Table of Contents

Question 1 ..... **2**

(a) Problem Statement *2*

(b) Introduction *2*

(c) Research Objectives *3*

(d) Analysis Results and Interpretations *4*

(i) Dataset Feature Description 4

(ii) Stepwise Multiple Linear Regression 7

(ii). Model Results Analysis (Model 2 to Model 7) 12

(iv) Stepwise Multiple Linear Regression Assumptions 16

(v) Significance of Regression Model 20

(vi) Hypothesis Testing for Coefficients of Regression 21

(e) Conclusions & Recommendations *22*

Question 2 ..... **23**

(a) Purpose of Factor Analysis *23*

(b) Exclusion of Non-Metric Independent Variable *23*

(c) Factor Analysis on SPSS *23*

(i) Communality 25

(ii) Improving factorability 25

(iii) Eigenvalue 26

(iv) Factor cross-loading 27

(v) Recommendations for cross-loading 28

(vi) Grouping of Metric Independent Variables 28

References **29**

Question 1

(a) Problem Statement

This report aims to predict the happiness score of countries through developing a multiple linear regression model based on variables measured within the World Happiness Report published by United Nations (World Happiness Report, 2019) spanning across 2016 to 2020 (Singh, 2021) . The multiple linear regression model is useful in analysing the contribution of each variable towards happiness. Additionally, factor analysis is implemented to group the variables and empower focus and clarity in analysing the true underlying factors affecting happiness.

(b) Introduction

The World Happiness Report is the averaged survey results of roughly 1000 individuals across 150 countries to describe happiness through the answers to the Cantril ladder question along with identified key variables to explain the variation in happiness. The Cantril ladder question and key variables identified are described in Table 1.

Measuring subjective well-being has been an important area of social research to contribute directly to accurately determining the happiness and life quality of each country's population. To identify the key aspects that determine happiness, researchers have incorporated results from the Gallup World Poll (GWP) and World Values Survey (WVS) and found that happiness is heavily dependent on income levels as compared to life satisfaction (Easterlin et al., 2010). However, as more data became readily available, it was shown that an average between income and life satisfaction is a more powerful indication of happiness (Fanning & O'Neill, 2019) (Helliwell et al., 2017). Through incorporating data from an established and reliable World Happiness Report, this study aims to model happiness through a Multiple Linear Regression and perform factor analysis to contribute towards subjective well-being measurement.

(c) Research Objectives

The aim of this study is to predict a country's happiness score through a Stepwise Multiple Linear Regression technique. To achieve this aim, three objectives are derived. These three objectives will be tested using it's respective set of proposed hypotheses.

Objective 1

- Determine if GDP per capita affects happiness score

H_o: The logged GDP per capita of a country has a significant relationship with happiness score

H_1: The logged GDP per capita of a country has no significant relationship with happiness score

Objective 2

- Determine if life expectancy affects happiness score

H_o: The life expectancy of a country's population has a significant relationship with happiness score

H_1: The life expectancy of a country's population has no significant relationship with happiness score

Objective 3

- Determine if perception of corruption affects happiness score

H_o: The perception of corruption within a country has a significant relationship with happiness score

H_1: The perception of corruption within a country has no significant relationship with happiness score

(d) Analysis Results and Interpretations

(i) Dataset Feature Description

The World Happiness Report dataset contains 613 instances whereby 24 instances consists of missing values. After investigation, it was found that the missing values are random with proportion being approximately 3.9%. As this falls below the threshold of 10% (Hair Jr et al., 2009), the missing values are not significant to the analysis and have been cleansed.

The cleaned dataset for analysis has dimensions of 589 instances and 11 variables, whereby 9 variables are metric continuous variables inclusive of the target variable. The dataset variables are described in Table 1 and summarized in the Variable View and Data view of the IBM SPSS Statistics software shown in Figure 1 and 2.

Variable	Variable Description
X_1 – Country	Country surveyed
X_2 – Year	Year of surveyed
X_3 – Happiness Score	National average where citizens were requested to rank their personal happiness level through a range of 0 (lowest) to 10 (highest)
X_4 – Logged GDP Per Capita	Natural logged GDP per capita of the country for adjustment of population growth
X_5 – Social Support	National average where citizens responded to whether social support is available in times of need, or not.
X_6 – Life Expectancy	National life expectancy based on data from World Health Organisation (WHO) Global Health Observatory data repository
X_7 – Freedom	National average where citizens responded to whether they are satisfied with freedom to make life choices.
X_8 – Generosity	National average where citizens responded to whether they have donated money to charity in the past month.
X_9 – Corruption	National average of citizen's perception of corruption within the country
X_{10} – Positive Affect	National average for positive emotions felt on the previous day
X_{11} – Negative Affect	National average for negative emotions felt on the previous day

Table 1. World Happiness Report Dataset Variable Description

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
x1	String	24	0	x1 - Country	None	None	12	Right	Nominal	Input
x2	Numeric	4	0	x2 - Year	None	None	8	Right	Scale	Input
x3	Numeric	5	3	x3 - Happiness Score	None	None	8	Right	Scale	Target
x4	Numeric	6	3	x4 - Logged GDP per Capita	None	None	8	Right	Scale	Input
x5	Numeric	5	3	x5 - Social Support	None	None	8	Right	Scale	Input
x6	Numeric	6	3	x6 - Life Expectancy	None	None	8	Right	Scale	Input
x7	Numeric	5	3	x7 - Freedom	None	None	8	Right	Scale	Input
x8	Numeric	6	3	x8 - Generosity	None	None	8	Right	Scale	Input
x9	Numeric	5	3	x9 - Corruption	None	None	8	Right	Scale	Input
x10	Numeric	5	3	x10 - Positive Affect	None	None	8	Right	Scale	Input
x11	Numeric	5	3	x11 - Negative Affect	None	None	8	Right	Scale	Input

Figure 1. Variable View of the World Happiness Report Dataset on SPSS

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11
1	Afghanistan	2016	4.220	7.697	.559	53.000	.523	.042	.793	.565	.348
2	Afghanistan	2017	2.662	7.697	.491	52.800	.427	-.121	.954	.496	.371
3	Afghanistan	2018	2.694	7.692	.508	52.600	.374	-.094	.928	.424	.405
4	Afghanistan	2019	2.375	7.697	.420	52.400	.394	-.108	.924	.351	.502
5	Albania	2016	4.511	9.437	.638	68.100	.730	-.017	.901	.675	.322
6	Albania	2017	4.640	9.476	.638	68.400	.750	-.029	.876	.669	.334
7	Albania	2018	5.004	9.518	.684	68.700	.824	.009	.899	.713	.319
8	Albania	2019	4.995	9.544	.686	69.000	.777	-.099	.914	.681	.274
9	Albania	2020	5.365	9.497	.710	69.300	.754	.007	.891	.679	.265
10	Algeria	2017	5.249	9.354	.807	65.700	.437	-.167	.700	.642	.289

Figure 2. Data View of the World Happiness Report Dataset on SPSS

To perform multiple regression analysis, the metric continuous variables are extracted to be independent variables while the dependent variable is Happiness Score, x_3. The variables used for analysis are shown in Table 2.

Dependent Variable, y	**Independent Variables**
X_3 – Happiness Score	X_4 – Logged GDP Per Capita
	X_5 – Social Support
	X_6 – Life Expectancy
	X_7 – Freedom
	X_8 – Generosity
	X_9 – Corruption
	X_{10} – Positive Affect
	X_{11} – Negative Affect

Table 2. Variables used for Multiple Regression Analysis

(ii) Stepwise Multiple Linear Regression

Preliminary Analysis – Correlation Analysis and Linearity

Invoking a 1:20 ratio (Hair Jr et al., 2009), the dataset is validated to have sufficient observations for model building. Data adequacy is further examined through correlation analysis and linearity of variables.

The output of a bivariate Pearson correlation analysis within SPSS in Figure 3 shows that 7 out of 8 of the variables have a correlation of greater than the absolute value threshold of 0.3 at a statistically significant level of 0.01. As majority of variables are correlated with the dependent variable, it is motivated to proceed with analysis of linearity of variables.

Correlations

		x3 - Happiness Score	x4 - Logged GDP per Capita	x5 - Social Support	x6 - Life Expectancy	x7 - Freedom	x8 - Generosity	x9 - Corruption	x10 - Positive Affect	x11 - Negative Affect
x3 - Happiness Score	Pearson Correlation	1	.804**	.757**	.780**	.524**	.048	-.456**	.460**	-.496**
	Sig. (2-tailed)		.000	.000	.000	.000	.249	.000	.000	.000
	N	589	589	589	589	589	589	589	589	589

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Figure 3.Bivariate Pearson Correlation Analysis between dependent variable and independent variables

Amongst the five important assumptions for a Multiple Linear Regression analysis (linearity, minimal multicollinearity, normality, homoscedasticity, and independence), linearity will first be tested through a pairwise scatter plot for all 9 variables. The pairwise plot in Figure 4 shows no alarming violations of linearity within the variables and the linearity assumption is satisfied. This dataset is deemed adequate to proceed with the Multiple Linear Regression analysis while the remaining four assumptions will be tested after the regression model is built.

Figure 4. Pairwise Plot

Stepwise Multiple Linear Regression Model Building

Stepwise Regression Analysis: Model 1

Through referring to Figure 3, it is deduced that x_4, logged GDP per capita is the first variable to be entered into the regression model as it has the highest zero-order Pearson correlation, R value of 0.804. Entering this variable produces the first regression model, Model 1, described in Figure 5 and 6.

Model Summary[h]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.804[a]	.646	.646	.666266	.646	1073.273	1	587	.000

a. Predictors: (Constant), x4 - Logged GDP per Capita

h. Dependent Variable: x3 - Happiness Score

Figure 5. Model Summary of Step 1 Results of Stepwise Multiple Linear Regression

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-1.832	.227		-8.073	.000					
	x4 - Logged GDP per Capita	.782	.024	.804	32.761	.000	.804	.804	.804	1.000	1.000

a. Dependent Variable: x3 - Happiness Score

Figure 6. Coefficients of Step 1 Results of Stepwise Multiple Linear Regression

Figure 5 shows the Model Summary which describes the fitness of Model 1 for the data. The R^2 value of 0.646 implies that x_4, logged GDP per capita explains 64.6% of variation in the target variable, x_3, Happiness Score.

Following this, Figure 6 describes the relationships between independent and dependent variables in Model 1. The unstandardized and standardized coefficients quantifies the change in dependent variable affected by the independent variable. They differ such that the unstandardized coefficient represents the raw unit scale of independent variable whereas standardized coefficients are the normalized unit-less coefficients. Through extracting b_0, -1.832 and b_4, 0.782, Model 1 is expressed as:

$$y = x_3 = -1.832 + 0.782x_4$$

This implies that for each unit increase in x_4 - Logged GDP per Capita, there is a 0.782 increase in x_3 - Happiness Score. As significance value shown is less than 0.05, it is concluded that b_4 is significant at a 95% confidence interval.

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	x5 - Social Support	.330[b]	8.927	.000	.346	.389	2.572	.389
	x6 - Life Expectancy	.325[b]	6.620	.000	.264	.233	4.288	.233
	x7 - Freedom	.276[b]	11.712	.000	.436	.878	1.139	.878
	x8 - Generosity	.158[b]	6.594	.000	.263	.982	1.018	.982
	x9 - Corruption	-.188[b]	-7.440	.000	-.294	.867	1.154	.867
	x10 - Positive Affect	.264[b]	11.462	.000	.428	.929	1.077	.929
	x11 - Negative Affect	-.071[b]	-2.404	.017	-.099	.691	1.447	.691

a. Dependent Variable: x3 - Happiness Score

b. Predictors in the Model: (Constant), x4 - Logged GDP per Capita

Figure 7. Excluded Variables from Model 1

As observed from Figure 7, x_7 – Freedom has the highest significant partial correlation of 0.436. Thus, x_7 is entered into the model to generate Model 2. This process is iterated to achieve the model of best fit for the data.

(ii). Model Results Analysis (Model 2 to Model 7)

Figure 8 shows that 7 models (Model 1 to Model 7) have been generated, with one additional predictor variable for each successive model, to achieve a linear regression model of best fit. The predictor variables are selected through evaluating the partial correlation of all remaining variables after each model.

The variables chosen to be entered into each model is observed in Figure 9. 7 out of 8 variables were entered into the regression model whereby the variable excluded is observed from Figure 8 to be x_8 – Generosity. Through invoking the parsimony principle, x_8 is excluded as the significance is greater than 0.05, it does not significantly improve the model performance.

Each variable added into the model improved the model performance as reflected by the increasing R^2 and adjusted R^2 values. Figure 8 shows the decreasing increment of the R^2 value while approaching the line of best fit. Fit of regression model improves as adjusted R^2 value improves from 64.6% of Model 1 to 77.3% of Model 7. Thus, it can be concluded that the 7 variables included into Model 7 successfully explains 77.3% of the variation in the dependent variable, x_3 – Happiness Score. Furthermore, the 0.3% difference in R^2 value and the adjusted R^2 value is small in showing that there is no overfitting and the model can be generalized to the population data.

Model Summary[h]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.804[a]	.646	.646	.666266	.646	1073.273	1	587	.000
2	.845[b]	.714	.713	.600267	.067	137.175	1	586	.000
3	.860[c]	.739	.738	.573339	.026	57.338	1	585	.000
4	.866[d]	.750	.748	.561716	.011	25.462	1	584	.000
5	.871[e]	.759	.757	.551705	.009	22.385	1	583	.000
6	.877[f]	.768	.766	.541683	.009	22.774	1	582	.000
7	.881[g]	.776	.773	.533491	.007	19.011	1	581	.000

a. Predictors: (Constant), x4 - Logged GDP per Capita

b. Predictors: (Constant), x4 - Logged GDP per Capita, x7 - Freedom

c. Predictors: (Constant), x4 - Logged GDP per Capita, x7 - Freedom, x5 - Social Support

d. Predictors: (Constant), x4 - Logged GDP per Capita, x7 - Freedom, x5 - Social Support, x9 - Corruption

e. Predictors: (Constant), x4 - Logged GDP per Capita, x7 - Freedom, x5 - Social Support, x9 - Corruption, x6 - Life Expectancy

f. Predictors: (Constant), x4 - Logged GDP per Capita, x7 - Freedom, x5 - Social Support, x9 - Corruption, x6 - Life Expectancy, x10 - Positive Affect

g. Predictors: (Constant), x4 - Logged GDP per Capita, x7 - Freedom, x5 - Social Support, x9 - Corruption, x6 - Life Expectancy, x10 - Positive Affect, x11 - Negative Affect

h. Dependent Variable: x3 - Happiness Score

Figure 8. Model Summary of Model 1 to Model 7

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	x4 - Logged GDP per Capita	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	x7 - Freedom	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
3	x5 - Social Support	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
4	x9 - Corruption	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
5	x6 - Life Expectancy	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
6	x10 - Positive Affect	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
7	x11 - Negative Affect	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: x3 - Happiness Score

Figure 9. Variables Entered/Removed for Stepwise Multiple Regression Model

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
7	x8 - Generosity	.034[h]	1.544	.123	.064	.807	1.239	.173

a. Dependent Variable: x3 - Happiness Score

h. Predictors in the Model: (Constant), x4 - Logged GDP per Capita, x7 - Freedom, x5 - Social Support, x9 - Corruption, x6 - Life Expectancy, x10 - Positive Affect, x11 - Negative Affect

Figure 10. Excluded Variable in Model 7

The standardized coefficients in Figure 11 are compared to identify that x_4 – Logged GDP per capita has the highest effect on happiness score as it has the largest standardized coefficient of 0.332. Through referring to the unstandardised coefficients, the expression for Model 7 is:

$$y = x_3 = -3.574 + 0.323x_4 + 0.967x_7 + 2.701x_5 - 0.797x_9 + 0.033x_6 + 1.529x_{10} + 1.514x_{11}$$

or

$$\begin{aligned} Happiness\ Score \\ &= -3.574 + 0.323(Logged\ GDP\ per\ Capita) + 0.967(Freedom) \\ &+ 2.701(Social\ Support) - 0.797(Corruption) + 0.033(Life\ Expectancy) \\ &+ 1.529(Positive\ Affect) + 1.514(Negative\ Affect) \end{aligned}$$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
7	(Constant)	-3.574	.389		-9.193	.000					
	x4 - Logged GDP per Capita	.323	.045	.332	7.131	.000	.804	.284	.140	.178	5.605
	x7 - Freedom	.967	.264	.103	3.664	.000	.524	.150	.072	.490	2.040
	x5 - Social Support	2.701	.330	.295	8.184	.000	.757	.322	.161	.298	3.355
	x9 - Corruption	-.797	.141	-.135	-5.663	.000	-.456	-.229	-.111	.683	1.463
	x6 - Life Expectancy	.033	.007	.210	5.048	.000	.780	.205	.099	.223	4.494
	x10 - Positive Affect	1.529	.288	.140	5.314	.000	.460	.215	.104	.557	1.795
	x11 - Negative Affect	1.514	.347	.116	4.360	.000	-.496	.178	.086	.543	1.841

a. Dependent Variable: x3 - Happiness Score

Figure 11. Coefficients Table for Model 7

(i*v*) Stepwise Multiple Linear Regression Assumptions

The 5 main assumptions associated with a regression model are linearity, minimal multicollinearity, normality, homoscedasticity, and independence. The linearity assumption has been explored during preliminary analysis where minimal violations was identified through Figure 4. As the final model, Model 7 is built, an assessment of the remaining assumptions are performed.

Minimal Multicollinearity

Multicollinearity refers to moderate to high intercorrelations amongst the predictor variables. Multicollinearity severely limits the model performance through the limitation of correlation to dependent variable (Dizney & Gromen, 1967) . The VIF for all predictor variables are found to below the threshold of 10 (Vittinghoff et al., 2005). Thus, it is concluded that the minimal multicollinearity assumption is satisfied in the final model.

Normality of Error Terms

The normality assumption requires the error terms to be normally distributed. This is visualized through the P-P and Q-Q plots in Figure 12 and 13, and further tested through a normality test. The P-P and Q-Q plots shows that the residuals do not divert from the normal distribution threshold by a great degree.

Figure 12. Normal P-P plot of Standardized Residual

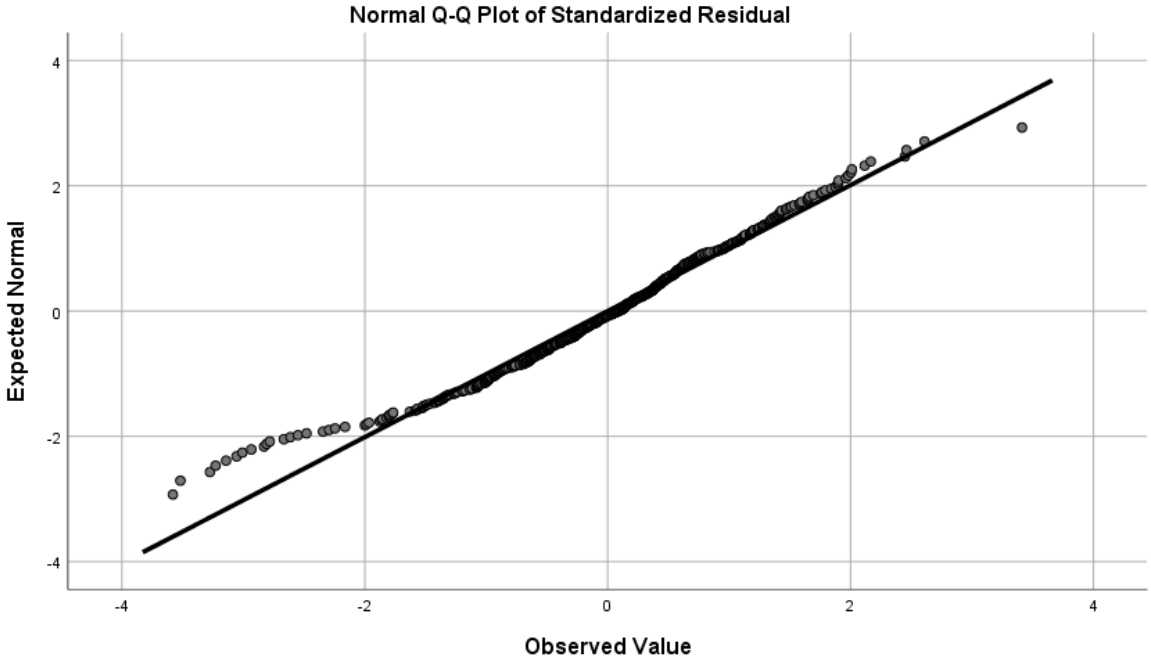

Figure 13. Normal Q-Q Plot of Standardized Residual

The normality test is used to clarify the normality of residuals with the following hypotheses:

H_0: The error terms are normally distributed.
H_1: The error terms are not normally distributed.

Through the output of Normality Test from SPSS shown in Figure 14, the Kolmogorov-Smirnov and Shapiro-Wilk tests shows that the significance values is less than 0.05. Thus, H_0 is rejected and it can be concluded that the error terms are not normally distributed and the normality assumption is violated.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.049	589	.002	.980	589	.000

a. Lilliefors Significance Correction

Figure 14. Normality Test Output from SPSS

Homoscedasticity of Error Terms

The homoscedasticity assumption expresses that variance for error terms are similar across all values of independent variables. This is observed through a plot of standardized residual values and predicted values of dependent variable in Figure 15. The plot shows that a dissimilar range of variation which is a violation of the homoscedasticity assumption

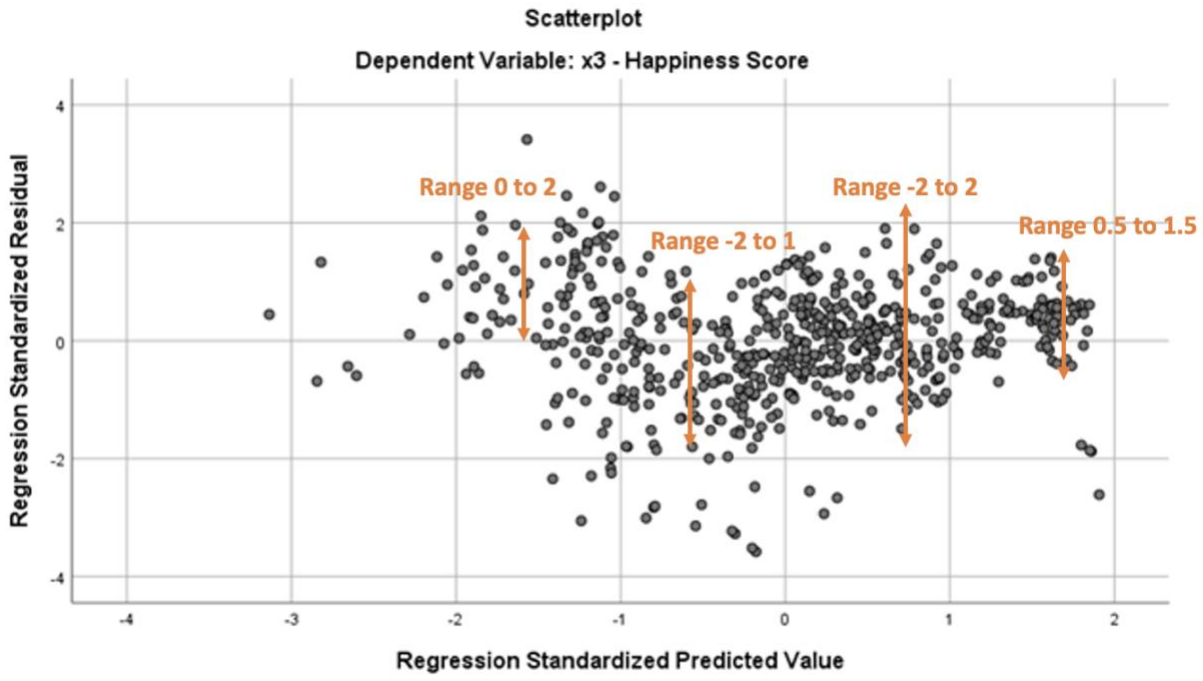

Figure 15. Residual Plot with marked variation range of error terms

Independence of Independent Variables

The independence of independent variables are determined through the Durbin Watson value of 0.806 as shown in Figure 16. At a 95% confidence level, the lower threshold and upper threshold for the Durbin Watson significance would be 1.697 and 1.841 respectively. As the Durbin Watson value of 0.806 lies below the lower threshold, there is positive autocorrelation between the independent variables and the independence assumption is violated.

Model Summary[h]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
7	.881[g]	.776	.773	.533491	.007	19.011	1	581	.000	.806

a. Predictors: (Constant), x4 - Logged GDP per Capita

g. Predictors: (Constant), x4 - Logged GDP per Capita, x7 - Freedom, x5 - Social Support, x9 - Corruption, x6 - Life Expectancy, x10 - Positive Affect, x11 - Negative Affect

h. Dependent Variable: x3 - Happiness Score

Figure 16. Model Summary with Durbin-Watson value

(v) Significance of Regression Model

The significance of Model 7 is tested through performing ANOVA testing with the following hypotheses:

H_0: Model 7 is not significant in predicting the target, i.e. all regression coefficients are equal to zero.

H_1: Model 7 is significant in predicting the target, i.e. at least one regression coefficient is non-zero.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
7	Regression	571.652	7	81.665	286.933	.000[b]
	Residual	165.360	581	.285		
	Total	737.011	588			

a. Dependent Variable: x3 - Happiness Score

h. Predictors: (Constant), x4 - Logged GDP per Capita, x7 - Freedom, x5 - Social Support, x9 - Corruption, x6 - Life Expectancy, x10 - Positive Affect, x11 - Negative Affect

Figure 17. ANOVA test output for Model 7

Through the output of ANOVA testing in Figure 11, an F statistics of 286.993 is observed with a p-value which lies below the 0.05 threshold. Thus, H_0 is rejected and it is concluded that the model is significant with a confidence interval of 95%.

(vi) Hypothesis Testing for Coefficients of Regression

The Student's t-test is implemented where the significance of t statistic is tested for each variable. The t-statistic is calculated through the formula, $t = \frac{\widehat{\beta}_i}{SE_{\beta_i}}$, while the significance represents the probability of obtaining errors.

To implement the t-test on x_4 coefficient, the following hypotheses are proposed:

H_0: There is no linear relationship between x_4 and the dependent variable (i.e $b_4 = 0$)

H_1: There is a relationship between x_4 and the dependent variable (i.e $b_4 \neq 0$)

Referring to Figure 18, it is seen that the p-value lies below the significance level of 0.05. Thus, H_0 is rejected and coefficient, x_4 is significant. Testing is repeated for all coefficients and constant to find that all the alternative hypothesis for all coefficients are accepted. As the sample size of data is larger, there are more training data for decrease in error. Hence, it is reasonable to obtain a low significance whereby it is concluded that all coefficients in Model 7 is significant.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
7	(Constant)	-3.574	.389		-9.193	.000					
	x4 - Logged GDP per Capita	.323	.045	.332	7.131	.000	.804	.284	.140	.178	5.605
	x7 - Freedom	.967	.264	.103	3.664	.000	.524	.150	.072	.490	2.040
	x5 - Social Support	2.701	.330	.295	8.184	.000	.757	.322	.161	.298	3.355
	x9 - Corruption	-.797	.141	-.135	-5.663	.000	-.456	-.229	-.111	.683	1.463
	x6 - Life Expectancy	.033	.007	.210	5.048	.000	.780	.205	.099	.223	4.494
	x10 - Positive Affect	1.529	.288	.140	5.314	.000	.460	.215	.104	.557	1.795
	x11 - Negative Affect	1.514	.347	.116	4.360	.000	-.496	.178	.086	.543	1.841

a. Dependent Variable: x3 - Happiness Score

Figure 18. Coefficients Table for Student's T-Testing

(e) Conclusions & Recommendations

The stepwise approach used to build the multiple linear regression model ensures accordance to the parsimony principle. The final model built has good fit and is statistically significant with the ability to explain 77.3% of the variation in Happiness Score through incorporating seven variables.

The study of relationships between independent variables and happiness score have found that the null hypothesis of all three objectives defined in Section (c) is accepted. The variables pertaining to all three objectives have been implemented for happiness score prediction while Logged GDP per Capita has the highest directly proportional relationship to Happiness Score. This information can be used to assist organisations and government bodies to increase population's happiness level.

Additionally, it was found that the Generosity variable did not contribute towards the modelling of Happiness Score. This leads to the recommendation for UN Sustainable Development Solutions Network to replace this variable with an alternative well-being measurement.

Post-model building, the testing of assumptions concluded that Model 7 is in violation of the normality, homoscedasticity, and independent assumptions. These violations imply that a Multiple Linear Regression is not sufficient for the prediction of Happiness Score. The prediction of Happiness Score may be more suitable through a Generalized Linear Model such as Poisson Regression and Logistic Regression.

Question 2

(a) Purpose of Factor Analysis

Factor Analysis refers to the process of reducing data dimensionality through identifying common relationships between a group of variables in a dataset. It is an interdependence technique which assists researchers to identify underlying constructs through main sources of variation within complex correlations of data.

In this study, factor analysis is implemented to reduce the number of metric independent variables within the World Happiness Report dataset. The reduced variable enables progress towards the declared problem statement of identifying and analysing the underlying factors affecting happiness.

(b) Exclusion of Non-Metric Independent Variable

Non-metric independent variables are problematic for standard factor analysis as it does not carry a meaningful scale to satisfy the linearity assumption. The linearity assumption expresses a linear relationship between the factors formed and observed variables. To satisfy this assumption, input variables are required to be scaled values with approximately equal intervals. This requirement carries meaning to the assumption such that a change in unit of factor reflects to a change in unit of observed variable with approximate equal intervals.

(c) Factor Analysis on SPSS

Prior to the implementation of factor analysis, preliminary analysis is performed to determine the factorability of variables within the dataset. First, the overall Measures of Sampling Adequacy (MSA) is determined and compared to a threshold of 0.5. From Figure 19, it is observed that the overall MSA, 0.779 is greater than 0.5 which indicates factorability. Following this, the Bartlett's Test of Sphericity is performed through the following hypotheses:

H_0 : Correlation matrix is equivalent to an identity matrix (i.e variables cannot be factorized)
H_1 : Correlation matrix is not equivalent to an identity matrix (i.e variables can be factorized)

The significance value of the Bartlett's Test observed in Figure 19 is less than 0.05. Thus, H_0 is rejected and it is concluded that there is existing correlation between the variables and variables can be factorized at a 95% significance level.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.779
Bartlett's Test of Sphericity	Approx. Chi-Square	2538.072
	df	28
	Sig.	.000

Figure 19. Overall MSA and Bartlett's Test Results

Following this, the individual MSA of each variable is observed through Figure 20 to show that the individual MSA of all variables are greater than 0.5. Thus, factorability is achieved.

Anti-image Matrices

		x4 - Logged GDP per Capita	x5 - Social Support	x6 - Life Expectancy	x7 - Freedom	x8 - Generosity	x9 - Corruption	x10 - Positive Affect	x11 - Negative Affect
Anti-image Covariance	x4 - Logged GDP per Capita	.173	-.092	-.135	.018	.065	.063	.007	.021
	x5 - Social Support	-.092	.296	-.029	-.020	-.036	-.100	-.061	.148
	x6 - Life Expectancy	-.135	-.029	.222	-.048	.002	.018	.023	-.015
	x7 - Freedom	.018	-.020	-.048	.487	-.051	.160	-.265	-.002
	x8 - Generosity	.065	-.036	.002	-.051	.807	.185	-.096	-.034
	x9 - Corruption	.063	-.100	.018	.160	.185	.638	-.011	-.116
	x10 - Positive Affect	.007	-.061	.023	-.265	-.096	-.011	.545	.066
	x11 - Negative Affect	.021	.148	-.015	-.002	-.034	-.116	.066	.542
Anti-image Correlation	x4 - Logged GDP per Capita	.740[a]	-.408	-.688	.061	.173	.189	.022	.068
	x5 - Social Support	-.408	.825[a]	-.114	-.053	-.073	-.231	-.153	.370
	x6 - Life Expectancy	-.688	-.114	.785[a]	-.146	.006	.048	.067	-.043
	x7 - Freedom	.061	-.053	-.146	.764[a]	-.082	.286	-.514	-.004
	x8 - Generosity	.173	-.073	.006	-.082	.626[a]	.258	-.145	-.051
	x9 - Corruption	.189	-.231	.048	.286	.258	.750[a]	-.019	-.198
	x10 - Positive Affect	.022	-.153	.067	-.514	-.145	-.019	.750[a]	.121
	x11 - Negative Affect	.068	.370	-.043	-.004	-.051	-.198	.121	.871[a]

a. Measures of Sampling Adequacy(MSA)

Figure 20. Factor Analysis SPSS Output - Anti Image Matrices

(i) Communality

After factorability is achieved, factor analysis output obtained through SPSS are analysed. The communalities table as shown in Figure 21 indicates the variance of the observed variable explained by the extracted factor. For instance, the extracted factor accounts for 87.2% of the variance in x_4 – Logged GDP per Capita and 78.3% of x_5 – Social Support. As the communality for x_9 – Corruption is below 50%, variable x_9 is removed from further steps of factor analysis.

Communalities

	Initial	Extraction
x4 - Logged GDP per Capita	1.000	.872
x5 - Social Support	1.000	.783
x6 - Life Expectancy	1.000	.805
x7 - Freedom	1.000	.686
x8 - Generosity	1.000	.591
x9 - Corruption	1.000	.479
x10 - Positive Affect	1.000	.606
x11 - Negative Affect	1.000	.560

Extraction Method: Principal Component Analysis.

Figure 21. Communalities Table of Factor Analysis from SPSS

(ii) Improving factorability

In addition to the KMO and Barlett's test, factorability is improved through the removal of variable x_9 where communality falls below 50% as observed in Figure 22. Furthermore, scaling and non-linear transformations can be implemented. The variables that do not follow Gaussian distributions can be scaled through a min-max transformation technique to maximise linearity within the data.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.783
Bartlett's Test of Sphericity	Approx. Chi-Square	2276.641
	df	21
	Sig.	.000

Figure 22. KMO and Bartlett's Test Output after improving factorability

(iii) Eigenvalue

The Total Variance Explained table in Figure 23 is used to determine the number of factors selected through eigenvalues. The eigenvalue reflects the amount of variance of observed variables by a factor. The number of factors chosen are traditionally compared to a minimum total eigenvalue threshold of 1 to ensure that the factor explains more variance than a single observed variable. Thus, two components will be used to represent the seven variables. These two components cumulatively account for 71.763% of the variance in the data variables. This results aligns with the Scree Plot where the curve flattens after factor 3 where eigenvalue is less than 1. Thus, only two factors are retained.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.543	50.613	50.613	3.543	50.613	50.613	3.259	46.561	46.561
2	1.481	21.150	71.763	1.481	21.150	71.763	1.764	25.202	71.763
3	.690	9.863	81.626						
4	.571	8.157	89.783						
5	.361	5.153	94.936						
6	.241	3.445	98.380						
7	.113	1.620	100.000						

Extraction Method: Principal Component Analysis.

Figure 23. Total Variance Explained Table of Factor Analysis from SPSS

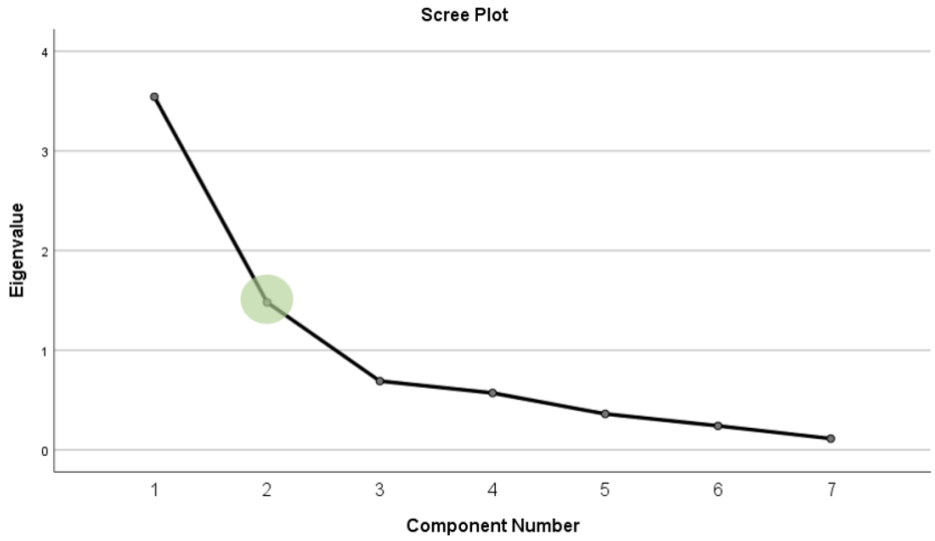

Figure 24. Scree Plot to determined number of components

(iv) Factor cross-loading

The component matrix shown in Figure 25 tabulates the Pearson correlation, known as factor loading of each variable within the extracted components. Factor cross-loading refers to the phenomenon whereby an observed variable is significantly explained by more than one component and the difference in loading value is less than 0.4. Cross-loading is problematic as the overlap in the extracted components prevents the components from representing distinct concepts. In Figure 25, it is observed that cross-loading exists within variables x_7 and x_{10}.

Component Matrix[a]

	Component	
	1	2
x5 - Social Support	.876	
x4 - Logged GDP per Capita	.871	
x6 - Life Expectancy	.849	
x11 - Negative Affect	-.748	
x7 - Freedom	.632	.547
x8 - Generosity		.762
x10 - Positive Affect	.579	.610

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Figure 25. Component Matrix

(v) Recommendations for cross-loading

To eliminate factor cross-loading, Varimax rotation is implemented to redistribute the factor loading as shown in Figure 26. It is seen that cross-loading is eliminated and grouping of variables into the extracted components can be performed.

Rotated Component Matrix[a]

	Component	
	1	2
x4 - Logged GDP per Capita	.933	
x6 - Life Expectancy	.896	
x5 - Social Support	.879	
x11 - Negative Affect	-.710	
x10 - Positive Affect		.781
x7 - Freedom		.742
x8 - Generosity		.719

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Figure 26. Rotated Component Matrix

(vi) Grouping of Metric Independent Variables

After elimination of cross-loading, the grouping of variables into components are performed according to Table X where the seven variables are grouped into two components which represents approximately 71.8% of the observed variables.

Component	**Variable**
1	x_3 – Logged GDP per Capita
	x_6 – Life Expectancy
	x_5 – Social Support
	x_{11} – Negative Affect
2	x_{10} – Positive Affect
	x_7 – Freedom
	x_8 – Generosity

Figure 27. Grouping of Metric Independent Variables

References

Dizney, H. F., & Gromen, L. (1967). Predictive Validity and Differential Achievement on Three Mla—Cooperative Foreign Language Tests. *Educational and Psychological Measurement*, *27*(4), 1127–1130. https://doi.org/10.1177/001316446702700465

Easterlin, R. A., McVey, L. A., Switek, M., Sawangfa, O., & Zweig, J. S. (2010). The happiness-income paradox revisited. *Proceedings of the National Academy of Sciences*, *107*(52), 22463–22468. https://doi.org/10.1073/pnas.1015962107

Fanning, A. L., & O'Neill, D. W. (2019). The Wellbeing–Consumption paradox: Happiness, health, income, and carbon emissions in growing versus non-growing economies. *Journal of Cleaner Production*, *212*, 810–821. https://doi.org/10.1016/j.jclepro.2018.11.223

Hair Jr, J. F., Black, W. C., Bajin, B. J., & Anderson, R. E. (2009). *Multivariate Data Analysis*. Prentice Hall.

Helliwell, J., Huang, H., Wang, S., De Neve, J.-E., Diener, E., Eaton, C., Exton, C., Fritjers, P., Gilbert, D., Goff, L., Gra-Ham, C., Grover, S., Hall, J., Layard, R., Mayraz, G., Rothstein, B., & Wiking, M. (2017). *THE SOCIAL FOUNDATIONS OF WORLD HAPPINESS*. http://ss835667.stars.ne.jp/shoko/happiness.pdf

Singh, A. (2021, March 21). *World Happiness Report 2021*. Kaggle.com. https://www.kaggle.com/ajaypalsinghlo/world-happiness-report-2021

Vittinghoff, E., Shiboski, S. C., Glidden, D. V., & McCulloch, C. (2005). Regression Methods in Biostatistics. In *Statistics for Biology and Health*. Springer New York. https://doi.org/10.1007/b138825

World Happiness Report. (2019). *Home*. Worldhappiness.report. https://worldhappiness.report/