# Jan_Smolen_PD1

March 9, 2021

## 1 Praca domowa nr 1

### 1.1 Jan Smoleń, 08-03-2021

Tematem poniższej pracy jest eksploracja zbioru danych dotyczącego pożarów lasów w parku Montesinho w północno-wschodnim rejonie Portugalii.

[52]:
```
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
import seaborn as sns
import pandas_profiling
```

[53]:
```
df = pd.read_csv("forest_fires_dataset.csv")
df.head()
```

[53]:

| | 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 |

[54]:
```
df.describe()
```

[54]:

| | 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 |

### 1.2 Kolumny

**X** - Współrzędne przestrzenne osi X na mapie parku Montesinho: od 1 do 9.

**Y** - Współrzędne przestrzenne osi Y na mapie parku Montesinho: od 2 do 9.

**month** - miesiąc

**day** - dzień tygodnia

**FFMC** - reprezentuje suchość paliwa ze najmniejszej ściółki leśnej w cieniu okapu lasu. Przyjmuje wartości od 0-101.

**DMC** - reprezentuje suchość paliwa rozłożonego materiału organicznego pod ściółką, do głębokośći 10cm. Przyjmuje dowolne dodatnie wartości.

**DC** - reprezentuje suchość dużego materiału organicznego na głębokości większej niż 10 cm. Przyjmuje wartości od 0 do 1000.

**ISI** - Szacuje potencjał rozprzestrzeniania się ognia, wynika z FFMC oraz prędkości wiatru. Przyjmuje dowolne dodatnie wartości.

**temp** - temperatura w stopniach Celsjusza

**RH** - względna wilgotność w procentach

**wind** - prędkość wiatru w km/h.

**rain** - opady deszczu w mm/m2.

**area** - spalony obszar lasu w hektarach.

Użyjemy teraz narzędzia do automatycznej eksploracji danych pandas-profiling żeby poznać trochę bardziej szczegółowe informacje na temat naszych danych.

[55]:
```
df.profile_report()
```

```
HBox(children=(HTML(value='Summarize dataset'), FloatProgress(value=0.0, max=26.
↪0), HTML(value='')))
```

```
HBox(children=(HTML(value='Generate report structure'), FloatProgress(value=0.0,␣
↪max=1.0), HTML(value='')))
```

```
HBox(children=(HTML(value='Render HTML'), FloatProgress(value=0.0, max=1.0),␣
↪HTML(value='')))
```

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

[55]:

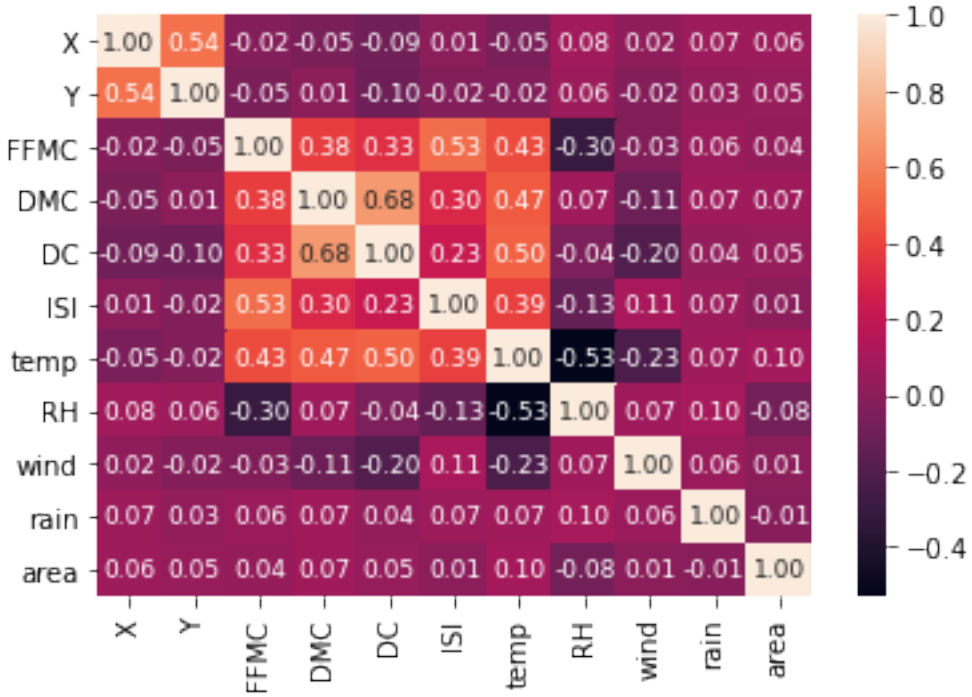
### 1.3 Wnioski/pytania

1. Nie występuje problem braku danych
2. Co oznacza, że spalony obszar wynosi zero? Czy to, że pożar się nie rozprzestrzenił czy może to, że nie pożar nie wystąpił wcale? Jeżeli ta druga opcja, to co to decyduje o wpisaniu danego dnia i obszaru do danych? Może pomiary były wykonywane częściej podczas okresów bardziej podatnych na pożary? Wydaje mi się to nie bez znaczenia w kontekście wykorzystania tych danych do przewidywania zmiennej celu.
3. Można się pozbyć zduplikowanych rzędów, ponieważ odzwierciedlają wyniki tych samych pomiarów.
4. Duża ilość zer w kolumnach area i rain powoduje, że przygotowane przez narządzie wykresy z tymi danymi są mało czytelne.
5. Główne ograniczenia tego narzędzia do automatyzowania eksploracji: dłuższy czas generowania, dużo nieinteresujących nas informacji, mało czytelne wykresy.

Usuniemy zduplikowane rzędy i przygotujemy wygodniejszą wersję heatmapy korelacji

[56]:
```
df_reduced=df.drop_duplicates()
sns.heatmap(df_reduced.corr(), annot=True, annot_kws={'size': 9}, fmt='.2f')
```

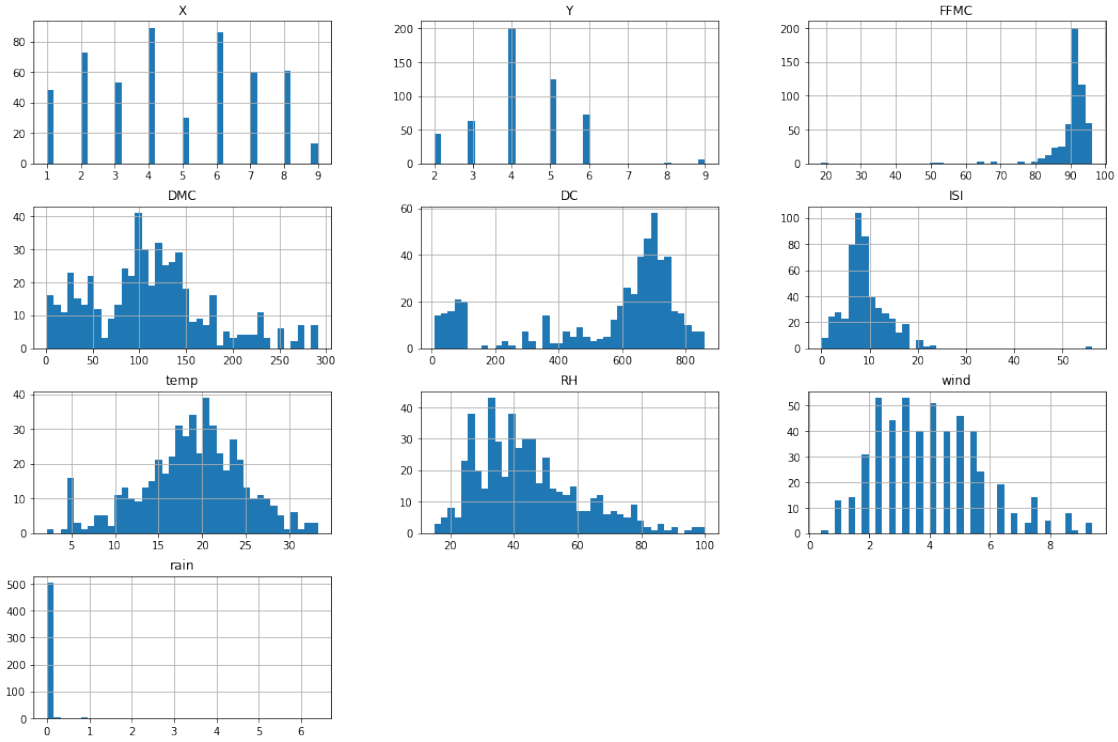
[56]: <AxesSubplot:>



Zgodnie z oczekiwaniami, wskaźniki oznaczające suchosć ściółki na różnych poziomach są dość mocno skorelowane zarówno między sobą, jak i z temperaturą. Po zrobieniu researchu okazuje się także, że wzór na ISI zależy jedynie od wartości FFMC i prędkości wiatru - może należałoby usunąć te kolumnę?

Spójrzmy teraz na rozkłady zmiennych objaśniających.

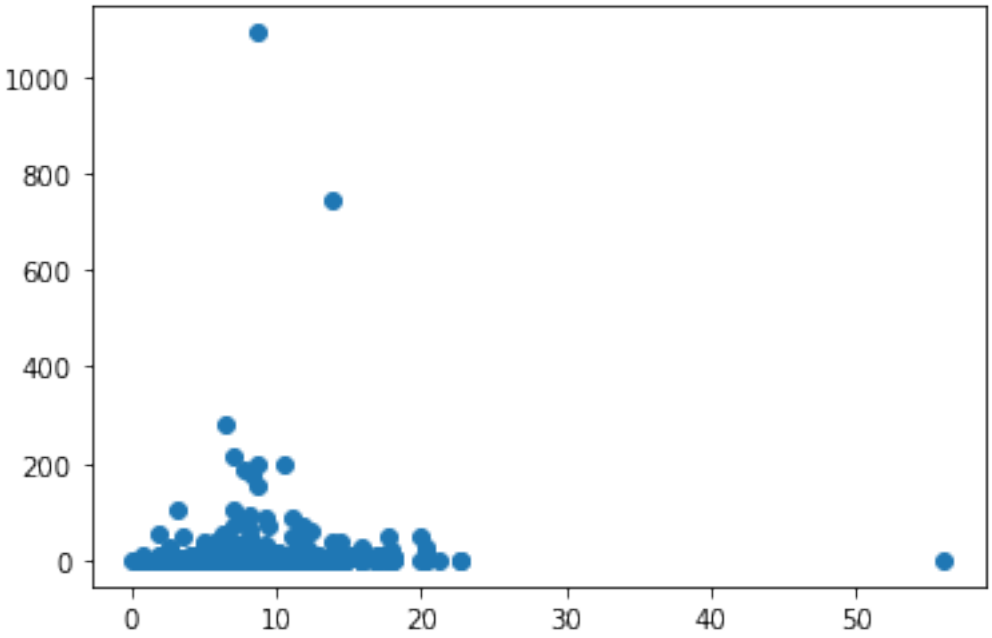
[57]:
```
df_reduced.drop(["area"], axis=1).hist(bins = 40, figsize=(18, 12))
plt.show()
```




Widzimy że zmienne temp, ISI, RH i wind mają rozkłady przypominające rozkłady normalne. W zmiennej rain dominują wartości zerowe. Uwagę zwracają pojedyncze wartość ISI i FFMC, która zdecydowanie odstają od reszty. Spójrzmy na wykres punktowy tych dwóch zmiennych i zmiennej celu area.

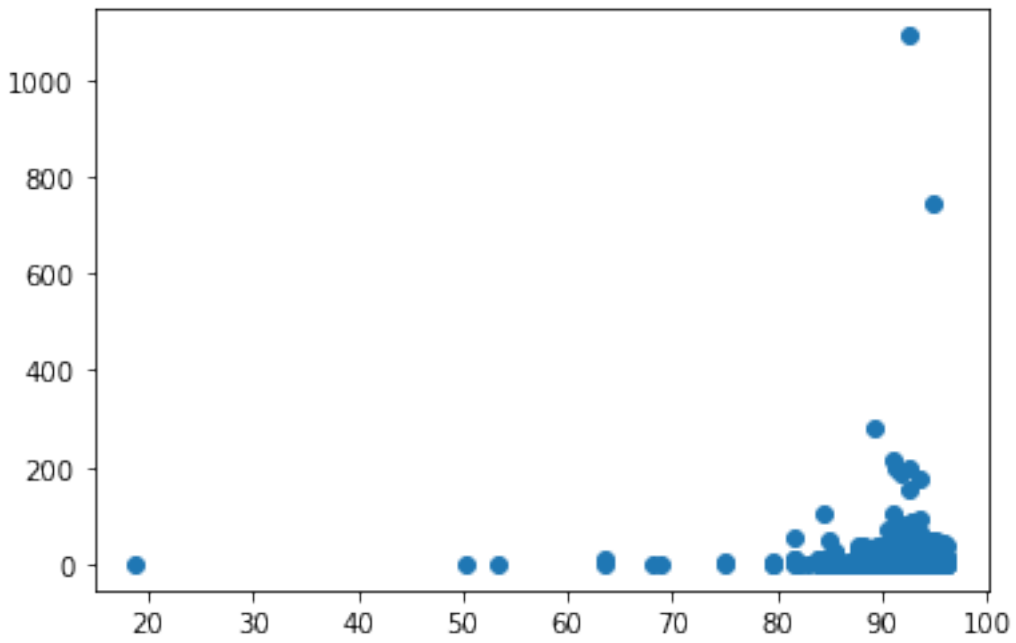
[58]:
```
plt.scatter(df_reduced["ISI"], df_reduced["area"])
```

[58]:
```
<matplotlib.collections.PathCollection at 0x1f094dabbb0>
```




[59]:
```
plt.scatter(df_reduced["FFMC"], df_reduced["area"])
```

[59]:
```
<matplotlib.collections.PathCollection at 0x1f0948fb2b0>
```




Ponieważ te dwa pomiary znacząco odstają od reszty i nie wpisują się szczególnie w jakiś widoczny

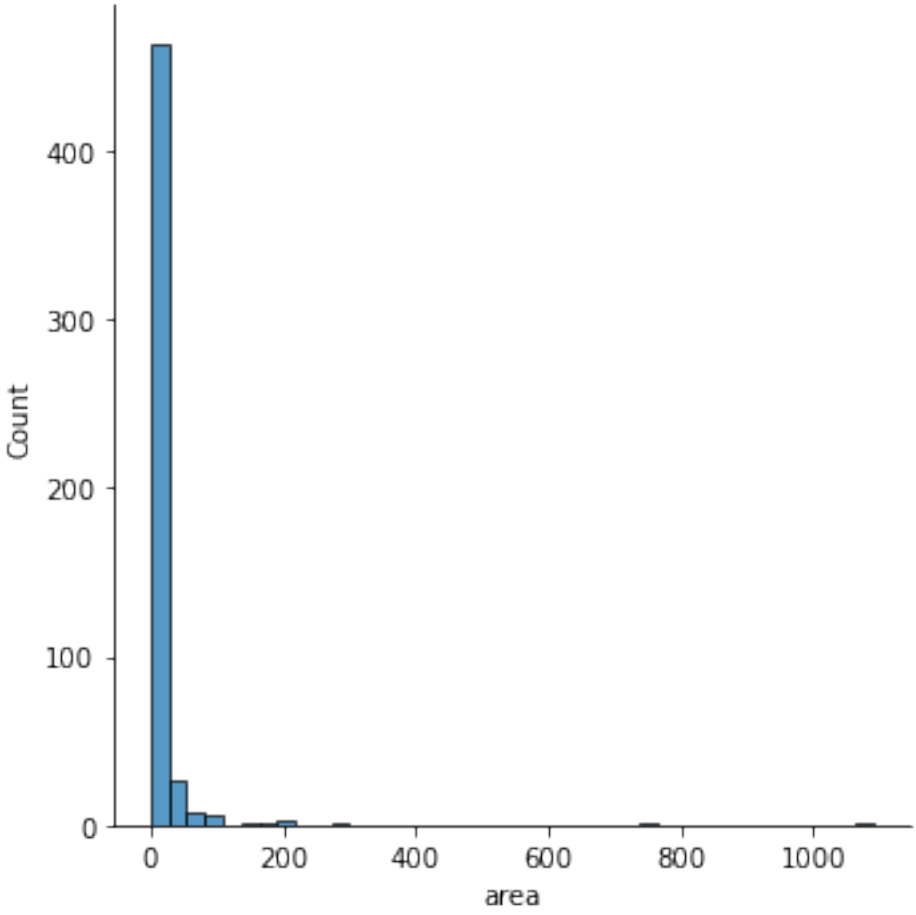
trend, usuniemy je z naszych danych. Możemy też się upewnić, że nie należą do tego samego rzędu - tak nie jest.

[60]:
```
df_reduced=df_reduced.drop([df_reduced["ISI"].idxmax()])
df_reduced=df_reduced.drop([df_reduced["FFMC"].idxmin()])
```

Przyjrzymy sie jeszcze trochę dokładniej rozkładowi zmiennej celu area.

[61]:
```
sns.displot(df_reduced["area"], bins=40)
```

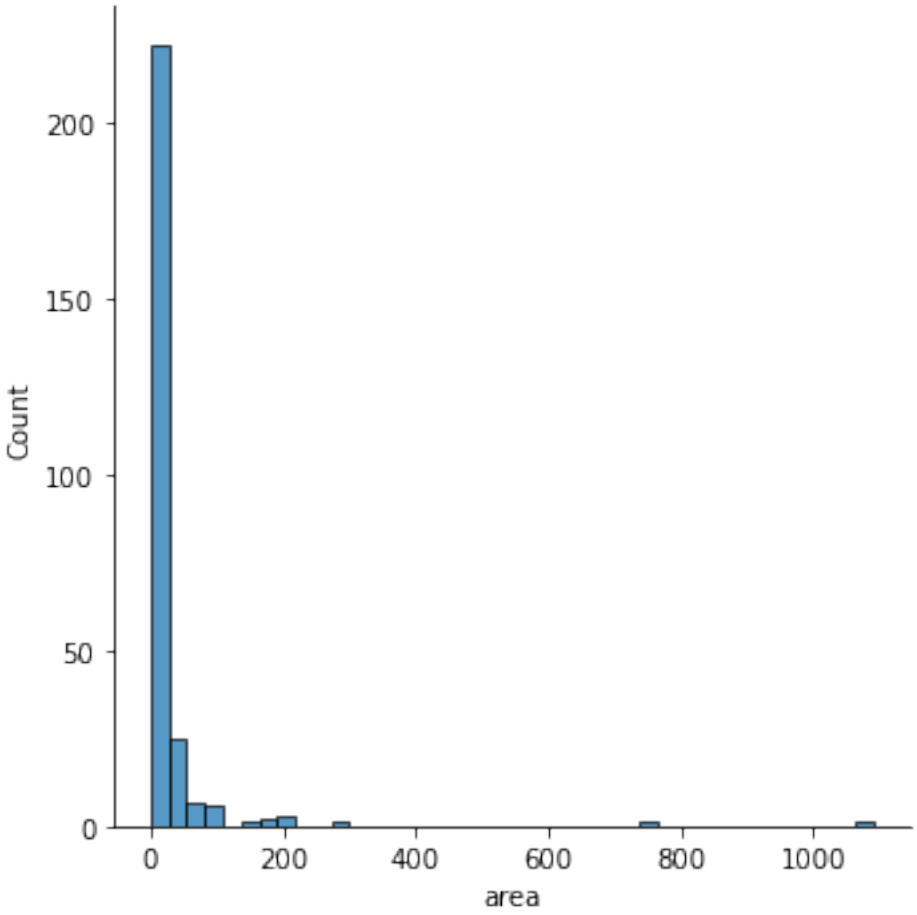
[61]:
```
<seaborn.axisgrid.FacetGrid at 0x1f0948c4af0>
```




Spójrzmy jeszcze, jak rozkładają się wielkości spalonego obszaru w sytuacji, w której wiemy już, że pożar w ogóle wystąpił/rozprzestrzenił się.

[62]:
```
df_drop0=df_reduced[df_reduced['area']!=0]
sns.displot(df_drop0["area"], bins=40)
```

[62]:
```
<seaborn.axisgrid.FacetGrid at 0x1f0949c0100>
```

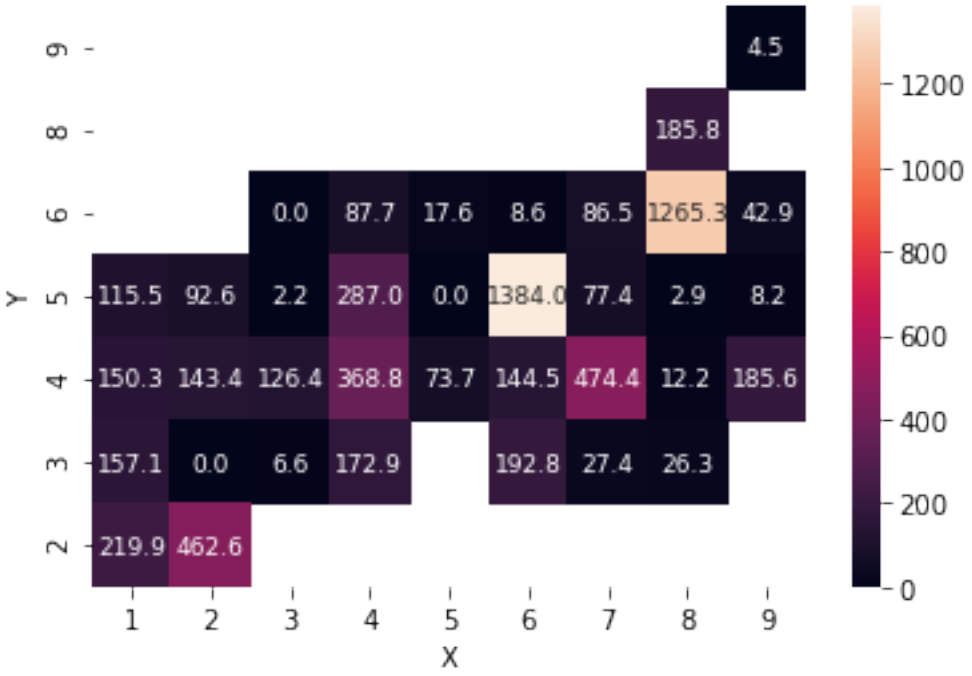



Zatem widzimy, że w przeważającej liczbie pomiarów pożar albo nie występuje wcale, albo występuje na małym obszarze. Rozkład jest bardzo przekrzywiony w lewo, być może na dalszym etapie do jego modelowania przydatna byłaby transformacja logarytmiczna. Dwie wartości bardzo odstają od reszty, ale moim zdaniem świadczą raczej o pożarach, które wymknęły się spod kontroli niż o błędach w pomiarze, więc je pozostawmy.

Biorąc pod uwagę, że wartości X i Y oznaczają położenie na mapie parku i raczej należałoby je rozpatrywać razem, spróbujemy odtworzyć te mapę i pokazać na niej sumę spalonego obszaru w danych sektorach. Dzięki temu też zobaczymy, jakie obszary w ogóle występują w parku.

[63]:
```
mapa=df_reduced.groupby(["X", "Y"]).agg({'area':'sum'}).reset_index()
mapa = mapa.pivot(index='Y', columns='X', values='area')
mapa=mapa.iloc[::-1]
sns.heatmap(mapa, annot=True, annot_kws={'size': 9}, fmt='.1f')
```

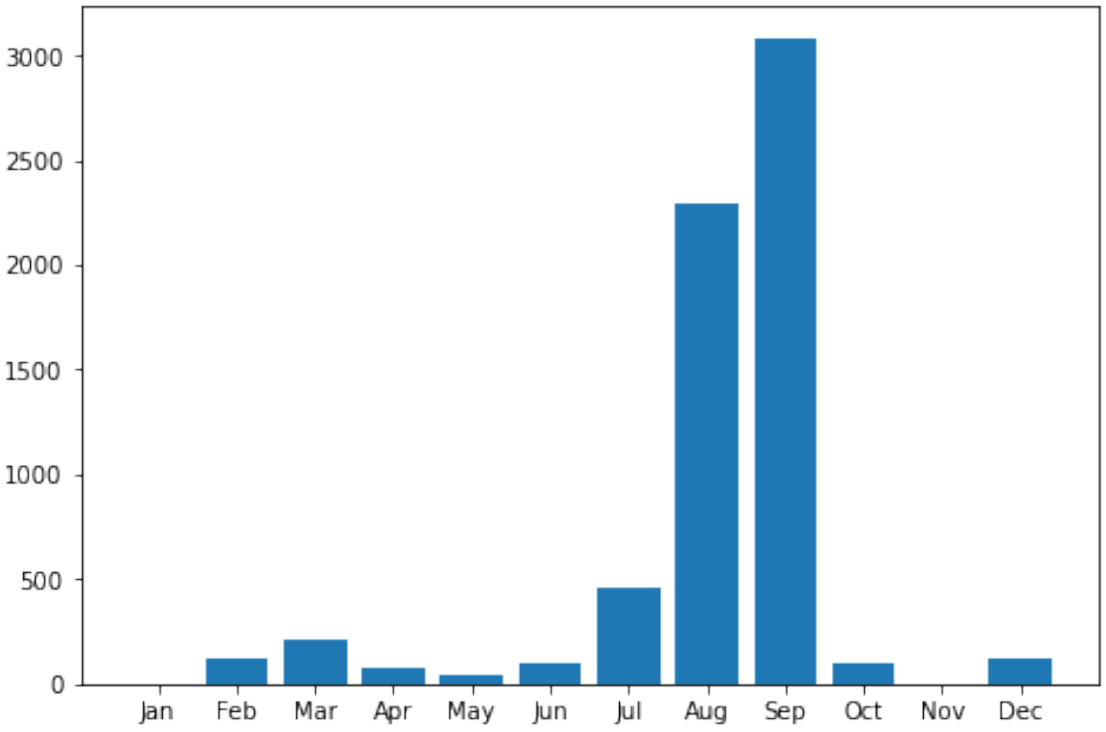
[63]:
```
<AxesSubplot:xlabel='X', ylabel='Y'>
```




Teraz zbadamy ilość spalonego obszaru w poszczególnych miesiącach i dniach tygodnia. Ze względu na bardzo liczne zera występujące w kolumnie area oraz nieznany nam sposób wyboru dni i obszarów zawartych w danych, posłużymy się wykresami słupkowymi pokazyjącymi łączną sumę spalonych obszarów.

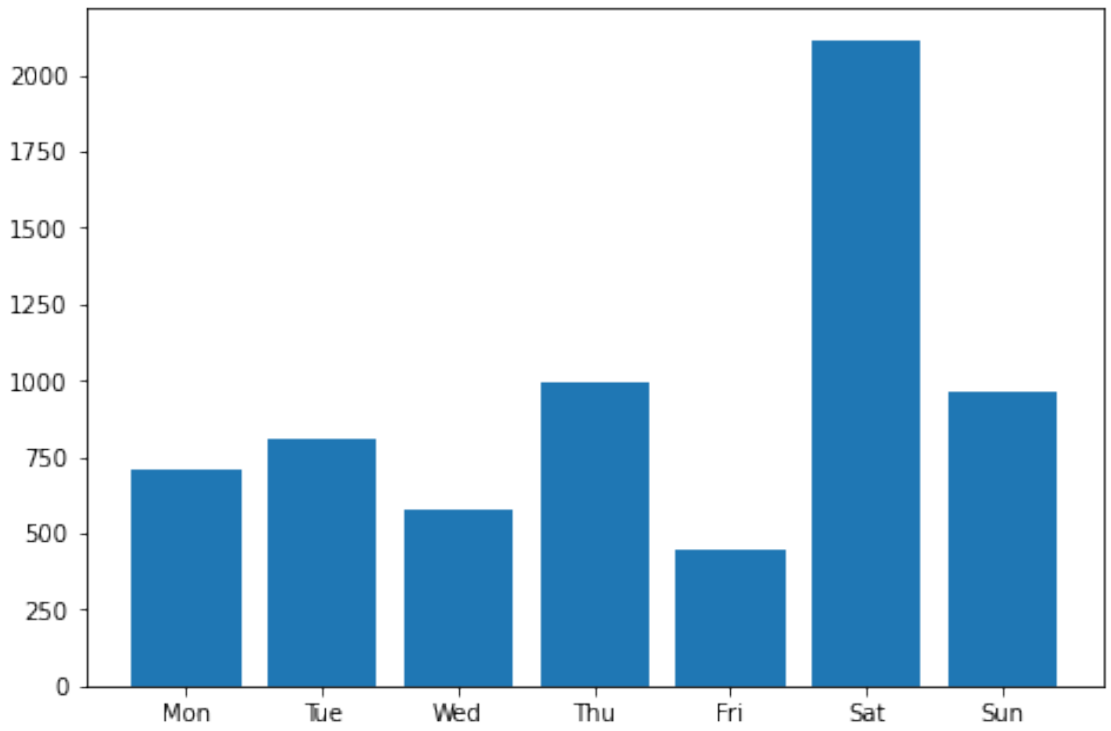
[64]:
```
tmp_month=df_reduced.groupby(["month"]).agg({'area':'sum'}).reset_index()
months = ["Jan", "Feb", "Mar", "Apr", "May", "Jun",
          "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"]
tmp_month["month"]=tmp_month["month"].str.capitalize()
tmp_month['month'] = pd.Categorical(tmp_month['month'], categories=months,␣
↪ordered=True)
tmp_month = tmp_month.sort_values(by="month")
fig = plt.figure()
ax = fig.add_axes([0,0,1,1])
ax.bar(tmp_month["month"], tmp_month["area"])
```

[64]:
```
<BarContainer object of 12 artists>
```




[65]:
```
tmp_day=df_reduced.groupby(["day"]).agg({'area':'sum'}).reset_index()
days=["Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun"]
tmp_day["day"]=tmp_day["day"].str.capitalize()
tmp_day['day'] = pd.Categorical(tmp_day['day'], categories=days, ordered=True)
tmp_day = tmp_day.sort_values(by="day")
fig = plt.figure()
ax = fig.add_axes([0,0,1,1])
ax.bar(tmp_day["day"], tmp_day["area"])
```

[65]:
```
<BarContainer object of 7 artists>
```




Jak widzimy, znacznie więcej powierzchni uległo pożarom w sierpniu i wrześniu niż w innych miesiącach. Co ciekawe, sobota wyróżnia się od pozostałych dni tygonia - być może zwiększona liczba spacerowiczów ma wpływ na więcej pożarów.