FinalProject

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

THIS CODE LOADS ALL THE LIBRARIES WARNING: ALL THESE LIBRARIES MUST BE INSTALLED BEFORE LOADING THEM

```
library (tidyverse)
```

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

```
## -- Attaching packages ---------------------------------------------------
- tidyverse 1.3.0 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.4
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

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

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

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

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

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

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

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

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

```
## -- Conflicts ------------------------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library (knitr)
```

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

```
library (dplyr)
library (rpart)
```

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

```
library (partykit)
```

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

```
## Loading required package: grid
```

```
## Loading required package: libcoin
```

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

```
## Loading required package: mvtnorm
```

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

```
library (maps)
```

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

```
##
## Attaching package: 'maps'
```

```
## The following object is masked from 'package:purrr':
##
##     map
```

```
library (readr)
library (rvest)
```

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

```
## Loading required package: xml2
```

```
## Warning: package 'xml2' was built under R version 3.5.3
```

```
##
## Attaching package: 'rvest'
```

```
## The following object is masked from 'package:purrr':
##
##     pluck
```

```
## The following object is masked from 'package:readr':
##
##     guess_encoding
```

THIS CODE LOADS SOME OF THE DATA SETS

```
path <-
"https://raw.githubusercontent.com/ntaback/UofT_STA130/master/Fall2018/Finalp
roject/"

democracyindex2017 <- read_csv(paste0(path,"democracyindex2017.csv"))
```

```
## Parsed with column specification:
## cols(
##   Rank = col_character(),
##   Country = col_character(),
##   Score = col_character(),
##   `Electoral processand pluralism` = col_character(),
##   `Functioning ofgovernment` = col_character(),
##   Politicalparticipation = col_character(),
##   Politicalculture = col_character(),
##   Civilliberties = col_character(),
##   Category = col_character()
## )
```

```
education_cia2017 <- read_csv(paste0(path,"education_cia2017.csv"))
```

```
## Parsed with column specification:
## cols(
##   Rank = col_double(),
##   Country = col_character(),
##   `(% OF GDP)` = col_double(),
##   `Date of Information` = col_double()
## )
```

```
gdpppp_cia2017 <- read_csv(paste0(path,"gdpppp_cia2017.csv"))
```

```
## Parsed with column specification:
## cols(
##   Rank = col_double(),
##   Country = col_character(),
##   `GDP - PER CAPITA (PPP)` = col_character(),
##   `Date of Information` = col_character()
## )
```

```
lifeexpect_cia2017 <- read_csv(paste0(path,"lifeexpect_cia2017.csv"))
```

```
## Parsed with column specification:
## cols(
##   Rank = col_double(),
##   Country = col_character(),
##   `(YEARS)` = col_double(),
##   `Date of Information` = col_character()
## )
```

```
healthexpend_cia2017 <- read_csv(paste0(path,"healthexpend_cia2017.csv"))
```

```
## Parsed with column specification:
## cols(
##   Rank = col_double(),
##   Country = col_character(),
##   `(% OF GDP)` = col_double(),
##   `Date of Information` = col_double()
## )
```

```
internetusers_cia2017 <- read_csv(paste0(path,"internetusers_cia2017.csv"))
```

```
## Parsed with column specification:
## cols(
##   Rank = col_double(),
##   Country = col_character(),
##   `INTERNET USERS` = col_number(),
##   `Date of Information` = col_character()
## )
```

```
telephonelines_cia2017 <- read_csv(paste0(path,"telephonelines_cia2017.csv"))
```

```
## Parsed with column specification:
## cols(
##   Rank = col_double(),
##   Country = col_character(),
##   `TELEPHONES - MAIN LINES IN USE` = col_number(),
##   `Date of Information` = col_character()
## )
```

```
population_cia2017 <- read_csv(paste0(path,"population_cia2017.csv"))
```

```
## Parsed with column specification:
## cols(
##   Rank = col_double(),
##   Country = col_character(),
##   POPULATION = col_number(),
##   `Date of Information` = col_character()
## )
```

```
world_regions <- read_csv(paste0(path,"world_regions.csv"))
```

```
## Parsed with column specification:
## cols(
##   Country = col_character(),
##   Region = col_character(),
##   `Global South` = col_character()
## )
```

THIS CODE LOADS THE OPTIONAL DATA SETS

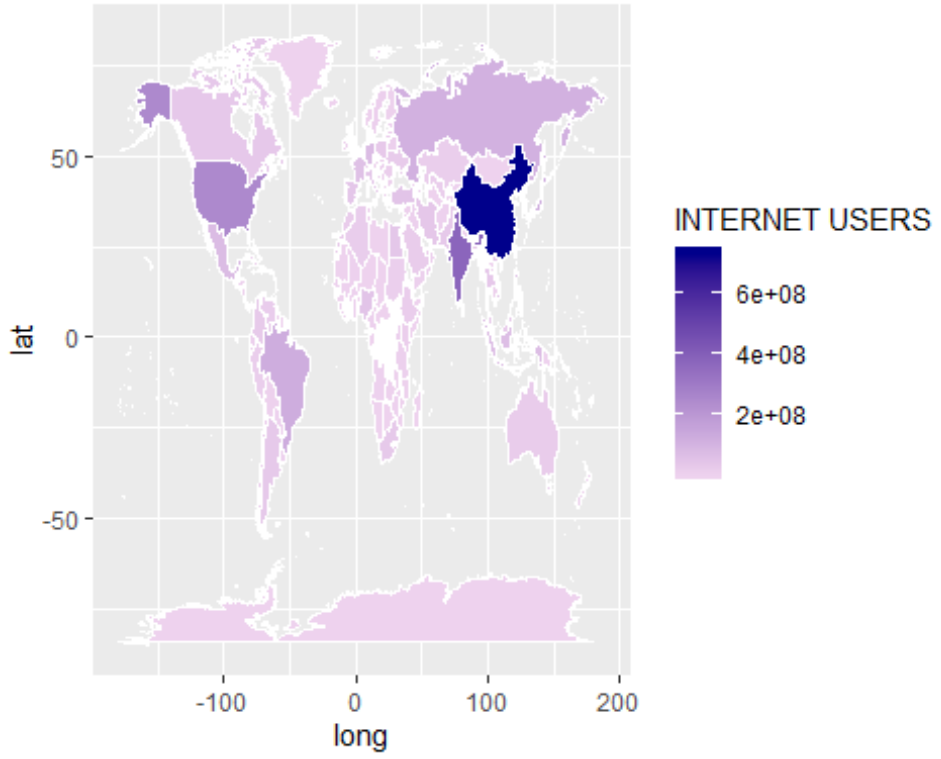
```
get_CIAWFB_data <- function(table_url){
  library(rvest)
  dat <- xml2::read_html(table_url) %>% rvest::html_table()
  dat[[1]]
}
medianage_cia2017 <-
get_CIAWFB_data("https://www.cia.gov/library/publications/resources/the-world-factbook/rankorder/2177rank.html")
```

Internet users world-wide geopgraphical area mapping

```r
world <- map_data("world")

#nternetusers_cia2017 <- read_csv("internetusers_cia2017.csv")

iu <- internetusers_cia2017 %>% rename(region = Country)

iu$region[4] <- "USA" # to match world map data

iu <- semi_join(iu, world, by = "region") #only keep countries according to
world map data

# code below is modified from
# https://stackoverflow.com/questions/29614972/ggplot-us-state-map-colors-
are-fine-polygons-jagged-r
gg <- ggplot()

gg <- gg + geom_map(
  data = world,
  map = world,
  aes(x = long, y = lat, map_id = region),
  fill = "#ffffff",
  color = "#ffffff",
  size = 0.20
  )
```

```
## Warning: Ignoring unknown aesthetics: x, y
```

```r
  gg <- gg + geom_map(
  data = iu,
  map = world,
  aes(fill = `INTERNET USERS`, map_id = region),
  color = "#ffffff",
  size = 0.15
  )

  gg <- gg + scale_fill_continuous(low = 'thistle2', high = 'darkblue',
  guide = 'colorbar')
  gg
```



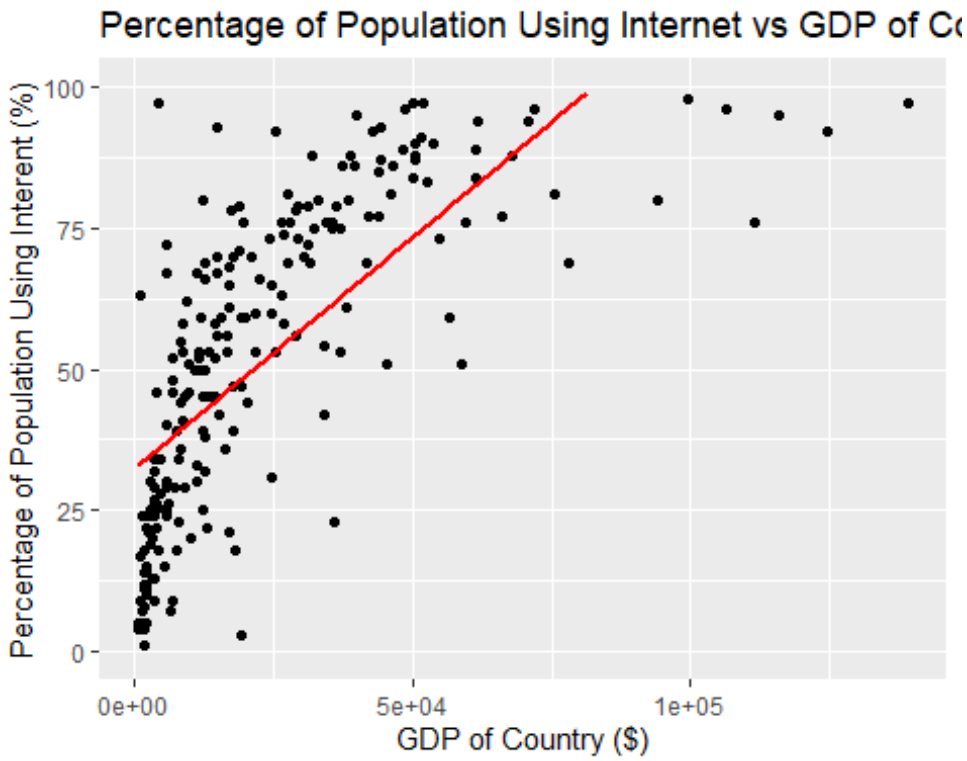
CALCULATE THE PERCENTAGE OF PEOPLE WHO ARE INTERNET USERS

```
internet_population <- inner_join(internetusers_cia2017, population_cia2017,
by = "Country")
internet_population <- internet_population %>%
  mutate(percentage = round((`INTERNET USERS`/POPULATION)*100))
internet_population$Rank.y <- NULL
internet_population$`Date of Information.x` <- NULL
internet_population$`Date of Information.y` <- NULL
```

IS THERE A RELATIONSHIP BETWEEN GDP AND PERCENTAGE OF PEOPLE WHO USE INTERNET?

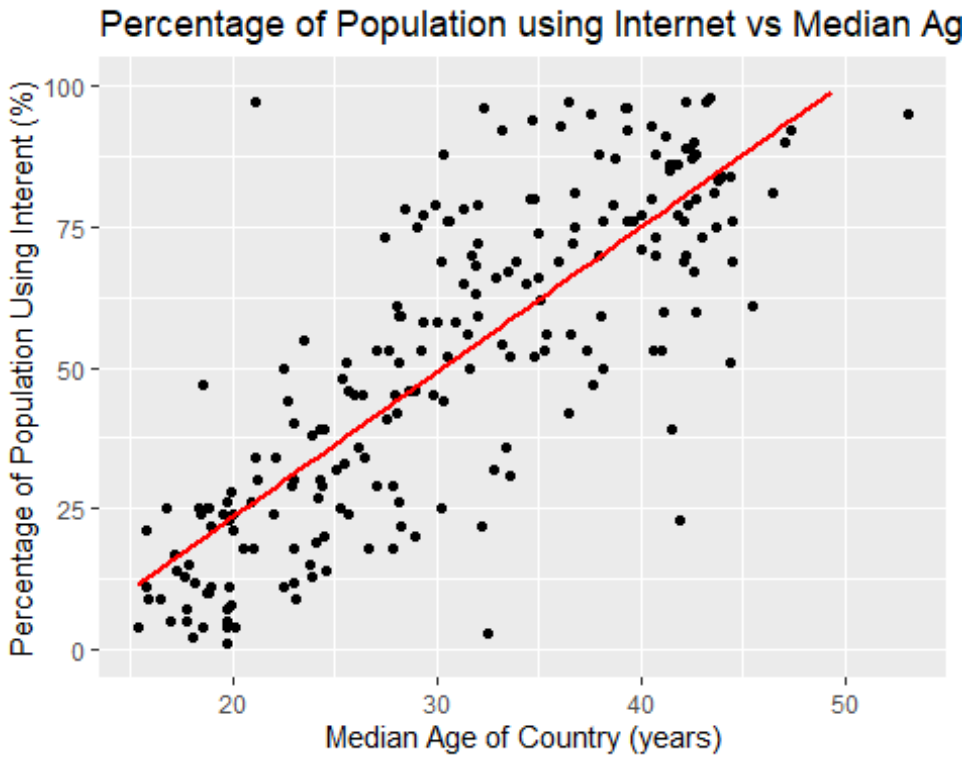
```
gdp_internetuser <- inner_join(gdpppp_cia2017,internet_population, by =
"Country")
gdp_internetuser$Rank.x<- NULL
gdp_internetuser$`Date of Information`<- NULL

gdp_internetuser <- gdp_internetuser %>%
  mutate (money = parse_number(`GDP - PER CAPITA (PPP)`))
gdp_internetuser$`GDP - PER CAPITA (PPP)`<- NULL

ggplot (data = gdp_internetuser, aes(x = money, y = percentage)) +
geom_point() + stat_smooth(method = "lm", colour = "Red", se = FALSE) +
ylim(0,100) + xlab("GDP of Country ($)") + ylab("Percentage of Population
Using Interent (%)") + labs(title = "Percentage of Population Using Internet
vs GDP of Country")
```

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).

## Warning: Removed 1 rows containing missing values (geom_point).

## Warning: Removed 33 rows containing missing values (geom_smooth).
```



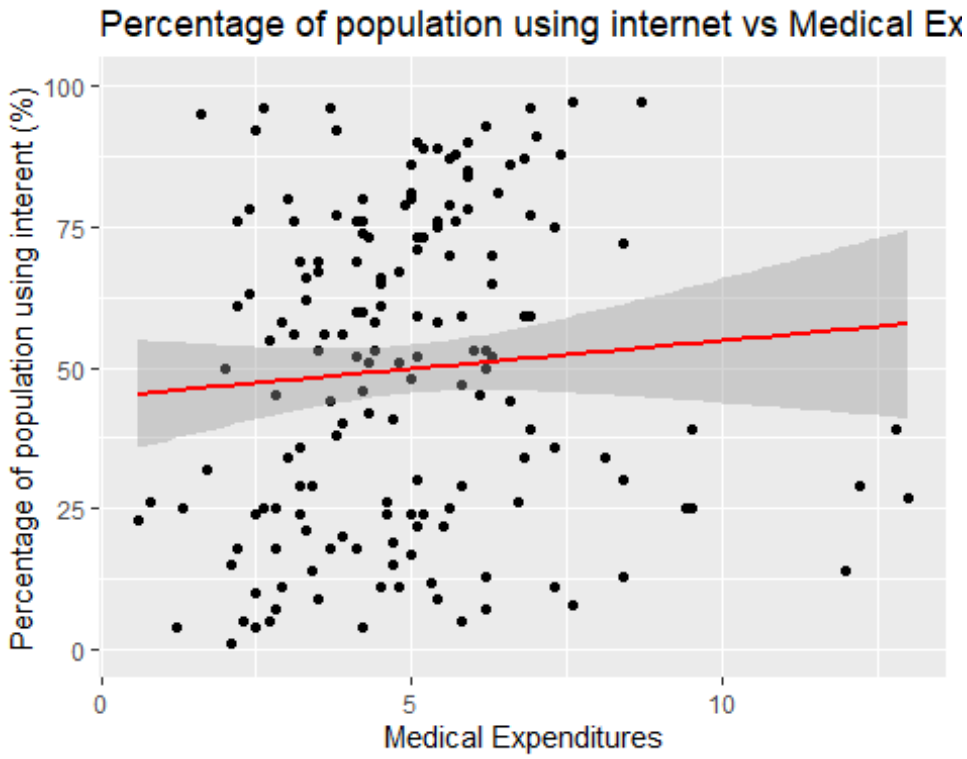
IS THERE A RELATIONSHIP BETWEEN MEDIAN AGE OF COUNTRY AND PERCENTAGE OF PEOPLE WHO USE INTERNET?

```
medianage_percentage <- inner_join(medianage_cia2017,internet_population, by
= "Country")
medianage_percentage$`Date of Information`<- NULL
medianage_percentage$Rank.x<- NULL

ggplot (data = medianage_percentage, aes(x = `MEDIAN AGE`, y = percentage)) +
geom_point() + stat_smooth(method = "lm", colour = "Red", se = FALSE) +
ylim(0,100) + xlab("Median Age of Country (years)") + ylab("Percentage of
Population Using Interent (%)") + labs(title = "Percentage of Population
using Internet vs Median Age of Country")

## Warning: Removed 2 rows containing non-finite values (stat_smooth).

## Warning: Removed 2 rows containing missing values (geom_point).

## Warning: Removed 8 rows containing missing values (geom_smooth).
```

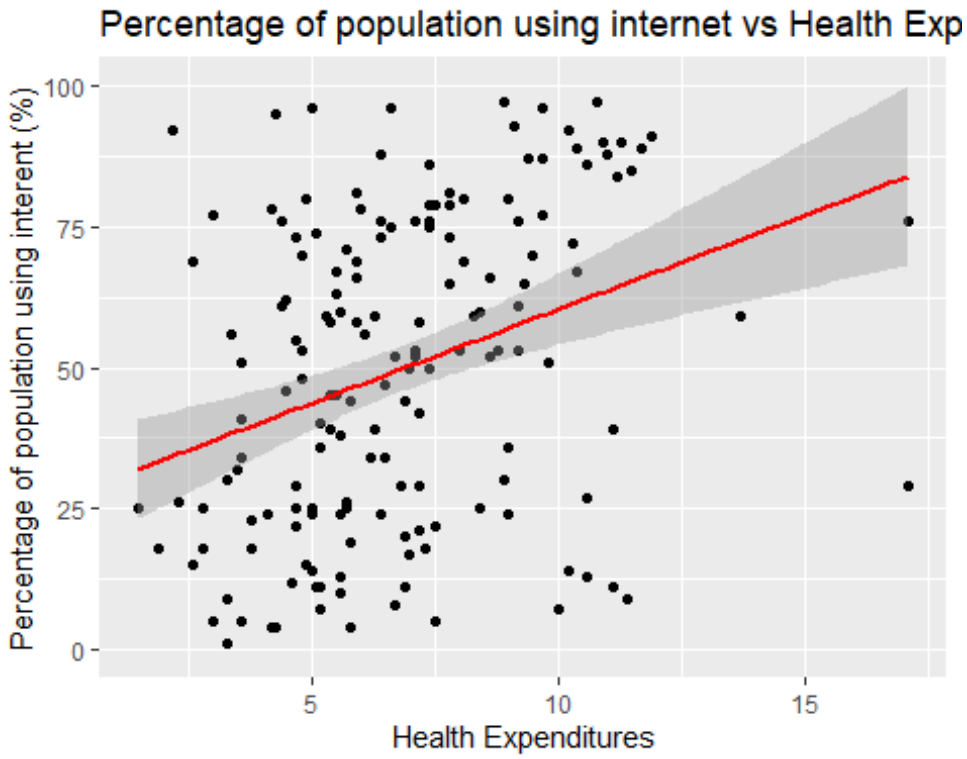


COMBINE HEALTH AND EDUCATION EXPENDITURES. ANALYZE THE DATA BY COMPARING IT WITH PERCENTAGE. (No relationship found)

```
education_health <- inner_join(education_cia2017, healthexpend_cia2017, by =
"Country")

education_health$Rank.y <- NULL
education_health$`Date of Information.x` <- NULL
education_health$`Date of Information.y` <- NULL
education_health <- education_health %>%
  mutate(`Total expenditure` = `(% OF GDP).x` + `(% OF GDP).y`)
expenditure_percentage <- inner_join(education_health, internet_population,
by = "Country")
expenditure_percentage$Rank.x.y <- NULL

ggplot(data = expenditure_percentage, aes(x = `(% OF GDP).x`, y =
percentage)) + geom_point() + stat_smooth(method = "lm", colour = "Red", se =
TRUE) + ylim(0,100) + xlab("Medical Expenditures") + ylab("Percentage of
population using interent (%)") + labs(title = "Percentage of population
using internet vs Medical Expenditures")
```



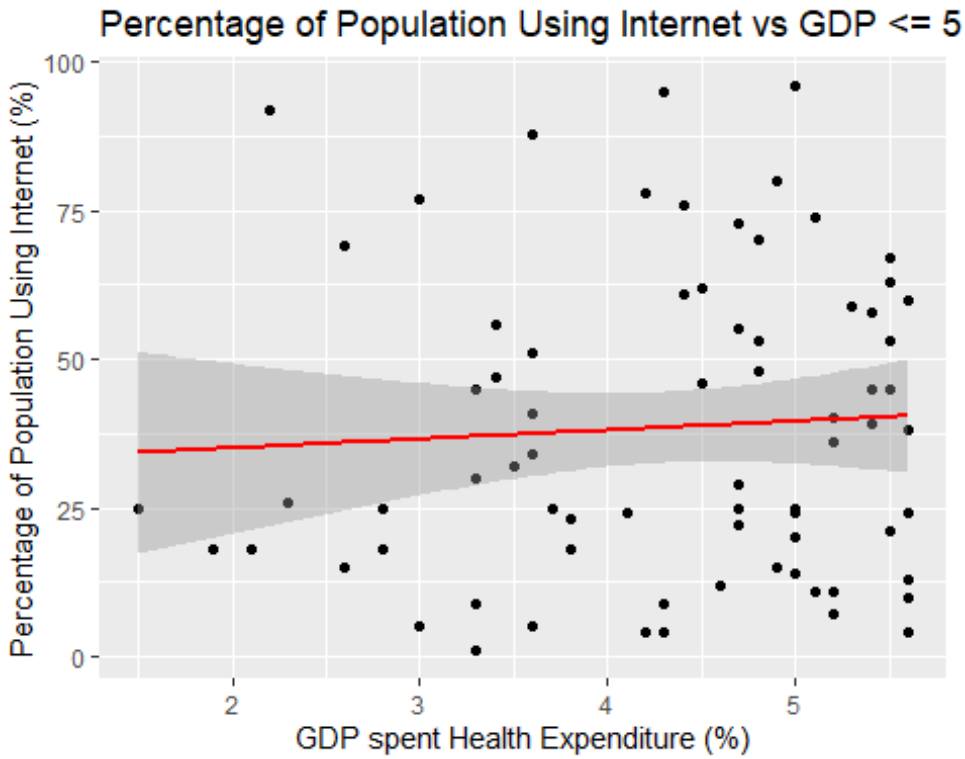

```
ggplot(data = expenditure_percentage, aes(x = `(% OF GDP).y`, y =
percentage)) + geom_point() + stat_smooth(method = "lm", colour = "Red", se =
TRUE) + ylim(0,100) + xlab("Health Expenditures") + ylab("Percentage of
population using interent (%)") + labs(title = "Percentage of population
using internet vs Health Expenditures")
```



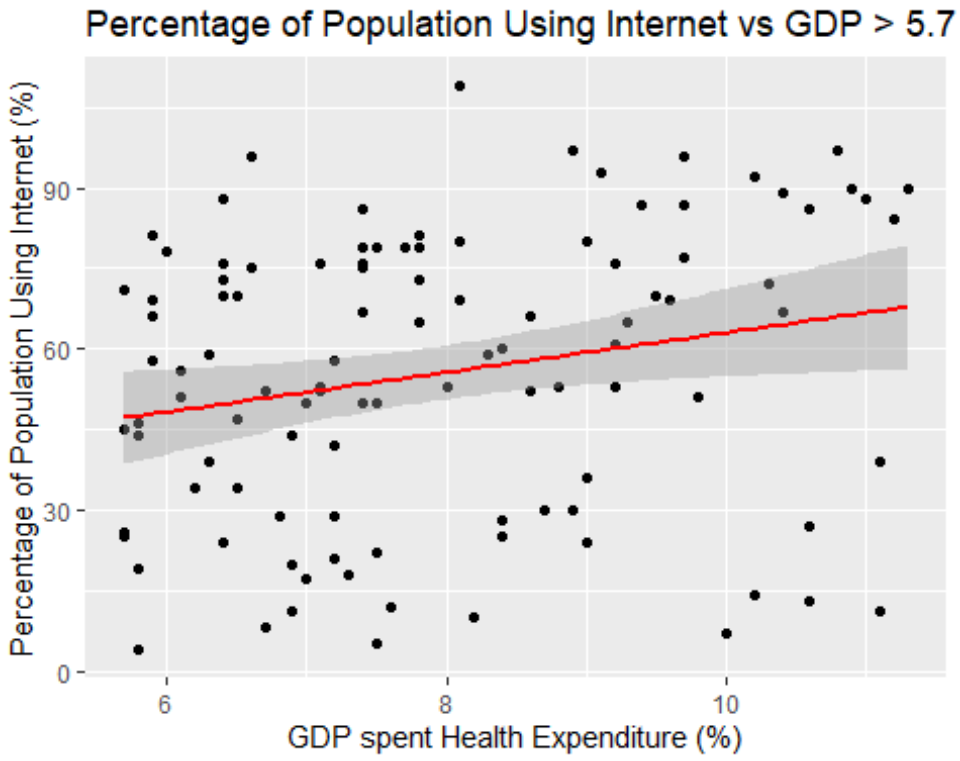
HEALTH SUBSETS

```
health_percentage <- inner_join(healthexpend_cia2017, internet_population, by
= "Country")

health_percentage <- health_percentage %>%
  mutate(group = ifelse(`(% OF GDP)` < 17.1/3, '1',ifelse(`(% OF GDP)` >=
17.1/3 & `(% OF GDP)` < (17.1*2)/3, '2','3')))

health_percentage <- health_percentage %>%
  group_by(group) %>%
  select (`(% OF GDP)`, Country, group, percentage)

health_gdp_1 <- health_percentage %>%
  filter(group == '1') %>%
  select (`(% OF GDP)`, group, percentage)

health_gdp_2 <- health_percentage %>%
  filter(group == '2') %>%
  select (`(% OF GDP)`, group, percentage)

health_gdp_3 <- health_percentage %>%
  filter(group == '3') %>%
  select (`(% OF GDP)`, group, percentage)


ggplot(data = health_gdp_1, aes(x = `(% OF GDP)`, y = percentage)) +
geom_point() + stat_smooth(method = "lm", colour = "Red") + labs(x = "GDP
```

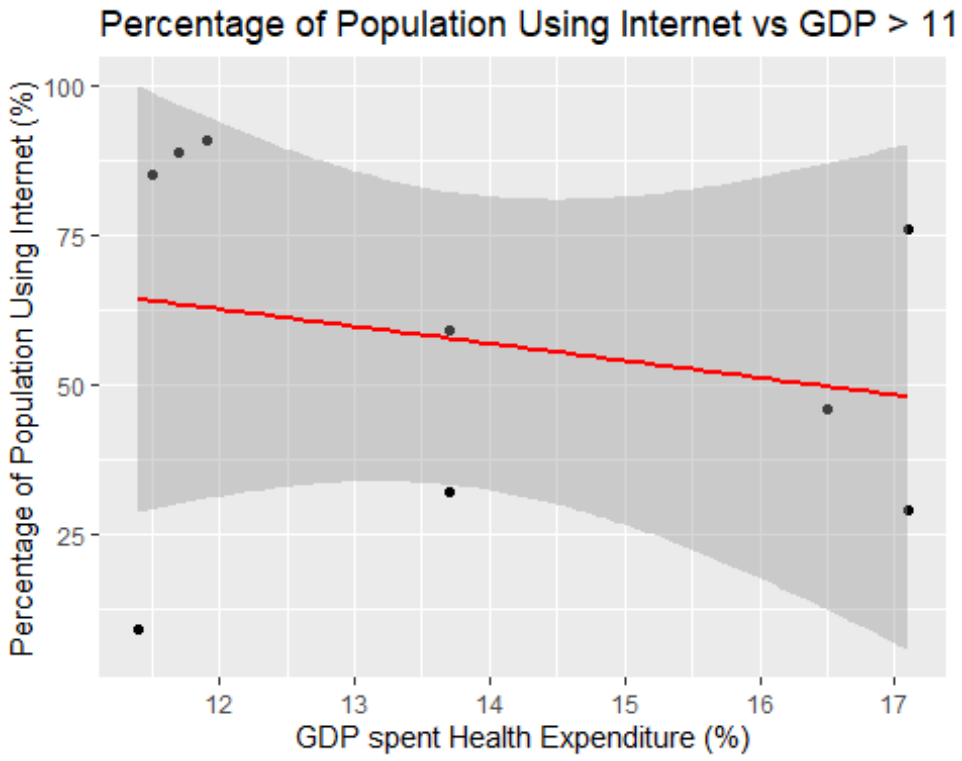
```
spent Health Expenditure (%)", y = "Percentage of Population Using Internet
(%)", title = "Percentage of Population Using Internet vs GDP <= 5.7")
```




```
ggplot(data = health_gdp_2, aes(x = `(% OF GDP)`, y = percentage)) +
geom_point() + stat_smooth(method = "lm", colour = "Red")+ labs(x = "GDP
spent Health Expenditure (%)", y = "Percentage of Population Using Internet
(%)", title = "Percentage of Population Using Internet vs GDP > 5.7 and <=
11.4")
```

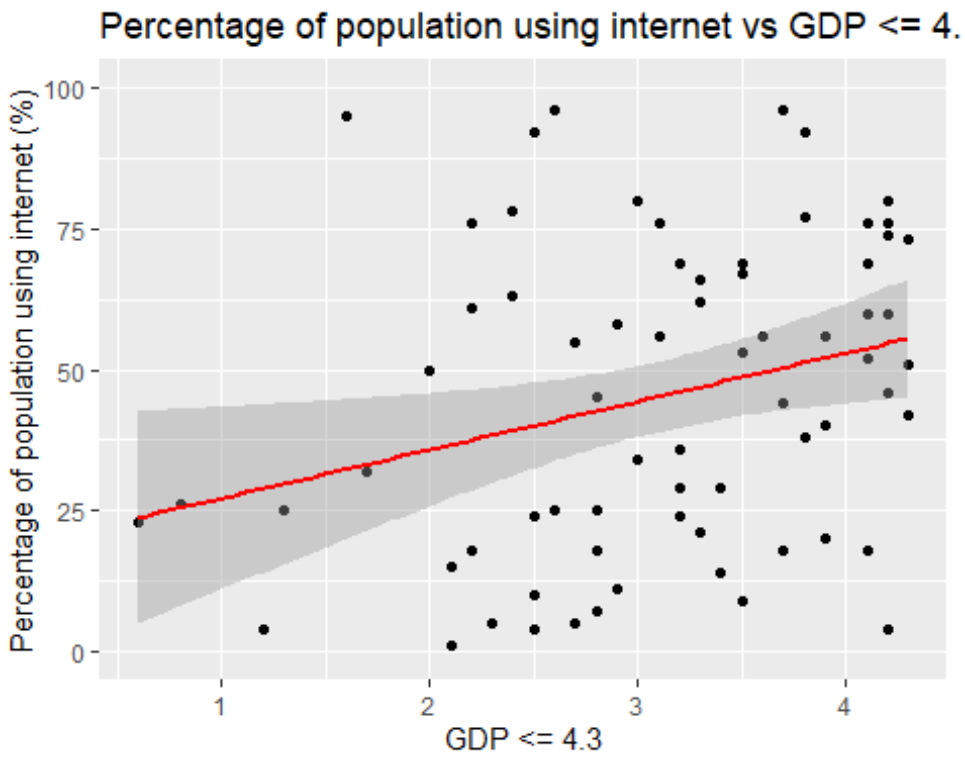



```
ggplot(data = health_gdp_3, aes(x = `(% OF GDP)`, y = percentage)) +
geom_point() + stat_smooth(method = "lm", colour = "Red")+ labs(x = "GDP
spent Health Expenditure (%)", y = "Percentage of Population Using Internet
(%)", title = "Percentage of Population Using Internet vs GDP > 11.4")
```

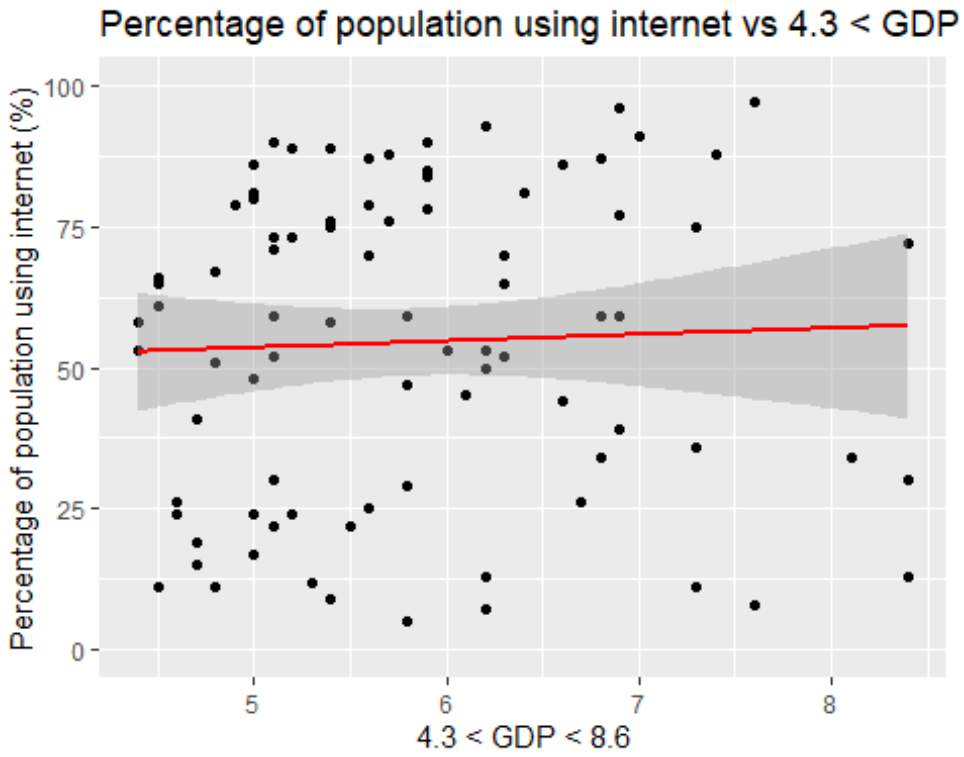


EDUCATION SUBSETS

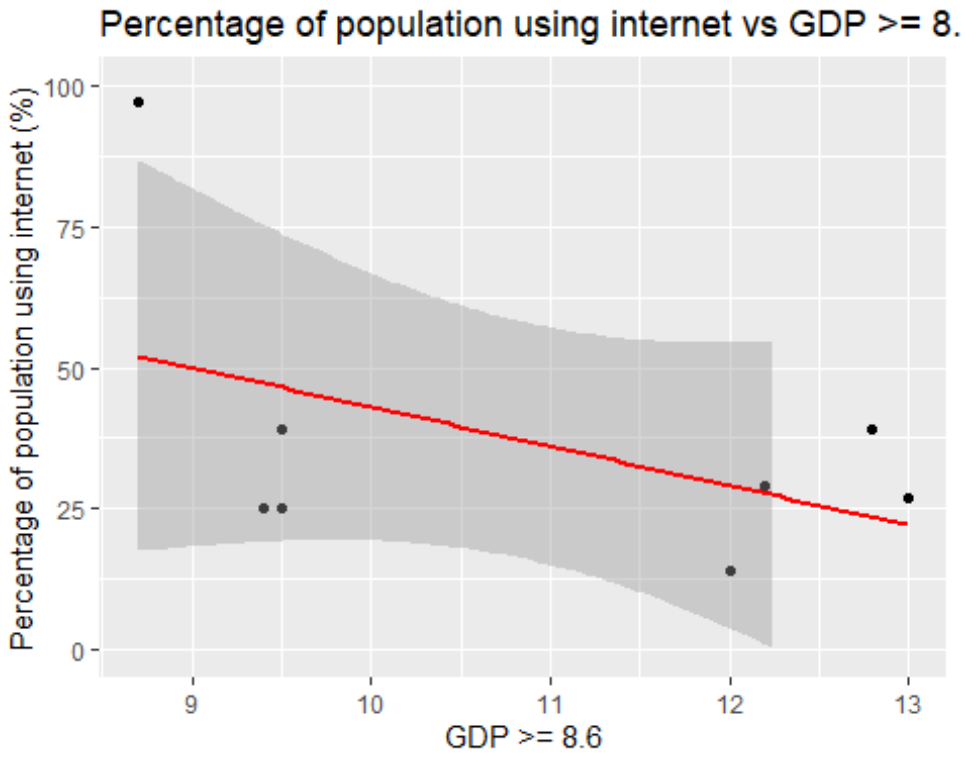
```
expenditure_percentage <- expenditure_percentage %>%
  mutate(group = ifelse(`(% OF GDP).x` <= 4.3, '1',ifelse(`(% OF GDP).x` >
4.3 & `(% OF GDP).x` < 8.6, '2', '3')))

education_gdp_1 <- expenditure_percentage %>%
  filter(group == '1') %>%
  select (`(% OF GDP).x`, group, percentage)
ggplot(data = education_gdp_1, aes(x = `(% OF GDP).x`, y = percentage)) +
geom_point() + stat_smooth(method = "lm", colour = "Red") + ylim(0,100) +
xlab("GDP <= 4.3") + ylab("Percentage of population using internet (%)") +
labs(title = "Percentage of population using internet vs GDP <= 4.3")
```




```
education_gdp_2 <- expenditure_percentage %>%
  filter(group == '2') %>%
  select (`(% OF GDP).x`, group, percentage)
ggplot(data = education_gdp_2, aes(x = `(% OF GDP).x`, y = percentage)) +
geom_point() + stat_smooth(method = "lm", colour = "Red") + ylim(0,100) +
xlab("4.3 < GDP < 8.6") + ylab("Percentage of population using internet (%)")
+ labs(title = "Percentage of population using internet vs 4.3 < GDP < 8.6")
```

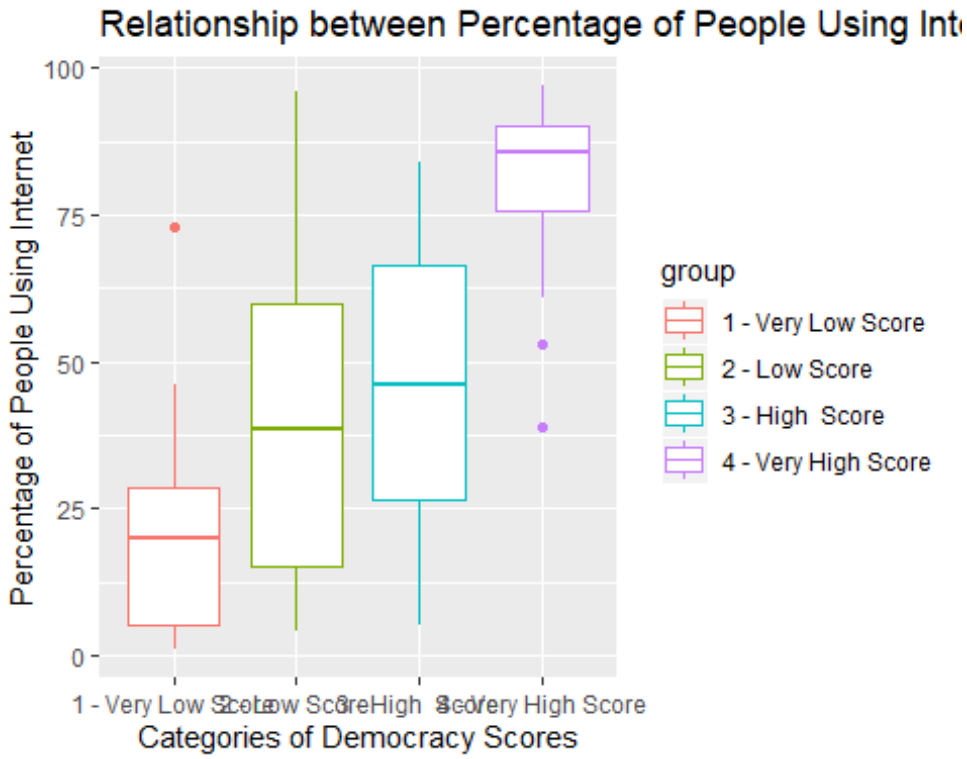



```
education_gdp_3 <- expenditure_percentage %>%
  filter(group == '3') %>%
  select (`(% OF GDP).x`, group, percentage)
ggplot(data = education_gdp_3, aes(x = `(% OF GDP).x`, y = percentage)) +
geom_point() + stat_smooth(method = "lm", colour = "Red") + ylim(0,100) +
xlab("GDP >= 8.6") + ylab("Percentage of population using internet (%)") +
labs(title = "Percentage of population using internet vs GDP >= 8.6")
```



Create a boxplot by linking different groups of democracy scores to percentage.

```
democracy_percentage <- inner_join(democracyindex2017, internet_population,
by = "Country")
democracy_percentage <- democracy_percentage %>%
  mutate (`Democracy Score` = parse_number(Score))

democracy_percentage <- democracy_percentage %>%
  mutate(group = ifelse(`Democracy Score` < 2.5, '1 - Very Low
Score',ifelse(`Democracy Score` >= 2.5 & `Democracy Score` < 5, '2 - Low
Score', ifelse(`Democracy Score` >= 5 & `Democracy Score` < 7.5, '3 - High
Score', '4 - Very High Score'))))

democracy_percentage <- democracy_percentage %>%
  group_by(group) %>%
  select (`Democracy Score`, Country, group)

# Combine democracy_percentage with internet_population and create a box plot
democracy_whisker <- inner_join(democracy_percentage, internet_population, by
= "Country")

ggplot(data = democracy_whisker, aes(x = group, y = percentage, color =
group)) + geom_boxplot() + labs(x = "Categories of Democracy Scores", y =
"Percentage of People Using Internet", title = "Relationship between
Percentage of People Using Internet and Categories of Democracy Scores")
```

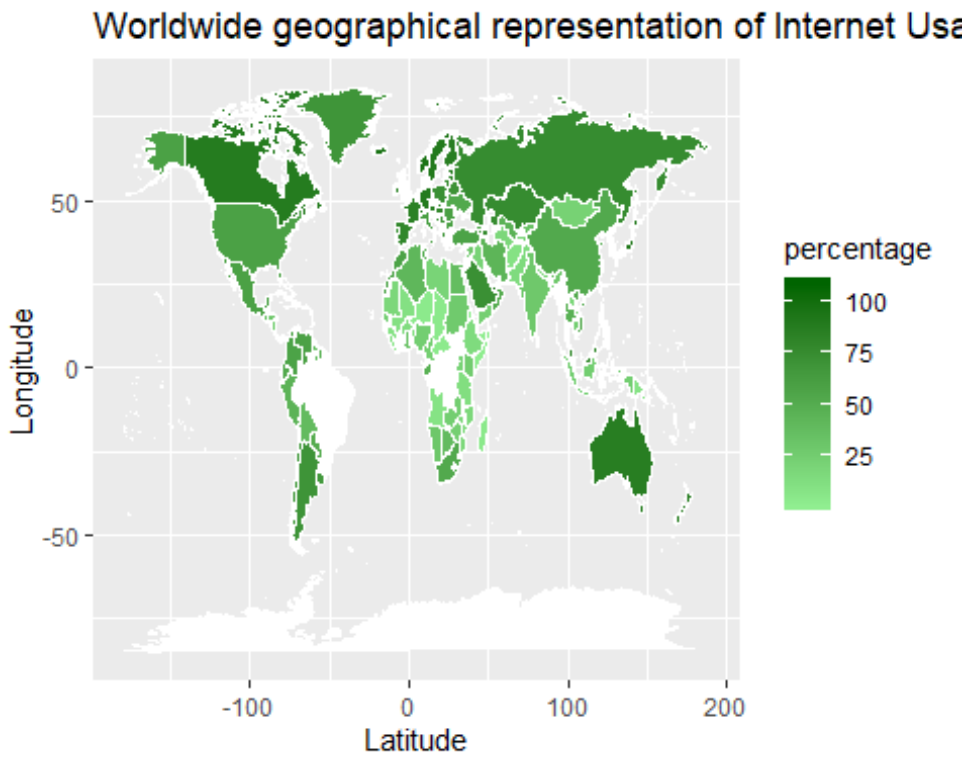


GEOGRAPHICAL MAPPING OF INTERNET USAGE

```
globe <- map_data("world")

#internetusers_cia2017 <- read_csv("internetusers_cia2017.csv")

in_per <- internet_population %>% rename(region = Country)

in_per$region[4] <- "USA" # to match world map data

in_per <- semi_join(in_per, globe, by = "region") #only keep countries
according to world map data

# code below is modified from
# https://stackoverflow.com/questions/29614972/ggplot-us-state-map-colors-
are-fine-polygons-jagged-r
gg <- ggplot()

gg <- gg + geom_map(
  data = globe,
  map = globe,
  aes(x = long, y = lat, map_id = region),
  fill = "#ffffff",
  color = "#ffffff",
  size = 0.20
  )
```

```
## Warning: Ignoring unknown aesthetics: x, y
```

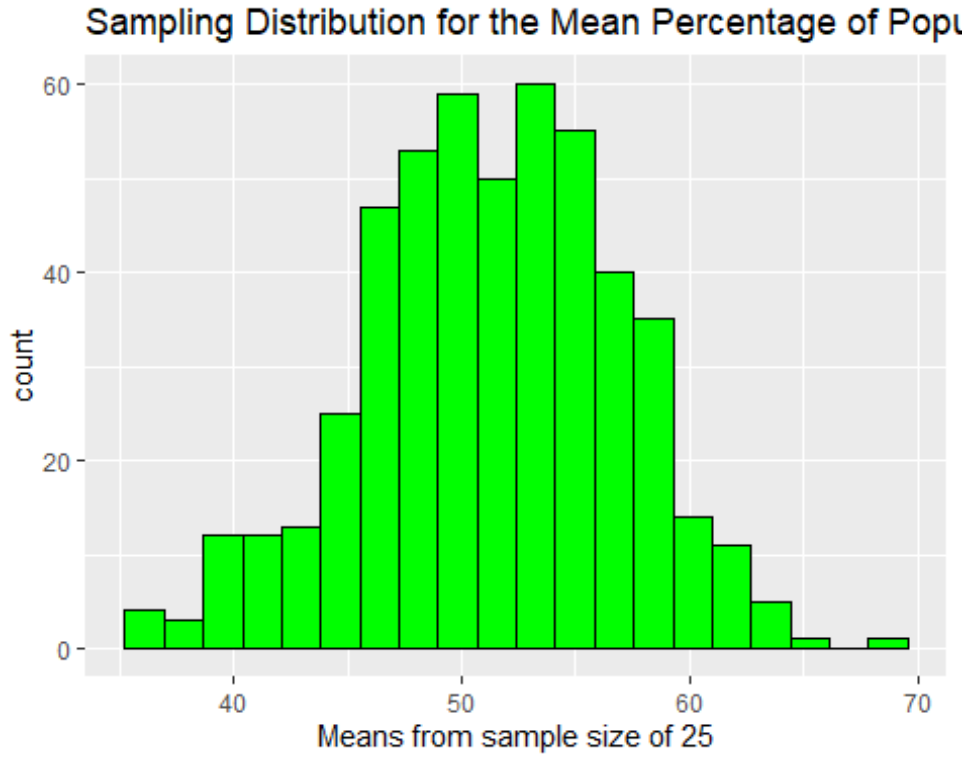
```
  gg <- gg + geom_map(
  data = in_per,
  map = globe,
  aes(fill = percentage, map_id = region),
  color = "#ffffff",
  size = 0.15
  )

  gg <- gg + scale_fill_continuous(low = 'lightgreen', high = 'darkgreen',
  guide = 'colorbar') + labs(x = "Latitude", y = "Longitude", title =
"Worldwide geographical representation of Internet Usage")
  gg
```



Percentile bootstrap method on percentage of internet users Note: 500 Replciations have been used instead of 5000 to reduce run time

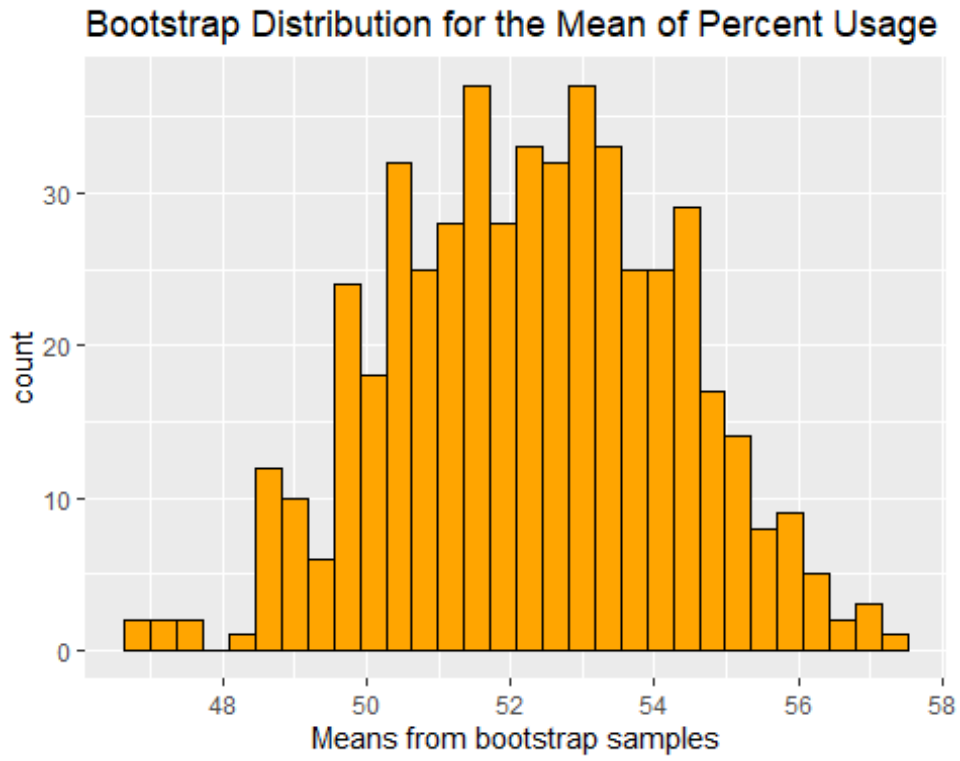
```
PER <- internet_population %>% filter(is.na(percentage) == FALSE &
percentage<= 100)

# Take a 25% sample of the original data
sample25 <- PER %>% sample_n(size = 25)

sample_means <- rep(NA, 500)
for (i in 1:500){
  sample25 <- PER %>% sample_n(size = 25)
```

```
  sample_means[i] <- as.numeric(sample25 %>%
                                  summarize(mean(percentage)))
}
sample_means <- data_frame(mean_percentage = sample_means)

## Warning: `data_frame()` is deprecated, use `tibble()`.
## This warning is displayed once per session.

ggplot(data = sample_means, aes(x = mean_percentage)) + geom_histogram(bins=
20, colour = "Black", fill = "Green") + labs(x = "Means from sample size of
25", title = "Sampling Distribution for the Mean Percentage of Population
Using Internet")
```

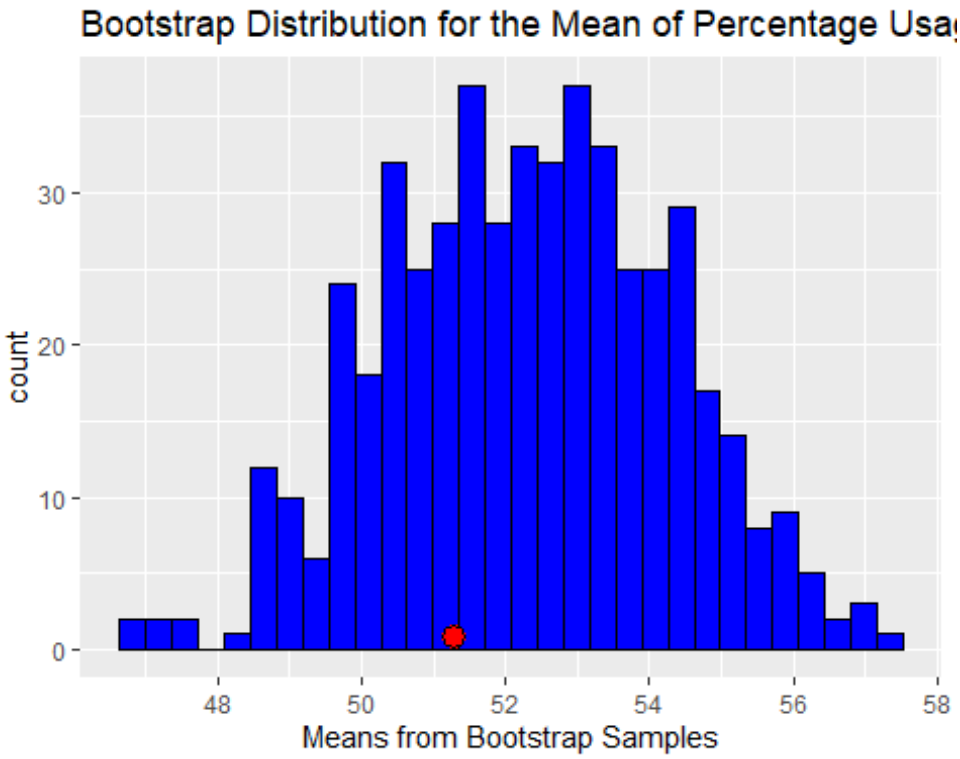



```
observed_data <- PER %>% sample_n(size = 200, replace = FALSE)
observed_mean <- as.numeric(observed_data %>%
                              summarize (mean(percentage)))

boot_samp <- observed_data %>% sample_n(size = 200, replace = TRUE)
boot_samp %>% summarize(mean_percentage = mean(percentage))

## # A tibble: 1 x 1
##   mean_percentage
##             <dbl>
## 1            49.3

boot_means <- rep(NA, 500)
for (i in 1:500){
  boot_samp <- observed_data %>% sample_n(size = 200, replace = TRUE)
```

```
  boot_means[i] <- as.numeric(boot_samp %>%
                                summarize(mean_percentage =
mean(percentage)))
}
boot_means <- data_frame(mean_percentage = boot_means)

ggplot(boot_means, aes(x = mean_percentage)) + geom_histogram(bins = 30,
colour = "Black", fill = "Orange") + labs(x = "Means from bootstrap samples",
title = "Bootstrap Distribution for the Mean of Percent Usage")
```




```
population_mean <- PER %>%
  summarize(population_mean_percentage = mean(percentage))
population_mean

## # A tibble: 1 x 1
##   population_mean_percentage
##                        <dbl>
## 1                       51.3

ggplot(boot_means, aes(x = mean_percentage)) + geom_histogram(bins = 30,
colour = "Black", fill = "Blue") + geom_dotplot(data = population_mean, aes(x
= population_mean_percentage), fill = "Red") + labs(x = "Means from Bootstrap
Samples", title = "Bootstrap Distribution for the Mean of Percentage Usage")

## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```




```r
# Generate 100 bootstrap samples
n_interval <- 100
perc_25 <- rep(NA, n_interval)
perc_975 <- rep(NA, n_interval)
sample_size <- 200
replications <- 500
for (i in 1:n_interval){
  observed_data <- PER %>%
    sample_n(size = sample_size, replace = FALSE)

  boot_means <- rep(NA, replications)
  for (j in 1:replications){
    boot_samp <- observed_data %>%
      sample_n(size = sample_size, replace = TRUE)
    boot_means[j] <- as.numeric(boot_samp %>%
                                  summarize(mean(percentage)))
  }

  perc_25[i] <- quantile(boot_means, 0.025)
  perc_975[i] <- quantile(boot_means, 0.975)

  print(c(i, perc_25[i], perc_975[i]))
}
```

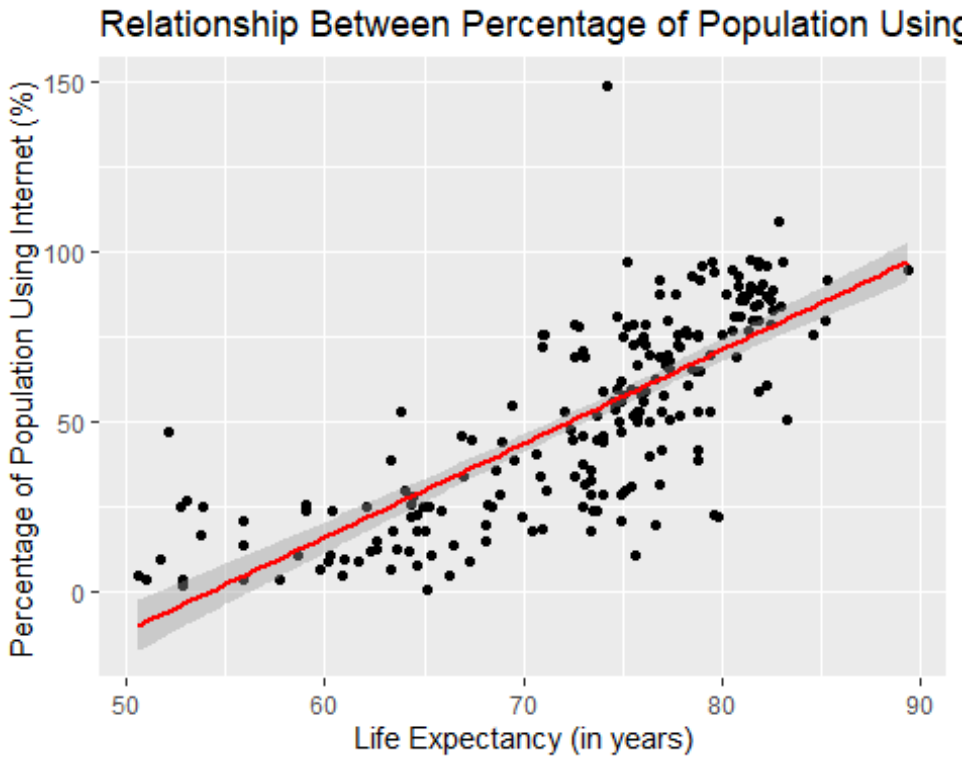
```
## [1]  1.00000 48.65975 56.03387
## [1]  2.00000 46.82425 55.03087
```

```
## [1]  3.00000 46.97825 54.16088
## [1]  4.00000 47.41875 55.24525
## [1]  5.0000 46.5450 54.0465
## [1]  6.00000 47.09875 54.43650
## [1]  7.00000 46.63200 54.32263
## [1]  8.0000 46.4870 54.4815
## [1]  9.00000 46.74688 54.58775
## [1] 10.00000 48.55938 56.34262
## [1] 11.00000 46.90950 54.34262
## [1] 12.00000 47.90825 55.18337
## [1] 13.00000 46.30438 54.38575
## [1] 14.00000 47.40237 55.19862
## [1] 15.00000 46.94237 54.77813
## [1] 16.00000 47.51400 55.07913
## [1] 17.00000 48.59238 56.28575
## [1] 18.00000 47.56687 56.15200
## [1] 19.00000 46.53975 53.98438
## [1] 20.00000 47.96688 55.45387
## [1] 21.0000 47.5985 55.2905
## [1] 22.00000 47.07537 54.85075
## [1] 23.00000 47.57725 55.20275
## [1] 24.00000 48.28713 55.96362
## [1] 25.00000 48.16137 56.24487
## [1] 26.00000 48.01825 55.96050
## [1] 27.00000 47.68275 55.02387
## [1] 28.00000 45.85600 53.89525
## [1] 29.00000 47.45238 55.01500
## [1] 30.00000 47.10350 54.39388
## [1] 31.00000 47.20338 54.63100
## [1] 32.00000 46.64475 54.39287
## [1] 33.00000 48.17850 55.60938
## [1] 34.00000 47.39113 55.02287
## [1] 35.00000 45.63612 53.67788
## [1] 36.00000 47.62425 55.51812
## [1] 37.00000 47.39462 54.53838
## [1] 38.00000 46.98650 54.63687
## [1] 39.00000 47.85750 55.63275
## [1] 40.0000 47.9785 55.9660
## [1] 41.00000 47.06475 55.11025
## [1] 42.00000 48.13137 55.89350
## [1] 43.00000 49.55088 56.85050
## [1] 44.00000 46.98450 54.81525
## [1] 45.00000 48.05738 56.16813
## [1] 46.00000 47.35725 54.84712
## [1] 47.00000 47.42750 55.60437
## [1] 48.00000 47.63825 55.51013
## [1] 49.00000 46.65713 54.29263
## [1] 50.00000 47.36100 55.43025
## [1] 51.00000 47.67975 55.42812
## [1] 52.00000 47.68462 55.22837
```

```
## [1] 53.00000 48.03612 55.61650
## [1] 54.0000 47.2495 55.2250
## [1] 55.00000 47.98213 55.68525
## [1] 56.00000 47.39725 54.77387
## [1] 57.00000 47.05663 54.40525
## [1] 58.00000 46.96213 54.47050
## [1] 59.00000 47.66950 55.13625
## [1] 60.00000 46.86662 54.87000
## [1] 61.00000 46.74237 54.57088
## [1] 62.00000 47.66900 55.78812
## [1] 63.00000 47.56737 55.39363
## [1] 64.00000 47.26687 55.02375
## [1] 65.00000 46.57000 53.88863
## [1] 66.0000 47.0745 55.0005
## [1] 67.00000 48.31162 55.89812
## [1] 68.00000 46.54138 53.94288
## [1] 69.00000 47.43712 55.16225
## [1] 70.00000 47.42450 54.97737
## [1] 71.00000 47.69550 55.84375
## [1] 72.00000 47.54762 55.23637
## [1] 73.00000 47.07400 54.56012
## [1] 74.00000 47.11463 54.66287
## [1] 75.00000 47.87162 55.62262
## [1] 76.00000 46.96425 55.28575
## [1] 77.00000 47.39475 55.13125
## [1] 78.00000 47.85712 55.76787
## [1] 79.00000 47.71688 55.40275
## [1] 80.00000 47.54725 54.96863
## [1] 81.00000 46.28775 54.27837
## [1] 82.000 46.814 54.176
## [1] 83.00000 47.55588 55.32362
## [1] 84.00000 47.63325 55.21263
## [1] 85.00000 46.05237 54.08263
## [1] 86.00000 47.18925 54.70763
## [1] 87.00000 47.62687 55.31387
## [1] 88.00000 48.41737 56.52263
## [1] 89.00000 46.95675 55.02625
## [1] 90.00000 47.48450 55.63587
## [1] 91.0000 47.5445 55.8875
## [1] 92.00000 47.45475 54.94675
## [1] 93.00000 47.06387 54.77950
## [1] 94.00000 48.41037 56.00562
## [1] 95.00000 47.88500 55.60787
## [1] 96.00000 47.30213 54.93937
## [1] 97.00000 46.50238 54.32337
## [1] 98.00000 48.20188 55.71125
## [1] 99.00000 47.96000 55.58637
## [1] 100.00000  47.43475  55.12525
```

Relationship between life expectency and percentage

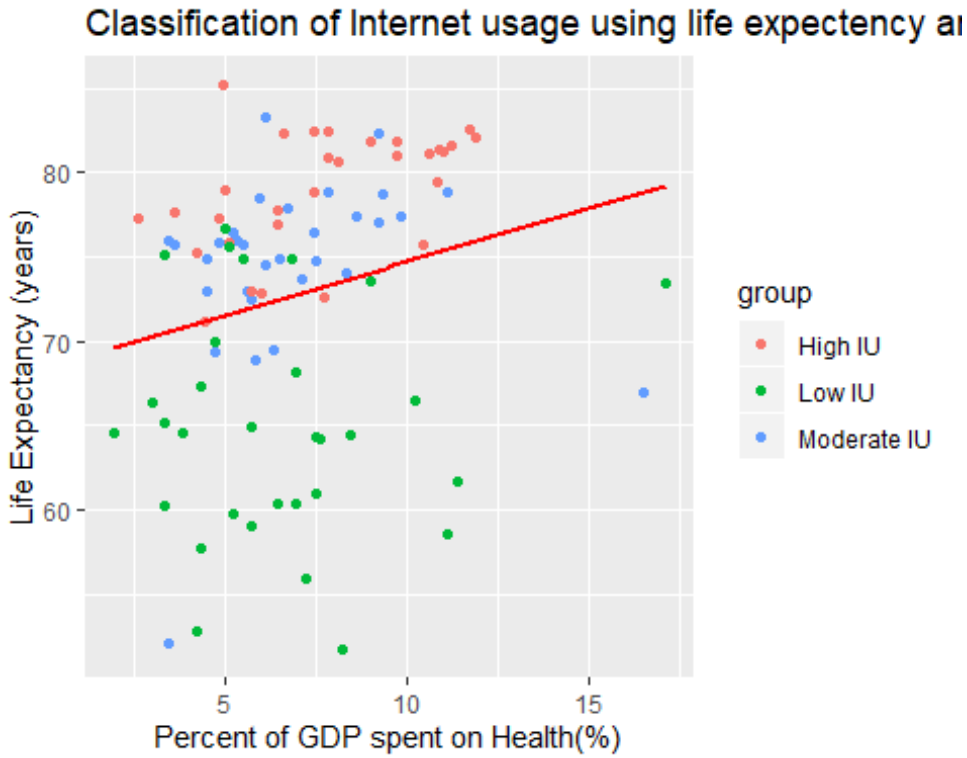
```r
life_percent <- inner_join(lifeexpect_cia2017, internet_population, by =
"Country")

ggplot(data = life_percent, aes(x = `(YEARS)`, y = percentage)) +
geom_point() + stat_smooth(method = "lm", colour = "Red", se = TRUE) + labs(x
= "Life Expectancy (in years)", y = "Percentage of Population Using Internet
(%)", title = "Relationship Between Percentage of Population Using Internet
and Life Expectancy")
```



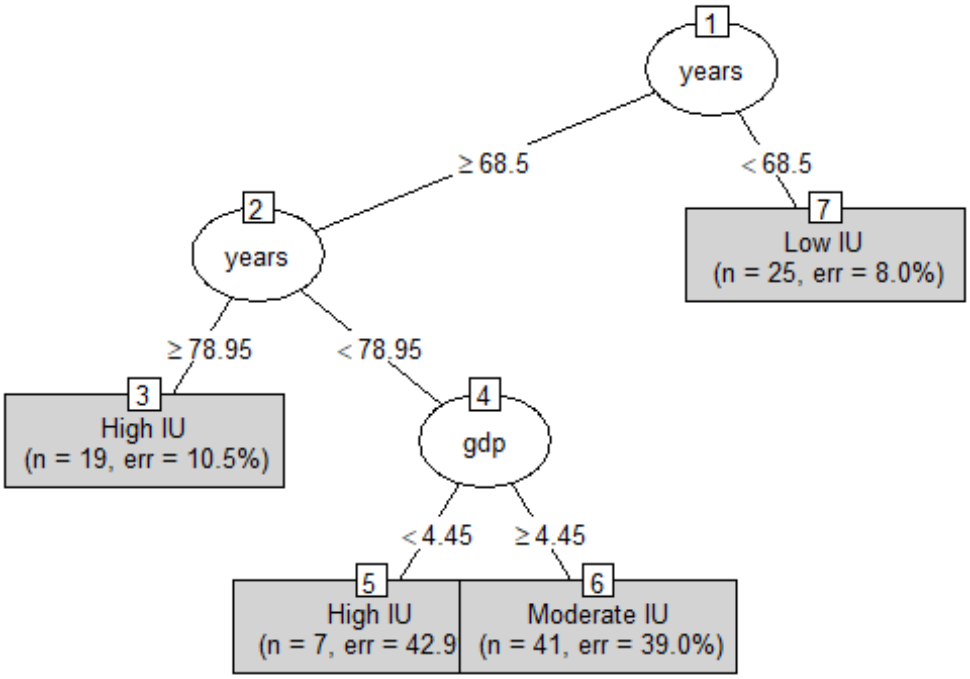
CLASSIFICATION TREE ON LIFE EXPECTENCY AND HEALTH EXPENDITURE

```r
set.seed(2150)
tree_df <- inner_join(lifeexpect_cia2017, healthexpend_cia2017, by =
"Country")
tree_df <- inner_join(tree_df, internet_population, by = "Country")

tree_df$`Date of Information.x` <- NULL
tree_df$`Date of Information.y` <- NULL
tree_df <- tree_df %>%
  rename(years = `(YEARS)`, gdp = `(% OF GDP)`) %>%
  mutate(group = ifelse(percentage <=33, 'Low IU',ifelse(percentage > 33 &
percentage <= 66, "Moderate IU", "High IU")))

tree_df <- tree_df %>% sample_n(size = 92)
```

```r
ggplot(data = tree_df, aes(x = gdp, y = years, colour = group)) +
geom_point() + stat_smooth(method = "lm", colour = "Red", se = FALSE) +
labs(x = "Percent of GDP spent on Health(%)", y = "Life Expectancy (years)",
title = "Classification of Internet usage using life expectency and Health
Expenditure")
```

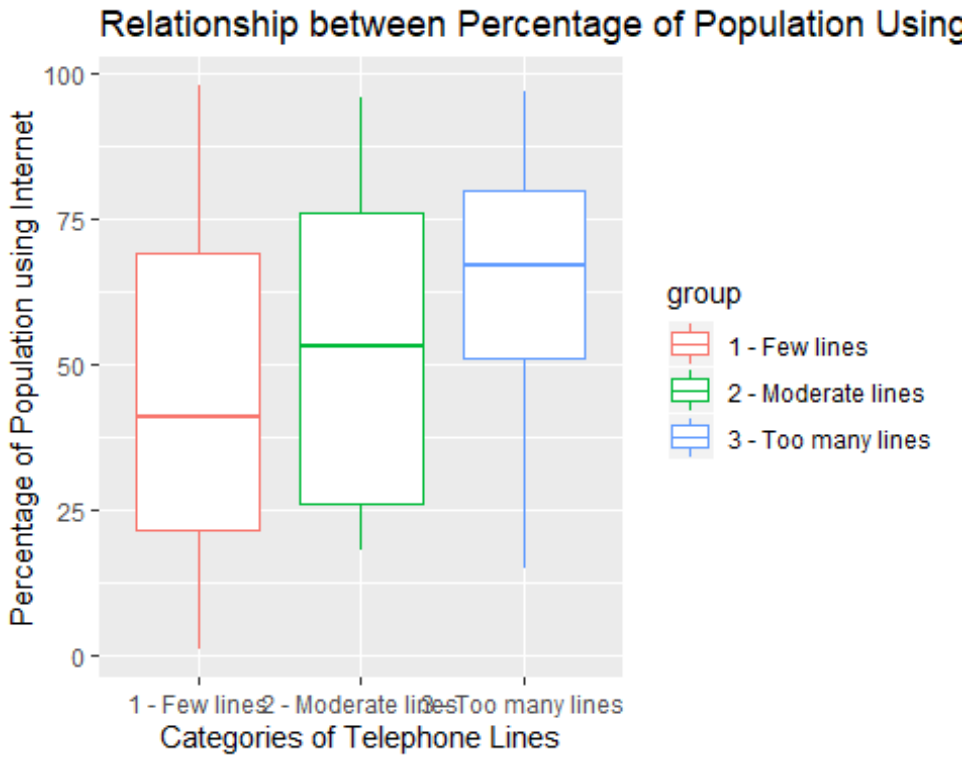



```r
# Create a classification tree
tree <- rpart(group ~ gdp + years, data = tree_df)
plot(as.party(tree), type = "simple", gp = gpar(cex = 0.8))
```



Relationship between telephone lines and percentage

```
tele_percentage <- inner_join(telephonelines_cia2017, internet_population,
by="Country")
tele_percentage <- tele_percentage %>%
  rename(lines = `TELEPHONES - MAIN LINES IN USE`)

#ggplot(data = tele_percentage, aes(x = lines, y = percentage)) +
geom_point() + stat_smooth(method = #"lm", colour = "Red", se = TRUE) +
xlim(0,1500000) + ylim(0,100)

tele_percentage <- tele_percentage %>%
  mutate(group = ifelse(lines <= 500000, "1 - Few lines", ifelse(lines >
500000 & lines <= 1000000, "2 - Moderate lines", "3 - Too many lines"))) %>%
  filter (percentage <= 100) %>%
  select(percentage, group, lines)

ggplot(data = tele_percentage, aes(x = group, y = percentage, color = group))
+ geom_boxplot() + labs(x = "Categories of Telephone Lines", y = "Percentage
of Population using Internet", title = "Relationship between Percentage of
Population Using Internet and Telephone Lines")
```



Note: According to a statistics article published by STATISTA (https://www.statista.com/statistics/266587/percentage-of-internet-users-by-age-groups-in-the-us/), most

Link to all datasets on World Factbook:
https://www.cia.gov/library/publications/resources/the-world-factbook/rankorder/rankorderguide.html

Link to definitions of words from Dataset:
https://www.cia.gov/library/publications/resources/the-world-factbook/docs/notesanddefs.html

- Data Wrangling (included)
- Exploratory data analysis (included)
- Plot and summary statistics (included)
- Confidence Intervals (included)
- Classification trees (included)
- Regression models(included)