**Project Report**

ETL Project – Best Quality Living Countries

Project was completed by Mohini Yadav, Phuong Han, Frederick Tan

**Objectives**

This project will deliver a review about the quality of living in each country. It might help users to determine and develop interest in which countries they would like to live in based on how the demographic factors are responsible

Our data source is coming from <https://www.kaggle.com/dumbgeek/countries-dataset-2020>

We would like to acknowledge other websites who have supported the datasets above

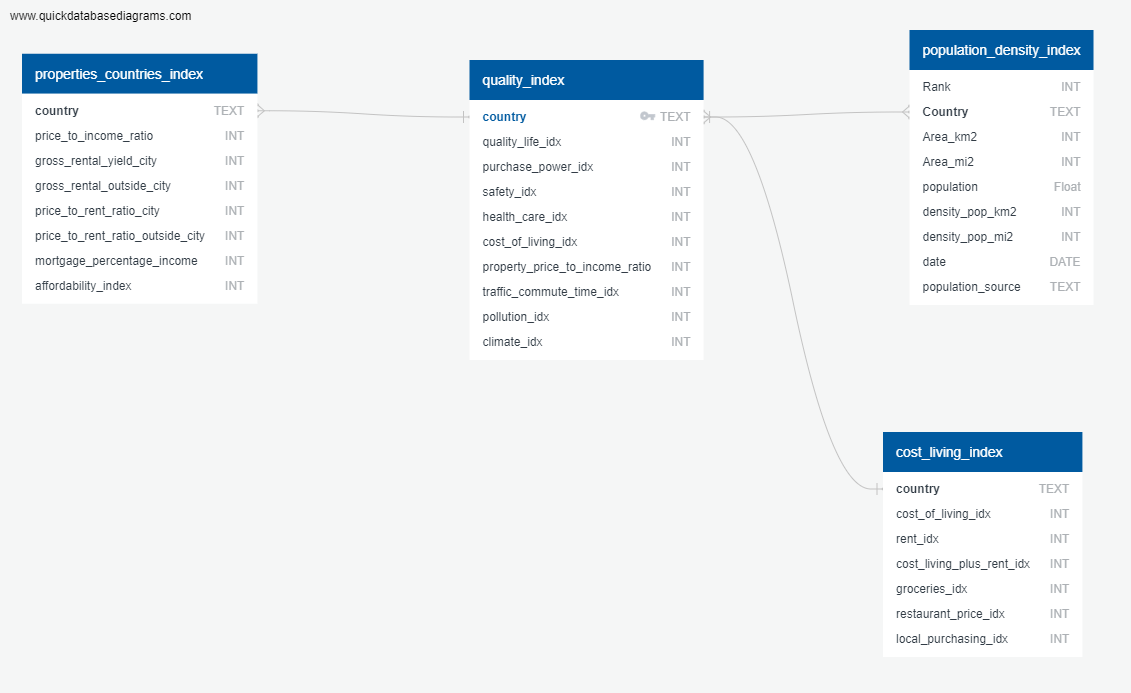
Acknowledgements Wikipedia : <https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population_density> Wikipedia : <https://en.wikipedia.org/wiki/List_of_countries_by_age_structure> Numbeo : [https://www.numbeo.com](https://www.numbeo.com/)

This project takes 4 datasets and eventually creates a structured database with 4 tables in PostgresSQL.

These are the steps that were executed

1. **SCHEMA DESIGN**

Design database schema using the ERD as follows:

[](https://user-images.githubusercontent.com/83207549/128088566-6634d1af-4bbf-4b9e-b943-e8b4a19fbe3b.png)

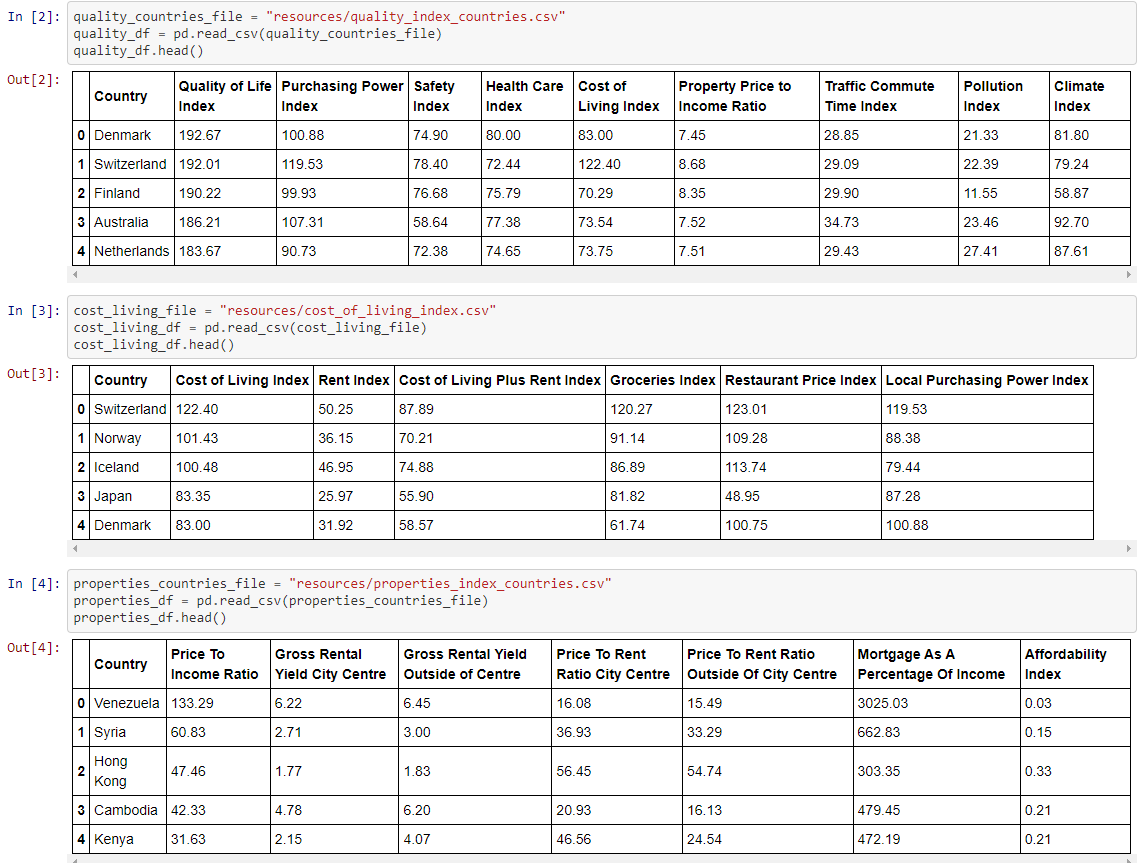
The ERD develop to predefine the database in Postgres SQL in order to run the ETL process

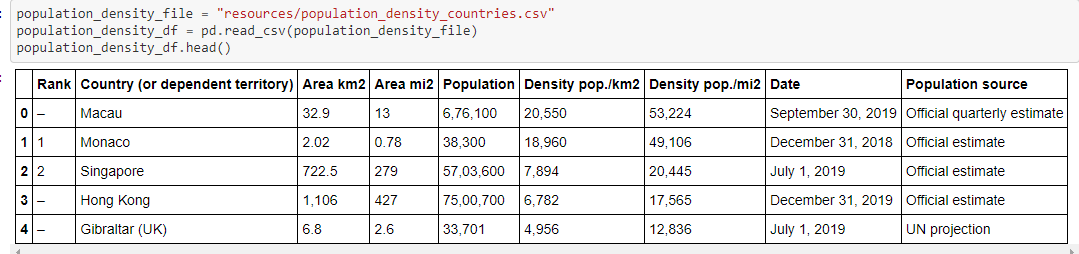
1. **EXTRACT**

Download the 4 csv files from Kaggle.com. These csv files can be found in the resources folder

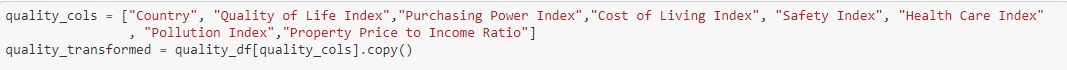
1. Cost\_of\_living\_index.csv
2. Population\_density\_countries.csv
3. Properties\_index.csv
4. Quality\_index\_countries.csv
5. **TRANSFORM**

* Load all those 4 csv files into Pandas as by using the read\_csv function and load them into a dataFrame



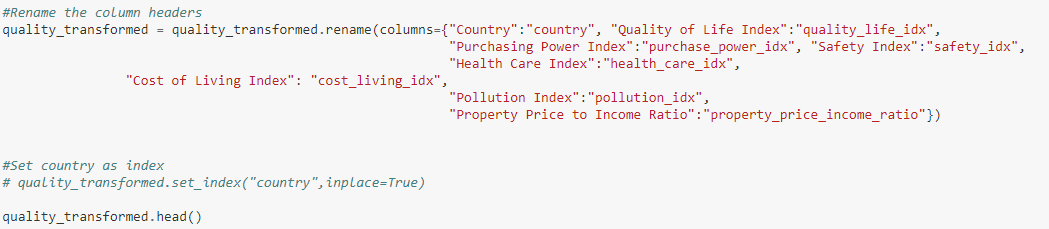


* Drop unwanted columns in each datasets by selecting only the necessary columns for information needed





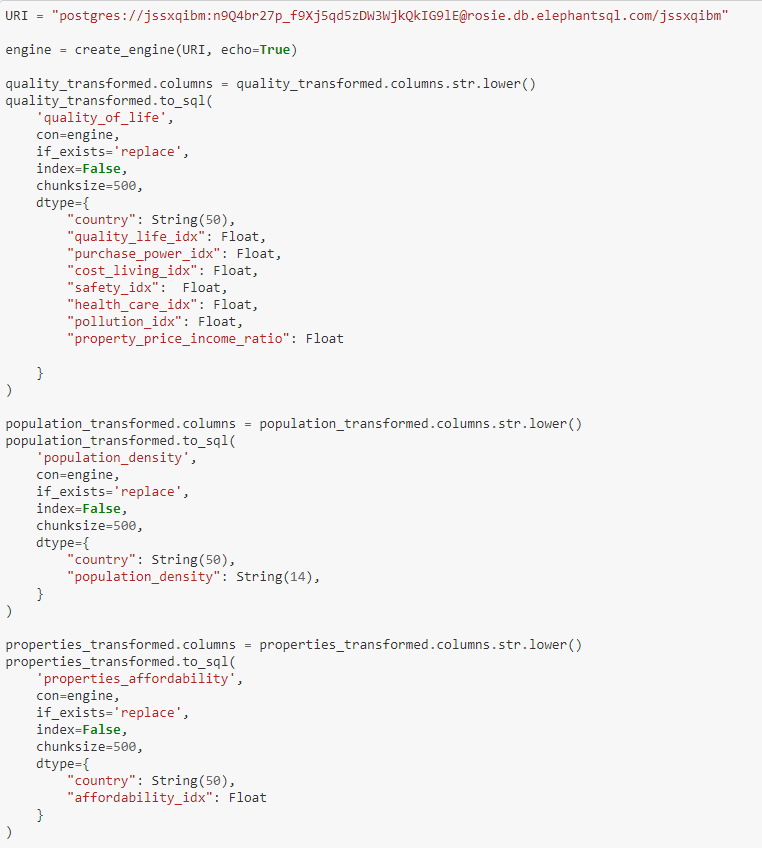
* Rename the column headers in each datasets by using the rename function to shorten the title header and removing any spaces

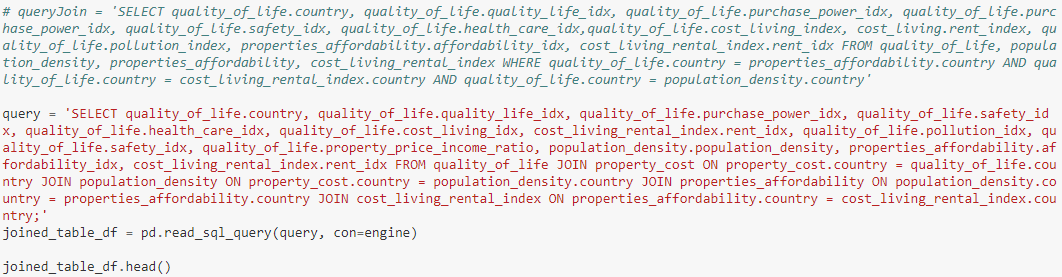


* Merge dataFrame based on country

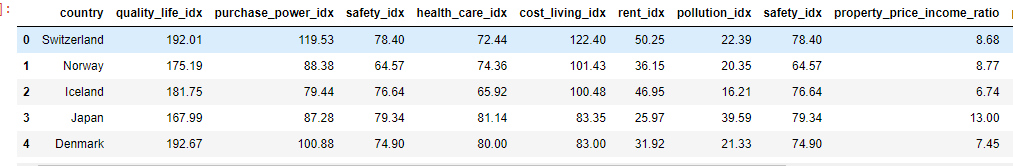
1. **LOAD**

* Create countries\_db in PostgresSQL with separate tables
* Create the join table queries using SQLAlchemy and import the data into tables directly from Pandas using SQL Alchemy Engine





**Screenshot of the final database**



For more detail information please look at the link below

https://github.com/fredtan30/Project-2-ETL-Countries-Cost-Living/blob/main/Transform.ipynb