UTRECHT UNIVERSITY

Department of Information and Computing Science

Applied Data Science Master Thesis

"Unveiling the Impact of Energy Labels on House Energy Consumption"

First examiner:

Bilgecag Aydogdu

Second examiner:

Albert Ali Salah

Candidate:

Mark Dielessen

In cooperation with:

Stater N.V.

July 14, 2023

Abstract

Energy efficiency and sustainability in residential properties are of increasing importance in addressing environmental concerns and reducing energy consumption. This study focuses on the relationship between energy labels and energy consumption in residential properties. Two datasets, the EP dataset and the Funda dataset, were utilized to analyze the potential cost savings associated with upgrading energy labels.

The aim was to determine the impact of energy label changes on primary fossil energy consumption and predict energy consumption based on relevant variables. Here we show that upgrading energy labels leads to a significant reduction in primary fossil energy consumption, with potential cost savings ranging from €19.73 to €123.78 per year. Linear regression model accurately predicts energy consumption and demonstrates the effectiveness of energy label improvements in achieving energy efficiency.

The findings highlight the importance of energy labels in promoting energy efficiency and provide insights for energy renovations in the future. This study contributes to the understanding of energy efficiency in the residential sector. Future research can expand to renovation loans and their interest rates to upgrade a residential home's energy label to A.

This research was conducted in collaboration with the Data Lab department at STATER N.V, a Mortgages Service Company (https://stater.nl/).

The code and datasets can be found on GitHub:

https://github.com/markdiel/Thesis_Mark.git

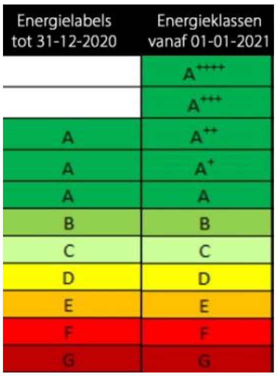
Table of contents

1. Introduction

Existing buildings represent a substantial contributor to greenhouse gas (GHG) emissions within the European Union, accounting for approximately 40% of total energy consumption (EC, 2020). The transition to sustainable and energy-efficient ways of operating has become an increasingly important agenda worldwide (EC, 2022). Even a bank like ABN AMRO is trying to support its customers in the transition to sustainability (n.d.). Making homes more sustainable is also becoming an increasingly important issue, says De Nederlandsche Bank (n.d.). They state that homeowners often refrain from investing in making their homes more sustainable, and this is partly due to financial bottlenecks.

These bottlenecks may potentially pose challenges due to the directives by the European Commission aimed at promoting enhanced energy performance in buildings, known as the Directive on Energy Performance of Buildings (EPBD) (EC, n.d.). Under EPBD, the Energy Performance Certificate (EPC) was established (EC, n.d.). EPC has labeled the energy performance of a building along a range from A to G and plays a key role in monitoring energy consumption in buildings.

In January 2021, a new calculation method was introduced to calculate a home's energy label. Called NTA 8800 (L'escaut,2021), it added the A+ categories. Figure 1 shows the old and new labels. *Appendix 4* shows the breakdown of the values that belong to each energy label. How the labels are divided into consumption categories will be explained later. The EU aims to be climate neutral by 2050 - an economy with zero net greenhouse gas emissions (EC, n.d.). It is therefore important to get energy labels of houses as high as possible to meet this goal.



Energielabels tot 31-12-2020	Energieklassen vanaf 01-01-2021
	A^{++++}
	A^{+++}
A	A^{++}
A	A^{+}
A	A
B	B
C	C
D	D
E	E
F	F
G	G

Figure 1 Energy label different methods

Literature review

Several existing studies have investigated the relationship between energy labels and their potential for energy savings. For instance, a notable study, titled "Uncertainty in Potential Savings from Improving Energy Labels" (Cozza et al., 2022), focused on modeling different scenarios for energy label-based retrofit targets in Switzerland's residential building sector for the year 2050. The findings of this study highlighted that adhering to a business-as-usual approach would not be sufficient to achieve the desired goals by 2050. Instead, it was concluded that a more aggressive policy, aiming to renovate all buildings to the highest energy efficiency standards, would be necessary to meet the targets effectively.

Further research was conducted in the area of "Financing energy-efficient housing" (IEA, 2007). The aim of this research was to examine policy measures and approaches aimed at addressing financial barriers to energy-efficient investments in existing housing. The research shows that simultaneously addressing multiple aspects of the financial barrier, promoting public-private sector cooperation, and implementing strong political will are essential for successful market transformation and increased private sector involvement in energy-efficient investments.

The findings of a study conducted by Brounen and Kok (2011) indicate that homebuyers demonstrate a willingness to pay a premium for homes that have been labeled as more energy-efficient or "green". This suggests that higher energy labels contribute to increased desirability and market value when selling a house. This can reduce the risk of a renovation. Another study conducted by Aydin et al. (2019) proved the positive relationship between energy labels and the sales process. This study shows that houses with an energy label for sale experienced a decrease in sales time. Moreover, having a high energy label speeds up sales compared to the lower label; they found that houses with label A experienced a 28 percent increase in sales speed.

The existing literature reveals a positive relationship between improvements in energy labels, reduced energy consumption, increased property value, and accelerated sales. These findings underscore the significance of risk management in the context of renovations aimed at achieving higher energy labels.

Research context

A research gap can be explored, necessitating an investigation into the impact of energy label upgrades on Dutch home energy savings. Consequently, the central research question of this thesis is as follows: "**What is the relationship between a home's energy label and energy consumption per square meter, and how much energy and cost savings can be achieved by upgrading the energy label**?" The outcome of this study can also be compared with the study conducted in Switzerland. This study only looks at energy consumption from electricity, not from other energy sources, such as gas. This is due to the timeframe set for the study.

The resulting data science questions are: "Can a causal relationship be established between increasing energy labels and energy consumption?", "Can energy consumption be accurately predicted based on relevant variables and models?", "What is the magnitude of the impact on energy consumption when energy labels are changed, specifically in terms of energy efficiency improvements?" and "What is the magnitude of the impact on energy consumption when energy labels are changed, specifically in terms of energy efficiency improvements?".

The hypothesis for this study is as follows: "There exists a significant relationship between a home's energy label and energy consumption per square meter, and upgrading the energy label results in both energy and cost savings." This relates to the research question and will be answered at the end of the study. With the aim of evaluating the feasibility and benefits of such a transition, this study focuses on investigating the relationship between a home's energy label and monthly energy consumption, looking specifically at the potential energy and cost savings that can be achieved by upgrading the energy label.

This study is in collaboration with another related study investigating the effect of energy label changes on house prices (Mawed, D., 2023). With this collaboration, a picture can be painted of the feasibility of issuing renovation loans to improve the energy label in the future. This will serve as a basis for further research to determine the financial risk in relation to renovation loans.

To answer the research question, this thesis begins by explaining the process of data collection and the subsequent steps of data cleaning and pre-processing. The methods and choices for analysis are then described in detail. These chosen methods lay the foundation for conducting the necessary analyses to quantify the relationship and achievable energy and cost savings in relation to energy labels. All this comes together in the conclusion following the discussion.

2. Data

This chapter focuses on the collection, description, and preparation of data for the analysis of the relationship between energy labels and energy consumption.

2.1 Data Collection

For this study, two datasets were used. One of the datasets was obtained from the government platform ep-online.nl (2023), which provides public data on energy labels. This dataset serves as a valuable resource for analyzing the relationship between energy labels and energy consumption, enabling the application of the selected research methods to address the research questions effectively.

To ensure the accuracy of this study, the most recent version of the ep-online.nl website was downloaded in June 2023. Furthermore, an older version of the dataset, collected in 2022, was also obtained. This 2022 dataset is not used for the analysis but for the difference in difference to have a time sample. These datasets encompass a comprehensive collection of house data, comprising five million rows, which includes information on the energy label assigned to each property as well as its corresponding fossil energy consumption. It is important to note that the focus of this study is on residential properties, as the aim is to analyze the potential monthly cost savings for homeowners. The dataset excludes business premises, as they fall outside the scope of this research. In-depth clarification regarding the structure and content of the EP data is presented in the subsequent subchapter or can be found in *Appendix 1*.

To calculate the primary fossil energy consumption per year for a house, it is crucial to consider the square meters of the property. However, the EP data does not include information on the square meters of the houses. Therefore, an additional data source needed to be incorporated. Funda, a prominent online platform for real estate listings in the Netherlands, was chosen as the supplementary data source (n.d.). Funda provides comprehensive details about properties, including their characteristics, prices, and transaction information. In order to access the required data for this project, scraping tools were explored and utilized. By combining the EP data with the relevant information obtained from Funda, the study was able to obtain the necessary data to calculate the fossil energy consumption per square meter per year for each house.

Data scraping was conducted using Funda Scraper, version 0.0.3, a Python library (Chien, W. 2023), to gather information on houses sold within a specific timeframe, spanning from the end of 2021 to the second quarter of 2023. The scraping process was performed city by city, as listed in *Appendix 2*. The datasets obtained for each city were subsequently merged into a consolidated dataset. Among the numerous variables available in the Funda data, the city, house type, price, price per square meter, living area, energy label, and house age were identified as the most pertinent for the purposes of this study. In-depth clarification regarding the structure and content of the Funda dataset is presented in the subsequent subchapter or can be found in *Appendix 3*.

2.2 Description of the data

2.2.1 EP-Data

For this study, we used the 2023 EP dataset. The Difference in Difference method was performed on two datasets. One is the EP dataset from 2023, and the other is the same dataset but from 2022. This is further explained in the Method and Analyses section. In total, the EP dataset contains 39 variables which relate to the information about the houses that were collected. Further explanation of what each variable refers to is included in *Appendix 1*. Not all the variables are needed for this study because not all of them contain relevant information to help answer the research question.

The most important variables retained in the study are: Building class, Calculation type, Energy class, Building type, Building subtype, Postcode, House number, Primary fossil energy, House letter, and House number addition. The energy class and Primary fossil energy are the most important variables for this study. Energy class contains the label of each house, which can be between A++++ and G for houses. These labels are stored as characters.

Primary fossil energy contains a value indicating energy consumption per square meter per year in kilowatt hours (KWH). This is done for each house in the dataset. The value is indicated by the term EP2. This means Primary fossil energy consumption with energy measures at area level quality statements, in kWh per m^2 of usable area per year (kWh/m^2.yr) (RVO, 2023). Primary fossil energy has missing values, which are further explained in the data preparation section.

In total, the 2023 EP dataset contains 5,0599,050 observations. This is therefore just over five million observations, divided between residential (W = 4,860,097) and non-residential (U = 199,853). For this study, it is only relevant to extract residential houses. To provide initial insights, all U values are removed. It contained nearly 200,000 observations. This is done because business premises are outside the scope of this study. Also because the square meters cannot be retrieved in this study. As a result, they are not included. From there, we looked at primary fossil energy. It had only 946,342 observations filled. We are not removing the missing rows as a prediction model will be used later. The prediction model serves to fill in the missing values. In order to find a good way if you don't know what the pledge primary fossil energy is, you can still predicate it to calculate the savings that can be gained for renovation to a better energy label.

When examining the primary fossil energy variable, it becomes apparent that there are 3,845,347 missing observations, leaving only 1,014,750 observations with primary fossil energy values filled. Although this may appear as a substantial amount of missing data, it is essential to assess the remaining dataset after merging with the Funda dataset. Moreover, it is worth noting that there is no missingness in the key variables that are crucial for the analysis. There is only missingness in the additions to the address. This is not a big concern because it has to do with the merging of the Funda and EP dataset. Further exposition of the EP 2023 dataset can be found in the *Appendix 5*.

2.2.2 Funda data

As previously explained in the collection part, there are several datasets that together represent the Funda dataset. To get one dataset, 23 city datasets had to be merged with each other. Eventually, after merging, a dataset of 48,148 observations and 17 variables emerged. With this, there is a large dataset with information on houses. It also makes this Funda dataset the main dataset of this study. Analyses of the Funda data can be found in *Appendix 6*. Not all variables are important in this study. Of the 17 original scrapped variables, 8 retained.

Those 8 independent variables are: House type, Living area, energy label, postcode (ZIP), House age, House number, House letter plus addition, and Extra. With these variables, living area is the most important. This is to calculate the final primary fossil energy consumed by a house per year. To discover the monthly cost and savings.

Now that the variables have been mentioned, it is important to see how many missing values appear in this dataset (*Figure 2).* Only in extra and house letter plus addition are there empty values. This does not matter because this can occur when houses have no extra letters. Other than that, the Funda dataset has no missingness.

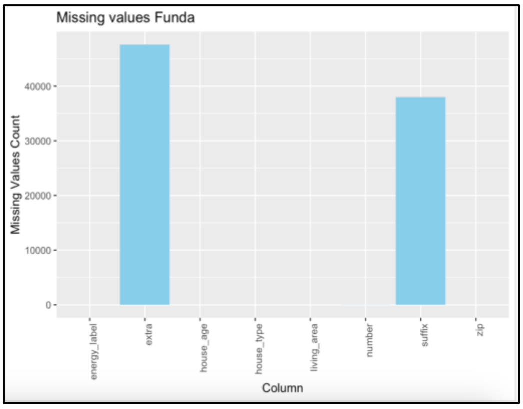


Figure 2 Missing values Funda

In addition to assessing the energy label distribution, it is important to examine the distribution of square meters in relation to the energy labels. This analysis allows us to evaluate whether there is sufficient representation across all square meter ranges. *Figure 3* illustrates a clear distribution pattern, with prominent peaks observed around the 100 square meter mark. Furthermore, it is evident that the Funda dataset contains a considerable number of properties with energy labels A and C. This distribution analysis provides valuable insights into the availability of data for different square meter ranges and energy labels.

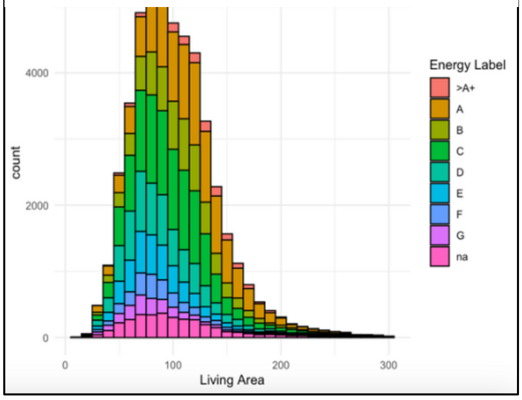


Figure 3 Distribution Energy Label and Living Area

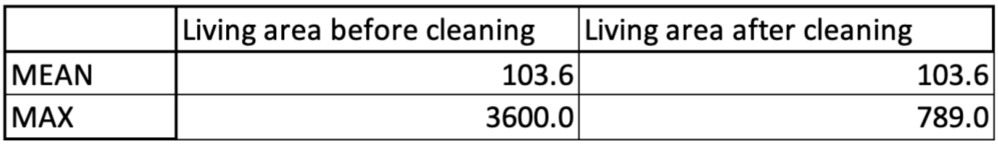
2.3 Data Preparation

Now that the two datasets have been described, they need to be merged. This is done based on three main independent variables. Those are Postcode, House number, and House letter plus number addition. Here, the Funda dataset is seen as the main dataset, with the EP dataset providing an addition. This is because the living area is needed for this study. This can only be found in the Funda dataset.

After merging the two datasets, 44,227 observations remain to be merged. This dataset is called mergedEPFU. The complete structure of the merged dataset can be found in *Appendix 7*. The number of observations is lower than the Funda dataset was initially. Thus, 4,000 observations were lost since they could not be matched. This small number has no major impact on the study. Furthermore, seven independent variables remain. These are the main features that will be used to work with the prediction model.

After the merge, cleaning needs to be done. First, to start with the outliers occurring in the merged dataset. It is also important to give interpretation to the outliers and how they may arise. Further, clean the independent variables for the prediction model. Outliers should also be considered in the merged dataset. Here, it is interesting to highlight three variables. These are Living area, House age, and fossil energy property. Looking at *Table 1,* the highest value is 3600 m2. This is obviously a very high value for living area. Anything above 1000 m2 is filtered out because this is a too extreme value to work with. It also occurs very rarely for this kind of square footage for a house. After adjusting, the maximum living area comes to 798 m2. The mean stays the same because of the low number of values that are removed.

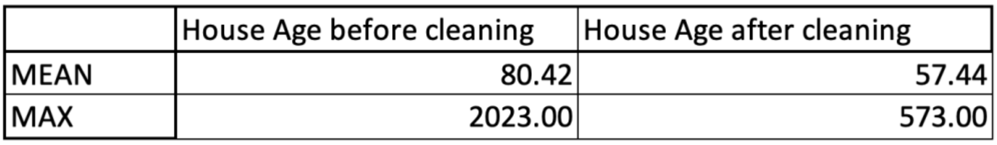
Table 1 Living area before and after cleaning



	Living area before cleaning	Living area after cleaning
MEAN	103.6	103.6
MAX	3600.0	789.0

House age also had outliers in the dataset. Initially, the maximum value was 2023 in the variable. This is shown in *Table 2*. It is not possible for a house to have been built 2023 years ago. Therefore, this is also an outlier that needs to be adjusted. The explanation for this is that the value when a house was built is not included. It can be because the calculation is done automatically in the system, resulting in 2023. It cannot be determined with certainty whether these are the reasons. Because this value is common, the house age is set to zero for those values. This is to avoid losing further information. As a result, the mean goes down. In the prediction model, less weight will be assigned to the independent variable house age because of this transformation.

Table 2 House age before and after cleaning



	House Age before cleaning	House Age after cleaning
MEAN	80.42	57.44
MAX	2023.00	573.00

Table 3 Primary fossil Energy beforecleaning

	Primary fossil Energie before cleaning
MEAN	3903
MAX	3385143362

The next thing to clean up is primary fossil energy. In *Table 3,* the outlier contains such a huge value that this is not possible for one square meter per year for a residential house. Why this big outlier is in the data couldn't be found. As a result, these outliers are not included in further analyses. Everything above 1000 KWH is going to be removed for better visualization.

Looking at the histogram in *Figure 4*. It is easy to see that after 550 KWH per square meter per year, almost no values occur. Following that, the boxplot in *Figure 5* shows that the outliers all fall into energy label G. These are exceptional values where a timeline is made for the study.

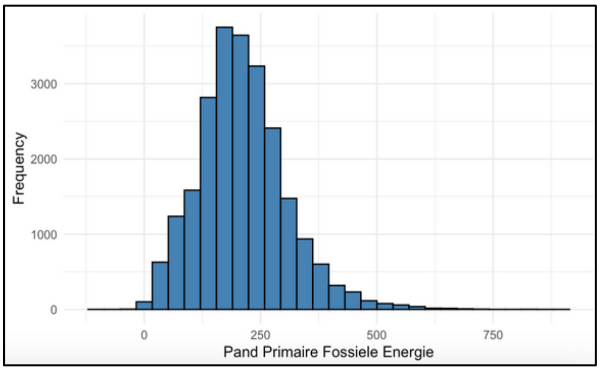


Figure 4 Histogram of frequency of Primary Fossil Energy

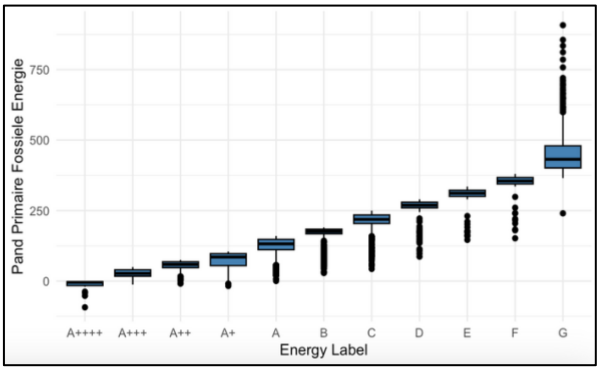
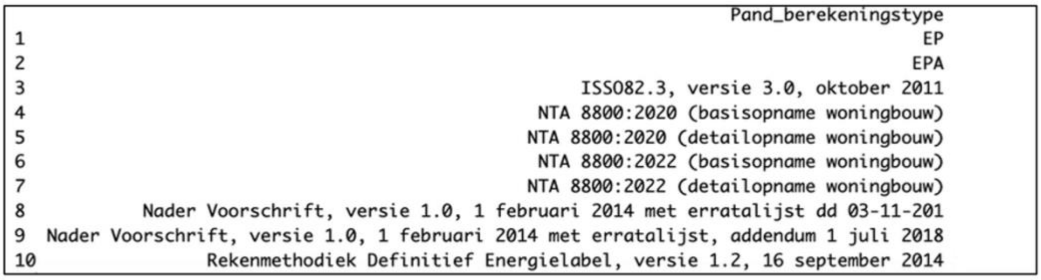


Figure 5 Boxplot Energy label

The last thing to be touched upon in the cleaning is the calculation type. There should be two measurement techniques in the dataset to calculate the energy label. These techniques are NEN 7120 (before 2021) and NTA 8800 (after 2021). If we visualize the list, it is easy to see that there are several derivatives of these two techniques. This can be seen in *Figure 6*. It is easy to see what is before 2021 and what is after. The EP and EPA are exceptions here that have been removed. These occurred only 200 times in total in the full dataset. It was not possible to determine what these terms meant.

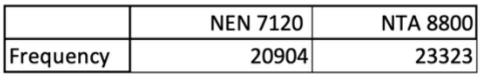
Figure 6 All calculation types for Energy Label



```
                                                         Pand_berekeningstype
1                                                                          EP
2                                                                         EPA
3                                             ISSO82.3, versie 3.0, oktober 2011
4                                          NTA 8800:2020 (basisopname woningbouw)
5                                         NTA 8800:2020 (detailopname woningbouw)
6                                          NTA 8800:2022 (basisopname woningbouw)
7                                         NTA 8800:2022 (detailopname woningbouw)
8           Nader Voorschrift, versie 1.0, 1 februari 2014 met erratalijst dd 03-11-201
9   Nader Voorschrift, versie 1.0, 1 februari 2014 met erratalijst, addendum 1 juli 2018
10                   Rekenmethodiek Definitief Energielabel, versie 1.2, 16 september 2014
```

These values have been transposed to two measurement techniques. There are ultimately 20,904 with the measuring technique NEN 7120 and 23,323 with the measuring technique NTA 8800 (*Table 4)*. This is very interesting because the missingness amounted is also 20,904 in the dataset. When further looking, we also see that all the measurements with NEN 7120 are empty.

Table 4 Amount of each calculation type

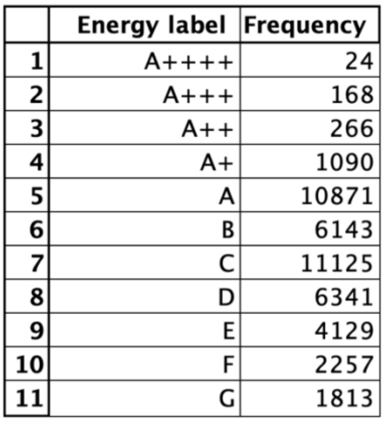


	NEN 7120	NTA 8800
Frequency	20904	23323

With this, it can be concluded that in the old measurement method, the primary fossil energy was not stored by the government. This may leave out many values that could be important in the study. For this, a prediction model has been made that will return in the following chapters.

Further attention should be dedicated to assessing the distribution of energy labels within the merged dataset, as it holds crucial implications for the interpretation of subsequent study outcomes. *Table 5* reveals an uneven distribution across the various energy label categories. Particularly notable is the scarcity of values beyond the A+ energy label in comparison to the lower labels. This discrepancy should be considered during the training and testing phases of the prediction models and when examining subsequent outcomes. However, it is important to highlight that the energy labels most pertinent to this study, specifically Labels A to G, demonstrate satisfactory data representation, with Labels A and C being particularly well-populated. For full analyses, see *Appendix 8.*

Table 5 Frequency of all Energy Labels



	Energy label	**Frequency**
1	A++++	24
2	A+++	168
3	A++	266
4	A+	1090
5	A	10871
6	B	6143
7	C	11125
8	D	6341
9	E	4129
10	F	2257
11	G	1813

3. Method

In this chapter, we explore the methods and models used to answer the research question. The data science questions are put out there to see how they can start to be answered. From there, we can work towards the results and conclusion.

3.1 Description of the method used

This study employed two distinct programming languages, namely Python and RStudio. RStudio predominantly facilitated the data cleaning process and training of the utilized models. Meanwhile, Python was employed specifically for implementing the Difference-in-Differences analysis, which will be explained further in subsequent sections to outline the rationale, methodology, and specific steps involved.

Based on the research questions of this study, we can translate them into the following data science questions. These can be answered using data science methods:

- Can a causal relationship be established between increasing energy labels and energy consumption?
- Can energy consumption be accurately predicted based on relevant variables and models?
- What is the magnitude of the impact on energy consumption when energy labels are changed, specifically in terms of energy efficiency improvements?
- What are the energy savings in terms of fossil fuel consumption when transitioning from a lower energy label to an energy label of A?

3.2 Difference in Difference

To investigate the causal effects of energy label changes on primary fossil energy consumption, a difference-in-differences (DiD) model was employed (Jiménez & Perdiguero, 2019). The DiD model facilitates the evaluation of causal effects by comparing a treatment group that underwent energy label changes with a control group that did not receive any label modifications. This approach enables the identification and assessment of the causal impact of energy label changes on primary fossil energy consumption. This is done to look at the correctness of the dataset.

This study utilized two EP datasets, with the starting point in 2022 and the endpoint in 2023. The focus was on identifying cases where the energy label transitioned to label A. This classification allowed for the establishment of a treatment group comprising the labels that transitioned to A, while the no treatment group consisted of the labels that did not undergo any changes.

To enhance comparability between the treatment and control groups, a propensity score matrix was employed (McMurry et al., 2015). The propensity score represents the probability of receiving one of the treatments being compared, considering the measured covariates. In this study, the energy labels that underwent changes were-matched with energy labels that remained unchanged but shared the same probability of achieving a similar outcome. This matching process significantly improves the accuracy of the difference-in-differences analysis by effectively addressing potential confounding factors and enhancing the comparability between the treatment and control groups.

Due to the constraints imposed by the Jupiter Notebook, the initial analysis was conducted using a sample size of 8,000 treatment and 8,000 control cases for the propensity score matrix. There were more cases that had a change to A but couldn't run. It is important to acknowledge that this limited sample size may introduce a potential for distorted representation. However, it is assumed that no systematic bias was introduced during this process. This is due to it not being a complex dataset. Furthermore, the energy labels were assigned numerical values, with Label A assigned as 1 and the highest label, Label G, designated as 7. As a result, the difference-in-differences analysis yielded negative results. It is worth noting that if the numerical values had been reversed, positive results would have been observed instead. Further explanations and the code can be found in *Appendix 9*.

If the analysis using this method fails to establish a causal relationship, it would prompt the conclusion that the EP dataset may not precisely reflect the true underlying relationship. This assumption is rooted in the expectation that a discernible causal relationship should be observable, given the clear differentiation between the different energy labels. It is anticipated that these energy labels adhere to distinct frameworks, thereby providing a foundational basis for inferring a causal impact.

3.3 Data modeling

To answer the second data science question, prediction models were used to predict primary fossil energy consumption based on relevant variables. With these models, post-renovation energy consumption could be estimated after energy label improvements. In addition, it was crucial to address missing data points in the dataset. Around 20,904 values were missing in the primary fossil energy column, while the other columns were sufficiently filled, as evidenced by the availability of 23,323 values in that column. These missing values must be predicted by the prediction model.

3.3.1 Variables

In training the prediction models, several variables were considered. The primary fossil energy consumption served as the dependent variable, while the independent variables included the energy class, property building sub-type, house type, living area, and house age. These independent variables were selected based on their logical relationship with energy consumption and associated costs. where the energy label is the main feature of the model, and the others make the variation between energy labels that can occur to predict an accurate primary fossil energy. Furthermore, some variables were aligned with those utilized in the study conducted in Switzerland, enhancing the comparability and potential insights derived from the findings.

3.3.2 Training and Test sets

To ensure a robust evaluation of the models' performance, the dataset was divided into two distinct subsets: a training dataset comprising 60% of the data and a test dataset comprising the remaining 40%. This was chosen because above the 60% training set, the results deteriorated in the prediction model. This is because the model becomes overfit. The training dataset was utilized to construct and train the models, and the testing dataset serves as an independent set for assessing the models' predictive performance.

There are two models employed, specifically **Linear regression** and **Gradient boosting**, to predict energy consumption based on the selected independent variables. The decision to utilize these models was informed by insights from the existing literature. One study delved into data-driven building energy consumption prediction studies, examining the efficacy of both simple and complex models in forecasting consumption patterns (Amasyali & El-Gohary, 2018). The findings indicate that a single universal model cannot be applied across all scenarios, emphasizing the need for tailored model development that considers specific application requirements. This entails a comprehensive analysis of data properties and the selection of suitable machine learning algorithms. Consequently, in this study, a simple and a complex model were chosen to determine the optimal fit for the given situation.

3.3.3 Linear regression

Linear regression is a statistical modeling technique that aims to establish a linear relationship between the dependent variable and independent variables (Su et al., 2012). By searching for the best fitting line that minimizes the differences between predicted and actual values, linear regression provides insights into the direction and magnitude of the impact of independent variables on the dependent variable. This model assumes a linear relationship between the variables, making it a suitable choice for this study given the assumption of linearity. Linear regression offers valuable insights into the relationships between all the variables and their impact on energy consumption patterns.

3.3.4 Gradient boosting

In contrast, gradient boosting is a machine learning algorithm that combines multiple weak models, decision trees in this case, to create a robust predictive model (Bentéjac et al., 2020). Unlike linear regression, gradient boosting can capture complex relationships between variables. Through an iterative process, it sequentially corrects the errors made by previous models, minimizing the loss function and optimizing the model's predictive performance. Maybe a complex relationship can be found through this model. A loop was created for this machine learning technique to determine the appropriate number of trees and depth for this model. By incorporating both Linear regression and Gradient boosting, this study offers the flexibility to choose between a simple and more advanced model. This allows for consideration of which model is more suitable for this scenario. The inclusion of both models enhances the comprehensiveness and accuracy of the predictions made in this study.

3.4 Data Analysis

To address the remaining questions, the predicted and known values are aggregated, and a comprehensive analysis is conducted on the dataset. Annual energy consumption amounts are calculated for each house, enabling further calculations to be performed on a monthly basis. It is important to note that the analysis does not account for the typical seasonal variation in energy consumption, where more energy is typically consumed during the winter months compared to the summer months. Instead, an average per year is considered.

In the calculations, a fixed value representing the current energy price is employed. This is done because of the energy ceiling that is in place (EZK, 2023). This means paying 40 cents per KWH of electricity that is consumed. It should be acknowledged that the monthly payment may fluctuate in the future due to changes in the fixed price. However, using this fixed value allows for the assessment of potential savings across all observations in the dataset on an average basis.

By conducting these calculations, the study provides insights into the energy savings achieved and their magnitudes across the entire dataset. This analysis offers valuable information on the potential cost savings associated with upgrading energy labels and serves as a basis for understanding the financial implications of energy efficiency improvements in residential buildings.

4. Results

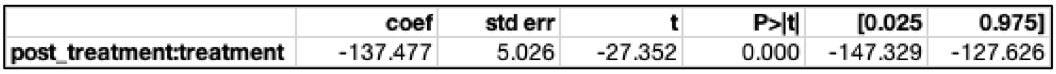
The Results section presents the findings obtained from various analyses conducted in this study, including the difference-in-differences method, the two employed machine learning techniques (linear regression and gradient boosting), and additional calculations to get to the savings. How these were implemented can be found in the GitHub *Appendix 9*.

4.1 Overview of the results

4.1.1 Difference in Difference

The purpose of the difference in difference (DiD) is to see if there is a causal relationship between energy label change and primary fossil energy. If there is a causal effect, then there should be a P-value ¡ below 0.05. If this is not the case and it falls above 0.05, then there is no causal effect between the treatment and no treatment group (Tan, S. 2010). The following results come from the DiD and can be seen in *Table 6.*

Table 6 Outcome DiD



	coef	std err	t	P>|t|	[0.025	0.975]
post_treatment:treatment	-137.477	5.026	-27.352	0.000	-147.329	-127.626

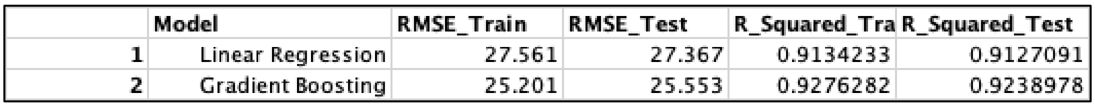
The estimated difference of -137.4775 was found to be statistically significant by the p-value < 0.001. These findings show that the energy label change had a substantial impact on the difference in primary fossil energy consumption between the treatment and control groups. More specifically, the results show that upgrading the energy label resulted in a noticeable reduction in primary fossil energy consumption in the treated group compared to the control group.

Based on this, it can be concluded that the EP-2023 dataset is reliable and accurate for further examination in this study. An underpinning prediction model can be built on this. The full DiD results can be found in *Appendix 10*.

4.1.2 Outcome prediction model

Two prediction models, linear regression and gradient boosting, were employed in this study using the available primary fossil energy data, encompassing approximately 23,254 observations. To assess the predictive performance of these models, the dataset was split into training and test sets. Initially, both models were trained on the training set to learn the underlying patterns and relationships. Here, gradient boosting was run on a loop to find the best adjustments for the model. Subsequently, the trained models were evaluated on the test set to measure their accuracy in predicting primary fossil energy consumption. The accuracy results of both models are depicted in *Table 7*, providing insights into their respective performance levels.

Table 7 Comparison both prediction models



	Model	RMSE_Train	RMSE_Test	R_Squared_Tra	R_Squared_Test
1	Linear Regression	27.561	27.367	0.9134233	0.9127091
2	Gradient Boosting	25.201	25.553	0.9276282	0.9238978

Both the linear regression and gradient boosting models achieved an accuracy of over 90% in predicting primary fossil energy consumption. These results demonstrate the effectiveness of both models in accurately forecasting missing values. Furthermore, it is crucial to examine the construction of these models in greater detail. *Figures 7 and 8* illustrate the predicted values plotted against the actual values, providing a visual representation of the models' performance and their ability to capture the underlying patterns and trends in the data.

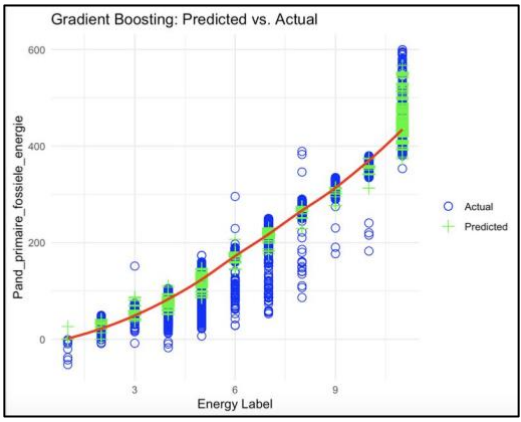


Figure 7 Linear regression

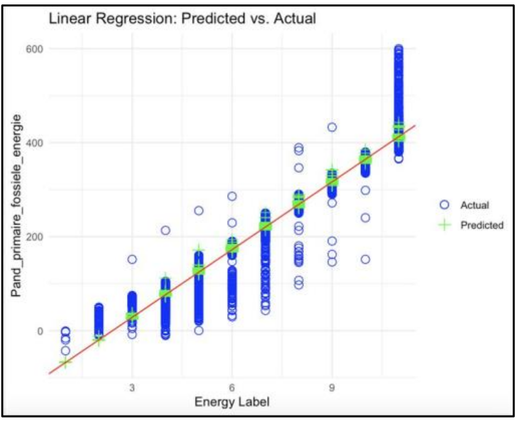
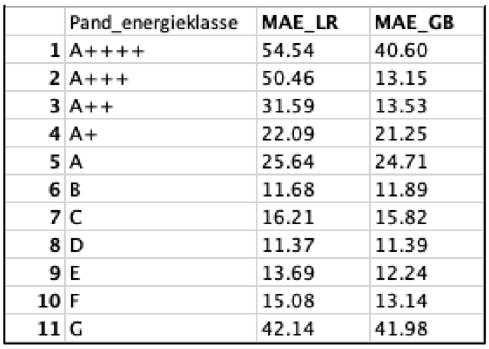


Figure 8 Gradient boosting

Upon closer examination of the figures, it becomes evident that the gradient boosting model exhibits a more ascending line, while the linear regression model closely aligns with the observations. Notably, the fact that linear regression dips below zero is of significance. This observation is particularly relevant because the highest energy label, A++++, often corresponds to primary fossil energy consumption below or equal to zero. Thus, the linear regression model's ability to capture this pattern is noteworthy.

In addition, a margin of error analysis for each energy label was performed for both prediction models. *Table 8* shows that energy labels above A+ show a significantly higher margin of error in both models. This observation can be attributed to the limited representation of data points in these higher energy label categories. It is also evident in label G. This is probably due to the wide range of values that the G label covers. It is important to note that the models perform well in predicting values between A and F, which is the focus of this study.

Table 8 Error margin for both prediction models



	Pand_energieklasse	MAE_LR	MAE_GB
1	A++++	54.54	40.60
2	A+++	50.46	13.15
3	A++	31.59	13.53
4	A+	22.09	21.25
5	A	25.64	24.71
6	B	11.68	11.89
7	C	16.21	15.82
8	D	11.37	11.39
9	E	13.69	12.24
10	F	15.08	13.14
11	G	42.14	41.98

Based on the insights gained from both models, it is evident that they yield similar performance. A discrepancy can only be seen in the higher energy labels. Where Linear regression dips below 0 and Gradient boosting does not have that. Consequently, in this study, the decision was made to proceed with the linear regression model for predicting the missing values.

4.1.3 Implementation prediction model

Before the predicted dataset was merged with the actual dataset, it was important to use a boxplot to visualize the distribution. *Figures 9 and 10* show that there is minimal difference between the two datasets, indicating that they can be merged seamlessly. This integration allows comprehensive analysis of the dataset, combining both predicted and actual values.

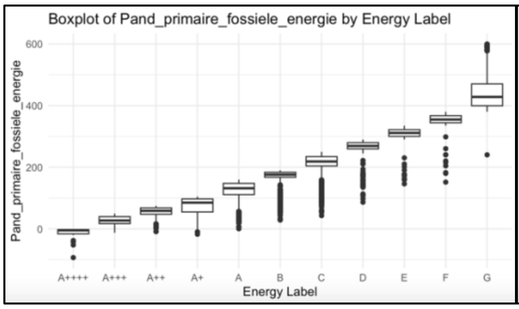


Figure 9 Boxplot primary fossil energy.

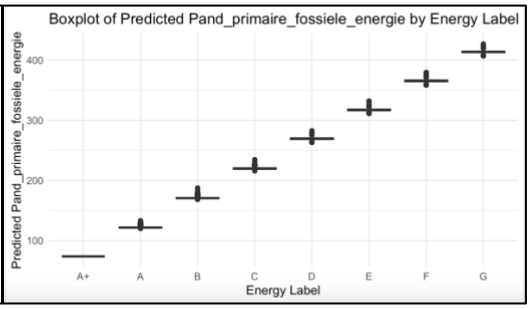


Figure 10 Boxplot predicted primary fossil energy.

Upon examination of *Table 9*, it becomes evident that there is a minimal discrepancy between the two datasets. When the datasets are segregated based on average energy labels, it can be observed that the differences are negligible, with values close to zero. This indicates that there is no significant distinction between the predicted dataset and the existing values.

Table 9 Difference in prediction and actual

	Difference, predicted and actual dataset
A	-2.65
B	-1.22
C	-8.63
D	-3.13
E	-9.72
F	-6.74
G	39.99

4.2 Overall analyses

According to the above results, further analysis is required to gain a deeper understanding of the specific impact of each energy label on primary fossil energy. By comparing the mean Primary fossil energy for each energy label, we can assess the average effect of changing the energy label on Primary fossil energy. These findings support the hypothesis that changes in the energy label variable indeed influence consumption, with higher energy label ratings being associated with lower energy consumption.

The below results show the percentage change in primary fossil energy per m2 per yr., when transitioning from the respective energy labels to energy label A

- Residential houses with Energy label B: Approximately 28.84% decrease in primary fossil energy when its energy label changed to A.
- Residential properties with Energy label C: Approximately 44.26% decrease in primary fossil energy when its energy label changed to A.
- Residential properties with Energy label D: Approximately 54.68% decrease in primary fossil energy when its energy label changed to A.
- Residential properties with Energy label E: Approximately 61.24% decrease in primary fossil energy when its energy label changed to A.
- Residential properties with Energy label F: Approximately 66.20% decrease in primary fossil energy when its energy label changed to A.
- Residential properties with Energy label G: Approximately 71.77% decrease in primary fossil energy when its energy label changed to A.

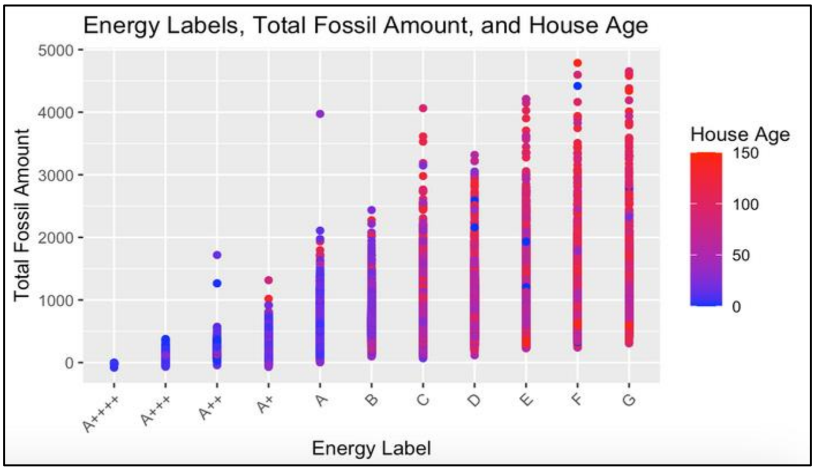
Additionally, to the savings percentages it is also important to express this in financial terms. In the calculation of potential savings when upgrading from one energy label to Label A, a price of 40 cents per kilowatt-hour (KWh) is considered. This value represents the average cost savings per square meter per year by transitioning to Label A. The savings can be seen below.

- Residential houses with Energy label B: Approximately €19,73 can be saved when its energy label changed to A.
- Residential houses with Energy label C: Approximately €38,67 can be saved when its energy label changed to A.
- Residential properties with Energy label D: Approximately €58,74 can be saved when its energy label changed to A.
- Residential properties with Energy label E: Approximately €76,92 can be saved when its energy label changed to A.
- Residential properties with Energy label F: Approximately €95,38 can be saved when its energy label changed to A.
- Residential properties with Energy label G: Approximately €123,78 can be saved when its energy label changed to A.

This analysis demonstrates that the potential savings can be substantial, particularly for larger residential properties with more square meters. The most significant jumps in cost savings occur when upgrading from energy labels C and D to Label A, as these are the energy labels commonly associated with residential houses eligible for energy renovations. By calculating the average savings for specific square meter ranges, a more accurate estimation of the potential cost savings benefits can be obtained.

Figure 11 visually presents the distribution of different house ages on energy label classes using distinct colors (blue and red). Red represents older houses, up to 150 years old, while blue represents newer houses. The transition between the colors is gradual. It is readily apparent that newer houses tend to have higher energy labels compared to older houses. This observation can be attributed to the regulatory requirements that mandate energy-efficient standards for newly constructed houses (Maessen, H. n.d.).

Table 11 Energy label compared with house age



5. Conclusion

This study aimed to analyze the relationship between energy labels and energy consumption in residential properties. By examining two merged datasets, namely the EP dataset and the Funda dataset, we collected comprehensive information on energy labels, primary fossil energy consumption, and relevant property characteristics. The collected data allowed us to address the research questions and explore the potential cost savings associated with upgrading energy labels.

The analysis employed various methods and models, including the Difference-in-Differences (DiD) method, Linear regression, and Gradient boosting. The results yielded valuable insights into the relationship between energy labels and energy consumption, as well as predictions for primary fossil energy consumption based on relevant variables.

The difference-in-differences analysis revealed a significant causal relationship between energy label changes and primary fossil energy consumption. Upgrading energy labels resulted in a noticeable reduction in energy consumption in the treated group compared to the control group. This finding provides strong evidence that improving energy labels can lead to energy efficiency improvements and associated cost savings for homeowners.

The prediction models, both Linear regression and Gradient boosting, demonstrated high accuracy in estimating primary fossil energy consumption based on selected independent variables. These models were able to predict missing values effectively, providing valuable insights into the potential energy consumption of residential properties. The Linear regression was further used to analyze the predicted and known values and it showed minimal discrepancy between the datasets. This integration allowed for a comprehensive analysis of the dataset and provided a basis for calculating potential cost savings.

The analysis of the data indicated that transitioning from lower energy labels to energy label A can result in significant energy savings. The percentage decrease in primary fossil energy consumption per square meter per year ranged from approximately 28.84% for houses with energy label B to 71.77% for houses with energy label G. Translating these savings into financial terms, homeowners can potentially save between €19.73 and €123.78 per square meter per year by upgrading their energy labels to Label A. Moreover, the examination of house ages in relation to energy label classes revealed that newer houses tend to have higher energy labels. This observation reflects the regulatory requirements for energy-efficient standards in newly constructed houses.

In conclusion, this study provides valuable insights into the relationship between energy labels and energy consumption in residential properties. The findings indicate that upgrading energy labels can lead to significant energy savings and associated cost reductions. The prediction models offer a means to estimate energy consumption and potential savings, empowering homeowners to make informed decisions about energy renovations.

5.1 Discussion

But there are also some limitations to this study and parts that can be discussed. This study only looked at energy consumption in the form of electricity, not gas. This may further widen the scope explored in this study in terms of savings. Also, limiting data was available above the A+ category.

There was also no further analysis at the individual level. By looking at the individual level, the outcomes could have been explored in even greater depth. In addition, it was not useful in this study because of the energy ceiling that is now in place. This made it more difficult to draw firm conclusions at the individual level. Further research can further expand all these factors for an even deeper understanding.

And as a final discussion point, the study did not look at the energy companies themselves. How these parties predict the consumption of a residential house for their gain. This was not considered because Stater was used to see what data was available. Overall, this study emphasizes the potential benefits of energy label improvements and provides valuable insights for stakeholders to look further into the risk of renovation loans. Further research should find out what loans and interest rates should be charged to get an energy label to A.

6. Bibliography

2022 juli - In 2023 naar Energielabel C. Bent u er klaar voor? - Centercon. (n.d.). https://centercon.nl/nieuws/2022-juli-in-2023-naar-energielabel-c-bent-u-er-klaar-voor

2050 long-term strategy. (n.d.). Climate Action. https://climate.ec.europa.eu/eu-action/climate-strategies-targets/2050-long-term-strategy_en

Amasyali, K., & El-Gohary, N. (2018). A review of data-driven building energy consumption prediction studies. *Renewable & Sustainable Energy Reviews*, *81*, 1192–1205. https://doi.org/10.1016/j.rser.2017.04.095

Aydin, E., Correa, S.B. and Brounen, D. (2019) 'Energy performance certification and time on the market', Journal of Environmental Economics and Management, 98, p. 102270. doi: 10.1016/j.jeem.2019.102270.

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2020). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Brounen, D., & Kok, N. (2011). On the economics of energy labels in the housing market. *Journal of Environmental Economics and Management*, *62*(2), 166–179. https://doi.org/10.1016/j.jeem.2010.11.006

Certificates and inspections. (n.d.). Energy. https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/certificates-and-inspections_en

Chien, W. (2023). Funda scraper (Version 0.0.3) [Library]. Available from https://pypi.org/project/funda-scraper/

Cozza, S., Patel, M. K., & Chambers, J. (2022). Uncertainty in potential savings from improving energy label: A Monte Carlo study of the Swiss residential buildings. *Energy and Buildings*, *271*, 112333. https://doi.org/10.1016/j.enbuild.2022.112333

Energieprestatie indicatoren - BENG. (n.d.). RVO.nl. https://www.rvo.nl/onderwerpen/wetten-en-regels-gebouwen/beng/indicatoren

Energy and the Green Deal. (2022, April 8). European Commission. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal/energy-and-green-deal_en

Energy performance of buildings directive. (n.d.). Energy. https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/energy-performance-buildings-directive_en

Financiering voor de verduurzaming van de woningvoorraad. (n.d.). https://www.dnb.nl/publicaties/publicaties-onderzoek/analyse/financiering-voor-de-verduurzaming-van-de-woningvoorraad/

In focus: Energy efficiency in buildings. (2020, February 17). European Commission. https://commission.europa.eu/news/focus-energy-efficiency-buildings-2020-02-17_en

Jiménez, J., & Perdiguero, J. (2019). Difference-in-Difference. In *Springer eBooks* (pp. 551–555). https://doi.org/10.1007/978-1-4614-7753-2_664

Maessen, H. (n.d.). *Verplichte toets energielabel bij oplevering nieuwbouw*. Atriensis. https://www.atriensis.nl/nieuwsbericht-data/verplichte-toets-energielabel-bij-oplevering-nieuwbouw#:~:text=Voor%20alle%20nieuwbouwwoningen%20is%20vanaf,%C3%A9n%20door%20Bouw%2D%20en%20woningtoezicht.

Mawed, D. (2023). *Exploring the Impact of Energy Labels on Residential Properties Prices: A Data-Driven Analysis* [MA, thesis].

McMurry, T. L., Hu, Y., Blackstone, E. H., & Kozower, B. D. (2015). Propensity scores: Methods, considerations, and applications in the Journal of Thoracic and Cardiovascular Surgery. *The Journal of Thoracic and Cardiovascular Surgery*, *150*(1), 14–19. https://doi.org/10.1016/j.jtcvs.2015.03.057

Ministerie van Economische Zaken en Klimaat. (2022, October 13). *Vanaf 1 januari lagere energierekening door verruimd prijsplafond*. Nieuwsbericht | Rijksoverheid.nl. https://www.rijksoverheid.nl/actueel/nieuws/2022/10/04/vanaf-1-januari-lagere-energierekening-door-verruimd-prijsplafond

nl.zhujiworld.com. (n.d.). Top 100 grootste steden in Nederland 2023. Copyright, nl.zhujiworld.com. All Rights Reserved. https://nl.zhujiworld.com/nl/largest-cities/

Openbare data energielabels - EP-Online. (n.d.). https://www.ep-online.nl/PublicData

Over funda. (n.d.). Funda. https://www.funda.nl/over-funda/

Su, X., Yan, X., & Tsai, C. (2012). Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(3), 275–294. https://doi.org/10.1002/wics.1198

Tan, S. H., & Tan, S. H. (2010). The correct interpretation of confidence intervals. *Proceedings of Singapore Healthcare*, *19*(3), 276–278. https://doi.org/10.1177/201010581001900316

Transitioning to sustainability together. (n.d.). ABN AMRO Bank. https://www.abnamro.com/en/about-abn-amro/landing-page/supporting-our-clients-transition-to-sustainability

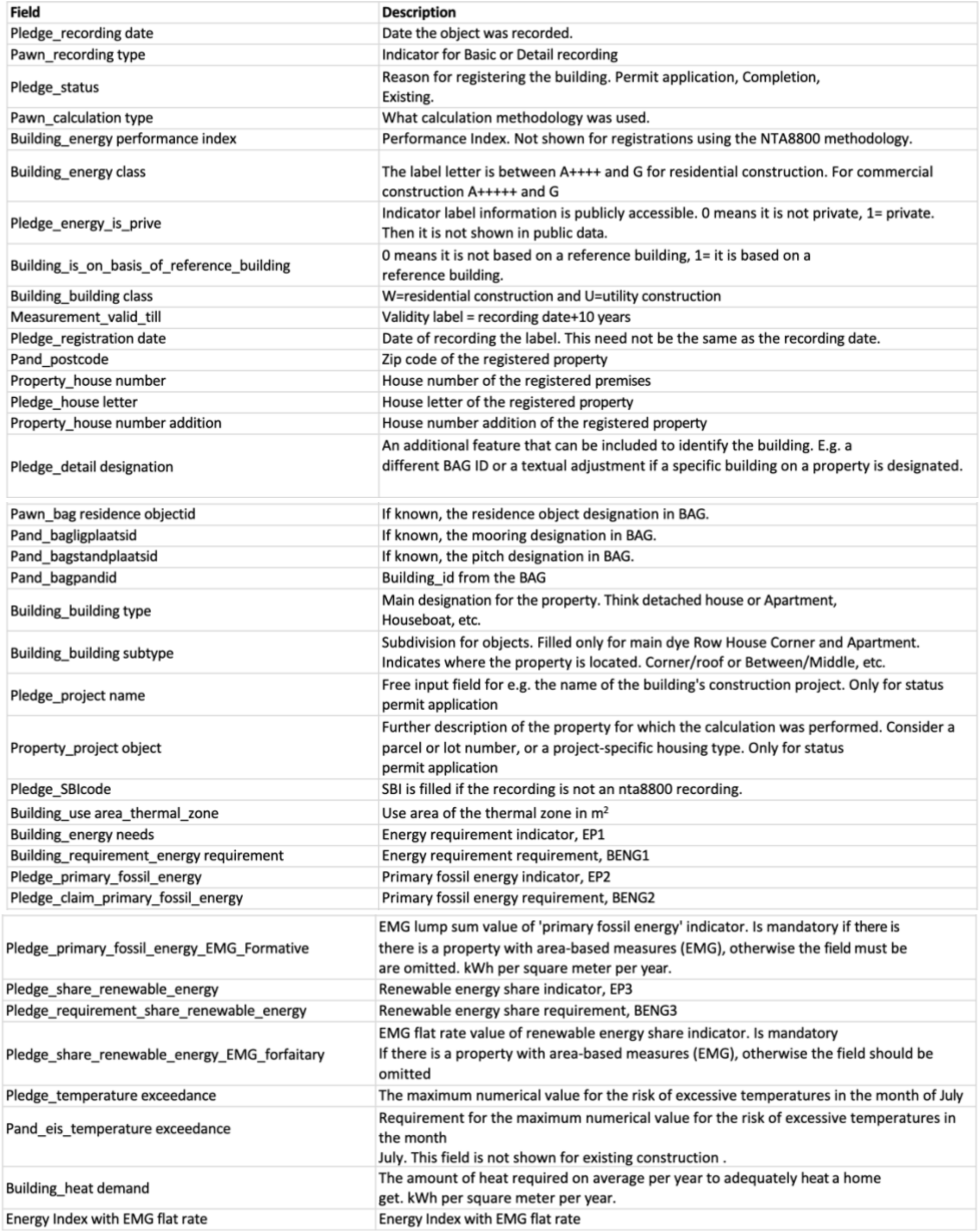
T. Werpy and G. Petersen, “Top value added chemicals from biomass: Volume i – results of screening for potential candidates from sugars and synthesis gas,” Aug. 2004. DOI: 10.2172/15008859.

Verandering per 1 januari 2021: Nieuwe rekenmethode bepaalt hoe energiezuinig jouw woning is. (2021, August 1). L’escaut. https://www.lescaut.nl/over-lescaut/actueel/nieuws/verandering-per-1-januari-2021-nieuwe-rekenmethode-bepaalt-hoe-energiezuinig-jouw-woning-is/

7. Appendices

7.1 Appendix 1 EP-data 2023 explanation

These appendices provide an explanation of the fields from the EP 2023 dataset.



Field	Description
Pledge_recording date	Date the object was recorded.
Pawn_recording type	Indicator for Basic or Detail recording
Pledge_status	Reason for registering the building. Permit application, Completion, Existing.
Pawn_calculation type	What calculation methodology was used.
Building_energy performance index	Performance Index. Not shown for registrations using the NTA8800 methodology.
Building_energy class	The label letter is between A++++ and G for residential construction. For commercial construction A+++++ and G
Pledge_energy_is_prive	Indicator label information is publicly accessible. 0 means it is not private, 1= private. Then it is not shown in public data.
Building_is_on_basis_of_reference_building	0 means it is not based on a reference building, 1= it is based on a reference building.
Building_building class	W=residential construction and U=utility construction
Measurement_valid_till	Validity label = recording date+10 years
Pledge_registration date	Date of recording the label. This need not be the same as the recording date.
Pand_postcode	Zip code of the registered property
Property_house number	House number of the registered premises
Pledge_house letter	House letter of the registered property
Property_house number addition	House number addition of the registered property
Pledge_detail designation	An additional feature that can be included to identify the building. E.g. a different BAG ID or a textual adjustment if a specific building on a property is designated.
Pawn_bag residence objectid	If known, the residence object designation in BAG.
Pand_bagligplaatsid	If known, the mooring designation in BAG.
Pand_bagstandplaatsid	If known, the pitch designation in BAG.
Pand_bagpandid	Building_id from the BAG
Building_building type	Main designation for the property. Think detached house or Apartment, Houseboat, etc.
Building_building subtype	Subdivision for objects. Filled only for main dye Row House Corner and Apartment. Indicates where the property is located. Corner/roof or Between/Middle, etc.
Pledge_project name	Free input field for e.g. the name of the building's construction project. Only for status permit application
Property_project object	Further description of the property for which the calculation was performed. Consider a parcel or lot number, or a project-specific housing type. Only for status permit application
Pledge_SBIcode	SBI is filled if the recording is not an nta8800 recording.
Building_use area_thermal_zone	Use area of the thermal zone in m^2
Building_energy needs	Energy requirement indicator, EP1
Building_requirement_energy requirement	Energy requirement requirement, BENG1
Pledge_primary_fossil_energy	Primary fossil energy indicator, EP2
Pledge_claim_primary_fossil_energy	Primary fossil energy requirement, BENG2
Pledge_primary_fossil_energy_EMG_Formative	EMG lump sum value of 'primary fossil energy' indicator. Is mandatory if there is there is a property with area-based measures (EMG), otherwise the field must be are omitted. kWh per square meter per year.
Pledge_share_renewable_energy	Renewable energy share indicator, EP3
Pledge_requirement_share_renewable_energy	Renewable energy share requirement, BENG3
Pledge_share_renewable_energy_EMG_forfaitary	EMG flat rate value of renewable energy share indicator. Is mandatory If there is a property with area-based measures (EMG), otherwise the field should be omitted
Pledge_temperature exceedance	The maximum numerical value for the risk of excessive temperatures in the month of July
Pand_eis_temperature exceedance	Requirement for the maximum numerical value for the risk of excessive temperatures in the month July. This field is not shown for existing construction .
Building_heat demand	The amount of heat required on average per year to adequately heat a home get. kWh per square meter per year.
Energy Index with EMG flat rate	Energy Index with EMG flat rate

7.2 Appendix 2 City list

This is the list of cities scrapped from Funda (ZhujiWorld, 2023)

- Alkmaar
- Almere
- Amersfoort
- Amsterdam
- Apeldoorn
- Arnhem
- Breda
- Den Bosch
- Den Haag
- Eindhoven
- Enschede
- Gouda
- Groningen
- Haarlem
- Heerenveen
- Maastricht
- Nijmegen
- Rotterdam
- Tilburg
- Utrecht
- Zoetermeer
- Zwolle

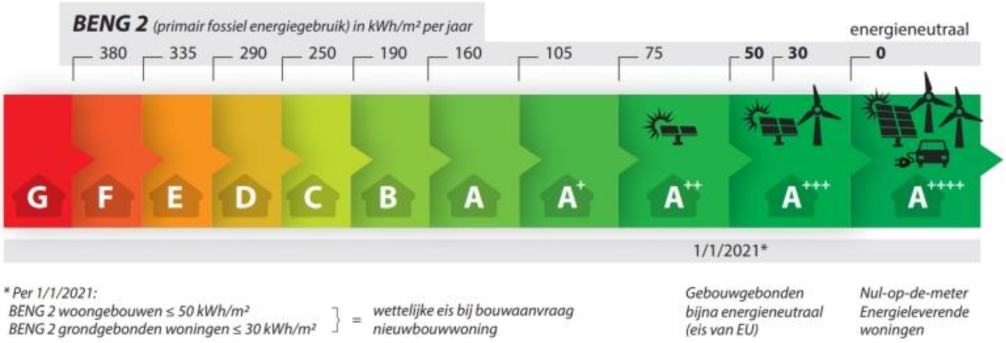
7.3 Appendix 3 Funda data explenation

These appendices provide an explanation of the fields from the FUNDA dataset.

Field	Description
city	The city where de house is in
house_type	Type of home (appartement or huis)
price	Price that we house was listed for
Price_m2	Sale price per square meter
Living_area	Living area of the house in m2
energy_label	The energy label of the home
zip	The Postcode or zip of the house (four numbers two figures)
address	The full address of a house with street, house number, add on
Year_built	The year the house was built in
house_age	How old a house is from date of construction to now
data_list	The data where the house was listed on Funda
term_days	How many days it was online on Funda
data_sold	The day the house was sold on Funda
street	The street where the house is in
number	The house number of the street
suffix	Addons of the house number
extra	Extra is needed to the house number to identify the house

7.4 Appendix 4 Energy labels

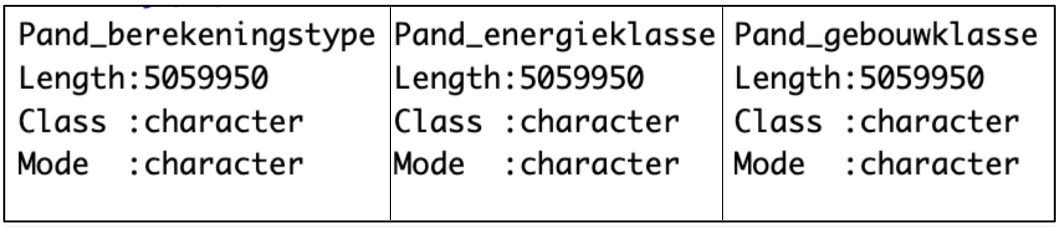
This is the current breakdown between energy labels (Centercon, 2023).



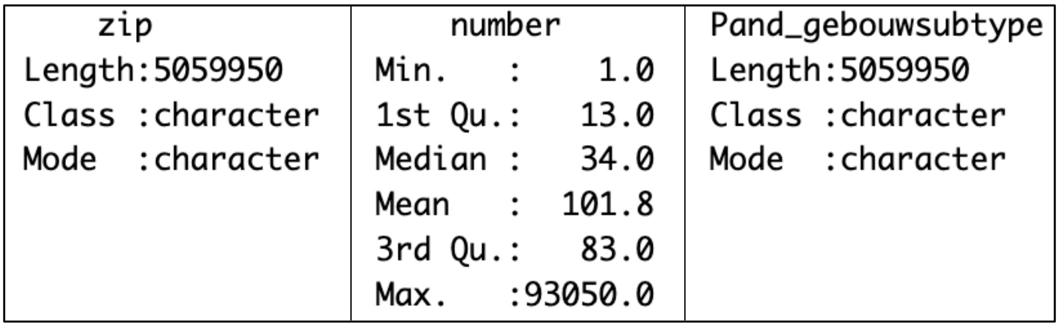
7.5 Appendix 5 Insights EP 2023 Data

Here are some additional insights into the EP 2023 data.

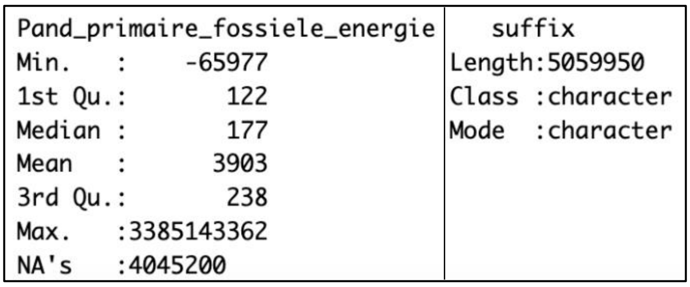
Summary of the important variables:



```
Pand_berekeningstype Pand_energieklasse Pand_gebouwklasse
Length:5059950       Length:5059950     Length:5059950
Class :character     Class :character   Class :character
Mode  :character     Mode  :character   Mode  :character
```

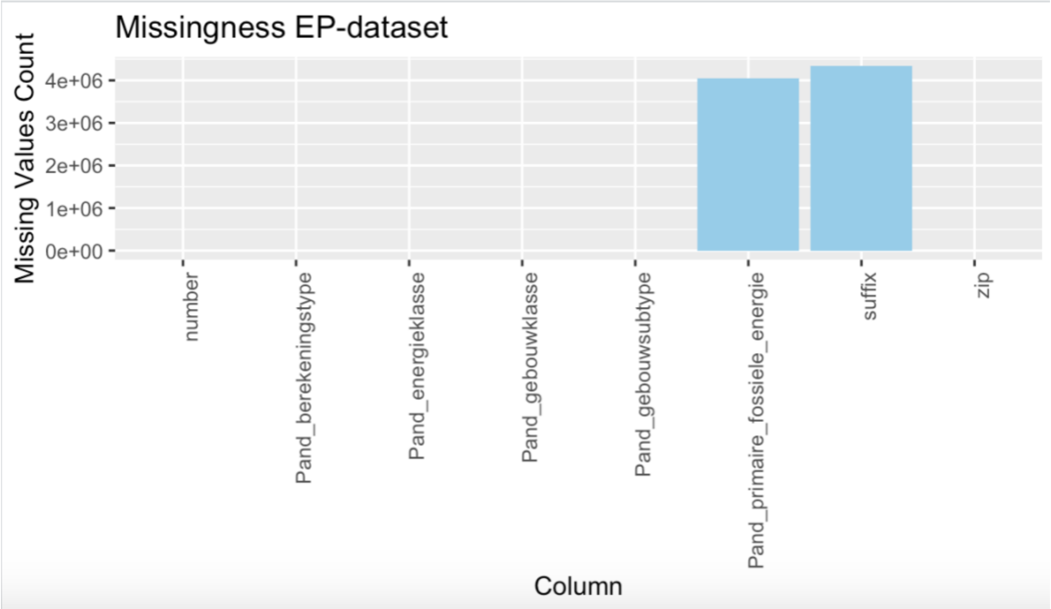


```
     zip               number          Pand_gebouwsubtype
 Length:5059950    Min.   :    1.0   Length:5059950
 Class :character  1st Qu.:   13.0   Class :character
 Mode  :character  Median :   34.0   Mode  :character
                   Mean   :  101.8
                   3rd Qu.:   83.0
                   Max.   :93050.0
```



```
Pand_primaire_fossiele_energie    suffix
Min.   :     -65977             Length:5059950
1st Qu.:        122             Class :character
Median :        177             Mode  :character
Mean   :       3903
3rd Qu.:        238
Max.   :3385143362
NA's   :4045200
```

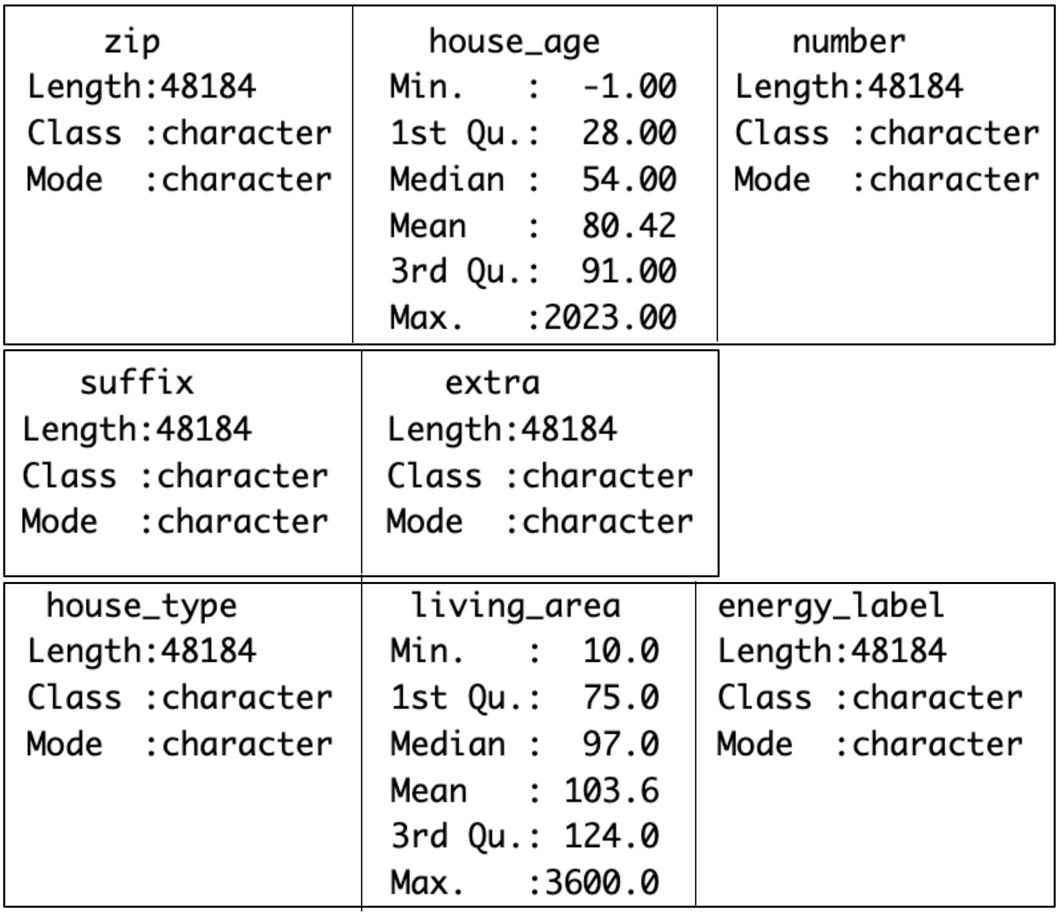
The missingness in de EP dataset:



7.6 Appendix 6 Insights Funda Data

Here are some additional insights into the FUNDA data.

Summary of the important variables:

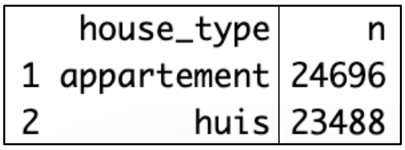


```
     zip                house_age             number
 Length:48184        Min.   :  -1.00     Length:48184
 Class :character    1st Qu.:  28.00     Class :character
 Mode  :character    Median :  54.00     Mode  :character
                     Mean   :  80.42
                     3rd Qu.:  91.00
                     Max.   :2023.00
```

```
    suffix              extra
 Length:48184        Length:48184
 Class :character    Class :character
 Mode  :character    Mode  :character
```

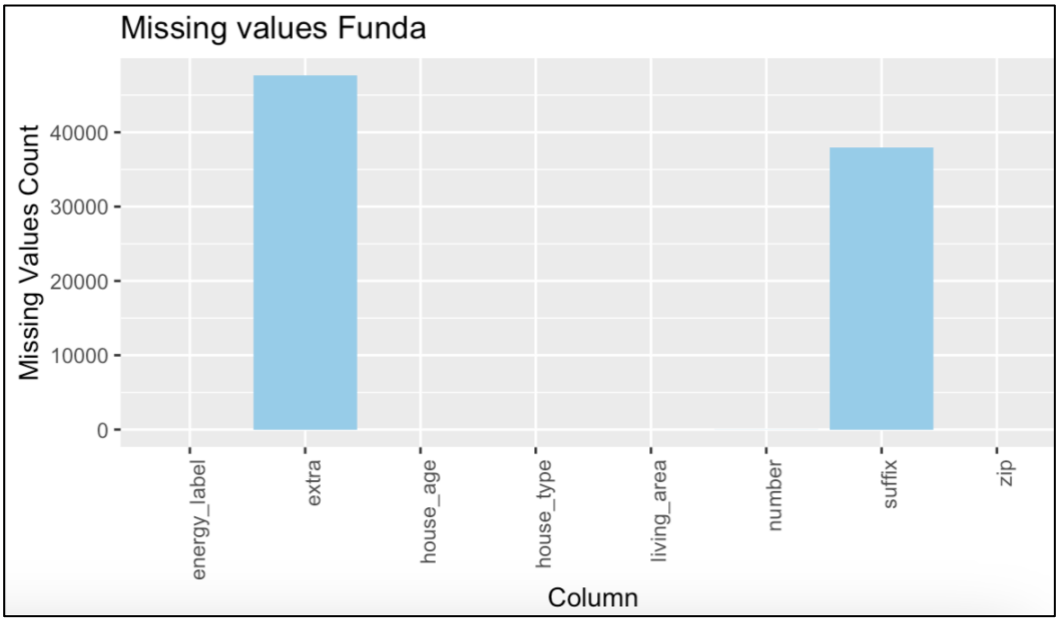
```
  house_type          living_area        energy_label
 Length:48184        Min.   :  10.0     Length:48184
 Class :character    1st Qu.:  75.0     Class :character
 Mode  :character    Median :  97.0     Mode  :character
                     Mean   : 103.6
                     3rd Qu.: 124.0
                     Max.   :3600.0
```

Distribution of the house type:

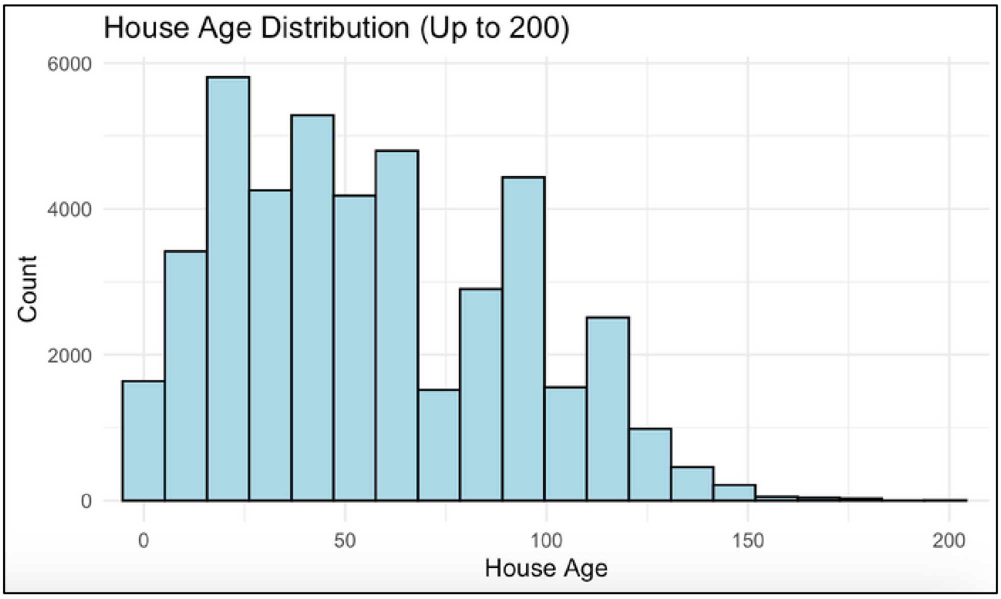


```
    house_type     n
1 appartement  24696
2        huis  23488
```

Missingness in the Funda dataset:



The house age distribution:



7.7 Appendix 7 merged data explanation

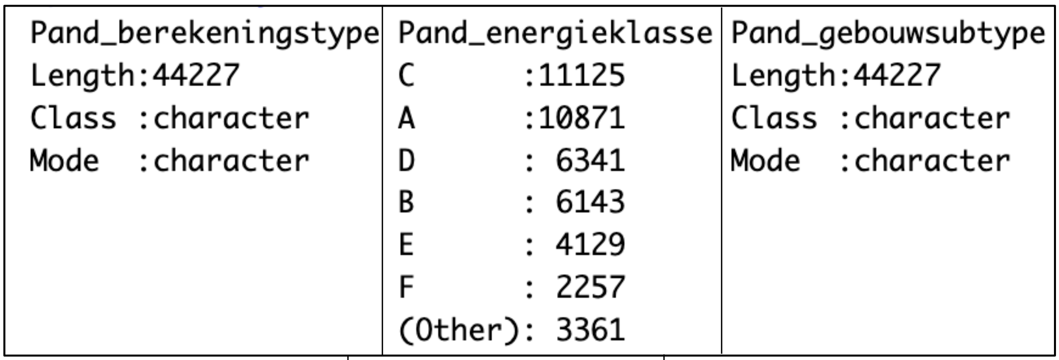
These appendices provide an explanation of the fields from the MERGED dataset.

Field	**Description**
Pand_berekeningstype	The calculation type for the energie label
Pand_energieklasse	Energy class of the house
Pand_gebouwsubstype	Further description of the building type of the house
House_type	House type (apartement or huis)
Living_area	How many square meter the residention building has
House_age	How old the house is from building year till now
Pand_primaire_fossiele_energie	How much electric energy consumption a house has per KWh, per square meter, per year

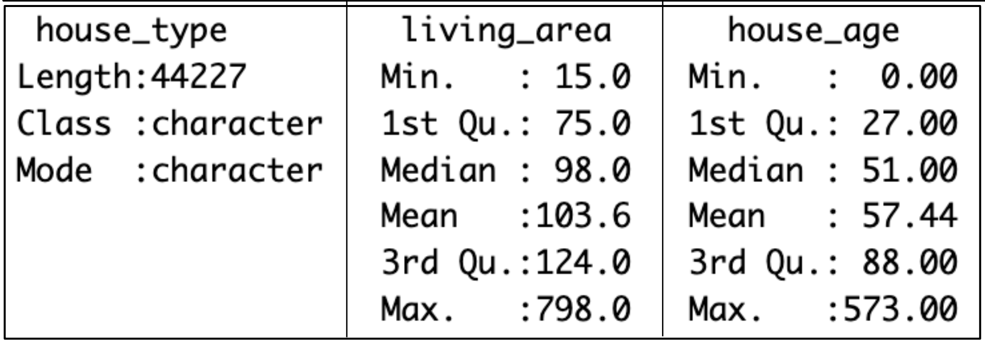
7.8 Appendix 8 Insights Merged data

Here are some additional insights into the MERGED data.

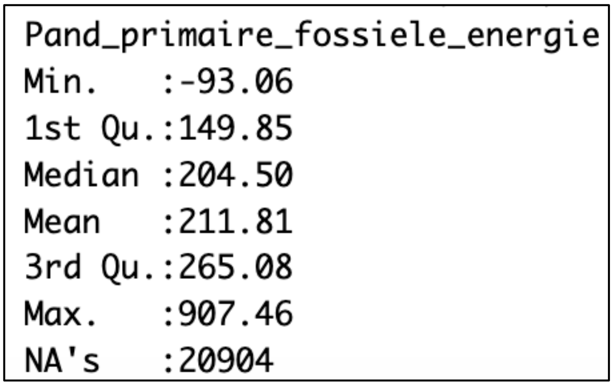
Summary of the variables:



```
 Pand_berekeningstype Pand_energieklasse Pand_gebouwsubtype
 Length:44227         C       :11125     Length:44227
 Class :character     A       :10871     Class :character
 Mode  :character     D       : 6341     Mode  :character
                      B       : 6143
                      E       : 4129
                      F       : 2257
                      (Other): 3361
```

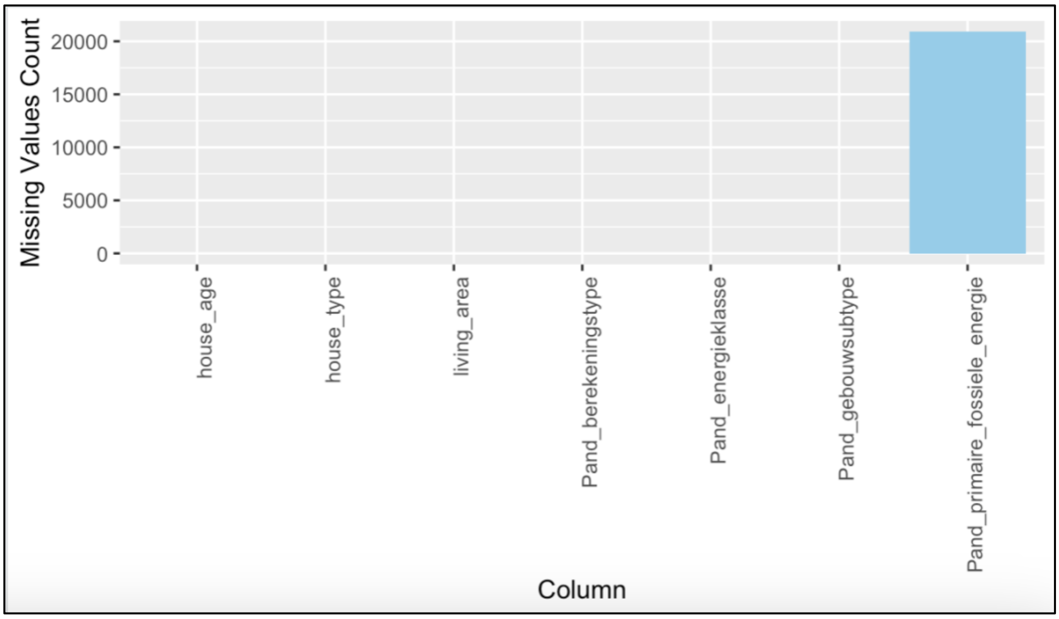


```
  house_type          living_area        house_age
 Length:44227       Min.   : 15.0    Min.   :  0.00
 Class :character   1st Qu.: 75.0    1st Qu.: 27.00
 Mode  :character   Median : 98.0    Median : 51.00
                    Mean   :103.6    Mean   : 57.44
                    3rd Qu.:124.0    3rd Qu.: 88.00
                    Max.   :798.0    Max.   :573.00
```

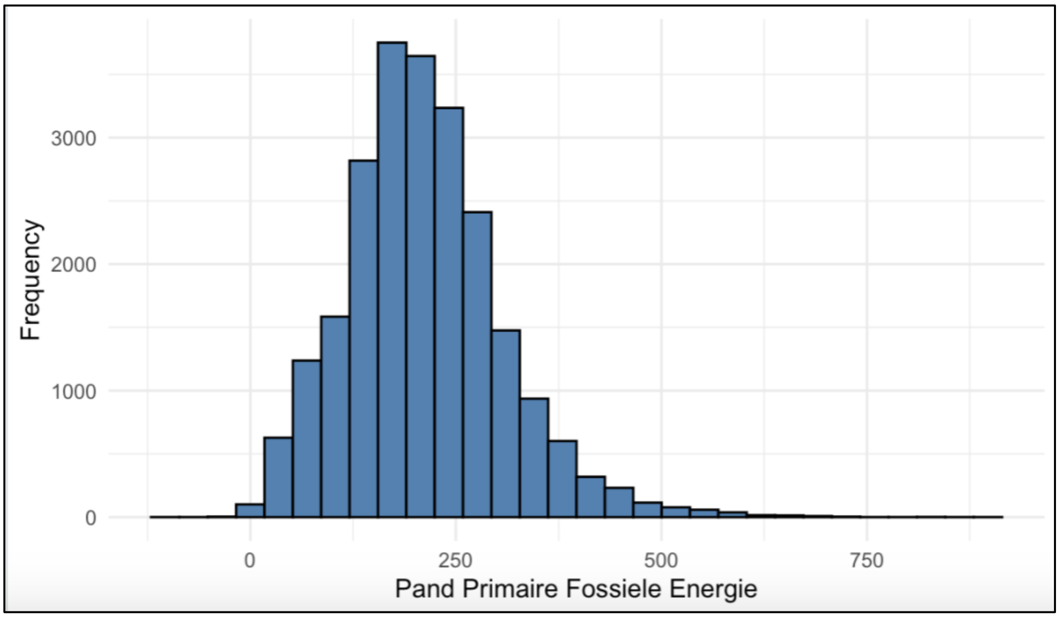


```
 Pand_primaire_fossiele_energie
 Min.   :-93.06
 1st Qu.:149.85
 Median :204.50
 Mean   :211.81
 3rd Qu.:265.08
 Max.   :907.46
 NA's   :20904
```

Missingness count in the merged dataset:



Distribution of primary fossil energy and how often it is in the dataset:



7.9 Appendix 9 GitHub Code

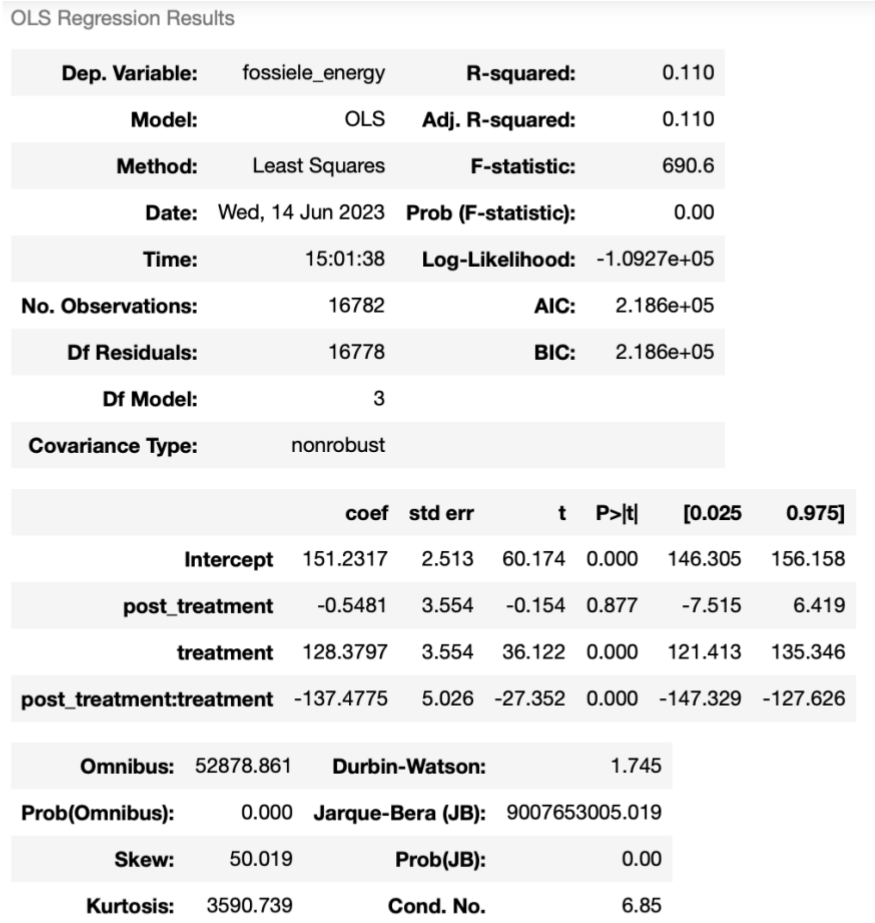
Here the Github link to get to the code and datasets this study is based on. Within this folder it contains the following components.

- All Datasets that are scaped from funda
- The funda dataset combined.
- The EP 2023 dataset
- The EP 2022 dataset
- The Merged dataset
- The python code for the difference in difference analyses
- The R-code for the cleaning and machine learning techniques
- Read me file where it is explained again what the folder is containing

Link: https://github.com/markdiel/Thesis_Mark.git

7.10 Appendix 10 Difference in Difference result

Below are the full results of the difference in difference analysis.



OLS Regression Results

Dep. Variable:	fossiele_energy	**R-squared:**	0.110
Model:	OLS	**Adj. R-squared:**	0.110
Method:	Least Squares	**F-statistic:**	690.6
Date:	Wed, 14 Jun 2023	**Prob (F-statistic):**	0.00
Time:	15:01:38	**Log-Likelihood:**	-1.0927e+05
No. Observations:	16782	**AIC:**	2.186e+05
Df Residuals:	16778	**BIC:**	2.186e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	151.2317	2.513	60.174	0.000	146.305	156.158
post_treatment	-0.5481	3.554	-0.154	0.877	-7.515	6.419
treatment	128.3797	3.554	36.122	0.000	121.413	135.346
post_treatment:treatment	-137.4775	5.026	-27.352	0.000	-147.329	-127.626

Omnibus:	52878.861	**Durbin-Watson:**	1.745
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	9007653005.019
Skew:	50.019	**Prob(JB):**	0.00
Kurtosis:	3590.739	**Cond. No.**	6.85