**Project 1 – 0222**

**Project 1.1**

After we review and discuss about our data, we decided to change the data set. Therefore, here are the answers to questions in HW2, HW3 and HW4.

**a. Partner:** Chunjingwen Cui..

**b. Include a link to your dataset**

<https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who?resource=download>

**c. What does each observation of the dataset represent?**

Each observation represents the data of a country in a particular year.

**d. What does each variable of the dataset represent?**

Country: 193 countries

Year: 2000-2015

Status: Developed or Developing status

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 1 000 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)

**e. Why are you interested in performing data analytics on this dataset?**

This dataset is about life expectancy, including data related to life expectancy in different years in different countries. This data set has a large amount of data, including 22 columns and 2938 rows, which meets the requirements of this project. In addition, many factors affect life expectancy, and this dataset lists immune-related, mortality, economic, and social factors. By analyzing these data, we can understand how these factors affect human life expectancy. We can also make recommendations to the government to improve the population's life expectancy based on the data analysis results.

**Project 1.2**

The following is the output of the console：

> ## Homework 3

> # 1. read data

> getwd()

[1] "/Users/corinne/Desktop"

> lifeExpectancy<- read\_csv("Life Expectancy Data.csv")

Rows: 2938 Columns: 22

── Column specification ────────────────────────────────────────────────────────────

Delimiter: ","

chr (2): Country, Status

dbl (20): Year, Life expectancy, Adult Mortality, infant deaths, Alcohol, percen...

ℹ Use `spec()` to retrieve the full column specification for this data.

ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

> View(lifeExpectancy)

> # 2. Use the str() function to figure out the number of rows and columns

> str(lifeExpectancy)

spc\_tbl\_ [2,938 × 22] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)

$ Country : chr [1:2938] "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...

$ Year : num [1:2938] 2015 2014 2013 2012 2011 ...

$ Status : chr [1:2938] "Developing" "Developing" "Developing" "Developing" ...

$ Life expectancy : num [1:2938] 65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...

$ Adult Mortality : num [1:2938] 263 271 268 272 275 279 281 287 295 295 ...

$ infant deaths : num [1:2938] 62 64 66 69 71 74 77 80 82 84 ...

$ Alcohol : num [1:2938] 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...

$ percentage expenditure : num [1:2938] 71.3 73.5 73.2 78.2 7.1 ...

$ Hepatitis B : num [1:2938] 65 62 64 67 68 66 63 64 63 64 ...

$ Measles : num [1:2938] 1154 492 430 2787 3013 ...

$ BMI : num [1:2938] 19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...

$ under-five deaths : num [1:2938] 83 86 89 93 97 102 106 110 113 116 ...

$ Polio : num [1:2938] 6 58 62 67 68 66 63 64 63 58 ...

$ Total expenditure : num [1:2938] 8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...

$ Diphtheria : num [1:2938] 65 62 64 67 68 66 63 64 63 58 ...

$ HIV/AIDS : num [1:2938] 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...

$ GDP : num [1:2938] 584.3 612.7 631.7 670 63.5 ...

$ Population : num [1:2938] 33736494 327582 31731688 3696958 2978599 ...

$ thinness 1-19 years : num [1:2938] 17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...

$ thinness 5-9 years : num [1:2938] 17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...

$ Income composition of resources: num [1:2938] 0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.405 ...

$ Schooling : num [1:2938] 10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...

- attr(\*, "spec")=

.. cols(

.. Country = col\_character(),

.. Year = col\_double(),

.. Status = col\_character(),

.. `Life expectancy` = col\_double(),

.. `Adult Mortality` = col\_double(),

.. `infant deaths` = col\_double(),

.. Alcohol = col\_double(),

.. `percentage expenditure` = col\_double(),

.. `Hepatitis B` = col\_double(),

.. Measles = col\_double(),

.. BMI = col\_double(),

.. `under-five deaths` = col\_double(),

.. Polio = col\_double(),

.. `Total expenditure` = col\_double(),

.. Diphtheria = col\_double(),

.. `HIV/AIDS` = col\_double(),

.. GDP = col\_double(),

.. Population = col\_double(),

.. `thinness 1-19 years` = col\_double(),

.. `thinness 5-9 years` = col\_double(),

.. `Income composition of resources` = col\_double(),

.. Schooling = col\_double()

.. )

- attr(\*, "problems")=<externalptr>

> # 3. Use the summary() function to summarize your variables

> summary(lifeExpectancy)

Country Year Status Life expectancy

Length:2938 Min. :2000 Length:2938 Min. :36.30

Class :character 1st Qu.:2004 Class :character 1st Qu.:63.10

Mode :character Median :2008 Mode :character Median :72.10

Mean :2008 Mean :69.22

3rd Qu.:2012 3rd Qu.:75.70

Max. :2015 Max. :89.00

NA's :10

Adult Mortality infant deaths Alcohol percentage expenditure

Min. : 1.0 Min. : 0.0 Min. : 0.0100 Min. : 0.000

1st Qu.: 74.0 1st Qu.: 0.0 1st Qu.: 0.8775 1st Qu.: 4.685

Median :144.0 Median : 3.0 Median : 3.7550 Median : 64.913

Mean :164.8 Mean : 30.3 Mean : 4.6029 Mean : 738.251

3rd Qu.:228.0 3rd Qu.: 22.0 3rd Qu.: 7.7025 3rd Qu.: 441.534

Max. :723.0 Max. :1800.0 Max. :17.8700 Max. :19479.912

NA's :10 NA's :194

Hepatitis B Measles BMI under-five deaths

Min. : 1.00 Min. : 0.0 Min. : 1.00 Min. : 0.00

1st Qu.:77.00 1st Qu.: 0.0 1st Qu.:19.30 1st Qu.: 0.00

Median :92.00 Median : 17.0 Median :43.50 Median : 4.00

Mean :80.94 Mean : 2419.6 Mean :38.32 Mean : 42.04

3rd Qu.:97.00 3rd Qu.: 360.2 3rd Qu.:56.20 3rd Qu.: 28.00

Max. :99.00 Max. :212183.0 Max. :87.30 Max. :2500.00

NA's :553 NA's :34

Polio Total expenditure Diphtheria HIV/AIDS

Min. : 3.00 Min. : 0.370 Min. : 2.00 Min. : 0.100

1st Qu.:78.00 1st Qu.: 4.260 1st Qu.:78.00 1st Qu.: 0.100

Median :93.00 Median : 5.755 Median :93.00 Median : 0.100

Mean :82.55 Mean : 5.938 Mean :82.32 Mean : 1.742

3rd Qu.:97.00 3rd Qu.: 7.492 3rd Qu.:97.00 3rd Qu.: 0.800

Max. :99.00 Max. :17.600 Max. :99.00 Max. :50.600

NA's :19 NA's :226 NA's :19

GDP Population thinness 1-19 years thinness 5-9 years

Min. : 1.68 Min. :3.400e+01 Min. : 0.10 Min. : 0.10

1st Qu.: 463.94 1st Qu.:1.958e+05 1st Qu.: 1.60 1st Qu.: 1.50

Median : 1766.95 Median :1.387e+06 Median : 3.30 Median : 3.30

Mean : 7483.16 Mean :1.275e+07 Mean : 4.84 Mean : 4.87

3rd Qu.: 5910.81 3rd Qu.:7.420e+06 3rd Qu.: 7.20 3rd Qu.: 7.20

Max. :119172.74 Max. :1.294e+09 Max. :27.70 Max. :28.60

NA's :448 NA's :652 NA's :34 NA's :34

Income composition of resources Schooling

Min. :0.0000 Min. : 0.00

1st Qu.:0.4930 1st Qu.:10.10

Median :0.6770 Median :12.30

Mean :0.6276 Mean :11.99

3rd Qu.:0.7790 3rd Qu.:14.30

Max. :0.9480 Max. :20.70

NA's :167 NA's :163

**Project 1.3**

The following is the output of the console：

> ## Homework 4

> library(tidyverse)

> # 1. At least one scatterplot showing the relationship between two quantitative

> # variables in your dataset, colored by a categorical variable in the dataset.

> lifeExpectancy %>%

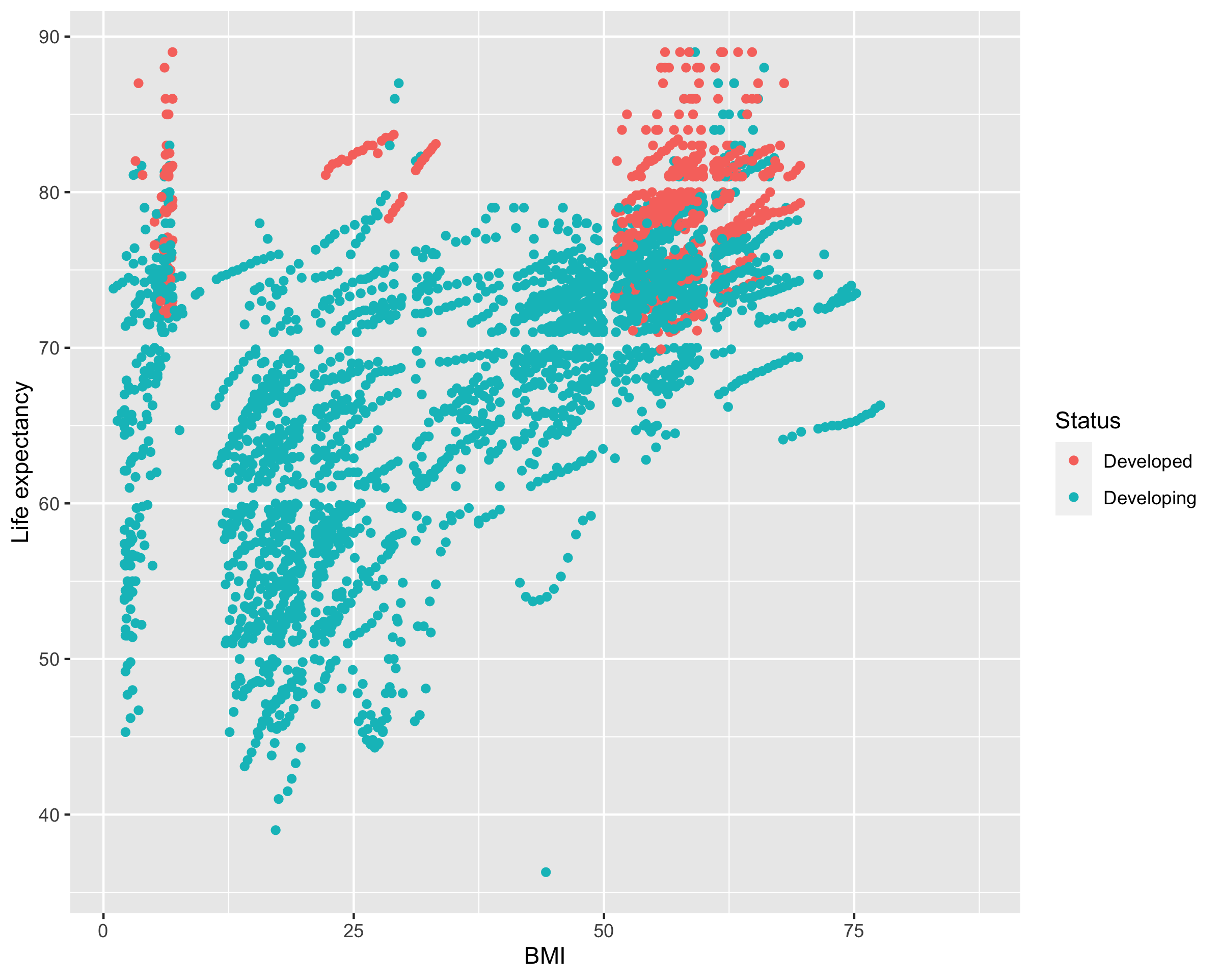
+ ggplot(aes(BMI, `Life expectancy`,

+ color = Status)) +

+ geom\_point()

> ggsave("Project1-scatter-1.png")

Saving 7.99 x 6.42 in image



> lifeExpectancy %>%

+ ggplot(aes(GDP, `Life expectancy`,

+ color = Status)) +

+ geom\_point()

> ggsave("Project1-scatter-2.png")

Saving 7.99 x 6.42 in image

Chart, scatter chart

Description automatically generated

> lifeExpectancy %>%

+ ggplot(aes(Schooling, `Life expectancy`,

+ color = Status)) +

+ geom\_point()

> ggsave("Project1-scatter-3.png")

Saving 7.99 x 6.42 in image

Chart, scatter chart

Description automatically generated

> lifeExpectancy %>%

+ ggplot(aes(Schooling, Alcohol,

+ color = Status)) +

+ geom\_point()

> ggsave("Project1-scatter-4.png")

Saving 7.99 x 6.42 in image



> # 2. At least one boxplot showing the relationship between a categorical and

> # quantitative variable in your dataset.

> lifeExpectancy %>%

+ ggplot(aes(Status, Alcohol)) +

+ geom\_boxplot()

> ggsave("Project1-box-1.png")

Saving 7.99 x 6.42 in image

Chart, box and whisker chart

Description automatically generated

> lifeExpectancy %>%

+ ggplot(aes(Status, `Life expectancy`)) +

+ geom\_boxplot()

> ggsave("Project1-box-2.png")

Saving 7.99 x 6.42 in image

Chart, box and whisker chart

Description automatically generated