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.

Dataset:

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

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"]
```

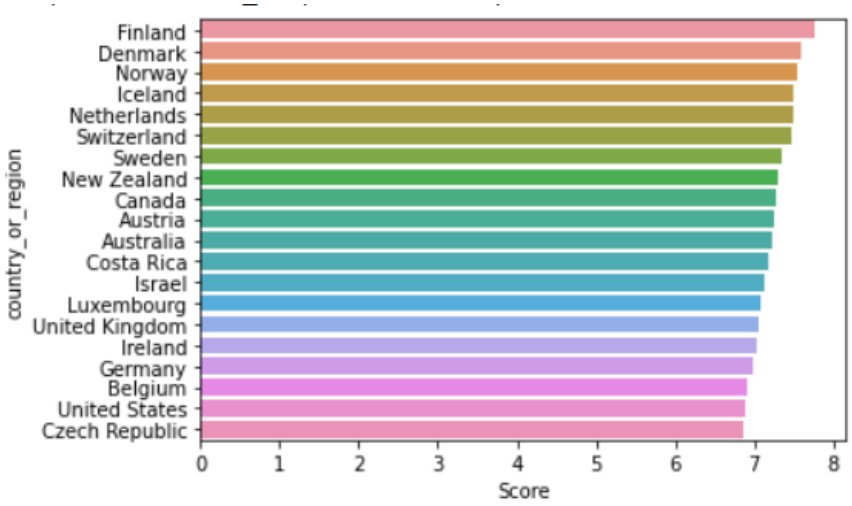
	overall_rank	country_or_region	Score	gdp_per_capita	social_support	healthy_life_expectancy	freedom_to_make_life_choices	Generosity	perceptions_of_corruption
19	20	United Arab Emirates	6.774	2.096	0.776	0.67	0.284	0.186	0.2532

Data Analysis and Visualization:

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

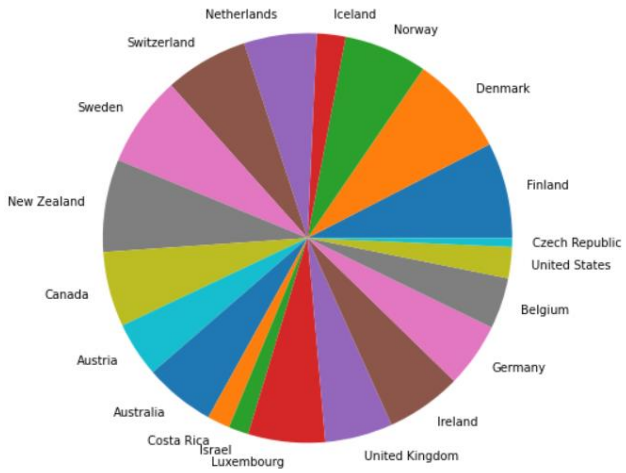
Happiness Score

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



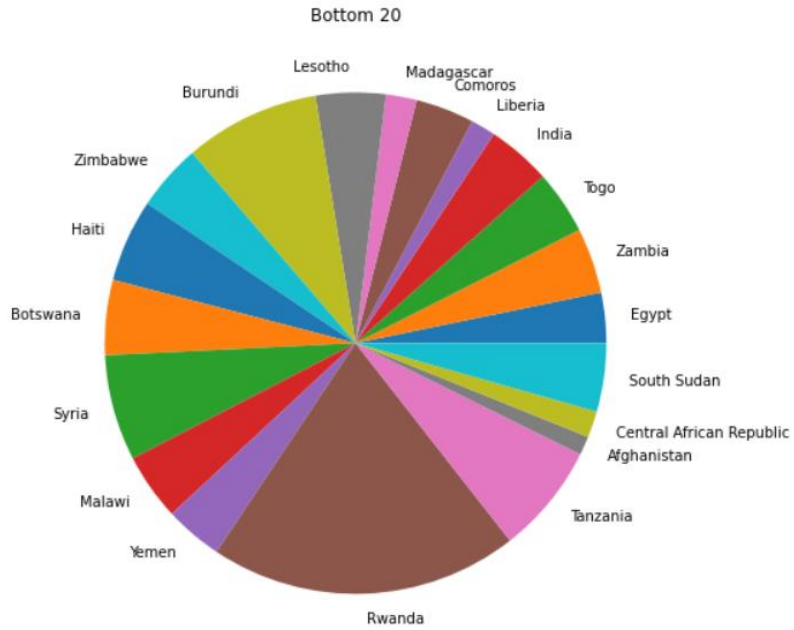
Corruption/ Trust in Government

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.

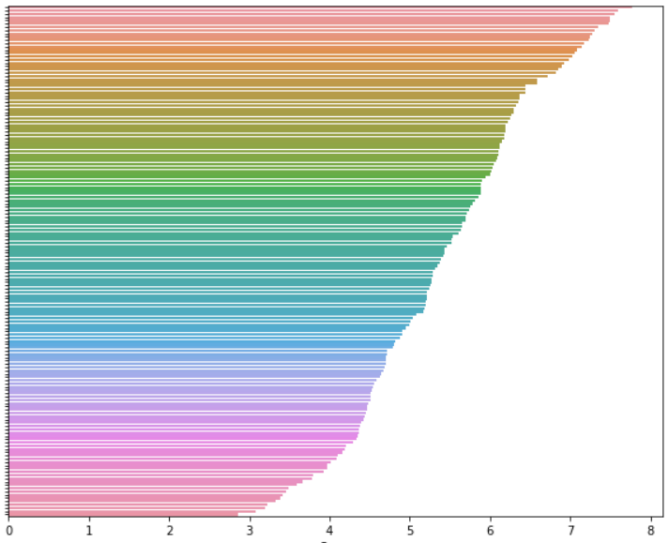


Bottom 20 Countries

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:

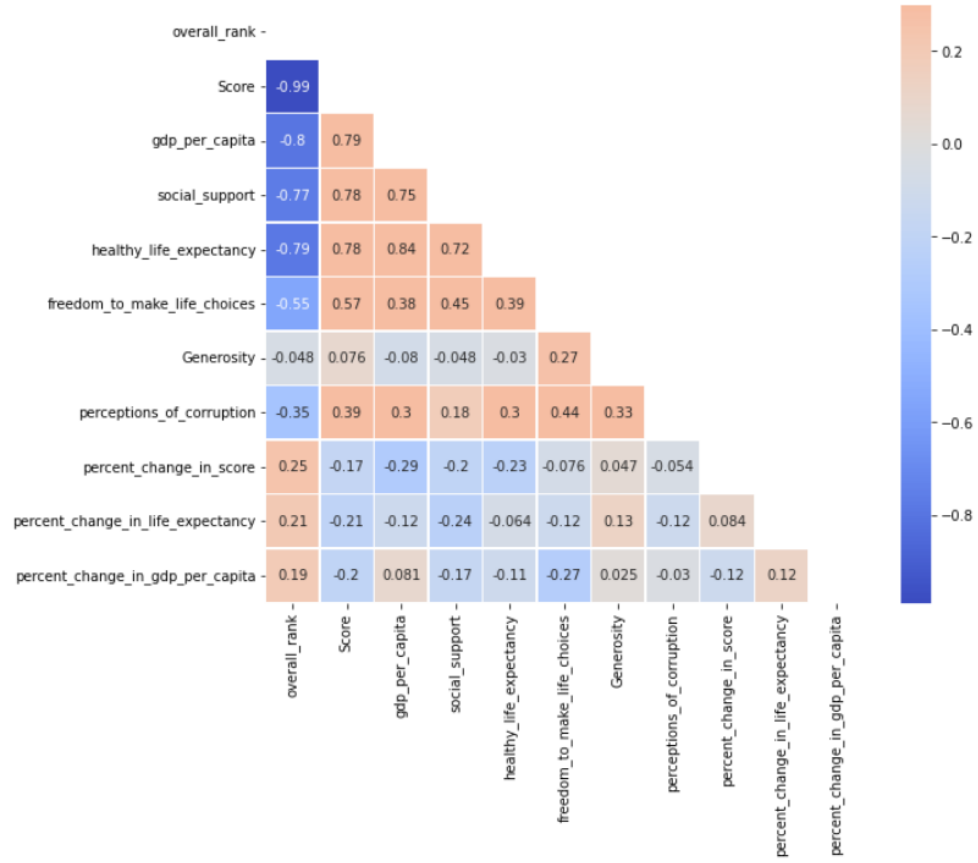
% = ((value in 2019 – value in 2018)/(value in 2018))*100

	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

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