

## **CHAPTER 3**

### **RESEARCH METHODOLOGY**

#### **3.1 Introduction**

This paper examines and reviews historical data from various types of electric vehicle markets, considering such major driving factors for the EV market: technology advancement, policy innovation, and consumer preference. This paper will apply ARIMA/ SARIMA to conduct appropriate forecasting of consumer preference for various types of vehicles with techniques like logistic regression, Random Forests, and SVM. These will provide useful insight into the trends of future development of various types of electric vehicles.

#### **3.2 Research Framework**

This research framework includes the following steps:

1. Problem Definition and Literature Review
2. Data Collection: Retrieve data from Kaggle using specific keywords.
3. Data Pre-processing: Cleaning and preparing data for further analysis.
4. Model Building: Building and training model.
5. Model Evaluation: Evaluation model performance using historical trends.
6. Trend Analysing: Exploring the trend of total annual sales of electric vehicles.

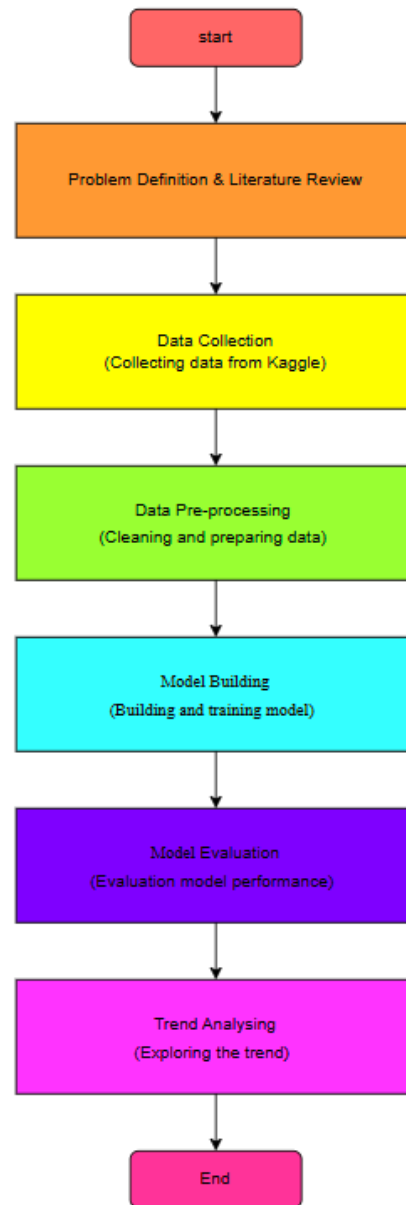


Figure 3.1 General Project Flow

### 3.3 Problem Formulation

This study will analyze the historical development trends of various electric vehicle (EV) markets and consider the main factors that affect the development trends, so as to make accurate predictions on the future development trends of the market. In order to ensure the accuracy of the prediction of the future development trends of the market, this study needs to solve the following two main problems:

- Detailed analysis of the historical development trends of various electric vehicle (EV) markets, and then derive the general development trends;
- Substitute historical data to verify the accuracy of the prediction results, and further revise and improve the prediction results by considering the influence of factors such as technology and policies.

### 3.4 Data Collection

The data required for this study comes from Kaggle, and through analysis and comparison, effective data cleaning is performed to obtain valid data.

The keywords used to capture the data are mainly:

- “Development trend of various types of electric vehicles”
- “Historical sales data of various types of electric vehicles in the world”
- “Historical sales data of the global automobile market”

The captured data mainly includes the following ranges:

- The time range covers the historical sales data of the electric vehicle market from 2010 to 2024
- Detailed information includes detailed historical statistical data of various types of electric vehicles, including car brands, car types, and locations;

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	VIN (1-10)	County	City	State	Postal Cod	Model Yea	Make	Model	Electric Vel	Clean Alter	Electric Rai	Base MSRP	Legislative	DOL Vehicle ID	Vehicle Loc	Electric Uti	2020 Census	Tract
2	5Y1YGDEE	King	Seattle	WA	98122	2020	TESLA	MODEL Y	Battery Ele	Clean Alter	291	0	37	125701579	POINT (-11	CITY OF SE	5.3E+10	
3	7SA1YGDDEE	Snohomish	Bothell	WA	98021	2023	TESLA	MODEL Y	Battery Ele	Eligibility u	0	0	1	244285107	POINT (-11	PUGET SOI	5.31E+10	
4	5YJSA1E4	King	Seattle	WA	98109	2019	TESLA	MODEL S	Battery Ele	Clean Alter	270	0	36	156773144	POINT (-11	CITY OF SE	5.3E+10	
5	5YJSA1E2	King	Issaquah	WA	98027	2016	TESLA	MODEL S	Battery Ele	Clean Alter	210	0	5	165103011	POINT (-11	PUGET SOI	5.3E+10	
6	5Y1YGDEE	Kitsap	Suquamish	WA	98392	2021	TESLA	MODEL Y	Battery Ele	Eligibility u	0	0	23	205138552	POINT (-11	PUGET SOI	5.3E+10	
7	3FA6P0SUI	Thurston	Yelm	WA	98597	2017	FORD	FUSION	Plug-in Hy	Not eligibl	21	0	2	122057736	POINT (-11	PUGET SOI	5.31E+10	
8	1N4A20CF	Yakima	Yakima	WA	98903	2013	NISSAN	LEAF	Battery Ele	Clean Alter	75	0	14	150126840	POINT (-11	PACIFICOR	5.31E+10	
9	KNAGV4L	Snohomish	Bothell	WA	98012	2018	KIA	OPTIMA	Plug-in Hy	Not eligibl	29	0	1	290605598	POINT (-11	PUGET SOI	5.31E+10	
10	1N4A20CF	Kitsap	Port Orcha	WA	98366	2015	NISSAN	LEAF	Battery Ele	Clean Alter	84	0	26	137322111	POINT (-11	PUGET SOI	5.3E+10	
11	5UXTA6C0	King	Auburn	WA	98001	2022	BMW	X5	Plug-in Hy	Clean Alter	30	0	47	240226332	POINT (-11	PUGET SOI	5.3E+10	
12	5Y1YGDEE	King	Seattle	WA	98144	2020	TESLA	MODEL Y	Battery Ele	Clean Alter	291	0	37	113323024	POINT (-11	CITY OF SE	5.3E+10	
13	WB8Y8P9C	Kitsap	Bainbridge	WA	98110	2019	BMW	I3	Plug-in Hy	Clean Alter	126	0	23	228846642	POINT (-11	PUGET SOI	5.3E+10	
14	1G1FZ6S0	Yakima	Yakima	WA	98908	2021	CHEVROLET	BOLT EV	Battery Ele	Eligibility u	0	0	14	156686106	POINT (-11	PACIFICOR	5.31E+10	
15	WA1E2AF	Snohomish	Lynnwood	WA	98036	2021	AUDI	Q5 E	Plug-in Hy	Not eligibl	18	0	1	168371122	POINT (-11	PUGET SOI	5.31E+10	
16	1N4A20CF	King	Seattle	WA	98119	2015	NISSAN	LEAF	Battery Ele	Clean Alter	84	0	36	126304132	POINT (-11	CITY OF SE	5.3E+10	
17	1N4A20CF	King	Seattle	WA	98107	2013	NISSAN	LEAF	Battery Ele	Clean Alter	75	0	43	100938848	POINT (-11	CITY OF SE	5.3E+10	
18	1N4A20CF	Snohomish	Lynnwood	WA	98087	2013	NISSAN	LEAF	Battery Ele	Clean Alter	75	0	21	139800496	POINT (-11	PUGET SOI	5.31E+10	
19	1N4B20CP	Snohomish	Bothell	WA	98021	2017	NISSAN	LEAF	Battery Ele	Clean Alter	107	0	1	348979466	POINT (-11	PUGET SOI	5.31E+10	

Figure 3.2 Initial Dataset

	A	B	C	D	E	F	G	H	I	J	K
1	region	category	parameter	mode	powertrain	year	unit	value			
2	Australia	Historical	EV sales sh	Cars	EV	2011	percent	0.0065			
3	Australia	Historical	EV stock sh	Cars	EV	2011	percent	0.00039			
4	Australia	Historical	EV sales	Cars	BEV	2011	Vehicles	49			
5	Australia	Historical	EV stock	Cars	BEV	2011	Vehicles	49			
6	Australia	Historical	EV stock	Cars	BEV	2012	Vehicles	220			
7	Australia	Historical	EV sales	Cars	BEV	2012	Vehicles	170			
8	Australia	Historical	EV stock sh	Cars	EV	2012	percent	0.0024			
9	Australia	Historical	EV sales sh	Cars	EV	2012	percent	0.03			
10	Australia	Historical	EV stock	Cars	PHEV	2012	Vehicles	80			
11	Australia	Historical	EV sales	Cars	PHEV	2012	Vehicles	80			
12	Australia	Historical	EV sales	Cars	PHEV	2013	Vehicles	100			
13	Australia	Historical	EV stock	Cars	PHEV	2013	Vehicles	180			
14	Australia	Historical	EV sales sh	Cars	EV	2013	percent	0.034			
15	Australia	Historical	EV stock sh	Cars	EV	2013	percent	0.0046			
16	Australia	Historical	EV sales	Cars	BEV	2013	Vehicles	190			
17	Australia	Historical	EV stock	Cars	BEV	2013	Vehicles	410			
18	Australia	Historical	EV stock	Cars	BEV	2014	Vehicles	780			
19	Australia	Historical	EV sales	Cars	BEV	2014	Vehicles	370			
20	Australia	Historical	EV stock sh	Cars	EV	2014	percent	0.014			
21	Australia	Historical	EV sales sh	Cars	EV	2014	percent	0.16			
22	Australia	Historical	EV stock	Cars	PHEV	2014	Vehicles	1100			
23	Australia	Historical	EV sales	Cars	PHEV	2014	Vehicles	950			
24	Australia	Historical	EV sales	Cars	PHEV	2015	Vehicles	1000			
25	Australia	Historical	EV stock	Cars	PHEV	2015	Vehicles	2100			
26	Australia	Historical	EV sales sh	Cars	EV	2015	percent	0.2			

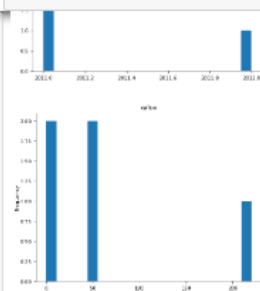
Figure 3.3 Initial Dataset

```
In [3]: # Load the dataset

from google.colab import drive
drive.mount('/content/drive')
file_path = '/content/drive/MyDrive/Colab Notebooks/IEA Global EV Data 2024.csv'
df = pd.read_csv(file_path)

# Display the first few rows
print(df.head())
```

```
In [ ]: df.head()
```



Categorical distributions



Figure 3.4 Loading and Showing the Initial Database

As shown in Figure 3.2, the original data contains more than 100,000 rows of data, including various electric vehicle sales data from 2010 to 2023 in various countries. This study will compare and analyze multi-source data to ensure the authenticity of the data.

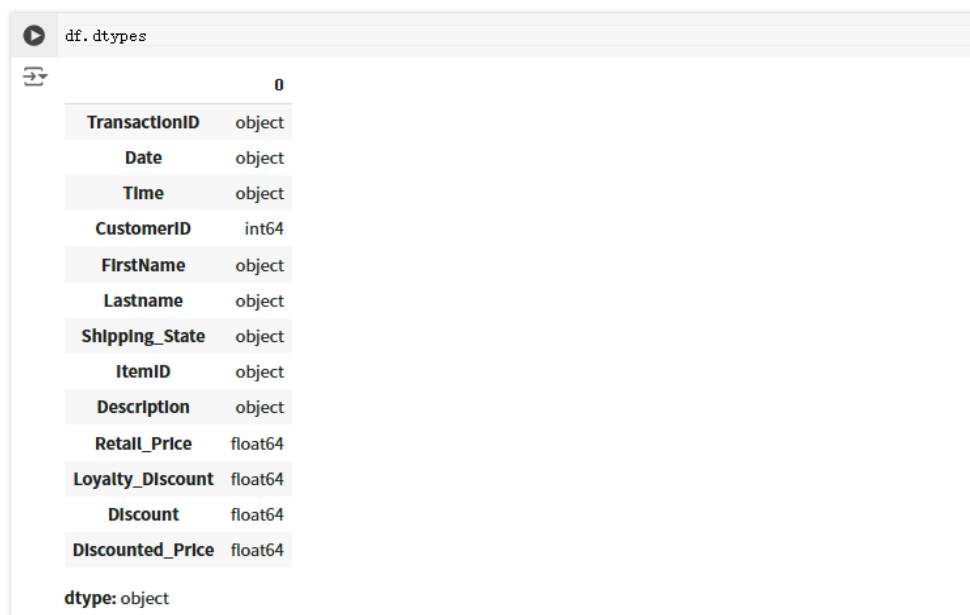
### 3.4.1 Data Pre-Processing and Initial Analysis

Before data cleaning, it is necessary to complete the preliminary analysis and processing of the original data.

The data preprocessing stage includes:

- a. 1. Visualize the original data, analyze the overall data, and understand the data characteristics.
- b. 2. Data conversion: convert the data types of different sources after merging into the same type of data for subsequent processing.

#### ✓ CHECKING FOR DATA TYPES



```
df.dtypes
```

	0
TransactionID	object
Date	object
Time	object
CustomerID	int64
FirstName	object
Lastname	object
Shipping_State	object
ItemID	object
Description	object
Retail_Price	float64
Loyalty_Discount	float64
Discount	float64
Discounted_Price	float64

dtype: object

#### ✓ CONVERT DATA TYPES FOR 'TransactionID' TO STRING OBJECT

```
[ ] df.TransactionID = df.TransactionID.astype(str)
df.ItemID = df.ItemID.astype(str)
df.CustomerID = df.CustomerID.astype(str)
```

Figure 3.5 Analyzing and convert Database

### 3.4.2 Data Cleaning

This operation realizes data cleaning, removes information that does not meet the research scope and erroneous information, extracts the data required for the research, and finally exports it into new usable data.

#### ▼ INSERT SPLITTED 'Timestamp' COLUMNS INTO df, RENAMING AS 'Date', 'Time' AT LOCATION [1] AND [2] RESPECTIVELY

```
[ ] df.insert(loc = 1, column = 'Date', value = dfsplit[0])
    df.insert(loc = 2, column = 'Time', value = dfsplit[1])
    df
```

显示隐藏的输出项

#### ▼ DROP 'Timestamp' COLUMN FROM DF

```
df = df.drop(['Timestamp'], axis=1)
df
```

显示隐藏的输出项

```
df.rename(columns = {'Name1': 'FirstName', inplace = True)
df.rename(columns = {'Surname': 'Lastname'}, inplace = True)
df
```

显示隐藏的输出项

#### ▼ CREATE NEW COLUMN 'Discount' AND 'Discounted\_Price' BY CALCULATING 'Retail\_Price' AND 'Loyalty\_Discount'

```
[ ] df['Discount'] = df['Retail_Price']*df['Loyalty_Discount']
```

```
[ ] df['Discounted_Price'] = df['Retail_Price']-df['Discount']
df
```

Figure 3.6 Extracting valid information



Figure 3.7 Checking missing values

df.describe()

	TransactionID	CustomerID	ItemID	Retail_Price	Loyalty_Discount	Discount	Discounted_Price
count	3455.000000	3.455000e+03	3.455000e+03	3455.000000	3455.000000	3455.000000	3455.000000
mean	111528.000000	1.797979e+08	5.276712e+09	58.526237	0.050457	2.968812	55.557323
std	997.516917	9.563412e+07	2.600486e+09	34.464217	0.032215	2.833184	32.728502
min	109801.000000	1.000000e+08	1.039855e+09	5.600000	0.000000	0.000000	5.040000
25%	110664.500000	1.000003e+08	2.963301e+09	31.800000	0.020000	0.777550	29.575000
50%	111528.000000	1.000009e+08	5.145202e+09	51.660000	0.050000	2.079000	49.810000
75%	112391.500000	2.000009e+08	7.645689e+09	79.800000	0.080000	4.374000	76.920000
max	113255.000000	4.000009e+08	9.916068e+09	159.800000	0.100000	14.382000	159.800000

Figure 3.8 Describing data frame

## EXPORT DATAFRAME AS .XLSX AND .CSV

```
[ ] # Export transformed table to .csv
df.to_csv(f"Transaction.csv", index=False)
```

```
# Export transformed table to .xlsx
df.to_excel(f"OLTP.xlsx", index=False)
```

Figure 3.9 Export the New Dataset

## 3.5 Data Modeling

The processed data is further analyzed and processed, mainly including visual display and trend prediction.

```
[1] import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
```

Generate 10 random numbers using numpy

```
# Load the dataset
from google.colab import drive
drive.mount('/content/drive')
file_path = '/content/drive/My Drive/content/dataset_ev/IEA_Global_EV_Data_2024_new.csv'
df = pd.read_csv(file_path)

# Display the first few rows
print(df.head())
```

	region	category	parameter	mode	powertrain	year	unit \
0	Austria	Historical	EV stock	Cars	BEV	2010	Vehicles
1	Austria	Historical	EV stock share	Cars	EV	2010	percent
2	Belgium	Historical	EV stock	Buses	BEV	2010	Vehicles
3	Belgium	Historical	EV sales	Vans	BEV	2010	Vehicles
4	Belgium	Historical	EV stock	Vans	BEV	2010	Vehicles

Figure 3.10 Loading Dataset

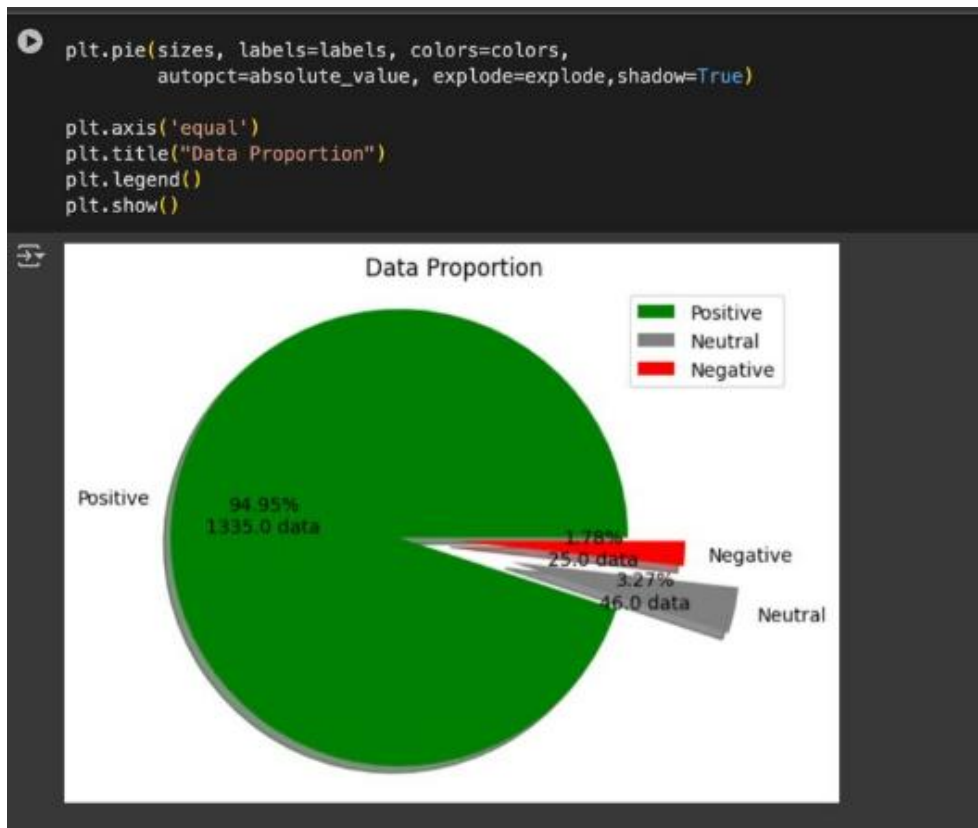


Figure 3.11 Data visualization

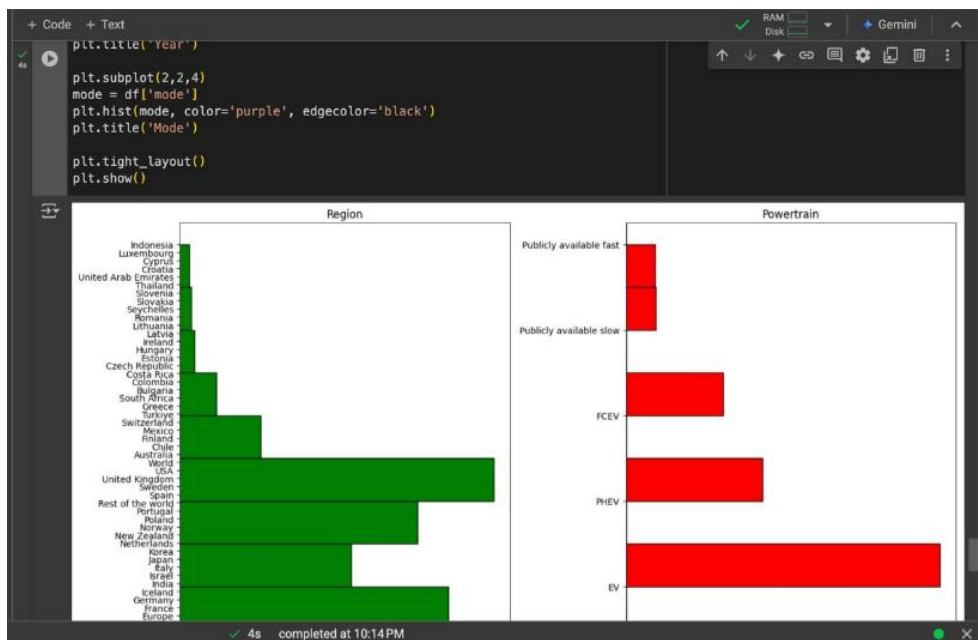


Figure 3.12 Data visualization



### **3.6 Summary**

This chapter details the research process from data collection to classification model evaluation. This process demonstrates the implementation of specific research methods such as data collection and analysis, result modeling and visualization.