Probability Theory and Statistics

ALY6010, WINTER 2022

Week -6

Module-6- R Practice Output

Submitted by: Abhinav Jain

NUID: 002938209

Submitted To: Prof. Tom Beur

Date: 04/02/2022

Introduction:

World Happiness Dataset

The world happiness report is a research project that looks at people's happiness all around the world. Approximately 155 nations contribute the report on the basis of world happiness data declared on the worldwide day of happiness, which was initially published in 2012, then again in 2013, then again in 2015, and lastly in 2017. Almost all government and commercial enterprises use this data to influence policymaking decisions in the leading sector, as well as to analyze population psychology and health, among other things. Examining the happiness report in the United States in order to boost the country's growth.

Imported the dataset: In this task imported the dataset through which we can compare the world happiness dataset performance with few variable.

	Country.name	Regional.indicator	Ladder.score	Standard.error.of.ladder.score	upperwhisker	lowerwhisker	Logged.GDP.per.capita	Social.support	Healthy.life.expectancy	Freedom.to.make.life.choices
1	Finland	Western Europe	7.842	0.032	7.904	7.780	10.775	0.954	72.000	0.949
2	Denmark	Western Europe	7.620	0.035	7.687	7.552	10.933	0.954	72.700	0.946
3	Switzerland	Western Europe	7.571	0.036	7.643	7.500	11.117	0.942	74.400	0.919
4	Iceland	Western Europe	7.554	0.059	7.670	7.438	10.878	0.983	73.000	0.955
5	Netherlands	Western Europe	7.464	0.027	7.518	7.410	10.932	0.942	72.400	0.913
6	Norway	Western Europe	7.392	0.035	7.462	7.323	11.053	0.954	73.300	0.960
7	Sweden	Western Europe	7.363	0.036	7.433	7.293	10.867	0.934	72.700	0.945

```
> str(df)
'data.frame':   149 obs. of  20 variables:
 $ Country.name                             : chr  "Finland" "Denmark" "Switzerland" "Iceland" ...
 $ Regional.indicator                       : chr  "Western Europe" "Western Europe" "Western Europe" "Western Europe" ...
 $ Ladder.score                             : num  7.84 7.62 7.57 7.55 7.46 ...
 $ Standard.error.of.ladder.score           : num  0.032 0.035 0.036 0.059 0.027 0.035 0.036 0.037 0.04 0.036 ...
 $ upperwhisker                             : num  7.9 7.69 7.64 7.67 7.52 ...
 $ lowerwhisker                             : num  7.78 7.55 7.5 7.44 7.41 ...
 $ Logged.GDP.per.capita                    : num  10.8 10.9 11.1 10.9 10.9 ...
 $ Social.support                           : num  0.954 0.954 0.942 0.983 0.942 0.954 0.934 0.908 0.948 0.934 ...
 $ Healthy.life.expectancy                  : num  72 72.7 74.4 73 72.4 73.3 72.7 72.6 73.4 73.3 ...
 $ Freedom.to.make.life.choices             : num  0.949 0.946 0.919 0.955 0.913 0.96 0.945 0.907 0.929 0.908 ...
 $ Generosity                               : num  -0.098 0.03 0.025 0.16 0.175 0.093 0.086 -0.034 0.134 0.042 ...
 $ Perceptions.of.corruption                : num  0.186 0.179 0.292 0.673 0.338 0.27 0.237 0.386 0.242 0.481 ...
 $ Ladder.score.in.Dystopia                 : num  2.43 2.43 2.43 2.43 2.43 2.43 2.43 2.43 2.43 2.43 ...
 $ Explained.by..Log.GDP.per.capita         : num  1.45 1.5 1.57 1.48 1.5 ...
 $ Explained.by..Social.support             : num  1.11 1.11 1.08 1.17 1.08 ...
 $ Explained.by..Healthy.life.expectancy    : num  0.741 0.763 0.816 0.772 0.753 0.782 0.763 0.76 0.785 0.782 ...
 $ Explained.by..Freedom.to.make.life.choices: num  0.691 0.686 0.653 0.698 0.647 0.703 0.685 0.639 0.665 0.64 ...
 $ Explained.by..Generosity                 : num  0.124 0.208 0.204 0.293 0.302 0.249 0.244 0.166 0.276 0.215 ...
 $ Explained.by..Perceptions.of.corruption  : num  0.481 0.485 0.413 0.17 0.384 0.427 0.448 0.353 0.445 0.292 ...
 $ Dystopia...residual                      : num  3.25 2.87 2.84 2.97 2.8 ...
```

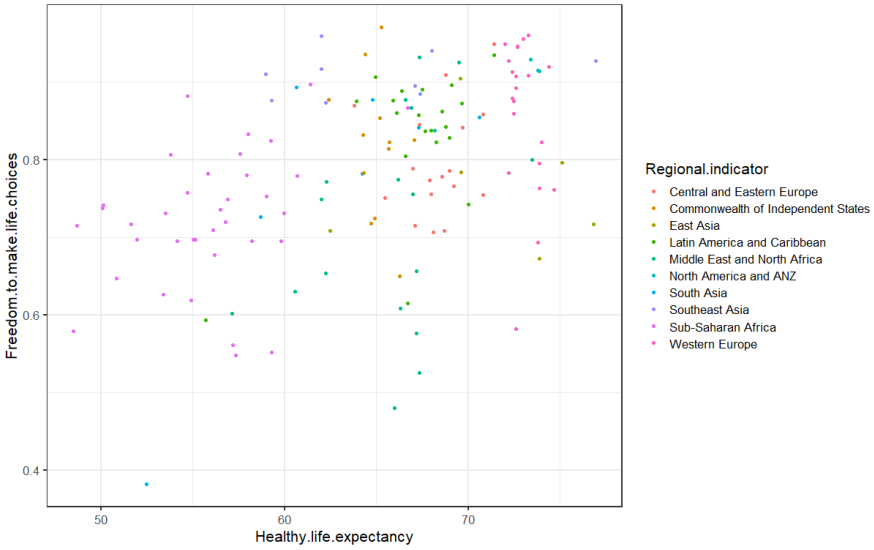
sumtable {vtable}

Summary Statistics

Summary Statistics

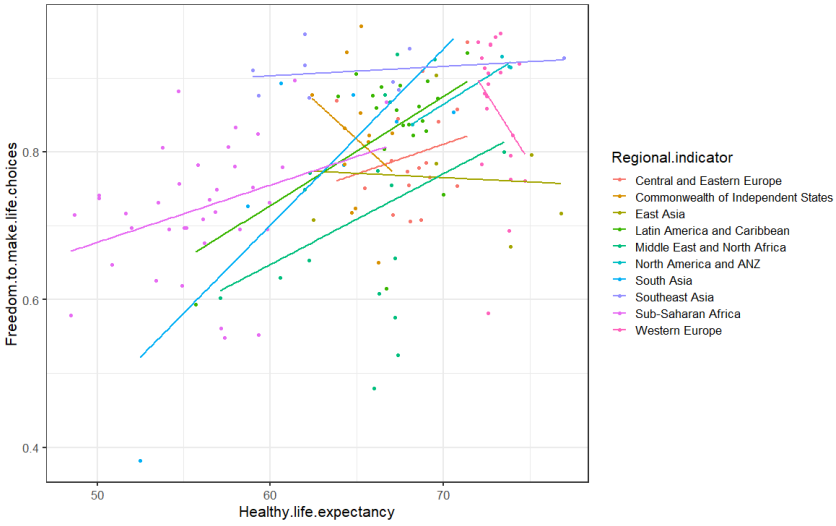
Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Passengers	5211	2536.33	4808.92	0	46	2989.5	67934
Seats	5211	3363.908	6131.183	0	100	4102	89927
Flights	5211	33.894	47.917	0	2	49	491
Distance	5211	692.542	570.302	0	284	920	4962
Fly_date	5211						
... 12/1/2009	5211	100%					
Origin_population	5211	5650104.4	7845570.21	13005	857903	6815696	38139592
Destination_population	5211	9101247.169	7016887.234	13346	1705075	19069796	19161134
Org_airport_lat	5197	37.65	5.903	20.899	33.637	41.709	64.815
Org_airport_long	5197	-90.943	15.628	-157.922	-95.894	-80.291	-68.828
Dest_airport_lat	5198	37.686	5.639	19.721	32.969	41.979	58.355
Dest_airport_long	5198	-87.788	11.528	-155.048	-95.888	-82.533	-68.828

Scatterplot between the Healthy Life Expectancy and the Freedom to Make Life Choices, using regional indicators as an extra variable, we may use a scatterplot to illustrate the related data in color.



Created a scatter plot with multiple regression lines with the categorical variable that is regional indicators

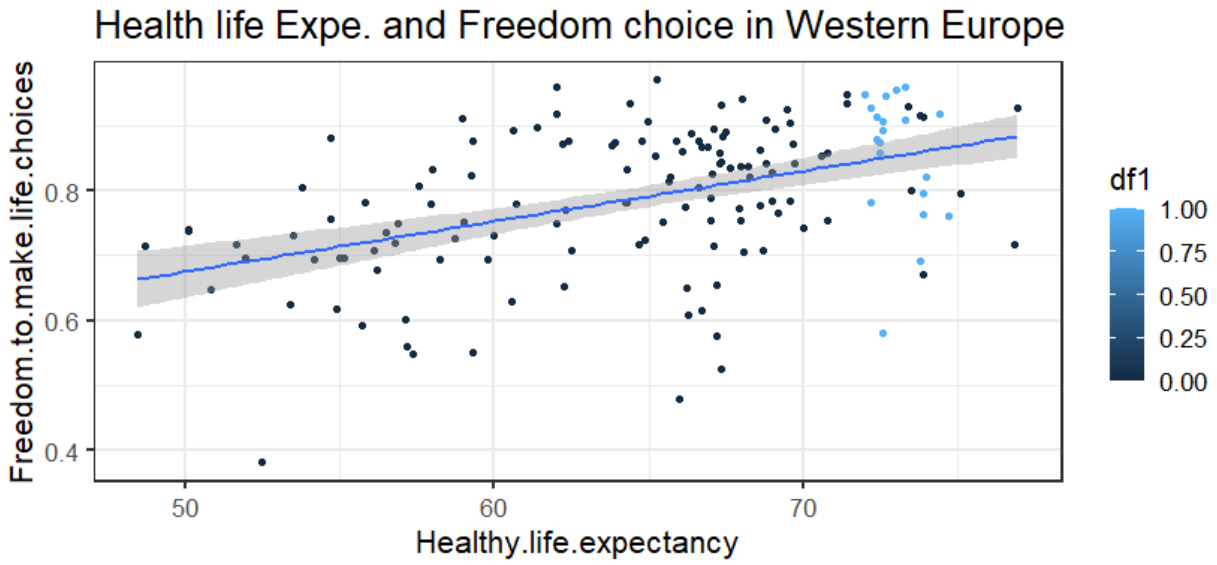
Multiple Regression Lines: To analyze the Healthy life expectancy and freedom to make the life choice, plotted the multiple regression of the different regional indicators which include 9 regions and our focus on Western Europe and South Asia with ladder



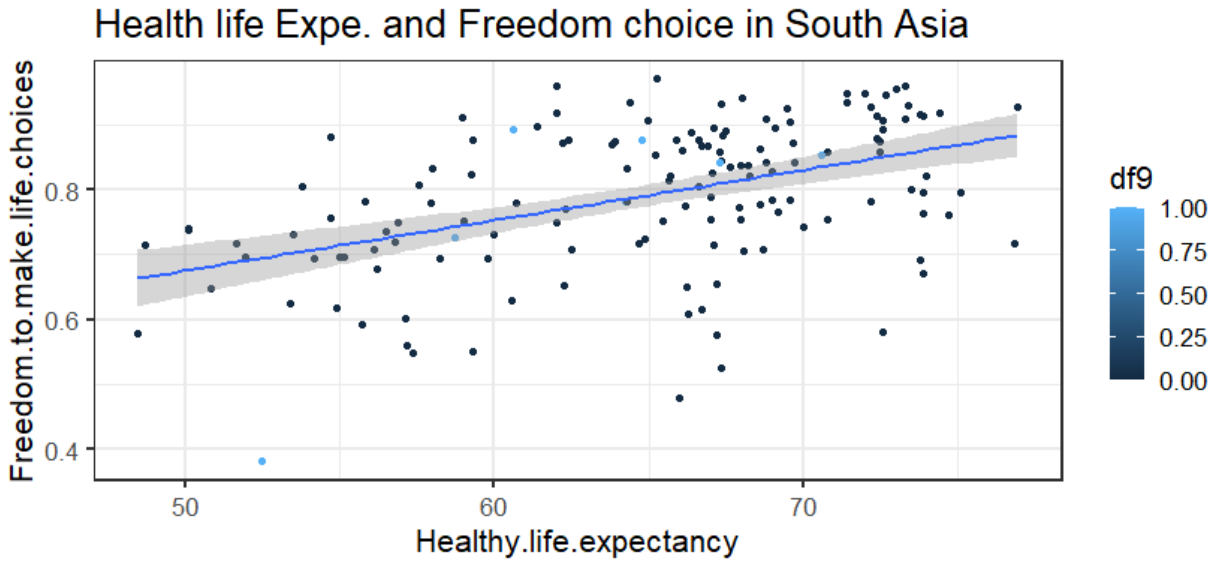
Subset: In this task, we have created the dummy variable by which we can understand the regression between various subsets and the dummy variable we have created of regional indicators.

	Healthy.life.expectancy	Freedom.to.make.life.choices	upperwhisker	Logged.GDP.per.capita	lowerwhisker	df1	df2	df3	df4	df5	df6	df7	df8	df9
1	72.000	0.949	7.904	10.775	7.780	1	0	0	0	0	0	0	0	0
2	72.700	0.946	7.687	10.933	7.552	1	0	0	0	0	0	0	0	0
3	74.400	0.919	7.643	11.117	7.500	1	0	0	0	0	0	0	0	0
4	73.000	0.955	7.670	10.878	7.438	1	0	0	0	0	0	0	0	0
5	72.400	0.913	7.518	10.932	7.410	1	0	0	0	0	0	0	0	0
6	73.300	0.960	7.462	11.053	7.323	1	0	0	0	0	0	0	0	0
7	72.700	0.945	7.433	10.867	7.293	1	0	0	0	0	0	0	0	0
8	72.600	0.907	7.396	11.647	7.252	1	0	0	0	0	0	0	0	0
9	73.400	0.929	7.355	10.643	7.198	0	1	0	0	0	0	0	0	0
10	73.300	0.908	7.337	10.906	7.198	1	0	0	0	0	0	0	0	0
11	73.900	0.914	7.265	10.796	7.102	0	1	0	0	0	0	0	0	0

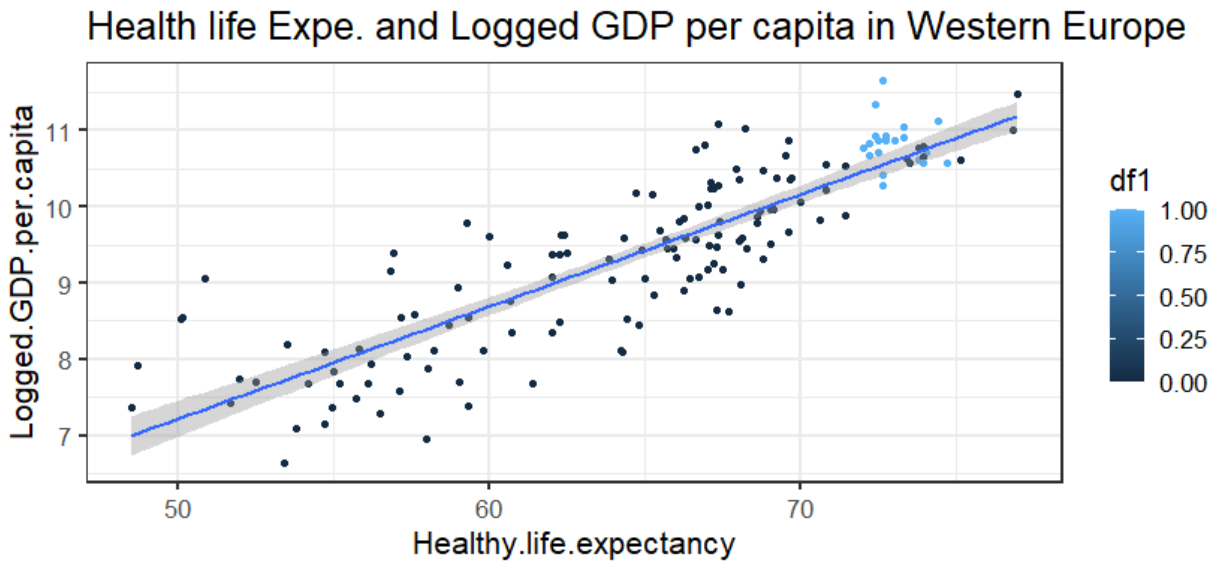
Subset 1: Plotted a regression line between the healthy life expectancy and the freedom to make life choices in the region of western Europe. Black dots signify the 0 whereas blue dots signify the 1 the dependent variable Health life expectancy depends on the freedom to make life choices.



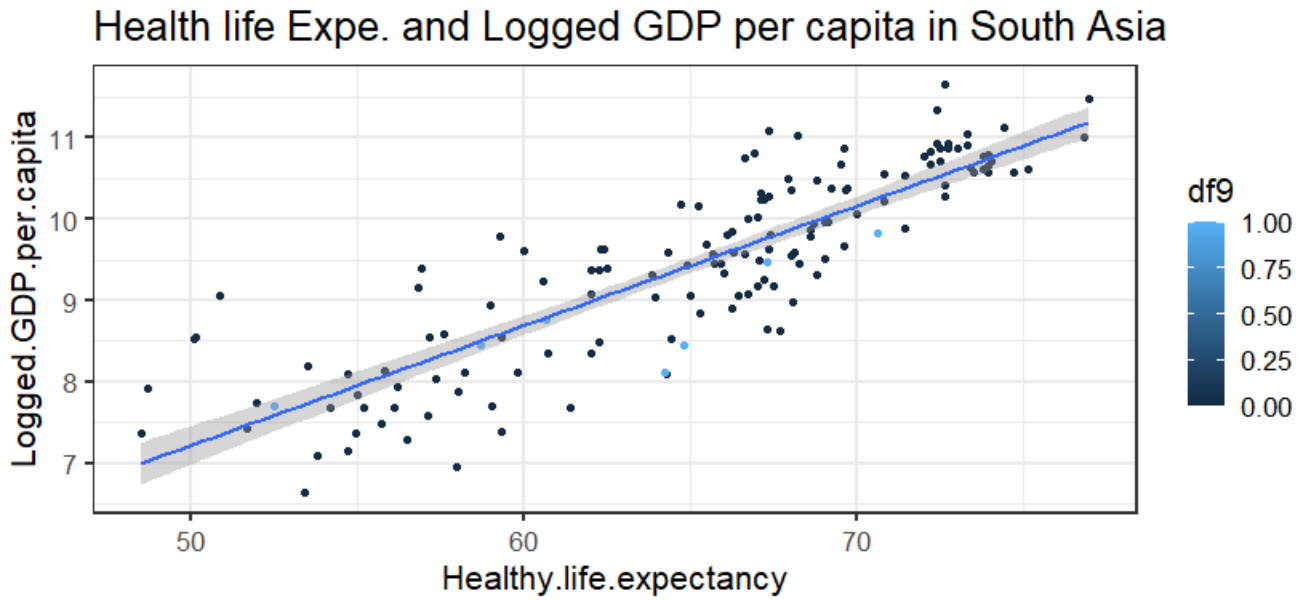
Subset2: Plotted a regression line between the healthy life expectancy and the freedom to make life choices in the region of south asia. Black dots signify the 0 whereas blue dots signify the 1 the dependent variable Health life expectancy depends on the freedom to make life choices.



Subset3: Plotted a regression line between the healthy life expectancy and the freedom to make life choices in the region of western europe. Black dots signify the 0 whereas blue dots signify the 1 the dependent variable Health life expectancy depends on the Logged GDP per capita.



Subset4: Plotted a regression line between the healthy life expectancy and the freedom to make life choices in the region of south asia. Black dots signify the 0 whereas blue dots signify the 1 the dependent variable Health life expectancy depends on the logged GDP per capita choices.



Summary:

The hypothesized relationship between the dependent and independent variables may be seen in the graphical representation of the world happiness data. By finding the interrelation between the dependent and independent variable analysts are able to forecast the behavior of the variable in the future. As per the requirement, we have first built a dummy variable from the regional indicator in this dataset. It aids in the analysis of the dependent and independent variables' regression. After that, we constructed a subset that allows us to gather data according to the requirements and another subset that allows us to examine the categorical variables on the regression.

In subset 1 we can see the relationship between the Western European region and Life expectancy. In the Western European region, I plotted a regression line between healthy life expectancy and freedom to make life choices. The dependent variable is represented by black dots, whereas the independent variable is represented by blue dots. The ability to make life choices affects life expectancy.

Whereas second subset 2 gives the understanding of the South Asia region In the South Asian area, I plotted a regression line between healthy life expectancy and life choice freedom. The dependent variable is represented by black dots while the independent variable is represented by blue dots. The ability to make life choices has an impact on life expectancy.

In the third subset 3, could be able to forecast the relationship between healthy life expectancy and logged GDP per capita. In the Western European region, I plotted a regression line between healthy life expectancy and freedom to make life choices. The dependent variable is represented by black dots, whereas the independent variable is represented by blue dots.

At last, defining the relationship between South Asia the dots are less dependent in south Asia than in western Europe. In the South Asian area, I plotted a regression line between healthy life expectancy and freedom to make life choices. The dependent variable is represented by black dots, whereas the independent variable is represented by blue dots. The recorded GDP per capita decisions affect health life expectancy.

References:

[1] How To Add Regression Line per Group to Scatterplot in ...

In this tutorial, we will learn how to add regression lines per group to scatterplot in R using ggplot2. In ggplot2, we can add regression lines using geom_smooth() function as additional layer to an existing ggplot2.

https://datavizpyr.com/add-regression-line-per-group-to-scatterplot-in-r

[2] What is a Regression Line? - Definition: Meaning: Example

https://www.myaccountingcourse.com/accounting-dictionary/regression-line#:~:text=Definition%3A%20In%20statistics%2C%20a%20regression,trend%20of%20a%20given%20data.

[3] NA values when regressing with dummy variable interaction term

Daniel ChoDaniel Cho 72711 gold badge66 silver badges1212 bronze badges & matteomatteo 28111 silver badge77 bronze badges

https://stackoverflow.com/questions/47976109/na-values-when-regressing-with-dummy-variable-interaction-term

[4] RPubs

just a test see http://rstudio-pubs-static.s3.amazonaws.com/25030_8e9c9ffc3b3c423d9381d81543423502.html

https://rpubs.com/TopQuirk67

[5] www.uncp.edu

Created Date: 1/30/2013 12:36:44 PM

https://www.uncp.edu/sites/default/files/2021-12/2012_W-2_Instructions.pdf

[6] How to Create Dummy Variables in R (with Examples)

Erik Marsja et al.

https://www.marsja.se/create-dummy-variables-in-r/

Appendix:

```
library(ggplot2)
library(tidyverse)
theme_set(theme_bw(base_size=16))


df<- read.csv("world_happiness_report_20220329_csv.csv")
df

summary(df)
str(df)
#Created Scatterplot with multiple regression line
df%>%
  ggplot(aes(x=Healthy.life.expectancy,y=Freedom.to.make.life.choices, color = Regional.indicator))+
  geom_point()

df%>%
  ggplot(aes(x=Healthy.life.expectancy,y=Freedom.to.make.life.choices, color = Regional.indicator))+
  geom_point()+
  geom_smooth(method = "lm",se=FALSE)
sc_plot+
  geom_smooth(method = "lm")

 # Created Dummy Varibale for subset the dataset
df1<-ifelse(df$Regional.indicator=="Western Europe",1,0)
df2<-ifelse(df$Regional.indicator=="North America and ANZ",1,0)
df3<-ifelse(df$Regional.indicator=="Middle East and North Africa",1,0)
df4<-ifelse(df$Regional.indicator=="Latin America and Caribbean",1,0)
df5<-ifelse(df$Regional.indicator=="Central and Eastern Europe",1,0)
df6<-ifelse(df$Regional.indicator=="East Asia",1,0)
df7<-ifelse(df$Regional.indicator=="Commonwealth of Independent States",1,0)
df8<-ifelse(df$Regional.indicator=="Sub-Saharan Africa",1,0)
df9<-ifelse(df$Regional.indicator=="South Asia",1,0)
df_dummy <- data.frame(Healthy.life.expectancy=df$Healthy.life.expectancy,
                       Freedom.to.make.life.choices=df$Freedom.to.make.life.choices,
                       upperwhisker=df$upperwhisker,
                       Logged.GDP.per.capita=df$Logged.GDP.per.capita,
                       lowerwhisker=df$lowerwhisker,
                       df1=df1, df2=df2,df3=df3,df4=df4,df5=df5,df6=df6,df7=df7,df8=df8,df9=df9)
```

```r
# Subset-1Health life Expe. and Freedom choice in Western Europe
sc_plot<- df_dummy%>%
  ggplot(aes(x= Healthy.life.expectancy , y= Freedom.to.make.life.choices, color=df1 ))+
  geom_point()
sc_plot+
  geom_smooth(method = "lm")+
  labs( title = "Health life Expe. and Freedom choice in Western Europe")

#Subset-2: Health life Expe. and Freedom choice in South Asia
sc_plot<- df_dummy%>%
  ggplot(aes(x= Healthy.life.expectancy , y= Freedom.to.make.life.choices, color=df9 ))+
  geom_point()
sc_plot+
  geom_smooth(method = "lm")+
  labs( title = "Health life Expe. and Freedom choice in South Asia")


#Subset-3: Health life Expe. and Logged GDP per capita in Western Europe
sc_plot<- df_dummy%>%
  ggplot(aes(x= Healthy.life.expectancy , y= Logged.GDP.per.capita, color=df1 ))+
  geom_point()
sc_plot+
  geom_smooth(method = "lm")+
  labs( title = "Health life Expe. and Logged GDP per capita in Western Europe")

#Subset-4: Health life Expe. and Logged GDP per capita in South Asia
sc_plot<- df_dummy%>%
  ggplot(aes(x= Healthy.life.expectancy , y= Logged.GDP.per.capita, color=df9 ))+
  geom_point()
sc_plot+
  geom_smooth(method = "lm")+
  labs( title = "Health life Expe. and Logged GDP per capita in South Asia")




#============================Finish=============================================
```