Happiness in the world

Pluto Zhang

2022-12-14

Introduction

Happiness level of a country is very important to the development of the country, and influences the all social sectors such as health, economy, social stability, so on and so forth. To figure out how the general social wellness of a country influences its overall happiness level, and how happy a country is with respect to its neighboring countries, is important.

My central questions are: 1. How does the happiness level of a country relate to its social wellness factors? Would the social wellness of a country predict its happiness level? 2. Do countries that locate geographically close to each other also have similarity in terms of happiness levels?

This project aims to 1. Generate a linear regression model to predict the level of happiness across countries in the world, using social wellness factors; 2. Assess whether the regions with similar happiness levels and social characters are geographically close, across 2020 to 2021. Use hierarchical clustering to figure out if the clusters of regions belong to the same continent or are different in 2020 and 2021.

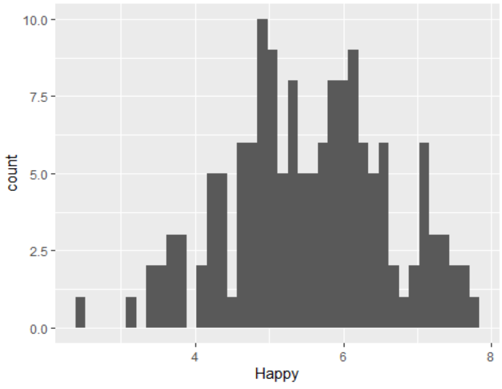
1. Datasets:

This project mainly uses two datasets: the 2020 and 2021 World Happiness Report datasets, both datasets are pulled from the World Happiness Report website, which provides numerical data. The variables included in the datasets: the adjusted happiness scores, log of the GDP per capita, the social support, healthy life expectancy at birth, freedom to make life choices, generosity and perception of corruption.

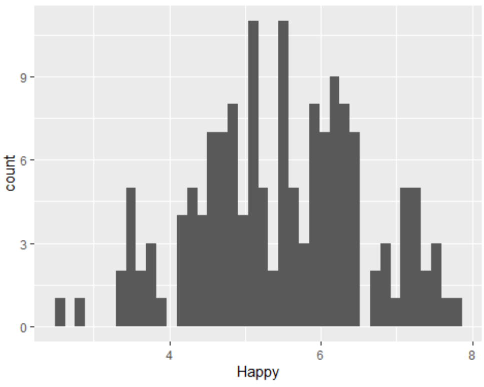
The 2020 dataset has data for 153 countries, while the 2021 dataset has data for 149 countries. since the number of countries are different in happy 20 and happy 21, for now, because the number of country difference is less than 5%, I removed the countries that differ from each other from the dataset, so each dataset now has 148 countries.

There're no NA values in the 2021 world happiness dataset, which is desirable and indicates that there is no missing data.

2. EDA:



(0.1)



(0.2)

From the summary histogram of 2021(graph 0.1), the happiness scores are approximately normally distributed, with most of the countries' happiness scores at around 5 and 6, a few below 3 and above 7. This finding indicates that most of the countries have moderate happiness scores among the 148 countries.

From the summary histogram 0.2, the 2020 happiness scores are approximately normally distributed, with most of the countries' happiness scores at around 4 to 6, a few below 3 and above 7. This finding indicates that most of the countries have moderate happiness scores among the 153 countries. Compared to 2021 happiness score data, the countries in 2020 have in general lower happiness score, with more countries hitting the values lower than 4 and less countries higher than 7.

The regions with the highest happiness score are North America and Western Europe, in both 2020 and 2021.

3. Analysis Methods

Two major analysis methods will be used in this project to address the two central questions:

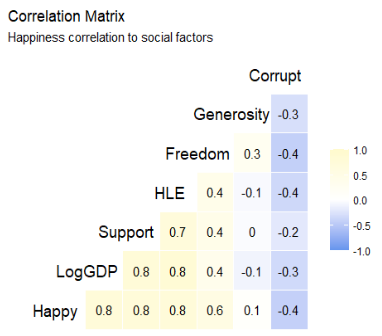
1. Linear Regression Model Fitting For the first question, we want to create a model that predicts the happiness score based on other social factors. For this part, I select the linear regression model to achieve the prediction, because it's a long proven and simple statistical method to assess the relationship between variables. First, I examined the correlation between the social factors with happiness, to see which social factors are the most relevant in foreseeing the happiness level of countries. Then I fit the model linearly, that the 2020 data will be used as the training set, while the 2021 data will be used as the test set of the model to evaluate the fit. I compared two different models by including in different covariates, and comapred their values of goodness of fit.

2. Hierarchical and K-means Clustering For the second question, I first want to see whether there exists any clusters of regions based on the happiness scores and social factors. I then plot these clusters on the world map, to see whether these clusters of regions geographically locate close to each other, or the regions that belong to the same cluster locate far apart. The hierarchical clustering method is used, because it's useful in dividing the regions into groups, so that it's easy to visualize on the world map how closely located these clusters of regions are.

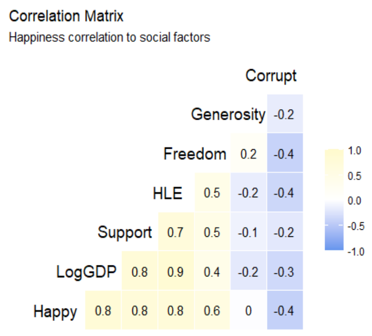
Results

Part1 Linear Regression Model

First, let's examine the strength of correlation of the factors with happiness is examined using pearson correlation matrices.



(1.1)

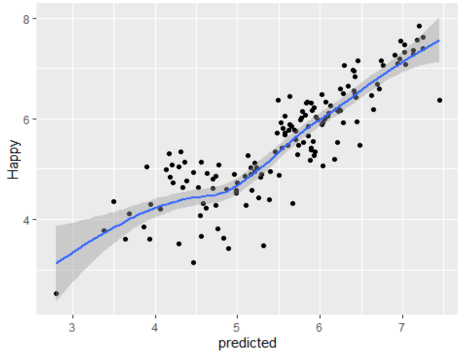


(1.2)

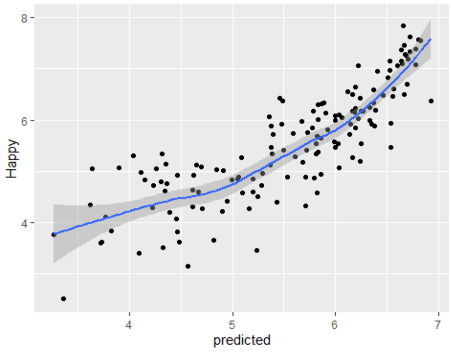
From the correlation matrices(1.1 for 2021 data, 1.2 for 2020 data)we can see that GDP, social support, healthy life expectation are the factors that most strongly correlates with the happiness levels of the countries, and all three indicates a strong positive correlation with happiness. Freedom indicates a moderate positive correlation with happiness score,
while the level of corruption indicates a moderate negative correlation with happiness score. The level of generosity did not present a strong correlation with happiness score.

Thus, for the final model, I chose Log of GDP, Support, Health Life Expectancy, Freedom and Perception of Corrupt as the predictors for happiness score.

The RMSE value at 0.535 and the R square of 0.752 indicates that the model is in general good at predicting the happiness score for 2021 happiness score.



(1.3)



(1.4)

From the graph 1.3 we can see that the prediction is generally accurate, with high correlation between the predicted values and the real happiness scores.

We can compare this model with another model that includes only the three most important variables, GDP, social support, healthy life expectation as the covariates.

This new model results in RMSE of 0.595, which is bigger than the RMSE of the full model at 0.535, indicating that this model is not as good as the full model in terms of fitting. It also has adjusted R squared value of 0.693, which is smaller than the adjusted R squared value of the full model at 0.752, which further supports the point that this model is worse in fitting as compared to the full model.

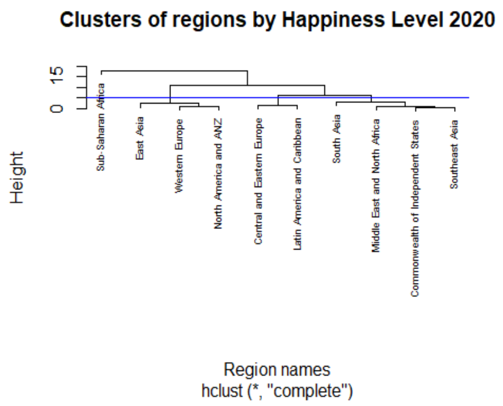
Looking at the predicted versus real happiness score value graph (graph 1.4) we can see that there are more scatters compared to the model with more covariates. Thus, the model that includes freedom and perception of corruption as covariates is better in predicting the happiness scores, and all the 5 covariates, LogGDP, Support, HLE, Freedom and Corrupt are positively correlated with happiness score, leading to a strong positive correlation.

Part 2 Clustering Analysis

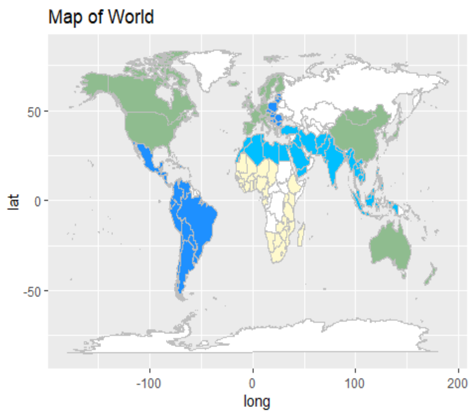
First, I generated data frames of the average of different social wellness factors by regions.

Part 2-1 Hierarchical clustering

Next, generate clusters of the regions based on the distance of each of the social factors:



(graph2.11)



(2.12)

From the 2020 dendrogram (graph2.11), there are 4 main clusters as divided by the blue line:

The first cluster: East Asia, Western Europe, North America and ANZ (Australia and New Zealand)

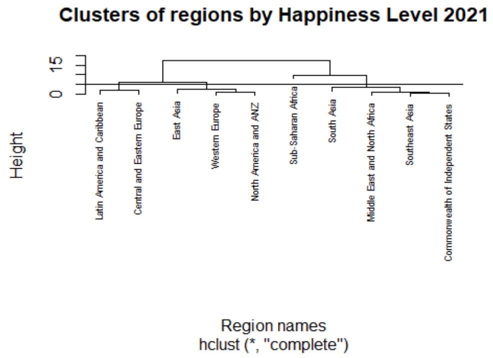
The second cluster: South Asia, Middle East and North Africa, Commonwealth and Independent States, Southeast Asia.

The third cluster: Central and Eastern Europe, Latin America and Caribbean.

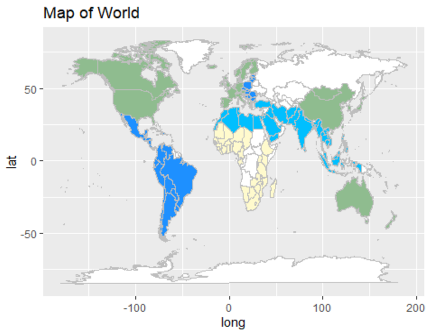
The second and the third cluster are closer to each other in terms of the social factors and happiness scores.

The forth cluster, which is also the cluster farthest away from the other three clusters, is Sub-saharan Africa.

When plotting 2020 clusters on the world map (2.12), it shows consistent colorings for continents such as the African and American continent, and the colorings usually spans may neighboring countries. It's interesting how eastern European and south American countries belong to the same cluster, despite the geographical distance.



(2.13)



(2.14)

From the 2021 dendrogram(2.13), there are 4 main clusters as divided by the line, and the 4 main clusters are the same as the 2020 dendrogram. For 2021 data, the first and third clusters are closer to each other, while the second and the forth clusters are closer to each other.

It's worth noticing that the regions within each cluster didn't change from 2020 to 2021.

From the plots above, we can see that: Under hierarchical clustering:

1. The clusters of regions with similarity in happiness score and social factor data did not change much from 2020 to 2021, as shown on the map.

2. The regions that belong to the same cluster, or have similar happiness level, do not necessarily locate close to each other or on the same continent. For countries on the American continent, it's clear that the North American countries are similar in happiness and socio-economic levels as a cluster, while south American countries form a cluster as well. Among Asian countries, however, the happiness levels vary a lot from subregions, such that the South Asian and Southeast Asian have very different happiness scores as compared to east Asian countries. However, each cluster tends to have some degree of spread and spanning across the same continent, including at least multiple neighboring countries at a time.

Part 2-2 K-means clustering

For k-means clustering with 4 clusters, we can see that: for 2020 data,

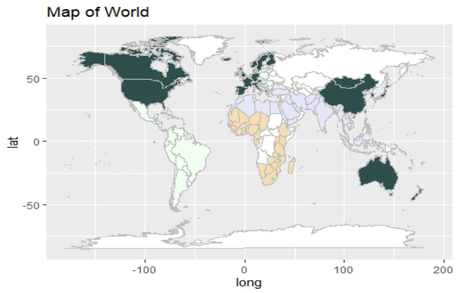
The first cluster is South Asia, Middle East and North Africa, Commonwealth and Independent States, Southeast Asia, the second cluster is Sub-Saharan Africa.

The third cluster: East Asia, Western Europe, North America and ANZ (Australia and New Zealand)

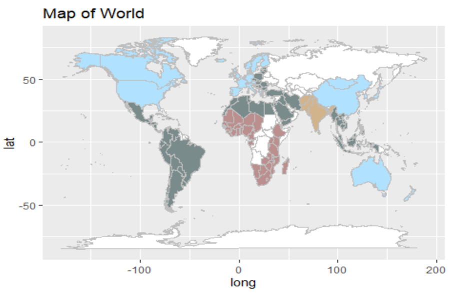
The fourth cluster: Central and Eastern Europe, Latin America and Caribbean.

The division of regions by clusters is the same as using hierarchical clustering.

Now visualize that on the world map:



(2.21)



(2.22)

As can see from the world map, for year 2020, the 4 clusters divided by the k-means approach are the same as the 4 clusters generated by the hierarchical approach.

For k-means clustering with 4 clusters, for 2021 data,

The first cluster is: Sub-Saharan Africa

The second cluster, South Asia

the third cluster: Middle East and North Africa, Commonwealth and Independent States, Southeast Asia, Central and Eastern Europe, Latin America and Caribbean.

The fourth cluster: East Asia, Western Europe, North America and ANZ (Australia and New Zealand)

This time, as can see from graph 2.22, the clusters have different contents as compared to the hierarchical approach on 2021 data, that South Asia itself consists of a single cluster, while Middle East and North Africa, Commonwealth and Independent States, Southeast Asia are put into the same cluster with Central and Eastern Europe, Latin America and Caribbean.

By looking at the social factors data of 2021(Appendix 2.3), the first and second clusters, Sub-Saharan Africa and South Asia, have the lowest happiness score values at 4.44 and 4.49, with relatively low GDP, Health life expectancy and Social Support, but relatively high or average in other wellness indicators. This explains why South Asia is separated from the cluster with other regions for the k-means clustering approach, because k-means clustering is better in ruling out clusters with different sizes as compared to hierarchical clustering. so despite South Asia is similar in other regions in terms of generosity and corrup scores, it is still singled out as a cluster with low happiness score, using k-means approach.

Conclusion

Our two research questions are,

1. How does the happiness level of a country relate to its social wellness factors? Would the social wellness of a country predict its happiness level?

2. Do regions that locate geographically close to each other also have similarity in terms of happiness levels?

To answer the first question, I used linear regression model fitting, using the 2020 data as the training set and the 2021 data as the test set, and it turns out that the linear regression model that includes the 5 covariates, LogGDP, Support, HLE, Freedom and Corrupt is an optimal model that can predict the happiness with relatively high accuracy, with a small RMSE value. And all 5 covariates are positively and strongly correlated with happiness score.

For the second question, I used hierarchical clustering and k-means clustering to figure out whether the global regions that locate geographically close to each other are also close in happiness level, using the average scores of social wellness factors by region in 2020 and 2021 as inputs, generating the clusters and then visualize the countries that belong to the same cluster by giving them same color on the world map.

Under hierarchical clustering, I found out that the clusters of regions similar in happiness level did not change much from 2020 to 2021, and that the regions that belong to the same cluster, or have similar happiness level, do not necessarily locate close to each other or on the same continent. But each cluster tends to spread across the at least multiple neighboring countries at the same continent. The k-means clustering results for 2020 is the same as the hierarchical clustering results, but for 2021 data, South Asia is ruled out as a single cluster, because its happiness score is the lowest in 2021, despite its similarity with other regions in some other social factors like generosity and perception of corruption, and this difference is caught by the k-means clustering approach. But in general, the division of global regions in terms of happiness level does not change much from 2020 to 2021, and that neighboring countries in African, North and South American tend to have similarities in social wellness, while in Asia, the differences in social wellness factors could be larger between different parts of Asian continent.

My approaches generally answered the two questions very well, because 1. the resulting regression model generates a good fit for the data and the prediction achieves relatively high accuracy. 2. For the clustering analysis, the outcomes are relatively consistent and coherent across differnet years and clustering approaches, and by plotting the clustering results on the world map, my approach is easy to visualize and interpret.

If I have more time, I will include in more datasets, such as the 2018 and 2019 datasets as well, to see whether the occurrences of COVID-19 have effects on the world happiness levels across countries. I would also consider adding in COVID infection data by country, to see whether that could also be a covariate that helps predict the happiness score by country.

References

The sources of 2020 and 2021 World Happiness Score data:

1. Helliwell, John F., et al. “World Happiness Report 2020.” The World Happiness Report, 20 Mar. 2020, worldhappiness.report/ed/2020/.

2. Helliwell, John F., et al. “World Happiness Report 2021.” The World Happiness Report, 20 Mar. 2021, worldhappiness.report/ed/2021/.

Appendices:

Table 0.1

Summary table of the 2021 data:

Region	happyavg
South Asia	4.441857
Sub-Saharan Africa	4.494472
Middle East and North Africa	5.219765
Southeast Asia	5.407556
Commonwealth of Independent States	5.467000
East Asia	5.810333
Latin America and Caribbean	5.908050
Central and Eastern Europe	5.984765
Western Europe	6.914905
North America and ANZ	7.128500

Graph 0.1: The histogram of the countries' happiness scores in 2021:

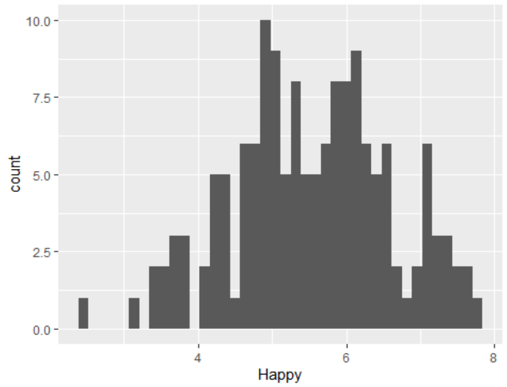


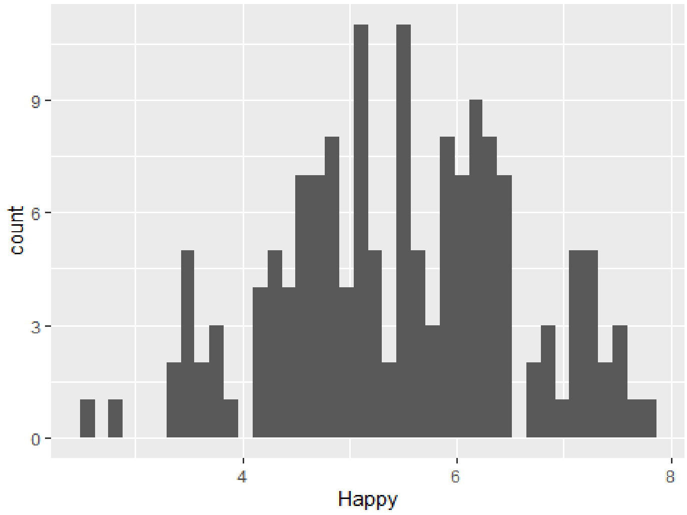
Table 0.2:

Summary table of the 2020 data:

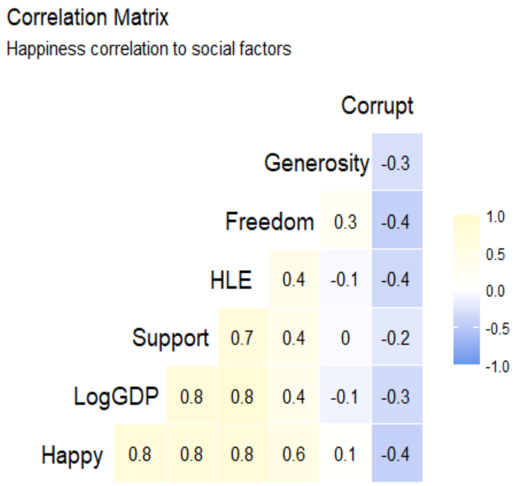
Region	happyavg
Sub-Saharan Africa	4.383495
South Asia	4.475443
Middle East and North Africa	5.227159
Commonwealth of Independent States	5.358342
Southeast Asia	5.383367
East Asia	5.714850

Region	happyavg
Central and Eastern Europe	5.883818
Latin America and Caribbean	5.981786
Western Europe	6.899219
North America and ANZ	7.173525

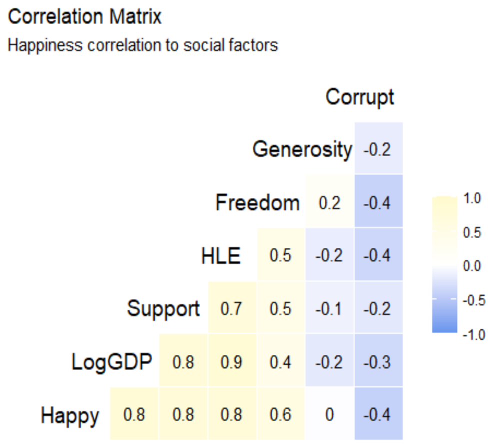
Graph 0.2: The histogram of the countries’ happiness scores in 2020



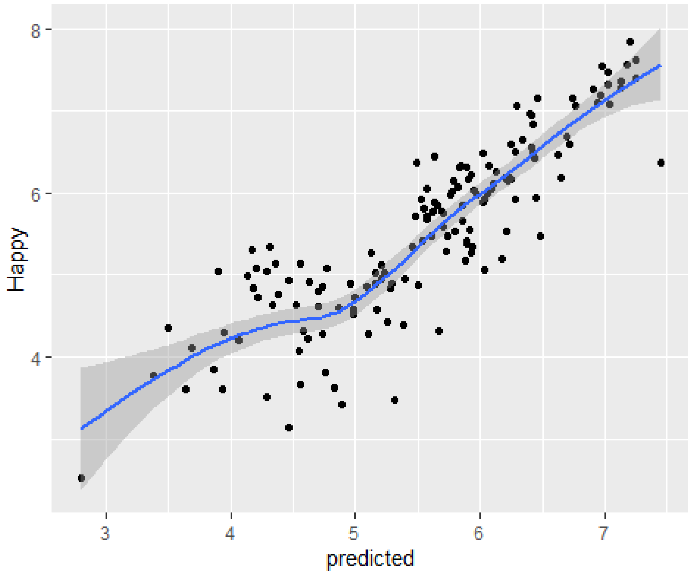
Graph 1.1: The correlation matrix of the social wellness factors and happiness score in 2020:



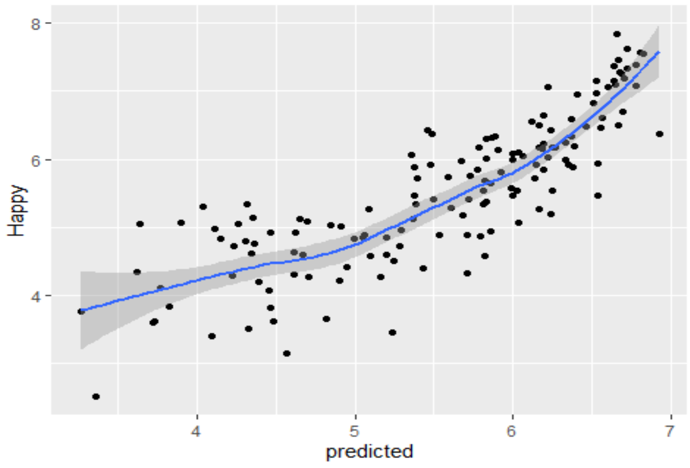
Graph 1.2: The correlation matrix of the social wellness factors and happiness score in 2021:



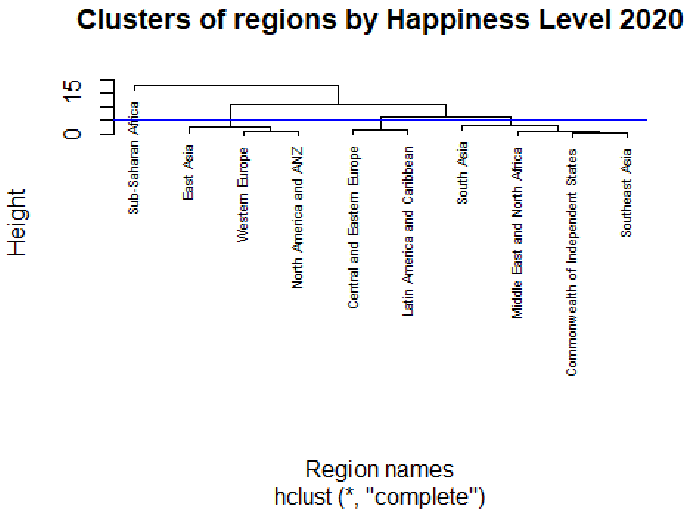
Graph 1.3: the real vs predicted happiness scores for 2021 data, using Log of GDP, Support, Health Life Expectancy, Freedom and Perception of Corrupt as covariates:



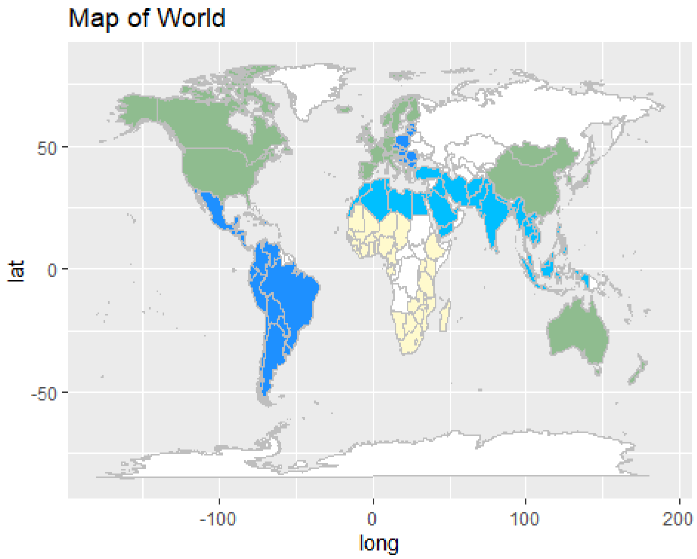
Graph 1.4: the real vs predicted happiness scores for 2021 data, using only log GDP, HLE and social support as covariates:



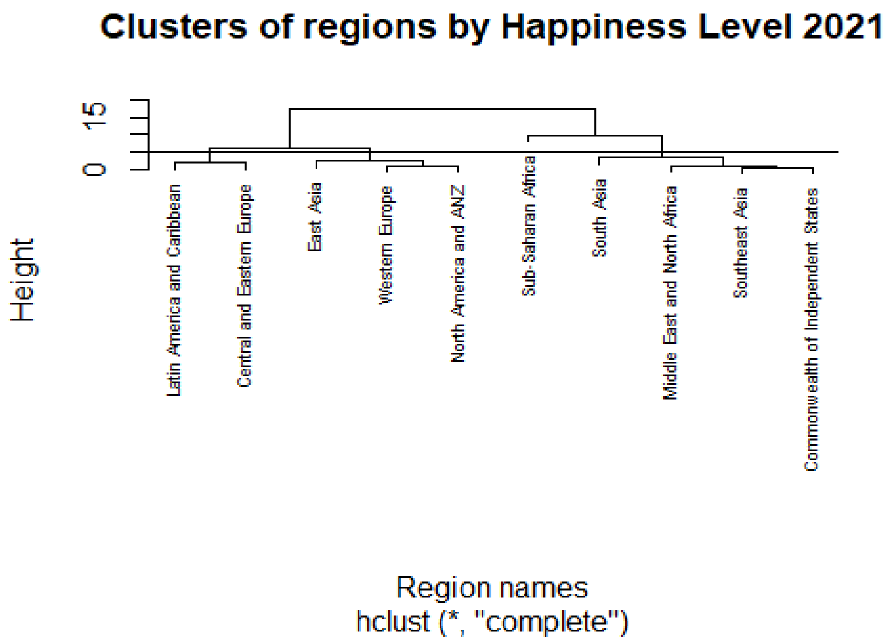
Graph 2.11: the dendrograms of the clusters of regions generated for 2020 data:



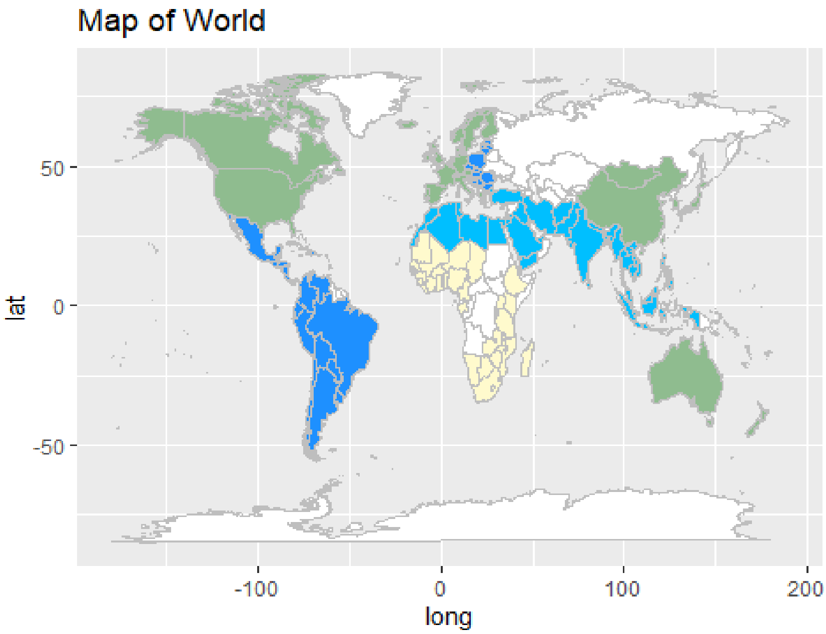
Graph 2.12: the hierarchical clustering results on 2020 data:



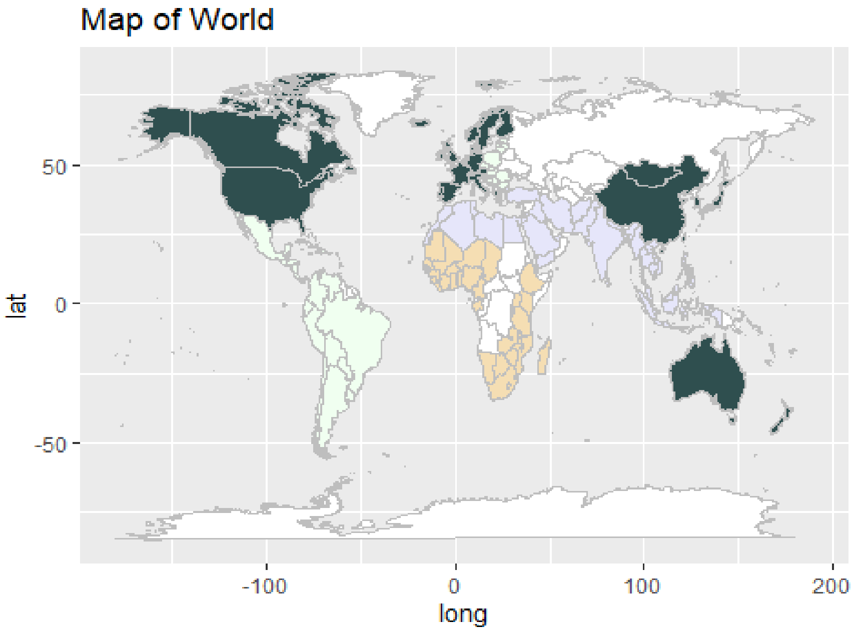
Graph 2.13: the hierarchical clustering dendrogram on 2021 data:



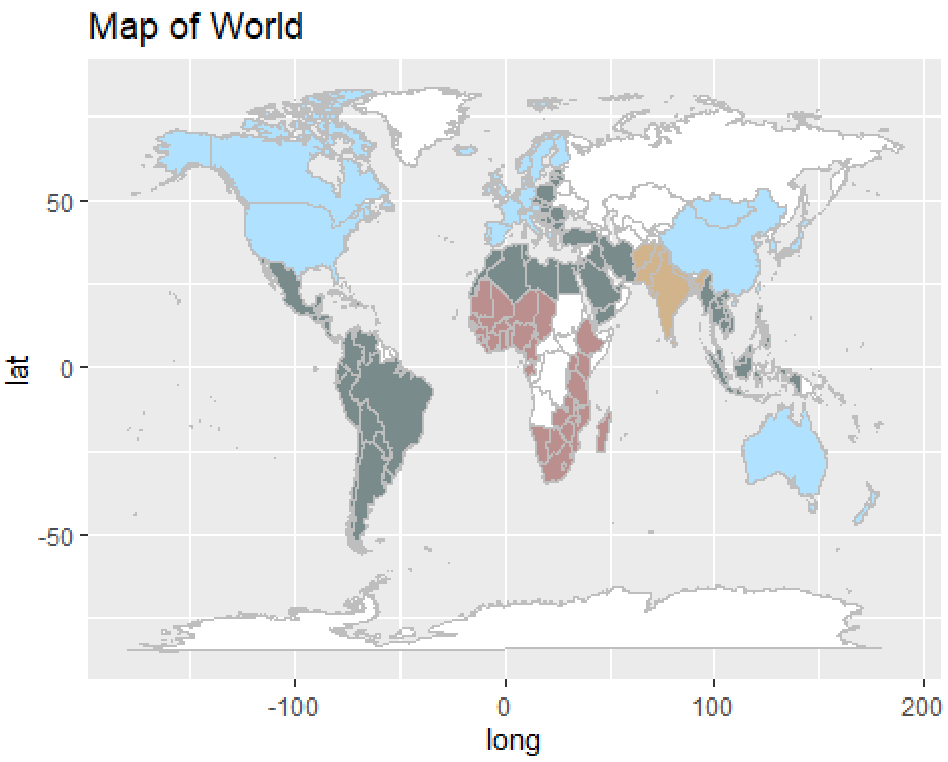
Graph 2.14: The hierarchical clustering results on 2021 data:



Graph 2.21: The k-means clustering results on 2020 data:



Graph 2.22: the k-means clustering results on 2021 data:



Appendix 2.3: The table for socio economic factors for different regions, on average in 2021:

Region	Happyavg	GDPavg	Supportavg	HLEavg	Freedomavg	Generosityavg	Corruptavg
South Asia	4.441857	8.682571	0.7034286	62.68100	0.7650000	0.0427143	0.7974286
Sub-Saharan Africa	4.494472	8.075194	0.6967500	55.88647	0.7231944	0.0134444	0.7659444
Middle East and North Africa	5.219765	9.666118	0.7976471	65.60912	0.7164706	-0.0797647	0.7622353
Southeast Asia	5.407556	9.421444	0.8203333	64.88844	0.9090000	0.1563333	0.7091111
Commonwealth of Independent States	5.467000	9.401833	0.8725000	65.00950	0.8169167	-0.0360000	0.7250833
East Asia	5.810333	10.367667	0.8605000	71.25217	0.7635000	-0.0623333	0.6833333
Latin America and Caribbean	5.908050	9.370000	0.8395000	67.07605	0.8317500	-0.0677000	0.7926000
Central and Eastern Europe	6.040000	10.135063	0.8925625	68.51744	0.7999375	-0.0862500	0.8471250

Western Europe	6.914905	10.822714	0.9144762	73.03310	0.8587143	-0.0031905	0.5230952
North America and ANZ	7.128500	10.809500	0.9335000	72.32500	0.8987500	0.1200000	0.4492500

Code:

```
knitr::opts_chunk$set(warning=FALSE, message = FALSE, echo=FALSE)
library(tidyverse)
library(dplyr)
library(dslabs)
library(caret)
library(ggplot2)
library(ggrepel)
library(plm)
library(knitr)

#explore the world happiness score dataset
happy21 <- read_csv('world-happiness-report-2021.csv')

#colnames(happy21)
#From the summary statistics: we can see that there is no NA values in the
2021 world happiness dataset, which is desirable and indicates that there is
no missing data.

summary(happy21)

#select the wanted columns and rename them in 2021 dataset
happy21 <- happy21 |> select('Country name','Regional indicator','Ladder
score','Logged GDP per capita', 'Social support',  'Healthy life expectancy',
'Freedom to make life choices',   'Generosity','Perceptions of corruption' )
colnames(happy21) <- c('Country', 'Region','Happy',
'LogGDP','Support','HLE','Freedom','Generosity', 'Corrupt')
#HLE stands for health life expectancy
#make sure that the countries in this dataset are unique from each other
length(unique(happy21$Country)) == nrow(happy21)

#EDA
# first make a histogram of average score of happiness by country and by
region

eda21 <- happy21 |> group_by(Region) |> summarize(happyavg = mean(Happy)) |>
arrange(happyavg)

kable(eda21)
happy21 |> ggplot() + geom_histogram(aes(x=Happy), bins=40)
#Now preprocess the 2020 happiness data:
#2020 happiness data
happy20 <- read_csv('WHR20_DataForFigure2.1.csv')

happy20 <- happy20 |> select('Country name','Regional indicator','Ladder
score', 'Logged GDP per capita', 'Social support',  'Healthy life
expectancy', 'Freedom to make life choices',   'Generosity','Perceptions of
corruption' )
colnames(happy20) <- c('Country', 'Region','Happy',
'LogGDP','Support','HLE','Freedom','Generosity', 'Corrupt')
#EDA
# first make a histogram of average have a general view of happiness by
country and by region

region20<- happy20 |> group_by(Region) |> summarize(happyavg = mean(Happy))
|> arrange(happyavg)

kable(region20)
happy20 |> ggplot() + geom_histogram(aes(x=Happy), bins=40)

#nrow(happy20)
#nrow(happy21)
#since the number of countries are different in happy 20 and happy 21, for
now, because the number of country difference is less than 5%, let's remove
the countries that differ from each other.

happy20 <- happy20 |> filter((Country %in% happy21$Country))
happy21 <- happy21 |> filter((Country %in% happy20$Country))
library(GGally)
happy20numeric <- happy20 |> select(-c('Country', "Region"))
happy21numeric <- happy21 |> select(-c('Country', "Region"))

ggcorr(happy20,
       method = c("everything", "pearson"),
       size = 5, hjust = 0.77,
       low = 'cornflowerblue', mid = 'white', high = "lemonchiffon",
       label = TRUE, label_size = 4,
       layout.exp = 1) +
labs(title = 'Correlation Matrix',
    subtitle = 'Happiness correlation to social factors')
ggcorr(happy21,
       method = c("everything", "pearson"),
       size = 5, hjust = 0.77,
       low = 'cornflowerblue', mid = 'white', high = "lemonchiffon",
       label = TRUE, label_size = 4,
       layout.exp = 1) +
labs(title = 'Correlation Matrix',
    subtitle = 'Happiness correlation to social factors')

library(caret)
# use the 2020 happiness data as the training set and the 2021 data as the
test set.
#take a look at the summary of the model fit
fit1 <- lm( Happy ~ LogGDP+Support+HLE+Freedom+Corrupt, data = happy20)
#summary(fit1)
y_hat1 <- predict(fit1, newdata = happy21)

#see the RMSE
rmse1 <- sqrt(mean((y_hat1-happy21$Happy)^2))
paste0('This is the RMSE value of the first model: ',round(rmse1, 3))
#R-square
r1<- R2(y_hat1, happy21$Happy)
paste0('This is the R square value of the first model: ',round(r1, 3))

modeldata<- happy21
modeldata$predicted <- y_hat1
modeldata |> ggplot(aes(predicted, Happy)) + geom_point() + geom_smooth()

fit2 <- lm( Happy ~ LogGDP+Support+HLE, data = happy20)
#summary(fit2)
y_hat2 <- predict(fit2, newdata = happy21)

rmse2 <- sqrt(mean((y_hat2-happy21$Happy)^2))
paste0('This is the RMSE value of the reduced model: ',round(rmse2, 3))
r2 <- R2(y_hat2, happy21$Happy)
paste0('This is the R squared value of the reduced model: ',round(r2, 3))
modeldata2<- happy21
modeldata2$predicted <- y_hat2
modeldata2 |> ggplot(aes(predicted, Happy)) + geom_point() + geom_smooth()
region20all <- happy20 |> group_by(Region) |>
  summarize(Happyavg = mean(Happy),
            GDPavg = mean(LogGDP),
            Supportavg=mean(Support),
            HLEavg = mean(HLE),
            Freedomavg = mean(Freedom),
            Generosityavg = mean(Generosity),
            Corruptavg = mean(Corrupt)) |>
  arrange(Happyavg)

region21all <- happy21 |> group_by(Region) |>
  summarize(Happyavg = mean(Happy),
            GDPavg = mean(LogGDP),
            Supportavg=mean(Support),
            HLEavg = mean(HLE),
            Freedomavg = mean(Freedom),
            Generosityavg = mean(Generosity),
            Corruptavg = mean(Corrupt)) |>
  arrange(Happyavg)
region21alltibble <- region21all
row20 <- region20all$Region
region20all <- region20all[,-1] |> as.matrix()
rownames(region20all) <- row20

row21 <- region21all$Region
region21all <- region21all[,-1] |> as.matrix()
rownames(region21all) <- row21


d20 <- dist(region20all)
d21 <- dist(region21all)

h20 <- hclust(d20)
h21 <- hclust(d21)
plot(h20, cex = 0.55, main = "Clusters of regions by Happiness Level 2020",
xlab = "Region names")
abline(h=5, col="blue")

#Now get the vectors of countries that belong to each cluster:
cluster1_2020 <- happy20 |> filter(Region %in% c('East Asia', 'Western
Europe', 'North America and ANZ')) |> select('Country')
cluster2_2020 <- happy20 |> filter(Region %in% c('South Asia', 'Middle East
and North Africa', 'Commonwealth and Independent States', 'Southeast Asia'))
|> select('Country')

cluster3_2020 <- happy20 |> filter(Region %in% c('Central and Eastern
Europe', 'Latin America and Caribbean')) |> select('Country')
cluster4_2020 <- happy20 |> filter(Region %in% c('Sub-Saharan Africa')) |>
select('Country')

#Plot the results on the world map
```

```
thismap = map_data("world")

cluster1_2020<- cluster1_2020 |> mutate(
  Country1 = case_when(Country =='Taiwan Province of China' ~ 'Taiwan',
                       Country == 'United States'~ 'USA',
                       Country == 'United Kingdom'~ 'UK',
                       TRUE ~ Country))

cluster3_2020 <- cluster3_2020 |> mutate(Country1 = Country)
cluster2_2020 <- cluster2_2020 |> mutate(Country1 =
                                           case_when(
                                             Country =='Palestinian
Territories' ~ 'Palestine',
                                             TRUE ~ Country))
cluster4_2020 <- cluster4_2020 |> mutate(Country1 = Country)

#set colors by the clusters of countries
thismap <- mutate(thismap, fill =ifelse(region %in% cluster1_2020$Country1
,'darkseagreen',
                                        ifelse(region %in%
cluster2_2020$Country1, "deepskyblue",
                                        ifelse(region %in%
cluster3_2020$Country1,"dodgerblue",
                                        ifelse(region %in%
cluster4_2020$Country1,"lemonchiffon",
                                        ifelse(str_detect('Congo',
region),'lemonchiffon','white')))))

# set the colors
ggplot(thismap, aes(long, lat, fill = fill, group=group)) +
  geom_polygon(colour="gray") + ggtitle("Map of World") +
  scale_fill_identity()
# the 2021 dendrogram:
plot(h21, cex = 0.55, main = "Clusters of regions by Happiness Level 2021",
xlab = "Region names")
abline(h=5, color = 'blue')

cluster1_2021 <- happy21 |> filter(Region %in% c('East Asia', 'Western
Europe', 'North America and ANZ')) |> select('Country')
cluster2_2021 <- happy21 |> filter(Region %in% c('South Asia', 'Middle East
and North Africa', 'Commonwealth and Independent States', 'Southeast Asia'))
|> select('Country')

cluster3_2021 <- happy21 |> filter(Region %in% c('Central and Eastern
Europe', 'Latin America and Caribbean')) |> select('Country')
cluster4_2021 <- happy21 |> filter(Region %in% c('Sub-Saharan Africa')) |>
select('Country')
```

```
thismap = map_data("world")

cluster1_2021<- cluster1_2021 |> mutate(
  Country1 = case_when(Country =='Taiwan Province of China' ~ 'Taiwan',
                       Country == 'United States'~ 'USA',
                       Country == 'United Kingdom'~ 'UK',
                       TRUE ~ Country))

cluster3_2021 <- cluster3_2021 |> mutate(Country1 = Country)
cluster2_2021 <- cluster2_2021 |> mutate(Country1 =
                                           case_when(
                                             Country =='Palestinian
Territories' ~ 'Palestine',
                                             TRUE ~ Country))
cluster4_2021 <- cluster4_2021 |> mutate(Country1 = Country)

#set colors by the clusters of countries
thismap <- mutate(thismap, fill =ifelse(region %in% cluster1_2021$Country1
,'darkseagreen',
                                        ifelse(region %in%
cluster2_2021$Country1, "deepskyblue",
                                        ifelse(region %in%
cluster3_2021$Country1,"dodgerblue",
                                        ifelse(region %in%
cluster4_2021$Country1,"lemonchiffon",
                                        ifelse(str_detect('Congo',
region),'lemonchiffon','white'))))))

# set the colors
ggplot(thismap, aes(long, lat, fill = fill, group=group)) +
  geom_polygon(colour="gray") + ggtitle("Map of World") +
  scale_fill_identity()
k20 <- kmeans(region20all, centers = 4)
k20$cluster
cluster3_2020 <- happy20 |> filter(Region %in% c('East Asia', 'Western
Europe', 'North America and ANZ')) |> select('Country')
cluster1_2020 <- happy20 |> filter(Region %in% c('South Asia', 'Middle East
and North Africa', 'Commonwealth and Independent States', 'Southeast Asia'))
|> select('Country')

cluster4_2020 <- happy20 |> filter(Region %in% c('Central and Eastern
Europe', 'Latin America and Caribbean')) |> select('Country')
cluster2_2020 <- happy20 |> filter(Region %in% c('Sub-Saharan Africa')) |>
select('Country')
thismap = map_data("world")

cluster3_2020<- cluster3_2020 |> mutate(
  Country1 = case_when(Country =='Taiwan Province of China' ~ 'Taiwan',
                       Country == 'United States'~ 'USA',
                       Country == 'United Kingdom'~ 'UK',
                       TRUE ~ Country))


cluster1_2020 <- cluster1_2020 |> mutate(Country1 =
                                           case_when(
                                             Country =='Palestinian
Territories' ~ 'Palestine',
                                             TRUE ~ Country))
cluster2_2020 <- cluster2_2020 |> mutate(Country1 = Country)
cluster4_2020 <- cluster4_2020 |> mutate(Country1 = Country)

#set colors by the clusters of countries
thismap <- mutate(thismap, fill =ifelse(region %in% cluster1_2020$Country1
,'lavender',
                                        ifelse(region %in%
cluster2_2020$Country1, "wheat",
                                        ifelse(region %in%
cluster3_2020$Country1,"darkslategrey",
                                        ifelse(region %in%
cluster4_2020$Country1,"honeydew",
                                        ifelse(str_detect('Congo',
region),'wheat','white'))))))

# set the colors
ggplot(thismap, aes(long, lat, fill = fill, group=group)) +
  geom_polygon(colour="gray") + ggtitle("Map of World") +
  scale_fill_identity()
k21 <- kmeans(region21all, centers = 4)
k21$cluster

cluster4_2021 <- happy21 |> filter(Region %in% c('East Asia', 'Western
Europe', 'North America and ANZ')) |> select('Country')
cluster2_2021 <- happy21 |> filter(Region %in% c('South Asia')) |>
select('Country')

cluster3_2021 <- happy21 |> filter(Region %in% c('Central and Eastern
Europe', 'Latin America and Caribbean', 'Middle East and North Africa',
'Commonwealth and Independent States', 'Southeast Asia')) |>
select('Country')
cluster1_2021 <- happy21 |> filter(Region %in% c('Sub-Saharan Africa')) |>
select('Country')


thismap = map_data("world")

cluster4_2021 <- cluster4_2021 |> mutate(
  Country1 = case_when(Country =='Taiwan Province of China' ~ 'Taiwan',
                       Country == 'United States'~ 'USA',
                       Country == 'United Kingdom'~ 'UK',
                       TRUE ~ Country))


cluster3_2021 <- cluster3_2021 |> mutate(Country1 =
                                           case_when(
                                             Country =='Palestinian
Territories' ~ 'Palestine',
                                             TRUE ~ Country))
cluster2_2021 <- cluster2_2021 |> mutate(Country1 = Country)
cluster1_2021 <- cluster1_2021 |> mutate(Country1 = Country)

#set colors by the clusters of countries
thismap <- mutate(thismap, fill =ifelse(region %in% cluster1_2021$Country1
,'rosybrown',
                                        ifelse(region %in%
cluster2_2021$Country1, "tan",
                                        ifelse(region %in%
cluster3_2021$Country1,"lightcyan4",
                                        ifelse(region %in%
cluster4_2021$Country1,"lightskyblue1",
                                        ifelse(str_detect('Congo',
region),'rosybrown','white'))))))

# set the colors
ggplot(thismap, aes(long, lat, fill = fill, group=group)) +
  geom_polygon(colour="gray") + ggtitle("Map of World") +
  scale_fill_identity()

kable(region21alltibble)
```