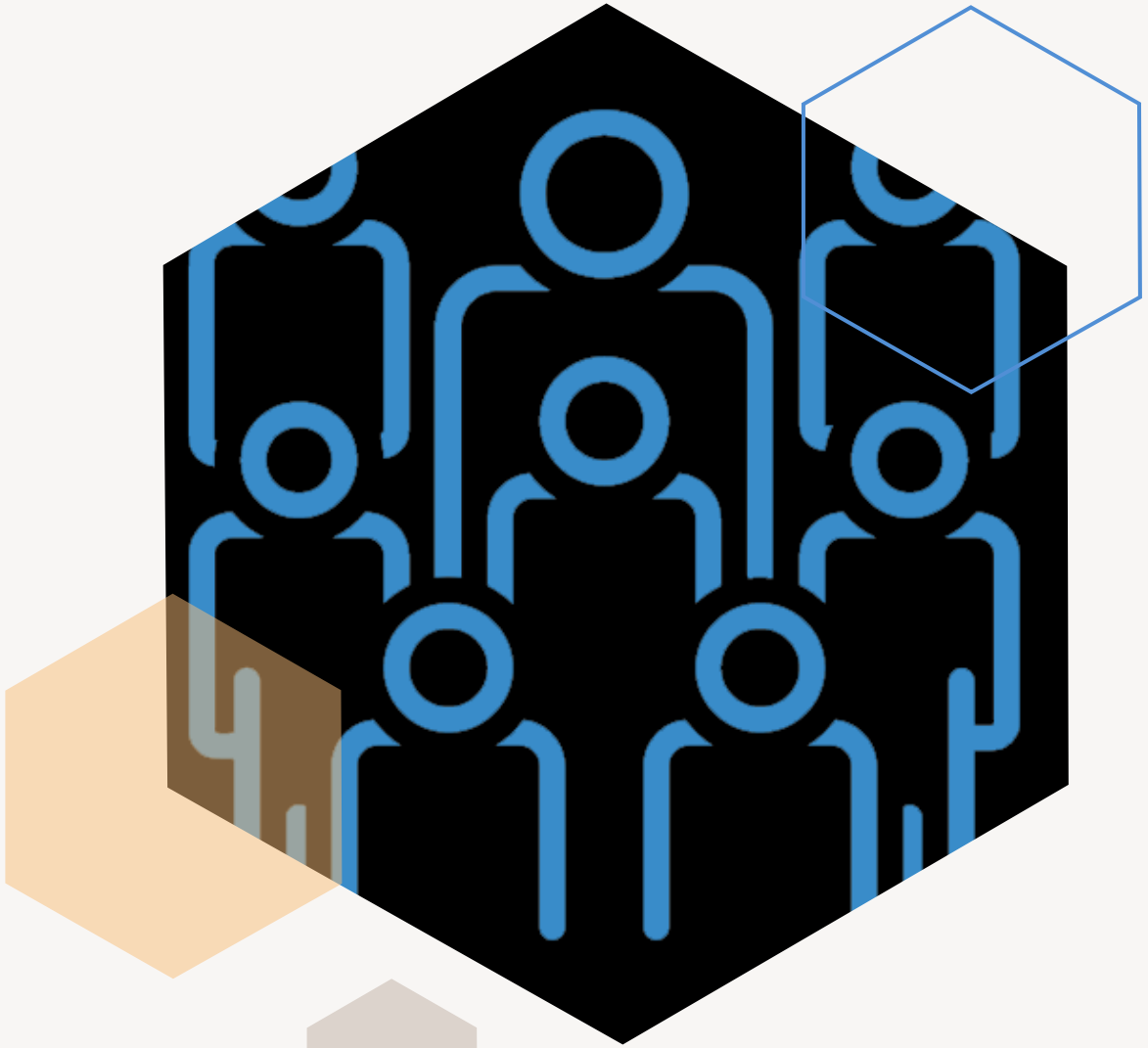
MSDS 6372
Project 1: Group 5

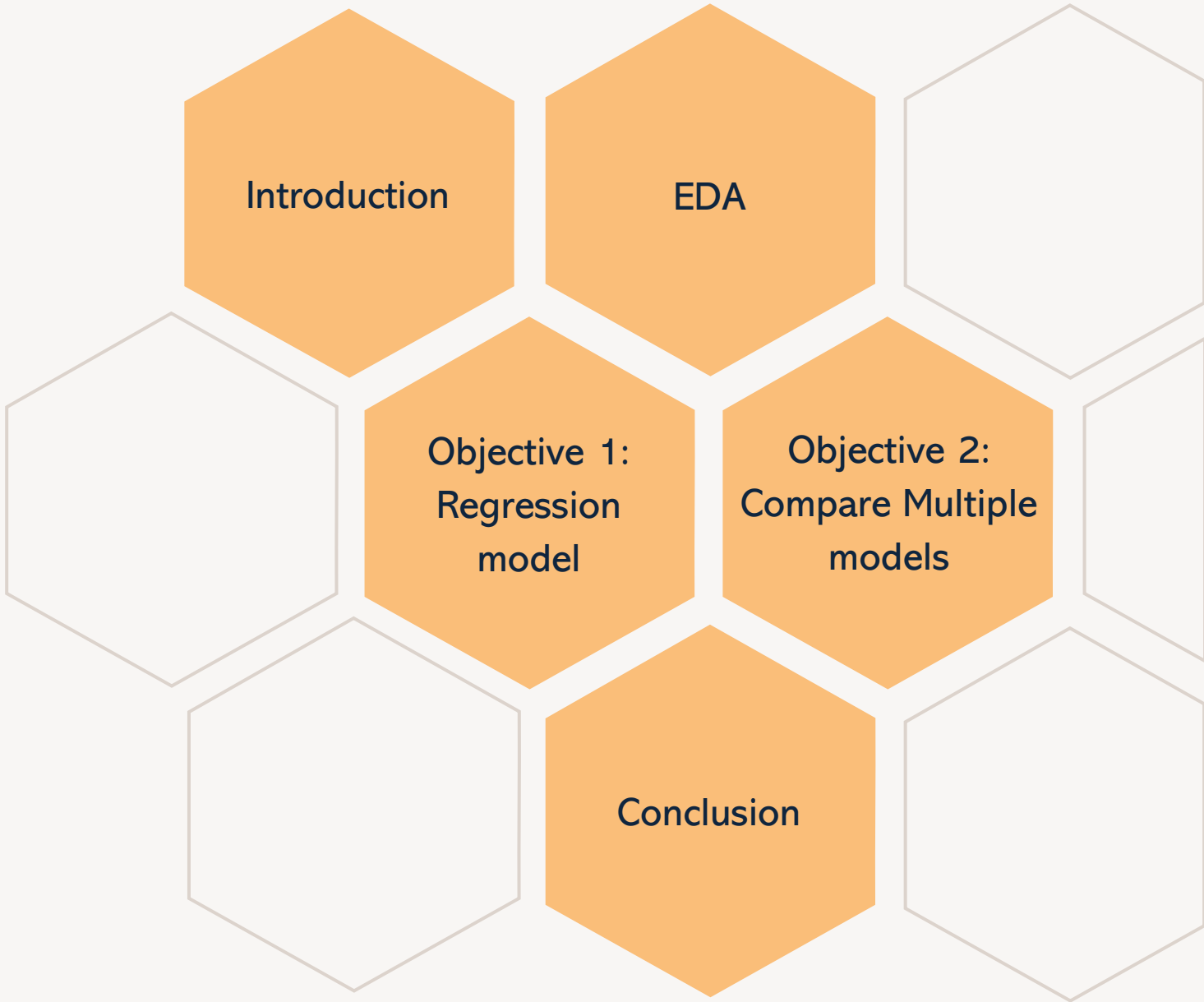
Ivan Chavez

Jessica McPhaul

Rafia Mirza



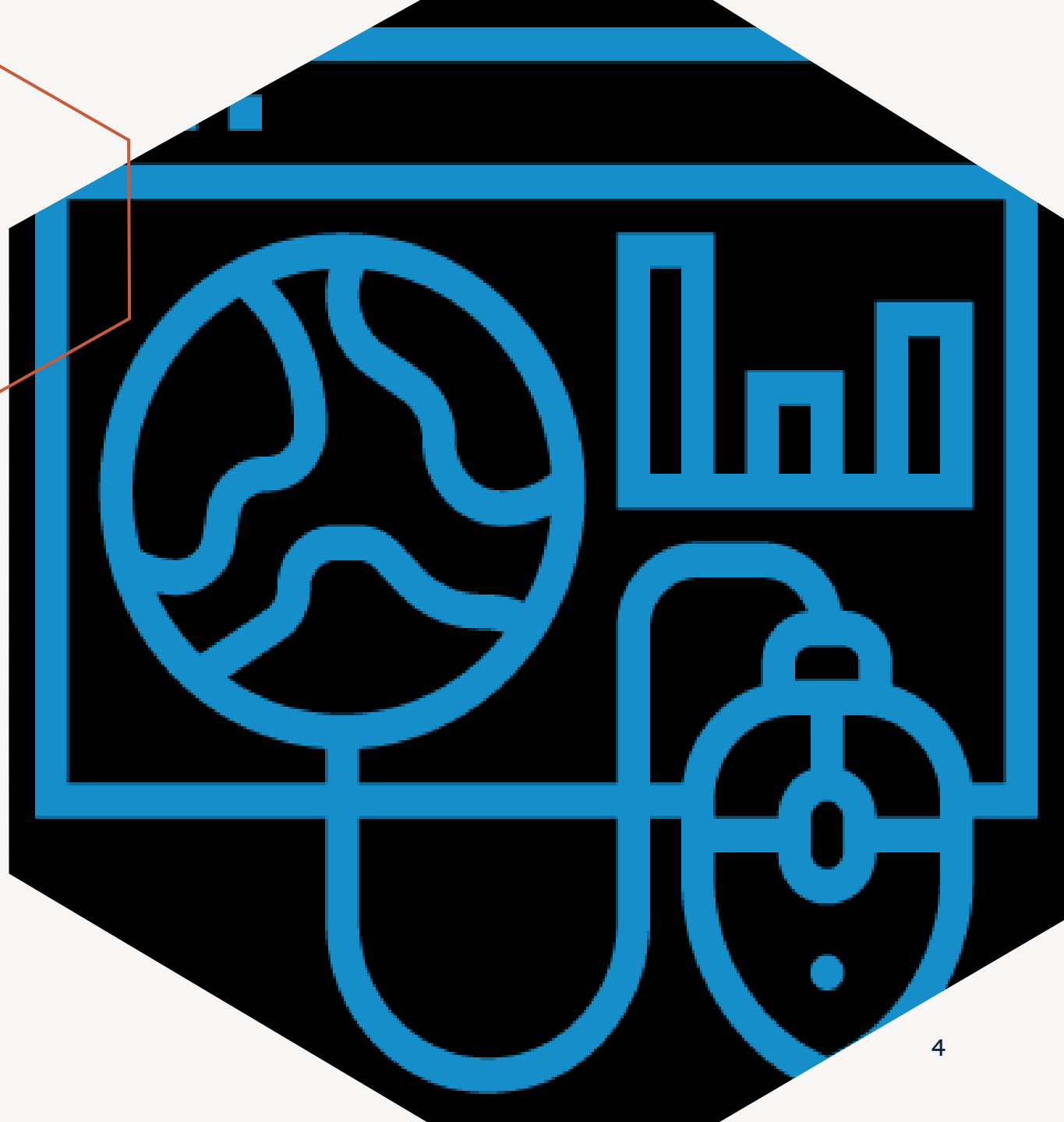
Agenda


Introduction
EDA
Objective 1: Regression model
Objective 2: Compare Multiple models
Conclusion

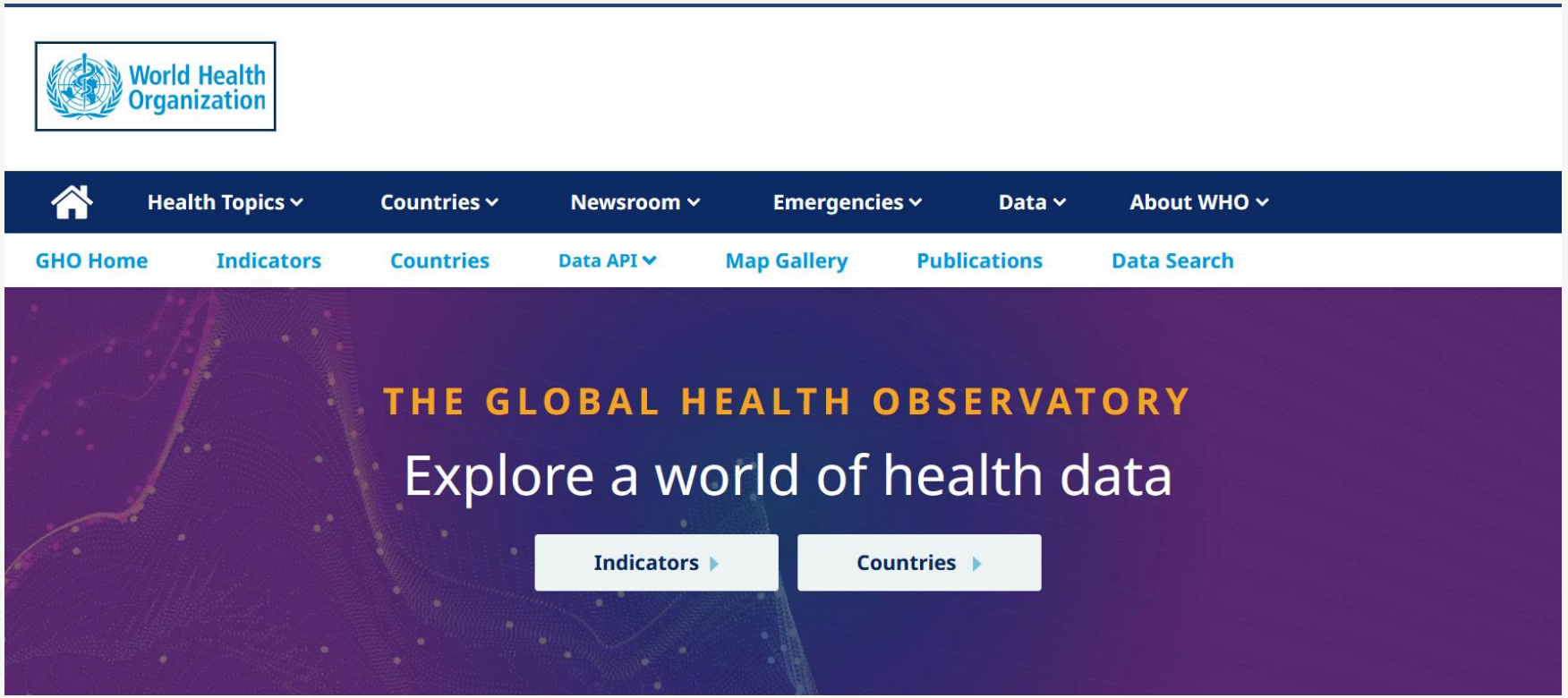
Click here for Final Analysis (Interactive)

Goal

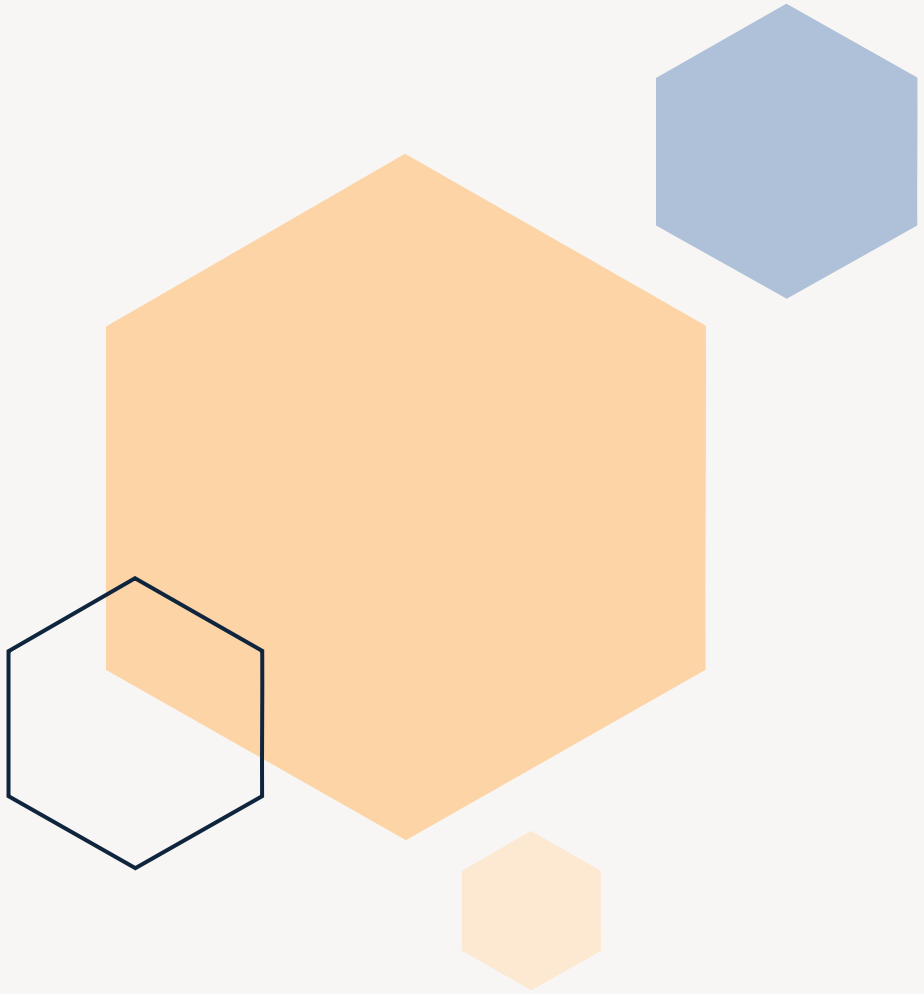
- A regression model that predicts life expectancy based on health indicators



Data Set: Health indicators



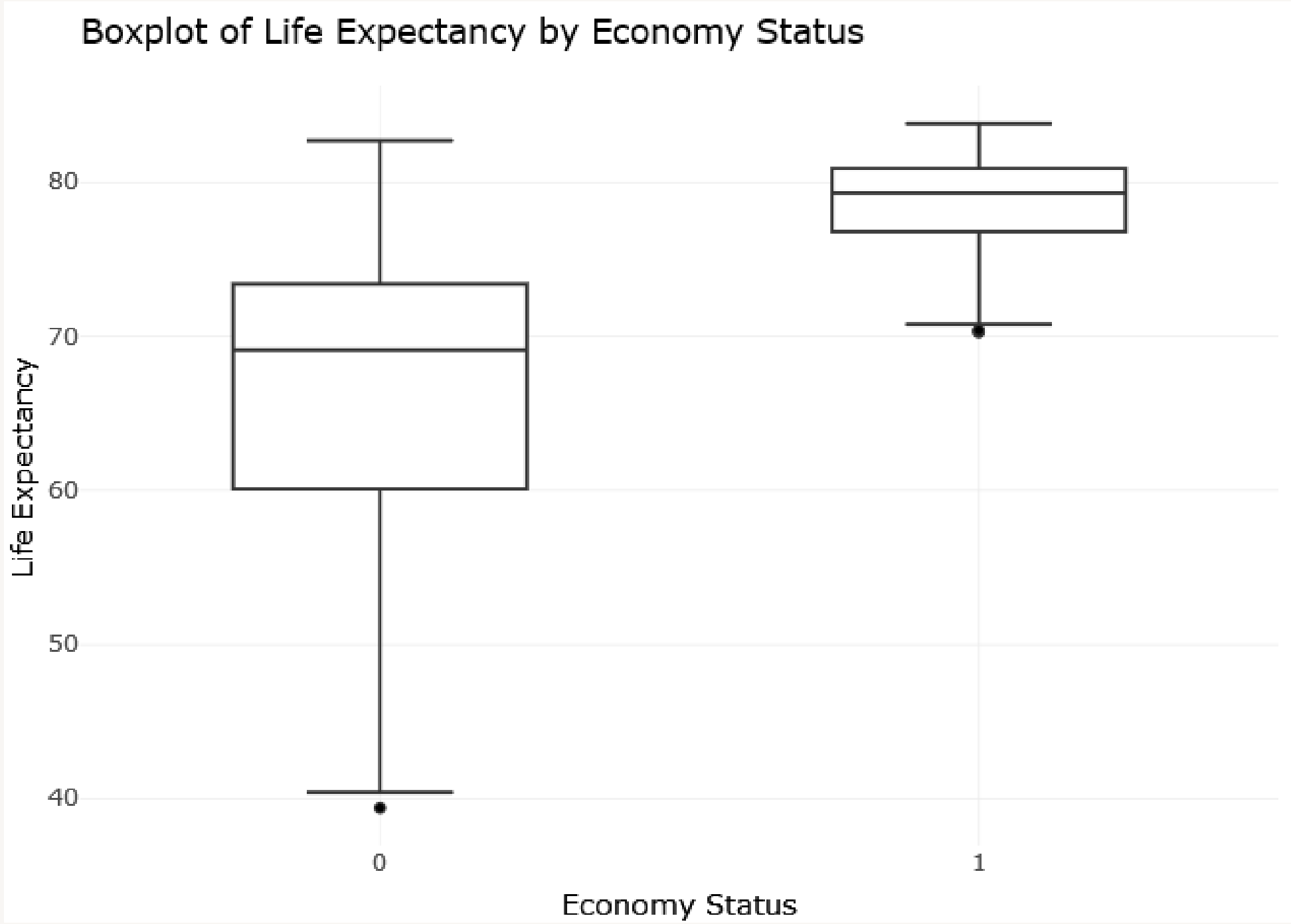
EDA

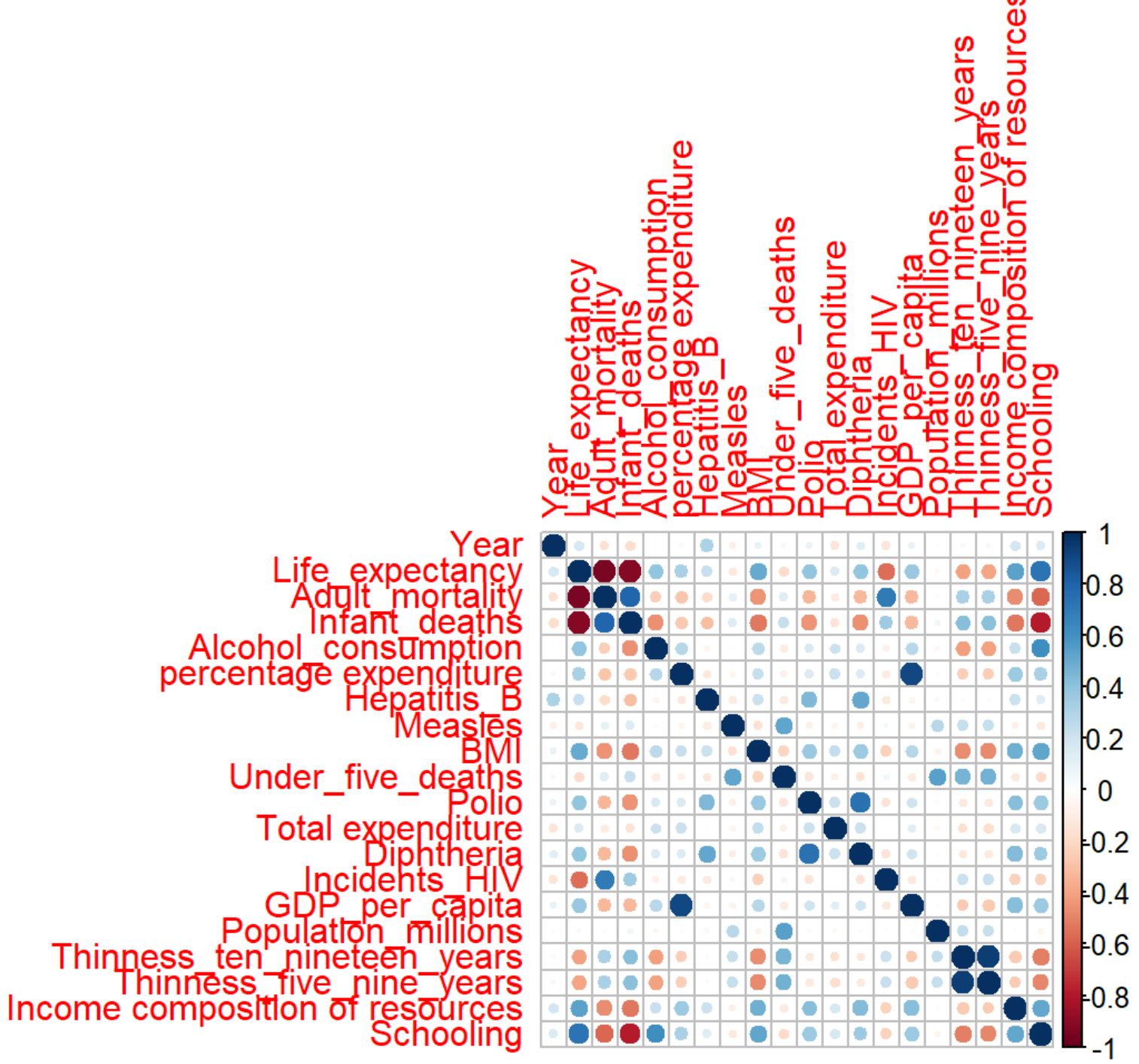


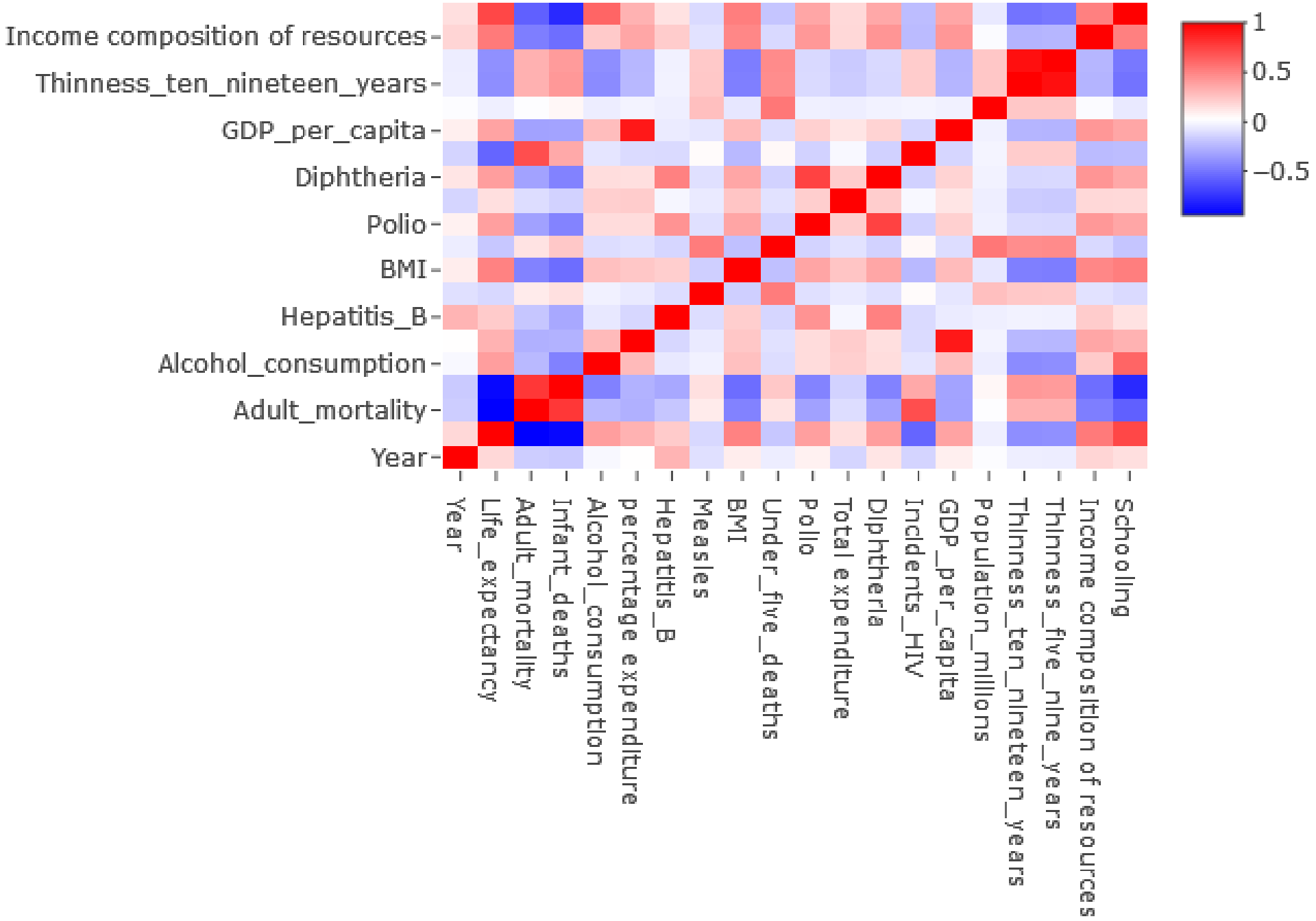
Missing Values

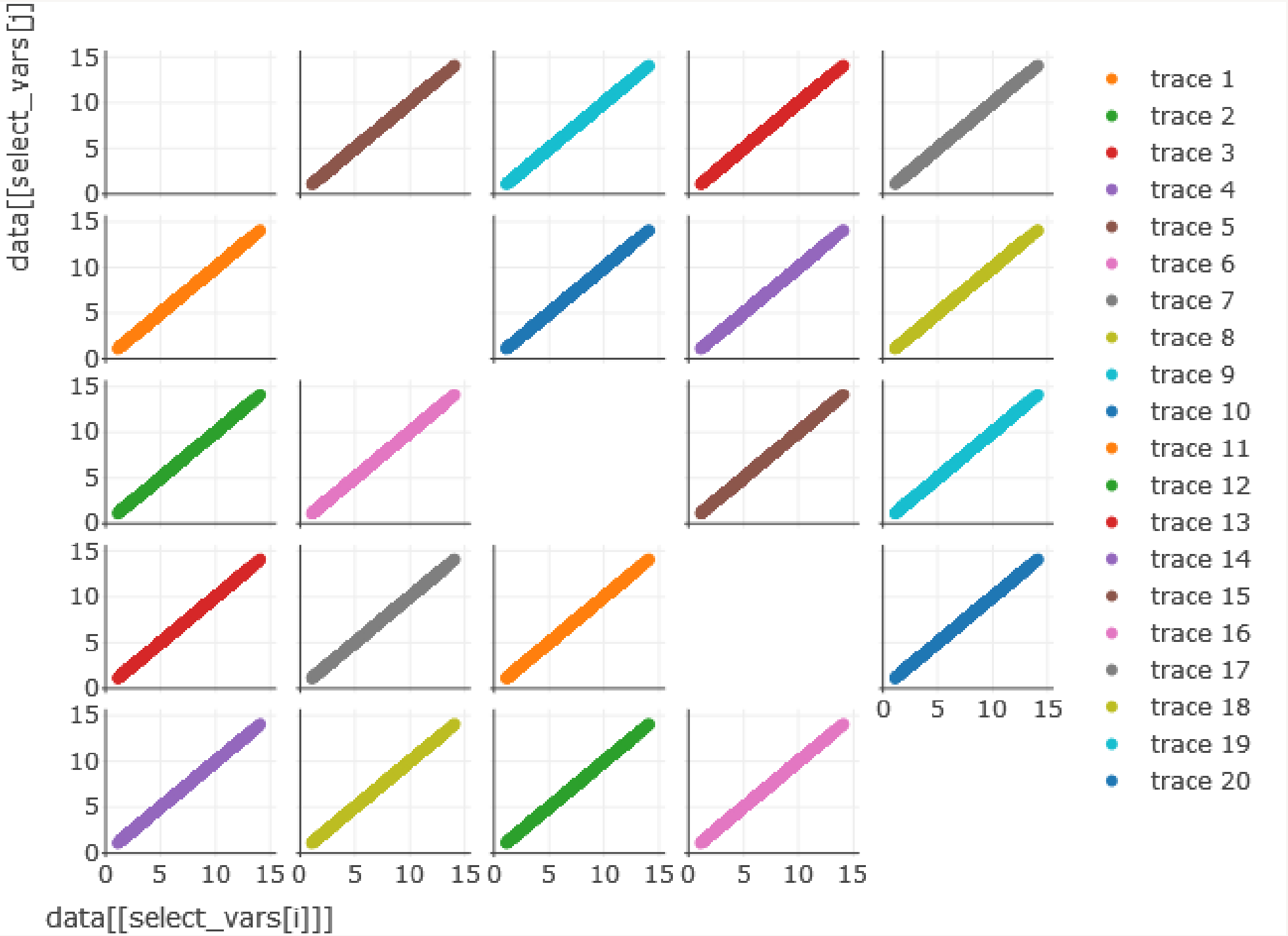
- Imputation
- Removal

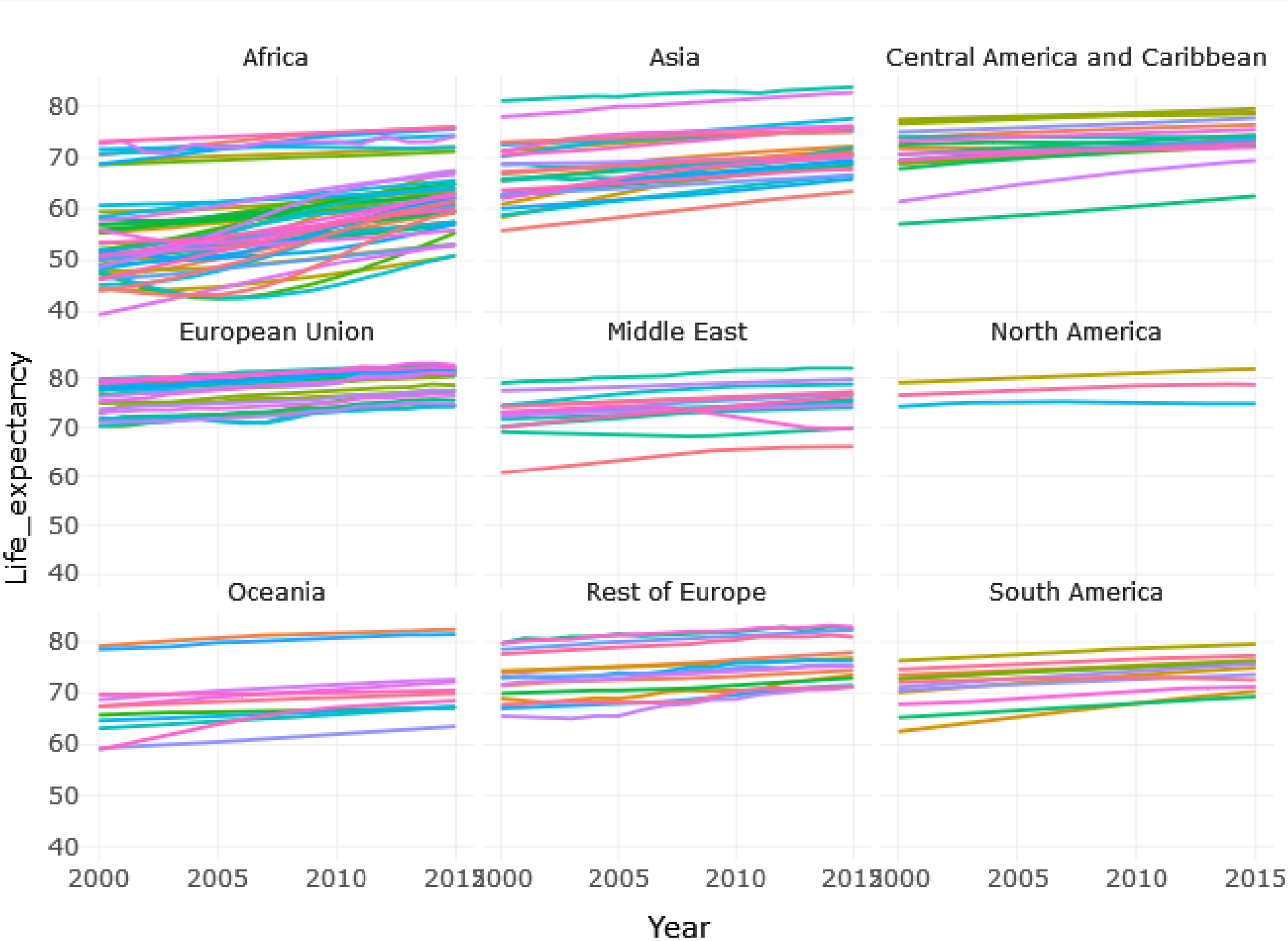
- Visualizations


Boxplot of Life Expectancy by Economy Status
80
70
60
50
40
Life Expectancy
0
1
Economy Status


Year
Life_expectancy
Adult_mortality
Infant_deaths
Alcohol_consumption
percentage expenditure
Hepatitis_B
Measles
BMI
Under_five_deaths
Polio
Total expenditure
Diphtheria
Incidents_HIV
GDP_per_capita
Population_millions
Thinness_ten_nineteen_years
Thinness_five_nine_years
Income composition of resources
Schooling
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


Income composition of resources
Thinness_ten_nineteen_years
GDP_per_capita
Diphtheria
Polio
BMI
Hepatitis_B
Alcohol_consumption
Adult_mortality
Year
Year
Life_expectancy
Adult_mortality
Infant_deaths
Alcohol_consumption
percentage expenditure
Hepatitis_B
Measles
BMI
Under_five_deaths
Polio
Total expenditure
Diphtheria
Incidents_HIV
GDP_per_capita
Population_millions
Thinness_ten_nineteen_years
Thinness_five_nine_years
Income composition of resources
Schooling
1
0.5
0
−0.5


data[[select_vars[j]]]
data[[select_vars[i]]]
0
5
10
15
trace 1
trace 2
trace 3
trace 4
trace 5
trace 6
trace 7
trace 8
trace 9
trace 10
trace 11
trace 12
trace 13
trace 14
trace 15
trace 16
trace 17
trace 18
trace 19
trace 20


Africa
Asia
Central America and Caribbean
European Union
Middle East
North America
Oceania
Rest of Europe
South America
Life_expectancy
80
70
60
50
40
2000
2005
2010
2015
Year

Objective 1: Regression model

Objective 1: Regression model

- We built the initial linear model from the basis of our EDA and running recursive feature elimination to get significant features for predicting life expectancy.
- There were 22 selection variables in the initial model, some of the features included were Alcohol consumption, Incidents of HIV, GDP, BMI, and others.

```
##  [1] "Adult_mortality"                "Infant_deaths"
##  [3] "Alcohol_consumption"            "Under_five_deaths"
##  [5] "Country"                        "Year"
##  [7] "Thinness_five_nine_years"       "Incidents_HIV"
##  [9] "Income composition of resources" "Total expenditure"
## [11] "Thinness_ten_nineteen_years"    "Region"
## [13] "Measles"                        "percentage expenditure"
## [15] "BMI"                            "Diphtheria"
## [17] "Schooling"                      "GDP_per_capita"
## [19] "Polio"                          "Hepatitis_B"
## [21] "Population_millions"            "Economy_status_Developed"
```

- When looking at our regression table of our model we will discuss two variables adult mortality and alcohol consumption their coefficients and how they influence life expectancy.
- The coefficient estimate for adult mortality is -.04070 this indicates that for each unit increase in adult mortality we expect a decrease in life expectancy by approx. .04070 years.
- This along with a statistically significant p-value of 2e-16 suggests that as adult mortality increases life expectancy decreases.
- For alcohol consumption we have a coefficient estimate of -.03304, and this indicates that for each unit increase in alcohol consumption we can expect a decrease in life expectancy by approx. .03304 years.
- Again, we have a statistically significant p-value of .0358 and suggests that there is a relation that when alcohol consumption goes up the life expectancy will fall.

```
## Analysis of Variance Table
##
## Model 1: Life_expectancy ~ Adult_mortality + Infant_deaths + Alcohol_consumption +
##     Under_five_deaths + Country + Year + Thinness_five_nine_years +
##     Incidents_HIV + `Income composition of resources` + `Total expenditure` +
##     Thinness_ten_nineteen_years + Region + Measles + `percentage expenditure` +
##     BMI + Diphtheria + Schooling + GDP_per_capita + Polio + Hepatitis_B +
##     Population_millions + Economy_status_Developed
## Model 2: Life_expectancy ~ Adult_mortality + Infant_deaths + Alcohol_consumption +
##     Under_five_deaths + Country + Year + Thinness_five_nine_years +
##     Incidents_HIV + `Income composition of resources` + `Total expenditure` +
##     Thinness_ten_nineteen_years + Region + Measles + `percentage expenditure` +
##     BMI + Diphtheria + Schooling + GDP_per_capita + Polio + Hepatitis_B +
##     Population_millions + Economy_status_Developed + I(Alcohol_consumption^2)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1799 503.62
## 2   1798 502.92  1   0.70149 2.5079 0.1135
```

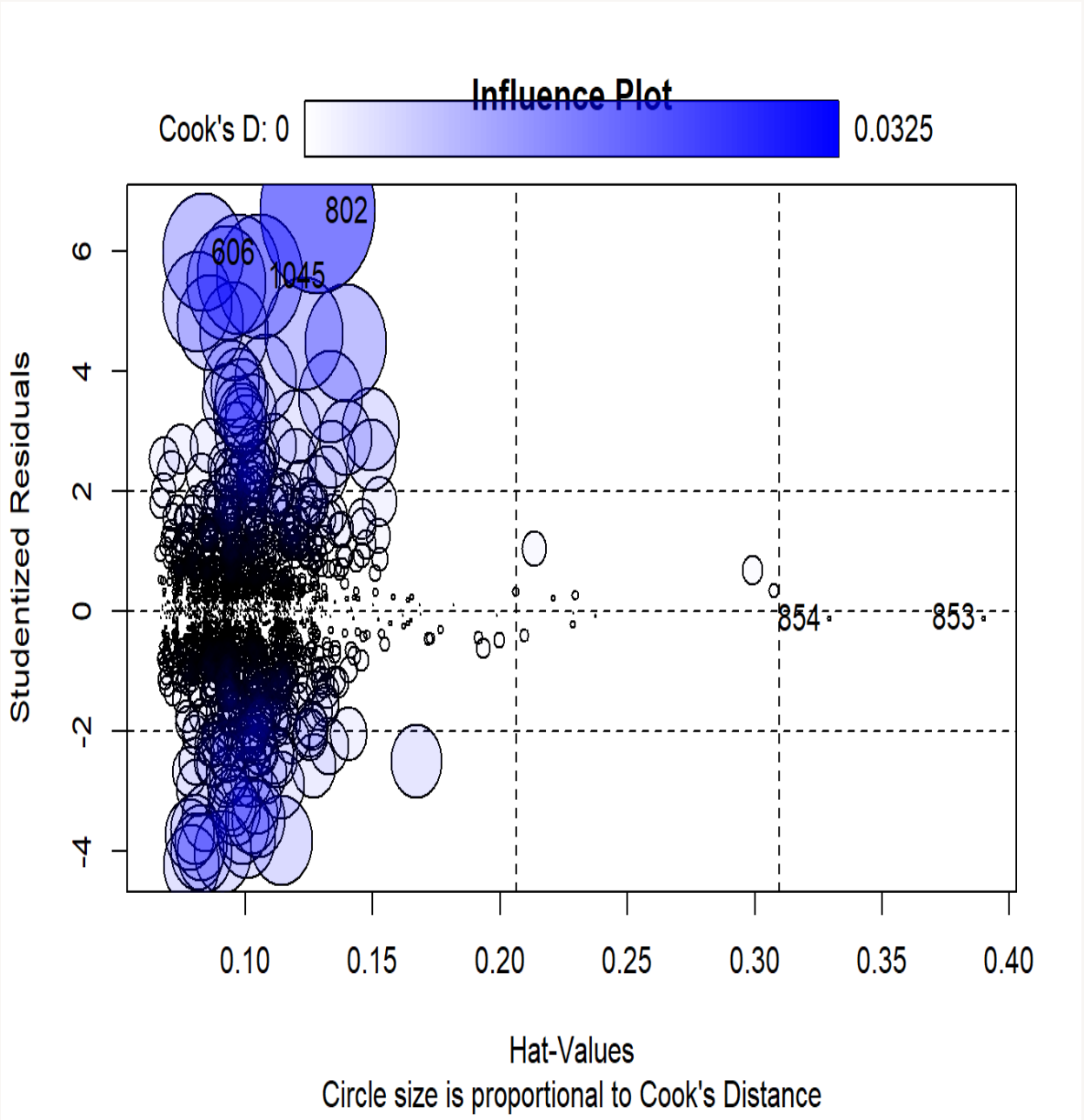
```
## Analysis of Variance Table
##
## Model 1: Life_expectancy ~ Adult_mortality + Infant_deaths + Alcohol_consumption +
##     Under_five_deaths + Country + Year + Thinness_five_nine_years +
##     Incidents_HIV + `Income composition of resources` + `Total expenditure` +
##     Thinness_ten_nineteen_years + Region + Measles + `percentage expenditure` +
##     BMI + Diphtheria + Schooling + GDP_per_capita + Polio + Hepatitis_B +
##     Population_millions + Economy_status_Developed
## Model 2: Life_expectancy ~ Year + Adult_mortality + Alcohol_consumption +
##     BMI
##   Res.Df     RSS   Df Sum of Sq     F    Pr(>F)
## 1   1799   503.6
## 2   1991 12779.8 -192    -12276 228.4 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The first screenshot is the anova table comparing our simple linear model with the enhanced model.
- In the top anova table we see that model 2 has all of the same predictors as model 1 but with an extra predictor being the quadratic term for Alcohol consumption.
- Our null hypothesis for this anova table is that the two models perform equally well in predicting the life expectancy.
- We see that with a p-value = .1135 that at the .05 significance level that there is no significant difference in the performance of Model 1 compared to Model 2 int terms of predicting the life expectancy.
- In the 2nd table we see that we are comparing the original simple linear model with another model containing only 4 variable predictors.
- Here we see that model 2 outperforms model 1 with the highly significant p-value of 2.2e-16.

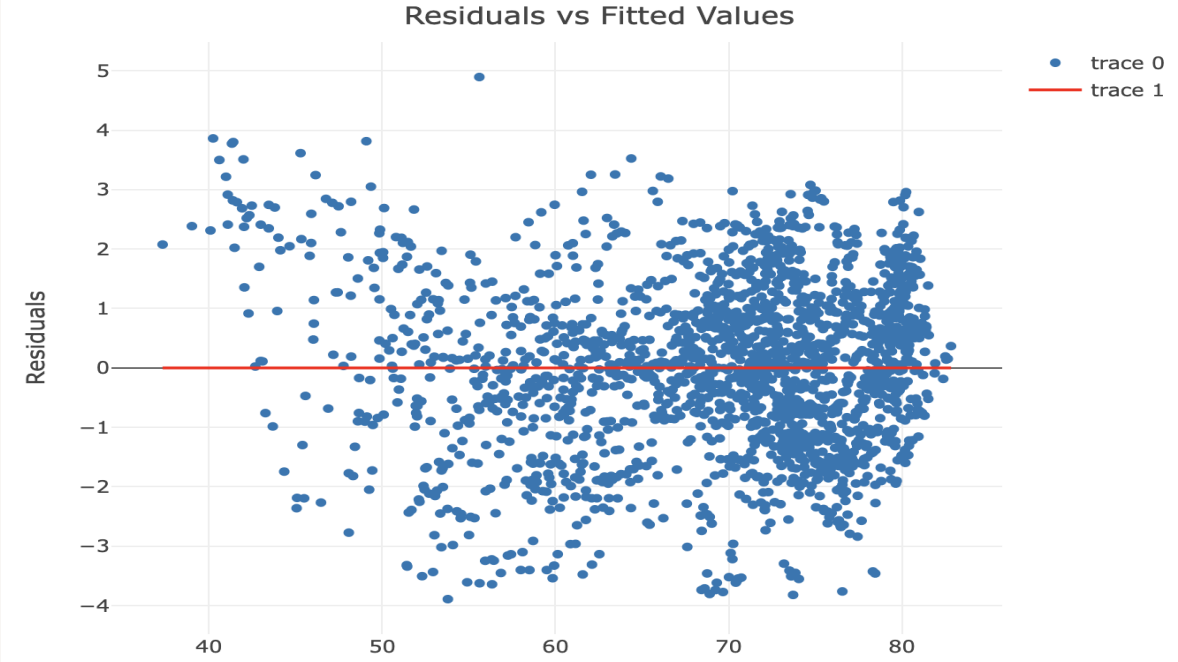
```
## Linear Regression
##
## 1996 samples
##   20 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1798, 1797, 1796, 1796, 1796, 1796, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   1.431348  0.9767833  1.143409
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
##                  Adult_mortality                      Infant_deaths
##                         7.247209                           6.823786
##               Alcohol_consumption                 Under_five_deaths
##                         2.661002                           2.315195
##                             Year           Thinness_five_nine_years
##                         1.205640                           8.755260
##                    Incidents_HIV `Income composition of resources`
##                         2.767337                           1.819810
##               `Total expenditure`       Thinness_ten_nineteen_years
##                         1.278385                           8.802770
##                          Measles           `percentage expenditure`
##                         1.414925                           5.806101
##                              BMI                        Diphtheria
##                         1.821269                           2.709575
##                        Schooling                   GDP_per_capita
##                         3.967074                           5.918139
##                            Polio                      Hepatitis_B
##                         2.434013                           1.770356
##               Population_millions          Economy_status_Developed
##                         1.504514                           2.705831
```

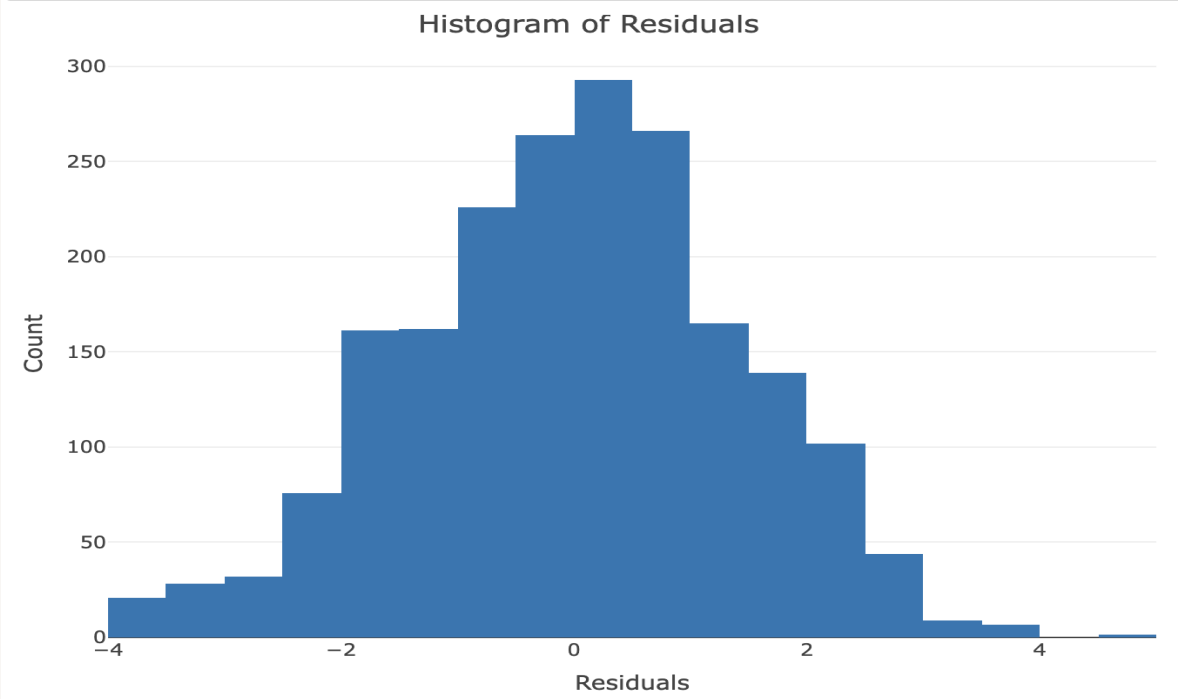
- Here we wanted to test how well our model performed by running it through a 10 fold cross validation.
- We performed the cross validation on our numeric values only.
- We can see that the model performed well based on the RMSE, Rsquared, and MAE values.
- Additionally, here we see the VIF values of our selected variables.
- Most of the variables had low levels but there were a few that had higher levels such as Thinness_ten_nineteen_years, Adult_mortality, and Thinness_five_nine_years.
- One step we could make in our future models could be removing one of the thinness variables.

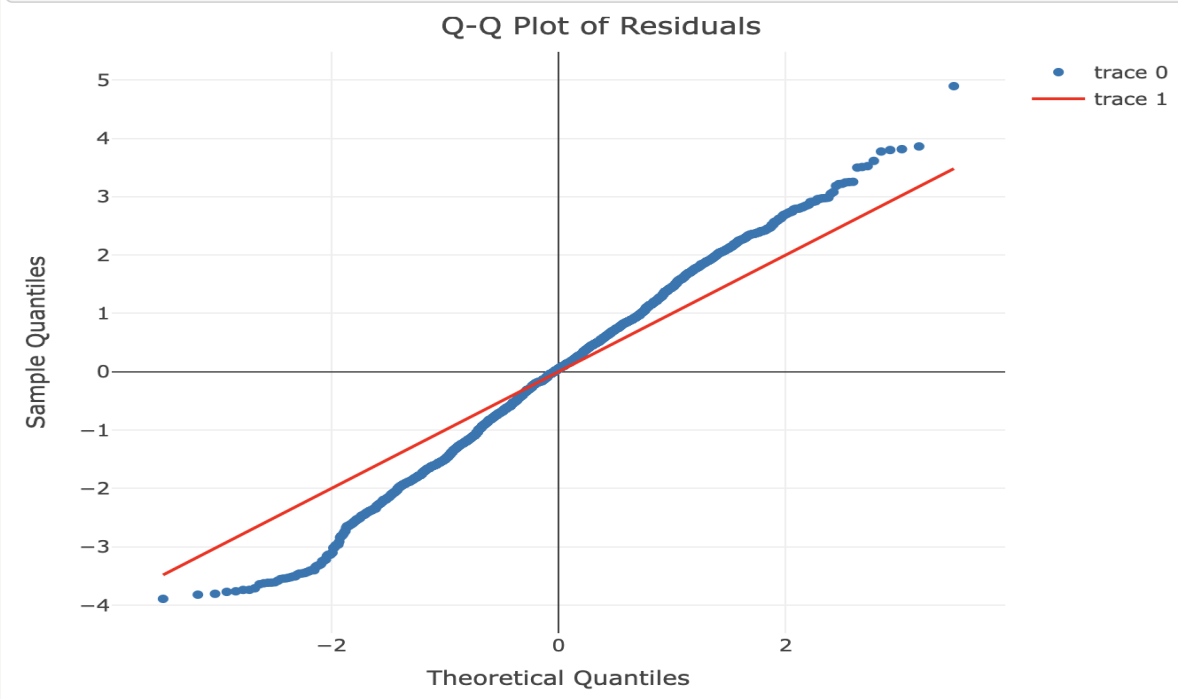


- In the graph to the left we display the points and the Cook's distance along with the influence and leverage of the points.
- We can see we get a good distribution of our points clustered in the area to the left with a few outliers farther out to the right.
- Overall the residuals and distribution of our model looks good and we will continue with our analysis.



- Here we see the various plots of our residual values.
- All of our plots for the model look to confirm the assumptions of normality for our residuals.
- These plots show that there is no evidence of heteroscedasticity since they are normally distributed.





Objective 2: Compare Multiple models

Conclusion

Thank you

Ivan Chavez

Jessica McPhaul

Rafia Mirza