Identification of factors influencing the population growth across the world via predictive modeling

Group A-Cube : Aman Bagla, Akanksha Devikar, Aman Thakkar

1 Introduction

Population growth over the world is dependent on parameters like demography, geography, climate, economic conditions, educational characteristics, ethno-cultural characteristics etc. Our aim is to identify region-specific factors affecting population growth. Based on studies, we consider the target value of population density to be 75 per square Km. We focus our study on two categories of countries:

(a) Countries where population density is below the target value
(b) Countries where population density is above the target value

If the countries in category (a) have a negative population growth or countries in category (b) have positive population growth, then we need to have some population control strategies. For this purpose, we considered population growth rate (in percent) and population density averaged over a decade (2005-2015) for each country to define a variable as our category variable as per following formula:

$$A = AveragePopulationGrowthRate(\%)$$

$$B = \frac{(AveragePopulationDensity - TargetValue)}{TargetValue}$$

$$PopulationControlRequirementIndex(PCRI) = A * B$$

A	B	PCRI	Requirement
Decreasing (-)	Decreasing (-)	Positive	Increase population growth rate
Decreasing (-)	Increasing (+)	Negative	-
Increasing (+)	Decreasing (-)	Negative	-
Increasing (+)	Increasing (+)	Positive	Decrease population growth rate

Table 1: Classes of PCRI.

If population density is lower than target value (75 per sq. Km) and population is decreasing, we need to find factors that can help in increasing the population. Similarly, if population density is greater than target value and population is increasing, we need to find factors that can help to curb the population growth. Also, higher the value of PCRI, more is the need to have population control policies as population density is farther from the target value.

Since population density of countries cannot be directly controlled, we developed a model on our entire data set to identify the factors affecting population growth rate in these countries. We use PCRI to identify countries that require population control policies and to provide more focused and directed recommendations. In conclusion, we hypothesize that population growth is a function of factors like demography, geography, climate, economic conditions, educational characteristics and ethno-cultural characteristics. To test our hypothesis, we have collected a total of 27 predictors in all categories. We want to use predictive modeling to test our hypothesis.

Population density is the measure of number of people per unit area or unit volume. Various studies estimate different ideal or optimum population densities. A world conference on transport research estimated the optimal density ranges from 20-40 residents per hectare, based on the criterion of transportation and municipal costs.

Litman,T.[5] estimates the population density ranges seems to be 2,801 inhabitants per sq. Km based on the criterion of paving and lighting sector dimensions, and to be 4,430 inhabitants per sq. Km based on the dimensions and sewerage and cleansing of residual waters.

A study published in Journal of Economics [4] carries out a detailed physics based economic analysis and came to the conclude that the optimum population density should be 300 people per sq. Km.

In this study the author assumes that all the wealth creation and resources come from the land. Earth has a population density of 6 per sq. Km, assuming world population to be 6 billion. The author then goes on to say that the closer a country is to this number the more a country is natural resources based, for example Canada. If a country has a significantly higher population density then this it tends to have poverty, for example India. When a country has this number higher but still has a healthy GDP, it means that the country is importing the goods (stealing others hinterland) or is a service-based economy. If the population density is lower, such as Australia, it means that the country can grow.

The problem of finding an optimum population density is a complex and a multivariate problem, with the need to take diverse factors into account. Deviation from this optimum population density can in turn affect the variables and can serve as an important measure that can give important insights.

2 Data

We have merged about 11 files .csv files to generate our final input data frame. These files contain our important predictors as described below and were merged based on the unique country codes generated for each of these files. After removing the missing data, we focus our study on 143 countries which cover majority of the world population-wise as well as geographically. Our response variable is population growth rate of each country averaged over 10 years (2005-2015). In correspondence, our predictors are also averaged over the 10 years horizon. All our predictors are either percentages or per-capita values. The final set of variables that we choose and fit our model onto are Birth Rate, Death Rate, Fertility Rate, Infant Mortality Rate, Life Expectancy at birth, Education levels, GDP per-capita, Employment by Industry, Unemployment Rate, CO2 emissions per capita, Fertile soil percentage, Proportion of desert land, Proportion of region with tropical climate, Proportion of coastal land, and Proportion of rugged terrain. [See Appendix for details]

Summary statistics is attached in Appendix

3 Exploratory Data Analysis

We have plotted the PCRI for all the countries in our data set.

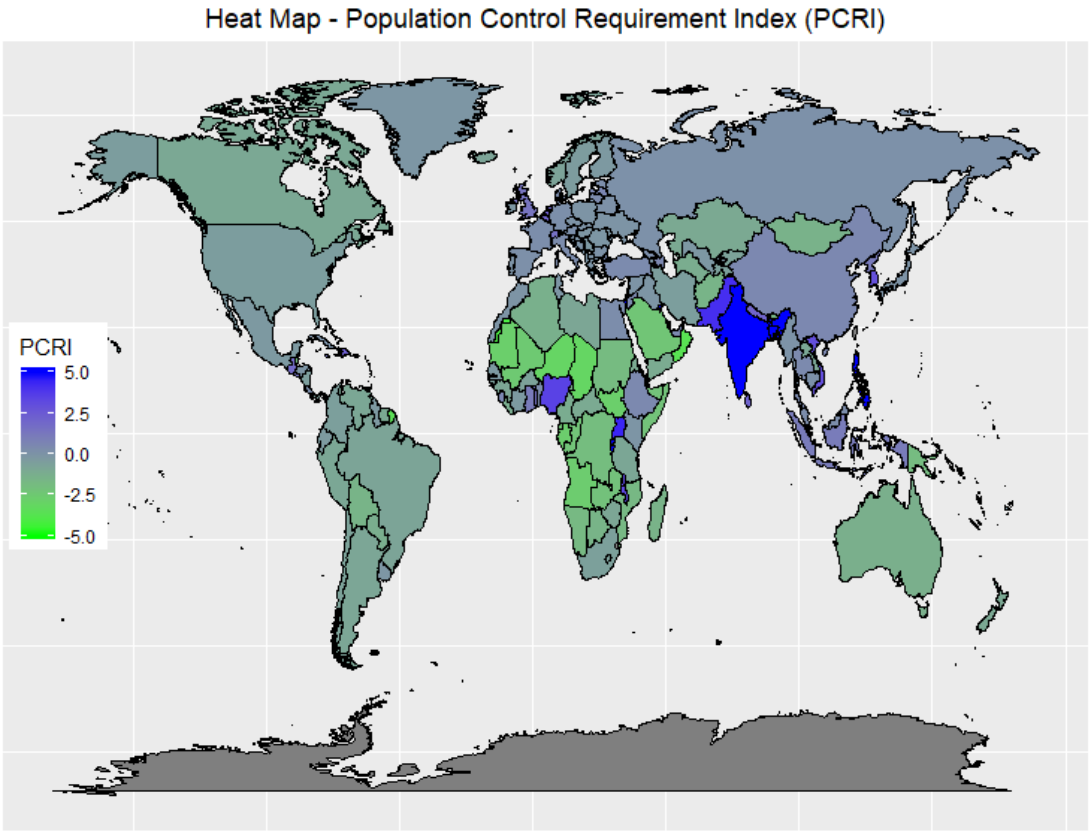

Figure 1: Heat map based on PCRI

Based on the value of PCRI, top 25 countries that need population control policies are: Monaco, Singapore, Bahrain, Maldives, Lebanon, Sint Maarten, Bangladesh, Palestine, Mayotte, Rwanda, Burundi, Comoros, Malta, Aruba, San Marino, Nauru, Qatar, India, Israel, Philippines, Haiti, Kuwait, Cayman Islands, Uganda and Pakistan.

Here we can see that countries with PCRI value greater than zero need to plan for population growth rate control. Also, if we see the African region near Sahara desert, it is green showing growth rate is already under control. Tropical countries like India, Nigeria have high value of PCRI which completely makes sense.

Some interesting findings from exploratory data analysis are as follows:

(i) From figure 2 (correlation plot), it can be seen that birth rate and fertility rate are highly correlated with population growth rate and also other predictors like infant mortality rate and employment sectors. However, countries do not have any real control over these factors through policy making. Furthermore these factors may suppress the importance of other variables which are actionable. Hence we exclude these variables from our models.

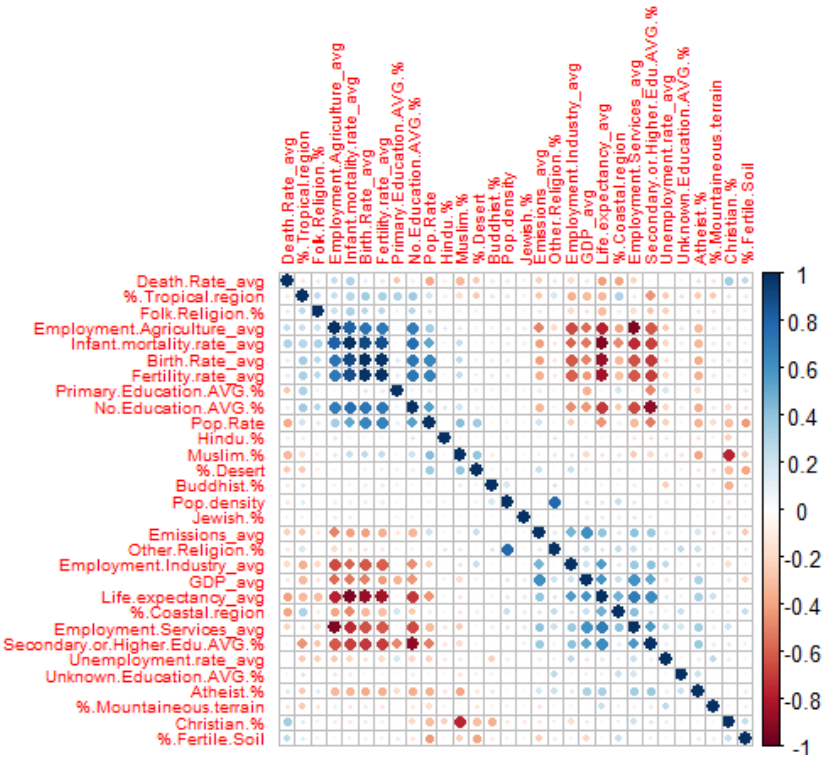

Figure 2: Correlation plot of all variables

(ii) From violin plot for percentage of people employed in agriculture, it can be seen that higher the number of people employed in agriculture higher the population growth rate. This can be attributed to the mindset in agrarian economies that more the members in the family more people can contribute to the farming.

(iii) We see that the infant mortality rate is also directly proportional to population growth rate. One of the possible reasons could be that usually higher infant mortality rate is due to inadequate healthcare and the mothers general health. Poorer and developing countries usually have higher population growth rate and weaker healthcare infrastructure.

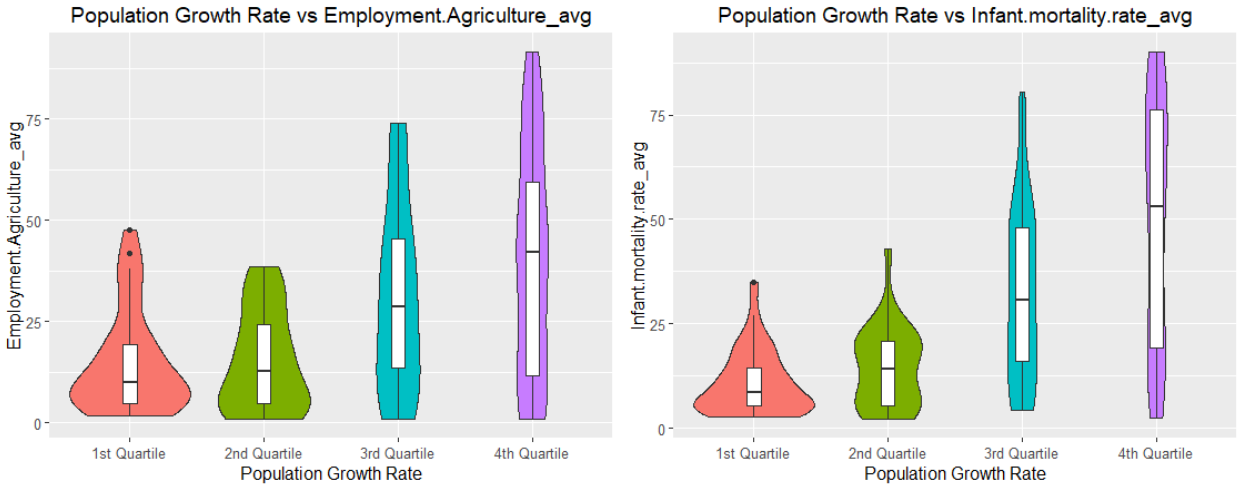

Figure 3: Violin plots of important variables

(iv) Violin plot of life expectancy is peculiar and there is negative correlation with population growth rate. More developed countries have better life expectancy and usually lower population growth rate compared to developing and poorer countries which have lower life expectancy and higher population growth rates.

(v) Another interesting finding is negative correlation of population growth rate and secondary education. This seems to be to in line with our previous observation of high population growth rate in countries with higher percentage of uneducated people. Additional violin plots are attached in appendix.

Figure 4: Violin plots of important variables

4 Modeling

After conducting our exploratory data analysis, we proceed to fit our covariates and response variables using different models explored in the class. Both our predictor and response variables are continuous in nature. We divide the 143 observations that we have into training and testing data using a 80/20 split ratio. We then proceed to fit nine models, and then tune hyperparameters in each model using 5-fold cross-validation and then select the best hyperparameters to fit our models. We compare each model using RMSE values on test data and select Gradient Boosting as our final model.

4.1 Linear Regression Model

We fit the linear regression model on our finalized variables against our response variable of Population growth rate. The adjusted R^2 is around 0.82 which explains a large amount of variation. However, we then find the RMSE value on training data to be **0.496** and on the test data to be **1.071** which is high and other models might give better performance. Linear model assumptions are checked by plotting diagnostic graphs. The residuals do not seem to have constant variance and linear model assumptions may not be valid. The plots are attached in the appendix.

4.2 Lasso and Ridge Regression

We use the glmnet package in R and fit a lasso and ridge regression model on our training data. We also tune our hyperparameter lambda using 5-fold cross-validation on training data, and choose the optimum lambda corresponding to lowest RMSE value. The relationship of lambda vs RMSE value is shown in the graph below. The final model is fitted on training set using $lambda_{min} = 5.542$ and $lambda_{min} = 0.082$ for ridge and lasso respectively. The RMSE values from the fitted values for ridge and lasso are are **0.967** and **0.608** on the training data, and **1.05** and **0.947** on the test data, respectively.

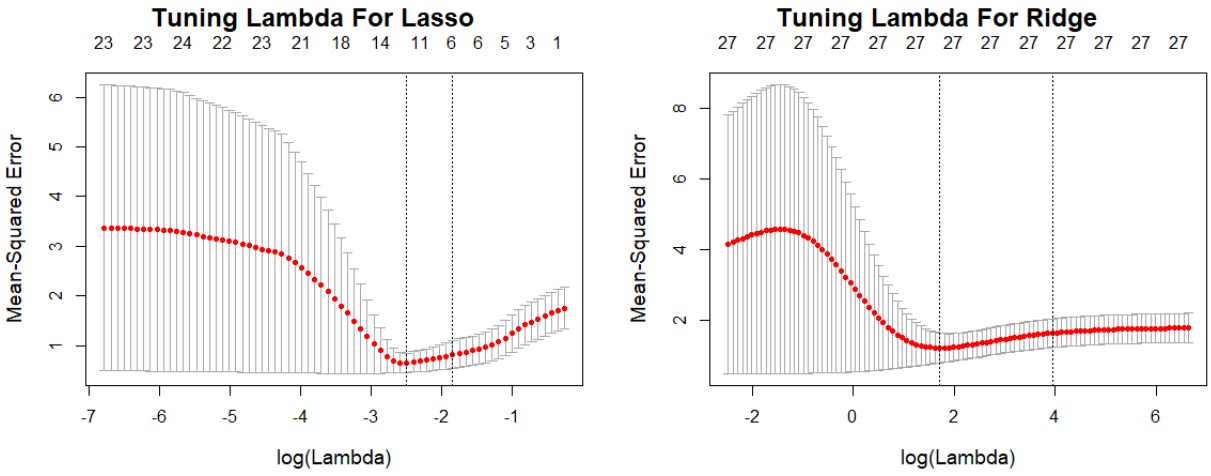

Figure 5: Lambda tuning for lasso and ridge

4.3 Generalized Additive Models (GAM)

We fit a GAM model and use spline function where the variable is non-linear. We find that the RMSE values for the fitted GAM model for training set is **0.218** and for test set is **1.145**.

4.4 Bayesian Additive Regression Trees (BART)

Using the bartMachine function and using 5-fold cross validation we tune our hyperparameters k, q , nu and m to find the least RMSE values for training and testing data. We vary k between (1,2,3) and check for 3 combinations of (q, nu) which are (0.75,10), (0.9,3), (0.9,3). Two values of m that are considered are 50 and 100. The optimal of values for k,q,nu and m are found to be **(1, 0.75, 10, 50)** as evident from Figure 6 below.

4.5 Multivariate Adaptive Regression Splines (MARS)

Using the earth package we find the optimum number of terms (including intercept) that remain after pruning. As evident from Figure 7 below, the optimum number of terms for which we get lowest RMSE value is 14. The RMSE found in MARS for training data is **0.485** and for the test data is **0.927**.

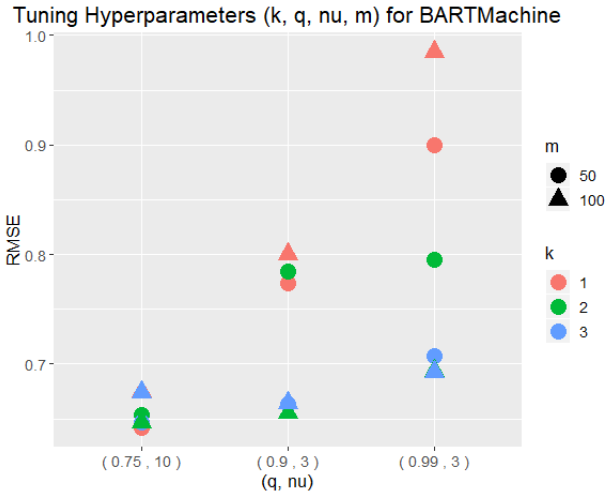

Figure 6: Hyperparameter tuning for BART

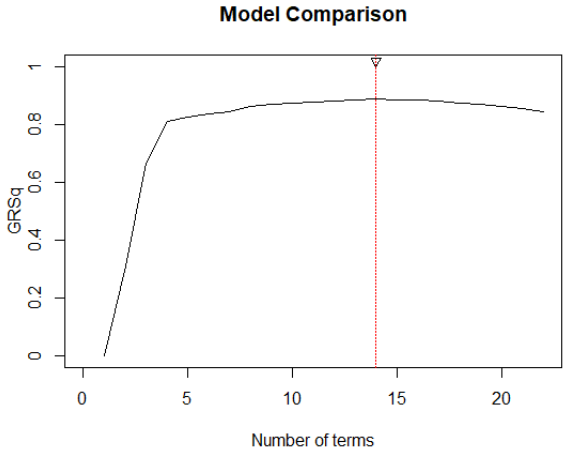

Figure 7: Optimal number of terms for MARS

4.6 Random Forest

We use 5-fold cross validation to find the optimal number of predictors to be sampled at each splitting node (mtry) at which out-of-bag error is the least. The number of trees used are 500. As evident from Figure 8, the optimal value for **mtry = 18**. The RMSE value on training data is found to be **0.294** and on the test data to be **0.866**.

4.7 Support Vector Regression (SVR)

Using tune function the hyperparameters epsilon (dictates the size of the errors that we admit in the solution) and cost (dictates the margin classification). Using 5-fold cross validation we find epsilon and cost to be 0.2 and 6 respectively,that results in the lowest RMSE values. The RMSE value of the SVM model for training data is **0.269** and on the test data is **1.255**

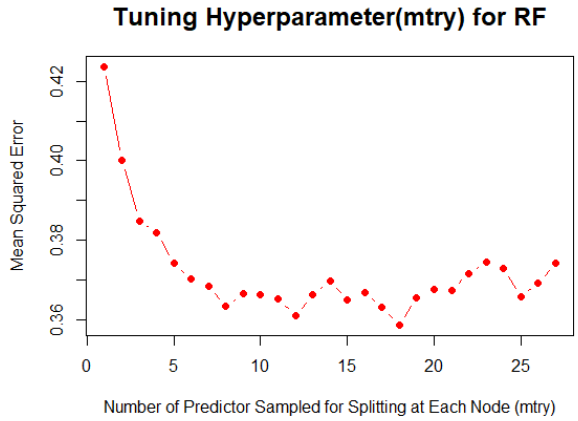

Figure 8: Tuning number of predictors for RF

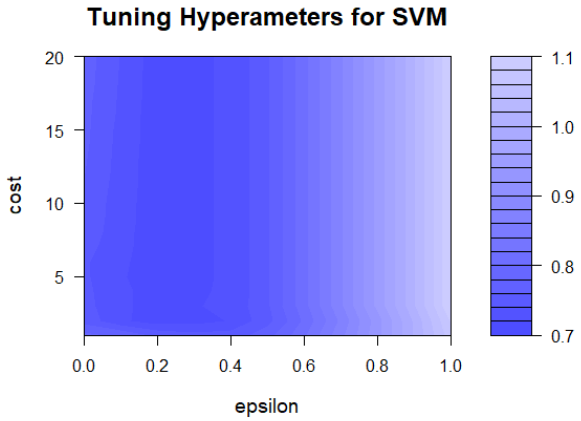

Figure 9: Tuning hyperparameters for SVR

4.8 Gradient Boosting Method (GBM)

We use 5-fold cross-validation to tune the shrinkage parameter and maximum depth of each tree (interaction.depth). We vary shrinkage between (0.01, 0.1, 0.3) and interaction.depth between (2, 3, 5). The optimal value is found to be 0.1 and 2 for shrinkage and maximum depth respectively. The number of trees (n.trees) are varied from 1500 to 2000. The best performance is found to be at $\mathbf{n.trees = 1500}$. Figure 10 shows the performance of boosting on both training and test set as a function of number of trees. The RMSE value of the boosting model for training data is **0.264** and on the test data is **0.906**

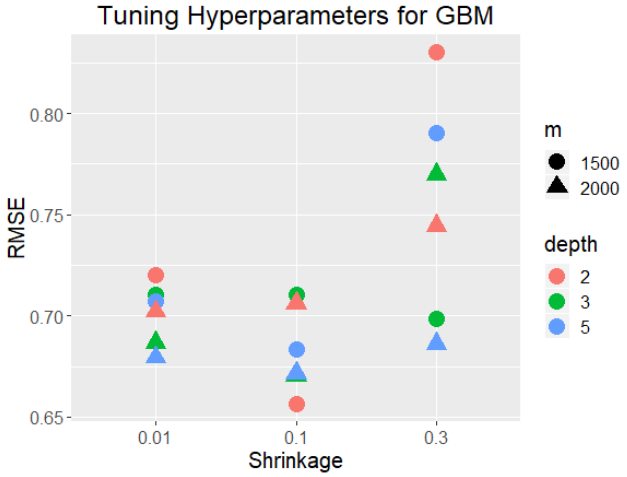

Figure 10: Tuning for Boosting

4.9 Model Comparison

Figure 11 shows the RMSE comparison between all fitted models on training and testing datasets. As we can see gradient boosting provides the best RMSE value on test data. This model also has good interpretability. Hence we select this method as our final model.

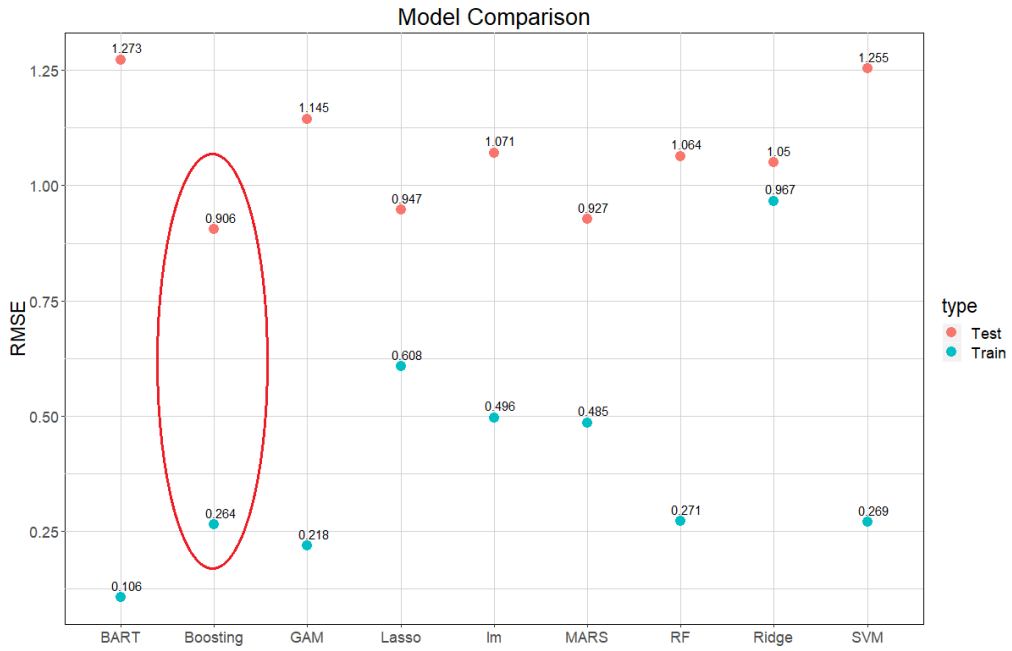

Figure 11: Comparison between all fitted models

5 Inferences

We computed the relative influence of each variable in the final fitted gbm model. Based on the plot (Figure 12), variables Percentage of uneducated people, Death Rate, Fertile soil percentage, Percentage of people practicing Islam, Life expectancy at birth seem to have most influence on the final model. Other variables influencing the model are in line with our initial exploratory data analysis.

Now to further understand the effect of these predictors, their individual marginal influence plots were studied (Figure 13). Looking at the marginal influence plot of Percentage of Uneducated People, it can be concluded that it has a strong positive influence on population growth rate which again make sense as uneducated people might not think towards family planning and can have a large family size leading to high population growth.

Also, if we look at our next important predictor i.e. Death Rate, it can be concluded the population growth rate is inversely proportional to death rate for the range 5 to 10. Hence, the countries requiring increase in population growth rate, need to work on controlling death rates i.e. enhancing safety of their citizens through various measures.

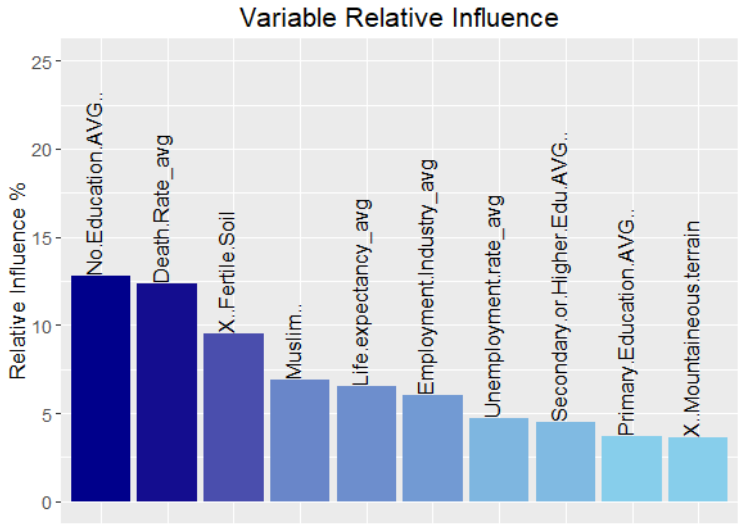

Figure 12: Variable Importance Plot for Boosting

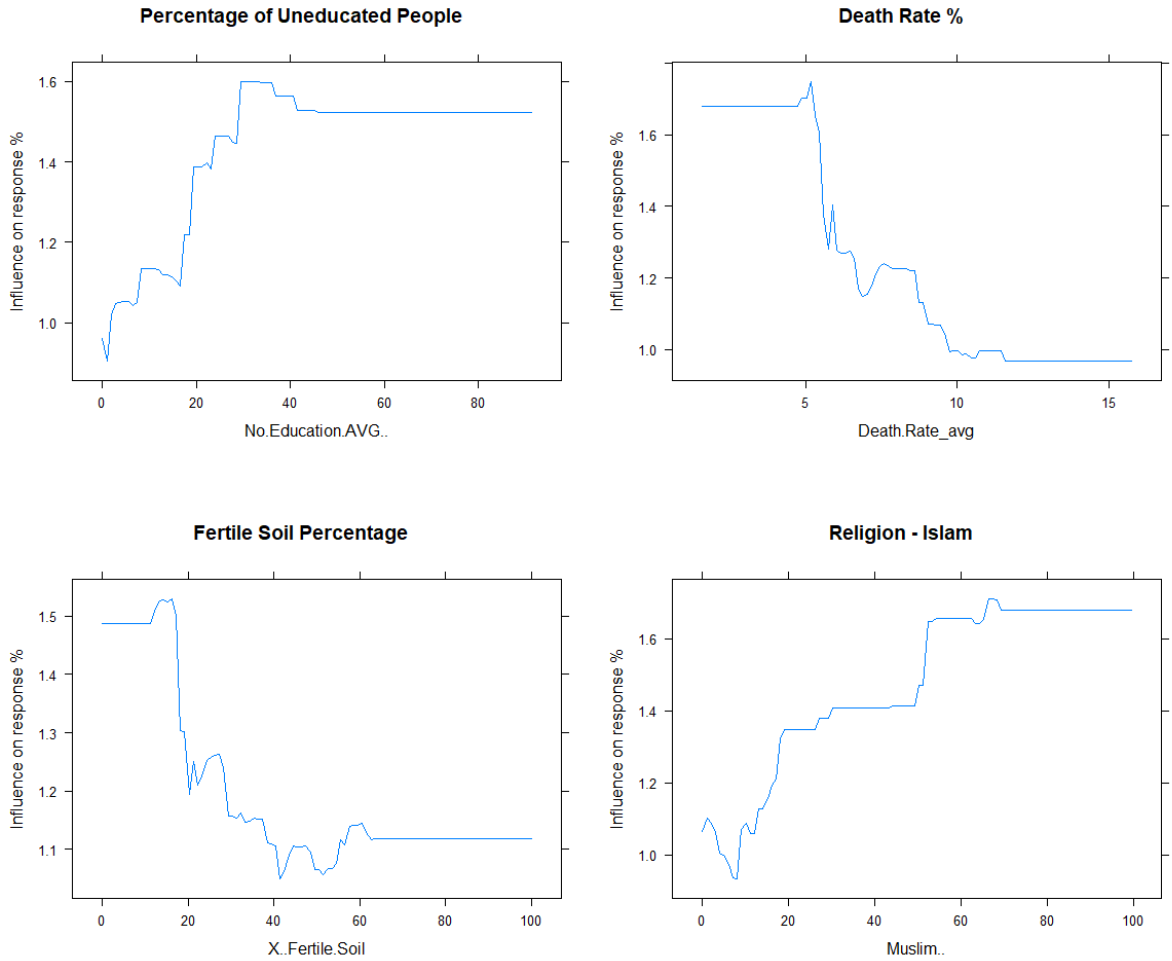

Figure 13: Marginal effect plots of important predictors

Looking the marginal effect plots of the fertile soil percentage, we see that higher this percent in the range of 10 to 65 percent, lesser the growth rate. One of the possible reasons for this trend might be that urban areas that have high concentration of populations don't have fertile soil in general.

Observing the plot of the religion Islam, it is observed that the population growth rate is directly proportional to the number of people following the religion. This can be attributed the fact that in Islamic culture it is common to have larger families.

From Figure 14, looking at the marginal influence plot of life expectancy at birth, we observe that higher the life expectancy lower the population growth rate. This goes against our intuition, but the reason might be that countries with higher life expectancy are usually developed countries with more awareness of family planning and higher level of education.

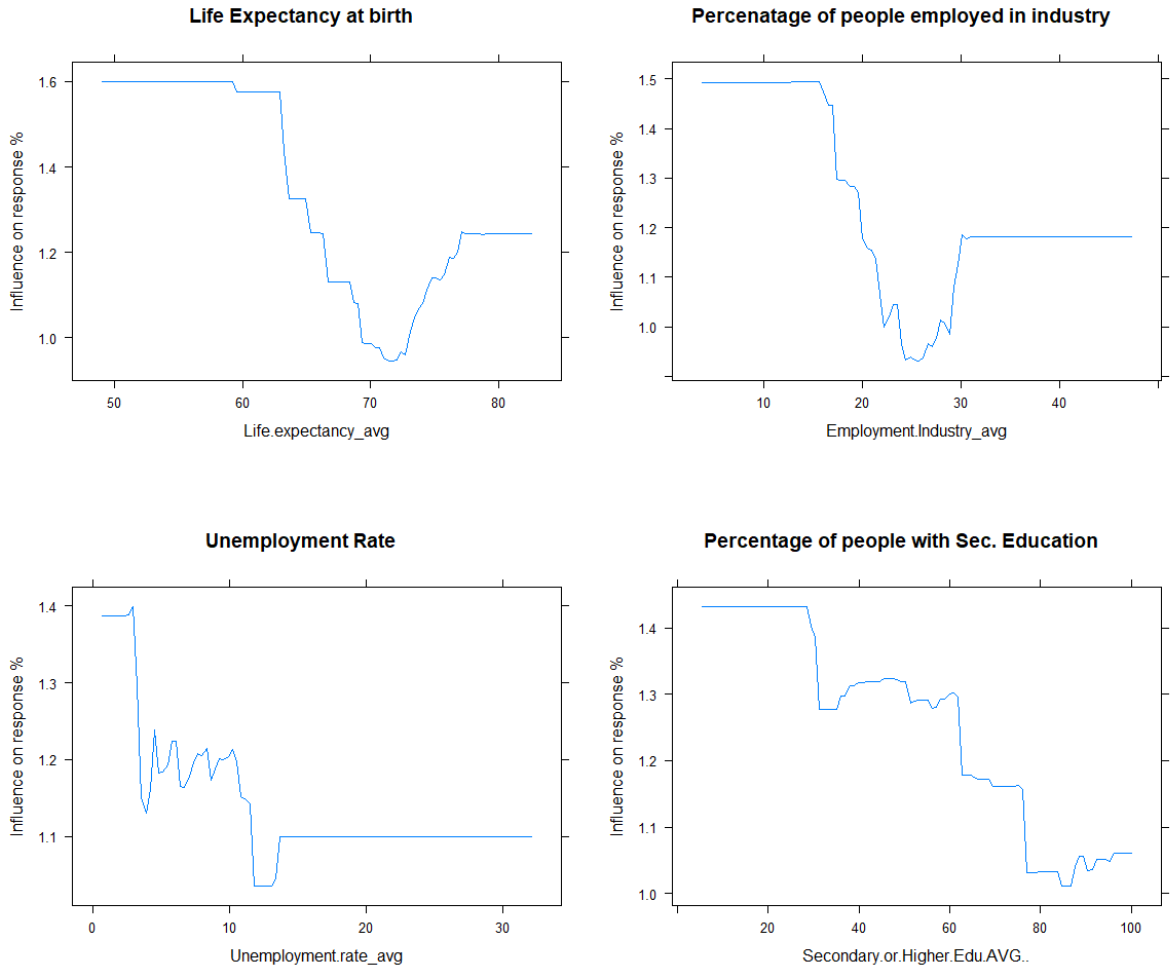

Figure 14: Marginal effect plots of important predictors

In the plot of unemployment rate, we observe that the population growth rate is inversely proportional to the unemployment rate, might be due to financial cost of raising children.

From the marginal influence plot of the people employed in industry sector, it is observed that population growth rate is inversely proportional to this predictor. This is in contrast to the people employed in agriculture sector where there is need to have more labor.

Lastly from the marginal influence plot of percentage of the people with secondary education, we observe that the population growth rate is inversely proportional. This might be due to the fact highly educated people will be more aware of family planning.

Some of the region specific inferences are tabulated below.

Region	Some Countries	Important Predictors
Africa	Uganda, Burundi, Rwanda, Nigeria	• Infant Mortality Rate • Percentage of uneducated people • Life expectancy at birth
Middle-East	Kuwait, Qatar, Bahrain	• Death Rate • Religion • Fertile Soil Percentage • Unemployment Rate
South Asia	India, Pakistan, Bangladesh	• Percentage of uneducated people • Religion

Table 2: Inferences based on Region.

6 References

[1] Nunn, and Puga. 2012. Ruggedness: The blessing of bad geography in Africa. *Review of Economics and Statistics* 94 (1): 20-36.

[2] Kottek, Markus, Jrgen Grieser, Christoph Beck, Bruno Rudolf, and Franz Rubel. 2006. World map of the Kppen-Geiger climate classification updated. *Meteorologische Zeitschrift* 15(3): 259-263.

[3] Riley, Shawn J., Stephen D. DeGloria, and Robert Elliot. 1999. A terrain ruggedness index that quantifies topographic heterogeneity. *Intermountain Journal of Sciences* 5(1-4): 23-27.

[4] Cusack PTE (2017) Physical Economics and Optimum Population Density. *J Glob Econ* 5: 244. doi:10.4172/2375-4389.1000244

[5] Litman, T. Determining Optimal Urban Expansion, Population and Vehicle Density, and Housing Types for Rapidly Growing Cities.*In Proceedings of the World Conference on Transport Research*, Shanghai, China, 1015 July 2016.

[6] Kelley, A. and R. Schmidt, (1995), Aggregate Population and Economic Growth Correlations: The Role of the Components of Demographic Change, *Demography*, 32, 543-555

[7] D. D. Zhang, P. Brecke, H. F. Lee, Y. Q. He, J. Zhang, Global climate change, war, and population decline in recent human history. *Proc. Natl. Acad. Sci. U.S.A.* 104, 1921419219 (2007). doi: 10.1073/pnas.0703073104; pmid: 18048343

7 Appendix

7.1 Data Description

Response	Description	Source
Population Increase Rate	Indicates annual rate of increase of population in percentage	UN Data (http://data.un.org/)
Population Density	Indicates the number of people living in per square km area	UN Data (http://data.un.org/)

Table 3: Details about response variables.

Predictors	Description	Source	Rationale
Birth Rate	Number of live births occurring during the year, per 1,000 population estimated at midyear	UN Data	A direct effect on the response variable
Death Rate	Number of deaths occurring during the year, per 1,000 population estimated at midyear	UN Data	A direct effect on the response variable
Fertility Rate	Total number of children that would be born to each woman if she were to live to the end of her child-bearing years	UN Data	Fertility rate gives a measure of the potential of population to bear children
Infant Mortality Rate	Number of deaths for both sexes per 1,000 live births of children under one year of age	UN Data	This should have an inverse relation to the response variables
Life Expectancy at birth	Indicates in years how long, on average, a newborn can expect to live, if current death rates do not change	UN Data	This is a more specific variable that will also be inversely proportional to the response variable
Education levels	3 variables indicating number of students enrolled in primary, secondary, tertiary education levels in thousands respectively	UN Data	Studies have shown educating women is effective in controlling population. Education also gives opportunities to people to have greater mobility, thus influencing population
GDP per-capita	Indicates gross domestic product per-capita in US dollars	UN Data	Economic status of person would affect the ability to raise children. It can also have an adverse effect especially in developing countries

Employment by industry	3 variables indicating percentage of people employed in agriculture, industry and services respectively	UN Data	Income would be different from community-based businesses and agriculture based economies which require more manpower to operate hence affecting response variable
Unemployment Rate	Percentage of unemployed workers in the total labor force	UN Data	This might lead to higher population growth rate with an individual's an attempt to increase the number of earning hands, or it might lead to a decrease in population growth rate if one thinks unemployment doesn't allow them to afford raising children.
CO2 emissions per capita	Indicates the carbon-dioxide emissions per capita in thousand metric tons	UN Data	This would have adverse effects on the health, which in turn might motivate people to move and can influence population density
Fertile soil percentage	Percentage of land surface area of each country that has fertile soil	Nunn, and Puga (2012)	This will influence the GDP of the agriculture-based economies and the growth
Proportion of desert land	Percentage of land surface area of each country covered by sandy desert, dunes, rocky or lava flows	Nunn, and Puga (2012)	The response variables will be inversely proportional to this variable
Proportion of region with tropical climate	Percentage of the land surface area of each country that has any of the four Kppen-Geiger tropical climates as defined by Kottek et al. (2006)	Nunn, and Puga (2012)	People tend to prefer living in the tropical climate. In addition to population density the population growth also seems to be directly correlated to this
Proportion of coastal land	Percentage of the land surface area of each country that is within 100km of the nearest ice-free coast	Nunn, and Puga (2012)	Coastal cities tend to have more temperate climates. This attracts people to move to the coastal regions
Proportion of rugged terrain	Percentage of a country's land area that is highly rugged	Nunn, and Puga (2012)	This would have an inverse relation to the population growth rate

Table 4: Details about predictor variables Source for UN Data : http://data.un.org/

7.2 Summary Statistics

Factors	Min.	1st Qu	Median	Mean	3rd Qu	Max.
GDP	305	1992	4892	12977	16443	91765
Emission	0	1.1	3.95	5.759	7.175	55.9
No.Education	0	1.589	13.606	23.088	38.515	91.294
Primary.Education	0	6.075	13.502	15.255	21.597	47.663
Secondary.or.Higher.Education	5.51	38.43	61.21	61.03	87.28	100
Unknown.Education	0	0	0.2306	0.975	0.9249	10.2425
Employment - Agriculture	0.9	4.9	17.35	23.7	37.27	86.6
Employment - Industry_avg	3.8	16.73	20.45	21.23	26.73	47.25
Employment - Services_avg	7.6	44.85	58.2	55.08	68.58	83.7
Infant.Mortality.rate_avg	2.05	6.55	17.5	26.65	37.62	101
Life.expectancy_avg	49.05	67.22	73	70.42	75.97	81.45
Population.Rate	-1.15	0.5	1.25	1.37	2.1	7.1
Unemployment.rate_avg	0.7	4.525	6.85	8.017	10.075	32.1
Mountainous%	0	1.609	9.723	17	25.628	90.249
Soil	0	19.87	41.6	39.63	56.78	100
Desert	0	0	0	2.7668	0.0225	39.21
Tropical	0	0	0	36.16	95.69	100
Coastal%	0	7.778	25.378	42.214	95.367	100
Death.Rate	1.641	5.659	7.495	8.229	10.125	15.75
Pop.Density	-0.97667	-0.54833	0.05333	1.55737	0.926	94.48867
Christian%	0.2	16.35	70.4	57.57	88.55	99.5
Muslim%	0	0.2	4.1	22.84	29.95	99.5
Atheist%	0	1.2	4.4	9.837	14.5	59.6
Hindu%	0	0	0	2.913	0.2	80.7
Buddhist%	0	0	0	3.996	0.3	96.9
Folk.Religion%	0	0	0.2	2.048	1.45	45.3
Other.Religion%	0	0	0.1	0.3823	0.3	9.7
Jewish%	0	0	0	0.7375	0.08	75.6

Figure 15: Summary statistics of input data

7.3 More Exploratory Data Analysis

Based on the density plot, the response seems to be fairly normally distributed except at the right tail.

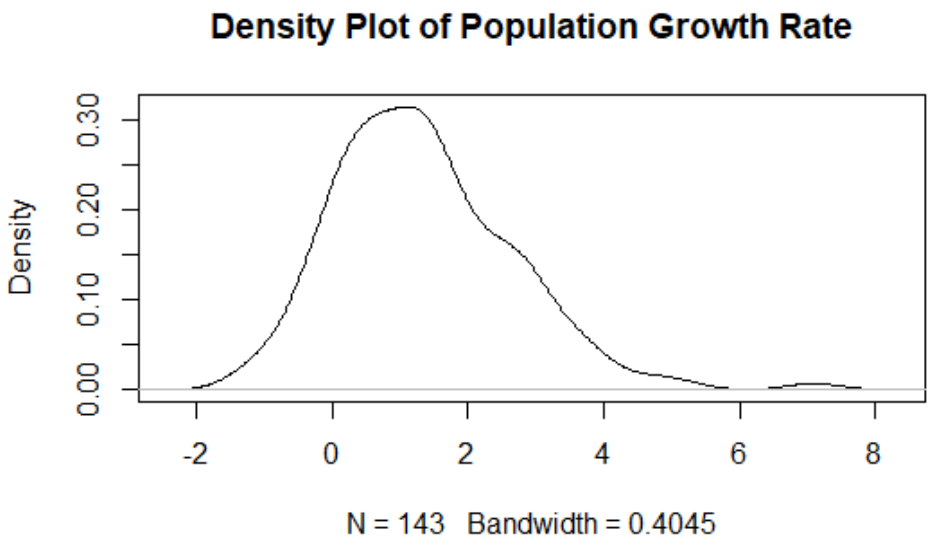

Figure 16: Density Plot of Population Growth Rate

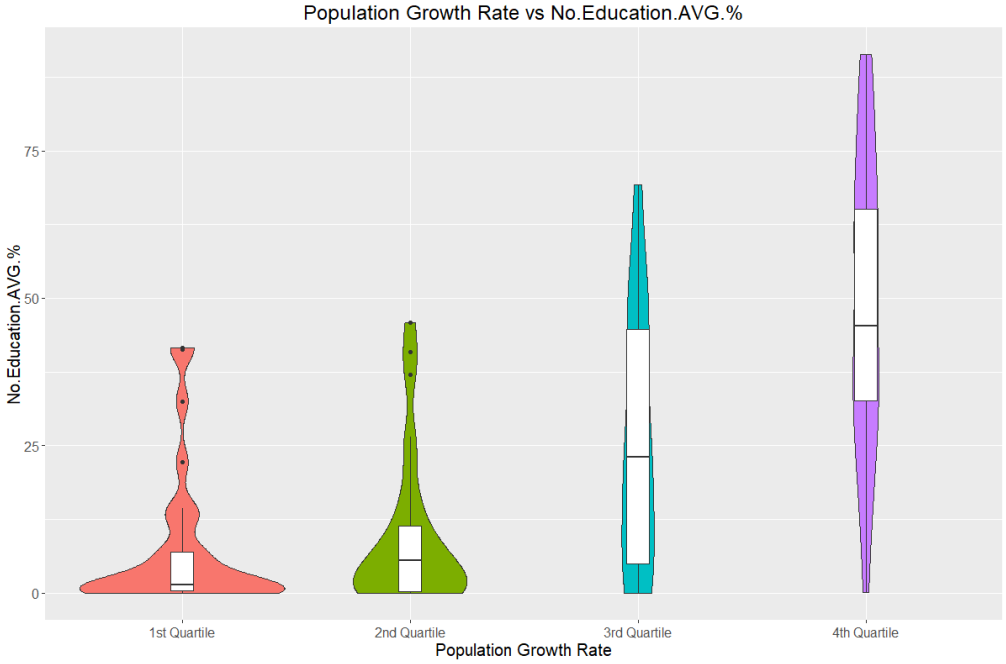

Figure 17: Violin Plot of Percentage of uneducated people

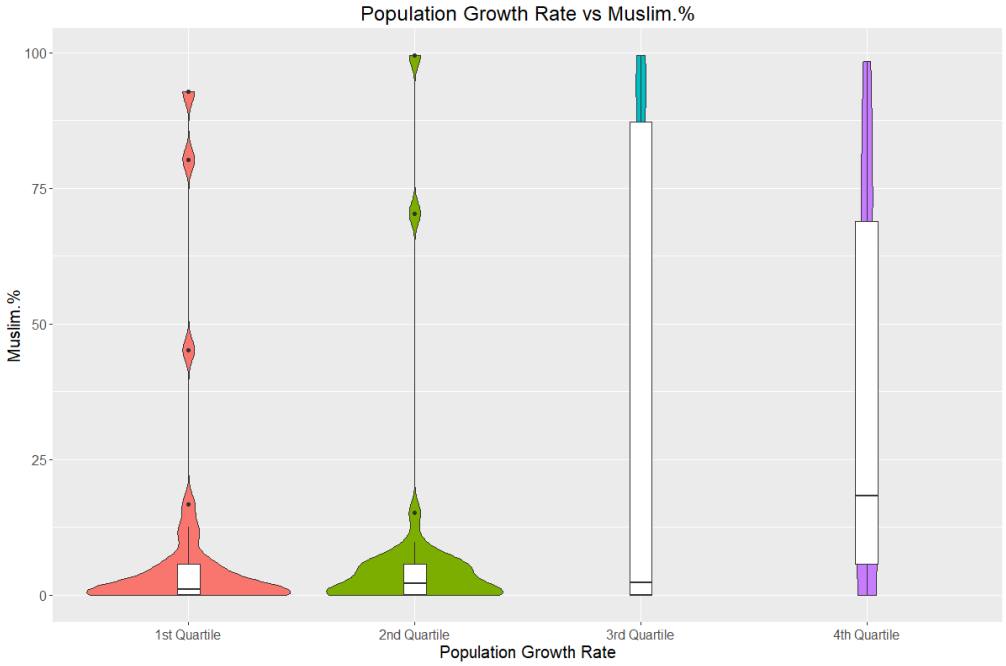

Figure 18: Violin Plot of Religion - Muslim

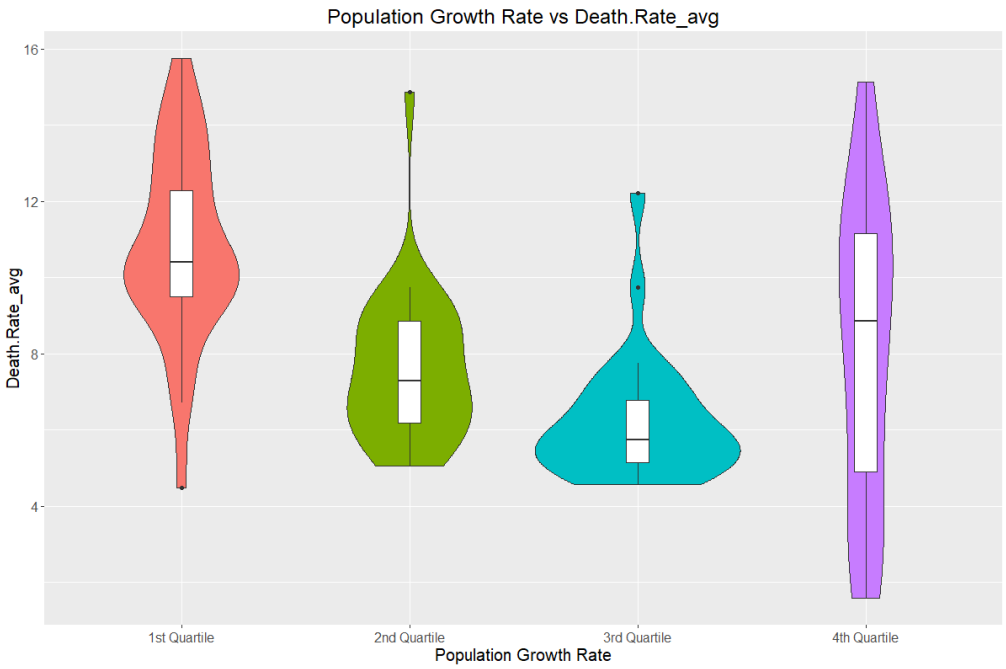

Figure 19: Violin Plot of Death Rate

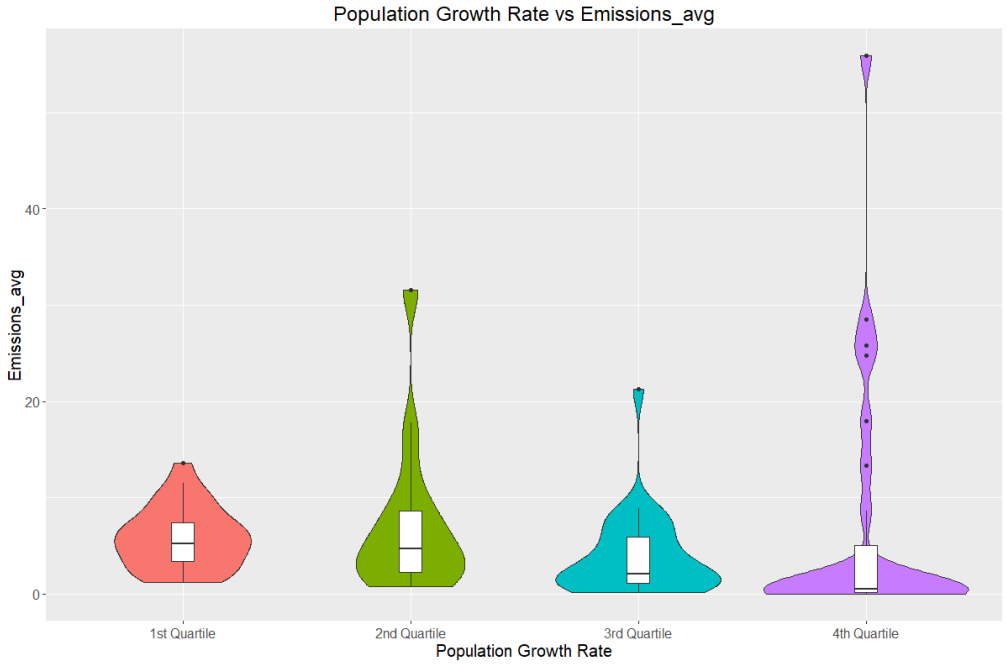

Figure 20: Violin Plot of Emissions

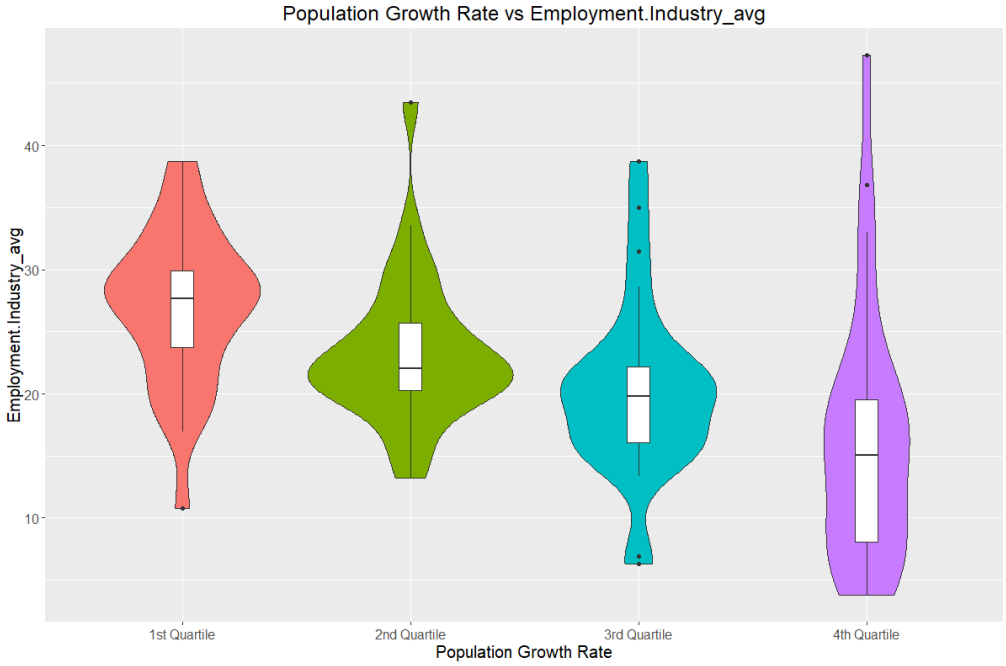

Figure 21: Violin Plot of Employment in industry sector

Figure 22: Violin Plot of Employment in service sector

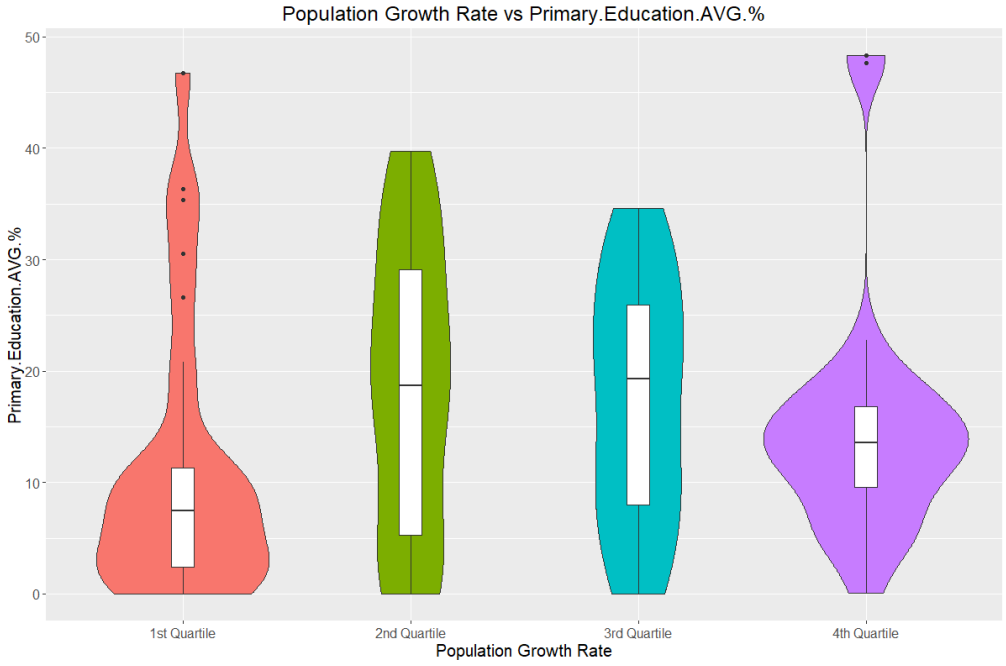

Figure 23: Violin Plot of Percentage of people educated upto primary level

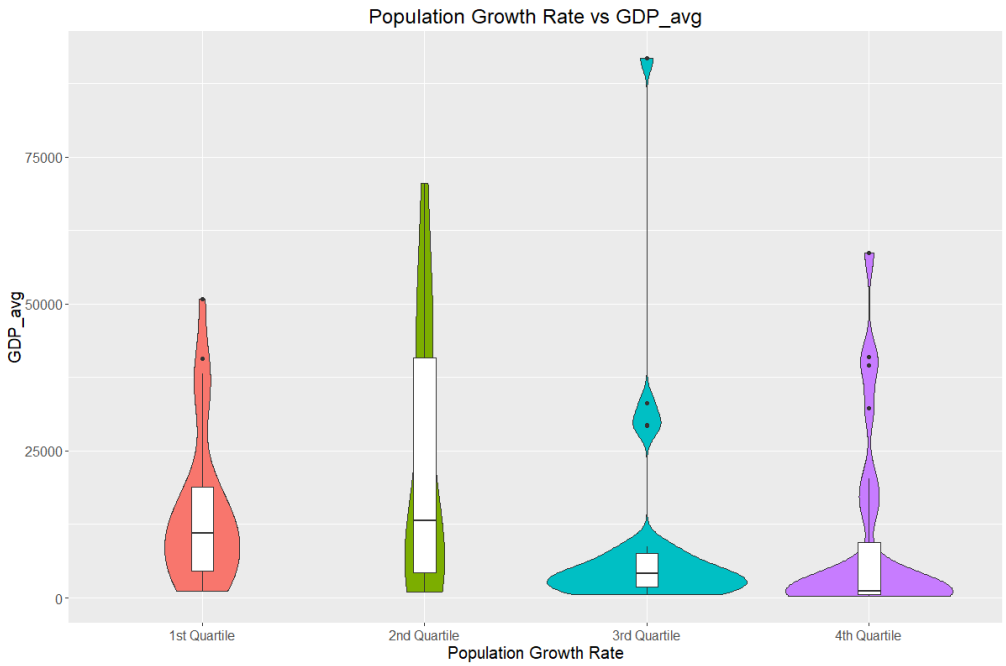

Figure 24: Violin Plot of Gross Domestic Product

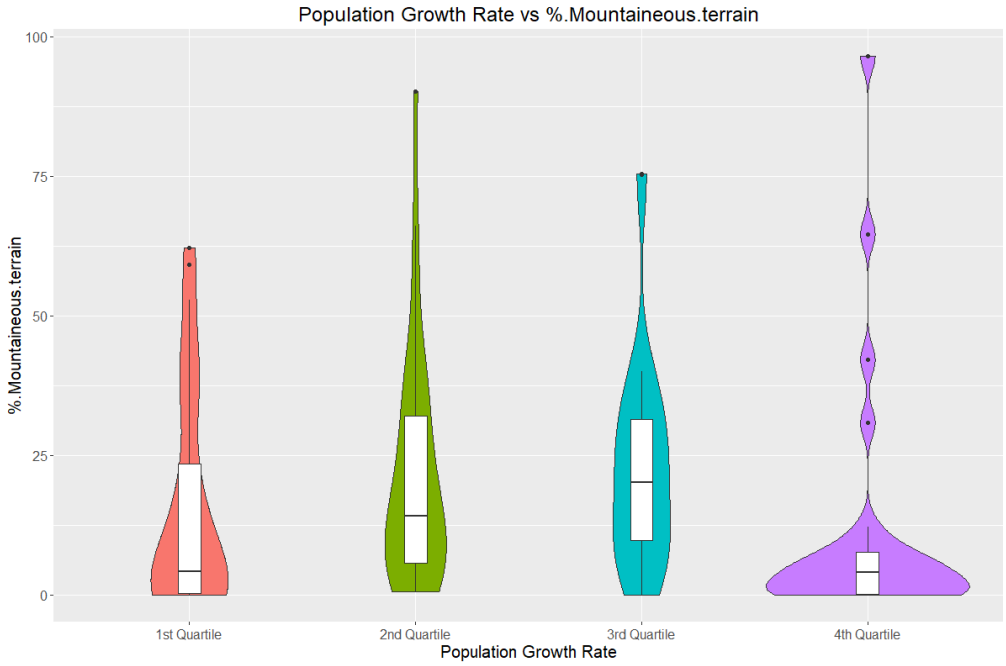

Figure 25: Violin Plot of Proportion of rugged terrain

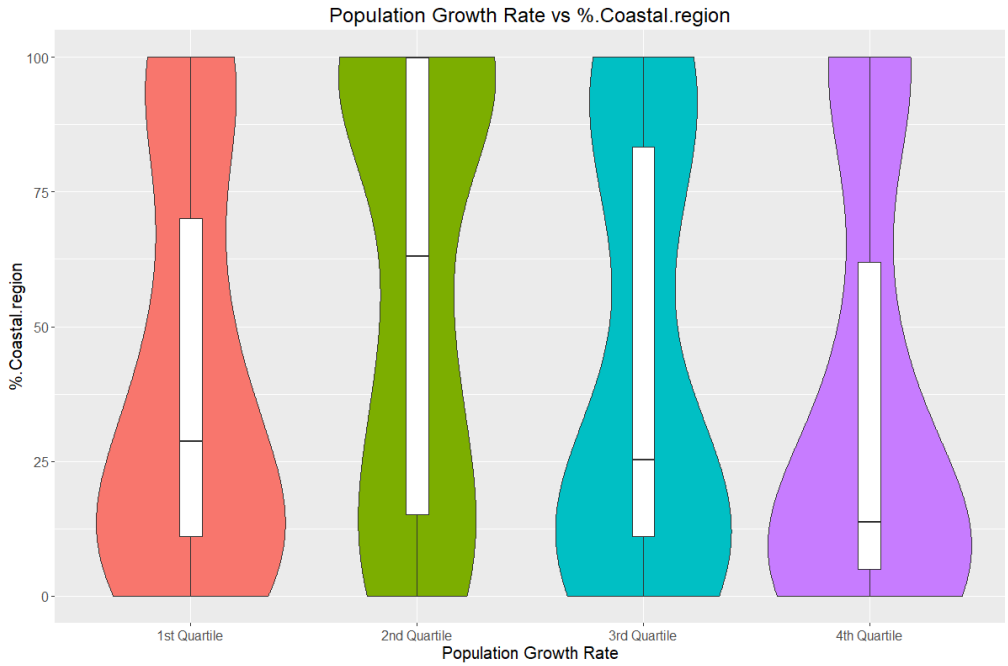

Figure 26: Violin Plot of Proportion of coastal land

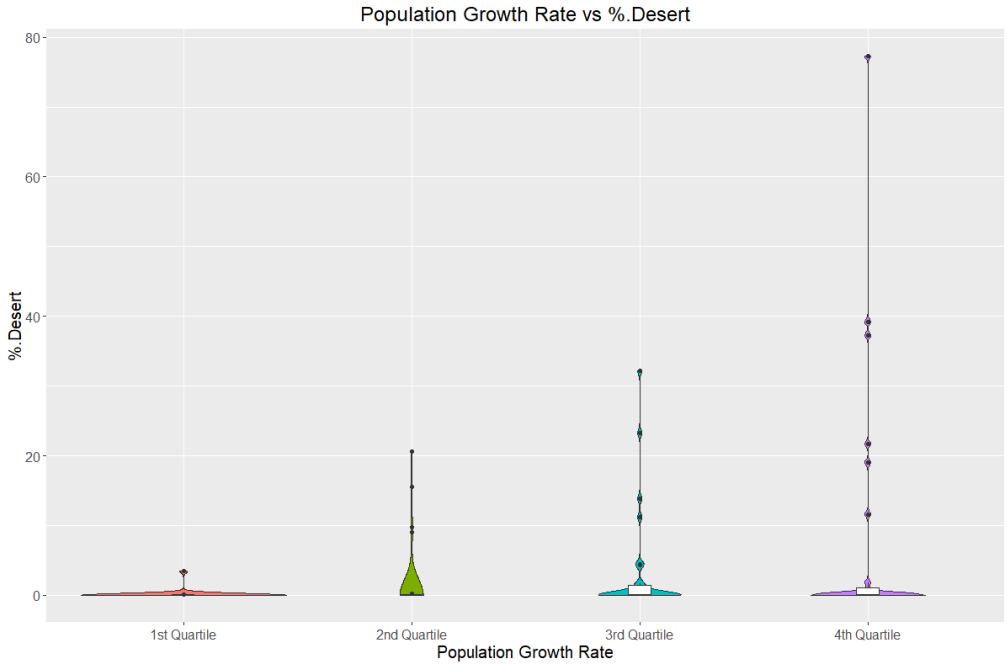

Figure 27: Violin Plot of Proportion of desert land

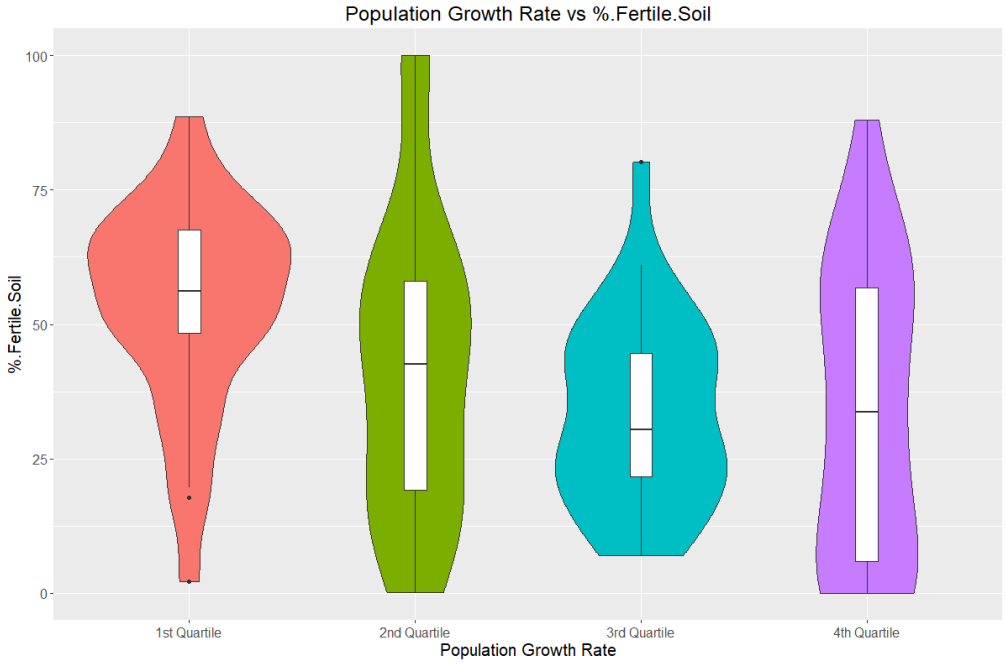

Figure 28: Violin Plot of Fertile Soil Percentage

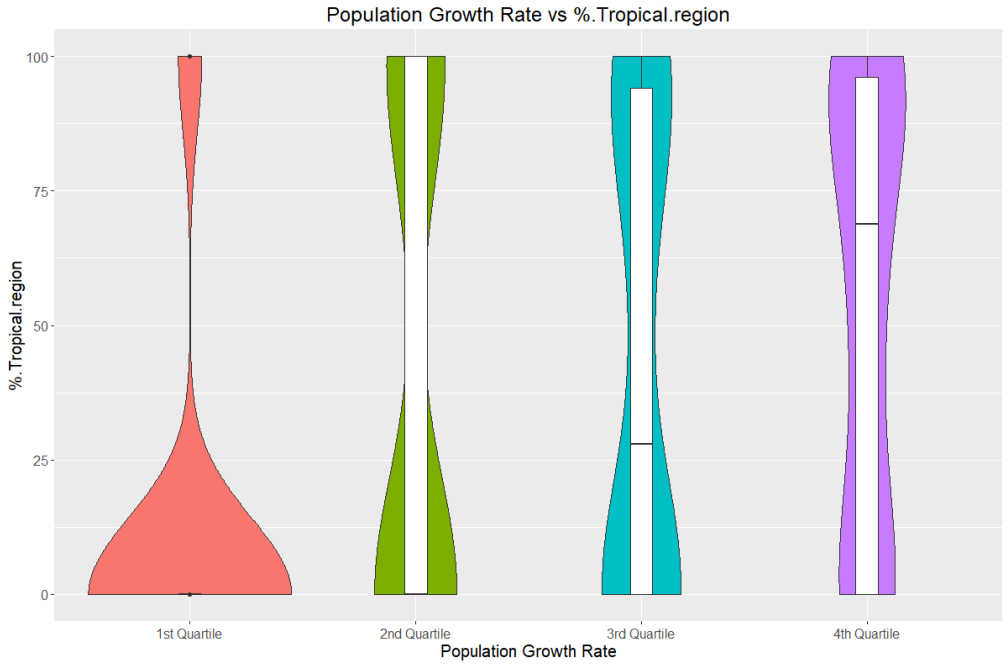

Figure 29: Violin Plot of Proportion of region with tropical climate

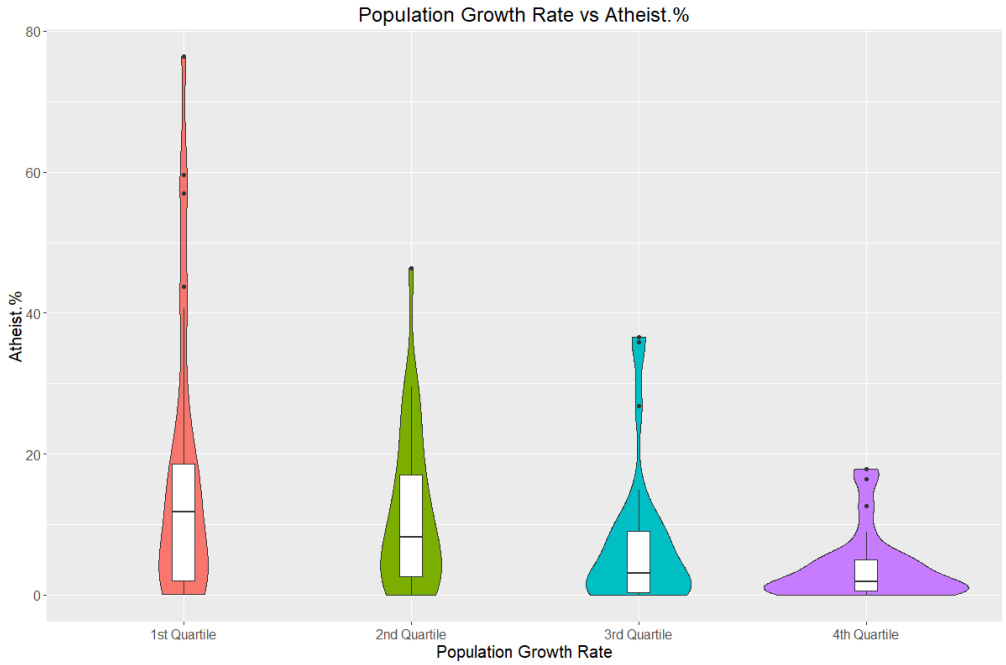

Figure 30: Violin Plot of Religion - Atheist

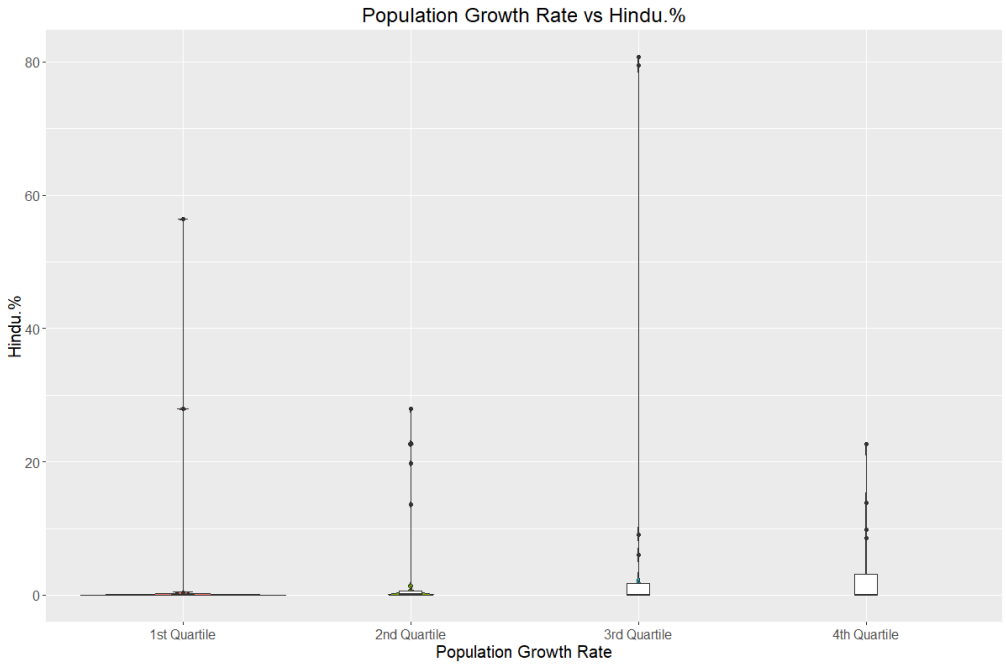

Figure 31: Violin Plot of Religion - Hindu

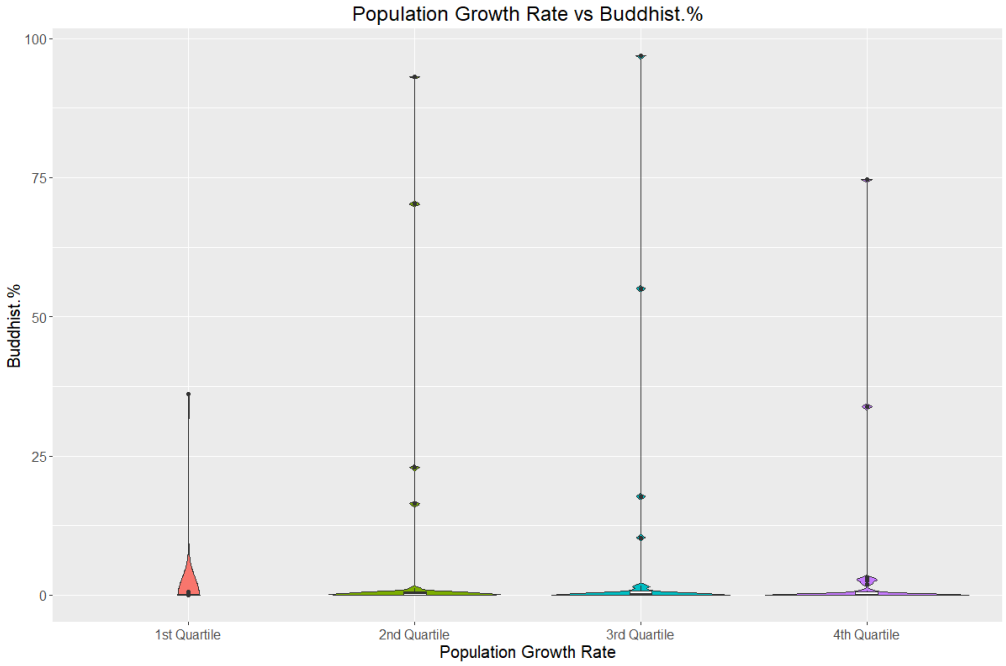

Figure 32: Violin Plot of Religion - Buddhist

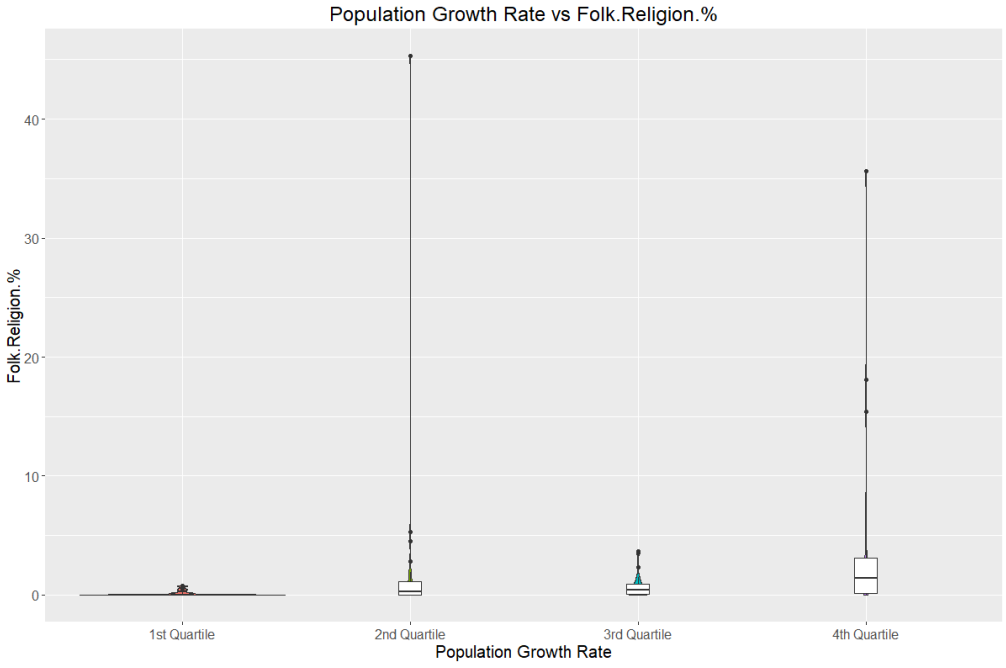

Figure 33: Violin Plot of Religion - Folk Religion

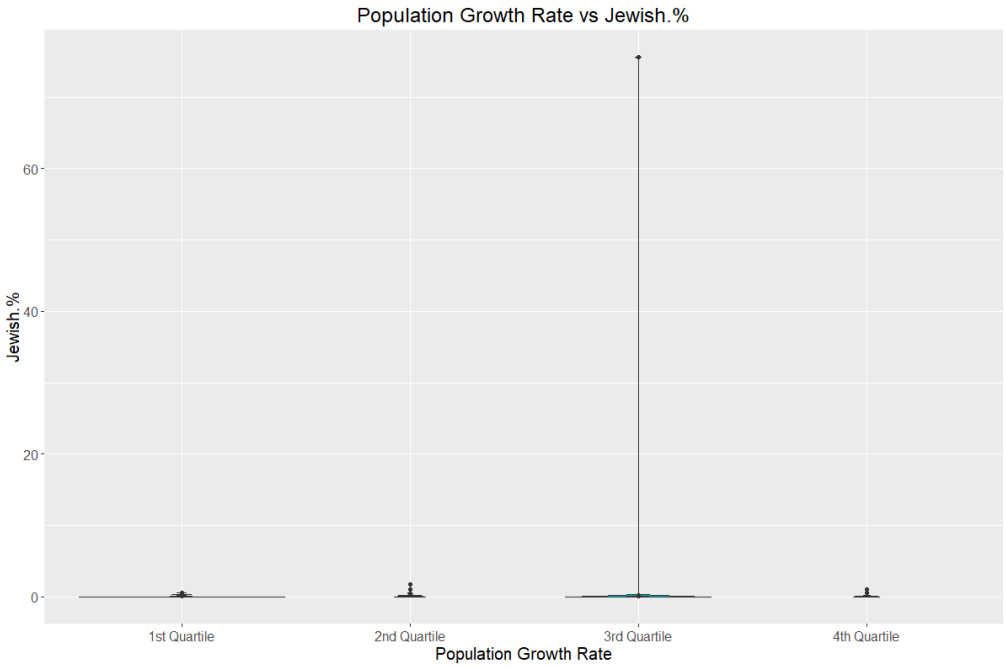

Figure 34: Violin Plot of Religion - Jewish

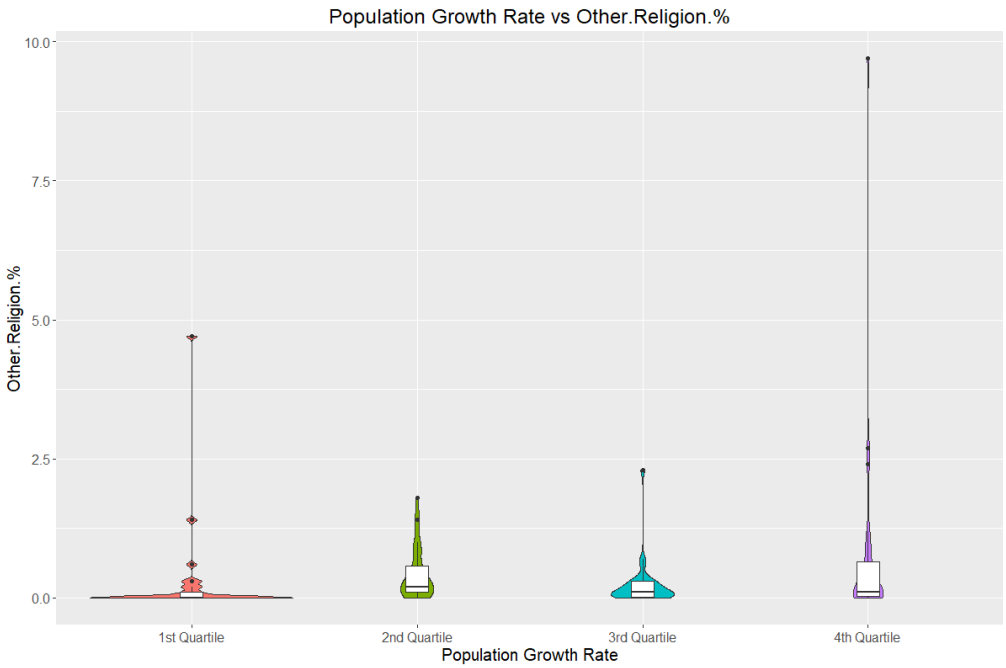

Figure 35: Violin Plot of Religion - Other Religion

Figure 36: Violin Plot of Unemployment Rate

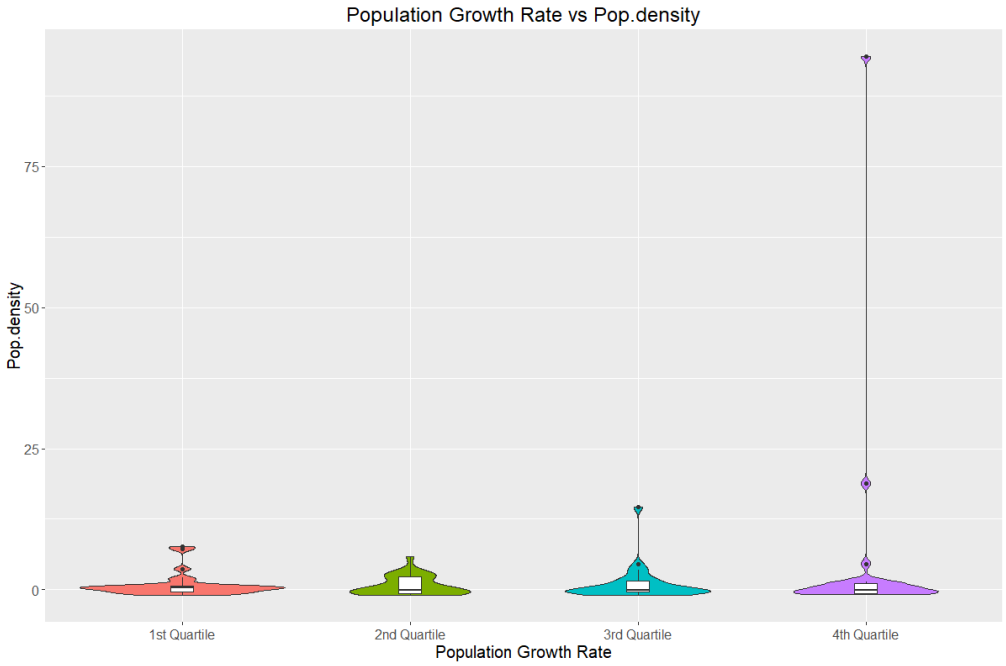

Figure 37: Violin Plot of Population Density

7.4 Linear Model Assumptions

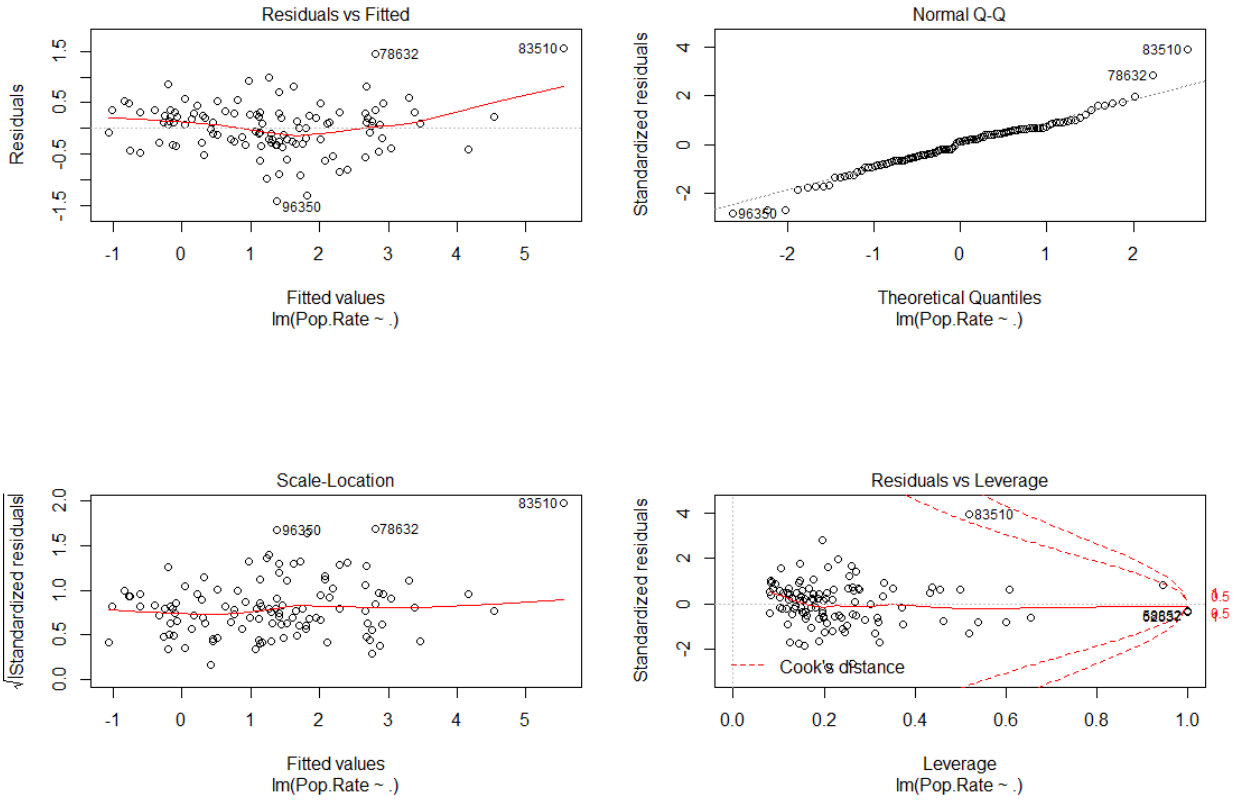

Figure 38: Linear Model Diagnostic Plots

7.5 Gradient Boosting Plots

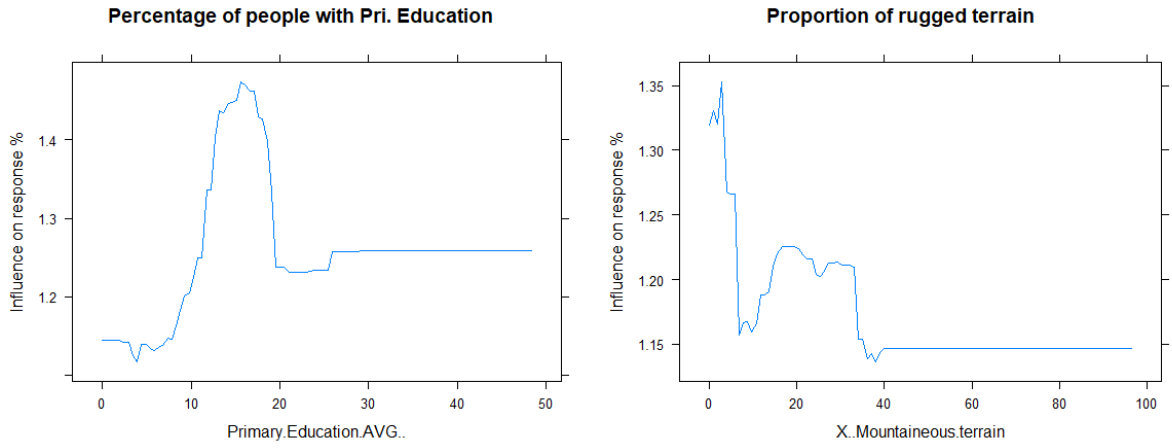

Figure 39: Marginal effect plots of other important predictors