**ETL Project**

By Syeda Fasahath Naaz, Aneida Winston

**Introduction:**

The purpose of this project was to demonstrate a complete Extract, Transform and Loading (ETL) Process. This process was used to blend data from multiple sources, and loading into a data warehouse . Our team decided to amalgamate data pertaining to the World Happiness Report.The happiness scores and rankings use data from the Gallup World Poll. The scores are based on answers to the main life evaluation question asked in the poll. This question, known as the Cantril ladder, asks respondents to think of a ladder with the best possible life for them being a 10 and the worst possible life being a 0 and to rate their own current lives on that scale. The scores are from nationally representative samples for the years 2013-2016 and use the Gallup weights to make the estimates representative. The columns following the happiness score estimate the extent to which each of six factors – economic production, social support, life expectancy, freedom, absence of corruption, and generosity – contribute to making life evaluations higher in each country than they are in Dystopia, a hypothetical country that has values equal to the world’s lowest national averages for each of the six factors. They have no impact on the total score reported for each country, but they do explain why some countries rank higher than others.

**Extraction:**

We used 6 data files from the public platform Kaggle and Data World. All of our data was based from Gallup World Poll for the World Happiness Report from 2015 to 2020. The data file for 2020 was extracted from Data World and the remaining data files were extracted from Kaggle.

**Transformations:**

The first step of cleaning the data was figuring out which variables were not relevant.

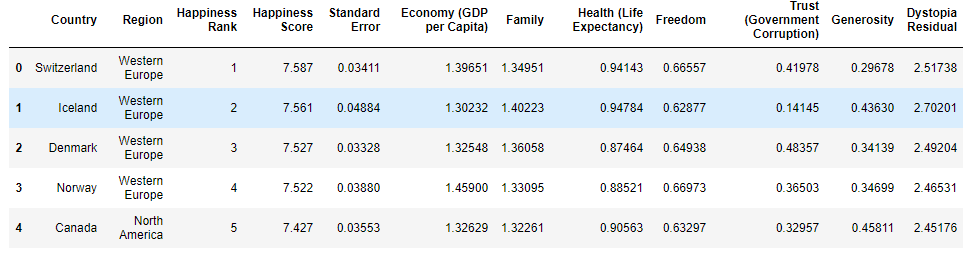


Figure1: 2015 file from kaggle

We dropped Standard Error and Dystopia Residual variables.

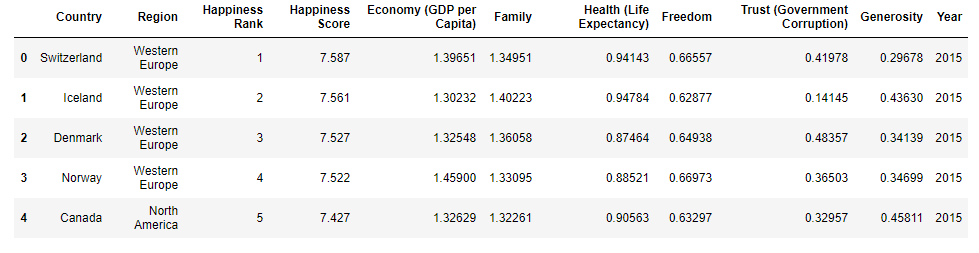


Figure 2: 2015 file after transformation

Similar transformations were done on the remaining files as well and transformed them in a way similar to the above figure.

The 2020 year file was from Data World as follows

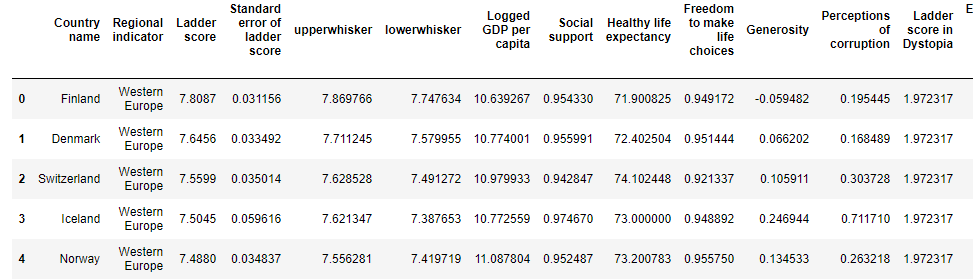


Figure 3: 2020 file from Data World

We dropped additional variables and renamed the columns. In addition the 2020 file did not include the Happiness Rank column which was required. We found a method in pandas Rank which computes numerical data rank (1 through n) along the axis. We used this method to calculate the Happiness Rank for 2020 file and named the column ‘Happiness Rank’.

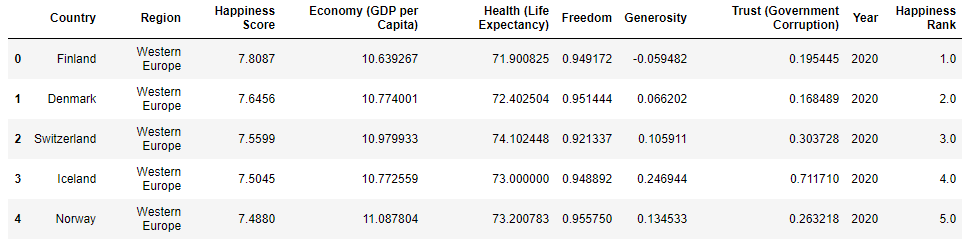


Figure 4: 2020 file after transformation

After cleaning the data files, we started merging the files into one file ‘Happiness\_report’. To start we added a new column ‘Year’ into each of the six files whose value was the year the data belongs to. Next the files were taken in a list and using a for loop we appended each file to the Happiness\_report file. Later we dropped the NaN values using the method ‘dropna’ in pandas.

The merged file is as follows:

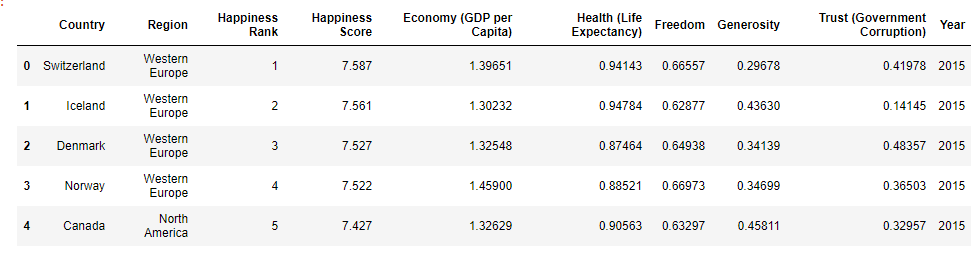


Figure5: Final merged data file ‘Happiness\_report’

**Load:**

The last step was to transfer our final data frame into a DataBase. We created DataBase and the respective table to match the columns from the final Pandas Dataframe using PostgresQL and then connected to the DataBase using SQLAlchemy and loaded the result.

References:

1. <https://www.kaggle.com/unsdsn/world-happiness>
2. https://data.world/makeovermonday/2020w19-world-happiness-report-2020