Team **Codebakers**

MahilAI Datathon 2.0

April 15,2022

Team Members –

- Nikhil Mohite
- Hithesh Patel

Problem Statement(s)

R1: Clean your dataset, perform Exploratory Data Analysis and come up with meaningful visualisations & statistical analysis to derive your initial hypothesis and note down your findings from the dataset. All the work should be done in a python notebook with clear markdowns which explain your findings.

R2: Using PCA(Principal Component Analysis) to be able to determine features that significantly determine the happiness index for that nation and provide reasonable explanation to some of the visible inferences that have come about due to these operations.

Dataset:

https://www.kaggle.com/datasets/unsdsn/world-happiness

Round 1

Data Pre-processing:

- Given dataset was preprocessed. All column names were renamed according to conventions, spaces were replaced with underscore(_) and converted to lowercase.

- Added 3 new columns, percentage change in score,GDP per capita and life expectancy

- Identification of rows/columns with NaN values.
  Only one row of 2018.csv dataset had NaN value, other datasets didn't have null/NaN values in them.

```
[10] # Where is the NaN value?
     df2[df2["perceptions_of_corruption"].isna()]
```

	overall_rank	country_or_region	Score	gdp_per_capita	social_support	healthy_life_expectancy	freedom_to_make_life_choices	Generosity	perceptions_of_corruption
19	20	United Arab Emirates	6.774	2.096	0.776	0.67	0.284	0.186	NaN

Data Cleaning:

- The NaN value was replaced with average value of the corruption indices of other years.

```
[41] corruption_value_2017=float(df1[df1.country=="United Arab Emirates"]["corruption_index"])
     corruption_value_2019=float(df3[df3.country_or_region=="United Arab Emirates"]["perceptions_of_corruption"])

     avg_value = round((corruption_value_2017 + corruption_value_2019)/2,4)

     df2.loc[df2.country_or_region=="United Arab Emirates","perceptions_of_corruption"] = avg_value

     df2[df2.country_or_region=="United Arab Emirates"]
```

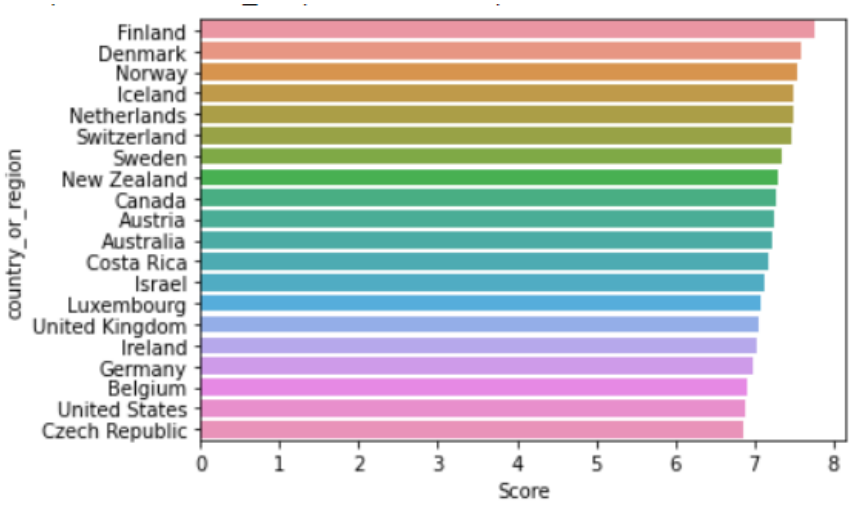
	overall_rank	country_or_region	Score	gdp_per_capita	social_support	healthy_life_expectancy	freedom_to_make_life_choices	Generosity	perceptions_of_corruption
19	20	United Arab Emirates	6.774	2.096	0.776	0.67	0.284	0.186	0.2532

Data Analysis and Visualization:

- Analysis was done for top_20, middle_20 and bottom_20 countries.

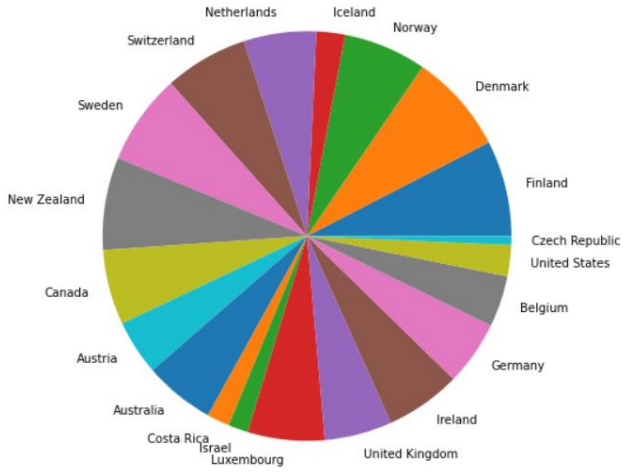
Happiness Score

Among Top 20 countries, **Finland** was first and **Czech Republic** was last.



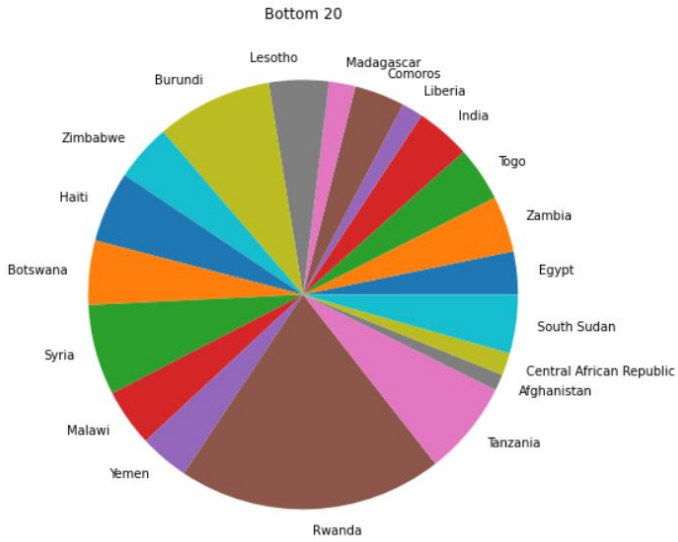
Corruption/ Trust in Government

Comparing the corruption levels in the top 20 countries, we found that **Denmark, New Zealand** and **Finland** have relatively high levels of corruption whereas **Czech Republic** has low levels of corruption.

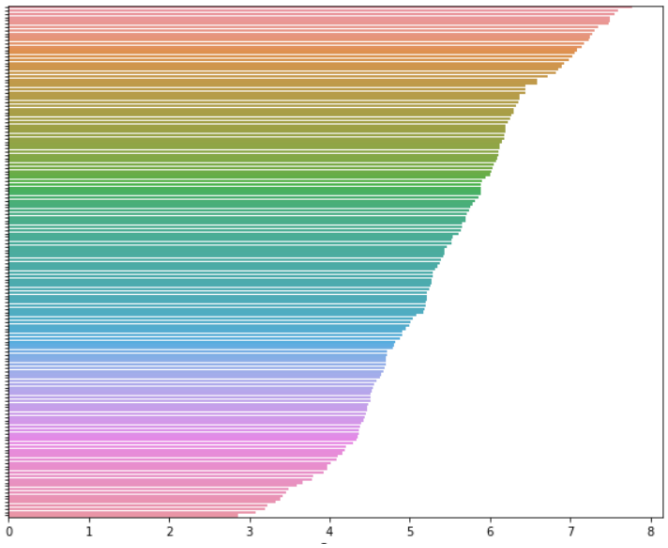


Bottom 20 Countries

Highest levels of corruption were shown by **Rawanda**



Difference in **Happiness Scores**:



Clearly the difference is very significant.

Percentage change in Happiness Score, GDP per Capita and Life Expectancy

Forumula:

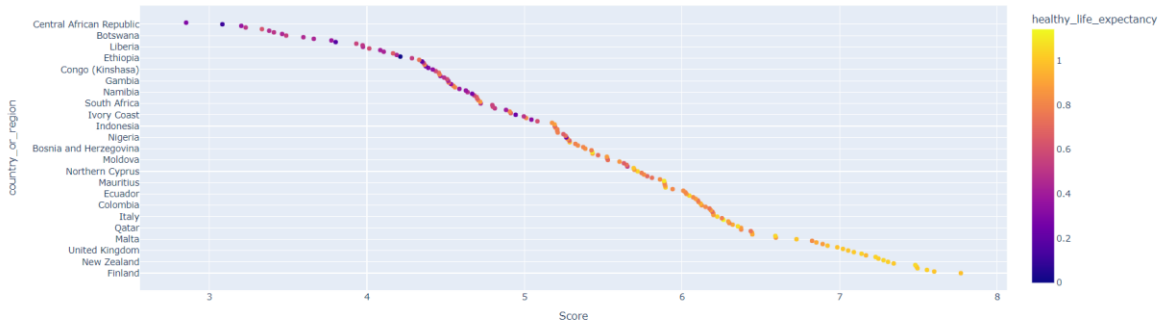
% = ((value in 2019 – value in 2018)/(value in 2018))*100

	country_or_region	percent_change_in_score	percent_change_in_life_expectancy
112	Namibia	1.487639	inf
90	Lebanon	0.814743	1597.916667
106	Albania	1.027617	992.500000
154	Central African Republic	0.000000	950.000000
130	Myanmar	1.371774	947.169811
140	Liberia	4.385504	460.759494
97	Ghana	0.281012	322.608696
139	India	1.286579	178.672986
117	Guinea	1.956375	146.710526
124	Bangladesh	1.804889	145.084746

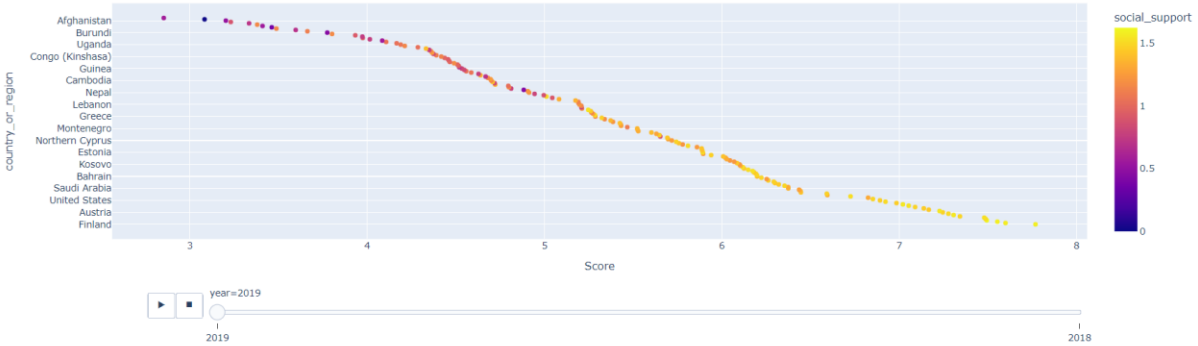
	country_or_region	percent_change_in_score	percent_change_in_gdp_per_capita
97	Ghana	0.281012	inf
148	Syria	-0.944206	714.473684
131	Chad	2.473498	407.246377
112	Namibia	1.487639	243.359375
155	South Sudan	-1.790017	236.263736
147	Botswana	-2.624232	230.476190
133	Ethiopia	2.880461	156.488550
134	Swaziland	1.225667	151.863354
139	India	1.286579	119.476744
87	Algeria	0.230814	111.392405

	country_or_region	percent_change_in_score
141	Comoros	4.690382
140	Liberia	4.385504
142	Madagascar	4.213037
144	Burundi	3.937225
132	Ukraine	3.389021
143	Lesotho	2.979415
133	Ethiopia	2.880461
131	Chad	2.473498
138	Togo	2.150538
116	Iran	2.064632

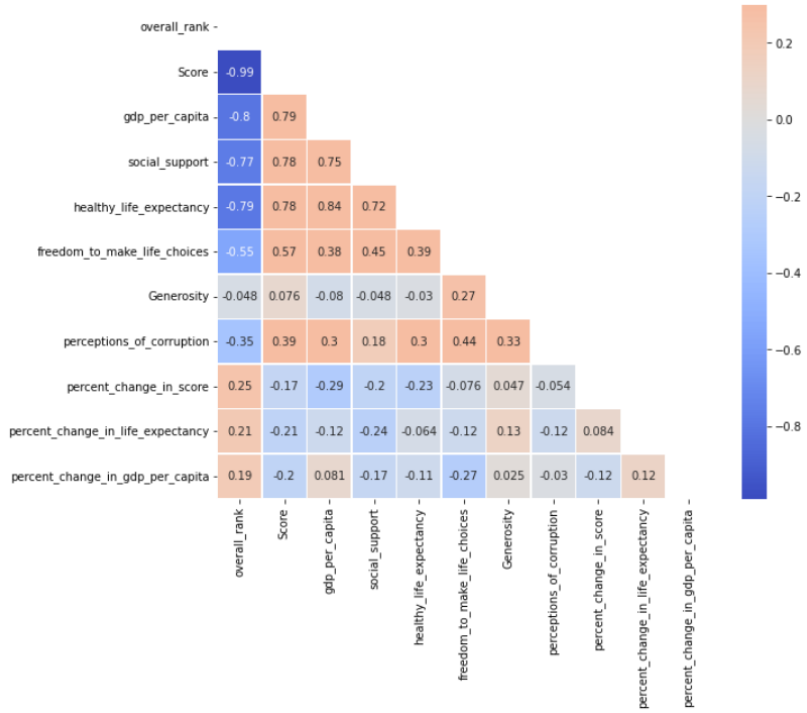
Interactive Graphs:



Comparison between 2018 and 2019



Correlation Graph/Matrix of features

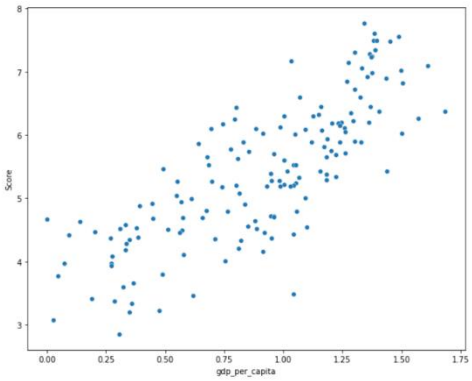


`Conclusions from Round 1:`

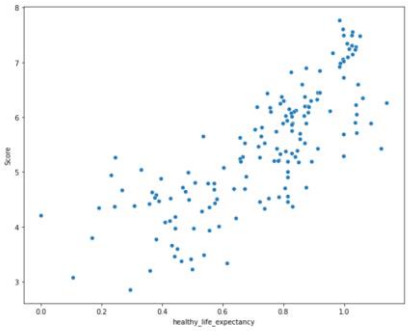
- **Only one NaN value** was found in 2018 dataset, other datasets didn't have any NaN values. This was **replaced with average value**.
- **3 new columns** were added in latest dataset i.e., 2019.csv
- These columns contain **percentage change in score, GDP per capita and life expectancy**.
- **Finland** was found to be the happiest country and **South Sudan** the least happiest country
- **Rwanda** turns out to be having highest levels of corruption.
- **Comoros** followed by **Liberia** show maximum increase in happiness score.
- From Correlation Graph, it can be said that the **Happiness Score is affected by all features except for Generosity which is having correlation value of almost 0**
- **Correlation graph/Matrix** helps us to filter features which are of high value/importance while making predictions

Round 2

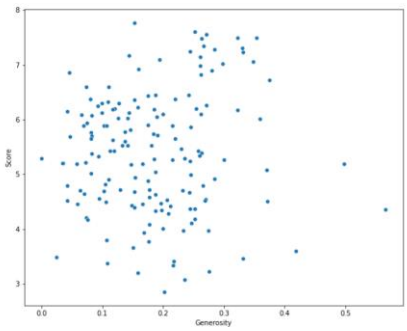
`Scatter Plots to describe correlation:`



Positive Correlation between `Score` and `gdp_per_capita` (r = 0.79)



Positive Correlation between `Score` and `healthy_life_expectancy` (r = 0.78)



No proper correlation between `Score` and `Generosity` (r=0.076)

Feature(s) Selection

Methodology –

- Correlation Matrix/Graph
- Permutation Graph(using Decision Tree)
- Feature Importance(using ExtraTreesRegr)

Why use Permutation Graph?

Using Permutation Graph in Data Analysis is very common and very effective. This Graph is used to observe the effect of variation of variables/features on the model. The variation is done by random shuffling.

Results from Permutation Graph:

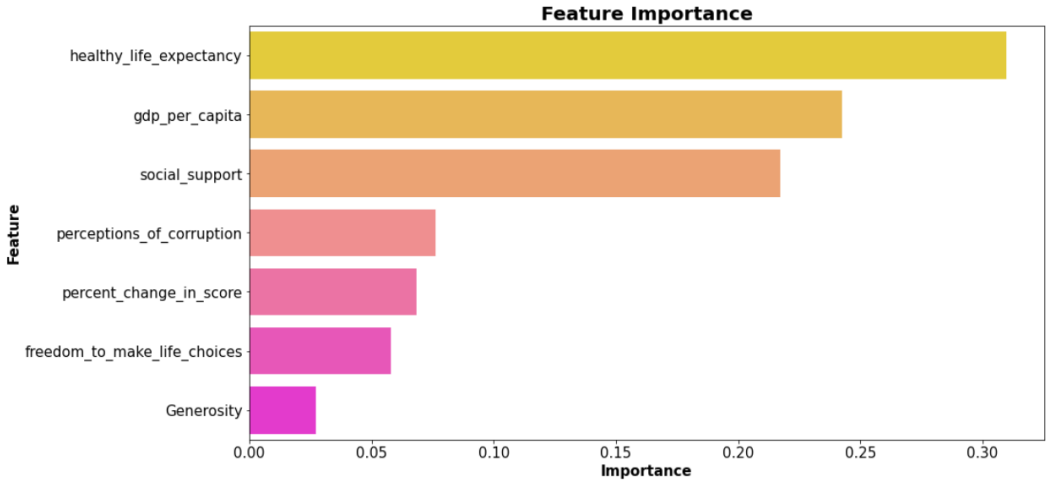
2019 Dataset

Weight	Feature
0.3214 ± 0.0875	social_support
0.0987 ± 0.0559	gdp_per_capita
0.0641 ± 0.0712	healthy_life_expectancy
0.0494 ± 0.0254	percent_change_in_score
0.0325 ± 0.0197	freedom_to_make_life_choices
-0.0125 ± 0.0148	Generosity
-0.0183 ± 0.0155	perceptions_of_corruption

2018 Dataset

Weight	Feature
0.3160 ± 0.0878	gdp_per_capita
0.1348 ± 0.0925	social_support
0.1206 ± 0.0927	healthy_life_expectancy
0.0025 ± 0.0577	freedom_to_make_life_choices
-0.0007 ± 0.0236	Generosity
-0.0053 ± 0.0129	perceptions_of_corruption

Results from Feature Importance Graph:



Conclusion/Inference:

Social Support, Health Life Expectancy and **GDP per capita** are most import features which are affecting Happiness Score.

Explanation:

Social Support - People will be happy when the government and other local authorities provide facilities/amenities to them. For example: Health Care, Transportation

Health Life Expectancy – Higher life expectancy means the people are living a healthy life. Living a healthy life is also indicator that people are happy.

GDP per capita is one of the major contributors in determining the poverty levels in a country. Higher is the GDP per capita, lesser are the poverty levels and people will have the power to buy basic necessities.

Round 3

Problem Statement(contd.):

- Given the parameters for any other country, estimate the happiness score. (Example: Multiple Linear Regression model)
- Find top 5 countries which support a high happiness index in terms of Economy(GDP) and Trust in the government.
- Model the Happiness Index over a time span of 3 years.
  - Which countries maintain a consistent HI value?
  - Which parameters vary largely with time?

1. Regression Model(s) to estimate Happiness Score

```
Multiple Linear Regression
Mean Absolute Error - 7.54382311604273e-16
Mean Squared Error - 1.1744419566511538e-30
Accuracy - 100.0%
```

```
Decision Tree Regressor
Mean Absolute Error - 0.01589743367219562
Mean Squared Error - 0.0008866920445581641
Accuracy - 99.92798%
```

```
Random Forest Regressor
Mean Absolute Error - 0.009416346659418812
Mean Squared Error - 0.000521676673037605
Accuracy - 99.95762%
```

2. Top 5 Countries over the span of 3 years

2017

	happiness_rank	country	Score	gdp_per_capita	perceptions_of_corruption
34	35	Qatar	6.375	1.870766	0.439299
25	26	Singapore	6.572	1.692278	0.464308
1	2	Denmark	7.522	1.482383	0.400770
8	9	Sweden	7.284	1.494387	0.384399
3	4	Switzerland	7.494	1.564980	0.367007

2018

	overall_rank	country_or_region	Score	gdp_per_capita	perceptions_of_corruption
33	34	Singapore	6.343	1.529	0.4570
2	3	Denmark	7.555	1.351	0.4080
19	20	United Arab Emirates	6.774	2.096	0.2532
8	9	Sweden	7.314	1.355	0.3830
0	1	Finland	7.632	1.305	0.3930

2019

	overall_rank	country_or_region	Score	gdp_per_capita	perceptions_of_corruption
33	34	Singapore	6.262	1.572	0.453
1	2	Denmark	7.600	1.383	0.410
0	1	Finland	7.769	1.340	0.393
6	7	Sweden	7.343	1.387	0.373
13	14	Luxembourg	7.090	1.609	0.316

Most frequently occurring in Top 5 :

`Singapore, Denmark, Sweden, Finland`

3. Countries which maintain consistent HI value

- Singapore(6.572,6.343,6.262)
- Denmark(7.522,7.555,7.6)
- Sweden(7.284,7.314,7.343)
- Finland(7.469,7.632,7.769)
- Norway(7.5944,7.594,7.554)

Consistency of HI value is relatively high in Top countries when compared to bottom countries.

The parameters which vary largely with time include

- GDP per capita: The profits made by a particular country increases(mostly) with time.
- Health Life Expectancy: The life expectancy of the people might get affected due to a disease/poor health infrastructure
- Social support: The facilities provided by the government/local authorities depends on the government.(which party in power now??)

Conclusion

All the 3 model(s) i.e, Multiple Linear Regression, Decision Tree Regr, Random Forest Regr, are working with very good accuracy.

The Model(s) present in this project are proof for our hypothesis that the important features selected at the end of Round 2 were True.

Colab Notebook:

https://colab.research.google.com/drive/1IdTDoTHLFqV3EEq26RDnuYiIB659ZBl6?usp=sharing