Metody planowania i analizy eksperymentów Zadanie domowe nr 1: Analiza opisowa wybranych danych.

Student: Tomasz Mroczko

Analizowane będą dane dotyczące sprzętu do tenisa stołowego. Dane pobrane zostały wydobyte ze strony Revspin, która jest bazą opinii użytkowników na temat sprzętu.

Do pobrania oraz analizy danych użyty został Python wraz z bibliotekami. BeautifulSoup (bs4) posłużyło do scrapowania danych o ocenach okładzin tenisa stołowego ze strony Revspin. Pandas zostało użyte do manipulacji oraz podstawowej analizy danych. Matplotlib oraz Seaborn pozwoliły na wizualizację danych.

1 Dane

1.1 Format danych

Pozyskane dane prezentują się następująco

	rank	name	speed	spin	control	tackiness	weight	hardness	gears	throw_angle	consistency	durable	overall	ratings	price	brand
0	1	Nittaku Hammond Z2	9.3	9.3	9.2	2.2	3.6	6.9	8.6	6.9	9.5	8.4	9.5	74	59.0	Nittaku
1	2	Donic BlueStar A1	9.5	9.5	8.9	2.5	6.8	9.7	9.3	6.1	9.4	8.2	9.5	20	65.0	Donic
2	3	Donic Bluestorm Pro AM	9.1	9.5	9.2	1.5	5.3	6.9	9.4	5.6	9.2	7.9	9.5	14	60.0	Donic
3	4	Victas V>22 Double Extra	9.2	9.5	9.5	2.9	5.4	6.9	8.0	5.6	9.3	6.4	9.5	34	42.0	Victas
4	5	Tibhar Hybrid K1 European Version	9.1	9.5	9.2	5.5	5.2	7.9	9.0	6.0	9.6	8.4	9.5	18	37.0	Tibhar

- rank - pozycja w rankingu według oceny ogólnej (*overall*)
- name - nazwa okładziny
- speed - oceniana przez użytkowników prędkość
- spin - oceniana przez użytkowników możliwość nadawania rotacji
- control - oceniana przez użytkowników łatwość kontroli uderzeń
- tackiness - oceniana przez użytkowników lepkość gumy
- weight - oceniana przez użytkowników waga
- hardness - oceniana przez użytkowników twardość podkładu
- hardness - oceniana przez użytkowników łatwość grania uderzeń o różnorodnej prędkości
- throw_angle - oceniana przez użytkowników ostrość paraboli lotu topspina
- consistency - oceniana przez użytkowników powtarzalność jakości okładziny
- durable - oceniana przez użytkowników trwałość okładziny
- overall - średnia wszystkich statystyk
- ratings - liczba wystawionych ocen
- price - cena w dolarach amerykańskich USD
- brand - dodana do danych kolumna oznaczająca firmę produkującą

Cechy w zestawieniu można podzielić na dyskretne i ciągłe. Do cech dyskretnych należą rank i ratings (liczby całkowite) oraz name. Pozostałe cechy są ciągłe, ponieważ opisują parametry oceniane w skali liczbowej, takie jak speed, spin, control czy price.

1.2 Wartości danych

	rank	speed	spin	control	tackiness	weight	hardness	gears	throw_angle	consistency	durable	overall
count	593.000000	593.000000	593.000000	593.000000	593.000000	593.000000	593.000000	593.000000	593.000000	593.000000	593.000000	593.000000
mean	297.000000	8.469140	8.770320	8.725464	3.354637	4.662732	5.240472	7.367622	4.989882	8.679427	7.277066	8.877234
std	171.328632	0.854445	0.786346	0.455495	2.241346	1.379499	2.066908	1.336923	1.163071	0.969492	1.071694	0.473669
min	1.000000	3.900000	2.500000	6.400000	0.000000	1.100000	0.200000	1.300000	0.300000	5.000000	3.100000	5.800000
25%	149.000000	8.200000	8.500000	8.500000	1.800000	3.700000	3.900000	6.600000	4.300000	8.100000	6.700000	8.600000
50%	297.000000	8.700000	9.000000	8.800000	2.500000	4.600000	5.400000	7.600000	5.100000	8.900000	7.400000	9.000000
75%	445.000000	9.100000	9.200000	9.100000	4.900000	5.700000	6.700000	8.400000	5.800000	9.400000	8.100000	9.200000
max	593.000000	9.500000	9.600000	9.500000	8.900000	8.900000	9.900000	9.400000	8.400000	10.000000	9.400000	9.500000

W sumie pozyskano statystyki 593 okładzin do tenisa stołowego. Nie ma żadnych wartości nieprawidłowych ani brakujących - wynika to z ograniczeń strony z której pobrano dane.

Wszystkie dane numeryczne ciągłe należą do zakresu $\langle 1, 10\rangle$ - są one średnią ocen użytkowników z tego zakresu.

Średnie wartości *speed, spin, spin, control, consistency, overall* są na wysokim poziomie 8.5. Ich odchylenie standardowe jest niewielekie co sugeruje dużą inflację ocen - użytkownicy oceniają głównie pozytywnie i wysoko. Może to być efekt pozytywnej selekcji - oceniane są głównie produkty, które wywarły pozytywne wrażenie. Warto zaznaczyć że na wielu portalach zbierających opinie użytkowników, oceny wahają się głównie pomiędzy $\langle 7, 10\rangle$ i jest to popularne zaburzenie skali.

1.3 Wartość grupująca

Jako wartość grupująca użyta zostanie wartość w kolumnie *brand*. Jest to kolumna utworzona poprzez wyciągnięcie pierwszych słów z nazw okładzin. W sumie wydzielono 41 producentów, z których pod uwagę zostanie wziętych będzie tylko 10 najpopularniejszych producentów okładzin. Pozostali producenci mieli zbyt małą ilość opinii.

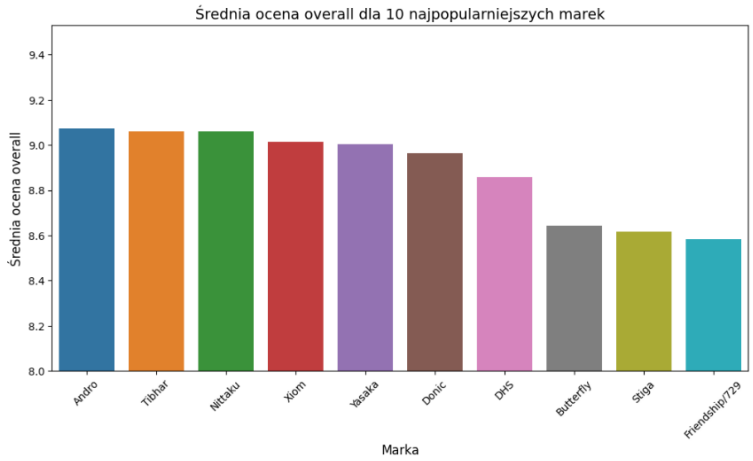
```
brand
Donic            52
Butterfly        50
Tibhar           48
Stiga            44
DHS              41
Andro            39
Friendship/729   35
Xiom             31
Yasaka           30
Nittaku          29
```

Po usunięciu okładzin pozostałych producentów pozostało aż 399 z 593 wpisów.

2 Analiza

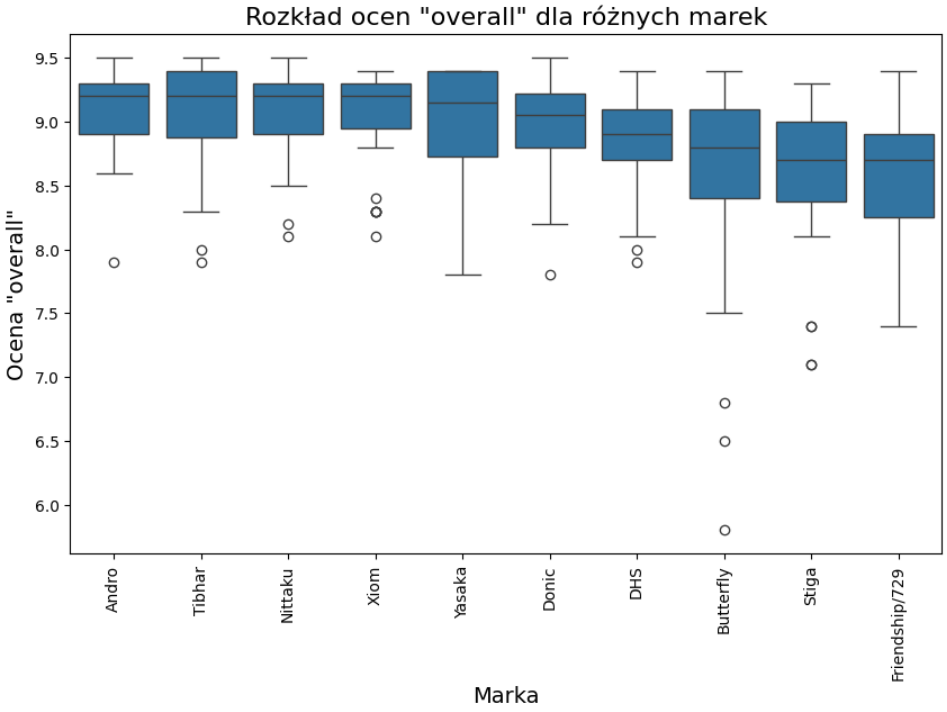
2.1 Średnie wartości overall

Zaczęto analizę od oceny średniej wartości *overall* dla poszczególnych marek.



Warto zaznaczyć że mediana była bardzo zgodna ze średnią. Oznacza to że średnia nie była mocno zawyżona w dół lub w górę przez wartości odstające.

2.2 Rozkład wartości overall producentów

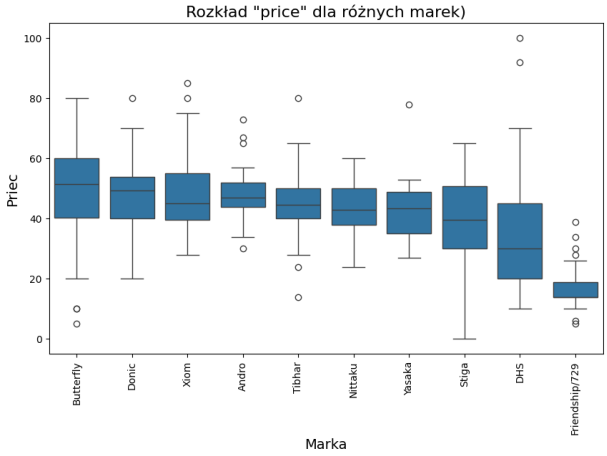


Analiza dystrybucji oceny *overall* gum różnych producentów potwierdziła, że większość ocen jset bardzo wysoka (inflacja ocen użytkowników). Z 10 topowych marek, można znaleźć bardzo niewiele gum ocenianych poniżej 7.

Warto zwrócić uwagę na *Butterfly*. Posiada ona największy rozrzut ocen. Pomimo wielu *bardzo* popularnych produktów, okładziny gorzej oceniane obniżyły jej pozycję w rankingu.

Prowadzące *Andro* posiada bardzo zwięzły rozkład - wszystkie produkty są bardzo wysoko oceniane

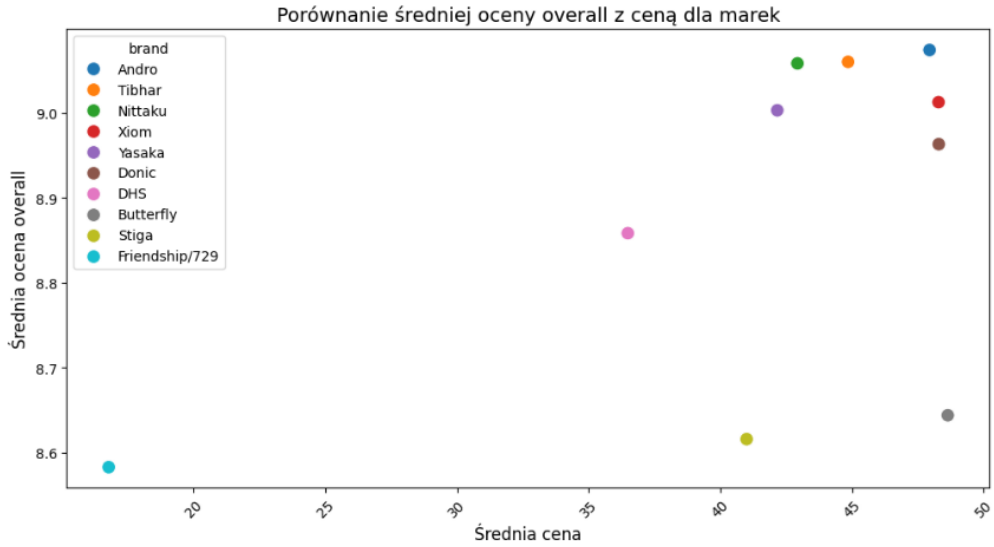
2.3 Rozkłady ceny producentów



Jak widać najdroższe produkty ma *Butterfly*. Chińskie marki *Friendship* oraz *DHS* są zdecydowanie najtańsze.

2.4 Średnie wartości overall a cena

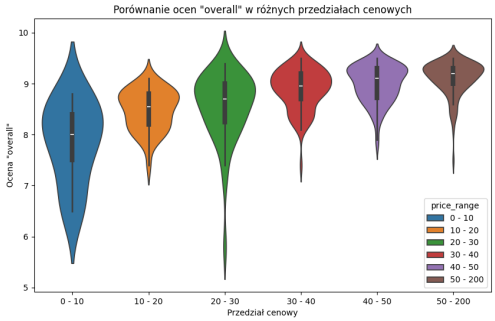
Następnie przeanlizowano średnią ocenę ogólną w stosunku do średniej ceny za okładzinę.



Wykres pokazuje dość jasno pozytywną korelację między średnią ceną gumy a średnią oceną. Warto zauważyć wyjątki, firmę *Butterfly* oraz *Stiga*, które pomimo wysokich cen posiadają dość niską średnią ocenę. Obie te firmy to klasyczne marki z dużą historią, uważane za wręcz prestiżowe. Być może wyższa stąd wynika nieproporcjonalnie wyższa cena do ocen użytkowników.

Marka *Butteryfly* tworzy sprzęt nastawiony zarówno na zawodowców jak na początkujących co także mogło negatywnie wpłynąć na jej niższą ocenę ogólną recenzje. Marki przodujące w rankingu, takie jak Andro oraz Tibhar tworzą głównie sprzęt profecsjonalny.

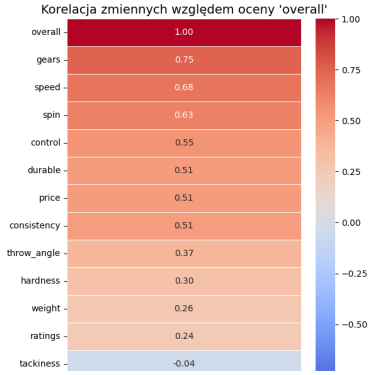
2.5 Średnie wartości overall a przedział ceneowy



Na wykresie rozkładu ocen ogólnych do przedziału cenowego widać że droższe okładziny są lepiej oceniane. Nie jest to jednak reguła i nawet w najtańszym przedziale znaleźć można bardzo dobrze oceniany sprzęt. Okładziny powyżej 30$ są konsekwentnie bardzo wysoko oceniane.

2.6 Korelacja cech z overall

Przenalizowano korelację wszystkich zgromadzonych cech gum względem oceny ogólnej *overall*.



Najwyższą korelację z oceną ogólną miała ocena *gears*, oznaczająca różnorodność prędkości jakie jest w stanie zapewnić guma przy uderzeniach oraz łatwość z jaką można to kontrolować. Może to oznaczać, że kiedy guma ma wysoki poziom *gears*, reszta jej współczynników jest także wysoka. Bardzo wysoką korelację miała także ocena *speed* oraz *spin*. Obie te cechy są fundamentalne dla jakości okładziny, bardzo często idą ze sobą w parze, znacząco wpływając na ocenę ogólną.

Parametr *control* wykazuje mniejszą korelację z ogólną oceną. Może wynikać to z faktu, że topowe okładziny są trudne w "okiełznaniu" dla początkujących graczy ze względu na eksplozywną naturę oraz wysoką czułość na rotację przeciwnika.

Co ciekawe *tackiness* ma negatywną korelację z oceną ogólną. Bardzo często lepką powierzchnią charakteryzują się gumy pochodzenia chińskiego (np. *DHS*), które wymagają bardzo dobrej techniki w celu wydobycia ich potencjału. Gumy europejskie oraz japońskie pozwalają łatwiej wyprowadzać wysokiej jakości uderzenia.