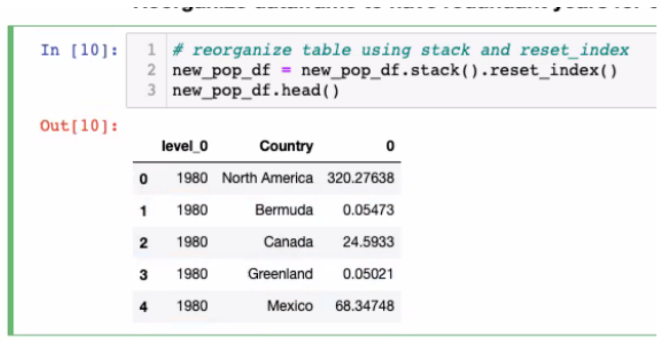
This project was an ETL (Extract-Transform-Load). The purpose of this was to bring two separate data sets together to help answer a question. A flow chart outlining the ETL process is included in the GitHub repository ReadMe file. This process is explained in more detail below.

The question to identify was whether countries win more Olympic medals based on their population. Two datasets from Kaggle that would identify which countries have won the most Olympic Medals by population were needed. For this, an Olympic medal data set that shows total medals won by each country and a population by country data set were used.

The next step was to transform the data. First, the medals csv file needed to be brought into python. The cleaning process included deleting columns that were not needed and renaming columns. The next step was to remove data that was not useful from this data set, based on the population data set. This involved removing all Olympic Data before 1980 due to the population data only going back that far. Once completed, there was a useful Olympic medal data set.

Next was to work with the population data. In order to transpose the data, a simple transpose function would not work. Instead, by using a stack and reset index function, the population dataset was able to match the layout of the medal data set. Below is how it looked. This converted the data frame from matrix form to table form.

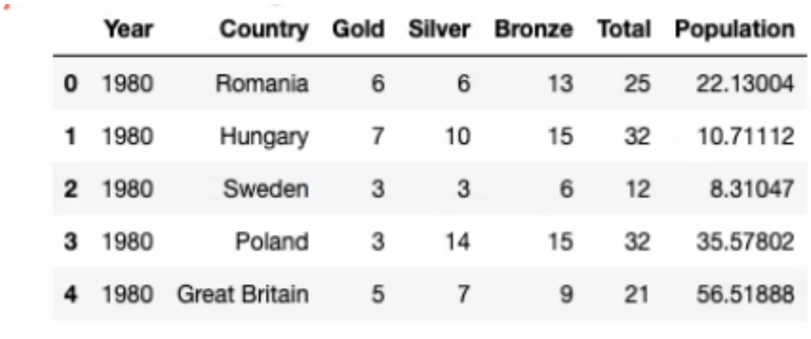


```
In [10]: # reorganize table using stack and reset_index
         new_pop_df = new_pop_df.stack().reset_index()
         new_pop_df.head()
```

```
Out[10]:
      level_0        Country           0
   0     1980  North America   320.27638
   1     1980        Bermuda     0.05473
   2     1980         Canada     24.5933
   3     1980      Greenland     0.05021
   4     1980         Mexico    68.34748
```

Once the population data set matched in terms of layout, the next cleaning steps were a bit more extraneous. First, the country data needed to be a renamed dictionary to identify any mismatches in how each table referred to countries, i.e., Cote de’ Ivoire and the Ivory Coast. The next step was to eliminate some outliers that would not match the other data set - such as the Unified Team and IOP, which were both one-off situations due to the breakup of the USSR.

This process above ensured the data was cleaned in a manner that truly identified each country that truly won a medal and identified their population during every Olympic games. The merged data set looked like this.



	Year	Country	Gold	Silver	Bronze	Total	Population
0	1980	Romania	6	6	13	25	22.13004
1	1980	Hungary	7	10	15	32	10.71112
2	1980	Sweden	3	3	6	12	8.31047
3	1980	Poland	3	14	15	32	35.57802
4	1980	Great Britain	5	7	9	21	56.51888

The last step once having the merged data set was to Load the data from Python into PgAdmin. To do this, a database was created in PgAdmin. The table schema was created, and data was loaded into the table. The scope of this exercise was to work on the ETL process, rather than to conduct analysis.

Future research and analysis, could include

- Including a medals per capita column
- Including countries that competed in the Olympic Games but did not win any medals (not in the starter dataset from Kaggle).
- Including the Winter Olympic games.

Having these three data points would help answer the question - *Do countries with a higher population win more medals, and are results similar for the Summer vs Winter games?*