UNIVERSIDAD
COMPLUTENSE
MADRID

Academic year 2022/2023

Master in Big Data and Data Science: Applications to Commerce, Business and Finance

Master's Degree Final Project in collaboration with Naturgy:

"Commercial Planning, Control and Administration:

Profit Margin Analysis".

AUTHOR:

Alexandra Diana Fediuc

NATURGY COORDINATOR:

Ramon Bono

Index

1. Introduction

1.1 Project Context and Collaboration with Naturgy

In the dynamic context of the energy sector, the ability to make an accurate assessment of the onerousness of operations is essential for effective financial decision-making. In this context, this Master's thesis is developed in collaboration with Naturgy, a leading energy supply company.

The entity faces the continuous challenge of assessing the onerousness of its transactions, where the constant fluctuation of production costs and selling prices adds a level of difficulty. The opportunity to collaborate with Naturgy offers the possibility to address a real issue and provides a unique opportunity to apply
advanced Data Science and Big Data techniques in the solution of industrial problems.

1.2 Objectives and Scope of Work

The main objective of this work is to develop a cost and generation allocation model that allows Naturgy to evaluate the onerousness of its operations in terms of the costs of renewable and conventional energy generation, as well as the costs of the generation of renewable and conventional energies.
industrial and retail sales prices. To achieve this, the following specific objectives are proposed:

- Generation of a model that determines, for each sales price range, the expected margin after taking into account the cost of generation, procurement and other operating costs.
- Automation of the process to obtain updated results from the
  The cost of each month, which facilitates and accurate decision-making.
- Creation of an interactive dashboard using Tableau, allowing monthly margin tracking and intuitive visualisation of onerousness.

1.3 Description of the Data and Initial Problem

To address the inherent complexity of assessing onerousness in the context of energy operations, Naturgy has provided a crucial set of data, which constitutes the starting point for this project. These data cover a specific period, from January 2023 to June 2024, and, although fictitious, serve as a fundamental starting point for the development of this work.

The shared data are organised in four vital files (see Annex 2: Original data shared by Naturgy), each one making its contribution
The only way to meet the challenge of accurately assessing the onerousness and, consequently, to make decisions in a financial and operational context:

- Renewable Energy Generation Costs: This file details the costs associated with renewable energy generation (Annex 2.3).
- Conventional Energy Generation Costs: Similar to the previous file, this dataset focuses on the costs related to conventional energy generation. conventional energy (Annex 2.4)
- Industrial Selling Prices: Within this file are the specific selling for the industrial sector (Annex 2.2).
- Retail Sales Prices: In contrast to industrial prices, this file reflects sales prices aimed at the retail market (Annex 2.1).

However, it is important to recognise that, despite the fictitious nature of the data, an initial issue arises that requires attention. The quality and consistency of the data shared may vary, which could directly affect the accuracy of the
evaluation of the model. Lack of uniformity and clarity in the initial data can complicate obtaining reliable results. Therefore, the first phase of the
The project focuses on the transformation of this data, ensuring that it is coherent and ready for subsequent analysis and modelling. Debugging and
preprocessing are critical elements to ensure that the dummy data become a solid and reliable basis for margin calculations and the achievement of the objectives set.

2. Literature Review

2.1 Concept of Onerousness in the Energy Sector

The concept of onerousness in the context of the energy sector refers to the
ratio between costs and revenues generated by power generation and sales operations. It is essential for companies in the sector to assess whether revenues from energy sales exceed the costs associated with generation and other operational activities.
Onerousness is a critical indicator for decision making.
informed and strategic financial management, as it provides a comprehensive overview of the
profitability of operations in relation to the various factors costs and selling prices.

2.2 Cost of Sales Allocation Methods

The process of cost allocation in sales focuses on how to allocate the operational costs between the different sales transactions performed. For this project, cost allocation will be carried out using an automation approach.
optimised using the Python programming language.

This automated approach increases the reliability and accuracy of the
assignment, while reducing the potential for human error. Before reaching this point, careful data cleaning and pre-processing will be carried out.
The data will be in Excel format, to ensure that the data is consistent and ready for the automation stage in Python.

The combination of manual debugging and subsequent automation with Python contributes to the creation of a comprehensive and effective workflow, ensuring a rigorous and optimal cost allocation in sales, in line with the objectives of this project.

3. Data Debugging and Preprocessing

3.1 Initial Manual Preprocessing with Excel

Given the nature of the data provided by Naturgy and its relatively complex structure, the first pre-processing step is carried out using Microsoft
Excel. This process involves cleaning the data, resolving inconsistencies and preliminary organisation to prepare it properly for further manipulation and analysis in Python. The choice of Excel at this stage is based on the need to
address specific challenges related to the structure of the data and ensure its adequate preparation for the next stage of the process.

Initially, the two data sets concerning the cost of energy production, comprising renewable and conventional energy, are consolidated. Although the period to be considered runs from July 2023 to June 2024, the available data mainly go up to December 2023. To address this limitation, several options are explored
strategies, including imputation with weighted average values. However, due to the fictitious nature of the data and its lack of impact on the final model, it is decided to repeat the values for the year 2023. In addition, the GWh figures are converted to MWh to facilitate future calculations. Finally, the initial structure of the tables is adjusted in order to optimise data manipulation (see Annex 3.3).

Subsequently, the Industrial Sales Price data is pre-processed. The
original file presents a row for each customer receiving energy supply. This row contains the month and year, the customer identifier, the consumption in GWh for that month, the total sales paid by the customer in the period and the type of contract, mostly Fixed Price. A column is added for the unit price per GWh, calculated by dividing the total sales by the consumption in GWh for each customer in that month. According to specific instructions from the company, the
total consumption in GWh for each customer divided by the number of days in the month, as this metric is used internally for calculations*. Likewise, columns such as CUPS (customer identification number), TYPE, the total sales for the period and the amount of GWh consumed in that month are suppressed, in accordance with the recommendation to work with the previously calculated daily consumption. This daily amount is converted from GWh

**Two weeks before the deadline, the company informed us that this division should be made by 12, and not by the number of days of the . However, given the fictitious nature of the data and that this does not affect the rest of the code, it has been decided not to apply this modification.*

to MWh to simplify the analysis. Finally, the sales prices are ordered from lowest to highest, first grouping the data by month and year (see Annex 3.2).

The last file submitted to pre-processing corresponds to Retail Sales Price. The data is presented in columns, where each column represents the monthly consumption in GWh and each row reflects a range of prices. These ranges cover values from
50/MWh, increasing by 1 from that point onwards. Given the small margin between the intervals, it is decided to use the average price of each range to facilitate the calculation. In addition, as has been done with the Industrial Sales Prices, the following is used proceeds to convert the data from GWh to MWh in order to simplify further calculations (see Annex 3.1).

4. Modelling and Automation

4.1 Development of the Cost Allocation and Generation Model

Once the data has been pre-processed in Excel, the next step is the
transformation and creation of variables using the Python programming language. This approach will allow the automation of repetitive tasks and the implementation of algorithms to generate new variables relevant to the onerousness analysis. The flexibility and data manipulation capabilities of Python are leveraged to ensure data consistency and quality prior to the construction of the cost allocation and generation model (see Annex 1: Full Jupyter code).
Notebook).

The initial phase of this process has focused on the analysis and processing of industrial and retail sales data, with the aim of achieving a uniform and coherent structure that allows for an effective comparison and linkage between the two sources.
of data.

We have started by examining the industrial sales file. In order to understand the distribution of prices in this dataset, a graph has been plotted of
price histogram. This visualisation has provided a clear understanding of how industrial sales prices are distributed in different ranges. By looking at this graph, the presence of peaks and troughs in certain ranges has been identified.

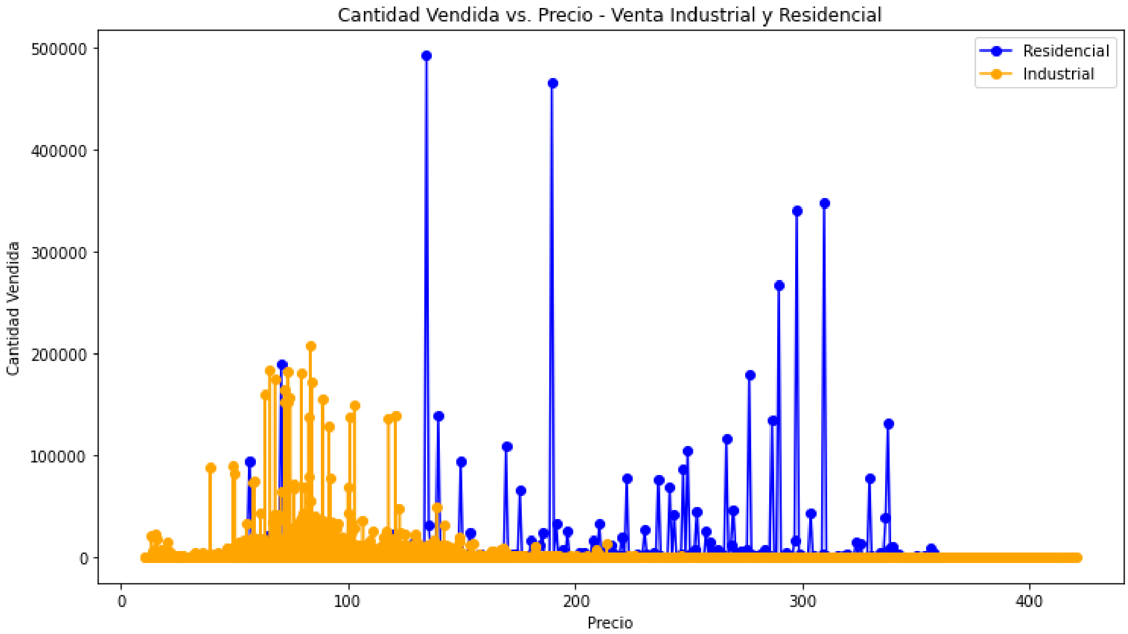
price segmentation. This preliminary analysis has led to consider different approaches price segmentation.

To address this issue, a more detailed analysis has been carried out. A scatter plot has been used to visualise the relationship between prices and quantities.
sold in industrial sales. This visualisation has allowed an understanding of how the quantity sold varies with different price levels. It has been observed that, in some price ranges, the quantity sold is significantly higher than in others, suggesting a correlation between prices and sales. This finding has led to consider the importance of selecting price ranges that allow a meaningful comparison between different groups.

In order to determine the optimal price ranges, graphs of quantity sold versus price were generated. These graphs have allowed a visual assessment of how the quantity sold is distributed across different price ranges. They have also helped to identify ranges with interesting variability and provide meaningful information on the relationship between prices and sales.
Before deciding on the width of the price ranges to work with, the same process will be followed for the retail sales data.

Following a similar process, retail sales data have been addressed. As with industrial sales, an exploratory analysis of prices has been carried out.
using a scatter plot and a histogram. It has been observed that the data also show clusters and concentrations in certain price ranges, which supports the idea of using price ranges to simplify the analysis and allow for comparison.

Combining the industrial and retail sales data in a single graph of quantity sold vs. price shows a notable difference between them. Industrial sales are concentrated in smaller price ranges, while retail sales have a higher price dispersion.



After a thorough analysis, intervals of 5 €/MWh have been chosen, as they provide a detailed overview of sales at different price levels.

Once the price ranges have been determined, the segmentation has been carried out. A new column has been created in each of the Dataframes in order to
indicate the price range to which each sale belongs. The data have then been grouped by month and ordered in ascending order according to price ranges.
price and months. This approach has ensured that the data is organised and ready for further analysis and concatenation.

Once the retail and industrial sales data have been prepared and
structured with uniform price intervals, the sets have been concatenated. An identifying column has been added to distinguish between industrial and retail sales. Then, using the concatenation function, an identifying column has been added to distinguish between industrial and retail sales.
merged both datasets into a single coherent table.

Table 1: Total sales grouped by month and price range

	MES	Rango_Precio	precio	volumen	Origen
0	2023-07-01	[10, 15)	13.208874	6045.816117	I
1	2023-07-01	[15, 20)	17.302245	38760.517078	I
2	2023-07-01	[20, 25)	22.211753	4589.759293	I
3	2023-07-01	[25, 30)	25.000000	77.538720	R
4	2023-07-01	[25, 30)	27.486343	3312.084203	I
...	...	...	...	...	...
1783	2024-06-01	[395, 400)	NaN	0.000000	I
1784	2024-06-01	[400, 405)	NaN	0.000000	I
1785	2024-06-01	[405, 410)	NaN	0.000000	I
1786	2024-06-01	[410, 415)	NaN	0.000000	I
1787	2024-06-01	[415, 420)	NaN	0.000000	I

1788 rows × 5 columns

Source: Naturgy. Own elaboration

In summary, the process of preparing and processing industrial and retail sales data has been a fundamental part of this project. Through a detailed analysis of prices and quantities sold, as well as the consideration of different
measures of price intervals, it has been possible to transform disparate data into one uniform and coherent structure. This process has paved the way for the analysis and subsequent modelling, allowing meaningful comparisons and a complete view of the data to achieve the objectives of this study.

4.2 Process Automation for Monthly Results

4.2.1 Sales data processing

In the sales processing process, we start by creating price ranges.
for industrial and retail sales. This involves grouping the data by month and ordering them from lowest to highest price. For industrial sales, the lowest price is 10.44 €/MWh, and for retail sales, it is. Price ranges are created from these minimum values and are increased by €5/MWh.

Next, a column called "Price_Range" is added, which indicates at which
price interval belongs to each sale. The date column is also converted

"MONTH" to the *datetime* data type for ease of manipulation. The data is sorted first by date and then by price range.

In both cases, an aggregation of data is performed to calculate the average price and volume sum for each month and price range. These results are stored in dataframes called "industrial_ranges" and "retail_ranges".

Then, a column called "Origin" is created in both Dataframes to distinguish whether the data comes from industrial ("I") or retail ("R") sales.

These Dataframes are combined into a single Dataframe called "total_sale" and sorted to maintain a proper organisation. The Dataframe is re-indexed to ensure indexes sequential. Furthermore, additional columns are added to the "total_sale" dataframe that will be important for future analysis, such as "Sale (€)", "Total Cost (€)", "Unit Margin (€/MWh)", and "Total Margin (€)".

Finally, in order to organise and prepare the data, a dictionary called "Monthly_Sale", which stores the Dataframes for each month, facilitating access to period-specific data in subsequent analyses.

4.2.2 Cost data processing

The processing of the cost data starts with the conversion of the column "Month" to the data type *datetime*, which facilitates the handling of dates.

It then iterates through the unique months in the original data and creates a monthly Dataframe for each of them. These Dataframes are stored in a dictionary called "monthly_generation_cost", as was done above with the sales data. Each of these Dataframes contains information on the volume and cost of different types of energy for a specific month.

After this step, the data structure is reorganised so that the indices correspond to the energy types (combi, hydro, renewables, nuclear) and the columns correspond to the price and volume of each of these energy types. In addition, the data is sorted according to price, from lowest to highest, for each month.

This process ensures that the data is properly structured and ready further analysis. An example of the data structure for each of the months can be seen in the table below.

Table 2: Generation costs in July 2023

	precio	volumen
combi	42.6	49000
nuclear	45.1	300000
renovables	60.2	358000
hidro	83.0	129000

Source: Naturgy. Own elaboration

4.2.3 Cost Allocation by Price Range

The cost allocation process by price range is detailed below. This process plays a fundamental role in the distribution of generation costs.
The main factors to be taken into account are the different selling prices and, , the assessment of profitability of each transaction.

Assignment Process Step by Step

1. Creating a Copy of the Data

To preserve the integrity of the original data, it starts by creating a copy of the monthly generation costs. This ensures that the original data remains unchanged during the process.

2. Cost of Sales Allocation

At this stage, costs are allocated to sales, considering both the availability of energy and the volumes required. The process is detailed :

- Iteration by Month and Sale: Each month and each point of sale is scoured for cost allocation.

- Initialisation of Variables: For each point of sale, following variables are initialised:
  - allocated_volume: The amount of energy allocated in this step.
  - Average cost: The weighted average cost allocated energy.
- Allocation Loop: The available generation costs are iterated over and costs are allocated to sales sequentially until the required volume for that specific sale is met.
- Assignment Control:
  - If the volume required is less than or equal to the available volume of a type of energy, allocate the full volume required and calculate the weighted average cost for that allocation.
  - If the required volume is greater than the available volume, all available volume from that source is allocated and the next energy source is moved on until the required volume is met.
- Calculation of the Weighted Average Cost: In each step, the average cost is calculated.

  The weighted energy allocation, taking into account the proportion of the volume allocated to the total volume.
- Finalisation of Allocation: Once the required volume has been allocated, the volume_allocated and the Average Cost are recorded in the corresponding sale record.

3. Verification

To ensure data integrity and consistency in allocation, two key checks are carried out:

- It is confirmed that the original generation cost data remains unchanged.
- It is checked that the sum of the allocated volumes coincides with the sum of the total volume to be allocated, ensuring that no volume has been lost.

  volume during the allocation process.

4. Results and Significance

At the end of this process, the following results are obtained for each point of sale:

- Sale: The sales value calculated for that point, taking into account the allocated price and volume.

- Total Cost: The total cost of the energy allocated at that point.
- Unit Margin: The difference between the unit price and the average cost.
- Total Margin: The difference between sales and total cost, if both values are available.

This cost allocation process is essential to assess the profitability of each sales transaction and to understand how generation costs are distributed across price ranges. These results are essential for decision making.
informed financial decisions in Naturgy.

Table 3: Scorecard for July 2023

	MES	Rango_Precio	precio	volumen	Origen	Venta	Coste Total	Margen unitario	Margen total	volumen_asignado	Coste medio
0	2023-07-01	[10, 15)	13.208874	6045.816117	I	79858.424366	2.575518e+05	-29.391126	-177693.342238	6045.816117	42.600000
1	2023-07-01	[15, 20)	17.302245	38760.517078	I	670643.951443	1.651198e+06	-25.297755	-980554.076069	38760.517078	42.600000
2	2023-07-01	[20, 25)	22.211753	4589.759293	I	101946.599406	1.965140e+05	-20.603995	-94567.377674	4589.759293	42.815748
3	2023-07-01	[25, 30)	25.000000	77.538720	R	1938.467999	3.496996e+03	-20.100000	-1558.528271	77.538720	45.100000
4	2023-07-01	[25, 30)	27.486343	3312.084203	I	91037.081840	1.493750e+05	-17.613657	-58337.915720	3312.084203	45.100000
...	...	...	...	...	...	...	...	...	...	...	...
144	2023-07-01	[395, 400)	397.560363	3.787721	I	0.000000	0.000000e+00	NaN	NaN	0.000000	0.000000
145	2023-07-01	[400, 405)	401.727642	23.264327	I	0.000000	0.000000e+00	NaN	NaN	0.000000	0.000000
146	2023-07-01	[405, 410)	409.279706	16.447313	I	0.000000	0.000000e+00	NaN	NaN	0.000000	0.000000
147	2023-07-01	[410, 415)	NaN	0.000000	I	0.000000	0.000000e+00	NaN	NaN	0.000000	0.000000
148	2023-07-01	[415, 420)	415.423771	5.210364	I	0.000000	0.000000e+00	NaN	NaN	0.000000	0.000000

149 rows × 11 columns

Source: Naturgy. Own elaboration

The table shows the structure of the data after all costs have been allocated and the total and unit margins have been calculated for each price range and energy type.

The last step of this code involves concatenating the dictionaries into a single Dataframe sorted by month and then exporting it to a csv file, which will be used as a data source in Tableau.

5. Creating Dashboards in Tableau

5.1 Data Preparation and Table Design

To start working in Tableau it is necessary to insert the previously obtained csv file. This will act as our data source.

Starting to work on the datasheets, each datasheet will be used for each different element that will later be added to the Dashboards:

1. Graph of the evolution over time of the total margin, with positive data in green and negative data in red.
2. Total data table, including the origin of the sale (residential or industrial), the total sale in euros, the total cost of generation allocated to the sales and the total margin. This table provides a global perspective of the data, allowing to evaluate the performance in terms of price ranges and the origin of the sale.
3. Sum total data. This sheet shares the same filters as the previous one, being in this case: *Price Ranges*, *Origin* and *Month*. While the previous sheet showed all the data that met the filter requirements, this one only shows the total sum of *Sales, Cost* and *Margin* of all the data. that meet these requirements.
4. Table of unit data, where the origin of the sale is included (either residential or industrial), the selling price, the average generation cost and the unit margin. This table provides a detailed view of the data at unit level.
5. Average unit data. This sheet shares the same filters as the previous one, being in this case: *Price Ranges*, *Origin* and *Month*. While the previous sheet showed all the data that meet the filter requirements, this one only shows the average of *Sales Price, Average Cost* and *Unit Margin* of all the data that meet these requirements.
6. Graph of total margins as a function of Price Range for Industrial and Residential. In order to observe how the total margin evolves as the price range increases, differentiating between industrial and residential sales.

7. Graph of unit margins as a function of the Price Range for Industrial and Residential. In order to observe how the unit margin evolves as the price range increases, differentiating between industrial and residential sales.

This structuring of data facilitates a clear and detailed representation of the information, allowing a comprehensive analysis of the influence of price on margins in the energy industry.

5.2 Interactive Dashboard and Analytics

In the next step, two interactive dashboards designed to simplify the analysis of the data are created. Both include interactive filters that
The filters allowed users to select the month, price range and origin (residential or industrial) of their choice. The application of these filters enables a dynamic analysis of the data, as the graphs and tables on the dashboard were updated in real time based on the selections made.

The first Dashboard (see Annex 4.1) shows two tables, one corresponding to the Total Margin and the other to the Unit Margin, which show according to the filters selected, all data relating to *Month, Origin of Sale, Total Sale or Unit Sale Price, Total Cost or Average Cost,* and *Total Margin or Unit Margin,* respectively. In addition, the totals calculation functionality is incorporated (*Sale, Cost* and *Margin*) based on the selected filters, as well as the calculation of the average unit costs (*Selling Price, Average Cost* and *Unit Margin*). This implies real-time performance of the calculations and users can observe how the calculations change as the filters are adjusted.

Finally, a graphical display of the monthly evolution of the total margin over time is included. This allows the analysis of trends and patterns in the total margin.

In the second Dashboard (see Annex 4.2) two graphs are presented, which correspond to the evolution of total and unit margins as a function of the range of
selling prices. It is shown that changes in sales price ranges have a significant impact on the margin, as it is observed that the higher the price range, the higher the margin.
small, the margin is smaller.

In summary, the process of creating dashboards in Tableau is focused on the
data preparation, table design, implementation of interactive filters and dynamic visualisation of results, which facilitates detailed analysis of the
influence of price ranges on the margin in the energy industry.

In addition, by applying a real-time connection to the data source, any changes applied to the original data or code are automatically reflected in the dashboards.

6. Analysis and results

6.1 Interpretation of Model Results

The results of the model developed during research provide a
in-depth understanding of the relationship between selling price ranges and profit margins in the energy industry. It has been shown that variation in selling prices directly impacts margins, which means that Naturgy can take specific actions depending on market conditions. For example, when price ranges are higher, the company can consider the
implementation of more dynamic pricing strategies. This could involve the ability to adjust prices more flexibly and nimbly to maximise margins at favourable times.

6.2 Data Visualisation in Tableau Dashboards

Data visualisation plays a crucial role in understanding and communication of the results of the study. The interactive dashboards created in Tableau allows for a clear and accessible representation of data. The graphs and tables in the Dashboards make it easy to identify patterns and trends. For example, through the visualisation, it is evident that the lowest prices are related to narrower margins. This information allows Naturgy to take more informed decisions on how to adjust their pricing strategies in real time to maintain healthy margins.

6.3 Naturgy's Decision Impact Assessment

The results of the study have a direct and significant impact on Naturgy's strategic decisions. The company can take advantage of these findings to improve its competitiveness in the energy market. For example, it can consider implementing more aggressive pricing strategies at times of favourable price ranges.
to maximise the overall margin. , it can adapt its pricing approach when ranges are less conducive to maintaining demand and revenues. At
In summary, the results not only provide valuable information, but also provide clear guidance for decision-making that can directly benefit the company in terms of profitability and market competitiveness.

7. Conclusions

7.1 Achievements in Relation to Project Objectives

The project has successfully achieved all previously set objectives. During the research and analysis process, an accurate identification of key trends in the energy data was achieved. An in-depth understanding of

how variations in selling price ranges are directly related to fluctuations in profit margins. These achievements provide strong support for
The project is designed to achieve the specific objectives set out at the beginning of the project in a solid and effective manner.

7.2 Contributions to Onerousness Management at Naturgy

The concrete contributions to onerousness management in Naturgy are remarkable. The results and analyses generated throughout the project provide a solid basis for the strategic decision making related to pricing strategies. The data and conclusions obtained are valuable resources that the company can use to
optimise the profit margin and improve their competitiveness in the energy market. These findings have a direct impact on the company's ability to make effective decisions.

7.3 Reflections on the Process and Lessons Learned

In the course of the development of this solo project, a number of different
challenges and obstacles that have been successfully overcome. One of the most notable aspects was the need for extensive pre-processing of the data prior to analysis. This process taught the critical importance of assuring the quality and
consistency of the data before working with it. I learned that cleaning and
Proper organisation of data is essential to obtain accurate and meaningful results.

A key part of this project was the development of Python code. It was
used advanced programming techniques, including loops and complex data structures such as dictionaries. This experience allowed me to improve my programming and problem-solving skills. In particular, the handling of data via
custom scripting in Python was essential for the efficient manipulation and analysis of complex datasets.

In addition, Tableau was used to create interactive dashboards that would allow for visualise the results in real time. This part of the project highlighted the importance of presenting data in an effective and accessible way. I learned how to design clear and attractive visualisations that make it easier to understand the results, in turn improves the ability to make informed decisions.

In summary, this project provided me with valuable experience in data pre-processing, Python programming and creating interactive dashboards in Tableau.
I learned that data quality and efficient programming are critical to success in data analytics projects, and that proper visualisation is essential.
to communicate results effectively. These lessons learned enrich my skill set and prepare me to tackle similar challenges in the future with confidence and effectiveness.