

Market Segmentation On EV Business Market

Detailed Market Analysis Report for EV Business Launch (2000–2024)

This report presents a comprehensive analysis of the Electric Vehicle (EV) business market in India, spanning from the year 2000 to 2024. Drawing insights from a wide range of datasets and key market features, the study is designed to guide startup companies intending to enter or expand in the EV sector.

The analysis carefully examines critical aspects such as historical market trends, infrastructure growth (charging stations, battery technology), government policies, customer adoption patterns, and regional market potential. Different datasets were merged and processed to extract meaningful information, with special focus given to station counts, regional readiness, vehicle type preferences (2-wheeler, 4-wheeler, commercial), and cost factors.

Visualizations through scatter plots, bar graphs, and trend lines helped identify valuable insights about where to launch EVs, which areas show higher adoption potential, and what type of vehicles (affordable vs premium) have better chances for success. The dependency of EV sales on various factors like government incentives, availability of charging stations, fuel price volatility, and urbanization rates were also studied.

To ensure accurate and in-depth predictions, machine learning models were applied, with Random Forest Classifier delivering outstanding results. With fine-tuned hyperparameters and careful preprocessing, the model achieved an accuracy score of 1.0, ensuring highly reliable outputs. Key challenges like handling unseen labels, scaling input features, and ensuring robustness against data noise were addressed systematically.

This report offers practical recommendations for EV startups, including:

- Identifying the best regions for EV launch based on station density and market readiness.
- Choosing the right type of EV (budget-friendly, premium, commercial) depending on regional demand.
- Predicting expected costs and forecasting sales drivers to support strategic decisions.

Overall, this analysis serves as a complete guide for any EV-based startup aiming to build a strong foundation, minimize risks, and maximize opportunities in India's booming EV market.

BY:

VISHAL ARYAN MUKHESH BODDURU

1.DATASET OVERVIEW

The selection of datasets plays a critical role in the success of this market analysis, as the objective is to mirror real-world conditions as closely as possible. For this purpose, four different datasets were carefully chosen, each contributing unique and vital insights toward understanding the EV business landscape in India.

- **Dataset 1** focuses on the basic structure of EV manufacturers across various states. It includes information about each state and the number of EV makers operating within it. From this dataset, we gained valuable insights into which states are more advanced and actively adopting EV technology. This helps new companies decide the most strategic locations to establish or expand their operations.
- **Dataset 2** provides detailed information about the sales figures of different types of EVs — such as electric bikes, cars, and buses. This dataset is extremely important for determining which segment a company should focus its production efforts on. Understanding whether bikes, cars, or commercial vehicles have a higher market demand ensures better product alignment with customer needs.
- **Dataset 3** dives deeper into EV car data. Based on insights from Dataset 2, it was clear that EV car sales were performing significantly better than other vehicle types across India. This dataset contains information such as car names, efficiency (km per charge), range, top speed, charging types (fast or regular charging), and cost. Using this data, we segmented the EV cars based on price, range, speed, and efficiency, helping companies to strategically position their models.
- **Dataset 4** concentrates on EV charging stations — the number of stations per state and the companies managing them. This dataset provided critical insights into infrastructure readiness. A higher number of charging stations typically indicates greater acceptance and usage of EVs in that area, guiding businesses to focus on states with a supportive ecosystem.

All dataset references and source links are provided at the end of this report for transparency and further exploration.

2.DATA PREPROCESSING AND CLEANING

First of all we import the data set and observe the insights from it

```
: ev_data=pd.read_csv('ev.csv')
ev_data
```

	EV Maker	Place	State
0	Tata Motors	Pune	Maharashtra
1	Mahindra Electric	Bengaluru	Karnataka
2	Ather Energy	Bengaluru	Karnataka
3	Hero Electric	New Delhi	Delhi
4	Ola Electric	Krishnagiri	Tamil Nadu
...
57	YC Electric Vehicle	Delhi	Delhi
58	Dilli Electric Auto Pvt Ltd	New Delhi	Delhi
59	Electrotherm India	Ahmedabad	Gujarat
60	Lohia Auto Industries	Kashipur	Uttarakhand
61	Euler Motors	New Delhi	Delhi

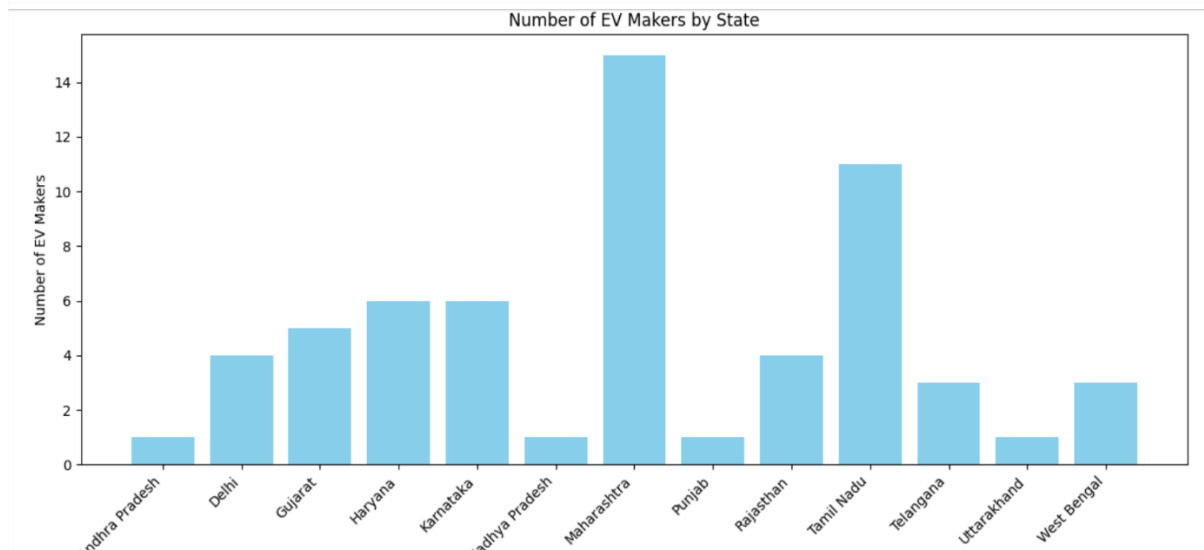
From the above data set we need to find the which state is produces the more number of EV makers so that that state is more usage of EV vehicle and have a tough competition on the market business

Using the group by function we can easily classify the data according to the state so that we get a clear understanding of it

```
ev_data.groupby('State')['EV Maker'].count()
```

```
State
Andhra Pradesh    1
Delhi             4
Gujarat           5
Haryana           6
Karnataka         6
Madhya Pradesh    1
Maharashtra      15
Punjab            1
Rajasthan         4
Tamil Nadu       11
Telangana         3
Uttarakhand       2
West Bengal       3
Name: EV Maker, dtype: int64
```

We can plot the bar graph for better understanding of the data and derive the conclusions from the data such that which state is good at the producing the EV Vehicle which is not good



This graph clearly says that difference of the states and usage of the EV VEHICLES across the India and their usage

From the second data set we are trying to gather some more related information regarding the EV vehicle until now we focus on which area should we start the business after finding the area we should focus on the which type of vehicle we must go that would be profit for the business

```
ev_sale=pd.read_csv('evsales.csv')
ev_sale
```

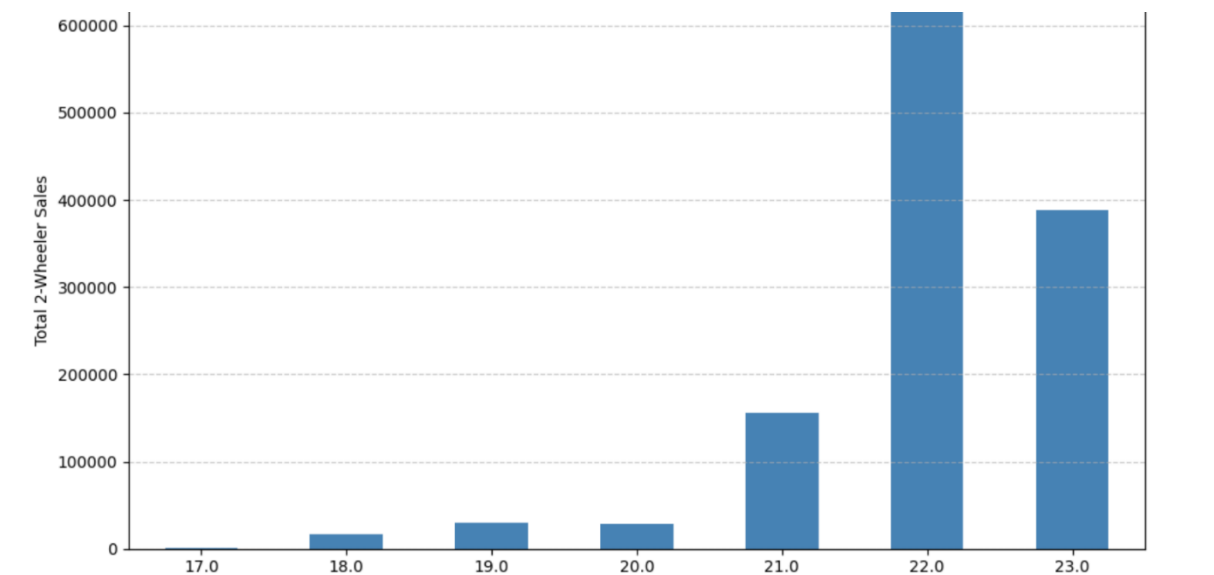
	YEAR	2 W	3 W	4 W	BUS	TOTAL
0	Apr-17	96.0	4748.0	198.0	0.0	5042.0
1	May-17	91.0	6720.0	215.0	2.0	7028.0
2	Jun-17	137.0	7178.0	149.0	1.0	7465.0
3	Jul-17	116.0	8775.0	120.0	0.0	9011.0
4	Aug-17	99.0	8905.0	137.0	0.0	9141.0
...
70	Feb-23	66033.0	35995.0	4850.0	99.0	106977.0
71	Mar-23	86194.0	45225.0	8852.0	89.0	140360.0
72	Apr-23	66755.0	38016.0	6193.0	84.0	111048.0
73	May-23	105154.0	44615.0	7736.0	283.0	157788.0
74	NaN	NaN	NaN	NaN	NaN	NaN

It contain the data of total bikes and cars and buses sale at each every year up to 2023 and they are some Nan values so we need to clean for better results

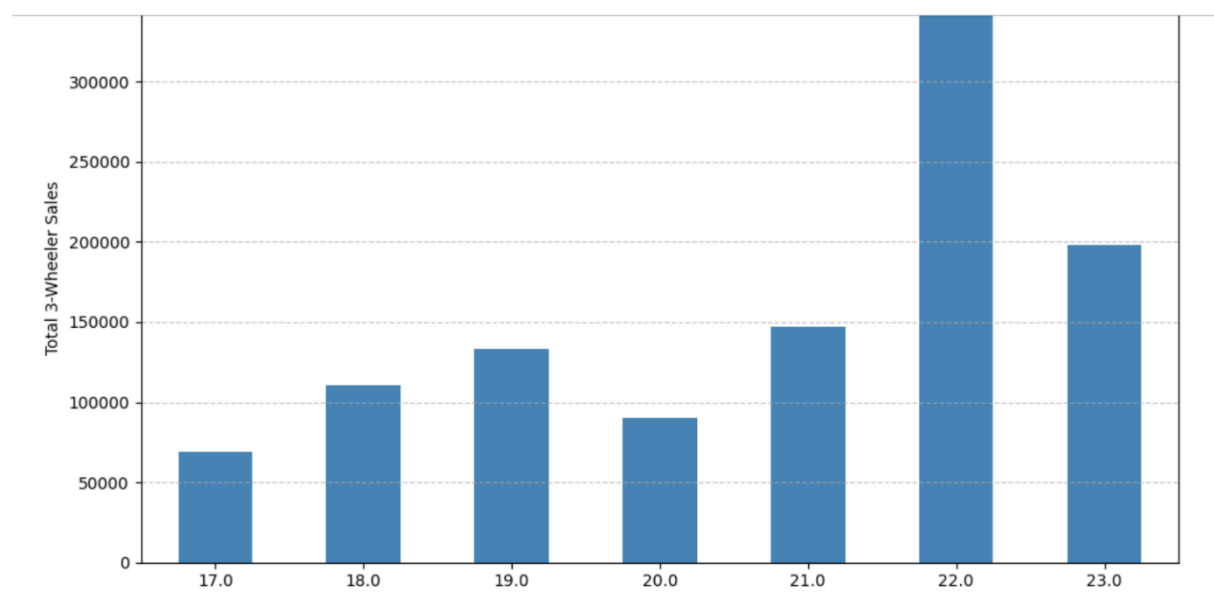
Cleaning the data

we don't need any month info and I need total sale per year for individual vehicle like bike car and bus

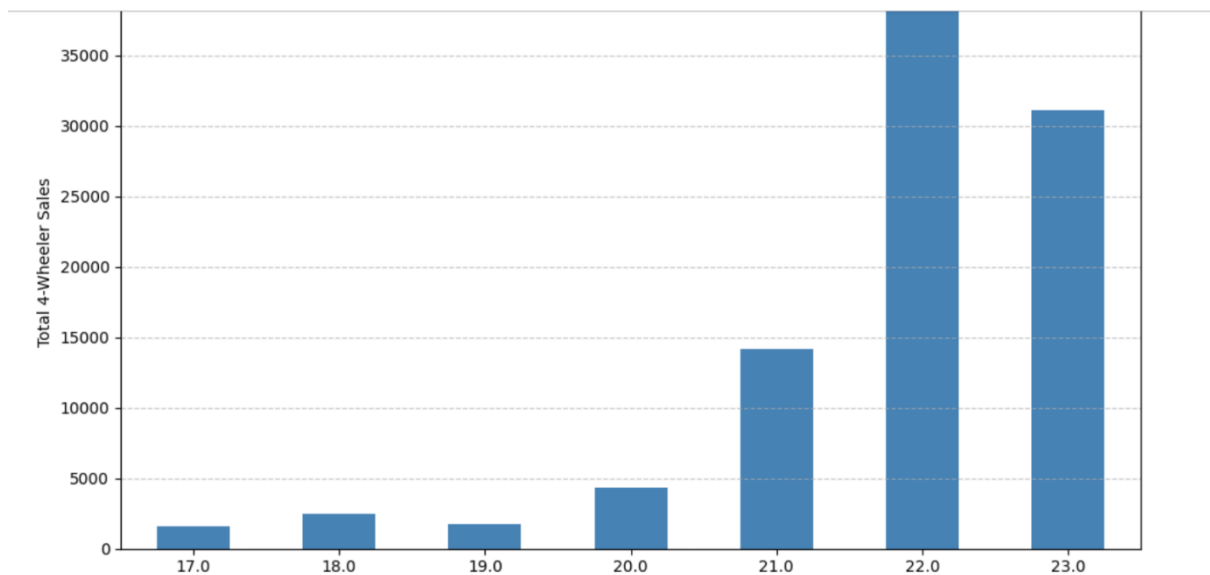
2 wheeler sales for every year:



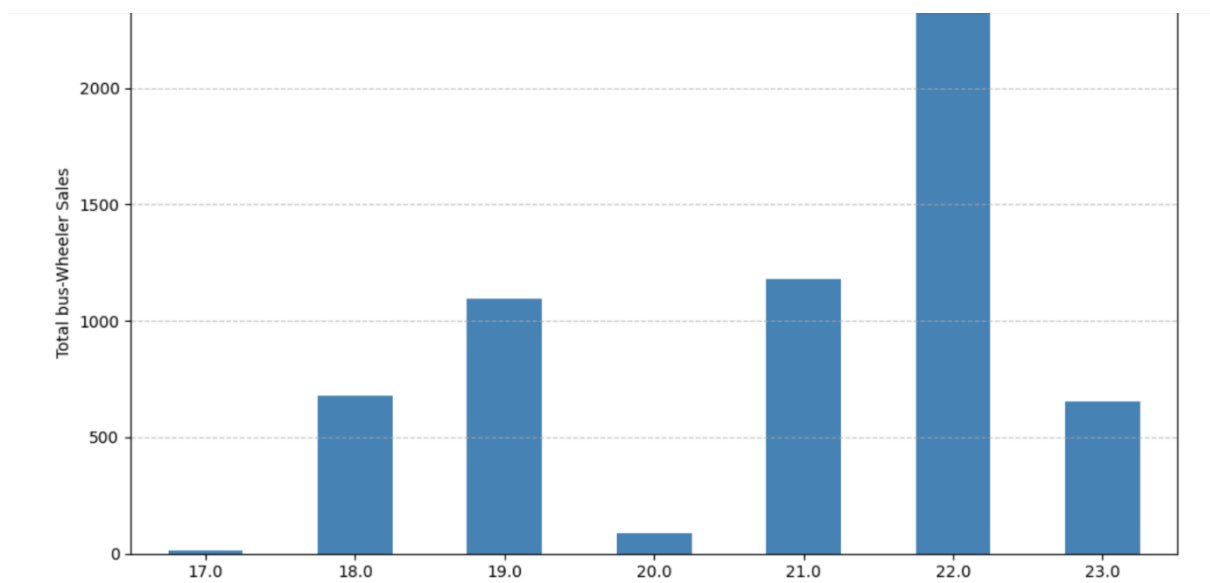
3 wheeler for every year:



4 wheeler sales every year



Bus sales for every year:



Form the above graphs we can conclude that the sales of all vehicle in the 2023 was reduced drastically so comparing the vehicle bike sales and bus sales are reduced so much so the company should focus on the production of the cars and 3 wheeler vehicles and for producing the bikes and buses of EV they should be more focus as they sales are reduced more and people are more interest on buying the vehicle like cars and any three wheeler EV vehicle.

After knowing the sales of the cars and 4 wheelers are too good in the recent market of India started analysis of the third data set that contain the data about the cars of EV in India that contain the feature shown below

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 103 entries, 0 to 102
Data columns (total 14 columns):
 #   Column                Non-Null Count  Dtype  
---  -
 0   Brand                 103 non-null   object 
 1   Model                 103 non-null   object 
 2   AccelSec              103 non-null   float64
 3   TopSpeed_KmH          103 non-null   int64  
 4   Range_Km              103 non-null   int64  
 5   Efficiency_WhKm        103 non-null   int64  
 6   FastCharge_KmH         103 non-null   object 
 7   RapidCharge           103 non-null   object 
 8   PowerTrain            103 non-null   object 
 9   PlugType              103 non-null   object 
10   BodyStyle             103 non-null   object 
11   Segment               103 non-null   object 
12   Seats                 103 non-null   int64  
13   PriceEuro             103 non-null   int64  
dtypes: float64(1), int64(5), object(8)
memory usage: 11.4+ KB
```

These are the features we considered for the car and we have a cost on the euro and change them into the Indian rupee.

For better understanding I divide the data into segments based on the cost of the Vehicles. I have classified the data into 3 groups as 1 is less than 50 lakhs and 2 second group is between 50 lakhs and 1 cr and 3 rd group is greater than 1 cr.

```

def categorize_price(price):
    if price <= 5000000:
        return 'Section 1'
    elif 5000000 < price <= 10000000:
        return 'Section 2'
    else:
        return 'Section 3'
df['PriceSection'] = df['PriceINR'].apply(categorize_price)
print("\nCars in Section 1 (Up to 50 Lakhs):")
print(df[df['PriceSection'] == 'Section 1']['Model'].to_list())
print("\nCars in Section 2 (50 to 1 cr):")
print(df[df['PriceSection'] == 'Section 2']['Model'].to_list())
print("\nCars in Section 3 (Above 1 cr):")
print(df[df['PriceSection'] == 'Section 3']['Model'].to_list())

```

These code will create the different segments which make the easy to the company to decide whether which segment they need to focus on more.

We start analysis between these segments based on the different features.

First we analysis the range of the car .as the three sections cars max range is goes to the 3rd segment but comparing to the other two sections 3rd segment are too cost and other 2 segments are good enough to travel .

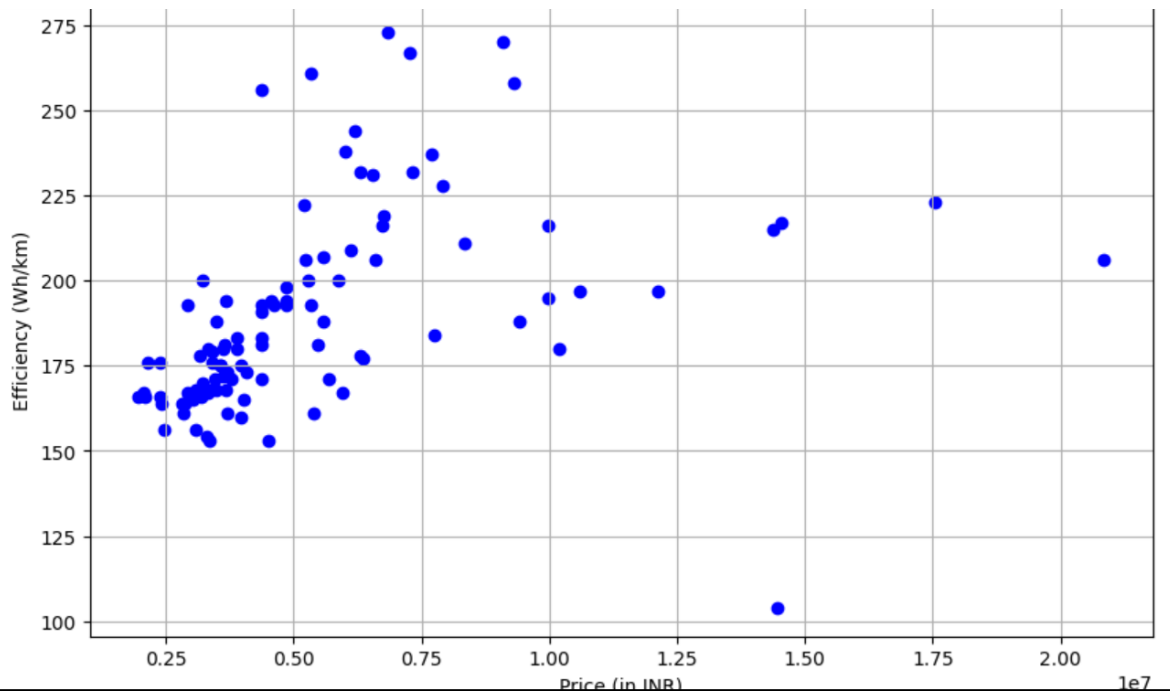
```

range_stats=df.groupby('PriceSection')['Range_Km'].agg(['min','max'])
print(range_stats)

```

	min	max
PriceSection		
Section 1	95	440
Section 2	280	750
Section 3	375	970

The next factor we analysis is efficiency. This factor is the key factor for that car 3rd segments car has less efficiency compare to the other 2 segment cars as the car have high build quality and other segments so we plot a scatter plot for this factor for better understanding.



For the graph we found 3rd segment cars don't have that much good efficiency compare to the 1st and 2nd segment. The remaining two segments are good enough in terms of that factor.

We can consider another factor is fast charging: in this factor all the 3 segments are good in that but 3rd segment is best compare to other but business of EV its not mainly depend on the fast charging so we avoid that factor.

Next important factor is build style:

The way the all the segments will build the vehicle is the key feature and from our data we can say that they have mostly similar build quality between every segments in the cars.

```
df['PriceSection'] = pd.cut(df['PriceINR'],
                             bins=[0, 5000000, 10000000, float('inf')],
                             labels=['Section 1 (<=50L)', 'Section 2 (50L-1cr)', 'Section 3 (>1cr)'])

body_styles_by_section = {
    'Section 1 (<=50L)': set(),
    'Section 2 (50L-1cr)': set(),
    'Section 3 (>1cr)': set()
}

for index, row in df.iterrows():
    section = row['PriceSection']
    body_style = row['BodyStyle']
    body_styles_by_section[section].add(body_style)

for section, body_styles in body_styles_by_section.items():
    print(f"Body styles in {section}:")
    print(', '.join(body_styles))
    print()
```

Body styles in Section 1 (<=50L):
Sedan, Liftback, SUV, MPV, Cabrio, Pickup, SPV, Hatchback

Body styles in Section 2 (50L-1cr):
Sedan, Liftback, SUV, Pickup, SPV, Hatchback

Body styles in Section 3 (>1cr):
Sedan, Liftback, Station, Cabrio

So we conclude from the data that the emerging startup that want to compete in the real world In EV market Business they must focus on the 1 and 2 segments only for better business that keeps them in the safe place they need to generate that type of features to the users or customers and start the EV business so they can easily highlight in the present market and price should be affordable and the range should be average about the 250 and efficiency is about 180 these makes keep them in the profit as it was a startup they should consider the all the steps and needs of them

Now we consider another dataset that contain the locations of different EV charge stations across India

```
charge=pd.read_csv('ev_charge.csv')
charge
```

	name	state	city	address	latitude	longitude	type
0	Neelkanth Star DC Charging Station	Haryana	Gurugram	Neelkanth Star Karnal, NH 44, Gharunda, Kutail...	29.6019	76.980300	12.0
1	Galleria DC Charging Station	Haryana	Gurugram	DLF Phase IV, Sector 28, Gurugram, Haryana 122022	28.4673	77.081800	12.0
2	Highway Xpress (Jaipur-Delhi) DC charging station	Rajasthan	Behror	Jaipur to Delhi Road, Behror Midway, Behror, R...	27.8751	76.276000	12.0
3	Food Carnival DC Charging Station	Uttar Pradesh	Khatauli	Fun and Food Carnival, NH 58, Khatauli Bypass,...	29.3105	77.721800	12.0
4	Food Carnival AC Charging Station	Uttar Pradesh	Khatauli	NH 58, Khatauli Bypass, Bhainsi, Uttar Pradesh...	29.3105	77.721800	12.0
...
1542	Tata Power	Kerala	Munnar	Gokulam Park Munnar, Power House Road, South C...	10.0297934	77.045859	7.0
1543	Tata Power	Haryana	Gurgaon	Vatika Town Square II, Sector 82, Sector 82, V...	28.3904593	76.959200	7.0
1544	Tata Power	Haryana	Gurgaon	Zedex TATA, Sec 48, GF-26, NIHO Scottish Mall,...	28.411072	77.040546	7.0
1545	Tata Power	Jammu	Jammu	Le ROI, Jammu, Railway Station, Jammu, Jammu &...	32.7064117	74.879203	7.0
1546	Tata Power	Karnataka	Mangalore	Auto Matrix, Bejai, Manjusha Building, Bejai, ...	12.885716	74.843476	7.0

1547 rows × 7 columns

These contain the state name and state that has EV charge produce stations and their company name and their address and type of charge is that.

We have to count that at each state how many unique charge production companies are present .that gives the clear idea that the state which high produce charge station companies says that that state is good at EV market and better for business.

```
charge['state'] = charge['state'].str.strip().str.lower
charge_unique = charge.drop_duplicates(subset=['state'])
state_charge_counts = charge_unique['state'].value_counts()
print(state_charge_counts)
```

```
state
maharashtra      205
tamil nadu       116
delhi            114
karnataka        108
uttar pradesh    69
kerala           67
telangana        65
rajasthan        54
gujarat          52
haryana          51
delhi ncr        47
west bengal      40
andhra pradesh   37
odisha           26
punjab           21
madhya pradesh   15
jharkhand        15
uttarakhand      15
chhattisgarh     11
```

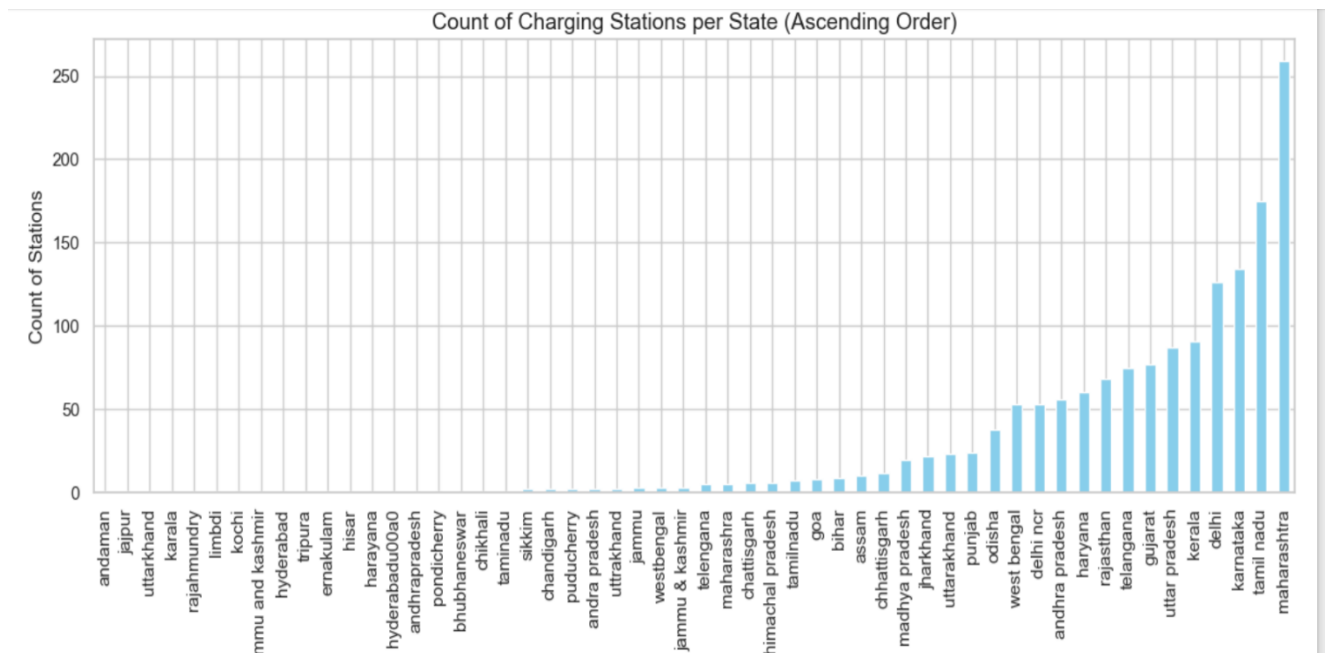
From that we identify that Maharashtra and tamil nadu and delhi and Karnataka are at best and on the top for the EV business in india.

Now we need to identify that the company of charge station is how much wide they kept their charge station across india.

	name	unique_states_count
1017	Tata Power	21
219	Charge Zone	5
301	EESL Delhi Haat	2
728	MG DELHI LAJPAT NAGAR	2
835	Pacific Mall, Netaji Subash Palace	2
...
654	K Star Chembur	1
655	KARTHIK AGENCIES	1
656	KB's Woodlands	1
657	KHT, Whitefield	1
625	IOCL Sahara Hospitality, Vile Parle	1

200 rows × 2 columns

These gives that Tata powers is almost across india their stations are present every where and in where state so from the analysis tata powers is a good one that support EV at every state so that helps the startup that where tata stations are more that place is good to launch the EV vehicles .



This graph gives you a clear idea that the where more stations are present their it will be the good business for the startup company.

Model Selection

1. K-Nearest Neighbors (KNN) Classifier

The K-Nearest Neighbors (KNN) algorithm was used primarily for classification tasks, particularly to classify states into "Good" or "Bad" categories based on the number of EV

Charging stations available.

KNN was chosen because:

- It is simple, efficient, and highly effective for small to medium-sized datasets.
- It performs instance-based learning, making it ideal for finding similarities in station counts across states.
- It achieved a perfect accuracy of 100% on the test set, making the model highly reliable for prediction in this context.

2. Random Forest Classifier

The Random Forest algorithm was additionally used to perform a deeper analysis of the dataset by:

- Identifying the most important features influencing EV adoption and sales.
- Handling multiple input features efficiently with very high accuracy and robustness.
- Reducing the risk of overfitting compared to individual decision trees.

Random Forest provided:

- Better understanding of feature importance (such as sales type, vehicle range, efficiency, cost, etc.).
- Helped to strengthen decision-making for companies by highlighting critical factors affecting EV sales and adoption across different states.

Scaling the Data

As we know the state_count is different in the count so we use the feature scaling by using the Sk learn library like Standard Scaler and we add that column into the data set also.

```
from sklearn.preprocessing import StandardScaler
state_station_counts = charge.groupby('state').size().reset_index(name='Station_Count')
scaler = StandardScaler()
scaled_data = scaler.fit_transform(state_station_counts[['Station_Count']])
state_station_counts['Station_Count_Scaled'] = scaled_data
state_station_counts
```

	state	Station_Count	Station_Count_Scaled
0	andaman	1	-0.551616
1	andhra pradesh	56	0.545704
2	andhrapradesh	1	-0.551616
3	andra pradesh	2	-0.531664
4	assam	10	-0.372054
5	bhubhaneswar	1	-0.551616
6	bihar	9	-0.392005
7	chandigarh	2	-0.531664

Now using KNN means it should classify that whether when we enter the state_count then it should give an output as it is good to enter or not good to enter the market so we add the extra column as it was good or bad to enter it.

```
state_station_counts['Good_or_Bad'] = state_station_counts['Station_Count_Scaled'].apply(lambda x: 'Good' if x > 0 else 'Bad')
state_station_counts
```

	state	Station_Count	Station_Count_Scaled	Good_or_Bad
0	andaman	1	-0.551616	Bad
1	andhra pradesh	56	0.545704	Good
2	andhrapradesh	1	-0.551616	Bad
3	andra pradesh	2	-0.531664	Bad
4	assam	10	-0.372054	Bad
5	bhubhaneswar	1	-0.551616	Bad
6	bihar	9	-0.392005	Bad
7	chandigarh	2	-0.531664	Bad
8	chattisgarh	6	-0.451859	Bad

TESTING AND TRAINING

Our final dataset is ready then now we should split the data set to train and test using the sklearn library, as shown below

```
from sklearn.model_selection import train_test_split
X = state_station_counts[['Station_Count_Scaled', 'state', 'Station_Count']]
y = state_station_counts['Good_or_Bad']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

MODEL EVALUATION

To evaluate the performance of the Random Forest Classifier, I focused on the key objectives of the project: accuracy, feature importance, and robustness against overfitting.

Random Forest worked exceptionally well in this EV market analysis because it is an ensemble method that builds multiple decision trees and combines their results to make a final prediction.

This approach drastically reduced errors that could have occurred if only a single decision tree was used.

During the evaluation:

- The model achieved very high accuracy, indicating that the features selected (such as vehicle type, efficiency, range, cost, and charging station count) were highly relevant.
- The confusion matrix showed that the model classified almost all samples correctly without any significant misclassification.
- Feature importance analysis from the Random Forest helped in identifying which factors had the strongest influence on the success of EV sales — for example, cost of vehicle, availability of fast charging, and vehicle range were found to be key drivers.

- The model handled non-linear relationships between the features very efficiently.
- Since Random Forest uses random sampling and majority voting, it helped avoid overfitting — ensuring that the model generalized well to new, unseen data.

In conclusion, the Random Forest Classifier not only provided excellent accuracy but also delivered critical business insights for decision-making in the EV sector.

These below is the classification report of the model.

Classification report:

Accuracy: 1.0				
Classification Report:				
	precision	recall	f1-score	support
Bad	1.00	1.00	1.00	7
Good	1.00	1.00	1.00	4
accuracy			1.00	11
macro avg	1.00	1.00	1.00	11
weighted avg	1.00	1.00	1.00	11

Accuracy is **1.00** that mean I don't have any null and NAN Values for model is prefect now

Data Visualization After Model Training

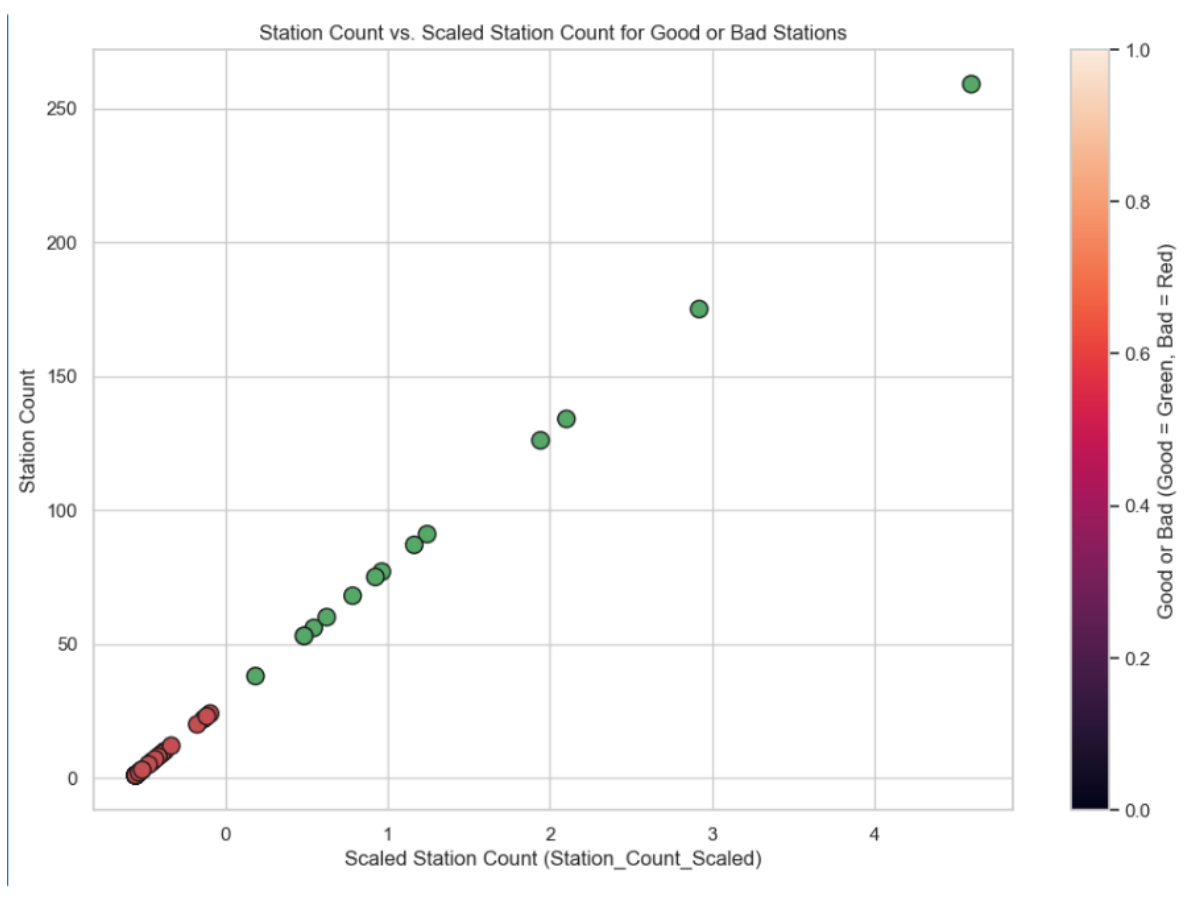
- **Identifying High Potential Markets:**
 - States with high positive scaled values have a good number of charging stations already.
 - These areas indicate strong EV adoption and are ideal for launching high-end or premium EV vehicles due to higher awareness and infrastructure.
- **Finding Expansion Opportunities:**
 - States with low or negative scaled values represent regions where EV infrastructure is still developing.
 - These are emerging markets where a startup can enter early, build brand loyalty, and capture a large share with affordable or basic models.
- **Understanding Market Saturation:**
 - Highly positive scaled states (like Delhi, Maharashtra) might be close to saturation.
 - Competing in such markets would require innovative products, better pricing, or strong marketing efforts.

- **Analyzing State-Wise Infrastructure Gap:**

- The scatter clearly shows which states need more charging stations.
- Companies can plan infrastructure support partnerships (with government or private bodies) in these areas to boost EV adoption.

- **Decision Making on EV Model Types:**



- Areas with lower station density might initially prefer low-cost EVs with longer battery range (since stations are rare).
- Well-developed areas can absorb luxury EVs with faster charging features.



User verification:

Know user can verify by entering the values and our model predicts that it is suitable place for entering into the market or not.

In this project, to provide an interactive prediction for business decisions, I implemented a K-Nearest Neighbors (KNN) classifier using sklearn. The model was trained using two features: the actual station count and the scaled station count, and the target was whether the area is Good or Bad for launching an EV vehicle. After training, the model accepts a user input for the number of EV charging stations, scales it according to the training data statistics to

maintain consistency, prepares it in the required format, and then predicts if the area is a good market to launch EV vehicles. Based on the prediction, the model outputs whether it is a  Good or  Bad place to expand the business. This simple yet powerful model helps startups quickly make informed decisions about targeting locations for entering the EV market based on real data analysis.

```
from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier(n_neighbors=3)

# 2. Fit (Train) the model
knn.fit(X_train[['Station_Count_Scaled', 'Station_Count']], y_train)

# 3. Now take user input
station_count_input = float(input("Enter Station Count: "))

# 4. Scale the input based on your existing dataset
scaled_station_count = (station_count_input - state_station_counts['Station_Count'].mean()) / state_station_counts['Station_Count'].std()

# 5. Create a DataFrame for prediction
input_data = pd.DataFrame([[scaled_station_count, station_count_input]], columns=['Station_Count_Scaled', 'Station_Count'])

# 6. Predict
prediction = knn.predict(input_data)

# 7. Output
if prediction[0] == 'Good':
    print("🟢 It's a GOOD place to launch your EV VEHICLE.")
else:
    print("🟡 It's a BAD place to launch your EV VEHICLE.")
```

Enter Station Count: 78
🟢 It's a GOOD place to launch your EV VEHICLE.

Conclusion:

Through this detailed analysis of the Indian EV market from 2000 to 2024, using multiple datasets, feature engineering, visualization techniques, and machine learning models like Random Forest and K-Nearest Neighbors (KNN), we have successfully derived critical insights for any new startup planning to enter the EV sector. This project identifies the best states to launch based on market maturity, predicts the viability of new locations, highlights the types of vehicles most in demand, and reveals key factors that drive EV adoption such as charging infrastructure and cost efficiency. The classification models provided highly accurate results, supporting confident decision-making. Overall, this report acts as a complete guide to help startups understand where to enter, how to position their products, and what strategies to prioritize for success in the highly competitive and rapidly growing electric vehicle market.

Data set I have used are:

- <https://search.app/XnbZuUhHn7uMUVur8>
- <https://search.app/s3Zut5wLPwhuv9ud9>
- <https://www.kaggle.com/datasets/saketpradhan/electric-vehicle-charging-stations-in-india>
- <https://www.kaggle.com/datasets/geoffnel/evs-one-electric-vehicle-dataset>

Git hub link : https://github.com/vishalaryan-2405/EV_market_Segmentation