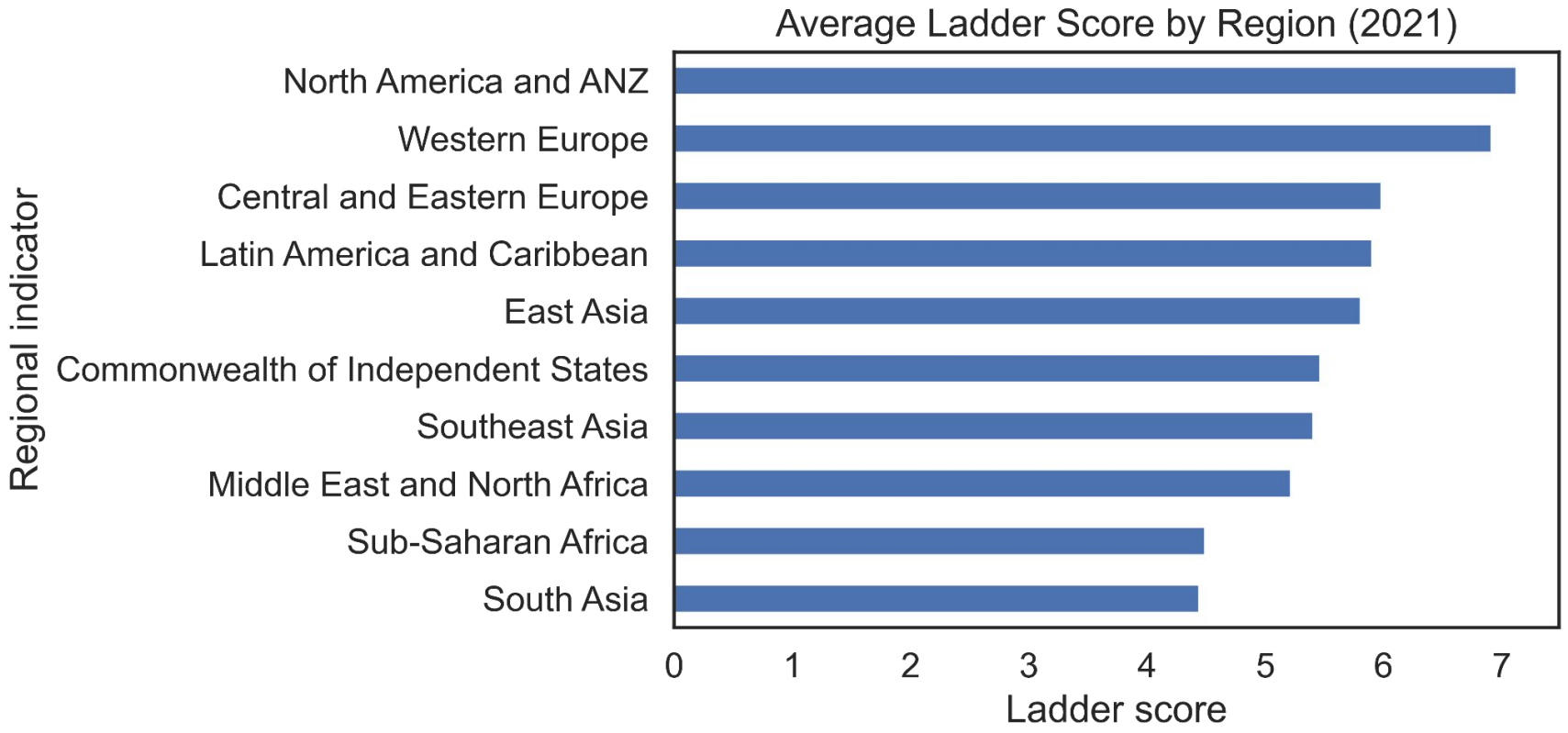
World Happiness Index

Varun Saini, Samson Dorfman, Tasfiq Ahmed

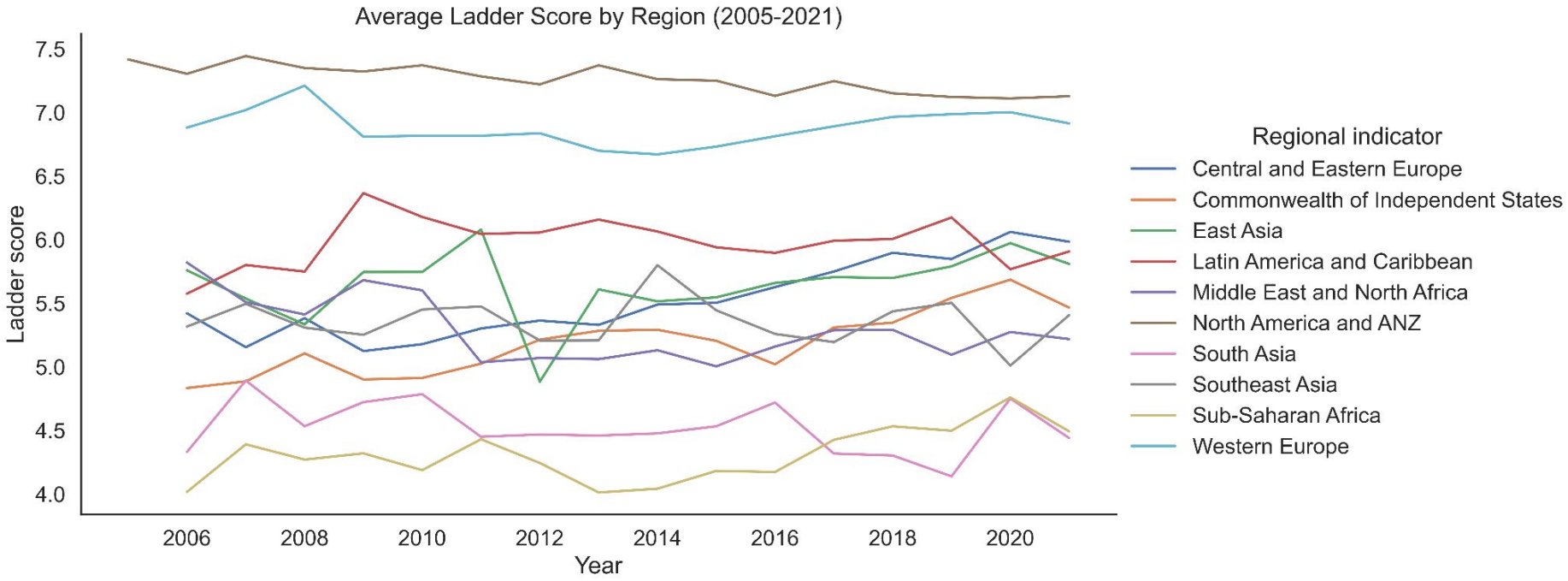
Overview

- **Goals**
  - Examine happiness (measured by ladder score) around the world
    - Compare regional averages
    - Create/compare models to predict ladder score given feature values
    - Pre-Covid vs. Post-Covid analysis
  - Group countries based on features (K-Means Clustering)
  - Consider impact of COVID
- **Features:** log GDP per capita, healthy life expectancy, social support, freedom (to make choices), generosity, perceptions of corruption
- **Target:** Ladder (happiness) score between 1 and 10
- **Dataset**: from World Happiness Reports
  - 1816 rows (after dropping missing values): 149 unique countries, data from 2005–2021

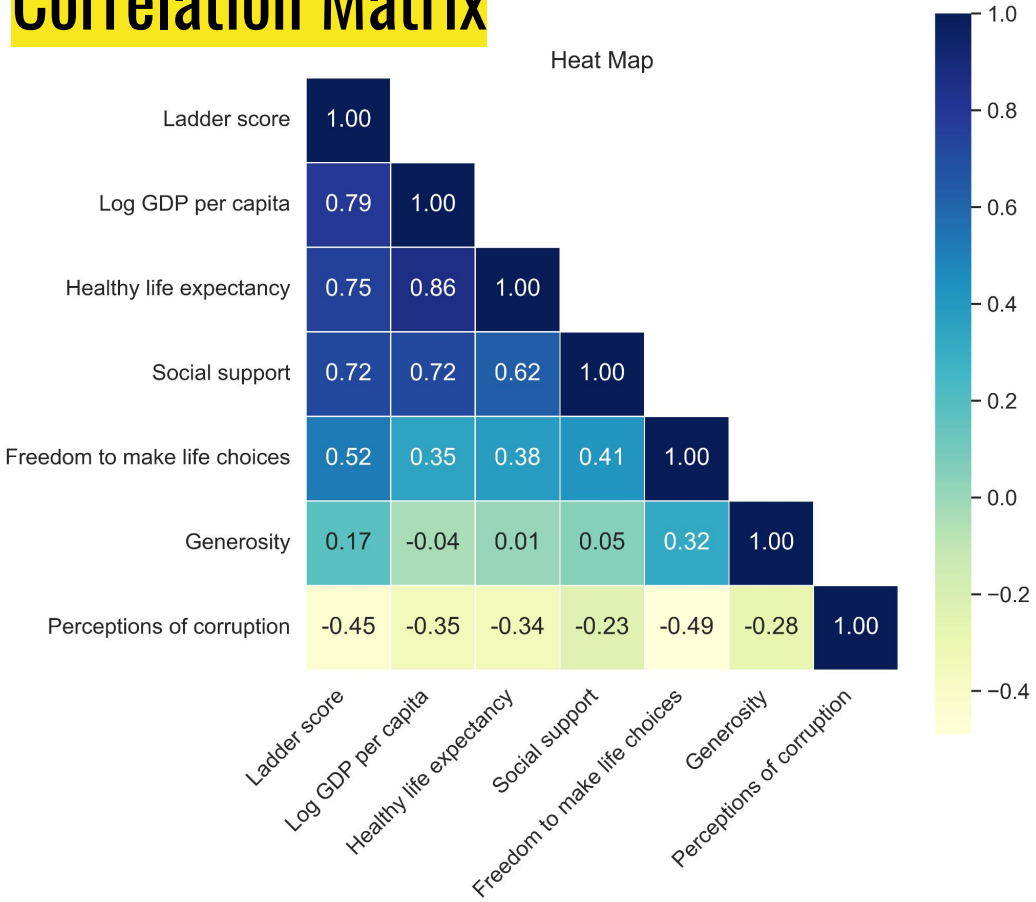
Happiness Index by Region

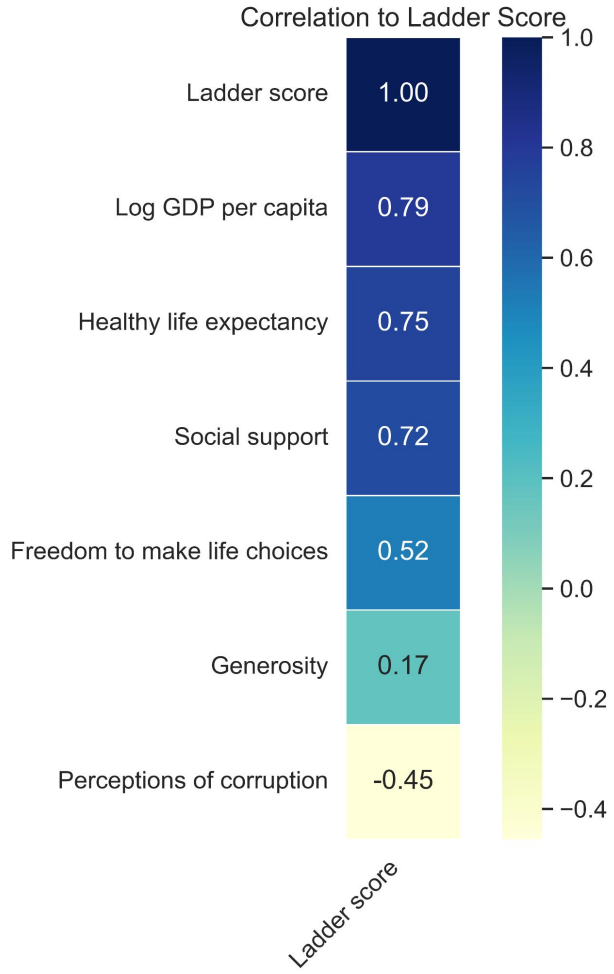


Change in Regions: 2005 – 2021



Correlation Matrix





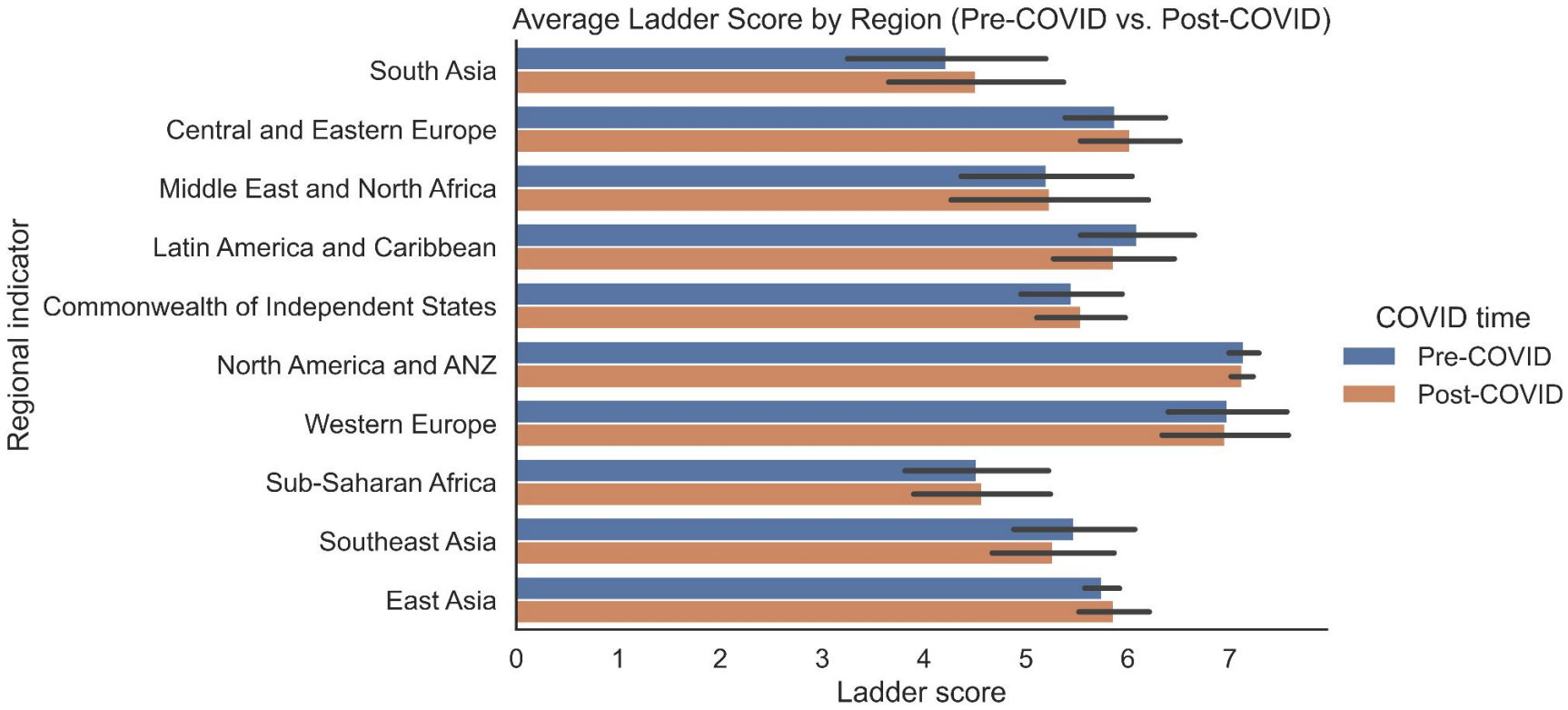
Missing Values

- Missing values in perceptions of corruption, generosity, and healthy life expectancy
- Perceptions of corruption
  - Constituted the greatest proportion of missing values
  - Objectivity of the feature also in question
- Began approach by dropping missing values, leaving 1816 rows

Pre vs Post COVID Analysis

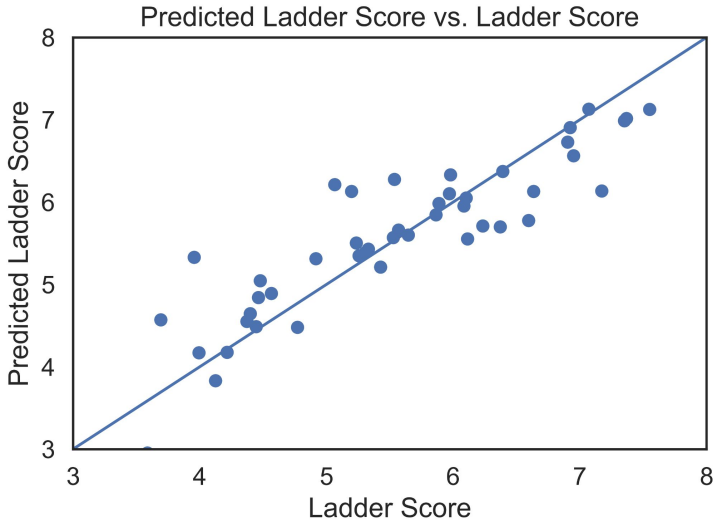
- Look at the impact Covid has had on our features
- Not a symmetric comparison - more data is available prior to the pandemic than after
- Ultimately, not enough data is present for a meaningful analysis

Pre vs Post COVID Analysis



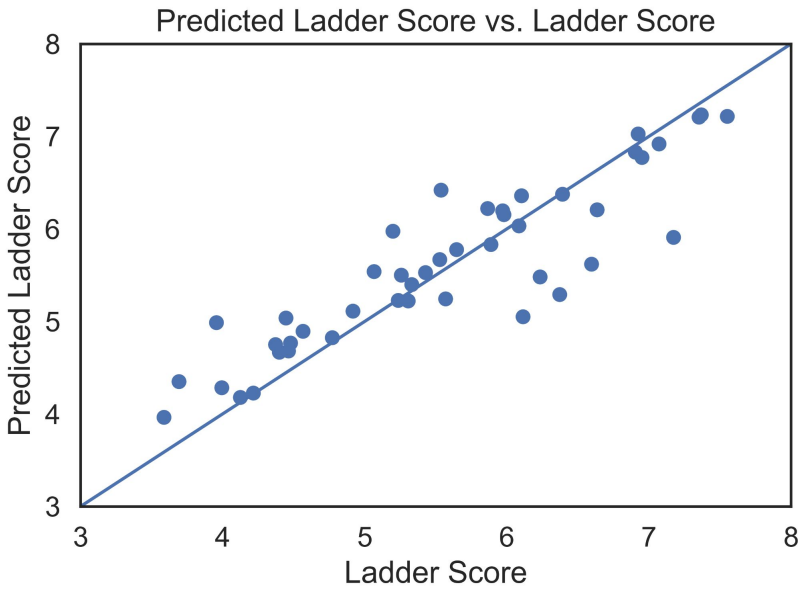
Multiple Linear Regression

- Scaling necessary as features are on different scales (for feature importance)
- As expected from the correlation matrix, the coefficients on log GDP per capita, healthy life expectancy, and social support were the highest
  - Small, negative coefficient on corruption
- *MSE* : 0.24
- *MAE*: 0.36
- R^2 : 0.79

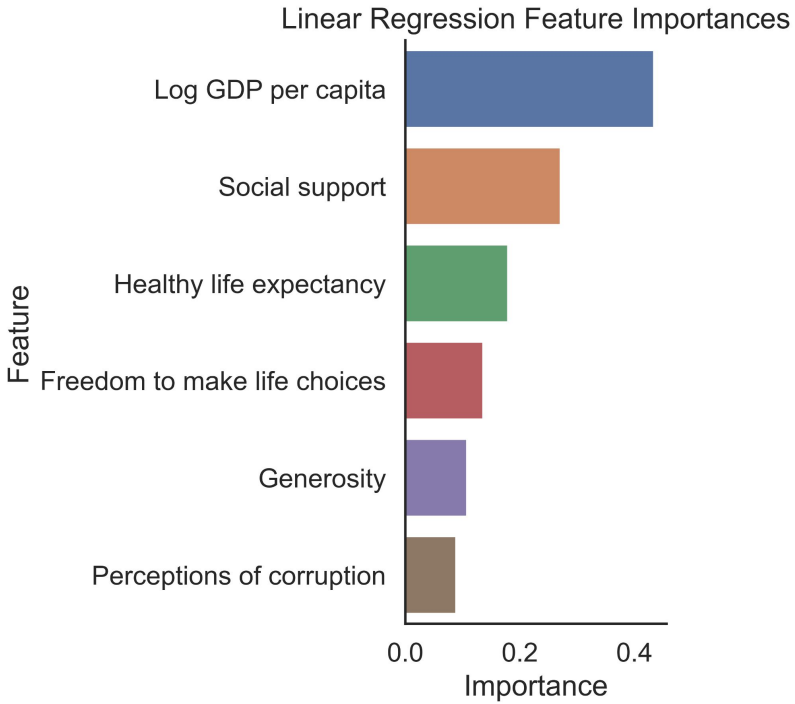


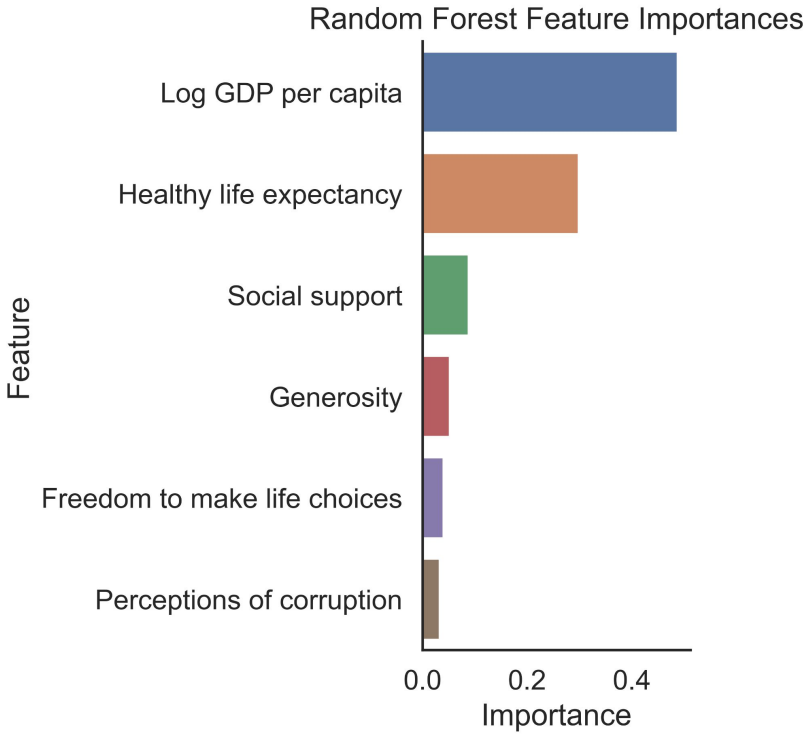
Random Forest

- Hyperparameters
  - Number of estimators : 22
  - Max leaf nodes : 30
  - Max depth : 10
- *MSE*: 0.21
- *MAE*: 0.33
- R^2 : 0.82



Feature Importances

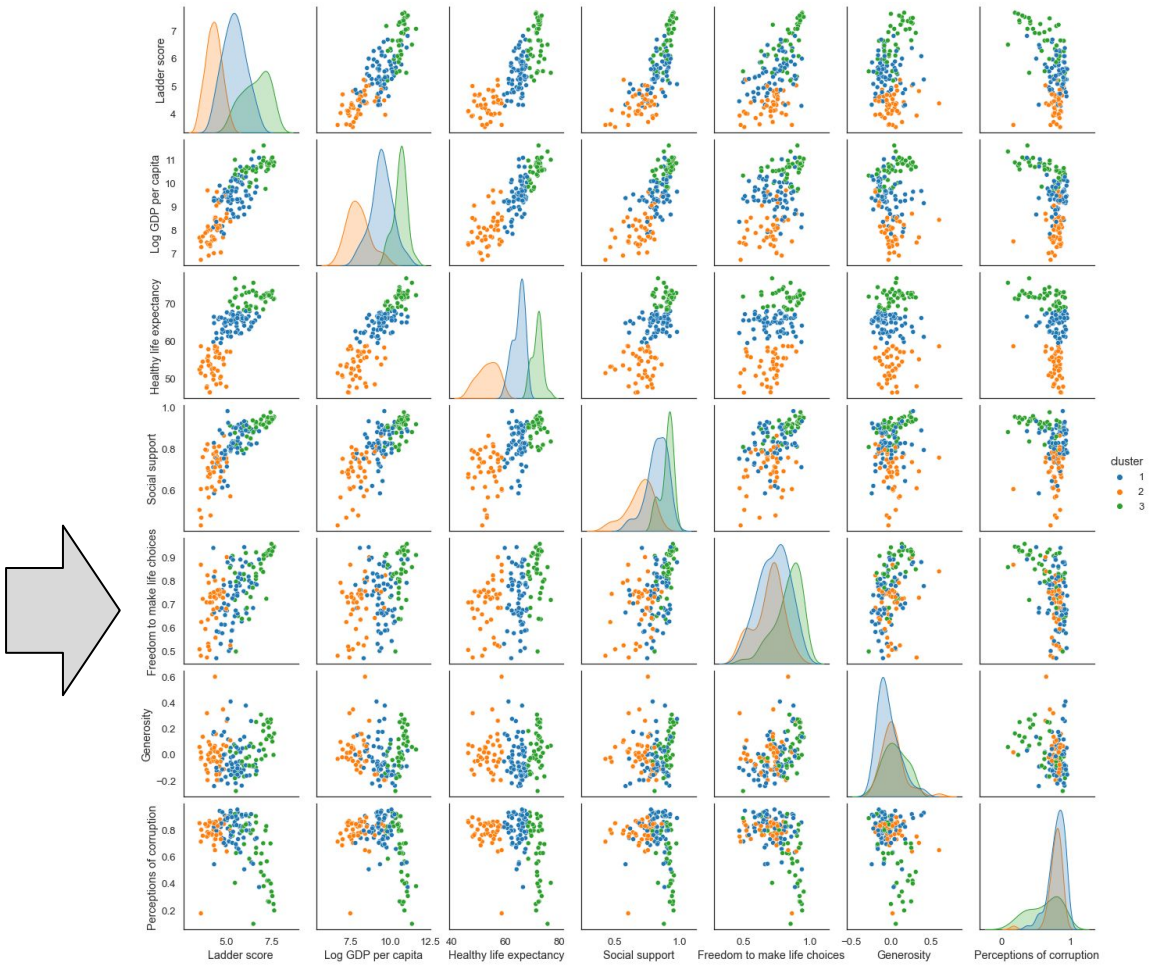




How can countries be grouped by features?

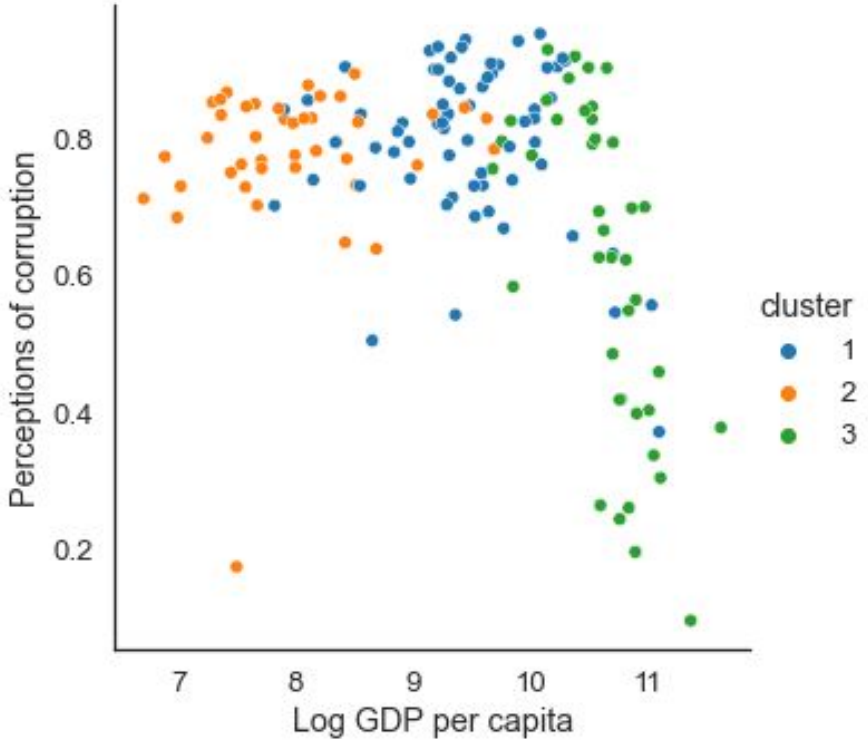
Clustering

1. **Using K-Means Clustering (with n=3) on the data**
2. **Generating a pair plot with colored clusters**
3. **Focusing on corruption and Log GDP per Capita (next slide)**



How can countries be grouped by features?

Clustering



The countries can all be sorted into three clusters.

The differences become more clear when looking at log GDP per capita (wealth) vs. perceptions of corruption.

KEY

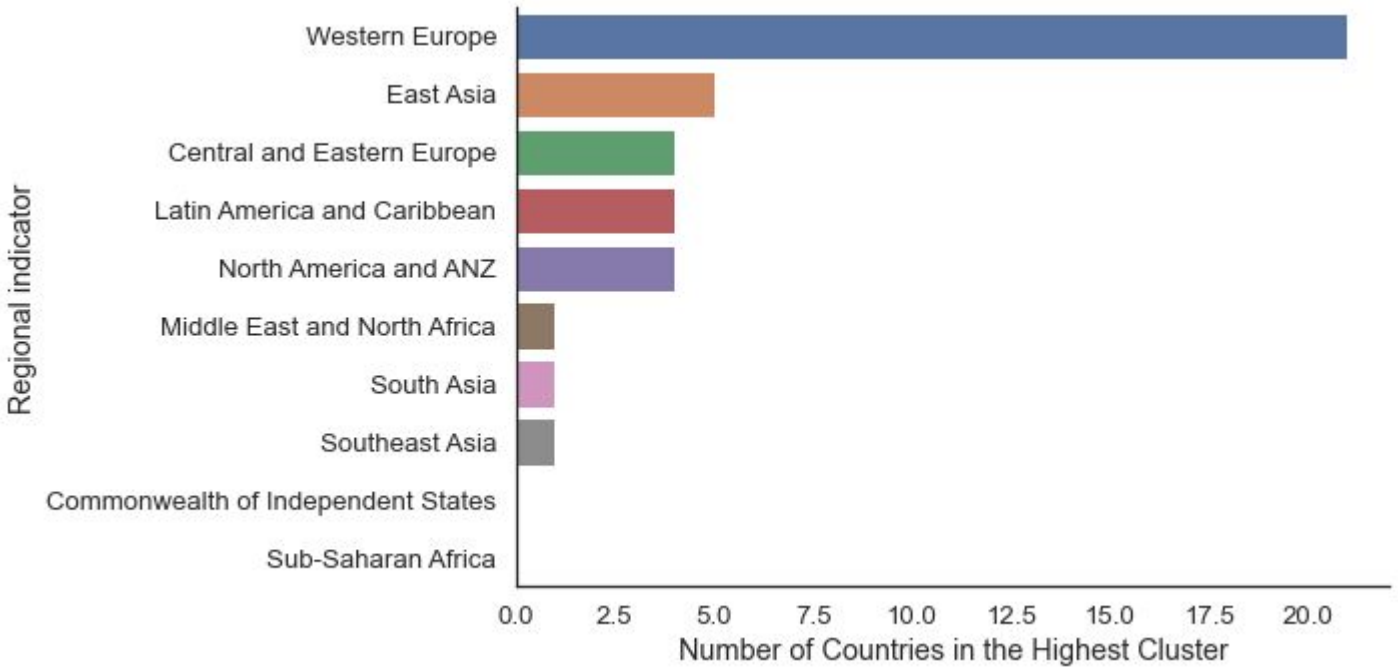
Poor, corrupt

Medium rich, corrupt

Rich, not (as) corrupt

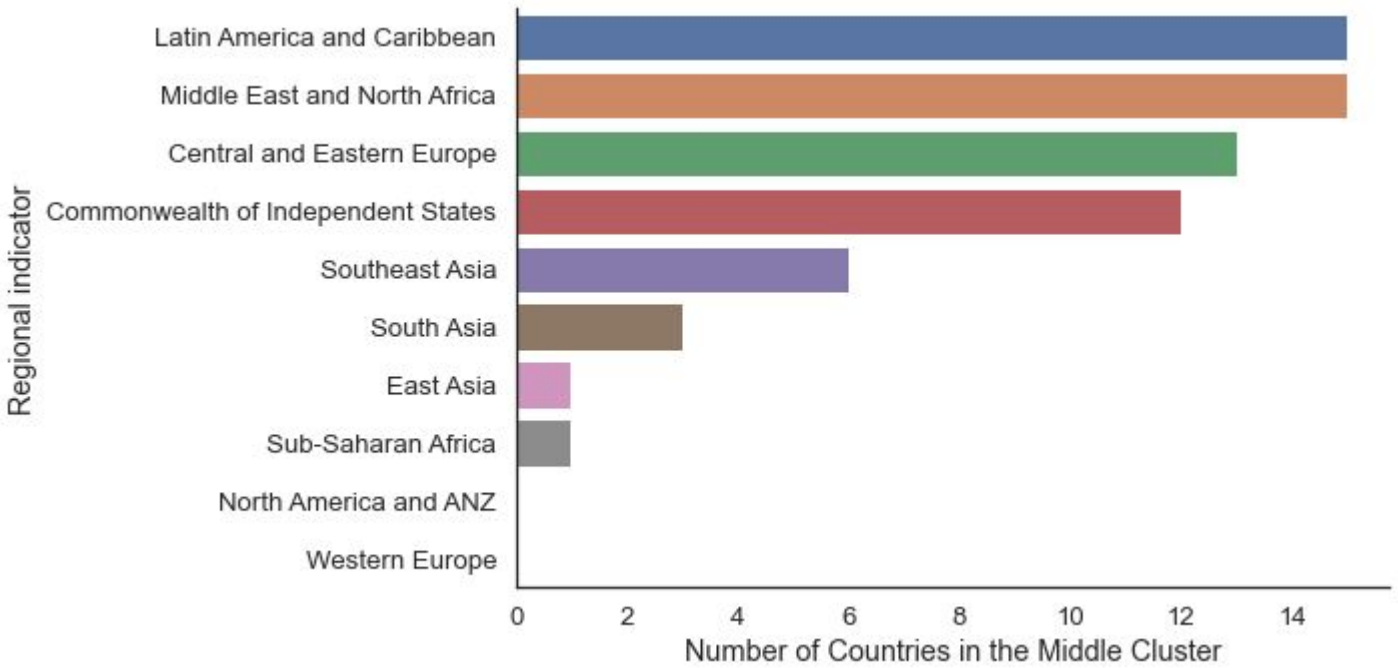
How can countries be grouped by features?

“Rich, not (as) corrupt” cluster



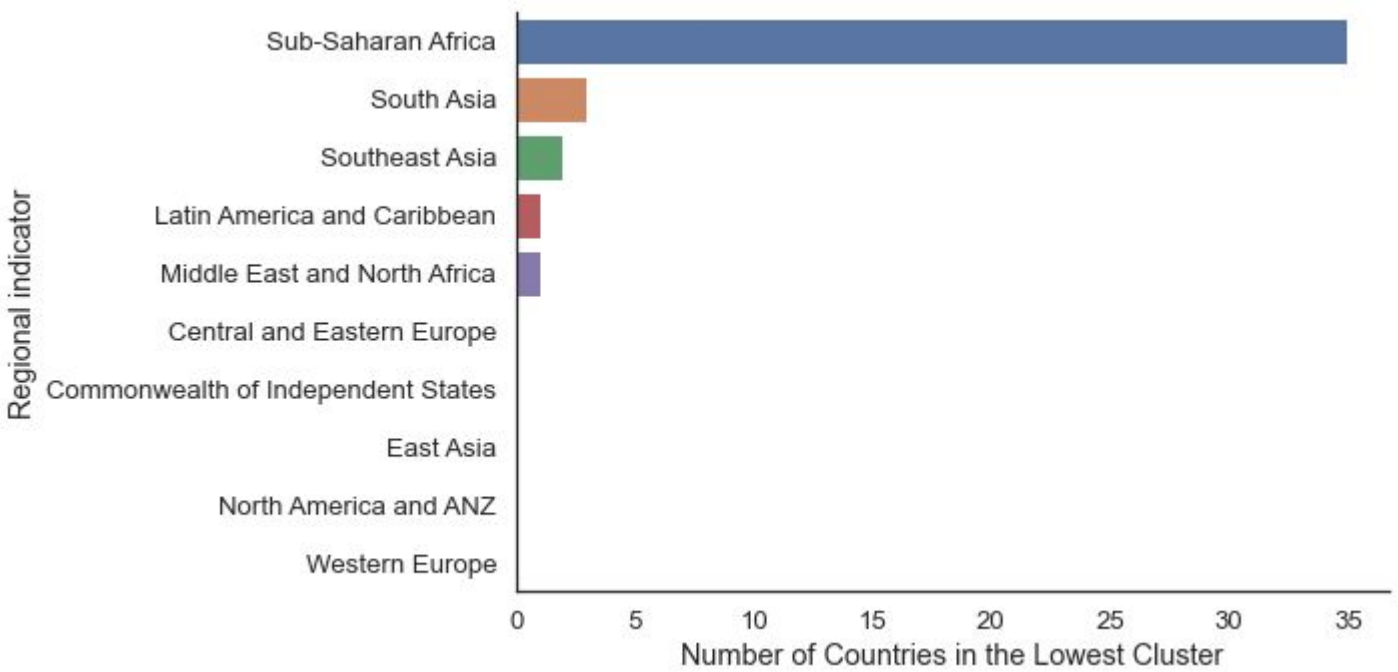
How can countries be grouped by features?

“Medium rich, corrupt” cluster



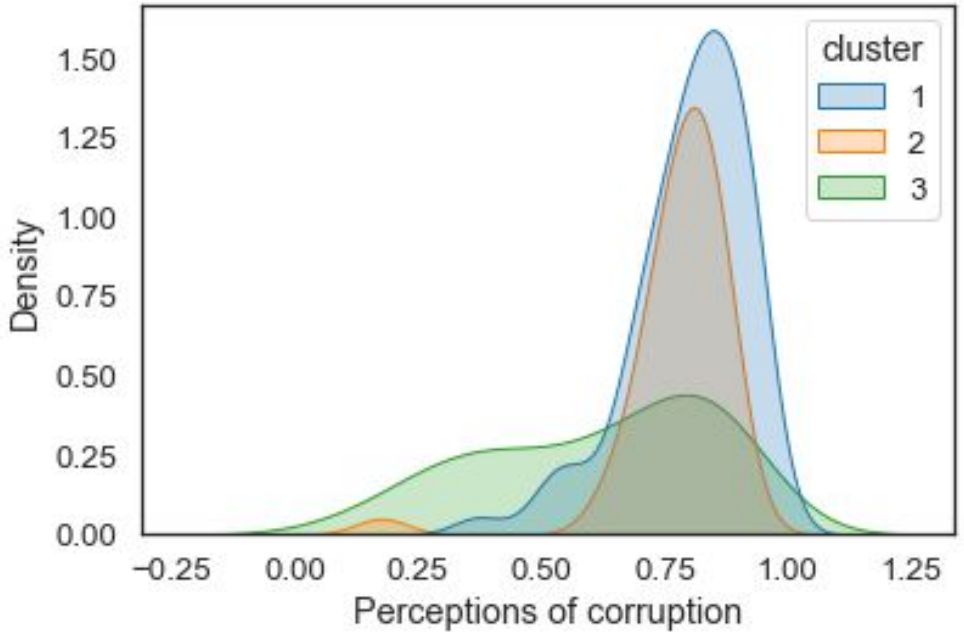
How can countries be grouped by features?

“Poor, corrupt” cluster



What does corruption look like in the highest grouping?

Clustering



KEY

Poor, corrupt

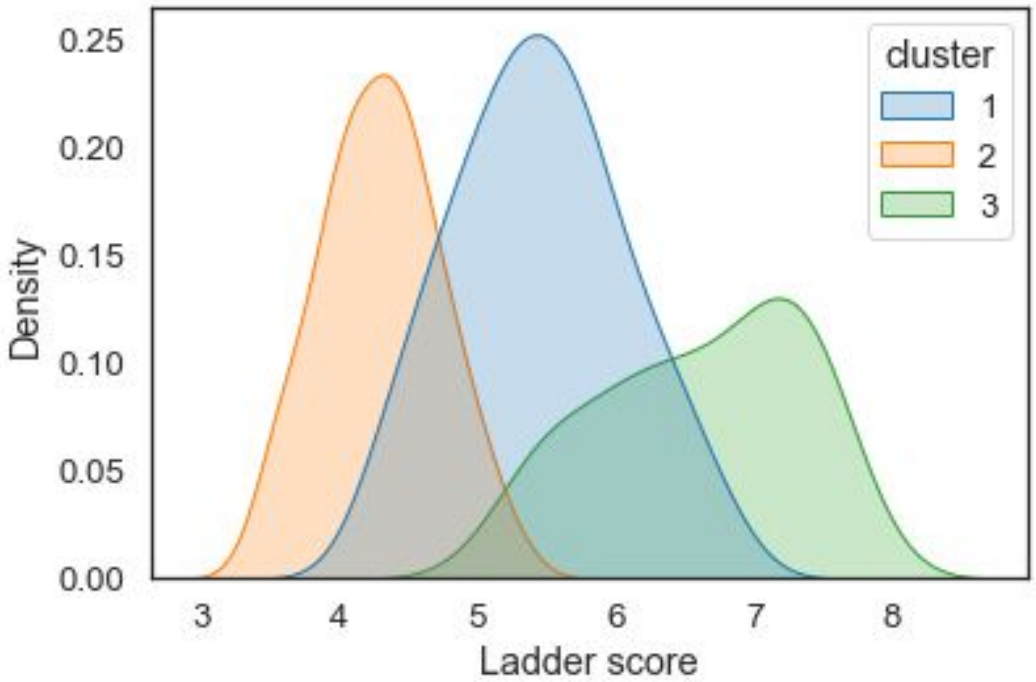
Medium rich, corrupt

Rich, not (as) corrupt

The distribution of corruption in the green countries drastically varies from the orange and blue countries' distributions.

How does corruption affect the happiness index?

Clustering



KEY

Poor, corrupt

Medium rich, corrupt

Rich, not (as) corrupt

What distinguishes the "medium-happy" countries from the top of the ladder?

Largely corruption (green countries have substantially less)

Conclusion

- Western European, North American, and ANZ countries have the highest ladder scores (happiness) over the years since 2005
- Sub-Saharan African countries have the lowest ladder scores
- Log GDP Per Capita was the most important feature for predicting the ladder score of a country
- “Perceptions of corruption” has a negative correlation with the ladder score (and with other features)
- Corruption is the least important feature in predicting ladder score, but… low corruption is a defining trait of the happiest countries (ladder > 7.5)