STAT 663

Analyzing World Development Indicators

Zulay Attanasio
Grace Shin
Ayushi Tiwari

Introduction

- World Development Indicators (WDI) is the World Bank’s premier compilation of cross-country comparable data on development.

- The dataset for this project was compiled by combining some of the World Bank’s development indicators.

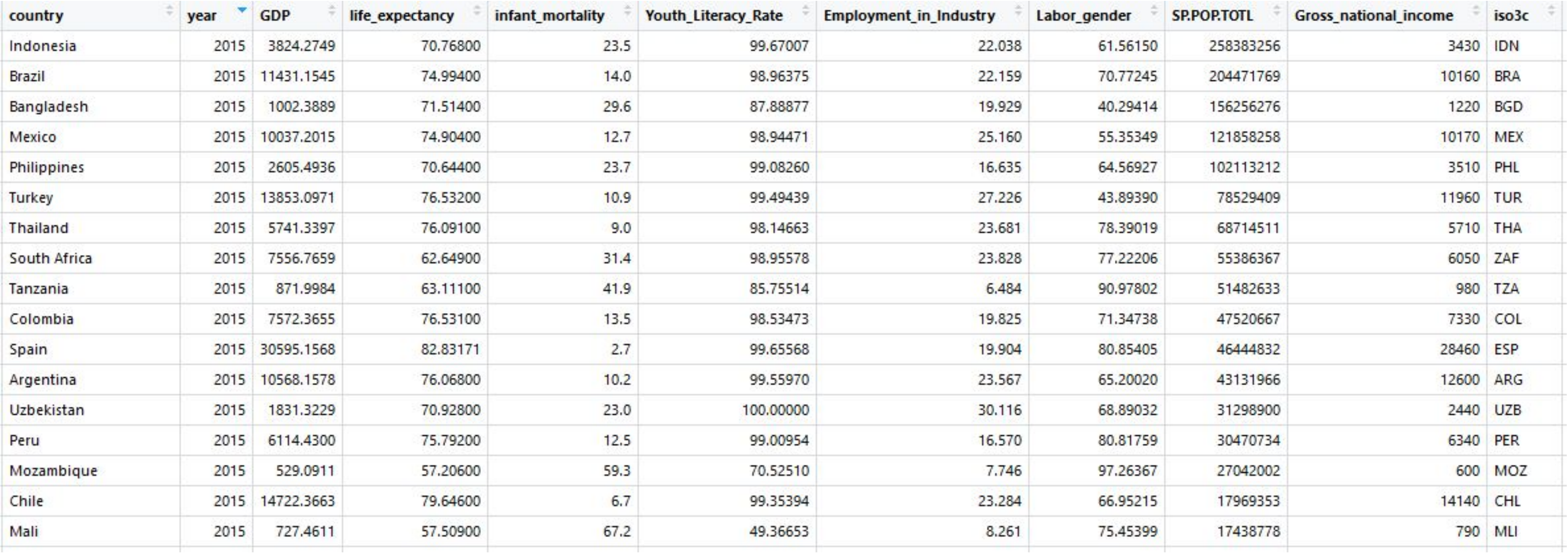
- The project aims to focus on the exploratory graphics, predictive models and analysis of results to predict the life expectancy.

- The exploratory data analysis starts with comparing indicators with each other and observing whether they are correlated.

Indicators used in the Dataset

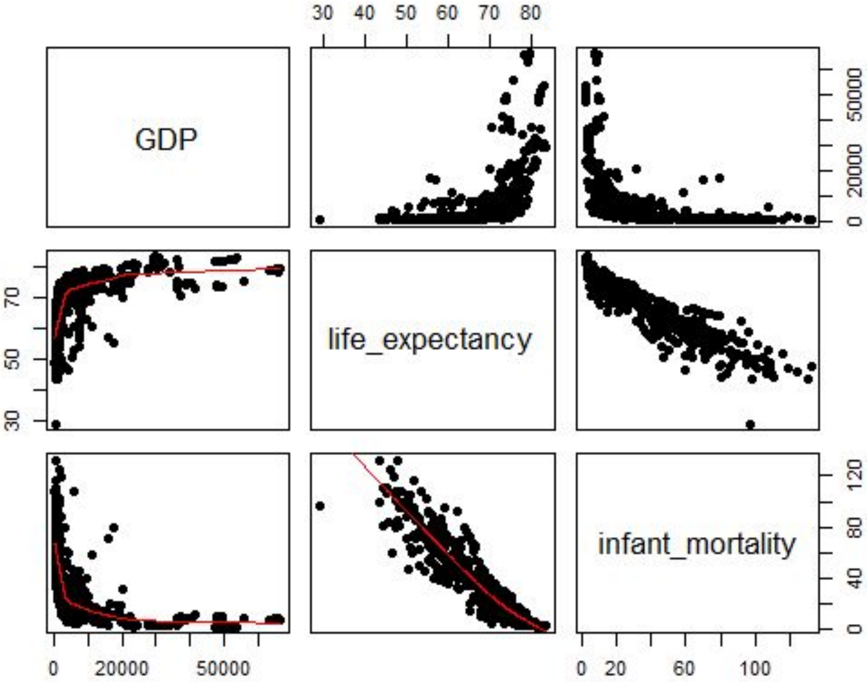
1. Population-total
2. Gross Domestic Product in US$.
3. Total Youth Literacy Rate (% of people aged 15-24)
4. Gross National Income (per capita) in US$
5. the Ratio of Female to Male Labor Force Participation Rate (%)
6. Employment in Industry (% of total employment)
7. Infant Mortality Rate (per 1000 live births)
8. Life Expectancy at Birth (in years)

Dataset Snippet

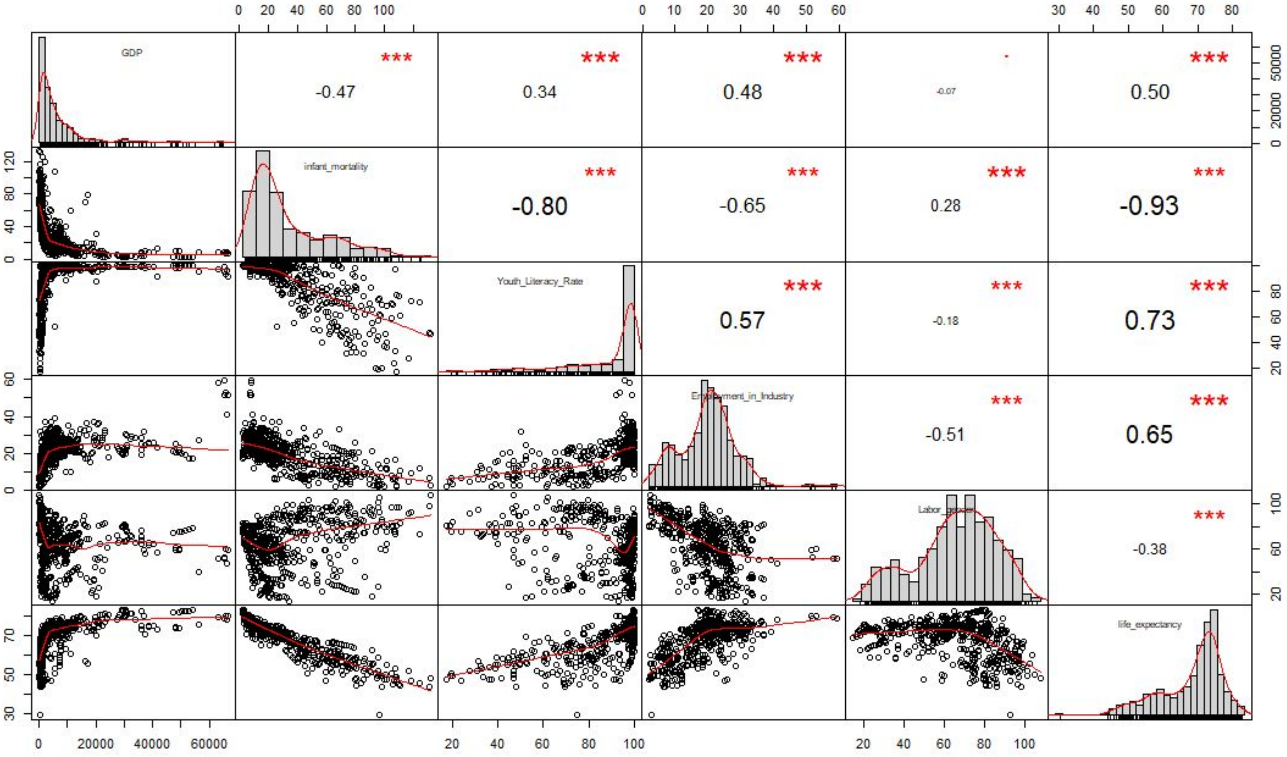


country	year	GDP	life_expectancy	infant_mortality	Youth_Literacy_Rate	Employment_in_Industry	Labor_gender	SP.POP.TOTL	Gross_national_income	iso3c
Indonesia	2015	3824.2749	70.76800	23.5	99.67007	22.038	61.56150	258383256	3430	IDN
Brazil	2015	11431.1545	74.99400	14.0	98.96375	22.159	70.77245	204471769	10160	BRA
Bangladesh	2015	1002.3889	71.51400	29.6	87.88877	19.929	40.29414	156256276	1220	BGD
Mexico	2015	10037.2015	74.90400	12.7	98.94471	25.160	55.35349	121858258	10170	MEX
Philippines	2015	2605.4936	70.64400	23.7	99.08260	16.635	64.56927	102113212	3510	PHL
Turkey	2015	13853.0971	76.53200	10.9	99.49439	27.226	43.89390	78529409	11960	TUR
Thailand	2015	5741.3397	76.09100	9.0	98.14663	23.681	78.39019	68714511	5710	THA
South Africa	2015	7556.7659	62.64900	31.4	98.95578	23.828	77.22206	55386367	6050	ZAF
Tanzania	2015	871.9984	63.11100	41.9	85.75514	6.484	90.97802	51482633	980	TZA
Colombia	2015	7572.3655	76.53100	13.5	98.53473	19.825	71.34738	47520667	7330	COL
Spain	2015	30595.1568	82.83171	2.7	99.65568	19.904	80.85405	46444832	28460	ESP
Argentina	2015	10568.1578	76.06800	10.2	99.55970	23.567	65.20020	43131966	12600	ARG
Uzbekistan	2015	1831.3229	70.92800	23.0	100.00000	30.116	68.89032	31298900	2440	UZB
Peru	2015	6114.4300	75.79200	12.5	99.00954	16.570	80.81759	30470734	6340	PER
Mozambique	2015	529.0911	57.20600	59.3	70.52510	7.746	97.26367	27042002	600	MOZ
Chile	2015	14722.3663	79.64600	6.7	99.35394	23.284	66.95215	17969353	14140	CHL
Mali	2015	727.4611	57.50900	67.2	49.36653	8.261	75.45399	17438778	790	MLI

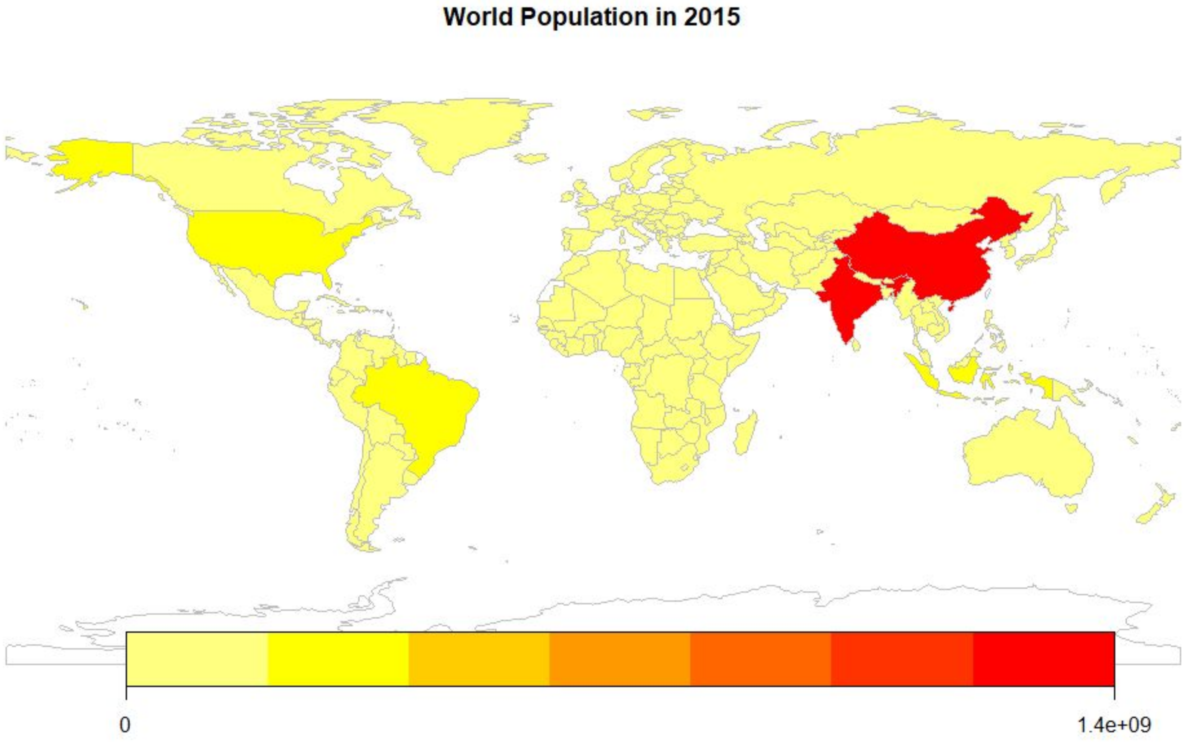
Scatter Plot Matrix



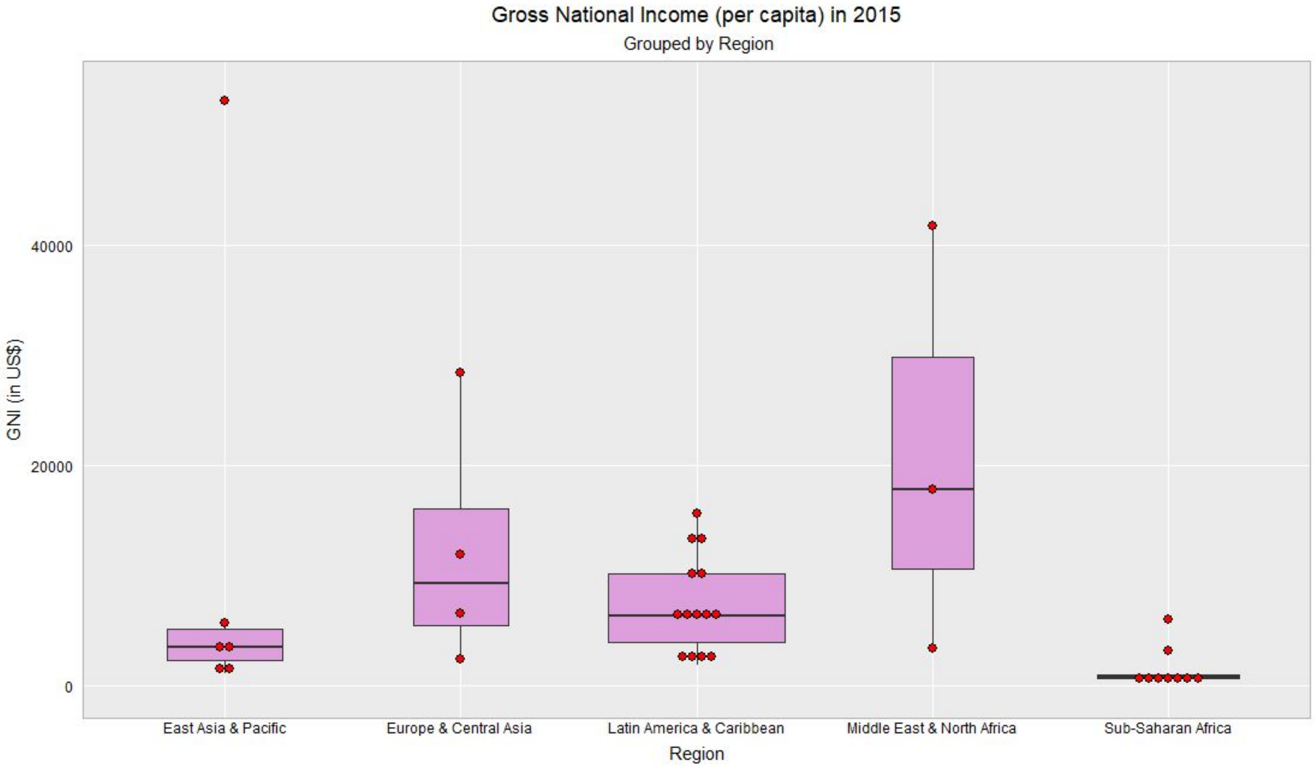
Correlation Chart



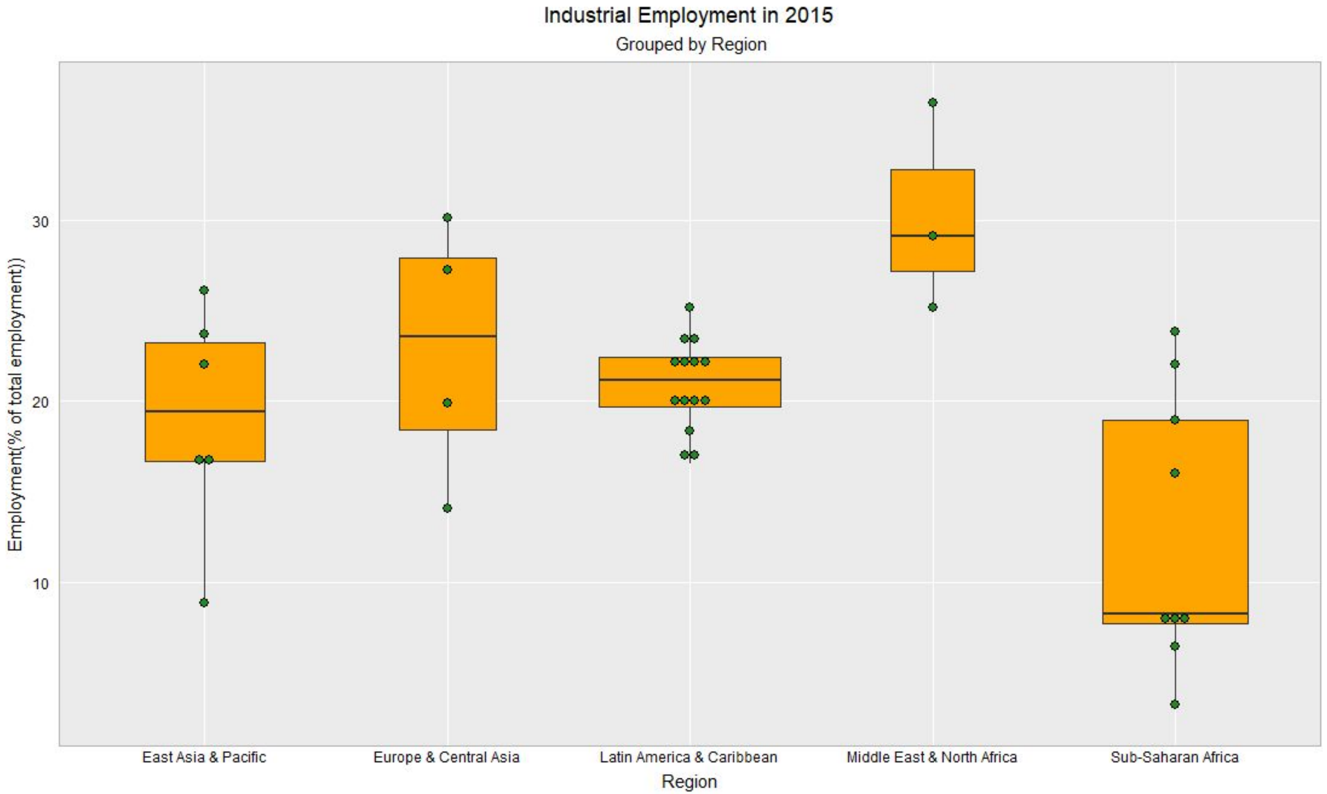
World population in 2015



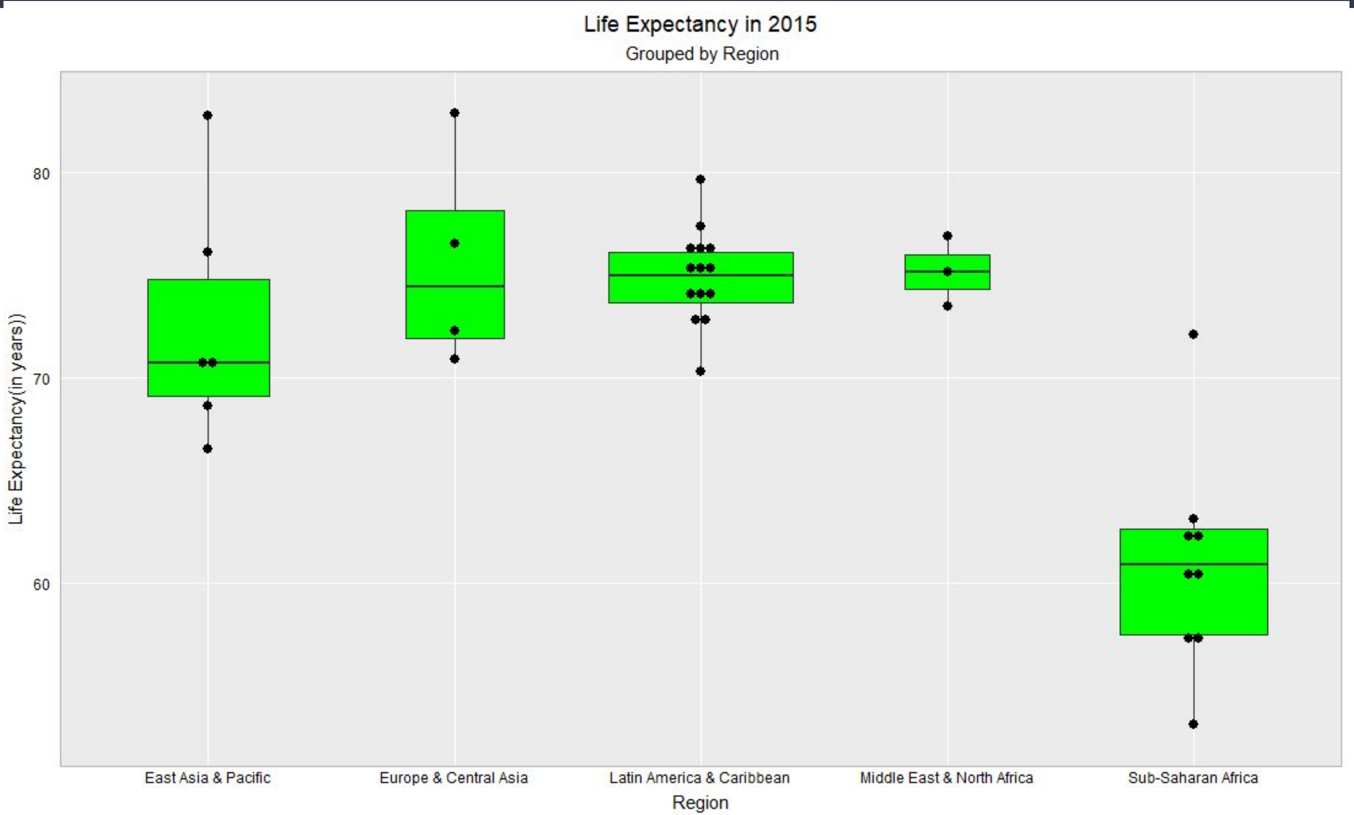
Gross National Income in 2015



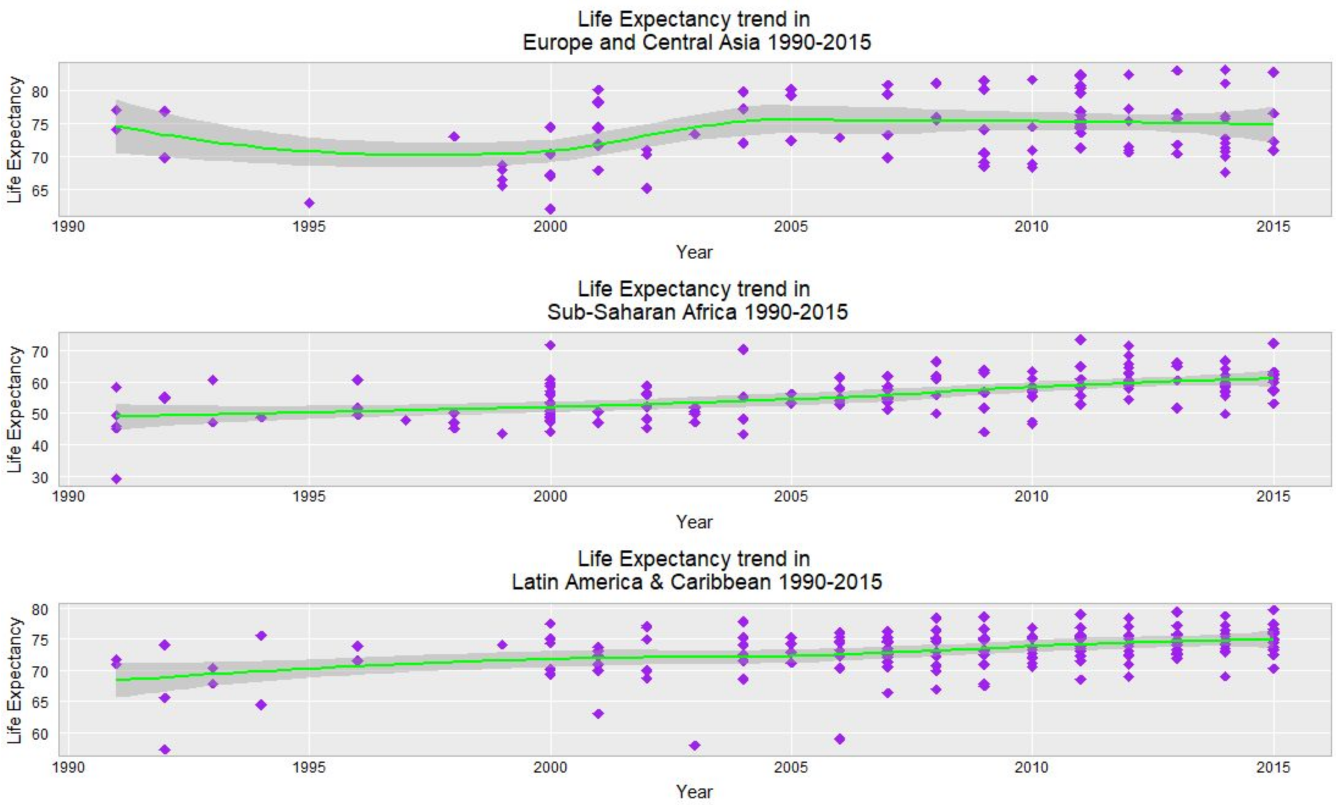
Distribution of Employment in the Industry



Life Expectancy in 2015



Trends in Life Expectancy 1990-2015



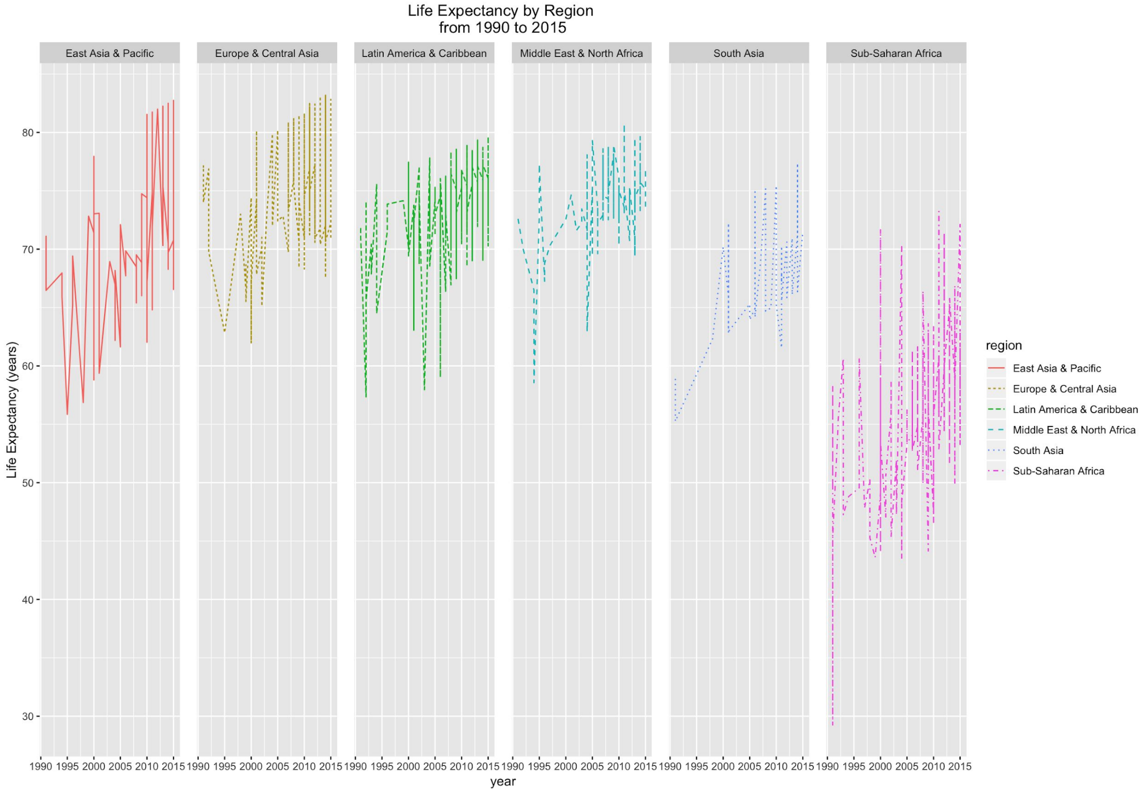
Life Expectancy trend in Europe and Central Asia 1990-2015

Life Expectancy trend in Sub-Saharan Africa 1990-2015

Life Expectancy trend in Latin America & Caribbean 1990-2015

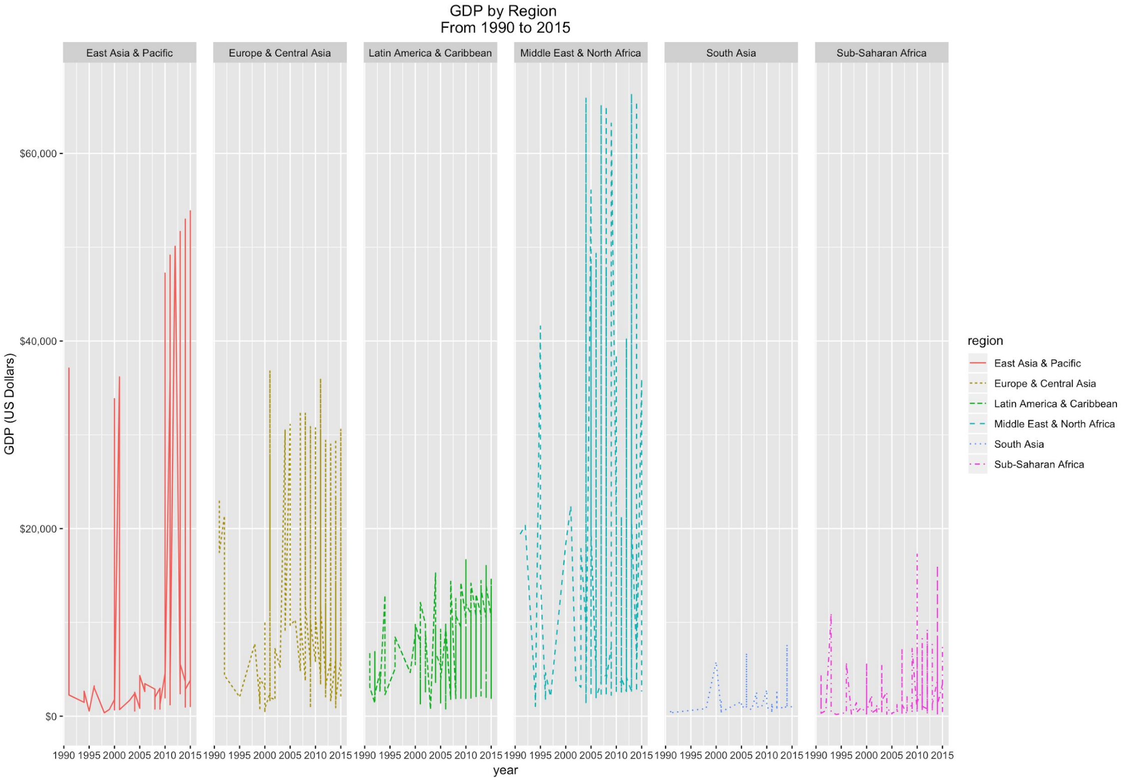
Time Series Plot: Life Expectancy (1990-2015)

- Regions with the highest life expectancy:
  - East Asia & Pacific
  - Europe & Central Asia
- Regions with the lowest GDP:
  - Sub-Saharan Africa
  - South Asia

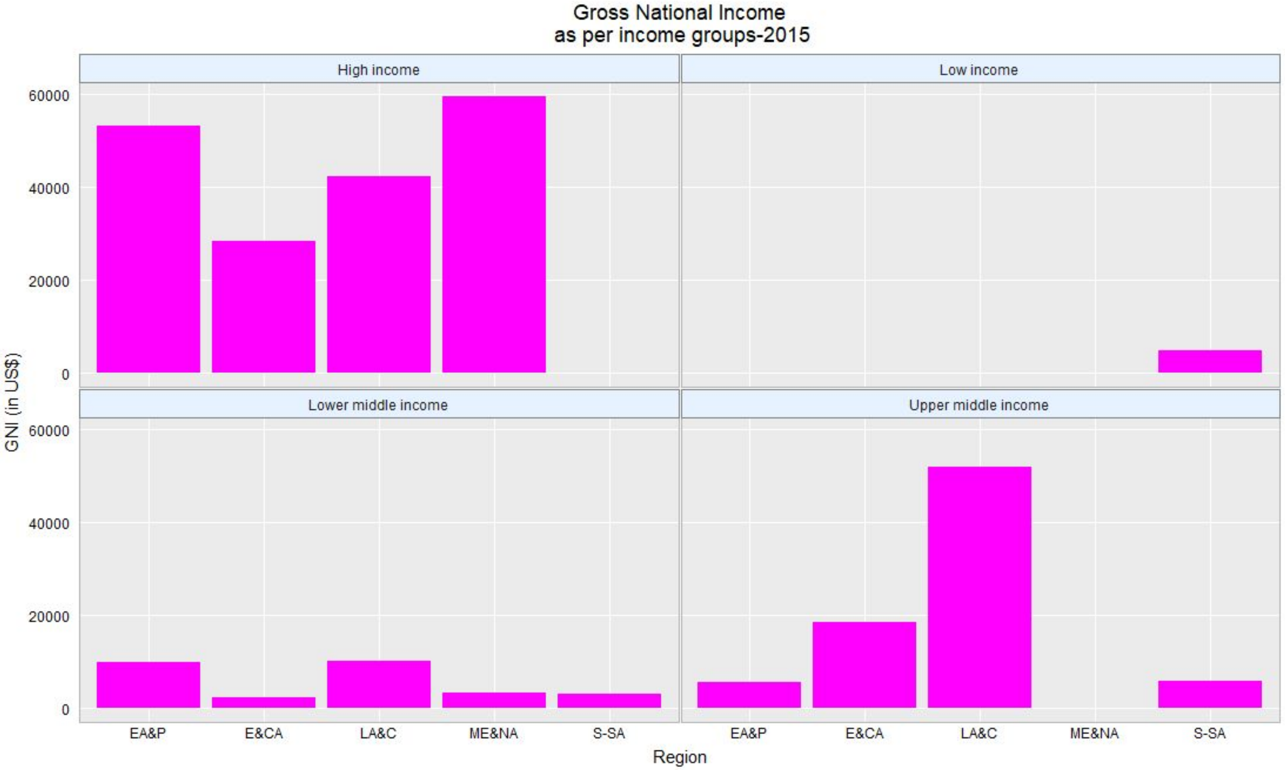


Time Series Plot: GDP (1990-2015)

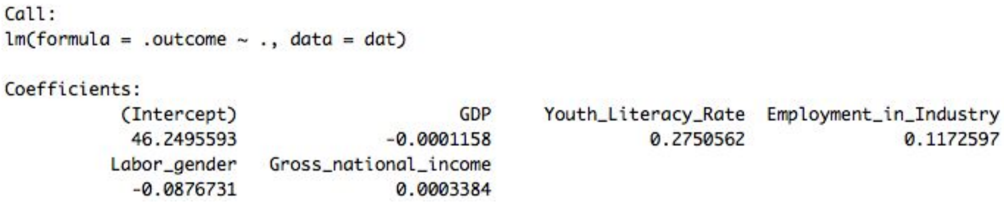
- Regions with the highest GDP:
  - East Asia & Pacific
  - Middle East & North Africa
- Regions with the lowest GDP:
  - Sub-Saharan Africa
  - South Asia
- From 2005: GDP for Middle East & North Africa significantly increased



Region wise comparison of Gross National Income

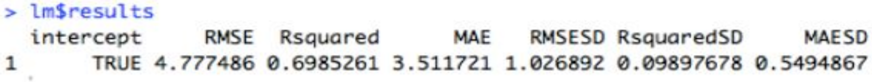


Models: Multiple Linear Regression (CV)

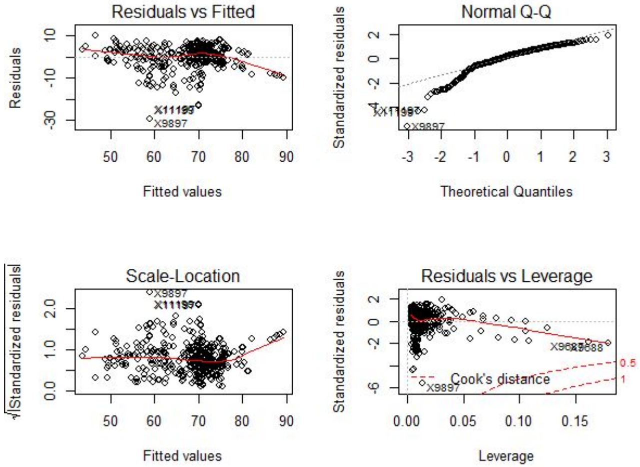


```
Call:
lm(formula = .outcome ~ ., data = dat)

Coefficients:
          (Intercept)                    GDP     Youth_Literacy_Rate  Employment_in_Industry
           46.2495593             -0.0001158               0.2750562               0.1172597
         Labor_gender  Gross_national_income
           -0.0876731              0.0003384
```

- ***Life expectancy*** is predicted to increase *by 0.0003384* years if the Gross National income rate is increased by 1%, holding all other predictors fixed.



```
> lm$results
  intercept     RMSE  Rsquared      MAE   RMSESD RsquaredSD     MAESD
1      TRUE 4.777486 0.6985261 3.511721 1.026892 0.09897678 0.5494867
```



- The constant variance assumption (homoscedasticity) is violated.

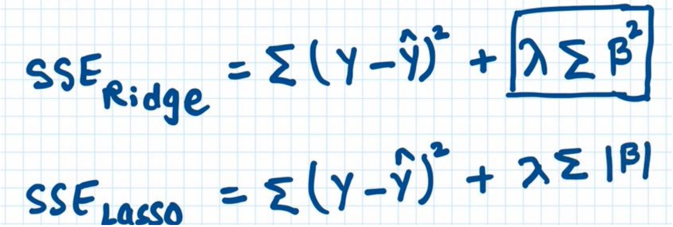
- Some points in the dataset do not lie in the straight line nearby to the tails. Transformation will be useful (Box-Cox).

Models

- If independent variables are highly correlated it creates multicollinearity problems..
- Collinearity leads to overfitting

<u>Solution:</u>

- ➢ Ridge regression: Shrink coefficients to non-zero values to prevent overfitting, but keeps all variables.
- ➢ Lasso regression: Shrink regression coefficients, with some shrunk to zero, helping with feature selection.

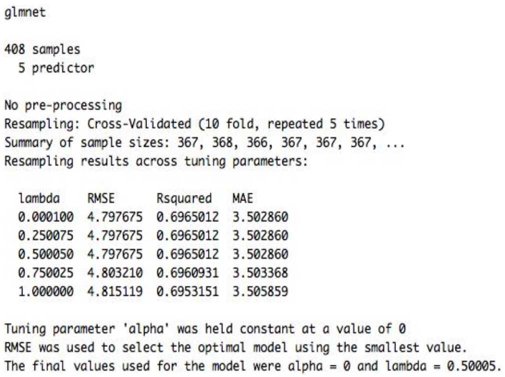


$$SSE_{Ridge} = \Sigma(y-\hat{y})^2 + \boxed{\lambda \Sigma \beta^2}$$

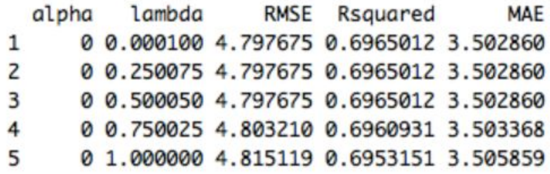
$$SSE_{Lasso} = \Sigma(y-\hat{y})^2 + \lambda \Sigma |\beta|$$

Models: Ridge Regression

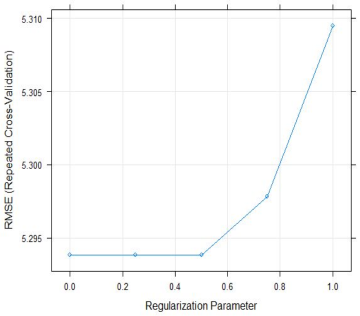
Adding more variables to the model incurred in a reduction of the RMSE and an increase in the R-square. Ridge fits a model containing all predictors and shrinks the coefficients towards zero.

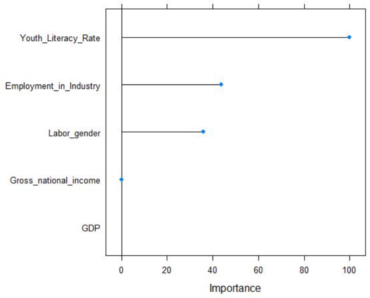


```
glmnet

408 samples
  5 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 367, 368, 366, 367, 367, 367, ...
Resampling results across tuning parameters:

  lambda    RMSE      Rsquared   MAE
  0.000100  4.797675  0.6965012  3.502860
  0.250075  4.797675  0.6965012  3.502860
  0.500050  4.797675  0.6965012  3.502860
  0.750025  4.803210  0.6960931  3.503368
  1.000000  4.815119  0.6953151  3.505859

Tuning parameter 'alpha' was held constant at a value of 0
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were alpha = 0 and lambda = 0.50005.
```



```
  alpha   lambda     RMSE  Rsquared      MAE
1     0 0.000100 4.797675 0.6965012 3.502860
2     0 0.250075 4.797675 0.6965012 3.502860
3     0 0.500050 4.797675 0.6965012 3.502860
4     0 0.750025 4.803210 0.6960931 3.503368
5     0 1.000000 4.815119 0.6953151 3.505859
```



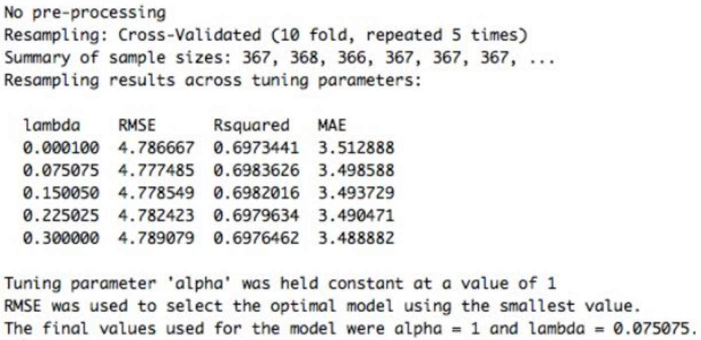


Results show that the RMSE is 4.7976 and R square is equal to 0.6965 when using a lambda equal to 0.50.

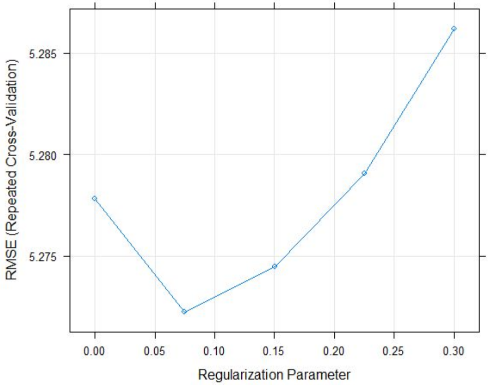
<u>Important Variables:</u>

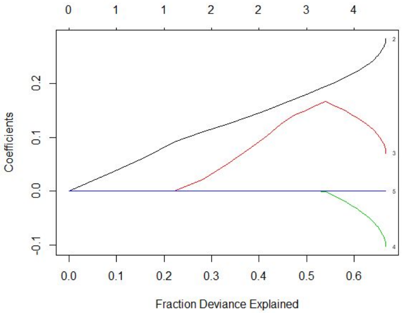
- Youth literacy
- Employment in industry
- Labor gender

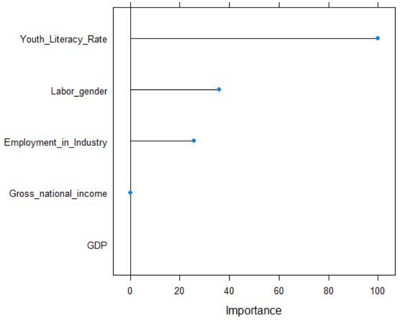
Models: LASSO Regression



```
No pre-processing
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 367, 368, 366, 367, 367, 367, ...
Resampling results across tuning parameters:

  lambda    RMSE      Rsquared   MAE
  0.000100  4.786667  0.6973441  3.512888
  0.075075  4.777485  0.6983626  3.498588
  0.150050  4.778549  0.6982016  3.493729
  0.225025  4.782423  0.6979634  3.490471
  0.300000  4.789079  0.6976462  3.488882

Tuning parameter 'alpha' was held constant at a value of 1
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were alpha = 1 and lambda = 0.075075.
```



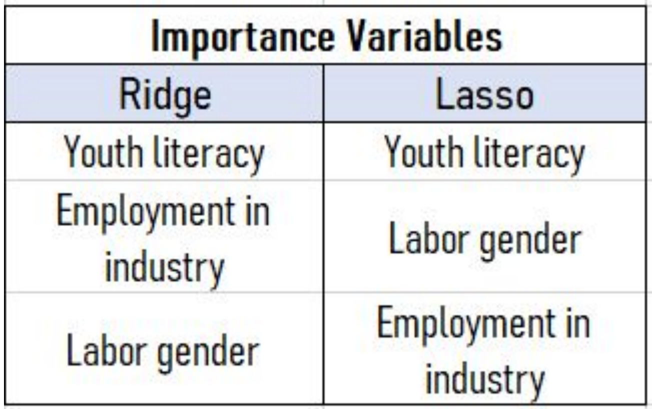




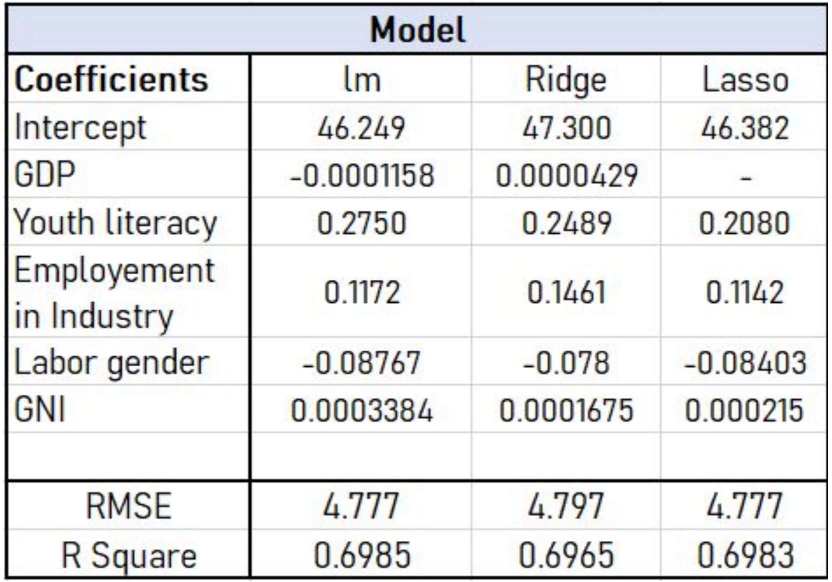
- *Results show that the RMSE is 4.77 and R square is equal to 0.6983 when using a small lambda (0.0750).*
- *60% of the variability is being explained only by four variables. The variable that grows rapidly next to the bottom corner has the least importance in the model compared to the variable at the top corner.*
- *The variable Youth literacy, Labor gender, and Employment in industry represent the most important variables.*

Models Comparison

- Models having the *best performance are “lm” and “Lasso”.*



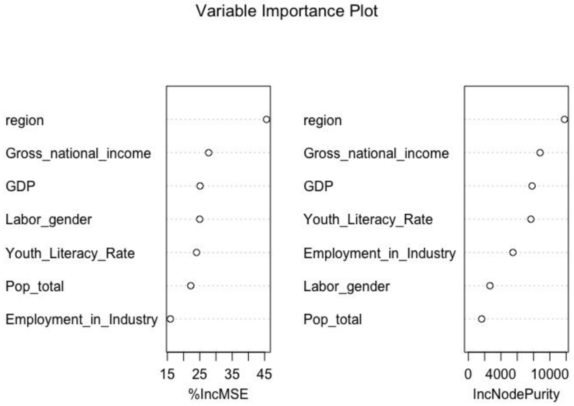
Importance Variables	
Ridge	Lasso
Youth literacy	Youth literacy
Employment in industry	Labor gender
Labor gender	Employment in industry

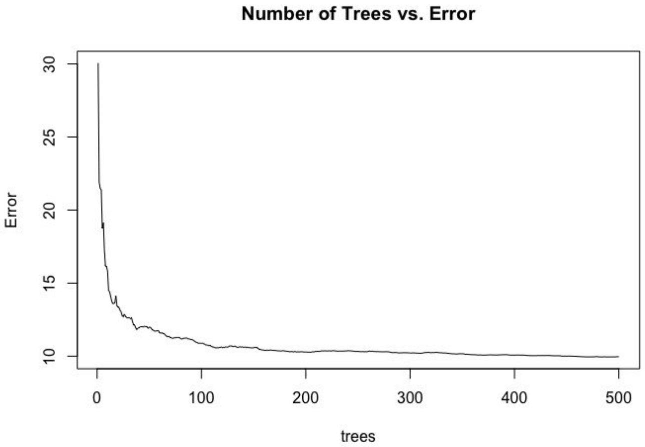


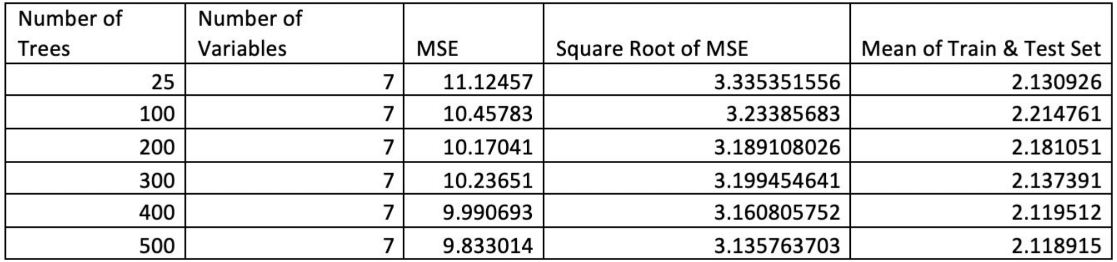
	Model		
Coefficients	lm	Ridge	Lasso
Intercept	46.249	47.300	46.382
GDP	-0.0001158	0.0000429	-
Youth literacy	0.2750	0.2489	0.2080
Employement in Industry	0.1172	0.1461	0.1142
Labor gender	-0.08767	-0.078	-0.08403
GNI	0.0003384	0.0001675	0.000215
RMSE	4.777	4.797	4.777
R Square	0.6985	0.6965	0.6983

Models: Random Forest

- Important Variables:
  - Region
  - Gross National Income
  - GDP
  - Labor gender ratio
- Error rate decreases at 100 trees and continues to decrease and stabilize at approximately 300 trees
- MSE (mean squared error) is 11.125 at 25 trees and 9.833 at 500 trees, does not change significantly at 400 and 500 trees







Number of Trees	Number of Variables	MSE	Square Root of MSE	Mean of Train & Test Set
25	7	11.12457	3.335351556	2.130926
100	7	10.45783	3.23385683	2.214761
200	7	10.17041	3.189108026	2.181051
300	7	10.23651	3.199454641	2.137391
400	7	9.990693	3.160805752	2.119512
500	7	9.833014	3.135763703	2.118915

Conclusion

- Linear Regression Mode (all other predictors fixed):
  - Life expectancy was predicted to increase by 0.0003384 years if the Gross National income rate is increased by 1%
  - Life expectancy increase by 0.2750 years if the variable Youth literacy increase by 1%
- Important Variables:
  - Ridge Regression: Youth literacy, Employment in industry and Labor gender
  - LASSO Regression: Youth literacy, Labor gender and Employment in Industry
  - Random Forest: Region, Gross National Income, GDP, and Labor Gender

Future Work

- Data analysis using normalization to improve the model’s performance
- Further analysis of the data in each of the variables to determine skewness and outliers to improve the models used
- Incorporate more indicators such as population and income
- Other models such as clustering