**Chapter 3: Data Collection and Pre-Processing**

**3.1 Overview**

This chapter outlines the collection, organization, and preparation of the data used to build the predictive energy density model in battery materials. Beginning with a description of the collected dataset ([Huang, S. and Cole, J. (2020)](https://shorturl.at/vGOmd)), this section also describes the workflow, which involved multiple stages, including data extraction from raw sources, transformation and cleaning through SAS and Python scripts, and consolidation into a final dataset suitable for analysis. Each workflow phase—from raw data acquisition to final dataset creation—is detailed below.

**3.2 Data Description**

The initial dataset, composed of raw battery material data, includes properties such as Voltage, Capacity, Coulombic Efficiency, Conductivity, and Energy. Each row describes a property of a compound/chemical composition and includes information regarding the Type of battery material (cathode, anode). Each row also has the compound's chemical name and composition, which was greatly useful in the cleanup process and model building alike. The initial dataset had about 292,000 data points. This was a huge amount of raw data; unsorted, full of missing data, and completely void of any links between each row. Even if a single compound had all the data, it was spread out across multiple rows and with many repetitions. It wasn’t possible to build any machine learning model based on this dataset. To solve this problem, I uploaded the data to SAS OnDemand for Academics platform. I chose this platform as it provides a structured environment with built-in tabular manipulation techniques. With SAS, I filtered the table into 5 separate tables based on the properties of the chemical compounds:

1. **Capacity**: Included charge/discharge capacities measured across various materials.
2. **Voltage**: Records of voltage readings of different battery materials.
3. **Coulombic Efficiency**: Represents the ratio of charge output to input, indicating energy conversion efficiency of each material.
4. **Conductivity**: Measured the ionic/electrical conductivity of materials under test.
5. **Energy**: Records of the energy density of various battery materials.

Each of these tables had key columns such as the name of the compound, the Extracted Name (the chemical composition), the type of battery material, the specific property that the table contained, and the unit of measurement of that property. These tables were exported as individual CSV files for further pre-processing in Python.

**3.3 Data Cleaning**

The data cleaning process involved several steps, including attempts that were not very favorable. Each exported CSV file underwent multiple steps of cleaning by utilizing Python scripts and was then finally merged for further analysis and model building. All the steps for cleaning and merging are given below:

**Data Preparation and Cleaning Workflow**

The initial phase of dataset preparation was conducted using a Python script (data\_cleaning.py) focused on removing a substantial volume of incomplete records. The original dataset comprised approximately 290,000 rows; however, after rigorous filtering of entries with missing or null values, this number was reduced to around 1,000 rows. This aggressive pruning ensured that subsequent analysis would be based on complete and reliable data.

Following the identification and removal of incomplete entries, cleaned subsets of the data were extracted into distinct tables. Each table corresponded to a specific material property—such as voltage or capacity—and was stored in a CSV file named using the convention “xxxx\_clean.csv,” where “xxxx” denotes the property in question. Simultaneously, the rows with missing or NaN entries were isolated and preserved in separate datasets for potential use in imputation or secondary analysis. Throughout this process, care was taken to retain the original column structure across all derived tables to ensure consistency.

**Feature Engineering: Molecular Weight Calculation**

An essential aspect of early feature engineering was the computation of the molecular weight for each chemical compound. This task was carried out using a separate Python script (data\_pre\_processing.py). The dataset included a column labeled Extracted Name, which contained the chemical formula for each compound. These formulas were parsed using regular expressions and string manipulation techniques to extract elemental components and their respective counts.

To enable this calculation, a predefined dictionary mapping atomic symbols to their atomic weights (e.g., H = 1.008, O = 16.00) was employed. For each compound, the molecular weight was computed by summing the weighted contributions of individual atoms based on their frequency within the formula. This derived feature added a quantifiable measure of chemical complexity to the dataset, serving as a valuable input variable for machine learning models aimed at predicting energy density. Furthermore, the inclusion of molecular weight enabled the use of a joint key—comprising Compound\_name and Molecular\_weight—to uniquely identify compounds across datasets.

**Data Integration Attempts**

With the availability of cleaned and enriched tables, efforts were made to integrate data across the five key property datasets. Two Python scripts were written for this purpose. The first, data\_combining\_vertical.py, attempted vertical stacking to unify entries across different battery material properties. This approach, however, was unsuccessful due to mismatched schemas and inconsistent compound identifiers across the files.

A more refined strategy was implemented in data\_merging\_for\_sql.py, which aimed for horizontal merging by aligning records using compound names or molecular weights. Despite being a more logical method for data fusion, this too failed due to naming inconsistencies and the absence of reliable relational keys. These failed integration attempts underscored the need for stricter standardization of column names and the elimination of irrelevant or redundant fields in each table.

**Dataset Refinement and Standardization**

The script further\_cleaning.py was developed to refine and harmonize the structure of the datasets. This included the application of consistent naming conventions across all tables and the removal of non-informative columns such as internal IDs and measurement units. Only essential fields were retained: Compound\_name, Extracted\_name, Molecular\_weight, Type (designating anode or cathode), and Property (such as voltage, capacity, conductivity, coulombic efficiency, or energy). Furthermore, data type conversions were applied to ensure that all numerical columns were appropriately formatted. These steps significantly improved the internal consistency and usability of the data.

**Normalization and Deduplication**

To further enhance data integrity and compatibility for merging operations, unique compound entries were extracted using the script find\_distinct\_compounds.py. This involved filtering each table to retain only distinct combinations of Compound\_name and Molecular\_weight. The output consisted of five de-duplicated CSV files, each representing one of the material property datasets. These normalized tables facilitated clearer compound-level analysis and laid the groundwork for final data consolidation.

**Final Merging, Feature Enhancement, and Cleanup**

The final phase of the data preparation pipeline involved two additional scripts: distinct\_data\_merging.py and final\_cleaning\_steps.py. The first successfully merged the five normalized tables into a single comprehensive dataset using an outer join on the compound name and molecular weight. This ensured the inclusion of all possible records, even those with partially missing property values. Subsequent refinements were executed in final\_cleaning\_steps.py. This included the removal of the Conductivity column due to an excessive number of missing entries. The Type column was standardized by replacing inconsistent labels—for instance, rows labeled "Cath" were updated to "Cathode," and those labeled "Anod" were corrected to "Anode." Such standardization was crucial to mitigate downstream issues related to categorical variable encoding and interpretation.

Moreover, all remaining rows containing NA values were removed to ensure the dataset was model-ready. A final and significant feature engineering step involved the computation of the Energy Density column, defined as:

This step was carried out in the later stages of cleaning, to impute the missing value for the Energy property of various materials. As Energy density itself is the most crucial column, it was necessary to have all relevant columns cleaned and free of NULL values. The final dataset was a complete one with 832 entries. This would be large number of data points to carry out model building and any further analysis. From now onwards, this dataset will be referred to as the ***prime dataset***.

The final cleaning steps were carried out one more time with a different approach, where instead imputing null values of the Voltage column (which happens to be one of the factors for calculating the Energy density) the null values were dropped altogether. This created an entirely different data set consisting of only 339 data points. From now onwards, this dataset will be referred to as the ***concise dataset***.

**3.4 Feature Scaling**

One of the most crucial transformations applied to the battery dataset was feature scaling. The dataset contained several numerical features that varied significantly in magnitude and units of measurement. Without scaling or normalization, these discrepancies would likely hinder the performance of most machine learning algorithms, particularly those sensitive to feature magnitude, such as gradient-based methods and distance-based models.

To address this, standardization was employed. This technique transforms the features so that they have a mean of zero and a standard deviation of one. As a result, all numerical variables were brought to a comparable scale, ensuring that no single feature disproportionately influenced the learning process. To implement this, the StandardScaler from the scikit-learn package from Python was utilized. With the help of this package, I was able to standardize the most significant numerical columns, which showed the most range in the data; Molecular\_weight, Capacity\_per\_gram\_in\_mAh, Voltage\_in\_V, Efficiency\_in\_percent. Using this a scaled Dataset was formed, which could be used for any distance based algorithms Additionally, I created a second version using MinMaxScaler also from the scikit-learn package from Python; a normalized version of the data, in which all the columns were scaled from 0 to 1 using the minimum and maximum values. This dataset is to be used for training a neural network for classification of battery material. Both the prime-dataset and the concise-dataset have undergone the feature scaling process as this process applies to all rows and columns.

**3.5 Exploratory Data Analysis**

The ***prime dataset*** consisting of 832 battery material candidates, each labeled as either an “Anode” or a “Cathode”, includes both physical and electrochemical descriptors such as molecular weight, capacity per gram (in mAh), voltage (in V), efficiency (in percent), and energy density (in Wh/kg). The purpose of this analysis is to examine the statistical structure of the dataset, identify patterns and correlations between key features, and establish a foundation for more advanced modeling and hypothesis generation in battery materials research. These attributes represent a broad range of material characteristics, making the dataset suitable for both classification and regression tasks in the context of materials informatics. Importantly, the dataset is complete, with no missing values across any of the eight columns. This suggests that the data was carefully curated and is ready for direct analysis without the need for imputation or data cleaning. The distribution of entries between the two types is relatively balanced, with 433 cathode materials and 399 anode materials. This balance is advantageous, as it enables a fair comparative analysis between the two material classes without introducing biases stemming from class imbalance.

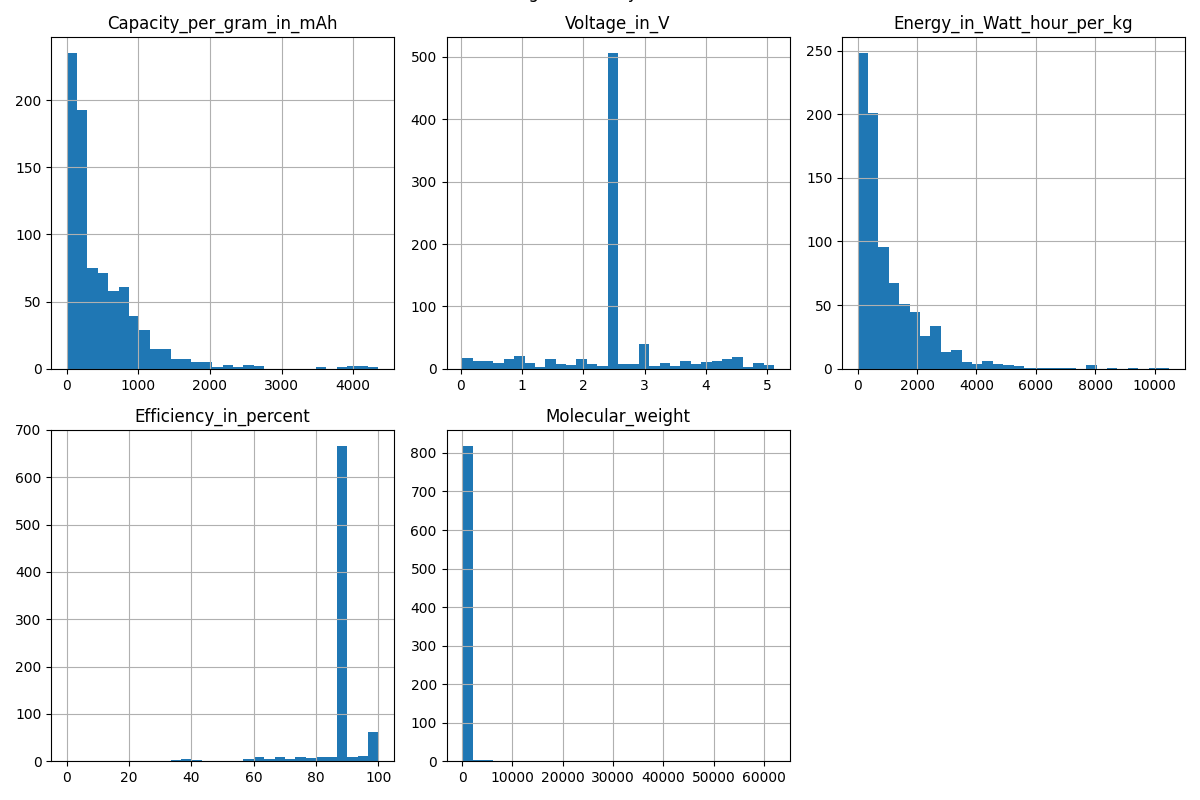
**3.5.1 Summary Statistics**

A preliminary statistical summary provides useful insights into the central tendencies and variability of the key numerical features. The average molecular weight of materials in the dataset is approximately 403.63 g/mol, but the standard deviation is unusually high at over 2400 g/mol. This indicates the presence of outliers or a long-tailed distribution, which is confirmed by the maximum molecular weight reaching over 61,000 g/mol. Most of the materials, however, fall below 500 g/mol, suggesting a heavily skewed distribution.

Capacity per gram exhibits an average value of 493.76 mAh/g with a standard deviation of about 564.24 mAh/g, again indicating a considerable spread. The maximum recorded capacity is an impressive 4352 mAh/g, whereas many materials cluster around the lower end, as shown by a median value of 272.5 mAh/g. Voltage, in contrast, is tightly distributed around a common value of 2.5V, with minimal variance across the dataset. This suggests a design preference or experimental standard in voltage configurations for both anode and cathode materials.

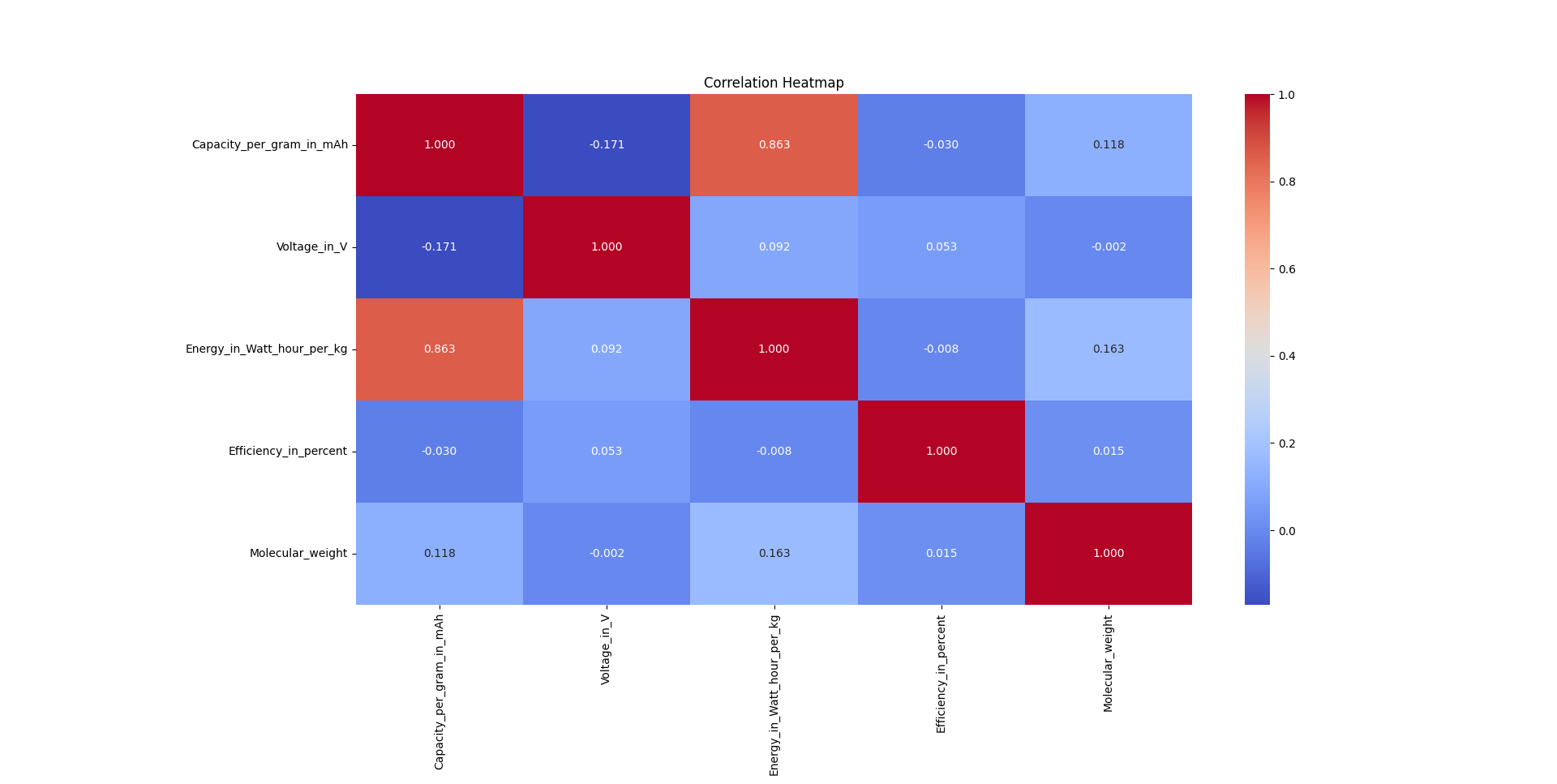
Efficiency is highly concentrated around 90%, with a small spread. The minimum efficiency recorded is near zero, which might reflect failed experimental attempts or highly suboptimal materials, though such entries are rare. Finally, the energy density feature shows a wide range of values, from 0.1 Wh/kg to over 10,000 Wh/kg, with a mean of approximately 1096 Wh/kg. The large standard deviation (1303 Wh/kg) again points to substantial variability across the dataset.

| **Prime-Dataset (#OB=832)** |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **Mean** | **Std Dev** | **Min** | **25%** | **50%** | **75%** | **Max** |
| Molecular\_weight | 403.63 | 2421.86 | 1.76 | 98.85 | 167.28 | 265.83 | 61958.81 |
| Capacity\_per\_gram\_in\_mAh | 493.76 | 564.24 | 0.04 | 135.00 | 272.50 | 706.28 | 4352.00 |
| Voltage\_in\_V | 2.48 | 0.94 | 0.02 | 2.50 | 2.50 | 2.50 | 5.11 |
| Efficiency\_in\_percent | 88.44 | 9.48 | 0.04 | 90.00 | 90.00 | 90.00 | 100.00 |
| Energy\_in\_Watt\_hour\_per\_kg | 1096.10 | 1303.44 | 0.10 | 304.35 | 600.00 | 1470.63 | 10500.00 |

**3.5.2 Univariate Analysis** To explore the distributions of individual features, histograms were generated for capacity, voltage, energy density, efficiency, and molecular weight. The histogram for capacity reveals a distinctly right-skewed distribution. Many materials exhibit capacities under 1000 mAh/g, while a smaller subset achieves extremely high capacities, extending the tail of the distribution.

The voltage histogram is quite unique in that it demonstrates a sharp peak at exactly 2.5V. This dominant mode suggests that a large proportion of materials, regardless of type, have been standardized to this voltage—mainly due to the (mean) imputation during the dataset cleaning process. While there are a few entries with voltages exceeding 3.0V, these are relatively rare.

There are clusters at both low and moderate-to-high energy densities, which may indicate the presence of multiple material families or chemistries. The histogram for Energy density shows that materials that appear in the lower mode represent early-stage or low-performing compounds, while those in the upper mode are more promising candidates for high-performance applications.

Efficiency is tightly clustered, with most values hovering around 90%. This reflects either consistently high performance across materials or potential rounding in reported values. Either way, efficiency offers limited discrimination between compounds due to its narrow variance. Molecular weight, as mentioned earlier, displays a heavily skewed distribution, with most values falling below 500 g/mol and a few extending into the tens of thousands.

**3.5.3 Bivariate Analysis**

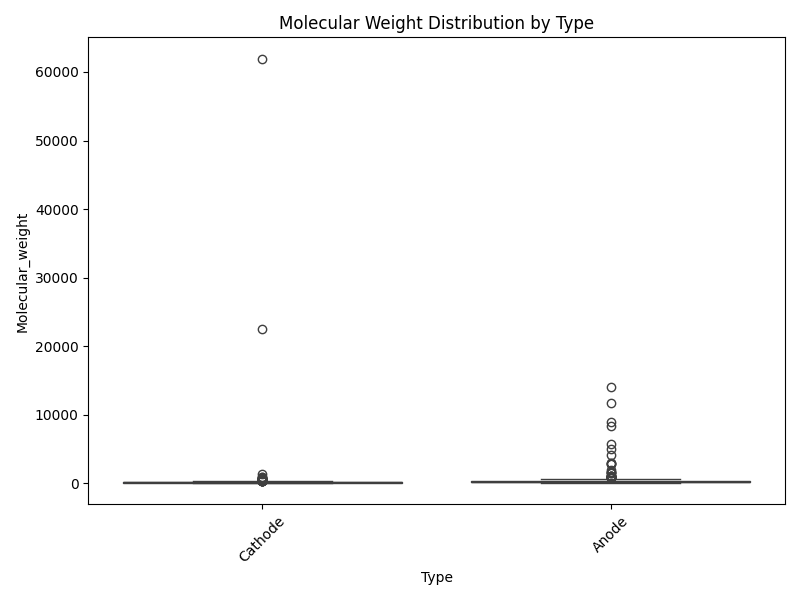
To understand how the different numerical features relate to one another, a correlation heatmap was created.

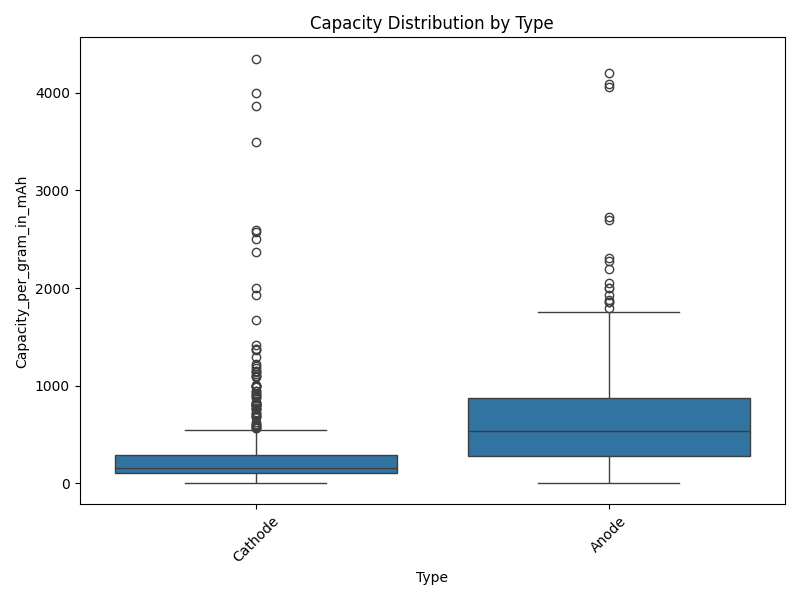
One of the most prominent relationships observed is the strong positive correlation between energy density and capacity, with a correlation coefficient near 0.863. This is expected, as energy is fundamentally a product of capacity and voltage, and in this dataset, capacity appears to be the more variable of the two inputs.

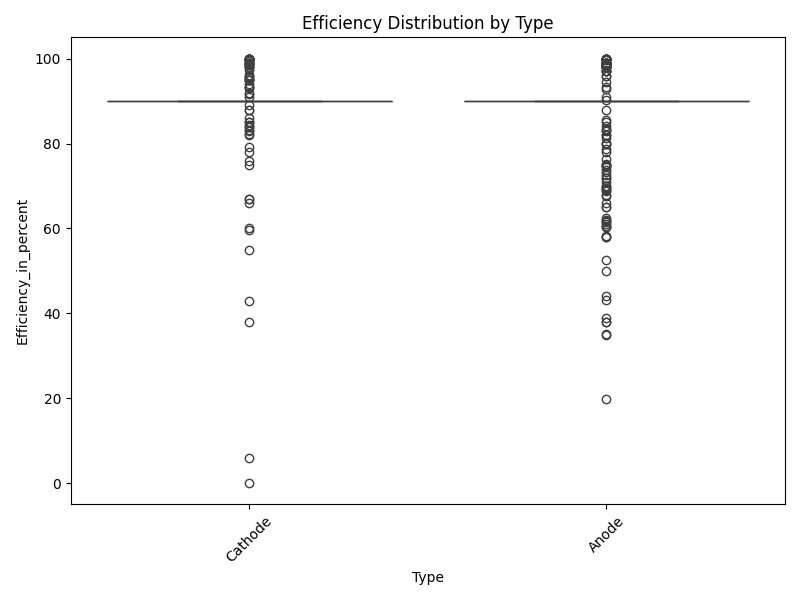
According to the ***prime-dataset***, there is little to no correlation between energy density and voltage (approximately 0.092), suggesting that voltage has very little contribution to the variance of energy density. The relationship between capacity and voltage is weaker still, indicating that high-capacity materials do not necessarily operate at high voltages, and vice versa.

Molecular weight, interestingly, exhibits weak or no correlation with any other feature. This suggests that molecular weight is not a good predictor of a material’s electrochemical properties, at least not in a direct or linear manner. Its strongest correlation (0.163) is with Energy density. This means that even though the effect is low, it does, to some degree, contribute to the variance of the Energy density of a material.

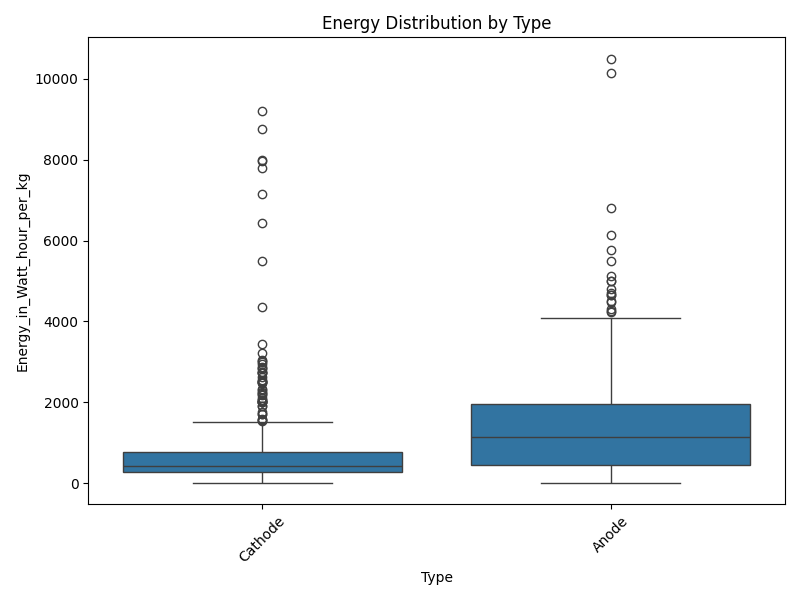
**3.5.4 Comparative Analysis by Material Type**

Boxplots were generated to compare the distributions of key features between anode and cathode materials. One of the clearest distinctions appears in molecular weight distributions. Cathode materials generally exhibit higher molecular weights than anodes, though this trend is somewhat confounded by a few outliers. This might reflect the use of heavier transition-metal compounds or polymeric frameworks in cathode designs.

The difference in capacity is even more pronounced. Cathodes tend to have both higher median capacities and a wider spread, indicating a broader range of performance among cathode candidates. Anodes, in contrast, show a more constrained distribution with lower average capacities. This suggests that the cathode space may offer more room for innovation or variability, whereas anode capacities are more tightly optimized.

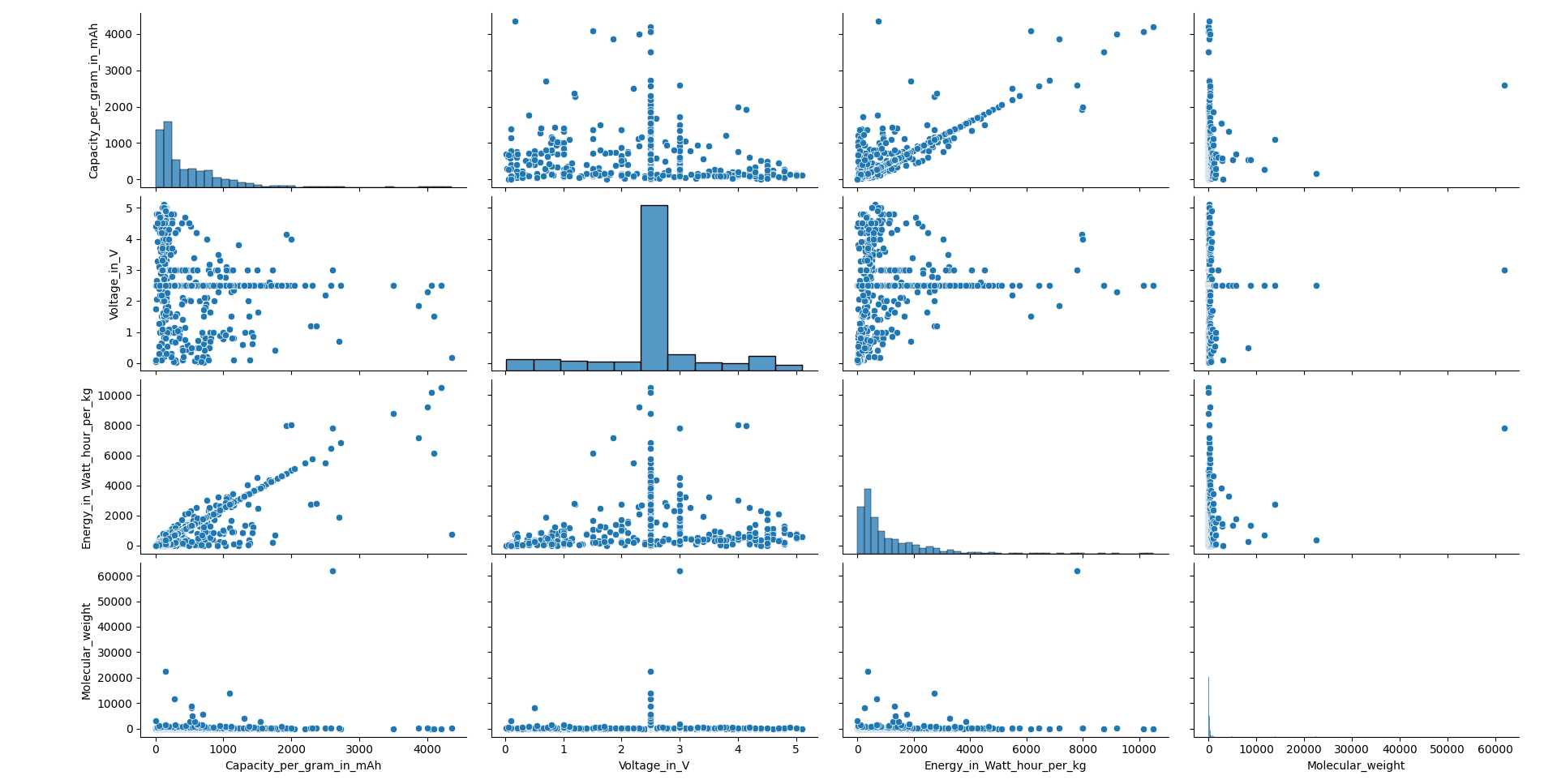
Efficiency values are similar across both types, with most materials clustering at or near 90%. There is no significant distinction in efficiency between anodes and cathodes, implying that this metric may not be particularly useful for classification or early-stage screening.

A diagram of voltage distribution

AI-generated content may be incorrect.When examining energy density, the superiority of cathodes becomes evident. Not only do they achieve higher average energy densities, but they also include several extreme outliers that suggest potential for ultra-high performance. Anodes, by comparison, exhibit a tighter and lower distribution. The voltage distributions follow a similar pattern, with cathodes having a slightly wider and higher voltage range, though both material Types peak around 2.5V.

**3.5.5 Multivariate Analysis**

To explore multivariate relationships and potential patterns, a pair plot of the top numerical features was constructed. This visualizes scatter plots of all pairwise feature combinations, along with their respective distributions. The most striking observation is the linear relationship between capacity and energy density, which is consistent with earlier correlation findings.

****Voltage introduces vertical spread in the energy vs. capacity plots, illustrating how voltage amplifies or attenuates the energy output of a material with a given capacity. However, voltage itself does not appear to provide strong discriminatory power when plotted against other features.

Molecular weight, once again, fails to exhibit strong linear associations with other variables, and its points appear scattered without forming clusters or gradients. No obvious material groupings or clusters are evident in the pair plot, indicating that the dataset may require more sophisticated dimensionality reduction techniques—such as Principal Component Analysis (PCA) or t-distributed Stochastic Neighbor Embedding (t-SNE)—to uncover latent structures.

This exploratory data analysis (EDA), encompassing 832 battery materials categorized as anodes and cathodes, provides significant insights into their electrochemical and structural properties. The dataset's robust curation and balanced distribution ensure the reliability of statistical comparisons. While most features, notably energy density and capacity, exhibit substantial variability, others, such as efficiency and voltage, demonstrate narrower distributions.

Cathode materials consistently surpass anode materials in both capacity and energy density, albeit often possessing higher molecular weights. The voltage distribution implies a degree of standardization across the dataset, potentially reflecting practical design constraints in battery development. A critical finding is the strong positive correlation between capacity and energy, indicating that capacity enhancement is a primary driver for improved energy performance.

Notably, molecular weight appears to have no direct influence on the analyzed performance metrics. This suggests that more intricate chemical features may be necessary for accurate modeling and prediction of battery performance.

This EDA establishes a robust foundation for future research, particularly in applying machine learning techniques to predict material performance or classify compounds. Subsequent work could involve unsupervised learning for cluster discovery or regression models to predict energy density from molecular descriptors. The analysis underscores the inherent complexity and rich structure within battery material datasets.