What are the effects of social support, healthy life expectancy, generosity, freedom to make life choices, perception of corruption on GDP per capita?

Introduction section:

SECTION A:
Why is this topic important?
People really focus on the economic factors of a country when they think of GDP, but the social factors also contribute a lot towards a country's development and GDP per capita is a measure of this (1). Hence, it is important to know how social factors like social support, healthy life expectancy, generosity, freedom to make life choices, perception of corruption affect GDP per capita. My goal for this paper is analyze and explain the effects of social support, healthy life expectancy, generosity, freedom to make life choices, perception of corruption on GDP per capita.

SECTION B:
Similar literature on this topic:
This paper's (2) goal is to assess which predictor is most significant among literacy, poverty, under slums, corruption, government effectiveness, political stability and adult morality when determining GDP by doing linear regression on multiple models and ranking the predictor based on the t-test of the coefficients. The result is that literacy is the most important among the predictors. The similarity to my research question here is that this paper is also looking at social factors for GDP, but the predictors involved are different.

The paper's (3) goal is to see if there is any correlation between GDP and final consumption by doing a simple linear regression with GDP as the response. The result is that final consumption will result in an increase of 1.22 units in the monetary value of GDP. Here, the similarity with my research question is that the GDP is the response variable and the difference is that only one economic factor is used to see the correlation.

The paper's (4) goal is to predict the GDP using economic factors: capital formation, total trade, interest rate, inflation rate, unemployment rate and stock exchange index using multiple linear regression. The result is that capital formation, total trade, interest rate, exchange rate and unemployment rate are found to be significant predictors of GDP and can explain GDP by 93% which is formulated using normal estimation. Here the similarity with my research question is that the GDP is the response variable and the difference is that only economic factors predict GDP.

Methods section:

Exploratory Data Analysis is done through scatterplots of the variables.

SECTION A:

Variable selection:
We do backward selection method which is fitting the model with all the predictors: social support, healthy life expectancy, generosity, freedom to make life choices and perception of corruption, and then try to reduce the model using anova, t-test, vif, partial f-test, aic and bic. We do ANOVA test on the whole model which tell whether there is any association between the response and any of the predictors included, under the null hypothesis that there is no association. If the p-value for the F-test from Anova is significant(less than the significance level 0.05) then we reject the null which means there is at least one significant predictor and continue with our analysis. We check for multicollinearity to see if there is any prominent relation between our predictors using a measure for multicollinearity: vif (Variance Inflation Factor). If any predictors have a vif value more than 5, then we should fit a model with reduced predictors to lower vif.

We assess the summary of the model, which gives the result for each predictor's t-test and p value. If any of the predictor's p-value is non-significant, we try to refit the model without some predictors and conduct partial F-test to see if the reduction is valid. The partial f test, under the null hypothesis says that there is no association between the removed predictors and response and we can fit a model without these predictors. We can reject or fail to reject the null based on the p-value of F-test being significant. We use AIC and BIC, which are metrics used to compare models, and the model which has a lower AIC or BIC is a better model. We also use R^2 (coefficient of determination) which determines the proportion of variance in the dependent variable, that can be explained by the independent variable.

SECTION B:
Model Violations and Diagnostics:
We are going to check four assumptions for each model we fit:
1) linearity
2) normality
3) uncorrelated error
4) common error variance

We assess the scatterplot of the variables by plotting them in same grid and we check for linearity (check for linear relationship between predictors and the response) and skewness in any variables (cluster on one side).
We plot a residual vs fitted values to check for linearity and constant variance. For linearity, if there is any pattern which is not linear, then try to transform the variable by finding an appropriate one using box cox which gives a lambda value that determines the type of function that is needed to be applied to that variable. Otherwise, we say there is a linear relationship. For constant variance, we look for any pattern, especially fanning where the residuals are gradually becoming more or less spread out. If there is, we use a variance stabilizing transformation to the response which will remove dependence of error variance on the predictors. For uncorrelated errors, we make sure the data points are independent of each other.
For normality, we plot a QQ plot(Quantile-Quantile). We check if all the points lie in a straight diagonal line. If not, we do box cox transformation which gives a lambda value to tell the type of function that is needed to be applied to the response.

Now we check for problematic points in the dataset: outliers (observations that are not near the trend of other points), leverage points (very distant from the center of all the predictors) and influential points (points that have a large impact on the regression line).

(For clarification, n refers to sample size, p refers to number of predictors in the paragraph below)
For leverage points, we find the hat matrix, which provides a measure of leverage.
It investigates whether one or more observations are outlying with regard to their X values and any values in hat matrix > $2 * \frac{p+1}{n}$, will be leverage points.
For outliers, we look at standardized residuals (r) and they quantify how large the residuals are in standard deviation units, and therefore, can be easily used to identify outliers. For small data set we look for r < -2 or r > 2 and for large data set r < - 4 or r > 4.
We look for influential points using cook's distance (Di) which is a measure of a single observation on a regression. For any point if Di > 50th percentile of F(p+1, n-p-1), it is influential. In order to look at all the observations, we do DFFITS (difference in fitted values). For any point |DFFITSi| > $2 * (\left(\frac{p+1}{n}\right)^{\frac{1}{2}})$, it is influential. DFBETAS (difference in betas) which looks at the direct effect that an observation has on the regression coefficient. For any point, |DFBETASj(i)| > $(2 * n^{1/2})$, it is influential.

Furthermore, we check for two additional condition:
Condition 1: Conditional mean response is a single function of a linear combination of the predictors.
Condition 2: Conditional mean of each predictor is a linear function with another predictor
We check Condition1 by checking the response against fitted values and see if there is a scatter around identity function which can be mapped by simple function.
We check Condition2 by checking for scatterplot of the predictors and check to make sure there is no non-linear exponential functions within any of the predictor and vif of the model.

SECTION C:
Model Validation:
In order to validate our model, initially before fitting the value we split the dataset into train and test dataset. We analyze and groom the training dataset and fit a model by checking assumption, transform variables if needed and check for problematic points, do variable selection and then compare the result from that by fitting a similar model in the test dataset, after doing all the similar preliminary checks in the test data as well. This makes sure that the results we got is not biased for the selected dataset.

Results section:

SECTION A:
Description of data:

The data is taken from Kaggle and consists of country, their happiness score, GDP per capita and their score of social support, healthy life expectancy, generosity, freedom to make life choices, perception of corruption. Some of the scores in this dataset are subjective because these are based on the country's citizen opinion.
Here is a numerical summary:

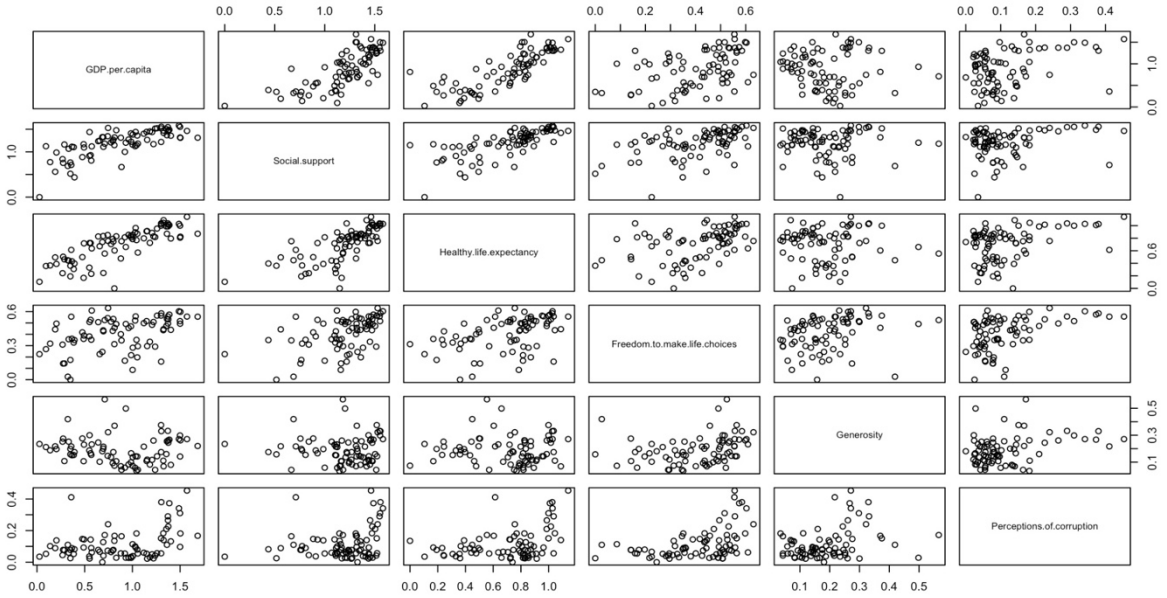
GDP.per.capita	Social.support	Healthy.life.expectancy	Freedom.to.make.life.choices	Generosity	Perceptions.of.corruption
Min. :0.0260	Min. :0.00001	Min. :0.00001	Min. :0.00001	Min. :0.0350	Min. :0.00001
1st Qu.:0.5495	1st Qu.:1.10650	1st Qu.:0.49600	1st Qu.:0.31425	1st Qu.:0.1095	1st Qu.:0.05325
Median :0.8825	Median :1.24450	Median :0.77700	Median :0.44100	Median :0.1750	Median :0.08100
Mean :0.8683	Mean :1.19181	Mean :0.70032	Mean :0.40703	Mean :0.1905	Mean :0.11623
3rd Qu.:1.2337	3rd Qu.:1.43325	3rd Qu.:0.87100	3rd Qu.:0.51975	3rd Qu.:0.2520	3rd Qu.:0.14625
Max. :1.6840	Max. :1.58200	Max. :1.14100	Max. :0.63100	Max. :0.5660	Max. :0.45300

Table 1: Numerical summary of all the variables included in the paper

By checking the min, max and the mean for the variable Social.support, we can assess that it is left-skewed. Similarly Healthy.life.expectancy and Freedom.to.make.life.choices is also left-skewed and Perception.of.corruption and Generosity is right-skewed. Visual representation of this is in figure 3.

SECTION B:
Model selection and analysis process:
The data is split 50-50 to make subsets: train and test data. We continue working on the train data and validate the final model using test data.
Then we do data exploration through scatterplot for the full model

Figure 1 Scatterplot of the variables.



The EDA highlights some issue with the model. The predictor Generosity and Perception.of.corruption is skewed, highlighting the potential to see maybe linearity problems or just poorly fitting models. We see no obvious violation of normality or constant variance.

In order to see if any transformation can be done or not, we use box cox transformation to find appropriate transformation for the variables and the result is:
Social.support has lambda value = 2, which means square transformation
Generosity has lambda value = 0, which means log transformation
Perception.of. corruption has lambda value = 0.5, which means square root transformation
Others have value 1 or close to 1.

After transforming the variables we check the scatterplot in figure 2 with the transformed variables and linearity seems to be satisfied. We then check for the other assumptions.

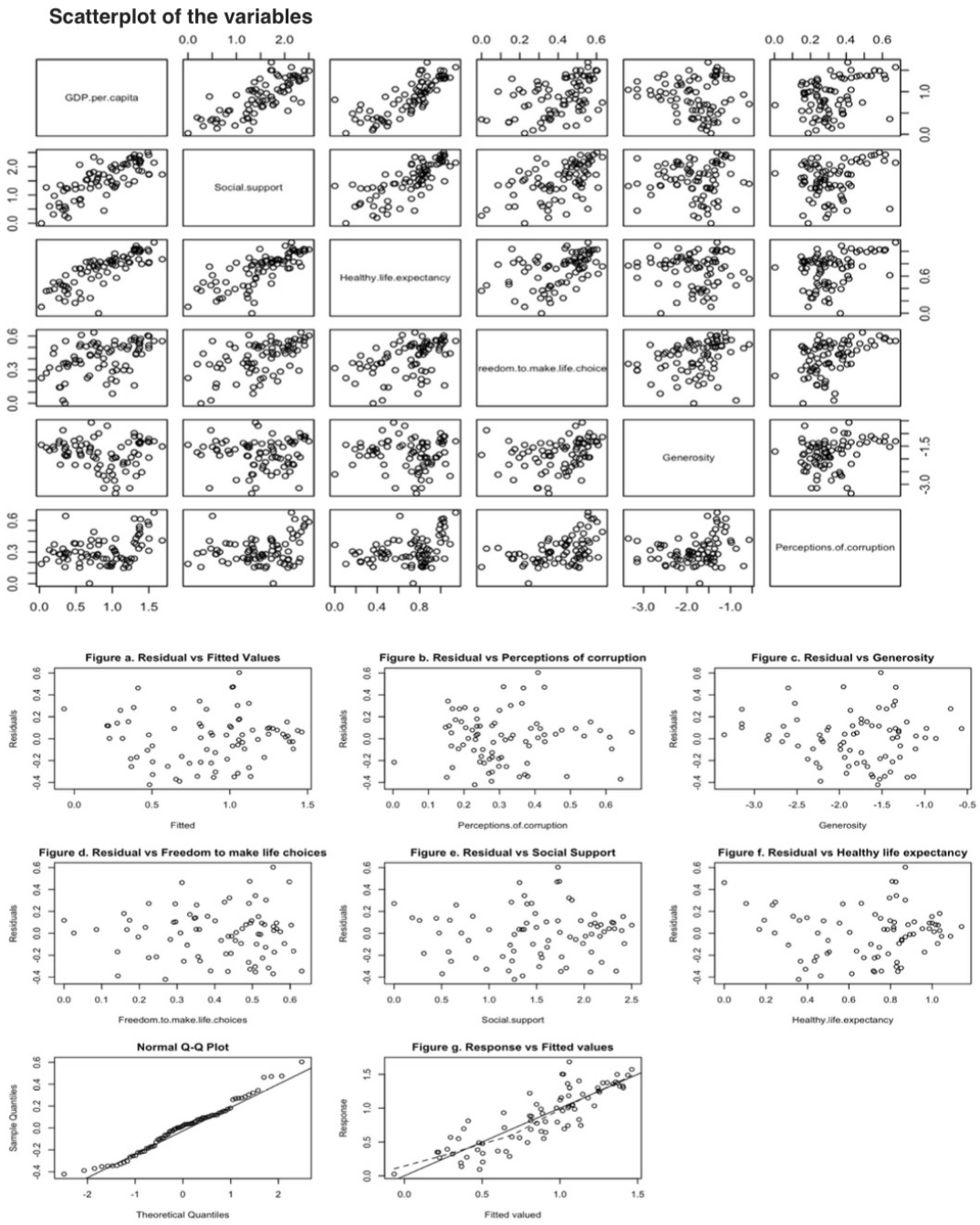


Figure 2 showing scatterplot of the variables on top half and qq plot and residual plots on the bottom.

We notice no violation of constant variance from figure 2a-e, uncorrelated errors as the data points are of different countries and are independent and do not influence each other. There is slight normality problem with the model as the points are deviating away from the diagonal line which can be seen from the normal QQ plot.
From figure 2g we see the model is satisfying condition1.

We check condition 2 by seeing the scatterplot which shows no predictors show any visible nonlinear/sinusoidal/exponential relationship with another predictor, which satisfies condition 2 and the results from vif which also show no multicollinearity present:

Social.support : 2.524678
Healthy.life.expectancy: 2.410481
Freedom.to.make.life.choices: 1.742079
Perceptions.of.corruption : 1.364310
Generosity : 1.182237

From the box plots in figure 4, we notice outliers in only freedom and perception variables. After checking for problematic points in our data, we identified 4 leverage points in the data that are distant from the rest of the observations in the predictor space. We also identified 4 outlier observations when considering the dataset as "small", but none when considering it as "large". No observations were identified as being influential on the entire regression surface, but we identified 4 who influenced their own fitted values and between 4-6 observations being influential on at least one estimated coefficient.

After all the preliminary checks, we look at the summary table of the full fitted model.
We look at the t-test of the predictors, Freedom.to.make.life.choices and Perceptions.of.corruption's p-value is not significant. We perform partial f- test to see if we can remove Freedom.to.make.life.choices. And the p value is 0.2745 and we can remove them. We perform another partial f-test to see if we can remove Perceptions.of.corruption and the p-value is 0.02899 which means we cannot remove this variable. We perform another partial f test to see if we can remove both , Freedom.to.make.life.choices and Perceptions.of.corruption and the p-value is 0.07889 which means we can remove them. But in order to decide if should remove both or either of them we look at AIC, BIC and R^2. We fit 2 extra models and compare with the full model after doing the assumption checks for each, which did not show any problems.
Model1: Without the variables Freedom.to.make.life.choices and Perceptions.of.corruption
Model2: Without the variables Freedom.to.make.life.choices
Model3: Full model

MODELS	AIC value	BIC value	R^2
Model 1	-3.142233	8.641311	0.718
Model 2	-5.342028	8.798225	0.7327
Model 3	-4.644911	11.852051	0.7372

Table 2 showing the three model and metric measure of goodness of the model

Model 2 is the best model among them.
Our final model:
GDP.per.capita = 0.01433 + 0.26427* $(Social.support)^2$ + 0.74033*Healthy.life.expectancy -0.48225*log(Generosity) + 0.44085* $(Perceptions.of.corruption)^{\frac{1}{2}}$

SECTION C:
Validation and goodness of the final model:
In order to validate our model we had split our original dataset into training and testing. We checked the assumptions for the test data and saws similar violations and needed same transformation as we did in the train data. Further checked the assumptions, conditions and problematic points then fit a similar model with the test data. Since the goal of the paper is to explain the effect we do not need to go in depth with our test data, to see similarity in

numerical summaries and approximately similar coefficients. However we do want the effects of the predictors to replicate the train data.
Model with test data:
GDP = -0.15406+ 0.16772 * $(Social.support)^2$ + 1.12840*Healthy.life.expectancy
-0.23020*log(Generosity) + 0.23175* $(Perceptions.of.corruption)^{\frac{1}{2}}$
We get almost similar model, except the intercept being a bit different, however the effect of the factors on GDP.per.capita is similar, which tells that our explanation of the predictors through train data is not biased to the dataset.

For knowing the predictive side of the model, we implemented K-nearest neighbour algorithm. We split the data into training and testing and we trained the k-NN model and evaluated its performance using Mean Squared Error (MSE). The MSE was too large and the accuracy was only 1.8% percent which means this is not a good model for prediction.

Discussion Section:

Final model interpretation:
The goal of this paper was to explain the effects of the social factors on the GDP.per.capita. The significant social factors are social support, healthy life expectancy , generosity and perception of corruption. GDP.per.capita increases with an increase in either Social.support or Healthy.life.expectancy or Perceptions.of.corruption. For example if everything else is constant it increases by a 1.12840 for every one unit increase in Healthy.life.expectancy. Similarly it decreases by -0.23020 for every one unit increase in logarithmic generosity. The model should correctly describe the effects of these factors given the assumptions are satisfied and there is no violation. The predictors in this model can explain GDP per capital by 73% normal estimation.

Limitation:
The model has slight issue with the normality assumption. Even trying to transform the variables did not help and trying to fix this issue, was making it hard to interpret the model and was violating other assumptions. This assumption allows us to correctly interpret p-values, use t and F tests for hypothesis testing because OLS estimators are normally distributed and to derive the probability distribution of it as for any linear function of a normally distributed variables itself is normally distributed. This could be a problem where we used p-values to reduce the model to find the significant predictors.

We had outliers, leverage and influential points in our data. They impact our ability to accurately measure means/centers and spread. They also have an effect on the regression line. However we did not remove them since they were not data error and each point represents GDP.per.capita and significant factors for a country, which should not be discarded.

Appendix:

Figure 3 Visual summary of the variables.

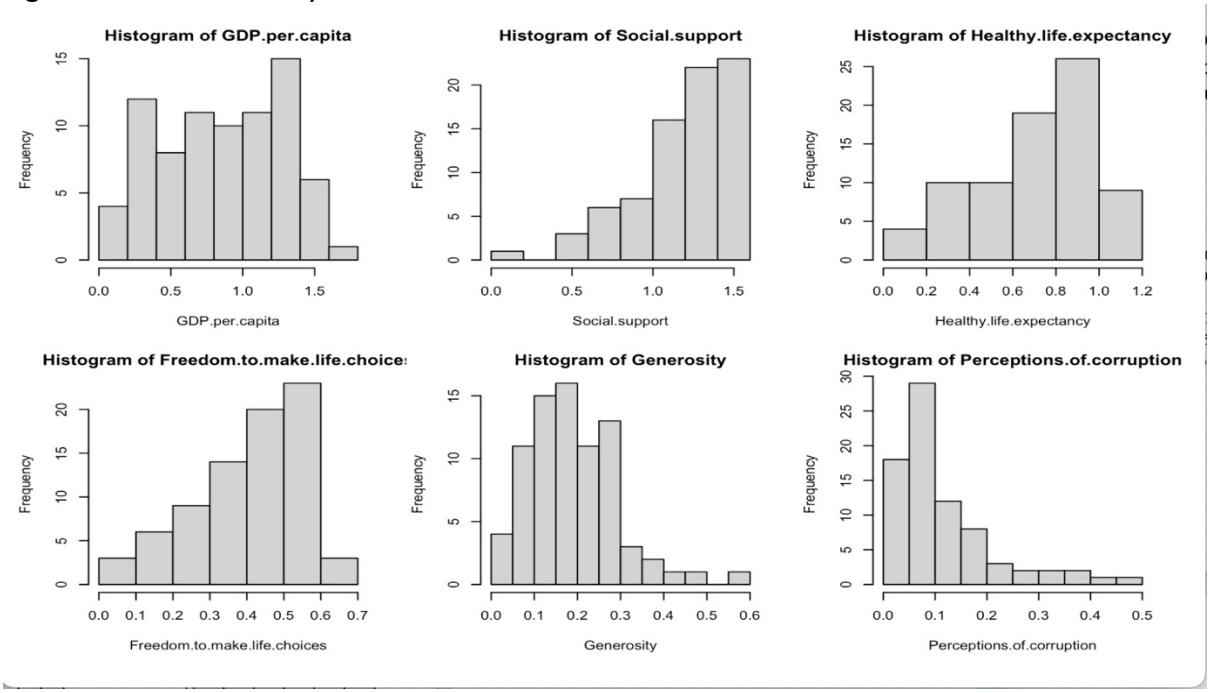
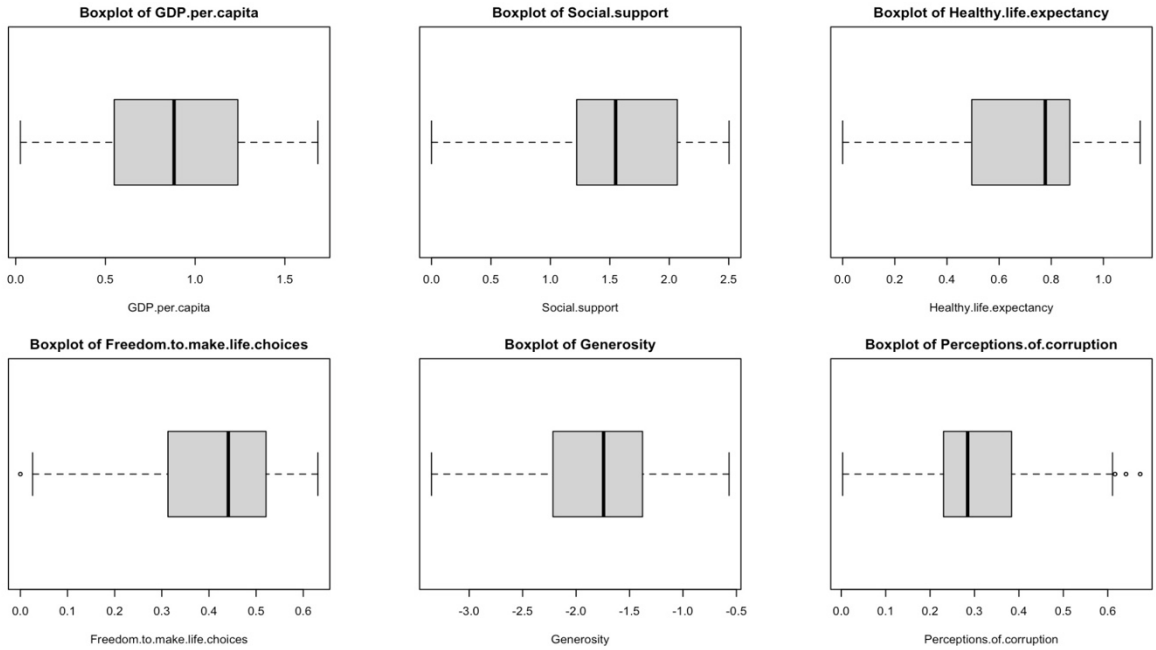


Figure 4 Boxplots of the variables.



Citation:

1. Bhatt, G., Birol, F., Yergin, D., Hausmann, R., Nordhaus, T., Lloyd, J., Graaf, T. V. D., Pescatori, A., Stuermer, M., Zettelmeyer, J., Tagliapietra, S., Zachmann, G., Heussaff, C., Celasun, O., Iakova, D., Bordoff, J., Henríquez, M., Khera, P., Ogawa, S., … Spilimbergo, A. (n.d.). *Finance and Development Magazine*. IMF. Retrieved December 21, 2022, from https://www.imf.org/en/Publications/fandd

2. Sinha, G. (2020, March 27). *Using regression to determine the most important social factors impacting GDP per capita*. Medium. Retrieved December 21, 2022, from https://towardsdatascience.com/using-regression-to-determine-the-most-important-social-factors-impacting-gdp-per-capita-2296f4f02dcf

3. *A statistical analysis of GDP and final consumption using simple linear ...* (n.d.). Retrieved December 21, 2022, from https://www.researchgate.net/publication/227382939_A_STATISTICAL_ANALYSIS_OF_GDP_AND_FINAL_CONSUMPTION_USING_SIMPLE_LINEAR_REGRESSION_THE_CASE_OF_ROMANIA_1990-2010

4. Urrutia, J. D., & Tampis, R. L. (n.d.). *Regression analysis of the economic factors of the gross domestic product in the Philippines*. Journal of Fundamental and Applied Sciences. Retrieved December 21, 2022, from https://www.ajol.info/index.php/jfas/article/view/168547