Generosity Around the World

Objective:

The primary objective of the model is to determine the extent to which certain socioeconomic factors influence the generosity of citizens of countries.

Data:

The datasets used are the 2018 and 2019 editions of the World Happiness Report conducted by Sustainable Development Solutions Network. The data shows scores of six socioeconomic factors for each country in the world as determined through surveys of citizens in each country. These scores are then summed to create an overall “Happiness” score for each country, which then is used to rank each country by “Happiness.” However, for this project, I wanted to have the dependent variable be the “Generosity” score rather than the “Happiness” score. The individual features are as follows.

1. Overall rank (int64)
2. Country or region (object)
3. Score (float64)
4. GDP per capita (float64)
5. Social support (float64)
6. Healthy life expectancy (float64)
7. Freedom to make life choices (float64)
8. Generosity (float64)
9. Perceptions of corruption (float64)

EDA & Data Cleaning:

For this analysis, I did not need to use the first three variables listed above, so I dropped them from the data.

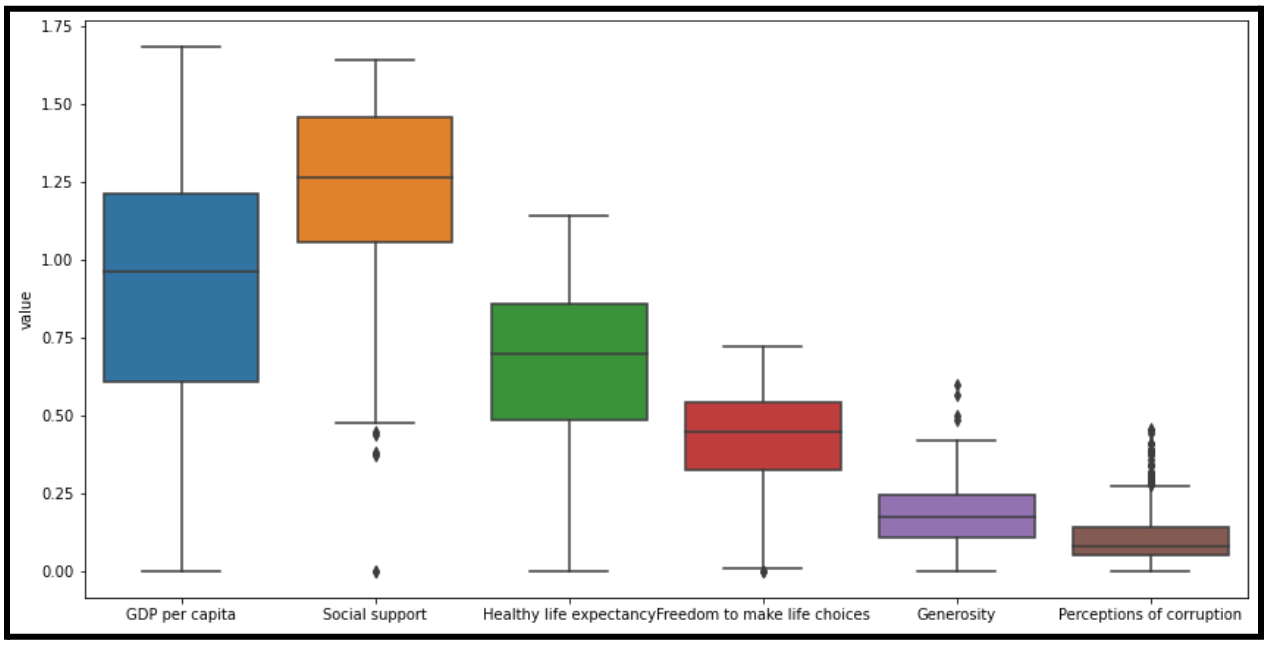
A search to see if any nulls were present indicates there are nulls present in “Perceptions of corruption.” A count of nulls in that specific feature showed there is only one record with a null value. Therefore I removed that one record.

```
Nulls Present?
GDP per capita                  False
Social support                  False
Healthy life expectancy         False
Freedom to make life choices    False
Generosity                      False
Perceptions of corruption        True
```

I then looked at a description of the data to get a better sense of what (if any) cleaning needed to be done.

	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
count	311.000000	311.000000	311.000000	311.000000	311.000000	311.000000
mean	0.894447	1.212424	0.661267	0.423987	0.182916	0.111299
std	0.389311	0.299774	0.253131	0.156074	0.096895	0.095365
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.608000	1.057000	0.487500	0.325500	0.108500	0.050000
50%	0.960000	1.266000	0.700000	0.450000	0.175000	0.082000
75%	1.214500	1.458000	0.859000	0.540500	0.245000	0.140500
max	1.684000	1.644000	1.141000	0.724000	0.598000	0.457000

Of note is that all features had a min value of zero. I was interested in understanding more about these zero values. I created a box plot for each variable to see if the zero values appeared to be outliers or within the whiskers of the box plots.



As shown above, the zero values in the ‘Social support’ variable are quite far outside of the whiskers, which indicates the values may not be reliable and therefore the records should be

removed. Consequently the zero values in the other variables are within the whiskers of their plots, which indicates those zero values are meaningful and should be retained for modeling. The box plot also shows outliers for some variables *above* the top whiskers. However, these appear to indicate data with a right skew rather than erroneous records.

The final step of data cleaning was to create an alternate dataframe where the variables were normalized using a Box-Cox transformation. This will be useful for models using linear regression without regularization methods, especially because this dataset does not have a large amount of records.

Regression Models (#1 - Simple Linear Regression):

As we are looking to create a model that prioritizes interpretability over predictive power, a simple linear regression model would be the best place to start. I used a backwards elimination procedure (with a 0.05 p-value threshold) in order to remove variables for the sake of simplifying the model as much as possible. Comparing the R^2 values for the model before and after backwards elimination, we see the backwards elimination model has slightly more predictive power, however both R^2 values are very low.

R^2 value (before backwards elimination): **0.171**
R^2 value (after backwards elimination): **0.185**

With the understanding that the model does not have high predictive power, we still want to examine which features remain after backwards elimination, and the extent to which they influence the "Generosity" feature. As seen below, the two largest influences on "Generosity" are the "Freedom to make life choices" and "Perceptions of corruption." Both features have a positive influence on "Generosity." Surprisingly "GDP per capita" has a negative influence on "Generosity." Perhaps as we make use of more elaborate models with more features, we will better understand the details of why "GDP per capita" has this negative correlation with "Generosity."

	coef	std err	t	P>|t|
GDP per capita	-0.2439	0.056	-4.384	0.000
Freedom to make life choices	0.2257	0.059	3.823	0.000
Perceptions of corruption	0.3261	0.058	5.618	0.000

Regression Models (#2 - Simple Linear Regression w/ Polynomial Features):

For the next model I wanted to add polynomial features to the model. This will inevitably complicate the model as it will create additional features without performing any sort of feature reduction process. However, it is worthwhile in order to determine if we can increase our predictive power.

I used a GridSearchCV on Polynomial Features to determine whether a degree of 2 or 3 would be best for a linear regression model. The result showed the best model used a degree of 2. The R^2 score of this model did show improvement over the simple linear regression model.

R^2 value (linear regression, polynomial features = 2): **0.302**
Number of Features (polynomial features = 2): **21**

The number of features in this model is 21, making it more complex than the 3 features shown in the simple linear regression.

Regression Models (#3 - Lasso Regression w/ Polynomial Features):

Since we determined in model #2 that polynomial features increases the predictive power, for the final model I wanted to use polynomial features again and couple it with a lasso regression in order to reduce the number of features. I again used GridSearchCV with polynomial features between 1 and 3, and lasso alphas between 1e-9 and 1.

The best model was determined to be as follows:
{'lasso__alpha': 0.001, 'poly_features__degree': 3}

R^2 value (lasso regression, alpha = 0.001, polynomial features = 3): **0.371**
Number of Features (lasso regression, alpha = 0.001, polynomial features = 3): **19**

The predictive power of this model shows a slight improvement over model #2. Despite creating polynomial features at degree=3, the lasso regression was able to remove enough features so the total count of features was less than that of the linear regression with polynomial features at degree=2.

Recommendations for model:

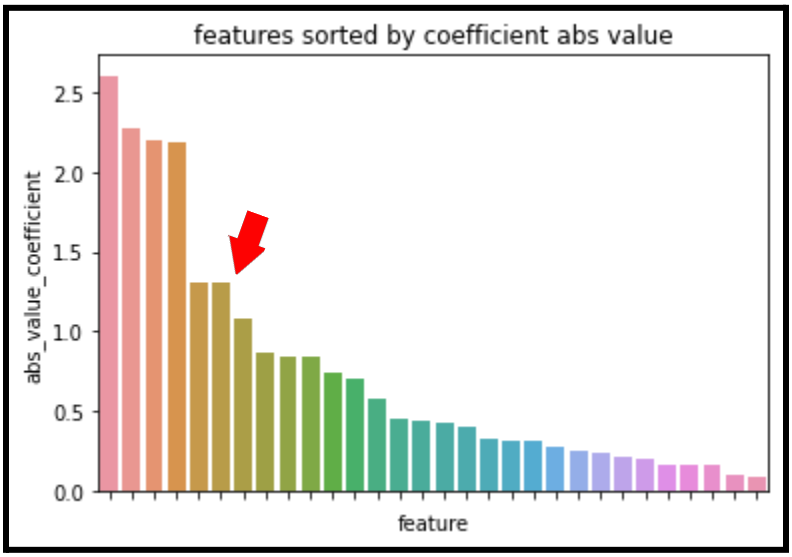
The intended focus for the model prioritizes interpretability over accuracy. The difficulty with these models is that the most interpretable model (model #1) has very little accuracy (R^2 = 0.185) and the most accurate model (model #3, R^2 = 0.371) is difficult to interpret due to the polynomial features at degree=3. If we were trying to find and append additional data features for the countries shown in the data, I would recommend using model #1 in order to maintain the greatest amount of interpretability. If we were solely examining the features provided in the datasets (as we are doing in this instance), I recommend using model #3 and attempting to interpret the results as best as possible. As we are not considering the addition of more features from external datasets, I will interpret the results of model #3.

Model interpretation:

For the sake of simplicity, the features will herein be referred to as follows:

x0: GDP per capita
x1: Social support
x2: Healthy life expectancy
x3: Freedom to make life choices
x4: Perceptions of corruption

The first step for interpretation is to take the absolute value of the model coefficients and order them in descending order so we can determine which features have the greatest impact in the model.



The bar graph above shows features x0 and x3 have the largest impact, which comports with what was seen in the simplistic backward elimination linear regression in model #1. Also of note is that feature x1 by itself was not retained after the feature elimination process. There is a noticeable inflection point after the first six features, therefore those fist six features will be the primary focus of the interpretation. The coefficients for those six features are as follows.

feature	coefficient	abs_value_coefficient
x0	-0.041161	0.041161
x3	0.035799	0.035799
x0 x2	0.025964	0.025964
x4^2	0.019014	0.019014
x0^2 x2	0.018574	0.018574
x1 x3 x4	0.013093	0.013093

Feature x0 (GDP per capita), as seen in model #1, has a negative correlation with generosity when not combined with other features. Only when it is combined with x2 (Healthy life expectancy) or squared then combined with x2 does it show positive correlation. The interpretation of this could be that increases in a country's GDP will have negative impacts on Generosity if the benefits of the GDP increase do not make a meaningful impact on a person's wellbeing. In other words, if citizens

feel as if the political and economic systems are not being generous to them, they are less likely to be generous to other citizens.

Feature x3 (Freedom to make life choices) is the feature with the largest *positive* correlation with generosity. Feature x4^2 (Perceptions of corruption) also has a large positive correlation. Both of these point to the interpretation that the more citizens feel as if they can trust the political system, economic system and institutions as a whole, they are more likely to be generous. This perhaps could transcend larger institutions and also be reflected in trust of fellow citizens being positively correlated with generosity.

Recommendations for further study:

As discussed above, there is little accuracy in the model without adding polynomial features. However, polynomial features decrease the ability to interpret the model. Therefore, the model could be more robust and interpretable if polynomial features were kept at degree= 1 or 2 while adding additional features from other data sources.