BSDS4999

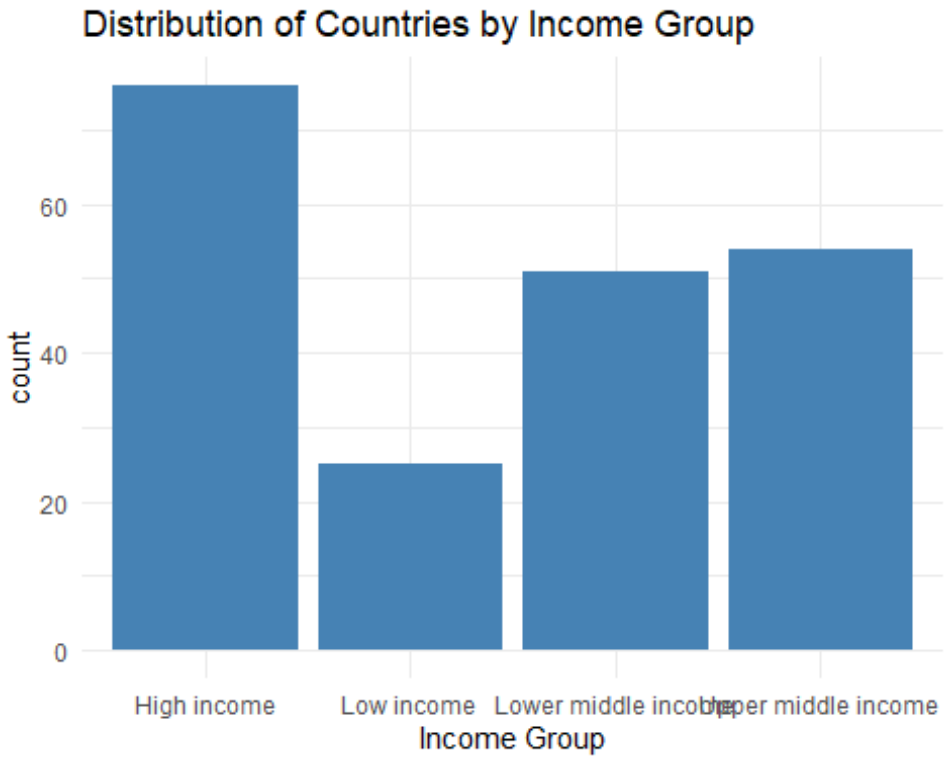
2025-05-14

##GDP and GNI

```
# Load required libraries
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

library(ggplot2)

#GDP
gdp = read.csv("GDP.csv")

# Calculate the mean GDP for each country
colnames(gdp)[4] <- "GDP"
gdp <- gdp %>%
  group_by(Code) %>%
  summarize(Mean_GDP = mean(GDP, na.rm = TRUE))



#GNI
gni = read.csv("GNI.csv")
gni_with_value = read.csv("GNI_Value.csv", skip=3)

# Calculate the mean GNI for each country and Filter out missed values
gni_with_value$Mean_GNI <- rowMeans(gni_with_value[,
5:ncol(gni_with_value)], na.rm = TRUE)
gni_with_value = gni_with_value[,c("Country.Code", 'Mean_GNI')]
gni_with_value = subset(gni_with_value, Mean_GNI!= 'NaN')
gni = subset(gni, IncomeGroup != '')
gni = merge(gni, gni_with_value, by = "Country.Code")

# Combine the data of GNI and GDP
gdpgni = merge(gdp, gni, by.x = "Code", by.y = "Country.Code")
gdpgni = gdpgni[,c("Code", "Mean_GDP", "Mean_GNI", "IncomeGroup",
'Region')]
```

##Bar plot for Income Group

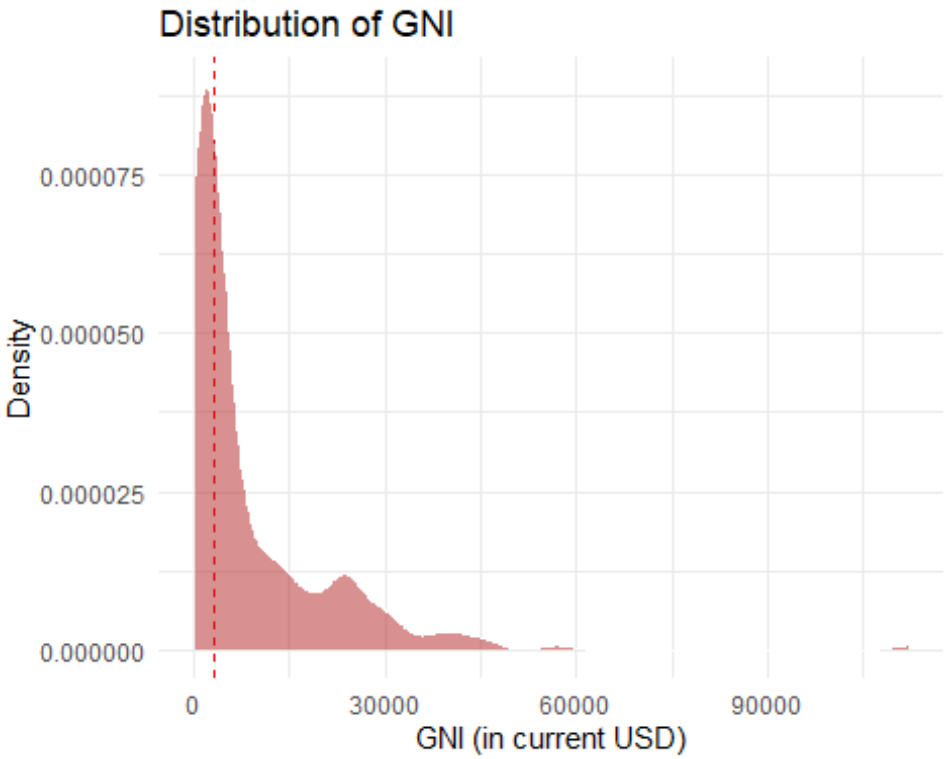
```
# Create a bar plot for GNI
ggplot(gni) +
 aes(x = IncomeGroup) +
 geom_bar(fill = "#4682B4") +
 labs(title = "Distribution of Countries by Income Group", x = 'Income
Group') +
 theme_minimal()
```



Density Plot for GNI

```
# Create a Density Plot for GNI
ggplot()+
#  scale_x_continuous(breaks = c(25000, 50000, 75000, 100000)) +
  scale_y_continuous(labels = scales::comma) +
#  geom_vline(xintercept = median(gdpgni$Mean_GDP), linetype =
'dashed', color ='green')+
  geom_vline(xintercept = median(gdpgni$Mean_GNI), linetype = 'dashed',
color ='red')+
#  geom_density(aes(x = gdpgni$Mean_GDP, fill="GDP"), color = 'white',
alpha=0.7) +
  geom_density(aes(x = gdpgni$Mean_GNI,fill="GNI"), color = 'white',
fill = '#B22222', alpha=0.5) +
  theme_minimal()+
  labs(title = "Distribution of GNI",
```

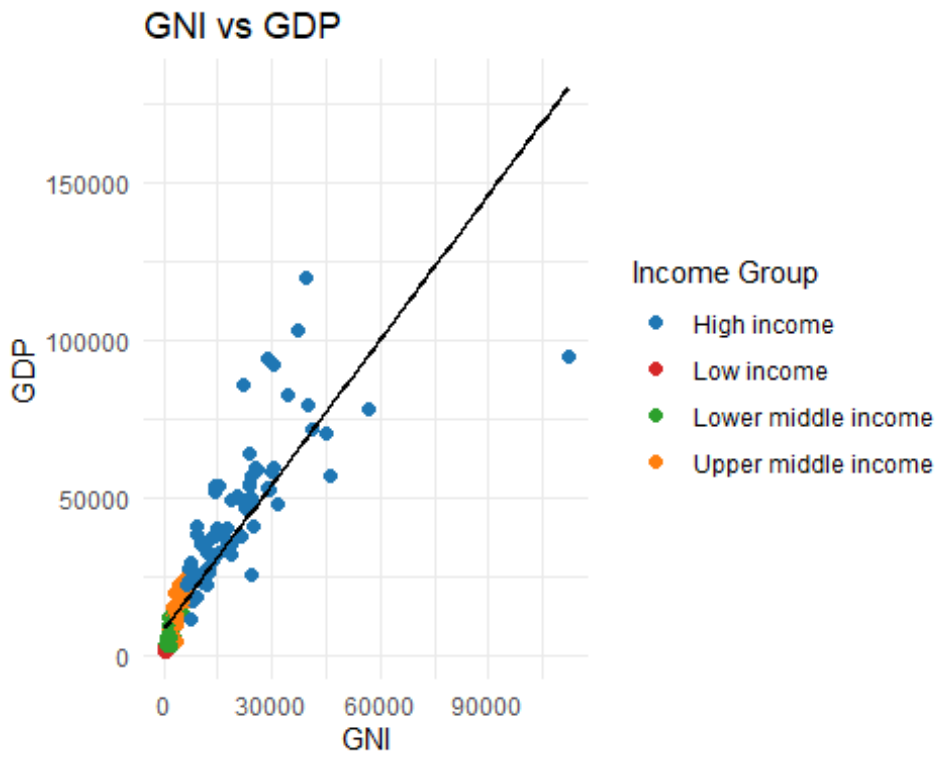
```
       x = "GNI (in current USD)",
       y = "Density")
```




```
#  scale_fill_manual(name = "", values = c("GDP" = "#69b3a2", "GNI" =
"#B22222"))
```

Relationship of GNP and GDP

```
ggplot(gdpgni, aes(x = Mean_GNI, y = Mean_GDP)) +
  geom_point(aes(color = IncomeGroup), size = 2) +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  labs(title = "GNI vs GDP",
       x = "GNI",
       y = "GDP") +
  theme_minimal() +
  scale_color_manual(name = 'Income Group', values = c("High income" =
"#1f77b4", "Upper middle income" = "#ff7f0e",
                                "Lower middle income" = "#2ca02c",
"Low income" = "#d62728"))

## `geom_smooth()` using formula = 'y ~ x'
```

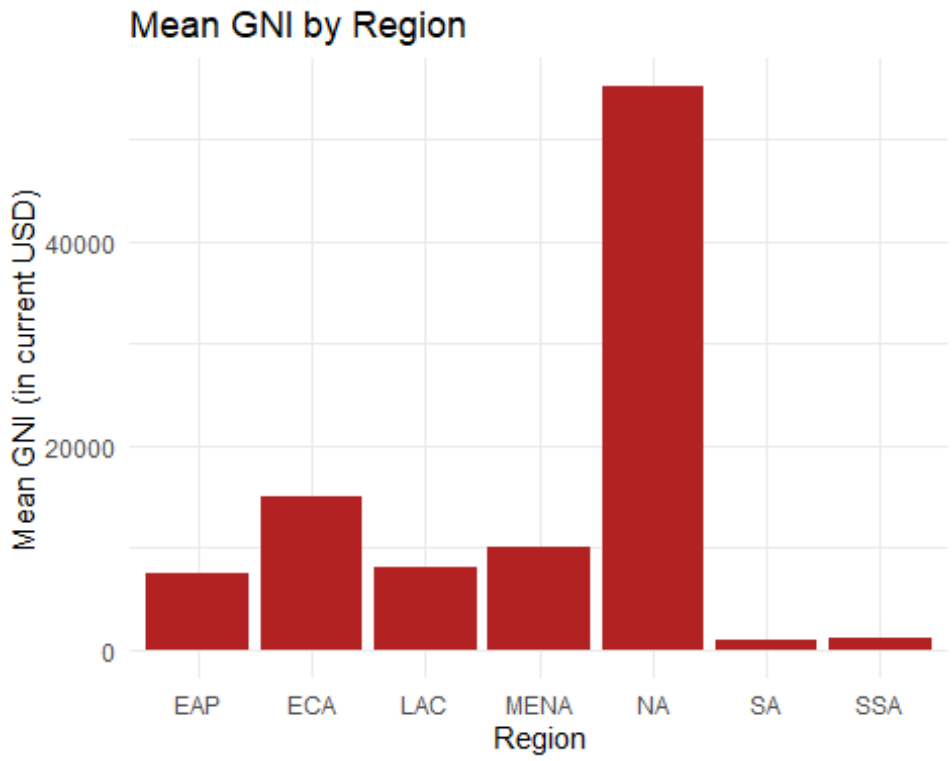



```
cor(gdpgni$Mean_GNI, gdpgni$Mean_GDP)

## [1] 0.868065
```

Association between poverty and Region

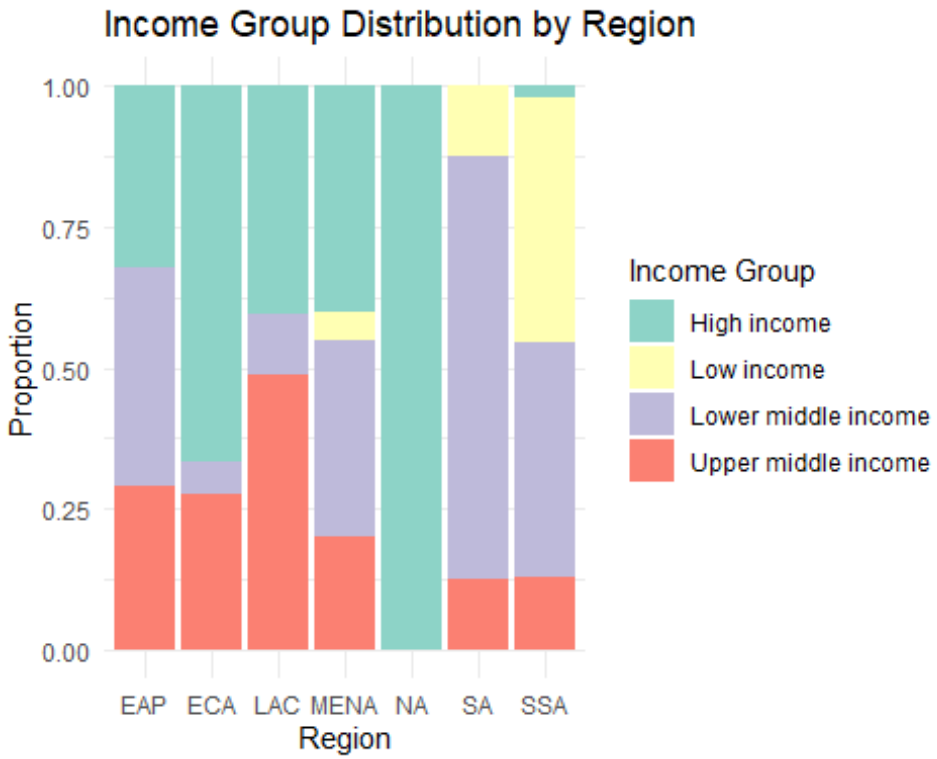
```
# Association Income group and Region
table <- table(gdpgni$Region, gdpgni$IncomeGroup)
chi_square <- chisq.test(table)

## Warning in chisq.test(table): Chi-squared approximation may be
incorrect

chi_square

##
##  Pearson's Chi-squared test
##
## data:  table
## X-squared = 129.31, df = 18, p-value < 2.2e-16

# Create Simplified Code for Region
gdpgni <- gdpgni %>%
  mutate(Simplified_Region = case_when(
    Region == "Latin America & Caribbean" ~ "LAC",
    Region == "South Asia" ~ "SA",
    Region == "Sub-Saharan Africa" ~ "SSA",
    Region == "Europe & Central Asia" ~ "ECA",
```

```
    Region == "Middle East & North Africa" ~ "MENA",
    Region == "East Asia & Pacific" ~ "EAP",
    Region == "North America" ~ "NA"))

# Calculate the mean of GNI of every region
mean_gni_region = gdpgni %>%
  group_by(Simplified_Region) %>%
  summarize(Mean_GNI_by_region = mean(Mean_GNI, na.rm = TRUE))

# Create a bar plot for Mean GDP by Region
ggplot(mean_gni_region) +
  aes(x = Simplified_Region, y = Mean_GNI_by_region) +
  geom_col(fill = "#B22222") +
  theme_minimal() +
  labs(title = "Mean GNI by Region",
       x = "Region",
       y = "Mean GNI (in current USD)")
```



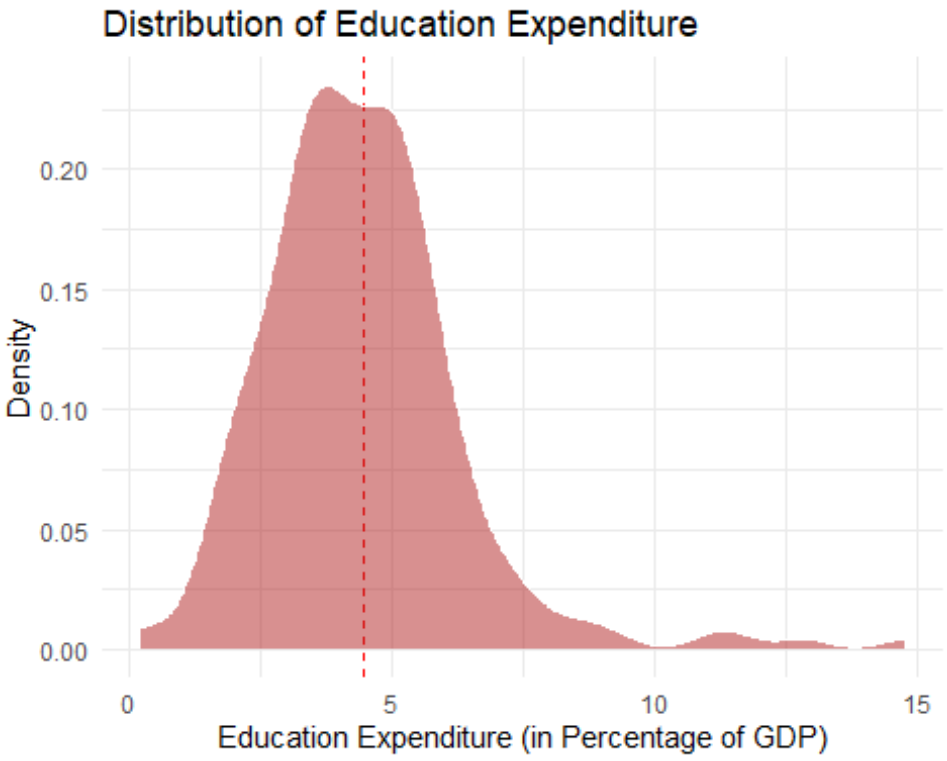

```
# Create a bar plot for Region and Income Group
ggplot(gdpgni, aes(x = Simplified_Region, fill = IncomeGroup)) +
  geom_bar(position = "fill") +
  labs(title = "Income Group Distribution by Region",
       x = "Region",
       y = "Proportion",
       fill = 'Income Group') +
```

```
  theme_minimal() +
  scale_fill_brewer(palette = "Set3")
```

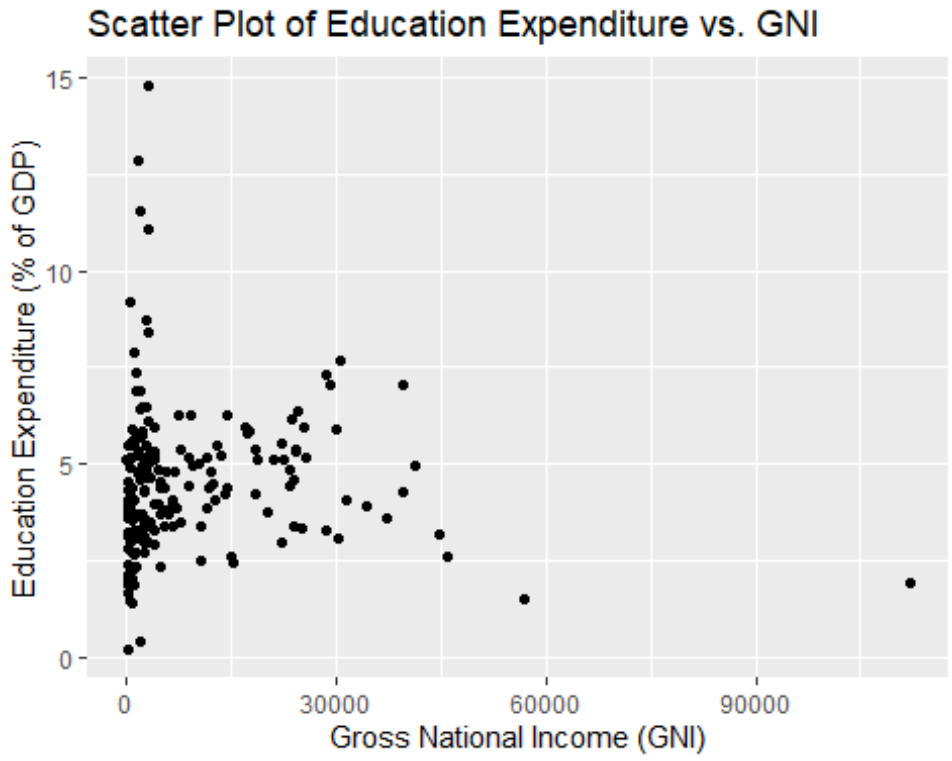


Education expenditure

```
# Read file and merge with GNI data
edu = read.csv("EDU.csv")
edu = edu %>%
  group_by(geoUnit) %>%
  summarize(Mean_Education = mean(`value`, na.rm = TRUE))
edu = merge(edu, gdpgni, by.x = "geoUnit", by.y = "Code")

# Create Density Plot
ggplot()+
  scale_y_continuous(labels = scales::comma) +
  geom_vline(xintercept = mean(edu$Mean_Education), linetype =
'dashed', color ='red')+
  geom_density(aes(x = edu$Mean_Education,fill="Education"), color =
'white', fill = '#B22222', alpha=0.5) +
  theme_minimal()+
  labs(title = "Distribution of Education Expenditure",
       x = "Education Expenditure (in Percentage of GDP)",
       y = "Density")
```




```r
# Create a scatter plot for Education Expenditure and GNI
ggplot(edu, aes(x = Mean_GNI, y = Mean_Education)) +
  geom_point() +
  labs(title = "Scatter Plot of Education Expenditure vs. GNI",
       x = "Gross National Income (GNI)",
       y = "Education Expenditure (% of GDP)")
```

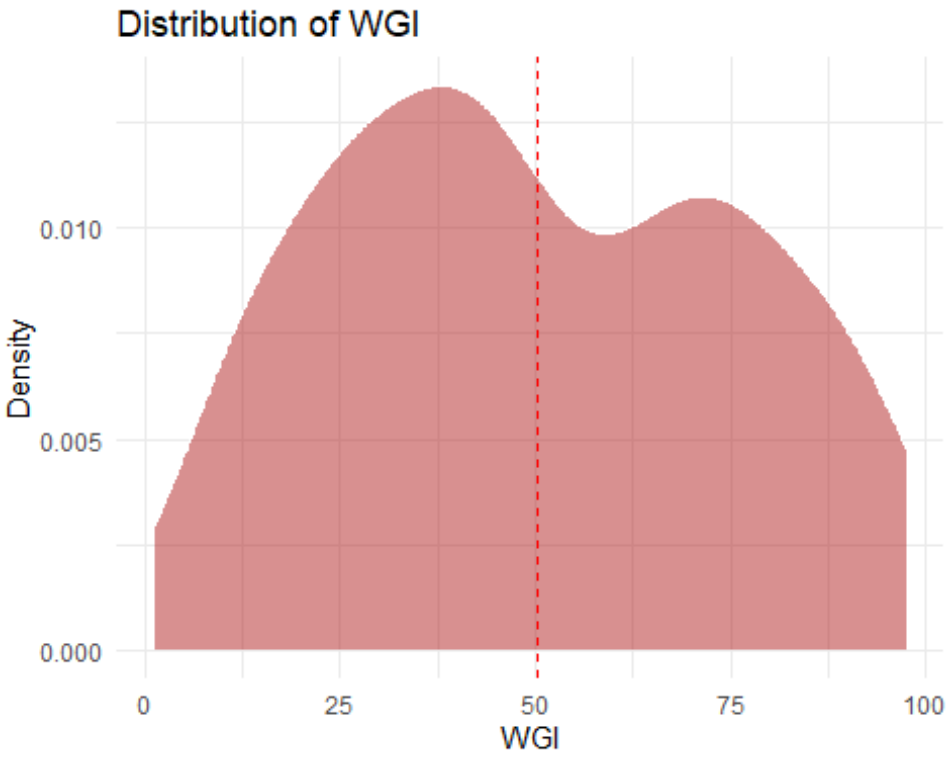



```
cor(edu$Mean_GNI, edu$Mean_Education)

## [1] 0.005454199
```

WGI

```
# Input the data of WGI
library(readxl)
wgi = read_excel("WGI.xlsx")
wgi$pctrank <- as.numeric(wgi$pctrank)

## Warning: NAs introduced by coercion

wgi = subset(wgi, pctrank != '')
wgi = wgi %>%
  group_by(code) %>%
  summarize(Mean_WGI = mean(pctrank, na.rm = TRUE))
wgi = merge( wgi, gdpgni, by.x = "code", by.y = "Code")

# Create a Density Plot for WGI
ggplot()+
  geom_density(aes(x = wgi$Mean_WGI), color = 'white', fill =
'#B22222', alpha=0.5) +
  geom_vline(xintercept = mean(wgi$Mean_WGI), linetype = 'dashed',
color ='red') +
  theme_minimal()+
  labs(title = "Distribution of WGI",
```

```
       x = "WGI",
       y = "Density")
```

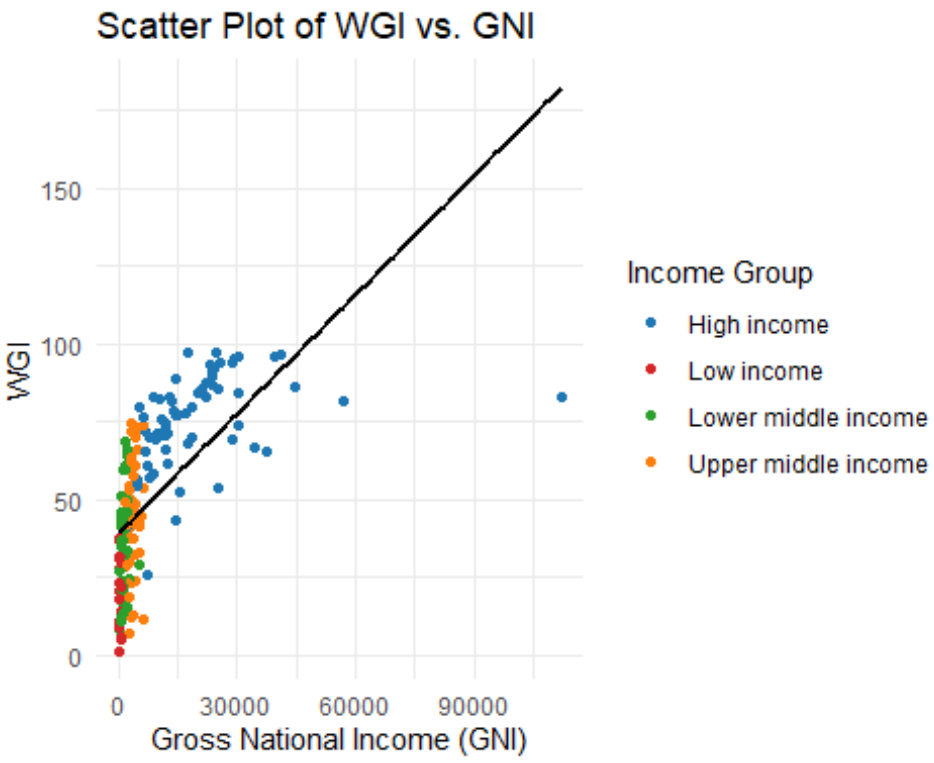



```
# Use Spearman's Rank Correlation to find out association WGI and GDP
cor(wgi$Mean_WGI, wgi$Mean_GNI, method = 'spearman')
```

```
## [1] 0.8101704
```

```
# Create a scatter plot for WGI and GNI
ggplot(wgi, aes(x = Mean_GNI, y = Mean_WGI, color = IncomeGroup)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "black")+
  labs(title = "Scatter Plot of WGI vs. GNI",
       x = "Gross National Income (GNI)",
       y = "WGI") +
  theme_minimal() +
  scale_color_manual(name = 'Income Group',
                     values = c("High income" = "#1f77b4",
                                "Upper middle income" = "#ff7f0e",
                                "Lower middle income" = "#2ca02c",
                                "Low income" = "#d62728"))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# Create a simple regression model to predict WGI based on GNI
model <- lm(Mean_GNI ~ Mean_WGI, data = wgi)
summary(model)

##
## Call:
## lm(formula = Mean_GNI ~ Mean_WGI, data = wgi)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13036  -4999  -1030   3187  93030
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7632.77    1592.15  -4.794 3.34e-06 ***
## Mean_WGI      321.63      28.24  11.388  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9947 on 186 degrees of freedom
## Multiple R-squared:  0.4108, Adjusted R-squared:  0.4076
## F-statistic: 129.7 on 1 and 186 DF,  p-value: < 2.2e-16
```

Multiple regression model

```
model2 <- lm(Mean_GNI ~ Mean_WGI + factor(Region), data = wgi)
summary(model2)
```

```
##
## Call:
## lm(formula = Mean_GNI ~ Mean_WGI + factor(Region), data = wgi)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -33575  -4121   -913   2599  57911
##
## Coefficients:
##                                          Estimate Std. Error t value
Pr(>|t|)
## (Intercept)                              -7810.46    2281.99  -3.423
0.000768
## Mean_WGI                                   273.99      29.05   9.431
< 2e-16
## factor(Region)Europe & Central Asia       4641.11    2008.07   2.311
0.021953
## factor(Region)Latin America & Caribbean   -442.82    2158.01  -0.205
0.837648
## factor(Region)Middle East & North Africa  6815.34    2563.95   2.658
0.008565
## factor(Region)North America              39242.16    5285.35   7.425
4.35e-12
## factor(Region)South Asia                  -319.48    3492.68  -0.091
0.927220
## factor(Region)Sub-Saharan Africa           574.71    2165.00   0.265
0.790963
##
## (Intercept)                              ***
## Mean_WGI                                 ***
## factor(Region)Europe & Central Asia      *
## factor(Region)Latin America & Caribbean
## factor(Region)Middle East & North Africa **
## factor(Region)North America              ***
## factor(Region)South Asia
## factor(Region)Sub-Saharan Africa
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8611 on 180 degrees of freedom
## Multiple R-squared:  0.5727, Adjusted R-squared:  0.5561
## F-statistic: 34.47 on 7 and 180 DF,  p-value: < 2.2e-16

library(broom)
coef_df <- tidy(model2, conf.int = TRUE)
ggplot(coef_df[-1, ], aes(x = estimate, y = term)) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "red") +
  geom_point(size = 3) +
  geom_errorbarh(aes(xmin = conf.low, xmax = conf.high), height = 0.2)
+
```

```
  labs(
    title = "Impact of Predictors on GNI",
    x = "Coefficient Estimate",
    y = "Predictor",
    caption = "Error bars show 95% confidence intervals"
  ) +
  theme_minimal()
```