

PROJEKT

PIWA

MARTA JAGOWDZIK

CEL BADANIA:

W JAKIM STOPNIU ZMIENNA „KOLOR” (OPISUJĄCA BARWĘ
BRZECZKI PIWNEJ) ZALEŻY OD INNYCH ZMIENNYCH.



Cel pracy:

Zbadanie, w jakim stopniu zmienna „kolor” (opisująca barwę brzeczki piwnej*) zależy od innych zmiennych. W tym celu stworzony zostanie model, który najlepiej odwzoruje zależność między objaśniającymi zmiennymi ilościowymi a zmienną „kolor”.

*Brzeczka to ciecz ekstrahowana z procesu zacierania podczas warzenia piwa lub whisky.

WSTĘPNA ANALIZA I OPIS ZMIENNYCH

Wybór konkretnych zmiennych do badania zostanie dokonany w dalszej części pracy – po wstępnej analizie danych. Większość zmiennych w zestawie nie ma braków danych. Jednak zestaw zawiera również zmienne z dużą liczbą braków danych — ponad 69 000 z łącznej liczby 73 861 obserwacji.

**W ZBIORZE
ZNAJDUJE SIĘ
73.861 WERSZY
I 23 OPISUJĄCE
JE ZMIENNE.**

LICZBA BRAKUJĄCYCH WARTOŚCI W KAŻDEJ KOLEJNEJ KOLUMNIE

BeerID	0
Name	1
URL	0
Style	596
StyleID	0
Size(L)	0
OG	0
FG	0
ABV	0
IBU	0
Color	0
BoilSize	0
BoilTime	0
BoilGravity	2990
Efficiency	0
MashThickness	29864
SugarScale	0
BrewMethod	0
PitchRate	39252
PrimaryTemp	22662
PrimingMethod	67095
PrimingAmount	69087
UserId	50490

Większość zmiennych w zbiorze nie posiada żadnej brakującej wartości. W zbiorze znajdują się jednak także zmienne z ogromną liczbą brakujących wartości - **sięgających ponad 69.000 na łączną liczbę 73.861 obserwacji.**

Zmienne wraz ze skalą pomiarową oraz jednostką miary:

Size(L):

Skala pomiarowa: ilorazowa

jednostka miary: litr

OG:

Skala pomiarowa: ilorazowa

jednostka miary: gęstość (w odniesieniu do cieczy)

FG:

Skala pomiarowa: ilorazowa

jednostka miary: gęstość (w odniesieniu do cieczy)

ABV:

Skala pomiarowa: ilorazowa

jednostka miary: procentowa zawartość alkoholu

IBU:

Skala pomiarowa: ilorazowa

jednostka miary: Skala IBU

Color:

Skala pomiarowa: ilorazowa

jednostka miary: Skala 'SRM'

BoilSize:

Skala pomiarowa: ilorazowa

jednostka miary: galon

BoilTime:

Skala pomiarowa: ilorazowa

jednostka miary: minuty

BoilGravity:

Skala pomiarowa: ilorazowa

jednostka miary: gęstość (w odniesieniu do cieczy)

Efficiency:

Skala pomiarowa: ilorazowa

jednostka miary: procenty

MashThickness:

Skala pomiarowa: ilorazowa

jednostka miary: litr wody na kilogram ziarna

PitchRate:

Skala pomiarowa: ilorazowa skokowa

jednostka miary: liczba komórek drożdży (wyrażona w mln) na ml

PrimaryTemp:

Skala pomiarowa: przedziałowa

jednostka miary: stopnie Fahrenheita / Celsjusza

PrimingAmount:

Skala pomiarowa: ilorazowa

jednostka miary: objętość



KOLEJNYM ETAPEM W PROCESIE PREPROCESSINGU DANYCH JEST USUNIĘCIE BRAKUJĄCYCH WARTOŚCI W KOLUMNIE 'NAME' ORAZ W KOLUMNIE 'STYLE' :

WYBRANO ZMIENNĄ DO BADANIA:

ZMIENNA OBJAŚNIAJĄCA: KOLOR

WYJAŚNIAJĄCE ZMIENNE:

OG

FG

ABV

IBU

BOILTIME

BOILGRAVITY

EFFICIENCY

MASHTHICKNESS


PITCHRATE

PRIMARYTEMP

Oprócz powyższych zmiennych, do badania można również dodać zmienne „StyleID” i „BrewMethod” (po zdekodowaniu). Odpowiednie algorytmy będą w stanie modelować zależność pomiędzy zmienną „kolor” a zmiennymi ilościowymi oraz wspomnianymi zmiennymi jakościowymi. Jednak dla uproszczenia nie uwzględnimy tych dwóch zmiennych jakościowych w naszym badaniu.

Kolejnym krokiem jest imputacja danych w kolumnach ze zmiennymi ilościowymi:

- BoilGravity
- MashThickness
- PitchRate
- PrimaryTemp



BeerID	0
Name	0
URL	0
Style	0
StyleID	0
Size(L)	0
OG	0
FG	0
ABV	0
IBU	0
Color	0
BoilSize	0
BoilTime	0
BoilGravity	0
Efficiency	0
MashThickness	0
SugarScale	0
BrewMethod	0
PitchRate	0
PrimaryTemp	0
PrimingMethod	66520
PrimingAmount	68510
UserId	50011
dtype:	int64

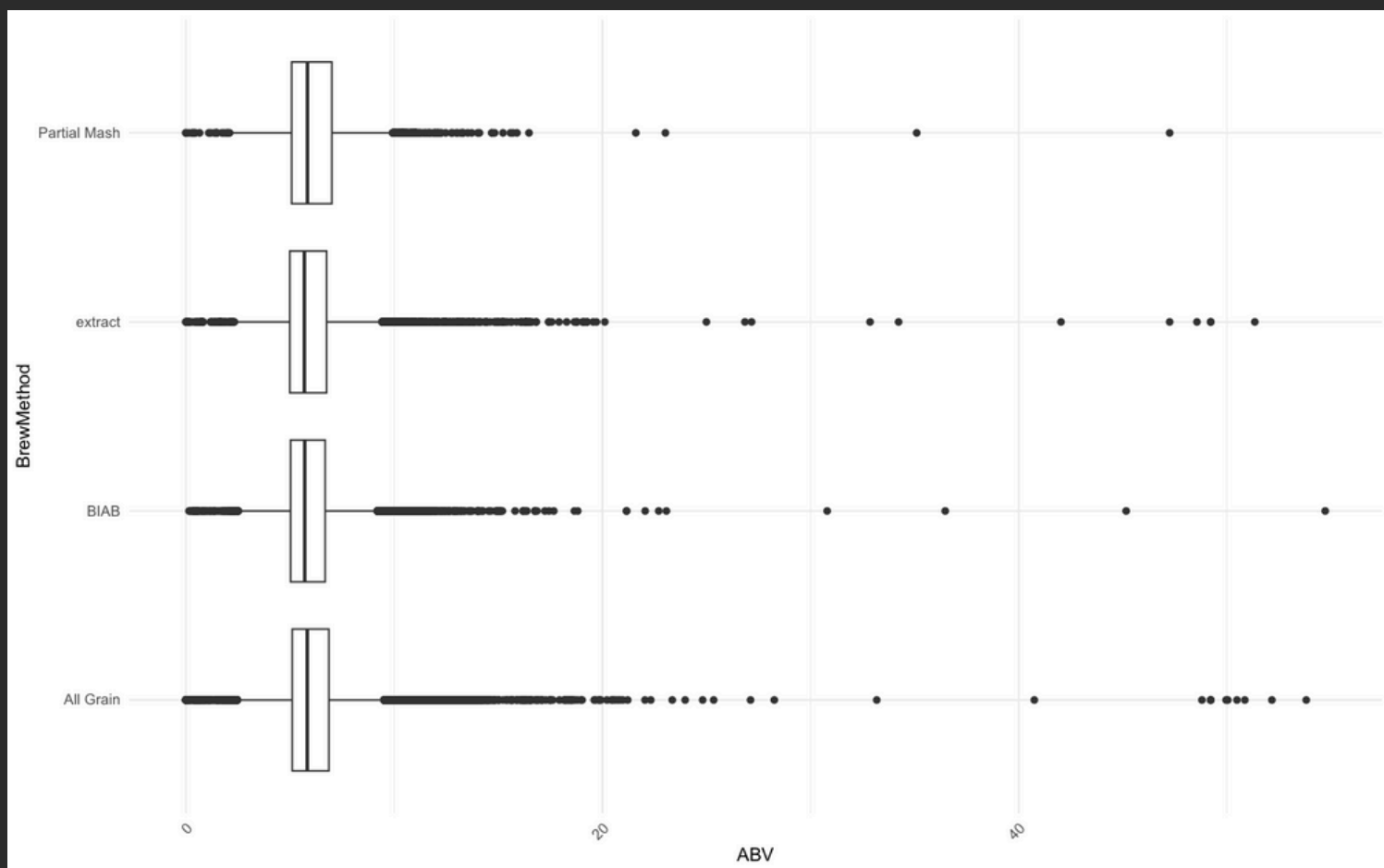
Po usunięciu brakujących danych w kolumnach Nazwa i Styl oraz podpisaniu danych w kolumnach zmiennymi ilościowymi, zostają nam trzy zmienne z dużą ilością brakujących danych:

- **Priming Method**- ma dużą liczbę brakujących wartości.
- **PrimingAmount** - zmienna przypisywalna, ponieważ jej wartości są wyrażone w różnych jednostkach i przedstawione w formie utrudniającej manipulację.
- **UserId** - zupełnie niepotrzebna zmienna badawcza.

Można by podjąć się zakodowania klas, a kolejno imputację daną techniką, jednakże występujący duży związek między wartościami obecnymi i brakującymi sprawia, iż można zadać sobie pytanie, czy będzie to skutecznie. **Uznano, że działania takie nie są potrzebne i ich nie zastosowano.**

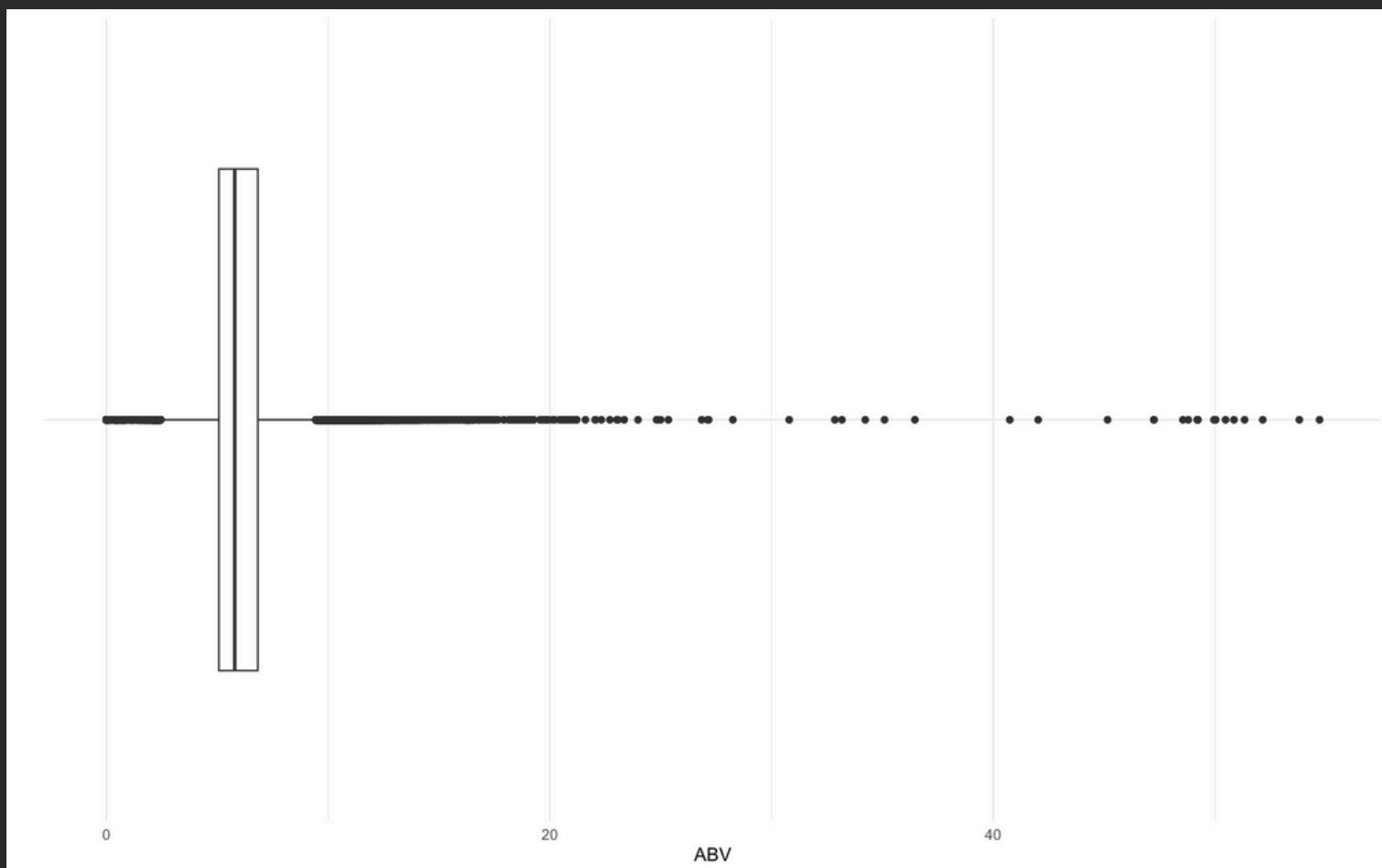
ZMIENNE MOŻNA WIZUALIZOWAĆ NA WIELE SPOSOBÓW:

Na przykład wykres pudełkowy dla zmiennej „ABV” w zależności od metody warzenia piwa (reprezentowanej przez kolumnę „BrewMethod”) pokazano poniżej.



Można zauważyć pewne niewspółmierności między różnymi metodami ważenia piwa. Można wizualnie zobaczyć, że metoda ważenia „**Częściowy zacier**” ma mniejszą liczbę wartości odstających w kolumnie zmiennej „**ABV**” niż inne metody.

PONIŻEJ ZAPREZENTOWANA ZOSTAŁA ZMIENNA 'ABV' NA WYKRESIE PUDEŁKOWYM BEZ PODZIAŁU NA KLASY W ZALEŻNOŚCI OD JAKIEJKOLWIEK INNEJ ZMIENNEJ.



Zmienna 'ABV' ma rozkład prawostronny.

Może to sugerować, że mediana (linia pozioma wewnątrz prostokąta) jest bliżej dolnego kwartyła (dolnego końca pudełka) niż górnego kwartyła (górnego końca pudełka). Ponadto, ogon "wąsów" boxplotu będzie dłuższy po prawej stronie (stronie górnej), co wskazuje na obecność wartości odstających lub ekstremalnie wysokich wartości.



WARTOŚCI SKRAJNE:

Nie usunięto z badania wartości ekstremalnych (pomimo tego, że są do tego podstawy).

Zamiasat tego wykonane zostanie **przykładowe wykrycie wartości skrajnych dla zmiennej 'OG'**. Zostanie ono wykonane metodą 3. odchyień standardowych na bazie średniej.

count	73264.000000
mean	1.406336
std	2.198066
min	1.000000
25%	1.051000
50%	1.058000
75%	1.069000
max	34.034500

PODSTAWOWE CHARAKTERYSTYKI ZMIENNEJ 'OG' POKAZUJE, ŻE WARTOŚCI ODSTAJĄCE MOGA DOSYĆ MOCNO WPŁYWAĆ NA SAM KSZTAŁT ZMIENNEJ.

JEJ STATYSTYKI OPISOWE. PODCZAS GDY ŚREDNIA WYNOSI 1.41, WARTOŚĆ MAKSYMALNA OBSERWACJI DLA ZMIENNEJ WYNOSI PONAD 34, A ODCHYLENIE STANDARDOWE ZMIENNEJ PRAWIE 2.2.

2.5606027516925094

WARTOŚCI SKRAJNE STANOWIĄ ŁĄCZNIE 2.56% ŁĄCZNEJ WARTOŚCI OBSERWACJI W KOLUMNIE ZE ZMIENNĄ 'OG'

Sprawdzenie, jak usunięcie zmiennych odstających wpłynie na statystyki opisowe dla zmiennej 'OG'

count	71388.000000
mean	1.061412
std	0.068161
min	1.000000
25%	1.050000
50%	1.057000
75%	1.067000
max	7.984240

USUNIĘCIE 1876 WARTOŚCI ODSTAJĄCYCH ZMNIJSZYŁOBY ŚREDNIĄ DLA ZMIENNEJ 'OG' Z 1.41 DO 1.06.

ODCHYLENIE STANDARDOWE TAKŻE BY SIĘ ZMNIJSZYŁO- Z 2.2 SPADŁOBY NA POZIOM 0.07.

WARTOŚĆ MINIMALNA NIE ULEGŁA ZMIANIE, JEDNAKŻE WARTOŚĆ MAKSYMALNA SPADŁA Z 34.03 DO 7.98

W projekcie dokonano także losowania prostego oraz losowania wielowarstwowego. wyniki znajdują się w programie rstudio.