Link to Our Work: Data Exploration

Link to github repo: https://github.com/hemanthreddy-1711/Data-Mining-Project-Unstoppable

1. Our Team Name and Team members along with emails

Name of the Team: Unstoppable

Team Members Names, UNH Email

a. Hemanth Reddy Reddy Battula - hredd2@unh.newhaven.edu
b. Kola Avinash - akola9@unh.newhaven.edu
c. Prem Niranjan Undavali - punda1@unh.newhaven.edu

2. Our Research question and Selected dataset

Research Question

How does the power demand of electric vehicle (EV) charging stations vary across different cities and what regional factors influence these variations?

About the data set: https://www.kaggle.com/datasets/omarsobhy14/supercharge-locations

By solving our research question, we will be able to understand the demand of EV across different locations. The dataset that we selected has the details of charging station locations, power demand, street address, city, state, zip, country, kilo watts, GPS, Elev. The data is from 25 EV charging stations and has 5000 sessions with data from more than 100 drivers. The data types in the data set are strings, timestamps, categorical, integer and more.

3. List of the exploration techniques, which you used in this work.

We have Performed Different Data Exploration Techniques:

1. **Performed Analysis for Univariate data:**
    a. **Histograms**: plotted histograms to visualize the numeric data distribution.
    b. **Box plots**: plotted box plots to identify the outliers
    c. **Summary statistics**: we have calculated the mean, median and mode for the data.

2. **Performed Analysis for Bivariate Analysis:**
    a. Performed correlation analysis: we have performed the correlation analysis to identify the relationship between the numeric values.
    b. Scatter plots: Similarly plotted the scatter plot to identify relationship.

3. **Performed Analysis on Categorical data using**
    a. Pie charts: plotted the pie chart to visualize the categorical data.
    b. Bar charts: plotted bar chart to compare the categorical data with one another.

4. **Performed Data Cleaning**
    a. We have calculated missing values and handled them using the imputation if mode.
    b. Then detected outliers and removed them using IQR method
    c. Then performed the data scaling using standard scalar.

4. **Describe your data explorations from different perspectives using varied visualization techniques.**

My Findings after data exploration:

1. **Missing Values:**
    a. When checked for existence of missing values in dataset we found that missing values are present in kW, State, Zip, and Open Date
    b. Then using imputed strategies using mean for KW. mode for categorical variables we have handled them.

```
# Check for missing values
missing_data = data.isnull().sum()
missing_data[missing_data > 0]
```

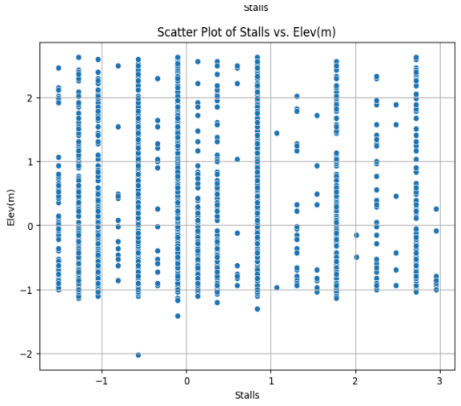
	0
State	122
Zip	1929
kW	6
Open Date	750

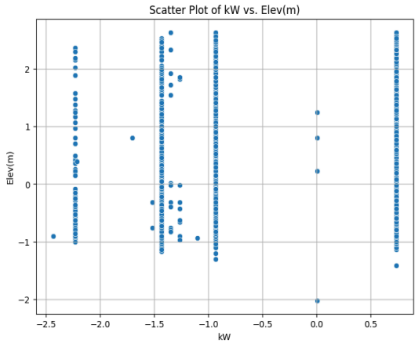
dtype: int64

```
# adding data to missing values
data['kW'] = data['kW'].fillna(data['kW'].mean())


data['State'] = data['State'].fillna(data['State'].mode()[0])
data['Zip'] = data['Zip'].fillna(data['Zip'].mode()[0])


data['Open Date'] = pd.to_datetime(data['Open Date'], errors='coerce')


data['Open Date'] = data['Open Date'].fillna(data['Open Date'].mode()[0])


# Check for missing values
missing_data = data.isnull().sum()

# Filter out columns with missing values
missing_data = missing_data[missing_data > 0]

if missing_data.empty:
    print("No missing values in the dataset.")
else:
    print("Missing values detected:")
    print(missing_data)
```

```
No missing values in the dataset.
```

2. After plotting the scatter plot we got to know that there is no linear relationship between KW, Elev and stalls.

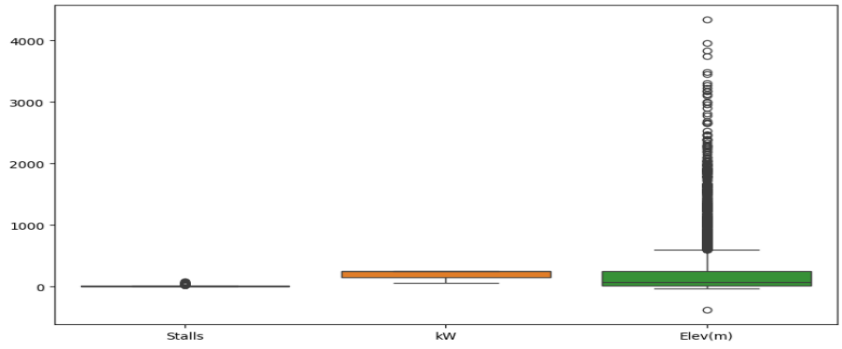




3. Outliers detection:

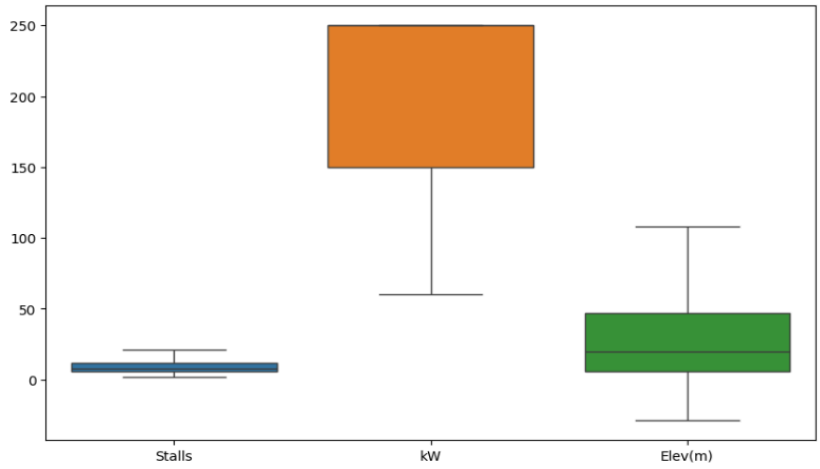
a. Using a box plot we found that there are outliers in the data. Then using the IQR method we have handled the outliers.

```
# Box plot to visualize outliers
plt.figure(figsize=(10, 6))
sns.boxplot(data=data)
plt.show()
```



b. Then performed the IQR method and handled outliers.

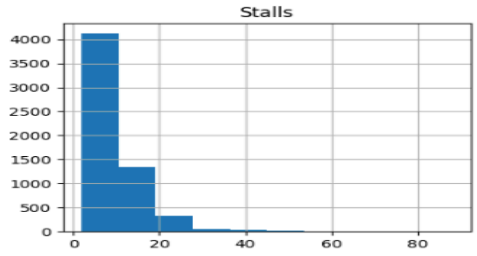
```
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np

cols_to_check = ['Stalls','kW','Elev(m)']

# Create a copy of the original DataFrame to avoid modifying it directly
data_filtered = data.copy()

for i in range(25):
  for col in cols_to_check:
      # Calculate quartiles (IQR)
      q1 = data_filtered[col].quantile(0.25)
      q3 = data_filtered[col].quantile(0.75)
      iqr = q3 - q1

      # Create a threshold for outlier detection (1.5 times IQR)
      lower_bound = q1 - 1.5 * iqr
      upper_bound = q3 + 1.5 * iqr

      # Filter outliers based on thresholds
      data_filtered = data_filtered[(data_filtered[col] >= lower_bound) & (data_filtered[col] <= upper_bound)]


plt.figure(figsize=(10, 6))
sns.boxplot(data=data_filtered) # Use data_filtered for plotting
plt.show()
```

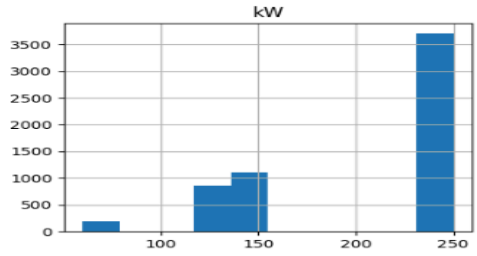


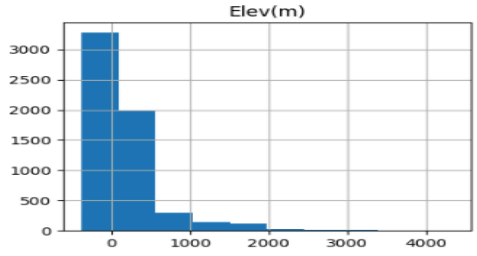
4. **Then data has been scaled using standardscaler**

```
from sklearn.preprocessing import StandardScaler

# Scaling numerical features
scaler = StandardScaler()
numerical_columns = data_filtered.select_dtypes(include=['float64', 'int64']).columns
data_filtered[numerical_columns] = scaler.fit_transform(data_filtered[numerical_columns])
```

5. After plotting histograms we found that the data is skewed mostly towards the right which may be due to the large number of locations that are having less number of stalls.



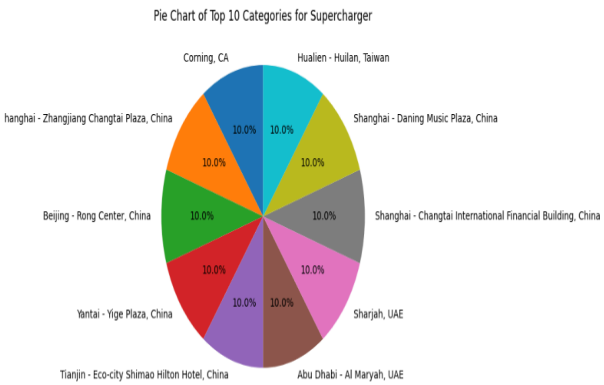


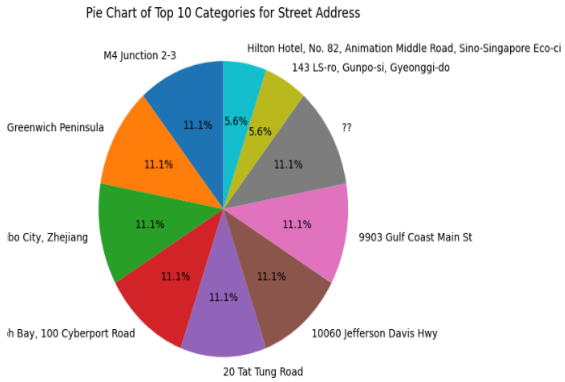


6. The histogram for Elev is also mostly skewed to the right, but to a lesser extent. This suggests a moderate for most locations, with a few having higher capacities.

7. The histogram is spread out with a wide range of power values for KW.

8. **analyzing categorical features:**





9. After seeing the pie charts we got to know that California has the highest number of charging stations, then Florida and texas.

10. Most of the charging stations were opened after 2015.

11. After seeing the correlation map we found a weak positive correlation between Stalls and KW. From that we can know that locations with more stalls have higher power capacities.