ETL Project: Annual Fishing Data

Gizem Yali-Pepper, Khalid Ettalii, Omar Eltorai

# **Executive Summary**

As a newly formed research team at the World Marine Biological Center (WMBC), we were tasked with investigating the impact that the fishing industries of different countries have on the fish species of the North Atlantic. In order to conduct our analysis, we found multiple datasets which contained this information. By combining the data in a single database, we were able to more readily conduct our analysis.

The following technical report outlines the steps taken by our team to clean, combine, store, and then access the data.

# **Extract:** Data sources

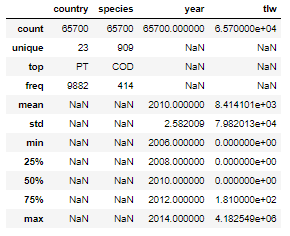
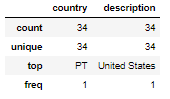
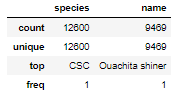
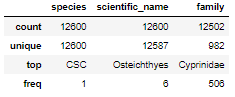
The datasets provide data of annual nominal catches of more than 200 species of fish and shellfish in the Northeast Atlantic region, which are officially submitted by 20 [International Council for the Exploration of the Sea (ICES)](http://www.ices.dk/) member countries between 2006 and 2014. The unit metric used to measure the catch quantity is total live weight (“TLW”).

* There are three primary datasets, in .csv and .xlsx file formats that were used:
  1. country\_code.csv
  2. fish\_catches.csv
  3. fish\_name.xlsx
* Link: <https://www.kaggle.com/victorgenin/ices-fish-catch/version/1#RECO_Export_08-04-2016.csv>

# **Transform:** Data transformation

Cleaning activities performed included primarily data dropping, transposing (using “melt” in pandas), aggregating, and merging. The majority of the data cleaning and transformation was performed using pandas.

For data dropping, we went through each file and determined which data fields were relevant and which were not relevant to this project. We dropped the fields that we determined were not relevant (shown in the table to the right).

After filtering for only the relevant data, we then performed data transformation as needed. In the “fish\_catches.csv” data set, we transposed the data so that the annual data was not broken into separate columns, but was put into a single column, and we added a new column that corresponded with the year. We were able to do so using the ‘melt’ functionality in pandas to unpivot a DataFrame from wide format to long format.

***Fish – Table Data Summary***

***Country – Table Data Summary***

***Fish Name – Table Data Summary***

***Summary of Data Kept vs Dropped***

***Catch – Table Data Summary***

After observing a discrepancy between the country data in the “fish\_catches.csv” and “country\_code.csv” data sets, we combined the two data sets in pandas to remove the catch data for the countries which were not included in the “country\_code.csv” set.

# **Load:** Final Production Database

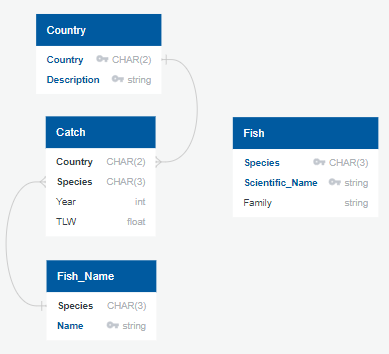
Once the three individual data sets were cleaned, we then loaded the data into a relational database using PostgreSQL. The database was named “etl\_db”. Four tables were created in the schema, and included (column header shown in parentheses):

1. catch (country, species, year, tlw) – contains country code, fish code, year, and total live weight (tlw) caught
2. country (country, description) – contains the country code and name of the country
3. fish (species, scientific name, family) – contains the fish code and taxonomic details for the fish
4. fish\_name (species, name) – contains the fish code and common name for the fish

With the tables created, the data from the csv files were imported to populate the newly created tables. The populated tables in the database were then accessed in Python libraries (pandas, matplotlib, etc.) using the SQLAlchemy ORM.

The diagram below maps the data base and the relationships between the different tables.

## Entity Relationship Diagram (ERD)



## Data Dictionary

**Country:** Two (2) letter country code based on the international naming convention.

**Description:** The full name of the country (e.g., BE=Belgium).

**Family:** A group of species with similar characteristics.

**Species:** Three (3) letter code corresponding to the fish species.

**Scientific\_Name:** The scientific name (taxonomy name) for the fish caught.

**TLW:** Tons live weight, the metric used to measure the size of catch.

**Year:** The year that corresponds to the timing of the catch.

## Table Schema

