engineering proceedings

MDPI

Proceeding Paper

ANOVA-Based Variance Analysis in Smart Home Energy Consumption Data Using a Case Study of Darmstadt Smart City Germany †

Yamini Kodali [1] and Yellapragada Venkata Pavan Kumar [2,*]

[1] School of Computer Science and Engineering, VIT-AP University, Amaravati 522241, Andhra Pradesh, India; yamini.24phd7012@vitap.ac.in
[2] School of Electronics Engineering, VIT-AP University, Amaravati 522241, Andhra Pradesh, India
* Correspondence: pavankumar.yv@vitap.ac.in; Tel.: +91-863-2370155

† Presented at the 11th International Electronic Conference on Sensors and Applications (ECSA-11), 26–28 November 2024; Available online: https://sciforum.net/event/ecsa-11.

Abstract: The evolution of smart grids (SG) is rapid and ubiquitous with the advent of information and communication technology. SGs enable utilities and prosumers to monitor energy consumption in real-time, thereby possessing effective supply and demand management. The subsets of SGs namely smart homes/smart buildings are tailored to take the benefits of SGs. These smart homes continuously record energy consumption data through smart meters, sensors, and smart appliances and facilitate consumers to track/manage their energy usage in real-time. Usually, the energy consumption of renewable energy-integrated smart homes depends on consumer behavior and weather conditions. These aspects lead to variance in the recorded energy consumption data from the desired levels. This variance in energy consumption impacts pattern finding, forecasting, financial risk, decision-making, and several other grid functionalities. Hence, comprehension of variance in energy consumption is essential to properly manage the energy. With this aim, this paper proposes the variance analysis on the smart home energy consumption readings using a statistical method named "Analysis of Variance (ANOVA)". It is implemented on the Tracebase dataset, which is a smart city database and contains data for ten months. The data were collected in the city of Darmstadt, Germany, in 2012. The proposed ANOVA is applied to all these months' data. As an initial step, the energy consumption readings recorded for every month at each day and at each hour are enumerated and this information is further used to perform day-wise variance analysis using ANOVA. The results show that there is a significant variance in several days in each month. Further, it is revealed that out of ten months, two months have high variability. Thus, this proposed variance analysis helps the stakeholders of SGs to take the necessary precautions for smooth grid functionalities as well as properly estimate future energy requirements.

Keywords: analysis of variance (ANOVA); energy consumption; smart city; smart grid; smart home; variance analysis

Citation: Kodali, Y.; Kumar, Y.V.P. ANOVA-Based Variance Analysis in Smart Home Energy Consumption Data Using a Case Study of Darmstadt Smart City Germany. *Eng. Proc.* **2024**, *6*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s): Name

Published: 26 November 2024

1. Introduction

Variance analysis plays a vital role in data analysis by providing insights into the data distribution and relationships within the data. The variance quantifies the data distribution in the dataset which helps understand the nature of the dataset. In the realm of smart home energy management, understanding the factors influencing energy consumption is crucial for optimizing efficiency and reducing costs. The energy consumption in the residential sectors has been captured for long [1]. For example, the Tracebase dataset is one such large database that was captured using smart home metering [2]. Using such databases, there are some machine learning models were developed to predict energy consumption in smart homes [3]. Similarly, there are certain methods namely SARIMA, LSTM RNN, and Fb Prophet models were used for forecasting energy prosumer behaviors [4,5]. There is especially a drastic change in energy consumption observed during and after COVID-19 [6]. The higher energy consumption and the customer behavior are the sources for various data anomalies during the capturing and creating inconsistency in the data [7]. Further, some studies find the energy readings of each appliance as well as the missing instances in the particular appliances [8]. Significant redundancy of data is also observed [9]. To address these anomalies in energy readings of smart homes, an Artificial Neural Network-based data imputation is performed [10]. However, this method leads to the creation of missing data anomalies [11]. This inconsistency further leads to variance in the data. Usually, variance is characterized as high variance and low variance. The high variance can impact the data analysis and concluding the data. This also will have a significant impact while building machine learning models that lead to issues such as complexity, overfitting, etc.

In such a case, the analysis of day-to-day energy-capturing is essential to observe the variance in the energy data. Thus, this paper performs the energy data variance analysis. The Analysis of Variance (ANOVA) is a powerful statistical method used to examine the differences among group means and determine whether these differences are statistically significant or not. When applied to smart home energy consumption data, it helps in identifying key factors that affect energy usage and enables more informed decision-making for energy management strategies. The basic idea is to partition the total variance observed in the data into components i.e., between-group variance and within-group variance. Several advantages such as comparing more than two groups, handling multiple factors, effective control of Type-I error, decomposing variability, and better interpretability of ANOVA allow its implementation than other techniques in different applications. The state-of-the-art literature works that discuss the implementation of ANOVA in various applications are given in Table 1.

Table 1. State-of-the-art literature works.

Ref	Year	Description of the Work
[12]	2022	The optimization of the thermostatic generator was achieved by ANOVA and response surface methodology.
[13]	2021	ANOVA was implemented to examine ceiling/floor effects.
[14]	2024	The combination of ANOVA and gradient-boosting was discussed for recognizing load in home energy systems.
[15]	2020	A comprehensive review of ANOVA simultaneous component analysis was conducted to understand its working.
[16]	2024	The Taguchi-ANOVA method was performed to carry out the parametric study on the microgrid energy.
[17]	2019	The Taguchi-ANOVA method was implemented to verify the performance optimization of refrigeration systems.
[18]	2024	The Taguchi-ANOVA method was implemented to optimize the phase frequency detector.
[19]	2022	ANOVA method was conducted to quickly assess the leakage near cryptographic modules.
[20]	2021	ANOVA was applied for fault diagnosis by observing the mean and standard deviation in the variable frequency drive-fed induction motors.
[21]	2024	ANOVA method was conducted for water quality assessment in Lake Burullus, Egypt.
[22]	2022	One-way ANOVA was performed to improve the evaluation of groundwater metals.
[23]	2022	One-way ANOVA was conducted to improve the Arabic tweets.
[24]	2022	ANOVA was applied to identify the adaptor proteins which play a key role in lymphocyte activation control.
[25]	2022	ANOVA method was implemented to optimize the thermoelectric cooler-based dehumidification system.

The abovementioned literature works showcase a wide range of implementations and the importance of ANOVA in various applications such as energy, water quality, metals, Arabic tweets, and healthcare. Although there are works on energy applications, they didn't focus on the variance analysis of the smart home energy consumption data. Hence, this paper contributes to calculating the day-wise variance analysis in each month of the energy consumption data of a smart city, Darmstadt in Germany.

2. Materials and Methods

This section discusses the dataset details and methodology used for the implementation of the proposed work.

2.1. Dataset Description

The Tracebase dataset [26] is a significant resource in the field of energy consumption research, particularly in the context of smart homes and building energy management. The smart home energy consumption was recorded in Darmstadt city of Germany. This dataset provides appliance-level energy consumption data and is invaluable for researchers, policymakers, and industry professionals aiming to understand and improve energy efficiency and sustainability in residential buildings. Comprehensive information on the Tracebase dataset is discussed in [2].

2.2. Methodology

The Analysis of Variance ANOVA is shown in Figure 1. The ANOVA analysis of Variance can be implemented in both one-way and two-way. Using One-way ANOVA, it can be performed only on a single independent variable while in two-way, the variance can be found using two or more independent variables. In the proposed methodology, ANOVA is implemented in one-way as the present study considers only one independent variable "Day" for conducting the monthly data variance analysis. Again the one independent variable can be applied on both (i) between the months, where the variance can be observed on different days, and (ii) within the month, where the variance can be observed on the same days/repeated days. The implementation involves three steps as described follows.

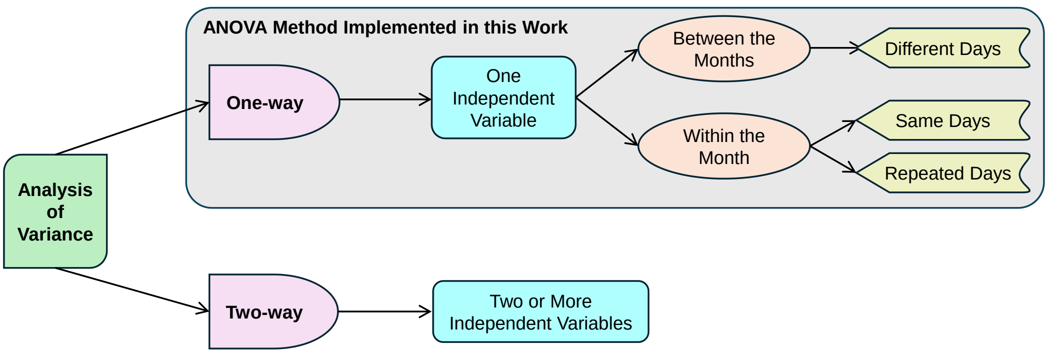


Figure 1. Types of ANOVA.

- Step-1: Finding the Variance:

The process starts with reading the dataset. To implement the proposed method, read the Tracebase dataset and read the monthly data readings "data_month <- read.csv". The data that is stored in the files can be stored in the comma-separated values format (.csv). Once the data is ready, implement the one-way ANOVA to find the variance in the data where the Reading value is a dependent variable, and the Day-wise is an independent variable, it will be converted as a main factor "data_month$date ← data_month$date". Then One-way ANOVA is performed "anova_result_day <- aov (No of readings – Date, data = data_month)" using the function "aov" resulting "anova_result_day".

- Step-2: Extracting F and p Values:

It calculates two variables, namely *F* and *p* values. These are the components that tell whether the ANOVA is statistically significant or not. The value of *p* is always depending on the value of *F*. The *F* value is the ratio of the variances between the variables and here it checks for the condition $p < 0.05$ "if (summary (anova_result _day){{1}}$'P (>F)'[1] < 0.05".

- Step-3: Tukey HSD Post-hoc Test:

If the variables in the summary of ANOVA are statistically significant where the *p*-value is less than the significant value ($p < 0.05$) then ANOVA determines that at least one group differs, but it doesn't provide the details of that group. Hence, the Tukey Honest Significant Difference (HSD) post-hoc test is performed, which helps in deciding which groups are different from each other in the data "Tukey _result <- Tukey HSD (anova_result_day)". This compares all probable pairs of group means. Further, it reduces the likelihood of incorrectly finding substantial differences. If the difference between the group means is higher than the Tukey HSD value, it considers the difference statistically significant" tukey_results[[as.character(data_month)]] <- tukey_result". Based on the differences observed, the day-wise variance analysis in all months is calculated by "print (paste ("Tukey HSD results for Day-wise variance", data_month)) and "Print (tukey_result)". If the variables are not statistically significant then the process is stopped. The workflow of the proposed methodology is observed in Figure 2.

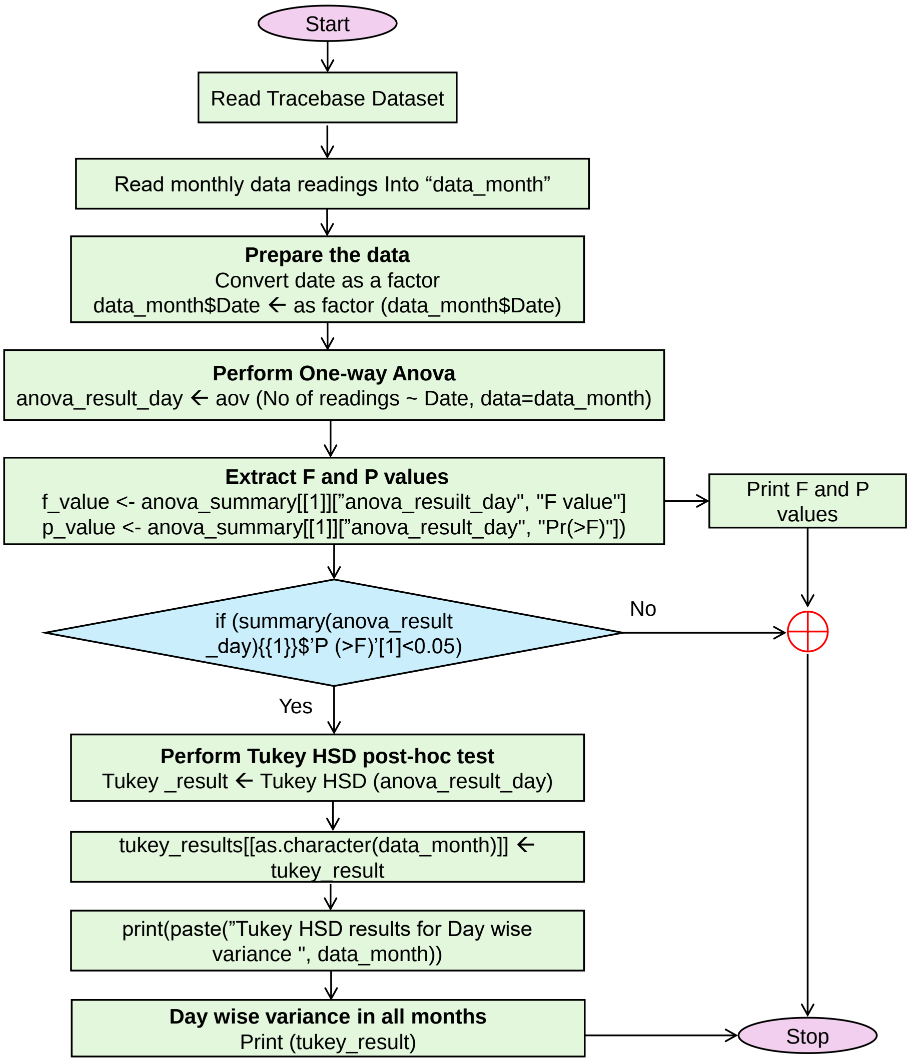
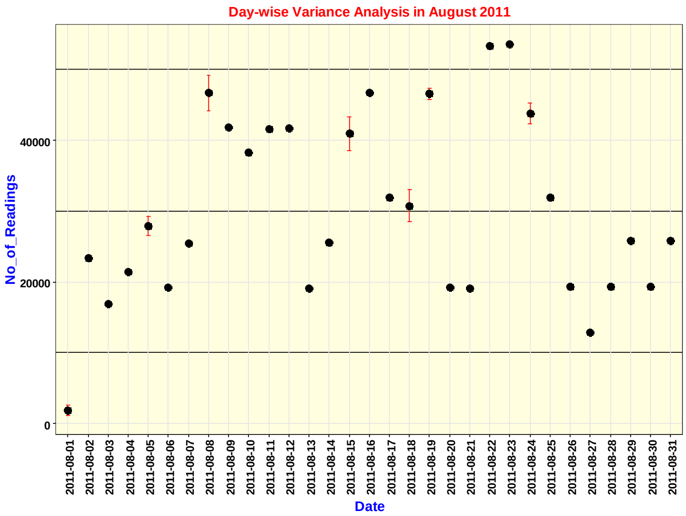


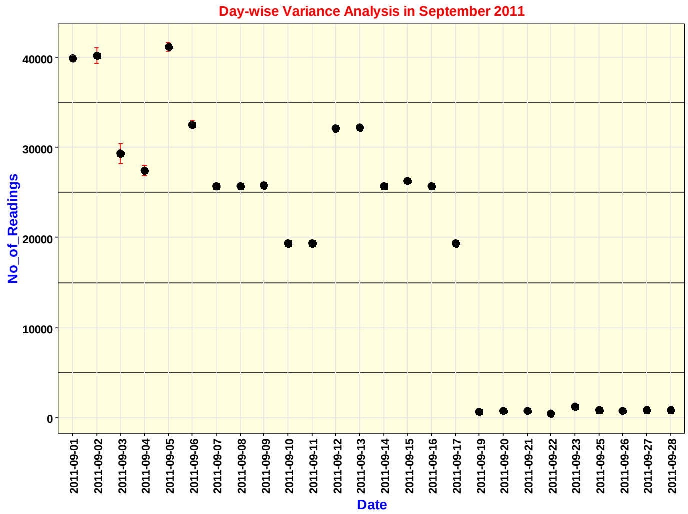
Figure 2. Workflow of the proposed methodology.

3. Results and Discussion

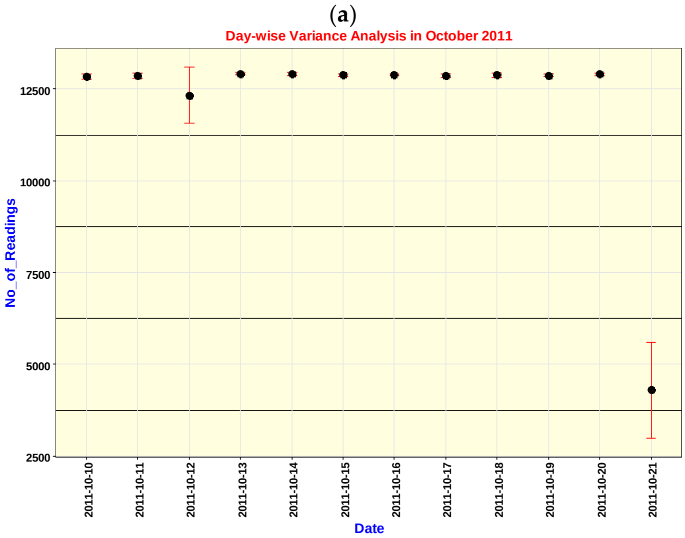
The results of the day-wise variance analysis are presented in Figure 3a–j. The date information is taken on the x-axis and the number of readings is taken on the y-axis. The presence of variance in the data is observed based on the error bar in the graph. The length of this bar (a vertical line on the data point) signifies the highest/lowest variances.



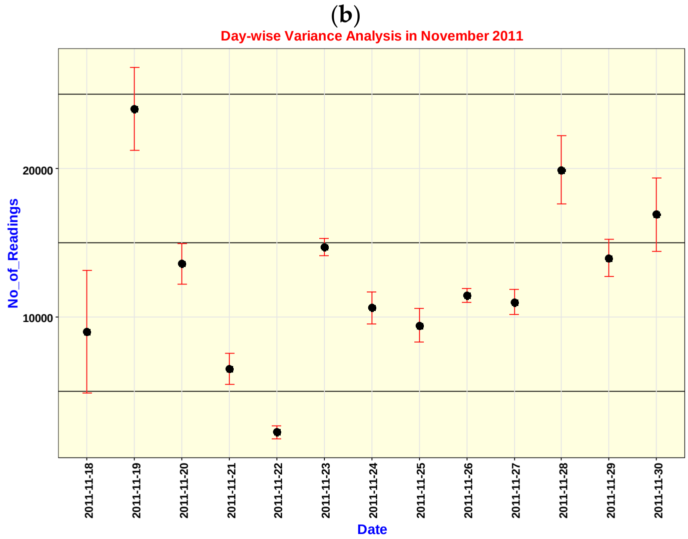
(**a**)



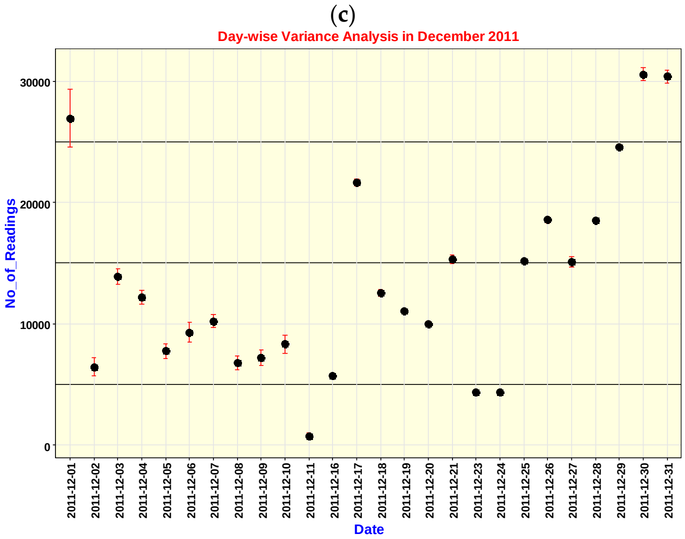
(**b**)



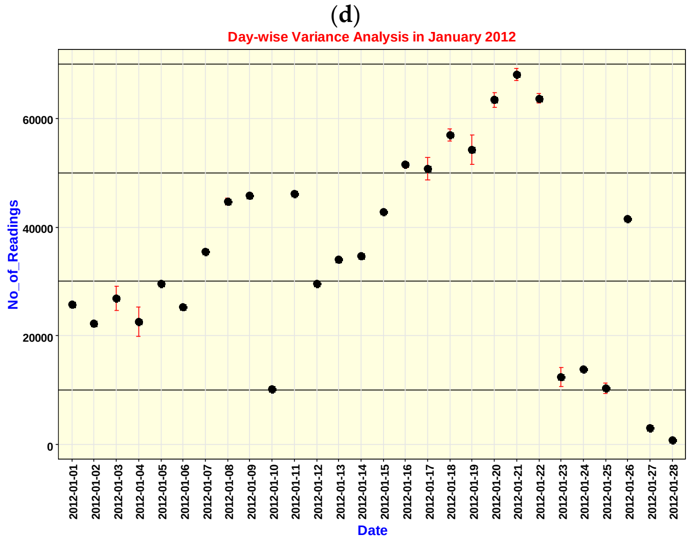
(**c**)



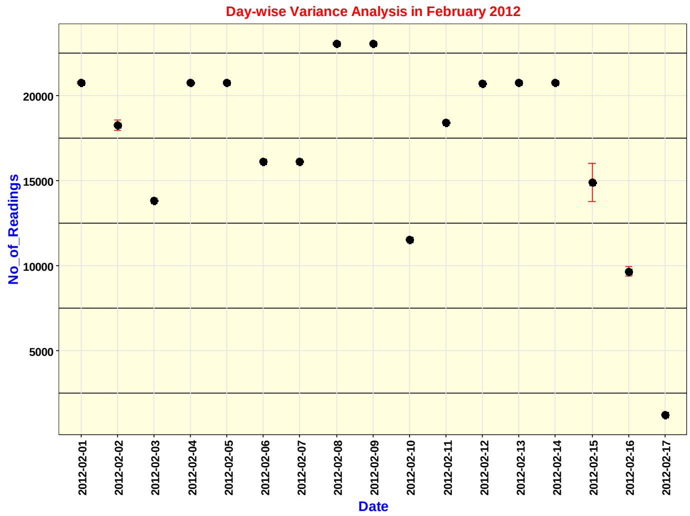
(**d**)



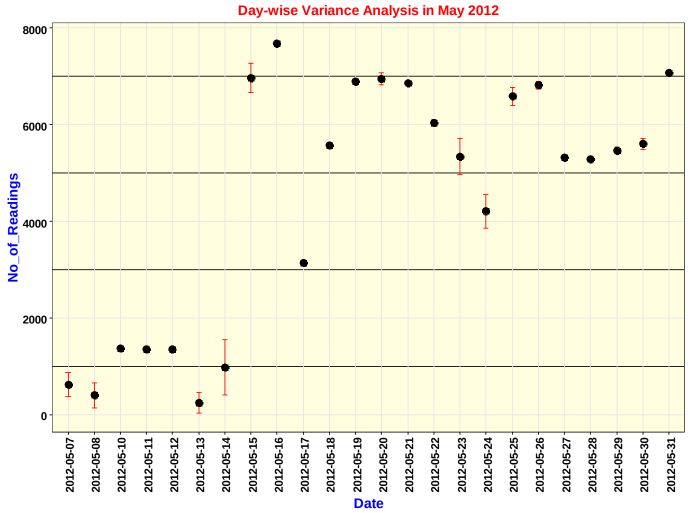
(**e**)



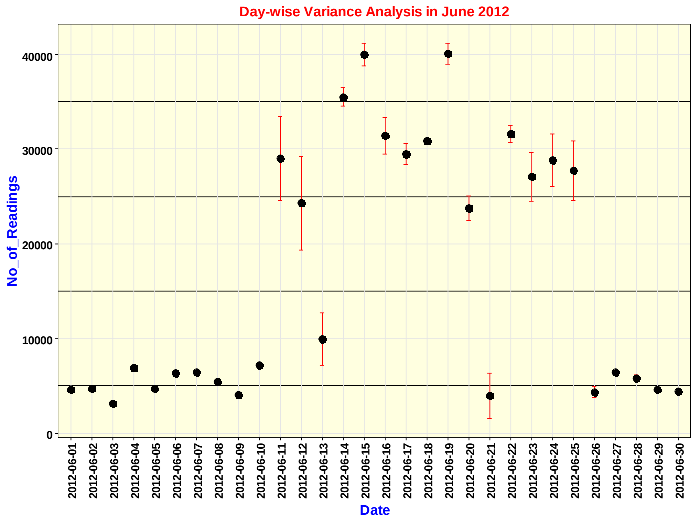
(**f**)



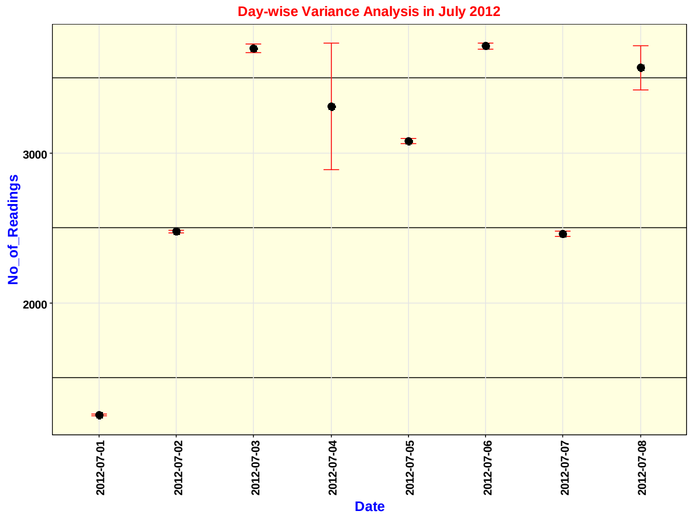
(**g**)



(**h**)



(**i**)



(**j**)

Figure 3. Day-wise variance analysis in various months: (**a**) August 2011; (**b**) September 2011; (**c**) October 2011; (**d**) November 2011; (**e**) December 2011; (**f**) January 2012; (**g**) February 2012; (**h**) May 2012; (**i**) June 2012; (**j**) July 2012.

In Figure 3a, day-wise variance is observed in August 2011 on the days 1 August 2011, 5 August 2011, 8 August 2011, 15 August 2011, 18 August 2011, and 24 August 2011 in which the highest variance is observed on 8 August 2011, 15 August 2011, 18 August 2011 and the lowest variance is observed on 1 August 2011. In Figure 3b, day-wise variance is observed on 2 September 2011, 3 September 2011, 4 September 2011, 5 September 2011, and 6 September 2011 in which the highest variance is observed on 3 September 2011 and the lowest variance is observed on 6 September 2011. In Figure 3c, day-wise variance is observed on all days from 10 October 2011 to 21 October 2011 in which the highest variance is observed on 12 October 2011 and the lowest variance is observed on all days, and medium variance is observed on 12 October 2011. In Figure 3d, day-wise variance is observed on all the days from 18 November 2011 to 30 November 2011, in which the highest variance is observed on 18 November 2011 and the lowest variance is observed on 22 November 2011. In Figure 3e, day-wise variance is observed on days 1 December 2011 to 10 December 2011, 21 December 2011, 27 December 2011, 30 December 2011, and 31 December 2011 in which the highest variance is observed on 1 December 2011 and the lowest variance is observed on 27 December 2011. In Figure 3f, day-wise variance is observed on 3 January 2012, 4 January 2012, 17 January 2012 to 23 January 2012, and 25 January 2012 in which the highest variance is observed on 4 January 2012 and 19 January 2012 and lowest variance is observed on 22 January 2012. In Figure 3g, day-wise variance is observed on 2 February 2012, 15 February 2012, and 16 February 2012 in which the highest variance is observed on 15 February 2012 and the lowest variance is observed on 2 February 2012 and 16 February 2012. In Figure 3h, day-wise variance is observed on 7 May 2012, 8 May 2012, 13 May 2012 to 15 May 2012, 20 May 2012, 23 May 2012 to 25 May 2012, 26 May 2012, 30 May 2012 in which highest variance is observed on 14 May 2012 and lowest variance is observed on 26 May 2012. In Figure 3i, day-wise variance is observed on 11 June 2012 to 17 June 2012, 19 June 2012, to 26 June 2012 in which the highest variance is observed on 12 June 2012 and the lowest variance is observed on 26 June 2012. In Figure 3j, day-wise variance is observed on all days where the highest variance is observed on 4 July 2012 and the lowest variance is observed on 1 July 2012. From this day-wise variance analysis, it is observed that November 2011 and July 2012 have the variance in all days. Further, this enables us to find the variance in the hours of the day.

4. Conclusions

To understand the degree of variability in smart home energy consumption, this paper implemented the Analysis of Variance (ANOVA) on a well-known smart home dataset named "Tracebase". This dataset is a case study of Darmstadt smart city in Germany. The proposed approach has successfully identified the day-wise variances using ANOVA and post-hoc test i.e., Tukey's Honest Significant Differences (HSD) in the considered smart home energy consumption data. The salient remarks of the analysis are as follows:

- It is observed that by performing one-way ANOVA, the variance was found from August 2011 to December 2011 and in January 2012, February 2012, May 2012 to July 2012.
- November 2011 and July 2012 have the variance in all days.
- The identification of this variance in energy consumption is useful in energy consumption pattern finding, forecasting, financial risk, decision-making, and several other grid functions, thereby providing valuable insights for optimizing energy. Further, the statistically significant variances determine specific distributions or anomalies in the data.

Hence, these insights are valuable for energy management and designing effective energy-saving strategies.

Author Contributions: Conceptualization, Y.K. and Y.V.P.K.; methodology, Y.K.; software, Y.K.; validation, Y.V.P.K.; formal analysis, Y.V.P.K.; investigation, Y.K.; resources, Y.K.; data curation, Y.K.; writing—original draft preparation, Y.K.; writing—review and editing, Y.V.P.K.; visualization, Y.K.; supervision, Y.K.; project administration, Y.V.P.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data considered for this study are openly available in [The Tracebase Data Set] at [http://www.tracebase.org] or [26].

Conflicts of Interest: The author declares no conflict of interest.

References

1. Bae, C.; Lee, J.Y.; Kim, D.; Chun, C. Understanding Residents' Energy Consumption Profiles: Behavioral Segmentation Analysis of Heat Energy Consumption in Korean Apartment Residences. *Energy Build.* **2024**, *319*, 114482. https://doi.org/10.1016/j.enbuild.2024.114482.

2. Purna Prakash, K.; Pavan Kumar, Y.V.; Reddy, Ch.P.; Pradeep, D.J.; Flah, A.; Alzaed, A.N.; Al Ahamdi, A.A.; Ghoneim, S.S.M. A Comprehensive Analytical Exploration and Customer Behaviour Analysis of Smart Home Energy Consumption Data with a Practical Case Study. *Energy Rep.* **2022**, *8*, 9081–9093. https://doi.org/10.1016/j.egyr.2022.07.043.
3. Priyadarshini, I.; Sahu, S.; Kumar, R.; Taniar, D. A Machine-Learning Ensemble Model for Predicting Energy Consumption in Smart Homes. *Internet Things* **2022**, *20*, 100636. https://doi.org/10.1016/j.iot.2022.100636.
4. Gellert, A.; Fiore, U.; Florea, A.; Chis, R.; Palmieri, F. Forecasting Electricity Consumption and Production in Smart Homes through Statistical Methods. *Sustain. Cities Soc.* **2022**, *76*, 103426. https://doi.org/10.1016/j.scs.2021.103426.
5. Chaturvedi, S.; Rajasekar, E.; Natarajan, S.; McCullen, N. A Comparative Assessment of SARIMA, LSTM RNN and Fb Prophet Models to Forecast Total and Peak Monthly Energy Demand for India. *Energy Policy* **2022**, *168*, 113097. https://doi.org/10.1016/j.enpol.2022.113097.
6. Chen, C.; Zarazua De Rubens, G.; Xu, X.; Li, J. Coronavirus Comes Home? Energy Use, Home Energy Management, and the Social-Psychological Factors of COVID-19. *Energy Res. Soc. Sci.* **2020**, *68*, 101688. https://doi.org/10.1016/j.erss.2020.101688.
7. Kasaraneni, P.P.; Pavan Kumar, Y.V. A Systematic Approach for Exploration, Behavior Analysis, and Visualization of Redundant Data Anomalies in Smart Home Energy Consumption Dataset. *Int. J. Renew. Energy Res.* **2022**, *12*, 109–123. https://doi.org/10.20508/ijrer.v12i1.12613.g8381.
8. Kasaraneni, P.P.; Pavan Kumar, Y.V. Analytical Approach to Exploring the Missing Data Behavior in Smart Home Energy Consumption Dataset. *J. Renew. Energy Environ.* **2022**, *9*, 37–48. https://doi.org/10.30501/jree.2021.313536.1277.
9. Kasaraneni, P.P.; Pavan Kumar, Y.V.; Moganti, G.L.K.; Flah, A. Analytical Enumeration of Redundant Data Anomalies in Energy Consumption Readings of Smart Buildings with a Case Study of Darmstadt Smart City in Germany. *Sustainability* **2022**, *14*, 10842. https://doi.org/10.3390/su141710842.
10. Purna Prakash, K.; Kumar, Y.V.P.; Ravindranath, K.; Pradeep Reddy, G.; Amir, M.; Khan, B. Artificial Neural Network-Based Data Imputation for Handling Anomalous Energy Consumption Readings in Smart Homes. *Energy Explor. Exploit.* **2024**, *42*, 1432–1449. https://doi.org/10.1177/01445987231221877.
11. Kasaraneni, P.P.; Kumar, Y.V.P. Simple and Effective Descriptive Analysis of Missing Data Anomalies in Smart Home Energy Consumption Readings. *J. Energy Syst.* **2021**, *5*, 199–220. https://doi.org/10.30521/jes.878318.
12. Chen, W.-H.; Carrera Uribe, M.; Kwon, E.E.; Lin, K.-Y.A.; Park, Y.-K.; Ding, L.; Saw, L.H. A Comprehensive Review of Thermoelectric Generation Optimization by Statistical Approach: Taguchi Method, Analysis of Variance (ANOVA), and Response Surface Methodology (RSM). *Renew. Sustain. Energy Rev.* **2022**, *169*, 112917. https://doi.org/10.1016/j.rser.2022.112917.
13. Liu, Q.; Wang, L. T-Test and ANOVA for Data with Ceiling and/or Floor Effects. *Behav. Res.* **2021**, *53*, 264–277. https://doi.org/10.3758/s13428-020-01407-2.
14. Cabral, T.W.; Neto, F.B.; De Lima, E.R.; Fraidenraich, G.; Meloni, L.G.P. Analysis of Variance Combined with Optimized Gradient Boosting Machines for Enhanced Load Recognition in Home Energy Management Systems. *Sensors* **2024**, *24*, 4965. https://doi.org/10.3390/s24154965.
15. Bertinetto, C.; Engel, J.; Jansen, J. ANOVA Simultaneous Component Analysis: A Tutorial Review. *Anal. Chim. Acta X* **2020**, *6*, 100061. https://doi.org/10.1016/j.acax.2020.100061.
16. Ma, X.; Joybari, M.M.; Luo, J.; Liu, J.; Lai, K. Conducting a Parametric Study on Microgrid Energy and Economic Performance with System Optimization Using Taguchi-ANOVA Approach. *J. Energy Storage* **2024**, *97*, 112960. https://doi.org/10.1016/j.est.2024.112960.
17. Canbolat, A.S.; Bademlioglu, A.H.; Arslanoglu, N.; Kaynakli, O. Performance Optimization of Absorption Refrigeration Systems Using Taguchi, ANOVA and Grey Relational Analysis Methods. *J. Clean. Prod.* **2019**, *229*, 874–885. https://doi.org/10.1016/j.jclepro.2019.05.020.
18. Sharma, J.; Ahmad, R.; Yadav, A.; Varma, T.; Boolchandani, D. Design and Optimization of Phase Frequency Detector through Taguchi and ANOVA Statistical Techniques for Fast Settling Low Power Frequency Synthesizer. *Integration* **2024**, *96*, 102162. https://doi.org/10.1016/j.vlsi.2024.102162.
19. Iyer, V.V.; Yilmaz, A.E. An ANOVA Method to Rapidly Assess Information Leakage Near Cryptographic Modules. *IEEE Trans. Electromagn. Compat.* **2022**, *64*, 915–929. https://doi.org/10.1109/TEMC.2022.3157664.
20. Shabbir, M.N.S.K.; Liang, X.; Chakrabarti, S. An ANOVA-Based Fault Diagnosis Approach for Variable Frequency Drive-Fed Induction Motors. *IEEE Trans. Energy Convers.* **2021**, *36*, 500–512. https://doi.org/10.1109/TEC.2020.3003838.
21. Hassan, A.; Samy, G.; Hegazy, M.; Balah, A.; Fathy, S. Statistical analysis for water quality data using ANOVA (case study–Lake Burullus Influent Drains). *Ain Shams Eng. J.* **2024**, *15*, 102652. https://doi.org/10.1016/j.asej.2024.102652.
22. Khan, I.; Umar, R. Improving Evaluation of Groundwater Heavy Metal(Loid)s Pollution Efficiencies: Insights from Novel Shannon Entropy-Weight and One-Way ANOVA Analysis. *Groundw. Sustain. Dev.* **2024**, *24*, 101052. https://doi.org/10.1016/j.gsd.2023.101052.
23. Alassaf, M.; Qamar, A.M. Improving Sentiment Analysis of Arabic Tweets by One-Way ANOVA. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 2849–2859. https://doi.org/10.1016/j.jksuci.2020.10.023.
24. Wang, Y.-H.; Zhang, Y.-F.; Zhang, Y.; Gu, Z.-F.; Zhang, Z.-Y.; Lin, H.; Deng, K.-J. Identification of Adaptor Proteins Using the ANOVA Feature Selection Technique. *Methods* **2022**, *208*, 42–47. https://doi.org/10.1016/j.ymeth.2022.10.008.

25. Eltaweel, M.; Heggy, A.H.; Yaseen, Z.M.; Alawi, O.A.; Falah, M.W.; Hussein, O.A.; Ahmed, W.; Homod, R.Z.; Abdelrazek, A.H. Application of the ANOVA Method in the Optimization of a Thermoelectric Cooler-Based Dehumidification System. *Energy Rep.* **2022**, *8*, 10533–10545. https://doi.org/10.1016/j.egyr.2022.08.193.
26. The Tracebase Data Set. Available online: http://www.tracebase.org (accessed on 21 November 2024).