CLOUD FRAMEWORKS EXAM (PART 2)

DATASET: WORLD HAPPINESS INDEX (WHI)

BABD
ZAKRA CHACHAR

DATE: 11 FEBRUARY 2024

INTRODUCTION AND ASSUMPTIONS:

1. The dataset analyzes the World’s Happiness based on Gallup World Poll for 155 countries.The dataset containes 8 variables which contribute towards a Happiness Score and a Happiness Rank, annually from 2015 – 2019

2. For the assignment, years 2015, 2017 and 2019 have been analyzed.

3. One key point to note is that 1 variable was changed from 2017 to 2019 and therefore that variable (Social Support) was dropped from the analyses.

4. It is assumed that the statistical inferences made in the assessment are observations that may guide further investigation. The analyses is conducted for the purpose of the Cloud Frameworks course.

QUERIES CONDUCTED FOR THE DATASET

- Data frames, data standardization and data cleaning
- Summarizing dataframes 2015 and 2019 and determining the countries and regions participating
- Association Measures for 2015, 2017, and 2019 to analyze the association between variables
- WHI trends in the 3 years
- Finding countries with greatest change in WHI
- Analyzing the change in variables for the countries with the highest increase and highest decrease in WHI
- Identifying the maximum change in each variable for 2015-2019 and 2017-2019
- Identifying the countries where the maximum change occurred for both time periods
- Trend in the happiness score for the countries with maximum change to determine if it increased or decreased.

Note: This presentation does not contain all visualizations or queries conducted, however, the assessment is present. The attached notebook contains the details of the queries.

Association Measures for 2015, 2017, and 2019 were measured through a scattermatrix.

Only the matrix for 2019 has been shared in this presentation.

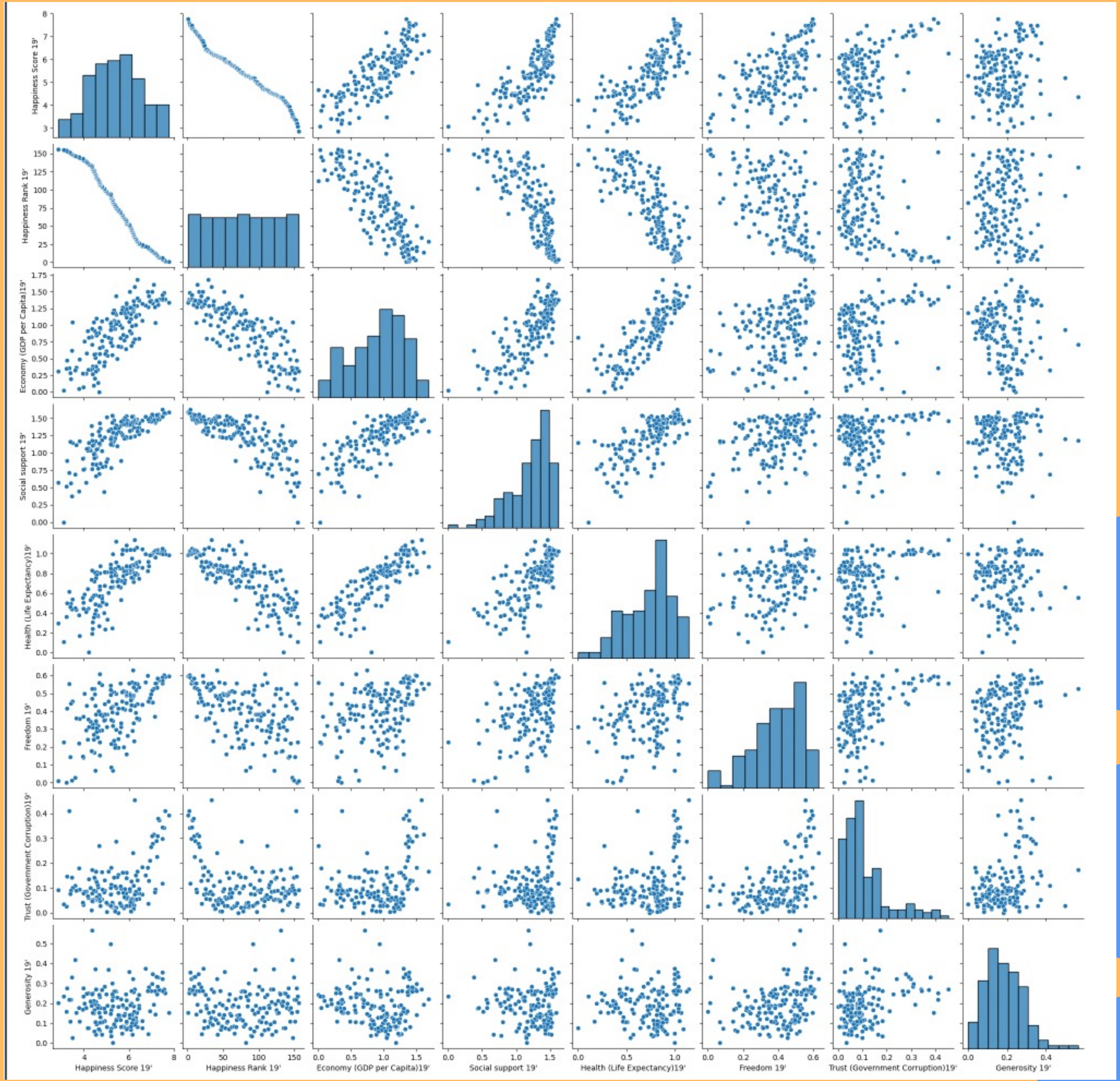
Note that Happiness Rank is showing a negative association because the rank takes an ascending order (from 1 to n).

GDP and Happiness Score have a positive association with each measure.

Freedom and Health are also positively associated.

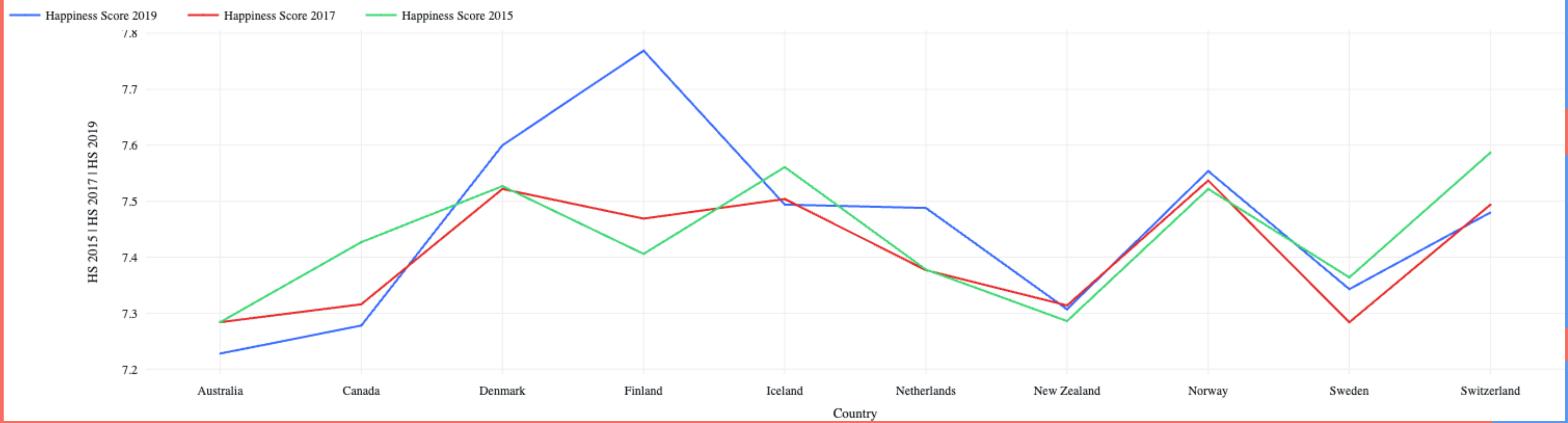
Trust and Generosity seem to have a concentrated range and suggest a higher coefficient value.

The association generally had a similar relationship for all 3 years



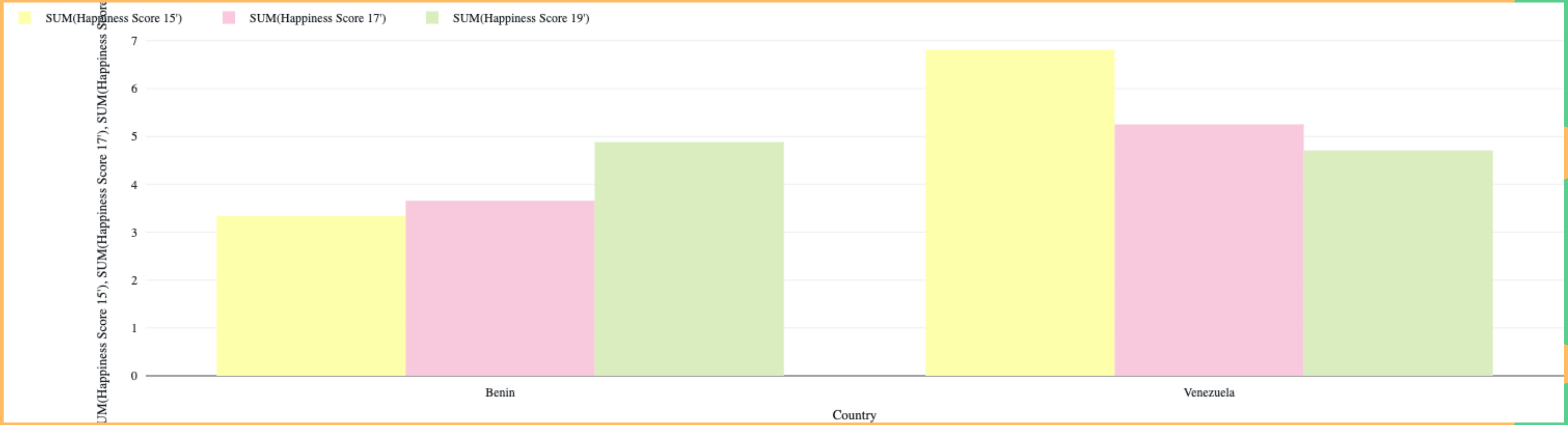
WHI trends in the 3 years

A joined dataframe shows the top 10 countries' happiness trend over the years. The same countries held the top spot. **FINLAND's** Happiness Score increased the most from 7.40 in 2015 to 7.76 in 2019. The last 10 countries were also analyzed, present in the notebook.



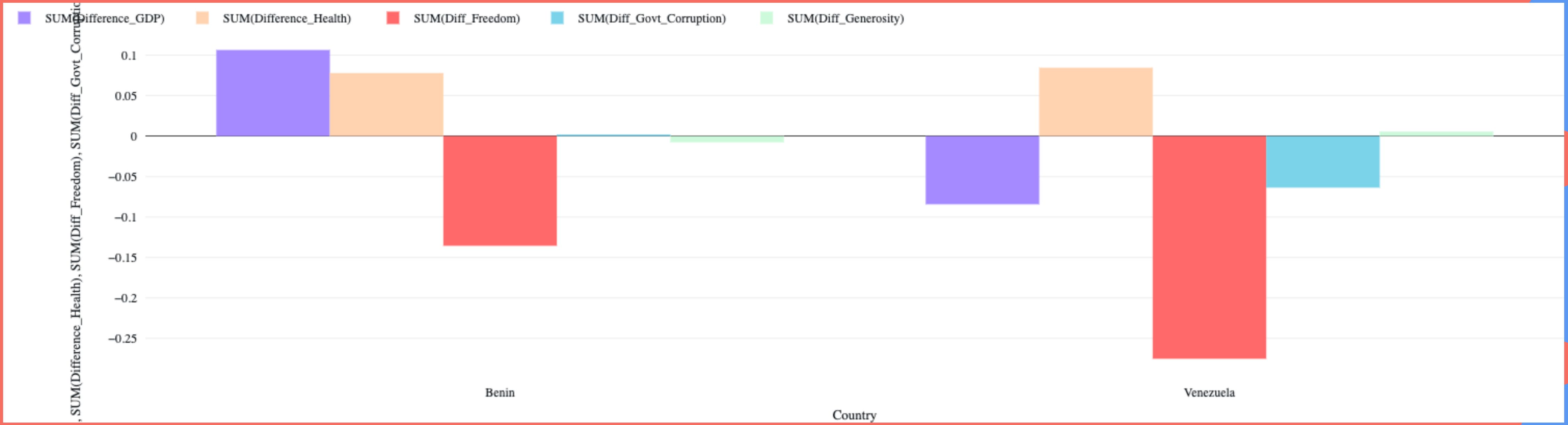
Finding countries with greatest change in WHI

A udf function was used to find the greatest change in WHI (2015 to 2019). **Venezuela** was found to have the greatest decrease in Happiness Score and Rank, and **Benin** had the most increase.



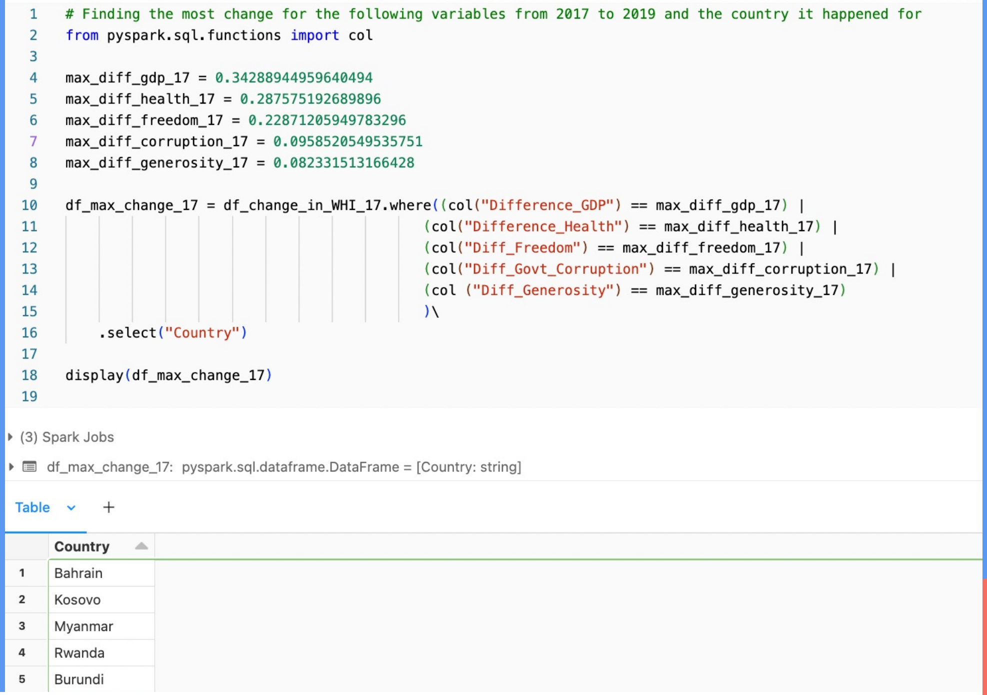
Analysing the delta in variables for the countries with the highest increase and highest decrease in WHI

The udf function was combined with a condition, where only Venezuela's and Benin's variables's difference was shown from 2015 to 2019. The graph shows that Venezuela's Freedom, Trust in Government and GDP all worsened in 2019. In comparision, Benin's GDP and Health increased.



Countries with greatest change in any single variable and their Happiness Score during that time

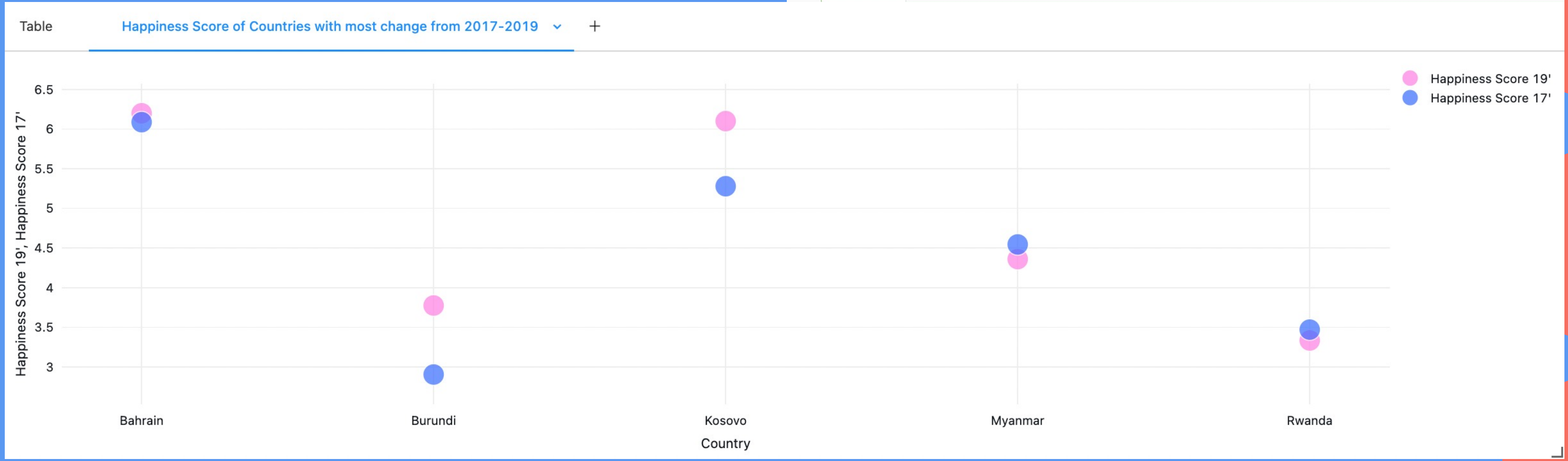
A difference was computed for each of the variables between 2017 and 2019. Each variable's maximum positive difference was identified. A boolean condition was used for each of the variable's maximum difference and the countries were identified. The graph below shows Happiness Rank for 2017 and 2019, representing comparatively greatest increase in Burundi's and Kosovo's HS. A similar anaylses for 2015-2019 is contained within the notebook



```
# Finding the most change for the following variables from 2017 to 2019 and the country it happened for
from pyspark.sql.functions import col

max_diff_gdp_17 = 0.34288944959640494
max_diff_health_17 = 0.287575192689896
max_diff_freedom_17 = 0.22871205949783296
max_diff_corruption_17 = 0.0958520549535751
max_diff_generosity_17 = 0.082331513166428

df_max_change_17 = df_change_in_WHI_17.where((col("Difference_GDP") == max_diff_gdp_17) |
                                              (col("Difference_Health") == max_diff_health_17) |
                                              (col("Diff_Freedom") == max_diff_freedom_17) |
                                              (col("Diff_Govt_Corruption") == max_diff_corruption_17) |
                                              (col ("Diff_Generosity") == max_diff_generosity_17)
                                              )\
    .select("Country")

display(df_max_change_17)
```

```
▸ (3) Spark Jobs
▸ df_max_change_17:  pyspark.sql.dataframe.DataFrame = [Country: string]
```

	Country
1	Bahrain
2	Kosovo
3	Myanmar
4	Rwanda
5	Burundi



CLOUD FRAMEWORKS EXAM (PART 2)

THANK YOU

NAME: ZAKRA CHACHAR
DATE: 11 FEBRUARY 2024