World Happiness Report Analysis

Computational Methods and Data Analysis – 4MM451

Andre Boiko

Contents

1. Dataset Overview

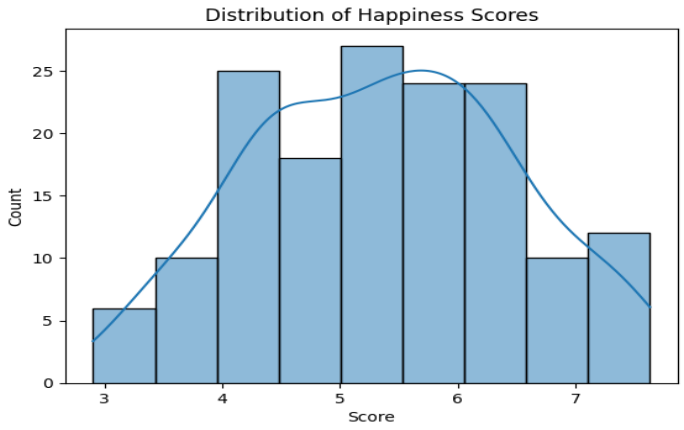
The dataset originates from the *World Happiness Report*, a global survey that ranks 156 countries based on self-reported well-being. The happiness scores are derived from the **Cantril ladder**, where respondents rate their current life on a scale from 0 (worst possible) to 10 (best possible). The data comes from the **Gallup World Poll** and reflects survey results collected between 2013 and 2016.
Each record represents a country and includes the following key variables:

- **Score** – the average life evaluation reported by citizens (*target variable*)
- **GDP per capita** – a measure of economic production per person
- **Social support** – whether individuals feel they have support in times of need
- **Healthy life expectancy** – expected years of life in good health
- **Freedom to make life choices** – perceived autonomy in daily life decisions
- **Generosity** – willingness to help others and donate
- *(Removed)* **Perceptions of corruption** – dropped due to missing data

The six explanatory variables were designed to account for differences in happiness scores between countries by comparing each nation to a theoretical baseline called **Dystopia –** benchmark country with lowest global scores.

2. Descriptive Statistics

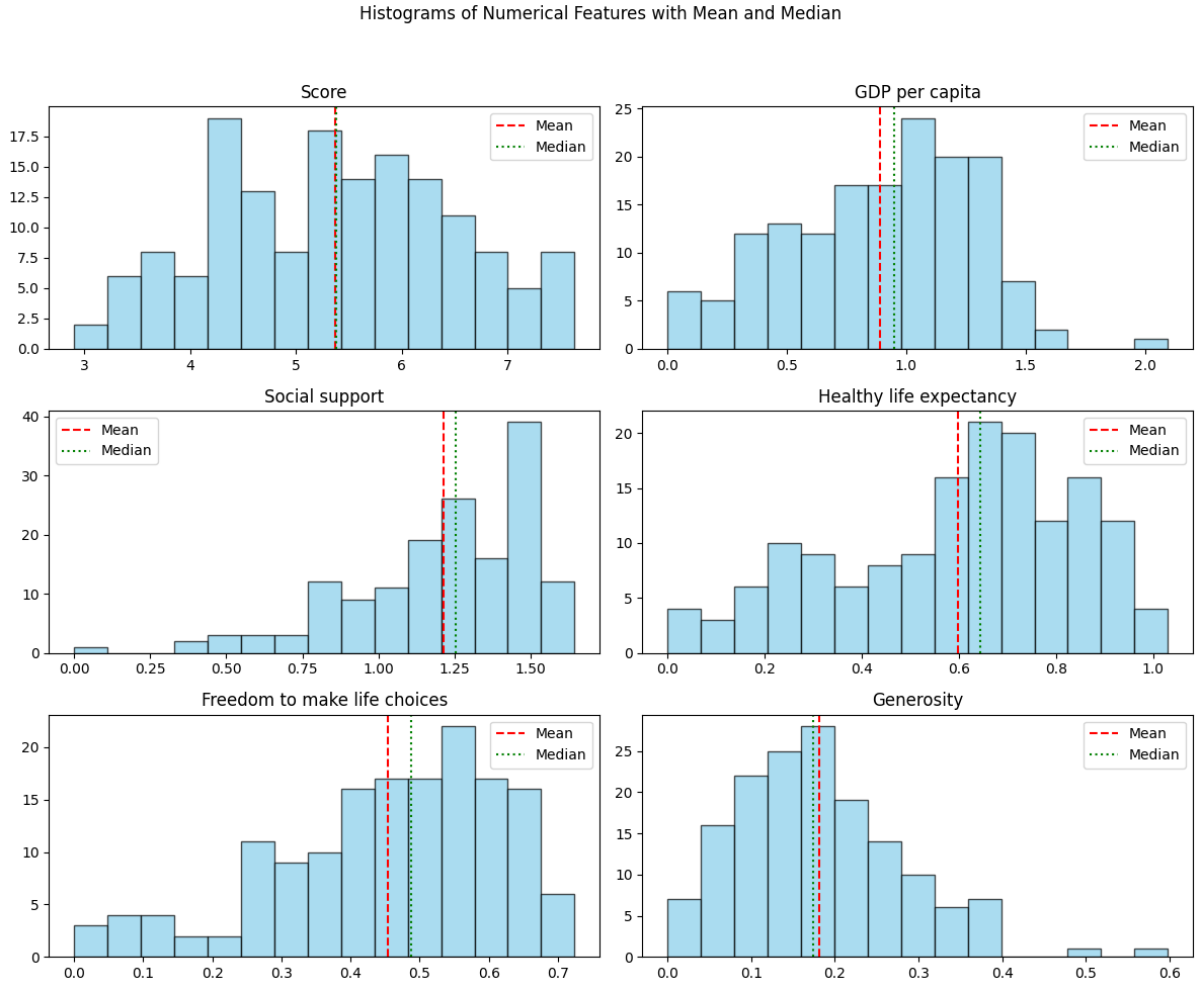
	Overall rank	Score	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
count	156.000000	156.000000	156.000000	156.000000	156.000000	156.000000	156.000000	155.000000
mean	78.500000	5.375917	0.891449	1.213237	0.597346	0.454506	0.181006	0.112000
std	45.177428	1.119506	0.391921	0.302372	0.247579	0.162424	0.098471	0.096492
min	1.000000	2.905000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	39.750000	4.453750	0.616250	1.066750	0.422250	0.356000	0.109500	0.051000
50%	78.500000	5.378000	0.949500	1.255000	0.644000	0.487000	0.174000	0.082000
75%	117.250000	6.168500	1.197750	1.463000	0.777250	0.578500	0.239000	0.137000
max	156.000000	7.632000	2.096000	1.644000	1.030000	0.724000	0.598000	0.457000



- Mean happiness score is 5.38, ranges from 2.91 to 7.63.
- GDP per capita: 0.00 to 2.10, high standard deviation, reflecting global economic inequality.
- Social support and healthy life expectancy show tighter clustering but still meaningful variation.
- Happiness scores are fairly symmetrically distributed, with most countries scoring between 4.5 and 6.5.

2. Exploratory Data Analysis (EDA)

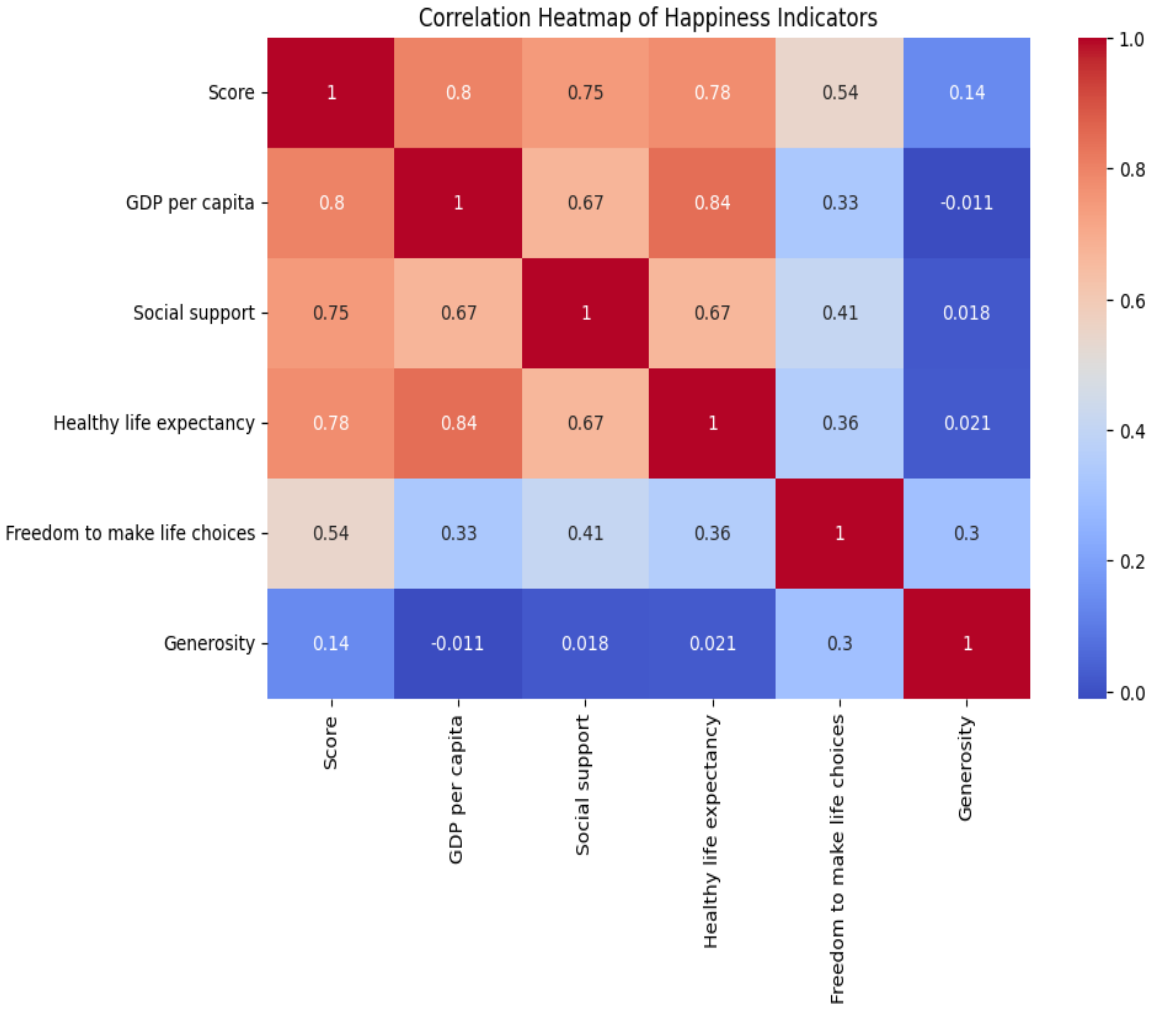
- Mean (red) and Median (green) help reveal skewness
- GDP per capita and Generosity are right-skewed
- Happiness Score is fairly symmetric
- Skewed variables may impact regression results
  - Skewed variables can distort regression results by violating assumptions, amplifying outliers, and making interpretation less reliable.



3. Correlation Analysis

Notable correlations:

- `*GDP per capita*` and `*Healthy life expectancy*` are highly correlated (0.79)
- `*Social support*` and `*Score*` also show strong correlation (0.74)
- This suggests that economic and health-related indicators are key drivers of happiness.



4. Hypothesis Testing

```
median_gdp = happiness['GDP per capita'].median()
high_gdp = happiness[happiness['GDP per capita'] > median_gdp]['Score']
low_gdp = happiness[happiness['GDP per capita'] <= median_gdp]['Score']
ttest_result = stats.ttest_ind(high_gdp, low_gdp, equal_var=False)
ttest_result
```

```
TtestResult(statistic=np.float64(11.67215758822544), pvalue=np.float64(6.027279635180903e-23), df=np.float64(153.8723723660498))
```

H0: No difference in happiness between high- and low-GDP countries

Result: p-value < 0.05 → Reject H0

Conclusion: High-GDP countries tend to be significantly happier.

5. Multiple Linear Regression

```
X_reg = sm.add_constant(X)
y_reg = happiness.loc[X.index, 'Score']
model = sm.OLS(y_reg, X_reg).fit()
model.summary()
```

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.8559	0.194	9.573	0.000	1.473	2.239
GDP per capita	1.1261	0.209	5.383	0.000	0.713	1.539
Social support	0.9813	0.201	4.880	0.000	0.584	1.379
Healthy life expectancy	0.8468	0.330	2.566	0.011	0.195	1.499
Freedom to make life choices	1.5020	0.303	4.957	0.000	0.903	2.101
Generosity	0.7572	0.453	1.671	0.097	-0.138	1.653

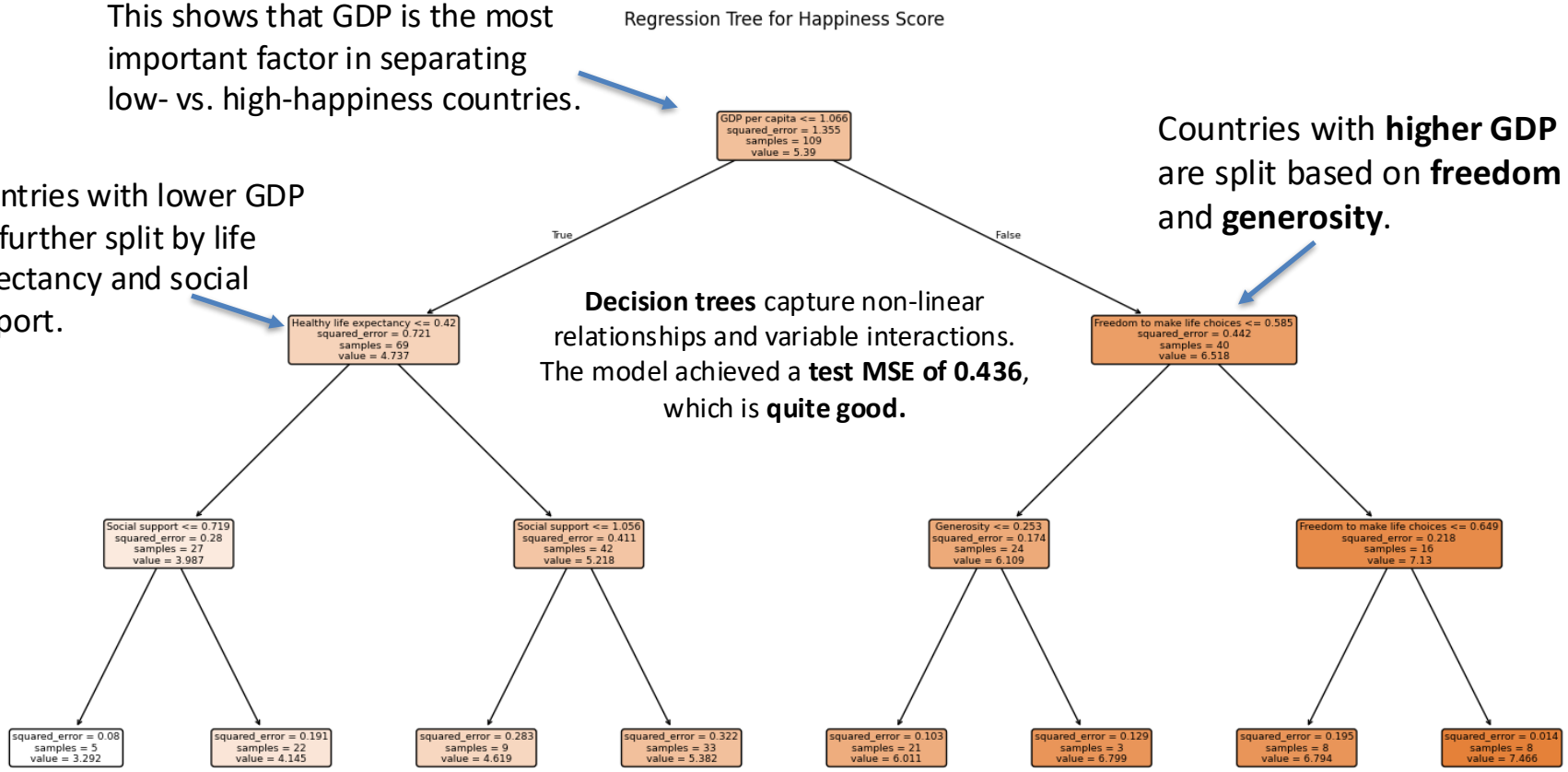
OLS Regression Results

Dep. Variable:	Score	**R-squared:**	0.787
Model:	OLS	**Adj. R-squared:**	0.780
Method:	Least Squares	**F-statistic:**	110.7
Date:	Sun, 04 May 2025	**Prob (F-statistic):**	1.59e-48
Time:	15:58:24	**Log-Likelihood:**	-117.91
No. Observations:	156	**AIC:**	247.8
Df Residuals:	150	**BIC:**	266.1
Df Model:	5		
Covariance Type:	nonrobust		

- The model has a strong fit ($R^2 = 0.787$)
- Significant predictors: **GDP per capita**, **Social support**, **Healthy life expectancy**, and **Freedom to make life choices**
- **Generosity** had a weaker effect and was not statistically significant at the 5% level

7. Regression Tree

This shows that GDP is the most important factor in separating low- vs. high-happiness countries.



Countries with lower GDP are further split by life expectancy and social support.

Countries with **higher GDP** are split based on **freedom** and **generosity**.

Decision trees capture non-linear relationships and variable interactions. The model achieved a **test MSE of 0.436**, which is **quite good.**

The final leaf nodes show **predicted happiness scores** for each subgroup.

score of ~7.4 for countries with **high GDP, high freedom**, and **good support**.

8. Conclusion

- High **GDP per capita**, **freedom**, and **social support** are the most influential variables
- Both regression models confirm consistent findings across methods
- The regression tree achieves a solid **MSE of 0.436**, indicating good predictive performance
  - RMSE: 0.66 - model's predictions are off by about 0.66 happiness points: reasonable (happiness 0-10 scale)
- While results are interpretable and aligned with theory, caution is needed due to possible endogeneity and omitted variables
  - explanatory variable is correlated with the error term
  - If happier people are also more productive, and that boosts GDP → Then GDP and happiness are mutually influencing each other meaning that GDP isn't strictly independent, and this creates endogeneity.
  - Or if a third variable (let's say "education" affects both GDP and happiness)
  - As we are doing just a explanatory analysis → we sohuldnt' worry

Thank you.

Computational Methods and Data Analysis – 4MM451

Andre Boiko