WHO Life Expectancy Data *Analysis*

Problem

Every country have a different life expectancy, and there are a lot of factors that affects a country's life expectancy

What factor affects a country's life expectancy?
Which country has the lowest life expectancy among any other countries?

Objective

- Find the average life expectancy
- Find out which country has the highest life expectancy.
- Find out which country has the lowest life expectancy
- Analyze life expectancy factors, especially education, country development, and healthcare.

About Dataset

Data contains life expectancy, health, immunization, economic, and demographic information

- Population : 179 Countries
- Year Range : 2000 - 2015
- Variable Count : 21
- Data Rows : 2.864

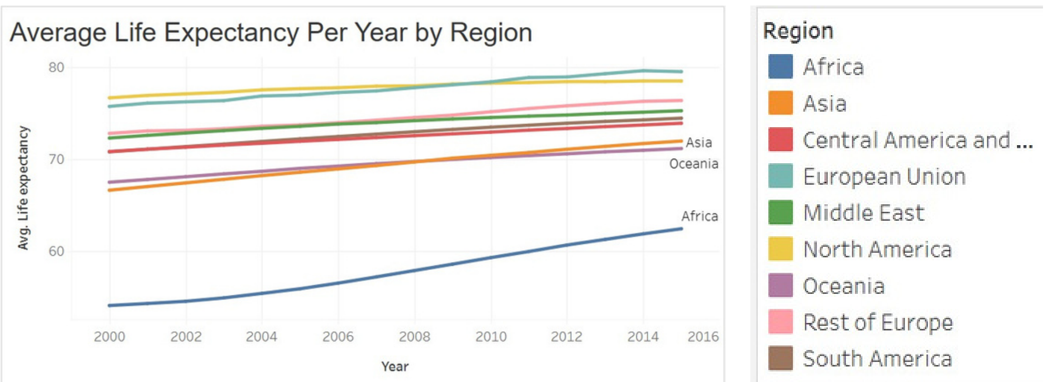
How Do We Analyze?

Some Graph & Code

we do **EDA (Exploratory Data Analysis)**
Through R & Tableau, with graphs to find patterns and insights!

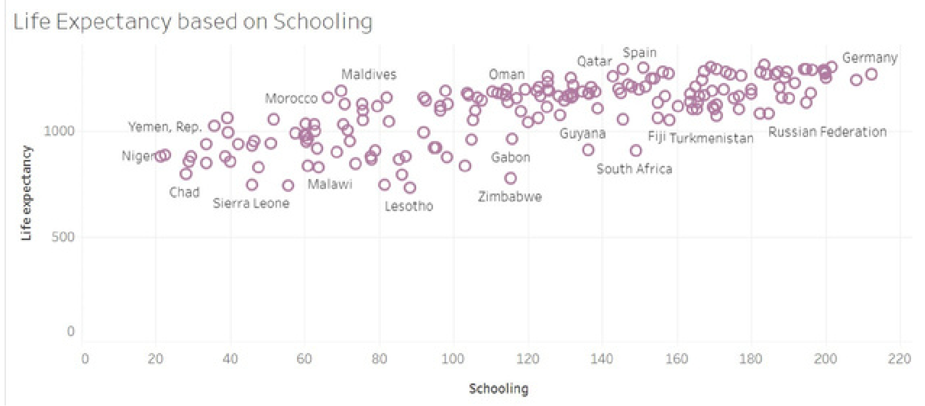
Exploratory Data Analysis (EDA)

Average Life Expectancy per year by Region



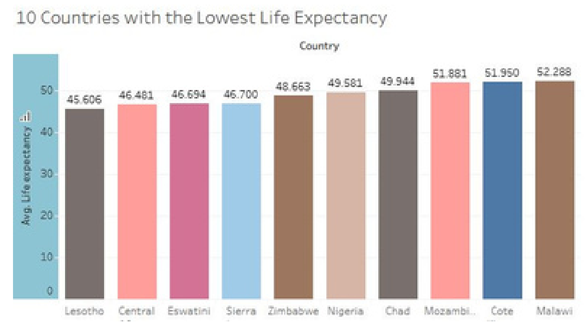
South America has the **highest** Average and **Africa** has the **lowest** average, far low from others. **So we analyze why Africa has the lowest?**

Factors that affects life expectancy - Schooling



Most of the country in Africa has the lowest number of schooling

Finding countries with the lowest Life Expectancy

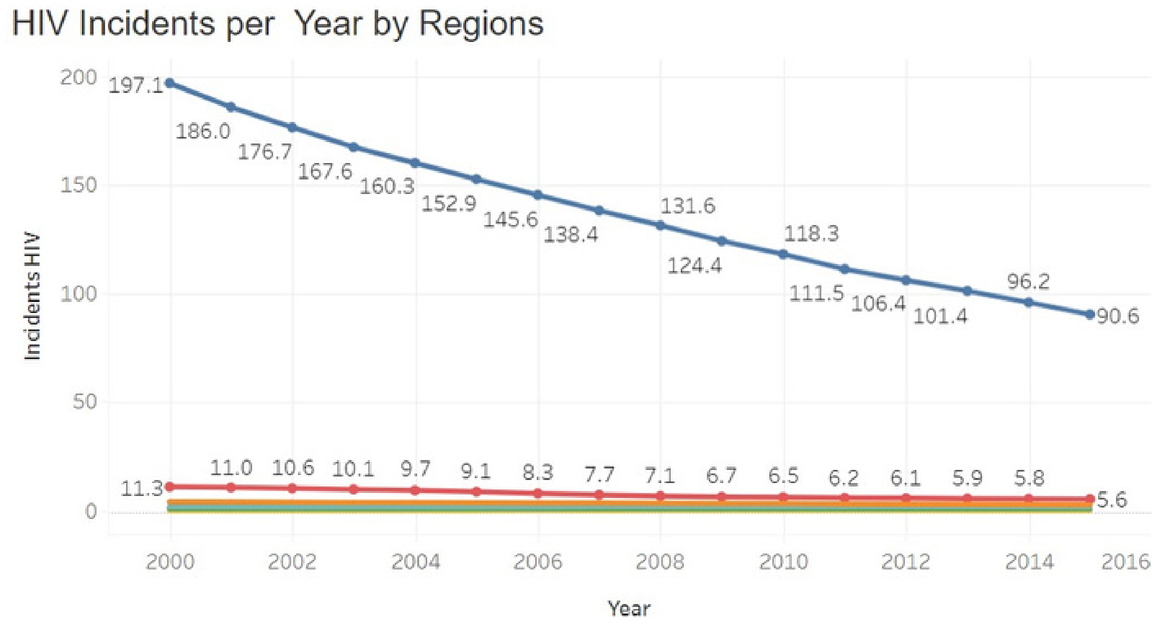


Top 5 Lowest :

1. Malawi
2. Cote
3. Mozambique
4. Chad
5. Nigeria

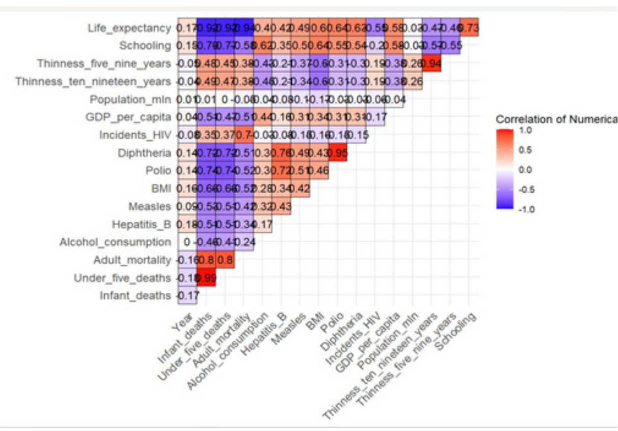
All country with Lowest life expectancy located Africa.

Factors that affects life expectancy - Sickness - HIV (2000 - 2016)



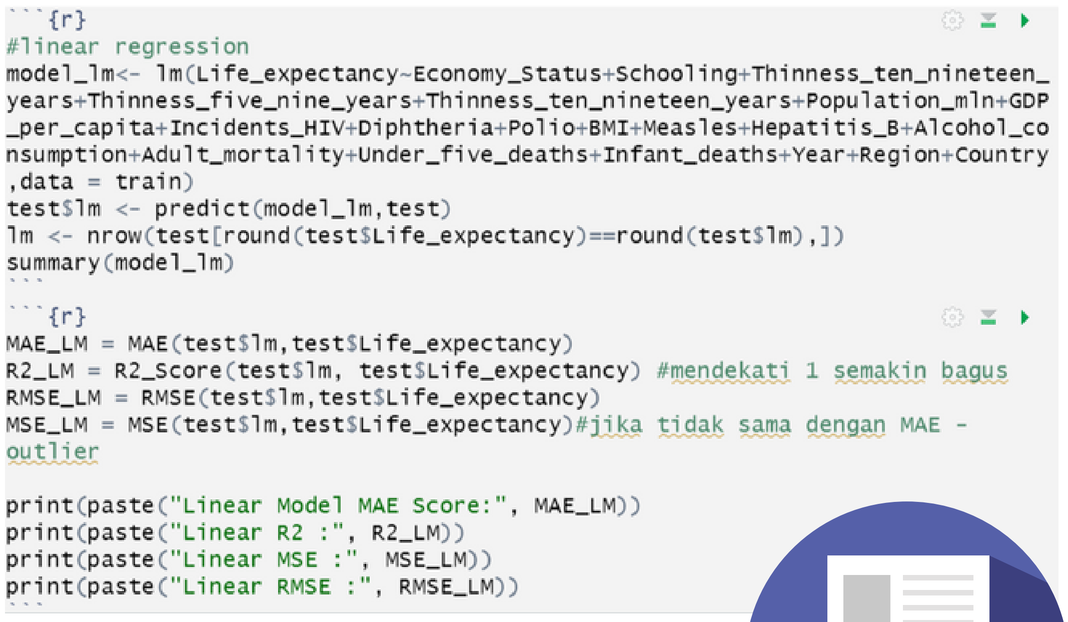
Africa has the highest HIV rate

Data Correlation



Correlation shown among every variable that exist in the dataset

Code - Linear Regression

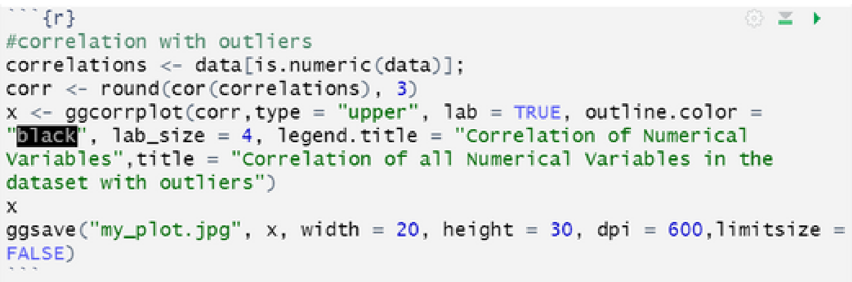


~~~
```{r}
#linear regression
model_lm<- lm(Life_expectancy~Economy_Status+Schooling+Thinness_ten_nineteen_
years+Thinness_five_nine_years+Thinness_ten_nineteen_years+Population_mln+GDP
_per_capita+Incidents_HIV+Diphtheria+Polio+BMI+Measles+Hepatitis_B+Alcohol_co
nsumption+Adult_mortality+Under_five_deaths+Infant_deaths+Year+Region+Country
,data = train)
test$lm <- predict(model_lm,test)
lm <- nrow(test[round(test$Life_expectancy)==round(test$lm),])
summary(model_lm)
~~~

```{r}
MAE_LM = MAE(test$lm,test$Life_expectancy)
R2_LM = R2_Score(test$lm, test$Life_expectancy) #mendekati 1 semakin bagus
RMSE_LM = RMSE(test$lm,test$Life_expectancy)
MSE_LM = MSE(test$lm,test$Life_expectancy)#jika tidak sama dengan MAE -
outlier

print(paste("Linear Model MAE Score:", MAE_LM))
print(paste("Linear R2 :", R2_LM))
print(paste("Linear MSE :", MSE_LM))
print(paste("Linear RMSE :", RMSE_LM))
```
```

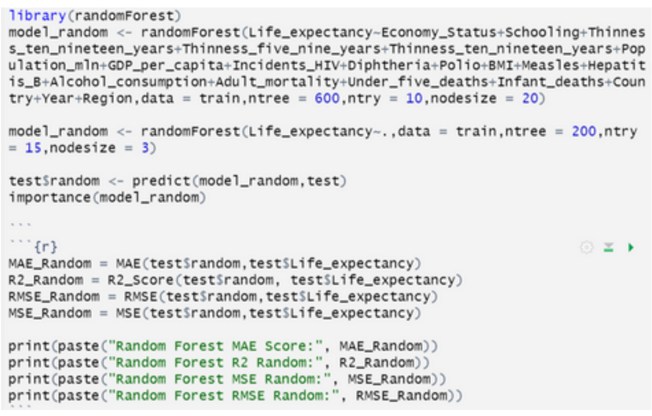
Code - Correlation



```
```{r}
#correlation with outliers
correlations <- data[is.numeric(data)];
corr <- round(cor(correlations), 3)
x <- ggcorrplot(corr,type = "upper", lab = TRUE, outline.color =
"black", lab_size = 4, legend.title = "Correlation of Numerical
Variables",title = "Correlation of all Numerical Variables in the
dataset with outliers")
x
ggsave("my_plot.jpg", x, width = 20, height = 30, dpi = 600,limitsize =
FALSE)
```
```

& Predictive Modelling

Code - Random Forest



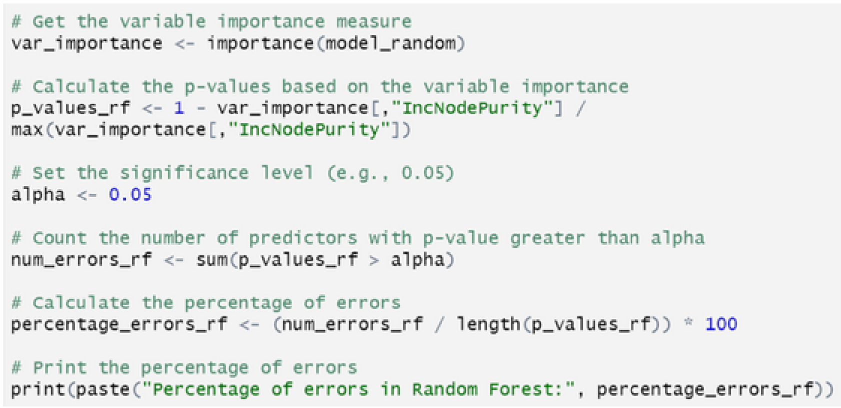
```
library(randomForest)
model_random <- randomForest(Life_expectancy~Economy_Status+Schooling+Thinnes
s_ten_nineteen_years+Thinness_five_nine_years+Thinness_ten_nineteen_years+Pop
ulation_mln+GDP_per_capita+Incidents_HIV+Diphtheria+Polio+BMI+Measles+Hepatit
is_B+Alcohol_consumption+Adult_mortality+Under_five_deaths+Infant_deaths+Coun
try+Year+Region,data = train,ntree = 600,ntry = 10,nodesize = 20)

model_random <- randomForest(Life_expectancy~.,data = train,ntree = 200,ntry
= 15,nodesize = 3)

test$random <- predict(model_random,test)
importance(model_random)

~~~
```{r}
MAE_Random = MAE(test$random,test$Life_expectancy)
R2_Random = R2_Score(test$random, test$Life_expectancy)
RMSE_Random = RMSE(test$random,test$Life_expectancy)
MSE_Random = MSE(test$random,test$Life_expectancy)

print(paste("Random Forest MAE Score:", MAE_Random))
print(paste("Random Forest R2 Random:", R2_Random))
print(paste("Random Forest MSE Random:", MSE_Random))
print(paste("Random Forest RMSE Random:", RMSE_Random))
~~~
```



```
# Get the variable importance measure
var_importance <- importance(model_random)

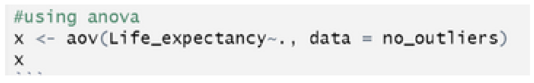
# Calculate the p-values based on the variable importance
p_values_rf <- 1 - var_importance[,"IncNodePurity"] /
max(var_importance[,"IncNodePurity"])

# Set the significance level (e.g., 0.05)
alpha <- 0.05

# Count the number of predictors with p-value greater than alpha
num_errors_rf <- sum(p_values_rf > alpha)

# Calculate the percentage of errors
percentage_errors_rf <- (num_errors_rf / length(p_values_rf)) * 100

# Print the percentage of errors
print(paste("Percentage of errors in Random Forest:", percentage_errors_rf))
```



```
#using anova
x <- aov(Life_expectancy~., data = no_outliers)
x
```
```

Conclusion

Africa has the lowest life expectancy, compared to other country, this data shown by the average graph and "lowest life expectancy" bar chart

The low life expectancy in Africa is caused by many factors, such as the schooling score, as in education, undeveloped countries, and even health factors such as sickness, especially HIV that happened in Africa.

There are 3 variables from the dataset that really affects the low life expectancy in Africa such as infant death, under 5 years death, and adult mortality, this data was taken from the correlation table showing the correlation of every available variable from the dataset

Random forest is the best predictive modelling in this case because it has a very high accuracy which is 0.99, and the value or r squared reaching almost 1, so it can be categorized as a good prediction model for this dataset.

Our Team

Phoebe Patricia Wibowo - 2602080825
Jennifer Ardelia Limicia - 2602105090
Anastasia Jocelyn Hilman - 2602073031