UNIVERSITI MALAYA

Centre for Continuing Education

MASTER IN DATA SCIENCE

GROUP ASSIGNMENT

(GROUP 10)

COURSE CODE : WQD7009

PROJECT TITLE : RENEWABLE ENERGY CONSUMPTION IN THE UNITED STATES

GROUP MEMBER : GROUP 10

1) NUR HIDAYAH BINTI AHMAD SHAFII (22120931)
2) CHOON YUE HUA (17152027)
3) THEN DAO QING (23057608)
4) LOW MENG FEI (23063305)
5) SYAIDATUL SALMAH NURBALQIS BIN SAIFUL (17140336)

LECTURER : DR. RIYAZ AHAMED ARIYALURAN HABEEB MOHAMED

Table of Contents

1 Introduction

The unpredictable changes in the Earth's climate have resulted in a rise in global challenges. In response to the urgency of the situation, The United Nation introduced Sustainable Development Goal (SDG) 13 in response to the pressing need to address climate change and reduce its effect. Climate change is mainly driven by the increase in greenhouse gases, which is largely due to human activities (Vishvakarma, 2022). Fossil fuel combustion in energy consumption is the primary source of greenhouse gas emissions in the United States (Delmas & Montes-Sancho, 2011). It remains the primary contributor to the energy sector despite the fact that fossil fuels are not sustainable and have severe environmental and health consequences (Olabi & Abdelkareem, 2022). The transition to renewable energy is crucial in the effort to reduce greenhouse gas emissions and slow down climate change. Therefore, it is important to understand the trends of energy consumption patterns to minimize climate change.

One of the effective methods to address this global issue is the development of forecasting models for renewable energy generation. These models are important for the decision-making process to increase sustainable energy implementation. Machine learning has been proven with its ability to forecast energy consumption. Nevertheless, there are still issues with choosing the best models and data quality for an accurate energy consumption forecasts. In this paper, we analyze the historical trends in renewable energy consumption in the United States to predict future trends using a machine learning model. Through this analysis, we developed a framework and visualization to identify the trends and provide insight to stakeholders.

The scope of the study focused on the monthly renewable energy consumption data in the United States from 1973 to 2024. The dataset consists of 17 columns and 3,065 rows. It has two integer columns (Year and Month), one category column (sector), and the remaining columns contain float data types. There is a consumption pattern across five economic sectors namely commercial, electric power, industrial, residential and transportation. The types of renewable energy sources in the dataset are Hydroelectric Power (including Conventional Hydroelectric Power), Geothermal Energy, Solar Energy, Wind Energy, Wood Energy, Waste Energy, Fuel Ethanol (excluding Denaturant), Biomass Losses and Co-products, Biomass Energy, Renewable Diesel Fuel, Other Biofuels, and Biodiesel.

2 Data Architecture

A data architecture is a systematic data-driven process that responsible to ensure organisational rules and policies are consistently applied to all data. Although the structure may be varied, it usually consists main component such as data sources, data ingestion, data processing, data storage and data sinks (Foidl et al., 2024). The data architecture used in this project is as illustrated in Figure 2.1. The architecture combines advanced technologies such as Google BigQuery, Google DataPrep, Google Colab, and Power BI, with Zapier for automation and monitoring.

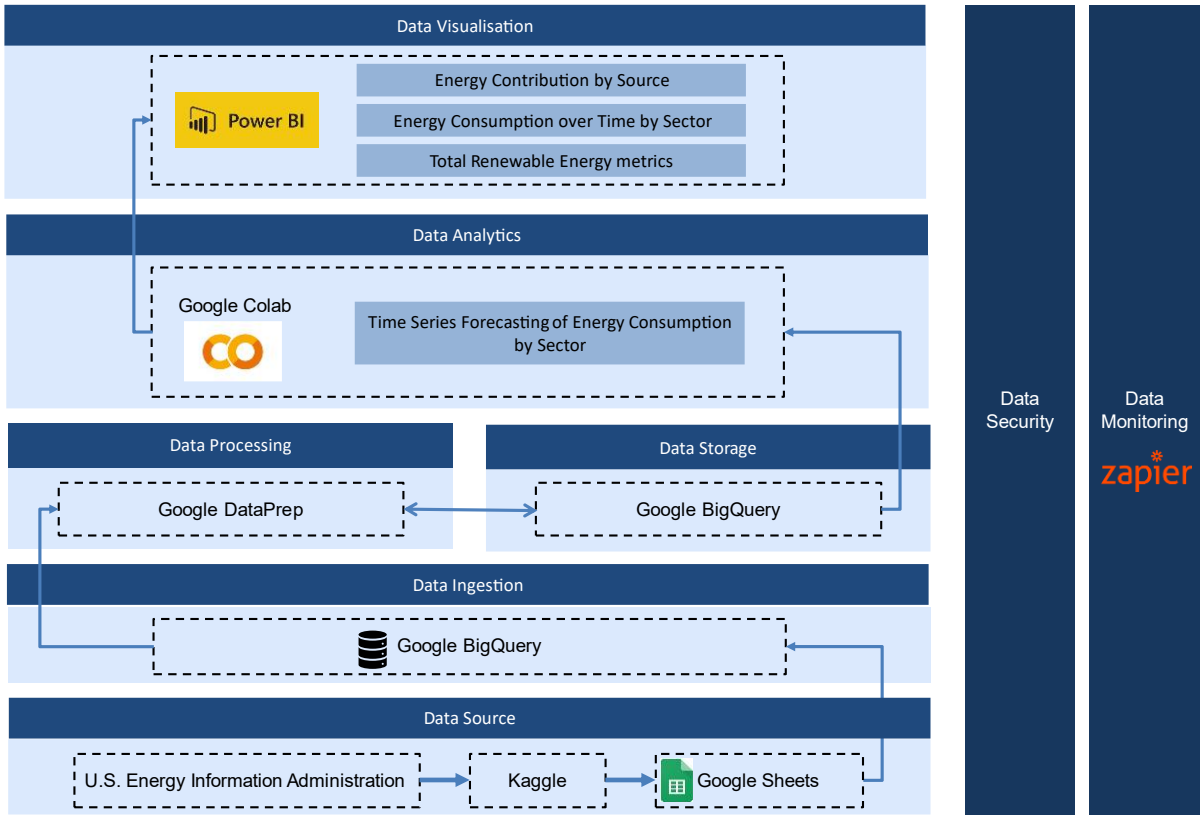


Figure 2.1 Data Architecture Diagram

The first layer of the architecture is the data source. The data is initially obtained from a reliable source, the United States Energy Information Administration. Then, it is compiled and made accessible through the Kaggle platform. From Kaggle, the raw data is imported into Google Sheets before ingesting to BigQuery. Google Sheets is used because it has a user-friendly interface that makes it easy for the non-technical team to update data in the future.

The second layer is data ingestion to Google BigQuery. Google BigQuery is chosen not only because of its speed and scability but also because of its ability to seamlessly integrate

with other platforms in the framework, such as Google Sheets, Google DataPrep, and Power BI. This seamless integration will ensure the data flows efficiently throughout the pipeline. Since the data in Google Sheets does not update automatically in BigQuery unless it is refreshed, we scheduled a data refresh using the free “Refresh options” feature in Google Sheets. The data is scheduled for monthly refresh since the updates are less frequent. The schedule refresh will be paused automatically if the user updates the existing data. The data partitioning can be implemented in BigQuery to further optimize query performance and reduce cloud cost. It helps to distribute data store across different storage types. For example, partition pruning to remove unnecessary time stamp data from consideration will helps the analyst to focus on relevant information.

The third layer is data processing using Google DataPrep. In this layer, a data flow is created to manage and organize the preprocessing tasks. Google DataPrep automatically validates the dataset and highlights if there are any missing or invalid values using the color-coded indicator in the column headers. This visual feedback allows users to identify and address problems within the dataset quickly. Once the issues are identified, the data is carefully reviewed and reformatted as needed. For example, the Year column is extracted from the date field to simplify time-based analysis. Additionally, irrelevant columns that do not contribute to predicting energy consumption by sector are removed. This preprocessing process will ensure that the data is cleaned before the modelling process.

The next layer is data storage using Google BigQuery. The cleaned data will be stored in Google BigQuery. The cleaned data is stored in a separate dataset to maintain a clear separation between the raw and processed data. This separation will ensure that the user works with the correct version of the date and reduce the risk of overwriting. On the other hand, users can access the dataset using SQL queries. This feature helps both technical and non-technical teams interact with the data.

Last but not least, the cleaned dataset will be imported into Google Collab for machine learning analytics. In this layer, we will forecast time series using models such as Simple Recurrent Neural Networks (Simple RNN) or Long Short-Term Memory (LSTM) networks. The target variable for the forecasting is the energy consumption of individual sectors such as Hydroelectric Power, Geothermal Energy, Solar Energy, Wind Energy, Wood Energy, and others. The input feature is a time-based variable, which is Year. Google Colab is chosen for its free access and ability to collaborate with teams. The performance of the models is evaluated

using metrics of RMSE, MAE, and R^2 to ensure accurate and reliable forecasts for decision-making.

The last layer of the framework is visualation using Microsoft PowerBI. The cleaned data from the Google Big Query is imported to Microsoft PowerBI. Since the tools are connected, any changes or updates to the cleaned data in BigQuery are automatically reflected in the Power BI dashboard. A color template based on the brand image is carefully selected to give the dashboard a professional look. A maximum of three colours is chosen to avoid the user feel overwhelmed. The interactive dashboard includes key visualizations designed for stakeholders such as energy contribution by source, energy consumption over time and total renewable energy metrics. Additionally, the use of slicers in Power BI will help to analyze specific subsets of data.

Throughout the process, DevOps will be actively involved in the data monitoring layer. Zapier automation tool will be used to notify the relevant team whenever updates are made to the Google Sheets. Additionally, it will send an alert email if there is any disruption in the data pipeline. This ensures that all changes are tracked in real time and minimizes the risk of errors. DevOps also monitors the overall system performance of Google BigQuery and Google DataPrep by using pay-as-you-go service of Google Cloud Monitoring. Google Cloud Monitoring will track the performance, uptime, and overall health of the cloud services. It will send alerts if there is high latency issues or failure. These alerts enable quick troubleshooting and reduce downtime.

Lastly, data security layer is the most important component of the process. In this layer, DevOps and IT security team is responsible for data integrity and prevent security incidents. DevOps will be responsible to implement Identity and Access Management (IAM) policies in the cloud environment. IAM Groups are used to manage permissions for multiple users. It is easier to manage user permission using this method as DevOps does not need to configure roles individually for many users. The architecture follows the least privilege permission model whereby users only grant necessary permission to perform their tasks. This approach will minimize the risk of unauthorized access and ensure the security of the data architecture.

Each layer of the data lifecycle process plays a vital role in the overall framework, with strong interconnections ensuring a seamless flow of data from source to visualization. By focusing on upgrading and optimizing each layer, the framework becomes more robust,

scalable, and efficient, capable of handling large and complex datasets. Together, these layers provide a comprehensive solution for analysing renewable energy consumption, enabling stakeholders to derive meaningful insights and make data-driven decisions. The importance of these layers lies not only in their individual functionalities but also in how they work together to form an integrated, reliable, and secure system.

3 Tools Selection Justification

i. Data Ingestion: Google Sheets and Google BigQuery

For the data ingestion layer, Google Sheets and Google BigQuery are selected as the primary tools. Google Sheets serves as an initial platform for handling raw data due to its user-friendly interface, which allows non-technical team members to review and update data effortlessly. It also facilitates quick and basic cleaning before transferring the dataset to more advanced tools. On the other hand, Google BigQuery, a serverless and highly scalable data warehouse, is chosen for its ability to seamlessly ingest and process large datasets. Its integration with Google Sheets ensures a smooth data flow, and its data partitioning capabilities optimize query performance while reducing storage costs. By scheduling data updates from Google Sheets, the system ensures consistent and timely ingestion into BigQuery.

ii. Data Processing: Google Dataprep

The data processing layer is essential for transforming raw data into a clean and structured format. Google Dataprep is a key tool in this process, offering automated data profiling to identify and resolve missing values, outliers, and inconsistencies. Its user-friendly interface and machine learning-powered suggestions simplify complex tasks like data validation, extraction, and filtering. This ensures that the dataset is accurate, consistent, and ready for analysis. Dataprep's integration with tools like BigQuery allows seamless data movement within the pipeline. High-quality data processing improves the reliability of analytics, as clean data forms the basis for accurate predictions and insights. Optimizations, such as reusable templates and automated anomaly detection, further enhance efficiency and maintain consistency across projects. This makes Dataprep a crucial component of any robust data analytics framework.

iii. Data Storage: Google BigQuery

The cleaned and processed data is stored in Google BigQuery, ensuring data integrity and accessibility for analytics. BigQuery's robust querying capabilities, which use standard SQL, make it easy for both technical and non-technical users to interact with the dataset. By separating raw and processed data into different datasets, the system ensures clarity and prevents accidental overwrites. Features such as query optimization and partitioning further enhance its efficiency, reducing computational costs and latency. BigQuery's scalability and compatibility with various tools make it an ideal choice for storing the renewable energy dataset.

iv. Data Analytics: Google Colab

For the analytics layer, Google Colab is employed to build and test machine learning models. Google Colab is a free, collaborative environment that supports Python-based libraries such as TensorFlow and Keras, making it an excellent platform for implementing forecasting models like Long Short-Term Memory (LSTM) and Simple Recurrent Neural Networks (Simple RNN). This tool is particularly beneficial as it provides a shared workspace where team members can collaborate on model development and evaluation in real time. Additionally, its integration with BigQuery enables seamless data retrieval for model training and testing. The model's performance is assessed using metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R^2 to ensure accurate predictions.

v. Data Visualization: Microsoft Power BI

The visualization layer is handled using Microsoft Power BI, which is known for its powerful and interactive dashboard capabilities. Power BI imports the cleaned dataset from BigQuery, ensuring that any updates in the dataset are automatically reflected in the dashboards. Its user-friendly interface allows the creation of clear and professional visualizations that highlight key metrics, such as energy consumption trends over time and contributions by different renewable energy sources. Advanced features like slicers enable stakeholders to filter and explore specific data subsets interactively. To maintain clarity and professionalism, a limited color scheme is applied, adhering to best visualization practices.

vi. Monitoring and Automation: Zapier

For monitoring and automation, Zapier is selected to streamline notifications and alerts within the data pipeline. Zapier automates updates, notifying relevant team members whenever changes occur in the Google Sheets source data or disruptions arise in the pipeline. For instance, it sends alerts if there are failures in the data refresh schedule or high latency in cloud operations. This tool reduces manual monitoring efforts, allowing the team to focus on more critical tasks while ensuring quick responses to potential issues.

vii. Data Security: Google Cloud Identity and Access Management (IAM)

Data security is managed using Google Cloud Identity and Access Management (IAM), which ensures that only authorized individuals have access to the dataset. IAM follows the principle of least privilege, granting users only the permissions necessary for their tasks. Role-based

access controls simplify the management of user permissions, particularly for larger teams. This approach minimizes the risk of unauthorized access and ensures compliance with data security policies, maintaining the integrity and confidentiality of the renewable energy data.

4 Framework Implementation

4.1 Data Preprocessing: Google Dataprep

The dataset consists of 17 parameters and 3065 rows. The table below shows the parameter descriptions:

Table 4.1: Data Description

Parameter	**Description**
Year	The year of the data entry
Month	The month of the data entry (e.g., 1 for January).
Sector	The energy consumption sector, which included Commercial, Electric Power, Industrial, Residential, and Transportation.
Hydroelectric Power	Hydroelectric power, in trillion BTUs
Geothermal Energy	Geothermal energy consumption, in trillion BTUs
Solar Energy	Solar energy consumption, in trillion BTUs
Wind Energy	Wind energy consumption, in trillion BTUs
Wood Energy	Wood energy consumption, in trillion BTUs
Waste Energy	Waste energy consumption, in trillion BTUs
Fuel Ethanol, Excluding Denaturant	Fuel ethanol (excluding denaturant) consumption, in trillion BTUs
Biomass Losses and Co-products	Biomass losses and co-products, in trillion BTUs
Biomass Energy	Total biomass energy consumption (sum of wood, waste, ethanol, and losses/co-products), in trillion BTUs
Renewable Diesel Fuel	Renewable diesel fuel consumption, in trillion BTUs
Other Biofuels	Other biofuels consumption, in trillion BTUs
Conventional Hydroelectric Power	Conventional hydroelectric power consumption, in trillion BTUs
Biodiesel	Biodiesel consumption, in trillion BTUs
Total Renewable Energy	Total renewable energy consumption, in trillion BTUs

i. Check and update data type

The data type is auto assigned to each parameter by Dataprep, which is as shown in the figure below. The data types are incorrectly assigned for several columns. For example, year is assigned to ‘datetime’ and some of the renewable energy consumption columns are assigned to ‘integer’. The data quality at the right end shows that there are mismatches value because the data type of these columns is assigned to integers (whole number) while the rows consist of decimal value. Therefore, these columns need to be updated to decimal (numeric) data type.

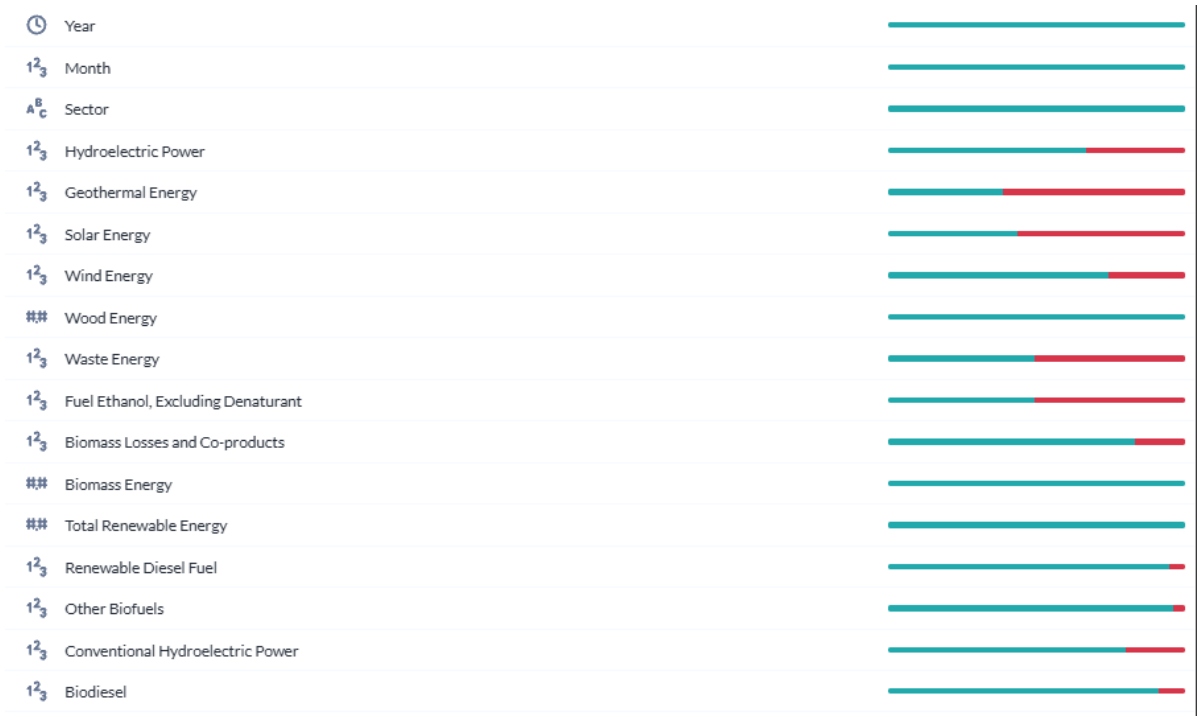


Figure 4.1: Auto-assigned data type, whereby some of them are inaccurate.

The data type of each column is updated as below:

- Integer: Year, Month
- Strings: Sector
- Decimal (Numeric): Hydroelectric Power, Geothermal Energy, Solar Energy, Wind Energy, Wood Energy, Waste Energy, Fuel Ethanol, Excluding Denaturant, Biomass Losses And Co-Product, Biomass Energy, Renewable Diesel Fuel, Other Biofuels, Conventional Hydroelectric Power, Biodiesel, Total Renewable Energy

After updating the column to the correct data type, there are no more mismatches values.

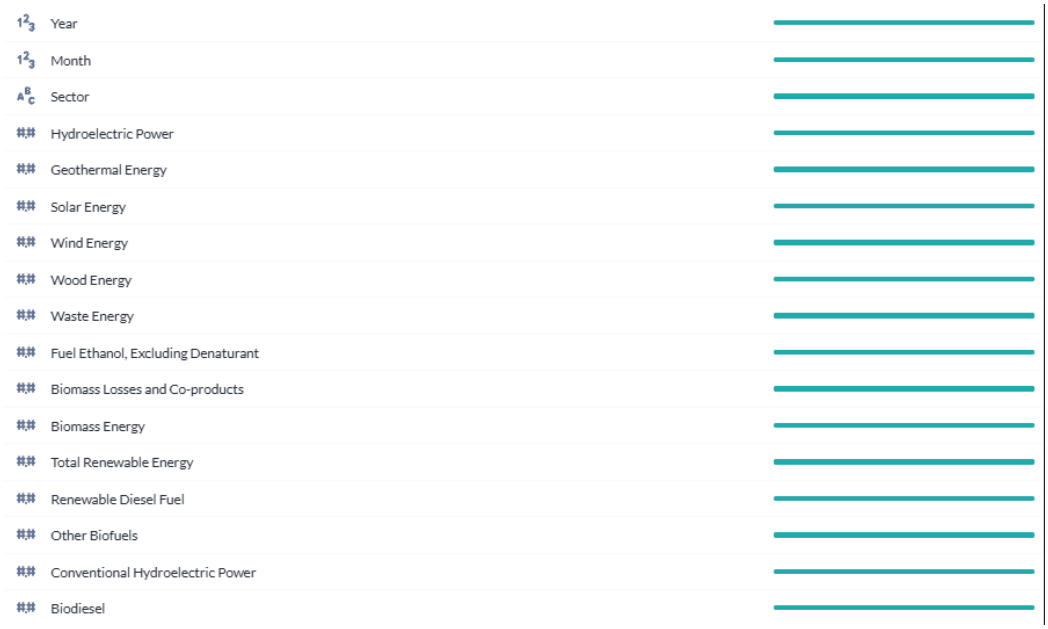


Figure 4.2: Updated data type.

2 **Lock** Year **type** to Integer

3 **Lock** Hydroelectric Power **type** to Decimal

4 **Lock** Geothermal Energy **type** to Decimal

5 **Lock** Solar Energy **type** to Decimal

6 **Lock** Wind Energy **type** to Decimal

7 **Lock** Waste Energy **type** to Decimal

8 **Lock** Biodiesel **type** to Decimal

9 **Lock** Conventional Hydroelectric Power **type** to Decimal

10 **Lock** Other Biofuels **type** to Decimal

11 **Lock** Renewable Diesel Fuel **type** to Decimal

12 **Lock** Biomass Losses and Co-products **type** to Decimal

13 **Lock** Fuel Ethanol, Excluding Denaturant **type** to Integer

14 **Lock** Fuel Ethanol, Excluding Denaturant **type** to Decimal

Figure 4.3: Breakdown of steps in recipe for updating data type.

ii. Delete redundant columns and create new category column

3 new columns will be created:

- total_biomassEnergy: Sum of wood energy, waste energy, fuel ethanol, excluding denaturant, biomass losses and co-product
- total_hydroelectricPower: Sum of hydroelectric power, conventional hydroelectric power
- total_biofuels: Sum of renewable diesel fuel, other biofuels, biodiesel

A new column, 'total_biomassEnergy', will be created to allow us to verify whether the original 'biomassEnergy' column accurately reflects the sum of these four columns.

As shown in the figure below, the number of unique zero values reduce to 96 in the newly created 'total_biomassEnergy' column, confirming that the data now accurately represents the sum of the respective energy sources. Therefore, the original 'biomassEnergy' column will be removed. The sum is then rounded to three decimal places and stored in a new column called 'round_total_biomassEnergy'. The original columns are deleted to prevent redundancy, and the rounded column is renamed back to 'total_biomassEnergy' for consistency.

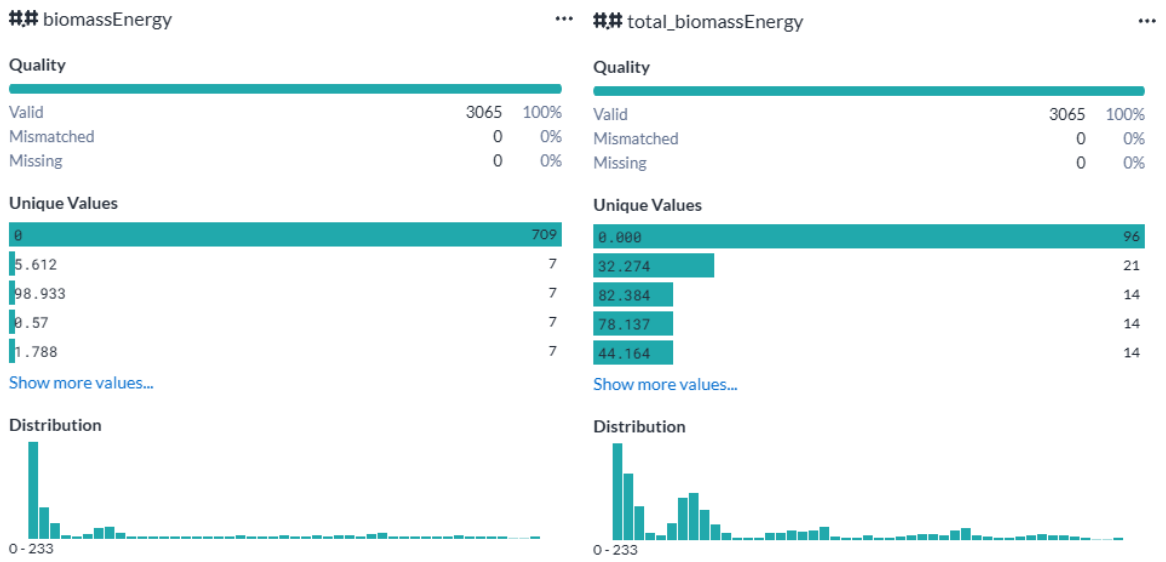


Figure 4.4: Original 'biomassEnergy' column does not correctly sum up the values of the other 4 columns. Unique value of zero is reduced from 709 to 96 in the newly created total_biomassEnergy.

```
16 Create total_biomassEnergy from {Wood Energy} + {Waste
   Energy} + {Fuel Ethanol, Excluding Denaturant} + {Biomass
   Losses and Co-products}

17 Create round_total_biomassEnergy from
   ROUND(total_biomassEnergy, 3)

18 Delete Biomass Energy

19 Delete total_biomassEnergy

20 Delete Biomass Losses and Co-products

21 Delete Fuel Ethanol, Excluding Denaturant

22 Delete Waste Energy

23 Delete Wood Energy

24 Rename round_total_biomassEnergy to
   'total_biomassEnergy'
```

Figure 4.5: Breakdown of steps in recipe for creating ‘total_biomassEnergy’.

The steps begin by creating a new column, ‘total_hydroelectricPower’, which is the sum of ‘Conventional Hydroelectric Power’ and ‘Hydroelectric Power’. This sum is then rounded to three decimal places and stored in the ‘round_total_hydroelectricPower’ column. The original columns, ‘Conventional Hydroelectric Power’, ‘Hydroelectric Power’, and the intermediate ‘total_hydroelectricPower’, are deleted to reduce redundancy. Finally, the rounded column is renamed back to ‘total_hydroelectricPower’ for clarity and consistency.

```
25 Create column1 from {Conventional Hydroelectric Power} +
   {Hydroelectric Power}

26 Rename column1 to 'total_hydroelectricPower'

27 Create round_total_hydroelectricPower from
   ROUND(total_hydroelectricPower, 3)

28 Delete Conventional Hydroelectric Power

29 Delete Hydroelectric Power

30 Delete total_hydroelectricPower

31 Rename round_total_hydroelectricPower to
   'total_hydroelectricPower'
```

Figure 4.6: Breakdown of steps in recipe for creating ‘total_hydroelectricEnergy’.

The steps start by creating a new column called ‘total_biofuels’, which is the sum of ‘Other Biofuels’, ‘Biodiesel’, and ‘Renewable Diesel Fuel’. This sum is then rounded to three decimal places and stored in the ‘round_total_biofuels’ column. The original columns, ‘Biodiesel’, ‘Other Biofuels’, ‘Renewable Diesel Fuel’, and the intermediate ‘total_biofuels’, are deleted to streamline the data. Finally, the rounded column is renamed to ‘total_biofuels’ for consistency and clarity.

```
32 Create total_biofuels from {Other Biofuels} + Biodiesel +
   {Renewable Diesel Fuel}
33 Create round_total_biofuels from ROUND(total_biofuels, 3)
34 Delete Biodiesel
35 Delete Other Biofuels
36 Delete Renewable Diesel Fuel
37 Delete total_biofuels
38 Rename round_total_biofuels to 'total_biofuels'
```

Figure 4.7: Breakdown of steps in recipe for creating 'total_biofuels'.

iii. Standardized column name and data format

Next, the column names are standardized, and the data are round up to 3 decimal places for data consistency.

```
39 Rename Year to 'year'
40 Rename Month to 'month'
41 Rename Sector to 'sector'
42 Rename Geothermal Energy to 'geothermalEnergy'
43 Rename Solar Energy to 'solarEnergy'
44 Rename Wind Energy to 'windEnergy'
45 Rename Total Renewable Energy to 'total_renewableEnergy'
```

Figure 4.8: Breakdown of steps in recipe for renaming columns.

```
46 Create round_geothermalEnergy from
   ROUND(geothermalEnergy, 3)
47 Create round_solarEnergy from ROUND(solarEnergy, 3)
48 Create round_windEnergy from ROUND(windEnergy, 3)
49 Delete geothermalEnergy
50 Delete solarEnergy
51 Delete windEnergy
52 Rename round_geothermalEnergy to 'geothermalEnergy'
53 Rename round_solarEnergy to 'solarEnergy'
54 Rename round_windEnergy to 'windEnergy'
```

Figure 4.9: Breakdown of steps in recipe for updating decimal places.

The table below shows the parameter descriptions after deleting redundant columns and column names are now standardized.

Table 4.2 New data table after data cleaning.

Parameter	Description
year	The year of the data entry
month	The month of the data entry (e.g., 1 for January).
sector	The energy consumption sector, which included Commercial, Electric Power, Industrial, Residential, and Transportation.
geothermalEnergy	Geothermal energy consumption, in trillion BTUs
solarEnergy	Solar energy consumption, in trillion BTUs
windEnergy	Wind energy consumption, in trillion BTUs
total_biomassEnergy	Total biomass energy consumption (sum of wood, waste, ethanol, and losses/co-products), in trillion BTUs
total_biofuels	Total biofuel energy consumption (sum of renewable diesel fuel, other biofuels and biodiesel), in trillion BTUs consumption, in trillion BTUs
total_hydroelectricPower	Hydroelectric power (sum of conventional and new method), in trillion BTUs
total_renewableEnergy	Total renewable energy consumption, in trillion BTUs

iv. Verifying data quality

We begin by checking for any missing values in each column by examining the column details in the preview section. Additionally, we review the right-hand panel for any "Missing" field indicators.

After reviewing, we find that there are no missing or mismatched values in the dataset. The column details pane also provides valuable insights into the data distribution for each column, allowing us to further assess the consistency of the data.

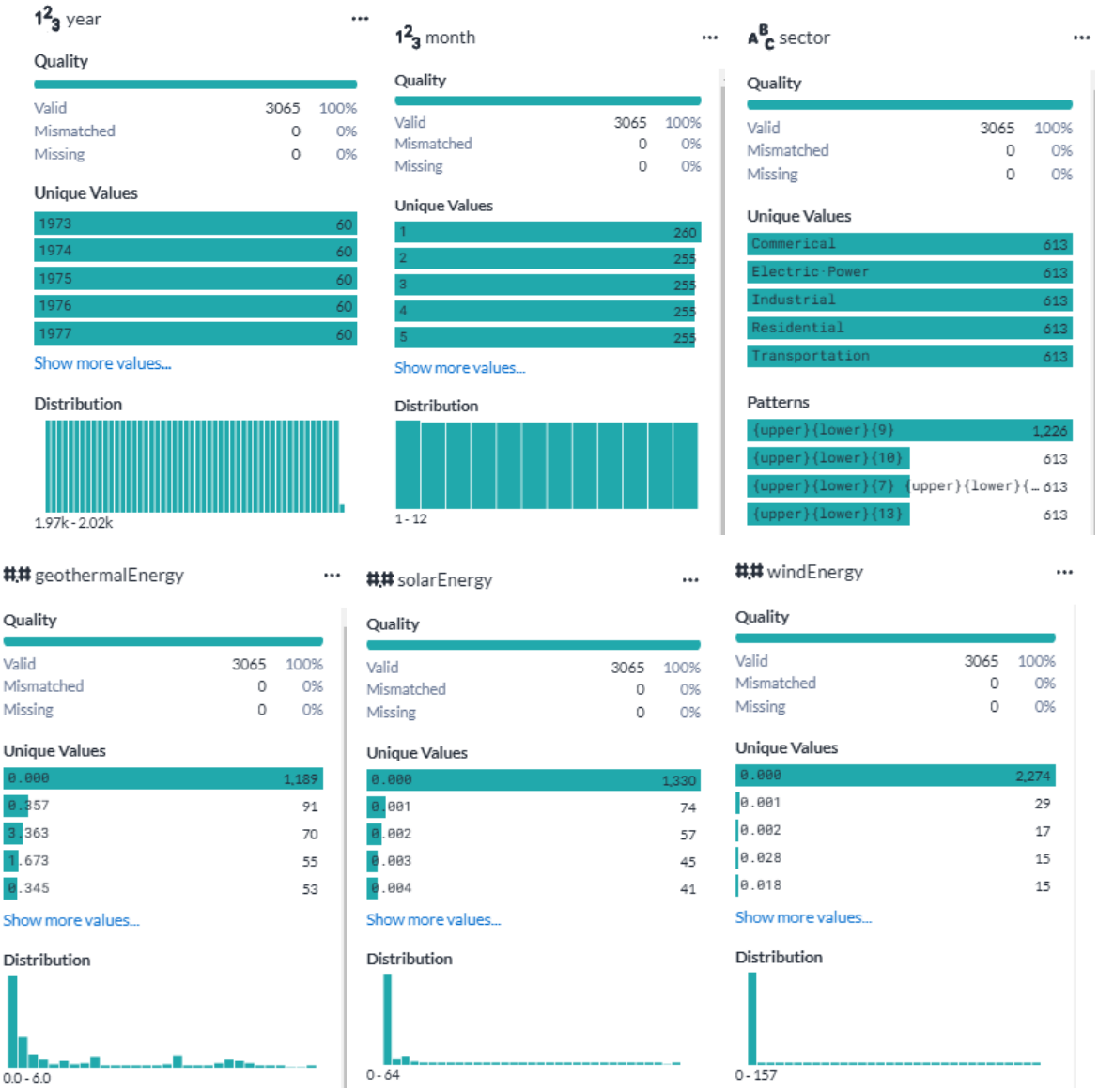


Figure 4.10 Data details of each column

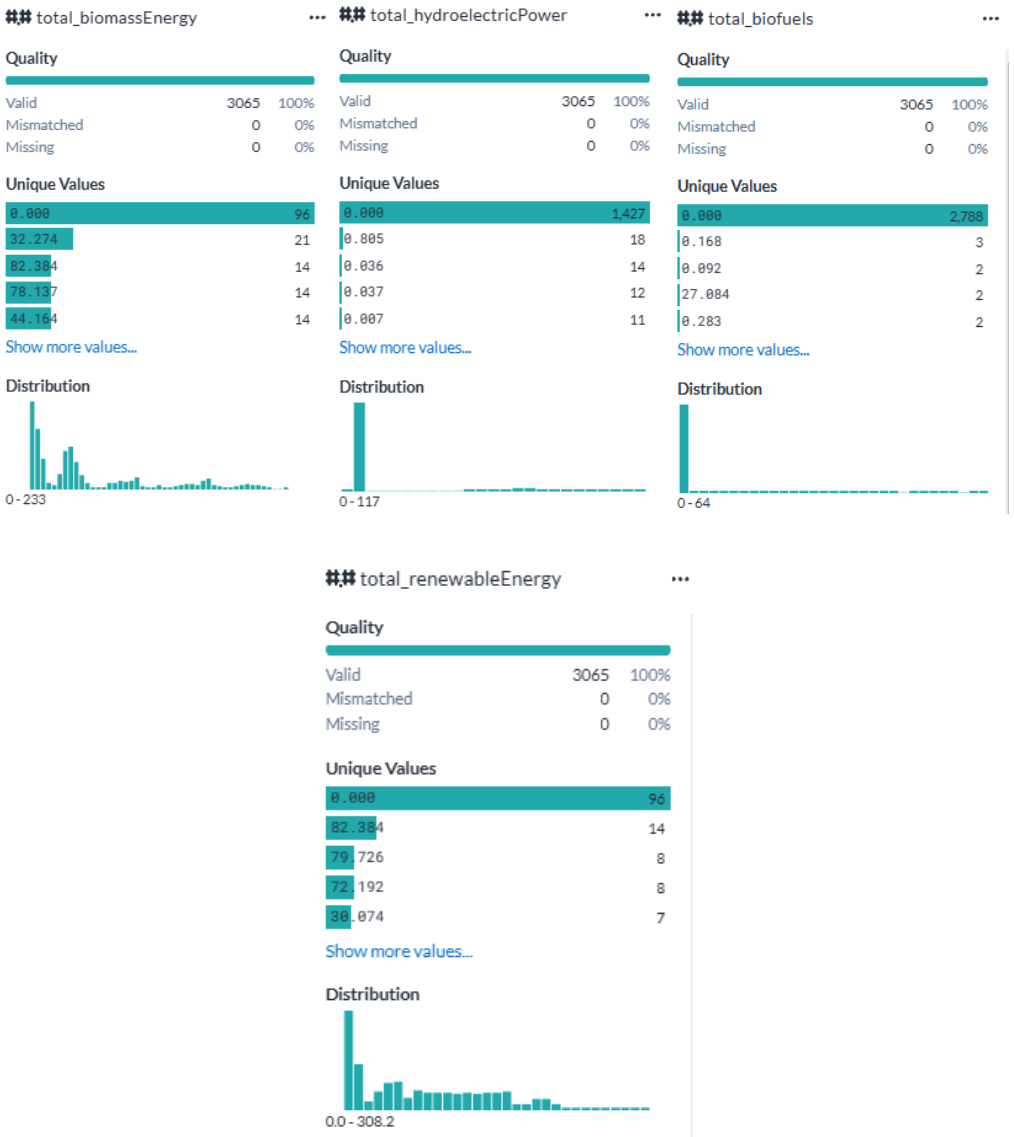


Figure 4.11: Data details of each column (continued)

v. Check if there is miscalculated value in total_renewableEnergy

In the 'total_renewableEnergy' column, we observe that there are 328 rows with a value of 0. To verify whether this is accurate, we will create a new column, 'new_total_renewableEnergy', which will be the sum of various energy sources: 'total_biofuels', 'geothermalEnergy', 'solarEnergy', 'total_biomassEnergy', 'total_hydroelectricPower', and 'windEnergy'. After performing this summation, we find that the number of zero values in the newly created 'new_total_renewableEnergy' column decreases from 328 to 96.

This suggests that the original ‘total_renewableEnergy’ column does not correctly reflect the sum of the energy sources. As a result, we will remove the inaccurate ‘total_renewableEnergy’ column.

Next, we round the values in ‘new_total_renewableEnergy’ to three decimal places and store these rounded values in a new column, ‘round_new_total_renewableEnergy’. To keep the dataset clean, we will delete the original ‘new_total_renewableEnergy’ and ‘total_renewableEnergy’ columns. Finally, we will rename the ‘round_new_total_renewableEnergy’ column to ‘new_total_renewableEnergy’ for consistency.

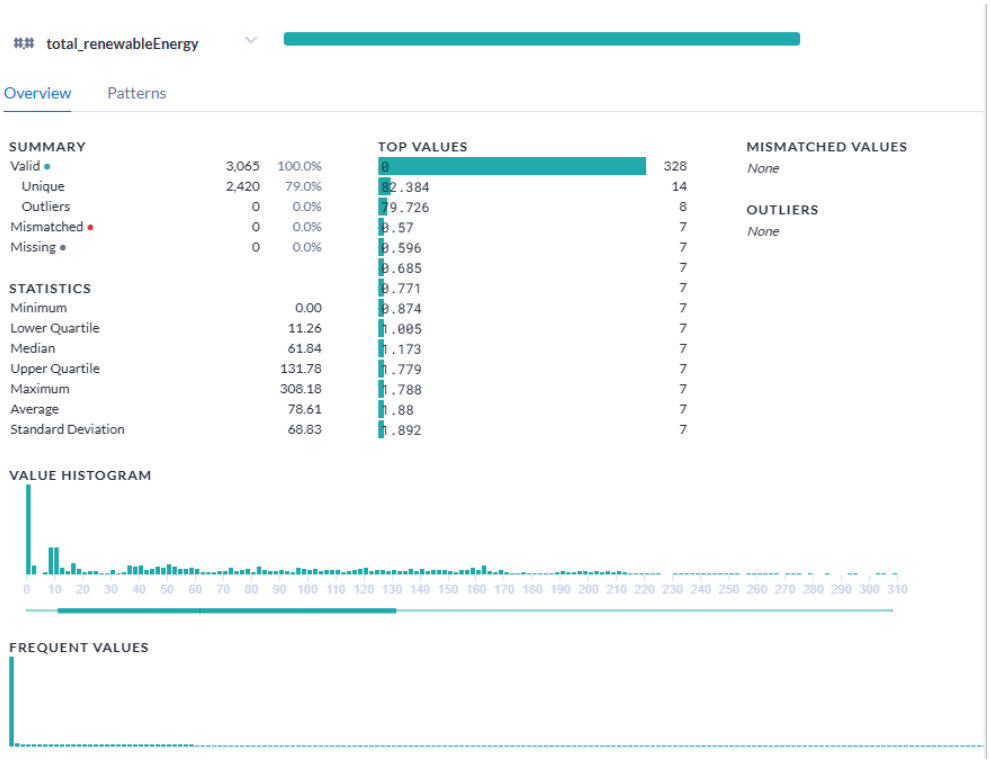


Figure 4.12: Original ‘total_renewableEnergy’ column detail.

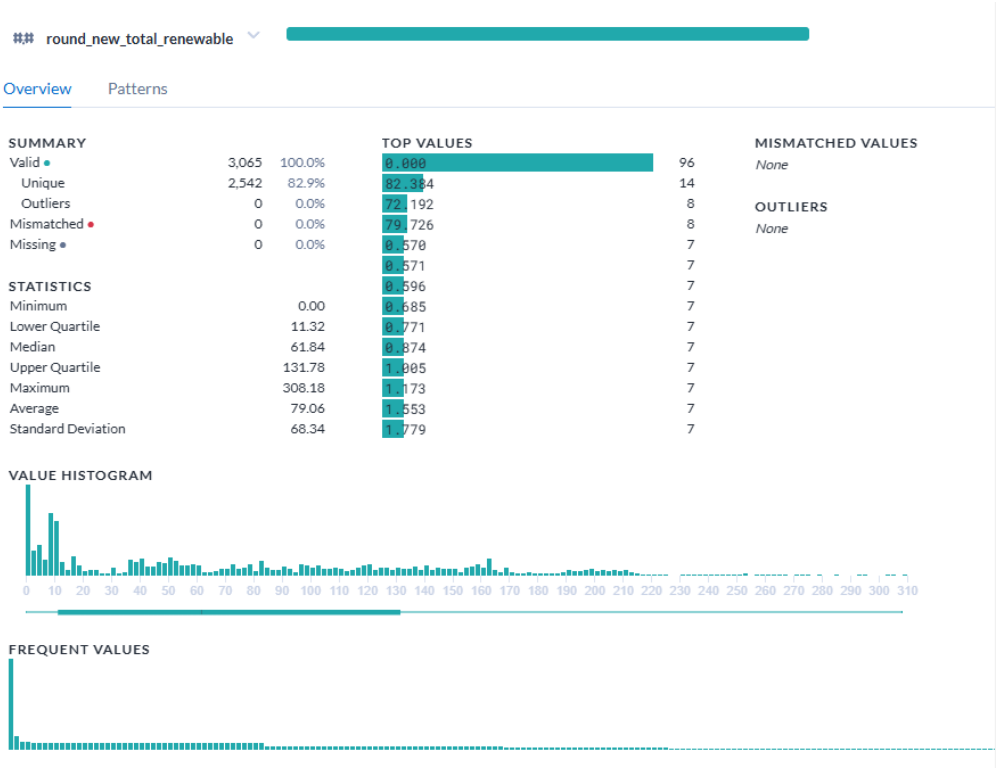


Figure 4.13: 'new_total_renewableEnergy' column detail. Number of 0 reduces to 96.

```
55 Create new_total_renewableEnergy from total_biofuels +
   geothermalEnergy + solarEnergy + total_biomassEnergy +
   total_hydroelectricPower + windEnergy

56 Create round_new_total_renewableEnergy from
   ROUND(new_total_renewableEnergy, 3)

57 Delete new_total_renewableEnergy

58 Delete total_renewableEnergy

59 Rename round_new_total_renewableEnergy to
   'new_total_renewableEnergy'
```

Figure 4.14: Breakdown of steps in recipe for creating new_total_renewableEnergy.

vi. Delete incomplete data

From the value histogram, we can observe that in 2024, the data is not complete. So, we will remove the data from 2024. After deleting 2024, the year column now consists of data from 1973 to 2023, number of rows is 3060.

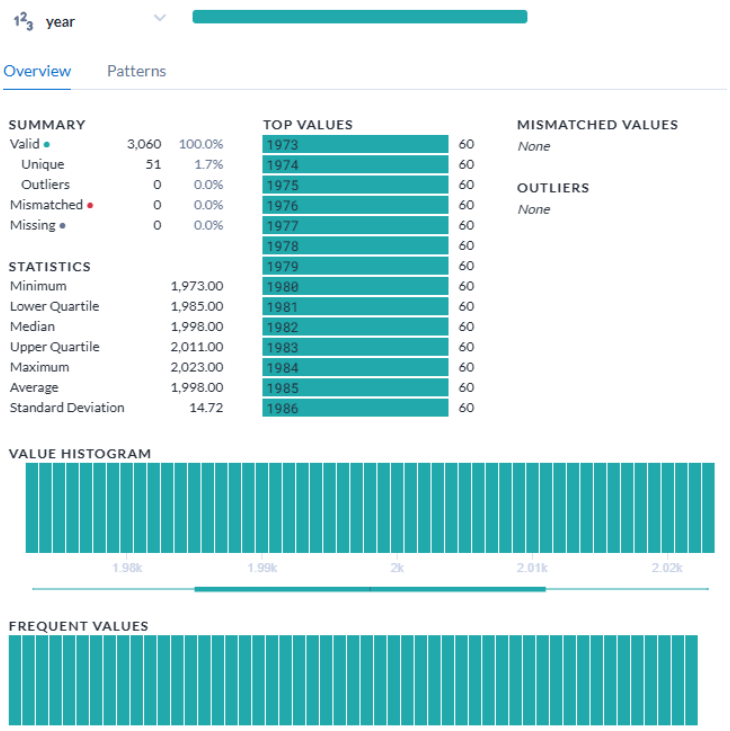


Figure 4.15: Column details for year.

vii. Overview

Schema Validation: This step checks the structure of the data, ensuring that the columns and their data types are correct. In this case, the validation has passed with 100% valid values, no mismatched values, and no missing values.

Transform with Profile: This step involves transforming the data using predefined operations, such as data cleansing, aggregations, and calculations. The profile step likely helps to examine the data distribution and quality. It shows that all values are valid with no issues.

Publish: This step involves writing the transformed data to the output destination, which is BigQuery. After transforming the data, it is published or stored in BigQuery for further analysis or reporting.

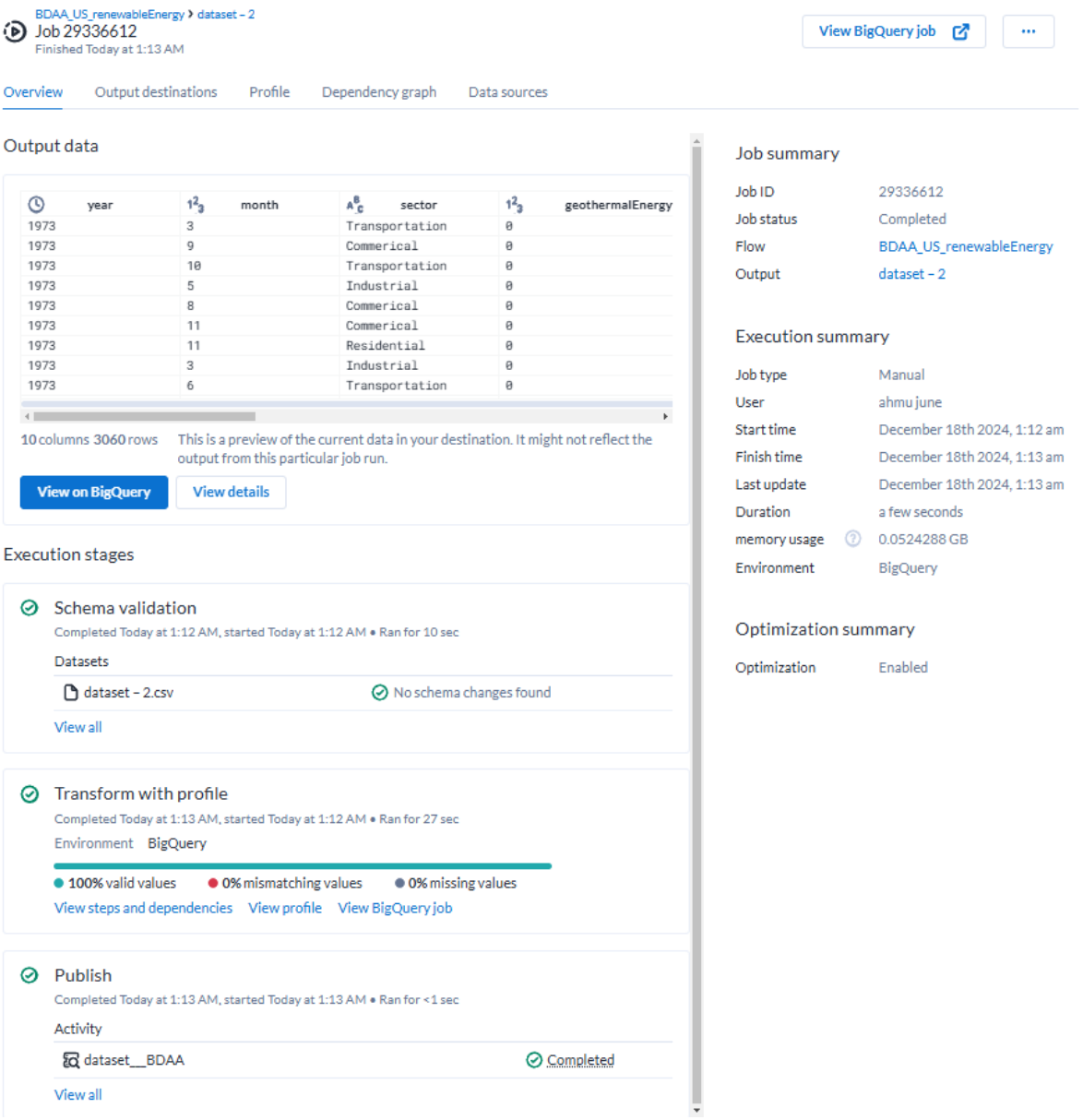


Figure 4.16: Job execution within Google DataPrep.

4.2 Data Visualization: Power BI

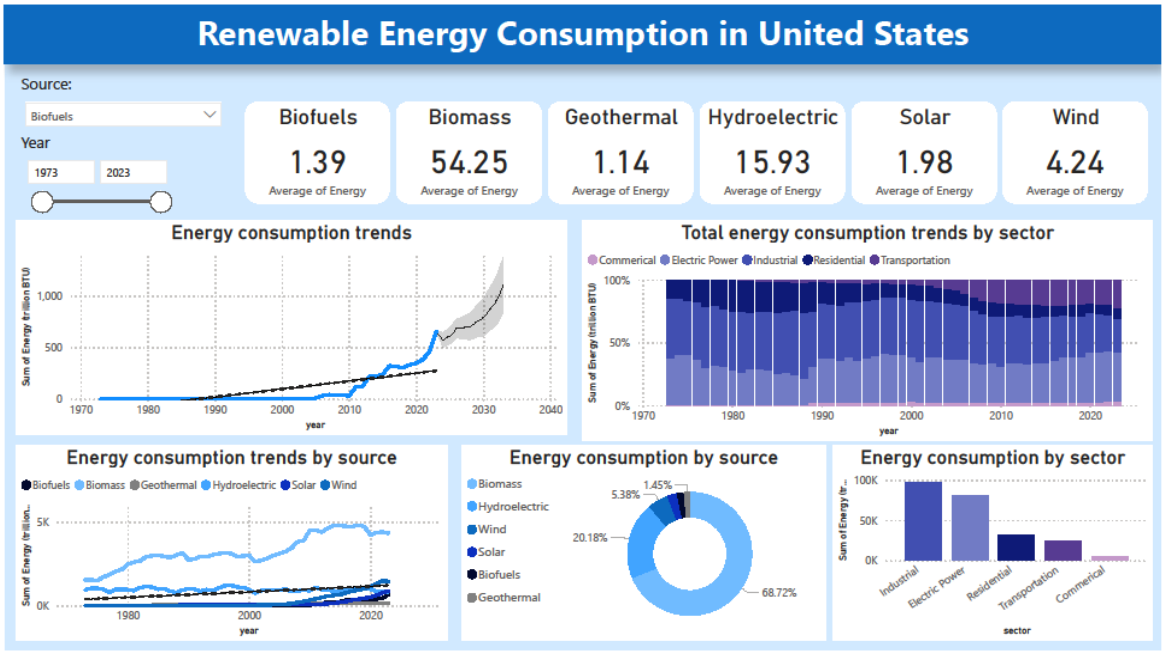


Figure 4.17: Dashboard created using Power BI.

The clean dataset is imported into Power BI for data visualization. Before the visualization starts, the columns with name geothermalEnergy, solarEnergy, windEnergy, total_biomassEnergy, total_hydroelectricPower, total_biofuels are renamed into Geothermal, Solar, Wind, Biomass, Hydroelectric and Biofuels, Wind. Then, these individual columns are unpivot and transformed into one column with name Energy consumption type. This purpose is to make the energy sources easier to setup for visualization that requires to show multiple energy sources in one graph.

For visualization, first the Source dropdown slicer on top left is used to edit the Energy consumption trends line graph only. By changing the dropdown value, the Energy consumption trends line graph will change accordingly. The Energy consumption trends line graph have a dotted line which shows the trend of energy consumption over the year and predicting energy consumption. The prediction shown in the Energy consumption trends line graph is 10 years which starts from 2023 to 2024, with the settings of 10 Seasonality and 95% of confidence interval.

Then, the year slicer with slider is used to edit the year of the data shown, it's applied to all the visualizations shown in the dashboard. By changing the year value, the average of

energy by every source and the 5 graphs below will display based on the year value input respectively.

Besides, the Total energy consumption trends by sector are visualized by using 100% stacked column chart. With this graph, energy consumption every year is always displayed 100%, total energy consumed by every sector can be clearly seen in this chart. For instance, electric power is the sector that uses the most energy in US. The industrial sector uses a large amount of energy used at the starts and slowly decreases over time. The transportation sector shows a low amount of energy used at the start but has increased drastically in the year 2008 and remains constant in the coming years.

Moreover, the Energy consumption trends by source line chart show the trends of the energy type used over the year, while the Energy consumption by source pie chart shows the percentage of the energy consumed. From the chart, we can observe that Biomass is the most used energy with peak usage nearly 5000 trillion BTU which occupied 68.72% out of the 6 energies. Vice versa, Geothermal energy is the least used energy with peak usage around 120 trillion BTU which occupied only 1.45% out of all the energies.

Lastly, the Energy consumption by sector bar chart shows the sum of the energy used by each of the sectors. Based on the bar chart, it indicates that the Industrial sector shows the highest energy consumption at around 100k trillion BTU, while Commercial sector shows the lowest energy consumption at only 5k trillion BTU.

Thus, by using this dashboard visualization, users can obtain the information of the average energy used over the year, energy consumption trend over the year, forecast of energy consumption in the next 10 years, total energy consumption trends by sector and source, and energy consumption by sectors.

5 Evaluation Metrics

The evaluation metrics were calculated programmatically using Google Colab and can be accessed through this link.

5.1 Data Preprocessing: Google Dataprep Evaluation

5.1.1 Data Preparation Performance

The performance of the data preparation process was evaluated by analyzing task-level execution times to identify potential bottlenecks in the pipeline.

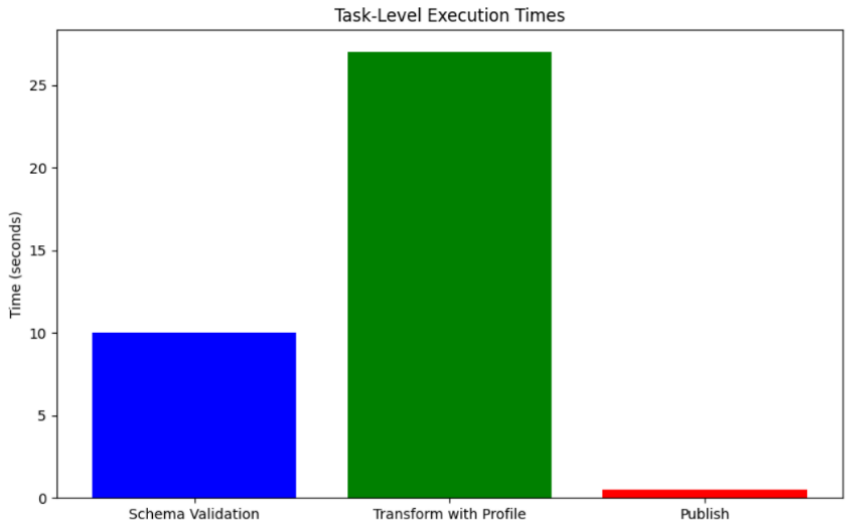


Figure 5.1: Bar chart analyzing the execution durations of different tasks in DataPrep.

A breakdown of task durations revealed that Transform with Profile had the longest execution time at 27 seconds. However, all task-level processes were completed in under one minute which is suitable for the size of the dataset. No optimization is required as the task durations are considered optimal for this context.

5.1.2 Transformation Accuracy

Three new columns — total_biomassEnergy, total_biofuels, and total_hydroelectricPower — were created by summing multiple columns and rounding the results to three decimal places. We aim to identify any discrepancies in these newly created columns by comparing the original calculated values with their rounded versions to three decimal places.

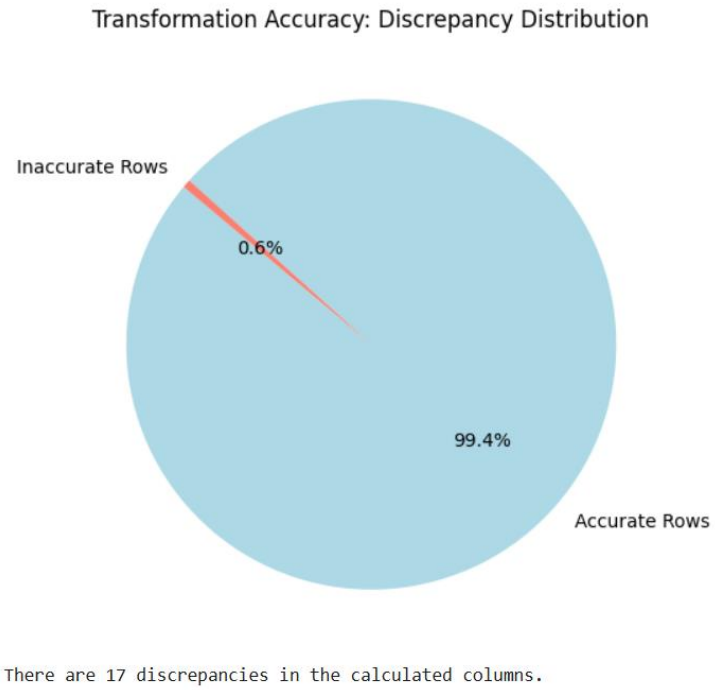


Figure 5.2: Pie chart showimg the discrepancy distributions of all the rows in the three newly created columns

The pie chart highlights 0.6% discrepancy distribution in the newly created columns involving of 17 inaccurate rows. Only minimal discrepancies were observed which confirmed accurate transformation and rounding processes to ensure data integrity throughout the data preparation pipeline.

5.1.3 Data Quality Validation

The descriptive statistics — specifically the mean and standard deviation — of the original dataset (using one of the original columns) and the cleaned dataset (featuring a newly created column derived from the summation of multiple columns) were compared.

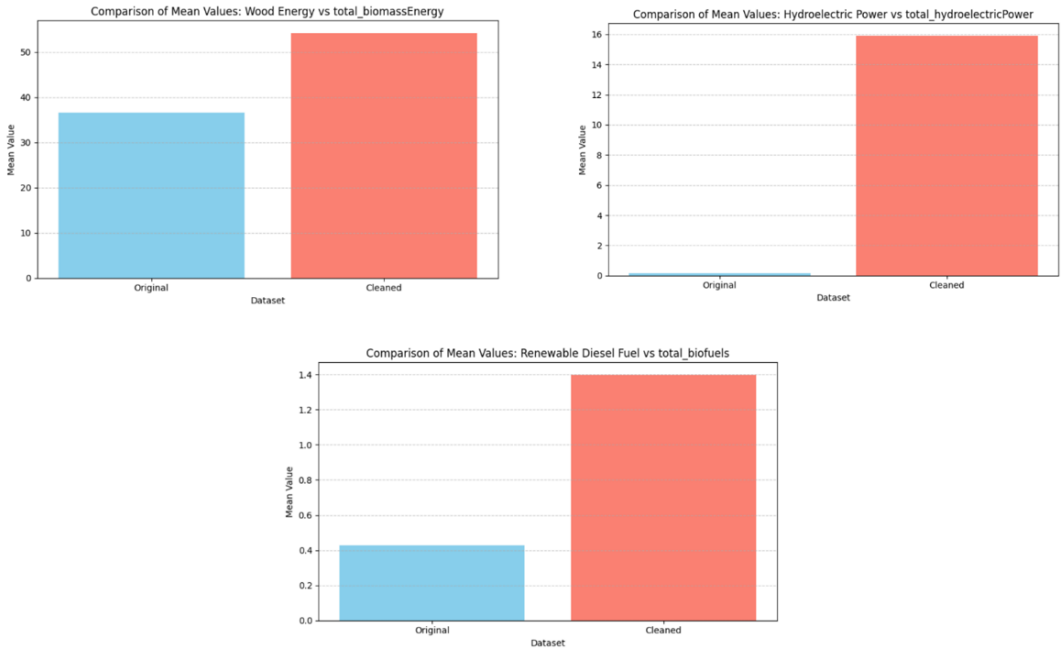


Figure 5.3: Bar charts showing the comparison of mean values between the original dataset and the cleaned dataset.

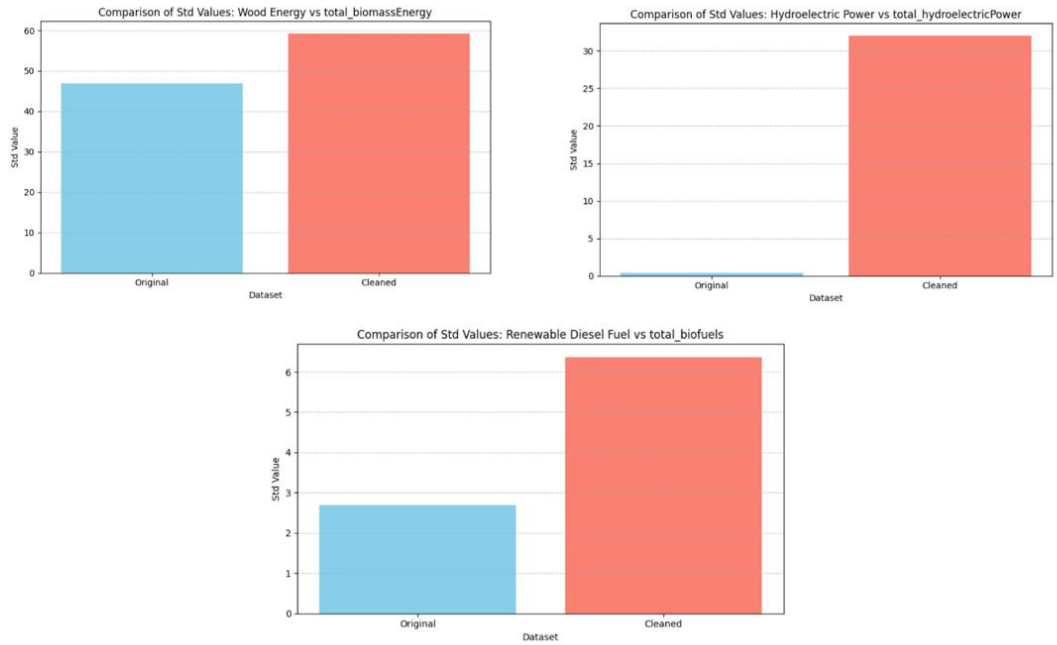


Figure 5.4: Bar charts showing the comparison of standard deviation values between the original dataset and the cleaned dataset.

The bar chart shows that the cleaned dataset increased value in mean and standard deviation for all columns. This indicates that the data transformation or summation of columns has altered the statistical properties of the dataset. A higher mean means that the cleaned data have effectively captured more comprehensive values that reflect a more accurate overall picture of total energy consumption. A higher standard deviation suggests increased variability around the mean indicating that the data spread is wider in the cleaned dataset compared to the original.

5.2 Data Visualization: Power BI Evaluation

5.2.1 Dashboard Interactive Performance

Data was extracted using Performance Analyzer in Power BI to assess the interactive performance of the dashboard and evaluate the event response times.

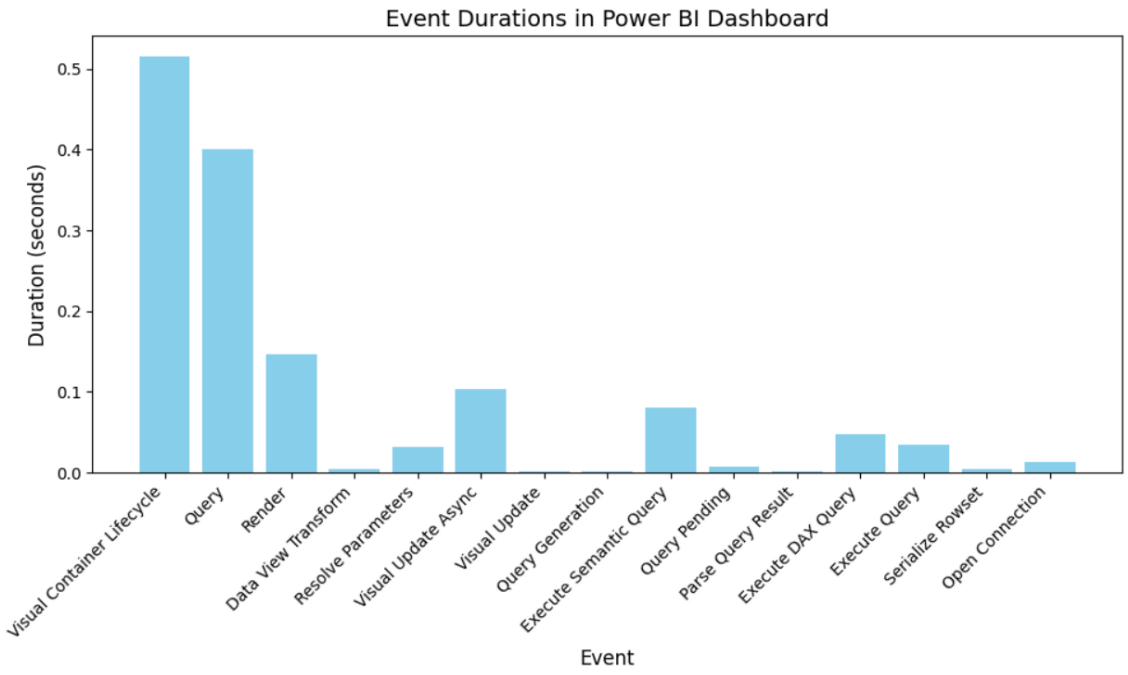


Figure 5.5: Bar chart displaying the event durations of Performance Analyzer.

The Visual Container Lifecycle that manages the rendering lifecycle of visuals takes the longest duration of 0.5 seconds. Next is Query that took 0.4 seconds which is likely due to data retrieval for filtered visualizations. Render which is responsible for drawing visual elements on screen took 0.15 seconds indicating efficient rendering process. Other events such as Data View Transform (data preparation for visuals) and Visual Update Async (asynchronous updates to visualizations) take only minimal time indicating streamlined transformation and update operations.

The duration of all events occurred in less than one second demonstrating optimal dashboard interactive performance. The smooth and responsive dashboard interactions ensure an effective and seamless user experience when applying filters and navigating between visuals.

5.2.2 Resource Usage Efficiency

The objective is to analyse the CPU and memory usage of the system during user interactions with the dashboard in 5 seconds intervals for 1 minute.

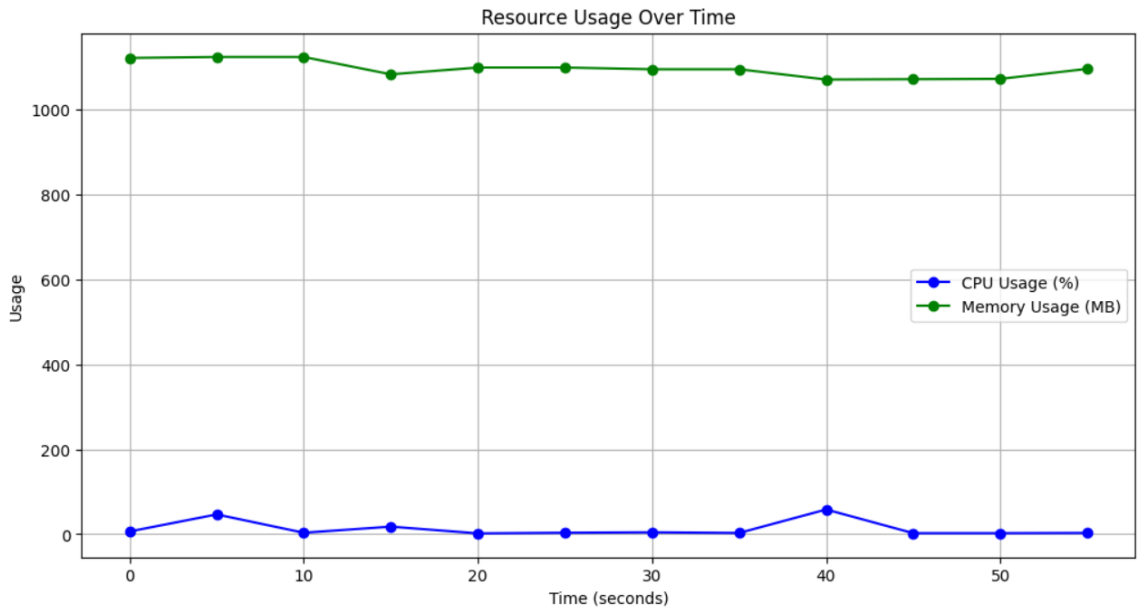


Figure 5.6: Line chart of CPU and memory usage of Power Bi over 1 minute.

The analysis of resource utilization indicates stable performance. No significant peaks or drops were observed that would necessitate increasing server capacity as the tasks remain non-resource intensive. Efficient resource management is crucial for maintaining system stability and optimal performance throughout dashboard usage.

5.2.3 Data Quality Validation

One of the graphs in the dashboard was validated by exporting the summarized data and was compared to the cleaned dataset to assess the consistency and accuracy of the energy consumption figures.

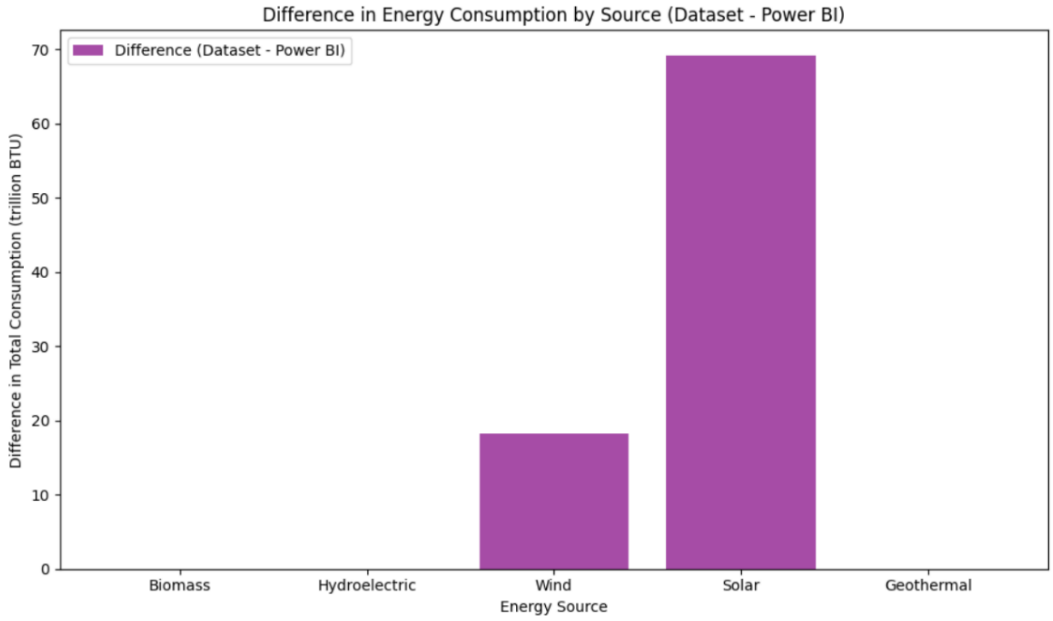


Figure 5.7:Bar chart of differences of energy consumption by source comparing data from cleaned dataset and data extracted from Power BI.

Minor discrepancies were reported in the exported data particularly in the total energy consumption for wind and solar energy sources. These differences are likely due to rounding and aggregation methods used in Power BI. Despite the minor variations, the dashboard still maintained high level of accuracy relative to the original dataset. This validation ensured that the visualization in the Power BI dashboard is based on reliable and consistent data.

6 Meeting Minutes Report

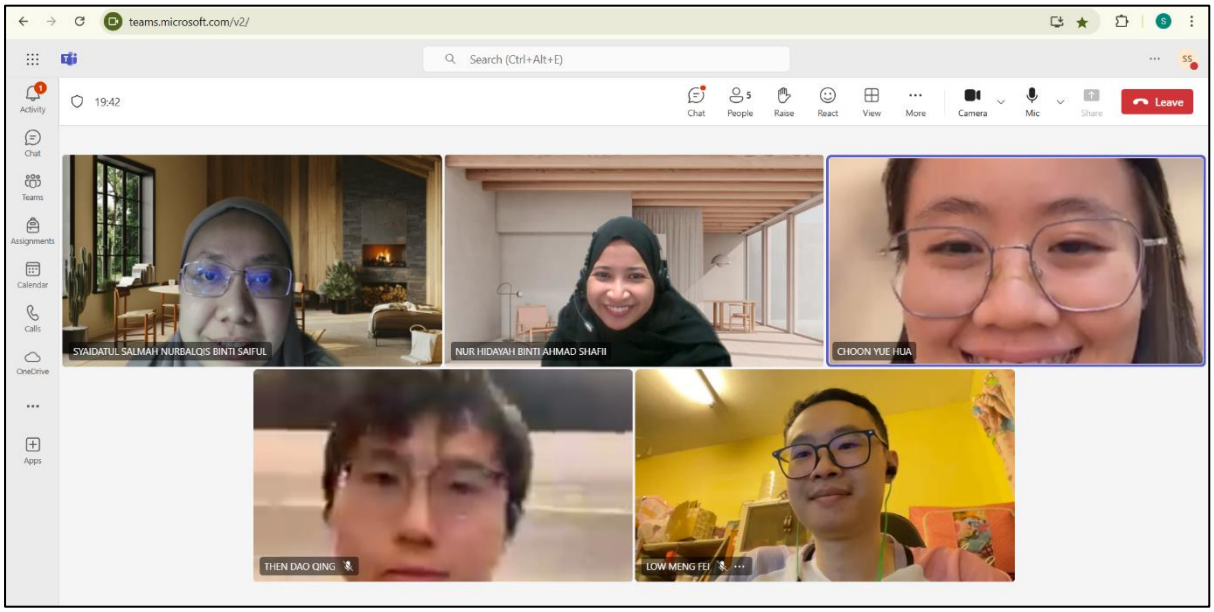
Meeting Minutes for Dataset Selection

Date: 8th December 2024 (Sunday)

Time: 8.00 p.m. - 9.00 p.m.

Participants:

- Nur Hidayah
- Then Dao Qing
- Choon Yue Hua
- Low Meng Fei
- Syaidatul Salmah



Agenda:

- Review and evaluate individual datasets.
- Discuss the relevance of each dataset for the group assignment.
- Select the most suitable dataset for the group assignment.

Discussion Points:

1. Size and Variety of Data

Each member presented their dataset that was used for the individual assignment:

1. Hidayah's dataset focused on agricultural data with potential insights into the impact of climate change on farming systems (link). However, the limited number of 10 columns and 50 rows might restrict the scope for diverse analyses.

2. Yue Hua's dataset was on renewable energy consumption in the U.S and is broken down by sectors and sources (link). The dataset has 17 columns and 3065 rows providing a variety of features for analysis.

3. Dao Qing's dataset also focused on renewable energy consumption (link). The dataset has 7 columns that might limit analysis but has a good number of records with 5695 rows.

4. Meng Fei's dataset was about climate insights that encompass a comprehensive collection of temperature records, CO2 emissions data, and sea level rise measurements (link). The dataset has limited column variety with 9 columns but has a good number of records with 7764 rows.

5. Salmah's dataset was on Japanese cherry blossoms has a comprehensive and rich historical data on phenology (link). The dataset size is small with 75 columns and 102 rows. It has a niche focus that might limit broader analysis about climate change.

Yue Hua's dataset stood out for its manageable size and variety which allows for both in-depth and broad analysis. Dao Qing's and Meng Fei's datasets are potential alternatives as both have a good number of records. Hidayah's and Salmah's datasets are smaller and have a highly specialized focus which is omitted from the project.

2. Data Quality

All datasets were deemed high-quality with no major issues in completeness or reliability reported during initial individual analyses.

3. Relevance to Domain

All datasets were relevant for the climate change or sustainability domain:

- Yue Hua's and Dao Qing's datasets are directly relevant as renewable energy is a critical area in sustainability.
- Meng Fei's dataset is directly relevant to climate change.

4. Data Exploration Potential

- Yue Hua's dataset has more columns that offer ample features to explore patterns in renewable energy consumption and trends over time.
- Dao Qing's and Meng Fei's datasets were seen to have limitation in exploration potential due to fewer variables.

5. Compatibility with Tools

All datasets were compatible with the tools that will be used for the project, such as Python, BigQuery, or HBase.

6. Potential for Innovation

- Yue Hua's dataset provides an excellent foundation for analyzing renewable energy trends and making predictions on renewable energy generation. The dataset's variety can support deeper statistical and machine learning analyses.
- Dao Qing's dataset offers potential to analyze renewable energy trends but with fewer columns, it might limit complex analysis.
- Meng Fei's dataset allows us to explore broader climate trends but lacks the column diversity to facilitate deeper analysis.

7. Interoperability

- Yue Hua's dataset has strong interoperability and can seamlessly integrate with other energy-related datasets. This can allow cross-domain analyses such as linking renewable energy trends to carbon emissions, if required. However, this will be confirmed in the next meeting based on the group's project goals.
- Other datasets may need additional preprocessing efforts to integrate with other datasets to increase data variety as they have fewer columns.

Decision:

The group unanimously agreed to use **Yue Hua's dataset (Renewable Energy Consumption in the U.S.)** for the group assignment based on:

- Strong relevance to sustainability domain.
- Balance of size and data variety.
- High data exploration potential.
- Flexibility for integration with other datasets (if needed).

Action Items:

- Yue Hua to share the dataset with all group members.
- Task division and role assignments to be finalized in the next meeting.

Minutes Prepared by: Syaidatul Salmah Nurbalqis Binti Saiful

Minutes Approved by: Nur Hidayah Binti Ahmad Shafii

7 References

Delmas, M., & Montes-Sancho, M. (2011). US State Policies for Renewable Energy: Context and Effectiveness. *Energy Policy*, *39*, 2273–2288. https://doi.org/10.1016/j.enpol.2011.01.034

Foidl, H., Golendukhina, V., Ramler, R., & Felderer, M. (2024). Data pipeline quality: Influencing factors, root causes of data-related issues, and processing problem areas for developers. *Journal of Systems and Software*, *207*, 111855. https://doi.org/https://doi.org/10.1016/j.jss.2023.111855

Olabi, A. G., & Abdelkareem, M. A. (2022). Renewable energy and climate change. *Renewable and Sustainable Energy Reviews*, *158*, 112111. https://doi.org/https://doi.org/10.1016/j.rser.2022.112111

Vishvakarma, T. (2022). Climate Change Forecasting using Machine Learning Algorithms. In *International Journal for Research in Applied Science & Engineering Technology (IJRASET)* (Vol. 10). www.ijraset.com