Techniki Wizualizacji Danych
Praca domowa 8

Julia Kaznowska

25.01.2022

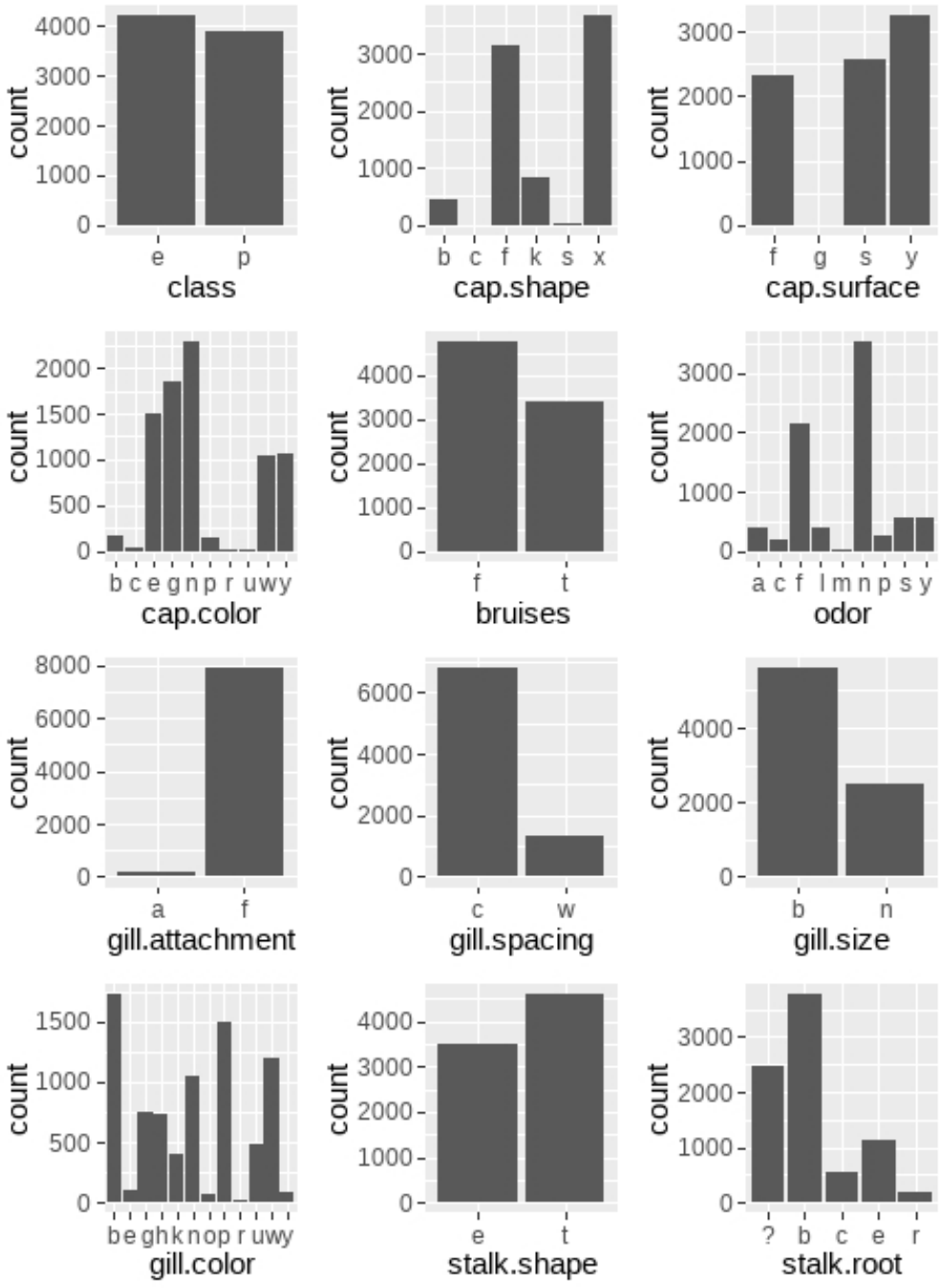
1 Wstępne informacje o danych

Moim zadaniem jest przygotowanie wstępnej analizy danych ze zbioru mushroom-classification. Zbiór ten składa się z 23 kolumn i 8124 obserwacji. Wszystkie dane są kategoryczne, co uniemożliwia wykorzystanie niektórych typów wykresów. Dzięki przyjrzeniu się unikalnym wartościom w poszczególnych kolumnach jesteśmy w stanie zauważyć jedną kolomnę, w której istnieją braki danych. Jest to kolumna `stalk.root`, a brakujące dane są oznaczone znakiem zapytania. Brakuje dokładnie 2480 rekordów, co stanowi około 30% wszystkich wartości. Kolumną z największą liczbą unikalnych danych jest `gill.color` (ma ich aż 12), natomiast najmniej unikalnych wartości posiada kolumna `veil.type` - jest tylko jedna.

2 Podstawowe wizualizacje

Na początku uznałam, że dobrym pomysłem będzie przedstawienie częstości występowania zmiennych w kolumnach, co można następnie wykorzystać do wyciągnięcia pewnych wniosków na temat wyników. Zacznijmy zatem od pokazania rozkładów w pierwszych 12 kolumnach.

2.1 12 pierwszych bohaterów



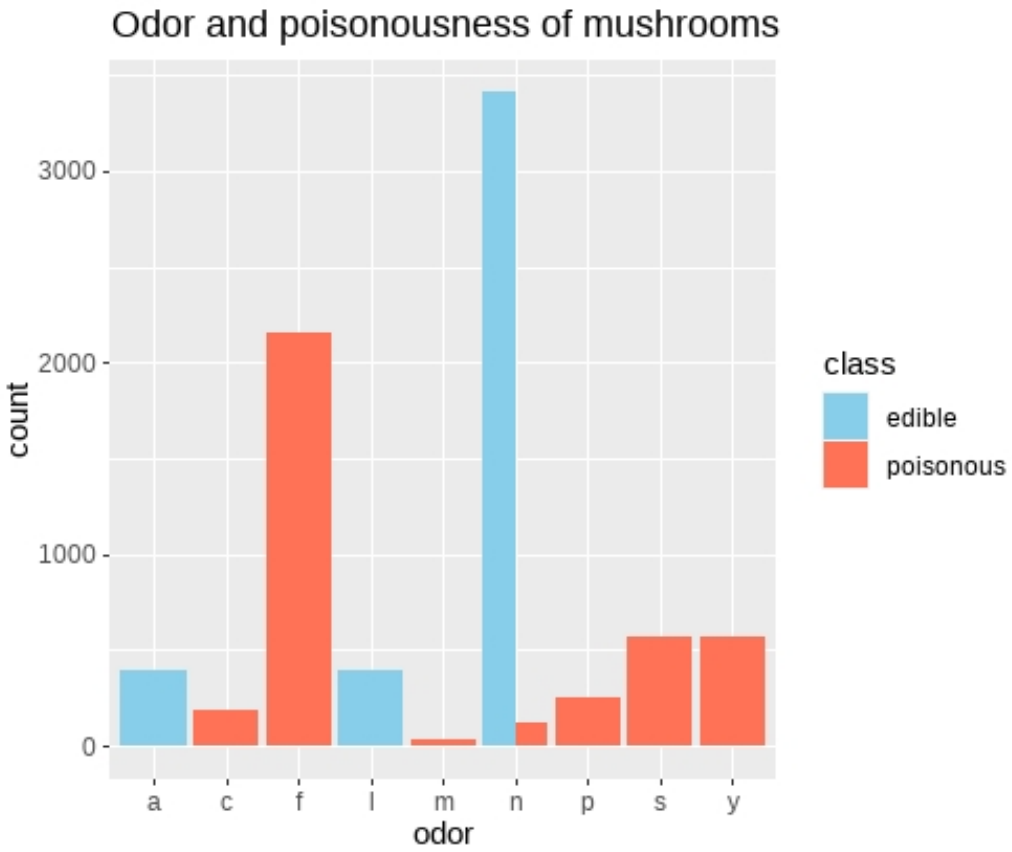
Krótko opiszmy sobie każdy z widocznych wykresów.

- Kolumna `class` mówi nam o tym czy grzyb jest jadalny, czy trujący. Widzimy, że obydwie liczby są porównywalne do siebie.
- Dzięki kolumnie `cap.shape` możemy dowiedzieć się, że większość grzybów ma wypukły bądź płaski kapelusz.
- Kolumna `cap.surface` informuje nas o tym, że grzyby ze żłobionym kapeluszem praktycznie nie występują - inne powierzchnie natomiast są równie częste.
- `Cap.color` jest jedną z bardziej zróżnicowanych kategorii - najczęstsze są brązowe, szare i czerwone kapelusze; z mniejszą częstością napotkamy żółte oraz białe kapelusze; inne kolory występują bardzo rzadko.
- Wartości w kolumnie `bruises` są w miarę równomiernie rozłożone
- Najwięcej grzybów nie posiada zapachu (`odor`). Jeśli natomiast grzyb ma zapach, to najczęściej jest on nieprzyjemny, cuchnący. Inne zapachy występują stosunkowo rzadko.
- Kolumna `gill.attachment` informuje nas o tym, że prawie wszystkie grzyby mają wolne (niełączące się z trzonem) blaszki.
- Zdecydowana większość grzybów posiada również blaszki w niewielkiej odległości (`gill.spacing`)
- Większość blaszek grzybów jest szeroka (`gill.size`)
- Kolor blaszek natomiast (`gill.color`) jest bardzo różnorodny i różnie rozłożony. Najczęściej występujące kolory to jasnobrązowo-żółty, różowy oraz biały.
- `Stalk.shape` informuje nas o tym, że trzony grzybów są w miarę równomiernie rozłożone na 2 wartości
- W kolumnie `stalk.root` natomiast, nie ma aż 30% danych, zatem trudno ją analizować.

2.2 Szukanie pierwszych powiązań

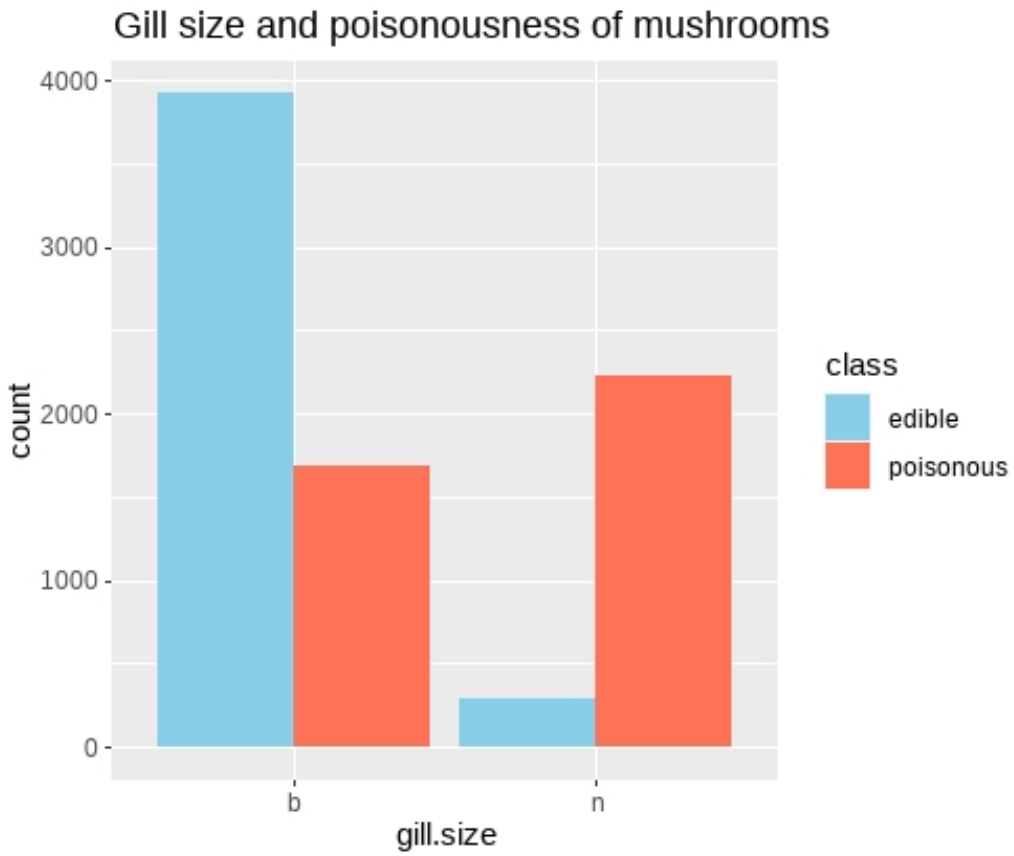
Byłoby to niezmiernie męczącym i czasochłonnym zadaniem - sprawdzenie każdej zależności (lub jej braku) w stosunku do dowolnej innej. Postanowiłam zatem poszukać tego, co może być najbardziej interesujące dla początkowego grzybiarza - w jaki sposób odróżnić grzyby jadalne od trujących?
Chociaż wiele z powyższych statystyk niewiele może nam powiedzieć, znalazłam 3 dość interesujące zależności.

2.2.1 Zapach



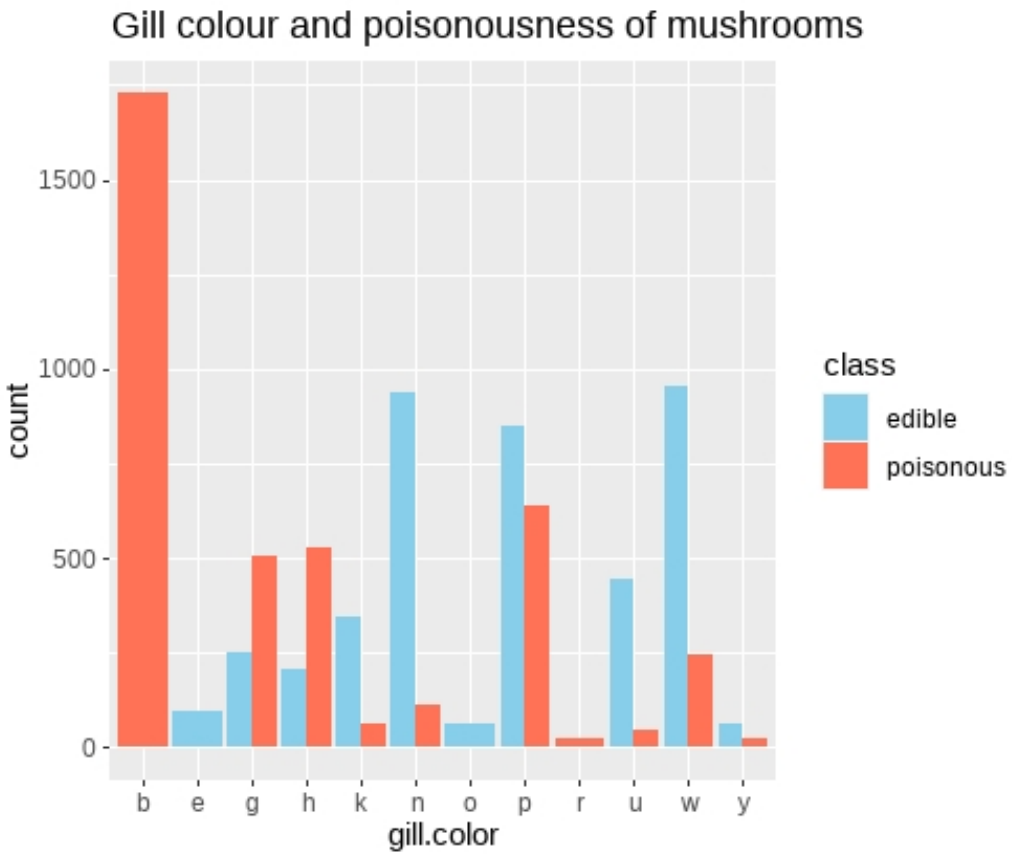
Na podstawie tego wykresu jesteśmy w stanie wywnioskować, że jeśli grzyb nie ma zapachu, to z dużym prawdopodobieństwem jest jadalny. Jeśli posiada on natomiast jakiś zapach (a już zwłaszcza cuchnący), to najlepiej będzie zostawić go w spokoju i zrezygnować z wrzucenia go do zupy czy sosu grzybowego.

2.2.2 Rozmiar blaszki



Dzięki powyższemu wykresowi wiemy, iż prawdopodobnie widok wąskiej blaszki powinien zachęcić nas do odłożenia zbieranego właśnie grzyba. Niestety, nie można powiedzieć tego samego w drugą stronę - mimo zdecydowanej przewagi przedstawicieli grzybów jadalnych pośród grzybów z blaszką szeroką, dalej istnieje prawie $\frac{1}{3}$ szans, iż osobnik taki będzie niezdatny do spożycia.

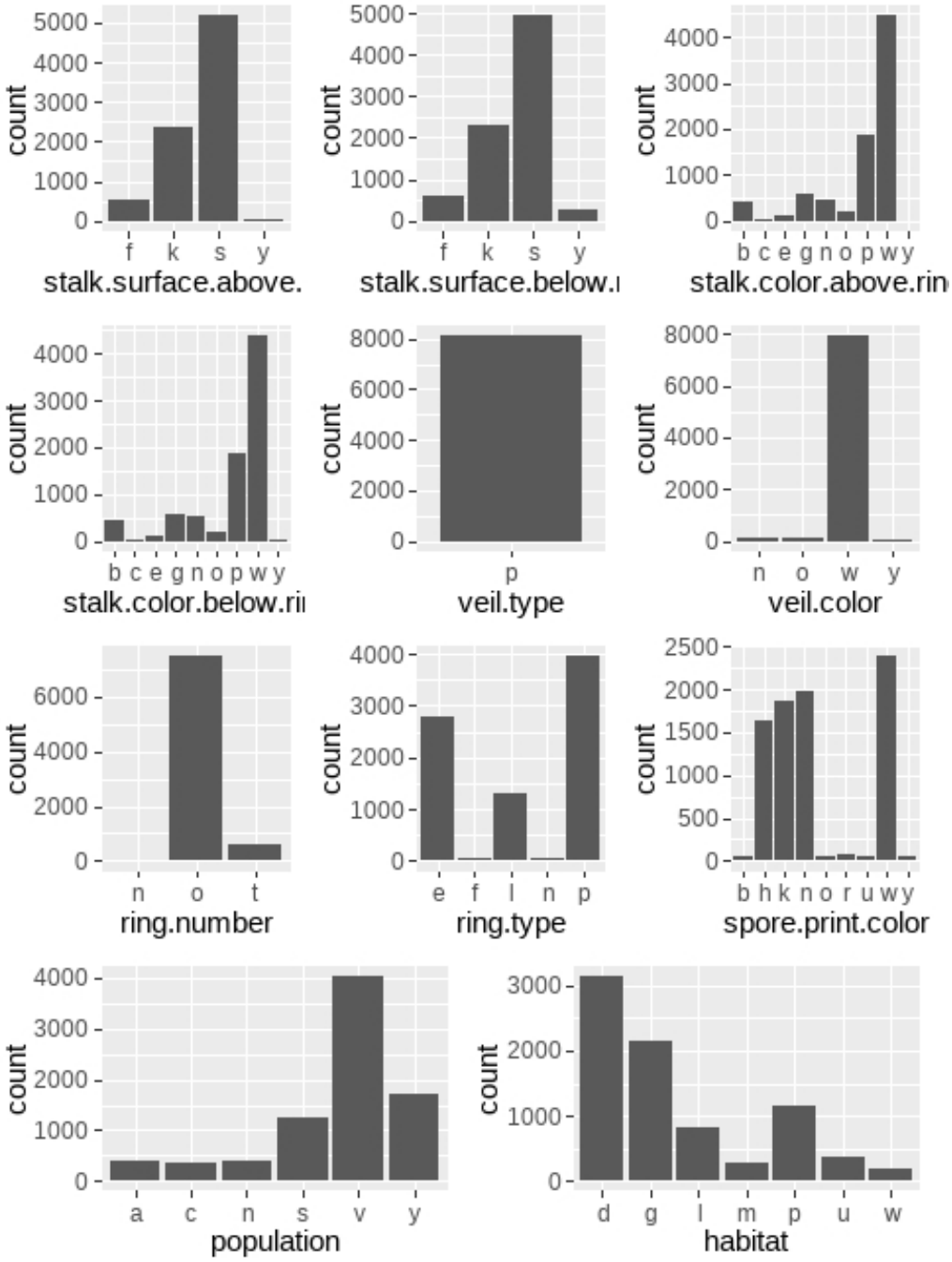
2.2.3 Kolor blaszek



W tym przypadku znaleźliśmy kolejny sposób na wykluczenie danego grzyba z naszego jadłospisu. Jeśli blaszki owego grzyba okażą się jasnobrązowo-żółte, to według naszych danych jest on z całą pewnością niejadalny. Zabawnym jest jednak fakt, iż kolor brązowy oznacza z dużą dozą prawdopodobieństwa możliwość zjedzenia danego grzyba bez większego uszczerbku na zdrowiu. Życzę z całego serca powodzenia początkującym grzybiarzom w odróżnianiu koloru brązowego od jasnobrązowo-żółtego, polecam jednak również kierowanie się innymi kategoriami podczas wybierania grzybów do swojego koszyczka.

2.3 11 kolejnych bohaterów

Po krótkiej przerwie na analizę pierwszych zależności powracamy z odczytywaniem danych z wykresów słupkowych.


count
stalk.surface.above.
f k s y
stalk.surface.below.
f k s y
stalk.color.above.rin
b c e g n o p w y
stalk.color.below.ri
b c e g n o p w y
veil.type
p
veil.color
n o w y
ring.number
n o t
ring.type
e f l n p
spore.print.color
b h k n o r u w y
population
a c n s v y
habitat
d g l m p u w

Krótki opis każdej kolumny:

- `Stalk.surface.above.ring` informuje nas o powierzchni trzonu grzyba powyżej jego pierścienia. Najwięcej grzybów ma gładką powierzchnię, najmniej natomiast łuskowatą. Powierzchnie jedwabiste i włókniste występują ze średnią częstością.

- Wykres kolumny `stalk.surface.below.ring` jest niezmiernie podobny do swojego poprzednika i przedstawia powierzchnię trzonu grzyba poniżej jego pierścienia. Liczba grzybów o łuskowatej powierzchni wzrosła jednak w stosunku do poprzedniego wykresu, nie są one zatem identyczne.

- `Stalk.color.above.ring` przedstawia kolor trzonu powyżej pierścienia. Dominującym kolorem jest biały, ponad dwukrotnie rzadziej występuje kolor różowy. Reszta kolorów występuje w niewielkim procencie.

- Podobnie jak w poprzedniej parze, wykres `stalk.color.below.ring` (kolor trzonu poniżej pierścienia) jest prawie identyczny w stosunku do swojego poprzednika, możemy zauważyć jednak drobne, niewielkie zmiany.

- Tak jak wcześniej zostało to zauważone, `veil.type` (typ błony czy osłony grzyba) posiada tylko jedną unikalną wartość.

- Kolor błony grzyba (`vail.color`) jest w przeważającej większości (wręcz prawie zawsze) biały. Reszta kolorów praktycznie nie występuje.

- Patrząc na wykres liczby pierścieni (`ring.number`) możemy zauważyć, że najczęściej grzyby mają tylko jeden. Rzadkością jest znalezienie grzyba z dwoma bądź brakiem pierścieni.

- `Ring.type` (typ pierścienia) jest zróżnicowany przede wszystkim na 3 możliwości: pedant, evanescent, large (w kolejności malejącej; niestety nie udało mi się wyszukać odpowiedniego polskiego tłumaczenia tych wartości bez przeglądania szeregu forów grzybiarskich, dlatego zostawię te oryginalne, angielskie)

- `Spore.print.color` czyli kolor zebranych na papierze zarodników grzyba. Wyróżniamy 4 najczęściej występujące kolory: biały, brązowy, czarny i czekoladowy. Inne kolory występują nieporównywalnie rzadziej.

- `Population` czyli czy grzyby występują w skupiskach i jeśli tak, to jakich? Najczęściej grzyby występują w skupiskach liczących kilka osobników.

- `Habitat`, czyli środowisko w jakim grzyby występują. Najczęstsze środowiska to lasy i obszary trawiaste.

2.4 Kolejne powiązania

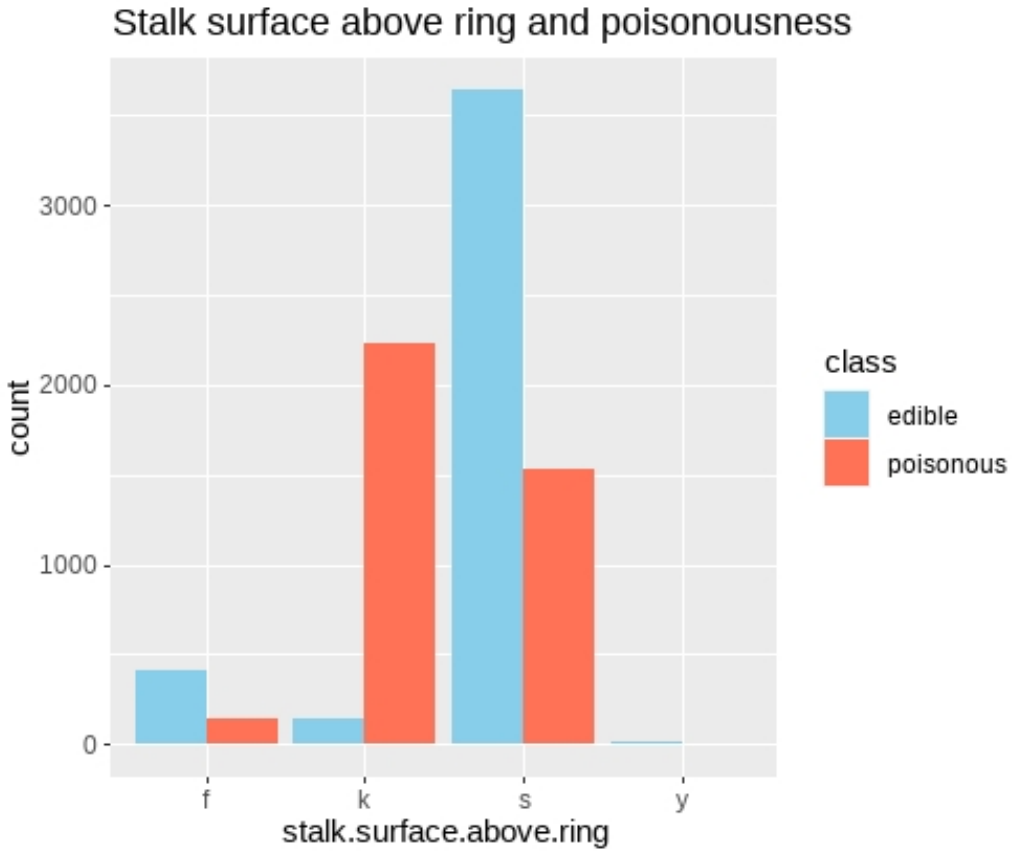
Tym razem postanowiłam sprawdzić coś, co zainteresowało również mnie, jako osobę analizującą dane, czyli porównanie w jakim procencie grzybów zmienia się powierzchnia i kolor trzona pod pierścieniem w stosunku do ponad pierścieniem. Dane postanowiłam przedstawić w tabelkach.

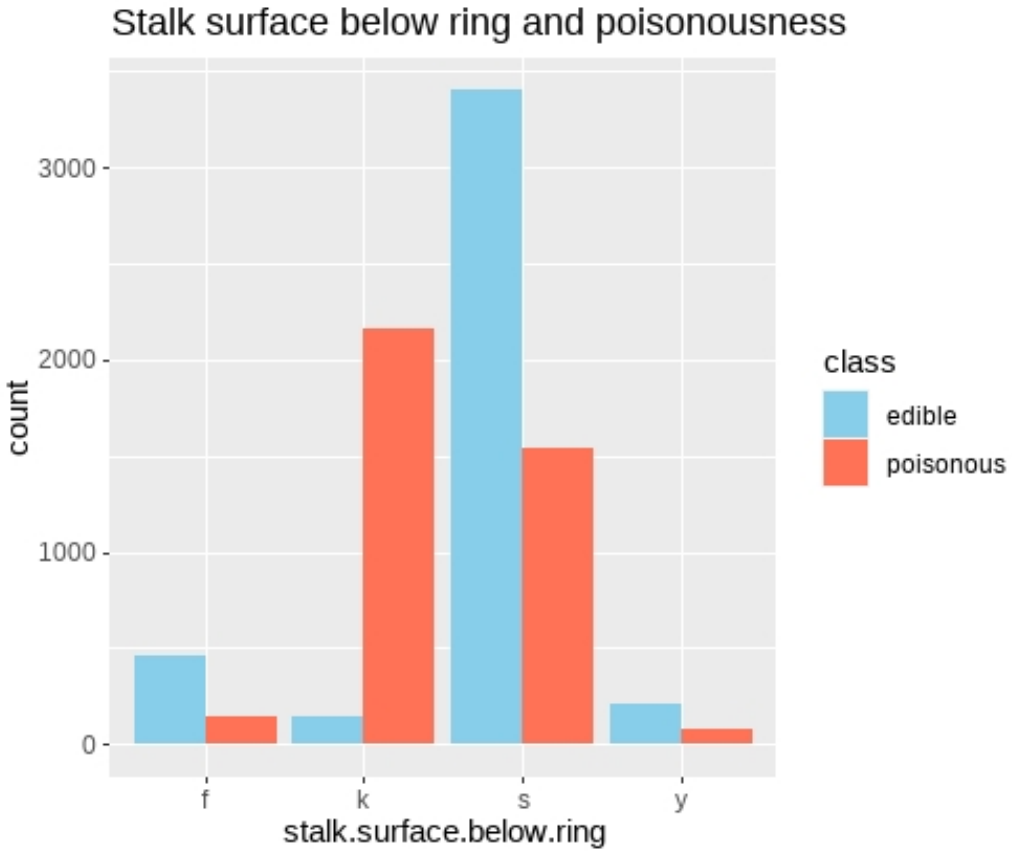
2.4.1 Powierzchnia trzona

	stalk.surface.above.ring	stalk.surface.below.ring	count
1	s	s	4156
2	k	k	1800
3	k	s	504
4	s	k	504
5	s	f	324
6	f	f	276
7	f	s	276
8	s	y	192
9	k	y	68
10	y	y	24

Jak widzimy, zdecydowana większość grzybów ma taką samą powierzchnię trzona ponad, jak i poniżej pierścienia (jest ich dokładnie 6256, czyli około 77%). Istnieją jednak mieszane trzony i bardzo interesującym jest, że wykresy ostatecznie się niewiele od siebie różnią.

Udało mi się również otrzymać 2 podobne wykresy, jeśli chodzi o zależność powierzchni trzona od jadalności grzybów:


Stalk surface above ring and poisonousness
count
3000
2000
1000
0
f
k
s
y
stalk.surface.above.ring
class
edible
poisonous



O ile wnioskowanie jadalności grzyba może być dość niebezpieczne na podstawie tych wykresów, o tyle można założyć, że jeśli część jego trzona jest jedwabista, to grzyb nadaje się do zjedzenia tylko raz.

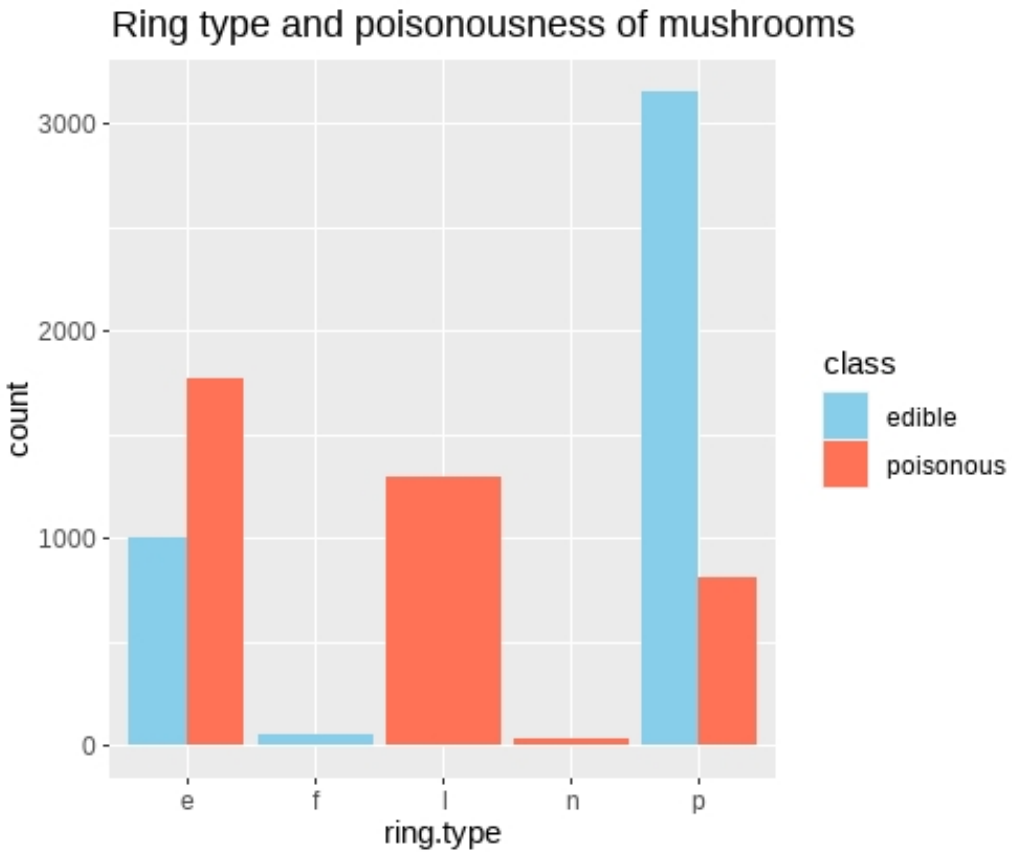
2.4.2 Kolor trzona

Analogiczną analizę przeprowadziłam dla koloru trzona:

	Above (color)	Below (color)	count
1	w	w	3520
2	p	p	768
3	p	w	624
4	w	p	624
5	g	g	192
6	g	p	192
7	g	w	192
8	o	o	192
9	p	g	192
10	w	g	192
11	n	n	160
12	b	b	144
13	b	n	144
14	b	p	144
15	n	b	144
16	n	p	144
17	p	b	144
18	p	n	144
19	w	n	64
20	e	e	48
21	e	w	48
22	w	e	48
23	c	c	36
24	w	y	16
25	y	y	8

Widzimy, że dla jedynie 9 unikatowych wartości kolorów mamy aż 25 różnych kombinacji kolorowania grzyba nad i pod pierścieniem. Kolory pokrywają się w około 60% przypadków (5068 grzybów). W tym przypadku jest to jeszcze bardziej interesujące niż w poprzednim - w końcu zmienia się całkiem sporo, a mimo tego - wykresy pozostają bardzo podobne.
Dla koloru trzona niestety nie ma ciekawego związku z jego jadalnością, niemniej jednak wykresy również są bardzo podobne (czego można było się spodziewać).

2.5 Typ pierścienia



Ten wykres natomiast pokazuje nam, że grzyby o `large` pierścieniach powinny zostać tam, gdzie były znalezione. Są one z całą pewnością niejadalne. Dla pierścieni `pedant` można założyć, że grzybek będzie dobry do zjedzenia, oczywiście jeśli nie boimy się odrobiny ryzyka.

2.6 Kolor zarodników

Jeśli ktoś wybrałby się do lasu ze sporym zapasem kartek papieru, następnie znalazł odpowiedniego grzyba - nie za starego ani nie za młodego, a później zrobił odcisk jego zarodników na tejże kartce, to mógłby w pewien sposób odkryć czy grzyb ów jest jadalny, czy nie. Z wykresu widzimy, że grzyby zostawiające kolory czarny oraz brązowy raczej powinny być jadalne, natomiast lepiej stronić od czekoladowych i białych kolorów.

3 Dodatkowa analiza poprawności szukania grzybów

W tej sekcji chciałam sprawdzić jaka jest szansa, że przez tę analizę ktoś mógłby niechcąco natknąć się na trującego osobnika podczas szukania podstawy do obiadu na najbliższe dni, bądź wręcz przeciwnie - ominąć całkiem niezły egzemplarz ze strachu, że jest on niejadalny. Zobaczmy zatem co mówią liczby:

3.1 Dobre grzyby ze wskazówek

Według mojej analizy dobre powinny być grzyby bezzapachowe, o `pedant` pierścieniu, brązowym kolorze blaszek oraz o czarnym bądź brązowym kolorze odcisku zarodników. Sprawdźmy jak prezentują się dane dla takich wartości: Okazuje się, że stosując się do wszystkich podanych wskazówek

	Edible?	count
1	true	472

nasz grzyb na pewno będzie się nadawał do zjedzenia. Ach, wspaniała statystyka!
Niestety okazuje się też, że w powyższy sposób zignorujemy bardzo dużo jadalnych grzybów, bo aż około 89%. No cóż, analiza statystyczna ma również swoje problemiki.

3.2 Złe grzyby ze wskazówek

Grzyby, których lepiej unikać, według moich wskazówek, powinny charakteryzować się okropnym zapachem, jasnobrązowo-żółtymi wąskimi blaszkami, jedwabistej powierzchni trzona, `large` pierścieniem oraz czekoladowym bądź białym kolorem odcisku zarodników. Brzmi jak wiele kryteriów, zatem sprawdźmy, jak bardzo dokładne one są.
Okazuje się, że po zastosowaniu wszystkich filtrów nie istnieje taki grzyb, który łączyłby w sobie wszystkie złe cechy. Jest to interesujący wynik, aczkolwiek nie do końca przydatny w codziennym życiu, zwłaszcza dla początkującego grzybiarza, który raczej nie ogląda się za trującymi grzybami, tylko pewnie próbuje ich unikać.

3.3 Kilka słów końcowych

Jak wiadomo, z tego zestawu danych można było wyciągnąć kilka wskazówek co do grzybów, które powinny być dobre do zjedzenia. Na swoje pierwsze grzybobranie jednak zawsze warto jest wybrać się z kimś bardziej doświadczonym. Osoba ta może pomóc w rozpoznawaniu grzybów, odróżnianiu ich od siebie, może posiadać nawet więcej informacji o grzybach, które nie były zawarte w tym zbiorze danych. Sama analiza była jednak ciekawym doświadczeniem, mimo początkowych trudności związanych z brakiem zmiennych ilościowych.