**ETL Project**

*Jaspreet Bhamra, Chao Wang*

**Introduction**

This is a ETL process guidance document for data analysis on GDP per capita, Life expectancy and International tourist departure numbers from 2000 to 2018 for Australia, United States, India and China. Following guidelines explain the process of downloading data, transforming it into desired structure and then loading it into Database.

**Extraction**

* Download GDP per capita and life expectation as .CSV from Kaggle @ <https://www.kaggle.com/mohamedbahgat/happiness-score-analysis>.
* Download International tourist departure data from Worldbank @ <https://databank.worldbank.org/source/world-development-indicators>.

Select above mentioned countries, series as international travel departure and time from 2000 – 2018.

**Transformation**

Jupyter notebook under Python 3.6.10 is used to import and transform the data. Dependencies imported are as follow: pandas, numpy, sqlalchemy.

* Import pandas as pd
* Import numpy as np
* From sqlalchemy import create\_engine

Read .CSV files in jupyter notebook using pandas pd.read\_csv and show data frames. Files would look like as following.

* GDP per capita • Life expectation

Table

Description automatically generatedTable

Description automatically generated

Figure 1 (left) top 5 rows of GDP per capita data, (right) top 5 rows of Life expectation data.

* International tourist departure

Graphical user interface, application

Description automatically generated

Figure 2 international tourist departure number for Australia, China, India, and the USA

Using .loc in GDP per capita data frame’s country column get data for above mentioned 4 countries.

Set country column as index by using .set\_index and have inplace value to True. Show data frame.

Use same method as above for Life expectation data frame.

Remove Unnamed column from the data frame.

Restructured and clean data frame for GDP per capita and Life expectancy would look like following.

Table

Description automatically generated Table

Description automatically generated

Figure 3: Cleaned and reindexed data frames for GDP per capita and life expectancy.

From travel data frame drop extra columns named Country Code, Series Name and Series Code by using **.drop code.** To transpose the data from horizontal layeout to vertical layout, .melt function in pandas is used.

Use . slice to change the string format data in Departure year column (e.g. 2000 [YR2000]) to integer format (e.g. 2000) by slicing the string from “0”th element and keep first “4”.

Use .sort function to sort the values in the data frame by columns “Country Name” and “Departure Year”.

Use .rename function to change column names of international travel departure data frame to match column name of GDP per capita data frame and life expectancy data frame. ‘Country Name’ to ‘country’ and ‘Departure Year’ to ‘year’.

Change year column in international data frame from string to integer using .apply(int) before merge with other data frames.

This is the transformed Dataframe of international travel departure data look like:

Table

Description automatically generated

Figure 4: Cleaned datasets for international tourist departure

.merge is used to merge GDP per capita data frame with Life expectancy data frames on “Country” and “Year” columns.

Similarly, merge International data frame on “Country” and “Year” columns to get one final combined data frame, as shown in Figure 5.



Figure 5 Final combined dataset.

**Load**

Finally, we would transfer our output into a SQL Database. We create a database diagram using Quick DBD and created tables in PostgreSQL respectively.

Config containing username and password for pgAdmin is imported.

Then connect to the database in Panda’s Data Frame using sqlalchemy. Use create\_engine to create database connection.Load DataFrames into SQL database using to\_sql function with condition con=engine, if\_exists='append', index=True.

**Summary**

The final output is a SQL database containing country, year, life\_expectancy, gdp\_per\_capita, and No. of travellers data. This database can be used to analyse the required task for Australia, United States, India and China.