Finding Heavy Traffic Indicators on I-94

- We're going to analyze a dataset about the westbound traffic on the I-94 Interstate highway.
- The goal of our analysis is to determine a few indicators of heavy traffic on I-94. These indicators can be weather type, time of the day, time of the week, etc. For instance, we may find out that the traffic is usually heavier in the summer or when it snows.

In [1]:
```
import pandas as pd
mtv = pd.read_csv('Metro_Interstate_Traffic_Volume.csv')
mtv.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48204 entries, 0 to 48203
Data columns (total 9 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   holiday              61 non-null     object 
 1   temp                 48204 non-null  float64
 2   rain_1h              48204 non-null  float64
 3   snow_1h              48204 non-null  float64
 4   clouds_all           48204 non-null  int64  
 5   weather_main         48204 non-null  object 
 6   weather_description  48204 non-null  object 
 7   date_time            48204 non-null  object 
 8   traffic_volume       48204 non-null  int64  
dtypes: float64(3), int64(2), object(4)
memory usage: 3.3+ MB
```

In [2]:
```
mtv.head()
```

Out[2]:

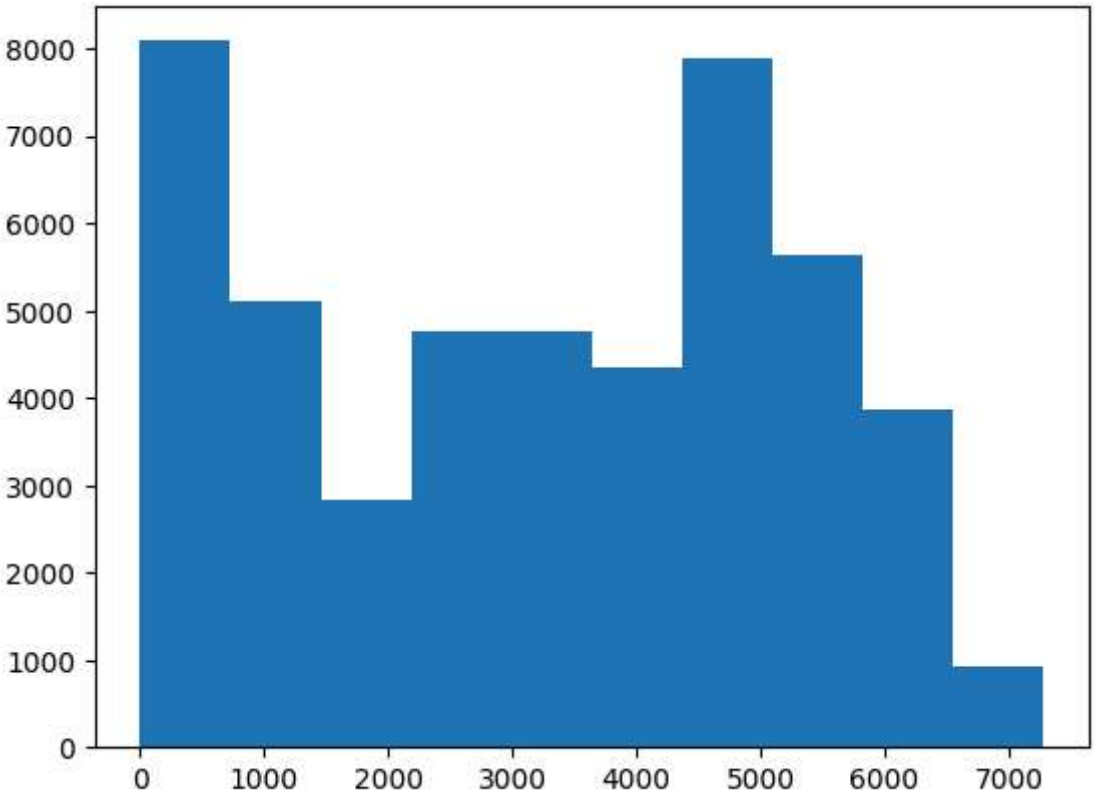
	holiday	temp	rain_1h	snow_1h	clouds_all	weather_main	weather_description	date_
0	NaN	288.28	0.0	0.0	40	Clouds	scattered clouds	2012 09:(
1	NaN	289.36	0.0	0.0	75	Clouds	broken clouds	2012 10:(
2	NaN	289.58	0.0	0.0	90	Clouds	overcast clouds	2012 11:(
3	NaN	290.13	0.0	0.0	90	Clouds	overcast clouds	2012 12:(
4	NaN	291.14	0.0	0.0	75	Clouds	broken clouds	2012 13:(

In [3]:
```
mtv.tail()
```

Out[3]:

	holiday	temp	rain_1h	snow_1h	clouds_all	weather_main	weather_description
48199	NaN	283.45	0.0	0.0	75	Clouds	broken clouds
48200	NaN	282.76	0.0	0.0	90	Clouds	overcast clouds
48201	NaN	282.73	0.0	0.0	90	Thunderstorm	proximity thunderstorm
48202	NaN	282.09	0.0	0.0	90	Clouds	overcast clouds
48203	NaN	282.12	0.0	0.0	90	Clouds	overcast clouds

In [4]:
```
import matplotlib.pyplot as plt
%matplotlib inline
plt.hist(mtv['traffic_volume'])
plt.show()
```



In [5]:
```
mtv['traffic_volume'].describe()
```

Out[5]:
```
count    48204.000000
mean      3259.818355
std       1986.860670
min          0.000000
25%       1193.000000
50%       3380.000000
75%       4933.000000
max       7280.000000
Name: traffic_volume, dtype: float64
```

Observing the data we can see that time can be an effecter to the traffic volume because:

- min-max: large distance => sometimes the road is very crowded, sometimes the road is very empty
- std vs mean => the number of vehicles fluctuates strongly over time
- mean vs median: has small deviation => right deviation => traffic volume fluctuates from low to medium => there are times when the road is empty (early morning or late night), sometimes it is very congested (rush hour) => the rest is average traffic time

Traffic Volume: Day vs. Night

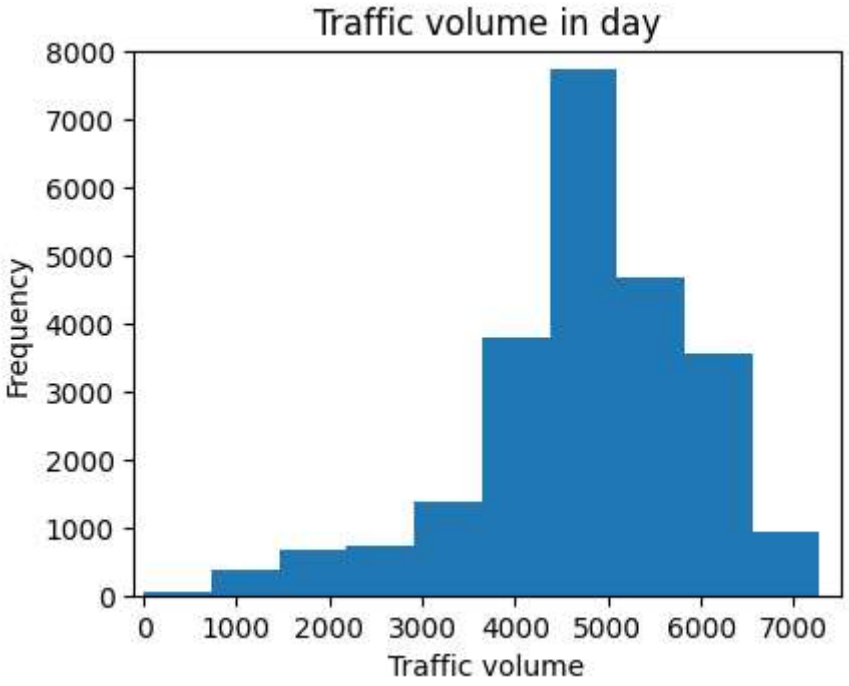
We'll start by dividing the dataset into two parts:

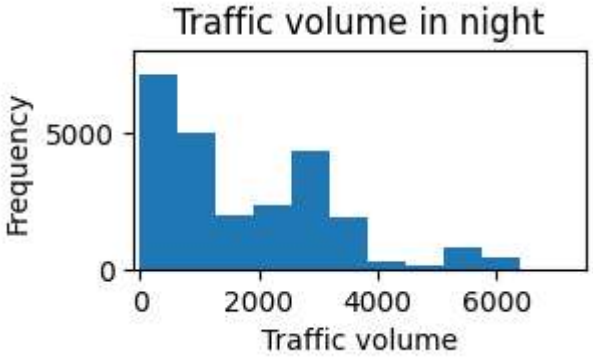
Daytime data: hours from 7 a.m. to 7 p.m. (12 hours) Nighttime data: hours from 7 p.m. to 7 a.m. (12 hours) While this is not a perfect criterion for distinguishing between nighttime and daytime, it's a good starting point.

In [6]:
```
'''converting the data_time to datatime'''
mtv['date_time'] = pd.to_datetime(mtv['date_time'])

day = mtv.copy()[(mtv['date_time'].dt.hour>=7) & (mtv['date_time'].dt.hour<19)]
print(day.shape)
night = mtv.copy()[(mtv['date_time'].dt.hour<7) | (mtv['date_time'].dt.hour >=19)]
print(night.shape)
```

```
(23877, 9)
(24327, 9)
```

In [7]:
```
plt.figure(figsize=(10, 12))
plt.subplot(3,2,1)
plt.hist(day['traffic_volume'])
plt.xlim([-100,7500])
plt.ylim([0,8000])
plt.title('Traffic volume in day')
plt.xlabel('Traffic volume')
plt.ylabel('Frequency')
plt.show()

plt.subplot(3,2,2)
plt.hist(night['traffic_volume'])
plt.xlim([-100,7500])
plt.ylim([0,8000] )
plt.title('Traffic volume in night')
plt.xlabel('Traffic volume')
plt.ylabel('Frequency')
plt.show()
```





In [8]:
```
day['traffic_volume'].describe()
```

Out[8]:
```
count    23877.000000
mean      4762.047452
std       1174.546482
min          0.000000
25%       4252.000000
50%       4820.000000
75%       5559.000000
max       7280.000000
Name: traffic_volume, dtype: float64
```

In [9]:
```
night['traffic_volume'].describe()
```

Out[9]:
```
count    24327.000000
mean      1785.377441
std       1441.951197
min          0.000000
25%        530.000000
50%       1287.000000
75%       2819.000000
max       6386.000000
Name: traffic_volume, dtype: float64
```

Daytime data chart is skewed left => most of the day has high traffic volume. 75% of the time has traffic volume above 4252. Meanwhile, nighttime data is skewed right => most of the night has low traffic volume. 75% of the time has traffic volume below 2819. => Although there are more than 5000 vehicles recorded, it is still insignificant => Focus on daytime data.

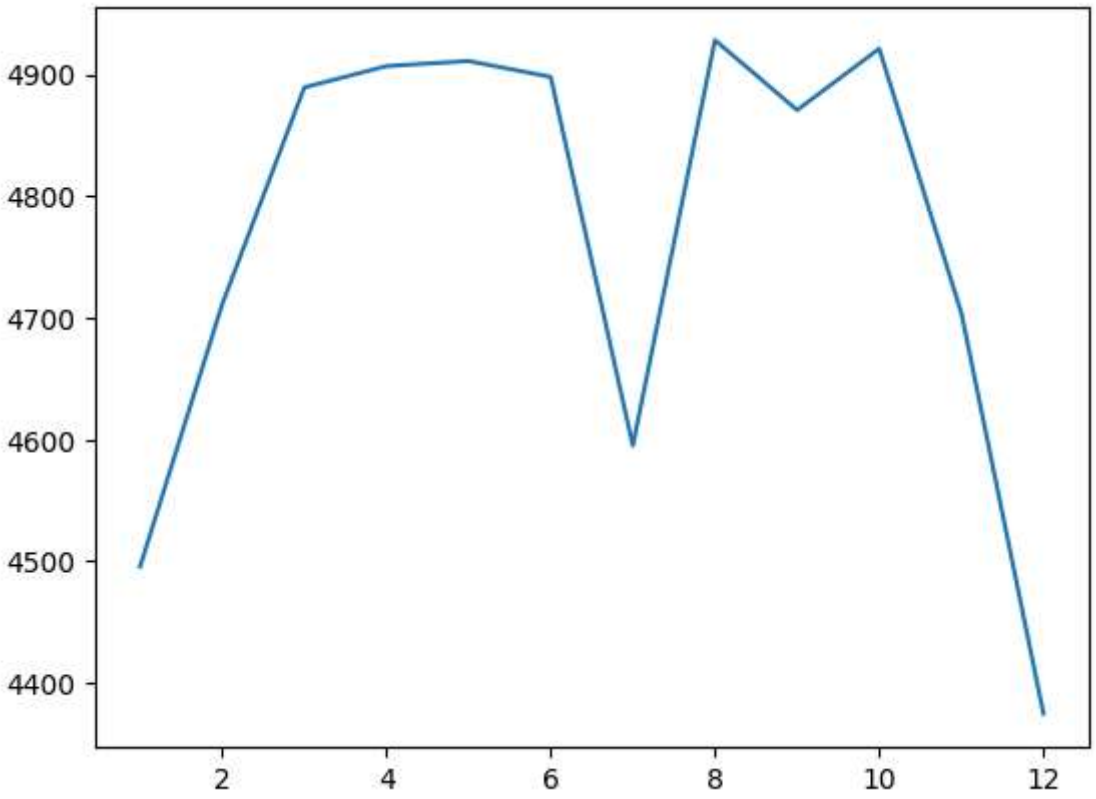
Time Indicators

One of the possible indicators of heavy traffic is time. There might be more people on the road in a certain month, on a certain day, or at a certain time of day.

We're going to look at a few line plots showing how the traffic volume changes according to the following:

Month Day of the week Time of day

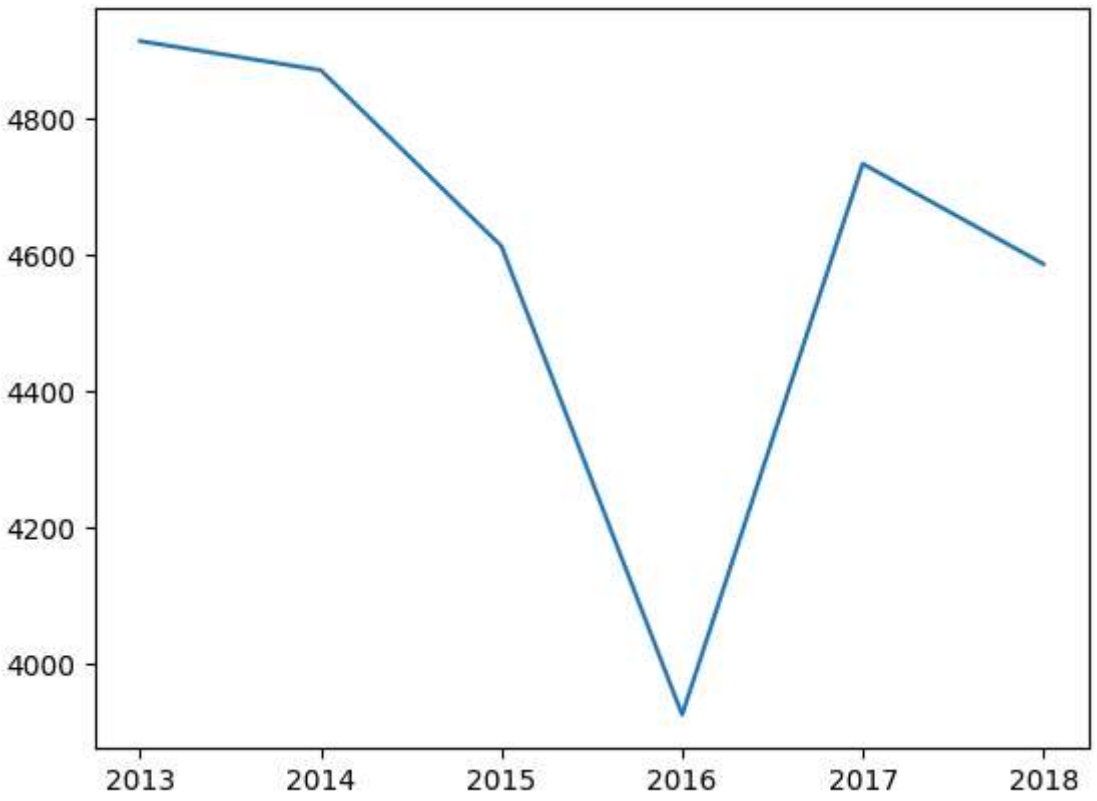
In [10]:
```
day['month'] = day['date_time'].dt.month
by_month = day.groupby('month').mean(numeric_only=True)
plt.plot(by_month['traffic_volume'])
plt.show()
```



The data shows that the months of March-June have a higher number of vehicles, similarly, August-October has. It could be due to the weather factor, these are the periods of pleasant weather, people go out more than other months It is possible that January, July, December have a much lower number of vehicles, these are the periods strongly affected by the weather:

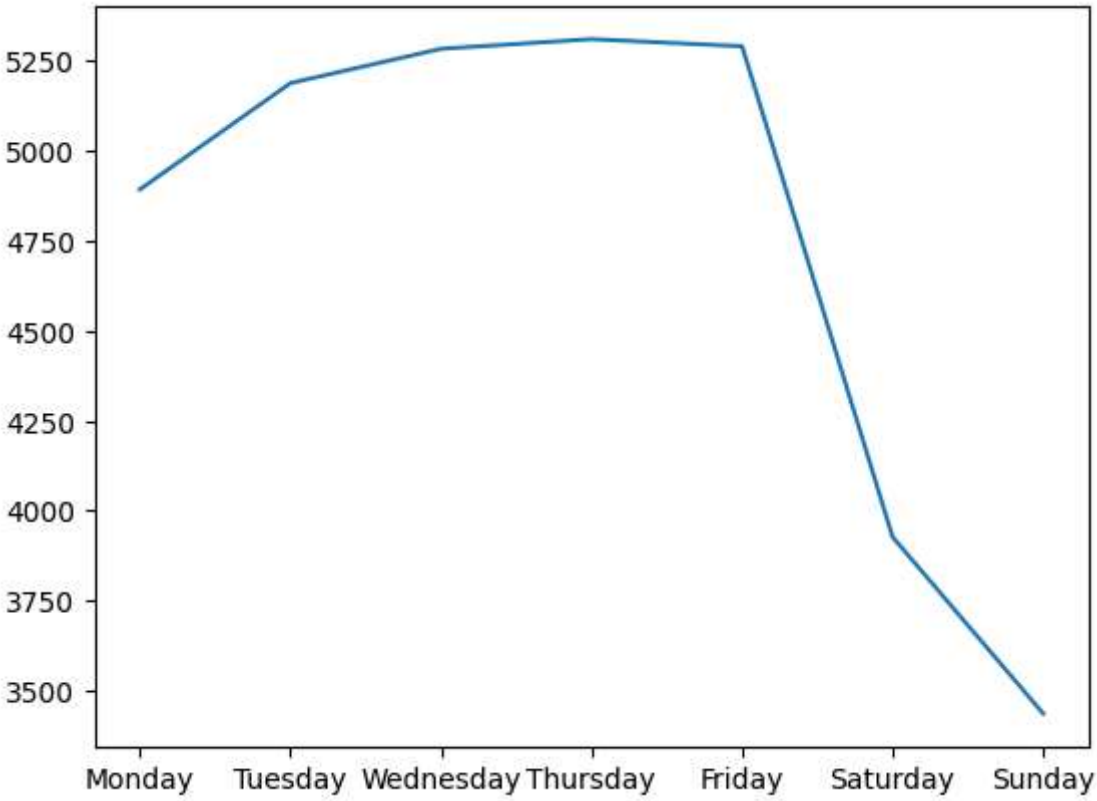
- Hot in summer (July)
- Snow in winter (December-January)
- =>But sunny weather is not a convincing reason. Whether there are some unexpections?

In [11]:
```
day['year']= day['date_time'].dt.year
by_year_july= day[day['month'] == 7].groupby('year').mean(numeric_only=True)
plt.plot(by_year_july['traffic_volume'])
plt.show()
```



We can see that in July, the traffic volume alwways over 4600 (high level) but in 2016 the traffic volume suddenly down under 4000. Therefor it effects significantly to the total data So We can conclude that in the whole warm months, the traffic volume is high

In [12]:
```
day['dayofweek'] = day['date_time'].dt.dayofweek
by_dayofweek = day.groupby('dayofweek').mean(numeric_only=True)
plt.plot(by_dayofweek['traffic_volume'])
plt.xticks(ticks=[0,1,2,3,4,5,6],
           labels=['Monday','Tuesday','Wednesday','Thursday','Friday','Saturday', '
plt.show()
```

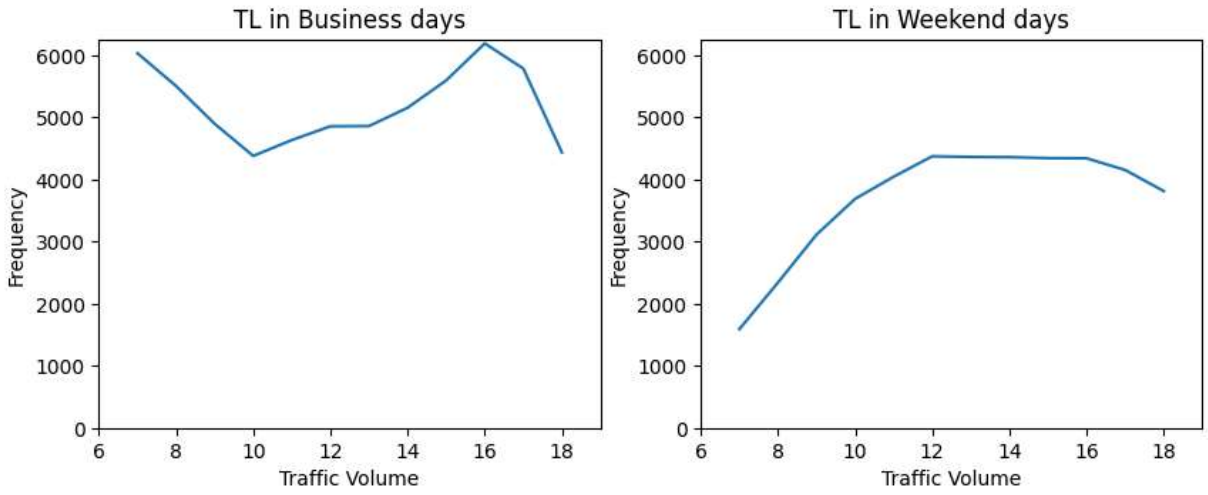


From Monday to Friday, The traffic volume is always high, at over 4800, but on 2 days of weekend, the traffic volume down significantly. => Another indecator for traffic jam is business day. People have to go to work.

In [20]:
```
day['hour']=day['date_time'].dt.hour
business_day=day.copy()[day['dayofweek'] <=4]
weekend_day=day.copy()[day['dayofweek'] >4]
by_hour_business = business_day.groupby('hour').mean(numeric_only=True)
by_hour_weekend = weekend_day.groupby('hour').mean(numeric_only=True)

print(by_hour_business)
plt.figure(figsize=(10,12))
plt.subplot(3,2,1)
plt.plot(by_hour_business['traffic_volume'])
plt.xlabel('Traffic Volume')
plt.ylabel('Frequency')
plt.xlim(6,19)
plt.ylim(0,6250)
plt.title('TL in Business days')


plt.subplot(3,2,2)
plt.plot(by_hour_weekend['traffic_volume'])
plt.xlabel('Traffic Volume')
plt.ylabel('Frequency')
plt.xlim(6,19)
plt.ylim(0,6250)
plt.title('TL in Weekend days')
```

hour	temp	rain_1h	snow_1h	clouds_all	traffic_volume	month	year	dayofweek
7	278.662639	0.145105	0.000068	50.538983	6030.413559	6.363390	2015.562712	1.984407
8	278.938443	0.144614	0.000135	53.666441	5503.497970	6.567659	2015.493234	1.989175
9	279.628421	0.156829	0.000139	53.619709	4895.269257	6.484386	2015.548924	1.981263
10	280.664650	0.113984	0.000033	54.781417	4378.419118	6.481283	2015.526738	1.957888
11	281.850231	0.151976	0.000000	52.808876	4633.419470	6.448819	2015.528275	1.979957
12	282.832763	0.090271	0.001543	53.855714	4855.382143	6.569286	2015.550000	1.989286
13	283.292447	0.092433	0.000370	53.325444	4859.180473	6.465237	2015.514053	1.982988
14	284.091787	0.102991	0.000746	55.326531	5152.995778	6.588318	2015.501056	1.990852
15	284.450605	0.090036	0.000274	54.168467	5592.897768	6.541397	2015.509719	1.962563
16	284.399011	0.118180	0.000632	54.444132	6189.473647	6.580464	2015.483486	1.995081
17	284.263033	7.299358	0.000000	55.204960	5784.827133	6.510576	2015.482859	1.994165
18	284.388061	0.121533	0.000125	54.183079	4434.209431	6.529126	2015.529126	1.988211

Out[20]: Text(0.5, 1.0, 'TL in Weekend days')



We can see clearly that 7-8h am and 16-17h are rush periods, the traffic volume is over 5000. In other side, these period are also is lowest traffic volume in weekend days.

=> we can conclude that around 7am and 16h when people travel between home and company are the densest.

form 10-14h on business day, the TL is maintained under 5000. And on weekend is around 4000.

Summarize all indecators:

- The traffic is usually heavier during warm months (March–October) compared to cold months (November–February).
- The traffic is usually heavier on business days compared to weekends.
- On business days, the rush hours are around 7 and 16.

Weather Indicators

Another possible indicator of heavy traffic is weather. The dataset provides us with a few useful columns about weather: temp, rain_1h, snow_1h, clouds_all, weather_main, weather_description.

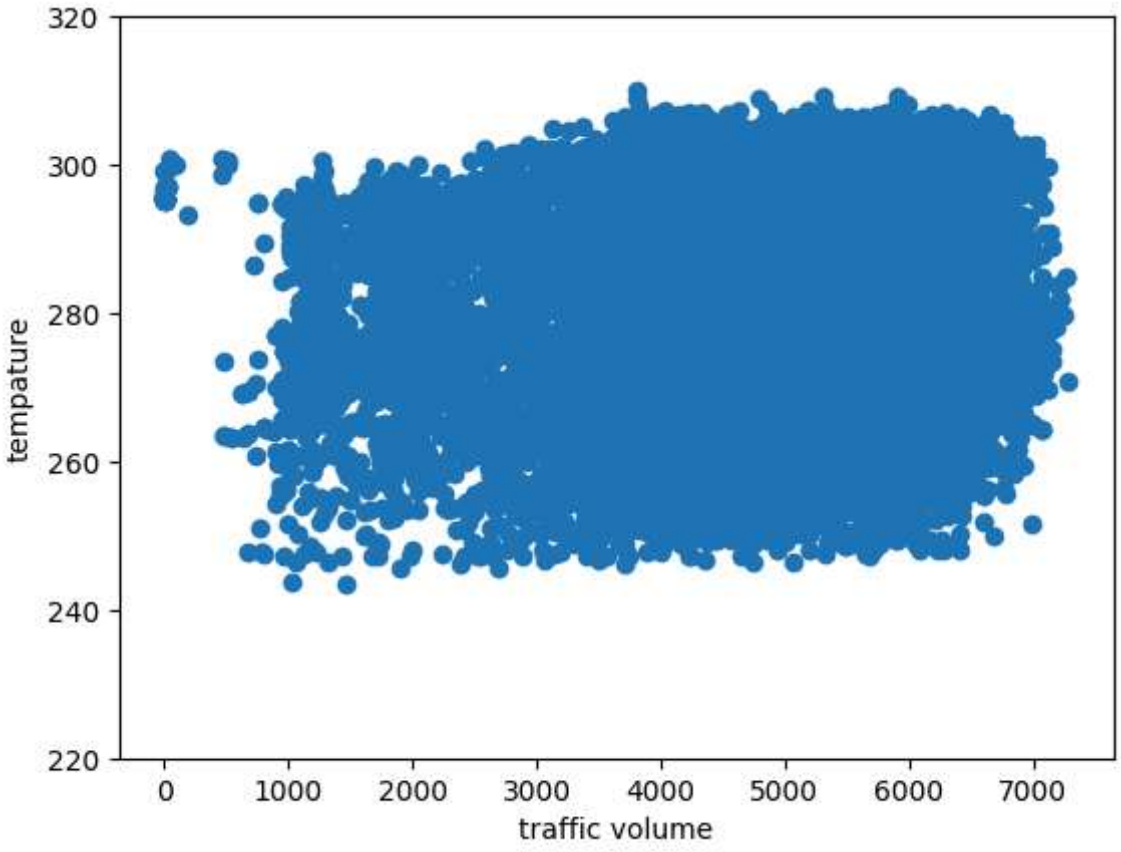
A few of these columns are numerical so let's start by looking up their correlation values with traffic_volume.

In [14]:
```
day.corr(numeric_only=True)['traffic_volume']
```

Out[14]:
```
temp              0.128317
rain_1h           0.003697
snow_1h           0.001265
clouds_all       -0.032932
traffic_volume    1.000000
month            -0.022337
year             -0.003557
dayofweek        -0.416453
hour              0.172704
Name: traffic_volume, dtype: float64
```

Excepting dayofweek, hour, and temp, other columes are not relevant with traffic volume.

In [15]:
```
plt.scatter(day['traffic_volume'], day['temp'])
plt.ylim(220,320)
plt.xlabel('traffic volume')
plt.ylabel('tempature')
plt.show()
```

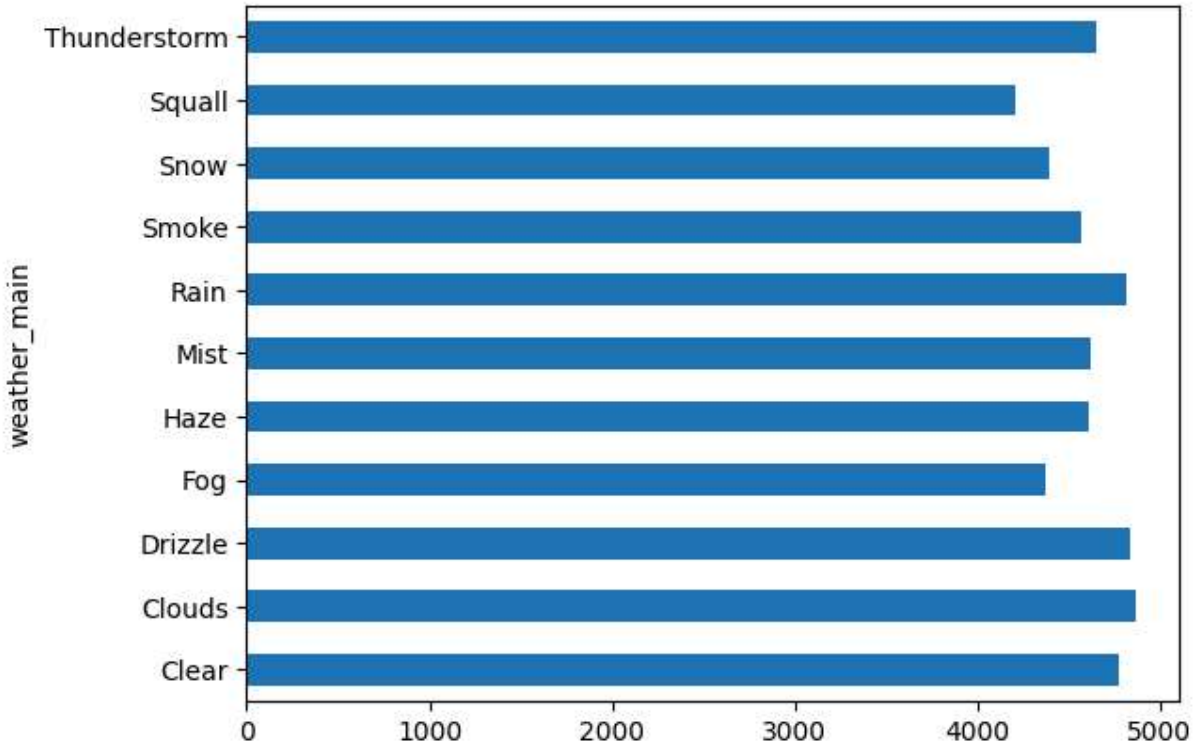


Provisional conclusion is that temp is not reliale indicator because, no sign shows that temp raise or down making traffic volume move.

To see if we can find more useful data, we'll look next at the categorical weather-related columns: weather_main and weather_description.

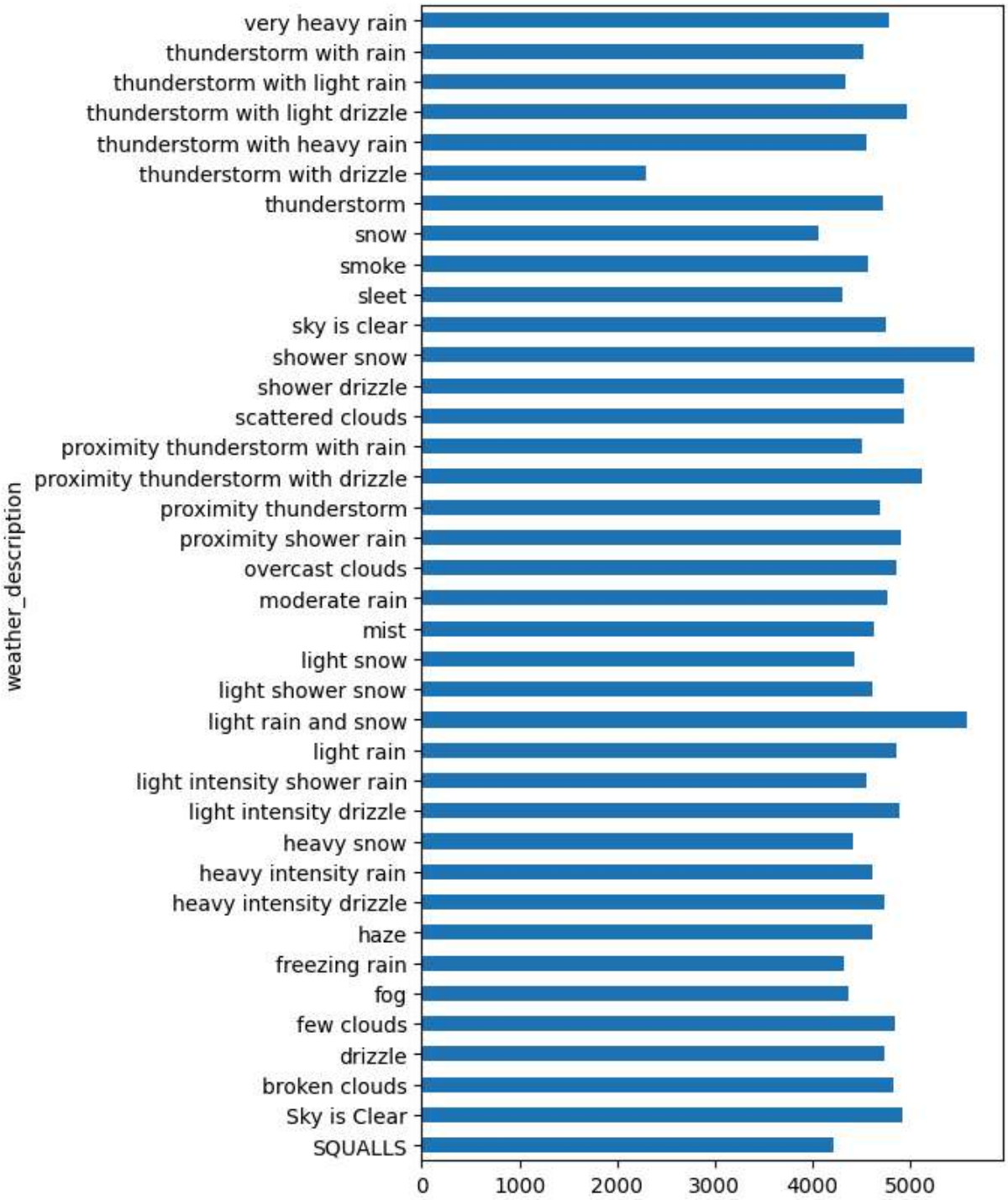
In [16]:
```
by_weather_main = day.groupby('weather_main').mean(numeric_only=True)
by_weather_description = day.groupby('weather_description').mean(numeric_only=True)
```

In [17]:
```
by_weather_main['traffic_volume'].plot.barh()
plt.show()
```



No colume exceeds 5000 cars and kinds of weather can effect lightly to the traffic volume, but it is not enough huge.

In [18]:
```
by_weather_description['traffic_volume'].plot.barh(figsize=(5,10))
plt.show()
```



'shower snow', 'light rain and snow', 'proximity thunderstorm with drizzle' are 3 indicators exceeding 5000 cars show that snow weather impact significantly to traffic

Conclusion

In this project, we tried to find a few indicators of heavy traffic on the I-94 Interstate highway. We managed to find two types of indicators:

Time indicators

- The traffic is usually heavier during warm months (March–October) compared to cold months (November–February).
- The traffic is usually heavier on business days compared to the weekends.
- On business days, the rush hours are around 7 and 16.

Weather indicators

- Shower snow
- Light rain and snow
- Proximity thunderstorm with drizzle