PROJECT REPORT

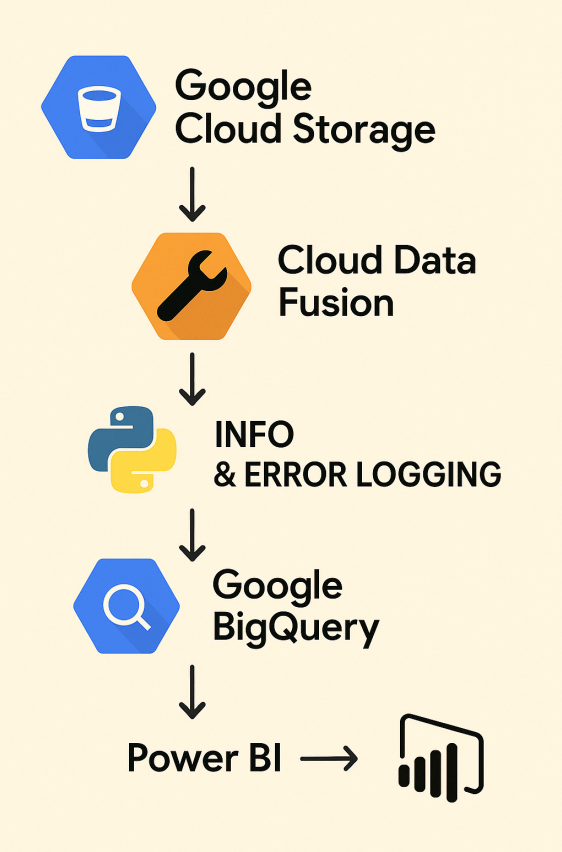
ETL Pipeline For Obesity Level Data Analysis Using GCP (Google Cloud Platform) And Power Bi

-Dinesh kumar raju kattunga

1.Introduction

This project focuses on building a scalable and automated ETL (Extract, Transform, Load) pipeline using Google Cloud Platform (GCP) to analyze obesity level data. Cloud-based solutions provide significant advantages in terms of scalability, automation, orchestration, and integration, which is why GCP was chosen as the primary platform.

The ETL process was executed using **Cloud Data Fusion**, while **BigQuery** was used for data warehousing and exploratory data analysis (EDA). Further, the processed data was visualized using **Power BI**, offering interactive dashboards and insightful visual analytics.



2.Tools and Technologies Used

- **GCP Cloud Storage** – For storing the raw dataset (.csv file).
- **Cloud Data Fusion** – For building a no-code/low-code ETL pipeline with built-in orchestration and version control.
- **BigQuery** – For scalable storage, querying, and EDA.
- **Power BI** – For building interactive dashboards and visualizations.
- **Cloud Dataproc & Log Explorer** – For job status monitoring and debugging.

Cloud Data Fusion's orchestration features and integrated version control made it easier to manage pipeline development and deployment efficiently.

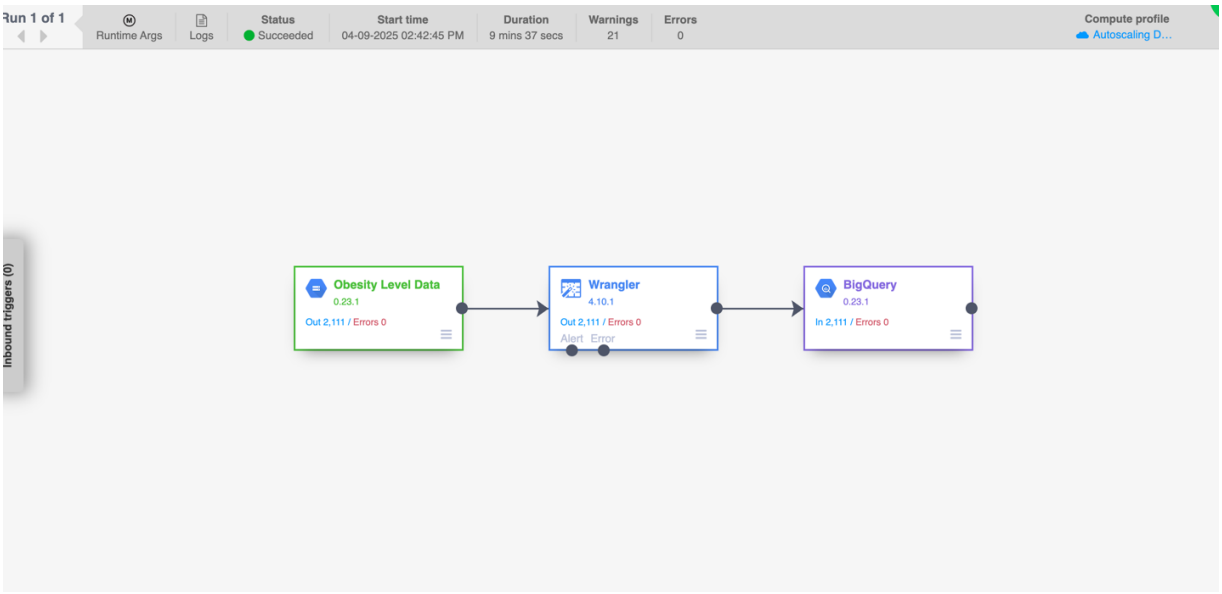
3.ETL Pipeline Overview

Why Cloud Data Fusion?

Cloud Data Fusion was chosen due to its visual interface, ease of use, and powerful capabilities:

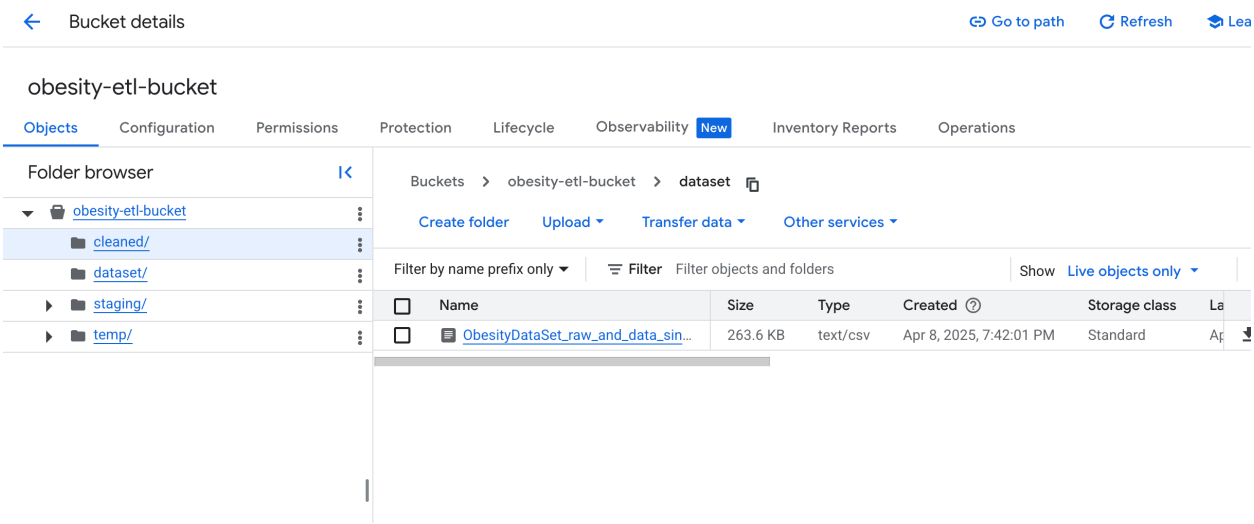
- **No-code/low-code pipeline building**, which accelerates development.
- **Built-in orchestration** which manages execution and job dependencies.
- **Version control** to track changes and enable collaborative development.
- **Seamless integration** with other GCP services like BigQuery and GCS.
- **Validation and data preview features** that help identify and fix issues early in the pipeline.

This tool enabled me to build an efficient and maintainable data pipeline without writing extensive code, aligning perfectly with the goals of this cloud-based data engineering project. Cloud Data usion was chosen due to its visual interface, ease of use, and powerful capabilities:



Step 1: Extract

The raw dataset was uploaded and stored in **Google Cloud Storage (GCS)**. This served as the data source for the ETL pipeline.



Step 2: Transform

The transformation was carried out using **Wrangler plugin** in Cloud Data Fusion, a visual tool that simplifies complex data transformations. The key transformations included:

- Cleaning column names for consistency.
- Converting categorical responses (e.g., "yes"/"no") into binary format (1/0) for columns like family history, FAVC, smoke, and SCC.
- Calculating **BMI** as weight / (height^2).
- Creating a **lifestyle score** using features such as FAF, TUE, CH2O, and behavioral inverses like (1 - FAVC), (1 - smoke), minus SCC.
- Adding **age group** classification (Youth, Young Adult, Adult, Senior).
- Updating data types for compatibility and accuracy.
- Renaming columns to improve clarity and analytical usefulness.

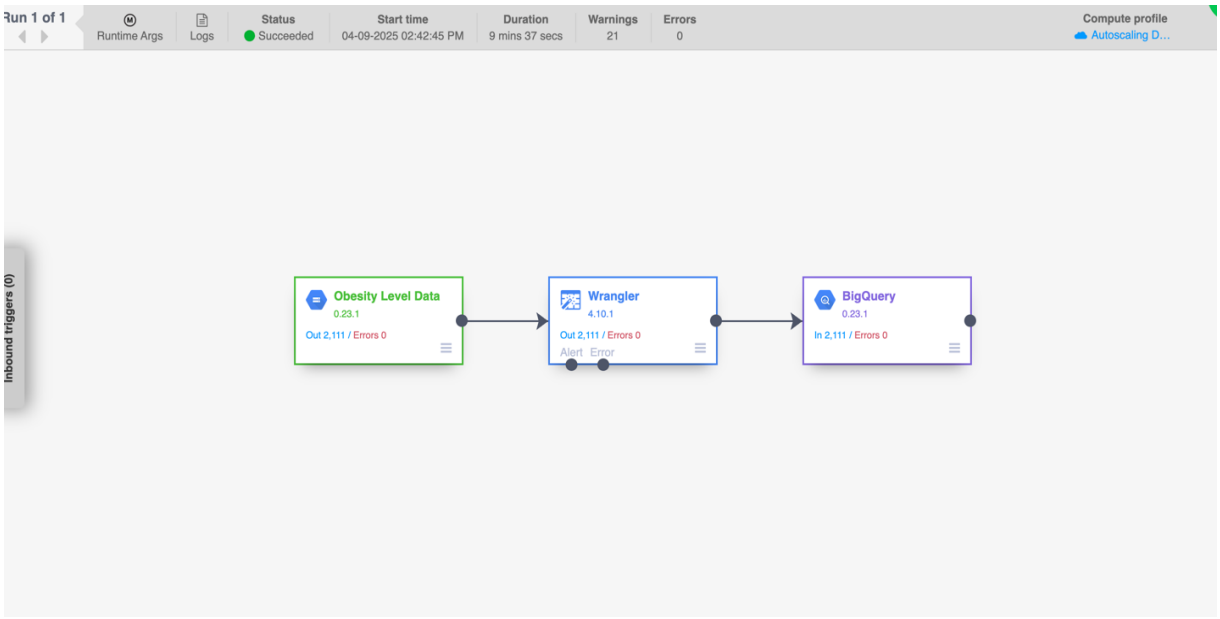
With this plugin, the required transformations were performed to prepare the dataset for both exploratory data analysis and potential machine learning applications.

These steps allowed the dataset to be ready for machine learning tasks and exploratory analysis. (Code Provided in txt file[wrangler_code.rtf])

Step 3: Load

The final cleaned dataset was loaded into **BigQuery**. A dataset named obesity_dataset and a table named obesity_cleaned were created to store the transformed data. The **BigQuery Sink** plugin in Cloud Data Fusion was used for this integration.

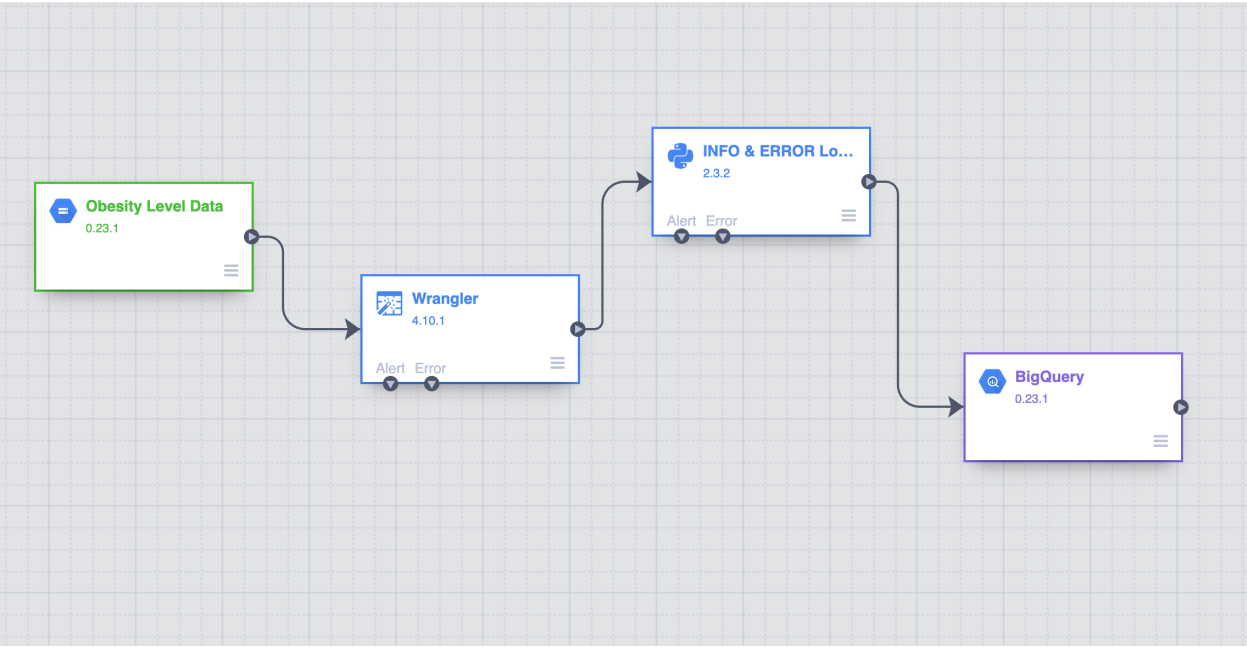
Before deployment, the pipeline was validated using Cloud Data Fusion's **preview and validation options** to ensure error-free execution. Once deployed, the job status was monitored through both Cloud Data Fusion and **Dataproc**, providing real-time feedback and logs.



4. Logging and Pipeline Monitoring

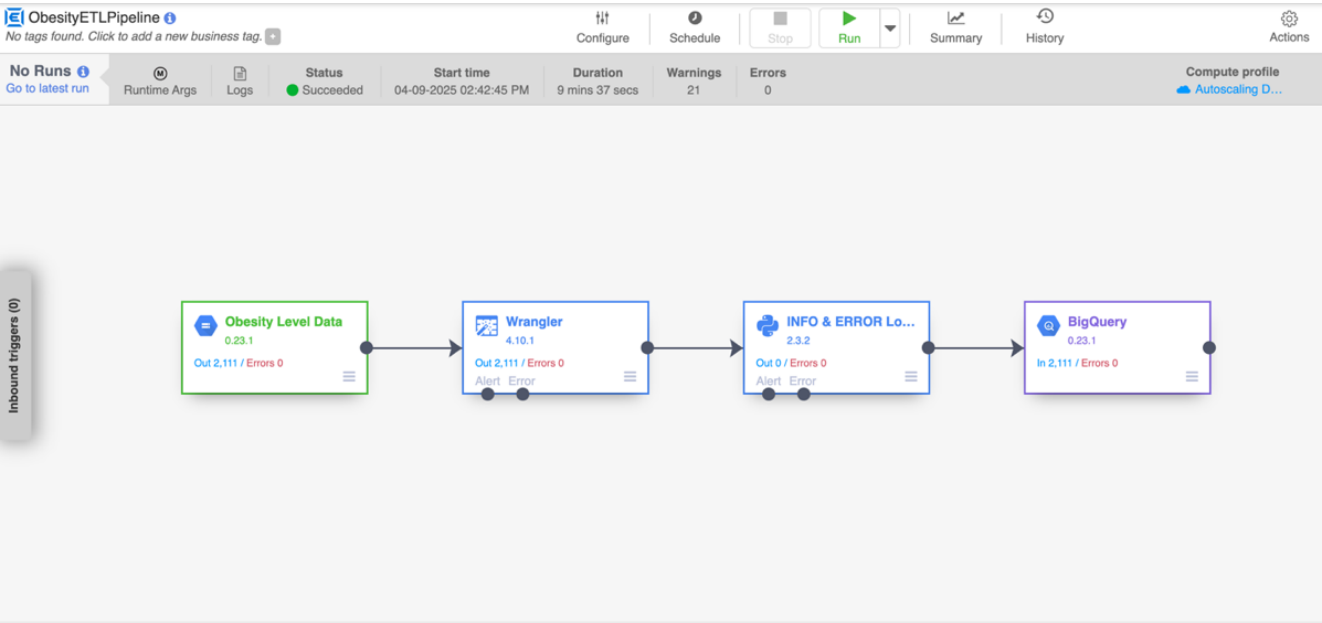
To prepare for future scaling and error tracing, **INFO and ERROR level logging** was implemented using Python Transform. This allowed tracking of important conditions (e.g., high BMI or missing data), categorizing records, and capturing exceptions—making the pipeline robust and maintainable, especially if data sources become dynamic in the future.

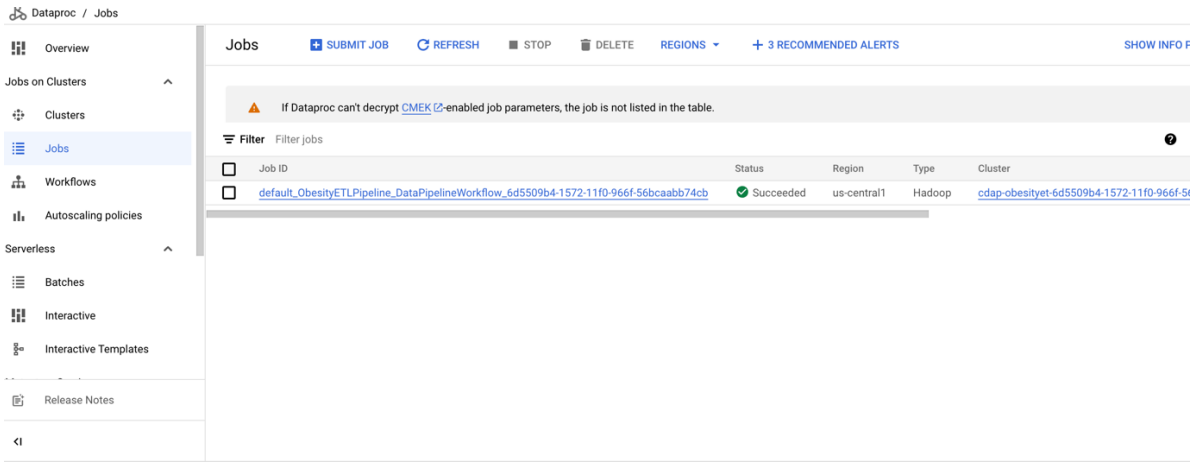
(Code provided in[Info_error_python.rtf])



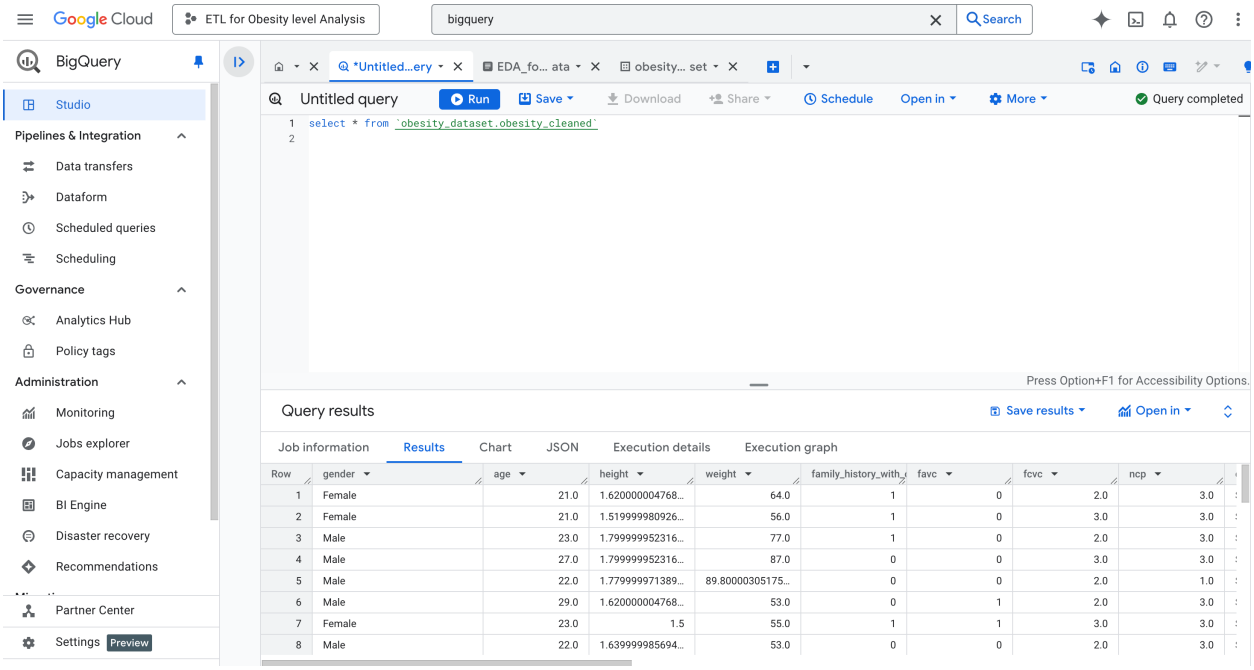
5. Redesign and Redeployment

Post-logging integration, the pipeline was redeployed successfully. The pipeline's success status was verified via Data Fusion and Dataproc UI. This step improved traceability and visibility into the ETL process.

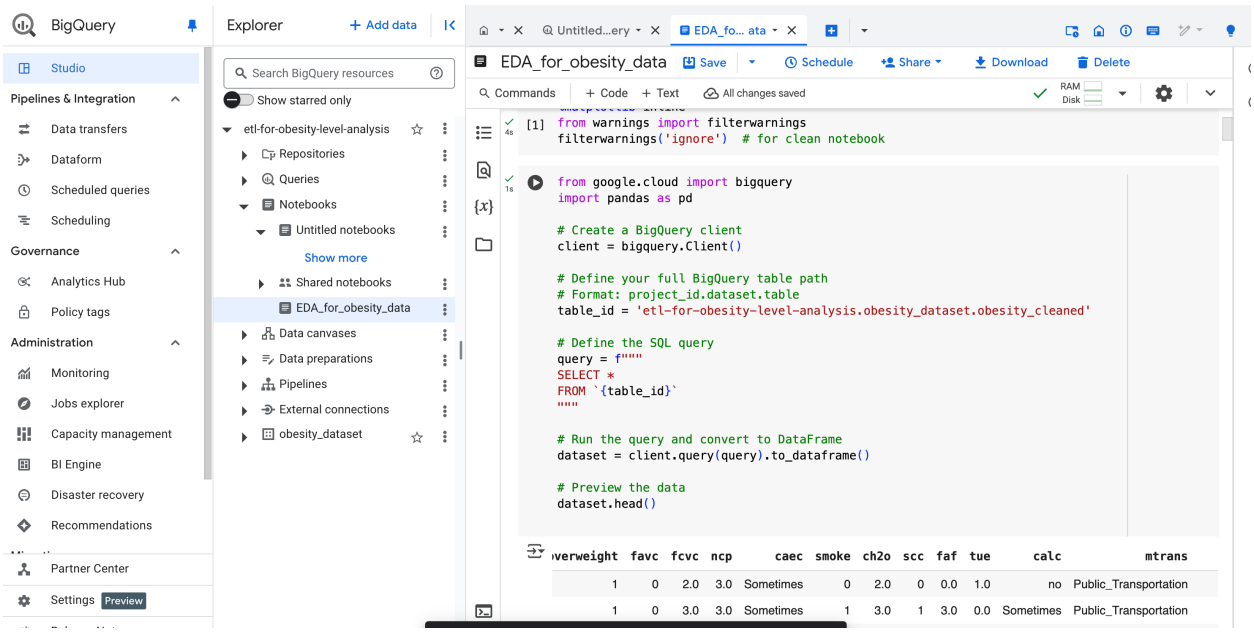




6. Data Validation and EDA in BigQuery:



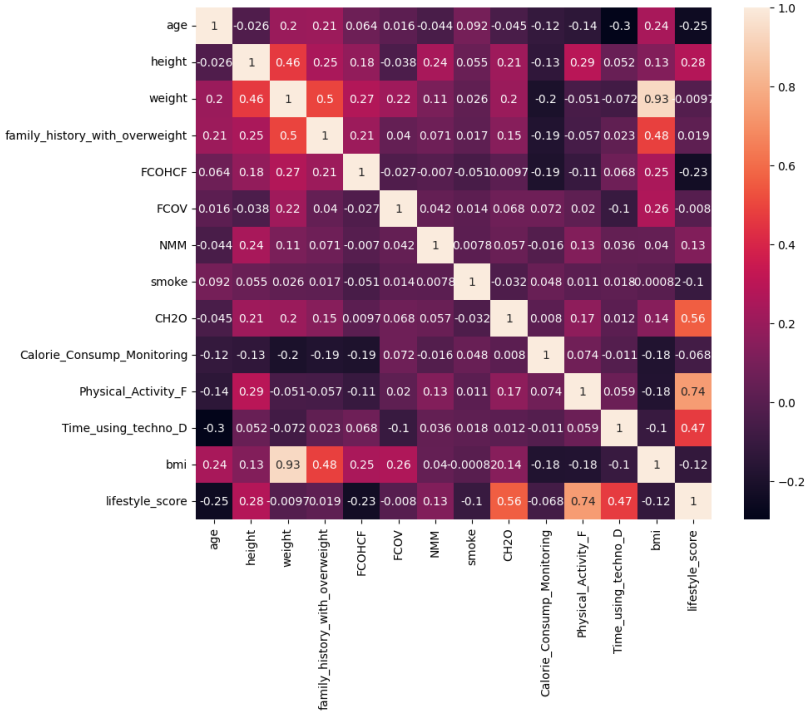
Once the data was loaded, **BigQuery Notebooks** were used to perform initial EDA directly within the GCP environment.



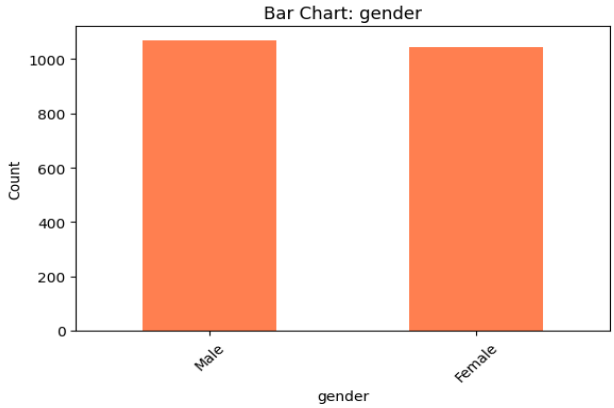
Key Findings from EDA: ([EDA_for_obesity_data.ipynb])

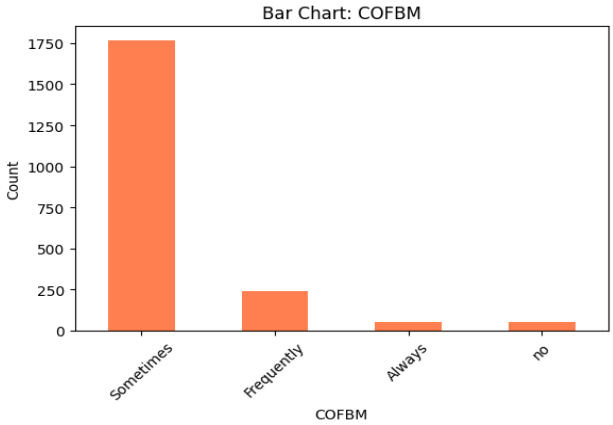
Correlation Matrix:

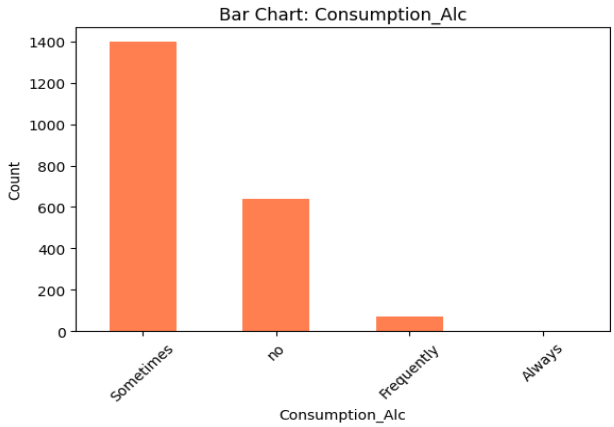
- BMI strongly correlates with weight.
- Family history shows a modest relationship with BMI.
- Lifestyle score correlates well with physical activity-related features.

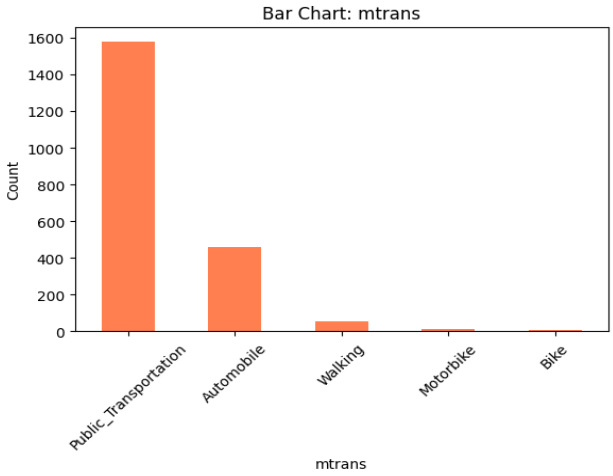


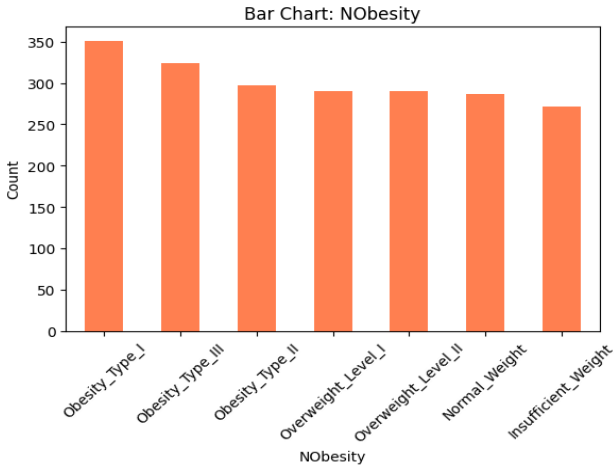
Categorical Distribution:

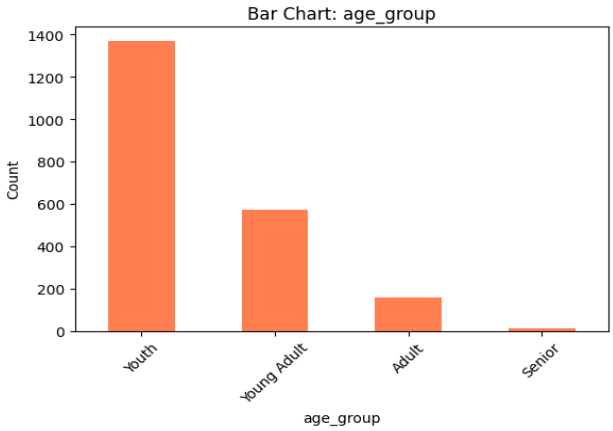






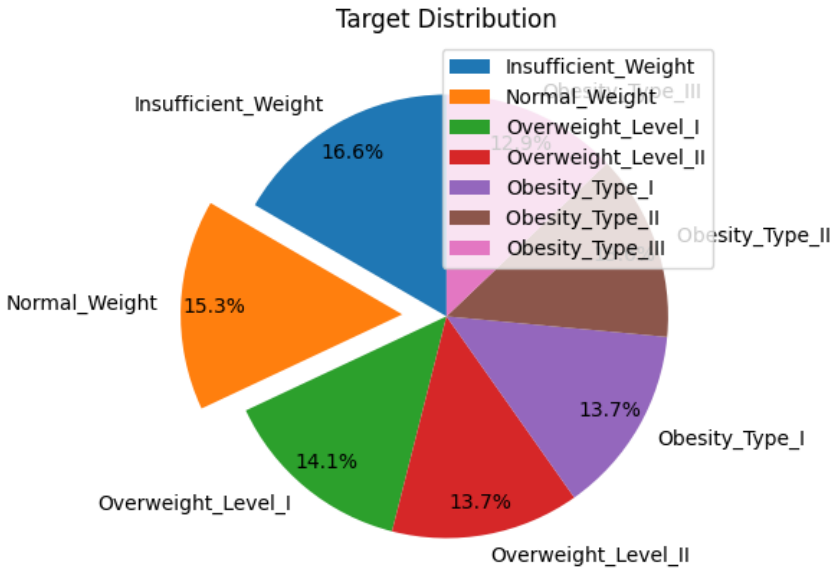






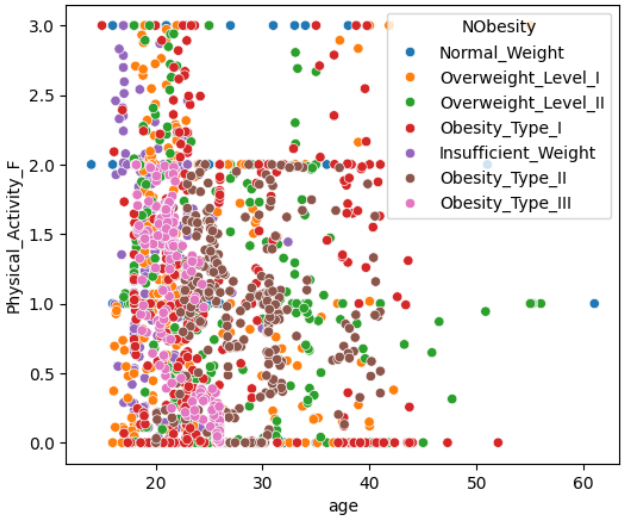
Target Distribution:

- Target class distribution showed the highest proportion in Overweight I.
- Noticeable presence of insufficient and normal weight categories.

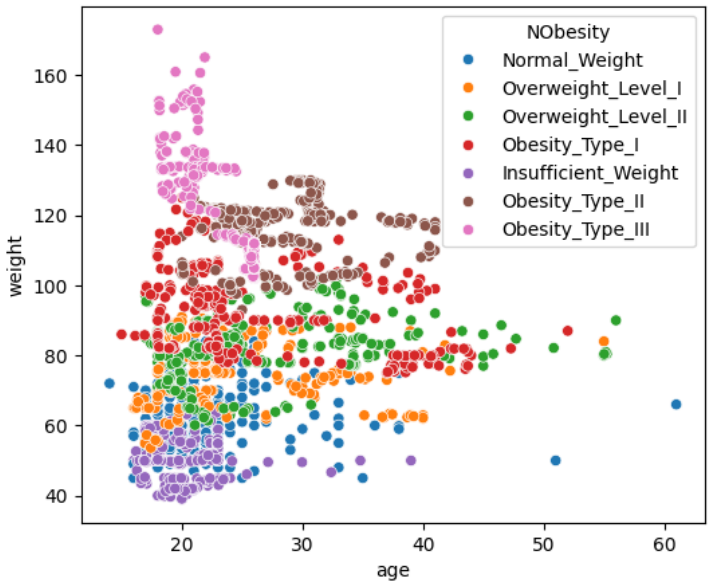


Scatter Plots:

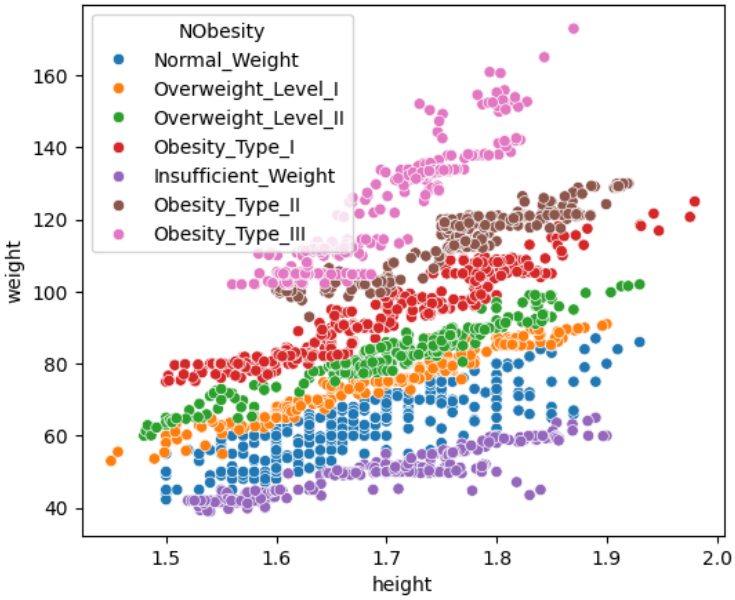
- **Age vs Physical Activity**:
  - Low physical activity corresponds to higher obesity levels.
  - Obesity Type III is common in the 15–30 age range.



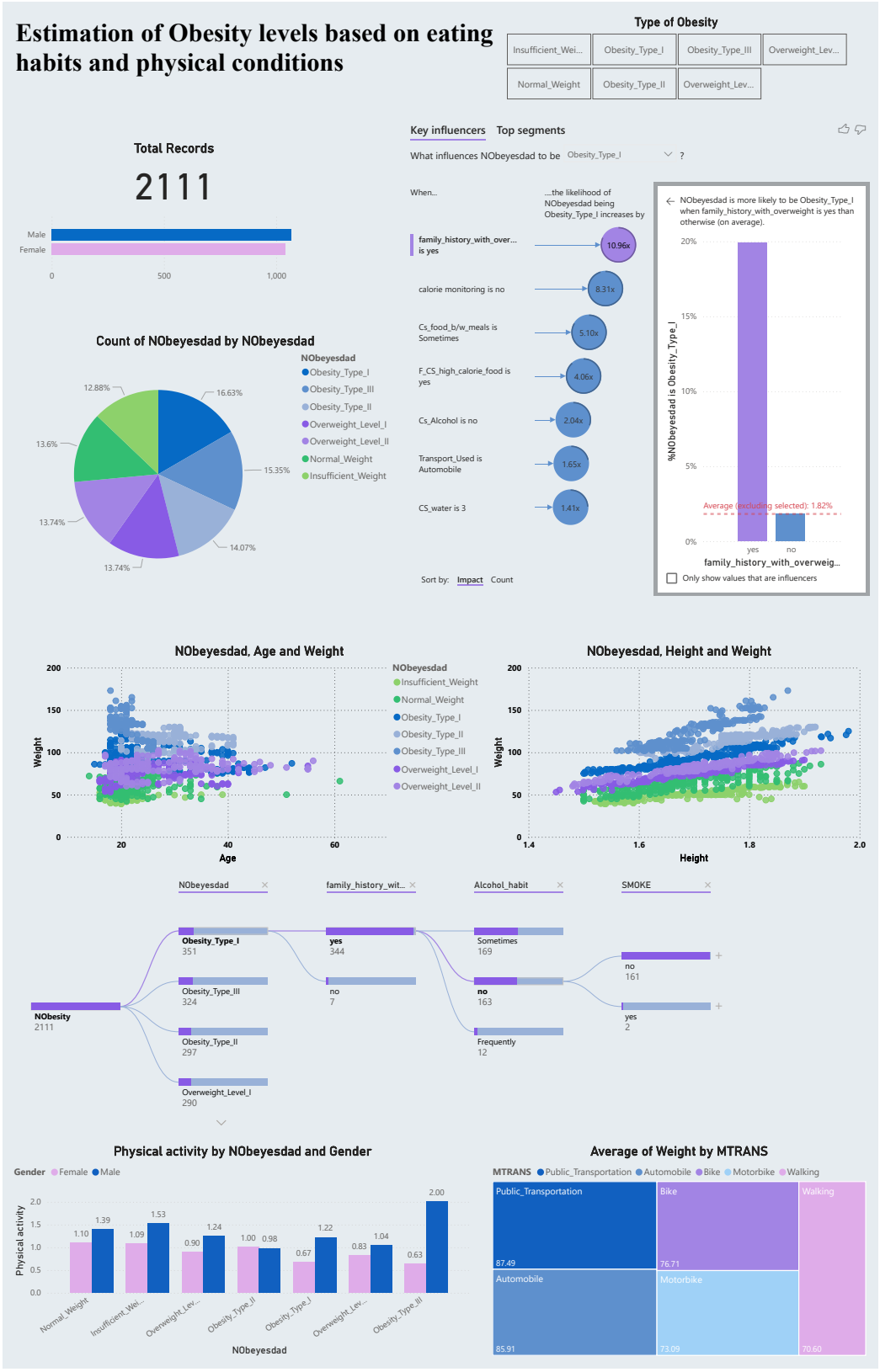
- **Age vs Weight**:
  - Subjects with obesity are mostly aged 15–25.



- **Weight vs Height**:
  - Obesity types show distinct separation in this plot, confirming consistent categorization.



Power Bi – Dashboard Creation:



7. Challenges and Resolutions

- While transforming the dataset, I encountered several syntax issues while using the Wrangler plugin. Referring to the Cloud Data Fusion documentation (wrangler-decimal-transformations) helped me resolve these challenges.
- Initial deployment faced repeated failures due to minor errors in configuration and transformation steps.
- Discovery of **GCP Log Explorer** played a key role in identifying the root causes and fixing the issues.
- The built-in **validation and data preview** in Cloud Data Fusion was instrumental in iterative error resolution.

8. Conclusion

This project successfully demonstrates the power and convenience of cloud-based ETL using GCP. It showcases how tools like Cloud Data Fusion, BigQuery, and Power BI can be orchestrated for end-to-end data analysis and visualization without requiring heavy code.

To finalize the analysis, a Power BI dashboard was created using the processed data from BigQuery. The dashboard includes:

- Visual representations of obesity level distributions,
- Age and weight-based comparisons,
- Key influences of target variable,
- Categorical breakdowns,
- Insights derived from physical activity, and family history.

These visualizations enhance the interpretability of the dataset, making it easier to communicate patterns and trends to non-technical stakeholders.

It also emphasizes the importance of logging, validation, and proper transformation in ensuring data readiness for further modeling or analytics.