Team Assignment Project Report

Team 13 - Data Destillateurs

Authors: Jascha Dahlhaus (6047181)
Tobias Derksen (5407583), Tim Hebestreit (7354304)
Denis Smolin (7364694), Evgenia Vdovichenko (7417843)
Supervisor: Univ.-Prof. Dr. Wolfgang Ketter
Co-Supervisor: Janik Muires

Analytics and Applications M.Sc. Course
Department of Information Systems for Sustainable Society
Faculty of Management, Economics and Social Sciences
University of Cologne

February 7, 2024

Executive Summary

Building up on the success of our Data Science project, this executive summary presents key insights and recommendations from our research.
The challenge of operating charging stations lies in the limited capacity of hubs, particularly in busy locations where not all charging stations can function at full capacity simultaneously. Therefore, it is essential to monitor, comprehend, plan, and potentially regulate the charging sessions of electric vehicles.
To optimize the operation at our charging stations we received a data set for two of our sites located near Burbank, from 2018-2021 covering data points like the duration of the charging session or how much kWh were charged.
We could find strong temporal patterns in the behavior of our customers. On average more of our customers visit our sites on weekdays than on weekends and more charging sessions occur in the morning than in the evening.
Moreover, we noticed a difference in user behavior between our operating sites. The first site restricted only to company employees, consistently operates at high capacity with recurring customers. In contrast, our second site, a public charging hub, experiences lower usage and greater fluctuation in utilization.
To better manage the capacity and continuously improve our operation we propose 4 KPIs: overall utilization, total kWh delivered / hour, charging time and parking time.
Moreover, we grouped the charging sessions into 4 groups by their similarity. We defined them as: New Regular Consumers, users charge close to work for long sessions; First and Flexible, early adopters that they stay for extended periods; Charging Professionals, rely on charging stations near their workplaces; First Fast Public Users, occasional daily users. We concluded our research by building a predictive model to forecast the hourly utilization rate with a complex and a simple model. We couldn't find a big difference in the predictive power of the two models. Due to explainability and computational burden we decided to propose usage of linear regression.
As a decisive point, we would like to recommend the following ideas:

- use our predictive model to dynamically adjusted prices by demand
- add additional incentives to free up charging stations when charging session are over

Lastly, we should monitor shifts in behavior, like working from home, caused by the effects of the covid-19 pandemic.

1 Problem Description

Our company operates charging stations for electric vehicles and generates its revenue from the sale of electricity from the charging stations to vehicle owners. Therefore, the company's aim is to maximize the number of sold kilowatt hours at the sites. A significant challenge for the company lies in the limited number of available charging stations for customers, considering the excessive costs and investments required to build new stations. Inefficiencies arise when a charging station is blocked by a vehicle not actively charging, or when malfunctioning stations reduce our total charging capacity. It is therefore important to minimize inefficiencies by carefully planning, monitoring, and controlling the use of the existing charging infrastructure. The most important metric from a business perspective is the site revenue, as this is directly reflected in the company's profit and loss. An alternative metric could be the utilization rate of the charging station and the kWh delivered as both metrics are closely tied to the station's revenue.

From a data science perspective, we need to use our domain expertise, programming skills, and knowledge of statistical methods to achieve the best results. Our first step is to understand the data and get an overview of all available features and if temporal, spatial or behavioral patterns can be spotted. We can expect that errors in the data exist and thus in the next step we should clean the data from outliers, build new features from the given data and enhance our data set with other data sets. If we understand the data in enough detail, we can start the unsupervised clustering. We can evaluate the result using such metrics as silhouette score, Davies-Bouldin index, and Calinski-Harabasz index. Also, we can analyze the feature distribution in the cluster if we observe clear patterns in the users behavior. In the last step we are going to build a predictive model to forecast the utilization of our stations. It is important for us to use cross validation to ensure a robust evaluation and validation and create a model that generalizes well on unseen data. Also, we should try to find a balance between performance and model complexity for the end user.

2 Data Description

We have been provided with data from two electric vehicle charging sites where one is restricted to employees only, and the other is publicly accessible. Additionally, we have a second dataset obtained from a nearby airport weather station. Given that the weather data provides us with the name Burbank we can expect the location to be in the area surrounding Los Angeles in California.

The first dataset, containing the data of the charging sessions, has a total of 66.450 entries while the second data set, containing weather data, has a total of 29.244 entries but did not cover the whole time period and ends on the 01.01.2021.

How we further handled the data is described in the next section.

3 Brief Data Preparation Details

Raw data must always be preprocessed before being used for analysis and prediction. We removed duplicate entries in the first place. After that we checked all values on feasibility and found out that there was a Space with ID '11900388' that had erroneous data. Empty values were filled with meaningful data. For example, where there was a guest user, we assigned the ID 0 and an empty list of user inputs. The data contains three timestamps which tell us when the user plugged in the vehicle, when it was done charging, and when the vehicle was plugged out. With this at hand, we found data that had curious entries where the vehicle was done charging before even being plugged in or done charging after being already plugged out. Such values were all located and removed or adjusted. There is missing data from August to November 2020 and we are not able to replace it.

The weather data was cleaned in an equivalent way. Here the observation was that almost nine months for 2021 of weather data was missing. We circumvented that problem by downloading, preprocessing, cleaning, and stacking up weather data from https://www.weather.gov/. With two prepared datasets, it took us one single line of code to join the weather data on the charging session data to retrieve a full ready-to-use dataset with all available data. In total our cleaned and merged data set has 63.883 rows where the features 'id' and 'sessionId' are distinct for every entry and can be used as unique identifiers for every entry as further described in Figure 1.

Further we unpacked the data in the column 'userInputs' that was in a JSON format and lead us to seven new features 'WhPerMile', 'kWhRequested', 'milesRequested', 'minutesAvailable', 'modifiedAt', 'paymentRequired', 'requestedDeparture', and a duplicate feature 'userID' that matched the feature with the same name already existing in the data set. Of our 63.883 rows 38.837 had user inputs connected to them.

For the KPI task we created the new features 'Hour', 'Weekday', 'Month', and 'Year' from the feature 'connectionTime'. Further by subtracting the 'connectionTime' from the 'disconnectTime' we could derive the duration of the session. By dividing the feature 'kwhDelivered' by the 'duration' we could get the average kWh charged per Hour of the session.

We encoded categorical feature with OrdinalEncoder, after we scaled features

with StandardScaler to normalize the data for optimal algorithm performance.

4 Data Analytics

As expected, we find strong temporal patterns in the data, as shown in Figure 2. Firstly, most charging sessions start in the morning, with a subsequent drop until midday and a small second peak. Secondly, there is a higher number of charging sessions on workdays compared to weekends.

Also looking at the number of charging sessions and mWh delivered to customers provided in Figure 3 we see a big structural break around March 2020 with a sharp drop off in both metrics and a slow recovery until the end of our observation period in 2021. This can be explained with the COVID-19 lockdowns and restrictions. This big shift in user behavior makes it difficult to sample meaningful results from a longer time horizon like months and quarters as we only have very few observations on this level.

Figure 4 and 5 allow us to compare site 1 and 2 and we find that site 1 has a bigger difference between the number of charging sessions on weekdays and weekends. Also, the charging sessions on site 1 are on average longer than on site 2. This shows us the differences between the private (site 1) and the public (site 2) and the resulting difference in the user behavior. Building on these observations we can also find that the users of site 1 are using the charging sites more often than the users of site 2 as shown in Figure 6.

To manage the charging park efficiently, we recommend four Key Performance Indicators (KPIs): **overall utilization**, **total kWh delivered** / hour, **charging** and **parking** time. Starting with overall utilization we see the general potential to grow (utilization is mostly less than 30%) without adding more stations (Figure 7). The hub owner should also have a closer look on site 2, the number of never connected stations is increasing (Figure 8). The other KPIs are indicators of how the revenue is generated with a closer look on parking time (without charging). If in the future the utilization would be remarkably high, increasing parking time costs could be interesting. To help the hub operator, we created interactive widgets and added even more KPIs to the notebook (Figure 9 and 10).

In our clustering analysis, we refined the dataset through feature engineering, focusing on temporal aspects and key metrics like 'Final Plugged Time', 'Charging Time', 'Parking time' and excluded irrelevant features to improve clustering. Outliers identified using the IQR method were removed to enhance accuracy. We applied PCA for dimensionality reduction, selecting ten components to maintain 95% data variance (Figure 11).

For K-Means we evaluated clustering on datasets with and without outliers,

determining 2 to 4 clusters and the dataset without outliers as optimal based on silhouette scores (0.4, 0.13, 0.21, respectively) and visualized data distributions in the clusters using various plots (Figure 12). Agglomerative clustering was tested for its ability to identify diverse clusters, with 5 and 6 clusters showing a silhouette score of 0.12 (Figure 13).

Comparing K-means (4 clusters) and Agglomerative clustering (5 clusters) through silhouette score, Davies-Bouldin index, and Calinski-Harabasz index, K-Means demonstrated superior performance across all metrics, suggesting better cluster definition and separation.

Best Clustering Result: Cluster 0: New Regular Consumers (users charge close to work for long sessions, to fully charge their batteries before going home); **Cluster 1**: First and Flexible (early adopters of the service, they prefer to fully charge, at stations close to their work, homes, or places where they stay for extended periods); **Cluster 2**: Charging Professionals (professionals who rely on charging stations near their workplaces); **Cluster 3**: First Fast Public Users (occasional daily users with diverse charging needs and routines).

We can apply two approaches, a neural network and a linear regression model, to create a prediction model that forecasts the future hourly electricity demanded. We assume that, in terms of business logic, utilization is how much electricity will be consumed is meaningful to the operator. We create in total four models, two for each site. Data is accumulated to get hourly values. We encoded categorical features to numerical values and for data such as wind speed, temperature, pressure, and precipitation average values are calculated. For weather we take the most occurred weather state and set it for the hour in question. All data is normalized down to values between 0 and 1, which is crucial to avoid divergence of models. We start off with a neural network model that takes in seven features (month, day, hour, and all weather data) and yields a prediction value of how much kWh will be demanded for the selected hour. The model is a minimalistic linear neural network with six hidden layers. We create test datasets for both sites, it is 10% of all data. The rest is split in a 4:1 ratio (80/20) for training and validation. The model runs 50 epochs, which means that it learns and evaluates itself 50 times. After this, the model loss is visualized, and test set performance is gained. This is done in a comparable way with the regression model. We apply a linear regression model with the same features and observe that the performance is almost identical. The performance of all models for both sites can be found in Figure 14. Cross-validation of models shows almost no significant difference in terms of performance. For the linear regression model we can see the coefficients in Figure 15. With these preconditions at hand, one may argue that neural networks are able to capture more complexity and are able to perform

better, especially in case of classification. In our use case we must predict electricity demand. The explainability and computational advantage of a regression model overrule, due to the similar performance. A neural network consumes more runtime to be trained and the model explainability is exceptionally low.

5 Conclusions

Our thorough analysis gives us a good understanding of the operations of the charging stations and the predictability of patterns. A limitation is given by the data, as there is a big structural change in the data induced by the exogenous COVID-19 shock in 2020. We managed to assign the charging sessions into 4 distinct groups that can be explained easily and showed that an easy linear model can compete with a neural network in certain tasks. Another drawback was the missing data from August to November 2020 reducing the number of observations to draw conclusions from.

Based on our results we propose the following measures. A better control system to verify that all stations are functioning and able to operate needs to implement, as we see in our KPIs that after 2020 a lot of stations on site 2 were not used anymore. Also, the behavioral shift towards more home office and less traveling could pose a risk to our operation on site 1 with less employees coming regularly to visit the site making it possible to allow outside users.

Furthermore, our KPIs show that a lot of time is spent parking on the sites while not charging. The company should think about incentivizing customers to free a charging site, when they are not charging anymore, by lowering the fees for shorter charging cycles or penalizing long standing. This could be tested cheaply by varying prices in the application and would not require costly investments in infrastructure.

Given the success of our prediction model, revenue could be leveraged by adopting dynamic pricing like classical gas stations do. When there is higher demand by electric vehicle owners, the prices can be slightly increased and vice versa. Also, a subscription model could increase the amount of user data which would allow us to gather better insights about user behavior.

This idea is reinforced by our clusters identified for the charging sessions. Different user groups with varying demands allow more flexibility in the pricing or to add services on top.

The variations between site 1 and 2 suggest structural differences in operating private and public charging hubs. As a result, they may require distinct KPIs and management strategies.

A Appendix

field	type	description
id	string	Unique identifier of the session record
connectionTime	datetime[a]	Time when the EV plugged in.
disconnectTime	datetime[a]	Time when the EV unplugged.
doneChargingTime	datetime[a]	Time when of the last non-zero current draw recorded.
kWhDelivered	float	Amount of energy delivered during the session.
sessionID	string	Unique identifier for the session.
siteID	string	Unique identifier for the site.
spaceID	string	Unique identifier of the parking space.
stationID	string	Unique identifier of the EVSE.
timezone	string	Timezone of the site. Based on pytz format.
userID	string	Unique identifier of the user, if provided.
userInputs	list	Inputs provided by the user. Since inputs can be changed over time, there can be multiple user input objects in the list.
WhPerMile[b]	float	Efficiency of the EV in Wh per mile.
kWhRequested[b]	float	Energy requested by the user in kWh.
milesRequested[b]	float	Number of miles requested by the user.
minutesAvailable[b]	float	Length of the session as estimated by the user.
modifiedAt[b]	datetime[a]	Time this user input was provided.
paymentRequired[b]	bool	If the user was required to pay for this session.
requestedDeparture[b]	datetime[a]	User estimated departure time.

[a] All datetimes are in UTC (GMT) see timezone field for the correct timezone of the site.
[b] Fields are optional, if user participates in app-based charging.

Table 1. Description of Fields of Charging Session Dataset

Figure 1: Cleaned data available for analysis

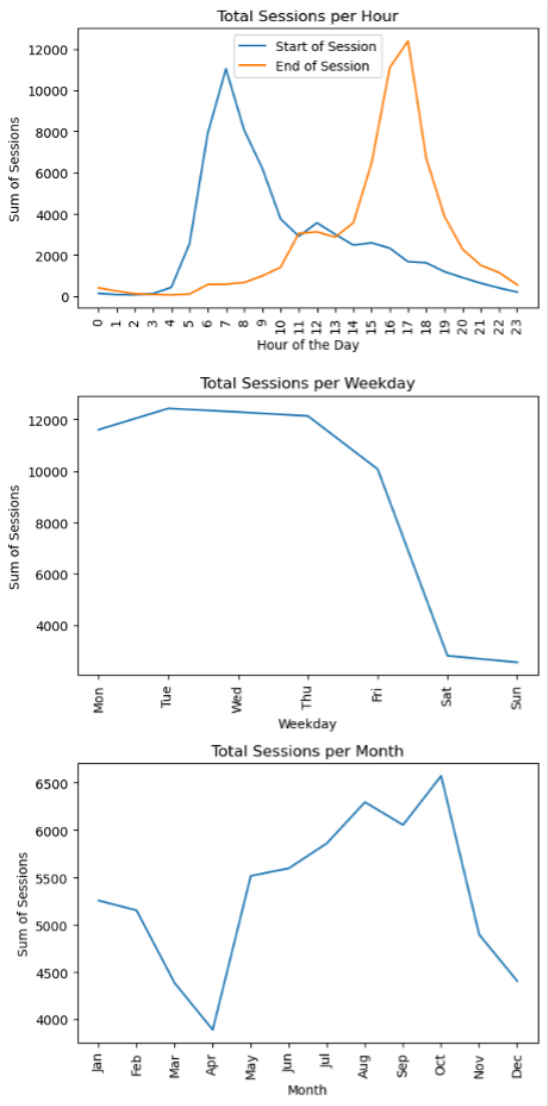

Figure 2: Daily, weekly, and monthly patterns by the number of charging sessions

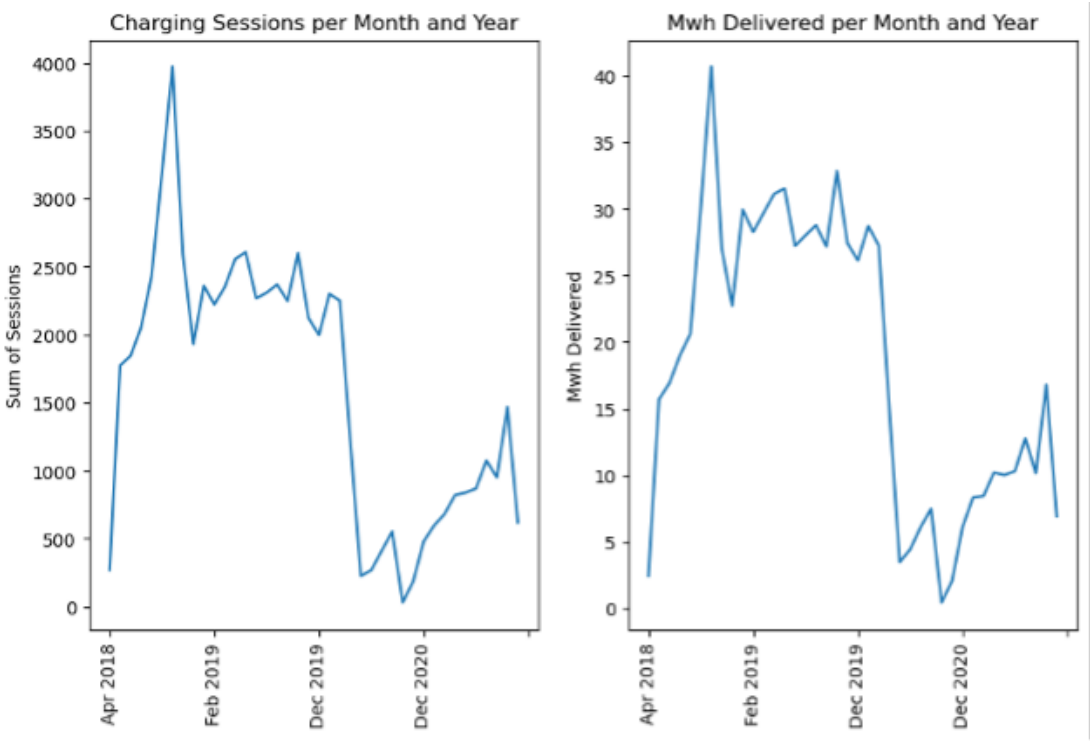

Figure 3: Monthly number of sessions and mWh delivered by the charging stations

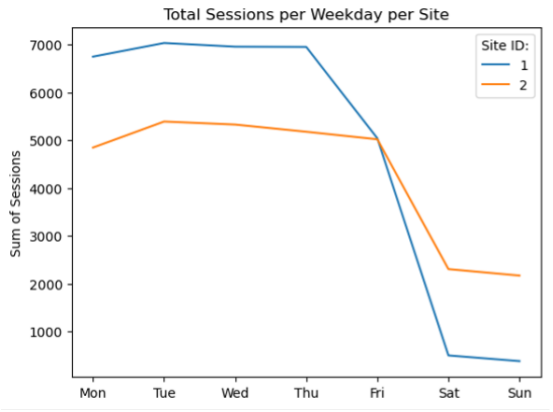

Figure 4: Charging Sessions by weekday for the different sites

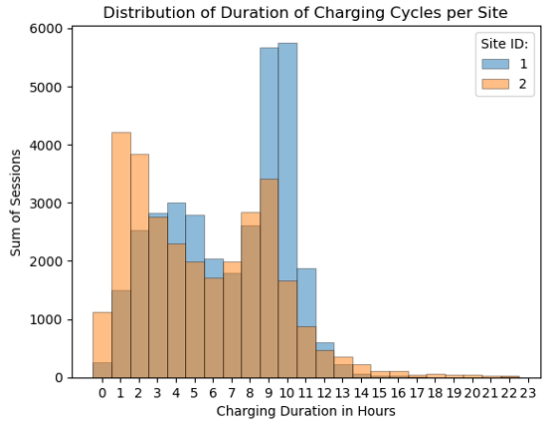

Figure 5: Duration of the charging sessions for the different sites

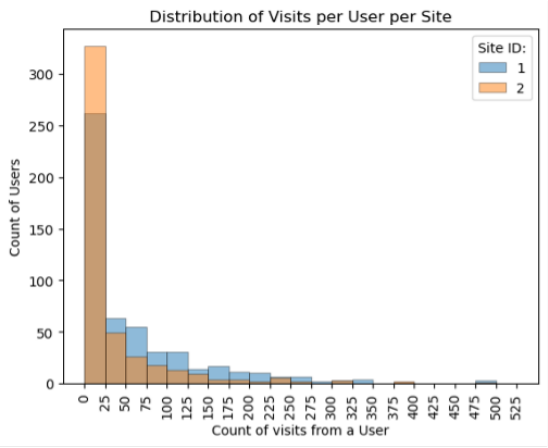

Figure 6: Number of recurring visits from the users for the different sites

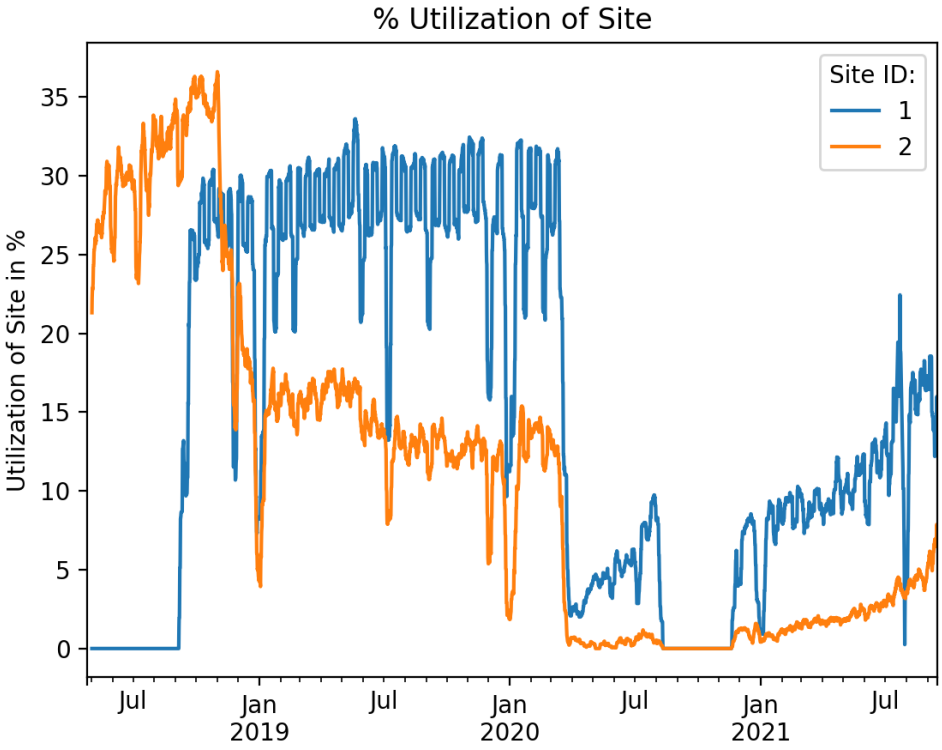

Figure 7: Utilization of Site

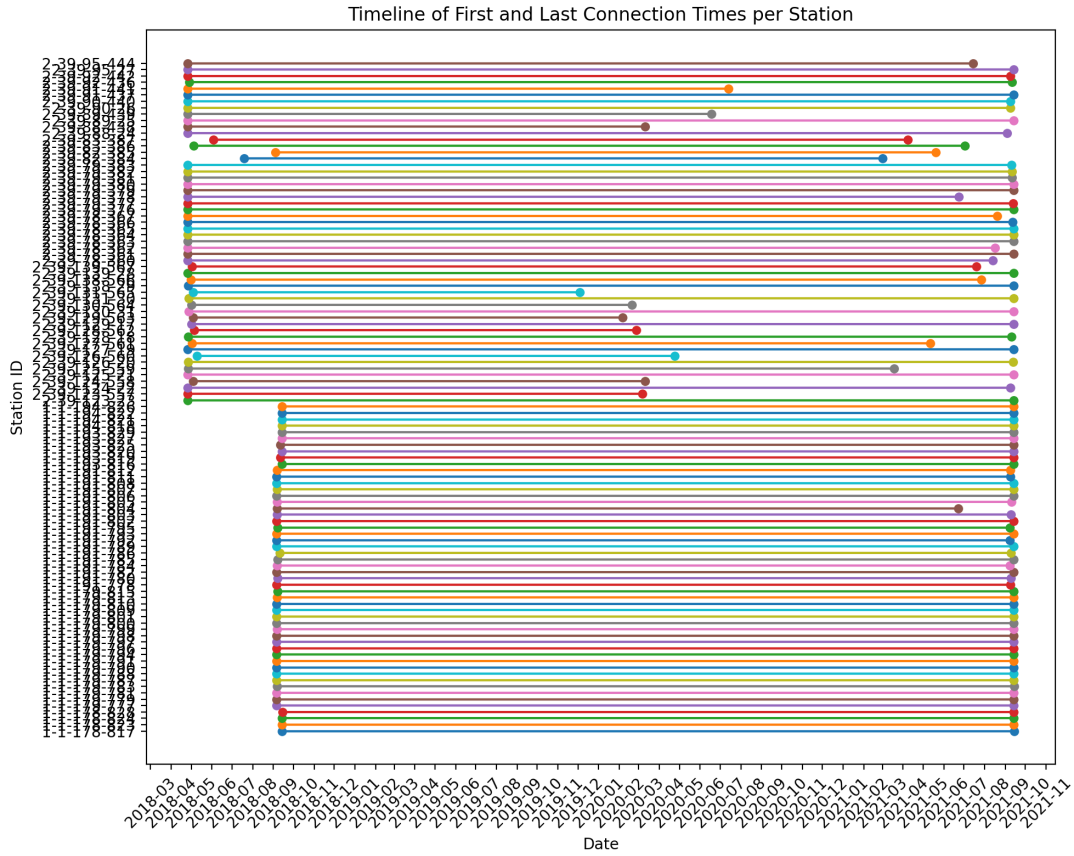

Figure 8: Timeline of first and last connection time per station

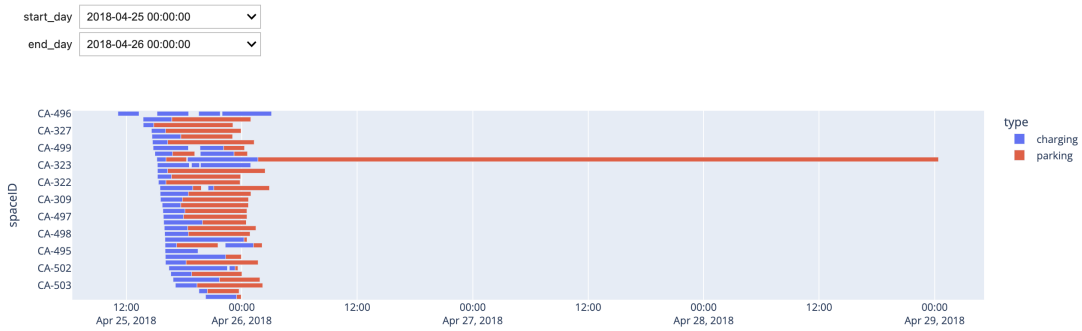

Figure 9: Interactive widget to display charging and parking time

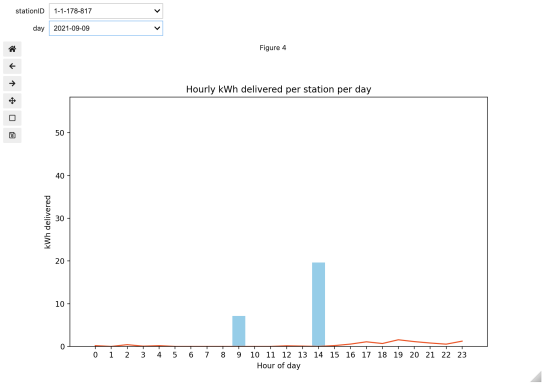

Figure 10: Interactive widget to display hourly kWh delivered per station per day

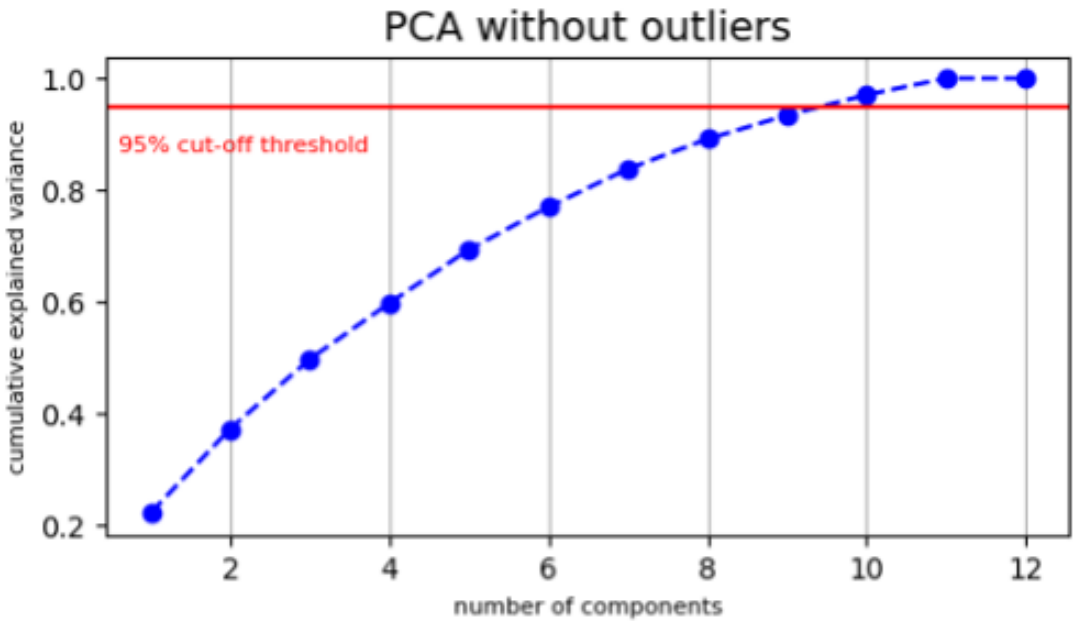

Figure 11: Checking number of components to reduce the data diminution

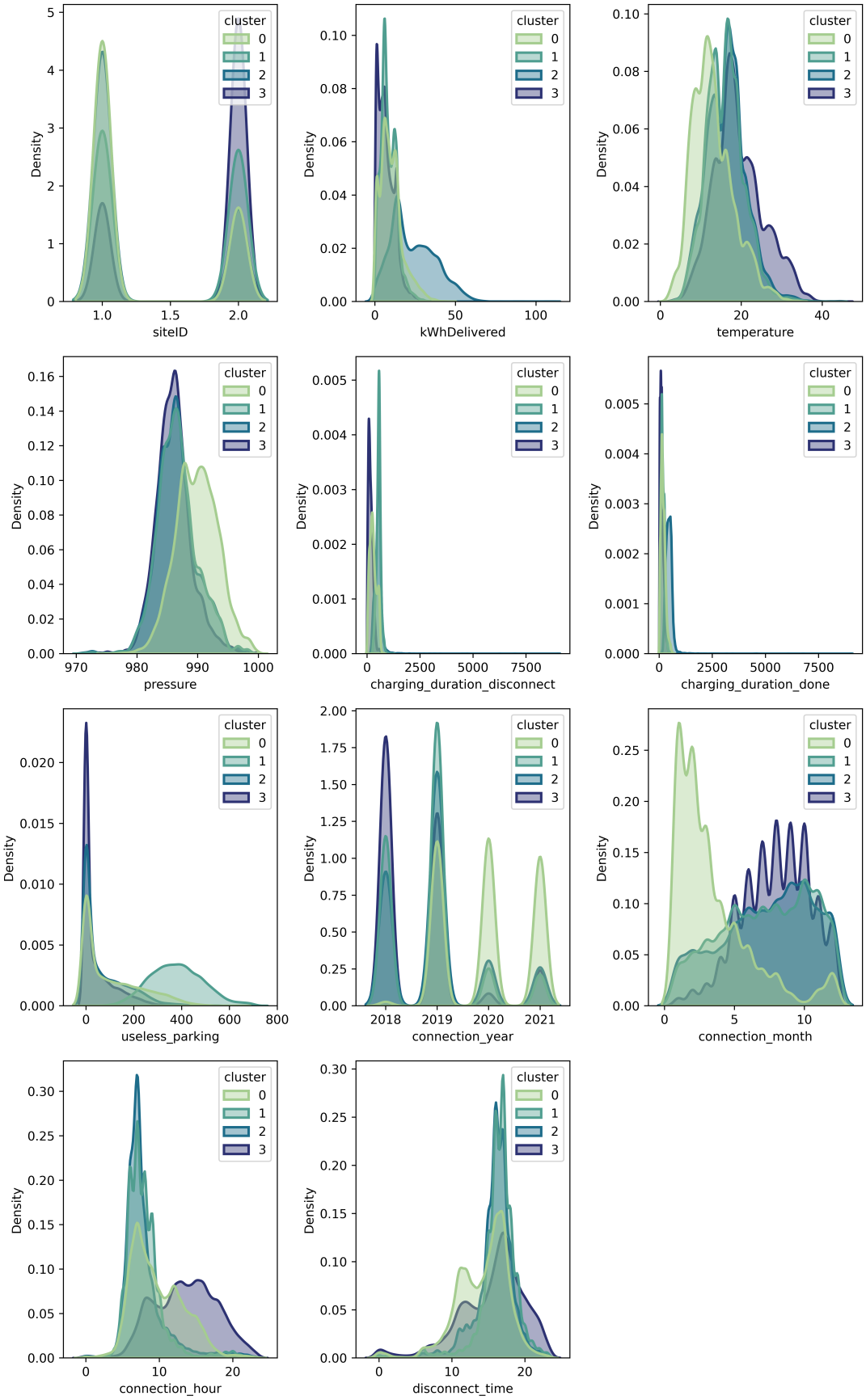

Figure 12: Visualization for numerical features in the clusters

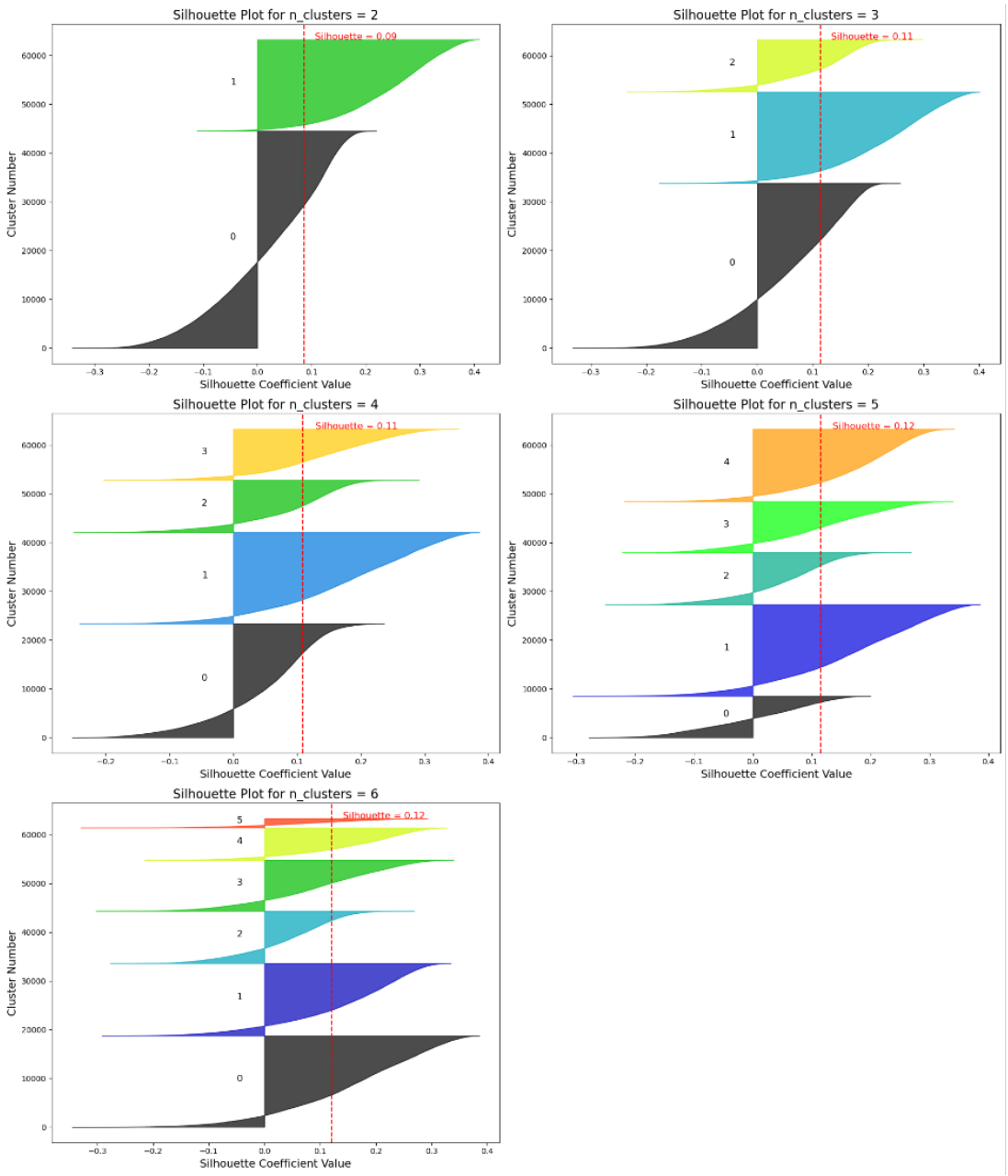

Figure 13: Silhouette score for Agglomerative Clustering

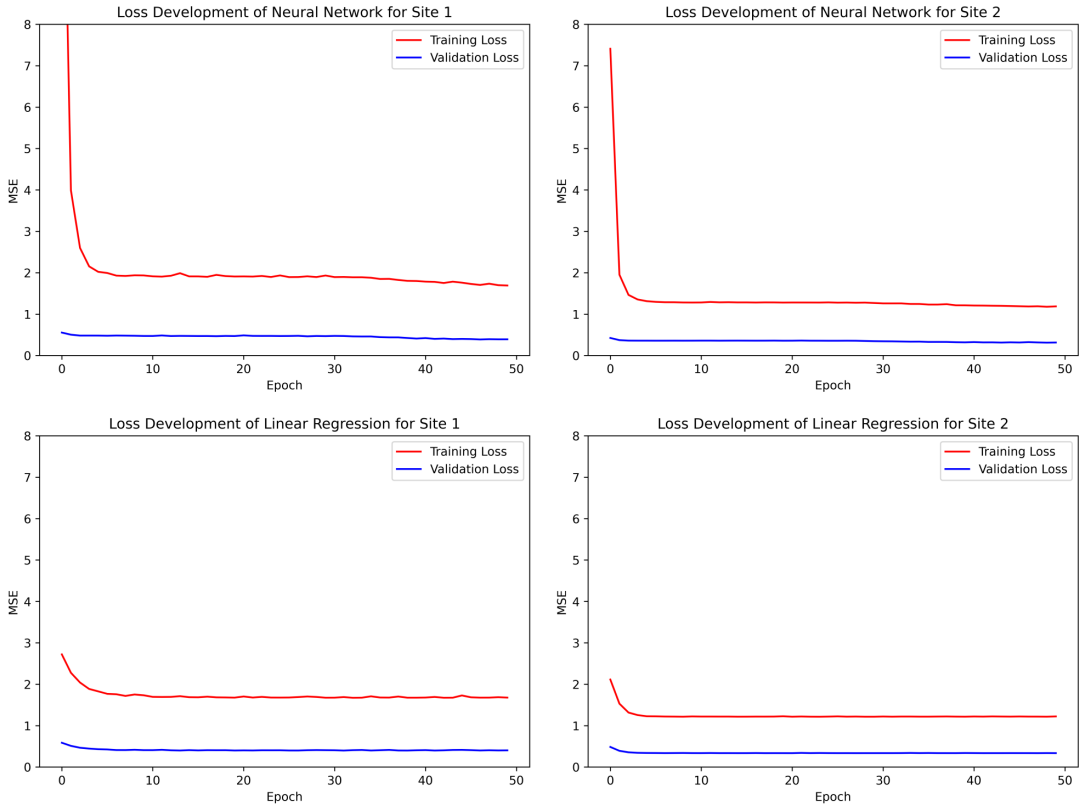

Figure 14: Loss development for prediction models

```
feature_names = df_site_1.columns
weights = model_reg_1.linear.weight.data[0].cpu().numpy()

for feature, weight in zip(feature_names, weights):
    print(f'{feature}: {weight}')
```

```
connection_month: 0.0052874451503157616
connection_day: -0.003787705209106207
connection_hour: -0.24217194318771362
temperature: -0.014410912990570068
pressure: -0.004087554290890694
weather: 0.045620452612638474
windspeed: 0.036312393844127655
precipitation: -0.06425146758556366
```

Site 2:

```
feature_names = df_site_2.columns
weights = model_reg_2.linear.weight.data[0].cpu().numpy()

for feature, weight in zip(feature_names, weights):
    print(f'{feature}: {weight}')
```

```
connection_month: 0.023451903834939003
connection_day: 0.004659086000174284
connection_hour: -0.08461172878742218
temperature: -0.007080324459820986
pressure: -0.04800762981176376
weather: -0.0018498190911486745
windspeed: -0.044084955006837845
precipitation: -0.018877575173974037
```

Figure 15: Weights of the linear regression