**ETL technical report**

**By : Hayley, John, Hatim, Tulfikha**

Background

The project is designed to create a database for the HR department at Youtube by extracting, transforming, and loading the world happiness index report from the last thirteen years from 2008 to 2020

The HR team wanted a stable database for the reporting dashboard to be able to pull data, demonstrate the changes of index through time and as well to track the results of it.

The interest in the data is mainly for promotional purposes but could be included with other global size datasets such as the UN poverty report in the future.

Extraction

We used 4 different datasets from Kaggle, a public platform.

The data in the 4 CSV files included the following economic indicators in measuring world happiness and the following information:

- Country codes for all countries

- Factors of world happiness index

- World happiness indication based on countries from year 1995 to 2019

- World happiness indicator for the year 2020

The sources for our dataset were:

* world-happiness-report.csv
* world-happiness-report-2021.csv
* population\_by\_country\_2020.csv
* wikipedia-iso-country-codes.cs

Transformation

Importing data

- Firstly, we have used pandas’ function in Jupyter notebook to load all 4 CSV files. We then reviewed the tables and transformed the tables into 4 data frames.

- We suspected that the data may be inconsistent and that there could potentially be columns with missing information, therefore we created a road map by using the function info() that displayed column information. A quick look at this text indicated that the data set was incomplete and contained null values. The road map also helped us to identify the type of columns as well.

- Another visualization we did was a heat map, this served as a quick indication as to which columns may need cleaning, filtering or dropping and a cross check to the road map we created as they essentially display the same information regarding the quality and completeness of the data sets chosen.

- Next, we had to decide if these null values should be dropped from the dataset, replaced by zero or another relevant value. In order to do so, we use the function pandas describe() to process the data to determine the desired value for null values which will replace the null values.

Now that we have a complete dataset,

We dropped duplicated columns and values, renamed columns for better readability and created columns as required in accordance with our structured ERD diagram.

We dropped some columns because we think that data is redundant.

- Negative affect

- Positive affect

We have also dropped a few values from the Country\_name columns due to the lack of information around the country and their population being less than 50,000 and is not in the list of countries that is approved and recognised by the United Nation.

LOAD

As part of normalization, a yearly calendar has been added as a reference table.

* Country used as Primary Key between data tables
* The use case dictates that this DB will expand

The last step was to transfer our final output into a Database. We created a database and respective table to match the columns from the final Panda's Data Frame using Postgres database using PG admin to store our original clean data sets. We reconnected to the database and generated additional tables for the data frames.

INPUT AND OUTPUT TABLES

-

| **Input tables** | **Output Tables** |
| --- | --- |
| **Table Name: 0 wikipedia-iso-country-codes.csv**  **--------------------------------------------------**    **Table Info:**  **-------------------------**  **<class 'pandas.core.frame.DataFrame'>**  **RangeIndex: 246 entries, 0 to 245**  **Data columns (total 5 columns):**  **# Column Non-Null Count Dtype**  **--- ------ -------------- -----**  **0 English short name lower case 246 non-null object**  **1 Alpha-2 code 245 non-null object**  **2 Alpha-3 code 246 non-null object**  **3 Numeric code 246 non-null int64**  **4 ISO 3166-2 246 non-null object**  **dtypes: int64(1), object(4)**  **memory usage: 9.7+ KB** | **Table Name: 0 wikipedia-iso-country-codes.csv**  **--------------------------------------------------**    **Table Info:**  **-------------------------**  **<class 'pandas.core.frame.DataFrame'>**  **Int64Index: 246 entries, 0 to 245**  **Data columns (total 7 columns):**  **# Column Non-Null Count Dtype**  **--- ------ -------------- -----**  **0 country\_id 246 non-null int32**  **1 english\_name 246 non-null object**  **2 alpha\_2\_code 246 non-null object**  **3 alpha\_3\_code 246 non-null object**  **4 numeric\_code 246 non-null int64**  **5 iso\_3166\_2 246 non-null object**  **6 land\_area 246 non-null float64**  **dtypes: float64(1), int32(1), int64(1), object(4)**  **memory usage: 14.4+ KB** |
| **Table Name: 1 population\_by\_country\_2020.csv**  **--------------------------------------------------**    **Table Info:**  **-------------------------**  **<class 'pandas.core.frame.DataFrame'>**  **RangeIndex: 235 entries, 0 to 234**  **Data columns (total 11 columns):**  **# Column Non-Null Count Dtype**  **--- ------ -------------- -----**  **0 Country (or dependency) 235 non-null object**  **1 Population (2020) 235 non-null int64**  **2 Yearly Change 235 non-null object**  **3 Net Change 235 non-null int64**  **4 Density (P/Km²) 235 non-null int64**  **5 Land Area (Km²) 235 non-null int64**  **6 Migrants (net) 201 non-null float64**  **7 Fert. Rate 235 non-null object**  **8 Med. Age 235 non-null object**  **9 Urban Pop % 235 non-null object**  **10 World Share 235 non-null object**  **dtypes: float64(1), int64(4), object(6)**  **memory usage: 20.3+ KB** | **Table Name: 1 population\_by\_country\_2020.csv**  **--------------------------------------------------**    **Table Info:**  **-------------------------**  **<class 'pandas.core.frame.DataFrame'>**  **Int64Index: 195 entries, 0 to 233**  **Data columns (total 8 columns):**  **# Column Non-Null Count Dtype**  **--- ------ -------------- -----**  **0 Yearly Change 195 non-null object**  **1 net\_change 195 non-null int64**  **2 migrants 195 non-null float64**  **3 fert\_Rate 195 non-null object**  **4 med\_age 195 non-null object**  **5 urban\_pop 195 non-null object**  **6 country\_id 195 non-null float64**  **7 yearly\_id 195 non-null int64**  **dtypes: float64(2), int64(2), object(4)**  **memory usage: 13.7+ KB** |
| **Table Name: 2 world-happiness-report.csv**  **--------------------------------------------------**    **Table Info:**  **-------------------------**  **<class 'pandas.core.frame.DataFrame'>**  **RangeIndex: 1949 entries, 0 to 1948**  **Data columns (total 11 columns):**  **# Column Non-Null Count Dtype**  **--- ------ -------------- -----**  **0 Country name 1949 non-null object**  **1 year 1949 non-null int64**  **2 Life Ladder 1949 non-null float64**  **3 Log GDP per capita 1913 non-null float64**  **4 Social support 1936 non-null float64**  **5 Healthy life expectancy at birth 1894 non-null float64**  **6 Freedom to make life choices 1917 non-null float64**  **7 Generosity 1860 non-null float64**  **8 Perceptions of corruption 1839 non-null float64**  **9 Positive affect 1927 non-null float64**  **10 Negative affect 1933 non-null float64**  **dtypes: float64(9), int64(1), object(1)** | **Table Name: 2 world-happiness-report.csv**  **--------------------------------------------------**    **Table Info:**  **-------------------------**  **<class 'pandas.core.frame.DataFrame'>**  **Int64Index: 1834 entries, 0 to 148**  **Data columns (total 9 columns):**  **# Column Non-Null Count Dtype**  **--- ------ -------------- -----**  **0 yearly\_id 1834 non-null int64**  **1 life\_ladder 1834 non-null float64**  **2 log\_gdp\_per\_capita 1834 non-null float64**  **3 social\_support 1834 non-null float64**  **4 healthy\_life\_expectancy\_at\_birth 1834 non-null float64**  **5 freedom\_to\_make\_life\_choices 1834 non-null float64**  **6 generosity 1834 non-null float64**  **7 perceptions\_of\_corruption 1834 non-null float64**  **8 country\_id 1834 non-null float64**  **dtypes: float64(8), int64(1)**  **memory usage: 143.3 KB** |
| **Table Name: 3 world-happiness-report-2021.csv**  **--------------------------------------------------**    **Table Info:**  **-------------------------**  **<class 'pandas.core.frame. DataFrame'>**  **RangeIndex: 149 entries, 0 to 148**  **Data columns (total 20 columns):**  **# Column Non-Null Count Dtype**  **--- ------ -------------- -----**  **0 Country name 149 non-null object**  **1 Regional indicator 149 non-null object**  **2 Ladder score 149 non-null float64**  **3 Standard error of ladder score 149 non-null float64**  **4 upperwhisker 149 non-null float64**  **5 lowerwhisker 149 non-null float64**  **6 Logged GDP per capita 149 non-null float64**  **7 Social support 149 non-null float64**  **8 Healthy life expectancy 149 non-null float64**  **9 Freedom to make life choices 149 non-null float64**  **10 Generosity 149 non-null float64**  **11 Perceptions of corruption 149 non-null float64**  **12 Ladder score in Dystopia 149 non-null float64**  **13 Explained by: Log GDP per capita 149 non-null float64**  **14 Explained by: Social support 149 non-null float64**  **15 Explained by: Healthy life expectancy 149 non-null float64**  **16 Explained by: Freedom to make life choices 149 non-null float64**  **17 Explained by: Generosity 149 non-null float64**  **18 Explained by: Perceptions of corruption 149 non-null float64**  **19 Dystopia + residual 149 non-null float64**  **dtypes: float64(18), object(2)**  **memory usage: 23.4+ KB** | **Have been added to other table** |
|  | **created**  **Table Name: 3 yearly**  **--------------------------------------------------**    **Table Info:**  **-------------------------**  **<class 'pandas.core.frame.DataFrame'>**  **RangeIndex: 99 entries, 0 to 98**  **Data columns (total 2 columns):**  **# Column Non-Null Count Dtype**  **--- ------ -------------- -----**  **0 yearly\_id 99 non-null object**  **1 past\_year 99 non-null object**  **dtypes: object(2)**  **memory usage: 1.7+ KB** |