# WHICH AUSTRALIAN HOUSEHOLDS SPEND THE MOST ON ELECTRICITY

*By*

*SAISRIKAR PARUCHURI*

*Comp 7801*

Department of Information Technology and Electrical Engineering

University of Queensland.

Submitted for the degree of

Master of Computer Science (Management)

in the division of Engineering, Architecture and Information Technology

November & 2018.


THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

ii

iii

2 Simon Street, Underwood
QLD, 4119
Mob. 0420554194
September 10, 2018

The Dean
School of Engineering
University of Queensland
St Lucia, Q 4072

Dear Professor Simmons,

In accordance with the requirements of the degree of Master of Computer Science (Management) in the division of Engineering, Architecture and Information Technology (EAIT), I present the following thesis entitled "Which Australian Households Spend the Most on Electricity". This work was performed under the supervision of Dr Mark Philip Griffin.

I declare that the work submitted in this thesis is my own, except as acknowledged in the text and footnotes, and has not been previously submitted for a degree at the University of Queensland or any other institution.

Yours sincerely,

*Author's Signature*

SAISRIKAR PARUCHURI.

iv

v

*To . . .*

# Acknowledgments

I first want to thank my supervisor Dr. Mark Philip Griffin from the Information Technology and Electrical Engineering department at The University of Queensland. Dr. Mark was always open whenever I get stuck in the project or had a question regarding my project. He reliably enabled this paper to be my own work, however my supervisor guided me in the right direction whenever it was necessary.

I likewise want to thank the University of Queensland for providing me with all the resources that were necessary for my thesis project.

At last, I should offer my exceptionally significant thanks to my parents and to my brother for providing me with the necessary financial support and consolation during my time of study. This achievement would not have been conceivable without the project supervisor and my family. Much obliged to you.

# Abstract

Existing literature has explored the variation of electricity consumption by household characteristics like household size, age, IT appliances usage, working from home etc. The goal of this project is to find the type of activities from the Australian households that lead to higher electricity usage. This project will use variables that are financial year wages and salaries, home repairs, physical functioning, physical activity, outdoor tasks, and unpaid work from the 2015 HILDA survey to explore the households that spent the most on electricity usage. The Household, Income and Labour Dynamics in Australia (HILDA) survey is a large, longitudinal survey that is the only survey of its kind that is nationally representative. This survey has followed approximately 17,000 people since 2001, with 15-time points of data released to date. This project uses statistical analysis to analyse data from the 2015 HILDA survey.

Table of Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

This project uses energy as an input to deliver the services that are needed for Australian households. Australian households represent a major group of consumers of energy resources like electricity [3]. Moreover, there is a need to focus on the electricity usage of the younger generation because future generations will be strongly affected by energy system changes [3]. Usage of electricity in households is an important political topic and an important topic in social science research [3]. For instance, electricity consumption seems to increase as the number of people in a household increase. House members have a crucial impact on their energy usage through behaviours such as their use of appliances [3]. Residential electricity usage can be explained by indirect and direct determinants [3]. Matthies et al. 2016 propose a system that can help explain the indirect sociodemographic and economic factors on electricity consumption by considering the direct behavioural and motivational components simultaneously.

The household members age is a relevant sociodemographic factor that is correlated with electricity consumption [3]. Several investigations have reported that when the number of household members in a household increases, electricity usage also increases [3]. Income has frequently been identified as a positive prediction of electricity consumption [3]. However, high correlations between the resident's income, house characteristics, and number of household members may lead to methodological problems such as multicollinearity [3].

To effectively target marketing campaigns, it is important to cluster the households into a manageable number of groupings so that each group can be

presented differently [4]. Currently, the utility companies use demographic data that is house size, family size, location etc. as the basis for clustering [4]. The work shown makes use of electricity meter data to explore whether useful clusters can be achieved based on the household's behaviour [4].

One stream of sustainable usage research is directed towards household consumption patterns related to environmental patterns [6]. Household expenditures are verified in a life cycle context, by identifying the environmental impact [6]. In the following, I have explained the goals, objectives, and significance of the project.

### 1.1 Goals and Objectives

The aim of this project is to identify the type of activities from Australian households that leads to higher electricity bills and explore the reasons for variation of the households in electricity consumption by using variables from the Household, Income and Labour Dynamics in Australia (HILDA) 2015 survey. Firstly, this project considered some variables from the Household, Income and Labour Dynamics in Australia (HILDA) 2015 survey that is household size, total children ever had, how often do you take care of your grandchildren, hours per week for paid employment, hours per week traveling to and from a place of employment, annual household expenditure on electricity bills, and where the household member works. However, after analysing all the variables through linear regression on a one-time basis it was proven that all the variables are not related to electricity usage more details are provided in the result section.

Again, I have chosen variables from the Household Income and Labour Dynamics in Australia (HILDA) 2015 survey that is of two types they are new person questionaries and self-completion questionaries. In new person questionaries the variables are enrolled in course of study to obtain qualification, currently enrolled in a course, currently receive income from wages/salary, current marital status, current living circumstances, k1 long term health condition, what are the main tasks and duties you undertake in your occupation, total gross amount of most recent pay before deductions, gross financial year wages and salaries, annual household expenditure.

In self-completion questionaries the variables are currently an active member of a sporting/hobby/community based club or association, hours/minutes per week-outdoor tasks, hours/minutes per week-volunteer/charity work, regularly pay someone to do housework, are you currently in paid work, physical functioning activities, self-assessed health, how often participate in physical activity, how often feel rushed or pressed for time, spare time that don't know what to do with, how often get together socially with friends/relatives not living with you, prosperity given current needs and financial responsibilities.

This project uses some of the above variables that are related to the electricity usage to explore the variation in electricity consumption by using statistical analysis like linear regression factor analysis, k-means clustering and linkage algorithm. Finally, with this research project, I will be able to find the type of activities from Australian households that leads to higher electricity bills.

### 1.2 Significance of Project

This research project will provide information on the household's activities that spend the most on electricity bills. The Household, Income and Labour Dynamics in Australia (HILDA) survey is a large, longitudinal survey that is the only survey of its kind that is nationally representative. This survey has followed approximately 17,000 people since 2001, with 15-time points of data released to date. I preferred HILDA dataset because the survey has followed so many people and it is nationally representative in Australia. I am the first person to find the activities of Australian households that leads to higher electricity bills by using 2015 HILDA dataset.

I will be using the new person questionaries and self-completion questionaries forms from The University of Melbourne website to choose variables for this project. Also, I have downloaded the 2015 HILDA Stata dataset, I got access to HILDA dataset by the UQ librarian. But the problem is 2015 HILDA Stata data contains only variable names, perhaps it is very difficult to identify the meaning of the variable. So, the University of Melbourne

provides HILDA online data dictionary in their website to provide details of variables. For instance, variable meaning, variable name etc for all the variables.

To analyse the variables from 2015 HILDA dataset firstly, I did linear regression analysis to know whether the variables are related or unrelated to electricity usage. Also, whether the variable is positive or negative towards electricity usage on a uni-variate basis. All the variables in this project, I have replaced negative values with NA and applied log transform which results in normal distribution to get better results. I have considered variables that are related to electricity usage from the linear regression analysis. For the selected variables I have applied advanced statistical analysis.

The above data can be useful for end users as well as the electricity providers. The end users can know the variation in household consumption as a function of an electricity usage on electric appliances, household size etc. Customers can also know whether they are getting a good electricity deal by comparing their electricity bills with other electricity providers. This research is useful for the energy providers so that they can know the usage of households and specifically target the households with good deals on electricity services and to improve the overall electricity network.

# Chapter 2

# Literature review / prior art

### 2.1 Electricity Usage

Extensive research has been done on various determinants of electricity usage. However, how specific socio-demographic, attitudinal determinants and behaviours influence residential electricity usage are still scarce [3]. Matthies et al. [3] used hierarchical regression analysis to systematically investigate the household engagement in electricity saving along with a wide range of other measures in some sample households. Special attention was given to households with young people and children by analysing the influence of the number of teenage people on electricity usage in a path model [3]. The researchers suggest that the use of behavioural information provides more detailed information on electricity usage [3]. Matthies et al. [3] aimed to explain the indirect influences of economic factors and socio-demographic variables on electricity usage by considering motivational components and behaviours simultaneously.

Matthies et al. [3] aim is to explain the indirect influences of economic factors and sociodemographic on electricity usage by finding the link between past purchasing behaviours and present usage behaviours. They assume that other activities would indirectly provide information about these behaviours. For example, the time residents spend at home was expected to increase the use of appliances, such as more use of warm water and frequent cooking [3]. First, they examined the influence of indirect factors such as income, number of adolescents on electricity consumption and number of residents [3].

Second, they considered that the influences of income, number of adolescents, and household characteristics on electricity usage would be mediated by the behaviours and activities [3]. Differences in electricity usage are primarily due to differences in purchasing and usage behaviours. Their research indicates that the number of adolescents in some households led to higher electricity usage because adolescents frequently use IT appliances that were

positively correlated with high household electricity usage [3]. Finally, their research suggests that having adolescents in a household led to higher electricity usage because adolescents spend more time at home [3].

Generally knowing, how a household varies their regular consumption of electricity is useful for an organization to allow accurate targeting of behaviour modifications with the aim of improving the quality of the electricity network. The variability of daily activities in a household is one possible way for the household to accept incentives to modify their behaviour [4]. To evaluate the variability measures of a household they validate the number of clusters indexes [4]. These indices are varying with several clusters, quality of attributes and the number of attributes [4]. The Cluster Dispersion Indicator (CDI) and the Davis-Boulden Indicator (DBI) are taken into consideration for the household behaviour variability [4].

To validate the household variability indices by the CDI and Davis-Boulden Indicator they have taken 180 households monitored over a year at an interval of 5 minutes. The time is taken from the peak electricity usage period that is 4pm to 8pm [4].

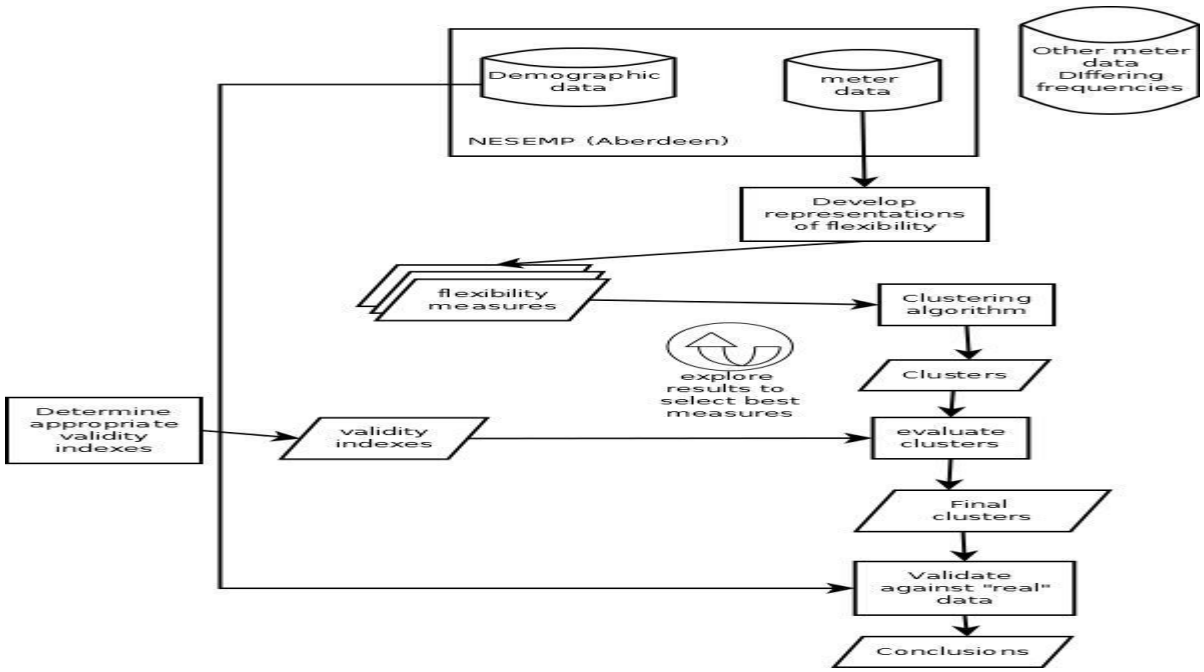



**Figure 1: Electricity Usage**

The authors have adopted the above approach to cluster the households according to the household's electricity usage [4].

### 2.2 Household/Family Structure

In today's culture, the electricity consumption is dependent on the household size, activities of the household, and age of the person (for example teenagers, kid's etc.) [3].

Generally knowing, the more time people spend at home the greater the electricity usage will be. Some people work from home at the time they will be using air-conditioners and other IT appliances, so the electricity bills will be higher for those households. Most households use IT appliances like refrigerators, air-conditioners, microwaves, heaters etc. for their daily needs. Matthies et al. [3] have considered the number of residents in a house and the time residents spend at home as indirect influences on electricity consumption.

M. Lenzen et al. [3] has discussed that socioeconomic-demographic factors generally have similar influences on energy requirements (that is age and household size). In some household's people maintain electric cars and they spend considerable electricity on charging their cars. Speidel et al. [10] say that in households 55% of electric vehicles are charged. If a household resident uses a vehicle for long distances, then the person must charge several times which results in higher electricity bills [10].

Lenzen et al. [8] describes strategies to reduce the household electricity usage. For that he has calculated the Sydney household's electricity consumption and considered household size, age, income, and degree of urbanity, lifestyle. All the above-mentioned household characteristics vary significantly from one household to another [8].

## 2.3 Employment

Employment is an important aspect of Australian households and some of the references in this project have considered income as a predictor of electricity usage. Matthies et al. [3] examined the impact of determinants such as income, number of residents, number of adolescents on electricity usage. In today's world because of technology several household residents work from home.

If the household residents work from home household residents may be using air-conditioners or heaters during their work, so the electricity bills are higher for those households. The authors have considered direct impacts like the number of IT appliances, hours of use of IT appliances, refrigerators etc. and indirect influences like the number of residents in a house and the amount of time residents spend at home to determine electricity consumption [3].

Lenzen et al. [6] examined most of the energy in energy technology allowing for large differences across countries. Some residents in the household are away due to work and others spend the most on working and traveling to and from work, households spend less on electricity bills. Some start-up companies work in garage depending on the field (For instance, software company they must set up an environment. Also, if more employees are working in the garage then the electricity bills will be higher).

Pears et al. [1] predict the future energy services by evaluating energy usage on different aspects like household electricity usage, an industry which includes mining and agriculture, transport, lifestyle. Pears et al. [1] have discussed how the Australian household's consumption varies in different fields like households, mining, agriculture, transport, and lifestyle. Lenzen et al. [8] research is on the relationship between energy consumption, household members size, age, income, and degree of urbanity.

## 2.4 Location

Location is an important factor in analysing the Australian household's electricity consumption. Based on the location electricity bills will be higher or lower (For instance, Queensland electricity bills are higher compared to New South Wales electricity bills [2]). South Australia has the highest annual electricity bills compared to New South Wales, Victoria, and Queensland [2]. This depends on the electricity provider because in some locations electricity providers will charge higher and, in other locations, electricity providers will charge lower rates per unit [2].

Household members travel to another place due to work, school, college etc. The information confirms that most charging for electric cars is handled at home locations (55%) and business [10]. However, charging stations are used for 33% of charging events [10]. If the household residents travel for long distances in daily life, they must charge their electric car several times.

If the resident works from home they will be using IT appliances for a long time, the electricity will be higher for those households and in other households both the Mother and Father will be working during the daytime and their kids or teenagers will be going to school or university, so they spend less time at home and, the electricity bills will be lower for those households.

## 2.5 Housework

Housework activities are part of people's daily life like personal care and grooming, food preparation, cleaning the house, gardening, laundry, ironing etc. Now-a-days most of the people regularly pay someone to do housework because of spending most of their time in employment, outdoor activities etc.

## 2.6 Outdoor Tasks

Outdoor workers in a household get higher electricity bills. For instance, because of lawn moving work they must regularly charge the battery of lawnmower depends on the number of houses they do lawn mowing, which consumes more electricity usage. Lawnmower consumes minimum 1000W (which is equal to the home air conditioner) and maximum 1400W [24].

## 2.7 Unpaid work

Generally, unpaid work like volunteering, social activities etc are more likely done by teenagers because their need some experience to find a job. Several investigations have reported that when the number of adolescents in a household increases, residential electricity consumption increases as well [3].

## 2.8 HILDA Dataset

The HILDA dataset [12] is publicly available de-identified data. As such privacy and confidentiality of all survey participants will be respected. This project will use data from the HILDA survey to explore the reasons for variation in household electricity usage. The Household, Income and Labour Dynamics in Australia (HILDA) survey is a large, longitudinal survey that is the only survey of its kind that is nationally representative. This survey has followed approximately 17,000 people since 2001, with 15-time points of data released to date.

The following are the variables considered for this project from HILDA dataset.

1. Income (Continuous)
2. Paying for housework (Categorical)
3. Bathing or Dressing (Categorical)
4. Physical Activity (Categorical)
5. Outdoor Task (Continuous)
6. Unpaid work (Continuous)

## 2.9 Statistical Methods and Software

### 2.9.1 Statistical Methods

Statistical methods are used for analysing, summarizing, and interpreting data. Statistical methods are used in economics, agricultural

sciences, and life sciences. Also, they have an important role in physical sciences for the measurement of errors (such as meteorological events) and to obtain appropriate results.

#### 2.9.1.1 Regression

In statistical methods, regression [13] is used to estimate the relationship between variables. The focus is on the one or more independent variables and dependent variables. Regression analysis is used to understand how the typical values of the different variables change when the independent variables are varied.

#### 2.9.1.2 Market Segmentation

Market segmentation [14] is the process of dividing the business market or a broad consumer (which consists of potential and existing customers) into sub-groups of consumers dependent on some type of the shared characteristics. In dividing the markets researchers typically look for some common characteristics (including common interests, similar lifestyles, shared needs and similar demographic profiles). The aim of the segmentation is to identify the segments that have high yields (that means the segments that are more likely to be profitable) so these segments can be selected for special attention.

#### 2.9.1.3 Scale Development using Factor Analysis

Factor analysis [15] is used to define the variability among observed and correlated variables in terms of a potentially low number of variables that are unobserved (called factors). The variables that are observed are modeled as linear combinations of potential factors and error terms. The main aim of factor analysis is to find the independent variables that are latent.

### 2.9.2 Software

#### 2.9.2.1 RStudio for this Project

This project uses RStudio [16], it is an integrated development environment for the R programming language. Mainly RStudio is used for graphics and statistical computing.

# Chapter 3

# Theory

### 3.1 Linear Regression

Linear Regression is a fundamental and ordinarily utilized kind of predictive analysis. The idea of linear regression is to inspect two things: (1) does a set of predictor variables do a good job in predicting a result (dependent) variable? (2) Which variables specifically are significant predictors of the result variable, and how do they– demonstrated by the magnitude and sign of the beta estimates– affect the result variable? These regression estimates are utilized to clarify the relationship between one dependent variable and one or more independent variables. The simple regression equation with one dependent and one independent variable is characterized by the equation y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Naming the Variables. There are numerous names for a regression's dependant variable. It might be called a result variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous factors, predictor variables, or regressors.

Three major uses for regression analysis are (1) determining the strength of the predictors, (2) forecasting an effect, and (3) trend forecasting [18].

To start with, the regression may be utilized to recognize the strength of the effect that the independent variable(s) have on a dependent variable. Typical questions are what the strength of relationship among dose and effect is, deals and marketing spending, or age and income.

Second, it can to be utilized to forecast impacts or effect of changes. That is, by the regression analysis, we can understand how much the dependent variable changes with a change in one or more independent variables. A typical question is, "how much additional income sales will I get for each additional $1000 spent on marketing?"[18]

Third, regression analysis predicts trends and future values. The regression analysis can be utilized to get point estimates. A typical question is, "what will the cost of gold be in a half year?" [18]

There are several sorts of linear regression analysis accessible to researchers.

**Simple linear regression**

1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)

**Multiple linear regression**

1 dependent variable (interval or ratio), 2+ independent variables (interval or ratio or dichotomous)

**Logistic regression**

1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)

**Ordinal regression**

1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)

**Multinominal regression**

1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)

**Discriminant analysis**

1 dependent variable(nominal), 1+ independent variable(s) (interval or ratio)

While choosing the model for the analysis, an essential consideration is model fitting. Adding independent variables to a linear regression model will always increase the clarified variance of the model (typically expressed as $R^2$) [18]. However, overfitting can happen by adding so many variables to the model, which diminishes model generalizability. Occam's razor depicts the issue extremely well – a basic model is generally desirable over a more complex model [18]. Statistically, if a model incorporates an extensive number of variables, a portion of the variables will be statistically significant because of chance alone.

**Assumptions of Linear Regression**

Linear regression analysis will evaluate whether one or more predictor variables explain the dependent variable.

First, linear regression needs the relationship between the independent and dependent variables to be linear. It is necessary to check for outliers since linear regression is delicate to outlier impacts. The linearity supposition can best be tested with scatter plots/ box plots, the following two examples depict two cases, where no and little linearity is available [19].

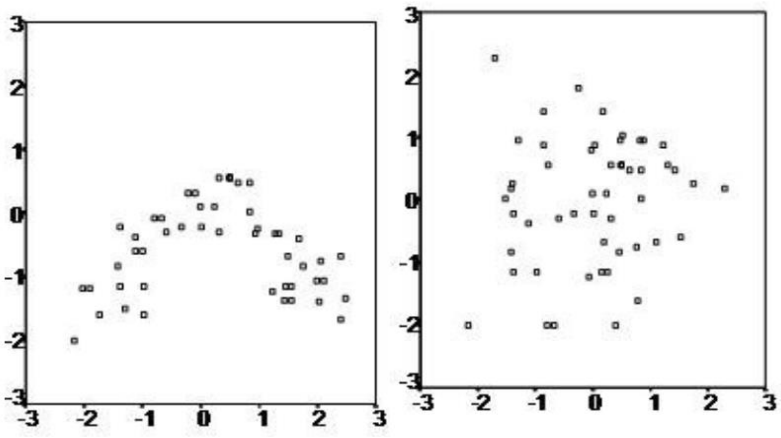



*Figure 2: plots of Linear Regression*

Also, the linear regression analysis requires all variables to be multivariate normal. This presumption can best be checked with a histogram or a Q-Q-Plot [19]. Typicality can be checked with a decency of fit test, e.g., the Kolmogorov-Smirnov test. At the point when the information isn't typically circulated a non-linear transformation (e.g., log-transformation) may settle this issue.

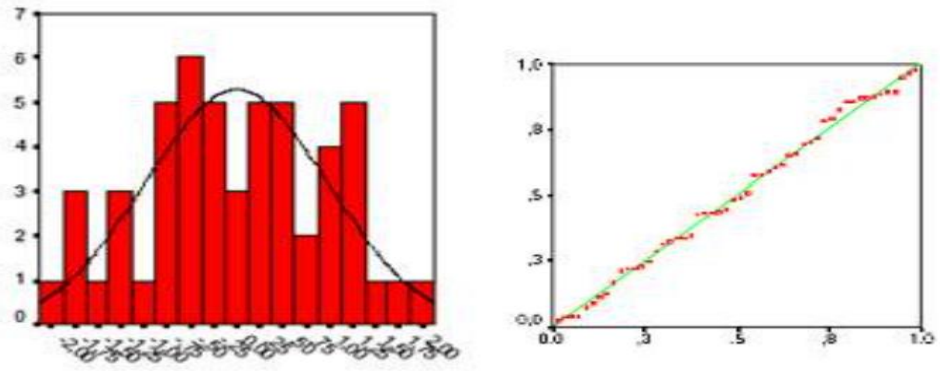


*Figure 3: Graphs of Linear Regression*

Thirdly, linear regression expects that there is little or no multicollinearity in the information. Multicollinearity happens when the independent variables are too highly associated with one another [19].

### 3.2 Factor Analysis

Exploratory Factor Analysis (EFA) is a statistical method that is utilized to distinguish the latent relational structure among bunch of variables and reduce to smaller number of variables. This basically implies that the variance of a greater number of variables can be portrayed by few variables' summary, i.e., factors. Here is an outline of exploratory factor analysis [20]:

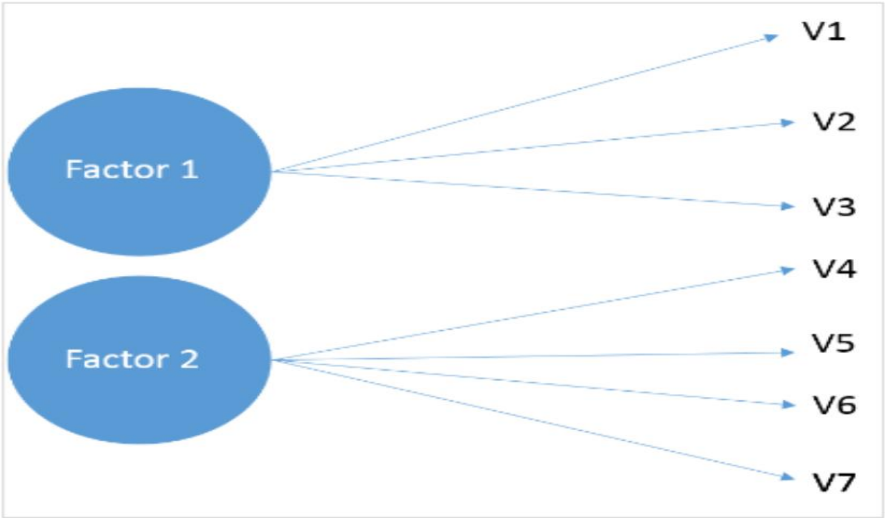



*Figure 4: Factor Analysis Diagram*

## Number of Factors

To discover the number of factors that we'll be choosing for factor analysis. This can be assessed through techniques, they are, `Parallel Analysis` and `eigenvalue`, and so forth [20].

## Parallel Analysis

We'll be utilizing `Psych` package's `fa. parallel` function to execute parallel analysis. Here we determine the data frame and factor technique (`miners` in our situation). Run the accompanying to discover acceptable number of factors and generate the `scree plot`:

parallel <-fa. parallel (data, fm = 'minres', fa = 'fa')

"Parallel investigation recommends that the number of factors = 5 and the number of components = NA"

Given underneath the `scree plot` created from the above code [20]:

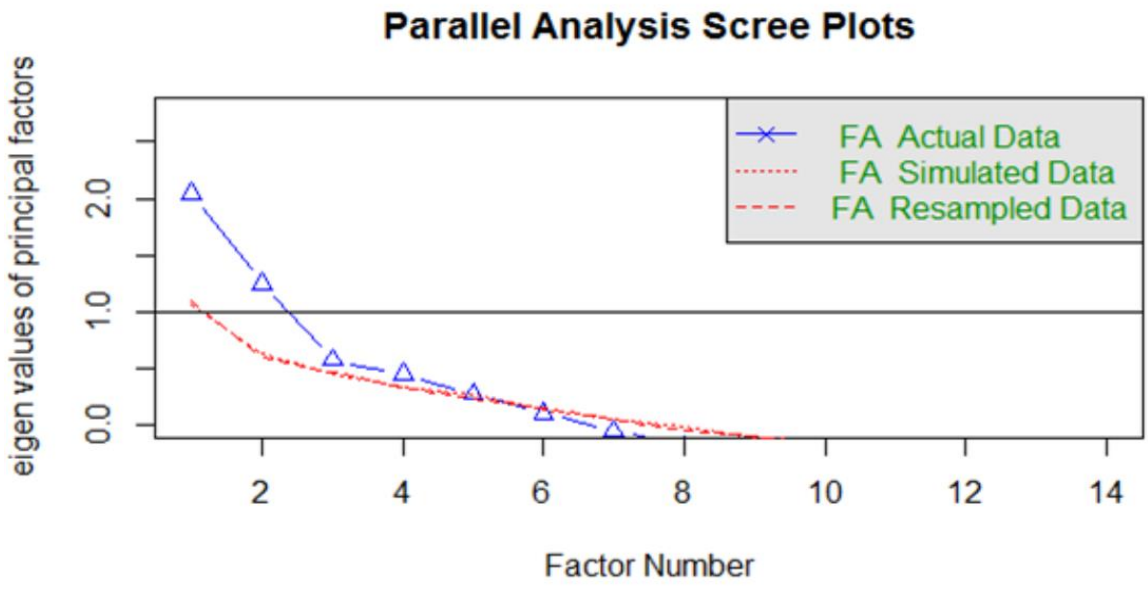



*Figure 5: Parallel Analysis Screen Plots*

The blue line indicates eigenvalues of real information and the two red lines (put over one another) demonstrate recreated and resampled information. Here we look at the huge drops in the real information and recognize the point where it levels off to the right. Likewise, we find the point of inflation – the point where the gap between simulated information and real information tends to be minimum.

The above plot and parallel analysis, it suggests anyplace between 2 to 5 factors elements would be great decision.

**Factor Analysis**

Since we've landed at possible number of factors, how about we begin off with 3 as the number of factors. With the end goal to perform factor analysis, we'll utilize `psych` package's `fa () function. Given underneath are the arguments we'll supply:

r – Raw data or correlation or covariance matrix

nfactors – Number of factors to extract

rotate – Although there are various types rotations, `Varimax` and `Oblimin` are generally prevalent

fm – One of the factor extraction strategies like `Minimum Residual (OLS)`, `Maximum Liklihood`, `Principal Axis` and so on.

For this situation, we will choose oblique rotation (rotate = "oblimin") as we trust that there is correlation in the factors. Note that Varimax rotation is utilized under the presumption that the factors are totally uncorrelated. We will utilize `Ordinary Least Squared/Minres` considering (fm = "minres"), as it is known to give results like `Maximum Likelihood` without expecting multivariate normal distribution and determines solutions through iterative eigen decomposition like principal axis.

Run the accompanying to begin the analysis:

threefactor <-fa (data, nfactors = 3, rotate = "oblimin",fm="minres")

print(threefactor)

In the following, output showing factors and loadings [20]

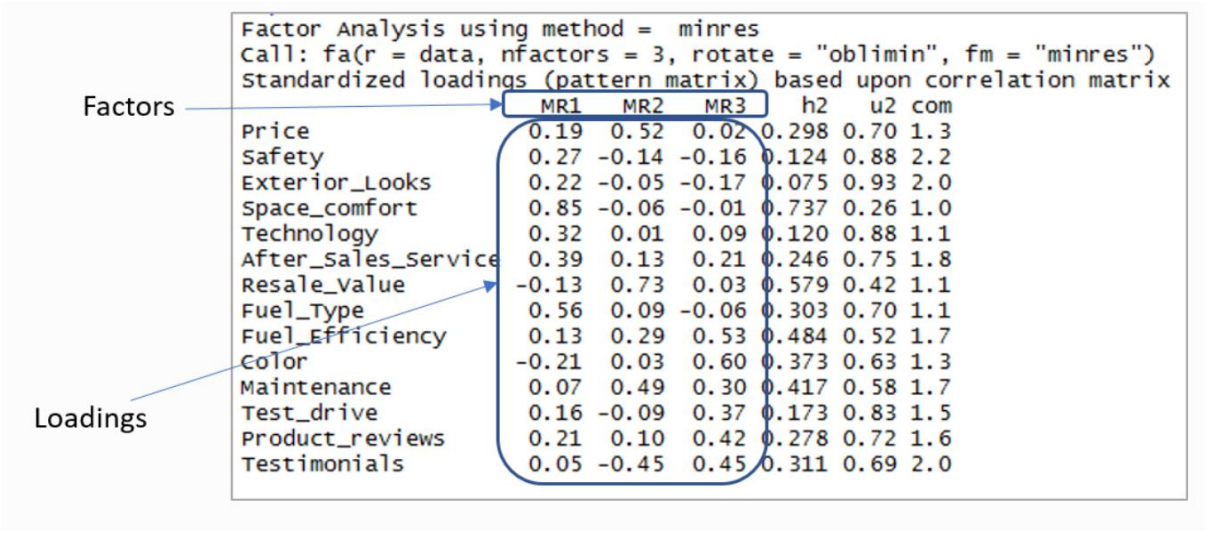



*Figure 6: Factor Loadings*

### 3.3 K-means clustering

**Introduction:**

K-means clustering is a kind of unsupervised learning, which is utilized when you have unlabelled information (i.e., information without defined categories or groups). The objective of this algorithm is to discover clusters in the information, with the number of groups represented by the variable K. The algorithm works iteratively to relegate every data point to one of K clusters in view of the highlights that are given. Data points are clustered based on the feature of similarity. The results of the K-means clustering algorithm are:

1. The centroids of the K groups, which can be utilized to label new information
2. Labels for the training data (every data point is doled out to a solitary cluster)

As opposed to defining groups before having a look at the information, grouping enables you to discover and analyse the groups that have formed naturally. The "Choosing K" section underneath depicts how the number of gatherings can be determined.

Every centroid of a cluster is a gathering of feature values which characterize the subsequent gatherings. Looking at the centroid

feature weights can be utilized to qualitatively interpret what sort of gathering each group represents.

**Algorithm:**

The K-means clustering algorithm utilizes iterative refinement to deliver the last outcome. The algorithm inputs are the number of K clusters and the data set. The data set is a collection of features for every data point. The algorithm begins with starting evaluations for the K centroids, which can either be arbitrarily created or haphazardly chosen from the data set. The algorithm works between two steps:

**Data assignment step**:

All the clusters are defined by each of its centroid. In data assignment step, all the data points are assigned to its nearest centroid, depending on the squared Euclidean distance.

**Centroid update step:**

In centroid update step, all the centroids are recomputed. This execution is done by calculating the mean of all the data points assigned to that centroids cluster.

The iteration of the algorithm between the above two steps will not stop until no data points change clusters [21]. This calculation is ensured to combine to an outcome. The outcome might be a local optimum (i.e. not really the most ideal result), implying that assessing in excess of one keep running of the calculation with randomized beginning centroids may give a superior result.

**Choosing K**

The algorithm portrayed above finds the clusters and data set labels for a specific pre-picked K. To locate the number of clusters in the information, the user needs to run the K-means clustering algorithm for a scope of K values and think about the outcomes. When all is said in done, there is no strategy for deciding definite estimation of K, yet a precise estimate can be acquired utilizing the accompanying methods.

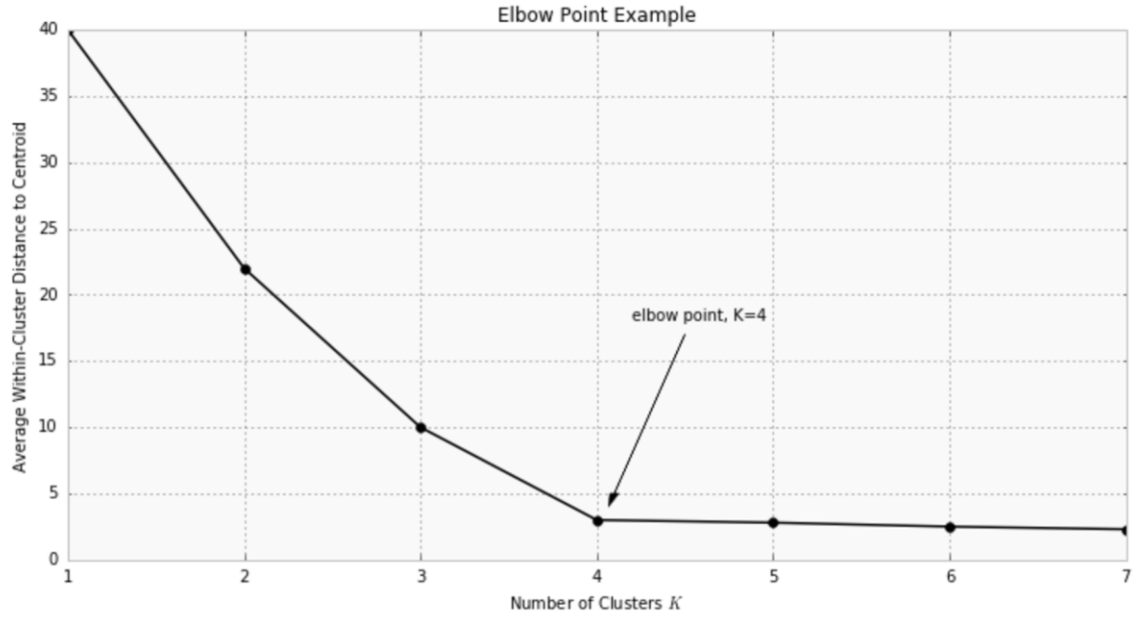



*Figure 7: Elbow Point Example*

One of the measurements that are ordinarily used to look at results crosswise over various estimations of K is the mean separation between information focuses and their group centroid. Since expanding the number of groups will dependably lessen the separation to information focuses, expanding K will dependably diminish this metric, to the extraordinary of achieving zero when K is the same as the quantity of information focuses. Along these lines, this metric can't be utilized as the sole target. Rather, mean separation to the centroid as a component of K is plotted and the "elbow point," where the rate of decrease sharply moves, can be utilized to generally decide K.

Various procedures exist for validating K, including cross-validation, data criteria, the data theoretic jump strategy, the silhouette technique, and the G-means algorithm. Furthermore, checking the conveyance of information focuses crosswise over groups gives knowledge into how the algorithm is splitting the information for every K [21].

### 3.4 Linkage Algorithm

Linkage Measure: The goal of the linkage measure is to merge the clusters that are near to each other. However, there are three types [22].

Single-Link, Complete-link, and Average-Link Clustering

**Single Linkage:**

Single linkage [22] defines the distance between 2 clusters as the minimum distance found between one case from the first cluster and one case from the second cluster. The problem with the single linkage clustering is It will only consider the nearest minimum distance between the one case of two different clusters and merge them which does not give the accurate results.

**Complete Linkage:**

Complete linkage [22] will search for the maximum distance between the cases from one cluster to another. This linkage will solve the chaining problem, but it creates another problem. If all the cases are near in 2 clusters and only one case is far from the other cases, then it will no longer merge 2 clusters.

**Average Linkage:**

To overcome the problems of single and complete linkage. The average linkage [22] will consider the average of the distance between the cases and then it will decide whether to merge the clusters or not. It will provide more accurate results.

# Chapter 4

# Methodology and Results

## 4.1 Goals and Objectives

The goal of this project is to find the type of activities in Australian households that lead to higher electricity bills. To evaluate the chosen variables from HILDA dataset, I have used statistical methods those are linear regression, factor analysis, k-means clustering and linkage algorithm in R programming language with RStudio as Integrated development environment to predict the type of variables that lead to higher electricity bills. In the following, I will provide the understanding of all the methodologies that I have used in this project.

### Linear Regression

Linear regression will compare the similarity between variable to variables from the HILDA dataset to determine which variable is more positively co-related to electricity usage. From the results of linear regression, we must focus on estimate and p value. If the estimate is positive, then the variable will more likely to increase the electricity usage. If the estimate is negative, then the variable will not affect the electricity consumption. In linear regression, if the P value is less than/equal to 0.05 then the variable is related to electricity usage. If the p value is greater than 0.05 then the variable is not related to electricity usage.

### Factor Analysis

In simple terms, factor analysis will combine the questions that belong to same category and gives unique value. Basically, factor analysis will analyse the given variables to determine the relationship between them. In this project, the goal of factor analysis is to relate the variables that are like each other by factor loadings. Each factor loading will talk about the variable relationship by producing values for each variable. In one factor loadings, if the variable loadings are high and positive then those variables are inter-related to each other. The objective of factor analysis in this project is to which variables are correlated to each other.

### K-Means Clustering

K-means clustering is used for data without defined categories or groups. The aim of k-means clustering is to find a group in the data where the number of clusters are represented by the variable k. Depending on k-size it will cluster all the data into k clusters. In this project, k-means will allocate all the data points into defined k-clusters by calculating the centroid of each data point. The goal of k-means clustering is to define the variable relationship by calculating the mean of each variable in each cluster to determine which variables are correlated in each cluster by observing the variables mean, if one variable mean is almost equal to other variables mean in a cluster then those variables are correlated with each other.

### Linkage Algorithm

In this project, I have used single linkage, complete linkage, average linkage to determine the relationship between the chosen variables from HILDA dataset. Single linkage measure will group the clusters by calculating the minimum possible distance between the data points. Complete linkage measure will create cluster by calculating the maximum possible distance between the data points. Average linkage measure will create cluster by calculating the average of the distance between the data points.

## 4.2 Initial Screening of Data

### 1. Income (Continuous)

| | |
|---|---|
| **Variable** | owsfga |
| **Label** | F32 Gross financial year wages and salaries ($) [weighted topcode] |
| **Form** | CPQ/NPQ |
| **Question No.** | F32 |
| **Questionnaire Text** | Last financial year, what was your total wage and salary income from all jobs before tax or anything else was deducted? |
| **Population** | Receives wage and salary earnings last financial year |
| **Subject Area** | INCOME - Wages and Salaries |
| **Survey Wave** | 15 |
| **Data File** | Responding Person File |
| **Frequency** | RP / Mean / 58,931<br>Std Dev / 50,666<br>N Obs / 10,466 |
| **Notes** | To preserve the weighted mean, the cases which exceed the threshold for the top-coded variable have a substituted value which is the weighted average value of all cases exceeding the threshold. This is always a value greater than the threshold. See the HILDA User Manual and HILDA Technical Paper 1/13 for details on the derivation of this variable. |
| **Variable Occurrence** | W1 W2 W3 W4 W5 W6 W7 W8 W9 W10 W11 W12 W13 W14 W15 W16 |

| Missing data values | |
|---|---|
| -1 | Not asked |
| -2 | Not applicable |
| -3 | Dont know |
| -4 | Refused/Not stated |
| -5 | Multiple response SCQ |
| -6 | Implausible value |
| -7 | Not able to be determined |
| -8 | No SCQ |
| -9 | Non-responding household |
| -10 | Non-responding person |

*Figure 8: Income Online Data Dictionary*

Income is a continuous variable because the income per person is continuous. I have chosen this variable because income or employment reducing electricity consumption could lead to a fall in income and/or employment [23]. For accuracy of data all the negative values I have replaced with NA. To check the data distribution for this variable I have taken histogram (which is skewed) and scatter plot. So, to make it normal distribution I have applied log transform for this variable value. To understand the relation between electricity bills and income I did linear regression analysis. Also, to understand the relationship between income and other chosen variables I have applied factor analysis, k-means and linkage algorithm.

## 2. Paying for housework (Categorical)

| | |
|---|---|
| **Variable** | olspayhw |
| **Label** | SCQ:B21 Regularly pay someone to do housework |
| **Form** | SCQ |
| **Question No.** | SCQ B21 |
| **Questionnaire Text** | Does your household regularly pay someone to do any of the housework (cleaning, washing, ironing, cooking, etc)? |
| **Population** | All |
| **Subject Area** | HEALTH - Lifestyle |
| **Survey Wave** | 15 |
| **Data File** | Responding Person File |
| **Frequency** | olspayhw / RP: [-4] Refused/Not stated 104; [-5] Multiple response SCQ 1; [-8] No SCQ 2095; [1] Yes 1817; [2] No 13589 |
| **Notes** | |
| **Variable Occurrence** | W5 W8 W11 W15 |

| Value | Label |
|---|---|
| 1 | Yes |
| 2 | No |

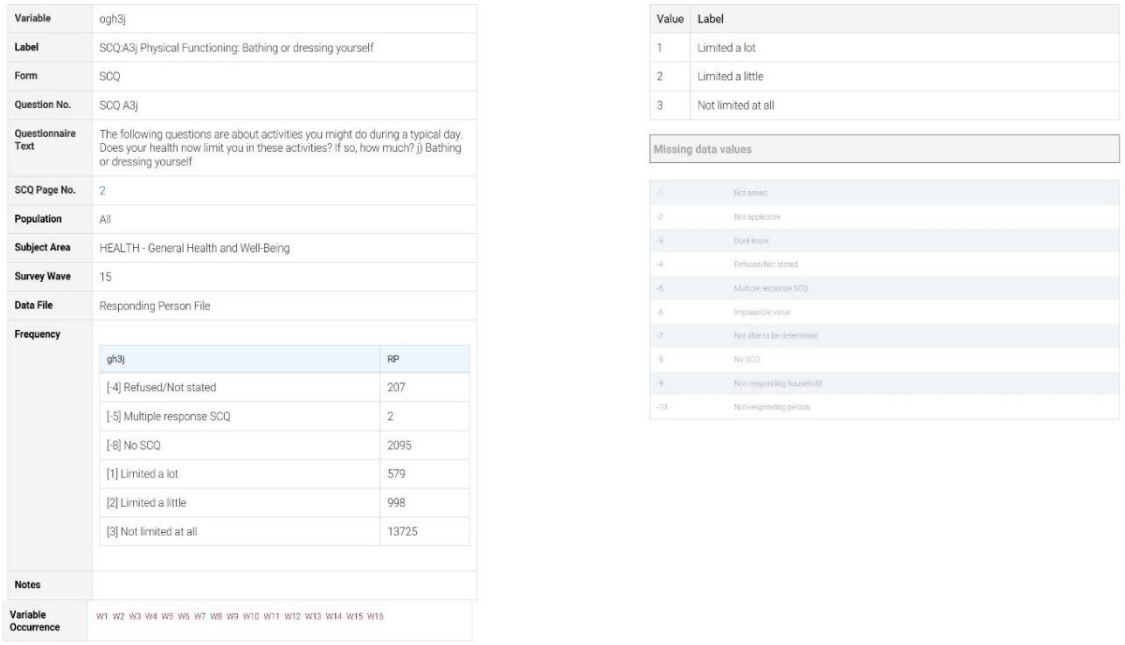
Missing data values

| | |
|---|---|
| -1 | Not asked |
| -2 | Not applicable |
| -3 | Dont know |
| -4 | Refused/Not stated |
| -5 | Multiple response SCQ |
| -6 | Implausible value |
| -7 | Not able to be determined |
| -8 | No SCQ |
| -9 | Non-responding household |
| -10 | Non-responding person |

*Figure 9: Paying for Housework Online Data Dictionary*

Paying for housework is a categorical variable because as you can see in the above image it has only two categories, for accuracy of data all the negative values are replaced with NA. To check the data distribution for this variable I have taken histogram and box plot. For categorical variables log transform is not required. To understand the relation between electricity bills and paying for housework I did linear regression analysis. Also, to understand the relationship between paying for housework and other variables I have applied factor analysis, k-means and linkage algorithm.

## 3. Bathing or Dressing (Categorical)



| | |
|---|---|
| Variable | ogh3j |
| Label | SCQ:A3j Physical Functioning: Bathing or dressing yourself |
| Form | SCQ |
| Question No. | SCQ A3j |
| Questionnaire Text | The following questions are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much? j) Bathing or dressing yourself |
| SCQ Page No. | 2 |
| Population | All |
| Subject Area | HEALTH - General Health and Well-Being |
| Survey Wave | 15 |
| Data File | Responding Person File |

Frequency

| gh3j | RP |
|---|---|
| [-4] Refused/Not stated | 207 |
| [-5] Multiple response SCQ | 2 |
| [-8] No SCQ | 2095 |
| [1] Limited a lot | 579 |
| [2] Limited a little | 998 |
| [3] Not limited at all | 13725 |

| | |
|---|---|
| Notes | |
| Variable Occurrence | W1 W2 W3 W4 W5 W6 W7 W8 W9 W10 W11 W12 W13 W14 W15 W16 |

| Value | Label |
|---|---|
| 1 | Limited a lot |
| 2 | Limited a little |
| 3 | Not limited at all |

Missing data values

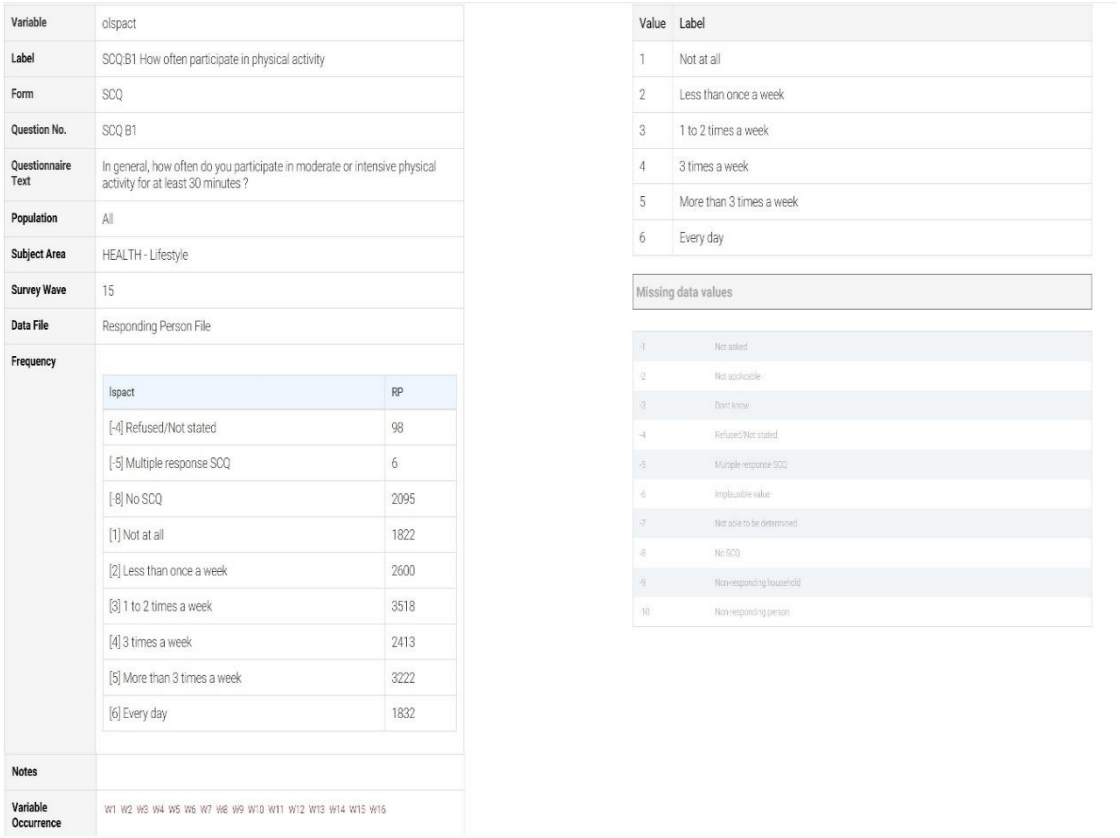
| | |
|---|---|
| -1 | Not asked |
| -2 | Not applicable |
| -3 | Don't know |
| -4 | Refused/Not stated |
| -5 | Multiple response SCQ |
| -6 | Implausible value |
| -7 | Not able to be determined |
| -8 | No SCQ |
| -9 | Non-responding household |
| -10 | Non-responding person |

*Figure 10: Bathing or Dressing Online Data Dictionary*

Bathing or Dressing is a categorical variable because as you can see in the above image it has only three categories, for accuracy of data all the negative values are replaced with NA. To check the data distribution for this variable I have taken histogram and box plot. For categorical variables log transform is not required. To understand the relation between electricity bills and bathing or dressing I did linear

regression analysis. Also, to understand the relationship between bathing or dressing and other variables I have applied factor analysis, k-means and linkage algorithm.

## 4. Physical Activity (Categorical)



| Variable | olspact |
|---|---|
| Label | SCQ:B1 How often participate in physical activity |
| Form | SCQ |
| Question No. | SCQ B1 |
| Questionnaire Text | In general, how often do you participate in moderate or intensive physical activity for at least 30 minutes ? |
| Population | All |
| Subject Area | HEALTH - Lifestyle |
| Survey Wave | 15 |
| Data File | Responding Person File |

Frequency

| lspact | RP |
|---|---|
| [-4] Refused/Not stated | 98 |
| [-5] Multiple response SCQ | 6 |
| [-8] No SCQ | 2095 |
| [1] Not at all | 1822 |
| [2] Less than once a week | 2600 |
| [3] 1 to 2 times a week | 3518 |
| [4] 3 times a week | 2413 |
| [5] More than 3 times a week | 3222 |
| [6] Every day | 1832 |

| Notes | |
|---|---|
| Variable Occurrence | W1 W2 W3 W4 W5 W6 W7 W8 W9 W10 W11 W12 W13 W14 W15 W16 |

| Value | Label |
|---|---|
| 1 | Not at all |
| 2 | Less than once a week |
| 3 | 1 to 2 times a week |
| 4 | 3 times a week |
| 5 | More than 3 times a week |
| 6 | Every day |

Missing data values

| | |
|---|---|
| -1 | Not asked |
| -2 | Not applicable |
| -3 | Don't know |
| -4 | Refused/Not stated |
| -5 | Multiple response SCQ |
| -6 | Implausible value |
| -7 | Not able to be determined |
| -8 | No SCQ |
| -9 | Non-responding household |
| -10 | Non-responding person |

*Figure 11: Physical Activity Online Data Dictionary*

Physical activity is a categorical variable because as you can see in the above image it has only six categories, for accuracy of data all the negative values are replaced with NA. To check the data distribution for this variable I have taken histogram and box plot. For categorical variables log transform is not required. To understand the relation between electricity bills and physical activity I did linear regression analysis. Also, to understand the relationship between physical activity and other variables I have applied factor analysis, k-means and linkage algorithm.

## 5. Outdoor Tasks (Continuous)

| | |
|---|---|
| Variable | olshrod |
| Label | SCQ:B19e Hours per week - Outdoor tasks |
| Form | SCQ |
| Question No. | SCQ B19e |
| Questionnaire Text | How much time would you spend on each of the following activities in a typical week? e) Outdoor tasks, including home maintenance (repairs, improvements, painting etc.), car maintenance or repairs and gardening |
| Population | All |
| Subject Area | HEALTH - Lifestyle |
| Survey Wave | 15 |
| Data File | Responding Person File |
| Frequency | RP: Mean 4; Std Dev 6; N Obs 14,890 |
| Notes | |
| Variable Occurrence | W1 W2 W3 W4 W5 W6 W7 W8 W9 W10 W11 W12 W13 W14 W15 W16 |

| Missing data values | |
|---|---|
| -1 | Not asked |
| -2 | Not applicable |
| -3 | Dont know |
| -4 | Refused/Not stated |
| -5 | Multiple response SCQ |
| -6 | Implausible value |
| -7 | Not able to be determined |
| -8 | No SCQ |
| -9 | Non-responding household |
| -10 | Non-responding person |

*Figure 12: Outdoor Tasks Hours Online Data Dictionary*

| | |
|---|---|
| Variable | olsmnod |
| Label | SCQ:B19e Minutes per week - Outdoor tasks |
| Form | SCQ |
| Question No. | SCQ B19e |
| Questionnaire Text | How much time would you spend on each of the following activities in a typical week? e) Outdoor tasks, including home maintenance (repairs, improvements, painting etc.), car maintenance or repairs and gardening |
| Population | All |
| Subject Area | HEALTH - Lifestyle |
| Survey Wave | 15 |
| Data File | Responding Person File |
| Frequency | RP: Mean 1; Std Dev 7; N Obs 14,890 |
| Notes | |
| Variable Occurrence | W2 W3 W4 W5 W6 W7 W8 W9 W10 W11 W12 W13 W14 W15 W16 |

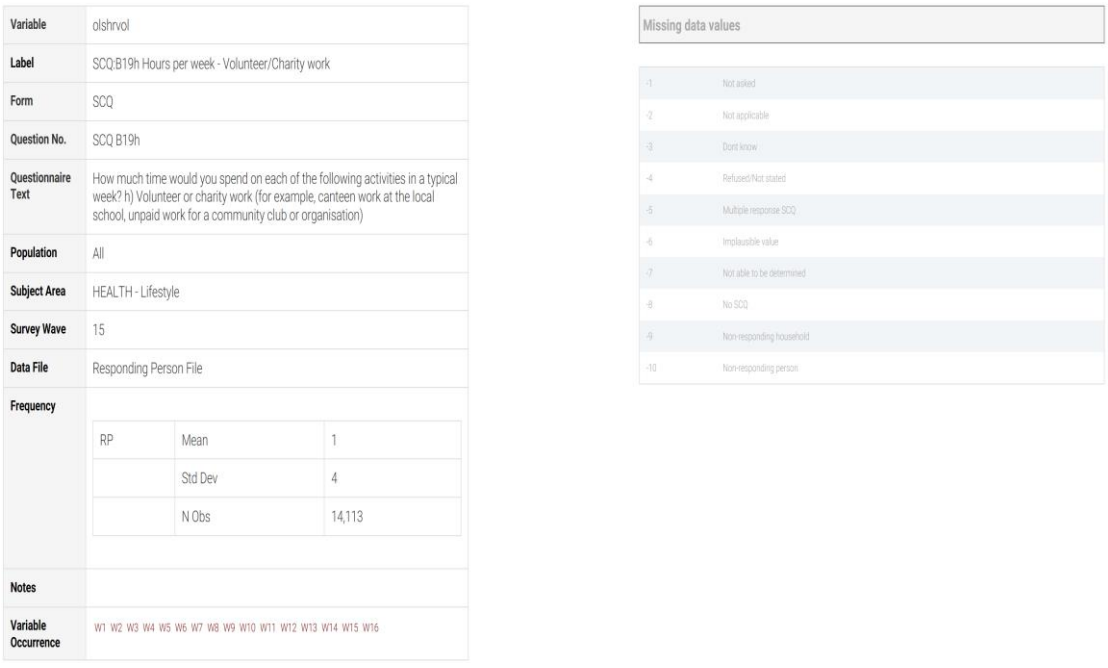
| Missing data values | |
|---|---|
| -1 | Not asked |
| -2 | Not applicable |
| -3 | Dont know |
| -4 | Refused/Not stated |
| -5 | Multiple response SCQ |
| -6 | Implausible value |
| -7 | Not able to be determined |
| -8 | No SCQ |
| -9 | Non-responding household |
| -10 | Non-responding person |

*Figure 13: Outdoor Tasks Minutes Online Data Dictionary*

Outdoor tasks are a continuous variable because hours/minutes per week on outdoor tasks per person is continuous. As you can see there are two variables for outdoor tasks that is olshrod (Hours per week on outdoor tasks)

and olsmnod (Minutes per week on outdoor tasks). So, I have combined both the variables into outdoor tasks by converting time into minutes. I have chosen this variable because some of the outdoor tasks are related to electricity usage for instance lawn moving with electric lawn mower. For accuracy of data all the negative values are replaced with NA. To check the data distribution for this variable I have taken histogram (which is skewed) and scatter plot. So, to make it normal distribution I have applied log transform for this variable value. To understand the relation between electricity bills and outdoor tasks I did linear regression analysis. Also, to understand the relationship between outdoor tasks and other chosen variables I have applied factor analysis, k-means and linkage algorithm.

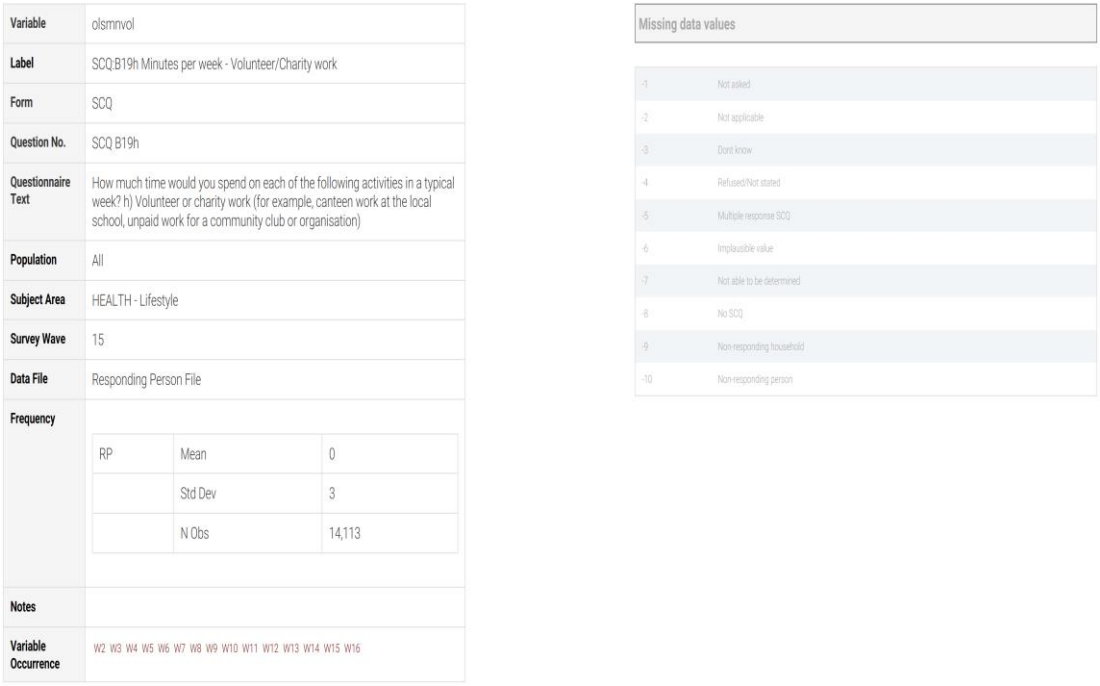
## 6. Unpaid Work (Continuous)



| | |
|---|---|
| **Variable** | olshrvol |
| **Label** | SCQ:B19h Hours per week - Volunteer/Charity work |
| **Form** | SCQ |
| **Question No.** | SCQ B19h |
| **Questionnaire Text** | How much time would you spend on each of the following activities in a typical week? h) Volunteer or charity work (for example, canteen work at the local school, unpaid work for a community club or organisation) |
| **Population** | All |
| **Subject Area** | HEALTH - Lifestyle |
| **Survey Wave** | 15 |
| **Data File** | Responding Person File |
| **Frequency** | RP Mean 1<br>Std Dev 4<br>N Obs 14,113 |
| **Notes** | |
| **Variable Occurrence** | W1 W2 W3 W4 W5 W6 W7 W8 W9 W10 W11 W12 W13 W14 W15 W16 |



*Figure 14: Unpaid Work Hours Online Data Dictionary*



| | |
|---|---|
| Variable | olsmnvol |
| Label | SCQ:B19h Minutes per week - Volunteer/Charity work |
| Form | SCQ |
| Question No. | SCQ B19h |
| Questionnaire Text | How much time would you spend on each of the following activities in a typical week? h) Volunteer or charity work (for example, canteen work at the local school, unpaid work for a community club or organisation) |
| Population | All |
| Subject Area | HEALTH - Lifestyle |
| Survey Wave | 15 |
| Data File | Responding Person File |
| Frequency | RP: Mean 0; Std Dev 3; N Obs 14,113 |
| Notes | |
| Variable Occurrence | W2 W3 W4 W5 W6 W7 W8 W9 W10 W11 W12 W13 W14 W15 W16 |



*Figure 15: Unpaid Work Minutes Online Data Dictionary*

Unpaid work is a continuous variable because hours/minutes per week on unpaid work per person is continuous. As you can see there are two variables for unpaid work that is olshrvol (Hours per week on unpaid work) and olsmnvol (Minutes per week on unpaid work). So, I have combined both the variables into unpaid work by converting time into minutes. I have chosen this variable because mostly unpaid work that is volunteer or charity work is done by teenagers to get an experience and built network for finding paid jobs. Several investigations have reported that when the number of adolescents in a household increases, residential electricity consumption increases as well [3]. For accuracy of data all the negative values are replaced with NA. To check the data distribution for this variable I have taken histogram (which is skewed) and scatter plot. So, to make it normal distribution I have applied log transform for this variable value. To understand the relation between electricity bills and unpaid work I did linear regression analysis. Also, to understand the relationship between unpaid work and other chosen variables I have applied factor analysis, k-means and linkage algorithm.

## 4.3 Linear Regression Results

| Question Number | Description | Page Number | Variable Name | Label | Type | P Value |
|---|---|---|---|---|---|---|
| NPQ A10 | Are you currently enrolled in a course of study for a trade certificate, diploma, degree or any other educational qualification? | 244 | oedcqen | NPQ: A10 Currently enrolled in a course | Categorical | P=0.0736 |
| F1b | CONFIRM: Do you currently receive income from wages or salary? | 289 | owschave | F1b Currently receive income from wages/salary | Categorical | P<2e-16 |
| NPQ H1 | Looking at SHOWCARD H4, which of these best describes your current marital status? And by "married" we mean in a registered marriage. | 362 | omrcms | CPQ:H4/NPQ:H1 Current marital status | Categorical | P<2e-16 |
| K1 | Looking at SHOWCARD K1, do you have any long-term health condition, impairment or disability (such as these) that restricts you in your everyday activities, and has lasted or is likely to last, for 6 months or more? | 383 | Ohelth | K1 Long term health condition | Categorical | P=5.8e-12 |

| | | | | | | |
|---|---|---|---|---|---|---|
| DV: C11 | What kind of work do you do in this job? That is, what is your occupation called and what are the main tasks and duties you undertake in this job? Please describe fully. | 257 | Ojbmo62 | DV: C11 Occupation 2-digit ANZSCO 2006 | Categorical | P=3.4e-14 |
| F3 | For your [IF F2=1: main job / ELSE job] what was the total gross amount of your most recent pay before tax or anything else was taken out? It will help to answer this question if you can refer to your last pay-slip. | 289 | Owscmga | F3 Total gross amount of most recent pay before deductions | Continuous | P<2e-16 |
| F32 | Last financial year, what was your total wage and salary income from all jobs before tax or anything else was deducted? | 304 | owsfga | F32 Gross financial year wages and salaries ($) [weighted topcode] | Continuous | P<2e-16(Selected) |
| SCQ B7 | Are you currently an active member of a sporting, hobby or community-based club or association? | 5 | olsclub | SCQ: B7 Currently an active member of a sporting/hobby/community-based club or association | Categorical | P=0.0045 |
| SCQ B19e | How much time would you spend on each of the following activities in a typical week? e) Outdoor tasks, including home maintenance | 10 | olshrod | SCQ: B19e Hours per week - Outdoor tasks | Continuous | P=5.79e-15 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | (repairs, improvements, painting etc.), car maintenance or repairs and gardening | | | | | |
| SCQ B19h | How much time would you spend on each of the following activities in a typical week? h) Volunteer or charity work (for example, canteen work at the local school, unpaid work for a community club or organisation) | 10 | olshrvol | SCQ: B19h Hours per week - Volunteer/Charity work | Continuous | P=0.0192 |
| SCQ B21 | Does your household regularly pay someone to do any of the housework (cleaning, washing, ironing, cooking, etc)? | 11 | olspayhw | SCQ: B21 Regularly pay someone to do housework | Categorical | P=1e-08(Selected) |
| SCQ E1 | Are you currently in paid work? | 17 | Ojopw | SCQ: E1 Is in paid work | Categorical | P<2e-16 |
| SCQ A3b | b) Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling or playing golf | 2 | Ogh3b | SCQ: A3b Physical Functioning: Moderate activities | Categorical | P=3.53e-13 |
| SCQ A3c | c) Lifting or carrying groceries | 2 | Ogh3c | SCQ: A3c Physical Functioning: Lifting or carrying groceries | Categorical | P<2e-16 |
| SCQ A3j | Bathing or Dressing | 2 | Ogh3j | | Categorical | P=3.17e-16(Selected) |
| SCQ A1 | In general, would you say your health is: | 2 | Ogh1 | SCQ: A1 Self- | Continuous | P=6.92e-06 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | assessed health | | |
| SCQ B1 | In general, how often do you participate in moderate or intensive physical activity for at least 30 minutes? | 5 | olspact | SCQ: B1 How often participate in physical activity | Categorical | P=0.00881 (Selected) |
| SCQ B10 | How often do you feel rushed or pressed for time? | 6 | olsrush | SCQ: B10 How often feel rushed or pressed for time | Continuous | P<2e-16 |
| SCQ B11 | How often do you feel you have spare time that you don't know what to do with? | 6 | Olsstime | SCQ: B11 Spare time that dont know what to do with | Continuous | P<2e-16 |
| SCQ B16 | In general, about how often do you get together socially with friends or relatives not living with you? | 8 | Olssocal | SCQ: B16 How often get together socially with friends/relatives not living with you | Continuous | P=0.0126 |
| SCQ B19e | How much time would you spend on each of the following activities in a typical week? e) Outdoor tasks, including home maintenance (repairs, improvements, painting etc.), car maintenance or repairs and gardening | 10 | Olsmnod | SCQ: B19e Minutes per week - Outdoor tasks | Continuous | P=0.000341 |
| SCQ B19 e Combine | Outdoor Tasks | | Olsmnod & the olshrod | | Continuous | P<2e-16(Selected) |
| SCQ B19h Combine | Unpaid work | | Olsmnvol & olshrvol | | Continuous | P=6.31e-05(Selected) |

| | | | | | | |
|---|---|---|---|---|---|---|
| SCQ C1 | Given your current needs and financial responsibilities, would you say that you and your family are: | 11 | ofiprosp | SCQ:C1 Prosperity given current needs and financial responsibilities | Continuous | P=1.7e-08 |

*Table 1: Linear Regression Results*

The above table contains all the variable details from HILDA 2015 questionnaires and its linear regression results. The above table contains all the selected variables after linear regression analysis based on the variables p value<0.1 but the total variables that I have chosen for conducting linear regression on a one-time basis is in the appendix.

Firstly, I have chosen the variables from 2015 HILDA questionnaires that are more likely related to electricity usage in general. After choosing the variables then I have determined whether it is continuous/categorical by nalysing HILDA online data dictionary. To better understand the data distribution of a variable I had look at histograms, scatter plots for continuous variables and box plots for categorical variables. Then I had replaced all the negative values with NA, applied log transform for continuous variables and combined some categories in categorical variable to get better results. After doing this I can produce better histograms, scatter plot for continuous variable and box plot for categorical variable.

In the linear regression analysis, I have compared all the chosen variables with electricity bills on a one-time basis. In this analysis I have selected all the variables whose p value<0.1. In the below table, all the variables are selected after conducting linear regression analysis based on p value. The lesser the p value the independent variable is more likely related to dependent variable.

First version of linear regression analysis

### 4.3 Final Linear Regression Analysis

Linear regression is used in this project to compare the similarity between electricity bills and the variables from the 2015 HILDA dataset to determine which Australian households spend the most on electricity. In this project, I have chosen some variables from the HILDA dataset by assuming the selected variables are related

to electricity usage according to my knowledge. I have conducted a linear regression analysis on a one-time basis and multivariable analysis. First, I did regression analysis on a one-time basis on how likely the variable is related to electricity usage by considering p value<=0.1. From the regression analysis on a one-time basis 30 variables are selected (P<=0.1) out of 34 variables. Then I did multivariable regression analysis by adding all the selected variables to compare the similarity with electricity bills. In the result it shows all the variables p-value then I have removed the variable from the list which has highest p-value and then did linear regression analysis. Again, in the result it will display all the variables p-value, I will eliminate another variable from the regression analysis that has highest p-value. I repeated the above procedure until I get all the variables p-value <=0.05 in the multivariable regression analysis. So, I am left with 6 variables out of 30 variables.

| **Variable Name** | **Estimate** | **P Value** | **Type** |
|---|---|---|---|
| Income | 0.020 | <0.001 | Continuous |
| Paying for housework | -0.207 | <0.001 | Categorical |
| Bathing or Dressing | -0.193 | <0.001 | Categorical |
| Physical activity | -0.014 | 0.031 | Categorical |
| Outdoor tasks | 0.052 | <0.001 | Continuous |
| Unpaid work | 0.015 | 0.001 | Continuous |

*Table 2: Results of Final Linear Regression*

Income, outdoor tasks, and unpaid work are positive estimates of electricity bills. Income or employment reducing electricity consumption could lead to a fall in income and/or employment [23]. So, increase in financial year wages and salaries increases electricity consumption. Generally, unpaid work like volunteering,

social activities etc are more likely done by teenagers because their need some experience to find a job. Several investigations have reported that when the number of adolescents in a household increases, residential electricity consumption increases as well [3]. Outdoor workers in a household get higher electricity bills. For instance, because of lawn moving work they must regularly charge the battery of lawnmower depends on the number of houses they do lawn mowing, which consumes more electricity usage. Lawnmower consumes minimum 1000W (which is equal to the home air conditioner) and maximum 1400W [24]**.** So, increase in income, spending more time on outdoor tasks and unpaid work will increase electricity bills. Whereas, physical activity, bathing or dressing and paying for housework are negative estimates of electricity usage that means an increase in physical activity, bathing or dressing and paying for housework doesn't affect electricity bills.

For the accuracy, I have considered 2014 HILDA data and conducted the above process it checks the accuracy of 2015 HILDA data. However, there is no paying for housework variable in 2014. To know the accuracy of the data I did multiple regression on variables whose $p<0.05$ with and without paying for housework in 2015 and the result of 2015 data is like 2014 data.

## 4.5 Factor Analysis

Basically, factor analysis will analyze the given variables to determine the relationship between them. In this project, the goal of factor analysis is to relate the variables that are like each other by factor loadings. Each factor loading will talk about the variable relationship by producing values for each variable. In one factor loadings, if the variable loadings are high and positive then those variables are inter-related to each other.

Firstly, I have selected some variables from 2015 HILDA dataset that is related to the electricity usage according to my assumption. Then I did a regression analysis for each selected variable on a one-time basis with the variable named oxputila (Annual household expenditure - Electricity bills, gas bills, and other heating fuel - Amount ($)). After the regression analysis on a one-time basis, I have considered some variables whose p value<0.1. Later, I have conducted multiple regression analysis with all the variables ($p<0.1$) at a time to oxputila (Annual household expenditure - Electricity bills, gas bills and other heating fuel - Amount ($)). In the above process, I

have eliminated one variable at a time whose "p" value is higher compare to other variables in this process. I have repeated this process until I get all the variables p value < 0.05.

I have binded all the selected six variables (p<0.05) in m and conducted factor analysis by using the command fa (m, nfactors=6, rotate=" none"). The aim of the factor analysis is it will combine the variables that belongs to similar category and produces unique value below is the table that describes about the factor analysis, from the below table, I will explain about the interpretation of factor analysis from MR1 to MR3.

| Variable Name | MR1 (Factor Loading 1) | MR2 (Factor Loading 2) | MR3 (Factor Loading 3) |
|---|---|---|---|
| Income | 0.08 | -0.40 | 0.23 |
| Paying for housework | 0.62 | 0.14 | 0.16 |
| Bathing or Dressing | -0.23 | 0.47 | 0.10 |
| Physical activity | 0.18 | -0.29 | -0.19 |
| Outdoor tasks | 0.71 | 0.10 | 0.03 |
| Unpaid work | 0.27 | 0.12 | -0.31 |

*Table 3: Factor Analysis*

If we observe MR1 there are two higher values that is paying for housework=0.62 and outdoor tasks= 0.71. When a variable that has higher value in the factor loadings that means, the factor loading is talking about that variable. Here the MR1 factor loading is talking about two variables that is paying for housework and outdoor tasks. So MR1 factor loading says that people who often pay some for

housework will spend most of their time on outdoor tasks perhaps they don't have time to do housework.

MR2 talks about the income=-0.40 and bathing or dressing=0.37 that is people who earn money through employment/business will do bathing or dressing by themselves. MR3 is related to unpaid work=-0.31 that means in factor loading MR3 talks about unpaid work and it is not related to any other variable.

### Multi linear regression for Factor loadings

Multi linear regression between factor scores and electricity bills. I have tested all the factor loadings by checking similarity against electricity bills. Except factor loadings 3, remaining all the factor loadings are related to electricity usage. Factor loadings 2,6 estimate is negative that means they do not affect electricity bills. Whereas, factor loadings 1,4,5 will affect electricity usage because their estimates are positive.

| Factor Scores | Estimate | P Value |
|---|---|---|
| Factor loadings 1 | 9.151e-02 | < 0.001 |
| Factor loadings 2 | -1.532e-01 | < 0.001 |
| Factor loadings 3 | -4.390e-03 | 0.842 |
| Factor loadings 4 | 9.897e-02 | 0.004 |
| Factor loadings 5 | 5.401e-01 | < 0.001 |
| Factor loadings 6 | -2.623e+13 | < 0.001 |

*Table 4: Linear regression of Factor Loadings*

## 4.6 K-Means Clustering

In this project, k-means will allocate all the data points into defined

k-clusters by calculating the nearest centroid of each data point. The goal of k-means clustering is to define the variable relationship by calculating the mean of each variable in each cluster to determine which variables are correlated in each cluster by observing the variables mean, if one variable mean is almost equal to other variables mean in a cluster then those variables are correlated with each other. The hard part in 'k' means is defining k value that is we must define the 'k' value before clustering.

While doing k-means clustering we have to remove the missing values in the variables else k-means doesn't work. To overcome the missing values either we must delete that missing value row or replace it with mean of the variable. In this project, I have replaced all the missing values with the mean of the variable.

The following table contain the results of k means clustering for six variables.

K-means clustering with 4 clusters of sizes 1507, 4200, 8902, 2997

| | **Income** | **Paying for housework** | **Bathing or Dressing** | **Physical Activities** | **Outdoor Tasks** | **Unpaid Work** |
|---|---|---|---|---|---|---|
| 1 | 10.433 | 1.861 | 1.355 | 1.484 | 4.254 | 5.094 |
| 2 | 0.010 | 1.875 | 1.321 | 1.437 | 5.508 | 1.459 |
| 3 | 10.599 | 1.623 | 1.153 | 1.258 | 3.139 | 0.001 |
| 4 | 0.057 | 1.286 | 0.888 | 0.913 | 0.016 | 0.411 |

*Table 5: K-means Clustering*

Interpretation of k-means clustering, the total number of data points are 17606 with 6 variables. Total number of data points in cluster one is 1507 in which most of the data points are from income and second most data points are from unpaid work. So, cluster one relates income and unpaid work. For instance, mostly teenagers will do unpaid work to get some experience and then get a job with the experience because it is very difficult to get a job without experience.

Cluster two of size 4200 data points in which most of the data points are from outdoor tasks and paying for housework. Cluster two relates to paying for housework and outdoor tasks. For instance, people will spend most of their time in doing outdoor tasks, so they regularly pay someone to do their housework.

Cluster three of size 8902 data points in which most of the data points are from income and outdoor tasks. Cluster 3 relates to income and outdoor tasks. For instance, people spend their time on outdoor tasks like lawn moving etc to get paid. The more hours people spend in outdoor tasks, higher the income.

Cluster four of size 2997 data points in which most of them are from paying for housework, physical activity and bathing or dressing. Cluster 4 relates to paying for housework, physical activity and bathing or dressing. For instance, most of the people do some physical activity like doing exercise, yoga etc early in the morning to maintain their health. Also, people who do physical activity can do bathing or dressing themselves.

## 4.7 Linkage Algorithm

In this project, I have used single linkage, complete linkage, average linkage to determine the relationship between the chosen variables from 2015 HILDA dataset.

### Single Linkage

In single linkage, it will group the clusters when the data points from each cluster has minimum possible distance. I have calculated single linkage and created a cluster dendrogram which show all the data points with its clusters. Below is the single linkage dendrogram.

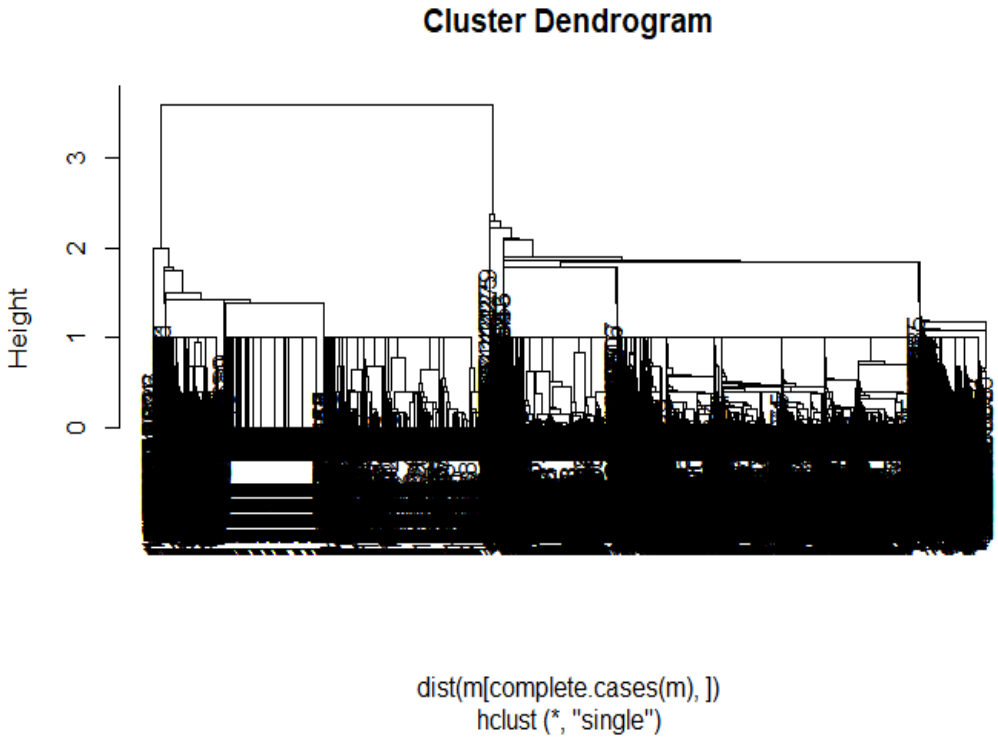



*Figure 16: Single Cluster Dendrogram*

However, single linkage dendrogram is not so clear. So, I cut the cluster dendrogram into four clusters for a better prediction. Below is the cluster cut of single linkage. Also, I have considered mean of the variable for each cluster. Here I have calculated mean for continuous variables.

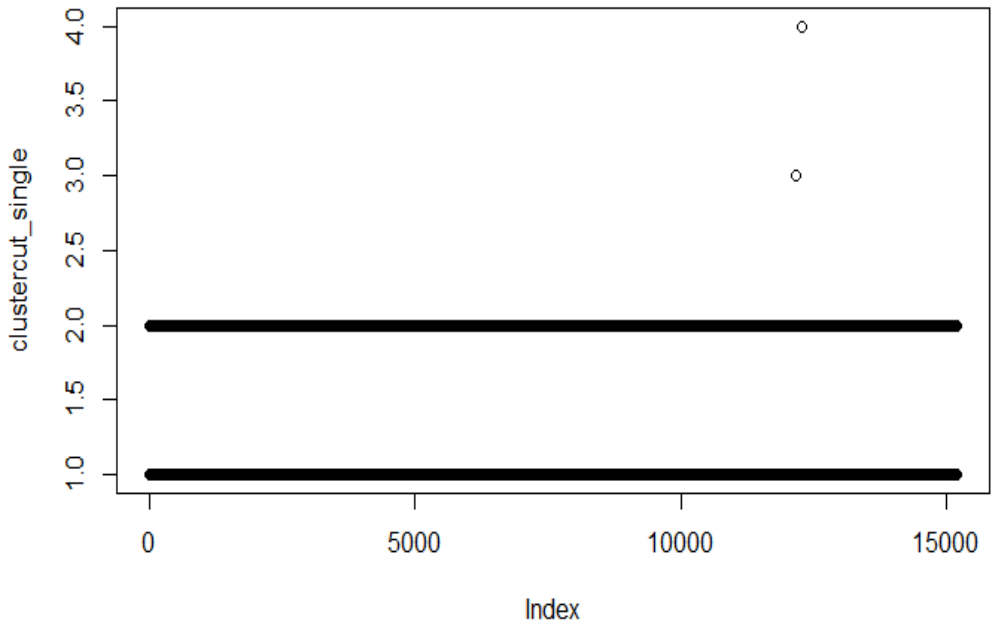



*Figure 17: Single Cluster Cut*

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| **Income** | 6.294 | 6.222 | 0 | 0 |
| **Outdoor tasks** | 3.242 | 3.307 | 4.110 | 5.484 |
| **Unpaid work** | 0.870 | 0.831 | 0 | 0 |

*Table 6: Mean of single cluster cut*

Cluster one and two in which most of the data points are from income and outdoor tasks. Cluster one and two relates to income and outdoor tasks. For instance, people spend their time on outdoor tasks like lawn moving etc to get paid. The more hours people spend in outdoor tasks, higher the income. Cluster 3 and 4 relates to outdoor tasks.

### Complete Linkage

In complete linkage, it will group the clusters when the data points from each cluster has maximum possible distance. I have calculated complete linkage and created a cluster dendrogram which show all the data points with its clusters. Below is the complete linkage dendrogram.

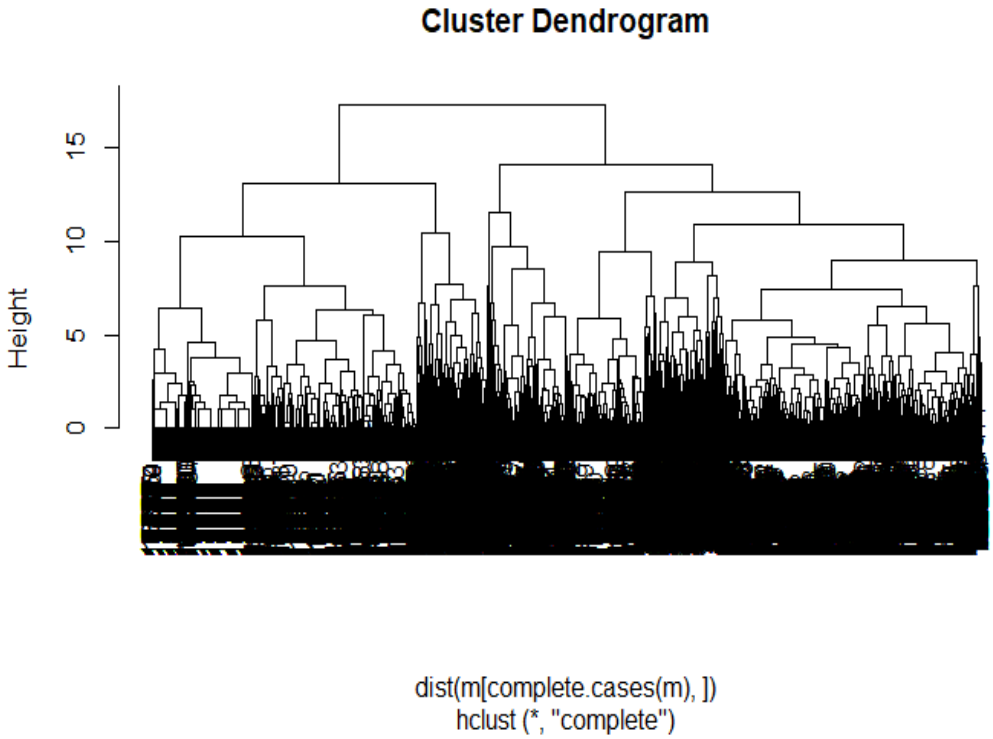



*Figure 18: Complete Cluster Dendrogram*

However, complete linkage dendrogram is not so clear. So, I cut the cluster dendrogram into four clusters for a better prediction. Below is the cluster cut of complete linkage. Also, I have considered mean of the variable for each cluster. Here I have calculated mean for continuous variables.

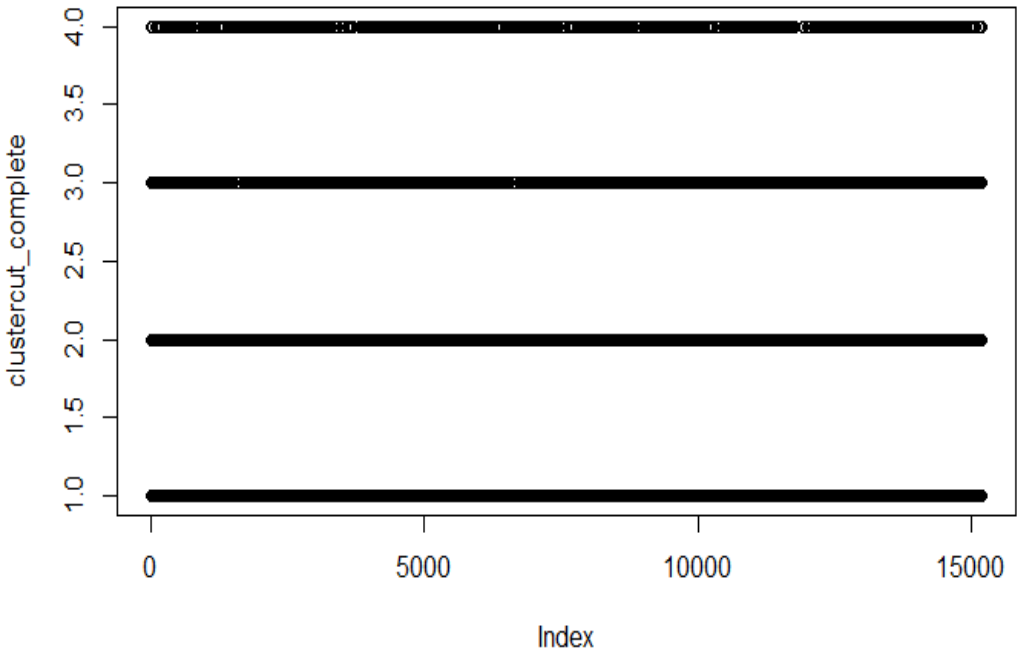



Figure 19: Complete Cluster Cut

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| **Income** | 6.314 | 6.233 | 6.176 | 6.190 |
| **Outdoor tasks** | 3.242 | 3.333 | 3.248 | 3.205 |
| **Unpaid work** | 0.859 | 0.821 | 0.930 | 0.865 |

*Table 7: Mean of complete cluster cut*

Cluster one, two, three and four in which most of the data points are from income and outdoor tasks. Cluster one and two relates to income and outdoor tasks. For instance, people spend their time on outdoor tasks like lawn moving etc to get paid. The more hours people spend in outdoor tasks, higher the income.

**Average Linkage:** To overcome the problems of single and complete linkage. The average linkage will consider the average of the distance between the cases and then it will decide whether to merge the clusters or not. It will

provide more accurate results. I have calculated average linkage and created a cluster dendrogram which show all the data points with its clusters. Below is the average linkage dendrogram.

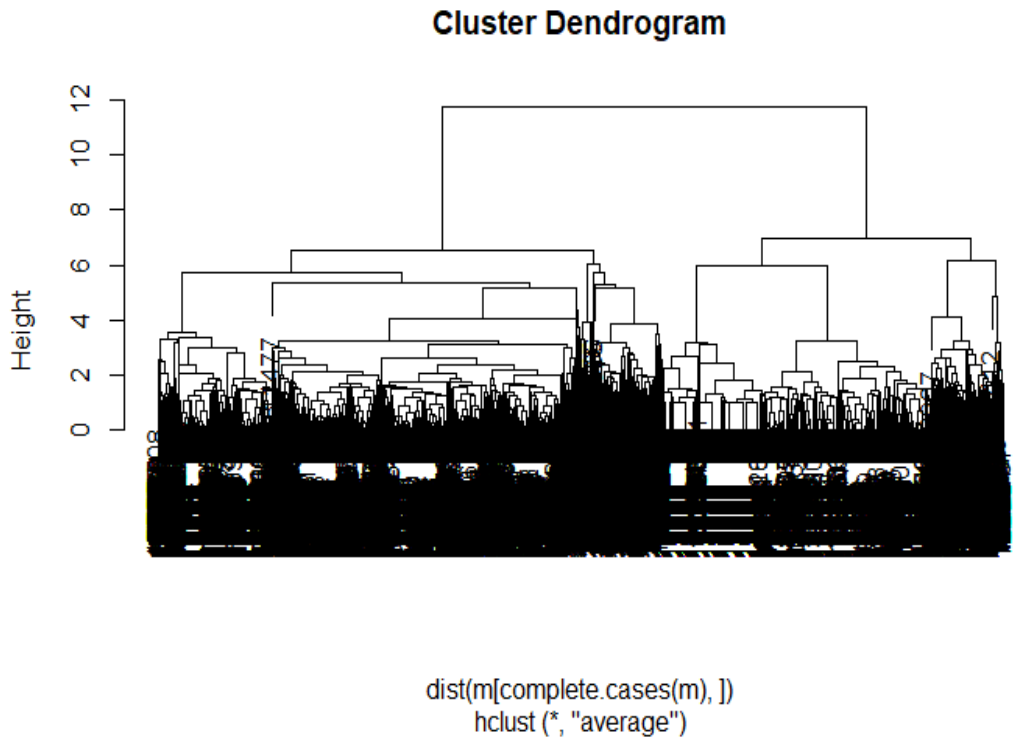



*Figure 20: Average Cluster Dendrogram*

However, average linkage dendrogram is not so clear. So, I cut the cluster dendrogram into four clusters for a better prediction. Below is the cluster cut of average linkage. Also, I have considered mean of the variable for each cluster. Here I have calculated mean for continuous variables.

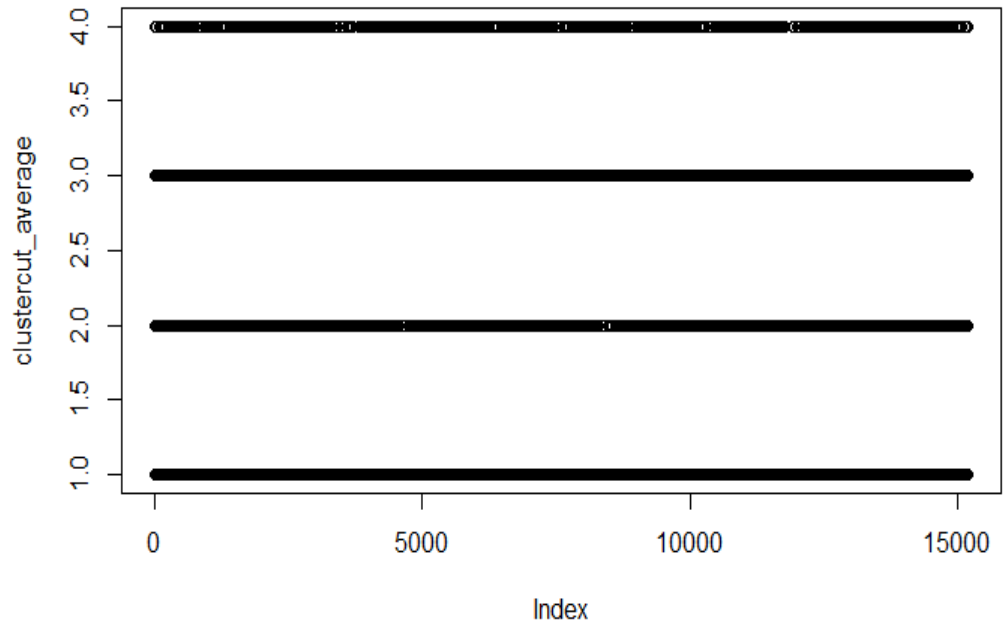



*Figure 21: Average Cluster Cut*

| | **1** | **2** | **3** | **4** |
|---|---|---|---|---|
| **Income** | 6.288 | 6.317 | 6.231 | 6.190 |
| **Outdoor tasks** | 3.241 | 3.245 | 3.335 | 3.208 |
| **Unpaid work** | 0.873 | 0.855 | 0.822 | 0.868 |

*Table 8: Mean of average cluster cut*

Cluster one, two, three and four in which most of the data points are from income and outdoor tasks. Cluster one and two relates to income and outdoor tasks. For instance, people spend their time on outdoor tasks like lawn moving etc to get paid. The more hours people spend in outdoor tasks, higher the income.

# Chapter 5
# Discussion

Matthies et al researched different determinants of power utilization. The various levelled system adhered to certain causal (fig. 22,) and methodological contemplations. The proposed system appeared to be fruitful at clarifying which (obtaining or potentially utilize) behaviours and activities of residents could represent the impacts of indirect variables, (for example, the quantity of teenagers in a household). This was reflected in our outcomes, which affirmed that the impact of the quantity of young people on power utilization was not any huger when the impacts of (intervening) behavioural variables were considered.

The Causal Order of Determinants of Electricity Consumption in Households

| indirect determinants | | direct determinants | |
|---|---|---|---|
| (past/present) house characteristics, sociodemographic & economic aspects, e.g.: | activities, e.g.: | (present) use behaviours, e.g.: | (past) purchasing behaviours, e.g.: |
| number of residents<br>household income<br>floor area in $m^2$<br>age of residents | time spent at home | frequency of use of appliances<br>electricity curtailment behaviours | number of appliances<br>efficiency of appliances |

*Figure 22: The Casual Order of Determinants of Electricity Consumption in Households*

Another progression offered by the present examination was a more differentiated investigation of sociodemographic factors, particularly the quantity of teenagers living in a household [3]. This variable had no critical impact on power utilization when acquiring and utilize behaviours were incorporated into the model. Subsequently, an extra investigation was conducted just in family families to inspect which specific variables would intervene the impact of youths on power utilization.

Our outcomes (way display, Fig. 23) demonstrated that young people impact the family household's number of IT machines. These (past) buying choices at that point appear to have an expanding effect on private power utilization. This conclusion

depends on the finding that the quantity of IT appliances (an immediate determinant) intervened the positive impact of the quantity of teenagers (a circuitous determinant) on power utilization in family households [3].

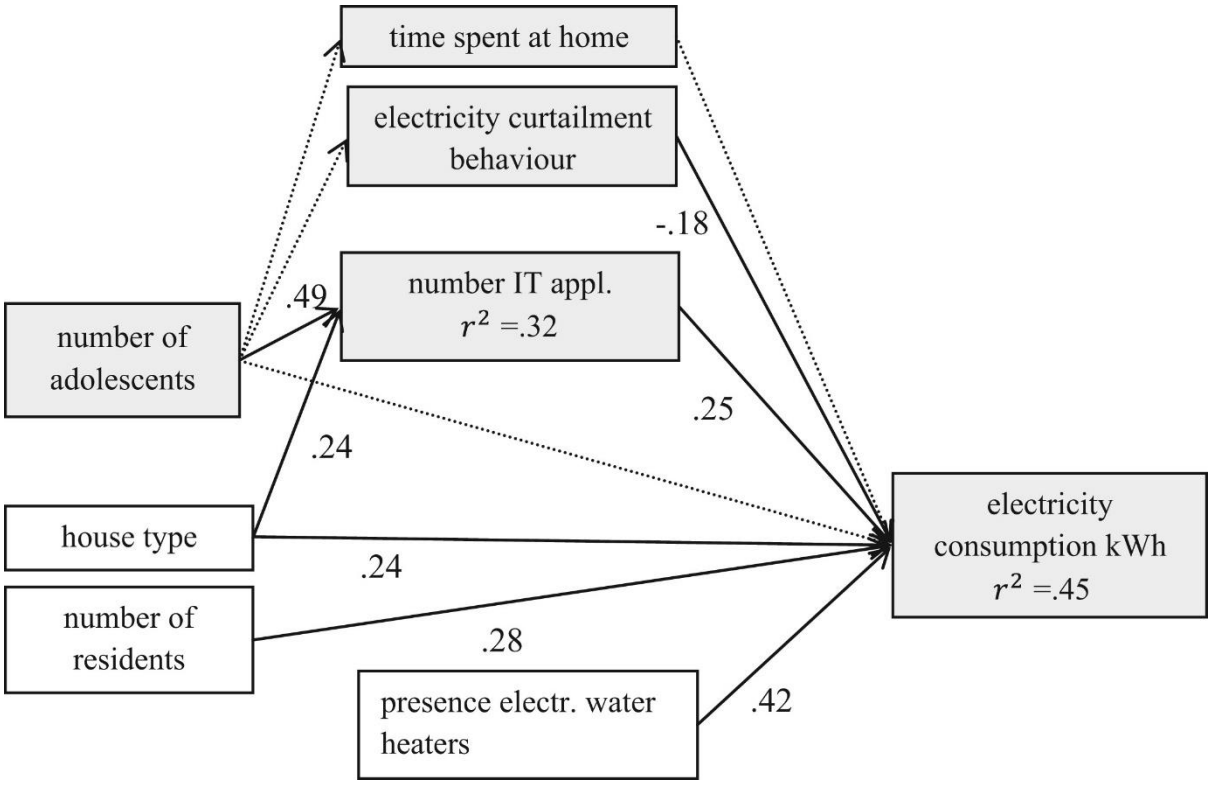



*Figure 23: Electricity Consumption in Family Households. Note: dotted line=nonsignificant path, solid line=significant path p<0.05*

Income or employment reducing electricity consumption could lead to a fall in income and/or employment [23]. So, increase in income increases electricity consumption. Generally, unpaid work like volunteering, social activities etc are more likely done by teenagers because their need some experience to find a job. Several investigations have reported that when the number of adolescents in a household increases, residential electricity consumption increases as well [3]. Outdoor workers in a household get higher electricity bills. For instance, because of lawn moving work they must regularly charge the battery of lawnmower depends on the number of houses they do lawn moving, which consumes more electricity usage. Lawnmower consumes minimum 1000W (which is equal to home air conditioner) and maximum 1400W [24].

# Chapter 6

# Conclusion

The statistical methods show that the list of some selected variables shows the variation of the electricity usage. The above analysis suggests that household members who earn high income, spending time in outdoor tasks and unpaid work will get higher electricity bills. Whereas, people spending their time in home repairs, physical functioning and physical activity will get less electricity bills.

# Appendix

**New Person Questionaries**

| Question Number | Description | Page Number | Variables Name | Label | Continuous/Categorical | Other comments |
|---|---|---|---|---|---|---|
| NPQ A6 | (Since leaving school (as a [child / teenager] )), have you ever enrolled in a course of study to obtain a trade certificate, diploma, degree or other educational qualification? | 239 | oedqenrn | NPQ: A6 Enrolled in course of study to obtain qualification | Categorical | P=0.579 (No) |
| NPQ A10 | Are you currently enrolled in a course of study for a trade certificate, diploma, degree or any other educational qualification? | 244 | oedcqen | NPQ: A10 Currently enrolled in a course | Categorical | P=0.0736 |
| F1b | CONFIRM: Do you currently receive | 289 | owschave | F1b Currently receive income from | Categorical | P<2e-16 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | income from wages or salary? | | | wages/salary | | |
| NPQH1 | Looking at SHOWCARD H4, which of these best describes your current marital status? And by "married" we mean in a registered marriage. | 362 | omrcms | CPQ:H4/NPQ:H1 Current marital status | Categorical | P<2e-16 |
| NPQH2 | Looking at SHOWCARD H5, which of the following best describes your current living circumstances? | 362 | Omrclc | CPQ:H5/NPQ:H2 Current living circumstances | Categorical | P=0.36(No) |
| K1 | Looking at SHOWCARD K1, do you have any long-term health condition, impairment or disability (such as these) that restricts | 383 | Ohelth | K1 Long term health condition | Categorical | P=5.8e-12 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | you in your everyday activities, and has lasted or is likely to last, for 6 months or more? | | | | | |
| DV: C11 | What kind of work do you do in this job? That is, what is your occupation called and what are the main tasks and duties you undertake in this job? Please describe fully. | 257 | Ojbmo62 | DV: C11 Occupation 2-digit ANZSCO 2006 | Categorical | P=3.4e-14 |
| F3 | For your [IF F2=1: main job / ELSE job] what was the total gross amount of your most recent pay before tax or anything else was taken out? It will help | 289 | Owscmga | F3 Total gross amount of most recent pay before deductions | Continuous | P<2e-16 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | to answer this question if you can refer to your last pay-slip. | | | | | |
| F32 | Last financial year, what was your total wage and salary income from all jobs before tax or anything else was deducted? | 304 | owsfga | F32 Gross financial year wages and salaries ($) [weighted topcode] | Continuous | P<2e-16(Selected) |

**Self-completion questionnaires**

| Question Number | Description | Page Number | Variable Name | Label | Conti/ Categ | Other Comments |
|---|---|---|---|---|---|---|
| SCQ B7 | Are you currently an active member of a sporting, hobby or community-based club or association? | 5 | olsclub | SCQ: B7 Currently an active member of a sporting/hobby/community-based club or association | Categorical | P=0.0045 |
| SCQ B19e | How much time would you spend on each of | 10 | olshrod | SCQ: B19e Hours per week - Outdoor tasks | Continuous | P=5.79e-15 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | the following activities in a typical week? e) Outdoor tasks, including home maintenance (repairs, improvements, painting etc.), car maintenance or repairs and gardening | | | | | |
| SCQ B19h | How much time would you spend on each of the following activities in a typical week? h) Volunteer or charity work (for example, canteen work at the local school, unpaid work for a community club or organisation) | 10 | olshrvol | SCQ: B19h Hours per week - Volunteer/Charity work | Continuous | P=0.0192 |
| SCQ B21 | Does your household | 11 | olspayhw | SCQ: B21 Regularly pay | Categorical | P=1e-08(Selected) |

| | | | | | | |
|---|---|---|---|---|---|---|
| | regularly pay someone to do any of the housework (cleaning, washing, ironing, cooking, etc)? | | | someone to do housework | | |
| SCQ E1 | Are you currently in paid work? | 17 | Ojopw | SCQ: E1 Is in paid work | Categorical | P<2e-16 |
| SCQ A3a | a) Vigorous activities, such as running, lifting heavy objects, participating in strenuous sports | 2 | Ogh3a | SCQ: A3a Physical Functioning: Vigorous activities | Categorical | P=0.124(No) |
| SCQ A3b | b) Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling or playing golf | 2 | Ogh3b | SCQ: A3b Physical Functioning: Moderate activities | Categorical | P=3.53e-13 |
| SCQ A3c | c) Lifting or carrying groceries | 2 | Ogh3c | SCQ: A3c Physical Functioning: Lifting or carrying groceries | Categorical | P<2e-16 |
| SCQ A3j | Bathing or Dressing | 2 | Ogh3j | | Categorical | P=3.17e-16(Selected) |
| SCQ A1 | In general, would | 2 | Ogh1 | SCQ: A1 Self-assessed health | Continuous | P=6.92e-06 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | you say your health is: | | | | | |
| SCQ B1 | In general, how often do you participate in moderate or intensive physical activity for at least 30 minutes? | 5 | olspact | SCQ: B1 How often participate in physical activity | Categorical | P=0.00881(Selected) |
| SCQ B10 | How often do you feel rushed or pressed for time? | 6 | olsrush | SCQ: B10 How often feel rushed or pressed for time | Continuous | P<2e-16 |
| SCQ B11 | How often do you feel you have spare time that you don't know what to do with? | 6 | Olsstime | SCQ: B11 Spare time that don't know what to do with | Continuous | P<2e-16 |
| SCQ B16 | In general, about how often do you get together socially with friends or relatives not living with you? | 8 | Olssocal | SCQ: B16 How often get together socially with friends/relatives not living with you | Continuous | P=0.0126 |
| SCQ B19e | How much time would | 10 | Olsmnod | SCQ: B19e Minutes per week - Outdoor tasks | Continuous | P=0.000341 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | you spend on each of the following activities in a typical week? e) Outdoor tasks, including home maintenance (repairs, improvements, painting etc.), car maintenance or repairs and gardening | | | | | |
| SCQ B19 e Combine | Outdoor Tasks | NA | Olsmnod & olshrod | Own created variable | Continuous | P<2e-16(Selected) |
| SCQ B19h | How much time would you spend on each of the following activities in a typical week? h) Volunteer or charity work (for example, canteen work at the local school, unpaid work for a communi | 10 | olsmnvol | SCQ: B19h Minutes per week - Volunteer/Charity work | Continuous | P=0.234(No) |

| | | | | | | |
|---|---|---|---|---|---|---|
| | ty club or organisation) | | | | | |
| SCQB19hCombine | Unpaid work | | Olsmnvol&olshrvol | Own created a variable | Continuous | P=6.31e-05(Selected) |
| SCQC1 | Given your current needs and financial responsibilities, would you say that you and your family are: | 11 | ofiprosp | SCQ:C1 Prosperity given current needs and financial responsibilities | Continuous | P=1.7e-08 |

*Table 9: Variables chosen for Linear Regression Analysis*

| Variable Name | Estimate | P value | Type |
|---|---|---|---|
| Income | 0.299 | <0.001 | Continuous |
| Bathing or Dressing | 0.718 | 0.000 | Categorical |
| Physical Activity | -0.136 | 0.063 | Categorical |
| Outdoor Tasks | 0.097 | <0.001 | Continuous |
| Unpaid work | 0.042 | 0.000 | Continuous |

*Table 10: 2014 Linear Regression Analysis*

# Bibliography


1. A. K. Pears, "Imagining Australia's energy services futures," *Futures,* vol. 39, no. 2, pp. 253-271, 2007/03/01/ 2007.
2. C. BLUE. (2018, January 15). What is the average electricity bill. Available: https://www.canstarblue.com.au/energy/electricity/average-electricity-bills/
3. E. Matthies, M. Nachreiner, and H. Wallis, "Adolescents and electricity consumption; Investigating sociodemographic, economic, and behavioural influences on electricity consumption in households. (Special Section on the European Union: Markets and Regulators)," *Energy Policy,* vol. 94, pp. 224-234, 2016.
4. I. Dent, T. Craig, U. Aickelin, and T. Rodden, "An Approach for Assessing Clustering of Households by Electricity Usage," 2014.
5. K. Atalay, S. Whelan, and J. Yates, "Housing Wealth and Household Consumption: New Evidence from Australia and Canada," E. School Of, Ed., ed, 2013.
6. M. Lenzen, M. Wier, C. Cohen, H. Hayami, S. Pachauri, and R. Schaeffer, "A comparative multivariate analysis of household energy requirements in Australia, Brazil, Denmark, India and Japan," *Energy,* vol. 31, no. 2, pp. 181-207, 2006/02/01/ 2006.
7. M. Lenzen, "Energy and greenhouse gas cost of living for Australia during 1993/94," Energy, vol. 23, no. 6, pp. 497-516, 1998/06/01/ 1998.
8. M. Lenzen, C. Dey, and B. Foran, "Energy requirements of Sydney households," Ecological Economics, vol. 49, no. 3, pp. 375-399, 2004/07/01/ 2004.
9. R. Kellogg, "Efficiency in energy production and consumption," S. Borenstein and J. Perloff, Eds., ed: ProQuest Dissertations Publishing, 2008.
10. S. Speidel and T. Bräunl, "Driving and charging patterns of electric vehicles for energy usage," Renewable and Sustainable Energy Reviews, vol. 40, pp. 97-110, 2014/12/01/ 2014.
11. Z. M. Chen and G. Q. Chen, "An overview of energy consumption of the globalized world economy," Energy Policy, vol. 39, no. 10, pp. 5920-5928, 2011/10/01/ 2011.

12. N. Watson and M. P. Wooden, "The HILDA survey: a case study in the design and development of a successful household panel survey," Longitudinal and Life Course Studies, vol. 3, no. 3, pp. 369-381, 2012.
13. M. J. Crawley, Regression. Chichester, UK: Chichester, UK: John Wiley & Sons, Ltd, 2012, pp. 449-497.
14. R. E. Frank, W. F. Massey, and Y. Wind, Market segmentation / Ronald E. Frank, William F. Massey, Yoram Wind. Englewood Cliffs, N.J.: Englewood Cliffs, N.J.: Prentice-Hall, 1972.
15. R. L. Gorsuch, Factor analysis / Richard L. Gorsuch, 2nd ed.. ed. Hillsdale, N.J.: Hillsdale, N.J. : Erlbaum Associates, 1983.
16. RStudio Team (2016). RStudio: Integrated Development for R. RStudio, Inc., B oston, MA URL http://www.rstudio.com/.
17. R. P. Cabeen, M. E. Bastin, and D. H. Laidlaw, "A Comparative evaluation of voxel-based spatial mapping in diffusion tensor imaging," NeuroImage, vol. 146, pp. 100-112, 2017/02/01/ 2017.
18. "What is Linear Regression? - Statistics Solutions", *Statistics Solutions*, 2018. [Online]. Available: https://www.statisticssolutions.com/what-is-linear-regression/. [Accessed: 08- Nov- 2018].
19. "Assumptions of Linear Regression - Statistics Solutions", *Statistics Solutions*, 2018. [Online]. Available: https://www.statisticssolutions.com/assumptions-of-linear-regression/. [Accessed: 08- Nov- 2018].
20. "Exploratory Factor Analysis in R | | PromptCloud", *Promptcloud.com*, 2017. [Online]. Available: https://www.promptcloud.com/blog/exploratory-factor-analysis-in-r/. [Accessed: 08- Nov- 2018].
21. A. Trevino, "Introduction to K-means Clustering", *Datascience.com*, 2016. [Online]. Available: https://www.datascience.com/blog/k-means-clustering. [Accessed: 08- Nov- 2018].
22. O. Yim and K. T. Ramdeen, "Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data," *The quantitative methods for psychology,* vol. 11, no. 1, pp. 8-21, 2015.
23. P. K. Narayan and R. Smyth, "Electricity consumption, employment and real income in Australia evidence from multivariate Granger causality tests," *Energy policy,* vol. 33, no. 9, pp. 1109-1116, 2005.

24. Daftlogic.com. (2018). *Power Consumption of Typical Household Appliances*. [online] Available at: https://www.daftlogic.com/information-appliance-power-consumption.htm [Accessed 11 Oct. 2018].