CSCI 6401-01
DATA MINING

Phase 4: Data exploration
Team: TSquad

1) **Team Members:**
    1. Kotha Priyatham Prem Kumar – pkoth10@unh.newhaven.edu
    2. Yamana Venkata Sai Sushmanth - vyama2@unh.newhaven.edu

Github link: https://github.com/Premkumar5225/TSquad/

2) **Research Question:**

This research question aims to investigate the feasibility of developing a predictive model that leverages insights from electric vehicle charging behavior to estimate the average charging time for a given zip code. By analyzing charging patterns and behavior data, this study seeks to provide a practical tool for users and stakeholders in the electric vehicle ecosystem to better plan and optimize charging experiences based on geographical location.

Dataset link:
https://dataverse.harvard.edu/file.xhtml?persistentId=doi:10.7910/DVN/NFPQLW/EQRRQH&version=1.0

The chosen dataset focuses on electric vehicle (EV) charging behavior, offering insights into 3,395 high-resolution charging sessions. This dataset encompasses 85 distinct EV drivers participating in a workplace charging program across 25 sites. These sites range from research and innovation centers to manufacturing, testing facilities, and office headquarters. The initiative is a part of the U.S. Department of Energy (DOE) workplace charging challenge. Data is available in a human and machine-readable CSV format, ensuring ease of import into various software platforms. The dataset provides detailed information with a resolution down to the nearest second, aligning with the analysis undertaken in this study.

Accessibility: The dataset is openly accessible and available for download from the designated repository or source. It does not require any special permissions or credentials for retrieval.

Data Collection Methods: The data was collected through a combination of on-site charging station monitoring and driver interaction with the charging infrastructure. This likely involved the deployment of monitoring equipment on the charging stations, as well as the incorporation of sensors to record session-specific data. Driver interaction may have been facilitated through RFID cards or mobile applications associated with the charging program.

Data Type: The dataset is structured and organized in a tabular format, with each row representing a distinct charging session. The columns encompass various attributes including timestamps, charging station IDs, driver information, location details, session durations, charging power levels, energy consumption, and potentially environmental conditions. These attributes are likely to be represented in appropriate data types, such as datetime for timestamps, numerical types for quantitative measurements, and categorical variables for driver and station identifiers.

3) In this project, we work with a dataset consisting of 3395 entries and 24 columns. Each column provided specific information about the sessions, including attributes such as session ID, total kilowatt-hours (kwhTotal), cost in dollars, timestamps for session creation and end times, charging duration in hours, weekday, platform, distance travelled, user ID, station ID, location ID, manager vehicle information, facility type, and individual days of the week.

Our exploration of the dataset involved several key techniques:

1. Pre-processing:

- We initiated the data exploration by undertaking a crucial pre-processing phase. This involved the removal of unnecessary data from the dataset and structuring it into meaningful columns, which included Session ID, Kwh Total, start time, end time, charge time, distance, station id. This step was instrumental in converting raw data into a format that was conducive for further analysis.

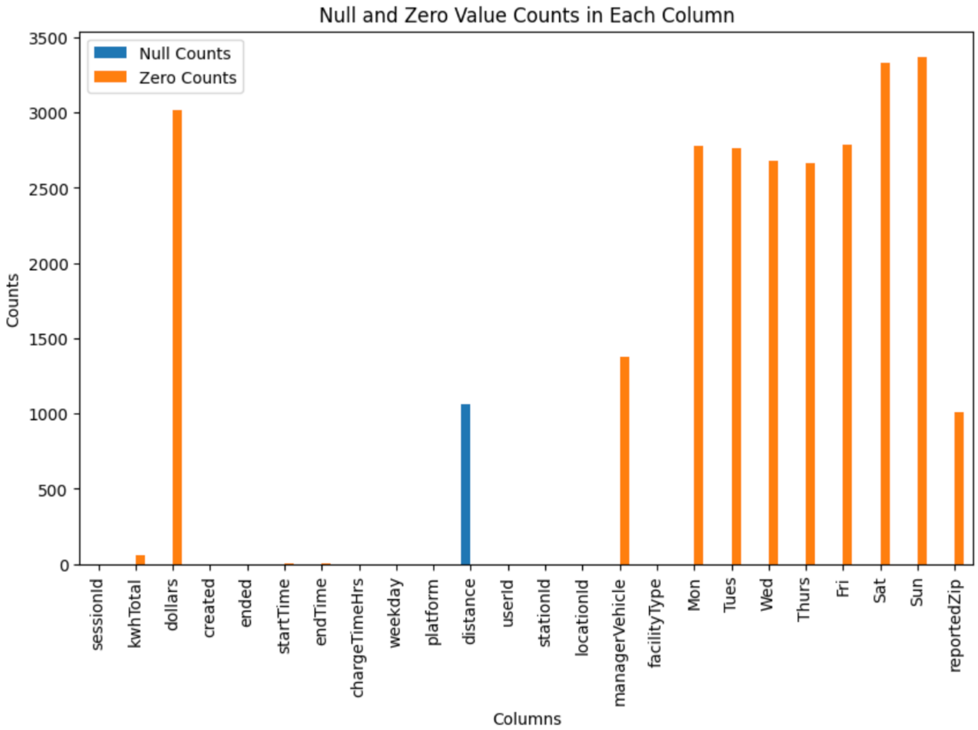
2. Data Cleaning:

- Data cleaning was a critical step where we removed any unnecessary or redundant information, ensuring that the remaining data was structured into well-defined columns. This process facilitated a more streamlined and effective analysis.

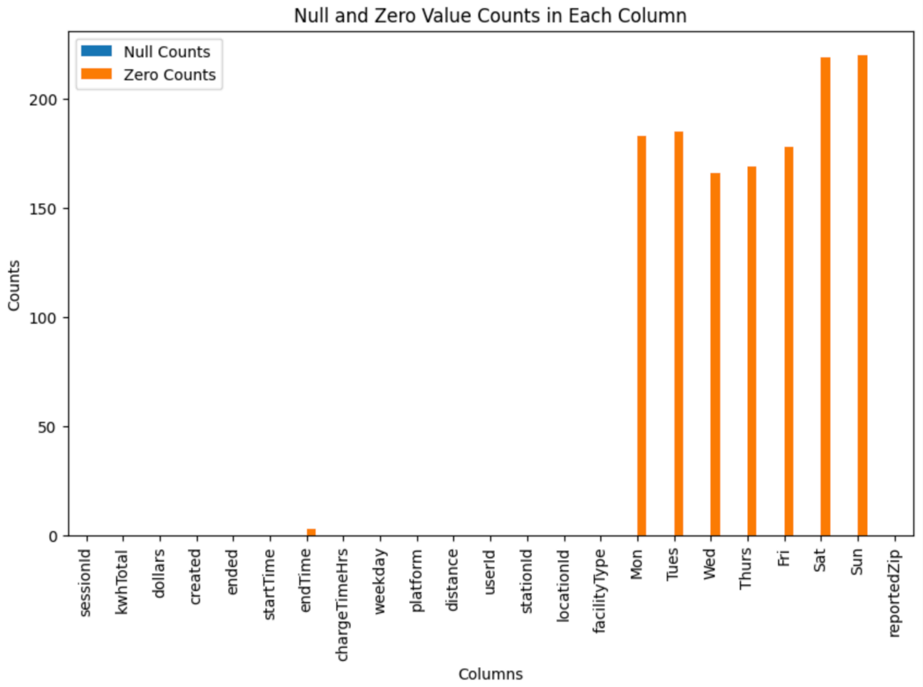
3. Null Value Handling:

- Any null values and empty cells in the dataset were identified and subsequently removed using the 'NOT NULL' function. This ensured that our analysis was based on complete and reliable data.

Before null value removal:

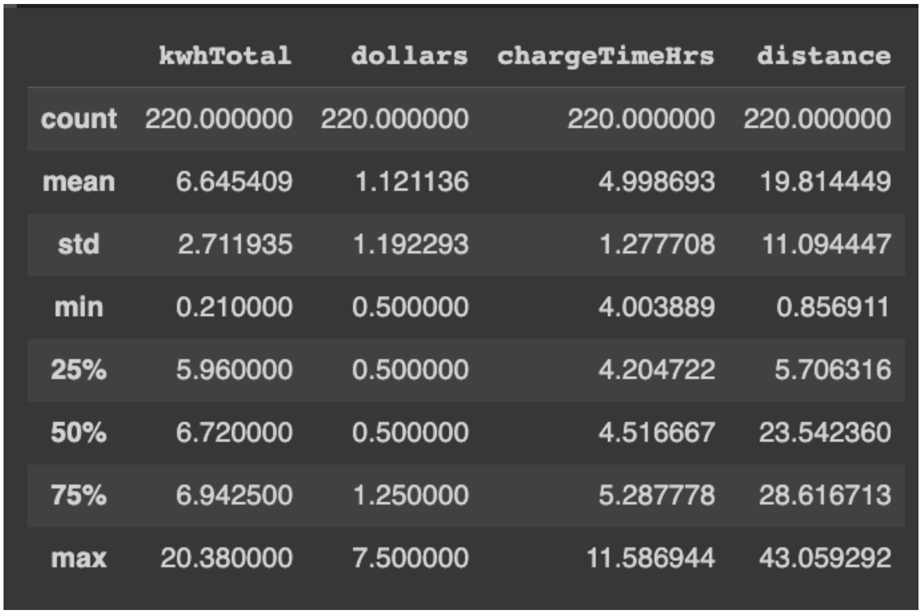


After removal of null values:



4. Descriptive Statistics:

- We computed descriptive statistics to gain a general overview of the data, including measures like mean, median, mode, and standard deviation. These statistics provided valuable insights into the central tendencies and variability within the dataset.



	kwhTotal	dollars	chargeTimeHrs	distance
count	220.000000	220.000000	220.000000	220.000000
mean	6.645409	1.121136	4.998693	19.814449
std	2.711935	1.192293	1.277708	11.094447
min	0.210000	0.500000	4.003889	0.856911
25%	5.960000	0.500000	4.204722	5.706316
50%	6.720000	0.500000	4.516667	23.542360
75%	6.942500	1.250000	5.287778	28.616713
max	20.380000	7.500000	11.586944	43.059292

5. Data Visualization:

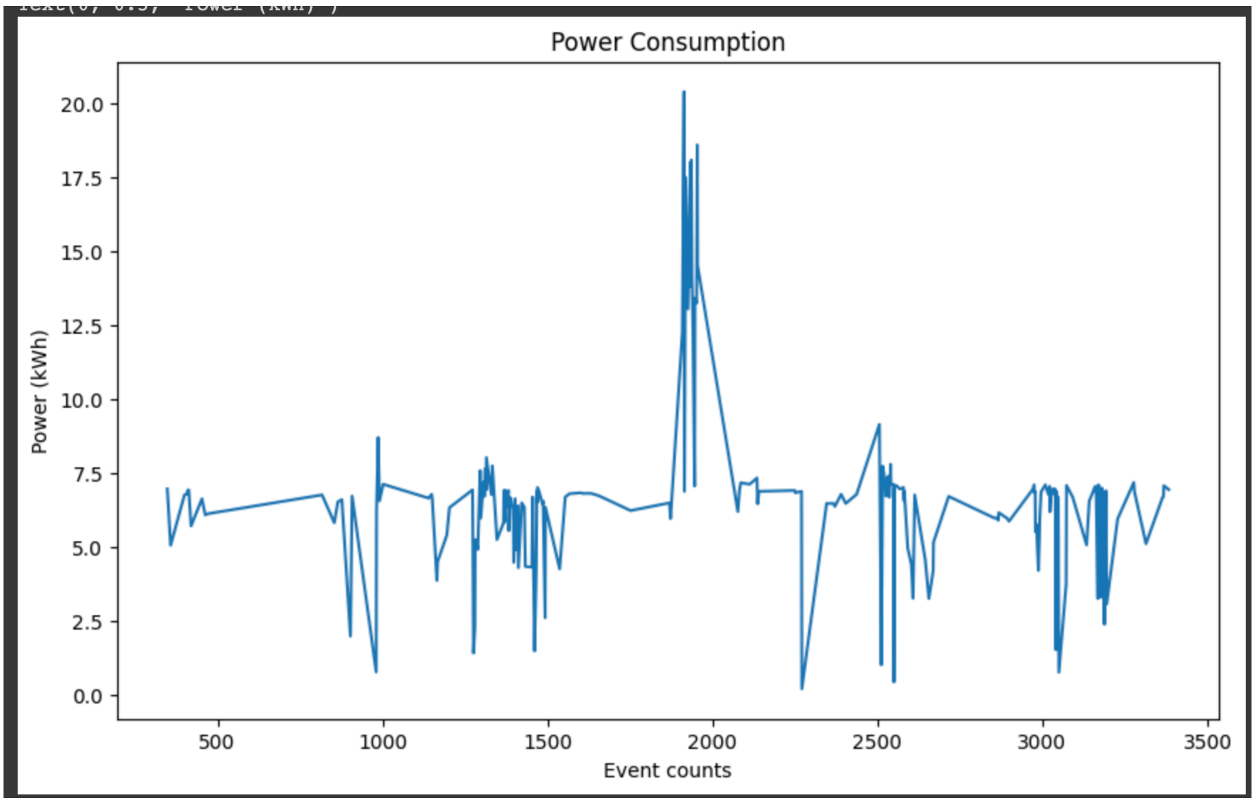
- To enhance our understanding of the data, we leveraged data visualization techniques. This involved creating various charts and graphs, including scatter plots bar, graphs, heatmap These visual representations provided a clearer picture of the trends and patterns in the data.

6. Cross-tabulation:

- We employed cross-tabulation to analyse the relationships between two or more variables. This technique allowed us to create contingency tables, enabling a deeper exploration of the interplay between different attributes.

These techniques collectively provided us with valuable insights and a comprehensive understanding of the dataset.

4) Please describe your data explorations from different perspectives using varied visualization techniques such as tables and charts. Finally, you should conclude your data exploration in a paragraph, which describes your findings based on the data exploration.

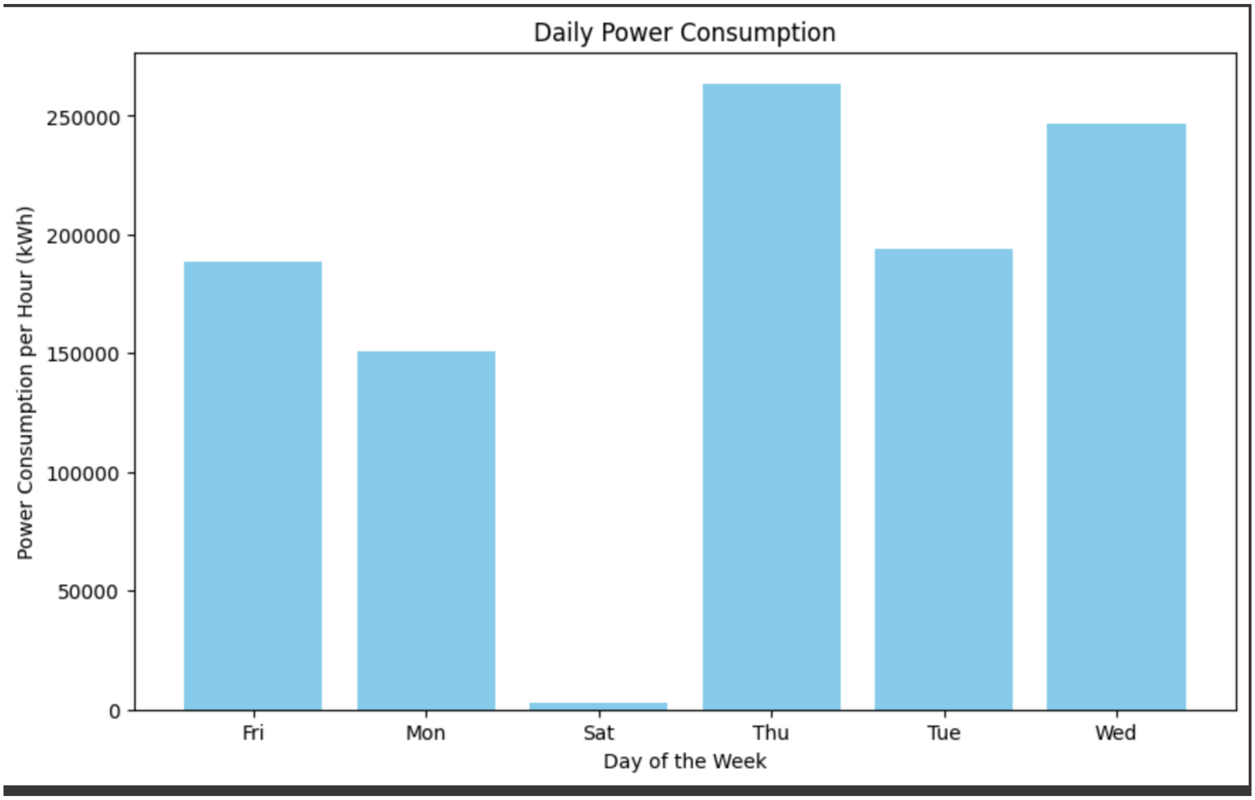


```
Text(0, 0.5, 'Power (kWh)')
```

□ The x-axis represents event counts, while the y-axis represents power consumption in kilowatts.

□ The blue line on the graph shows how power consumption changes with the number of events. There's a peak in power consumption at around 1500 event counts, after which it gradually decreases.

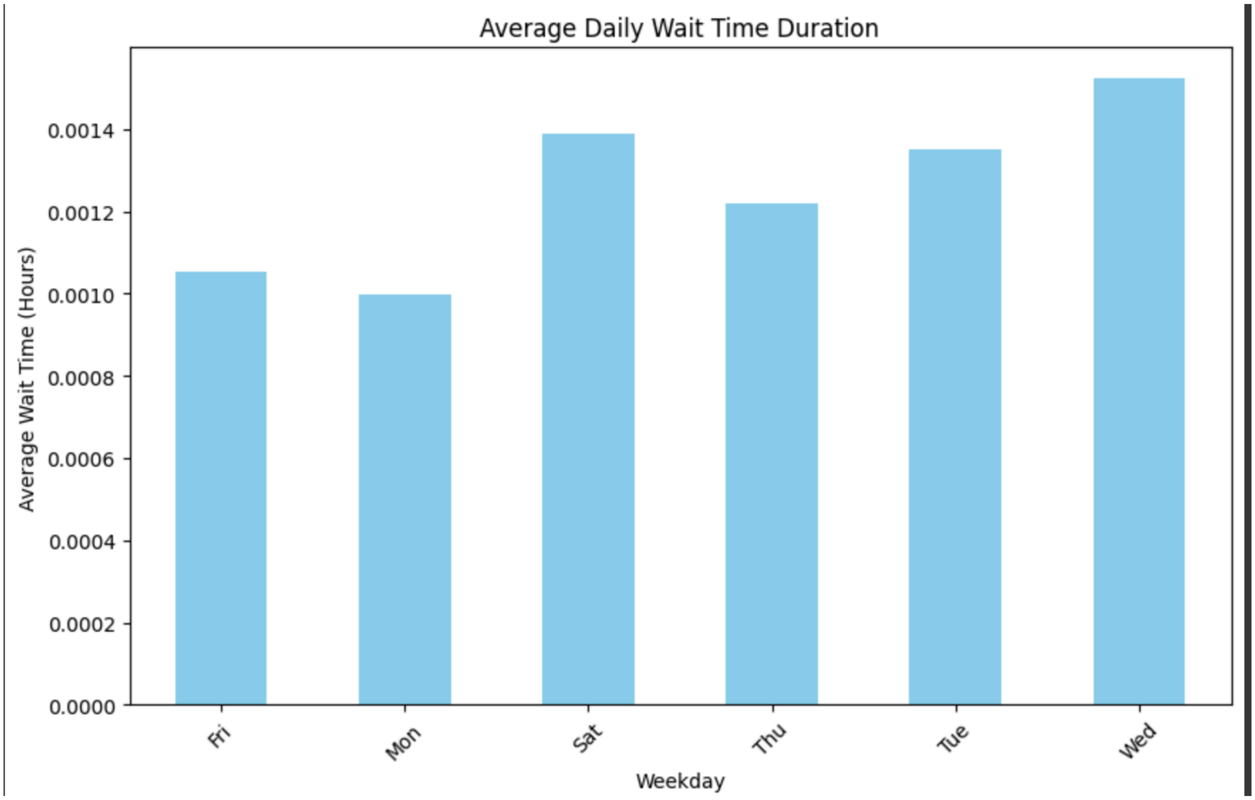
□ This could represent a system where power usage increases with the number of events up to a point, after which efficiency measures or other factors cause the power usage to decrease even as more events occur.



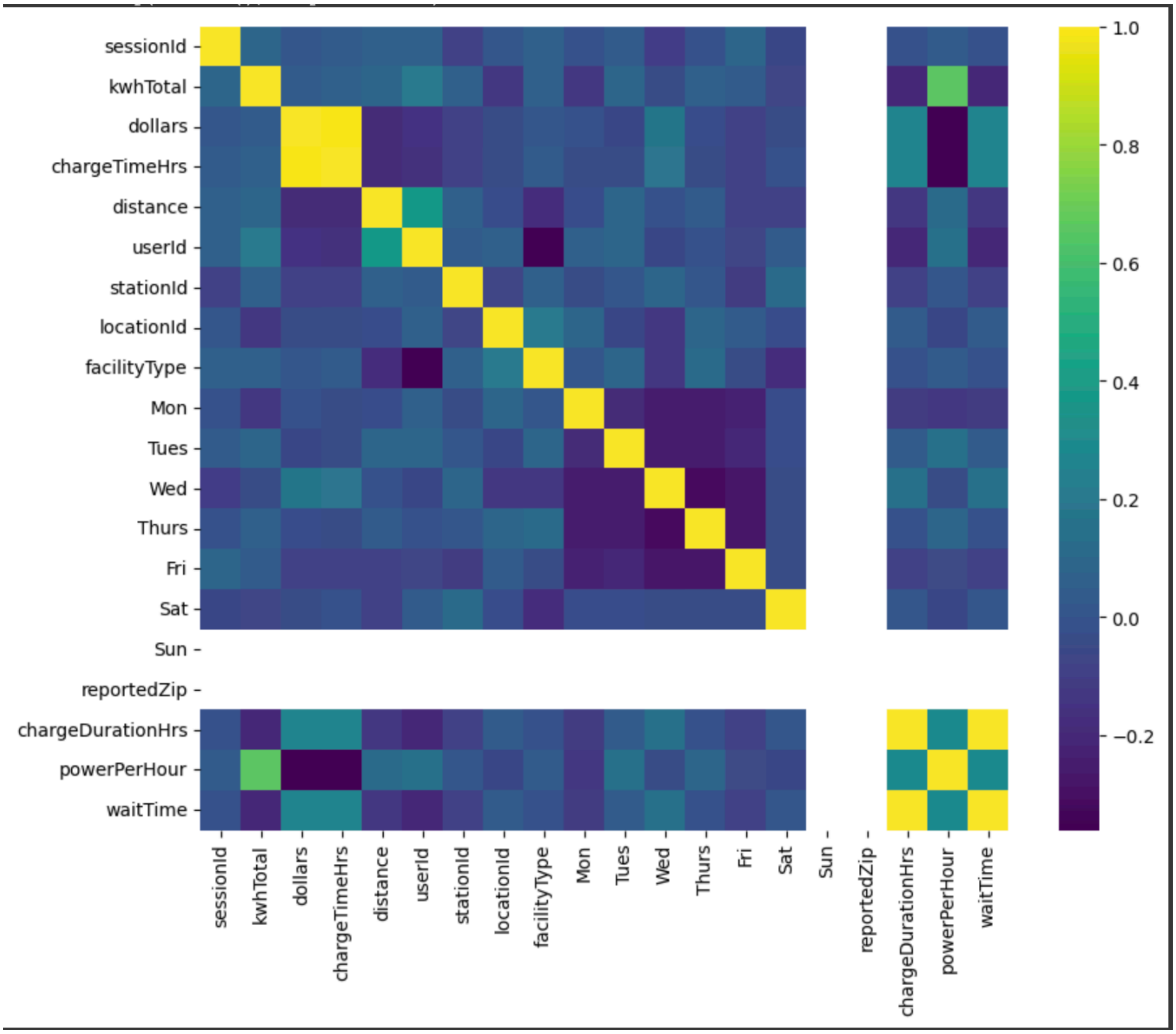
□ Daily Power Consumption It shows the power consumption in kilowatt hours (kWh) for each day of the week.

□ From the graph, we can see that power consumption varies throughout the week. The highest consumption occurs on Wednesday, while the lowest is on Saturday. This could suggest that whatever system or activity this data is tracking uses more power in the middle of the week and less towards the end of the week.

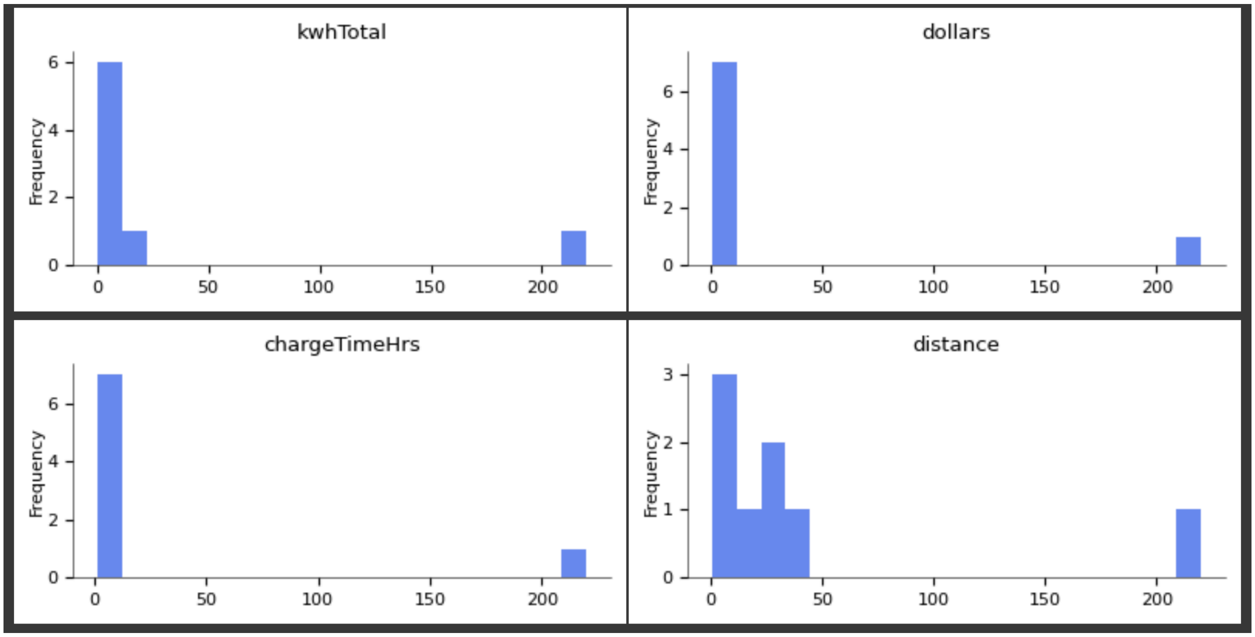
□ This data can be very useful for energy management. By understanding when power usage is at its highest and lowest, it’s possible to make adjustments to activities or systems to reduce overall energy consumption. For example, non-essential activities could be scheduled for times when power usage is typically lower.



- Average Daily Wait Time Duration It shows the average wait time in hours for a particular service for each day of the week.
- From the graph, we can see that the wait times vary throughout the week. The highest wait time occurs on Wednesday, while the lowest is on Friday. This could suggest that the service is busiest in the middle of the week and less busy towards the end of the week.
- This kind of data can be very useful for service management. By understanding when wait times are at their highest and lowest, it's possible to make adjustments to staffing or processes to reduce overall wait times.

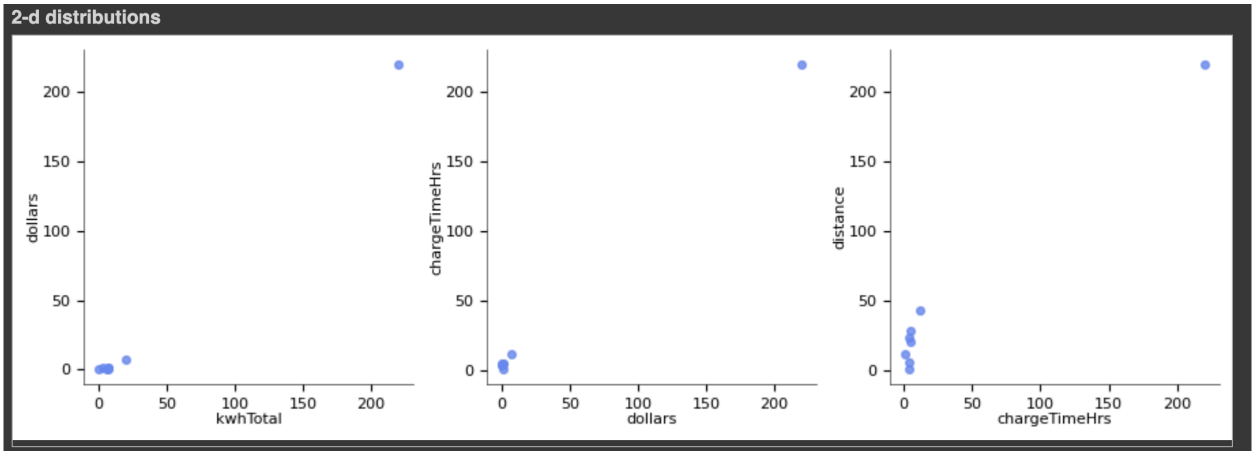


- heatmap is showing the correlation between different variables.
- The variables include sessionID, kWhTotal, dollars, distanceInMiles, userID, locationID, facilityType, Mon, Tues, Wed, Thurs, Fri, Sat, Sun, reportedZip, chargeDurationHours, powerInkW, and waitTime.
- Each square in the heatmap represents a pair of variables. The color of the square indicates the strength of the correlation between those two variables.
- Darker colors represent stronger correlations and lighter colors represent weaker correlations.
- This kind of chart can be very useful for identifying patterns and relationships in large datasets. For example, if there's a strong correlation between two variables represented by a dark square, it might suggest that those variables are closely related in some way.



There are four histograms representing four different variables: kWhTotal, dollars, chargeTimeHrs, and distance:

- This histograms represents the distribution of total kilowatt-hours (kWh) consumed, cost in dollars, charging time in hours, distance covered. Each bar in the histogram represents the frequency number of occurrences of a particular range of kWh values.



This image shows 2-dimensional distributions of three variables: kWhTotal, dollars, and charge time in hrs.

- **kWhTotal**: This panel represents the distribution of total kilowatt-hours (kWh) consumed. The x-axis represents the range of kWh values, and the y-axis represents the frequency of those values. Each blue circle represents a data point.
- **dollars**: This panel represents the distribution of cost in dollars. The x-axis represents the range of dollar values, and the y-axis represents the frequency of those values. Each blue circle represents a data point.
- **Charge Time in hrs**: This panel represents the distribution of charging time in megawatt-hours (MWh). The x-axis represents the range of charge time values, and the y-axis represents the frequency of those values. Each blue circle represents a data point.

From our exploration of the data, we've uncovered some noteworthy patterns. We observed a peak in power consumption at around 1500 event counts, followed by a gradual decrease. This suggests that power usage varies over time. Specifically, Wednesdays see the highest consumption, while Saturdays have the lowest. This insight implies that the system or activity tracked in this data tends to use more power in the middle of the week and less towards the end. This information is invaluable for effective energy management. By identifying peak usage times, we can adjust activities or systems, potentially reducing overall energy consumption. Additionally, we noticed variations in wait times across the week, with the longest waits occurring on Wednesdays and the shortest on Fridays. This indicates that the service experiences higher demand mid-week and less towards the weekend. This knowledge can be leveraged for service management, allowing for staffing or process adjustments to minimize overall wait times. The heatmap visually represents correlations between different variables, offering insights into their relationships. Each square in the heatmap corresponds to a pair of variables, with color intensity indicating the strength of their correlation. This analytical approach provides a deeper understanding of how these factors interrelate.