate, total (births per woman)** | -0.456907 | 1.000000 | -0.460377 | -0.860470 | 0.628376 | 0.306615 | -0.662347 |
| **GDP per capita (current US$)** | 0.333742 | -0.460377 | 1.000000 | 0.520003 | -0.147486 | -0.122496 | 0.471849 |
| **Life expectancy at birth, total (years)** | 0.413737 | -0.860470 | 0.520003 | 1.000000 | -0.409827 | -0.252995 | 0.737912 |
| **Population growth (annual %)** | -0.211552 | 0.628376 | -0.147486 | -0.409827 | 1.000000 | 0.339017 | -0.245712 |
| **Rural population growth (annual %)** | -0.056852 | 0.306615 | -0.122496 | -0.252995 | 0.339017 | 1.000000 | -0.294378 |
| **Urban population (% of total)** | 0.223303 | -0.662347 | 0.471849 | 0.737912 | -0.245712 | -0.294378 | 1.000000 |

# Training and testing the model

I use the sci-kit-learn model to train the model using the remaining five indicators. I have Fixed the random state so that the results can be reproduced: Looking at the prediction and test-values we can see that the model does a decent job of predicting the values, but is totally off for others, ex. Row 2 (0.56 and 1.41)

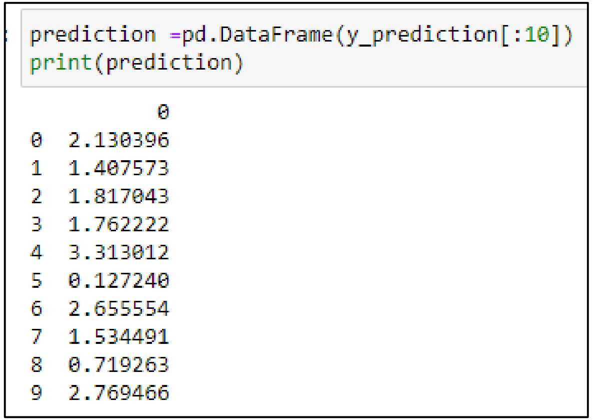
Test first 10 observations:



```
print(y_test[:10])
```

```
       Population growth (annual %)
4934                       2.143048
3506                       0.564524
11074                      2.288147
7613                       2.602004
12666                      3.149467
4982                       0.016584
10381                      2.048656
6640                       1.485625
3350                       0.695364
9809                       2.742993
```

Preditcted first 10 observations



```
prediction =pd.DataFrame(y_prediction[:10])
print(prediction)
```

```
          0
0  2.130396
1  1.407573
2  1.817043
3  1.762222
4  3.313012
5  0.127240
6  2.655554
7  1.534491
8  0.719263
9  2.769466
```

# Root Mean SquareError

Looking at the Root mean squareerror (RMSE) we find that the error is 0,91. Given that the standard deviation is 1,36, this means that the RMSE is quite close to the standard deviation. The model does not fit the data very well.

By using a Decision Tree Regressor at depht 10 I was able to reduce the RMSE to 0,53. While this is better, the model is still not very satisfactory.

# Limitations

The biggest limitation for the project, was missing values. This severely limited the analysis and I was forced to drop many of my chosen indicators. Given time and effort this could have been explored further.

# Conclusions

The model clearly had some limitations for predicting the population growth. This comes as no surprise. The most obvious reason is that I did not include anything about rate of deaths.

I did however find that while the model is not perfect, it still did a decent job of predictiong population growth. I belive that (given more time) I would be able to predict population growth based on many of the indicators.

# Acknowledgements

I did not have any formal or informal input.

# References

If applicable, report any references you used in your work.  For example, you may have used a research paper from X to help guide your analysis.  You should cite that work here. If you did all the work on your own, please state this.