Report for DAIS bonus project

by Nicolas Webersinke and Kevin Kohler

3.2 Get to know the data

Task 3.2.1

group	initially allocated	with consumption data
Residential	4,225	3,639
SME	485	427
Other	1,735	570
Total	6,445	4,636

Table 1: This table shows the number of residential households, small and medium enterprises (SME) and other entities initially for the project allocated and with electricity consumption data.

The entire following analysis is based only on participants with reported electricity consumption data as the project's objective is the performance assessment of smart meters and their impact on energy consumption.

Task 3.2.2

group	space or water electric heating systems		
	yes	no	n/a
Residential	48.31	33.94	17.75
SME	40.05	27.87	32.08
Other	20.53	13.51	65.96

Table 2: This table displays the fraction of residential households, SMEs and ‘others’ having installed space or water electric heating systems. All values are displayed in percent. Space and water electric heating systems of any kind were considered, since they need lots of electricity and were not separable for SMEs (see Q4551).

Task 3.2.3

According to the residential pre survey all people under 15 years are children (see Q410, Q43111). Thus, 22.53% of residential households have children.

Task 3.2.4

	mean	median	variance	std. dev.
Residential	3.14	3	1.70	1.31

Table 3: This table reports summary statistics of the number of people living in a residential household. Includes adults and children (see Q420 and Q43111).

Table 3 shows that on average, 3.14 people (adults and children) live in a residential household.

Task 3.2.5

While there are standard definitions of company size (e.g. by the European Commission), the companies that participated in the project appear to be very small in terms of annual turnover. Therefore, we have included as small companies those that generate an annual turnover of less than or equal to 2 million euros and as medium companies those that generate an annual turnover of more than or equal to 3 million euros. The upper limit of turnover for medium companies is set by the sample, since only small and medium enterprises were included as project participants.

size by annual turnover	number of companies
small	97
medium	11
refused or n/a	319
Total	427

Table 4: This table describes the categorization of participating enterprises by size, based on annual turnover in euros. Noteworthy, a large proportion of enterprises surveyed did not want to disclose their annual turnover (see Q6201).

Task 3.2.6

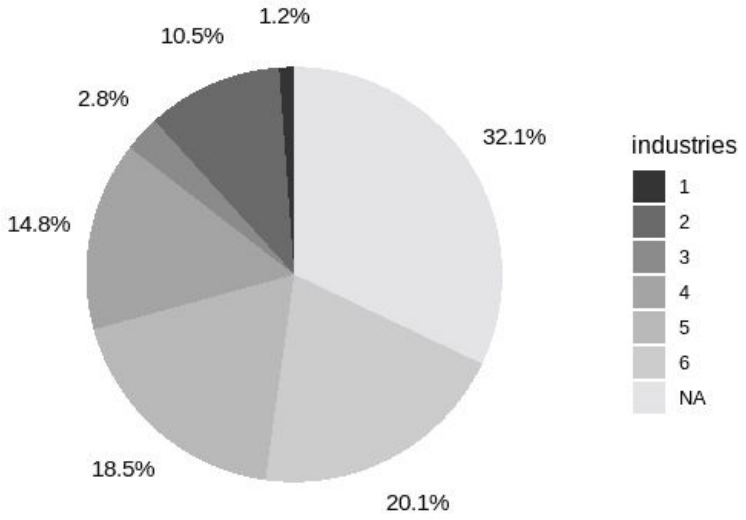

Figure 1: This figure shows the industry distribution of participating SMEs. Industry classification following survey question Q611 (1: Agriculture, forestry and fishing; 2: Industry Mining and quarrying, Manufacturing Electricity, gas, steam and air conditioning supply, Water supply, sewerage, waste management and remediation activities; 3: Construction; 4: Wholesale and retail trade, repair of motor vehicles and motorcycles; 5: Business and Professional Services, Information and communication, Financial and insurance activities, Real estate activities, Professional, scientific and technical activities, Administrative and support service activities, Public administration and defence, compulsory social security, Education Human health and social work activities; 6: Other Transportation and storage, Accommodation and food service activities, Leisure hotels Arts, entertainment and recreation, Other service activities, Activities of households as employers, undifferentiated goods- and services producing activities of households for own use, Activities of extraterritorial organisations and bodies).

Task 3.2.7

Satisfaction indicators	mean	median	std. dev.
Quality of electricity supply	1.68	1	0.90
Number of suppliers competing	2.86	3	1.19
Range of different tariffs available	2.90	3	1.10
Percentage of electricity generated from renewables	3.25	3	1.16
Number of estimated bills your organisation receives	2.64	2	1.25
Level of wastage of electricity in the home or office	3.14	3	1.19
Overall cost of electricity	3.69	4	1.16
Environmental damage assoc. with electricity used	3.49	4	1.03
Total	2.96	3	1.27

Table 5: This table shows summary statistics of various satisfaction metrics derived from survey question Q5512. Summary statistics of total satisfaction were calculated using pooled survey data. Answers are in the range from 1 = very satisfied to 6 = very dissatisfied.

We calculated summary statistics for each electricity market satisfaction question. To get an overall impression of satisfaction, we pooled the responses and calculated summary statistics again. All answers are in range from 1 = very satisfied to 6 = very dissatisfied. A mean of about 3 and a median of 3 can be interpreted as medium satisfied, according to the scale. Enterprises seem to be most satisfied with the quality of electricity supply, while they are more dissatisfied about the overall cost of

electricity. For the remaining indicators, there is a medium level of satisfaction with a fairly low standard deviation.

3.3 Investigate the data

Task 3.3.1

group	mean	variance	std. dev.
Residential	8,711.60	19,007,356	4359.74
SME	36,547.14	1,331,332,711	36,487.43
Other	13,092.48	676,390,893	26,007.52

Table 6: This table shows summary statistics of annual electricity consumption for each group. Electricity consumption is measured in kWh.

Task 3.3.2

household size	mean	variance	std. dev.
small	8,386.03	11,734,607	3,425.58
medium	11,561.13	17,595,356	4,194.68
large	13,822.85	22,297,317	4,722.00

Table 7: This table reports summary statistics of residential households' annual electricity consumption. Household sizes are categorized into small (1-3 people), medium (4-5 people) and large (> 6 people). Electricity consumption is measured in kWh.

The mean annual electricity consumption increases with household size, which is in line with the naive expectation. Interestingly though, the standard deviation also increases with household size. This implies that large households differ more in their annual electricity consumption than smaller households, which can cause problems in later regression models in the form of heteroskedasticity.

Task 3.3.3

(a) Average monthly electricity consumption of 100 residential households

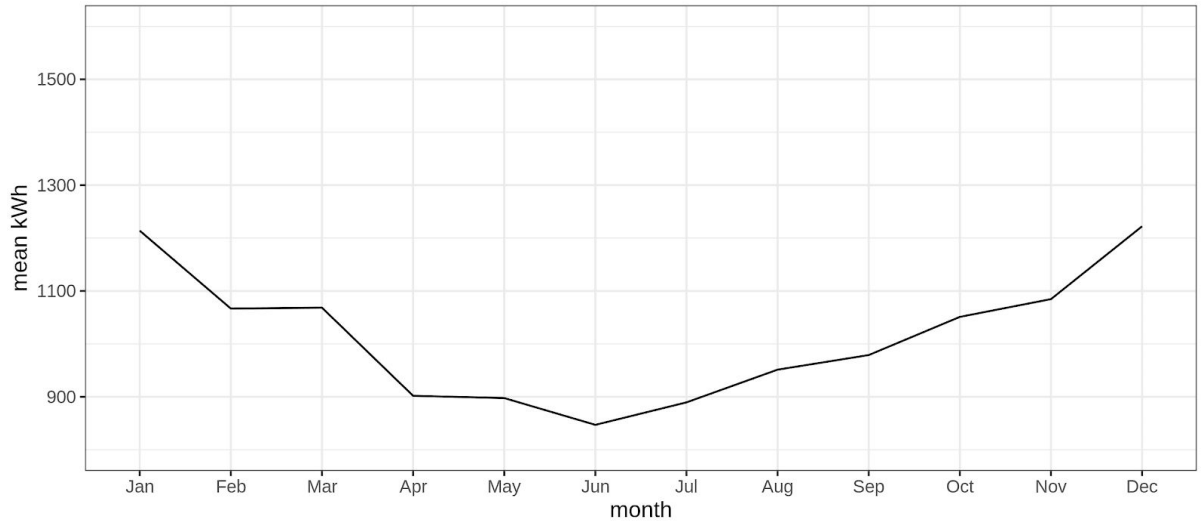

Figure 2: This figure illustrates the average monthly energy consumption of 100 randomly selected residential households over the course of one year. Electricity consumption is measured in kWh/month.

We can see a sharp decline in consumption during the spring months and a strong increase in electricity consumption during the autumn months. The reason for this is probably the beginning and end of the heating period. Moreover, people might stay at home longer during the colder and darker months of the year.

(b) Average monthly electricity consumption of 100 SMEs

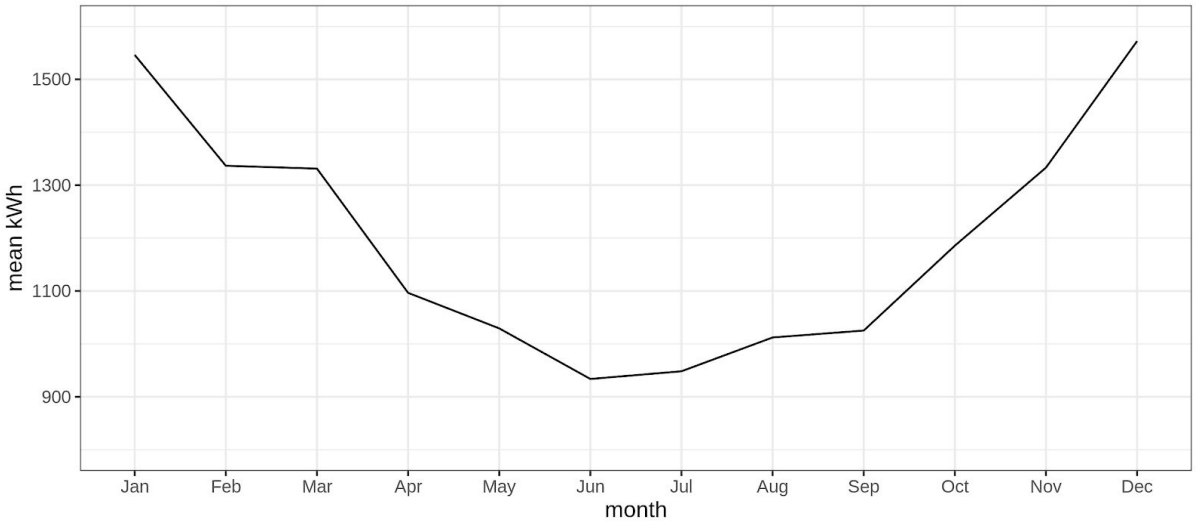

Figure 3: This figure shows the average monthly electricity consumption of 100 randomly selected SMEs over the course of one year. Electricity consumption is measured in kWh/month.

We can see a sharp decline in consumption during the spring months and a strong increase in energy consumption during the autumn months. The reason for this is probably the beginning and end of the heating period. Interestingly, the drop in consumption in the spring is much more significant for enterprises than for residential households. Reasons for this could be higher heating needs or lighting systems with higher consumption compared to residential households.

(c) Average daily load profile of 100 residential households and 100 SMEs

i. one week in the summer

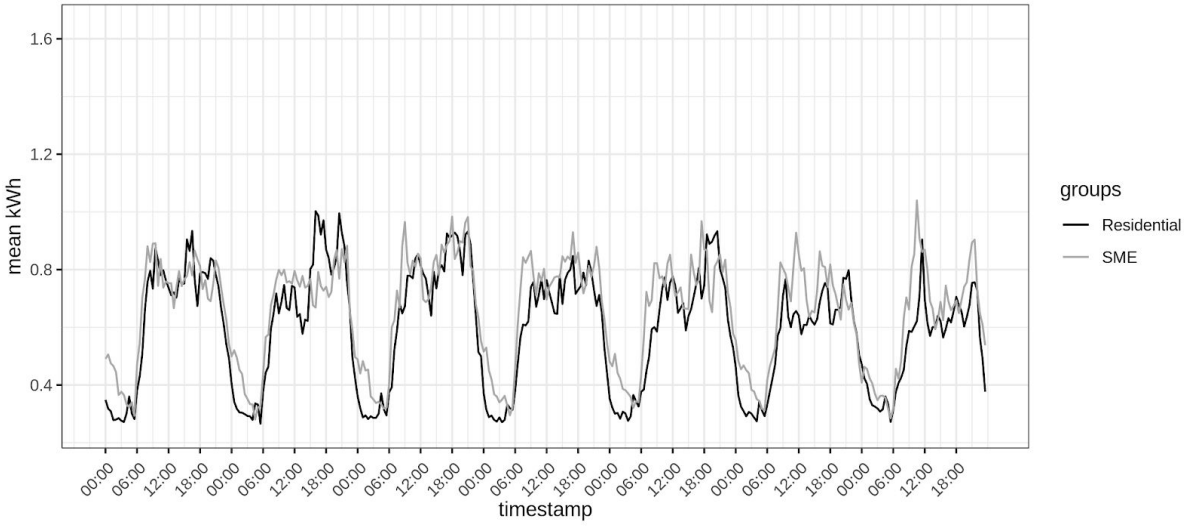

Figure 4: This figure shows average daily load profiles of 100 randomly selected residential households (black) and 100 randomly selected SMEs (grey) over the course of one week in the summer. Load profile is measured intraday in kWh per 30 minutes.

From the above load profile depiction, it can be inferred that SMEs have on average a slightly higher consumption than residential households over the entire period. The SME load profile is also less volatile compared to the profile of residential households. These observations could be explained by typical patterns of residential households sleeping during nighttime and leaving for work or education from morning to afternoon, while SMEs show on average a higher base load due to permanently switched on machines, for instance. The weekend electricity consumption of residential households is on average slightly lower than during the week. This does not seem to be the case for enterprises.

ii. one week in the winter.

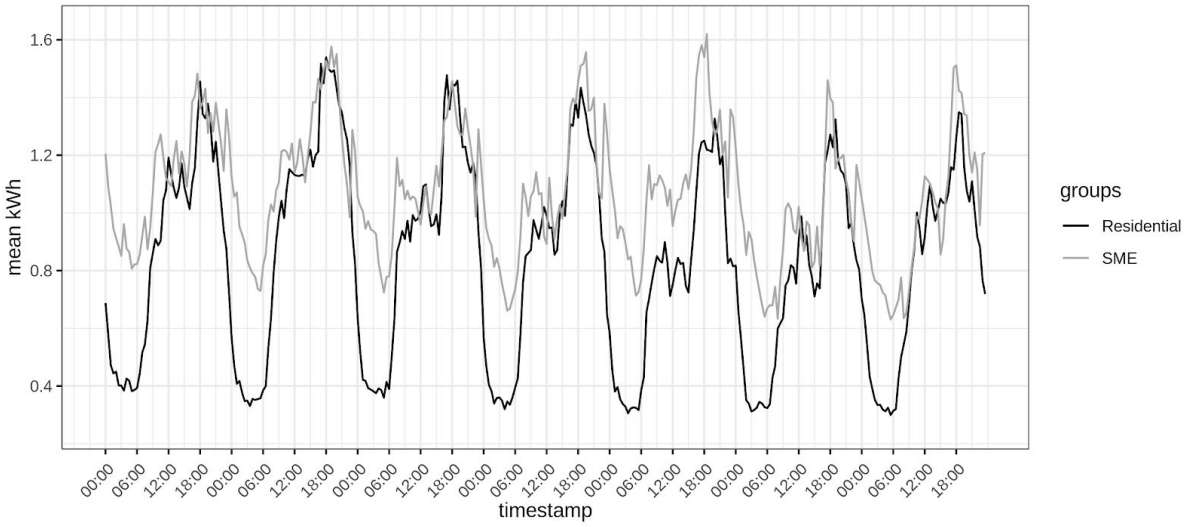

Figure 5: This figure illustrates average daily load profiles of 100 randomly selected private households (black) and 100 randomly selected SMEs (grey) over the course of one week in the winter. Load profile is measured intraday in kWh per 30 minutes.

From the above load profile illustration, it can be derived that SMEs have a considerably higher load profile than residential households over the entire period (especially during nighttime). The intraday patterns are similar to those of the week in the summer. However, peaks appear more extreme. Again, this could be due to the heating period, with greater nighttime reductions for households than for enterprises. As could be expected, total load profiles for both groups are generally higher in winter than in summer due to heating or less daylight.

iii. Three days of choice

Tuesday in April

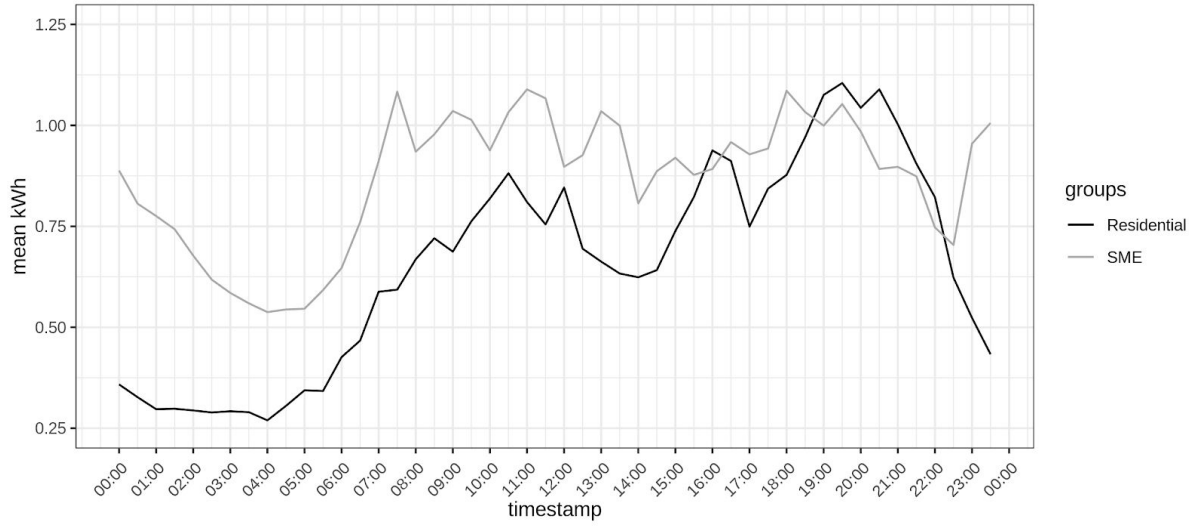

Figure 6: This figure illustrates the average daily load profiles of 100 randomly selected residential households (black) and 100 randomly selected SMEs (grey) over the course of one weekday (Tuesday in April). Load profile is measured intraday in kWh per 30 minutes.

Thursday in April

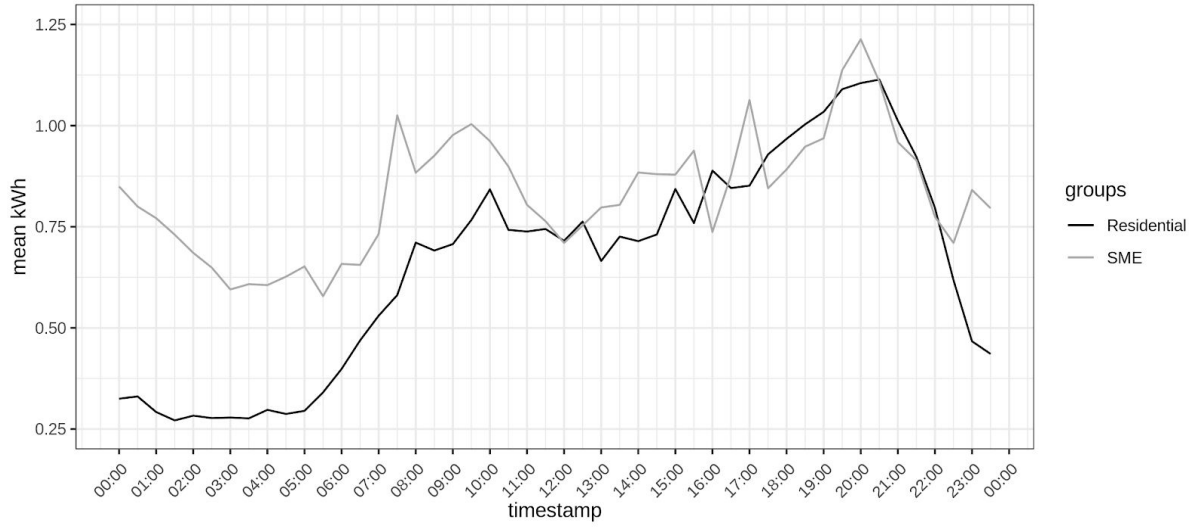

Figure 7: This figure illustrates the average daily load profiles of 100 randomly selected residential households (black) and 100 randomly selected SMEs (grey) over the course of one weekday (Thursday in April). Load profile is measured intraday in kWh per 30 minutes.

Sunday in April

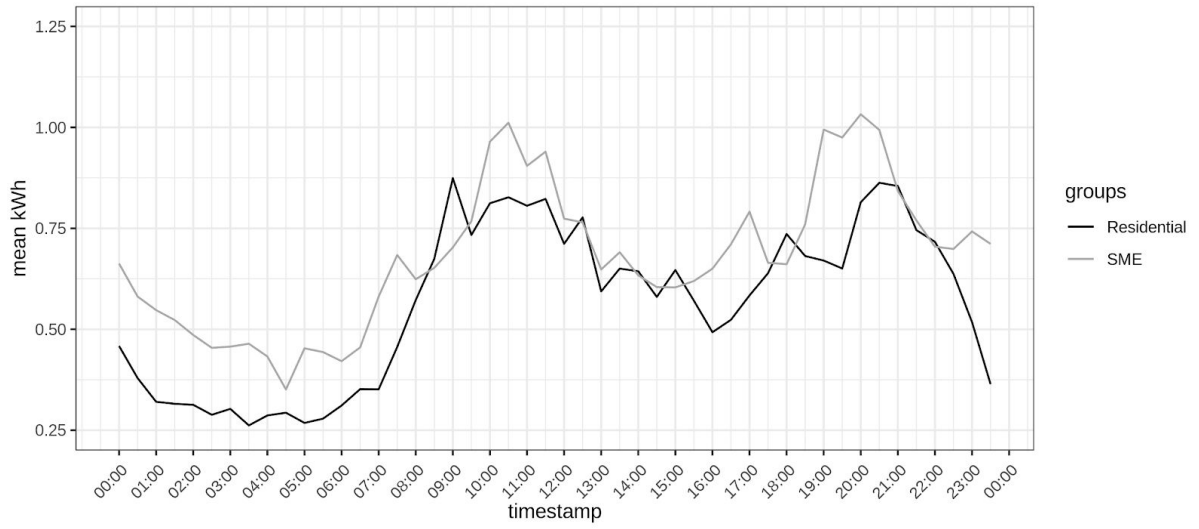

Figure 8: This figure illustrates the average daily load profiles of 100 randomly selected residential households (black) and 100 randomly selected SMEs (grey) over the course of one weekend day (Sunday in April). Load profile is measured intraday in kWh per 30 minutes.

Derivable from figures 6, 7 and 8 are following observations: first, both the groups (residential households and SMEs) have the highest loads during weekdays. Both groups across all days observed have their lowest loads during the night hours from from 1 a.m. to 6 a.m. This is not very surprising given normal personal and business schedules. However, the SME drop in load is not as significant as the residential one, which might be due to permanently switched on machines, for instance. The highest loads experienced in both the groups occur from morning hours (around 7 a.m.) until evening (circa. 8 p.m.) on all days observed. Interestingly, on weekdays, the level of load through daytime is relatively constant for SMEs, while residential households experience more of a gradual increase with a spike at around 8 p.m. (which might potentially be workday-related). When it comes to the weekend, though, both SMEs and residential households share two peaks at around 10 a.m. as well as 8 p.m. In general, and apart from a few short-term observations around the afternoons, SMEs have a generally higher daily load profile compared to residential households. The gap between SMEs and residential load is the largest during the night hours (probably due to machinery running overnight) and smallest during the late afternoon hours (as residential loads are now increasing, e.g. due to laundry).

(d) Average daily load profile for Christmas week

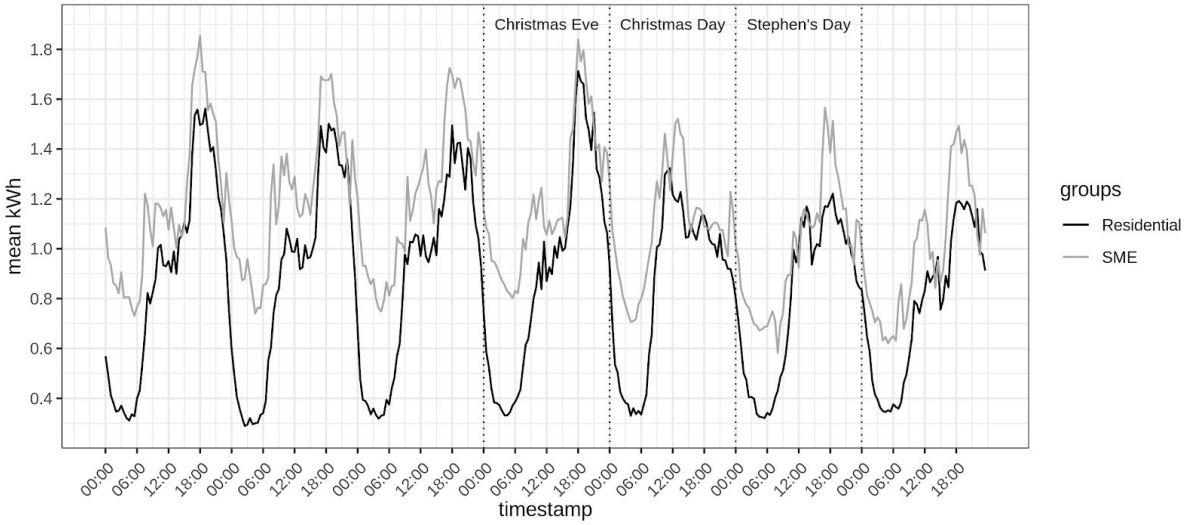

Figure 9: This figure shows the average daily load profiles of 100 randomly selected residential households (black) and 100 randomly selected SMEs (grey) over the course of one weekend day (Sunday in April). Load profile is measured intraday in kWh per 30 minutes.

Over the course of the Christmas week, both SMEs and residential households show the same pattern of having the highest loads during the afternoon hours (approx. 4 p.m. to 6 p.m.) except of Christmas day (25th of December) and their lowest loads at night, shortly between 12 a.m. and 6 a.m. After Christmas Eve (24th of December), both SMEs and residential households have lower loads over day and night (most likely due to public holidays from 25th to 26th of December), with residential households experiencing a stronger decrease in load than SMEs (this might be due to the happening of many family reunions during the holidays, when more people share the same heated space).

3.4 Answer questions with the data

Task 3.4.1

mean / mean difference	Residential	SME	Other
Residential (*N* = 3,639)	8,711.60		
SME (*N* = 427)	27,835.54***	36,547.14	
Other (*N* = 570)	4,380.87***	23,454.67***	13,092.48

Table 8: This table shows in the diagonal means of the annual electricity consumption for each group. The other cells represent mean differences. All values in kWh per annum. Significance levels are calculated based on a Welch two-sample t-test for equal means. ***, ** and * indicate significance at the 1%, 5% and 10% level.

Table 8 demonstrates that there are highly significant differences between the average annual electricity consumptions of residential households, SMEs and others, both in terms of effect size and statistical significance. The difference between residential households and SMEs is the largest, the difference between households and others the smallest.

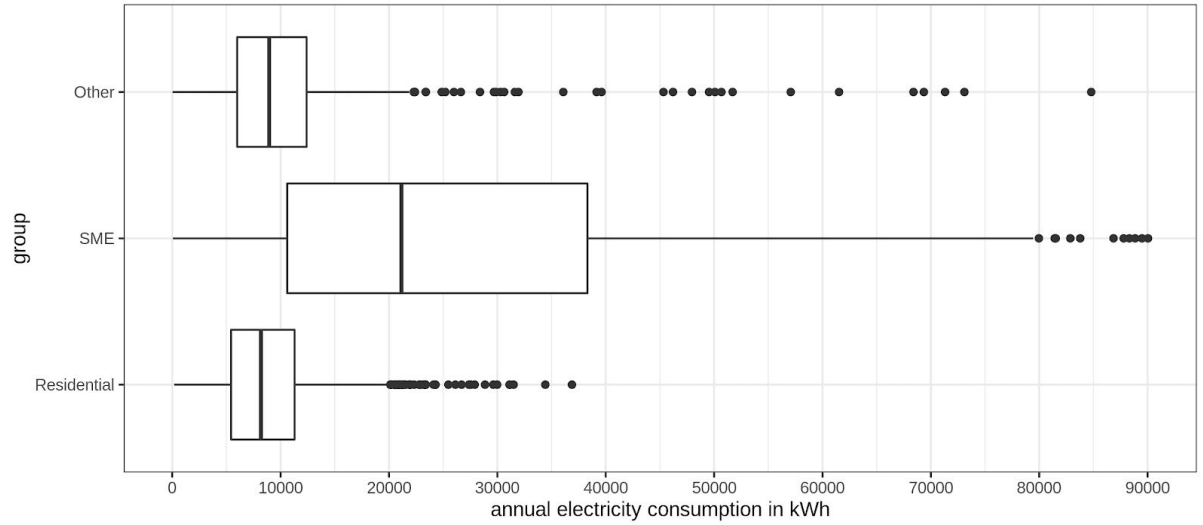

Figure 10: This figure shows a boxplot depiction of the annual electricity consumption in kWh for each group separately.

The boxplot reveals that households are the most homogeneous group. The distribution of SMEs' annual electricity consumption shows a clear right skew. The others show a large heterogeneity with many outliers. However, this is in the nature of a residual group. Most likely, this group contains both households and enterprises that could not be assigned to the actual groups.

Task 3.4.2

independent variables	coefficient	t-value	p-value
	Panel A: Model 1		
intercept	5,155.49***	26.22	<0.01
household_size	1,431.05***	24.75	<0.01
	Panel B: Model 2		
intercept	4,485.09***	15.76	<0.01
household_size	1,730.20***	15.57	<0.01
children	1,023.10	1.46	0.14
household_size * children	-430.52**	-2.24	0.03

Table 9: This table shows regression results for model 1 (Panel A) and model 2 (Panel B). The formula for model 1 is electricity_consumption = a + b * household_size + e and for model 2 electricity_consumption = a + b * household_size + c * children + d * (household_size * children) + e with children being a dummy variable for having at least one child. Following White (1980), we adjust standard errors for heteroskedasticity. ***, ** and * indicate significance at the 1%, 5% and 10% level.

Looking at the coefficient of household_size in table 9 Panel A and the regression line in figure 11, it can be derived that the marginal yearly electricity consumption of an additional household member is ceteris paribus (c. p.) on average 1,431.05 kWh. However, this effect interacts with whether a household has children or not. This can be seen in Panel B. The interaction coefficient is highly significant, both in terms of effect size and statistical significance. We control for heteroskedasticity by adjusting standard errors following White. The interaction term household_size * children can be interpreted in a way that if the additional household member is a child, c. p. on average only 1,730.20 - 430.52 = 1,299.68 kWh electricity more is consumed. On the other hand, the presence of at least one child in a household implies c. p. on average an additional electricity consumption of 1,023.10 kWh compared to households without children. However, this effect is not statistically significant.

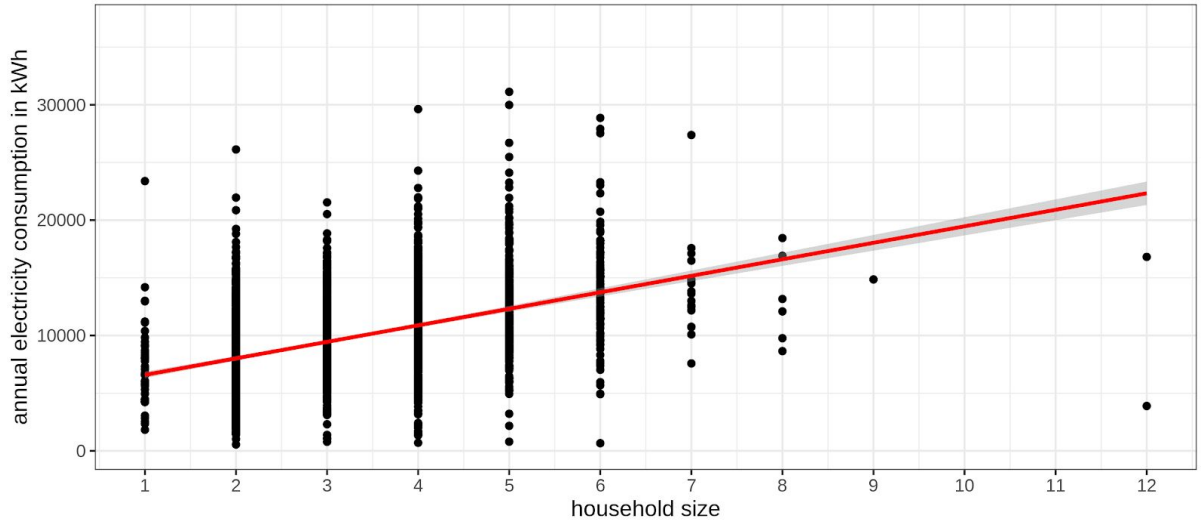

Figure 11: This figure shows regression results for a residential household's average annual electricity consumption, depending on its household size. Annual electricity consumption is measured in kWh.

Task 3.4.3

independent variables	coefficient	t-value	p-value
	Panel A: Model 1		
intercept	7.25***	156.16	<0.01
household_size	0.07***	5.32	<0.01
	Panel B: Model 2		
intercept	7.20***	47.94	<0.01
household_size	0.05***	2.75	<0.01
income_2	0.01	0.08	0.94
income_3	0.13	0.88	0.38
income_4	0.12	0.82	0.41
income_5	0.13	0.91	0.37
	Panel C: Model 3		
intercept	7.49***	75.56	<0.01
employees_medium	0.66***	2.78	<0.01
	Panel D: Model 4		
intercept	7.37***	36.05	<0.01
employees_medium	-0.24	-0.98	0.33
turnover_medium	0.52*	1.90	0.06

Table 10: This table shows regression results for residential households for model 1 (Panel A) and model 2 (Panel B) and for SMEs for model 3 (Panel C) and model 4 (Panel D). The dependent variables floor_area (residential households) and premise_area (SMEs) are log transformed for better model fit and to be more consistent with the assumptions of normal distribution. Formulas on request. Following White (1980), we adjust standard errors for heteroskedasticity. ***, ** and * indicate significance at the 1%, 5% and 10% level.

Floor area and premise area are measured in square foot in this analysis because the conversion to the metric system in Ireland did not occur until 2005 and about 90% of participants therefore reported

area in square foot. We convert the remaining 10% from square meters to square feet by multiplying by the factor 10.764. For all models in table 10, the dependent variable area is log-transformed for better model fit and to be more consistent with the assumption of normal distribution of residuals. Panel A of table 10 shows that household size is a significant predictor of the floor area. On average, the floor area is c. p. 7 % higher for an additional household member. This effect is highly statistically significant. Controlling for income through several income category dummy variables, this effect decreases but remains significant. The income variable has a larger effect on floor area (at least for higher income) than household size in terms of effect size. However, the coefficients of the income dummy variables are not statistically significant.

Performing the same analysis for SMEs with employees and/or turnover as independent dummy variables (employees_medium equals to 1 if the company employs more than 10 people, turnover_medium equals to 1 if the company makes more than 500k euro annual turnover), Panel C of table 10 shows that employees_medium is a significant predictor of premise area. Companies with more than 10 employees have c. p. on average 66 % larger premise areas. Controlling for turnover in Panel D, this effect disappears. Turnover_medium seems to explain the variety of premise areas better. Enterprises making more than 500,000 euros annual turnover c. p. on average have a 52 % larger premise area.

Task 3.4.4

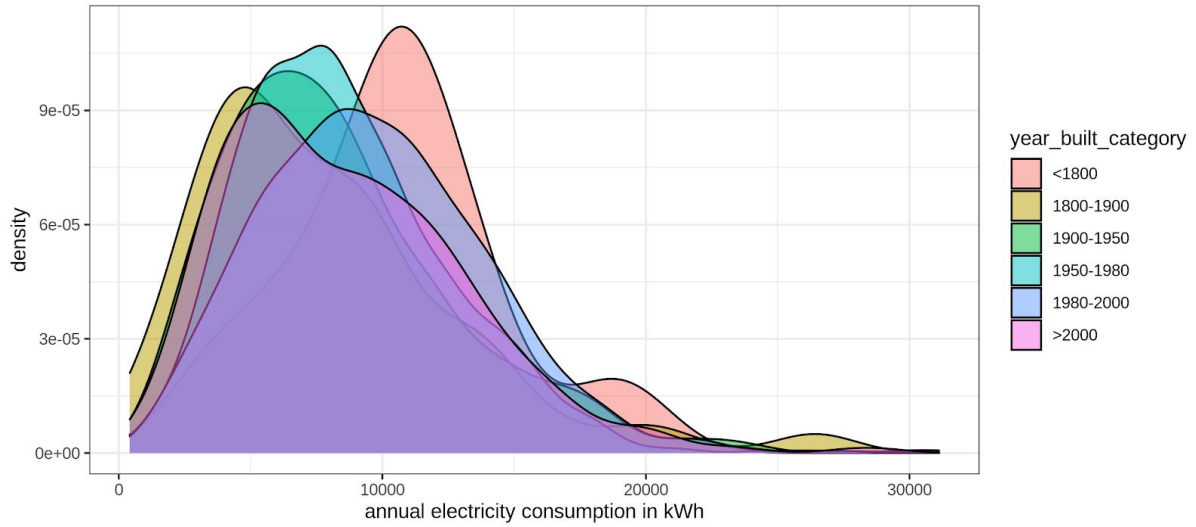

Figure 12: This figure shows the density of the distribution of residential household's average annual electricity consumption for different years of construction of the houses in which the households live. Annual electricity consumption is measured in kWh and the category of the building year ranges from pre-1800 to post-2000 with a total of six categories.

Based on the density plot shown above, the general trend can be derived that the older a house is, the higher its annual electricity consumption is, except for the time frame between 1800 - 1950. This can be explained due to the fact that houses from older time periods were not constructed and equipped as well-insulated and energy-efficient as today's houses. Older buildings may thus suffer from higher consumption due to heating. In addition, it could be that older houses more often use electric heating, since the subsequent installation of gas heating, for instance, is very expensive and complex. Even with most of the single categories showing a pretty decent normal distribution, some high tail values can be detected - this might be due to highly luxurious and spacious elder buildings, as well as faulty constructions of some houses.

Thinking of possible confounding variables, we identified the presence of electric heating systems and the household size. Electric heating systems should influence the electricity consumption directly

compared to the construction year. We expect household size to be even more important, as it is a good proxy for usage of a house or apartment.
Indeed, table 11 shows that household size and electric heating are highly significant. However, if a building was built after 2000, the households living in it consume c. p. on average 1,636.77 kWh less electricity than households living in a building constructed before 1800.

independent variables	coefficient	t-value	p-value
intercept	5,721.51***	7.04	<0.01
household_size	1,466.95***	20.52	<0.01
electric_heating	972.86***	6.28	<0.01
year_built_1800_1900	-721.04	-0.73	0.46
year_built_1900_1950	-1,318.53	-1.63	0.10
year_built_1950_1980	-1,173.73	-1.49	0.14
year_built_1980_2000	-1,060.16	-1.34	0.18
year_built_2000	-1,636.77**	-2.03	0.04

Table 11: This table shows regression results for model 1 (Panel A) and model 2 (Panel B). The formula for model 1 is electricity_consumption = a + b * household_size + e and for model 2 electricity_consumption = a + b * household_size + c * children + d * (household_size * children) + e. Following White (1980), we adjust standard errors for heteroskedasticity. ***, ** and * indicate significance at the 1%, 5% and 10% level.

Task 3.4.5

Figure 13: This figure describes the residential households' annual electricity consumption distribution depending on the home category. Annual electricity consumption is measured in kWh, different homes are categorized into three groups: Apartments, bungalows and houses.

From figure 13 and table 12, a number of observations and interpretations can be observed: Apartments have the lowest average electricity consumption across all groups, both in terms of mean and median and the households living in an apartment are also relatively equal in their consumption patterns (showing low variance). This can be explained by a number of characteristics applying for apartments, namely their low average household size and floor area. Only a relatively small number of apartments (probably more spacious and luxurious ones) show higher consumptions. The other two groups' electricity consumptions are relatively similarly distributed. Bungalows show the highest mean annual electricity consumption, houses the second highest. Their mean difference is marginal, but still significant. The distribution of annual electricity consumption for houses has more outliers compared

to the distribution for bungalows, which is probably due to the fact that there are several types of houses, which we have grouped together. All mean differences in table 12 are statistically significant.

mean / mean difference	Apartment	Bungalow	House
Apartment (*N* = 57)	4,942.06		
Bungalow (*N* = 812)	4,208.10***	9,150.16	
House (*N* = 2,312)	3,807.90***	400.20**	8,749.96

Table 12: This table shows in the diagonal means of the annual electricity consumption for each home category. The other cells represent mean differences. All values in kWh per annum. Significance levels are calculated based on a Welch two-sample t-test for equal means. ***, ** and * indicate significance at the 1%, 5% and 10% level.

Task 3.4.6

independent variables	coefficient	t-value	p-value
	Panel A: Model 1		
intercept	1,646.47***	5.44	<0.01
bedrooms	2,058.21***	22.88	<0.01
R^2	0.1662		
Adj. R^2	0.1659		
	Panel B: Model 2		
intercept	1,085.07***	3.67	<0.01
bedrooms	1,709.19***	18.77	<0.01
electric_heating	582.09***	4.17	<0.01
tumble_dryer	2,096.77***	14.53	<0.01
R^2	0.2201		
Adj. R^2	0.2193		

Table 13: This table shows regression results for model 1 (Panel A) and model 2 (Panel B). The formula for model 1 is electricity_consumption = a + b * bedrooms + e and for model 2 electricity_consumption = a + b * bedrooms + c * electric_heating + d * tumble_dryer + e. Following White (1980), we adjust standard errors for heteroskedasticity. ***, ** and * indicate significance at the 1%, 5% and 10% level.

Panel A of table 13 shows that the number of bedrooms has a positive and significant correlation with the annual average electricity consumption. In panel B the dummy variables electric_heating and tumble_dryer were added to increase the proportion of the variance in the dependent variable that is predictable from the independent variables. Electric_heating is set to 1 if there is at least one electric heating system in the household, both water and space. Tumble_dryer is set to 1 if there is at least one tumble dryer in the household. The two new variables in model 2 increase both R^2 and adjusted R^2.

Adjusted R^2 penalizes models with many variables, which is why R^2 and adjusted R^2 differ for models with many variables.

Task 3.4.7

independent variables	coefficient	t-value	p-value
intercept	7.55***	34.27	<0.01
log(floor_area)	0.20***	6.90	<0.01

Table 14: This table shows regression results for the formula log(electricity_consumption) = a + b * log(floor_area) + e. We log transform the data for better model fit and to be more consistent with the assumptions of normal distribution. Following White (1980), we adjust standard errors for heteroskedasticity. ***, ** and * indicate significance at the 1%, 5% and 10% level.

As can be seen in Table 14, the floor area of a residential household has a significant effect on the annual average energy consumption, both in statistical terms and in terms of the size of the effect. The relationship is positive, which implies that on average, homes with larger floor area have a higher electricity consumption.

Task 3.4.8

This could introduce multicollinearity into the model, as the number of bedrooms is bound to go hand in hand with a larger floor area. However, this does not seem to be the case with this dataset. The correlation between the two variables is low and the variance inflation factors do not signal a multicollinearity problem.

3.5 Explore the patterns in the data

Task 3.5.1

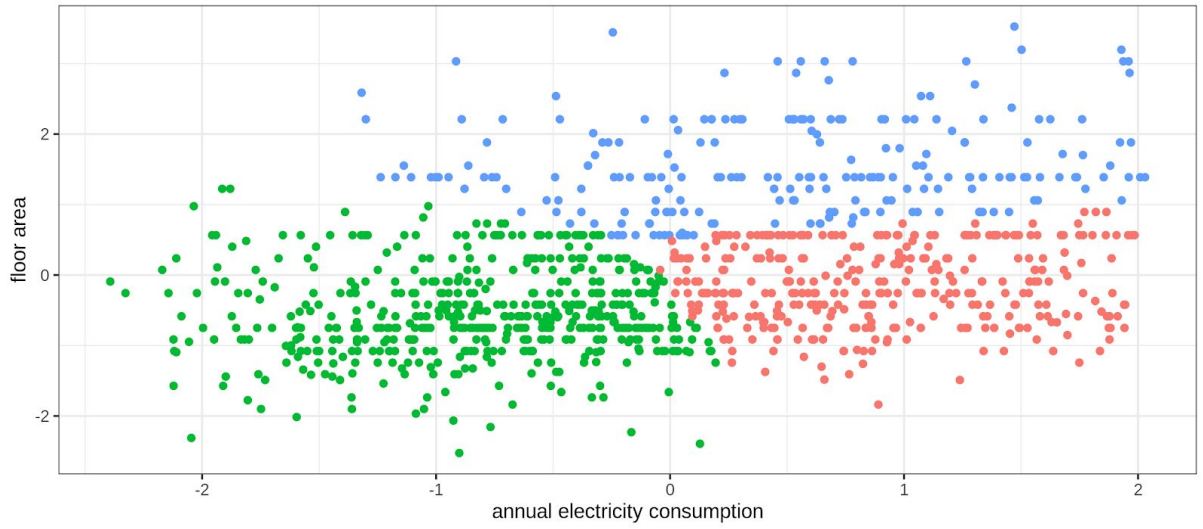

Figure 14: This figure displays clusters (individual data points with similar characteristics) of residential households derived by applying k-means clustering with k = 3 based on the two features floor area and annual electricity consumption, both normalized. Accumulations of single data points with similar values were indicated with a common colour (green, red or blue). Extreme outliers (see below) were removed beforehand.

Before clustering, extreme outliers above the 95th percentile are removed due to the well-known outlier sensitivity of the k-means algorithm. We include these for the clustering analysis using the DBSCAN algorithm, which is less prone to outliers. All data are normalized. Figure 14 shows cluster results using k-means clustering. The number of 3 clusters in total was determined using the elbow method. According to this, 2 clusters would be too few and a fourth cluster would only explain a little more. Therefore we chose k = 3 clusters. The green cluster seems to represent households with

smaller floor area and on average less annual electricity consumption. These might be small-sized households living in apartments. In the red cluster in the bottom right corner are households with smaller floor area and high electricity consumption. According to the correlations found in the previous analyses, these could be households in older buildings and with electric heating systems, for instance. In the upper right corner indicated with the blue cluster are households that tend to live in homes with larger floor area and higher average electricity consumption due to their size or home type.

Figure 15 shows the results of the DBSCAN clustering. Since this algorithm is density-based, it is less prone to outliers. The initially truncated outliers are thus now inluded again. Two clusters were identified, one that essentially combines extreme values of floor area and electricity consumption, and one that includes the remaining households with normal values. The results are hence very different from those of k-means clustering. They could be useful for detecting errors. For instance, the households in the red cluster in figure 15 might be more likely to be enterprises because of the large floor area or high electricity consumption. On the other hand, the results from k-means clustering may be more useful for marketing campaigns, as they also further subdivide the contiguous cyan cluster in figure 15.

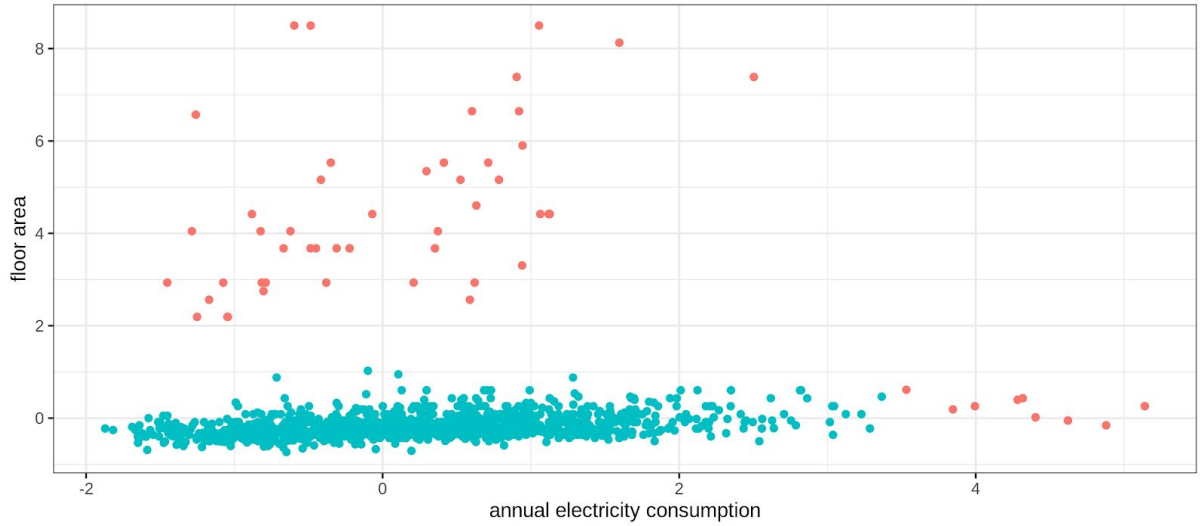

Figure 15: This figure displays clusters (individual data points with similar characteristics) of residential households derived by applying DBSCAN clustering based on the two features floor area and annual electricity consumption, both normalized. Accumulations of single data points with similar values were indicated with a common colour (red or cyan).

Task 3.5.2

Panel A: Confusion matrix					
predicted / real	1	2	3	4	5
1	0	0	0	0	0
2	7	267	104	67	23
3	2	21	15	25	5
4	0	27	50	61	36
5	1	1	7	8	6
Panel B: Performance evaluation metrics					
class	1	2	3	4	5
sensitivity	0.00	84.49	8.52	37.89	8.57
specificity	100.00	51.80	90.49	80.25	97.44
accuracy			47.61		
Cohen's Kappa			0.19		

Table 15: Panel A shows the confusion matrix and Panel B various metrics for classifier performance evaluation. Column values in Panel A are real labels from the test dataset, row values are predicted labels by the classifier. Household sizes larger than 5 were excluded due to too few occurrences.

We predict the number of residential household members using the features annual electricity consumption in kWh, number of bedrooms and the number of washing machines loads per day. We use a support vector machine (SVM) for classification and drop all data with more than 5 household members and an annual electricity consumption of more than 20,000 kWh due to too few occurrences, leaving us with N = 2,221 observations. Moreover, we use ⅔ of the data for training and ⅓ for testing and classifier performance evaluation.

The classifier performance in table 15 seems to be fairly decent, given that it is a multi-class problem and relatively little data is available. It should also be noted that, as Panel A indicates, when the classifier's prediction is wrong, it is mostly at least not totally wrong. However, accuracy and Cohen's Kappa in Panel B indicate a moderate classification performance which implies that the model clearly needs improvement. We performed feature engineering, analyzed various features from the survey, and tried several classification algorithms, but could not achieve any improvement. This leads us to the conclusion that the granularity at which the data from the survey responses are available is too coarse for prediction accurate to the household member. It is likely that the survey data are only sufficient for a binary classifier, such as small vs. large households.

Task 3.5.3

Panel A: Confusion matrix		
predicted / real	1	2
1	400	23
2	0	34
Panel B: Performance evaluation metrics		
sensitivity	100.00	
specificity	59.65	
accuracy	94.97	
Cohen's Kappa	0.7213	

Table 16: Panel A shows the confusion matrix and Panel B various metrics for classifier performance evaluation. Column values in Panel A are real labels from the test dataset, row values are predicted labels by the classifier.

To classify the observations labeled as "other" we apply the k-nearest neighbors algorithm with k = 11, determined through exhaustive hyperparameter tuning. We remove outliers for residential households with electricity consumptions of more than 20,000 kWh per year, since those households are probably rather enterprises than households. Again, we use ⅔ of the data for training and ⅓ for testing and classifier performance evaluation. Looking at the confusion matrix in Panel A of table 16, it is noteworthy that all residential households of the test dataset were classified correctly. However, only half of the SMEs were correctly assigned to their code. Obviously, the dataset seems to be imbalanced, which might cause the classifier to perform poorly for the minor class. This problem could be mitigated in a further analysis by resampling methods, for instance. Alternatively, it could simply be that some small SMEs are very similar to residential households in terms of the features used. This problem could be addressed by adding further features. Still, classifying half of the SMEs correctly using only two features is not bad. This is also supported by the metrics in Panel B, which means that our classifier performs a lot better than a random or naive classifier which assigns only one class. However, specificity shows that classifying SMEs remains fuzzy.