Statistical analyses and visualization in R: I - HT21

Final Project

Sanjiv Kumar

12 January 2022

Introduction

We are analyzing the data set created by combining data from Gapminder (Ref. 1). The data was downloaded in the form of CSV files for the population, Human Development Index (HDI), Income per person (Income), Life Expectancy (LifeExp), The Sustainable Development Index (SDI), Region and Income Group of various countries. These parameters are defined as follows:

Population: Data for the population was downloaded from Gapminder population data (Ref. 2).

Human Development Index (HDI): HDI is used to rank various countries by the level of human development and includes level of health, education and living standard (Ref. 3).

Income per person (Income): This is calculated as gross domestic product per person adjusted for inflation and is converted to dollars (from 2017) using power parity rates (Ref. 4).

Life Expectancy (LifeExp): It is the average number of years a newborn child would live, considering the current mortality patterns (Ref. 5).

The Sustainable Development Index (SDI): This index assesses the ecological efficiency of nations in delivering human development considering "development index" and "ecological impact index" (Ref. 6).

GDP per capita (GDPPC): is calculated as GDP divided by midyear population, data is in constant of dollars (from 2010) (Ref. 7).

Metadata: Income group and Region in the data were taken from the classification for The World Bank GNI (Ref. 8).

Note: Rmarkdown for the report is attached at the end of the document as Appendix.

Data cleaning

All data imported from CSV were read into R and cleaned (i.e., values in K, M, B etc. were converted to e3, e6, e9 respectively. The data is converted to tidy format using tidyverse (Ref. 9). Since, after combining all the data it became too large (>2.5 Gb) and therefore it was filtered only for few selected years (i.e., 1990, 1995, 2000, 2005, 2015, 2019). Missing data (i.e., `NA`) were removed.

Examining measures of cental tendency across various parameters

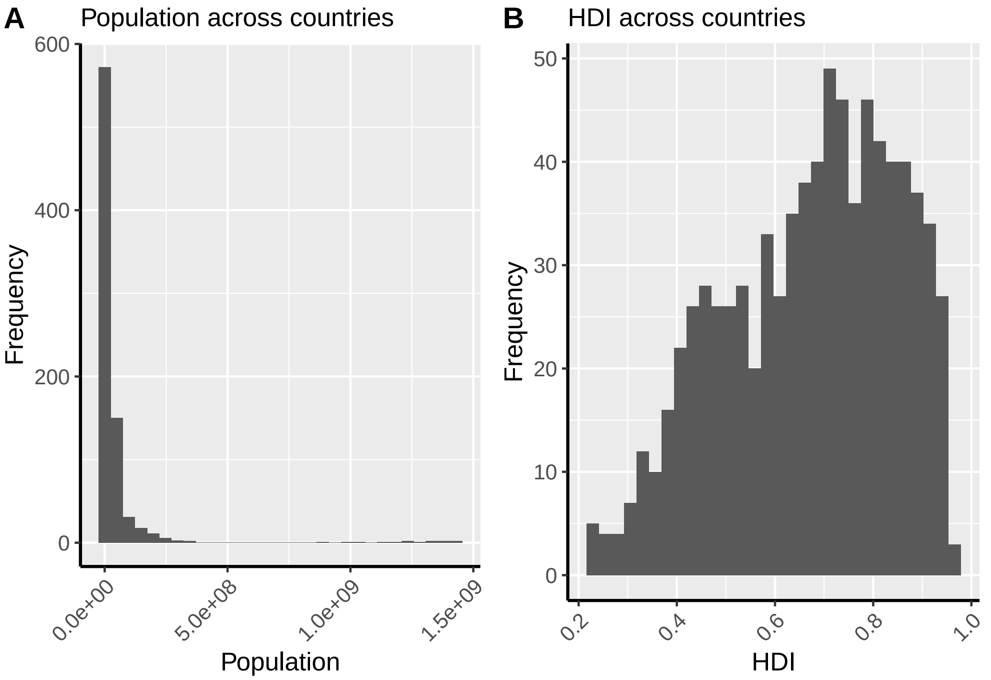
Measures of central tendency (i.e., Mean, Median, Mode, SD, StdErr, and Var) were calculated (Table 1).

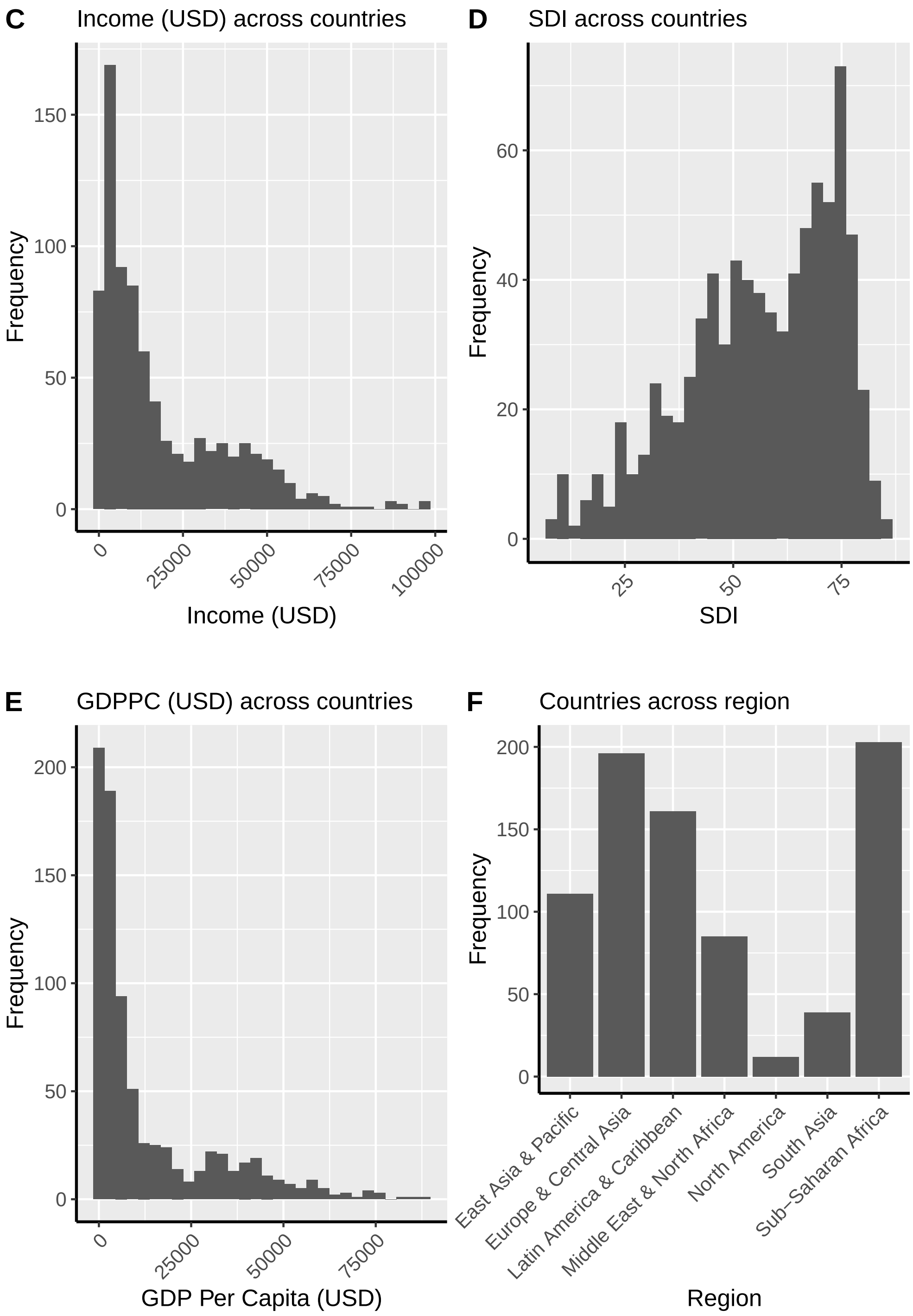
Table 1: Measures of central tendency across various parameters
(From 1990 to 2019)

Mean	Median	Mode	SD	StdErr	Var
Population					
47723866.171	1.00e+07	1.02e+07	1.649468e+08	5806402.721	2.720745e+16
HDI					
0.671	6.97e-01	6.24e-01	1.750000e-01	0.006	3.000000e-02
Income					
18263.062	1.04e+04	1.01e+04	1.891111e+04	665.703	3.576302e+08
LifeExp					
69.992	7.26e+01	7.68e+01	9.533000e+00	0.336	9.087000e+01
SDI					
56.091	5.85e+01	7.59e+01	1.755400e+01	0.618	3.081600e+02
GDPPC					
13141.736	4.69e+03	1.81e+04	1.735073e+04	610.775	3.010477e+08

Data visualization (Histogram) across various parameters

Histograms were generated to examine the distribution of data (Figure 1).




C
Income (USD) across countries
Frequency
0
50
100
150
0
25000
50000
75000
100000
Income (USD)
D
SDI across countries
Frequency
0
20
40
60
25
50
75
SDI
E
GDPPC (USD) across countries
Frequency
0
50
100
150
200
0
25000
50000
75000
GDP Per Capita (USD)
F
Countries across region
Frequency
0
50
100
150
200
East Asia & Pacific
Europe & Central Asia
Latin America & Caribbean
Middle East & North Africa
North America
South Asia
Sub−Saharan Africa
Region

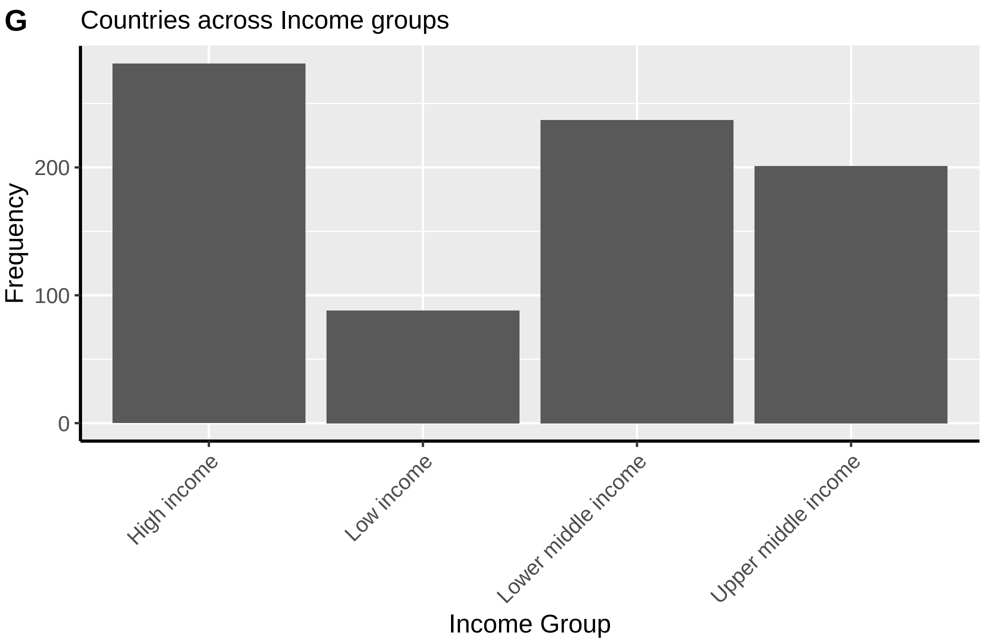


Figure 1: Distribution of data across various parameters

Questions

---

1. At least one of the questions must be a trend analysis, e.g. using correlation or linear regression.

Has HDI increased globally from 1990 to 2019? The Q-Q plot, the residuals appear to be on straight line, and Residuals vs Fitted graphs shows spread in the largest group is not more than three times the spread in the smallest group, therefore, data was not transformed. Regression model was generated for HDI over Years from 1990 to 2019. From the figure (Figure 2 A) it is evident that there is a liner increase in HDI from 1990 to 2019 globally. Statistics Report:

- r-squared: 0.0888317
- b (regression coefficient): -10.1682316
- SEb: 6.0596425×10^{-4}
- t: 8.9206862
- df: 2, 805, 2
- p: $3.0762017 \times 10^{-18}$

Has Income increased world-wide from 1990 to 2019? The Q-Q plot, the residuals appear to be on straight line, and Residuals vs Fitted graphs shows spread in the largest group is not more than three times the spread in the smallest group, therefore, data was not transformed. Regression model was generated for Income over Years from 1990 to 2019. From the figure (Figure 2 B) it is evident that there is a liner increase in Income from 1990 to 2019 globally. Statistics Report:

- r-squared: 0.0318663

- b (regression coefficient): -6.9382504×10^5
- SEb: 67.6833817
- t: 5.2468728
- df: 2, 805, 2
- p: 1.979894×10^{-7}

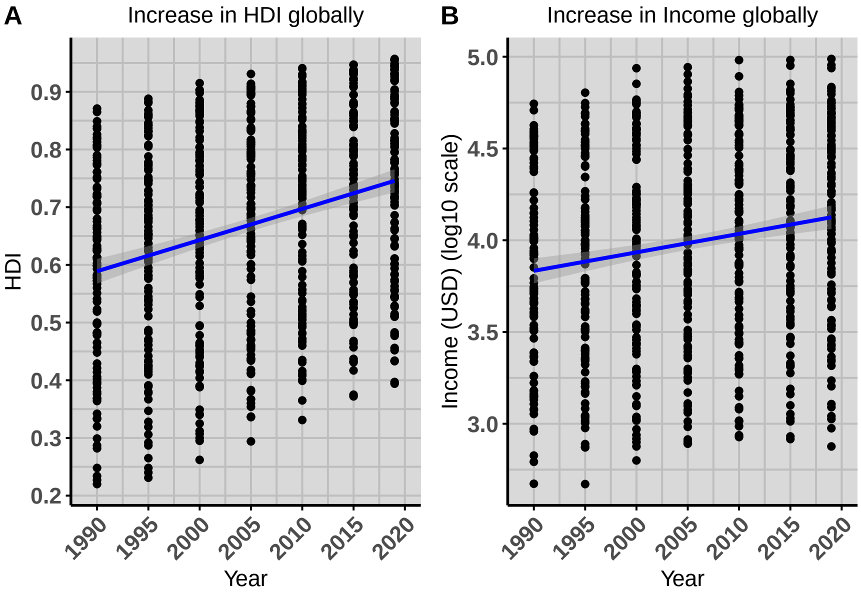
Has LifeExp increased globaly from 1990 to 2019? The Q-Q plot, the residuals appear to be on straight line, and Residuals vs Fitted graphs shows spread in the largest group is not more than three times the spread in the smallest group, therefore, data was not transformed. Regression model was generated for LifeExp over Years from 1990 to 2019. From the figure (Figure 2 C) it is evident that there is a liner increase in LifeExp from 1990 to 2019 globally. Statistics Report:

- r-squared: 0.065619
- b (regression coefficient): -440.0939368
- SEb: 0.0335173
- t: 7.5896745
- df: 2, 805, 2
- p: $8.8769847 \times 10^{-14}$

Has GDPPC increased from 1990 to 2019?

The Q-Q plot, the residuals appear to be on straight line, and Residuals vs Fitted graphs shows spread in the largest group is not more than three times the spread in the smallest group, therefore, data was not transformed. Regression model was generated for GDPPC over Years from 1990 to 2019. From the figure (Figure 2 D) it is evident that there is a liner increase in GDPPC from 1990 to 2019 globally. Statistics Report:

- r-squared: 0.0133422
- b (regression coefficient): -4.2047853×10^5
- SEb: 62.6899978
- t: 3.4495319
- df: 2, 805, 2
- p: 5.9076414×10^{-4}



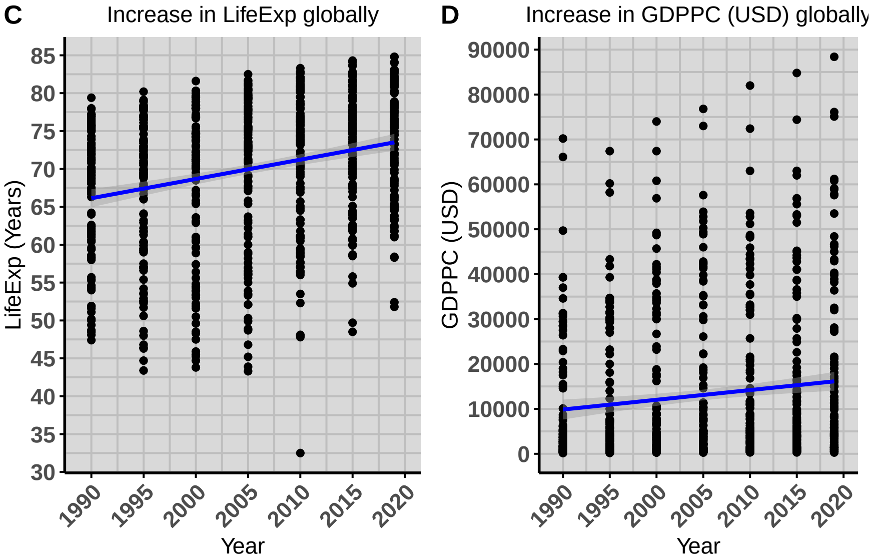


Figure 2: Linear Regression analysis

---

2. At least one of the questions must deal with differences between two groups, e.g. using t-test or non-parametric alternatives

---

Non-parametric test: Wilcoxon's rank-sum test

Is there's a difference in the Life expectancy between Low Income and High Income group (IncomeGroup)? The life expectancy data is not normally distributed data, it is left skewed (Figure 3A) and therefore, non-parametric test Wilcoxon's rank-sum test would be used in this case. This is also supported by shapiro.test, with p-value < 2.2e-16.

Wilcoxon rank-sum test Since, the p-value $1.7531938 \times 10^{-45}$ is less then than the significance level 0.05, it is therefore concluded that there is significant difference between the life expectancy between the two group i.e. low and high income. W = 24724.

Distribution by Income Group

Graph (Figure 3B) representing the difference between low and high income countries. The boxplot plot shows that the life expectancy of high income group is higher than low income group.

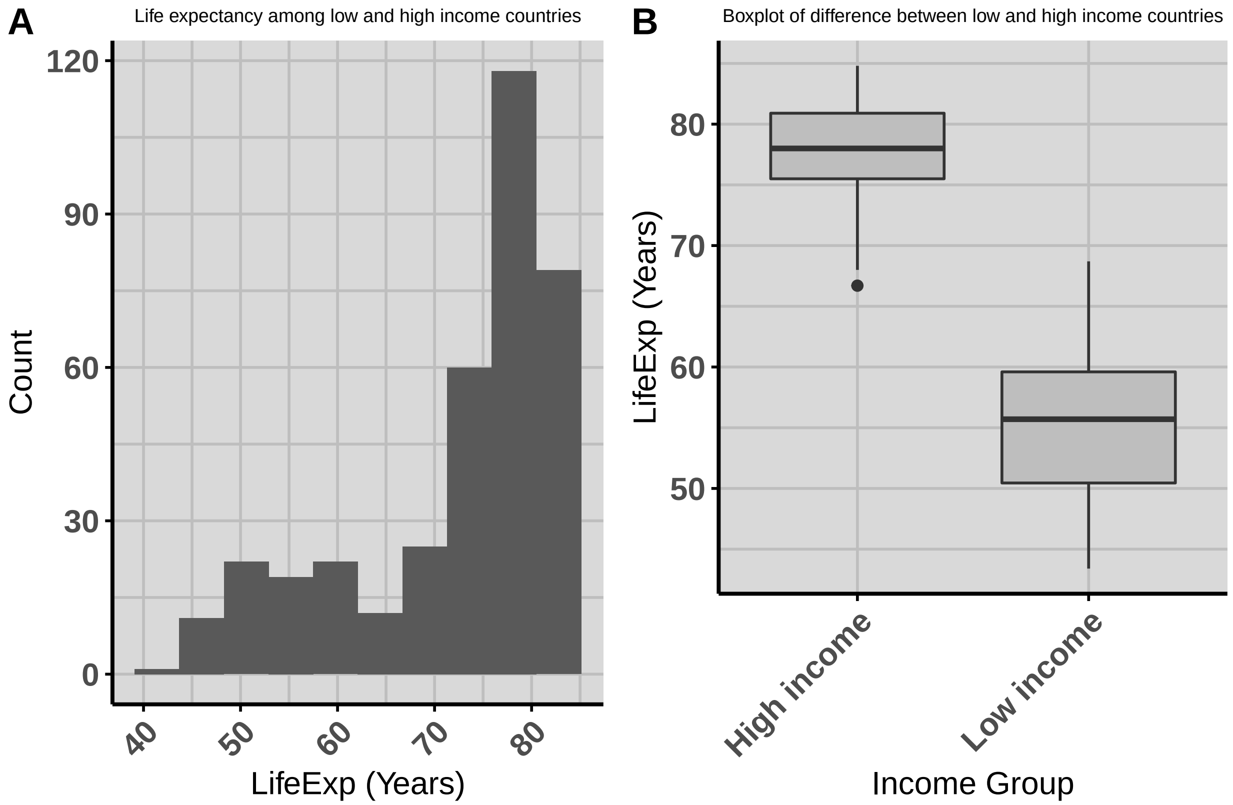


Figure 3: Histogram and boxplot of difference between low and high income countries

---

3. At least one of the questions must deal with differences between more than two groups, e.g. using different versions of ANOVAs or non-parametric alternatives.

Is there is a effect of `Region` and `IncomeGroup` on life expectancy? Since we are examining the effect of two different categorical independent variable (`Regions` and `IncomeGroup`) and one continuous dependent variable (`LifeExp (Years)`) (Table 2), two-way ANOVA was performed. The data was visvualized using boxplot (Figure 4).

Table 2: Effect of Region and IncomeGroup on Life Expectancy
Measured central tendencies

Region	IncomeGroup	variable	n	mean	sd
East Asia & Pacific	High income	LifeExp (Years)	28	80.714	2.711
East Asia & Pacific	Lower middle income	LifeExp (Years)	55	66.696	4.553
East Asia & Pacific	Upper middle income	LifeExp (Years)	28	72.079	3.877
Europe & Central Asia	High income	LifeExp (Years)	168	78.199	3.460
Europe & Central Asia	Upper middle income	LifeExp (Years)	28	73.725	3.107
Latin America & Caribbean	High income	LifeExp (Years)	28	75.075	2.648
Latin America & Caribbean	Lower middle income	LifeExp (Years)	42	68.840	7.944
Latin America & Caribbean	Upper middle income	LifeExp (Years)	91	74.915	3.486
Middle East & North Africa	High income	LifeExp (Years)	45	75.471	4.647

Middle East & North Africa	Lower middle income	LifeExp (Years)	21	72.610	3.225
Middle East & North Africa	Upper middle income	LifeExp (Years)	19	72.989	3.672
North America	High income	LifeExp (Years)	12	79.100	2.234
South Asia	Low income	LifeExp (Years)	4	60.825	2.435
South Asia	Lower middle income	LifeExp (Years)	35	66.886	5.329
Sub-Saharan Africa	Low income	LifeExp (Years)	84	55.011	6.123
Sub-Saharan Africa	Lower middle income	LifeExp (Years)	84	58.139	6.308
Sub-Saharan Africa	Upper middle income	LifeExp (Years)	35	62.251	7.315

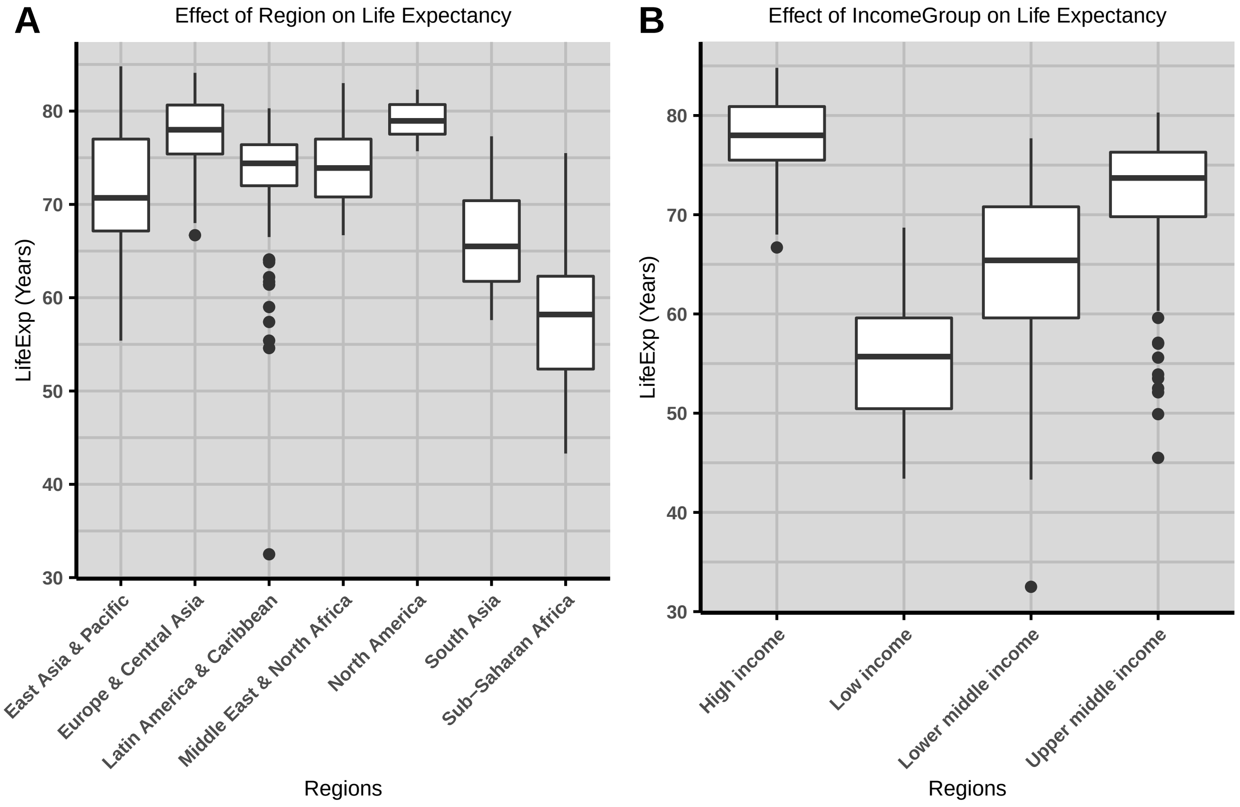


Figure 4: Effect of Region and IncomeGroup on Life Expectancy

In `Q-Q plot` looks good but the difference of the spread in largest group is more than three times the spread in the group with the smallest spread as appears from `Residuals vs Fitted`. Therefore, the data was log transformed.

Anova Table:

```
## Anova Table (Type II tests)
##
## Response: log10(‘LifeExp (Years)‘)
##              Sum Sq  Df F value    Pr(>F)
## Region      0.45112   6  61.606 < 2.2e-16 ***
## IncomeGroup 0.23270   3  63.556 < 2.2e-16 ***
## Residuals   0.97268 797
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Both Region and IncomeGroup has significant effect on life expectancy.

Statistics Report:

- Test: Two-way ANOVA
- F-value: 288.7311579
- df: 3
- p-value: $5.0687813 \times 10^{-196}$

Pair-wise comparison using Tukey post-hoc test using glht from multcomp. The table below show the significant differences among the groups.

Tukey post-hoc test:

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = log10(‘LifeExp (Years)‘) ~ Region + IncomeGroup,
##     data = all_datatidy_90_19_noNA, na.action = na.omit)
##
## Linear Hypotheses:
##                                                               Estimate
## Europe & Central Asia - East Asia & Pacific == 0             0.008798
## Latin America & Caribbean - East Asia & Pacific == 0         0.005142
## Middle East & North Africa - East Asia & Pacific == 0        0.003660
## North America - East Asia & Pacific == 0                     0.014908
## South Asia - East Asia & Pacific == 0                       -0.009450
## Sub-Saharan Africa - East Asia & Pacific == 0               -0.070242
## Latin America & Caribbean - Europe & Central Asia == 0      -0.003657
## Middle East & North Africa - Europe & Central Asia == 0     -0.005138
## North America - Europe & Central Asia == 0                   0.006110
## South Asia - Europe & Central Asia == 0                     -0.018248
## Sub-Saharan Africa - Europe & Central Asia == 0             -0.079040
## Middle East & North Africa - Latin America & Caribbean == 0 -0.001481
## North America - Latin America & Caribbean == 0               0.009767
## South Asia - Latin America & Caribbean == 0                 -0.014591
## Sub-Saharan Africa - Latin America & Caribbean == 0         -0.075384
## North America - Middle East & North Africa == 0              0.011248
## South Asia - Middle East & North Africa == 0                -0.013110
## Sub-Saharan Africa - Middle East & North Africa == 0        -0.073902
## South Asia - North America == 0                             -0.024358
## Sub-Saharan Africa - North America == 0                     -0.085150
## Sub-Saharan Africa - South Asia == 0                        -0.060792
## Low income - High income == 0                               -0.075757
## Lower middle income - High income == 0                      -0.051274
## Upper middle income - High income == 0                      -0.019508
## Lower middle income - Low income == 0                        0.024483
## Upper middle income - Low income == 0                        0.056250
## Upper middle income - Lower middle income == 0               0.031766
##                                                             Std. Error t value
## Europe & Central Asia - East Asia & Pacific == 0              0.004826   1.823
## Latin America & Caribbean - East Asia & Pacific == 0          0.004429   1.161
## Middle East & North Africa - East Asia & Pacific == 0         0.005168   0.708
## North America - East Asia & Pacific == 0                      0.011002   1.355
## South Asia - East Asia & Pacific == 0                         0.006742  -1.402
## Sub-Saharan Africa - East Asia & Pacific == 0                 0.004697 -14.956
## Latin America & Caribbean - Europe & Central Asia == 0        0.004494  -0.814
## Middle East & North Africa - Europe & Central Asia == 0       0.004720  -1.088
## North America - Europe & Central Asia == 0                    0.010404   0.587
## South Asia - Europe & Central Asia == 0                       0.007341  -2.486
## Sub-Saharan Africa - Europe & Central Asia == 0               0.005339 -14.804
## Middle East & North Africa - Latin America & Caribbean == 0   0.004884  -0.303
## North America - Latin America & Caribbean == 0                0.010892   0.897
## South Asia - Latin America & Caribbean == 0                   0.006735  -2.167
## Sub-Saharan Africa - Latin America & Caribbean == 0           0.004481 -16.824
## North America - Middle East & North Africa == 0               0.010917   1.030
## South Asia - Middle East & North Africa == 0                  0.007326  -1.789
## Sub-Saharan Africa - Middle East & North Africa == 0          0.005405 -13.673
## South Asia - North America == 0                               0.012332  -1.975
## Sub-Saharan Africa - North America == 0                       0.011272  -7.554
## Sub-Saharan Africa - South Asia == 0                          0.006361  -9.557
## Low income - High income == 0                                 0.006214 -12.191
## Lower middle income - High income == 0                        0.004396 -11.663
## Upper middle income - High income == 0                        0.003936  -4.956
## Lower middle income - Low income == 0                         0.004914   4.983
## Upper middle income - Low income == 0                         0.005515  10.199
## Upper middle income - Lower middle income == 0                0.003726   8.526
##                                                             Pr(>|t|)
## Europe & Central Asia - East Asia & Pacific == 0                0.597
## Latin America & Caribbean - East Asia & Pacific == 0            0.951
## Middle East & North Africa - East Asia & Pacific == 0           0.998
## North America - East Asia & Pacific == 0                        0.885
## South Asia - East Asia & Pacific == 0                           0.864
## Sub-Saharan Africa - East Asia & Pacific == 0                  <0.001 ***
## Latin America & Caribbean - Europe & Central Asia == 0          0.995
## Middle East & North Africa - Europe & Central Asia == 0         0.967
## North America - Europe & Central Asia == 0                      1.000
## South Asia - Europe & Central Asia == 0                         0.189
## Sub-Saharan Africa - Europe & Central Asia == 0                <0.001 ***
## Middle East & North Africa - Latin America & Caribbean == 0     1.000
## North America - Latin America & Caribbean == 0                  0.991
## South Asia - Latin America & Caribbean == 0                     0.358
## Sub-Saharan Africa - Latin America & Caribbean == 0            <0.001 ***
## North America - Middle East & North Africa == 0                 0.977
## South Asia - Middle East & North Africa == 0                    0.622
## Sub-Saharan Africa - Middle East & North Africa == 0           <0.001 ***
## South Asia - North America == 0                                 0.487
## Sub-Saharan Africa - North America == 0                        <0.001 ***
## Sub-Saharan Africa - South Asia == 0                           <0.001 ***
## Low income - High income == 0                                  <0.001 ***
## Lower middle income - High income == 0                         <0.001 ***
## Upper middle income - High income == 0                         <0.001 ***
## Lower middle income - Low income == 0                          <0.001 ***
## Upper middle income - Low income == 0                          <0.001 ***
## Upper middle income - Lower middle income == 0                 <0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

The differences among the income groups is significant. On the contrary, the difference among the regions is only significant for few cases. Plot of means (Figure 5).

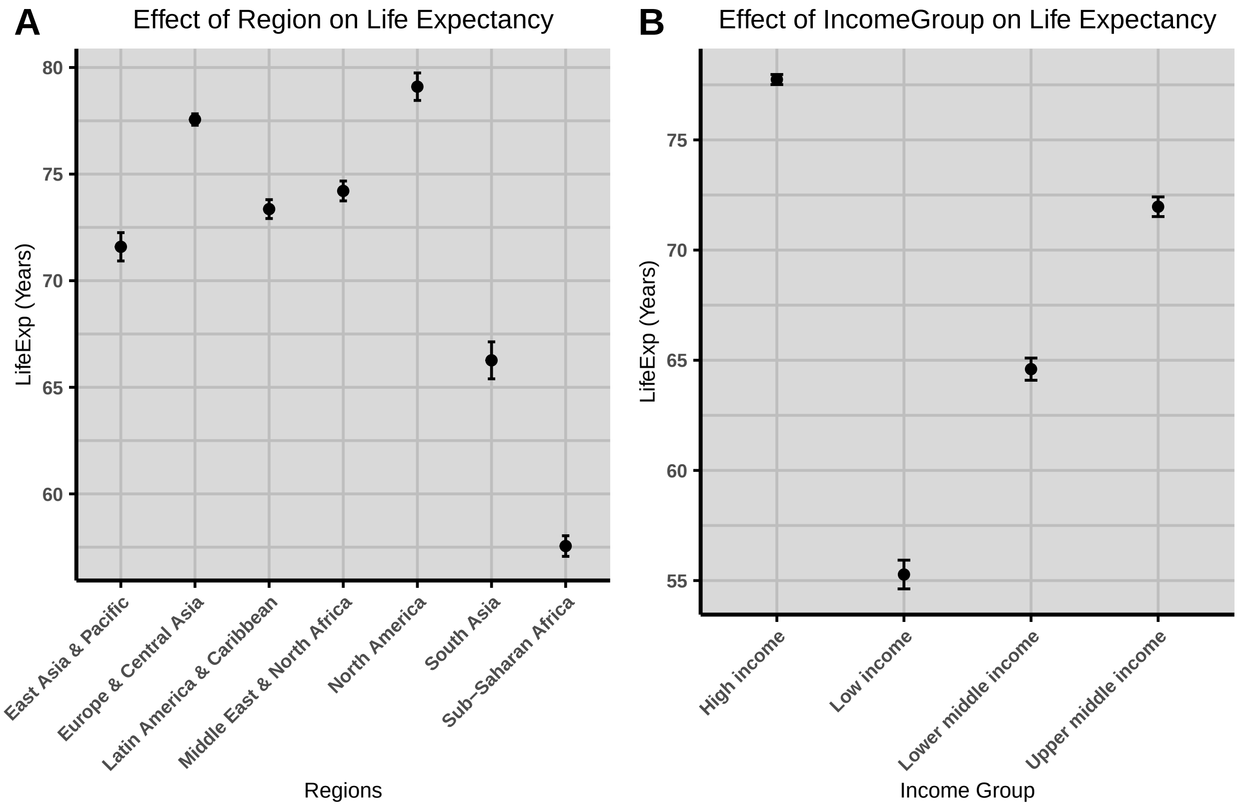


Figure 5: Plot of means, effect of Region and IncomeGroup on Life Expectancy

Post-hoc test to evaluate pairwise differences among the different `Regions` with holms-correction.

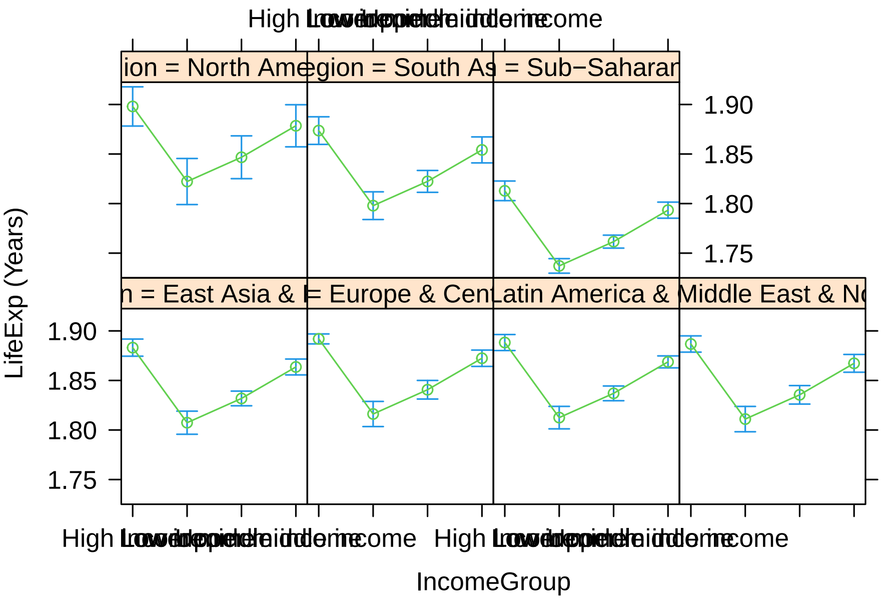
```
## F Test:
## P-value adjustment method: holm
##                                                          Value  Df  Sum of Sq
##            East Asia & Pacific-Europe & Central Asia -0.008798   1    0.00406
##        East Asia & Pacific-Latin America & Caribbean -0.005142   1    0.00164
##       East Asia & Pacific-Middle East & North Africa -0.003660   1    0.00061
##                    East Asia & Pacific-North America -0.014908   1    0.00224
##                       East Asia & Pacific-South Asia  0.009450   1    0.00240
##               East Asia & Pacific-Sub-Saharan Africa  0.070242   1    0.27299
##      Europe & Central Asia-Latin America & Caribbean  0.003657   1    0.00081
##     Europe & Central Asia-Middle East & North Africa  0.005138   1    0.00145
##                  Europe & Central Asia-North America -0.006110   1    0.00042
##                     Europe & Central Asia-South Asia  0.018248   1    0.00754
##             Europe & Central Asia-Sub-Saharan Africa  0.079040   1    0.26746
## Latin America & Caribbean-Middle East & North Africa  0.001481   1    0.00011
##              Latin America & Caribbean-North America -0.009767   1    0.00098
##                 Latin America & Caribbean-South Asia  0.014591   1    0.00573
##         Latin America & Caribbean-Sub-Saharan Africa  0.075384   1    0.34544
##             Middle East & North Africa-North America -0.011248   1    0.00130
##                Middle East & North Africa-South Asia  0.013110   1    0.00391
##        Middle East & North Africa-Sub-Saharan Africa  0.073902   1   0.22817
##                             North America-South Asia  0.024358   1   0.00476
##                     North America-Sub-Saharan Africa  0.085150   1   0.06964
##                        South Asia-Sub-Saharan Africa  0.060792   1   0.11146
## Residuals                                                      797   0.97268
##                                                              F    Pr(>F)
##            East Asia & Pacific-Europe & Central Asia   3.3241    0.8237
##        East Asia & Pacific-Latin America & Caribbean   1.3475    1.0000
##       East Asia & Pacific-Middle East & North Africa   0.5016    1.0000
##                    East Asia & Pacific-North America   1.8361    1.0000
##                       East Asia & Pacific-South Asia   1.9643    1.0000
##               East Asia & Pacific-Sub-Saharan Africa 223.6844 < 2.2e-16 ***
##      Europe & Central Asia-Latin America & Caribbean   0.6621    1.0000
##     Europe & Central Asia-Middle East & North Africa   1.1847    1.0000
##                  Europe & Central Asia-North America   0.3449    1.0000
##                     Europe & Central Asia-South Asia   6.1793    0.1969
##             Europe & Central Asia-Sub-Saharan Africa 219.1552 < 2.2e-16 ***
## Latin America & Caribbean-Middle East & North Africa   0.0920    1.0000
##              Latin America & Caribbean-North America   0.8040    1.0000
##                 Latin America & Caribbean-South Asia   4.6938    0.4280
##         Latin America & Caribbean-Sub-Saharan Africa 283.0522 < 2.2e-16 ***
##             Middle East & North Africa-North America   1.0617    1.0000
##                Middle East & North Africa-South Asia   3.2021    0.8237
##        Middle East & North Africa-Sub-Saharan Africa 186.9629 < 2.2e-16 ***
##                             North America-South Asia   3.9013    0.6317
##                     North America-Sub-Saharan Africa  57.0607 1.855e-12 ***
##                        South Asia-Sub-Saharan Africa  91.3312 < 2.2e-16 ***
## Residuals
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Post-hoc test to evaluate pairwise differences among the different `IncomeGroup` with holms-correction.

```
## F Test:
## P-value adjustment method: holm
##                                             Value  Df Sum of Sq       F
##                   High income-Low income  0.075757   1   0.18139 148.626
##          High income-Lower middle income  0.051274   1   0.16600 136.020
##          High income-Upper middle income  0.019508   1   0.02998  24.562
##           Low income-Lower middle income -0.024483   1   0.03030  24.828
##           Low income-Upper middle income -0.056250   1   0.12696 104.029
## Lower middle income-Upper middle income -0.031766   1   0.08871  72.688
## Residuals                                         797   0.97268
##                                            Pr(>F)
##                   High income-Low income < 2.2e-16 ***
##          High income-Lower middle income < 2.2e-16 ***
##          High income-Upper middle income 1.538e-06 ***
##           Low income-Lower middle income 1.538e-06 ***
##           Low income-Upper middle income < 2.2e-16 ***
## Lower middle income-Upper middle income 2.271e-16 ***
## Residuals
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Effect plot of effect of Region and IncomeGroup on Life Expectancy (Figure 6).

Fig. 6: Effect plot of Region and IncomeGroup on Life Expectancy



---

4. At least one of the questions must deal with count data, e.g using chi-square tests.

If there is an association between distribution of countries by `Region` and `IncomeGroup`?

Following is the distribution of the `Regions` as per the `IncomeGroup` (Table 3).

Table 3: Region and IncomeGroup

Region	IncomeGroup	n
East Asia & Pacific	High income	28
Europe & Central Asia	High income	168
Latin America & Caribbean	High income	28
Middle East & North Africa	High income	45
North America	High income	12
South Asia	High income	0
Sub-Saharan Africa	High income	0
East Asia & Pacific	Low income	0
Europe & Central Asia	Low income	0
Latin America & Caribbean	Low income	0
Middle East & North Africa	Low income	0
North America	Low income	0
South Asia	Low income	4
Sub-Saharan Africa	Low income	84
East Asia & Pacific	Lower middle income	55

Europe & Central Asia	Lower middle income	0
Latin America & Caribbean	Lower middle income	42
Middle East & North Africa	Lower middle income	21
North America	Lower middle income	0
South Asia	Lower middle income	35
Sub-Saharan Africa	Lower middle income	84
East Asia & Pacific	Upper middle income	28
Europe & Central Asia	Upper middle income	28
Latin America & Caribbean	Upper middle income	91
Middle East & North Africa	Upper middle income	19
North America	Upper middle income	0
South Asia	Upper middle income	0
Sub-Saharan Africa	Upper middle income	35

Contingency Table From the data provided a contingency table was prepared with the sums at the margins (Table 4). The table shows that the smallest count is less then 5 (i.e., 0), Fisher's exact test was used for further analysis.

Table 4: Contingency Table

	High income	Low income	Lower middle income	Upper middle income	Sum
East Asia & Pacific	28	0	55	28	111
Europe & Central Asia	168	0	0	28	196
Latin America & Caribbean	28	0	42	91	161
Middle East & North Africa	45	0	21	19	85
North America	12	0	0	0	12
South Asia	0	4	35	0	39
Sub-Saharan Africa	0	84	84	35	203
Sum	281	88	237	201	807

Fisher's exact test Since the smallest count (expected frequencies) are less then 5, i.e. 0 (see Table 4), Fisher's exact test was performed instead of Chi-square test (χ^2). Here, since the values were too small, `workspace` was increased, but that also gave error with a suggestion to use 'simulate.p.value=TRUE', which was incorporated. `B` value, corresponding to number of replicates used in Monte Carlo was increased to 200000.

```
##
##  Fisher’s Exact Test for Count Data with simulated p-value (based on
##  2e+05 replicates)
##
## data:  country_by_region_table
## p-value = 5e-06
## alternative hypothesis: two.sided
```

Statistics Report:

- Type of test: Fisher's Exact Test for Count Data with simulated p-value (based on 2e+05 replicates)
- p-value: 4.999975×10^{-6}

Data visualization Data was visualized using ggplot (Figure 7).

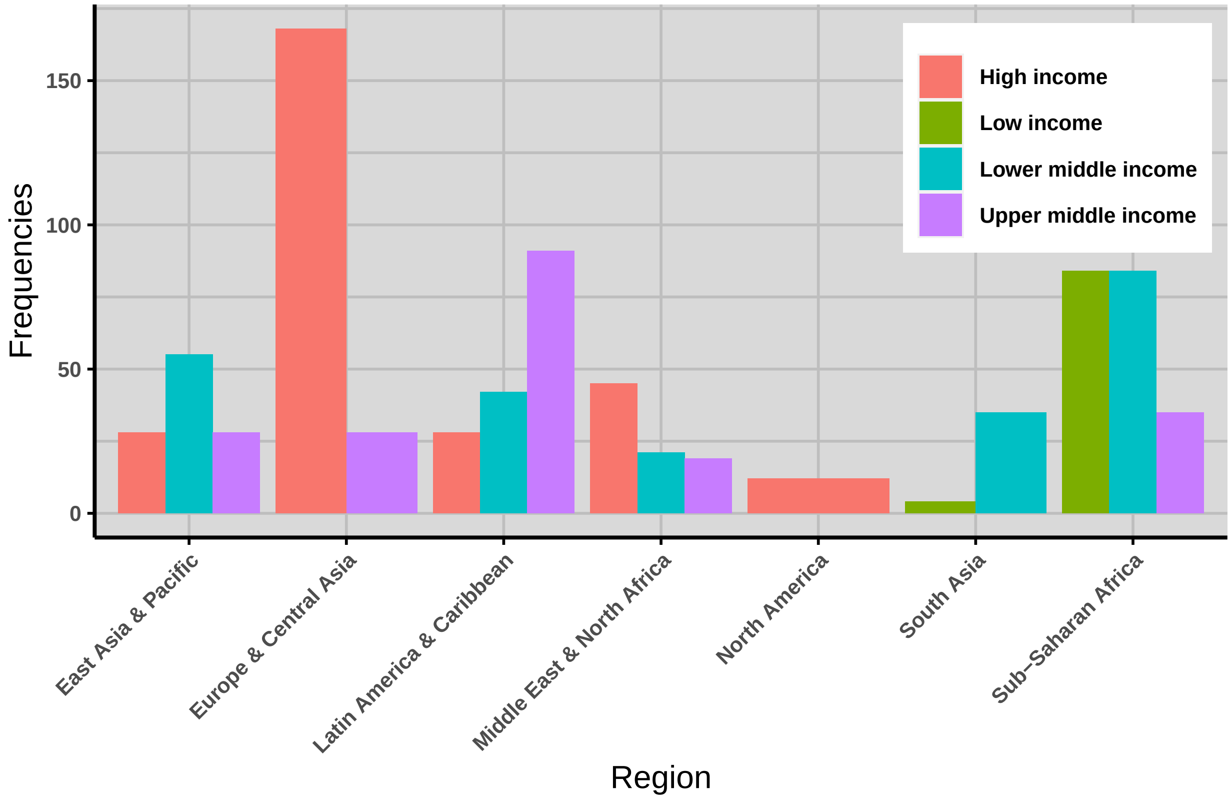


Figure 7: Distribution of regions as per the IncomeGroup type

Fisher's exact test with p-value (4.999975×10^{-6}) hence $p<0.05$, therefore, we reject the H0 (null hypothesis) and we conclude that there is significant association between the distribution of the regions as per their income groups.

References

1. Gapminder: https://www.gapminder.org/data/
2. Population: http://gapm.io/dpop
3. HDI: https://hdr.undp.org/en
4. Income: http://gapm.io/dgdppc
5. LifeExp: http://gapm.io/ilex
6. SDI: http://gapm.io/dsdi
7. GDPPC: https://data.worldbank.org/indicator/NY.GDP.PCAP.KD
8. Metadata: https://data.worldbank.org/indicator/NY.GNP.MKTP.PP.CD
9. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M. Welcome to the Tidyverse. Journal of open source software. 2019 Nov 21;4(43):1686.

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(gt)
library(pracma) # for mode
library(ggpubr) # for ggarrange
library(memisc) # mtable
library(multcomp) # for Tukey post-hoc test
library(Rmisc) # for summarySE
library(phia) # for testInteractions
library(effects) # for effect
library(knitr) # for kable
library(broom) # for tidy glht
library(jtools) # for summ
# Reading files

# Population
pop <- read.csv("data/population_total.csv")
pop_tidy <- pop %>%
  pivot_longer("X1800":"X2100", names_to = "year", values_to = "population")
pop_tidy$year <- sub("X", "", pop_tidy$year)
pop_tidy$population <- as.numeric(str_replace_all(pop_tidy$population,
                                                  regex(c("M" = "e6", "K" = "e3",
                                                          "B" = "e9"),
                                                        ignore_case = TRUE)))
names(pop_tidy)[1] <- "country"
# Selected data from 1990 to 2019 with interval of ~5 years
pop_tidy_90_19 <- pop_tidy %>%
  filter(year %in% c(1990, 1995, 2000, 2005, 2010, 2015, 2019))

# HDI
hdi <- read.csv("data/hdi_human_development_index.csv")
hdi_noNA <- na.omit(hdi)
hdi_noNA_tidy <- hdi_noNA %>%
  pivot_longer("X1990":"X2019", names_to = "year", values_to = "HDI")
hdi_noNA_tidy$year <- sub("X", "", hdi_noNA_tidy$year)
names(hdi_noNA_tidy)[1] <- "country"
hdi_noNA_tidy_90_19 <- hdi_noNA_tidy %>%
  filter(year %in% c(1990, 1995, 2000, 2005, 2010, 2015, 2019))

# Income
income <- read.csv("data/income_per_person_gdppercapita_ppp_inflation_adjusted.csv")
income_noNA <- na.omit(income)
income_noNA_tidy <- income_noNA %>%
  pivot_longer("X1800" : "X2050", names_to = "year", values_to = "Income",
               values_transform = list(Income = as.character))
income_noNA_tidy$year <- sub("X", "", income_noNA_tidy$year)
income_noNA_tidy$Income <- as.numeric(str_replace_all(income_noNA_tidy$Income,
                                                      regex(c("M" = "e6", "K" = "e3",
                                                              "B" = "e9"),
                                                            ignore_case = TRUE)))
names(income_noNA_tidy)[1] <- "country"
```

```
income_noNA_tidy_90_19 <- income_noNA_tidy %>%
  filter(year %in% c(1990, 1995, 2000, 2005, 2010, 2015, 2019))

# Life Expectancy
lifeExp <- read.csv("data/life_expectancy_years.csv")
lifeExp_noNA <- na.omit(lifeExp)
lifeExp_noNA_tidy <- lifeExp_noNA %>%
  pivot_longer("X1800" : "X2100", names_to = "year", values_to = "LifeExp")
lifeExp_noNA_tidy$year <- sub("X", "", lifeExp_noNA_tidy$year)
names(lifeExp_noNA_tidy)[1] <- "country"
lifeExp_noNA_tidy_90_19 <- lifeExp_noNA_tidy %>%
  filter(year %in% c(1990, 1995, 2000, 2005, 2010, 2015, 2019))

# SDI
sdi <- read.csv("data/sdi.csv")
sdi_noNA <- na.omit(sdi)
sdi_noNA_tidy <- sdi_noNA %>%
  pivot_longer("X1990" : "X2019", names_to = "year", values_to = "SDI")
sdi_noNA_tidy$year <- sub("X", "", sdi_noNA_tidy$year)
names(sdi_noNA_tidy)[1] <- "country"
sdi_noNA_tidy_90_19 <- sdi_noNA_tidy %>%
  filter(year %in% c(1990, 1995, 2000, 2005, 2010, 2015, 2019))

# GDP per capita
gdppc <- read.csv("data/gdppercapita_us_inflation_adjusted.csv",
                  na.strings=c(NA,"NA"," NA"))
gdppc_noNA <- na.omit(gdppc)
gdppc_noNA_tidy <- gdppc_noNA %>%
  pivot_longer("X1960" : "X2020", names_to = "year", values_to = "GDPPC")
gdppc_noNA_tidy$year <- sub("X", "", gdppc_noNA_tidy$year)
gdppc_noNA_tidy$GDPPC <- as.numeric(str_replace_all(gdppc_noNA_tidy$GDPPC,
                                                    regex(c("M" = "e6", "K" = "e3",
                                                            "B" = "e9"),
                                                          ignore_case = TRUE)))
gdppc_noNA_tidy_90_19 <- gdppc_noNA_tidy %>%
  filter(year %in% c(1990, 1995, 2000, 2005, 2010, 2015, 2019))
names(gdppc_noNA_tidy_90_19)[1] <- "country"
gdppc_noNA_tidy_90_19 <- na.omit(gdppc_noNA_tidy_90_19)

# Metadata
metadata <-  read.csv("data/Metadata_Country_API_NY.GNP.MKTP.PP.CD_DS2_en_csv_v2_3479540.csv")
metadata_fewCols <- metadata[c(5, 2, 3)]
metadata_fewCols_noNA <- na.omit(metadata_fewCols)
names(metadata_fewCols_noNA)[1] <- "country"

# Combining all data
adding_data <- list(pop_tidy_90_19, hdi_noNA_tidy_90_19, income_noNA_tidy_90_19,
                    lifeExp_noNA_tidy_90_19, sdi_noNA_tidy_90_19,
                    gdppc_noNA_tidy_90_19) %>%
    reduce(left_join, by = c("country", "year"))
all_datatidy_90_19 <- list(adding_data, metadata_fewCols_noNA) %>%
  reduce(left_join, by = "country")
names(all_datatidy_90_19) <- c("Country", "Year", "Population", "HDI",
                               "Income (USD)", "LifeExp (Years)", "SDI",
                               "GDPPC (USD)", "Region", "IncomeGroup")
# which(is.na(all_datatidy_90_19_noNA))
all_datatidy_90_19_noNA <- na.omit(all_datatidy_90_19)

all_datatidy_90_19_noNA$Region <- as.factor(all_datatidy_90_19_noNA$Region)
all_datatidy_90_19_noNA$IncomeGroup <- as.factor(all_datatidy_90_19_noNA$IncomeGroup)


# Measures of central tendency (Mean, Median, Mode, SD, StdErr, and Var)

# Population
data_table_pop <- all_datatidy_90_19_noNA %>%
  group_by(Region) %>%
  summarise("Mean" = mean(Population),
            "Median" = median(Population),
            "Mode" = Mode(Population),
            "SD" = sd(Population),
            "StdErr" = sd(Population)/sqrt(length(Population)),
            "Var" = var(Population)) %>%
  # mutate_if(is.numeric, ~round(., 3)) %>%
  gt() %>%
  tab_header(title = paste("Table 1: Distribution of population across various regions",
                           sep = "\n"), subtitle = "(From 1990 to 2019)")
# HDI
data_table_hdi <- all_datatidy_90_19_noNA %>%
  group_by(Region) %>%
  summarise("Mean" = mean(HDI),
            "Median" = median(HDI),
            "Mode" = Mode(HDI),
            "SD" = sd(HDI),
            "StdErr" = sd(HDI)/sqrt(length(HDI)),
            "Var" = var(HDI)) %>%
  mutate_if(is.numeric, ~round(., 3)) %>%
  gt() %>%
  tab_header(title = paste("Table 2: Distribution of Human Development Index (HDI) across various regio
                           sep = "\n"), subtitle = "(From 1990 to 2019)")
# Income
data_table_Income <- all_datatidy_90_19_noNA %>%
  group_by(Region) %>%
  summarise("Mean" = mean(`Income (USD)`),
            "Median" = median(`Income (USD)`),
            "Mode" = Mode(`Income (USD)`),
            "SD" = sd(`Income (USD)`),
            "StdErr" = sd(`Income (USD)`)/sqrt(length(`Income (USD)`)),
            "Var" = var(`Income (USD)`)) %>%
  mutate_if(is.numeric, ~round(., 3)) %>%
  gt() %>%
  tab_header(title = paste("Table 3: Distribution of Income (USD) across various regions",
                           sep = "\n"), subtitle = "(From 1990 to 2019)")
# Life Expectancy
data_table_LifeExp <- all_datatidy_90_19_noNA %>%
  group_by(Region) %>%
  summarise("Mean" = mean(`LifeExp (Years)`),
            "Median" = median(`LifeExp (Years)`),
            "Mode" = Mode(`LifeExp (Years)`),
            "SD" = sd(`LifeExp (Years)`),
            "StdErr" = sd(`LifeExp (Years)`)/sqrt(length(`LifeExp (Years)`)),
            "Var" = var(`LifeExp (Years)`)) %>%
  mutate_if(is.numeric, ~round(., 3)) %>%
  gt() %>%
  tab_header(title = paste("Table 4: Distribution of Life Expectancy (Years) across various regions",
                           sep = "\n"), subtitle = "(From 1990 to 2019)")
# SDI
data_table_sdi <- all_datatidy_90_19_noNA %>%
  group_by(Region) %>%
  summarise("Mean" = mean(SDI),
            "Median" = median(SDI),
            "Mode" = Mode(SDI),
            "SD" = sd(SDI),
            "StdErr" = sd(SDI)/sqrt(length(SDI)),
            "Var" = var(SDI)) %>%
  mutate_if(is.numeric, ~round(., 3)) %>%
  gt() %>%
  tab_header(title = paste("Table 5: Distribution of The Sustainable Development Index (SDI) across vari
                           sep = "\n"), subtitle = "(From 1990 to 2019)")
# GDP per capita
data_table_gdppc <- all_datatidy_90_19_noNA %>%
  group_by(Region) %>%
  summarise("Mean" = mean(`GDPPC (USD)`),
            "Median" = median(`GDPPC (USD)`),
            "Mode" = Mode(`GDPPC (USD)`),
            "SD" = sd(`GDPPC (USD)`),
            "StdErr" = sd(`GDPPC (USD)`)/sqrt(length(`GDPPC (USD)`)),
            "Var" = var(`GDPPC (USD)`)) %>%
  mutate_if(is.numeric, ~round(., 3)) %>%
  gt() %>%
  tab_header(title = paste("Table 6: Distribution of GDP per capita (USD) across various regions",
                           sep = "\n"), subtitle = "(From 1990 to 2019)")
centraltendencyTable = gt(rbind(data_table_pop$`_data`, data_table_hdi$`_data`,
                                data_table_Income$`_data`, data_table_LifeExp$`_data`,
                                data_table_sdi$`_data`, data_table_gdppc$`_data`)) %>%
  tab_row_group(label = "Population", rows = 1) %>%
  tab_row_group(label = "HDI", rows = 2) %>%
  tab_row_group(label = "Income", rows = 3) %>%
  tab_row_group(label = "LifeExp", rows = 4) %>%
  tab_row_group(label = "SDI", rows = 5) %>%
  tab_row_group(label = "GDPPC", rows = 6) %>%
  row_group_order(groups = c("Population", "HDI", "Income", "LifeExp", "SDI", "GDPPC")) %>%
  tab_header(title = "Table 1: Measures of central tendency across various parameters",
             subtitle = "(From 1990 to 2019)")
centraltendencyTable

all_hist_Pop <- all_datatidy_90_19_noNA %>%
  group_by(Country) %>%
  ggplot(., aes(Population)) +
  geom_histogram(bins = 30) +
  labs(title="Population across countries", x="Population", y="Frequency") +
  theme(axis.line = element_line(size = 0.7, color = "black"),
        text = element_text(size = 12),
        plot.title = element_text(size= 12),
        axis.text.x = element_text(angle = 45, hjust = 1))

all_hist_hdi <- all_datatidy_90_19_noNA %>%
  group_by(Country) %>%
  ggplot(., aes(HDI)) +
  geom_histogram(bins = 30)  +
  labs(title="HDI across countries", x="HDI", y="Frequency") +
  theme(axis.line = element_line(size = 0.7, color = "black"),
        text = element_text(size = 12),
        plot.title = element_text(size= 12),
        axis.text.x = element_text(angle = 45, hjust = 1))

all_hist_Income <- all_datatidy_90_19_noNA %>%
  group_by(Country) %>%
  ggplot(., aes(`Income (USD)`)) +
  geom_histogram(bins = 30) +
  labs(title="Income (USD) across countries", x="Income (USD)", y="Frequency") +
  theme(axis.line = element_line(size = 0.7, color = "black"),
        text = element_text(size = 12),
        plot.title = element_text(size= 12),
        axis.text.x = element_text(angle = 45, hjust = 1))

all_hist_sdi <- all_datatidy_90_19_noNA %>%
  group_by(Country) %>%
  ggplot(., aes(SDI)) +
  geom_histogram(bins = 30) +
  labs(title="SDI across countries", x="SDI", y="Frequency") +
  theme(axis.line = element_line(size = 0.7, color = "black"),
        text = element_text(size = 12),
        plot.title = element_text(size= 12),
        axis.text.x = element_text(angle = 45, hjust = 1))

all_hist_GDPPC <- all_datatidy_90_19_noNA %>%
  group_by(Country) %>%
  ggplot(., aes(`GDPPC (USD)`)) +
  geom_histogram(bins = 30) +
  labs(title="GDPPC (USD) across countries", x="GDP Per Capita (USD)", y="Frequency") +
  theme(axis.line = element_line(size = 0.7, color = "black"),
        text = element_text(size = 12),
        plot.title = element_text(size= 12),
        axis.text.x = element_text(angle = 45, hjust = 1))

all_hist_Region <- all_datatidy_90_19_noNA %>%
  group_by(Country) %>%
  ggplot(., aes(Region)) +
  geom_histogram(stat = "count") +
  labs(title="Countries across region", x="Region", y="Frequency") +
  theme(axis.line = element_line(size = 0.7, color = "black"),
        text = element_text(size = 12),
        plot.title = element_text(size= 12),
        axis.text.x = element_text(angle = 45, hjust = 1))

all_hist_IncomeGroup <- all_datatidy_90_19_noNA %>%
  group_by(Country) %>%
  ggplot(., aes(IncomeGroup)) +
  geom_histogram(stat = "count") +
  labs(title="Countries across Income groups", x="Income Group", y="Frequency") +
  theme(axis.line = element_line(size = 0.7, color = "black"),
        text = element_text(size = 12),
        plot.title = element_text(size= 12),
        axis.text.x = element_text(angle = 45, hjust = 1))

ggarrange(all_hist_Pop, all_hist_hdi, nrow = 1, ncol = 2,
          labels = c("A", "B"))

ggarrange(all_hist_Income, all_hist_sdi, nrow = 1, ncol = 2,
          labels = c("C", "D"))

ggarrange(all_hist_GDPPC, all_hist_Region, nrow = 1, ncol = 2,
          labels = c("E", "F"))

ggarrange(all_hist_IncomeGroup, nrow = 1, ncol = 1,
          labels = c("G")) %>%
          annotate_figure(bottom = text_grob("Figure 1: Distribution of data across various parameters"
                                             size = 16))
Reg.Model_HDI_Global_normal <- lm(HDI ~ as.integer(Year), data=all_datatidy_90_19_noNA)
summary(Reg.Model_HDI_Global_normal)
par(mfrow = c(2, 2))
plot(Reg.Model_HDI_Global_normal)
par(mfrow = c(1, 1))
Reg.Model_HDI_Global_normal_ggplot <- ggplot(all_datatidy_90_19_noNA,
                                             aes(as.integer(Year), HDI)) +
  geom_point() +
  geom_smooth(method = "lm", colour="blue") +
  labs(title="Increase in HDI globally", x="Year", y="HDI") +
  scale_y_continuous(breaks = seq(0, 1, 0.1)) +
  scale_x_continuous(breaks = seq(0, 2050, 5), limits = c(1989, 2020)) +
  theme(panel.grid.major = element_line(size = 0.5, color = "grey"),
        panel.grid.minor = element_line(size = 0.5, color = "grey"),
        plot.title = element_text(hjust = 0.5, size = 12),
        axis.line = element_line(size = 0.7, color = "black"),
        text = element_text(size = 12),
        axis.text = element_text(angle=0, vjust=0.5, size=12, face = "bold"),
        axis.text.x = element_text(angle = 45, hjust = 1),
        panel.background = element_rect(fill = "grey85",
                                        size = 0.5, linetype = "solid"),
        axis.ticks = element_line(size = 0.5, colour = "black"))

Reg.Model_HDI_Global_normal_ggplot
Reg.Model_income_Global_normal <- lm(`Income (USD)` ~ as.integer(Year),
                                     data=all_datatidy_90_19_noNA)
```

```r
summary(Reg.Model_income_Global_normal)
par(mfrow = c(2, 2))
plot(Reg.Model_income_Global_normal)
par(mfrow = c(1, 1))
Reg.Model_income_Global_normal_ggplot <- ggplot(all_datatidy_90_19_noNA,
                                                aes(as.integer(Year),
                                                    log10(`Income (USD)`))) +
  geom_point() +
  geom_smooth(method = "lm", colour="blue") +
  labs(title="Increase in Income globally", x="Year",
       y="Income (USD) (log10 scale)") +
  scale_y_continuous(breaks = seq(0, 100, 0.5)) +
  scale_x_continuous(breaks = seq(0, 2050, 5), limits = c(1989, 2020)) +
  theme(panel.grid.major = element_line(size = 0.5, color = "grey"),
        panel.grid.minor = element_line(size = 0.5, color = "grey"),
        plot.title = element_text(hjust = 0.5, size = 12),
        axis.line = element_line(size = 0.7, color = "black"),
        text = element_text(size = 12),
        axis.text = element_text(angle=0, vjust=0.5, size=12, face = "bold"),
        axis.text.x = element_text(angle = 45, hjust = 1),
        panel.background = element_rect(fill = "grey85",
                                        size = 0.5, linetype = "solid"),
        axis.ticks = element_line(size = 0.5, colour = "black"))

Reg.Model_income_Global_normal_ggplot
Reg.Model_LifeExp_Global_normal <- lm(`LifeExp (Years)` ~ as.integer(Year), data=all_datatidy_90_19_noNA
summary(Reg.Model_LifeExp_Global_normal)
par(mfrow = c(2, 2))
plot(Reg.Model_LifeExp_Global_normal)
par(mfrow = c(1, 1))
Reg.Model_LifeExp_Global_normal_ggplot <- ggplot(all_datatidy_90_19_noNA, aes(as.integer(Year), `LifeExp
  geom_point() +
  geom_smooth(method = "lm", colour="blue") +
  labs(title="Increase in LifeExp globally", x="Year", y="LifeExp (Years)") +
  scale_y_continuous(breaks = seq(0, 100, 5)) +
  scale_x_continuous(breaks = seq(0, 2050, 5), limits = c(1989, 2020)) +
  theme(panel.grid.major = element_line(size = 0.5, color = "grey"),
        panel.grid.minor = element_line(size = 0.5, color = "grey"),
        plot.title = element_text(hjust = 0.5, size = 12),
        axis.line = element_line(size = 0.7, color = "black"),
        text = element_text(size = 12),
        axis.text = element_text(angle=0, vjust=0.5, size=12, face = "bold"),
        axis.text.x = element_text(angle = 45, hjust = 1),
        panel.background = element_rect(fill = "grey85",
                                        size = 0.5, linetype = "solid"),
        axis.ticks = element_line(size = 0.5, colour = "black"))

Reg.Model_LifeExp_Global_normal_ggplot
Reg.Model_gdppc_Global_normal <- lm(`GDPPC (USD)` ~ as.integer(Year),
                                     data=all_datatidy_90_19_noNA)
summary(Reg.Model_gdppc_Global_normal)
par(mfrow = c(2, 2))
plot(Reg.Model_gdppc_Global_normal)
```

```
par(mfrow = c(1, 1))
Reg.Model_gdppc_Global_normal_ggplot <- ggplot(all_datatidy_90_19_noNA,
                                                aes(as.integer(Year),
                                                    `GDPPC (USD)`)) +
  geom_point() +
  geom_smooth(method = "lm", colour="blue") +
  labs(title="Increase in GDPPC (USD) globally", x="Year",
       y="GDPPC (USD)") +
  scale_y_continuous(breaks = seq(0, 1000000, 10000)) +
  scale_x_continuous(breaks = seq(0, 2050, 5), limits = c(1989, 2020)) +
  theme(panel.grid.major = element_line(size = 0.5, color = "grey"),
        panel.grid.minor = element_line(size = 0.5, color = "grey"),
        plot.title = element_text(hjust = 0.5, size = 12),
        axis.line = element_line(size = 0.7, color = "black"),
        text = element_text(size = 12),
        axis.text = element_text(angle=0, vjust=0.5, size=12, face = "bold"),
        axis.text.x = element_text(angle = 45, hjust = 1),
        panel.background = element_rect(fill = "grey85",
                                        size = 0.5, linetype = "solid"),
        axis.ticks = element_line(size = 0.5, colour = "black"))

Reg.Model_gdppc_Global_normal_ggplot
ggarrange(Reg.Model_HDI_Global_normal_ggplot, Reg.Model_income_Global_normal_ggplot,
          nrow = 1, ncol = 2, labels = c("A", "B"))
ggarrange(Reg.Model_LifeExp_Global_normal_ggplot, Reg.Model_gdppc_Global_normal_ggplot,
          nrow = 1, ncol = 2, labels = c("C", "D")) %>%
          annotate_figure(bottom = text_grob("Figure 2: Linear Regression analysis",
                                             size = 14))
shapiro.test(subset(all_datatidy_90_19_noNA,
                    IncomeGroup %in% c("Low income",
                                       "High income"))$`LifeExp (Years)`)
income_group_hist <- ggplot(subset(all_datatidy_90_19_noNA,
                                   IncomeGroup %in% c("Low income", "High income")),
                            aes(`LifeExp (Years)`)) +
  geom_histogram(bins = 10) +
  labs(title="Life expectancy among low and high income countries",
       x="LifeExp (Years)", y="Count") +
    # scale_y_continuous(breaks = seq(0, 200, 30, limits = c(0, 200))) +
  # scale_x_continuous(breaks = seq(0, 100, 10), limits = c(30, 100)) +
  theme(panel.grid.major = element_line(size = 0.5, color = "grey"),
        panel.grid.minor = element_line(size = 0.5, color = "grey"),
        plot.title = element_text(hjust = 0.5, size = 7),
        axis.line = element_line(size = 0.7, color = "black"),
        text = element_text(size = 12),
        axis.text = element_text(angle=0, vjust=0.5, size=12, face = "bold"),
        axis.text.x = element_text(angle = 45, hjust = 1),
        panel.background = element_rect(fill = "grey85",
                                        size = 0.5, linetype = "solid"),
        axis.ticks = element_line(size = 0.5, colour = "black"),
        plot.caption = element_text(hjust = 0.5, size = 12))
wilcox_LifeExp_IncomeGroup <- wilcox.test(subset(all_datatidy_90_19_noNA,
                                                 IncomeGroup %in% c("Low income",
                                                                    "High income"))$`LifeExp (Years)` ~
```

```
                                            subset(all_datatidy_90_19_noNA,
                                                   IncomeGroup %in% c("Low income",
                                                                      "High income"))$IncomeGroup)
income_group_boxplot <- ggplot(subset(all_datatidy_90_19_noNA,
                                     IncomeGroup %in% c("Low income", "High income"))) +
  aes(x = IncomeGroup, y = `LifeExp (Years)`) +
  geom_boxplot(fill = "grey") +
  labs(title="Boxplot of difference between low and high income countries",
       x="Income Group",
       y="LifeExp (Years)") +
   theme(panel.grid.major = element_line(size = 0.5, color = "grey"),
        panel.grid.minor = element_line(size = 0.5, color = "grey"),
        plot.title = element_text(hjust = 0.5, size = 7),
        axis.line = element_line(size = 0.7, color = "black"),
        text = element_text(size = 12),
        axis.text = element_text(angle=0, vjust=0.5, size=12, face = "bold"),
        axis.text.x = element_text(angle = 45, hjust = 1),
        panel.background = element_rect(fill = "grey85",
                                        size = 0.5, linetype = "solid"),
        axis.ticks = element_line(size = 0.5, colour = "black"),
        plot.caption = element_text(hjust = 0.5, size = 12))
ggarrange(income_group_hist, income_group_boxplot,
          nrow = 1, ncol = 2, labels = c("A", "B")) %>%
          annotate_figure(bottom = text_grob("Figure 3: Histogram and boxplot of difference between low
                                             size = 12))
all_datatidy_90_19_noNA %>%
  group_by(Region, IncomeGroup) %>%
  get_summary_stats(`LifeExp (Years)`, type = "mean_sd") %>%
  gt() %>%
    tab_header(title = "Table 2: Effect of Region and IncomeGroup on Life Expectancy",
    subtitle = "Measured central tendencies")
## Plot of means
region_lifeExp_boxplot <- ggplot(all_datatidy_90_19_noNA, aes(x = Region , y = `LifeExp (Years)`)) +
   geom_boxplot() +
   labs(title="Effect of Region on Life Expectancy",
        x ="Regions", y = "LifeExp (Years)") +
    theme(panel.grid.major = element_line(size = 0.5, color = "grey"),
        panel.grid.minor = element_line(size = 0.5, color = "grey"),
        plot.title = element_text(hjust = 0.5, size = 8),
        axis.line = element_line(size = 0.7, color = "black"),
        text = element_text(size = 8),
        axis.text = element_text(angle=0, vjust=0.5, size=7, face = "bold"),
        axis.text.x = element_text(angle = 45, hjust = 1),
        panel.background = element_rect(fill = "grey85",
                                        size = 0.5, linetype = "solid"),
        axis.ticks = element_line(size = 0.5, colour = "black"),
        plot.caption = element_text(hjust = 0.5, size = 8))

income_lifeExp_boxplot <- ggplot(all_datatidy_90_19_noNA, aes(x = IncomeGroup , y = `LifeExp (Years)`))
   geom_boxplot() +
   labs(title="Effect of IncomeGroup on Life Expectancy",
        x ="Regions", y = "LifeExp (Years)") +
    theme(panel.grid.major = element_line(size = 0.5, color = "grey"),
        panel.grid.minor = element_line(size = 0.5, color = "grey"),
        plot.title = element_text(hjust = 0.5, size = 8),
        axis.line = element_line(size = 0.7, color = "black"),
        text = element_text(size = 8),
        axis.text = element_text(angle=0, vjust=0.5, size=7, face = "bold"),
        axis.text.x = element_text(angle = 45, hjust = 1),
        panel.background = element_rect(fill = "grey85",
                                        size = 0.5, linetype = "solid"),
        axis.ticks = element_line(size = 0.5, colour = "black"),
        plot.caption = element_text(hjust = 0.5, size = 8))

ggarrange(region_lifeExp_boxplot, income_lifeExp_boxplot,
          nrow = 1, ncol = 2, labels = c("A", "B")) %>%
          annotate_figure(bottom = text_grob("Figure 4: Effect of Region and IncomeGroup on Life Expecta
                                             size = 12))
lm_lifeExp <- lm(`LifeExp (Years)` ~ Region * IncomeGroup,
                 all_datatidy_90_19_noNA, na.action=na.omit)
summary(lm_lifeExp)

par(mfrow = c(2, 2))
plot(lm_lifeExp)
par(mfrow = c(1, 1))
lm_lifeExp_log10 <- lm(log10(`LifeExp (Years)`) ~ Region + IncomeGroup,
                       data = all_datatidy_90_19_noNA, na.action=na.omit)
summary(lm_lifeExp_log10)

par(mfrow = c(2, 2))
plot(lm_lifeExp_log10)
par(mfrow = c(1, 1))
Anova(lm_lifeExp_log10)
R1 <- glht(lm_lifeExp_log10, mcp(Region = "Tukey"))$linfct
R2 <- glht(lm_lifeExp_log10, mcp(IncomeGroup = "Tukey"))$linfct

glht_sum <- summary(glht(lm_lifeExp_log10, linfct = rbind(R1, R2)))

glht_sum
sum_Region <- summarySE(all_datatidy_90_19_noNA, measurevar= "LifeExp (Years)",
                        groupvars="Region")
sum_Region_ggplot <- ggplot(sum_Region, aes(x= Region, y= `LifeExp (Years)`)) +
  geom_errorbar(aes(ymin= `LifeExp (Years)` - se, ymax=`LifeExp (Years)` + se),
                width=.1) +
  geom_point() +
   labs(title="Effect of Region on Life Expectancy",
        x ="Regions", y = "LifeExp (Years)") +
    theme(panel.grid.major = element_line(size = 0.5, color = "grey"),
        panel.grid.minor = element_line(size = 0.5, color = "grey"),
        plot.title = element_text(hjust = 0.5),
        axis.line = element_line(size = 0.7, color = "black"),
        text = element_text(size = 8),
        axis.text = element_text(angle=0, vjust=0.5, size=7, face = "bold"),
        axis.text.x = element_text(angle = 45, hjust = 1),
        panel.background = element_rect(fill = "grey85",
                                        size = 0.5, linetype = "solid"),
        axis.ticks = element_line(size = 0.5, colour = "black"))

sum_IncomeGroup <- summarySE(all_datatidy_90_19_noNA, measurevar= "LifeExp (Years)",
                        groupvars= "IncomeGroup") #summarySE

sum_IncomeGroup_ggplot <- ggplot(sum_IncomeGroup, aes(x= IncomeGroup, y= `LifeExp (Years)`)) +
  geom_errorbar(aes(ymin= `LifeExp (Years)` - se, ymax=`LifeExp (Years)` + se),
                width=.1) +
  geom_point() +
   labs(title="Effect of IncomeGroup on Life Expectancy",
        x ="Income Group", y = "LifeExp (Years)") +
    theme(panel.grid.major = element_line(size = 0.5, color = "grey"),
        panel.grid.minor = element_line(size = 0.5, color = "grey"),
        plot.title = element_text(hjust = 0.5),
        axis.line = element_line(size = 0.7, color = "black"),
        text = element_text(size = 8),
        axis.text = element_text(angle=0, vjust=0.5, size=7, face = "bold"),
        axis.text.x = element_text(angle = 45, hjust = 1),
        panel.background = element_rect(fill = "grey85",
                                        size = 0.5, linetype = "solid"),
        axis.ticks = element_line(size = 0.5, colour = "black"))

ggarrange(sum_Region_ggplot, sum_IncomeGroup_ggplot,
          nrow = 1, ncol = 2, labels = c("A", "B")) %>%
          annotate_figure(bottom = text_grob("Figure 5: Plot of means, effect of Region and IncomeGroup
                                             size = 12))
testInteractions(lm_lifeExp_log10, pairwise = "Region", adjustment="holm")
testInteractions(lm_lifeExp_log10, pairwise = "IncomeGroup", adjustment="holm")
plot(effect(term = "Region:IncomeGroup", mod = lm_lifeExp_log10, se=TRUE,
            x.var= "IncomeGroup"),
     ylab="LifeExp (Years)",
     xlab= "IncomeGroup",
     main="Fig. 6: Effect plot of Region and IncomeGroup on Life Expectancy",
     colors = c(3,4))

country_by_region <- all_datatidy_90_19_noNA[, 9:10]

as_tibble(table(country_by_region)) %>%
  gt() %>%
    tab_header(title = "Table 3: Region and IncomeGroup",
    subtitle = "")
addmargins(table(country_by_region)) %>%
  kable(caption = "Table 4: Contingency Table")
country_by_region_table = table(country_by_region)
country_by_region_table_fi = fisher.test(country_by_region_table,
                                         simulate.p.value=TRUE,
                                         B = 200000)
country_by_region_table_fi
ggplot(country_by_region, aes(x=Region, fill = IncomeGroup)) +
  geom_bar(position = "dodge") +
  labs(caption="Figure 7: Distribution of regions as per the IncomeGroup type",
       x="Region", y="Frequencies") +
  theme(panel.grid.major = element_line(size = 0.5, color = "grey"),
        legend.position = c(0.85, 0.75),
        # legend.title = element_text(colour="black", size=12, face="bold"),
        legend.title = element_blank(),
        legend.text = element_text(colour="black", size=8, face="bold"),
        panel.grid.minor = element_line(size = 0.5, color = "grey"),
        plot.title = element_text(hjust = 0.5),
        axis.line = element_line(size = 0.7, color = "black"),
        text = element_text(size = 12),
        axis.text = element_text(angle=0, vjust=0.5, size=8, face = "bold"),
        axis.text.x = element_text(angle = 45, hjust = 1),
        panel.background = element_rect(fill = "grey85",
                                        size = 0.5, linetype = "solid"),
        axis.ticks = element_line(size = 0.5, colour = "black"),
        plot.caption = element_text(hjust = 0.5, size = 12))
```