TU DS Challenge

Milestone 2 (Data Importance)

- Dimensions: (34722, 26) => 34722 rows, 26 columns
- 3 Features (Provider, Location, Charger)

```
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Betreiber                34722 non-null  object
 1   Straße                   34722 non-null  object
 2   Hausnummer               34722 non-null  object
 3   Adresszusatz             4847 non-null   object
 4   Postleitzahl             34722 non-null  int64
 5   Ort                      34722 non-null  object
 6   Bundesland               34722 non-null  object
 7   Kreis/kreisfreie Stadt   34722 non-null  object
 8   Breitengrad              34722 non-null  object
 9   Längengrad               34722 non-null  object
 10  Inbetriebnahmedatum      34722 non-null  object
 11  Anschlussleistung        34722 non-null  object
 12  Normalladeeinrichtung    34722 non-null  object
 13  Anzahl Ladepunkte        34722 non-null  int64
 14  Steckertypen1            34722 non-null  object
 15  P1 [kW]                  34722 non-null  object
 16  Public Key1              3402 non-null   object
 17  Steckertypen2            29072 non-null  object
 18  P2 [kW]                  29069 non-null  object
 19  Public Key2              2845 non-null   object
 20  Steckertypen3            1825 non-null   object
 21  P3 [kW]                  1844 non-null   object
 22  Public Key3              166 non-null    object
 23  Steckertypen4            993 non-null    object
 24  P4 [kW]                  993 non-null    object
 25  Public Key4              122 non-null    object
dtypes: int64(2), object(24)
```

Milestone 2 (Data Important) - Codebase

+ Code + Markdown

```python
# --------------
# Import the Data
# --------------
import os
import pandas as pd
```

Python

```python
# Get the current working directory & define tha data path
print(os.getcwd())
path = str(os.getcwd())+'/data/data.csv'
```

Python

```python
# Define the data frame
df = pd.read_csv(path, sep=';', encoding='ISO-8859-1', header=10)
print(df.info())
```

Python

```python
# Log the shape (rows, columns) of the df [would work with: print(len(df.columns)) as well]
print(df.shape)
```

Python

```python
# Log the column types
print(df.dtypes)
```

Python

Milestone 3 (Data Preprocessing)

Problems in the Dataset:

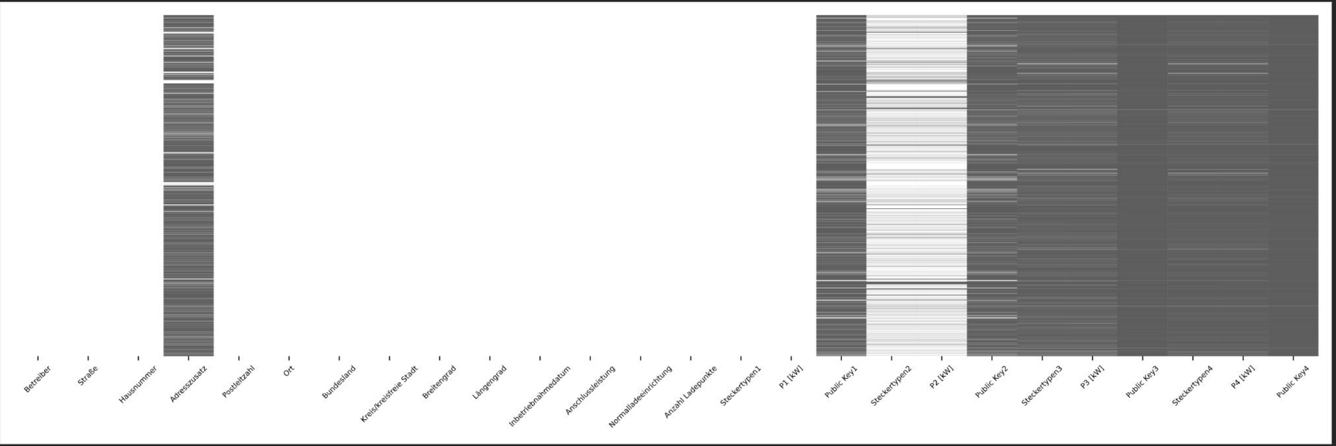
- dtypes are mixed up (resulting in object)
- Hausnummer has 0
- many NaN entries (null values)
- longitudes etc -> proper floats
- Public key columns?

Data Preprocessing:

- Adjust missing values
- Data scaling with techniques such as Label encoding or one hot encoding
- Outlier detection
- Correlation of features (heatmap)

Milestone 3 - Pre Processing Report (1/2)

Missing Values:



	index	0
0	Betreiber	0
1	Straße	0
2	Hausnum...	0
3	Adresszus...	29875
4	Postleitzahl	0
5	Ort	0
6	Bundesland	0
7	Kreis/kreis...	0
8	Breitengrad	0
9	Längengrad	0
10	Inbetriebn...	0
11	Anschlussl...	0
12	Normallad...	0
13	Anzahl La...	0
14	Steckertyp...	0
15	P1 [kW]	0
16	Public Key1	31320
17	Steckertyp...	5650
18	P2 [kW]	5653
19	Public Key2	31877
20	Steckertyp...	32897
21	P3 [kW]	32878
22	Public Key3	34556
23	Steckertyp...	33729
24	P4 [kW]	33729
25	Public Key4	34600

Milestone 3 - Pre Processing Report (2/2)

Prevalence of missing values:

- The following columns have only a few entries: Adresszusatz, 2+ Outlet details

Incorrect data types:

- Nearly all data got set as objects in the first place
- A parser can fix this problem
- Remaining problems are:
  - Hausnummer has string values due to values like “10C” or “13 - 15” 0> add regex check to remove everything that is not a number (stop as soon as we hit a character or a whitespace)
  -

Milestone 4: EDA (Exploratory Data Analysis)

Summary statistics for the variables/features:

```
[69] df.describe()
```

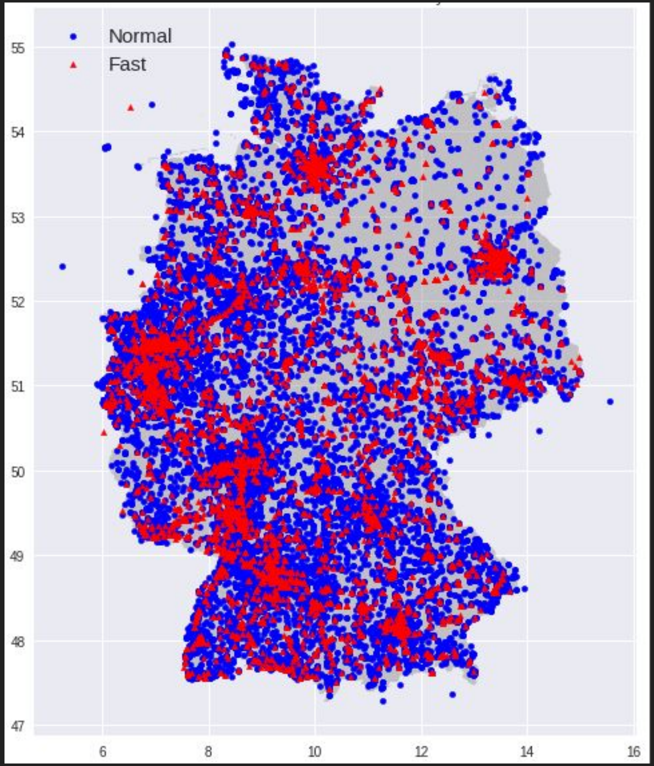
	Postleitzahl	Breitengrad	Längengrad	Anschlussleistung	Anzahl Ladepunkte	P1 [kW]	P2 [kW]	P3 [kW]	P4 [kW]
count	34722.000000	34722.000000	34722.000000	34722.000000	34722.000000	34722.000000	29069.000000	1844.000000	992.000000
mean	54541.003082	50.596165	9.693615	86.742523	1.921750	44.322247	38.344971	32.784706	27.978830
std	27178.574850	1.809136	1.995396	221.380981	0.543078	74.440018	64.349365	55.057880	32.992744
min	1067.000000	47.287800	5.243745	2.000000	1.000000	2.000000	2.000000	2.000000	3.000000
25%	31640.750000	48.921440	8.234325	22.000000	2.000000	22.000000	22.000000	22.000000	22.000000
50%	55411.000000	50.724632	9.441630	44.000000	2.000000	22.000000	22.000000	22.000000	22.000000
75%	78315.000000	52.034355	11.208292	50.000000	2.000000	22.000000	22.000000	22.000000	22.000000
max	99991.000000	55.019600	15.543810	5299.000000	4.000000	2175.000000	2175.000000	1125.000000	375.000000

> The above result shows the statistics of the continuous variables (consider Anschlussleistung Anzahl Ladepunkte P1 [kW] P2 [kW] P3 [kW] P4 [kW] only)

Milestone 4: EDA (Exploratory Data Analysis)

Power Distribution in Germany

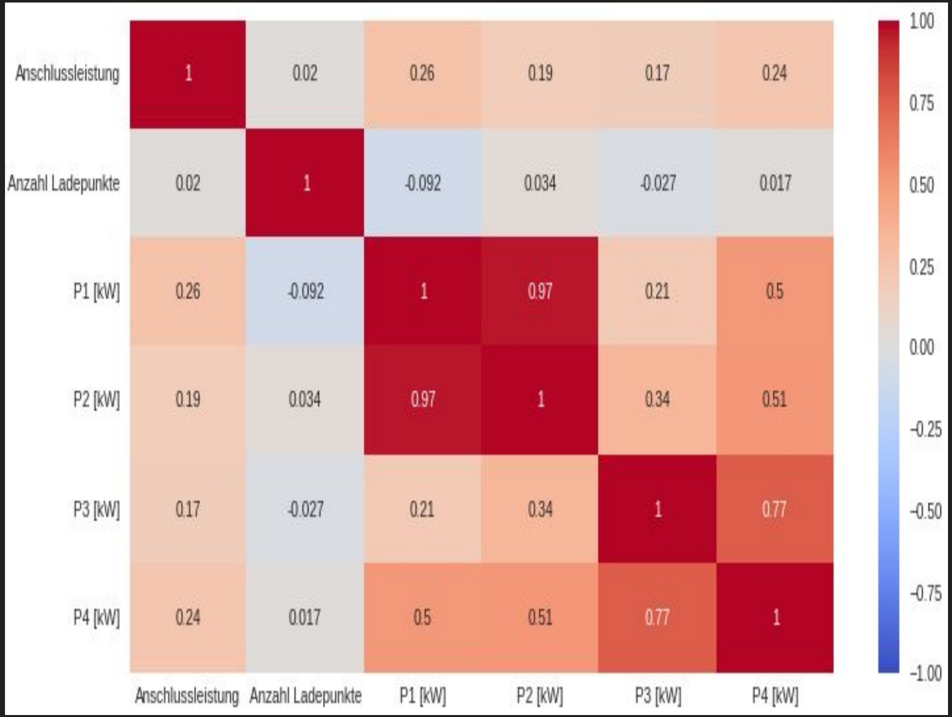
- Various fast chargers are clustered in some areas in Germany.



Milestone 4: EDA (Exploratory Data Analysis)

Correlation of Continuous Features (via Heatmap)

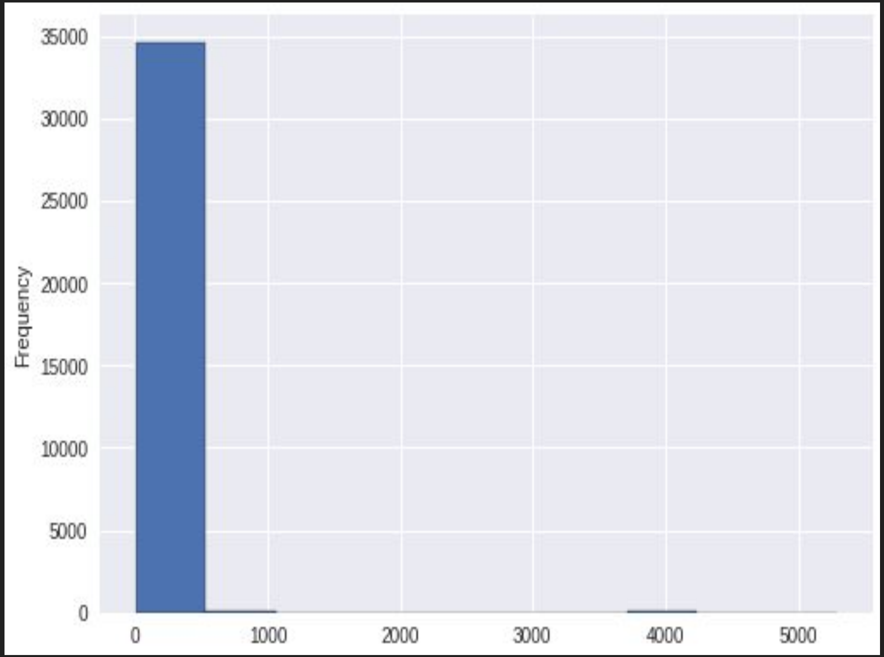
- Some powers are highly correlated with each other
  - E.g. P1 & P2, P3 & P4
- Very low correlation between max. power supply for cars & number of charging points



Milestone 4: EDA (Exploratory Data Analysis)

Distribution of max power supply for cars

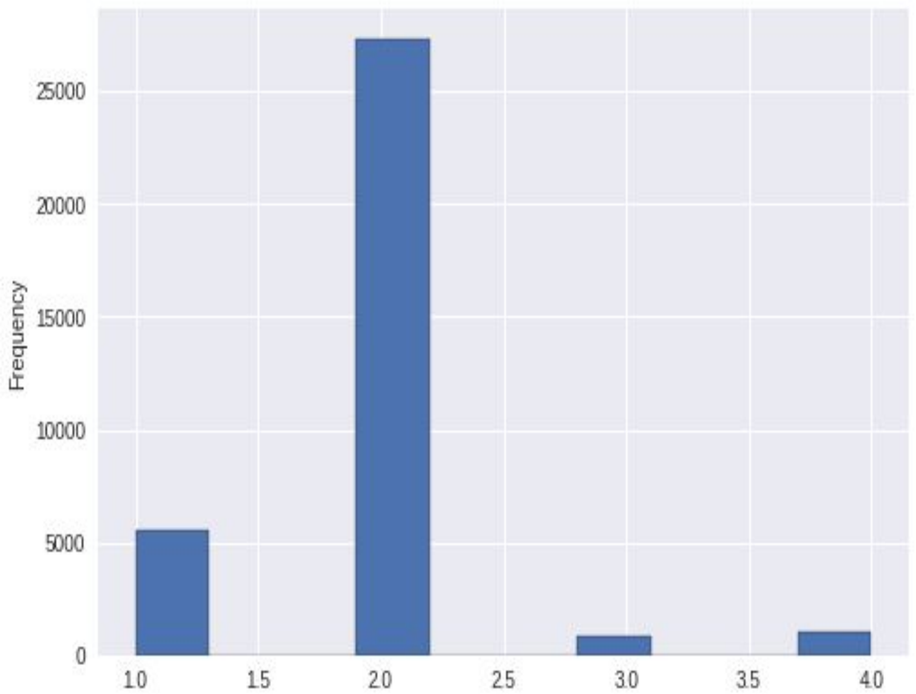
- Most of the max power supply for cars range from 0 to 500.



Milestone 4: EDA (Exploratory Data Analysis)

Distribution of charging points

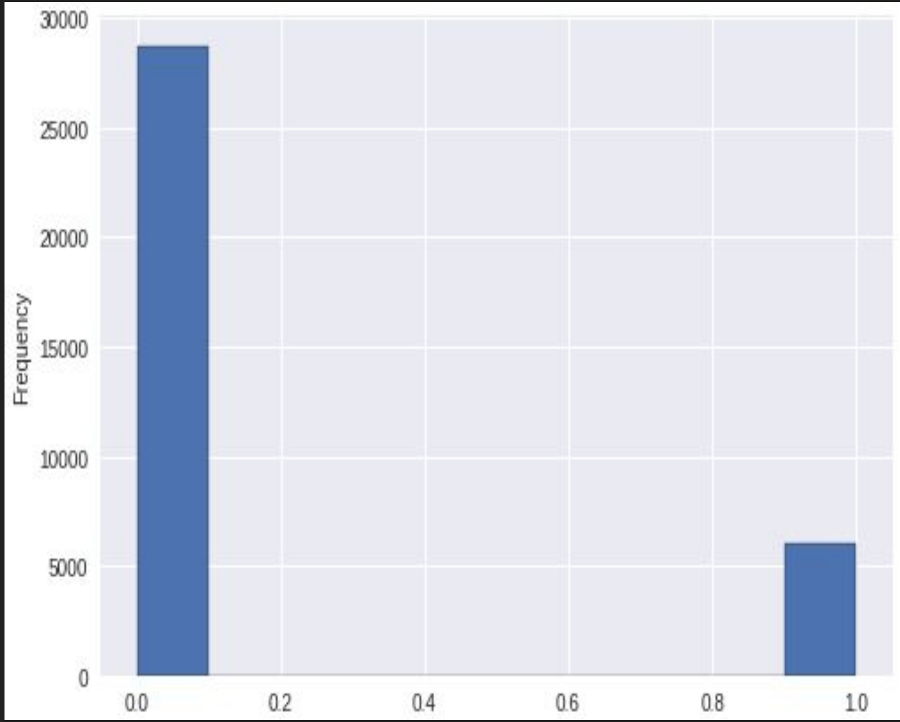
- 2 is the most number of charging point.



Milestone 4: EDA (Exploratory Data Analysis)

Distribution of charger types

- There are more normal chargers compared to fast chargers.



Milestone 4: EDA (Exploratory Data Analysis)

Major findings:

- Fast chargers are not evenly distributed across areas in Germany.
- There are fewer fast chargers compared to normal chargers.

Overview

Goal:

- **Build a model to predict regions in Germany with the highest potential charging stations demand**

Business benefit:

- Have a priority list of charging stations investment needs

Approach:

- Identify potential delta per region of charging stations *(EV:charging points)*

Technique:

- Unsupervised learning (K-Means)

Features:

- Charging station distribution, per longitude & latitude in Germany in 2020
- Electric vehicle distribution based on population density & EV per person, per longitude & latitude in Germany in 2020

Approach

Approach`s Rationale:

- The algorithm works by **grouping data points into clusters based on their similarity**, with each cluster represented by the mean (or "centroid") of the data points within it.

- In the case of electric vehicle charging stations, k-means can be used to **group the locations of charging stations into clusters**.

- By calculating the ratio of EV numbers to existing charging points, regions with lowest ratio are identified, which can then be used to determine the areas where new charging stations are needed.

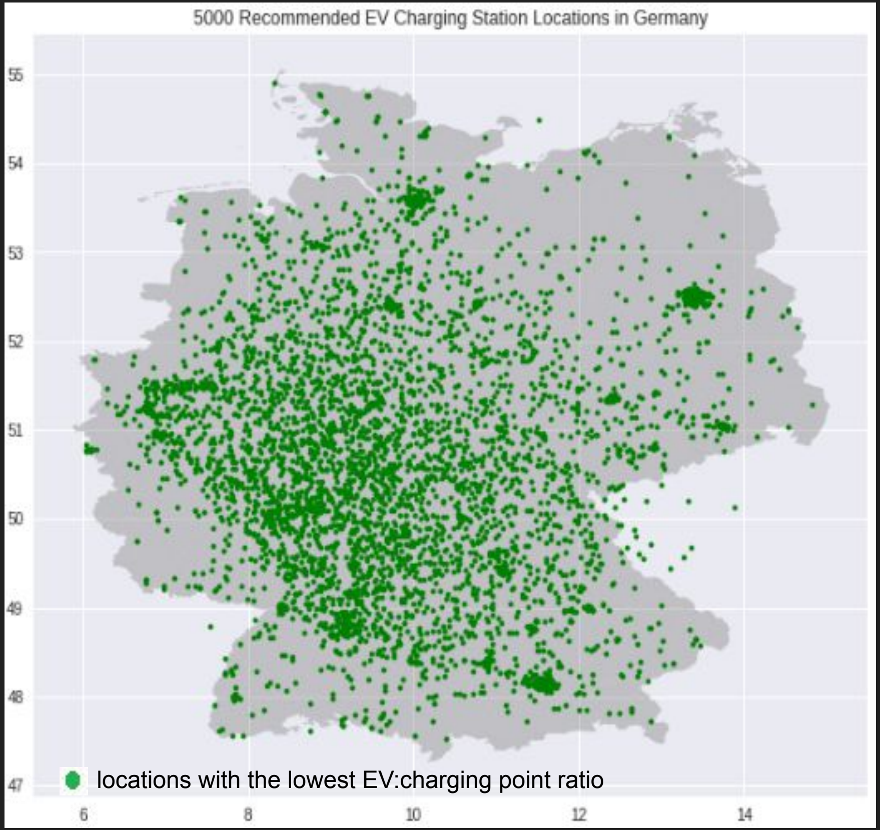
- This can help ensure that charging stations are evenly distributed based on population density throughout Germany, making it easier for electric vehicle owners to find a charging station when they need one.

Milestone 5: Findings

- By feeding the selected features to K-means, **5000 regions with the currently highest demand for more charging stations are identified**

- By using our model and different sets of data, the potential **future EV charging points demand** can be determined, however the model training time would take longer

- Our model can assist in **planning the EV charging infrastructure development** and, as a result, reaching Germany´s target of 15.8 million EV on the roads by 2030 and reducing CO_2 emissions from the transport sector



Performance

Model Performance

- Using statistical metrics, good scores are achieved in the performance of the model.
    - Silhouette coefficient: 0.515 (range -1 to 1)
    - Calinski-Harabasz index: 1454334.822 (0-infinity)
    - Dunn index: 0.593 (0-infinity)

Closing Remarks

- Our result is the model rather than the information, since it can be used to provide information about high charging point demand locations based on the datasets provided. It is therefore able to make predictions with the right database.

- The Datasets that were used were rather an abstraction of actual data and outdated (2020 & calculated EV per person based on total EVs)

- We used densities for 1 sqkm areas, which could be more accurate

- There are additional factors that affect a proper charging station placement like:
    - Availability of parking spaces
    - Proximity of highways/major roads
    - Availability of public transportation
    - Presence of amenities (restaurants etc.)
    - Proximity to residential areas