# Predicting Solar Energy Production

This project aims to predict solar energy production using machine learning models based on various factors. Solar energy is a key renewable resource for sustainable power generation. The goal is to optimize solar power generation, improve grid efficiency, and support better decision-making in solar energy investments.

by Rishita Shah

# The Importance of Solar Energy

### Objective :

The main objective of predicting solar energy production is to **enhance energy efficiency, optimize resource planning, and improve grid integration** using data-driven insights.

### Dataset :

The dataset contains **218,115** records with **17 columns** .Key observations:

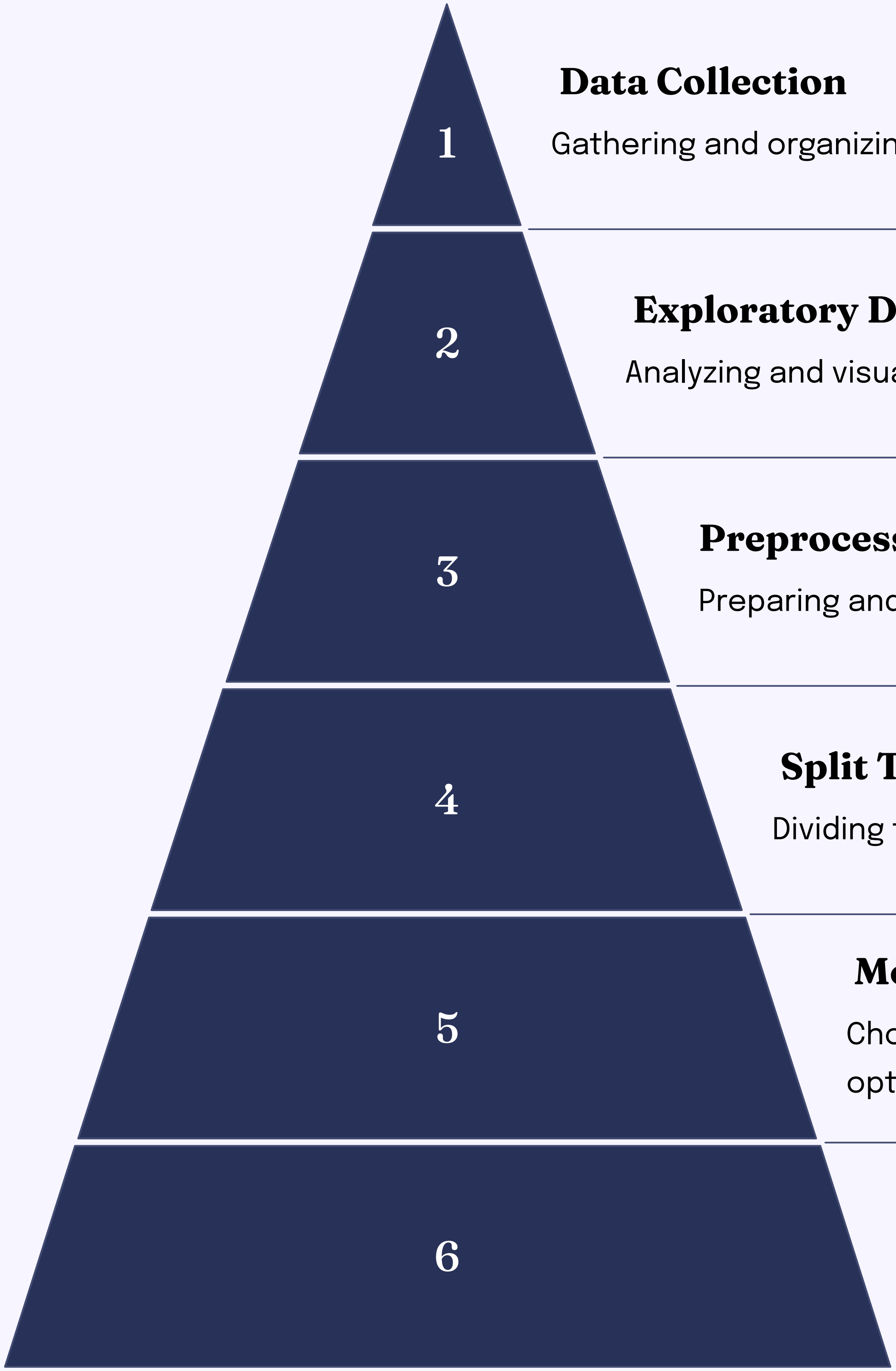
✅ **Target Variable:**

- `Estimated Annual PV Energy Production (kWh)` → This is the **dependent variable** we aim to predict.

### Importance :

- Improves energy efficiency and grid stability.
- Enhances decision-making for power distribution.
- Helps in cost reduction and resource optimization.

# Steps :



### Data Collection

Gathering and organizing data to train machine learning models.

### Exploratory Data Analysis

Analyzing and visualizing data patterns to understand its characteristics.

### Preprocessing

Preparing and cleaning data to enhance its quality and suitability.

### Split Train and Test Data

Dividing the dataset into training and testing sets to evaluate.

### Model Selection and Model Training

Choosing a suitable machine learning algorithm and optimizing its parameters.

### Model Evaluation

Assessing the performance of machine learning model using metrics.

# Exploratory Data Analysis

1 **Descriptive Statistics :**

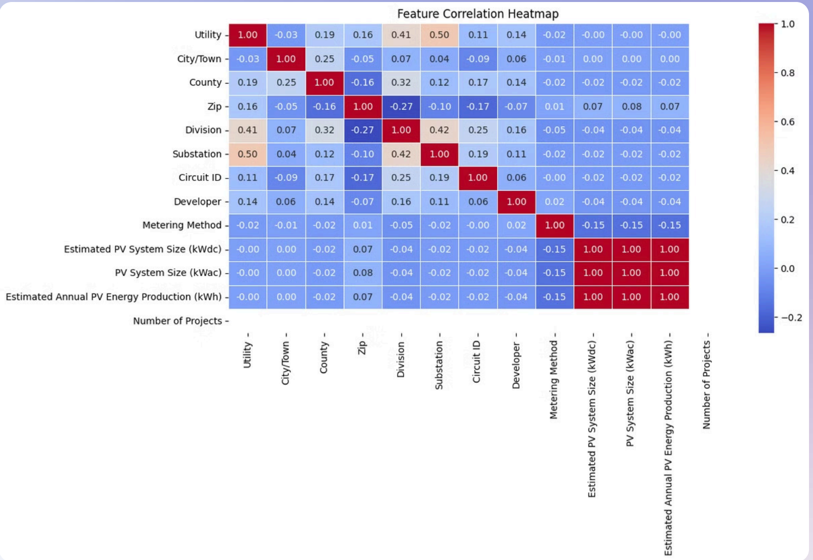
This step involves calculating basic statistics like mean, median, and standard deviation of key features, to understand data characteristics.

2 **Correlation Analysis :**

Examining correlations between features reveals potential relationships and helps identify which features are most relevant for segmentation.

3 **Visualizations :**

Creating visualizations like histograms, scatter plots, and box plots helps understand data patterns and outliers in the data.



| | Utility | City/Town | County | Zip | Division | Substation | Circuit ID | Developer | Metering Method | Estimated PV System Size (kWdc) | PV System Size (kWac) | Estimated Annual PV Energy Production (kWh) | Number of Projects |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Utility | 1.00 | -0.03 | 0.19 | 0.16 | 0.41 | 0.50 | 0.11 | 0.14 | -0.02 | -0.00 | -0.00 | -0.00 | |
| City/Town | -0.03 | 1.00 | 0.25 | -0.05 | 0.07 | 0.04 | -0.09 | 0.06 | -0.01 | 0.00 | 0.00 | 0.00 | |
| County | 0.19 | 0.25 | 1.00 | -0.16 | 0.32 | 0.12 | 0.17 | 0.14 | -0.02 | -0.02 | -0.02 | -0.02 | |
| Zip | 0.16 | -0.05 | -0.16 | 1.00 | -0.27 | -0.10 | -0.17 | -0.07 | 0.01 | 0.07 | 0.08 | 0.07 | |
| Division | 0.41 | 0.07 | 0.32 | -0.27 | 1.00 | 0.42 | 0.25 | 0.16 | -0.05 | -0.04 | -0.04 | -0.04 | |
| Substation | 0.50 | 0.04 | 0.12 | -0.10 | 0.42 | 1.00 | 0.19 | 0.11 | -0.02 | -0.02 | -0.02 | -0.02 | |
| Circuit ID | 0.11 | -0.09 | 0.17 | -0.17 | 0.25 | 0.19 | 1.00 | 0.06 | -0.00 | -0.02 | -0.02 | -0.02 | |
| Developer | 0.14 | 0.06 | 0.14 | -0.07 | 0.16 | 0.11 | 0.06 | 1.00 | 0.02 | -0.04 | -0.04 | -0.04 | |
| Metering Method | -0.02 | -0.01 | -0.02 | 0.01 | -0.05 | -0.02 | -0.00 | 0.02 | 1.00 | -0.15 | -0.15 | -0.15 | |
| Estimated PV System Size (kWdc) | -0.00 | 0.00 | -0.02 | 0.07 | -0.04 | -0.02 | -0.02 | -0.04 | -0.15 | 1.00 | 1.00 | 1.00 | |
| PV System Size (kWac) | -0.00 | 0.00 | -0.02 | 0.08 | -0.04 | -0.02 | -0.02 | -0.04 | -0.15 | 1.00 | 1.00 | 1.00 | |
| Estimated Annual PV Energy Production (kWh) | -0.00 | 0.00 | -0.02 | 0.07 | -0.04 | -0.02 | -0.02 | -0.04 | -0.15 | 1.00 | 1.00 | 1.00 | |
| Number of Projects | | | | | | | | | | | | | |

# Feature Correlation Heatmap.

- The color bar on the right shows the correlation scale from **-1 (blue, strong negative correlation)** to **1 (red, strong positive correlation)**.
- The diagonal values are all **1.00** (deep red) because each variable is perfectly correlated with itself.
- **Estimated PV System Size (kWdc), PV System Size (kWac), and Estimated Annual PV Energy Production (kWh)** are **highly correlated** (values close to 1), which makes sense as larger PV systems generate more energy.
- **Substation & Division** show a moderate positive correlation (0.42), indicating some level of dependency.
- **Zip Code & Division have a weak negative correlation (-0.27)**, meaning geographical areas may have different energy profiles.

# Histogram for Numerical Variables.

- These histograms help in understanding how each numerical variable is distributed.
- It also provides insights into possible **outliers, skewness, and data concentration**.

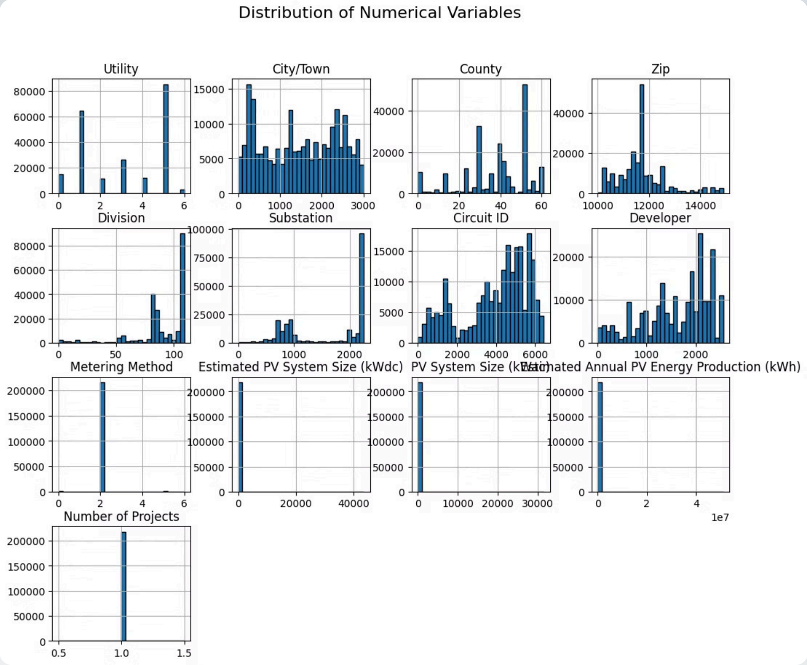
## Analysis of Specific Variables:

### 1. Categorical-Like Numerical Variables:

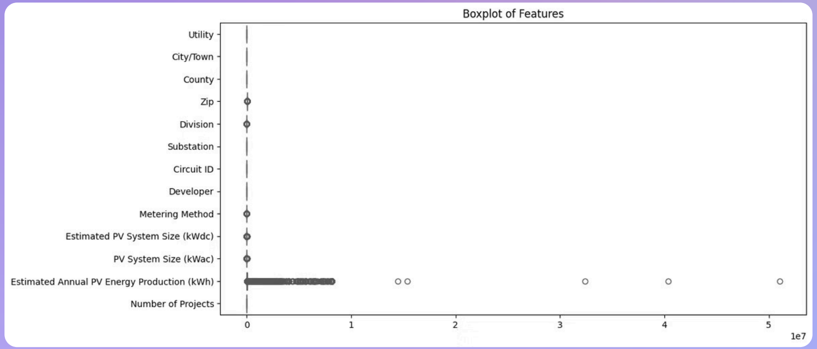
- Some variables like Utility, Division, Metering Method, and Number of Projects have very few unique values, suggesting they are categorical (even though they are represented numerically).
- Metering Method and Number of Projects seem to have only one or very few distinct values, indicating low variance.

### 2. Geographical Variables:

- Variables like City/Town, County, Zip, and Substation show a **multi-modal** distribution, meaning multiple peaks exist. This makes sense, as different locations have different frequencies of solar energy projects.



# Boxplot to Detect Outliers.

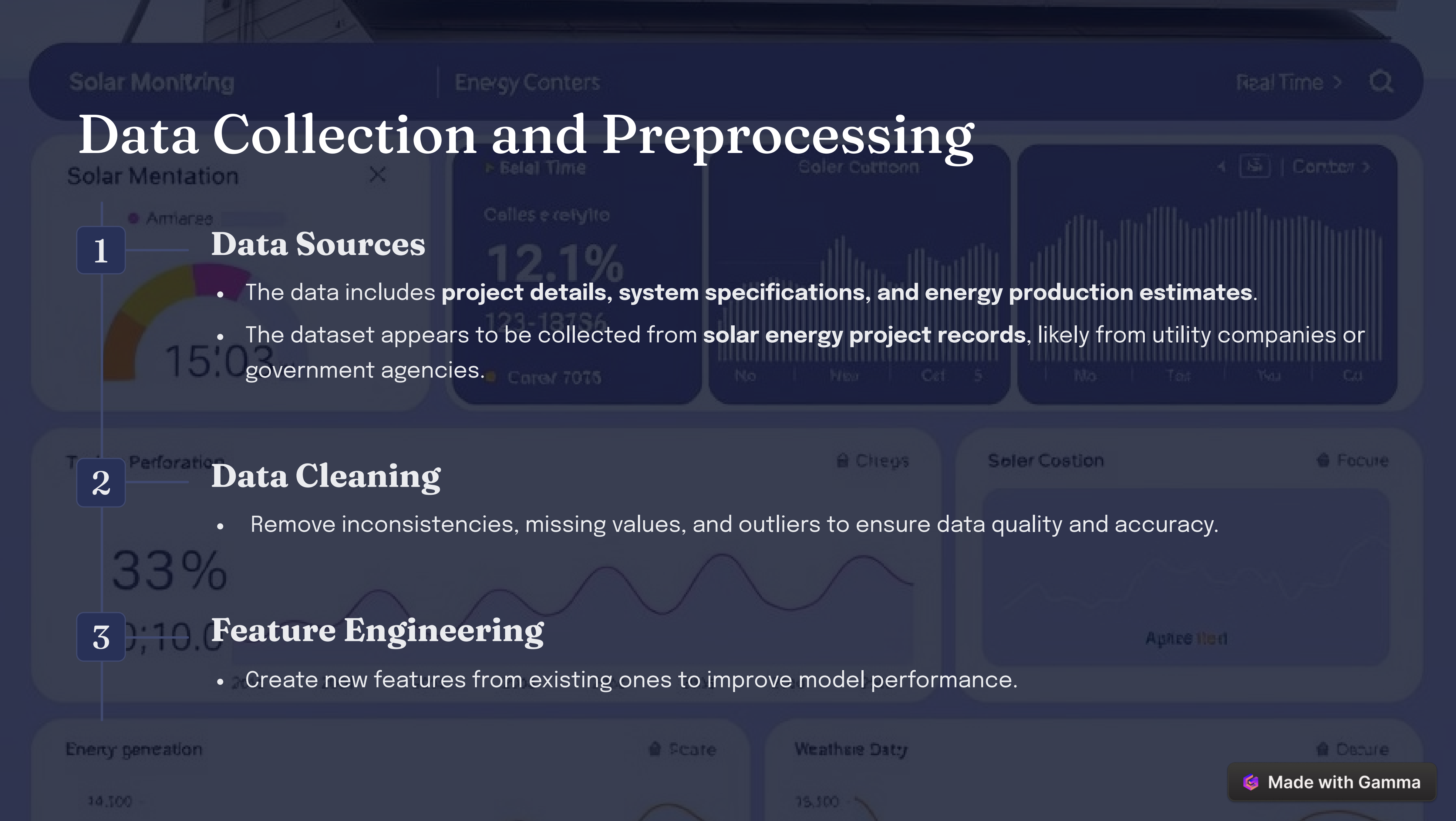


## Purpose:

- This boxplot helps visualize the distribution and detect outliers in numerical variables.

## Outliers Observed:

- The **Estimated Annual PV Energy Production (kWh)** has extreme outliers, suggesting the presence of very high-value solar energy projects.
- Other features like **PV System Size (kWdc, kWac)** also exhibit some outliers.
- **Categorical-like variables** (e.g., Utility, Division, and Substation) show small variances but some dispersed values.



# Data Collection and Preprocessing

1

## Data Sources

- The data includes **project details, system specifications, and energy production estimates**.
- The dataset appears to be collected from **solar energy project records**, likely from utility companies or government agencies.

2

## Data Cleaning

- Remove inconsistencies, missing values, and outliers to ensure data quality and accuracy.

3

## Feature Engineering

- Create new features from existing ones to improve model performance.

# The Future of Solar Energy Prediction

Advancements in technology, AI, and data analysis are enabling more accurate and sophisticated solar energy prediction models. This will lead to a more reliable and efficient solar energy grid.

# Thank You !!