R and visualization

Goal : -Become familiar with data visualization

Part 1 - ggplot

1. Load and explore the data from “Who.csv”  
   *#This set contains the world health organization data*
2. Install ggplot2 package using : install.packages(“ggplot2”)
3. Load package using : library(ggplot2)
4. Create your ggplot of Fertility Rate depending on GNI, using ggplot function. Name it scatterplot  
   *# we need at least three things to create a plot using ggplot-- data, an aesthetic mapping of variables in the data frame to visual output, and a geometric object.*
5. Add the geometric objects to ur plot using geom\_... functions ( points, lines)
6. Add a color, shape and size to your points.
7. Add a title using ggtitle function
8. Export your plot as a pdf using pdf(“File name”) ( don’t forget to turn off the process with dev.off() )

Part 2 – visualizing a third parameter

1. Add colors to your points based on Regions  
    *(you should add a color parameter to your aes() function)*
2. Now, color your points based on LifeExpectancy
3. Let’s explore the relation between under15 percentage and FertilityRate, what do we expect ? To do so create a new ggplot with x=GNI, Y=FertilityRate. What type of relation do we see ?
4. Replot with x=log(FertilityRate). What type of relation can we see
5. Create a linear regression model based on the log of FertilityRate
6. Add the regression model line to your scatter plot by adding : +stat\_smooth(method = "lm")  
   *#for a fine line add : ,se=FALSE / to change line color add : , color=”red”*

Part 3 – Heatmaps

1. Load and explore murders set and the US map data by copying :  
   murders=read.csv("murders.csv")  
   statesMap=map\_data("state")
2. Use the long , lat and group variables to ggplot the US map  
   *#Use group=group, and add +geom\_polygon(fill=”white”, color=”black”)*
3. Copy this command on your console  
   ggplot(murderMap,aes(x=long,y=lat,group=group,fill=Murders)) +geom\_polygon(color="black") +scale\_fill\_gradient(low="black",high="red",guide="legend")   
   *#california and texas have the highest percentage of murders*
4. Add MurderRate (murder/population) to your dataset ( \*10000 to have it per 10000 population)
5. Replot according to MurderRate instead of murder
6. To solve the outlier problem(washington DC) add limits to your plot

ggplot(murderMap,aes(x=long,y=lat,group=group,fill=MurderRate)) +geom\_polygon(color="black") +scale\_fill\_gradient(low="black",high="red",guide="legend",limits=c(0,10))