MATH 456
Final Project
World Happiness Score

By Jakub Dlugosz & Oguzcan Adabuk

Data Source: https://www.kaggle.com/unsdsn/world-happiness

I Introduction:

One of the many important questions regarding countries, is what makes a country perform better compared to the rest of the world? In order to answer that question, we simply have to look at the world happiness score, and rank these countries based on this score. In this article, "Understanding the determinants of happiness through Gallup World Poll", which is a study conducted by Vidushi Jaswal , Kamal Kishore , Muniraju M, Nidhi Jaswal , Rakesh Kapoor, tried to understand the determinants for the world happiness score. The first time this happiness score became known, was back in 2012, when the first world happiness report was conducted using the Gallup World Survey, measuring across 120 different countries that looked at factors such as GDP, Trust, Life Expectancy, etc. (4826). The researchers looked at these factors and concluded that Freedom and Family are the strongest predictors of happiness and that trust variables do not have a significant relationship with happiness (4826). For this project, we would like to conduct a similar study and try to to answer what factors contribute to a country's happiness score. For this process, we will be considering variables, such as GDP, Life Expectancy etc, and try to develop a multiple regression model to better predict a country's happiness score based on those types of variables.

II Data Description:

The dataset we will be working with, World Happiness Dataset, are world happiness scores from 156 different countries across the world. This dataset consists of 9 different variables:

Overall Rank: This is an ordinal variable that will show how the country is ranked in terms of happiness score compared to other countries.

Country or region: This variable shows the name of the country which the happiness score is measured in. There are 156 different countries for this column.

Score: This is a continuous variable that is an overall happiness score for a country, where 10 is the happiest while 0 is the unhappiest country.

GDP per capita: This scores the overall economic well being for a country

Social Support: This measure describes how well people are supported especially in families.

Healthy life expectancy: This is a continuous variable that measures the overall health expectancy for a person.

Freedom: This measures how much freedom people believe to have within their country.

Generosity: This is a measure of how generous people are to one another within their country.

Corruption: This is the measure on how much people trust in their government.

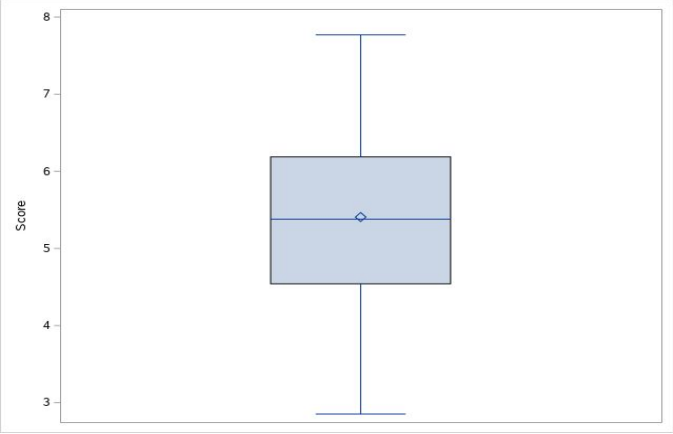
For this project, we will be considering all variables except for Rank and Country, since we feel that this would not impact our regression model.

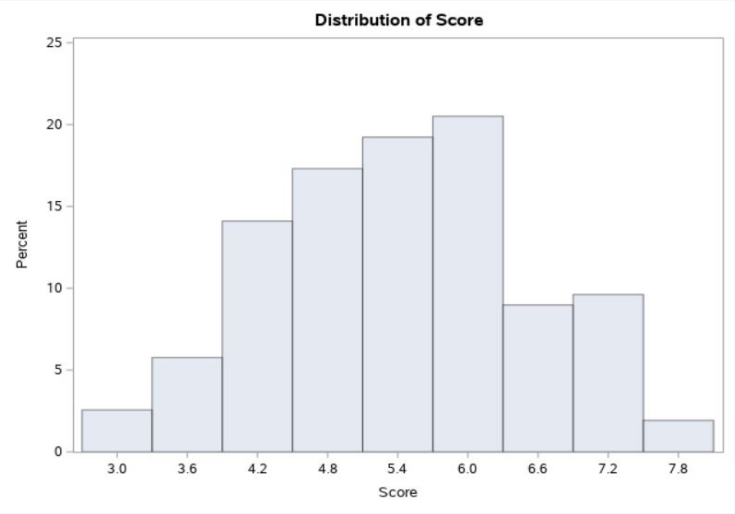
III Data Analysis:

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
Overall_rank	156	78.5000000	45.1774280	1.0000000	156.0000000
Score	156	5.4070962	1.1131199	2.8530000	7.7690000
GDP_per_capita	156	0.9051474	0.3983895	0	1.6840000
Social_support	156	1.2088141	0.2991914	0	1.6240000
Healthy_life_expectancy	156	0.7252436	0.2421240	0	1.1410000
Freedom_to_make_life_choices	156	0.3925705	0.1432895	0	0.6310000
Generosity	156	0.1848462	0.0952544	0	0.5660000
Perceptions_of_corruption	156	0.1106026	0.0945378	0	0.4530000

Above, we have some summary statistics for the various variables we are using for this analysis. One thing to note is that we have variables that only range from 0 to 10 score, and looking at the world happiness score, we have a mean score of around 5, while the rest of our variables have a mean score of roughly 1 or closer to 0. For this analysis we would like to look at the Overall score as our response variables and looking at the histogram, we can see that this looks approximately normal. We also test for normality using the Shapiro Wilk's test with a p-value of 0.1633, which we fail to reject H0 and conclude that our response variable is normally distributed.





Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.987201	Pr < W	0.1633
Kolmogorov-Smirnov	D	0.057802	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.061926	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.423097	Pr > A-Sq	>0.2500

Pearson Correlation Coefficients, N = 156
Prob > |r| under H0: Rho=0

	Score	GDP_per_capita	Social_support	Healthy_life_expectancy	Freedom_to_make_life_choices	Generosity	Perceptions_of_corruption
Score	1.00000	0.79388 <.0001	0.77706 <.0001	0.77988 <.0001	0.56674 <.0001	0.07582 0.3468	0.38561 <.0001
GDP_per_capita	0.79388 <.0001	1.00000	0.75491 <.0001	0.83546 <.0001	0.37908 <.0001	-0.07966 0.3229	0.29892 0.0002
Social_support	0.77706 <.0001	0.75491 <.0001	1.00000	0.71901 <.0001	0.44733 <.0001	-0.04813 0.5508	0.18190 0.0230
Healthy_life_expectancy	0.77988 <.0001	0.83546 <.0001	0.71901 <.0001	1.00000	0.39039 <.0001	-0.02951 0.7146	0.29528 0.0002
Freedom_to_make_life_choices	0.56674 <.0001	0.37908 <.0001	0.44733 <.0001	0.39039 <.0001	1.00000	0.26974 0.0007	0.43884 <.0001
Generosity	0.07582 0.3468	-0.07966 0.3229	-0.04813 0.5508	-0.02951 0.7146	0.26974 0.0007	1.00000	0.32654 <.0001
Perceptions_of_corruption	0.38561 <.0001	0.29892 0.0002	0.18190 0.0230	0.29528 0.0002	0.43884 <.0001	0.32654 <.0001	1.00000

Next, we want to construct and look at the correlation matrix to see how strongly related these variables are to one another. Looking at this correlation matrix, we have strong positive correlations with GDP per Capita, social support, health life expectancy, with mild correlations with freedom of choice and perceptions of corruption.

Partial F-Test for $B_5=B_6=0$

H_0:$B_5=B_6=0$ vs H_a: B_5 and B_6 are not both 0

Full Model: Y=B0 + B1X1 + B2X2 + B3X3 + B4X4 + B5X5 + B6X6

Reduced Model: Y=B0 + B1X1 + B2X2 + B3X3 + B4X4

SSR(F) = 149.63884, SSR(R) = 149.36323

Test Statistic: F* = ((SSR(F) - SSR(R))/2) / MSE(F) = (0.27654/2) / 0.28464 = 0.485771

F(1-0.05, 2, 150) =3.056366, F* < F, conclude H_0. Meaning that generosity and perception of corruption does not add significant information to the model.

IV Model Selection and Diagnostics:

We will first generate our full model:

$$Score_i = GDPX_1 + SocialX_2 + LifeX_3 + FreedomX_4 + GenerosityX_5 + CorruptionX_6 + \varepsilon_i .$$

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	149.63884	24.93981	87.62	<.0001
Error	149	42.41171	0.28464		
Lack of Fit	149	42.41171	0.28464	.	.
Pure Error	0	0	.		
Corrected Total	155	192.05056			

Root MSE	0.53352	R-Square	0.7792
Dependent Mean	5.40710	Adj R-Sq	0.7703
Coeff Var	9.86701		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	1.79522	0.21107	8.51	<.0001	1.37814	2.21230
GDP_per_capita	1	0.77537	0.21823	3.55	0.0005	0.34416	1.20659
Social_support	1	1.12419	0.23690	4.75	<.0001	0.65607	1.59231
Healthy_life_expectancy	1	1.07814	0.33454	3.22	0.0016	0.41709	1.73920
Freedom_to_make_life_choices	1	1.45483	0.37534	3.88	0.0002	0.71316	2.19650
Generosity	1	0.48978	0.49775	0.98	0.3267	-0.49377	1.47333
Perceptions_of_corruption	1	0.97228	0.54236	1.79	0.0751	-0.09943	2.04399

With our full model, we see that at alpha = 0.05 significance, that all variables were significant except for generosity and perception of corruption . However, we will want to run a stepwise regression in order to detect to see if we can eliminate variables to come up with a better model for this analysis.

Bounds on condition number: 3.9561, 45.579

Stepwise Selection: Step 5

Variable Perceptions_of_corruption Entered: R-Square = 0.7777 and C(p) = 5.9683

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	149.36323	29.87265	104.97	<.0001
Error	150	42.68732	0.28458		
Corrected Total	155	192.05056			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1.86887	0.19734	25.52447	89.69	<.0001
GDP_per_capita	0.74545	0.21607	3.38722	11.90	0.0007
Social_support	1.11803	0.23679	6.34425	22.29	<.0001
Healthy_life_expectancy	1.08402	0.33445	2.98963	10.51	0.0015
Freedom_to_make_life_choices	1.53401	0.36657	4.98360	17.51	<.0001
Perceptions_of_corruption	1.11755	0.52183	1.30525	4.59	0.0338

After running our stepwise regression algorithm, we find that the best model with the highest R-Squared value, and C(p) value less than p=7, will include these variables: GDP, Social Support, Health Expectancy, Freedom and Corruption.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	149.36323	29.87265	104.97	<.0001
Error	150	42.68732	0.28458		
Lack of Fit	150	42.68732	0.28458	.	.
Pure Error	0	0	.		
Corrected Total	155	192.05056			

Root MSE	0.53346	R-Square	0.7777
Dependent Mean	5.40710	Adj R-Sq	0.7703
Coeff Var	9.86597		

Parameter Estimates									
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	1	1.86887	0.19734	9.47	<.0001	.	0	1.47896	2.25879
GDP_per_capita	1	0.74545	0.21607	3.45	0.0007	0.24777	4.03594	0.31851	1.17239
Social_support	1	1.11803	0.23679	4.72	<.0001	0.36580	2.73374	0.65015	1.58591
Healthy_life_expectancy	1	1.08402	0.33445	3.24	0.0015	0.27999	3.57159	0.42317	1.74486
Freedom_to_make_life_choices	1	1.53401	0.36657	4.18	<.0001	0.66547	1.50270	0.80970	2.25832
Perceptions_of_corruption	1	1.11755	0.52183	2.14	0.0338	0.75442	1.32552	0.08648	2.14863

After running our stepwise regression, we now perform some diagnostics with the new model that excludes the variable, Generosity. For this new model, all variables are significant at alpha = 0.05 with an R-squared value of 0.777. If we compare it to the full model we see that our R-squared value remains roughly the same, so even though this did not improve significantly, it still allowed us to eliminate any unnecessary variables. We also test to see if we have multicollinearity within our model, and as we can see that for all of our variables, we have VIF values that are less than 10 and tolerance values greater than 0.1. Since the VIF values are less than 10 and tolerance values greater than 0.1, we can safely assume that we do not have multicollinearity for this fitted model.

Partial F-Test for $B_5=B_6=0$

H_0:$B_5=B_6=0$ vs H_a: B_5 and B_6 are not both 0

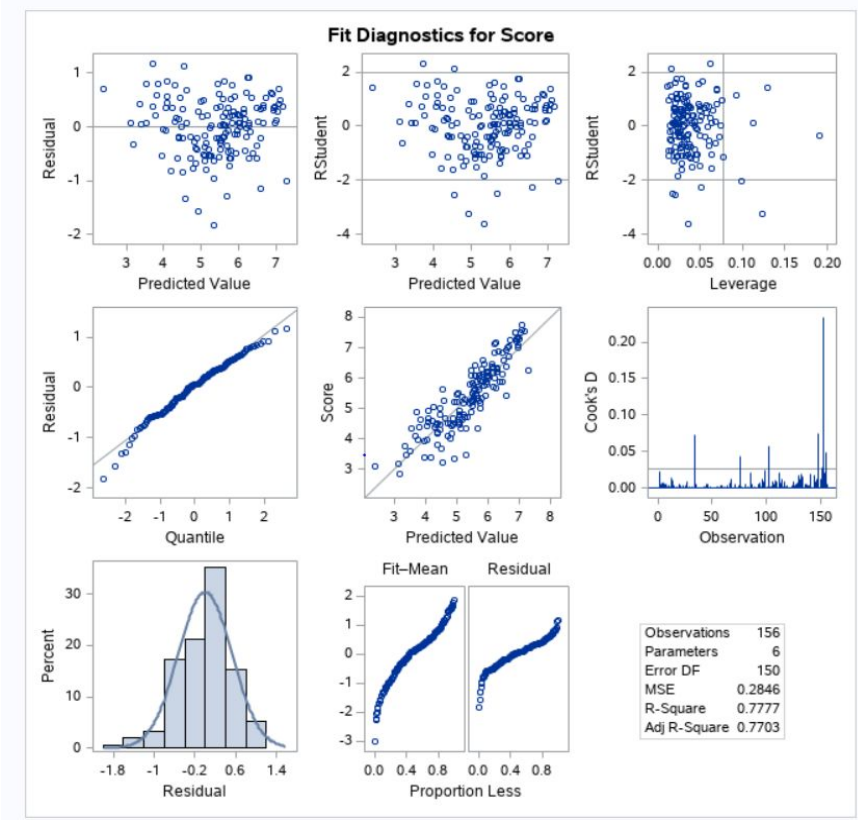
Full Model: Y=B0 + B1X1 + B2X2 + B3X3 + B4X4 + B5X5 + B6X6

Reduced Model: Y=B0 + B1X1 + B2X2 + B3X3 + B4X4

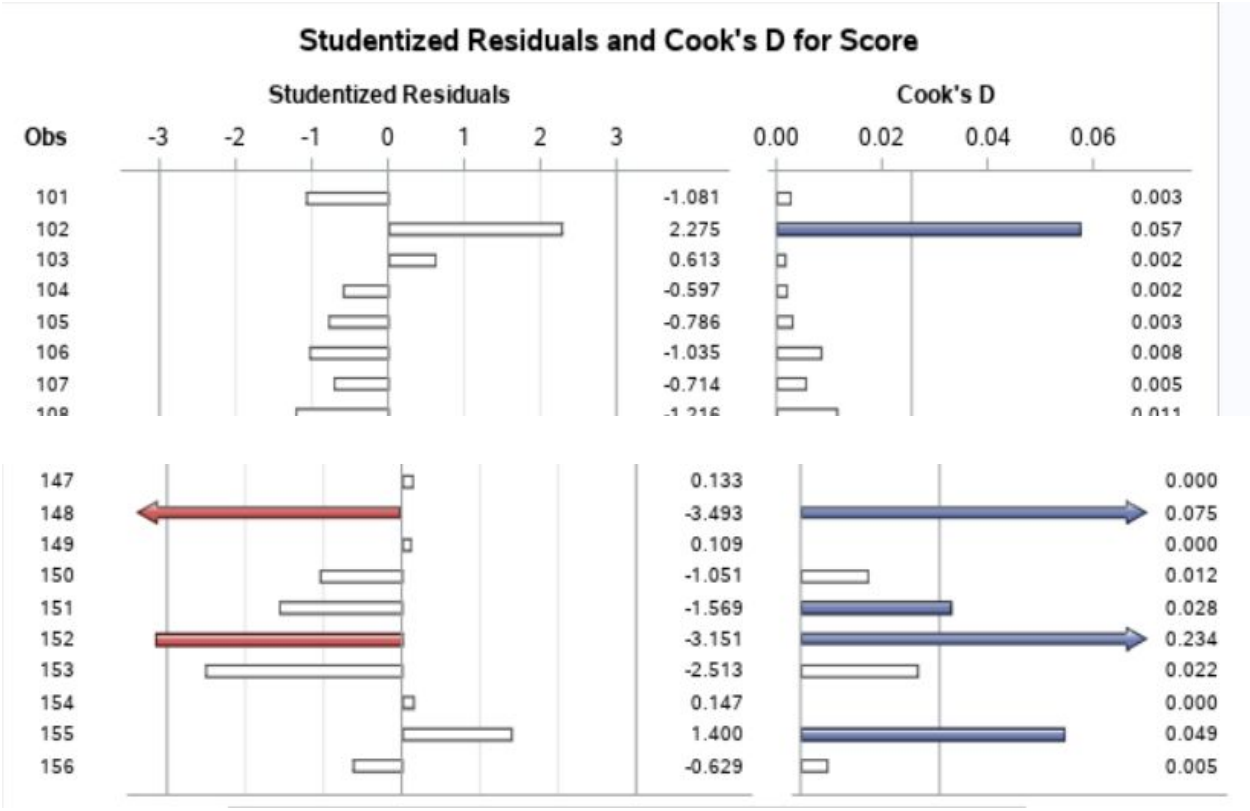
SSR(F) = 149.63884, SSR(R) = 149.36323

Test Statistic: F* = ((SSR(F) - SSR(R))/2) / MSE(F) = (0.27654/2) / 0.28464 = 0.485771

F(1-0.05, 2, 150) =3.056366, F* < F, conclude H_0. Meaning that generosity and perception of corruption does not add significant information to the model.



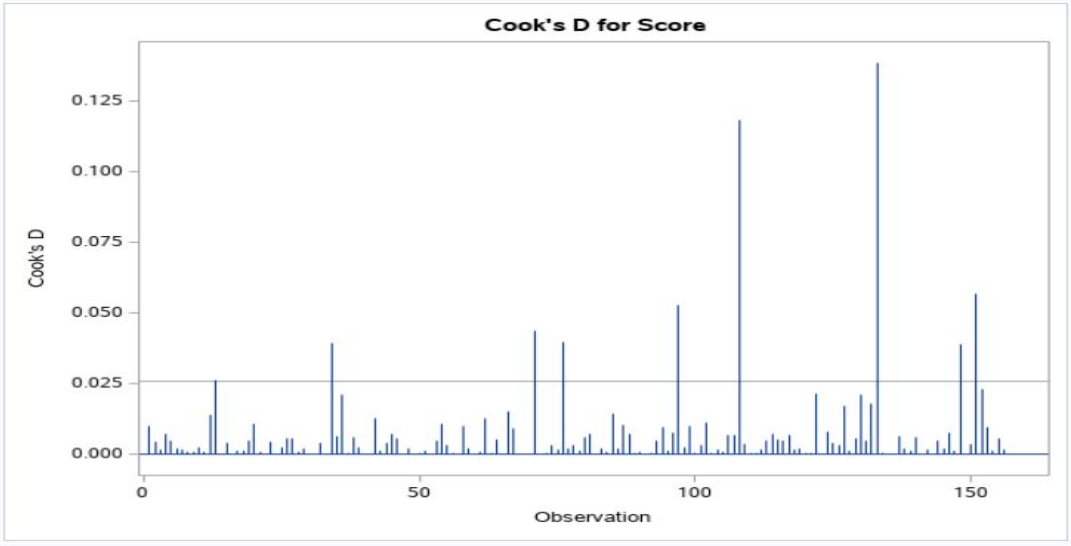
Looking at our diagnostics, we can see for our residuals plotted against Score, that there seems to be no visible pattern and the residuals are scattered, which indicates that these residuals do not violate the constant variance assumption. We then look at our qqplot to see if the residuals are normally distributed, and we can see that a majority of them fall on the line with some residuals trailing off at the tails, which can mean that we might have some observations that might be influential and should be worth investing and eliminating from the data set. However, looking at the normal distribution plot for our residuals, we can see that the residuals appear to approximately normal.



Looking at the Cook's distance plots, we can see that observations 102,148,151,152, and 155 are influential which can explain the outliers we see in our qqplot.

V Weighted Least Squares Estimation

After applying weighted least squares, we can see that the confidence intervals have slightly shrunk for some predictor variables but increased for some others. A possible reason for WLS not performing significantly better can be attributed to homoscedastic data. In other words the error variance is constant. Therefore the use of WLS cannot be justified. All predictors have similar p-values and no predictor needs to be discarded in this model. Fit diagnostics plots look mostly similar except the Cook's distance plot which shows several more outliers that have strong influence.



The REG Procedure
Model: MODEL1
Dependent Variable: Score

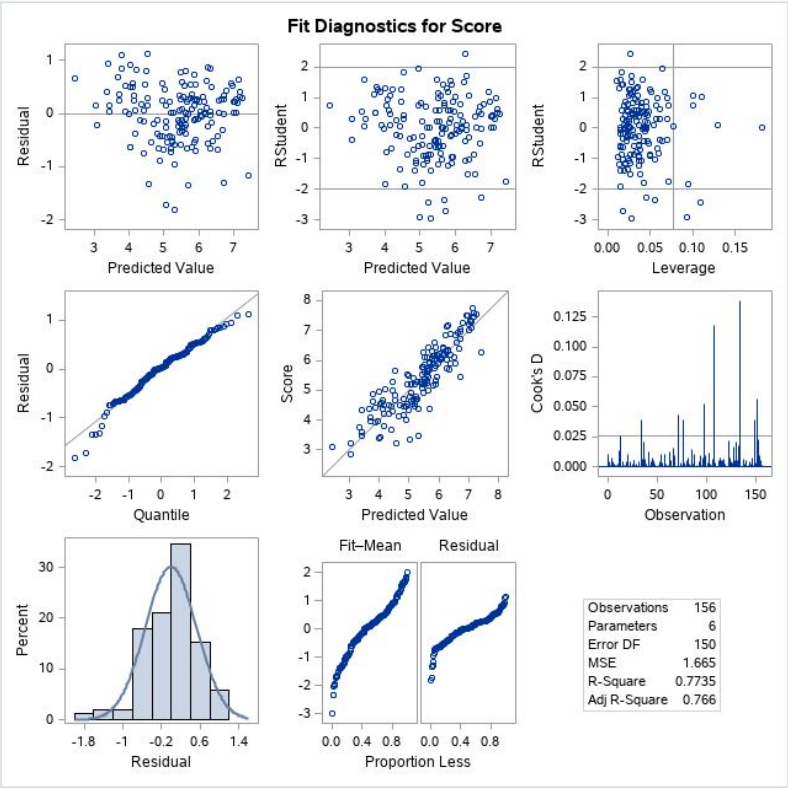
Number of Observations Read	156
Number of Observations Used	156

Weight: wt

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	853.09756	170.61951	102.47	<.0001
Error	150	249.75096	1.66501		
Corrected Total	155	1102.84851			

Root MSE	1.29035	R-Square	0.7735
Dependent Mean	5.47414	Adj R-Sq	0.7660
Coeff Var	23.57174		

Parameter Estimates								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation	95% Confidence Limits	
Intercept	1	1.80479	0.22369	8.07	<.0001	0	1.36281	2.24678
GDP_per_capita	1	0.79974	0.21662	3.69	0.0003	3.86376	0.37173	1.22776
Social_support	1	0.84504	0.23580	3.58	0.0005	2.25297	0.37913	1.31096
Healthy_life_expectancy	1	1.39095	0.32855	4.23	<.0001	3.29322	0.74175	2.04014
Freedom_to_make_life_choices	1	1.84818	0.29964	6.17	<.0001	1.38659	1.25611	2.44024
Perceptions_of_corruption	1	1.14954	0.54885	2.09	0.0379	1.30199	0.06506	2.23402



VI Conclusion

The score of happiness relies on GDP per Capita, Social Support, Healthy Life Expectancy, Freedom to make life choices and Perception of Corruption. The multiple regression model for happiness score with an R-squared value of 0.77 is:

SCORE = 1.86887 + 0.74545(GDP) + 1.11803(SOCIAL_SUPPORT) + 1.08402(HEALTHY_LIFE) + 1.53401(FREEDOM_OF_LIFE_CHOICES) + 1.11755(PERCEPTION_OF_CORRUPTIONS)

All predictors except PERCEPTION_OF_CORRUPTIONS are mild to strongly positively correlated with each other. PERCEPTION_OF_CORRUPTIONS is weakly correlated. On the other hand, GENEROCITY (not included in the final model) is negatively correlated with GDP per Capita, Social Support and Healthy life expectancy. We can conclude that the biggest

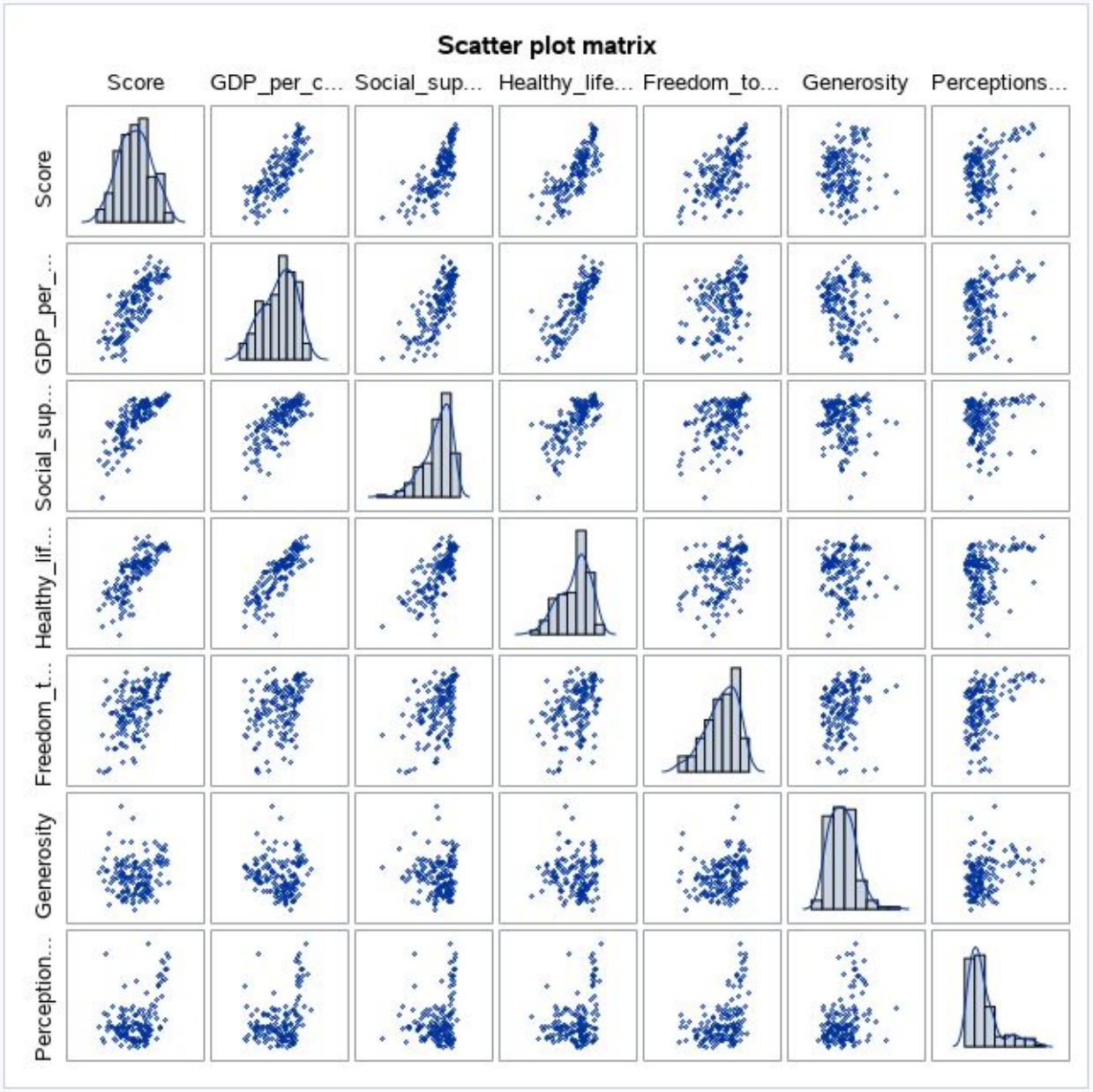
contributing factor to happiness score is Freedom to make life choices, followed by Healthy life expectancy, Perception of Corruptions, Social Support and the GDP per capita, in that order.

VII Future Work

1. Incorporate Important Predictors: Some other important predictors that can make a difference can be the climate, educational opportunities and personal time. Collecting data for other possible predictors can improve our model and provide a better insight to people's happiness measurement.
2. Different cultures focus on different priorities: It may make sense to conduct separate studies of happiness in different geographical areas to find alternative local models. The current model takes many countries from all around the world into account. Separate regression models can be conducted for Europe, Asia, North America and other regions.

Appendix

I.Scatter Plot Matrix


Scatter plot matrix
Score
GDP_per_c...
Social_sup...
Healthy_life...
Freedom_to...
Generosity
Perceptions...
Score
GDP_per_...
Social_sup...
Healthy_lif...
Freedom_t...
Generosity
Perception...

SAS Code:

```
/*this is how I imported the data, use the option New import Dataset*/
FILENAME REFFILE '/folders/myshortcuts/SASUniversityEdition/MATH 456 SAS/2019.csv';

PROC IMPORT DATAFILE=REFFILE
        DBMS=CSV
        OUT=WORK.IMPORT;
        GETNAMES=YES;
RUN;

/*Run simple statistics for all variables*/
proc means data = work.import;
Run;

/*Get Histograms for all variables, mainly looked at Score since that will be our Y variables*/
proc univariate data =work.import normal;
var Score
GDP_per_capita
Social_support
Healthy_life_expectancy
Freedom_to_make_life_choices
Generosity
Perceptions_of_corruption;
histogram Score
GDP_per_capita
Social_support
Healthy_life_expectancy
Freedom_to_make_life_choices
Generosity
Perceptions_of_corruption;
Run;

/*ran a boxplot for score to check for outliers*/
proc sgplot data=work.import ;
vbox Score;
Run;

/*here is a correlation matrix for variables*/
proc corr data=work.import plots=matrix(histogram);
var Score
GDP_per_capita
Social_support
```

```
Healthy_life_expectancy
Freedom_to_make_life_choices
Generosity
Perceptions_of_corruption;
run;

/*Here is a full model that is constructed*/
proc reg data=work.import;
model Score = GDP_per_capita Social_support Healthy_life_expectancy
Freedom_to_make_life_choices Generosity Perceptions_of_corruption/clb lackfit;
output out=a2 p=pred r=resid;

/*here is the stepwise regression selection*/
proc reg data=work.import;
model Score = GDP_per_capita Social_support Healthy_life_expectancy
Freedom_to_make_life_choices Generosity Perceptions_of_corruption /
selection = stepwise;
Run;

/*here is the new model after the stepwise selection, includes VIF, Cooks D for diagnostics*/
proc reg data=work.import  plots=COOKSD;
model Score = GDP_per_capita Social_support Healthy_life_expectancy
Freedom_to_make_life_choices  Perceptions_of_corruption  / vif clb lackfit
influence tol r partial ;

/*Possible box cox in case we need to transform our Y variable*/
proc transreg;
 model boxcox(Score) = identity(GDP_per_capita) identity(Social_support)
identity(Healthy_life_expectancy) identity(Freedom_to_make_life_choices)
                    identity(Perceptions_of_corruption) ;
run;

/* WLS*/
FILENAME happy '/home/u49763051/MAT456/happy.txt';

data a1;
        infile happy;
        input Score GDP_per_capita Social_support Healthy_life_expectancy
Freedom_to_make_life_choices Generosity Perceptions_of_corruption;

ods graphics on;
proc reg data=a1;
```

```
 model Score=GDP_per_capita Social_support Healthy_life_expectancy
Freedom_to_make_life_choices Perceptions_of_corruption / clb;
 output out=a2 r=resid;
run;
ods graphics off;

data a2;
 set a2;
 absr=abs(resid);
 sqrr=resid*resid;

proc reg data = a2;
model absr = GDP_per_capita Social_support Healthy_life_expectancy Freedom_to_make_life_choices
Perceptions_of_corruption;
output out=a3 p=shat;
data a3;
set a3;
wt=1/(shat*shat);

ods graphics on;
proc reg data = a3 plots=COOKSD;
model Score = GDP_per_capita Social_support Healthy_life_expectancy
Freedom_to_make_life_choices Perceptions_of_corruption /clb VIF;
weight wt;
output out = a4 r = resid1;
run;
ods graphics off;

/*create scatter plot matrix*/
proc sgscatter data = a1;
title "Scatter plot matrix";
matrix Score GDP_per_capita Social_support Healthy_life_expectancy Freedom_to_make_life_choices
Generosity Perceptions_of_corruption /diagonal=(histogram kernel) ;
run;
proc corr data = a1;run;
```

Works Cited

Kishore, Kamal, et al. “Understanding the Determinants of Happiness through Gallup World Poll.” *Journal of Family Medicine and Primary Care*, vol. 9, no. 9, 2020, p. 4826.

Network, Sustainable Development Solutions. “World Happiness Report.” *Kaggle*, 27 Nov. 2019, www.kaggle.com/unsdsn/world-happiness.