

Mapping the Statistical Significance of Factors Contributing to the World Happiness Report

Karan Bhowmick, Charuchith Ranjit



Abstract: This paper aims to delineate findings of the statistical significance of the factors contributing to the happiness score. The happiness score also termed as ladder score, is a metric used by the United Nations Sustainable Development Solutions Network to metricize the happiness of the citizens in a country. To tackle this issue, we use regression and data visualization. We perform a survey on the factors affecting ladder score and how these factors can be used for predictive analytics. We use Linear Regression, Polynomial Regression, Lasso Regression with cross-validation, and Ridge Regression with cross-validation. Next, we use evaluation metrics like MSE, RMSE, Adjusted r-squared, and r-squared value for the evaluation of the factors on the predictive model. Then, we plot the countries mentioned in the report on a geographical scale based on their happiness index scores. Furthermore, we plot the statistical significance of these factors on a continental scale, to reveal insightful patterns over a larger geographical domain. We aim to bring to light the trends of the aforementioned factors and produce the significance of these results on a world map. The results will help elucidate the global patterns formed by these metrics. An additional application is an extrapolation of the results procured. To augment the metrics of the World Happiness Report in a statistically comprehensive way. Furthermore, through this evaluation, the world happiness report can be revised to accommodate more inclusive factors and mitigate the redundancy of the factors.

Keywords: RMSE, MSE, VIF, Ladder Score, GDP, Cross Validation, Overfitting, Underfitting.

I. INTRODUCTION

The World Happiness Report is a publication of the UN that ranks countries based on happiness, quantified by ratings from citizens in relation to different factors that affect a citizen's life including economic and social factors. The 2021 report had a special emphasis on the effect COVID-19 played. This ranking is done by the Cantril Ladder Survey. The final scores are on a scale of 1 to 10, with 10 being the happiest a citizen can be. The different scores. The scoring for the different metrics is done with the help of the Gallup World Poll. The Gallup World Poll is a global survey done in the individual countries with more than a hundred questions. There are 6 metrics currently used for getting the ladder score

and subsequently ranking the countries: Log GDP per capita, social support, healthy life expectancy, freedom to make life choices, Generosity, perceptions of corruption

We will be conducting a data-incentivized empirical study to conduct the multivariate analysis of different regression techniques. We evaluate the results of the model predicted using Multiple Linear Regression, Polynomial Regression, Ridge Regression and Lasso Regression. These techniques will be used in conjunction with Cross Validation for regularization and Principal Component Analysis for dimensionality reduction.

II. LITERATURE SURVEY

World Happiness Report (2021) delineates the procedures and advances made by the UN in ranking countries based on ladder score, refining the processes for data collection, and the progress they have made throughout the years. Furthermore, the document explicates the significance of the sociological and economic factors with a special emphasis on the effects of the pandemic. However, they do not analyze the relationship between the factors through data analysis techniques.[1]

John F. Helliwell et al. (2021), Statistical Appendix 1 for Chapter 2 of World Happiness Report 2021. This additional file demonstrates the statistical analyses performed in the study for variables already. It plots the progression of the reports throughout the years; they have also plotted the regression tables to evaluate the change in significance of the variables throughout the years. We take this research and build upon it in our research paper, with a more comprehensive emphasis on regression.[2]

Karan Bhowmick (2021) has presented Clustering Analysis for Residential Areas based on Neighborhood Amenities. This paper uses clustering analysis to segregate the neighborhoods of Mumbai, India. The paper uses the data available to create a new parameter for evaluation of the neighborhoods in the study.[4]

Jan-Emmanuel De Neve and George Ward (2017) have presented Does Work Make You Happy? Evidence from the World Happiness Report. This paper correlates job satisfaction and happiness, citing evidence from the World Happiness Report. They plot the job satisfaction in different geographical regions with a "life evaluation" metric This paper clearly describes how to extrapolate the data given by the World Happiness Report for statistical analyses and data visualization.[3]

Manuscript received on July 09, 2021.

Revised Manuscript received on July 15, 2021.

Manuscript published on August 30, 2021.

* Correspondence Author

Karan Bhowmick, Department of Information Technology, Vellore Institute of Technology, Vellore (Tamil Nadu), India. E-mail: bhowmickkaran@gmail.com

Charuchith Ranjit*, Department of Computer Science with Specialization in Bioinformatics, Vellore Institute of Technology, Vellore Tamil Nadu, India. E-mail: charuchith@gmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

III. RESEARCH BACKGROUND

The dataset has been obtained from the official World Happiness Report website. In this section we summarize the different variables that we will be working with.

Dataset Exploration

We will be using Data Analysis Techniques for exploring the relationship between the different variables. The dependent variable here will be the Ladder score. The independent variables are:

- Logged GDP per capita
- Social support
- Healthy life expectancy
- Freedom to make life choices
- Generosity
- Perceptions of corruption

Logged GDP per capita: This metric has been calculated with actual GDP values.

Social Support: This metric has been calculated based on the responses to the GWP question “If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?”

Healthy Life Expectancy: This metric has been calculated with the help of World Health Organization’s Global Health Observatory data repository.

Freedom to make life choices: This metric has been calculated based on the responses to the GWP question “Are you satisfied or dissatisfied with your freedom to choose what you do with your life?”.

Generosity: This metric has been calculated based on the responses to the GWP question “Have you donated money to a charity in the past month?” on GDP per capita.

Corruption Perception: This metric has been calculated based on the responses to the two GWP questions: “Is corruption widespread throughout

the government or not” and “Is corruption widespread within businesses or not?”

IV. RESEARCH METHODOLOGY

We will now describe the methodology that we have followed in our study. We explicate the data analytics and statistical methods used in this study.

A. Descriptive Statistics of the Dataset

The dataset has six independent variables that encapsulate a nation’s sociological and economic competencies relating to a citizen’s quality of life. The dependent variable is the ladder score, the dataset procured also mentions the significance of each variable scaled down for determining the ladder score. In this study, we look at 149 countries from different continents.

First, we checked the levels of multicollinearity between the variables in the dataset. We use a heat map to plot correlation among the variables as shown in Figure 1.

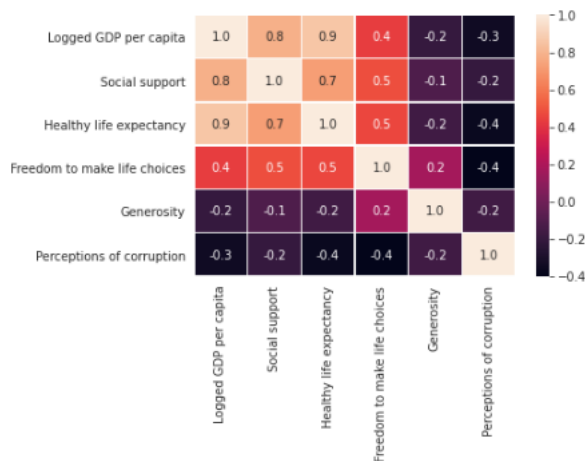


Fig. 1. Correlation heatmap

We use correlation to check for multicollinearity among the variables, since multicollinearity can impair the results obtained via regression. Here, we can see that Logged GDP per capita shows high correlation with Social Support and Health Life Expectancy. To remove the discrepancies caused by multicollinearity, we assess the quality of variables using Variance Inflation Factor or VIF.

The first instance of VIF calculation for the variables at hand can be seen in the figure below, Figure 2.

	variables	VIF
0	Explained by: Log GDP per capita	35.026571
1	Explained by: Social support	29.338650
2	Explained by: Healthy life expectancy	28.664292
3	Explained by: Freedom to make life choices	18.042354
4	Explained by: Generosity	4.346052
5	Explained by: Perceptions of corruption	3.230472

Figure 2: VIF for the independent variables

The acceptable score for VIF of independent variables should be less than 10. Good quality variables are those that have a VIF<5. Here, we can see that all the independent variables exceed the acceptable amount except Generosity and Perceptions of Corruption. So we use a standard scaler to scale the data and remove structural multicollinearity.

	variables	VIF
0	Explained by: Log GDP per capita	5.103236
1	Explained by: Social support	2.969610
2	Explained by: Healthy life expectancy	4.099477
3	Explained by: Freedom to make life choices	1.585310
4	Explained by: Generosity	1.181310
5	Explained by: Perceptions of corruption	1.366464

Fig. 3. VIF for the independent variables after scaling

Now, we have independent variables that are fit for regression and can produce meaningful statistical results. We can now proceed with Exploratory Data Analysis using Data Visualization.

B. Exploratory data analysis

We will use Exploratory Data Analysis to gain insights and produce from the dataset after preprocessing.

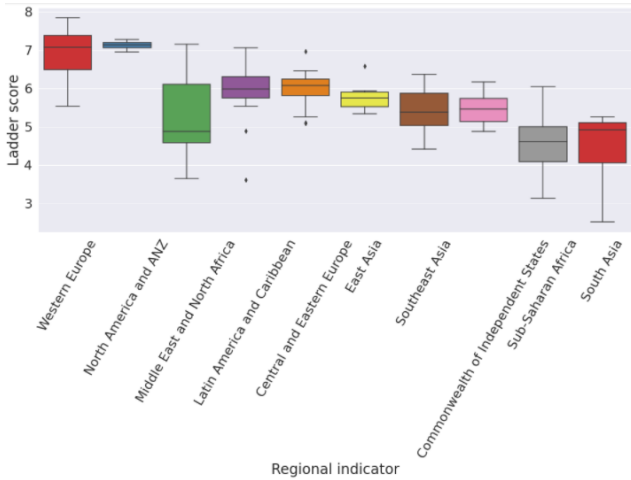


Fig. 4. Box plot between Regional Indicator and Ladder score

To understand the variation in Ladder score values across the different Regions, we have drawn a box plot of the same. We can clearly see the trend that Western European countries are happier while the South Asian and Sub-Saharan Africa countries did worse.

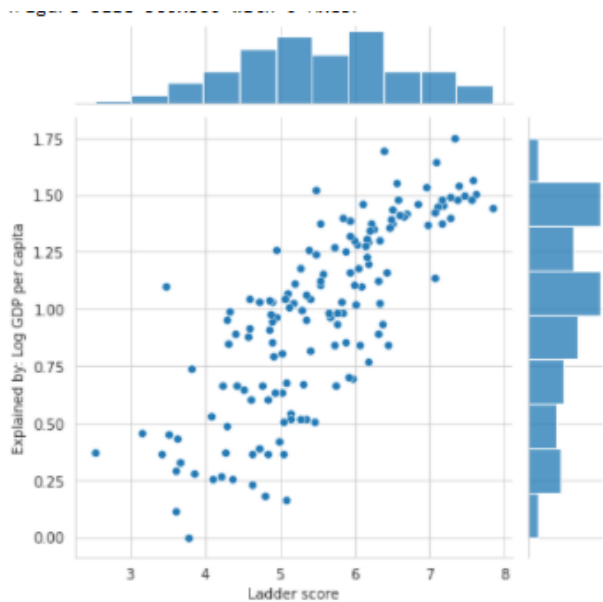


Fig. 6: Scatter plot between Ladder Score and Log GDP per capita

The scatter plot between Ladder Score and Log GDP per capita shows the high correlation between the 2 variables.

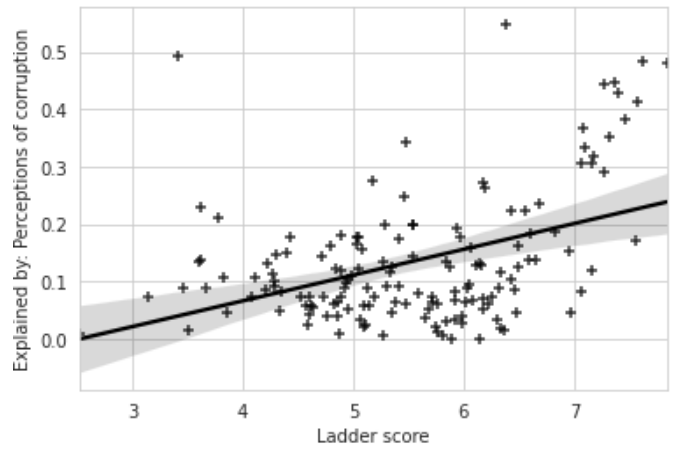


Fig. 6. Regression plot between Ladder score and Perceptions of corruption

Similarly, there is a slight positive correlation between Ladder score and Perception of corruption.

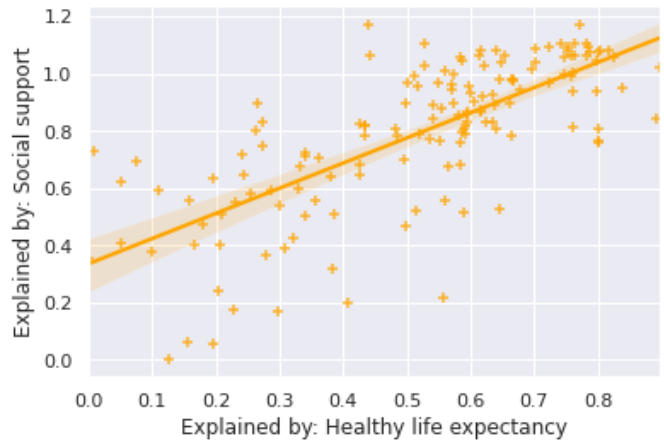


Fig. 7. Regression plot between Health Life Expectancy and Social Support

When drawing a regression plot between Health Life Expectancy and Social Support, we see that there is some positive correlation between the 2 independent variables.



Fig. 8. Regression plot between Explained by: Generosity and Perceptions of corruption

The plot shows a slight positive correlation between the independent variables Generosity and Perceptions of Corruption.

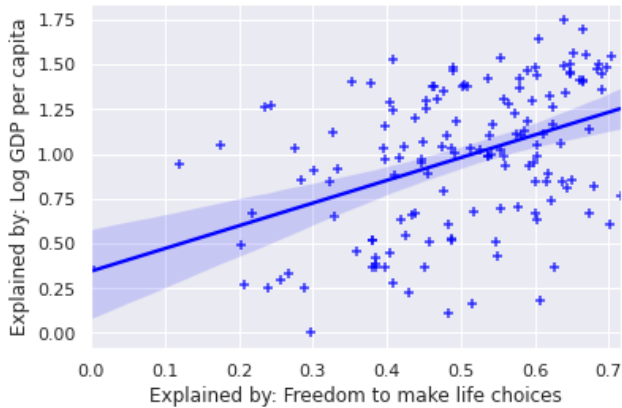


Fig. 9. Regression plot between Freedom to make life choices and Log GDP per capita

Similarly, this plot shows a slight positive correlation between the independent variables Log GDP per capita and Freedom to make life choices.

C. Regression Analytics

We will now use different regression algorithms to develop models to help predict the ladder score based on the independent variables. For this study, we use four algorithms for prediction, namely, Multiple Linear Regression, Polynomial Regression, Lasso Regression and Ridge Regression.

For this analysis, we divide the dataset for testing and training for use in developing each model. As per the industry standard, we use 30% of the data set for testing and the rest for training the algorithm. We also randomize the data to receive a more diverse range of data for testing and training.

Multiple Linear Regression:

Multiple linear regression predicts the dependent variable linearly in relation to the independent variables. The result of the equation produced in our results is given below:

Ladder score = $-0.00658629 + 0.26766075 * \text{Explained by: Log GDP per capita} + 0.27142922 * \text{Explained by: Social support} + 0.23810014 * \text{Explained by: Healthy life expectancy} + 0.19023918 * \text{Explained by: Freedom to make life choices} + 0.05889578 * \text{Explained by: Generosity} + 0.13026696 * \text{Explained by: Perceptions of corruption}$

Polynomial Regression:

Polynomial Regression uses equations of different degrees to predict the dependent variable. It comes into use when a linear model fails to provide accurate results for the data used.

The intercept and coefficients for the equation produced are:

Intercept: [0.02404635]

Coefficients: [[0.36682425 0.45053607 0.0679401 0.10784694 0.10220071 0.14544205 0.08909033 0.05051476 -0.03542378 -0.28216551 0.28298223 0.1375595 -0.08226707 0.1972162 0.11014726 0.08498324 0.26356126 0.00897961 0.14368769 -0.30761585

$-0.39740647 -0.09167526 -0.12906472 0.007051 -0.03617692 -0.00179268 -0.02373343]]$

Ridge Regression with Cross Validation:

Ridge regression is an extension of Multiple Linear Regression with an adaptation to the cost function. The cost function has been limited by the square of the weights and a regularization parameter 'alpha'. This is used to increase the dynamism of the model developed and increase reactivity of the model to new data. Cross validation is used to pick the best value of the regularization parameter. The equation produced using the model is given below:

Ladder score = $-0.00614776 + 0.253413245 * \text{Explained by: Log GDP per capita} + 0.26060027 * \text{Explained by: Social support} + 0.23563863 * \text{Explained by: Healthy life expectancy} + 0.18788742 * \text{Explained by: Freedom to make life choices} + 0.05108426 * \text{Explained by: Generosity} + 0.12713408 * \text{Explained by: Perceptions of corruption}$

Lasso Regression with Cross Validation

Lasso Regression is similar to Ridge Regression in that it has a cost function that is adapted differently. In the essence of cost function, we use the modulus of the weights.

Ladder score = $-0.007051432 + 0.26195068 * \text{Explained by: Log GDP per capita} + 0.26919622 * \text{Explained by: Social support} + 0.23423063 * \text{Explained by: Healthy life expectancy} + 0.18988816 * \text{Explained by: Freedom to make life choices} + 0.04877323 * \text{Explained by: Generosity} + 0.12451701 * \text{Explained by: Perceptions of corruption}$

D. Model Evaluation

The evaluation metrics that we have used for our models are:

MSE, RMSE, R-Squared Value and Adjusted R-Squared Value.

MSE: Mean Squared Error is the difference in the average values between the actual values and the predicted and the square of that.

RMSE: It is the square root of the MSE.

RMSE gives a larger weightage to bigger errors and thus RMSE helps indicate larger errors and outliers in a more proficient way.

R-Squared Value: This value shows the amount of variance that can be explained by the model between predicted values and expected values.

Adjusted R-Squared Value: Adjusted R-Squared takes the R-Squared value and changes it to accommodate the metric in terms of the number of independent variables.

V. EXPERIMENTAL ANALYSIS

In this section, we will present the findings and results of our statistical and regression analysis. We evaluate the regression models with the aforementioned evaluation metrics. We also present the findings of the report with respect to geographic scale through data visualization. This is done to elucidate the trends followed in larger portions of the world.

A. Comparative Analysis of Regression models

First, we start with analyzing the Multiple Linear Regression model, here we predict the values of Ladder score based on the six independent variables mentioned in the Research Methodology section.

We expatiate our findings through a distribution plot, which also serves as a way for us to look for instances of overfitting and underfitting. Furthermore, we evaluate the model with the evaluation metrics mentioned.

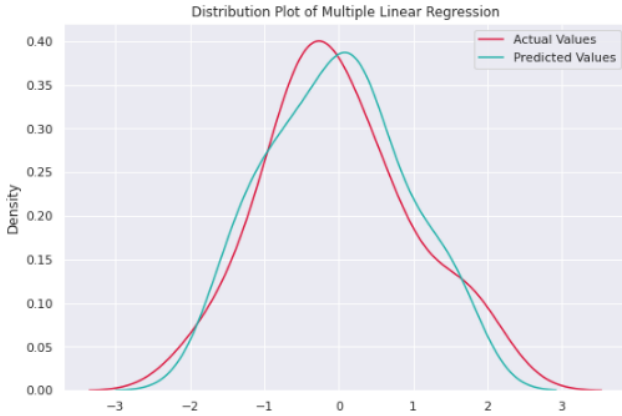


Fig. 10. Distribution Plot of the actual and predicted values for Multiple Linear Regression

As we can see the model is not overfitting and larger parts of the graph can be explained by the predicted value.

Now, let us take a look at the results of model evaluation.

MSE: 0.23559
RMSE: 0.48538
R-squared value: 0.75034
Adjusted R-Squared Value 0.71092

Fig. 11. Model Evaluation for MLR

Next, we see the results of the Polynomial Regression Model, here we use a polynomial regression of degree 2. We use the Elbow Method to find the best fit for the degree that produces the least amount of error.

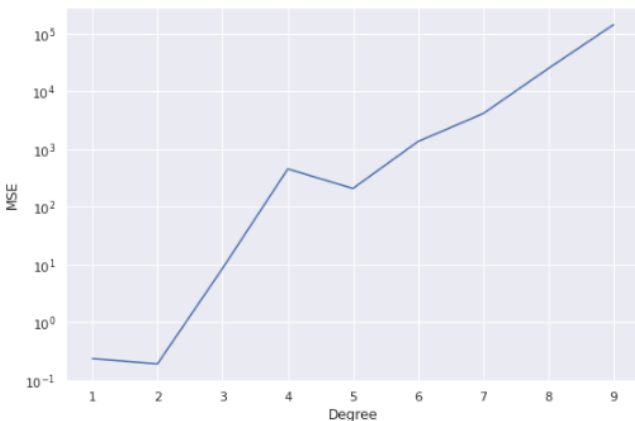


Fig. 12. Elbow Method to choose the best value of the degree

From the graph above, we can see that the best value of the degree is 2. For which the mean squared error is the least.

Next, we plot the distribution plot for Polynomial regression of degree 2.

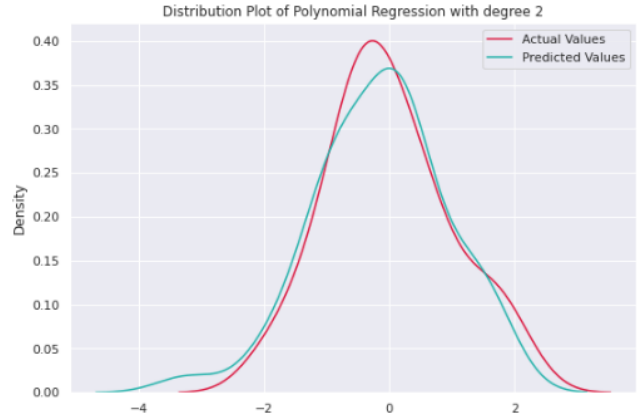


Fig. 13. Distribution Plot of the actual and predicted values for Polynomial Regression

We can see that there is a small instance of underfitting, but this model produces the most statistically valid results, as shown in Figure 14.

MSE: 0.19173
RMSE: 0.43787
R-squared value: 0.79682
Adjusted R-Squared Value 0.76474

Fig. 14. Model Evaluation for Polynomial Regression

Now, we move to the model developed by Ridge Regression. Here we enter a range of values for cross validation, between 0.01 and 10. So the model gives us the best value of the regularization parameter 'alpha' in this range and proceeds to develop the model with said value.

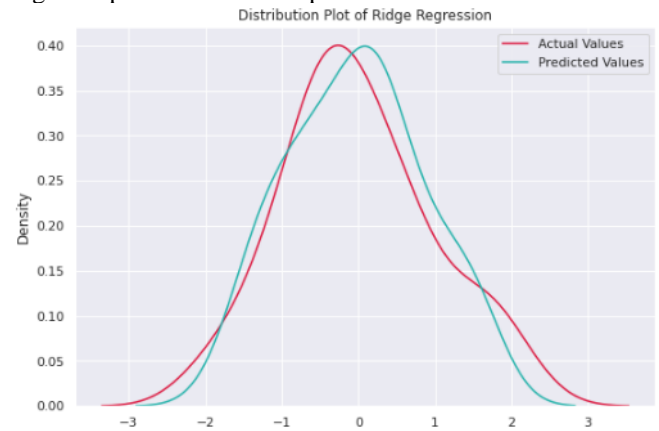


Fig. 15. Distribution Plot of the actual and predicted values for Ridge Regression

The evaluation metrics for Ridge Regression are shown below.

MSE: 0.23343
RMSE: 0.48315
R-squared value: 0.75263
Adjusted R-Squared Value 0.71357

Fig. 16. Model Evaluation for Ridge Regression

As we can see, it produces better results in terms of model performance than Multiple Linear Regression but it does not come close to Polynomial Regression.

Finally, we look at the results produced by Lasso Regression. We follow the same procedure for cross validation and evaluate the model.

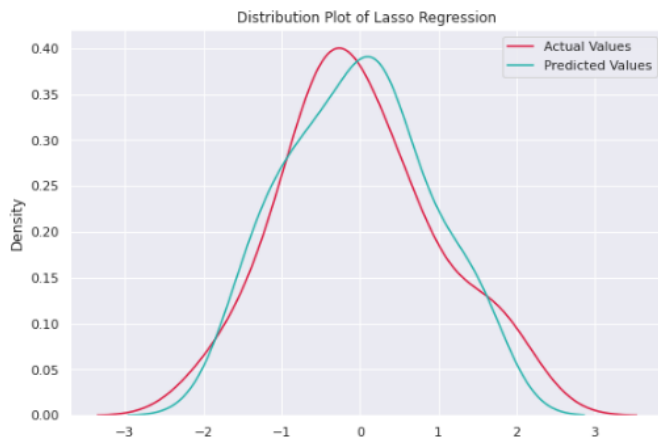


Fig. 17. Distribution Plot of the actual and predicted values for Lasso Regression

MSE: 0.23325
 RMSE: 0.48296
 R-squared value: 0.75282
 Adjusted R-Squared Value 0.71379

Fig. 18. Model Evaluation for Lasso Regression

We can see in the values of the different evaluation metrics that Lasso Regression produces the second-best results.

Thus, through our experimental analysis, we can see that Polynomial regression is the best predictor, followed by Lasso Regression, Ridge Regression and finally Multiple Linear Regression. We will be encompassing our results in a table to elucidate the findings of the model development for Ladder Score.

B. Data Visualization to analyze trends

We will plot different charts to analyze the trends that the different regions follow with respect to the different metrics.

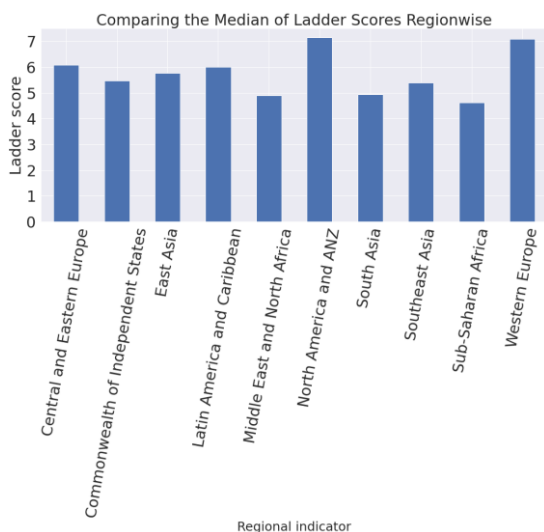


Fig. 19. Bar Charts comparing median of Ladder score against Regions

From Figure 19, we notice that Northern America & ANZ countries and Western European countries show a higher median Ladder Score which shows that the citizens of these countries feel happier. Countries in South Asia and Sub-Saharan Africa do comparatively poorly. Within the Asian Regions, countries in Eastern Asia do the best, followed by Southeast Asia and then lastly South Asia.

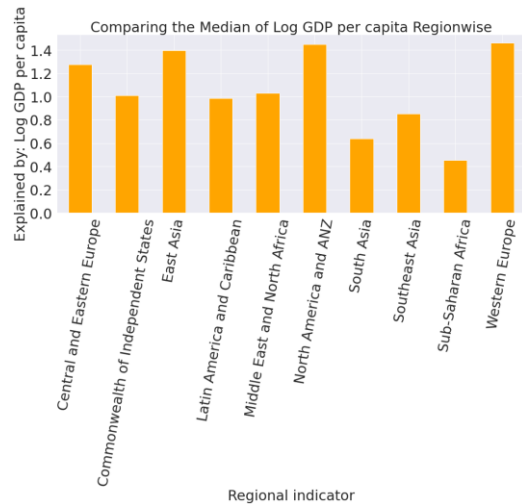


Fig. 20. Bar Charts comparing median of GDP per capita against Regions

We see that in terms of GDP, again the Western European countries and Northern European lead the pack, while East Asia and Central and Eastern Europe follow closely with a healthy metric. We see that even within a continent, like Asia, there is a huge variation between South Asia and East Asia, while the countries in the European region have performed similarly

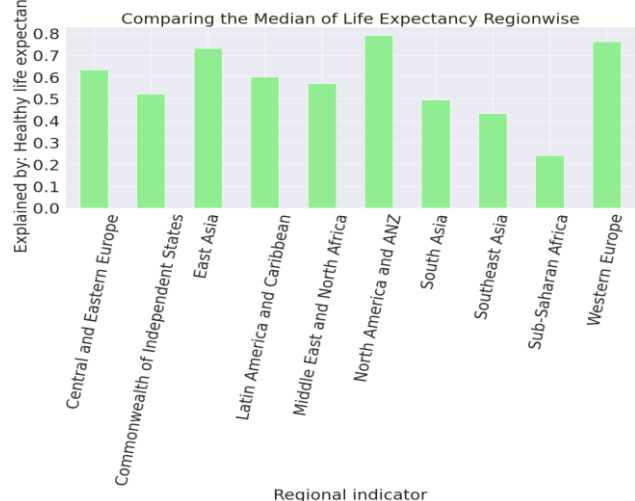


Fig. 21. Bar Charts comparing median of Life Expectancy against Regions

Life Expectancy follows a similar trend to the GDP graph with countries in East Asia, Western Europe and Northern America & ANZ performing well. But, countries in Sub-Saharan Africa have severely underperformed compared to the other regions.

Across the same continent of Africa, the Northern African countries have performed much better than their counterparts.

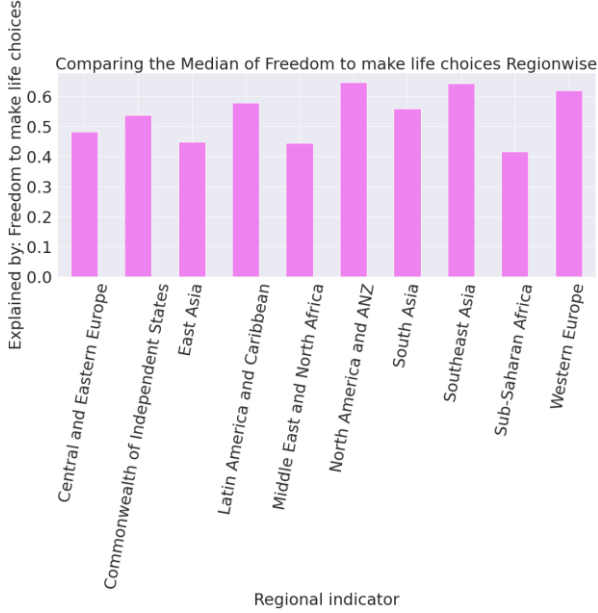


Fig. 22. Bar Charts comparing median of Freedom to make life choices against Regions

The Middle East and North Africa have shown a down trend in this graph as they are the worst performing in terms of giving the people the freedom to make life choices. East Asia, which has been a consistent top performer for the previous metrics also underperforms in this. In contrast, South Asia and Southeast Asia do well in the metric.

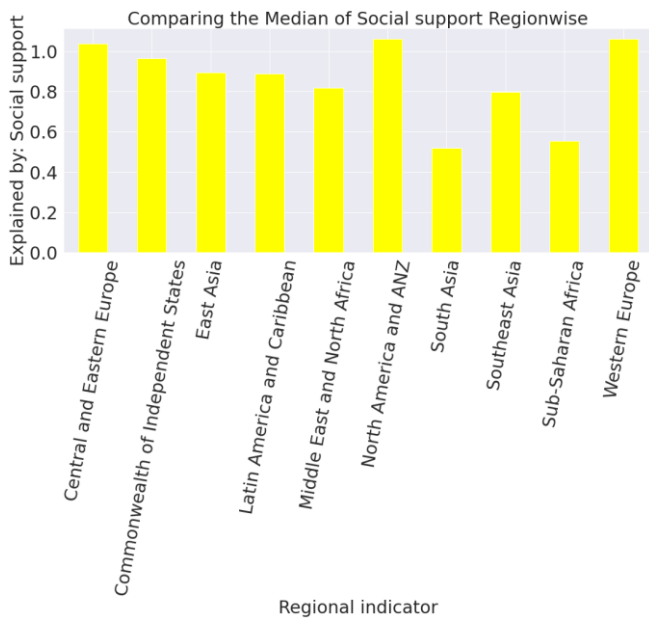


Fig. 23. Bar Charts comparing median of Social Support against Regions

In terms of Social Support, the European countries and North America lead from the front, while it is again South Asia and Sub-Saharan Africa that trail way behind. The Southern part of Asia also keeps underperforming relative to the rest of Asia. Similar to the GDP graph, the countries across Europe perform similarly in this metric.

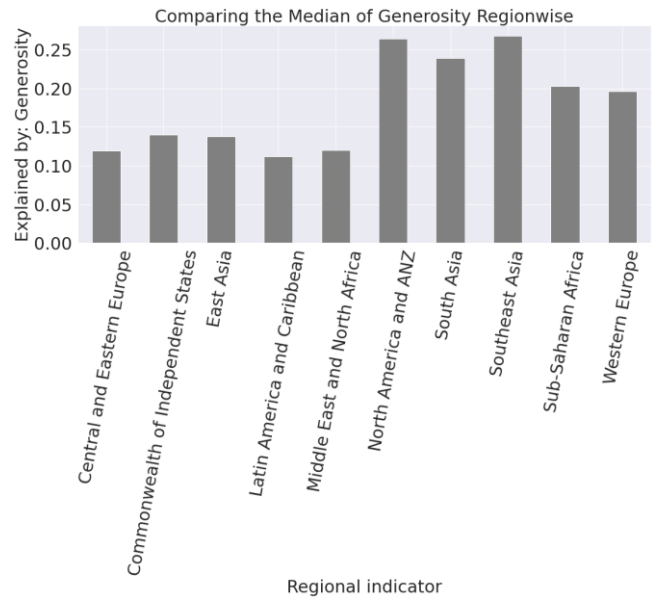


Fig. 24. Bar Charts comparing median of Generosity against Regions

Europe as a whole underperforms in terms of the generosity metric, while countries in Southeast Asia and South Asia have performed better in this metric compared to the others. This is very different from how Europe has performed in the other metrics. This is also the first time Sub-Saharan Africa has outperformed North Africa and the Middle East.

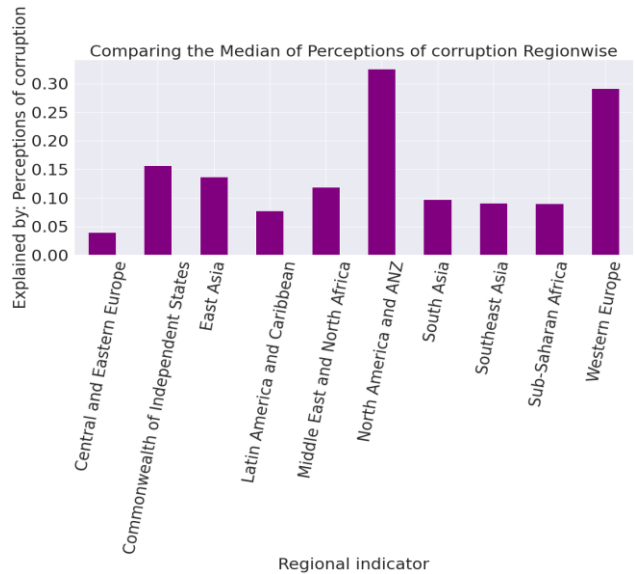


Fig. 25. Bar Charts comparing median of Perceptions of corruption against Regions

In terms of corruption, countries in Northern America & ANZ and Western Europe seem to do much better in terms of perception in their countries. Even within the same continent, Central and Eastern European countries do much worse in the same metric compared to their Western counterparts. All the Asian regions have underperformed in this metric.



Fig. 26. Visualization of the Ladder score among various continents

We classify countries into four different clusters, based on the range of the ladder score values. We have given a Color Index below to help differentiate the Ladder Score values from country to country.

Colour Index:

Pink - Ladder score < 3.56

Dark salmon - LS between 3.56 and 4.62

Dark orange - LS between 4.62 and 5.68

Red - LS between 5.68 and 6.74

Maroon - Ladder score > 6.74

We use the gradients of pink and orange to show the increasing levels of Ladder score on the world map. Now, we take a look at the countries marked by the circle marker in the different continents to get a better perspective of the geographic trends.

The trends in continents are explicated through these data visualizations, we first take a look at North America.

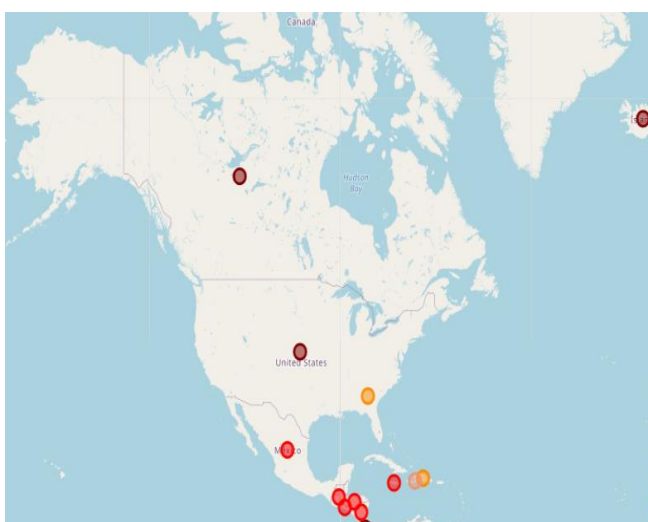


Fig. 27. Visualization of Ladder score of countries in North America

We can see here that Canada and USA have been marked by the maroon marker, suggesting the highest strata of happiness or Ladder Score. The country in the southern region of North America shows high values of the Ladder score as well and is marked by red i.e. Mexico.

Next, we take a look at South America.



Fig. 28. Visualization of Ladder score of countries in South America

These countries show red orange and maroon markers, which indicate an above average ladder score, signaling towards greater contentment of its citizens. These countries include El Salvador, Guatemala, Panama, Columbia, Ecuador, Peru and Bolivia marked by red. Costa Rica marked by maroon is an outlier and so are Paraguay, Venezuela and Dominican Republic marked by orange.

Now, we move to Africa.

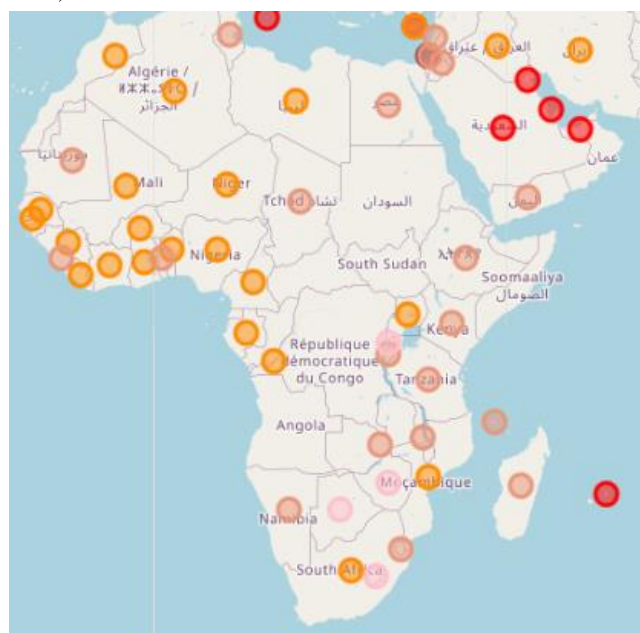


Fig 29. Visualization of Ladder score of countries in Africa

This continent has sporadic markers of pink, dark salmon and orange markers, indicating a low-mid level of happiness among the citizens. Mauritius is the only outlier with a red marker indicating high levels of happiness. Ivory Coast, Algeria, Niger, Nigeria and Gabon are marked with orange indicating better happiness among the north and northwestern regions of the continent. Madagascar, Kenya, Namibia, Sierra Leone and others are marked with dark salmon.

Here Botswana, Rwanda Mozambique and Zimbabwe produce low levels of happiness. The south and southeastern regions of the continent show lesser levels of ladder scores than the rest.

We now move on to Asia.



Fig. 30. Visualization of Ladder score of countries in Asia

By far the most diversity is shown in the countries of Asia. The middle east shows red, orange and dark salmon markers. With one outlier being maroon. Kazakhstan, Uzbekistan, North Korea, South Korea, Japan, Philippines and Singapore among others show high levels of happiness. China, Russia, Nepal, Bangladesh and Pakistan among others show average levels of happiness whereas, India and Myanmar show low levels of happiness and Afghanistan marked with pink shows extremely low levels of happiness. Mapping out trends in Asia is difficult since the markers are sporadic and follow no general trend.

Further, we now analyze Europe.



Fig. 31. Visualization of Ladder score of countries in Europe

Europe shows the best happiness scores out of all the continents and has the greatest number of countries marked by maroon. Some are marked with red and fewer with orange. Austria, UK, Ireland, Germany, Netherlands, Iceland, Finland and Sweden are marked with maroon. Italy, France, Spain, Portugal, Romania and others are marked with red, showing high levels of happiness. North Macedonia, Albania, Bulgaria and a few others have average levels of happiness, marked hence with orange.

Finally, we look at Australia.



Fig. 32. Visualization of Ladder score of countries in Australian continent

Both countries in this continent are marked by maroon, denoting the highest level of happiness among the citizens.

Thus, we have an overview of different trends shown in various regions of the globe. Extensive geographic trend mapping can be done extensively with the availability of a few more parameters.

VI. RESULT AND DISCUSSION

We have presented the insights gained by regression analytics on the model developed in the table below. The columns are the results of the evaluation of each model in terms of the variance explained and the errors produced by each of them.

Table 1: Comparison of the different Regression Algorithms

Algorithm	MSE value	RMSE value	R-Squared value	Adjusted R - Squared value
Multiple Linear Regression	0.2356	0.4854	0.7503	0.711
Polynomial Regression	0.1917	0.4379	0.7968	0.765
Ridge Regression	0.2334	0.4832	0.75263	0.71357
Lasso Regression	0.2333	0.483	0.7528	0.71379

Furthermore, we have successfully mapped the trends of Ladder Score, Logged GDP per capita, Perceptions of corruption, Generosity etc. on a wider scale. To accentuate trends on a global scale, in terms of the sociological and economic factors that affect countries the most.

VII. CONCLUSION AND FUTURE WORK

This paper highlights the different trends that the factors of ladder score have towards the happiness of the citizens in a country. We performed regression analysis on the dataset and evaluated the different models developed with evaluation and performance metrics. Finally, we visualized the ladder score of the different regions to demarcate trends on a global scale.

In the future, independent variables that were not part of the happiness report, like crime rate, unemployment or housing crisis, can also be taken into account to analyze their effects on the ladder score.

REFERENCES

1. World Happiness Report (2021), <https://worldhappiness.report/ed/2021/>
2. John F. Helliwell, Haifang Huang, Shun Wang and Max Norton (2021), Statistical Appendix 1 for Chapter 2 of World Happiness Report 2021.
3. Jan-Emmanuel De Neve and George Ward (2017), Does Work Make You Happy? Evidence from the World Happiness Report.
4. Karan Bhowmick (2021), Clustering Analysis for Residential Areas Based on Neighbourhood Amenities, *Int. J. of Adv. Res.* 9 (Jan). 957-965] (ISSN 2320-5407).

AUTHORS PROFILE



Karan Bhowmick, currently pursuing Bachelor of Technology in Information Technology in Vellore Institute of Technology, Vellore. The author has published two papers before.
Email-ID: bhowmickkaran@gmail.com



Charuchith Ranjit, currently pursuing Bachelor of Technology in Computer Science with specialization in Bioinformatics in Vellore Institute of Technology, Vellore. Email-ID: charuchith@gmail.com