The Blue Zones: Why do some areas of the world have higher life expectancies than others?

Group 9

2022-11-22

Questions we cover in the project

(1) A statement of the question or purpose. What problems or questions did you set out to analyse? What were the key issues raised?

(1) Our goal is to analyze different factors, such as; nutrition, lifestyle, exercise, climate, genetics, hygiene,... and the impact they have on longevity in different parts of the world.

More specifically we aim to answer these questions:

a. Why do people living in blue zones live longer than others ?
b. How and why did longevity change over time for these countries ?
c. Which factors have the biggest influence on life expectancy ?

(2) The background and preparation for conducting the project. How did you prepare for the project? What sources or background readings did you consult? What information did you use in developing your ideas from the conceptual stage to the finished project?

We read through some articles about Blue Zones. Blue zones are regions in the world where people are speculated to live longer than average; the term was coined by Gianni Pes, Michel Poulain and Dan Buettner. There are five blues zones: Okinawa (Japan), Sardinia (Italy), Nicoya (Costa Rica), Icaria (Greece) and Loma Linda (California). We also looked up some research about some potential factors that could contribute to life expectancy in a region (eg. infant mortality rate, GDP, education level, and different deseases)

(3) Methodology. What did you do, and how did you do it? What statistical techniques did you use — for instance, scatterplots, correlation, confidence intervals, linear/logistic regression?

We used hypothesis testing to check correlations between factors affecting life expectancy and life expectancy. We used linear regressions to identify variables that are significantly impacting life expectancy in a particular country in different time points, and then we adjusted the regression model based on the initial regression result to increase its explaining power. We also used cluster analysis based on GDP and life expectancy to cluster world nations.

(4) Results and conclusions. This is where you summarise and present your data analyses and communicate your main results. What did you find out? This might include tables, graphs, or verbal summaries. What did you learn about the problem or question you set out to investigate?

Data Cleaning

```
df1 <- read_csv(file = 'data viz dataset/(NEEDS CLEAN)-infant-inmortality.csv',skip=4, show_col_types =
df1 <-df1[,-67]
df1 <-df1[,-66]

df1 <- df1 %>%
  drop_na() %>%
  pivot_longer(names_to = "year",
               values_to = "infant_inmortality",
               cols = -(1:4)) %>%
  janitor::clean_names()

glimpse(df1)
```

```
## Rows: 7,015
## Columns: 6
## $ country_name       <chr> "United Arab Emirates", "United Arab Emirates", "Un~
## $ country_code       <chr> "ARE", "ARE", "ARE", "ARE", "ARE", "ARE", "ARE", "A~
## $ indicator_name     <chr> "Mortality rate, infant (per 1,000 live births)", "~
## $ indicator_code     <chr> "SP.DYN.IMRT.IN", "SP.DYN.IMRT.IN", "SP.DYN.IMRT.IN~
## $ year               <chr> "1960", "1961", "1962", "1963", "1964", "1965", "19~
## $ infant_inmortality <dbl> 133.2, 126.9, 120.7, 114.7, 108.4, 102.1, 95.6, 89.~
```

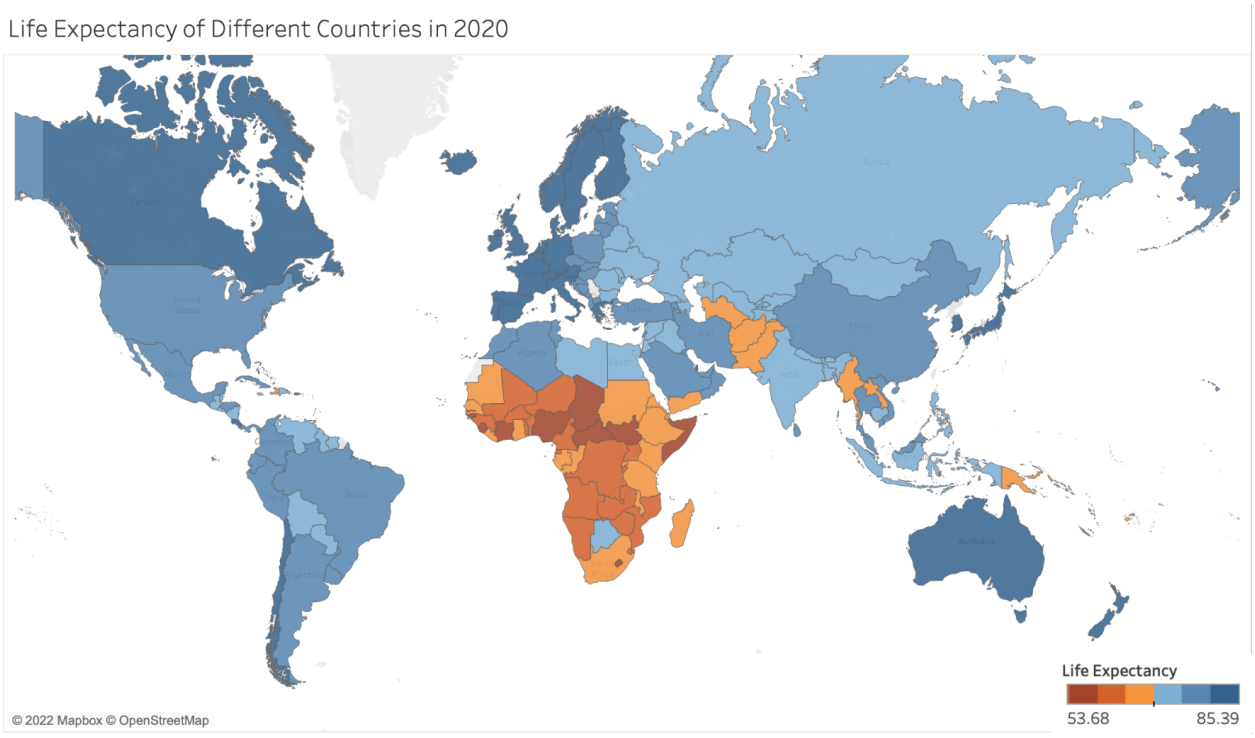
```
write.table(df1, file = "data viz dataset/CLEANED-infant-mortality.csv", row.names = FALSE, na = "", se
#storing the data in csv file, removing row name and column name
```

```
#read in data, remove the first 3 rows
df2 <- read_csv(file = 'data viz dataset/(NEEDS CLEAN) life-expectancy-2020.csv',skip=4, show_col_types
df2 <-df2[,-67]
df2 <-df2[,-66]

df2 <- df2 %>%
  drop_na() %>%
  pivot_longer(names_to = "year",
               values_to = "life_expectancy",
               cols = -(1:4)) %>%
  janitor::clean_names()

glimpse(df2)
```

```
## Rows: 14,457
## Columns: 6
## $ country_name    <chr> "Aruba", "Aruba", "Aruba", "Aruba", "Aruba", "Aruba", ~
## $ country_code    <chr> "ABW", "ABW", "ABW", "ABW", "ABW", "ABW", "ABW", "ABW"~
## $ indicator_name  <chr> "Life expectancy at birth, total (years)", "Life expec~
## $ indicator_code  <chr> "SP.DYN.LE00.IN", "SP.DYN.LE00.IN", "SP.DYN.LE00.IN", ~
## $ year            <chr> "1960", "1961", "1962", "1963", "1964", "1965", "1966"~
## $ life_expectancy <dbl> 65.7, 66.1, 66.4, 66.8, 67.1, 67.4, 67.8, 68.1, 68.4, ~
```

```
write_csv(df2, file = "data viz dataset/CLEANED-life-expectancy-2020.csv", na = "")
#storing the data in csv file, removing row name and column name
```

The Trend of Life Expectancy around the world



In the graph, we can see that there is a clear difference in life expectancy around the globe, with Europe and North America higher than the rest of the world, and Africa lower than average.

```
library(gapminder)
gapminder %>% ggplot(aes(x = continent, y=lifeExp, group_by = year, colour = continent))+
  facet_wrap(~year)+
  geom_boxplot()+
  labs(
    title ="Life Expectancy in Asia and Africa is growing fastest",
    subtitle = "Life Expectancy of Different Continents Over the Years")+
  theme_bw()+
  theme(legend.position = "None")
```

According to the boxplot over the years, we found that life expectancy in different continents has changed over time, with Africa and Asia showing a significant increase, while the number of the rest of the world increase steadily.

Regression on life-expectancy - potential factor dataset

Preparing data for regression

```
lc <- read.csv('data viz dataset/CLEANED-Life Expectancy - potential factors.csv') %>%
  clean_names()
```

```r
lc_clean <- na.omit(lc)

#filtering data to the latest available year (2014)
lc_2014=lc_clean%>%filter(year==2014)

lc_2004=lc_clean%>%filter(year==2004)

lc_2000=lc_clean%>%filter(year==2000)
```

linear model on year 2014

```r
lm1 <-lm(life_expectancy ~ bmi + polio + percentage_expenditure + hepatitis_b +measles +total_expenditu

summary(lm1)
```

```
##
## Call:
## lm(formula = life_expectancy ~ bmi + polio + percentage_expenditure +
##     hepatitis_b + measles + total_expenditure + diphtheria +
##     hiv_aids + gdp + income_composition_of_resources + schooling +
##     alcohol, data = lc_2014)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.965  -1.772   0.274   1.849   9.347
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      4.02e+01   2.24e+00   18.00  < 2e-16 ***
## bmi                             -6.45e-03   1.95e-02   -0.33    0.741
## polio                           -3.39e-03   2.22e-02   -0.15    0.879
## percentage_expenditure           5.69e-04   4.91e-04    1.16    0.250
## hepatitis_b                      1.13e-02   2.96e-02    0.38    0.705
## measles                         -1.80e-05   3.15e-05   -0.57    0.570
## total_expenditure                2.99e-01   1.34e-01    2.24    0.027 *
## diphtheria                       6.69e-03   3.62e-02    0.18    0.854
## hiv_aids                        -1.36e+00   2.32e-01   -5.87  4.1e-08 ***
## gdp                             -7.00e-05   6.98e-05   -1.00    0.318
## income_composition_of_resources  4.69e+01   6.15e+00    7.63  6.6e-12 ***
## schooling                       -2.03e-01   2.86e-01   -0.71    0.479
## alcohol                         -5.27e-02   9.99e-02   -0.53    0.599
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.42 on 118 degrees of freedom
## Multiple R-squared:  0.857,  Adjusted R-squared:  0.842
## F-statistic: 58.9 on 12 and 118 DF,  p-value: <2e-16
```

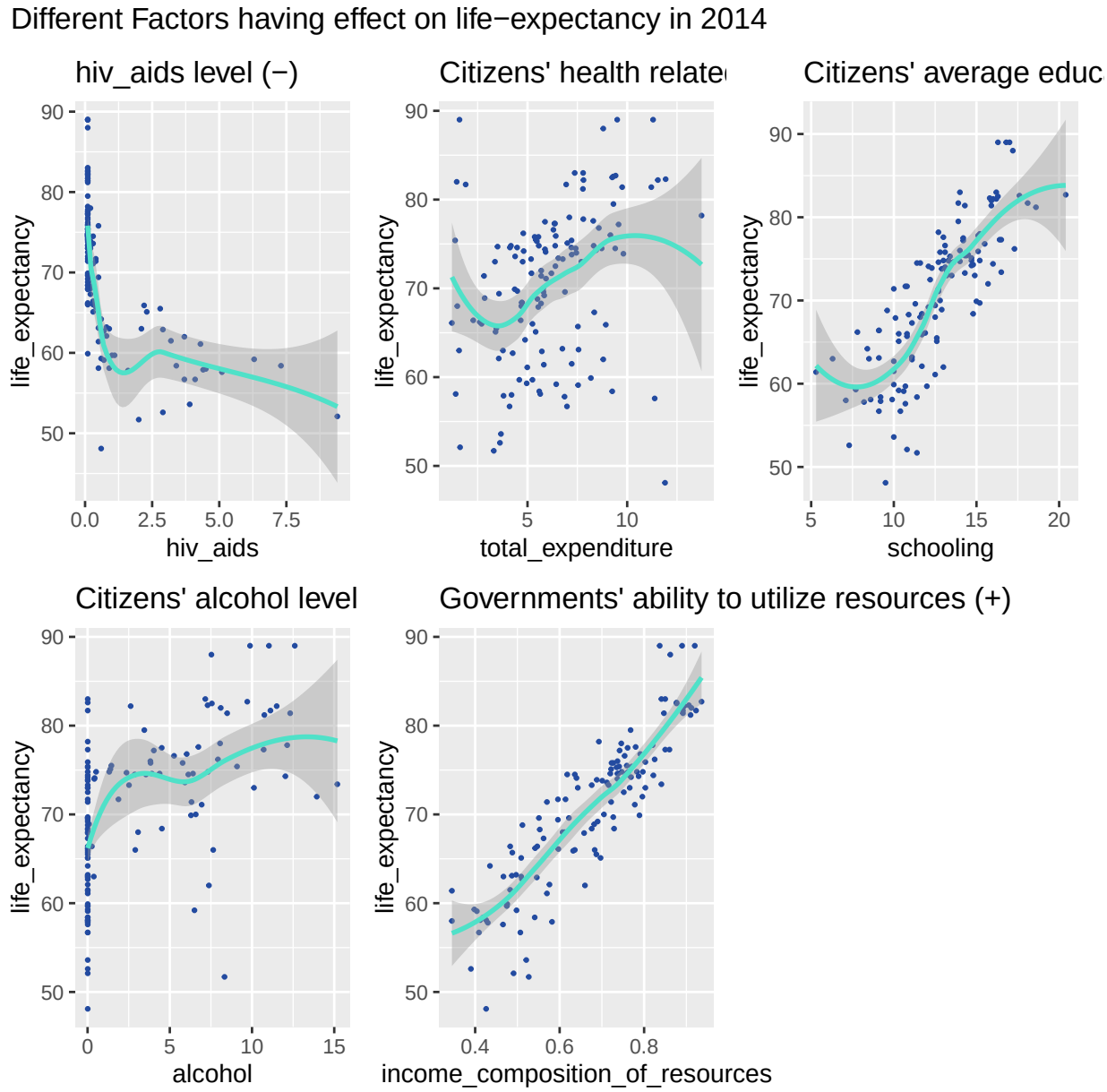
```r
#adjusting the linear model

lm_new1=lm(life_expectancy ~  total_expenditure + hiv_aids + income_composition_of_resources, lc_2014)
```

```r
summary(lm_new1)
```

```
##
## Call:
## lm(formula = life_expectancy ~ total_expenditure + hiv_aids +
##     income_composition_of_resources, data = lc_2014)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.317  -1.766   0.205   1.809   9.663
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       41.189      1.632   25.25  < 2e-16 ***
## total_expenditure                  0.335      0.120    2.79   0.0061 **
## hiv_aids                          -1.389      0.214   -6.50  1.7e-09 ***
## income_composition_of_resources   42.419      2.277   18.63  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.34 on 127 degrees of freedom
## Multiple R-squared:  0.852,  Adjusted R-squared:  0.849
## F-statistic:  244 on 3 and 127 DF,  p-value: <2e-16
```

```r
#visualizing influence of significant features in the regression model

a <- ggplot(lc_2014, aes(x = hiv_aids, y = life_expectancy))+
  geom_point(colour="#234BA0",size=0.5)+
  geom_smooth(colour="#4FE1C8")+
  labs(title="hiv_aids level (-)",
       )


b <- ggplot(lc_2014, aes(x = total_expenditure, y = life_expectancy))+
  geom_point(colour="#234BA0",size=0.5)+
  geom_smooth(colour="#4FE1C8")+
  labs(title="Citizens' health related expenditure (+)")

c <- ggplot(lc_2014, aes(x = schooling, y = life_expectancy))+
  geom_point(colour="#234BA0",size=0.5)+
  geom_smooth(colour="#4FE1C8")+
  labs(title="Citizens' average education level (+)")

d <- ggplot(lc_2014, aes(x =alcohol, y = life_expectancy))+
  geom_point(colour="#234BA0",size=0.5)+
  geom_smooth(colour="#4FE1C8")+
  labs(title="Citizens' alcohol level (+)")

#Income composition of resources is a index that measures the government's ability to utilize its natur
e <- ggplot(lc_2014, aes(x =income_composition_of_resources, y = life_expectancy))+
  geom_point(colour="#234BA0",size=0.5)+
  geom_smooth(colour="#4FE1C8")+
  labs(title="Governments' ability to utilize resources (+)")
```

```
a+b+c+d+e + plot_annotation(
  title = "Different Factors having effect on life-expectancy in 2014"
)
```



Based on multi-variable linear regression result, we can see that, in 2014, Income Composition of Resources, and HIV rate significantly impacts the average life expectancy in a particular country. We left out Adult Mortality to avoid perfect fit problem)

#linear model on year 2004

```r
lm2 <-lm(life_expectancy ~ bmi + polio + percentage_expenditure + hepatitis_b +measles +total_expenditu

summary(lm2)
```

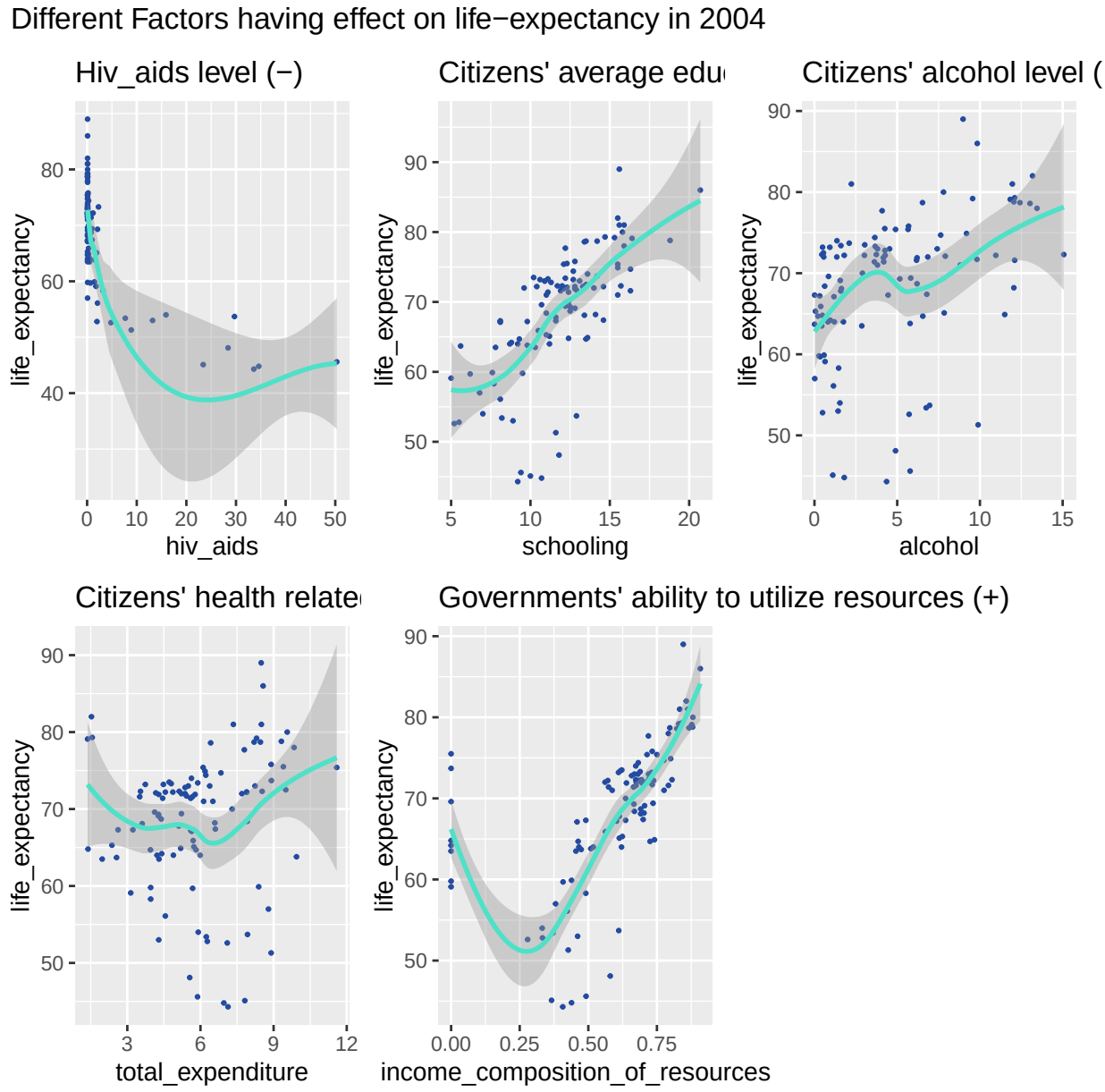
```
##
## Call:
## lm(formula = life_expectancy ~ bmi + polio + percentage_expenditure +
##     hepatitis_b + measles + total_expenditure + diphtheria +
##     hiv_aids + gdp + income_composition_of_resources + schooling +
##     alcohol, data = lc_2004)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.254 -2.170  0.059  2.872  9.835
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      4.64e+01   2.45e+00   18.94  < 2e-16 ***
## bmi                              5.74e-02   2.64e-02    2.17    0.032 *
## polio                            3.24e-04   1.86e-02    0.02    0.986
## percentage_expenditure          -5.56e-04   9.37e-04   -0.59    0.555
## hepatitis_b                      4.81e-03   1.60e-02    0.30    0.765
## measles                          5.04e-05   4.14e-05    1.22    0.227
## total_expenditure                2.20e-01   1.93e-01    1.14    0.256
## diphtheria                       2.34e-02   2.30e-02    1.02    0.311
## hiv_aids                        -5.82e-01   4.99e-02  -11.68  < 2e-16 ***
## gdp                              2.56e-04   1.50e-04    1.70    0.092 .
## income_composition_of_resources  4.82e+00   2.45e+00    1.97    0.052 .
## schooling                        1.32e+00   2.13e-01    6.21  1.6e-08 ***
## alcohol                         -3.75e-01   1.45e-01   -2.58    0.011 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.84 on 90 degrees of freedom
## Multiple R-squared:  0.842,  Adjusted R-squared:  0.821
## F-statistic: 39.9 on 12 and 90 DF,  p-value: <2e-16
```

```r
#adjusting the linear model

lm_new2=lm(life_expectancy ~ bmi + hiv_aids + income_composition_of_resources +gdp +alcohol+schooling,l

summary(lm_new2)
```

```
##
## Call:
## lm(formula = life_expectancy ~ bmi + hiv_aids + income_composition_of_resources +
##     gdp + alcohol + schooling, data = lc_2004)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.674 -2.184 -0.074  2.915 10.300
##
```

```
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      4.93e+01   1.87e+00   26.35  < 2e-16 ***
## bmi                              4.82e-02   2.53e-02    1.90  0.05980 .
## hiv_aids                        -5.88e-01   4.79e-02  -12.28  < 2e-16 ***
## income_composition_of_resources  5.51e+00   2.38e+00    2.32  0.02245 *
## gdp                              1.64e-04   4.09e-05    4.02  0.00012 ***
## alcohol                         -3.25e-01   1.39e-01   -2.35  0.02109 *
## schooling                        1.38e+00   2.06e-01    6.71  1.4e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.82 on 96 degrees of freedom
## Multiple R-squared:  0.833,  Adjusted R-squared:  0.822
## F-statistic: 79.8 on 6 and 96 DF,  p-value: <2e-16
```

```
a <- ggplot(lc_2004, aes(x = hiv_aids, y = life_expectancy))+
  geom_point(colour="#234BA0",size=0.5)+
  geom_smooth(colour="#4FE1C8")+
  labs(title="Hiv_aids level (-)")

b <- ggplot(lc_2004, aes(x = schooling, y = life_expectancy))+
  geom_point(colour="#234BA0",size=0.5)+
  geom_smooth(colour="#4FE1C8")+
  labs(title="Citizens' average education level (+)")

c <- ggplot(lc_2004, aes(x =alcohol, y = life_expectancy))+
  geom_point(colour="#234BA0",size=0.5)+
  geom_smooth(colour="#4FE1C8")+
  labs(title="Citizens' alcohol level (+)")

d <- ggplot(lc_2004, aes(x = total_expenditure, y = life_expectancy))+
  geom_point(colour="#234BA0",size=0.5)+
  geom_smooth(colour="#4FE1C8")+
  labs(title="Citizens' health related expenditure (+)")

e <- ggplot(lc_2004, aes(x =income_composition_of_resources, y = life_expectancy))+
  geom_point(colour="#234BA0",size=0.5)+
  geom_smooth(colour="#4FE1C8")+
  labs(title="Governments' ability to utilize resources (+)")

a+b+c+d+e + plot_annotation(
  title = "Different Factors having effect on life-expectancy in 2004"
)
```



However, when we run multi-variable linear regression on year 2004 dataset, we can see that, in 2004, Income Composition of Resources no longer has an impact on life expectancy. In its place, Schooling (education level), HIV rate, and under-five deaths are significantly impacting average life expectancy in a country.

#linear model on year 2000

```
lm3 <- lm(life_expectancy ~ bmi + polio + percentage_expenditure + hepatitis_b +measles +total_expenditu

summary(lm3)
```

```
##
## Call:
## lm(formula = life_expectancy ~ bmi + polio + percentage_expenditure +
##     hepatitis_b + measles + total_expenditure + diphtheria +
```

```
##     hiv_aids + gdp + income_composition_of_resources + schooling +
##     alcohol, data = lc_2000)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.558 -1.858  0.108  2.061  5.056
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      5.10e+01   3.26e+00   15.64  < 2e-16 ***
## bmi                              5.15e-02   2.69e-02    1.91  0.06171 .
## polio                           -1.77e-02   2.86e-02   -0.62  0.53937
## percentage_expenditure          -1.66e-03   1.87e-03   -0.89  0.37937
## hepatitis_b                      2.92e-02   1.71e-02    1.71  0.09378 .
## measles                          9.95e-05   5.28e-05    1.88  0.06558 .
## total_expenditure               -2.04e-02   2.26e-01   -0.09  0.92821
## diphtheria                       2.22e-02   2.04e-02    1.09  0.28265
## hiv_aids                        -5.03e-01   4.66e-02  -10.80  1.9e-14 ***
## gdp                              3.59e-04   3.13e-04    1.15  0.25764
## income_composition_of_resources  2.37e+00   2.55e+00    0.93  0.35644
## schooling                        1.12e+00   2.75e-01    4.06  0.00018 ***
## alcohol                          1.01e-01   1.78e-01    0.57  0.57254
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.28 on 48 degrees of freedom
## Multiple R-squared:  0.846,  Adjusted R-squared:  0.807
## F-statistic: 21.9 on 12 and 48 DF,  p-value: 1.78e-15
```

```
lm_new3=lm(life_expectancy ~ hepatitis_b + hiv_aids +schooling +income_composition_of_resources ,lc_200

summary(lm_new3)
```

```
##
## Call:
## lm(formula = life_expectancy ~ hepatitis_b + hiv_aids + schooling +
##     income_composition_of_resources, data = lc_2000)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.505 -2.392  0.099  2.992  5.903
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     52.4704     2.4574   21.35  < 2e-16 ***
## hepatitis_b                      0.0163     0.0151    1.08    0.284
## hiv_aids                        -0.5197     0.0463  -11.23  5.7e-16 ***
## schooling                        1.2966     0.2135    6.07  1.2e-07 ***
## income_composition_of_resources  3.9731     2.3247    1.71    0.093 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.41 on 56 degrees of freedom
## Multiple R-squared:  0.806,  Adjusted R-squared:  0.792
```

```
## F-statistic: 58.2 on 4 and 56 DF,  p-value: <2e-16
```

```
a<-ggplot(lc_2000, aes(x = schooling, y = life_expectancy))+
  geom_point(colour="#234BA0",size=0.5)+
  geom_smooth(colour="#4FE1C8")+
  labs(title="Citizens' average education level (+)")


b<-ggplot(lc_2000, aes(x =alcohol, y = life_expectancy))+
  geom_point(colour="#234BA0",size=0.5)+
  geom_smooth(colour="#4FE1C8")+
  labs(title="Citizens' average consumption of alcohol (+)")


c<-ggplot(lc_2000, aes(x =income_composition_of_resources, y = life_expectancy))+
  geom_point(colour="#234BA0",size=0.5)+
  geom_smooth(colour="#4FE1C8")+
  labs(title="Governments' ability to utilize resources (+)")

a+b+c + plot_annotation(
  title = "Different Factors having effect on life-expectancy in 2000"
)
```

Based on the multi-variable regression result and its visualization, we've noticed two interesting and counter-intuitive facts.

(1). Citizens' average alcohol consumption is positively impacting life expectancy in this country. We propose the theory that this is because alcohol can (a) improve cardiovascular health, (b) improve individuals' emotion, (c) indicate that these people are well financially.

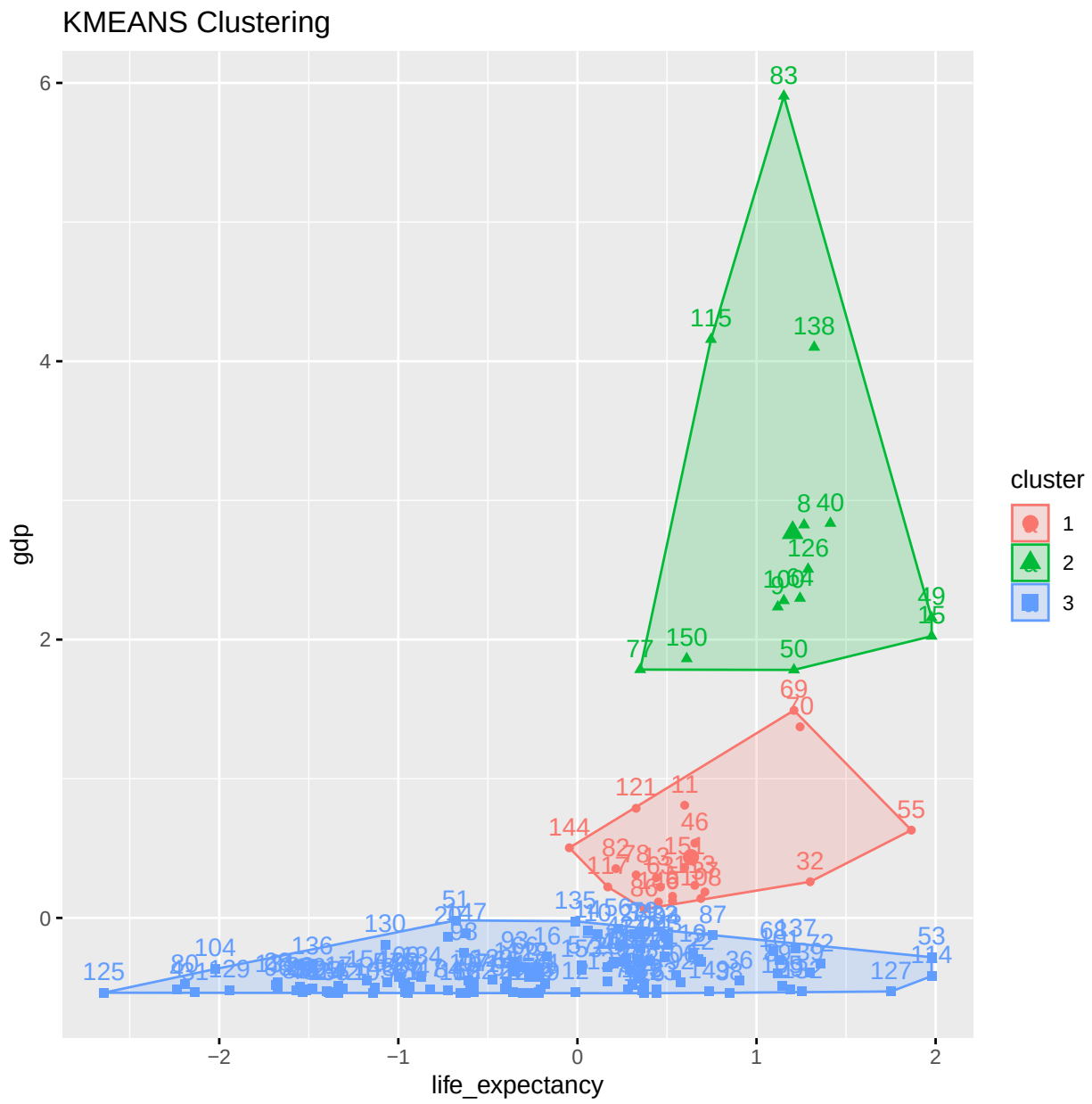
(2). Countries' GDP level is not significantly affecting this country's average life expectancy. We think it's because of potential outliers and will do further analysis in the next part.

Further Visualization

Before doing the actual analytics, we assumed that GDP would be correlated with life-expectancy, but the linear regression result suggested otherwise. Therefore, we did further clustering analysis on GDP and life-

expectancy and tried to identify the reasons behind why GDP and life-expectancy is not linearly correlated.

```
library(factoextra)

lifeExp_factors <- read_csv(file = "data viz dataset/CLEANED-Life Expectancy - potential factors.csv",sh

lifeExp_factors_cluster <- lifeExp_factors %>%
  clean_names() %>%
  filter(year==2014) %>%
  select("life_expectancy", "gdp") %>%
  na.omit()

# lifeExp_factors %>%
#   clean_names() %>%
#   filter(year==2014) %>%
#   select("country", "life_expectancy", "gdp") %>%
#   slice_max(gdp, n=10) %>%
#   arrange(desc(gdp))
#
# lifeExp_factors %>%
#   clean_names() %>%
#   filter(year==2014) %>%
#   select("country", "life_expectancy", "gdp") %>%
#   slice_max(life_expectancy, n=10) %>%
#   arrange(desc(life_expectancy))

model_kmeans <- eclust(lifeExp_factors_cluster, "kmeans", k=3, graph = TRUE)
```

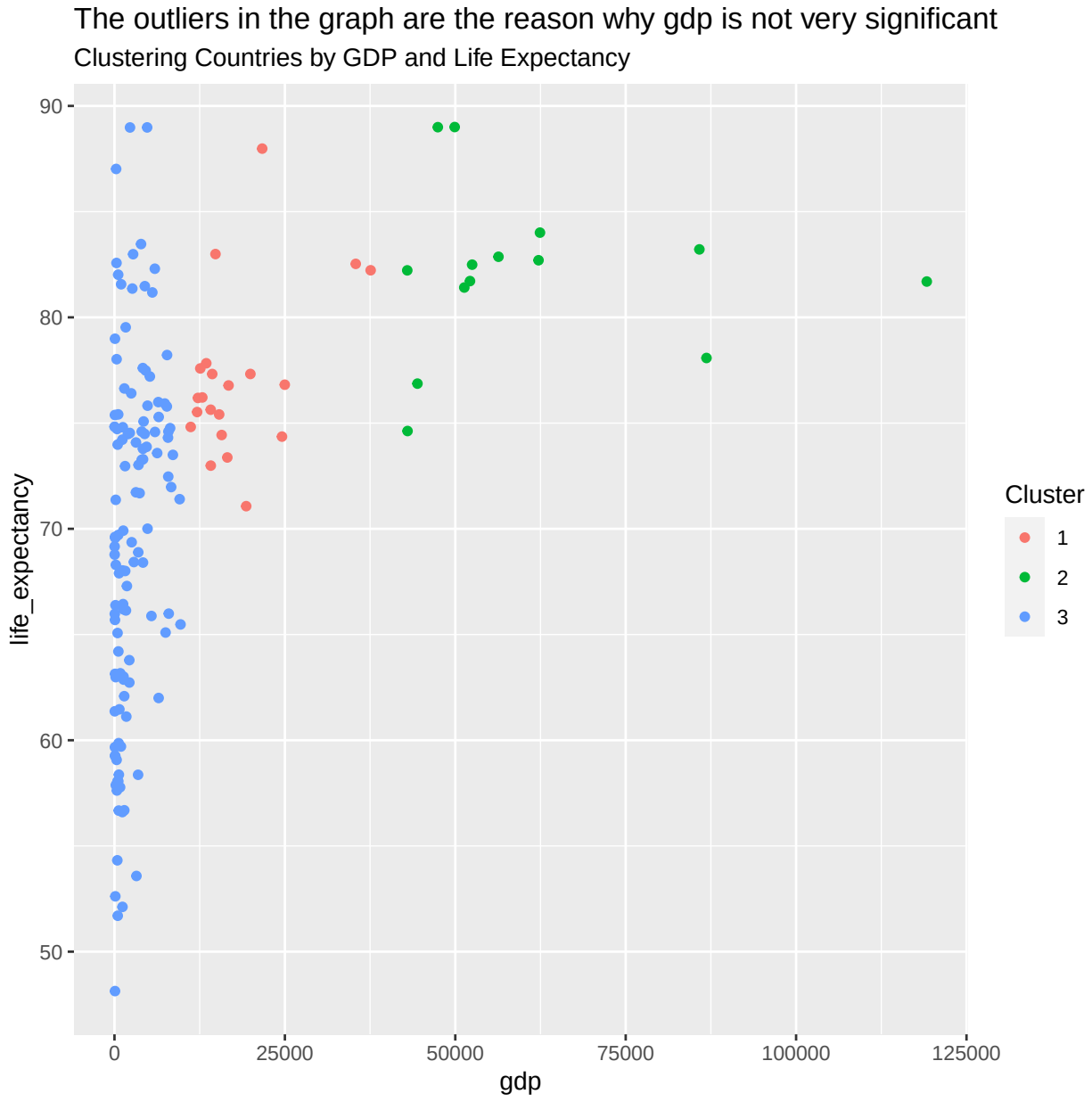



```
#Let's check the components of this object.
summary(model_kmeans)
```

```
##              Length Class  Mode
## cluster      155    -none- numeric
## centers        6    -none- numeric
## totss          1    -none- numeric
## withinss       3    -none- numeric
## tot.withinss   1    -none- numeric
## betweenss      1    -none- numeric
## size           3    -none- numeric
## iter           1    -none- numeric
## ifault         1    -none- numeric
```

```
## clust_plot     9    gg     list
## silinfo        3    -none- list
## nbclust        1    -none- numeric
## data           2    tbl_df list
```

```
library(ggpubr)

lifeExp_factors_withClusters<-mutate(lifeExp_factors_cluster, cluster = as.factor(model_kmeans$cluster))

ggplot(lifeExp_factors_withClusters, aes(x = gdp, y = life_expectancy, color =  as.factor(cluster))) +
  geom_jitter()+
  labs(color = "Cluster",
       title = "The outliers in the graph are the reason why gdp is not very significant",
       subtitle = "Clustering Countries by GDP and Life Expectancy")
```



Using kmeans method, we can tell that there are three groups of countries: a) countries with low gdp and lower life expectancy; b) countries with median gdp and higher life expectancy; c) countries with high gdp and high life expectancy. However, there are outliers in the (a) group. Citizens in some countries have high life expectancy despite low gdp. After examining these countries, we found that they are exercising a healthier life style. Maybe that's the reason why gdp is not significant in the regression models.

(5) Discussion and critique. What did you learn about the process of carrying out your project? What went wrong, and how could you improve it next time? For instance, did any sources of bias creep into your survey or experiment? What advice would you give future students?

(1)At first, we used every variable in the data set to do the multi-variable regression, without assessing potential perfect fit and co-linearity problem. For example, "Schooling" and "Income" are likely to be linearly related; "Infant mortality" and "under-five deaths" are likely to be linearly relater; and "Adult

mortality" is likely to perfectly explain the dependent variable "life expectancy". Therefore, we took these variables out of the model.

(2)For features like "infant mortality", and "under-five deaths", even if they are significantly related to the dependent variable, it cannot generate any useful explanatory result. There fore we should focus on variables that could potentially be well explained.

For future references, we should be more careful when dealing with feature choice and feature engineering, especially focus on variable's explaining power and multi-colinearity problem.