Data Analysis Project Report

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

January 8, 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:

- Correlation analysis:

- Key visualizations:

3.3 Statistical Analysis

- Probability analysis:

- Law of Large Numbers demonstration:

- Central Limit Theorem application:

  - Q-Q plot analysis (if applicable):

- Regression analysis:

  - Model selection:
  - Model fitting and validation:
  - Cross-validation (if applicable):

4 Key Findings

4.1 Statistical Insights

- Distribution characteristics:
- Significant correlations:
- Probability analysis results:

4.2 Pattern Analysis

- Temporal patterns:
- Variable relationships:
- Identified anomalies:

4.3 Advanced Analysis Results

- Interactive visualization insights:
- Regression performance:
- Additional findings:

5 Summary and Conclusions

- Main insights:
- Limitations:
- Future analysis suggestions: