STA 141B Statistics Data Technologies - Fall 2022

Dr. Peter Kramlinger

Final Project

Topic: Global Happiness Level before & after Covid Outbreak

Group 23

Shutong Gu - 917109366

Qinyi Qiu - 917405772

Sihe Long - 917046076

Weihan Zheng - 918839473

Yuqi Shi - 917615449

December 07, 2022

Abstract

We are interested in how the global happiness level changed during the Covid outbreak over the past few years. Firstly, we extract each year's happiness scores between 2015 to 2022 according to the dataset from the World Happiness Report. We used some visualization like dot plots and map plots to see the changes in a clear pattern. Then, we used a non-parametric test to examine whether there is significant evidence to show the difference of the Happiness score among different regions. On the other hand, we are interested in which factor impacts the happiness score most. For this question, we first use a correlation matrix to see the relationship between happiness and other variables. Then, we also use multiple linear regression and the method of increment of R square to see which variable has the largest effect on happiness.

Background

Covid started in January 2020. People were quite panicked at the beginning of the pandemic because nobody knows how to prevent it effectively and how harmfully this virus would affect human health. Most of the businesses in every country started to shut down gradually. People started to get stuck in their homes. There are a huge proportion of people in the world who are lacking in food and materials. People's happiness levels certainly will be affected. We are trying to see how the happiness level would change because of Covid-19. We believe the change could go either way. For example, people may find more freedom to work from home and get subsidies from the government. These will increase happiness levels. Or, people are frustrated and devastated because they can not go out and travel. In this project, we gathered information from over 100 countries in an 8-year period and tried to see the differences in each country over multiple years.

Dataset description

Different years' data contain similar information with minor differences.

Using 2015 data as an example:

There are 13 columns and 133 rows which contain 132 unique values. The first column is the country name. Countries are ranked from showing the highest happiness score to showing the lowest happiness score. The relevant variables and values we are focusing on are countries and happiness scores.

Here are all the variables:

Country: Name of the country.

Happiness_score: Average of responses to the primary life evaluation question from the Gallup World Poll (GWP). 0-10

Gdp_per_capita: The extent to which GDP contributes to the calculation of the Happiness Score.

Family: The extent to which Family contributes to the calculation of the Happiness Score

Health: The extent to which Life expectancy contributed to the calculation of the Happiness Score

Freedom: The extent to which Freedom contributed to the calculation of the Happiness Score.

Generosity: A numerical value calculated based on poll participants' perceptions of generosity in their country.

Government_trust: The extent to which Perception of Corruption contributes to Happiness Score.

Dystopia_residual: A score based on a hypothetical comparison to the world's saddest country.

Continent: Region of the country.

Topic I: How did the global Happiness level change after the covid outbreak at the beginning of 2020?

Introduction

Due to the outbreak of covid at the beginning of 2020, many countries have experienced lockdowns and quarantines. The distribution led by the virus may strongly influence the daily life of people from different countries. Knowing the effect of the virus outbreak on people's happiness levels may help us to reveal the influence of covid-19 from a completely new point of view.

Data cleaning

Since our datasets, World Happiness Report, range from 2015 to 2022, we specifically pick out two columns from each year's dataset to form new data frames, like "data2015_score". Two columns are "Country" and "Happiness Score". Then these data frames are merged based on the "Country" names. The newly formed data frame is named "merge_score". After these processes, the "Region" variable from a single-year dataset (2015data is used here) is merged into the "merge_score". And the column's names for scores are reassigned to fit the year they are from. After some data cleaning and data type transformation, the final data frame looked like this:

```
merge_score
```

	Country	2015	2016	2017	2018	2019	2020	2021	2022	Region
0	Switzerland	7.587	7.509	7.494	7.487	7.480	7.5599	7.571	7.512	Western Europe
1	Iceland	7.561	7.501	7.504	7.495	7.494	7.5045	7.554	7.557	Western Europe
2	Denmark	7.527	7.526	7.522	7.555	7.600	7.6456	7.620	7.636	Western Europe
3	Norway	7.522	7.498	7.537	7.594	7.554	7.4880	7.392	7.365	Western Europe
4	Canada	7.427	7.404	7.316	7.328	7.278	7.2321	7.103	7.025	North America
...	...	...	...	...	...	...	...	...	...	...
112	Ivory Coast	3.655	3.916	4.180	4.671	4.944	5.2333	5.306	5.235	Sub-Saharan Africa
113	Burkina Faso	3.587	3.739	4.032	4.424	4.587	4.7687	4.834	4.670	Sub-Saharan Africa
114	Afghanistan	3.575	3.360	3.794	3.632	3.203	2.5669	2.523	2.404	Southern Asia
115	Benin	3.340	3.484	3.657	4.141	4.883	5.2160	5.045	4.623	Sub-Saharan Africa
116	Togo	2.839	3.303	3.495	3.999	4.085	4.1872	4.107	4.112	Sub-Saharan Africa

117 rows × 10 columns

We will then do an analysis and data description using this data frame.

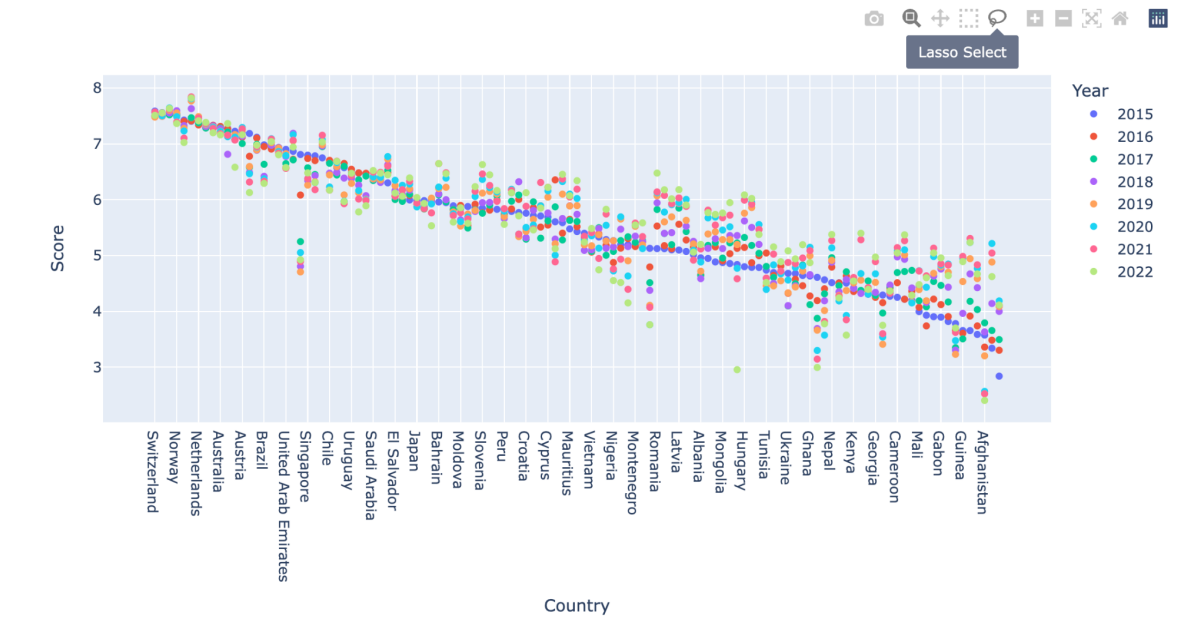
Data Visualization

To have a better understanding of how the world happiness score changed over the years, we want to build a plot that can directly demonstrate the variation of the scores for every country over the years. A dot plot is used here. We choose to build the plot using a package called "plotly". This package requires the data frame to be in a certain format. Then the "merge_score" data frame is converted into the following format using the "stack" function:

	Country	Year	Score	Region
0	Switzerland	2015	7.587	Western Europe
1	Switzerland	2016	7.509	Western Europe
2	Switzerland	2017	7.494	Western Europe
3	Switzerland	2018	7.487	Western Europe
4	Switzerland	2019	7.48	Western Europe
...	...	...	...	...
931	Togo	2018	3.999	Sub-Saharan Africa
932	Togo	2019	4.085	Sub-Saharan Africa
933	Togo	2020	4.1872	Sub-Saharan Africa
934	Togo	2021	4.107	Sub-Saharan Africa
935	Togo	2022	4,112	Sub-Saharan Africa

This new data frame is called "stack_score_df ". Then the dot-plot is built with x = "Country", y = "Score", and color = "Year":

```
import plotly.express as px
df = stack_score_df
fig = px.scatter(df, x="Country", y="Score", color="Year")
fig.show()
```



Through the dot plot above, we find it is still hard to directly observe how the score changes over the year for a certain country. So we decided to build a Map plot. The “plotly” package is still used here, and the “3 letters country codes” are merged into the “stack_score_df” for each country using another package called “country_converter”. And the updated data frame that will be used to draw a Map looked like this:

```
for i in range(936):
    stack_score_df.iloc[i,4] = cc.convert(stack_score_df.iloc[i,0], to='ISO3')
```

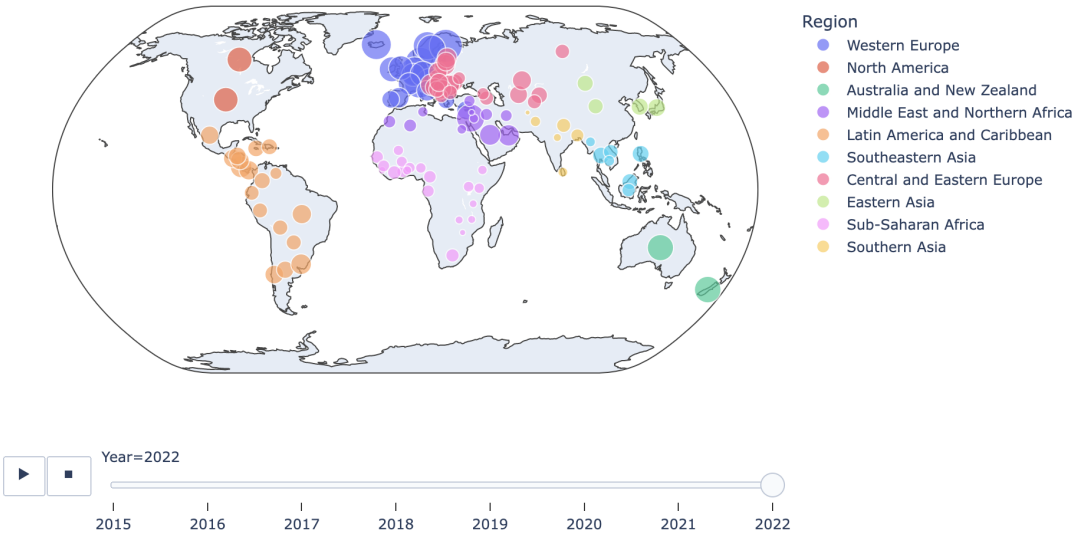
```
stack_score_df
```

	Country	Year	Score	Region	ISO3
0	Switzerland	2015	7.5870	Western Europe	CHE
1	Switzerland	2016	7.5090	Western Europe	CHE
2	Switzerland	2017	7.4940	Western Europe	CHE
3	Switzerland	2018	7.4870	Western Europe	CHE
4	Switzerland	2019	7.4800	Western Europe	CHE
...	...	...	...	...	...
931	Togo	2018	3.9990	Sub-Saharan Africa	TGO
932	Togo	2019	4.0850	Sub-Saharan Africa	TGO
933	Togo	2020	4.1872	Sub-Saharan Africa	TGO
934	Togo	2021	4.1070	Sub-Saharan Africa	TGO
935	Togo	2022	4.1120	Sub-Saharan Africa	TGO

Then an interactive Map plot is built:

```
# Map:

import plotly.express as px
df = stack_score_df.copy()
df['Score'] = (2**df['Score']) #Square the score to show difference
fig = px.scatter_geo(df, locations="ISO3",
                     hover_name="Country", size="Score",color='Region',
                     animation_frame="Year",
                     projection="natural earth")
fig.show()
```



Data Analysis

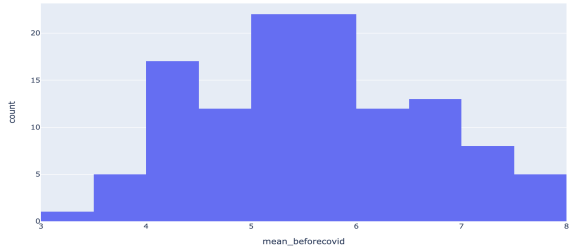
Since we are interested in whether there is a significant difference in the World Happiness Score before and after Covid Outbreak, we decide to first separate our data into two periods based on the time period. We choose to group 2015-2019 as the data before the covid outbreak, and group 2020-2022 as the data after the covid outbreak. We then calculate the mean Happiness level for each country in these two periods. And the difference between the two periods for each country is calculated. An updated data frame called “pt” looked like this:

```
pt = merge_score.copy()
pt['mean_beforecovid'] = pt.iloc[:,1:6].mean(axis=1)
pt['mean_duringcovid'] = pt.iloc[:,6:9].mean(axis=1)
pt['difference'] = pt.iloc[:,11] - pt.iloc[:,10]
pt
```

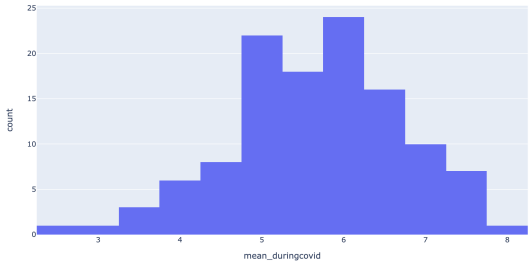
	Country	2015	2016	2017	2018	2019	2020	2021	2022	Region	mean_beforecovid	mean_duringcovid	difference
0	Switzerland	7.587	7.509	7.494	7.487	7.480	7.5599	7.571	7.512	Western Europe	7.5114	7.547633	0.036233
1	Iceland	7.561	7.501	7.504	7.495	7.494	7.5045	7.554	7.557	Western Europe	7.5110	7.538500	0.027500
2	Denmark	7.527	7.526	7.522	7.555	7.600	7.6456	7.620	7.636	Western Europe	7.5460	7.633867	0.087867
3	Norway	7.522	7.498	7.537	7.594	7.554	7.4880	7.392	7.365	Western Europe	7.5410	7.415000	-0.126000
4	Canada	7.427	7.404	7.316	7.328	7.278	7.2321	7.103	7.025	North America	7.3506	7.120033	-0.230567
...	...	...	...	...	...	...	...	...	...	...	...	...	...
112	Ivory Coast	3.655	3.916	4.180	4.671	4.944	5.2333	5.306	5.235	Sub-Saharan Africa	4.2732	5.258100	0.984900
113	Burkina Faso	3.587	3.739	4.032	4.424	4.587	4.7687	4.834	4.670	Sub-Saharan Africa	4.0738	4.757567	0.683767
114	Afghanistan	3.575	3.360	3.794	3.632	3.203	2.5669	2.523	2.404	Southern Asia	3.5128	2.497967	-1.014833
115	Benin	3.340	3.484	3.657	4.141	4.883	5.2160	5.045	4.623	Sub-Saharan Africa	3.9010	4.961333	1.060333
116	Togo	2.839	3.303	3.495	3.999	4.085	4.1872	4.107	4.112	Sub-Saharan Africa	3.5442	4.135400	0.591200

117 rows × 13 columns

Then we check the assumption for a paired t-test: Normality. We use histograms to draw the distribution of “mean_beforecovid” and “mean_duringcovid”, and a Shapiro test is used to confirm the normality of both variables.



(distribution for variable mean_beforecovid)



(distribution for variable mean_duringcovid)

Two variables can be considered as normal distribution and the normality assumption for the paired t-test is met. Then a paired t-test is constructed:

```
from scipy import stats
```

```
stats.ttest_rel(pt['mean_duringcovid'],pt['mean_beforecovid'])
```

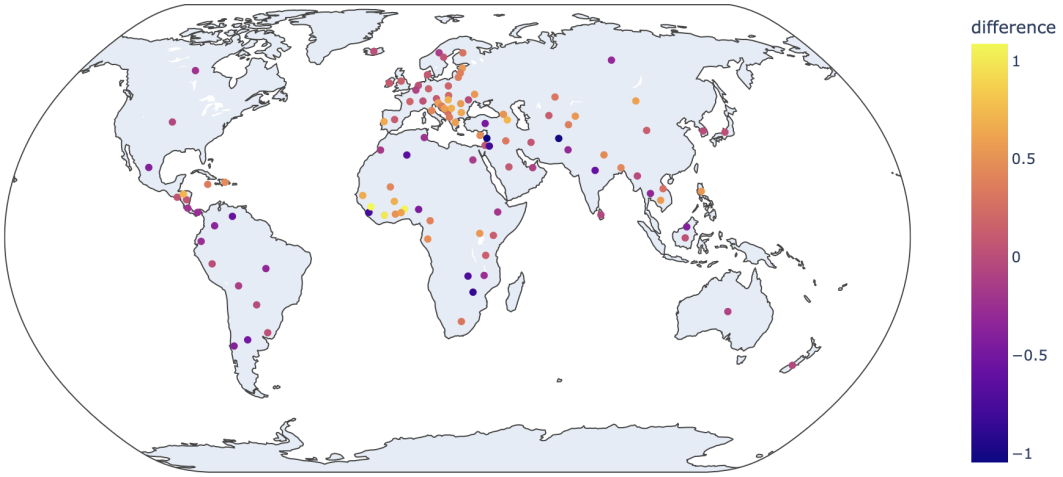
```
Ttest_relResult(statistic=2.9449217964942904, pvalue=0.0039040403356459313)
```

The p-value for this test is 0.0039 which indicates that there is a significant difference in the world happiness level before and after the covid outbreak. However, the test also shows that the test statistic is positive which means that the world happiness level actually increased after the covid outbreak in general. It is a result out of our expectation since we expect the world happiness level to decrease after the covid outbreak in general. We then try to draw another map that indicates the happiness level difference for each country.

```
pt["ISO3"] = cc.convert(pt.iloc[:,0], to='ISO3')
```

hide_code

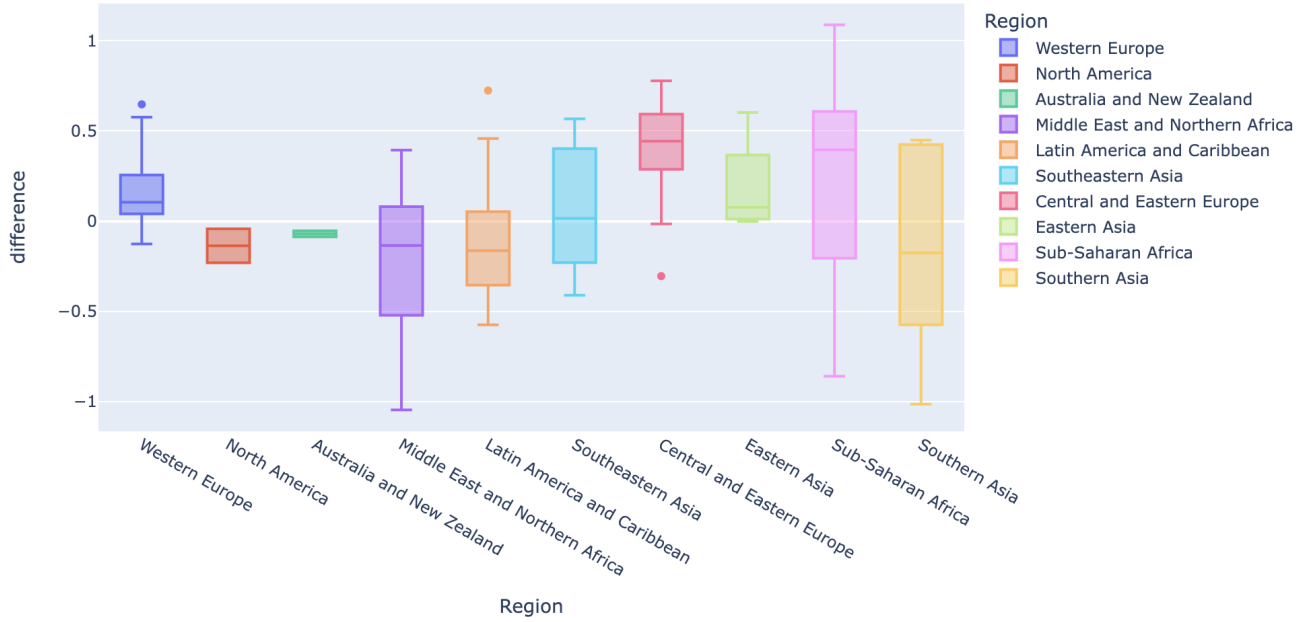
```
import plotly.express as px
df = pt.copy()
fig = px.scatter_geo(df, locations="ISO3",
                     hover_name="Country", color="difference",
                     projection="natural earth")
fig.show()
```



The Map shows that there is a clear variation in the happiness level difference among various regions. The European countries seem to have a positive reaction toward the Covid outbreak since in general, their happiness score increased. And the Southern American countries seem to have a negative reaction toward the Covid outbreak since in general, their happiness score dropped.

Based on this fact, we try to understand whether there is significant evidence to show that different regions' happiness levels changed differently before and after the covid outbreak.

Then a multiple samples comparison test should be used. We first consider using the ANOVA test to tell whether the differences among regions are significant. However, we realize that the inner group variance for different regions is not the same by observing the following boxplot:



So we then choose to use a non-parametric Kruskal Wallis test. To prepare the dataset for the test, we abstract the "difference" variable based on the regions and store them in the corresponding variables. After preparing the test data for a Kruskal Wallis test, the test is set:

```
#Set Up Kruska Wallis test
stats.kruskal(WesternEU,NorthAmerica,Australia_NZ,MiddleEast,Latin,SoutheasternEU,Cen
```

```
KruskalResult(statistic=33.974043753619014, pvalue=9.026275173896719e-05)
```

The p-value indicates that there is a significant difference among the regions.

Topic II: Which factor impacts the happiness score most?

Introduction

In this question, we are interested in the happiness score in 2022. We want to learn about which factor has the biggest contribution to the happiness score. In this question, we may understand what will bring the most happiness to people and if we want to improve people's happiness, what we should do.

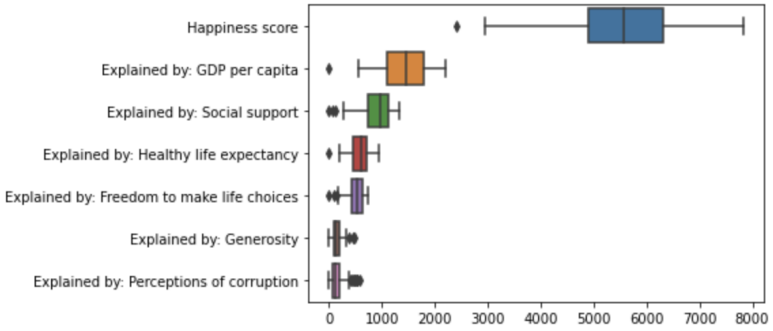
Data Cleaning

We applied the World Happiness Report (2022) from Kaggle to this question. The dataset includes a total of twelve variables. However, there are some variables that are irrelevant to our question. Hence, we first exclude those irrelevant variables. Then we unify units and formats of all variables and exclude missing values. So in this question, we include eight variables: country, happiness score, GDP per capita, social support, healthy life expectancy, freedom, generosity, and corruption.

	Country	Happiness score	Explained by: GDP per capita	Explained by: Social support	Explained by: Healthy life expectancy	Explained by: Freedom to make life choices	Explained by: Generosity	Explained by: Perceptions of corruption
0	Finland	7821	1892	1258	775	736	109	534
1	Denmark	7636	1953	1243	777	719	188	532
2	Iceland	7557	1936	1320	803	718	270	191
3	Switzerland	7512	2026	1226	822	677	147	461
4	Netherlands	7415	1945	1206	787	651	271	419
...	...	...	...	...	...	...	...	...
141	Botswana*	3471	1503	815	280	571	12	102
142	Rwanda*	3268	785	133	462	621	187	544
143	Zimbabwe	2995	947	690	270	329	106	105
144	Lebanon	2955	1392	498	631	103	82	34
145	Afghanistan	2404	758	0	289	0	89	5

Data Description

In our dataset, we totally have 146 countries and 1168 data. We plot a boxplot for each variable to see their distribution. We can see from the plot that the happiness score has the largest value. The range of the happiness score is between 3000 to 8000, and the average happiness score across all countries is about 5500. In the box plot, we can see that the range of the perceptions of corruption and generosity is small and the value of these two variables is also smaller than other variables.



Analysis

To investigate which variables have the biggest impact on happiness scores, we use three methods to achieve our result. First, we try to explore the relationship between each variable and happiness score using a correlation matrix. Second, we fit a multiple linear regression, investigating the p-value and coefficient from the model. In addition, we look at the increment of R square in the multiple linear regression.

Correlation Matrix

The correlation matrix is a table that shows the correlation coefficient between each variable. In this matrix, we can explore the correlation between each variable and happiness score. A high correlation coefficient means that the relationship between these two variables is strong.

	Happiness score	Explained by: GDP per capita	Explained by: Social support	Explained by: Healthy life expectancy	Explained by: Freedom to make life choices	Explained by: Generosity	Explained by: Perceptions of corruption
Happiness score	1.000000	0.763677	0.777889	0.740260	0.624822	0.063785	0.416216
Explained by: GDP per capita	0.763677	1.000000	0.722421	0.815386	0.458591	-0.164472	0.377589
Explained by: Social support	0.777889	0.722421	1.000000	0.666760	0.480466	-0.002339	0.223352
Explained by: Healthy life expectancy	0.740260	0.815386	0.666760	1.000000	0.433166	-0.098133	0.362626
Explained by: Freedom to make life choices	0.624822	0.458591	0.480466	0.433166	1.000000	0.176800	0.402474
Explained by: Generosity	0.063785	-0.164472	-0.002339	-0.098133	0.176800	1.000000	0.096107
Explained by: Perceptions of corruption	0.416216	0.377589	0.223352	0.362626	0.402474	0.096107	1.000000

The correlation matrix above is separated by red cells, the diagonal of the matrix, which is the correlation coefficient of the variables themselves. Each cell represents the correlation between the variable on the top row and the first left column. For example, the cell on the top left corner, the intersection of the first row and first column, represents the correlation between the happiness score itself, so the value in the cell is one. Since we are interested in the relationship between each variable and happiness score, we can only focus on the first column. We can see that there are few variables that have a relatively large correlation with happiness score, GDP per capita (0.764), social support (0.778), and healthy life expectancy (0.740). From this matrix, we can predict that GDP per capita, social support, and healthy life expectancy are variables that are important in happiness scores. And social support is the one that has the biggest impact.

Multiple Linear Regression

We try to fit all variables into a multiple linear regression and check whether they are significant to the happiness score by looking at the p-value of each of them. In addition, we look at the coefficient of the multiple linear regression and see which variable will lead to the biggest changes in happiness score for a unit change.

$$Y_{happiness} = \beta_0 + \beta_1 * X_{GDP} + \beta_2 * X_{Social\ support} + \beta_3 * X_{Life} + \beta_4 * X_{Freedom} + \beta_5 * X_{Generosity} + \beta_6 * X_{Corruption}$$

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Happiness score   R-squared:                       0.774
Model:                            OLS   Adj. R-squared:                  0.764
Method:                 Least Squares   F-statistic:                     79.20
Date:                Wed, 30 Nov 2022   Prob (F-statistic):           2.16e-42
Time:                        15:21:50   Log-Likelihood:                -1118.9
No. Observations:                 146   AIC:                             2252.
Df Residuals:                     139   BIC:                             2273.
Df Model:                           6
Covariance Type:            nonrobust
=========================================================================================================
                                            coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
const                                  1667.7124    205.467      8.117      0.000    1261.469    2073.956
Explained by: GDP per capita              0.5513      0.207      2.661      0.009       0.142       0.961
Explained by: Social support              1.4094      0.242      5.815      0.000       0.930       1.889
Explained by: Healthy life expectancy     1.2735      0.441      2.885      0.005       0.401       2.146
Explained by: Freedom to make life choices  1.6032      0.375      4.270      0.000       0.861       2.345
Explained by: Generosity                  0.9689      0.568      1.705      0.090      -0.154       2.092
Explained by: Perceptions of corruption   0.7305      0.396      1.843      0.068      -0.053       1.514
==============================================================================
Omnibus:                        9.121   Durbin-Watson:                   1.517
Prob(Omnibus):                  0.010   Jarque-Bera (JB):                9.350
Skew:                          -0.617   Prob(JB):                      0.00932
Kurtosis:                       3.113   Cond. No.                     9.06e+03
==============================================================================
```

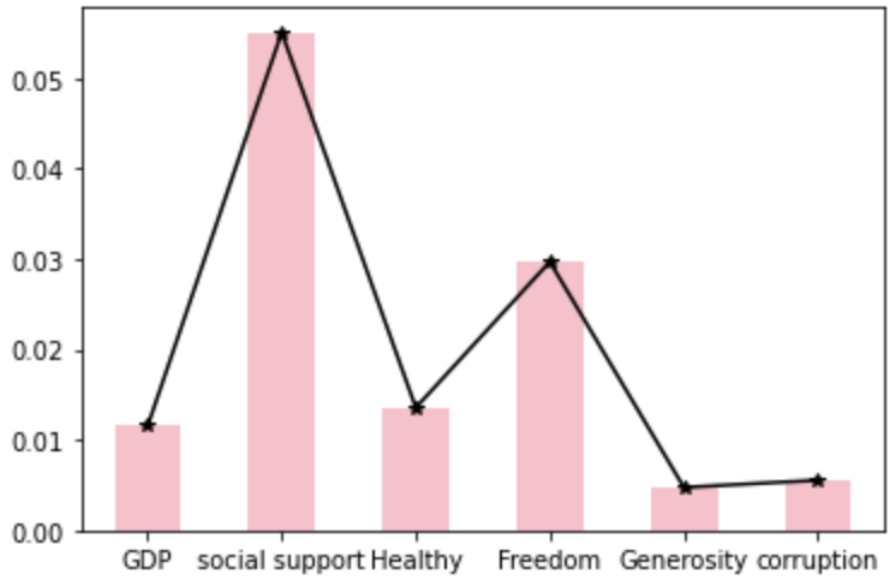
According to the result we get from the multiple regression, we can see most of the p-value for variables are very small. If we select α=0.05, the p-value of all variables except generosity and perceptions of corruption are smaller than 0.05. We can conclude that all variables except generosity and perceptions of corruption are significant. Then we check the coefficient of each variable and it shows that freedom to make life choices has the largest coefficient 1.6032. That means, compared to the other variables, the happiness score will have the largest changes as the freedom to make life choices change for a unit when the other variable stays the same. According to the coefficient of the multiple linear regression, we get a different result from the correlation matrix. However, the coefficient may be affected by the magnitude and unit of the variable, it may not be an efficient way to evaluate the importance of a variable.

Increment of R square

Increment of R square indicates that the variation of a dependent variable is explained by the independent variables in the regression model. According to our multiple linear regression model above, we can say our total six independent variables can explain about 77.4% of the happiness score.

So our basic idea here is to investigate the changes in the R-square to see which variable has the biggest contribution to explaining our dependent variable, the happiness score. For each time, we exclude one of the independent variables from the full model and compare the R square with the new model, seeing how many percentages decrease for the explanation of the happiness score.

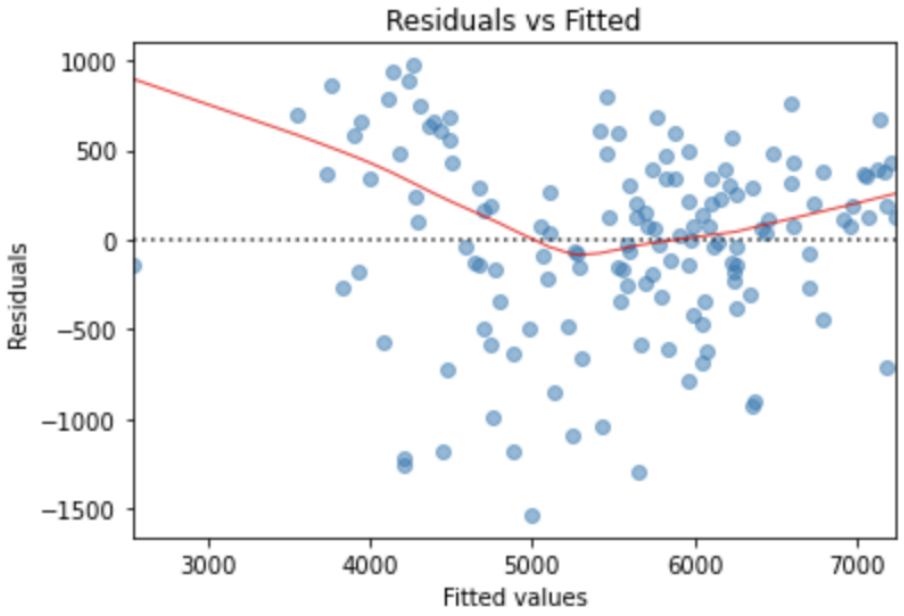
	R-increament
GDP	0.011527
social support	0.055057
Healthy	0.013553
Freedom	0.029693
Generosity	0.004735
corruption	0.005528



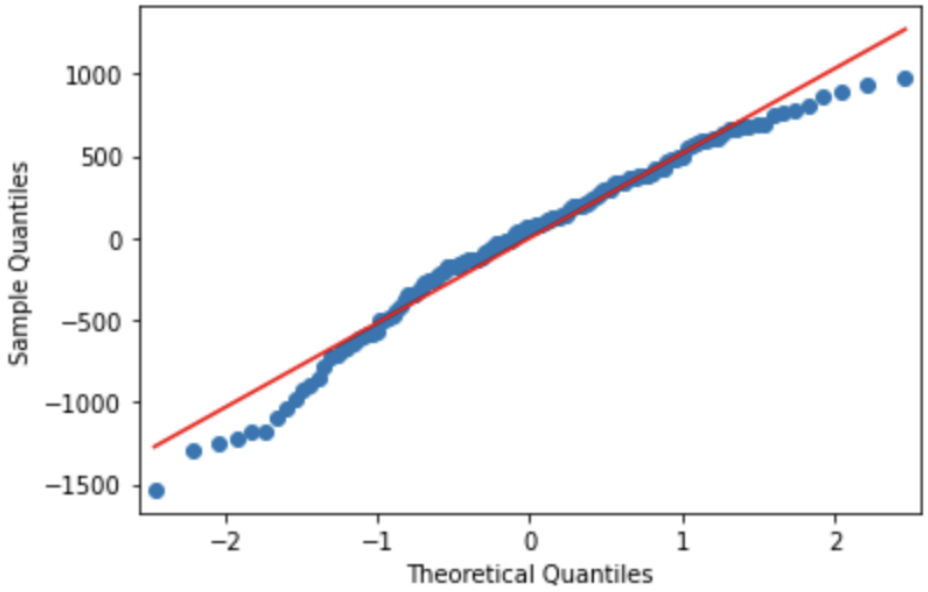
According to the result we get, we find that the R-square has the biggest decrease when we exclude social support, 0.055. That means, when we exclude social support in our model, our model explains about 5.5% less than the full model of the happiness score. We can conclude from the changes in R-square that social support has the biggest contribution to happiness score, so it is the most important variable.

Check for Assumptions

To ensure our multi-linear regression is valid in this case, we check its assumption of it. First, we plot the residual and fitted value for our model to see if our residuals are randomly scattered. Then we plot the QQ plot to check whether our data satisfy the normality assumption.



According to this residual plot, we can see our residuals do not have an obvious pattern. And the residuals relatively randomly scatter around the horizontal line in zero. Thus we can say multi-linear regression is an appropriate method for our data.



According to our QQ plot, we can see that most quartile points lie on the theoretical normal line. Hence, we can conclude that our dataset is similar to the normal distribution and it satisfies the normality assumption.

Results

From the correlation matrix, we can see that the happiness score has a strong positive correlation with GDP per capita, social support, and healthy life expectancy, where social support has the strongest correlation. Then, from the multiple linear regression analysis, we

can see that the happiness score is influenced mostly by the variable of freedom to make life choices. At the same time, from the analysis of the increment of R square, we can see that the variable social support has the biggest contribution to the happiness score as well.

Conclusion

In this project, we explored two questions. The first question is how the global happiness level changed during the Covid outbreak over the past few years. From our analysis, we found that there is a significant difference in the world happiness level before and after the covid outbreak. However, different from our expectation, the test statistic also showed that the world happiness level actually increased after the covid outbreak. In this way, we then drew another map and get the result that there is a clear variation in the happiness level difference among various regions. The European countries seem to have a positive reaction toward the Covid outbreak since in general, their happiness score increased. And the Southern American countries seem to have a negative reaction toward the Covid outbreak since in general, their happiness score dropped. Then from the non-parametric Kruskal Wallis test, we confirm the result that there is a significant difference among the regions.

Also, our second question is which factor impacts the global happiness score most. According to our analysis above, we found that GDP per capita, healthy life expectancy, social support, and freedom are some significant influential factors to the global happiness scores. And through the analysis results of different analysis methods, we conclude that the factor of social support is the most influential factor for the global happiness score. In this way, we might conclude that the improvement of social support may bring the most happiness to people.