Statistical Analysis Of Energy Consumption Big Data

Suman Bidarahalli
TU Darmstadt
Darmstadt, Germany 64289
Email: suman.bidarahalli@stud.tu-darmstadt.de

Saba Sabrin
TU Darmstadt
Darmstadt, Germany 64289
Email: saba.sabrin@stud.tu-darmstadt.de

***Abstract*—Energy conservation is an exigency. In order to optimize energy consumption, it is crucial to understand energy usage and find out various parameters which impact the energy consumption. Big data analysis is one way of analyzing the energy consumption. To gain insights about the energy consumption, huge amount of data is collected via various sources like sensors, social media, mobile phones etc. The two data sets under analysis are, the data set from three real homes (UMass Smart* Home Data Set) and the data set collected by the Commission for Energy Regulation (CER), Ireland. Big data analysis needs a suitable environment which enables fast processing of huge amount of data and fetch accurate results. To aid the big data analysis, we use a distributed computing platform called as the *IBM Analytics for Hadoop* provided by IBM. Through big data analysis, some of the parameters which have an impact on energy consumption were found and understood.**

***Index Terms*—Big Data,Correlation,PaaS,Distributed Computing**

I. Introduction

Global consumption of energy is increasing day by day, as a consequence of rapid improvements in standard of living across the globe. It is predicted that the consumption of energy is going to increase by 56% until 2040 [1]. Energy conservation is undoubtedly the highest priority for each of us. To do this efficiently and effectively, we need to first understand the various parameters which impact the consumption. In our approach, we are analyzing the energy consumption data to gain insights and understand the possible reasons for the observed energy consumption. This data is collected via sensors [2] fitted in the homes and companies whose energy consumption is under study. Such huge amount of raw, unstructured data with varied complexity is called as Big Data [3].

Big data analysis is done to extract useful information from the raw data. It is a new research field which presents us with an opportunity to work on huge amounts of data and derive significant results [4]. Big data analysis adds value to the business sector, lets the analysts predict the possible trends, as per [5] more than 100 billion Euros ($149 billion) could be saved in the operational efficiency improvements by the government administrators in the developed European economy by using Big Data solutions.

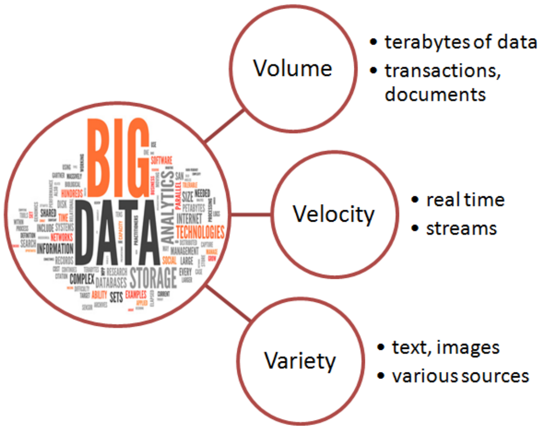

Fig. 1. Scale for Coefficient Correlation r

Big data is characterized by three Vs (as seen in Figure 1)- velocity, volume and variety [6]. These characteristics of big data make it extremely difficult to perform the analysis correctly and efficiently [7]. So, to deal with such difficulties, suitable environment for big data analysis is used. One of such environment is the distributed computing platform provided by IBM Bluemix. More detailed discussion about the technical support for big data analysis is done in the section III.

The data under analysis with various demographic features, is discussed thoroughly in the section III. The section V provides the detailed information about the Smart* Data Statistical Analysis followed by the section VI, which details the CER Data Statistical Analysis. Section VII provides the conclusion drawn from the analysis.

II. Related Work

There is currently a lot of work happening in the field of Big Data Analytics in statistical research as well as domains related to data analyzing and processing. It is a current hype in the cloud and database technology too. As of in any type of software, there is always a moderate and sometimes too big parts which includes data transformation and analyzing. From the last section, enough details has been discussed about the Big Data Analytics and other related technologies. The basic motivation. All these raw or unprocessed data which we ultimately call Big Data could be in some specific format or possibly depend on some demographic factors. There are some more works carried out before and still continued in the area of big data processing. One of them is, Time Series Prediction of Smart Energy Data. In that research, the goal was to process huge telemetry data based on massive timestamps generated from the Smart Metering Project by Commission for Energy Regulation (CER) and to forecast on some energy efficiency techniques. Here they have used different kind of statistical analysis and scientific algorithms for processing the Big Data, such as, k-Means algorithm for feature selection and behavioral analysis of the data. They have also used Linear or Polynomial Regression model to predict the future using the pas data analysis. The dataset used by them is one of the dataset we have used too for our analysis purpose. But, our approach towards the dataset is entirely on a different direction. We have not worked on the direction of time-series analysis, rather than we worked on both the energy and the performed survey data provided by CER. There has been some more researches conducted on the areas of energy efficiency and performance upgrade for computing environments of using Big Data based applications depending on sorting the huge amount of raw data each time by the CPU. New datasets are rising faster on the horizon of Big Data merely by the researchers and different institutions. Such as, a very powerful dataset of Big Microdata on population research having over 750 million records describing individuals regarding different factors like, age, gender, profession and other behavioral aspects of human being. It is expected to increase the size of this dataset to 2 billion by 2018. These kind of datasets are a big motivation towards the researchers to work and analyze more on the Big Data analytics.

Our research basically started with the Smart* project. But besides that, there were some more motivational researches conducted on Smart Meter for saving energy in homes, cutting the electricity bills and using renewable energy efficiently mostly for Smart homes. The Smart* (Smart Star) project was mainly initialized to optimize home energy consumption, mainly focusing on the Smart homes. The project provides a wide range of real-life data gathered from the heavily instrumented smart homes and thus opens a broad door to experiment the data with the help of new data processing techniques, tools and algorithmic modeling. We will describe in details about this dataset on the later section.

III. Data Sets

The Umass Smart* Home data set is the high resolution data set from three homes - Home A, Home B and Home C along with microgrid data which consists of observations from 400 homes located in Western Massachusetts [8]. The three homes differ in the area (measured in square foot), number of full time occupants, number of rooms, mix of sensors deployed, appliances used in the house etc. Home A data set has circuit data file which has the power load observed in Watts for each circuit in the home. The power load is measured every second. Each circuit is identified by a unique circuit ID and the grid (whose circuit ID is 1) corresponds to the total load, which should ideally be equal to sum of all individual circuit loads. Along with circuit data, motion sensors readings, thermostat sensor readings contribute to the motion data and the environment data. All the files consists of values, separated with commas and in .csv format. The meta-data of the file is in the text file named as *format*, for example the environment data has the meta-data which describes the columns of the data i.e. the parameters of the data like inside temperature, outside temperature, inside humidity, outside humidity etc. There are variety of data types i.e. numeric (real power consumption), boolean (fridge door is open/close i.e.1/0), string (name of the circuit: fridgeRange corresponds to the fridge circuit).

The data set from CER group, is much vast compared to the first data set. The CER data set has readings for 7444 meter Ids for electricity consumption out of which 4225 meter Ids correspond to the home, 485 meter Ids correspond to the *S*mart Metering Electricity (SME) and 1735 meter Ids belong to other group.The CER data set also has observations of gas consumption of 2575 meter Ids, out of which 543 houses belong to the control group. All the above information is available in the file named as *r*esidential allocations provided in the zipped folder.

The homes are characterized by numerous demographic features like varying employment status of the occupants, number of people living in the home, living with or without families, families with or without kids, people belonging to different age groups etc. Each observation in the data file is a tuple of three variables namely -

1) MeterID - Meter identifiers.
2) DT - Date and Time expressed 5 digit string,first 3 digits depict the day of the year and remaining 2 depict time in 30 min time slots which gives 48 readings each day
3) Usage - Electricity / Gas usage in KwH.

There are two files which describe the pre-survey and the post-survey questions along with answers from the participants of the survey. Using the answers to the pre-survey or the post-survey questions, it is possible to deduce some information of the house having the corresponding meterID.

IV. Hardware And Software Tools

We have used different software based tools and almost no hardware tools, other than a local machine (personal computer) with high configuration memory of at least 4 GB (Gigabyte). The need of high end memory (RAM) is basically used for huge amount of raw data processing. It was optionally used as an alternative approach though. Figure 2 basically points on the design of our Big Data processing and also the connectivity between all software tools or components we have used.

*A. **Programming Languages and Tools***

1) R Language
2) BigR is an API (Application Programming Interface) provided by IBM Analytics. It offers an end-to-end integration for R within IBM InfoSphere BigInsights. We have used this API as a library in the R platform for connecting to the IBM Server for data accessing.
3) Data Formatting tools, Data Archiving tools, Windows command tools for data transformation

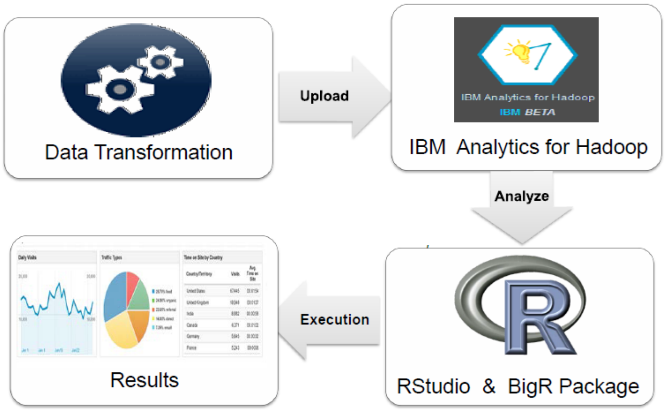

Fig. 2. Interaction of Software Tools

B. *Development Infrastructure*

1) **IBM Bluemix :** IBM Bluemix is a digital innovation platform from IBM. It provides all types of computing technologies through different cloud based services. It is basically a cloud based platform which has app-centric run time environment for easier and flexible application execution. As of there are a lot of different services available inside the IBM Bluemix, our point of concern was specifically on the IBM Analytics for Hadoop service. We have used this service to process our data through the powerful cloud based servers. For that purpose, we had to transform the data and then it was uploaded to the IBM cloud server using the mentioned service. The whole mechanism works in a simple way. If a data processing request has been made from the local machine using the authentication through the BigR API to the IBM Bluemix server, there was a processing of those uploaded data using the Map-Reduce technology.
2) **R Studio :** R Studio is a platform for analytical and statistical based data processing. We have used this platform for all of our Big Data processing through connecting with the IBM Analytics platform or either processing all the data from the In-Memory configuration. In-Memory computation means, copying all the data inside the cache memory used from the main memory of the local machine. R does this for faster data processing while availing all the library packages to analyze the data. There are multiple packages available for plotting data as graphs, or simulating the data using math functions and others for specific purposes.

V. Smart* Data Statistical Analysis

The Smart* data analysis is details the correlation between the power consumption and the different parameters of the data of Home A. The correlation coefficient, r, is a measure which helps to understand the extent of the statistical relationship between two variables [9]. The Figure 3 summarizes the possible coefficient correlation(r) values and the correlation types [9] .

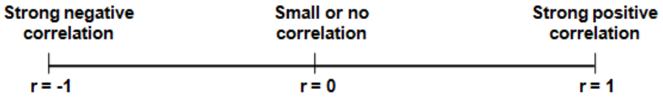

Fig. 3. Scale for Coefficient Correlation r

The circuits data of Home A consists of the observed load for the all the circuits in the house. The door data contains the observation of the motion of the fridge door. Correlation test helps to understand the correlation of the fridge door motion on the fridge load. R-studio offers the cor function to find the correlation between the variables passed to it as parameters. The result of the correlation test between the fridge load and the fridge door motion showed almost weak negative correlation with correlation coefficient value of -0.37. Whenever the fridge door opens, the temperatures inside the fridge tends to increase [8]. If the temperature of the fridge increases beyond the threshold, the thermostat kicks in and triggers the compressor to pull more energy to bring the temperature below the threshold and thus maintaining the temperature of the fridge [8]. This is one possible reason for almost weak negative correlation as the opening of the fridge door may not lead to increase in the load each time unless the temperature of fridge crosses the threshold.

To check if the result of the cor function is significant or by chance, hypothesis testing is done on the correlation coefficient r. R-studio provides the cor.test function which performs the hypothesis testing for the correlation coefficient between the variables given as input. The cor.test uses the Pearson's product-moment correlation. It considers:

1) Null hypothesis H_0- There exists no correlation between fridge door motion and fridge load
2) Alternate hypothesis H_1- There is a correlation between them

The result of the cor.test are summarized in the Table I.

TABLE I
Test of significance for fridge load and fridge door motion

Test of significance	p-value	correlation coefficient (r)	Interpretation
cor.test on fridge load and fridge door motion	2.2e-16	-0.3711785	As the p-value <<0.05, the probability of getting r = -0.37, given the null hypothesis is true, is very low which means that we reject the null hypothesis

One more aspect of the Smart dataset was to analyze the power consumption dependency over environmental features. Initially, Smart Home A power consumption data has a power value measured against each circuit of the house. As we have analyzed some of the circuits, there were sometimes pattern seen or not seen at all. But, the main purpose was to find correlations among the power consumption and other factors. Now, regarding the environment data measured for the Home A consists of different factors like, Inside Temperature, TimestampUTC, inside-Temp, outside-Temp, inside-Humidity, outside-Humidity, wind-Speed, wind-Direction-Degrees, wind-Gust, wind-Gust-Direction-Degrees, rain-Rate, rain, wind-Chill and heat-index .

So, as we proceed with the analysis, the first step was to transform the complete circuit data from multiple files to a combined file so that, average power consumption can be calculated for further processing. The second step was to transform the data column ”TimestampUTC” to ”DateTime” in a format so that it can be used to calculate the Monthly or Daily average power consumption. The next step was to merge the circuit and the environment data depending on this common ”DateTime” column. Then according to the subsequent processing, the Monthly Average was calculated and correlation was measured through the cor function. From Figure 3, we can depict the range of the correlation coefficient (r). The value calculated from this function is, 0.2107487 which indicates a very weak positive correlation between the inside temperature and power consumption. The final step is to determine if the value provided from the correlation measurement was entirely by chance or significant. To test that aspect, cor.test was used depending on the significance level and P-value. So before initializing the test, the hypothesis was to be defined as below,

1) Null hypothesis H_0- There is no correlation between the Inside or Outside Temperature and Power Consumption
2) Alternate hypothesis H_1- There is an existing correlation between the Inside or Outside Temperature and Power Consumption

The result of the cor.test are summarized in the Table II.

TABLE II
TEST OF SIGNIFICANCE BETWEEN AVERAGE POWER CONSUMPTION AND INSIDE / OUTSIDE TEMPERATURE OF HOME

Test Factor	p-value	correlation coefficient (r)	Interpretation
Inside Temperature	0.2551	0.2107487	As the p-value >0.05, so the probability of getting r = 0.2107, which means the Null hypothesis is true
Outside Temperature	1.335e-06	0.7477364	As the p-value <<0.05, the probability of getting r = 0.7477, which means the Null hypothesis is failed to be accepted and Alternate is thus accepted

VI. CER Data Statistical Analysis

The analysis of the CER Data is divided in to two sections:

1) Finding the association between the demographic features and energy consumption
2) Finding the association between the environmental parameters and energy consumption

The demographic features considered are the gender, employment status, age groups. The only environmental parameter considered is the season i.e. energy consumption is compared for different time of the year.

A. Association between demographic features and energy consumption

Given a population, checking if the gender has an impact on the gas consumption is the most basic test which could be done.The first experiment done is to compare the retired male to retired female living alone gas consumption.Before performing a test on the data, the selection of the test depends on the distribution of the data. To understand the distribution of the data, box-plots are plotted for the usage of energy for each gender. The Figure 3 summarizes the data for retired male vs retired female gas consumption.

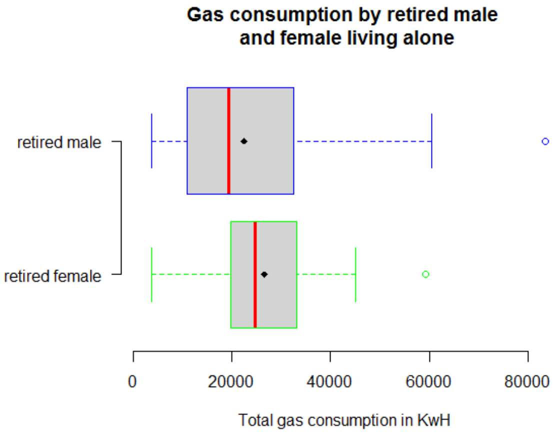

Fig. 4. Distribution of total gas consumption data for retired male and female living alone

The data distribution seen in the Figure 3 shows that the data is not normally distributed. To perform the normality test, the R language function provided is the *shapiro test*. The results of the *shapiro test* are summarized in the Table III.

TABLE III
SHAPIRO TEST FOR NORMALITY

Gas Consumption data	P-value	Interpretation
retired male living alone	0.0004004	As the p-value <<0.05, the null hypothesis that the data is normally distributed is rejected
retired female living alone	0.7152	As the p-value >>0.05, the null hypothesis that the data is normally distributed is accepted

As the data is not normally distributed, the non-parametric test for comparison of two groups is used [10]. This test is called the *Wilcoxon rank sum test* and the R-language function provided is *wilcox.test* [10]. Given μ_m is the mean of total gas consumption of male retired living alone, μ_f is the mean of total gas consumption of female retired living alone,this test considers

1) Null hypothesis H_0 is $\mu_\mathrm{m} - \mu_\mathrm{f} = 0$

2) Alternate hypothesis H_1 is $\mu_m - \mu_f \neq 0$

The results of the *Wilcoxon test* are summarized in Table IV.

TABLE IV
WILCOXON TEST RESULTS

Data	P-value	Interpretation
retired male living alone toal gas consumption	0.06645	As the p-value >0.05, but is statistically insignificant to accept the null hypothesis

The next demographic feature considered is the employment status. The comparison of the gas consumption of retired male and female living alone is done with the gas consumption of the employed male and female living alone. The distribution of the data for the mentioned two groups is seen in the box-plot of Figure 5. The test for the normality is performed using the *shapiro test* and the results show that the data is not normal. Hence, the wilcox.test is used and result of the *wilcox.test* is summarized in Table V.

TABLE V
WILCOXON TEST FOR RETIRED VS EMPLOYED PEOPLE LIVING ALONE GAS CONSUMPTION

Data	P-value	Interpretation
retired vs employed	2.036e-05	As the p-value <<0.05, the null hypothesis that the data is normally distributed is rejected.

The last demographic feature considered is the age groups. For the given CER gas pre-survey data, there are 7 age groups seen. Out of these 7 age groups , the last group belongs to the people who do not wish to disclose their age. Hence, the last group is not considered in the analysis. The first age group is also removed from the analysis as the number of observations were too low.The distribution of the gas consumption data for the age groups for march month is depicted in the Figure 6.

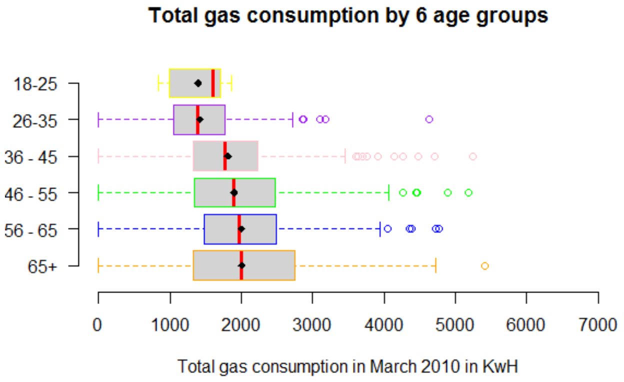

Fig. 5. Distribution of total gas consumption data for the age groups

From 2nd age group to the 6th age group, the distribution of the data is normal as seen in the Figure 6 and also confirmed by the *shapiro test*. As the comparison of gas consumption is between 5groups, the parametric test called as *Analysis of Variance test* is done [11]. *Aov.test* is the function offered by R language for the same. To know more about which groups differ and p-value obtained for each group comparison, the R function is *TukeyHSD* used [12]. The result of the *aov.test* is used as the input parameter to *TukeyHSD test*.To know more about which groups differ and p-value obtained for each group comparison, the R function is *TukeyHSD* used [12]. The result of the *aov.test* is used as the input parameter to *TukeyHSD test*. The results of the *TukeyHSD test* are summarized in the Table VI.

TABLE VI
TUKEYHSD TEST: COMPARISON OF GAS CONSUMPTION BY AGE GROUPS

Age groups	P-value	Interpretation
2-3 2-4 2-5 2-6	0.0000122 0.0000001 0.0000001 0.0000000	As the p-value <<0.05, the null hypothesis is rejected which means that the gas consumption between the 2nd age group and rest of the age groups differs
3-4 3-5 3-6	0.6879002 0.1103307 0.0860217	As the p-value >0.05, the null hypothesis is true which means that the gas consumption of the 3rd age group does not differ when compared to 4th , 5th and 6th group.
4-5 4-6	0.7347049 0.6824941	As the p-value >>0.05,the null hypothesis is true which means that the gas consumption of the 4th age group does not differ when compared to 5th and 6th group
5-6	0.9999960	As the p-value >>0.05, the null hypothesis is true which means that the gas consumption of the 5th age group does not differ when compared to the 6th group

The results of all the above tests are discussed in the section VII.

B. Association between the environmental parameters and energy consumption

To understand if the changes in the weather causes change in the energy consumption, the comparison of the gas consumption within the age group for March and December month is done. As the data for each group is found out to be normal, the comparison within each group for March and December is done by using the *Independent t-test* [13]. The R language function for the *Independent t-test* is called *t.test*. Given μ_m is the mean of total gas consumption of March, μ_d is the mean of total gas consumption of December, the *Independent t-test* considers

1) Null hypothesis H_0 is $\mu_m - \mu_d = 0$
2) Alternate hypothesis H_1 is $\mu_m - \mu_d \neq 0$

The results of the *t.test* are summarized in Table VII for each age group along with the mean values. The interpretation of the t.test results are discussed in the next section.

VII. CONCLUSION AND DISCUSSION

The gender for the given population does not affect the gas consumption. The employment status does affect the gas

TABLE VII
INDEPENDENT T-TEST FOR COMPARING THE GAS CONSUMPTION IN MARCH AND DECEMBER

Group (Size)	Mean in March (KwH)	Mean in December (KwH)	P-value
2-2 (195)	1418.751	1683.985	0.0003992
3-3 (371)	1813.964	2127.933	1.934e-06
4-4 (330)	1904.556	2310.468	3.131e-06
5-5 (222)	2002.357	2436.144	1.286e-05
6-6 (240)	2007.943	2405.068	0.000373

consumption as the p-value is incredibly small leading to the rejection of the null hypothesis. Age group 2 (26-35 years) utilizes the minimum gas as compared to rest of the age group. The variation in the season has a significant impact on the gas consumption within each group. The difference in the means of the gas consumption of each age group between March and December is at least 200 kWh. The conclusion drawn from the tests is summarized in Table VIII.

TABLE VIII
SUMMARY OF THE RESULTS

Demographic feature	Example	Impacts the gas consumption
Gender	Female/ Male	No
Employment status	Retired/Employed	Yes
Age group	25-36,65+	Yes-for some age groups

Through big data analysis, more significant results can be achieved, for example: comparison between the post survey data and the pre-survey data will help us to know if the trial made any difference the energy consumption.The conclusion of the analysis could then put into good use to discover new ways of energy conservation.

ACKNOWLEDGMENT

We would like to thank our mentor, Alaa Alhamoud, for guiding us through our analysis as well as reviewing the report draft. We also thank our peers who have helped us improvise.

REFERENCES

[1] M. Zhou, "World energy consumption to increase 56 percent by 2040 led by asia," July 2013.

[2] M. van Rijmenam, "Understanding the various sources of big data - infographic," July 2013.

[3] S. Sagiroglu and D. Sinanc, "Big data: A review," in *Collaboration Technologies and Systems (CTS), 2013 International Conference on*, May 2013, pp. 42–47.

[4] D. woo Nam, D. woo Kang, and S. Kim, "Process of big data analysis adoption: Defining big data as a new is innovation and examining factors affecting the process," in *System Sciences (HICSS), 2015 48th Hawaii International Conference on*, January 2015, pp. 4792–4801.

[5] J. Manyika, M. Chui, B. Brown, R. Bughin, J.and Dobbs, C. Roxburgh, and A. Byers, "Big data: The next frontier for innovation, competition,and productivity," 2011.

[6] T. Chardonnens, P. Cudre-Mauroux, M. Grund, and B. Perroud, "Big data analytics on high velocity streams: A case study," in *Big Data, 2013 IEEE International Conference on*, October 2013, pp. 784–787.

[7] A. Katal, M. Wazid, and R. Goudar, "Big data: Issues, challenges, tools and good practices," in *Contemporary Computing (IC3), 2013 Sixth International Conference on*, August 2013, pp. 404–409.

[8] S. Barker, A. Mishra, D. Irwin, E. Cecchet, P. Shenoy, and J. Albrecht, "Smart*: An open data set and tools for enabling research in sustainable homes," in *Proceedings of the 2012 Workshop on Data Mining Applications in Sustainability (SustKDD 2012)*, August 2012.

[9] P. Gingrich, "Association between variables," in *Introductory Statistics for the Social Sciences*, 1992.

[10] P. Sedgwick, "Parametric v non-parametric statistical tests," *BMJ*, vol. 344, 2012.

[11] Shapiro and S. Samuel, "Analysis of variance test," in *An Analysis of Variance Test for Normality (complete Samples)*, 1964, pp. 591–611.

[12] Abdi, Herv, and J. Lynne Williams, "Tukeys honestly significant difference (hsd) test," in *Encyclopedia of Research Design*, 2010.

[13] P. Chen and P. Popovich, "Correlation," 2002.