"Happiness and Mental Health Disorders | Exploratory Data Analysis | Aarya Desai | Jeremy Tan | Cassie Kang | Osama Ahmed"

Data Overview: This assignment was a continuation of a research proposal submitted which outlined two potential datasets that we could use and further explore to ask two research questions. Based on the feedback, this assignment will be dedicated to conducting exploratory data analysis on the Happiness and Mental Health Dataset chosen from the previous report. This dataset is a combination of two datasets, namely, the World Happiness Dataset, 2022 retrieved from the official World Happiness Report website which conducts a yearly survey and analysis of world happiness metrics across the globe (i.e., each country), including different variables such as, life ladder, log GDP per capita, social support, life expectancy at birth, freedom, perception of corruption, generosity, positive affect and negative affect. The second dataset used was a World Mental Health Disorders dataset retrieved from the University of Oxford website containing datasets related to mental health metrics across the globe (i.e., each country) spanning over multiple years (). The latter dataset included metrics for mental health disorders recorded and analyzed on a Likert Scale (1 = Least Like to have given mental health disorder to 5 = Most likely to have given mental health disorder), and included this metric for the following disorders: Alcohol Use Disorders, Drug Use Disorders, Schizophrenia, Bipolar Disorders, Anxiety Disorders, Eating Disorders, and Depression. The resultant dataset had 1462 observations and 16 unique variables. These variables could be categorized as follows:

The following is also a brief overview of the descriptive statistics for each of these variables, including the mean, median, standard deviation, and confidence intervals of each.

		Mean	Median	Standard Deviation	Confidence Interval
1	Year	2011.813	2012	3.440	2011.636, 2011.989
2	Schizophrenia (%)	0.212	0.200	0.045	0.210, 0.214
3	Bipolar disorder (%)	0.743	0.708	0.151	0.735, 0.751
4	Eating disorders (%)	0.263	0.201	0.172	0.254, 0.272
5	Anxiety disorders (%)	4.008	3.556	1.277	3.942, 4.073
6	Drug use disorders (%)	0.879	0.683	0.521	0.852, 0.906
7	Depression (%)	3.460	3.479	0.641	3.427, 3.493
8	Alcohol use disorders (%)	1.668	1.510	0.994	1.617, 1.719
9	Life Ladder	5.448	5.347	1.142	5.389, 5.506
10	Log GDP per capita				9.280, 9.399
11	Social support				0.806, 0.818
12	Healthy life expectancy at birth	62.750	64.620	7.202	62.381, 63.120
13	Freedom to make life choices				0.726, 0.741
14	Generosity				-0.006, 0.012
15	Perceptions of corruption				0.742, 0.762
16	Positive affect				0.650, 0.661
17	Negative affect				0.258, 0.267

Table 1: Descriptive Statistics

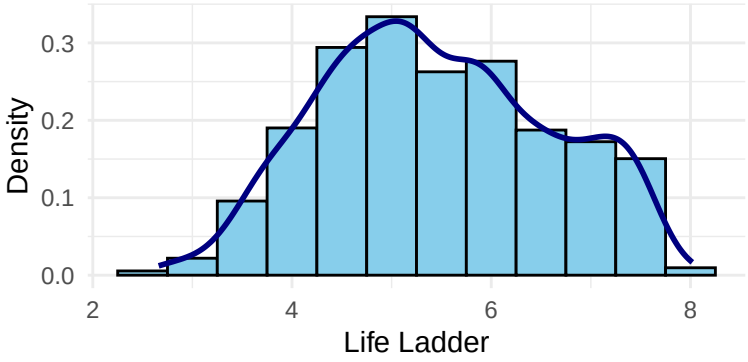
Primary relationships of interest: These datasets were then merged on corresponding Country and Year in each dataset. Based on this merged dataset, our team wanted to investigate the following research questions:

1. To what extent do happiness metrics (predictor variables: log GDP per capita, social support, life expectancy, freedom, perception of corrpution, generosity) predict mental health disorders, specifically, depression (outcome variable: continuous) in countries?

2. To what extent do mental health disorders and happiness metrics (predictor variables: schizophrenia, bipolar disorder, eating disorders, anxiety disorders, depression, log GDP per capita, social support, freedom, perception of corruption, generosity) predict life ladder (outcome variable: categorical)?

For the second research question, a binary outcome variable needed to be calculated from the original, continuous variable Life Ladder. This was to gain a more discerning outlook on the influence of various variables on how *good* or *bad* a countries' overall quality of life is. Therefore, for the same, Life Ladder was converted into a categorical variable named Lifestyle Category with two levels: Subpar Quality of Life and Satisfactory Quality of Life. This was done by assigning the prior factor to observations falling below the median of Life Ladder and the latter for those observations with a Life Ladder value above the median ($Mdn = 5.35$).

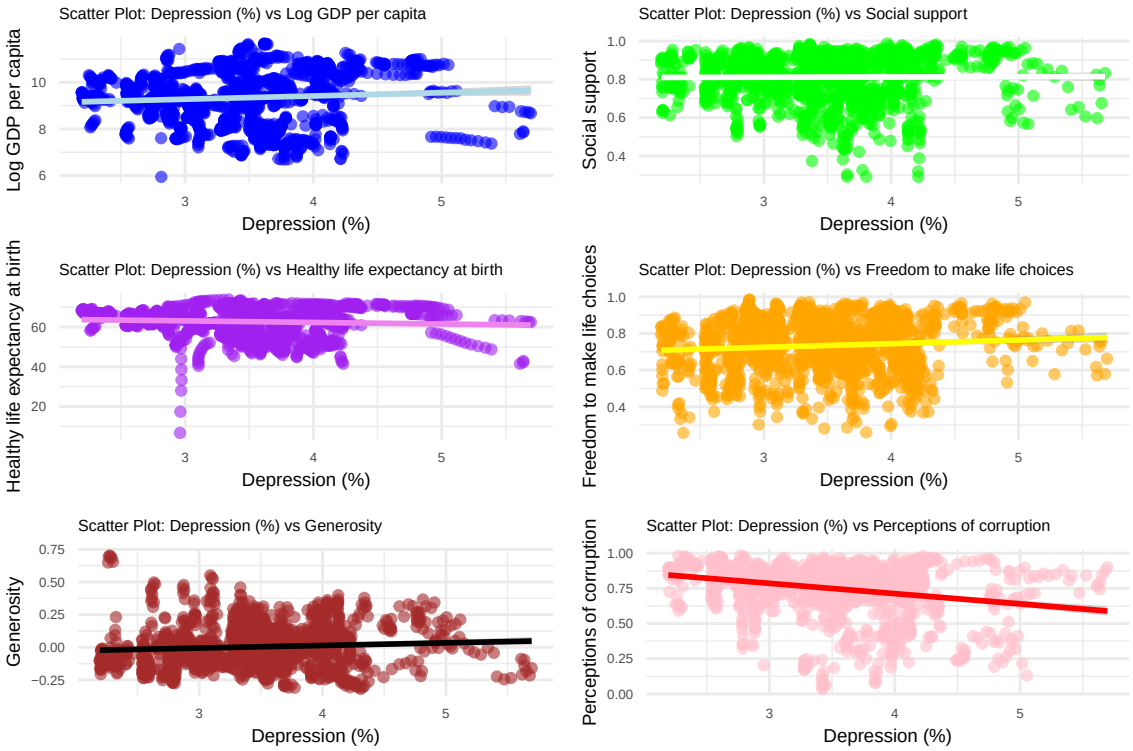
Figure 1. Life Ladder Distribution



Outcome variables

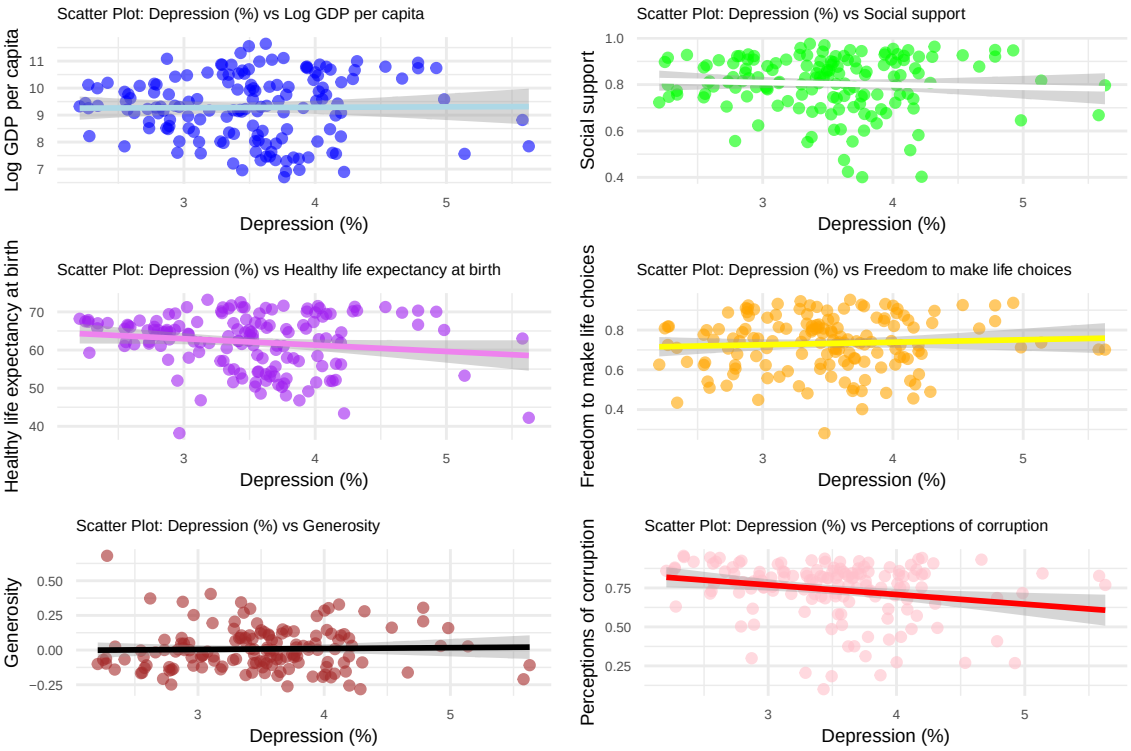
Research Question 1 (RQ1): The following are plots to visually represent the predictors variables for this research question against depression.

Figure 2. RQ1 predictor variables against the outcome variable depression



As seen in the figures above, the data is showing distinct patterns in a clustered fashion. This is because the data is being inherently group based on the country, changing very minimally by the year. Therefore, to have a better understanding of the relationship without having sub-trends in the graphs, the data was transformed, where the mean of each variable was calculated per country in the dataset. Though this reduced the number of observations from 1462 to 152, this was a necessary step required to gain some more understandable exploratory information regarding the data without accounting for the confound of the Year variable. Below are the new plots with the averaged variables plotted against depression.

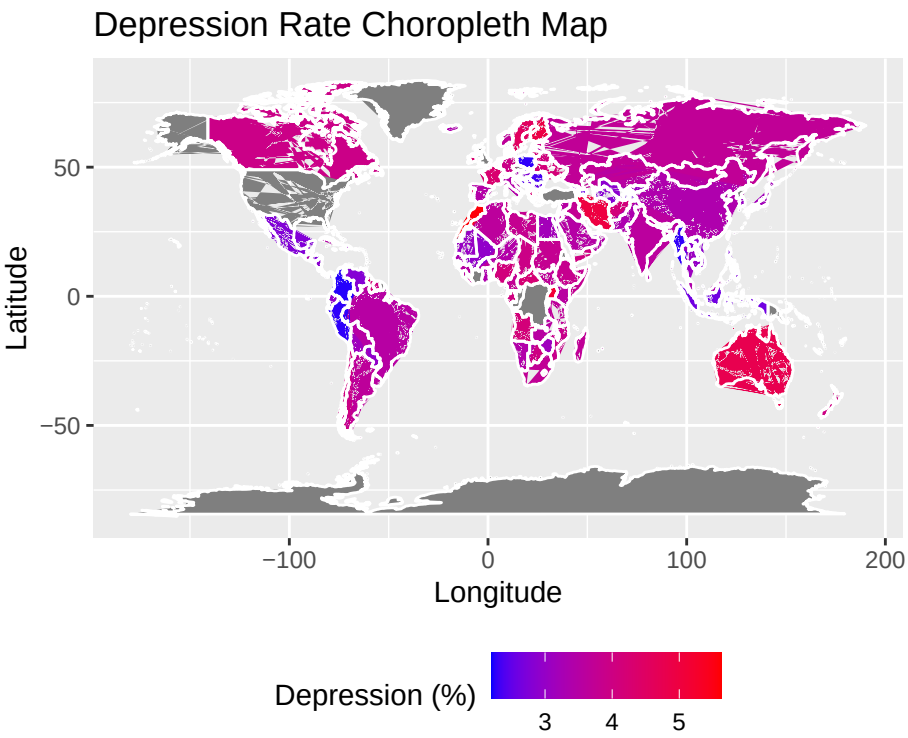
Figure 3. RQ1 predictor variables against the outcome variable depression, with variables grouped and averaged by country



These grid plots reveal that the connections between depression and the chosen predictor variables are not notably pronounced. The only variable that shows a weak, negative relationship with depression is 'Perception of Corruption'. On the other hand, the remaining variables exhibit relatively flat relationship lines, indicating a more subtle correlation, and providing us with limited additional insights into their interrelationship with depression. Therefore, these plots do not provide much insightful information regarding RQ2.

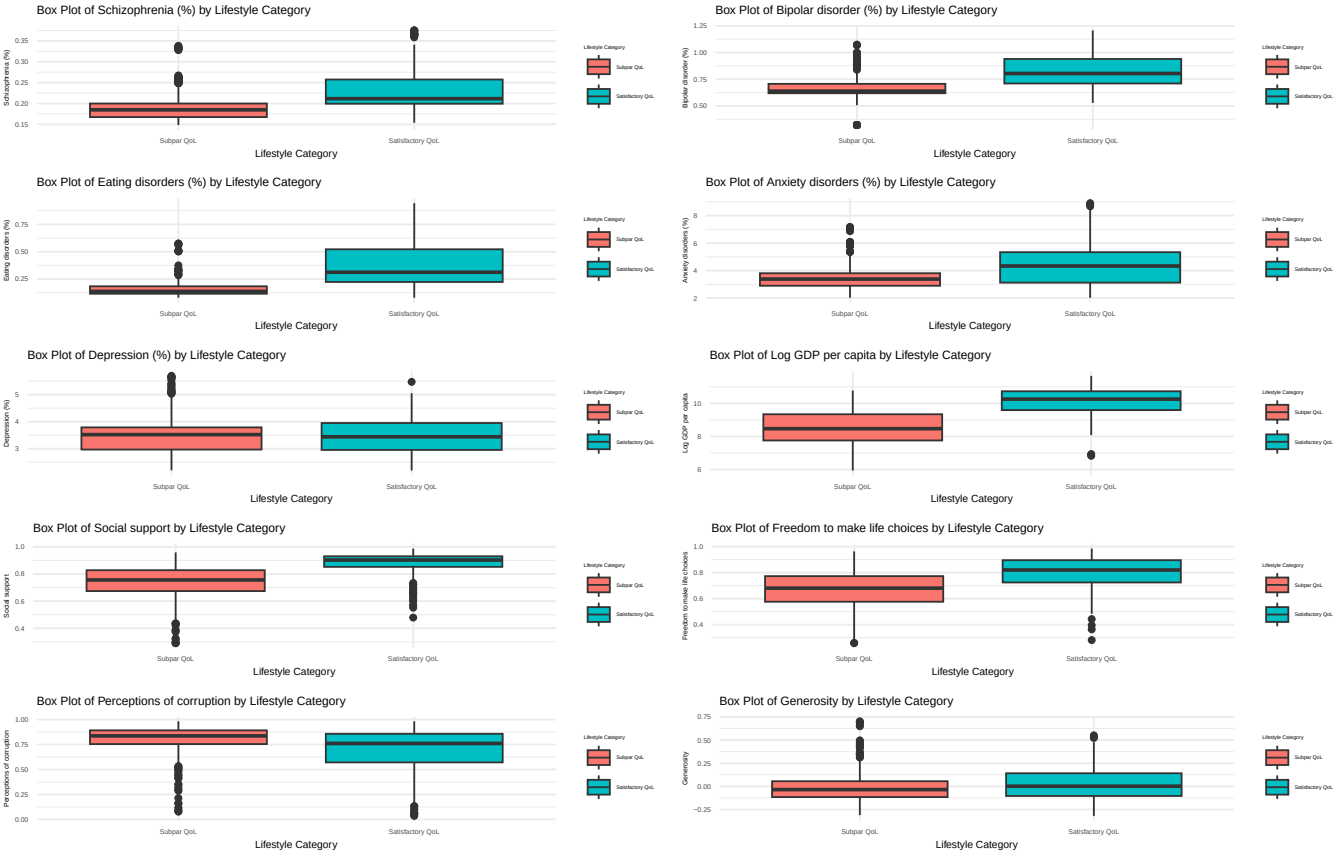
Lastly, for some exploratory understanding of the general distribution of depression rates across the globe, a map plot was made to see the variation in depression rates between countries (Figure 4). As seen in the figure, there is a notable variation in the depression metric between countries. Therefore, our model may benefit from using Country as a factor or confound, to gain a more in-depth understanding of RQ1.

Figure 4. Heatmap to visualize depression rates across countries.



Research Question 2 (RQ2): To gain a better understanding of the relationship between the binary outcome variable, Lifestyle Category, and the continuous predictor variables chosen for RQ2, bar plots to show the difference in Subpar vs Satisfactory Quality of Life (QoL) were plotted (Figure 5).

Figure 5.



Other Characteristics: In the analysis of the World Happiness Report and World Mental Health Data, it was essential to consider variables that would not be included in our model. These variables are "Country"/"Entity" and "Year."

Year is a categorical variable representing the years for which the data was collected. It has 13 unique values with varying repetitions. The "Country" variable represents the countries for which the data was collected. It is a categorical variable with 153 unique countries, each with varying repetitions.

Potential Challenges:

1. Assumptions about Data Homogeneity: It is assumed that the data is uniform across countries and years, disregarding potential influences from these variables. This is crucial because factors like GDP and cultural nuances can significantly impact happiness metrics.
2. Addressing NAs: There's a need to understand why there are missing values (NAs). This could be due to merging processes or a specific pattern (e.g., related to certain years, countries, or occurring randomly). It's important to address this issue, as it could affect the accuracy of the results or model.
3. Binary Variable Creation: Creating a binary variable based on the median may not capture nuanced differences between groups. This might limit the depth of insights derived from the analysis.
4. Limited Insights from EDA for RQ1: Initial exploratory data analysis (EDA) for Research Question 1 suggests weak relationships. This could potentially lead to a model with limited predictive power and insights.
5. Promising Aspects of RQ2: Research Question 2 holds more promise, but there are still important considerations, starting with an examination of how the binary variable was derived.
6. Multicollinearity Concerns: Given the nature of predictor variables in the World Happiness Dataset, there's a possibility of high correlation among them. This can lead to multicollinearity in regression models, complicating the interpretation of individual predictor effects on the outcome.