Introduction

We explored the 2005-2020 world happiness report data set, which contains several countries' happiness statistics. This data set contains 10 variables, one of which is categorical variable “Country name” and nine of which are numerical variables “Life Ladder”,”Log GDP per capita”,”social support”, “Healthy life expectancy at birth”,”Freedom to make life choices”, “Generosity”,”Perceptions of corruption”, “Positive affect”, “Negative affect”. However, “Positive affect” and “Negative affect” two variables won’t be used in this project because it is useless to analyze countries' happiness. To exclude the time effect, we choose to compare the happiness score of 2008 and 2019.

Background

What is happiness and why is it important? We all know that happiness is significant in people’s daily life. People’s living environment infrastructure, convenient air transportation, personal safety, working environment atmosphere and so on will affect people’s well-being. The most obvious example is the economic crisis of 2008, when people were staying in a very bad situation. Most factories were shutting down because a lot of people were unemployed without income and unable to maintain a normal life. More and more, a huge backlog of goods. A broken credit relation. It was an extremely chaotic and paralyzed socio-economic society in 2008. There is no denying that people’s happiness must not have been high at that time. As time went on, society began to get on track and people were put into a normal state of life. Rate of unemployment is also falling, and people can afford their daily lives without feeling anxiety. Compared to 2008, people now have higher happiness.

Statistical Question of Interest

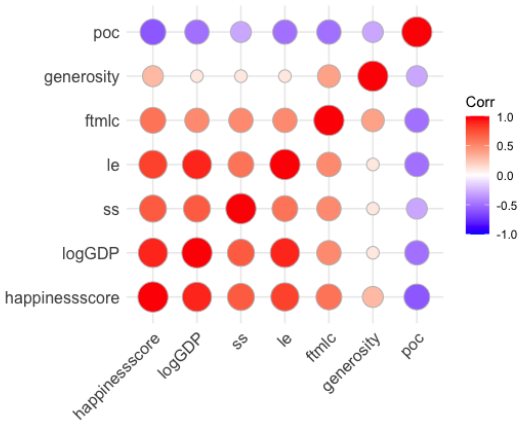
The primary goal of analyzing our dataset is to find which variables will have the most influence on a country's happiness. We are also interested in the percentage change of the top 10 and least 10 countries' happiness in 2008 and 2019 which can help the countries with lower happiness to improve their people's life.

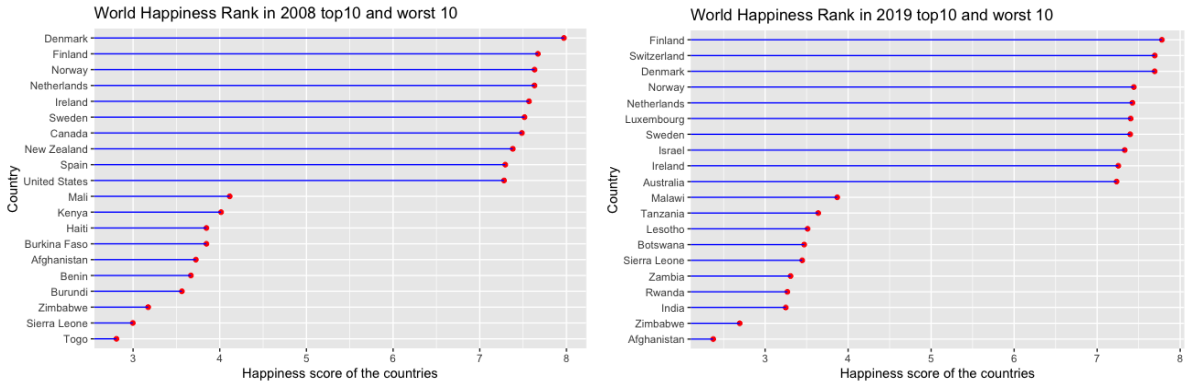
1. Which factors are the main effects of the countries' happiness?
2. What is the percentage change of the top 10 and least 10 countries' happiness in 2008 and 2019, and why?
3. Which country has the highest happiness score in the world in 2008, 2019 respectively?
4. Determine if GDP is the most impact factor of happiness
5. What is the regional happiness contrast for 2008 and 2019?

Statistical Analysis

Correlation:

After running the 'cor' function in R between Happiness Scores and all other numeric variables in our data, we found that the variables that are most correlated with Happiness Scores are logGDP, Social support and Life expectancy. The correlation between Happiness Scores and logGDP is 0.75943; the correlation between Happiness scores and Social support is 0.75036; and the correlation between Happiness Scores and Life expectancy is 0.77135. But before choosing the best model, we will look into which 10 countries have top and least happiness in 2008 and 2019.



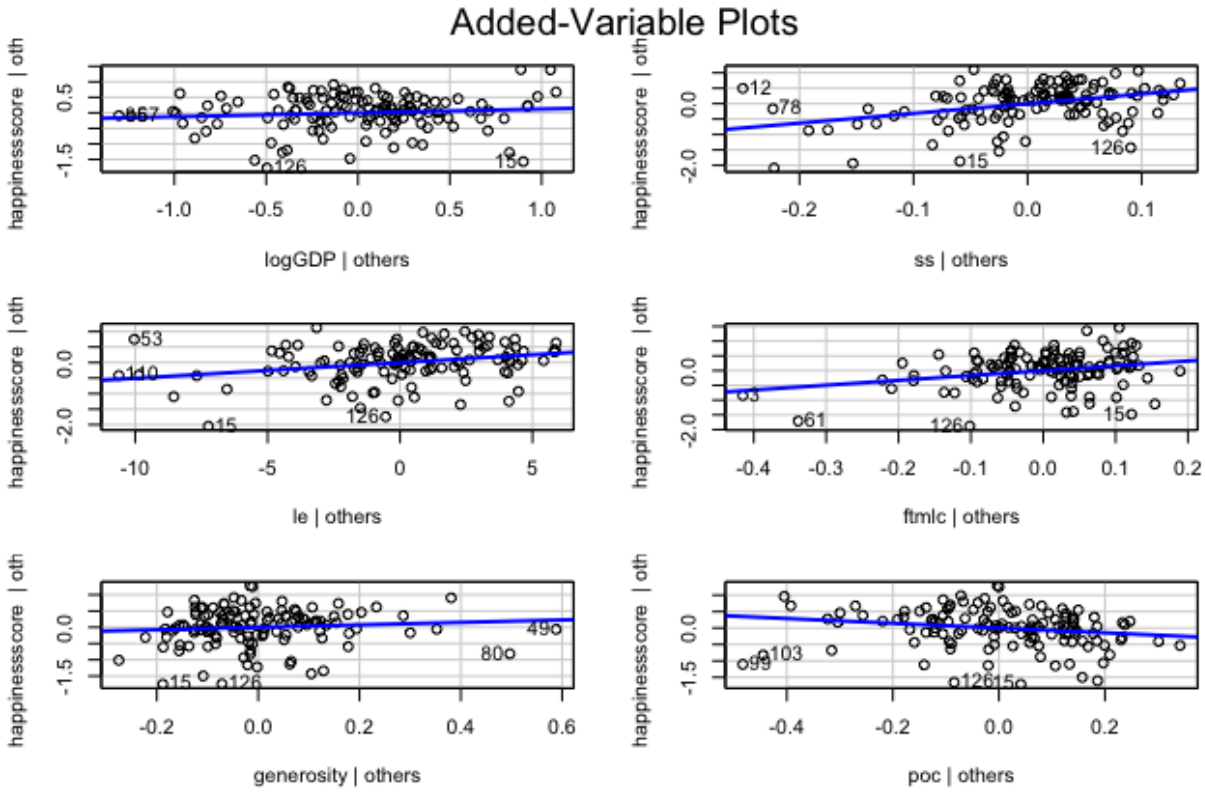


The graph on top shows that in 2008, Denmark, Finland, Norway, Netherlands, Ireland, Sweden, Canada, New Zealand, Spain and United States are the top 10 countries with happiness scores over 7; however, Mali, Kenya, Haiti, Burkina Faso, Afghanistan, Benin, Burundi, Zimbabwe, Sierra Leone and Togo are the least 10 countries with happiness scores below 4. While the graph on the bottom shows that in 2019, Finland, Switzerland, Denmark, Norway, Netherlands, Luxembourg, Sweden, Israel, Ireland and Australia are the top 10 countries with happiness scores over 7, but Malawi, Tanzania, Lesotho, Botswana, Sierra Leone, Zambia, Rwanda, India, Zimbabwe and Afghanistan are the least 10 countries with happiness scores less than 4.

Multiple Linear Regression:

After running the multiple linear regression of Happiness Scores on variables, we notice that GDP and generosity may have little to no effect for the happiness score. The p-value for

logGDP is 0.25, and p-value for generosity is 0.32. Both of them are greater than 0.05. Thus, we conclude that they are not significant. The F-statistic is 56.83, further validating that our model is useful (there is a significant relationship between Happiness scores, social support, life expectancy, freedom to make life choices and Perceptions of corruption). The multiple R squared value for the model is 0.7413 and the adjusted R squared value is 0.7283, indicating that our regression explains more than 70 percent variability in the Happiness scores.



Data Transformation and Backward Stepwise:

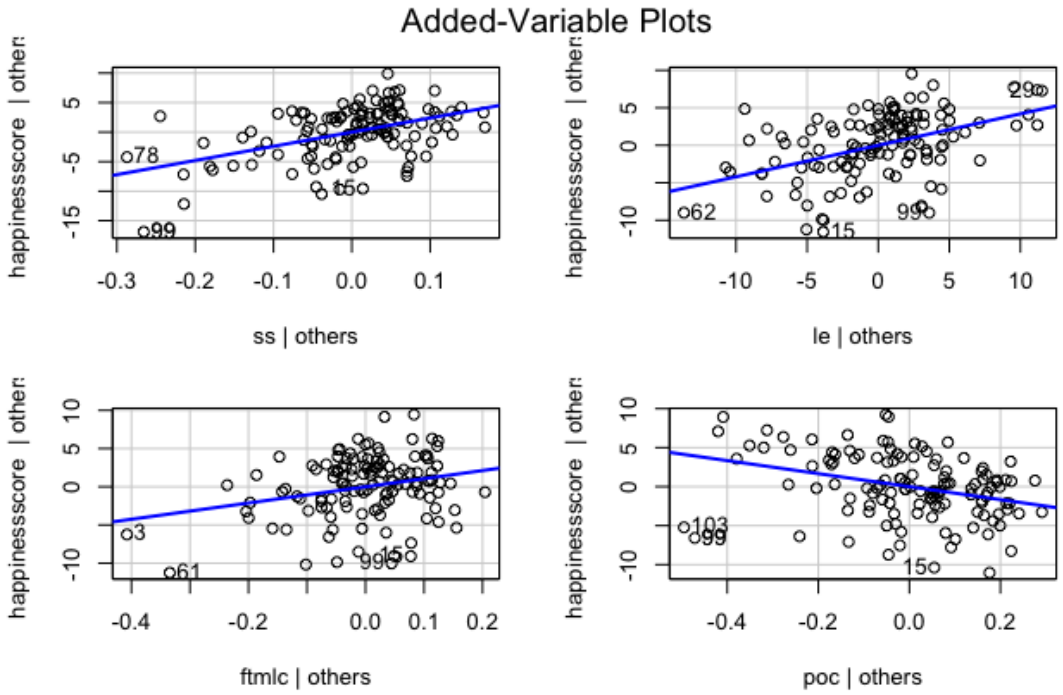
We then plotted the diagnostics for this model. The residual plot indicates that the data are simulated in a way that meets the linearity assumption not so well. So, we will apply box cox transformation for the happiness score. After transfer, the variance looks better, and the qq plot shows the residuals are normally distributed.

In order to choose the most useful model, we use backward stepwise methods. The result shows that we get the highest adjusted R2 dropping the logGDP and generosity.

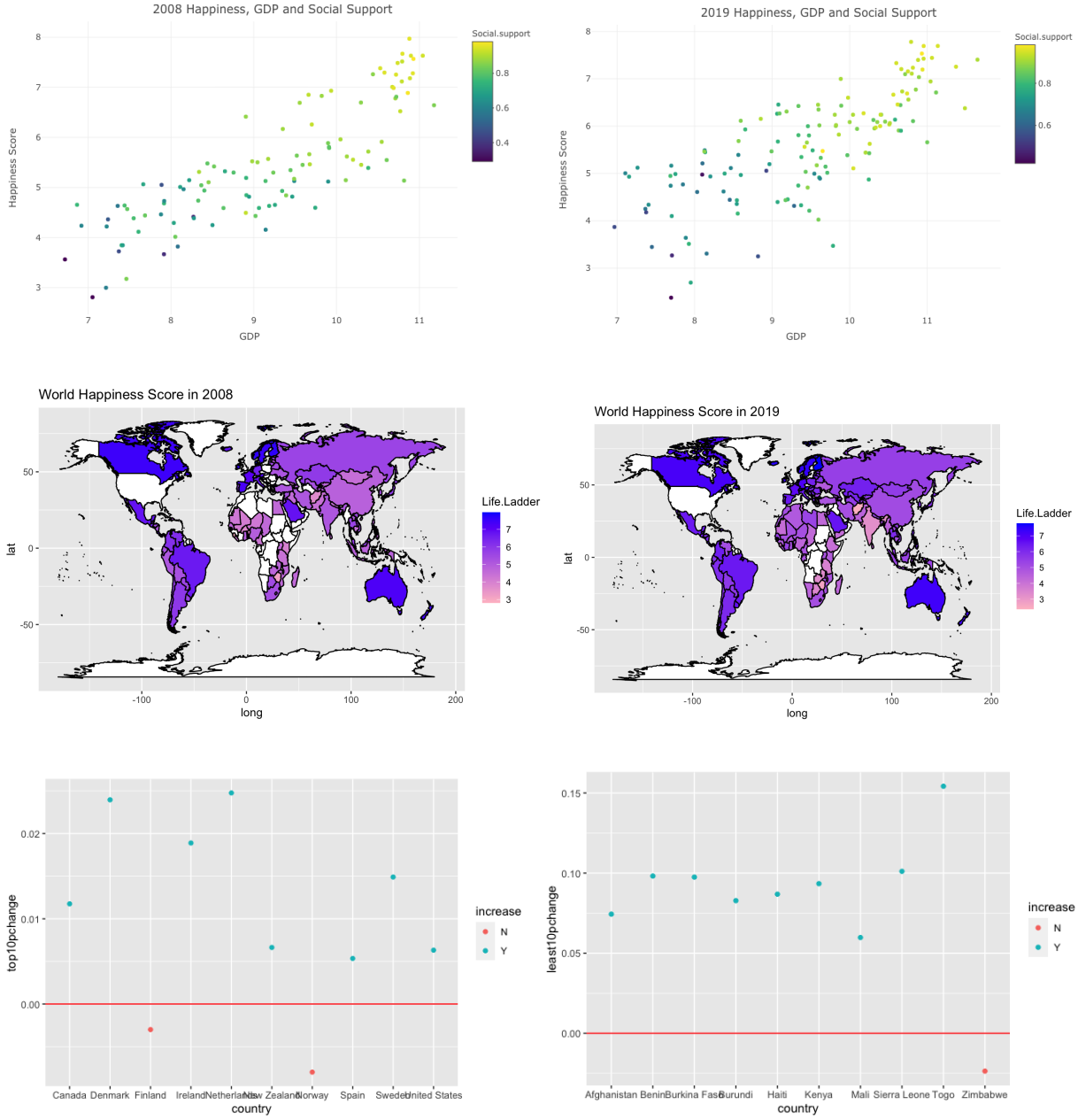
```
> cbind(summary(modelback)$which,"adjusted r^2" = summary(modelback)$adjr2)
  (Intercept) logGDP ss le ftmlc generosity poc adjusted r^2
1           1      1  0  0     0          0   0    0.7283256
2           1      1  0  0     0          0   1    0.7675804
3           1      1  1  0     0          0   1    0.8068960
4           1      1  1  0     1          0   1    0.8241881
5           1      1  1  0     1          1   1    0.8315697
6           1      1  1  1     1          1   1    0.8331945
```

The final model is:

$Y=\beta 0+\beta_{ss}+\beta_{le}+\beta_{ftmlc}+\beta_{poc}$



Results:



Social support, life expectancy, freedom to choose the life to live, and perceptions of corruption are the main factors affecting the countries' happiness.

Denmark has the highest happiness score in the world in 2008, while Finland has the highest happiness score in 2019.

From the last two plots, we can see there is about a 10% change in most of the least 10 countries' happiness. However, there's only 1 or 2 percent change in most of the top 10 countries' happiness, which indicates that people in the countries which they do not feel happy are changing. GDP is not one of the reasons that makes people happy. In other words, you can not buy happiness using monetary items.

Based on the map plot, we know that western countries' people are more satisfied with their life compared with the eastern countries' people.

Conclusion

In conclusion, we used multiple linear regression to analyze which factor has more influence on happiness score. We were able to show that there is a significant relationship between Happiness scores, social support, life expectancy, freedom to make life choices and Perceptions of corruption. After box cox transformation and backyard stepwise, we get a better model. The percentage change of happiness score of the top 10 countries told us that people are getting more happiness worldwide.

Contributions

Cheng liu: Data cleaning, debug

Xiaojin Yan, Cheng liu (partial): Data visualization

Chunqiu Li: Introduction and Background

Xinmei Wang, Cheng Liu (partial): Statistical analysis

Reference

https://www.kaggle.com/ajaypalsinghlo/world-happiness-report-2021?select=world-happiness-report.csv

Appendix

```
setwd("~/Documents")
whrorigional <- read.csv("world-happiness-report.csv")
names(whrorigional) <-
c("country","year","happinessscore","logGDP","ss","le","ftmlc","generosity","poc","pa","na")
whr <- whrorigional[complete.cases(whrorigional), ]
library(dplyr)
library(leaps)
library(MASS)
library(ggiraph)
library(ggiraphExtra)
library(plyr)
library(car)
library(ggplot2)
library(ggcorrplot)
library(corrplot)
library(plotly)
#2019 world happiness report
df <- whr %>% dplyr::select("country":"poc")%>% filter(year == 2019)

plot(df[3:9])
cor(df[3:9])
#top10 countries 2019
top102019 <- df %>% arrange(desc(happinessscore)) %>% filter(happinessscore >
happinessscore[11])
#least10 countries 2019
least102019 <- df %>% arrange(happinessscore) %>% filter(happinessscore <
happinessscore[11])
#mlr
model <- lm(happinessscore ~ logGDP + ss + le + ftmlc + generosity + poc, data = df)
summary(model)

plot(model,which = 1)
plot(model,which = 2)
avPlots(model)


modelback <- regsubsets(happinessscore ~ logGDP + ss + le + ftmlc + generosity + poc, data =
df, method = "backward")
cbind(summary(modelback)$which,"adjusted r^2" = summary(modelback)$adjr2)
```

```
modelnew <- lm(happinessscore ~ ss + le + ftmlc + poc, data = df)
par(mfrow = c(1,2))
plot(modelnew,which = 1)
plot(modelnew,which = 2)
par(mfrow = c(1,1))
boxcox(happinessscore ~ ss + le + ftmlc + poc, data = df,seq(0,5,1))
df$happinessscore <- df$happinessscore^(1.8)

modelnew1 <- lm(happinessscore ~ ss + le + ftmlc + poc, data = df)
summary(modelnew1)
par(mfrow = c(1,2))
plot(modelnew1,which = 1)
plot(modelnew1,which = 2)
par(mfrow = c(1,1))
avPlots(modelnew1)

#top10 countries 2008
df <- whr %>% filter(year == 2008)
top102008 <- df %>% arrange(desc(happinessscore)) %>% filter(happinessscore >
happinessscore[11])
#least 10 countries 2008
least102008 <- df %>% arrange(happinessscore) %>% filter(happinessscore <
happinessscore[11])

#percentage change
top10pchange
<-(top102008$happinessscore-top102019$happinessscore)/top102008$happinessscore
increase <- as.factor(ifelse(top10pchange > 0, "Y","N"))
top102008 <- cbind(top102008,top10pchange, increase)
top102008 <- group_by(top102008, increase)

least10pchange
<-(least102008$happinessscore-least102019$happinessscore)/least102008$happinessscore
increase <- as.factor(ifelse(least10pchange > 0, "Y","N"))
least102008 <- cbind(least102008,least10pchange, increase)
least102008 <- group_by(least102008, increase)

#percentage plot for top10 and low10
```

```
ggplot(data = top102008,aes(x = country, y = top10pchange,colour=increase)) + geom_point() +
geom_hline(yintercept = 0,col = "red")

ggplot(data = least102008,aes(x = country, y = least10pchange,colour=increase)) + geom_point()
+ geom_hline(yintercept = 0,col = "red")

#correlation plot

corr <- round(cor(df[3:9]), 1)
ggcorrplot(corr,method = "circle")

##World Happiness Rank in 2019 with score greater than 6
ggplot(whr %>% filter(year==2008) %>% filter(happinessscore>6.5), aes(x= happinessscore
                                                  ,y= reorder(country,happinessscore))) +
  geom_point(colour = "red", alpha = 1) +
  geom_segment(aes(yend=reorder(country, happinessscore
  )), xend = 0, colour="blue", alpha = 1) +
  theme(axis.text.y = element_text(angle = 0, hjust = 1)) +
  labs(title = "World Happiness Rank in 2019 with score greater than 6", y = "Country", x =
"Happiness score of the countries")

##World Happiness Rank in 20019 with score less than 6
ggplot(whr %>% filter(year==2019) %>% filter(happinessscore<5), aes(x= happinessscore
                                                  ,y= reorder(country,happinessscore))) +
  geom_point(colour = "red", alpha = 1) +
  geom_segment(aes(yend=reorder(country, happinessscore
  )), xend = 0, colour="blue", alpha = 1) +
  theme(axis.text.y = element_text(angle = 0, hjust = 1)) +
  labs(title = "World Happiness Rank in 20019 with score less than 6", y = "Country", x =
"Happiness score of the countries")


#World Happiness Rank in 2008 top10 and worst 10
library("dplyr")
whr_2008_top10 <- whr %>%  filter(year==2008) %>%
  arrange(desc(happinessscore)) %>%
  slice(1:10)
whr_2008_low10 <- whr[order(as.numeric(whr$happinessscore)), ] %>%  filter(year==2008)
whr_2008_low10<-whr_2008_low10[1:10,]
whr_2008_T_L_10<-bind_rows(whr_2008_top10,whr_2008_low10)
```

```
ggplot(whr_2008_T_L_10 %>% filter(year==2008) , aes(x= happinessscore
                                        ,y= reorder(country,happinessscore))) +
  geom_point(colour = "red", alpha = 1) +
  geom_segment(aes(yend=reorder(country, happinessscore
  )), xend = 0, colour="blue", alpha = 1) +
  theme(axis.text.y = element_text(angle = 0, hjust = 1)) +
  labs(title = "World Happiness Rank in 2008 top10 and worst 10", y = "Country", x =
"Happiness score of the countries")

#World Happiness Rank in 2019 top10 and worst 10
library("dplyr")
whr_2019_top10 <- whr %>%  filter(year==2019) %>%
  arrange(desc(happinessscore)) %>%
  slice(1:10)
whr_2019_low10 <- whr[order(as.numeric(whr$happinessscore)), ] %>%  filter(year==2019)
whr_2019_low10<-whr_2019_low10[1:10,]
whr_2019_T_L_10<-bind_rows(whr_2019_top10,whr_2019_low10)
whr_2019_T_L_10

ggplot(whr_2019_T_L_10 %>% filter(year==2019) , aes(x= happinessscore
                                        ,y= reorder(country,happinessscore))) +
  geom_point(colour = "red", alpha = 1) +
  geom_segment(aes(yend=reorder(country, happinessscore
  )), xend = 0, colour="blue", alpha = 1) +
  theme(axis.text.y = element_text(angle = 0, hjust = 1)) +
  labs(title = "World Happiness Rank in 2019 top10 and worst 10", y = "Country", x =
"Happiness score of the countries")

#2008 Happiness, GDP and Social Support
plot_ly(data = whr %>% filter(year==2008),
        x=~logGDP, y=~happinessscore, color=~ss, type = "scatter",
        text = ~paste("Country:", country)) %>%
  layout(title = "2008 Happiness, GDP and Social Support",
         xaxis = list(title = "GDP"),
         yaxis = list(title = "Happiness Score"))
#2008 Happiness, GDP and Social Support
plot_ly(data = whr %>% filter(year==2019),
        x=~logGDP, y=~happinessscore, color=~ss, type = "scatter",
```

```
        text = ~paste("Country:", country)) %>%
  layout(title = "2019 Happiness, GDP and Social Support",
         xaxis = list(title = "GDP"),
         yaxis = list(title = "Happiness Score"))

#World Happiness Score in 2008
library(maps)
world <- map_data('world')
whr_2008 <- whrorigional %>% select(country, happinessscore, year) %>% filter(year == 2008)

ggplot() +
  geom_map(data=world, map=world,
           aes(x=long, y=lat, group=group, map_id=region),
           fill="white", colour="black") +
  geom_map(data=whr_2008, map=world,
           aes(fill=happinessscore, map_id=country),
           colour="black") +
  scale_fill_continuous(low="pink", high="blue",
                        guide="colorbar") +
  labs(title = "World Happiness Score in 2008")

#World Happiness Score in 2019
whr_2019 <- whrorigional %>% select(country, happinessscore, year) %>% filter(year == 2019)

ggplot() +
  geom_map(data=world, map=world,
           aes(x=long, y=lat, group=group, map_id=region),
           fill="white", colour="black") +
  geom_map(data=whr_2019, map=world,
           aes(fill=happinessscore, map_id=country),
           colour="black") +
  scale_fill_continuous(low="pink", high="blue",
                        guide="colorbar") +
  labs(title = "World Happiness Score in 2019")
```