**Assignment I.**

**Analysis of photovoltaic (PV) power plant located near Tallinn.**

**Ilja Matjas**

**Taltech / Tallinn 2024**

[**ilja.matjas@taltech.ee**](mailto:ilja.matjas@taltech.ee)

**Introduction**

Home assignment (hereinafter HA) is based on analysis of photovoltaic (hereinafter PV) power plant (hereinafter PP) located near Tallinn. Input data and Scope of Work (hereinafter SoW) were provided by course instructors.

Provided data is time-series and includes hourly electricity production data and temperature data.

Multiple analytical and coding tasks, as defined in SoW, are done within HA. Result, finding and discussions are presented in this report.

1. **Business understanding**

Analysis and modelling PV production can be useful at least in terms of commercially valuable insights and scientific research (R&D). For sure, there are still dozens more useful stuff to be done with such data, but I will stick to these two.

* Commercial value:
  + accurately predicting production of PV for next hour, day, etc., to minimise production-consumption balancing costs
  + accurately predicting production of PV to make correct decisions on energy futures markets
  + further investment decisions (shall I invest in one more PV PP?)
* Scientific research:
  + insights for planning energy power system adequacy (adequacy means that power system generation shall be able to cover power system load on every timestep)
  + insights for research and modelling if and how to use PV PPs in frequency supporting services (FCR, FRR). PV PP is of-course not 100% reliable energy source, but maybe there is some way to use them as reserves based on probabilistic models.
  + use such data as a measurement tool to track trends related to nature phenomenon. Maybe it can be used to derive some conclusions based on how production is changing from year-to-year?

1. **Data collection**
   1. **Understanding available data**

Data is provided in tabular form and contains 17016 rows and 6 columns. Columns included in dataset are represented in Table 1

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Value example** | **Count** | **Data type** |
| timestamp | 1 612 224 000 000 | 17 016 | int |
| raw | 178 582.47 | 17 016 | float |
| temperature | -3.9 | 17 016 | float |
| modified1 | 0.60 | 15 825 | float |
| modified2 | 172 644.91 | 15 641 | float |
| modified3 | 17.34 | 15 879 | float |

Table 1 – Description of provided data columns

From SoW, Table 1 and data itself we can understand what kind of data is provided.

2.1.1 Timestamp

Timestamp is a Unix-timestamp given in milliseconds. We can understand that these are milliseconds from power of values e.g. 1.6e+12. In seconds, the order would be e.g. 1.6e+9. Time step is 1 hour or 3 600 seconds, however it is not always the case in provided data. Different time steps and their counts are given in Table 2.

|  |  |  |
| --- | --- | --- |
| **Time step (s)** | **Time step (h)** | **Count** |
| 3 600 | 1 | 17 004 |
| 90 000 | 25 | 10 |
| 954 000 | 265 | 1 |

Table 2 – Time steps and their counts of provided data

From Table 2, we can conclude that for ten times data records are completely missing daily. And once, records are missing for 11 days. If we will plot occurrence of daily missing records (see Chart 1), we will see that some pattern of missing records is rather present in the data, which, for example, could be monthly planned outages which took place during the first year after commissioning PV PP.

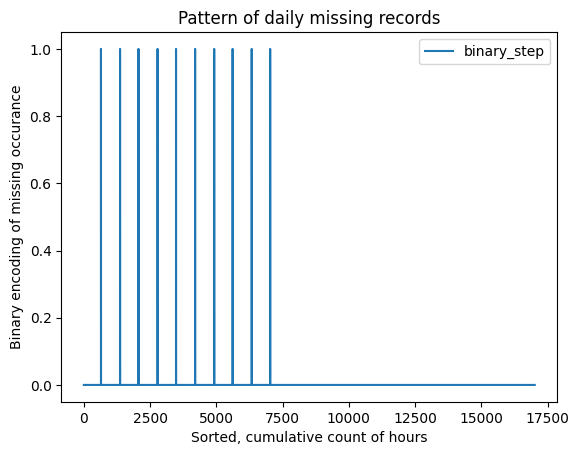


Chart 1 – pattern of daily missing records

2.1.2 Raw

Raw is an hourly production of PV PP. As maximum value of raw equals to 301 630, we can with very high confidence conclude that measuring unit is Wh (Watt-hour). Raw cannot be given in kWh, because physically there is no 300 MW large PV PP installed in Estonia. As it is more convenient to work in kWh scale, I converted raw column to kWh.

2.1.3. Temperature

Temperature is given in Celsius. This can be concluded from deriving minimum and maximum values of the column. Minimum value is -24.5 and maximum 30.1. Thus, we can be confident that data is not provided in Fahrenheit or Kelvin.

2.1.4. Modified

Modified1, modified2 and modified3 columns are randomly or using some logic, modified raw columns. Thus, these values are also provided in Wh.

* 1. **External data sources**

It depends on scope of analysis and modelling if any additional sources might be needed or not.

To understand what provided data is, what are the measuring units, etc. I didn’t need to use any additional external resources.

To catch more patterns or to end up with nice predictive models, some additional sources shall be very useful. For example, cloudiness has very large effect of PV PP production due to reduced level of sun insolation. Also wind strength, as wind is acting as a cooling, it allows to generate more.

As we are dealing with PV PP, sun map and sun-clock information might be very useful to handle zeros. However I haven’t used sun-map data and tried to derive night hours from data itself. I needed to do that under ‘feature engineering’ part of work.

Information related to planned (and maybe unplanned) outages of PV PP might be very useful, especially if outages are in correlation with weather conditions or happen periodically. Electricity price data might be useful, as plant could be just turned off if the price went negative (cheaper to shut down the plant instead of paying for production). So, quite a lot of additional data might be used. However, to complete HA, due to time-limitations and due to the fact that it is a school project, I haven’t used any additional external data sources.

1. **Data cleaning**

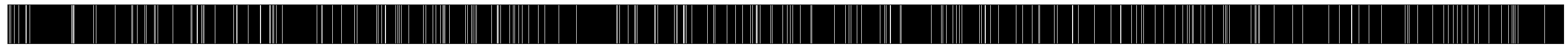
After reading the data, analysis of missing values was done. Results of quick check is represented in Table 3.

|  |  |  |
| --- | --- | --- |
| **Column** | **Count of missing values** | **% of missing values** |
| Timestamp | 0 | 0 |
| Raw | 0 | 0 |
| Temperature | 0 | 0 |
| modified1 | 1 191 | 7.00 |
| modified2 | 1 375 | 8.08 |
| modified3 | 1 137 | 6.68 |

Table 3 – missing values of provided data

As it can be seen from the Table 3, only columns modified1, modified2 and modified3 contain missing values.

To get better understanding regarding the nature of missing values some steps were taken. First, a binary visualisation was done (Chart 2). Top – modified1 and bottom – modified3.



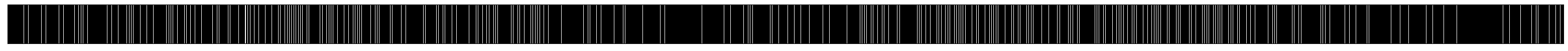




Chart 2 – binary representation of missing values

White strips on a black font are representing occurrence of missing value (data is sorted based on timestamp). From Chart 2 we can see that nature of missing values of modified3 is something different compared to modified1 and modified2. However, from such representation, we cannot conclude mechanism of missing values. To collect more information, I’ve done correlation analysis for missing values.

For correlation analysis I used temperature as a feature and additionally generated three more time-based features: day of week, month and hour.

To catch correlations between missing values and features mentioned above, I’ve binary encoded modified columns such, that values with present data were set to 1 and with missing values were set to 0. For time-based features (day of week, month and hour) one-hot-encoding was applied. Temperature was left as is. Top 6 correlations (based on absolute values) for each modified feature are represented in Table 4.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **modified1 correlations** | | **modified2 correlations** | | **modified3 correlations** | |
| day\_of\_week\_3 | 0.026 | hour\_10 | -0.271 | temperature | -0.501 |
| hour\_6 | 0.017 | hour\_11 | -0.238 | month\_6 | -0.197 |
| hour\_7 | 0.017 | hour\_9 | -0.233 | month\_7 | -0.145 |
| hour\_23 | 0.016 | hour\_12 | -0.144 | month\_8 | -0.109 |
| day\_of\_week\_1 | -0.015 | hour\_8 | -0.136 | month\_9 | 0.06 |
| hour\_16 | -0.013 | temperature | -0.116 | month\_3 | 0.06 |

Table 4 – correlations between missing values of modified columns and other features

From Table 4 we can see that missing values of modified1 are not in correlation with any another feature. Thus, I conclude that modified1 is missing completely at random (**MCAR**).

Missing values of modified2 is strongly correlating with certain hours (8-12) with peak at hour 10 and is correlating with temperature. As data is never missing during night, evening or twilight hours, and distribution of missing data is very similar to distribution of mean hourly production of modified2 (Charts 3 and 4), it can be concluded that there is a relationship between missing values and level of production of PV PP. A hypothesis might be raised, that messing values are related to measurement malfunctions due to high sun insolation intensity at peak hours. Such missing data mechanism can be concluded as missing at random or (**MAR**).

A graph of a number of missing data

Description automatically generated A graph of a number of hours

Description automatically generated

Chart 3 – missing values hourly distribution Chart 4 – production hourly distribution

Missing values of modified3 are very strongly correlating with temperature. Chart 5 shows distribution of missing values with respect to temperature. Starting from 21C data is not being measured or transmitted at all. This could have happened because with high temperatures, generation and production of PV PP is so high, that measurement system cannot handle such high datapoints (or PV PP is shutting down) and thus data is not being transmitted. If so, then the reason of missing data is the data point itself. Such missing data mechanism can be concluded as (**MNAR**).

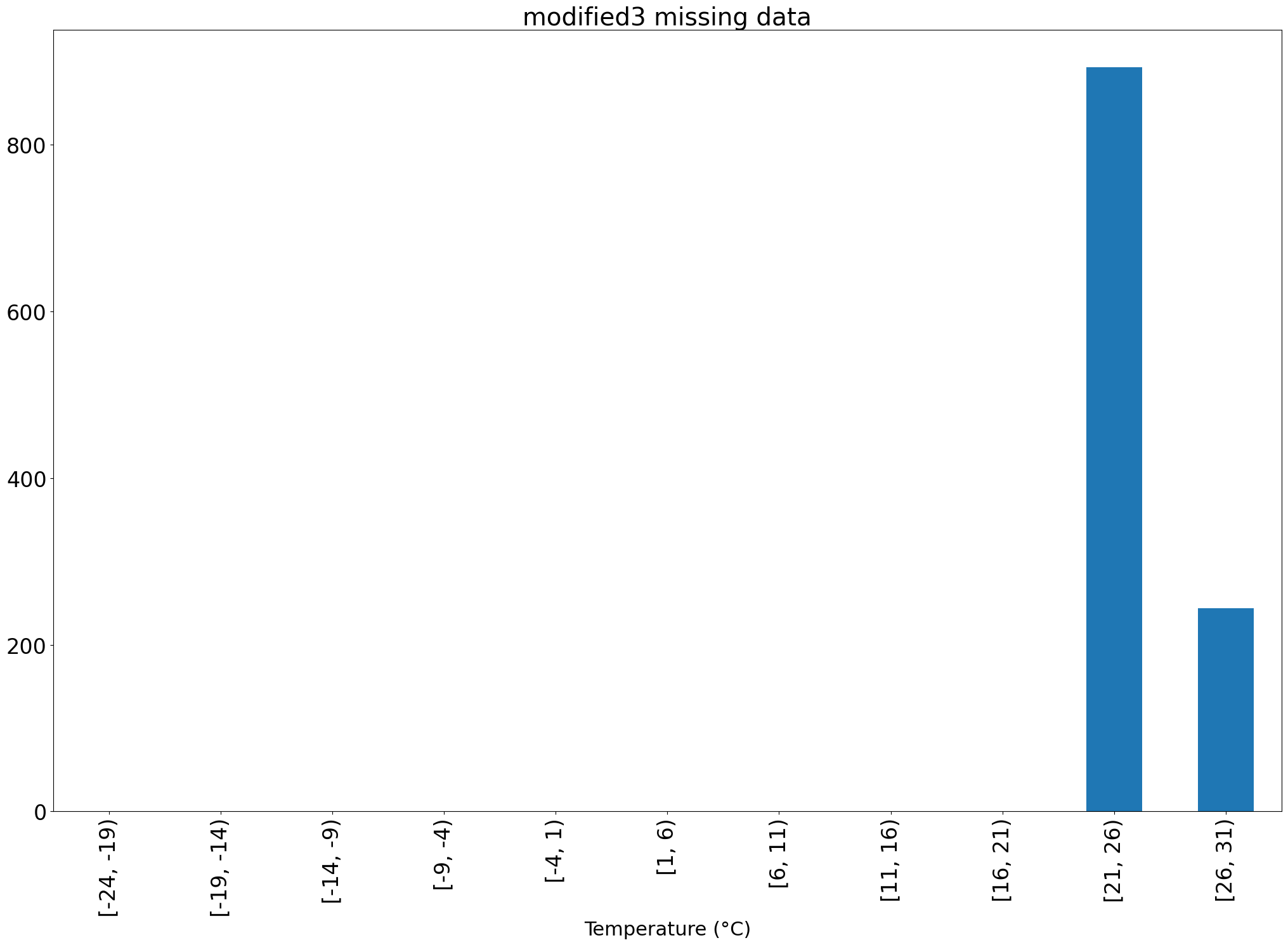


Chart 5 – missing values hourly distribution

* 1. **Dealing with missing data of modified1**

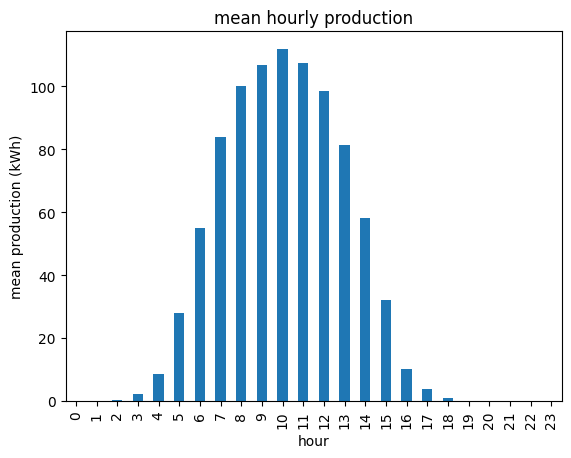
First and very simple way to deal with missing data is just to drop it. As stated in Table 3, after dropping missing data of modified1, dataset shall lose 7% of data.

More advanced way to impute missing values would be to build a model which then would fill missing values. To make it accurate, natural patterns of the data and mechanism of missing values shall be considered. We already concluded that missing values are MCAR, so no sophisticated modelling approach shall be required.

We also know that we are dealing with sequential data. In sequential data, points located closer to each other, often carry most information regarding neighbouring points.

Thus, it would be a good idea to do single imputation grouping data daily e.g. (to fill ’2022-01-01’ missing values, ’2022-01-01’ hourly data shall be used) and for multivariate imputation averaging same hours of neighbouring days e.g. (to fill ’2022-01-01’ missing values, ’2021-12-31’, ’2022-01-01’ and ’2022-01-02’ mean hourly data shall be used).

From Chart 6 we can see hourly distribution of daily PV PP production based on modified1. It is quite accurately bell-shaped. Even-order polynomial would be a good choice to fit such data. To mimic bell-shape, 6’th order polynomial was used and result of fitting mean hourly production is shown on Chart 7.

A graph of a normal distribution

Description automatically generated

Chart 6 – modified1 hourly distribution Chart 7 – polynomial fit of Chart 6

Polynomial fit will also return negative numbers. In our case, negative numbers can be converted to zeros, as we have the knowledgebase that during whole period of measurements PV PP was not acting like a consumer, thus modified1 production shall not go below zero. This fact can be used in modelling, we can simply ‘shave’ negative values by converting them to zeros.

Both univariate and multivariate imputations were applied, and results of performance metrics and standard errors given in Table 5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Imputation Method** | **Negative value shaving** | **MSE / SE** | **MAE / SE** | **MAPE / SE** |
| **Single** | No | 78.30 / 16.50 | 1.19 / 0.067 | 5.93 / 3.76 |
| Yes | 35.94 / 4.0 | 0.81 / 0.046 | 2.46 / 1.64 |
| **Multivariate** | No | 62.37 / 5.65 | 1.17 / 0.06 | 3.26 / 1.91 |
| Yes | 60.61 / 5.65 | 1.04 / 0.059 | 1.67 / 1.15 |

Table 5 – performance metrics and SE of different imputation methods

MAPE metric shall be considered quasi-MAPE, as all zeros from actual values were discarded prior calculate MAPE (otherwise MAPE would have been infinity).

From MAE and MSE metrics we can see that multivariate imputation performed a bit better compared to univariate imputation. It is because model can generalise better as more observations was used to build model. ‘Shaving’ negative part also helped, especially in scenario with single imputation, and that also makes sense as no negative values should be present in PV PP production. Multivariate model was also more robust in terms of not going to negative part (as it had more data to generalise over hours).

Large values of MSE are signalising that some missing points were badly predicted. One example of good and bad predictions is shown on Charts 8 and 9.

A graph of a bad actual vs predicted production

Description automatically generated with medium confidence A graph of a graph

Description automatically generated

Chart 8 – example of bad fit Chart 9 – example of good fit

Overall, results of imputation can be concluded as satisfying.

1. **Data exploration and Feature engineering of raw column**

Statistical description of raw (production) data is given in Table 6.

From statistical description we can see that standard deviation is almost twice as high as mean. It indicates on large variability in the data such as high spread, possible outliers or skewness. If we look further, we can also notice that median is zero. It means that at least 50% of datapoints are equal to zero. Thus, data is having high right-skewed distribution. Maximum value is almost 8 times larger and 90% quantile almost 5 times than mean. Thus, there is a spread in the data as well.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Metric** | **Count** | **Mean** | **Std** | **Min** | **Max** | **Q 1/4** | **Q 1/2** | **Q ¾** | **Q 9/10** |
| **Value** | 17 016 | 36.91 | 66.91 | 0 | 301.63 | 0 | 0 | 38.26 | 154.99 |

Tabel 6

Distribution of raw represented on Chart 10 confirms our conclusions above. Data is indeed right skewed with significant spread and zeros occurring in more than half of all occurrences.

**A graph of a distribution of production data

Description automatically generated**

Chart 10 – distribution of production (raw)

Such zero-rich skewness can be considered as extremely high skewness. I am not sure if there is a way to apply a transformation on such data to get data close to normally distributed and avoid data loss (e.g. be able to represent 100% of data via back-transform). During HA I couldn’t come up with good transformation keeping that large number of zeros.

But we know that during non-sunny hours there is no PV PP production. So, first, simple and robust idea would be to drop non-sunny hours. For that, additional data might be used. Alternatively, we can derive night-hours from data itself. Charts 11 and 12 are basically represent ‘monthly-hourly’ zero hours which could be interpreted as night hours and dropped.

After dropping night hours, distribution became better (Chart 13), but still contains quite many zeros. To be able to transform data as more-or-less normal distributed I had to discard remaining zero values as well (after discarding night-hours there were still too many zeros left).

A graph of a number of bars

Description automatically generated with medium confidence

Chart 11 – hourly box plot of production (raw) data

A screen shot of a graph

Description automatically generated

Chart 12 –daily-hourly heat map of production (raw) data

A graph of a number of production data

Description automatically generatedA graph of a distribution of production data

Description automatically generated

Chart 13 – distribution of raw wo night Chart 14 - distribution of raw after tr-form

I applied nested transformation (power 1/3 -> power ½ -> log\_e (3 – data) and was able to come up with distribution represented on Chart 14. I also tested back transformation and managed to get original data. Thus, transformation did not result in any data loss, which can be considered as good transforming practice. Overall results of transformation can be concluded as satisfying and dealing with zeros must include some additional/special modelling or transformation process.

1. **Additive classical decomposition**

To understand trend, seasonal and residual components, I plotted monthly production data (Chart 15).

A graph of a graph showing the production of a product

Description automatically generated with medium confidence

Chart 15 – PV PP monthly production (raw)

From Chart 15 we can conclude that seasonality is at least monthly based (should be also quarter based).

To get a trend from such data we must do yearly aggregation. However, we have only two years, so we would have only 2 datapoints to calculate trend. We’ve could make a linear approximation of the trend between these two points, but that would be rather approximation than a comprehensive trend over a longer period. Thus, in terms of this HA I neglect trend component and set it to zero T(t) = 0.

!!!!!!!!!!!!!!!!!!!!!!!!!!!!! FROM HERE!!!!

To calculate seasonal component S(t), points representing each season were averaged. As I concluded that trend is neglected, no de-trading prior calculating seasonal component was needed. Results of S(t) are given on Chart XXX.

Finally, residual component can be calculated by subtracting both trend T(t) and seasonal S(t) components. Result of R(t) is given on Chart XXX

A graph with numbers and lines

Description automatically generated

A graph with blue lines

Description automatically generated