REGRESJA

Projekt praktyczny

Rozpoznanie danych

Epicurious - przepisy z oceną i wartościami odżywczymi:

Oceny z zakresu 0 - 5

Wartości odżywcze:

- Kalorie,
- Tłuszcze,
- Białko,
- Sód

674 kolumn z tagami i ponad 22,000 rekordów w pliku CSV

Problematyka

Analizowane problemy:

- Wpływa wartości odżywczych na ocenę
- Wpływ występowania mięsa w daniu na ocenę
- Wpływ ilości tagów przepisu na ocenę

Pozostałe hipotezy:

- Wpływ typu posiłku (śniadanie, obiad, kolacja) na ocenę
- Najczęściej występujące słowa w nazwach dań

Zapoznanie się z bazą danych

Wstępna obróbka danych

1. Usunięcie wierszy z wartościami NaN
2. Usunięcie wierszy z oceną 0
3. Czyszczenie outlierów (quantiles)

Dostosowanie danych do analizy:

1. Stworzenie kolumny z informacją o zawartości mięsa w daniu
2. Stworzenie kolumny z liczbą tagów dla każdego dania
3. Usunięcie zbędnych kolumn

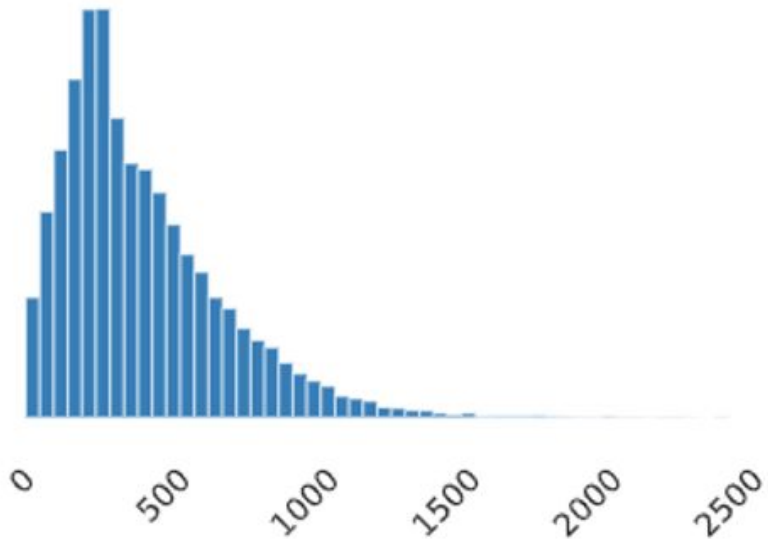
Ogólne statystyki przygotowanych danych

	rating	calories	protein	fat	sodium	with_meat	breakfast	lunch	dinner	num_tags
count	8858.000000	8858.000000	8858.000000	8858.000000	8858.000000	8858.000000	8858.000000	8858.000000	8858.000000	8858.000000
mean	4.081409	403.966358	15.194965	22.653534	438.542335	0.235042	0.045496	0.074622	0.119553	12.899639
std	0.628286	198.702606	12.322600	15.018741	363.788044	0.424049	0.208400	0.262795	0.324456	4.741075
min	1.250000	80.000000	3.000000	1.000000	15.000000	0.000000	0.000000	0.000000	0.000000	1.000000
25%	3.750000	250.000000	6.000000	12.000000	150.000000	0.000000	0.000000	0.000000	0.000000	9.000000
50%	4.375000	364.000000	10.000000	19.000000	332.000000	0.000000	0.000000	0.000000	0.000000	13.000000
75%	4.375000	529.000000	23.000000	30.000000	641.000000	0.000000	0.000000	0.000000	0.000000	17.000000
max	5.000000	985.000000	49.000000	85.000000	1592.000000	1.000000	1.000000	1.000000	1.000000	33.000000

Co można wynieść ze statystyk?

Według National Health Service (NHS) w Wielkiej Brytanii, przeciętny dorosły mężczyzna potrzebuje około 2500 kalorii dziennie, aby utrzymać swoją masę ciała na stałym poziomie, podczas gdy dorosłe kobiety potrzebują średnio około 2000.

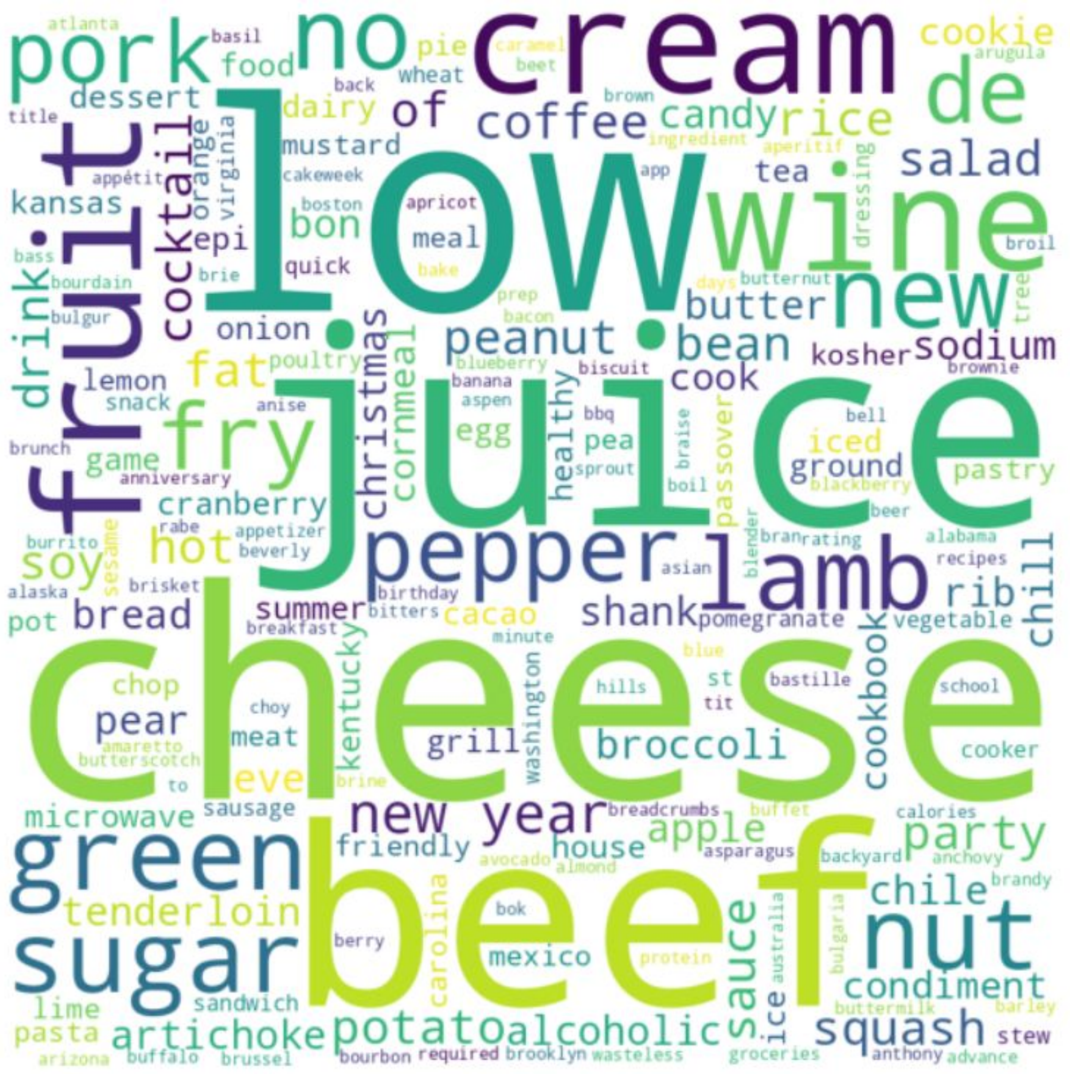
Średnie zawartość kalorii w daniach w zbiorze wyniosła 404kcal. Biorąc pod uwagę 5 posiłków dziennie bilans kaloryczny zostaje spełniony



Rozkład kalorii

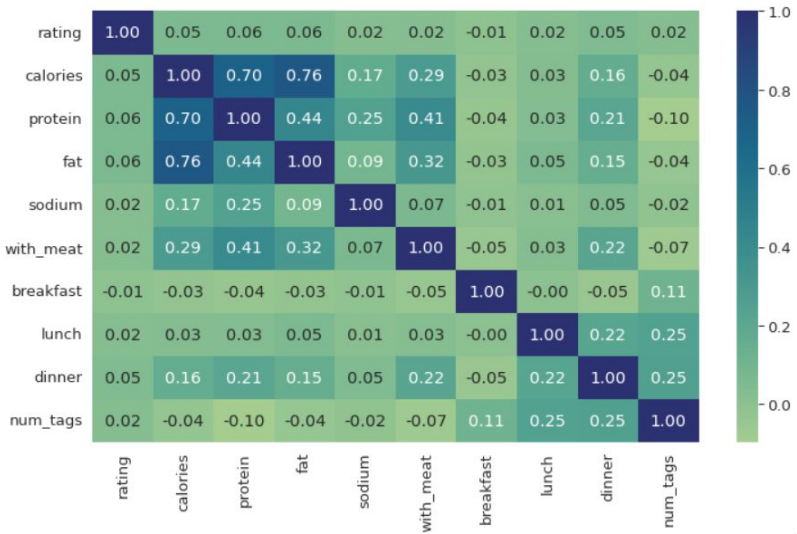
Mapa słów

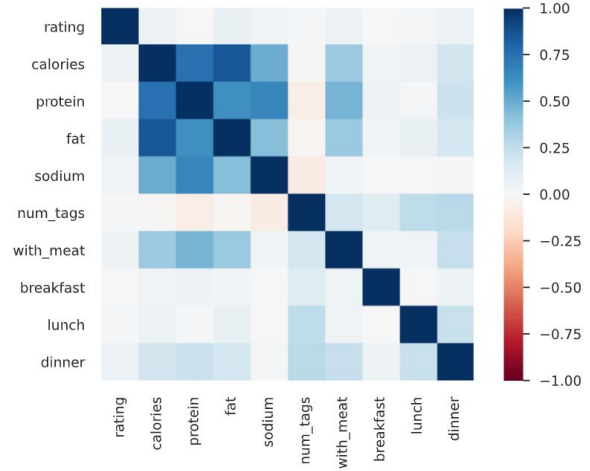
Mapa przedstawia najczęściej występujące słowa w nazwach dań w tabeli

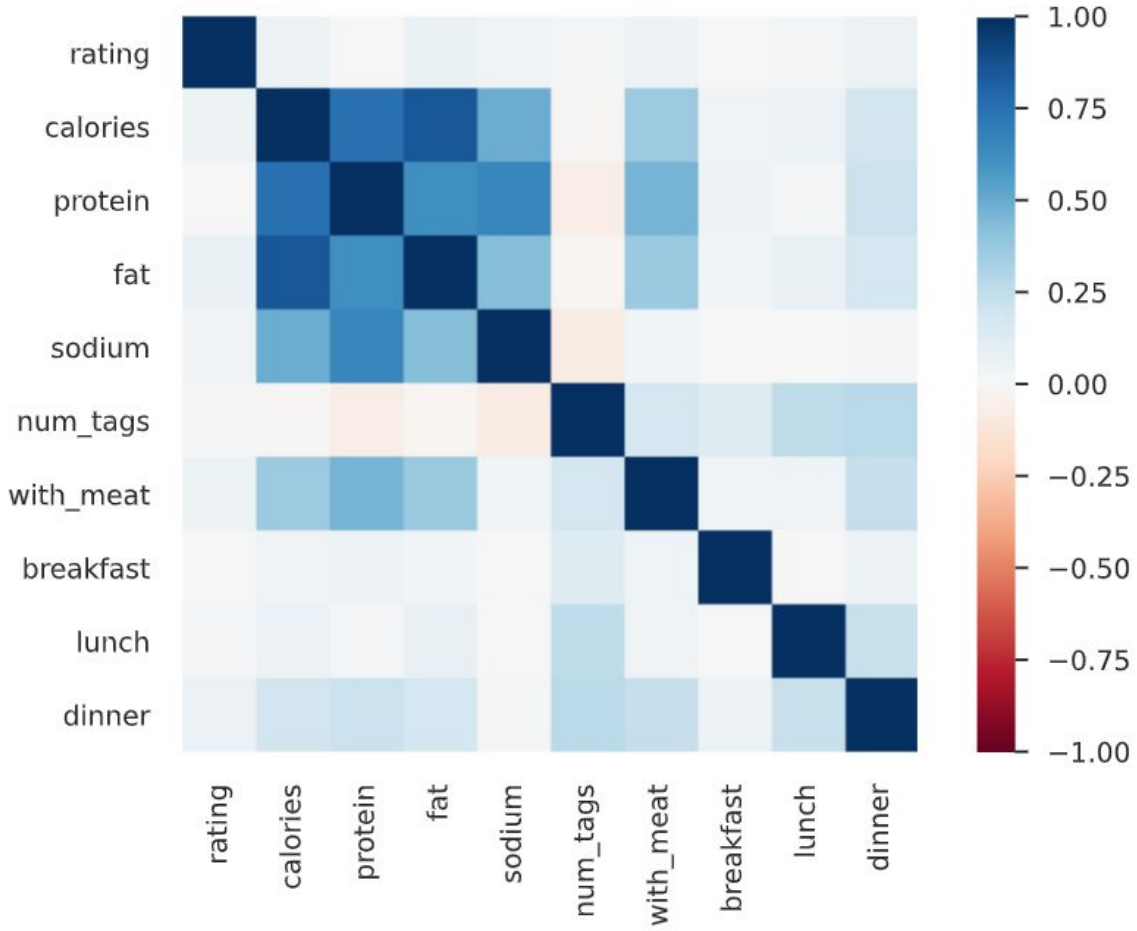


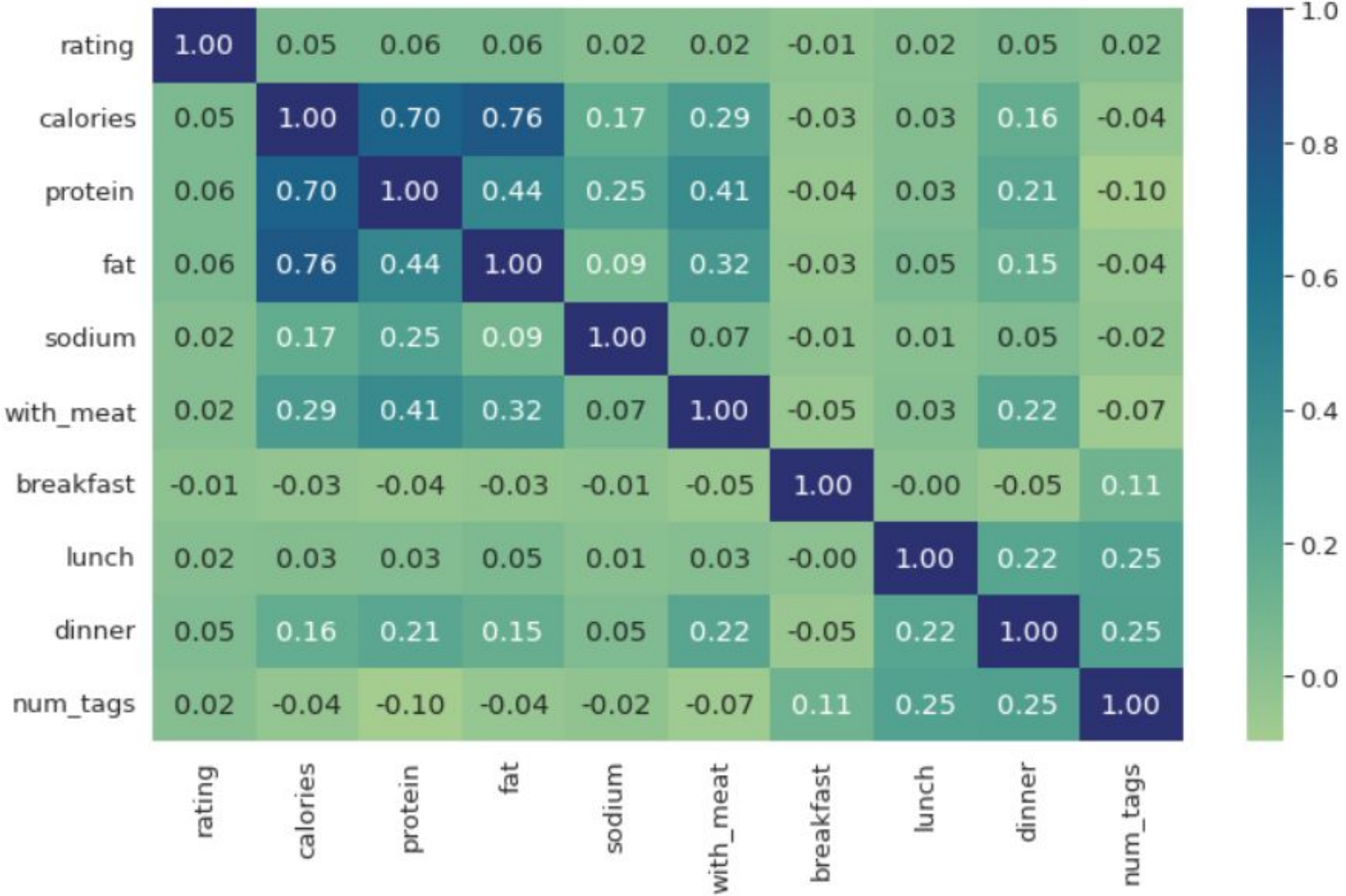
Heatmapa

Heatmapa to narzędzie do analizy danych, które w graficzny sposób (najczęściej wykorzystując kolory) reprezentuje to, jak zmienne są ze sobą skorelowane.






rating
calories
protein
fat
sodium
num_tags
with_meat
breakfast
lunch
dinner
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

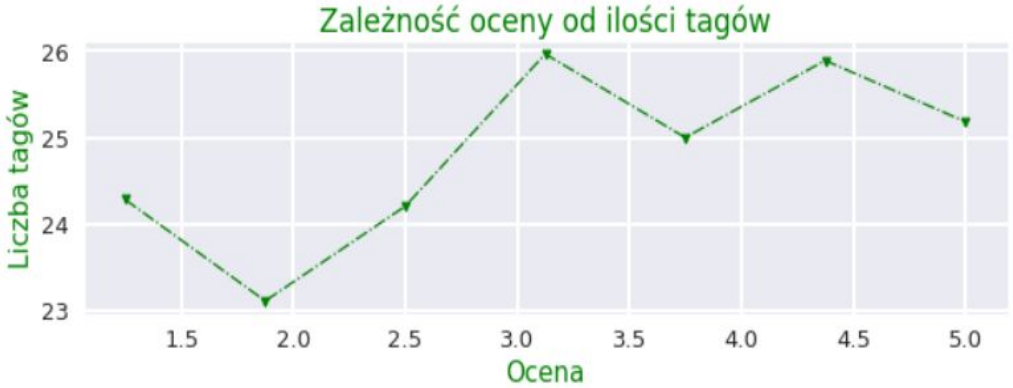

rating 1.00 0.05 0.06 0.06 0.02 0.02 -0.01 0.02 0.05 0.02
calories 0.05 1.00 0.70 0.76 0.17 0.29 -0.03 0.03 0.16 -0.04
protein 0.06 0.70 1.00 0.44 0.25 0.41 -0.04 0.03 0.21 -0.10
fat 0.06 0.76 0.44 1.00 0.09 0.32 -0.03 0.05 0.15 -0.04
sodium 0.02 0.17 0.25 0.09 1.00 0.07 -0.01 0.01 0.05 -0.02
with_meat 0.02 0.29 0.41 0.32 0.07 1.00 -0.05 0.03 0.22 -0.07
breakfast -0.01 -0.03 -0.04 -0.03 -0.01 -0.05 1.00 -0.00 -0.05 0.11
lunch 0.02 0.03 0.03 0.05 0.01 0.03 -0.00 1.00 0.22 0.25
dinner 0.05 0.16 0.21 0.15 0.05 0.22 -0.05 0.22 1.00 0.25
num_tags 0.02 -0.04 -0.10 -0.04 -0.02 -0.07 0.11 0.25 0.25 1.00
rating
calories
protein
fat
sodium
with_meat
breakfast
lunch
dinner
num_tags
1.0
0.8
0.6
0.4
0.2
0.0

Korelacje zmiennych

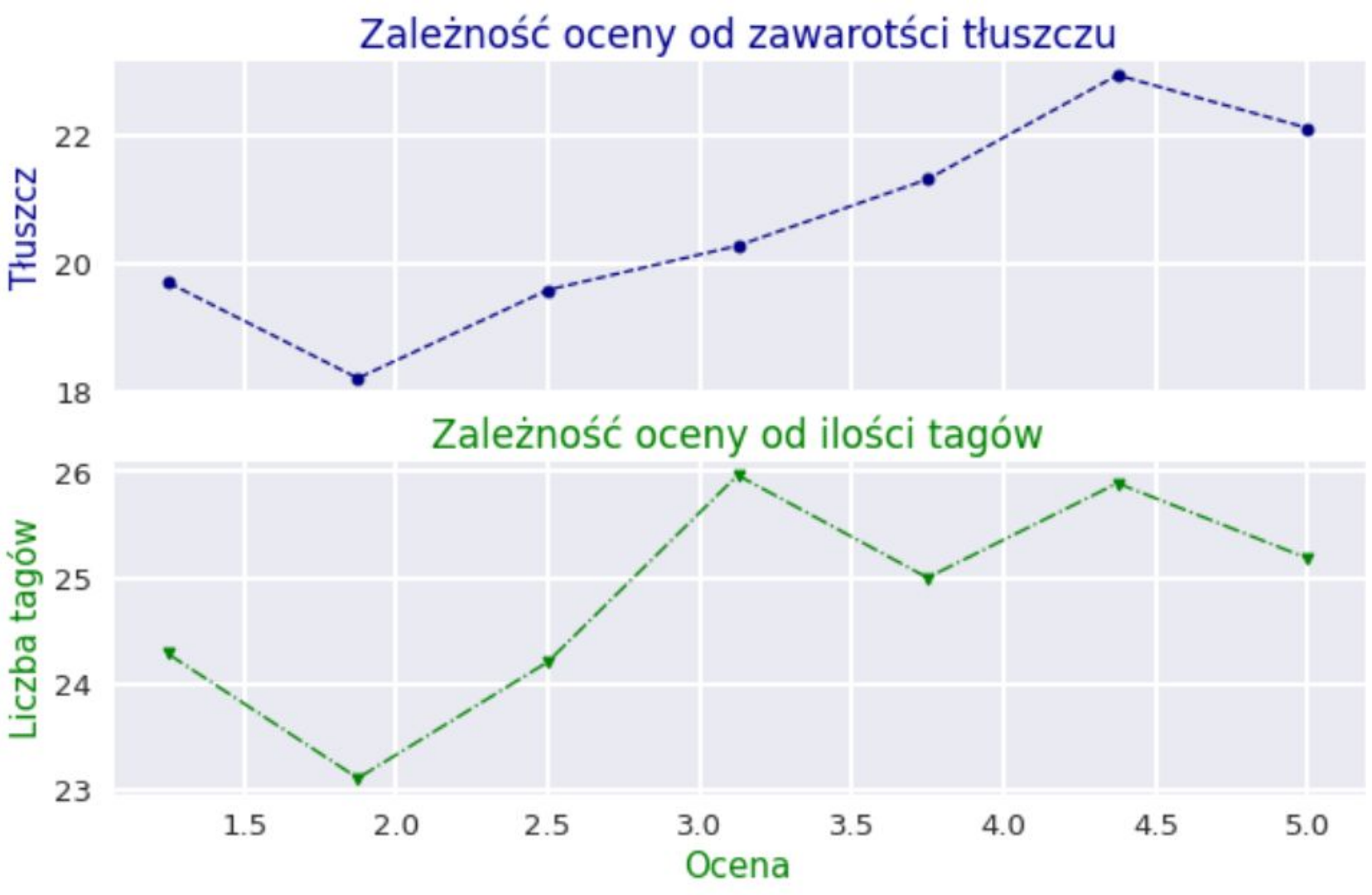
`title` has a high cardinality: 11640 distinct values	High cardinality
`calories` is highly overall correlated with `protein` and 1 other fields	High correlation
`protein` is highly overall correlated with `calories` and 2 other fields	High correlation
`fat` is highly overall correlated with `calories` and 1 other fields	High correlation
`sodium` is highly overall correlated with `protein`	High correlation
`breakfast` is highly imbalanced (77.4%)	Imbalance
`lunch` is highly imbalanced (65.4%)	Imbalance
`sodium` is highly skewed (γ1 = 30.2760962)	Skewed
`title` is uniformly distributed	Uniform
`protein` has 222 (1.7%) zeros	Zeros

Wykresy liniowe zależności oceny

W celu uzyskania maksymalnej oceny optymalną wartością tagów jest 25



Przy ilości 22g tłuszczów najbardziej prawdopodobne jest uzyskanie maksymalnej oceny


Zależność oceny od zawartości tłuszczu
Tłuszcz
22
20
18
Zależność oceny od ilości tagów
Liczba tagów
26
25
24
23
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Ocena

Jak rozkładają się oceny dań?

Distinct	7	**Minimum**	1.25
Distinct (%)	0.1%	**Maximum**	5
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	4.0838572	**Memory size**	202.6 KiB

Regresja

- **Zmienne niezależne:**

  "calories", "protein", "fat", "sodium", "with_meat", "breakfast", "lunch", "dinner", "num_tags"

- **Zmienna zależna:**

  “rating”

- **Podział na dane testowe i treningowe:**

  X_train, X_test, y_train, y_test

Regresja liniowa

Wyniki

- **Mean Absolute Error** (Średni błąd predykcji):

  0.49

- **Root Mean Squared Error**:

  0.64

- **R2 score** (W ilu % nasz model przewiduje dobrze):

  0.01

Drzewo decyzyjne

- Najlepsze parametry z wykorzystaniem kroswalidacji:

  max_depth': 14,

  'min_samples_split': 160

- **Mean Absolute Error** (Średni błąd predykcji):

  0.49

- **Root Mean Squared Error**:

  0.65

- **R2 score** (W ilu % nasz model przewiduje dobrze):

  -0.86

Wyniki

Żaden z dwóch modeli nie przewiduje oceny w zadowalającym stopniu.

Model liniowy jest zdecydowanie lepszy od drzewa decyzyjnego, dla którego R2 Score wyszedł ujemny.

Ocena dania na podstawie wykorzystanych przez nas danych jest niemożliwa do przewidzenia

Inny sposób agregacji i przekształceń danych mogą pomóc w przewidywaniu oceny.