Analysis and Modeling of Energy Consumption using Dimensionality Reduction and t-SNE visualization

Manisha Ramesh
Computer Science
West Virginia University
Morgantown, West Virginia USA
mr0151@mix.wvu.edu

Abstract— This project is a study of energy consumption in residential households of the United States to identify the key factors or components for the usage and how to better optimize the performance with minimal use of resources. The building sector is a major source of energy consumption and greenhouse gas emissions in urban regions. Several studies have explored energy consumption analysis, and the value of the knowledge extracted is directly related to the quality of the data used. We are making use of the U.S Energy Information Administration to analyze consumption patterns across residential households and identify the factors responsible for usage. This paper visualizes Residential Energy Consumption Survey (RECS) 2015 (n-dimensional data) and uses feature reduction techniques to find what are the factors which have the most impact on residential energy usage.

Keywords—dimensionality reduction, energy consumption, modeling, t-SNE

I. Introduction

Understanding factors concerning Energy Consumption in the United States has been a topic of discussion many times since it involves the environment, people and economy. Among all the different contributors to energy consumption, it is observed that the residential energy consumption accounts to 17% of total global carbon emissions directly. Thus, there becomes a need to conduct a survey on the household energy consumption and identify characteristics and parameters that affect the most. Energy management and control on the supply-side conduct comprehensive analyses based on electricity usage, weather forecasts and the characteristics of the heating and cooling systems used in the buildings to determine the optimal operation and control scheme. In addition, the demand-side management aims to guide the users' electricity usage in a scientific and reasonable way by adjusting the user loads or users' behavior of electricity consumption through economic subsidies, compulsory legal means or publicity means. Due to the impact of dynamic real-time changes on both supply and demand sides, it is important to classify and predict the energy consumption of residential buildings from historical data to provide a sufficient decision-making basis for planning power transmission configuration patterns that meet regional characteristics. The energy consumption prediction is of decisive importance for the improvement of the power grid quality and the rational allocation of the power supply, which contributes to the enhancement of life quality and the optimization of energy usage. Therefore, there are efforts in related works by researchers around the world that are geared towards improving energy consumption predictions.

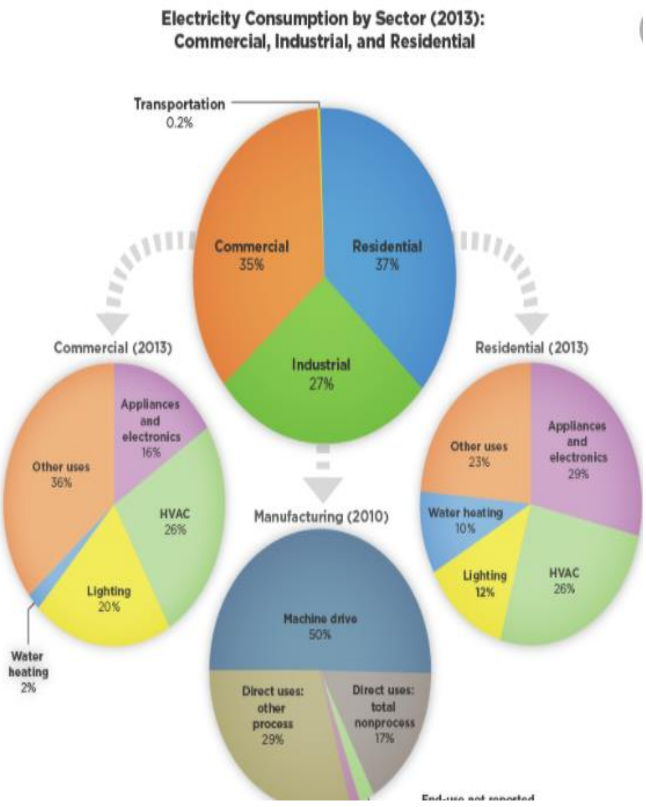


Fig 1: Energy Consumption in 2013 across sectors proving residential is the highest utilized.

It is difficult for the power sector to use energy simulations for predicting the energy consumption of an entire region due to the unavailable of some detailed data which requires large quantities of various sensors. As such, the use of open data for modeling research and predictions is a good choice.

The data is described along with the analytical approach. Following this, we present and discuss the findings, offering concluding comments at the end. In a nutshell, the findings show mixed relationships between energy efficiency measures and residential energy consumption. They reveal positive relationships between the measures of affluence and energy consumption. Importantly, they indicate state context and factors do influence residential energy consumption.

II. RELATED WORK

"Beckel, C.; Sadamoria, L.; Staake, T.; Santini, S. Revealing Household Characteristics from Smart Meter Data. Energy 2014, 78, 397–410". This paper creates awareness and improves efficiency in available energy usage by identifying individual patterns of energy usage and thereby helping end users understand their energy consumption. It talks about how utilities are currently deploying smart electricity meters in millions of households worldwide to collect fine-grained electricity consumption data. The paper presents an approach to automatically analyze the data to enable personalized and scalable energy efficiency programs for private households. They develop and evaluate a system that uses supervised machine learning techniques to automatically estimate specific "characteristics" of a household from its electricity consumption. The characteristics are related to a household's socio-economic status, its dwelling, or its appliance stock. Evaluation of the system is done by analyzing smart meter data collected from 4,232 households in Ireland at a 30-minute granularity over a period of 1.5 years. Analysis results showed that revealing characteristics from smart meter data is feasible, and the method achieved an accuracy of more than 70% over all households for many of the characteristics and even exceeds 80% for some of the characteristics. The findings are applicable to all smart metering systems without making changes to the measurement infrastructure. On the basis of these promising results, the paper discusses the potential for utilities as well as policy and privacy implications

"*Even for the environment, context matters! States, households, and residential energy consumption*" This paper is about a multi-level approach to examine the extent to which state and household-level factors shape residential energy consumption in the United States, focusing on efficiency improvement and affluence.

This study adopts a multi-level approach to examine the extent to which state- and household-level factors shape residential energy consumption in the United States, focusing on efficiency improvement and affluence. They have analyzed the 2009 Residential Energy Consumption Survey, state-level energy efficiency data from the American Council for an Energy-Efficient Economy (ACEEE), and other sources, we find that state context significantly influences energy consumption at the household level. Households in states with high value on energy efficiency consume significantly less residential energy than those in states scoring low on the measure. At the household level, the analysis revealed mixed relationship between investment in energy efficiency technologies and residential energy consumption, as some measures of efficiency technology are negatively related to residential energy consumption, others are positively related to it. In regard to affluence, state-level measures do not emerge as significant predictors of residential energy consumption. At the household level, however, affluence drives residential energy consumption, which, in turn, is a significant driver of carbon dioxide emissions. This paper makes an important contribution to the social scientific literature on energy consumption, illuminating distinct relationships at different levels. It is considered to be the first study that simultaneously examines the impacts of factors measured at both the household (micro) and state (meso) levels.

"Modeling and analysis of energy data: state-of-the-art and practical results from an application scenario by Maria Riveiro, Ronnie Johansson and Alexander Karlsson", presents a comprehensive summary of the state-of-the-art of energy efficiency research. The literature review carried out focuses on the application of data mining and data analysis techniques to energy consumption data, as well as descriptions of tools, applications and research prototypes to manage the consumption of energy. Moreover, preliminary results of the application of a clustering technique to energy consumption data illustrate the review.

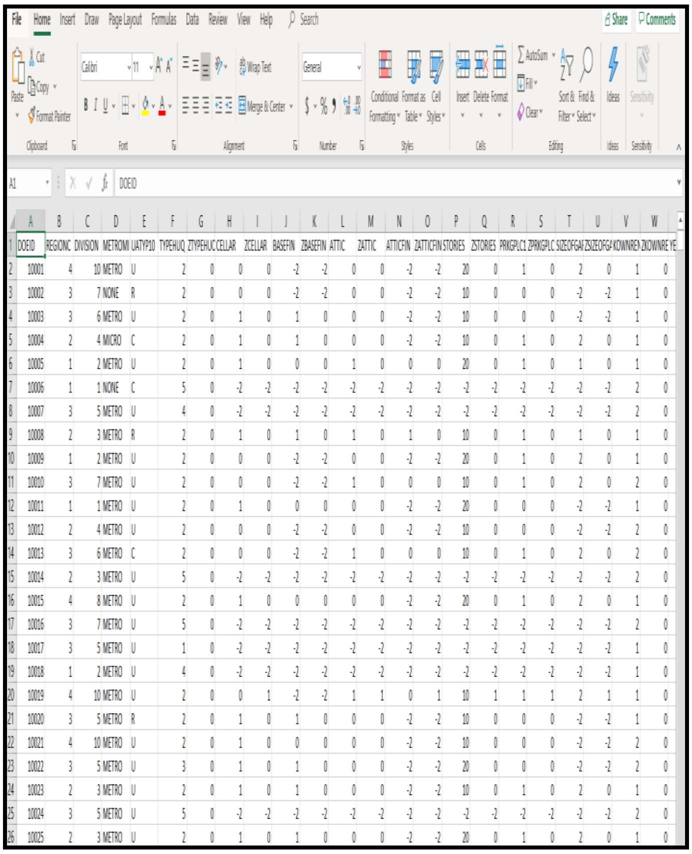
III. METHODOLOGY

This project visualizes Residential Energy Consumption Survey (RECS) 2015 (n-dimensional data) and uses feature reduction techniques to find what are the factors which have the most impact on residential energy usage

Project flow

1. Prepare a RECS 2015 data, which is a high-dimensional dataset.
2. Understand challenges with visual interpretation of the raw data.
3. Reduce the dataset to 2 dimensions by using t-SNE for plotting and interpretation.
4. Assess if the type of housing unit feature's impacts on the energy consumption pattern of US households.

The dataset is a sample of 12000+ households selected randomly using a multi-stage probability design representing the 113.6 million US households. The initial step is to load the data. The dataset consists of 931 features from the climate, type of housing unit, HDD, CDD till the sizes of every room in the house. We consider only the numeric features and discard the 9-character features from the dataset.



DOEID	REGIONC	DIVISION	METROMI	UATYP10	TYPEHUQ	ZTYPEHUC	CELLAR	ZCELLAR	BASEFIN	ZBASEFIN	ATTIC	ZATTIC	ATTICFIN	ZATTICFIN	STORIES	ZSTORIES	PRKGPLC1	ZPRKGPLC	SIZEOFGA	ZSIZEOFGA	KOWNREN	ZKOWNRE
10001	4	10	METRO	U	2	0	0	0	-2	-2	0	0	-2	-2	20	0	1	0	2	0	1	0
10002	3	7	NONE	R	2	0	0	0	-2	-2	0	0	-2	-2	10	0	0	0	-2	-2	1	0
10003	3	6	METRO	U	2	0	1	0	1	0	0	0	-2	-2	10	0	0	0	-2	-2	1	0
10004	2	4	MICRO	C	2	0	1	0	1	0	0	0	-2	-2	10	0	1	0	2	0	1	0
10005	1	2	METRO	U	2	0	1	0	0	0	1	0	0	0	20	0	1	0	1	0	1	0
10006	1	1	NONE	C	5	0	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	2	0
10007	3	5	METRO	U	4	0	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	2	0
10008	2	3	METRO	R	2	0	1	0	1	0	1	0	1	0	10	0	1	0	1	0	1	0
10009	1	2	METRO	U	2	0	0	0	-2	-2	0	0	-2	-2	20	0	1	0	2	0	1	0
10010	3	7	METRO	U	2	0	0	0	-2	-2	1	0	0	0	10	0	1	0	2	0	2	0
10011	1	1	METRO	U	2	0	1	0	0	0	0	0	-2	-2	20	0	0	0	-2	-2	1	0
10012	2	4	METRO	U	2	0	0	0	-2	-2	0	0	-2	-2	10	0	0	0	-2	-2	2	0
10013	3	6	METRO	C	2	0	0	0	-2	-2	1	0	0	0	10	0	1	0	2	0	2	0
10014	2	3	METRO	U	5	0	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	2	0
10015	4	8	METRO	U	2	0	1	0	0	0	0	0	-2	-2	20	0	1	0	2	0	1	0
10016	3	7	METRO	U	5	0	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	2	0
10017	3	5	METRO	U	1	0	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	1	0
10018	1	2	METRO	U	4	0	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	1	0
10019	4	10	METRO	U	2	0	0	1	-2	-2	1	1	0	1	10	1	1	1	2	1	1	0
10020	3	5	METRO	R	2	0	1	0	1	0	0	0	-2	-2	10	0	0	0	-2	-2	1	0
10021	4	10	METRO	U	2	0	1	0	0	0	0	0	-2	-2	10	0	0	0	-2	-2	2	0
10022	3	5	METRO	U	3	0	1	0	1	0	0	0	-2	-2	20	0	0	0	-2	-2	2	0
10023	2	3	METRO	U	2	0	1	0	1	0	0	0	-2	-2	10	0	1	0	2	0	1	0
10024	3	5	METRO	U	5	0	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	2	0
10025	2	3	METRO	U	2	0	1	0	1	0	0	0	-2	-2	20	0	1	0	2	0	1	0

Fig 2: Dataset information dataset for energy consumption across several states in USA for the year 2015

DATA MODELLING

The steps involved in modelling the data are:

1. Import data and libraries 1.1 Defining Local Functions 1.2 Load Data
2. Raw Data Exploration 2.1 View class group sizes 2.2 View distributions of feature values
3. Dimensionality Reduction 3.1 Check for degeneracy and compressability (using scipy svd method) 3.2 Feature Reduction (using scikit-learn TruncatedSVD method)
4. Visualisation of dataset with reduced dimensionality 4.1 Pairs Plots 4.2 3D Scatterplot
5. t-SNE Reduction and Visualisation 5.1 Observe scaling on basic t-SNE 5.2 Calculate t-SNE representation 5.3 View t-SNE representation with and without class labels

There are ~360 features in the dataset which are just imputation flags for the other features. They are thus creating degeneracy in the data. By a quick visual inspection, we have determined that the names of all the imputation flag variables start with the letter 'Z'. So, we remove those features from our dataset.

	DOEID	REGIONC	DIVISION	REPORTABLE_DOMAIN	TYPEHUQ	NWEIGHT	HDD65	CDD65	HDD30YR	CDD30YR	...	SCALEEL	KAVALNG	PERIODNG
0	1	2	4	12	2	2471.679705	4742	1080	4953	1271	...	0	-2	-2
1	2	4	10	26	2	8599.172010	2662	199	2688	143	...	0	1	1
2	3	1	1	1	5	8969.915921	6233	505	5741	829	...	0	3	5
3	4	2	3	7	2	18003.639600	6034	672	5781	868	...	3	3	5
4	5	1	1	1	3	5999.605242	5388	702	5313	797	...	0	1	1

5 rows × 931 columns

Fig 3: After loading dataset in pandas framework

Original raw dataset shape (12083, 931)

After discarding the features, the shape of the dataset (12083, 572)

VIEWING CLASS GROUP SIZES

HYPOTHESIS

The initial assumption is to consider "Type of Housing Unit-TYPEHUQ" as a feature to classify data and check for any patterns across other features that are characterized with TYPEHUQ.

There are 5 types of housing units in this dataset.

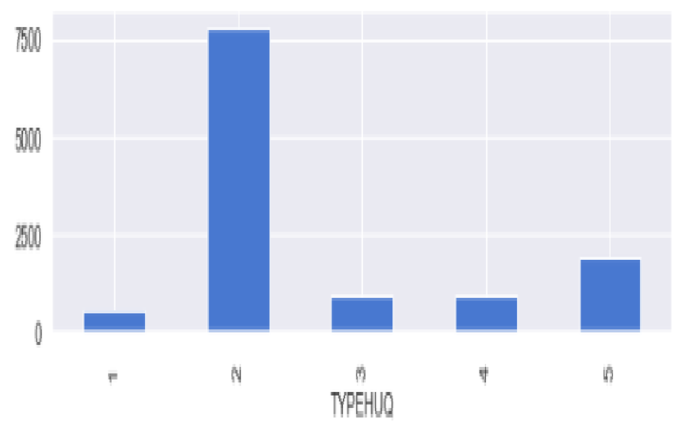
1- Mobile Home

2- Single-Family Detached

3- Single-Family Attached

4- Apartment in Building with 2 - 4 Units

5- Apartment in Building with 5+ Units



From the below graph we observe, the classes are heavily skewed, with TYPEHUQ 2 (Single family-detached), being most of the houses in this entire dataset.

VIEWING DISTRIBUTION OF INDEPENDENT VARIABLES

We determine 21 major features to create a boxplot visualization which shows how houses of different types TYPEHUQs classification cluster together.

MAJOR FEATURES

- HDD65 Heating degree days in 2015, base temperature 65F
- CDD65 Cooling degree days in 2015, base temperature 65F
- WALLTYPE Major outside wall material
- BEDROOMS Number of bedrooms
- TOTROOMS Total number of rooms in the housing unit
- STOVEN Number of stoves (one appliance with cooktop and an oven)
- STOVEFUEL Fuel used by most-used separate cooktop
- AMTMICRO Microwave usage
- AGERFRI1 Age of most-used refrigerator
- ESFRIG Energy Star most-used refrigerator
- HEATHOME Space heating equipment used
- EQUIPAGE Age of main space heating equipment
- AUTOHEATNITE Programmable thermostat lowers temperature at night
- AUTOHEATDAY Programmable thermostat lowers temperature during the day
- TEMPGONE Temperature when no one is home during the day (winter)
- BTUNGSPH Natural Gas usage for space heating, in thousand BTU, 2015
- TOTALBTUCOL Total usage for air conditioning, in thousand BTU, 2015
- NHSLDMEM Number of household members
- EMPLOYHH Employment status of householder
- AGEAUD Year of home energy audit
- WINDOWS Number of windows in heated areas

In the box plot below, HDD65: Heating degree days in 2015, base temperature 65F is the first feature. The boxplot shows that TYPEHUQ 3 houses are in warmer parts of the country. But it's hard to state how different are all the other basic features and how does that impact the energy usage in the houses. It is also hard to visualize the differences between household types classifications across all the 572 features.
Thus, we need a better visualization for identifying correlation across multiple dimensions.

Heating degree days are defined relative to a base temperature—the outside temperature above which a building needs no heating.
To resolve this issue, we go in for Dimensionality reduction.

.

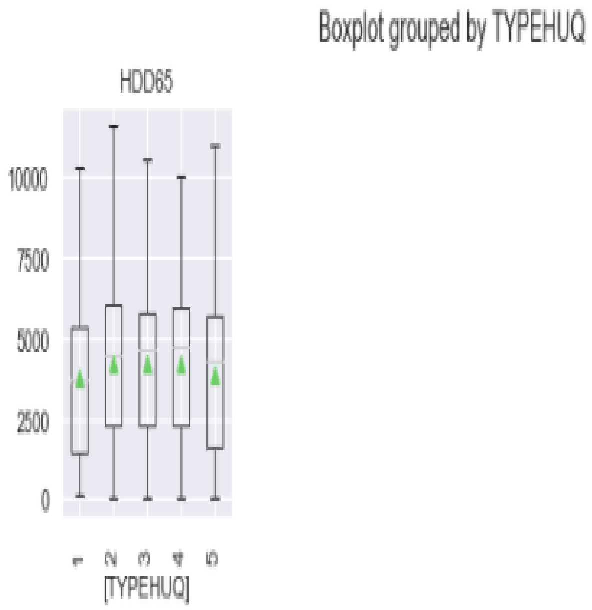


Fig 5: Box Plot for Heating degree days on Type of Housing Units.

DIMENSIONALITY REDUCTION

1. NORMALIZATION AND STANDARIZATION

This step is an important when dealing with parameters of different units and scales. All parameters should have the same scale for a fair comparison between them because some data mining techniques use the Euclidean distance. Two methods well methods for rescaling data are:

1. Normalization - scales all numeric variable s in the range [0,1].
2. Standardization - transforms the data to ha ve zero mean and unit variance.

Out[12]:

	REGIONC	DIVISION	REPORTABLE_DOMAIN	TYPEHUQ	NWEIGHT	HDD65	CDD65	HDD30YR	CDD30YR	Climate_Region_Pub	...	SCALEEL	KAV
0	2	4	12	2	2471.679705	4742	1080	4953	1271	4	...	0	
1	4	10	26	2	8599.172010	2662	199	2688	143	5	...	0	
2	1	1	1	5	8969.915921	6233	505	5741	829	1	...	0	
3	2	3	7	2	18003.639600	6034	672	5781	868	1	...	3	
4	1	1	1	3	5999.605242	5388	702	5313	797	1	...	0	

5 rows × 569 columns

Fig 6: After replacing NaN and missing values with 0.

Out[14]:

	REGIONC	DIVISION	REPORTABLE_DOMAIN	TYPEHUQ	NWEIGHT	HDD65	CDD65	HDD30YR	CDD30YR	Climate_Region_Pub	...	SCALEEL	KAV
0	0.333333	0.333333	0.423077	0.25	0.020939	0.378603	0.197080	0.371122	0.237260	0.75	...	0.0	
1	1.000000	1.000000	0.961538	0.25	0.085234	0.212535	0.036314	0.201409	0.026694	1.00	...	0.0	
2	0.000000	0.000000	0.000000	1.00	0.089124	0.497645	0.092153	0.430166	0.154751	0.00	...	0.0	
3	0.333333	0.222222	0.230769	0.25	0.183914	0.481756	0.122628	0.433163	0.162031	0.00	...	1.0	
4	0.000000	0.000000	0.000000	0.50	0.057957	0.430180	0.128102	0.398097	0.148777	0.00	...	0.0	

5 rows × 569 columns

Fig 7: After normalization and standarization.

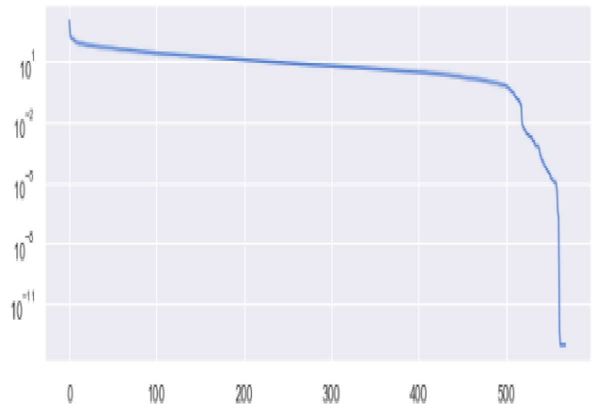
2. SINGULAR VALUE DECOMPOSITION

DEGENERACY CHECK AND COMPRESSION

Using the SVD function from the Scipy package, calculate the SVD and observe the singular values. If any are very close to zero then we have some degeneracy in the full dataset which we should definitely try to remove or avoid.

$$M = U \sum V^*$$

```
0 SVs are NaN
8 SVs are less than 1e-12
```



568 original features could be compressed into 300 components (eigenvectors) with only 0.57% loss of variance. There's clearly some near degeneracy in the data.

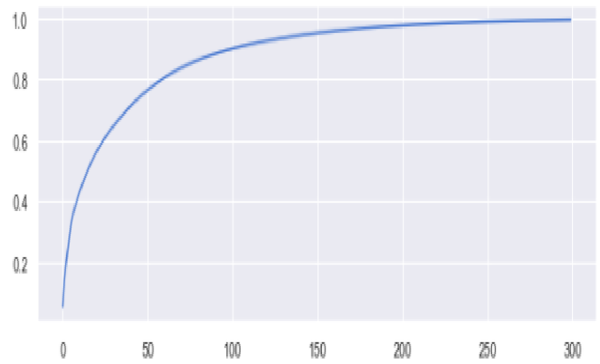
In extreme cases, we could consider representing the dataset using only the first 50 components and still maintain ~61% of the variance.

3. TRUNCATED SVD

Using the TruncatedSVD method in the scikit-learn package, transform the full dataset into a representation using the top 300 components, thus preserving variance in the data but using fewer dimensions/features to do so. This has a similar effect to Principal Component Analysis (PCA) where we represent the original data using an orthogonal set of axes rotated and aligned to the variance in the dataset.

With 300 as the number of features in the reduction process preserves a lot of variance (~99%) and is still too large to easily visualize.

```
Variance preserved by first 300 components == 99.41%
```

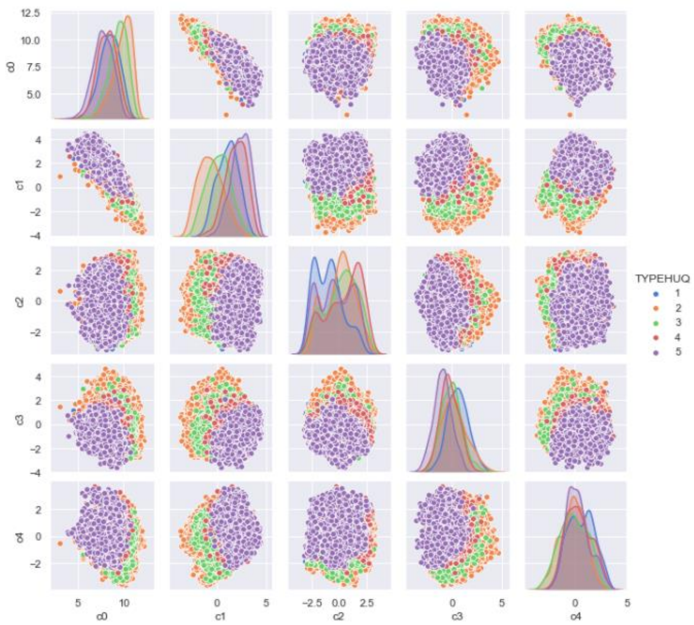


VISUALIZATION

Visualize the data with the compressed dataset, represented by the top 300 components of an SVD.

PAIR PLOTS

Pairs-plots are a simple representation using a set of 2D scatterplots, plotting each component against another component, and coloring the datapoints according to their classification.

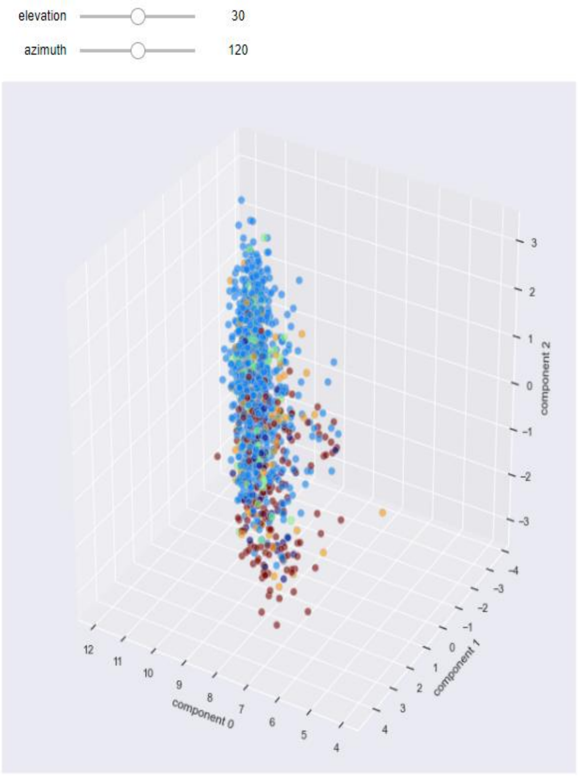


- We can clearly see that datapoints from the same type of housing units tend to cluster together
- It's possible to see distinct TYPEHUQs on different sides of the plots, e.g.
    - TYPEHUQ 2 and TYPEHUQ 5 seem to predominate hiding away the information related to other TYPEHUQs

- Pair-plot suffers several issues for visualization:
  - It shows only the first 5 components of the available 300, we only see a small percentage of the full variance
  - If we try to show more components, the overall plot will get very large.
  - It's hard to get a full appreciation of the differences and similarities between datapoints across all the components, suitable for the viewer to interpret results.
  - Classes with lower counts e.g. TYPEHUQ 1, TYPEHUQ 3 and TYPEHUQ 4 are hard to see.

3D SCATTER PLOT

As an alternative to the pairs-plots, we could view a 3D scatterplot, which lets us see more dimensions at once and possibly get an interactive feel for the data.



Using Python Notebook interactive package, we can create an interactive plot with controls for elevation and azimuth

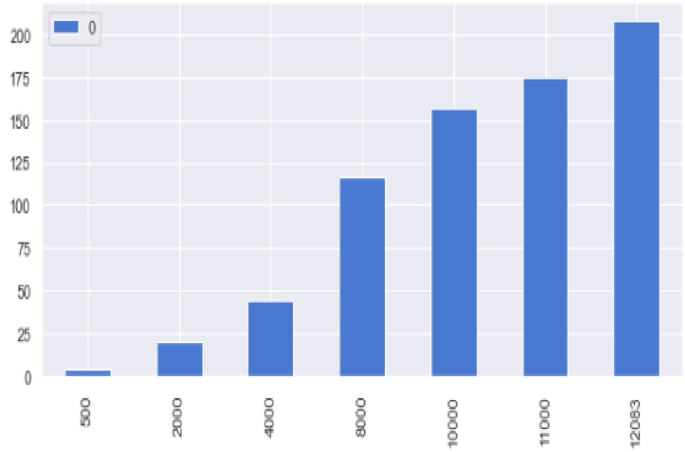
We can use these controls to interactively change the view of the top 3 components and investigate their relations. This appears to be more informative than pairs-plots.

The same major limitations of the pairs-plots; we lose a lot of the variance and must hold a lot in our heads when viewing.

We need a better visualization technique to view the ~12000 datapoints and their 300 dimensions and to validate our hypothesis and say if TYPE of Housing Unit is a major factor to classify households for their energy consumption patterns.

t-SNE VISUALIZATION

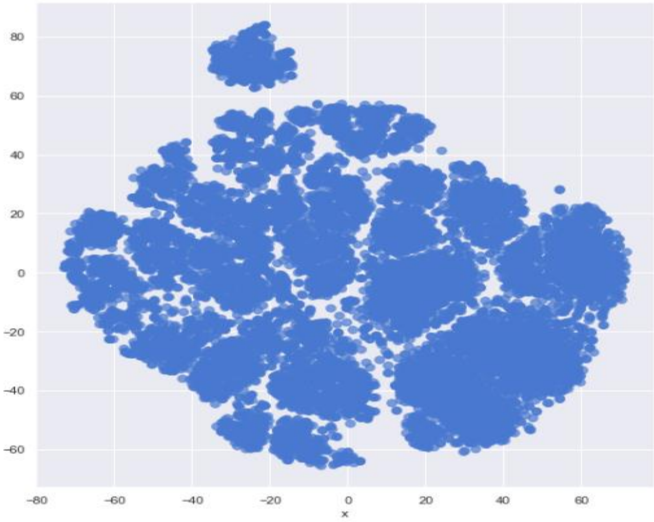
We will use the basic implementation available in scikit-learn which has O(n^2) complexity. Ordinarily this prohibits use on real-world datasets (and we would instead use the Barnes-Hut implementation O(n*log(n))), but for out 675 datapoints it's no problem.



From the above diagram, we see that basic t-SNE scales badly O(n^2). However, this dataset is of manageable size and calculation is quick on my laptop, so I won't use the faster Barnes-Hut implementation O(n*log(n)).

CALCULATE t-SNE REPRESENTATION

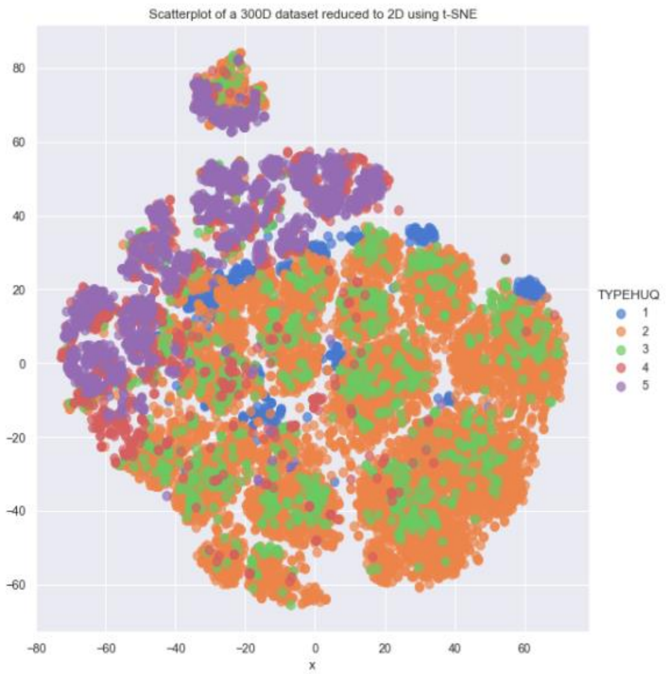
We'll calculate the t-SNE 2D representation for our 300D dataset and view the results without coloring the classes.



This has reduced our top 300 SVD components to 2 dimensions and we can clearly see clusters of datapoints. When doing unsupervised learning (without class labels) we might expect to get this far and then backtrack to run clustering algorithms in the 300 dimensional space.

The next step is to view the t-SNE representation with TYPEHUQ classification class labels

t-SNE REPRESENTATION WITH CLASS LABELS



The above plot shows us that the type of housing unit alone doesn't classify/categorize the households for their energy consumption patterns. Nonetheless, there is a good degree of clustering for Type 2 (Orange) and Type 5 (Green) houses.

CONCLUSION

This paper presents an overview of how Residential Energy Consumption affects the total carbon emissions and contributes to green-house effect indirectly. We have made use of RECS2015 data to identify characteristics that forms patterns based of energy consumption. We read from and wrote to different file format (csv) and databases (sql)We prepared the data by cleaning (removing character features values, replacing nans) and normalizing. We applied transformation during the feature reduction stage. We then visualized the data in the reduced dimensionality and ultimately applied t-SNE algorithm to reduce the data into two dimensions and visualize effectively. It was observed that the hypothesis assumed was not true always since “TypeOfHousingUnit” alone doesn’t categorize the households for their energy consumption patterns.

REFERENCES

[1] Daswin Desilva, Xinghuo. Yu, Damminda. Alahakoon and Grahame. Holmes, "A data mining framework for electricity consumption analysis from meter data", *Industrial Informatics IEEE Transactions on*, vol. 7, no. 3, pp. 399-407, 2011..

[2] Ian. Dent, Uwe. Aickelin and Tom. Rodden, "The application of a data mining framework to energy usage profiling in domestic residences using uk data", *arXiv preprint arXiv*, pp. 1307-1380, 2013.

[3] G. Asimakopoulou, Y. Papagrigorakis, A. Dimeas, P. Aristidou, and N. Hatziargyriou. A review of customer management tools: The energy industry. In Nikos Hatziargyriou, Aris Dimeas, Thomai Tomtsi, and Anke Weidlich, editors, Energy-Efficient Computing and Networking. Revised selected papers. First International Conference, E-Energy 2010, Athens, Greece, pages 64–72. Springer Heidelberg Dordrecht London New York, October 2011

[4] L. Bartram, J. Rodgers, and K. Muise. Chasing the negawatt: Visualization for sustainable living. Computer Graphics and Applications, IEEE, 30(3):8–14, 2010.

[5] C. Comito, D. Talia, and P. Trunfio. “An Energy-Aware Clustering

Scheme for Mobile Applications”. IEEE Scalcom, pp. 15–22, 2011.

[6] [1] W. Poortinga, L. Steg, C. Vlek, G. Wiersma, Household preferences for energy-saving measures: A conjoint analysis, Journal of Economic Psychology 24 (1) (2003) 49–64.

[7] Adua L and Sharp J S 2011 Explaining residential energy consumption: a focus on location and race differences in natural gas use *J. Rural Soc. Sci.* **26** 107–41

[8] Adua L, York R and Schuelke-Leech B-A 2016 The human dimensions of climate change: a micro-level assessment of views from the ecological modernization, political economy and human ecology perspectives *Soc. Sci. Res.* **56** 26–43