Maozhu tang, James le, Steven Bouios  
ETL, Assignment 2, Group 2

1 August 2022

# Overview

## Project Background and Description

|  |  |
| --- | --- |
|  | Monash University Bootcamp, Data Analytics, Assignment 2, Group 2. Team members: Maozhu Tang, James Le, Steven Bouios.  Purpose of project was to identify multiple datasets, that could be amalgamated to provide a richer dataset, ready to be analyzed. The multiple datasets are to be extracted (using either csv, JSON, pgadmin4. pandas or web scraping), transformed via MySQL, pandas, pgadmin4) and loaded into a final production database (relational or non-relational).  Finally, to document the whole process and load into GitHub with the raw, transformed, and create the final database.  Topic chosen: **Australian Tourism** |

## Source of data

|  |  |
| --- | --- |
|  | List sources of data that was extracted from. |

1. Australian inbound international tourist survey data:

* International survey data, from Tourism Research Australia (TRA), formed under the Australian Trade and Investment Commission. TRA provides statistics and research to assist the government, tourism industry and Australian businesses.
  + International Visitor Survey results, March 2022:
    - https://www.tra.gov.au/data-and-research/reports/international-visitor-survey-results/international-visitor-survey-results

1. National travel survey data:
   * National survey data, March 2022:
     + https://www.tra.gov.au/data-and-research/reports/national-visitor-survey-results/national-visitor-survey-results
2. Australian population data, birth by country:

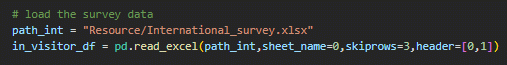
* Australia’s Population by Country of Birth, 2021, as released by Australian Bureau of Statics:
  + - <https://www.abs.gov.au/statistics/people/population/australias-population-country-birth/2021>

## Extraction process

|  |  |
| --- | --- |
|  | How the original data sources were extracted. |

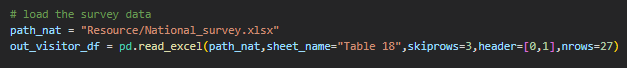
1. Australian inbound international tourist survey data:

* Data was obtained in excel format (xlsx).
* Accessed from local drive via Python Pandas modules inbuilt excel reading.
* Example code:



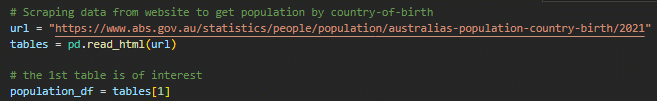
1. National travel survey data:

* Data was obtained in excel format (xlsx).
* Accessed from the local drive via Python Pandas module inbuilt excel reading.
* Example code:



1. Australian population data, birth by country:

* Scaped from the ABS website (https://www.abs.gov.au/statistics/people/population/australias-population-country-birth/2021).
* Scraped using Python, Pandas module inbuilt scaping for tabular data from websites.
* Example code:



## Transformation process

|  |  |
| --- | --- |
|  | Describe how you plan to implement the project. For example, will all parts of the project be rolled out at once or will it be incremental? What will be included in each release? |

1. Australian inbound international tourist survey data:

* As data was in summary format (not raw), so needed to:
  + select which sheet within the excel file was required,
  + how many rows to skip,
  + specify how many rows the data spans within the sheet, so as not to pick up unrelated summary data below.
  + Subsequently data was retrieved into a pandas dataframe.
  + Header included 2 rows, which pandas couldn’t handle, so the two rows were joined to create one row as a header.
  + Columns names were renamed.
  + Two columns were kept for final dataframe.
  + Values in dataframe were divided by 1000, to be more legible.

1. National travel survey data:

* Once again data was in a summary format (not raw), so needed to:
  + Perform the step mentioned above on this data as well.
  + In addition, removing rows where there is null/np values.
  + Converting values to float.

1. Australian population data, birth by country:

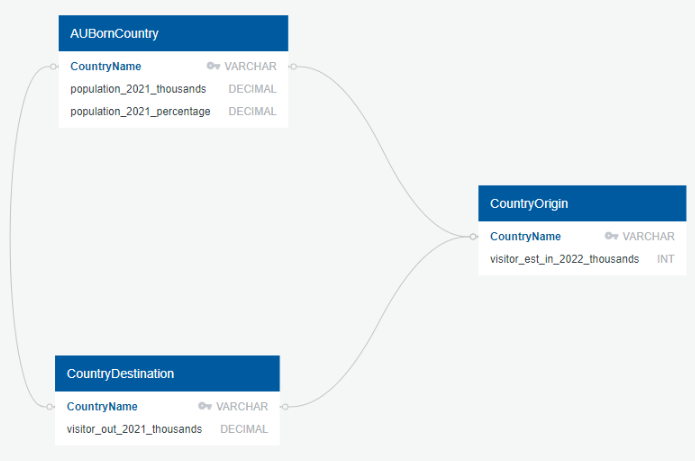
* Tabular data was scraped from website, so needed to:
  + Select 2nd table which had the relevant data.
  + Header included 2 rows, which pandas couldn’t handle, so the two rows were joined to create one row as a header.
  + Renamed columns titles.

## Load: final database

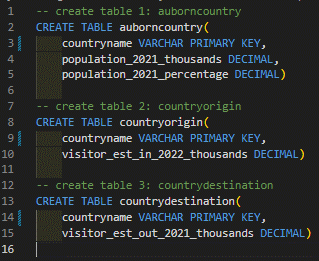
|  |  |
| --- | --- |
|  | Why the final database, tables, collections, were chosen. |

To ensure a structured data with specified primary keys and the ability to later join all three tables, to create a richer amalgamated dataset, SQL was chosen to load data into a database.

An entity relationship diagram (ERD) from quickdatabasediagrams.com, was created to data model each table and how the three tables would be linked, taking into consideration the primary keys. Here:



Here is the SQL schema representation:



The three data frames were then loaded into individual SQL tables, utilising the SQLAlchemy module within python to commit the data frames to SQL. Example code here:



## GitHub repository

|  |  |
| --- | --- |
|  | Final report, raw data, and code loaded onto GitHub. |

Link to GitHub: <https://github.com/MZt92-ui/Project-2---ETL.git>