Household Power Project Documentation

**Team Member:**

* **Asmaa Hamdy Abdelaleem**
* **Fatma Ayoub Ahmed**
* **Salma Mahdy Mohamed**

**Project Overview**

**Title**: House Power

**Brief Description**:  
"House Power" is a data engineering and machine learning project that focuses on analyzing and predicting household energy consumption. It integrates large datasets from smart meters and IoT sensors, using advanced data pipelines and machine learning models to provide actionable insights for optimizing energy usage.

**Goals**:  
The primary goal of the "House Power" project is to create a system that enables homeowners and energy providers to:

* Understand energy usage patterns.
* Predict future energy consumption.
* Implement strategies for energy optimization and cost reduction.

**Key Features**:

**Data-Driven Insights**: Analyze energy patterns and detect inefficiencies.

* **Predictive Modeling**: Use machine learning algorithms to forecast energy usage and optimize future consumption.
* **Scalable Data Pipelines**: Handle large volumes of data efficiently with automated data engineering workflows.

**Technologies Used**:

* **Data Engineering**: SQL, and Python for handling data extraction, transformation, and loading (ETL).
* **Machine Learning**: Python libraries (scikit-learn) for building and training predictive models.
* **Data Storage**: PostgreSQL and MySQL for large-scale data management.
* **Visualization**: Matplotlib for visualizing energy trends and predictions.

**Impact**:  
By delivering accurate energy consumption predictions and analysis, "House Power" allows homeowners to reduce energy waste, lower utility bills, and contribute to a more sustainable environment. Energy providers benefit from improved load management and better customer insights.

**Introduction:**

**Problem Statement**

In today's world, energy efficiency is becoming increasingly crucial as the demand for power rises and energy costs fluctuate. Households often struggle to monitor and manage their energy consumption, leading to inefficiencies, higher bills, and excessive carbon footprints. With the growing availability of smart devices and sensors, large amounts of household energy data are being collected, yet many homeowners lack the tools to effectively analyze and predict their energy usage patterns.

**Solution**

The "House Power" project aims to address this challenge by leveraging data engineering and machine learning techniques to provide a comprehensive solution for analyzing and predicting household energy usage. The project uses a large dataset of energy consumption records collected from various sources, such as smart meters and IoT sensors. By building robust data pipelines and predictive models, "House Power" helps homeowners and energy providers make data-driven decisions to optimize energy use, reduce costs, and minimize environmental impact.

**Objectives**

-**Analyze**: The project focuses on transforming raw energy data into meaningful insights, highlighting trends, peak usage periods, and areas of inefficiency.

-**Predict**: Using machine learning models, "House Power" forecasts future energy consumption based on historical data, allowing users to plan for optimal energy usage.

-**Optimize**: With the predictions in hand, homeowners and energy providers can implement strategies to reduce waste, manage peak loads, and potentially lower energy costs.

**Target Audience**

"House Power" is designed for a wide range of users, including:

* **Homeowners**: Those interested in monitoring their energy consumption and reducing their bills.
* **Energy Providers**: Companies looking to analyze large-scale consumption data to improve energy distribution and load management.
* **Data Scientists**: Professionals in the energy sector who are exploring innovative ways to apply machine learning to real-world energy challenges.

**Data Engineering Workflow**

**Data Sources:**The raw data for the "House Power" project originated from three separate Excel (XLSX) files, each containing different aspects of household energy usage:

1-**Voltage Dataset**: Captured voltage measurements over time.

**Attribute Information:**

1-date: Date in format dd/mm/yyyy

2-time: time in format hh:mm:ss

3-voltage: minute-averaged voltage (in volt)

4-global\_intensity: household global minute-averaged current

intensity (in ampere)

2- **Power Dataset**: Recorded global active and reactive power consumption.

**Attribute Information:**

* 1. global\_active\_power: household global minute-averaged active

power (in kilowatt)

* 1. global\_reactive\_power: household global minute-averaged reactive power (in kilowatt)

3-**Sub-metering Dataset**: Monitored energy consumption for specific household appliances

**Attribute Information:**

1. global\_active\_power: household global minute-averaged active

power (in kilowatt)

2-sub\_metering\_1: energy sub-metering No. 1 (in watt-hour of active energy). It corresponds to the kitchen, containing mainly a dishwasher, an oven and a microwave (hot plates are not electric but gas powered).

3-sub\_metering\_2: energy sub-metering No. 2 (in watt-hour of active energy). It corresponds to the laundry room, containing a washing-machine, a tumble-drier, a refrigerator and a light.

4-sub\_metering\_3: energy sub-metering No. 3 (in watt-hour of active energy). It corresponds to an electric water-heater and an air-conditioner.

To streamline data processing, the files were converted into CSV format using Python, which allowed for easier manipulation in later steps.

**Data Extraction:** After converting the XLSX files to CSV, the data was extracted into Pandas DataFrames.

**Data Cleaning**: The raw data required some preprocessing to ensure consistency and correctness: Time Column Transformation

* The time column in the Voltage dataset was initially in string format. Using a Python lambda function, it was converted into a proper datetime format for easier time-based analysis
* Handling Missing Values: Any missing or incorrect data entries were cleaned during this step to ensure the data was ready for analysis.

**Data Transformation:** To make the datasets more informative and ready for analysis, several transformations were performed:

* Column Renaming: To maintain clarity and uniformity, the column names were updated to include the units of measurement

**Feature Engineering:**One of the critical transformations involved adding a new column to the power dataset:

* Apparent Power Calculation: The Apparent\_power\_KVA column was derived using the formula for apparent power in kilovolt-amperes (KVA). This was calculated by combining the global active and reactive power

**Data Analysis Using Python and SQLAlchemy**

**Objective**

* The goal of this analysis is to extract meaningful insights from household energy consumption data by performing database queries using SQLAlchemy. These insights focus on understanding patterns in energy usage, peak consumption times, and overall energy consumption during specified periods.

**Database Setup**

* We used SQLAlchemy to connect to the PostgreSQL database containing the processed data. The data includes key metrics such as Global\_active\_power measured in kilowatts (KW)

**Analysis Queries and Insights**

**1. What is the Average Global Active Power Consumption per Day?**

* **Purpose**: This query calculates the daily average global active power consumption. It helps identify typical usage patterns for each day.
* **Insight**: This provides a general idea of household power usage trends on a daily basis.

**2. What is the Total Energy Consumed for Each Specified Time Period?**

* **Purpose**: This query sums the total energy consumed within a defined time period (e.g., daily, weekly, or monthly). It’s useful for understanding the energy usage over time.
* **Insight**: By specifying different date ranges, we can analyze energy consumption trends during particular periods.

**3. What is the Average Energy Consumption per Hour?**

* **Purpose**: This query calculates the average energy consumption per hour to capture usage patterns throughout the day.
* **Insight**: Understanding hourly energy consumption patterns allows for analysis of daily fluctuations and can help identify peak hours.

**4. What is the Maximum Energy Consumption Recorded During a Day?**

* **Purpose**: This query finds the maximum energy consumption recorded on each day. It’s important for detecting energy usage spikes and potential overuse during certain periods.
* **Insight**: Identifying maximum consumption per day can help in understanding when energy usage reaches its highest point, allowing for better load management

**5. What is the Energy Consumption During Peak Hours (6 PM to 9 PM)?**

* **Purpose**: This query isolates energy consumption during peak hours (6 PM to 9 PM), which are typically high usage periods for households.
* **Insight**: Peak hour consumption analysis helps in understanding how much energy is used during the evening when household activity tends to be highest

**Machine Learning Pipeline**

**Data Preprocessing**: Before training the machine learning model, several preprocessing steps were taken to prepare the data

* **Dropping Date and Time Columns**: Since the model focused on predicting active power (KW) and did not require time-specific analysis, the date and time columns were dropped from the dataset to simplify the features.
* **Feature Scaling:** To ensure that all features contributed equally to the model and to prevent bias from features with larger ranges, a MinMax scaler was applied. This scaled the feature values to a range between 0 and 1

**Model Selection**: Chose Linear Regression to predict Global\_active\_power\_KW.

-To ensure robust model evaluation and prevent overfitting, I applied 5-fold Cross-Validation to validate the performance across different data subsets. Cross-validation divides the dataset into five subsets and trains the model five times, each time using a different subset for validation

**Model Evaluation:** The model was evaluated using the R² score, which measures how well the regression predictions match the actual data. A score close to 1.0 indicates a good fit

- Applied 5-fold cross-validation, with the average R² score across the folds being 99.9%, confirming the model’s accuracy and consistency

**Visualization :**After training and evaluating the model, visualizations were generated to better understand the model's performance and the data's behavior

**System Architecture Tools & Technologies Overview**

**Tools & Technologies Overview:**

* **Python**: Used for data extraction, transformation, and loading (ETL) processes, as well as for developing and training the machine learning model.
* **SQLAlchemy:** A Python library used to connect to and interact with the database, allowing seamless integration between the Python scripts and the database.
* **PostgreSQL:** Chosen as the data warehouse to store and manage large volumes of household energy data. It provided robust support for handling complex queries and large-scale data.
* **Microsoft SSIS:** In addition to Python, Microsoft SQL Server Integration Services (SSIS) was used to create an alternative ETL package for the project. This package performed similar data extraction and transformation steps and loaded the data into a MySQL database**.**
* **MySQL**: While PostgreSQL was the primary data warehouse, MySQL was used to store the data from the SSIS ETL package, offering flexibility in how the data was stored and accessed

**System Architecture**

The overall architecture of the project was structured into key components:

1. **Data Ingestion:** The raw data, originally in Excel files, was converted to CSV and ingested into the system using Python and SQLAlchemy, which connected to a PostgreSQL database.
2. **Data Transformation & Feature Engineering:** Data cleaning, type transformations, and feature engineering (e.g., calculating Apparent\_power\_KVA) were performed using Python scripts
3. **Machine Learning Pipeline**: The prepared data was then used to train a machine learning model (Linear Regression) for predicting household power consumption.
4. **ETL Package (SSIS):** A separate ETL pipeline was created using SSIS, which mirrored the Python-based ETL workflow and stored data in a MySQL database
5. **Data Warehousing & Storage:** The PostgreSQL database served as the main data warehouse, housing all cleaned and transformed data for downstream analysis and machine learning tasks.

This architecture ensured flexibility, scalability, and efficient management of the large dataset, allowing for robust analysis and predictive modeling.

**Results and Performance**

The primary objective of the "House Power" project was to develop a machine learning model capable of accurately predicting household power consumption. After data preprocessing and model training, the Linear Regression model performed exceptionally well, yielding the following results: **Model Performance**

* R² Score: The model achieved an impressive R² score of 99.9%, indicating that the model's predictions were almost perfectly aligned with the actual power consumption data. This high accuracy demonstrates the effectiveness of the linear regression model for predicting energy usage in households.

**Key Insights**

* The model was able to capture the relationships between active power and other variables in the dataset with great precision.
* including the addition of the Apparent\_power\_KVA column, played a significant role in improving the model's predictive accuracy.

These results confirm that the "House Power" system can reliably be used to forecast household energy consumption, enabling users to optimize their power usage and reduce inefficiencies

**Conclusion**

* The "House Power" project successfully achieved its objective of analyzing and predicting household energy consumption through a well-defined data engineering and machine learning pipeline. The data, initially sourced from three Excel files, was processed and transformed using Python. Key transformations included converting time columns, renaming columns

with measurement units, and feature engineering to create the Apparent\_power\_KVA column

* The machine learning model chosen for prediction, a linear regression model, demonstrated outstanding performance with an R² score of 99.9%, indicating highly accurate predictions. The system architecture employed a combination of Python for ETL and machine learning, SQLAlchemy for database connections, PostgreSQL as the primary data warehouse, and Microsoft SSIS for creating an additional ETL pipeline loaded into a MySQL database
* Overall, the project showcased a robust and scalable workflow for processing large datasets and building accurate predictive models, providing actionable insights into energy usage.