IMT 572 Final Project:
Predicting Life Expectancy

Group 1 - Nivedita Arvind, Saman Ateeq, and Makenna Barton

PROBLEM STATEMENT

The dataset our group found and would like to dive deeper into is the WHO Life Expectancy Dataset (found at https://www.kaggle.com/kumarajarshi/life-expectancy-who). This dataset includes observations related to life expectancy and related factors from 193 different countries from the years 2000-2015. This dataset has both categorical and continuous variables.

Based on an initial analysis (shown below in the Data Exploration section) of this dataset, there appeared to be an observable difference between the life expectancies of individuals from developed and developing countries (78 years for the former and 67 for the latter), and an observable difference in immunization coverage of Polio, Hepatitis B, Measles, and Diphtheria.

RESEARCH QUESTION

After considering the variables in this dataset, initial analysis, and our curiosity surrounding life expectancy and immunization, we formulated the following research question:

Is there a statistically significant relationship between Immunization coverage and life expectancy based on correlation analysis?

Variables of interest:

Dependant variable: Life expectancy
Independent variables: Hepatitis B, Polio, Measles and Diphtheria

HYPOTHESIS

H0: There is no statistically significant relationship between the independent and dependant variables at a 95% confidence
H1: There is a statically significant relationship between life expectancy and immunization (Hepatitis B, Polio, Measles and Diphtheria) at a 95% confidence interval.

DATA EXPLORATION

The original dataset consists of 2938 observations of 22 variables. The following is a list of all the variables in the set, the highlighted features are the variables of interest for our analysis:

- **'Country'** = Country
- **'Year'** = Year of observation
- **'Status'** = Country status, Developing or Developed
- **'Life expectancy** ' = Life expectancy in age
- **'Adult Mortality'** = Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
- **'infant deaths**' = Number of Infant Deaths per 1000 population
- **'Alcohol'** = Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
- **'percentage expenditure'** = Expenditure on health as a percentage of Gross Domestic Product per capita(%)
- **'Hepatitis B'** = Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
- **'Measles** ' = Measles - number of reported cases per 1000 population
- ' **BMI** ' = Average Body Mass Index of entire population
- **'under-five deaths '** = Number of under-five deaths per 1000 population
- **'Polio'** = Polio (Pol3) immunization coverage among 1-year-olds (%)
- **'Total expenditure'** = General government expenditure on health as a percentage of total government expenditure (%)
- **'Diphtheria '** = Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
- **' HIV/AIDS'** = Deaths per 1000 live births HIV/AIDS (0-4 years)
- **'GDP'** = Gross Domestic Product per capita (in USD)
- **'Population'** = Population of the country
- **' thinness 1-19 years'** = Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
- **' thinness 5-9 years'** = Prevalence of thinness among children for Age 5 to 9(%)
- **'Income composition of resources'** = Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
- **'Schooling'** = Number of years of Schooling(years)

The following are the key components from the initial exploration and analysis we did of the WHO Life Expectancy dataset. This included cleaning of the dataset as well.

A sample of 5 observations from the dataset:

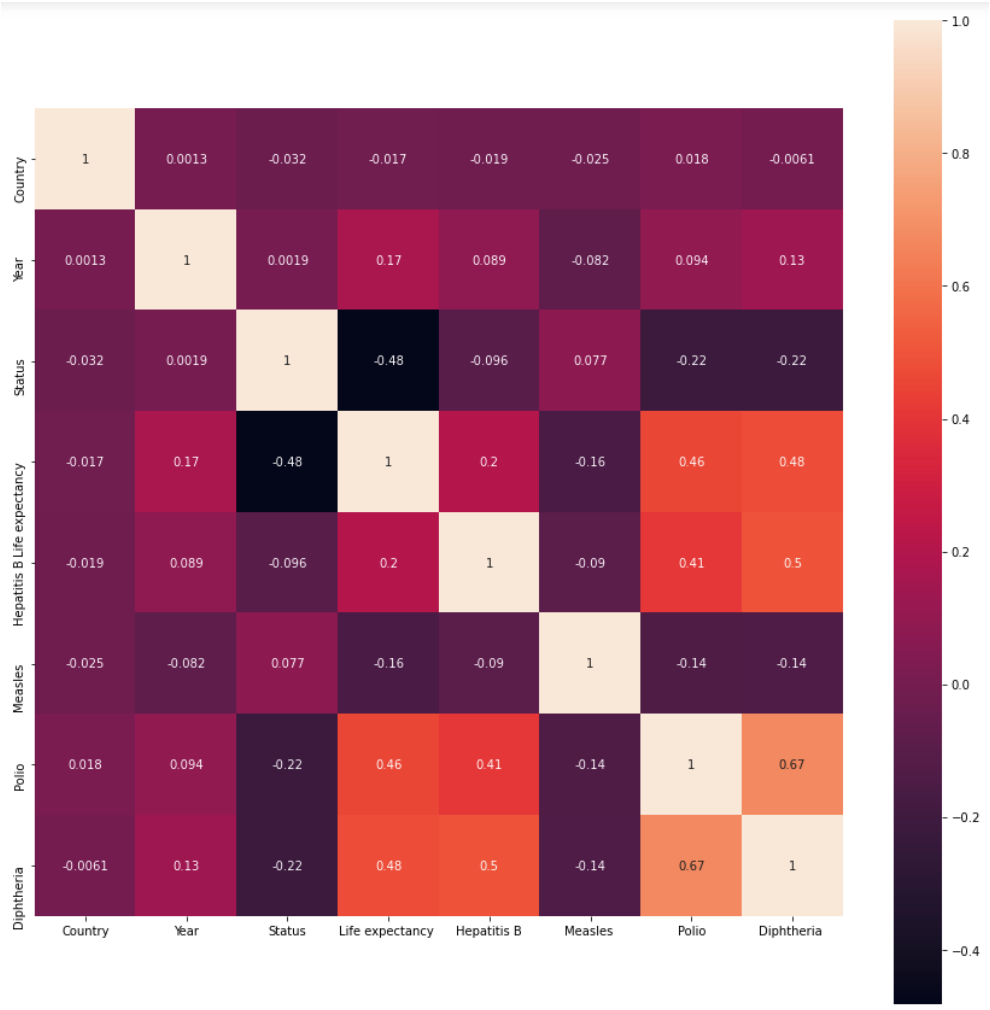
	Country	Year	Status	Life expectancy	Hepatitis B	Measles	Polio	Diphtheria
1144	Honduras	2009	Developing	73.4	97.0	0	97.0	97.0
89	Argentina	2006	Developing	75.2	84.0	0	92.0	91.0
204	Bangladesh	2003	Developing	66.8	5.0	4067	9.0	87.0
764	Djibouti	2004	Developing	58.1	NaN	71	64.0	64.0
1598	Malaysia	2003	Developing	73.1	95.0	632	96.0	96.0

A table of general descriptive statistics of the entire dataset:

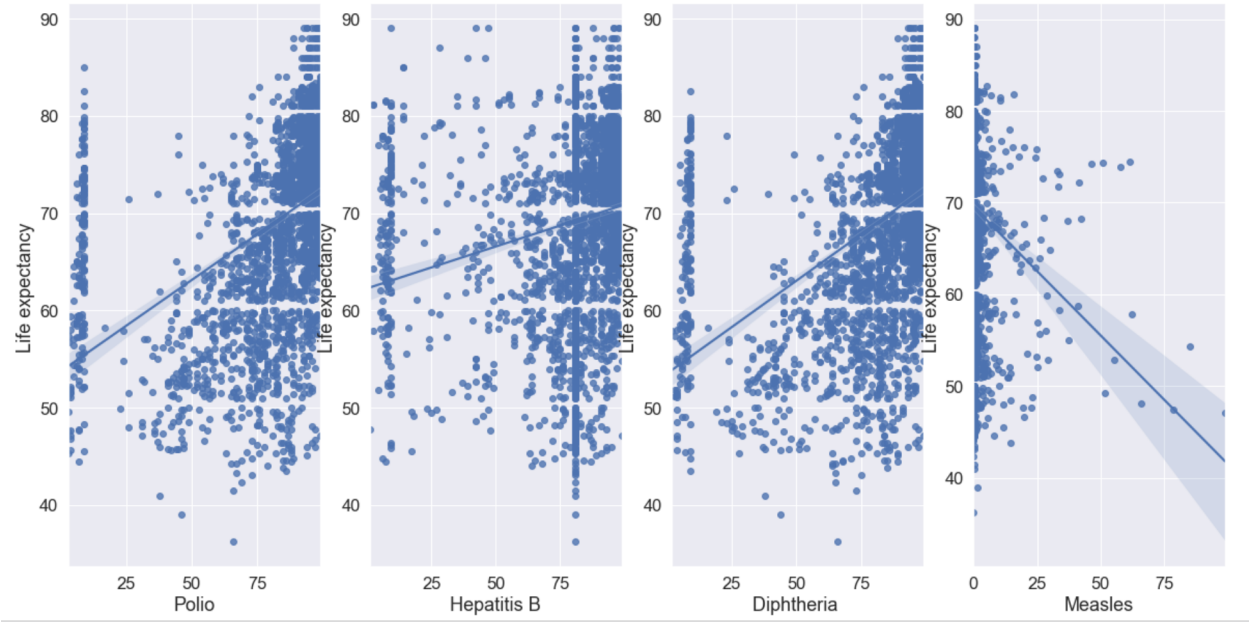
	Country	Year	Status	Life expectancy	Hepatitis B	Measles	Polio	Diphtheria
count	2938.000000	2938.000000	2938.000000	2938.000000	2938.000000	2938.000000	2938.000000	2938.000000
mean	95.096324	2007.518720	0.825732	69.224932	80.940461	2419.592240	82.550188	82.324084
std	56.244904	4.613841	0.379405	9.507640	22.586855	11467.272489	23.352143	23.640073
min	0.000000	2000.000000	0.000000	36.300000	1.000000	0.000000	3.000000	2.000000
25%	46.000000	2004.000000	1.000000	63.200000	80.940461	0.000000	78.000000	78.000000
50%	93.000000	2008.000000	1.000000	72.000000	87.000000	17.000000	93.000000	93.000000
75%	145.000000	2012.000000	1.000000	75.600000	96.000000	360.250000	97.000000	97.000000
max	192.000000	2015.000000	1.000000	89.000000	99.000000	212183.000000	99.000000	99.000000

We could observe that Measles has values in total figures, and rest of the variables of our interest are in percentage, so we scaled the Measles column values using MinMax scaler ranging between 0-99.

A heatmap of the correlation between each of the variables of interest:



A series of plots showing the correlation between our predictors of interest and Life Expectancy:



From the above data exploration, we found that there was a positive correlation between Polio, Hepatitis B, and Diphtheria immunization coverage and life expectancy, and a strong negative correlation between number of Measles cases and life expectancy. This initial observation was the basis for our hypothesis and research motivation. We look to see if this correlation is statistically significant or by random chance.

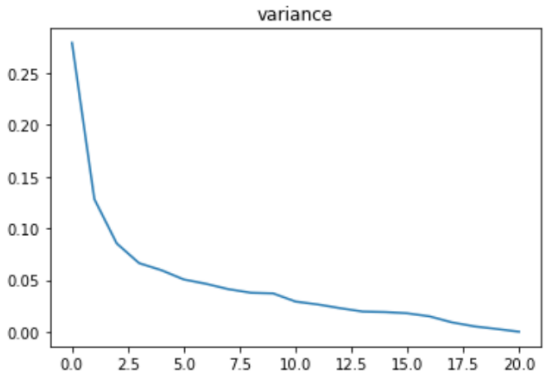
REGRESSION MODEL ANALYSIS

PRINCIPAL COMPONENT ANALYSIS

Table 1

METHOD	R^2 Value (PCA 0.75)	R^2 Value (PCA 0.95)
Logistic Regression	.92	.92
Linear Regression	.78	.81
Decision Tree	.82	.82

To the right you will see the Elbow Analysis plot for the explained variance from fitting our data to a PCA model.



During analysis of our regression models for predicting life expectancy, we ran a PCA at 0.75 and 0.95 on each model. At 0.75, we eliminated ⅔ of the variables and found that 7 principal components explained 75% of variance. However when we set the PCA threshold at 0.95, we were left with 14 principle components for 95% of variance. As you can see in Table 1, the R^2 values for each of the regression models did not vary significantly between the .75 and .95 PCA thresholds. Because of this, we decided to go with the 0.75 PCA threshold as we did not want to overcomplicate our model. In the Logistic Regression, it is also interesting to see that the R2 values are exactly the same at both the 0.75 and 0.95 thresholds.

ERROR ANALYSIS

To evaluate the effectiveness of each of the Regression Models we used during our data analysis, we calculated the following error statistics depicted in Table 2:

Table 2

METHOD	Mean Absolute Error	Mean Squared Error	RMSE
Logistic Regression	0.0782	0.0782	0.2797
Linear Regression	3.3537	19.7096	4.4395
Decision Tree	3.3434	21.5571	4.6430

Based on the error values, we determined that the regression model that best suited this dataset and most accurately predicted Life Expectancy was the Decision Tree Regressor model since there is a level of collinearity between our dependent and independent variables. In the heat map, there is some level of collinearity between the dependent and independent variables, and in this case we would expect the Decision Tree would perform better. However, when it comes to predicting Life Expectancy, it is interesting to see that the MAE, RMSE and the R2 values for the Decision Tree Regression and Linear Regression are almost identical, and neither of the models are over performing the other.

We had also built a Logistic Regression model to predict a categorical variable, Status. This model was the most accurate, however it considered the various factors to predict if an individual was either from a Developed or Developing country, and the target variable was different from the Linear and Decision Tree Regression models.

MULTIVARIATE LINEAR REGRESSION MODEL ANALYSIS

We ran a multivariate linear regression analysis on our variables of interest using an OLS Regression model with the following equation:

$$Life\ Expectancy\ =\ intercept\ +B_1(x_1)+\ B_2(x_2)+B_3(x_3)+B_4(x_4)$$

Where x_1, x_2, x_3, $x_4$represent HepB, Measles, Polio, and Diphtheria respectively and B_1, B_2, B_3, B_4 are their weights.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	52.7445	1.255	42.020	0.000	50.281	55.208
HepB	-0.0114	0.013	-0.909	0.363	-0.036	0.013
Measles	-0.2668	0.063	-4.206	0.000	-0.391	-0.142
Polio	0.1165	0.015	7.590	0.000	0.086	0.147
Diphtheria	0.1323	0.016	8.285	0.000	0.101	0.164

Dep. Variable:	Life_Expectancy	**R-squared:**	0.289
Model:	OLS	**Adj. R-squared:**	0.286
Method:	Least Squares	**F-statistic:**	103.6
Date:	Thu, 09 Dec 2021	**Prob (F-statistic):**	4.98e-74
Time:	23:44:14	**Log-Likelihood:**	-3544.5
No. Observations:	1024	**AIC:**	7099.
Df Residuals:	1019	**BIC:**	7124.
Df Model:	4		
Covariance Type:	nonrobust		

The resulting multivariate model for Life Expectancy from the OLS Regression is:

$$Life\ Expectancy\ =\ 52.7445\ -\ .0114(x_1)\ -\ .2668(x_2)\ +\ .1165(x_3)\ +\ .1323(x_4)$$

Based on the F-statistic of this model we can conclude that there is a statistically significant relationship between Life Expectancy and the predictor variables in the model.

Hypothesis Testing

We performed ANOVA (Analysis of Variance) test to statistically validate our hypothesis. In our dataset, there are 512 instances for developed countries and 2426 instances for developing countries. In order to perform statistical analysis, we created a new sample containing an equal number of data records for developing and developed countries respectively. Using ANOVA, we tested the impact of vaccination related features (Hepatitis B, Polio, Measles and Diphtheria) on life expectancy. We could find that the p-value of the scenario is less than .05, hence our null hypothesis (H0) can be rejected.

CONCLUSION

Based on our analyses, and the statistically significant relationship between the independent immunization variables and the dependent variable Life Expectancy, we are able to reject the null hypothesis of 'There is no statistically significant relationship between the independent and dependent variables at a 95% confidence'.

From our investigation of the WHO Life Expectancy dataset, we think that a major contributing factor to the discrepancy between the life expectancy in developing countries as compared to developed countries is a result of lower immunization rates.