Data Visualization Assignment

Haleh Hashemi, 216498834

Nadia Mehdizadeh, 216252694

Department of Psychology, York University

PSYC 3031 : Intermediate statistics laboratory

Dr. Monique Herbert

November 4, 2020

Description of dataset:
The dataset we use in this visualization assignment demonstrate the happiness scores of countries around the world measured by the Gallup World Poll on a scale from 0 to 10. In addition, this dataset includes six other factors which may contribute to the overall life evaluation. These factors include economic production (GDP), social support, healthy life expectancy, freedom to make life choices, absence of corruption and generosity and are on a scale from 0 to 2. Additionally, we added an extra variable to this dataset called Region which specifies which region each country belongs to and will further assist us in answering our visualization question. All the variables in this dataset are numerical except for Country and Region which are categorical.

Visualization question:
Through this visualization, we are aiming to explore the relationship between happiness, social support and life expectancy among three world regions; Western Europe, Latin America and Sub Saharan Africa. Our initial visualization question is if social support and life expectancy are good predictors of a nation's happiness. Moreover, is there any correlation between the regional location of a country and its state of happiness? Finally, are countries in the same regions rank similarly in these three variables?

Goals/outcomes
Through this visualization, we are trying to show any possible patterns that there are between these three measured variables. Although they do not impact the happiness score, social support and healthy life expectancy may explain why some regions rank higher and others lower. Since the data was so large and country-based, we decided to compare world regions to be able to better visualize the information. Thus we categorized all countries into world regions and chose Western Europe, Latin America and Sub Saharan Africa to demonstrate graphically. We then randomly selected several countries to represent each region. The reason we chose these regions is due to their unique culture and lifestyle.

In order to have an understanding of how these two factors are correlated with happiness score across the world, we first plotted two scatter plots. According to Graph 1 and 2, happiness score is positively correlated with social support and healthy life expectancy. Furthermore, to compare all the three variables at once, a bubble plot was created for all countries worldwide. As shown in Graph 3, the relationship between social support, life expectancy and happiness score is fairly linear. Finally, to answer our visualization questions, Graph 4 was plotted to illustrate the relationship between out three variables among three selected regions.

According to Graph 4, it can be understood that based on the accumulation of circles, the region of Western Europe not only ranks higher in all three variables, but it also has the most congested data; meaning that Western European countries are similar to each other in terms of these variables.
On the other hand, Sub Saharan African countries rank the lowest in terms of all the variables and they also have the most scattered data; which means there is a significant difference between each country of this region. Finally, Latin American countries lie in between the two other regions.

Limitations:

Bubble plots are good means to visualize datasets with three numerical variables and one categorical variable. Since one of the numerical variables (happiness score) is shown with circles with different sizes, it is difficult to quantify the data as their actual values. Therefore, we can only interpret the size of happiness scores across regions and see which regions rank higher; however, we cannot know the actual scores that each region has. Moreover, as shown in Graph 3, bubble graphs are not a good fit for visualizing datasets with large amounts of data as there will be too many overlaps so the distinction between scores are hard to see. Finally, bubble plots limit us in terms of the number of variables that can be used since they are only capable of comparing three numerical variables at once.

The graph was plotted slightly different each time, due to random selection of the representative countries. Hence there was an issue of consistency that limited us in terms of interpreting the results.

DataVisualizationAssignment.R

macbookpro

2020-11-03

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────────────────────
──────── tidyverse 1.3.0 ──
```

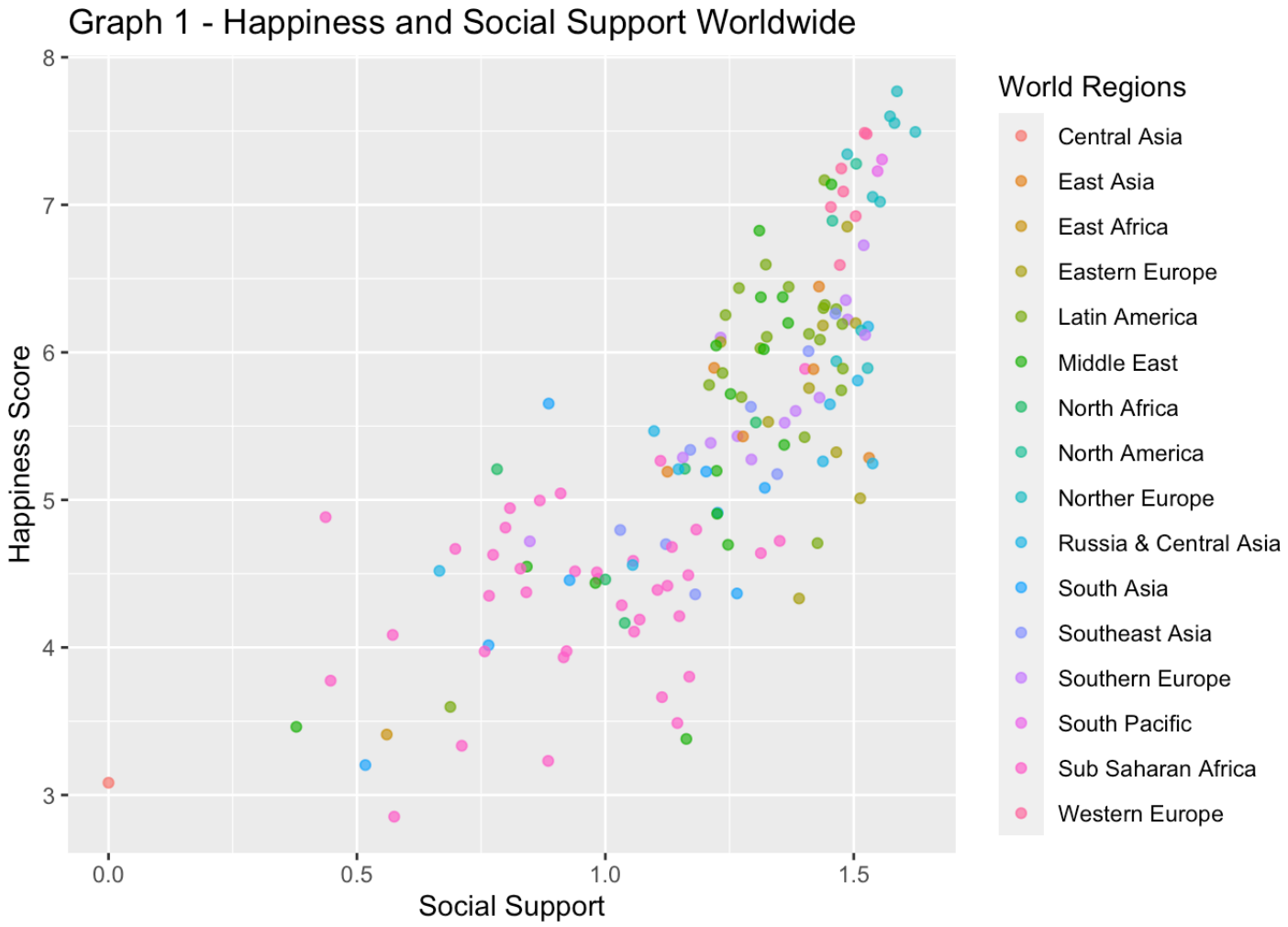
```
## ✓ ggplot2 3.3.2     ✓ purrr   0.3.4
## ✓ tibble  3.0.3     ✓ dplyr   1.0.2
## ✓ tidyr   1.1.2     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.5.0
```

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

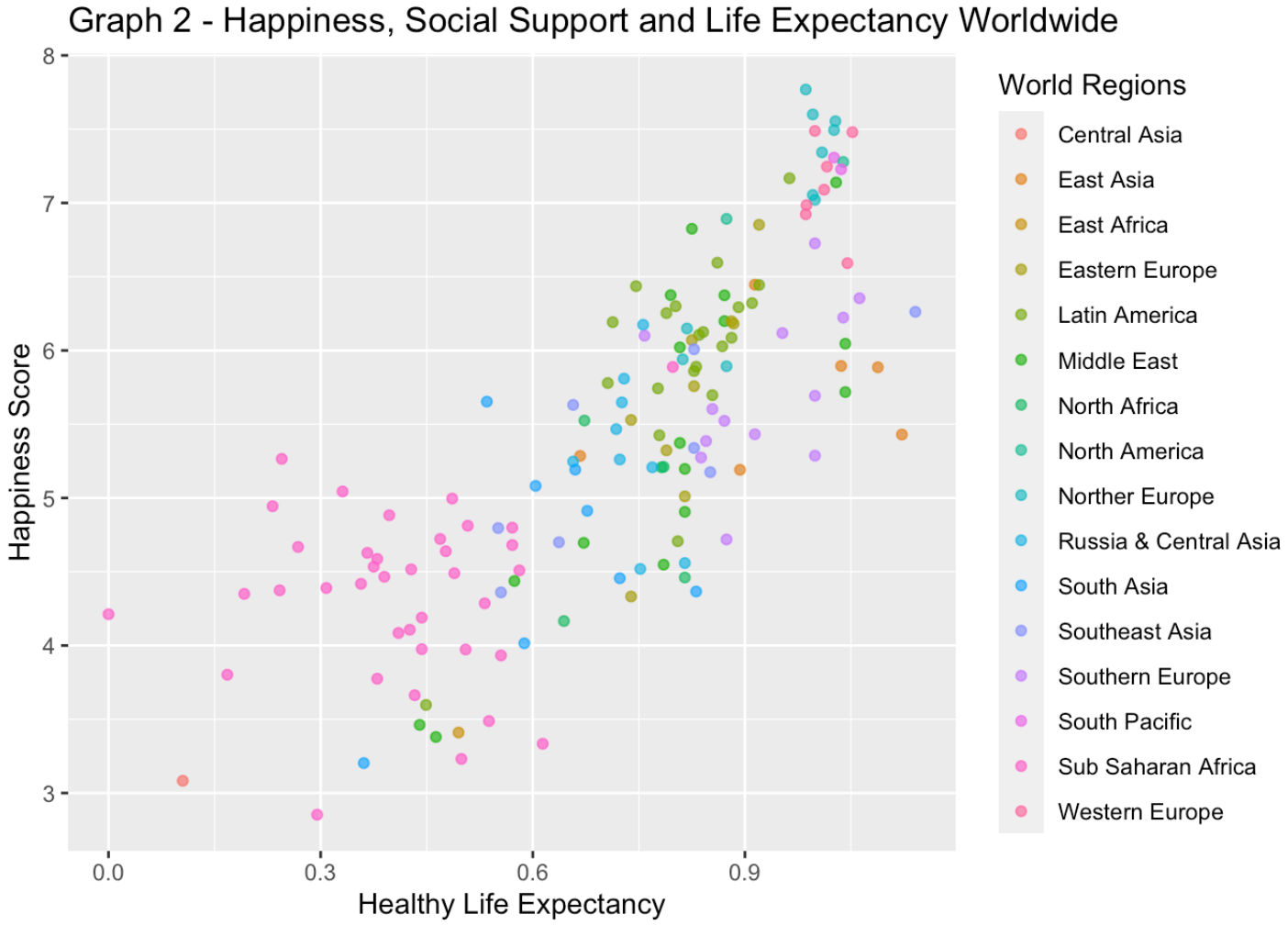
```
library(ggplot2)
#Importing the data set into RStudio, which we named "WorldHappiness"
WorldHappiness<-read_csv("Data/2019(3).csv")
```

```
## Parsed with column specification:
## cols(
##   Overall_rank = col_double(),
##   Region = col_character(),
##   Country = col_character(),
##   Score = col_double(),
##   GDP = col_double(),
##   Social_support = col_double(),
##   Healthy_life_expectancy = col_double(),
##   Freedom = col_double(),
##   Generosity = col_double(),
##   Perceptions_of_corruption = col_double()
## )
```

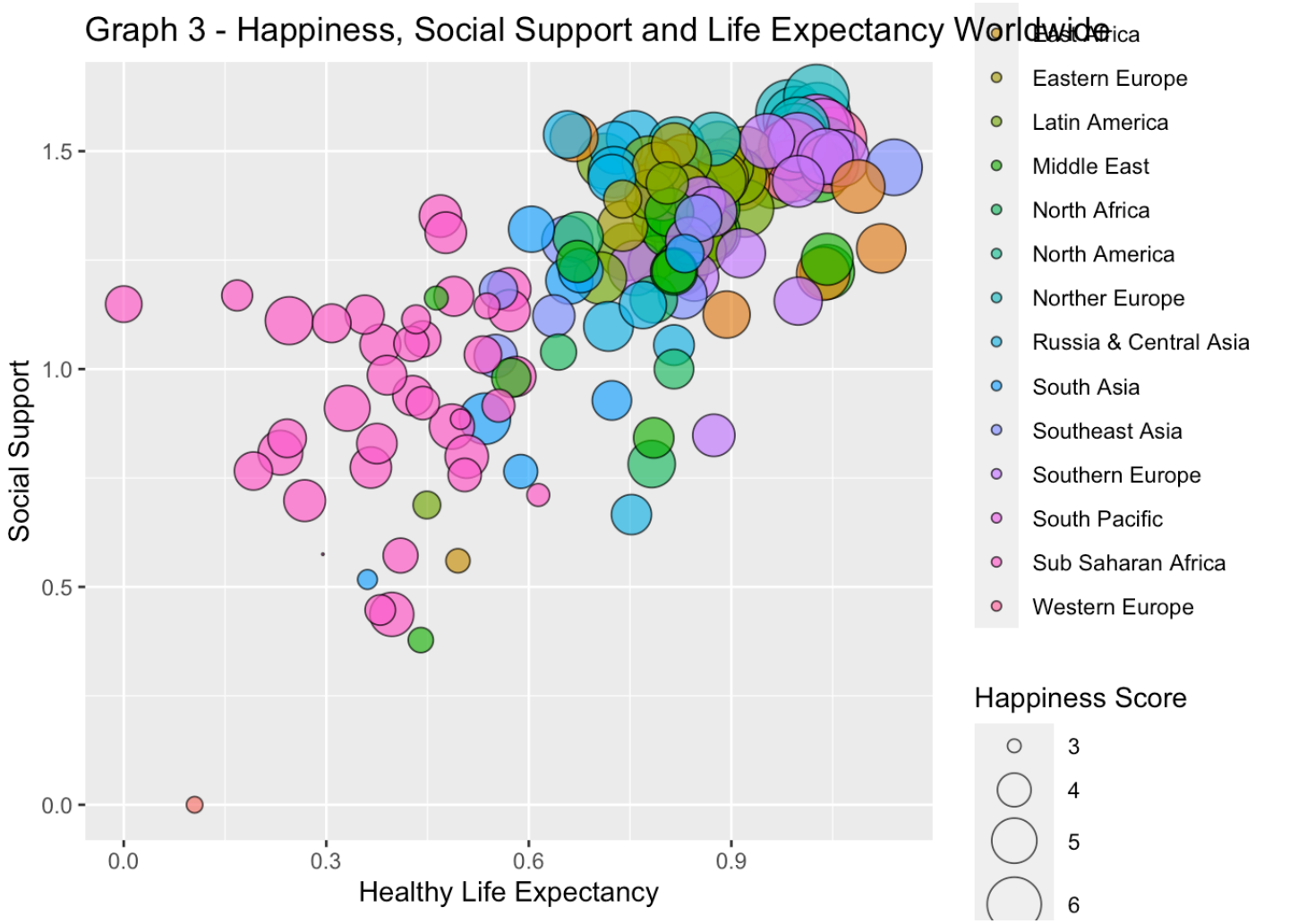
```
#Transforming the acronyms into full name in the Region column
WorldHappiness<- mutate(WorldHappiness,
                        Region = as.character(Region),
                        Region = fct_recode(Region,"Middle East"="ME","Western Europe
"= "WEU","Sub Saharan Africa" = "SSA", "Latin America"="LAC", "North Africa"="NAF","S
outh Pacific"="SP","Southeast Asia"="SEA","East Asia"="EA","South Asia"="SA","Russia
& Central Asia"="RCA","North America"="NAM", "Norther Europe"="NEU","Eastern Europe"=
"EEU","Southern Europe"="SEU","Central Asia"="CA","East Africa"="EAF"))
#Viewing the data set
view(WorldHappiness)
#Separating countries in the Western Europe region
WesternEurope<-filter(WorldHappiness, Region=="Western Europe")
#Randomly selecting 7 countries to represent Western Europe
WesternEurope<-sample_n(WesternEurope,7)
#Separating countries in the Latin America region
LatinAmerica<-filter(WorldHappiness, Region=="Latin America")
#Randomly selecting 7 countries to represent Latin America
LatinAmerica<-sample_n(LatinAmerica,7)
#Separating countries in the Sub Saharan region
SubSaharanAfrica<-filter(WorldHappiness,Region=="Sub Saharan Africa")
#Randomly selecting 7 countries to represent Sub Saharan Africa
SubSaharanAfrica<-sample_n(SubSaharanAfrica,7)
#Combining all randomly selected countries into one data set
RegionHappiness<-rbind(WesternEurope,LatinAmerica,SubSaharanAfrica)
#Plotting a scatter plot to show the relationship between social support and happines
s worldwide
ggplot(WorldHappiness, aes(Social_support, Score, colour = Region)) +
  geom_point(alpha=0.7)+labs(title="Graph 1 - Happiness and Social Support Worldwide
",
                           x= "Social Support",
                           y="Happiness Score",
                           color="World Regions")
```



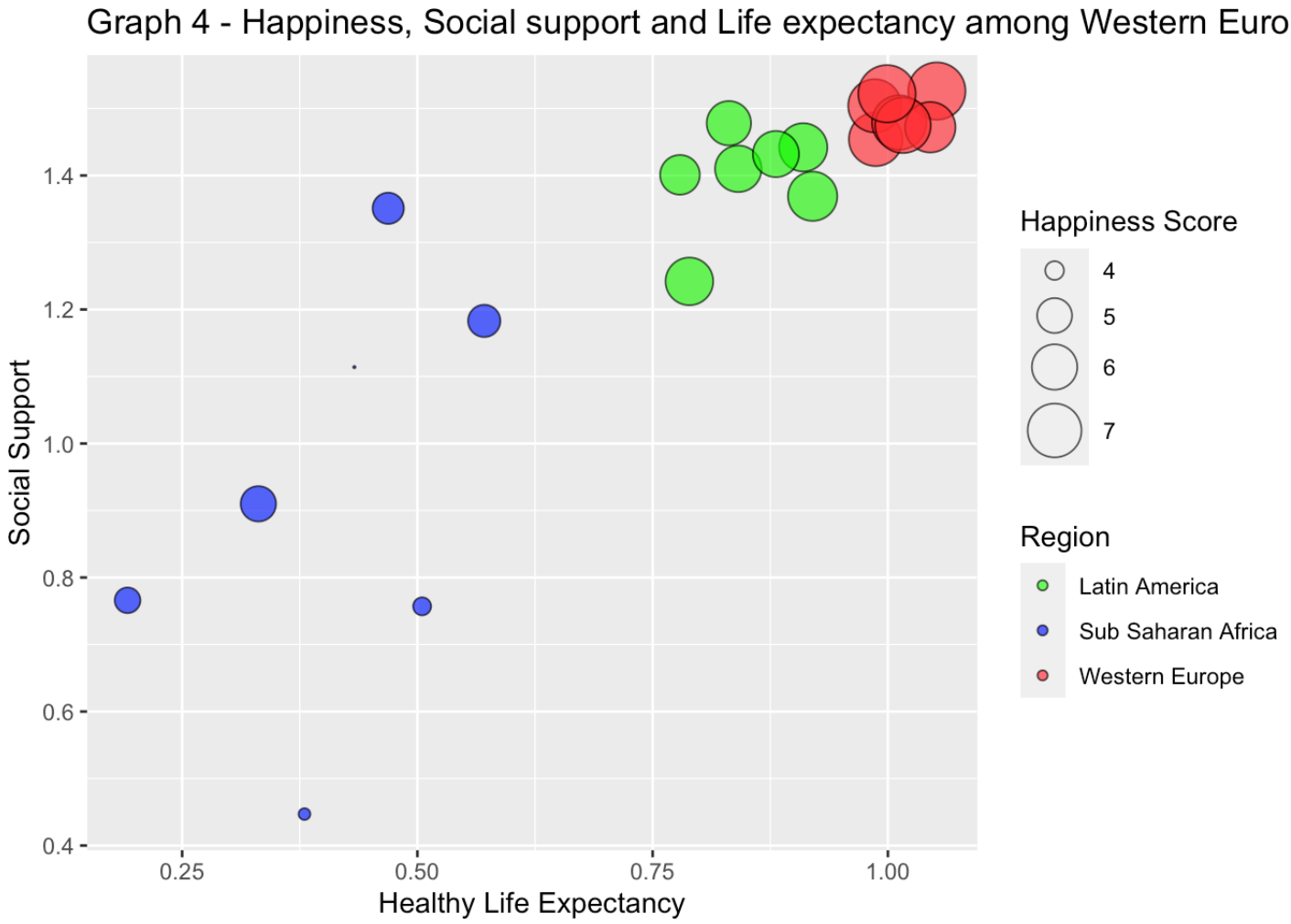

```
##Plotting a scatter plot to show the relationship between healthy life expectancy an
d happiness worldwide
ggplot(WorldHappiness, aes(Healthy_life_expectancy, Score, colour = Region)) +
  geom_point(alpha=0.7)+labs(title="Graph 2 - Happiness, Social Support and Life Expe
ctancy Worldwide",
                           x= "Healthy Life Expectancy",
                           y="Happiness Score",
                           color="World Regions")
```




```
#Plotting a bubble plot to show the relationship among social support, healthy life e
xpectancy and happiness across the world
WorldHappiness %>%
  ggplot(aes(x=Healthy_life_expectancy, y=Social_support, size=Score, fill=Region)) +
  geom_point(alpha=0.7, shape=21, color="Black")+
  scale_size(range = c(.1, 12), name="Happiness Score") +
  labs(title="Graph 3 - Happiness, Social Support and Life Expectancy Worldwide",
       x= "Healthy Life Expectancy",
       y="Social Support",
       color="World Regions")
```




```
#Plotting a bubble plot to show the relationship among social support, healthy life e
xpectancy and happiness across Western Europe, Latin America and Sub Saharan Africa
RegionHappiness%>%
  ggplot(aes(x=Healthy_life_expectancy, y=Social_support, size=Score, fill=Region)) +
  geom_point(alpha=0.7, shape=21, color="Black")+
  scale_size(range = c(.1, 10), name="Happiness Score") +
  scale_fill_manual(values = c("#00FF00", "#0033FF", "#FF3333"))+
  labs(title="Graph 4 - Happiness, Social support and Life expectancy among Western E
urope, Latin America and Sub Saharan Africa",
       x= "Healthy Life Expectancy",
       y="Social Support",
       color="World Regions")
```


Graph 4 - Happiness, Social support and Life expectancy among Western Europ
Social Support
1.4
1.2
1.0
0.8
0.6
0.4
0.25
0.50
0.75
1.00
Healthy Life Expectancy
Happiness Score
4
5
6
7
Region
Latin America
Sub Saharan Africa
Western Europe