Country’s quality of life vs (perception of corruption, GDP per capita, etc.)

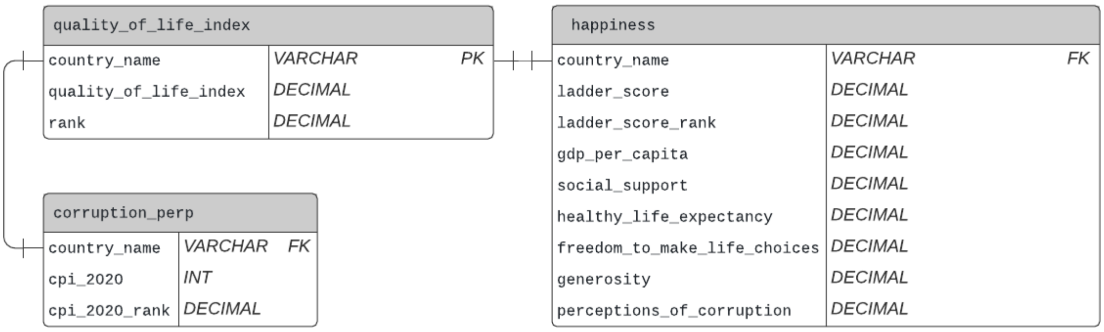
This project was completed by Hayley Lim, Nicholas Chua, Sanuli Lohara and Tamer Abdelaal

In this project, we are loading a database with data about quality of life for more than 100 countries, along with data about some factors that can be affecting it such as Perception of Corruption Index, GDP per capita, Healthy Life Expectancy, Social Support and Freedom to make life choices.

To make this study, we used 3 csv files:

1. country_quality_of_life_score_data
2. corruption_perception_index_dataset
3. corruption_perception_index_dataset

We used (country_quality_of_life_score_data) as a master file to get the list of countries used in the study, then we merged the other datasets with it to do further cleaning before loading to Postgresql database. This relation between the data sets can be seen in the diagram below



quality_of_life_index		
country_name	VARCHAR	PK
quality_of_life_index	DECIMAL	
rank	DECIMAL	

corruption_perp		
country_name	VARCHAR	FK
cpi_2020	INT	
cpi_2020_rank	DECIMAL	

happiness		
country_name	VARCHAR	FK
ladder_score	DECIMAL	
ladder_score_rank	DECIMAL	
gdp_per_capita	DECIMAL	
social_support	DECIMAL	
healthy_life_expectancy	DECIMAL	
freedom_to_make_life_choices	DECIMAL	
generosity	DECIMAL	
perceptions_of_corruption	DECIMAL	

The ETL Process:

Extract:

The source files were extracted from the following websites, then kept in the Resources folder:

1. country_quality_of_life_score_data.csv from this Kaggle page
2. corruption_perception_index_dataset.csv as “CPI2020_SignificantChanges_210125.xlsx” from transparency.org
3. happiness_index_data.csv from this Kaggle page

Transform:

Country quality of life:

1. Using Pandas, we imported the csv file as a data frame, and renamed the columns
2. Removed the column that shows country name as per its native language
3. Created another column showing each country’s rank based on its quality of life index

Corruption Perception Index:

1. The original file was in Excel, but we had many errors while trying to import it in Pandas, so we had to convert it to a simpler csv file before importing.
2. As the data in (Country quality of life) was for the year 2020, we kept only the columns of (country, CPI 2020, CPI rank 2020)
3. As some of the country names were different than (Country quality of life) file, we renamed them

Happiness Index:

1. Using Pandas, we imported the csv file as a data frame
2. Of all the columns, we decided to keep only (Country name, Ladder score, Logged GDP per capita, social support, Healthy life expectancy, Freedom to make life choices, Generosity, Perceptions of corruption)
3. As some of the country names were different than (Country quality of life) file, we renamed them

Further cleaning for all files:

1. Merged (Corruption perception index) file with (Country quality of life) file to remove all the countries that do not exist in the quality of life file
2. Added column "cpi_2020_rank" to re-rank the corruption perception index for the countries that remained after the merge
3. For (Happiness index) file, we did step 1 above, then added a ranking based on the "ladder score"

Load:

1. We created an Entity Relational Diagram for the 3 tables on (Lucidchart.com)
2. From the diagram we exported the Postgresql code to use for loading the tables to Postgresql
3. We connected to the local database, checked for successful connection to the database and confirmed that the 3 tables (quality_of_life_index, corruption_perp, happiness) have been created.
4. We then used pandas to load csv converted dataframe into PostgreSQL relational database.
5. The database can be used for future analysis or business use.