ĐỀ TÀI 7: DỰ BÁO PHÁT TRIỂN CỦA XE ĐIỆN TRÊN THẾ GIỚI BẰNG MÔ HÌNH ARIMA

Thành viên nhóm:

Số thứ tự	Mã số sinh viên	Họ tên
1	21028411	Trương Dương Minh Nhật
2	21080071	Nguyễn Thanh Thuận
3	21017351	Trương Quốc Bảo
4	22643391	Nguyễn Hải Tâm
5	21024541	Nguyễn Đức Tài

I. LÝ DO CHỌN ĐỀ TÀI

- Với xu hướng phát triển phương tiện điện ngày càng phổ biến dần, nó được coi là xu hướng của tương lai và có thể thay thế các phương tiện sử dụng xăng, dầu.
- Dựa vào sự phát triển và phổ biến, có thể dự báo được sự phát triển của phương tiện ở các thời điểm tương lai.
- Có thể dự báo được rằng, phương tiện điện nào là phương tiện sẽ là phổ biến mạnh nhất (trend) trong tương lai dựa vào doanh thu.

II. DỮ LIỆU

- Dữ liệu đề tài của nhóm em sử dụng sẵn trên trang Kaggle, với sự tìm hiểu về bộ dữ liệu thì bộ dữ liệu của nhóm em bao gồm: 12654 dòng và 9 cột (bao gồm: 3 biến định lượng và 6 biến định tính)
- Giải thích từng cột dữ liệu:
    - Region (khu vực): dùng để mô tả khu vực kinh doanh xe điện – *biến định tính*
    - Category (loại): phân loại danh mục dữ liệu – *biến định tính*
    - Parameter (tham số): phân loại kiểu của dữ liệu – *biến định tính*
    - Mode (chế độ): mô tả loại xe (hay phương tiện) điện nào. – *biến định tính*

- Powertrain (năng lượng chạy): mô tả loại động cơ điện mà xe (hay phương tiện sử dụng) – *biến định tính*
- Year (năm): năm mà kinh doanh – *biến định lượng*
- Unit (đơn vị): mô tả đơn vị đo lường dữ liệu – *biến định tính*
- Value (giá trị): mô tả giá trị thu nhập (hay là doanh thu) của dữ liệu – *biến định lượng*
- Percentage (phần trăm): cũng giống như value nhưng mô tả mức phần trăm (công thức: value * 100 = percentage) – *biến định lượng*

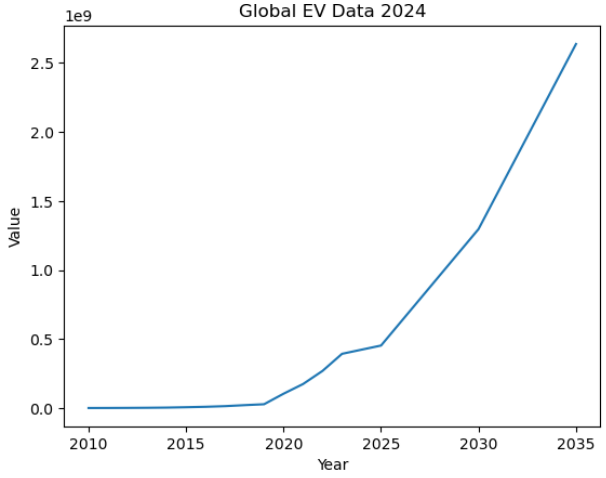
III. MÔ HÌNH MÁY HỌC

1. Chọn mô hình

Dựa vào đề tài của nhóm em và cũng như là bộ dữ liệu mà chúng em đã phân tích, nhóm chúng em quyết định sử dụng mô hình máy học **ARIMA** – mô hình xử lý chuỗi thời gian Time – Series. So với các mô hình máy học khác, nhóm chúng em cho rằng mô hình này rất phù hợp cho yêu cầu cũng như bộ dữ liệu mà chúng em sẽ xử lý.

2. Lý do chọn mô hình

Lý do nhóm chúng em chọn mô hình này, vì trong bộ dữ liệu có chuỗi thời gian là năm (year) ứng với các doanh thu (value).



Dựa vào có thể thấy rằng doanh thu của xe (phương tiện) điện có sự thay đổi qua các năm, cụ thể là dấu hiệu tăng dần của sự phát triển theo thời gian nên việc nhóm em nghĩ rằng mô hình ARIMA rất phù hợp cho bộ dữ liệu này.

IV. EDA DỮ LIỆU

1. Xử lý định dạng dữ liệu, loại bỏ các giá trị null:

```
data['value'] = pd.to_numeric(data['value'], errors='coerce')
data['percentage'] = pd.to_numeric(data['percentage'], errors='coerce')
```
✓ [3] 34ms

```
data['value']    data
```
✓ [4] < 10 ms

1-10 Length: 12654, dtype: float64

	value
Missing	3233
Count	9421
0.3 – 440000000	
0	350.0
1	NaN
2	3.0
3	7.0
4	62.0
5	48.0
6	61.0
7	8.0
8	NaN
9	1.0

```
data.isnull().sum()
```
✓ [5] 10ms

9 rows Length: 9, dtype: int64

	<unnamed>
Missing	0
Count	9
0 – 12616	
region	0
category	0
parameter	0
mode	0
powertrain	0
year	0
unit	0
value	3233
percentage	12616

2. Gom nhóm dữ liệu

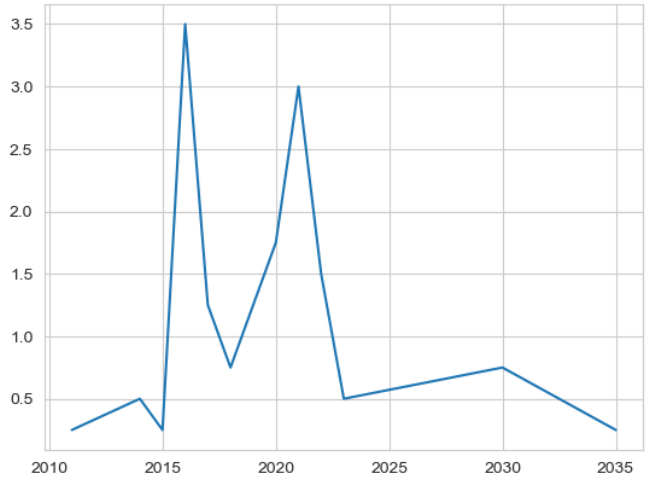
```
data_group = data[["year", "value"]].groupby("year").sum().reset_index()
data_group
```
✓ [7] 15ms

1-10 13 rows × 2 columns

	year	value
Missing	0	0
Count	13	13
2011 – 2035		0.3 – 3.5
0	2011	0.25
1	2014	0.50
2	2015	0.25
3	2016	3.50
4	2017	1.25
5	2018	0.75
6	2019	1.25
7	2020	1.75
8	2021	3.00
9	2022	1.50

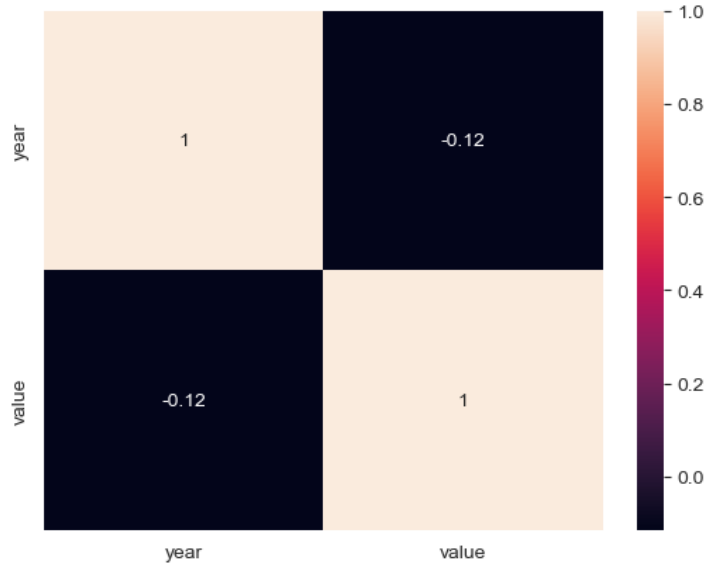
3. Trực quan dữ liệu:

```
plt.plot(data_group["year"], data_group["value"])
plt.show()
✓ [8] 337ms
```



4. Biểu diễn độ tương quan:

```
data_corr = data_group.corr()
sns.heatmap(data_corr, annot = True)
plt.show()
✓ [9] 226ms
```



V. ÁP DỤNG MÔ HÌNH MÁY HỌC

Để có thể áp dụng được mô hình máy học ARIMA trong đồ án, nhóm em đã chia ra và phân tích các bước làm để có thể áp dụng được mô hình cũng như là dự báo tương lai của sự phát triển.

- Bước 1: Đọc, thu thập bộ dữ liệu để phân tích
- Bước 2: Phân tích, xử lý và làm sạch dữ liệu
- Bước 3: Lọc và lấy dữ liệu cần thiết
- Bước 4: Xác định chuỗi time – series sau khi lọc dữ liệu có tính dừng hay không (stationary), có 3 cách kiểm tra:
    - Cách 1: Vẽ biểu đồ đường để thể hiện
    - Cách 2: Tính ACF và PACF
    - Cách 3: Tính ADFuller
    - Bước 4.1: Vẽ đồ thị đường để thể hiện, dựa vào đồ thị đường nếu các đường có dấu hiệu lên và xuống liên tục quanh một giá trị trung bình ở cột y => Chuỗi time – series có tính dừng (ngược lại không có tính dừng thì buộc phải tính các cách khác). Với cách này độ chính xác không tin cậy.
    - Bước 4.2: Tính ACF và PACF (tính độ tương quan và độ tương quan 1 phần) – việc tính độ tương quan để chỉ rõ giữa các chân ở thời điểm này so với thời điểm trước có quan hệ tương quan, tuyến tính gì không? Thông thường, với cách này rất khó để nhận thấy được.
    - Bước 4.3: Tính ADFuller, việc tính bằng cách này giúp chúng ta có thể nhận thấy được mô hình chuỗi Time – Series có tính dừng hay không, dựa vào giả thuyết H0 và H1.
        - Giả sử H0: Chuỗi time – series không có dừng.
        - Giả sử H1: Chuỗi time – series có tính dừng.
        - Nếu tính giá trị ADFullter, cho ra p_value có giá trị < 0.05 (mức ý nghĩa 5%), đồng nghĩa chúng ta bác bỏ giả thuyết H1 lấy H0 và ngược lại.
- Bước 5: Xác định các tham số p (AR), d (I) và q (MA) – để có thể áp dụng mô hình ARIMA, cần phải xác định được 3 tham số quan trọng là đầu vào để mô hình có thể train được bộ dữ liệu.
    - Xác định p và q: dựa vào đồ thị tương quan và tương quan 1 phần (ACF và PACF) đã được tính ở bước 4, để có thể xác định được ta dựa vào các giả thuyết sau:
        - Nếu đồ thị ACF có dấu hiệu tăng đột biến ở các chân đầu nhưng không quá đột biến quá mức và tan rã ở các chân sau, cùng với đồ thị PACF không có dấu hiệu đột biến ở các chân đầu nhưng

phân rã mạnh mẽ về sau => Chúng ta có thể xác định: q = 0 và p sẽ là hệ số dựa vào thứ tự các chân

- Nếu đồ thị PACF có dấu hiệu tăng đột biến ở các chân đầu nhưng không quá đột biến quá mức và tan rã ở các chân sau, cùng với đồ thị ACF không có dấu hiệu đột biến ở các chân đầu nhưng phân rã mạnh mẽ về sau => Chúng ta có thể xác định: p = 0 và q sẽ là hệ số dựa vào thứ tự các chân.
- Nếu trong trường hợp, cả ACF và PACF đều có dấu hiệu tăng đột biến hoặc phân rã ở các chân đầu, lúc này hệ số p và q dựa vào thứ tự của các chân có độ tương quan mạnh.

- Xác định d (I): dựa vào số lần sai phân – để tính được số lần sai phân thì cần phải dựa vào giá trị p_value tính được ở ADFuller, nếu lần đầu mà p_value <= 0.05 (bác bỏ giả thuyết H0) thì mặc định d = 0. Nhưng nếu p_value > 0.05 thì cần phải lấy sai phân để sao cho chuỗi Time – Series phải có tính dừng. Nếu lần sai phân tiếp theo mà vẫn chưa thoả mãn thì tiếp tục lấy sai phân tiếp ở lần sai phân trước đó. Thông thường sai phân tối đa mức 2 – 3 lần là ổn nhất, có thể hơn ở mọi trường hợp xấu.

- Bước 6: Áp dụng mô hình ARIMA
- Bước 7: Dự báo xu hướng phát triển

Lưu ý: trong trường hợp không thể xác định được 3 tham số p, d và q một cách thủ công có thể sử dụng một thư viện hỗ trợ auto_arima, thư viện này giúp cho việc tự động tìm tham số p, d và q phù hợp để thực hiện train mô hình.

1. Chia tập train, tập test

```
train = data_group[data_group["year"] <= 2023]
test = data_group[data_group["year"] >= 2023]
train, test
✓ [10] 16ms
```

```
   2   2015   0.25
   3   2016   3.50
   4   2017   1.25
   5   2018   0.75
   6   2019   1.25
   7   2020   1.75
   8   2021   3.00
   9   2022   1.50
   10  2023   0.50,
       year  value
   10  2023   0.50
   11  2030   0.75
   12  2035   0.25)
```

2. Kiểm tra chuỗi time – series có tính dừng không?

```
adfuller(train["value"])    train
✓ [31] 11ms
```

```
(-2.781185328699488,
 0.061011975232483,
 1,
 9,
 {'1%': -4.473135048010974,
  '5%': -3.28988060356653,
  '10%': -2.7723823456790124},
 19.272979508238105)
```

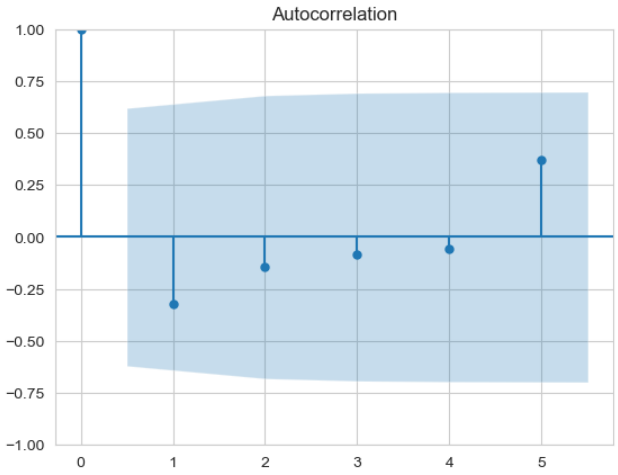
3. Lấy sai phân và kiểm lại độ ổn định của chuỗi time – series:

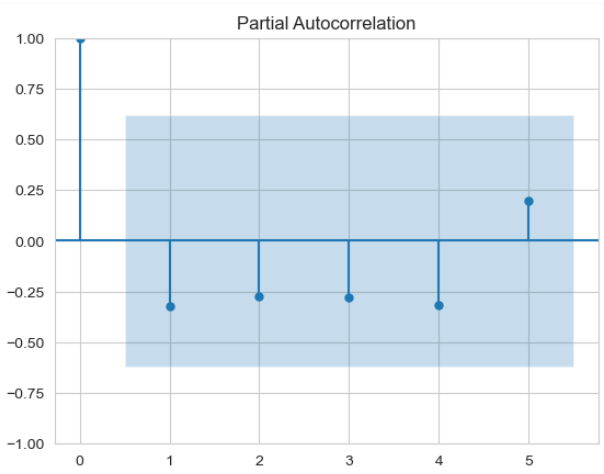
Lúc này có thể thấy p_value có giá trị 0.01 < 0.05 => Ngay sai phân 1 chuỗi mới có tính dừng

```
train_diff_1 = train["value"].diff().dropna()
adfuller(train_diff_1)
✓ [32] 11ms
```

```
(-3.322376869678028,
 0.013907381219126266,
 3,
 6,
 {'1%': -5.354256481481482,
  '5%': -3.6462381481481483,
  '10%': -2.901197777777778},
 13.386246362519806)
```

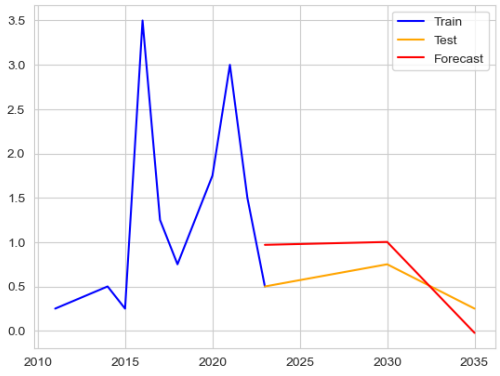
4. Kiểm tra độ tương quan và tương quan 1 phần giữa các chuỗi time - series:





5. Tiến hành train mô hình, trực quan và đánh giá mô hình:

- Dựa vào mô hình có thể thấy, sau khi train thì có thể thấy rằng giữa tập test và tập dự báo có xu hướng khá khít dần (fitting), điều này cho thấy rằng mô hình đang có xu hướng hoạt động tốt khi chọn tham số là p = 5, d = 1 và q = 5.
- Mô hình cũng không quá underfitting cũng như là overfitting nhưng dựa vào trực quan đồ thị thì có thể nghiêng về phân overfitting.
- Ngoài các tham số lựa chọn p, d và q thì nhóm chúng em có lựa chọn các tham số khác nhưng sau khi test nhiều tham số thì nhóm chúng em cho rằng p = 5, d = 1 và q = 5 có độ sai lệch thấp nhất (dựa vào 3 độ đo MSE, MAE và MAPE) so với các tham số khác.



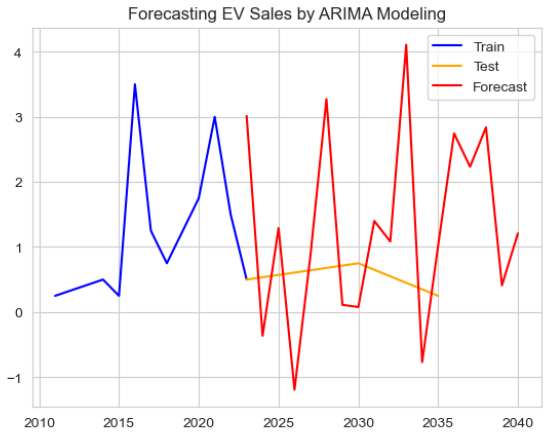

```
mean_squared_error(test["value"], forecast), mean_absolute_error(test["value"], forecast), mean_absolute_percentage_error(test["value"], forecast)
# MSE - MAE - MAPE
✓ [35] < 10 ms
```

```
(0.11953850401929784, 0.33161931024197416, 0.7895807069961313)
```

```
ARIMA(3, 0, 2): (1.0244163286939914, 0.8274959591149648, 2.3628053091261276)
ARIMA(1, 1, 2): (0.40697182114692954, 0.5722009145098567, 1.3275423228337007)
ARIMA(1, 0, 1): (0.37306307270709005, 0.5418028490845743, 1.0486537413045798)
ARIMA(0, 0, 0): (0.2672964185877666, 0.4750050019958736, 0.9389011159899132)
ARIMA(2, 1, 1): (0.2531815420983404, 0.4832155621571108, 1.0562684452227664)
ARIMA(1, 0, 0): (0.23168380763421972, 0.37097182184772115, 0.6798463135539675)
ARIMA(1, 1, 1): (0.2240230576390695, 0.39240985221580865, 0.744984671530755)
ARIMA(5, 1, 5): (0.11953850401929784, 0.33161931024197416, 0.7895807069961313)
```

6. Dự báo tương lai:

- Sau khi kiểm tra mô hình, nhóm chúng em tiến hành dự báo tương lai cho sự phát triển của doanh thu.
- Dựa vào mô hình có thể nói rằng, mô hình hoạt động không thực sự tốt khi khoảng thời gian giữa tập test và tập train không thực sự khít nhau và gần như là undefitting.
- Khi kiểm tra các độ đo thì gần như chỉ có tham số p = 5, d = 2 và q = 5 là cho ra độ sai lệch nhỏ nhất so với các tham số còn lại nhưng về độ sai sót vẫn khá lớn nên rất khó cho rằng mô hình dự báo chính xác.
- Về dự báo thì doanh thu có xu hướng dao động liên tục không có ngừng nghỉ ở tương lai.




```
mean_squared_error(forecast_years, forecast), mean_absolute_error(forecast_years, forecast), mean_absolute_percentage_error(forecast_years, forecast)
```

✓ [39] < 10 ms

```
(4121745.6942141172, 2030.201794474657, 0.9993612754698193)
```

{} Code | M↓ Markdown | SQL

ARIMA(5, 1, 5): (4126892.6617287826, 2031.4681717732476, 0.9999841028224498)
ARIMA(1, 1, 1): (4126808.942116284, 2031.4482102828506, 0.9999744832907275)
ARIMA(5, 1, 1): (4126565.781916346, 2031.387602500944, 0.9999443919334582)
ARIMA(3, 0, 2): (4126448.4943599724, 2031.358903730733, 0.9999303484982894)
ARIMA(5, 2, 5): (4121745.6942141172, 2030.201794474657, 0.9993612754698193)