

NHẬP MÔN KHOA HỌC DỮ LIỆU BÁO CÁO ĐỒ ÁN CUỐI KỲ

DỰ BÁO THỜI TIẾT Ở TPHCM

Nhóm 14:

19120472 - Nguyễn Văn Tuấn Đạt

19120543 - Hoàng Mạnh Khiêm

19120501 - Nguyễn Nhật Hảo

20120237 - Hà Nguyễn Thảo Vy

NỘI DUNG ĐÔ ÁN

- ~~ Giới thiệu đồ án
- ~~ Thu thập dữ liệu
- ~~ Khám phá và Tiền xử lý dữ liệu
- ~~ Xây dựng mô hình
- ~~ Đánh giá mô hình và tinh chỉnh



GIỚI THIỆU ĐỒ ÁN

Câu hỏi đặt ra?

Từ các thông số thời tiết ở TPHCM, tiến hàng dự đoán nhiệt độ trung bình ngày tiếp theo.

Input: dữ liệu về thời tiết TPHCM từ ngày 01/01/2010 đến 31/12/2021

Output: Thông số nhiệt độ trung bình ngày tiếp theo



Ho Chi Min »

GIỚI THIỆU ĐỒ ÁN

Dữ liệu được thu thập tại trang web:

Ho Chi Minh City Historical Weather

<https://www.worldweatheronline.com/ho-chi-minh-city-weather-history/vn.aspx>

Dữ liệu mà nhóm thu thập đã được
cho phép.

Ho Chi Minh City Historical Weather

VN
(Not the location you were looking for? Other matching results or Interactive Map Search)

Time in Ho Chi Minh City is Fri 09th Dec 4:42 am
2718413:3467426

[Weather »](#) [Hourly »](#) [3 Hour »](#) [History »](#) [Graphs »](#) [Averages »](#) [Widgets »](#) [Weather Maps](#) [** Alert **](#)

07/12/2022

Weather History

Wednesday, 07 December 2022

Min/max 23°/33°C Patchy rain possible 

Moonrise: 04:48 PM Moonset: 05:00 AM  05:59 AM  05:31 PM

Time	Forecast	Rain	Rain %	Cloud	Pressure	Wind	Gust	Dir
00:00 25 °C		28 °C	0.1 mm	0%	88%	1008 mb	4 km/h	6 km/h 
03:00 24 °C		27 °C	0.1 mm	0%	79%	1007 mb	8 km/h	14 km/h 

THU THẬP DỮ LIỆU

Dữ liệu thu thập được bao gồm hơn 4000 dòng dữ liệu (từ 01/01/2010 - 31/12/2021) và 10 cột (chưa tiền xử lý dữ liệu) bao gồm:

- **Date:** Ngày thu thập
- **Weather Type:** Loại thời tiết
- **Average Temperature:** Nhiệt độ trung bình trong ngày (°C)
- **Highest Temperature:** Nhiệt độ cao nhất trong ngày (°C)
- **Lowest Temperature:** Nhiệt độ thấp nhất trong ngày (°C)
- **Wind Speed:** Tốc độ gió (km/h)
- **Rain:** Lượng mưa (mm)
- **Humidity:** Độ ẩm (%)
- **Cloud:** Độ che phủ của mây (%)
- **Pressure:** Áp suất không khí (mb)

Ho Chi Minh City Historical Weather on 10th December over the years

Year	Weather	Max	Min	Wind	Rain	Humidity	Cloud	Pressure
2009		35 °c	23 °c	6 km/h ESE	0.0 mm	70%	5%	1011 mb
2010		30 °c	23 °c	4 km/h SW	0.0 mm	79%	33%	1009 mb
2011		30 °c	22 °c	11 km/h E	0.1 mm	79%	33%	1010 mb
2012		33 °c	24 °c	6 km/h SW	0.4 mm	76%	26%	1010 mb
2013		34 °c	23 °c	6 km/h E	0.0 mm	76%	7%	1008 mb
2014		31 °c	21 °c	9 km/h S	0.0 mm	73%	11%	1011 mb
2015		33 °c	26 °c	8 km/h ESE	3.4 mm	75%	38%	1011 mb
2016		30 °c	24 °c	6 km/h SE	1.3 mm	82%	48%	1008 mb
2017		32 °c	24 °c	4 km/h SW	0.0 mm	70%	8%	1013 mb
2018		34 °c	26 °c	11 km/h E	0.0 mm	68%	19%	1010 mb
2019		31 °c	22 °c	7 km/h WSW	0.0 mm	76%	12%	1012 mb
2020		32 °c	23 °c	6 km/h SE	13.4 mm	77%	64%	1009 mb
2021		32 °c	22 °c	9 km/h E	0.0 mm	70%	24%	1012 mb

THU THẬP DỮ LIỆU

Các trường dữ liệu: Weather Type, Highest Temperature, Lowest Temperature, Wind Speed, Rain, Humidity, Cloud, Pressure được thu thập từ bảng thống kê thông số thời tiết của ngày qua các năm.

Saturday, 10 December 2022

Min/max

23°/31°c

Light rain shower



Moonrise: 07:19 PM

Moonset: 07:38 AM

06:01 AM

05:32 PM

Time	Forecast	Rain	Rain %	Cloud	Pressure	Wind	Gust	Dir
00:00	27 °c	0.0 mm	0%	36%	1009 mb	6 km/h	10 km/h	↗
03:00	26 °c	0.0 mm	0%	23%	1007 mb	4 km/h	7 km/h	↖
06:00	26 °c	0.0 mm	0%	35%	1008 mb	4 km/h	6 km/h	↖
09:00	30 °c	0.0 mm	0%	50%	1010 mb	4 km/h	4 km/h	↖
12:00	36 °c	0.1 mm	0%	93%	1008 mb	2 km/h	2 km/h	↗
15:00	34 °c	1.6 mm	0%	88%	1006 mb	1 km/h	1 km/h	↖
18:00	29 °c	1.4 mm	0%	84%	1007 mb	10 km/h	21 km/h	↖
21:00	26 °c	0.1 mm	0%	80%	1009 mb	10 km/h	19 km/h	↗

THU THẬP DỮ LIỆU

Riêng trường dữ liệu Average Temperature được thu thập từ 8 mốc nhiệt độ trong ngày, sau đó tính được nhiệt độ trung bình trong ngày đó.

THU THẬP DỮ LIỆU

Dữ liệu được lưu trữ dưới dạng CSV để phục vụ cho quá trình tiền xử lý, khám phá và mô hình hóa dữ liệu.

Date	Weather Type	Average Temperature	Highest Temperature	Lowest Temperature	Wind Speed	Rain	Humidity	Cloud	Pressure
1/1/2010	Partly cloudy	28.4	34	24	6	0	59	20	1010
2/1/2010	Cloudy	27	32	24	7	0	61	48	1010
3/1/2010	Partly cloudy	28.1	33	26	7	0	65	37	1010
4/1/2010	Partly cloudy	28.2	35	25	9	0.1	61	42	1009
5/1/2010	Partly cloudy	28.8	35	26	10	0	57	35	1009
6/1/2010	Sunny	28.5	34	25	12	0	56	13	1010
7/1/2010	Sunny	28.6	35	25	10	0	55	12	1012
8/1/2010	Sunny	28.2	35	24	6	0	57	14	1012
9/1/2010	Partly cloudy	28	33	24	5	0	64	35	1011
10/1/2010	Patchy light driz	28.4	33	25	6	1.3	62	54	1011

1	Feature	Description
2	Date	Ngày lấy dữ liệu (yyyy-mm-dd)
3	Weather Type	Loại thời tiết
4	Average Temperature	Nhiệt độ trung bình trong ngày (°C)
5	Highest Temperature	Nhiệt độ cao nhất trong ngày (°C)
6	Lowest Temperature	Nhiệt độ thấp nhất trong ngày (°C)
7	Wind Speed	Tốc độ gió (km/h)
8	Rain	Lượng mưa (mm)
9	Humidity	Độ ẩm (%)
10	Cloud	Độ che phủ của mây (%)
11	Pressure	Áp suất không khí (mb)

KHÁM PHÁ VÀ TIỀN XỬ LÝ DỮ LIỆU

In [3]:

```
#shape data  
weather_df.shape
```

Out[3]: (4380, 10)

In [4]:

```
#check null  
weather_df.isnull().sum()
```

Out[4]: Date 0
Weather Type 0
Average Temperature 0
Highest Temperature 0
Lowest Temperature 0
Wind Speed 0
Rain 0
Humidity 0
Cloud 0
Pressure 0
dtype: int64

Dữ liệu không có cột nào chứa giá trị thiếu

In [5]:

```
#check duplicated  
np.any(weather_df.duplicated())
```

Out[5]: False

KHÁM PHÁ VÀ TIỀN XỬ LÝ DỮ LIỆU

Kiểm tra số dòng, số cột và có giá trị thiếu, trùng lặp hay không?

KHÁM PHÁ VÀ TIỀN XỬ LÝ DỮ LIỆU

Chuyển đổi đúng kiểu dữ liệu của cột

In [6]:

```
weather_df.dtypes
```

```
Out[6]: Date          object
Weather Type    object
Average Temperature float64
Highest Temperature int64
Lowest Temperature int64
Wind Speed      int64
Rain            float64
Humidity        int64
Cloud            int64
Pressure         int64
dtype: object
```

Chuyển cột "Date" từ object sang datetime64[ns]

In [7]:

```
weather_df['Date'] = pd.to_datetime(weather_df['Date'])
```

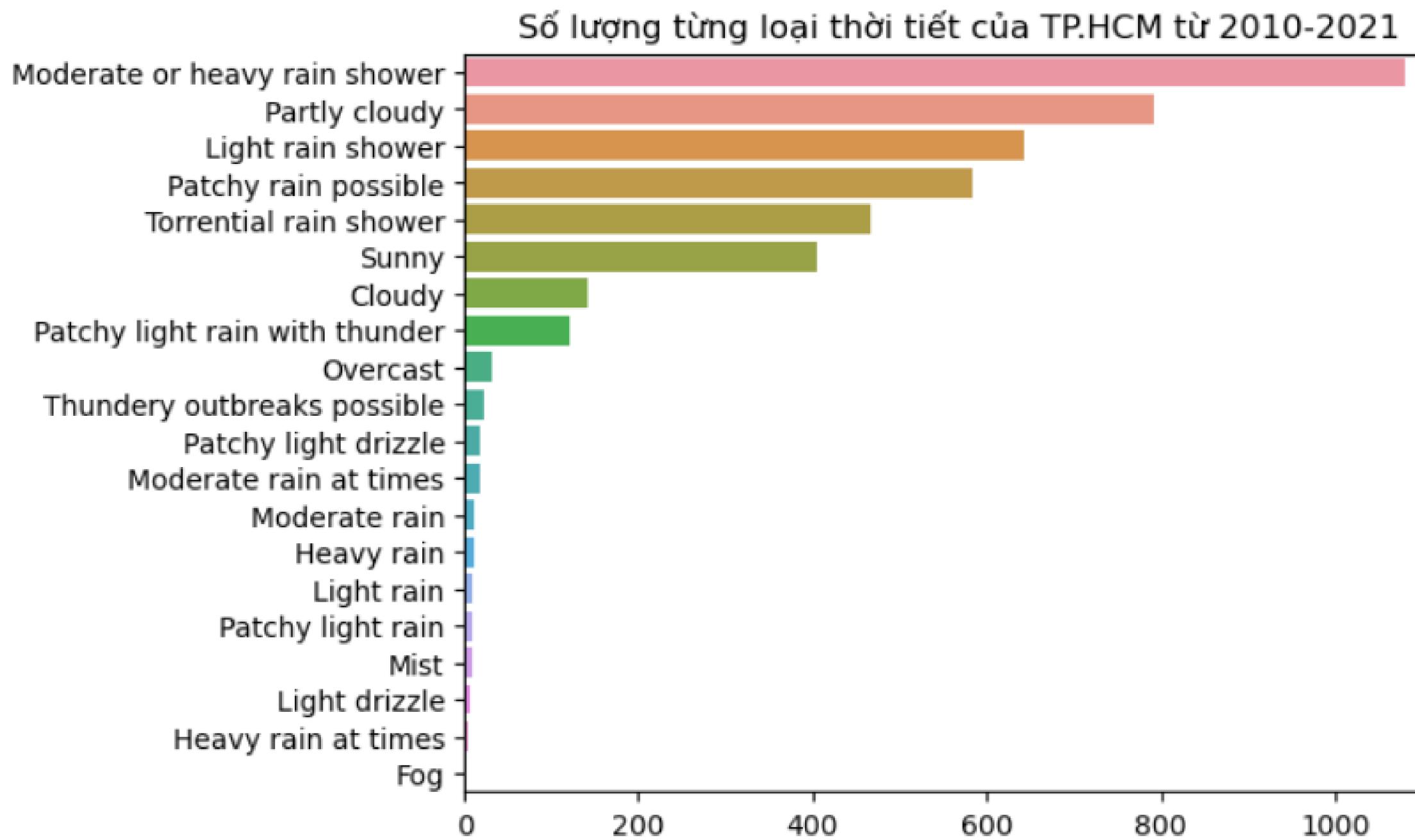
Chuyển đổi các cột có kiểu dữ liệu không phù hợp

-> Chuyển cột Date từ object sang
datetime64[ns]

KHÁM PHÁ VÀ TIỀN XỬ LÝ DỮ LIỆU

Với cột có kiểu dữ liệu phân loại ("Weather Type")

-> Thống kê số lượng và các loại thời tiết ở TPHCM



KHÁM PHÁ VÀ TIỀN XỬ LÝ DỮ LIỆU

Với cột có dữ liệu số:

-> Kiểm tra min, lower_quartile, median, upper_quartile, max, mean

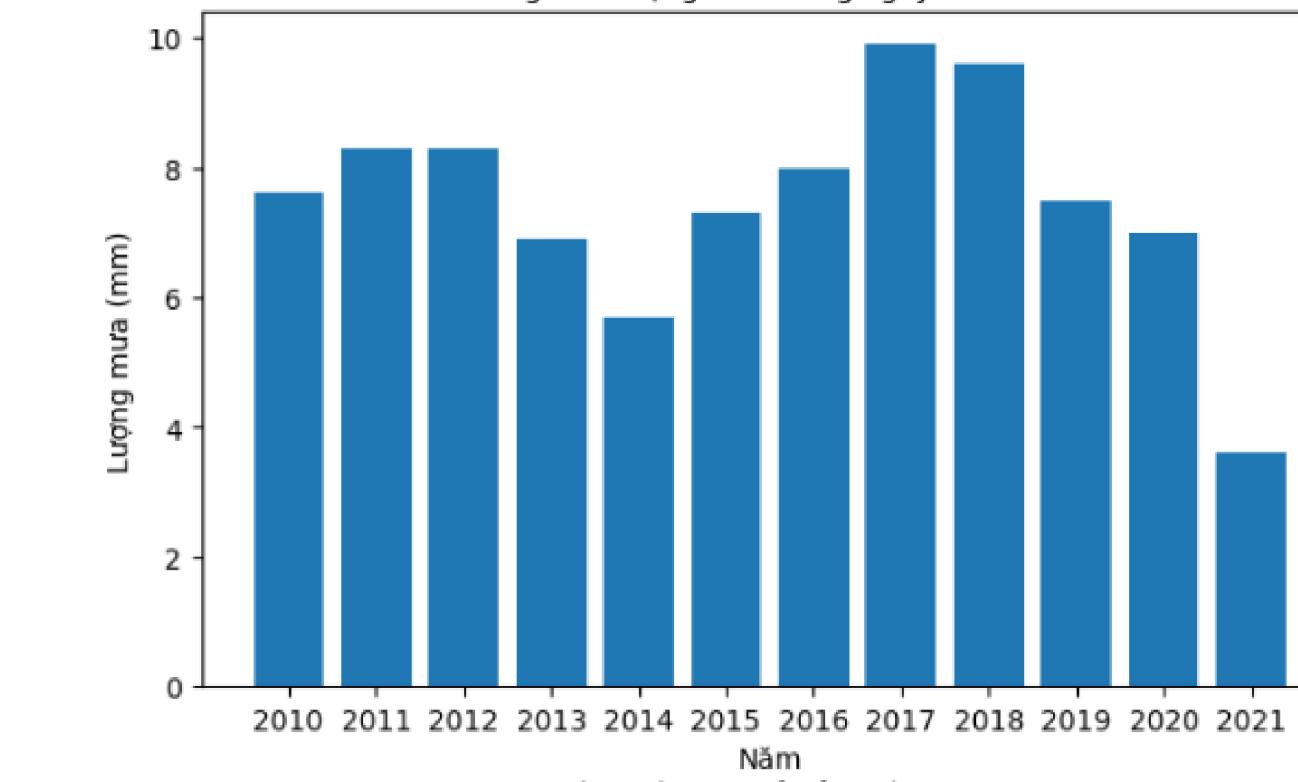
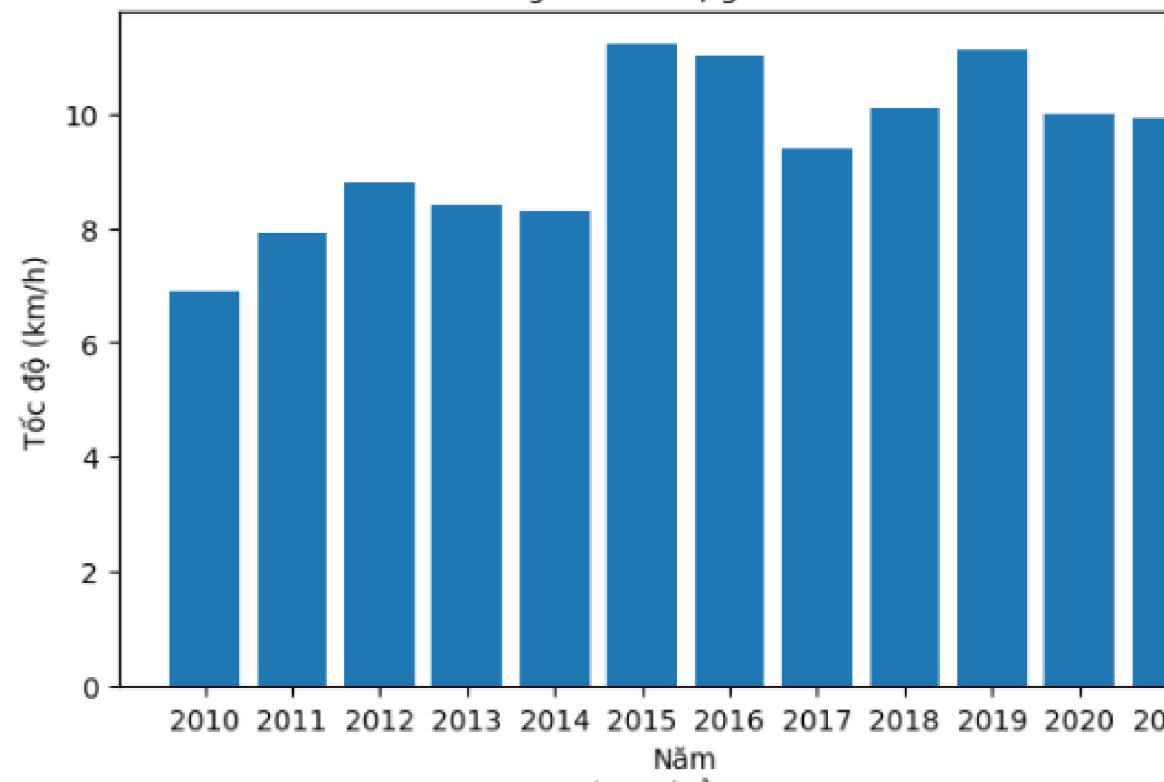
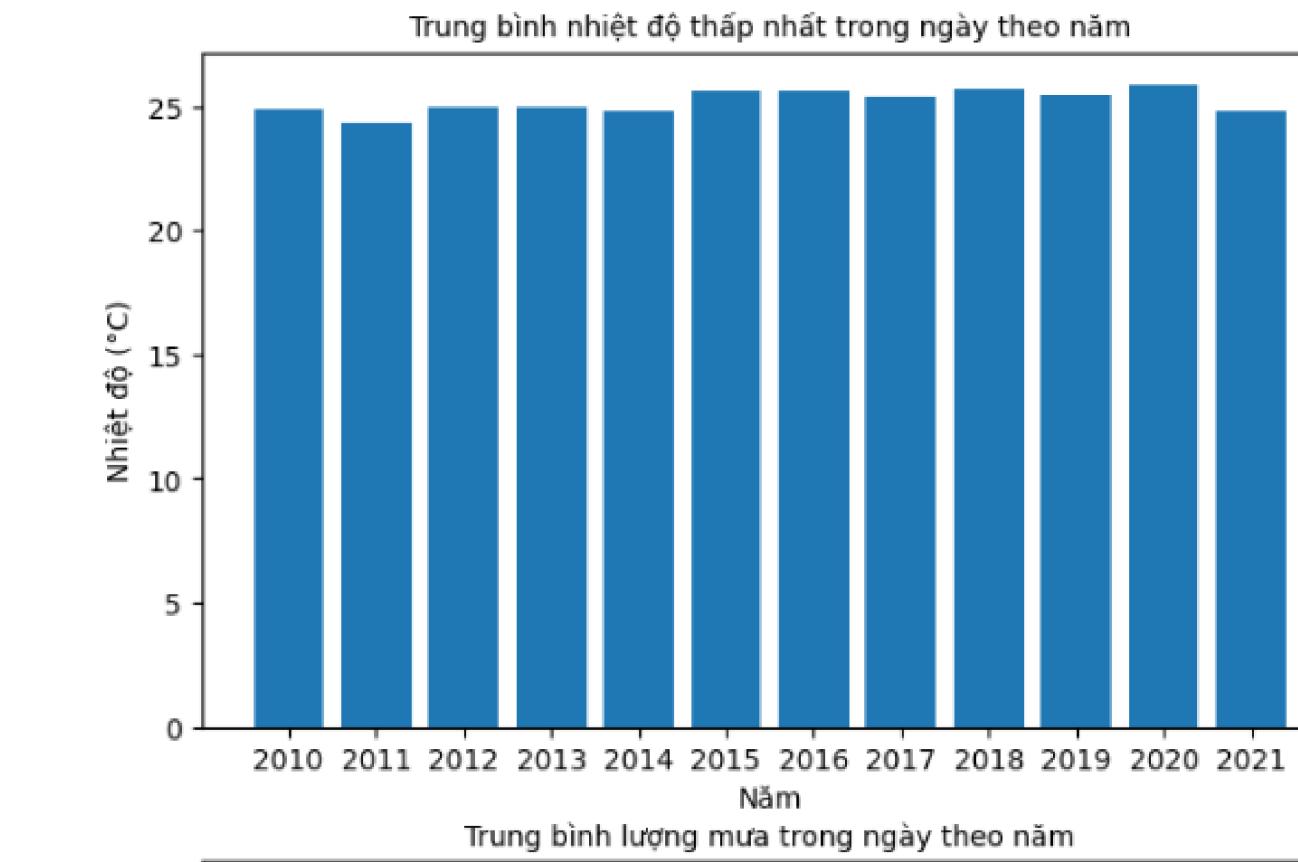
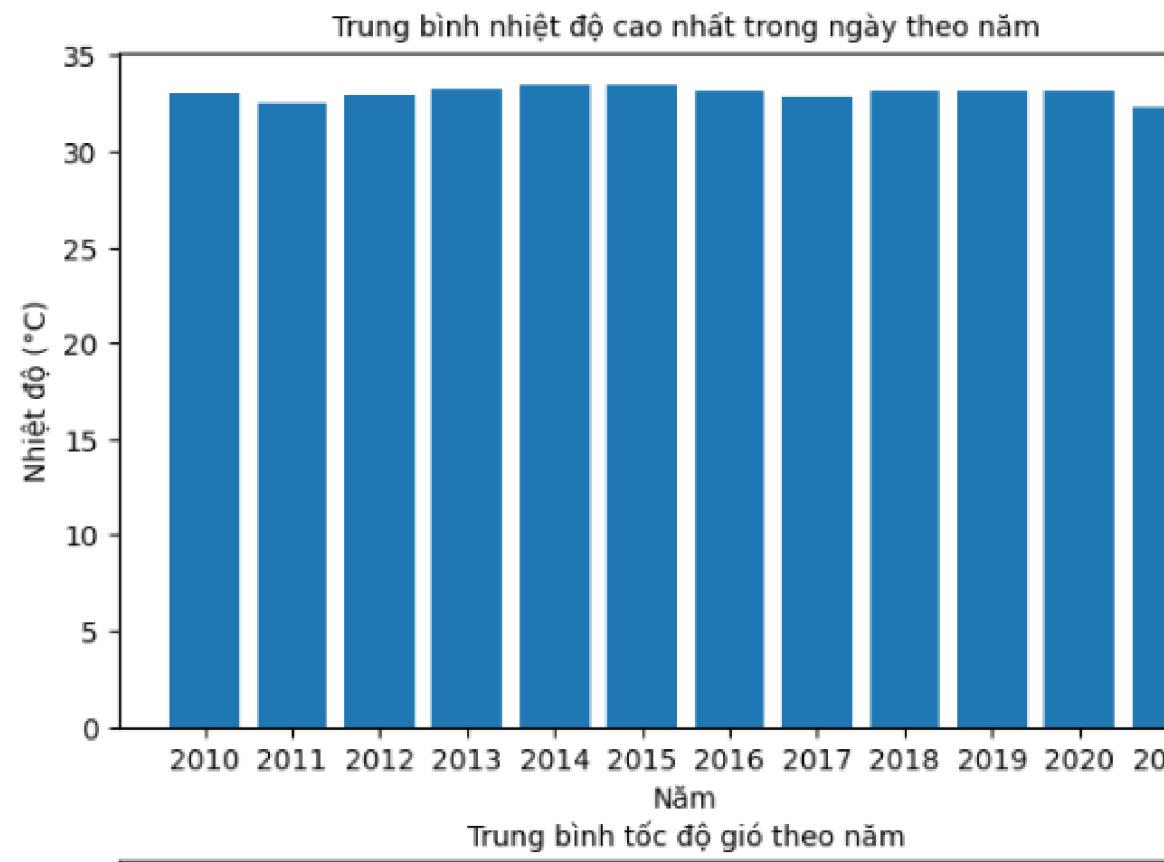
Out[12]:

	Average Temperature	Highest Temperature	Lowest Temperature	Wind Speed	Rain	Humidity	Cloud	Pressure
count	4380.000000	4380.000000	4380.000000	4380.000000	4380.000000	4380.000000	4380.000000	4380.000000
mean	28.376712	33.035616	25.200228	9.413014	7.477968	73.594521	37.042466	1009.256849
std	1.655659	2.474554	1.647616	3.431909	10.980748	10.010817	18.737816	2.076964
min	21.400000	22.000000	16.000000	3.000000	0.000000	45.000000	0.000000	1003.000000
25%	27.400000	32.000000	24.000000	7.000000	0.100000	66.000000	23.000000	1008.000000
50%	28.200000	33.000000	25.000000	9.000000	3.300000	75.000000	34.000000	1009.000000
75%	29.500000	35.000000	26.000000	12.000000	11.100000	82.000000	50.000000	1011.000000
max	33.800000	41.000000	30.000000	26.000000	253.300000	96.000000	99.000000	1016.000000

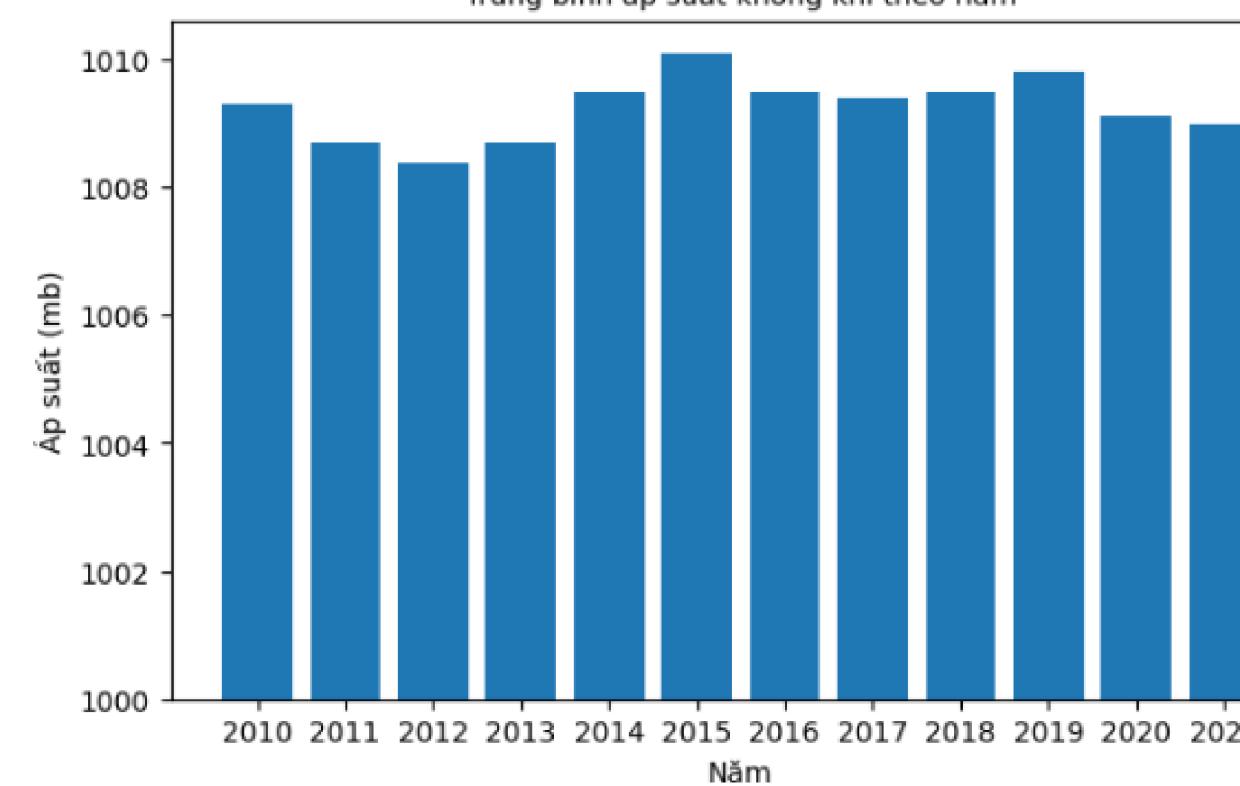
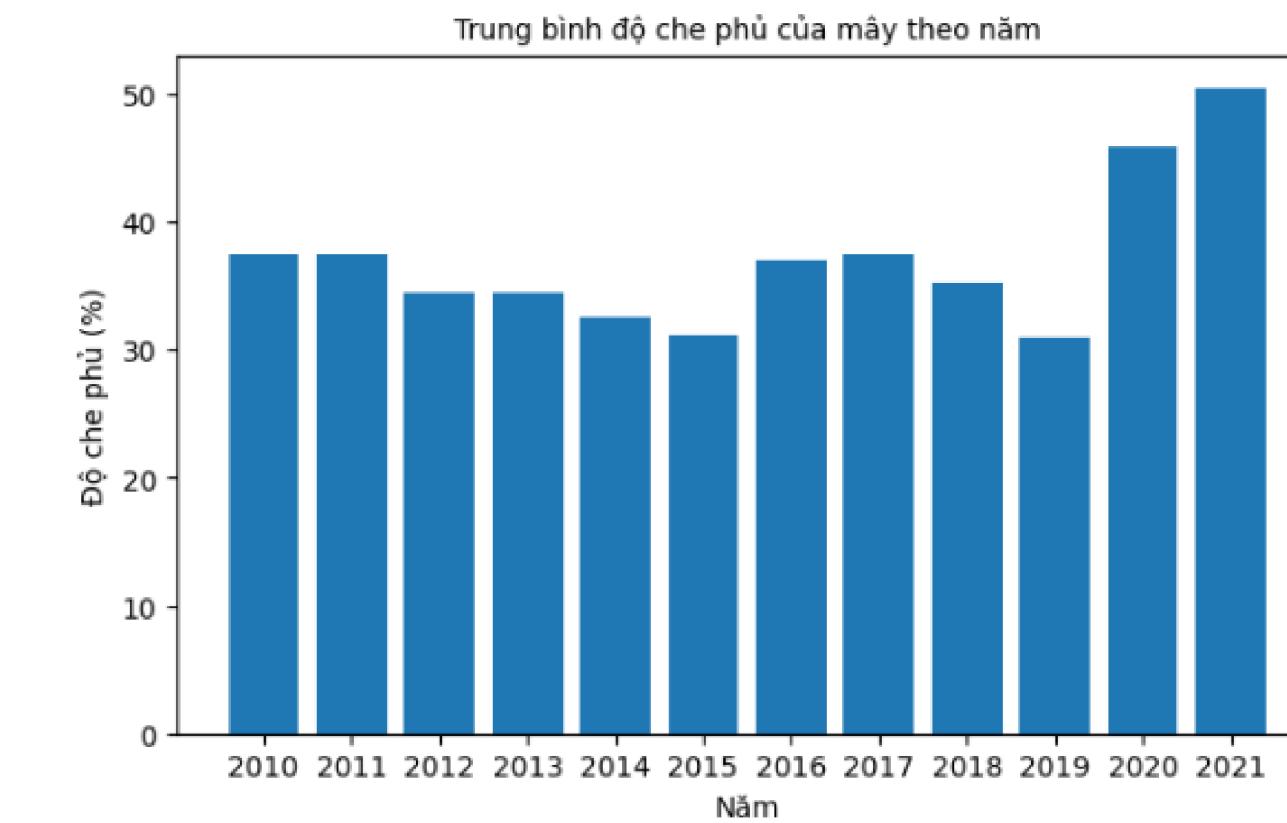
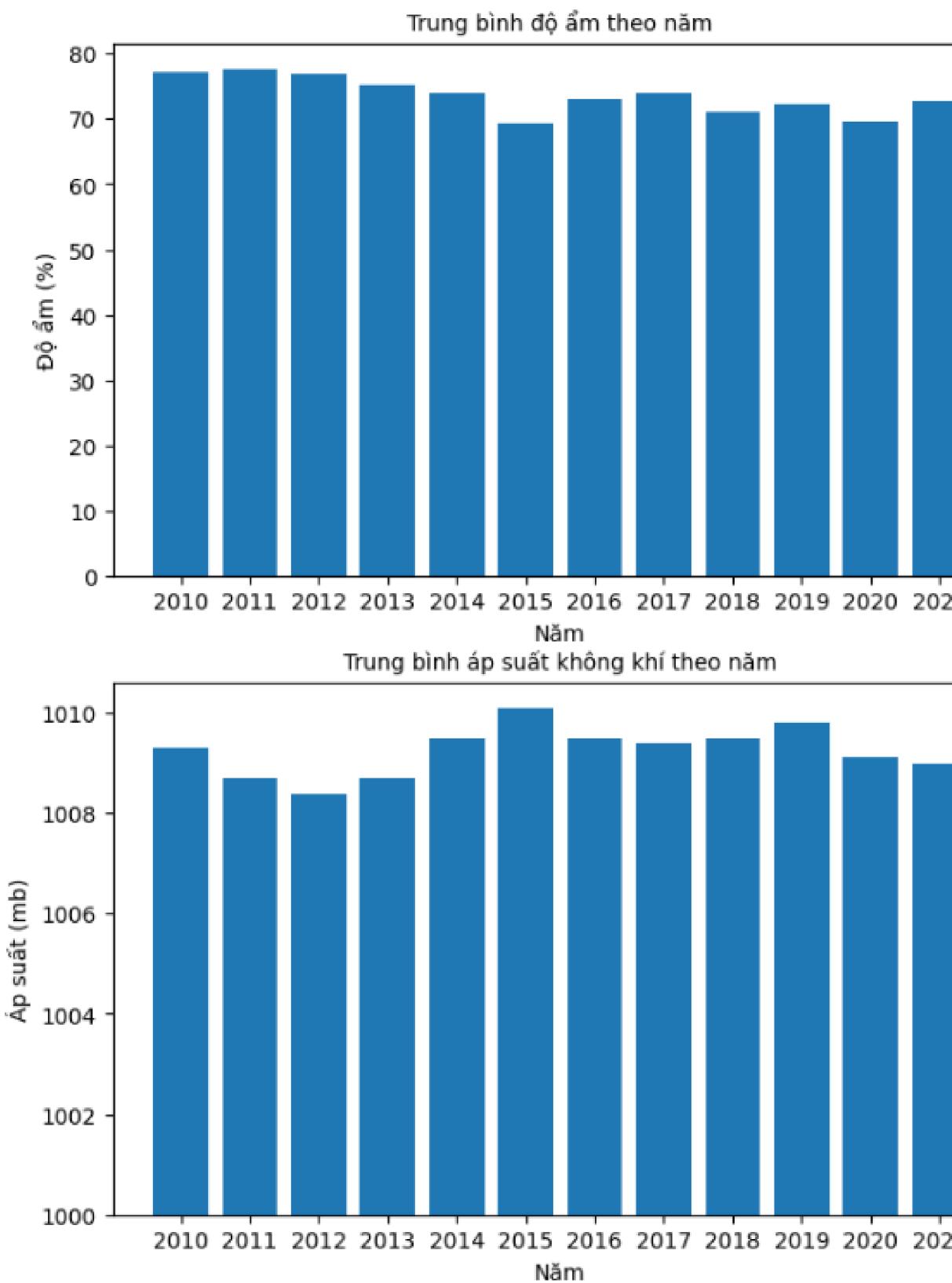
Ta thấy giá trị min và max của các attribute có sự chênh lệch

ĐẶT CÂU HỎI VÀ TRẢ LỜI BẰNG DỮ LIỆU

1. NHIỆT ĐỘ CAO NHẤT TRONG NGÀY, NHIỆT ĐỘ THẤP NHẤT TRONG NGÀY, TỐC ĐỘ GIÓ, LƯỢNG MƯA, ĐỘ ẨM, ĐỘ CHE PHỦ CỦA MÂY, ÁP SUẤT KHÔNG KHÍ TRUNG BÌNH THEO TỪNG NĂM THAY ĐỔI NHƯ THẾ NÀO?



1. NHIỆT ĐỘ CAO NHẤT TRONG NGÀY, NHIỆT ĐỘ THẤP NHẤT TRONG NGÀY, TỐC ĐỘ GIÓ, LƯỢNG MƯA, ĐỘ ẨM, ĐỘ CHE PHỦ CỦA MÂY, ÁP SUẤT KHÔNG KHÍ TRUNG BÌNH THEO TỪNG NĂM THAY ĐỔI NHƯ THẾ NÀO?



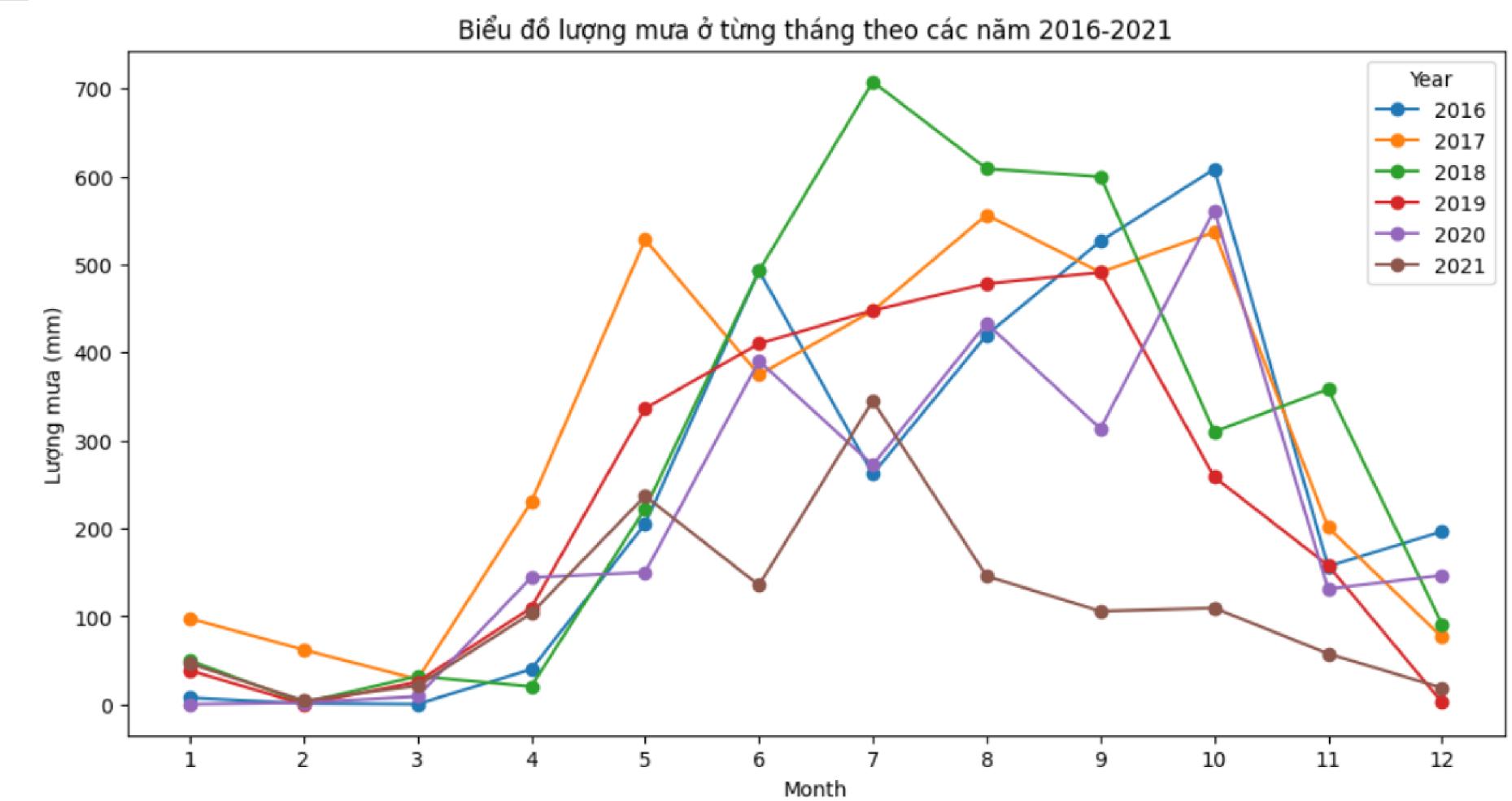
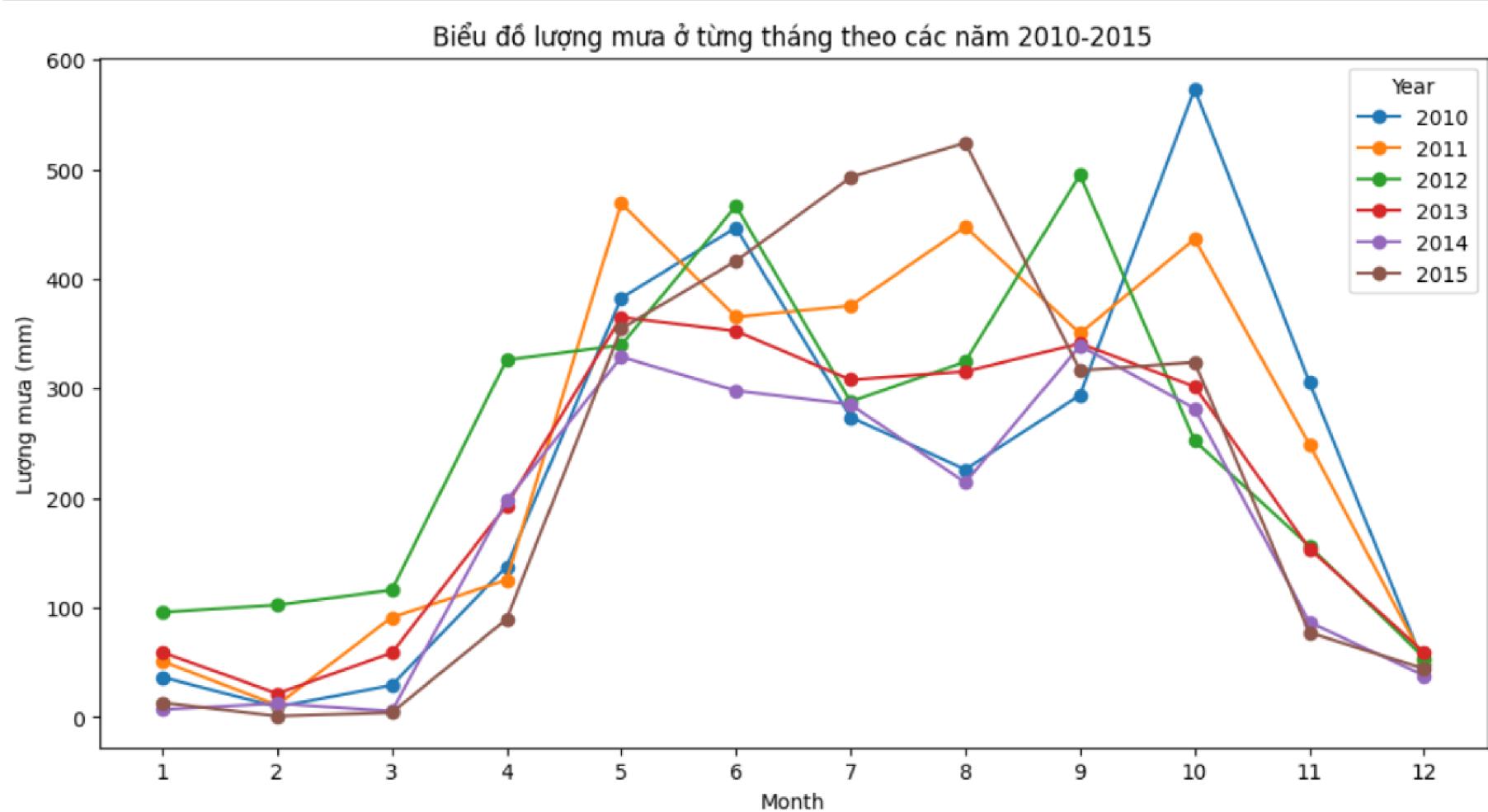
1. NHIỆT ĐỘ CAO NHẤT TRONG NGÀY, NHIỆT ĐỘ THẤP NHẤT TRONG NGÀY, TỐC ĐỘ GIÓ, LƯỢNG MƯA, ĐỘ ẨM, ĐỘ CHE PHỦ CỦA MÂY, ÁP SUẤT KHÔNG KHÍ TRUNG BÌNH THEO TỪNG NĂM THAY ĐỔI NHƯ THẾ NÀO?

- Ta có thể thấy trung bình nhiệt độ cao nhất/thấp nhất trong ngày, độ ẩm, áp suất không có biến động nhiều qua các năm.
- Tốc độ gió có xu hướng tăng dần và cao nhất vào năm 2015.
- Lượng mưa biến động thất thường, năm thấp nhất là 2021 lượng mưa không bằng một nửa của 2015.
- Độ che phủ của mây có sự tăng đột ngột vào năm 2020 và 2021.

Những phân tích trên nhằm để có cái nhìn tổng quát về sự thay đổi của những chỉ số thời tiết qua các năm.

2. LƯỢNG MƯA CỦA TỪNG THÁNG TRONG CÁC NĂM BIẾN ĐỔI NHƯ THẾ NÀO?

Để dễ quan sát, nhóm chia thành 2 biểu đồ 2010-2015 và 2016-2021

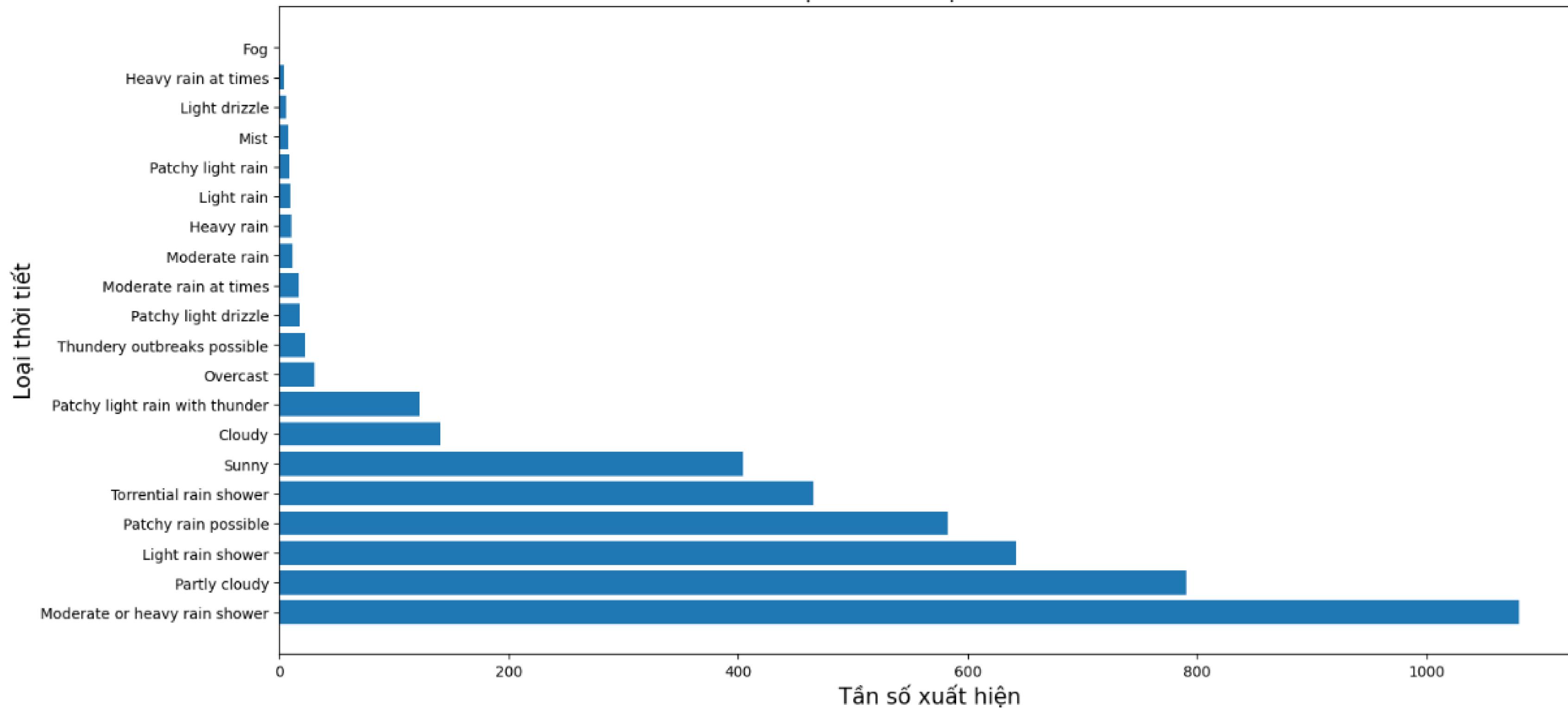


2. LƯỢNG MƯA CỦA TỪNG THÁNG TRONG CÁC NĂM BIẾN ĐỔI NHƯ THẾ NÀO?

- Nhận xét:
 - Tuy có sự khác biệt phần nào giữa các năm. Nhưng nhìn chung lượng mưa duy trì ở mức cao và đạt định trong các tháng từ 5-10.
 - Năm khác biệt nhất là 2021 khi lượng mưa chỉ ở mức cao vào tháng 5 và 7 và cũng tương đối thấp nếu so với cùng kỳ của các năm khác
- Lợi ích đạt được từ câu hỏi:
 - Ta có thể thấy rõ thời gian của mùa mưa và mùa khô ở Tp HCM
- Nguồn cảm hứng:
 - Có thể biết được vào những tháng nào thì ta nên nêu chuẩn bị tâm lý xe sẽ chết máy khi lội nước trên những con đường ở tp HCM chẳng hạn :v

3. LOẠI THỜI TIẾT NÀO THƯỜNG XUYÊN XUẤT HIỆN NHẤT Ở TP HCM?

Tần số xuất hiện của các loại thời tiết từ 2010-2021



3. LOẠI THỜI TIẾT NÀO THƯỜNG XUYÊN XUẤT HIỆN NHẤT Ở TP HCM?

Nhận xét:

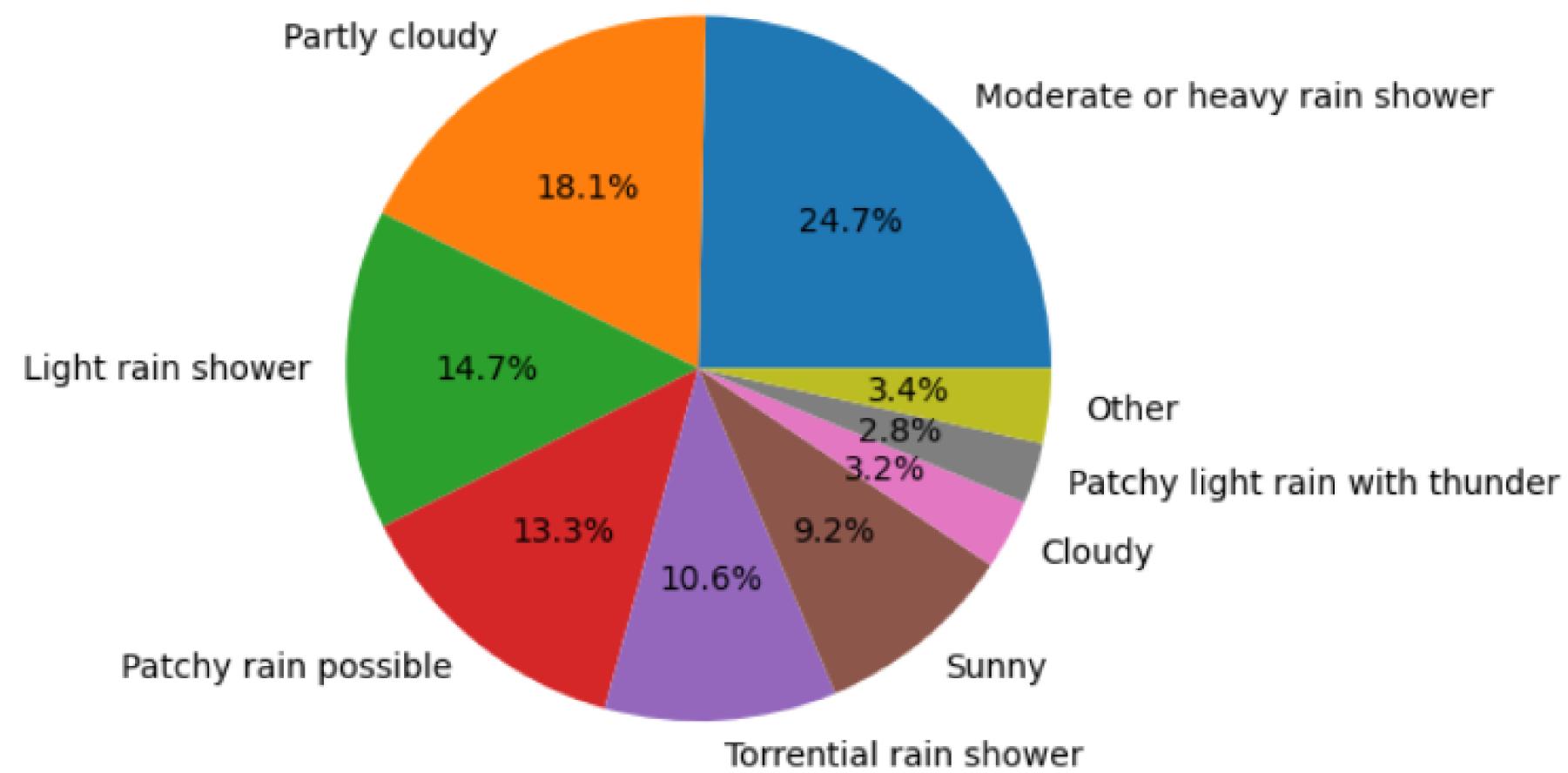
- Có 6 kiểu thời tiết phổ biến là: Moderate or heavy rain shower, Partly cloudy, Light rain shower, Patchy rain possible, Torrential rain shower, Sunny chiếm hơn 90%
- Theo quan sát kiểu thời tiết thì thường như hầu hết không có mưa cũng sẽ có mây

Lợi ích đạt được từ câu hỏi:

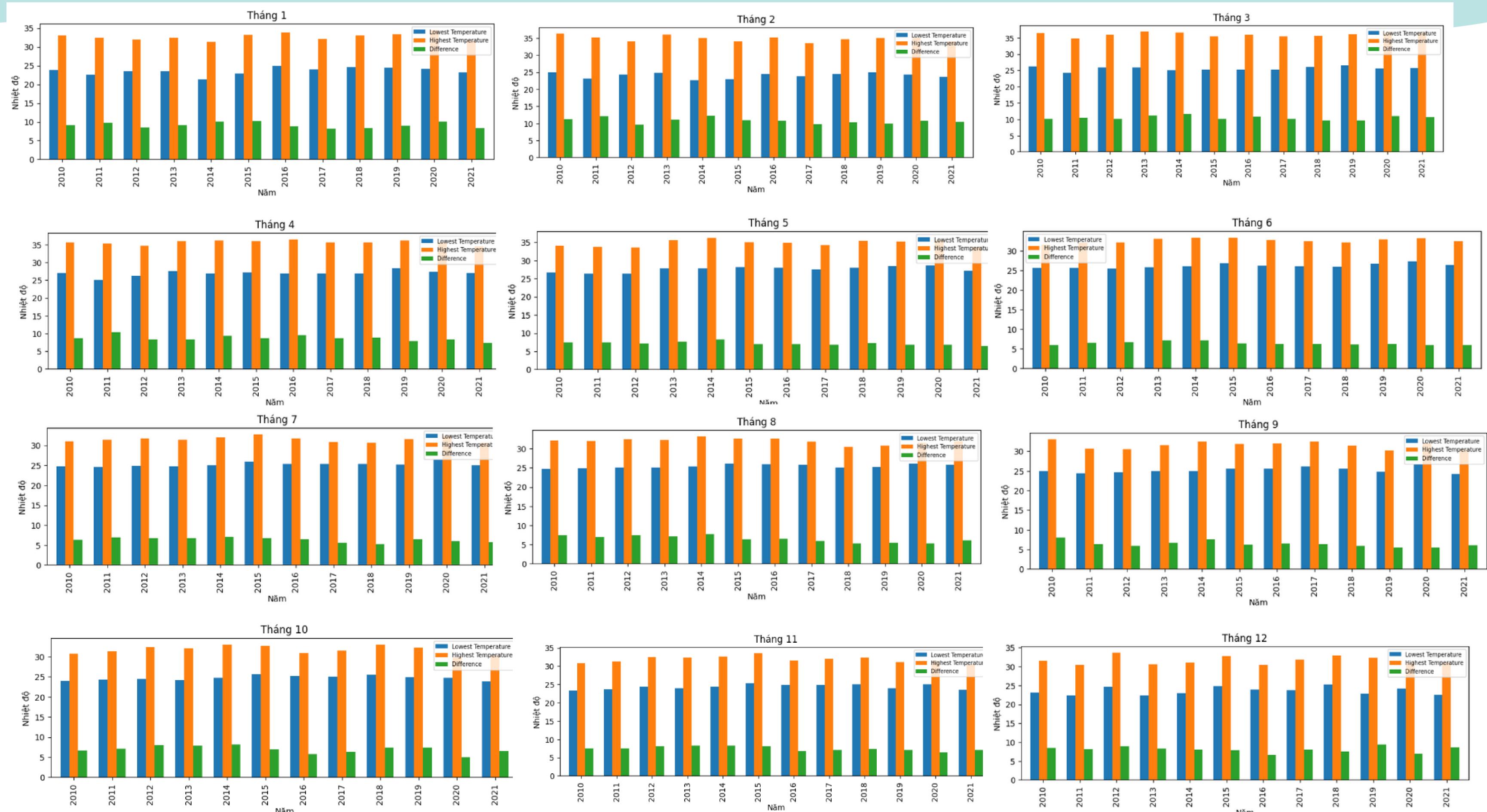
- Hiểu được kiểu thời tiết ở Tp HCM

Nguồn cảm hứng:

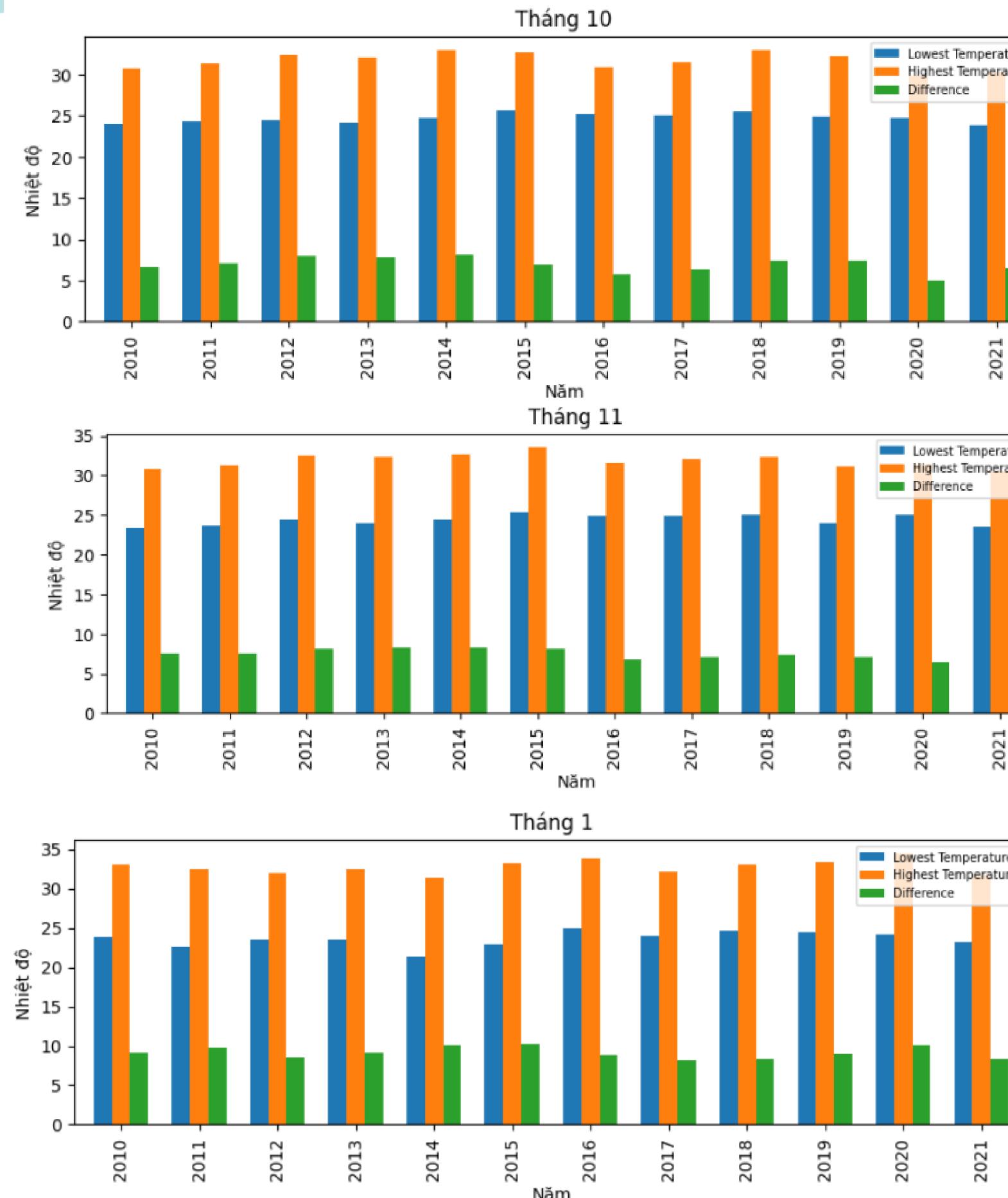
- Người ta thường nói Sài Gòn hay có những cơn mưa bất chợt



4. SO SÁNH SỰ CHÊNH LỆCH NHIỆT ĐỘ CAO NHẤT VÀ THẤP NHẤT

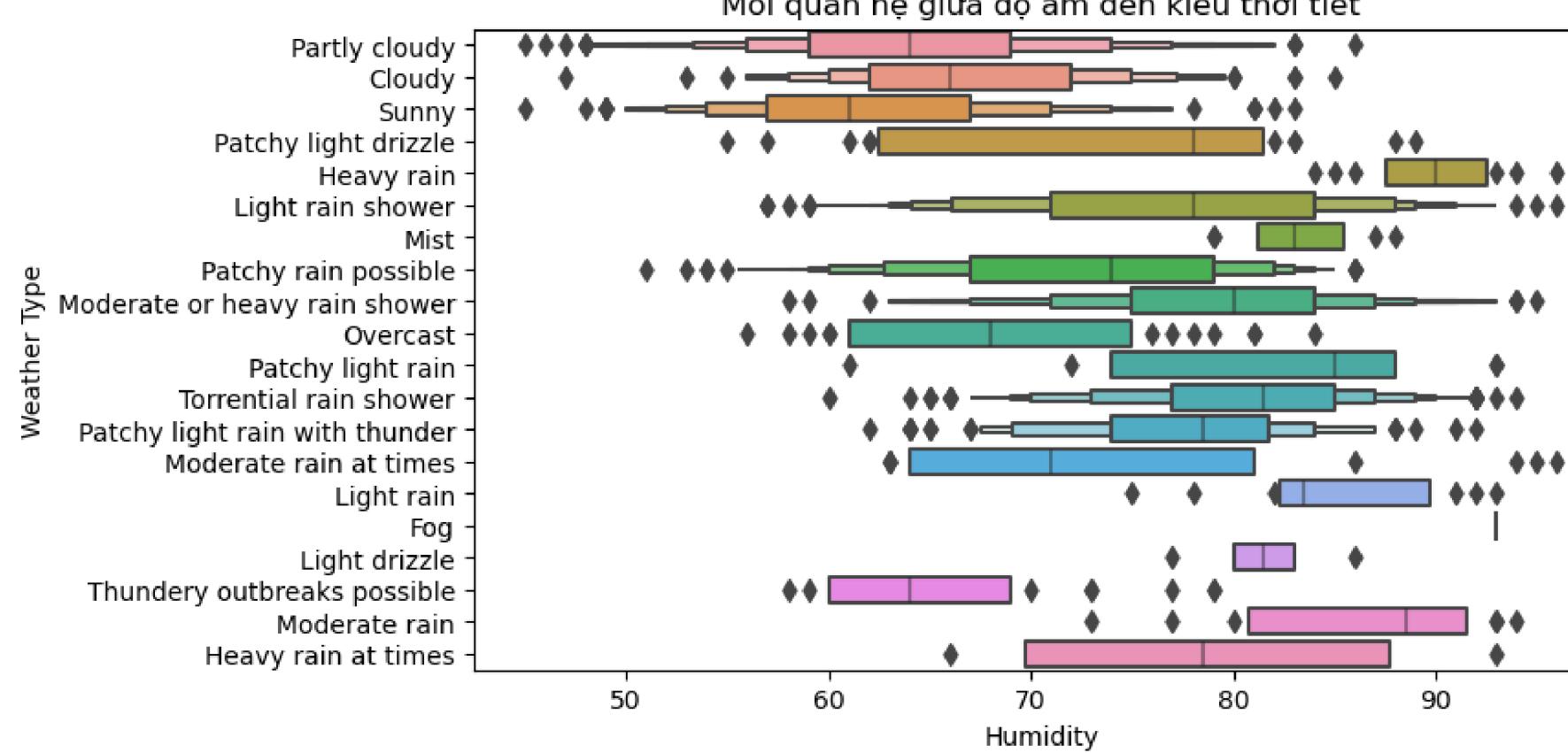
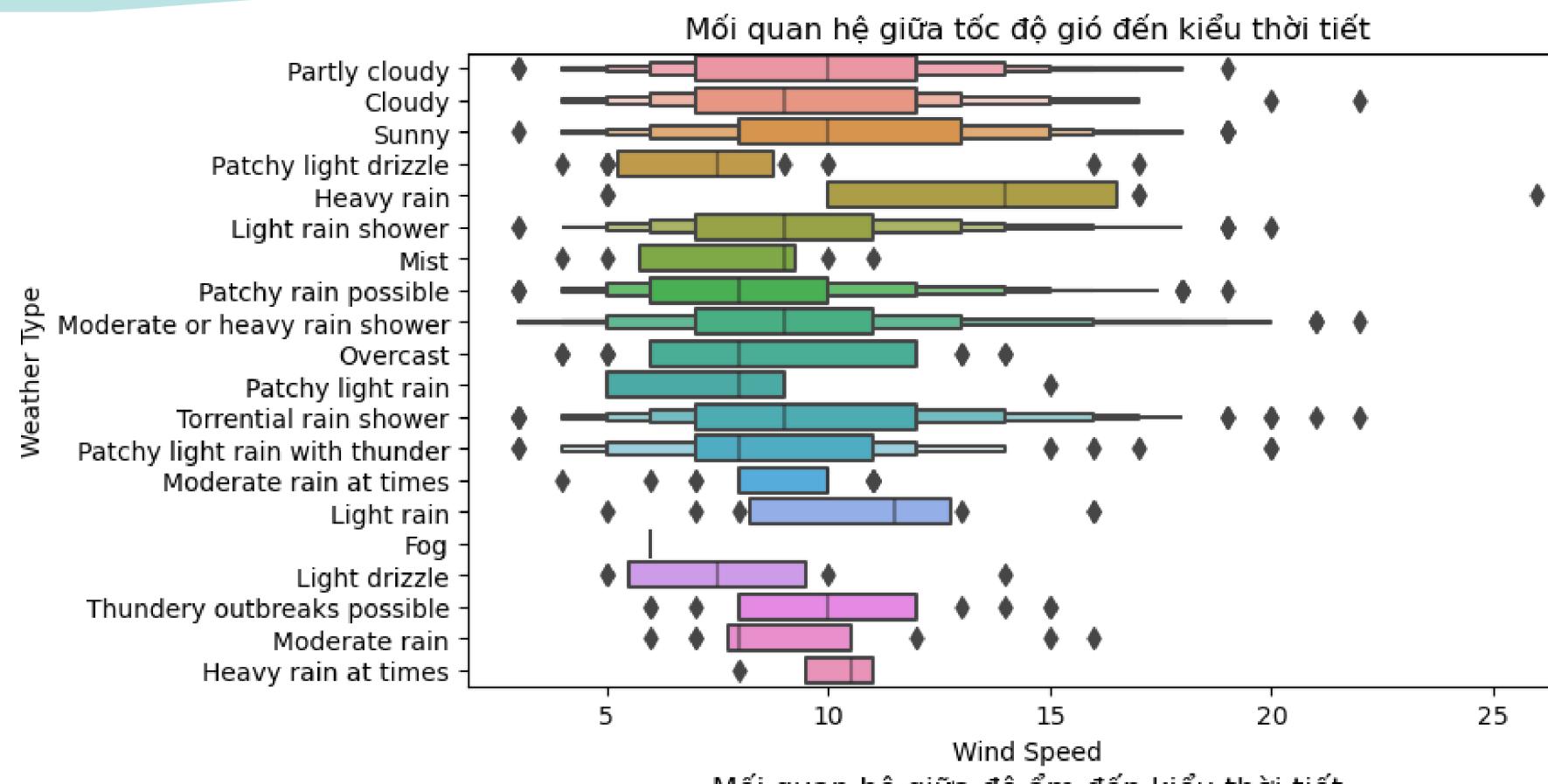


4. SO SÁNH SỰ CHÊNH LỆCH NHIỆT ĐỘ CAO NHẤT VÀ THẤP NHẤT



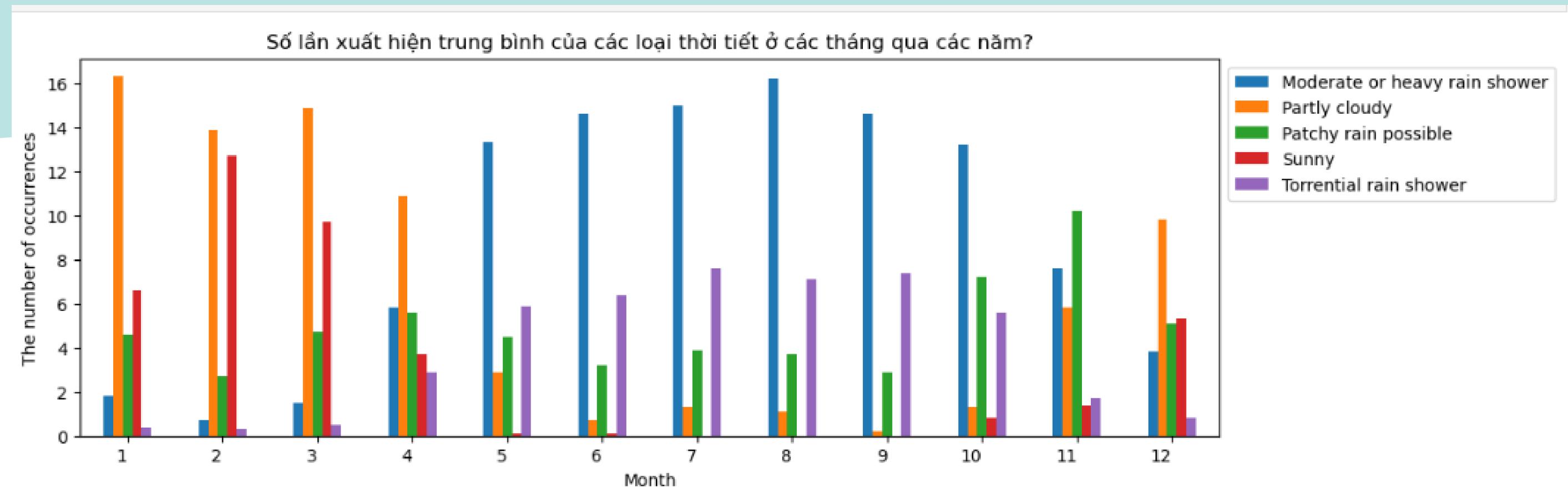
Ta có thể thấy sự biến động về nhiệt độ cao nhất và thấp nhất trong ngày giữa các năm lớn hơn vào ở các tháng cuối năm như 10,11,1.

5. MỐI LIÊN HỆ GIỮA TỐC ĐỘ GIÓ VÀ ĐỘ ẨM ĐẾN LOẠI THỜI TIẾT



Kết quả cho thấy có sự tương quan

6. SỐ LẦN XUẤT HIỆN TRUNG BÌNH CỦA CÁC LOẠI THỜI TIẾT Ở CÁC THÁNG QUA CÁC NĂM?

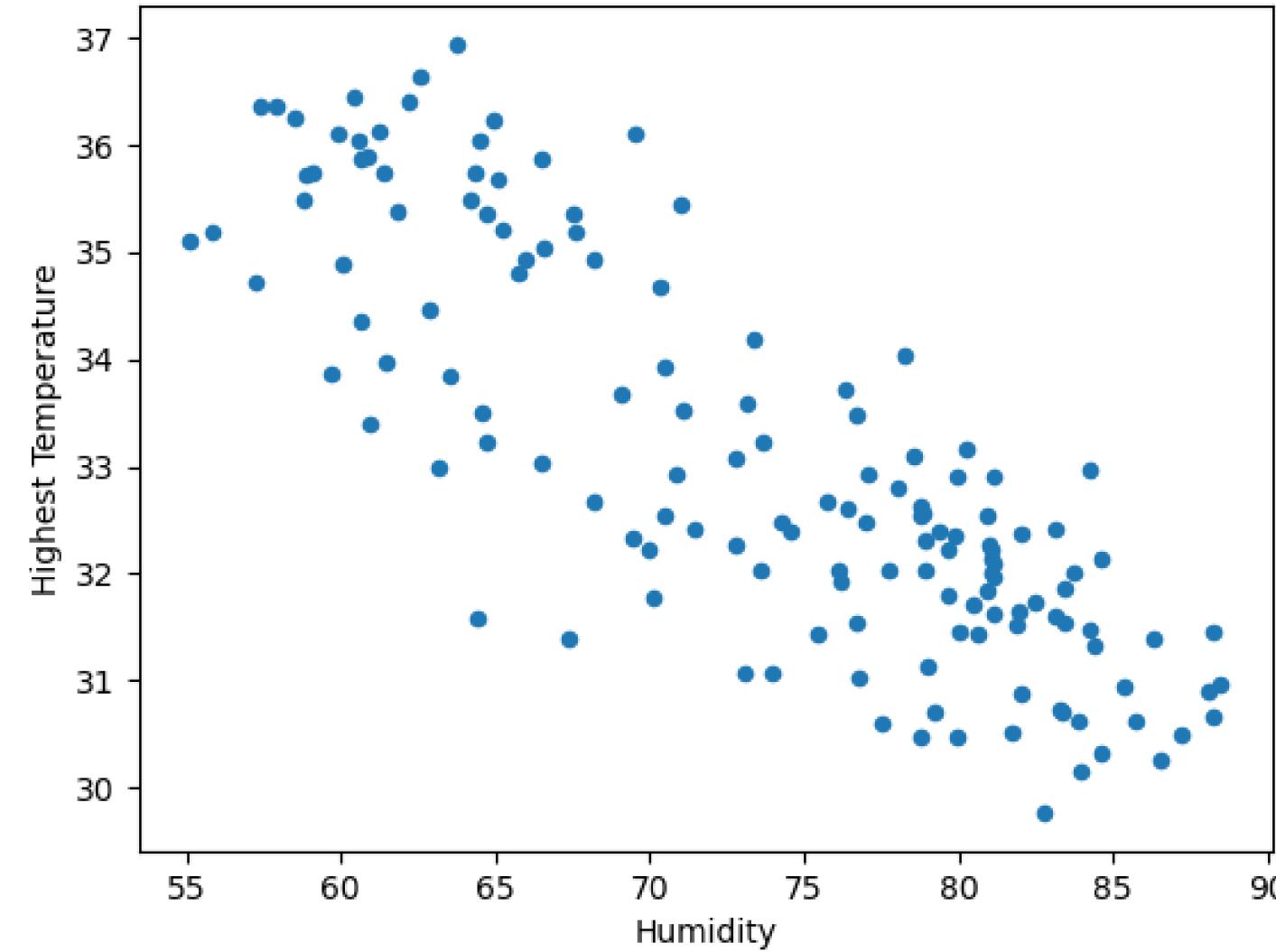


Ở đây em chỉ định các loại thời tiết là 'Partly cloudy','Moderate or heavy rain shower', 'Sunny', 'Torrential rain shower','Patchy rain possible'.

- Lợi ích từ câu hỏi: Biết được các tháng mà các loại thời tiết này thường xuất hiện mà ta có thể chủ động trong các tình huống như:
 - Lựa chọn việc di chuyển bằng các phương tiện công cộng
 - Có sự lưu ý về những tháng có nhiều mưa.
 - Chọn được những tháng có thời tiết đẹp để đi dạo phố.
- Như biểu đồ ta có thể thấy là những tháng có nhiều ngày đẹp trời thường rơi vào tháng 1 đến tháng 3. Đó là khoảng thời gian Tết của Việt Nam, rất thích hợp cho việc đi chơi!

7. SỰ TƯƠNG QUAN GIỮA NHIỆT ĐỘ CAO NHẤT VÀ ĐỘ ẨM TRUNG BÌNH Ở CÁC THÁNG QUA CÁC NĂM

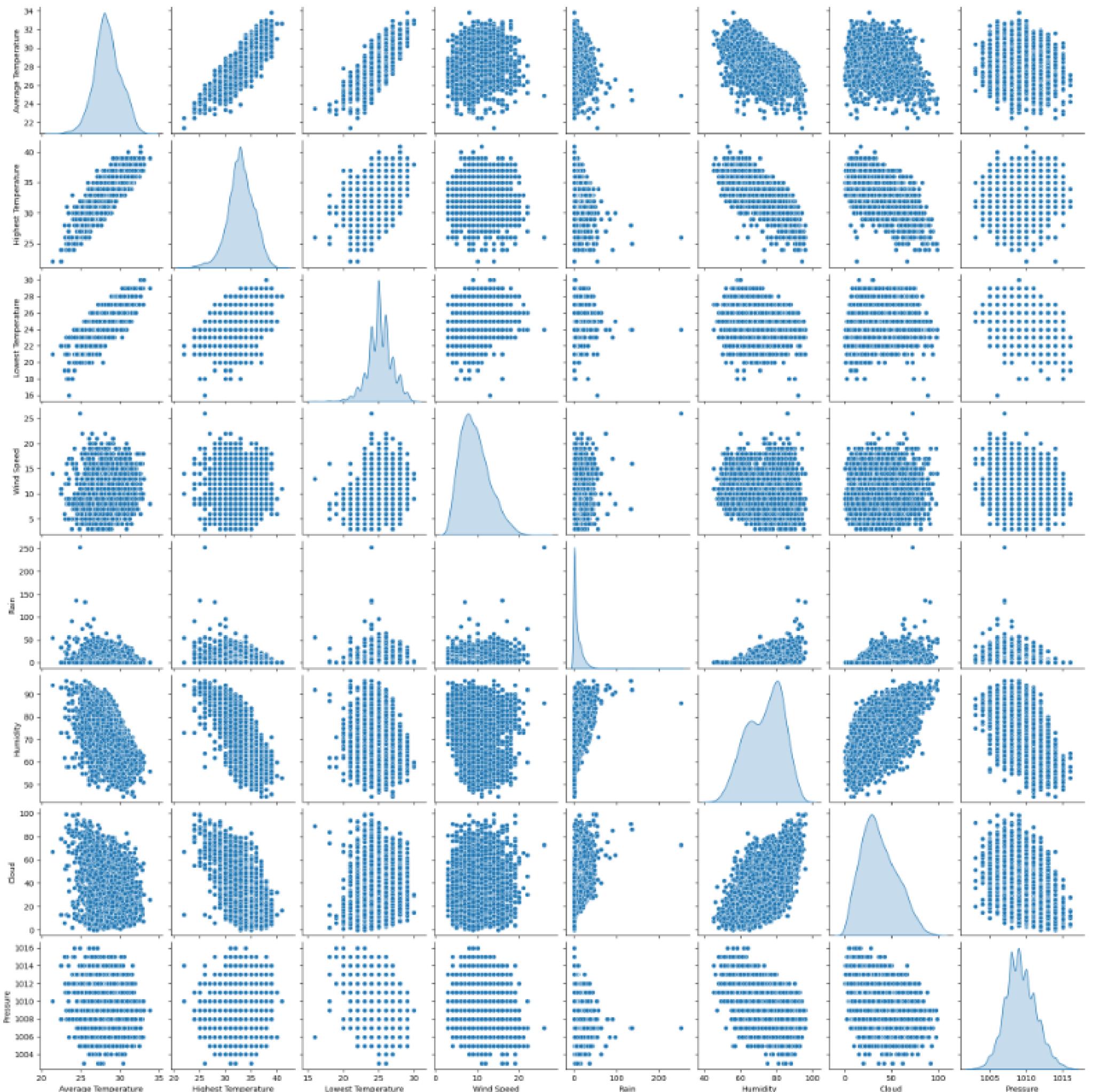
Sự tương quan giữa Nhiệt độ cao nhất và Độ ẩm trung bình ở các tháng qua các năm



Từ biểu đồ ta có thể thấy Khi nhiệt độ tăng đồng nghĩa tốc độ bay hơi của nước ngày càng nhanh, độ ẩm trong không khí sẽ giảm và ngược lại. Đây là hiện tượng quen thuộc vào những ngày nóng nực mùa hè.

KHÁM PHÁ VÀ TIỀN XỬ LÝ DỮ LIỆU CHO MÔ HÌNH HỌC MÁY

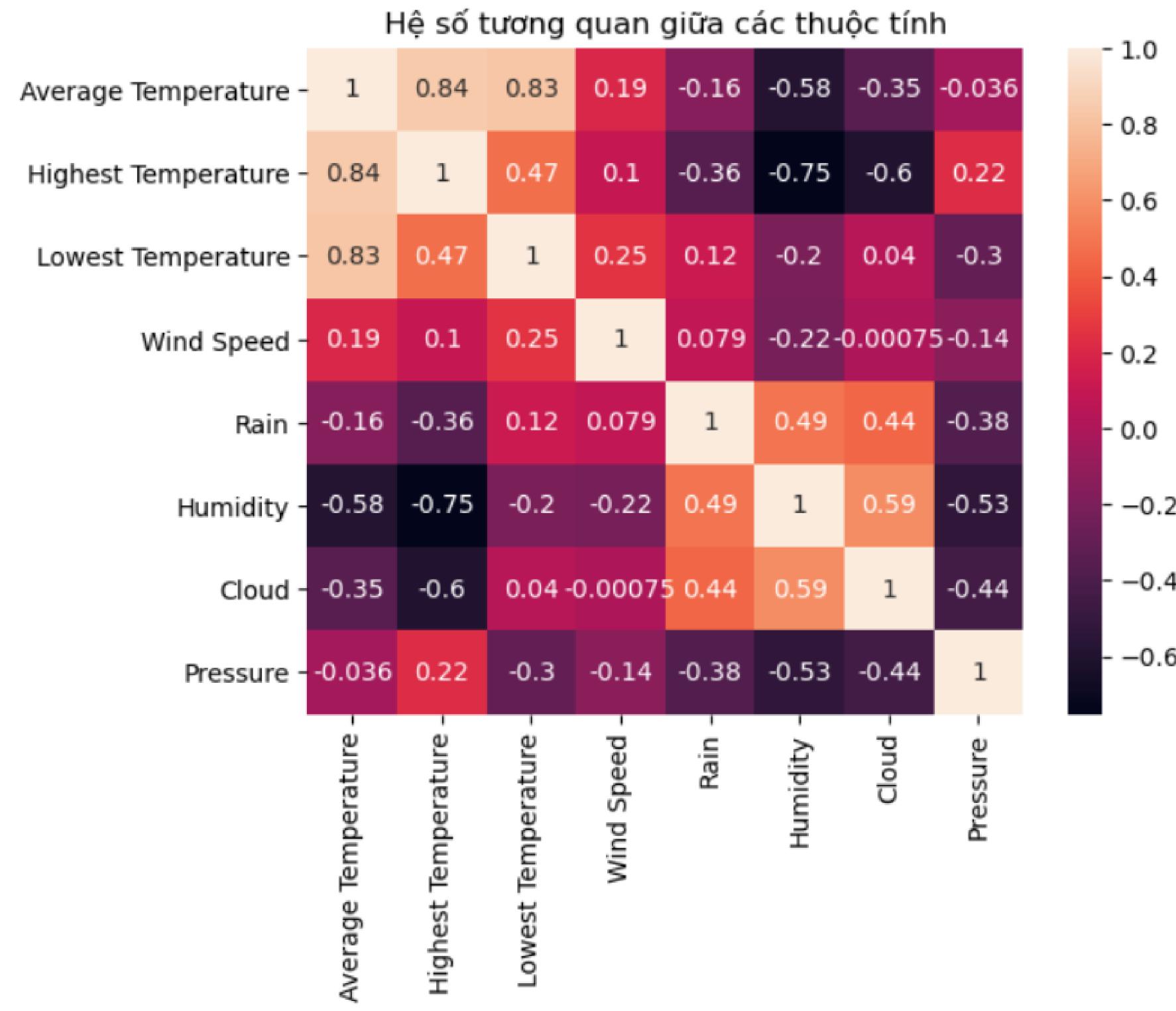
vẽ pairplot để xem
phân bố của từng cặp
thuộc tính



Từ pairplot trên, ta thấy có 1 vài thuộc tính có dữ liệu là phân phối lệch

KHÁM PHÁ VÀ TIỀN XỬ LÝ DỮ LIỆU CHO MÔ HÌNH HỌC MÁY

Vẽ heatmap để xem hệ số tương quan giữa các cặp thuộc tính

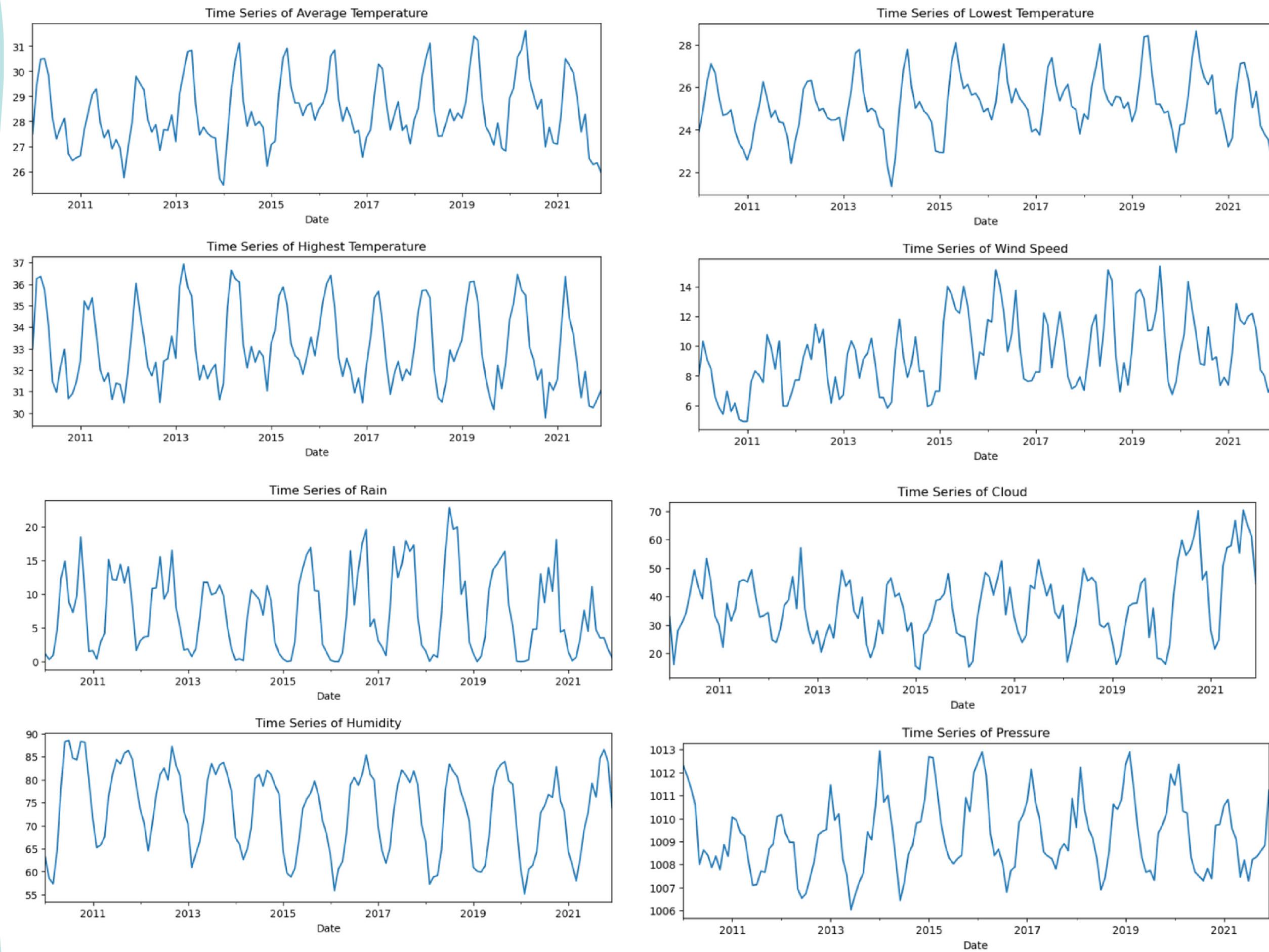


Từ heatmap trên, ta có thể thấy:

- "Highest Temperature", "Lowest Temperature", "Humidity" là 3 thuộc tính có độ lớn của giá trị tương quan lớn nhất với "Average Temperature"
- "Pressure" có độ lớn của giá trị tương quan với "Average Temperature" rất nhỏ nhưng lại có độ lớn của giá trị tương quan với "Humidity" lớn

KHÁM PHÁ VÀ TIỀN XỬ LÝ DỮ LIỆU CHO MÔ HÌNH HỌC MÁY

vẽ lineplot để xem
timeseries của từng
thuộc tính theo tháng
qua từng năm



KHÁM PHÁ VÀ TIỀN XỬ LÝ DỮ LIỆU CHO MÔ HÌNH HỌC MÁY

Tạo dataset cho
training và testing
của Time Series

- Ta sẽ sử dụng N ngày trước đây để dự đoán nhiệt độ ngày hôm nay. Hàm `extract_n_past_days` dùng để tạo ra dataframe đấy. Ở đây, ta sẽ dùng N=3.
- Dữ liệu thời tiết từ 2010-2021 gồm 12 năm. Ta chia bộ dữ liệu thành 10 năm đầu cho training và 2 năm cuối cho testing.

KHÁM PHÁ VÀ TIỀN XỬ LÝ DỮ LIỆU CHO MÔ HÌNH HỌC MÁY

Xử lí cột có kiểu dữ liệu phân loại, xử lí phân phối lệch và chuẩn hóa dữ liệu.

Bây giờ ta sẽ xử lí các vấn đề được nêu ở trên của bộ dữ liệu, bao gồm:

- Chuẩn hóa dữ liệu: dùng StandardScaler của thư viện sklearn.
- Xử lí cột có kiểu dữ liệu phân loại thành dạng one hot: dùng OneHotEncoder của thư viện sklearn.
- Xử lí phân phối lệch: Viết custom class HandleSkewedDistribution để xử lí.
- Ta sẽ dùng Pipeline và ColumnTransformer của sklearn để tạo pipeline xử lí cả 3 vấn đề trên.

MÔ HÌNH HÓA DỮ LIỆU

MÔ HÌNH HÓA DỮ LIỆU

- Dữ liệu thời tiết trong 12 năm tách 10 năm cho train và 2 năm cho test.
- Các mô hình sử dụng Linear Regression, SVR, XGBoostRegressor.
- Mô hình đạt kết quả tốt nhất là mô hình SVR với R2 Score = 0.748 và RMSE=0.99

	LR	SVR	XGBR
r2_score_train	0.822	0.835	0.874
r2_score_test	0.739	0.748	0.746
mse_test	1.018	0.980	0.988
rmse_test	1.009	0.990	0.994

ĐÁNH GIÁ MÔ HÌNH

NHẬN XÉT

Nhìn chung cả 3 mô hình ở trên đều cho ra các kết quả xấp xỉ nhau dựa trên các độ đo như r2 score, mse, rmse.

Nếu được deploy thì mô hình Linear Regression sẽ được chọn vì mô hình vừa đơn giản và vừa có kết quả tốt như mô hình phức tạp khác.



CẢM ƠN THẦY VÀ CÁC
BẠN ĐÃ LẮNG NGHE

