Phân tích dữ liệu

I. Về nguồn dữ liệu

Bộ dữ liệu được lấy từ trang web Estesparkweather.net. Estesparkweather.net cung cấp thông tin chi tiết về thời tiết theo các tháng, thời tiết dài hạn cũng như có bộ database chứa thông tin thời tiết ở quá khứ của thành phố Estes Park từ năm 2010 cho đến nay (14 năm).

May 24 Average and Extremes	
Average temperature	48.7°F
Average humidity	30%
Average dewpoint	17.4°F
Average barometer	29.6 in.
Average windspeed	12.2 mph
Average gustspeed	18.2 mph
Average direction	245° (WSW)
Rainfall for month	0.37 in.
Rainfall for year	4.30 in.
Maximum rain per minute	0.00 in. on day 24 at time 23:59
Maximum temperature	60.9°F on day 24 at time 15:52
Minimum temperature	38.3°F on day 24 at time 02:37
Maximum humidity	49% on day 24 at time 23:59
Minimum humidity	16% on day 24 at time 15:25
Maximum pressure	29.666 in. on day 24 at time 12:16
Minimum pressure	29.545 in. on day 24 at time 04:11
Maximum windspeed	29.9 mph on day 24 at time 05:06
Maximum gust speed	38.0 mph from 267 °(W) on day 24 at time 05:04
Maximum heat index	60.9°F on day 24 at time 15:52

24 Hour Graph of this day is not available (20240524.gif)

May 25 Average and Extremes	
Average temperature	45.9°F
Average humidity	58%
Average dewpoint	30.0°F
Average barometer	29.5 in.

Dữ liệu thu thập được gồm 20 cột :

1. date : Ngày /tháng / năm
2. avg_temp: Nhiệt độ trung bình
3. avg_humidity: Độ ẩm trung bình
4. avg_dewpoint: Nhiệt độ điểm sương trung bình
5. avg_barometer: Áp suất không khí trung bình
6. avg_winspeed: Tốc độ gió trung bình
7. avg_gustspeed: Tốc độ gió giật trung bình
8. avg_direction: Hướng gió trung bình
9. month_rainfall: Lượng mưa trong tháng
10. year_rainfall: Lượng mưa trong năm
11. max_rain_per_minute: Lượng mưa lớn nhất trong 1 phút
12. max temp: Nhiệt độ cao nhất trong ngày
13. min_temp: Nhiệt độ thấp nhất ngày
14. max_humidity: Độ ẩm lớn nhất ngày
15. min_humidity: Độ ẩm thấp nhất trong ngày
16. max_pressure: Áp suất không khí lớn nhất trong ngày
17. min_pressure: Áp suất không khí nhỏ nhất trong ngày
18. max_windspeed: Tốc độ gió lớn nhất trong ngày
19. max_gustspeed: Tốc độ gió giật lớn nhất trong ngày
20. max_heat_index: Chỉ số nóng bức tối đa

Những đặc trưng trên có đặc điểm thống kê được mô tả ở bảng sau:

8 rows × 19 columns pd.DataFrame

CSV

	avg_temp	avg_humidity	avg_dewpoint	avg_barometer	avg_windspeed	avg_gustspeed	avg_direction
count	4923.000000	4923.000000	4923.000000	4923.000000	4923.000000	4923.000000	4923.00000
mean	44.131018	50.082876	23.754012	29.868068	5.564249	9.002722	213.96506
std	15.446007	17.594224	13.894416	0.251392	3.893761	10.416759	93.94868
min	0.100000	9.000000	0.000000	28.200000	0.000000	0.000000	0.00000
25%	32.500000	37.000000	12.300000	29.700000	2.600000	4.300000	118.50000
50%	44.500000	49.000000	22.400000	29.800000	4.300000	6.800000	246.00000
75%	57.600000	62.000000	35.500000	30.000000	7.700000	11.600000	281.00000
max	76.300000	94.000000	55.100000	31.000000	22.300000	240.400000	360.00000

II. Xử lý dữ liệu

- Chuyển dữ liệu ngày tháng năm trong trường date về dạng datetime.

- Chuyển các đại lượng về nhiệt độ đang để ở đơn vị °F về đơn vị °C

- Chuyển các đại lượng về lượng mưa từ đơn vị in sang mm

- Từ giá trị của trường year_rainfall tính giá trị cho một trường mới day_rainfall

- Từ giá trị trường mới day_rainfall vừa tìm được ta thêm vào bộ dữ liệu một trường mới là weather_code (Với 2 nhãn là “Rain” và “No Rain”)

- Thêm trường weather_code vào bộ dữ liệu.

- Lấy mẫu trung bình theo tháng.

Kỹ thuật này thường áp dụng cho việc phân tích dữ liệu theo thời gian. Lấy mẫu trung bình theo tháng/ năm giúp ta nhận biết được xu hướng của đặc trưng trong bài toán và loại bỏ các nhiễu khi phân tích.

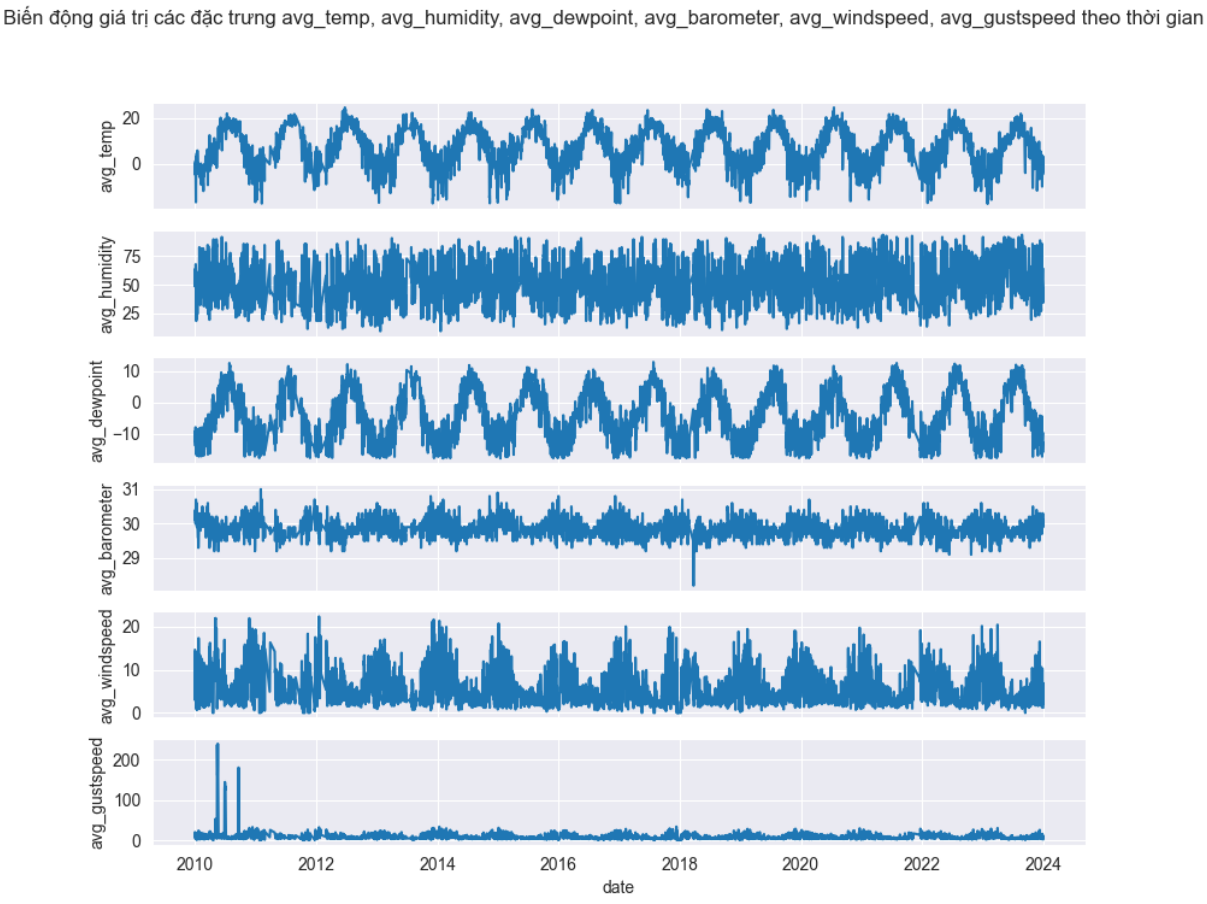
1-100 » 173 rows × 2 columns pd.DataFrame

date	avg_temp	day_rainfall
2010-01-31	-2.043871	0.057097
2010-02-28	-4.587500	0.453571
2010-03-31	0.996129	0.597742
2010-04-30	3.810667	2.082000
2010-05-31	7.465000	1.439000
2010-06-30	15.063667	1.829000
2010-07-31	17.631333	2.056667
2010-08-31	16.882258	0.892258
2010-09-30	15.053103	0.148621
2010-10-31	7.827097	0.867742

III. Trực quan hóa dữ liệu

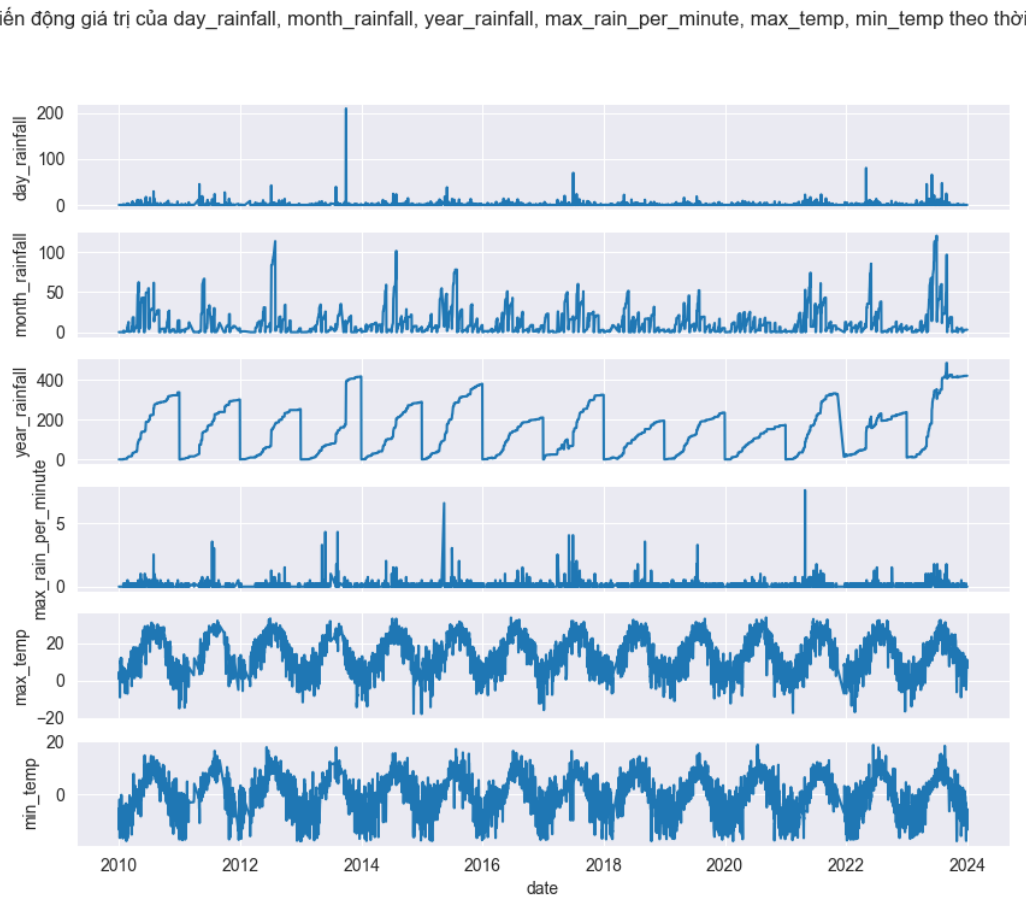
1. Phân tích về độ biến động của các đặc trưng theo thời gian

Cách phân tích này dùng để phân loại các nhóm đặc trưng ảnh hưởng theo địa lý; đặc trưng ảnh hưởng theo thời gian (ngày, tháng và năm)



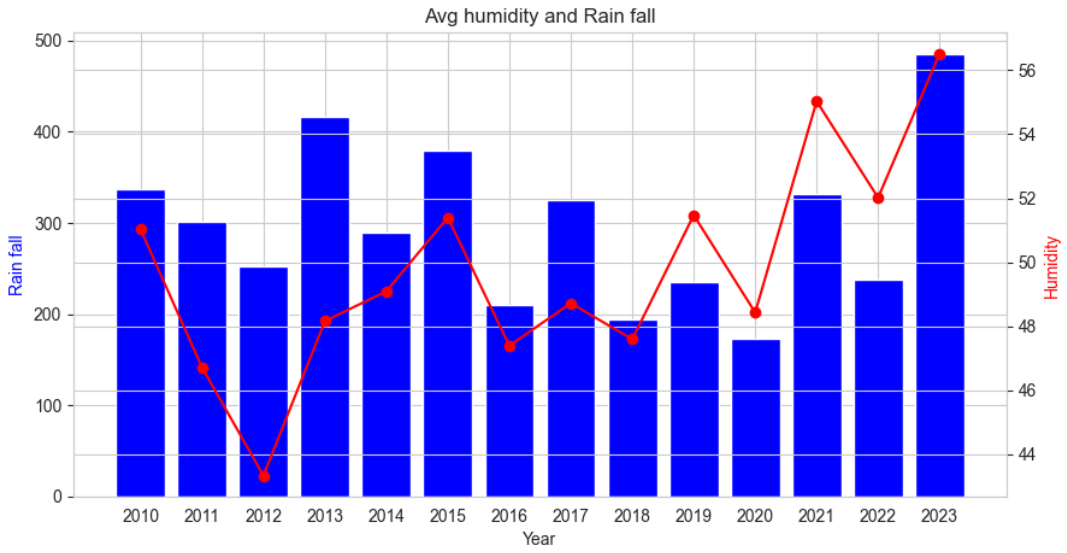
Hình 1: Biến động giá trị các đặc trưng avg_temp, avg_humidity ,avg_dewpoint, avg_barometer, avg_winspeed, avg_gustspeed theo thời gian'

Quan sát: Những đặc trưng khí tượng như avg_temp, avg_dewpoint, avg_baro meter, avg_winspeed có xu hướng đối xứng.Đây là những đặc trưng diễn ra theo chu kỳ trong năm, do đó ít trơn mịn hơn.Với avg_humidity thì khá dày và biên độ đều, không theo xu hướng. Avg_gustspeed khá mịn, có tính chu kì năm



Hình 2: Biến động giá trị của month_rainfall, year_rainfall, max_rain_per_minute, max_temp, min_temp, uv theo thời gian

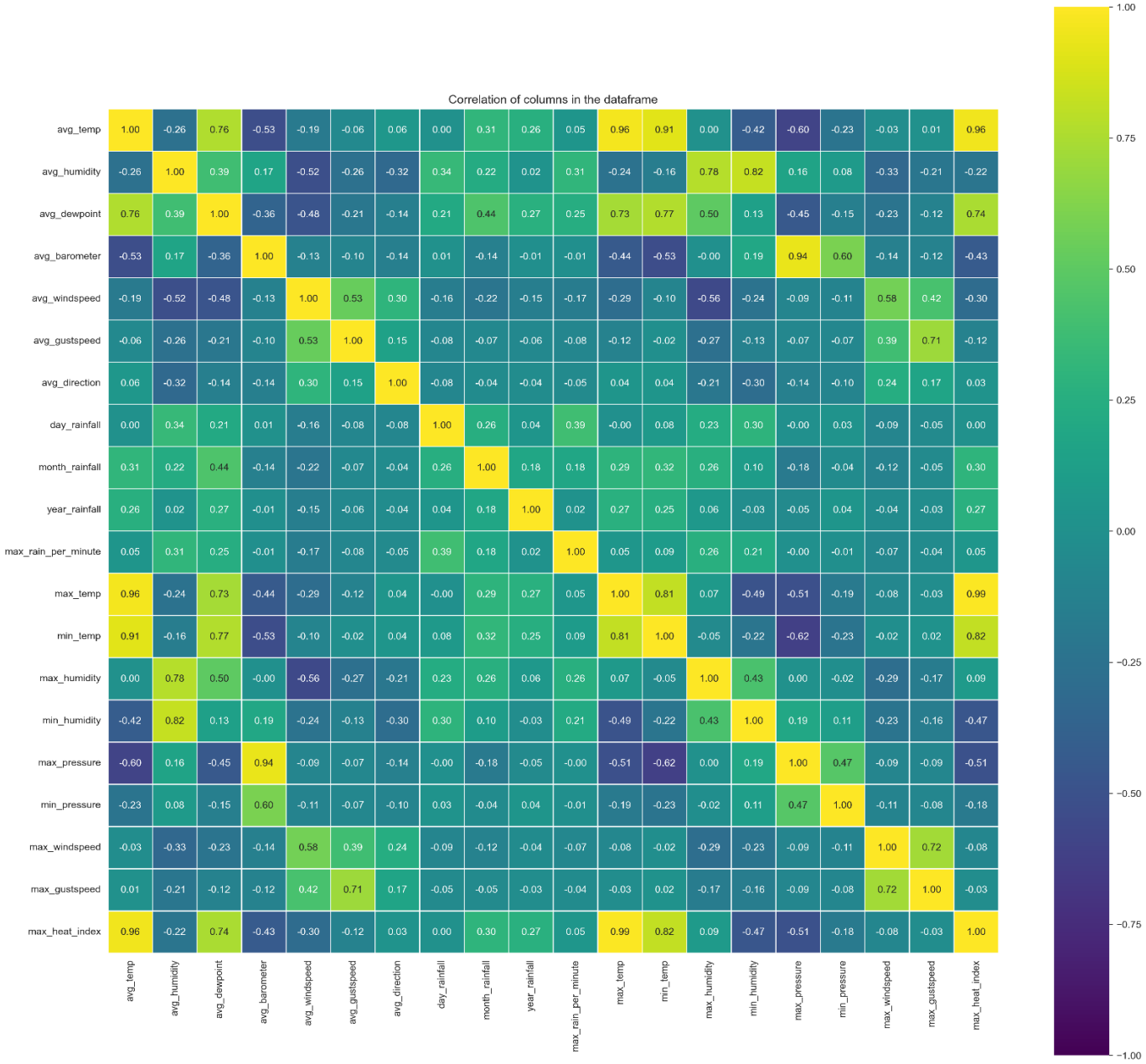
Quan sát: Những đặc trưng khí tượng như max_temp, min_temp, year_rainfall, có xu hướng đối xứng. Đây là những đặc trưng diễn ra theo chu kỳ trong năm, do đó ít trơn mịn hơn. Day_rainfall khá mịn khi xét trên khoảng thời gian 13 năm



Hình 3: Biểu đồ tương quan lượng mưa và độ ẩm

Quan sát: Từ quan sát trên ta thấy được sự tương quan giữa lượng mưa trong năm và độ ẩm trung bình của năm đó . Ta thấy đa số các năm khi độ ẩm tăng thì kèm theo lượng mưa trong năm đó cũng tăng và ngược lại

2. Phân tích mối quan hệ giữa các đặc trưng



Hình 4: Biểu đồ nhiệt

Các cặp đặc trưng tỉ lệ nghịch với nhau lớn <-0,45:

- avg_temp & avg_barometer, avg_temp & max_pressure
- avg_humidity & avg_winspeed, avg_dewpoint & avg_winspeed
- avg_barometer & avg_mintemp, avg_barometer & avg_winspeed

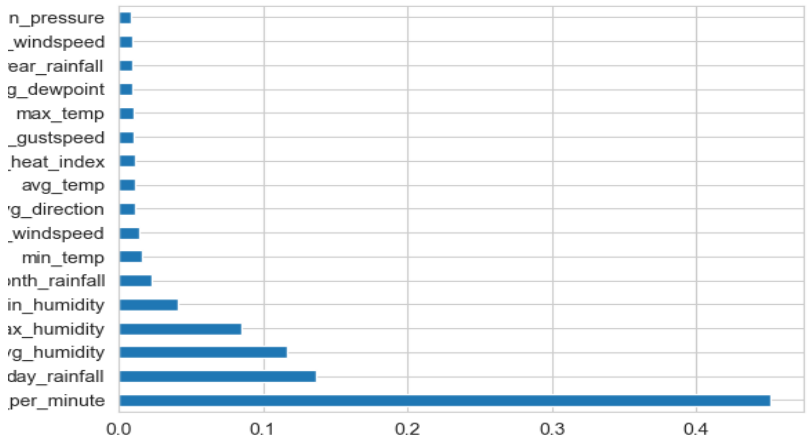
- avg_winspeed & max_humidity, max_temp & max_pressure,max_temp &min_humidity
-

Các đặc trưng tỉ lệ thuận với nhau > 0,45:

- avg_temp & max_temp, avg _temp & min_temp ,avg_temp & avg_dewpoint , avg_temp & max_heat_index
- avg_humidity & min_humidity , avg_humidity & max_humidity, ,…

Từ đó, một số đặc trưng ít liên quan đến như 'avg_direction', 'day_rainfall', 'month_rainfall','year_rainfall','max_rain_per_minute' trở nên thành những đặc trưng quan trọng vì tính độc lập của chúng.

Ta kiểm nghiệm độ quan trọng của các đặc trưng sử dụng tree_based classifiers



Hình 5: Độ quan trọng các đặc trưng thời tiết

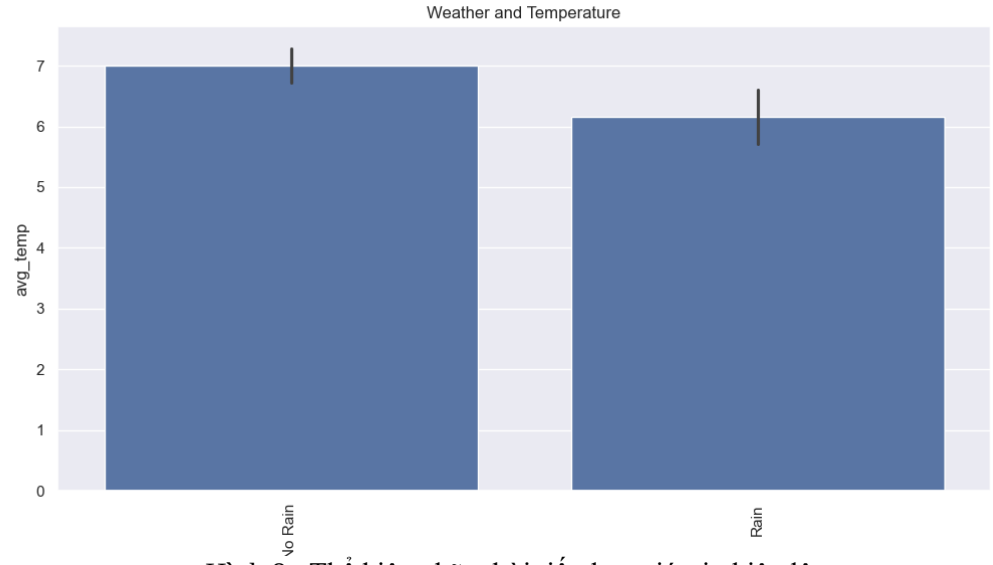
Ta thấy xếp hạng độ quan trọng từ max_rain_per_minute> day_rainfall>humidity>.... Biểu đồ nhiệt giúp ta chỉ ra được những đặc trưng quan trọng một cách trực quan.

3. Phân tích tương quan nhãn dữ liệu và các đặc trưng

	Weather	Count
0	No Rain	3411
1	Rain	1512

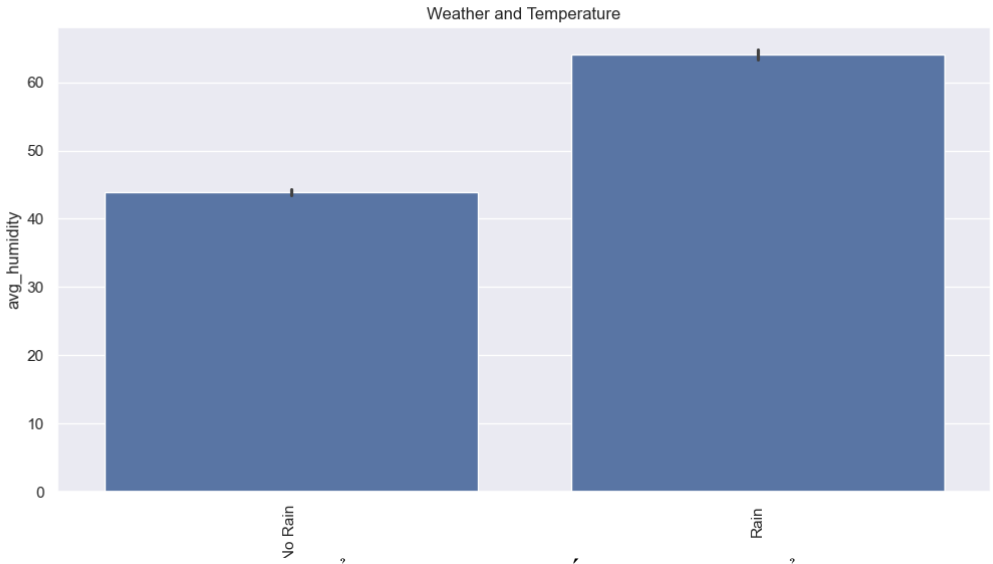
Hình 7: Số lượng từng nhãn thời tiết.

Chúng ta có 2 nhãn thời tiết, số lượng mẫu 'Rain' chiếm khoảng 30,7% còn lại là mẫu "No Rain" Chiếm khoảng 69.3%



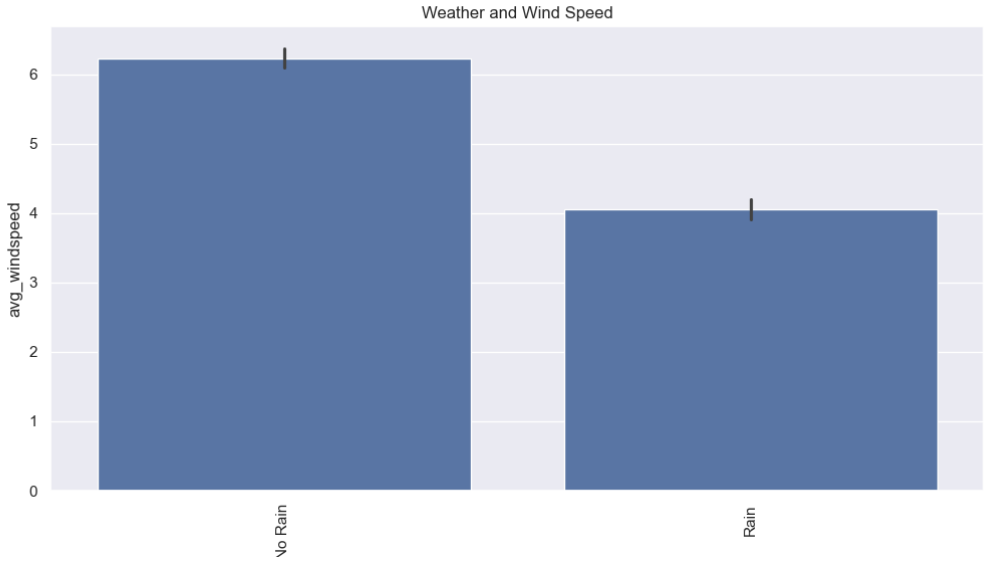
Hình 8: Thể hiện nhãn thời tiết theo giá trị nhiệt độ

Quan sát: Đối với đặc trưng temp ta thấy những ngày không mưa nhiệt độ thường cao hơn những ngày không mưa



Hình 9: Thể hiện nhãn thời tiết theo giá trị độ ẩm

Quan sát: Đặc trưng humidity cho thấy những ngày có mưa thì độ ẩm thường sẽ cao hơn những ngày không mưa

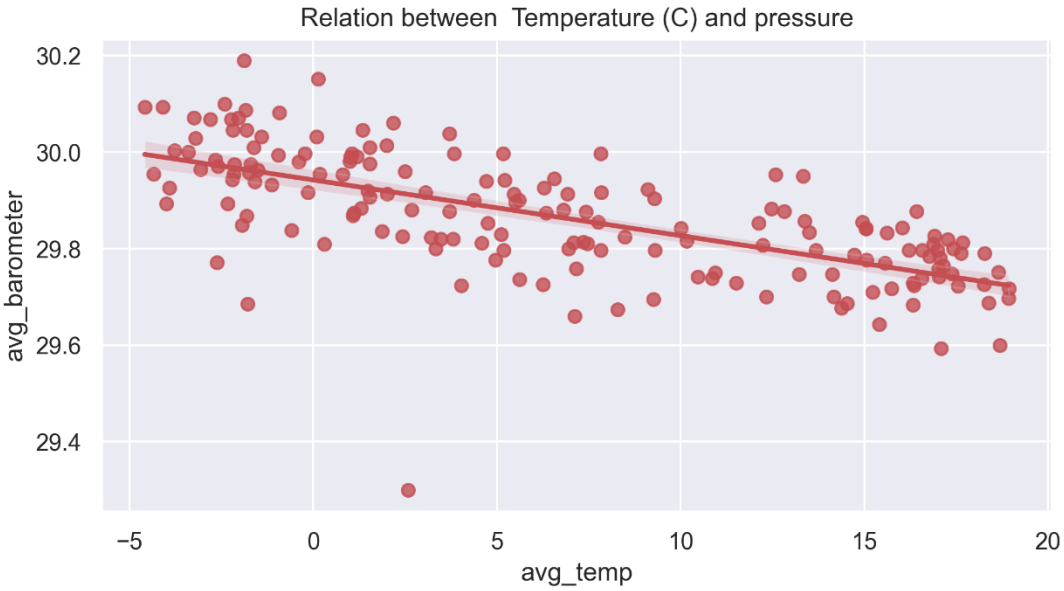


Hình 10: Thể hiện nhãn thời tiết theo giá trị mây phủ bức xạ tán xạ

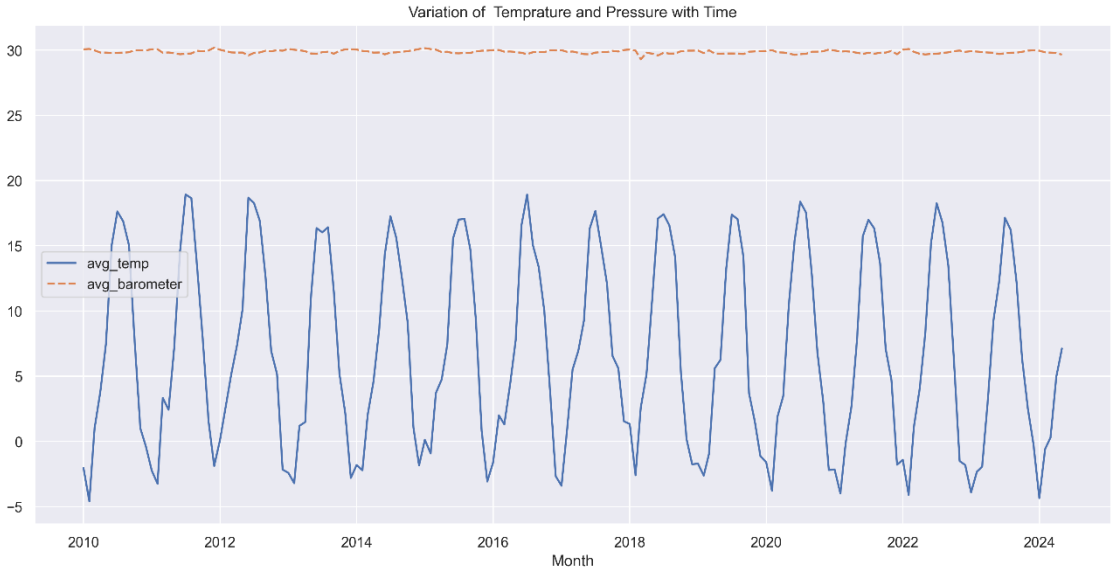
Quan sát: Đối với đặc trưng winspeed ta thấy tốc độ gió trong những ngày không mưa thường cao hơn tốc độ gió ngày có mưa

4. Phân tích theo cặp đặc trưng dựa trên hồi quy

Các thao tác sau đây đều dựa trên việc lấy mẫu trung bình theo tháng

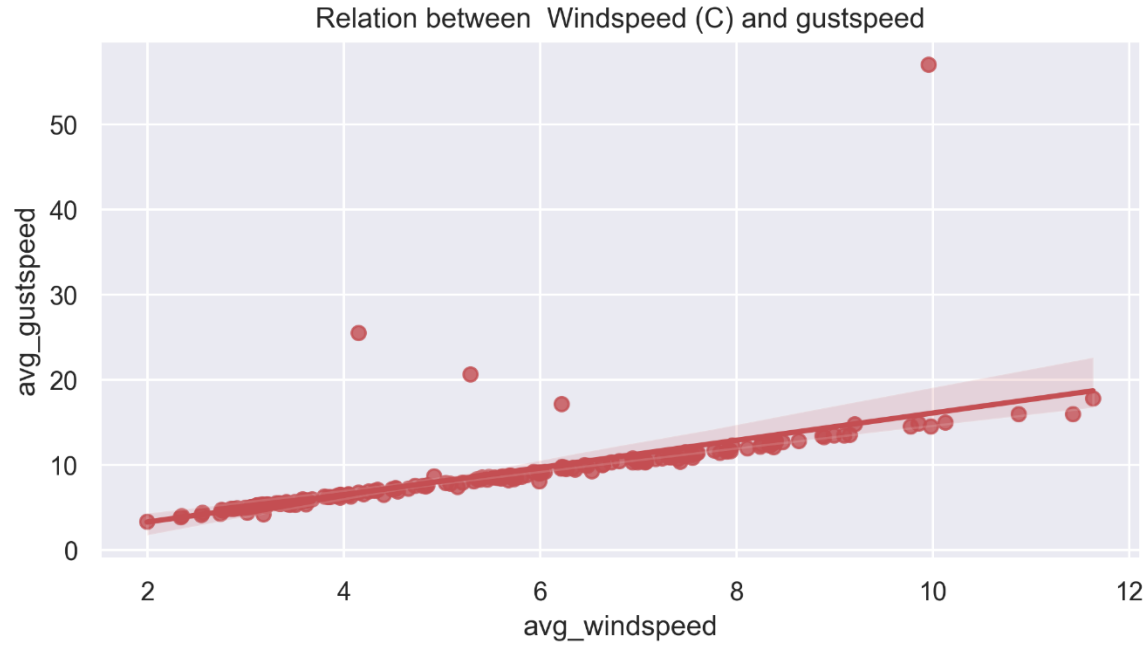


Hình 13: Mô tả quan hệ hồi quy giữa nhiệt độ và áp suất

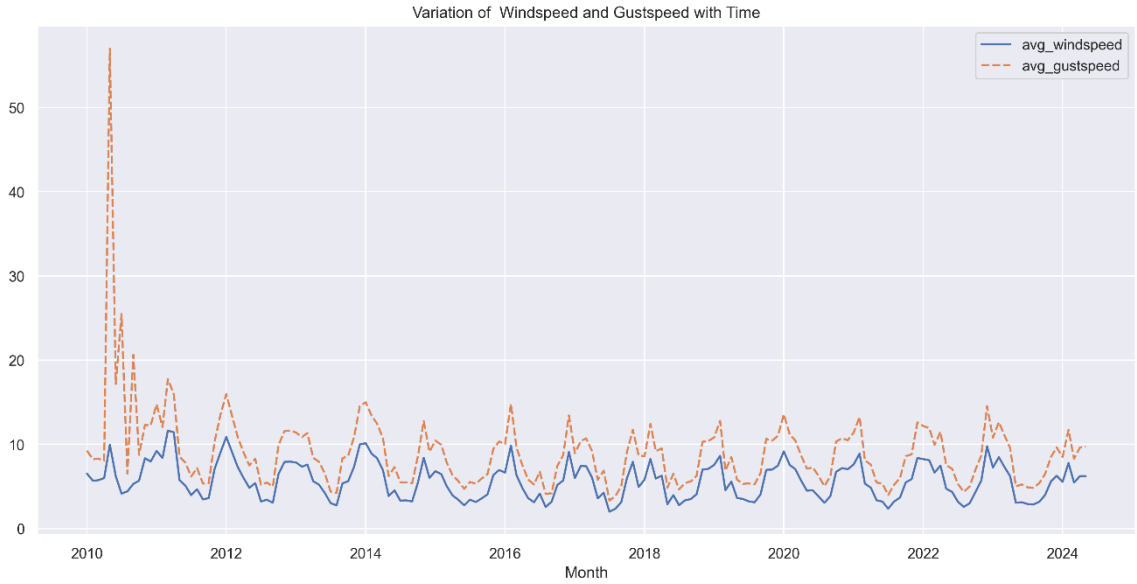


Hình 13: Mô tả quan hệ hồi quy giữa nhiệt độ và áp suất

Quan sát: Đây là mối quan hệ tỉ lệ nghịch. Càng nhiệt độ cao, áp suất KK càng giảm (khoảng tháng 5 đến tháng 11 năm 2016 thể hiện tương đối rõ). Một kiểm chứng khoa học: Khi nhiệt độ càng tăng thì phân tử nhận được năng lượng chuyển động ra xa khỏi nhau dần, sự va chạm lẫn nhau giảm đi khiến áp suất không khí giảm (ở 1 độ cao nhất định)



Hình 14: Mô tả quan hệ hồi quy giữa windspeed và gustspeed

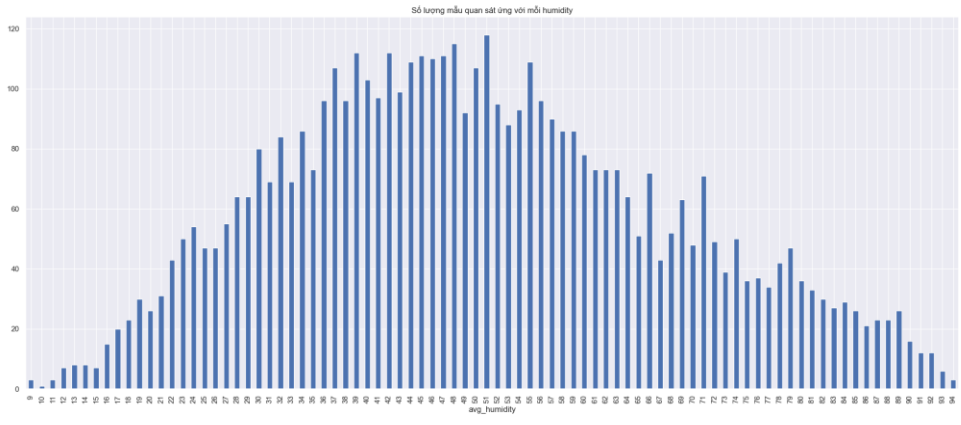


Hình 15: Mô tả trực quan về sự thay đổi giá trị windspeed và gustspeed

Hai đặc trưng windspeed và gustspeed có sự tương quan tỉ lệ thuận cao (heatmap: 0.84) . Tốc độ gió càng cao thì tốc độ gió giật càng mạnh

5. Phân tích ảnh hưởng đặc trưng lên phán đoán dựa trên kiểm định giả thiết

Ở đây, ta lấy đặc trưng rh làm ví dụ, ta muốn kiểm định xem độ ẩm (rh) có ảnh hưởng lên nhiệt độ hay không? Đầu tiên ta xem đặc trưng này có bao nhiêu giá trị và số lượng của từng giá trị.



Hình 18: Số lượng mẫu quan sát ứng với mỗi giá trị độ ẩm.

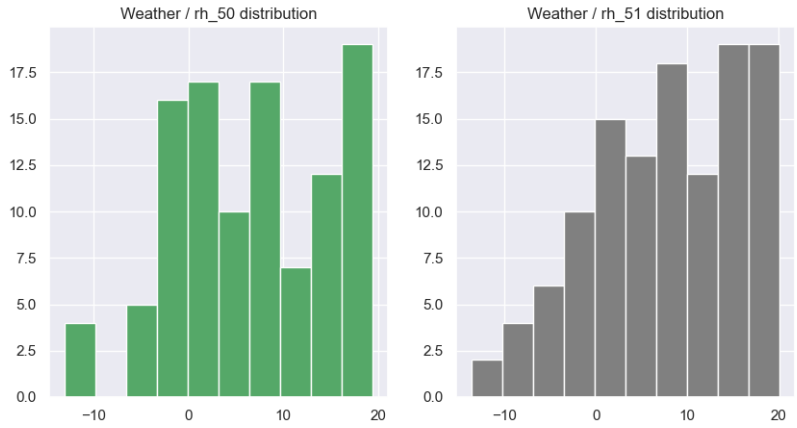
Sau đó, ta kiểm tra xem 2 giá trị liên tiếp nhau có cùng phân phối chuẩn hay không đối với các giá trị nhiệt độ. Lưu ý dữ liệu ở dạng liên tục.

```
[ ]  ### Null hypothesis: dữ liệu tuân theo phân phối chuẩn ###
     ### Nếu pValue < 0.05 ==> phản bác null hypothesis ###
     from scipy import stats
     rh_50_dist = stats.shapiro(rh_50)
     rh_51_dist = stats.shapiro(rh_51)

     print('pvalue for rh_50 distribution: ', rh_50_dist[1])
     print('pvalue for rh_51 distribution: ', rh_51_dist[1])

     pvalue for rh_50 distribution:  9.965465708921991e-17
     pvalue for rh_51 distribution:  2.7548309300413746e-16
```

Rõ ràng, phân phối của 2 giá trị này khác phân phối chuẩn (với độ tin cậy 95%)



Hình 19: Minh họa phân phối của rh=50 và rh=51 đối với avg_temp.

Không cùng phân phối chuẩn, ta sẽ kiểm tra với giả thuyết null: các giá trị rh không có cùng phân phối đối với nhiệt độ của kiểm thử Man Whitnet U .

```
#Null: 2 phân phối như nhau
different = stats.mannwhitneyu(rh_50, rh_51  , alternative='two-sided')


if different[1] < 0.05:
  print('Nhiệt độ không giống nhau về mặt thống kê với 2 giá trị rh 50 và 51 với độ tin cậy 0.05- Tức là độ ẩm có ảnh hưởng đến nhiệt độ')
else:
  print('Không thể kết luận nhiệt độ đối với 2 giá trị rh như trên là khác nhau với độ tin cậy 0.05.')


Nhiệt độ không giống nhau về mặt thống kê với 2 giá trị rh 50 và 51 với độ tin cậy 0.05- Tức là độ ẩm có ảnh hưởng đến nhiệt độ
```

Vậy ta có thể cho rằng độ ẩm có ảnh hưởng lên nhiệt độ, tức là phân phối hai giá trị độ ẩm 50 và 51 khác nhau đối với avg_temp.