College: Engineering and Information Technology
Department: Information Technology
Program: Data Analytics

Big Data Technologies course Project

Big Data Analysis of Household Energy Consumption for Climate Change Mitigation

Prepared by:
202011025 Maryam Hesham Raweh Ali

Supervised by:
Prof. Salam Fraihat

Academic Year 2023- 2023 – Fall

LIST OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Big Data Analysis of Household Energy Consumption for Climate Change Mitigation

Maryam Ali
Depatment of Information Technology
Ajman University
Ajman, U.A.E
202011025@ajmanuni.ac.ae

Abstract – the research study concentrates on examining the energy consumption patterns of households with the help of data analytics. Smart meters data from 5,567 London residences involved in the UK Power Networks' Low Carbon London project is applied. This study takes on data preprocessing, feature selection, and machine learning algorithms to segment consumption patterns, break down the electricity load curve, connect consumption data with ACORN classification information, and forecast household electricity consumption. This research shares insights supporting the reduction of climate change and provides sustainable energy solutions.

Keywords—Big data, Analytics, Household Energy Consumption, Smart meters, Climate change, Patterns, Machine Learning, Apache Spark, PySpark, London, Data Preprocessing, ACORN classification.

I. INTRODUCTION

It is now more critical to address the solutions available for more sustainable energy consumption in different households. We are facing a rapidly changing climate and environmental concerns are increasing by the day. A challenge that humans currently face is the task of reducing carbon emission and reducing the impact it has on the earth.

Unfortunately, Household energy consumption plays a significant role in shaping our carbon footprint – so everything from powering our homes to the chargers and activities we do daily. Those actions we take by choice have contributed to widespread consequences for the planet. To address these problems, we must have a deep understanding of how and why energy is consumed at the household level, then search for potential for change.

The smart meter technology emergence has unlocked new opportunities. When gathering data on energy consumption, smart meters provide an unparalleled chance to explore the complexity of energy usage habits. Not only this but when analyzed through big data analytics, this amount of information was able to cover the invaluable insights which aims in revolutionizing the method of energy consumption and climate change mitigation.

II. TASK AND DATA

The dataset which was examined in the data engineering task concentrated on climate change. This data was extracted from Kaggle and contained numerous tables and folders.[1] One of the data key files worked with was the "Halfhourly energy consumption" folder which consisted of numerous files.

To preprocess the data, Kaggle's API has been used to obtain, and download the data relevant for the problem. The raw data had to be combined to form a singular view of each table. After the tables have been achieved with the appropriate structure, PySpark was utilized to perform the cleaning and transformation of the data (this includes tasks like removing missing data and duplicates, as well as defining the appropriate data type and making the data encoded and normalized), in addition to PySpark.ml for the machine learning task and achieve the primary tasks of this project.

A. Primary tasks of this project

i. Analyzing the yearly, monthly, daily, hourly energy consumption patterns.
ii. Observe the relationship between energy consumption and other climate variables.
iii. Explore the possibility of reducing energy consumption and reducing its impact on climate change.
iv. Evaluate a linear regression model using RMSE, MAE, R-squared, and MSLE.

B. Data sources of this project

Table I: Data Sources Information [1]

Source	***Type***	***Size***
Halfhourly dataset	Zip folder, containing block files	7.33 GB
Weather hourly darksky	CSV file	1.9 MB
Information households	Csv file	229 KB
UK bank holidays	CSV file	786 Bytes

The integration of smart meter and weather data enables an extensive analysis of household's energy consumption patterns and its relationship/correlation with the climate variables. Moreover, the ACORN classification allows examining consumption patterns across different demographic groups.

Using A Classification of Residential Neighborhood (ACORN), the UK segments people into 18 categories starting from A-Q that shows how wealthy those people are based on income, housing, and lifestyle. As you go forward with the letters, the group goes from rich to poor. This dataset also adds ACORN-U. [3]

Table II: ACORN Segmentation

Acorn-grouped	***Values***
Affluent	ACORN-A, ACORN-B, ACORN-C, ACORN-E

Comfortable	ACORN-F, ACORN-G, ACORN-H, ACORN-I, ACORN-J
Adversity	ACORN-K, ACORN-L, ACORN-M, ACORN-N, ACORN-O, ACORN-P, ACORN-Q
ACORN-U	ACORN-U
ACORN-	ACORN-

III. METHODOLOGY

This report discusses the pipeline and results of the experience taken place using PySpark, which is a distributed data processing framework. This project is done through the Google Colab environment, following a specific framework to achieve the best results possible. [Figure II]

A. *Big data tools and technologies*

Apache Spark, especially through the python-based API, PySpark, played an important role in the report's pipeline allowing for parallel processing and analysis. Spark DataFrame was used for structured data analysis, PySpark.ml made it possible to apply machine learning tasks in a distributed manner, in addition to Spark SQL and Resilient Distributed Datasets (RDDs) that were used for distributed data processing.

PySpark works in a distributed cluster environment, where there exists one main computer (driver). Within the cluster, there exists more worker computers (nodes). Some of the nodes will store the data (data nodes), whereas others will do the computations (worker nodes). The driver divides the work into smaller tasks, where each task acts like a small job that needs to be executed. The tasks are sent to the worker nodes for processing.

PySpark was chosen as it is great at handling big data by splitting the data into smaller chunks and distributing them on the different data nodes. Each worker node processes their assigned data chunk and sends the results back to the driver. PySpark also caches data in memory for faster access.

Supervisors make sure that the worker nodes are doing their jobs correctly and report any issue back to the driver. For resilience, PySpark by its nature is designed to be resilient. If one worker node falls into an issue, the task can be reassigned to another node.

Size: In reality, the size that is being dealt with is 83,925,602 rows.

Please note that this work has been done using google colab, meaning that it not utilizing multiple worker nodes because of Colab's nature. However, this can be scaled up using any local machine and creating a cluster that fits the needs using the local machine or cloud service providers like Databricks, Microsoft Azure HDInsight, or Google Cloud Dataproc, among others. My attempts can be viewed in the appendix section. [appendix A]

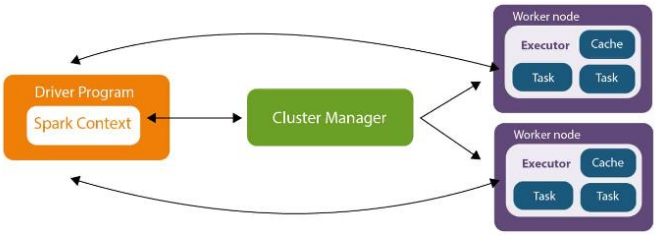

Figure I: PySpark architecture [2]

B. *Data preprocessing and cleaning*

i. Data collection: The data desired for the analysis was fetched using Kaggle API.

ii. Data cleaning: PySpark was manipulated to clean the data from any missing data and duplicates that may affect the analysis.

iii. Data transformation: PySpark was used to transform the data types into the desired form, joining the different files available for appropriate use, encoding categorical data into numerical form for applying the ML models efficiently.

Please note that the data used in this project has gone through stratified sampling in respect to the 'Acorn_grouped' attribute. The purpose of this is to just take a sample and fasten the execution of code for analysis. The fraction of each class used is 0.1.

C. *Feature selection and engineering*

i. Feature selection: After exploring the data and understanding how different attributes contribute to the problem, Correlation matrixes, as well as personal knowledge were used to get rid of those features that are of less value to the analysis.

ii. Feature engineering: As the data contains temporal data, a very efficient technique to analyze the data in a magnified view was to split those columns in addition to adding a 'day of week' column so that it is easier to understand the behavior in the yearly, monthly, daily, and hourly level.

D. *Visualizations*

i. Energy consumption per year: The data is available represents the following duration: from 2011-11-23 till 2014-02-28. That means that only two months are considered for both 2011 and 2014 whereas its full years for 2012 and 2013. Nevertheless, Figure II shows a good look for the overall statistics on the different years.

year	avg(sum_energy)	median(sum_energy)	stddev(sum_energy)
2011	0.52	0.3	0.63
2012	0.41	0.24	0.54
2013	0.42	0.24	0.56
2014	0.47	0.27	0.63

Figure II: Energy Consumption per year

ii. Energy consumption per Month: As illustrated in Figure III, people tend to consume more energy at the beginning and end of the month, but less in the middle of the month. This could be because of many reasons:

a. Budgeting: people may aim to save energy in the middle of the month to avoid high costs at the end of the billing period.

b. Seasonal variations: As it is cold in January, February, November, and

December, then people are probably staying at home instead of going out.

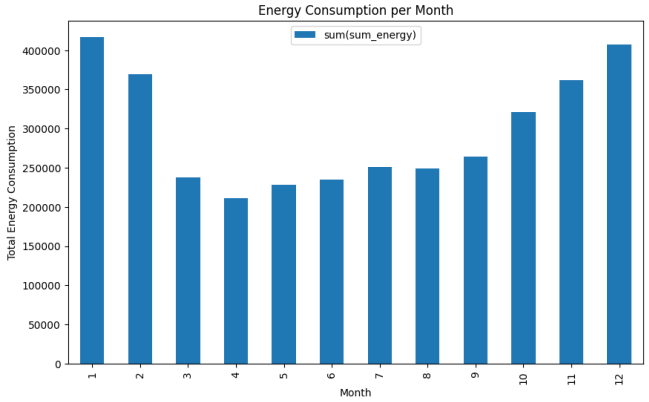

Figure III: Energy Consumption per Month

iii. Energy consumption per Hour: As illustrated by Figure IV, the peak in energy consumption appears to be around the evening hours (4:00 PM to 10:00 PM). In contrast, the least amount of energy is consumed in the morning hours (from 1:00 AM to 6:00 AM). This could be because of different factors in London households.

<u>High energy consumption hours:</u>

a. Daily routines: People return home from school and work around the late afternoon and early evening, and start cooking, using lighting, T.V, among other things that result in higher energy consumption.
b. Lighting: As daylight decreases, people rely on house lighting.

<u>Low energy consumption hours:</u>

c. Less activities: people are mostly asleep in morning hours, meaning less energy consumption.
d. Off-peak hours: in some regions, people are charged based on the time of the day. If it is night, it becomes cheaper. This is the concept of Time-of-Use (ToU).

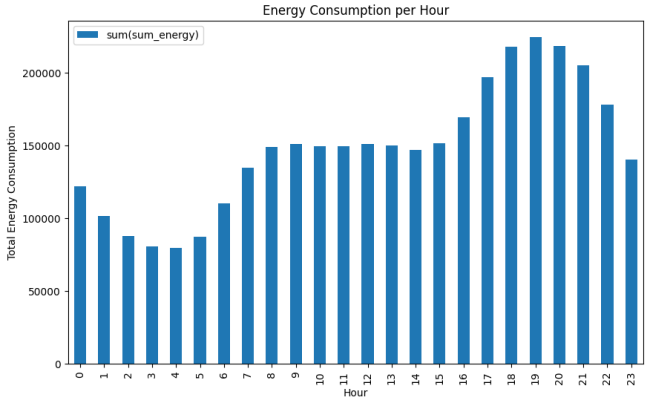

Figure IV: Energy Consumption per Hour

iv. Energy Consumption per Day: The distribution is uniform, that shows that no matter what day it is, that does not affect the amount of energy consumed.

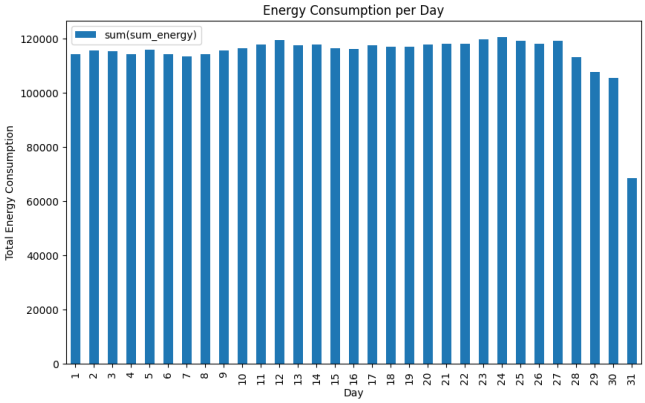

Figure V: Energy Consumption per Day

v. Energy Consumption Patterns by Day of the Week: not much can be derived from how day of the week influences energy consumption.

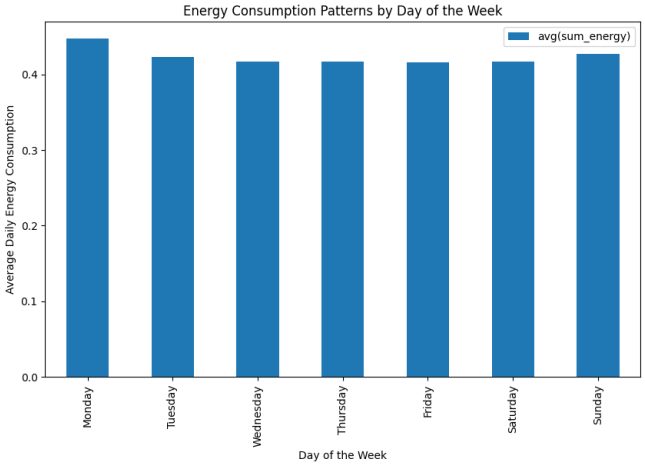

Figure VI: Energy Consumption Patterns by Day of the Week

Please note that more visualizations were made and constructed using Streamlit. [6] There exists more interactive visualizations to view information about each Acorn group and observe the highest and lowest points of energy consumption of each group over the years. [Appendix B]

E. Machine Learning Models

i. Regression model to predict the hourly energy consumption given temporal data only.
ii. Regression model to predict the hourly energy consumption given temporal, in addition to weather information.
iii. Classification models to predict the 'Acorn_grouped' based on energy consumption, and weather information.

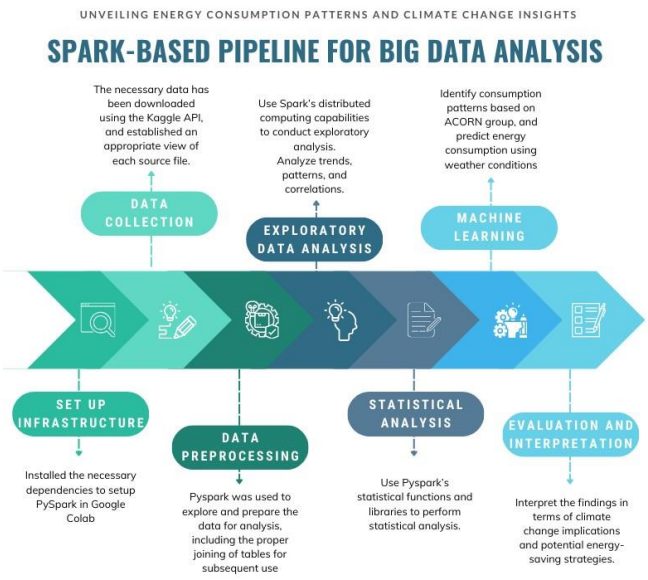

Figure VII: Project Pipeline

IV. EXPERIMENTS

A. Motivation

The motivation behind this experiment is to emphasis the urgent need for sustainable energy solutions. As we are now dealing with a huge and rapid change in climate, we should start examining and understanding household energy consumption patterns as it plays a big role in shaping carbon footprints and the impacts on the environment. By analyzing these patterns, we can gain valuable insights on how to transform our approaches to energy usage and reduce its effect on our earth.

This research study's findings carry substantial suggestions for different stakeholders. Policymakers can leverage those insights to make more effective policies and strategies that places both efficiency and sustainability as a priority. In addition, the research can also encourage and empower individuals by raising awareness as well as enhancing behavioral changes towards more sustainable activities.

B. Description

In this project, linear regression model was utilized, beginning with creating a linear regression model to predict energy consumption using only temporal data, this meant doing feature extraction to gain more knowledge about the year, month, day, day of the week, and hour of that record and later assemble, and normalize the data.

A pipeline was created to first assemble, normalize, then perform the linear regression model.

Two models have been used using only the temporal data, and another model has been utilized to be trained from all the data including the weather information available in our data.

Table III: Experiment Configuration Overview

Experiment Description	***Hyperparameter settings***
Data preprocessing	
Data source	Kaggle, Smart meters in London
Data cleaning and feature selection	Discussed deeply in the jupyter file [4]
Data splitting (training and test)	80% training and 20% testing.
Model Architecture 1	
Type of model	Linear regression model
Grid hyperparameters (model tuning)	Regularization parameters: 0.1, 0.01, 0.001 Elastic net mixing parameters: 0.0, 0.5, 1.0
Model selection and hyperparameter tuning	TrainValidationSplit with 0.8 train ratio.
Model Architecture 2 ***(only trained with 500000 due to the slow execution)***	
Type of model	Ensemble linear regression model
Hyperparameters	Regularization parameters: 0.1, 0.01, 0.001 Elastic net mixing parameters: 0.0, 0.5, 1.0
Approach	Model averaging of three linear regression models
Model Architecture 3 ***(same as Model Architecture 2 but with more weather data)***	
Model Architecture 4 ***(Uses RandomForestRegressor DecisionTreeRegressor LinearRegression)***	
Evaluation metrics	
Meretrics used for model evaluation	Root Mean Squared Error, Mean Absolute Error, R-squared, Mean Squared Logarithmic Error.
Reproducibility	
seed	42

C. Results

The results have been compared to the work of a competitor on Kaggle [5]. His experiment used daily data instead of hourly data and leveraged the power of pandas instead of PySpark. Therefore, this comparison could be beneficial in terms of comparing the Pandas vs PySpark. Nevertheless, this project's experiment took the bigger step by leveraging and taking the hourly data into use to make the model more precise at its predictions.

Note how the linear regression model is not bad compared to the other experiment, please keep in mind that not the entire data is being considered for training/testing, just a small portion of it.

The smaller the RMSE, MAE, and MSLE, the better the model is performing.

RMSE, Root Mean Squared Error, measures the average magnitude of errors between the predicted and the actual values.

MAE, Mean Absolute Error, measures the average absolute errors between the predicted and the actual values.

MSLE, Mean Squared Logarithmic Error, measures the average logarithmic errors between the predicted and the actual values.

On the other hand, R-squared is a measure of the proportion of the variance in the dependent variable that is predictable from the independent variables. It should be between 0 and 1, where 1 indicates a good performance.

Table IV: Results Comparison of Linear Regression Models

Experiment	***RMSE***	***MAE***	***R^2***	***MSLE***
Model Architecture 1	0.5498	0.3267	0.0335	0.3013
Model Architecture 2	0.3519	0.3331	-6.2201	0.1238
Model Architecture 3	0.2911	0.2711	-5.0774	0.0847
Other's Experiment	20.884		0.3834	

Since PySpark.ml does not contain ARD Regressor nor Ridge, this experiment only included Random Forest and Decision Tree in addition to the Linear regression.

Table V: Result Comparison of Radom Forest Regressor Model

Experiment	***RMSE***	***MAE***	***R^2***	***MSLE***
Model Architecture 4	0.6142	0.3451	0.0427	0.3773
Other's Experiment	14.251		0.7129	

Table VI: Result Comparison of Decision Tree Regressor Model

Experiment	***RMSE***	***MAE***	***R^2***	***MSLE***
Model Architecture 4	0.3449	0.3449	0.0377	0.3792
Other's Experiment	14.251		0.7061	

D. Interpretation and discussion

An extensive analysis and understanding of potential areas are necessary for improvement to handle and solve the problem of energy consumption at household energy level.

As a solution, this project investigates a machine learning algorithm that can predict energy consumption on an hourly basis. Moreover, since the main objective is to deal with big data, the model can adapt to real time response and helps in predicting the energy consumption of subsequent hours. (This can be implemented using LSTM and the attention mechanism).

Nevertheless, the linear regression model, in addition to the ensemble approach, was great in predicting energy consumption patterns. The success of the linear regression model could be because of how simple it is and its ability to be interpreted easily. This made it easier to understand the direct impact of time-based features on energy consumption.

As cons of this model, is its inability to accommodate non-linear relationships, and any challenging temporal dependencies within the data. That is why, as an opportunity this report will explore more complex models like RNNs or LSTMs in hope of capturing those complex patterns that linear models are ignorant of.

One potential solution that people may need to investigate further is energy storage technologies. This includes compressed air energy storage, pumped hydropower energy storage, flywheels, and batteries (like lead-acid, and lithium-ion).

The benefits of such energy storage solutions include the following:

i. Load shifting which saves the energy generated during the periods of low demands and giving you the opportunity of using it during high demand periods.

ii. Backup power which helps maintain continuous access to energy when needed or unforeseen events like technical issues.

iii. Renewable energy integration that utilizes renewable energy sources, like solar and/or wind power, Therefore, enhancing the sustainability of energy systems by leveraging the power of nature through technological adaptations.

iv. Grid support that manages and distributes the flow of energy to meet the demand and prevents grid congestion, especially during peak hours. This ensures availability and sense of control.

Yet, limitations must also be considered. Those might include the following:

i. Those energy storage systems require ongoing operational costs as well as physical space.

ii. They also have specific lifespan and degradation characteristics.

The following must also be considered as opportunities and as this project evolves:

i. Policies and regulations related to energy storage in the country of choice.

ii. The scalability of the chosen energy solution.

iii. Safety characteristics of the chosen energy systems and their corresponding environmental impacts.

iv. Simplicity and agility of the selected interface to help with the monitoring and control for homeowners.

v. Making a good community to raise awareness about how beneficial energy storage is.

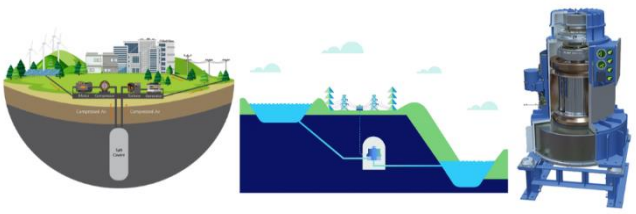

Figure VIII: Energy Storage Systems

V. Conclusion

In conclusion, this project successfully navigated the different paths and possibilities when it comes to energy consumption prediction and how humans can start to benefit from underlying data regarding the hourly energy consumption to make better decisions to benefit both parties (consumers and providers). This report is nothing but a step towards a bigger solution and exploring more insights for a better sustainable future that will reduce the impact energy consumption affects climate change.

This project will also go further into exploring the different Acorn groups and how each group's behavior is contributing to energy consumption. This could also include involving a machine learning model that is able to price individuals according to how much energy they consume per hour in a better and flexible manner, taking into consideration the hour of the day, as well as any additional information that can be collected.

VI. References

[1] J. M. D, "Smart meters in London," *www.kaggle.com*, May 23, 2022. https://www.kaggle.com/datasets/jeanmidev/smart-meters-in-london (accessed Oct. 16, 2023).

[2] R. Kumar, "Getting Started With PySpark," *www.c-sharpcorner.com*, Feb. 16, 2023. https://www.c-sharpcorner.com/article/getting-started-with-pyspark/ (accessed Nov. 07, 2023).

[3] CACI, *How Acorn Works*. Accessed: Nov. 07, 2023. [Online]. Available: https://www.google.com/url?sa=i&url=https%3A%2F%2Facorn.caci.co.uk%2Fhow-acorn-works%2F&psig=AOvVaw1eIEtnJvC6awl7ZsULjtNn&ust=1699643857058000&source=images&cd=vfe&ved=0CBIQjRxqFwoTCIDz0qHSt4IDFQAAAAAdAAAAABAU

[4] Maryam Ali, "Analyzing Energy Consumption," Jupyter Notebook, 2023. Available: ./Analyzing Energy Consumption.ipynb

[5] MILDCOFFEE, "Energy Consumption Prediction," *kaggle.com*, Apr. 22, 2023. https://www.kaggle.com/code/mildcoffee/energy-consumption-prediction (accessed Nov. 11, 2023).

[6] Maryam Ali, "Visualizations," Python file, 2023, Available: ./more_visualizations/visualizations.py

[7] robinson-wn, "Running a spark application in Docker and on Kubernetes," *GitHub*, Nov. 08, 2023. https://github.com/robinson-wn/k8spark (accessed Oct. 18 , 2023).

VII. Appendix

A. Appendix A

My first attempt was to create a spark cluster through Kubernetes. I have set up all configurations needed for the experiment like installing Ubuntu (WSL2), installing helm, enabling Kubernetes through docker, and set up the files and dependencies needed to run the application using PyCharm as my IDE. [7]

However, due to my limited experience, this took me 11 days to achieve, understand and learn, yet I still encountered issues with the resources of the cluster, which put me under pressure, and I had to change my solution into a less complicated application.

The next thing I have thought about leveraging is creating a Hadoop cluster that had a spark image, this worked perfectly until I had to place my files and create a directory inside the cluster so that I can use it.

Therefor and under the time pressure I was dealing with, the only solution was to utilize google colab alone and continue my work on there.

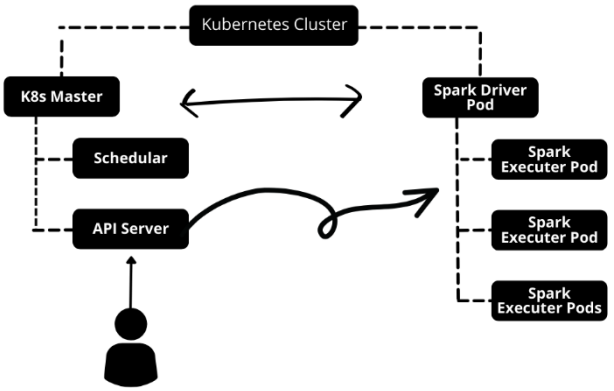

Figure IX: Initial Kubernetes cluster plan

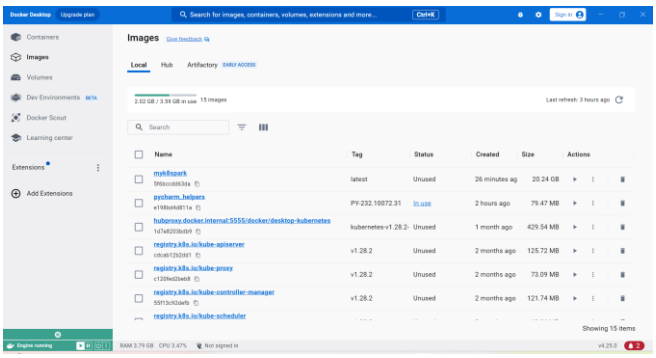

Figure X: Spark Image using Kubernetes

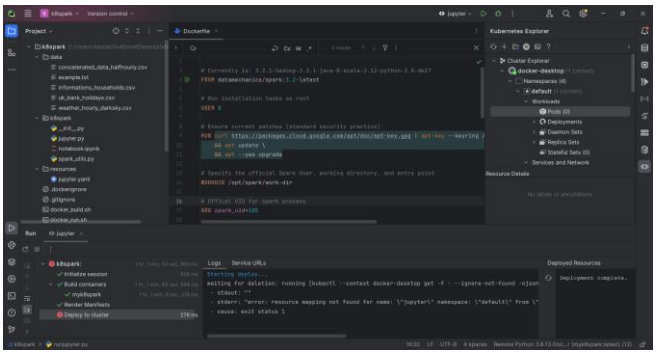

Figure XI: Spark via Kubernetes Application

Note that I have already enabled Kubernetes cluster and all the necessary configurations, yet issues remained unsolved.

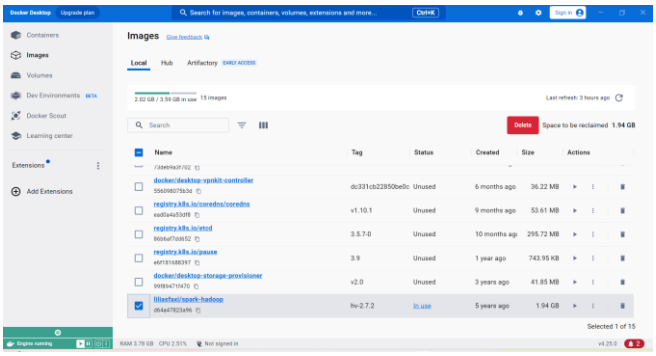

Figure XII: Hadoop Image

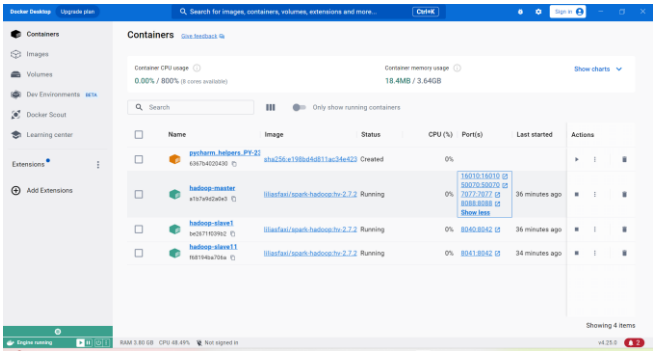

Figure XIII: Hadoop Master and Slave Nodes

B. Appendix B

As being done, visualizations are very important to understand why and where the problem is, and present that in a nice graphical and user-friendly way. Using Streamlit, I was able to demonstrate a nice-looking website-like page showing interactive and nice visualizations using a sample of the data.[6] To view the data please follow the following:

i. Open the Anaconda PowerShell.

ii. Enter the path where the file exists.

iii. Run Streamlit file.

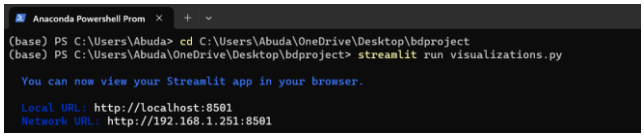

Figure XIV: Opening Streamlit