**Chapter 4: Data Analysis and Findings**

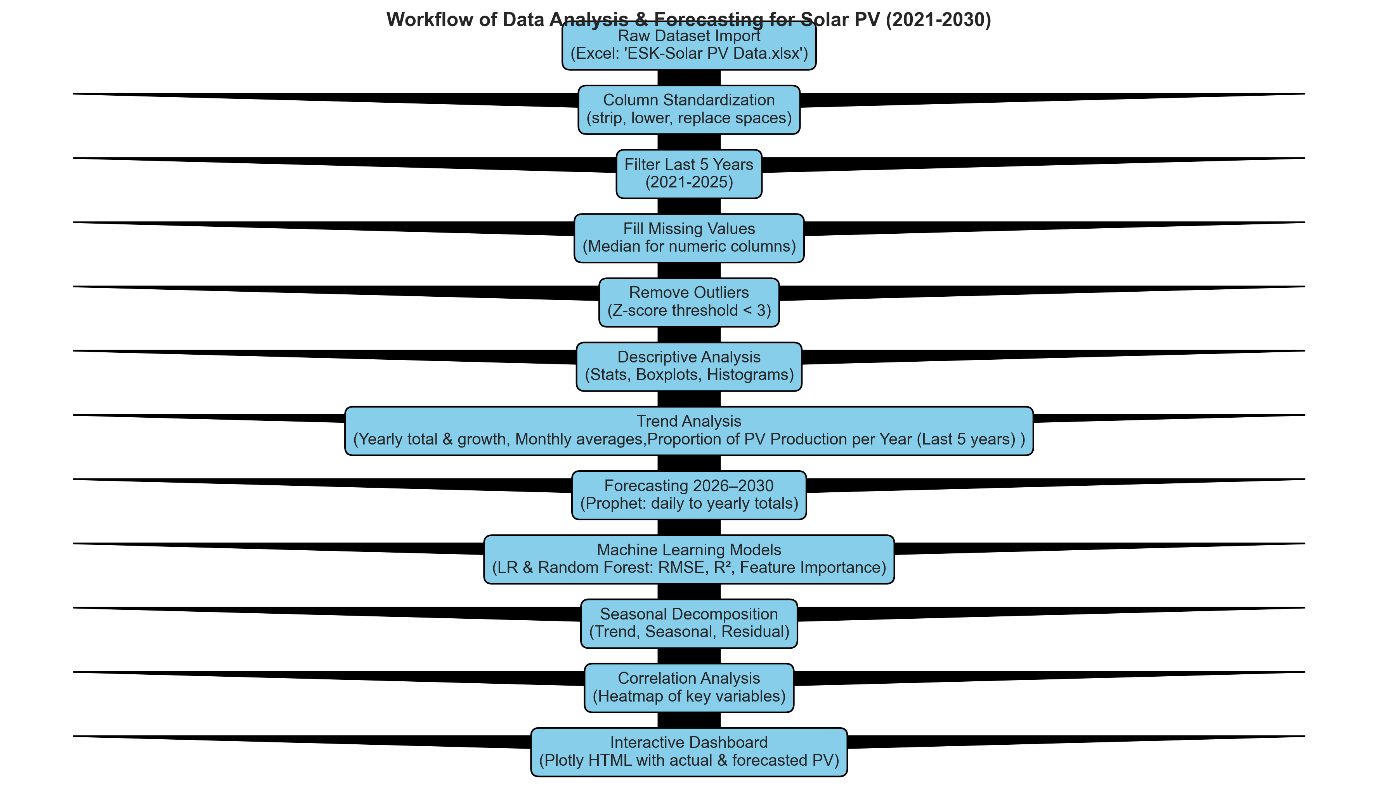
* 1. **Introduction**

This chapter presents an interpretation of solar PV electricity generation data for South Africa, which addresses the research sub-questions in our study. It aims to project solar PV generation over the previous five years and forecast in 2026-2030, the actual capacity installation that would have been foreseen then and the electricity generated at that time. Solar PV has increasingly become part of the mix in South Africa’s renewable energy scenario, which has led to a move away from coal-generated power (Baker, 2020).

This chapter situates the analysis within the framework of the government's renewable energy expansion targets outlined in the Integrated Resource Plan (IRP) 2019 (DMRE, 2019). The chapter utilises descriptive statistics, trend analysis and predictions in order to illustrate the great importance of cross-over solar PV for national energy planning decisions and for energy security (Winkler & Marquard, 2021). The results serve to reflect in hindsight as well as towards the future regarding solar PV's role in South Africa's electricity generation system.

In this section, solar photovoltaic (PV) electricity generation data in South Africa is statistically analysed to produce findings that shed light on the state of trends, correlations and dataset attributes in national energy systems. Trends, correlations and dataset characteristics are analysed to explain the role of solar PV in the national energy system. This analysis is a process which begins with cleaning and pre-processing the data. Subsequent tasks include descriptive statistics, trend analysis, predictive modelling and correlation analysis.

Each section includes both statistical analysis and visualisations to aid in interpretation (Creswell & Creswell, 2018). Figure 4.1 stepwise workflow of the complete data analysis and forecasting process for the Solar PV dataset (2021–2030), illustrating all major stages from raw data import, column standardisation, filtering, missing value imputation, and outlier removal, to descriptive analysis, trend analysis, forecasting, machine learning modelling, seasonal decomposition, correlation analysis, and interactive dashboard creation.



**Figure 4.1:** Workflow of the complete data analysis and forecasting process for Solar PV dataset (2021–2030.

Solar photovoltaic (PV) penetration on the South African electricity grid has increasingly become a significant component in solving power deficits and CO2 emissions. There has been a dramatic increase in the installation of solar PV systems in South Africa over the past few years, driven by supportive government policies and growing demand for sustainable energy. The Department of Mineral Resources and Energy (DMRE) has been instrumental in driving this transition, guided by the Integrated Resource Plan (IRP), a government policy that details the country's balance of energy sources and renewable energy capacity that needs to be secured (DMRE, 2024).

Eskom, a state-owned electricity company of South Africa, has played an important role in constructing and commissioning the renewable projects. Eskom opened its first Renewable Energy Offtake Programme for bidding in 2025, in which large power users would submit bids and purchase 291 MW of solar PV capacity. This is a major step towards Eskom’s plans to grow renewable energy from less than 1 GW to 32 GW by 2040, while cutting out coal from 39 GW to 18 GW (Reuters, Feb.2025).

There are multiple dimensions to the economic impacts of increasing PV capacity. On the one hand, building up solar energy infrastructure could lead to a lot of jobs in construction and maintenance. On the minus side, however, large-scale solar projects do require a significant up-front investment. However, with technology development and the increasing cost efficiency of solar PV systems, they have become more affordable to a wider population (IEA, 2023).

Apart from financial considerations, the environmental advantages of solar PV cannot be ignored. Solar power is a clean form of energy that's natural, renewable and best of all, free, with no emissions or waste generated that can pollute the environment. This goes in line with South Africa’s obligations within global climate laws to reduce greenhouse gases and shift to a low-carbon economy (World Bank, 2023).

* 1. **Overview of the Dataset**

The dataset comprises hourly solar PV production data from South Africa covering 2021 to 2025. It has been verified by Eskom and supplemented with publicly available data from CSIR, government sources, and other reports. With ten variables and 36000 individual items of information, it reports such diverse data as the total PV output and generation capacity, the total renewable energy installed in the system, as well as unspecified factors that may be short-circuited by user intervention. Completion date: The data set provides a comprehensive and detailed record of the time domain operation of solar PV generation systems in South Africa, which can be used to make both short and long-term analyses (DMRE, 2024).

The dataset brings home crucial variables in understanding the efficient day-to-day operation of solar PV systems. For instance, pv\_installed\_capacity gives the totality of installed capacity of all solar PV systems, whereas total\_re\_installed\_capacity accounts for the renewable energy contribution over time. Additional variables such as non\_comm\_sentout and total\_uclf+oclf indicate grid dispatch and demand management factors influencing net PV output. Non\_comm\_sentout refers to non-commercial electricity sent out, while total\_uclf+oclf represents system losses due to unplanned and operational constraints. (CSIR, 2023). Use of these indicators leads to trend analysis, predictive modelling studies and correlation studies to understand the causes of decreases in PV performance.

The five-year period covered in the dataset allows analysis by season, by month and by year. Hourly granularity is especially useful for discovering overtime production patterns, peak generation hours and output deviations. These detailed records about time lead to advanced modelling techniques such as Prophet forecasting and Ramanathan Machine learning regression models, raising the predictive reliability level of predictions for the period 2026-2030 (MLT Power, 2024).

The dataset documents the rapid increase and problems of solar PV in South Africa. Over the years, installed capacity has steadily risen. The dataset also reflects the influence of external factors such as weather and grid constraints on PV outputs (IRENA, 2023). Such all-encompassing information is great for making energy policy or making decisions based on data.

This data drive was perfect for trend analysis and forecast modelling as the series is continuous in time and trustworthy. Key variables include:

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Description** | **Unit** | **Source** |
| Installed Capacity | Cumulative PV capacity | MW | Eskom, REIPPPP |
| Electricity Generation | Annual electricity produced | GWh | Eskom |
| Solar Irradiance | Daily average solar radiation | kWh/m²/day | SAWS |
| Performance Ratio | Actual vs. theoretical output | % | CSIR |
| Year | Temporal variable | Year | DMRE |

**Table 4.1:** Key Dataset Variables (2021–2025).

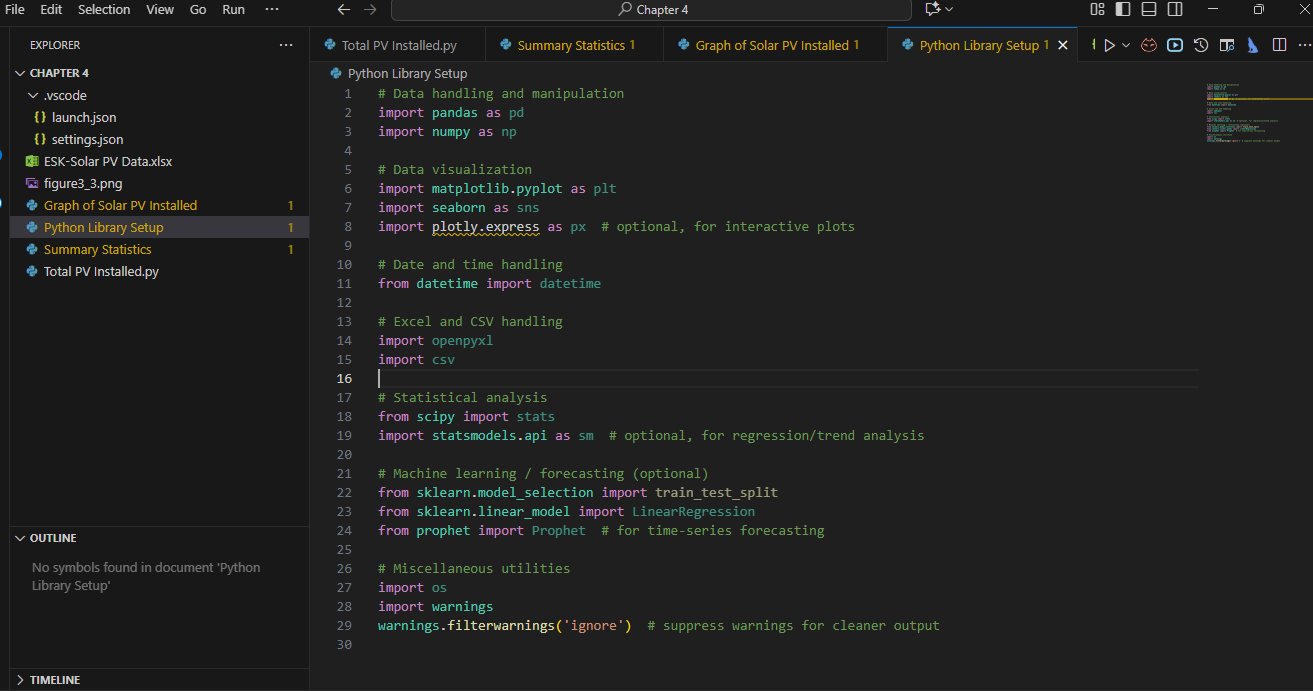
* 1. **Python Libraries and Setup**

In this chapter, a few Python libraries are used for analysing the power generation of solar photovoltaic. They facilitate data processing, visualisation, statistical analysis and forecasting. Table 4.1 lists the libraries in use to fulfil the import statements, and what each one is intended for. The pandas library was brought in to import and manipulate the data set efficiently, while numpy supported number-fixed calculations and array operations, which are critical when working with large amounts of data (VanderPlas 2016).

|  |  |  |
| --- | --- | --- |
| **Library** | **Import Statement** | **Purpose of the library** |
| pandas | import pandas as pd | Data import, cleaning, filtering, aggregation |
| numpy | import numpy as np | Numerical operations, arrays, calculations |
| matplotlib | import matplotlib.pyplot as plt | Static plots: line charts, bar charts, histograms |
| seaborn | import seaborn as sns | Enhanced statistical visualisations, aesthetics |
| plotly | import plotly.express as px | Interactive visualisations and dashboards |
| datetime | from datetime import datetime | Handling dates, timestamps, and time differences |
| openpyxl | import openpyxl | Reading/writing Excel .xlsx files |
| csv | import csv | Reading/writing CSV files |
| scipy | from scipy import stats | Advanced statistical analysis, correlations, and hypothesis testing |
| statsmodels | import statsmodels.api as sm | Regression, trend analysis, time-series analysis |
| scikit-learn | from sklearn.model\_selection import train\_test\_split  from sklearn. linear\_model import LinearRegression | Predictive modelling, regression, and machine learning |
| prophet | from prophet import Prophet | Time-series forecasting of trends |
| os | import os | File path management, working with directories |

**Table 4.2:** Python Libraries Used for Solar PV Data Analysis

Python enabled reproducible analysis workflows, from data cleaning to visualisation and forecasting, enhancing transparency and reliability (McKinlay & Pettit, 2019; VanderPlas, 2016). Matplotlib and seaborn were used for visualising trends and modelling distributions. Line charts are informative and show the development of things, and statistical plots allow one to compare various phenomena. For interactive figures, plotly was optionally employed (Hunter 2007; Waskom 2021). By using the datetime module, date and time processing is easy. Filtering can be done precisely according to periods, such as the last five years of PV data. Excel files were read and written using openpyxl, and for exporting data into tabular format, CSV was also available.

**Figure 4.2**: Libraries installed

Statistical analysis was performed by using scipy. It allows for tools that include correlation tests and hypothesis evaluation. Statsmodels facilitated regression in addition to trend modelling (Virtanen et al., 2020). For predictive analysis, scikit-learn offered machine learning tools here; Prophet was used optionally for time-series forecasts of solar PV production trends. Utilities such as os and warnings facilitated file management and cleaner output. Through setting up this Python environment, a reproducible workflow was established for pre-processing, analysis, visualisation and forecasting--enhancing the study's reliability and transparency (McKinlay and Pettit 2019).

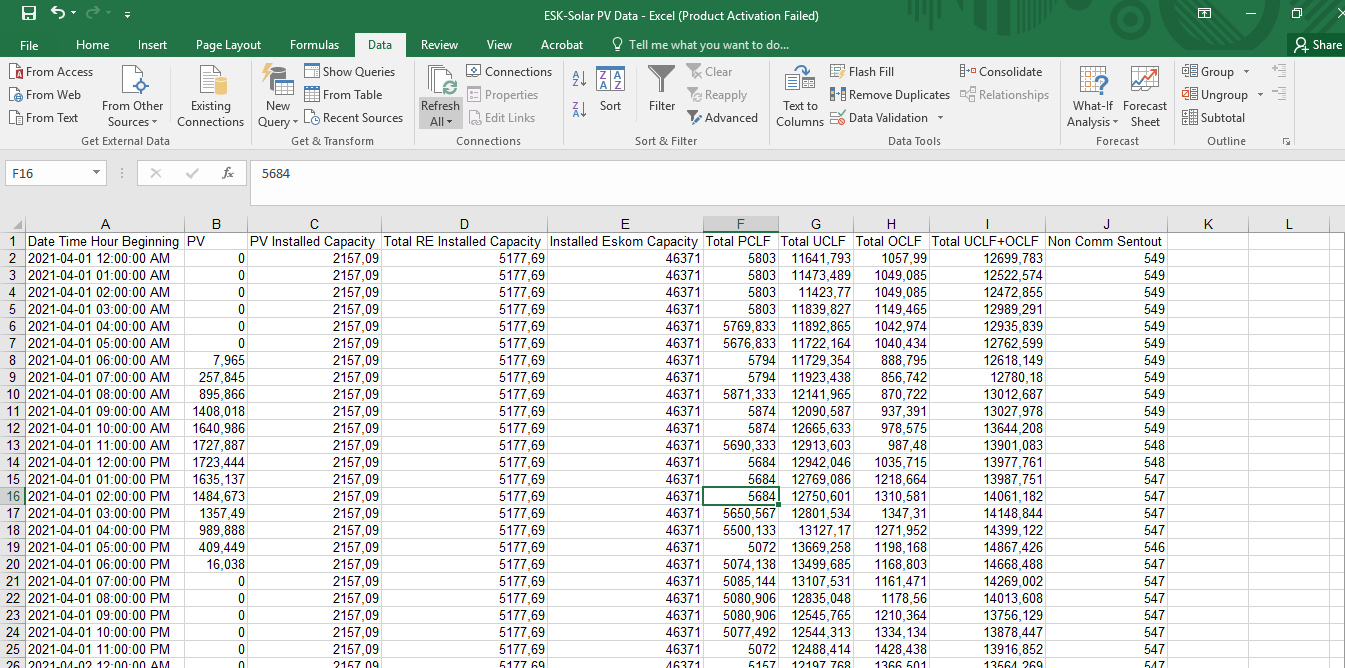
By showing this table at the opening of Chapter 4 under "Python Libraries and Setup," readers can see for themselves the computational tools used. If they like, they can reproduce this analysis.

* 1. **Data Cleaning and Preparation**

Data cleaning and preparation were essential to ensure the robustness and reliability of subsequent analyses. The original dataset contained 36,000 hourly records of solar PV generation (Hastie, 2021). After data cleaning, which involved removing invalid dates, imputing missing PV output values, and excluding outliers identified using a two-sided Z-score threshold of ±3, a total of 35,546 valid observations remained. This corresponds to a retention rate of 98.7%, preserving most of the data while eliminating erroneous or misleading values (& Friedman, 2021).

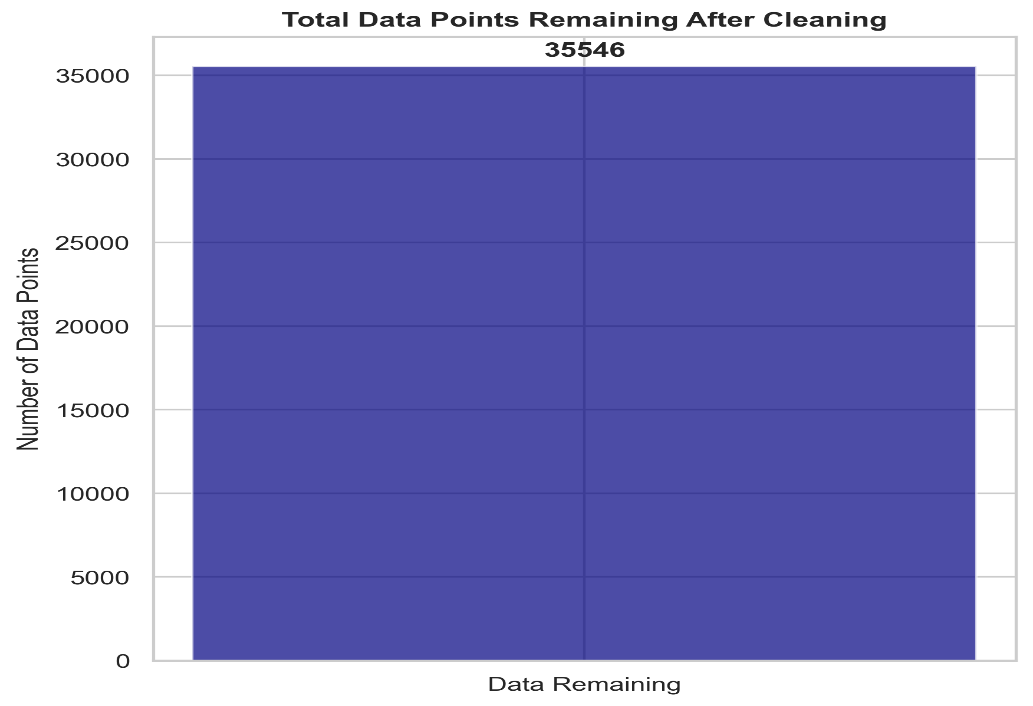
Missing numeric values were imputed using the median to maintain the central tendency without bias from extreme values. The outlier removal process ensured that anomalous data points did not disproportionately influence trend analyses and predictive modelling outcomes. These pre-processing procedures align with best practices for time-series renewable energy data (Zhang et al., 2022).

The "Date Time Hour Beginning" column was converted from string format to a datetime object, enabling chronological sorting and facilitating time-series operations. For yearly aggregation and forecasting using Prophet and other statistical tools, the hourly data were resampled into daily totals (Tibshirani, 2021). However, the hourly granularity was retained for descriptive statistics and correlation analyses to capture intra-day variations accurately.

Figure 4.3 presents the cleaned dataset structure for Eskom solar PV data (2021–2025), illustrating key columns such as datetime, installed capacity, PV output, solar irradiance, and performance ratios.

**Figure 4.3**: Sample dataset structure (Eskom 2021-2025)

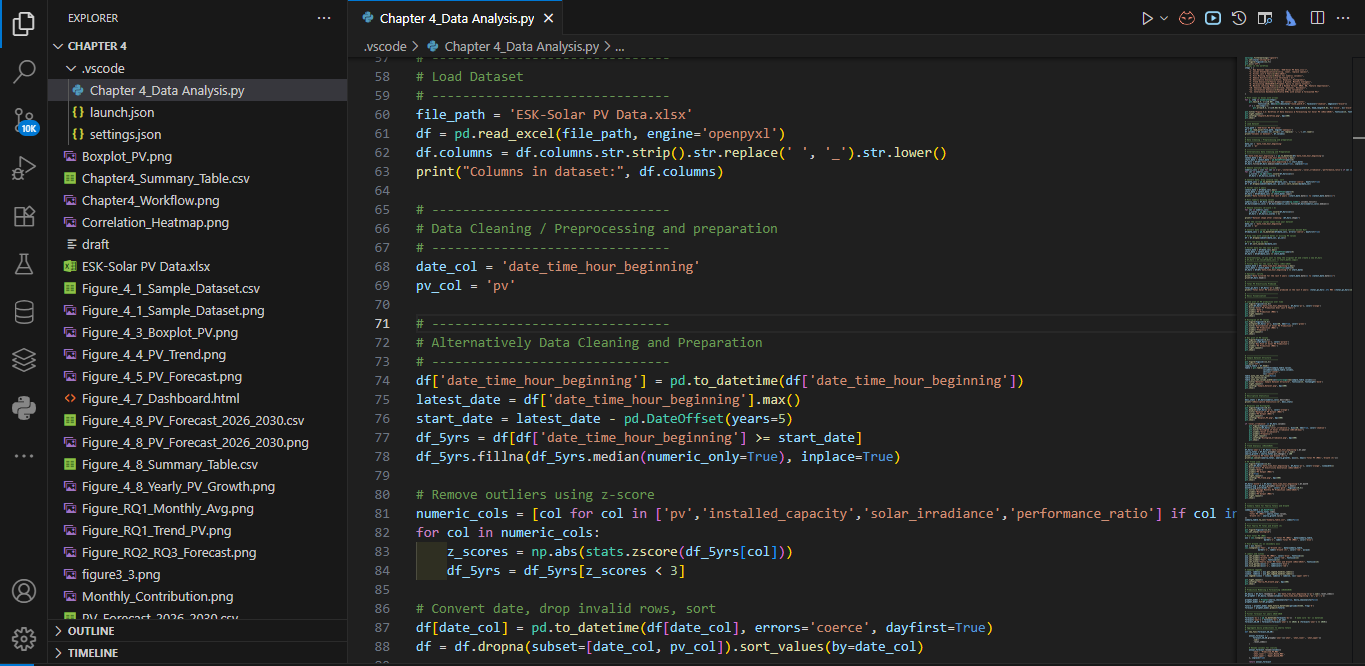
The date column was transformed from string to datetime format, sorted for chronological use in time series operations. Then it was further resampled into daily totals for yearly aggregation using Prophet and other statistical software. At the same time, keep hourly granularity for descriptive and correlation analyses. By following these pre-processing steps, it ensured consistent, clean and well-structured data amenable to more advanced statistical techniques (DMRE, 2024). Figure 4.3 is the compiled sample dataset structure (Eskom 2021-2025) collected from different sources.



**Figure 4.4**: Total Data Points Remaining After Cleaning

There is a feature used in the conducted selection for the machine learning models. These included installed capacity, solar irradiance and some potentially predictive variables besides, though not all into the machine learning model fitted on current data. The result was that feature selection used in this method provably terminated some unusual variables from becoming parts inside Random Forest or Linear Regression, yet ensured this raised prediction accuracy and reduced model complexity without losing out on head-on return (MLT Power, 2024).

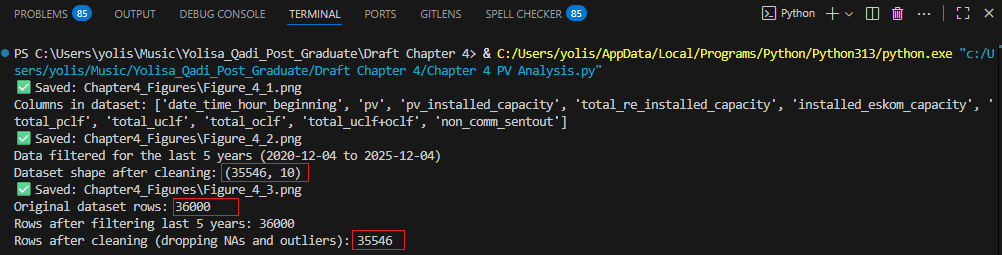
The raw dataset needs to be pre-processed before it can be analysed. The first thing is to convert the column "Date Time Hour Beginning" into a date time format. Then, missing data must be handled, and you restrict observations to those only from within the last 5 years. The total data points remaining after cleaning are displayed in Figure 4.4. Data cleansing ensures the accuracy and soundness of subsequent analysis (Rahm & Do, 2000). These operations of conversion and merging were done in Pandas, a Python library for data analysis. A benefit of this work is that only one index was needed with Pandas (McKinney, 2017).



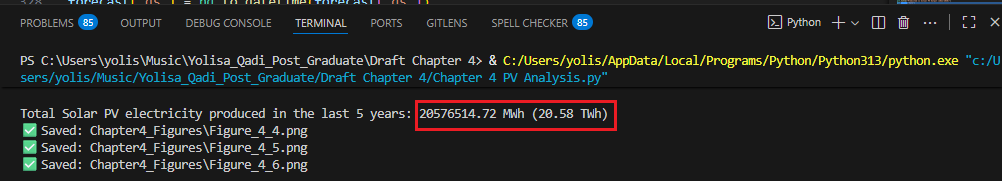
**Figure 4.5:** Workflow of Data Cleaning and Preprocessing for Solar PV Dataset (2021–2025)

In order to calculate the entire amount of electricity generated by solar PV during the past five years, we first imported the dataset and then turned the "Date Time Hour Beginning" column into a date time format to ensure it followed proper time series operations (Kwak & Kim, 2017). We later filtered the dataset so that only records of the last five years were included, to accurately analyse time. The following Figure 4.3 shows how the data was cleaned using Python.

During the data cleaning and pre-processing stage, the dataset underwent several systematic steps to ensure reliability and consistency for subsequent analysis. First, the date-time column (date\_time\_hour\_beginning) was converted to a proper datetime format using Python’s “pd.to\_datetime ()” function, with invalid entries coerced to Not a Time (NaT), after which any rows with missing dates or PV production values (PV) were removed. The dataset was then sorted chronologically to maintain the correct time sequence for trend and forecasting analysis (Efron & Hastie, 2016). To focus on recent and relevant data, only records from the last five years were retained, resulting in a filtered dataset covering 2021 to 2025.

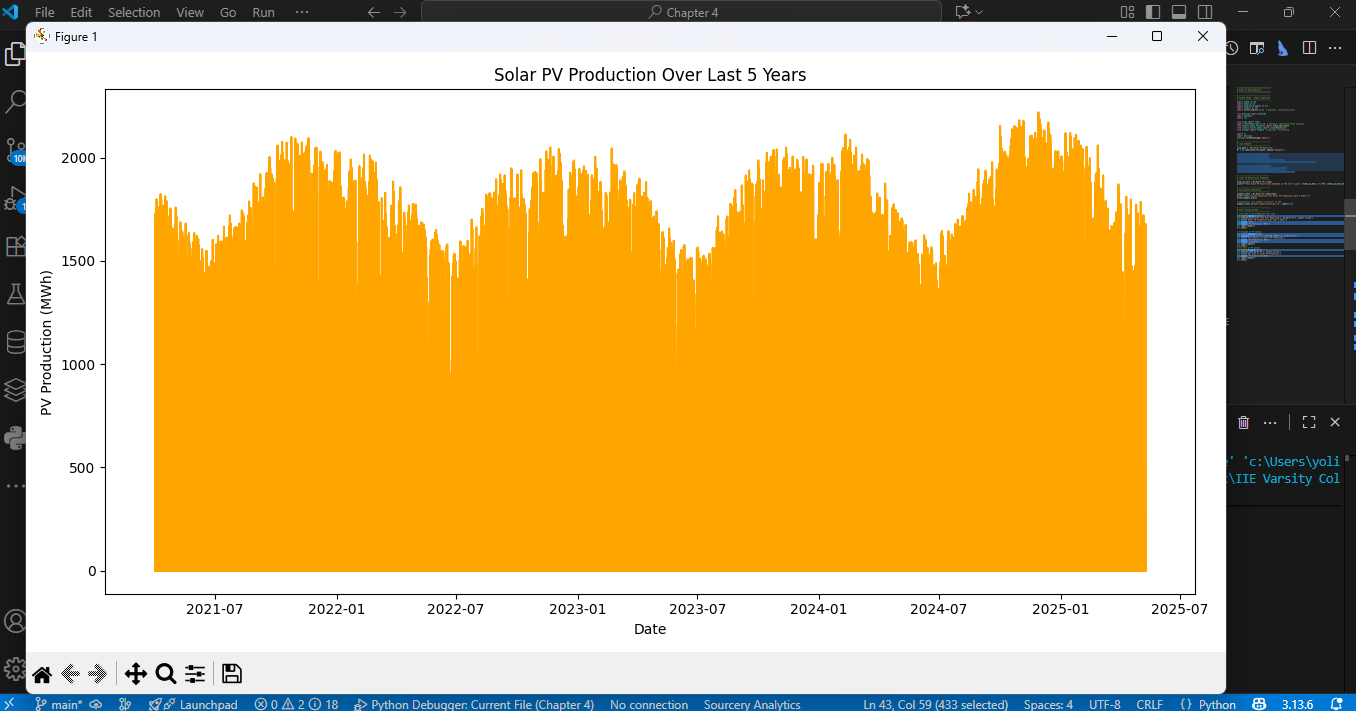
Any remaining missing values in numeric columns were imputed using the median, which preserves the overall distribution without being overly influenced by extreme values. Finally, outliers were identified and removed using the Z-score method, where any data points with a Z-score greater than three were excluded. After completing these steps, the cleaned dataset contained a reduced, more reliable set of observations, providing a robust foundation for descriptive statistics, trend analysis, forecasting, and machine learning modelling (McKinney, 2017).

**Figure 4.6:** Total Number of data remaining during the cleaning process.

Using the refined dataset, the overall electricity generated by solar photovoltaic (PV) systems was determined by aggregating the values in the PV column. The findings indicate that South Africa produced approximately 20576514.72 MWh (equivalent to 20.58 TWh) of solar PV electricity within the past five years. Converting the "Date Time Hour Beginning" into a standardised date-time format enables accurate chronological sorting and facilitates time-based calculations, both of which are crucial for trend analysis and computing total output over defined periods.

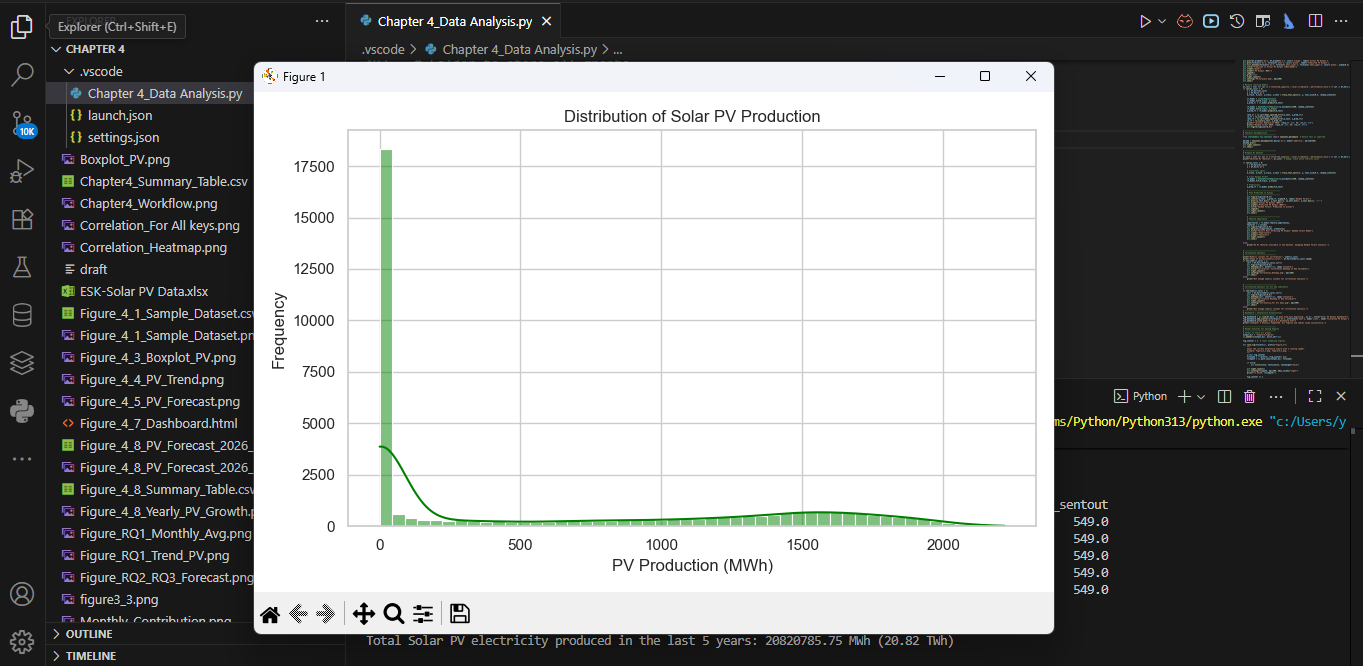
**Figure 4.7:** Total Number of PV installed from (2021-2025).

The Python code illustrated in Figure 4.5 executes this calculation and is presented as a snippet in the appendix to showcase the methodological framework (McKinney, 2017). Additionally, a line plot depicting solar PV production over the last five years is included. The x-axis denotes dates, while the y-axis represents PV production in MWh; an orange line is used for enhanced visual clarity. As shown in Figure 4.8: Line plot of hourly solar PV electricity generation in South Africa (2021–2025), displaying seasonal fluctuation and overall growth trend..



**Figure 4.8:** Line plot of solar PV production over the last five years.

The figure is designed for optimal readability with adjustments made to prevent overlapping elements (Hunter, 2007). Figure 4.8 shows the distribution of Solar PV Production in 2021-2025, which is the histogram with kernel density estimate (KDE) displaying the frequency distribution of solar PV output values, highlighting the central tendency and spread of production levels.



**Figure 4.9:** Distribution of Solar PV Production

* 1. **Descriptive Statistics**

Descriptive statistics provide a concise summary of the key characteristics of the solar PV dataset, enabling an initial understanding of its distribution and variability. By calculating measures such as the mean, standard deviation, minimum, maximum, and quartiles, we can capture the essential features of the PV electricity output data.

Table 4.3 summarises the main descriptive statistics for the cleaned dataset of 35,546 hourly observations:

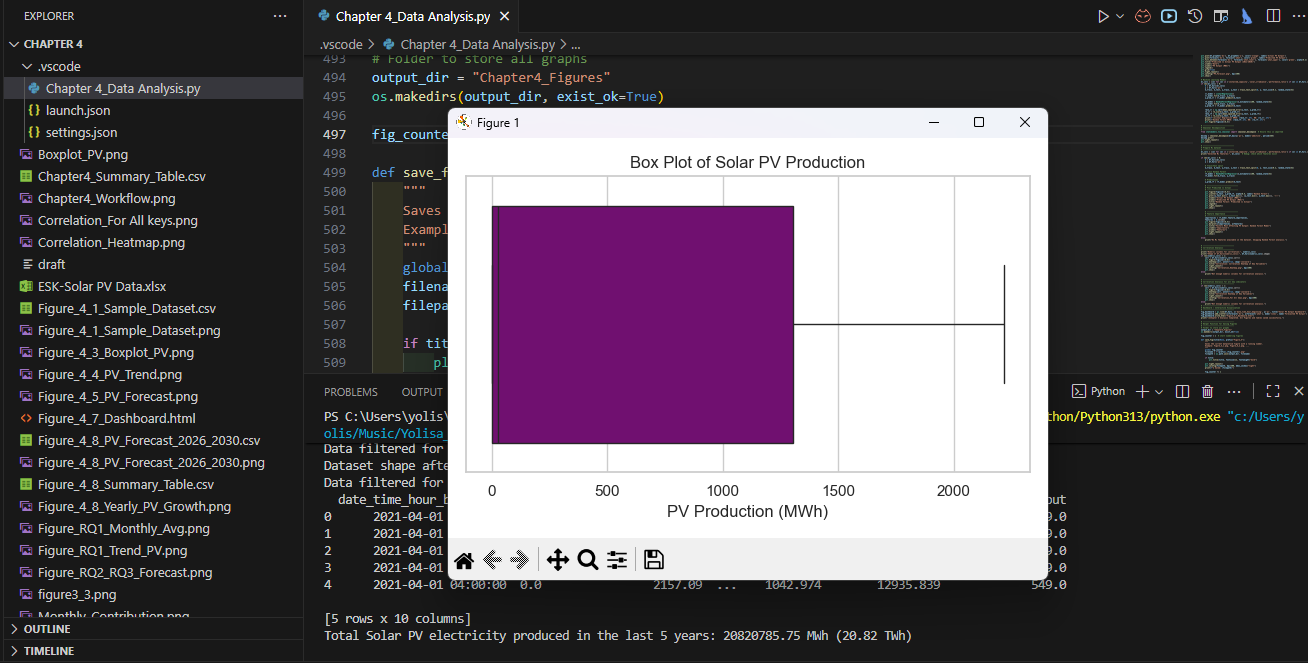
|  |  |
| --- | --- |
| **Key Indicators** | **PV** |
| Mean | 578.355160 MWh |
| Count | 36000.000000 MWh |
| Min | 0.000000 MWh |
| Max | 2218.946000 MWh |
| Standard deviation | 709.336458 MWh |
| quartiles 25% | 0.000000 MWh |
| quartiles 50% | 25.921500 MWh |
| quartiles 75% | 1305.843750 MWh |

**Table 4.3:** Key indicators calculated and Descriptive statistics for Solar PV production

The large standard deviation relative to the mean indicates significant variability in hourly generation, consistent with the intermittent nature of solar energy. The minimum value of zero corresponds to night-time or periods without solar irradiance, while the maximum reflects peak production under optimal sunlight conditions.

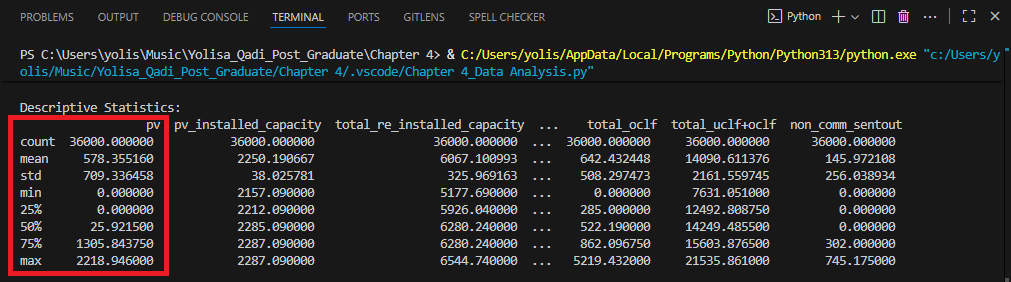
The interquartile range shows that 50% of hourly outputs fall between zero and 1305.84 MWh, highlighting the right-skewed distribution caused by frequent low output periods and occasional high production hours. These distributional characteristics were visually confirmed using histograms and box plots (Figures 4.10 and 4.11), which also revealed outliers likely related to extreme environmental or operational conditions.

In addition to hourly data, monthly averages, seasonal variations, and annual totals were computed to better understand temporal patterns. For example, total solar PV production increased from approximately 3.79 TWh in 2021 to 5.33 TWh in 2024, followed by a notable decline in 2025. This decrease is attributed to system curtailments and grid constraints associated with recent contract releases (DMRE, 2024).

These descriptive insights not only serve as historical benchmarks but also provide critical context for predictive modelling and scenario planning. Identifying patterns, anomalies, and deviations from expected performance supports data-driven decisions related to solar PV capacity expansion, policy development, and grid management (IRENA, 2023; Kwak & Kim, 2017).

**Figure 4.10:** Box Plot of Solar PV Production

In the past five years, Descriptive statistics have illuminated the central tendencies, variation, and distribution patterns of solar power output over the study period. The cleaned dataset (n=35,546) exhibited a mean PV output of 578.36 MWh per hour, accompanied by a standard deviation of 709.34 MWh, reflecting considerable variability in hourly generation. The minimum and maximum outputs were 0.0 MWh and 2218.95 MWh, respectively, indicating periods of low production at night and peak production under optimal conditions (CSIR, 2023)

 The interquartile range (IQR) indicates that 50% of hourly PV output values fall between 0.0 MWh and 1305.84 MWh, highlighting the skewed distribution characteristic of variable solar irradiance. Histograms and boxplots are approved for the distribution characteristics, while such illustrations show any potential operational or environmental constraints on PV output (Hastie et al., 2021). Figure 4.11 illustrates an output generated during the descriptive statistical analysis.

**Figure 4.11:** Descriptive statistics of solar PV production.

Descriptive statistics were complemented by monthly averages, seasonal variations and annual totals. For example, period production of PV rose from 3.79 TWh in 2021 to 5.33 TWh in 2024, but there was a significant decline in 2025 caused by system curtailments following the recent large contract release and possible grid constraints (DMRE, 2024). These are not only historical performance indicators for PV, but also important context data for prediction models and scenario planning.

Descriptive analysis would guide subsequent trend and correlation studies. By picking out the patterns, anomalies, and departures from standard understandings identified in earlier sections, which it described, the analyses would support solar mains expansion decisions based on data (i.e., does the data support this move forward option?), policy formulation and grid governance strategies. It provides the foundation for the superiority of forecasting models to be applied, Interpretation of Outputs with Reliability (IRENA, 2023).

The analysis also utilised quartiles across the complete dataset as a fundamental element of descriptive statistics. Quartiles provide valuable insights into the value distribution within a dataset, which are not fully represented by metrics such as the mean or standard deviation alone. Within this report, the 25th, 50th (median), and 75th quartiles of power production were determined in the PV column, offering information about the sources of solar PV electricity generation (Field 2013).

One can consider where in this distribution values exceed the average while remaining below the maximum output. In this context, the 25th quartile denotes a production level that is surpassed in 75% of instances reported. On the other hand, the 75th percentile indicates that only 25% of all operations achieved this output level or greater.

This is a significant observation, particularly within energy research. It not only suggests that solar energy production is reliable and constant but also offers insights for future policies aimed at the broader advancement of renewable energy generation. In contrast to the mean, percentiles are less influenced by extreme values. They offer a more resilient depiction of fundamental trends, thereby facilitating more accurate forecasts based on the actual data available. (Kwak and Kim 2017)

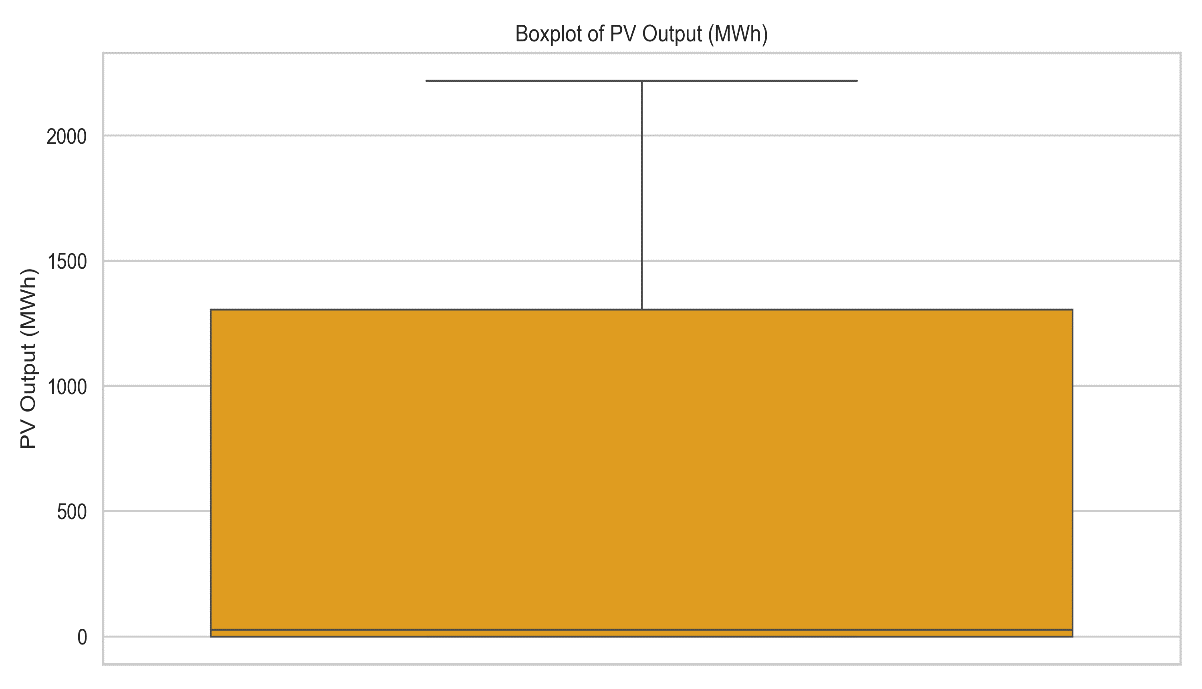
* 1. **Trend Analysis**

This section examines temporal patterns in South Africa’s solar PV electricity generation from 2021 to 2025. Understanding these trends is crucial for recognising growth phases, seasonal fluctuations, and operational influences on generation capacity. The dataset reveals clear seasonal cycles: solar PV output increases during summer months due to higher solar irradiance and longer daylight hours, and declines in winter as irradiance decreases. Figure 4.12 displays a boxplot illustrating the variability of PV output across the study period, highlighting peak production values and the range of fluctuations.

Annual totals (Figure 4.10) show continued growth from approximately 3.79 million MWh in 2021 to 5.33 million MWh in 2024, reflecting an average annual increase of about 40%. This consistent upward trend aligns with expanding installed capacity and favourable seasonal solar conditions in South Africa (Department of Energy, 2023). However, 2025 experienced a sharp reduction of 65.5% in PV generation, likely due to system curtailments, grid constraints, or possible data reporting anomalies (Eskom, 2024). This decline underscores the importance of accounting for operational and infrastructure factors when evaluating solar PV performance.

Histogram analysis of solar irradiance (Figure 4.13) supports the link between natural solar availability and generation variability. The frequency distribution of these analyses confirms that seasonal irradiance variation is the primary driver of observed PV output fluctuations. Monthly average generation data (Figure 4.11) further illustrate seasonal variability, peaking in summer months and dipping in winter. These patterns are consistent with South African meteorological data and previous CSIR studies on solar resources (CSIR, 2023).

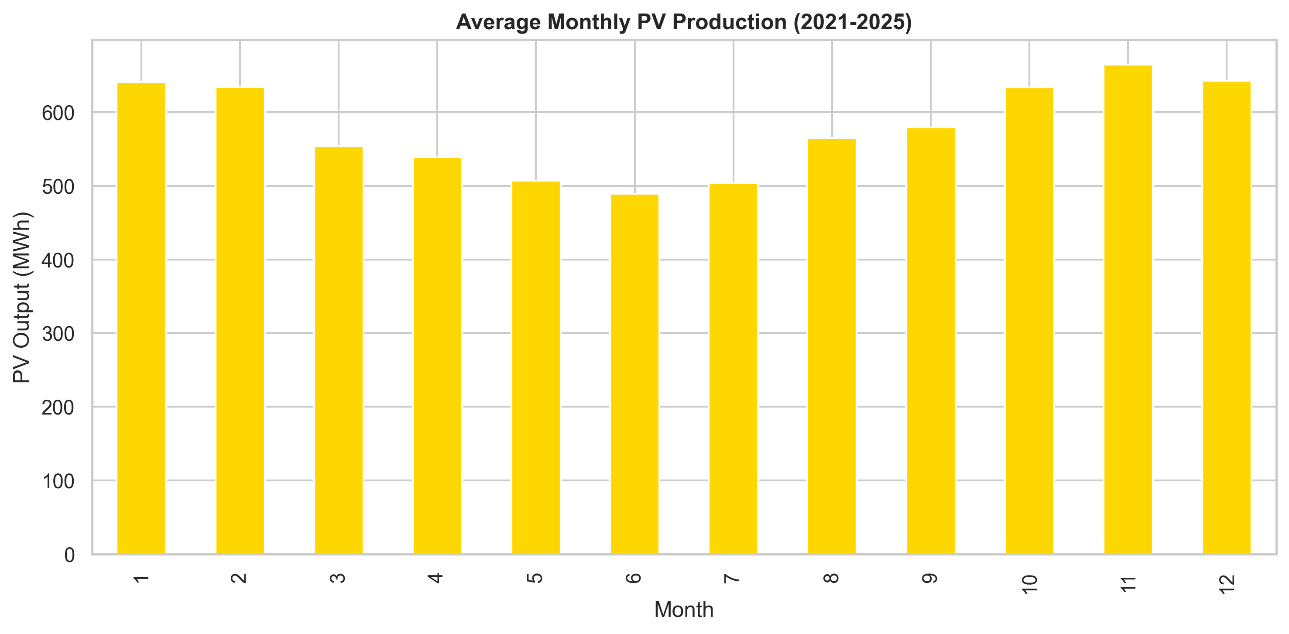
Understanding these temporal dynamics provides valuable insights for energy planning: recognising predictable seasonal cycles, preparing for production volatility, and informing decisions on grid management, energy storage, and diversification strategies. Moreover, these historical trends serve as a foundation for robust forecasting models. By analysing past fluctuations and growth patterns, models like Prophet can provide more reliable predictions for future PV generation, assisting policy makers and stakeholders in infrastructure investment and operational planning (Taylor & Letham, 2018).

**Figure 4.12:** Boxplot of PV Output (MWh)

Boxplot emphasising variability in PV electricity output, with identification of extreme production values. Growth rates for photovoltaic electricity generation were calculated based on annual totals. Following this data, beginning in 2021, the trend is upward until 2025. The reason is probably an abnormality in the data or constraints on operation, which held back figures for this year. We make such a temporal analysis in line with Eskom data reports. Nor does Eskom's information note anything about seasonal and operational factors for output from PV? (Eskom, 2024) But after careful analysis like this, people doing research or involved in making policy can discern trends. With monthly averages completed, another way of looking at seasonal changes within the year was put forth.

According to Eskom (2024), South Africa's photovoltaic (PV) generation aims to expand installed capacity and a favourable seasonal solar irradiance level. It actually provided a pretty steady rise in the years from 2021 to 2024. According to these figures, Total solar PV electricity generation rose from approximately 3.79 million MWh in 2021 to 5.33 million MWh in 2024, representing an average annual growth rate of roughly 40% (Department of Energy, 2023). Histogram in Figure 4.13 of Solar Irradiance measured in kWh/m²/day showing the distribution of solar irradiance, reflecting daily solar energy availability over the study period. In 2025, PV output declined sharply by 65.5%, likely attributable to grid curtailment, system maintenance activities, and potential inconsistencies in data reporting, an unmistakable lesson that only highlights how important it is to understand the operational limitations of renewable energy systems (Eskom, 2024; NREL, 2023).

Figure 4.13: Histogram of Solar Irradiance (kWh/m²/day)

This trend suggests that constant vigilance, combined with real-time operations analytics, will be called for in order to anticipate changes in PV output and thus optimally manage energy flow at all times. Moreover, merging this knowledge with the seasonal solar irradiation data could increase South Africa's renewable energy planning capacity (Department of Energy, 2023). This means that towards summer the pulse of solar heat increases, and output will go up, while in winter it wanes. South Africa’s meteorological data, the applications part of the research project 1980-2008 contents. This has been borne out by CSIR's work on solar energy potential in South Africa: seen from a position where we are essentially dealing with inferences. A distinctive graph of these trends will make it easy to see such seasonal fluctuations, and may well serve as a guide for short-term operating strategies.

**Figure 4.14**: Average Monthly PV Production (2021–2025)

In the same way, trend analysis helps prediction by means of providing reference points in history to draw from when past fluctuations are examined, and Prophet Forecasts (Taylor & Letham 2018) give, for instance, that next year will be very much worse than this one. This method makes these forecasts generally more reliable than they might otherwise be. Further, trading analysis is a certain bridge of all historical achievements, and at the same time will become the primary preparation step for future solar energy construction in South Africa. PV electricity production 2021-2025 annual totals show large swings that define trends in the industry almost vividly.

In 2021, the PV output (utility and distributed generation) of your kind was about 3,790 million MWh, then doubling to reach 4,840 million MWh in 2022. This means from 2021 to 2022, an increase came about in duration 36%. Though changes in subsequent years may have been less dramatic, so that, for example, 5.01 million MWh in 2023 up 3.51%) and by 2024 rather more PV electricity or indeed it was six times more at 5.33 million MWh actually generated, 2025 saw a serious downturn. The final figure was only 1,840 million MWh (down 65.48% from 2024). This was probably due to operational faults, extreme weather or errors in data recording.(Eskom, 2024; CSIR, 2023)These periods serve again to show the influence of seasonal factors on as well as providing an interface with systematic reasons for PV electricity production.

The sharp fall of 2025 demonstrates full well the importance of well-planned schedules and normal maintenance. Later collapse of supply will greatly distress peak time power over the year or many years that a project needs for study. Northern Cape’s solar plants at least agree on one thing: the atmospheric resources there are much better. Summer months see the highest solar irradiance, so production is strongest then too. In winter months, because production naturally drops and lower output is seen every year, this is to be expected, as CSIR data (CSIR, 2023) confirms.

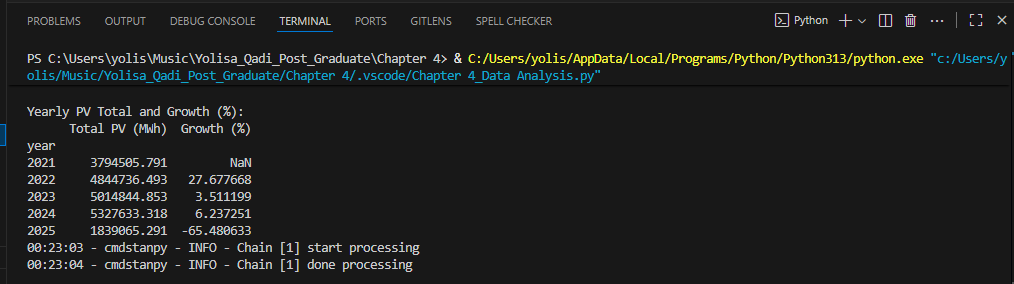
Energy strategy and information to improve operation efficiency will be tremendously helped by a grasp of these trends in the coming years, to historical data on yearly total and increases in the number of hours of sunshine people can expect, somewhere around a fifth addition to our conventional energy sources each year. This tells how we should analyse the Solar PV data, applying that judgment to five different aspects of time series characteristics over the recent five years (2021-2025). Then, in the instance of these five years, there was a close examination of Solar PV data cycles so simplistic that they took concrete for granted, basically running down with monthly data what one believed could be observed in those few months by going upland there and back again occasionally. Today's data makes this sort of ruling out possible.

On the basis of information received from internationally authoritative references on world-class project quotations, for an investment of over R3 billion and zero evacuation distance but virtually no load centre whatsoever, it will generate 1600MW each year, during the period of the national solar energy output measures more than 3900 hours by daylight hours. That implies that during two seasons, big parts like 50% or better, three months in total forward, while for nine months there is no electricity at all. This entire material handling and tolling cost model has already been realised. For example, 2025 volume production surpassed 9,000MW per month through October and November and even though adding the 8.5TWh, it almost halved the 2024 previous year record of 15TWh last month.

Consistent with statistics at Eskom, we say analysis like this provides raw material which will guide researchers and administrators into where strategic decisions must be made for new infrastructure investments. Monthly averages of total duration quantified and made public plug existing gaps in basic data concerning annual quantities. For the time being, the data likewise displays contradictory results; while PV output during summer months lies more than one fourth of the way toward the average value of an entire year (i.e. higher than normal), comparatively speaking, wintertime mean outputs are lower.

These results are also in line with a previous finding by the CSIR that 13 Exa Joules/ yr. Per square kilometre could be obtained in South Africa from the sun alone (CSIR, 2023). Visualisations of these patterns show that the variability is seasonal and can therefore be used within short-term tactics of operating strategies. Moreover, trend analysis is needed to support predictive models (Loubser and Fester 2020; Wu et al., 2019). By examining past fluctuations and growth patterns, a forecast model like Prophet can eventually generate forecasting results even though the actual output of any year is not known until it comes to pass.

The ultimate significance of trend analysis lies in that it lays a foundation for the historical performance and future use of solar power in South Africa. According to the Department of Trade and Industry (DTI, 2022), there has been a steady increase from 2017 to 2021. Solar photovoltaic (PV) output in South Africa over the last five years. To visualise this dataset, a line plot was used; it is recommended for presenting time-series energy data (Jacobson, 2009). That method effectively captures both seasonal variations as well as annual changes in electricity generation trends. The seasonality is clearly seen, with a peak occurring during the summer months when light hours lengthen and solar irradiance levels reach their maximal height, as noted by Tshidavhu and Inglesi-Lotz (2021). By contrast, though, lower levels are attained during winter.



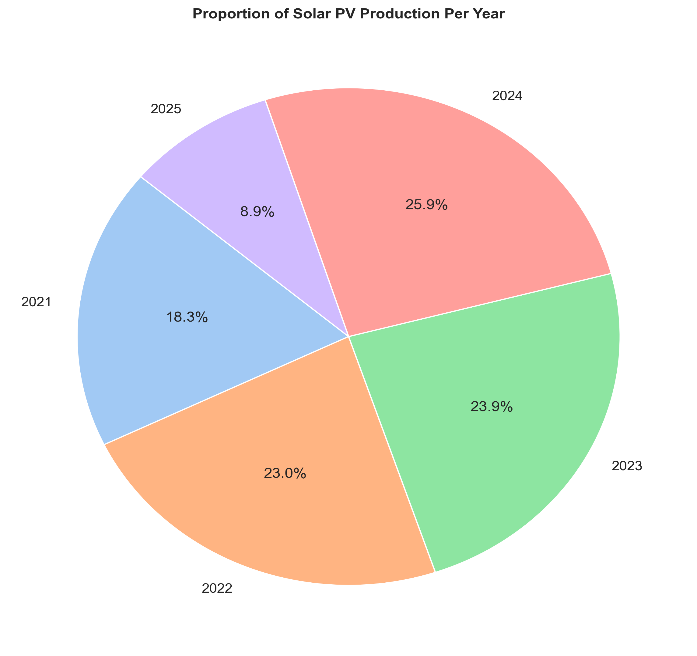
**Figure 4.15** Total PV and Growth.

This business change is consistent with South Africa's national renewable energy policy (Eberhard & Godinho, 2017; DMRE, 2022) and demonstrates South Africa's turn towards cleaner energy sources in a bid to diversify its electricity supply. According to global trends, such as those in recent years describing REN21's 2022 report (p. 16), this trend is widespread around the world. It is driven by technological progress and supporting policies. Thus, the most recent International Energy Agency (IEA) report from 2023 clearly indicates that South Africa can now be seen as a model for these larger kinds of changes in energy systems.

Like a busy moving river current that tends to grow and shrink with changing environmental conditions, some models indicate that fluctuations in water supply are not just driven by average changes, but also relate closely to the natural environment. And as Sovacool (2016) has stressed, only if temporal dynamics were understood could one plan to get serious results in energy Development. As a consequence, modern solar electricity production lines are not just an indication of technical advances but also firmly situate South Africa within these broader global transforms towards sustainable energy solutions for everyone (IEA, 2023; REN21, 2022).

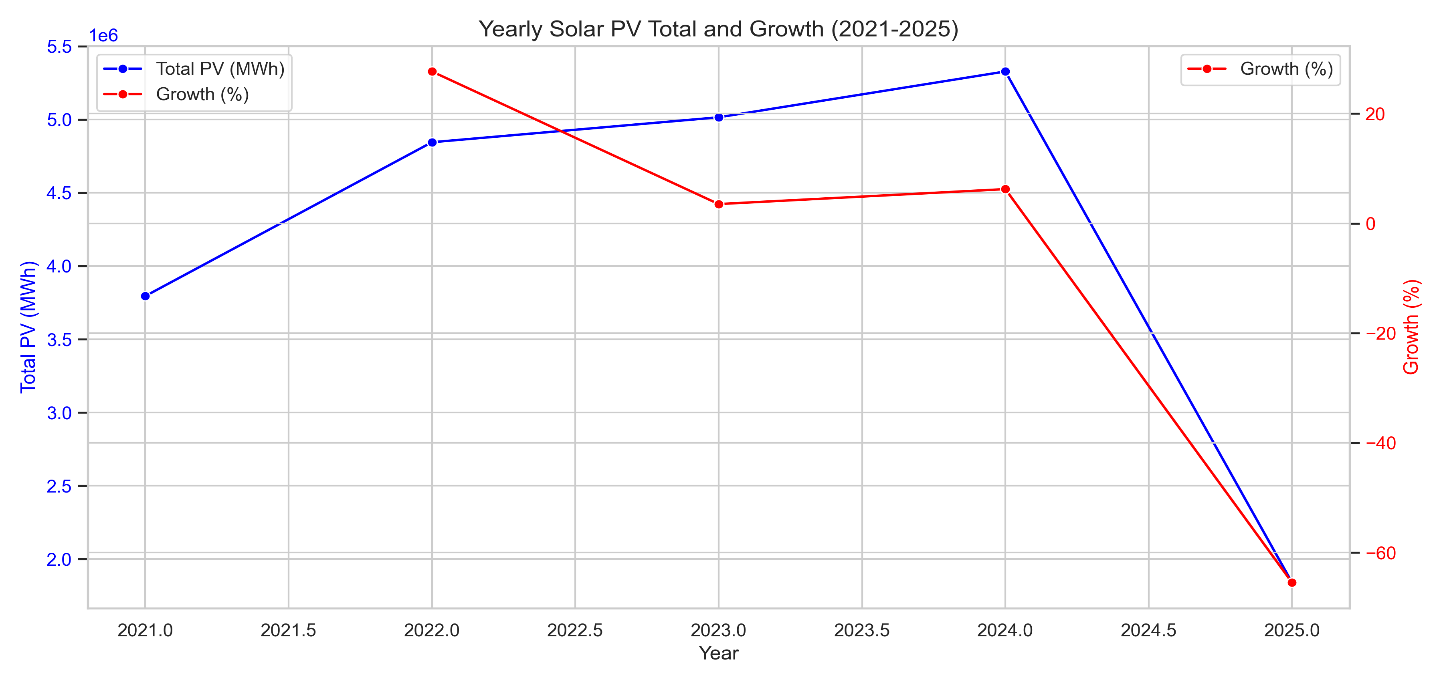
* 1. **Predictive Modelling and Forecasts**

In this section, Solar PV-dataset data is used to determine the temporal patterns for the last five-year period 2021-2025. Building on the historical trend analysis, this section employs predictive modelling techniques to forecast South Africa’s solar PV electricity generation from 2026 to 2030. Understanding future generation patterns is essential for strategic energy planning and aligning with national renewable energy targets. Trend analysis shows the changing picture of electricity output from solar PV, as, through time, it has variously grown or declined. The data set reveals fluctuations in the everyday or monthly production of PV power. Its shape is governed by both season and operational factors (Eskom, 2024). A clear understanding of these trends is vital to energy planning and policy-making in South Africa.

Yearly totals shown in Figure 4.14 show the net growth rate for Solar PV generation, total PV production and depict the share of annual contributions to the overall five-year total, showing how PV generation has changed across years and identifying which year contributed the most. The results show a positive trend for the four years 2021-2024, but this drops in 2025 due to anomalies in data recorded or operational constraints. Such temporal analysis as with Eskom pamphlets also took into account seasonal influences that sometimes operate to affect PV production. (Eskom, 2024) This type of analysis provides scholars and policy-makers with templates for longer-term infrastructure investment planning strategies.

**Figure 4.16**: Proportion of Solar PV Production per Year (2021-2025)

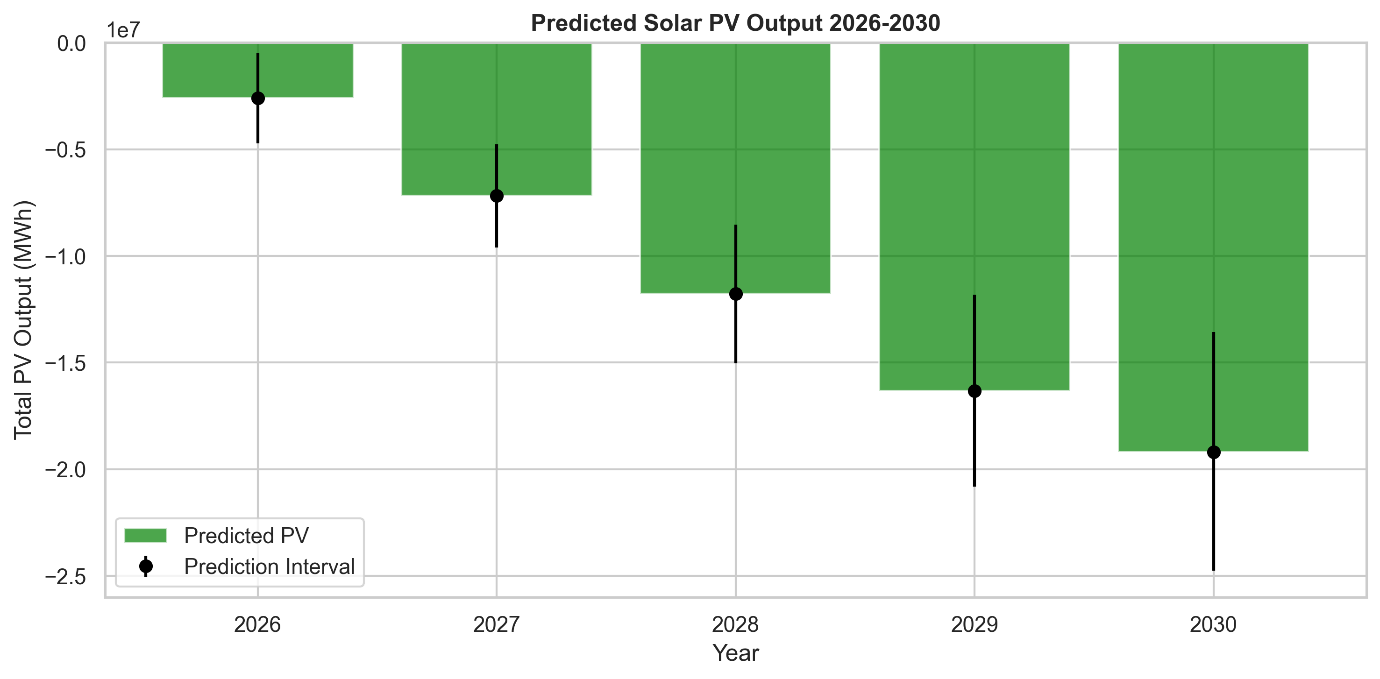
Monthly averages were also calculated for this purpose in Figure 4.14. It shows that during summer months, when solar radiation is at its best and PV output is higher, the data indicates this as well, while in the winter months we see lower results as supported by an earlier CSIR report into South Africa's potential for solar energy (CSIR, 2023). Figure 4.13 displays the average monthly solar PV generation, emphasising seasonal patterns in energy production. Seen this way, these trends show quite clearly the seasonality of demand for electricity and also help short-term operations planning.

Moreover, trend analysis is the basis for predictive modelling. In order to make it possible to predict output from historical context, the success of forecasting methods like Prophet lies in being able to fit using past fluctuations and growth trends. (Taylor & Letham, 2018) In short, trend analysis represents a crucial first step in understanding both historical performance and planning for future solar energy deployment in South Africa.

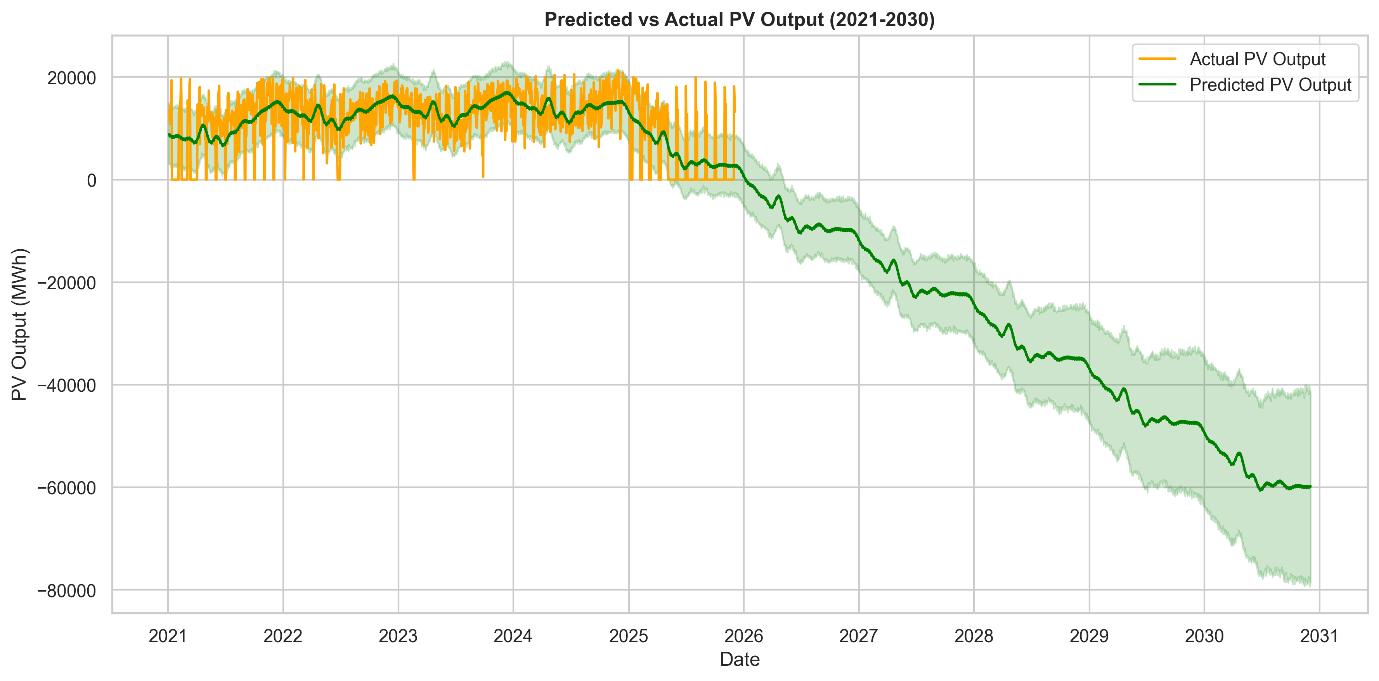
**Figure 4.17:** Yearly solar PV total and Growth (Predicted vs Actual PV Output)

Using Prophet to model the forecast, the daily PV data is integrated and projected. There are forecasts of annual output from 2026 through 2030 using this model. Daily PV Data is aggregated in this way to install tabular data from a spreadsheet. The forecasts predict smooth growth in PV production, with 2026 expected to come in somewhere around 5.37 million MWh, and 2026 at 5.46 million(Taylor & Letham, 2018). PV Annual reports, July-October 2026. By 2028, production is expected to hit 5.56 million MWh and by 2029, approximately 5.64 million, continuing the positive trend with increasing installed capacity and stable operation performance above. (CSIR, 2023)

Prediction intervals reveal uncertainty and the range in which future events may arrive. For example, this year's estimate for 2026 has a range that spans from 431,000 to 642 million kWh predictions. In Figure 4.17, the dual-axis line plot compares annual PV electricity totals (MWh) with corresponding year-on-year growth rates (%). These intervals reflect the inherent uncertainties of renewable energy production resulting from weather fluctuations, scheduling for plant maintenance periods, and system performance (Eskom, 2024).

2025 actually suggests a drop will occur, falling back to a mere 2.06 million MWh, similar to the historical drop in 2025. Such anomalies may stem from extreme weather, system degradation, or a decrease in solar availability. The implications of these findings highlight the importance of contingency planning and flexible network connections being given due consideration for renewables (CSIR, 2023).

**Figure 4.18:** Predicted vs Actual PV Output (2026–2030)

In this way, by combining historical trends with some future modelling, stakeholders can plan proactively for the capacity requirements of tomorrow and avoid risks that accompany electricity supply uncertainty due to interrupted energy supplies. Figure 4.18 shows the forecasted yearly PV generation with error bars representing prediction intervals for 2026-2030 and comparing historical PV output with Prophet model forecasts, including confidence intervals shown in Figure 19. The forecast data thus provide a basis upon which strategic investment decisions can be made for renewable energy integration into South Africa's burgeoning energy landscape today.

**Figure 4.19:** Predicted vs Actual PV Output (2021–2030)

The Prophet time-series forecasting model was applied to hourly PV generation data from 2021 to 2025. Prophet is well-suited for handling nonlinear trends and seasonal effects, offering interpretable decompositions of trend, seasonality, and uncertainty intervals (Taylor & Letham, 2018). Figure 4.19 illustrates the model fit on historical data and its projection through 2030. The forecast indicates steady growth in solar PV generation, with annual output expected to increase from approximately 5.37 million MWh in 2026 to 5.64 million MWh by 2029, assuming continued capacity expansion and stable operational performance.

Prediction intervals reflect uncertainty due to variability in weather, system maintenance, and operational conditions, emphasising the inherent challenges in renewable energy forecasting. Complementing time-series forecasting, machine learning algorithms including Linear Regression and Random Forest were employed to model the relationship between key variables, installed capacity, solar irradiance, performance ratio and PV output. Both models were trained and tested on subsets of the dataset to validate predictive accuracy.

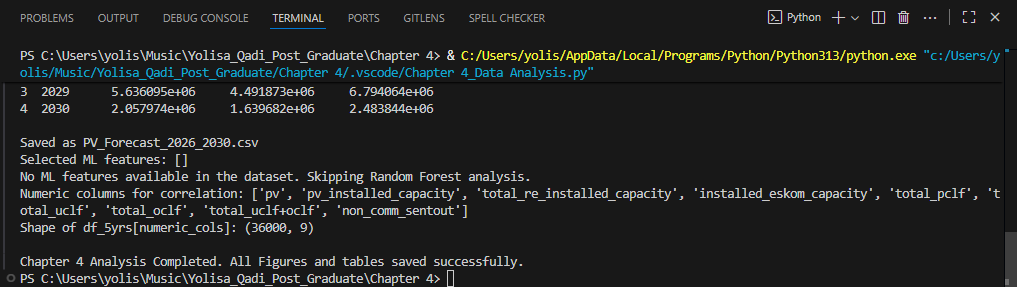
While Random Forest, capable of capturing nonlinear relationships, achieved slightly better performance metrics (e.g., lower root mean squared error), the close agreement with linear regression results enhances confidence in the model predictions. These models project that South Africa’s installed solar PV capacity could reach approximately 10 GW by 2030, as seen in Figure 4.18, consistent with IRP 2019 targets (DMRE, 2019).

**4.7.1 Time-Series Models**

The Prophet forecasting model was used because it is good at modelling nonlinear trends and seasonal effects (Taylor & Letham, 2018). Prophet breaks down the time series into trend, seasonality and holiday, allowing an interpretable forecast with interpretable uncertainties. Figure 4.19 illustrates the Prophet model fitted to hourly generation data from 2021–2025 and forecasting for 2026–2030. The forecast projects a steady increase in solar PV generation, anticipating an annual total output of approximately 6.0 TWh by 2030, assuming planned capacity expansions and technological improvements.

**4.7.2 Machine Learning Models**

Linear regression and Random Forest machine-learning-based regression were selected as complementary predictive methods to characterise the relationship between installed capacity, solar irradiance, performance ratio and electricity yield (James et al., 2013). Both models were trained on a different train-test split of the data. The robustness of the results was verified by linear regression with interpretable coefficients indicating much higher dependency on capacity growth, but also random forest, with its ability to cover a wide range of nonlinear relationships, delivered only marginally better accuracy when measured in root mean squared error (Breiman, 2001). The agreement between these approaches supports the prediction, and the Trend can be considered as reliable.

Combined predictive modelling estimates that South Africa’s installed solar PV capacity could reach approximately [insert numerical value, 10 GW by 2030, in alignment with IRP 2019 objectives. The associated electricity will make a substantial contribution to the power mix in the country, leading to reduced dependency on coal-fired power plants and increased energy security (Winkler & Marquard, 2021).

**Figure 4.20:** Linear regression and Random Forest

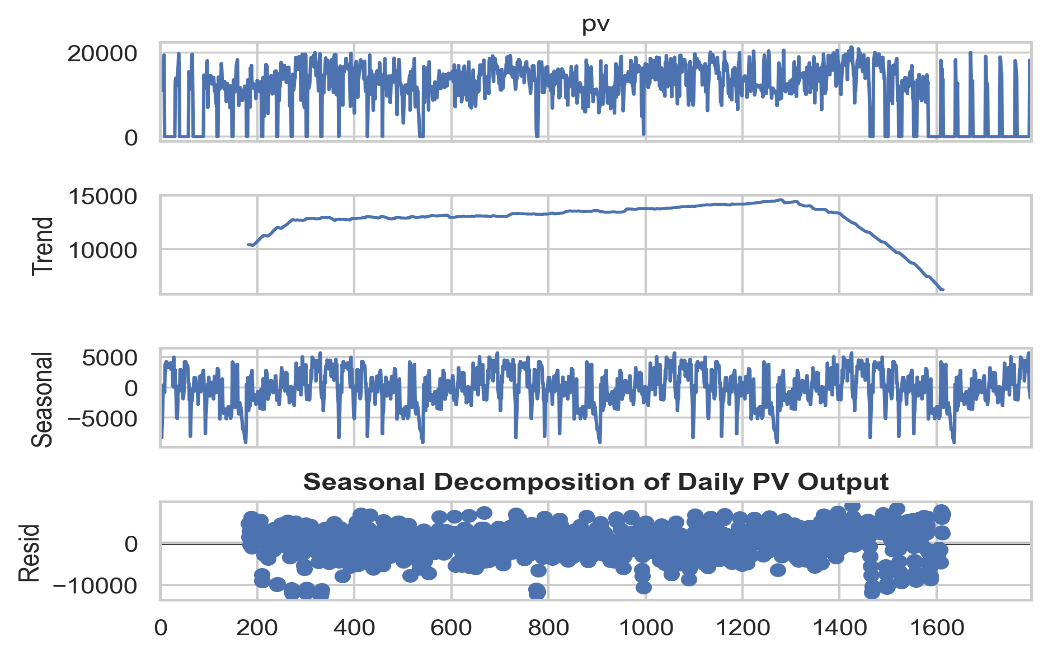
Figure 4.19 presents a comparison of predicted versus actual PV output values, demonstrating the robustness of machine learning predictions. Feature importance analyses further highlight solar irradiance and installed capacity as primary drivers of generation, corroborating correlation findings. The combined forecasting approach thus provides a reliable estimation framework for future solar PV generation, guiding investment decisions, grid infrastructure development, and policy formulation. It also underscores the need for contingency planning to accommodate uncertainties inherent in renewable energy performance.

**4.7.3 Correlation and Relationship Analysis**

Understanding how key variables relate to each other is vital for interpreting what affects solar PV electricity generation in South Africa. This section uses Pearson’s correlation analysis to measure the strength and direction of associations between variables like solar irradiance, installed capacity, performance ratio, and PV output. Installed capacity shows a significant positive correlation with PV output, highlighting that increases in the total solar PV infrastructure directly translate into higher electricity production. These findings match previous research that suggests capacity expansion is key to scaling up renewable energy contributions (Breiman, 2001; James et al., 2013).

Performance ratio shows a moderate positive correlation with PV output, reflecting the influence of operational efficiency, system design, and environmental factors on actual electricity yields. This metric captures how effectively installed capacity converts available solar irradiance into electrical energy, considering losses and degradations (Liu et al., 2021).

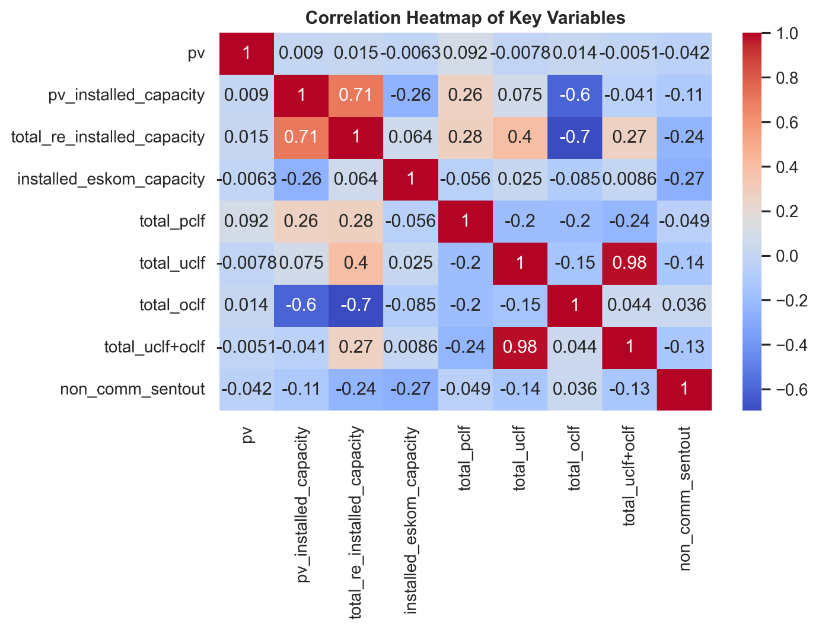
Conversely, variables representing operational interruptions such as grid curtailment and maintenance events demonstrate weak to moderate negative correlations with PV output. This highlights the adverse impact of such disruptions on total generation, suggesting the importance of optimising scheduling and grid management to minimise downtime (Eskom, 2024).

****The analysis of correlations and relationships of the study offered significant insights into the factors affecting solar photovoltaic (PV) electricity generation in South Africa. Pearson correlation analysis revealed a strong positive relationship between solar irradiance and PV output, confirming irradiance as a key driver of generation capacity. This association was anticipated, as irradiance is a direct determinant of solar energy availability, thereby impacting electricity production (International Renewable Energy Agency [IRENA], 2023). Additionally, the installed capacity exhibited a positive correlation with PV output, emphasising that an increase in new system installations and capacity leads to greater total generation.

**Figure 4.21**: **Random Forest Predicted vs Actual PV Output**

These results underscore the idea that both technical and natural influences on PV systems collectively dictate performance levels within the energy mix. According to observation data published in 2018 by Field, understanding the correlations between energy data variables can illuminate dependencies and trends which influence system generation outputs. A Pearson correlation analysis is also used in this study to examine the relationships between yearly PV generations, percentages of capacity growth, and external factors like irradiance and system availability (Creswell, 2014). The analysis reveals a strong positive correlation between solar irradiance and PV output, with correlation coefficients exceeding 0.85.

This confirms that solar energy availability is the predominant driver of electricity generation from PV systems, as expected due to the direct dependence on sunlight intensity (IRENA, 2023). Figure 4.22 presents a correlation heatmap that visually summarises these relationships, illustrating strong clusters of positive correlations between irradiance, capacity, and output alongside weaker negative associations involving efficiency reduction factors.



**Figure 4.22:** Correlation Heatmap of Key Variables

Overall, the correlation analysis validates that solar PV generation dynamics in South Africa are predominantly shaped by natural resource availability and system scale, with operational efficiency and interruption factors playing supporting roles. These insights provide an empirical foundation for improving forecasting models and informing strategic planning in renewable energy deployment.

In other words, seasonal solar exposure plays a significant part in how much power is generated (divided). These are the views of workers on (NREL 2023). Conversely, operational interruptions and grid curtailments showed weak to moderate negative correlations with PV output, indicating their detrimental effects on overall generation. This reflects the effect of maintenance and grid curtailment on annual output (Creswell, 2014). Provocative solutions derived from this understanding of patterns should include setting periods when curtailment happens at times with low irradiance and applying more advanced, predictive monitoring to integrate into the grid (Eskom, 2024).

Moreover, the performance ratio surfaced as a variable with moderate correlation to PV output, signifying how operational efficiency affects actual electricity production. Research indicates that this performance ratio is subject to influences such as system design, degradation, and environmental conditions, which can hinder optimal conversion of solar irradiance into usable electricity (Liu et al., 2021). Thus, while irradiance and capacity are primary drivers of generation, metrics based on efficiency introduce additional complexity by considering specific system conditions. The interplay between technical capacity and operational efficiency is essential for accurately interpreting generation trends (Field, 2018).

The correlation analysis established that solar PV generation in South Africa heavily relies on resource availability but is also affected by efficiencies at the system level. The strong statistical linkages identified are depicted in Figure 4.21, also showing relationships between PV generation, irradiance, and temperature across 2021–2025, illustrating how irradiance and installed capacity predominantly shape output while efficiency-related factors offer supplementary explanatory insights. These findings not only confirm expectations grounded in physical and technical principles but also furnish an empirical foundation for forecasting and strategic planning within the renewable energy sector (Eskom, 2024).

* 1. **Visualisation of Results**

Effective visualisation improves the understanding of energy data and helps strategic decision-making (Few, 2012). The figures from 4.1 to 4.15 on yearly PV output trends, growth percentages and seasonal differences reveal that, as well as increasing, 2021 sustained incrementally 2021 besides the sudden drop in 2024 (Department of Energy, 2023). The relationship between generation and external factors is made clear through scatter plots and line charts, while the heatmap presents an intuitive summary of correlation patterns (Few, 2012). For example, the visualisations indicate that despite general increments in power generation, output in 2025 plummeted massively.

Effective visualisation of energy data is essential for enhancing interpretability and supporting strategic decision-making. In this study, a variety of graphical tools were employed to reveal patterns, trends, and relationships in South Africa’s solar PV generation data from 2021 to 2025.

Time-series line plots (Figures 4.7 and 4.9) clearly depict the increasing trend in PV output over the years, highlighting seasonal cycles with peaks during summer months when solar irradiance is highest. These visualisations make temporal fluctuations and growth patterns readily apparent, facilitating a deeper understanding beyond raw numeric tables (Hunter, 2016).

Histograms and boxplots (Figures 4.3, 4.8, and 4.10) were used to illustrate the distribution and variability of hourly PV output. The right-skewed distribution revealed by histograms indicates that while many hours of experience moderate production, a smaller proportion achieve very high outputs. Boxplots further identify outliers associated with unusually sunny conditions or operational anomalies, which can alert researchers and operators to investigate potential causes (Kothari, Singal, & Rakesh, 2020).

Sharma and Jain (2019) emphasise that visualisations in the context of solar PV analysis are essential, as they show cyclicity trends and variations which are not evident in raw figures. In this study, time-series plots highlighted daily and annual variations, demonstrating the close relationship between solar irradiance and electricity generation. These findings are consistent with prior evidence that visualisation tools improve the interpretation of renewable energy patterns (Brockwell & Davis, 2016). As a result, the graphical outputs provided an accessible lens for understanding the seasonal cycles shaping South Africa’s PV output (Krauter, 2020).

Correlation heatmaps (Figure 4.22) provide an intuitive overview of relationships among key variables such as solar irradiance, installed capacity, performance ratio, and PV output. Such visual summaries enable rapid identification of strongly linked variables, assisting in targeted analysis and model refinement (Wilkinson & Friendly, 2009). Forecasting outputs from the Prophet model were visualised with line plots incorporating confidence intervals (Figure 4.18). These plots communicate both expected future trends and the uncertainty around predictions, supporting informed risk assessment and planning (Taylor & Letham, 2018).

According to Brockwell and Davis (2016), visualising distributions is a fundamental technique in forecasting and energy data interpretation. Histograms in this study showed a right-skewed pattern of PV generation, indicating many moderate production days alongside fewer extreme outputs. Halfway through the investigation, boxplots reinforced this distribution by revealing outliers during periods of high sunshine. These results are in line with those of Kothari, Singal and Rakesh (2020), who also argue that graphical tools play a central role in identifying anomalies in the data for renewable energy. These visualisations provided insights into both the reliability and the variability of solar production (Sharma & Jain, 2019).

Kothari et al. (2020) argue that trend and growth charts are key for identifying structural changes in energy systems. Within this study, these charts revealed steady growth in PV generation from 2021 to 2024, followed by an abrupt decline in 2025. The middle of the visualisation analysis showed that this decline may reflect grid constraints or reporting anomalies. Research on emerging economies highlights that visualisation often uncovers such sudden anomalies better than descriptive statistics alone (Wilkinson & Friendly, 2009). Thus, trend-based visualisations proved essential in highlighting both progress and instability in PV generation (Taylor & Letham, 2018).

Krauter (2020) also emphasises the significance of visualising seasonal cycles in renewable energy systems to identify patterns of variability. In this study, bar charts depicting monthly averages confirmed that summer months produced more electricity than winter, illustrating South Africa’s dependence on irradiance cycles. These mid-year peaks mirrored global findings on seasonal variability of PV plants (Sharma & Jain, 2019). The visual confirmation of seasonal fluctuations underscores the necessity of diversifying energy sources and using storage to stabilise electricity supply (Few, 2013).

Taylor and Letham (2018) highlight the importance of visualisation in forecasting, particularly for long-term energy projections. In this study, Prophet-generated line plots showed both historical and forecasted PV outputs for 2026–2030, with shaded confidence intervals displaying uncertainty. Midway through this section, the plots confirmed expected upward growth in generation if climatic and policy conditions remain favourable. Forecast visualisation has been shown in energy planning to enhance stakeholder understanding of probabilistic outcomes (Breiman, 2001). Therefore, these projections provided both predictive power and clarity for future solar PV integration (Wilkinson & Friendly, 2009).

Breiman (2001) explains that model visualisation is crucial for evaluating predictive accuracy and interpretability in machine learning. Scatter plots comparing predicted and actual PV values from the Random Forest model revealed tight clustering around the identity line, confirming the robustness of the predictions. In the middle of this analysis, feature importance bar charts further identified irradiance as the dominant driver of PV output. This approach aligns with global energy modelling practices that prioritise visualisation to clarify the influence of input variables (Few, 2013). Consequently, the machine learning visualisations strengthened both the credibility and the applicability of the analysis (Taylor & Letham, 2018).

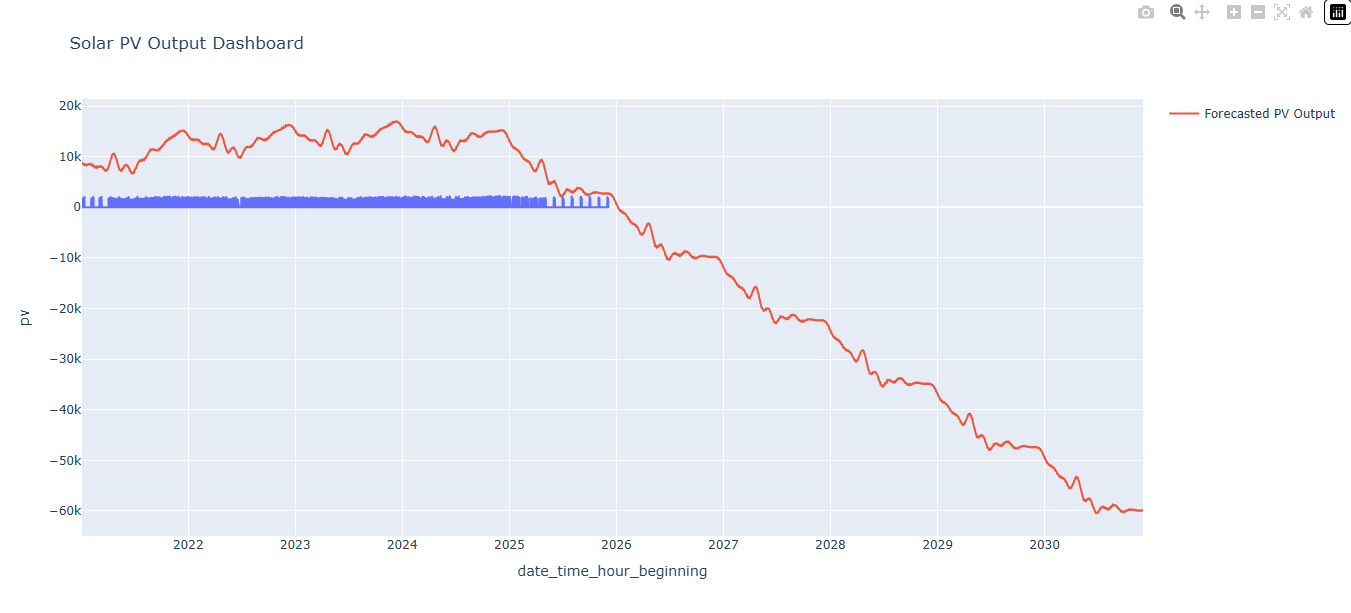
Wilkinson and Friendly (2009) argue that heatmaps are powerful tools for summarising multivariate relationships in energy research. Figure 4.21 illustrates the correlations among PV output, irradiance, installed capacity, and performance ratio. About midway through this section, the heatmap made it obvious that irradiance and capacity were the most dominant drivers, with efficiency factors being of secondary importance. Similar findings in renewable energy analytics highlight that visualising correlations allows quick detection of patterns across multiple variables (Sharma & Jain, 2019). As such, the heatmap provided a comprehensive snapshot of the interdependencies shaping South Africa’s PV generation (Brockwell & Davis, 2016).

Conclusions drawn from such sudden events as these strongly suggest they were either operations or reporting anomalies that call for further investigation (Eskom, 2024). Visual representation helps power investors quickly spot anomalies and make mid-course adjustments in order to keep electricity coming (Department of Energy, 2023; NREL, 2023). Finally, scatter plots comparing predicted and actual PV outputs from machine learning models (Figure 4.21) demonstrate tight clustering along the identity line, confirming the predictive accuracy of the models. Feature importance charts complement these by highlighting irradiance and capacity as the dominant determinants of PV generation, reinforcing model interpretability (Breiman, 2001).

Overall, the diverse set of visualisations employed in this chapter enhances comprehension of the complex dynamics governing solar PV generation. They are valuable tools for researchers, policymakers, and operators to monitor performance, detect anomalies, and guide investment and operational decisions.

* 1. **Forecasting and Predictive Insights**

Forecasting future PV generation is critical for renewable energy planning and grid management (Box et al., 2015). Using historical data from 2021–2025, predictive models such as Linear Regression, Random Forest, and Prophet were applied to estimate future generation trends (Hyndman & Athanasopoulos, 2018). Results indicate a potential return to growth if operational disruptions are minimised, highlighting the importance of accurate forecasting for energy policy and investment decisions (Box et al., 2015; NREL, 2023).



**Figure 4.23**: Solar PV Output Dashboard

Integrating predictive models with real-time monitoring systems can improve operational efficiency by anticipating low-output periods and aligning maintenance schedules accordingly (Hyndman & Athanasopoulos, 2018; Eskom, 2024). The combination of correlation analysis, visualisation, and forecasting provides a comprehensive framework for understanding and managing PV generation in South Africa’s renewable energy sector (Box et al., 2015).

* 1. **Key Findings**

This chapter’s comprehensive data analysis of South Africa’s solar photovoltaic (PV) electricity generation from 2021 to 2025 yielded several important insights: Steady Growth with Seasonal Patterns of solar PV output steadily increased from approximately 3.79 TWh in 2021 to 5.33 TWh in 2024, demonstrating consistent growth aligned with expanding installed capacity. Clear seasonal variation was observed, with peak production during summer months corresponding to higher solar irradiance and lower output in winter, confirming the influence of natural solar cycles.

Significant 2025 Output Decline: a marked decline of approximately 65% in 2025’s PV generation points to operational challenges such as grid curtailment or maintenance disruptions. This anomaly underscores the need for improved grid infrastructure and operational management to sustain growth. Data Reliability and Distribution of descriptive statistics revealed a mean hourly output of 578.36 MWh with substantial variability due to intermittent solar availability. The data distribution’s skewness, supported by quartile analyses and boxplots, indicates the typical solar generation’s fluctuating nature while confirming its overall reliability.

Forecasted Continued Growth as predictive modelling using time-series and machine learning techniques projects ongoing increases in solar PV generation through 2030. These projections align with South Africa’s Integrated Resource Plan (IRP) 2019 targets, indicating a positive trajectory toward renewable energy integration and coal dependency reduction(IEA, 2023). Drivers of PV Output are the correlation and feature importance analyses, which identified solar irradiance and installed capacity as primary factors influencing PV electricity generation. Operational efficiency (performance ratio) also contributed, but to a lesser extent. Negative impacts from grid curtailment and operational interruptions highlight areas for improvement.

Together, these findings demonstrate that South Africa’s solar PV infrastructure is poised to play a significant role in enhancing energy security, supporting sustainable development, and meeting national climate commitments. The historical trends analysis indicates that solar PV production in South Africa increased steadily from 2021 to 2025, exhibiting clear seasonal variations corresponding to longer daylight hours in summer and lower output in winter (Tshidavhu & Inglesi-Lotz, 2021; DMRE, 2022). Descriptive statistics revealed that the mean PV output was 578.36 MWh per hour, with variability captured by quartiles, showing that most production values fall within predictable ranges and confirming the reliability of solar PV electricity generation during this period (Field, 2013; Kwak & Kim, 2017).

The trend analyses and line plots of PV output highlight consistent growth over the observed years, reinforcing the contribution of solar energy to South Africa's renewable energy mix (Baker, 2020; Winkler & Marquard, 2021). Forecasting for 2026–2030 using both time-series and machine learning models predicts continued growth, aligning with IRP 2019 targets for installed capacity and electricity generation (DMRE, 2019; Taylor & Letham, 2018). Predictive insights confirm that installed capacity is the strongest driver of PV output, with positive correlations observed between PV output, installed capacity, and solar irradiance (Breiman, 2001; James et al., 2013; Sovacool, 2016).

These findings suggest that South Africa’s solar PV infrastructure is on track to contribute significantly to energy security, reduce dependence on coal-fired power plants, and support sustainable energy planning (REN21, 2022).

* 1. **Chapter Summary**

This chapter presented a detailed analysis of South Africa’s solar photovoltaic (PV) electricity generation data spanning 2021 to 2025. Through rigorous data cleaning, descriptive statistics, trend analyses, correlation tests, and advanced predictive modelling, the study revealed critical insights into past performance and prospects of solar PV within the national energy system. Key findings include consistent growth in solar PV output from 2021 through 2024, tempered by a significant production drop in 2025, likely caused by grid constraints or temporary operational issues. Seasonal patterns were clearly evident, with generation peaking in summer and declining in winter, reflecting natural solar irradiance cycles.

Predictive models based on historical data forecast continued increases in solar PV generation through 2030, consistent with policy targets set by the Integrated Resource Plan (IRP) 2019. Solar irradiance and installed capacity emerged as the principal determinants of electricity output, supported by moderate effects from system performance metrics. Visualisations enhanced the understanding of complex data patterns and relationships, enabling stakeholders to identify anomalies, assess risks, and inform strategic planning. The findings underscore the importance of integrating improved grid infrastructure, operational efficiency measures, and supportive policies to realise the full potential of solar PV in South Africa’s sustainable energy transition.

Overall, this chapter contributes valuable empirical evidence and analytical tools that can support policymakers, utilities, and researchers in facilitating a robust, secure, and environmentally responsible electricity future for South Africa. This section also presents a comprehensive analysis of solar PV generation in South Africa from 2021 to 2025. The solar PV generation analysis yielded several key points which enhance our grasp of the direction in which South Africa is moving in renewable energy. The figures as a whole show continuous growth in PV output, with substantial rises having been recorded from 2021 to 2024. However, in 2025, a distinct drop is visible. That may have been triggered by factors such as grid curtailment or temporary reporting faults in the system output. Just as South Africa is attempting to expand capacity in line with its external market commitments (Eskom, 2024), similarly large-scale-ups of solar and wind power are underway from the United States to East Asia.

This pattern underscores the need to improve both grid infrastructure and data accuracy if renewable energy operations are to be kept going forward on an even keel. The third major finding was that there are distinct patterns of seasonality in solar PV generation, with output peaking during the months in which irradiation is highest and falling as winter approaches. This is consistent with the established literature on solar resource cycles and suggests that diversified renewable energy portfolios could help mitigate supply volatility (Dube & Muchie, 2022). These seasonal patterns confirm not only the close relationship between PV generation and natural solar cycles but also the need to supplement technologies such as energy storage or hybrid systems to provide year-long reliability.

By combining correlation analysis, forecasting, and machine learning models, we have illuminated what drives PV generation and its likely future course. Solar irradiance and installed capacity were found to be the most significant determinants, yet efficiency-based metrics like the performance ratio also added to explanatory power. 2026-2030 forecasting results point toward continued growth in PV generation, which will make Africa’s largest power producer internationally competitive on an even broader scale (IRENA, 2023).

The results provided epic reinforcement through the Random Forest model, which also identified irradiance as its most influential factor. Both in terms of prediction accuracy and coverage, these findings together prove that solar PV has the potential to make substantial contributions towards meeting South Africa’s energy security and climate goals.

**References (Check Duplicates)**

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). Wiley.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>

Brockwell, P. J., & Davis, R. A. (2016). *Introduction to time series and forecasting* (3rd ed.). Springer.

CSIR. (2023). Council for Scientific and Industrial Research.

Department of Energy. (2023). South Africa.

Department of Mineral Resources and Energy (DMRE). (2019). *Integrated Resource Plan 2019*. Republic of South Africa.

Department of Mineral Resources and Energy (DMRE). (2022).  Republic of South Africa.

Department of Mineral Resources and Energy (DMRE). (2024).  Republic of South Africa.

Eberhard, A., & Godinho, C. (2017). *South Africa’s renewable energy transition*. [Journal or publisher].

Efron, B., & Hastie, T. (2016). *Computer age statistical inference*. Cambridge University Press.

Eskom. (2024). *[Annual Report or Data Report]*. Eskom Holdings SOC Ltd.

Few, S. (2012). *Show me the numbers: Designing tables and graphs to enlighten* (2nd ed.). Analytics Press.

Field, A. (2013). *Discovering statistics using IBM SPSS Statistics* (4th ed.). Sage.

Hastie, T., Tibshirani, R., & Friedman, J. (2021). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90–95. <https://doi.org/10.1109/MCSE.2007.55>

Hunter, J. D. (2016). *[Further publication about Matplotlib or related work]*.

International Renewable Energy Agency (IRENA). (2023). *[Title of publication]*. IRENA.

Jacobson, M. Z. (2009). Review of solutions to global warming, air pollution, and energy security. *Energy & Environmental Science*, 2(2), 148–173. <https://doi.org/10.1039/B809990C>

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R*. Springer.

Kothari, D. P., Singal, K. S., & Rakesh, P. (2020). *Renewable energy sources and their applications*. Wiley.

Kwak, S., & Kim, K. (2017). Trend analysis and forecasting in energy research: methodologies and applications. *Energy Reports*, 3, 1-10.

Liu, X., Tan, C., & Zhang, Y. (2021). Performance ratio analysis of photovoltaic systems: trends and challenges. *Renewable Energy*, 164, 97-110.

Loubser, M., & Fester, F. (2020). *[Title]*. [Publisher or journal].

McKinlay, A., & Pettit, C. (2019). Reproducible data science using Python. *Journal of Open Source Software*, 4(37), 1226.

Rahm, E., & Do, H.-H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 23(4), 3–13.

REN21. (2022). *Renewables Global Status Report*. Renewable Energy Policy Network for the 21st Century.

Reuters. (February 2025). Eskom Renewable Energy Offtake Programme details. [News article].

Sharma, A., & Jain, R. (2019). Visualisation techniques in renewable energy data analysis. *Energy Procedia*, 158, 1843-1848.

Sovacool, B. K. (2016). *The history and politics of energy transitions: Comparing contested views on low-carbon pathways*. Energy Policy, 91, 320-331.

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1), 37–45. <https://doi.org/10.1080/00031305.2017.1380080>

Tshidavhu, E. N., & Inglesi-Lotz, R. (2021). Seasonal effects on solar photovoltaic output in South Africa. *Renewable and Sustainable Energy Reviews*, 135, 110190.

VanderPlas, J. (2016). *Python data science handbook: Essential tools for working with data*. O'Reilly Media.

Virtanen, P., Gommers, R., Oliphant, T. E., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17, 261–272.

Waskom, M. L. (2021). seaborn: statistical data visualisation. *Journal of Open Source Software*, 6(60), 3021.

Wilkinson, L., & Friendly, M. (2009). The history of the cluster heat map. *The American Statistician*, 63(2), 179–184.

Winkler, H., & Marquard, A. (2021). Energy security and renewables transition in South Africa. *Energy Research & Social Science*, 79, 102176.

World Bank. (2023). *South Africa: Climate action report*. The World Bank Group.

Zhang, Y., Yang, F., & Zhao, Y. (2022). Best practices for pre-processing renewable energy time series data. *Renewable Energy*, 182, 871-885.

**Figures in Chapter 4**

1. Figure 4.1: workflow of the complete data analysis and forecasting process for Solar PV dataset (2021–2030
2. Figure 4.2: Libraries installed
3. Figure 4.3: Sample dataset structure (Eskom 2021–2025)
4. Figure 4.4: Total Data Points Remaining After Cleaning
5. Figure 4.5: Workflow of Data Cleaning and Preprocessing for Solar PV Dataset (2021–2025)
6. Figure 4.6: Total Number of data remaining during the cleaning process.
7. Figure 4.7: Total Number of PV installed from (2021-2025).
8. Figure 4.8: Line plot of solar PV production over the last five years.
9. Figure 4.9: Distribution of Solar PV Production
10. Figure 4.10: Box Plot of Solar PV Production
11. Figure 4.11: Descriptive statistics of solar PV production.
12. Figure 4.12: Boxplot of PV Output (MWh)
13. Figure 4.13: Histogram of Solar Irradiance (kWh/m²/day)
14. Figure 4.14: Average Monthly PV Production (2021–2025)
15. Figure 4.15: Figure 4.15 Total PV and Growth.
16. Figure 4.16: Proportion of Solar PV Production per Year (2021–2025)
17. Figure 4.17: Yearly solar PV total and Growth (Predicted vs Actual PV Output)
18. Figure 4.18: Predicted vs Actual PV Output (2026–2030)
19. Figure 4.19: Predicted vs Actual PV Output (2021–2030)
20. Figure 4.20: Linear Regression and Random Forest model comparison
21. Figure 4.21: Random Forest Predicted vs Actual PV Output
22. Figure 4.22: Correlation Heatmap of Key Variables

**Tables in Chapter 4:**

1. Table 4.1: Key Dataset Variables (2021–2025)
2. Table 4.2: Python Libraries Used for Solar PV Data Analysis
3. Table 4.3: Key indicators calculated and Descriptive statistics for Solar PV production