Data Analysis Project Report

Team: I love data analysis
Peter Felber & Andreas Heindl & Jakob Hütter

January 9, 2025

1 Contributions

The following contributions were made by each team member:

- Peter Felber:
  - Data preprocessing tasks
  - Initial visualization development
- Andreas Heindl:
  - Statistical analysis implementation
  - Regression analysis
- Jakob Hütter:
  - Advanced visualizations
  - Report writing and documentation

2 Dataset Description

- Dataset name and source: Solar Power Generation Data by Ani Kannal from Kaggle
- Time period and sampling frequency: data has been collected over a period of 34 days with a sampling frequency of 15 minutes
- Key variables analyzed: DC_Power, AC_Power, Ambient Temperature, Module Temperature, Irradiation
- Basic statistical properties:
  - Number of observations: 3134
  - Missing values: 130 (should be 34 days * 24 hours * 4 observations per hour = 3264)
  - Key statistics of cleaned dataset:

Variable	Mean	Median	Min	Max	Std
DC_Power in MW	67.540	8.632	0.000	269.097	85.798
AC_Power in MW	66.060	8.344	0.000	262.392	83.858
Ambient Temperature in °C	25.5	24.7	21.1	33.8	3.3
Module Temperature in °C	31.1	24.8	19.2	60.3	12.1
Irradiation kW/m^2	0.2273	0.0289	0.0000	0.999	0.2950

3 Methods and Analysis

3.1 Data Preprocessing

- Cleaning procedures: Fix AC_Power wrong factor to get correct kW values, synchronize Datetime format
- Outlier handling: Replace outliers with missing values, but remove rows with 6 consecutive outliers, to decrease time frame of interpolation
- Missing value treatment: Interpolate them with plausible values
- Data transformations: Split original dataframe to seperate inverters to different columns

3.2 Exploratory Data Analysis

- Distribution analysis: For total power we can observe a lean towards lower values, possible due to night time. This represents an inverse gaussian distribution. This of course correlates with the IR-Radiation distribution. The ambient temperature shows multi-modal tendencies, with clear bumps around 23 and 28 degrees. For the module temperatures this is less pronounced, with the bumps at 22 and 45 degrees.
- Time series patterns: From the daily pattern plots we can clearly observe the relation between time of day and IR-Radiation and the Power gain coming with this. With peaks during times of 12:00 to 15:00 o'clock. When it comes to time series analysis over multiple days we can tell which days were sunny and which days were cloudy, based on the power gain and temperatures respectively.
- Correlation analysis: Looking at the correlation matrix we can observe strong correlations between the different power outputs and the IR-Radiation. This is of course because the photovoltaic panels produce more energy in direct sunlight.

3.3 Statistical Analysis

- Probability analysis:
  - The probability of Total_AC exceeding the threshold valu of 120 MW is approximately 0.15.
    - This value changes: the higher the threshold, the lower the probability.
  - The cross tabulation analysis shows the distribution of Total_AC exceeding the threshold across different levels of Irradiation.
    - Depending on the set threshold level, the distribution changes. When set to a higher Total_AC, the more likely it is with a higher irradiation level to be over the threshold.
  - The conditional probability analysis reveals the likelihood of Total_AC exceeding the threshold given different Irradiation levels.
    - For this analysis, only the probabilities around the threshold have a probability not to be 1 or 0. Higher levels have a probability of 1, and lower levels have a probability of 0.
- Law of Large Numbers demonstration: The Law of Large Numbers states that as the number of trials increases, the sample mean will tend to be closer to the population mean. In this case, the Law of Large Numbers is demonstrated by calculating the sample mean of Total_AC exceeding the threshold value of 90,000 for different sample sizes. As the sample size increases, the sample mean tends to be closer to the population mean, which is the probability of Total_AC exceeding the threshold value of 90,000.
- Central Limit Theorem application: The Central Limit Theorem states that the sampling distribution of the sample mean will be approximately normally distributed, regardless of the population distribution, as the sample size increases. In this case, the Central Limit Theorem is applied by calculating the sample mean of Total_AC exceeding the threshold value of 90,000 for different sample sizes and plotting the sampling distribution. As the sample size increases, the sampling distribution tends to be closer to a normal distribution.

  - Q-Q plot analysis: Data from the total yield does not follow a normal disribution as shown in the Q-Q plot. While in the middle section the data is somewhat normally distributed the tails show a clear deviation from the normal distribution.

- Regression analysis:
  - Model selection: We selected a polynomial regression model of degree 5 to predict the daily pattern the ambient temperature follows.
  - Model fitting and validation: We validated the model by comparing the predicted values to the actual values.
  - Cross-validation: We performed a 5-fold cross-validation to evaluate the model's performance. The model achieved an R-squared value of 0.965, indicating a good fit.

4 Key Findings

4.1 Statistical Insights

- Distribution characteristics: The distribution of power output shows a strong skew towards lower values, with peaks during midday hours.
- Significant correlations: There is a strong positive correlation between power output and irradiation levels, indicating that higher irradiation leads to higher power generation.
- Probability analysis results: The probability of Total_AC exceeding 120 MW is approximately 0.15, with higher thresholds resulting in lower probabilities.

4.2 Pattern Analysis

- Temporal patterns: The analysis of temporal patterns revealed that power generation peaks during midday hours, corresponding with higher irradiation levels.
- Variable relationships: There is a strong positive correlation between power output and irradiation levels, as well as a noticeable relationship between ambient temperature and module temperature.
- Identified anomalies: Several anomalies were identified in the dataset, including periods of unexpectedly low power output and temperature readings, which were addressed during the data preprocessing stage. E.g.: the AC_Power had a wrong factor applied to it, which was corrected.

4.3 Advanced Analysis Results

- Interactive visualization insights: The interactive visualizations provide an easy way to analyze different metrics. We compared the Total_AC to the IR-Radiation and the Ambient Temperature with the Module Temperature.
- Regression performance: A polynomial regression model of degree 5 was chosen to predict the daily pattern of ambient temperature. The model was validated by comparing predicted values with actual values. A 5-fold cross-validation was conducted to assess the model's performance, resulting in an R-squared value of 0.965, indicating a strong fit.

5 Summary and Conclusions

- Main insights: The analysis revealed strong correlations between power output and irradiation levels, as well as significant daily patterns in temperature and power generation.
- Limitations: The dataset had missing values and outliers which were addressed, but may still affect the overall analysis. Additionally, the analysis was limited to a 34-day period, which may not capture long-term trends.
- Future analysis suggestions: Future work could include a longer time period to capture seasonal variations, and the use of more advanced machine learning models to improve prediction accuracy.