pd1

March 9, 2021

1 Praca domowa 1

1.0.1 Przemysław Olender

[191]:
```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
```

[192]:
```
fires = pd.read_csv('forest_fires_dataset.csv')

fires.head()
```

[192]:

	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
0	7	5	mar	fri	86.2	26.2	94.3	5.1	8.2	51.0	6.7	0.0	0.0
1	7	4	oct	tue	90.6	35.4	669.1	6.7	18.0	33.0	0.9	0.0	0.0
2	7	4	oct	sat	90.6	43.7	686.9	6.7	14.6	33.0	1.3	0.0	0.0
3	8	6	mar	fri	91.7	33.3	77.5	9.0	8.3	97.0	4.0	0.2	0.0
4	8	6	mar	sun	89.3	51.3	102.2	9.6	11.4	99.0	1.8	0.0	0.0

Analizujemy zbiór danych o pożarach w północnej Portugalii. Według opisu autora cała ramka służy po przewidywania powierzchni pożary za pomocą pozostałych zminnych, są to:

X - x-owa współrzedna wewnątrz parku

Y - y-owa współrzędna wewnątrz parku

month - miesiąc

day - dzień tygodnia

FFMC - wskaźnik wilgotności paliw* ze ściółki leśnej

DMC - wskaźnik wilgotności paliw* znajdujących się poniżej ściółki leśnej

DC - wskaźnik wilgotności paliw* wgłąb gleby

ISI - wskaźnik łączocy wilgotność martwych paliw* i prędkości wiatru, pomaga w oszacowaniu prędkości roznoszenia sie pożaru

temp - temperatura w stopniach celsiujsza

RH - względna wilgotność powietrza wyrażona w %

wind - prędkość wiatru w km/godzinę

rain - wielkość opadów w mm/m^2

area - spalona powierzchnia w hektarach

FFMc, DMC, DC, ISI to indeksy z międzynaroowego systemu FWI (Fire Weather Index) określającego zagrożenie pożarowe (informazje ze strony https://www.nwcg.gov/publications/pms437/cffdrs/fire-weather-index-system).

* - paliwo oznacza materiał podatny na spalenie, w tym przyapdku rośliny

Źródło danych: https://www.apispreadsheets.com/datasets/129

[193]:
```
fires.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 517 entries, 0 to 516
Data columns (total 13 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   X       517 non-null    int64  
 1   Y       517 non-null    int64  
 2   month   517 non-null    object 
 3   day     517 non-null    object 
 4   FFMC    517 non-null    float64
 5   DMC     517 non-null    float64
 6   DC      517 non-null    float64
 7   ISI     517 non-null    float64
 8   temp    517 non-null    float64
 9   RH      517 non-null    float64
 10  wind    517 non-null    float64
 11  rain    517 non-null    float64
 12  area    517 non-null    float64
dtypes: float64(9), int64(2), object(2)
memory usage: 48.5+ KB
```

W danych nie ma żadnych braków, nie będziemy musieli się poźniej martwić wypełniem dziur.

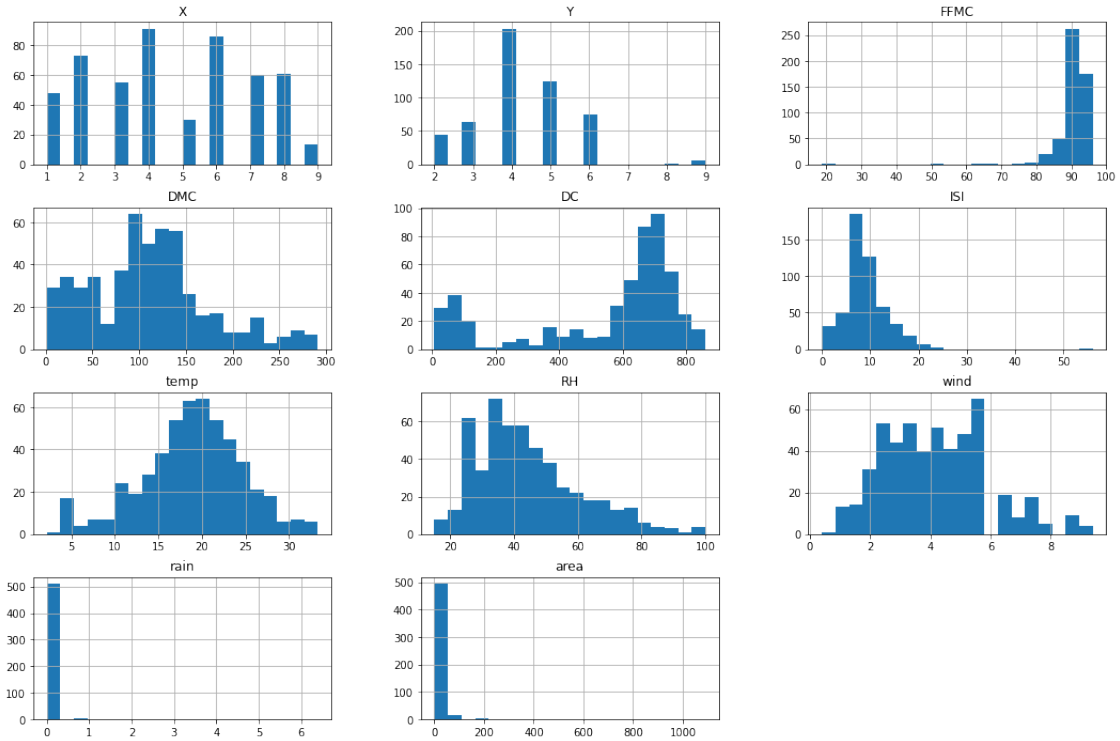
[194]:
```
fires.describe()
```

[194]:

	X	Y	FFMC	DMC	DC	ISI	\
count	517.000000	517.000000	517.000000	517.000000	517.000000	517.000000	
mean	4.669246	4.299807	90.644681	110.872340	547.940039	9.021663	
std	2.313778	1.229900	5.520111	64.046482	248.066192	4.559477	
min	1.000000	2.000000	18.700000	1.100000	7.900000	0.000000	
25%	3.000000	4.000000	90.200000	68.600000	437.700000	6.500000	
50%	4.000000	4.000000	91.600000	108.300000	664.200000	8.400000	
75%	7.000000	5.000000	92.900000	142.400000	713.900000	10.800000	
max	9.000000	9.000000	96.200000	291.300000	860.600000	56.100000	

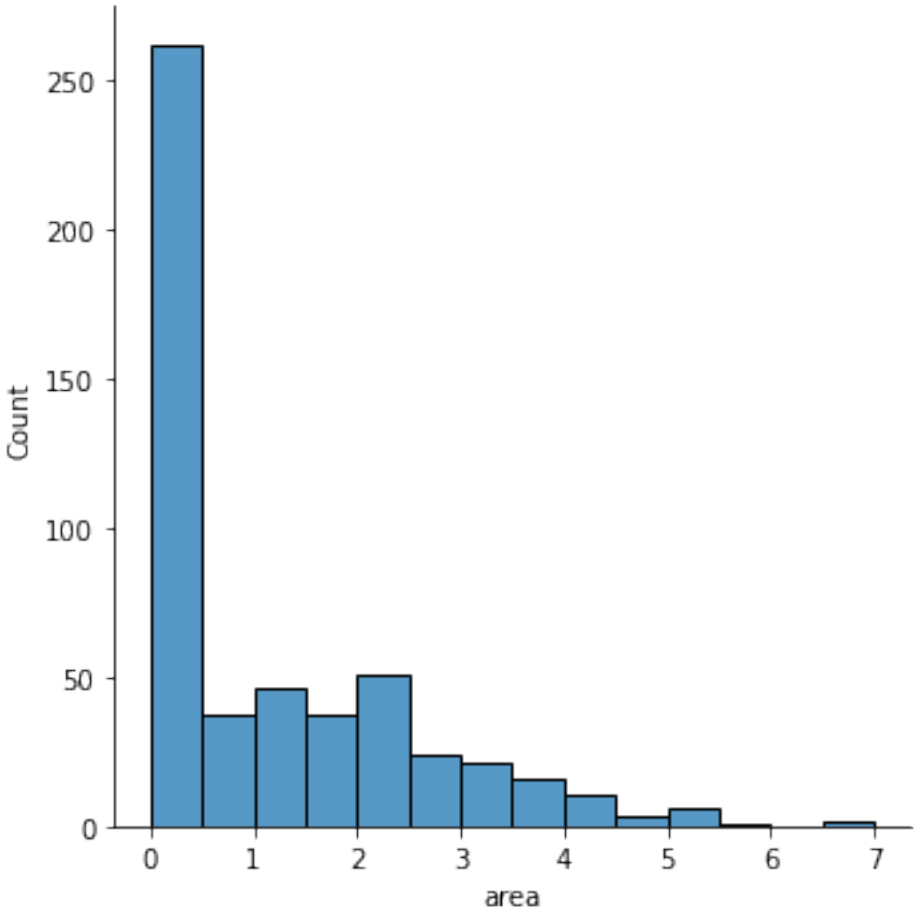
	temp	RH	wind	rain	area
count	517.000000	517.000000	517.000000	517.000000	517.000000
mean	18.889168	44.288201	4.017602	0.021663	12.847292
std	5.806625	16.317469	1.791653	0.295959	63.655818
min	2.200000	15.000000	0.400000	0.000000	0.000000
25%	15.500000	33.000000	2.700000	0.000000	0.000000
50%	19.300000	42.000000	4.000000	0.000000	0.520000
75%	22.800000	53.000000	4.900000	0.000000	6.570000
max	33.300000	100.000000	9.400000	6.400000	1090.840000

[195]:
```
fires.hist(bins = 20, figsize=(18, 12))

plt.show()
```



Jak widać na wykresach, temperatura ma rozkład bliski normalnemu, wskaźniki FWI i RH mają rozkłady skośne. Prawie wszytskie wartości w kolumnie area są badrzo bliskie zero, użyjmy więc skali logarytmicznej, żeby lepiej przyjrzeć się wykresowi.

[196]:
```
sns.displot(np.log1p(fires['area']))
plt.show()
```



Sprawdźmy ile wartości jest równych 0:

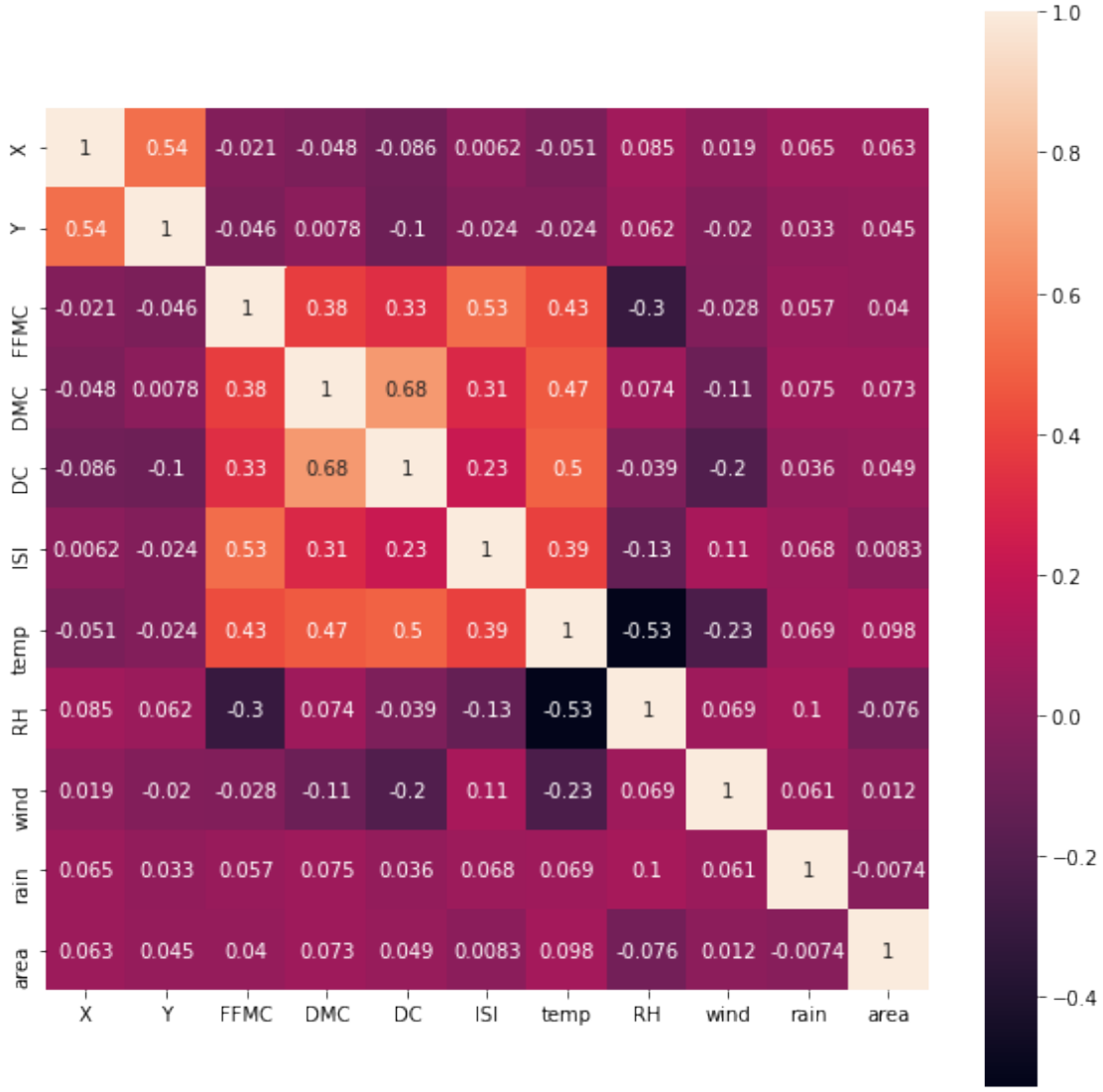
[197]:
```
sum(fires['area'] == 0)
```

[197]: 247

Aż 247 z 517 wartości w kolumnie area to 0, może to oznaczać błąd w danych lub to, że większość pożarów w zaookrągleniu nie stawiło nawet jednego hektara lasu.

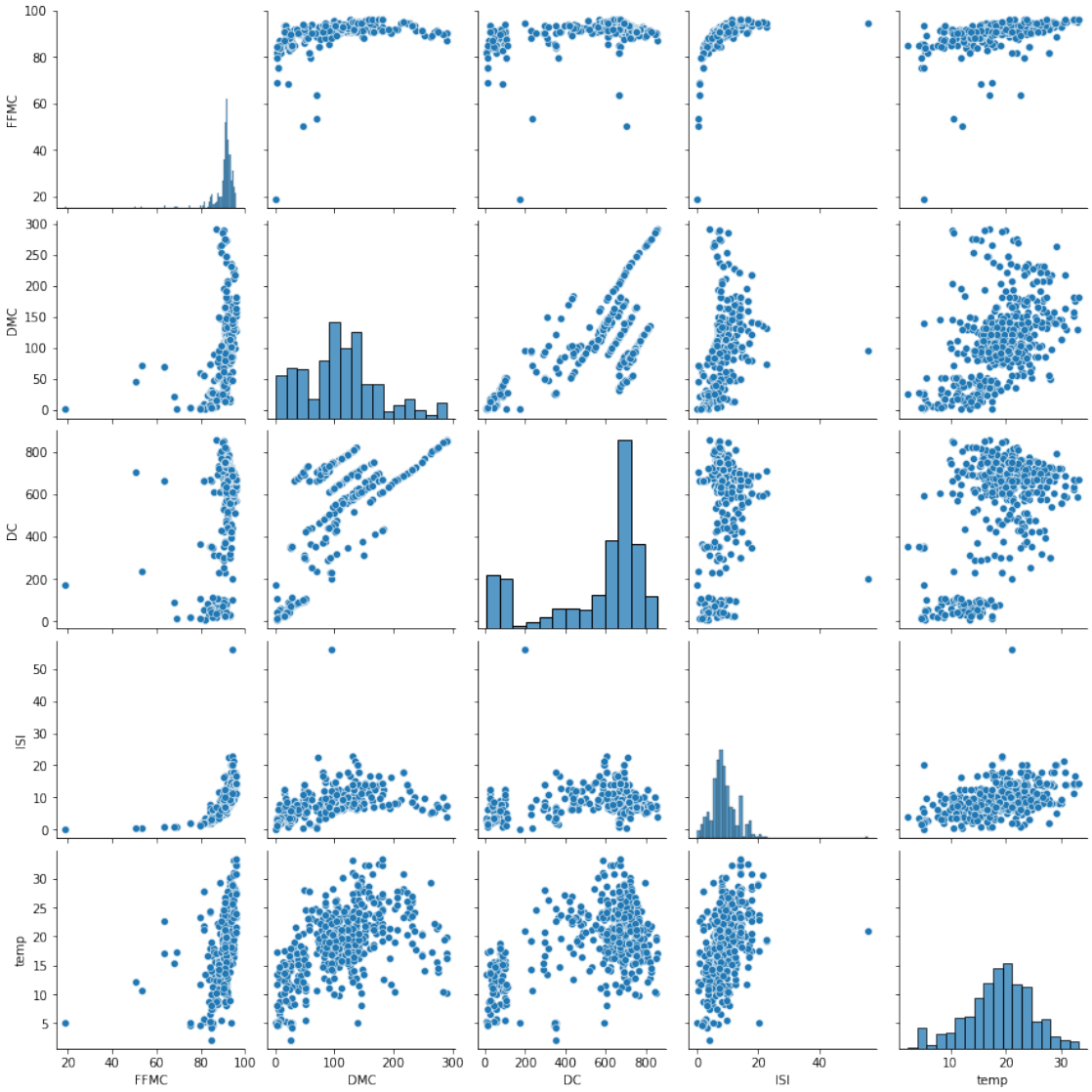
Sprawdźmy jakie korelacje zachodzą miedzy danymi:

[198]:
```
fig, ax = plt.subplots(figsize=(10, 10))
sns.heatmap(fires.corr(), annot=True, square=True)
#fig = plt.figure(fizsize= (10, 10))
plt.show()
```



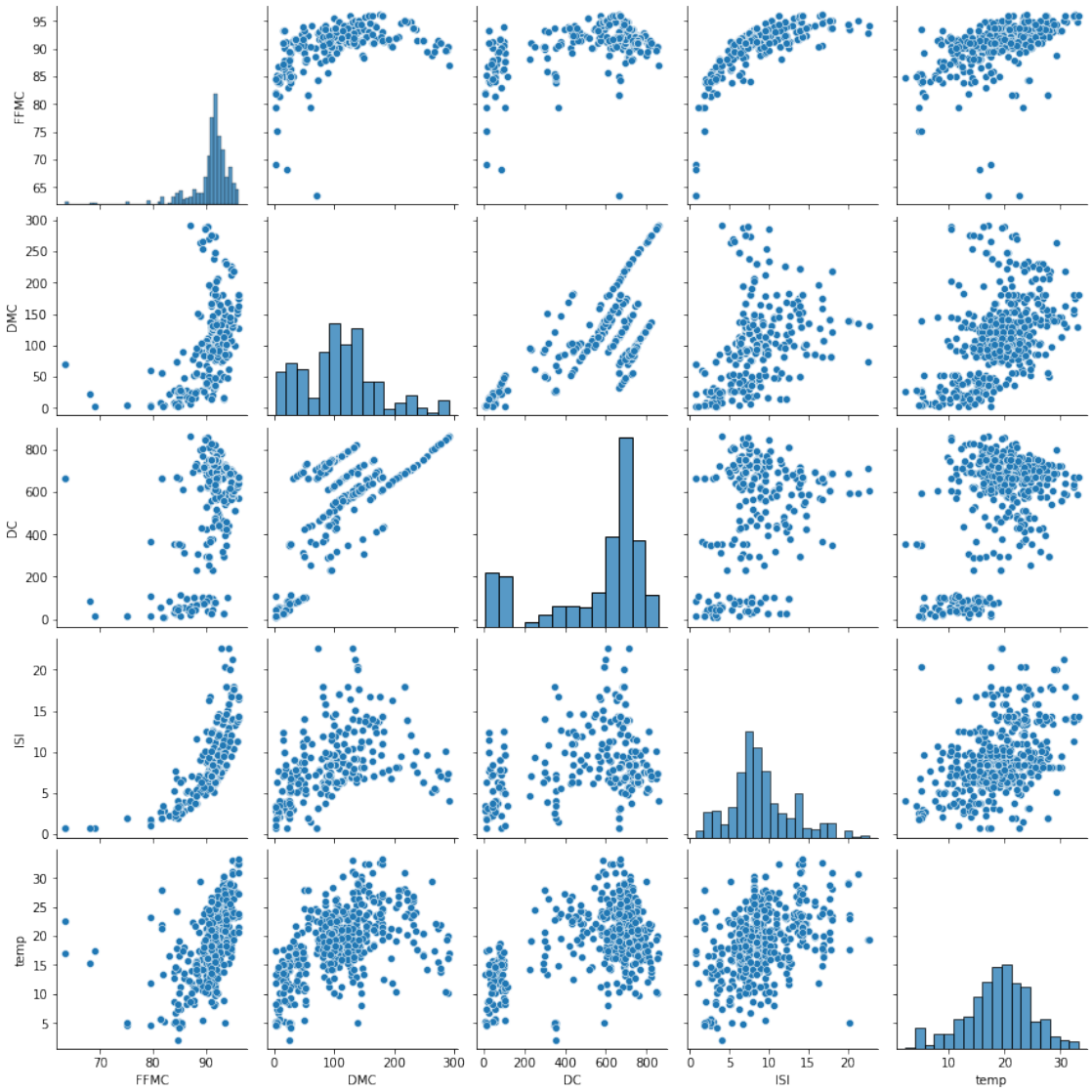
Największe korelacje zachodzą między wskaźnikami FWI i temperaturę, nie jest to duże zaskoczenie, ponieważ im wyższa temperatura tym rośliny są suchsze. Podobnej korelacji można było sie spodziewać między wielkościa opadów (kolumna rain) a wkaźnikamim jednak jak widzieliśmy wcześniej opady prawie nie występują w badanym czasie. Sprawdźmy jak wyglądają scatterploty dla tych zmiennych:

[199]:
```
cols = ['FFMC', 'DMC', 'DC', 'ISI', 'temp']

sns.pairplot(fires, y_vars=fires[cols], x_vars=fires[cols])
plt.show()
```



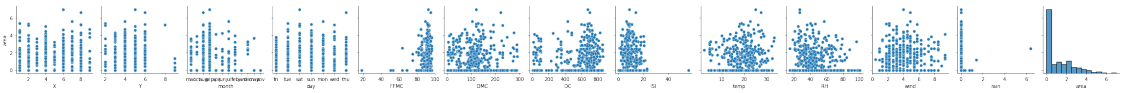
Na wykresach ze zminnymi ISI i FFMC można zauważyć odstające wartość, usuńmy je, żeby nie zaburzały wykresów.

[201]:
```
tmp = fires[fires['ISI'] < 50 ]
tmp = tmp[tmp['FFMC'] > 60]


sns.pairplot(tmp, y_vars=fires[cols], x_vars=fires[cols])
plt.show()
```



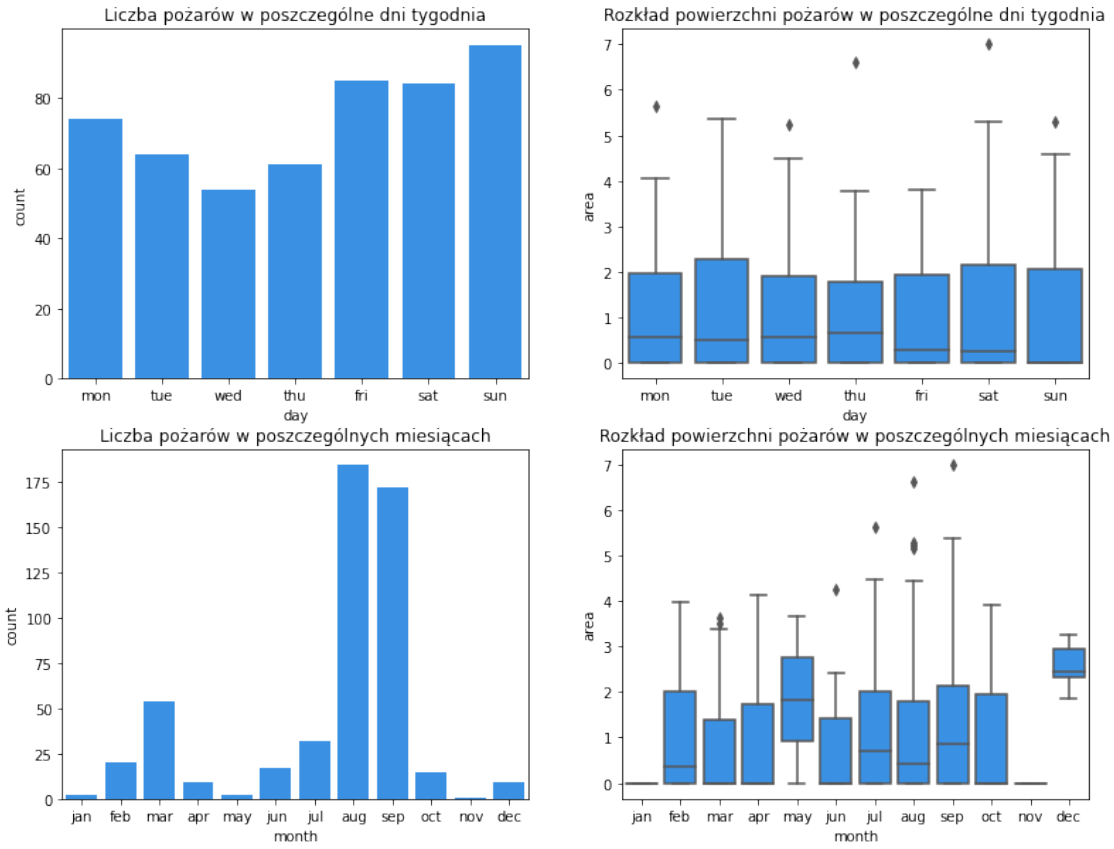
Przyjzyjmy się jeszcze zależnościom zmiennej area od innych zmiennych numerycznych, użyjemu loarytmicznego przekształcenia powirzchni.:

[202]:
```
fires_log = fires
fires_log['area'] = np.log1p(fires['area'])

sns.pairplot(fires_log, y_vars="area", x_vars=fires.columns.values)
plt.show()
```



Sprawdźmy jak pożary o danej powierzchni rozkładaja się w poszczególnych dniach tygodnia i miesiącach. Zoabczmy też, ile pożarów występuje w danych dnia i miesiącach.

[203]:
```
day_order = ['mon', 'tue', 'wed', 'thu', 'fri', 'sat', 'sun']
month_order = ['jan', 'feb', 'mar', 'apr', 'may', 'jun', 'jul', 'aug', 'sep',␣
 ↪'oct', 'nov', 'dec']


fig, ((ax1, ax2), (ax3, ax4)) = plt.subplots(2, 2, figsize=(13, 10))
sns.countplot(data=fires, x='day', order = day_order, color = 'dodgerblue', ax␣
 ↪= ax1)
sns.boxplot(data=fires_log, x='day', y = 'area', order = day_order, color =␣
 ↪'dodgerblue', ax = ax2)
sns.countplot(data=fires, x='month', order = month_order, color = 'dodgerblue',␣
 ↪ax = ax3)
sns.boxplot(data=fires_log, x='month', y= 'area', order = month_order, color =␣
 ↪'dodgerblue', ax = ax4)

ax1.set_title('Liczba pożarów w poszczególne dni tygodnia')
ax2.set_title('Rozkład powierzchni pożarów w poszczególne dni tygodnia')
ax3.set_title('Liczba pożarów w poszczególnych miesiącach')
ax4.set_title('Rozkład powierzchni pożarów w poszczególnych miesiącach')

plt.show()
```



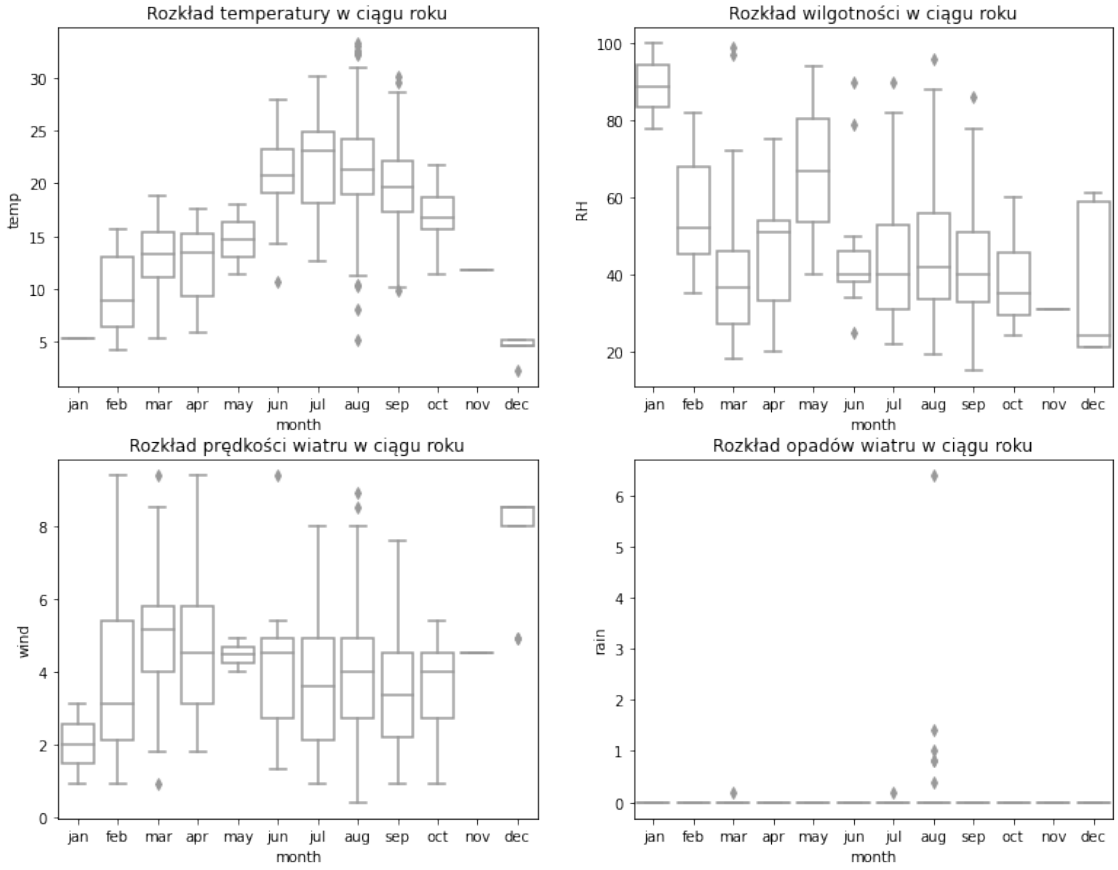
Liczba pożarów jest największa w weekeny, może się to wiązać z działalnością człowieka i większą liczbą odwiedzających park w weekend.

Jeśli chodzi o rozłożenie pożarów w ciągu roku, to w lipicu i sierpieniu liczba pożarów znacznie większa niż w pozostałych miesiącach. Sprawdźmy jak wygląda rozłożenie temperatury i wilgotności w ciągu roku, żeby ocenić powiązanie tych cech z liczbą pożarów.

Jenak i w ciągu poszczególnych dni jak i miesięcy powierzchnia pożarów utrzymuje się na podobnym poziomie.

[204]:
```
fig, ((ax1, ax2), (ax3, ax4)) = plt.subplots(2, 2, figsize=(13, 10))
sns.boxplot(data = fires, x = 'month', y = 'temp', order = month_order, ax =␣
↪ax1, color = 'white')
sns.boxplot(data = fires, x = 'month', y = 'RH', order = month_order, ax = ax2,␣
↪color = 'white')
sns.boxplot(data = fires, x = 'month', y = 'wind', order = month_order, ax =␣
↪ax3, color = 'white')
sns.boxplot(data = fires, x = 'month', y = 'rain', order = month_order, ax =␣
↪ax4, color = 'white')

ax1.set_title('Rozkład temperatury w ciągu roku')
```

```
ax2.set_title('Rozkład wilgotności w ciągu roku')
ax3.set_title('Rozkład prędkości wiatru w ciągu roku')
ax4.set_title('Rozkład opadów wiatru w ciągu roku')
plt.show()
```



Jak widać temperatura w lipcu i sierpniu jest najwyższa w ciągu rokum jednak nie jest dużo większa niż w czerwcu i wrześniu. Wilgotność również znacząco nie odstaje od innych miesięcy letnich.

Widać korelację dużej prekości wiatru w grudniu i dużej powierzchni strawionej pożarem (z poprzedniego wykresu), nawet przy małej liczbie pożarów (mimo, że pożarów było mało to każdy był stosunkowo duży).

Sprawdźmy jeszcze rozłożenie wskaźników FWI w ciągu roku.

[205]:
```
fig, ((ax1, ax2), (ax3, ax4)) = plt.subplots(2, 2, figsize=(13, 10))
sns.boxplot(data = fires, x = 'month', y = 'FFMC', order = month_order, ax =␣
↪ax1, color = 'white')
sns.boxplot(data = fires, x = 'month', y = 'DMC', order = month_order, ax =␣
↪ax2, color = 'white')
sns.boxplot(data = fires, x = 'month', y = 'DC', order = month_order, ax = ax3,␣
↪color = 'white')
```

```
sns.boxplot(data = fires, x = 'month', y = 'ISI', order = month_order, ax =␣
 ↪ax4, color = 'white')

ax1.set_title('Rozkład wskaźnika FFMC w ciągu roku')
ax2.set_title('Rozkład wskaźnika DMC w ciągu roku')
ax3.set_title('Rozkład wskaźnika DC w ciągu roku')
ax4.set_title('Rozkład wskaźnika ISI w ciągu roku')

plt.show()
```