

# DATA CURATION PROJECT

ELECTRIC VEHICLE CHARGING STATIONS

PRESENTED BY

Sagar Shekhargouda Patil (A20501427)

Yuvraj Nikam (A20501952)

Marut Pandya (A20518560)

UNDER GUIDANCE OF DR.BORIS GLAVIC

# Agenda:

- Objective
- Dataset Introduction
- Data Quality Overview
- Framework used in Data Curation Process
- Data Cleaning and Transformation Tasks
- Tools used
- Results and Insights
- Conclusion

# Objective

- The primary objective of this project is to harness the power of data curation techniques, specifically focusing on data cleaning and transformation processes.
- Through meticulous data curation, our aim is to refine and cleanse the dataset, ensuring its accuracy and reliability.
- By employing various data cleaning methods and transformative approaches, we intend to enhance the overall quality of the dataset.

# Dataset Introduction :

Data source: Electric Vehicle Charging Stations

[https://afdc.energy.gov/fuels/electricity\\_locations.html#/analyze?fuel=ELEC](https://afdc.energy.gov/fuels/electricity_locations.html#/analyze?fuel=ELEC)

Data source metadata:

[https://afdc.energy.gov/data\\_download/alt\\_fuel\\_stations\\_format](https://afdc.energy.gov/data_download/alt_fuel_stations_format)

	Original Dataset	Modified Dataset
Attributes	74	32
Size	22.7MB	568 KB
Data Format	CSV	CSV

## Column reduction due to:

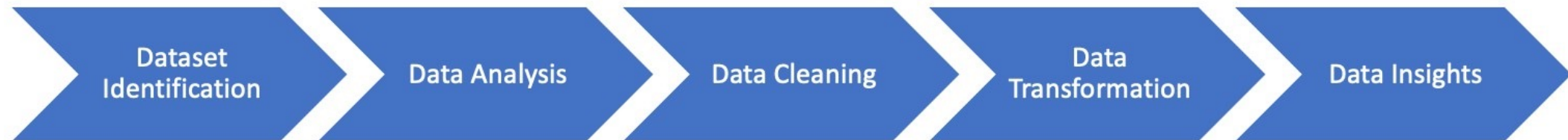
- Deprecated columns as per metadata
- No curation tasks in the removed columns
- Many null/blank values

# Data Quality Overview:

- Data quality relates to its accuracy, completeness, consistency, and validity. Identifying data quality issues and correcting them, we would need data that is ideal for use.
- Data quality issues arises due to following factors:
  - Inaccurate Data
  - Incomplete Data
  - Duplicate Data
  - Inconsistent Data
  - Unstructured Data
  - Irrelevant Data
- We have identified such problems in our source dataset which we will discuss in further slides.

# Framework used in Data Curation Process:

- We have used framework in data curation that uses a wide range of activities and processes done to create, manage, maintain, validate and showcase.
- Process Diagram :



# Tools used for Problem Identification and Curation

## Identification:

1. Microsoft Excel Spreadsheet
2. Vizier visualization

## Curation :

1. Vizier Tool
2. Jupyter Notebook
3. Python Libraries :
  - Pandas - Pandas is a Python library for data analysis and data manipulation. It offers data structures and operations for manipulating numerical tables and time series.
  - Re - Regex library for performing regular expression related operations over data.

# Data Cleaning and Transformation Tasks:

## 1. Handling Missing Values:

- Imputed missing values in numeric columns ('EV\_DC\_Fast\_Count').
- Filled categorical columns with appropriate defaults ('EV\_Network,' 'Provider,' 'Owner\_Type\_Code,' etc.).

## 2. Column Renaming:

- Renamed 'EV\_Network\_Web' to 'Provider'.

## 3. Regex Pattern Matching:

- Categorized 'Groups\_With\_Access\_Code' values as 'Public' or 'Private' using regex patterns.

## 4. Time Information:

- Modified 'Access\_Days\_Time,' categorized as "24 hours daily" or "Less than 24 hours"

## 5. Credit Card Information:

- Transformed 'Cards\_Accepted' into binary columns ('Credit,' 'Debit,' 'Cash,' 'No Payment Info').

## 6. Special Character Removal:

- Removed special characters from 'Station\_Name'

## 7. Value Replacement:

- Replaced specific values like 'NAME?' with 'NIU' in 'Station\_Name' and null values with 'Not Available' in 'Station\_Phone'

## 8. Handling Multiple Phone Numbers:

- Retained the first phone number if multiple in 'Station\_Phone'

## 9. Filling Missing Phone Numbers:

- Replaced missing phone numbers in 'Station\_Phone' with 'Not Available'



# Challenges during data curation:

- **Data Quality Issues:**
  1. Incomplete or missing data.
  2. Inconsistent data formats.
  3. Data inaccuracies or errors.
- **Data Heterogeneity:**
  - Dealing with diverse data types and formats.
- **Metadata Complexity:**
  1. Developing and managing comprehensive metadata.
  2. Ensuring consistency in metadata across datasets.
- **External Dependencies:**
  - Managing changes or disruptions in external data such as deprecations of columns.

# Solutions to data curation problems:

```
In [14]: 1 import pandas as pd
2 df = pd.read_csv('Cleaned_Dataset.csv')
3 # Replace missing values in 'EV_DC_Fast_Count' column with 0
4 df['EV_DC_Fast_Count'] = df['EV_DC_Fast_Count'].fillna(0)
5 df.to_csv('Cleaned_Dataset.csv', index=False)
```

```
In [15]: 1 import re
2 import numpy as np
3 df['EV_Network'] = df['EV_Network'].fillna('Non-Networked')
4 df=df.rename(columns={'EV_Network_Web': 'Provider'})
5 df['Provider']=df['Provider'].fillna('Not Applicable')
6 df['Owner_Type_Code']=df['Owner_Type_Code'].fillna('Unknown')
7 df['LPG_Primary']= df['LPG_Primary'].fillna('FALSE')
8 df['E85_Blender_Pump']= df['E85_Blender_Pump'].fillna('FALSE')
9 df['EV_Connector_Types']= df['EV_Connector_Types'].fillna('NA')
10 df['Access_Detail_Code']= df['Access_Detail_Code'].fillna('CALL')
11 df['CNG_Dispenser_Num']=df['CNG_Dispenser_Num'].fillna(0)
12 df['Restricted_Access']=df['Restricted_Access'].fillna('TRUE')
13 df['EV_Workplace_Charging']=df['EV_Workplace_Charging'].fillna('FALSE')
14 df['Facility_Type']=df['Facility_Type'].fillna('OTHER')
15
16 # Define regex patterns for Public and Private for Groups With Access Code Column
17 public_pattern = re.compile(r'\bpublic\b', flags=re.IGNORECASE)
18 private_pattern = re.compile(r'\bprivate\b', flags=re.IGNORECASE)
19 # Applying regex patterns to categorize values as Public and Private
20 df['Groups_With_Access_Code'] = df['Groups_With_Access_Code'].apply(lambda x: 'Public' if public_pattern.search(x) else 'Pri
21
```

# Solutions to data curation problems(Continued):

```
21
22 df["Access_Days_Time"] = df["Access_Days_Time"].fillna("NULL")
23 df["Access_Days_Time"] = np.where(df["Access_Days_Time"].str.contains("24 hours daily"), "24 hours daily", "Less than 24
24
25 #Cards_Accepted column transformation into 4 columns i.e. Credit,Debit,Cash,No Payment Info
26 df["Credit"] = np.where(df["Cards_Accepted"].str.contains("CREDIT"), 1, 0)
27 df["Cash"] = np.where(df["Cards_Accepted"].str.contains("Cash"), 1, 0)
28 df["Debit"] = np.where(df["Cards_Accepted"].str.contains("Debit"), 1, 0)
29 df["No Payment Info"] = np.where(df["Cards_Accepted"].isna(), 1, 0)
30 df=df.drop("Cards_Accepted", axis=1)
31
32 #Removing the Special Character from Station Name
33 df['Station_Name'] = df['Station_Name'].apply(lambda x: re.sub(r'^[a-zA-Z0-9]+', '', x))
34 # Replace 'Name?' with 'NIU' if present
35 df['Station_Name'] = df['Station_Name'].replace('NAME?', 'NIU')
```

```
In [21]: 1 # Filling missing values in Station_Phone column with 'Not Available'
2 df['Station_Phone'] = df['Station_Phone'].fillna('Not Available')
3 # If there are multiple phone numbers, choose one (here we are choosing the first one)
4 df['Station_Phone'] = df['Station_Phone'].apply(lambda x: x.split()[0] if ' ' in str(x) else x)
5 # Replace 'Not' with 'Not Available' in the 'Station_Phone' column
6 df['Station_Phone'] = df['Station_Phone'].replace('Not', 'Not Available')
```

```
In [22]: 1 df.to_csv('Cleaned_Dataset.csv', index=False)
```

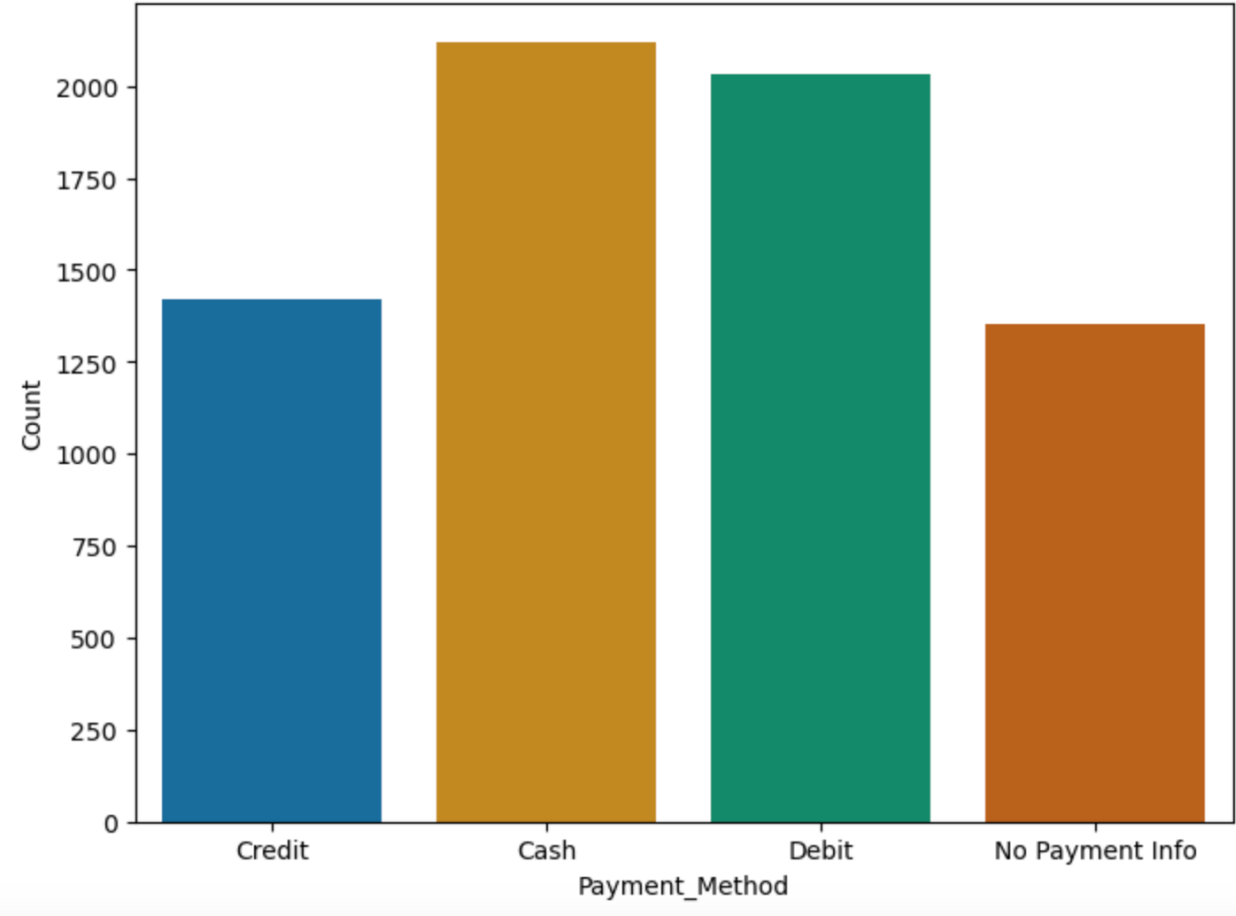
# Results and Insights:

- Cleaned and transformed Data Results

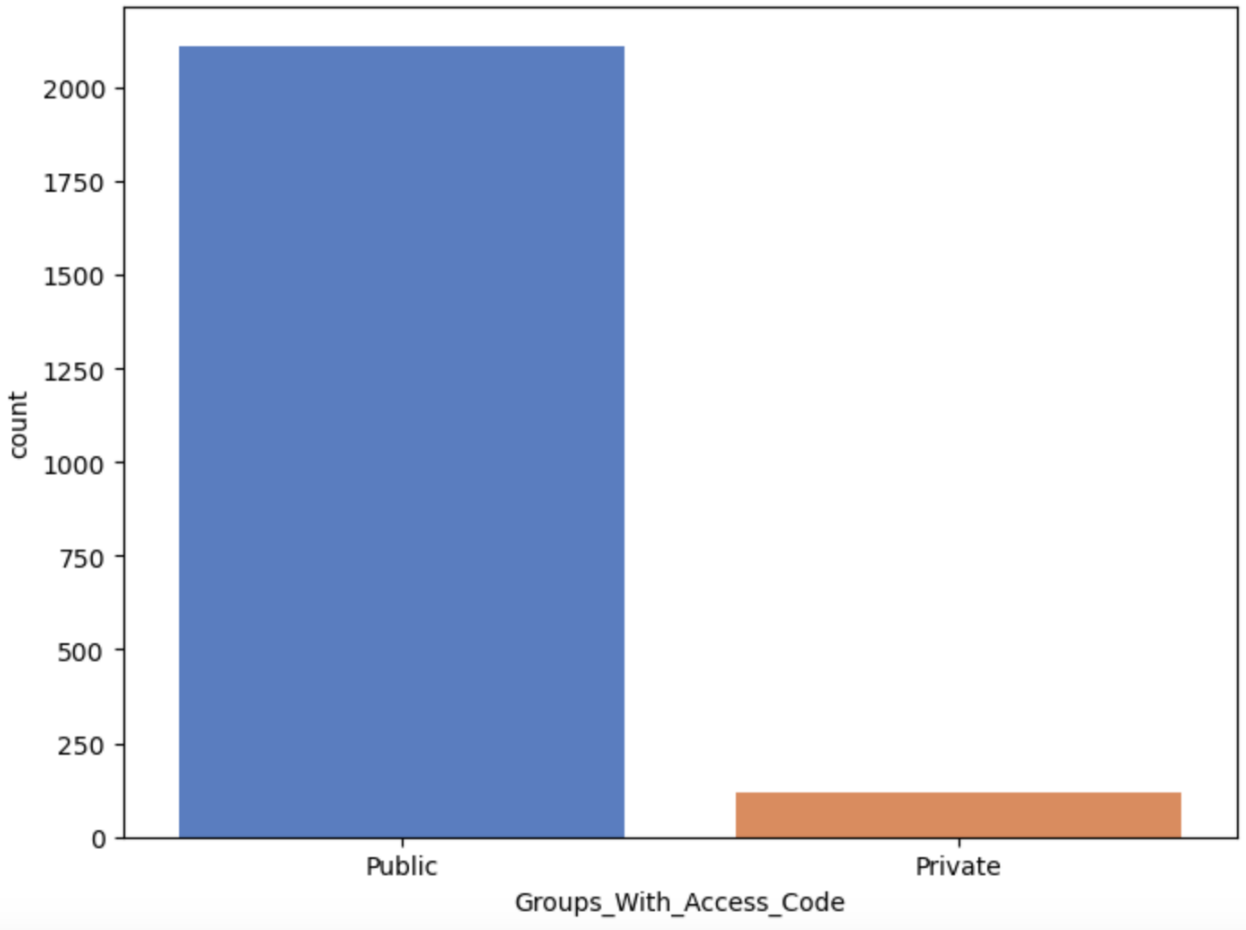
	Fuel_Type_Code	Station_Name	Street_Address	City	ZIP	Station_Phone	Status_Code	Groups_With_Access_Code	Access_Days_Time
0	LPG	Ferrellgas	10522 N 2nd St	Machesney Park	61115	815-877-7333	E	Public	Less than 24 hours
1	LPG	U-Haul	1560 Mount Prospect Rd	Des Plaines	60018	847-298-1170	E	Public	Less than 24 hours
2	LPG	U-Haul	2915 W 159th St	Markham	60428	708-333-7840	E	Public	Less than 24 hours
3	LPG	U-Haul	4650 W 95th St	Oak Lawn	60453	708-422-2332	E	Public	Less than 24 hours
4	LPG	U-Haul	1700 N Cicero Ave	Chicago	60639	773-889-8194	E	Public	Less than 24 hours
5	LPG	U-Haul	4301 N Cicero Ave	Chicago	60641	773-286-4507	E	Public	Less than 24 hours
6	LPG	U-Haul	1650 E 71st St	Chicago	60649	773-493-1206 708-389-0852	E	Public	Less than 24 hours
7	LPG	U-Haul	5027 W Cermak Rd	Cicero	60804	708-656-8890	E	Public	Less than 24 hours
8	LPG	U-Haul	306 E University Ave	Champaign	61820	217-351-7040	E	Public	Less than 24 hours
9	LPG	U-Haul	410 N Bruns Ln	Springfield	62702	217-546-2730	E	Public	Less than 24 hours
10	E85	Citgo	4070 N Clark St	Chicago	60613	773-528-3040	E	Public	24 hours daily

# Results and Insights: (Continued)

Payment Methods Accepted at Fuel Stations

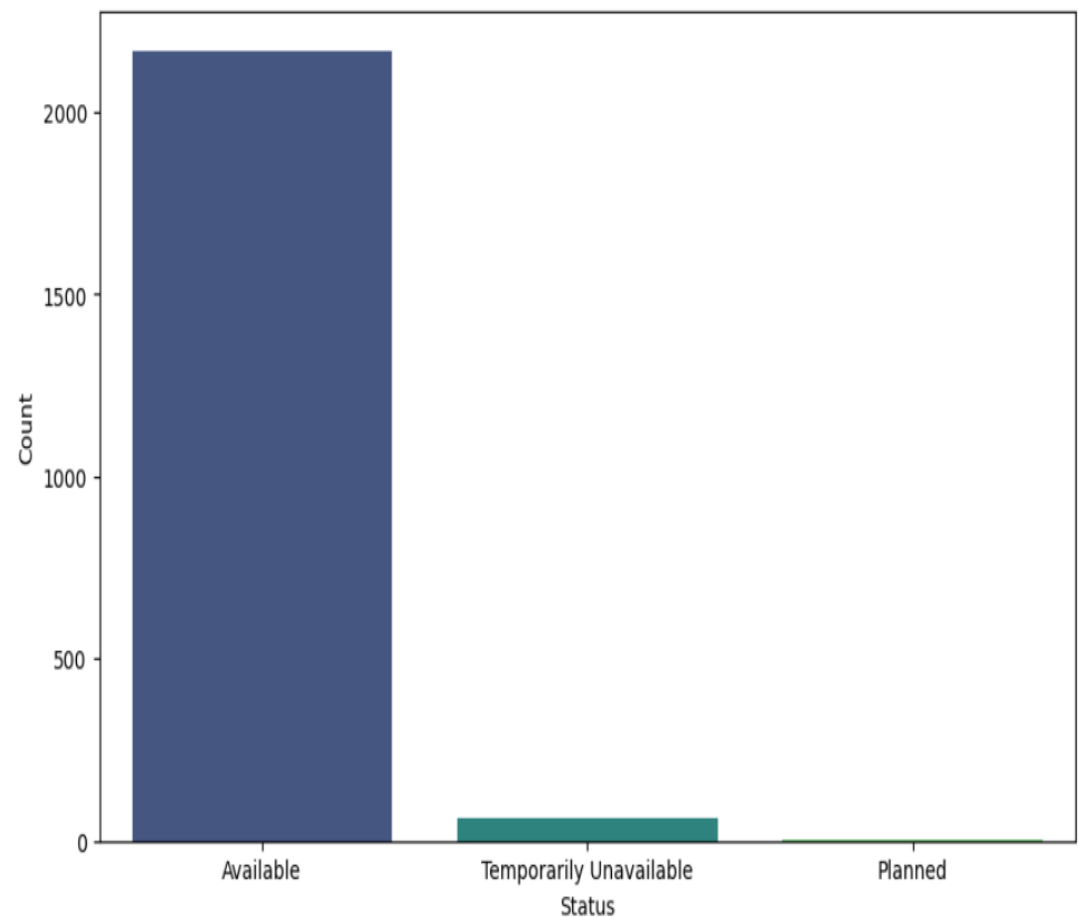


Access Code Distribution at Fuel Stations

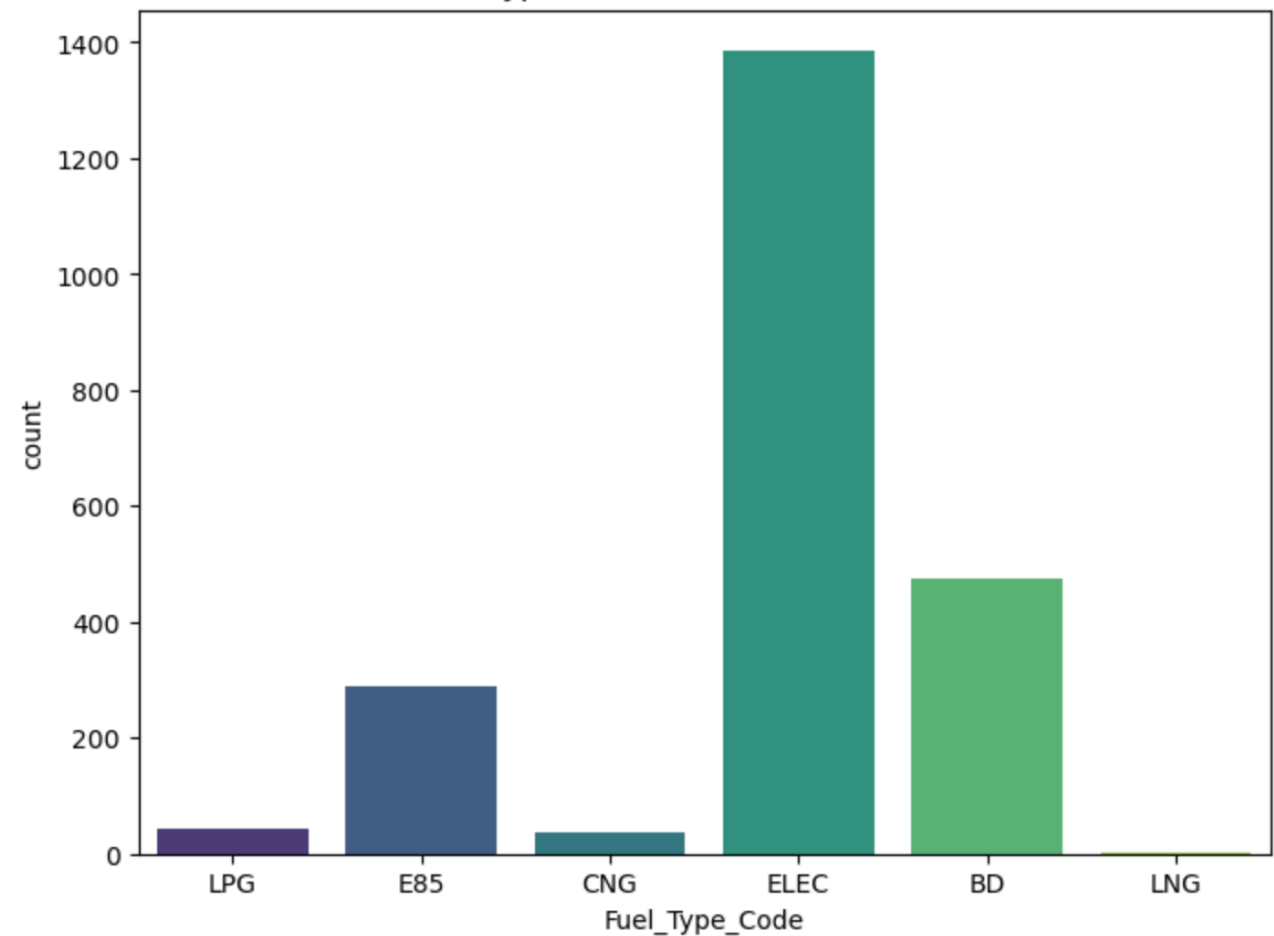


# Results and Insights: (Continued)

Station Status Distribution

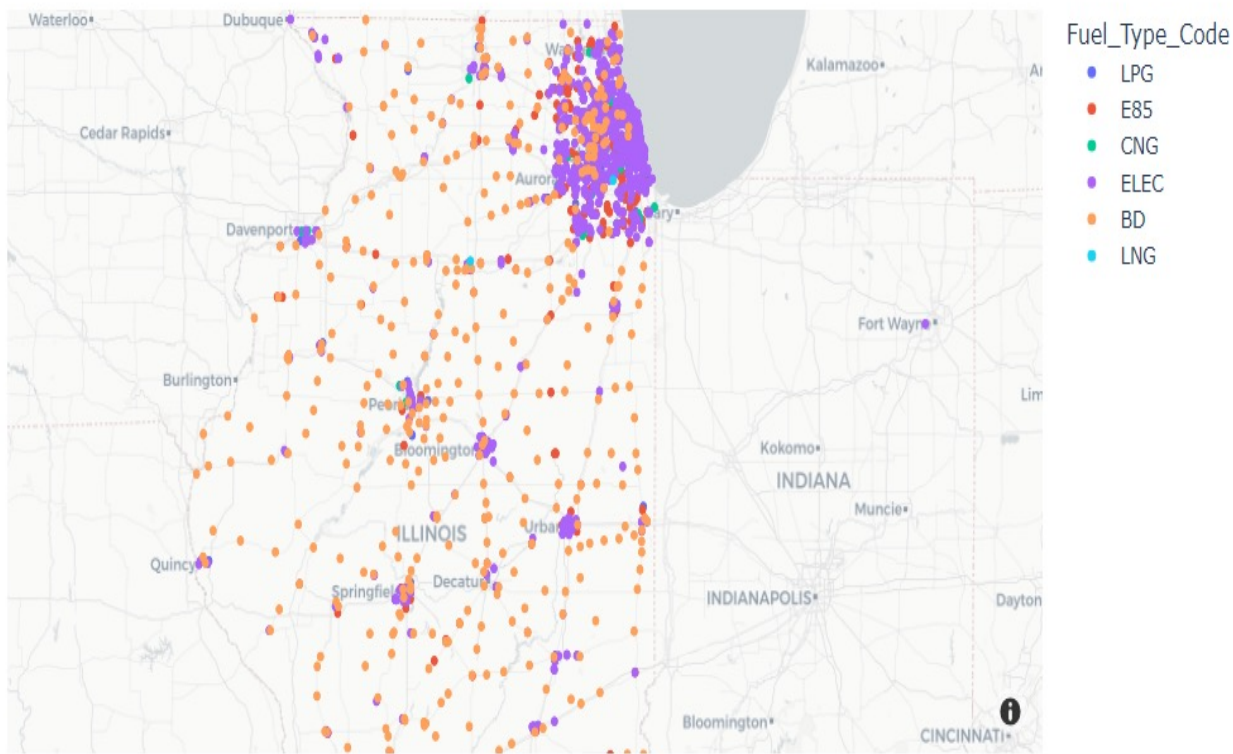


Fuel Type Distribution at Fuel Stations

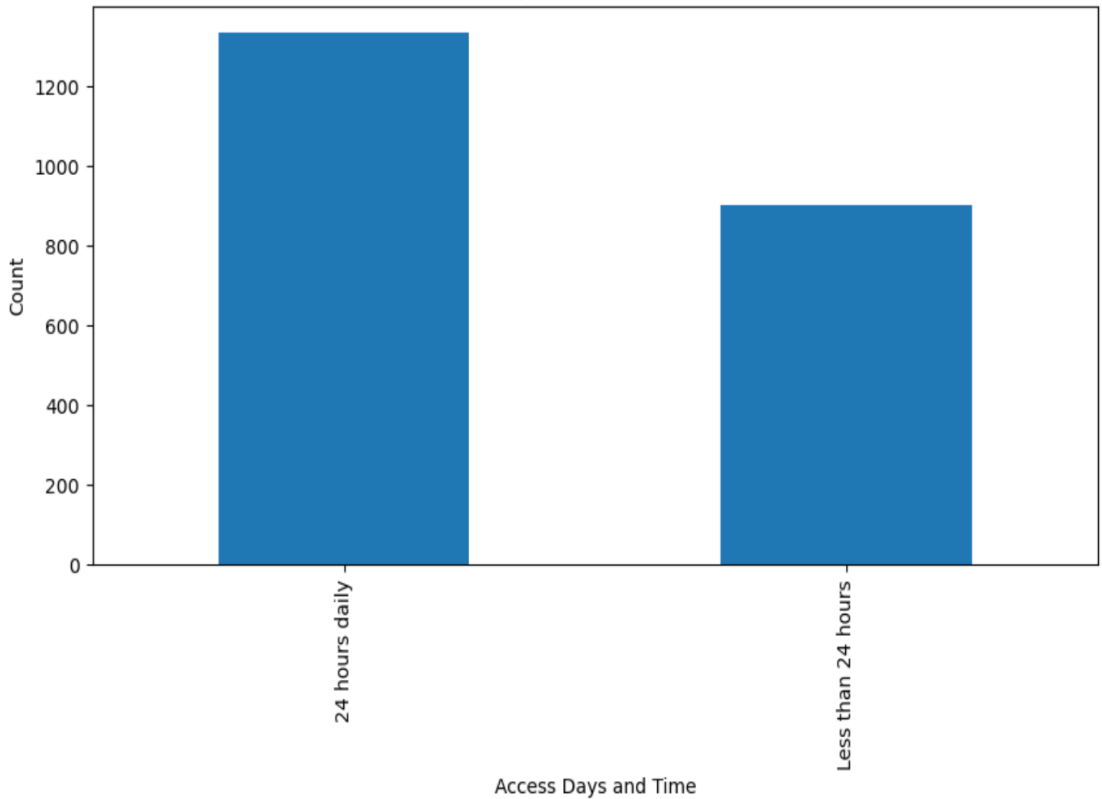


# Results and Insights: (Continued)

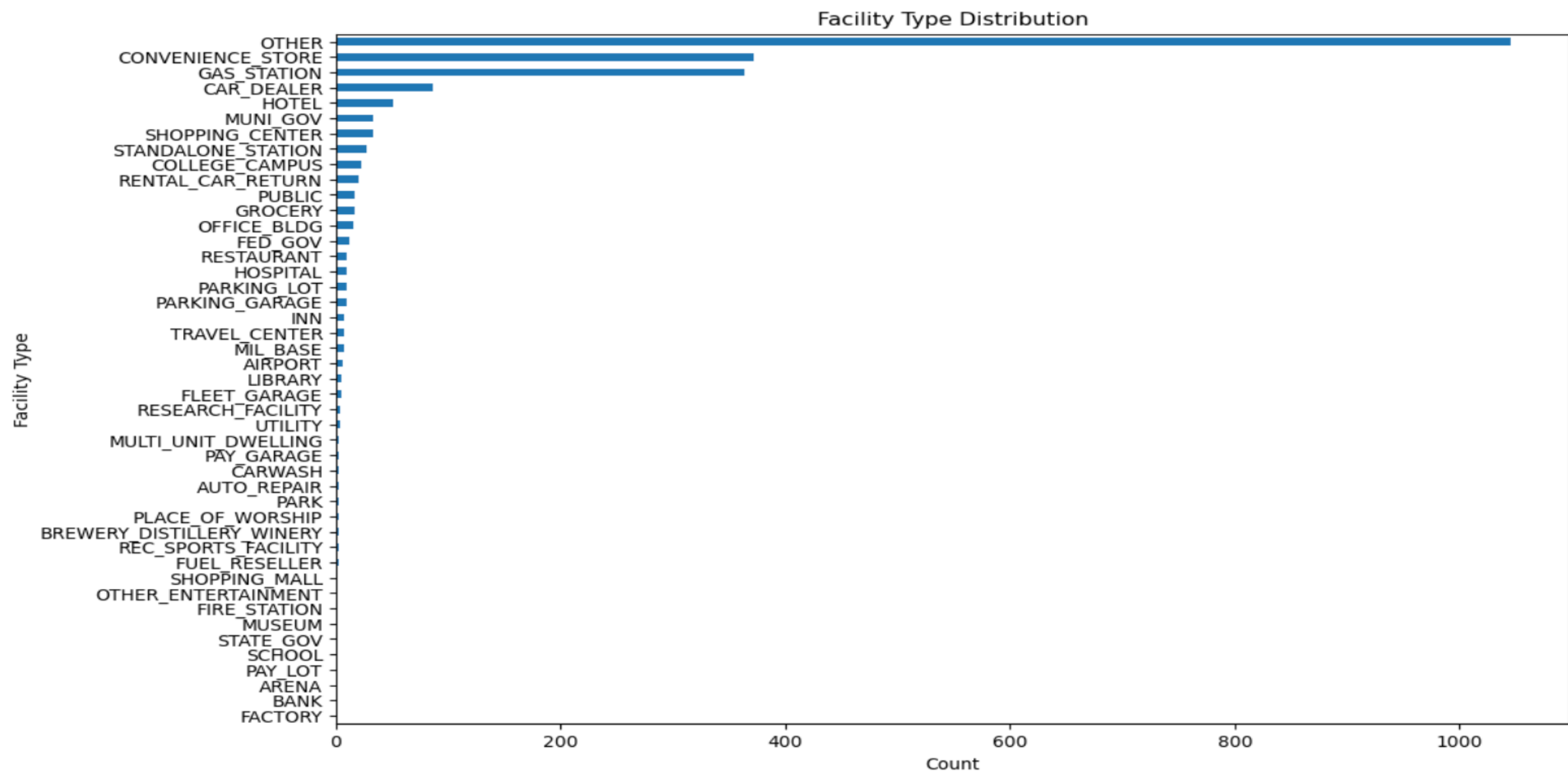
Geospatial Distribution of Fuel Stations



Access Days and Time Distribution

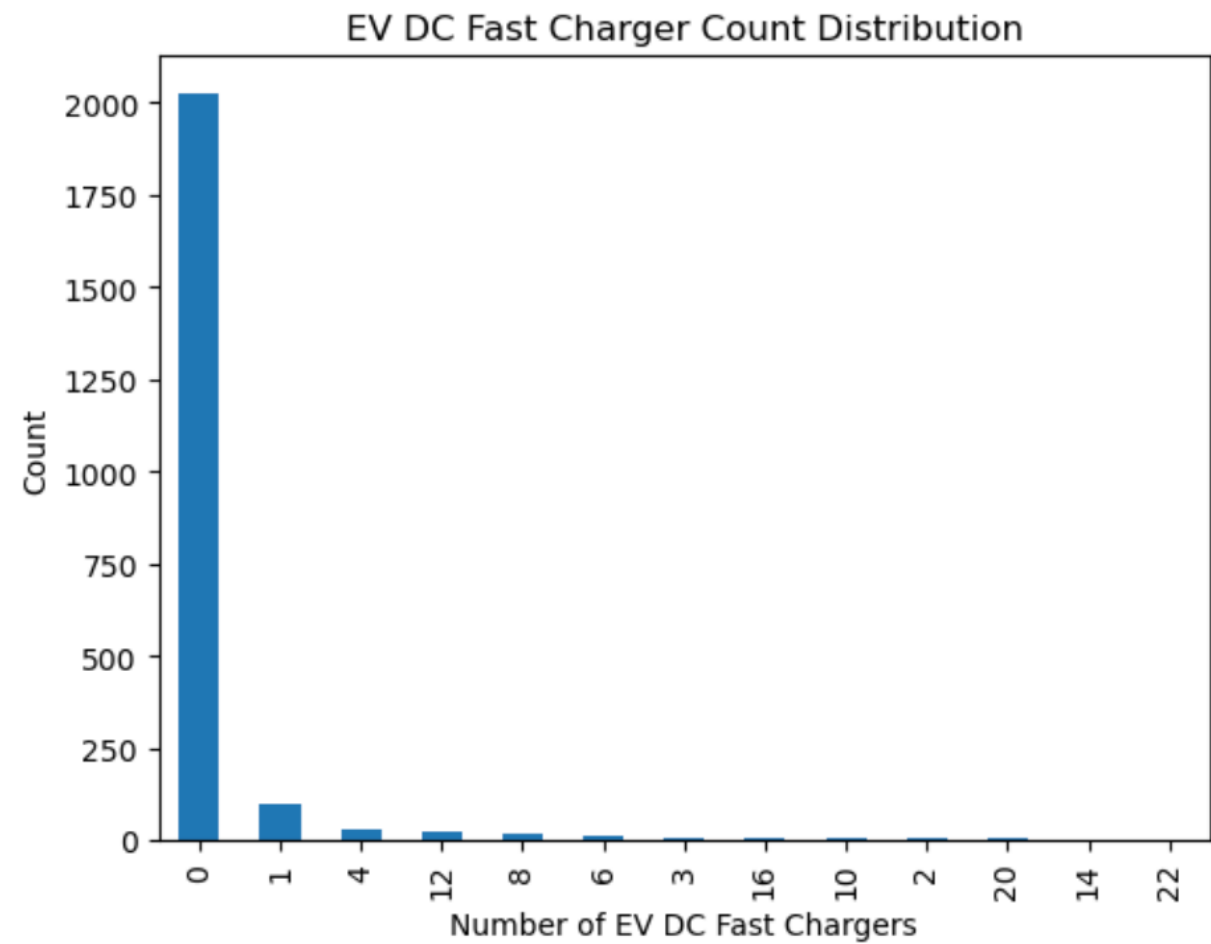


# Results and Insights:(Continued)





# Results and Insights: (Continued)



# Conclusion:

The data cleaning and transformation efforts have successfully prepared the dataset for meaningful analysis.

## Key accomplishments include:

- **Data Integrity:** Ensured data integrity by addressing missing values and applying consistent formatting.
- **Enhanced Readability:** Renaming columns for clarity and utilizing regex patterns improved the readability and interpretability of categorical data.
- **Standardized Time Information:** Normalized the representation of time information in the 'Access\_Days\_Time' column.
- **Improved Accessibility:** The transformation of credit card information into distinct columns facilitates a clear understanding of payment methods.
- **Standardized Naming:** Special character removal and value replacement enhanced the standardization of station names, contributing to a cleaner dataset.
- **Enhanced Phone Number Information:** Filled missing phone numbers, handled multiple entries, and replaced ambiguous values to ensure uniformity in the 'Station\_Phone' column.