Statistical Analysis of Life Expectancy Dataset

Project by :

- Dejan Dichoski
- Marija Cveevska

Mentor :

- Prof. A. Roverato

University of Padova

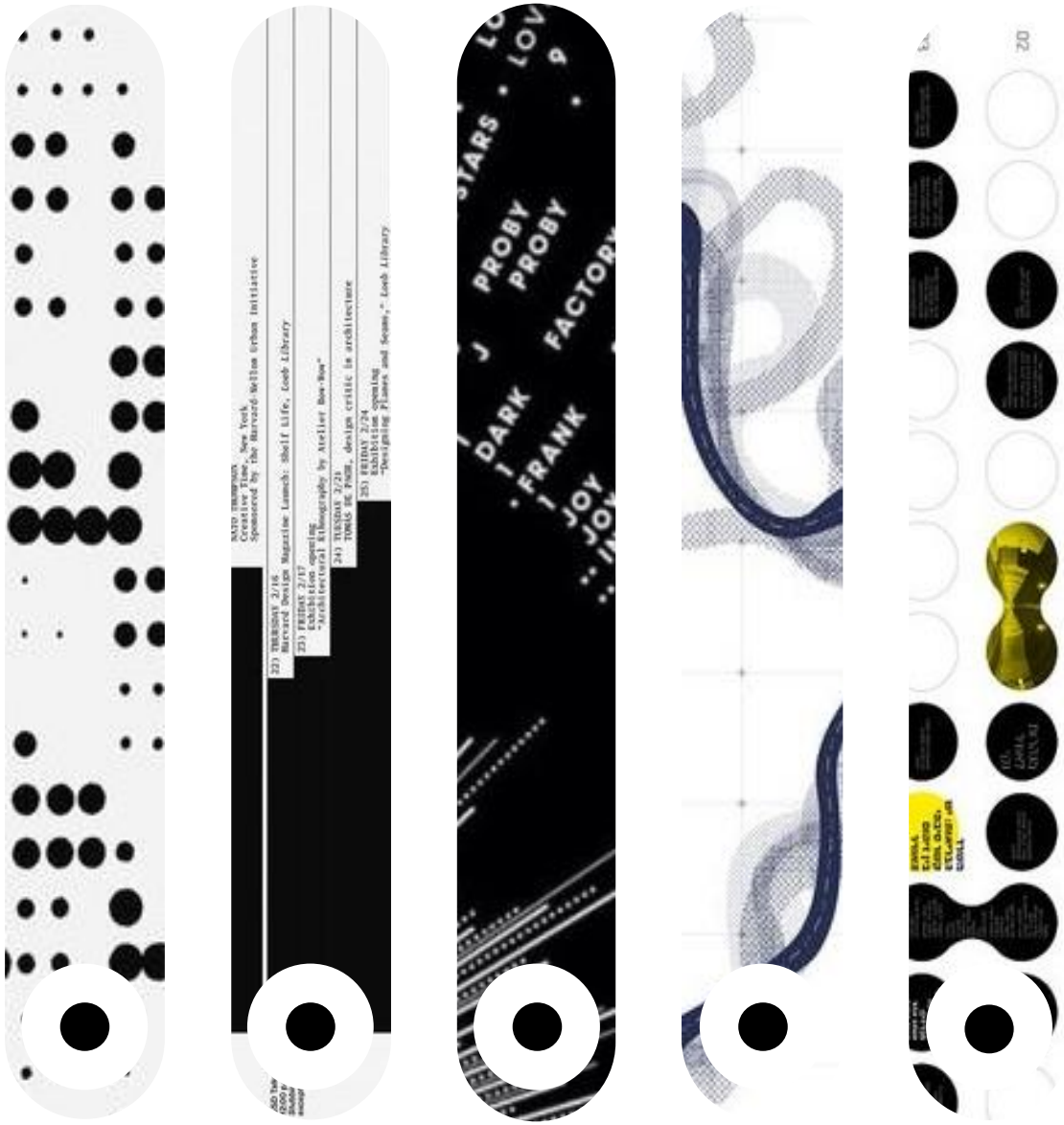
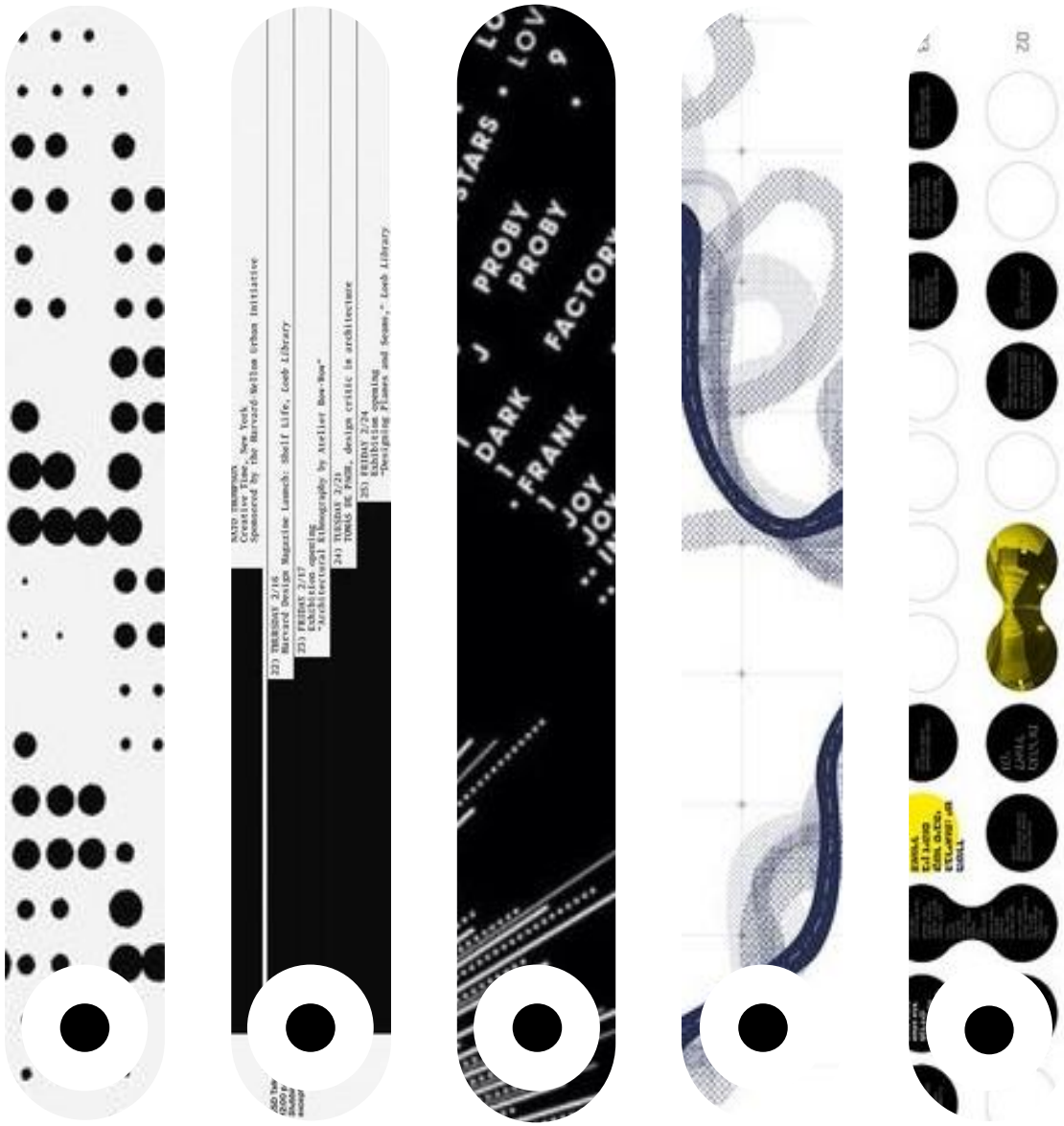
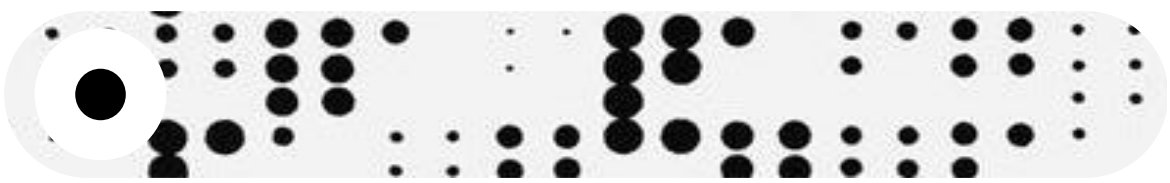


Table of Contents






NATO THOMPSON
Creative Time, New York
Sponsored by the Harvard-Mellon Urban Initiative
22) THURSDAY 2/16
Harvard Design Magazine Launch: Shelf Life, Loeb Library
23) FRIDAY 2/17
Exhibition opening
"Architectural Xilnography by Atelier Bow-Wow"
24) TUESDAY 2/21
TOMAS DE PAOR, design critic in architecture
25) FRIDAY 2/24
Exhibition opening
"Designing Planes and Seams," Loeb Library


DARK
FRANK
JOY
PROBY
PROBY
FACTORY
STARS


02

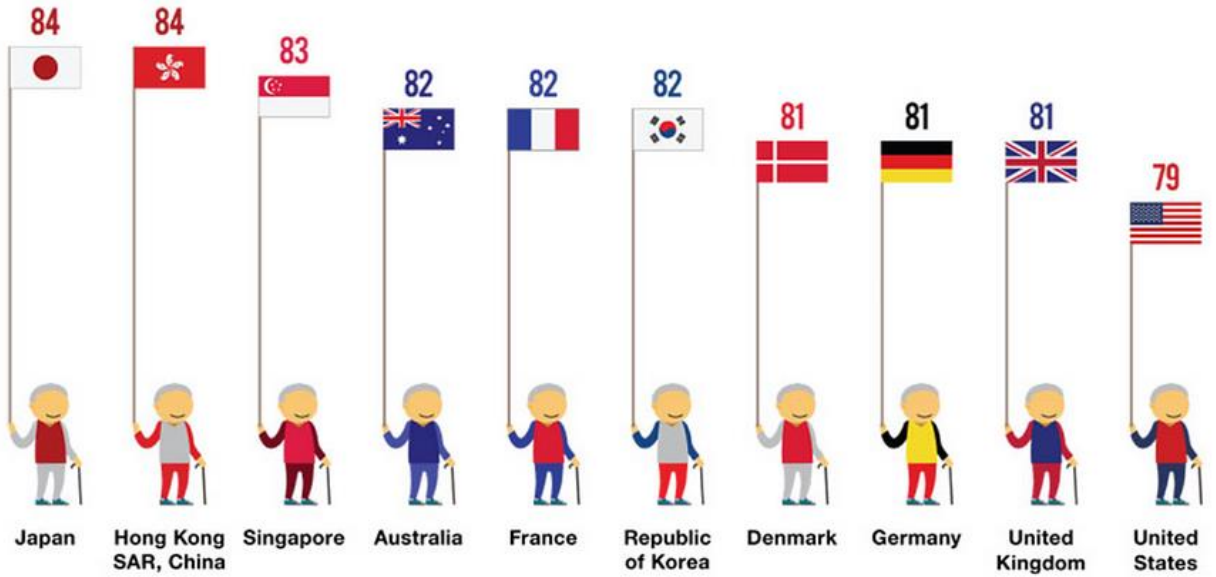
INTRODUCTION

Project Description

Understanding **life expectancy**, which refers to the average duration a person is projected to live, holds significance as it serves as a comprehensive measure of community well-being.

In addition, life expectancy finds various applications in the financial domain, encompassing areas such as life insurance, pension planning, and social security benefits.

- Insurance businesses:
  - Life insurance
  - Pension planning
- Bank loans:
  - Mortgage and home loans
  - Personal loans

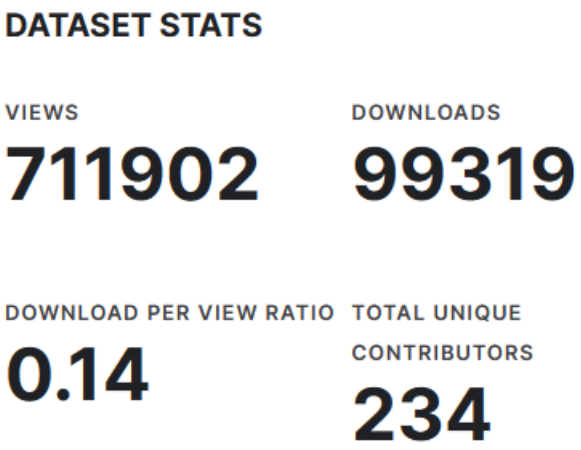


About the Dataset

Life Expectancy (WHO) dataset:

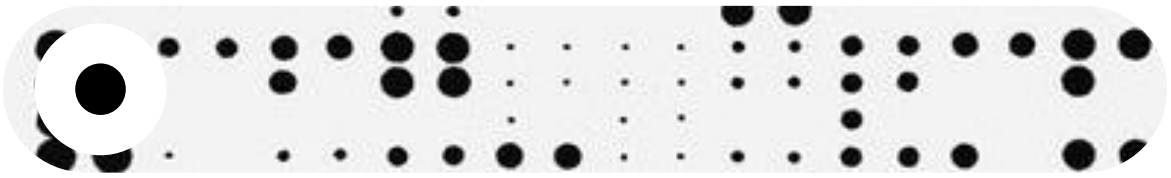
- Obtained from **Kaggle**:
  - collected from the website of WHO
  - economic data collected from UN website

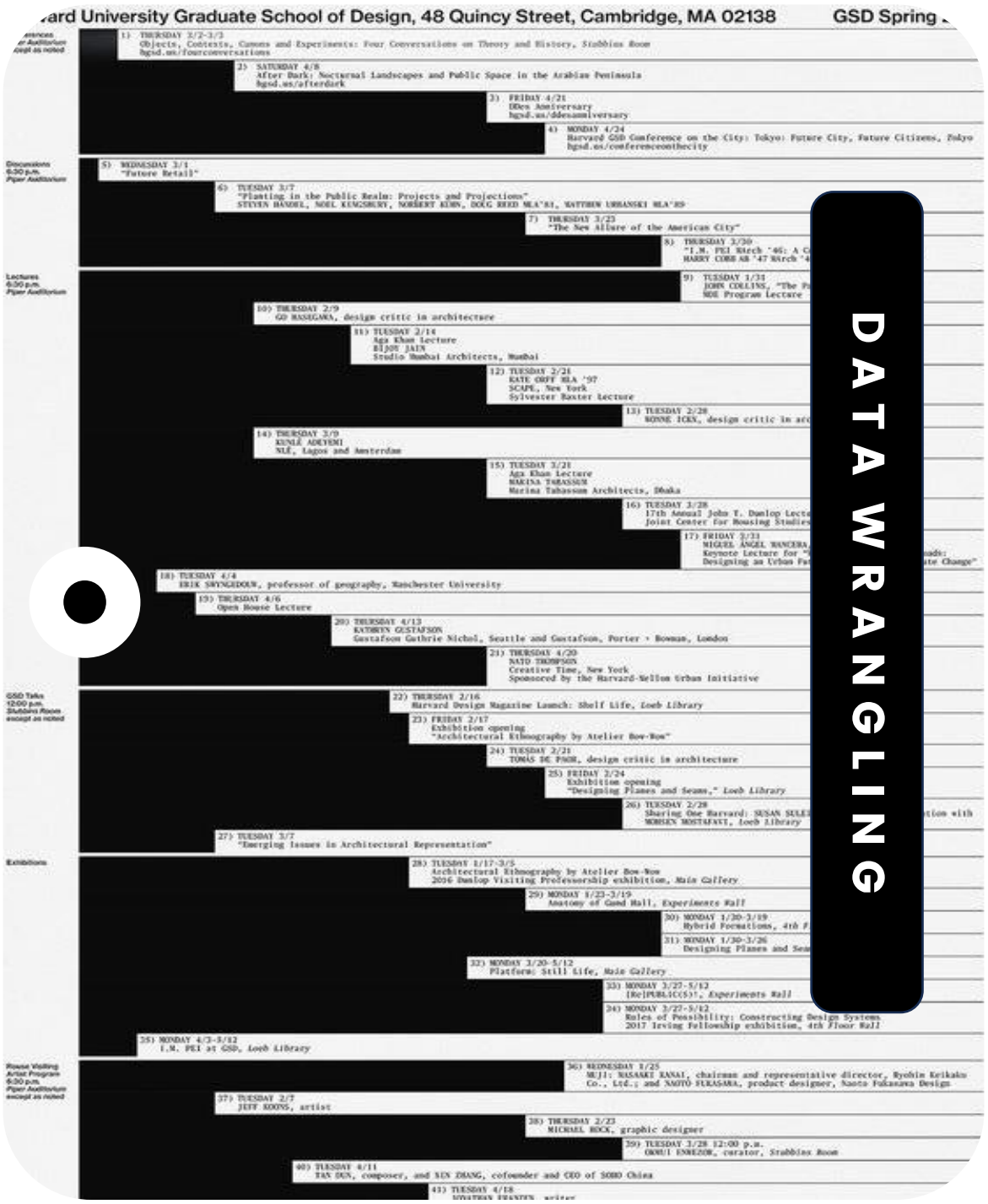
- Data for **193 countries**, in a span of **15 years**
- Number of rows: 2938
- Number of **columns: 22**



Column Name	Type	Description
Country	Nominal	Country
Year	Ordinal	Data is collected from 2000 - 2015 years
Status	Nominal	Developed or Developing status
Life expectancy	Ratio	Life Expectancy in age
Adult Mortality	Ratio	Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
Infant deaths	Ratio	Number of Infant Deaths per 1000 population
Alcohol	Ratio	Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
Percentage expenditure	Ratio	Expenditure on health as a percentage of Gross Domestic Product per capita(%)
Hepatitis B	Ratio	Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
Measles	Ratio	Number of reported cases of Measles per 1000 population
BMI	Ratio	Average Body Mass Index of the entire population
Under-five deaths	Ratio	Number of under-five deaths per 1000 population
Polio	Ratio	Polio (Pol3) immunization coverage among 1-year-olds (%)
Total expenditure	Ratio	General government expenditure on health as a percentage of total government expenditure (%)
Diphtheria	Ratio	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
HIV/AIDS	Ratio	Deaths per 1 000 live births due to HIV/AIDS (0-4 years)
GDP	Ratio	Gross Domestic Product per capita (in USD)
Population	Ratio	Population of the country
Thinness 1-19 years	Ratio	Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
Thinness 5-9 years	Ratio	Prevalence of thinness among children for Age 5 to 9 (%)
Income composition of resources	Ratio	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
Schooling	Ratio	Number of years of Schooling (years)

INTRODUCTION





DATA WRANGLING

Missing values

- **Detection**
  - Null values
  - Check for wrong entries (**inexplicit nulls**)
- **Dealing**
  - Fill with the mean value
  - Dropping
  - **Interpolation**

Inexplicit nulls

	Min	Median	Mean	Max	NA's
Adult.Mortality	1	144	164	723	10
GPD	1.68	1766.95	7483.16	119172.74	448
infant.deaths	0	3	30.3	1800	0
under.five.deaths	0	4	42.04	2500	0
Population	34	1,387,000	12,750,000	1,294,000,000	652
BMI	1	43.50	38.32	87.30	34

Dealing with missing values

```
## [3] Life.expectancy has 10 null values: 0.34% null
## [4] Adult.Mortality has 155 null values: 5.28% null
## [5] infant.deaths has 848 null values: 28.86% null
## [6] Alcohol has 194 null values: 6.6% null
## [8] Hepatitis.B has 553 null values: 18.82% null
## [10] BMI has 1456 null values: 49.56% null
## [11] under.five.deaths has 785 null values: 26.72% null
## [12] Polio has 19 null values: 0.65% null
## [13] Total.expenditure has 226 null values: 7.69% null
## [14] Diphtheria has 19 null values: 0.65% null
## [16] GDP has 448 null values: 15.25% null
## [17] Population has 652 null values: 22.19% null
## [18] thinness.10.19.years has 34 null values: 1.16% null
## [19] thinness.5.9.years has 34 null values: 1.16% null
## [20] Income.composition.of.resources has 167 null values: 5.68% null
## [21] Schooling has 163 null values: 5.55% null
## Out of 22 total columns, 16 contain null values; 72.73% columns contain null values.
```

drop

imputation by year

Outlier analysis

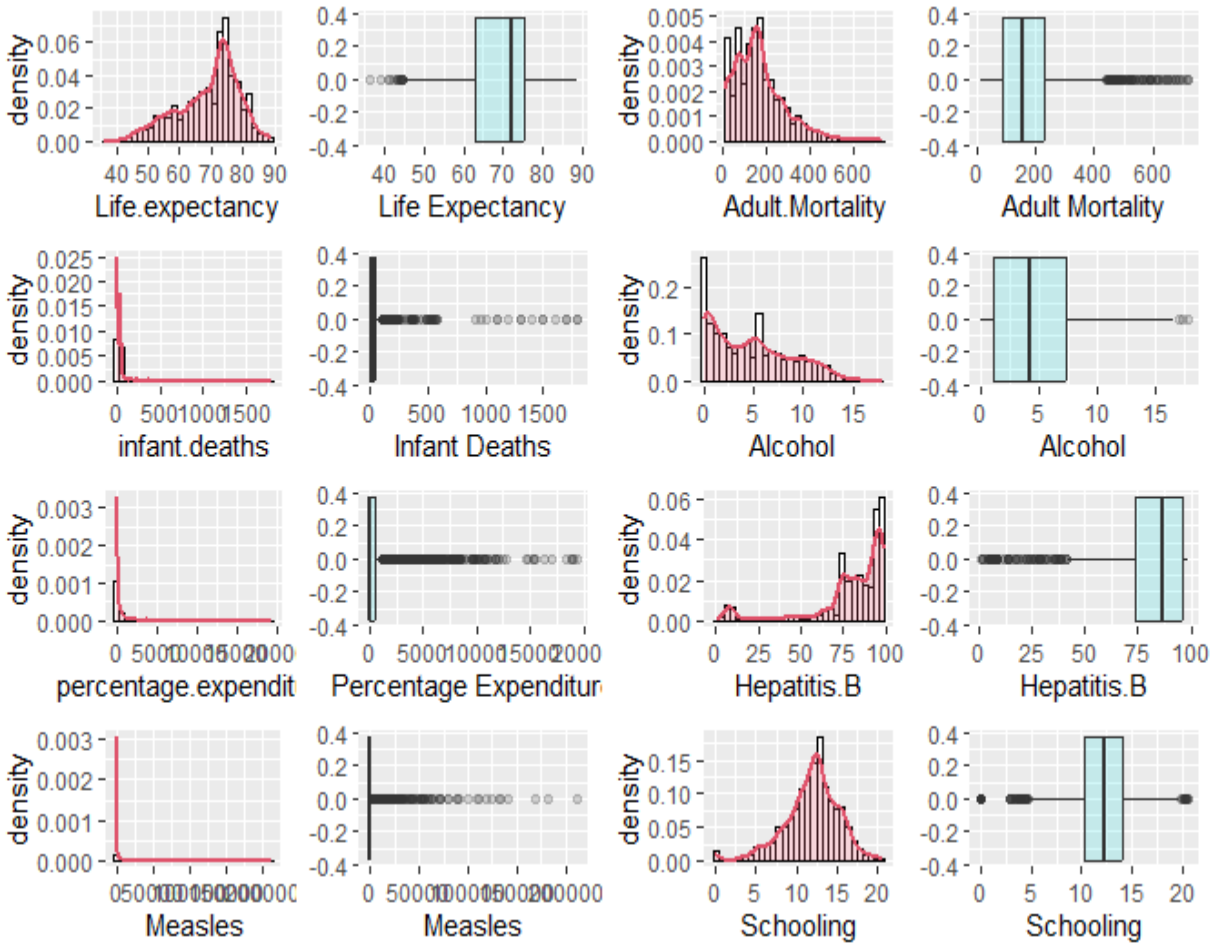
Detection

- Visualization
- Tukey's method

Treatment

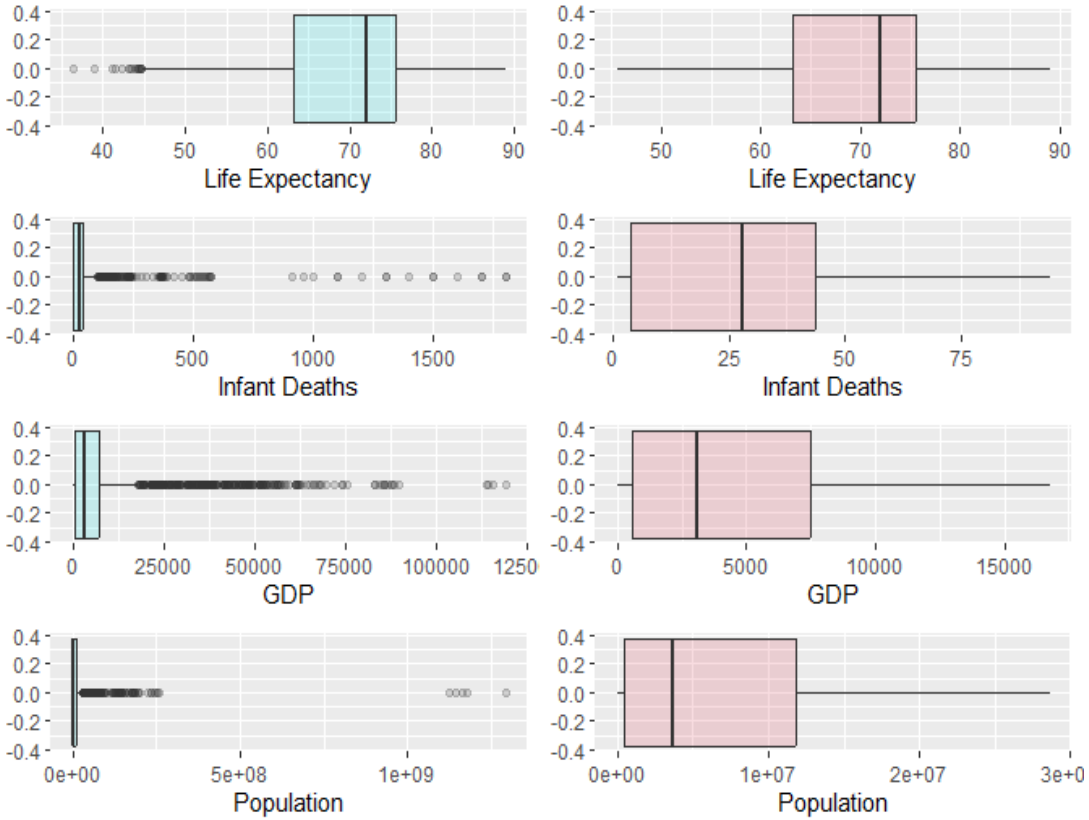
- ❑ Discard outliers
- ✓ Apply boundaries (Winsorization)
- ❑ Data transformation

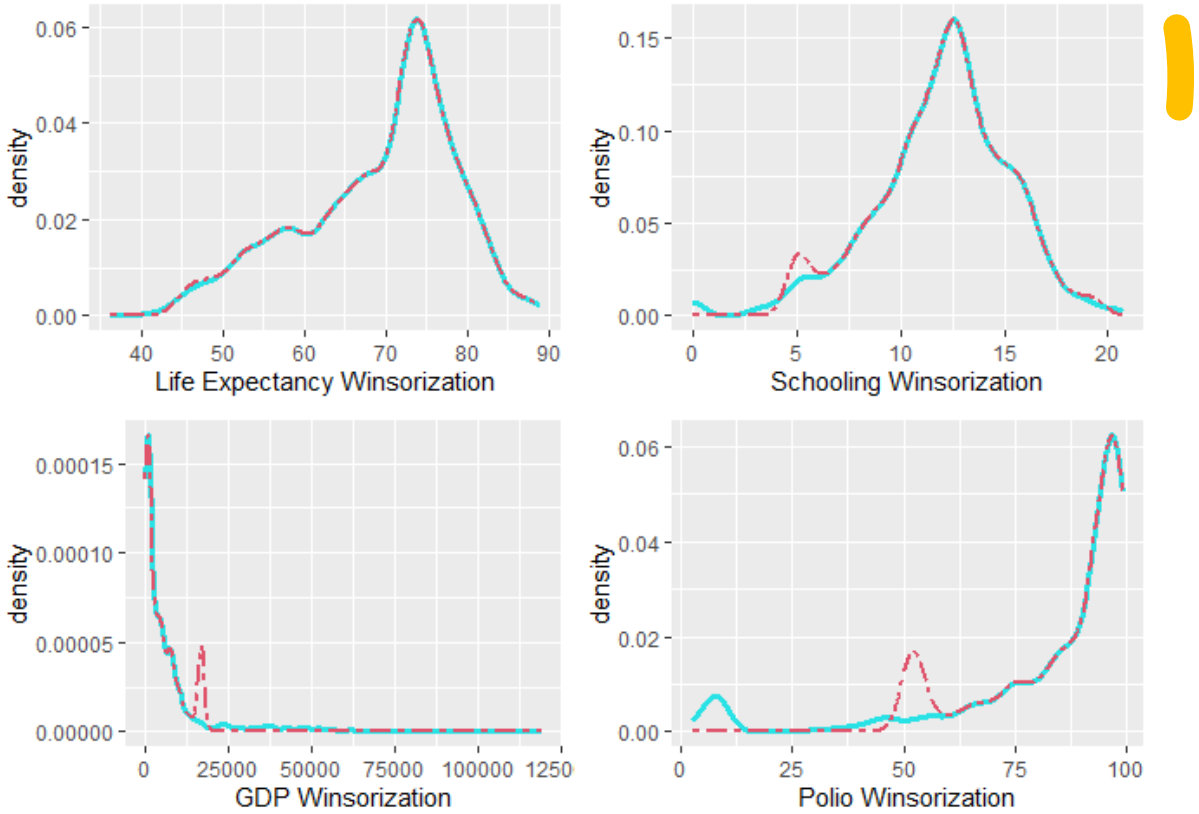
Outlier detection

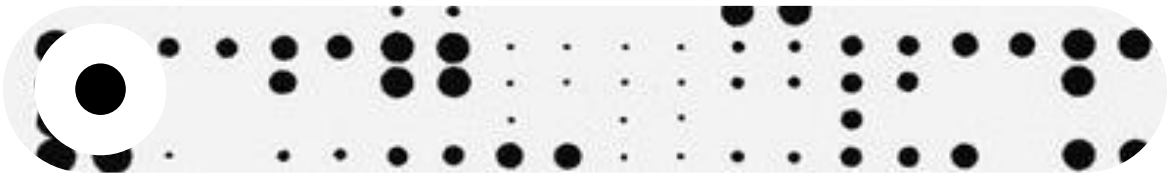


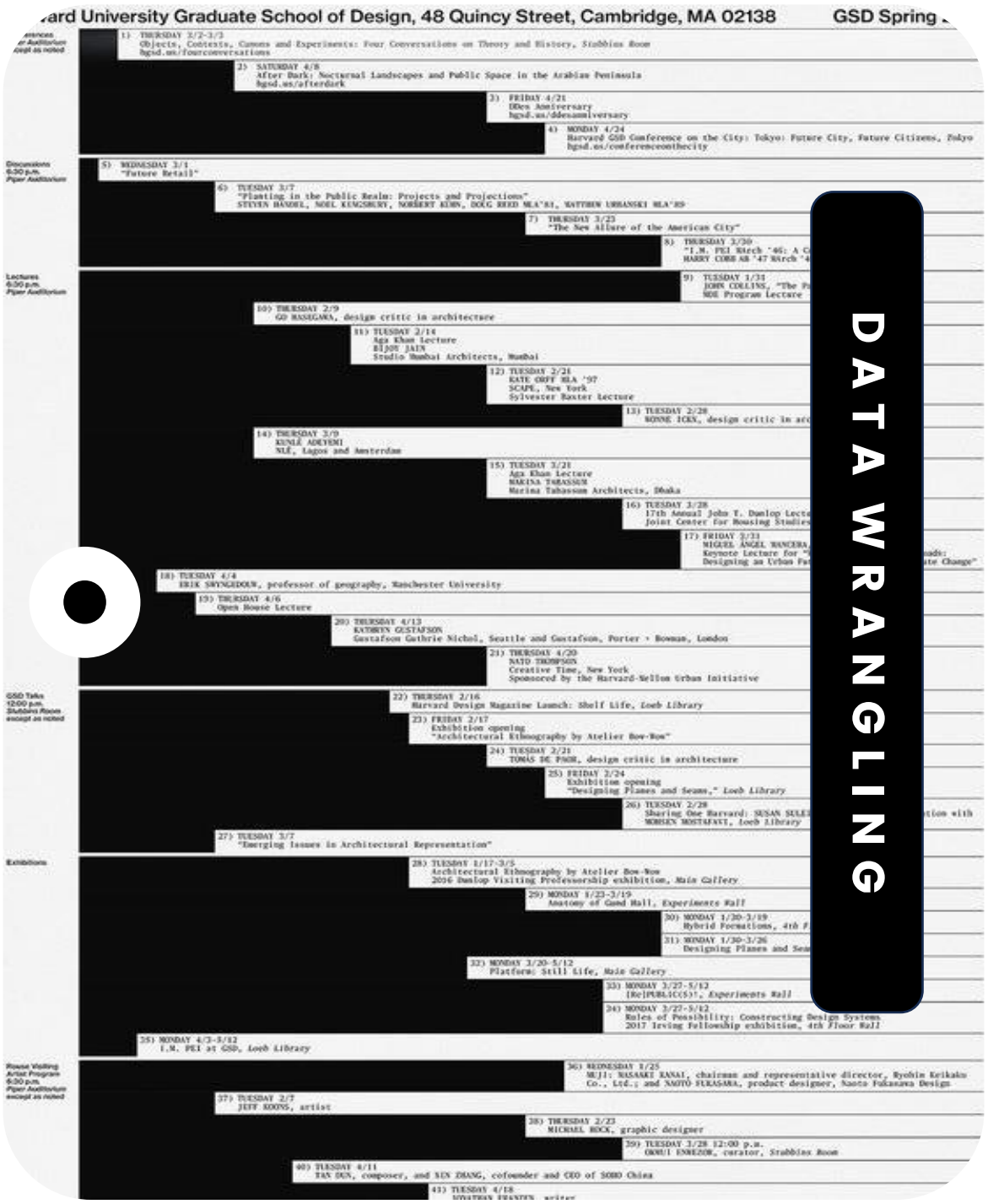
	OutlierCount	OutlierPercent
Measles	542	18.4479238
HIV.AIDS	542	18.4479238
percentage.expenditure	389	13.2402995
GDP	300	10.2110279
Diphtheria	298	10.1429544
Polio	279	9.4962560
Hepatitis.B	222	7.5561607
Population	203	6.9094622
under.five.deaths	142	4.8332199
infant.deaths	135	4.5949626
Income.composition.of.resources	130	4.4247788
thinness.10.19.years	100	3.4036760
thinness.5.9.years	99	3.3696392
Adult.Mortality	97	3.3015657
Schooling	77	2.6208305
Total.expenditure	51	1.7358747
Life.expectancy	17	0.5786249
Alcohol	3	0.1021103

Outlier treatment

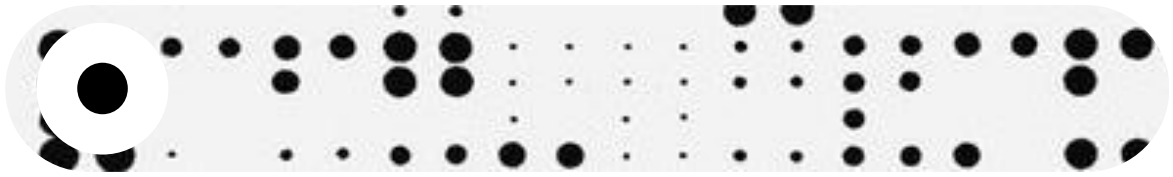




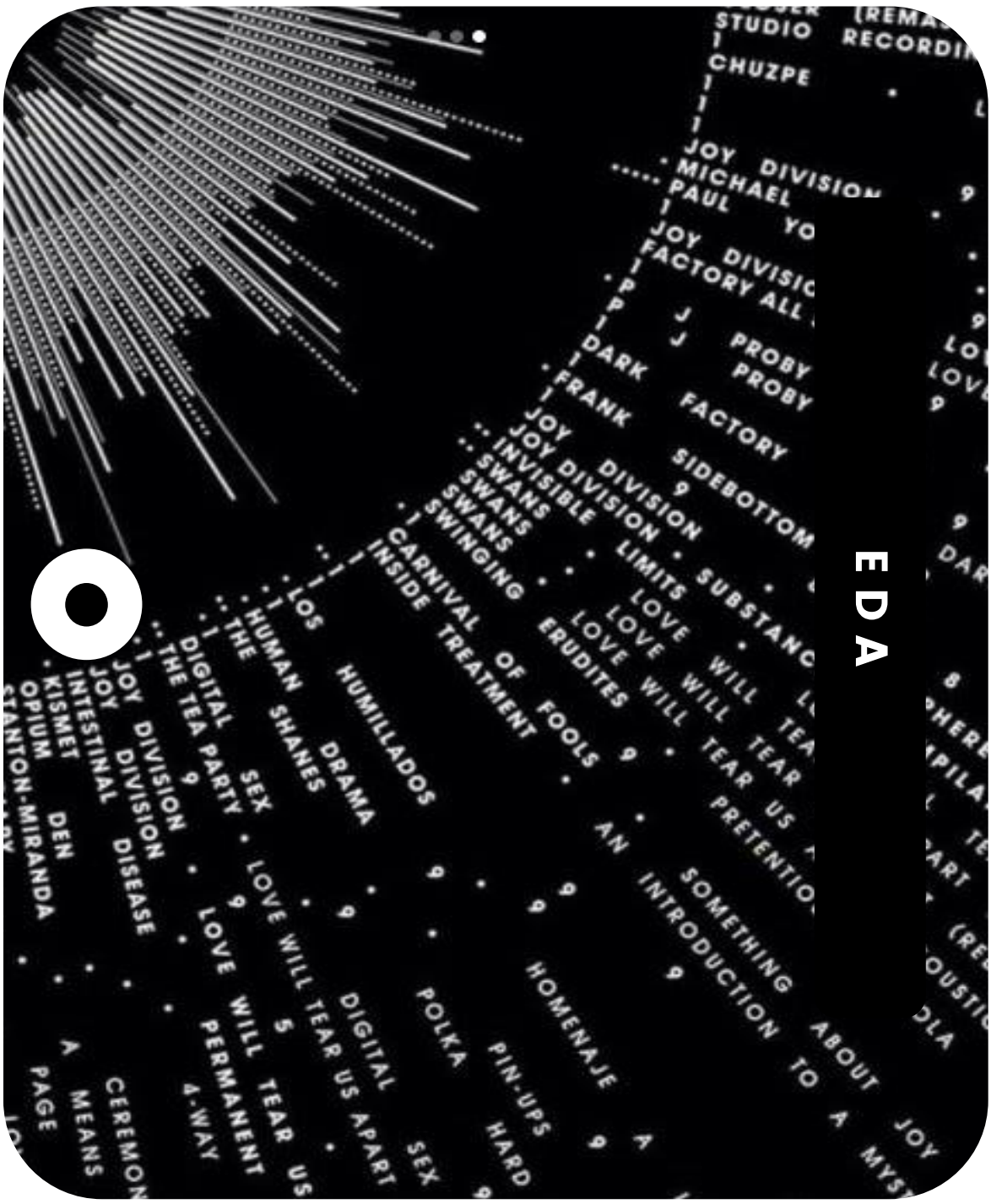




DATA WRANGLING




NATO THOMPSON
Creative Time, New York
Sponsored by the Harvard-Wellon Urban Initiative
22) THURSDAY 2/16
Harvard Design Magazine Launch: Shelf Life, Loeb Library
23) FRIDAY 2/17
Exhibition opening
"Architectural Kibnography by Atelier Bow-Wow"
24) TUESDAY 2/21
TOMAS DE PAUW, design critic in architecture
25) FRIDAY 2/24
Exhibition opening
"Designing Planes and Seams," Loeb Library


STUDIO
RECORDI
CHUZPE
JOY DIVISION
MICHAEL
PAUL
JOY DIVISIO
FACTORY ALL
J J
PROBY
PROBY
DARK
FACTORY
FRANK
SIDEBOTTOM
JOY DIVISION
JOY DIVISION
INVISIBLE
SWANS
SWANS
SWANS
SWINGING
CARNIVAL OF FOOLS
INSIDE TREATMENT
LOS HUMILLADOS
HUMAN SHANES
DRAMA
THE SEX
DIGITAL
THE TEA PARTY
JOY DIVISION
JOY DIVISION
INTESTINAL
KISMET
OPIUM
DEN
STANTON-MIRANDA
DISEASE
LIMITS
LOVE WILL TEAR US
LOVE WILL TEAR
LOVE WILL TEAR
ERUDITES
LOVE WILL TEAR US
PRETENTIO
SUBSTANC
AN INTRODUCTION TO A
SOMETHING ABOUT JOY
HOMENAJE A
PIN-UPS
HARD
POLKA
SEX
DIGITAL
LOVE WILL TEAR US APART
LOVE WILL TEAR US
PERMANENT
4-WAY
CEREMON
A MEANS
PAGE
EDA


02

Exploratory Data Analysis

Goal: extract valuable insights from the data

- Uncovering patterns
- Inspiring new directions
- Detecting errors
- Assessing assumptions

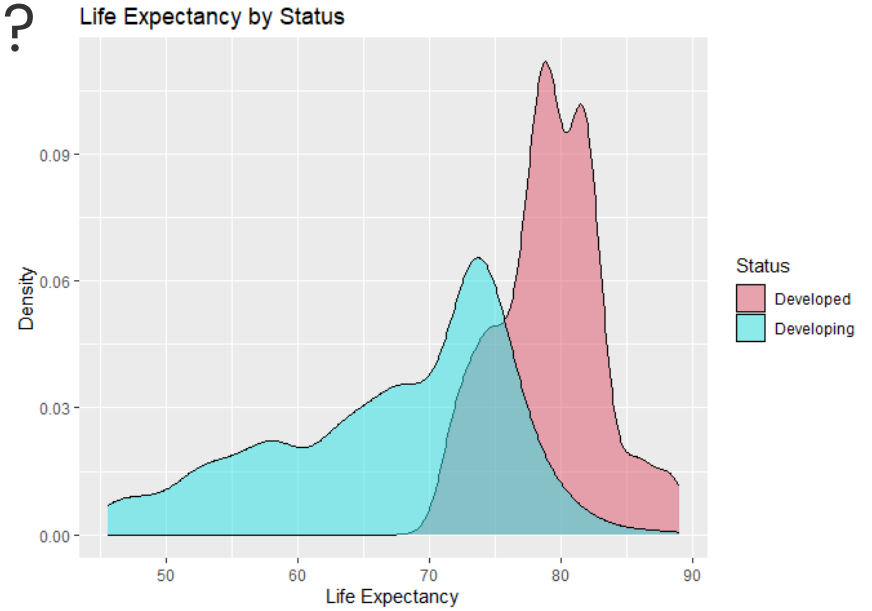
Relevant questions:

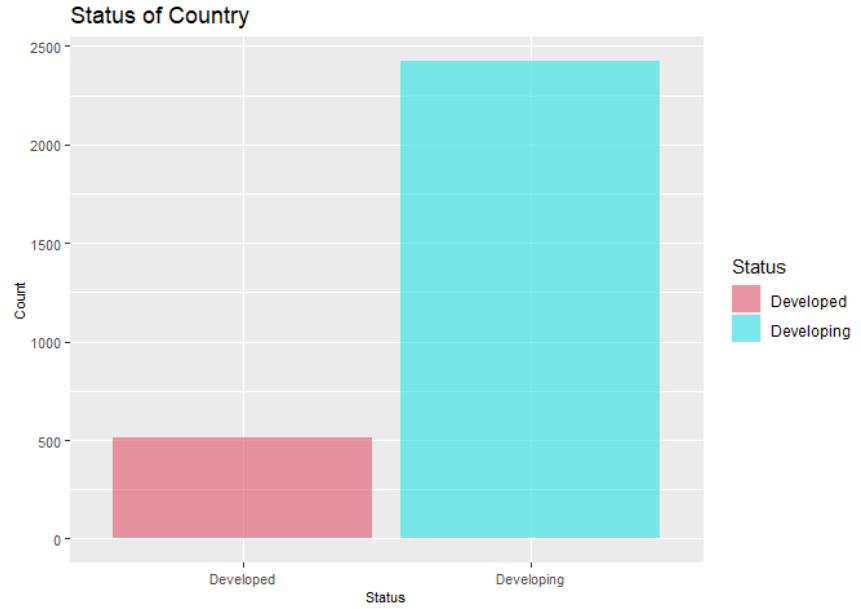
1. *Can we say that Developed countries have more average life expectancy than Developing countries?*
2. *Is there a statistically significant relationship between the average number of schooling years and life expectancy?*
3. *Check if countries that spend a higher proportion of their resources on human development have a higher life expectancy?*
4. *Italian Government has claimed that they have spent an average of around 8.41% of their total expenditure on health for the year 2000–2015. Can we test their claim?*
5. *What is the correlation of Life expectancy with Alcohol drinking habits?*
6. *Correlation between Life Expectancy and Immunization.*

Q1. Can we say that Developed countries have higher average life expectancy than Developing countries?

```
 F test to compare two variances

data:  Developed_Y$Life.expectancy and Developing_X$Life.expectancy
F = 0.14793, num df = 31, denom df = 160, p-value = 4.472e-08
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.08968881 0.27049424
sample estimates:
ratio of variances
         0.1479263
```

```
 Welch Two Sample t-test

data:  Developed_Y$Life.expectancy and Developing_X$Life.expectancy
t = 13.541, df = 126.15, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 10.36548      Inf
sample estimates:
mean of x mean of y
 79.19785  67.38706
```





Q2. Is there a statistically significant relationship between the average number of schooling years and life expectancy?

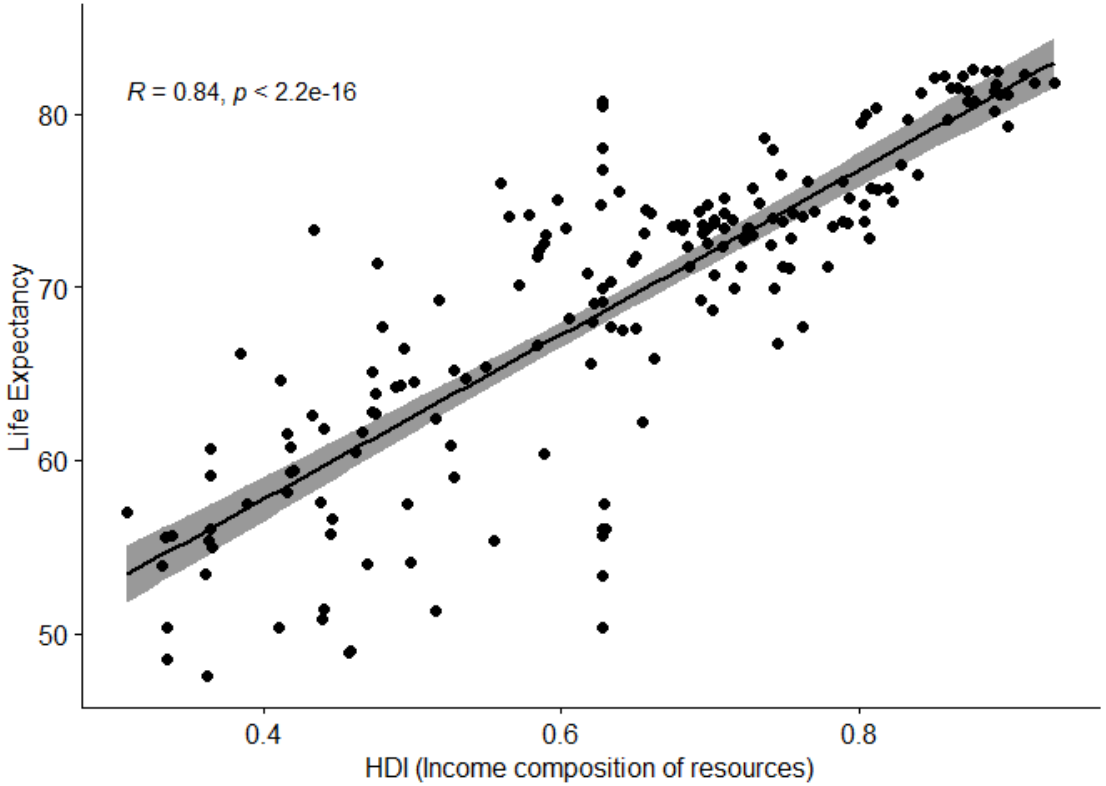
- We used the **ANOVA** test to test the significance of education on life expectancy.
- Countries were categorized into one of the three categories, depending upon the country's average schooling years:
  - '**Low**' (≤ 8),
  - '**Medium**' (> 8 and ≤ 12)
  - '**High**' (> 12)

```
Anova_Results <- aov(Life.expectancy ~ Education, data = Schooling_Y)
summary(Anova_Results)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Education     2    391  195.33   2.977 0.0535 .
## Residuals   176  11549   65.62
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value is slightly above the significance threshold of 0.05.

Q3. Check if countries that spend a higher proportion of their resources on human development have a higher life expectancy?



Pearson's product-moment correlation

Correlation coefficient = **0.838** => strong positive linear relationship between the variables being correlated.

Q4. Italian Government has claimed that they have spent an average of around 8.41% of their total expenditure on health for the year 2000–2015. Can we test their claim?

- Used **one-Sample t-test** to test this claim. We decided to also check **India**'s claim of 5.2 % and make a comparison.

```
 One Sample t-test

data:  Italy_Y
t = 1.5893, df = 15, p-value = 0.1328
alternative hypothesis: true mean is not equal to 8.41
95 percent confidence interval:
 8.320883 9.021617
sample estimates:
mean of x
  8.67125
```

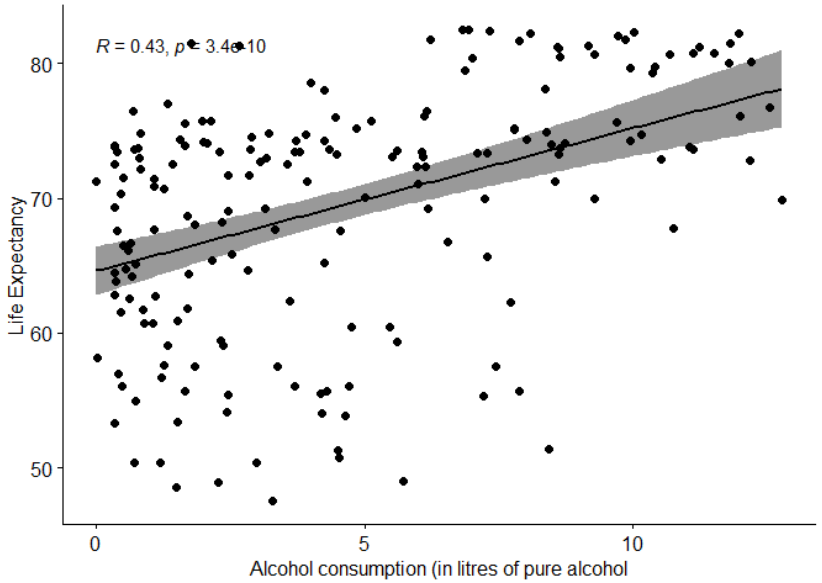
Italy's claim of 8.4 % lies in the 95% CI range.

```
 One Sample t-test

data:  India_Y
t = -3.881, df = 15, p-value = 0.001477
alternative hypothesis: true mean is not equal to 5.2
95 percent confidence interval:
 4.160101 4.897399
sample estimates:
mean of x
  4.52875
```

India's claim of 5.2 % doesn't lie in the 95% CI range.

Q5. What is the correlation of Life expectancy with Alcohol drinking habits?

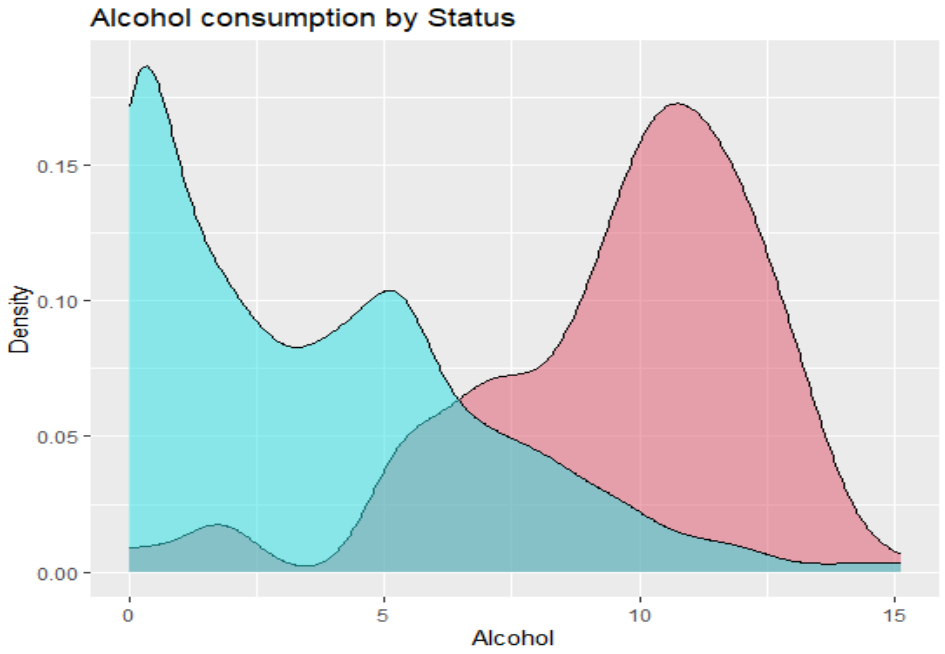
```
 Pearson's product-moment correlation

data:  Alc_X$Alcohol and Alc_X$Life.expectancy
t = 6.6274, df = 191, p-value = 3.397e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3100919 0.5406181
sample estimates:
      cor
0.4323943
```

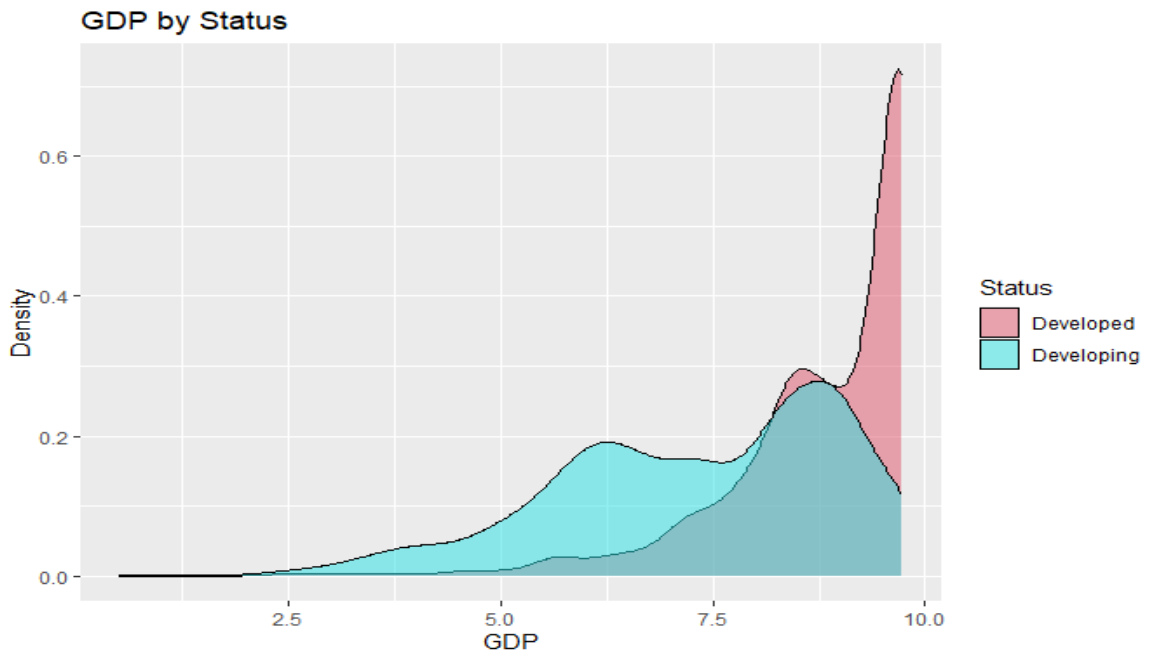


Pearson's product-moment correlation

Correlation coefficient = **0.43** => moderate positive linear relationship between the variables being correlated.

Q5. What is the correlation of Life expectancy with Alcohol drinking habits?





Developed countries tend to have higher levels of alcohol consumption compared to developing countries. This can be attributed to various factors:
higher income levels, greater access to alcohol and more established alcohol industries.

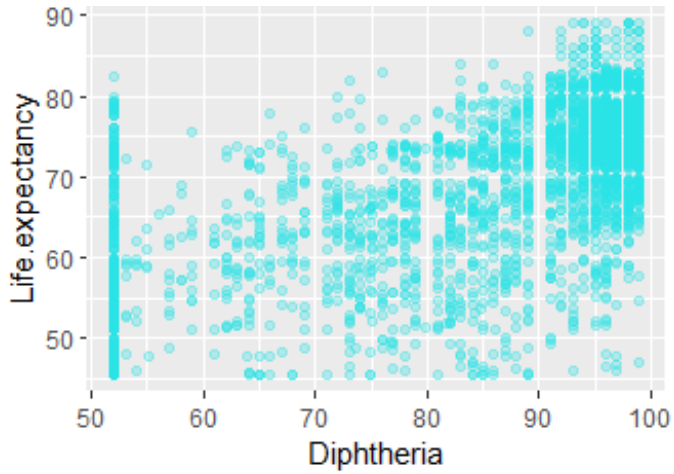
Higher GDP at Developed countries can influence alcohol consumption.
As countries experience economic growth and an increase in GDP, is often an associated rise in income levels and increased spending power.

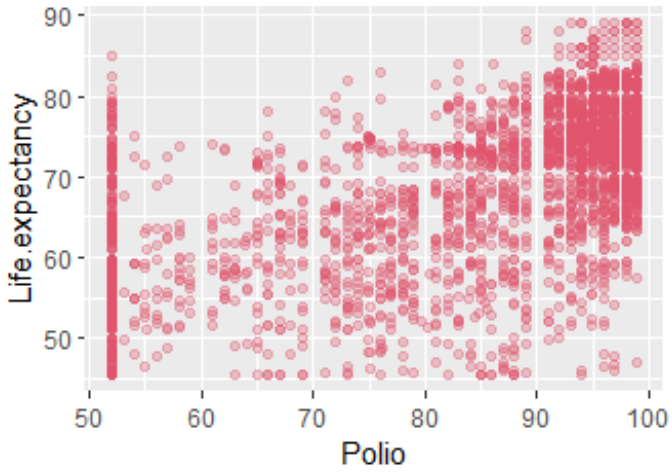
Q6. Correlation between Life Expectancy and Immunization.

- Does immunization against Polio and Diphtheria have a significant effect on life expectancy? We will use a **two-way ANOVA test**.
- If the immunization coverage ≥ than the median value: '**High**' else '**Low**'.

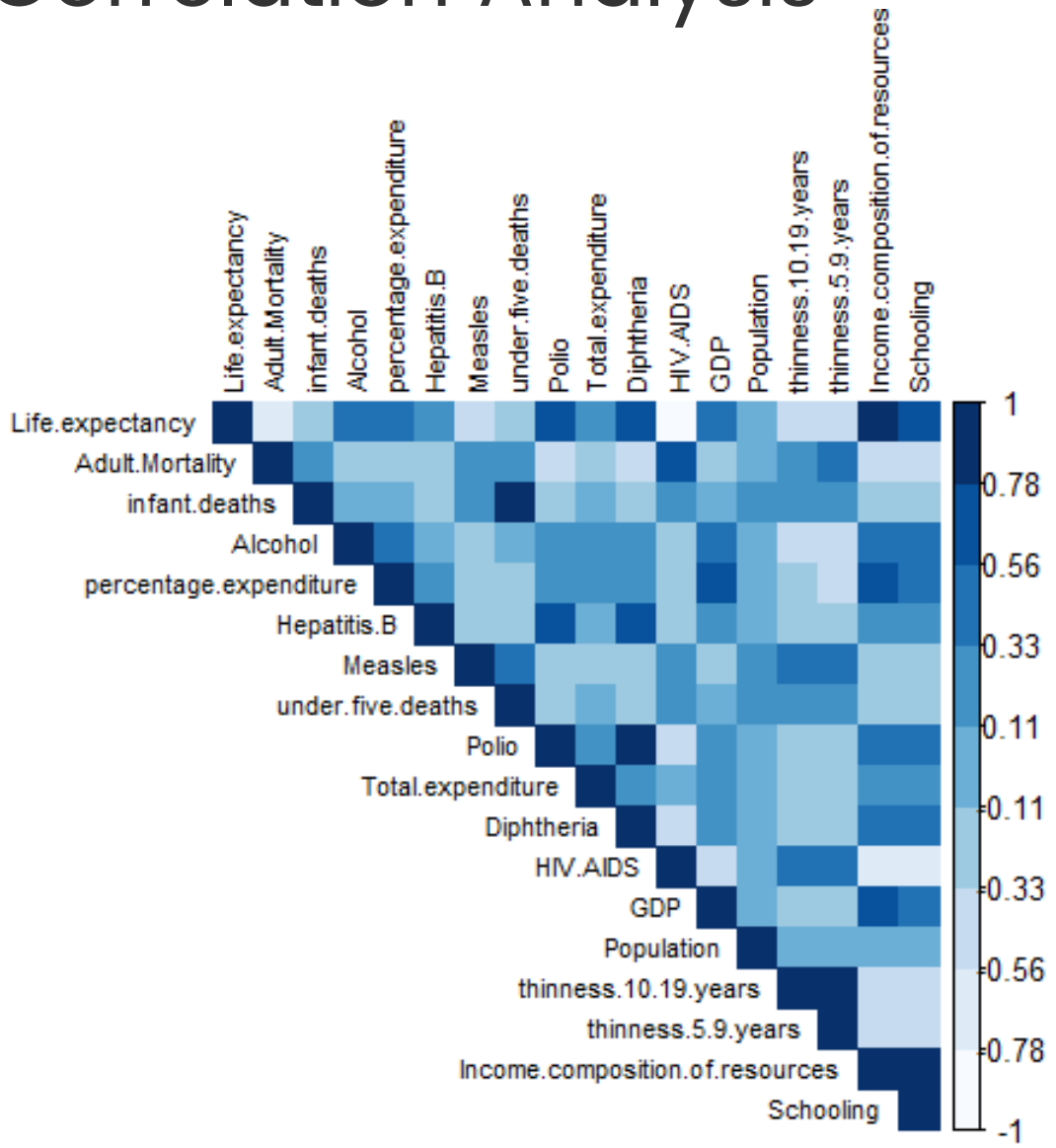
```
Anova_Results_1 <- aov(Life.expectancy.x ~ Polio + Diphtheria, data = Immun_Y)
summary(Anova_Results_1)
```

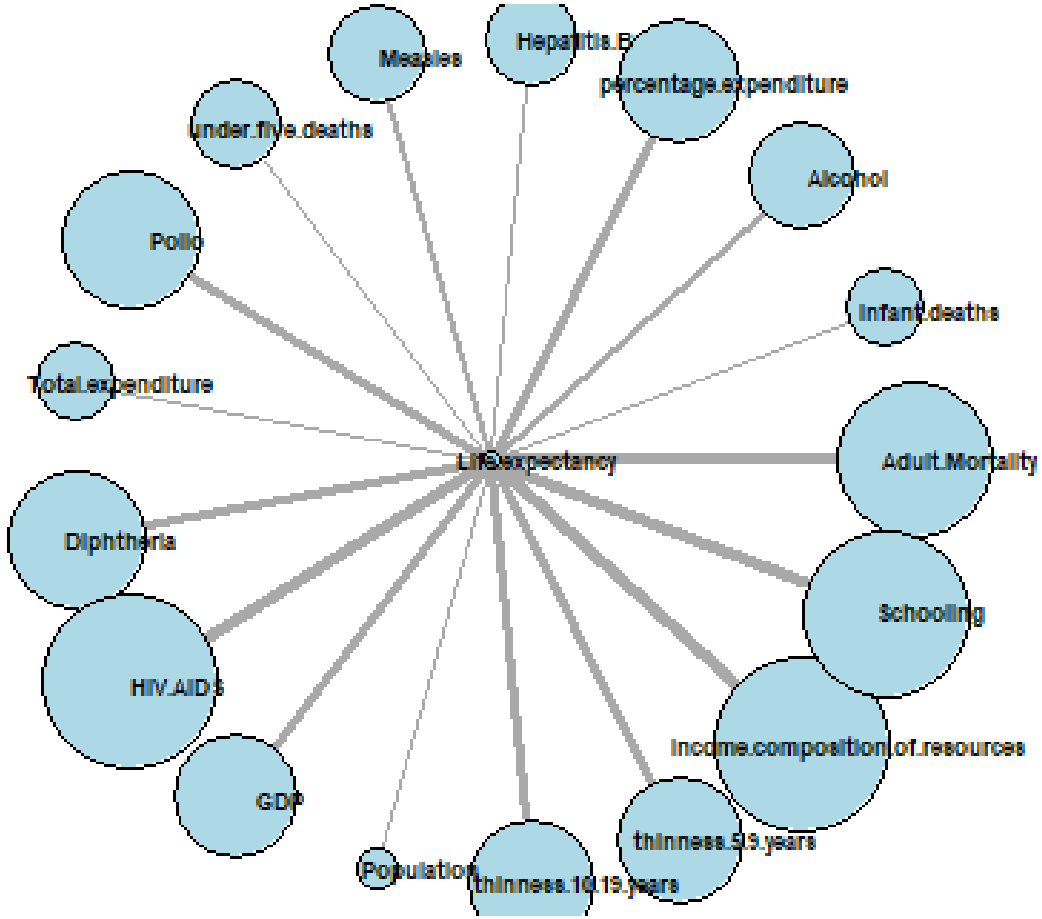
```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Polio         1   5597    5597  117.46  < 2e-16 ***
## Diphtheria    1    631     631   13.25 0.000352 ***
## Residuals   190   9053      48
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

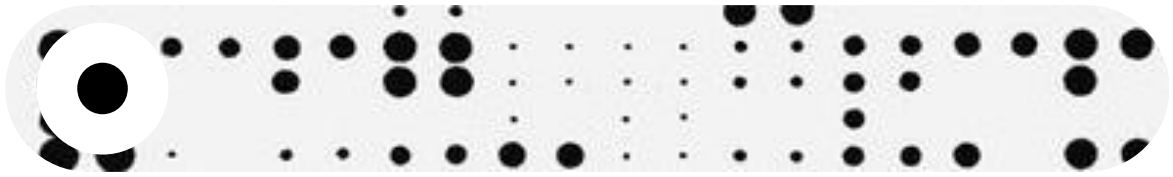




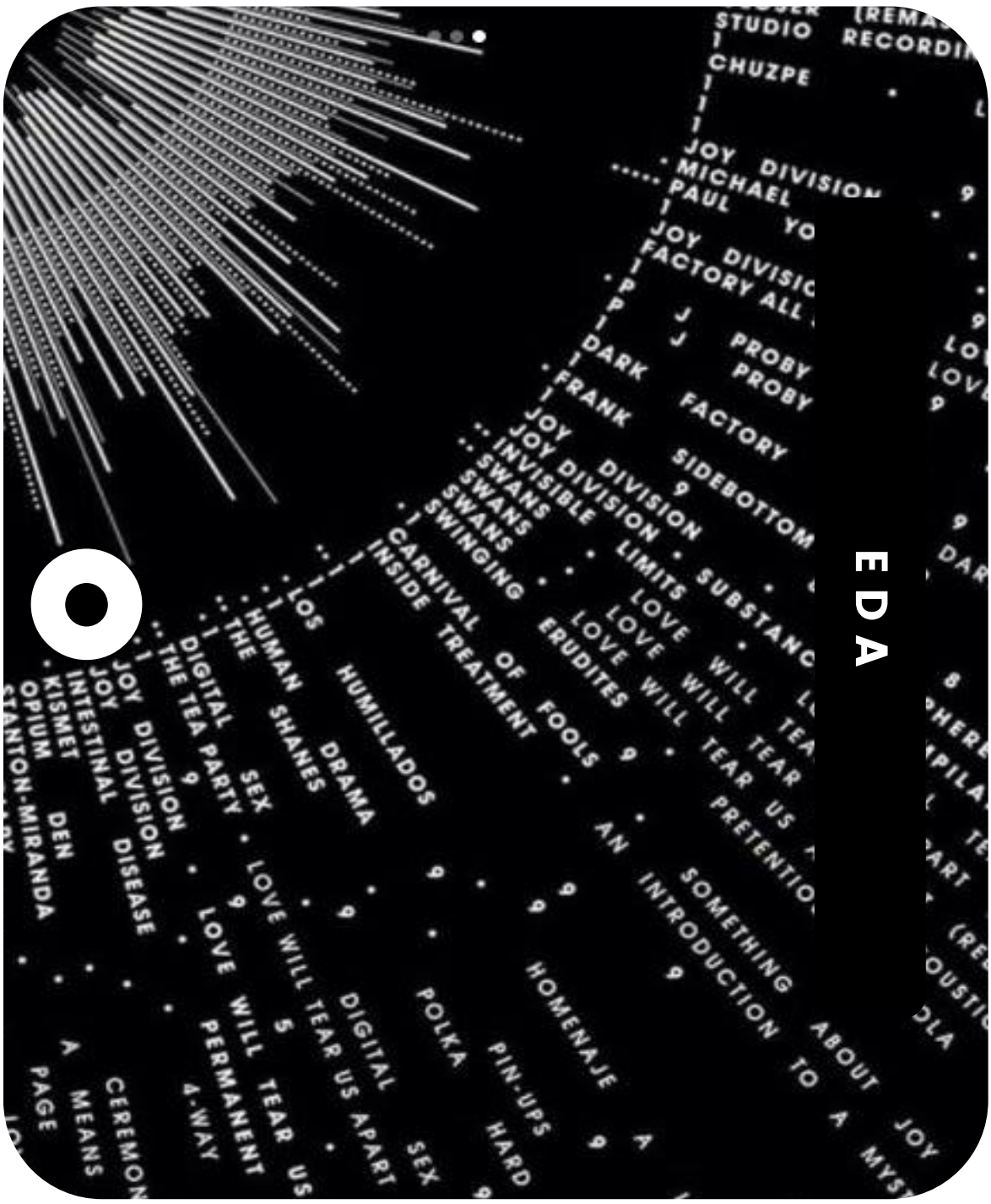
Correlation Analysis



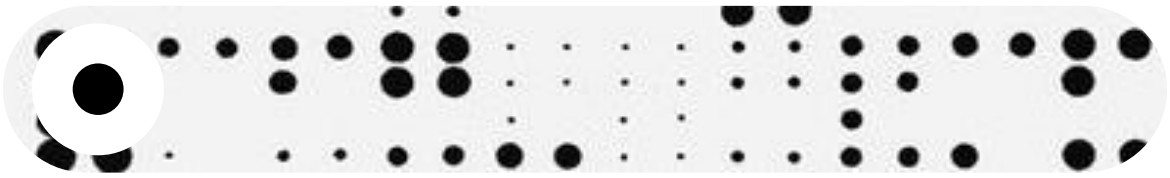


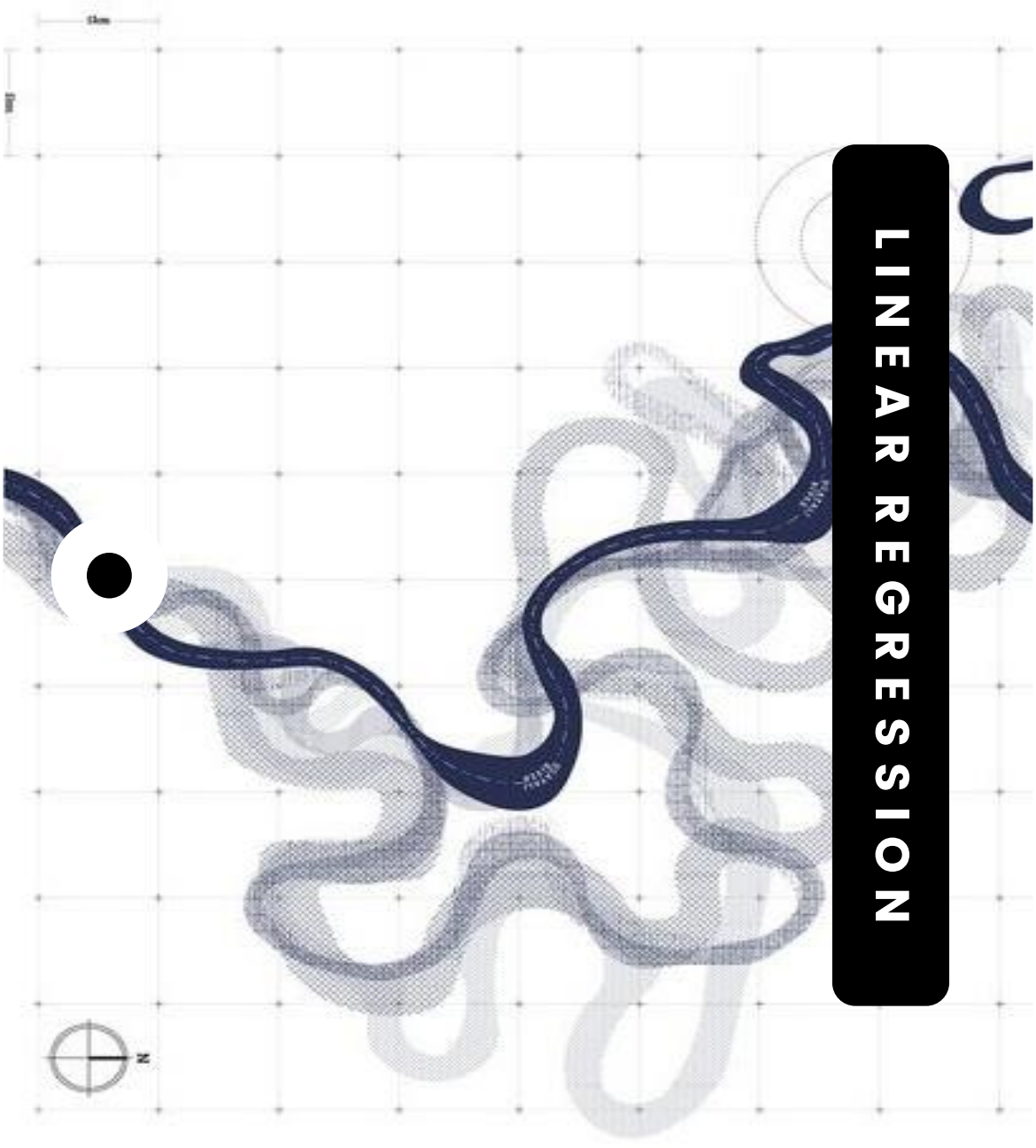



NATO THOMPSON
Creative Time, New York
Sponsored by the Harvard-Wellon Urban Initiative
22) THURSDAY 2/16
Harvard Design Magazine Launch: Shelf Life, Loeb Library
23) FRIDAY 2/17
Exhibition opening
"Architectural Kibnography by Atelier Bow-Wow"
24) TUESDAY 2/21
TOMAS DE PAUW, design critic in architecture
25) FRIDAY 2/24
Exhibition opening
"Designing Planes and Seams," Loeb Library


STUDIO
RECORDI
CHUZPE
JOY DIVISION
MICHAEL
PAUL
JOY DIVISIO
FACTORY ALL
J J
PROBY
PROBY
DARK
FACTORY
FRANK
SIDEBOTTOM
JOY DIVISION
JOY DIVISION
INVISIBLE
SWANS
SWANS
SWANS
SWINGING
CARNIVAL OF FOOLS
INSIDE TREATMENT
LOS HUMILLADOS
HUMAN SHANES
DRAMA
THE SEX
DIGITAL
THE TEA PARTY
JOY DIVISION
JOY DIVISION
INTESTINAL
KISMET
OPIUM
DEN
STANTON-MIRANDA
DISEASE
LIMITS
LOVE WILL TEAR US
LOVE WILL TEAR
LOVE WILL TEAR
ERUDITES
LOVE WILL TEAR US
PRETENTIO
SUBSTANC
AN INTRODUCTION TO A
SOMETHING ABOUT JOY
HOMENAJE A
PIN-UPS
HARD
POLKA
SEX
DIGITAL
LOVE WILL TEAR US APART
LOVE WILL TEAR US
PERMANENT
4-WAY
CEREMON
A MEANS
PAGE
EDA

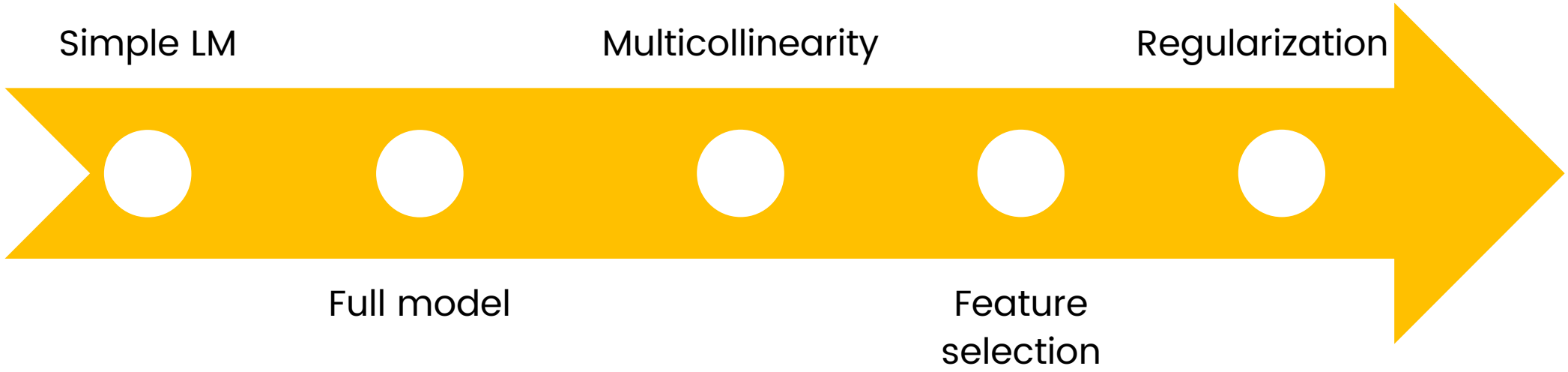

02





LINEAR REGRESSION

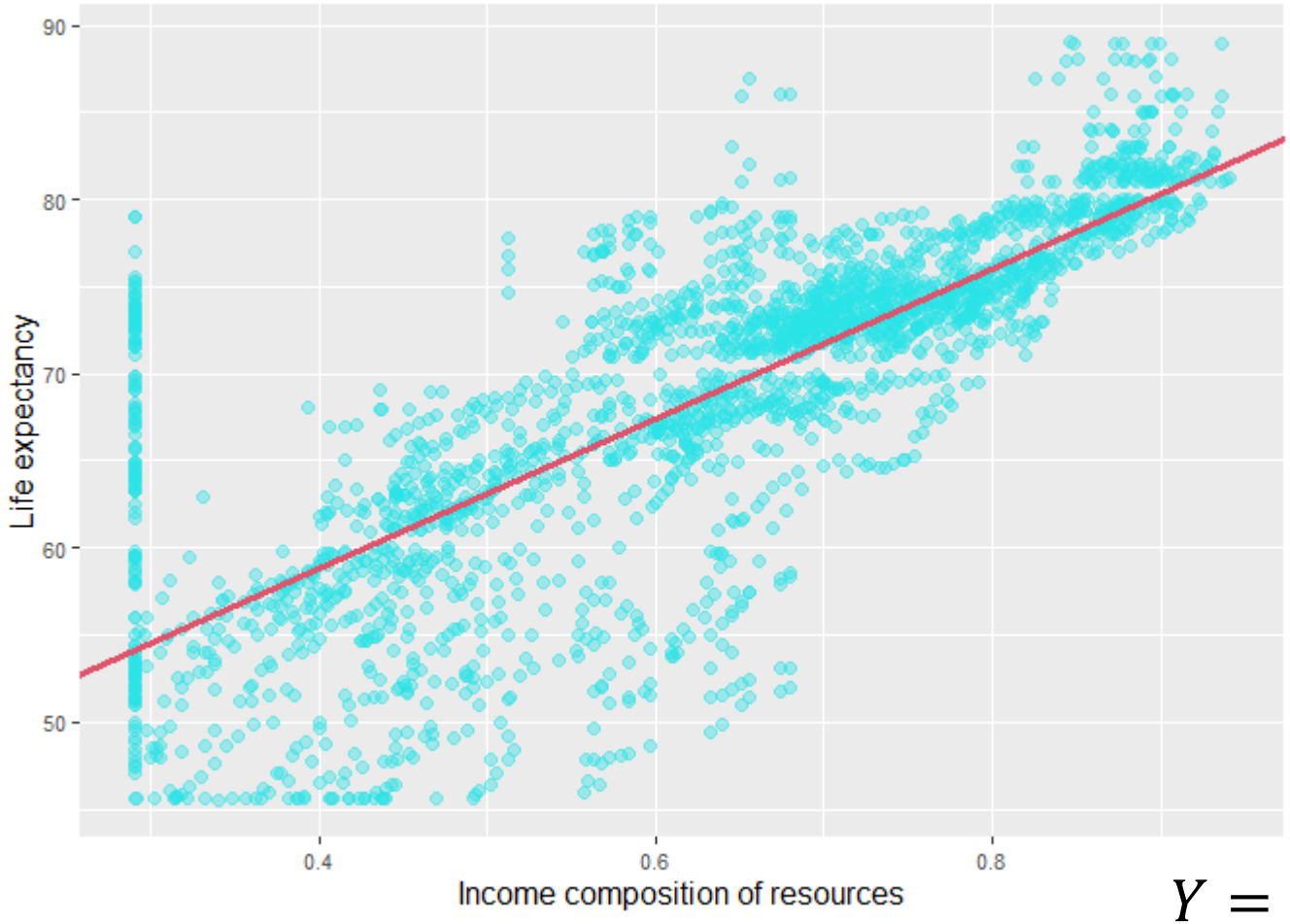
Linear Regression



check Linear regression assumptions

Simple Linear Regression

Adj R^2 : 0.6071



$$Y = \beta_0 + \beta_1 \cdot x + \varepsilon$$

Life.expectancy = 42.9837 · Income.composition.of.resources + 41.6751

Multiple Linear Regression

Full Model

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.501  -2.142   0.054   2.097  14.789
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     6.289e+01  8.827e-01  71.246  < 2e-16 ***
## StatusDeveloping               -1.559e+00  2.815e-01  -5.538 3.40e-08 ***
## Adult.Mortality                -1.631e-02  8.921e-04 -18.286  < 2e-16 ***
## infant.deaths                   1.327e-04  1.231e-02   0.011  0.99140
## Alcohol                         4.322e-02  2.751e-02   1.571  0.11632
## percentage.expenditure          2.052e-03  3.343e-04   6.137 9.82e-10 ***
## Hepatitis.B                    -3.905e-02  7.169e-03  -5.447 5.64e-08 ***
## Measles                        -1.206e-03  2.740e-04  -4.402 1.12e-05 ***
## under.five.deaths              -1.238e-02  8.604e-03  -1.439  0.15024
## Polio                           2.736e-02  9.939e-03   2.753  0.00595 **
## Total.expenditure               4.768e-02  3.651e-02   1.306  0.19165
## Diphtheria                      6.145e-02  1.023e-02   6.006 2.19e-09 ***
## HIV.AIDS                       -5.658e+00  1.626e-01 -34.797  < 2e-16 ***
## GDP                            -1.102e-05  2.357e-05  -0.468  0.64008
## Population                      1.119e-08  9.796e-09   1.142  0.25363
## thinness.10.19.years            1.066e-01  5.833e-02   1.827  0.06782 .
## thinness.5.9.years             -2.276e-01  5.734e-02  -3.969 7.43e-05 ***
## Income.composition.of.resources 1.223e+01  9.915e-01  12.331  < 2e-16 ***
## Schooling                       1.086e-01  5.413e-02   2.006  0.04502 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.743 on 2360 degrees of freedom
## Multiple R-squared:  0.8506, Adjusted R-squared:  0.8495
## F-statistic: 746.5 on 18 and 2360 DF,  p-value: < 2.2e-16
```

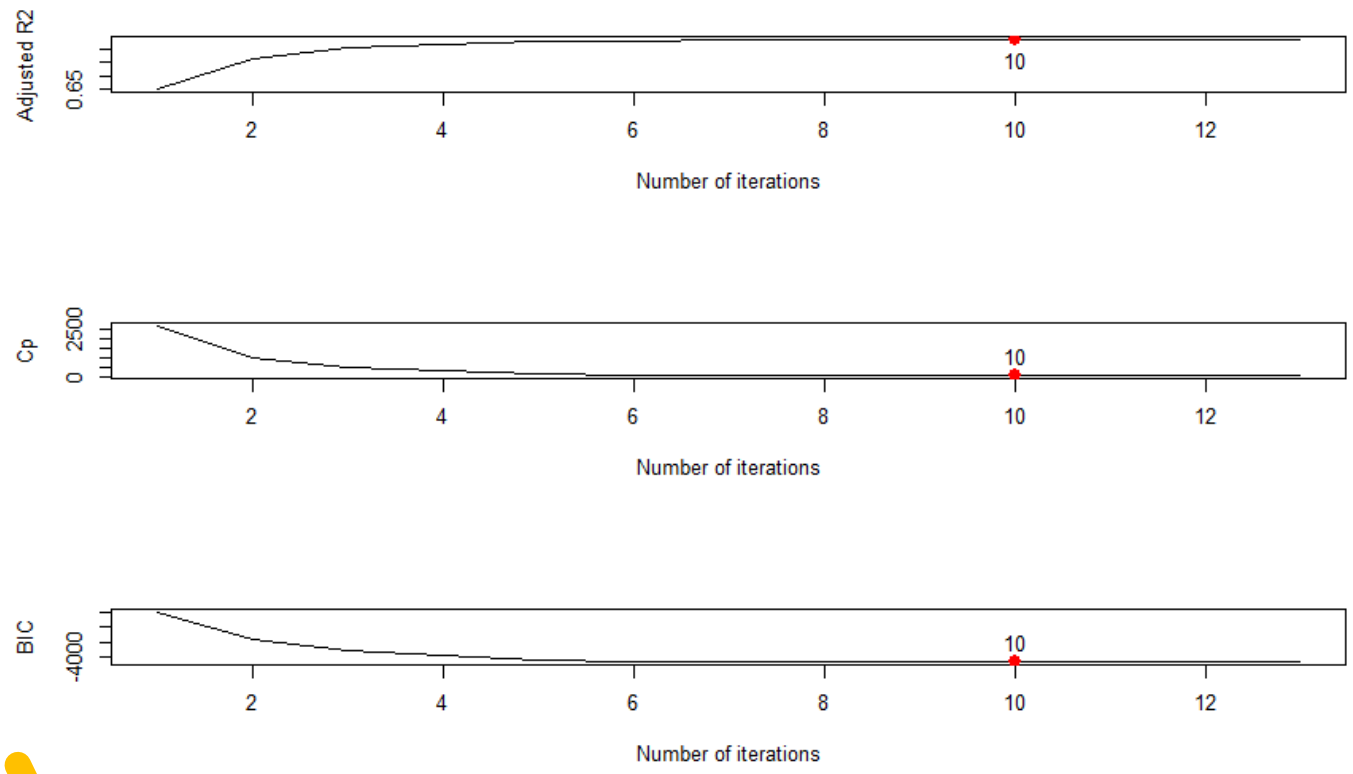
Adj R²: 0.8495

Multicollinearity check

Adj R²: 0.8343

	VIF
under.five.deaths	18.741587
infant.deaths	17.574420
thinness.10.19.years	9.144831
thinness.5.9.years	9.014553
Income.composition.of.resources	5.103967
Schooling	4.828482
Diphtheria	4.223635
Polio	3.963522
percentage.expenditure	2.818810
GDP	2.754848
HIV.AIDS	2.156189
Alcohol	2.043490
Status	1.941817
Hepatitis.B	1.648238
Adult.Mortality	1.638067
Measles	1.437989
Total.expenditure	1.215365
Population	1.159776

Feature selection

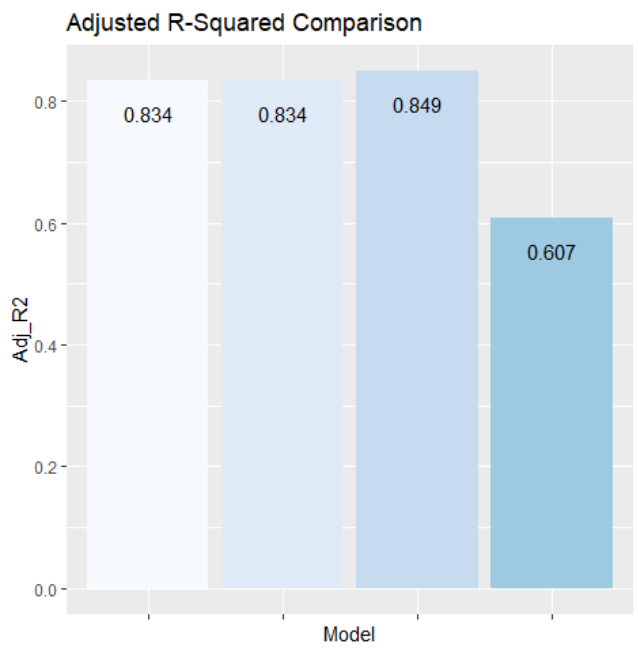


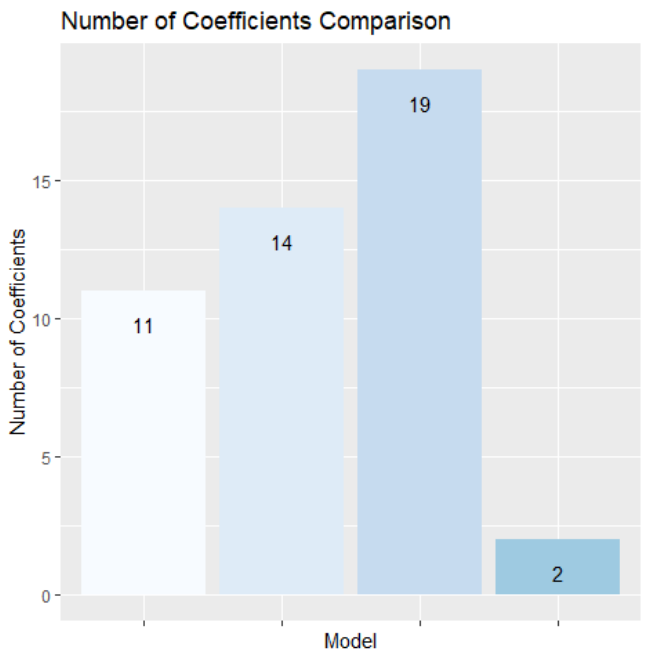
Adj R² : 0.8345

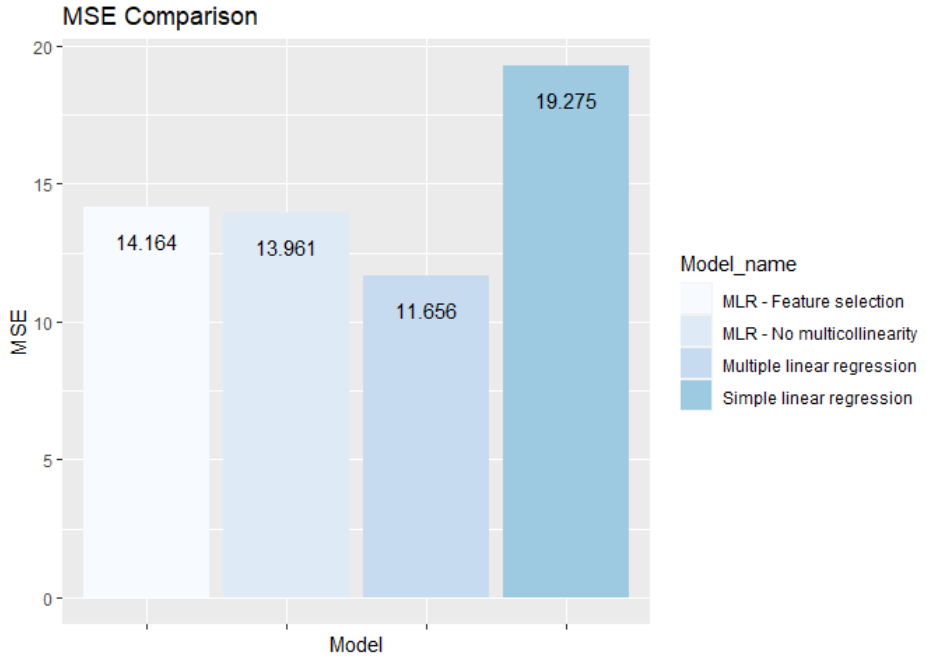
Feature	AdjR2	Cp	BIC
StatusDeveloping	1	1	1
Adult.Mortality	1	1	1
Alcohol	1	1	1
percentage.expenditure	1	1	1
Hepatitis.B	1	1	1
Measles	1	1	1
Polio	1	1	1
Total.expenditure	0	0	0
Diphtheria	1	1	1
HIV.AIDS	1	1	1
GDP	0	0	0
Population	0	0	0
Schooling	1	1	1

Linear Regression - Model comparison

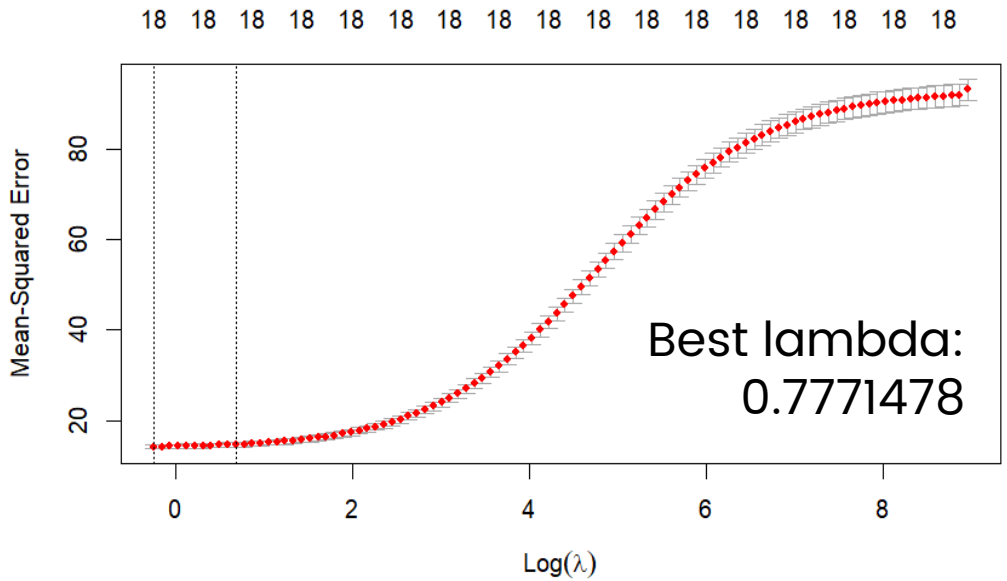
Model_name	Adj_R2	AIC	BIC	n_coef	MSE	RMSE	n_RMSE_sd	n_RMSE_range
Multiple linear regression	0.8494650	13052.44	13167.93	19	11.65604	3.414094	0.4123672	0.0834742
MLR - Feature selection	0.8344579	13270.57	13339.86	11	14.16370	3.763469	0.4545661	0.0920164
MLR - No multicollinearity	0.8343365	13275.30	13361.92	14	13.96116	3.736464	0.4513042	0.0913561
Simple linear regression	0.6071161	15317.74	15335.06	2	19.27546	4.390383	0.5302871	0.1073443

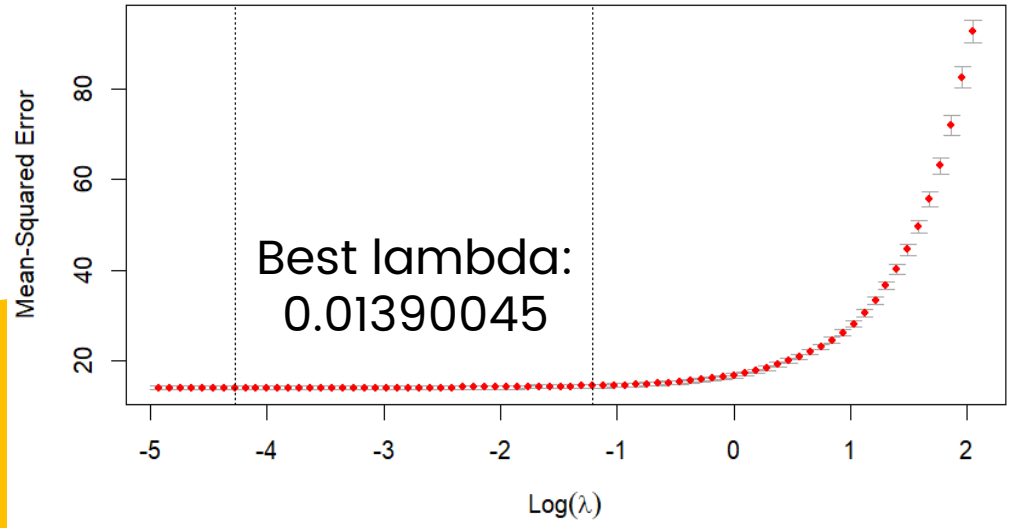






Linear Regression - Regularization

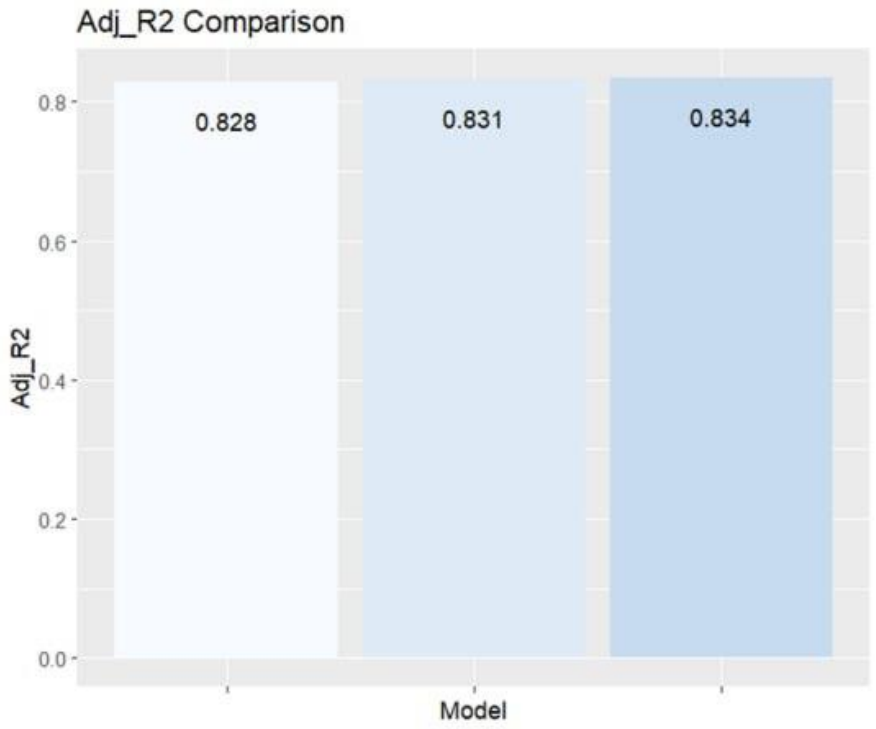


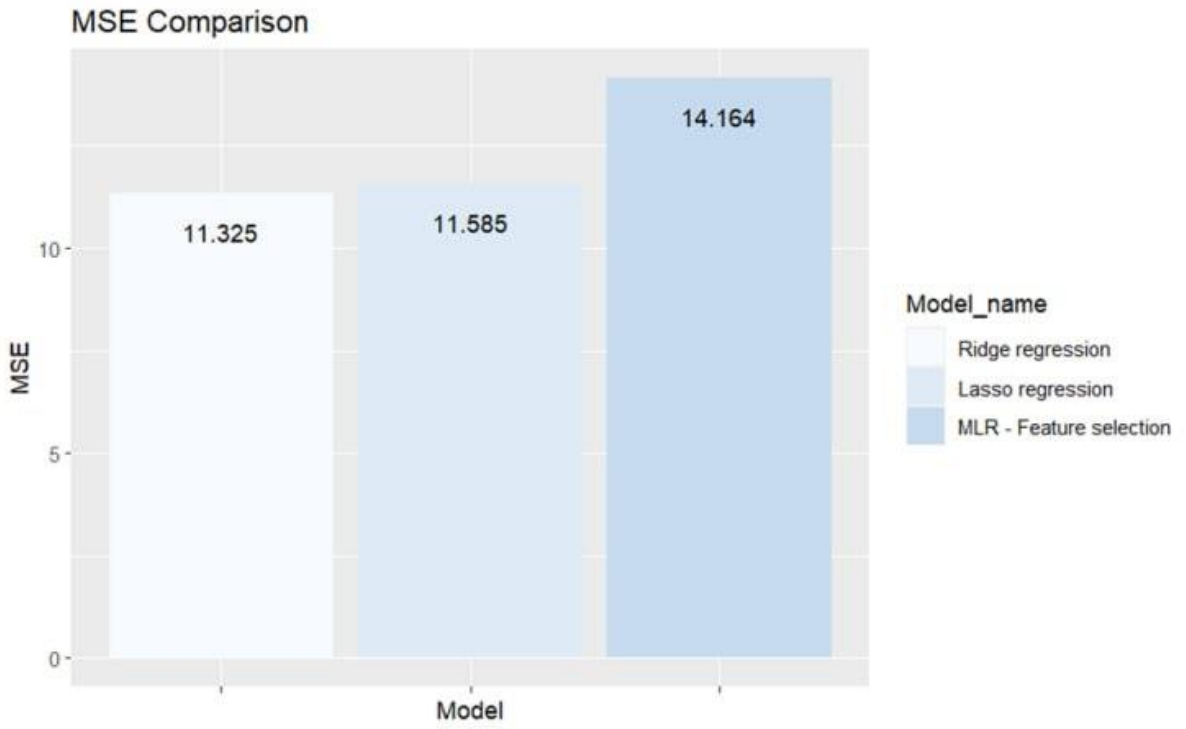


Least important features

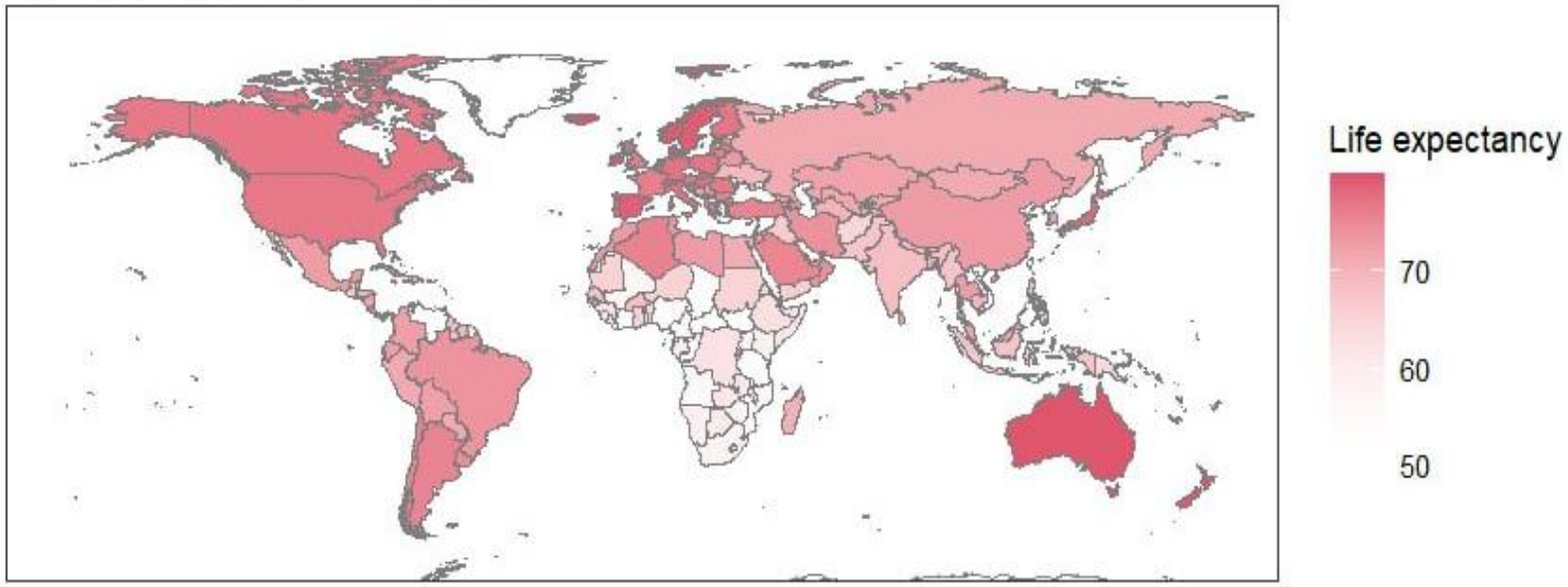
- GDP
- Population
- Measles
- percentage.expenditure
- Infant.deaths

Linear Regression – Model Comparison

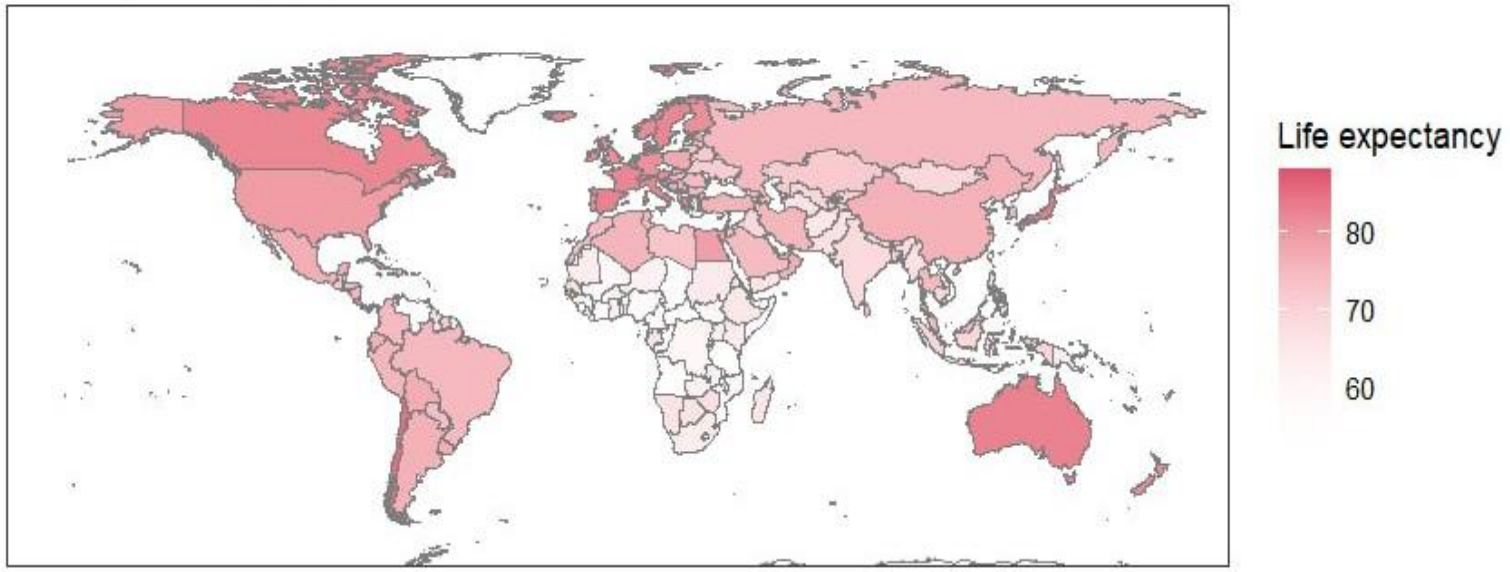




Life expectancy predictions

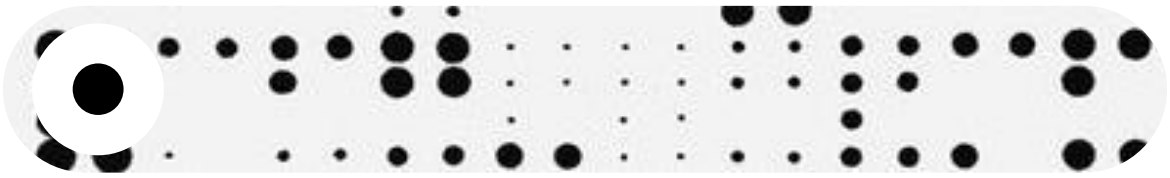


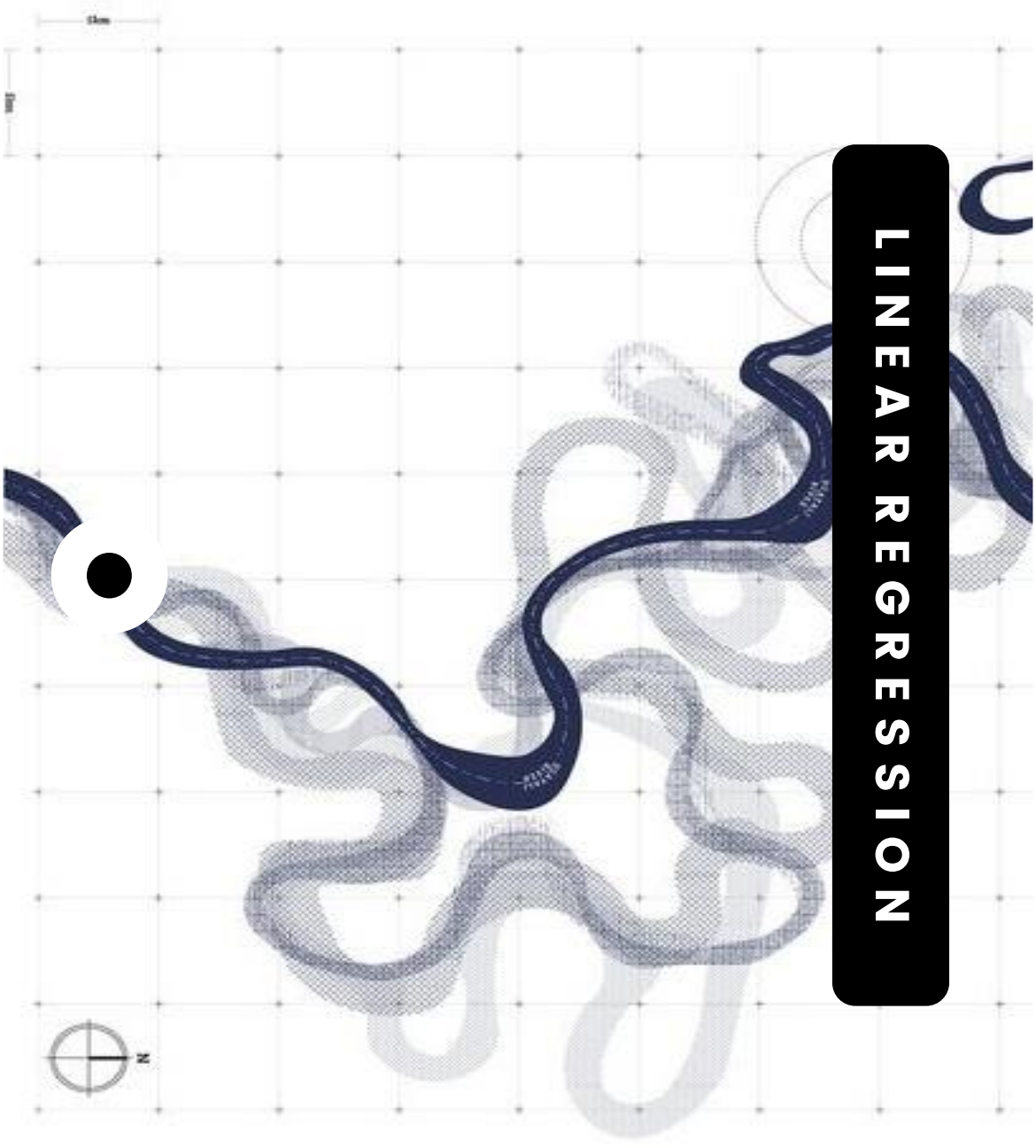
Life expectancy true values



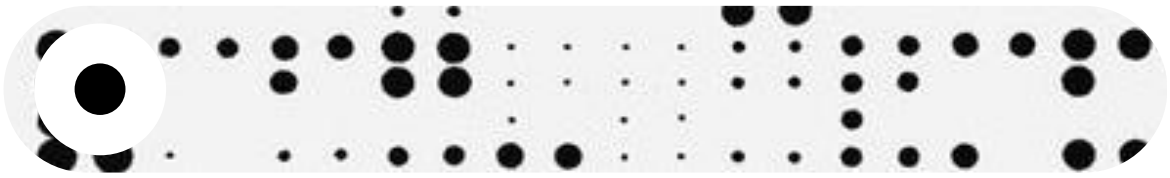
Linear Regression - Conclusions

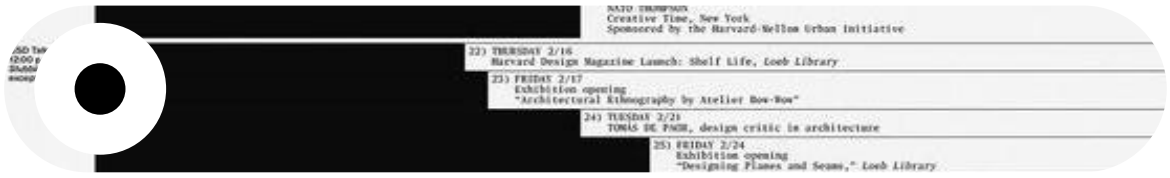
Variable	Effect on Life Expectancy
Alcohol	
Percentage Expenditure	
Polio	
Diphtheria	
GDP	
Schooling	
StatusDeveloping	
Adult Mortality	
Hepatitis B	
Measles	
HIV/AIDS	



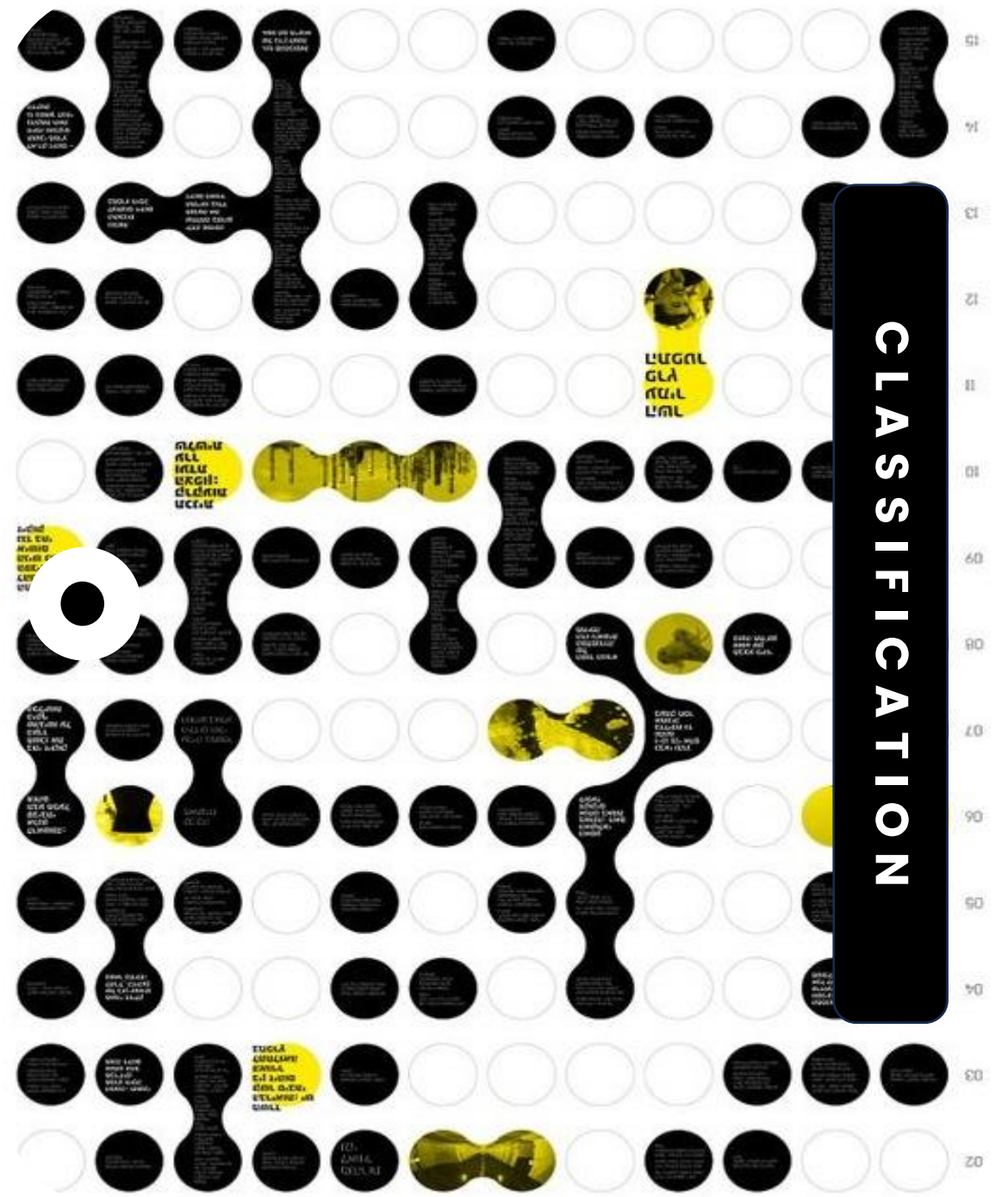


LINEAR REGRESSION




NATO THOMPSON
Creative Time, New York
Sponsored by the Harvard-Mellon Urban Initiative
22) THURSDAY 2/16
Harvard Design Magazine Launch: Shelf Life, Loeb Library
23) FRIDAY 2/17
Exhibition opening
"Architectural Kinography by Atelier Bow-Wow"
24) TUESDAY 2/21
TOMAS DE PAOR, design critic in architecture
25) FRIDAY 2/24
Exhibition opening
"Designing Planes and Seams," Loeb Library


DARK
FRANK
JOY
PROBY
PROBY
FACTORY
STARS


CLASSIFICATION

Classification

Binary classification problem :
Country's life expectancy is below or above the (Italian) **pension threshold of 67 years**

```
##      Above      Below
## 0.6589517 0.3410483
```

Unbalanced data : **Above <u>66 : 34</u> Below**

Test set : years 2013, 2014, 2015

```
## Train size:  80.97345
## Test size:  19.02655
```

```
##      Above      Below
## 0.6469105 0.3530895
```

```
##      Above      Below
## 0.7101968 0.2898032
```

Classification

Algorithms

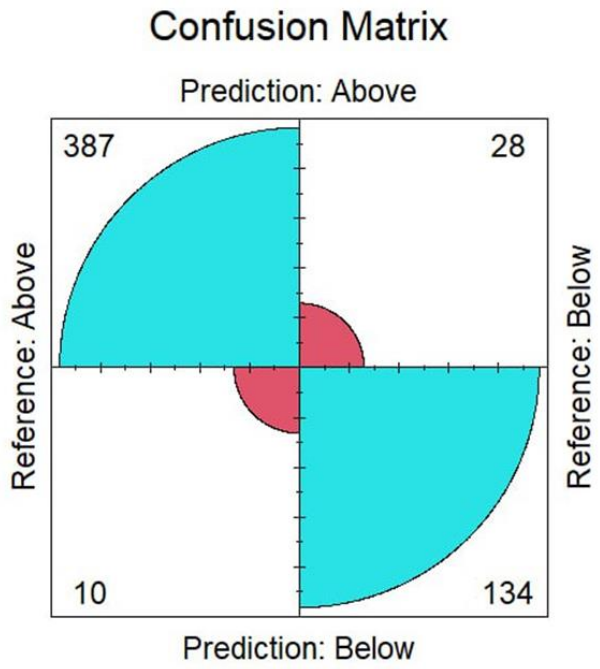
- Logistics Regression
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- Naive Bayes
- K-Nearest Neighbors

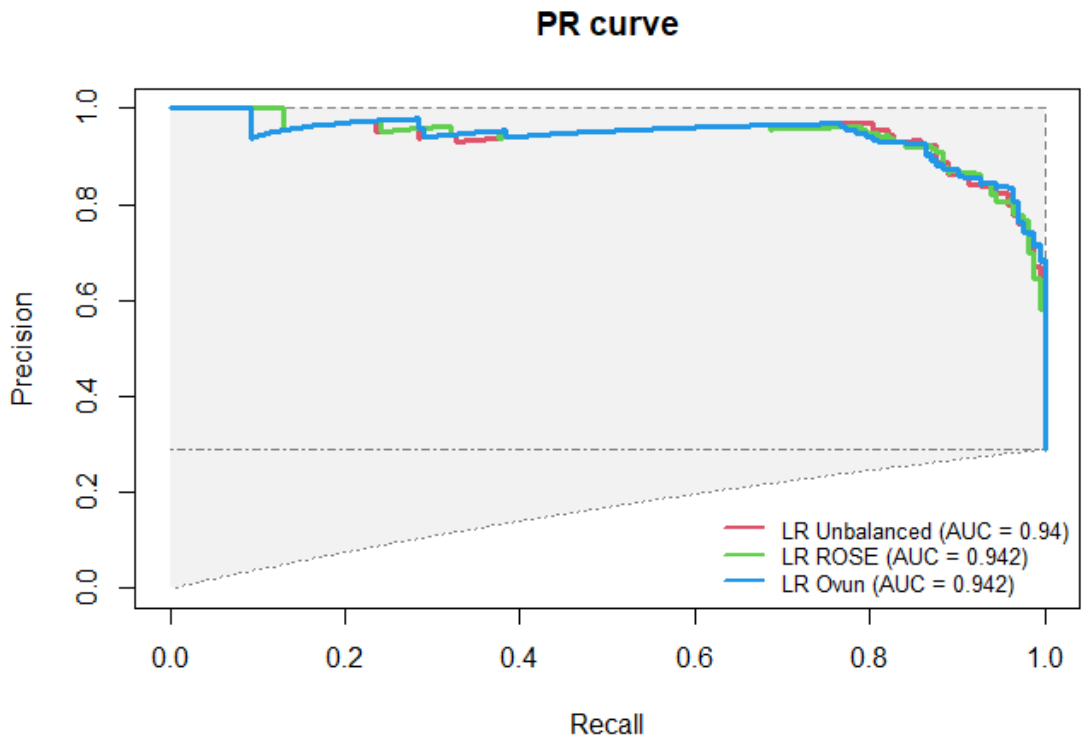
Balancing approaches

- ROSE
- Ovun.sample

Logistic Regression

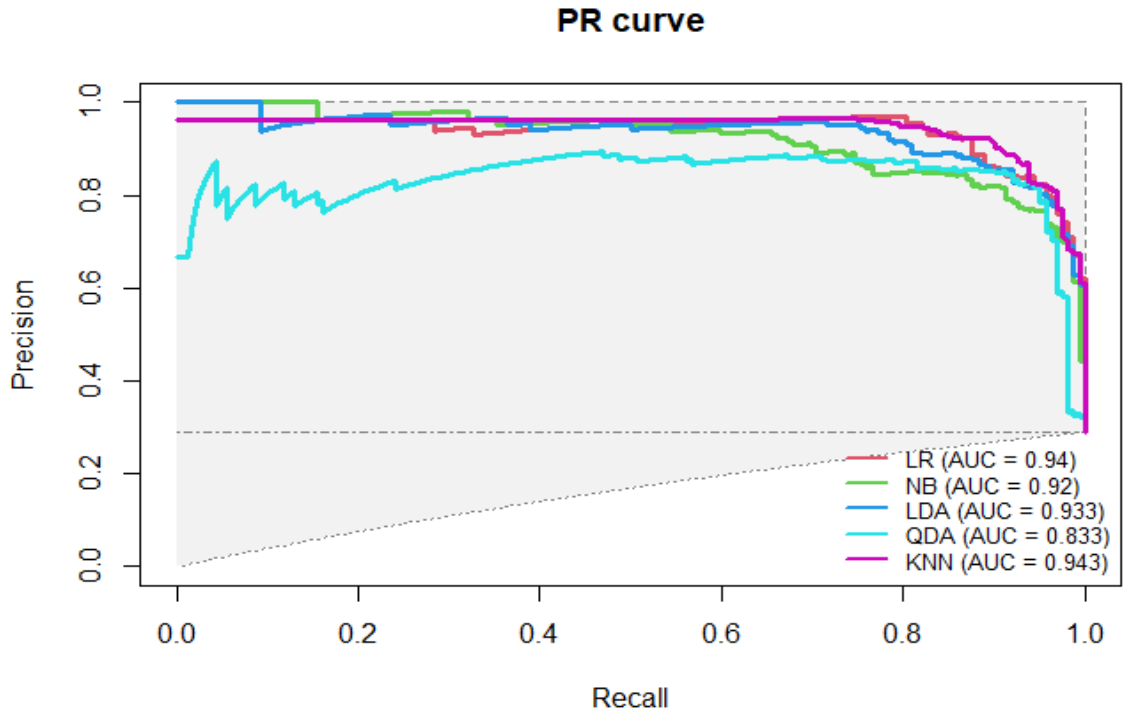
Model_name	AIC	McFaddens_R2	Accuracy	Precision	Recall	F1_score	PR_AUC
Logistic regression - Ovun	904.580	0.737	0.934	0.950	0.957	0.954	0.942
Logistic regression - ROSE	1091.475	0.681	0.930	0.954	0.947	0.951	0.942
Logistic regression	929.451	0.711	0.932	0.933	0.975	0.953	0.940





Classification - Model comparison

Model_name	Accuracy	Precision	Recall	F1_score	PR_AUC
Logistic regression	0.934	0.950	0.957	0.954	0.942
Quadratic Discriminant Analysis	0.921	0.949	0.940	0.944	0.833
Linear Discriminant Analysis	0.919	0.929	0.960	0.944	0.933
K-Nearest Neighbors	0.909	0.989	0.882	0.932	0.943
Naive Bayes	0.905	0.962	0.902	0.931	0.920



Note: These models utilize the balanced dataset (ovun.sample).

Thank you
for your attention

You can find the project at the following link