In [13]:
```
import pandas as pd

# Load full dataset
ev_df = pd.read_csv("ev_charging_station_data.csv")

# Create a small sample (500 rows)
ev_small = ev_df.sample(n=500, random_state=42)

# Save the sample dataset
ev_small.to_csv("ev_charging_station_small.csv", index=False)

# Display first rows
ev_small.head()
```

Out[13]:

	timestamp	station_id	station_name	network	city	state	latitude	long
669199	2025-08-02 01:30:00	EV00077	Volta - Los Angeles #17	Volta	Los Angeles	CA	33.911204	-118.15
81467	2025-08-20 01:00:00	EV00010	Volta - San Diego #10	Volta	San Diego	CA	32.859690	-117.06
196547	2025-09-07 06:30:00	EV00023	Tesla Supercharger - Chicago #3	Tesla Supercharger	Chicago	IL	41.823295	-87.68
894968	2025-12-08 01:30:00	EV00102	Tesla Supercharger - Seattle #2	Tesla Supercharger	Seattle	WA	47.513232	-122.25
316362	2025-07-03 03:00:00	EV00037	Electrify America - Portland #17	Electrify America	Portland	OR	45.479089	-122.53

5 rows × 33 columns

In [15]:
```
ev_small = ev_df.sample(n=500, random_state=42)
ev_small.head()
```

Out[15]:

	timestamp	station_id	station_name	network	city	state	latitude	long
669199	2025-08-02 01:30:00	EV00077	Volta - Los Angeles #17	Volta	Los Angeles	CA	33.911204	-118.15
81467	2025-08-20 01:00:00	EV00010	Volta - San Diego #10	Volta	San Diego	CA	32.859690	-117.06
196547	2025-09-07 06:30:00	EV00023	Tesla Supercharger - Chicago #3	Tesla Supercharger	Chicago	IL	41.823295	-87.68
894968	2025-12-08 01:30:00	EV00102	Tesla Supercharger - Seattle #2	Tesla Supercharger	Seattle	WA	47.513232	-122.25
316362	2025-07-03 03:00:00	EV00037	Electrify America - Portland #17	Electrify America	Portland	OR	45.479089	-122.53

5 rows × 33 columns

In [16]:
```
station_details = ev_small[
    ['station_id', 'station_name', 'network', 'city', 'state', 'latitude'
].drop_duplicates()

station_details.head()
```

Out[16]:

	station_id	station_name	network	city	state	latitude	longitude	loca
669199	EV00077	Volta - Los Angeles #17	Volta	Los Angeles	CA	33.911204	-118.153220	
81467	EV00010	Volta - San Diego #10	Volta	San Diego	CA	32.859690	-117.067481	Hotel/
196547	EV00023	Tesla Supercharger - Chicago #3	Tesla Supercharger	Chicago	IL	41.823295	-87.682445	Urb
894968	EV00102	Tesla Supercharger - Seattle #2	Tesla Supercharger	Seattle	WA	47.513232	-122.254816	
316362	EV00037	Electrify America - Portland #17	Electrify America	Portland	OR	45.479089	-122.538465	Hotel/

In [17]:
```
charging_details = ev_small[
    [
        'station_id',
        'charger_type',
        'power_output_kw',
        'ports_total',
        'ports_available',
        'ports_occupied',
        'ports_out_of_service',
        'utilization_rate',
        'estimated_wait_time_mins',
        'avg_session_duration_mins',
        'current_price',
        'pricing_type'
    ]
].drop_duplicates()

charging_details.head()
```

Out[17]:

	station_id	charger_type	power_output_kw	ports_total	ports_available	ports_occupi
669199	EV00077	Level 2	7.2	8	7	
81467	EV00010	Level 2	7.2	4	3	
196547	EV00023	Tesla DC Fast	350.0	4	4	
894968	EV00102	Tesla DC Fast	250.0	6	5	
316362	EV00037	Hyper-Fast	150.0	4	4	

In [18]:
```
ev_integrated = pd.merge(
    station_details,
    charging_details,
    on='station_id',
    how='inner'
)

ev_integrated.head()
```

Out[18]:

	station_id	station_name	network	city	state	latitude	longitude	location_type
0	EV00077	Volta - Los Angeles #17	Volta	Los Angeles	CA	33.911204	-118.153220	Airport
1	EV00077	Volta - Los Angeles #17	Volta	Los Angeles	CA	33.911204	-118.153220	Airport
2	EV00077	Volta - Los Angeles #17	Volta	Los Angeles	CA	33.911204	-118.153220	Airport
3	EV00010	Volta - San Diego #10	Volta	San Diego	CA	32.859690	-117.067481	Hotel/Hospitality
4	EV00010	Volta - San Diego #10	Volta	San Diego	CA	32.859690	-117.067481	Hotel/Hospitality

In [19]:

```
ev_part1 = ev_small.iloc[:250]
ev_part2 = ev_small.iloc[250:]

ev_concat_rows = pd.concat([ev_part1, ev_part2], ignore_index=True)
ev_concat_rows.head()
```

Out[19]:

	timestamp	station_id	station_name	network	city	state	latitude	longitude
0	2025-08-02 01:30:00	EV00077	Volta - Los Angeles #17	Volta	Los Angeles	CA	33.911204	-118.153220
1	2025-08-20 01:00:00	EV00010	Volta - San Diego #10	Volta	San Diego	CA	32.859690	-117.067481
2	2025-09-07 06:30:00	EV00023	Tesla Supercharger - Chicago #3	Tesla Supercharger	Chicago	IL	41.823295	-87.682445
3	2025-12-08 01:30:00	EV00102	Tesla Supercharger - Seattle #2	Tesla Supercharger	Seattle	WA	47.513232	-122.254816
4	2025-07-03 03:00:00	EV00037	Electrify America - Portland #17	Electrify America	Portland	OR	45.479089	-122.538465

5 rows × 33 columns

In [20]:

```
ev_part1 = ev_small.iloc[:250]
ev_part2 = ev_small.iloc[250:]

ev_concat_rows = pd.concat([ev_part1, ev_part2], ignore_index=True)
ev_concat_rows.head()
```

Out[20]:

	timestamp	station_id	station_name	network	city	state	latitude	longitude
0	2025-08-02 01:30:00	EV00077	Volta - Los Angeles #17	Volta	Los Angeles	CA	33.911204	-118.153220
1	2025-08-20 01:00:00	EV00010	Volta - San Diego #10	Volta	San Diego	CA	32.859690	-117.067481
2	2025-09-07 06:30:00	EV00023	Tesla Supercharger - Chicago #3	Tesla Supercharger	Chicago	IL	41.823295	-87.682445
3	2025-12-08 01:30:00	EV00102	Tesla Supercharger - Seattle #2	Tesla Supercharger	Seattle	WA	47.513232	-122.254816
4	2025-07-03 03:00:00	EV00037	Electrify America - Portland #17	Electrify America	Portland	OR	45.479089	-122.538465

5 rows × 33 columns

In [21]:
```
ev_concat_cols = pd.concat([station_details, charging_details], axis=1)
ev_concat_cols.head()
```

Out[21]:

	station_id	station_name	network	city	state	latitude	longitude	loca
669199	EV00077	Volta - Los Angeles #17	Volta	Los Angeles	CA	33.911204	-118.153220	
81467	EV00010	Volta - San Diego #10	Volta	San Diego	CA	32.859690	-117.067481	Hotel/
196547	EV00023	Tesla Supercharger - Chicago #3	Tesla Supercharger	Chicago	IL	41.823295	-87.682445	Urb
894968	EV00102	Tesla Supercharger - Seattle #2	Tesla Supercharger	Seattle	WA	47.513232	-122.254816	
316362	EV00037	Electrify America - Portland #17	Electrify America	Portland	OR	45.479089	-122.538465	Hotel/

5 rows × 21 columns

In [27]:
```
data.describe()
```

Out[27]:

	latitude	longitude	power_output_kw	ports_total	ports_available	ports_o
count	1.317750e+06	1.317750e+06	1.317750e+06	1.317750e+06	1.317750e+06	1.317
mean	3.744355e+01	-1.016788e+02	1.345880e+02	7.306667e+00	4.199560e+00	2.393
std	5.894349e+00	1.751827e+01	1.285099e+02	4.483968e+00	3.420840e+00	3.323
min	2.561943e+01	-1.228127e+02	7.200000e+00	2.000000e+00	0.000000e+00	0.000
25%	3.348356e+01	-1.172766e+02	1.920000e+01	4.000000e+00	2.000000e+00	0.000
50%	3.697384e+01	-1.050262e+02	1.000000e+02	6.000000e+00	4.000000e+00	1.000
75%	4.197794e+01	-8.452968e+01	2.500000e+02	8.000000e+00	6.000000e+00	3.000
max	4.772777e+01	-7.091298e+01	3.500000e+02	2.400000e+01	2.400000e+01	2.300

In [ ]:
```
data.dtypes
```

In [37]:
```
ev_df["station_id"].dtype
ev_df["station_name"].dtype
ev_df["city"].dtype
ev_df["state"].dtype
ev_df["charger_type"].dtype
ev_df["power_output_kw"].dtype
ev_df["ports_total"].dtype
ev_df["ports_available"].dtype
ev_df["station_status"].dtype
ev_df["current_price"].dtype
ev_df["is_weekend"].dtype
```

Out[37]:
```
dtype('bool')
```

In [42]:
```
print(data.dtypes)
```

```
timestamp                    object
station_id                   object
station_name                 object
network                      object
city                         object
state                        object
latitude                    float64
longitude                   float64
location_type                object
charger_type                 object
power_output_kw             float64
amenities_nearby             object
ports_total                   int64
ports_available               int64
ports_occupied                int64
ports_out_of_service          int64
utilization_rate            float64
station_status               object
estimated_wait_time_mins      int64
avg_session_duration_mins     int64
current_price               float64
pricing_type                 object
temperature_f               float64
precipitation_mm            float64
weather_condition            object
gas_price_per_gallon        float64
traffic_congestion_index      int64
local_event                  object
is_weekend                     bool
is_peak_hour                   bool
hour_of_day                   int64
day_of_week                   int64
month                         int64
dtype: object
```

In [43]:
```
data.isnull().sum()
```

Out[43]:
```
timestamp                    0
station_id                   0
station_name                 0
network                      0
city                         0
state                        0
latitude                     0
longitude                    0
location_type                0
charger_type                 0
power_output_kw              0
amenities_nearby             0
ports_total                  0
ports_available              0
ports_occupied               0
ports_out_of_service         0
utilization_rate             0
station_status               0
estimated_wait_time_mins     0
avg_session_duration_mins    0
current_price                0
pricing_type                 0
temperature_f                0
precipitation_mm             0
weather_condition            0
gas_price_per_gallon         0
traffic_congestion_index     0
local_event                  0
is_weekend                   0
is_peak_hour                 0
hour_of_day                  0
day_of_week                  0
month                        0
dtype: int64
```

In [44]:
```
station_details = data[['station_id', 'station_name', 'network',
                        'city', 'state', 'latitude', 'longitude',
                        'location_type', 'station_status']].drop_duplicat

station_details.head()
```

Out[44]:

	station_id	station_name	network	city	state	latitude	longitude	location_
0	EV00001	ChargePoint - Los Angeles #1	ChargePoint	Los Angeles	CA	34.012362	-118.114786	Shop Ce
1	EV00001	ChargePoint - Los Angeles #1	ChargePoint	Los Angeles	CA	34.012362	-118.114786	Shop Ce
10	EV00001	ChargePoint - Los Angeles #1	ChargePoint	Los Angeles	CA	34.012362	-118.114786	Shop Ce
11	EV00001	ChargePoint - Los Angeles #1	ChargePoint	Los Angeles	CA	34.012362	-118.114786	Shop Ce
8785	EV00002	Electrify America - Seattle #2	Electrify America	Seattle	WA	47.506806	-122.433202	Reside

In [45]:
```
charger_details = data[['station_id', 'charger_type', 'power_output_kw',
                        'ports_total', 'ports_available', 'ports_occupied
                        'ports_out_of_service', 'utilization_rate']]

charger_details.head()
```

Out[45]:

	station_id	charger_type	power_output_kw	ports_total	ports_available	ports_occupied	p
0	EV00001	DC Fast Charge	19.2	6	1	0	
1	EV00001	DC Fast Charge	19.2	6	6	0	
2	EV00001	DC Fast Charge	19.2	6	6	0	
3	EV00001	DC Fast Charge	19.2	6	6	0	
4	EV00001	DC Fast Charge	19.2	6	6	0	

In [9]:
```
print("Total duplicates:", data.duplicated().sum())

data = data.drop_duplicates()

print("After removing duplicates:", data.duplicated().sum())
```

```
Total duplicates: 0
After removing duplicates: 0
```

In [ ]:
```
data transformation
```

In [16]:
```
# Step 1: Wide to Long (ports related columns)
ports_long = pd.melt(
    data,
    id_vars=['timestamp', 'station_id', 'station_name', 'network', 'city'
    value_vars=['ports_total', 'ports_available', 'ports_occupied', 'port
    var_name='port_status',
    value_name='count'
)

ports_long.head()
```

Out[16]:

	timestamp	station_id	station_name	network	city	state	port_status	count
0	2025-07-01 00:00:00	EV00001	ChargePoint - Los Angeles #1	ChargePoint	Los Angeles	CA	ports_total	6
1	2025-07-01 00:30:00	EV00001	ChargePoint - Los Angeles #1	ChargePoint	Los Angeles	CA	ports_total	6
2	2025-07-01 01:00:00	EV00001	ChargePoint - Los Angeles #1	ChargePoint	Los Angeles	CA	ports_total	6
3	2025-07-01 01:30:00	EV00001	ChargePoint - Los Angeles #1	ChargePoint	Los Angeles	CA	ports_total	6
4	2025-07-01 02:00:00	EV00001	ChargePoint - Los Angeles #1	ChargePoint	Los Angeles	CA	ports_total	6

In [17]:
```
ports_wide = ports_long.pivot_table(
    index=['timestamp', 'station_id', 'station_name', 'network', 'city',
    columns='port_status',
    values='count'
).reset_index()

ports_wide.head()
```

Out[17]:

port_status	timestamp	station_id	station_name	network	city	state	ports_availa
0	2025-07-01	EV00001	ChargePoint - Los Angeles #1	ChargePoint	Los Angeles	CA	
1	2025-07-01	EV00002	Electrify America - Seattle #2	Electrify America	Seattle	WA	
2	2025-07-01	EV00003	Electrify America - Atlanta #3	Electrify America	Atlanta	GA	
3	2025-07-01	EV00004	ChargePoint - Minneapolis #4	ChargePoint	Minneapolis	MN	
4	2025-07-01	EV00005	Blink - Phoenix #5	Blink	Phoenix	AZ	

In [18]:
```
agg = data.groupby('charger_type').agg({
    'power_output_kw': ['mean', 'sum', 'max', 'min'],
    'ports_total': ['mean', 'sum', 'max', 'min'],
    'utilization_rate': ['mean', 'max', 'min']
})

agg
```

Out[18]:

	power_output_kw				ports_total				util
	mean	sum	max	min	mean	sum	max	min	mean
charger_type									
DC Fast Charge	154.101449	93410905.0	350.0	7.2	6.608696	4005960	24	2	0.406909
Hyper-Fast	222.727273	21523250.0	350.0	150.0	6.000000	579810	12	2	0.365382
Level 2	29.697959	12783932.0	125.0	7.2	7.346939	3162600	24	2	0.437067
Tesla DC Fast	269.047619	49635250.0	350.0	150.0	10.190476	1879990	24	4	0.424425

In [21]:
```
scaling['power_standardized'] = (
    scaling['power_output_kw'] - scaling['power_output_kw'].mean()
) / scaling['power_output_kw'].std()

scaling[['power_output_kw', 'power_standardized']].head()
```

Out[21]:

	power_output_kw	power_standardized
0	19.2	-0.897892
1	19.2	-0.897892
2	19.2	-0.897892
3	19.2	-0.897892
4	19.2	-0.897892

In [23]:

```
from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()

# Encoding categorical columns
data['charger_type_encoded'] = le.fit_transform(data['charger_type'])
data['station_status_encoded'] = le.fit_transform(data['station_status'])
data['location_type_encoded'] = le.fit_transform(data['location_type'])
data['pricing_type_encoded'] = le.fit_transform(data['pricing_type'])
data['weather_condition_encoded'] = le.fit_transform(data['weather_condit

data[['charger_type', 'charger_type_encoded',
      'station_status', 'station_status_encoded']].head()
```

Out[23]:

	charger_type	charger_type_encoded	station_status	station_status_encoded
0	DC Fast Charge	0	under_maintenance	3
1	DC Fast Charge	0	operational	1
2	DC Fast Charge	0	operational	1
3	DC Fast Charge	0	operational	1
4	DC Fast Charge	0	operational	1

In [24]:

```
data_final = data.drop(columns=[
    'timestamp',
    'station_name',
    'station_id',
    'charger_type',
    'station_status',
    'location_type',
    'pricing_type',
    'weather_condition'
])

data_final.head()
```

Out[24]:

	network	city	state	latitude	longitude	power_output_kw	amenities_nearby	p
0	ChargePoint	Los Angeles	CA	34.012362	-118.114786	19.2	Hotel	
1	ChargePoint	Los Angeles	CA	34.012362	-118.114786	19.2	Hotel	
2	ChargePoint	Los Angeles	CA	34.012362	-118.114786	19.2	Hotel	
3	ChargePoint	Los Angeles	CA	34.012362	-118.114786	19.2	Hotel	
4	ChargePoint	Los Angeles	CA	34.012362	-118.114786	19.2	Hotel	

5 rows × 34 columns

In [ ]: