ETL Technical Report

The steps below will reproduce the process which produced the final result of my ETL project.

1. importing of pandas and pymongo to clean the data and export it to a database.
2. reading in the four csv files
3. converting csv files into pandas dataframes

Specific cleaning of the COVID-19 dataset included

1. dropping the "Country/Region", "Confirmed", "Deaths", "Recovered" columns
2. performing a .groupby on the "Country/Region" column and aggregating by sum
3. arranging the dataframe by descending order based on most confirmed cases of COVID-19 by country
4. dropping all but the top 10 most affected countries from dataframe
5. creating a “country code” list and appending it as an additional column, anticipating a merge with the population age datasets.
6. formatting the index and column order

Specific cleaning of the population age datasets included

1. removing all columns save "Country Name", "Country Code", "2018" the most recent year of data
2. renaming the “2018” column on each dataset to make clear what portion of the population they reflected (ex. "2018": "Pop % 65 and over (2018)")
3. filtering by the top 10 countries most affected by COVID-19 as calculated in the COVID-19 dataset
4. merging all three population age datasets
5. performing a final merge with the COVID-19 dataset on the “Country Code” column

Once the final dataset was produced the final steps were

1. connecting to MongoDB
2. transforming the pandas dataset into a list of dictionaries to be readable by MongoDB
3. inserting the final data into the MongoDB database