# World Happiness Report

Ezra Nyabuti

Evan Ma

Hitesh Mantrabuddi

Varun Aravapalli

# Introduction

The world happiness index is a component the World Happiness Report  published by the UN annually using factors such as:

1. GDP per capita
2. Social Support
3. Healthy life expectancy at birth
4. Freedom to make life choices
5. Generosity
6. Perception of corruption

These report continue to gain global recognition which in turn help the government in policy making decisions.

# Agenda

Using Exploratory Data Analysis to explore factors that are mostly correlated with happiness.

Establish hypothesis about what affects the happiness score

Use machine learning models to predict the happiness score

# Questions:

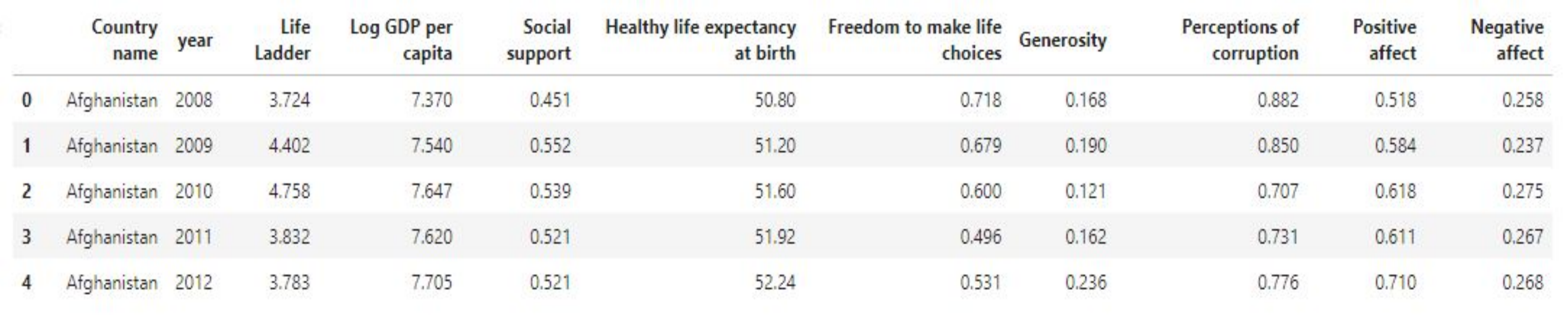
What factors have a direct influence on the happiness score?

How does the Happiness score differ based on country / region?

# Data Cleaning

The first step of our EDA was to clean our data and remove unnecessary/ irrelevant column.

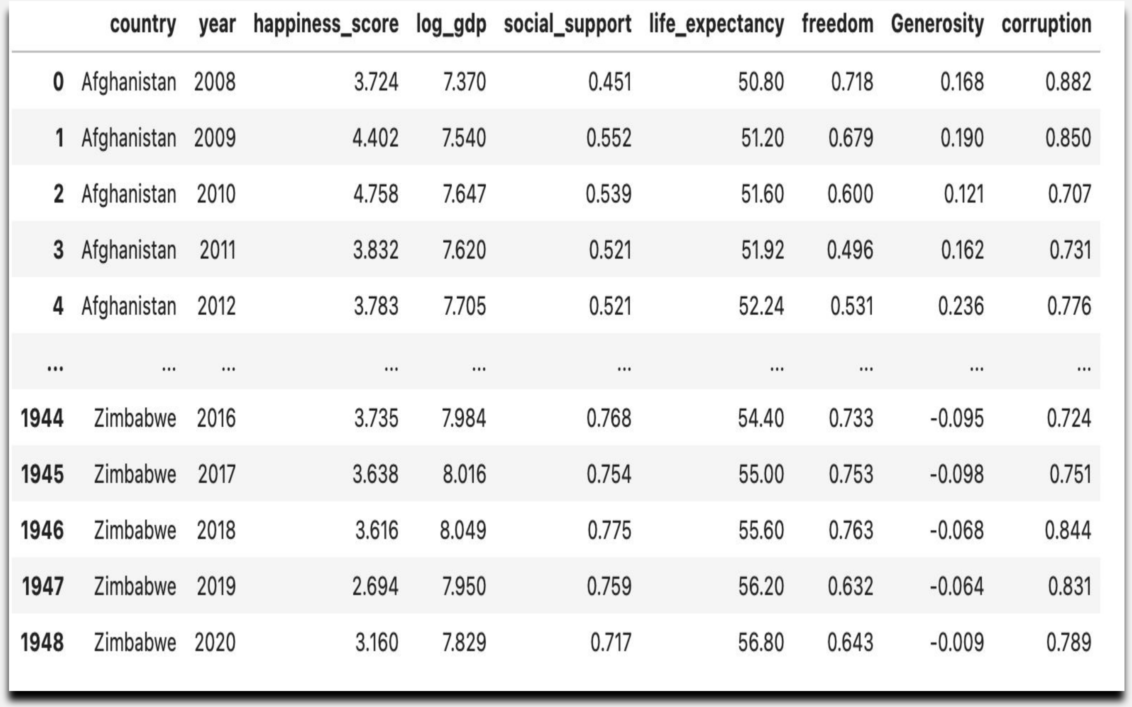
We also changed some of the column name to refer to each column easily



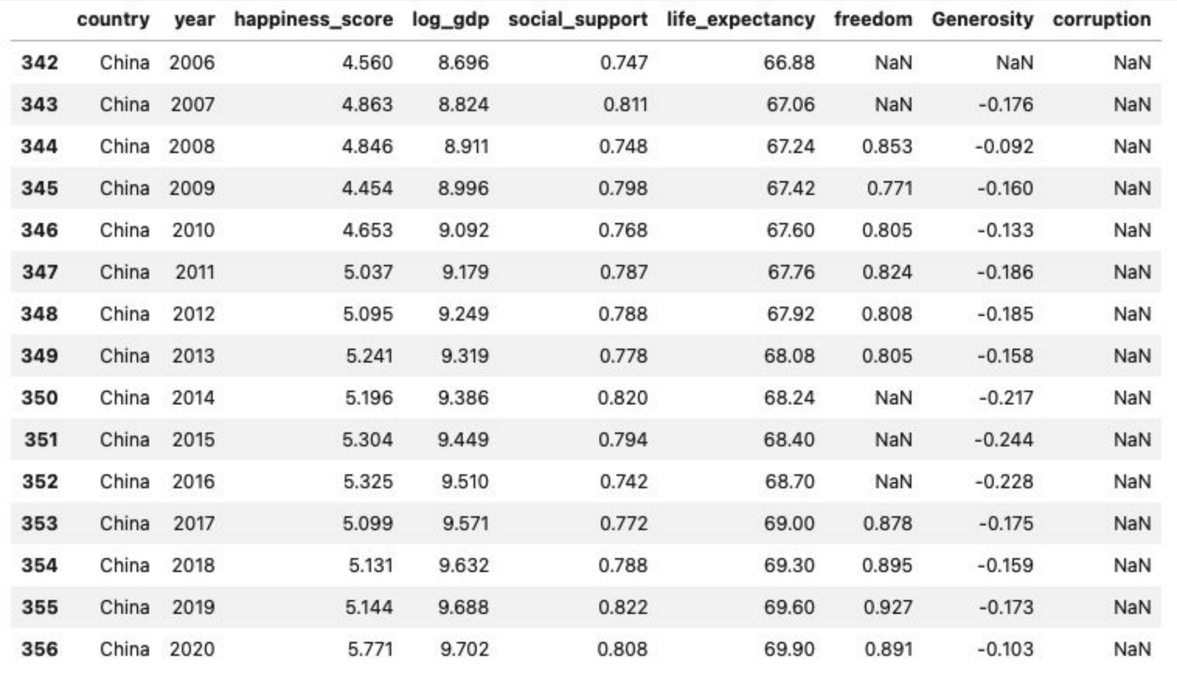
| | Country name | year | Life Ladder | Log GDP per capita | Social support | Healthy life expectancy at birth | Freedom to make life choices | Generosity | Perceptions of corruption | Positive affect | Negative affect |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Afghanistan | 2008 | 3.724 | 7.370 | 0.451 | 50.80 | 0.718 | 0.168 | 0.882 | 0.518 | 0.258 |
| **1** | Afghanistan | 2009 | 4.402 | 7.540 | 0.552 | 51.20 | 0.679 | 0.190 | 0.850 | 0.584 | 0.237 |
| **2** | Afghanistan | 2010 | 4.758 | 7.647 | 0.539 | 51.60 | 0.600 | 0.121 | 0.707 | 0.618 | 0.275 |
| **3** | Afghanistan | 2011 | 3.832 | 7.620 | 0.521 | 51.92 | 0.496 | 0.162 | 0.731 | 0.611 | 0.267 |
| **4** | Afghanistan | 2012 | 3.783 | 7.705 | 0.521 | 52.24 | 0.531 | 0.236 | 0.776 | 0.710 | 0.268 |

# Data Cleaning

This is the updated dataset.



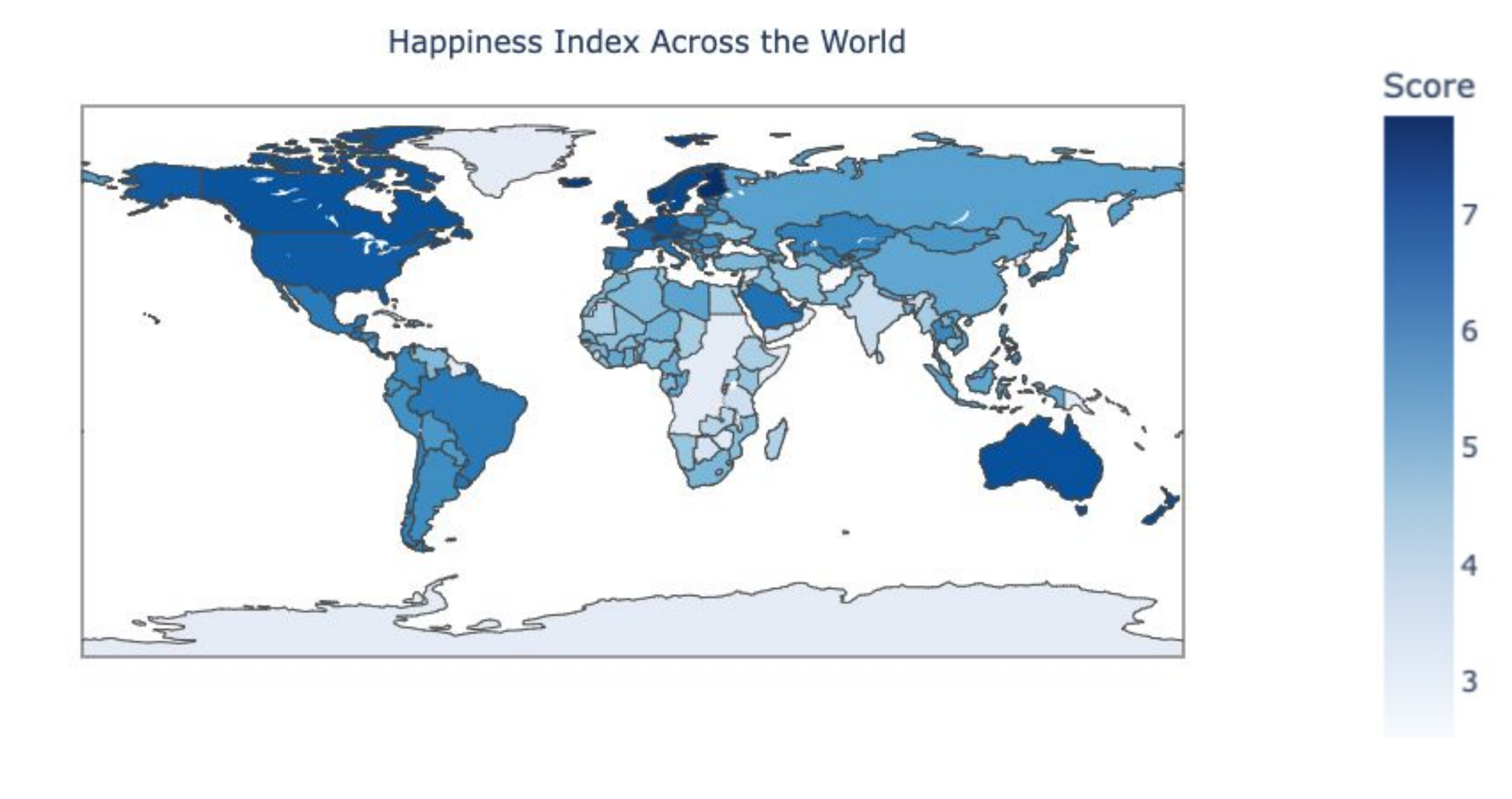
| | country | year | happiness_score | log_gdp | social_support | life_expectancy | freedom | Generosity | corruption |
|---|---|---|---|---|---|---|---|---|---|
| **0** | Afghanistan | 2008 | 3.724 | 7.370 | 0.451 | 50.80 | 0.718 | 0.168 | 0.882 |
| **1** | Afghanistan | 2009 | 4.402 | 7.540 | 0.552 | 51.20 | 0.679 | 0.190 | 0.850 |
| **2** | Afghanistan | 2010 | 4.758 | 7.647 | 0.539 | 51.60 | 0.600 | 0.121 | 0.707 |
| **3** | Afghanistan | 2011 | 3.832 | 7.620 | 0.521 | 51.92 | 0.496 | 0.162 | 0.731 |
| **4** | Afghanistan | 2012 | 3.783 | 7.705 | 0.521 | 52.24 | 0.531 | 0.236 | 0.776 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **1944** | Zimbabwe | 2016 | 3.735 | 7.984 | 0.768 | 54.40 | 0.733 | -0.095 | 0.724 |
| **1945** | Zimbabwe | 2017 | 3.638 | 8.016 | 0.754 | 55.00 | 0.753 | -0.098 | 0.751 |
| **1946** | Zimbabwe | 2018 | 3.616 | 8.049 | 0.775 | 55.60 | 0.763 | -0.068 | 0.844 |
| **1947** | Zimbabwe | 2019 | 2.694 | 7.950 | 0.759 | 56.20 | 0.632 | -0.064 | 0.831 |
| **1948** | Zimbabwe | 2020 | 3.160 | 7.829 | 0.717 | 56.80 | 0.643 | -0.009 | 0.789 |



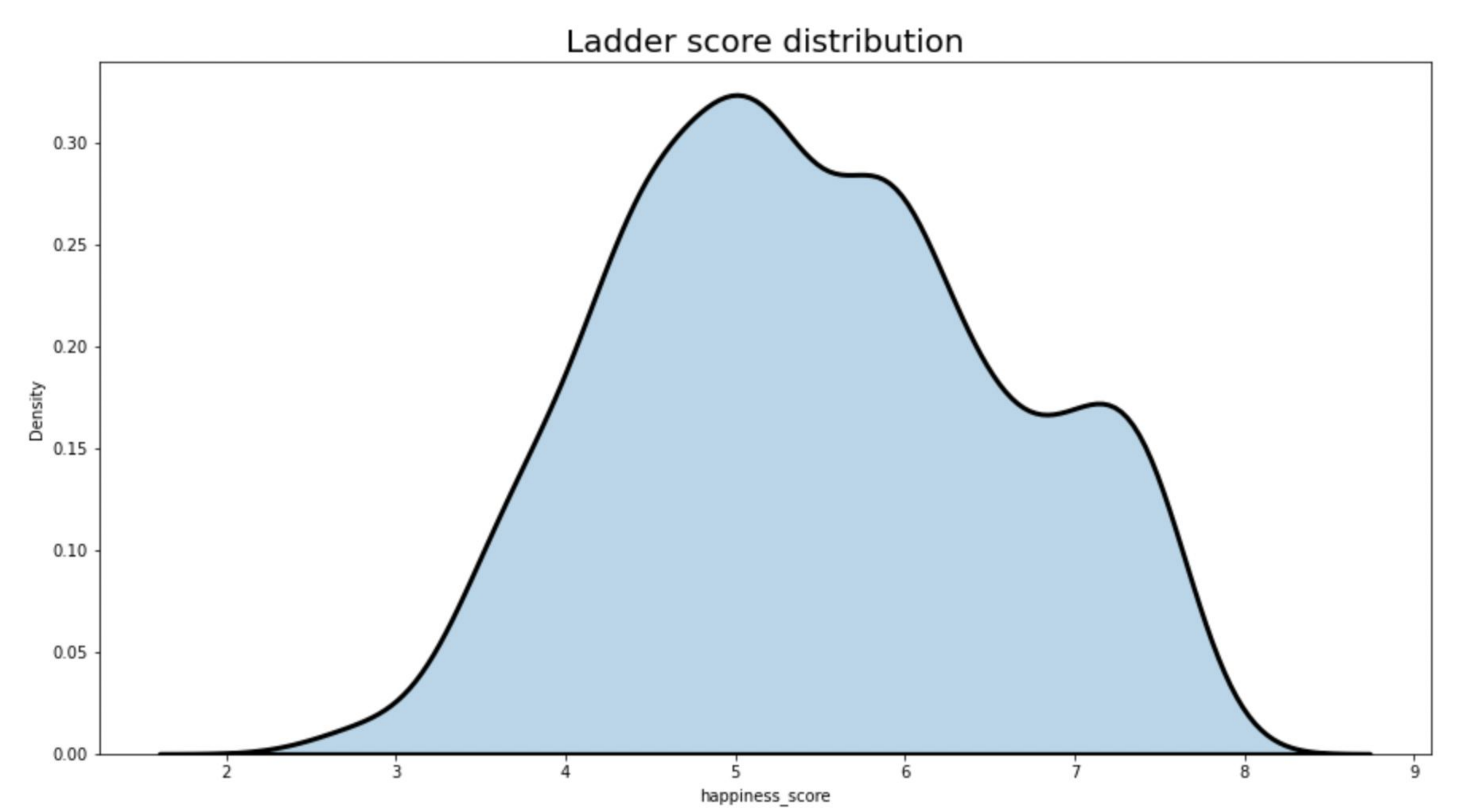
| | country | year | happiness_score | log_gdp | social_support | life_expectancy | freedom | Generosity | corruption |
|---|---|---|---|---|---|---|---|---|---|
| **342** | China | 2006 | 4.560 | 8.696 | 0.747 | 66.88 | NaN | NaN | NaN |
| **343** | China | 2007 | 4.863 | 8.824 | 0.811 | 67.06 | NaN | -0.176 | NaN |
| **344** | China | 2008 | 4.846 | 8.911 | 0.748 | 67.24 | 0.853 | -0.092 | NaN |
| **345** | China | 2009 | 4.454 | 8.996 | 0.798 | 67.42 | 0.771 | -0.160 | NaN |
| **346** | China | 2010 | 4.653 | 9.092 | 0.768 | 67.60 | 0.805 | -0.133 | NaN |
| **347** | China | 2011 | 5.037 | 9.179 | 0.787 | 67.76 | 0.824 | -0.186 | NaN |
| **348** | China | 2012 | 5.095 | 9.249 | 0.788 | 67.92 | 0.808 | -0.185 | NaN |
| **349** | China | 2013 | 5.241 | 9.319 | 0.778 | 68.08 | 0.805 | -0.158 | NaN |
| **350** | China | 2014 | 5.196 | 9.386 | 0.820 | 68.24 | NaN | -0.217 | NaN |
| **351** | China | 2015 | 5.304 | 9.449 | 0.794 | 68.40 | NaN | -0.244 | NaN |
| **352** | China | 2016 | 5.325 | 9.510 | 0.742 | 68.70 | NaN | -0.228 | NaN |
| **353** | China | 2017 | 5.099 | 9.571 | 0.772 | 69.00 | 0.878 | -0.175 | NaN |
| **354** | China | 2018 | 5.131 | 9.632 | 0.788 | 69.30 | 0.895 | -0.159 | NaN |
| **355** | China | 2019 | 5.144 | 9.688 | 0.822 | 69.60 | 0.927 | -0.173 | NaN |
| **356** | China | 2020 | 5.771 | 9.702 | 0.808 | 69.90 | 0.891 | -0.103 | NaN |

# Exploratory Data Analysis

- Countries without happiness scores in Central Africa
- Varies by region
  - North America, Europe highest overall
  - Africa lowest overall
- Clusters with similar score

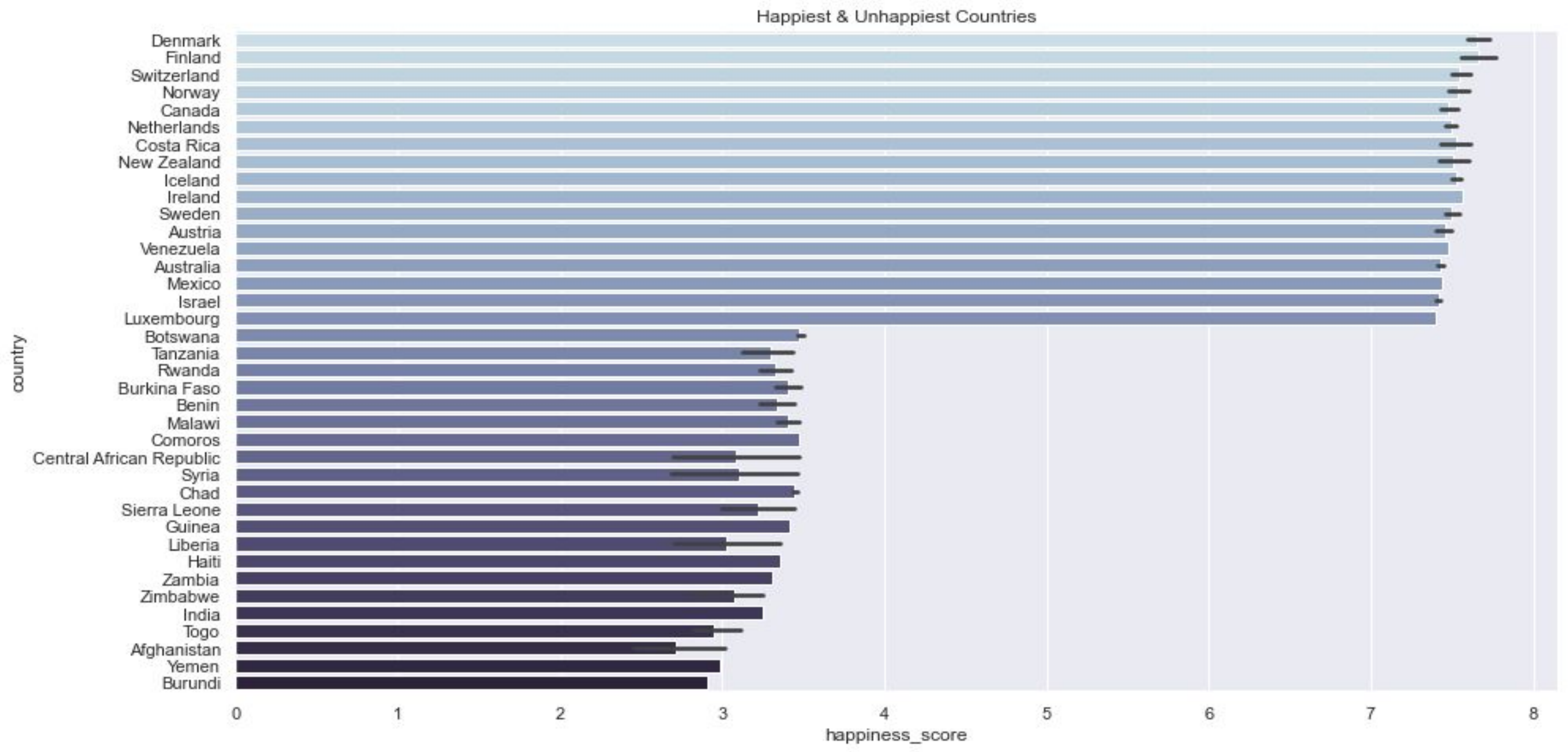


# EDA Continued



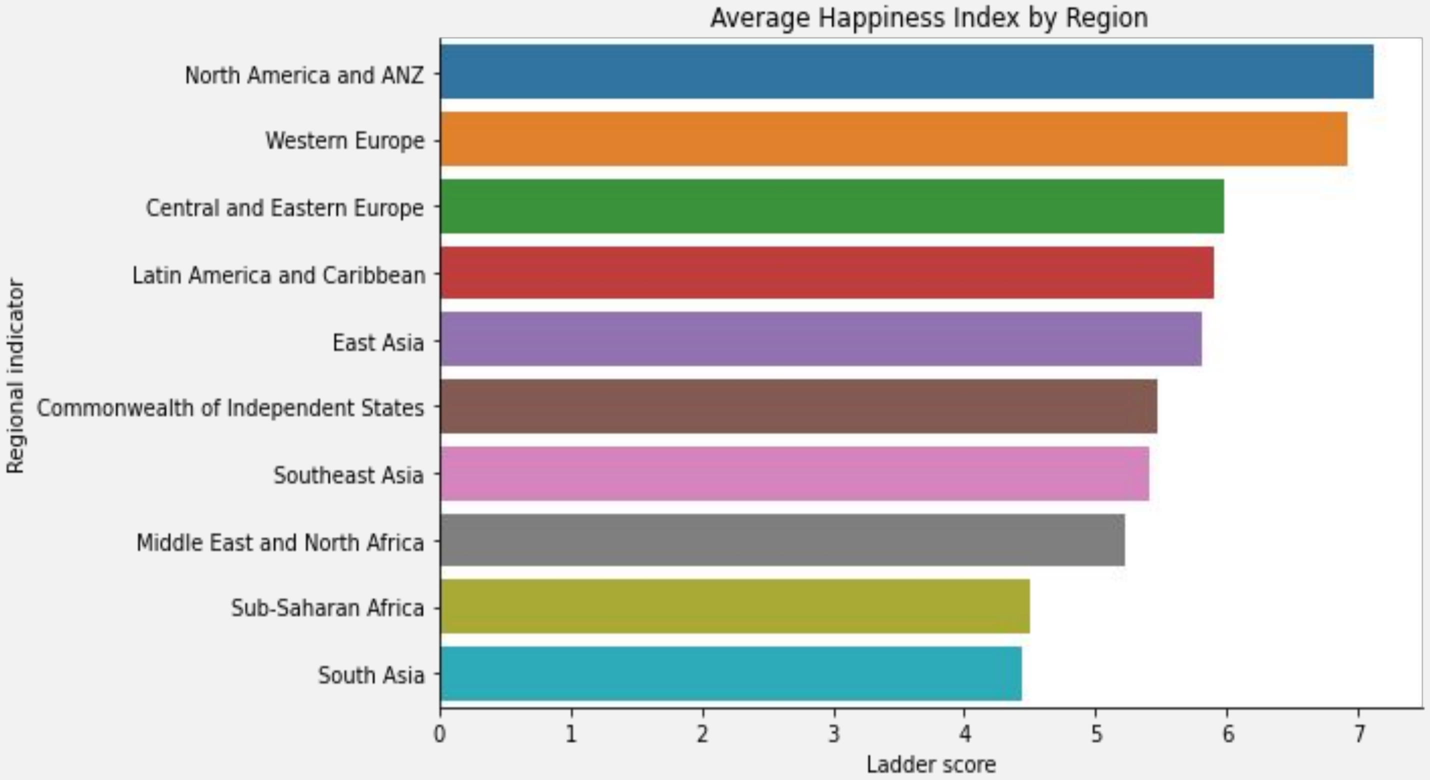
# EDA Continued

These are the top happiest and unhappiest countries based on the happiness score.



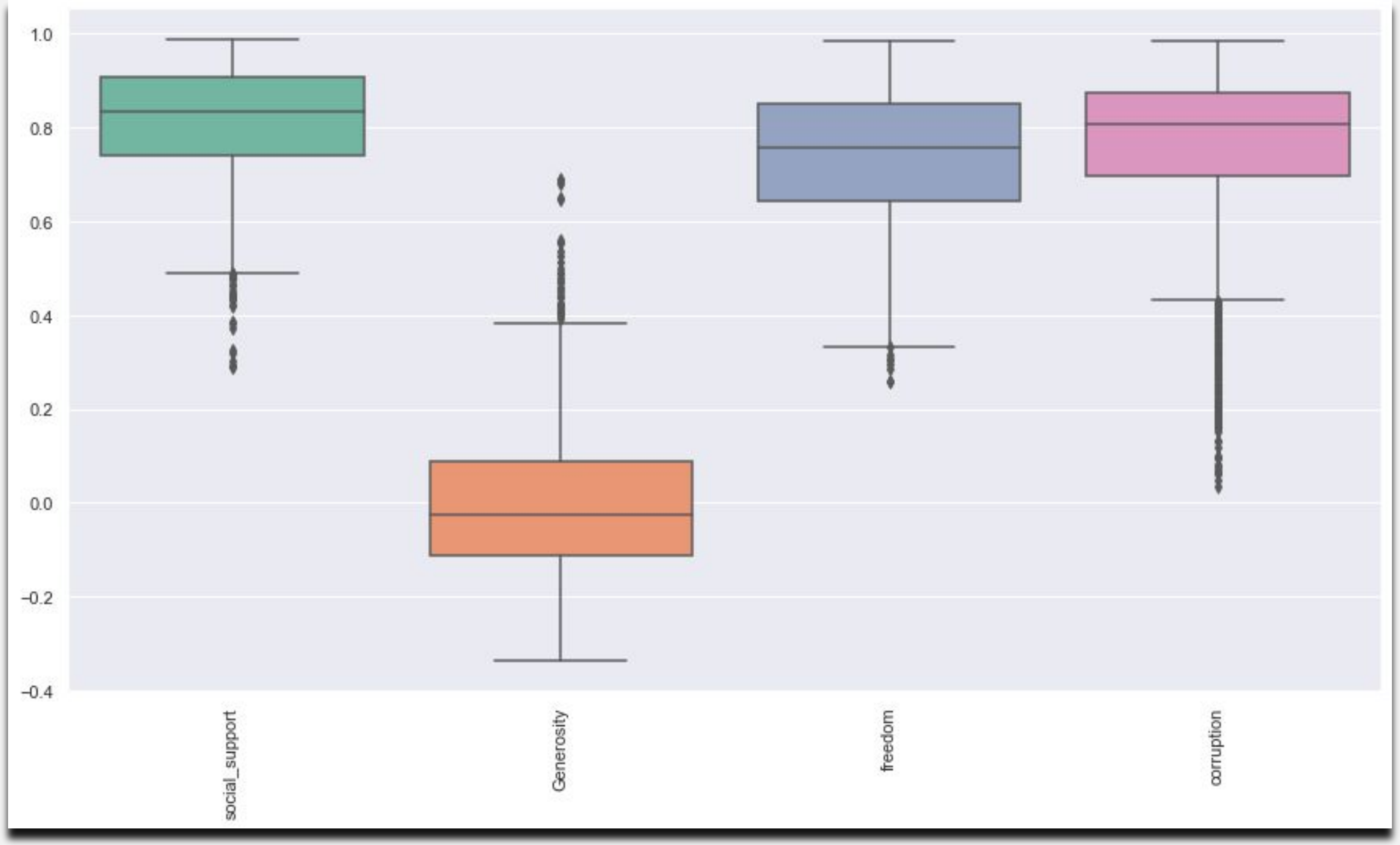
# EDA Continued

- These are the average regional indicator of happiness score across the globe.
- South Asia has the least score
- North America has the happiest ladder score.



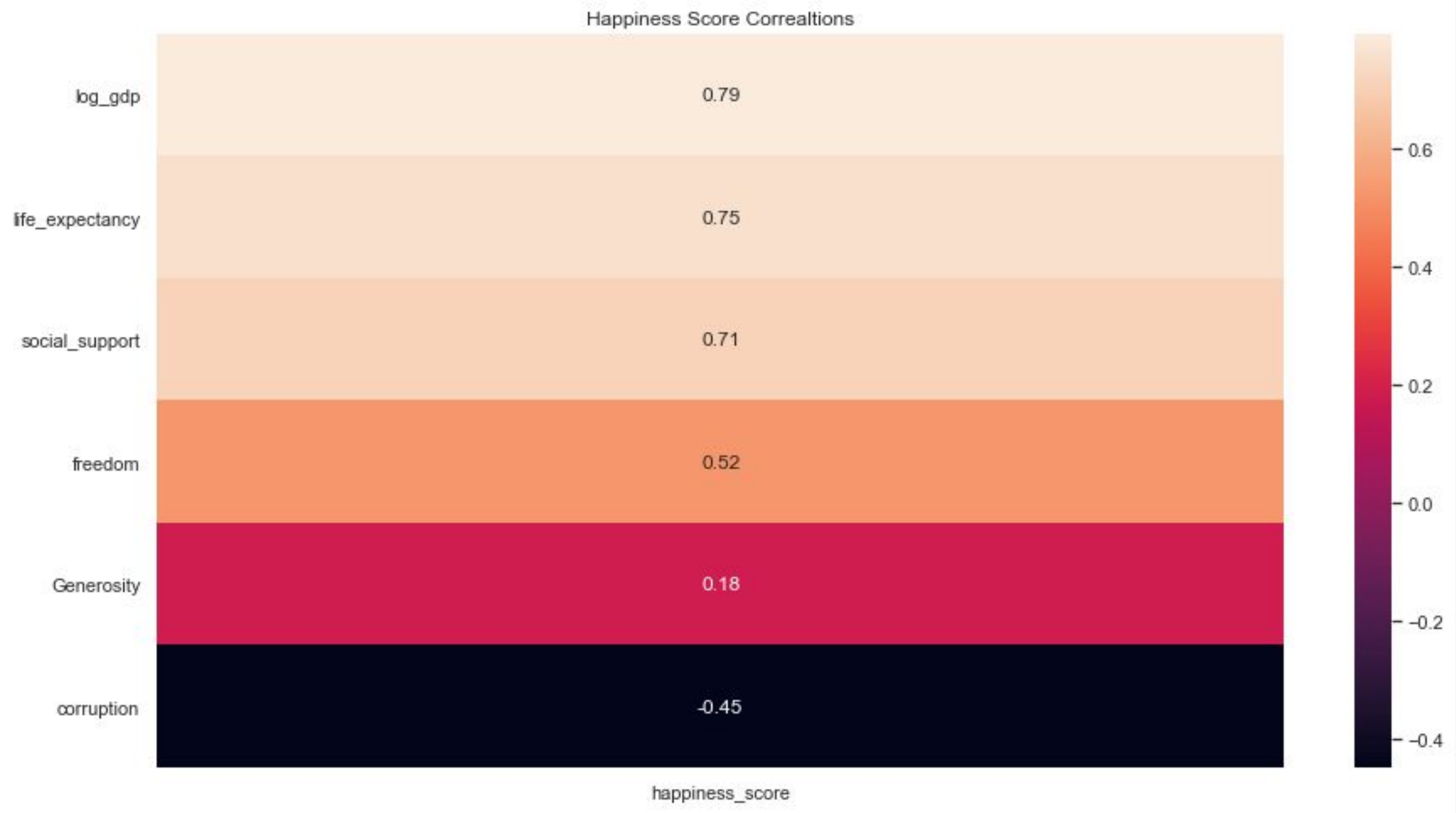
# EDA Continued

This is the distribution of features with a set of 1 and the since they’re all in the same range we use this variable to see the variation of plots.



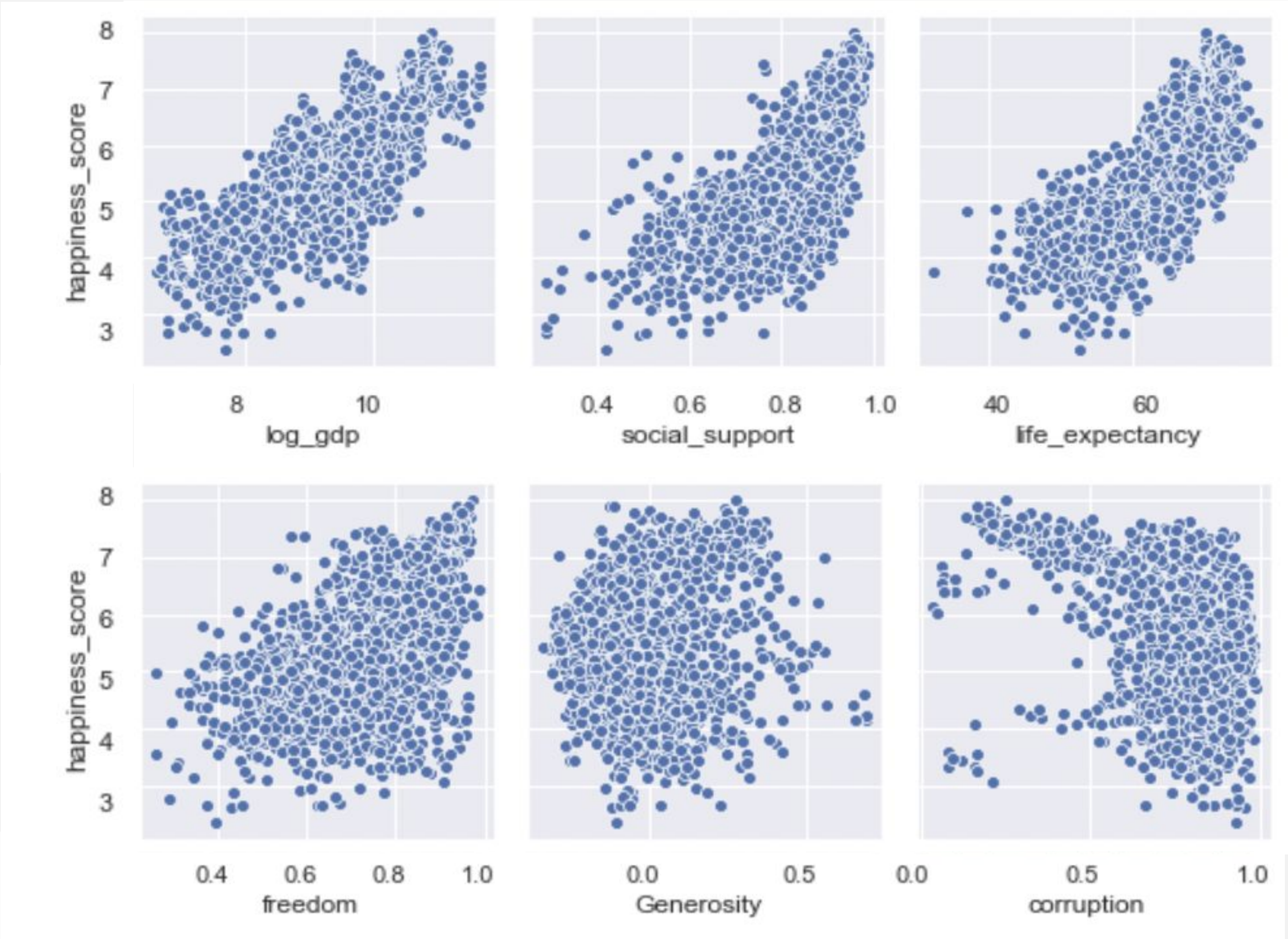
# EDA Continued

- This is a chart that shows correlation between happiness score and other factors
- Corruption has a significant effect on the score



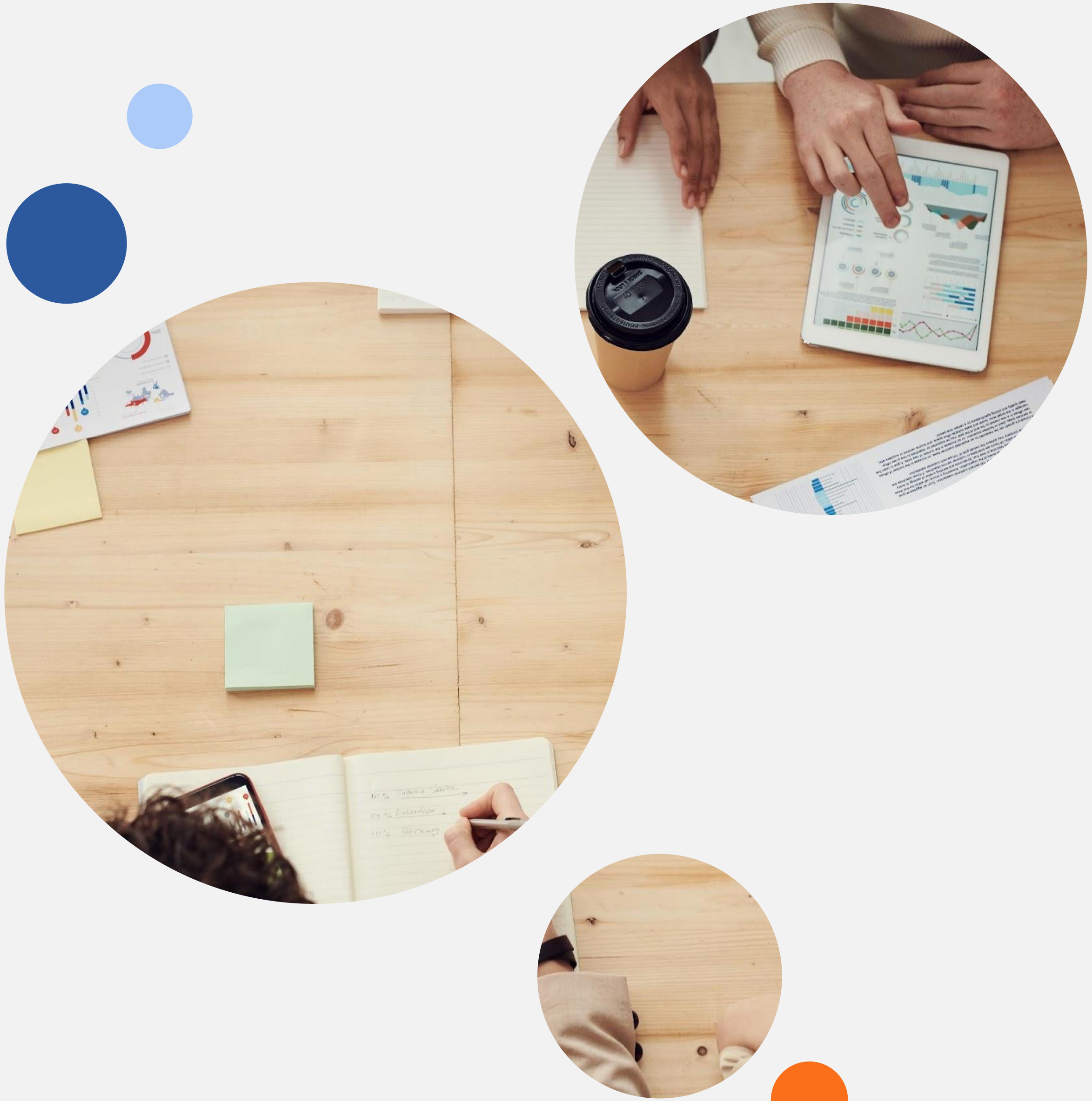
# EDA Continued

- Strong positive association with happiness score
  - log GDP per capita, social support and life expectancy
- Weak positive association with happiness score
  - freedom
- Unclear association with happiness score:
  - generosity
- Negative association with happiness score
  - corruption



# Hypothesis

Based on our EDA, we hypothesize that social support, log GDP, freedom, and life expectancy have a positive correlation and corruption has a negative correlation with the happiness score of a country.

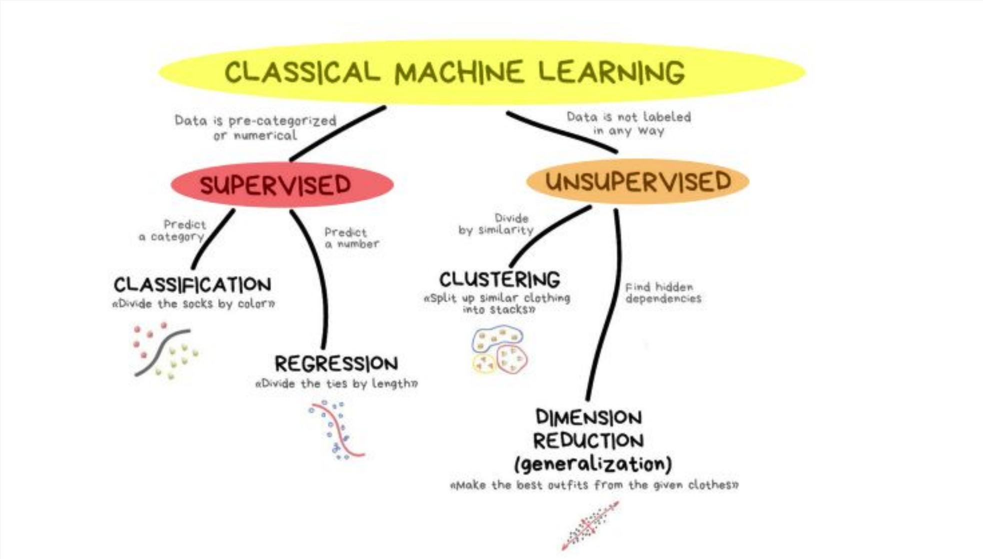


# Multiple Linear Regression

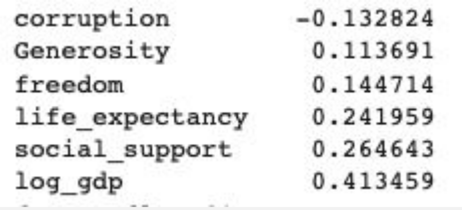
- Predict happiness score
  - Data is numerical and want to predict a numerical value

- Acquire coefficients from the fitted model to see if our hypothesis holds true

- Performed a 80-20 train test split and standardized our features


CLASSICAL MACHINE LEARNING
Data is pre-categorized or numerical
Data is not labeled in any way
SUPERVISED
UNSUPERVISED
Predict a category
Predict a number
Divide by similarity
Find hidden dependencies
CLASSIFICATION
«Divide the socks by color»
REGRESSION
«Divide the ties by length»
CLUSTERING
«Split up similar clothing into stacks»
DIMENSION REDUCTION (generalization)
«Make the best outfits from the given clothes»

**Values from model:**


```
corruption        -0.132824
Generosity         0.113691
freedom            0.144714
life_expectancy    0.241959
social_support     0.264643
log_gdp            0.413459
```

**Correlation Values from EDA:**

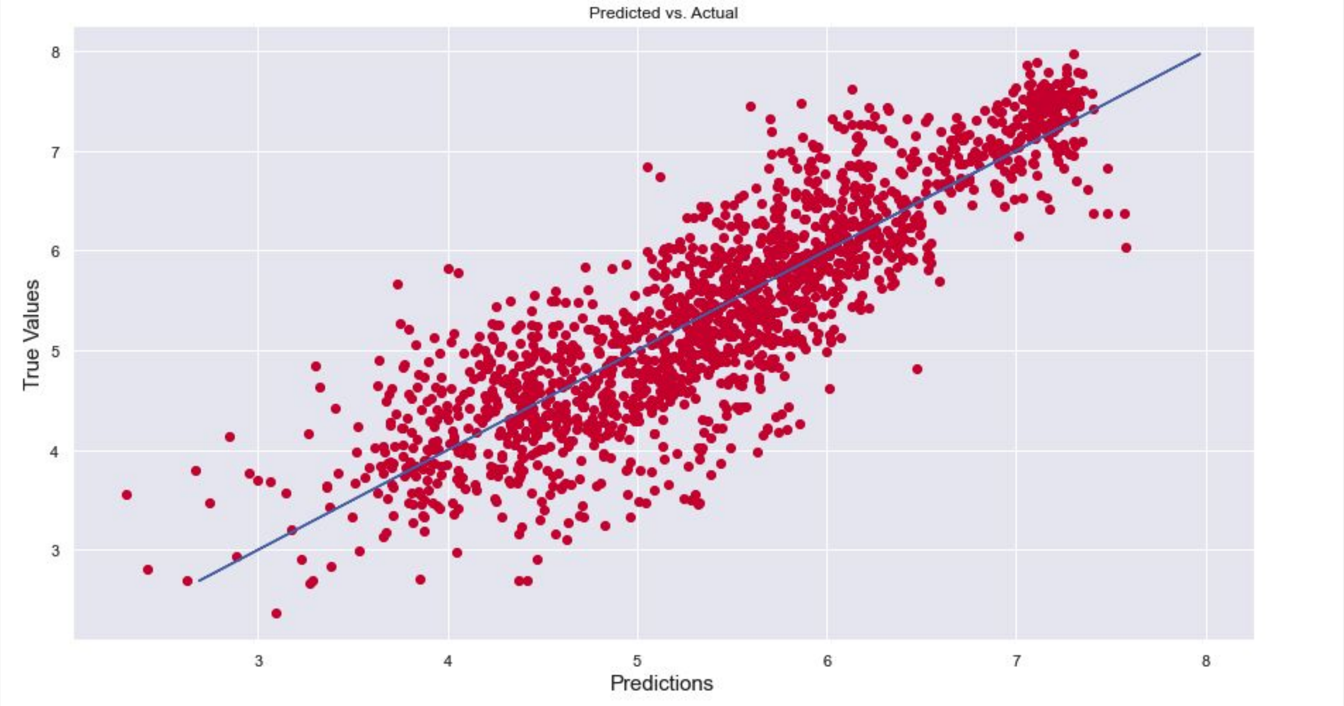
-0.45
0.18
0.52
0.75
0.71
0.79

# Multiple Linear Regression

- Predicted happiness values on testing data
- Linear relationship

**5-fold Cross Validation:**

- MSE for the full model: 0.216
- MSE for the simple model with only log GDP and life expectancy: 0.467
- MSE for the simple model with only log GDP: 0.515

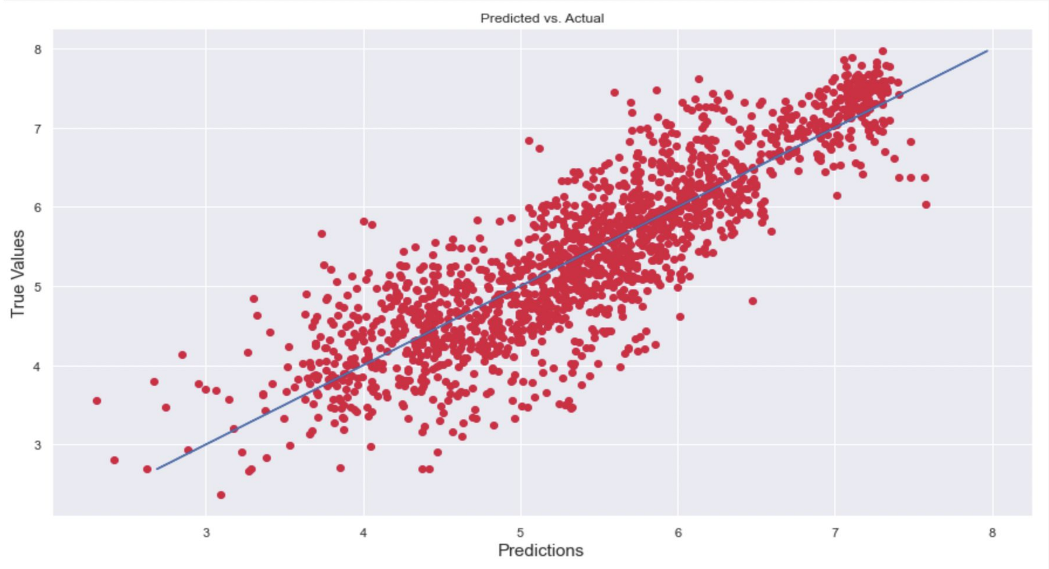


# K Nearest Neighbors

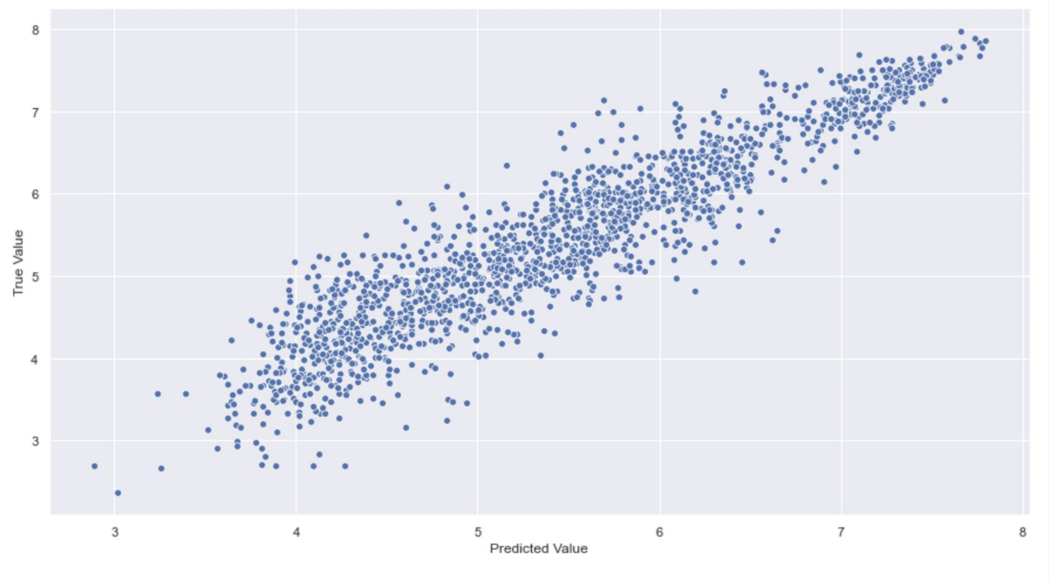
- Compare to linear regression to determine the best model for predicting the happiness score

- Performed 80-20 train test split
  - MSE: 0.29
  - MAE: 0.41
  - R-squared: 0.77

OLS

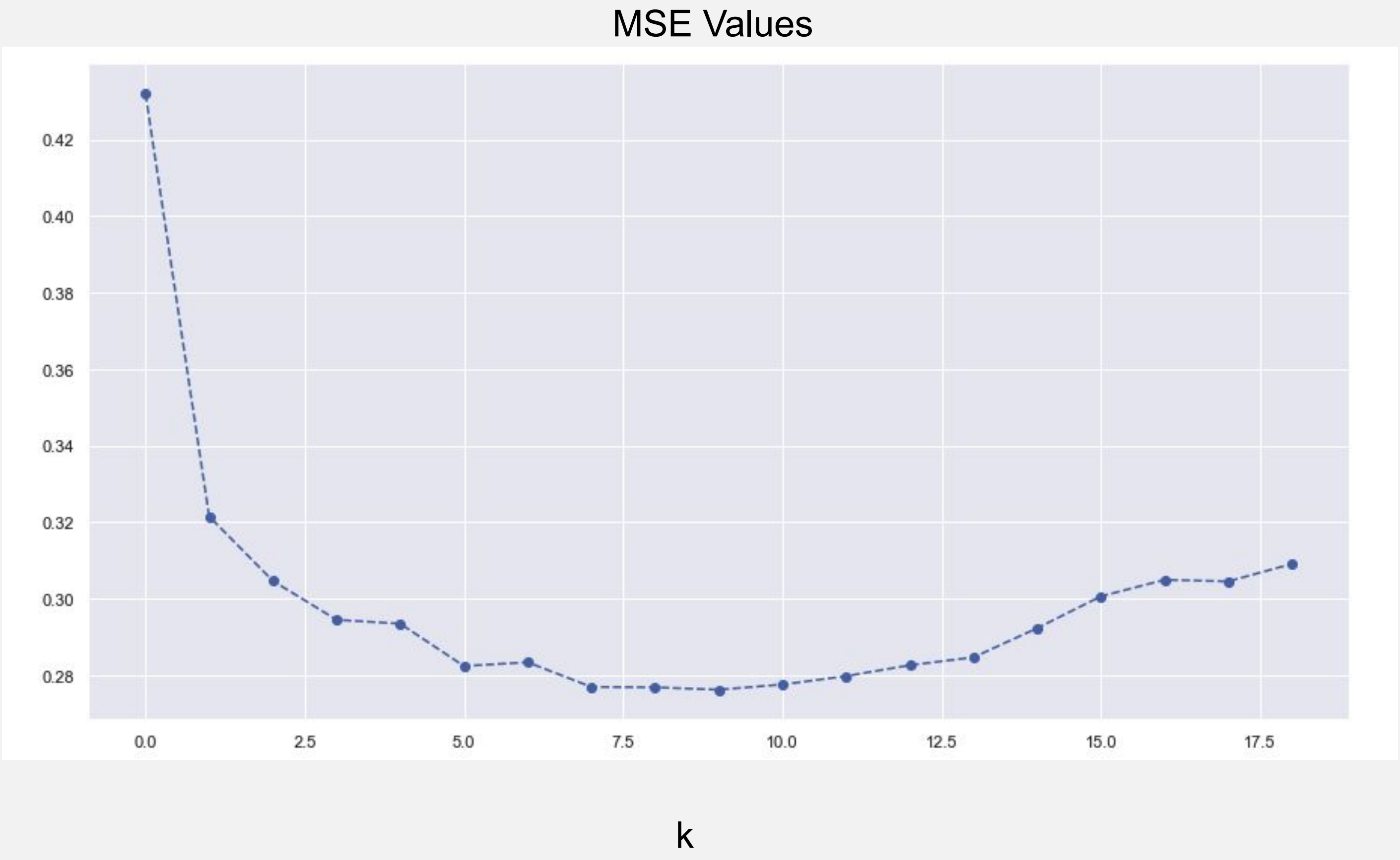


KNN



# KNN

- Iterated through 20 values for k
  - k=9 resulted in the best model
  - MSE: 0.27
  - MAE: 0.40
  - r-squared: 0.78



# Model Evaluation

## Linear Regression

R squared = 0.753

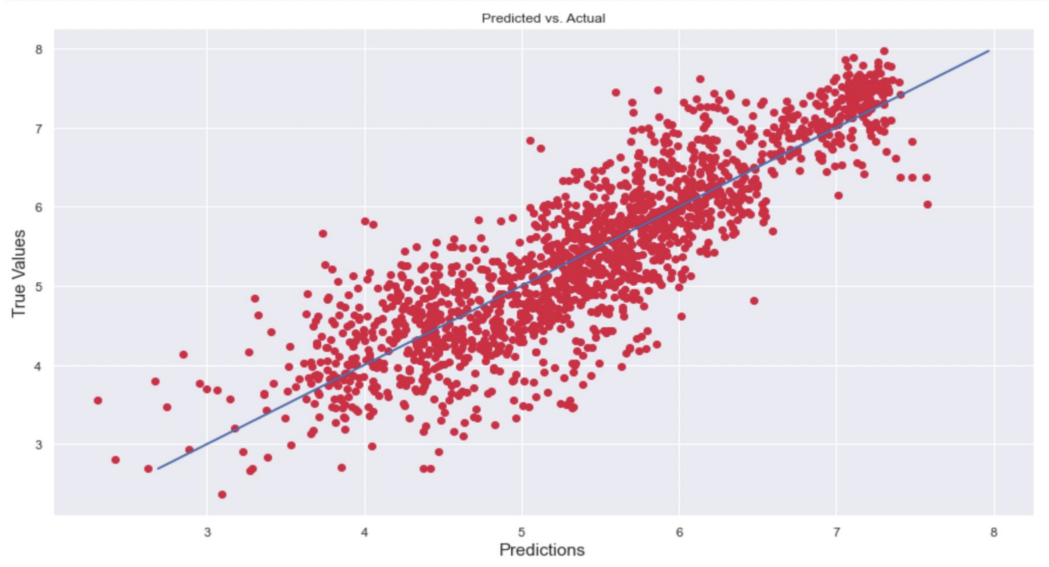
MSE = 0.216

## K-Nearest Neighbors

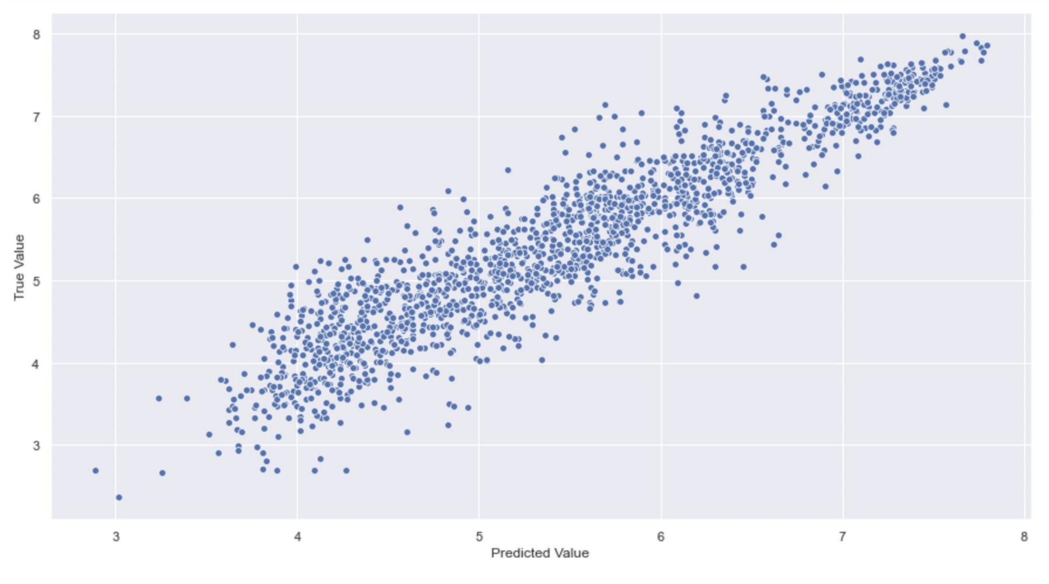
R squared = 0.78

MSE = 0.27

OLS



KNN



# Summary

## Our hypothesis was true

Corruption -

Generosity +

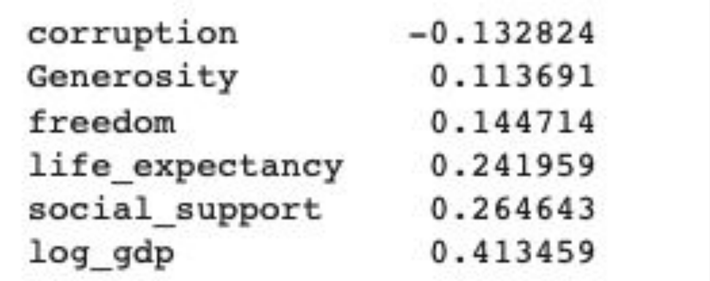
Freedom +

Life Expectancy +

Social Support +

GDP +

## Different relative magnitudes of impact



```
corruption         -0.132824
Generosity          0.113691
freedom             0.144714
life_expectancy     0.241959
social_support      0.264643
log_gdp             0.413459
```

## Both models have similar performance

R squared = 0.753 vs. R squared = 0.771

Linear Regression is better for interpretability

## Next steps and future questions

Running the same models for each year to see how the factors' influence changed overtime

Impact of Covid on happiness score?