Project 3 – Group 4

EV charging stations App

Willian Ruiz

Luisa Murillo, Thaissa Champagne, Jason Johnson, Ugur Cetin

Introduction:

We discovered this project topic by exploring Kaggle. We came across the "Global EV Charging Stations" dataset (accessible at this link). We believe the future of transportation is electric, and we wanted to create a convenient website where people can easily find EV charging stations. Whether you're on a long road trip or just need a quick charge during your daily commute, our platform provides reliable, real-time data at your fingertips—including station ratings for an even better user experience.

In addition to using the dataset, we researched how other data scientists have approached similar projects, finding inspiration from four different individuals who worked with the same dataset. Their visualizations, which effectively conveyed insights from the data, served as a major influence on our approach and helped shape the visual aspects of our project.

Data organization and cleaning

The dataset consisted of a single CSV file that initially appeared clean and well-organized. However, upon closer inspection and based on our professor's suggestion, we decided to drop a column called **"Availability"** as it contained random and unclear data. We then saved and continued using the new CSV file, **"cleaned_data.csv,"** for further analysis and visualization.

```
df = pd.read_csv("cleaned_data.csv")
df.head()
```

	Station ID	Latitude	Longitude	Address	Charger Type	Cost (USD/kWh)	Distance to City (km)	Usage Stats (avg users/day)	Station Operator	Charging Capacity (kW)	Connector Types	Installation Year	Renewable Energy Source	Reviews (Rating)	Parkin Spo
0	EVS00001	-33.400998	77.974972	4826 Random Rd, City 98, Country	AC Level 2	0.27	4.95	35	EVgo	350	CCS, CHAdeMO	2013	Yes	4.0	
1	EVS00002	37.861857	-122.490299	8970 San Francisco Ave, San Francisco	DC Fast Charger	0.19	4.96	83	EVgo	350	Tesla, Type 2	2010	Yes	3.9	
2	EVS00003	13.776092	100.412776	5974 Bangkok Ave, Bangkok	AC Level 2	0.48	8.54	24	ChargePoint	50	Type 2, CCS	2019	No	3.6	
3	EVS00004	43.628250	-79.468935	6995 Toronto Ave, Toronto	AC Level 1	0.41	13.28	70	Greenlots	350	Type 2	2010	Yes	4.2	
4	EVS00005	19.119865	72.913368	5704 Mumbai Ave, Mumbai	AC Level 2	0.11	9.76	19	EVgo	350	CCS	2015	Yes	3.7	

As we moved forward with creating the website, especially the interactive map, we noticed that some addresses were missing, while others were inaccurate or fake. This affected the quality of our visualizations, making the map unreliable. To improve the dataset, one of our team members took the initiative to further clean the data, ensuring better accuracy and consistency.

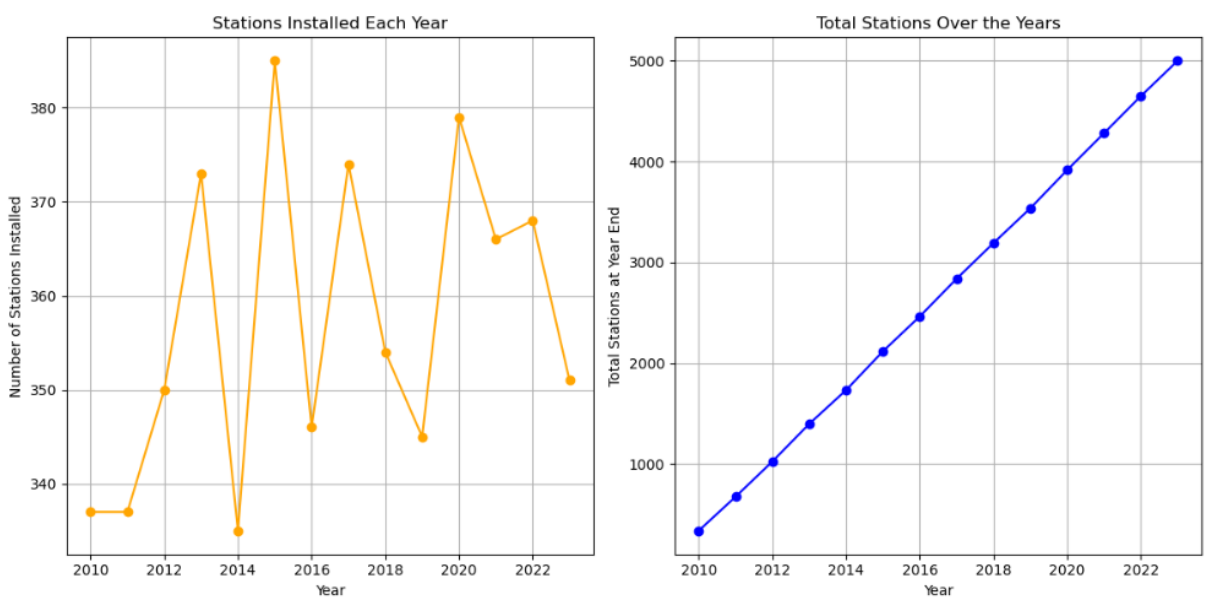
Then, we created a second file in Jupyter Notebook called *Test Visualizations* to explore and generate some of the graphs we wanted to include on the website.

We started by analyzing the number of charging stations installed per year to understand growth trends over time.

```
#RAW SQL
query = text("""SELECT "Installation Year", COUNT(*) AS station_count
    FROM ev_charging_stations
    GROUP BY "Installation Year"
    ORDER BY "Installation Year";
""")

df1 = pd.read_sql(query, con=conn)
df1.head(10)
```

	Installation Year	station_count
0	2010	337
1	2011	337
2	2012	350
3	2013	373
4	2014	335
5	2015	385
6	2016	346
7	2017	374
8	2018	354
9	2019	345



We first created the front end adding the appy.

Limits & Bias of data

The bias and limitations can play a factor in introducing raw data along with visualizations. One of the limitations that we ran into as a group is during the data cleaning process, there are some addresses that are fictitious and do not present a physical location. Altogether, there are 500 rows of raw data that was included in the data cleaning that resulted in addresses with the street name “Random Rd” which can cause an import output of incorrect data as far as actual physical locations for EV Chargers. This data alone can result in making it possible for consumers to go look for these actual locations to only find out that the locations don’t exist. Running into issues such as this are beyond our control, but we use this information get into specifics as to why it’s important clean, analyzed and visualize raw data in this form so that we can come to a conclusion as to why the issue occurred and what we can and have done to make sure that it didn’t interfere with the common goal. The data with Randon Road as the street name will come as inconclusive as there may be physical location, but there is not a definite answer on the location