EDA Report: Life Expectancy

I-Ting Cheng

June 15, 2019

Import your data

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## Warning: package 'ggplot2' was built under R version 3.5.3

## Warning: package 'corrplot' was built under R version 3.5.2

## corrplot 0.84 loaded

##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha

## Warning: package 'tidyr' was built under R version 3.5.3

## Warning: package 'tinytex' was built under R version 3.5.3
```

Set up your questions

In this study, we would like to explore how different factors are influencing the life expectancy. We would like to identify the relationships between these factors and life expectancy (Is it strong or weak? Is it positive or negative?) Below are some questions we'd like to answer after conducting our exploratory data analysis: 1. Does life expectancy changes throughout the years? 2. Does different country have distinct life expectancy? 3. Does population impact the life expectancy? 4. Does income(GDP) have to do with life expectancy? 5. Does every region have different life expectancy?

Describe your data

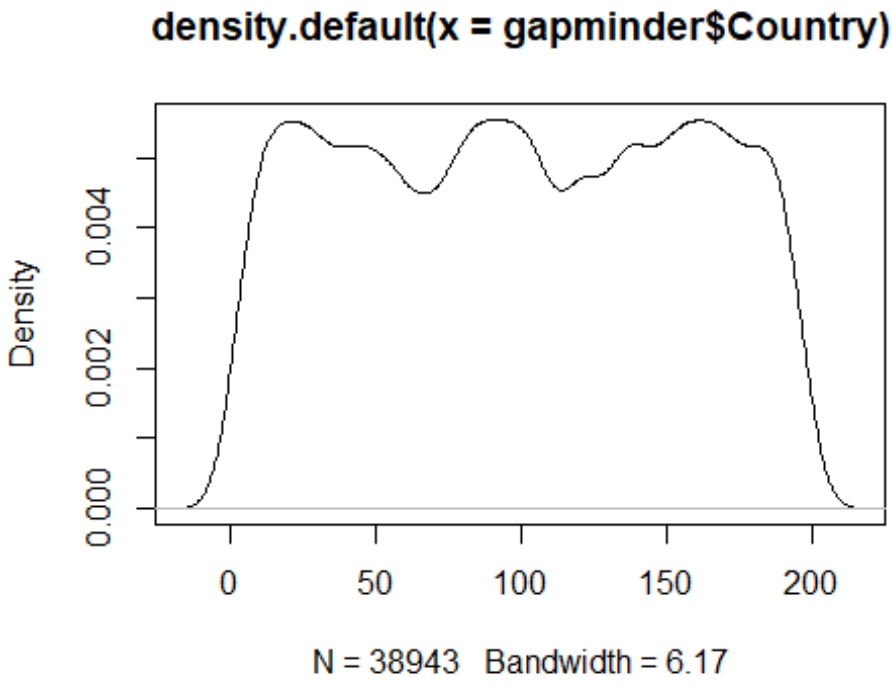
This dataset has 41284 observations with 6 variables. In order to identify which factors have the most impact on life expectancy, we will conduct an correlation plot first to see how different factors impact life expectancy. We will then conduct exploratory analysis to visualize how life expectancy is influenced by country, year, population, income and region.

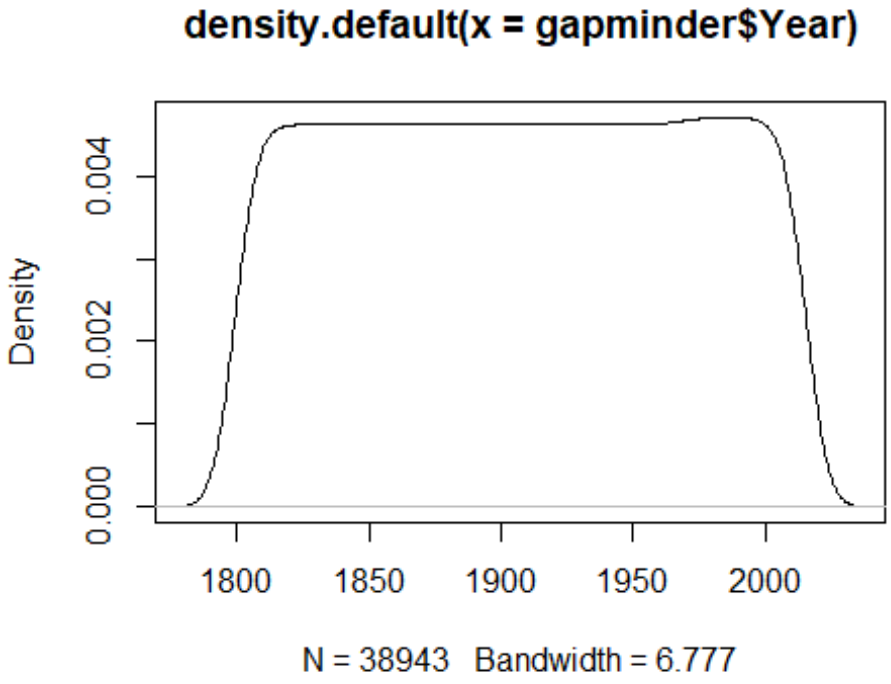
o how many observations: 41284 observations o how many variables: 6 variables o type of variables: $ Country : Factor // $ Year : Integer // $ life : Numeric // $ population: Factor // $ income : Integer // $ region : Factor // o how disperse is your data, range of variables $ Country : 197 countries in the dataset // $ Year : ranging from 1800 - 2015 // $ life : ranging from 1 - 84.10, the mean is 42.88// $ population: a lot of missing values, will wrangle // $ income : ranging from 142 - 182668, the mean is 4571 // $ region : America - 7961 data, East Asia & Pacific - 6256, Europe & Central - Asia has 10468, Middle East & North Africa has 4309, South Asia - 1728, Sub-Saharan Africa - 10562 // o data wrangling: Viewing the dataset, theres's a lot of missing data in here, so what we're going to do is to impute the missing data.

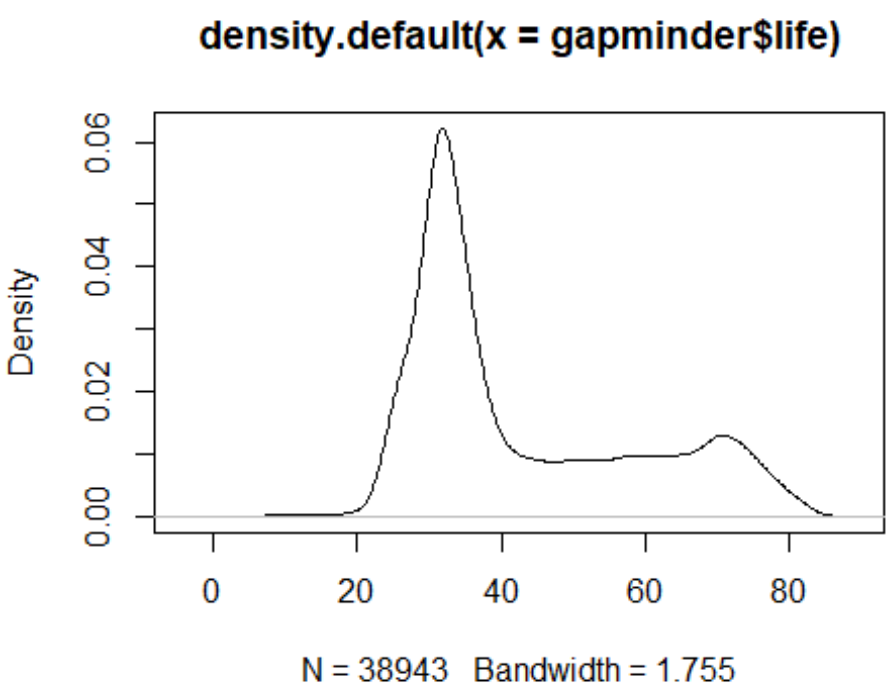
```
## Warning: NAs introduced by coercion
```

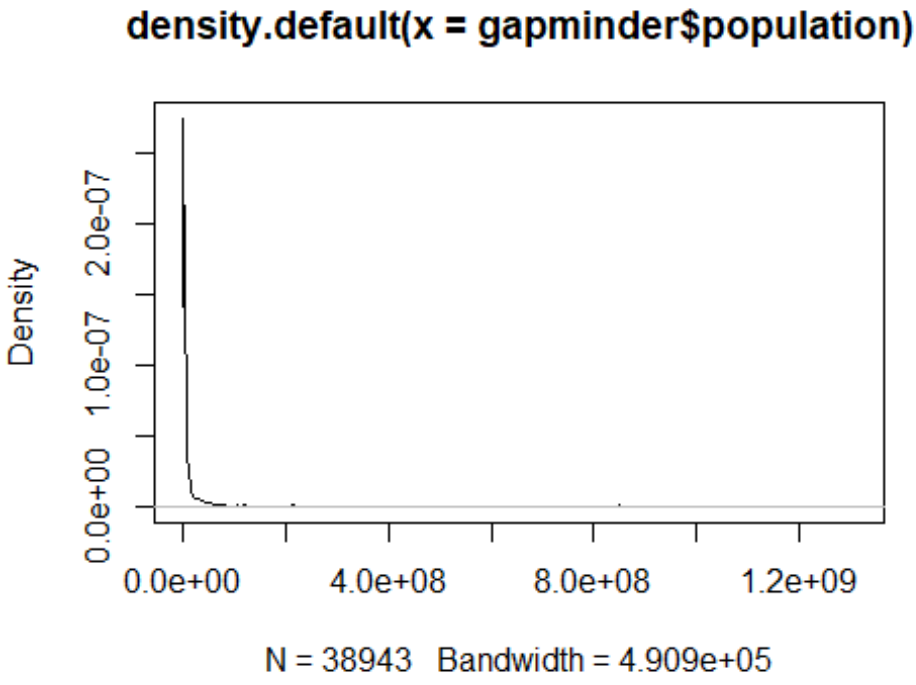
o preprocessing steps: In order to identify different variables' relationships with life expectancy, I first transfer all the variables into numeric and run a correlation plot.

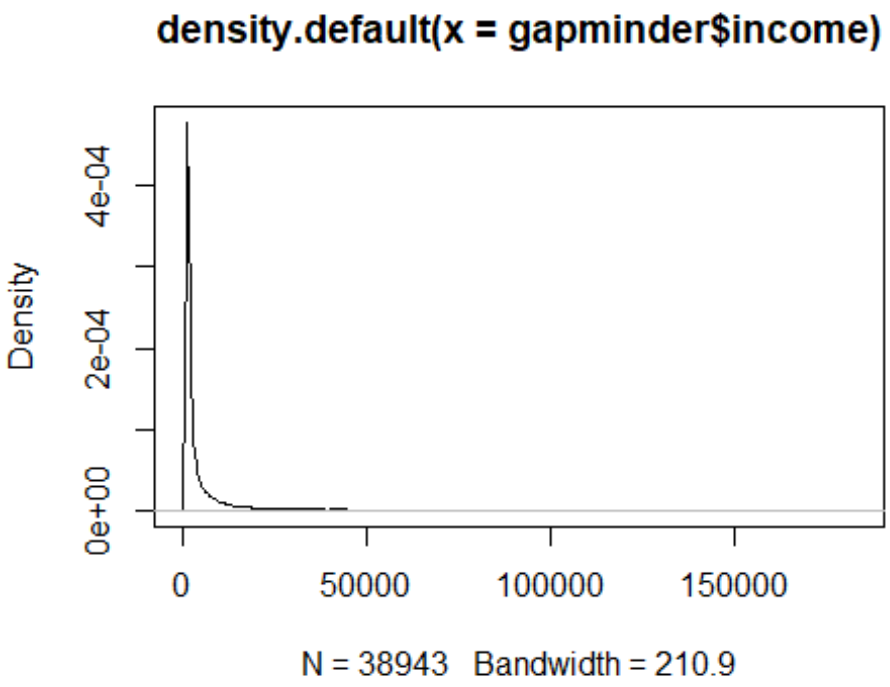
Plot out your data

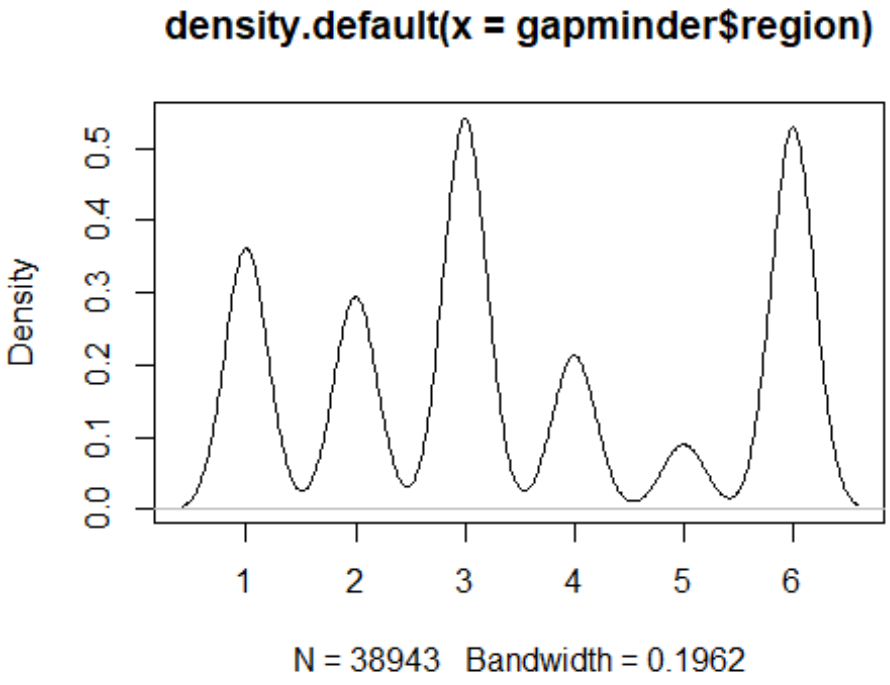





density.default(x = gapminder$life)
Density
0.00
0.02
0.04
0.06
0
20
40
60
80
N = 38943 Bandwidth = 1.755


density.default(x = gapminder$population)
Density
0.0e+00
1.0e-07
2.0e-07
0.0e+00
4.0e+08
8.0e+08
1.2e+09
N = 38943 Bandwidth = 4.909e+05


density.default(x = gapminder$income)
Density
0e+00
2e-04
4e-04
0
50000
100000
150000
N = 38943 Bandwidth = 210.9


density.default(x = gapminder$region)
Density
0.0
0.1
0.2
0.3
0.4
0.5
1
2
3
4
5
6
N = 38943 Bandwidth = 0.1962

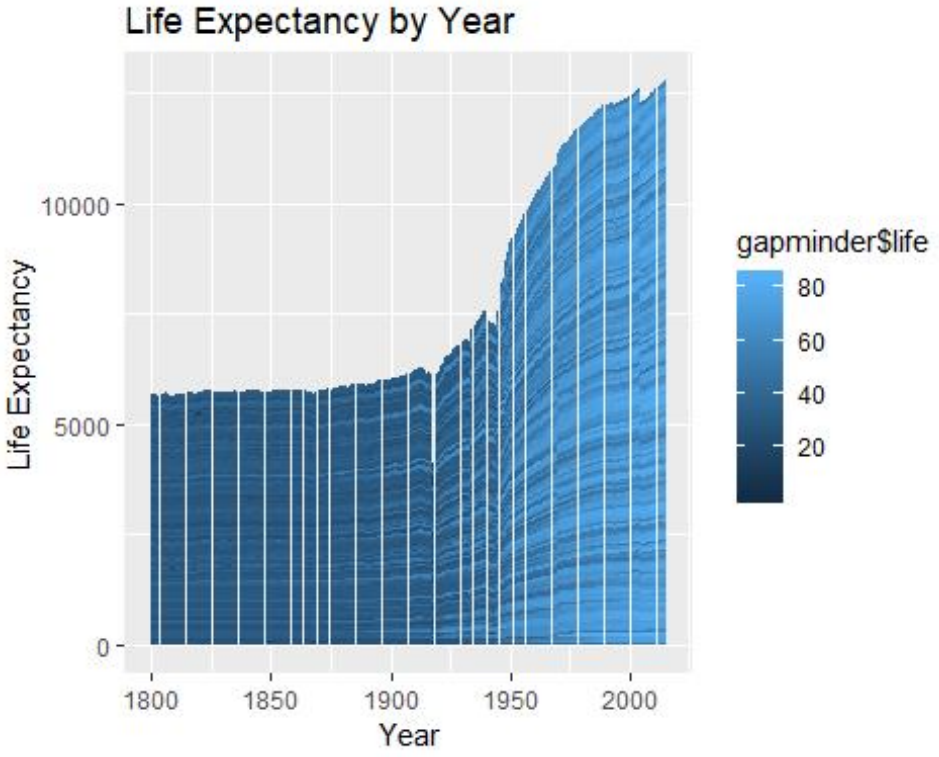
correlation plot

In this correlation plot, we can easily see which variables have the most impact on life expectancy. In this plot, year has the most significant impact on life expectancy,followed by income, region and population. Our next step is to see how these variables interact with life expectancy and visulize the result.

Life Expectancy by Year

When viewing the plot, we can see that life expectancy has been growing years over years. But in year 1920s, there's a huge decrease in life expectancy, perhaps it has something to do with the economic background during that period. In 1950 - 2000, the life expectancy bloom dramatically, meaning people have a greater life

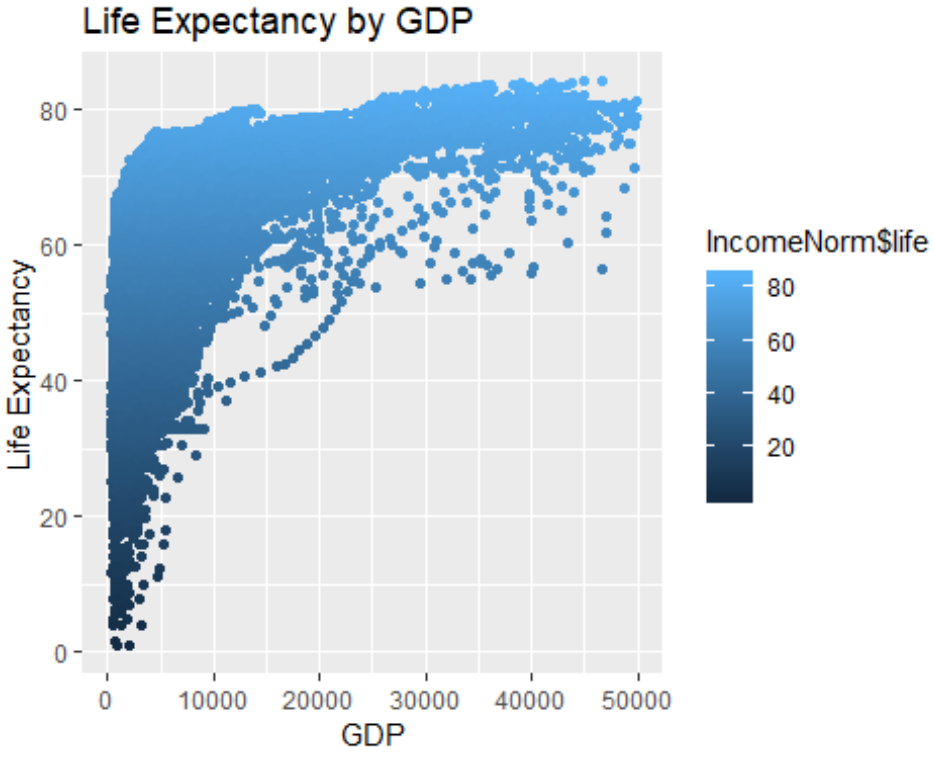
expectancy during recent years.



Life Expectancy by Income

When viewing the plot, we can see that when GDP gets higher, the life expectancy gets higher as well. However, in this dataset, most of the GDP are ranging from 883

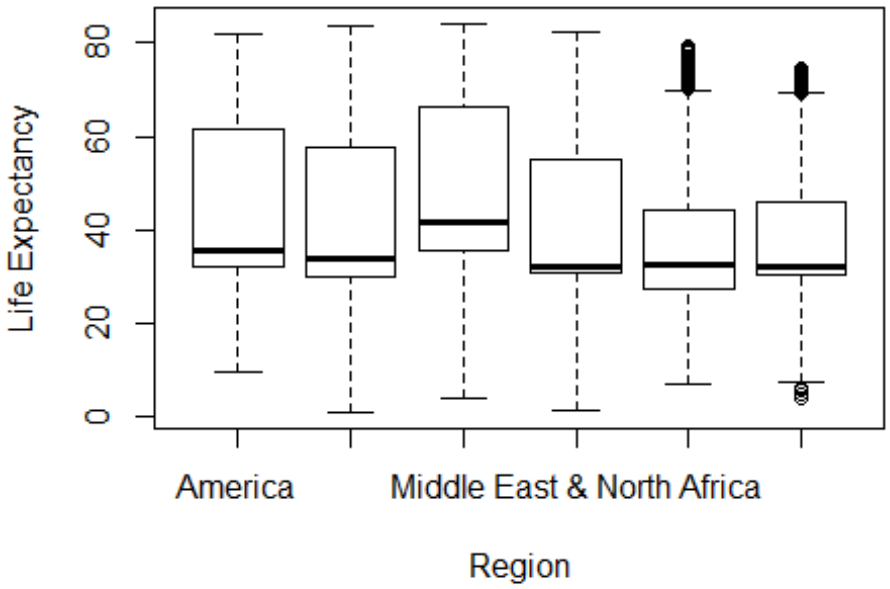
(1st Qu) - 3483 (3rd Qu) which reflects the normal phenomemon.



Life Expectancy by region

When viewing the plot, we can see that Region 3 (Europe & Central Asia) has the higher average life expectancy and other regions have very identical average life expectancy. That is to say, region doesn't play an significant part of influencing

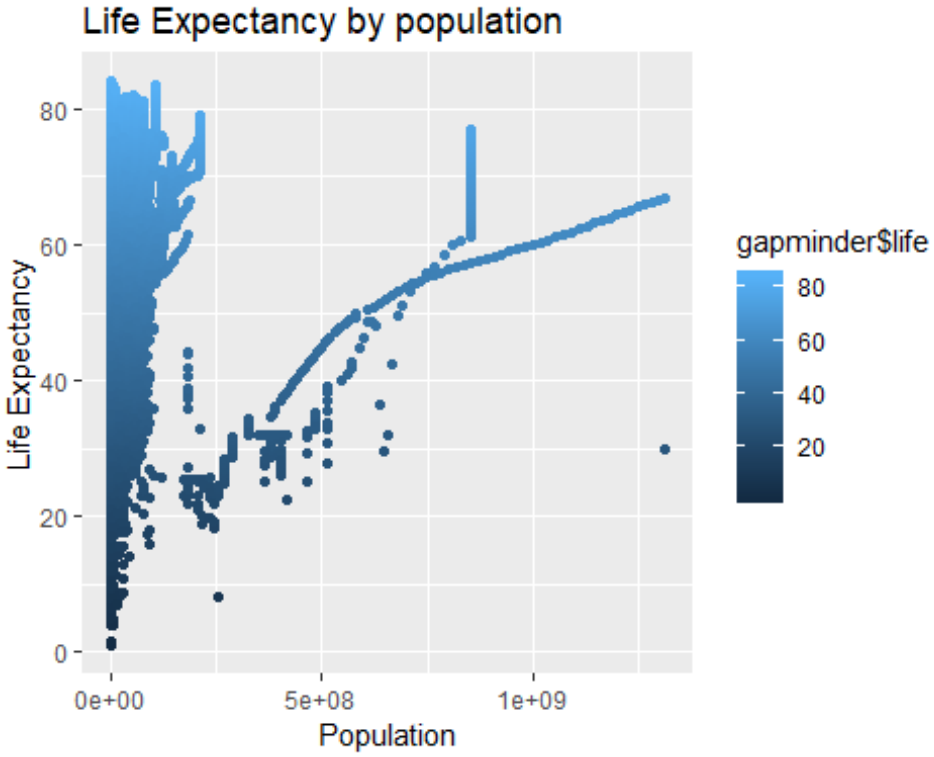
people's life expectancy.



Life Expectancy by population

When viewing the plot, we can see that when population gets bigger, the life expectancy doesn't necessarily grow higher. And most of the population falls within the same range and the life expectancy varies among different population.

Therefore, no clear clue that population has something to do with life expectancy.



Interpret the result

1. Does life expectancy changes throughout the years? Yes, life expectancy has been growing years over years, especially in 1950 - 2000, life expectancy grew dramatically than it'd ever been.

2. Does different country influence the life expectancy? We don't see a significant relationship between country and life expectancy when viewing the corelation plot. We believe that country has little impact on life expectancy.

3. Does population influence life expectancy? We do see a little relationship between population and life expectancy when viewing the corelation plot, however, when viewing into the Life Expectancy by population plot, most of the population falls within the same range yet the life expectancy varies in that range. Therefore, we believe population doesn't have a significant impact on life expectancy.

4. Does income(GDP) have to do with life expectancy? Yes, when viewing the Life expectancy by GDP plot, we can see that when GDP gets higher, the life expectancy gets higher. That is to say, GDP is a significant index for life expectancy.

5. Does every region have different life expectancy? Not necessarily, most of the regions have a similar average life expectancy. Although region 3 (Europe &

Central Asia) has a higher average life expectancy, we cannot make a conclusion that region is a significant factor influencing the life exepectancy.