Akademia Górniczo-Hutnicza

Wydział Geologii, Geofizyki, i Ochrony Środowiska
Kierunek Inżynieria i Analiza Danych

Uczenie Maszynowe

Raport

Projekt dotyczący Eksploracyjnej Analizy Danych i Inżynierii Cech

---

Izabela Karczewska

Kraków, Maj 2024

1 Biblioteka Seaborn

Seaborn jest biblioteką w języku Python zbudowana na pakiecie Matplotlib. Umożliwia tworzenie bardzo jakościowych wizualizacji danych, oraz modyfikowanie ich w przystępny sposób. Seaborn oferuje wiele funkcji do tworzenia różnych rodzajów wykresów, a następnie do modyfikacji wielu parametrów.

Biblioteka ta staje się standardem w obszarze analizy danych - pozwala wygenerować atrakcyjne i efektowne wizualizacje w przystępny sposób. Wiele funkcji z pakietu Seaborn dobrze współpracuje z danymi w formatach z pakietu Pandas.

2 Strona dane.gov.pl

Strona dane.gov.pl jest serwisem rządowym udostępniającym publicznie dane z różnych obszarów. Jak możemy przeczytać na stronie głównej "Korzystaj z danych bezpłatnie, również do celów komercyjnych".

Zbiory danych podzielone są na kilkanaście kategorii, takich jak:

- Energia
- Nauka i technologia
- Edukacja, kultura i sport
- Gospodarka i Finanse itp.

Dane są udostępniane jako pliki lub jako podgląd gotowych tabeli. Portal umożliwia filtrację wyników. Można zobaczyć, że do wyboru są dane w różnych formatach, z czego największy procent stanowi format CSV.Każdy zbiór jest dodatkowo oceniony przy użyciu gwiazdek od 1 do 5, mówi to o poziomie przygotowania danych do dalszego przetwarzania. Część zbiorów można zobaczyć na stronie w przyjaznej dla oczu formie tabeli, bez uprzedniego pobierania ich.

Zasoby są dostępne poprzez API, plik źródłowy lub z poziomu strony dostawcy danych - tę opcję również wybieramy podczas filtracji. Dane możemy pobierać dzięki różnym API - przykładem będzie API Zabytek, kolejne zbiory są udostępniane przez Ministerstwo Finansów, czy Główny Urząd Statystyczny,a także takie API jak radon.nauka.gov.pl.

3 Wybór danych

Do realizacji projektu wybrane zostały dane udostępnione przez Komendę Główną Państwowej Straży Pożarnej. Mimo sporej ilości zbiorów, wiele z nich było ubogich, na przykład na dwie kolumny, lub w formacie trudnym do przetworzenia. Z kolei ten zbiór zawiera zestawienie informacji ze zdarzeń z 2023 roku - zdarzeń, w których interweniowały zastępy Państwowej Straży Pożarnej. Dane określone są jako te o wysokiej wartości i zawierają ocenę 3/5 w kontekście poziomu otwartości. Zestaw jest dostępny pod tym adresem.

4 Pierwszy etap pipline'u ML

4.1 Pobranie danych

Dane zostały pobrane ze strony ręcznie jako plik z rozszerzeniem .csv. W tym przypadku nie był dostępny widok tabelaryczny na stronie dane.gov.pl.

Rysunek 1: Fragment wczytanych danych

	ID_MELDUNEK	F_BEZ_JOP	OPERATION_TYPE	RODZAJ	WLK	F_POZ_LAS_1	F_POZ_LAS_2	F_POZ_LAS_3	F_POZ_LAS_4	F_MZ_RODZ_1	...	F_DOST_1	F_DOST_2
0	0117001-19/2023	N	Ratownicze	MZ	L	N	N	N	N	N	...	N	N
1	0117001-21/2023	N	Ratownicze	MZ	L	N	N	N	N	N	...	N	N
2	0117001-68/2023	N	Ratownicze	MZ	L	N	N	N	N	N	...	N	N
3	0117001-178/2023	N	Ratownicze	MZ	L	N	N	N	N	N	...	N	N
4	0117001-281/2023	N	Ratownicze	MZ	L	N	N	N	N	N	...	N	N

5 rows × 305 columns

Na podglądzie pierwszych pięciu rekordów można zobaczyć ogólną strukturę oraz to, że dostępnych jest aż 305 kolumn, dlatego potrzebne było przejrzenie drugiego, dodatkowo udostępnionego pliku, czyli opisu kolumn. Bazowe nazwy cech są nazwami skrótowymi i często ciężko wywnioskować co tak naprawdę oznaczają.

4.2 Do czego można użyć wybranych danych?

Dane dostarczają wiele interesujących informacji o zaistniałych zdarzeniach zagrożenia - między innymi o przyczynach tych wydarzeń, przebiegu oraz skutkach.

Zastanówmy się nad możliwościami wykorzystania wybranych danych w ujęciu uczenia maszynowego. W kontekście uczenia nienadzorowanego można wykonać klasteryzację zdarzeń na podstawie podobieństwa cech. Pozwoliłoby to na identyfikację przyszłych zdarzeń na podstawie udostępnionych parametrów, co może pomóc w optymalizacji zarządznia procedurami. Wykonana może być także analiza skupisk w kontekście geograficznym w poszukiwaniu obszarów o wyższym poziomie zagrożenia, co można wykorzystać do lepszego ulokowania zasobów i jednostek straży.

Przechodząc do metod uczenia nadzorowanego, dane mogą posłużyć do wykonania prognozowania czasu poszczególnej akcji ratowniczej. Metody regrsyjne dają również możliwość estymowania liczby osób poszkodowanych, co może pomóc w poprawie działania służb ratowniczych i zjawienia się odpowiednich jednostek.

4.3 Inżynieria cech - Feature Engineering

Jako pierwsze należy dokonać dokładnego przejrzenia danych, szczególnie przy takim rozmiarze. Wybrany zbiór zawiera ponad 400 000 wierszy oraz 305 kolumn. Wybrano 31 z nich, z czego większość to zmienne kategoryczne. Przykładowe z nich opisują mijesce zdarzenia (województwo i gmina), wielkość zdarzenia określoną słownie, jak również powierzchniowo, czy liczbę osób biorących udział w akcji z poszczególnych organów, takich jak OSP czy ZSP. Tak wysoka liczba kolumn wynika także z faktu, że dostawca zbioru w przypadku części zmiennych zapewnił ich zakodowanie, ale dla lepszego wglądu do danych wybrane zostały obie wersje - opisowa i zakodowana.

Rysunek 2: Dane po wybraniu kolumn

```
  RODZAJ WLK   WOJEWODZTWO     GMINA  \
0     MZ   L  dolnośląskie    Milicz
1     MZ   L  dolnośląskie    Milicz
2     MZ   L  dolnośląskie    Milicz
3     MZ   L  dolnośląskie  Krośnice
4     MZ   L  dolnośląskie    Milicz

                                  OBIEKT_OPIS_1  OBIEKT_KOD_1  \
0                                 Wielorodzinne         209.0
1                                 Wielorodzinne         209.0
2  Płyty manewrowe i pasy lotnisk, szlaki kolejow...         816.0
3  Płyty manewrowe i pasy lotnisk, szlaki kolejow...         816.0
4  Drogowe - samochody ciężarowe, maszyny drogowe...         503.0

           DATA_ZGL  POMOC_ALL  \
0  17.01.2023 19:32          0
1  21.01.2023 11:22          1
2  10.02.2023 16:53          0
3  13.04.2023 07:24          0
4  24.06.2023 10:18          0

                                PRZYCZYNA_OPIS  PRZYCZYNA_KOD  ...  \
0      Nieprawidłowa eksploatacja urządzeń gazowych            6.0  ...
1  Wady elektrycznych urządzeń ogrzewczych, w szc...            3.0  ...
2                          Wady środków transportu           16.0  ...
3                                  Inne przyczyny            38.0  ...
4                          Wady środków transportu           16.0  ...

  WOP_LUDZ ZSP_LUDZ ZSR_LUDZ  WYP_INNI_S  WYP_INNI_R  WYP_DZ_S  WYP_DZ_R  \
0        0        0        0           0           0         0         0
1        0        0        0           0           1         0         1
2        0        0        0           0           0         0         0
3        0        0        0           0           0         0         0
4        0        0        0           0           0         0         0

   WLK_POW  WLK_KUB  ZL
0      200        0   0
1       50        0   0
2     3000        0   0
3      100        0   0
4      200        0   0

[5 rows x 27 columns]
```

Dane zawierały trzy zduplikowane wiersze, a więc zostały one usunięte. Sprawdzone zostało ile jest wartości brakujących dla poszczególnych kolumn. Ze względu na czytelność wydruku postanowiono wyświetlić tylko te zmienne dla których ich liczba jest większa od 0.

Rysunek 3: Wartości brakujące (NA)

	Brakujące wartości	Procent
OBIEKT_OPIS_1	46863	9.64%
OBIEKT_KOD_1	46863	9.64%
PRZYCZYNA_OPIS	46866	9.64%
PRZYCZYNA_KOD	46866	9.64%
DATA_ZGL	4	0.00%
DATA_ZAU	45332	9.33%
DATA_POW	45156	9.29%
WLK_POW	63	0.01%
WLK_KUB	2351	0.48%

Rozwiązanie problemu brakujących danych rozpoczęto od kolumny `DATA_ZGL`, w której braków

jest na tyle mało, że całe rekordy zostaną usunięte. Okazuje się również, że dzięki temu zmniejszyła się liczba pozostałym wartości NA, czyli usunięte rekordy były mocno niekompletne. Podobnie postąpiono z kolumnami opisującymi powierzchnię i kubaturę, gdyż procentowy udział wartości brakujących w całości danych wynosił mniej niż 0.5%.

Można dostrzec, że dwie pary kolumn OBIEKT_OPIS_1 i OBIEKT_KOD_1 mają po tyle samo wartości brakujących, co wskazuje, że braki występują w tych samych przypadkach. Taka sama sytuacja występuje w zmiennych opisujących przyczynę wypadku, czyli PRZYCZYNA_KOD i PRZYCZYNA_OPIS.

Przechodząc dalej, w przypadku informacji o przyczynie pozstanowiono zamienić wartości NA na wartość najczęściej występującą dla danego województwa. Dla wszystkich województw najczęstszą przyczyną okazały się "Inne przyczyny"i odpowiadający temu kod 38, więc w taki sposób te wartości zostały zmienione. Ta operacja spowodowała znaczący wzrost liczebności dla tej przyczyny, natomiast wydaje się to być najlepszą możliwą opcją.

Z opisem obiektu postąpiono w następujący sposób - sprawdzono, że w tej kolumnie istnieje opcja "Nieznane", zamieniono wartości NA na taką wartość. Obiekt w wersji zakodowanej zamieniono na liczbę wcześniej nie występującą w tej kolumnie.

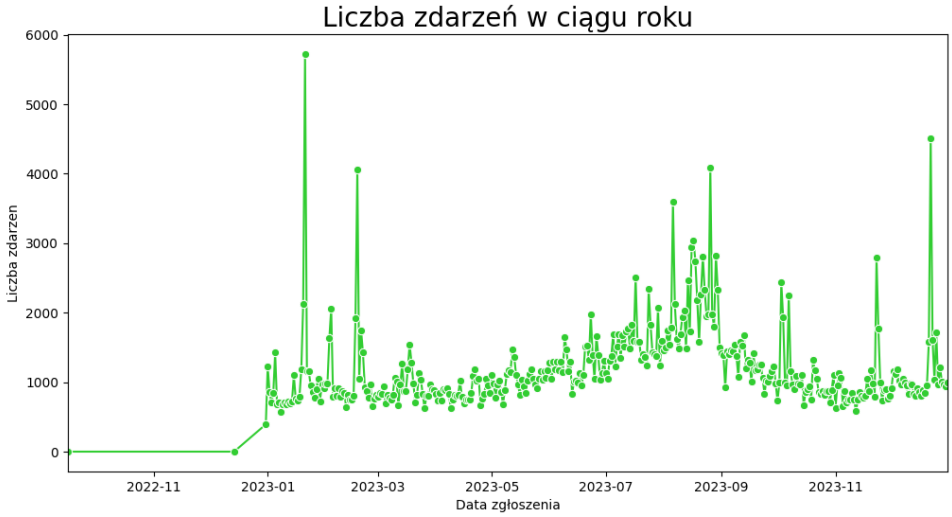
Pozostało zająć się wartościami brakującymi w kolumnach dotyczących dat. W przypadku predykcji czasu akcji nie będą one nam potrzebne, gdyż istnieje kolumna opisująca całkowity czas akcji w sekundach (SUM_CZAS). Stworzona zostanie dodatkowa tablica bez wierszy niezawierających te daty, aby móc analizować na przykład trendy w ciągu roku. Liczebność naszych danych dalej pozostaje spora, co nie powinno utrudnić analizy.

Postanowiono zwizualizować liczbę akcji w zależności od daty. Potrzebna była do tego konwersja naszych dat z typu string na typ datowy.

Rysunek 4: Zamiana dat

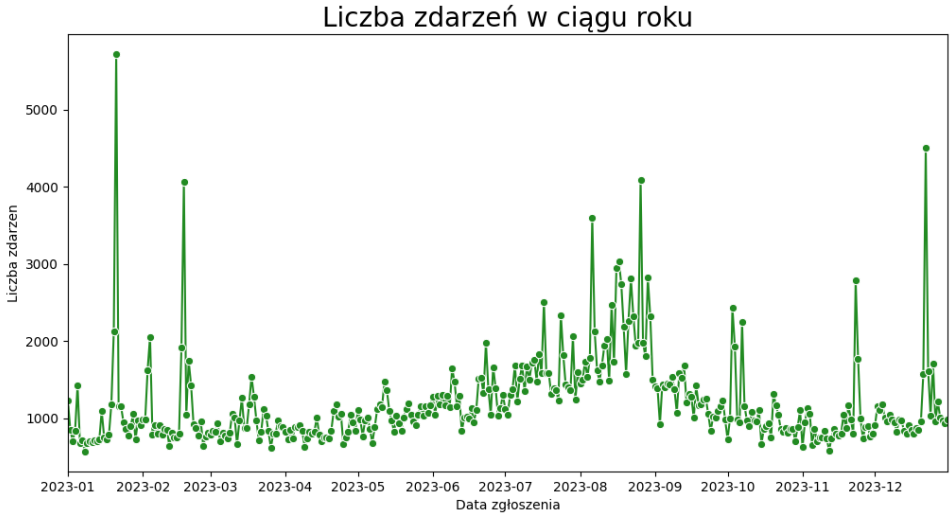
	DATA	DATA_ZGL
486048	2023-05-02	02.05.2023 14:58
486049	2023-07-26	26.07.2023 15:23
486050	2023-01-29	29.01.2023 16:06
486051	2023-07-11	11.07.2023 21:20
486054	2023-09-02	02.09.2023 01:24

Rysunek 5: Wykres przedstawiający liczbę zdarzeń w ciągu roku



Dzięki tej wizualizacji dostrzeżono, że zbiór zawiera obserwacje jeszcze z roku 2022, pomimo nazwy zbioru sugerującej, że chodzi tylko o rok 2023. Usunięto wiersze niedotyczące roku 2023 i powtórzono wykonanie wizualizacji.

Rysunek 6: Wykres po usunięciu danych z roku 2022



W tym momencie jest to już dużo lepiej widoczne i bardziej przygotowane do dalszej analizy. Przechodząc do kolumny opisującej rodzaj zdarzenia zauważono, że skrótowy zapis nie jest do końca jasny. Znaleziono plik o nazwie "Zasady ewidencjonowania zdarzeń w SWD PSP"dostępny na stronie gov.pl. Według informacji tam zawartych:

- P to pożar
- MZ to miejscowe zagrożenie
- AF to fałszywy alarm

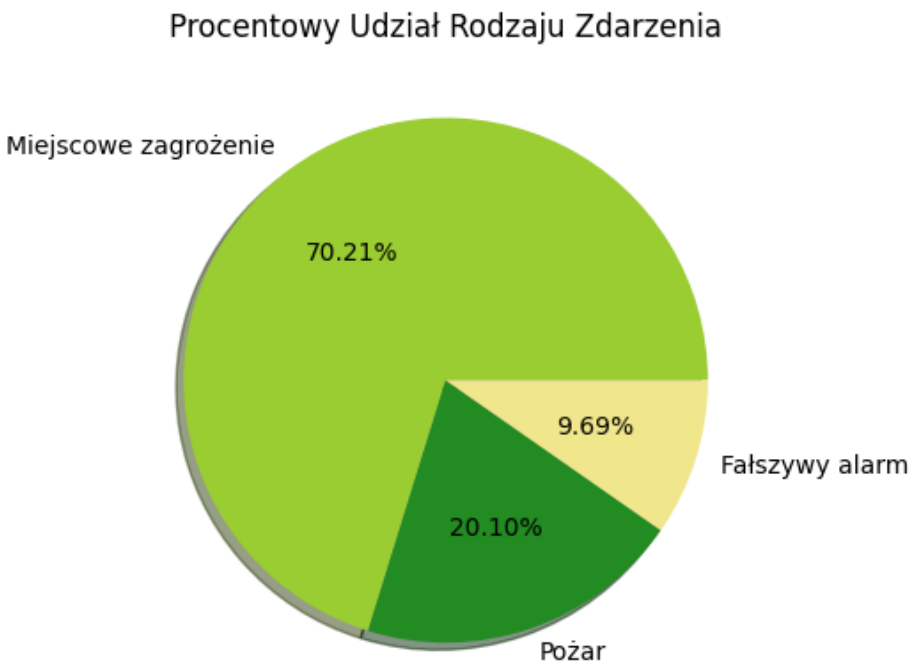
Dokonano tradycyjnego kodowania zmiennych, co może się okazać przydatne w kontekście dalszego przetwarzania danych lub tworzenia modelu.

Rysunek 7: Zakodowanie rodzaju zdarzenia

	RODZAJ	RODZAJ_KOD
0	MZ	0
1	P	1
2	AF	2

Aby sprawdzić ilościowy rozkład poszczególnego typu zdarzenia wykonano wykres kołowy.

Rysunek 8: Badanie rozkładu rodzaju zdarzeń



Jak widać na wyżej zaprezentowanym wykresie, aż 70% zdarzeń są miejscowymi zagrożeniami. Wniosek kolejny jest taki, że spora liczba, bo aż 10% wszystkich zdarzeń okazuje się być fałszywym alarmem.

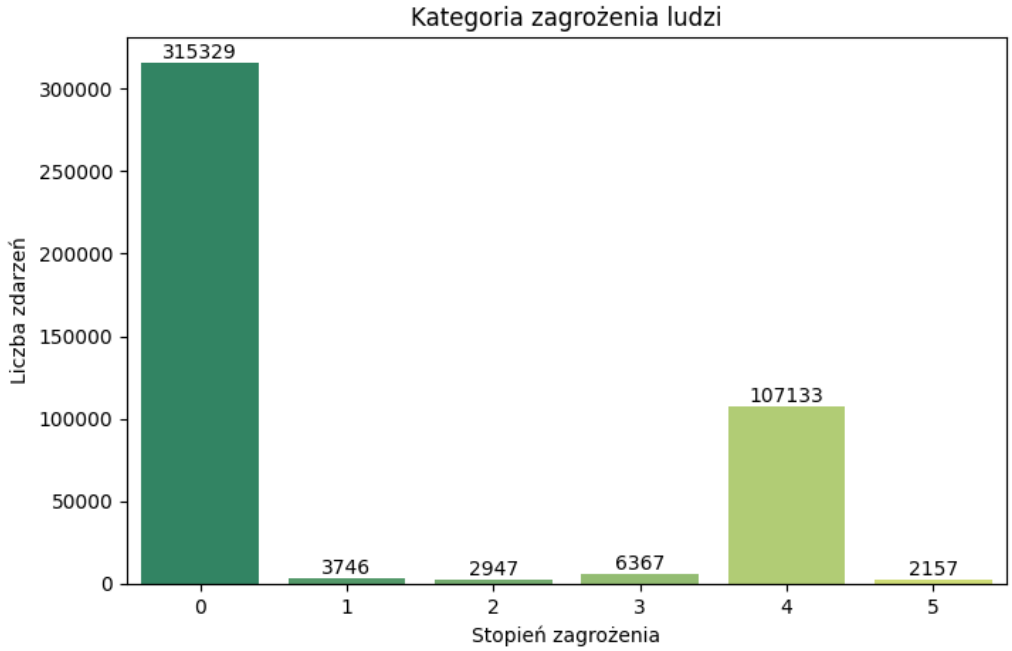
Pozostając w tematyce kodowania zmiennych kategorycznych, to samo należało dokonać dla zmiennej opisującej wielkość zdarzenia. Na ten moment nie wiadomo na ile ta zmienna jest zależna przykładowo od powierzchni zniszczeń. W tym wypadku, skrótowe nazwy były na tyle nieintuicyjnie rozpisane przez dostawcę, że postawnowiono nie zagłębiać się w ich znaczenie.

Rysunek 9: Kodowanie zmiennej WLK

	Wielkość Zdarzenia	Wielkość zakodowana
0	L	0
1	S	1
2	M	2
3	D	3
4	BD	4
5	IW	5
6	DW	6
7	Z	7

Przyjrzano się także kolumnie ZL, która według dostawcy zbioru opisuje kategorię zagrożenia ludzi. Jest ona w skali 0-5 (od braku informacj do największego zagrożenia). Zmienna ta domyślnie jest zakodowane nie trzeba więc tego robić.

Rysunek 10: Liczebność kategorii zagrożenia ludzi



Widać, że największą liczebność ma kategoria 0 - niestety jest to brak informacji. Jako następna w kolejnośći pod względem liczby zdarzeń jest 4 kategoria, która znacząco różni się od reszty.

Wracając do inżynierii cech, zdecydowano się na wykonanie sumowania w trzech przypadkach.

Pierwszy z nich dotyczy wszystkich służb uczestniczących w akcji. Pozostawiano kolumny bazowe z myślą, że mogłyby one być przydatne. Jako dwa następne wykonano sumowanie śmiertelnych przypadków oraz osób rannych.

Rysunek 11: Wykonane sumowanie dotyczące liczby osób biorących udział w akcji, liczby ranncyh i liczby przypadków śmiertelnych

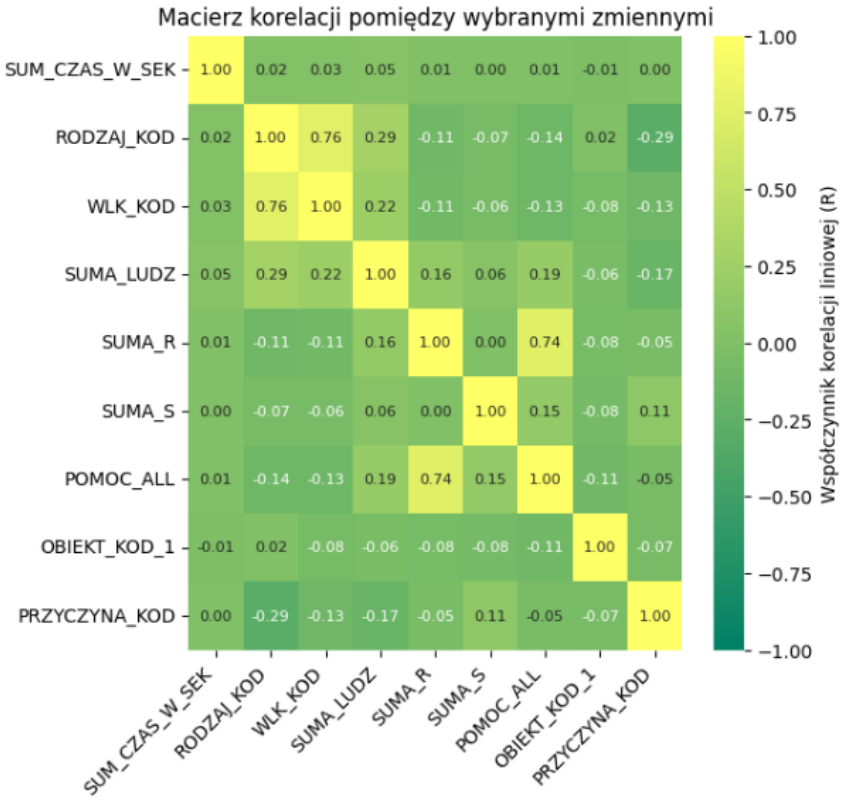
	SUMA_LUDZ	SUMA_R	SUMA_S
count	437679.00000	437679.000000	437679.000000
mean	7.96773	0.137523	0.025402
std	5.91967	0.503383	0.164068
min	0.00000	0.000000	0.000000
25%	5.00000	0.000000	0.000000
50%	6.00000	0.000000	0.000000
75%	10.00000	0.000000	0.000000
max	574.00000	34.000000	7.000000

Postanowiono wyświetlić zliczenia zdarzeń w poszczególnych województwach. Pozwoli to na wgląd na to, czy informacje mamy dostępne równomiernie względem województw.

Rysunek 12: Przedstawienie liczby zdarzeń w poszczególnych wojewodztwach

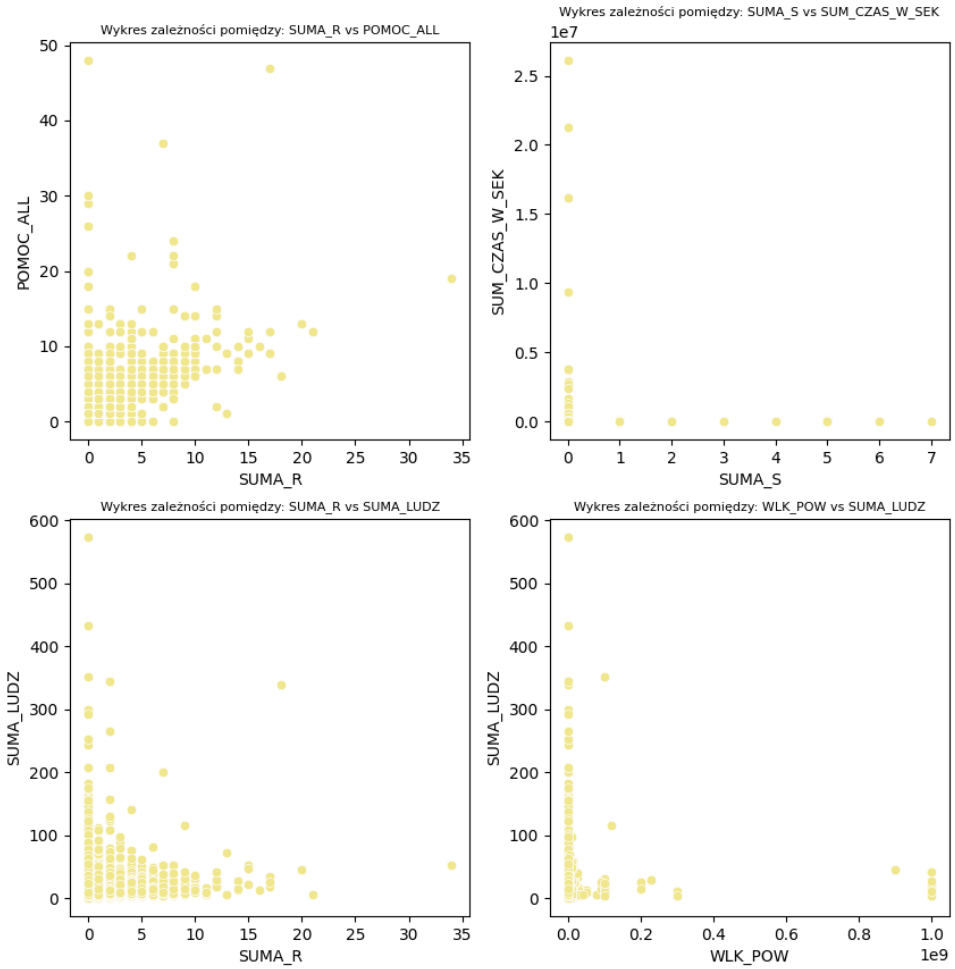
Da się zauważyć, że liczby te nie są jednakowe dla wszystkich województw. Pora sprawdzić teraz przypuszczalne zależności pomiędzy danymi. Wybranych zostało kilka kolumn i wyświetlona została macierz korelacji.

Rysunek 13: Wykres przedstawiający zależność pomiędzy zmiennymi



Wykres ten miał pokazać przewidywane zależności. Okazuje się, że są one w większości niewielkie, aczkolwiek warto było to zwizualizować. Jako następne wybrano cztery pary kolumn i wykonano wizualizację punktową. Miała na celu umożliwienie bardziej szczegółowego zrozumienia związków między nimi poprzez prezentację danych w formie punktów na wykresie.

Rysunek 14: Wykres przedstawiający zależność pomiędzy zmiennymi



4.4 Wybór kolumn TARGET oraz FEATURES

Przy tworzeniu modelu, jako zmienną objaśnianą, czyli tą którą będziemy chcieli przewidywać wybrano kolumnę `SUM_CZAS`, czyli sumaryczny czas akcji ratowniczej. Jest to korzystny wybór, ponieważ może ona być istotnym wskaźnikiem efektywności interwencji ratowniczych. Jeszcze większym plusem takiego wyboru mogłoby być lepsze odpowiedzi na sytuacje kryzysowe. W przyszlości może to również pomóc w lepszym zarządzaniu zasobami ludzkimi i środkami pomocy.

Jako zmienne FEATURES postanowiono wybrać następujący zestaw kolum:

- GMINA
- WLK - wielkość zdarzenia
- RODZAJ
- PRZYCZYNA
- `'WLK_POW'` - Wielkość powierzchniowa.
- `SUMA_R` - liczba osób rannych

Wybór padł na te zmienne, gdyż wydają się one wpływać na czas trwania. Słuszność tego wyboru może okazać się jednak wątpliwa, na przykład ze względu na przykład na wartość współczynnika korelacji.

Wizualizacji danych i ich przetworzenia można było wykonać na wiele różnych sposobów. Prawdopodobnie, nie wszystko co przydałoby się dalej zostało wykonane, aczkolwiek pozwoliło to wyciągnąć istotne informację na temat struktury naszych danych.