Data Analysis Final Assignment Report

Team: TEAMNAME
Oswald Lackner & *Christoph Stocker*

1 Contributions

Clearly state each team member's specific contributions. Be concrete.

- Oswald Lackner:
  - Dataset selection and acquisition
  - Data quality analysis and preprocessing pipeline
  - Filtering visualization
  - Dataset overview
  - Raw hexbin-plots, Power share plots, Seasonal plots
  - Probability calculation and cross-tab
  - Regression
- Christoph Stocker:
  - Dataset overview, Statistical analysis, Original data analysis
  - Distribution analysis with histograms, Correlation analysis, Periodic pattern analysis, summary of observed patterns
  - Law of Large Numbers (LLN)
  - Central Limit Theorem (CLT)

2 Dataset Description

- Dataset name and source (Kaggle, Hugging Face, Westermo tests, etc.):

  Hourly power generation of European countries, sourced from the **ENTSO-E** (European Network of Transmission System Operators) Transparency Platform. (Downloaded from Kaggle)

- Why it is suitable for time-series analysis:

  The dataset provides periodic observation of power consumption for the countries Germany, Italy, France, Spain at an sample rate of 15 minutes to 60 minutes over a time of about 8 years.

- Time period covered and sampling frequency:

  - Period: 01.01.2015 to 26.09.2022.
  - Germany is sampled every 15 minutes;
  - France, Spain, and Italy are sampled every 60 minutes.

- Key variables analysed (signals, sensors, physical quantities):

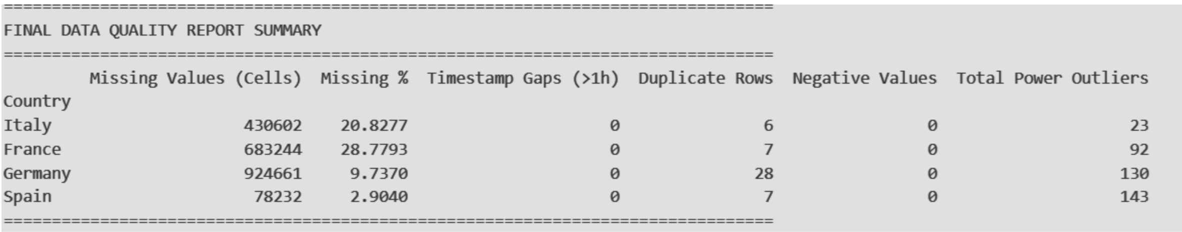
  Aggregated hourly power generation in MW of Germany, France, Spain and Italy Furthermore the different generation methods like Biogas, Onshore and Offshore Wind turbines, Nuclear, Hydropower an much more.

- Size and structure:

  - Number of observations (rows): 59,064 to 271,296 datapoints per country
  - Number of features (columns): 30
  - Target variable(s) if any:

    - Total Power Germany
    - Total Power Spain
    - Total Power Italy
    - Total Power France

- Missing data summary:



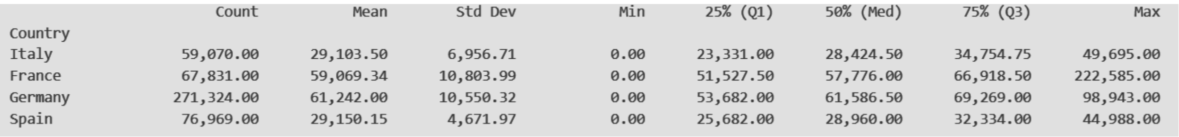
```
================================================================================
FINAL DATA QUALITY REPORT SUMMARY
================================================================================
         Missing Values (Cells)  Missing %  Timestamp Gaps (>1h)  Duplicate Rows  Negative Values  Total Power Outliers
Country
Italy                    430602    20.8277                     0               6                0                    23
France                   683244    28.7793                     0               7                0                    92
Germany                  924661     9.7370                     0              28                0                   130
Spain                     78232     2.9040                     0               7                0                   143
================================================================================
```

3 Task 1. Data Preprocessing and Basic Analysis

3.1 Basic statistical analysis using pandas

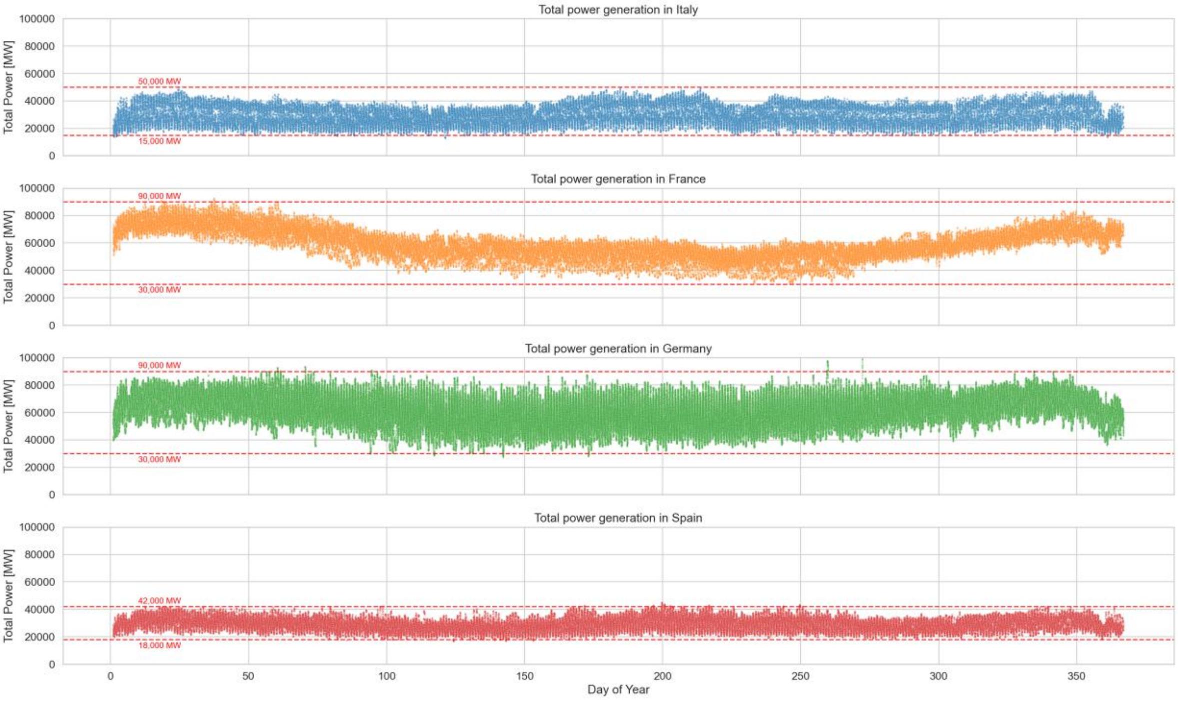
- Descriptive stats (mean, std, min, max, quantiles) for key variables:

  Key variables: total power in of every country in MW



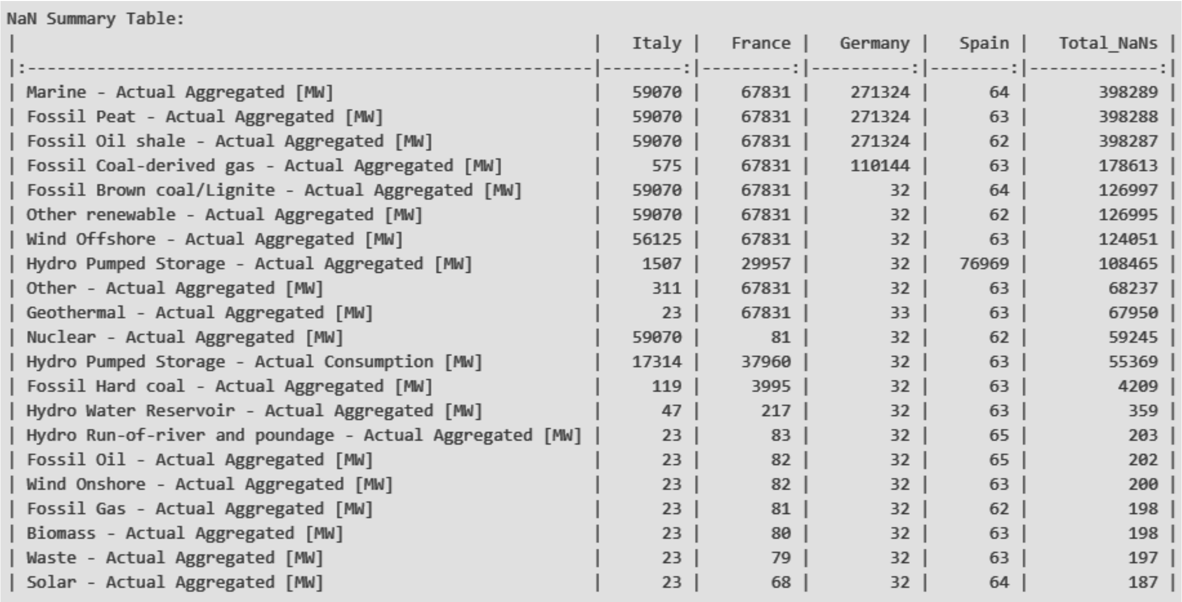
```
                 Count        Mean     Std Dev    Min    25% (Q1)    50% (Med)    75% (Q3)         Max
Country
Italy        59,070.00   29,103.50    6,956.71   0.00   23,331.00    28,424.50   34,754.75   49,695.00
France       67,831.00   59,069.34   10,803.99   0.00   51,527.50    57,776.00   66,918.50  222,585.00
Germany     271,324.00   61,242.00   10,550.32   0.00   53,682.00    61,586.50   69,269.00   98,943.00
Spain        76,969.00   29,150.15    4,671.97   0.00   25,682.00    28,960.00   32,334.00   44,988.00
```

- Grouped summaries where relevant (by day, device, category, test run):

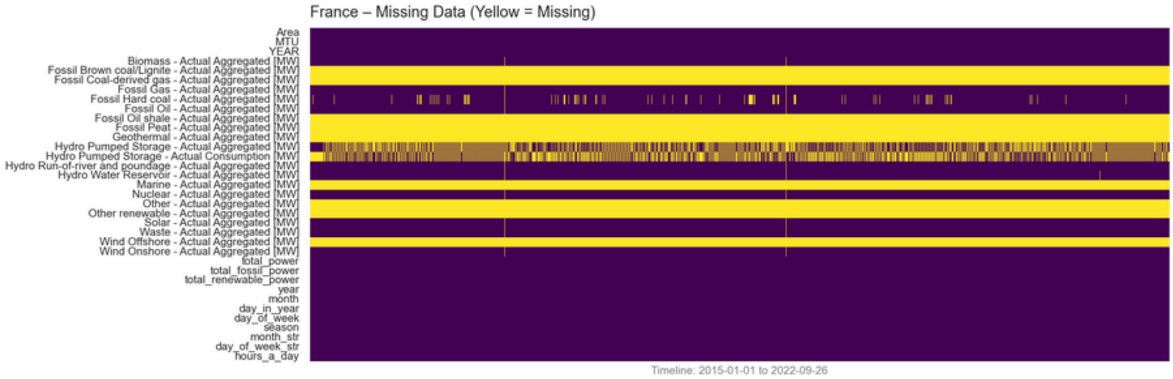


3.2 Original data quality analysis including visualization

- Missingness patterns (counts, heatmap, timeline gaps):
  Timestamp was missing for very high extend of the data but could be constructed out of the field MTU. According to country power mix also several power sources were missing. Some power source showed wired occurrence our disappearance which doesn‘t make sense without further information.



```
NaN Summary Table:
|                                                         |   Italy |   France |   Germany |   Spain |   Total_NaNs |
|:--------------------------------------------------------|--------:|---------:|----------:|--------:|-------------:|
| Marine - Actual Aggregated [MW]                         |   59070 |    67831 |    271324 |      64 |       398289 |
| Fossil Peat - Actual Aggregated [MW]                    |   59070 |    67831 |    271324 |      63 |       398288 |
| Fossil Oil shale - Actual Aggregated [MW]               |   59070 |    67831 |    271324 |      62 |       398287 |
| Fossil Coal-derived gas - Actual Aggregated [MW]        |     575 |    67831 |    110144 |      63 |       178613 |
| Fossil Brown coal/Lignite - Actual Aggregated [MW]      |   59070 |    67831 |        32 |      64 |       126997 |
| Other renewable - Actual Aggregated [MW]                |   59070 |    67831 |        32 |      62 |       126995 |
| Wind Offshore - Actual Aggregated [MW]                  |   56125 |    67831 |        32 |      63 |       124051 |
| Hydro Pumped Storage - Actual Aggregated [MW]           |    1507 |    29957 |        32 |   76969 |       108465 |
| Other - Actual Aggregated [MW]                          |     311 |    67831 |        32 |      63 |        68237 |
| Geothermal - Actual Aggregated [MW]                     |      23 |    67831 |        33 |      63 |        67950 |
| Nuclear - Actual Aggregated [MW]                        |   59070 |       81 |        32 |      62 |        59245 |
| Hydro Pumped Storage - Actual Consumption [MW]          |   17314 |    37960 |        32 |      63 |        55369 |
| Fossil Hard coal - Actual Aggregated [MW]               |     119 |     3995 |        32 |      63 |         4209 |
| Hydro Water Reservoir - Actual Aggregated [MW]          |      47 |      217 |        32 |      63 |          359 |
| Hydro Run-of-river and poundage - Actual Aggregated [MW] |      23 |       83 |        32 |      65 |          203 |
| Fossil Oil - Actual Aggregated [MW]                     |      23 |       82 |        32 |      65 |          202 |
| Wind Onshore - Actual Aggregated [MW]                   |      23 |       82 |        32 |      63 |          200 |
| Fossil Gas - Actual Aggregated [MW]                     |      23 |       81 |        32 |      62 |          198 |
| Biomass - Actual Aggregated [MW]                        |      23 |       80 |        32 |      63 |          198 |
| Waste - Actual Aggregated [MW]                          |      23 |       79 |        32 |      63 |          197 |
| Solar - Actual Aggregated [MW]                          |      23 |       68 |        32 |      64 |          187 |
```



- Outliers and suspicious values (plots and rule used):

- Consistency checks (timestamps order, duplicates, impossible values):

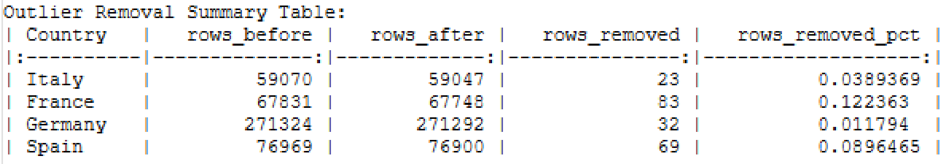
3.3 Data preprocessing

- Cleaning steps, missing-value treatment and outlier handling:

  Out of the available power sources a total power was calculated, cleaning was done using this calculated value by removing line where power production was far beyond typical expectation (upper and lower range limit). Due to very low number of outliers data was dropped. Cleaning of the other sources without deeper knowledge was not possible.

- Feature engineering: datetime reconstruction from MTU, total power was calculated

- Final dataset shape after preprocessing:

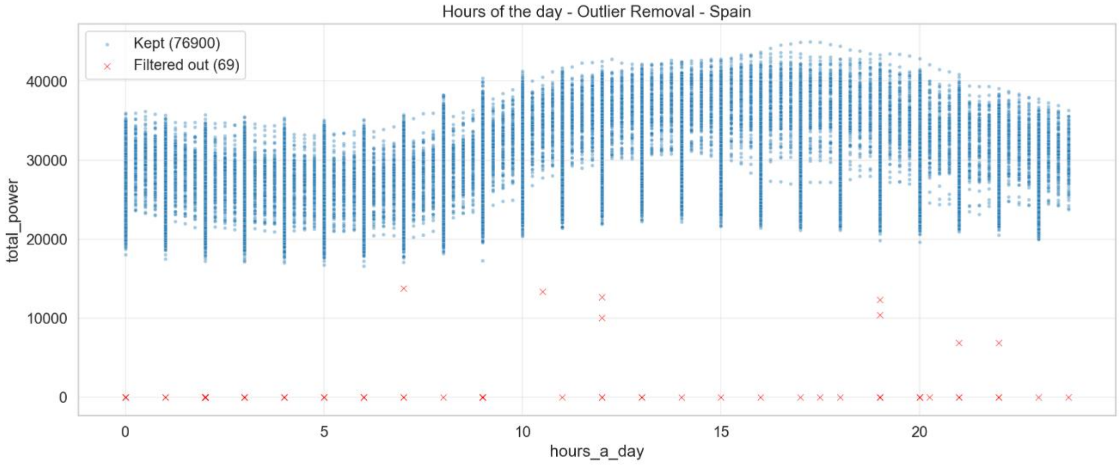


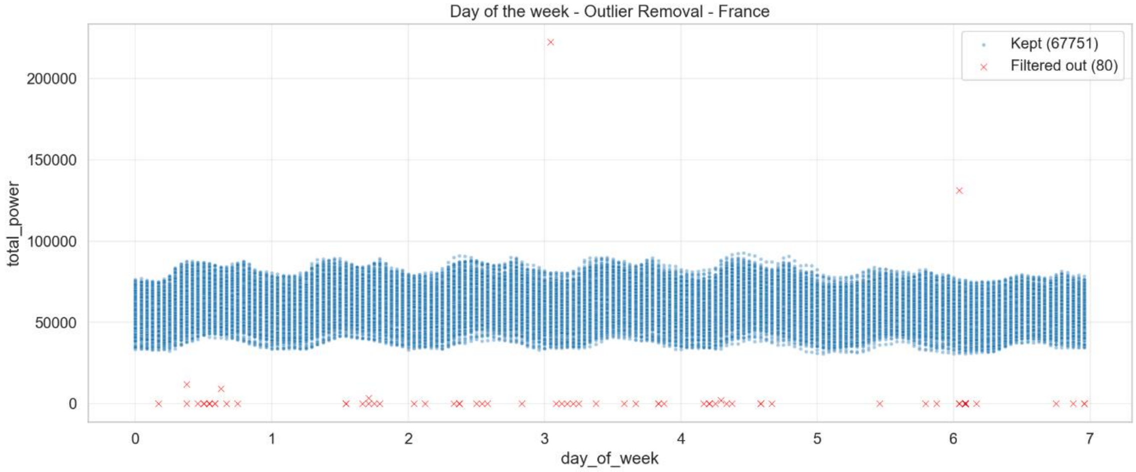
```
Outlier Removal Summary Table:
| Country    |   rows_before |   rows_after |   rows_removed |   rows_removed_pct |
|:-----------|--------------:|-------------:|---------------:|-------------------:|
| Italy      |         59070 |        59047 |             23 |          0.0389369 |
| France     |         67831 |        67748 |             83 |          0.122363  |
| Germany    |        271324 |       271292 |             32 |          0.011794  |
| Spain      |         76969 |        76900 |             69 |          0.0896465 |
```

3.4 Preprocessed vs original data visual analysis

- Before vs after comparison plots (at least 2 to 3 key variables):

- What improved and what trade-offs exist:

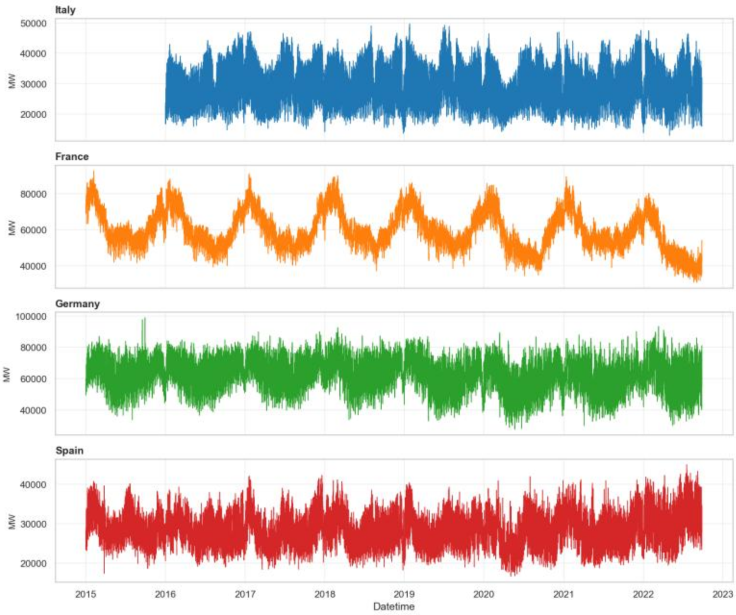




4 Task 2. Visualization and Exploratory Analysis

4.1 Time series visualizations

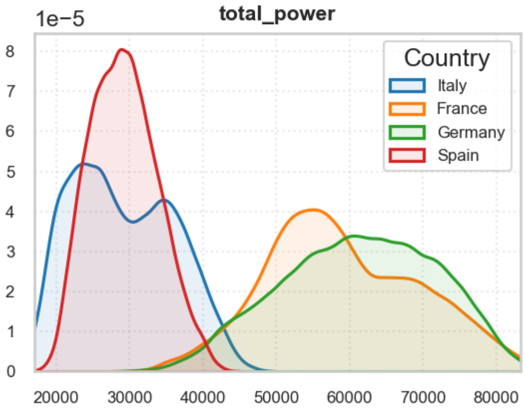
- Plot of main variable(s) over time:



- Annotations for notable events or pattern shifts (if applicable):
  - The timeline reveals a significant missing data period for Italy at the start of the series(2015), where no generation was recorded.
  - France exhibits the most distinct "wave" pattern, with massive winter peaks and deep summer valleys
  - Unlike the smoother curves of France (Nuclear based), Germany's power signal is noisier, reflecting the dependency on renewable sources

4.2 Distribution analysis with histograms

- Histograms for key numeric variables:

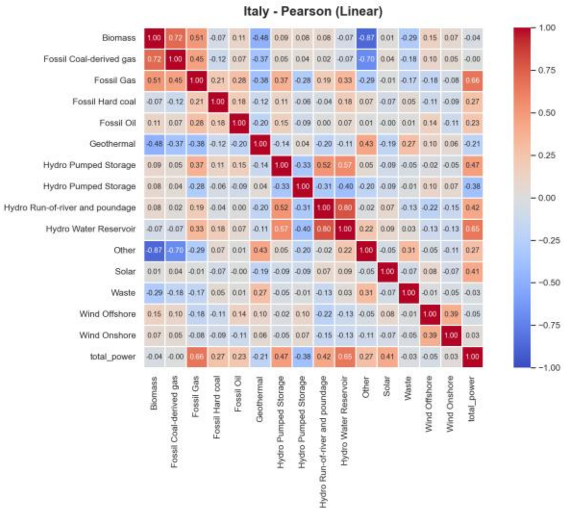


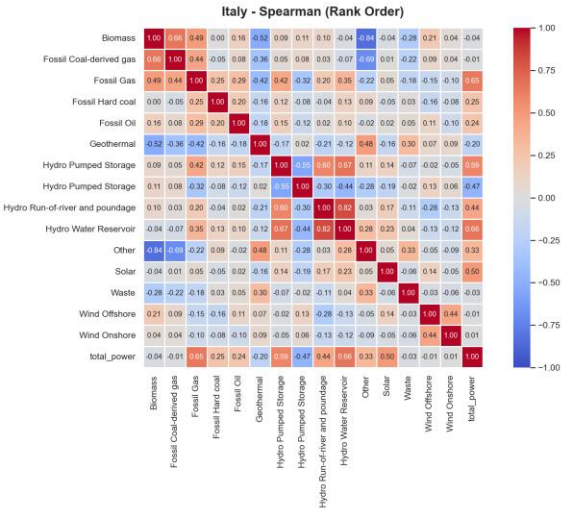
- Notes on skewness, heavy tails, multi-modality:

  <u>ITALY</u>
  Two clear peaks visible, for example day and night or winter and summer peaks
  <u>FRANCE</u>
  Extremely broad distribution
  <u>SPAIN</u>
  Extreme peak, but on the lower half of the spectrum
  <u>GERMANY</u>
  Also broad distribution like France, but much more even, also no sharp peaks

4.3 Correlation analysis and heatmaps

- Correlation type used (Pearson or Spearman) and why:





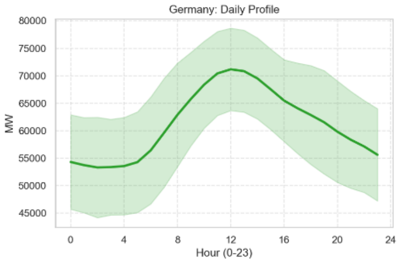
Both diagrams are in our jupyter notebook, to have a better comparison on the nonlinear (Spearman) and linear (Pearson) heatmaps.

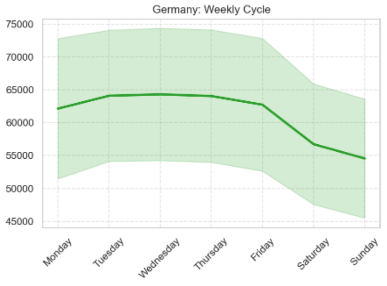
- Heatmap and top correlated pairs with short interpretation:
  <u>Hydro Run of River & Hydro run of Reservoir</u>
  r~0.80 -> strong link (shared cycles like rainfall and snow-melt)
  <u>Biomass & Coal-delivered Gas</u>
  r~0.72 -> similar operation cycles (maybe switched on at high power times e.g. mid-day
  <u>Other & Coal-derived Gas</u>
  r~0.72 -> strong inverse relationship (cleaner sources are likely displacing coal)

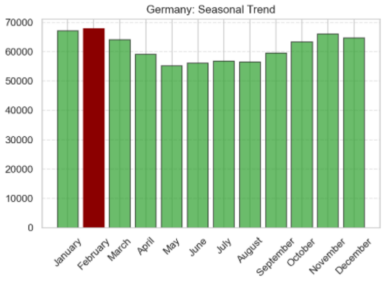
4.4 Daily pattern analysis

- Aggregation method (hourly means, day-of-week, rolling averages):

- Plots showing daily cycles or weekday-weekend differences:







- What patterns are stable vs noisy:
  - Stable: The weekly cycle is highly predictable and on the weekend, demand drops always
  - Noisy: The Daily Profile shows high difference during daylight hours. While the average trend is clear, the actual daily values change wildly due to the solar and wind energy production in Germany

4.5 Summary of observed patterns, similar to True/False questions

Write short, testable statements and answer them based on evidence. Example format below.

GERMANY

1. Significant consumption drop on Weekends (>5%) [TRUE]

-> Weekends are 12.3% lower than Weekdays (Average: 55GW vs 63GW)

2. Grid load is dominated by Winter heating demand [TRUE]

-> Winter load is 18.0% higher on average
Temperature (average): summer max 20°C, winter min 1.7°C

3. Strong correlation between Solar Generation and Total Load [FALSE]

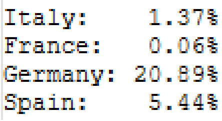
-> Pearson Correlation coefficient is 0.37
Good correlation between Wind offshore and onshore (0.64)

5 Task 3. Probability Analysis

5.1 Threshold-based probability estimation

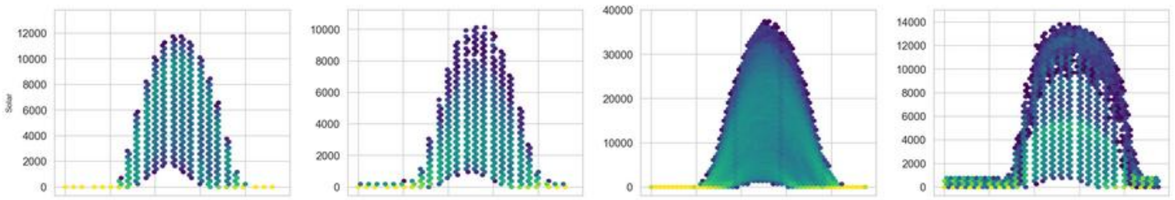
- Define threshold(s) and justify choice: Threshold of 10MW Solar power, to show the high extend of photovoltaic power sources of Germany which easily reaches this threshold at daytimes whereas a country like France with similar Industrial level will hardly reach this threshold at best times. A comparison of electrical energy prices, countryside funding and network frequency stability would be interesting here.

- Estimate probabilities of exceeding the threshold of 10 MW:



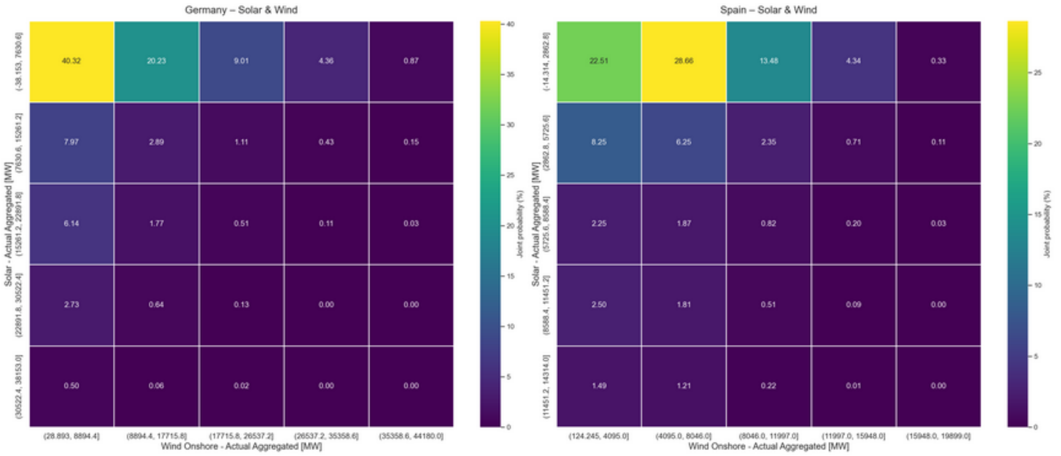
```
Italy:     1.37%
France:    0.06%
Germany:  20.89%
Spain:     5.44%
```

- Solar Power production over day time(0-24) at Italy, France, Germany and Spain:



5.2 Cross tabulation analysis

- Interesting event is the availability of either wind or solar energy for a power production by renewable energy sources.



5.3 Conditional probability analysis

- Compute and interpret $P(A)$, $P(B)$, $P(A \mid B)$, $P(B \mid A)$:

- Include at least one meaningful comparison and conclusion:

5.4 Summary of observations from each probability task

- Key takeaway from threshold probability:

- Key takeaway from crosstab: The probability of wind and solar energy operating both at their lowest operation point at same time, being one fifth of nominal power is already 22.5% (Spain) to 44.3% (France) which creates already a very high necessity to save the energy produced in the other partial of the time.

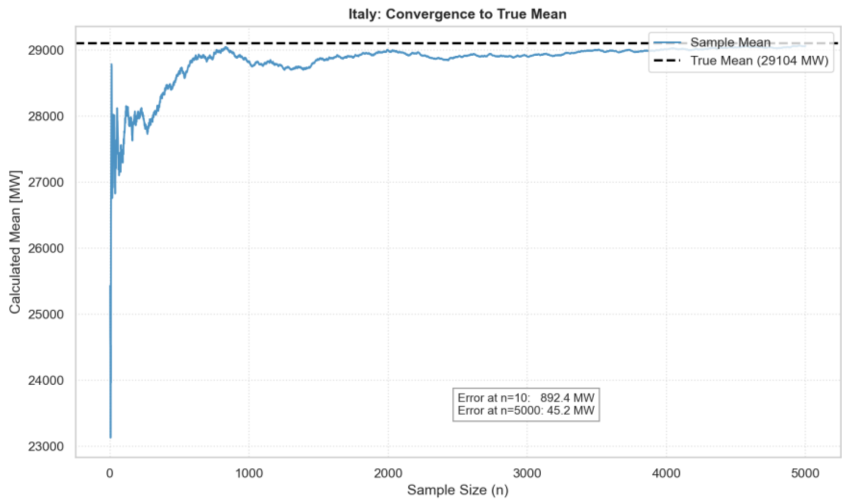
6 Task 4. Statistical Theory Applications

6.1 Law of Large Numbers (LLN) demonstration

- Variable chosen and why it makes sense:

The “Total Generation” variables have a relatively high spectrum, and high values, which it makes interesting to use such a variable

- Experiment: show sample mean as *n* increases:

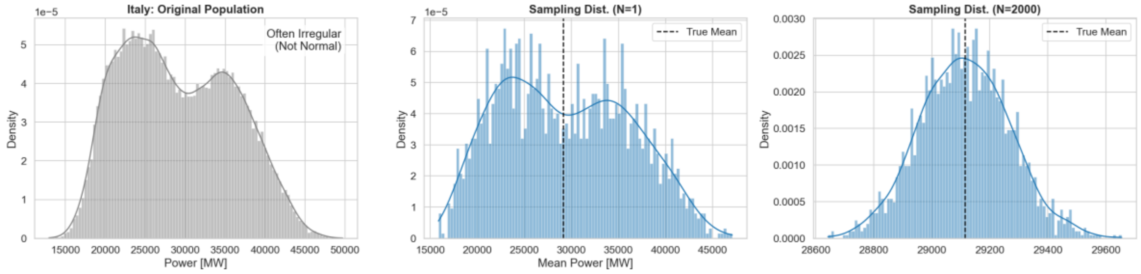


- Plot and short interpretation:

- The jagged line starts wildly volatile because 'n' is small (small sample size).
- As 'n' increases (moving right), the coloured line flattens and is getting closer to the mean.
- This proves larger datasets generates more reliable statistics.

6.2 Central Limit Theorem (CLT) application

- Sampling procedure (sample size, number of trials, with or without replacement):
  General procedure
  Randomly draw n data points (sample) and calculate the mean
- Show distribution of sample means for increasing *n*:
  N=1
  Here no distinct bell shape is visible, more a distribution with two peaks
  N=2000
  The distribution transforms into a nearly perfect bell curve around the true mean.
- Plot(s): histogram(s) of sample means and comparison to normal shape:



6.3 Result interpretation

- What LLN showed in your data context:

  The experiment confirmed that the power generation is highly unstable on an hourly basis. But the long-term average is a stable and deterministic value.

  It demonstrates that a sample size of about 2000 values is required to cancel out the random noise of the chaotic world (like weather, day night cycles) and get a precise estimate of the average

- What CLT showed, and any deviations and why:

  Generally the CTL shows, that we can use the normal distribution even on a dataset, which is non-normal (with two peaks)
  At N=2000 no significand deviations of the theorem was observed. The convergence to a bell curve was confirmed despite the underlying data doesn't have a normal distribution.

7 Task 5. Regression Analysis

7.1 Linear or Polynomial model selection

- Define target *y* and predictors *X*:

- Motivation for linear vs polynomial: To show the continuous rise in power consumption linear regression would have been nice, but due influence of 2020 year and therefore lower power consumption this wasn't working out.

- Any train-test split rationale (time-aware split if relevant):

7.2 Model fitting and validation

- Fit procedure and preprocessing (scaling, feature selection):

- Validation method (holdout, time-series split, etc.):

- Metrics reported (RMSE, MAE, R^2) and why:

- Residual analysis (at least one plot recommended):

7.3 Result interpretation and analysis

- Main effects and practical meaning:
  - LLN - Long term averages are stable
  - CLT proves usage of simple standard statistical tools for complex data sets

- Failure cases or where model performs poorly:
  - Short term predictions are often faulty

8 Key Findings and Conclusions

- Main findings from preprocessing and EDA:
- Main findings from probability tasks:
- Main findings from LLN and CLT:
  - LLN -> Long term averages are stable, regardless of high hourly frequency and noise
  - Small samples are unreliable (n~2000 is ok)
  - CLT -> sample means always form a bell curve
  - Random sampling broke time correlation and ensured independence
- Main findings from regression: degree of 1 was searched, steep improvement with higher degree with a certain point where no relevant improvement could be found any more. In our case a point of worsening with higher degree was not found. Curves are steeper in transition with higher order.
- Limitations:
- What you would do next if you had more time:
  - Add annotations for known deactivated for power plants
  - Calculate power production per citizen
  - Add country specific network frequency

9 Reproducibility Notes

- Source of Data: Kaggle: Hourly Power Generation of Europe (date: 2026-01-16)
- Key libraries : `numpy, pandas, matplotlib, seaborn, scipy, scikit-learn, statsmodels, jupyter, ipykernel, tqdm`
- How to run the notebook end-to-end:
  - Create venv and install packages with delivered requirements.txt
  - Start jupyter server in your venv
  - Go to cell 11, uncomment line 4, comment out lines 2-3