Analysis of Global Life Expectancy and Related Factors

Echo Chen, Andrew Kroening, Pooja Kabber, Dingkun Yang

November 23rd, 2022

Report: Your report will be an 8-10 page self-contained document describing your analysis. It should be written as a professional document that can be understood by someone with limited statistics background (e.g., a client). You are also required to submit an RMD file that includes your code for the EDA and analysis. The report should be organized as follows:

intro - 2 pages eda - 1 page question 1 - 3 pages question 2 - 3 pages conclusion - 1 page

Abstract

This analysis is conducted to understand which factors influence life expectancy and the development status of countries around the world. The dataset used for this research consists of national disease, economic, and social factors and was compiled by the World Health Organization (WHO). [Need to link to the dataset from Kaggle] To conduct the analysis, two research objectives are formulated: one prediction question and one inference question. Unique models are fit to each approach and the results are analyzed for utility. [Add a concluding punchline]

Introduction

Provide more background on the data and research questions. Be sure to cite the data and background information appropriately (APA style is fine)

This analysis uses data from the WHO to better understand the drivers of life expectancy around the globe. We also attempt to find inferential value from the data for determining the developmental status of a given country. From the two research questions we aim to improve insights into factors that drive a country's developmental status, and the population health indicators that lead to improved life expectancy.

The particular dataset for this analysis contains national-level observations of variables related to life expectancy around the globe for a period spanning the early portion of the 21st century. The complete dataset includes observations beginning in the year 2000 and ending in the year 2015. As a full dataset, there are 2,938 observations for 22 variables. Practically, each country has approximately one observation each year, averaging 183 for each of the 16 years encompassed by the data. The dataset effectively contains 20 variables for each country and year combination, covering significant disease, economic, and social factors. Below are the questions we aim to answer in this analysis:

(Cite data here)

Question #1 (Prediction)

"How did major disease, economic, and social factors impact life expectancy around the globe in 2014?"

Question #2 (Inference)

"How did disease and mortality rates, along with national economic factors, contribute to a country's development status in 2014?"

Data

[Data introduction: address missing, factors, cleanliness, subsetting, and eda]

During the course of this initial rouund of EDA, the team identified a number of missing values from the dataset. These missing values included: 41x Population, 10x Hepatitis B, and 10x Schooling observations, with the large number of missing population values the most concerning. The team considered several approaches for mitigating the problems posed by this data, as it precludes a number of potentially influential countries from being included in this analysis. We considered multiple imputation as well as scholastic imputation for ways to mitigate these issues.

An alternative approach for missing data was identified as theoretically feasible early in the analysis process. Because of the national-level of our dataset, it is possible that we could find suitable replacement values from another source with high integrity in these areas, such as the World Bank, the International Monetary Fund, or the CIA World Factbook. While those sources had existing data for some of our missing values, we opt to not use them for replacement. Most replacement candidates we found did not match the surrounding data points (i.e. GDP figures for the country/year in question were not close enough to consider a match), and thus we have low-confidence that those values would be consistent with the WHO's data collection methods.

After the initial treatments to factor variables, our dataset is reduced to complete cases only and subset for the two years of interest in our analysis. We ultimately decide to preserve the original integrity of the data and bypass any available imputation methods. While there are certainly options available, the team assesses that the potential gain from the inclusion of the additional countries does not offset the possible bias or skew introduced from imputation. At the conclusion of these steps and decisions, we subset the data to make two sub-datasets: one for the year 2014 and one for the year 2013 which we will use in our second research question. These two datasets are nearly ideintical in size, with the 2014 dataset consisting of 131 observations, and the 2013 dataset having 130.

Methods

Describe the process you used to conduct analysis. This includes EDA and any relevant data cleaning information (e.g., did you exclude missing values? If so, how many? Did you collapse categories for any variables?) Then describe the models you fit, and any changes you made to improve model fit (e.g., did you exclude any influential points? Did you do have to address multicollinearity issues? Did you transform any variables?). Also describe model diagnostics. The organization of this section may depend on your particular dataset/analysis, but you may want to break it into subsections such as "Data," "Models," and "Model assessment." Note that you do not present any results in this section.

(The general methodology of our analysis centered around the ability to draw insights from the dataset without significant transformations or imputations. - do we need this?) To analyse this data / accomplish this objective, the team began with a priori variable selection, after which we conducted exploratory data analysis (EDA) to examine the distributions of key variables.

(A Priori variable selection) These are the variables we selected as part of our initial a priori variable selection which included selecting the variables whose relationship with Life expectancy we were interested in: Status + Population + Life.expectancy + Measles + Polio + HIV.AIDS + GDP + Income.composition.of.resources + BMI + Total.expenditure + percentage.expenditure + Schooling

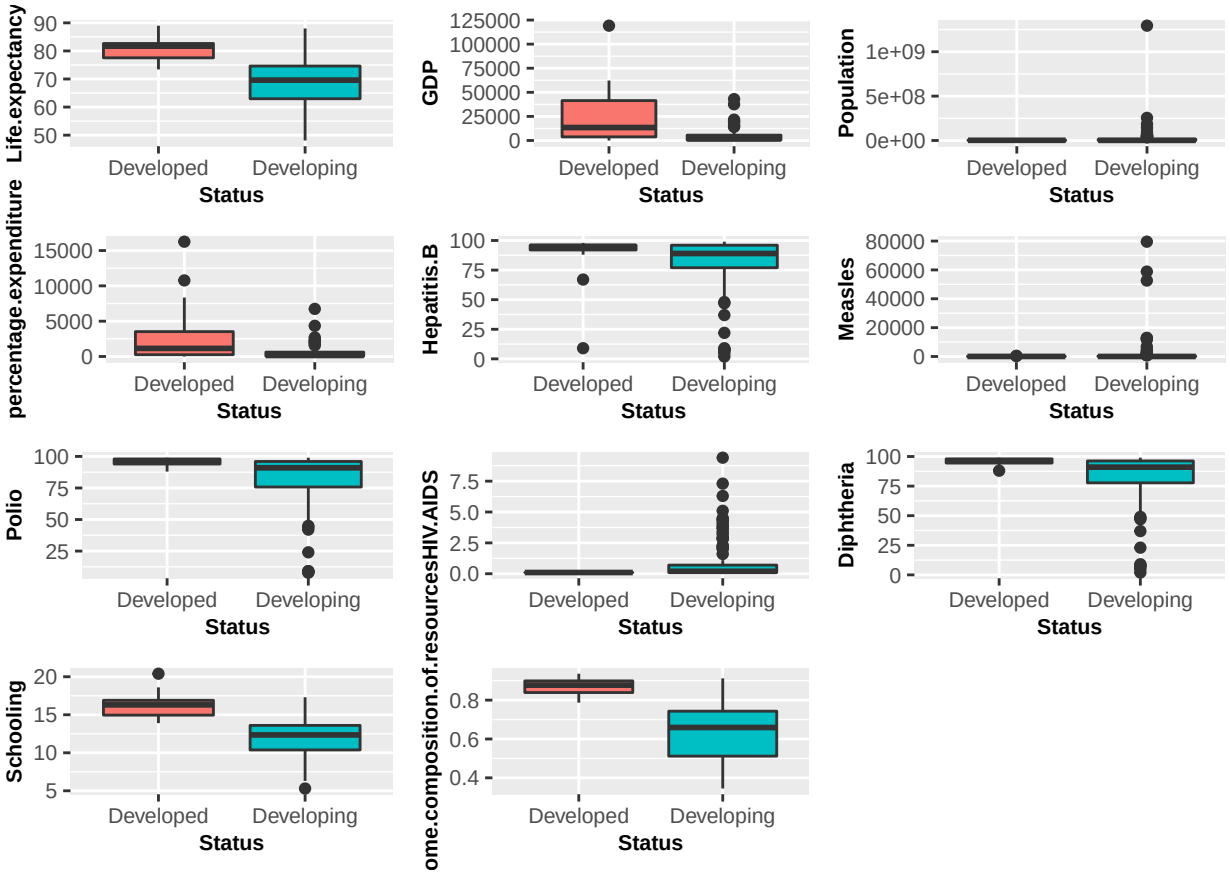
(Dealing with missing data) From the EDA, missing data points were identified. Since imputing the missing values using mean or regression methods is not valid in this context where each value is for a country and using these methods may create an estimate significantly distant from the true value, we considered finding the missing data from other sources. We tried sources like (). Here we discovered that the data points for factors we had differed in these alternate datasets. Since the data collection time and methods may have been different for these sources, we did not use this method to impute the missing data. In the end we decided to drop the 52 rows with null values. At conclusion of data preparation, we subset for the year 2014, our focal point for this analysis, and the year 2013 as a testing validation dataset used in our second research question.

(Correlation and multicollinearity) While conducting EDA, we found that a few of the variables in our data were highly correlated. To analyse this correlation, we used a correlation plot and VIF. We found that these variables were highly correlated:

1. Infant deaths and Under five deaths
2. Percentage expenditure and GDP
3. Hepatitis B and Diphtheria
4. Thinness variables ()
5. Schooling and Income composition of resources

Since percentage expenditure is highly correlated with GDP and schooling is highly correlated with income composition of resources, we dropped percentage expenditure and schooling from our initial list of apriori variables.

(EDA interpretation)



[Describe Table 1 - feels like we should add more variables]

Models

Question 1

We then proceeded to model fitting. To answer the first research question, we fit a simple linear regression model with life expectancy as our response variable. We fit three models to study the relationship between the independent variables and life expectancy, including only our a priori variables, including all variables in the data and using backward stepwise selection. To find the best fit parsimonious model, we evaluated the models using adjusted R^2 and BIC. In our final model, we included the variables selected by backward stepwise selection and our a priori variables but excluded the variables with high collinearity that we discovered during EDA.

Table 1: Summary of Variables

Statistic	N	Mean	St. Dev.	Min	Max
Population	131	22,269,096.000	116,699,866.000	41.000	1,293,859,294.000
Life.expectancy	131	70.520	8.605	48.100	89.000
percentage.expenditure	131	850.874	2,071.444	0.443	16,255.160
Measles	131	2,042.863	9,842.341	0	79,563
Polio	131	83.496	20.966	8	99
HIV.AIDS	131	0.810	1.562	0.100	9.400
GDP	131	7,256.847	14,741.400	12.277	119,172.700
Schooling	131	12.676	2.750	5.300	20.400
Income.composition.of.resources	131	0.670	0.151	0.345	0.936
BMI	131	40.476	20.734	2.000	77.100
Total.expenditure	131	6.107	2.533	1.210	13.730

Finally, we obtain diagnostic information about the performance of our models.

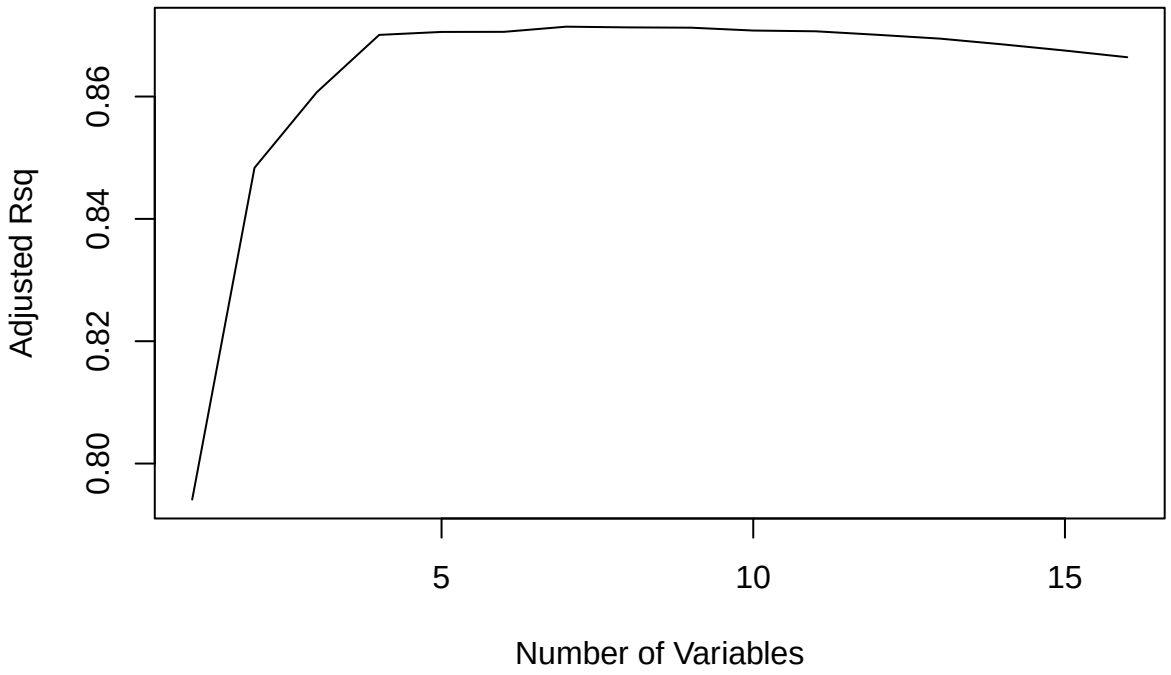
Model selection based on AICc:

```
K   AICc Delta_AICc AICcWt Cum.Wt      LL
```

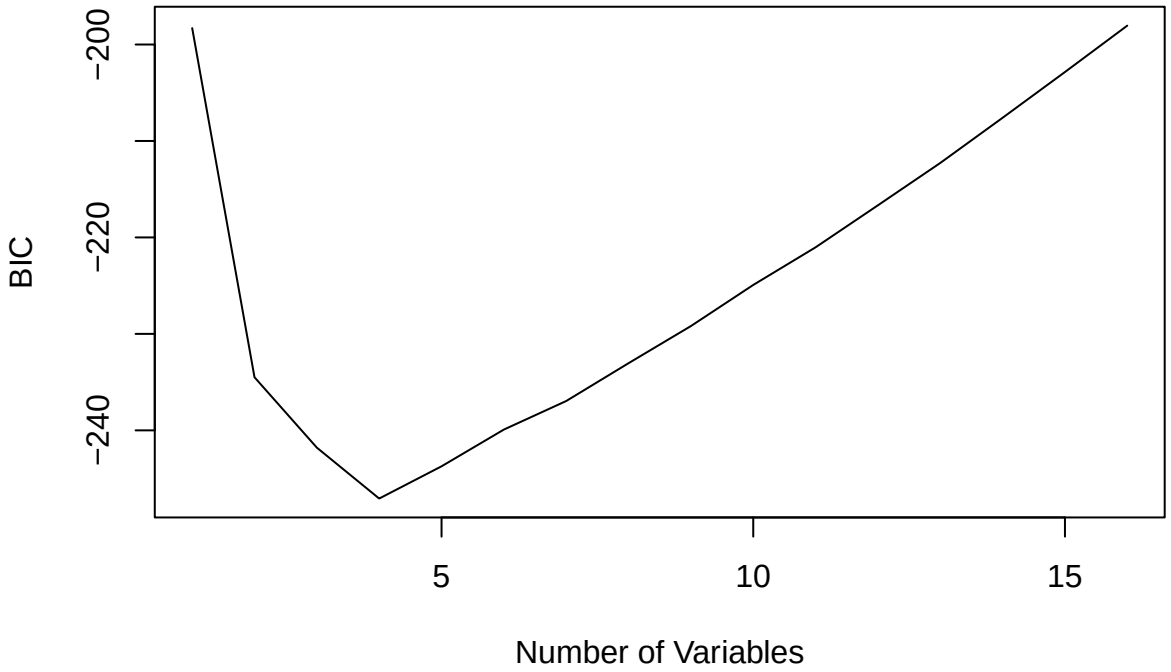
backwardstepwise 8 677.87 0.00 1 1 -330.34 all 21 704.17 26.30 0 1 -326.85 apriori 11 706.28 28.41 0 1 -341.03

% latex table generated in R 4.2.1 by xtable 1.8-4 package % Sun Nov 27 14:22:25 2022

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	49.5586	2.9829	16.61	0.0000
StatusDeveloping	-1.0507	0.9883	-1.06	0.2898
Population	0.0000	0.0000	0.33	0.7416
Measles	-0.0000	0.0000	-0.38	0.7043
Polio	0.0026	0.0148	0.18	0.8609
HIV.AIDS	-0.8647	0.2390	-3.62	0.0004
GDP	0.0000	0.0000	0.21	0.8356
Income.composition.of.resources	34.6152	3.5565	9.73	0.0000
BMI	0.0001	0.0180	0.00	0.9964
Total.expenditure	0.3286	0.1168	2.81	0.0057
Adult.Mortality	-0.0179	0.0040	-4.52	0.0000



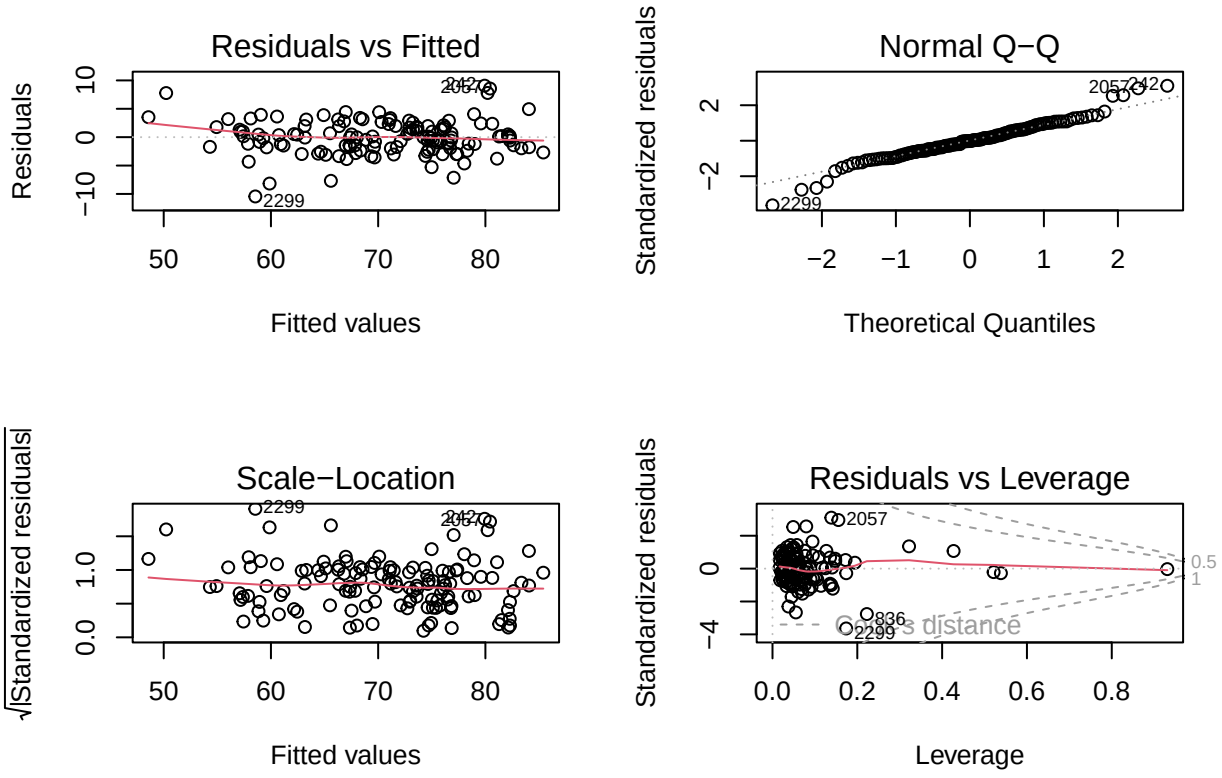

```
## [1] 7
```




```
## [1] 4
```

Model Assessment

To check if the model that we selected follows the linear regression assumptions, we will plot the model summary plots and also interpret the F-statistic in the model results.



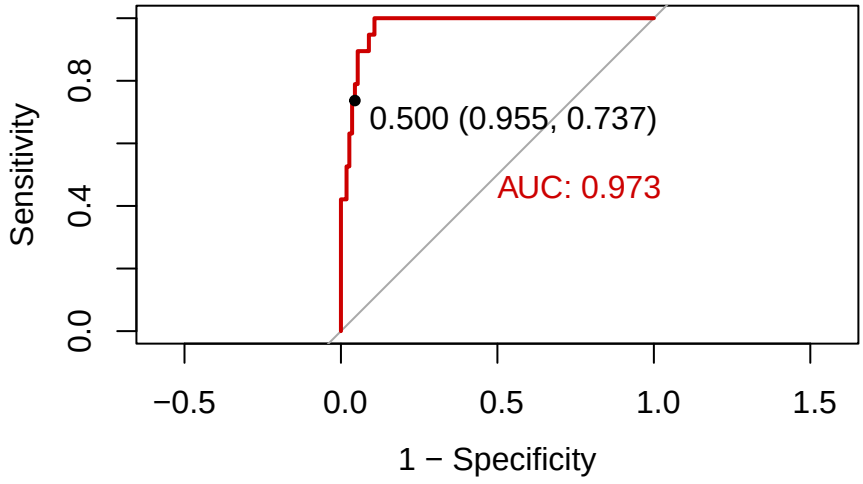
Question 2

[The second question, the inferential one, will be approached with a logistic regression model]

[Dingkun, what is all of this]

[Multicolinearity concerns]

Percentage expenditure and GDP are the ones have higher VIF being over 10, WE NEED TO DECIDE WHICH ONE TO INCLUDE IN THE MODEL



Income composition of resources VIF is near 5, with p-value less than 0.05 (CANNOT DELETE, but maybe delete total expenditure)

Table 2: Logistic Regression Model

	Dependent variable:
	Status_num
Life.expectancy	−0.07 (0.12)
	p = 0.57
percentage.expenditure	−0.0002 (0.0005)
	p = 0.62
BMI	−0.04 (0.03)
	p = 0.14
Total.expenditure	0.22 (0.18)
	p = 0.22
HIV.AIDS	−152.86 (23,893.84)
	p = 1.00
GDP	0.0000 (0.0001)
	p = 0.84
Income.composition.of.resources	41.25 (15.60)
	p = 0.01***
Schooling	−0.06 (0.39)
	p = 0.89
Constant	−12.15 (2,389.40)
	p = 1.00
Observations	131
Log Likelihood	−20.18
Akaike Inf. Crit.	58.36
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3: Variance Inflation Factors

Life.expectancy	percentage.expenditure	BMI	Total.expenditure	HIV.AIDS	GDP	Income.composition.of.resources	Scho
2.04	10.61	1.29	1.20	1.00	10.56	4.23	2.

Table 4: Confusion Matrix for Full Model

	True Developed	True Developing
Predicted Developed	14	5
Predicted Developing	5	107

Table 5: Logistic Regression Model

	Dependent variable:
	Status_num
Life.expectancy	−0.07 (0.12)
	p = 0.53
BMI	−0.04 (0.03)
	p = 0.14
Total.expenditure	0.20 (0.17)
	p = 0.24
HIV.AIDS	−153.11 (23,950.77)
	p = 1.00
GDP	−0.0000 (0.0000)
	p = 0.52
Income.composition.of.resources	40.86 (15.68)
	p = 0.01***
Schooling	−0.03 (0.40)
	p = 0.95
Constant	−11.67 (2,395.09)
	p = 1.00
Observations	131
Log Likelihood	−20.32
Akaike Inf. Crit.	56.64
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 6: Variance Inflation Factors

Life.expectancy	BMI	Total.expenditure	HIV.AIDS	GDP	Income.composition.of.resources	Schooling
2.03	1.28	1.09	1.00	1.62	4.34	2.07

Table 7: Confusion Matrix for Model 1

	True Developed	True Developing
Predicted Developed	14	4
Predicted Developing	5	108

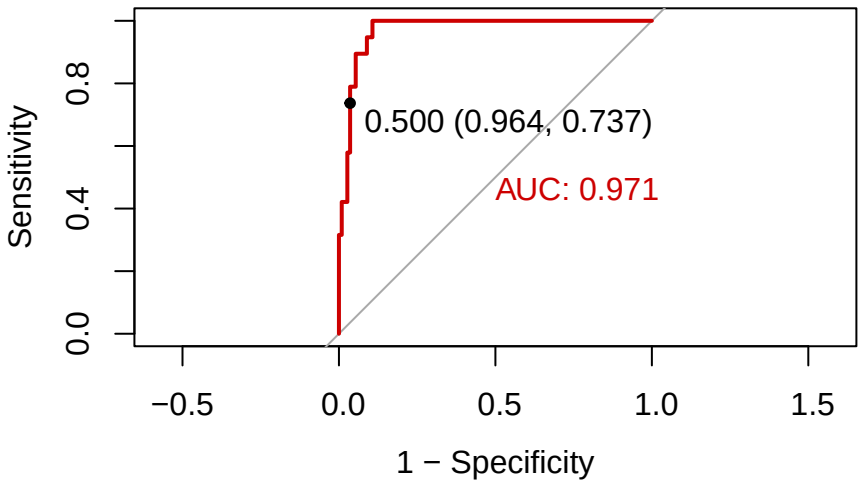


Table 8: Analysis of Deviance: Full Model vs Model 1

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	122	40.36			
2	123	40.64	-1	-0.28	0.59

Without GDP

Model 2

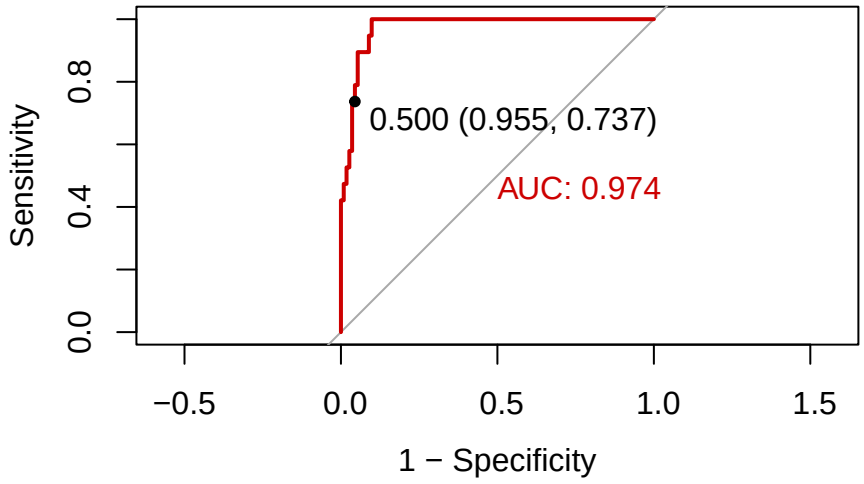
Table 9: Logistic Regression Model

	Dependent variable:
	Status_num
Life.expectancy	−0.07 (0.12)
	p = 0.55
BMI	−0.04 (0.03)
	p = 0.13
percentage.expenditure	−0.0001 (0.0002)
	p = 0.41
Total.expenditure	0.22 (0.17)
	p = 0.22
HIV.AIDS	−153.15 (23,916.24)
	p = 1.00
Income.composition.of.resources	41.73 (15.53)
	p = 0.01***
Schooling	−0.05 (0.39)
	p = 0.89
Constant	−12.26 (2,391.64)
	p = 1.00
Observations	131
Log Likelihood	−20.20
Akaike Inf. Crit.	56.40
Note:	*p<0.1; **p<0.05; ***p<0.01

% Error: Unrecognized object type.

Table 10: Variance Inflation Factors

Life.expectancy	BMI	percentage.expenditure	Total.expenditure	HIV.AIDS	Income.composition.of.resources	Schooling
2.02	1.31	1.63	1.14	1.00	4.21	2.08



Model 1 vs Model 2

Table 11: Analysis of Deviance: Model 1 (W/O percentage expenditure) vs Model 2 (W/O GDP)

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	123	40.64			
2	123	40.40	0	0.24	

Full Model vs Model 2

Table 12: Analysis of Deviance: Full Model vs Model 2

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	122	40.36			
2	123	40.40	-1	-0.04	0.83

Using Model 1 (w/o percentage expenditure) to predict out-of-sample (Year 2013) probabilities

Table 13: Confusion Matrix 4

	True Developed	True Developing
Predicted Developed	14	6
Predicted Developing	5	105

Using Model 2 (w/o GDP) to predict out-of-sampl e(Year 2013) probabilities

Table 14: Confusion Matrix 5

	True Developed	True Developing
Predicted Developed	14	5
Predicted Developing	5	106

Model 2 is slightly better, numerical wise. But not statistically different than Model 1

Tested, not good

Without Total Expenditure and percentage expenditure.

Model 3

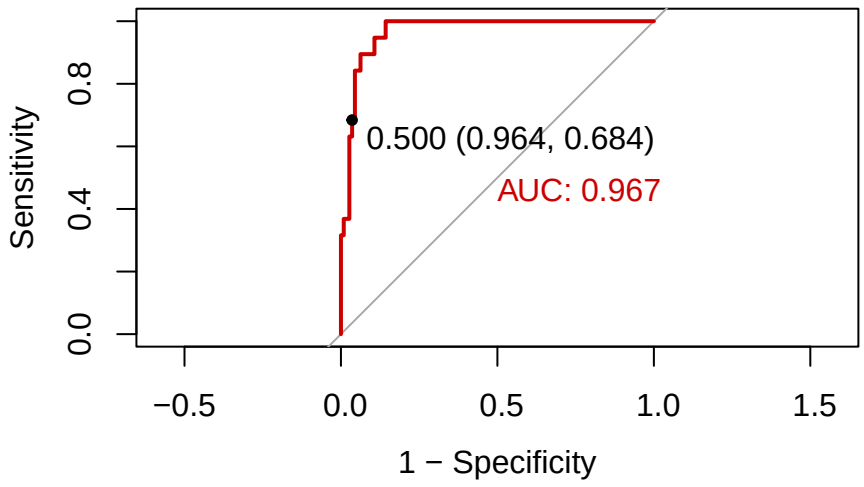
Table 15: Logistic Regression Model

	Dependent variable:
	Status_num
Life.expectancy	−0.04 (0.11)
	p = 0.74
BMI	−0.03 (0.03)
	p = 0.19
HIV.AIDS	−152.18 (24,634.34)
	p = 1.00
Income.composition.of.resources	35.30 (13.99)
	p = 0.02**
Schooling	0.01 (0.39)
	p = 0.98
Constant	−9.94 (2,463.44)
	p = 1.00
Observations	131
Log Likelihood	−21.27
Akaike Inf. Crit.	54.54
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 16: Variance Inflation Factors

Life.expectancy	BMI	HIV.AIDS	Income.composition.of.resources	Schooling
2.01	1.19	1.00	3.62	2.13

% Error: Unrecognized object type.



Model 1 vs Model 3

Table 17: Analysis of Deviance: Model 1 vs Model 3

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	123	40.64			
2	125	42.54	-2	-1.90	0.39

Full Model vs Model 3

Table 18: Analysis of Deviance: Full Model vs Model 3

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	122	40.36			
2	125	42.54	-3	-2.18	0.54

Model Assessment

[How did we assess the linear model and its assumptions? Plots, four key assumptions, etc.]

[How did we assess the validity of the logistic model?]

Results

Here you should present results for all aspects of the analysis. The structure of this section should mirror the structure of the methods section. For example, you can start with a few key EDA results (e.g., a table of descriptive statistics), then present model results, then address assessment. This is the section where you will primarily refer to tables and figures. You should have at least 1 figure for each research question that illustrates a key result of the analysis.

[General insights. Were the models effective, set the stage for the discussion below]

Question 1: *"How did major disease, economic, and social factors impact life expectancy around the globe in 2014?"*

Model Results / Interpretation

[What does the model output tell us]

From the evaluation of the three models, (state adjusted R^2 and BIC values).

(model interpretation)

Model Assessment

[Assess the validity of the outputs]

We can see from the model summary plots that the linear regression assumptions for the model are met.

1. **Check for linearity:** The residuals vs. fitted plot is used to check the assumption that the variables have a linear relationship. The mean line is almost horizontal with no distinct pattern or trend in the data, which is an indication of a linear relationship.

2. **Check for homoscedasticity of residuals or equal variance of errors:** To check for this assumption, we must study plots 1 and 4, that show how the residuals vary as the fitted values increase. Residuals are the unexplained variance. While they are not the same as model error, they are calculated from it, so seeing a bias in the residuals would indicate a bias in the error. In plot 1, the mean line is close to horizontal, with no visible trend. In plot 4, there is no visible trend due to high influence points. The model does not break the assumption of equal variance of errors.
3. **Check for normal distribution of residuals:** The normal Q-Q plot is used to examine if the residuals are normally distributed. If the real residuals from the model match the theoretical residuals from a perfect model, the plot will be a straight line. Since the plot is close to the straight line, we can accept the assumption that the residuals are normally distributed.

From the model summary plots, we can also see that there are no leverage points that we need to be concerned about.

From the model summary, we can see that the F-statistic is a very small value, less than 0.05, which can be interpreted as ().

Question 2: *"How did disease and mortality rates, along with national economic factors, contribute to a country's development status in 2014?"*

Model Results

[What does the model output tell us]

Assessment

[What kid of accuracy did we find from the model]

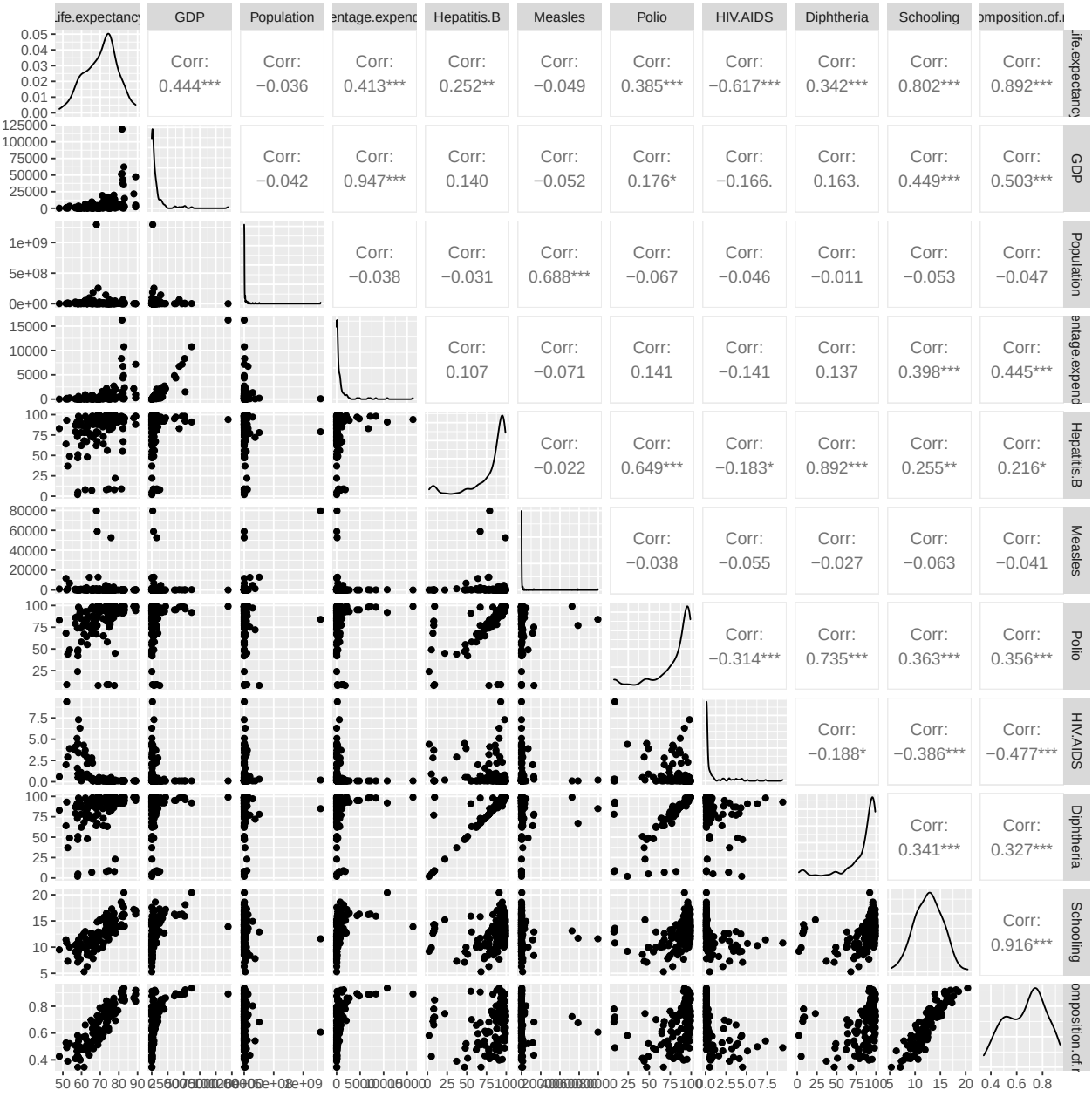
Conclusion

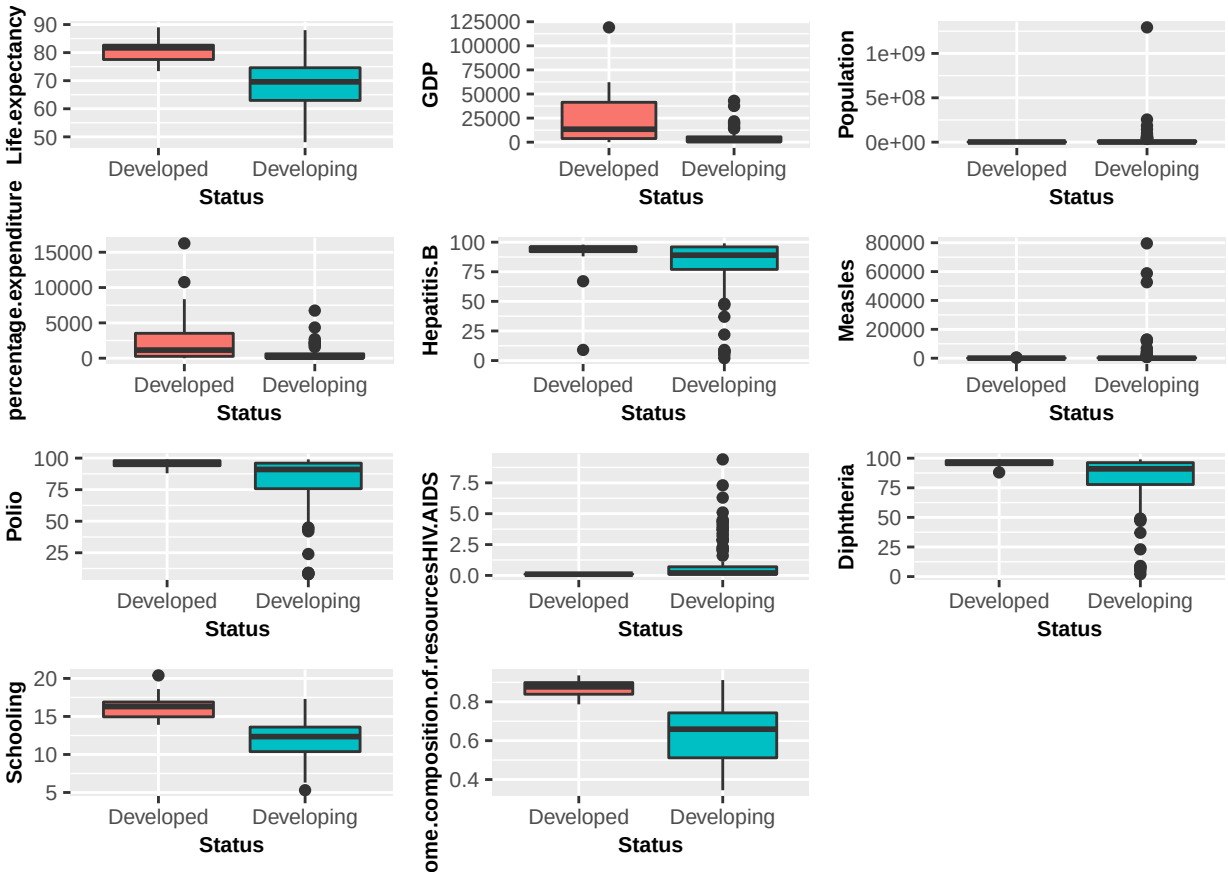
Describe the key takeaways from your analysis, limitations, and future work that can be done to advance knowledge in this area.

[What is the impact of this analysis, do we think it is insightful or not?]

Appendix

[Presently, a dumping ground for all our images and lots until we know what we want to keep]




Life.expectancy
90
80
70
60
50
Developed
Developing
Status
GDP
125000
100000
75000
50000
25000
0
Developed
Developing
Status
Population
1e+09
5e+08
0e+00
Developed
Developing
Status
percentage.expenditure
15000
10000
5000
0
Developed
Developing
Status
Hepatitis.B
100
75
50
25
0
Developed
Developing
Status
Measles
80000
60000
40000
20000
0
Developed
Developing
Status
Polio
100
75
50
25
Developed
Developing
Status
ome.composition.of.resourcesHIV.AIDS
7.5
5.0
2.5
0.0
Developed
Developing
Status
Diphtheria
100
75
50
25
0
Developed
Developing
Status
Schooling
20
15
10
5
Developed
Developing
Status
0.8
0.6
0.4
Developed
Developing
Status

A few things to keep in mind: • You should never refer to actual variable names in the text, tables, or figures. For example, if a variable for height is called "ht___cm," you should always say "height," and the first time you mention it you should state that it is measured in cm. In plots and tables, it should say "height (cm)" • The report should be produced in R Markdown and knit to PDF. This may mean you need to create tables "manually" with knitr. I recommend this anyway because you can customize the labels and formatting. • Someone should be able to read the abstract and look at the tables and figures and have a pretty good idea of 1) the goals of your analysis, and 2) the key results. • I recommend using colorblind-friendly color palettes in your figures. It can be even better to differentiate with line types or symbols instead of relying on color.

Keep you audience in mind! A non-statistician should be able to read your report and have a good idea of what you did. • You can have an appendix if tables or figures are too large to fit into the main text. For example, if you have several predictors, you may want to put a table of model results in the appendix.