Finding Heavy Traffic Indicators on I-94

In this project, I am going to analyze a dataset about the westbound traffic on the I-94 Interstate highway.

The goal of this 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.

The I-94 Traffic Dataset

Dataset link: https://archive.ics.uci.edu/dataset/492/metro+interstate+traffic+volume

In [2]:
```
import pandas as pd
traffic = pd.read_csv('Metro_Interstate_Traffic_Volume.csv')
traffic.head(5)
```

Out[2]:

	holiday	temp	rain_1h	snow_1h	clouds_all	weather_main	weather_description	date_time	traffic_volume
0	None	288.28	0.0	0.0	40	Clouds	scattered clouds	2012-10-02 09:00:00	5545
1	None	289.36	0.0	0.0	75	Clouds	broken clouds	2012-10-02 10:00:00	4516
2	None	289.58	0.0	0.0	90	Clouds	overcast clouds	2012-10-02 11:00:00	4767
3	None	290.13	0.0	0.0	90	Clouds	overcast clouds	2012-10-02 12:00:00	5026
4	None	291.14	0.0	0.0	75	Clouds	broken clouds	2012-10-02 13:00:00	4918

In [3]:
```
traffic.tail(5)
```

Out[3]:

	holiday	temp	rain_1h	snow_1h	clouds_all	weather_main	weather_description	date_time	traffic_volume
48199	None	283.45	0.0	0.0	75	Clouds	broken clouds	2018-09-30 19:00:00	3543
48200	None	282.76	0.0	0.0	90	Clouds	overcast clouds	2018-09-30 20:00:00	2781
48201	None	282.73	0.0	0.0	90	Thunderstorm	proximity thunderstorm	2018-09-30 21:00:00	2159
48202	None	282.09	0.0	0.0	90	Clouds	overcast clouds	2018-09-30 22:00:00	1450
48203	None	282.12	0.0	0.0	90	Clouds	overcast clouds	2018-09-30 23:00:00	954

In [4]:
```
traffic.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48204 entries, 0 to 48203
Data columns (total 9 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   holiday              48204 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
```

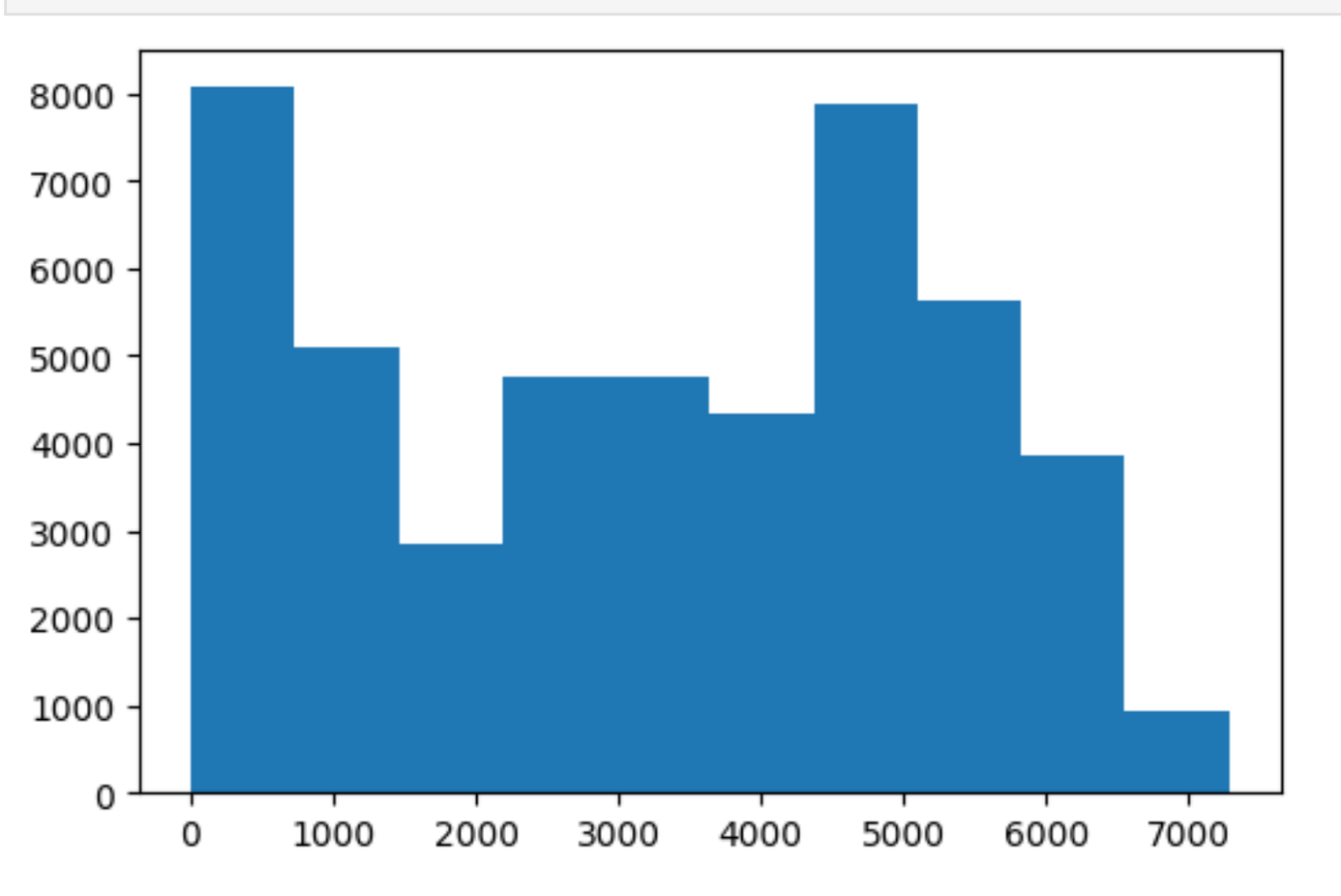
The dataset has 48,204 rows and 9 columns, and there are no null values. Each row describes traffic and weather data for a specific hour — we have data from 2012-10-02 09:00:00 until 2018-09-30 23:00:00.

A station located approximately midway between Minneapolis and Saint Paul records the traffic data. For this station, the direction of the route is westbound (i.e., cars moving from east to west). This means that the results of this analysis will be about the westbound traffic in the proximity of the station. In other words, I will avoid generalizing the results for the entire I-94 highway.

Analyzing Traffic Volume

I am going to start this analysis by examining the distribution of the `traffic_volume` column.

In [5]:
```
import matplotlib.pyplot as plt
%matplotlib inline

plt.hist(traffic['traffic_volume'])
plt.plot()
plt.show()
```



In [6]:
```
traffic.describe()
```

Out[6]:

	temp	rain_1h	snow_1h	clouds_all	traffic_volume
count	48204.000000	48204.000000	48204.000000	48204.000000	48204.000000
mean	281.205870	0.334264	0.000222	49.362231	3259.818355
std	13.338232	44.789133	0.008168	39.015750	1986.860670
min	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
max	310.070000	9831.300000	0.510000	100.000000	7280.000000

Between 2012-10-02 09:00:00 and 2018-09-30 23:00:00, the hourly traffic volume varied from 0 to 7,280 cars, with an average of 3,260 cars.

About 25% of the time, there were only 1,193 cars or fewer passing the station each hour — this probably occurs during the night, or when a road is under construction. However, about 25% of the time, the traffic volume was four times as much (4,933 cars or more).

This observation gives this analysis an interesting direction: comparing daytime data with nighttime data.

Traffic Volume: Day vs. Night

I will start by dividing the dataset into two parts:

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

In [7]:
```
pd.to_datetime(traffic['date_time'])
```

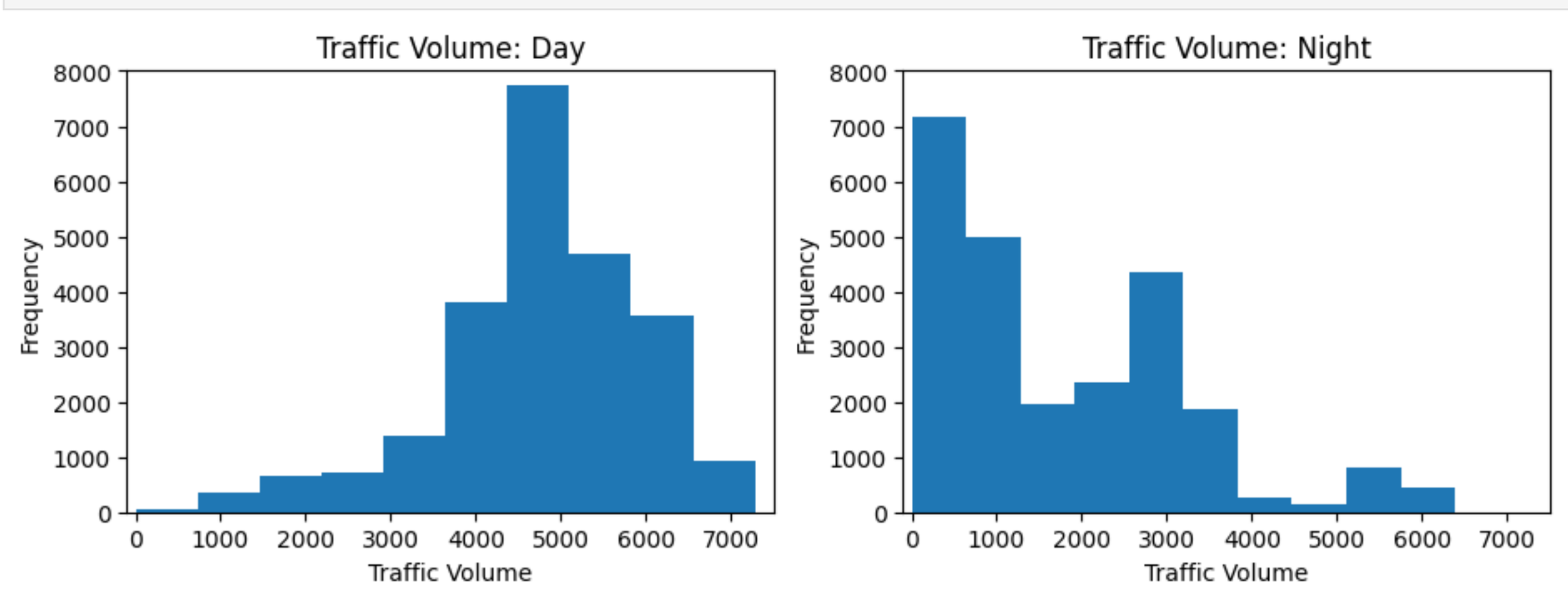
Out[7]:
```
0       2012-10-02 09:00:00
1       2012-10-02 10:00:00
2       2012-10-02 11:00:00
3       2012-10-02 12:00:00
4       2012-10-02 13:00:00
                ...
48199   2018-09-30 19:00:00
48200   2018-09-30 20:00:00
48201   2018-09-30 21:00:00
48202   2018-09-30 22:00:00
48203   2018-09-30 23:00:00
Name: date_time, Length: 48204, dtype: datetime64[ns]
```

In [8]:
```
traffic['date_time'] = pd.to_datetime(traffic['date_time'])
traffic['hour'] = traffic['date_time'].dt.hour
daytime_data = traffic[(traffic['hour'] >= 7) & (traffic['hour'] < 19)]
nighttime_data = traffic[(traffic['hour'] < 7) | (traffic['hour'] >= 19)]

print(daytime_data.shape)
print(nighttime_data.shape)
```

```
(23877, 10)
(24327, 10)
```

This significant difference in row numbers between day and night is due to a few hours of missing data. For instance, looking at rows 176 and 177 (i_94.iloc[176:178]), there's no data for two hours (4 and 5).

In [9]:
```
plt.figure(figsize=(11,3.5))

plt.subplot(1, 2, 1)
plt.hist(daytime_data['traffic_volume'])
plt.xlim(-100, 7500)
plt.ylim(0, 8000)
plt.title('Traffic Volume: Day')
plt.ylabel('Frequency')
plt.xlabel('Traffic Volume')

plt.subplot(1, 2, 2)
plt.hist(nighttime_data['traffic_volume'])
plt.xlim(-100, 7500)
plt.ylim(0, 8000)
plt.title('Traffic Volume: Night')
plt.ylabel('Frequency')
plt.xlabel('Traffic Volume')

plt.show()
```



In [10]:
```
daytime_data['traffic_volume'].describe()
```

Out[10]:
```
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 [11]:
```
nighttime_data['traffic_volume'].describe()
```

Out[11]:
```
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
```

The histogram that shows the distribution of traffic volume during the day is left skewed. This means that most of the traffic volume values are high — there are 4,252 or more cars passing the station each hour 75% of the time (because 25% of values are less than 4,252).

The histogram displaying the nighttime data is right skewed. This means that most of the traffic volume values are low — 75% of the time, the number of cars that passed the station each hour was less than 2,819.

Although there are still measurements of over 5,000 cars per hour, the traffic at night is generally light. The goal is to find indicators of heavy traffic, so I will only focus on the daytime data moving forward.

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.

I am 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

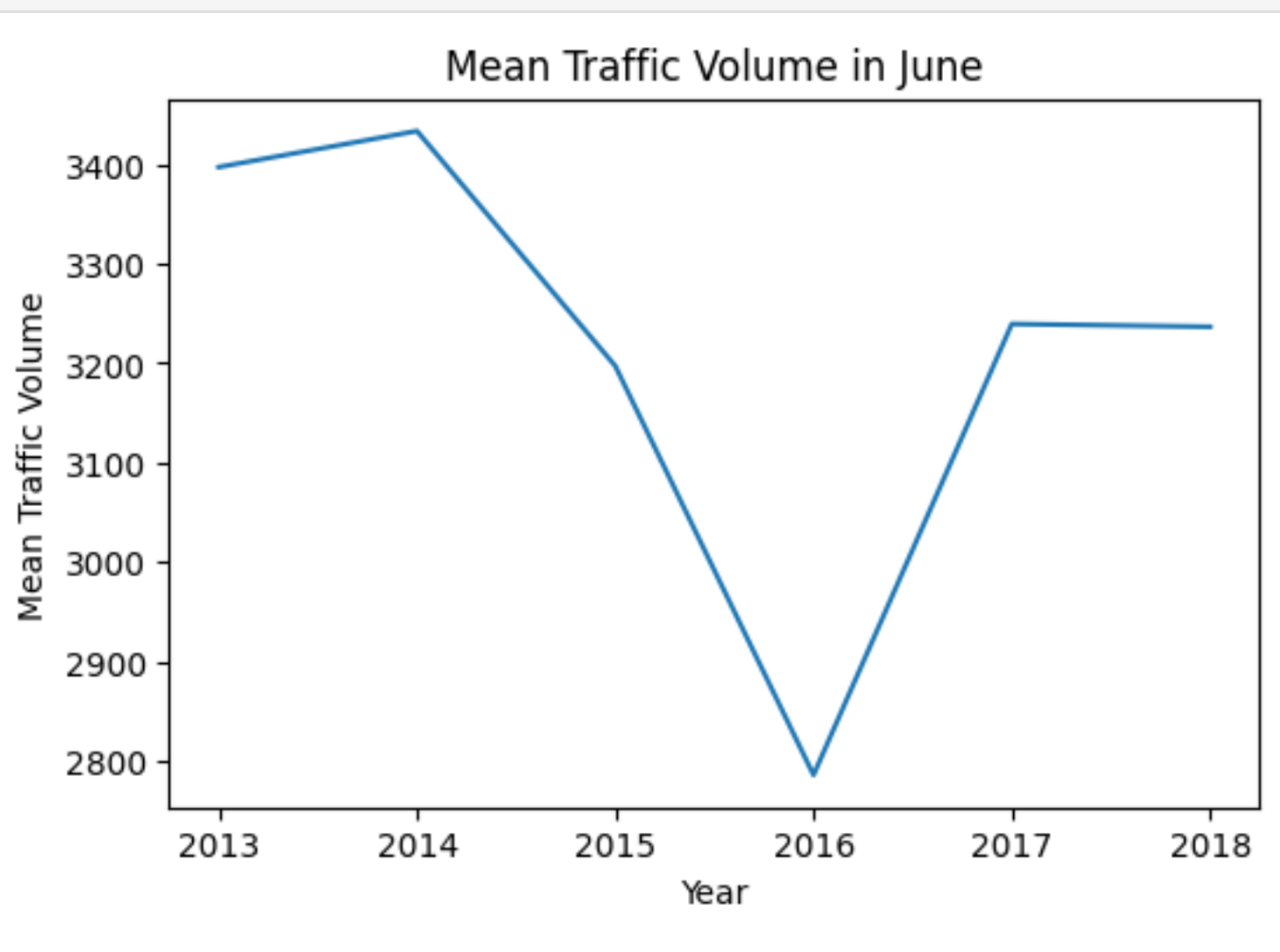
In [12]:
```
daytime_data = daytime_data.copy()
daytime_data['month'] = daytime_data['date_time'].dt.month
by_month = daytime_data.groupby('month').mean()
by_month['traffic_volume']
```

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

The traffic looks less heavy during cold months (November–February) and more intense during warm months (March–October), with one interesting exception: July. Is there anything special about July? Is traffic significantly less heavy in July each year?

To answer the last question, I will see how the traffic volume changed each year in July.

In [14]:
```
june_data = traffic[traffic['date_time'].dt.month == 7]
grouped_data = june_data.groupby(june_data['date_time'].dt.year)['traffic_volume'].mean()

plt.plot(grouped_data.index, grouped_data.values)

plt.xlabel('Year')
plt.ylabel('Mean Traffic Volume')
plt.title('Mean Traffic Volume in June')


plt.show()
```

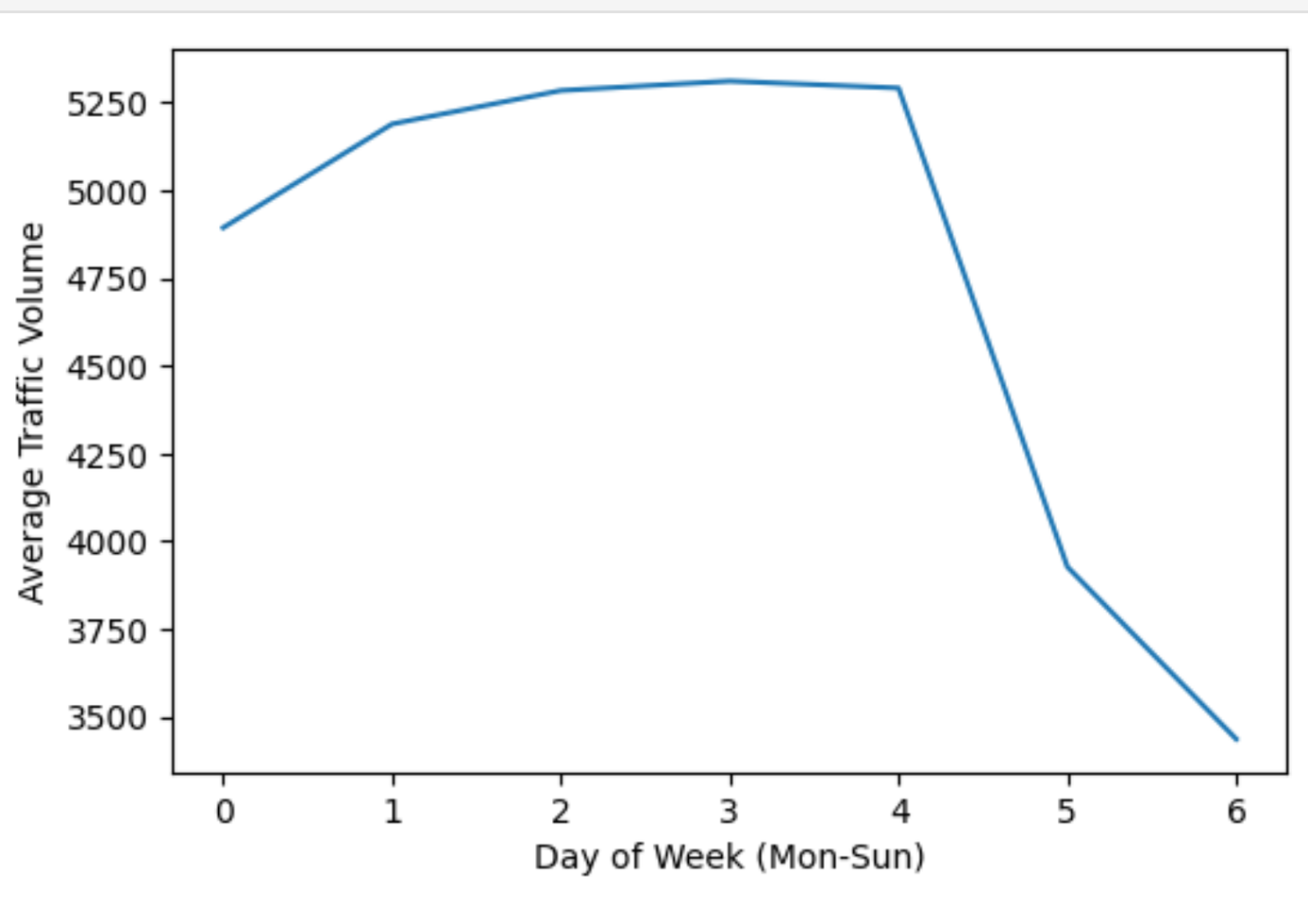


Typically, the traffic is pretty heavy in July, similar to the other warm months. The only exception we see is 2016, which had a high decrease in traffic volume. One possible reason for this is road construction — this article (https://www.crainsdetroit.com/article/20160728/NEWS/160729841/weekend-construction-i-96-us-23-bridge-work-i-94-lane-closures-i-696) from 2016 supports this hypothesis.

As a tentative conclusion here, I can say that warm months generally show heavier traffic compared to cold months. In a warm month, you can can expect for each hour of daytime a traffic volume close to 5,000 cars.

Now I will look at a more granular indicator: day number.

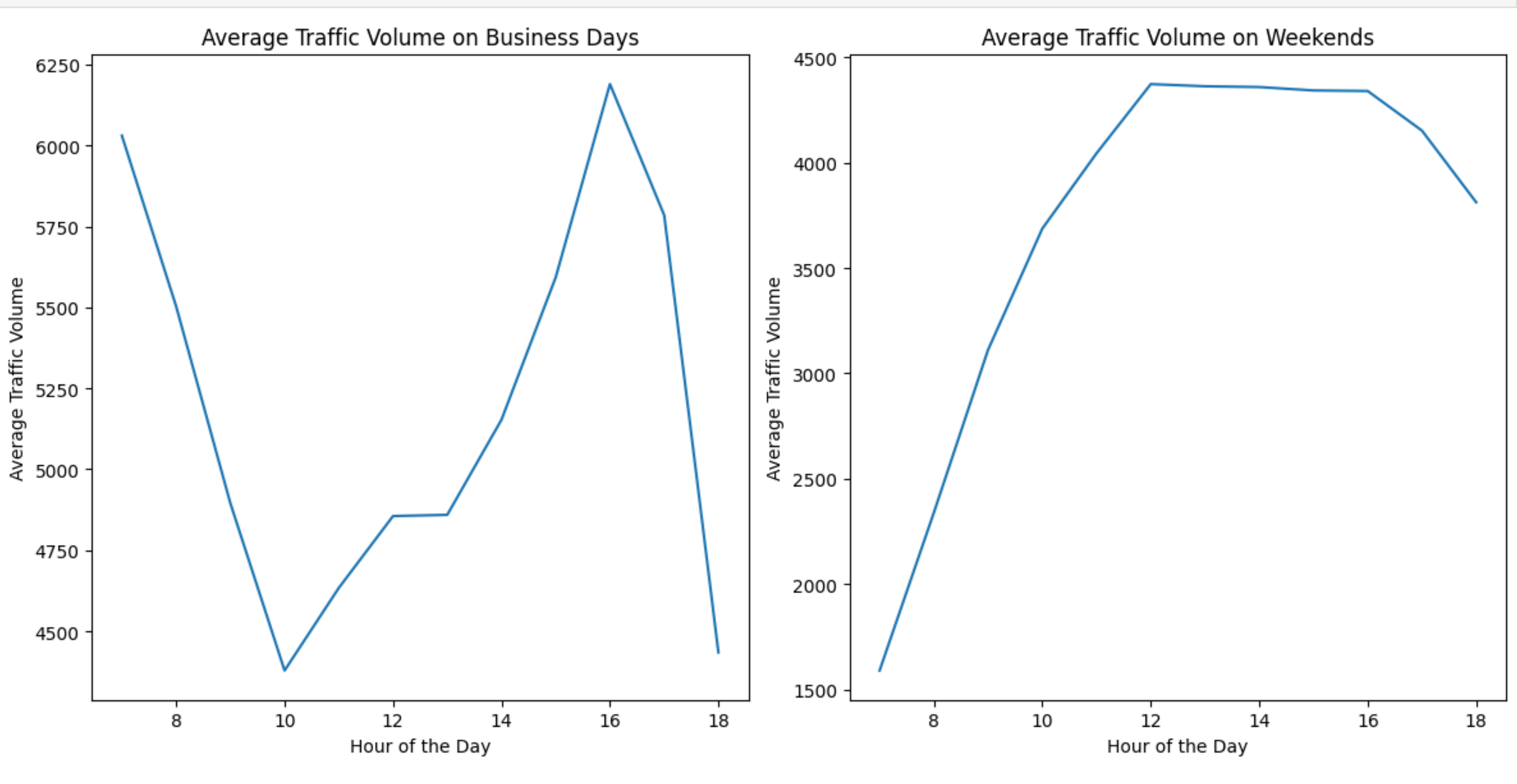
In [15]:
```
daytime_data['dayofweek'] = daytime_data['date_time'].dt.dayofweek
by_dayofweek = daytime_data.groupby('dayofweek').mean()
by_dayofweek['traffic_volume']  # 0 is Monday, 6 is Sunday

plt.plot(by_dayofweek['traffic_volume'])
plt.xlabel('Day of Week (Mon-Sun)')
plt.ylabel('Average Traffic Volume')
plt.show()
```



Traffic volume is significantly heavier on business days (Monday – Friday). Except for Monday, we only see values over 5,000 during business days. Traffic is lighter on weekends, with values below 4,000 cars.

Now I will see what values we have based on time of the day. The weekends, however, will drag down the average values, I am going to look only at the averages separately.

In [16]:
```
# Splitting data into business days and weekends
business_days = daytime_data[daytime_data['dayofweek'] < 5]  # Monday to Friday
weekends = daytime_data[daytime_data['dayofweek'] >= 5]  # Saturday and Sunday

# Create a figure with two subplots
fig, axes = plt.subplots(1, 2, figsize=(12, 6))

# Plot for business days
axes[0].plot(business_days.groupby('hour')['traffic_volume'].mean().sort_index())
axes[0].set_xlabel('Hour of the Day')
axes[0].set_ylabel('Average Traffic Volume')
axes[0].set_title('Average Traffic Volume on Business Days')

# Plot for weekends
axes[1].plot(weekends.groupby('hour')['traffic_volume'].mean().sort_index())
axes[1].set_xlabel('Hour of the Day')
axes[1].set_ylabel('Average Traffic Volume')
axes[1].set_title('Average Traffic Volume on Weekends')

# Adjust spacing between subplots
plt.tight_layout()

# Show the plots
plt.show()
```



At each hour of the day, the traffic volume is generally higher during business days compared to the weekends. As somehow expected, the rush hours are around 7 and 16 — when most people travel from home to work and back. We see volumes of over 6,000 cars at rush hours.

To summarize, there are a few time-related indicators of heavy traffic:

- 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

In [17]:
```
traffic.head(5)
```

Out[17]:

	holiday	temp	rain_1h	snow_1h	clouds_all	weather_main	weather_description	date_time	traffic_volume	hour
0	None	288.28	0.0	0.0	40	Clouds	scattered clouds	2012-10-02 09:00:00	5545	9
1	None	289.36	0.0	0.0	75	Clouds	broken clouds	2012-10-02 10:00:00	4516	10
2	None	289.58	0.0	0.0	90	Clouds	overcast clouds	2012-10-02 11:00:00	4767	11
3	None	290.13	0.0	0.0	90	Clouds	overcast clouds	2012-10-02 12:00:00	5026	12
4	None	291.14	0.0	0.0	75	Clouds	broken clouds	2012-10-02 13:00:00	4918	13

Another possible indicator of heavy traffic is weather. The dataset provides 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 I will start by looking up their correlation values with traffic_volume.

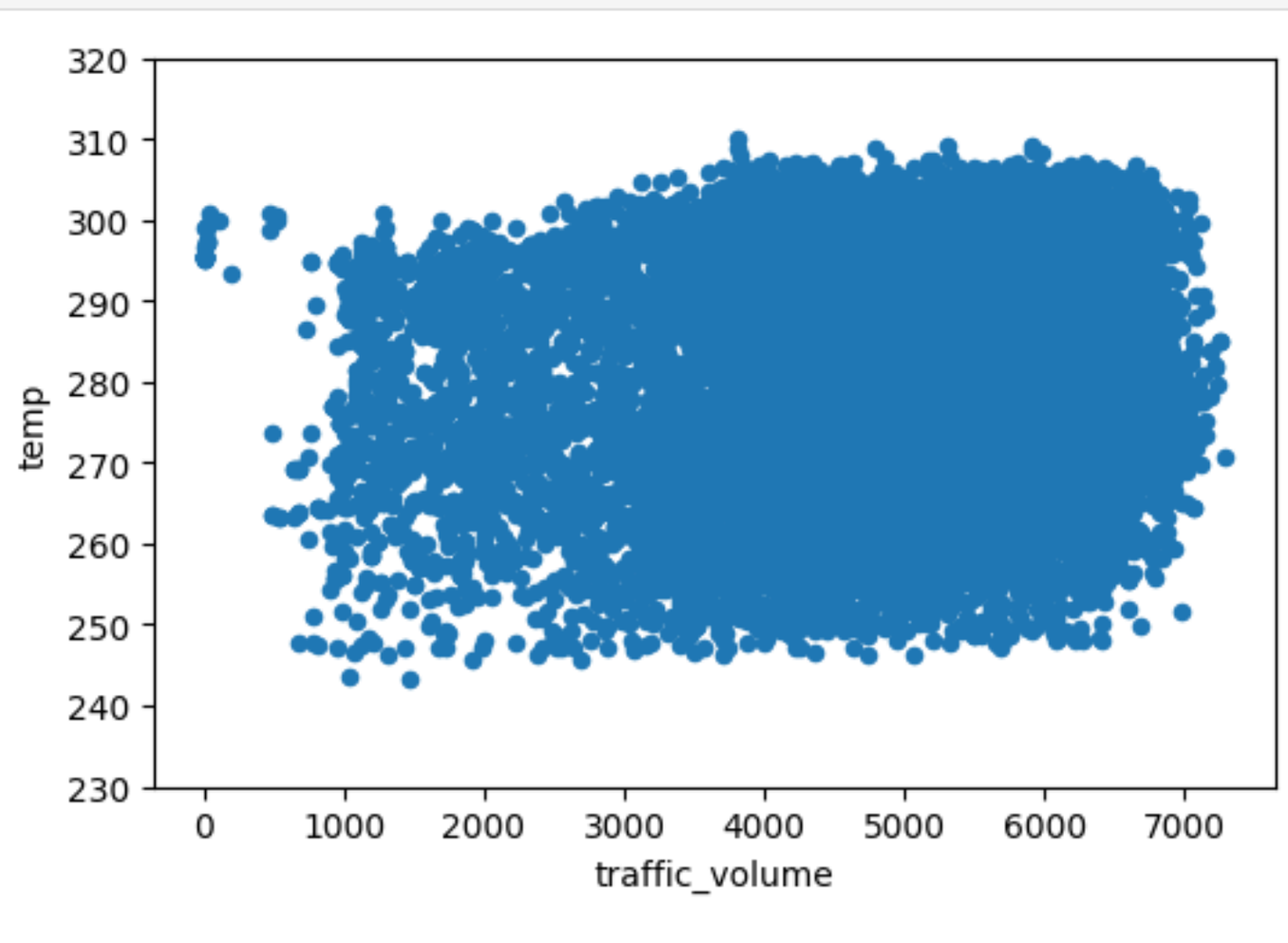
In [18]:
```
daytime_data.corr()['traffic_volume']
```

Out[18]:
```
temp              0.128317
rain_1h           0.003697
snow_1h           0.001265
clouds_all       -0.032932
traffic_volume    1.000000
hour              0.172704
month            -0.022337
dayofweek        -0.416453
Name: traffic_volume, dtype: float64
```

Temperature shows the strongest correlation with a value of just +0.13. The other relevant columns (`rain_1h`, `snow_1h`, `clouds_all`) don't show any strong correlation with traffic_value.

Now, I will generate a scatter plot to visualize the correlation between temp and traffic_volume.

In [26]:
```
daytime_data.plot.scatter('traffic_volume', 'temp')
plt.ylim(230, 320) # two wrong 0K temperatures mess up the y-axis
plt.show()
```

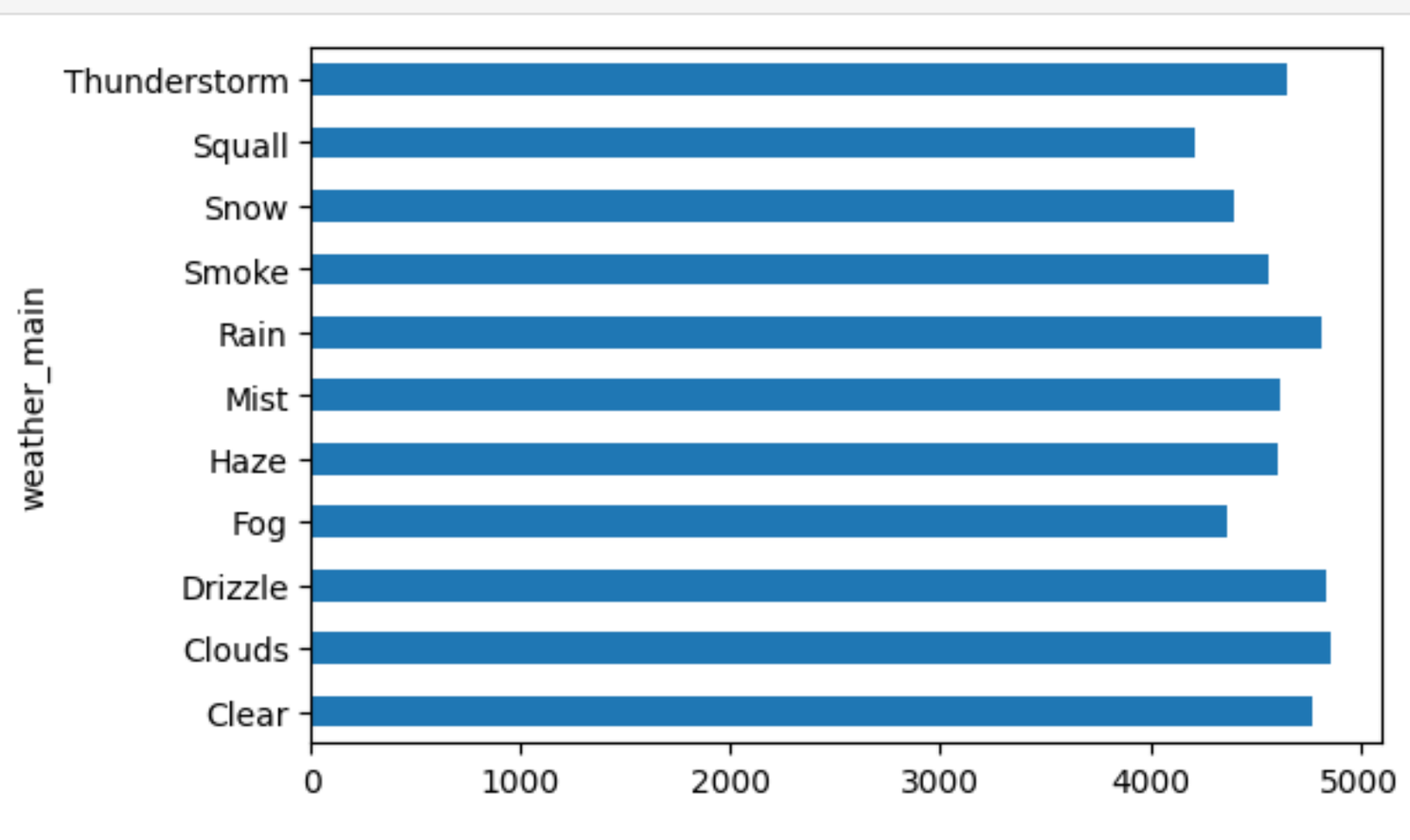


This scatterplot shows that temperature doesn't look like a solid indicator of heavy traffic.

Now, I will look at the other weather-related columns: `weather_main` and `weather_description`.

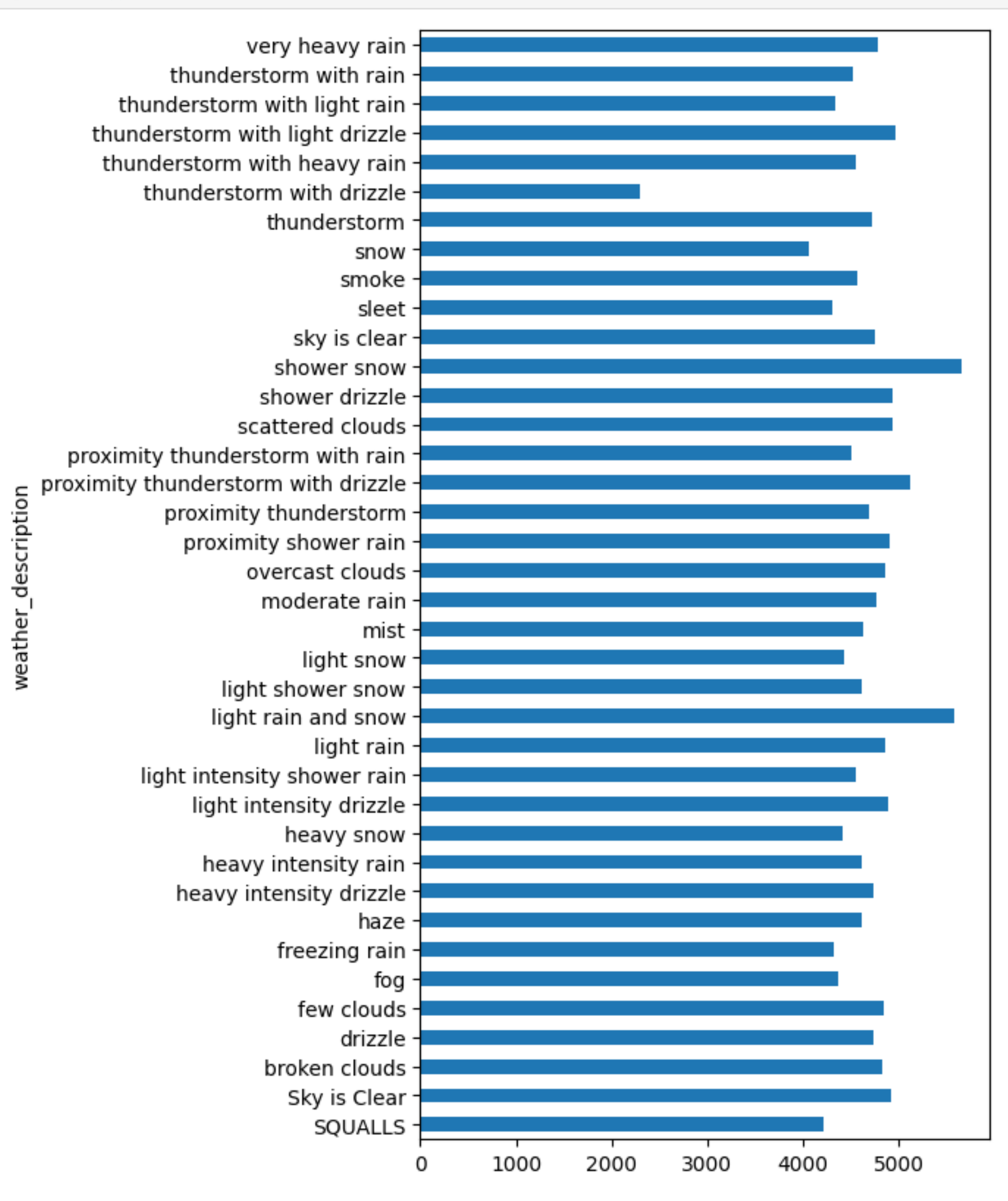
Weather Types

In [30]:
```
by_weather_main = daytime_data.groupby('weather_main').mean()
by_weather_main['traffic_volume'].plot.barh()
plt.show()
```



It looks like there's no weather type where traffic volume exceeds 5,000 cars. This makes finding a heavy traffic indicator more difficult. I will also group by `weather_description`, which has a more granular weather classification.

In [31]:
```
by_weather_description = daytime_data.groupby('weather_description').mean()
by_weather_description['traffic_volume'].plot.barh(figsize=(5,10))
plt.show()
```



It looks like there are three weather types where traffic volume exceeds 5,000:

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

It's not clear why these weather types have the highest average traffic values — this is bad weather, but not that bad. Perhaps more people take their cars out of the garage when the weather is bad instead of riding a bike or walking.

Conclusion

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

1. 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.

1. Weather indicators

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