Data Project Part 1, 2 & 3: Group 2

Sharicka Zutshi (SID: 3034761627), David Zhong (SID:3034783363), Kayla Leiber (SID:3033188277), Edward Bian (SID: 3034608279), Patrick Udenyi (SID:3034018810)

7/14/2020

Part 1:

Question 1:

[2 marks] What is the problem your are addressing with these data? State the question you are trying to answer and let us know what type of question this is in terms of the PPDAC framework.

Answer: We are addressing the problem of how the health care expenditures of different countries as a percentage of GDP is spread across the world with our data. This is a descriptive problem because we are trying to analyze the spread of healthcare expenditures as an attribute in a population consisting of the world's countries.

Question 2:

[2 marks] What is the target population for your project? Why was this target chosen ie what was your rationale for wanting to answer this question in this specific population?

Answer: The target population is all the countries of the world because we are analyzing the reasoning behind their decision of allocating a certain percentage of their GDP they used for health care programs. We want to assess patterns, variations, or general trends accross the regions, and economies accross the world.

Question 3: [2 marks] What is the sampling frame used to collect the data you are using. Describe why you think this sampling strategy is appropriate for your question. To what group(s) would you feel comfortable generalizing the findings of your study and why.

Answer: The sampling frame we are using is the available data on government spending on healthcare, for all countries in the world. Our sampling strategy is appropriate to our question as available data can make patterns and spread of trends over a time period evident. Looking at the percentage spent of GDP allows us to make this observation unbiasedly. Since we are encompassing the population of the world in our data through individual countries, we would want to extend our findings to other factors affected by GDP allocation, or healthcare per capita spending, time periods from 2017 till date in the same population, and also to countries for whom such data is unavailable.

Question 4:

[2 marks] Write a brief description (1-4 sentences) of the source and contents of your dataset. Provide a URL to the original data source if applicable. If not (e.g., the data came from your internship), provide 1-2 sentences saying where the data came from. If you completed a web form to access the data and selected a subset, describe these steps (including any options you selected) and the date you accessed the data.

Answer:

We got our data on percent GDP spending by country from: https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS

The data consists of how much each country spent on healthcare (as a percent of total GDP), the income brackets and geographical regions that the World Bank classifies the countries into. The data was found through online searching. We acquired the data by clicking on the CSV link in the "Download" section. The data was grouped into folders upon downloading, and the folder titled "Metadata_Country" allowed us to get the income group for each country as well as we consolidated all data into one spreadsheet. We trust this data as it is from the World Bank, an international organization and reliable source.

Question 5:

[1 mark] Write code below to import your data into R. Assign your dataset to an object.

```
library(readxl)
GDP_data <- read_excel("GDP_HC_project1.xlsx")
GDP_data
```

```
## # A tibble: 264 x 6
##    Country_Name Country_Code Income_Group Indicator_Name GDP_percen_2017
##    <chr>        <chr>        <chr>        <chr>                     <dbl>
##  1 Aruba        ABW          High income  Current healt~            NA
##  2 Afghanistan  AFG          Low income   Current healt~            11.8
##  3 Angola       AGO          Lower middl~ Current healt~             2.79
##  4 Albania      ALB          Upper middl~ Current healt~            NA
##  5 Andorra      AND          High income  Current healt~            10.3
##  6 Arab World   ARB          <NA>         Current healt~             4.86
##  7 United Arab~ ARE          High income  Current healt~             3.33
##  8 Argentina    ARG          Upper middl~ Current healt~             9.12
##  9 Armenia      ARM          Upper middl~ Current healt~            10.4
## 10 American Sa~ ASM          Upper middl~ Current healt~            NA
## # ... with 254 more rows, and 1 more variable: Region <chr>
```

Question 6:

[3 marks] Use code in R to answer the following questions:

i) What are the dimensions of the dataset?

```
library(dplyr)
```

```
##
## 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
```

```
dim(GDP_data)
```

```
## [1] 264   6
```

ii) Provide a list of variable names.

```
names(GDP_data)
```

```
## [1] "Country_Name"    "Country_Code"    "Income_Group"    "Indicator_Name"
## [5] "GDP_percen_2017" "Region"
```

iii) Print the first six rows of the dataset.

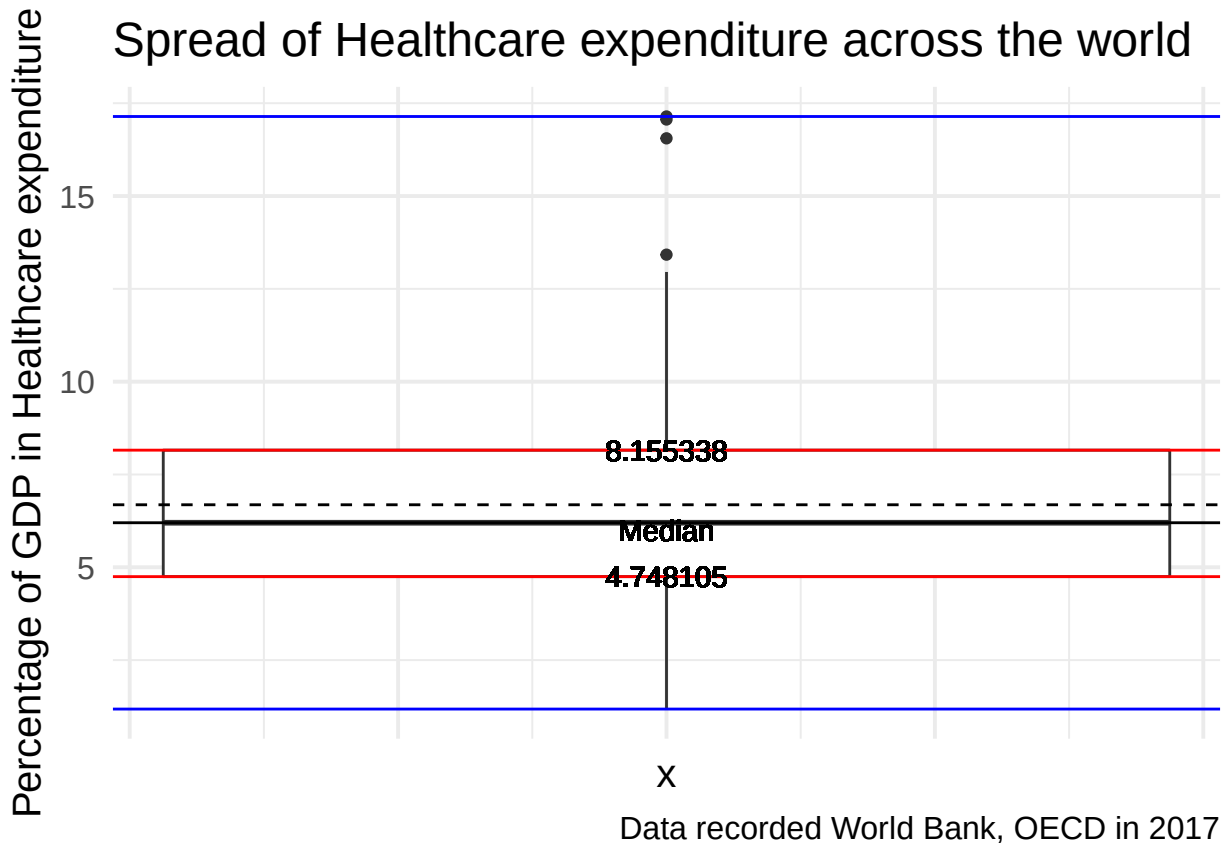
```
head(GDP_data)
```

```
## # A tibble: 6 x 6
##   Country_Name Country_Code Income_Group Indicator_Name GDP_percen_2017
##   <chr>        <chr>        <chr>        <chr>                    <dbl>
## 1 Aruba        ABW          High income  Current healt~           NA
## 2 Afghanistan  AFG          Low income   Current healt~           11.8
## 3 Angola       AGO          Lower middl~ Current healt~            2.79
## 4 Albania      ALB          Upper middl~ Current healt~           NA
## 5 Andorra      AND          High income  Current healt~           10.3
## 6 Arab World   ARB          <NA>         Current healt~            4.86
## # ... with 1 more variable: Region <chr>
```

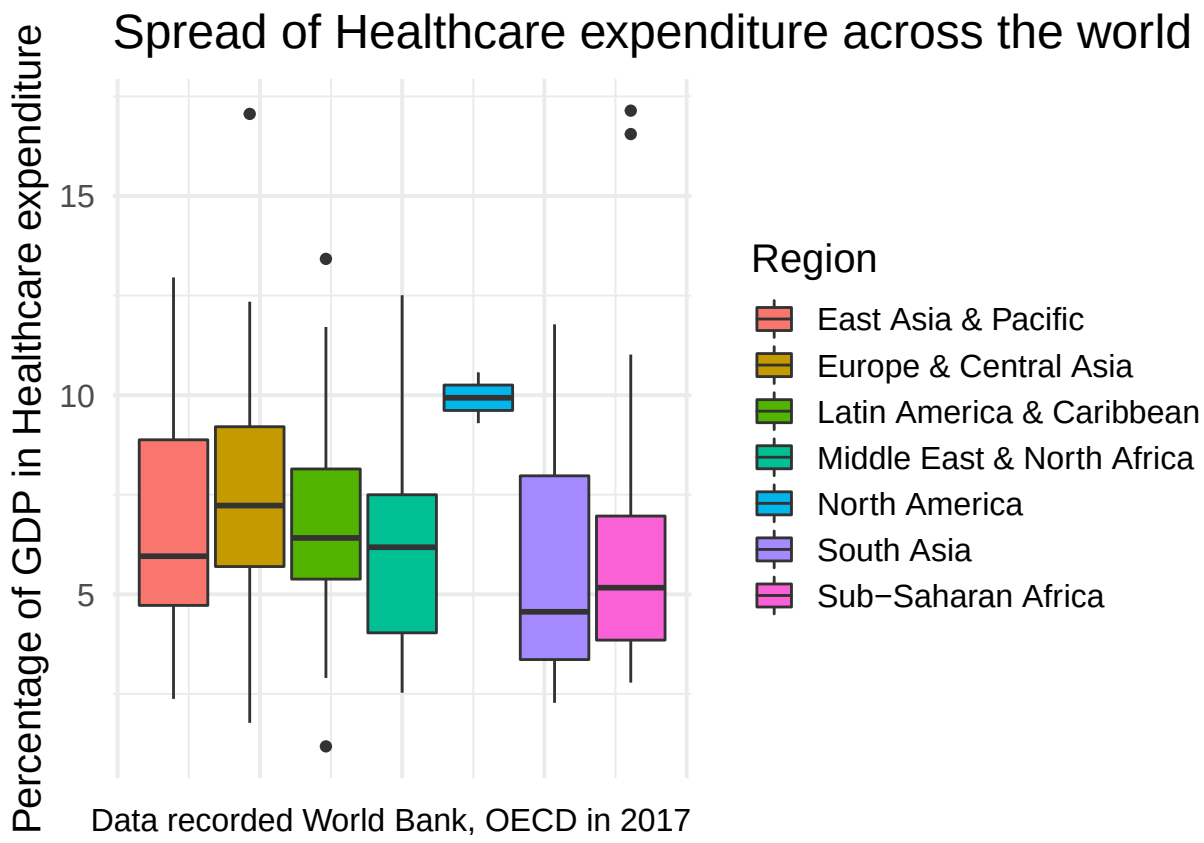
Question 7: [4 marks] Use the data to demonstrate a statistical concept from Part I of the course. Describe the concept that you are demonstrating and interpret the findings. This should be a combination of code and written explanation.

Code:

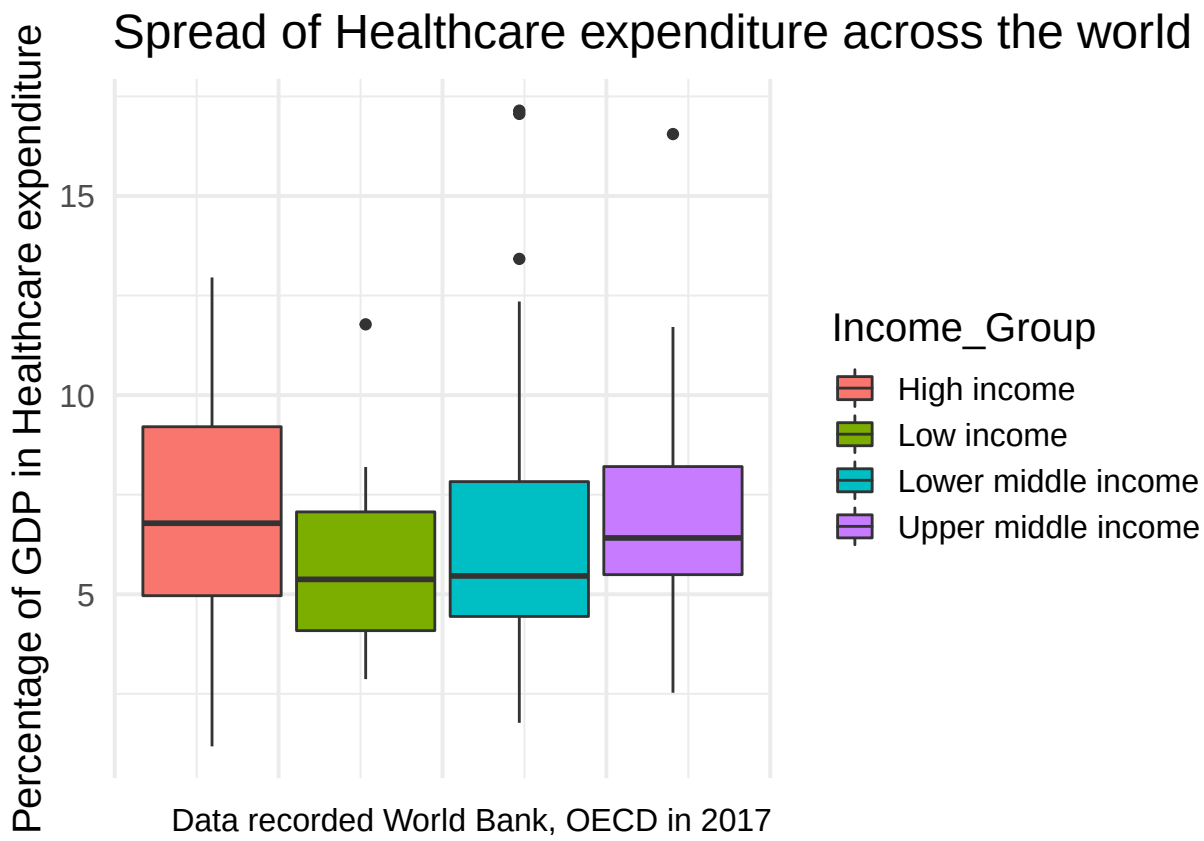
```
library(ggplot2)
library(dplyr)
library(ggrepel)

GDP_data<- GDP_data %>% na.omit(GDP_percen_2017)

summary <- GDP_data %>% summarize(min= min(GDP_percen_2017),
                                  mean= mean(GDP_percen_2017),
                                  Q1 = quantile (GDP_percen_2017, 0.25),
                                  median = median(GDP_percen_2017),
                                  Q3 = quantile (GDP_percen_2017, 0.75),
                                  max = max(GDP_percen_2017))

GDP_plot <- ggplot(GDP_data, aes(y= GDP_percen_2017)) +
  geom_boxplot(na.rm=TRUE) +
  ylab("Percentage of GDP in Healthcare expenditure")+
  labs(title = "Spread of Healthcare expenditure across the world",
       caption = "Data recorded World Bank, OECD in 2017") +
  theme_minimal(base_size= 15)+
  scale_x_continuous(labels =NULL)+
  geom_hline(aes(yintercept = 1.18121), col = "blue") +
geom_hline(aes(yintercept = 17.14256), col = "blue") +
  geom_hline(aes(yintercept = 6.683917), col = "black", lty = 2) +
  geom_hline(aes(yintercept = 6.199512), col = "black") +
   geom_text(aes(0, 6, label = "Median")) +
  geom_hline(aes(yintercept = 8.155338), col = "red") +
  geom_text(aes(0, 8.155338, label = 8.155338)) +
  geom_hline(aes(yintercept =   4.748105), col = "red") +
 geom_text(aes(0, 4.748105, label = 4.748105))

GDP_plot2 <- ggplot(GDP_data, aes(y= GDP_percen_2017, fill= Region)) +
  geom_boxplot(na.rm=TRUE) +
  ylab("Percentage of GDP in Healthcare expenditure")+
  labs(title = "Spread of Healthcare expenditure across the world",
       caption = "Data recorded World Bank, OECD in 2017") +
  theme_minimal(base_size= 15)+
  scale_x_continuous(labels =NULL)

GDP_plot3 <- ggplot(GDP_data, aes(y= GDP_percen_2017, fill = Income_Group)) +
  geom_boxplot(na.rm=TRUE) +
  ylab("Percentage of GDP in Healthcare expenditure")+
  labs(title = "Spread of Healthcare expenditure across the world",
       caption = "Data recorded World Bank, OECD in 2017") +
  theme_minimal(base_size= 15)+
  scale_x_continuous(labels =NULL)

GDP_plot
```


Spread of Healthcare expenditure across the world
Percentage of GDP in Healthcare expenditure
15
10
5
8.155338
Median
4.748105
X
Data recorded World Bank, OECD in 2017

```
GDP_plot2
```


Spread of Healthcare expenditure across the world
Percentage of GDP in Healthcare expenditure
15
10
5
Data recorded World Bank, OECD in 2017
Region
East Asia & Pacific
Europe & Central Asia
Latin America & Caribbean
Middle East & North Africa
North America
South Asia
Sub−Saharan Africa

```
GDP_plot3
```



Explanation:

In this analysis, we aim to demonstrate the spread of healthcare expenditure in different countries through the statistical concept of box plot and measures of spread. We are using a few parameters to assess this distribution. First, we find the measures of spread for all of the data combined, divide it according to region, and finally, according to income group expenditure.

From the first plot, the findings that we have suggest that the min = 1.18121, mean= 6.683917, first quartile = 4.748105, median = 6.199512, Third quartile (Q3) = 8.155338, and maximum value = 17.14256%. This implies that the interquartile range IQR = Q3 - Q1 = 3.407233. 50% of the world's countries, for which we have data available, spend about 4.7% to 8.2% of their GDP on healthcare. The minimum value is 1%; this implies that 75% of the world spends about 1%-8% of their GDP on healthcare. Given that the maximum value of GDP expenditure is 17.14%, this implies about only 25% of the world spends in the range of 8%-17.4%. Plots 2 and 3 give more insight into what countries could be in the top 25%.

The second plot compares countries according to region. Interestingly, North American, European, and central Asian countries have the highest median expenditure, while Sub-saharan and South Asian countries have the lowest median expenditure. The spread of North American countries is relatively small, with the smallest interquartile range, indicating that all countries of this subgroup spend high amounts on healthcare. The third plot analyses the distribution of countries according to income level. The healthcare expenditure of Low income and lower-middle-income countries is about the same. High-income countries have the highest median expenditure and most extensive spread, whereas upper-middle-income countries are between high and low-income countries.

Overall, the primary plot is the first one, as it confirms the suspicion that most of the world spends on healthcare in a reasonable range. Only one-quarter of all the countries in the world spend above 8% of their GDP on healthcare. These may be high- income countries from North America and Europe. The average expenditure of GDP on healthcare across the world is about 6.7%.

Part 2:

Question 1: [1 mark] Include your work for Part I.

Answer: Included above

Question 2: Describe a quantity you will estimate as an outcome in your problem using probability notation. Are you planning to calculate marginal probabilities? Conditional probabilities?

Answer:

We will be looking at the amount spent on healthcare that is considered with respect to the mean value amongst different global regions. The goal is to analyse the probability of a country to have a GDP expenditure on healthcare higher than the mean, and then replicate that analysis for each individual geographical region to find patterns in the geography of healthcare expenditure.

We will analyze this in terms of conditional probabilities of countries having a GDP expenditure of healthcare higher than the mean given their geographical region. The conditional probability is calculated the amount of countries over the mean out of the countries specifically in that specific geographical region.

The probability notation for the above is: **P(GDP expenditure > mean | Region)**

```
regions <- GDP_data %>% arrange(Region)
```

```
LA_Carribean <- regions %>% filter(Region %in% c("Latin America & Caribbean"))
LA_Carribean
```

```
## # A tibble: 35 x 6
##    Country_Name Country_Code Income_Group Indicator_Name GDP_percen_2017
##    <chr>        <chr>        <chr>        <chr>                    <dbl>
##  1 Argentina    ARG          Upper middl~ Current healt~            9.12
##  2 Antigua and~ ATG          High income  Current healt~            4.53
##  3 Bahamas, The BHS          High income  Current healt~            5.76
##  4 Belize       BLZ          Upper middl~ Current healt~            5.64
##  5 Bolivia      BOL          Lower middl~ Current healt~            6.44
##  6 Brazil       BRA          Upper middl~ Current healt~            9.47
##  7 Barbados     BRB          High income  Current healt~            6.78
##  8 Chile        CHL          High income  Current healt~            8.98
##  9 Colombia     COL          Upper middl~ Current healt~            7.23
## 10 Costa Rica   CRI          Upper middl~ Current healt~            7.33
## # ... with 25 more rows, and 1 more variable: Region <chr>
```

```
South_Asia <- regions %>% filter(Region %in% c("South Asia"))
South_Asia
```

```
## # A tibble: 7 x 6
##   Country_Name Country_Code Income_Group Indicator_Name GDP_percen_2017
##   <chr>        <chr>        <chr>        <chr>                    <dbl>
## 1 Afghanistan  AFG          Low income   Current healt~           11.8
## 2 Bangladesh   BGD          Lower middl~ Current healt~            2.27
## 3 Bhutan       BTN          Lower middl~ Current healt~            3.19
## 4 India        IND          Lower middl~ Current healt~            3.53
## 5 Madagascar   MDG          Upper middl~ Current healt~            5.50
## 6 Norway       NOR          Lower middl~ Current healt~           10.4
## 7 Other small~ OSS          Lower middl~ Current healt~            4.56
## # ... with 1 more variable: Region <chr>
```

```
Europe_CA <- regions %>% filter(Region %in% c("Europe & Central Asia"))
Europe_CA
```

```
## # A tibble: 49 x 6
##    Country_Name Country_Code Income_Group Indicator_Name GDP_percen_2017
##    <chr>        <chr>        <chr>        <chr>                     <dbl>
##  1 Andorra      AND          High income  Current healt~            10.3
##  2 Armenia      ARM          Upper middl~ Current healt~            10.4
##  3 Austria      AUT          High income  Current healt~            10.4
##  4 Azerbaijan   AZE          Upper middl~ Current healt~             6.65
##  5 Belgium      BEL          High income  Current healt~            10.3
##  6 Bulgaria     BGR          Upper middl~ Current healt~             8.10
##  7 Bosnia and ~ BIH          Upper middl~ Current healt~             8.93
##  8 Belarus      BLR          Upper middl~ Current healt~             5.93
##  9 Switzerland  CHE          High income  Current healt~            12.3
## 10 Cyprus       CYP          High income  Current healt~             6.68
## # ... with 39 more rows, and 1 more variable: Region <chr>
```

```
EA_Pacific <- regions %>% filter(Region %in% c("East Asia & Pacific"))
EA_Pacific
```

```
## # A tibble: 31 x 6
##    Country_Name Country_Code Income_Group Indicator_Name GDP_percen_2017
##    <chr>        <chr>        <chr>        <chr>                     <dbl>
##  1 Australia    AUS          High income  Current healt~             9.21
##  2 Brunei Daru~ BRN          High income  Current healt~             2.37
##  3 China        CHN          Upper middl~ Current healt~             5.15
##  4 Fiji         FJI          Upper middl~ Current healt~             3.50
##  5 Micronesia,~ FSM          Lower middl~ Current healt~            12.4
##  6 Indonesia    IDN          Upper middl~ Current healt~             2.99
##  7 Jordan       JOR          High income  Current healt~             8.12
##  8 Kyrgyz Repu~ KGZ          Lower middl~ Current healt~             6.19
##  9 Cambodia     KHM          Lower middl~ Current healt~             5.92
## 10 St. Kitts a~ KNA          High income  Current healt~             5.04
## # ... with 21 more rows, and 1 more variable: Region <chr>
```

```
SS_Africa <- regions %>% filter(Region %in% c("Sub-Saharan Africa"))
SS_Africa
```

```
## # A tibble: 43 x 6
##    Country_Name Country_Code Income_Group Indicator_Name GDP_percen_2017
##    <chr>        <chr>        <chr>        <chr>                     <dbl>
##  1 Angola       AGO          Lower middl~ Current healt~             2.79
##  2 Burundi      BDI          Low income   Current healt~             7.52
##  3 Benin        BEN          Lower middl~ Current healt~             3.72
##  4 Burkina Faso BFA          Low income   Current healt~             6.92
##  5 Botswana     BWA          Upper middl~ Current healt~             6.13
##  6 Central Afr~ CAF          Low income   Current healt~             5.82
##  7 Cote d'Ivoi~ CIV          Lower middl~ Current healt~             4.45
##  8 Cameroon     CMR          Lower middl~ Current healt~             4.67
##  9 Congo, Dem.~ COD          Low income   Current healt~             3.98
## 10 Congo, Rep.  COG          Lower middl~ Current healt~             2.93
## # ... with 33 more rows, and 1 more variable: Region <chr>
```

```
ME_NA <- regions %>% filter(Region %in% c("Middle East & North Africa"))
ME_NA
```

```
## # A tibble: 18 x 6
##    Country_Name Country_Code Income_Group Indicator_Name GDP_percen_2017
##    <chr>        <chr>        <chr>        <chr>                     <dbl>
##  1 United Arab~ ARE          High income  Current healt~             3.33
##  2 Bahrain      BHR          High income  Current healt~             4.75
##  3 Djibouti     DJI          Lower middl~ Current healt~             3.32
##  4 Algeria      DZA          Lower middl~ Current healt~             6.37
##  5 Egypt, Arab~ EGY          Lower middl~ Current healt~             5.29
##  6 Ireland      IRL          Upper middl~ Current healt~             7.18
##  7 Iran, Islam~ IRN          Upper middl~ Current healt~             8.66
##  8 Iceland      ISL          High income  Current healt~             8.33
##  9 Jamaica      JAM          Upper middl~ Current healt~             5.99
## 10 Korea, Rep.  KOR          High income  Current healt~             7.60
## 11 Lao PDR      LAO          Upper middl~ Current healt~             2.53
## 12 Liberia      LBR          Upper middl~ Current healt~             8.16
## 13 Mali         MLI          High income  Current healt~             3.79
## 14 OECD members OED          High income  Current healt~            12.5
## 15 Paraguay     PRY          Lower middl~ Current healt~             6.65
## 16 South Asia   SAS          High income  Current healt~             3.46
## 17 Seychelles   SYC          Low income   Current healt~             5.01
## 18 Trinidad an~ TTO          Lower middl~ Current healt~             6.98
## # ... with 1 more variable: Region <chr>
```

```
North_America <- regions %>% filter(Region %in% c("North America"))
North_America
```

```
## # A tibble: 2 x 6
##   Country_Name Country_Code Income_Group Indicator_Name GDP_percen_2017
##   <chr>        <chr>        <chr>        <chr>                     <dbl>
## 1 Canada       CAN          High income  Current healt~            10.6
## 2 Uruguay      URY          High income  Current healt~             9.30
## # ... with 1 more variable: Region <chr>
```

```
LA_Carribean_prob <- LA_Carribean %>% filter(GDP_percen_2017 >= 6.683917)
LA_Carribean_prob
```

```
## # A tibble: 16 x 6
##    Country_Name Country_Code Income_Group Indicator_Name GDP_percen_2017
##    <chr>        <chr>        <chr>        <chr>                     <dbl>
##  1 Argentina    ARG          Upper middl~ Current healt~             9.12
##  2 Brazil       BRA          Upper middl~ Current healt~             9.47
##  3 Barbados     BRB          High income  Current healt~             6.78
##  4 Chile        CHL          High income  Current healt~             8.98
##  5 Colombia     COL          Upper middl~ Current healt~             7.23
##  6 Costa Rica   CRI          Upper middl~ Current healt~             7.33
##  7 Cuba         CUB          Upper middl~ Current healt~            11.7
##  8 Ecuador      ECU          Upper middl~ Current healt~             8.26
##  9 Honduras     HND          Lower middl~ Current healt~             7.86
## 10 Haiti        HTI          Low income   Current healt~             8.04
```

```
## 11 Italy        ITA          Upper middl~ Current healt~            8.84
## 12 Kiribati     KIR          High income  Current healt~           10.8
## 13 Panama       PAN          Upper middl~ Current healt~            7.32
## 14 Portugal     PRT          Upper middl~ Current healt~            8.97
## 15 Sierra Leone SLE          Lower middl~ Current healt~           13.4
## 16 Eswatini     SWZ          High income  Current healt~            6.93
## # ... with 1 more variable: Region <chr>
```

```
South_Asia_prob <- South_Asia %>% filter(GDP_percen_2017 >= 6.683917)
South_Asia_prob
```

```
## # A tibble: 2 x 6
##   Country_Name Country_Code Income_Group Indicator_Name GDP_percen_2017
##   <chr>        <chr>        <chr>        <chr>                    <dbl>
## 1 Afghanistan  AFG          Low income   Current healt~            11.8
## 2 Norway       NOR          Lower middl~ Current healt~            10.4
## # ... with 1 more variable: Region <chr>
```

```
Europe_CA_prob <- Europe_CA %>% filter(GDP_percen_2017 >= 6.683917)
Europe_CA_prob
```

```
## # A tibble: 30 x 6
##    Country_Name Country_Code Income_Group Indicator_Name GDP_percen_2017
##    <chr>        <chr>        <chr>        <chr>                    <dbl>
##  1 Andorra      AND          High income  Current healt~           10.3
##  2 Armenia      ARM          Upper middl~ Current healt~           10.4
##  3 Austria      AUT          High income  Current healt~           10.4
##  4 Belgium      BEL          High income  Current healt~           10.3
##  5 Bulgaria     BGR          Upper middl~ Current healt~            8.10
##  6 Bosnia and ~ BIH          Upper middl~ Current healt~            8.93
##  7 Switzerland  CHE          High income  Current healt~           12.3
##  8 Cyprus       CYP          High income  Current healt~            6.68
##  9 Czech Repub~ CZE          High income  Current healt~            7.23
## 10 Germany      DEU          High income  Current healt~           11.2
## # ... with 20 more rows, and 1 more variable: Region <chr>
```

```
EA_Pacific_prob <- EA_Pacific %>% filter(GDP_percen_2017 >= 6.683917)
EA_Pacific_prob
```

```
## # A tibble: 13 x 6
##    Country_Name Country_Code Income_Group Indicator_Name GDP_percen_2017
##    <chr>        <chr>        <chr>        <chr>                    <dbl>
##  1 Australia    AUS          High income  Current healt~            9.21
##  2 Micronesia,~ FSM          Lower middl~ Current healt~           12.4
##  3 Jordan       JOR          High income  Current healt~            8.12
##  4 Latin Ameri~ LAC          Lower middl~ Current healt~            7.97
##  5 Malta        MLT          Lower middl~ Current healt~            9.34
##  6 Malawi       MWI          Upper middl~ Current healt~            9.65
##  7 Namibia      NAM          High income  Current healt~            8.55
##  8 Nauru        NRU          High income  Current healt~           11.0
##  9 Palau        PLW          Lower middl~ Current healt~           12.0
## 10 Post-demogr~ PST          High income  Current healt~           13.0
```

```
## 11 Latin Ameri~ TLA          Lower middl~ Current healt~            7.96
## 12 Middle East~ TMN          Upper middl~ Current healt~            6.68
## 13 World        WLD          Upper middl~ Current healt~            9.88
## # ... with 1 more variable: Region <chr>
```

```
SS_Africa_prob <- SS_Africa %>% filter(GDP_percen_2017 >= 6.683917)
SS_Africa_prob
```

```
## # A tibble: 12 x 6
##    Country_Name Country_Code Income_Group Indicator_Name GDP_percen_2017
##    <chr>        <chr>        <chr>        <chr>                    <dbl>
##  1 Burundi      BDI          Low income   Current healt~            7.52
##  2 Burkina Faso BFA          Low income   Current healt~            6.92
##  3 Comoros      COM          Lower middl~ Current healt~            7.38
##  4 Guinea-Biss~ GNB          Low income   Current healt~            7.24
##  5 Lebanon      LBN          Low income   Current healt~            8.20
##  6 Moldova      MDA          Low income   Current healt~            7.01
##  7 North Ameri~ NAC          Upper middl~ Current healt~           16.6
##  8 Niger        NER          Lower middl~ Current healt~            7.74
##  9 San Marino   SMR          Low income   Current healt~            7.36
## 10 Sweden       SWE          Lower middl~ Current healt~           11.0
## 11 Tuvalu       TUV          Lower middl~ Current healt~           17.1
## 12 South Africa ZAF          Lower middl~ Current healt~            8.11
## # ... with 1 more variable: Region <chr>
```

```
ME_NA_prob <- ME_NA %>% filter(GDP_percen_2017 >= 6.683917)
ME_NA_prob
```

```
## # A tibble: 7 x 6
##   Country_Name Country_Code Income_Group Indicator_Name GDP_percen_2017
##   <chr>        <chr>        <chr>        <chr>                    <dbl>
## 1 Ireland      IRL          Upper middl~ Current healt~            7.18
## 2 Iran, Islam~ IRN          Upper middl~ Current healt~            8.66
## 3 Iceland      ISL          High income  Current healt~            8.33
## 4 Korea, Rep.  KOR          High income  Current healt~            7.60
## 5 Liberia      LBR          Upper middl~ Current healt~            8.16
## 6 OECD members OED          High income  Current healt~           12.5
## 7 Trinidad an~ TTO          Lower middl~ Current healt~            6.98
## # ... with 1 more variable: Region <chr>
```

```
North_America_prob <- North_America %>% filter(GDP_percen_2017 >= 6.683917)
North_America_prob
```

```
## # A tibble: 2 x 6
##   Country_Name Country_Code Income_Group Indicator_Name GDP_percen_2017
##   <chr>        <chr>        <chr>        <chr>                    <dbl>
## 1 Canada       CAN          High income  Current healt~           10.6
## 2 Uruguay      URY          High income  Current healt~            9.30
## # ... with 1 more variable: Region <chr>
```

Calculations:

Conditional Probabilities

Latin America & Carribean:

P(countries > mean | LA countries) = (16 countries above mean/42 total)*100 = 38.09% of countries spend above mean amount of GDP on healthcare in this region

South Asia:

P(countries > mean | SA countries)= (2 countries above mean/8 total)*100 = 25% of countries spend above mean amount of GDP on healthcare in this region

Europe & Central Asia:

P(countries > mean | Europe CA countries) = (30 countries above mean/58 total)*100 = 51.72% of countries spend above mean amount of GDP on healthcare in this region

Eastern Asia & Pacific:

P(countries > mean | EA Pacific) = (13 countries above mean/37 total)*100 = 35.14% of countries spend above mean amount of GDP on healthcare in this region

Sub-Saharan Africa:

P(countries > mean | SS African countries) = (12 countries above mean/48 total)*100 = 25% of countries spend above mean amount of GDP on healthcare in this region

Middle East & Northern Africa:

P(countries > mean | ME NA countries) = (7 countries above mean/21 total)*100 = 33.33% of countries spend above mean amount of GDP on healthcare in this region

North America:

P(countries > mean | North American countries) = (2 countries above mean/3 countries in region total)*100 = 66.67% of countries spend above mean amount of GDP on healthcare in this region

Question 3: [3 marks] Describe the type of theoretical distribution that is relevant for your data.

1. What type of variable are you estimating (continuous, categorical, ordinal, etc)?
2. What theoretical distribution that we have talked about would potentially be appropriate to use with these data (Normal, Binomial, Poisson. . .)
3. Why is this an appropriate model for the data you are studying?

Answer:

We are estimating the variable "GDP_percen_2017" which describes the annual expenditure on healthcare in GDP spending for all countries in our dataset. This variable is a continuous quantitative variable.

The theoretical distribution most appropriate to use with these data is the Normal distribution because we are estimating a continuous variable with parameters of mean and standard deviation being relevant, where the mean and median are very close to each other (In Normal distributions, mean= median).

```
GDP_data<- GDP_data %>% na.omit(GDP_percen_2017)

why_normal <- GDP_data %>% summarize(mean= mean(GDP_percen_2017),
                                     median = median(GDP_percen_2017))
why_normal
```

```
## # A tibble: 1 x 2
##    mean median
##   <dbl>  <dbl>
## 1  6.68   6.20
```

We cannot use binomial distributions since the data does not examine an experiment with a number of trials that have two possible, mutually exclusive outcomes. We can also not use Poisson because the data does not describe the count of occurrences of an event in fixed finite intervals of space and time.

Further, our normal distribution analysis confirms our hypothesis that Normal ditribution is most appropriate for our data.

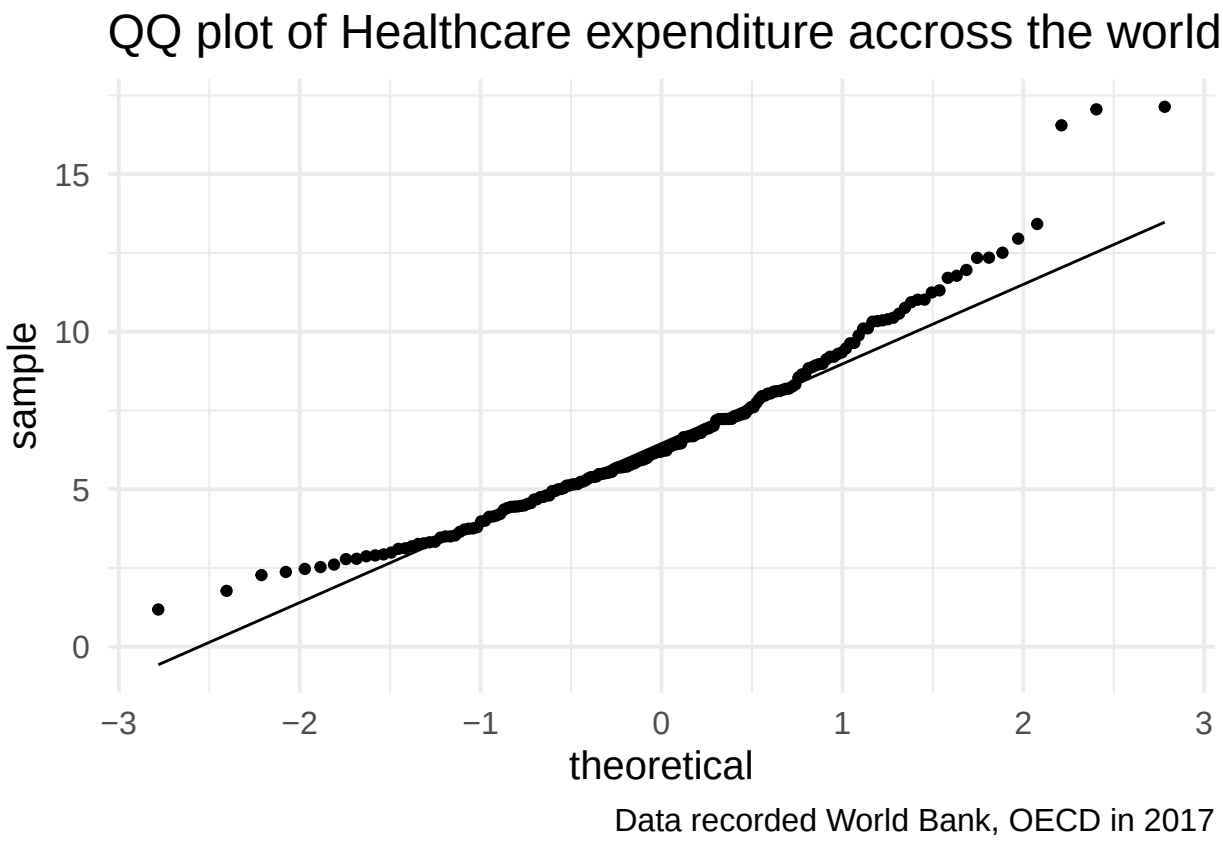
Question 4: [4 marks] Use the data you have to demonstrate a statistical concept from Part II of the course. Describe the concept that you are demonstrating and interpret the findings. This should include code in R, a visual of some kind and text interpretation.

Answer: Since we proposed that the most appropriate theoretical distribution for these data is Normal distribution, we are using these data to see how close it is to the afformentioned by perfoming a Normal distribution analysis.

```
library(dplyr)
library(ggplot2)

GDP_data_Normal <- GDP_data %>% arrange(GDP_percen_2017) %>%
  mutate (quantile= row_number()/n(),
          Z_score = qnorm(quantile, mean = 0, sd = 1))

head(GDP_data_Normal)
```

```
## # A tibble: 6 x 8
##   Country_Name Country_Code Income_Group Indicator_Name GDP_percen_2017
##   <chr>        <chr>        <chr>        <chr>                    <dbl>
## 1 Venezuela, ~ VEN          High income  Current healt~            1.18
## 2 Monaco       MCO          Lower middl~ Current healt~            1.77
## 3 Bangladesh   BGD          Lower middl~ Current healt~            2.27
## 4 Brunei Daru~ BRN          High income  Current healt~            2.37
## 5 Papua New G~ PNG          High income  Current healt~            2.47
## 6 Lao PDR      LAO          Upper middl~ Current healt~            2.53
## # ... with 3 more variables: Region <chr>, quantile <dbl>, Z_score <dbl>
```

```
GDP_data_Normal_plot <- ggplot(GDP_data_Normal, aes(sample=GDP_percen_2017))+
  stat_qq() + stat_qq_line() + theme_minimal(base_size=15) +
  labs(title = "QQ plot of Healthcare expenditure accross the world",
       caption= "Data recorded World Bank, OECD in 2017")

GDP_data_Normal_plot
```



Our objective is to find out how "normal" our data is so we decided on a normal quantile plot (QQ plot). We chose the QQplot because it is a better way of determining "normality" compared to visual inspection of a bar chart or histogram. From the plot shown above, our data is not perfectly normally distributed because although the middle section of the plotted points is fairly coincide to the line, the tails curve off the 45 degree line. The tails curving off the 45 degree line shows that our data has a large standard deviation than a normally distributed variable.

Part 3:

Question 5:

5.a

[2 marks] Identify the statistical test that you applied to your data (must be a concept we covered in part III of the course).

The statistical test we will apply to our data is Kruskal-Wallis.

It is the non- parametric alternative to the ANOVA test which helps compare statistics of 3 or more groups, and we are using it due to comparing 4 income groups from our dataset. The Kruskal Wallis test compares the medians of different groups, and not the means, to test if at least one is significantly different from the other.

5.b

[2 marks] What assumptions are required by the method you chose in 5.a)? Describe how you assessed whether these assumptions are met by your dataset.

The assumptions needed for conducting a Kruskal Wallis test is that the data needs to have 3 or more comparison groups, the sample data is a simplr random sample (SRS) from the population, and each observation must be mutually independent. A large sample size or Normal distribution of the data are not required.

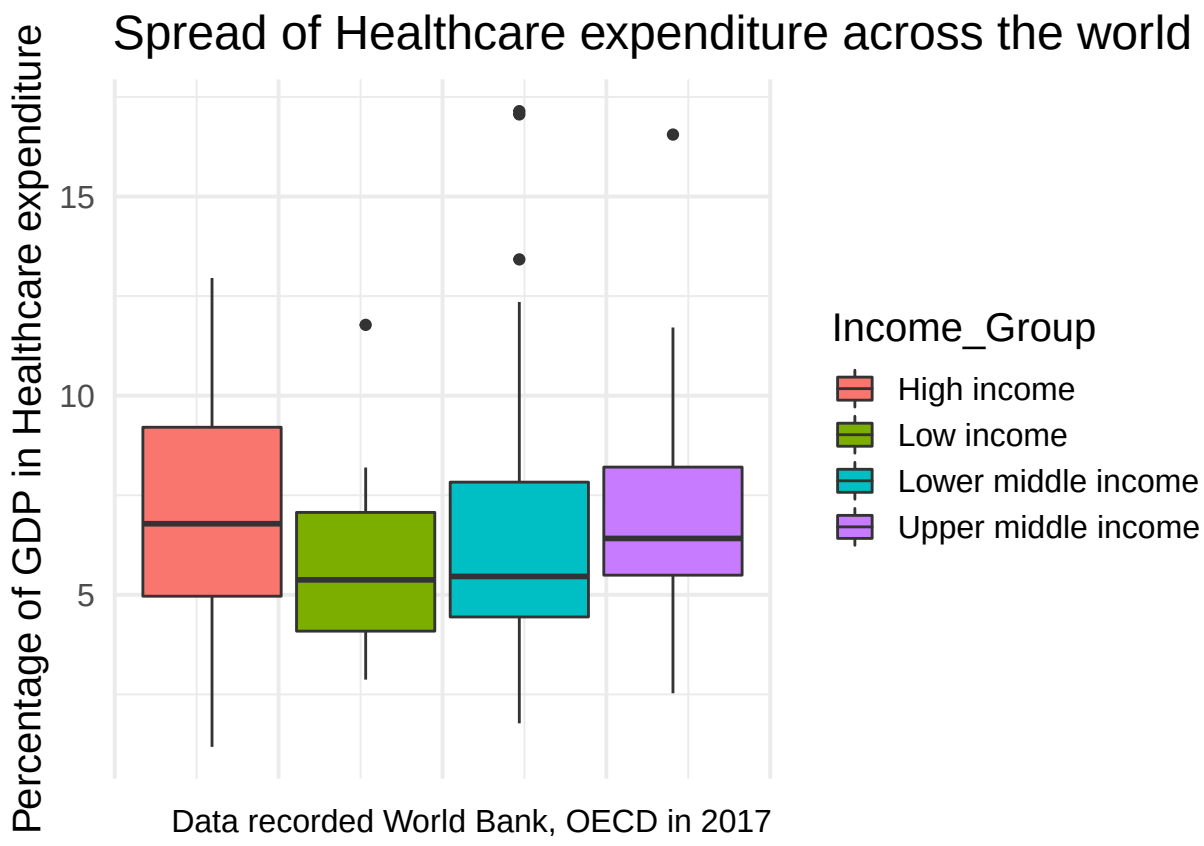
We know the data is a simple random sample and each observation is mutually independent due to data being collected from countries across the world whose GDP per capita expenditure on healthacre is mutually independent from each other for all practical purposes. Also, our observations contain data for countries whose healthcare expenditure. data is available, making it a simple random sample of the overall population. The comparison group assumption is satisfied as we are using it due to comparing 4 income groups from our dataset.

5.c [2 marks] Explain why this test is appropriate for the data you have and the question you are trying to answer. Use at least one visualization technique and include both the output and the r code that generated it.

We selected the Kruskal-Wallis test as it is a non-parametric test and does not require that the distribution of data be normal and is tolerant to large variance between group statistics we are trying to compare. This is important as our QQ plot from Part II indicated that the set is not perfectly normal (due to the tails curving off the 45 degree line in GDP_data_Normal_plot). Although ANOVA is tolerant to certain deviance from Normality, due to the difference in variance of the High income group and low income groups compared to all other groups, as depicted by the boxplot. (GDP_plot3)

We decided to use Kruskal Wallis instead as ANOVA assumes that the difference in variance amongst all groups being compared is minimal. The Kruskal-Wallis test does have less statistical power than ANOVA, but it was the only test that undoubtedly fit our specifications.

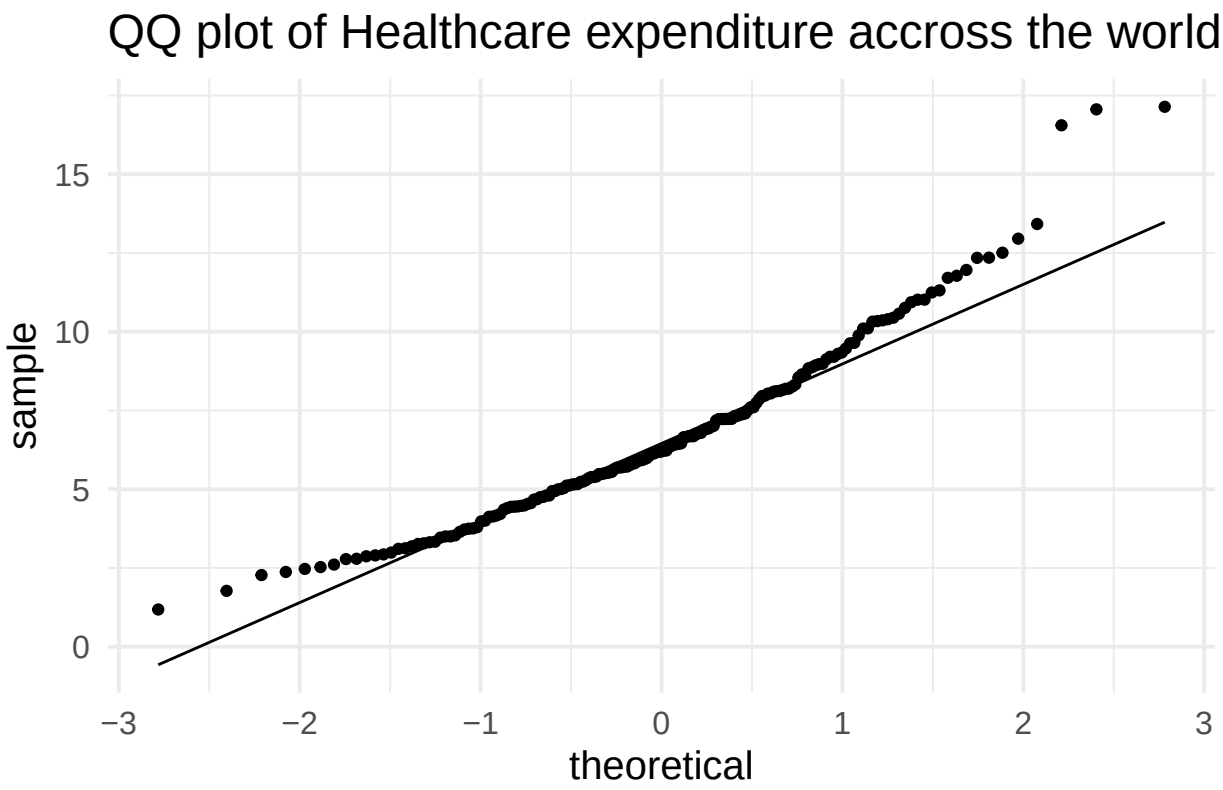
```
GDP_plot3 <- ggplot(GDP_data, aes(y= GDP_percen_2017, fill = Income_Group)) +
  geom_boxplot(na.rm=TRUE) +
  ylab("Percentage of GDP in Healthcare expenditure")+
  labs(title = "Spread of Healthcare expenditure across the world",
       caption = "Data recorded World Bank, OECD in 2017") +
  theme_minimal(base_size= 15)+
  scale_x_continuous(labels =NULL)

GDP_plot3
```



```r
library(dplyr)
library(ggplot2)

GDP_data_Normal <- GDP_data %>% arrange(GDP_percen_2017) %>%
  mutate (quantile= row_number()/n(),
          Z_score = qnorm(quantile, mean = 0, sd = 1))

head(GDP_data_Normal)
```

```
## # A tibble: 6 x 8
##   Country_Name Country_Code Income_Group Indicator_Name GDP_percen_2017
##   <chr>        <chr>        <chr>        <chr>                    <dbl>
## 1 Venezuela, ~ VEN          High income  Current healt~            1.18
## 2 Monaco       MCO          Lower middl~ Current healt~            1.77
## 3 Bangladesh   BGD          Lower middl~ Current healt~            2.27
## 4 Brunei Daru~ BRN          High income  Current healt~            2.37
## 5 Papua New G~ PNG          High income  Current healt~            2.47
## 6 Lao PDR      LAO          Upper middl~ Current healt~            2.53
## # ... with 3 more variables: Region <chr>, quantile <dbl>, Z_score <dbl>
```

```r
GDP_data_Normal_plot <- ggplot(GDP_data_Normal, aes(sample=GDP_percen_2017))+
  stat_qq() + stat_qq_line() + theme_minimal(base_size=15) +
  labs(title = "QQ plot of Healthcare expenditure accross the world",
       caption= "Data recorded World Bank, OECD in 2017")

GDP_data_Normal_plot
```



Data recorded World Bank, OECD in 2017

5.d [2 marks] Clearly state the null and alternative hypotheses for this test.

Null hypothesis: There is no difference in the medians of the percentage of their GDP a country spends on healthcare amongst the High income, low income, lower-middle and Upper-middle income groups.

Alternative hypothesis: There is a statically significant difference in the medians of the percentage of their GDP a country spends on healthcare amongst the High income, low income, lower-middle and Upper-middle income groups.

6. [2 marks] Include the R code you used to generate your results - annotate your code to help us follow your reasoning.

```
kruskal.test(GDP_percen_2017~ Income_Group, GDP_data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  GDP_percen_2017 by Income_Group
## Kruskal-Wallis chi-squared = 6.7433, df = 3, p-value = 0.08054
```

```
#ran kruskal test for income group and percentage spent with the format kruskal.test(outcome ~ group, d
#The outcome is the percentage of GDP spent on healthcare whereas the groups we are comparing are incom
```

7. [4 marks] Present your results in a clear summary. This should include both a text summary and a table or figure with appropriate labelling.

From the table "results" we know that our degrees of freedom are n-1 = 4-1 = 3. We also have a chi-squared value whic is the H statistic which is named the "Kruskal-Wallis chi-squared" because it follows a chi-square distribution. The H statistic is not necessary for our hypothesis testing. Finally, the most relevant outcome is the p- value of 0.08054 or approximately 0.081, which is not-significant given an alpha significance level assumed to be 5 % or 0.05.

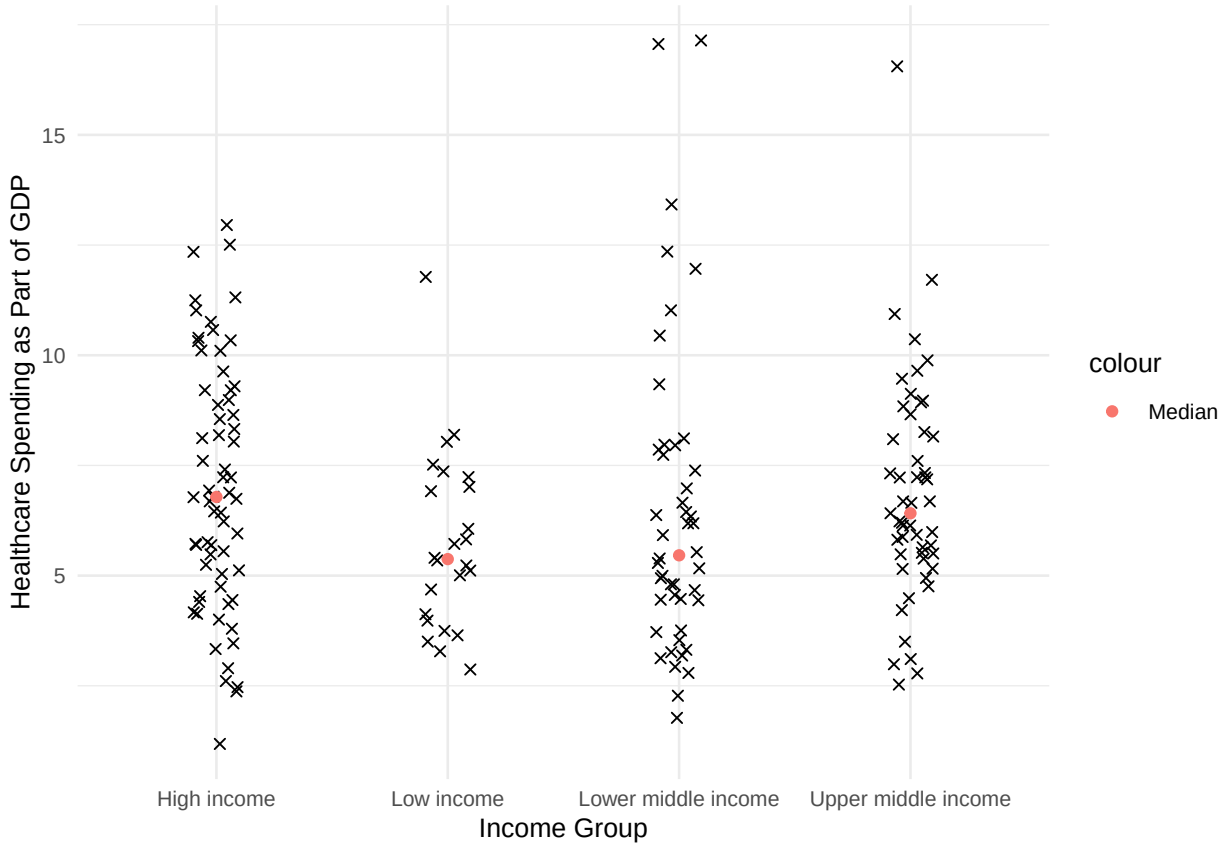
```
results <- data.frame("Kruskal-Wallis chi-squared" = 6.7433,
                      "df" = 3,
                      "p-value" = 0.08054)
results
```

```
##   Kruskal.Wallis.chi.squared df p.value
## 1                     6.7433  3 0.08054
```

```
GDP_data<- GDP_data %>% na.omit(GDP_percen_2017)
summary_stats <- GDP_data %>%
  group_by(Income_Group) %>%
  summarize(median_spend = median(GDP_percen_2017))
summary_stats
```

```
## # A tibble: 4 x 2
##   Income_Group        median_spend
##   <chr>                      <dbl>
## 1 High income                 6.78
## 2 Low income                  5.37
## 3 Lower middle income         5.46
## 4 Upper middle income         6.41
```

```
plot2 <- ggplot(GDP_data, aes(x = Income_Group, y = GDP_percen_2017)) +
  geom_jitter(pch = 4, width = 0.1) +
  geom_point(data = summary_stats, aes(y = median_spend,
                                       col = "Median"), pch = 19) +
  labs(y = "Healthcare Spending as Part of GDP", x = "Income Group") +
  theme_minimal(base_size = 10)

plot2
```


Healthcare Spending as Part of GDP
15
10
5
High income
Low income
Lower middle income
Upper middle income
Income Group
colour
Median

8. [4 marks] Interpret your findings. Include a statement about the strength of this testing, your conclusions and the generalizability of your findings.

Based on our p-value calculated by our Kruskall Wallis test, which was 0.08054, we fail to reject the null hypothesis for alpha = 0.05 of no difference between the medians of percentage of GDP based on income group.

Our results are not as strong as with an ANOVA test, but our data does not strongly support the assumptions needed for an ANOVA test. We can generalise these results to more income group sub-divisions in High income and Low income groups. Our conclusions are that the median expenditure on healthcare as a percentage of GDP for all economic groups in the countries are similar, due to failing to reject the Null hypothesis, which implies that better or worse economic status does not necessarily drive healthcare expenditure as a percent of GDP. Possible factors that do drive healthcare expenditure are government policies, availability of medical resources, and healthcare industry structures of countries.