# General introduction

This dataset contains information about solar power plants and their various geographic and technical attributes. The dataset supports energy planning, renewable energy development, and spatial analysis, providing insights into the distribution and characteristics of solar projects in California. It is particularly valuable for utility companies, policy analysts, and researchers focused on renewable energy infrastructure.The dataset consists of 5397 rows and 20 columns (excluding the OBJECTID index column). We have missing values on four columns, especially related to unique identifiers (HIFLD IDs) for certain substations (e.g., HIFLD ID (GTET 100 Max Voltage), HIFLD ID (GTET 200 Max Voltage) and HIFLD ID (CAISO)). The following columns are given:

1. **County**: The name of the Californian county where the data point is located, represented as a string.
2. **Acres**: The area in acres (1 ac = 4046.8564224 m2) associated with each data point, represented as a float.
3. **Install Type**: The type of solar installation (modes: 'Rooftop', 'Parking', 'Ground'), represented as a string.
4. **Urban or Rural**: A classification of whether the location is in an urban or rural area (modes: 'Urban', 'Rural'), represented as a string.
5. **Combined Class**: A combination of the install type and the column “Urban or Rural” (modes: 'Rooftop - Urban', 'Parking - Urban', 'Ground - Urban', 'Ground - Rural', 'Rooftop - Rural', 'Parking - Rural'), represented as a string.
6. **Distance to Substation (Miles) GTET 100 Max Voltage**: The distance from the data point to a substation with a GTET 100 Max Voltage, in miles (1 mi = 1,609.344 m), represented as a float.
7. **Percentile (GTET 100 Max Voltage)**: The percentile ranking related to the GTET 100 Max Voltage distance, represented as a string.
8. **Substation Name GTET 100 Max Voltage**: The name of the substation with a GTET 100 Max Voltage, represented as a string.
9. **HIFLD ID (GTET 100 Max Voltage)**: A unique identifier for the GTET 100 Max Voltage substation, represented as a float, though some values are missing.
10. **Distance to Substation (Miles) GTET 200 Max Voltage**: The distance from the data point to a substation with a GTET 200 Max Voltage, in miles (1 mi = 1,609.344 m), represented as a float.
11. **Percentile (GTET 200 Max Voltage)**: The percentile ranking related to the GTET 200 Max Voltage distance, represented as a string.
12. **Substation Name GTET 200 Max Voltage**: The name of the substation with a GTET 200 Max Voltage, represented as a string.
13. **HIFLD ID (GTET 200 Max Voltage)**: A unique identifier for the GTET 200 Max Voltage substation, represented as a float, though some values are missing.
14. **Distance to Substation (Miles) CAISO**: The distance from the data point to a CAISO substation, in miles (1 mi = 1,609.344 m), represented as a float.
15. **Percentile (CAISO)**: The percentile ranking related to the CAISO substation distance, represented as a string.
16. **Substation CASIO Name**: The name of the CAISO substation, represented as a string, though a few values are missing.
17. **HIFLD ID (CAISO)**: A unique identifier for the CAISO substation, represented as a float, though some values are missing.
18. **Solar Technoeconomic Intersection**: A string indicating the technoeconomic intersection related to solar energy (modes: 'Within', 'Outside'), referring to areas with high or low solar potential or feasibility.
19. **Shape\_\_Area**: The area of the geographical shape in squaremetres, represented as a float.
20. **Shape\_\_Length**: The length of the geographical shape, represented as a float.

**Key Features:**

* Location information, including county and urban/rural classification.
* Proximity to high-voltage substations (≥100 kV and ≥200 kV) and CAISO substations.
* Solar installation type (e.g., rooftop) and area in acres.
* Percentile rankings for distances to substations.
* Spatial details, including shape area and length for GIS applications.

**Target:**

The **Solar Technoeconomic Intersection** column suggests that this data may be used for analysis related to solar energy generation. This column could be used as target variable because it classifies the area’s suitability for solar technology.

This dataset could be valuable for analyzing energy infrastructure in relation to geographic regions, urban/rural areas, and distances to various types of substations, potentially useful for planning, optimization, and analysis in energy systems or grid management.

**Source:**

1. Data: <https://www.kaggle.com/datasets/vijayveersingh/california-sunny-spaces>
2. Data visualized with interactive map: <https://gis.data.cnra.ca.gov/datasets/CAEnergy::solar-footprints-in-california/about>

# Overview and Preliminary Data Exploration

The dataset under consideration provides crucial insights into the characteristics of solar power plants across California, focusing on key geographic and technical attributes. The data spans multiple features, including the type of solar installations, proximity to various substations, and the feasibility of solar energy projects in different areas based on their technoeconomic potential. By analyzing this data, the project aims to shed light on the distribution of solar energy infrastructure and assist in the optimization and development of energy systems.

Initial data processing involved cleaning and exploring the dataset using various Python libraries, with a focus on handling missing values, managing categorical and numerical variables, and preparing the dataset for deeper analysis. The process of understanding the structure and quality of the data is crucial for any further exploration and modeling.

## Data Cleaning and Preprocessing

The first step in the data analysis was examining the basic structure and attributes of the dataset. The dataset contains 5397 rows and 20 columns, with several key attributes that provide insights into the solar energy infrastructure in California. Some of the notable features include Install Type (such as 'Rooftop', 'Parking', 'Ground'), Urban or Rural classification, and distance-related attributes to substations.

A detailed review of the dataset revealed some important issues:

1. **Missing Values**: Four columns related to substation identifiers (e.g. HIFLD ID) had missing values. The share of missing values was below 10% for those columns. Becase of the lack of informative value for the modeling these columns were dropped.
2. **Data Redundancy**: Some columns, such as the Shape\_\_Area column, were redundant as they provided the same information as other columns (e.g., Acres), but in different units. These redundant columns were dropped to streamline the dataset.
3. **Long Column Names**: To simplify data processing and visualization, long column names, particularly those related to substations, were shortened using an abbreviation dictionary. For example, columns like "Distance to Substation (Miles) GTET 100 Max Voltage" were renamed to "Distance to GTET 100" for ease of access and clarity.
4. **Additional columns**: Because of the geographical nature of the dataset it was plausible to add more geographical information like latitude and longitude as well as the population density as columns to the dataset. This was achievd by using the geopandas library and open source internet resources.

Finally, the target variable “Solar Technoeconomic Intersection” was converted to binary values (1 for "Within" and 0 for "Outside") to facilitate subsequent statistical and machine learning analysis. This step ensured that the data was in an appropriate format for modeling.

## Exploratory Data Analysis (EDA)

**Summary Statistics**

After cleaning the dataset, an initial exploration of the numerical features was performed. Summary statistics were generated to better understand the distribution and spread of key attributes like Acres, Distance to Substations, and pop\_density. These statistics provided a useful overview of the data's central tendencies and variability, highlighting any outliers or skewed distributions.

**Missing Values Analysis**

The missing values across the dataset were systematically analyzed to identify the proportions of missing data in each column. Some columns, especially those related to substation identifiers, exhibited significant missing data. This is important because these missing values could affect downstream analyses, such as any spatial or proximity-based assessments related to the distribution of substations and solar installations.

**Duplicate Data Detection**

The dataset was also examined for duplicated entries. It was confirmed that there were no significant duplicates present, which ensures the integrity of the dataset for further analysis.

**Categorical Variables Analysis**

A categorical feature analysis was performed for columns such as Install Type, Urban or Rural, and Combined Class. The unique values of these features were reviewed to gain insights into the distribution of solar installations across urban and rural areas, as well as to determine the relative frequency of different installation types (rooftop, parking, or ground).

**Correlation Analysis**

A correlation matrix was computed to examine the relationships between numerical features. This step was instrumental in identifying potential relationships between key attributes, such as the correlation between the distance to substations and other geographic factors.

### Data Visualization

Key visualizations were created to further explore and communicate the data’s insights.

**Correlation Heatmap**

A heatmap visualization was generated to represent these correlations visually, allowing for a clear identification of positive or negative relationships. The heatmap indicated interesting trends, such as the inverse relationship between the distance to substations and the potential solar technoeconomic viability of different locations. This provides an initial indication that areas closer to substations may have better feasibility for solar installations, due to easier grid access. The strongest correlation with the target variable are given with “pop\_density” (0.27) and “Distance to GTET 200” (-0.27).

This visualization also helped identify key variables that might influence the location and type of solar installations, such as proximity to substations or the size of the area. The distance features (to substation GTET 100, GTET 200 and CAISO) show some moderate to weak correlations with each other, with the strongest being between the distance to GTET 100 and to GTET 200 (0.70). The geographic features in the dataset like longitude and latitude are very strongly negatively correlated with each other which is because of the elongated, rectangular shape of the state of California which stretches from north west to south east. The distance measures like GTET and CAISO have a weak correlation whereas using domain knowledge (i.e. here the "business" perspective) it is understandable that the distances of the solar power installations to the different substation types indicate that installations closer to one GTET location tend to be closer to the other as well.

**Boxplot Matrix**

A boxplot matrix was created to compare the distances to various substations (GTET 100, GTET 200, and CAISO) by installation type. This boxplot provided an intuitive view of how different installation types are distributed in terms of their proximity to high-voltage substations. The results showed that Rooftop installations tend to be closer to substations compared to Ground installations, which typically have more variable distances (outliers).

**Geospatial Analysis**

One of the key aspects of this dataset is the geographic distribution of solar installations across California counties. To support this, the geolocation of each county was retrieved using the geopy library. While the geocoding process encountered some rate-limiting issues due to the number of requests, it was planned that this step would allow for the addition of latitude and longitude information for further spatial analysis. These visualizations allowed for a better understanding of the geographic patterns of solar installations in California and their proximity to critical energy infrastructure.

## Statistical Analysis and Tests

We explored several relationships between the target variable ("Solar Technoeconomic Intersection" – a binary variable) and other features in the dataset, using appropriate statistical tests. The goal was to assess if there were significant associations between the target and other variables, including continuous, categorical, and binary features.

1. Point-Biserial Correlation: Relationship between Pop Density and Solar Technoeconomic Intersection

Since "Pop Density" is a continuous variable and "Solar Technoeconomic Intersection" is binary (0 or 1), we used Point-Biserial Correlation to assess the strength and direction of the association between these variables. Before proceeding with the correlation, we verified whether "Pop Density" follows a normal distribution using the Shapiro-Wilk test for normality. The result showed that "Pop Density" is not normally distributed (p-value < 0.05), which led us to apply a non-parametric test.

The Point-Biserial correlation(essentially a special case of Pearson’s correlation) is a suitable test for measuring the strength and direction of the association between a binary variable and a continuous variable. The Point-Biserial Correlation was computed, with a correlation coefficient and p-value indicating the strength of the relationship. If the p-value is below the significance threshold (0.05), it would imply a significant relationship. The result showed that there was no significant correlation between "Pop Density" and the target variable (p-value > 0.05).

2. Point-Biserial Correlation: Relationship between Distance to GTET 200 and Solar Technoeconomic Intersection

Next, we assessed the relationship between "Distance to GTET 200" (a continuous variable) and the binary target variable. Similarly to the previous analysis, we used the Shapiro-Wilk test to check for normality. The results indicated that "Distance to GTET 200" was not normally distributed. Consequently, we proceeded with the Point-Biserial Correlation, yielding a p-value greater than 0.05. This result suggested that there is no significant relationship between the distance to the GTET 200 and the Solar Technoeconomic Intersection.

3. Chi-Square Test of Independence: Relationship between Install Type and Solar Technoeconomic Intersection

One can use also the Chi-Square test of independence to check if there's a relationship between the categorical variables and a binary target. While the correlation coefficient is often used to measure the strength of association, the Chi-Square test helps testing whether the proportions of a binary target differ by category (e.g., do "Urban" and "Rural" categories have different proportions of target=1 or target=0).

Thus, for categorical variables, we used the Chi-Square Test of Independence to test if there was a significant association between "Install Type" (a categorical variable) and "Solar Technoeconomic Intersection." The contingency table for the two variables was created, followed by the Chi-Square test. The test statistic and p-value were computed to determine if the variables were independent or associated. The p-value from this test (which was below 0.05) indicated a significant relationship between "Install Type" and the target variable.

4. Chi-Square Test of Independence: Relationship between Urban or Rural and Solar Technoeconomic Intersection

We applied the Chi-Square Test of Independence once again, this time to explore the relationship between "Urban or Rural" (a categorical variable) and "Solar Technoeconomic Intersection." A contingency table was constructed for these variables, and the Chi-Square test was performed. The p-value of the test suggested a significant relationship between "Urban or Rural" and the target variable (p-value < 0.05), meaning that the classification of "Urban or Rural" could influence the target variable.

**Visualizing Relationships**

We used Violin Plots to visualize the distribution of continuous variables (e.g., "Pop Density" and "Distance to GTET 200") against the binary target variable ("Solar Technoeconomic Intersection"). Violin plots are particularly useful as they display the distribution, central tendency, and range of the data, while also visualizing the presence of outliers.

For instance, we observed the distribution of "Pop Density" and "Distance to GTET 200" against "Solar Technoeconomic Intersection" in separate plots. These plots provided insight into how the values of continuous variables differ across the binary groups (0 or 1).

**Conclusion of the statistical analysis**

Through these statistical tests, we were able to evaluate key relationships between the target variable and other variables in the dataset. While the Point-Biserial Correlation tests showed no significant relationships for the continuous variables, the Chi-Square tests revealed significant associations between the categorical variables ("Install Type" and "Urban or Rural") and the target. The visualizations further aided in understanding these relationships.

### Interim conclusion

As interim conclusion, the preliminary steps of data cleaning, exploration, and visualization have laid a solid foundation for the project's next stages. The dataset's rich geographic and technical details offer a promising opportunity to advance the analysis of solar energy projects, with significant implications for energy infrastructure planning and renewable energy development in California.