MATH 208 Final Project

Chuqi Wang 260842578

Preparation: Setting working directory and loading the data.

```
setwd("~/Desktop/MATH 208/Final Project") # set working directory.
library(readr) # install readr package to read data.
# assign Happiness data.
Happiness <- read_csv("~/Desktop/MATH 208/Final Project/Project_Happiness_data.csv")
head(Happiness) # view first 6 rows of the data.
```

```
# A tibble: 6 x 9
  Country Region GDPperCap TrustGov Family Freedom HealthLifeExp Generosity
  <chr>   <chr>      <dbl>    <dbl>  <dbl>   <dbl>         <dbl>      <dbl>
1 Switze~ Weste~      1.40    0.420   1.35   0.666         0.941      0.297
2 Iceland Weste~      1.30    0.141   1.40   0.629         0.948      0.436
3 Denmark Weste~      1.33    0.484   1.36   0.649         0.875      0.341
4 Norway  Weste~      1.46    0.365   1.33   0.670         0.885      0.347
5 Canada  Ameri~      1.33    0.330   1.32   0.633         0.906      0.458
6 Finland Weste~      1.29    0.414   1.32   0.642         0.889      0.234
# ... with 1 more variable: Year <dbl>
```

Task 1: Family, Generosity and Freedom by Region in 2019.

Part(a) Using the appropriate plots and numerical summaries, evaluate the strength of the association of each of the six scores with the region variable.

For variable Economic Production score(GDPperCap), the density plot, box plot and summary table are shown below.

```
library(tidyverse) #install tidyverse package to make plots and summaries.
# change the levels names of the Region to shorthand version and choose data only in year 2019.
Happiness2019 = Happiness %>% filter(Year==2019) %>%
  mutate(short_region=fct_recode(Region, "Saharan Africa"="Sub-Saharan Africa",
                                 "ME/N Africa"="Middle East and Northern Africa",
                                 "C/E Europe"="Central and Eastern Europe"))
# Density plot.
GDP_density = ggplot(Happiness2019, aes(x=GDPperCap, group=short_region))+
  ggtitle("Density Plot")+geom_histogram(aes(y=..density..,fill=short_region))+
  geom_density()+facet_wrap(~short_region)+xlab("GDPperCap score")
# Box plot.
GDP_boxplot = ggplot(Happiness2019, aes(x=short_region, y=GDPperCap))+
  stat_boxplot(geom = "errorbar",width=0.35)+geom_boxplot()+
```

```
   xlab("Region")+ylab("GDPperCap score")+ggtitle("Box Plot")
library(ggpubr) # install ggpubr package to combine ggplots together.
# Combine two plots together.
ggarrange(GDP_density, GDP_boxplot, nrow=2)
```

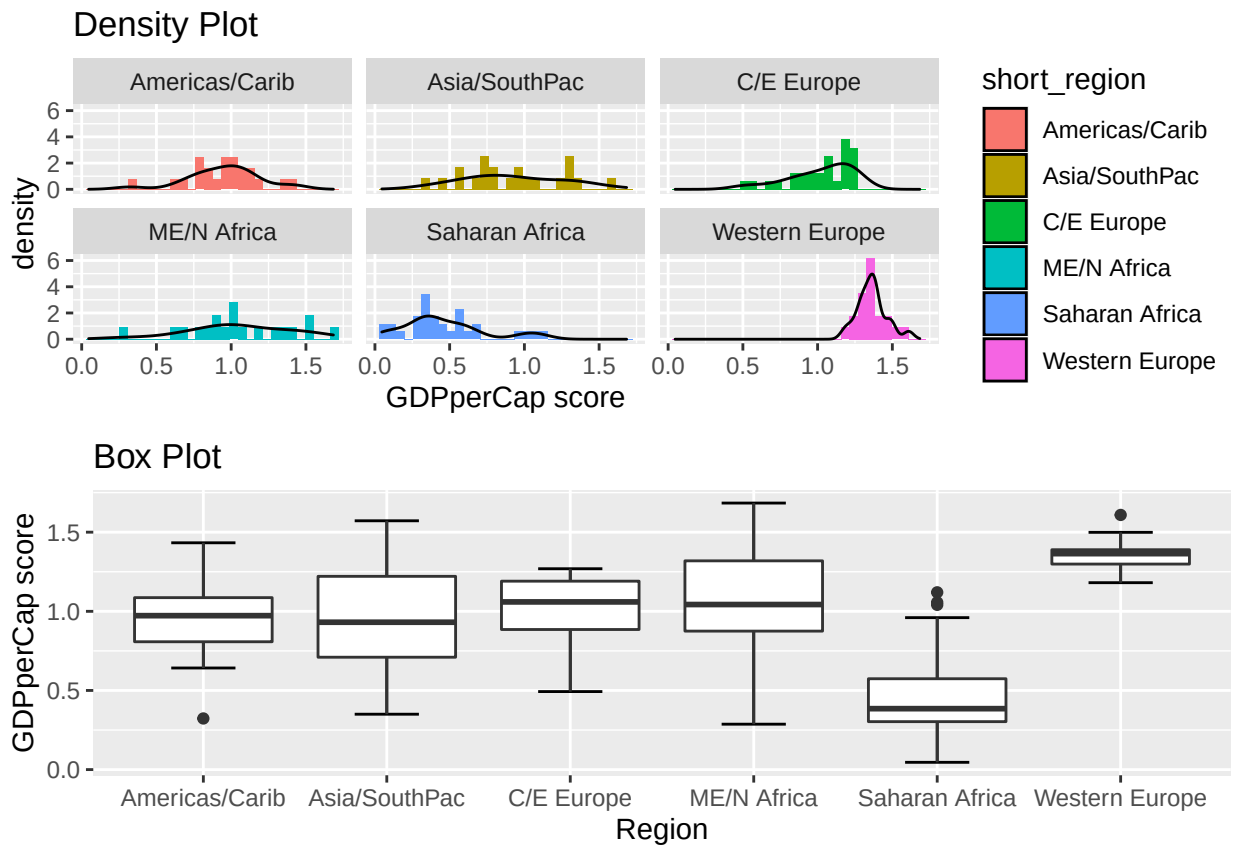


Figure 1: GDPperCap vs. Region Plots in Year 2019

```
library(knitr) # install knitr package to make tables.
# summary table.
GDP_summary = Happiness2019 %>% group_by(Region) %>%
  summarise(n=n(), avg=mean(GDPperCap), med=median(GDPperCap), sd=sd(GDPperCap)) %>%
  arrange(desc(avg))
kable(GDP_summary, digits = 4, caption = "Summary table for GDPperCap in Year 2019",
      col.names = c("Region", "Count", "Average", "Median", "Stardand Deviation"))
```

Table 1: Summary table for GDPperCap in Year 2019

Region	Count	Average	Median	Stardand Deviation
Western Europe	20	1.3620	1.3645	0.0990
Middle East and Northern Africa	19	1.0591	1.0430	0.3474
Central and Eastern Europe	28	1.0224	1.0595	0.2161
Americas/Carib	22	0.9535	0.9725	0.2414
Asia/SouthPac	21	0.9256	0.9310	0.3310
Sub-Saharan Africa	31	0.4662	0.3850	0.2819

From Figure 1 we found that the average GDPperCap score for Western Europe is the highest socre around 1.4 while the average score for Sub-Saharan Africa is the lowest around 0.45 among these six regions and its graph shows right skewed. For the rest of 4 regions, the mean values of score for GDPperCap have the range[0.90, 1.10]. For region Asia/SouthPac and Middle Eastern/Northern Africa, the distribution of the data is wide. The summary table is shown below which contains the total numbers of investigation, average score and median for Economic Production score. Table 1 shows similar association for GDPperCap score with Region, the average score of Western Europe is 1.3620 which is the highest score in six regions while the average score of Sub-Saharan Africa is 0.4662 which is the lowest.

For variable Trust in Government score(TrustGov), the density plot, boxplot and summary table are shown below.

```
# Density plot.
TrustGov_density = ggplot(Happiness2019, aes(x=TrustGov, group=short_region))+
  ggtitle("Density Plot")+geom_histogram(aes(y=..density..,fill=short_region))+
  geom_density()+facet_wrap(~short_region)
# Box plot.
TrustGov_boxplot = ggplot(Happiness2019, aes(x=short_region, y=TrustGov))+
  stat_boxplot(geom = "errorbar", width=0.35)+geom_boxplot()+
   xlab("Region")+ylab("TrustGov score")
ggarrange(TrustGov_density, TrustGov_boxplot, nrow=2)
```

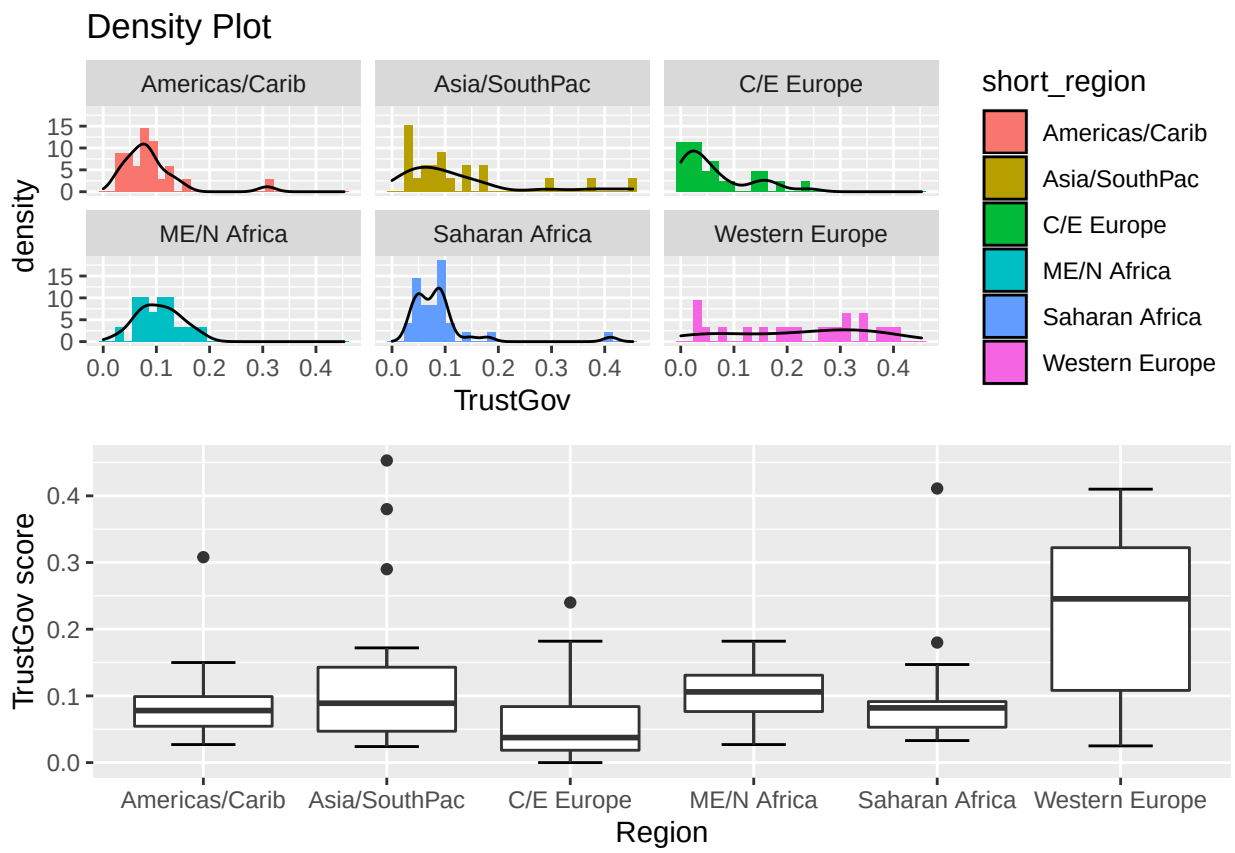


Figure 2: TrustGov vs. Region Plots in Year 2019.

From Figure 2 we found that the average scores for Region are both very low except for Western Europe, the

mean value of TrustGov score is also the highest but the interquartile range is large, while the mean value for Central and Estern Europe is very low and the density plots looks like left skewed. Expect for Middle Estern Africa and Western Europe, there are outliers in boxplots for the other four regions.

```
# summary table.
TrustGov_summary = Happiness2019 %>% group_by(Region) %>%
  summarise(n=n(),avg=mean(TrustGov), med=median(TrustGov), sd=sd(TrustGov)) %>%
  arrange(desc(avg))
kable(TrustGov_summary, digits = 4,
      caption = "Summary table for Trust in Government in Year 2019.",
      col.names = c("Region", "Count", "Average", "Median", "Stardand Deviation"))
```

Table 2: Summary table for Trust in Government in Year 2019.

Region	Count	Average	Median	Stardand Deviation
Western Europe	20	0.2212	0.2455	0.1308
Asia/SouthPac	21	0.1240	0.0890	0.1168
Middle East and Northern Africa	19	0.1050	0.1060	0.0403
Americas/Carib	22	0.0879	0.0780	0.0589
Sub-Saharan Africa	31	0.0876	0.0820	0.0682
Central and Eastern Europe	28	0.0628	0.0375	0.0646

From table 2 we found that the average score for TrustGov of Western Europe is 0.2212 which is the highest comparing to other regions whereas the lowest value for average score is 0.0628 from Central and Eastern Europe. And for region Sub-Saharan Africa, the mean value score is 0.0876 which is the second lowest value.

For variable Family Support score(Family), the density plot, boxplot and summary table are shown below.

```
# Density plot.
Family_density = ggplot(Happiness2019, aes(x=Family, group=short_region))+
  ggtitle("Density Plot")+geom_histogram(aes(y=..density..,fill=short_region))+
  geom_density()+facet_wrap(~short_region)
# Box plot.
Family_boxplot = ggplot(Happiness2019, aes(x=short_region, y=Family))+
  stat_boxplot(geom = "errorbar", width=0.35)+geom_boxplot()+
   xlab("Region")+ylab("Family score")
ggarrange(Family_density, Family_boxplot, nrow=2)
```

```
# summary table.
Family_summary = Happiness2019 %>% group_by(Region) %>%
  summarise(n=n(),avg=mean(Family), med=median(Family), sd=sd(Family)) %>%
  arrange(desc(avg))
kable(Family_summary, digits = 4,
      caption = "Summary table for Family support score in Year 2019.",
      col.names = c("Region", "Count", "Average", "Median", "Stardand Deviation"))
```

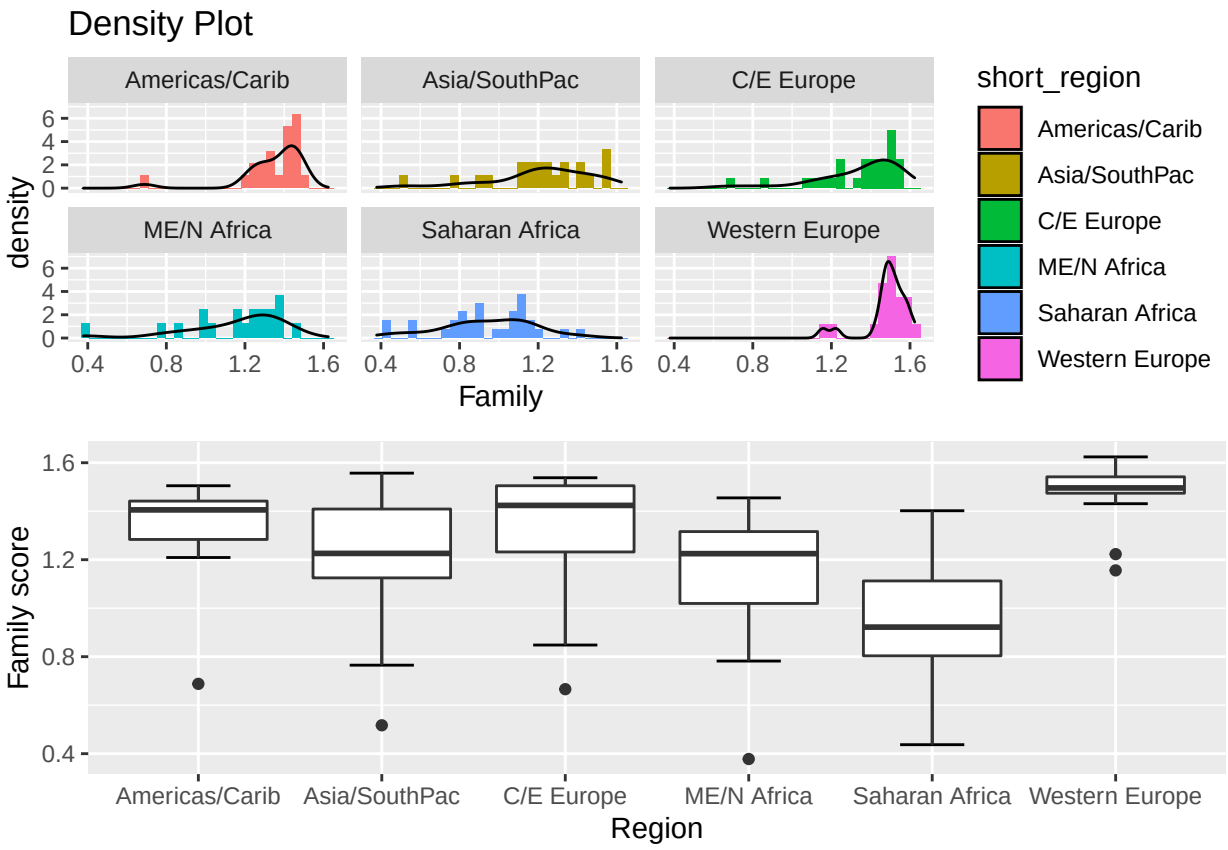


Figure 3: Family vs. Region Plots in Year 2019.

Table 3: Summary table for Family support score in Year 2019.

Region	Count	Average	Median	Stardand Deviation
Western Europe	20	1.4839	1.4960	0.1124
Americas/Carib	22	1.3463	1.4055	0.1721
Central and Eastern Europe	28	1.3404	1.4240	0.2171
Asia/SouthPac	21	1.2140	1.2260	0.2651
Middle East and Northern Africa	19	1.1487	1.2250	0.2627
Sub-Saharan Africa	31	0.9381	0.9220	0.2383

Figure 3 indicates the mean family score for Western Europe is the highest among six regions and the interquartile range is very small. The average score for Saharan Africa is the lowest again and the distribution of the data is wide. The plots of C/E Europe and Americas/Carib show obvious left skewed distributions. Table 3 shows the highest average family score of Western Europe is 1.4839 whereas the lowest score of region Sub-Saharan Africa is 0.9381.

For variable Perceived Freedom score(Freedom), the density plot, box plot and summary table are shown below.

```
# Density plot.
Freedom_density = ggplot(Happiness2019, aes(x=Freedom, group=short_region))+
  ggtitle("Density Plot")+geom_histogram(aes(y=..density..,fill=short_region))+
  geom_density()+facet_wrap(~short_region)
# Box plot.
Freedom_boxplot = ggplot(Happiness2019, aes(x=short_region, y=Freedom))+
  stat_boxplot(geom = "errorbar", width=0.35)+geom_boxplot()+
   xlab("Region")+ylab("Freedom score")
ggarrange(Freedom_density, Freedom_boxplot, nrow=2)
```

From Figure 4 we found the average freedom score for Western Europe and Asia/SouthPac are both very high, the lowest value is shown in Middle Eastern and Northern Africa. However, the boxplot of ME/N Africa indicates much wider distribution than the other plots. In addition, the mean value of freedom score for Sub-Saharan Africa is about 0.35 and the score is still quite small comparing to other regions.

```
Freedom_summary = Happiness2019 %>% group_by(Region) %>%
  summarise(n=n(),avg=mean(Freedom), med=median(Freedom), sd=sd(Freedom)) %>%
  arrange(desc(avg))
kable(Freedom_summary, digits = 4,
      caption = "Summary table for Perceived Freedom score in Year 2019.",
      col.names = c("Region", "Count", "Average", "Median", "Stardand Deviation"))
```

Table 4: Summary table for Perceived Freedom score in Year 2019.

Region	Count	Average	Median	Stardand Deviation
Western Europe	20	0.4826	0.5210	0.1349
Asia/SouthPac	21	0.4588	0.5080	0.1491
Americas/Carib	22	0.4458	0.4805	0.1281
Central and Eastern Europe	28	0.3580	0.3400	0.1190
Sub-Saharan Africa	31	0.3422	0.3520	0.1013
Middle East and Northern Africa	19	0.3179	0.3050	0.1667

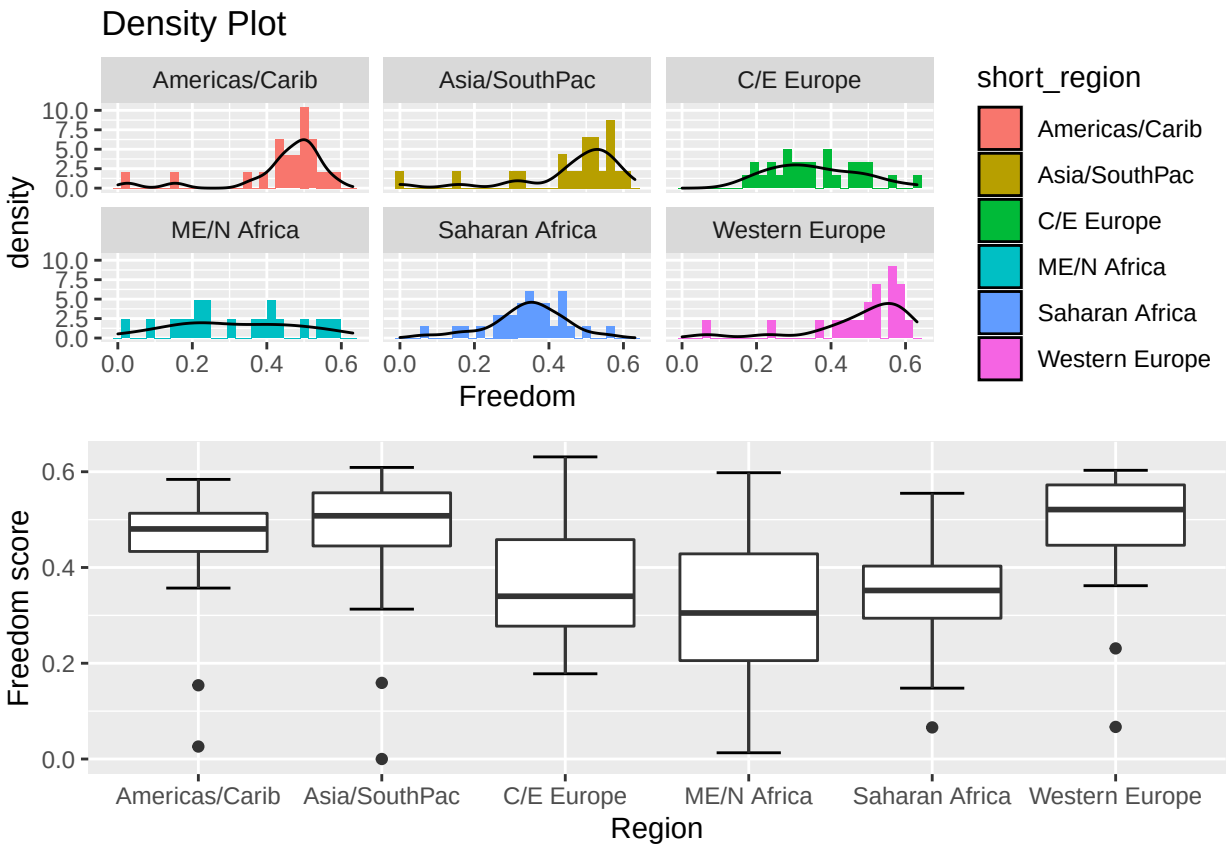


Figure 4: Freedom vs. Region Plots in Year 2019.

Table 4 shows that the average freedom score for Western Europe is the highest and the value is 0.4826 while the average score for Middle East and Northern Africa is 0.3179 and it's the lowest among the six regions. In addition, the mean vlaue of freedom score for Sub-Saharan Africa is the second lowest and has value 0.3422.

For variable Perceived Health and Life Expectancy score(HealthLifeExp), the density plot, box plot and summary table are shown below.

```
# Density plot.
HealthLifeExp_density = ggplot(Happiness2019, aes(x=HealthLifeExp, group=short_region))+
  ggtitle("Density Plot")+geom_histogram(aes(y=..density..,fill=short_region))+
  geom_density()+facet_wrap(~short_region)
# Box plot.
HealthLifeExp_boxplot = ggplot(Happiness2019, aes(x=short_region, y=HealthLifeExp))+
  stat_boxplot(geom = "errorbar", width=0.35)+geom_boxplot()+
   xlab("Region")+ylab("HealthLifeExp score")
ggarrange(HealthLifeExp_density, HealthLifeExp_boxplot, nrow=2)
```

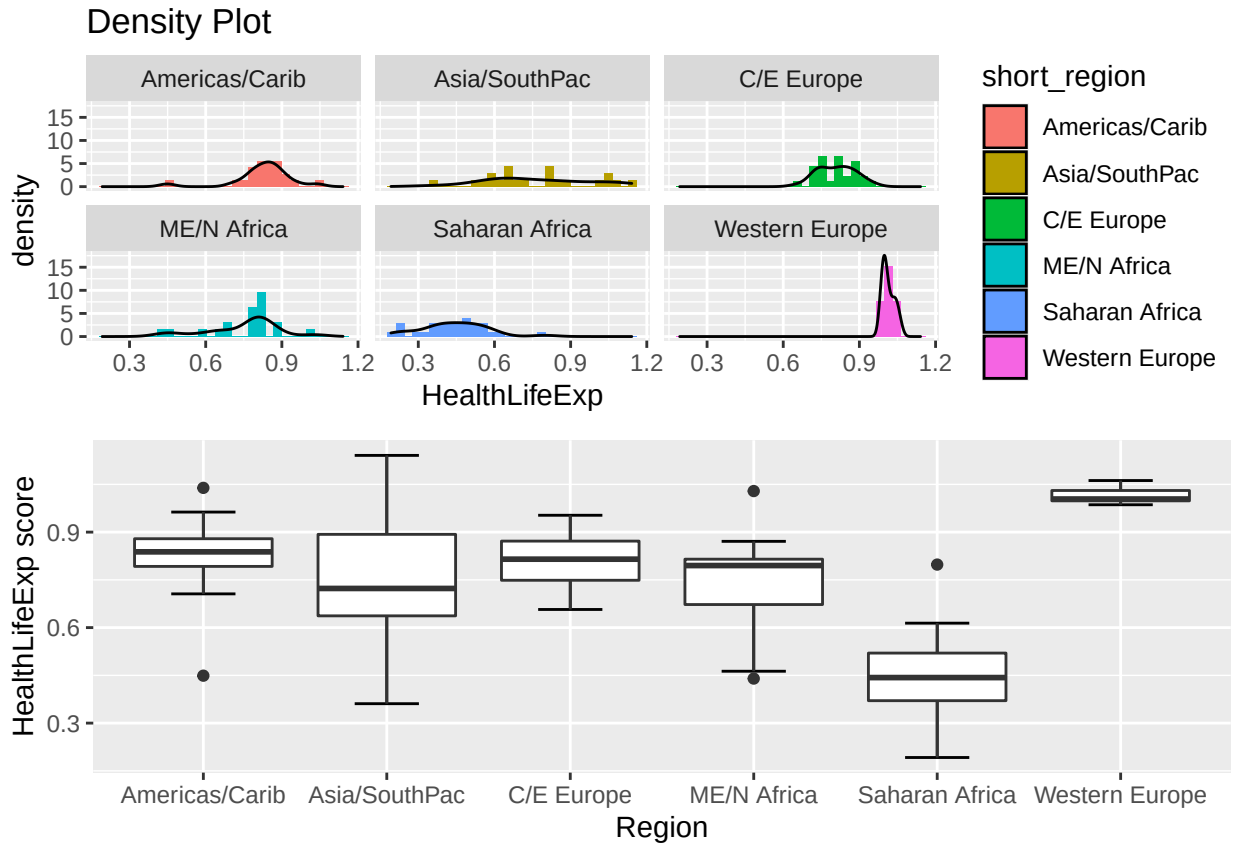


Figure 5: HealthLifeExp vs. Region Plots in Year 2019.

```
# summary table.
Health_summary = Happiness2019 %>% group_by(Region) %>%
  summarise(n=n(),avg=mean(HealthLifeExp), med=median(HealthLifeExp),
            sd=sd(HealthLifeExp)) %>% arrange(desc(avg))
kable(Health_summary, digits = 4,
```

```
      caption = "Summary table for Perceived Health and Life Expectancy score in Year 2019.",
      col.names = c("Region", "Count", "Average", "Median", "Stardand Deviation"))
```

Table 5: Summary table for Perceived Health and Life Expectancy score in Year 2019.

Region	Count	Average	Median	Stardand Deviation
Western Europe	20	1.0138	1.004	0.0236
Americas/Carib	22	0.8295	0.838	0.1120
Central and Eastern Europe	28	0.8085	0.815	0.0730
Asia/SouthPac	21	0.7725	0.723	0.2087
Middle East and Northern Africa	19	0.7511	0.795	0.1434
Sub-Saharan Africa	31	0.4408	0.443	0.1289

Figure 5 shows that the highest average HealthLifeExp score is in region Western Europe and the density plot for this region has concentrated distribution. The average score for Sub-Saharan Africa is the lowest again. The average scores for the other four regions are between 0.75 and 0.80. If we looke at Table 5, the average score for Western Europe is 1.0138 while that for Sub-Saharan Africa is only 0.4408.

For variale Perceived Generosity score(Generosity), the density plot, box plot and summary table are shown below.

```
# Density plot.
Gen_density = ggplot(Happiness2019, aes(x=Generosity, group=short_region))+
  ggtitle("Density Plot")+geom_histogram(aes(y=..density..,fill=short_region))+
  geom_density()+facet_wrap(~short_region)
# Box plot.
Gen_boxplot = ggplot(Happiness2019, aes(x=short_region, y=Generosity))+
  stat_boxplot(geom = "errorbar", width=0.35)+geom_boxplot()+
   xlab("Region")+ylab("Generosity score")
ggarrange(Gen_density, Gen_boxplot, nrow=2)
```

Figure 6 illustrates that the average Generosity scores for both Western Europe and Asia/SouthPac are high, the plots for these two regions has wide distribution. And the average scores for other four regions are below 0.2 which is very small.

```
# summary table.
Gen_summary = Happiness2019 %>% group_by(Region) %>%
  summarise(n=n(),avg=mean(Generosity), med=median(Generosity),
            sd=sd(Generosity)) %>% arrange(desc(avg))
kable(Gen_summary, digits = 4,
      caption = "Summary table for Perceived Generosity score in Year 2019.",
      col.names = c("Region", "Count", "Average", "Median", "Stardand Deviation"))
```

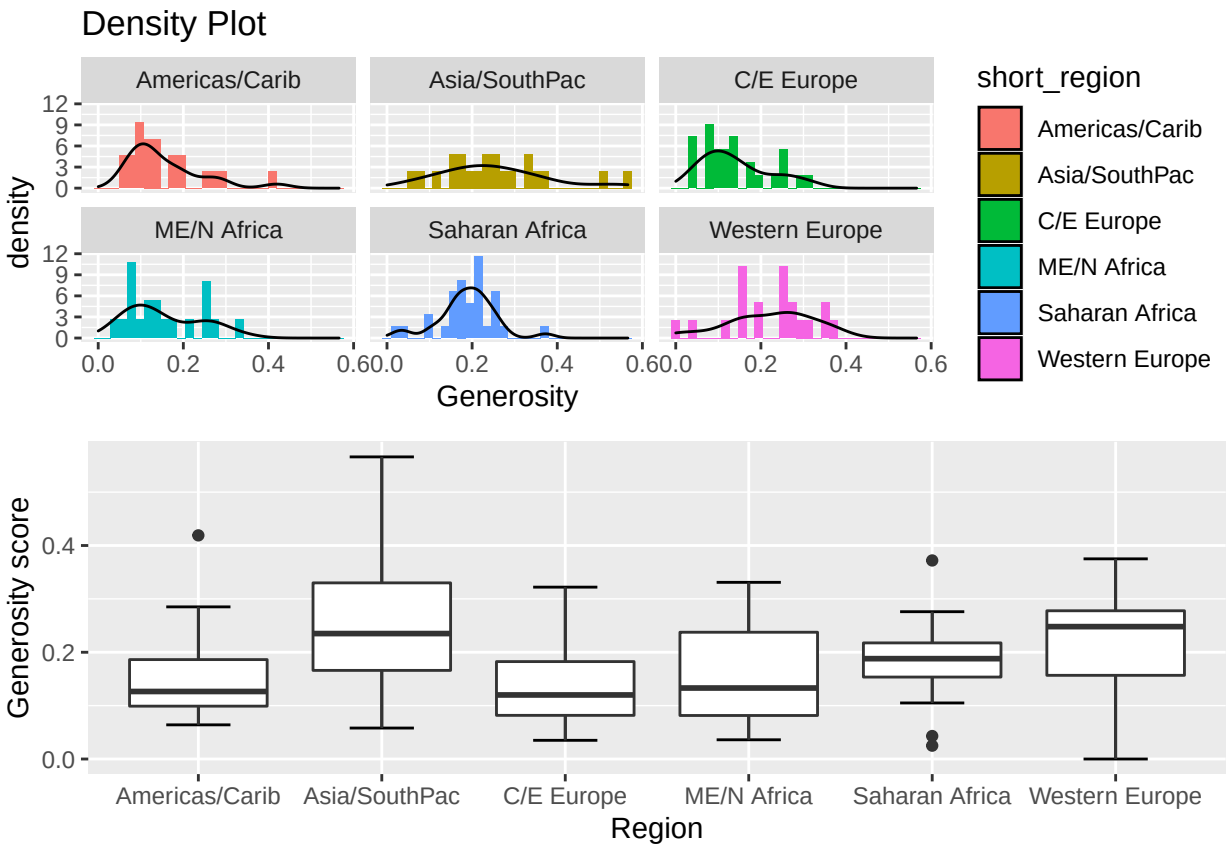


Figure 6: Generosity vs. Regions Plots in Year 2019.

Table 6: Summary table for Perceived Generosity score in Year 2019.

Region	Count	Average	Median	Stardand Deviation
Asia/SouthPac	21	0.2520	0.2350	0.1276
Western Europe	20	0.2210	0.2480	0.1008
Sub-Saharan Africa	31	0.1883	0.1880	0.0668
Americas/Carib	22	0.1553	0.1265	0.0876
Middle East and Northern Africa	19	0.1535	0.1330	0.0869
Central and Eastern Europe	28	0.1412	0.1200	0.0808

The highest average score is 0.2520 from Asia/SouthPac and the second highest score is 0.2210 from Western Europe while for Central and Eastern Europe the average score is only 0.1412.

Part(b) Does the association of between score and region vary amongst the scores or do all the scores seem to seem to show the same pattern? Explain briefly the reasons behind your assessment.

From part(a) we found that the high average scores of the six variables always show in region Western Europe, and the low scores are always from region Sub-Saharan Africa and Middle East and North Africa in 2019. We can't say the association is true for all the six variables. For instance, if we look table 6:Summary table for variable Perceived Generosity score, it shows that the average Generosity score for Sub-Saharan Africa is not very low comparing to average scores from other regions. However, for most of variables scores, the summary table and plots indicate the same pattern which high average scores from Western Europe and low average scores from Sub-Saharan Africa and Middle East and North Africa in Year 2019.

Task 2: Predicted Happiness score

Part(a) Compute the happiness score for each country for each year.

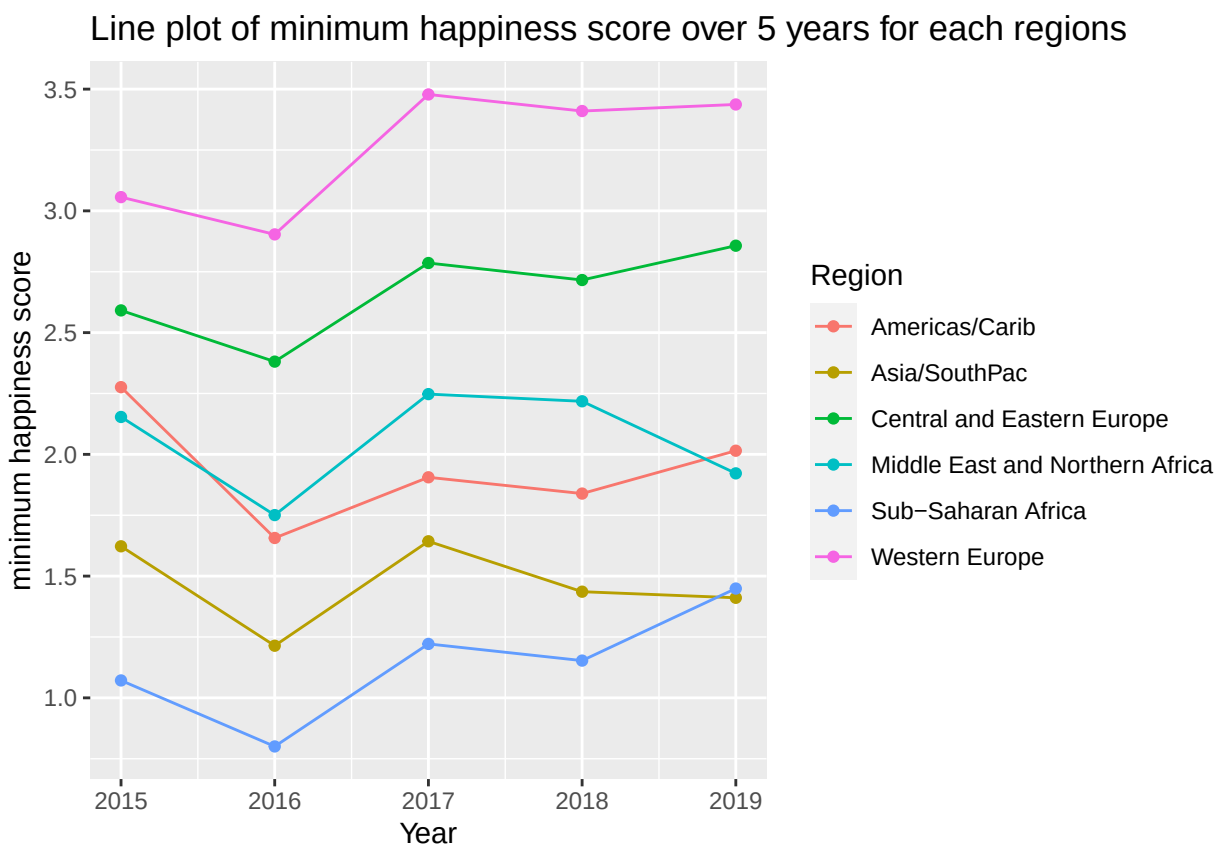
The happiness score for each country for each are shown below in the tibble.

```
Happiness %>% group_by(Year, Country) %>%
  mutate(happiness=GDPperCap+TrustGov+Family+Freedom+HealthLifeExp+Generosity) %>%
  summarise(happiness) %>% pivot_wider(names_from = Year, values_from=happiness)
```

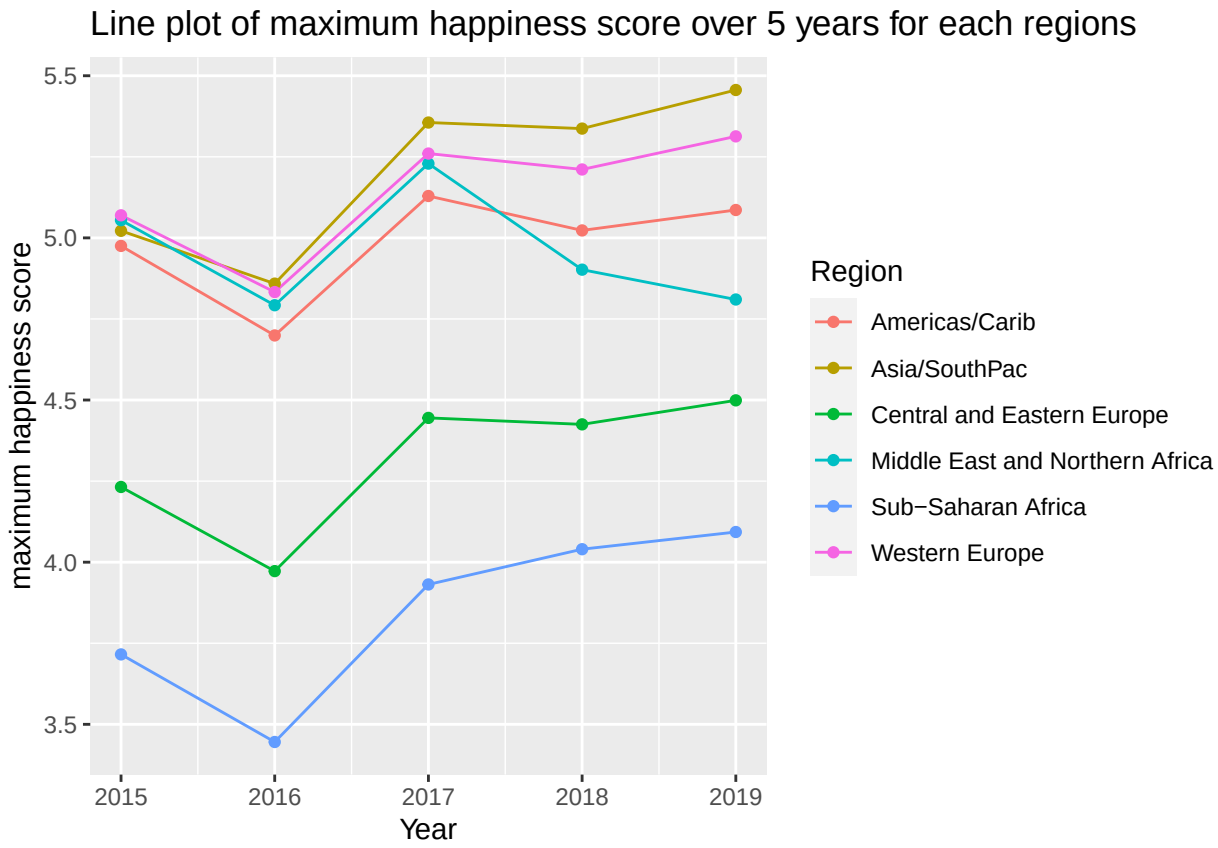
```
## # A tibble: 141 x 6
##    Country     `2015` `2016` `2017` `2018` `2019`
##    <chr>        <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
##  1 Afghanistan   1.62   1.21   1.64   1.44   1.41
##  2 Albania       3.06   2.73   3.15   3.12   3.26
##  3 Algeria       3.17   2.95   3.30   3.09   3.22
##  4 Argentina     3.74   3.52   3.98   3.97   3.99
##  5 Armenia       2.59   2.38   2.85   2.84   3.16
##  6 Australia     5.02   4.77   5.22   5.13   5.14
##  7 Austria       4.67   4.43   4.87   4.82   4.87
##  8 Azerbaijan    3.21   3.02   3.47   3.42   3.53
##  9 Bahrain       4.22   3.94   4.43   4.36   4.50
## 10 Bangladesh    2.18   2.03   2.63   2.84   3.05
## # ... with 131 more rows
```

Part(b) Compute the minimum and maximum happiness score over all countries for each region in each year. In other words, you should have a minimum and maximum happiness score for each region for each year. Plot how the minimum and maximum happiness score for each region varies over time, i.e. with year on the x-axis and lines for the minimum and maximum for each region on the y-axis. Hint: You will want to have multiple panels in your plot, either one each for the minimum and maximum OR one for each region. Do the minimum and/or maximum happiness scores change much over time? Explain your answer.

```
min_max_happiness = Happiness %>% group_by(Year,Region) %>% drop_na() %>%
  mutate(happiness=GDPperCap+TrustGov+Family+Freedom+HealthLifeExp+Generosity) %>%
  summarise(min=min(happiness), max=max(happiness))
# Line plot.
ggplot(min_max_happiness, aes(x=Year, y=min, group=Region, col=Region))+
  geom_line()+geom_point()+
  labs(y="minimum happiness score",
       title = "Line plot of minimum happiness score over 5 years for each regions")
```




```
ggplot(min_max_happiness, aes(x=Year, y=max, group=Region, col=Region))+
  geom_line()+geom_point()+
  labs(y="maximum happiness score",
       title = "Line plot of maximum happiness score over 5 years for each regions")
```



- For minimum happiness score of these 6 regions, the line plot illustrates that for all six regions the minimum happiness scores decrease between 2015 and 2016 and then increase in 2017. From 2017 to 2019, the minimum happiness scores for most of regions have a slight increasing except for Middle Ease and Northern Africa and Asia/SouthPac. Over the five years, the minimum happiness scores for Western Europe ranked first and increase from 3.0 to almost 3.5. The similar increasing was found in Central/Eastern Europe and Sub-Saharan Africa, the minimum scores rise from 2.6 to 2.8 and from 1.1 to 1.5 respectively. For the other regions, the minimum happiness scores have decreasing trend over the five years. In conclusion, for Western Europe, Central and Eastern Europe and Sub-Saharan Africa, the minimum happiness scores increase obviously while or the other three regions, the minimum happiness scores decrease signficantly between 2015 and 2019.
- For maximum happiness score of the 6 regions, the line plot indicates that for all six regions the maximum happiness scores have a signficant decreasing from 2015 to 2016 and then reincrease obviously between 2016 and 2017. Between the period of 2017 and 2019, except for Middle East and Northern Africa, the maximum happiness scores for the other five regions increase slightly. In addition, the maximum scores for Middle East and Northern Africa is the only one that decreased from 5.0 to 4.76 approximately over the half decade while the maximum score for other five regions have an obvious increasing. Overall, except for Americas/Carb region, the maximum happiness scores for the other regions change much between 2015 and 2019.

Task 3: Happiest((well, predicted happiest) countries over time

Part(a) Create a table with the 10 countries who have the highest average happiness scores over the five years of the data.

```
# Use the tibble from task 2(a) and assign it.
country_happiness = Happiness %>% group_by(Year, Country) %>%
  mutate(happiness=GDPperCap+TrustGov+Family+Freedom+HealthLifeExp+Generosity) %>%
  summarise(happiness) %>% pivot_wider(names_from = Year, values_from=happiness)
# Rename the column names.
country_happiness = rename(country_happiness, c("Year2015"="2015", "Year2016"="2016",
                                                "Year2017"="2017","Year2018"="2018",
                                                "Year2019"="2019"))
hightop10= country_happiness %>%
  mutate(avg=(Year2015+Year2016+Year2017+Year2018+Year2019)/5) %>%
  select(Country, avg) %>% arrange(desc(avg)) %>% filter(avg>=4.982)
kable(hightop10, caption = "Top 10 highest average happiness scores table",
      col.names = c("Country", "Average score"), digits = 4)
```

Table 7: Top 10 highest average happiness scores table

Country	Average score
Singapore	5.1613
Norway	5.1348
New Zealand	5.0995
Switzerland	5.0957
Denmark	5.0839
Australia	5.0541
Luxembourg	5.0515
Ireland	5.0447
Sweden	5.0231
Canada	4.9825

Table 7 is shown above that contains the top 10 countries that have highest average happiness score over the five years of the data.

Part(b) Create a table with the 10 countries who have the largest positive change in happiness score from 2015 to 2019, i.e. the 10 countries with the largest positive value of the difference between their 2019 happiness score and their 2015 happiness score.

```
# Use the tibble from task 3(a).
difftop10 = country_happiness %>% mutate(diff = Year2019-Year2015) %>%
  select(Country, diff) %>% arrange(desc(diff)) %>% filter(diff>=0.5979)
kable(difftop10, caption = "Top 10 largest positive change in happiness score table",
      col.names = c("Country", "Change score"), digits = 4)
```

Table 8: Top 10 largest positive change in happiness score table

Country	Change score
Bangladesh	0.8727
Myanmar	0.8200
Kosovo	0.8062
Croatia	0.7676

Country	Change score
Pakistan	0.6424
Montenegro	0.6099
Nepal	0.6040
India	0.6015
Serbia	0.6009
Bosnia and Herzegovina	0.5980

Table 8 is shown above that indicates the top 10 countries who have the largest positive change in happiness score from 2015 to 2019.