# World Happiness Report during year 2015 ~ year 2020

Mark Wu, Ruhani Arora, Abhishek Grewal, Yumian Cui

05 May, 2021

Introduction

This data contains the Happiness Score for 153 countries along with the factors used to explain the score.T he happiness score of a country is dependent on the following factors : GDP per capita,, Healthy Life Expectancy ,Social support, Freedom to make life choices, Generosity, Corruption Perception, and Residual error which is given in the data and on the basis of which countries are provided with the rankings with Finland being at the top. We believe that this data source is interesting because happiness is one of the most important factor for growth in any country. By studying these factors it can help us gain knowledge and insight about the important factors that influence the happiness index of a country. Apart from the happiness index, the data gives information about the GDP per capita, Healthy Life Expectancy ,Social support, Freedom to make life choices, Generosity, and Corruption Perception on the basis of which we can see much progress a country is making. Following the trend and patterns of the factors which creates a positive growth in the happiness can help counties at the bottom of the list to make improvements in the factors affecting the happiness index.Since we are studying the data of 2020, the year when the Covid-19 pandemic hit the world it can provide a lot more detailed about how countries manage to maintain their happiness index. In years prior to 2020 or COVID, happiness score is expected to increase year by year with GDP per capita increasing in all countries, and GDP per capita plays a bigger role in affecting the happiness index in those developed countries(high latitudes). And in 2020, accounting for all the rest of factors determining the happiness score, GDP per capita is seen as a less significant predictor because GDP per capita growth for developed countries is also assumably slowed down, we believe people tend to migrate to countries with greater happiness scores and per capita GDP in search of better opportunities and countries with more people migrating where harder hit by COVID.

Data

1. https://www.kaggle.com/mathurinache/world-happiness-report?select=2015.csv
2. https://www.kaggle.com/mathurinache/world-happiness-report?select=2016.csv
3. https://www.kaggle.com/mathurinache/world-happiness-report?select=2017.csv
4. https://www.kaggle.com/mathurinache/world-happiness-report?select=2018.csv
5. https://www.kaggle.com/mathurinache/world-happiness-report?select=2019.csv
6. https://www.kaggle.com/mathurinache/world-happiness-report?select=2020.csv
7. https://www.kaggle.com/tanuprabhu/population-by-country-2020

First six dataset credit to original author: John Helliwell, Richard Layard, Jeffrey D. Sachs, and Jan Emmanuel De Neve.

Last dataset credit to original author: Tanu N Prabhu

First we did the data cleaning and separate in 5 group based on happiness rank in each interval 30 ranks in each interval.

```
head(data)
```

```
#       Country happiness_rank happiness_score     gdp life_expectancy freedom
# 1 Switzerland              1           7.587 1.39651         0.94143 0.66557
# 2     Iceland              2           7.561 1.30232         0.94784 0.62877
# 3     Denmark              3           7.527 1.32548         0.87464 0.64938
# 4      Norway              4           7.522 1.45900         0.88521 0.66973
# 5      Canada              5           7.427 1.32629         0.90563 0.63297
# 6     Finland              6           7.406 1.29025         0.88911 0.64169
#     trust generosity year level
# 1 0.41978    0.29678 2015 First
# 2 0.14145    0.43630 2015 First
# 3 0.48357    0.34139 2015 First
# 4 0.36503    0.34699 2015 First
# 5 0.32957    0.45811 2015 First
# 6 0.41372    0.23351 2015 First
```

Methods

1. Linear Regression
2. Tree
3. K means Clustering

First,the reason why we use "Linear Regression Model" is that we want to predict the dependent variable (Happiness score) base on the predictors variable(GDP/captia,Life expectancy,Freedom, Trust, Generosity) and to see which is significant predictors of the outcome variable.

Secondly,the reason why we use "Tree Model" is that we want to show the audience can easily understand the which predictor variable are more significant by observing the dendrogram.

Third, the reason why we use "K-Means Clustering" is that it is hard to observe a definite pattern in the scatterplots, to get a better idea of how countries can be grouped, we tried to using the unsupervised learning method to group them.

Results

**Show the summary of happiness stat in each year during 2015-2020**

```
data_x
```

```
#    Country          happiness_rank   happiness_score      gdp
#  Length:934         Min.   :  1.00   Min.   :2.567   Min.   :0.0000
#  Class :character   1st Qu.: 40.00   1st Qu.:4.537   1st Qu.:0.5995
#  Mode  :character   Median : 78.50   Median :5.350   Median :0.9738
#                     Mean   : 78.48   Mean   :5.393   Mean   :0.9070
#                     3rd Qu.:117.00   3rd Qu.:6.197   3rd Qu.:1.2272
#                     Max.   :158.00   Max.   :7.809   Max.   :1.8708
#  life_expectancy     freedom           trust          generosity
#  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
#  1st Qu.:0.4493   1st Qu.:0.3182   1st Qu.:0.05455   1st Qu.:0.1270
#  Median :0.6672   Median :0.4361   Median :0.09266   Median :0.1995
#  Mean   :0.6255   Mean   :0.4198   Mean   :0.12630   Mean   :0.2138
```

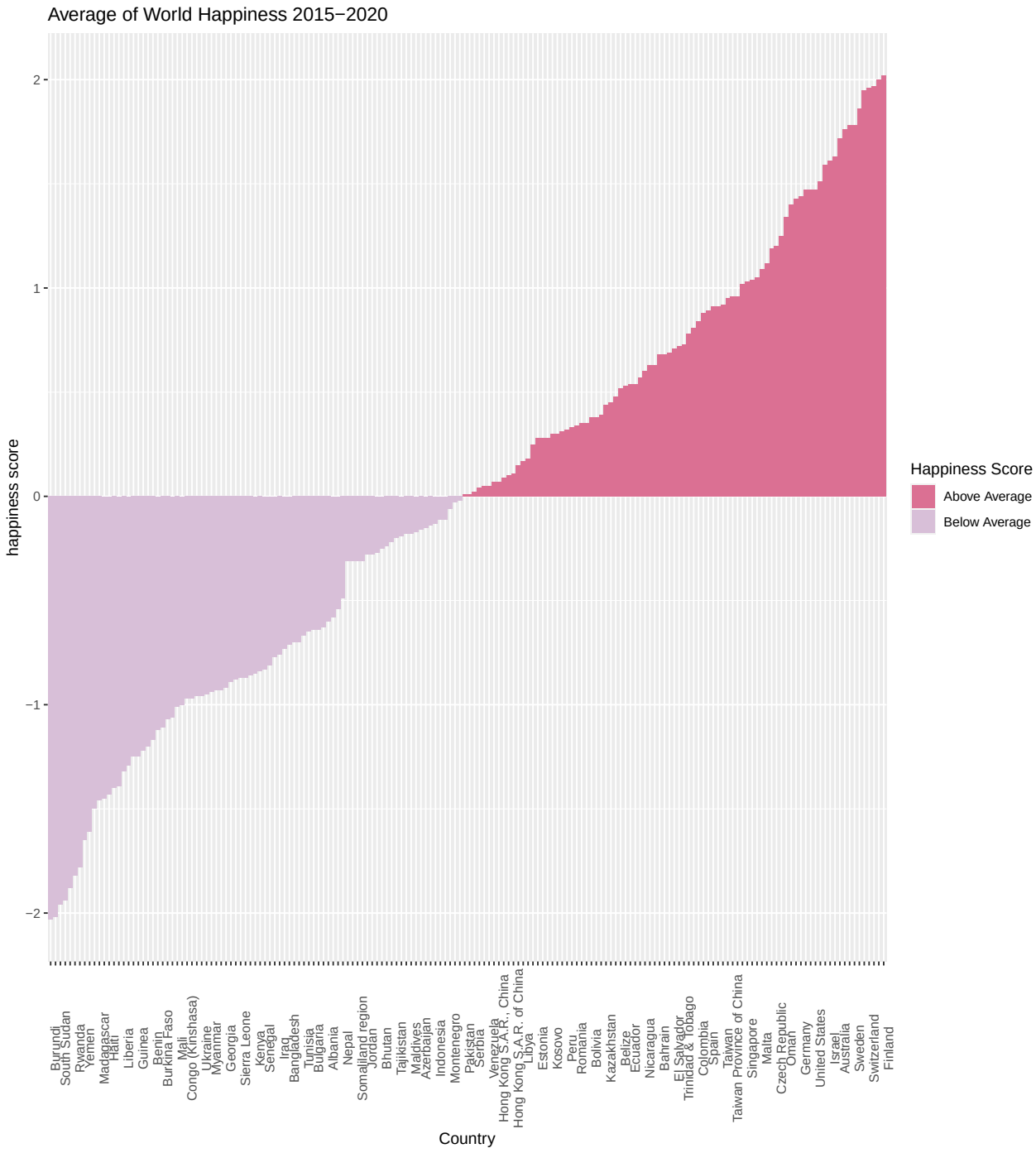
```
#  3rd Qu.:0.8148   3rd Qu.:0.5399   3rd Qu.:0.15787   3rd Qu.:0.2715
#  Max.   :1.1410   Max.   :0.7240   Max.   :0.55191   Max.   :0.8381
#    year        level
#  2015:158   First :179
#  2016:157   Second:180
#  2017:155   Third :180
#  2018:155   Fourth:180
#  2019:156   Fifth :215
#  2020:153
```

**Show the summary of happiness score in 2015-2020**

```
happiness_stats
```

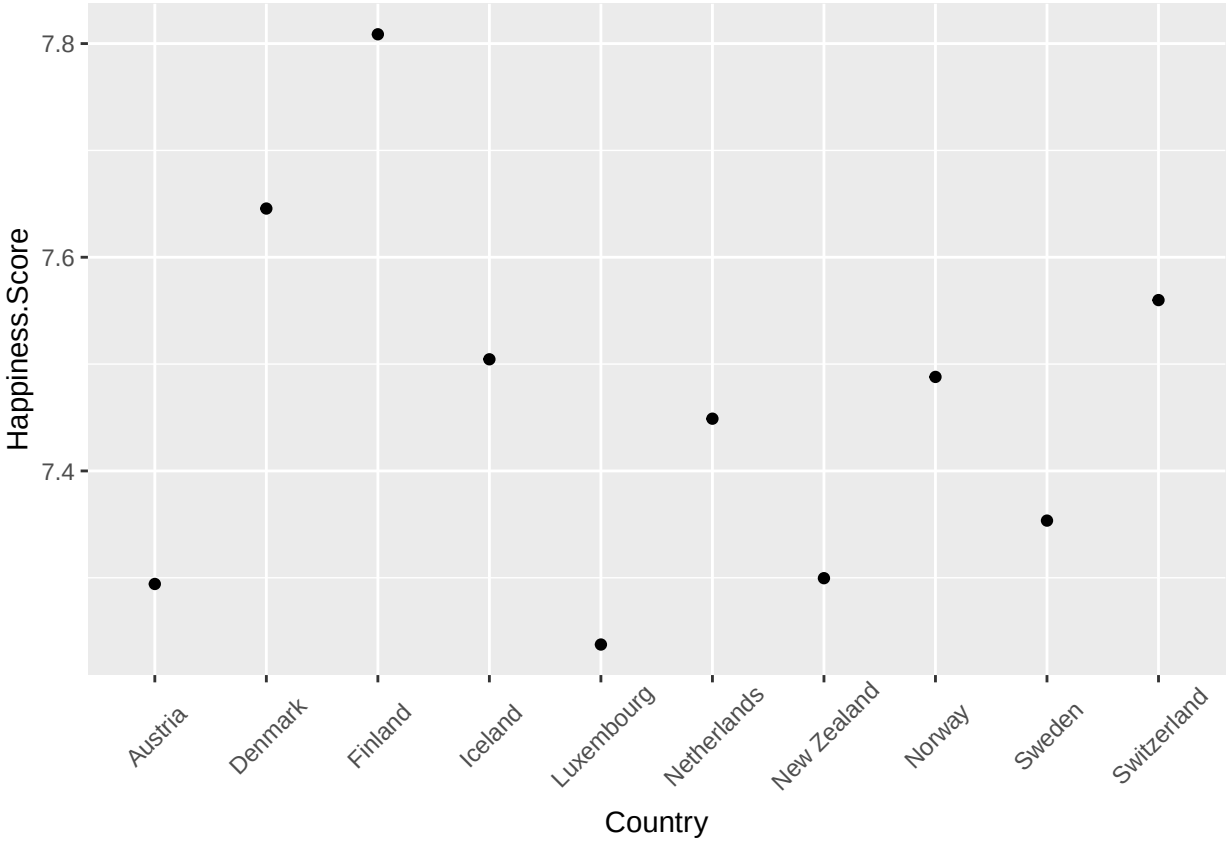
```
#   mean(happiness_score) median(happiness_score) sd(happiness_score)
# 1              5.392959                 5.35025             1.12463
#   min(happiness_score) max(happiness_score)
# 1               2.5669               7.8087
```

**To show if the happiness score above or below the average**

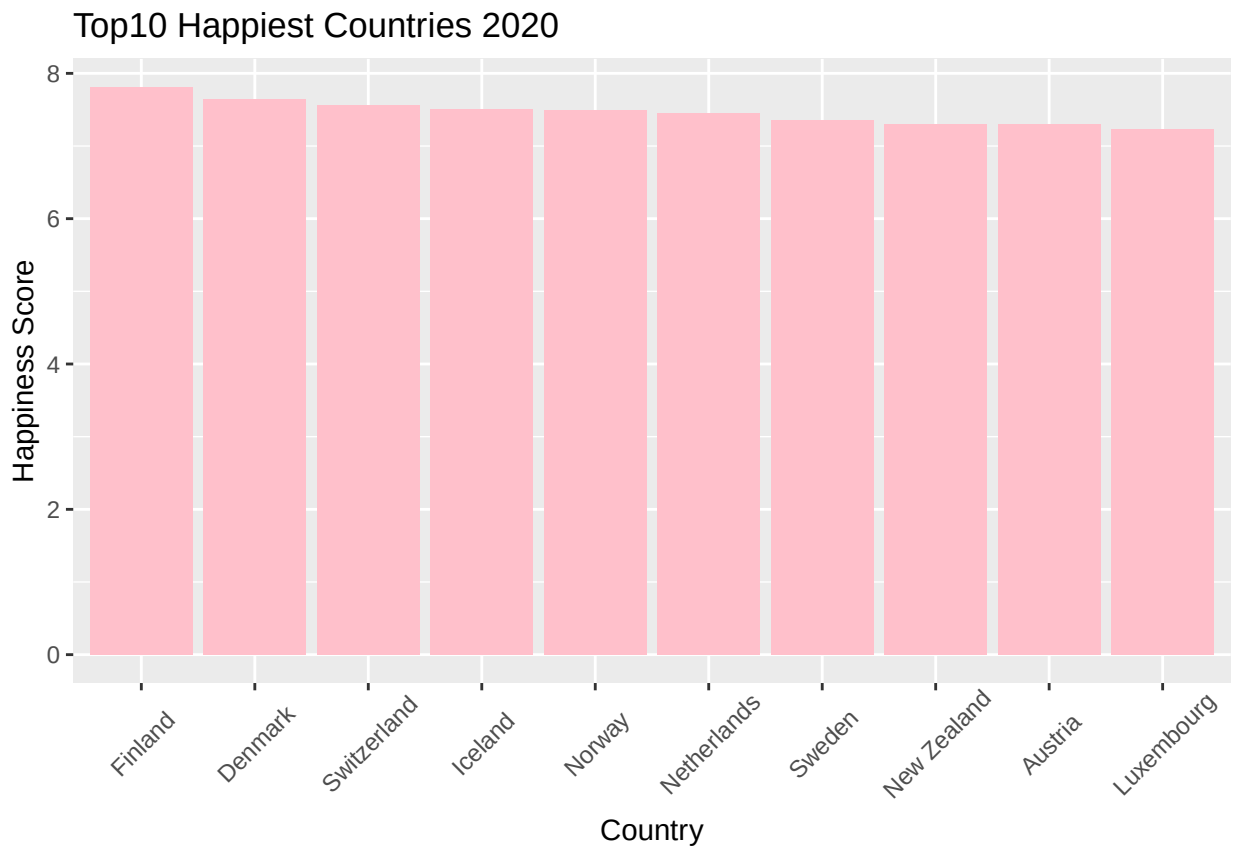


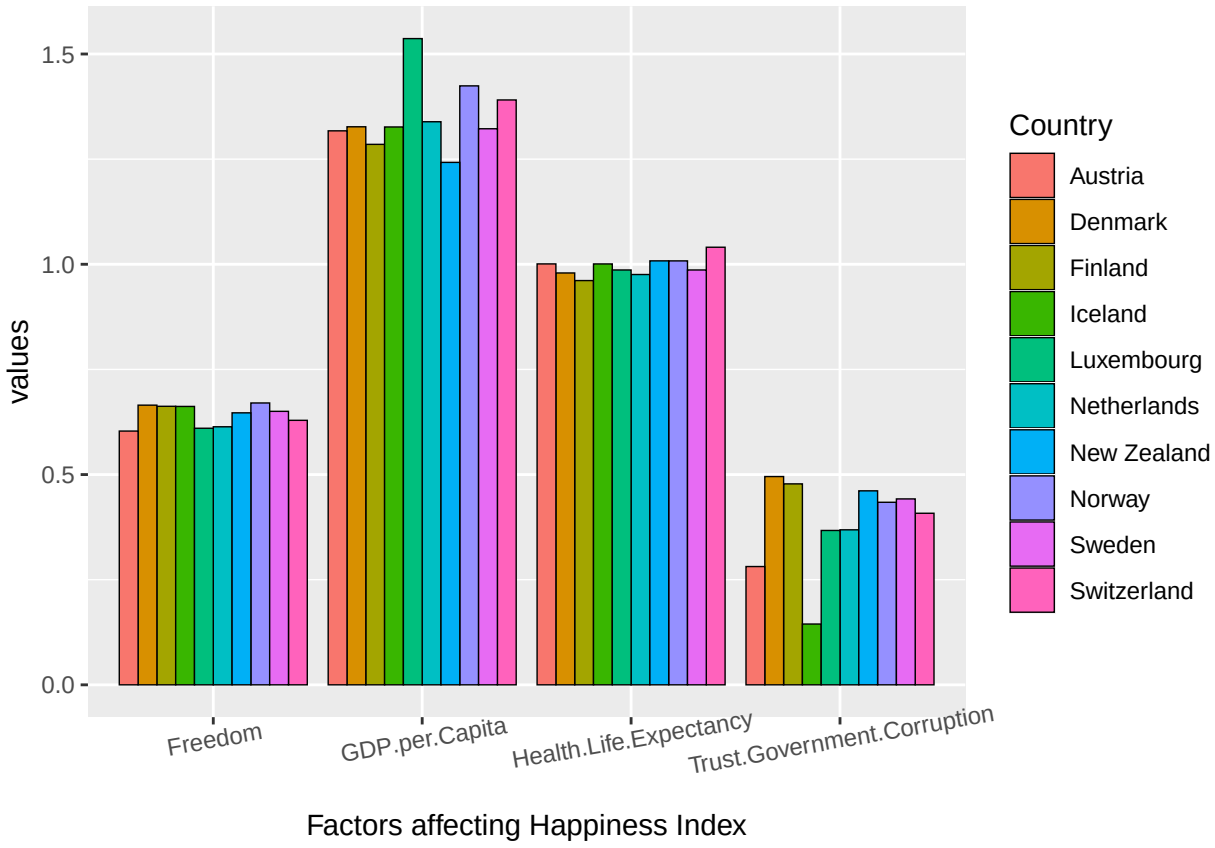

From our above graph,we order the country by happiness score, and we easily found that "Burndi" is the least happiest country during 2015~2020, and Finland is the happiest country during 2015~2020 through observing the plot.

**The distribution of Top 10 in 2020**

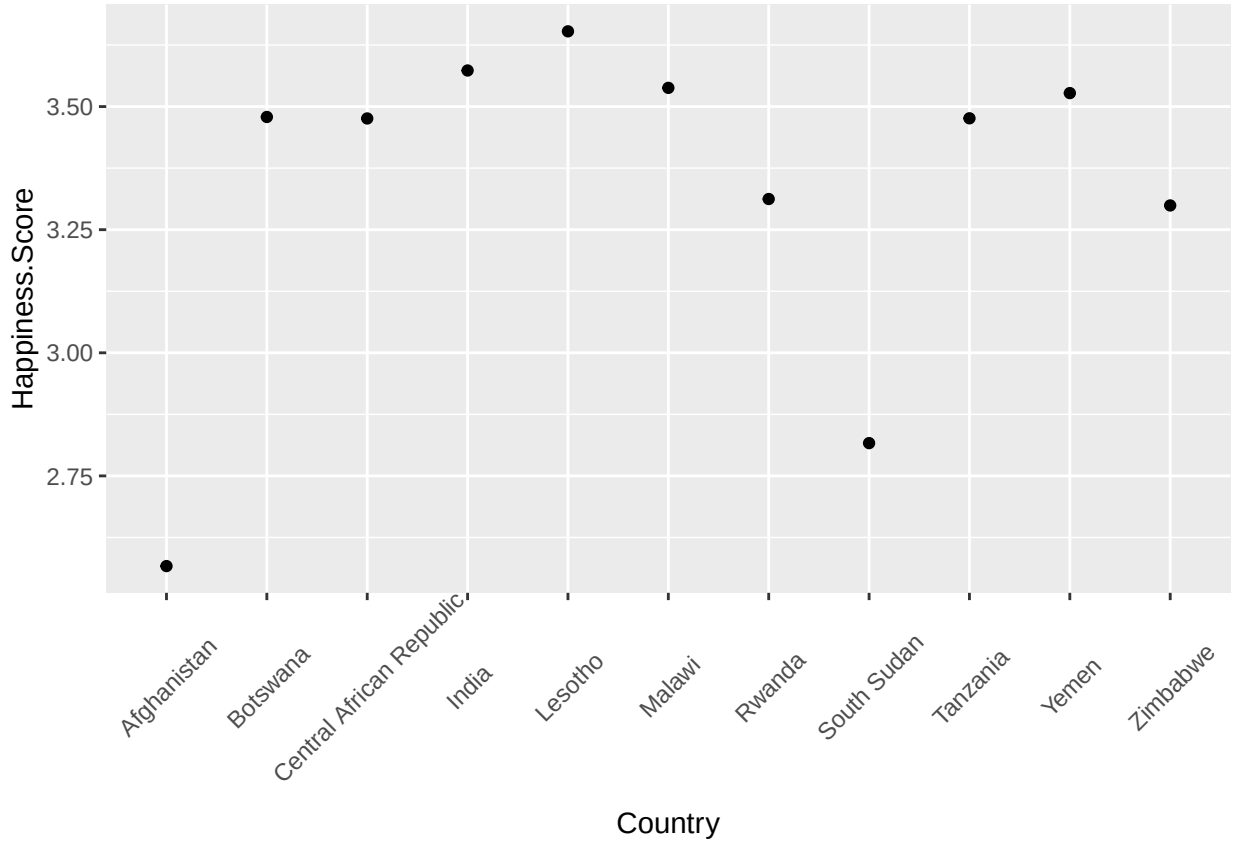


**Order the distribution of Top 10 in 2020**

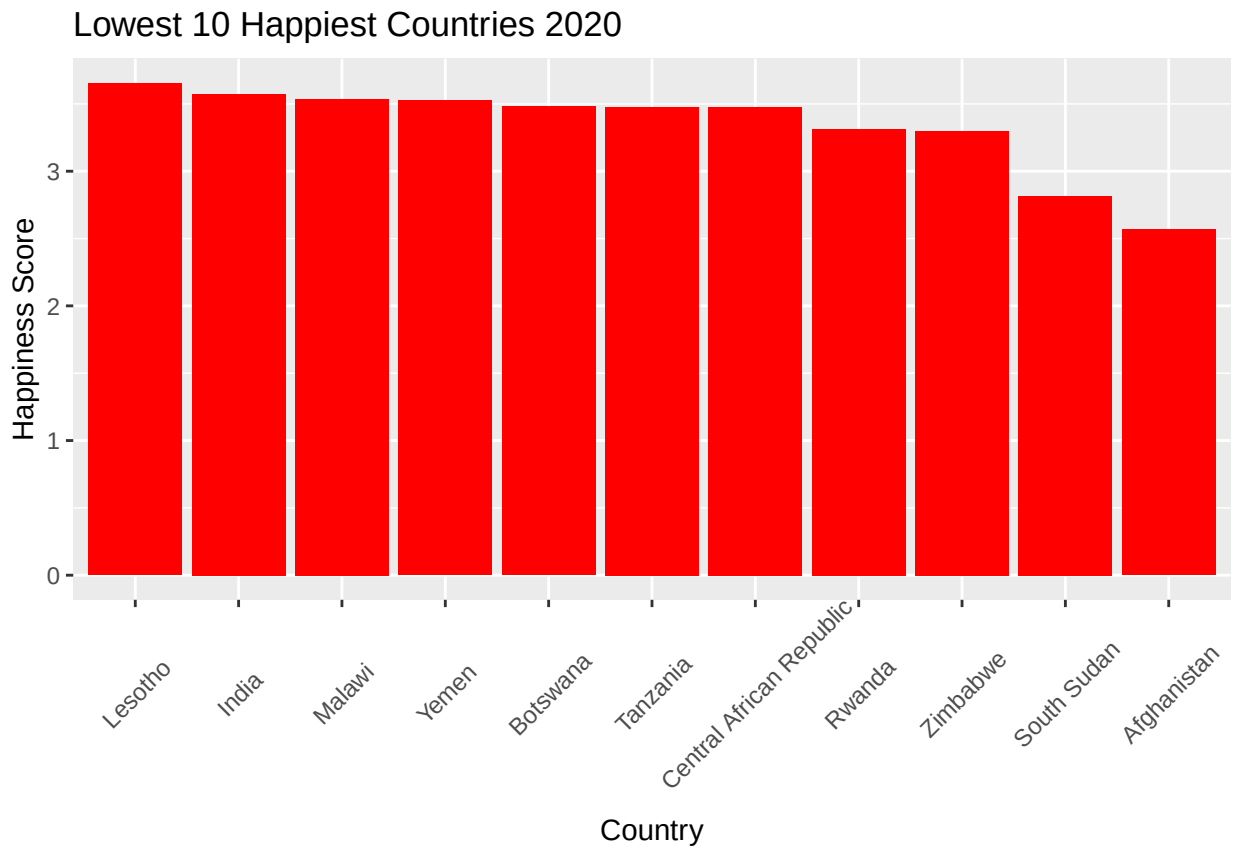


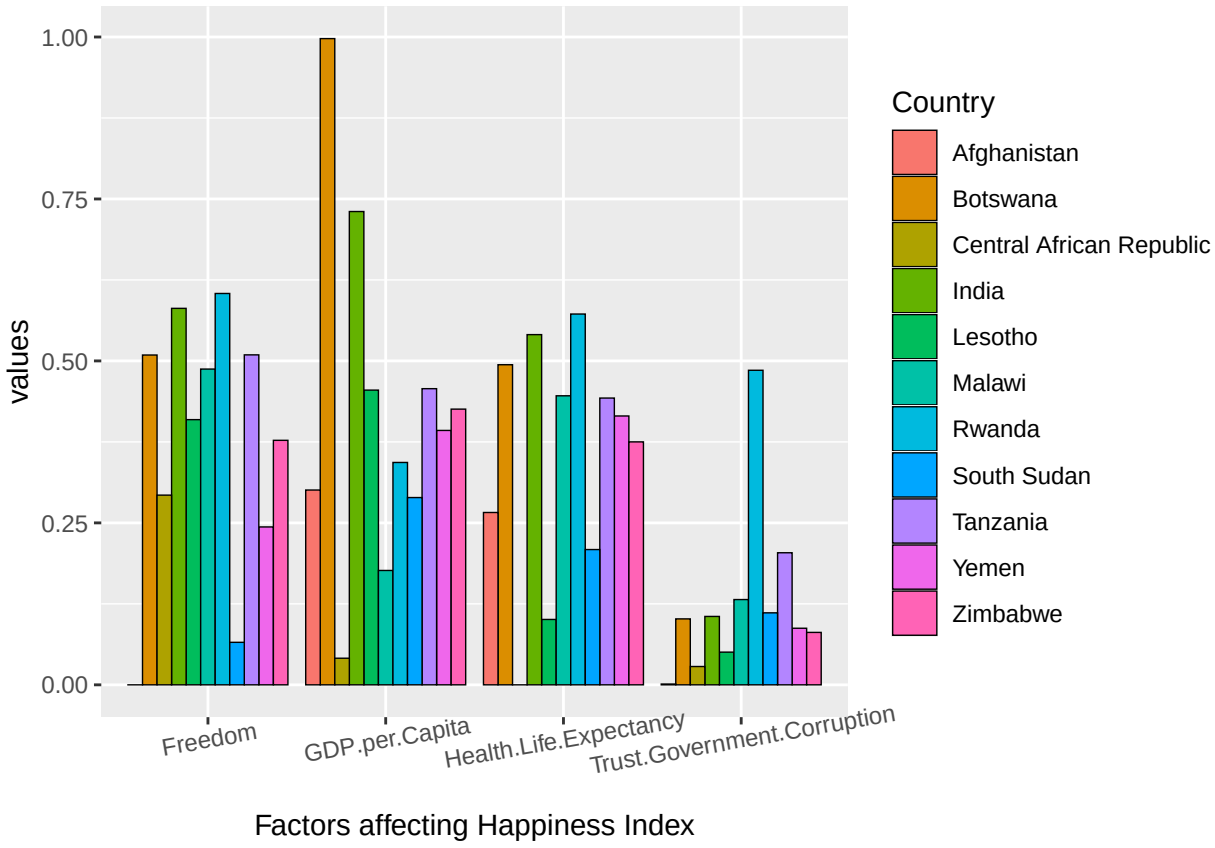

values
1.5
1.0
0.5
0.0
Freedom
GDP.per.Capita
Health.Life.Expectancy
Trust.Government.Corruption
Factors affecting Happiness Index
Country
Austria
Denmark
Finland
Iceland
Luxembourg
Netherlands
New Zealand
Norway
Sweden
Switzerland

**The distribution of low 10 in 2020**



**OrderThe distribution of low 10 in 2020**

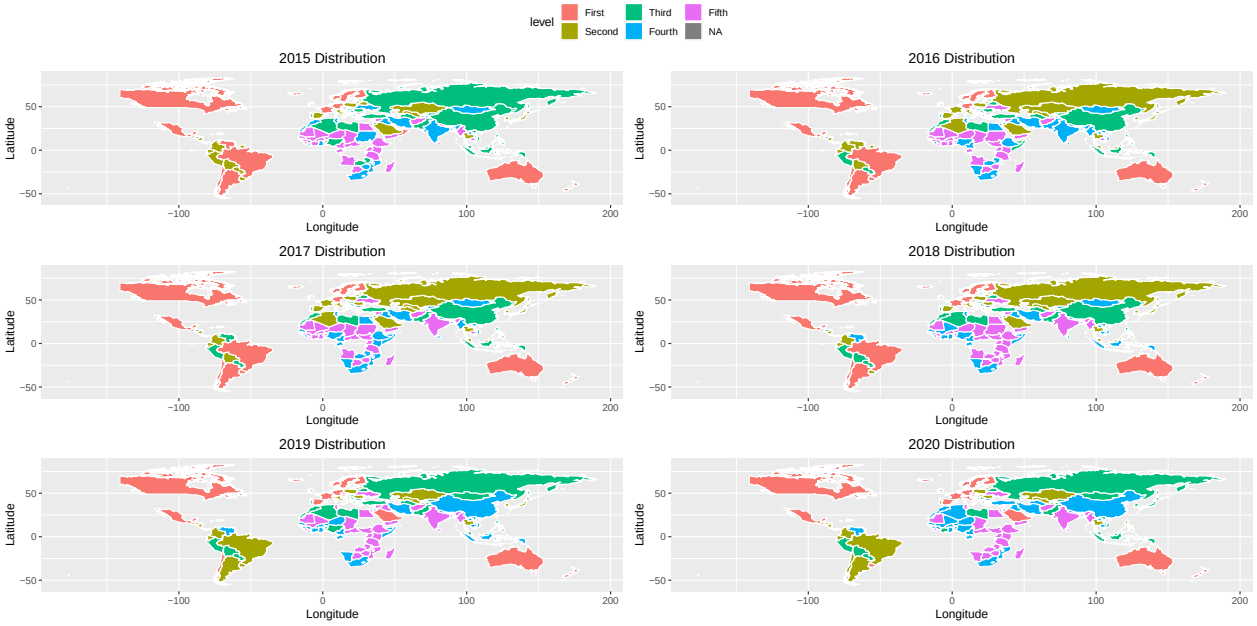





Factors affecting Happiness Index

After analyzing the happiness index of the world from the past 5 year we can come to this conclusion that GDP has the most impact on a country's happiness score and rank. While plotting the graph of happiness score of top 10 countries and the bottom 10 countries, we can see that the countries in the list of top 10 have a high GDP followed by high life expectancy whereas countries on the bottom have low gpd value.

**Show the distribution of happiness score by the 5 group in the world map**




We can see the the development of happiness score in each country, we can obviously see that in the nations with High latitudes would always in the first level, which is the first 30 rank. Also, most of the African country are in the fifth rank, which showed that they're not really happy.

**Using linear regression in years prior to 2020 or COVID**

```
lm1<-lm(happiness_score~.*.,prior_20)
summary(lm1)
```

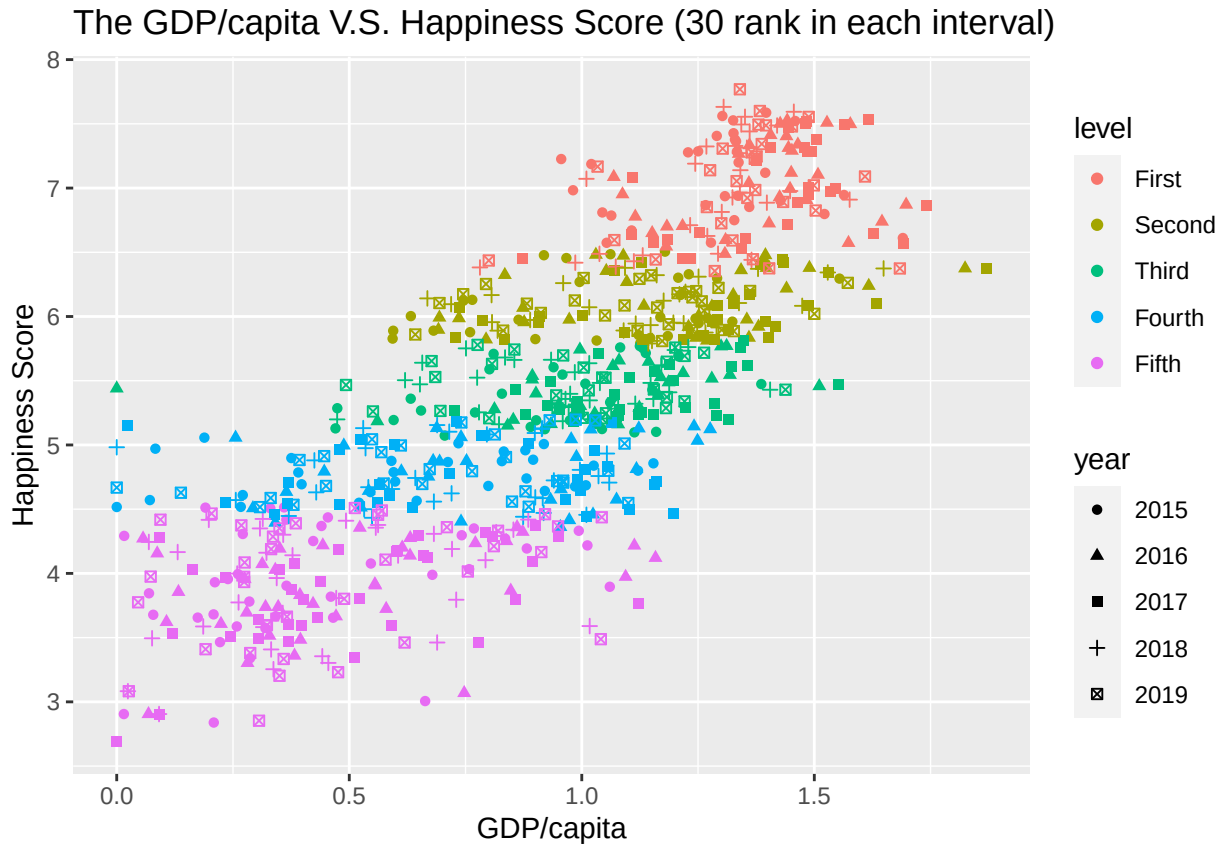
```
#
# Call:
# lm(formula = happiness_score ~ . * ., data = prior_20)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -1.80972 -0.36198  0.05453  0.34493  1.49798
#
# Coefficients:
#                             Estimate Std. Error t value Pr(>|t|)
# (Intercept)                   3.4680     0.2247  15.437  < 2e-16 ***
# gdp                           0.8358     0.2867   2.915  0.00366 **
# life_expectancy               0.1151     0.4742   0.243  0.80829
# freedom                       0.2422     0.5677   0.427  0.66976
# trust                        -1.2425     1.1318  -1.098  0.27263
# generosity                   -0.3215     0.7263  -0.443  0.65813
# gdp:life_expectancy           0.6204     0.2623   2.365  0.01828 *
# gdp:freedom                  -0.4213     0.6750  -0.624  0.53271
# gdp:trust                    -0.2215     0.8264  -0.268  0.78869
# gdp:generosity                1.5020     0.7880   1.906  0.05702 .
```

```
# life_expectancy:freedom      2.7840     0.9843   2.828  0.00480 **
# life_expectancy:trust       -0.3078     1.6037  -0.192  0.84784
# life_expectancy:generosity  -2.5586     1.4889  -1.718  0.08613 .
# freedom:trust                2.2952     1.8960   1.211  0.22643
# freedom:generosity           1.3892     1.3160   1.056  0.29148
# trust:generosity             2.0554     1.9291   1.065  0.28700
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.5613 on 765 degrees of freedom
# Multiple R-squared:  0.7568, Adjusted R-squared:  0.752
# F-statistic: 158.7 on 15 and 765 DF,  p-value: < 2.2e-16
```

**According to the model in years prior to 2020 or COVID, what happiness score would I expect on average with a country that has gdp=1.5,life_expectancy=0.9,freedom=0.45,trust=0.3,generosity=0.5?**

```
test_data =data.frame(gdp=1.5,life_expectancy=0.9,freedom=0.45,trust=0.3,generosity=0.5)
predict(lm1,test_data)
```

```
#        1
# 6.804502
```




From our linear model above, we found that the GDP factor are significant in this model, then we made a plot to see the The GDP/capita V.S. Happiness Score. However, we have strong evidenve to say that in

years prior to 2020 or COVID, happiness score is expected to increase with GDP per capita increasing in all countries but not incresing year by year via the above plot.

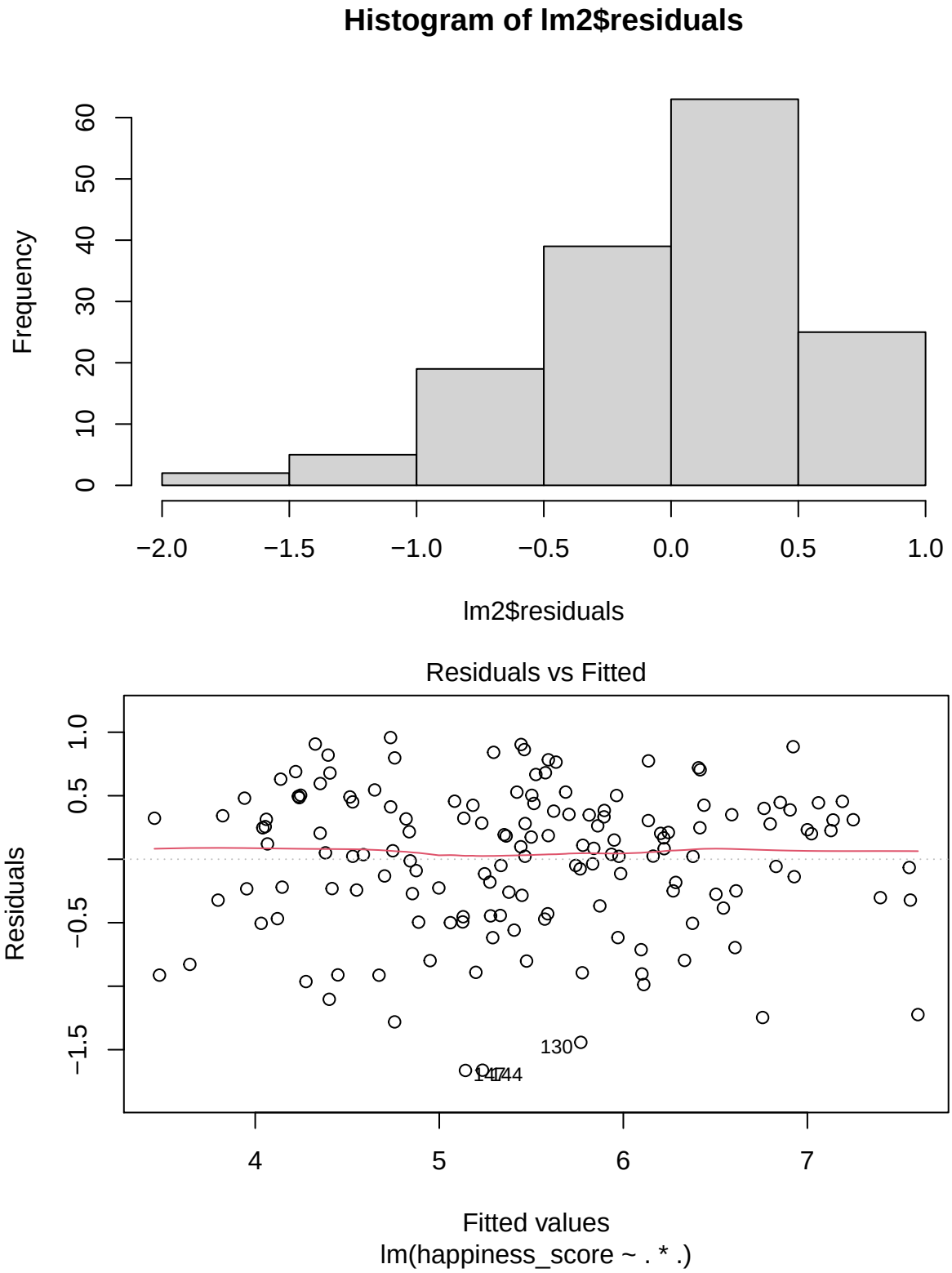
**Using linear regression in years in 2020**

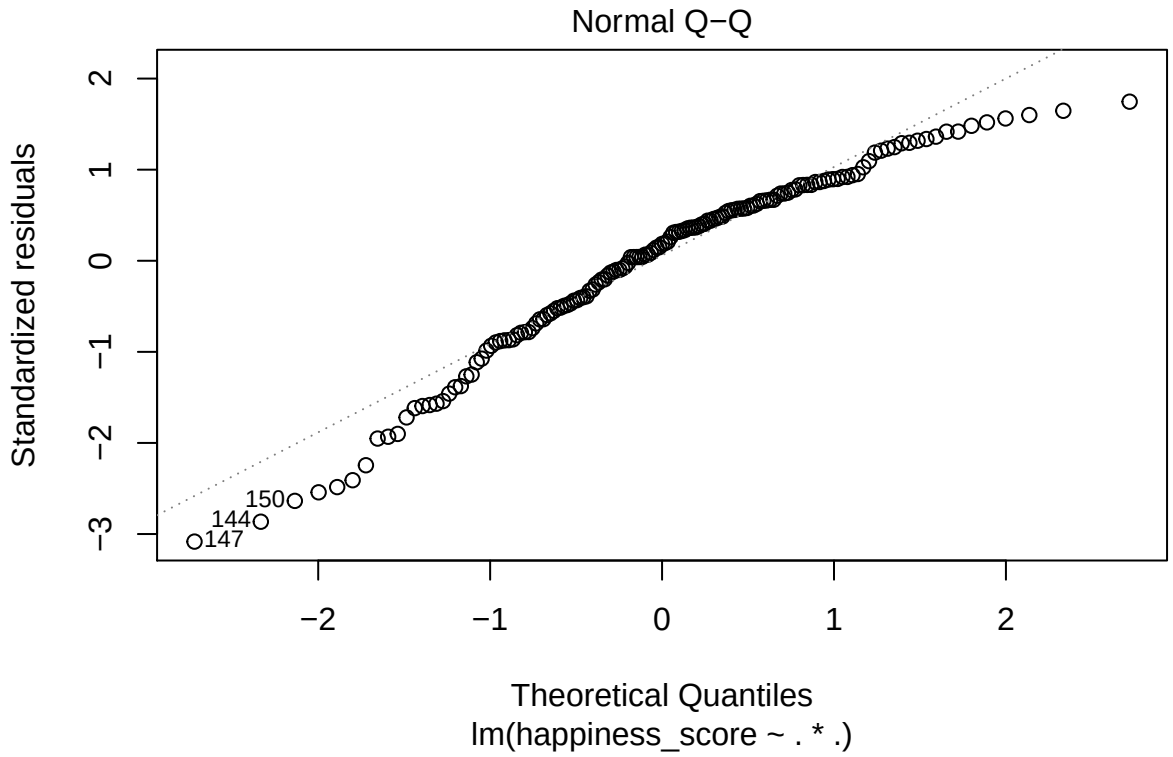
```
lm2<-lm(happiness_score~.*.,in_20)
summary(lm2)
```

```
#
# Call:
# lm(formula = happiness_score ~ . * ., data = in_20)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -1.66342 -0.32197  0.09837  0.39993  0.95791
#
# Coefficients:
#                              Estimate Std. Error t value Pr(>|t|)
# (Intercept)                   3.13856    0.66408   4.726 5.61e-06 ***
# gdp                           0.15722    1.11752   0.141   0.8883
# life_expectancy              -0.07613    1.58240  -0.048   0.9617
# freedom                       2.49939    1.63052   1.533   0.1276
# trust                        -0.63145    3.04823  -0.207   0.8362
# generosity                    0.95815    2.37881   0.403   0.6877
# gdp:life_expectancy           1.47377    0.76259   1.933   0.0554 .
# gdp:freedom                  -2.43216    2.39574  -1.015   0.3118
# gdp:trust                     7.36480    3.12525   2.357   0.0199 *
# gdp:generosity                1.73255    3.25847   0.532   0.5958
# life_expectancy:freedom       3.29574    3.32535   0.991   0.3234
# life_expectancy:trust       -11.72444    5.79825  -2.022   0.0451 *
# life_expectancy:generosity   -0.06540    4.63561  -0.014   0.9888
# freedom:trust                 3.47158    5.30157   0.655   0.5137
# freedom:generosity           -4.71901    4.43673  -1.064   0.2894
# trust:generosity              2.85043    6.01811   0.474   0.6365
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.5899 on 137 degrees of freedom
# Multiple R-squared:  0.7465, Adjusted R-squared:  0.7187
# F-statistic:  26.9 on 15 and 137 DF,  p-value: < 2.2e-16
```

From our second linear model, in year 2020, we found that the GDP factor are NOT significant in this model anymore, which can support our thesis, GDP per capita is seen as a less significant predictor because GDP per capita growth is also assumably slowed down.

**Multiple linear regression in year 2020**

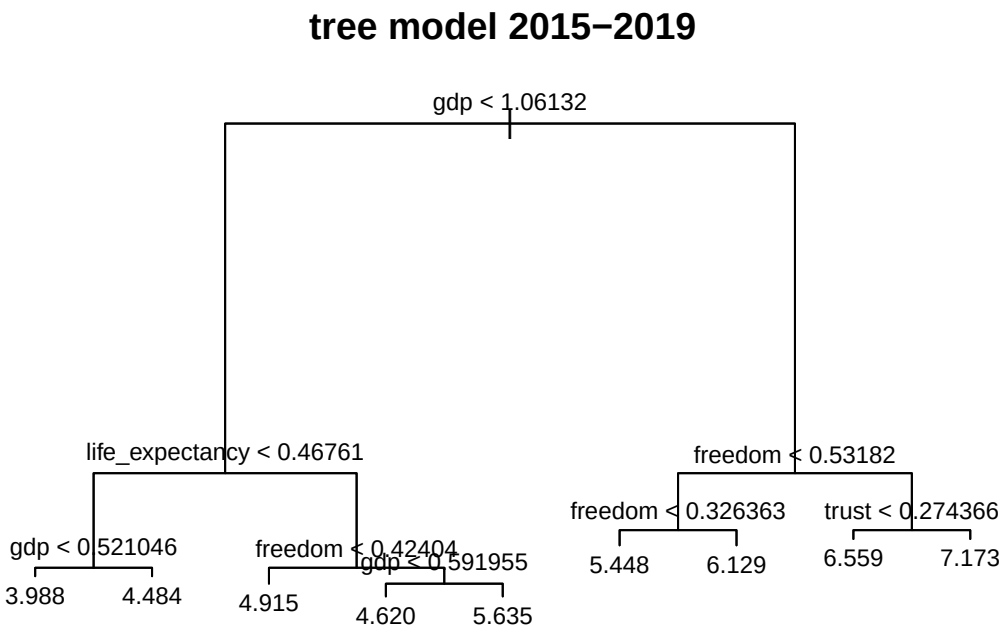


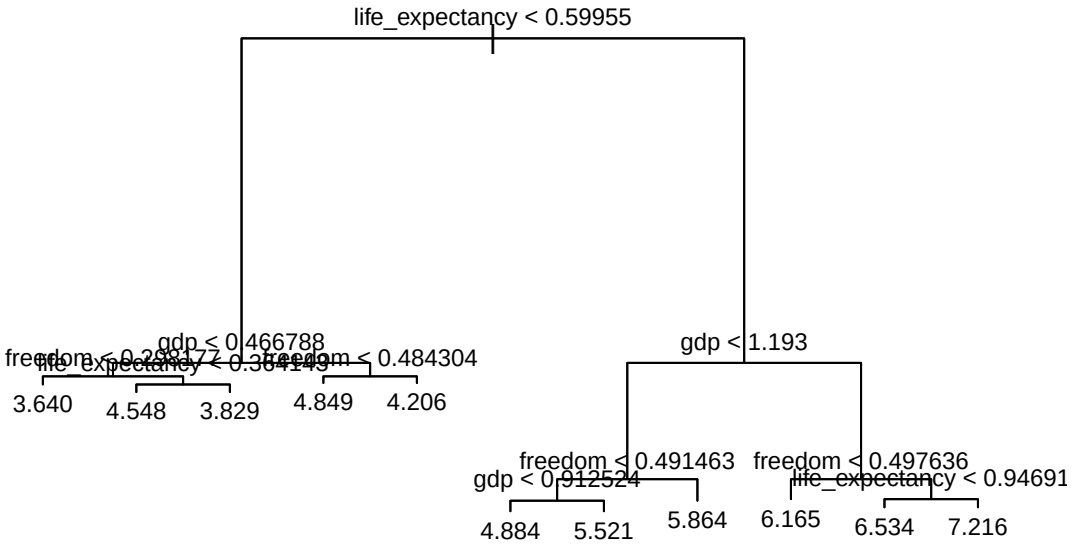



From the Residual VS Fitted, we can see the curve showed the linearity property.

From normal QQ plot we can see that the residual is closed to the fitted line, which showed it satisfied the Normality property.

**Tree**



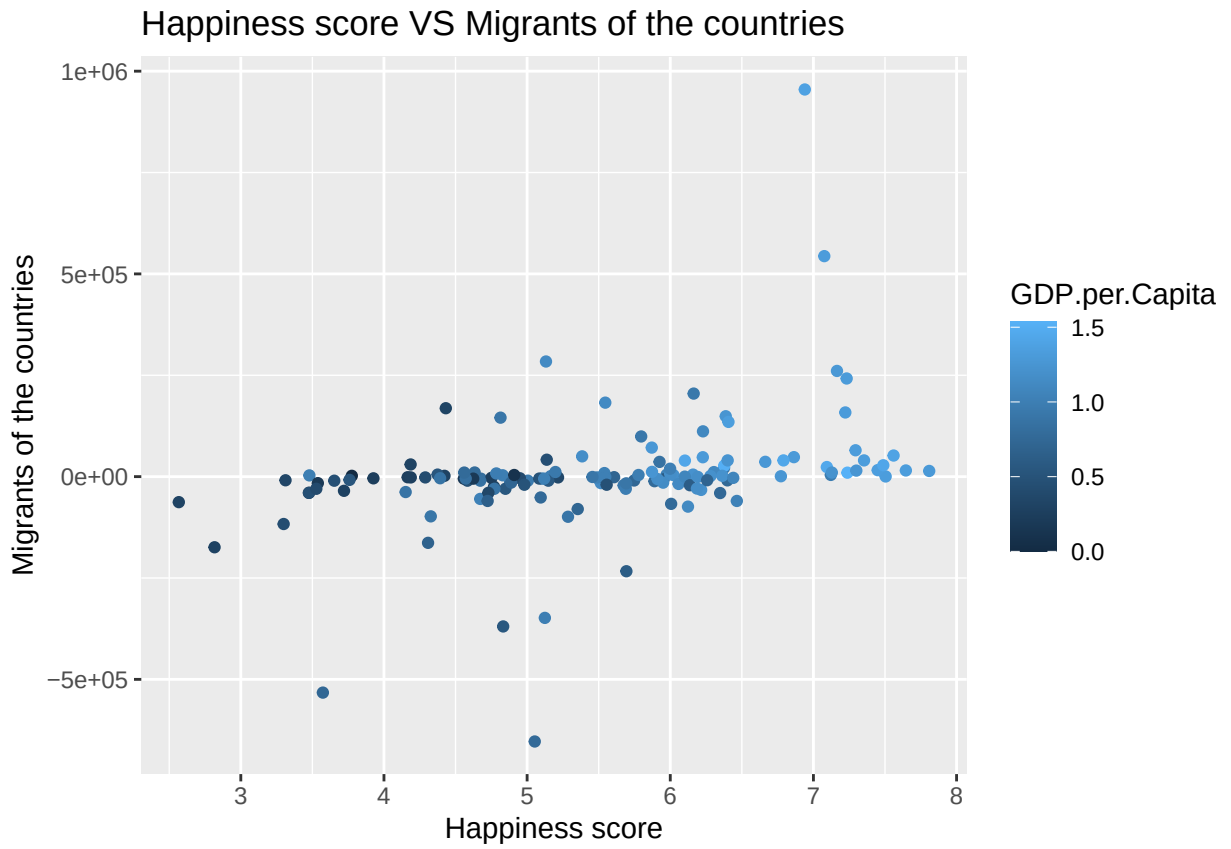



We think this two tree models fitting on data before 2020 and in 2020 corresponds to what has been stated in thesis that GDP is a less significant predictor in 2020 alone than in all previous years. In 2020, it seems that life expectancy is more significant. It may be reasonable to guess that this has to do with the Covid outbreak.

**Migration**

We imported a dataset displaying information about the population of each country. We where interested to know wether people tend to move to places where they can be more happy (measured by happiness score) and places that are more developed and provide more oppertunities (measured by per capita GDP). So we extracted the net migration (number of people entering the country - number of people leabing) for each country and merged it with per capita GDP and happiness score for each country.

**Happiness score VS Migrants of the countries**

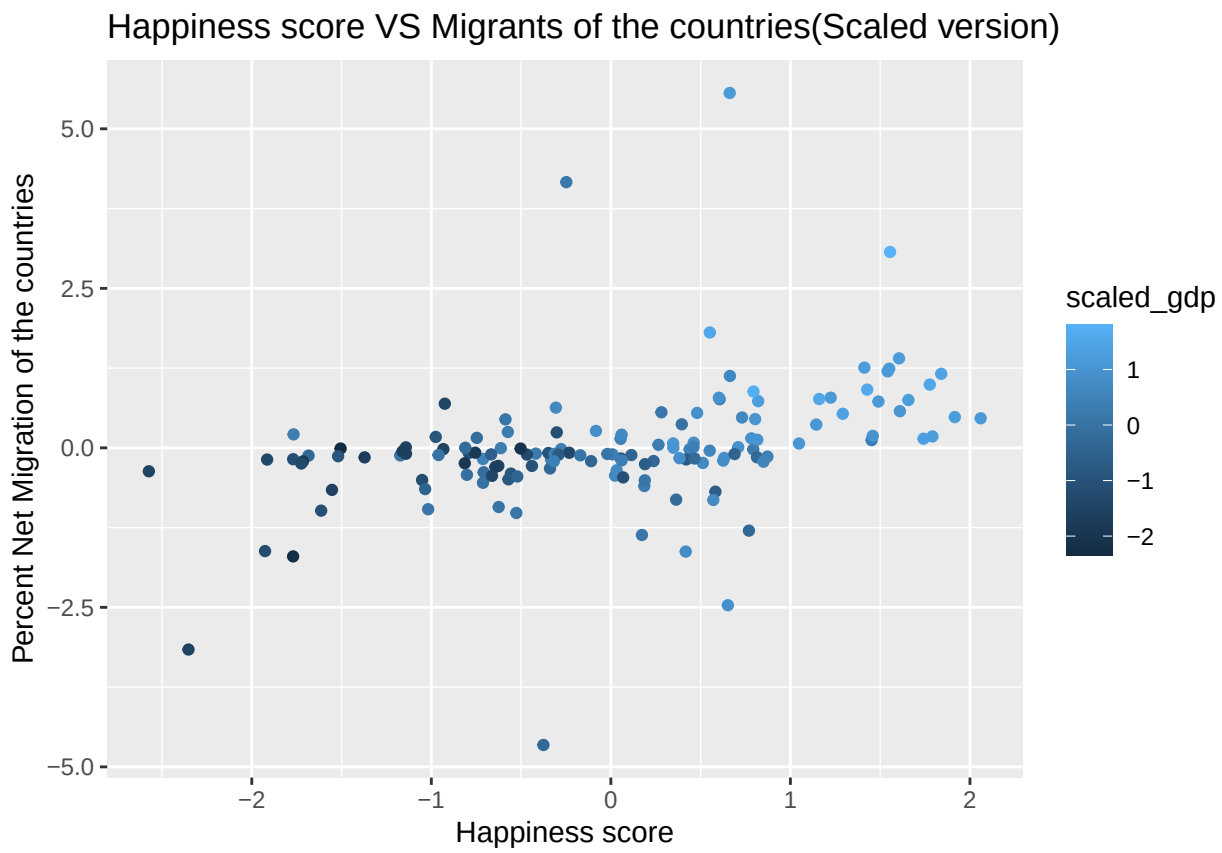



To get an initial idea, we ploted the net migration and happiness score with color dimension as per capita GDP on a scatter plot. We observe that countries with happiness scores below five have negative net migrations staying mostly under 250000 people leaving with only 2 countries exceeding 250000 leaving. For Happiness scores between 5 and 6.5, it is really hard to observe any relevent patterns as there are many countries with negative and positive migration. Countries with happiness scores greater 6.5 all have positive migration with two countries exceeding 500000 and one country close to 1000000 incoming migrants. Countries appearing lighter blue corresponding to higher GDP are found towards the top of the plot signaling that more people want to move to more developed places where they are also likely to be more happy. Another interesting trend observerd is that countries with mass negative migrations exceeding or close to -250000 migrants have a per capita GDP close to 0.5 to 1.0 units.

**Scaled migration data**

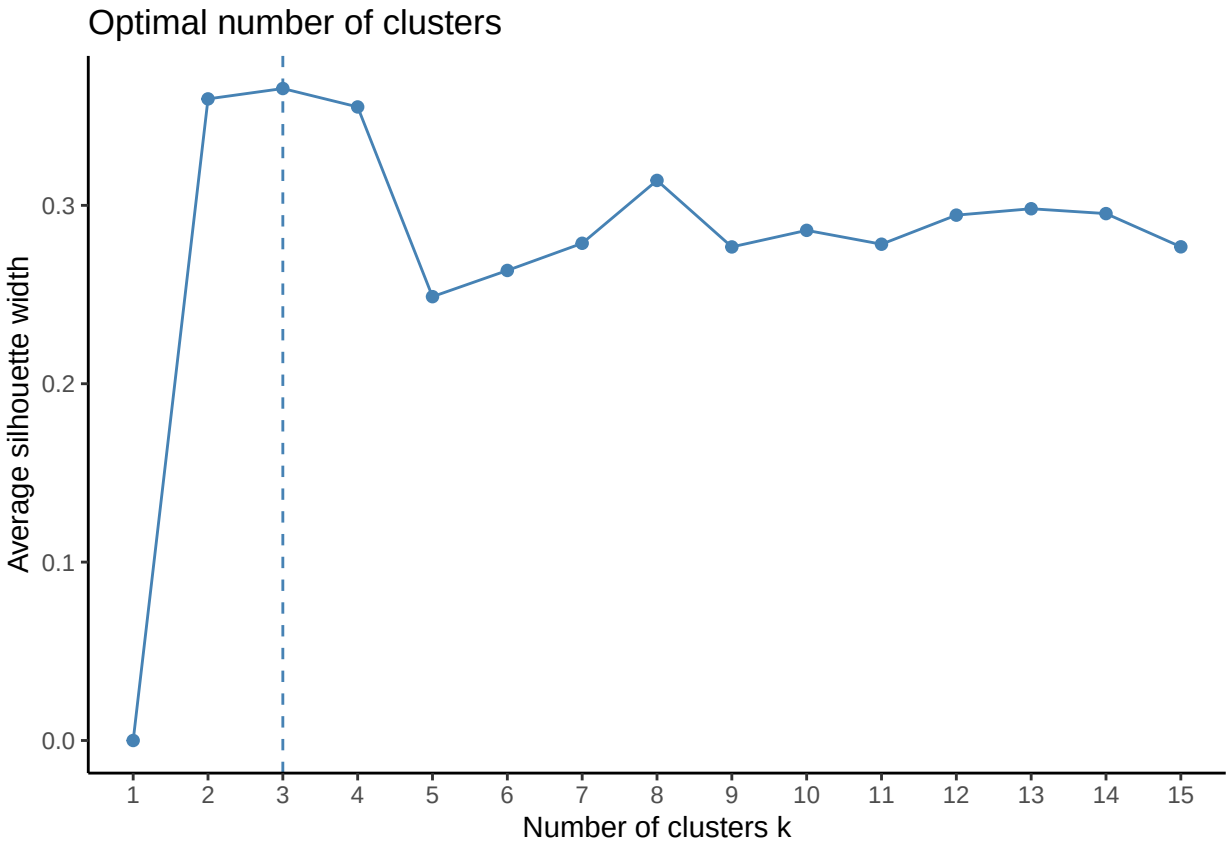
There is no definitive pattern for instance we can not fit a linear or even polynomial eqaution to this plot. To get a better idea we used the unsupervised learning technique of kmeans clustering to group countries into cenrtain categories based on migration, GDPm and happiness score. Inorder to conduct kmeans, we must first scale the data to one standard deviation, which we have done above.

**Happiness score VS Migrants of the countries(Scaled version)**



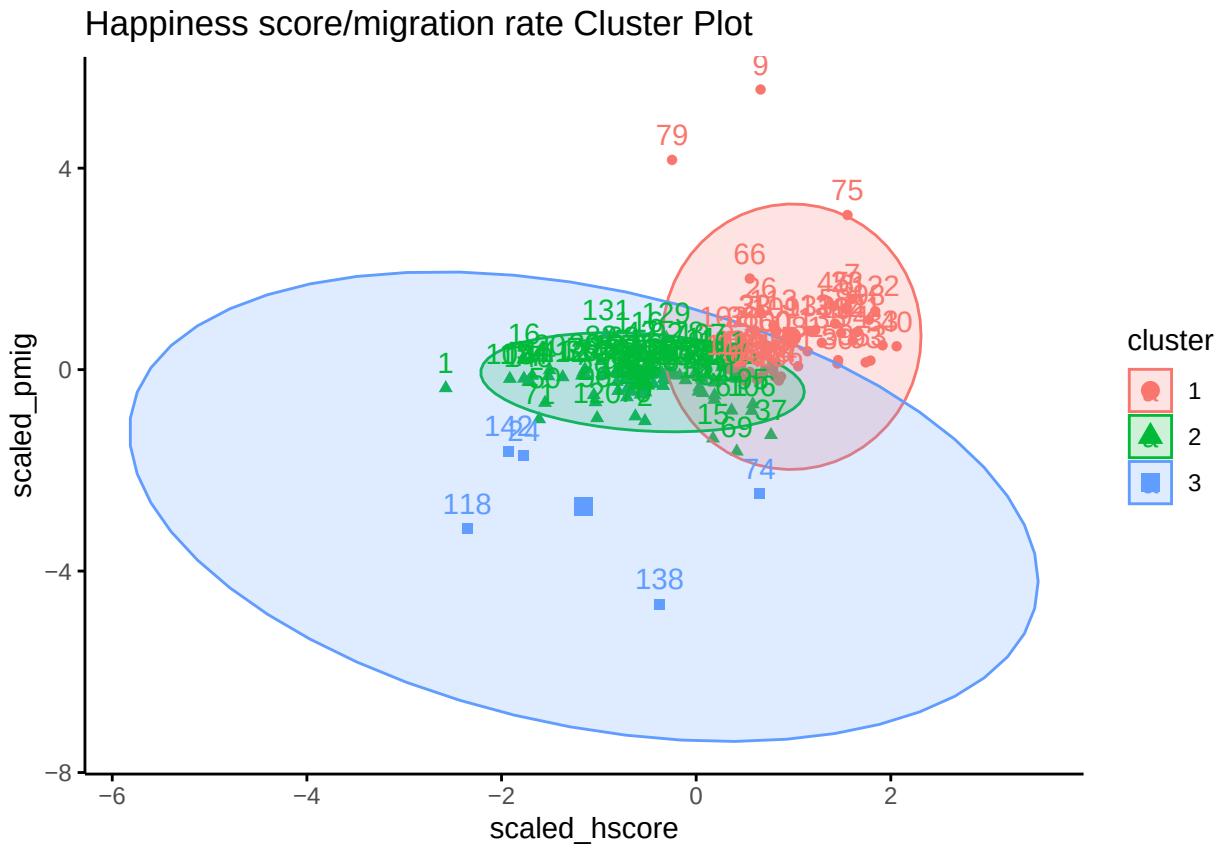

In our previous plot, we used net migration which might have made highly populated countries seem like they lose a lot of people. Now we use what proportion of the total population accounts for the migrants. We also used the scaled version of our happiness score, percent migration, and percent per capita gdp. We now see a much more definitive pattern. Countries with a less than average happiness scores are losing proportions of their populations to migration with only a 4 countries exceeding the proportion of 1.75. Countries within one standard deviation of mean happiness score have both positive and negative migration with the number of positive migrations increaseing as happiness score increases. And a higher net positive migration exceeding or close the proportion of 1.75 is associated with really high per capita gdp. countries with scaled happiness score greater than 1 exclusively have net positive migrations ranging from 0 to 1.75 with one country exceeding 2.5. These countries tend to have the highest per capita GDP as well.

**Choose the optimal number of clusters**




We use Silhouette method which computes the average silhouette of observations for different values of k since it gave the exact optimal number, this suggests an optimal of 3 clusters.

**K means happiness score VS migration rate**




The kmeans clusters of Happiness scores vs percent migration:

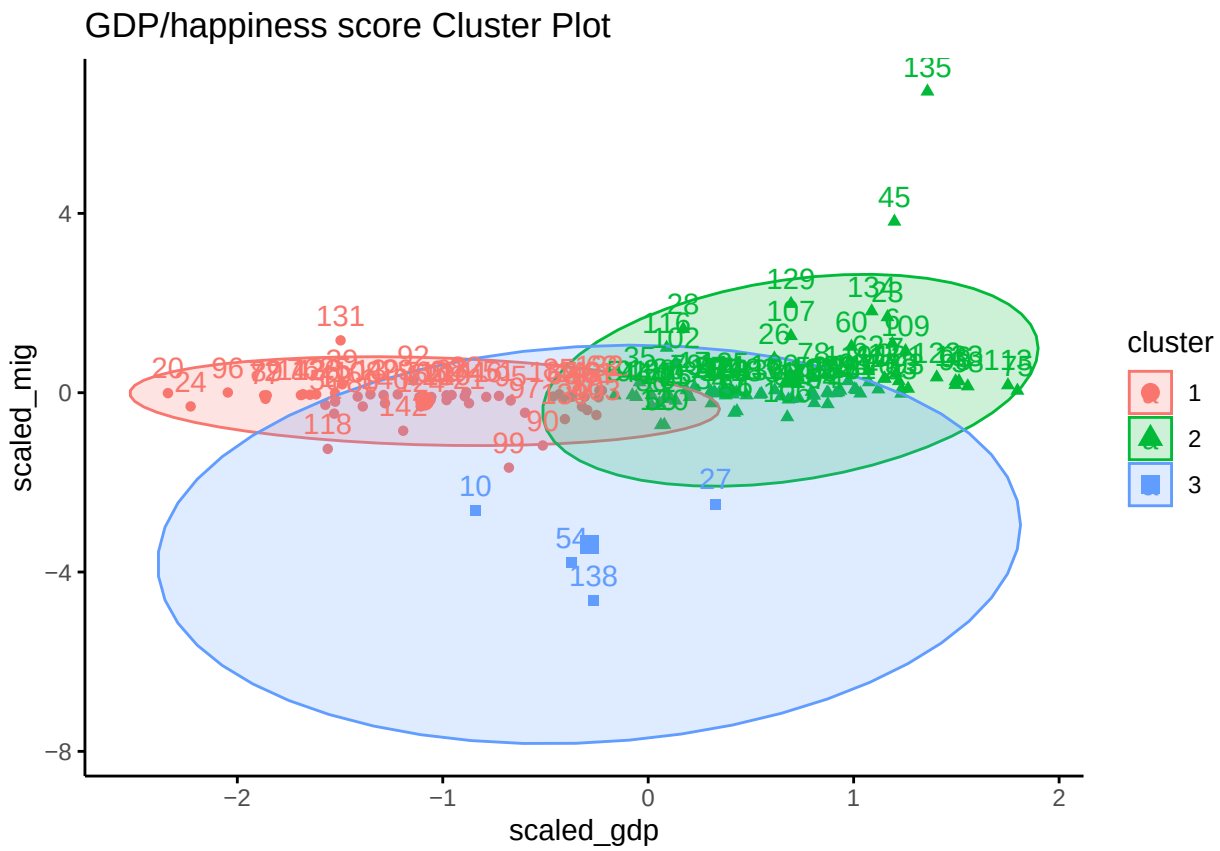
1. The first group is of the countries with happiness scores greater than the mean happiness scores. We can see that the proportion of people that migrated to these countries is positive. This seems reasonable as people want to move to places where they can be happy. There are a handful of exceptions that have negative migration.

2. The second group consists of countries with happiness scores less than average with few countries having a happiness slighlty greater than the average. This group is the most popular group and there is no definate migration pattern except that slightly more countries appear to have an postive migration.

3. The third group has similar range of happiness scores to the second group, but these countries have extremely negative migration compared to other countries.

What we can tell from these groups is that people are likely to move to countries with higher than average happy scores probably because of factors that happiness score accounts for like less corruption and better life expentancy. We need more information than just happiness score to predict migration of country which has less than average happyness score or a happiness slightly over the average. For future analysis we can look at education rates relating to more skilled workers to export, proximity to other countries which would make migration easy and cheap (so low income groups can afford it) or strictness of immigration laws.

**K means GDP vs happiness score**




The kmeans clusters of Happiness scores vs percent migration:

1. The first group is of countries ranges approximately from -0.2 to -0.25 standard deviations of the mean per capita GDP. Most countries in this group have negative migration, but there are still quite a few countries with positive migration. Almost all the countries in this group are within one standard deviation of the mean net migration.

2. the second group consists of countries with more than average per capita GDP. A good majority of these countries have a positive migration with some countries going beyond 1 standard deviations.

3. The third group consists of countries near the average per capita GDP. These countries have migration rates beyond -2 standard deviations.

So, countries that have per capita GDP close to the average will have more people leaving to countries close to +1 standard deviation. We need to look at more factors to understand why this happens. An educated guess would be to assume countries near the average per capita GDP have good enough infrastructure to export their educated population to obtain highskill jobs in countries with higher per capita GDP.

```
growth2020%>%
  group_by(mig_activity)%>%
  summarise("mean_change" = mean(gdp_change))
```

```
 # `summarise()` ungrouping output (override with `.groups` argument)
```

```
# # A tibble: 2 x 2
#   mig_activity mean_change
#   <chr>              <dbl>
# 1 Active           -0.0459
# 2 Passive          -0.0397
```

It is logical to beleive that the more people are going to travel, the more covid is going to spread. Having looked at the the scatter plots of net migration we assumed countries with more than a net number 15000 people leaving or entering to be active countries. We calculated the mean GDP drop of these countries to be -0.0459 while that of passive countries to be -0.0397. We use the loss of per capita GDP for how hard a country is hit my COVID. From our results we conclude that countries with more migration are harder hit. We expected this difference to be much greater. Perhaps there are factors buffering this loss of GDP like crisis resistant economies for instance an economy with more factory jobs is likely to be hit harder than an economy with technical jobs that are easier to do remotely (Usually technical economies import and export more workers). For future analysis we would incorporate factors such as type of economy and health care.

Conclusion

From our graphical analysis we were able to infer that GDP and happiness score has a strong correlation which was proved by the positive slope of the linear regression. This was further confirmed by the high correlation coefficient between these two factors indicating that the GDP has the greatest affect on a country's happiness score followed by health life expectancy. By looking at this analysis countries can decipher what makes it's citizens happy and prioritize the aspects that are required to be worked upon in order to make its citizens happiness and increase the happiness score. If happiness (happiness score), high opportunities and development (GDP) are so important to people, They will go out in search for it to other countries if they are not able to find. Our scatterplots and kmeans clusters show that people tend to migrate from countries with less than average happiness scores and GDP to countries with greater than average GDP and happiness scores. The general trend is that countries 2 standard deviations to the left of mean happiness scores lose a relatively large proportion of their population. As we come near the mean the there is no general trend, but 2 standard deviations to the right countries gain a large proportion of their population as migrants. We got curious as to how migration may effect the spread of COVID and found that the mean GDP drop of these countries to be -0.0459 while that of passive countries to be -0.0397. This shows countries with busier immigration and emigration where greater hit by COVID. However, we need more information than just happiness score to predict migration of country which has less than average happiness score or a happiness slightly over the average. For future analysis we can look at education rates relating to more skilled workers to export, proximity to other countries which would make migration easy and cheap (so low income groups can afford it) or strictness of immigration laws. Additionally, in order for any country to grow, the happiness of the citizens is required. Ultimately, we are able to attain the true pursuit of happiness, which we as human beings aim for.