GITHUB URL:

https://github.com/BrianRiera/UCDPA_BrianRiera

Datasets:
https://www.kaggle.com/ajaypalsinghlo/world-happiness-report-2021
https://www.kaggle.com/tanuprabhu/population-by-country-2020

Abstract

Analysis was carried out on the ‘World happiness 2021 report’ dataset.
A report which measured happiness based on the ‘Cantril ladder’: which asks respondents to think of a ladder, with the best possible life for them being a 10, and the worst possible life being a 0. They are then asked to rate their own current lives on that 0 to 10 scale.

This was measured across 140 countries worldwide, grouping countries by regions. Initial research discovered that Finland scored highest in the Cantril Ladder and that fourteen out of the twenty countries that scored as ‘Thriving’ were from Western Europe, while Afghanistan and Zimbabwe were the only countries to be categorised as ‘Suffering’.
Additional research was conducted to analyze the contributing factors to ladder scores. A Pearson correlation found a large positive correlation with logged GDP per capita, social support, healthy life expectancy and ladder scores.

Introduction

This dataset was chosen, as it focused on the effects of COVID-19 on happiness and how countries have differed in their success in maintaining connected and healthy societies. The effects of the pandemic on happiness,mental health, social connections and the workplace respectively have been widely reported worldwide making this a very ideal and relevant dataset for data analysis.

The ladder scores and rankings use data from the Gallup World Poll . The columns following the happiness score estimate the extent to which each of six factors – economic production, social support, life expectancy, freedom, absence of corruption, and generosity – contribute to making life evaluations higher in each country than they are in Dystopia, a hypothetical country that has values equal to the world’s lowest national averages for each of the six factors.

The World Happiness Report 2021 is a small csv file(149 rows, 20 columns) located on Kaggle, with various columns containing either floats or strings making it an ideal candidate for a data analysis project.
A second csv dataset from Kaggle listing country names and their respective population was used, to add further data to each row from the initial dataset.

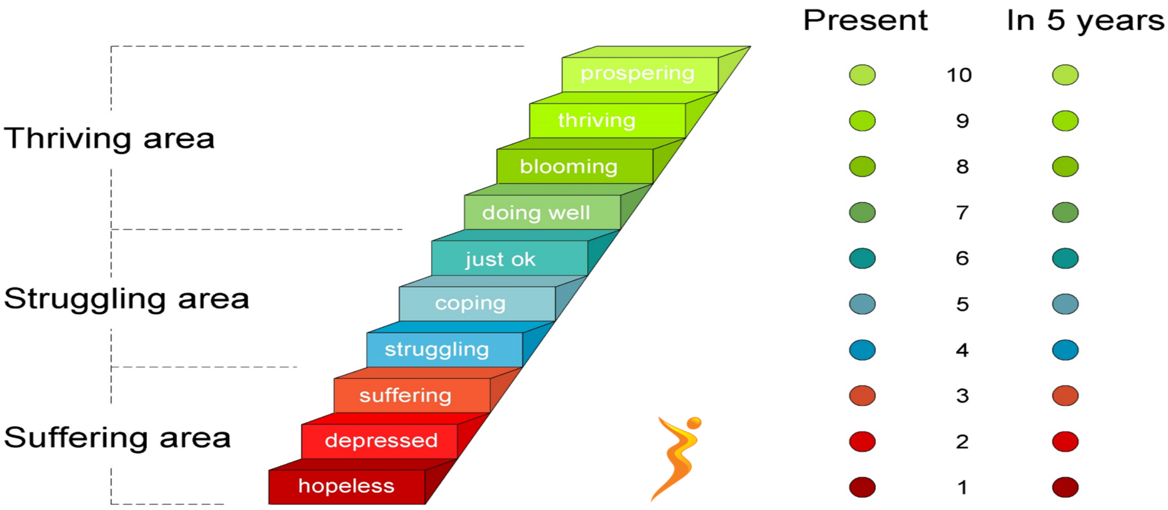
Implementation Process & Results

As both datasets are csv files, pandas was imported and used to load and convert the files into dataframes (df & pop) to enable easier analysis of the data. Following this it was important to perform a preliminary examination on the dataframes (df.head,info,shape etc)

The pop dataframe was sliced and stored as a new dataframe, containing only the country names and subsequent populations of each country, this was then merged onto the df dataframe using an outer left join on the ‘country name’ column from each, this was carried out as it was necessary to keep all columns from the df dataframe.
After merging, the new dataframe was sorted alphabetically by country name, the index was reset; and columns were renamed, removing white space while also abbreviating some of the longer named columns, this was done to enable tidier analysis when using python on the data.

The new data was then checked for missing values which were subsequently dropped. As the data was numeric data from the population column, it was decided to drop rather than forward/backfill, an alternative option would have been to research the populations of the NaN values and fill them in using a list.

After researching the Cantril Ladder see figure 1 below:

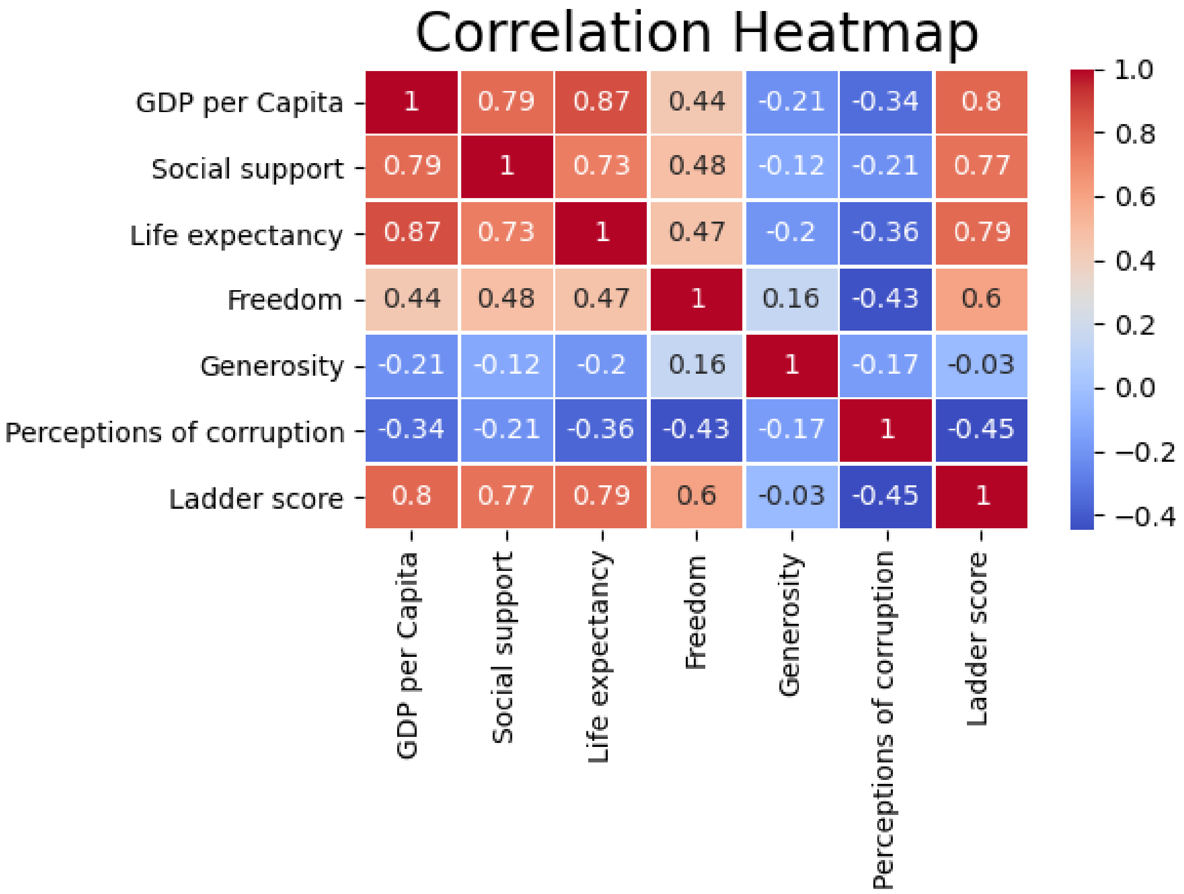


It was decided to add a new column to the dataframe, ‘Ladder_Category’ which would assign countries to their Cantril Ladder group score: Thriving, Struggling, Suffering; based on their ladder score, this was done using a for loop and the iterrows() method was used as it was a dataframe, however if the dataset was much larger apply() might have been considered for efficiency and speed.

Upon completion this was saved as a new csv file to be loaded onto a second file page in python for the data to be analyzed. Initial analysis was carried out to find out how many regions were listed in the dataset and the subsequent count of countries in each region, following this it was important to find out how these regions scored in the newly added ladder category column. From the 140 countries listed, 2 scored as ‘suffering’, 18:’Thriving, and 118:’struggling’, most notable from this research was that 14 of the 18 countries listed as ‘Thriving’ belonged to Western Europe.
A function was created utilizing Pandas’ groupby() function to examine a variable of choice (‘Regional Indicator’), in order to avoid the groupby output having an index/multi-index on rows, “as_index=False” was passed to the groupby operation, following this the last part of the syntax;
agg() function was chosen to enable multiple statistical calculations to be passed.

For the purpose of passing a select few rows(GDP, social support, etc) through this function , the dataframe was sliced using iloc and stored as a new data frame. A for loop was then used to iterate through this new dataframe of interest using the function created, this was chosen over a ‘while loop’ as it was known how many times the code needed to be performed.

After performing aggregated statistics on the selected columns, a correlation was conducted between the Ladder score and the columns of interest, corrwith was used as the data was separated into a Series and data frame. This was chosen over df.corr as the desired output was Ladder score vs selected columns, rather than all the columns against each other, however seaborn was utilized to visualize a heatmap of the latter, see figure 2 below:



The ‘coolwarm’ palette was chosen for the heatmap to highlight both the highs and lows, as can be seen from the above figure; GDP, Social Support, Life Expectancy, and Ladder score all had strong positive correlations amongst themselves. This heatmap was used initially to determine what areas of interest could be researched/visualized further. In order for the labels to not be cut off plt.tight_layout() was applied.

Figure 3:

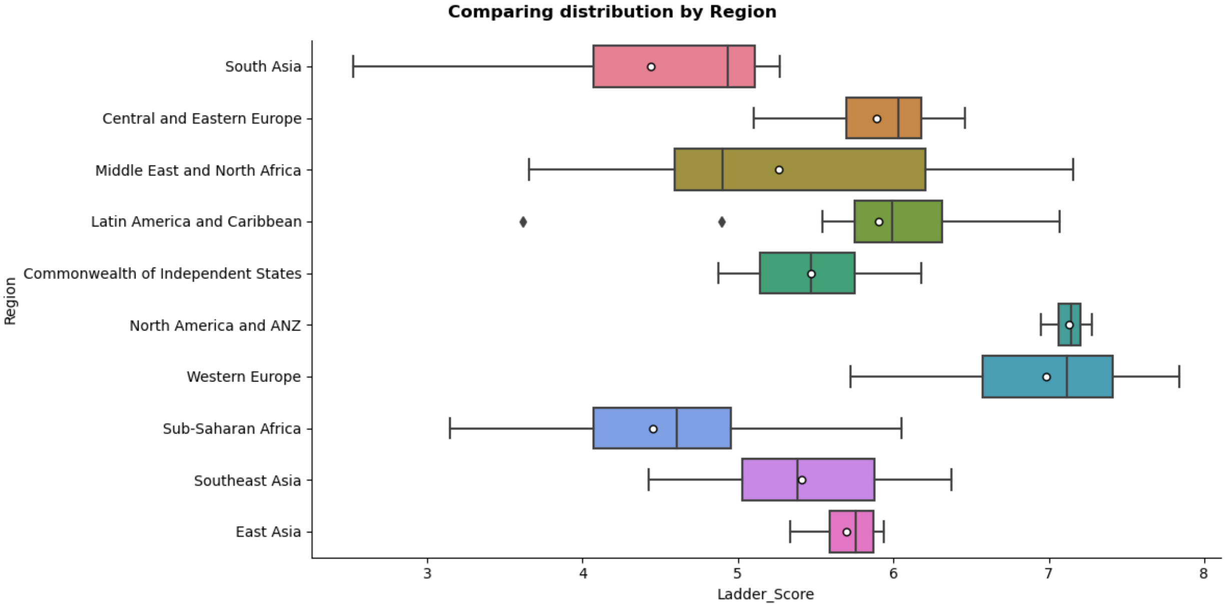
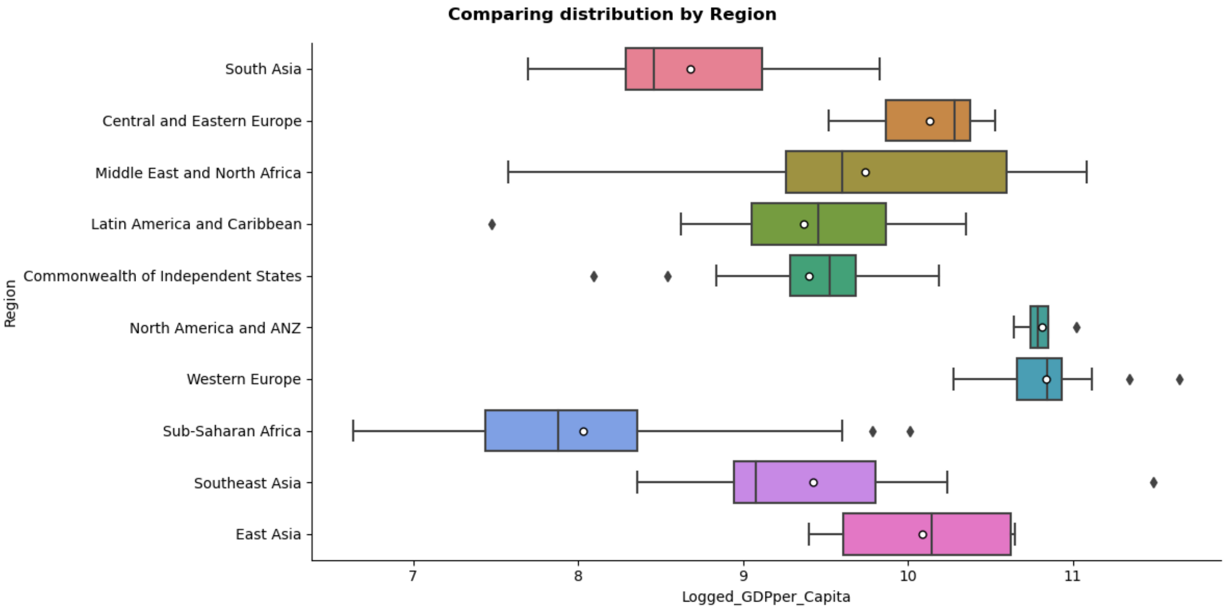
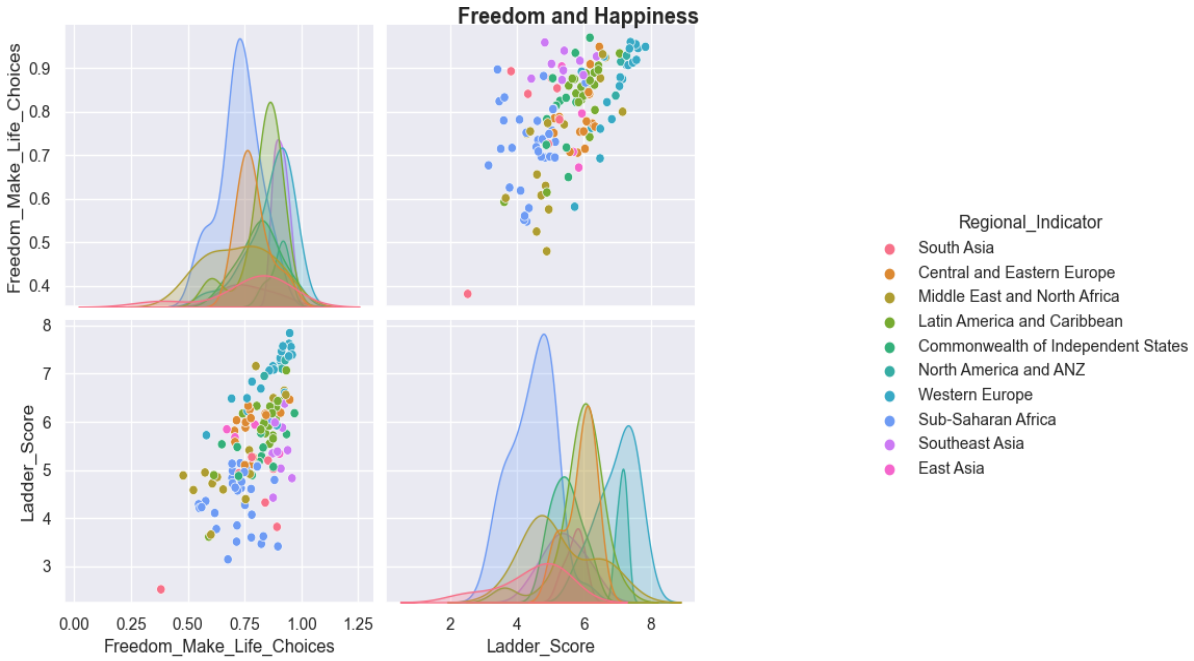


Figure 4:



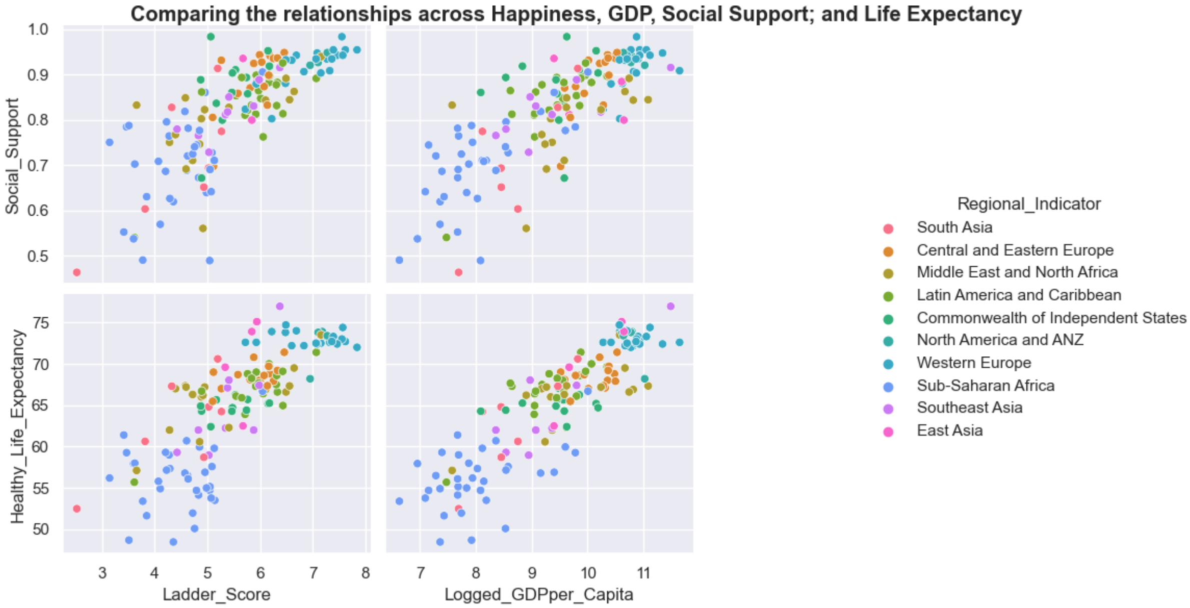
Multiple columns from the dataframe were passed through a function to plot the regional distribution of each column, a box plot was chosen to visualize this providing an indication of each region's symmetry and skewness. If the number of observations was larger, a swarmplot or striplot would've been used to visualize the distribution, however this dataset was too small to utilize this effectively so the boxplot was chosen. The white circle was added to each region's box to show their respective mean score.

Figure 5:



After the regional distribution analysis across various columns, further research was conducted using Seaborn’s pairplot to investigate the probability density and relationship between two numeric columns, plotting a Kernel density estimate and scatterplot to visualize this.

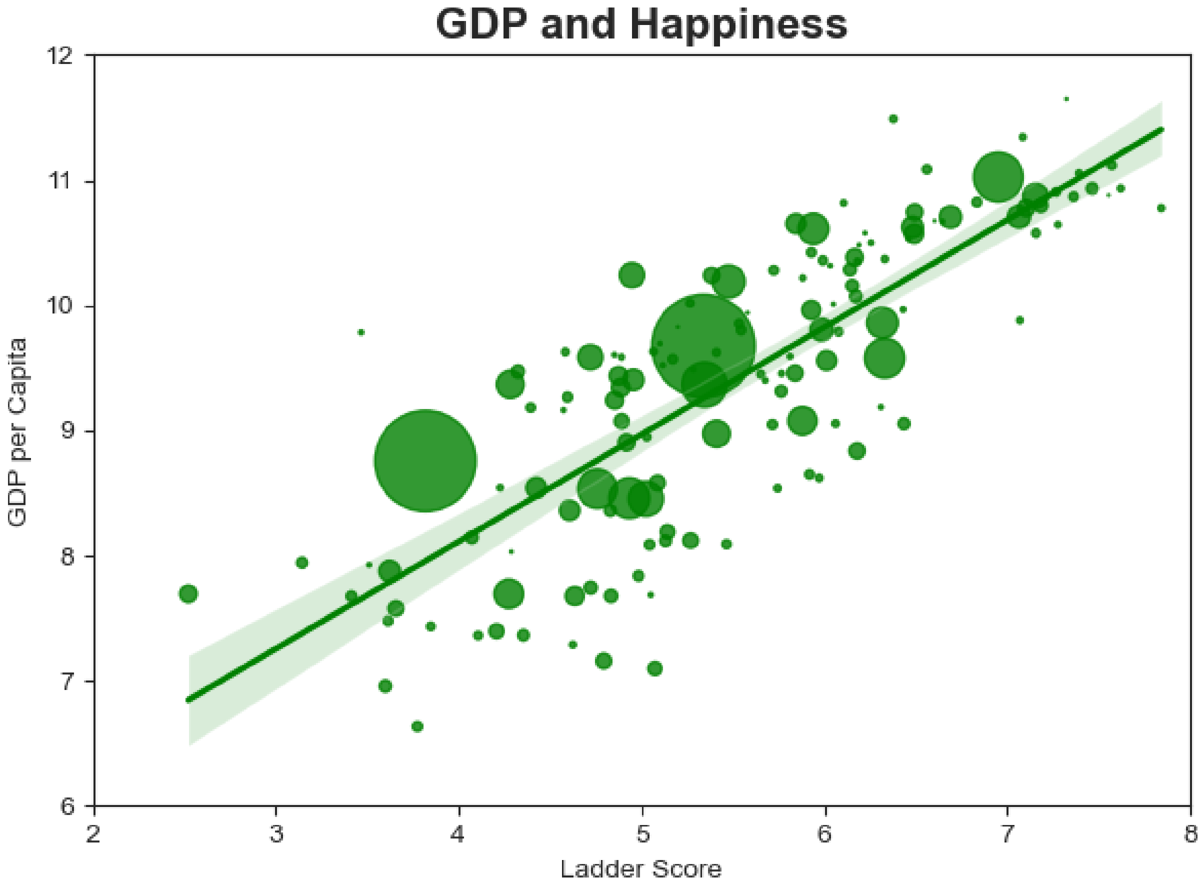
Figure 6:



A second pairplot was used to further examine the relationship between the four variables that had shown the strongest correlations in figure 1.

In figures 3-6, the ‘husl’ palette was used to distinguish between regions.

Figure 7:



Finally a regplot was used to plot the data and a linear regression model fit between GDP and ladder score, the bubble size was set to correspond to the population of each country, the x and y axis limits were adjusted to fit in all observations and alpha was set to 0.8 to ensure visibility of each observation.

Insights

- Figure 2, displays the relationship between each column and highlighted where further research could be carried out, showing a range of values from 1 to -1,leading to a focus on visualizing the strong positive correlations annotated in the graph, however further research could be conducted on each and also on the negative correlations shown in blue

- It can be seen from both boxplots, that there is a disparity between regions. Most notably the difference between the sub saharan african region and western europe, the former is distributed on the lower side of both GDP and ladder score, while western european countries are in contrast to this much higher, with the majority of countries lying higher than the highest sub saharan african country

- When viewing the distribution more closely in figure 5, it can be seen from the KDE and scatter plots, that the majority of countries are densely populating a range of 0.5 - 1.0 in 'Freedom to make life choices' with Western Europe being on the higher end of this distribution, and that there appears to be a positive linear relationship between the two variables

- Figure 6, provides an insight into the relationship between 4 variables, and shows a strong positive correlation in each graph. As each graph groups points by region, it can be seen that once again the majority of the 33 sub saharan african countries are scoring lower than the rest of the regions.

- Finally Figure 7, further investigated the relationship between GDP and ladder score, using a linear regression model fit represented by the line passing through the graph and the translucent band lines describe a bootstrap confidence interval generated for the estimate, highlighting the positive correlation observed.

- It can be seen from many of the graphs above that there were many positive influences on ladder scores in each region, and that generally a higher logged gdp per capita equalled a higher ladder score observed, this was also shown with life expectancy, social support and freedom to make life choices

.