cp100mj

June 17, 2023

1 PSTAT 100 Course Project

1.1 Madhan Jeganathan, Imran Azmi

1.1.1 Author Contributions

Madhan worked on tidying the data, running and fitting models, and constructing key visualizations. Imran worked on providing the documentational background of the WHR, preparing the questions of interest, and creating other key visualizations.

1.2 Data Description

The World Happiness Report (WHR) is a publication of the Sustainable Development Solutions Network, powered by the Gallup World Poll data. TheWHR reflects a worldwide demand for more attention to happiness and well-being as criteria for government policy and initiative. WHR reviews the state of happiness in the world today by nation and shows how the science of happiness explains personal and national variations in happiness.

The dataset we are analyzing has a list of 165 countries from the years of 2005 to 2022. Along with `Life Ladder`, the measurement of overall happiness in a nation, we are presented with multiple contributing variables. These variables include `Log GDP per capita`, `Social support`, `Healthy life expectancy at birth`, `Generosity`, `Freedom to make life choices`, `Perceptions of corruption`, `Postive affect`, and `Negative affect`, all of which explain themselves. This totals to 9 measured variables per year per country in the dataset if none are missing. Below is a quick look at the first few rows of the raw dataset.

[318]:
```
happy.head(3)
```

[318]:
```
  Country name  year  Life Ladder  Log GDP per capita  Social support
0  Afghanistan  2008        3.724               7.350           0.451  \
1  Afghanistan  2009        4.402               7.509           0.552
2  Afghanistan  2010        4.758               7.614           0.539

   Healthy life expectancy at birth  Freedom to make life choices  Generosity
0                              50.5                         0.718       0.168  \
1                              50.8                         0.679       0.191
2                              51.1                         0.600       0.121

   Perceptions of corruption  Positive affect  Negative affect
0                      0.882            0.414            0.258
```

1	0.850	0.481	0.237
2	0.707	0.517	0.275

1.3 Question of Interest

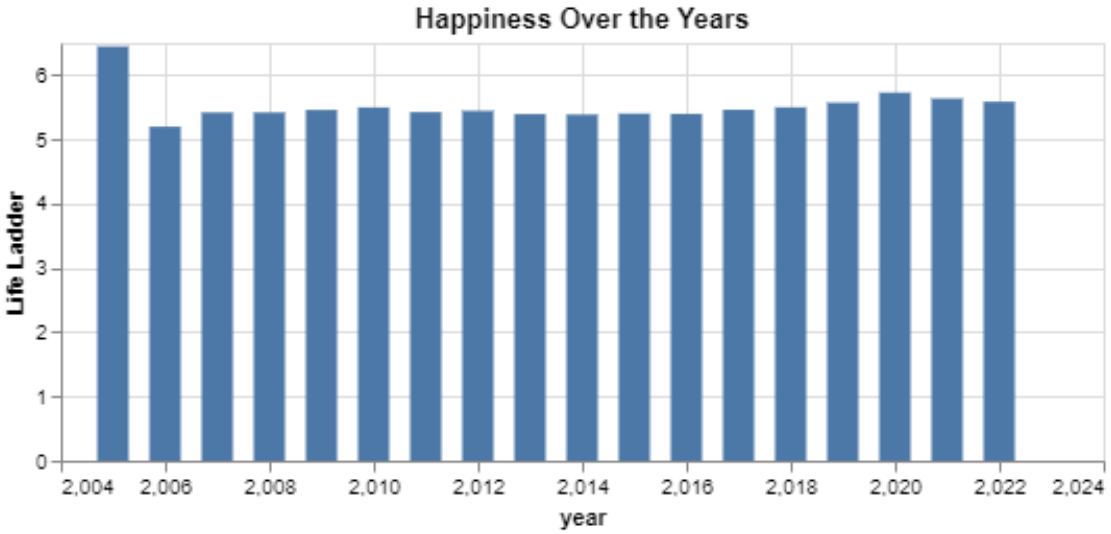
We understand that `Life Ladder` changes annually for countries, with some having little to no change, and others having extreme increases or decreases over time. Our goal is to view the general happiness of the world over time, identify the countries with the largest change in happiness over time, and identify the countries with the highest and lowest average happiness over time. We also want to see which variables affect `Life Ladder` the most positively and negatively, and see if we can relate those variables to the countries with the most change in happiness.

1.4 Data Analysis

Let's start our analysis with a simple plot of happiness over the years.

[313]:
```
year_bar
```

[313]:



From this plot, we can see that from 2006 to 2022, general happiness has been near constant. However, in 2005, it was significantly higher. This may be biased, though, becuase there is much less data from 2005 than all of the other years. Below is a breakdown of the proportion of missing values from each year.

[315]:
```
happy_missing
```

[315]:
```
year
2005    0.836364
2006    0.460606
2007    0.381818
2008    0.333333
```

```
2009    0.309091
2010    0.248485
2011    0.115152
2012    0.145455
2013    0.175758
2014    0.127273
2015    0.139394
2016    0.145455
2017    0.109091
2018    0.145455
2019    0.133333
2020    0.296970
2021    0.260606
2022    0.309091
dtype: float64
```

Because of this, we decided to use 2007 as the starting point for the next few tables (2006 also has a large proportion of missing data).

Below are the countries with the highest increase in `Life Ladder` from 2007 to 2022, followed by the countries with the highest decrease.

[345]:
```
happiest_dir
```

[345]:
```
year Country name   2007   2022  Progress
44        Georgia  3.707  5.293     1.586
16       Bulgaria  3.844  5.378     1.534
91      Nicaragua  4.944  6.392     1.448
69         Latvia  4.667  6.055     1.388
72      Lithuania  5.808  7.038     1.230
37    El Salvador  5.296  6.492     1.196
83       Mongolia  4.609  5.788     1.179
66         Kosovo  5.104  6.160     1.056
103       Romania  5.394  6.437     1.043
38        Estonia  5.332  6.357     1.025
```

It is interesting to see that many of the most improved countries in terms of happiness are in Eastern Europe, a few in Central America, and the rest elsewhere.

[325]:
```
unhappiest_dir.head(10)
```

[325]:
```
year   Country name   2007   2022  Progress
75           Malawi  4.891  3.356    -1.535
36            Egypt  5.541  4.024    -1.517
63           Jordan  5.598  4.356    -1.242
7        Bangladesh  4.607  3.408    -1.199
54            India  5.027  3.930    -1.097
46            Ghana  5.220  4.191    -1.029
```

```
108    Sierra Leone  3.585  2.560    -1.025
97           Panama  6.894  5.979    -0.915
105    Saudi Arabia  7.267  6.382    -0.885
129   United States  7.513  6.693    -0.820
```

It is interesting to see the U.S. in the table with the countries with the highest decrease in happiness. Along with them are India, one of the most populous countries in the world, some countries in Africa and the Middle East, and a few others.

Below are the countries with the highest average happiness from 2007 to 2022, followed by the countries with the lowest.

[348]:
```
happiest
```

[348]:
```
    Country name  Life Ladder
39       Denmark     7.673529
48       Finland     7.619067
111       Norway     7.481750
142  Switzerland     7.474583
63       Iceland     7.458600
105  Netherlands     7.452000
141       Sweden     7.377176
25        Canada     7.323647
106  New Zealand     7.278500
69        Israel     7.265647
```

Many of these countries are in Northern Europe. This is due to the factors of high GDP, low corruption, high social support, and more that we will further discuss later in the analysis.

[349]:
```
unhappiest
```

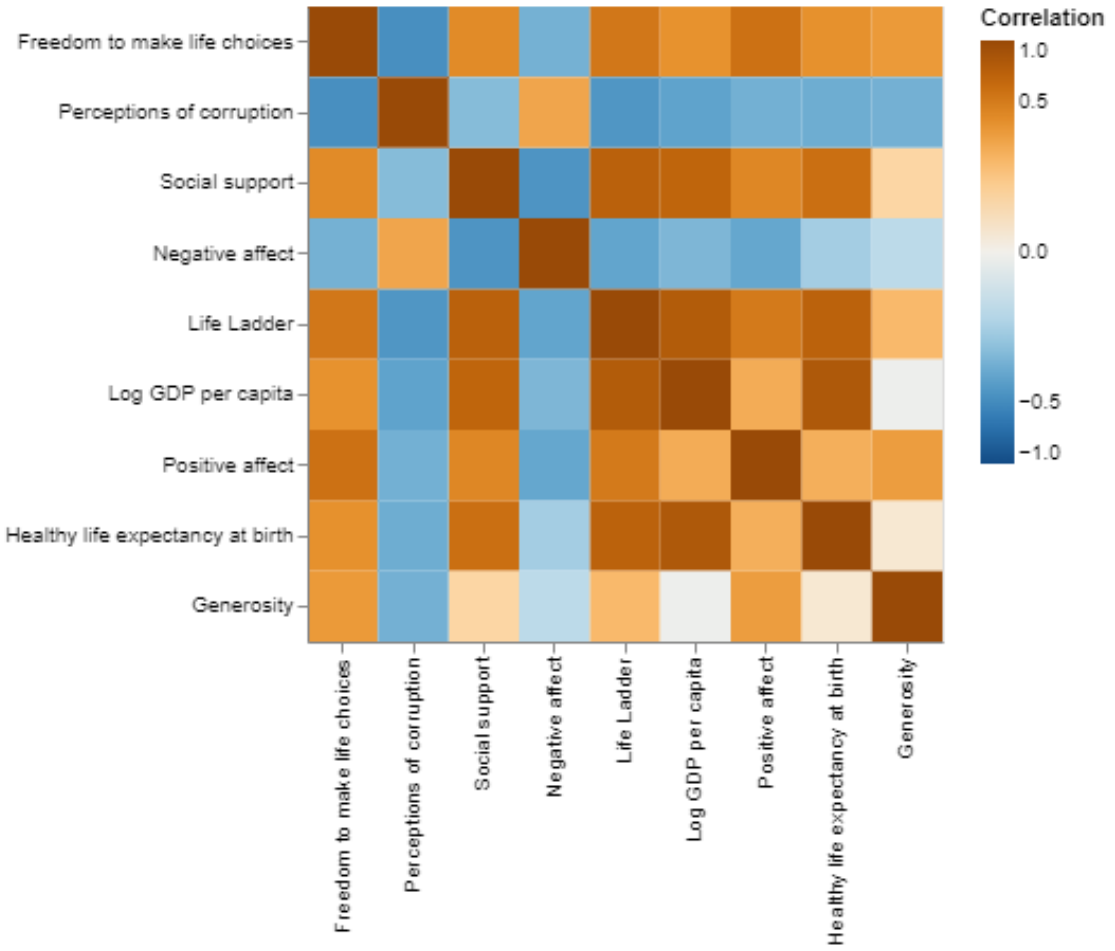
[349]:
```
                 Country name  Life Ladder
0                 Afghanistan     3.346643
135               South Sudan     3.402000
26   Central African Republic     3.515000
22                    Burundi     3.548200
123                    Rwanda     3.654417
148                      Togo     3.661000
146                  Tanzania     3.691588
164                  Zimbabwe     3.805294
31                    Comoros     3.887000
162                     Yemen     3.912250
```

All of these countries are in Africa or the Middle East. We can see that Afghanistan at the top of the list for least happy on average over the years. Afghanistans situation can be a direct result of the withdrawl of U.S troops from the country, leading to the country being under Taliban control.

1.4.1 Correlation

[304]:
```
happy_corr_plot
```

[304]:



Above is a correlation plot with all of the different measured variables. Dark brown indicates high positive correlation and dark blue indicates high negative corelation. Let's take a closer look at `Life Ladder`.

[357]:
```
life_ladder_corr
```

[357]:
```
Perceptions of corruption          -0.431500
Negative affect                    -0.339969
Generosity                          0.181630
Positive affect                     0.518169
Freedom to make life choices        0.534493
Healthy life expectancy at birth    0.713499
Social support                      0.721662
Log GDP per capita                  0.784868
```

```
Life Ladder                          1.000000
Name: Life Ladder, dtype: float64
```

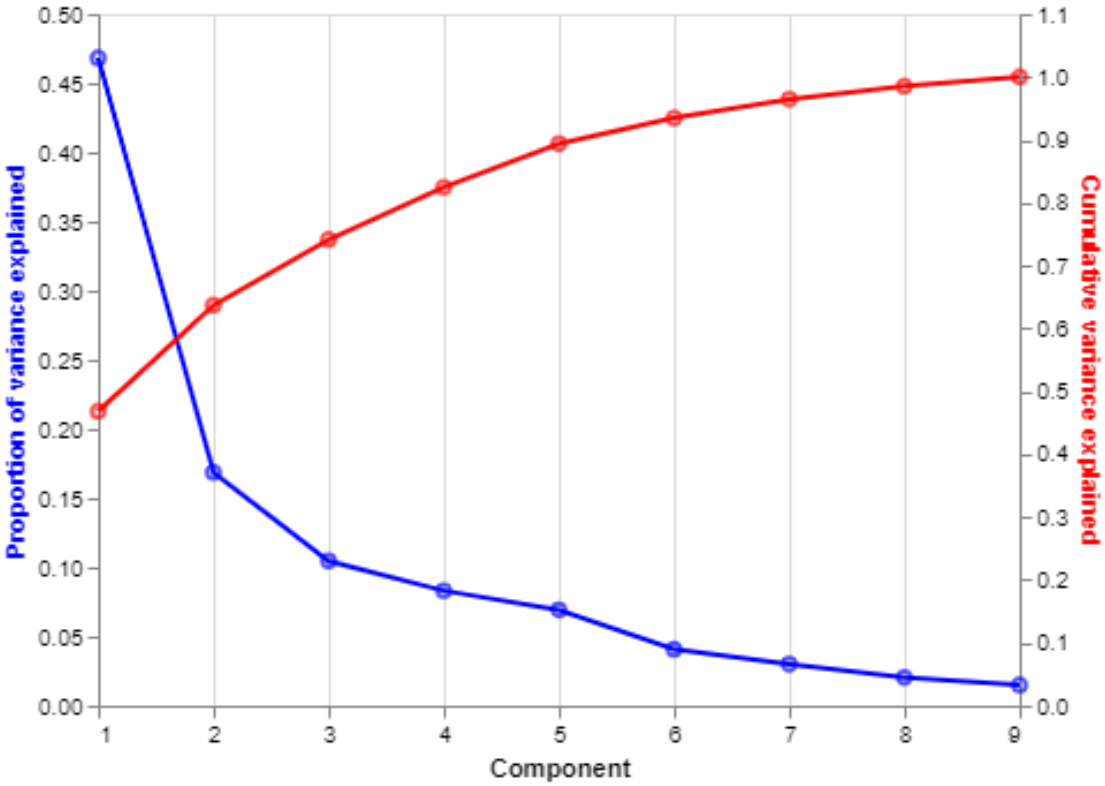
The table above shows the specific correlation values of other variables with `Life Ladder`. We see that `Perceptions of corruption` is the most negatively correlated while `Log GDP per capita` is the most positively corelated, with `Social support` and `Healthy life expectancy at birth` very close behind.

1.4.2 PCA

Now we will run a principal component analysis of the data.

[290]:
```
var_explained_plot
```

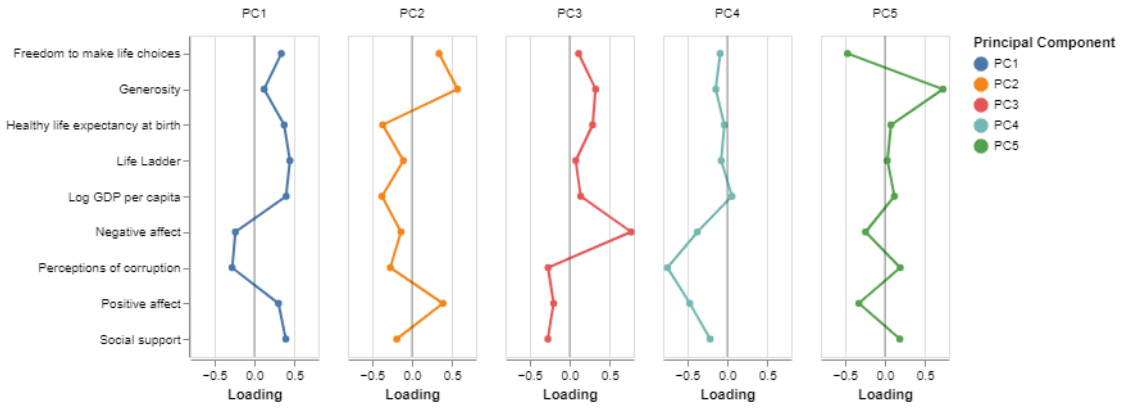
[290]:



The plot above shows the variance explained for each of the principal components (PCs). There are 5 PCs that explain more than 5% of the total variation individually. Together they capture 89.34% of the total variance. Let's take a look at these 5 PCs.

[289]:
```
PC_plot
```

[289]:



The more positive the loading value, the more positively the variable affects the PC value, and the more negative the loading value, the more negatively the variable affects the PC value. For example, PC5 is most positively affected by `Generosity` and most negatively affected by `Freedom to make life choices`. Variables with loadings close to 0 have little to no effect on that PC.

[300]:
```
score_df.head()
```

[300]:

Country	PC1	PC2	PC3	PC4	PC5
Afghanistan	-0.047931	0.007681	0.016692	0.022302	0.021633
Albania	-0.009556	-0.019860	0.008848	0.002836	0.004098
Algeria	-0.008656	-0.028186	-0.003758	0.025352	0.016203
Angola	-0.028231	-0.009611	-0.005549	-0.005519	0.007217
Argentina	0.011662	-0.017272	-0.008648	-0.023098	-0.020437

Above, we can see loading scores for a few countries. Let's take a look at one. Afghanistan's PC1 loading value is about -0.048. Compared to the other values in the table, this is very low. Referring back to the plot above this table, we can see that this must be because in Afghanistan, `Perception of corruption` is high and things like `Social support` and `Log GDP per capita` are low.

1.5 Summary of Findings

Analyzing our findings, we can identify several factors that contribute to the overall happiness levels of a nation. One of the key findings is the negative correlation between perceptions of corruption and happiness. A higher perception of corruption is associated with lower levels of happiness. This suggests that transparency, trust, and integrity in a goverment is crucial for fostering happiness within a society. Countries with lower levels of corruption tend to have higher levels of happiness as citizens feel more secure and have confidence in their institutions. Furthermore, the freedom to make life choices and social support demonstrate positive correlations with happiness. Countries that provide individuals with greater freedom to make life choices tend to have higher levels of happiness. This suggests that the ability to shape one's life contribute to overall well-being and satisfaction. Similarly, social support support programs promote happiness by providing security, meeting basic needs, reducing stress, and creating a sense of fairness. Health-related factors also have a significant impact on happiness. Healthy life expectancy at birth shows a strong positive

correlation with happiness. This implies that countries with longer life expectancy and better overall health tend to have higher levels of happiness. Individuals who enjoy good health are more likely to lead fulfilling lives and experience greater happiness. Lastly, economic prosperity, as measured by log GDP per capita, shows the highest positive correlation with happiness. Countries with higher levels of economic development and wealth tend to have higher levels of happiness. Economic factors play a crucial role in providing individuals with opportunities, resources, and a sense of security, all of which contribute to overall well-being and happiness. This all adds up when looking at the tables of countries with that are the happiest and unhappiest. Many of the unhappiest countries are war torn and corrupt, while many of the happiest countries are the healthiest, freest, and most financially stable. Countries that have improved in happiness have become less corrupt and more free, resulting in being more free and financially stable. Countries that have become less happy have had the opposite happen, becoming more corrupt and in turn becoming less free and financialyl stable.

1.6 Code

[327]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import altair as alt
import seaborn as sns
from scipy import linalg
from statsmodels.multivariate.pca import PCA
alt.renderers.enable('mimetype')
```

[327]:
```
RendererRegistry.enable('mimetype')
```

[316]:
```
happy = pd.read_csv('data/whr-2023.csv')
```

[319]:
```
happy_missing = happy.pivot_table(
    index = 'Country name',
    columns = 'year',
    values = 'Life Ladder'
).isna().mean()
```

[320]:
```
happy_year = happy.drop(columns = 'Country name').groupby('year').mean().
 ↪reset_index()
year_bar = alt.Chart(happy_year).mark_bar(size = 15).encode(
    x = 'year',
    y = 'Life Ladder'
).properties(
    width = 500,
    height = 200,
    title = 'Happiness Over the Years'
)
```

[344]:
```
happy_dir = happy[(happy['year'] == 2007) | (happy['year'] == happy['year'].
 ↪max())]
happy_dir = happy_dir.pivot_table(
    index = 'Country name',
    columns = 'year',
    values = 'Life Ladder'
).reset_index()
happy_dir['Progress'] = happy_dir[happy['year'].max()] - happy_dir[2007]
happiest_dir = happy_dir.sort_values(by="Progress", ascending=False).head(10)
```

[346]:
```
unhappiest_dir = happy_dir.sort_values(by="Progress", ascending=True).head(10)
```

[350]:
```
happiest = happy_agg[['Country name', 'Life Ladder']].sort_values(by='Life␣
 ↪Ladder', ascending=False).head(10)
```

[351]:
```
unhappiest = happy_agg[['Country name', 'Life Ladder']].sort_values(by='Life␣
 ↪Ladder', ascending=True).head(10)
```

[352]:
```
happy_quant = happy.drop(columns = ['Country name', 'year'])
corr_happy = happy_quant.corr()
corr_happy_long = corr_happy.reset_index().rename(
    columns = {'index': 'row'}
).melt(
    id_vars = 'row',
    var_name = 'col',
    value_name = 'Correlation'
)

# construct plot
happy_corr_plot = alt.Chart(corr_happy_long).mark_rect().encode(
    x = alt.X('col', title = '', sort = {'field': 'Correlation', 'order':␣
 ↪'ascending'}),
    y = alt.Y('row', title = '', sort = {'field': 'Correlation', 'order':␣
 ↪'ascending'}),
    color = alt.Color('Correlation',
                      scale = alt.Scale(scheme = 'blueorange',
                                        domain = (-1, 1),
                                        type = 'sqrt'),
                      legend = alt.Legend(tickCount = 5))
).properties(width = 300, height = 300)
```

[353]:
```
life_ladder_corr = corr_happy.loc[:, 'Life Ladder'].sort_values()
```

[354]:
```
pca = PCA(happy_quant.dropna(), standardize = True)
var_ratios = pca.eigenvals/pca.eigenvals.sum()
pca_var_explained = pd.DataFrame({
    'Component': np.arange(1, 10),
```

```
    'Proportion of variance explained': var_ratios})
pca_var_explained['Cumulative variance explained'] = var_ratios.cumsum()
base = alt.Chart(pca_var_explained).encode(
    x = 'Component')
prop_var_base = base.encode(
    y = alt.Y('Proportion of variance explained',
              axis = alt.Axis(titleColor = 'blue'))
)
cum_var_base = base.encode(
    y = alt.Y('Cumulative variance explained', axis = alt.Axis(titleColor =␣
 ↪'red'))
)
prop_var = prop_var_base.mark_line(stroke = 'blue') + prop_var_base.
 ↪mark_point(color = 'blue')
cum_var = cum_var_base.mark_line(stroke = 'red') + cum_var_base.
 ↪mark_point(stroke = 'red')
var_explained_plot = alt.layer(prop_var, cum_var).resolve_scale(y =␣
 ↪'independent')
```

[355]:
```
num_pc = np.sum(pca_var_explained['Proportion of variance explained'] > 0.05)
var_explained = var_ratios[0:num_pc].sum()
loading_df = pca.loadings.iloc[:, 0:num_pc]
loading_df = loading_df.rename(columns = dict(zip(loading_df.columns, ['PC' +␣
 ↪str(i) for i in range(1, num_pc + 1)])))
loading_plot_df = loading_df.reset_index().melt(
    id_vars = 'index',
    var_name = 'Principal Component',
    value_name = 'Loading'
).rename(columns = {'index': 'Variable'})
loading_plot_df['zero'] = np.repeat(0, len(loading_plot_df))
base = alt.Chart(loading_plot_df)
loadings = base.mark_line(point = True).encode(
    y = alt.X('Variable', title = ''),
    x = 'Loading',
    color = 'Principal Component'
)
rule = base.mark_rule().encode(x = alt.X('zero', title = 'Loading'), size = alt.
 ↪value(0.05))
loading_plot = (loadings + rule).properties(height = 250, width = 100)
PC_plot = loading_plot.facet(column = alt.Column('Principal Component', title =␣
 ↪''))
```

[356]:
```
score_df = pca.scores.iloc[:, 0:num_pc]
score_df = score_df.rename(
    columns = dict(zip(score_df.columns, ['PC' + str(i) for i in range(1,␣
 ↪num_pc + 1)])))
```

```
)
score_df['Country'] = happy['Country name']
score_df = score_df.groupby('Country').mean()
```