Analiza wydatkow na alkohol we Wloszech

Projekt na zaliczenie przedmiotu Wstep do Analizy Danych

Michal Rajda

27-01-2021

Wstep do projektu

W ramach projektu wykorzystuje duzy zbior danych, obejmujacy dane jednostkowe badania budzetu gospodarstw domowych we Wloszech w roku 2011.

Jego celem jest mozliwie dobre przygotowanie danych do badania i wstepna analiza opisowa oraz graficzna dotyczaca wybranego przeze mnie tematu - Alkoholu.

Aby ulatwic wstepne rozmyslenia dotyczace mojej pracy postanowilem wypisac kilka pomocniczych hipotez wybranych na podstawie wlasnych przemyslen:

- Jak wyglada nasza proba?
- Kto kupuje alkohol?
- Jak wygladaja wydatki na alkohol?

Obrobka danych

Standardowo zaczynam poprzez wczytanie danych i selekcje interesujacych nas danych.

```
alkohol <- dane %>%
  select(Mese, Regione, Eta4_1, Sesso1, Titstu1, C_1803, C_1804, C_1805) %>%
  rename(Miesiac = Mese, Region = Regione, Wiek = Eta4_1, Plec = Sesso1, Edukacja = Titstu1,
         Wino = C_1803, Piwo = C_1804, Inny_alkohol = C_1805)
```

W obecnej chwili dane prezentuja sie nastepujaco:

Miesiac	Region	Wiek	Plec	Edukacja	Wino	Piwo	Inny_alkohol
7	7	3	1	6	0.00	12.61	0
11	3	3	2	6	0.00	9.59	0
7	7	3	1	4	0.00	0.00	0
9	15	3	1	4	6.06	0.00	0
3	10	2	2	7	0.00	0.00	0

```
## 'data.frame':    23074 obs. of  8 variables:
##  $ Miesiac     : num  7 11 7 9 3 6 1 6 3 9 ...
##  $ Region      : num  7 3 7 15 10 15 12 9 1 3 ...
##  $ Wiek        : int  3 3 3 3 2 3 2 4 3 3 ...
```

```
##  $ Plec        : int  1 2 1 1 2 1 2 2 1 1 ...
##  $ Edukacja    : int  6 6 4 4 7 7 4 7 4 6 ...
##  $ Wino        : num  0 0 0 6.06 0 ...
##  $ Piwo        : num  12.61 9.59 0 0 0 ...
##  $ Inny_alkohol: num  0 0 0 0 0 0 0 0 0 0 ...
```

Interpretacja zmiennych:

Miesiac - miesiac przeprowadzenia ankiety Region - region zamieszkania osoby ankietowanej Wiek - przedzial wiekowy do ktorego nalezy osoba ankietowana. Wyrozniamy cztery takie grupy:

- osoby ponizej 18 roku życia
- osoby majace od 18 do 34 lat (mloda_dorosla)
- osoby majace od 35 do 64 lat (dorosla)
- osoby majace powyzej 64 lat (starsza)
- Wino - wydatek osoby ankietowanej na wino (w euro)
- Piwo - wydatek osoby ankietowanej na piwo (w euro)
- Inny_alkohol - wydatek osoby ankietowanej na inne alkohole (w euro)

Obszar Wloch podzielilismy na 3 grupy:

Polnoc:

- Piemont i Valle d'Aosta
- Lombardia
- Trentino Alto Adige
- Veneto
- Friuli Venezia Giulia
- Liguria
- Emilia Romagna

Srodek:

- Toskania
- Umbria
- Marki
- Lazio
- Abruzja
- Molise

Poludnie:

- Kampania
- Apulia
- Basilicata
- Kalabria
- Sycylia
- Sardynia

Dla wykresow przedstawiajacych, czy dana grupa kupuje alkohol czy tez nie:

- 0 - nie
- 1 - tak

Braki danych

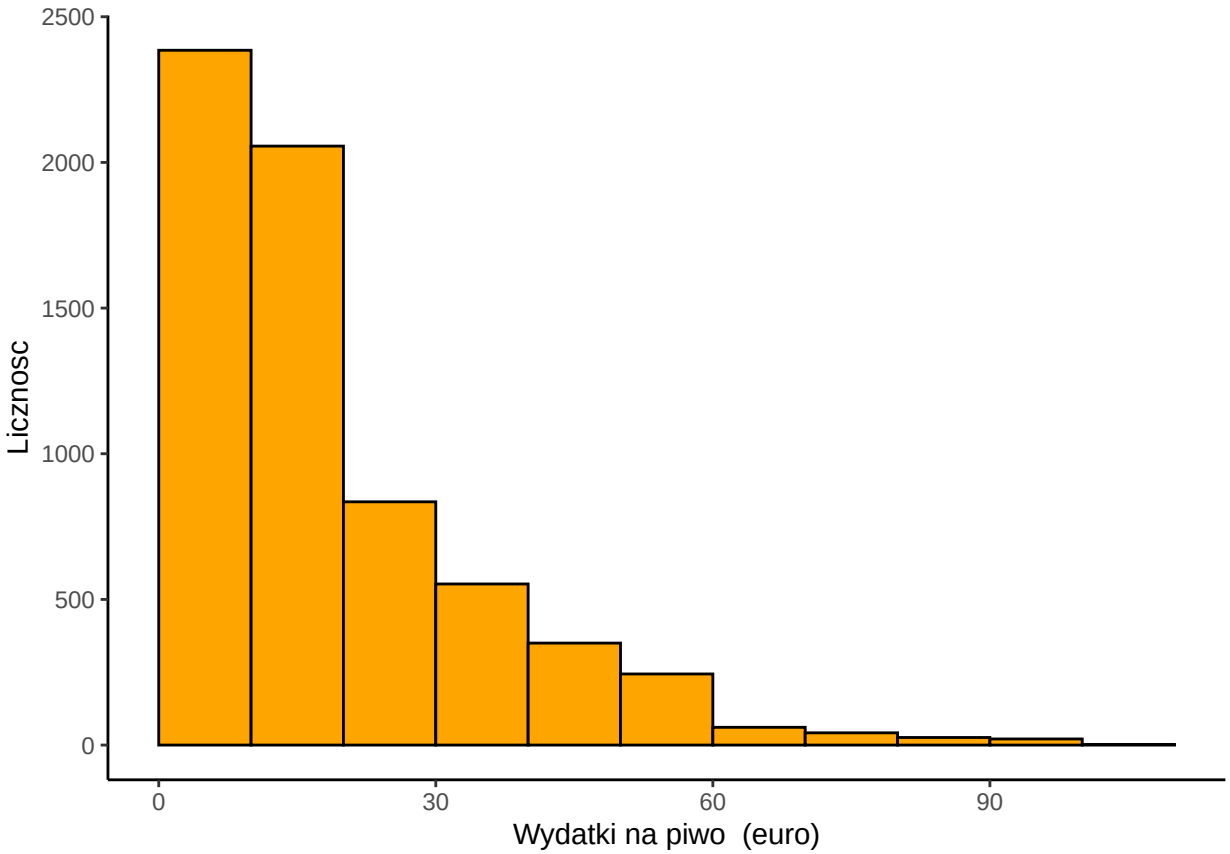
```
colSums(is.na(alkohol))
```

```
##      Miesiac       Region         Wiek         Plec     Edukacja         Wino
##            0            0            0          226            0            0
##         Piwo Inny_alkohol
##            0            0
```

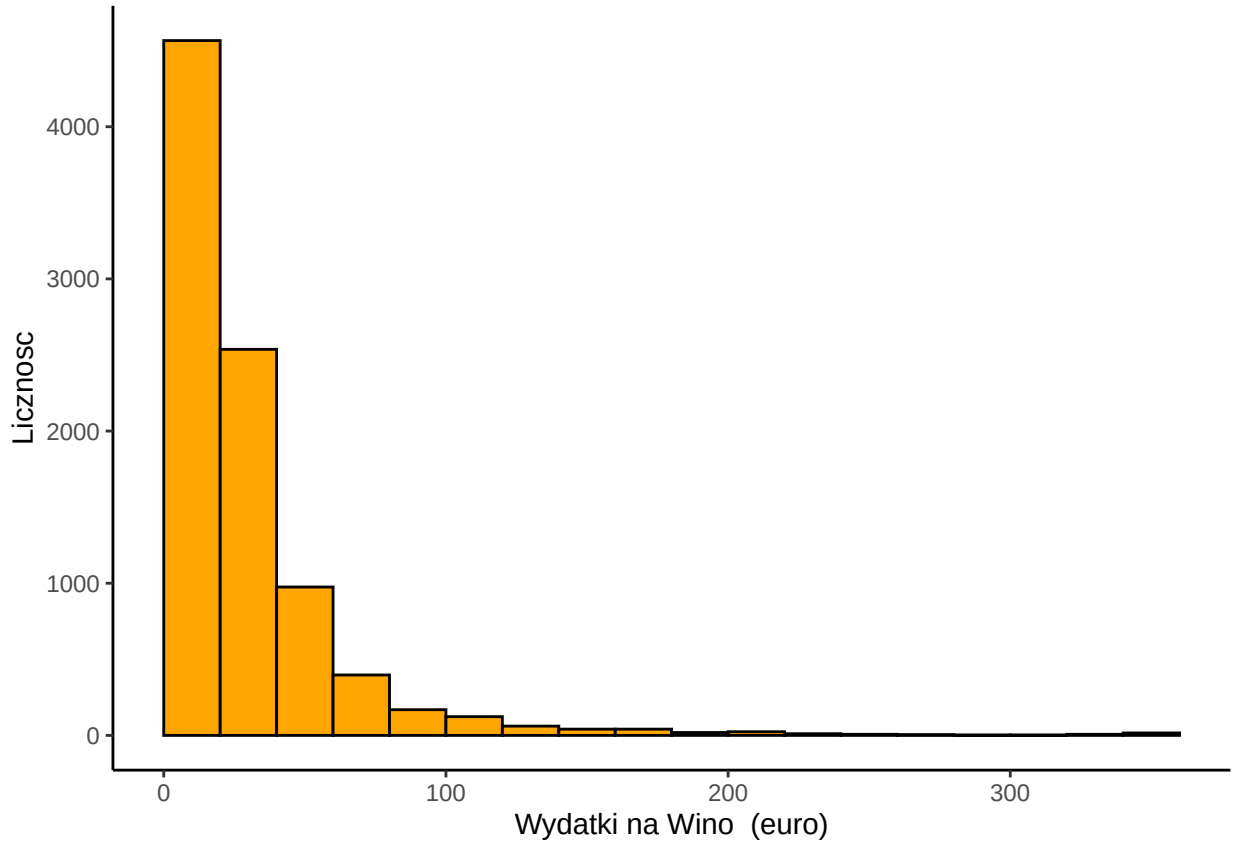
Ze wzgledu na niewielka liczbe brakow danych postanowilem usunac je z proby za pomoca funkcji na.omit().

Analiza danych

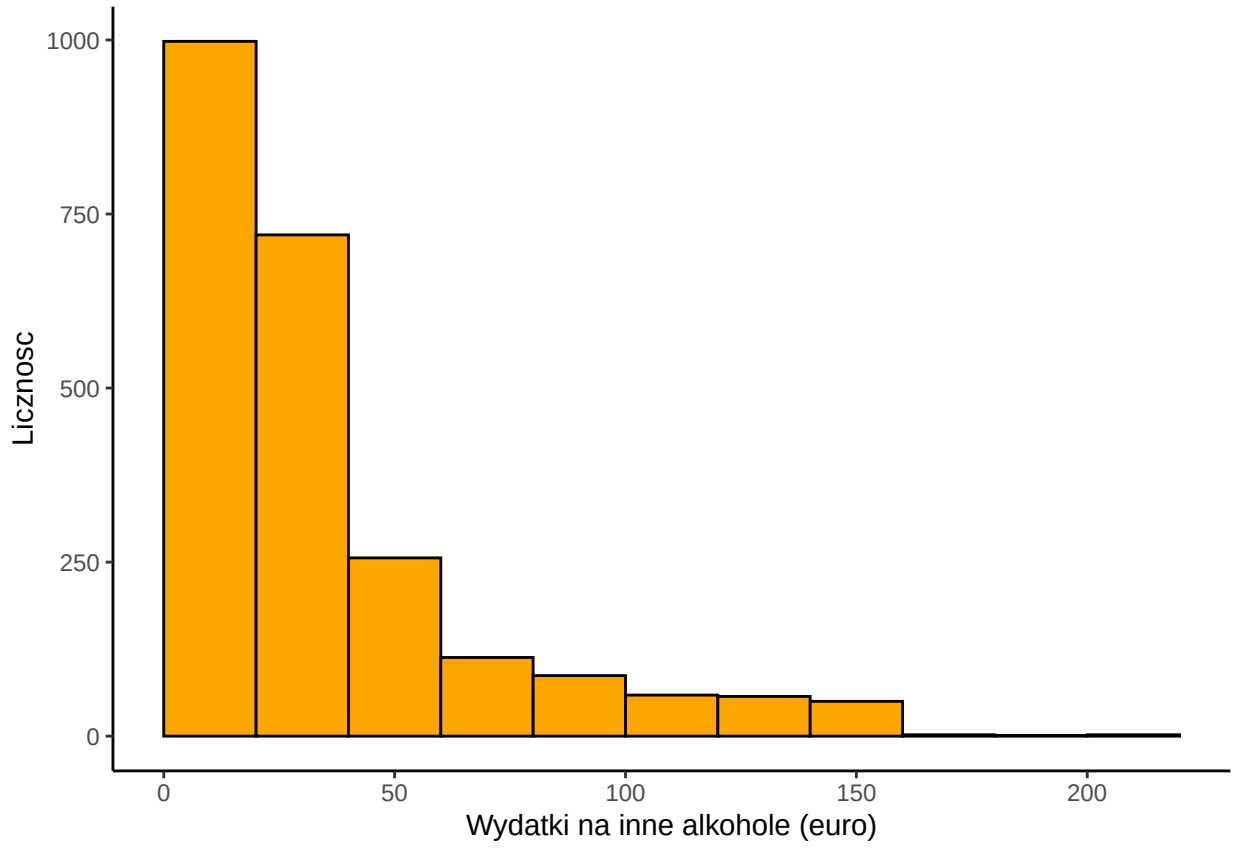
```
par(mfcol = c(2,2))
alkohol %>%
  filter(Piwo > 0) %>%
    ggplot()+
    geom_histogram(aes(Piwo), binwidth = 10, center = 5, color="black", fill="orange") +
    labs(x="Wydatki na piwo  (euro)", y = "Licznosc") +
    theme_classic()
```



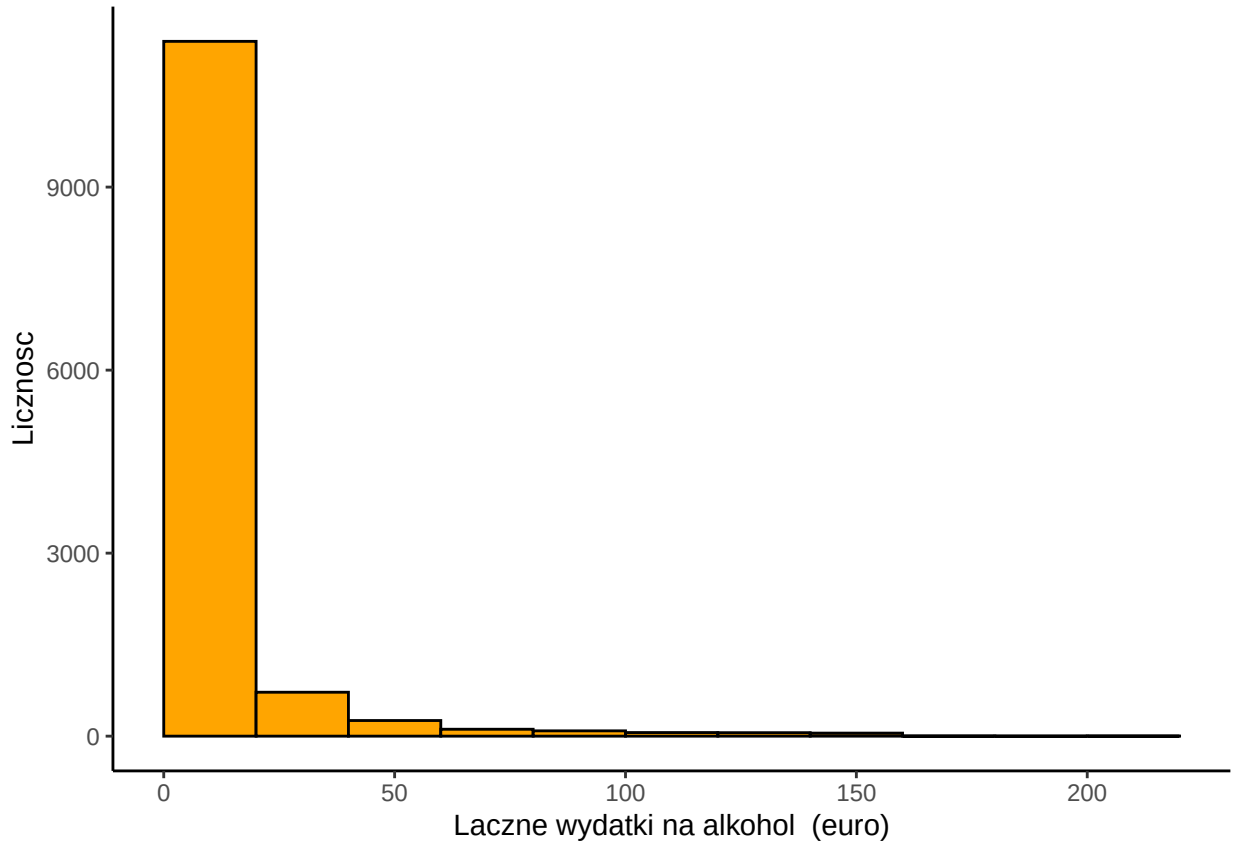
```
alkohol %>%
  filter(Wino> 0) %>%
    ggplot()+
    geom_histogram(aes(Wino), binwidth = 20, center = 10, color="black", fill="orange") +
    labs(x="Wydatki na Wino  (euro)", y = "Licznosc") +
    theme_classic()
```




```
alkohol %>%
  filter(Inny_alkohol > 0) %>%
    ggplot()+
    geom_histogram(aes(Inny_alkohol), binwidth = 20, center = 10, color="black",
                   fill="orange") +
    labs(x="Wydatki na inne alkohole (euro)", y = "Licznosc") +
    theme_classic()
```




```
alkohol %>%
  filter(Wydatki_alkohol > 0) %>%
    ggplot()+
    geom_histogram(aes(Inny_alkohol), binwidth = 20, center = 10, color="black",
                   fill="orange") +
    labs(x="Laczne wydatki na alkohol  (euro)", y = "Licznosc") +
    theme_classic()
```



Wedlug wlasnego poczucia 'estetyki' danych postanowilem za punkty graniczne odstajacych danych wybrac nastepujace kwantyle

```
quantile(alkohol$Wino, 0.995)
```

```
##  99.5%
## 169.45
```

```
quantile(alkohol$Piwo, 0.997)
```

```
##    99.7%
## 75.50065
```

```
quantile(alkohol$Inny_alkohol, 0.997)
```

```
##    99.7%
## 135.7331
```

```
quantile(alkohol$Wydatki_alkohol, 0.99)
```

```
##      99%
## 168.5459
```

Podstawy badania

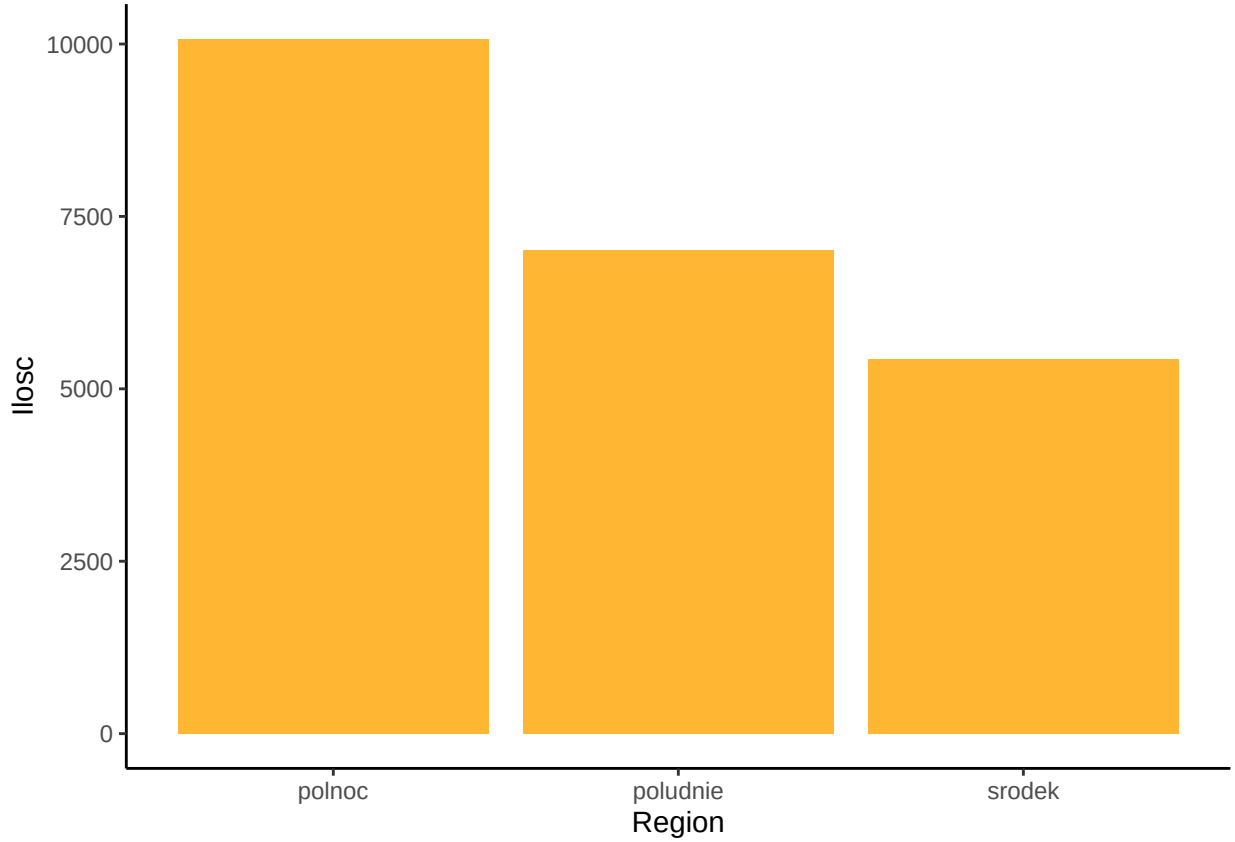
Dane zawieraja 22518 obserwacji - liczba przeprowadzonych ankiet.

Miesiac	Region	Wiek	Plec	Edukacja	Wino	Piwo	Inny_alkohol	Wydatki_alkohol	Kupuje
7	polnoc	dorosla	M	podstawowe	0.00	12.61	0	12.61	1
11	polnoc	dorosla	K	podstawowe	0.00	9.59	0	9.59	1
7	polnoc	dorosla	M	srednie	0.00	0.00	0	0.00	0
9	poludnie	dorosla	M	srednie	6.06	0.00	0	6.06	1
3	srodek	mloda_dorosla	K	podstawowe	0.00	0.00	0	0.00	0
6	poludnie	dorosla	M	podstawowe	0.00	0.00	0	0.00	0
1	srodek	mloda_dorosla	K	srednie	0.00	0.00	0	0.00	0
6	srodek	starsza	K	podstawowe	0.00	0.00	0	0.00	0
3	polnoc	dorosla	M	srednie	23.93	0.00	0	23.93	1
9	polnoc	dorosla	M	podstawowe	0.00	0.00	0	0.00	0

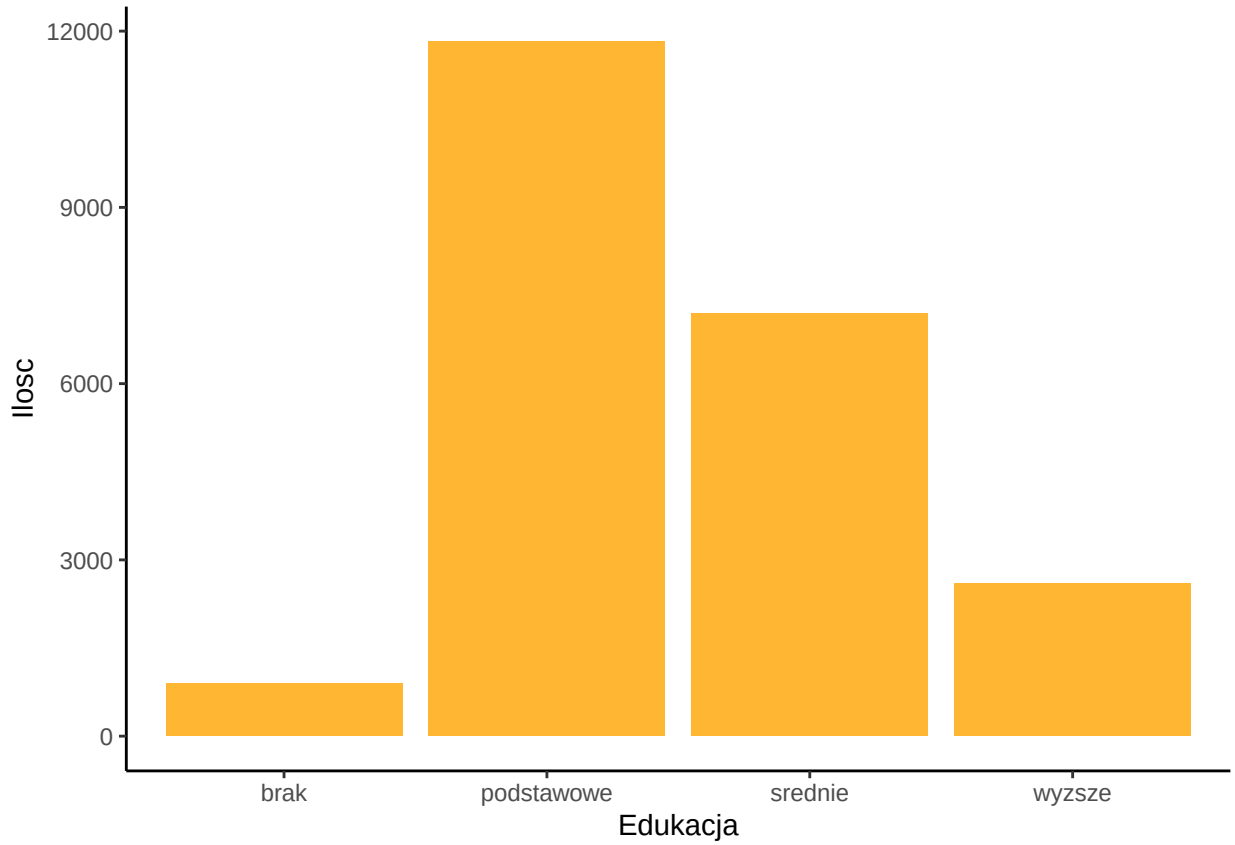
Statystyki opisowe zmiennych ilosciowych

```
##       Wino              Piwo          Inny_alkohol     Wydatki_alkohol
##  Min.   :  0.00   Min.   : 0.000   Min.   :  0.000   Min.   :  0.00
##  1st Qu.:  0.00   1st Qu.: 0.000   1st Qu.:  0.000   1st Qu.:  0.00
##  Median :  0.00   Median : 0.000   Median :  0.000   Median :  6.68
##  Mean   : 10.06   Mean   : 5.153   Mean   :  2.944   Mean   : 18.16
##  3rd Qu.: 13.66   3rd Qu.: 5.560   3rd Qu.:  0.000   3rd Qu.: 26.21
##  Max.   :164.12   Max.   :75.070   Max.   :135.600   Max.   :168.53
```

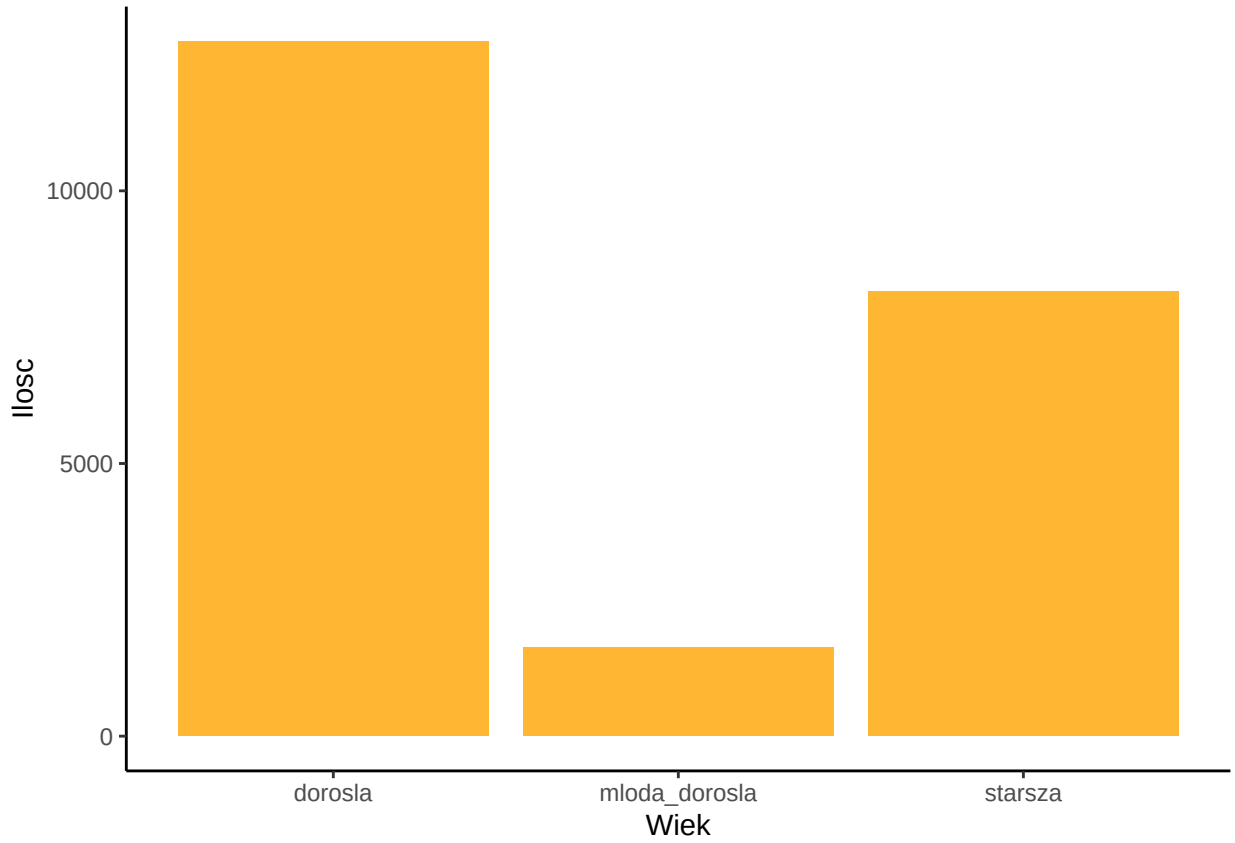
Jak charakteryzuje sie proba w badaniu?



Na podstawie przeprowadzonych operacji i utworzonego wykresu można zauważyć, że najwięcej osób, które wzięły udział w ankiecie pochodziło z północy Włoch.

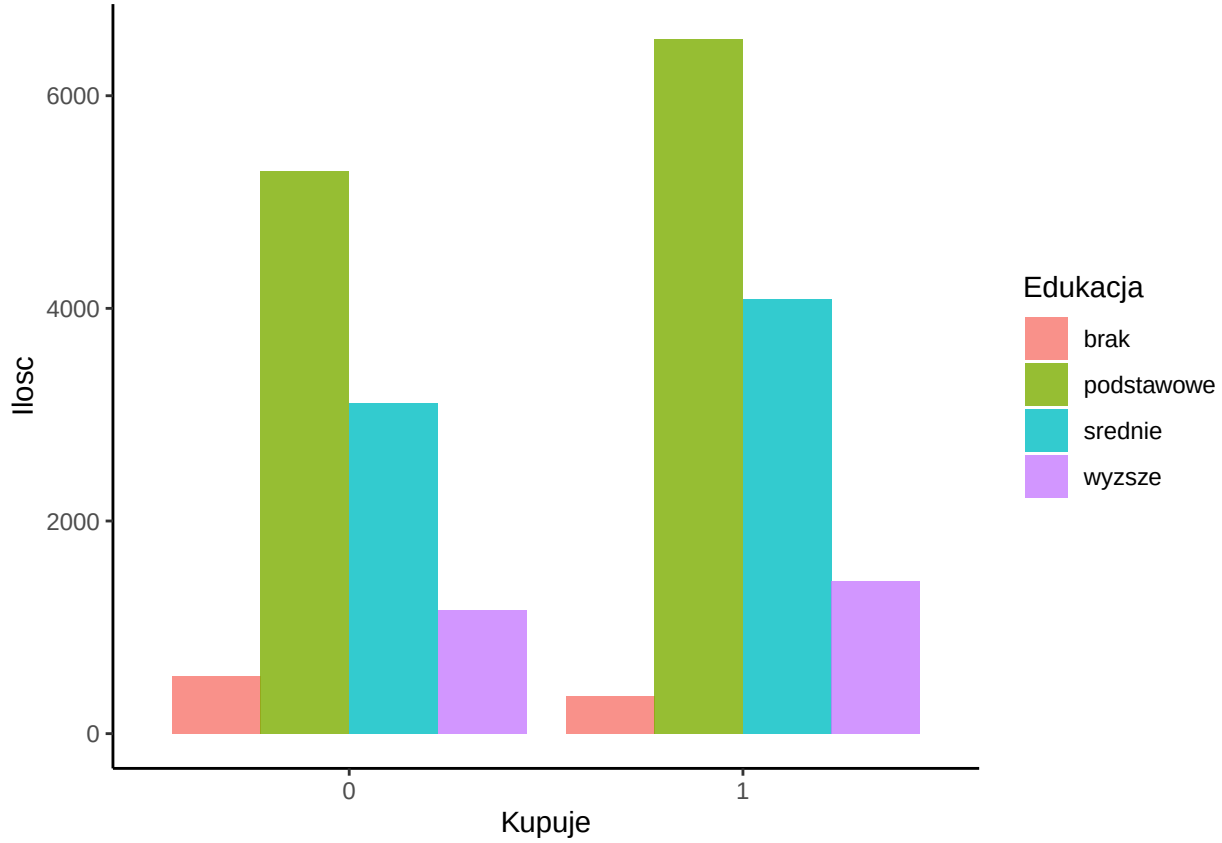


Analizując powyższy wykres możemy łatwo zauważyć, że najczęście w ankiecie brały udział osoby z wykształceniem podstawowym, a najrzadziej te, które nie mogą pochwalić się posiadaniem jakiegokolwiek wykształcenia.

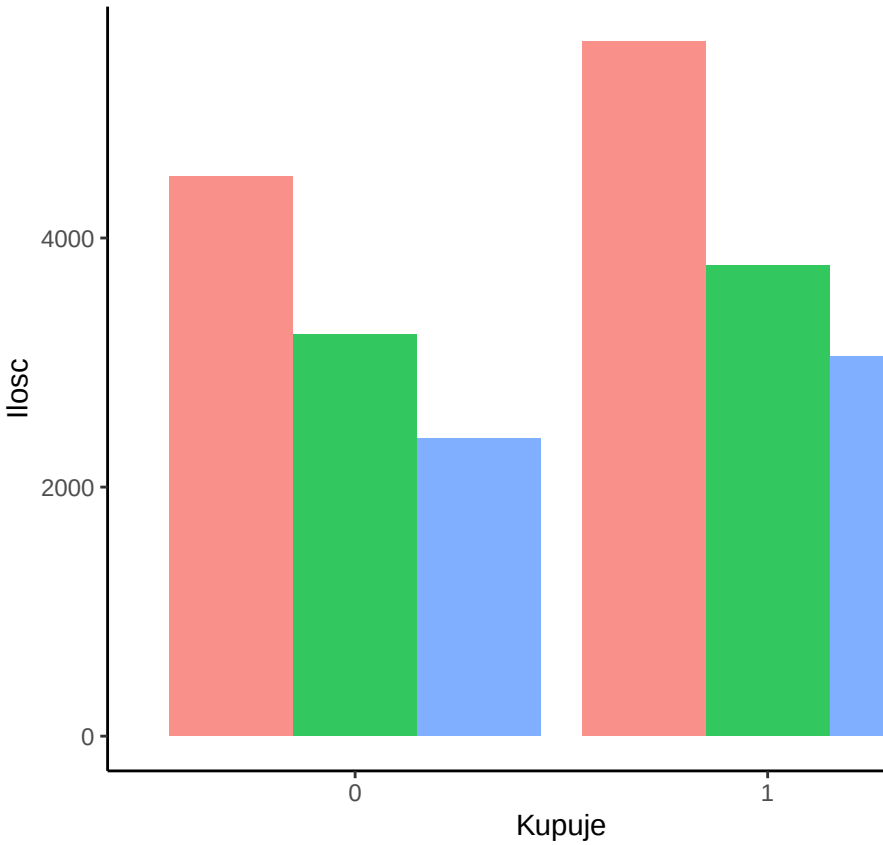


Jak widać na załączonym wykresie wiek odgrywał znaczną role w tym jak często osoby brały udział w ankiecie, w naszym przypadku najwięcej ankiet zostało wypełnionych przez osoby w wieku od 35 do 64 lat.

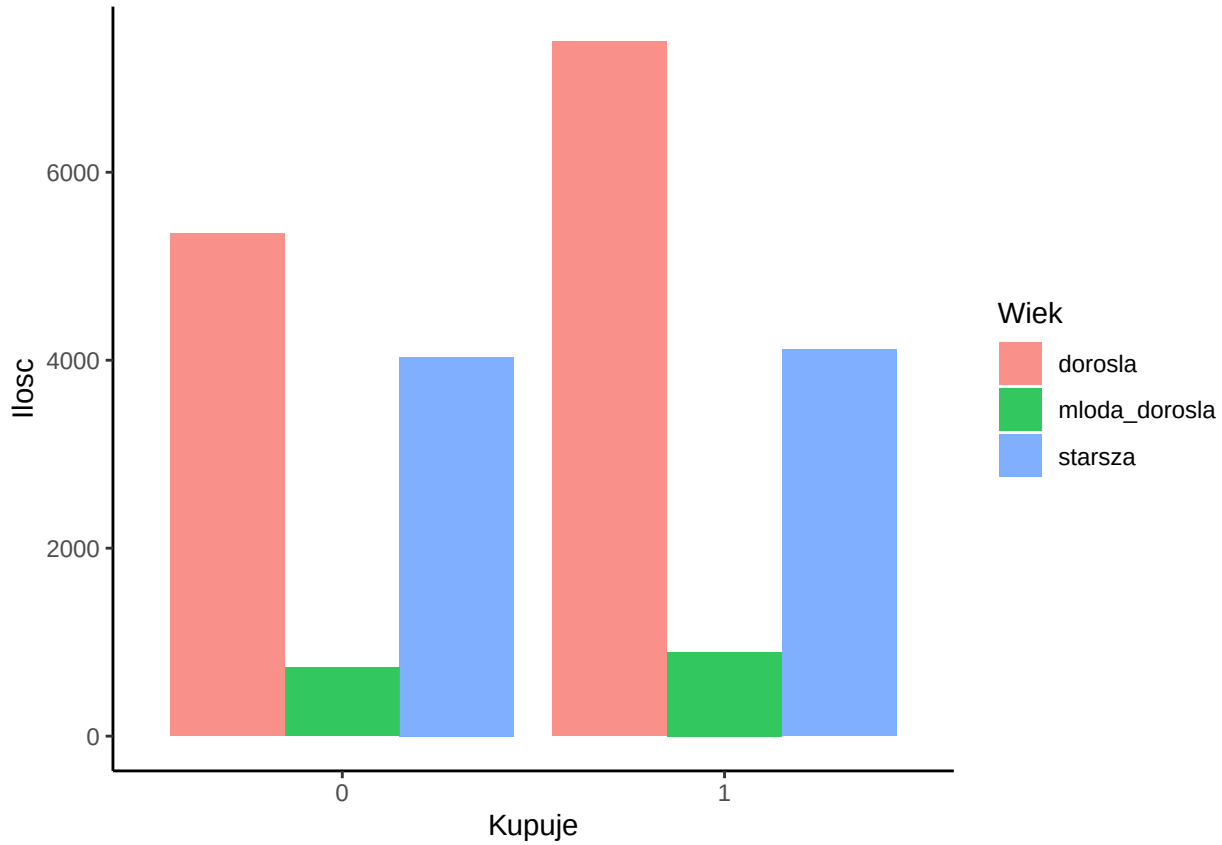
Jak duzo osob wydaje pieniadze na alkohol?



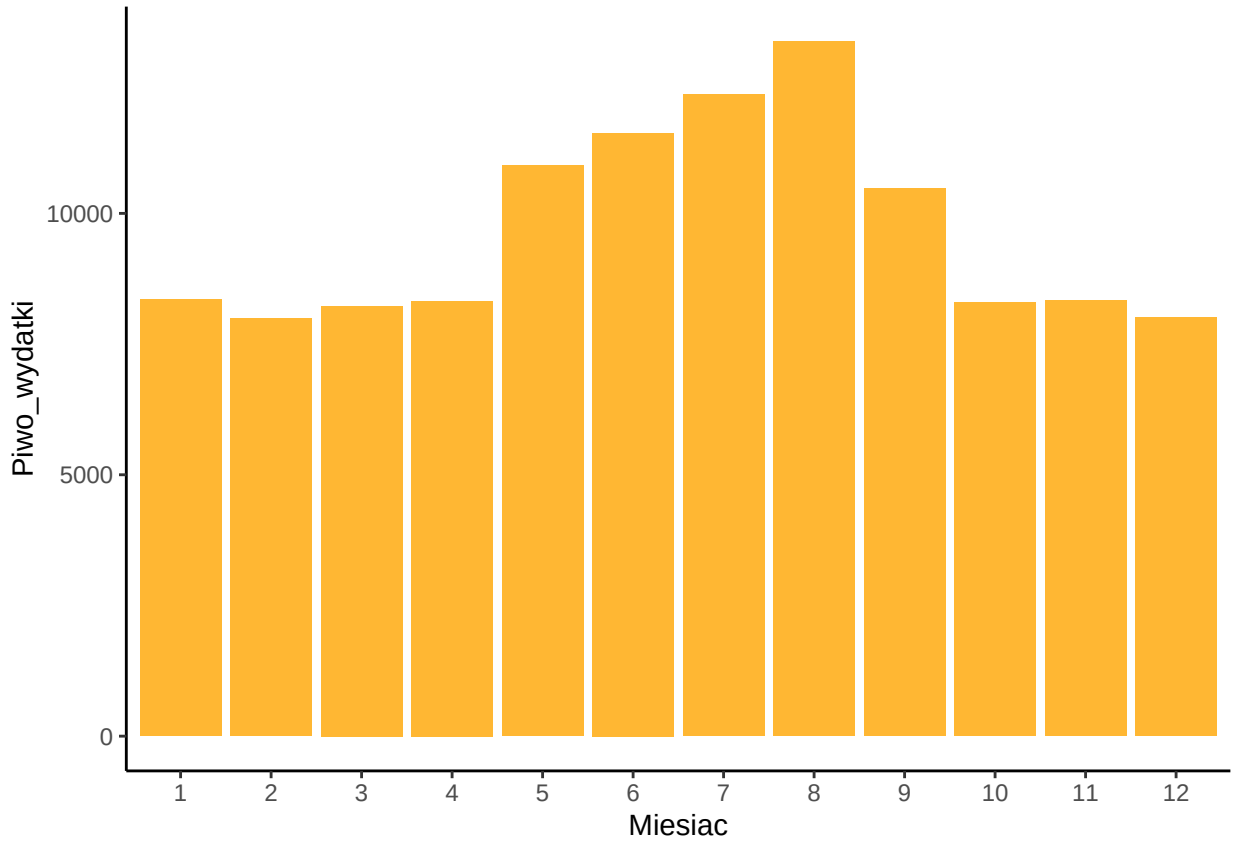
Z powyższego wykresu wynika, że najwięcej osób zarówno kupujących jak i nie kupujących alkoholu występuje



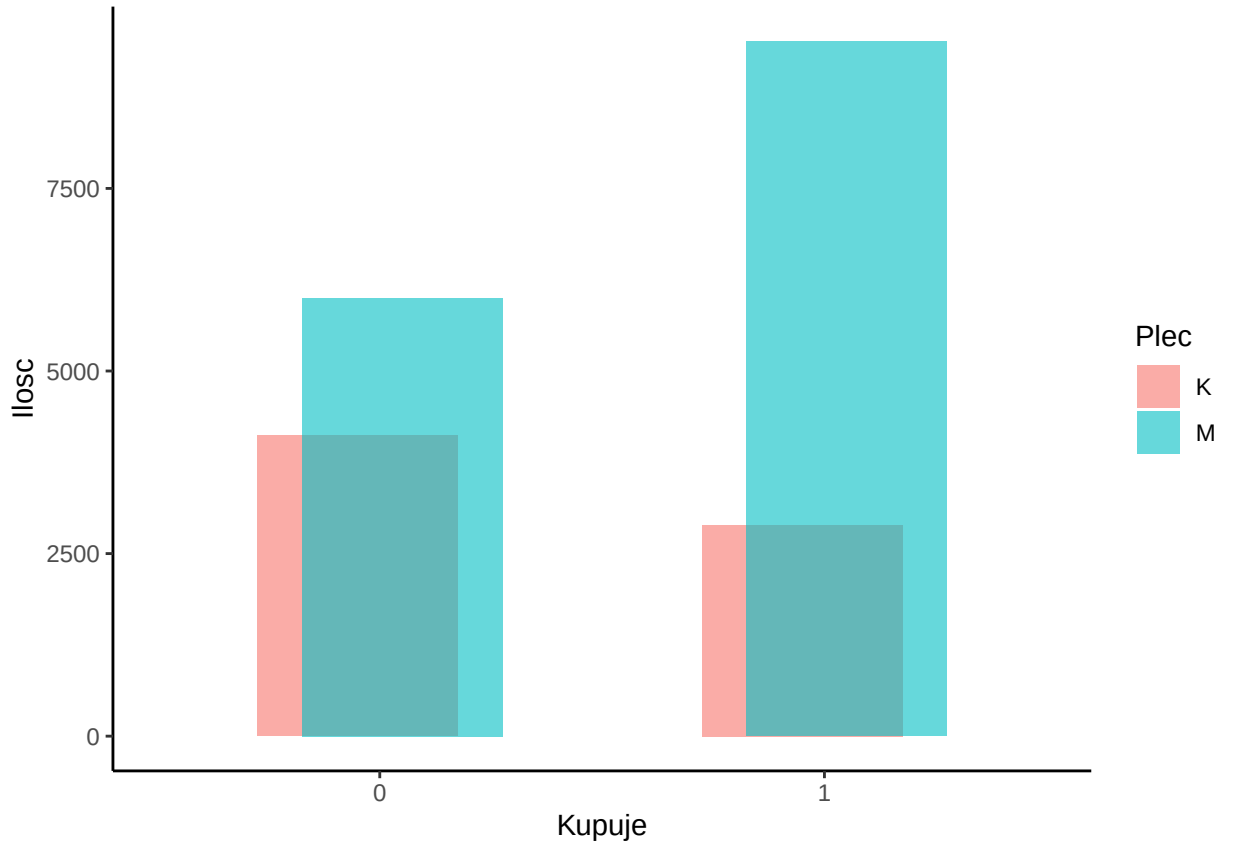
w grupie osób z wykształceniem podstawowym.
Analizując załączony wykres można zauważyć że najwięcej alkoholu kupują osoby mieszkające na północy Włoch, a najmniej te pochodzące z częsci środkowej.



Jak wynika z powyższego wykresu najwięcej abstynentów (osób nie kupujących alkoholu) oraz osób kupujących alkohol występuje w grupie osób dorosłych, czyli tych z przedziału 35 - 64 lata.

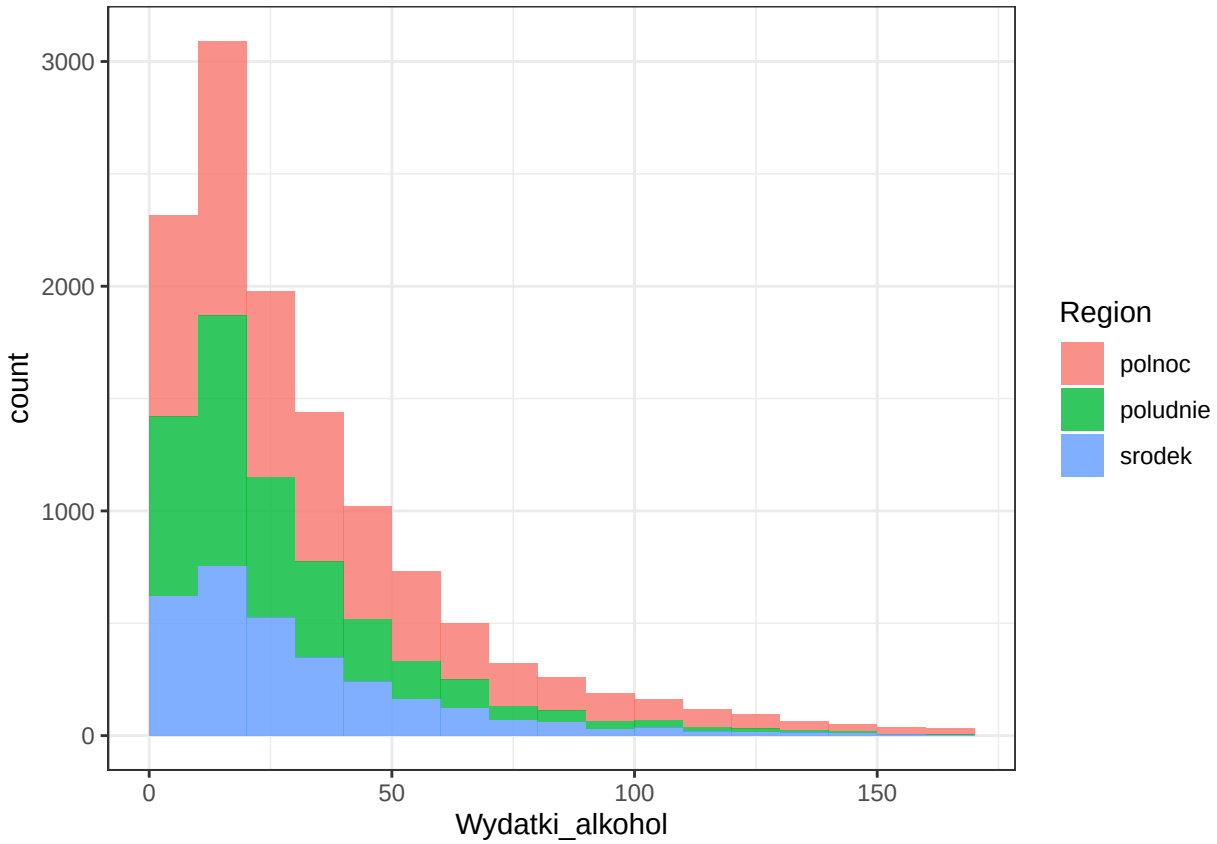


Na podstawie powyższego wykresu można potwierdzić nasza hipotezę, że mieszkańcy Włoch wydają średnio na piwo znacznie więcej w mieisącach letnich niż zimowych (najwięcej w sierpniu a najmniej w lutym i grudniu).

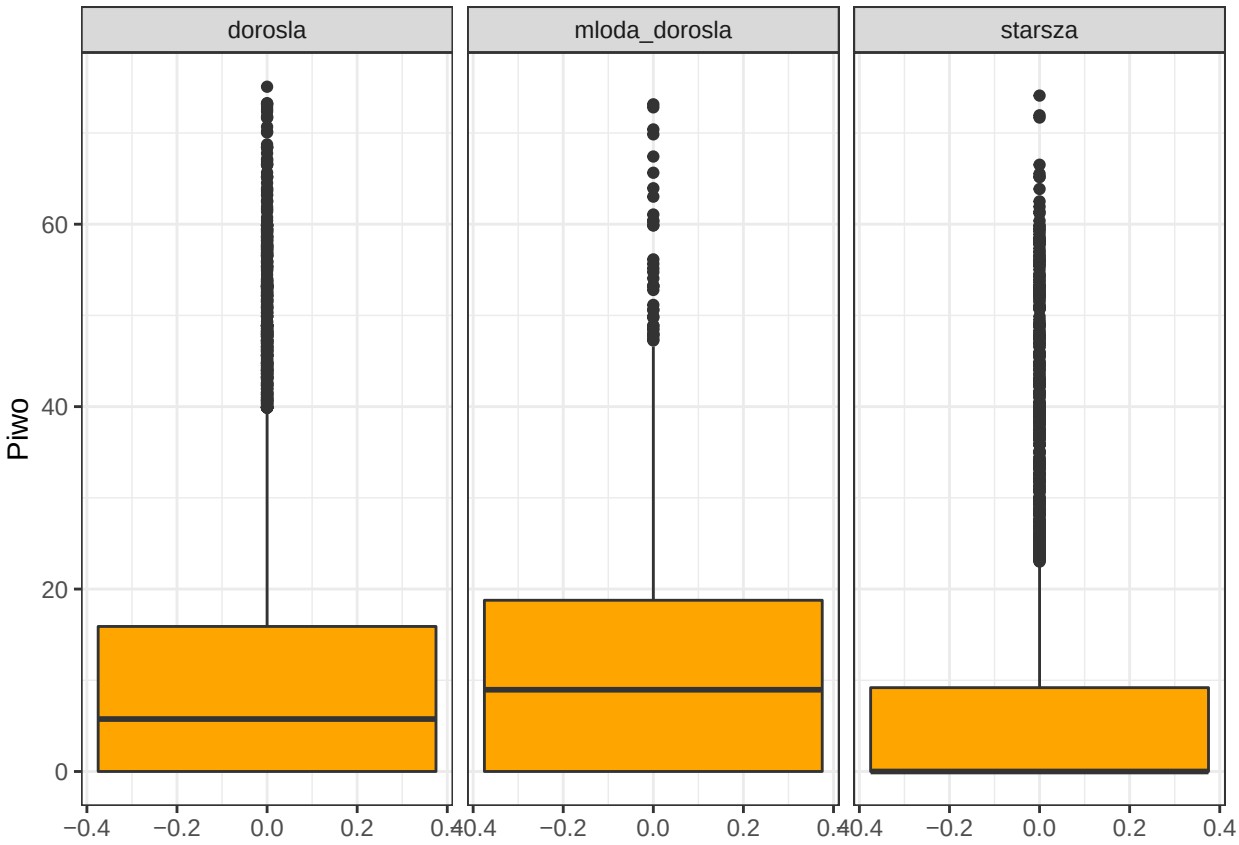


Analizując powyższy wykres można dojść do wniosku, że odsetek kobiet, które nie kupują alkoholu jest znacznie wyższy niż analogiczna statystyka dotycząca mężczyzn.

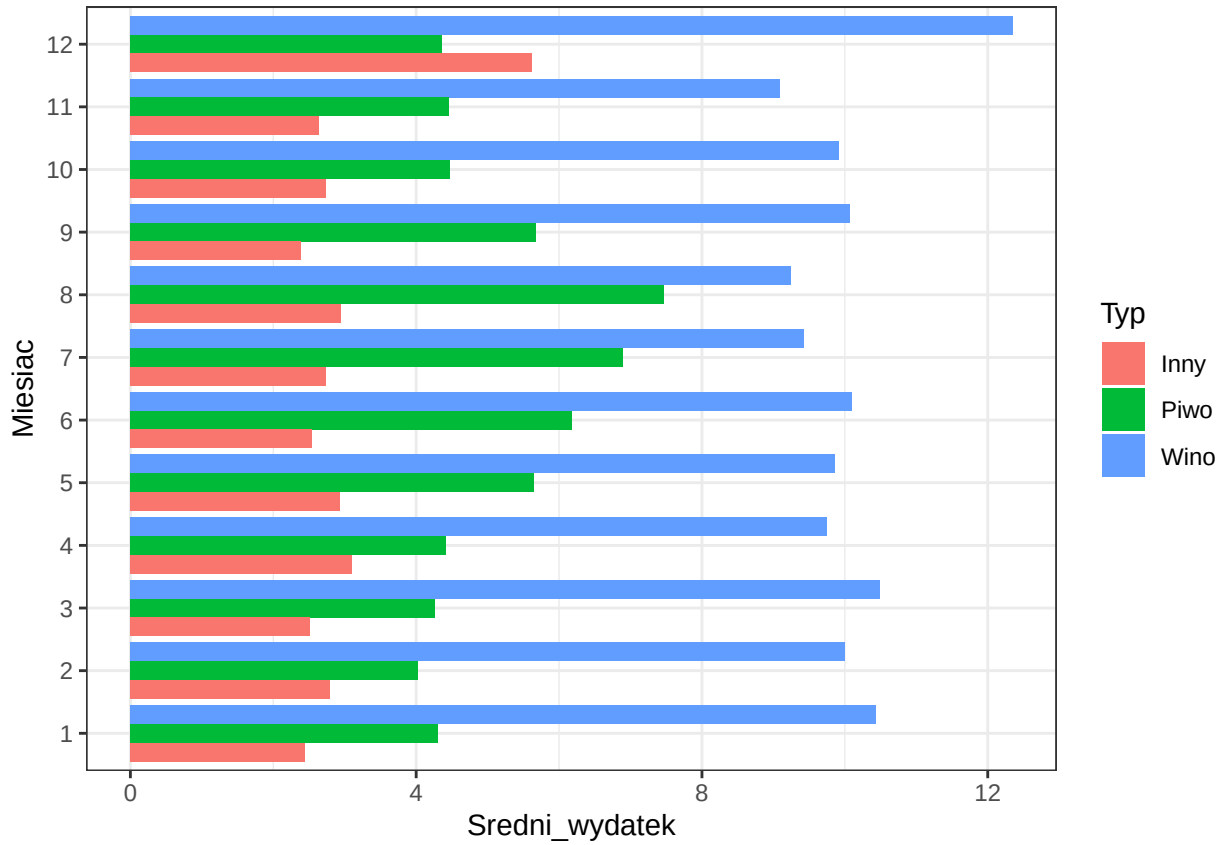
Jak wygladaja wydatki na alkohol w roznych grupach?



Na powyższym wykresie widać, że zdecydowana większości ankietowanych nie przeznacza na alkohol więcej niż 50 euro miesięcznie. Kolejną informacją, którą możemy uzyskać jest to że najwiecej osób, które deklarują wydatki na alkohol pochodzi z północy Włoch.



Na podstawie mediany wydatków na piwo osób w róznym wieku można zauważyć, że osoby w podeszłym wieku o wiele rzadziej decydują się na zakup piwa niż osoby w innych grupach wiekowych.



Jak widać na załączonym wykresie, w każdym z miesięcy, spośród analizowanych alkoholi, Włosi najchętniej kupują wino. Potwierdza to nasza hipotezę, że obywatele Włoch zdecydowanie najwięcej wydają na zakup wina.

Region	Suma_wydatkow	Liczba_ankietowanych	Wspolczynnik
polnoc	210677.68	10075	20.91094
poludnie	104612.70	7006	14.93187
srodek	93667.77	5437	17.22784

Z powyzszej tabelki mozemy odczytac ile wynosi przecietny wydatek na alkohol jednego mieszkanca wzgledem poszczegolnych regionow Wloch.

Regresja logistyczna

```
##
## Call:
## glm(formula = Kupuje ~ Region + Plec + Wiek + Edukacja, family = "binomial",
##     data = alkohol_czyste)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4375  -1.3106   0.9431   1.0021   1.5429
##
## Coefficients:
```

```
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.58798    0.08054  -7.300 2.87e-13 ***
## Regionpoludnie     -0.07388    0.03224  -2.291  0.02194 *
## Regionsrodek        0.04590    0.03462   1.326  0.18495
## PlecM               0.77607    0.02988  25.972  < 2e-16 ***
## Wiekmloda_dorosla  -0.03868    0.05407  -0.715  0.47431
## Wiekstarsza        -0.16583    0.03149  -5.265 1.40e-07 ***
## Edukacjapodstawowe  0.35932    0.07414   4.847 1.26e-06 ***
## Edukacjasrednie     0.34582    0.07804   4.431 9.36e-06 ***
## Edukacjawyzsze      0.32096    0.08428   3.808  0.00014 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 30982  on 22517  degrees of freedom
## Residual deviance: 30119  on 22509  degrees of freedom
## AIC: 30137
##
## Number of Fisher Scoring iterations: 4
```

Na podstawie regresji logistycznej mozemy wyroznic 5 zmiennych ktore maja bardzo istnotny wplyw na to czy ktos kupuje alkohol:

- Mezczyzni decyduja sie na zakup alkoholu 2.16 razy czesciej od kobiet.
- Osoby w podeszlym wieku kupuja 1.18 razy rzadziej alkohol niz osoby dorosle.
- Zarowno na poziomie edukacji podstawowej, sredniej jak i wyzszej osoby z kazdej tej grupy kupuja okolo 1.42 razy czesciej alkohol niz osoby bez wyksztalcenia.