PD1_EksploracjaDanych

March 30, 2021

1 Eksploracja zbioru danych

1.1 Wczytanie danych

Aby uatwi analize zamienimy nazwy dni i miesicy na wartoci liczbowe.

[2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from pandas_profiling import ProfileReport
sns.set()

import calendar

df_fire = pd.read_csv('forest_fires_dataset.csv')
df_fire_att = pd.read_csv('attributes_forest_fires.csv')
```

[3]:
```
days_name = [each_string.lower() for each_string in calendar.day_abbr]
months_name = [each_string.lower() for each_string in calendar.month_abbr]

def dayToNum(day):
    return days_name.index(day) + 1
def monthToNum(month):
    return months_name.index(month)

df_fire.day = df_fire.day.apply(dayToNum)
df_fire.month = df_fire.month.apply(monthToNum)
```

Naszym zadaniej jest przeprowadzenie eksploracji danych zbioru dotyczcego poarów w pónocno-wschodnim regionie Portugalii. Na pocztku rzumy okiem na nasze dane

[4]:
```
df_fire.head()
```

[4]:

	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
0	7	5	3	5	86.2	26.2	94.3	5.1	8.2	51.0	6.7	0.0	0.0
1	7	4	10	2	90.6	35.4	669.1	6.7	18.0	33.0	0.9	0.0	0.0
2	7	4	10	6	90.6	43.7	686.9	6.7	14.6	33.0	1.3	0.0	0.0
3	8	6	3	5	91.7	33.3	77.5	9.0	8.3	97.0	4.0	0.2	0.0
4	8	6	3	7	89.3	51.3	102.2	9.6	11.4	99.0	1.8	0.0	0.0

Opisy poszczególnych kolumn znajduj si w tabeli poniej Dokadne objanienie skrótów FFMC, DMC, DC, ISI znajduj si pod tym linkiem: https://www.nwcg.gov/publications/pms437/cffdrs/fire-weather-index-system

[5]:
```
df_fire_att
```

[5]:

	name	type	description
0	X	integer	x-axis spatial coordinate within the Montesinh...
1	Y	integer	y-axis spatial coordinate within the Montesinh...
2	month	string	month of the year: 'jan' to 'dec'
3	day	string	day of the week: 'mon' to 'sun'
4	FFMC	float	FFMC index from the FWI system: 18.7 to 96.20
5	DMC	float	DMC index from the FWI system: 1.1 to 291.3
6	DC	float	DC index from the FWI system: 7.9 to 860.6
7	ISI	float	ISI index from the FWI system: 0.0 to 56.10
8	temp	float	temperature in Celsius degrees: 2.2 to 33.30
9	RH	float	relative humidity in %: 15.0 to 100
10	wind	float	wind speed in km/h: 0.40 to 9.40
11	rain	float	outside rain in mm/m2 : 0.0 to 6.4
12	area	float	the burned area of the forest (in ha): 0.00 to...

[6]:
```
df_fire.describe()
```

[6]:

	X	Y	month	day	FFMC	DMC
count	517.000000	517.000000	517.000000	517.000000	517.000000	517.000000
mean	4.669246	4.299807	7.475822	4.259188	90.644681	110.872340
std	2.313778	1.229900	2.275990	2.072929	5.520111	64.046482
min	1.000000	2.000000	1.000000	1.000000	18.700000	1.100000
25%	3.000000	4.000000	7.000000	2.000000	90.200000	68.600000
50%	4.000000	4.000000	8.000000	5.000000	91.600000	108.300000
75%	7.000000	5.000000	9.000000	6.000000	92.900000	142.400000
max	9.000000	9.000000	12.000000	7.000000	96.200000	291.300000

	DC	ISI	temp	RH	wind	rain
count	517.000000	517.000000	517.000000	517.000000	517.000000	517.000000
mean	547.940039	9.021663	18.889168	44.288201	4.017602	0.021663
std	248.066192	4.559477	5.806625	16.317469	1.791653	0.295959
min	7.900000	0.000000	2.200000	15.000000	0.400000	0.000000
25%	437.700000	6.500000	15.500000	33.000000	2.700000	0.000000
50%	664.200000	8.400000	19.300000	42.000000	4.000000	0.000000
75%	713.900000	10.800000	22.800000	53.000000	4.900000	0.000000
max	860.600000	56.100000	33.300000	100.000000	9.400000	6.400000

	area
count	517.000000
mean	12.847292
std	63.655818
min	0.000000
25%	0.000000

```
50%       0.520000
75%       6.570000
max    1090.840000
```

Z powyszej tabeli momy zauway, e kolumny rain i area s gownie wypenione zerami.

[33]:
```
print('Liczba rekordów, gdzie w kolumnie area nie ma zera: ' ,df_fire[df_fire.
 ↪area != 0 ].shape[0])
print('Liczba rekordów, gdzie w kolumnie rain nie ma zera: ' ,df_fire[df_fire.
 ↪rain != 0 ].shape[0])
```

```
Liczba rekordów, gdzie w kolumnie area nie ma zera:  270
Liczba rekordów, gdzie w kolumnie rain nie ma zera:  8
```

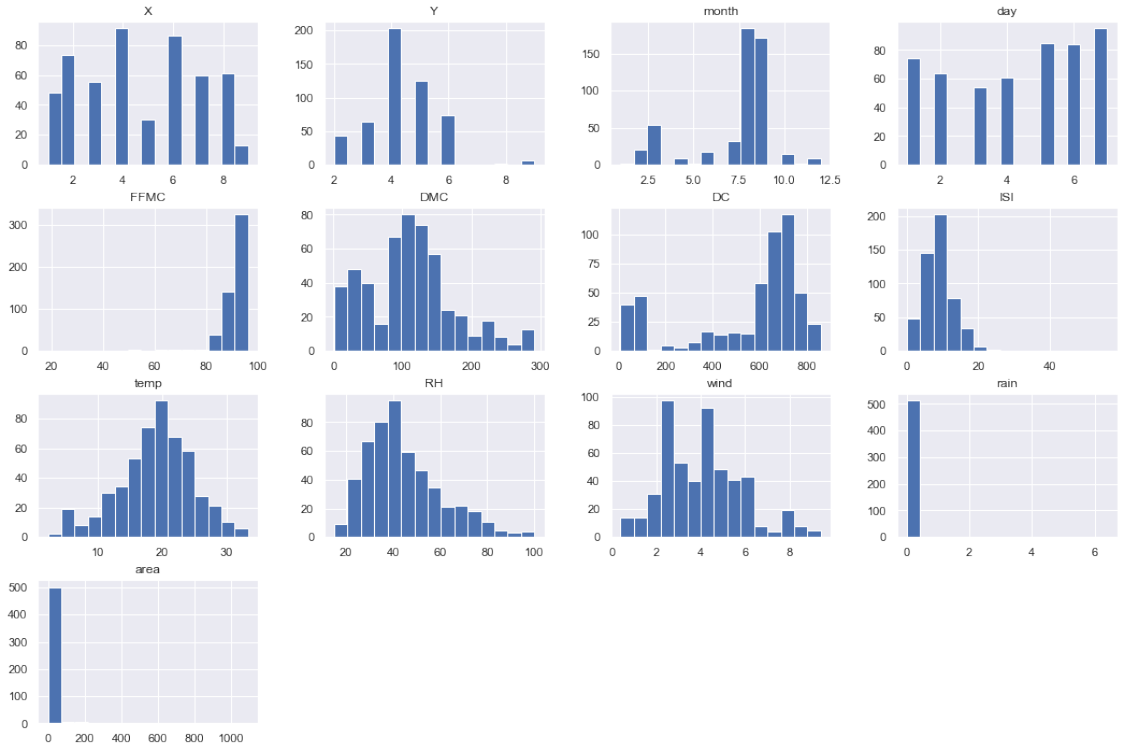
Powysze obliczenia potwierdzaj to. Mniej wicej poowa rekordów zawiera zero w kolumnie area i tylko 8 z 517 zawiera co innego ni zero w kolumnie rain.

[6]:
```
df_fire.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    int64
 3   day     517 non-null    int64
 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(4)
memory usage: 52.6 KB
```

Dziki funkcji info() wiemy, e w naszej ramce danych nie ma wartoci null

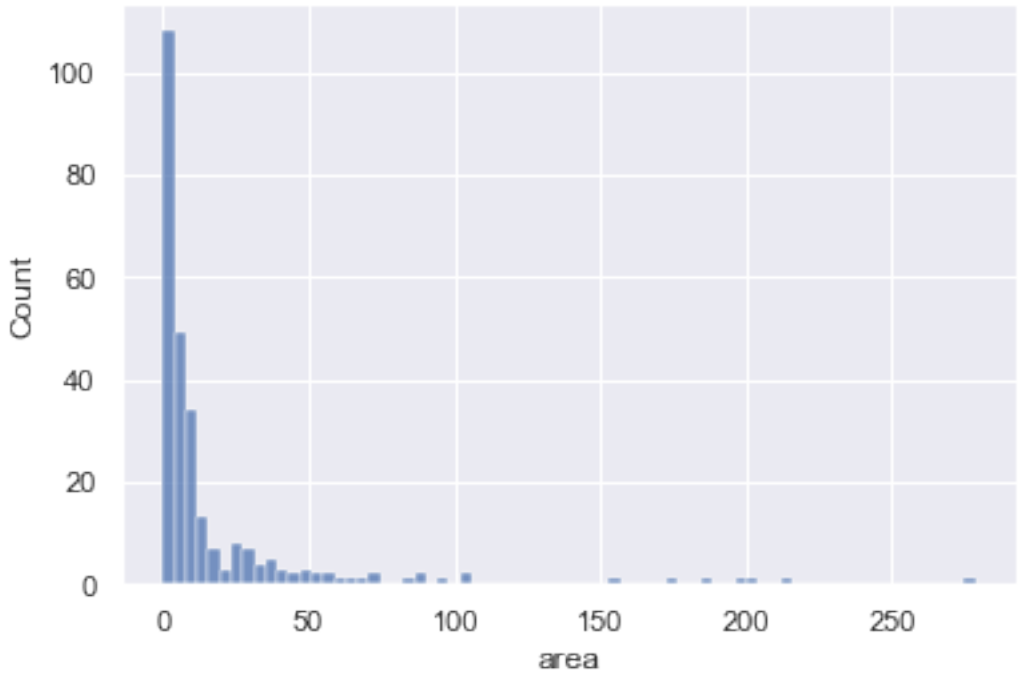
[7]:
```
df_fire.hist(bins = 15, figsize=(18, 12))
plt.show()
```



Powysze histogramy obrazuj rozkady poszczególnych kolumn. Wida w nich wyranie to co rzucio nam si w oczy przy uyciu funkcji describe(), tzn kolumny area i rain s wypenione prawie cakowicie zerami. Dodatkowo moemy zauway, e rozkad temperatury jest mocno zbliony do rozkadu normalnego. Oprócz tego moemy zauway, e cecha dotyczca uczszczania do lasu przez ludzi ma rozkad skony prawostronny.

[8]:
```
sns.histplot(df_fire.area[(df_fire.area < 300) & (df_fire.area != 0)])
```

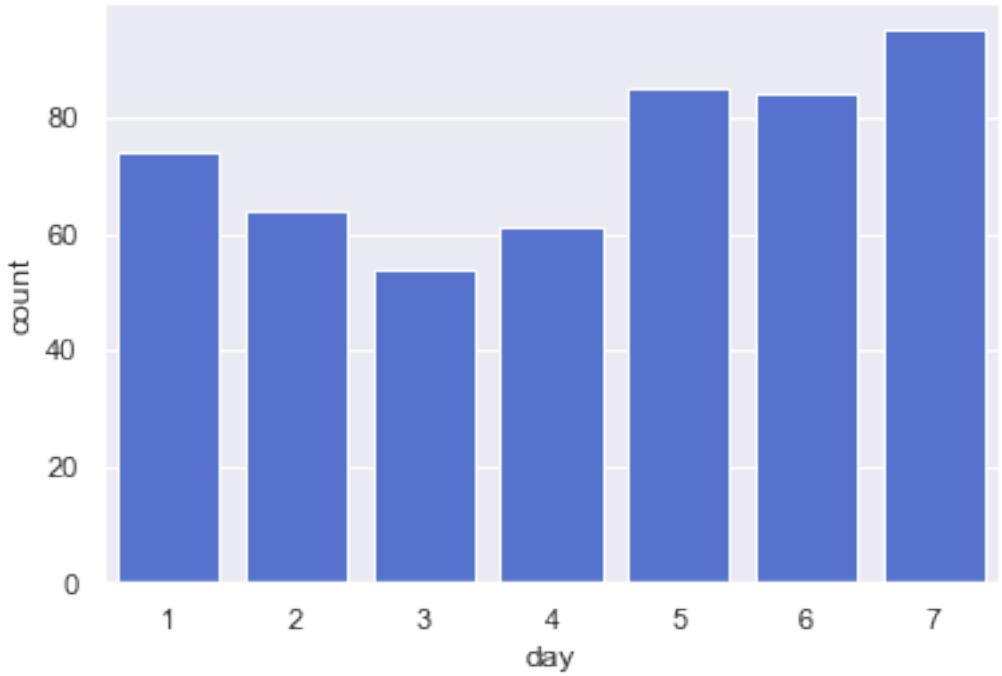
[8]:
```
<AxesSubplot:xlabel='area', ylabel='Count'>
```



Po odrzuceniu skrajnie duych wartoci i wartoci zerowych widzimy, e rozkad obszaru objtego poarem ma ju rozkad podobny do 1/x.

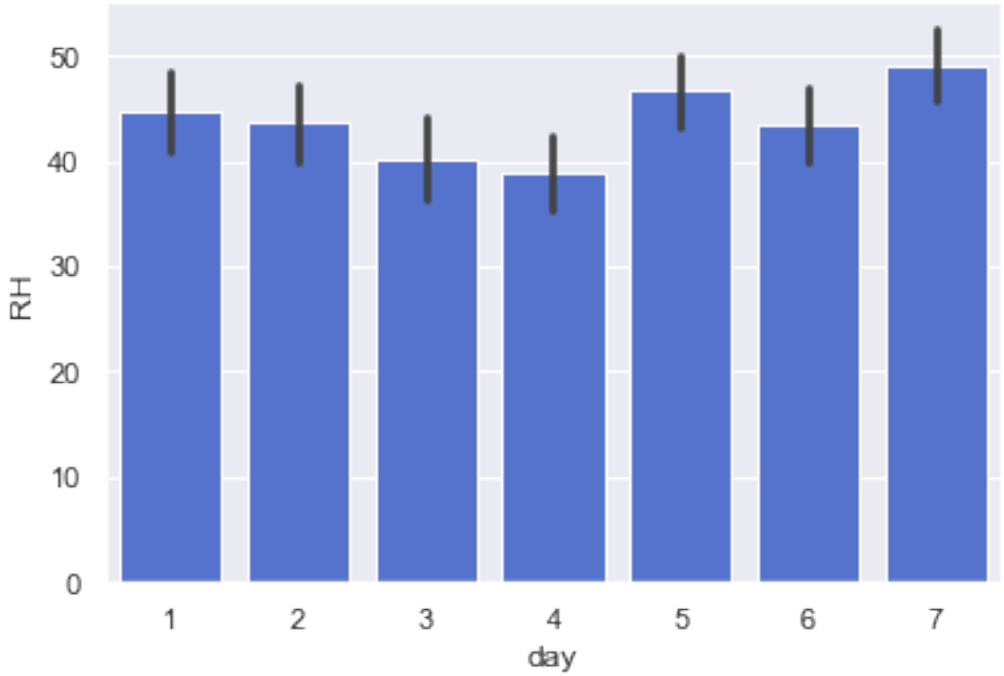
[9]:
```
sns.countplot(data = df_fire, x  = 'day', color="RoyalBlue")
```

[9]:
```
<AxesSubplot:xlabel='day', ylabel='count'>
```



[10]:
```
sns.barplot(data = df_fire, x = 'day', y = 'RH', color="RoyalBlue" )
```

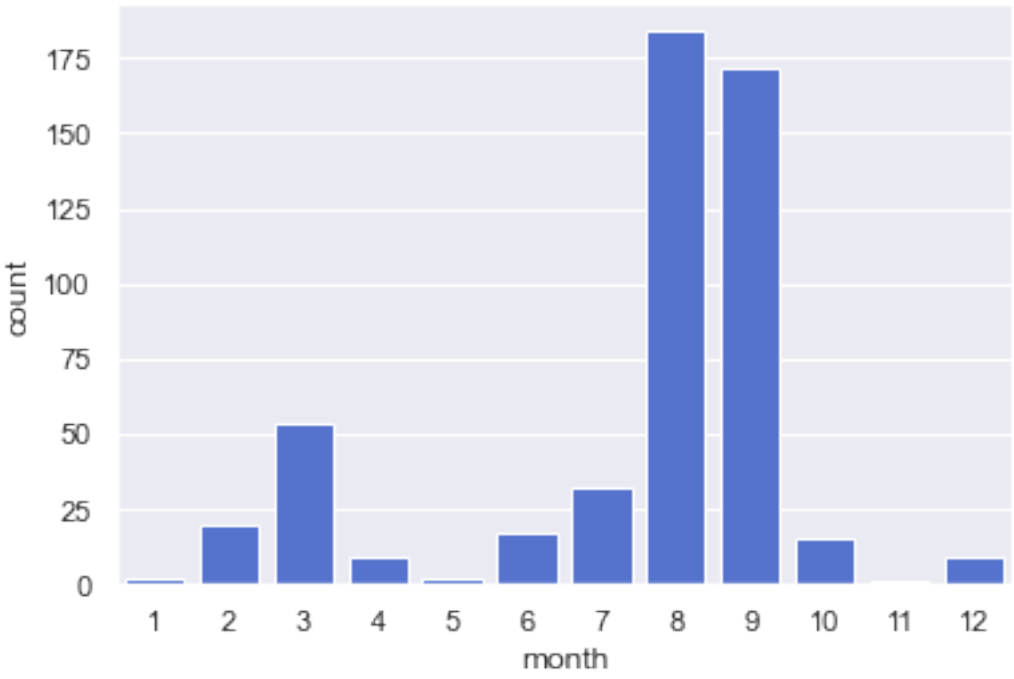
[10]:
```
<AxesSubplot:xlabel='day', ylabel='RH'>
```



Z dwóch powyszych wykresów wida to co byo wida ju wczeniej. Najwicej poarów miao miejsce w pitek sobote i niedziele. Jest to do logiczne zwaszcza, gdy spojrzymy na drugi wykres. Wida z niego e najwicej ludzi przychodzi do lasu wanie w weekendy. To wanie ludzie mog by zatem przyczynom wielu poarów.

[11]:
```
sns.countplot(data = df_fire, x  = 'month', color="RoyalBlue")
```

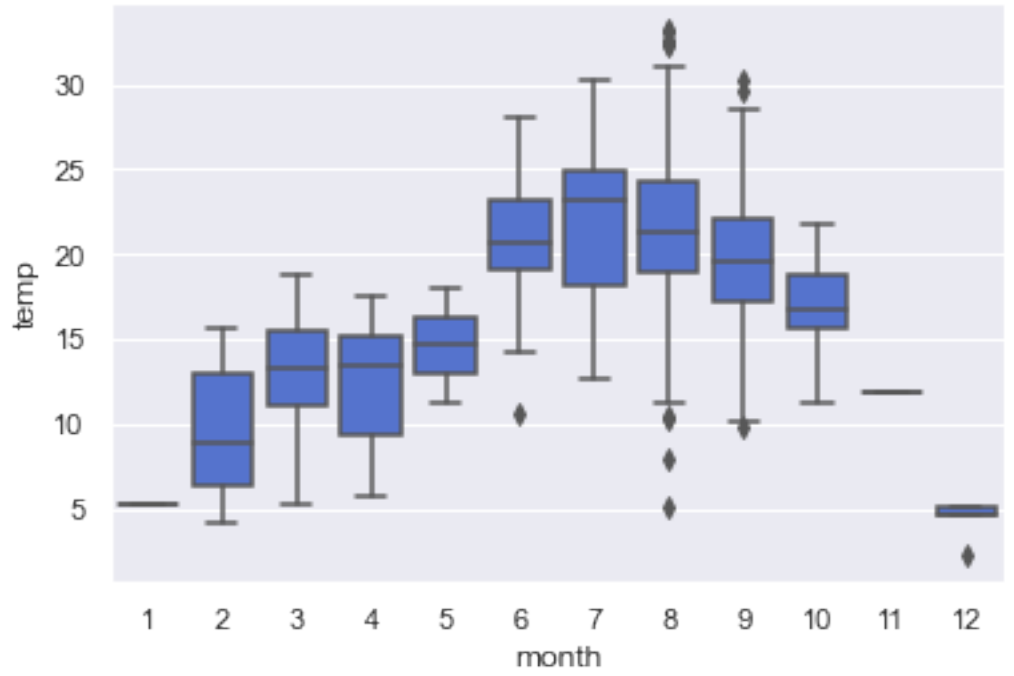
[11]:
```
<AxesSubplot:xlabel='month', ylabel='count'>
```



Ten wykres pokazuje natomiast ilo poarów w rónych miesicach. Szczególnie duo poarów jest w sierpniu i wrzeniu. Zaskakujco duo poarów jest równie w marcu.

[25]:
```
sns.boxplot(x="month", y="temp", data=df_fire, color="RoyalBlue")
```

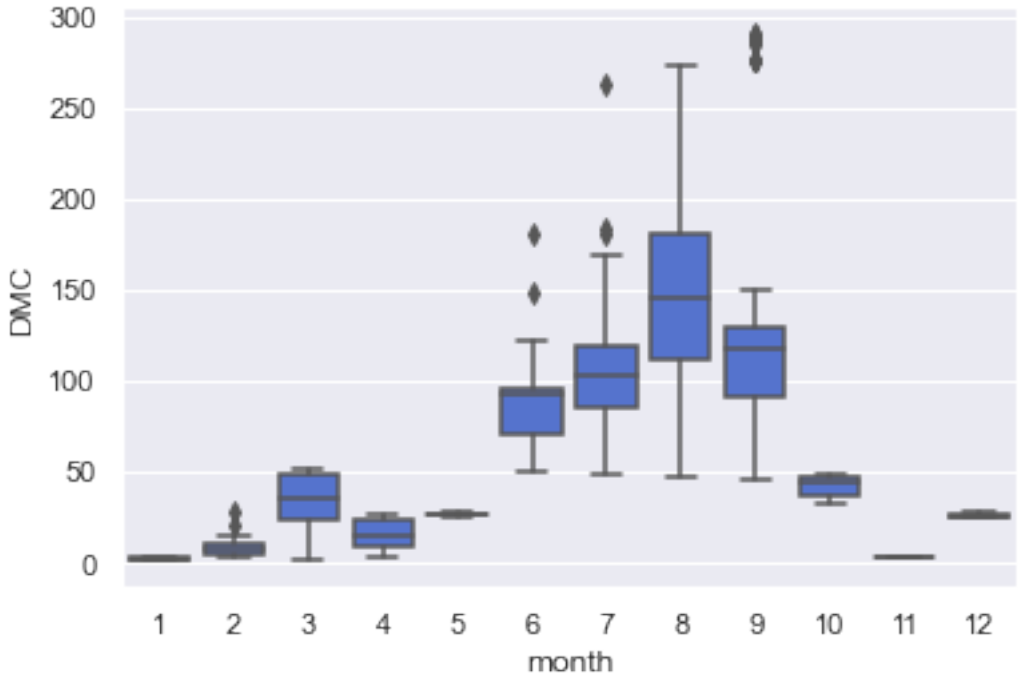
[25]:
```
<AxesSubplot:xlabel='month', ylabel='temp'>
```



Widzimy, e temperatura jest zapewne istotnym czynnikiem w poarach, widzimy jednak, e nie zaley to tylko od temperatury. W czerwcu i libcu temperatury s porównywalne z temperaurami w sierpniu i wrzeniu, a jednak ilo poarów w tych miesicach jest znaczca.

[32]:
```
sns.boxplot(x="month", y="DMC", data=df_fire, color="RoyalBlue")
```

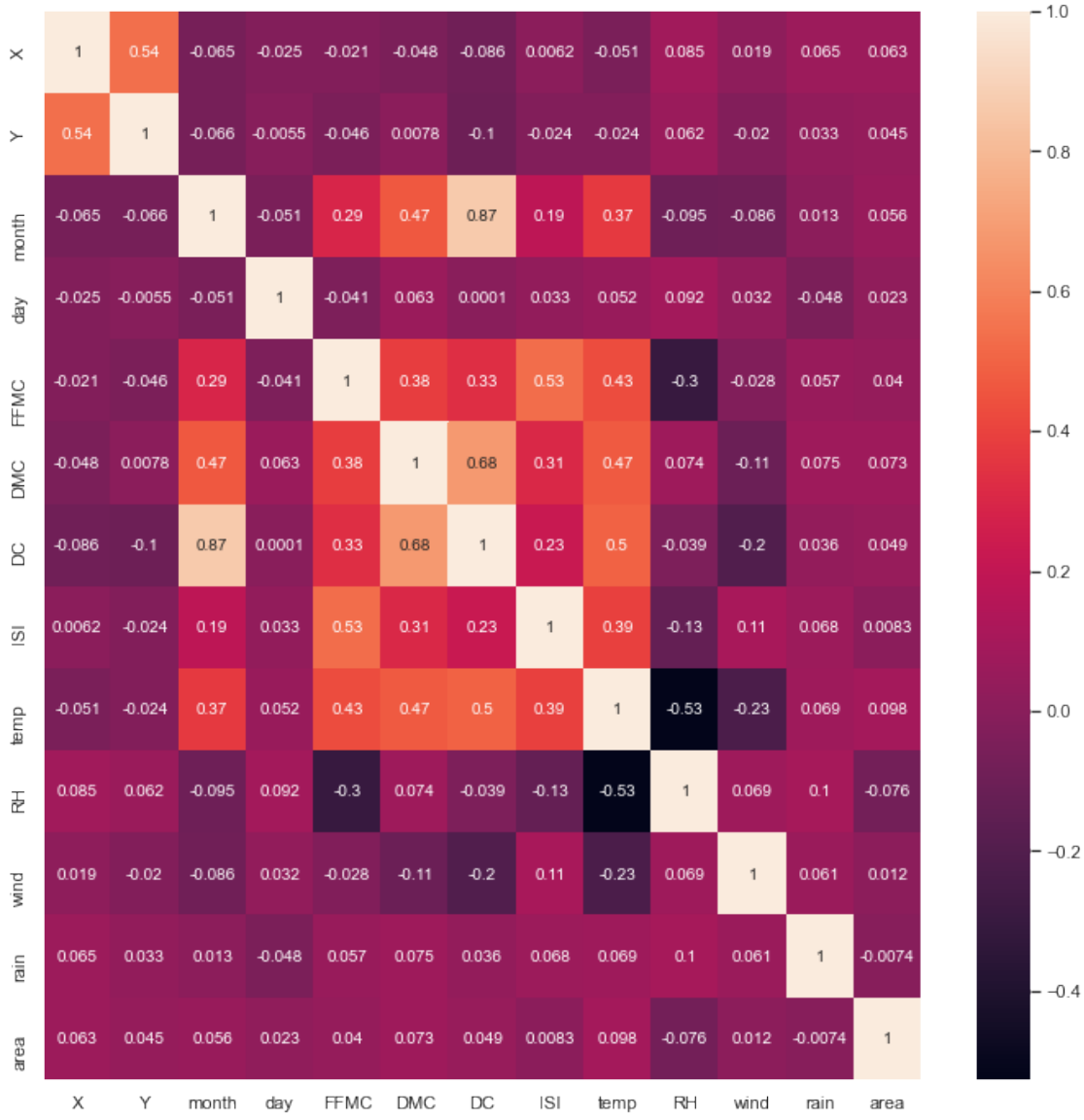
[32]:
```
<AxesSubplot:xlabel='month', ylabel='DMC'>
```



Po obejrzeniu powyszego wykresu moemy wnioskowa, e to odpowiednia temperatura w poczeniu z nisk wilgotnocia sprzyja poarom. Warto ta jest cakiem wysoka w marcu co zgadza si ze wzrotem poarów w tym miesicu.

[12]:
```
correlation = df_fire.corr()
fig, ax = plt.subplots(figsize=(12,12))
sns.heatmap(correlation, annot = True)
```

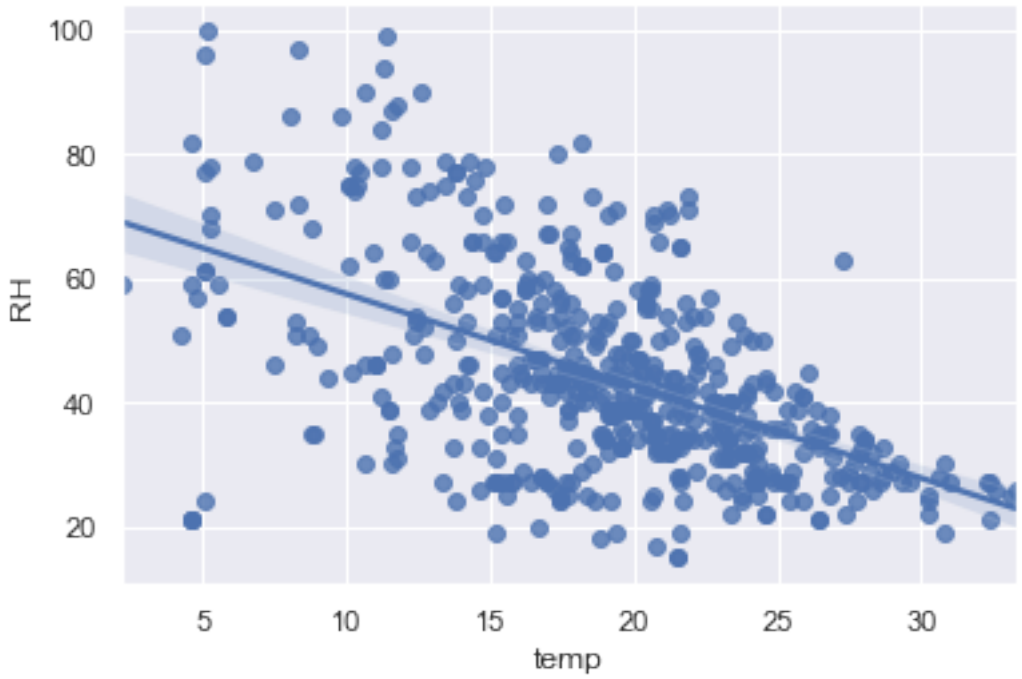
[12]:
```
<AxesSubplot:>
```



Powysza macierz korelacji pokazuje nam które wartoci mog by ze sob skorelowane. Wida z niej na przykad, e temperatura jest skoreloawna ujemnie z iloci ludzi w lasach. Oznacza to, e im wiksza temperatura tym mniej ludzi chodzi do lasu. Jest to do zaskakujce, gdy w Polsce tendencja jest raczej odwrotna. Nie jest zaskoczeniem, e cechy oznaczajce wilgotno s ze sob skorelowane. Na ogó jednak kolumny nie s jednak ze sob skorelowane.

[13]:
```
sns.regplot(data = df_fire, x = 'temp', y = 'RH')
```

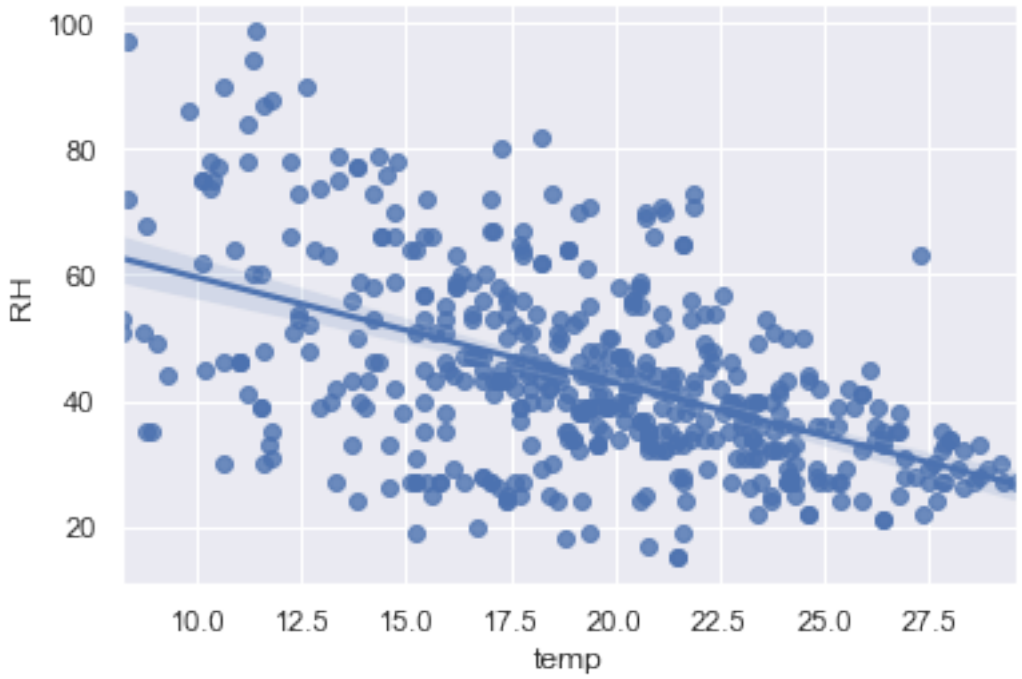
[13]:
```
<AxesSubplot:xlabel='temp', ylabel='RH'>
```



Widzimy z wykresu e wraz ze wzrostem temperatury maleje liczba lidzi w lesie. Spróbujmy jednak pozby si wartoci odstajcych aby zobaczy czy to zmieni tendencje.

[14]:
```
sns.regplot(data = df_fire[(df_fire.temp > 8) & (df_fire.temp <= 30)], x =␣
↪'temp', y = 'RH')
```

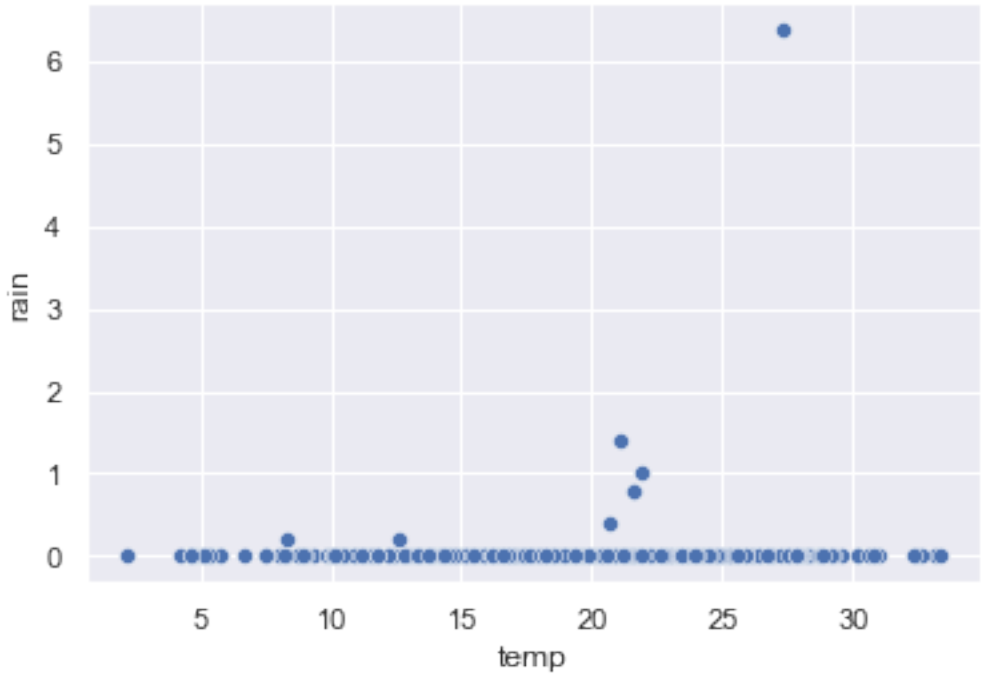
[14]:
```
<AxesSubplot:xlabel='temp', ylabel='RH'>
```



Wypaszcza to nieco krzyw dopasowanie, jednak nieznaczenie. By moe, gdy jest cieplej to wicej pada i std mniejsze zainteresowanie spacerami do lasu.

[15]:
```
sns.scatterplot(data = df_fire, x = 'temp', y = 'rain')
```

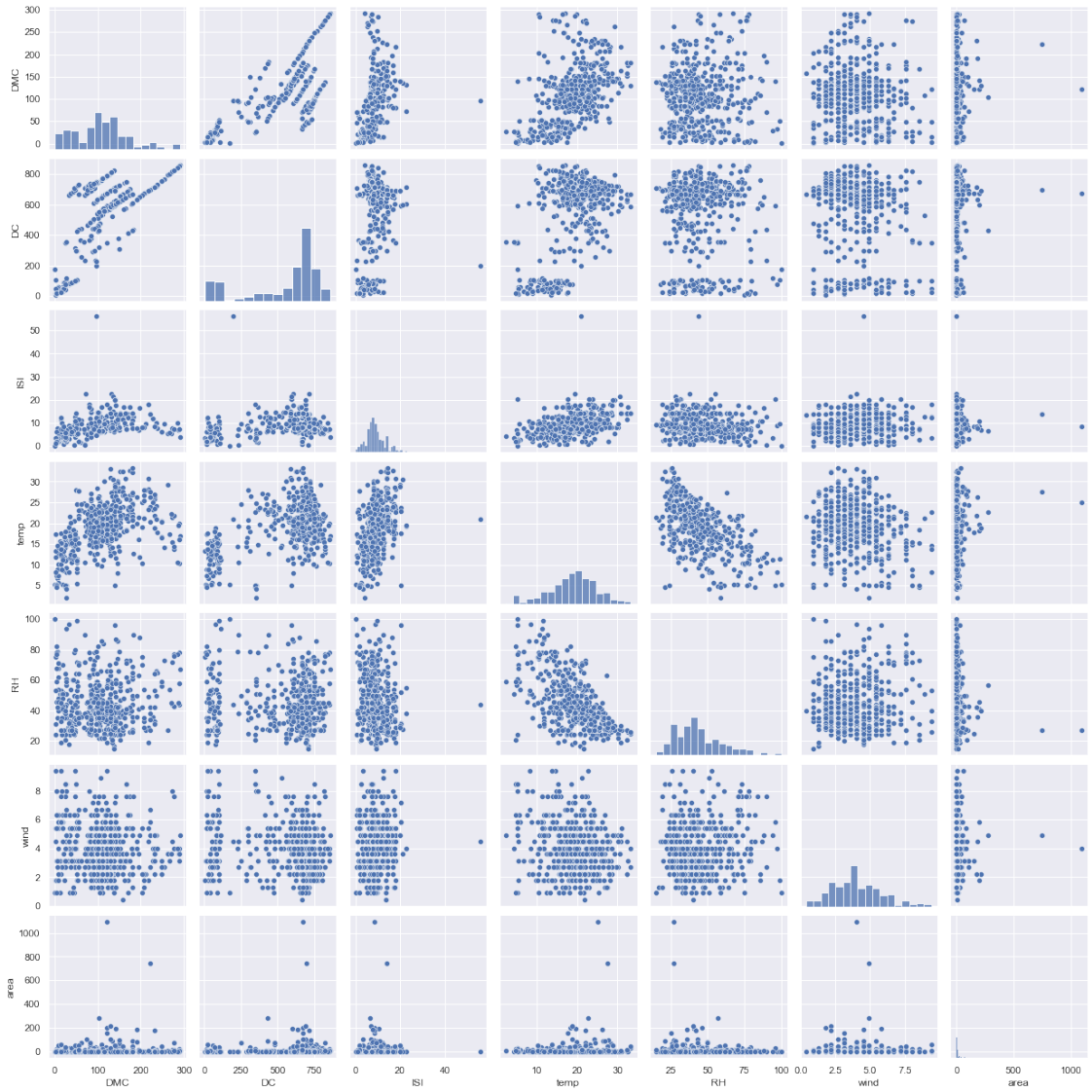
[15]:
```
<AxesSubplot:xlabel='temp', ylabel='rain'>
```



Nie moemy tego stwierdzi, gdy w wikszoci rekordów w kolumnie rain jest zerem. Sprawdmy jak koreluj ze sob inne wasnoci cige

[26]:
```
sns.pairplot(df_fire[['DMC', 'DC','ISI','temp','RH','wind', 'area']])
```

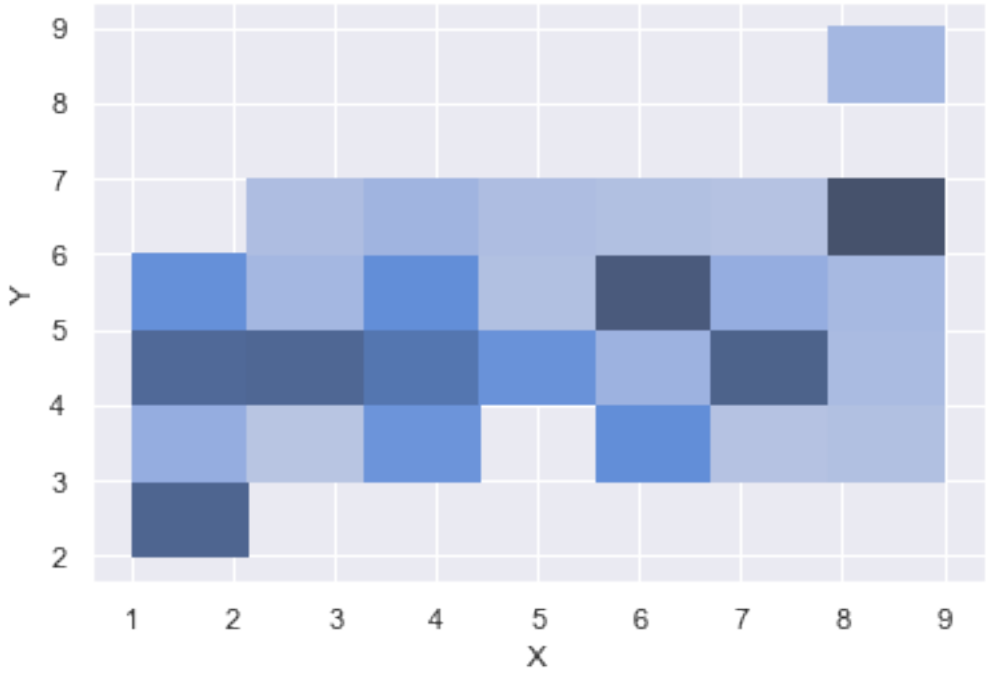
[26]:
```
<seaborn.axisgrid.PairGrid at 0x13a1148d0>
```



Oprucz analizowanej wczeniej zalenoci temperatury i liczby ludzi w lesie, ciekawe zalenoci wida pomidzy DC i DMC. adnie wyglda równie zaleno midzy DMC I DC a ISI. Relacja midzy temperatur a ISI przypomina troch relacje liniow. Na wykresach wyej wida równie zalenoc pomidy DMC a temoeratur, potwierdza to nasze przypuszczenia dotyczce ich zwizku z poarami w danych miesicach. Powysze wykresy punktowe potwierdzaj niejako to co mówia macie korelacji.

[30]:
```
sns.histplot(data = df_fire, x = 'X', y = 'Y', bins = 7)
```

[30]:
```
<AxesSubplot:xlabel='X', ylabel='Y'>
```



Powysza mapa prezentuje nam, w których obszarach lasu najczciej dochodzio do poarów

1.2 Narzdzia do automatycznej eksploracji danych

[18]:
```
profile = ProfileReport(df_fire, title='Pandas Profiling Report')
profile
```

```
Summarize dataset:   0%|          | 0/26 [00:00<?, ?it/s]

Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]

Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]

<IPython.core.display.HTML object>
```

[18]:

Profile report uatwia zdecydowanie eksploracje danych. Ma on jednak troche ogranicze. Za jego pomoc nie mona na przykad bliej przyjrze si poszczególnym rozkadom, nie bylibymy w stanie zobaczy, e rozkad zmiennej area ma posta 1/x. Drugim problemem jest ograniczona wizualizacja zmiennych kategorycznych, nie mona na przykad zobaczy boxplotów, które lepiej odzwierciedlaj rozkady z podziaem na kategori. Myle e pandas_profiling to wietne narzdzie, do rzucenia okiem na dane. Zapewnie nam ono podstawowe statystyki i wykresy. Powinnimy jednak zwróci uwage na to co jest ciekawe lub niepasujce i samodzielnie to pogebi.

[ ]: