**ETL Project: Sugar Analysis**

A picture containing indoor

Description automatically generated

By: Caitlan Beachey, Amaris Hassan, Cecilia Leung, Hillary Mandich, Kapir Pundhir

Date: December 19th, 2020

Contents

[**1. EXECUTIVE SUMMARY** 3](#_Toc59022705)

[**2. EXTRACTION** 4](#_Toc59022706)

[2.1 FIND THE OPTIMAL DATASETS 4](#_Toc59022707)

[2.1 RELATIONSHIP OF DATASETS 4](#_Toc59022708)

[**3. TRANSFORMATON** 6](#_Toc59022709)

[**DATA CLEAN-UP FOR CSV FILES** 6](#_Toc59022710)

[3.1 DELETING UNWANTED COLUMNS 6](#_Toc59022711)

[3.2 COMBINING USEFUL INFORMATION FOR ANALYSIS 6](#_Toc59022712)

[3.3 ADD COUNTRY CODE 7](#_Toc59022713)

[3.4. CONNECT DATA BY INSERTING COUNTRY CODE 7](#_Toc59022714)

[3.5 NORMALIZING 7](#_Toc59022715)

[3.6 REFORMATTING FROM STRING TO FLOAT 7](#_Toc59022716)

[**B. DATA CLEAN UP USING PANDAS** 8](#_Toc59022717)

[3.7 MELTING 8](#_Toc59022718)

# **1. EXECUTIVE SUMMARY**

ETL stands for Extract, Transform and Load. Extracting involves collecting, reading, and migrating large volumes of raw data from various sources into one easily accessible database. Transformation makes data meaningful by reformatting, filtering, transposing, merging, and joining. Finally, loading uses various applications/software to load the data into the faster and produce faster results.

ETL provides numerous benefits as it breaks down data silos by gathering all relevant data into one database. Another advantage is it greatly assist the Data Analyst to analyze the data and turn it into business intelligence. Finally, with data readily available, it helps you to make better decisions in a timely manner.

In this proposal, we have conducted the ETL process through our Sugar Analysis Datasets. We will explain how we completed the extraction, transformation and loading the data in the SQL Web Server using various techniques and software applications. Finally, we will provide a summary of the process that will explain why we did what we did, limitations, and next steps.

# **2. EXTRACTION**

**This process involves finding and collecting datasets from various types of sources.**

## 2.1 FIND THE OPTIMAL DATASETS

We used a total of 5 CSV Datasets from Kaggle that came from various sources like WHO, FAO, World Bank and Wikipedia. We were able to collect the data for at least 190 countries in each dataset.

The sources of our datasets are as follows:

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Source** | **Website** |
| Sugar Consumption | FAO | https://www.kaggle.com/angelmm/healthteethsugar?select=sugar\_consumption.csv |
| Health Expenditure | WHO | https://www.kaggle.com/angelmm/healthteethsugar?select=healthexpend.csv |
| Income | Worldbank | https://www.kaggle.com/frankmollard/income-by-country |
| Obesity | WHO | https://www.kaggle.com/amanarora/obesity-among-adults-by-country-19752016?select=obesity-cleaned.csv |
| Country Codes | Wikipedia | https://www.kaggle.com/andradaolteanu/iso-country-codes-global?select=wikipedia-iso-country-codes.csv |

## 2.1 RELATIONSHIP OF DATASETS

Illustrating an ERD diagram is vital in database design because it:

* Increases understanding of how relationships and entities of the data.
* Decreases ambiguities and unnecessary processes.
* Outlines what applications and methods are needed to transform and load data.

ERD Diagram:

Diagram

Description automatically generated

# **3. TRANSFORMATON**

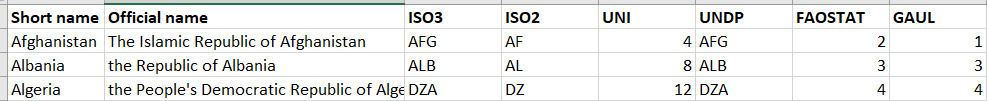
**Transformation is the process of converting the extracted data from its previous form into the form it needs to be in so that it can be placed into another database.**

## **DATA CLEAN-UP FOR CSV FILES**

## 3.1 DELETING UNWANTED COLUMNS

**COUNTRY TABLE**

* Only one set of country code is necessary for performing data analysis.
* We only kept the ISO3 column and deleted the rest.



## 3.2 COMBINING USEFUL INFORMATION FOR ANALYSIS

**COUNTRY TABLE**

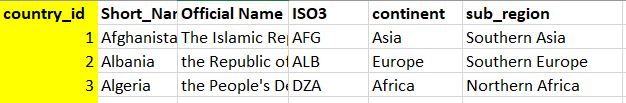
* Add columns, Continent and Sub-Regions through lookup tales
* [Need Hillary to confirm where those info came from]



## 3.3 ADD COUNTRY CODE

**COUNTRY TABLE**

* Country code will be inserted in the first column of the Country table to identify each country as a unique value.
* First row will be 1, second row will be 2, and so forth.



## 3.4. CONNECT DATA BY INSERTING COUNTRY CODE

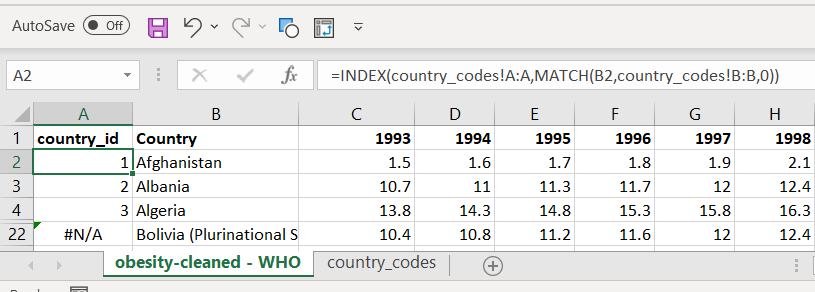
**INCOME TABLE, OBESITY TABLE, SUGAR CONSUMPTION TABLE, HEALTH EXPENDITURE TABLE, COUNTRY TABLE**

* Open the country table, copy and move tab to each of the 4 tables.
* Add a country\_id column on the leftmost side of each of the 4 tables.
* Conduct a LOOKUP to obtain the country ID for each country.
* Cannot have #N/A for country ID
  + If there is #N/A, determine if the country names on the respective tables match the country table.

RESULTS:

**Country Table (Lookup Value)**

**Obesity Table (Insert INDEX MATCH Formula on first column)**



ELIMINATING #N/A TO OBTAIN COUNTRY CODE:

Country Table (Lookup Value)

Obesity Table (Get rid of #N/A by matching country name)

## 

## 3.5 NORMALIZING

**INCOME TABLE, OBESITY TABLE, SUGAR CONSUMPTION TABLE, HEALTH EXPENDITURE TABLE**

* Since we have a country code as a unique identifier for each table, we do not need to have all the country names that may take up storage space.
* The unique identifier have linked all the tables together for SQL.

## 3.6 REFORMATTING FROM STRING TO FLOAT

Obesity

[Replace country codes with country names by linking tables to the country table]

## **B. DATA CLEAN UP USING PANDAS**

## 3.7 RENAMING COLUMNS

## 3.8 FILL NA

## 3.8 MELTING

By Country Tables

[moved the years as a column that will be easier for querying]