

WHO Notebook

```
#install.packages("ggplot2")
```

One of the advantages of software such as R is that you can add and install packages that provided added functionality. The `ggplot2` packages is one such package that improves on the basic plotting in R. In this package, you build a graph by adding layer upon layer where you specify the data, aesthetic mapping, statistical transformations (if any) and geometric object (type of plot).

We will use a data set from <http://www.who.int/gho/en>. The dataset contains data on 194 countries from WHO where the variables are: 1. Country - name of country, 2. Region - region the country lies in, 3. Population - population in thousands, 4. Under15 - Percentage of the population under 15 years of age, 5. Over60 - Percentage of the population over 60 years of age, 6. FertilityRate - Average number of children per woman, 7. LifeExpectancy - Life expectancy in years, 8. LiteracyRate - Literacy rate among adults at least 15 years of age, 9. GNI - Gross national income per capita

First we have a look at the data set

```
library(ggplot2)
WHO <- read.csv("WHO.csv")
str(WHO)
```

```
## 'data.frame': 194 obs. of 9 variables:
## $ Country : chr "Afghanistan" "Albania" "Algeria" "Andorra" ...
## $ Region : chr "Eastern Mediterranean" "Europe" "Africa" "Europe" ...
## $ Population : int 29825 3162 38482 78 20821 89 41087 2969 23050 8464 ...
## $ Under15 : num 47.4 21.3 27.4 15.2 47.6 ...
## $ Over60 : num 3.82 14.93 7.17 22.86 3.84 ...
## $ FertilityRate : num 5.4 1.75 2.83 NA 6.1 2.12 2.2 1.74 1.89 1.44 ...
## $ LifeExpectancy: int 60 74 73 82 51 75 76 71 82 81 ...
## $ LiteracyRate : num NA NA NA NA 70.1 99 97.8 99.6 NA NA ...
## $ GNI : num 1140 8820 8310 NA 5230 ...
```

```
which(WHO$Country == "Singapore")
```

```
## [1] 155
```

```
# WHO[155,]
summary(WHO$Over60)
```

```
## Min. 1st Qu. Median Mean 3rd Qu. Max.
## 0.81 5.20 8.53 11.16 16.69 31.92
```

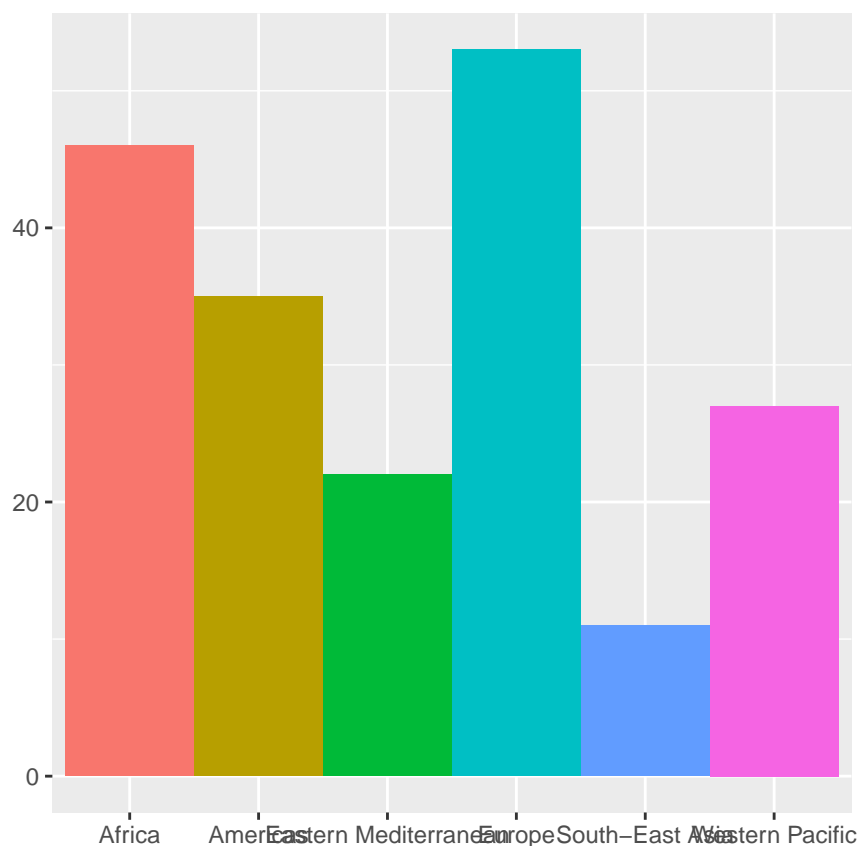
```
summary(WHO$Under15)
```

```
## Min. 1st Qu. Median Mean 3rd Qu. Max.
## 13.12 18.72 28.65 28.73 37.75 49.99
```

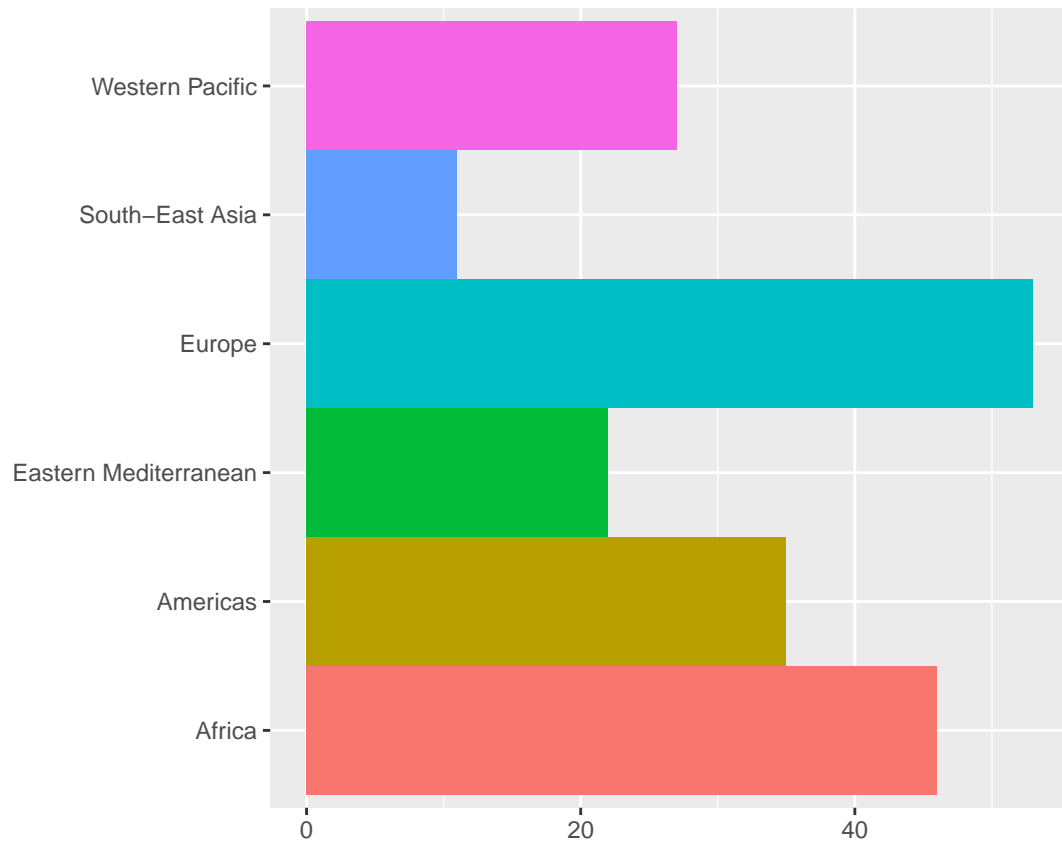
Singapore has a high fraction of population over 60 and low fraction of population under 15 as compared to many countries - ageing population and low fertility rate issues.

Visualisation Let us see how many countries are there in each region

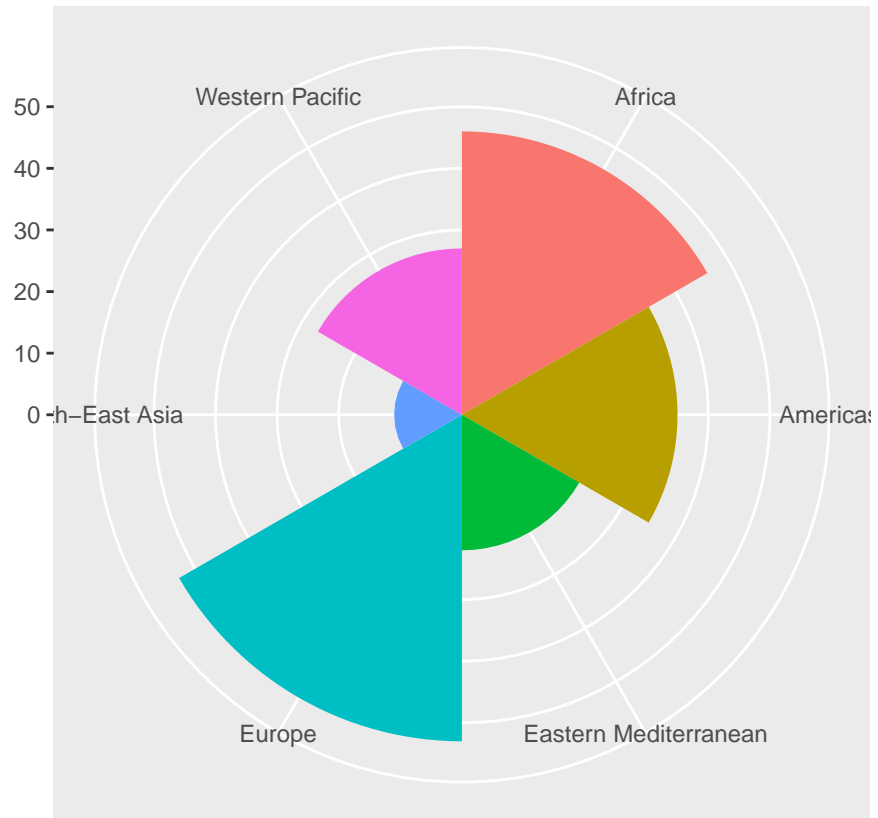
```
bar <- ggplot(WHO) + geom_bar(mapping = aes(x = Region, fill = Region), show.legend = FALSE, width = 1)
bar
```



```
bar + coord_flip() # flip x and y to get horizontal bar chart
```



```
bar + coord_polar() # polar chart for bar charts. ie. radial chart
```



Now we compare the following

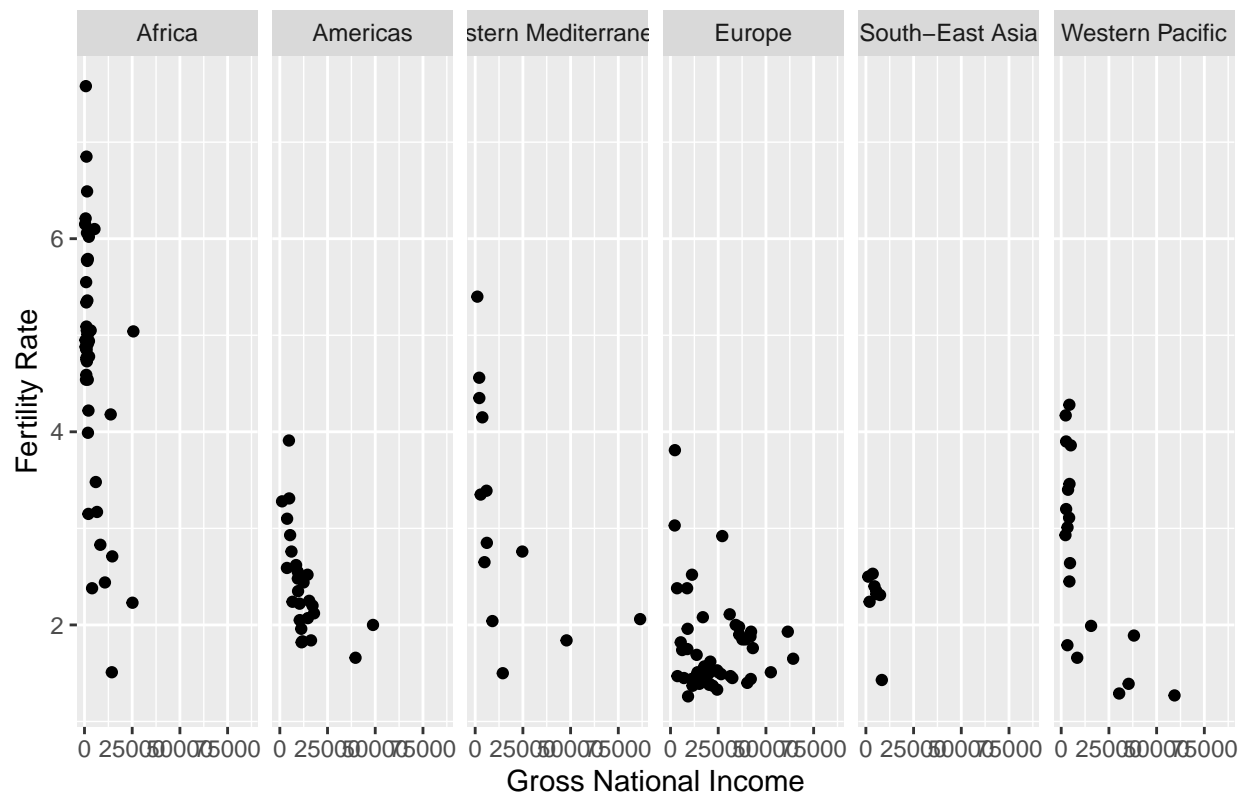
1. GNI vs Fertility rate
2. GNI vs Literacy Rate
3. GNI vs Life Expectancy

We also subgroup by region sometimes

```
# 1. GNI vs Fertility rate
a <- ggplot(WHO, aes(x = GNI, y = FertilityRate))

a + geom_point(na.rm = TRUE) + facet_grid(.~Region) + ggtitle("Fertility Rate vs. Gross National Income")
```

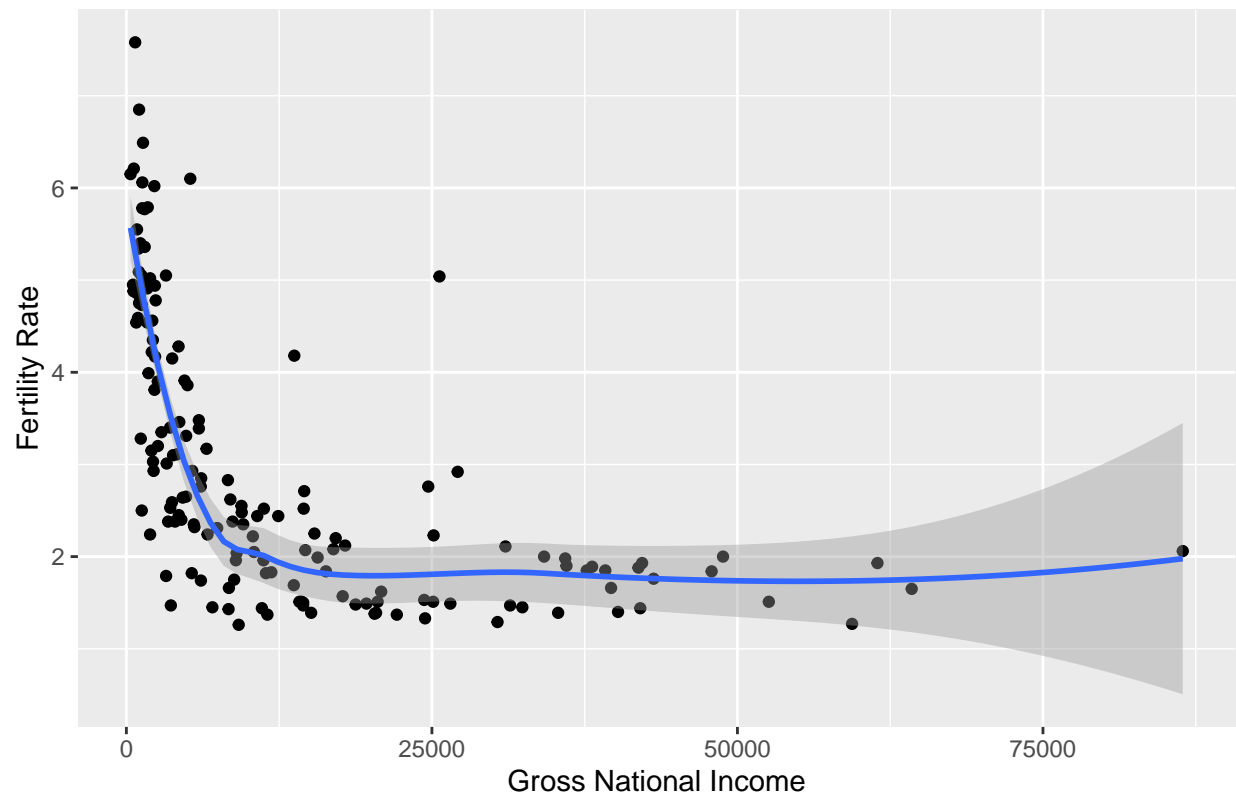
Fertility Rate vs. Gross National Income



```
# facet_grid will give different panels of plots (instead of one combined plot), plots separated based
a + geom_point(na.rm = TRUE) + geom_smooth(na.rm = TRUE) + ggtitle("Fertility Rate vs. Gross National Income")
```

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

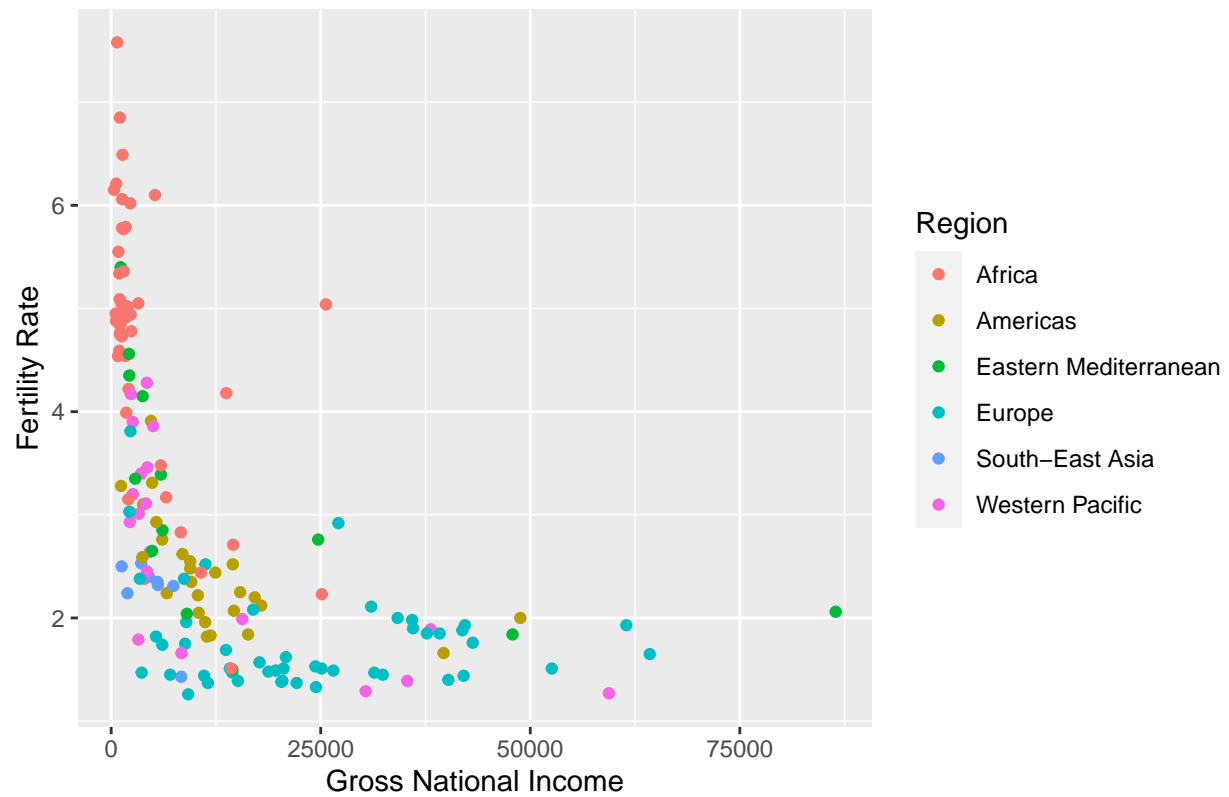
Fertility Rate vs. Gross National Income



```
# color by region
acol <- ggplot(WHO, aes(x = GNI, y = FertilityRate, color = Region))

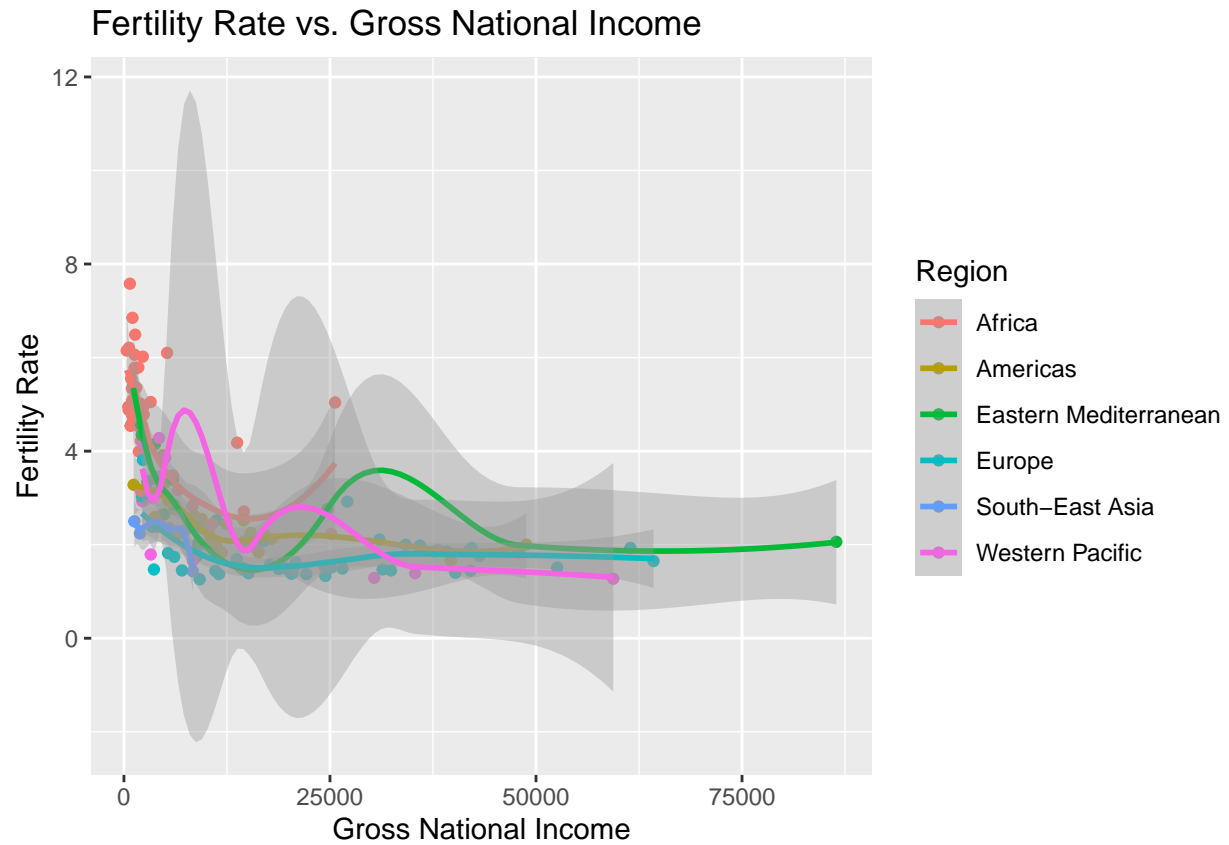
acol + geom_point(na.rm = TRUE) + ggtitle("Fertility Rate vs. Gross National Income") + xlab("Gross Nat
```

Fertility Rate vs. Gross National Income



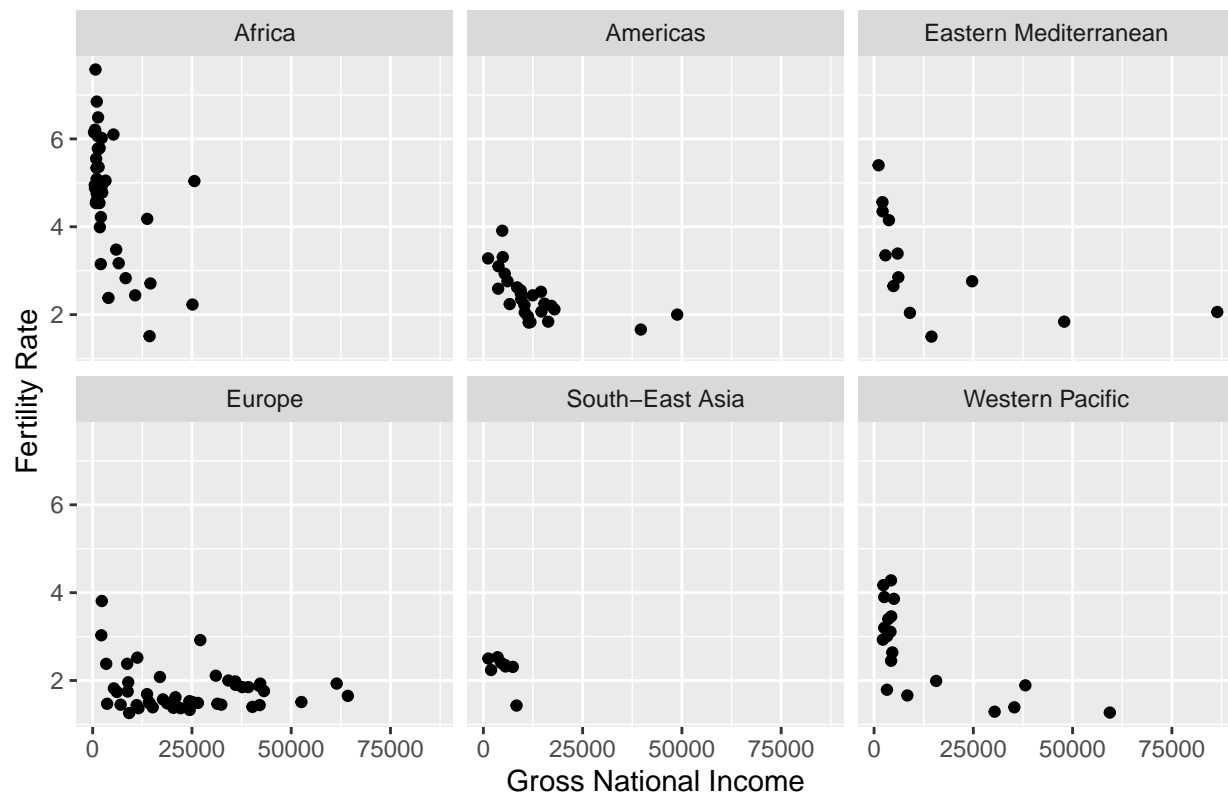
```
# add geom_smooth
acol + geom_point(na.rm = TRUE) + geom_smooth(na.rm = TRUE) + ggtitle("Fertility Rate vs. Gross National Income")

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



```
# multiple panels of plots, split by Region column
a + geom_point(na.rm = TRUE) + facet_wrap(~Region, nrow = 2) + ggtitle("Fertility Rate vs. Gross National Income")
```

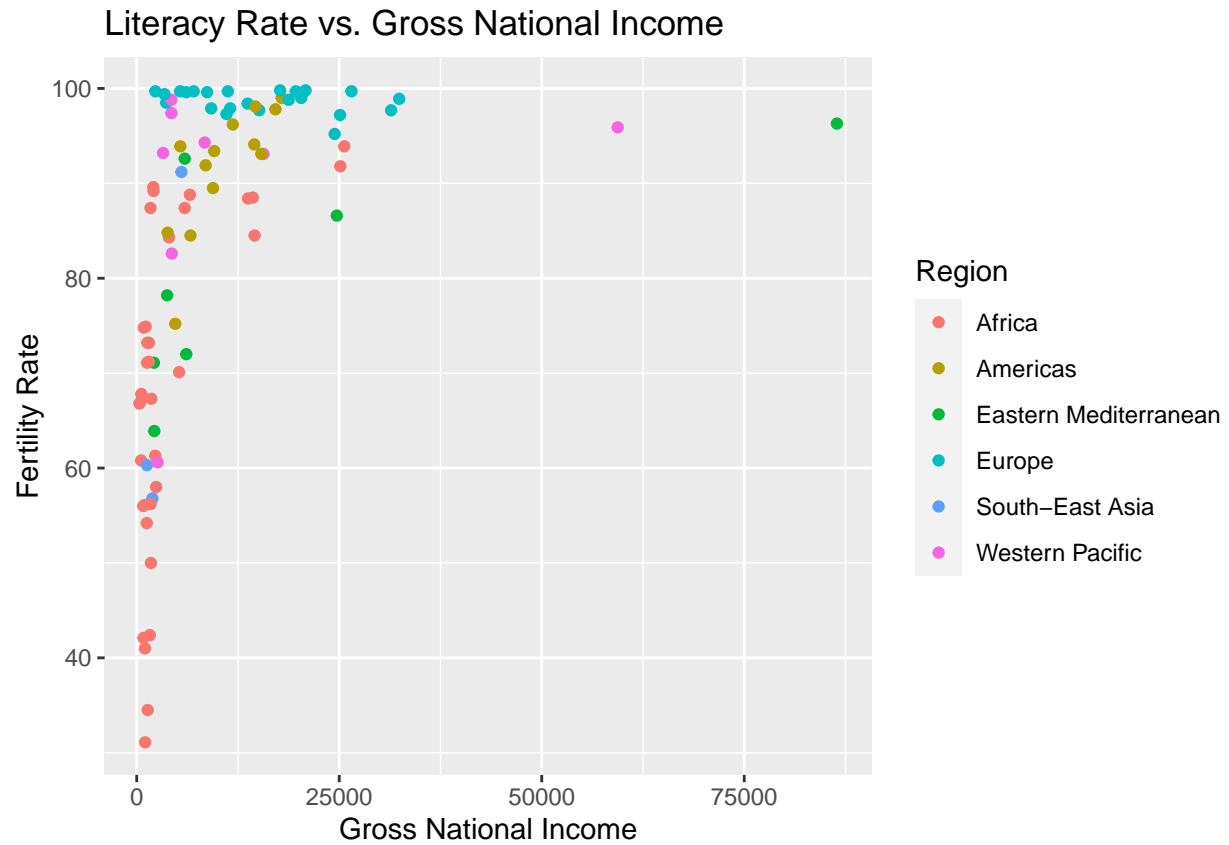

Fertility Rate vs. Gross National Income



*# the difference between facet_grid and facet_wrap is that grid will display all the plots in one row
while using wrap, we can define the number of rows to display the plots for better screen dimensions*

2. GNI vs Literacy rate

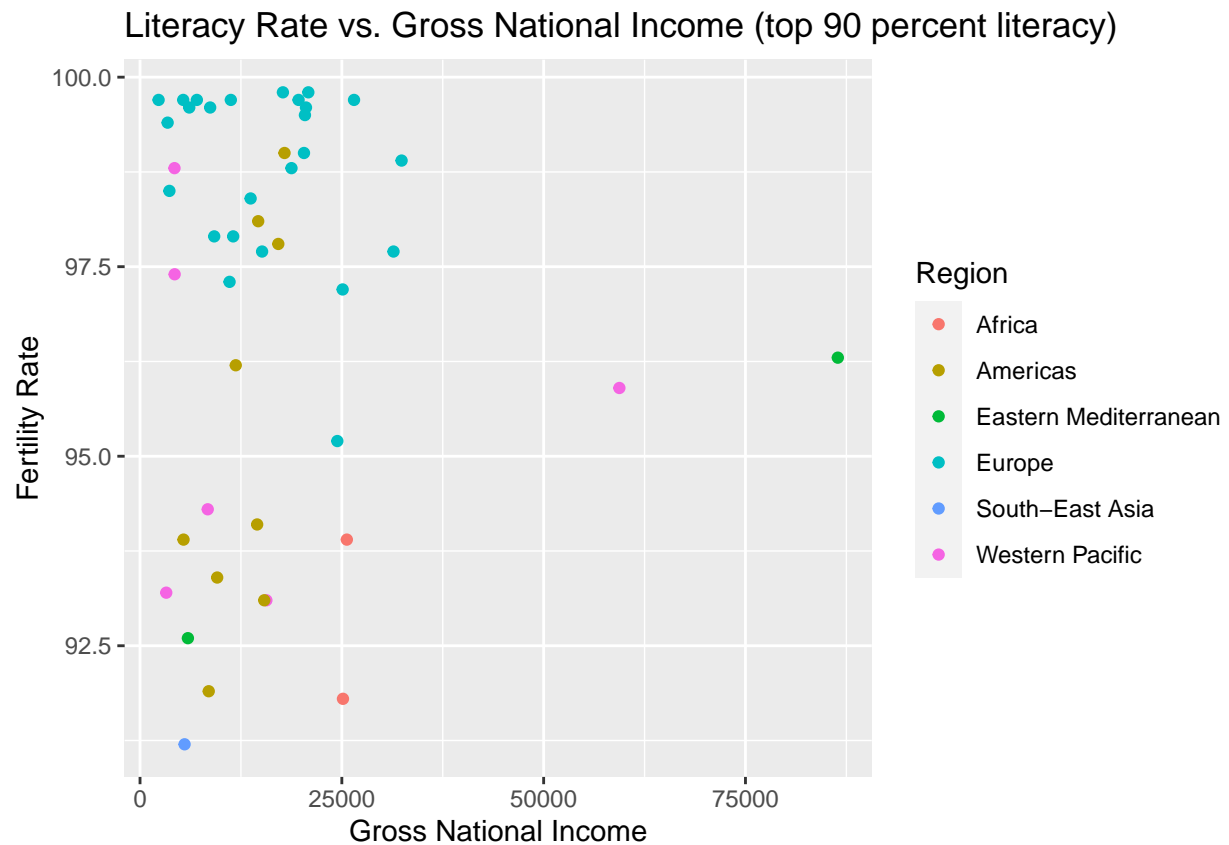
```
b <- ggplot(WHO, aes(x = GNI, y = LiteracyRate, color = Region)) + ggtitle("Literacy Rate vs. Gross Nat.  
b + geom_point(na.rm = TRUE)
```



```
# filter out a selected part of the original dataset
WHOL <- subset(WHO, LiteracyRate > 90)

c <- ggplot(WHOL, aes(x = GNI, y = LiteracyRate, color = Region))

c + geom_point(na.rm = TRUE) + ggtitle("Literacy Rate vs. Gross National Income (top 90 percent literacy)")
```

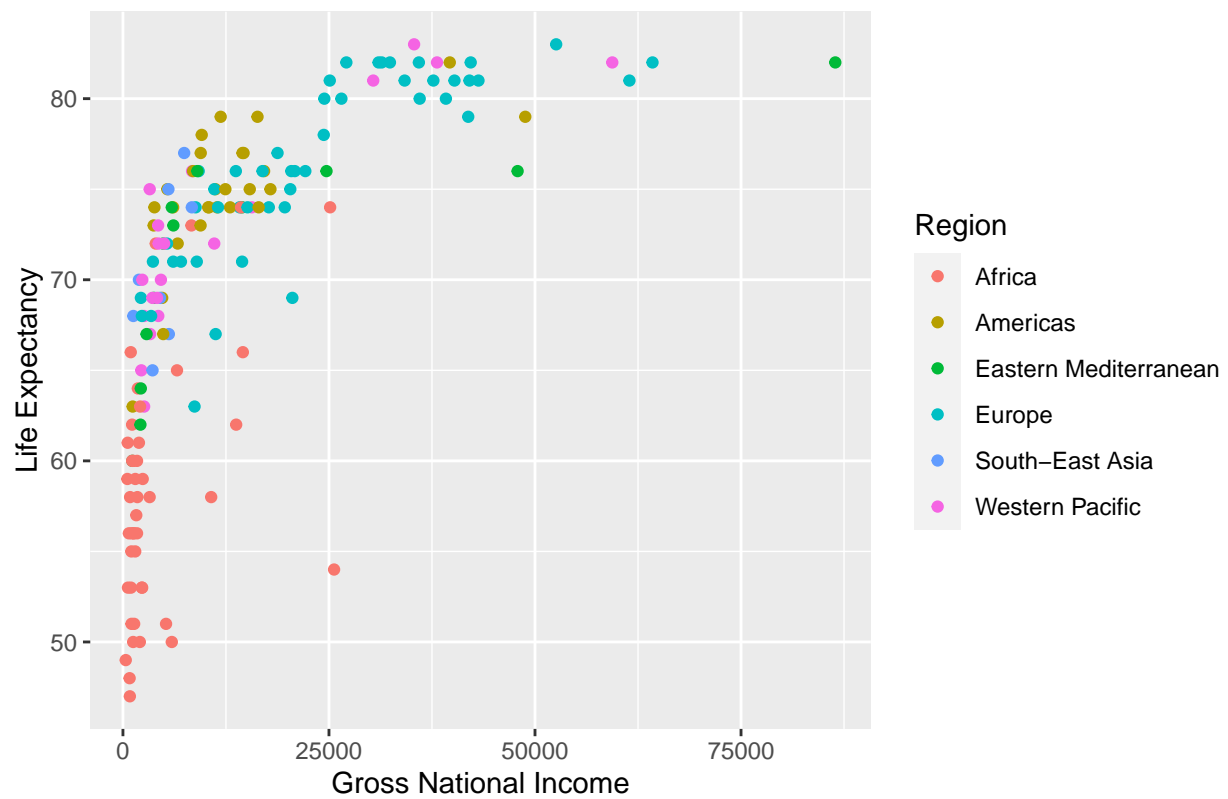


```
# 3. GNI vs Life expectancy
```

```
d <- ggplot(WHO, aes(x = GNI, y = LifeExpectancy, color = Region))
```

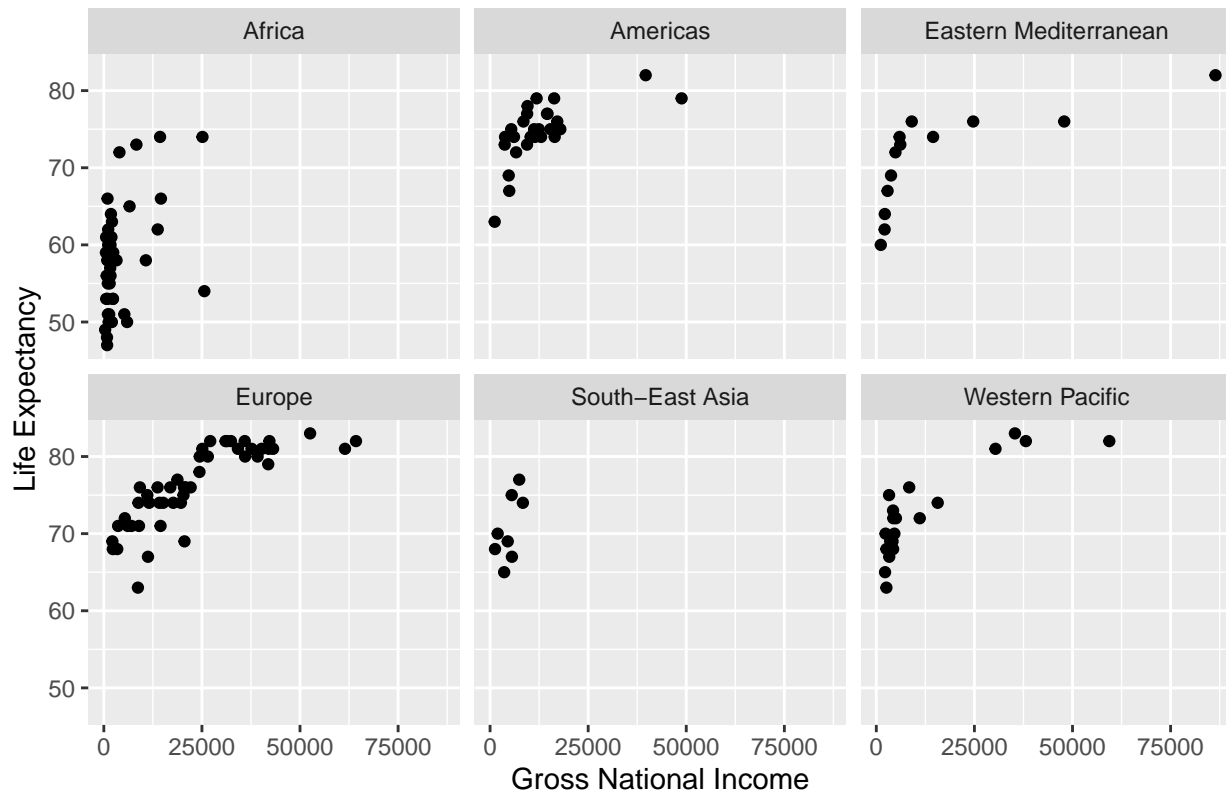
```
d + geom_point(na.rm = TRUE) + ggtitle("Literacy Rate vs. Gross National Income") + xlab("Gross National Income")
```

Literacy Rate vs. Gross National Income



```
ggplot(WHO, aes(x = GNI, y = LifeExpectancy)) + geom_point(na.rm = TRUE) + facet_wrap(~Region, nrow = 2)
```

Literacy Rate vs. Gross National Income



For single variable plots you can use either `facet_grid()` or `facet_wrap()`.

- `facet_wrap()` : “wraps” a 1d ribbon of panels into 2d.
- `facet_grid()` : produces a 2d grid of panels defined by variables which form the rows and
- `facet_wrap(~variable)` will return a symmetrical matrix of plots for the number of levels of variable.
- `facet_grid(.~variable)` will return facets equal to the levels of variable distributed horizontally.
- `facet_grid(variable~.)` will return facets equal to the levels of variable distributed vertically.

Certain observations from the plots above

1. The plots above indicate an inverse relation between income and fertility. This suggests that reproductive restraint arises often as a consequence of economic progress or lower fertility leads to more resources being available per child, making them more productive (causality is not observable). This is often referred to as the *demographic-economic paradox*.
2. We also color the points as per the regions by adding the color argument to the aesthetic option. This helps see the trend by regions. For example Europe has high GNI, low fertility rates while Africa has low GNI and high fertility rates.
3. We can similarly analyze the other pairwise plots we see.