Data Analytics Project Plan

This project plan outlines a structured approach to analyze the correlation between lifestyle factors and obesity levels, utilizing both Python for data manipulation and visualization, and SQL for data querying and aggregation.

# 1. Data Cleaning and Preparation with Python

- Load Dataset: Import the dataset using pandas.  
- Data Inspection: Check for missing values, data types, and summary statistics to understand the data.  
- Handle Missing Values: Decide on a strategy for dealing with missing data (e.g., imputation, removal).  
- Encode Categorical Variables: Convert categorical variables into a numerical format suitable for analysis, using methods like one-hot encoding or label encoding.  
- Normalize/Standardize Numerical Data: (Optional) Apply normalization or standardization to numerical columns if necessary, especially if you plan to perform any machine learning tasks.  
- Save Cleaned Dataset: Export the cleaned DataFrame to a new CSV file for SQL importation.

# 2. Importing Data into an SQL Database

- Create Database and Table: Use SQL commands to create a database and a table that matches the schema of your cleaned dataset.  
- Import Data: Import the cleaned CSV file into the SQL table. This can be done using SQL commands or through Python libraries like pandas with a connector like SQLAlchemy.

# 3. SQL Queries for Data Analysis

- Aggregate Queries: Write SQL queries to aggregate data on key metrics by obesity levels, such as average physical activity, average technology use time, and average water consumption.  
- Proportion Queries: Calculate the proportion of individuals preferring different modes of transportation and having a smoking habit across obesity levels.  
- Grouping and Ordering: Use GROUP BY and ORDER BY clauses to organize your results based on obesity levels and the metrics of interest.

# 4. Statistical Analysis and Visualization with Python

- Load SQL Query Results: Use Python to load the results of your SQL queries into pandas DataFrames.  
- Statistical Testing: Perform statistical tests (e.g., ANOVA, Chi-Square Test) to assess the significance of the differences observed in lifestyle factors across obesity levels.  
- Visualization: Create visualizations (using seaborn or matplotlib) to illustrate the distribution of key metrics across different obesity levels and to highlight significant findings.

# 5. Insights and Recommendations

- Interpret Results: Analyze the statistical and visual results to draw conclusions about the relationship between lifestyle factors and obesity levels.  
- Recommendations: Based on the analysis, propose recommendations for lifestyle changes that could potentially help in preventing or reducing obesity.

# Deliverables

- A Jupyter notebook containing the Python code for data cleaning, statistical analysis, and visualization.  
- SQL script files containing the queries used for data aggregation and analysis.  
- A report or presentation summarizing the findings, insights, and recommendations.