Tìm chỉ báo giao thông đông đúc trên I-94

Trong dự án này, chúng ta sẽ phân tích tập dữ liệu về giao thông đi về phía tây trên đường cao tốc Liên bang I-94 (https://en.wikipedia.org/wiki/Interstate_94) kết nối Great Lakes và các vùng phía bắc Great Plains của Hoa Kỳ Bộ dữ liệu do John Hogue cung cấp và có thể tải xuống từ kho lưu trữ này (https://archive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume).

Mục tiêu phân tích của chúng tôi là xác định một vài chỉ số về mật độ giao thông trên I-94, chẳng hạn như loại thời tiết, ngày trong tuần, giờ, v.v.

Tóm tắt kết quả

Chúng tôi phát hiện ra rằng lưu lượng truy cập cao nhất vào ban ngày, những tháng ấm áp và ngày làm việc, đặc biệt là từ 6:00-8:00 và 16:00-17:00. Nhiệt độ không ảnh hưởng đến cường độ giao thông, trong khi một số điều kiện thời tiết tương đối nhẹ thì có. Lưu lượng truy cập trung bình thấp nhất liên quan đến năm 2016, tiếp theo là cao điểm nhất vào năm 2017. Trong tất cả các ngày lễ, lưu lượng truy cập lớn nhất liên quan đến Ngày Columbus, lưu lượng truy cập nhẹ nhất liên quan đến Ngày Giáng sinh và Năm mới.

Tải xuống tập dữ liệu và phân tích ban đầu

Trong [11]:
```
# Nhập thư viện
nhập gấu trúc dư ới dạng pd
nhập numpy dư ới dạng np
nhập matplotlib.pyplot dư ới dạng plt
nhập seaborn dư ới dạng sns

%matplotlib nội tuyến
#Khi chúng tôi sử dụng Matplotlib bên trong Jupyter, chúng tôi cũng cần thêm %matplotlib vào
```

Trong [12]:
```
# Tải dữ liệu vào
mitv = pd.read_csv('Metro_Interstate_Traffic_Volume.csv')
```

Trong [13]:
```
mitv.head(5)
```

Hết[13]:

	nhiệt độ ngày lễ mư a_1 giờ tuyết_1 giờ mây_mọi thời tiết_thời tiết chính_mô tả ngày_giờ					
0	Không có 288,28	0,0	0,0	40	Mây	những đám mây rải rác	2012-10-02 09:00:00
1	Không 289.36	0,0	0,0	75	Mây	những đám mây tan vỡ	2012-10-02 10:00:00
2	Không có 289,58	0,0	0,0	90	Mây	những đám mây u ám	2012-10-02 11:00:00
3	Không có 290,13	0,0	0,0	90	Mây	những đám mây u ám	2012-10-02 12:00:00
4	Không có 291,14	0,0	0,0	75	Mây	những đám mây tan vỡ	2012-10-02 13:00:00

Trong [14]:
```
mitv.tail(5)
```

Hết[14]:

	nhiệt độ	ngày lễ	mư a_1 giờ	tuyết_1 giờ	mây_mọi	thời tiết_thời tiết chính	mô tả ngày_t	
48199	Không có	283,45	0,0	0,0	75	Mây	những đám mây tan vỡ	2018 19:00
48200	Không có	282,76	0,0	0,0	90	Mây	những đám mây u ám	2018 20:00
48201	Không có	282,73	0,0	0,0	90	Giông bão	sự gần gũi dông	2018 21:00
48202	Không có	282.09	0,0	0,0	90	Mây	những đám mây u ám	2018 22:00
48203	Không có	282.12	0,0	0,0	90	Mây	những đám mây u ám	2018 23:00

Trong [15]:
```
mitv.info()
```

```
<lớp 'pandas.core.frame.DataFrame'>
RangeIndex: 48204 mục, 0 đến 48203
Các cột dữ liệu (tổng cộng 9 cột):
 # Cột Non-Null Count Dtype
--- ------ -------------- -----
 0 ngày lễ 1 48204 đối tư ợng không null
 nhiệt độ 48204 float64 không null
 2 mư a_1h 3 48204 float64 không null
 tuyết_1h 4 48204 float64 không null
 đám mây_all 48204 không null int64 5 thời tiết_main 48204
 đối tư ợng không null 6 weather_description 48204 đối tư ợng
 không null 7 date_time 48204 đối tư ợng không null 8
 traffic_volume 48204 non null int64 dtypes: float64(3),
 int64(2), đối tư ợng(4)

sử dụng bộ nhớ: 3,3+ MB
```

Trong [16]:
```
mitv.describe()
```

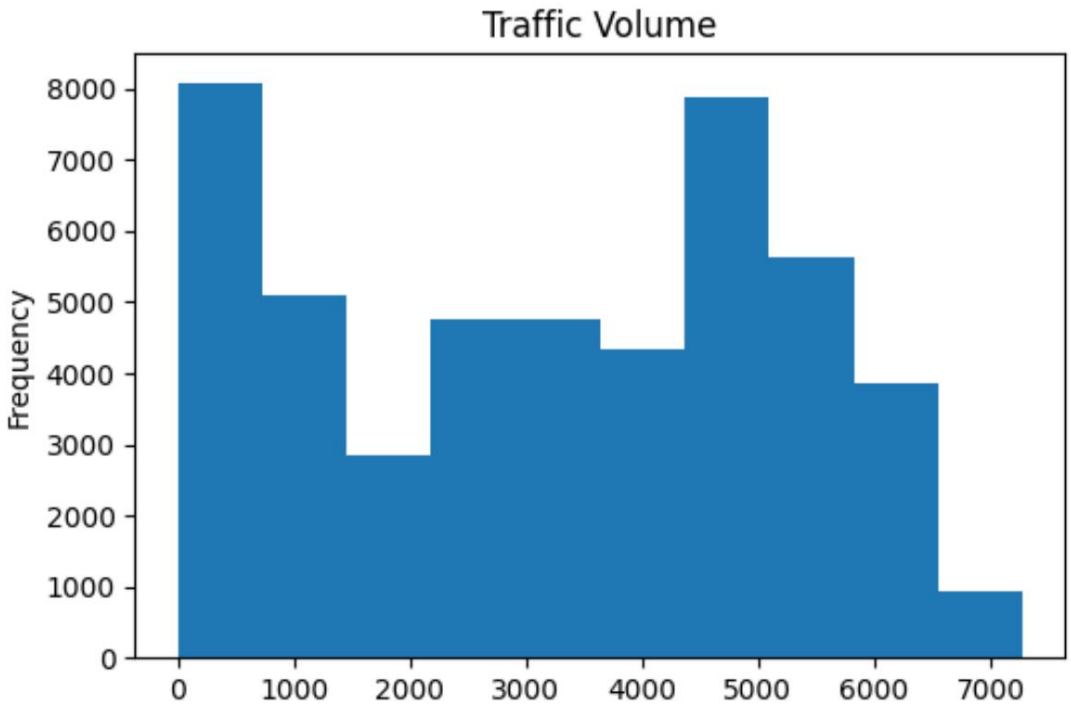
Hết[16]:

	nhiệt độ	mư a_1h	tuyết_1 giờ	đám mây_tất cả lư u lư ợng truy cập_âm lư ợng
đếm	48204.000000	48204.000000	48204.000000	48204.000000	48204.000000
nghĩa là	281.205870	0,334264	0,000222	49.362231	3259.818355
tiêu chuẩn	13.338232	44.789133	0,008168	39.015750	1986.860670
tối thiểu	0,000000	0,000000	0,000000	0,000000	0,000000
25%	272,160000	0,000000	0,000000	1.000000	1193.000000
50%	282.450000	0,000000	0,000000	64.000000	3380.000000
75%	291.806000	0,000000	0,000000	90.000000	4933.000000
tối đa	310.070000	9831.300000	0,510000	100.000000	7280.000000

Trong [17]:
```
mitv['holiday'].value_counts()
```

Hết[17]:
```
Không                                48143
Ngày lao đô ng                           7
Ngày Martin Luther King Jr.              6
ngày lễ Tạ Ơ n                           6
ngày Giáng Sinh                          6
Ngay đâu năm                             6
sinh nhật Washington                     5
ngày cựu chiến binh                      5
ngày kỷ niệm                             5
Ngày Columbus                            5
Ngày Quốc Khánh                          5
hội chợ bang                             5
Tên: kỳ nghỉ, dtype: int64
```

Trong [18]:
```
mitv['traffic_volume'].plot.hist()
plt.title(' Lưu lượng truy cập')
plt.show()
```



Trong [19]:
```
mitv['traffic_volume'].describe()
```

Ra[19]:

đếm	48204.000000
nghĩa là	3259.818355
tiêu	1986.860670
chuẩn	0,000000
tối thiểu 25%	1193.000000
50%	3380.000000
75%	4933.000000
tối đa	7280.000000

Tên: traffic_volume, dtype: float64

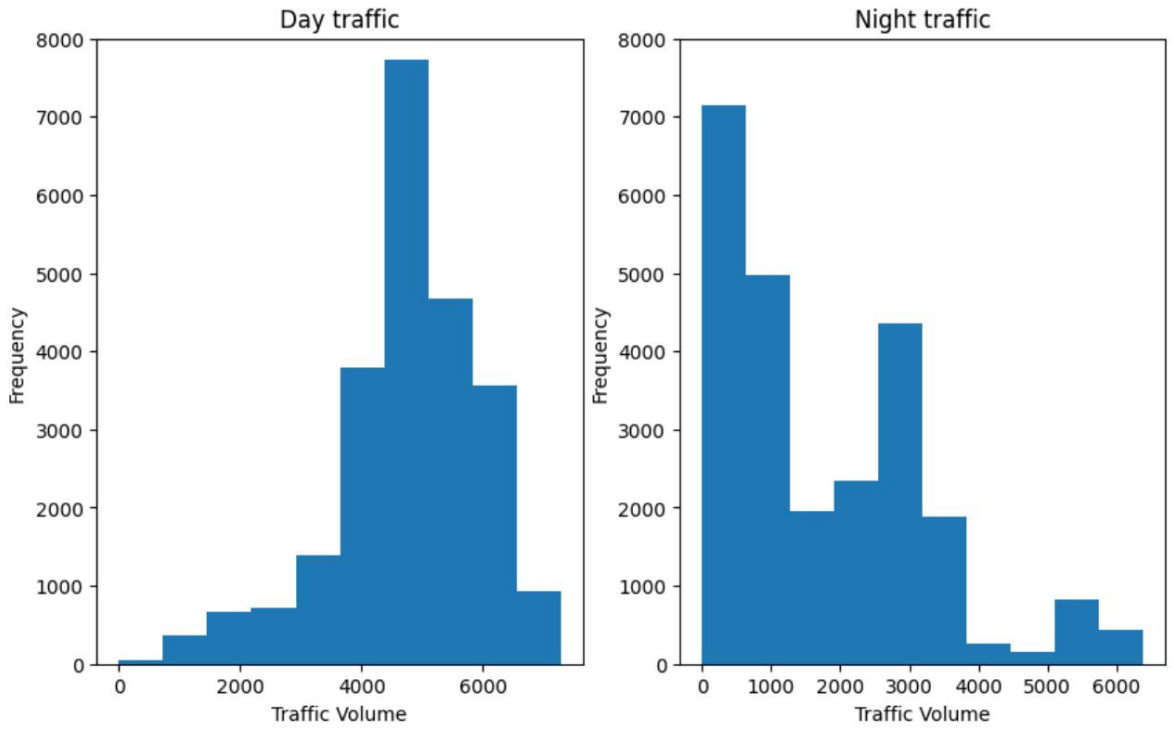
Lưu lượng truy cập phân bổ khá đều mỗi giờ, có thời điểm tăng lên khoảng 7000 giao thông trong giờ cao điểm.

Trong [20]:
```
mitv['date_time'] = pd.to_datetime(mitv['date_time']) # nó từng là đối tượng
mitv['hour'] = mitv['date_time'].dt.hour

# Cô lập ngày và đêm
ngày = mitv.copy()[(mitv['hour'] >= 7) & (mitv['hour'] < 19)]
đêm = mitv.copy()[(mitv['hour'] < 7) | (mitv['giờ'] >= 19)]

# Giá trị duy nhất trong tập dữ liệu
print('Ngày giờ: \n', ngày['giờ'].unique())
in ('-' * 40)
print('Giờ đêm: \n', đêm['giờ'].unique())
```

```
Ngày giờ:
 [ 9 10 11 12 13 14 15 16 17 18 8 7]
----------------------------------------

Giờ ban đêm:
 [19 20 21 22 23 0 1 2 3 4 5 6]
```

Trong [21]:
```
plt.figure(figsize = (10, 6))

# Ô con đầu tiên - day
plt.subplot(1, 2, 1)
plt.title(' Lưu lượng truy
cập trong ngày') plt.hist(ngày['lưu
lượng truy cập']) plt.xlabel('
Lưu lượng truy cập')
plt.ylabel(' Tần số') plt.ylim([0, 8000]) # cùng dải

# Ô con thứ hai - ban đêm
plt.subplot(1, 2, 2)
plt.title(' Giao thông ban đêm
') plt.hist(night['traffic_volume'])
plt.xlabel(' Lưu lượng giao
thông') plt.ylabel(' Tần
suất') plt.ylim([0, 8000])
```

Hết[21]: (0,0, 8000,0)



Trong [22]:
```
# Thống Kê Ngày Đêm

print(" Giao thong trong day:", "\n", day["traffic_volume"].describe())
in("-" * 40)
print("Giao thông ban đêm:", "\n", night["traffic_volume"].describe())
```

```
Lư u lư ợng truy cập
  trong ngày: đếm   23877.000000
nghĩa là          4762.047452
tiêu chuẩn        1174.546482
tối thiểu             0,000000
25%               4252.000000
50%               4820.000000
75%               5559.000000
tối đa            7280.000000
Tên: traffic_volume, dtype: float64
----------------------------------------

Giao thông ban
 đêm: đếm 24327.000000
nghĩa là          1785.377441
tiêu              1441.951197
chuẩn                 0,000000
tối thiểu 25%      530.000000
50%               1287.000000
75%               2819.000000
tối đa            6386.000000
Tên: traffic_volume, dtype: float64
```

75% trong ngày là hơ n 5559 lần vận chuyển như ng trong đêm chỉ là 2819. Mọi số liệu thống kê trong ngày đều lớn hơ n những ngư ời trong đêm.

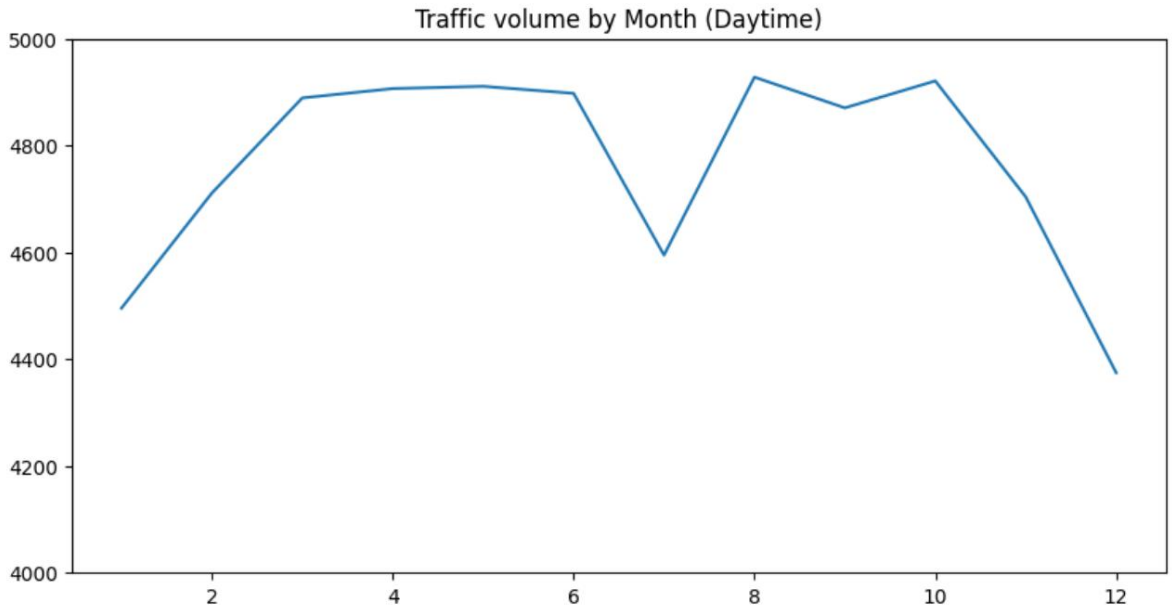
Chỉ báo thời gian

Trong [23]:
```
ngày['tháng'] = ngày['ngày_giờ'].dt.tháng
by_month = day.groupby('month').mean()
theo_tháng['traffic_volume']
```

Hết[23]:
```
tháng
1       4495.613727
2       4711.198394
3       4889.409560
4       4906.894305
5       4911.121609
6       4898.019566
7       4595.035744
        4928.302035
8 9     4870.783145
10      4921.234922
11      4704.094319
12 4374.834566
Tên: traffic_volume, dtype: float64
```

Trong [30]:
```
plt.figure(figsize=(10,5))
plt.plot(theo_tháng['traffic_volume'])
plt.title(" Lư u lư ợng truy cập theo Tháng (Ban ngày)")
plt.ylim(4000,5000)
```

Ra[30]: (4000.0, 5000.0)



Lư u lư ợng phư ơ ng tiện vào các tháng ấm trong năm như tháng 3 - 6 và tháng 8 - 10 đông đúc hơ n đó trong những tháng lạnh giá. Ngoài tháng 7, có thể có các kỳ nghỉ hè và trẻ em không phải đi học.

Ngày trong tuần

Trong [31]:
```
day['dayofweek'] = day['date_time'].dt.dayofweek
by_dayofweek = day.groupby('dayofweek').mean()
by_dayofweek['traffic_volume'] # 0 là Thứ Hai, 6 là Chủ Nhật
```

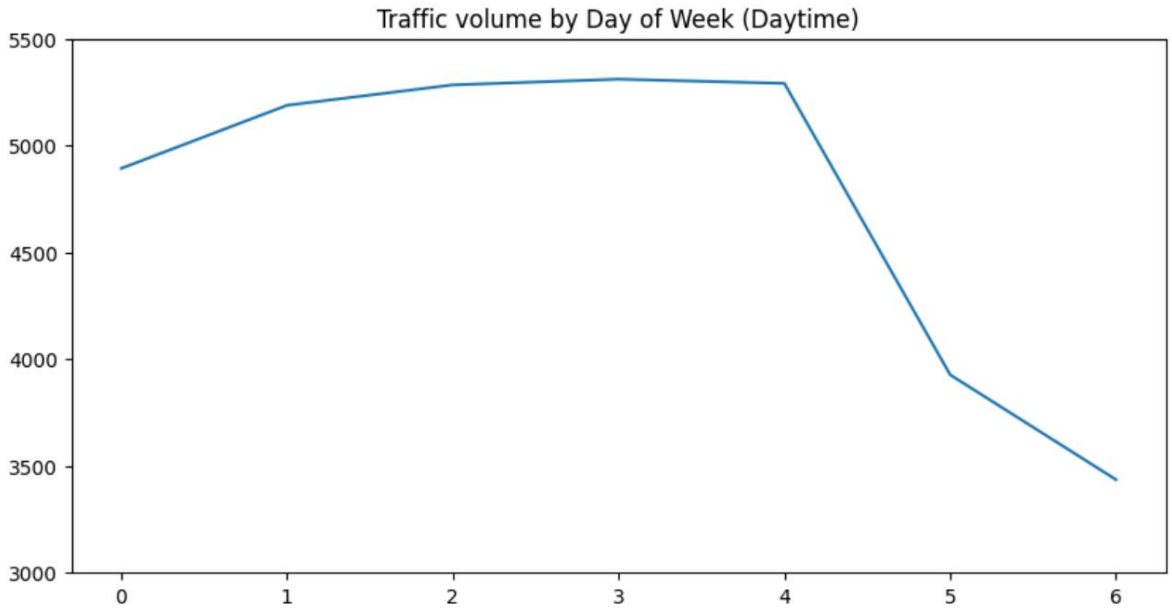
Hết[31]: ngày trong tuần

0	4893.551286
1	5189.004782
2	5284.454282
3	5311.303730
4	5291.600829
5	3927.249558
6	3436.541789

Tên: traffic_volume, dtype: float64

Trong [32]:
```
plt.figure(figsize=(10,5))
plt.plot(by_dayofweek['traffic_volume'])
plt.title("Lưu lượng truy cập theo ngày trong tuần (Ban ngày)")
plt.ylim(3000, 5500)
```

Ra[32]: (3000.0, 5500.0)



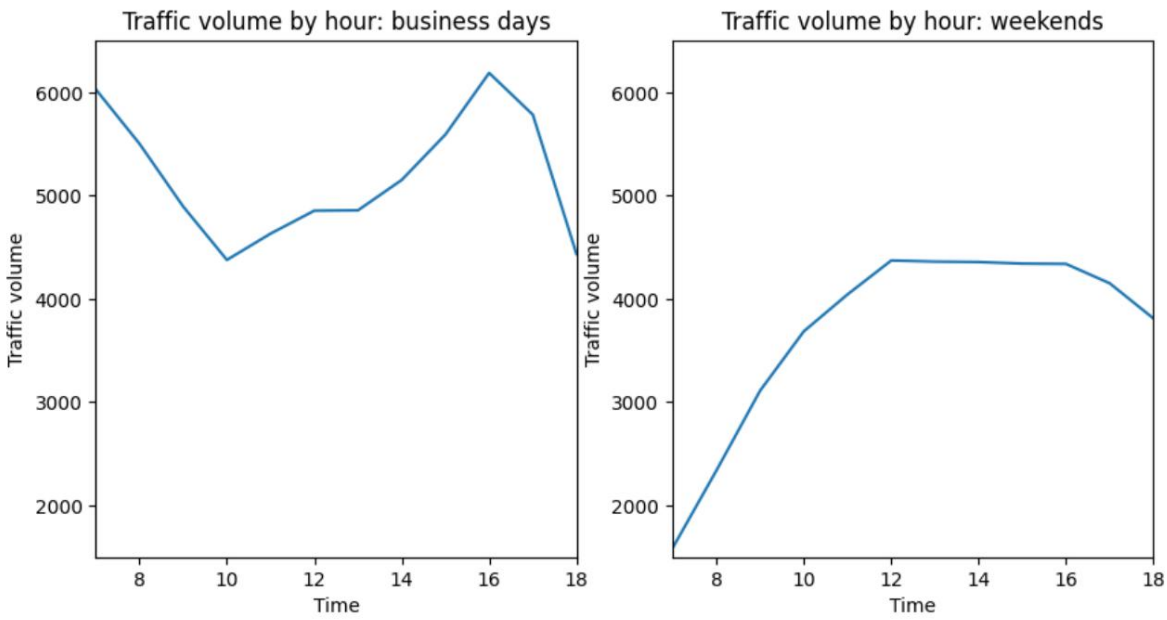
Lưu lượng giao thông cao hơn đáng kể vào các ngày làm việc (0-4) so với các ngày cuối tuần (5, 6).

Trong [33]:
```
day['hour'] = day['date_time'].dt.hour
bussiness_days = day.copy()[day['dayofweek'] <= 4] # 4 == Thứ sáu
cuối tuần = day.copy()[day['dayofweek'] >= 5] # 5 == Thứ bảy
by_hour_business = bussiness_days.groupby('hour').mean()
by_hour_weekend = cuối tuần.groupby('hour').mean()

in(by_hour_business['traffic_volume'])
in(theo_giờ_cuối tuần['traffic_volume'])
```

```
giờ
7      6030.413559
       5503.497970
9      4895.269257
10     4378.419118
11     4633.419470
12     4855.382143
13     4859.180473
14     5152.995778
15     5592.897768
16     6189.473647
17     5784.827133
18 4434.209431
Tên: traffic_volume, dtype: float64
giờ
7      1589.365894
       2338.578073
8 9    3111.623917
10     3686.632302
11     4044.154955
12     4372.482883
13     4362.296564
14     4358.543796
15     4342.456881
16     4339.693805
17     4151.919929
18 3811.792279
Tên: traffic_volume, dtype: float64
```

Trong [45]:
```
plt.figure(figsize=(10,5))
plt.subplot(1, 2, 1)
plt.title(' Lưu lượng truy cập theo giờ: ngày làm việc')
plt.plot(by_hour_business['traffic_volume' ]) plt.xlabel('Thời
gian') plt.ylabel(' Lưu
lượng truy cập') plt.ylim([1500,
6500]) plt.xlim(7,18)


plt.subplot(1, 2, 2)
plt.title(' Lưu lượng truy cập theo giờ: cuối tuần')
plt.plot(theo_hour_weekend['lưu lượng truy cập'])
plt.xlabel('Thời gian')
plt.ylabel('Lưu lượng truy cập' )
plt.ylim([1500, 6500])
plt.xlim(7,18)
```

Hết[45]: (7.0, 18.0)



Giao thông đông đúc hơn vào các ngày làm việc trong hầu hết các giờ ban ngày so với các ngày cuối tuần. Đối với ngày làm việc, có 2 giờ cao điểm rõ ràng: 7h-8h và 16h-17h, đều liên quan đến giờ cao điểm khi mọi người đi làm và về. Đối với các ngày cuối tuần, không có cao điểm trên biểu đồ và lưu lượng truy cập tăng dần từ 7:00 đến 12:00, khi đạt đến mức ổn định và từ 16:00 bắt đầu giảm.

Nói chung, chúng tôi đã tìm thấy các chỉ báo thời gian sau đây về lưu lượng truy cập cao hơn:

- tháng ấm áp,
- ngày làm việc,
- thời gian:
  - 7.00-8.00 và 16.00-17.00 vào các ngày làm việc,
  - 12.00-16.00 vào cuối tuần.

Ngoài ra, chúng tôi phát hiện lư u lư ợng giao thông giảm mạnh trong năm 2016, có lẽ là do đư ờng công trình mở rộng, tiếp theo là đỉnh cao nhất vào năm 2017.

Chỉ báo thời tiết

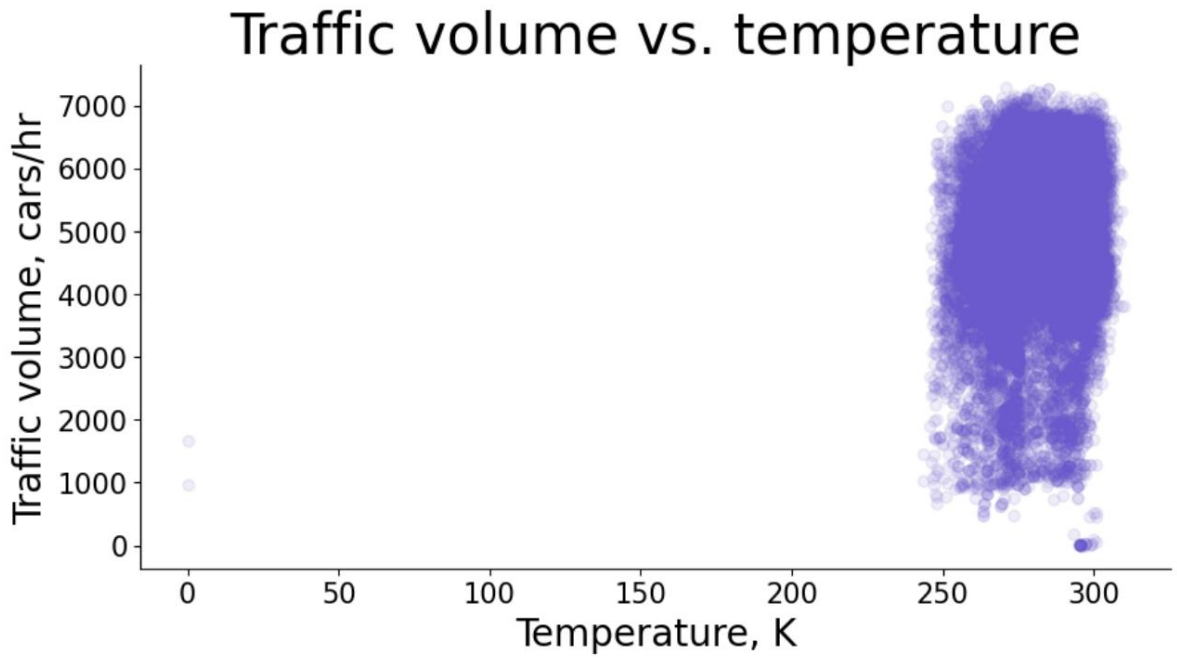
Một chỉ báo có thể khác về lư u lư ợng truy cập lớn là thời tiết. Chúng ta có thể tìm thấy thông tin về thời tiết trong các cột sau: nhiệt độ, mư a_1h, tuyết_1h, mây_all,
thời tiết_chính, mô tả thời tiết. 4 cái đầu tiên trong số chúng là số, vì vậy hãy thử

Trong [49]:
```
round(day.corr()['traffic_volume'][['temp', 'rain_1h', 'snow_1h', 'clouds_all'
```

Hết[49]:
```
nhiệt độ 0,128
mư a_1h 0,004
tuyết_1h 0,001
đám mây_all -0,033
Tên: traffic_volume, dtype: float64
```

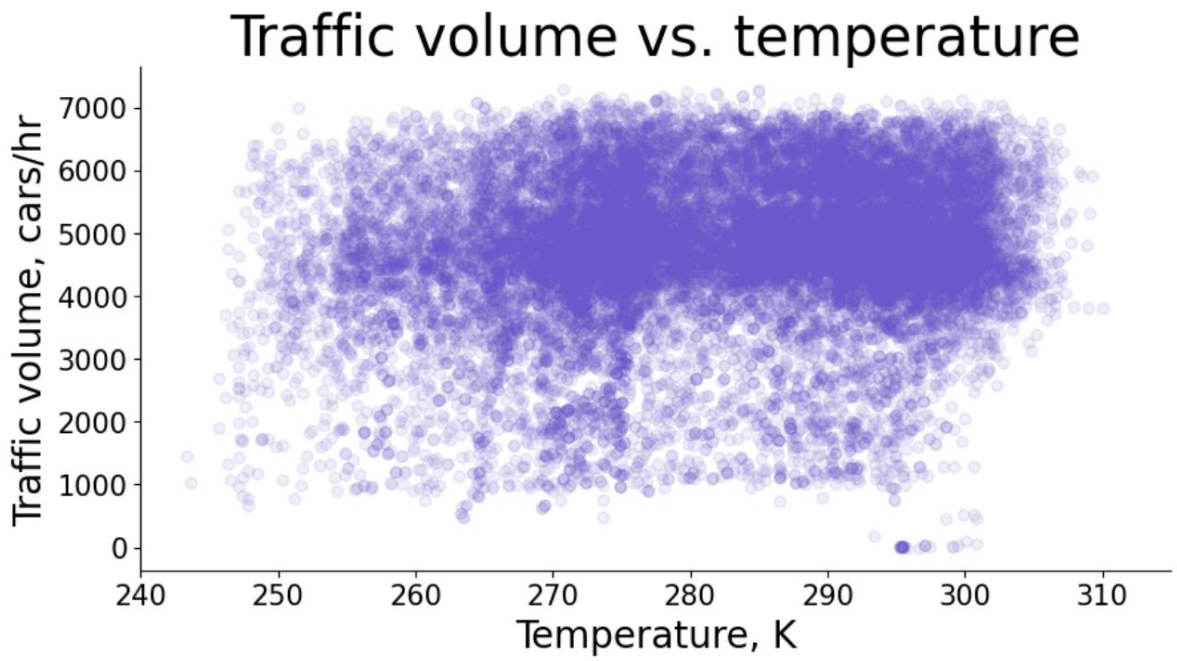
Nhiệt độ cho thấy mối tư ơ ng quan mạnh nhất (mặc dù rất thấp) với lư u lư ợng truy cập âm lư ợng. Hãy vẽ hai biến này với nhau:

Trong [53]:
```
def create_scatter_plot(df, column, title, xlabel, xmin=None, xmax=None):
    plt.figure(figsize=(10,5))
    plt.scatter(df[column], df['traffic_volume '], color='slateblue', alpha=0,1 plt.title(title,
    fontsize=30) plt.xlabel(xlabel,
    fontsize=20) plt.ylabel(' Lư u lư ợng
    giao thông, ô tô/giờ', fontsize=20) plt .xlim(xmin,xmax)
    plt.xticks(fontsize=15)
    plt.yticks(fontsize=15)
    sns.despine()


# Vẽ biểu đồ lư u lư ợng truy cập so với nhiệt độ
create_scatter_plot(df=day, column='temp', title='Lư u
                        lư ợng truy cập so với nhiệt độ', xlabel='Nhiệt
                        độ, K')
```



Có 2 giá trị sai của nhiệt độ đư ợc bỏ qua.

Trong [54]:
```
# Vẽ biểu đồ lư u lư ợng truy cập so với nhiệt độ
create_scatter_plot(df=day, column='temp', title='Lư u
                            lư ợng truy cập so với nhiệt độ', xlabel='Nhiệt độ,
                            K', xmin=240, xmax=315)
```

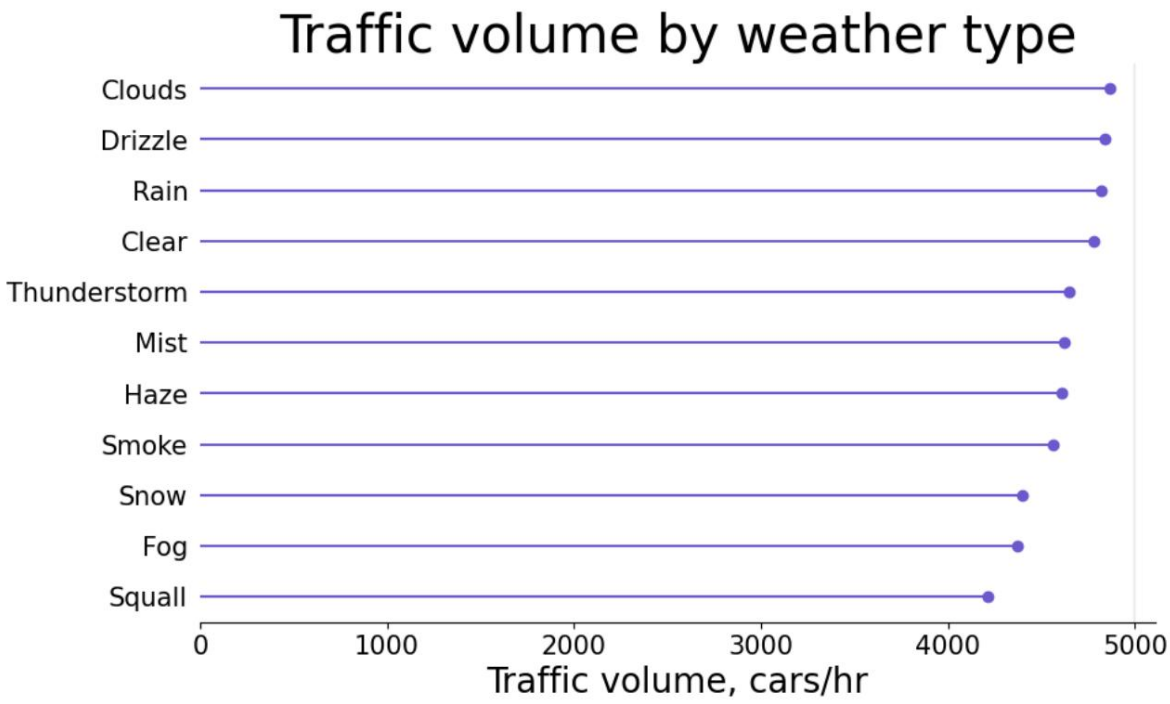


Bây giờ chúng ta có thể kết luận rằng thực sự không có mối tư ơ ng quan hợp lệ nào giữa nhiệt độ và lư u lư ợng giao thông, nghĩa là nhiệt độ không phải là chỉ báo đáng tin cậy cho lư u lư ợng giao thông đông đúc, chư a kể đến 3 cột thời tiết số khác (rain_1h , snow_1h và Clouds_all) cho thấy rất thấp hệ số tư ơ ng quan Pearson. Để xem liệu chúng ta có thể tìm thấy nhiều dữ liệu hữu ích hơ n hay không, tiếp theo chúng ta sẽ xem xét các cột thời tiết phân loại: weather_main và weather_description .

Loại thời tiết

Chúng tôi sẽ tính toán và vẽ biểu đồ lư u lư ợng giao thông trung bình đư ợc liên kết với từng loại thời tiết, tức là mỗi giá trị duy nhất trong các cột weather_main và weather_description .

Trong [55]:
```
by_weather_main = day.groupby('weather_main').mean().sort_values('traffic_volu by_weather_description
= day.groupby('weather_description').mean().sort_values

def create_stem_plot(df, fig_height,
                           title='Lư u lư ợng truy cập theo loại thời tiết',
                           ymin=None, ymax=None, vert_line=5000):
     plt.figure(figsize=(10,fig_height))
     plt.hlines(y=df.index,
                   xmin=0, xmax =df['traffic_volume'],
                   color='slateblue')
     plt.plot(df['traffic_volume'], df.index, 'o',
                c='slateblue')
     plt.title(title, fontsize=30)
     plt .xlabel(' Lư u lư ợng truy cập, ô tô/giờ', fontsize=20)
     plt.ylabel(Không)
     plt.xlim(0,None)
     plt.ylim(ymin,ymax)
     plt.tick_params(left=False)
     plt.axvline( x=vert_line, color='grey', linewidth=0.2)
     plt.xticks(fontsize=15)
     plt.yticks(fontsize=15)
     sns.despine(left=True)

# Vẽ biểu đồ lư u lư ợng giao thông theo loại thời
tiết create_stem_plot(df=by_weather_main, fig_height=6)
```



Không có loại thời tiết nào mà lư u lư ợng giao thông vư ợt quá 5.000 ô tô/giờ, vì vậy chúng tôi không thể xác định bất kỳ chỉ báo giao thông đông đúc nào từ cột weather_main. Thay vào đó, hãy vẽ kết quả cho cột weather_description:

Trong [56]:
```
# Biểu đồ lư u lư ợng giao thông theo loại thời tiết (chi tiết)
create_stem_plot(df=by_weather_description, fig_height=20,
                 ymin=-1, ymax=38)
```

Trong trư ờng hợp này, chúng tôi có thể xác định 3 loại thời tiết sau đây dẫn đến lư u lư ợng truy cập lớn hơ n 5.000 xe/giờ:

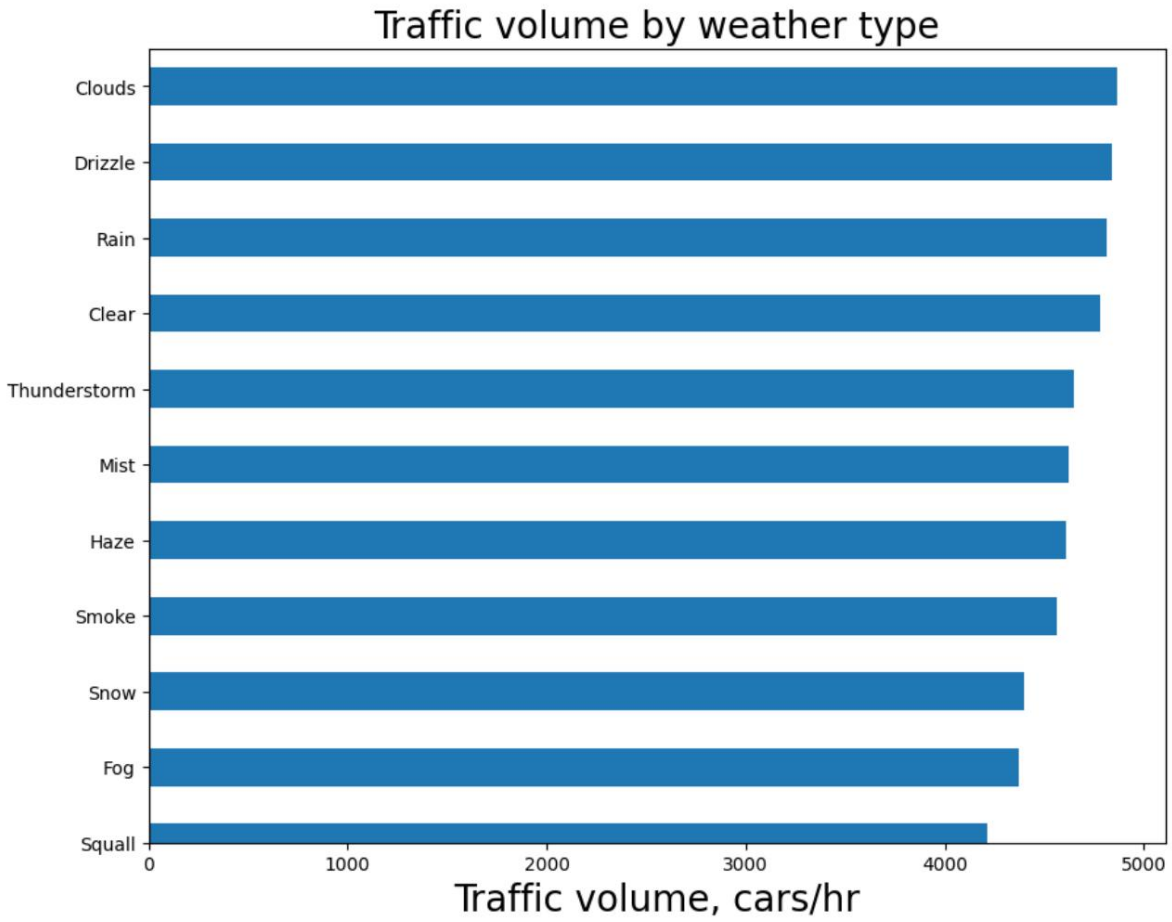
- mưa tuyết,
- mưa nhẹ và tuyết,
- giông bão gần với mưa phùn.

Kết quả có vẻ đáng ngạc nhiên: rõ ràng là có nhiều loại thời tiết khác trong bộ dữ liệu đại diện cho thời tiết tồi tệ hơn nhiều khi giao thông ít hơn nhiều. Một lời giải thích khả dĩ ở đây là các điều kiện thời tiết thực sự xấu (giông bão, mưa rất lớn, gió giật, v.v.) thường được dự báo trước, vì vậy mọi người cố gắng hết sức để không di chuyển bằng ô tô vào những ngày như vậy.

Giải pháp thứ hai của lần khám phá cuối cùng:

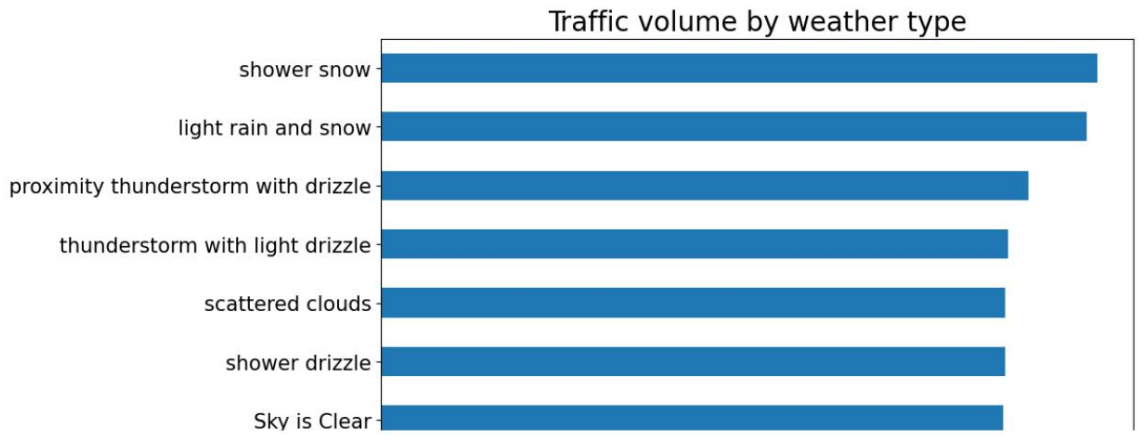
Trong [85]:
```
by_weather_main = day.groupby('weather_main').mean().sort_values('traffic_volu
by_weather_description = day.groupby('weather_description').mean().sort_values

plt.figure(figsize=(10,8))
by_weather_main['traffic_volume'].plot.barh()
plt.title(' Lư u lư ợng giao thông theo loại thời tiết', fontsize=20)
plt.xlabel('Lư u lư ợng giao thông, ô tô/ hr', fontsize=20)
plt.ylabel(None)
plt.xlim(0,None)
plt.ylim(0,None)
plt.xticks(fontsize=10)
plt.yticks(fontsize=10)
```

Ra[85]: (mảng([0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]),
<danh sách 11 đối tư ợng nhãn đánh dấu chính của Văn bản>)



Trong [87]:
```
plt.figure(figsize=(10,30))
by_weather_description['traffic_volume'].plot.barh() plt.title('
Lưu lượng giao thông theo loại thời tiết', fontsize=20) plt.xlabel('
Lưu lượng giao thông , ô tô/giờ', fontsize=20) plt.ylabel(None)
plt.xlim(0,None)
plt.ylim(0,None)
plt.xticks(fontsize=15)
plt.yticks(fontsize=15)
```

Ra[87]:
```
(mảng([ 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1
6,
         17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 3
3,
         34, 35, 36, 37]),
 <danh sách 38 đối tượng nhãn đánh dấu chính của Văn bản>)
```



TRONG [ ]: