Group Project: World Happiness Report

by group pd.csr

Project Goal

We have been hired by the UN to build a model that can predict the happiness of an average person in each country based on scores for 8 categories provided by the UN and an additional 4 factors that we added. Secondly, can we predict what region of the world this average person is from, based on the scores from the features.

Data Gathering

United Nations Happiness Reports (2015-2017) (Kaggle)

- 1000 people sampled per country

https://ourworldindata.org/

- Country
- Region
- Year
- Happiness Score (aggregate)
- **Economy (GDP per cap)**
- **Family**
- **Health**
- **Freedom**
- **Trust (Government)**
- **Generosity**
- Dystopia Residual

- Food Supply (kcal per capita/day)
- Crude Birth Rate (per 100k)
- Deaths - Unsafe Water (per 100k)
- Deaths - Conflict and Terrorism

Data Cleaning

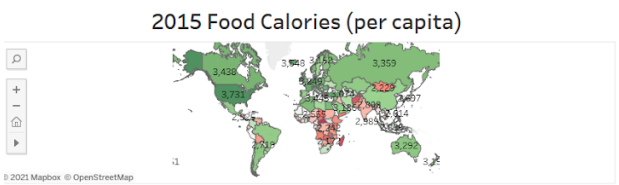
01 Only Data from OurWorldInData had missing values for several years

02 90% + good data, not missing!

03 Since every region is very different, all missing values were imputed based on the mean of that respective region

Exploratory Data Analysis

I wanted to explore the relationship between Regions, their Calorie Intake, and how they valued Freedom.



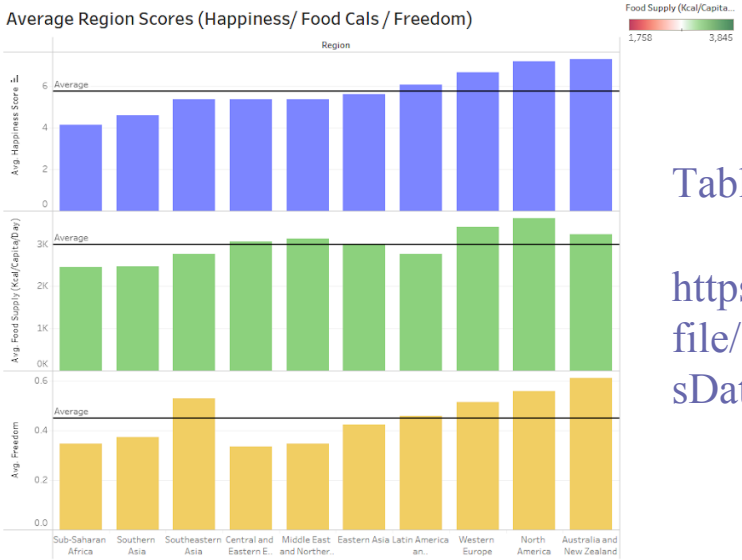
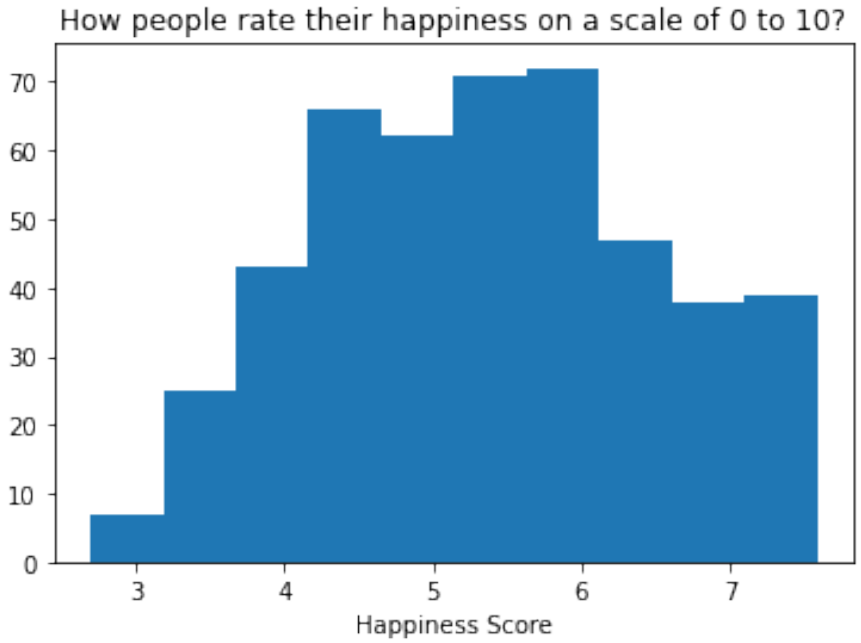


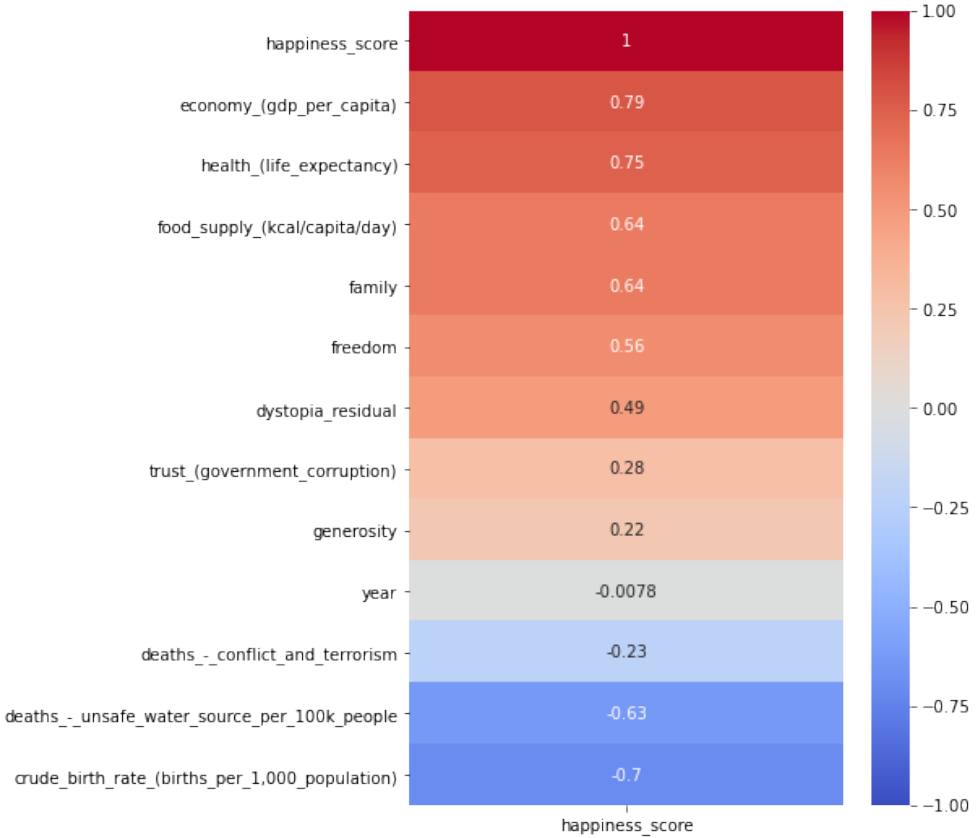
Tableau Visual

https://public.tableau.com/app/profile/spencer.buckner/viz/HappinessData_16396021350410/Dash1

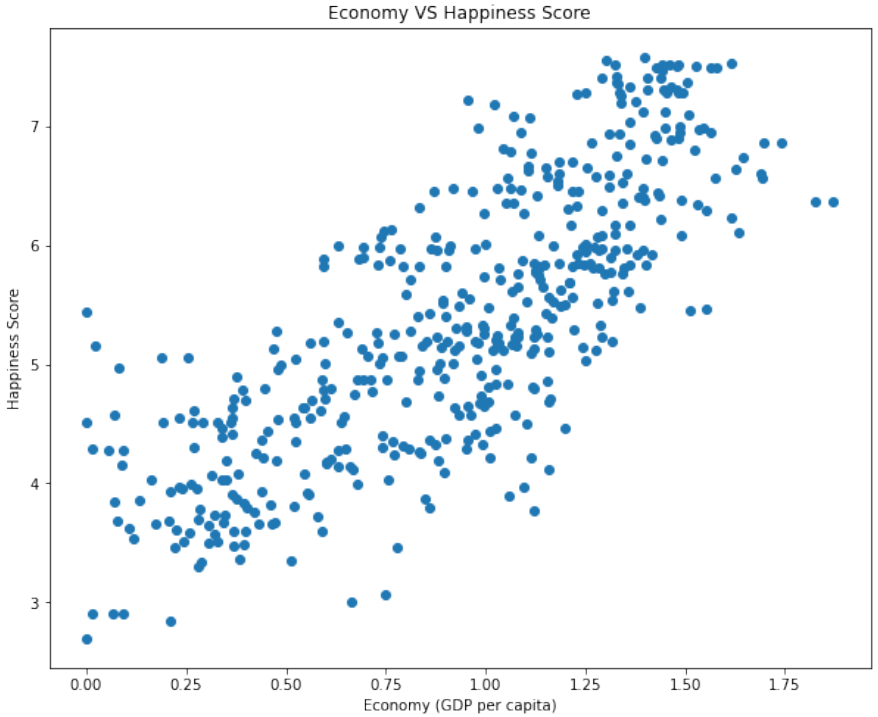
Happiness Score Distribution



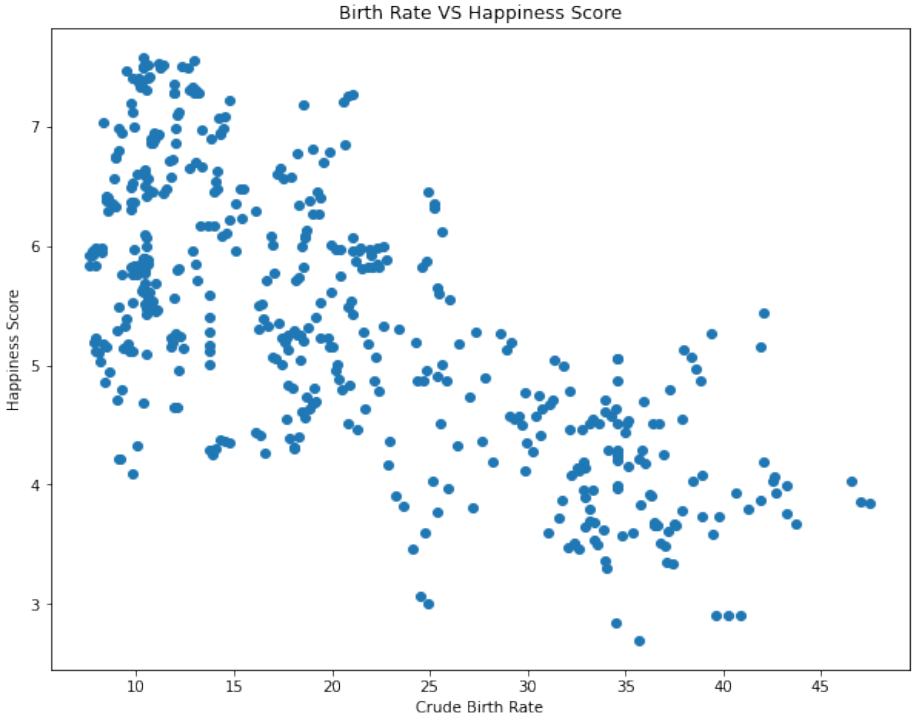
Happiness Score Heatmap



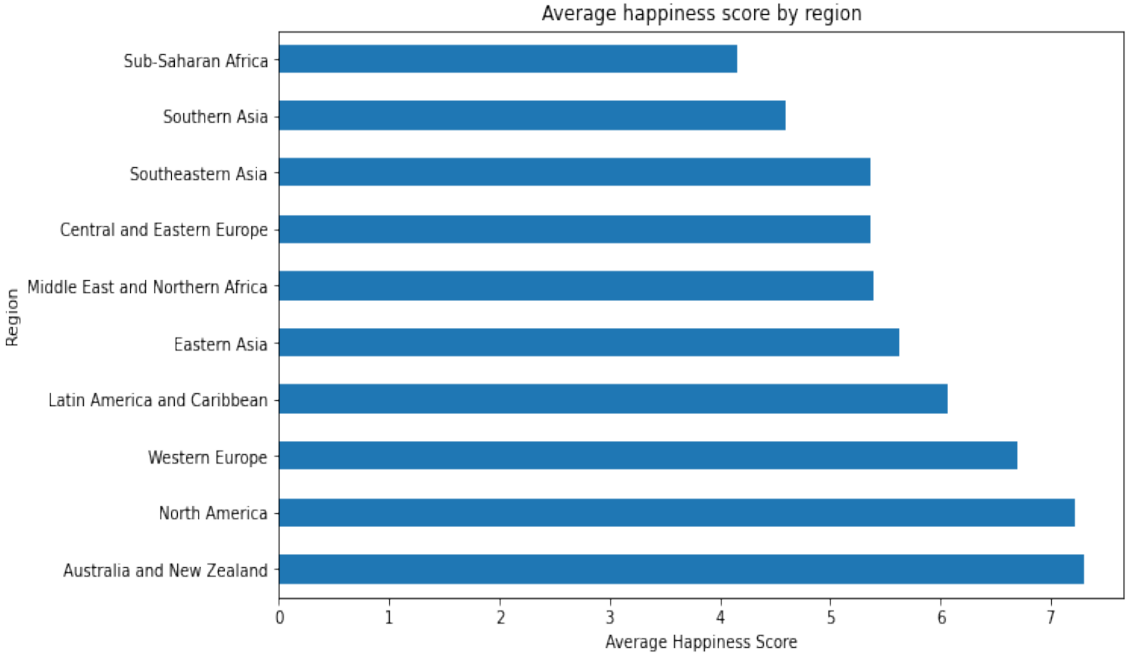
Economy VS Happiness Score



Birth Rate VS Happiness Score



Happiness score by region



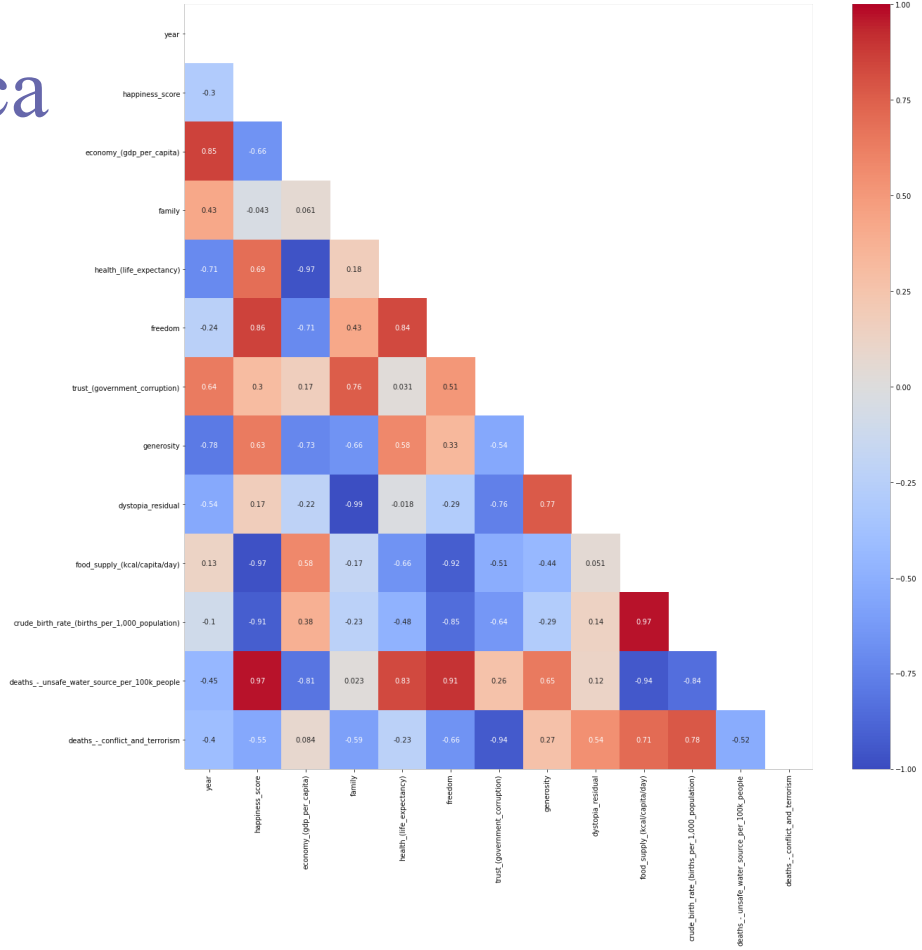
Ranking

Australia and New Zealand 7.302500
North America 7.227167
Western Europe 6.693000
Latin America and Caribbean 6.069074
Eastern Asia 5.632333
Middle East and Northern Africa 5.387879
Central and Eastern Europe 5.371184
Southeastern Asia 5.364077
Southern Asia 4.590857
Sub-Saharan Africa 4.150957

Heatmap of North America

Most Positive Correlations with Happiness:

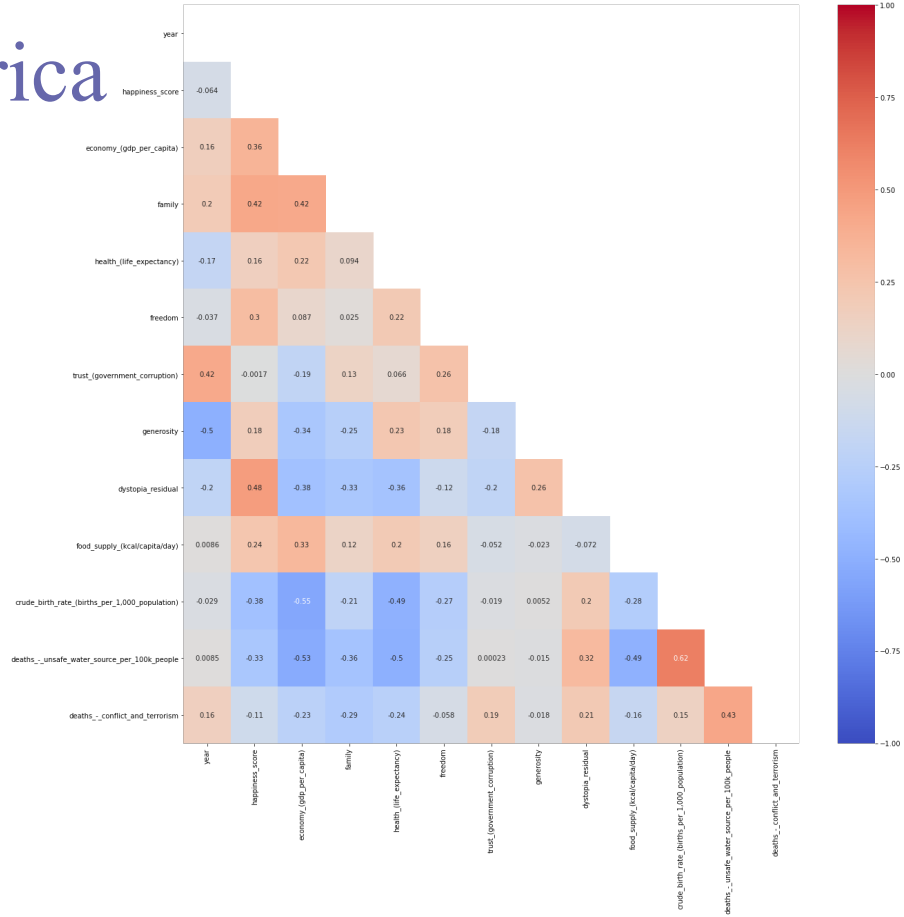
- Deaths due to unsafe drinking water
- Freedom
- Health (Life Expectancy)



Heatmap of Sub-Saharan Africa

Most Positive Correlations with Happiness:

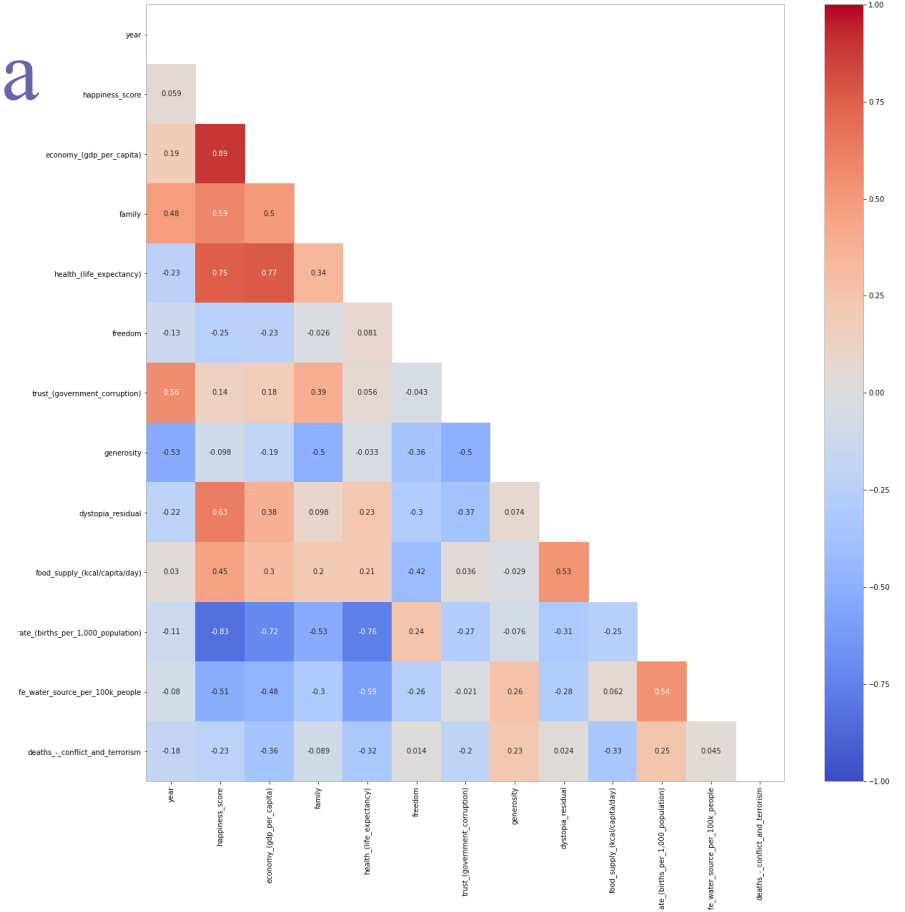
- Family (Social Support)
- Freedom
- Economy (GDP per Capita)



Heatmap of Southeast Asia

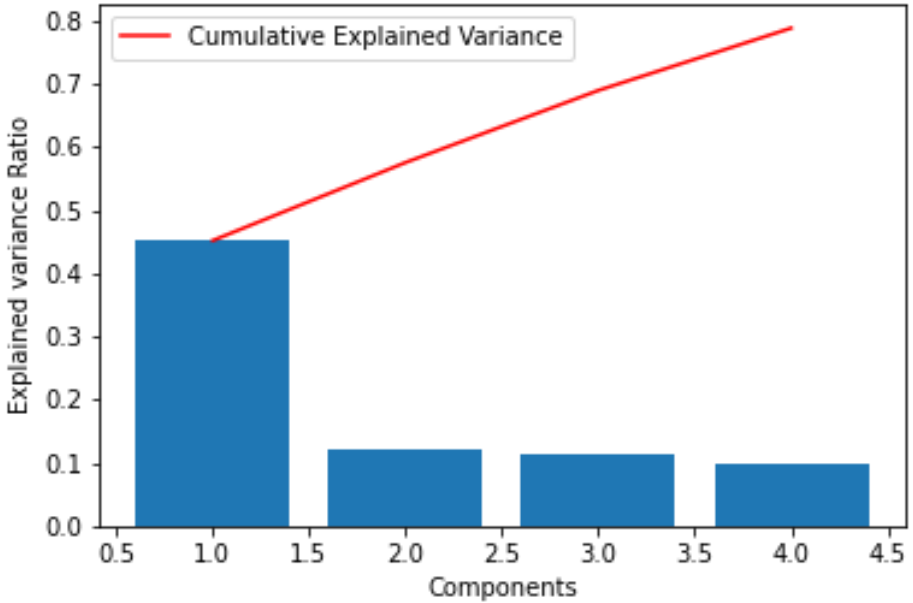
Most Positive Correlations with Happiness:

- Economy (GDP per Capita)
- Health (Life Expectancy)
- Family (Social Support)

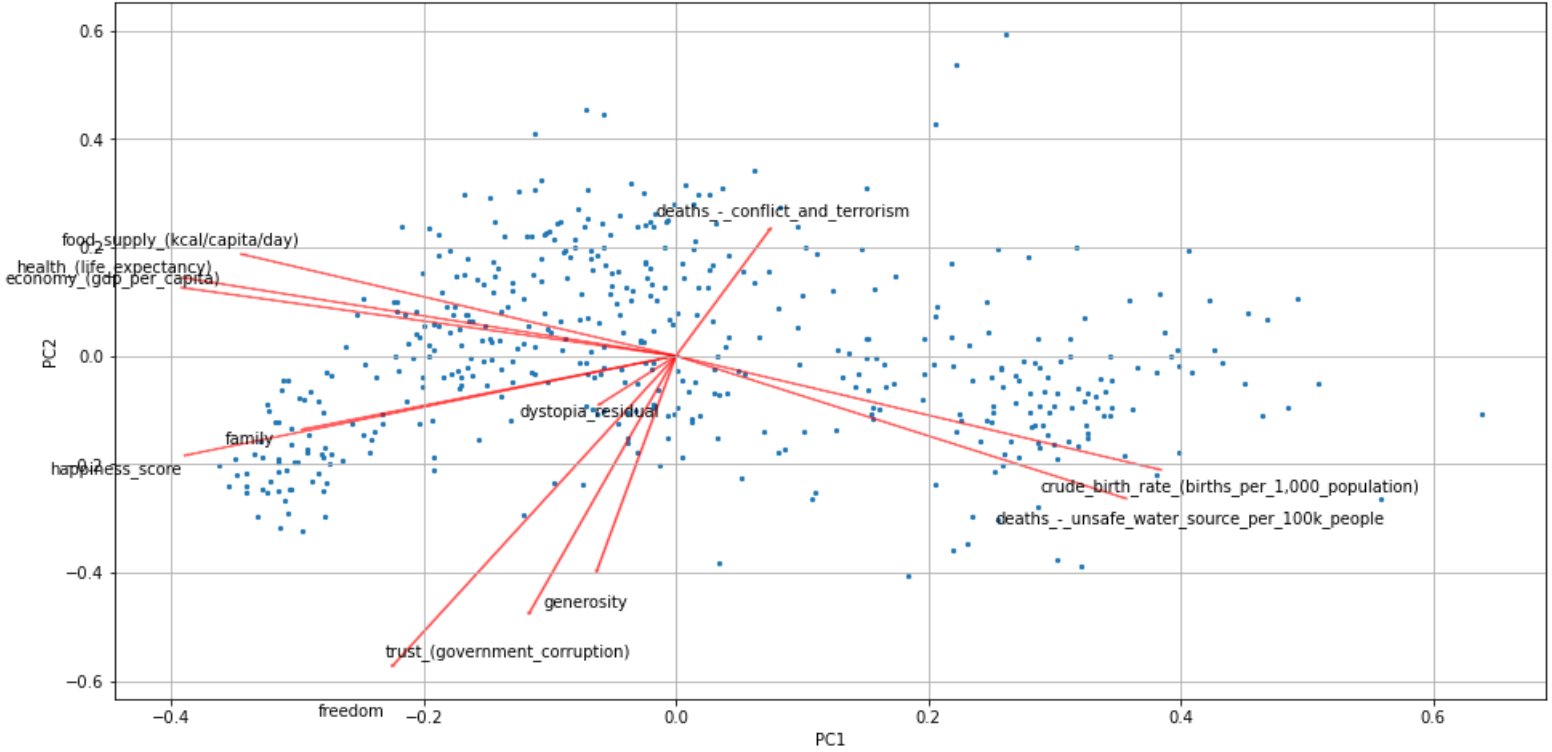


Cumulative Explained Difference (ratio)

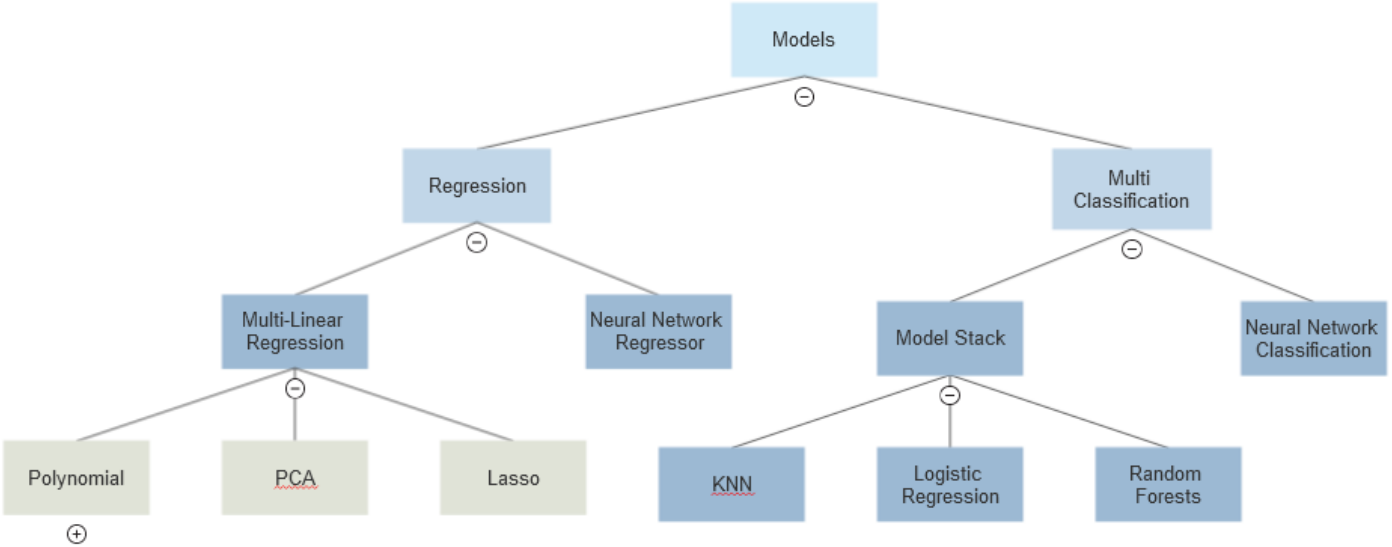
Principal Component 1 has nearly 50% of explained difference



Principle Component Map



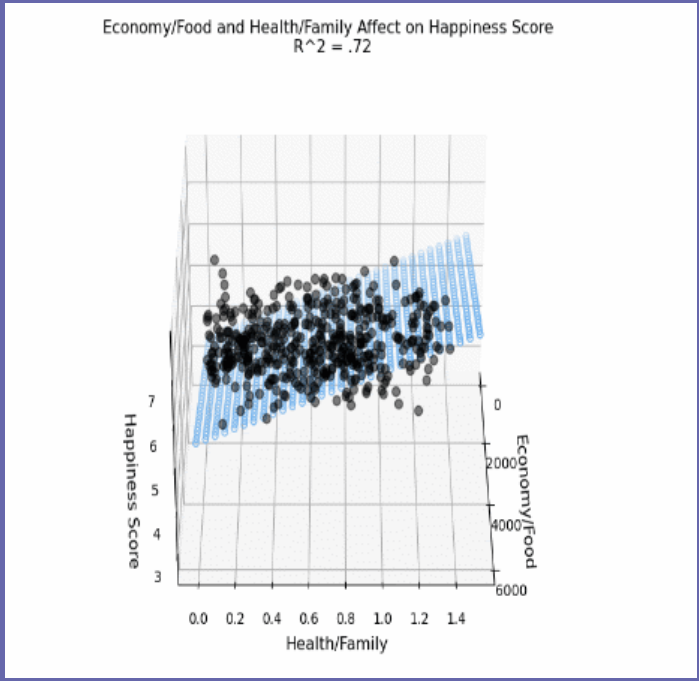
Model Map


Models
Regression
Multi Classification
Multi-Linear Regression
Neural Network Regressor
Model Stack
Neural Network Classification
Polynomial
PCA
Lasso
KNN
Logistic Regression
Random Forests

MLR 3D Model

Feature Engineering 2 new Features:
Economy/Food
- economy_(gdp_per_capita)
- food_supply_(kcal/capita/day)

Health/Family
- health_(life_expectancy)
- family



R2 Score 0.72

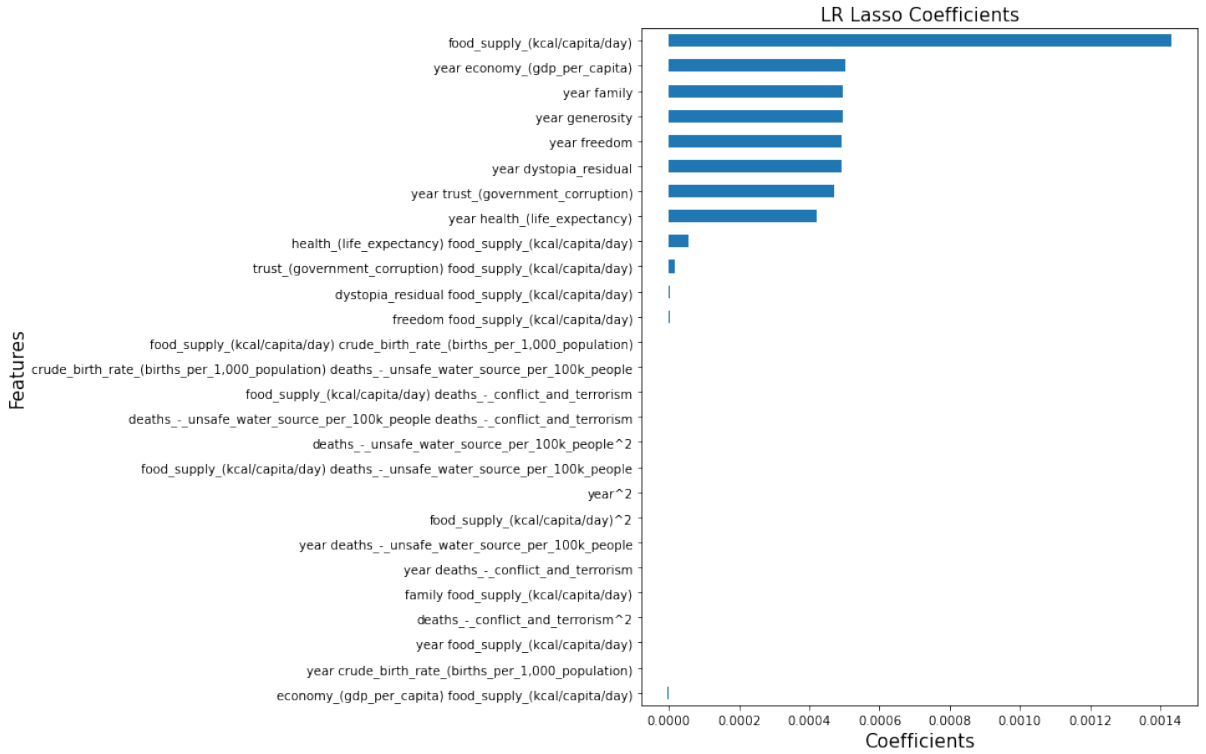
Linear Regression – Polynomial/Lasso/PCA

X = df[features]
y = df['happiness_score']

90 polynomial features

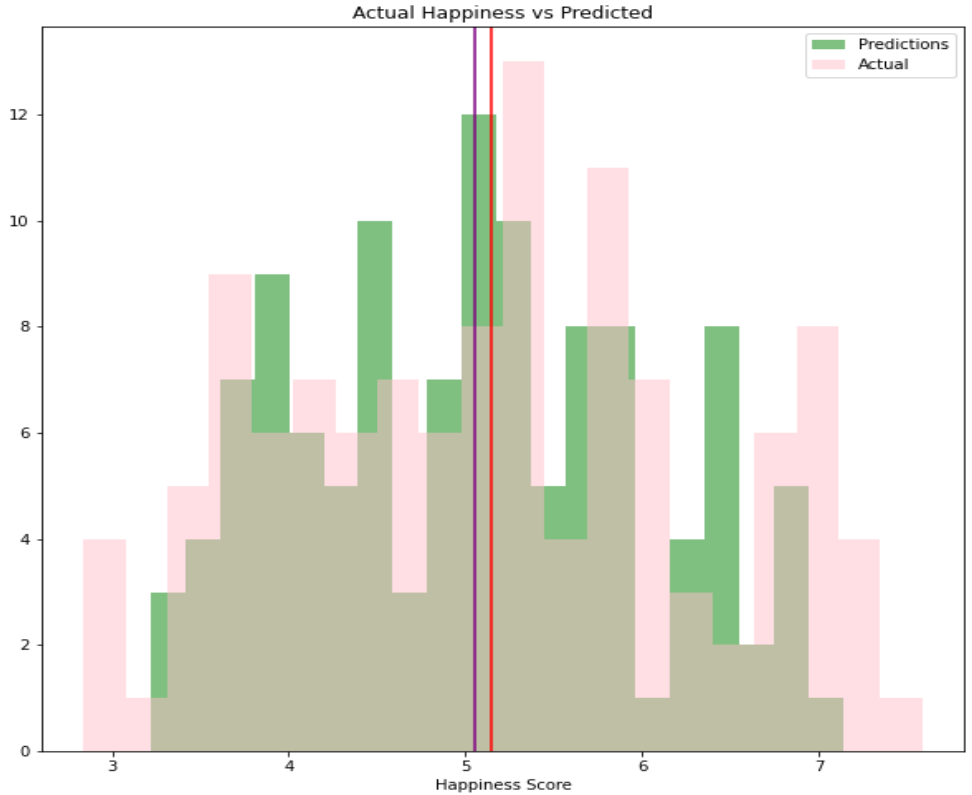
27 lasso features

polynomial.score: 0.9999
lasso.score: 0.9999
PCA.score: 0.98

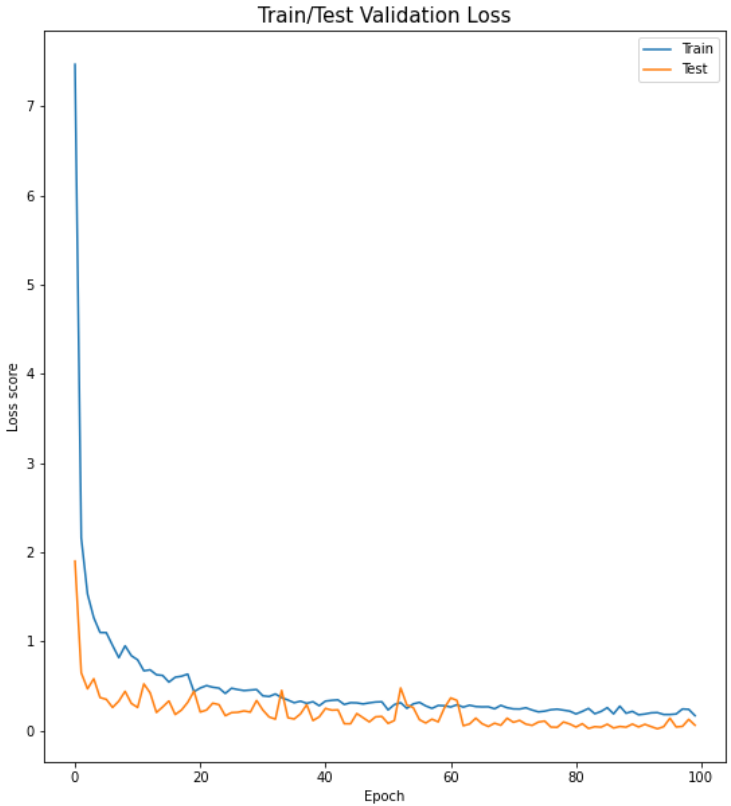


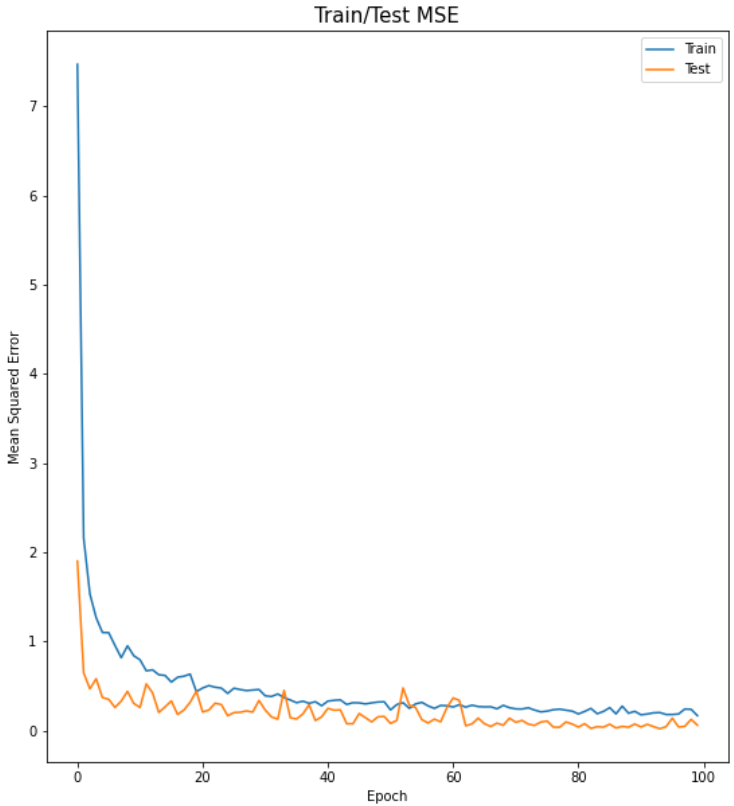
Neural Network Regressor

R2 score: 0.95408



NN Regressor Loss/MSE



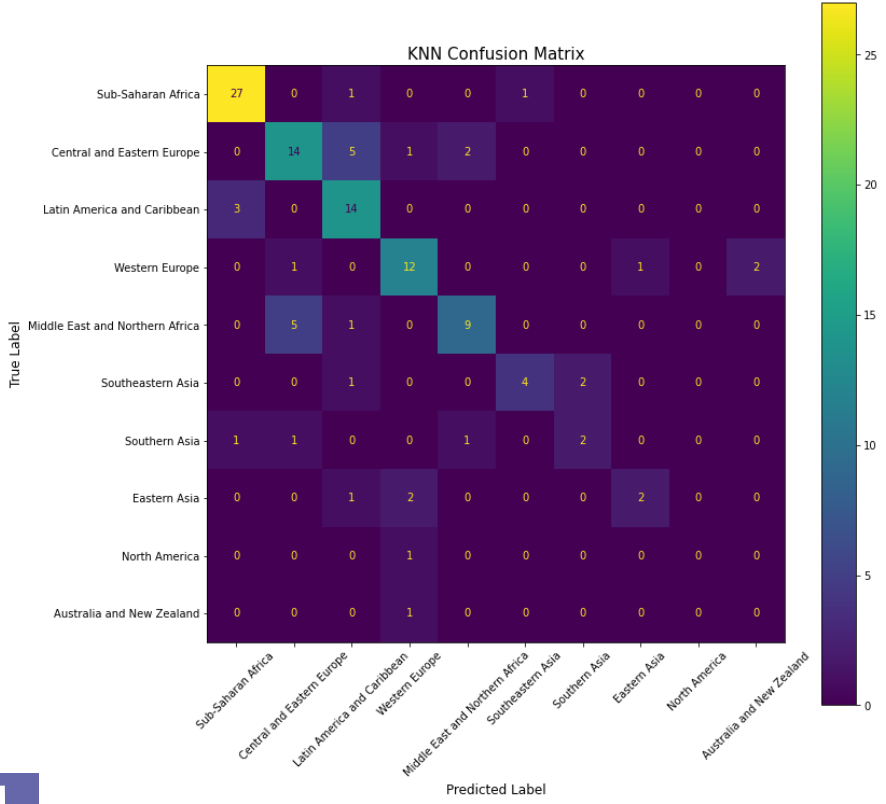


Multi Classification by Region

Can we correctly predict the region of each country?

- KNN
- Logistic Regression
- Random Forest
- Model Stack of all 3
- Neural Network Classifier

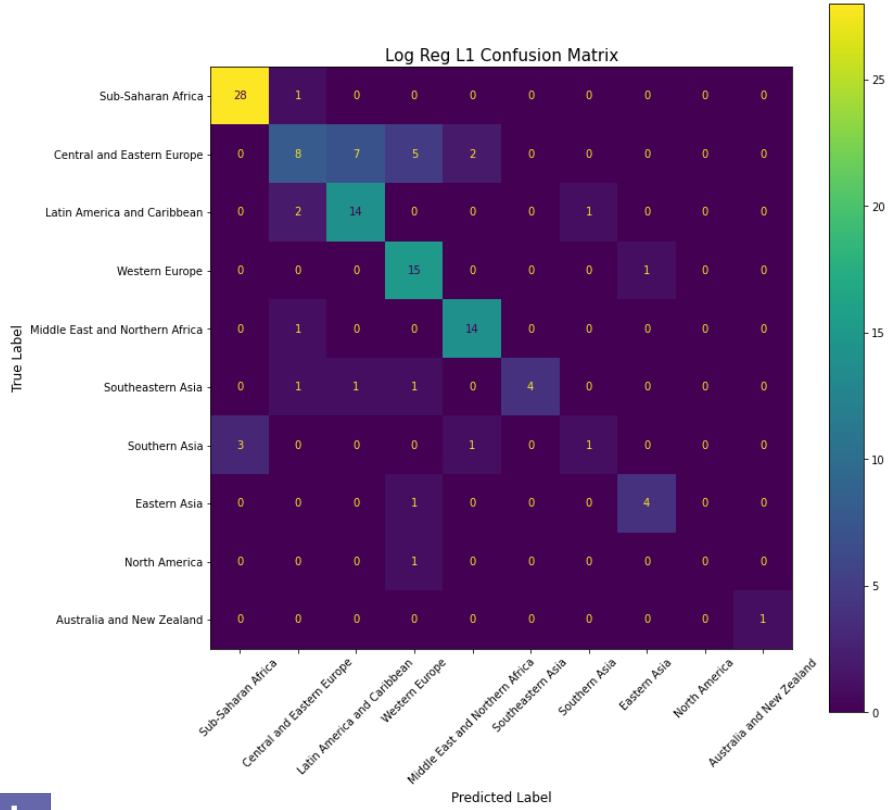
KNN



Classification Report

	precision	recall	f1-score	support
Sub-Saharan Africa	0.87	0.93	0.90	29
Central and Eastern Europe	0.67	0.64	0.65	22
Latin America and Caribbean	0.61	0.82	0.70	17
Western Europe	0.71	0.75	0.73	16
Middle East and Northern Africa	0.75	0.60	0.67	15
Southeastern Asia	0.80	0.57	0.67	7
Southern Asia	0.50	0.40	0.44	5
Eastern Asia	0.67	0.40	0.50	5
North America	0.00	0.00	0.00	1
Australia and New Zealand	0.00	0.00	0.00	1
accuracy			0.71	118
macro avg	0.56	0.51	0.53	118
weighted avg	0.71	0.71	0.71	118

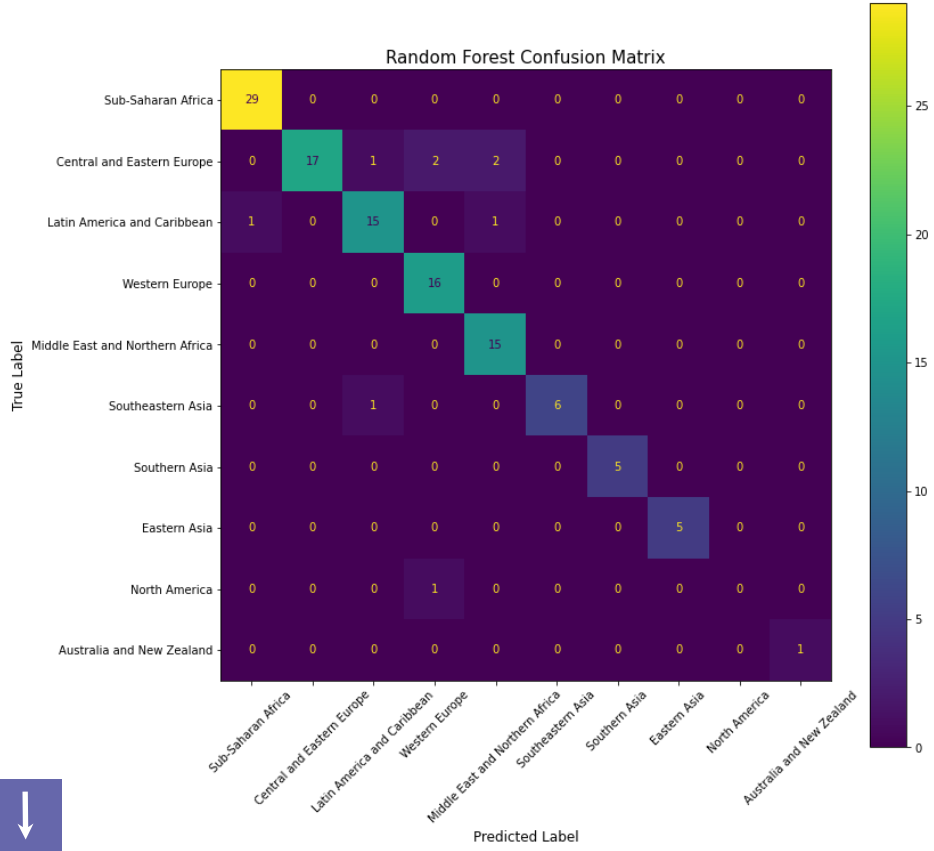
Log Regression



Classification Report

	precision	recall	f1-score	support
Sub-Saharan Africa	0.90	0.97	0.93	29
Central and Eastern Europe	0.62	0.36	0.46	22
Latin America and Caribbean	0.64	0.82	0.72	17
Western Europe	0.65	0.94	0.77	16
Middle East and Northern Africa	0.82	0.93	0.87	15
Southeastern Asia	1.00	0.57	0.73	7
Southern Asia	0.50	0.20	0.29	5
Eastern Asia	0.80	0.80	0.80	5
North America	0.00	0.00	0.00	1
Australia and New Zealand	1.00	1.00	1.00	1
accuracy			0.75	118
macro avg	0.69	0.66	0.66	118
weighted avg	0.74	0.75	0.73	118

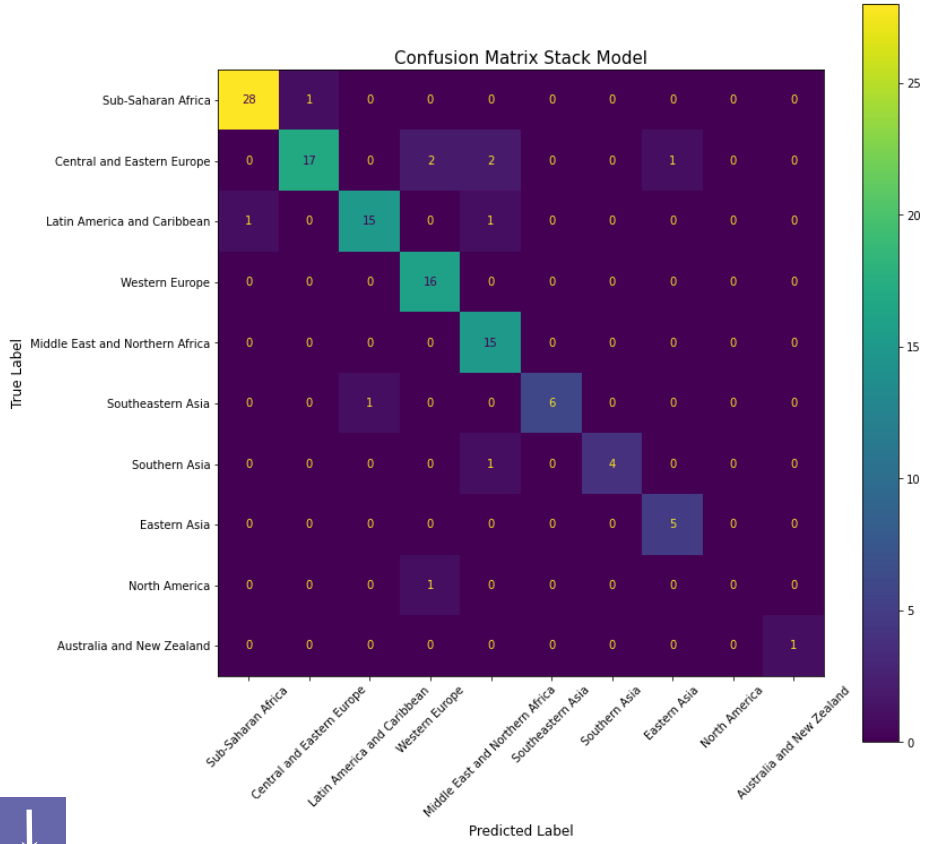
Random Forest



Classification Report

	precision	recall	f1-score	support
Sub-Saharan Africa	0.97	1.00	0.98	29
Central and Eastern Europe	1.00	0.77	0.87	22
Latin America and Caribbean	0.88	0.88	0.88	17
Western Europe	0.84	1.00	0.91	16
Middle East and Northern Africa	0.83	1.00	0.91	15
Southeastern Asia	1.00	0.86	0.92	7
Southern Asia	1.00	1.00	1.00	5
Eastern Asia	1.00	1.00	1.00	5
North America	0.00	0.00	0.00	1
Australia and New Zealand	1.00	1.00	1.00	1
accuracy			0.92	118
macro avg	0.85	0.85	0.85	118
weighted avg	0.92	0.92	0.92	118

Model Stack



Classification Report

	precision	recall	f1-score	support
Sub-Saharan Africa	1.00	0.97	0.98	29
Central and Eastern Europe	0.94	0.77	0.85	22
Latin America and Caribbean	0.83	0.88	0.86	17
Western Europe	0.84	1.00	0.91	16
Middle East and Northern Africa	0.83	1.00	0.91	15
Southeastern Asia	0.83	0.71	0.77	7
Southern Asia	0.80	0.80	0.80	5
Eastern Asia	1.00	1.00	1.00	5
North America	0.00	0.00	0.00	1
Australia and New Zealand	1.00	1.00	1.00	1
accuracy			0.90	118
macro avg	0.81	0.81	0.81	118
weighted avg	0.90	0.90	0.89	118

Confusion Matrix for NN Multi-Classifier

```
<tf.Tensor: shape=(10, 10), dtype=int32, numpy=
array([[39,  0,  0,  0,  1,  0,  0,  0,  0,  0],
       [ 0, 14,  0,  1,  2,  0,  0,  0,  0,  0],
       [ 0,  0, 17,  0,  0,  0,  0,  0,  0,  0],
       [ 0,  0,  0, 12,  0,  0,  0,  1,  1,  1],
       [ 0,  0,  0,  1, 11,  0,  0,  0,  0,  0],
       [ 0,  0,  2,  0,  0,  4,  0,  0,  0,  0],
       [ 1,  0,  1,  0,  0,  0,  3,  0,  0,  0],
       [ 0,  1,  0,  1,  1,  0,  0,  1,  0,  0],
       [ 0,  0,  0,  1,  0,  0,  0,  0,  0,  0],
       [ 0,  0,  0,  1,  0,  0,  0,  0,  0,  0]],
```

Conclusion

Regression Models

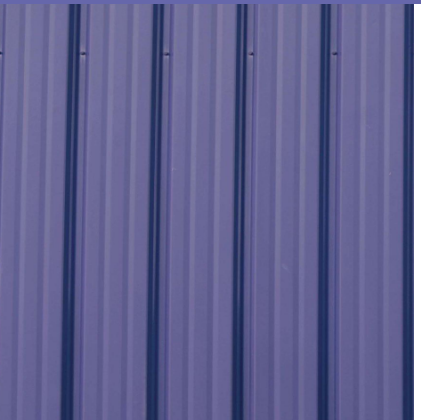
Since all of the features were known and that that there were no unknown variables that would affect happiness, our polynomial model produced the highest R2 score. When reducing features with a Lasso, we were able to reduce down from 90 to 27 features. We also discovered that the most important coefficient was the food calorie supply, which was not information supplied by the UN. It was also interesting that the neural network performed worse than the Linear Regression, even though the Neural Network was also given polynomial features. We believe that this is likely due to the fact that this is a small data set, and neural networks were designed to handle larger data sets. If we were to add more years, and therefore more data, we believe it is likely that the Neural Network would match the effectiveness of the simple linear regression model.

Conclusion

Classification Models

The worst model was the KNN model with 71%, followed by the Logistic Regression model with an accuracy of 75%. The best model was the Random Forest model with an accuracy of 92%, which was even better than when we stacked all of the models. The random forest was able to perfectly predict 3 of the regions with 100% precision and recall scores. All models struggled to predict Central and Eastern Europe countries. The Neural Network again was out performed by the simpler models, and again we believe this is the case because we had a small data set. It should noted that all models out performed the baseline accuracy, which was 25%.

Thank you



Do you have any questions?