COMBED DATA SET ANALYSIS

RUST GRAPHICAL ANALYSIS IMPLEMENTATION PROJECT - WRITE UP

> **DS 210 / ABDUL RAFAY / COLLABORATORS : None.**

DATA SET OF CHOICE : Commercial Building Energy Dataset (COMBED)

> https://combed.github.io/

GOAL OF THE PROJECT : Implement graphical analysis using RUST language to deduce insights on various energy consumption parameters' relationship in the context of well documented academic research building.

WHY COMBED? → real world relevance, detailed energy consumption patterns and research paper context.

Upon short listing the links containing various datasets, I was interested in seeing what, COMBED a commercial building energy dataset contained. It's description was interesting, with dataset demonstrating some unique qualities, hence why I decided to include it.

- COMBED is the first of it's kind energy dataset, where commercial building energy data was sampled more than once per minute. → high frequency sampling, suitable for implementing higher grid separation which requires large dataset.
- There are **200 smart meters** recording various power related parameters every 30 seconds, in **IIITD's academic building**. → smart meters, potentially can provide us with rich data with multiple dimensions of data. Also the academic setting of the building could yield interesting insights related to energy consumptions in academic settings.
- The meters are Schneider electric based and the data collection system consists of RaspberryPi based controller and sMAP for data visualization and archiving. → another technical hint of how versatile and richly the data is collected.
- On top of that there were some other perks, such as their entire analysis scripts being **open sourc**e and accessible on **Github**, potentially providing us with tools for much more in-depth analysis of our interest. → open Github dataset, a valuable resource which could potentially become foundation for my data analysis.
- Also, COMBED was released as a part of following paper : **A comparison of non-intrusive load monitoring methods for commercial and residential buildings** Nipun Batra, Oliver Parson, Mario Berges, Amarjeet Singh, Alex Rogers. → the fact that the dataset was published with research paper, adds further credibility to the dataset, on which we can build on top for analysis according to our goals.

So, overall the versatility and tools provided with the dataset, which allows to explore various aspects of energy consumption in a academic setting, became my main reason in choosing this dataset. I'm pretty hopeful in potentially finding trends, relationships between parameters or anomalies in the context of energy consumption of academic building.

INSIGHTS OF INTEREST (as of proposal date):

- Identify peak energy consumption periods.
- Explore the relationships between different meters and power parameters.
- Investigate patterns in energy consumption associated with specific building systems or areas.
- Assess the overall energy efficiency of the academic building.

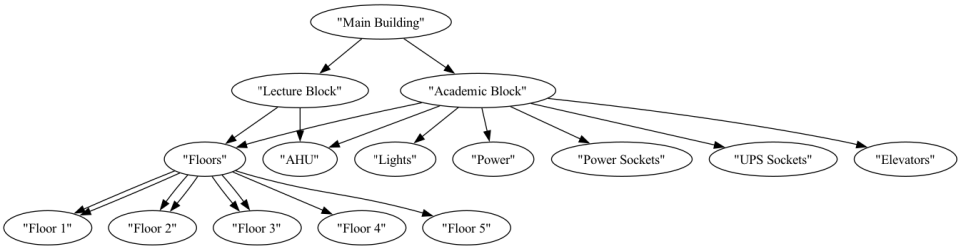
DATA LOADING/PREPROCESSING

First of it's kind dataset, **COMBED**, is truly unique dataset in it's data collection. It leveraged from newly constructed academic campus, where the researchers deployed 200 smart energy meters, by Schneider electric. Each of them were capable of multi-

dimensional data collection, with each meter collecting parameters of **current, energy and power** consumptions. Data was collected over a span of month, with each data point being collected **every 30 seconds by each of 200 meters**. This generated huge amount of data, with potential of valuable insights regarding energy consumption patterns of a commercial academic building. The dataset was released by the researchers, alongside their research paper, which contained 75 individual **.csv** files. So, I tried to organize them into one document using Excel, which I was able to do, but due to the complexity and size of the file, was not able to use it whole for analysis. **The merged file, contained 150 columns, with 4 levels of header rows, and 86000 rows of data.** Which is where I decided to first understand the relationship of these files, of which I created a datasets' hierarchy network diagram, using RUST. Result was as following. (code for network creation is in `main.rs`)

NETWORK OF DATASET HIERARCHY:

- Nodes represent entities (e.g., buildings, floors, appliances).
- Directed edges represent relationships or connections between entities.



Upon looking at the network, it was easier to understand the hierarchy of datasets. But the overlapping of the sub datasets, was the most valuable insight from the graph. I was then able to focus on three potential analysis topics which I decided to work on.

- **FOR OVERALL ANALYSIS:**
  - To assess overall energy efficiency, identify peak consumption periods, I focused on the data files with overlap by both lecture and academic blocks. As they seemed to be most reasonable option to look at the general trend. Overlap of topic was in following sub-datasets:
    - Building Total Main → each with `energy.csv, current.csv, power.csv`
- **FOR APPLIANCE-LEVEL ANALYSIS:**
  - I was then also interested in understanding the consumption patterns of air handling units (AHU), which also showed an overlapping by both lecture and academic blocks. From which I hoped to deduce valuable trend of appliance's energy consumption pattern in both blocks:
    - AHU → also each with `energy.csv, current.csv, power.csv`
- **FOR DETAILED APPLIANCE LEVEL ANALYSIS:**
  - Lastly I wanted to seem how specific floor's energy consumption trends differed in both blocks. I performed comparison for floors 1, 2, and 3. Academic block had two additional floors, but that wasn't of interest as we were focusing on overlapping regions for analysis.
    - Floor Levels→ also each with `energy.csv, current.csv, power.csv`

GRAPH ALGORITHM - KMeans CLUSTERING (HISTOGRAM)

Also for each analysis, **KMeans clustering** was implemented using RUST as needed. It was implemented mainly for **Pattern Recognition.** Because it is effective in identifying patterns and we could leverage from it potentially grouping the similar energy consumption profiles together. Which could be helpful in our anticipated insights of peak energy consumption hours, low-demand periods, or any potentially distinct consumption patterns. y-axis is number of clusters, while x-axis is the energy consumption value.

FORMAT OF RESULTS

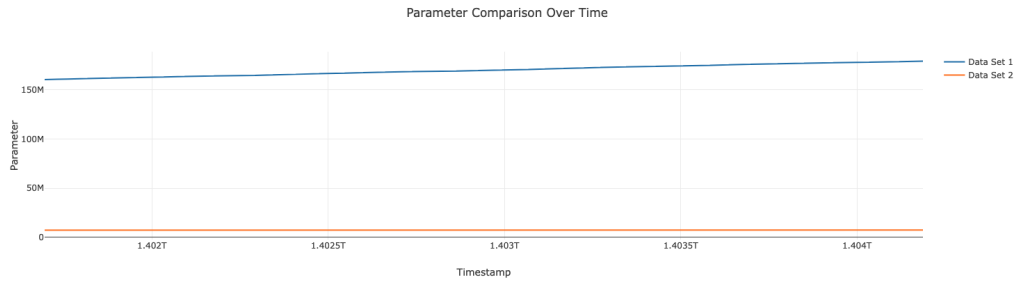
For each of the analysis performed, the two-line graph was created for comparison of overall trend of datasets of interest. Though images of the graphs are inserted, link for originally interactive plot is also shared below the graph images. Below the graph is

graphs appendix explaining the units, and labels of the graph. Following is basic statistical calculations computed using RUST. ARIMA analysis was performed for each dataset and is included in the program code, but the it was omitted in the writeup, because of too many values being generated. lastly KMeans clustering algorithm, was performed, visualized with histogram. For each implementation of algorithm, clusters were set to 10, with 1,000 iterations.

For each graph :

- **Parameter :** Energy Consumption (kWh)
- **Timestamp :** total of 1 month in which the data was collected, (per 30s).

BUILDING TOTAL MAIN ENERGY CONSUMPTION TREND ANALYSIS:



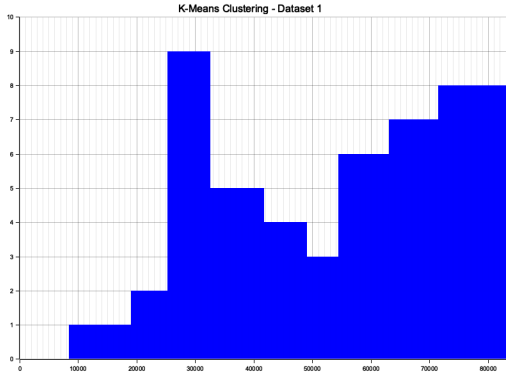
> **Interactive plot link:**
> file:///private/var/folders/37/ngxw00414mb8kkyzf41yl35m0000gn/T/plotly_VDx95i9u5uFquPReriMzcd.ht

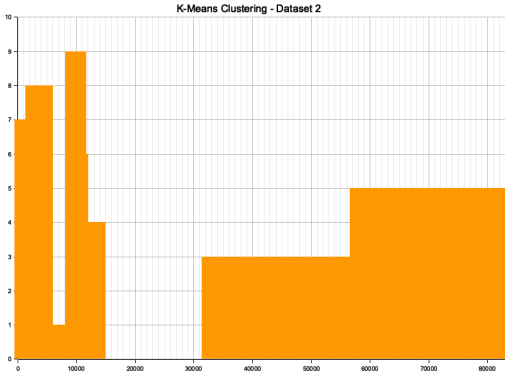
Data set 1 : Academic Block's Total Main Energy Consumption

- Mean 1: 169953229.75581393
- Variance 1: 30275976191574.95
- Std. Deviation 1: 5502360.965219835

Data set 2: Lecture Block's Total Main Energy Consumption

- Mean 2: 7411109.836887946
- Variance 2: 1059775244.5944803
- Std. Deviation 2: 32554.189355511226

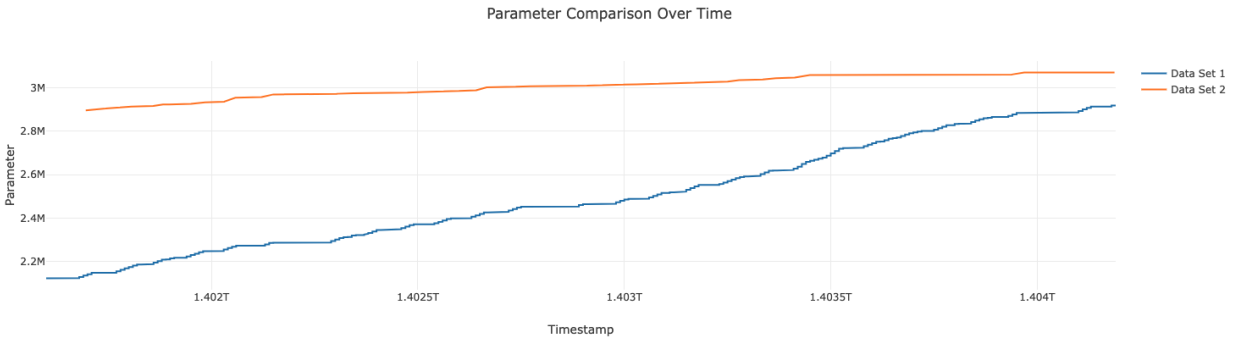




INSIGHTS:

- Total main energy consumption of Academic block shows a steady increase over time.
- Whereas, total main energy consumption of lecture block seems to be constant over time.
- Even from the beginning, Academic block's energy consumption was significantly large compared to lecture block, which does work well with academic block being equipped with heavy consumption appliances and equipments compared to lecture block. Mean of the Academic blocks energy consumption is approximately 20 times more than that of lecture block.

AHU ENERGY CONSUMPTION TREND ANALYSIS:



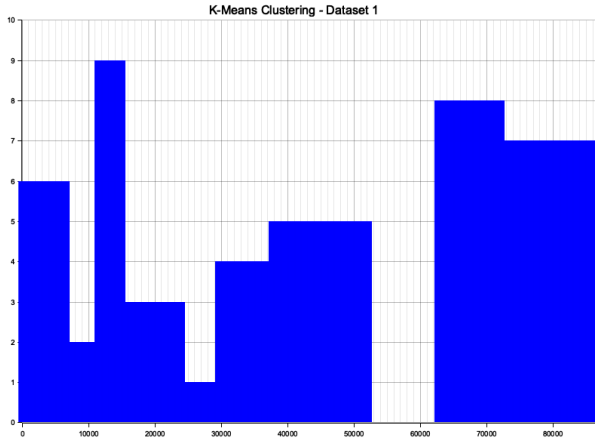
> **Interactive plot link:**
> file:///private/var/folders/37/ngxw00414mb8kkyzf41yl35m0000gn/T/plotly_5SJD5qIzNK2oikqKRCUp31.h

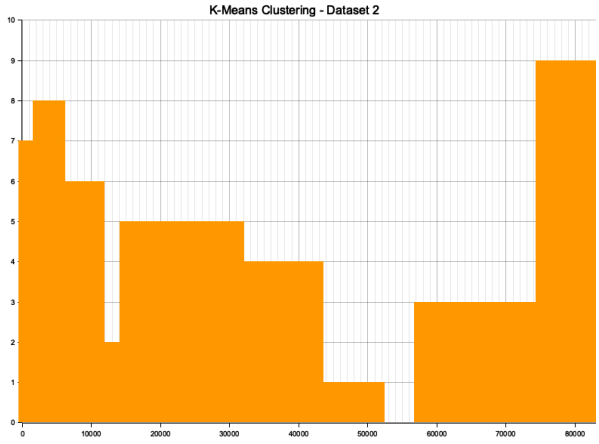
Data set 1 : Academic Block's AHU Energy Consumption

- Mean 1: 2498454.547144756
- Variance 1: 57470942394.74027
- Std. Deviation 1: 239730.97921365997

Data set 2: Lecture Block's AHU Energy Consumption

- Mean 2: 3007465.286190848
- Variance 2: 2453781046.0995636
- Std. Deviation 2: 49535.654291626786





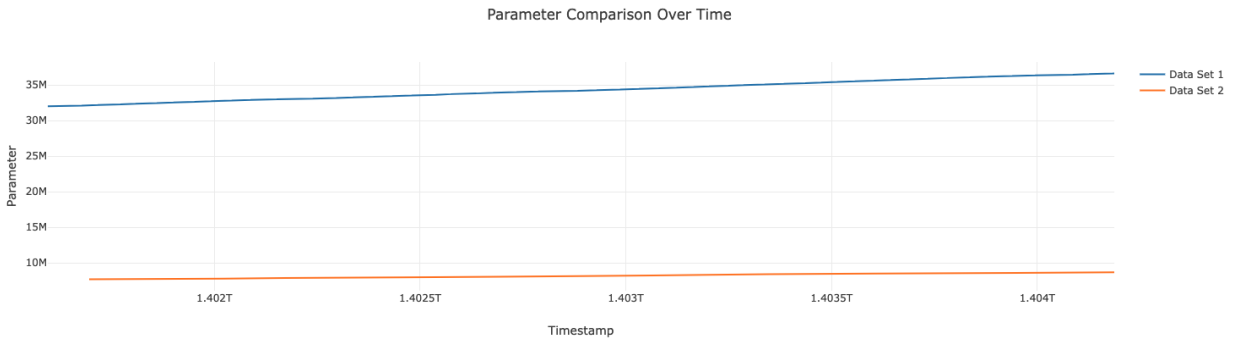
INSIGHTS:

- AHU's energy consumption of academic block shows a steady increase over time. However, unlike total main energy consumption, Academic block have lower energy consumption overall.
- AHU's main energy consumption of lecture block seems to be constant over time.
- At the beginning, Academic block's energy consumption was significantly smaller compared to lecture block, which by the end of data collection period almost reaches lecture block's energy consumption.
- Mean of both energy consumption does not show large difference as before, like in the case of total main energy consumption.
- Clusters seems to be more distinctly created compared to total main energy consumption's case.

- Also, where total main had very large values on y-axis, ranging from 0 to 150, this time it stays in the range of 2.2 to 3.
- Energy consumption of lecture block seems to be in the similar range in both analysis, both showing similar trend approximately constant consumption over time.

FLOOR SPECIFIC ENERGY CONSUMPTION TREND ANALYSIS:

FLOOR 1 :



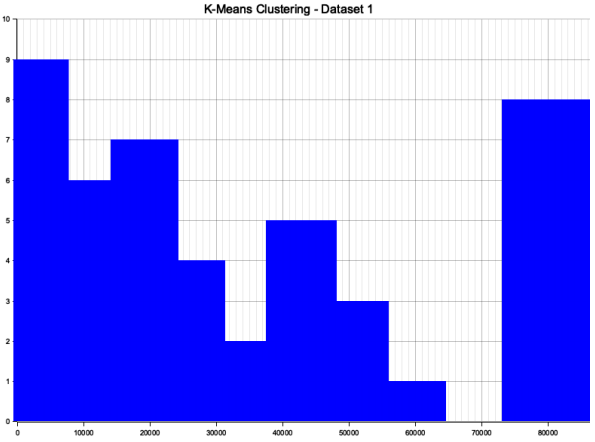
> **Interactive plot link:**
> file:///private/var/folders/37/ngxw00414mb8kkyzf41yl35m0000gn/T/plotly_pYrsLYv7uDyiuw6mJU3WuQ.

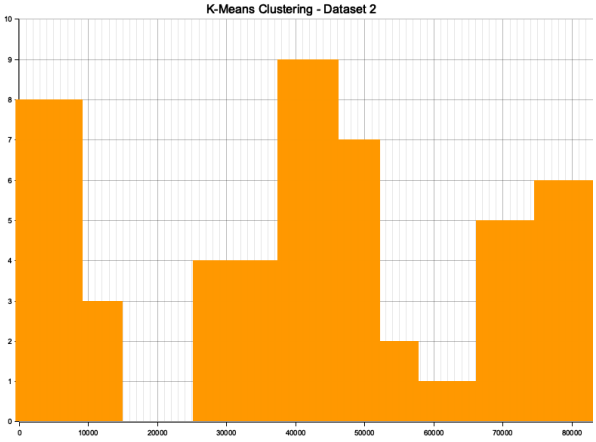
Data set 1 : Academic Block's Floor 1 Energy Consumption

- Mean 1: 34300669.49134484
- Variance 1: 1838901426846.278
- Std. Deviation 1: 1356060.9967277572

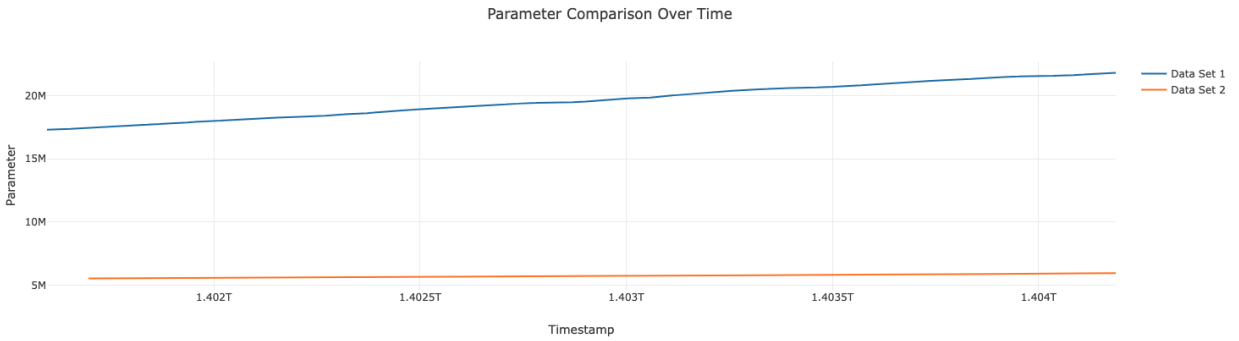
Data set 2: Lecture Block's Floor 1 Energy Consumption

- Mean 2: 8205900.373291382
- Variance 2: 95188919758.79445
- Std. Deviation 2: 308527.0162543216





FLOOR 2 :



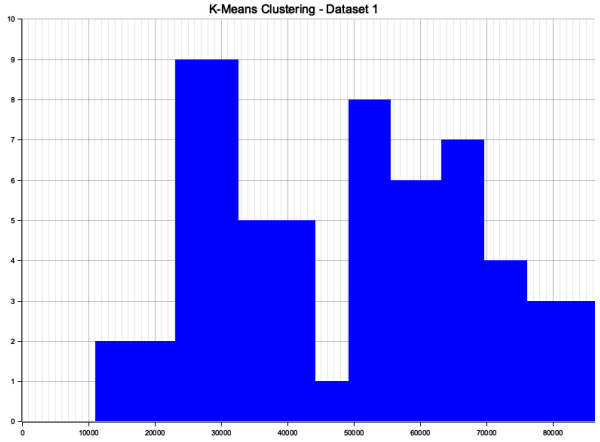
> **Interactive plot link:**
> file:///private/var/folders/37/ngxw00414mb8kkyzf41yl35m0000gn/T/plotly_2119TESpsAFrNorrbhIbS7.ht

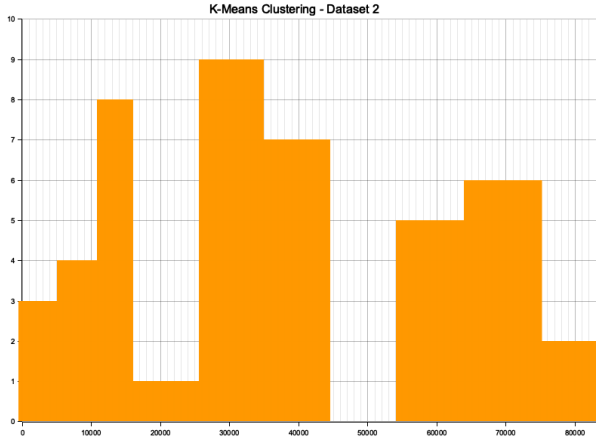
Data set 1 : Academic Block's Floor 2 Energy Consumption

- Mean 1: 19608378.31661991
- Variance 1: 1818309349015.2454
- Std. Deviation 1: 1348447.0137959612

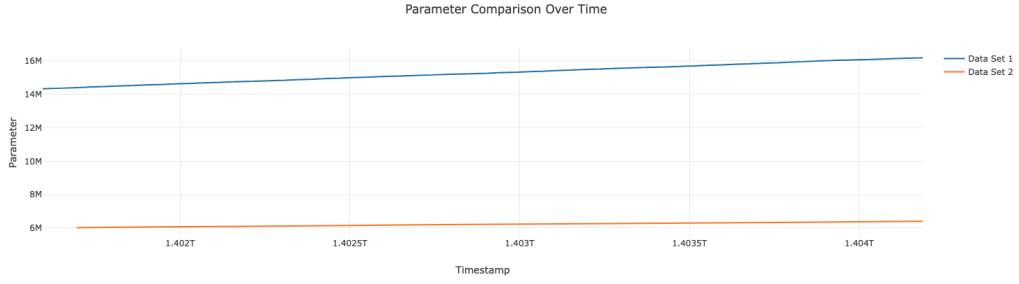
Data set 2: Lecture Block's Floor 2 Energy Consumption

- Mean 2: 5742360.313547334
- Variance 2: 14271887383.241827
- Std. Deviation 2: 119465.00484761981





FLOOR 3 :



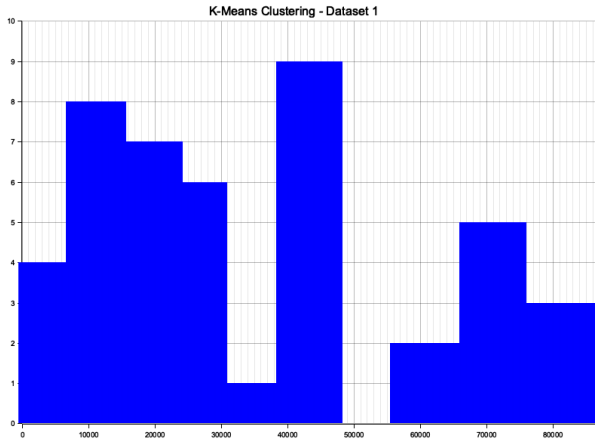
> **Interactive plot link:**
> file:///private/var/folders/37/ngxw00414mb8kkyzf41yl35m0000gn/T/plotly_J1iP5VPSh3kJz0xiN0wugT.ht

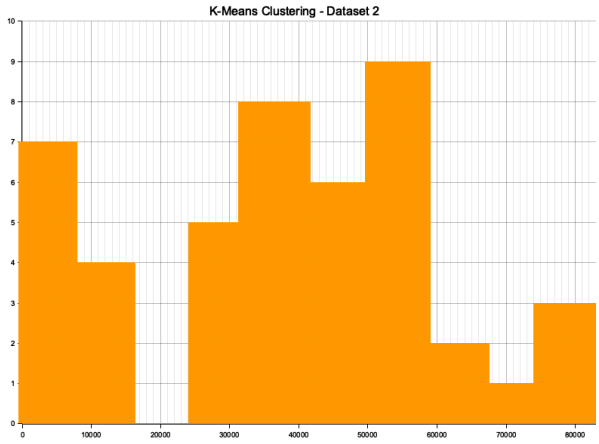
Data set 1 : Academic Block's Floor 3 Energy Consumption

- Mean 1: 15268391.172024565
- Variance 1: 288975573952.5856
- Std. Deviation 1: 537564.4835297302

Data set 2: Lecture Block's Floor 3 Energy Consumption

- Mean 2: 6216584.345770137
- Variance 2: 11775754815.24465
- Std. Deviation 2: 108516.15002037554





INSIGHTS:

- In all three floors academic block had higher energy consumption.
- In all three floors lecture block seems to have similar constant consumption trend in similar range.
- Whereas, floor 1 have significantly higher energy consumption compared to floor 2 and 3, in the case of academic block. Though not very large difference, floor 2's energy consumption is higher compared to floor 3's consumption, again in the case of academic block. Such, trend is also verified through mean values calculated for each.
- In all three academic block's trend there was a steady increase in consumption over time.
- In all three analysis, KMeans graph seemed to have clustered similar to each other for academic and lecture block.

So, I was able to implement such analysis in RUST, implementing a code structure as following:

- Modules
  - `arima_module.rs` → ARIMA based predictions on the datasets
  - `clustering.rs` → KMeans Clustering implementation
  - `csv_reader.rs` → reads two **.csv** files of interest for analysis
  - `graph_creator.rs` → creates a two trend line graph
  - `scatter_plot.rs` → scatter plot for individual dataset
  - `statistical_analysis.rs` → statistical measures, e.g. mean, var., std. dev.
- `main.rs` → all modules tied together, create network visualization, etc

---

Code specific and other details of the project implementation can be found at my public git-hub repository, at following link : https://github.com/Abdul03Rafay/COMBED-ANALYSIS

SOURCES:

- **Git-hub:**

  https://combed.github.io/

- **Research Paper :**

  A comparison of non-intrusive load monitoring methods for commercial and residential buildings Nipun Batra, Oliver Parson, Mario Berges, Amarjeet Singh, Alex Rogers

  BIBTEX @article{batra2014comparison,title={A comparison of non-intrusive load monitoring methods for commercial and residential buildings},author={Batra, Nipun and Parson, Oliver and Berges, Mario and Singh, Amarjeet and Rogers, Alex},year={2014},journal= {arXiv:1408.6595} }

REFERENCES:

- https://arxiv.org/pdf/1408.6595v1.pdf - COMBED study research paper excerpt
- https://rust-ml.github.io/book/3_kmeans.html - RUST KMeans Basics
- https://docs.rs/csv/latest/csv/tutorial/ - RUST CSV module guidelines
- https://docs.rs/plotly/latest/plotly/ - RUST Plotly guidlines
- https://en.wikipedia.org/wiki/Hyperparameter_optimization - Hyperparameter optimization