Happiness Formula Project Report

CIS 240 Project: A statistical analysis of which factors affect the citizens' happiness the most across different countries.

Department_Team ID:
General_30

Team Members:

محمد هشام صلاح عبدالحميد
Seat Num.: 20201701245 Section Num.: 26

مازن محسن إبراهيم سعد
Seat Num.: 20201701229 Section Num.: 22

عمر ياسر حامد رمزي عبدالعزيز
Seat Num.: 20201701195 Section Num.: 19

يحيى إدريس مختار صديق
Seat Num.: 20201700982 Section Num.: 33

Dr. Ghada Hamed *2021/2022*

Table of Contents

Introduction

Motivation:

According to the Merriam-Webster Dictionary, Happiness is the state or quality of being happy that occurs as a result of encountering a joyful experience.

Definitions aside, when it comes to describing developed and/or developing countries, the Citizens' Happiness, and the state of well-being and contentment is one of the most important metrics. So, let's have a closer look into it.

Research has shown that there are six factors which affect Global Happiness. These factors listed below:

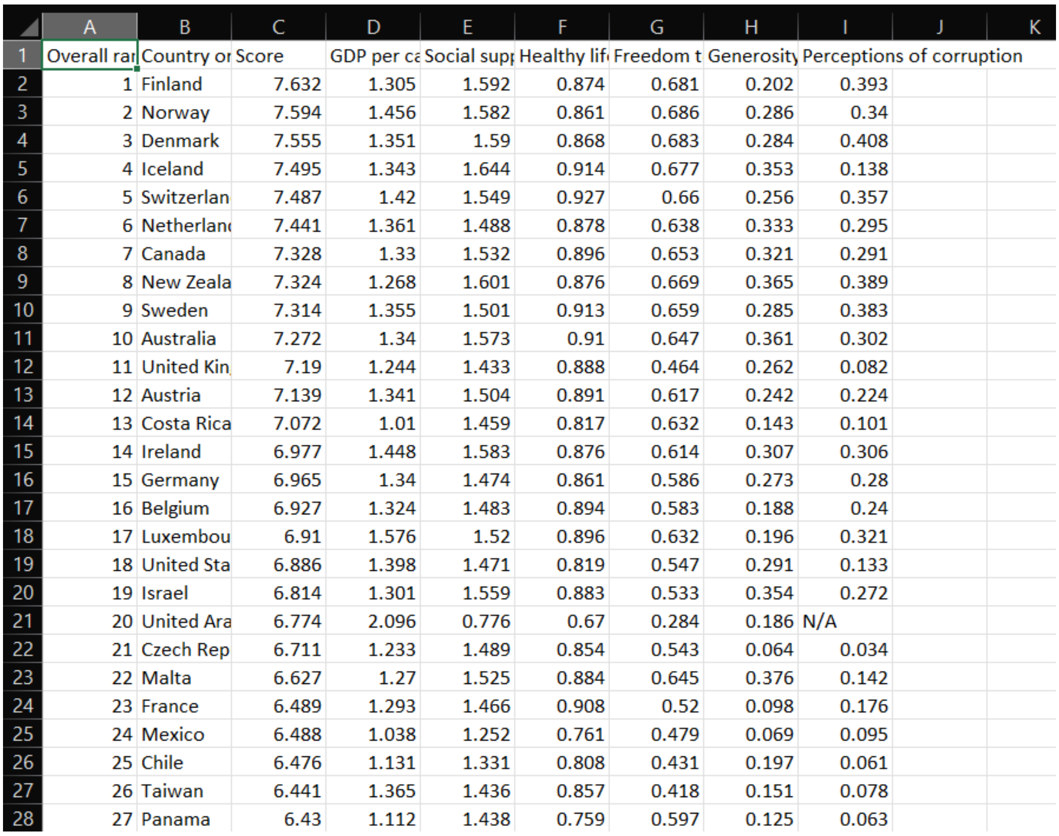
- Economic Production
- Social Support
- Life Expectancy
- Freedom
- Absence of Corruption
- Generosity

But the question of interest remains: **Which of the aforementioned factors affect the Global Citizens' Happiness the most?**

Happiness Formula is a project that aims to analyse data collected through surveys conducted across different countries concerning the citizens' happiness with their lives, with a single purpose in mind: determining which one of the six factors has the biggest influence on global happiness.

Direction:

The chosen dataset is the "Happiness 2018" dataset, one of Kaggle's datasets. The dataset ranks 156 countries/regions around the globe by their happiness level. Each country/region is given a score based on the six factors: GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity, perceptions of corruption.



	A	B	C	D	E	F	G	H	I	J	K
1	Overall ra	Country or	Score	GDP per ca	Social sup	Healthy lif	Freedom t	Generosity	Perceptions of corruption		
2	1	Finland	7.632	1.305	1.592	0.874	0.681	0.202	0.393		
3	2	Norway	7.594	1.456	1.582	0.861	0.686	0.286	0.34		
4	3	Denmark	7.555	1.351	1.59	0.868	0.683	0.284	0.408		
5	4	Iceland	7.495	1.343	1.644	0.914	0.677	0.353	0.138		
6	5	Switzerlan	7.487	1.42	1.549	0.927	0.66	0.256	0.357		
7	6	Netherlan	7.441	1.361	1.488	0.878	0.638	0.333	0.295		
8	7	Canada	7.328	1.33	1.532	0.896	0.653	0.321	0.291		
9	8	New Zeala	7.324	1.268	1.601	0.876	0.669	0.365	0.389		
10	9	Sweden	7.314	1.355	1.501	0.913	0.659	0.285	0.383		
11	10	Australia	7.272	1.34	1.573	0.91	0.647	0.361	0.302		
12	11	United Kin	7.19	1.244	1.433	0.888	0.464	0.262	0.082		
13	12	Austria	7.139	1.341	1.504	0.891	0.617	0.242	0.224		
14	13	Costa Rica	7.072	1.01	1.459	0.817	0.632	0.143	0.101		
15	14	Ireland	6.977	1.448	1.583	0.876	0.614	0.307	0.306		
16	15	Germany	6.965	1.34	1.474	0.861	0.586	0.273	0.28		
17	16	Belgium	6.927	1.324	1.483	0.894	0.583	0.188	0.24		
18	17	Luxembou	6.91	1.576	1.52	0.896	0.632	0.196	0.321		
19	18	United Sta	6.886	1.398	1.471	0.819	0.547	0.291	0.133		
20	19	Israel	6.814	1.301	1.559	0.883	0.533	0.354	0.272		
21	20	United Ara	6.774	2.096	0.776	0.67	0.284	0.186	N/A		
22	21	Czech Rep	6.711	1.233	1.489	0.854	0.543	0.064	0.034		
23	22	Malta	6.627	1.27	1.525	0.884	0.645	0.376	0.142		
24	23	France	6.489	1.293	1.466	0.908	0.52	0.098	0.176		
25	24	Mexico	6.488	1.038	1.252	0.761	0.479	0.069	0.095		
26	25	Chile	6.476	1.131	1.331	0.808	0.431	0.197	0.061		
27	26	Taiwan	6.441	1.365	1.436	0.857	0.418	0.151	0.078		
28	27	Panama	6.43	1.112	1.438	0.759	0.597	0.125	0.063		

Figure (1): Happiness 2018 dataset.

Here, the population consists of 156 countries, and a random sample was taken. The sample consisted of 30 countries to ensure that the sample follows the central limit theorem, which states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed regardless of whether the source population is normal or skewed, provided the sample size is sufficiently large (usually n >= 30).

The sample was visualised using both pie and bar charts for a better understanding of the taken data sample. To analyse the standard and the influence of each factor on the happiness score, central tendencies and variability were calculated where central tendencies illustrate where most of the data points lie, but variability how far apart the points are from each other. Central Tendency and Variability include Mean $\bar{x}$, where $\bar{x} = \sum x/n$, Variance s^2,where $s^2 = \sum(x - \bar{x})^2/(n - 1)$, and Standard Deviation s,where s = $\surd(\sum(x - \bar{x})^2/(n - 1))$.

Linear Regression Models were made to describe relationships between variables by fitting a line to the observed data using the most commonly used type of correlation, Pearson Correlation Coefficient r, where $r = \sum(ZxZy)/(n - 1)$, which is the measure of association of variables. Regression allows you to estimate how a dependent variable, happiness score, changes as the independent variables change which are the six aforementioned factors . Using the Regression Equation, $\hat{y} = \beta1.(x) + \beta0$, where $\beta1 = r(Sy/Sx)$, and $\beta0 = \bar{y} - \beta1.(\bar{x})$, the happiness score can be predicted given any factor with a coefficient of determination r^2, the proportion of the variation in the dependent variable that is predictable from the independent variable.

Statistical inference is the process of analysing the result and making conclusions from data subject to random variation. It is also called inferential statistics. Hypothesis testing and confidence intervals are the applications of statistical inference. With a confidence interval of 95%, using statistical inference, the means of population μ for each factor is then estimated to be between $\mu = \bar{x} \pm 1.96\,\sigma$ if the standard deviation of the population is known, or $\mu = \bar{x} \pm 1.96.(s)$ if not. This holds as the sample is large (usually n >= 30).

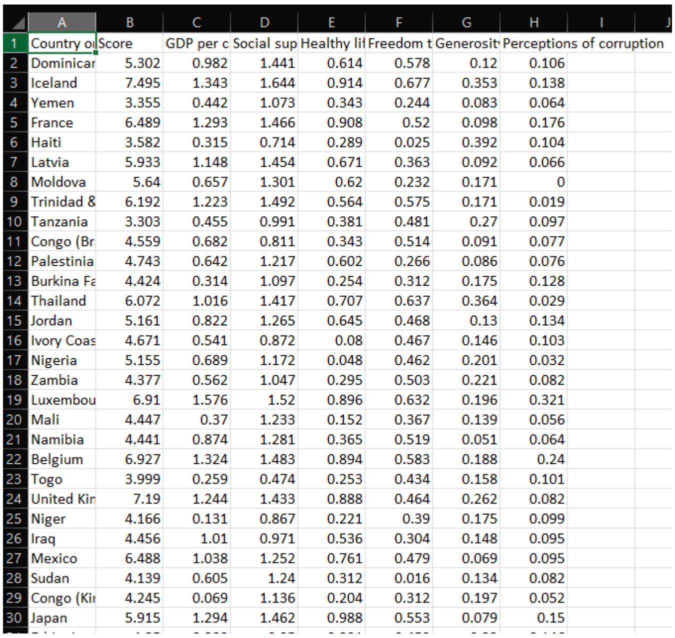
Applying Machine Learning, multiple linear regression was used to estimate the relationship between all factors (independent variables): including economic production, social support, life expectancy, freedom, absence of corruption, and generosity, and happiness score (dependent variable). A third of the dataset was set for training and the rest of the two thirds were set for testing. The random state was set to zero to ensure a consistent result that doesn’t change as a result of the random variability. To evaluate the model, a correlation coefficient r of the regression model was calculated and presented in the figure.

At last, the happiness score for each country/region was predicted given different values in the dataset of each factor with a coefficient of determination r^2 equals approximately 0.785 on average. This means that the study is now applicable, given the values of the factors, it can be determined which factor affects the happiness score the most.

Summary of Research

Exploratory Data Analysis:

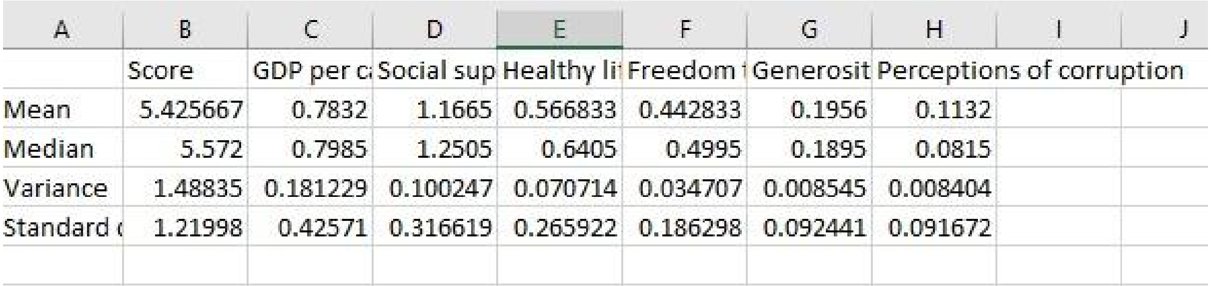
A simple random sample of 30 countries of each factor was taken for analysis, as shown in figure (2).



	A	B	C	D	E	F	G	H	I	J
1	Country o	Score	GDP per c	Social sup	Healthy li	Freedom t	Generosit	Perceptions of corruption		
2	Dominicar	5.302	0.982	1.441	0.614	0.578	0.12	0.106		
3	Iceland	7.495	1.343	1.644	0.914	0.677	0.353	0.138		
4	Yemen	3.355	0.442	1.073	0.343	0.244	0.083	0.064		
5	France	6.489	1.293	1.466	0.908	0.52	0.098	0.176		
6	Haiti	3.582	0.315	0.714	0.289	0.025	0.392	0.104		
7	Latvia	5.933	1.148	1.454	0.671	0.363	0.092	0.066		
8	Moldova	5.64	0.657	1.301	0.62	0.232	0.171	0		
9	Trinidad &	6.192	1.223	1.492	0.564	0.575	0.171	0.019		
10	Tanzania	3.303	0.455	0.991	0.381	0.481	0.27	0.097		
11	Congo (Br	4.559	0.682	0.811	0.343	0.514	0.091	0.077		
12	Palestinia	4.743	0.642	1.217	0.602	0.266	0.086	0.076		
13	Burkina Fa	4.424	0.314	1.097	0.254	0.312	0.175	0.128		
14	Thailand	6.072	1.016	1.417	0.707	0.637	0.364	0.029		
15	Jordan	5.161	0.822	1.265	0.645	0.468	0.13	0.134		
16	Ivory Coas	4.671	0.541	0.872	0.08	0.467	0.146	0.103		
17	Nigeria	5.155	0.689	1.172	0.048	0.462	0.201	0.032		
18	Zambia	4.377	0.562	1.047	0.295	0.503	0.221	0.082		
19	Luxembou	6.91	1.576	1.52	0.896	0.632	0.196	0.321		
20	Mali	4.447	0.37	1.233	0.152	0.367	0.139	0.056		
21	Namibia	4.441	0.874	1.281	0.365	0.519	0.051	0.064		
22	Belgium	6.927	1.324	1.483	0.894	0.583	0.188	0.24		
23	Togo	3.999	0.259	0.474	0.253	0.434	0.158	0.101		
24	United Kir	7.19	1.244	1.433	0.888	0.464	0.262	0.082		
25	Niger	4.166	0.131	0.867	0.221	0.39	0.175	0.099		
26	Iraq	4.456	1.01	0.971	0.536	0.304	0.148	0.095		
27	Mexico	6.488	1.038	1.252	0.761	0.479	0.069	0.095		
28	Sudan	4.139	0.605	1.24	0.312	0.016	0.134	0.082		
29	Congo (Kir	4.245	0.069	1.136	0.204	0.312	0.197	0.052		
30	Japan	5.915	1.294	1.462	0.988	0.553	0.079	0.15		

Figure (2): Simple Random Sample of 30 countries

The central tendency and variability parameters were calculated, as shown in figure (3), for happiness score and each of the six factors including economic production, social support, life expectancy, freedom, absence of corruption, and generosity, and happiness score.



A	B	C	D	E	F	G	H	I	J
	Score	GDP per c	Social sup	Healthy li	Freedom t	Generosit	Perceptions of corruption		
Mean	5.425667	0.7832	1.1665	0.566833	0.442833	0.1956	0.1132		
Median	5.572	0.7985	1.2505	0.6405	0.4995	0.1895	0.0815		
Variance	1.48835	0.181229	0.100247	0.070714	0.034707	0.008545	0.008404		
Standard (	1.21998	0.42571	0.316619	0.265922	0.186298	0.092441	0.091672		

Figure (3): Central Tendency and Variability Parameters.

The sample was visualised using pie and bar charts for a better understanding of the taken data sample, as shown in figure (4) and figure (5).

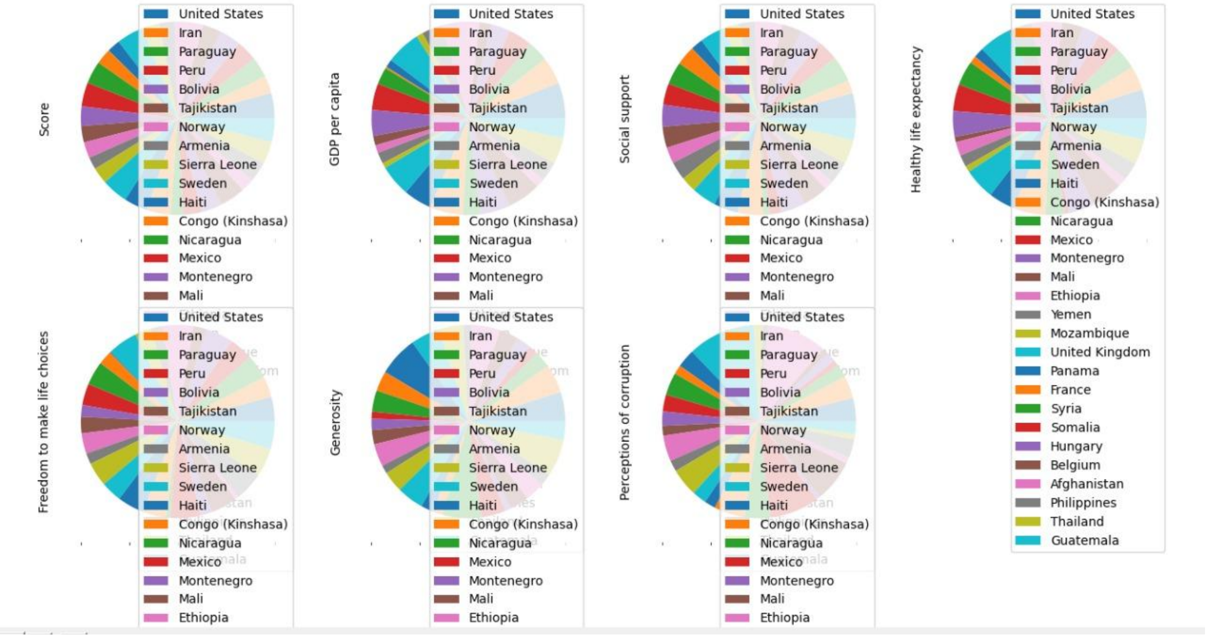


Figure (4): Visual representation of the sample using pie charts.

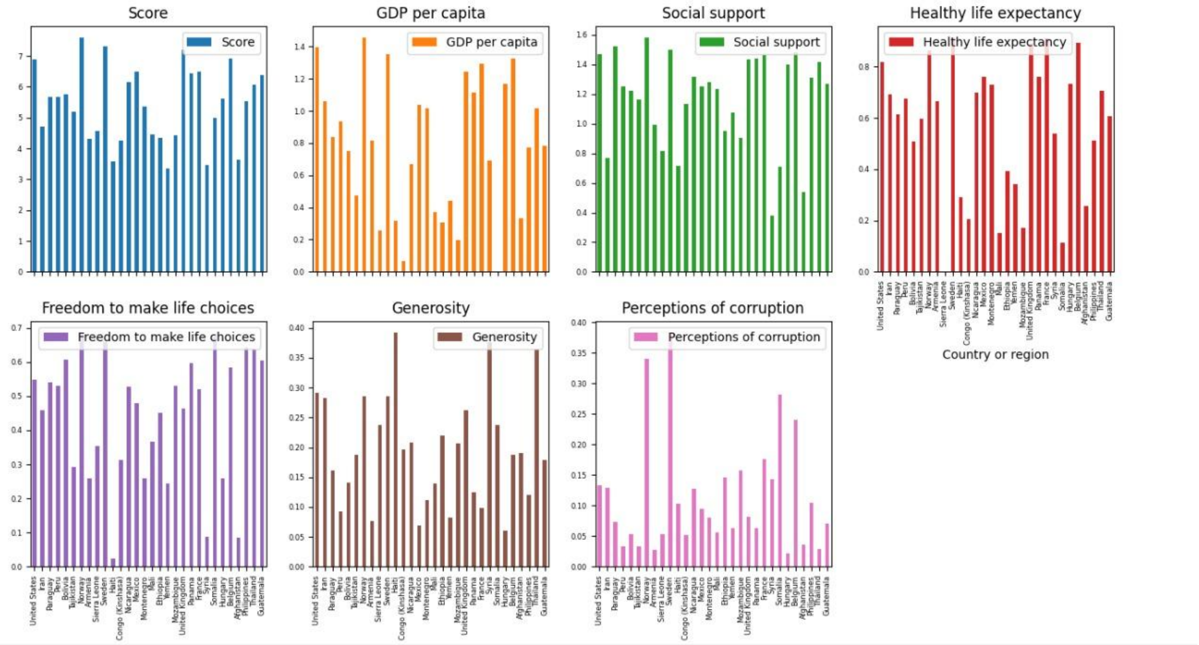


Figure (5): Visual representation of the sample using bar charts.

Correlations and Linear Regression:

Linear Regression Models then were made, as shown in the following figure, to describe relationships between variables by fitting a line to the observed data using Pearson Correlation Coefficient and Regression Equation for each factor as independent variable and happiness factor as the dependent one.

It can be seen that not all factors have a strong correlation with happiness. Some have a very weak correlation such as in figure (9).

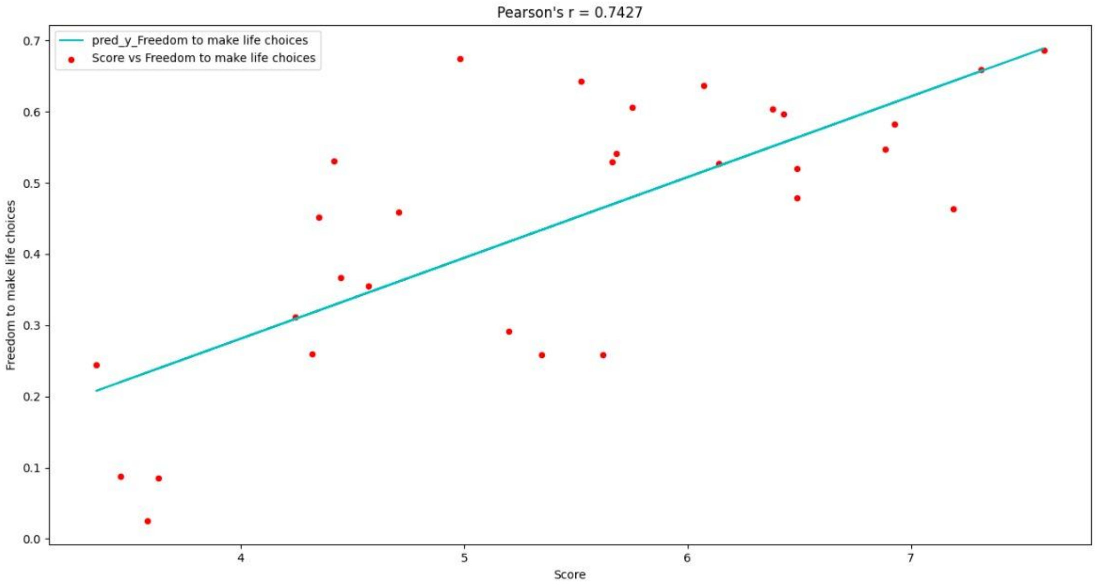


Figure (6): Regression model between Happiness Score and Freedom to Make Life Choices.

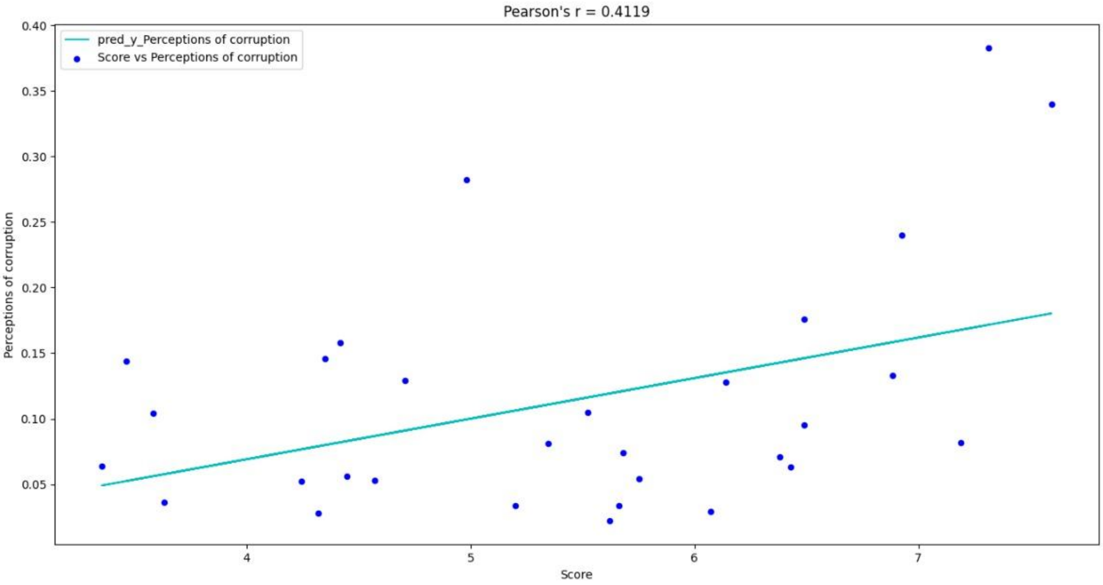


Figure (7): Regression model between Happiness Score and Perception of Corruption.

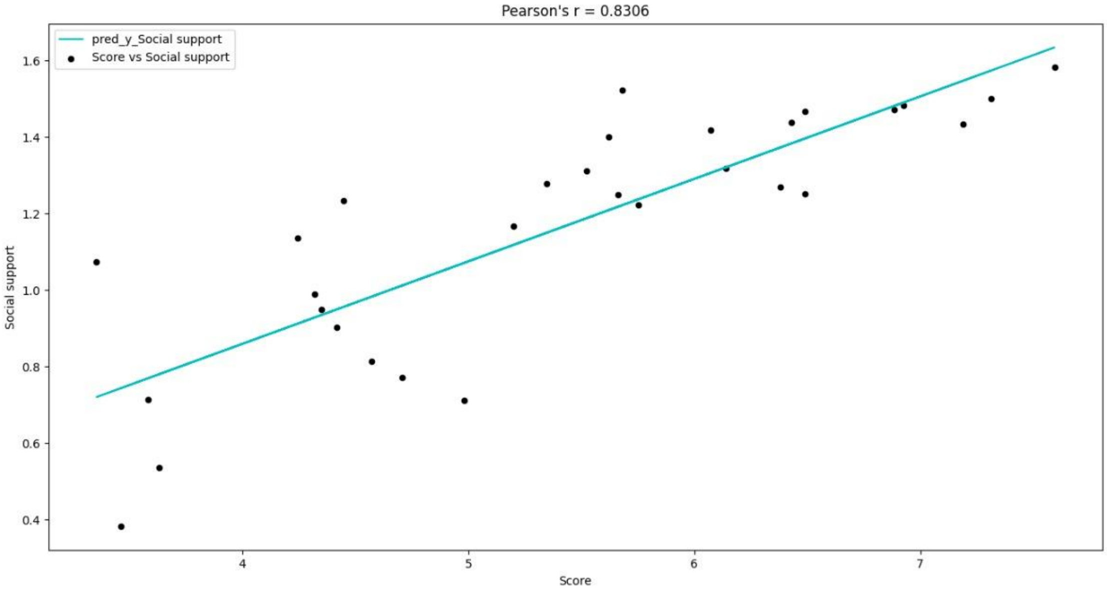


Figure (8): Regression model between Happiness Score and Social Support.

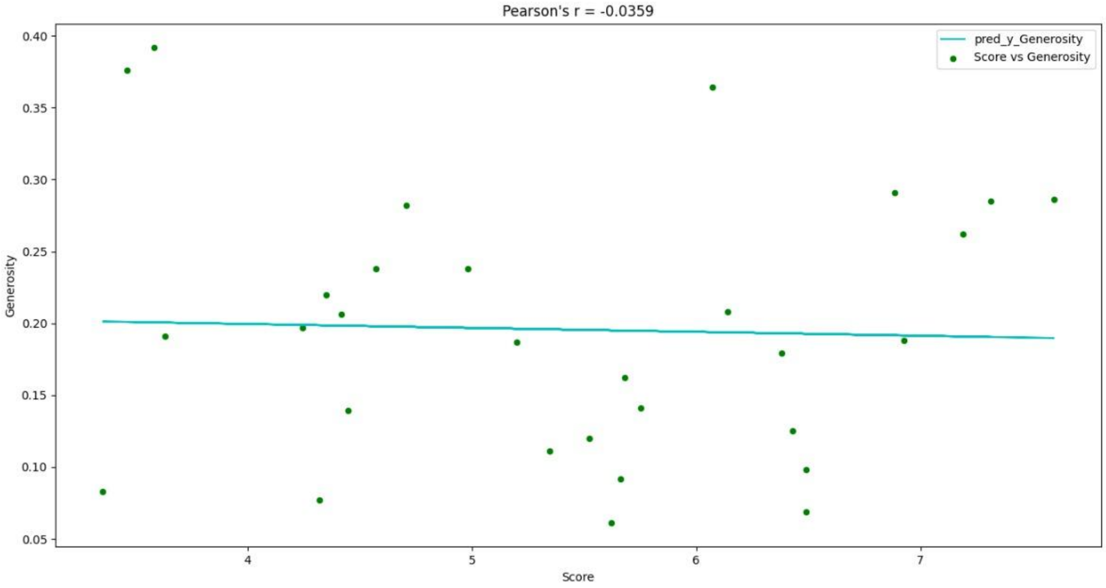


Figure (9): Regression model between Happiness Score and Generosity.

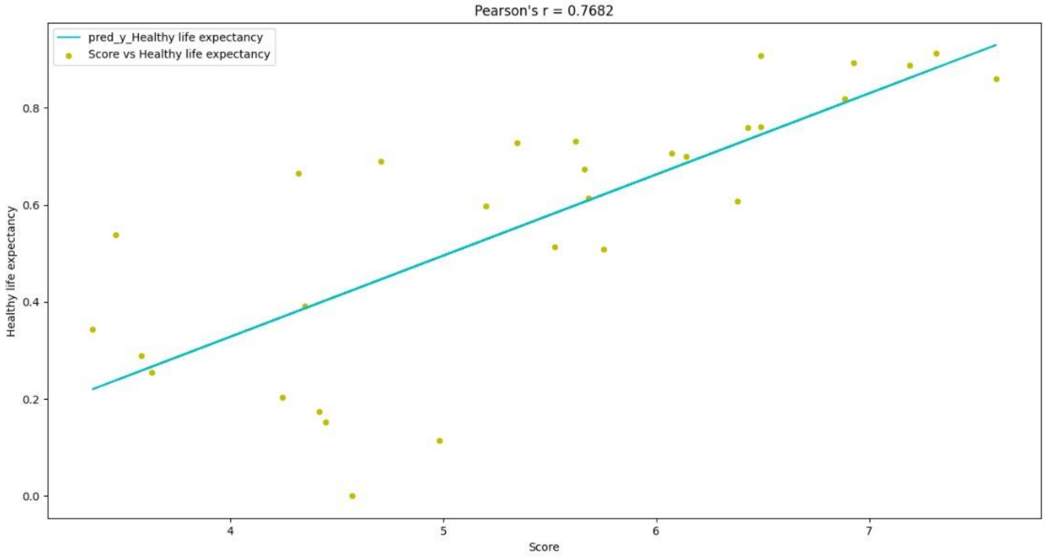


Figure (10): Regression model between Happiness Score and Health Life Expectancy.

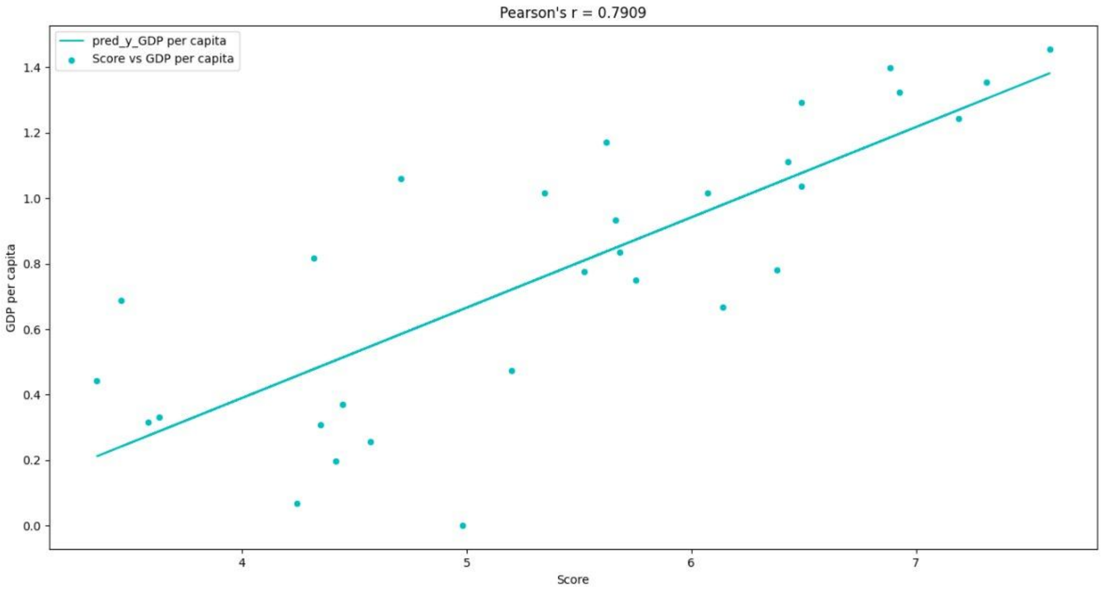
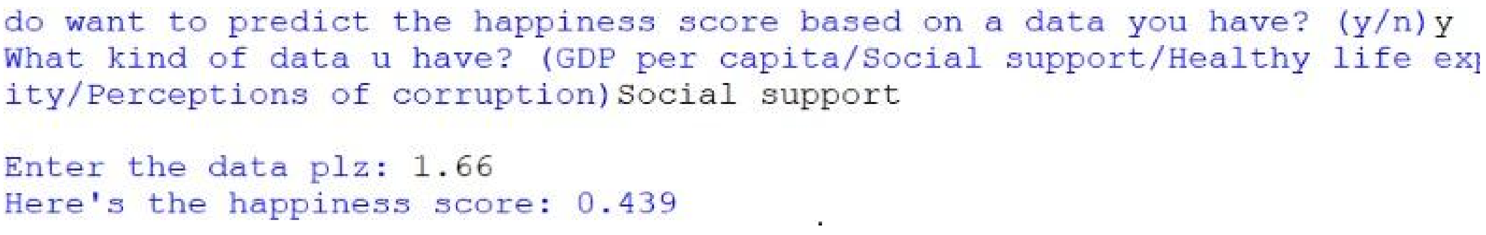


Figure (11): Regression model between Happiness Score and GDP per capita.

Using Linear Regression Equation for each Regression Model, the happiness score can be predicted given any factor value, as shown in figure (12).



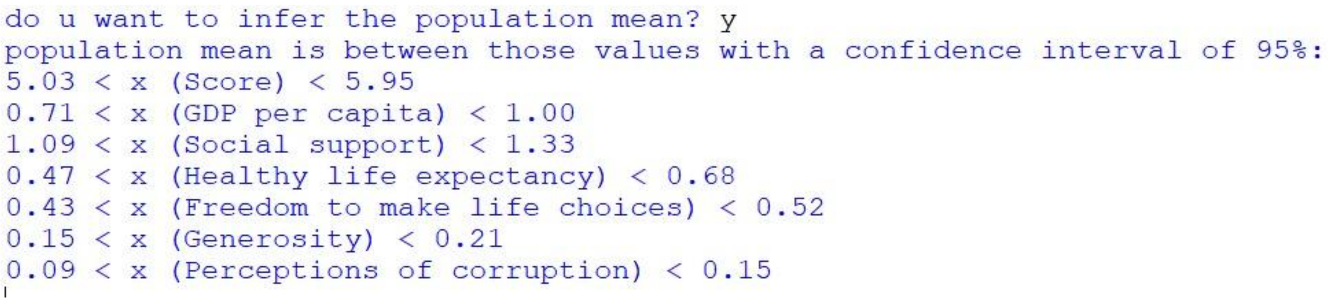
```
do want to predict the happiness score based on a data you have? (y/n)y
What kind of data u have? (GDP per capita/Social support/Healthy life ex
ity/Perceptions of corruption)Social support

Enter the data plz: 1.66
Here's the happiness score: 0.439
```

Figure (12): Happiness Score prediction given Social Support factor value

Inference:

The means of population μ for each factor is then estimated using statistical inference with a confidence interval of 95%, as shown in figure (13).



```
do u want to infer the population mean? y
population mean is between those values with a confidence interval of 95%:
5.03 < x (Score) < 5.95
0.71 < x (GDP per capita) < 1.00
1.09 < x (Social support) < 1.33
0.47 < x (Healthy life expectancy) < 0.68
0.43 < x (Freedom to make life choices) < 0.52
0.15 < x (Generosity) < 0.21
0.09 < x (Perceptions of corruption) < 0.15
```

Figure (13): The estimation of population means using statistical inference.

Multilinear Regression:

At last, the happiness score for each country/region was predicted given different values in the dataset of each factor using multiple linear regression. Then, a linear regression was made between the actual values of happiness score and the predicted values to evaluate the given model, as shown in figure (14).

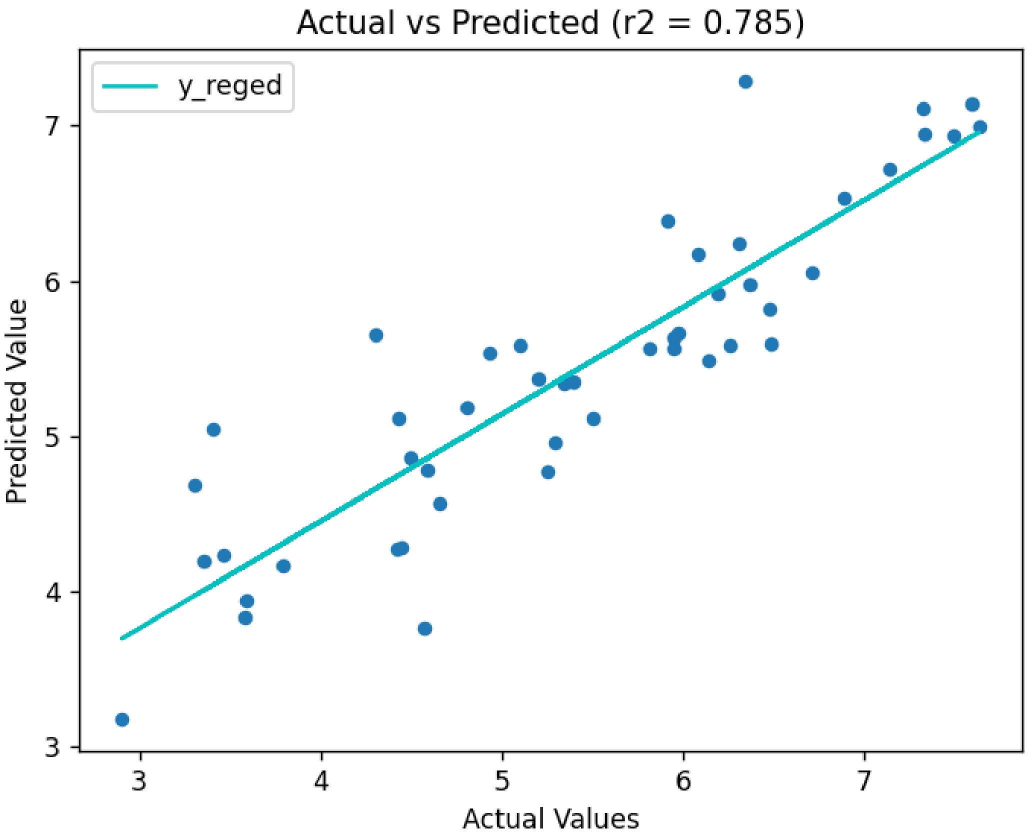


Figure (14): Regression model between the actual values of happiness score and the predicted values.

Conclusion:

After analysing the data, it can be seen that the two factors that are correlated the most with happiness are social support, and GDP per capita, respectively. While correlation doesn’t necessarily equal causation, one can argue that those two factors combined are what affects the overall happiness score the most.

Citations

- World Happiness Report. (n.d.). Retrieved December 20, 2021, from https://www.kaggle.com/unsdsn/world-happiness.

- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2

- McKinney, W., & others. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56).

- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science &amp; Engineering, 9(3), 90–95.

- Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.