

Cardiovascular disease



Manuel Alejandro Gruezo

Jhonatan Steven Morales

Carol Dayana Varela

ETL PROJECT 1

Teacher: Javier Alejandro Vergara Zorrilla

Universidad Autónoma de Occidente

Faculty of Engineering

Santiago de Cali

2024

1. Introducción

This project aims to demonstrate knowledge in data management and visualization creation. The project involves using a dataset to perform an ETL (Extract, Transform, Load) process, followed by exploratory data analysis (EDA) and creating visualizations to help interpret the data stored in a relational database.

2. Cardiovascular Disease project

This Data Engineering project was conducted with the objective of efficiently handling, transforming, and visualizing data related to cardiovascular disease. The project involved several phases, including data setup, exploratory data analysis (EDA), data processing, dashboard creation using Power BI, and repository management through GitHub.

The data used comes from two datasets:

- **Cardiovascular Disease dataset:** A dataset containing 70,000 records about patients and 12 health-related features.
- **Causes of Death dataset:** A dataset detailing causes of death by country and year, with a focus on cardiovascular diseases.

3. Tools Used:

Before starting, ensure you have the following installed:

- **Python**
- **Jupyter**
- **SQLAlchemy**
- **PostgreSQL**
- **Power BI**
- **GitHub**

Additionally, you need access to PostgreSQL database credentials, which should be stored in a `.env` file with the following variables:

- **PGDIALECT** = The database dialect or type. In this case it is set to postgres
- **PGUSER** = Your PostgreSQL database username.
- **PGPASSWORD** = Your PostgreSQL database password.
- **PGHOST** = The host address or IP where your PostgreSQL database is running.
- **PGPORT** = The port on which PostgreSQL is listening.
- **PGDB** = The name of your PostgreSQL database.
- **WORK_DIR** = the location for you root of the project

4. Configuración del Entorno

1. Clone the project repository.
2. Create and activate a virtual environment

```
python -m venv venv  
.\venv\Scripts\Activate.ps1
```

3. Install dependencies:

```
pip install -r requirements.txt
```

Module Descriptions

- **db_connection.py** This module establishes a connection to the PostgreSQL database using SQLAlchemy. Connection credentials are obtained from the .env file.
 - Functions:
 - **getconnection():** Establishes and returns a connection to the database engine.
- **models.py** This module defines ORM (Object-Relational Mapping) models using SQLAlchemy. Models represent the structures of tables in the database.
 - Classes:
 - **CardioTrain:** Represents the table storing the original data from the cardiovascular disease dataset.
 - **CauseOfDeaths:** Represents the table storing data from the causes of death dataset.
 - **CardioTrainNormalize:** Represents the table with normalized data from the cardiovascular disease dataset.
 - **CholesterolTypes:** Reference table for cholesterol levels.
 - **GlucoseTypes:** Reference table for glucose levels.
- **transformation.py** This module handles data transformation from CSV files before loading them into the database.
 - Classes and Methods:
 - **DataTransform:** Class responsible for transforming data from the cardiovascular disease dataset.
 - **gender_by_category():** Transforms gender column into categories.
 - **cholesterol_by_category():** Transforms cholesterol column into categories.
 - **gluc_by_category():** Transforms glucose column into categories.
 - **bmi():** Calculates the Body Mass Index (BMI) and adds it as a new column.
 - **days_to_age():** Converts age from days to years.
 - **normalize_gluc():** Normalizes glucose levels.
 - **normalize_cholesterol():** Normalizes cholesterol levels.
 - **StandardizeBloodPressure():** Standardizes blood pressure readings.
 - **CategorizeBMI():** Categorizes BMI.
 - **categorize_blood_pressure():** Categorizes blood pressure.

- **CalculatePulsePressure():** Calculates pulse pressure.
- **DataTransformCauseOfDeaths:** Class responsible for transforming data from the causes of death dataset.
 - **insert_id():** Inserts an ID column.
 - **drop_code():** Drops unnecessary columns.
 - **total_deaths():** Calculates total deaths.

6. Notebook Descriptions

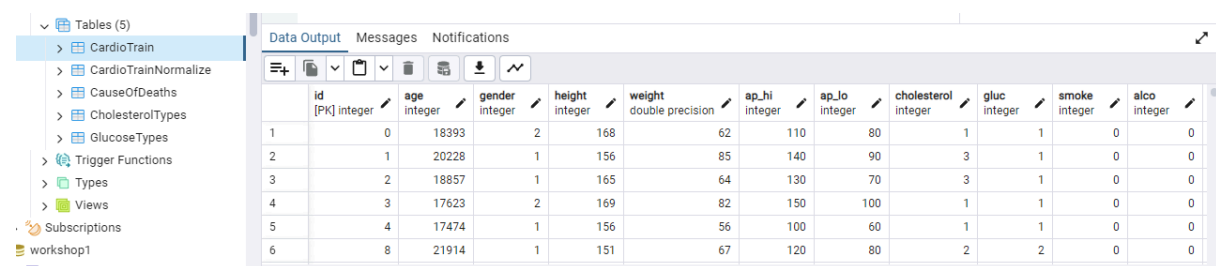
DataSetup.ipynb

This notebook focuses on the initial setup of the project, including:

- Connecting to the PostgreSQL database.
- Creating the necessary tables.
- Basic transformation of data from the CSV file.
- Inserting transformed data into the database..

The project began with the importation of a CSV file containing critical cardiovascular health data. This file was carefully selected for its relevance to the project, featuring various columns with structured information that needed efficient storage in a database. SQLAlchemy was employed to handle the interaction between Python and a PostgreSQL database.

A connection to a local PostgreSQL database was configured using SQLAlchemy, leveraging its capabilities to manage the communication between the programming environment and the database. The CSV file was read using pandas, and the data was inserted into pre-defined tables within the database. This setup ensured that the data was organized and readily accessible for the subsequent phases of the project.



	id [PK] integer	age integer	gender integer	height integer	weight double precision	ap_hi integer	ap_lo integer	cholesterol integer	gluc integer	smoke integer	alco integer
1	0	18393	2	168	62	110	80	1	1	0	0
2	1	20228	1	156	85	140	90	3	1	0	0
3	2	18857	1	165	64	130	70	3	1	0	0
4	3	17623	2	169	82	150	100	1	1	0	0
5	4	17474	1	156	56	100	60	1	1	0	0
6	8	21914	1	151	67	120	80	2	2	0	0

EDA.ipynb

Following data storage, an Exploratory Data Analysis (EDA) was conducted to gain a comprehensive understanding of the cardiovascular dataset. The EDA phase included:

- **Statistical Analysis:** Calculating metrics such as mean, median, standard deviation, and frequency distribution of key variables.
- **Correlation Analysis:** Evaluating relationships between different variables to identify significant correlations that could influence cardiovascular health.

- **Outlier and Error Detection:** Identifying and addressing outliers, missing data, and potential errors to ensure data quality.
- **Data Visualization:** Utilizing tools like pandas and matplotlib to create histograms, scatter plots, and box plots, which provided clear visual insights into the data distribution.

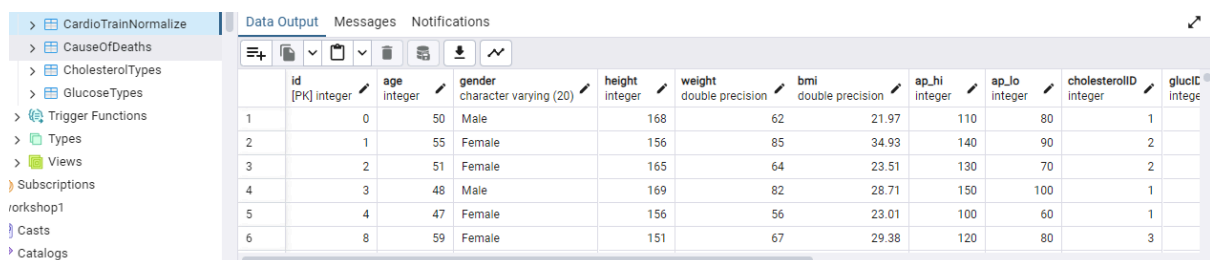
The EDA laid the groundwork for planning the necessary data transformations, optimizing data for the creation of a star schema used in later stages.

DataProcessed.ipynb

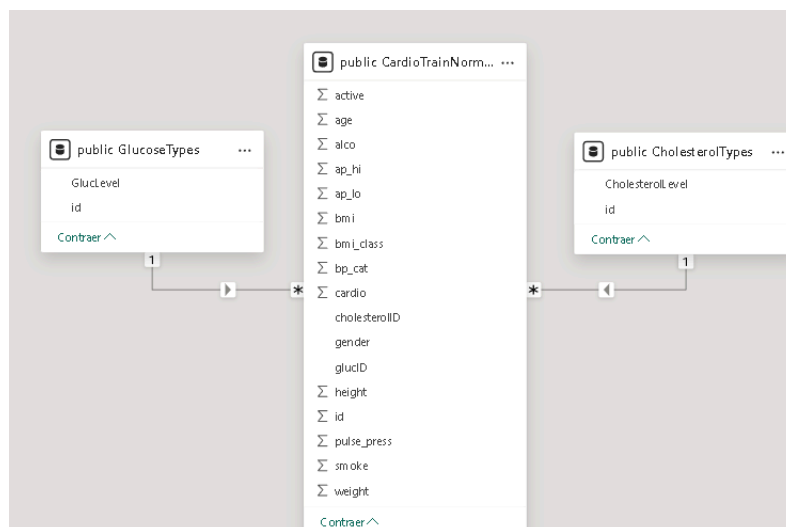
This notebook focuses on normalizing data before importing it into the database:

- Creating tables for normalized data.
- Executing complex transformations (categorization, normalization, calculation of new metrics).
- Inserting normalized data into the database.

Based on the insights from the EDA, the data was transformed into a star schema model, widely used in data warehouses to optimize analytical queries. The transformation involved creating fact tables to store events or transactions and dimension tables to contain descriptive attributes, significantly improving query performance and facilitating report generation



	id [PK] Integer	age Integer	gender character varying (20)	height Integer	weight double precision	bmi double precision	ap_hi Integer	ap_lo Integer	cholesterolID Integer	glucID Integer
1		0	50	Male	168	62	21.97	110	80	1
2		1	55	Female	156	85	34.93	140	90	2
3		2	51	Female	165	64	23.51	130	70	2
4		3	48	Male	169	82	28.71	150	100	1
5		4	47	Female	156	56	23.01	100	60	1
6		8	59	Female	151	67	29.38	120	80	3

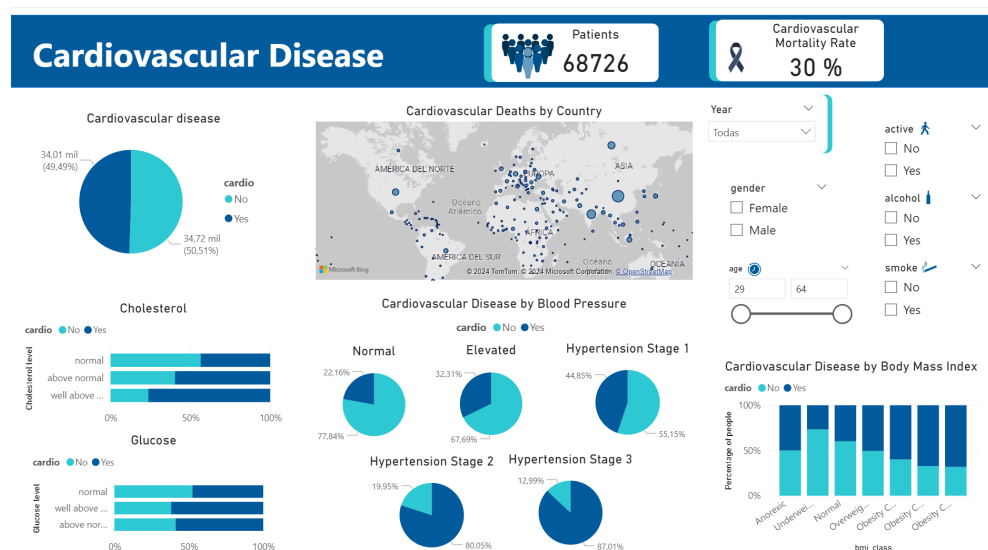


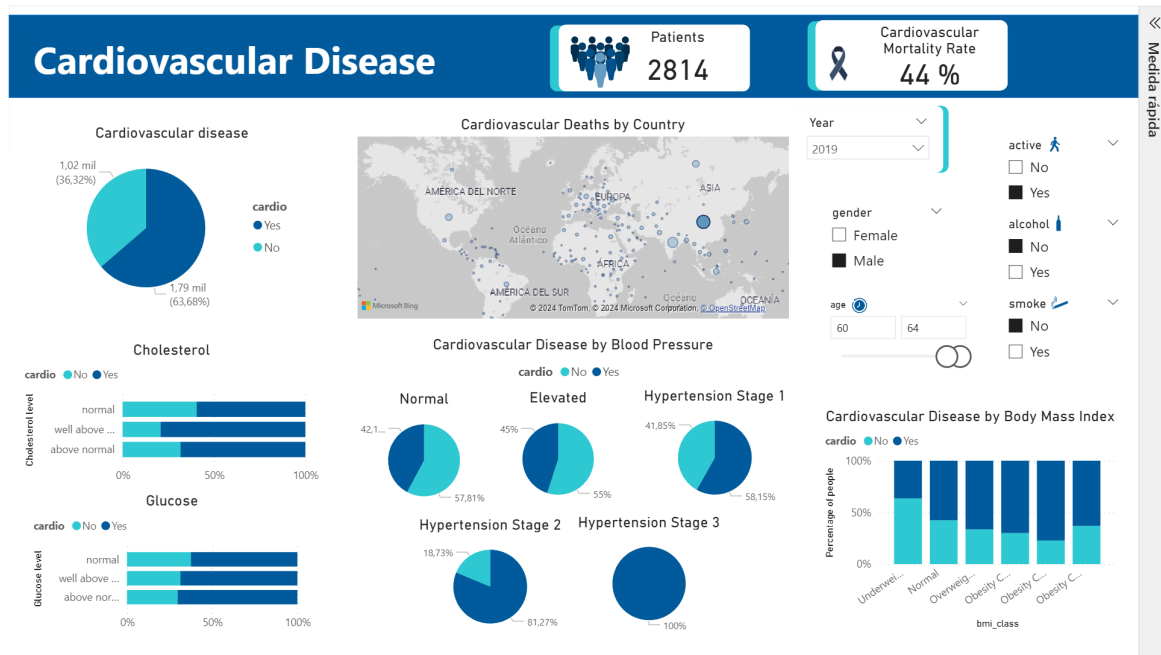
8. Power BI Dashboard

After completing the ETL process and running all the mentioned notebooks, a dashboard was created in Power BI to visualize and analyze results interactively. Various visualizations were used to represent key information from the dataset. Notable visual elements include:

- **Distribution of Cardiovascular Diseases:** A pie chart showing the proportion of patients with and without cardiovascular diseases.
- **Deaths from Cardiovascular Diseases by Country:** A world map visualizing deaths related to cardiovascular diseases, categorized by country and population size.
- **Relationship of Cardiovascular Diseases with Risk Factors:** Various visualizations exploring how factors such as cholesterol, glucose, BMI, and blood pressure influence the presence of cardiovascular diseases.
- **Interactive Filters:** Tools that allow data segmentation by gender, age, physical activity status, alcohol consumption, and smoking habits.

This dashboard facilitates dynamic exploration of the data, allowing users to identify patterns, correlations, and key trends effectively. It is a powerful tool for informed decision-making in public health and medical research.





9. Documentation and Git Repository

The entire project was meticulously documented, ensuring that every aspect of the workflow is clearly outlined and accessible for future reference. This documentation process involved capturing detailed technical specifications, including the configuration of the database, the code used for data extraction, transformation, and loading (ETL), as well as the settings and design choices made for the Power BI dashboard.

All this documentation was organized and stored in a Git repository, providing a centralized and version-controlled environment for the project. The repository includes a comprehensive README file that serves as a guide for users, offering step-by-step instructions on how to set up and run the project in various environments. This ensures that the project can be easily replicated or modified by others, facilitating collaboration and knowledge sharing.

Commits			
develop		All users	All time
Commits on Aug 28, 2024			
English Titles			
caroldvarela committed 14 hours ago	4f42ab5		
Commits on Aug 27, 2024			
Dashboard			
caroldvarela committed yesterday	Verified 561281f		
Blood Pressure transformation			
caroldvarela committed 2 days ago	f4d6342c		
Commits on Aug 26, 2024			

10. Conclusions

The project demonstrates a comprehensive ETL (Extract, Transform, Load) and data analysis workflow utilizing a Python and PostgreSQL-based environment. By establishing a robust database schema, performing thorough data transformation and normalization, and conducting an in-depth analysis, this project lays a strong foundation for advanced analytics and data-driven decision-making. The careful orchestration of these elements ensures not only the accuracy and reliability of the data but also enhances its usability for future applications. Through this process, the project showcases the critical steps required to transform raw data into meaningful insights, ultimately enabling more informed and strategic decisions within the domain of cardiovascular health or any other applicable field.