Part 1: Exploring, processing, gaining insight from the data by Watt The Data

We have a dataset containing information on approximately 100,000 electric car charging sessions. The training data spans from January 2022 to June 2024 and contains sessions using fast chargers by Eveny. The chargers in use have nominal power ratings ranging from 50 DC to 400 DC, providing a diverse set of data points for analysis. The dataset includes following features: IDs, sub IDs, State of Charges (SoC), charging powers, nominal_powers, location IDs and timestamps.

Electric vehicles (EVs) have been gaining traction, and understanding the charging behavior of these vehicles is critical for predicting power consumption, planning infrastructure, and supporting a shift toward renewable energy use.

Overview of the Dataset

Before diving into the predictions, it is essential to understand the structure and distribution of the data. Below are some exploratory data analyses and visualizations that provide insights into various aspects of the charging sessions, from time trends to seasonal power consumption patterns.
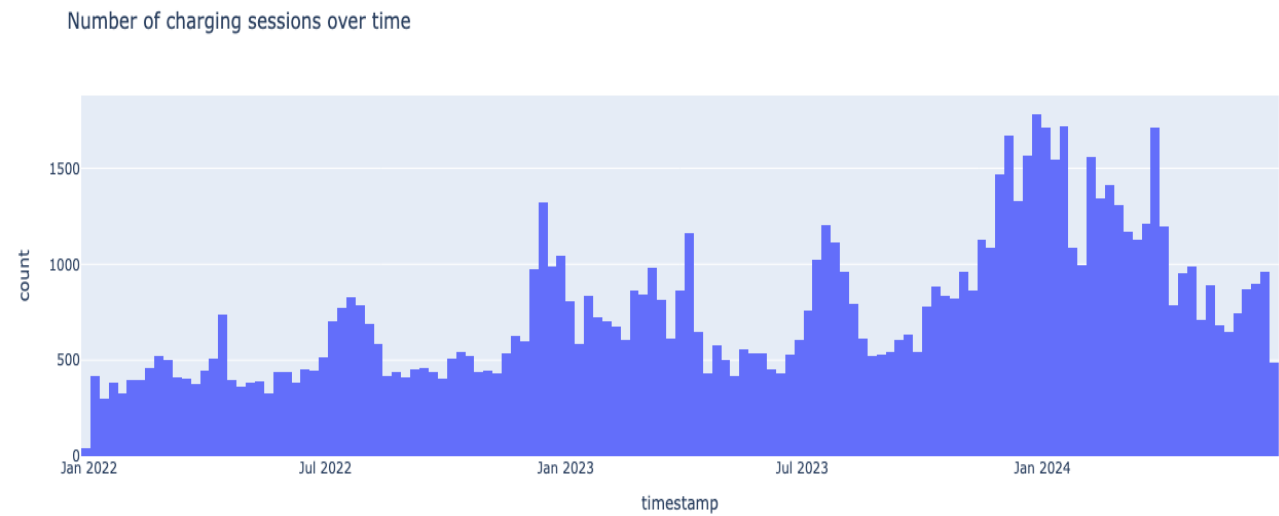

This plot shows the number of charging sessions over the dataset's time span. The general upward trend indicates an increasing adoption of electric vehicles and growing demand for fast charging stations. The data shows how charging activity has evolved, potentially influenced by factors such as increased EV sales, seasonal variations, and even shifts in user behavior.

The pie chart illustrates the distribution of nominal power categories across the dataset. Fast chargers with nominal power ratings ranging from 50 DC to 400 DC are represented. Here we can see that we have a wide distribution of different nominal chargings used.


Distribution of Nominal Power Categories
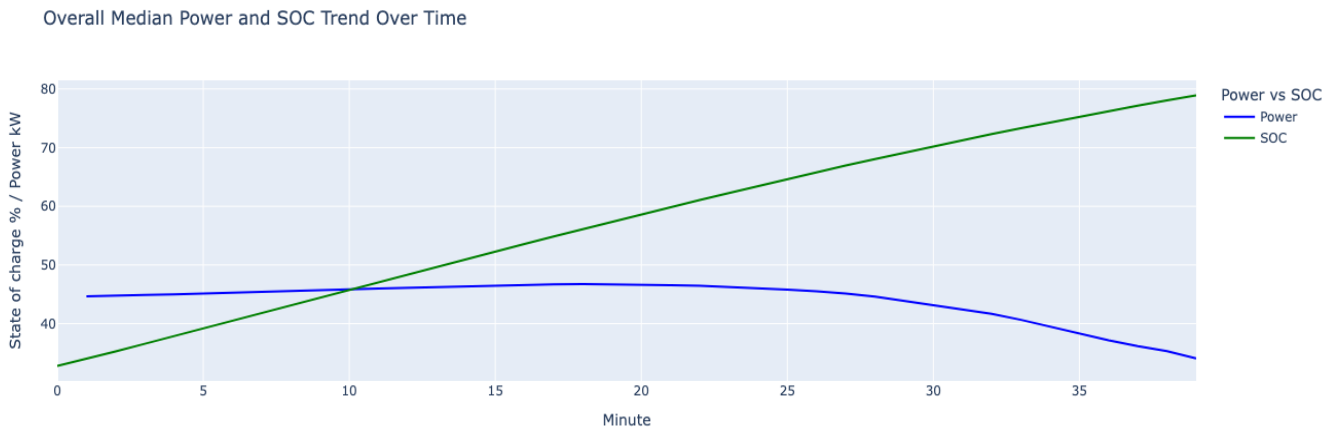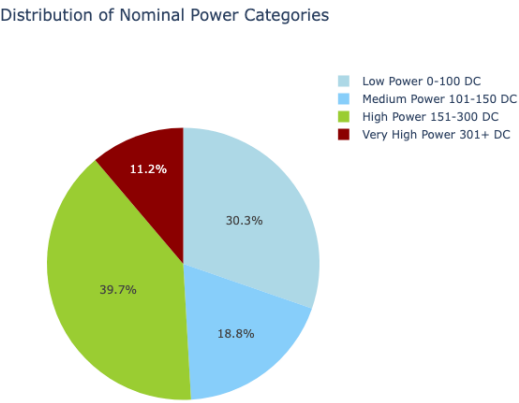
This line plot visualizes the relationship between State of Charge (SoC) and charging power over time. The trend shows that as SoC increases, charging power gradually decreases. This behavior is expected, as fast charging is most effective when a battery is empty, and power is gradually reduced as the battery fills up. A sharper decline in charging power is noticeable as SoC approaches 80%, which is a typical charging curve for many EV batteries.


Median Power for Each Season
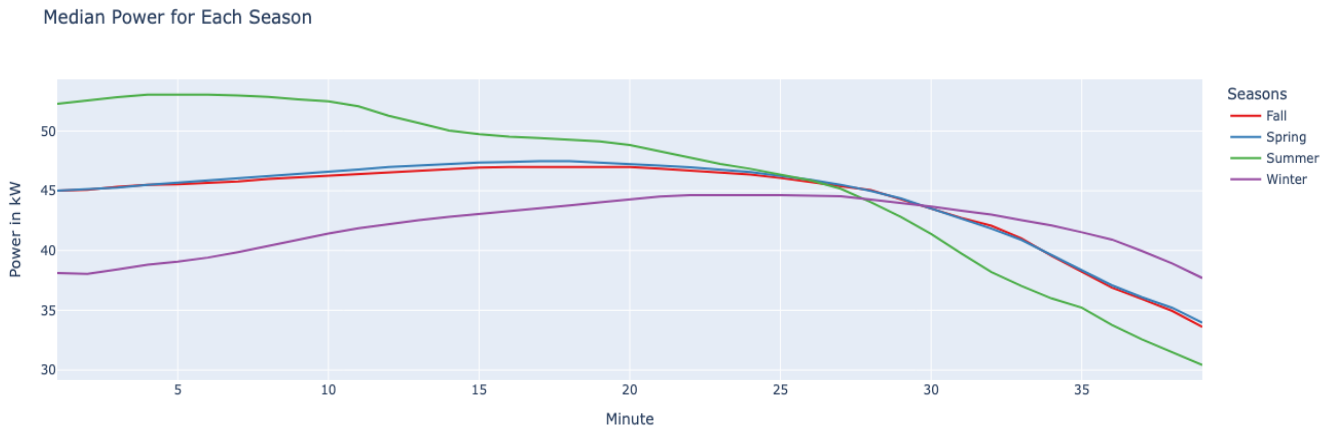
This line plot displays the median charging power at each minute, categorized by season. It shows how the charging behavior varies depending on the time of year. Winter charging starts at the lowest power level and follows a relatively flat curve, possibly due to colder temperatures affecting the efficiency of the

charging process. On the other hand, summer starts at the highest power and decreases more steeply over time. Spring and autumn follow nearly identical curves, suggesting that weather conditions may play a significant role in power consumption patterns during transitional seasons.
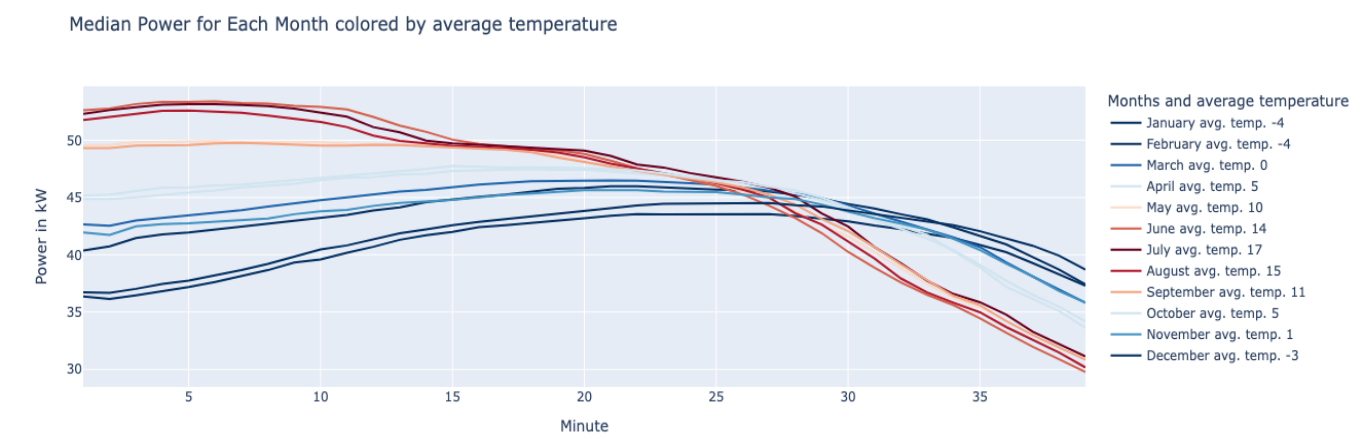
In this line plot, we visualize the average charging power per month, with temperature as a color-coded variable. We see a clear seasonal trend: warmer months start with higher power and gradually decline as the session progresses. In contrast, colder months begin with lower power and increase until reaching a peak, after which the power consumption gradually declines toward the end. This suggests that temperature, both ambient and within the vehicle, can impact the efficiency of the charging process and the overall power consumption.
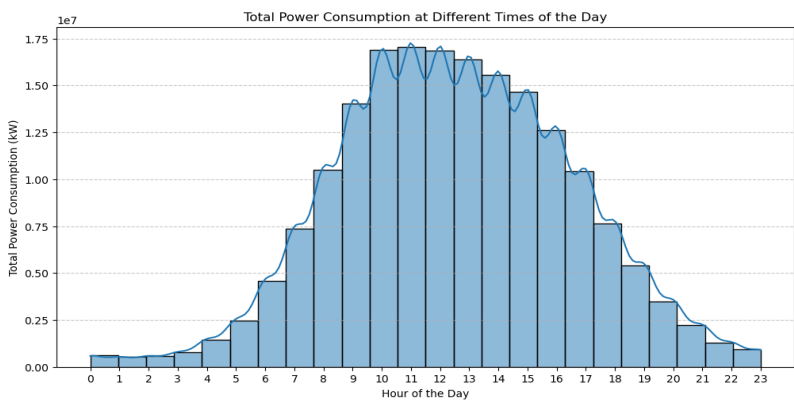

The histogram above shows total power consumption by the hour of the day. The distribution is skewed toward midday, with a peak around 12:00 PM, reflecting typical energy usage patterns during the day. This suggests that charging is concentrated during specific times, likely corresponding to when people are most active or available to charge their vehicles, such as during lunch breaks or after work hours. The overall pattern also seems to resemble a normal distribution, indicating consistent behavior in power consumption across the day.

Inconsistencies

The dataset includes a few inconsistencies, particularly in the State of Charge (SOC) values. The describe function revealed that the maximum SOC value is 104%, which is clearly invalid since SOC cannot exceed 100%. Upon further investigation, we identified and removed the session with this incorrect SOC value. Ensuring that the SOC is within a valid range is critical, as this directly impacts the modeling process and the accuracy of any power consumption predictions.

To examine the distribution of these errors, we looked at power readings that exceeded the nominal power rating of the chargers. While minor deviations (0.1–0.2 kW above the nominal power) were retained, as they are unlikely to significantly impact model performance, we decided to discard any observations where the power exceeded the nominal power threshold. This approach maintains data integrity while ensuring that anomalous readings don't interfere with the analysis.
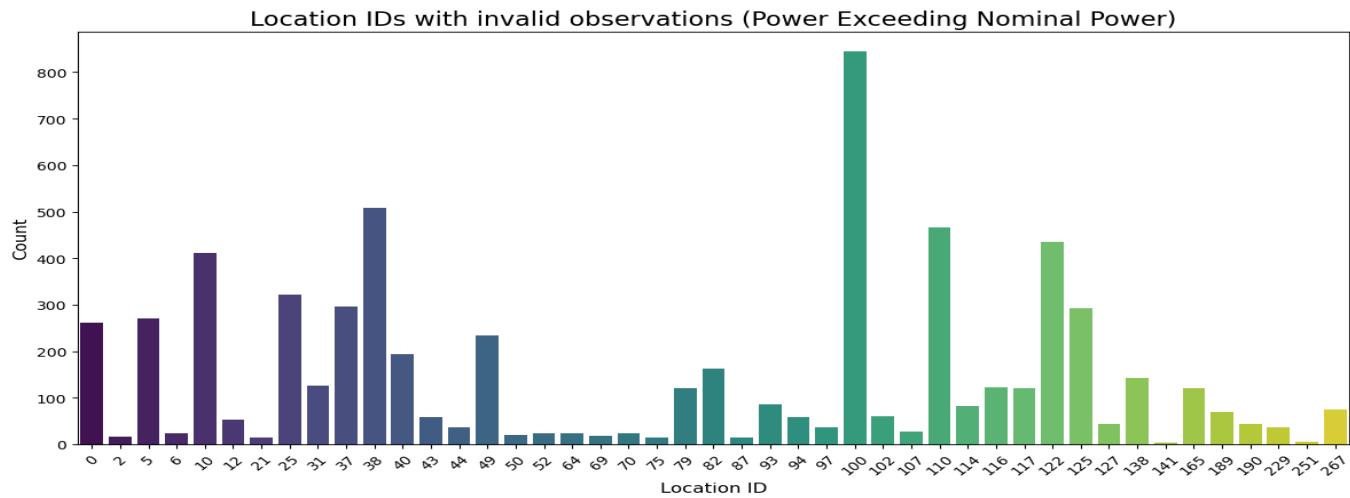
The histogram illustrates the distribution of inconsistencies across all location IDs. This visualization gives insight into the frequency of these anomalies, highlighting whether certain locations have a higher incidence of such issues. These inconsistencies could be indicative of malfunctioning chargers or data collection errors.

Missing values

During the data cleaning process, we identified that State of Charge (SOC) and Power were the only variables with missing values. These gaps are not uncommon in large datasets, and we need to carefully handle them to ensure accurate modeling.
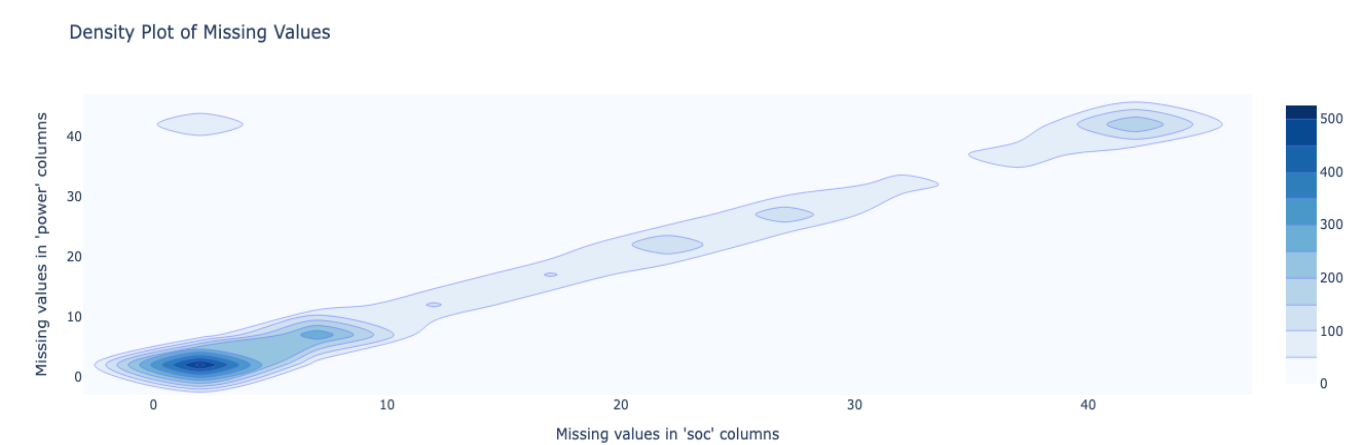
The heatmap illustrate the extent of missing data for both SOC and Power across the charging sessions. To simplify imputation and reduce complexity, we decided to remove sessions with more than a certain percentage of missing values. This strategy ensures that we are working with a clean and reliable dataset, without discarding too many observations that could be useful for prediction.

Data preperation

After cleaning the data and addressing missing values and inconsistencies, we reshaped the dataset so that each row represents a charging session of 40 minutes. This restructuring facilitates an easier modeling process, as it ensures that each session is properly aligned with a specific time window. By having consistent time intervals, we can more accurately capture trends and patterns in power consumption during the charging process.