Final ETL Report

Extraction

The data extracted came from four separate csv files, one from Kaggle (<https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset/data#covid_19_data.csv>) and three others from the World Bank’s data files. (<https://data.worldbank.org/indicator/SP.POP.0014.TO.ZS?view=chart>, <https://data.worldbank.org/indicator/SP.POP.1564.TO.ZS?view=chart>, <https://data.worldbank.org/indicator/SP.POP.65UP.TO.ZS?view=chart>)

Transform

The data was transformed in such a way as to compare the age demographics (percent of population under 15, between 15 and 64, and 65 and over) of the 10 countries with the largest number of confirmed COVID-19 cases. This required several types of transformations to reach the final dataset.

1. Cleaning: several columns and rows in all datasets were dropped, indexes were reset and reformatted, and columns were renamed for improved comprehension.
2. Joining: the three population datasets from the World Bank were joined after each was cleaned and this merged population dataset was merged with the COVID-19 dataset to produce the final result.
3. Filtering: filter functions were used to isolate the top 10 countries by number of confirmed COVID-19 cases in all four datasets.
4. Aggregation – as the COVID-19 dataset was spread out to identify growing number of recorded cases by day, all cases were aggregated by their sums from first reported date to the most recent date.

Load

While the original datasets might have been more suited to relational databases, the final table was merged and filtered to a degree that a single nonrelational database was suitable. As the query of interest centered around fast developing events, the risk of the data being compiled and/or its structure lacking validity or facing large amounts of new incoming data also favored the greater flexibility of a nonrelational database. Even so, the final data was structured in such a way as to allow for a relational structure via a clearly formatted index column (primary key) and the supplemental ‘country code’ column (potential foreign key)

report should include

* descr original data sources, its format (csv)
* what data cleaning was required
* the final database and why it was chosen