EXPLORATORY DATA ANALYSIS for MACHINE LEARNING

Khobie Maseko

25/10/2021

Project Background

- This data was obtained from a dataset by a Kaggle user named Michael Londeen titled **World Happiness Report 2020**. It illustrates the different levels of the perceived general happiness of different populations in 153 countries and the different factors that contribute to that happiness in the year 2020.
- Dateset source: https://www.kaggle.com/londeen/world-happiness-report-2020
- The dataset consists of 153 columns and 20 rows.
- For purposes of this project I will clean the data and explore the relationship between the target (**Dystopia + residual levels of happiness**) and potential predictors.

Data Exploration Plan

The following preliminary steps are the methods we will utilise to attempt to build a baseline model to find out if there are any strong correlations between the different factors and the target variable:

Data Overview (1) - Data Shape

- Dataset has **153 rows** and **20 columns**

```
Number of rows in the data: 153
Number of columns in the data: 20
```

Data Overview (2) - Data Columns

```
['Country name', 'Regional indicator', 'Ladder score', 'Standard error of ladder score', 'upperwhisker', 'lowerwhisker', 'Logged GDP per capita', 'Social support', 'Healthy life expectancy', 'Freedom to make life choices', 'Generosity', 'Perceptions of corruption', 'Ladder score in Dystopia', 'Explained by: Log GDP per capita', 'Explained by: Social support', 'Explained by: Healthy life expectancy', 'Explained by: Freedom to make life choices', 'Explained by: Generosity', 'Explained by: Perceptions of corruption', 'Dystopia + residual']
```

- All the different columns in the dataset
- Our target variable is '**Distopia + residual**'

Data Overview (3) - Data Types

```
Country name                                  object
Regional indicator                            object
Ladder score                                 float64
Standard error of ladder score               float64
upperwhisker                                 float64
lowerwhisker                                 float64
Logged GDP per capita                        float64
Social support                               float64
Healthy life expectancy                      float64
Freedom to make life choices                 float64
Generosity                                   float64
Perceptions of corruption                    float64
Ladder score in Dystopia                     float64
Explained by: Log GDP per capita             float64
Explained by: Social support                 float64
Explained by: Healthy life expectancy        float64
Explained by: Freedom to make life choices   float64
Explained by: Generosity                     float64
Explained by: Perceptions of corruption      float64
Dystopia + residual                          float64
dtype: object
```

- All the different data types of the dataset.
- We notice that all but two of the columns are float64.
- The two columns ’Country name’ and ‘Regional indicator’ are objects.

Data Overview (4) - Missing Values

```
...  Country name                                    0
     Regional indicator                              0
     Ladder score                                    0
     Standard error of ladder score                  0
     upperwhisker                                    0
     lowerwhisker                                    0
     Logged GDP per capita                           0
     Social support                                  0
     Healthy life expectancy                         0
     Freedom to make life choices                    0
     Generosity                                      0
     Perceptions of corruption                       0
     Ladder score in Dystopia                        0
     Explained by: Log GDP per capita                0
     Explained by: Social support                    0
     Explained by: Healthy life expectancy           0
     Explained by: Freedom to make life choices      0
     Explained by: Generosity                        0
     Explained by: Perceptions of corruption         0
     Dystopia + residual                             0
     dtype: int64
```

- We check for any missing values that might be in the dataset.
- We conclude that there are no missing values in the dataset.

Data Feature Engineering (1) - Descriptive Statistics

	Ladder score	Standard error of ladder score	upperwhisker	lowerwhisker	Logged GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Ladder score in Dystopia	Explained by: Log GDP per capita	Explained by: Social support	Explained by: Healthy life expectancy	Explained by: Freedom to make life choices	Explained by: Generosity	Explained by: Perceptions of corruption	Dystopia + residual
count	153.00000	153.000000	153.000000	153.000000	153.000000	153.000000	153.000000	153.000000	153.000000	153.000000	1.530000e+02	153.000000	153.000000	153.000000	153.000000	153.000000	153.000000	153.000000
mean	5.47324	0.053538	5.578175	5.368304	9.295706	0.808721	64.445529	0.783360	-0.014568	0.733120	1.972317e+00	0.868771	1.155607	0.692869	0.463583	0.189375	0.130718	1.972317
std	1.11227	0.018183	1.096823	1.128631	1.201588	0.121453	7.057848	0.117786	0.151809	0.175172	2.227738e-16	0.372416	0.286866	0.254094	0.141172	0.100401	0.113097	0.563638
min	2.56690	0.025902	2.628270	2.505530	6.492642	0.319460	45.200001	0.396573	-0.300907	0.109784	1.972317e+00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.257241
25%	4.72410	0.040698	4.826248	4.603149	8.350645	0.737217	58.961712	0.714839	-0.127015	0.683019	1.972317e+00	0.575862	0.986718	0.495443	0.381457	0.115006	0.055805	1.629928
50%	5.51500	0.050606	5.607728	5.430644	9.456313	0.829204	66.305145	0.799805	-0.033665	0.783122	1.972317e+00	0.918549	1.203987	0.759818	0.483293	0.176745	0.098435	2.046272
75%	6.22850	0.060677	6.363886	6.138881	10.265124	0.906747	69.289192	0.877709	0.085429	0.849151	1.972317e+00	1.169229	1.387139	0.867249	0.576665	0.255510	0.163064	2.350267
max	7.80870	0.120590	7.869766	7.747634	11.450681	0.974670	76.804581	0.974998	0.560664	0.935585	1.972317e+00	1.536676	1.547567	1.137814	0.693270	0.569814	0.533162	3.440810

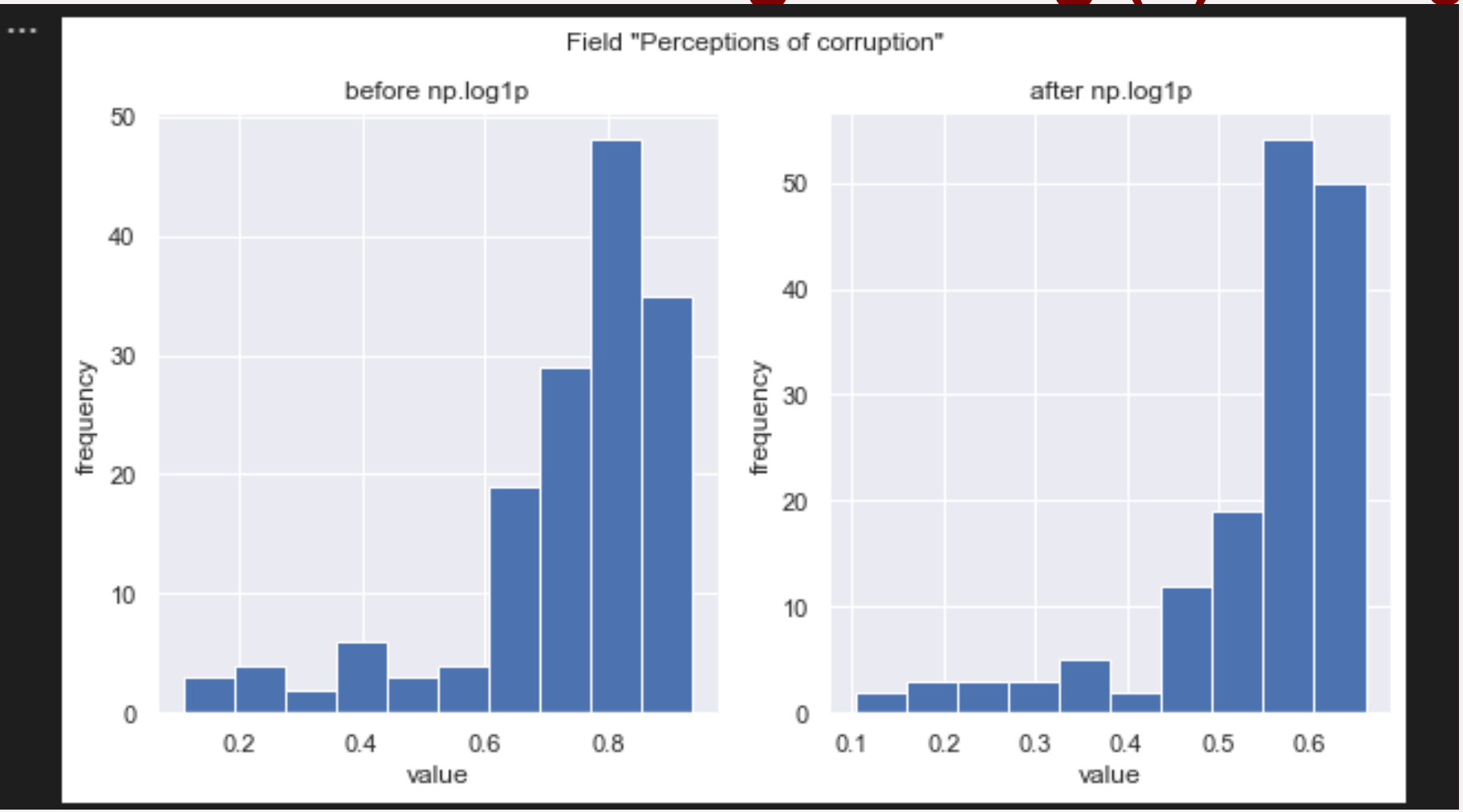
- We look for the descriptive statistics of the dataset

Data Feature Engineering (2) - Categorical Data

	Skew
Explained by: Perceptions of corruption	1.677125
Standard error of ladder score	1.450652
Explained by: Generosity	0.848921
Generosity	0.848921
Social support	-1.165826
Explained by: Social support	-1.165826
Perceptions of corruption	-1.677125

- We cleaned the data by using hot encoding to convert categorical variables to dummies
- We then logged transformed skew variables
- We then show the skewed columns

Data Feature Engineering (3) - Categorical Data



- We then looked at what happens to one of these variables when we applied np.log1p visually.
- We chose the variable "Perceptions of corruption" to apply np.log1p on and show the “before” and “after” of its application.

Data Feature Engineering (1) - Numerical Data

	count	mean	std	min	25%	50%	75%	max
Logged GDP per capita	153.0	9.295706	1.201588e+00	6.492642	8.350645	9.456313	10.265124	11.450681
Social support	153.0	0.590240	7.033987e-02	0.277222	0.552284	0.603881	0.645399	0.680401
Healthy life expectancy	153.0	64.445529	7.057848e+00	45.200001	58.961712	66.305145	69.289192	76.804581
Freedom to make life choices	153.0	0.783360	1.177863e-01	0.396573	0.714839	0.799805	0.877709	0.974998
Generosity	153.0	-0.025914	1.490434e-01	-0.357972	-0.135837	-0.034244	0.081976	0.445111
Perceptions of corruption	153.0	0.544100	1.122217e-01	0.104166	0.520589	0.578366	0.614727	0.660410
Ladder score in Dystopia	153.0	1.972317	2.227738e-16	1.972317	1.972317	1.972317	1.972317	1.972317
Explained by: Log GDP per capita	153.0	0.868771	3.724157e-01	0.000000	0.575862	0.918549	1.169229	1.536676
Explained by: Social support	153.0	0.757885	1.491880e-01	0.000000	0.686484	0.790268	0.870096	0.935139
Explained by: Healthy life expectancy	153.0	0.692869	2.540935e-01	0.000000	0.495443	0.759818	0.867249	1.137814
Dystopia + residual	153.0	1.972317	5.636378e-01	0.257241	1.629928	2.046272	2.350267	3.440810

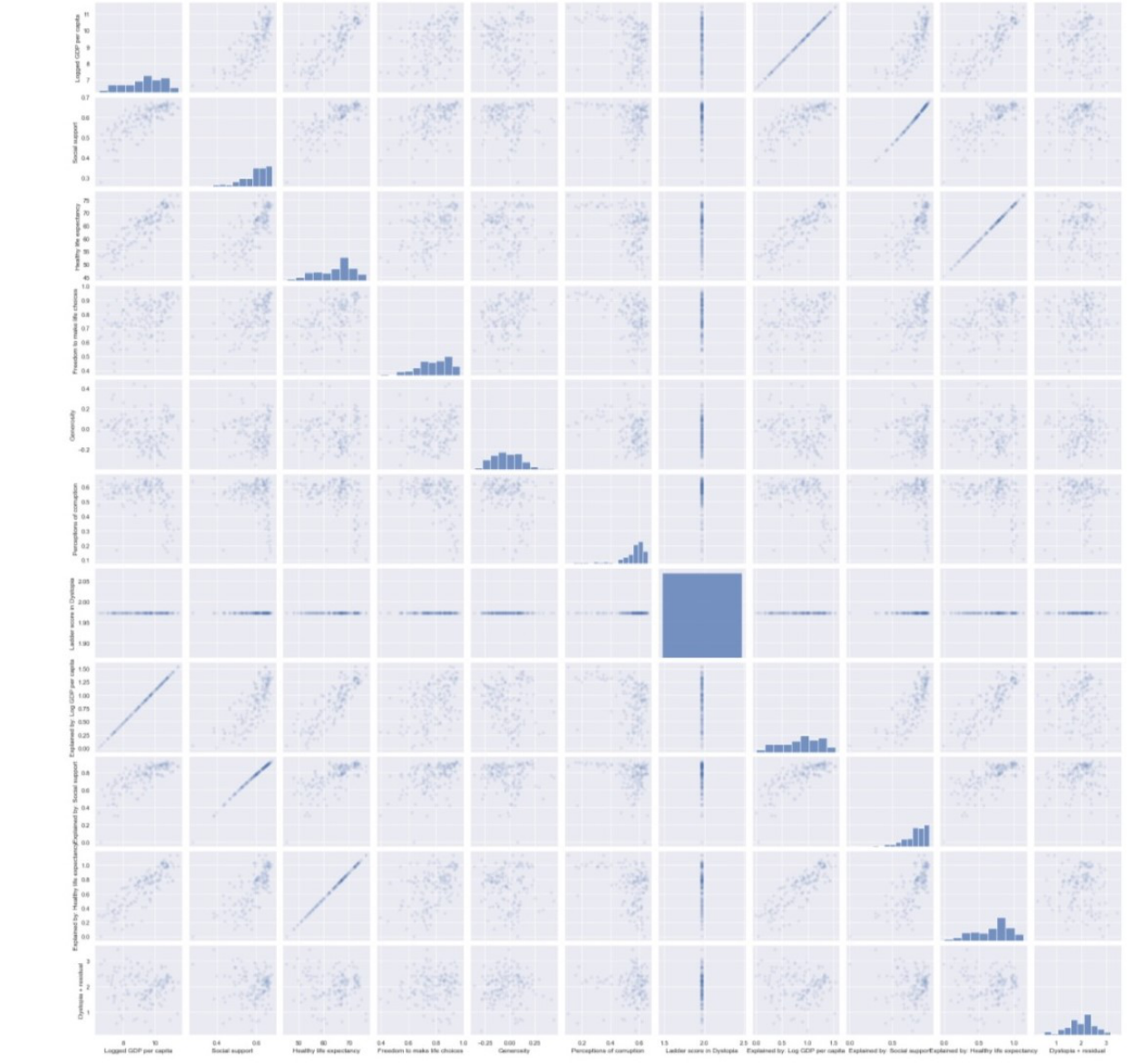
- We performed skew transformations on numerical data.
- We picked out a few numeric columns to illustrate basic feature transformations.
- Then took a look at the summary statistics of the subset data

Data Feature Engineering (2) - Numerical Data

```
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Logged GDP per capita                 153 non-null    float64
 1   Social support                        153 non-null    float64
 2   Healthy life expectancy               153 non-null    float64
 3   Freedom to make life choices          153 non-null    float64
 4   Generosity                            153 non-null    float64
 5   Perceptions of corruption             153 non-null    float64
 6   Ladder score in Dystopia              153 non-null    float64
 7   Explained by: Log GDP per capita      153 non-null    float64
 8   Explained by: Social support          153 non-null    float64
 9   Explained by: Healthy life expectancy 153 non-null    float64
 10  Dystopia + residual                   153 non-null    float64
dtypes: float64(11)
memory usage: 13.3 KB
```

- We noted that there were no NaN values in the data set and concluded that our dataset was perfectly filtered.

Data Feature Engineering (3) - Numerical Data



- We then generated pairplot visuals to better understand the target and feature-target relationships
- We saw that the target variable does not seem to have a linear relationship with any of the features.
- As far as the relationship between the various features, we wanted to ensure that there was not a excessive multi-collinearity between each of them, as that could throw off our interpretation of something like linear regression.

Data Feature Engineering (4) - Numerical Data

```
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Logged GDP per capita                 153 non-null    float64
 1   Social support                        153 non-null    float64
 2   Healthy life expectancy               153 non-null    float64
 3   Freedom to make life choices          153 non-null    float64
 4   Generosity                            153 non-null    float64
 5   Perceptions of corruption             153 non-null    float64
 6   Ladder score in Dystopia              153 non-null    float64
 7   Explained by: Log GDP per capita      153 non-null    float64
 8   Explained by: Social support          153 non-null    float64
 9   Explained by: Healthy life expectancy 153 non-null    float64
dtypes: float64(10)
memory usage: 12.1 KB
```

- We separated our features from our target.
- This gave us feature/target data X, y which meant that were nearly ready to fit and evaluate a baseline model using our current feature set.

Data Feature Engineering (5) - Numerical Data

	1	Healthy life expectancy	Explained by: Healthy life expectancy	Healthy life expectancy^2	Healthy life expectancy Explained by: Healthy life expectancy	Explained by: Healthy life expectancy^2
0	1.0	71.900825	0.961271	5169.728708	69.116208	0.924043
1	1.0	72.402504	0.979333	5242.122581	70.906130	0.959092
2	1.0	74.102448	1.040533	5491.172727	77.106056	1.082709
3	1.0	73.000000	1.000843	5329.000000	73.061569	1.001688
4	1.0	73.200783	1.008072	5358.354600	73.791652	1.016209
...	...	...	...	...	...	...
148	1.0	45.200001	0.000000	2043.040069	0.000000	0.000000
149	1.0	61.098846	0.572383	3733.069036	34.971958	0.327623
150	1.0	55.617260	0.375038	3093.279608	20.858564	0.140653
151	1.0	51.000000	0.208809	2601.000000	10.649261	0.043601
152	1.0	52.590000	0.266052	2765.708116	13.991650	0.070783

153 rows × 6 columns

- We then created a **train/validation split** before we fitted and scored the model.
- We repeatedly split X, y into the same train/val partitions and fitted new models as we updated our feature set.
- We then added quadratic **polynomial transformations** to features that seem to have linear relationships with each other.

Hypothesis Formulation

From the above analysis, we got some insight about the data. Since this data can be used to make a prediction model that can determine which factors have strong or weak correlation with the Happiness level of a country. I made several hypothesis regarding to this matter.

- **Hypothesis 1**:

H0 = Healthy life expectancy has a positive correlation with citizens' happiness.

Ha = Healthy life expectancy has a negative correlation with citizens' happiness.

- **Hypothesis 2**:

H0 = Logged GDP per capita has a positive correlation with citizens' happiness.

Ha = Logged GDP per capita has a negative correlation with citizens' happiness.

- **Hypothesis 3**:

H0 = The Freedom to make life choices has a positive correlation with citizens' happiness.

Ha = The Freedom to make life choices has a negative correlation with citizens' happiness

Hypothesis Testing (1)

	Healthy life expectancy	Logged GDP per capita	Freedom to make life choices	Dystopia + residual
0	71.900825	10.639267	0.949172	2.762835
1	72.402504	10.774001	0.951444	2.432741
2	74.102448	10.979933	0.921337	2.350267
3	73.000000	10.772559	0.948892	2.460688
4	73.200783	11.087804	0.955750	2.168266

- We made a new dataframe and obtained a preview of it.
- We noted that all values were numerical.

Hypothesis Testing (2)

	Healthy life expectancy	Logged GDP per capita	Freedom to make life choices	Dystopia + residual
Healthy life expectancy	1.000000	0.848469	0.448846	-0.039948
Logged GDP per capita	0.848469	1.000000	0.419019	-0.062063
Freedom to make life choices	0.448846	0.419019	1.000000	0.062571
Dystopia + residual	-0.039948	-0.062063	0.062571	1.000000

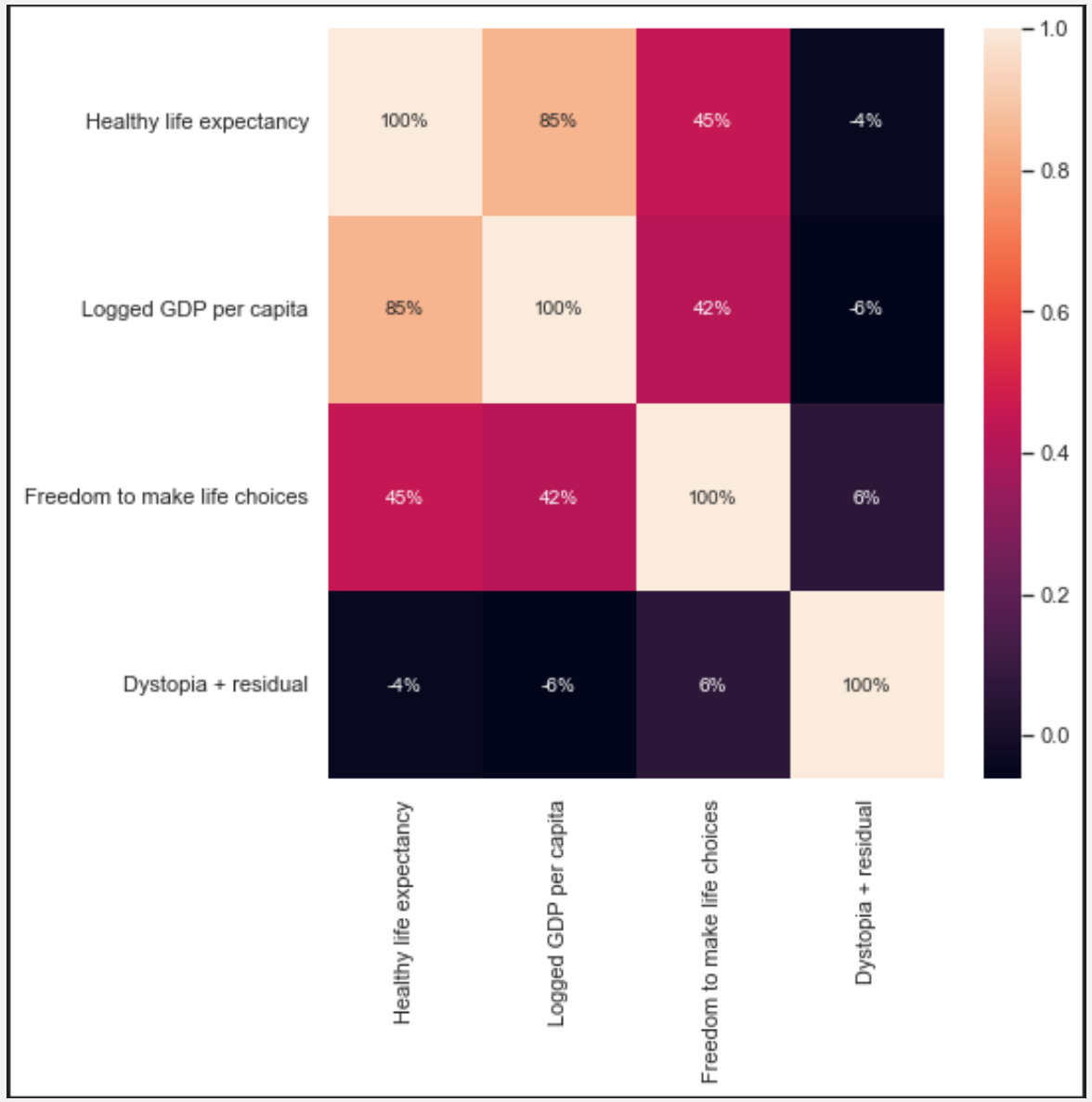
- We obtained the correlation table of the dataset.
- We used a correlation coefficient formulas to find how strong a relationship is between the data. The formulas return a value between -1 and 1, where:

  1 indicates a strong positive relationship.

  -1 indicates a strong negative relationship.

  0 indicates no relationship at all.

Hypothesis Testing (3)



- We generated a visualization of the correlation using a heatmap.
- As we can see from the heatmap, we can conclude that:

1. Healthy life expectancy and Dystopia + residual have a **negative and weak correlation**,
2. Logged GDP per capita and Dystopia + residual have a **negative and weak correlation**, and
3. Freedom to make life choices and Dystopia + residual have a **positive but weak correlation**.

Further Data Engineering and Analyzing

- For further analysis, I suggest that we do more insight finding for each variable by visualization as well as more in-depth feature engineering.
- As far as data engineering goes, I suggest the following:

1. Apply Backward Stepwise Regression.
2. We calculate deviance of a row's feature value from the mean value of the category that row belongs to.
3. I also recommend that we obtain a more comprehensive dataset with all 195 countries included in it to gain a more accurate picture of world happiness.

Beyond this point, additional feature engineering can provide significant, but potentially diminishing returns. Whether it's worth it depends on the use case for the model.

Conclusion

- This analysis has shown us that linear regression might not be a good fit for this particular dataset.
- It may however, make for a good base model.
- Further internet exploration of the World Happiness Report(s) for the year 2020 might also prove to be beneficial for our data gathering (API's, etc.)

The Jupyter Notebook for this project can be viewed at:

https://github.com/KhobieMaseko/IBM-Machine-Learning-Professional-Certificate-/blob/main/Project%201%20EDA%20on%20World%20Happiness%20Report%202020.ipynb

THANK YOU