# World Happiness Report

By Xinyu Meng xm73

Data 101

March 5th 2023

[World Happiness Report | Kaggle](https://www.kaggle.com/datasets/unsdsn/world-happiness)

Background:

As a computer science student who is very interested in big data, I chose to do research on the happiness value of the world because this data set not only takes into account the happiness scores of different countries, but also takes into account various factors including economics, corruption, etc. scope.

The report is significant because it provides insights into the factors that contribute to happiness and well-being around the world. It can be used to inform policies and interventions aimed at improving happiness and well-being in different countries. In this analysis, we have will explore the 2019 World Happiness Report dataset to investigate whether there is a significant difference in the happiness score between the top 50% and bottom 50% countries. Our analysis has shown that there is indeed a significant difference in the happiness score between these two groups of countries, with the top 50% countries having a higher median happiness score than the bottom 50% countries.

Introduction:

The World Happiness Report is a landmark survey of the state of global happiness. The report ranks 156 countries by how happy their citizens perceive themselves to be. The data is released on a yearly basis and is based on the Gallup World Poll. In this analysis, we will be exploring the 2019 World Happiness Report dataset, which can be found on Kaggle.

Data Set:

Text

Description automatically generated

Number of attributes/columns: 9

8 numerical + 1 categorical

Distribution of Happiness Scores:

Chart, histogram

Description automatically generated

As shown in the above graph, the Happiness Score ranges from 3 to 8. The distribution of the Happiness Score is slightly skewed to the right, with the majority of the scores falling between 4.5 and 6.5

Splitting Dataset:

First we split the dataset into two groups, where the first group contains data with top 10% life expectancy and the other group contains data with bot 10% life expectancy.

Barplots for each group are shown below.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedWe can see that the happiness score for the top 10% life expectancy is evenly distributed around 5.5 to 7, while the happiness score for the bot 10% life expectancy for the bot 10% life expectancy is evenly distributed across 3-5.

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

The boxplot for happiness scores between the two groups also shows that the outliers for each range of the health life expectancy are relatively trivial.

Statistical Analysis:

Permutation Testing

Null Hypothesis 1: There is no significant difference between the happiness score of top 10% healthy life expectancy and the happiness score of bot 10% healthy life expectancy.

Alternative Hypothesis 1: The happiness score of top 10% healthy life expectancy is significantly larger than the score of bot 10% life expectancy.

First, add a group column to each dataframe with ‘top’ or ‘bot’ keyword. Then, merge the two dataframe using the rbind function.

Chart, histogram

Description automatically generated

After conducting the permutation test, the p value I obtained is 0. Thus, we can immediately reject the null hypothesis, indicating that the difference in happiness score between top 10% life and bot 10% life expectancy is significant.

Conclusion:

Consequently, it can be concluded that healthy life expectancy has played a large part in the happiness score of each country. While our analysis provides strong evidence for the difference, it is important to consider potential sources of bias or hidden variables that may be affecting my result. For instance, the Gallup World Poll may not be representative of all individuals within a country, and cultural differences in how happiness is perceived and reported may also affect our result. Since the Poll is conducted mostly online, it is obvious that the result might not consider the ideas for other people in the world without access to internet and computers.

Alternative hypothesis2: The happiness score of top 10% gdp and social support combined is significantly larger than the score of bot 10% gdp and social support combined.

Null hypothesis 2: There is no significant difference between the happiness score of top 10% gdp and social support combined and the happiness score of bot 10% gdp and social support combined.

We also start by splitting the data set into two groups, on is the data for top 10% gdp and social support combined, the other is for happiness score of bot 10% gdp and social support combined. The following plots will show the distribution of the data.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

A picture containing text

Description automatically generated

Chart, scatter chart

Description automatically generated

First, add a group column to each dataframe with ‘top’ or ‘bot’ keyword. Then, merge the two dataframe using the rbind function.

Chart, histogram

Description automatically generated

After conducting the permutation test using the library, I reached the p-value of 0, which means that we can immediately reject the null hypothesis as the data clearly shows that people with better gdp and social support is indeed happier than the other.

Conclusion:

Consequently, it can be concluded that gdp and social support has played a large part in the happiness score of each country. While our analysis provides strong evidence for the difference, it is still worth considering why people with higher gdp and social support has better happiness score. Also, during the research I conducted, most people with higher gdp seems also has higher social support, I think maybe it's the people who have lower gdp who need more social help to get out of their situation.