# Zadania z kursu języka R

Wojciech Sobczuk

Wszystkie zadania będą bazowały na tych samych danych, z plików `dane_badane.csv` oraz `dane_kontrolne.csv`. Dane pochodzą z ankiet dotyczących reakcji dzieci na COVID-19. Przygotowaliśmy dwie osobne, choć podobne ankiety, dla rodziców dzieci z ASD (grupa badana) i dla dzieci neurotypowych (grupa kontrolna).

*Zad. 1*

Załaduj pliki `dane_badane.csv` oraz `dane_kontrolne.csv`. Ze zbioru danych wybierz zmienne opisujące wiek dziecka, płeć dziecka, typ szkoły do której uczęszcza dziecko oraz poziom lęku dziecka w poszczególnych okresach. Dokonaj pobieżnej analizy eksploracyjnej - upewnij się, że wszystkie zmienne mają odpowiedni typ (tzn. numeryczny, kategoryczny lub tekstowy), jeśli nie mają to je zmień (np. za pomocą funkcji as.factor). Na koniec, ogranicz zbiory danych względem wieku do przedziału [4;19] tak, aby powstał nowy data.frame na którym będziesz pracował/a przez resztę zadań. Dodatkowy punkt za użycie biblioteki dplyr.

**Uwaga** Użyj funkcji `read.csv` z parameterm `encoding = "UTF-8"`, nazwa zmiennej dla płci do przekopiowania - X.U.FEFF.Plec.dziecka

```
d.bad <- read.csv("dane_badane.csv", dec = ".", sep = ",",encoding="UTF-8")
d.kon <- read.csv("dane_kontrolne.csv", dec = ".", sep = ",",encoding="UTF-8")

d.bad = d.bad %>% select(X.U.FEFF.Plec.dziecka, Wiek.dziecka,
                         Czy.dziecko.uczeszcza.do.szkoly.publicznej.,
                         Jak.w.skali.1.5.mozna.ocenic.poziom.leku.dziecka.przed.rozpoczeciem.epidemii.,
                         Jak.oceniasz.poziom.leku.dziecka.w.trakcie.tygodnia.przez.zamknieciem.szkól.,
                         Jaki.jest.poziom.leku.dziecka.obecnie..w.trakcie.trwania.kwarantanny.)

d.kon = d.kon %>% select(X.U.FEFF.Plec.dziecka, Wiek.dziecka,
                         Czy.dziecko.uczeszcza.do.szkoly.publicznej.,
                         Jak.w.skali.1.5.mozna.ocenic.poziom.leku.dziecka.przed.rozpoczeciem.epidemii.,
                         Jaki.jest.poziom.leku.dziecka.obecnie..w.trakcie.trwania.kwarantanny.)

colnames(d.bad) = c("Plec", "Wiek", "Publiczne", "Lek.ok.1", "Lek.ok.2","Lek.ok.3")
colnames(d.kon) = c("Plec", "Wiek", "Publiczne", "Lek.ok.1", "Lek.ok.2","Lek.ok.3")

str(d.bad)
```

```
## 'data.frame':    294 obs. of  6 variables:
##  $ Plec     : chr  "M" "M" "Nie wiem" "K" ...
##  $ Wiek     : chr  "13" "13" "placzliwosc" "8" ...
##  $ Publiczne: chr  "Tak" "Nie" "Nie" "Nie" ...
##  $ Lek.ok.1 : int  1 NA NA 2 3 3 1 2 4 1 ...
##  $ Lek.ok.2 : int  4 NA NA 2 3 3 1 2 4 1 ...
##  $ Lek.ok.3 : int  2 NA NA 2 3 3 1 2 3 1 ...
```

```
str(d.kon)
```

```
## 'data.frame':    372 obs. of  6 variables:
##  $ Plec     : chr  "M" "K" "M" "K" ...
```

```
##  $ Wiek     : int  11 4 13 6 6 8 4 9 6 4 ...
##  $ Publiczne: chr  "Tak" "Nie" "Tak" "Nie" ...
##  $ Lek.ok.1 : int  1 1 2 3 1 1 1 1 3 2 ...
##  $ Lek.ok.2 : int  1 1 2 3 1 1 1 1 3 2 ...
##  $ Lek.ok.3 : int  1 2 2 2 1 1 1 1 3 2 ...
```

```
d.bad = d.bad %>% mutate(Plec = as.factor(Plec), Wiek = as.numeric(Wiek),
                         Publiczne = as.factor(Publiczne)) %>%
  filter(Wiek>=4 & Wiek <=19)
```

```
## Warning: Problem with `mutate()` input `Wiek`.
## i NAs introduced by coercion
## i Input `Wiek` is `as.numeric(Wiek)`.
```

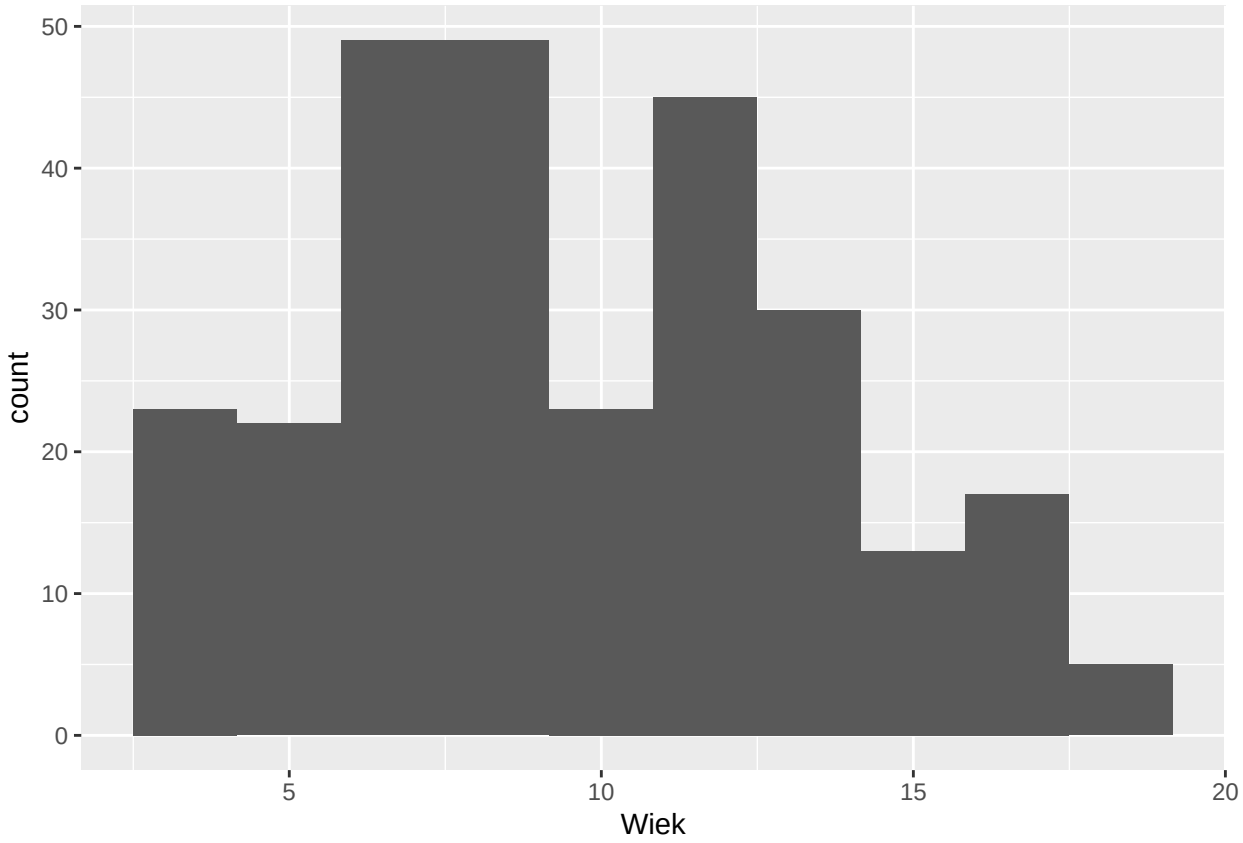
```
## Warning in mask$eval_all_mutate(dots[[i]]): NAs introduced by coercion
```

```
d.kon = d.kon %>% mutate(Plec = as.factor(Plec),
                         Publiczne = as.factor(Publiczne)) %>%
  filter(Wiek>=4 & Wiek <=19)
```

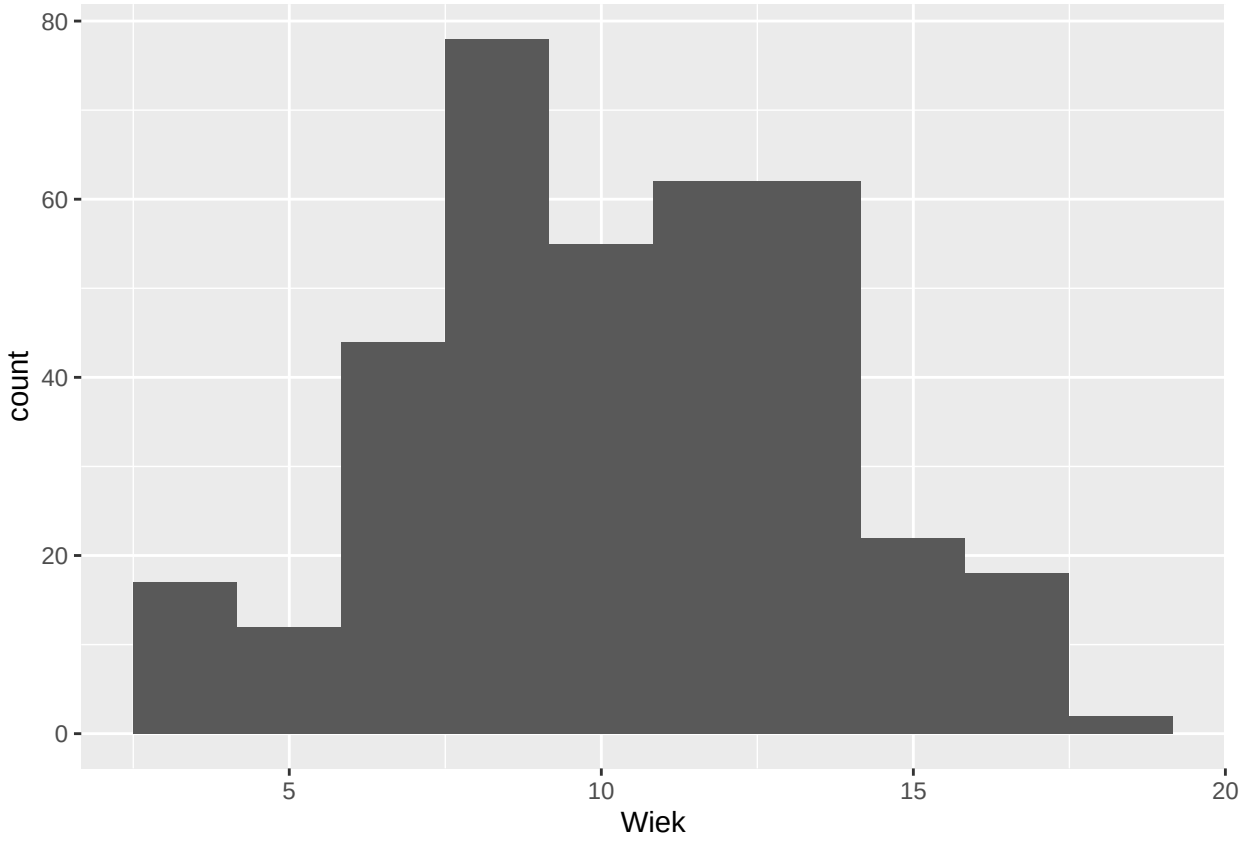
*Zad. 2*

Narysuj histogram rozkładu wieku dziecka osobno dla grupy badanej oraz dla grupy kontrolnej. Oblicz i wyświetl podstawowe miary statystyczne dla wieku (średnia, odch. std., min, max). Sprawdź testem statystycznym czy średnie wieku dziecka są równe między grupami. (dodatkowy punkt za użycie biblioteki ggplot2)

```
ggplot(d.bad, aes(x = Wiek)) + geom_histogram(bins = 10)
```



```
ggplot(d.kon, aes(x = Wiek)) + geom_histogram(bins = 10)
```




```
summary(d.bad$Wiek)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00    7.00    9.00    9.62   12.00   19.00
```

```
summary(d.kon$Wiek)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00    8.00   10.00   10.31   13.00   19.00
```

```
shapiro.test(d.bad$Wiek)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  d.bad$Wiek
## W = 0.9596, p-value = 5.87e-07
```

```
wilcox.test(d.bad$Wiek, d.kon$Wiek)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  d.bad$Wiek and d.kon$Wiek
## W = 44775, p-value = 0.005207
## alternative hypothesis: true location shift is not equal to 0
```

*Zad. 3*

*a* Stwórz tabele zliczające liczbę chłopców i dziewczynek w obu grupach osobno, oblicz i wyświetl odsetek chłopców oraz sprawdź testem statystycznym (t-Studenta), czy te odsetki są równe między grupami. (pomiń sprawdzanie założeń testu statystycznego)

*b* Stwórz tabelę zliczającą liczbę dzieci uczęszczających do szkół publicznych i niepublicznych w obu grupach. Wylicz odsetek dzieci uczęszczających do szkół publicznych i sprawdź testem statystycznym (t-Studenta), czy te odsetki są równe między grupami. (pomiń sprawdzanie założeń testu statystycznego)

```
table(d.bad$Plec)
```

```
##
##        K        M Nie wiem
##       54      222        0
```

```
table(d.kon$Plec)
```

```
##
##   K   M
## 177 195
```

```
d.bad = d.bad %>% mutate(Zad.3a = ifelse(d.bad$Plec == "M",1,0),
                         Zad.3b = ifelse(d.bad$Publiczne == "Tak",1,0))
d.kon = d.kon %>% mutate(Zad.3a = ifelse(d.kon$Plec == "M",1,0),
                         Zad.3b = ifelse(d.kon$Publiczne == "Tak",1,0))

var.test(d.bad$Zad.3a,d.kon$Zad.3a)
```

```
##
##  F test to compare two variances
##
## data:  d.bad$Zad.3a and d.kon$Zad.3a
## F = 0.63156, num df = 275, denom df = 371, p-value = 6.007e-05
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5072808 0.7891795
## sample estimates:
## ratio of variances
##           0.631559
```

```
t.test(d.bad$Zad.3a,d.kon$Zad.3a,var.equal = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  d.bad$Zad.3a and d.kon$Zad.3a
## t = 7.9413, df = 642.94, p-value = 8.953e-15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2108802 0.3494284
## sample estimates:
## mean of x mean of y
## 0.8043478 0.5241935
```

```
var.test(d.bad$Zad.3b,d.kon$Zad.3b)
```

```
##
##  F test to compare two variances
```

```
##
## data:  d.bad$Zad.3b and d.kon$Zad.3b
## F = 2.4642, num df = 275, denom df = 371, p-value = 8.882e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.979308 3.079220
## sample estimates:
## ratio of variances
##           2.464217
```

```
t.test(d.bad$Zad.3b,d.kon$Zad.3b,var.equal = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  d.bad$Zad.3b and d.kon$Zad.3b
## t = -7.7877, df = 436.21, p-value = 5.019e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3224614 -0.1924988
## sample estimates:
## mean of x mean of y
## 0.6376812 0.8951613
```

*Zad. 4*

*a* Stwórz data.frame (jeden dla obu grup), w którym będą informacje o poziomie lęku w każdym z okresów, grupie z której pochodzi dziecko. Na jego podstawie oblicz średni poziom lęku w każdym z okresów w obu grupach.

*b* Zwizualizuj średnie z poprzedniego podpunktu na wykresie słupkowym łączonym (tj. dla okresu I ma być słupek dla grupy badawczej i kontrolnej obok siebie, w różnych kolorach, powtórzyć dla każdego okresu).

*c* Sprawdź testem statystycznym (t-Studenta) czy te poziomy są sobie równe między grupami dla każdego z okresów osobno. (pomiń sprawdzanie założeń testu statystycznego)

```
wyk.df = as.data.frame(d.bad$Lek.ok.1)
colnames(wyk.df) = c("Okres.1")
wyk.df = wyk.df %>% mutate(Grupa = "bad", Okres.2 = d.bad$Lek.ok.2, Okres.3 = d.bad$Lek.ok.3)

temp = as.data.frame(d.kon$Lek.ok.1)
colnames(temp) = c("Okres.1")
temp = temp %>% mutate(Grupa = "kon", Okres.2 = d.kon$Lek.ok.2, Okres.3 = d.kon$Lek.ok.3)

poz.oba = rbind(wyk.df,temp)

wyk.lek = poz.oba %>% select(Grupa, Okres.1, Okres.2, Okres.3) %>% group_by(Grupa) %>%
  summarise(mu.1 = mean(Okres.1, na.rm = T),
              mu.2 = mean(Okres.2, na.rm = T), mu.3 = mean(Okres.3, na.rm = T))
```

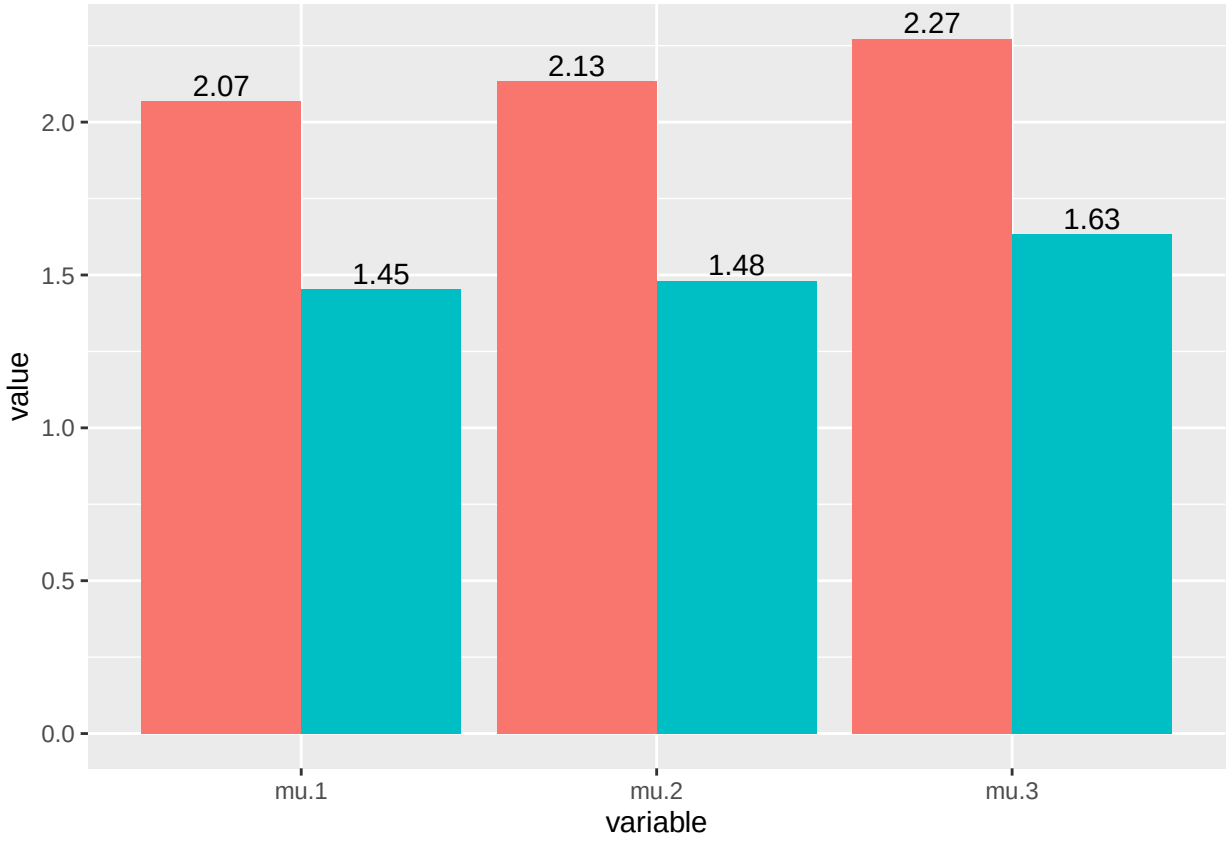
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
wyk.lek = melt(wyk.lek)
```

```
## Using Grupa as id variables
```

```
ggplot(wyk.lek, aes(x=variable, y=value, fill = Grupa)) +
  geom_bar(stat='identity', position='dodge') +
```

```
  theme(legend.position = "none") +
  geom_text(aes(label=round(value,2)), position=position_dodge(width=0.9), vjust=-0.25)
```




*Zad. 5*

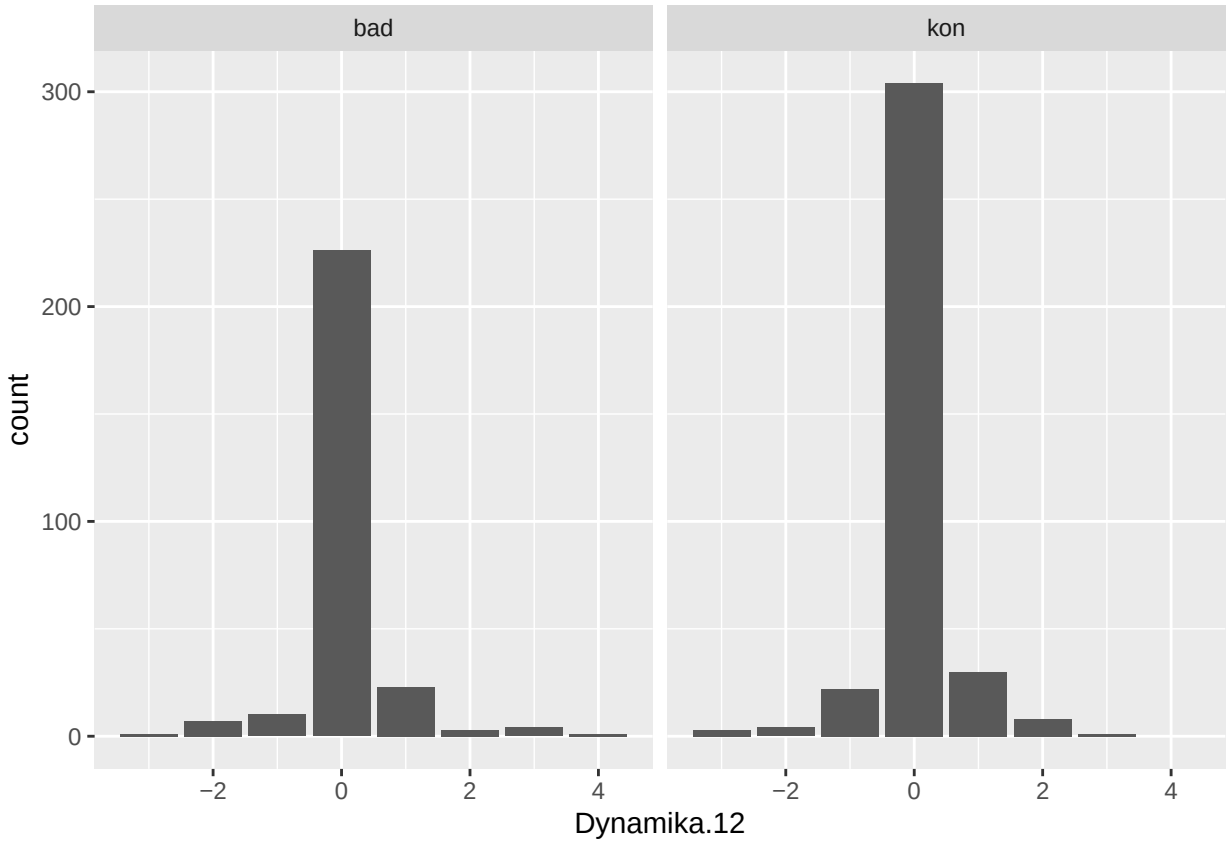
Do data.frame stworzonego w poprzednim zadaniu dodaj zmienne, które będą odpowiadały zmianie poziomu lęku między okresami 1-3, 2-3 i 1-2. Zmianę poziomu lęku nazwiemy dalej dynamikami lęku (tj. zmiana 1-3 będzie dynamiką 1-3) Wybierz jedną z dynamik i narysuj histogram rozkładu wartości dynamiki dla obu grup na dwóch wykresach koło siebie (tzn. wykres dla grupy badanej po jednej stronie, wykres dla grupy kontrolnej po drugiej stronie).

```
poz.oba = poz.oba %>% mutate(Dynamika.12 = Okres.2 - Okres.1,
                             Dynamika.13 = Okres.3 - Okres.1,
                             Dynamika.23 = Okres.3 - Okres.2)

wyk2.df = poz.oba %>% select(Grupa, Dynamika.12)

ggplot(wyk2.df, aes(x = Dynamika.12)) + facet_wrap(~Grupa) + geom_bar()
```

```
## Warning: Removed 1 rows containing non-finite values (stat_count).
```


bad
kon
300
200
100
0
count
−2
0
2
4
Dynamika.12