ETL-PROJECT

Obesity database

BY Grant Thornburgh, Muhammad Angel, and Luminda Kulasiri

Data Boys, Inc.

**Table of Contents**

1. **Introduction**
   1. Summary 2
   2. Scope 2
   3. Technologies and Resource Contributions 2
   4. Definitions, Acronyms and Abbreviations 2

**2. ETL Details**

2.1 Data Import/Extract Sources and Method 3

2.2 Data Acquisition 3

2.3 Data Transform 3

2.4 Data Integrity 4

2.5 Data Refresh Frequency 6

2.6 Data Security 6

2.7 Data Loading and Availability 7

**3. Data Quality**

1. **Introduction**
   1. Summary

The goal of our project is to look at the relationship between a country’s economic conditions and its level of obesity. In order to draw the relationship, GDP as an indication of a country’s economic standing and BMI as an indication of the rate of obesity in said country can be used. Two data sources have been selected and can be used to obtain rates of obesities and GDP for all the countries for specific years, and are given as follows. We merged the two data sets on country name, as this is a unique identifier.

* 1. Scope

Our dataset from Kaggle examines country’s levels of obesity. These values are given as a percentage and show an average across the time period of 1975-2016. Our data set from Wikipedia shows a country’s GDP per capita. These values are listed in dollars.

<https://www.kaggle.com/amanarora/obesity-among-adults-by-country-19752016>

[https://en.wikipedia.org/wiki/List\_of\_countries\_by\_GDP\_(nominal)\_per\_capita](https://slack-redir.net/link?url=https%3A%2F%2Fen.wikipedia.org%2Fwiki%2FList_of_countries_by_GDP_(nominal)_per_capita)

* 1. Technologies and Resource Contributions

The technologies we used to format this data were Python and Postgress Server with PG Admin 4. Luminda scraped and cleaned the country GDP data from Wikipedia, he also merged the clean data sets for GDP and obesity levels by country. Angel worked on cleaning and formatting the country obesity level data from Kaggle. Grant reviewed the initial and managed the team, he also wrote the final report and presented it to the client.

* 1. Definitions, Acronyms and Abbreviations

GDP stands for Gross Domestic Product, the value represents the overall economic status level of the country. GDP per capita is a country’s GDP is divided by the number of citizens in the country, giving a relative value. BMI stands for Body Mass Index and is a person’s weight in kilograms divided by the square of height in meters. We initially thought that a high BMI can be an indicator of high body fatness. A BMI greater than 30 is considered obese, but the table values are percentages. The values used in this set are actually age-standardized rate, which is a better overall indicator. A more detailed explanation of calculating the age-standardized rate can be found at,

<https://meteor.aihw.gov.au/content/index.phtml/itemId/327276>

**2. ETL Details**

2.1 Data Import/Extract Sources and Method

We downloaded the country obesity data as a CSV file and imported the file into python. We were able to scrape the GDP data from Wikipedia and imported it directly into a pandas dataframe in python.

2.2 Data Acquisition

This data was acquired from publicly available sources, so we did not have any issues with this process.

2.3 Data Transform

Once our data was imported to python we cleaned the data by only keeping data from 2016 and 2017 and dropping all of the rest of the data. We then dropped all unnecessary columns. We also changed the headers of the columns to “total”, “male” and “female” for the obesity values and “GDP” for the GDP values. We also had to make slight adjustments to some of the country values so that they were consistent in both data sets. After we merged both data sets we searched for and dropped the null values that resulted from the merge.

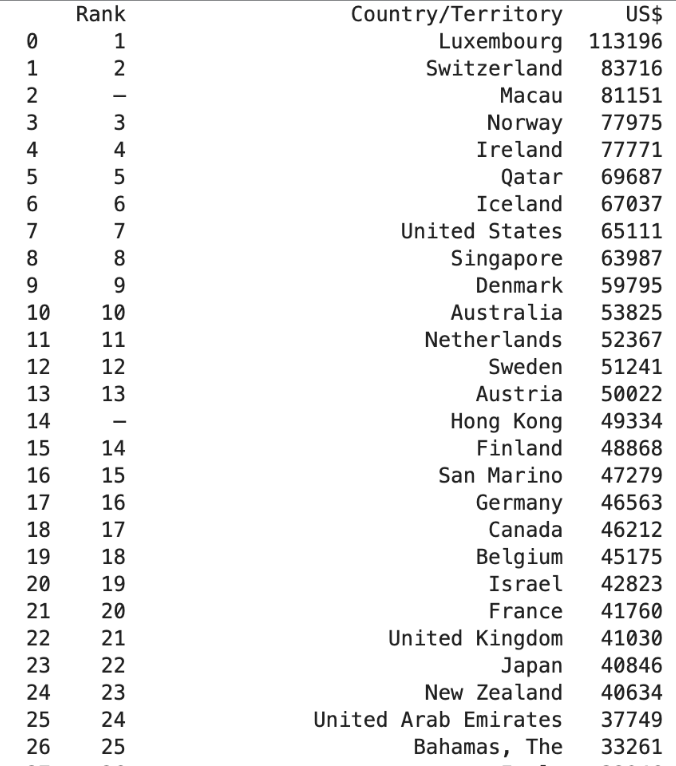


Figure 1: GDP data before cleaning up

A screenshot of a cell phone

Description automatically generated

Figure 2: Obesity data table before cleaning up

2.4 Data Integrity

In our final dataset we had 165 countries with values for obesity and GDP, we found this to be an acceptable dataset for further analysis. We saved these values to a Database in PostgreSQL called Obesity. Our table names are: income, obesity, combined data. Our columns in income are: country, GDP. Our columns in obesity are: country, Total, Male, Female. Our columns in combined data: country, Total, Male, Female, GDP.

A screenshot of a cell phone

Description automatically generated

Figure : Entity Relationship Diagram for the 'obesity' database

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Figure 4: Two tables after cleaning up (GDP-left and Obesity-right)

2.5 Data Refresh Frequency – N/A

We have pulled all of the necessary data into the database, because our data is static refreshing the data is not necessary. The data provided is from years 2016 and 2017 which is fairly recent, we do not believe that this data will need to be updated for at least another year.

2.6 Data Security

The data used for this project is publicly available so there were no issues with security. These data sets did not list personal information and were designed to give population data.

A screenshot of a cell phone

Description automatically generated

Figure 5: Table 'obesity' was created by merging the two tables

2.7 Data Loading and Availability

Our client can access the database using both python dataframes and PostgreSQL.

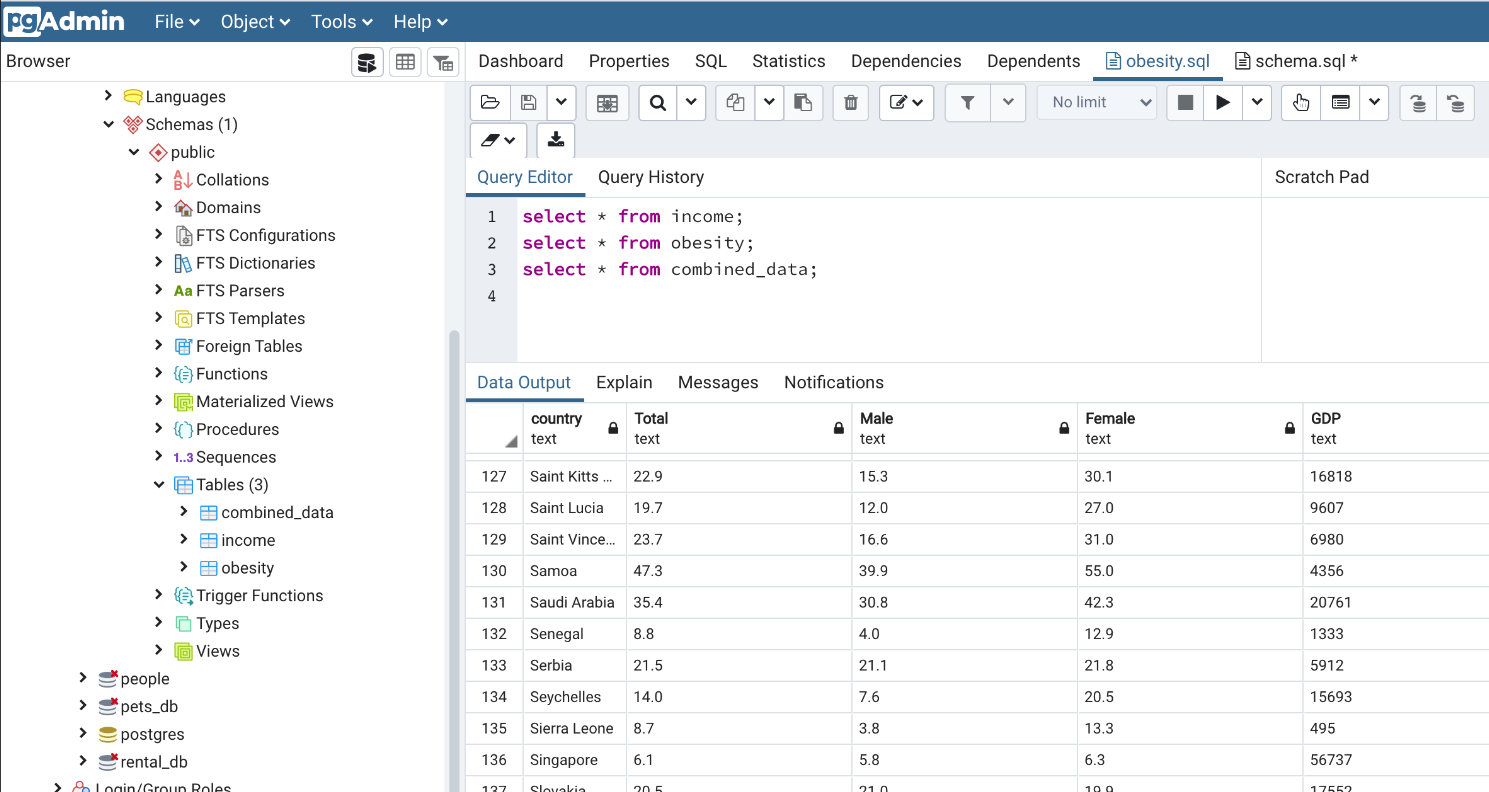


Figure 6: This is an example of the database in PG Admin

3.0 Data Quality

We tested the database by creating a plot for Average Obesity vs. GDP per capita. The following figure shows the results of the plotted data.

A close up of a map

Description automatically generated

Figure : A plot, Average Obesity vs. GDP per capita