**ETL Project – Happiness & Life Expectancy**

Group members: Jasjeet Kaur, Saleha Khatun, Erika Hoshino, Phimnarin Aksupak

**Project Outline:**

Our group aimed to explore the happiness rankings of countries across the world, focusing on specific factors such as the freedom to make life choices, social support and GDP per capita. In addition, we investigated the life expectancy rankings of countries worldwide. Further analysis can be performed to determine any correlations between happiness factors and countries with higher life expectancies.

**EXTRACTION:**

We utilised 2 datasets from the public platform Kaggle; the ‘World Happiness Report’ and the ‘Human Life Expectancy Around the World’ report. The World Happiness Report dataset listed information for 158 countries between the years of 2015 to 2021, and the Human Life Expectancy dataset comprised of a collection of 186 countries ranging from 1990 to 2019.

We focused on extracting the data from 2019 from both datasets for consistency.

The sources are listed below:

* World Happiness Report 2015-2021 (Kaggle)
  + <https://www.kaggle.com/mathurinache/world-happiness-report-20152021?select=2019.csv>
* Human Life Expectancy Around the World (Kaggle)
  + <https://www.kaggle.com/deepcontractor/human-life-expectancy-around-the-world>

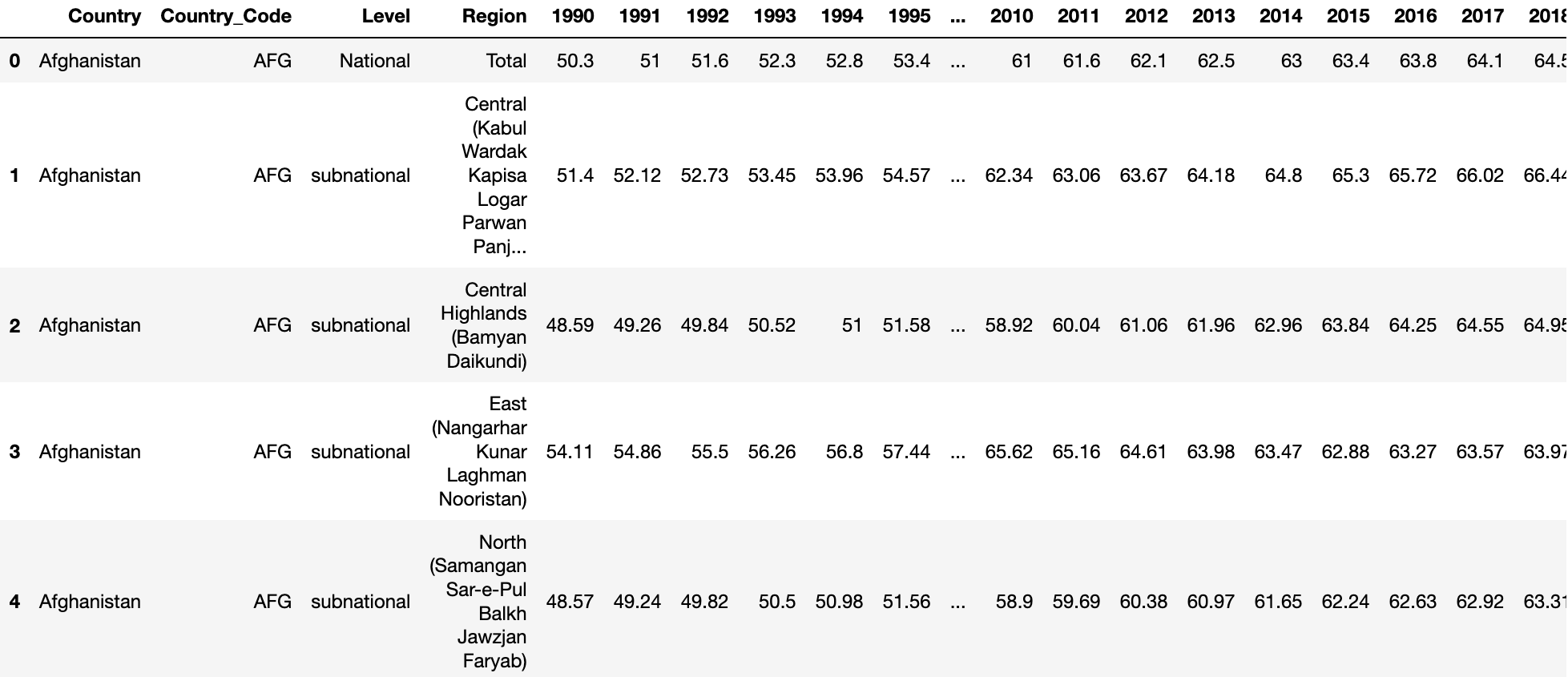
The factors of interest from the World Happiness Report are as follows:

* Freedom to make life choices
* Social support
* GDP per capita
* Happiness rank
* Happiness score

1. Following the importing of dependencies, firstly we read the csv of the World Happiness and Human Life Expectancy dataset into a pandas data frame using Jupyter notebook (fig.1, 2)



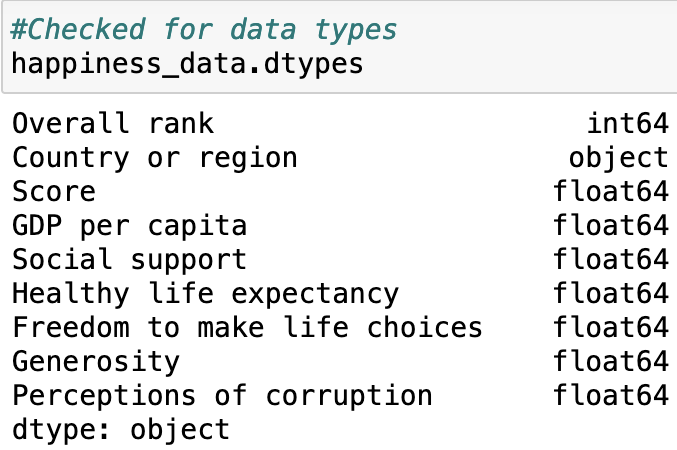
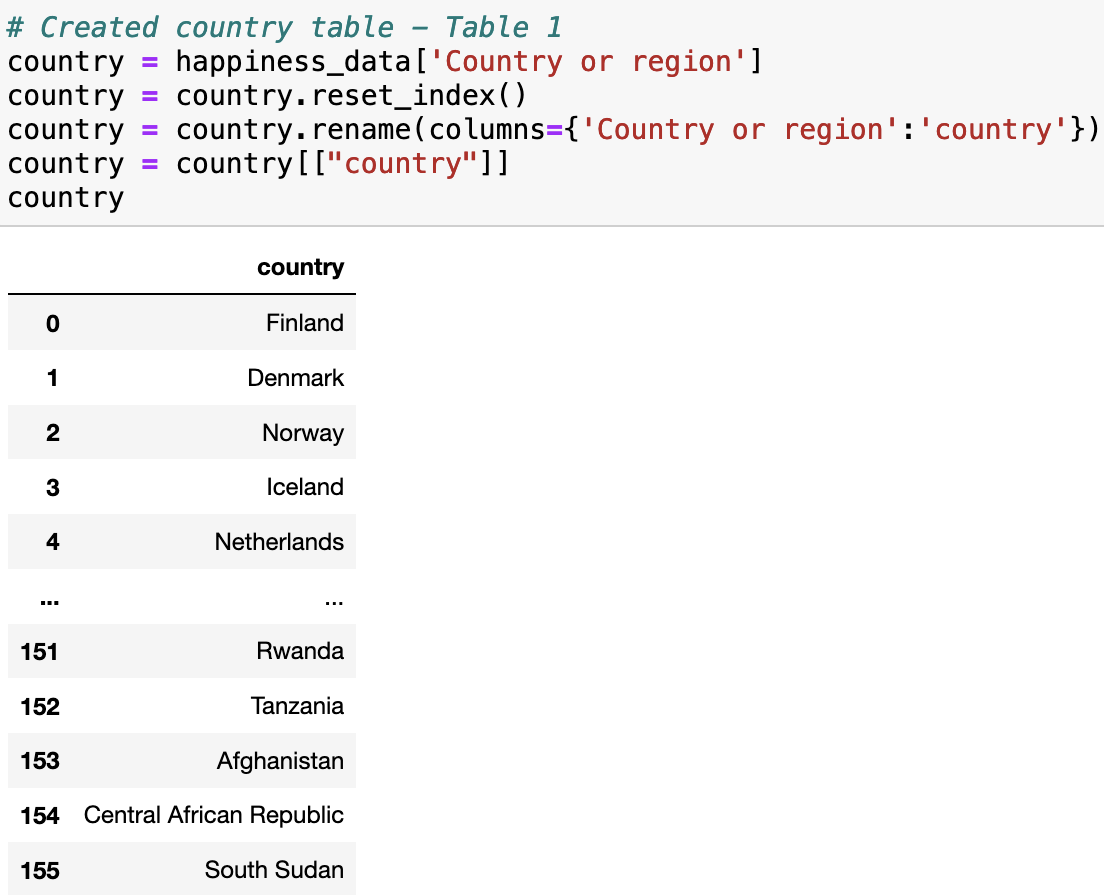
**Figure 1.** Original uncleaned Happiness Ranking dataset



**Figure 2.** Original uncleaned Human Life Expectancy dataset

**TRANSFORM:**

1. We checked for the data types contained within the dataset (fig.3) and created a ‘Country’ dataframe, renaming the ‘Country or region’ column (fig.4)

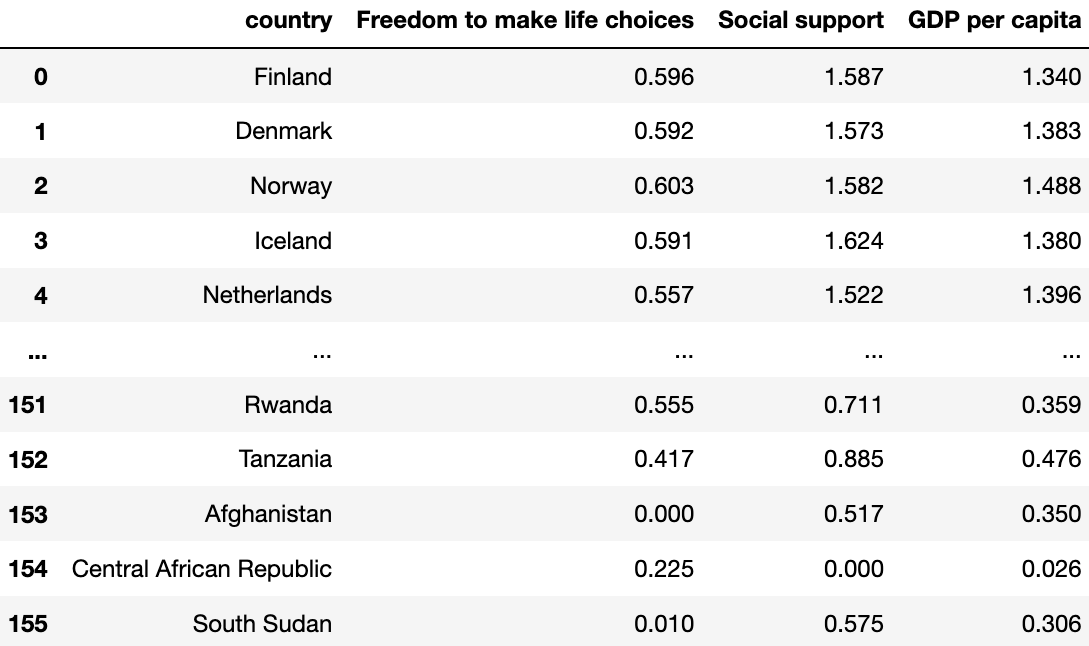
**Figure 3.** List of data types **Figure 4.** Renamed ‘Country’ table

1. We then started the dataset cleaning process by creating a separate data frame, selecting the Overall rank & Score and renaming the columns accordingly (fig.5)



**Figure 5.** Filtered dataset & renamed columns for Happiness rank/score

1. We created another separate data frame, selecting the factors of interest and renaming the columns for better comprehension (fig.6)



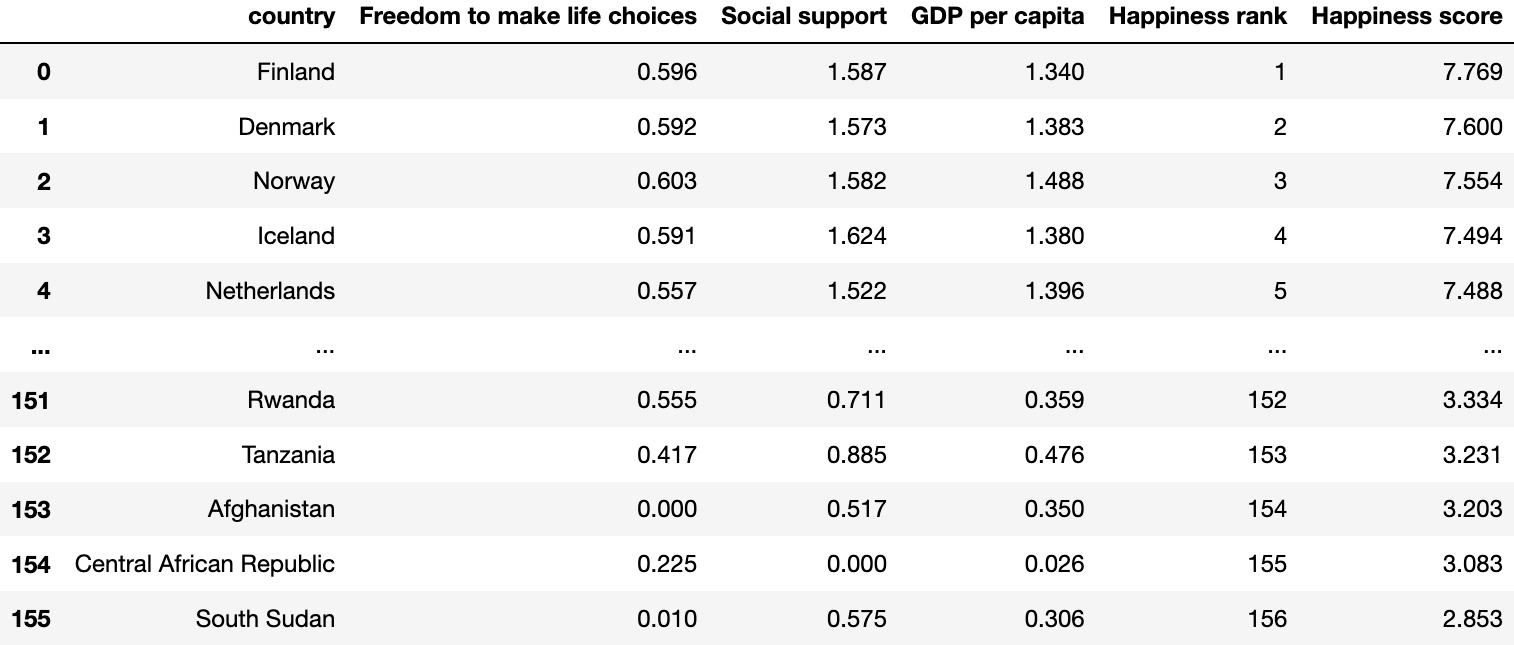
**Figure 6.** Filtered dataset & renamed columns for Happiness factors of interest

1. Additionally, we filtered the ‘Life Expectancy’ dataset for 2019 and selected the life expectancy data by each country’s region total (fig.7)

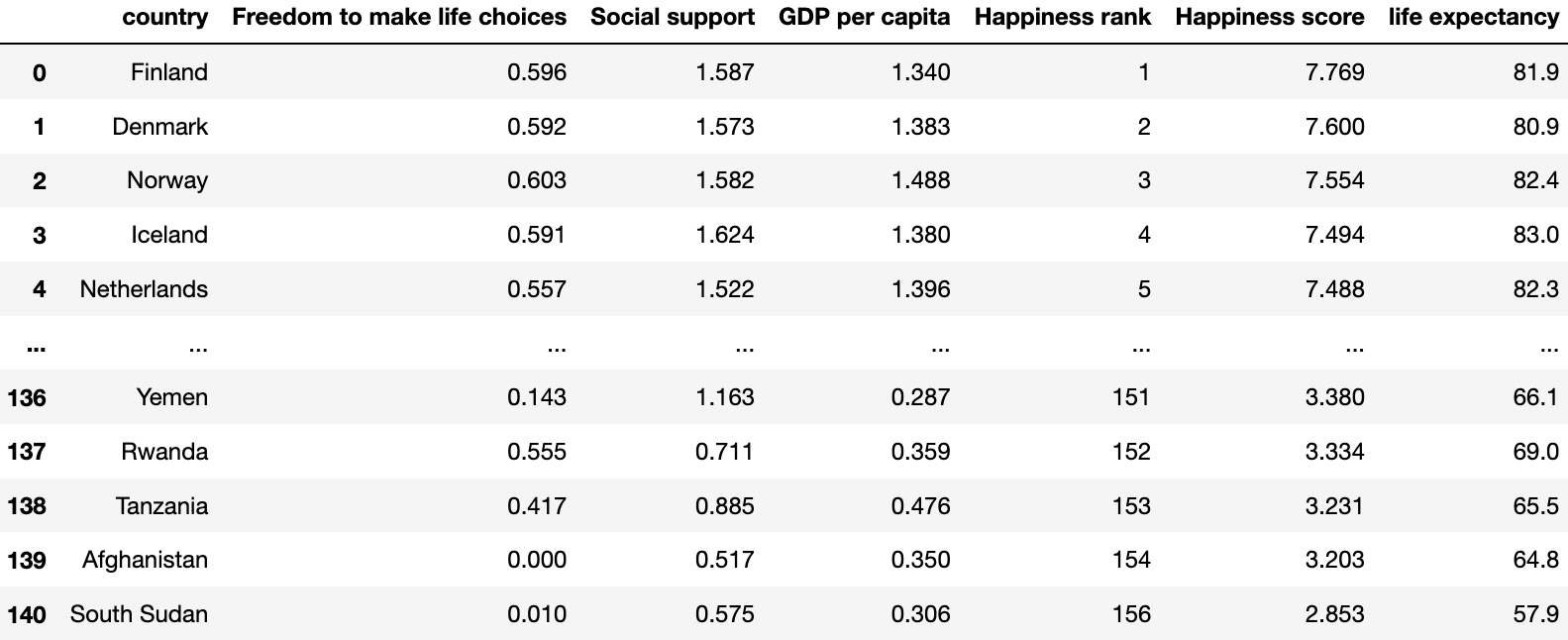
 

**Figure 7.** Dataset filtered for 2019 & country by region total

1. We combined the Happiness Factor & Happiness Score data frames (fig.8) and then with the Human Life Expectancy data frame (fig.9)



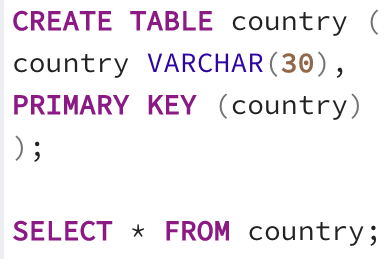
**Figure 8.** Combined Happiness Factor & Score data frame



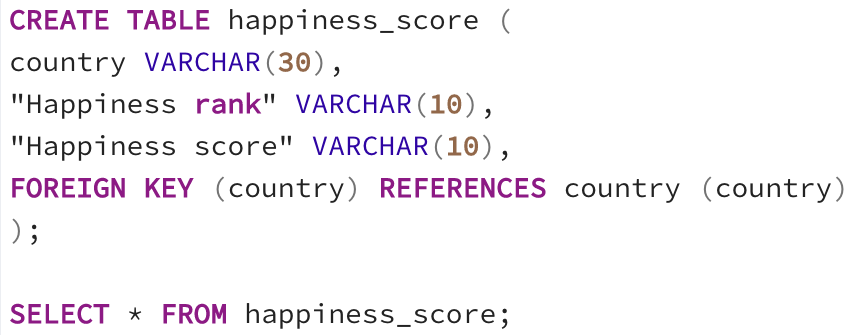
**Figure 9.** Combined Happiness & Life Expectancy data frame

**LOAD:**

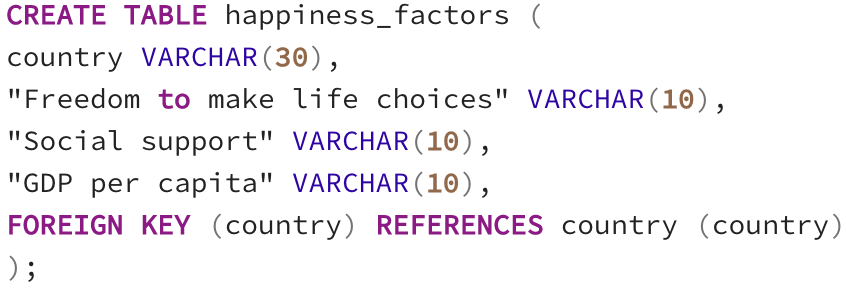
1. The final process involved loading the Happiness & Life Expectancy data frame into a database. This was achieved by establishing a connection with the local SQL database, creating a database using PgAdmin 4 and generating tables with columns matching the respective pandas data frames (fig.10, 11, 12, 13)

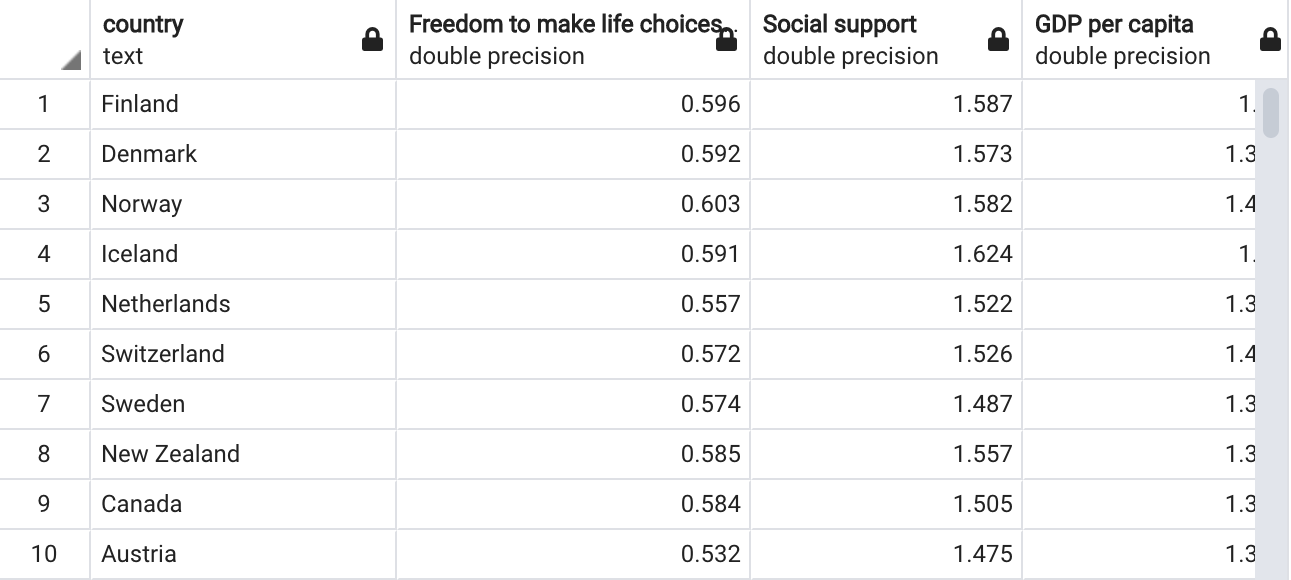
 

**Figure 10.** Country SQL database

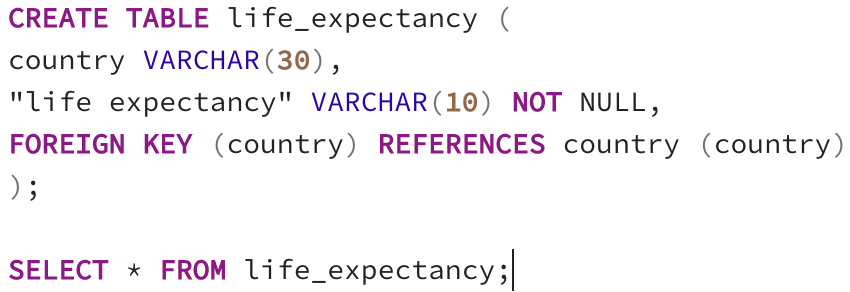
 

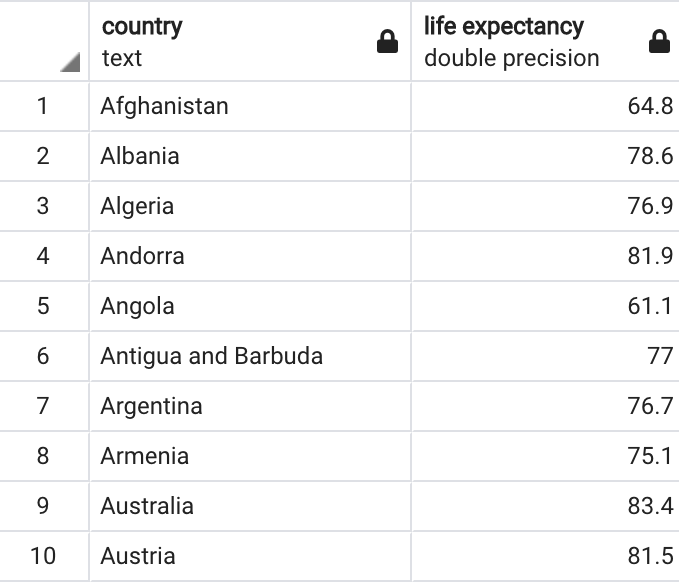
**Figure 11.** Happiness Score SQL database





**Figure 12.** Happiness Factors SQL database

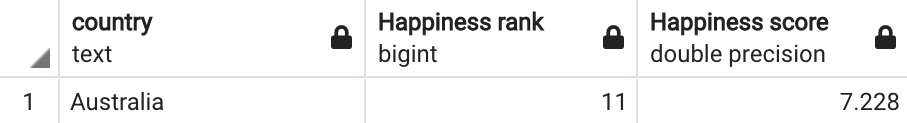




**Figure 13.**  Life Expectancy SQL database

1. Using this database, we are able to query for specific countries or other desired criteria (fig.11)





**Figure 11.** Example SQL query that returns the happiness rank and score for specific country variable, Australia

**SUMMARY:**

Our group selected two datasets to identify the specific happiness factors (freedom to make life choices, social support and GDP per capita) and life expectancy rankings across various countries worldwide. The final SQL database output will enable us to determine the countries with the highest and lowest happiness rankings based on our selected factors.

Moreover, further analysis can be performed to determine whether there is a correlation between country happiness ranking and life expectancy. This can be achieved by generating scatter plots and running linear regressions using the merged data frame (fig.9) from the transformation process.