

2017-03-16

*P.Krzysztoporska, M.Koralewski, B.Pawliczak*

*16 marca 2017*

## Typy prezentacji danych

Podstawowe 4 typy prezentacji danych:

- Porównanie
- Kompozycja
- Rozkład
- Relacje między danymi

## Pytania, które warto sobie zadać:

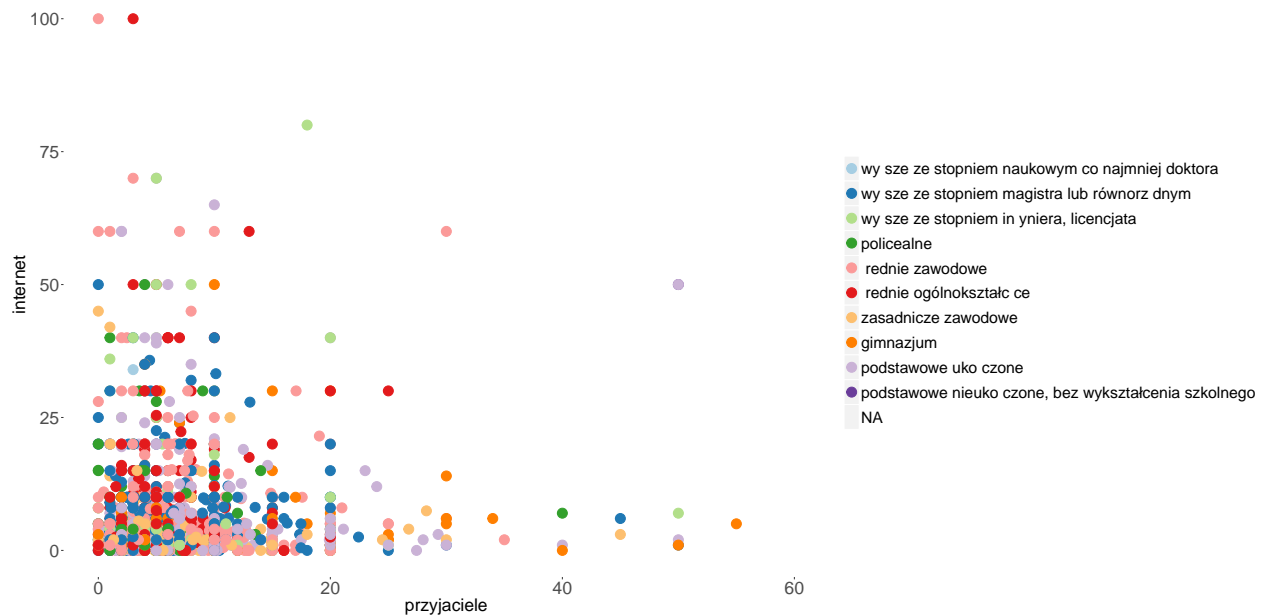
- Jak dużo zmiennych chcemy pokazać na jednym wykresie?
- Jak wiele punktów danych pokażemy?
- Czy dane są zmienne w czasie? A może zmieniają się w zależności od grup?

## Czy zawsze potrzebujemy wykresu?

**Tabela** dobrze prezentuje dane, gdy - chcemy znaleźć i porównać pojedyncze wartości, - pokazujemy dokładne wartości, - mamy do czynienia z różnymi jednostkami, - pokazujemy ilościowe informacje, a nie trend.

**Wykres** będzie odpowiedni, gdy - będzie przekazywał wiadomość poprzez kształt, - chcemy pokazać relacje między wieloma zmiennymi.

## Wykres punktowy (scatter plot)



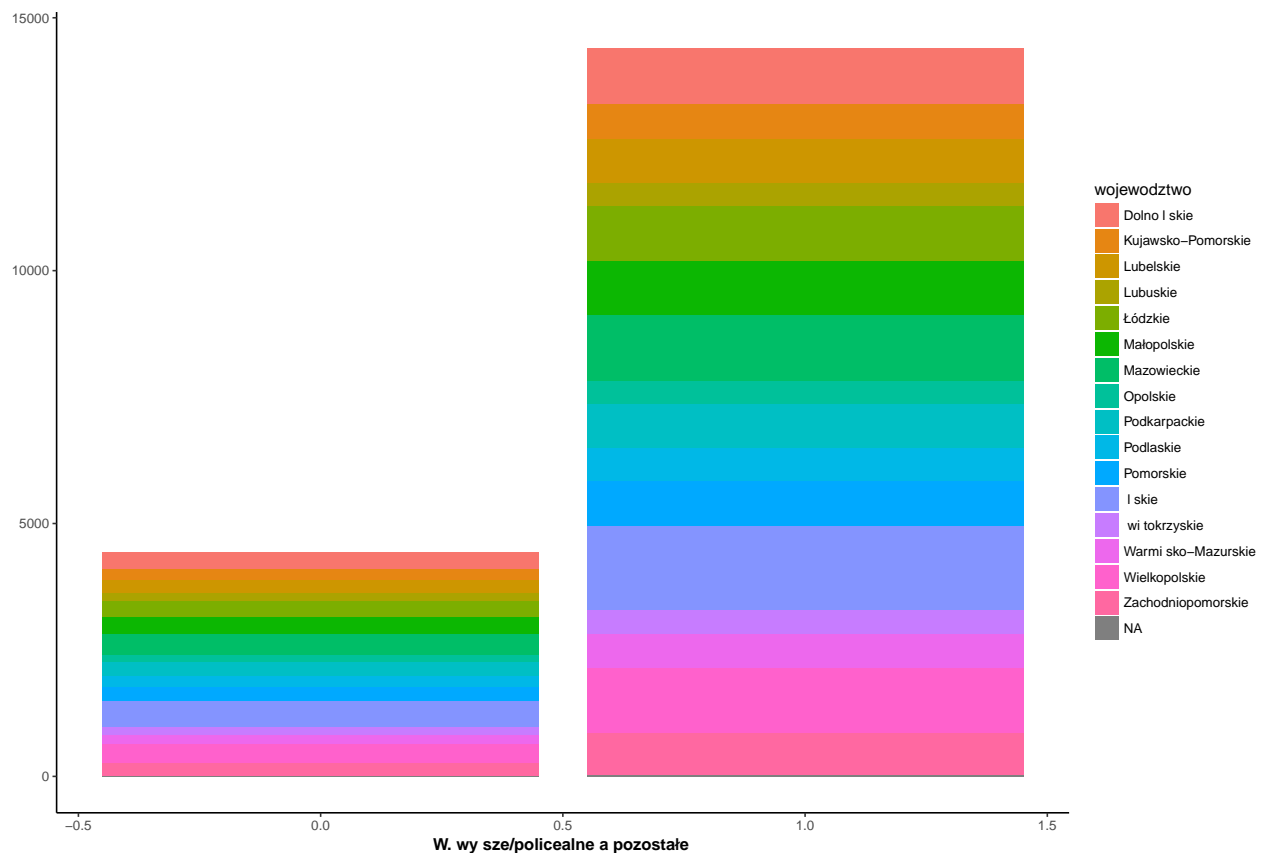
## Wykres kołowy (pie chart)

## Wykres słupkowy (bar plot)

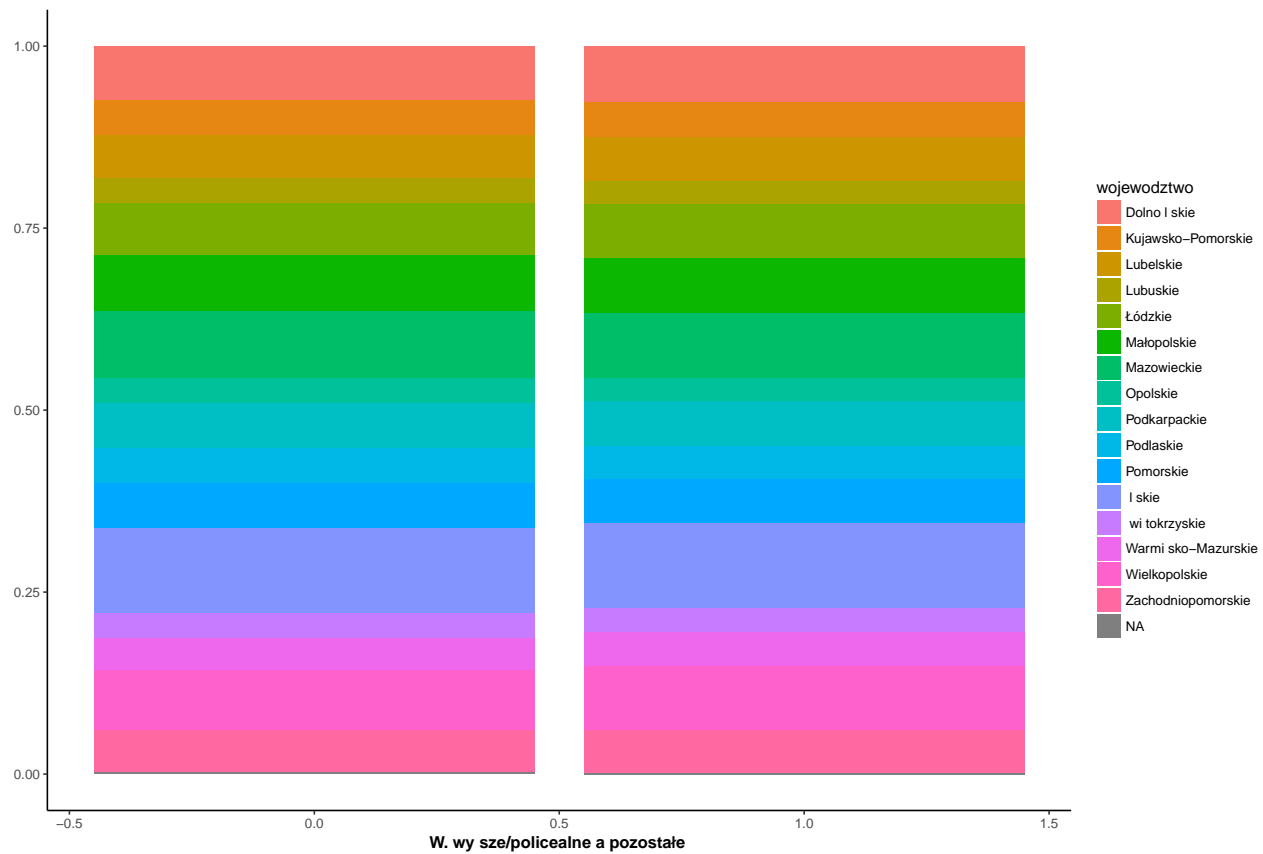
- w pozycji horyzontalnej najlepiej obrazuje 5,7 zmiennych,
- jeśli zmienną jest czas, umieszczamy ją na osi x,
- osie startują z zera!

## Wykres słupkowy (bar plot)

```
daneNaSzybko <- read.csv("edu_dat.csv", encoding="UTF-8")
Wykształcenie<-as.numeric(daneNaSzybko$edukacja!="wyższe i policealne")
ggplot(daneNaSzybko,aes(Wykształcenie,fill=województwo))+
  geom_bar(position="stack")+xlab(expression(bold("W. wyższe/policealne a pozostałe")))+
  theme_classic()+ylab("")
```



```
ggplot(daneNaSzybko,aes(Wykształcenie,fill=województwo))+
  geom_bar(position="fill")+xlab(expression(bold("W. wyższe/policealne a pozostałe")))+
  theme_classic()+ylab("")
```



```
wykres<-ggplot(spotkan, aes(x = plec, y = waga, fill = psycholog)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~ rok, ncol = 1)
wykres+ggtitle(expression(atop("Wizyty u psychologa",
                                atop(italic("Dane z Diagnozy Społecznej"), ""))))+standard_bp+labs(y="")
```

## Wizyty u psychologa

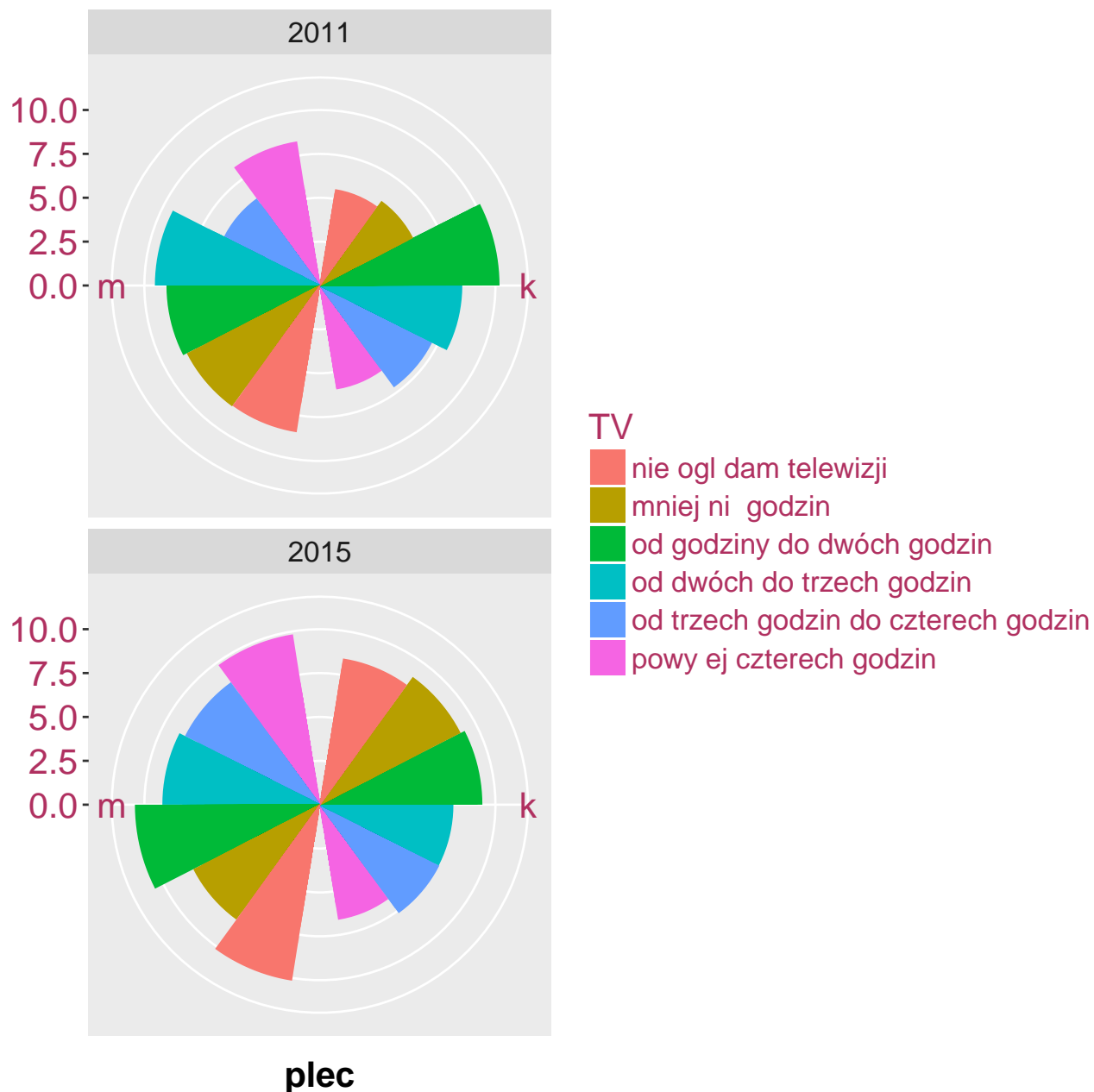
Dane z Diagnozy Społecznej



```
wykres<-ggplot(spotkan2, aes(x = plec, y = waga, fill = TV)) +
  geom_bar(stat = "identity", position = "dodge") + coord_polar() +
  facet_wrap(~ rok, ncol = 1)
wykres+ggtitle(expression(atop("Częstość oglądania TV",
                                atop(italic("Dane z Diagnozy Społecznej"), ""))))+standard_bp+labs(y="")
```

## Cz sto ogl dania TV

*Dane z Diagnozy Społecznej*



### Wykres liniowy (line plot)

- Używany najczęściej wtedy, kiedy chcemy pokazać zmienność cechy w czasie.
- Bardzo podatny na zmniejszenie czytelności przez dużą liczbę cech

```
temp = osoby %>%
  select(ap83_1, wojewodztwo, podregion58, eduk4_2000, waga_2000_ind) %>%
```

```

mutate_each(funs(as_factor), -waga_2000_ind) %>%
group_by(ap83_1, eduk4_2000, podregion58, wojewodztwo) %>%
summarise(waga = sum(waga_2000_ind, na.rm = TRUE), rok = 2000) %>%
filter(!is.na(ap83_1)) %>%
filter(!is.na(eduk4_2000)) %>%
rename(czy_pali = ap83_1, wyksztalcenie = eduk4_2000)
temp$rok = 2000
### (...)
palenie = rbind(temp, temp2, temp3, temp4, temp5, temp6, temp7, temp8)

#### more factors
palenie_a_podregion_region_wyksztalcenie_i_rok = palenie %>%
  select(czy_pali, wyksztalcenie, podregion58, wojewodztwo, rok, waga) %>%
  mutate_each(funs(as_factor), -waga)

# change values to numeric
palenie_a_podregion_region_wyksztalcenie_i_rok$czy_pali = gsub(
  'TAK', 1, palenie_a_podregion_region_wyksztalcenie_i_rok$czy_pali)
palenie_a_podregion_region_wyksztalcenie_i_rok$czy_pali = gsub(
  'NIE', 0, palenie_a_podregion_region_wyksztalcenie_i_rok$czy_pali)
palenie_a_podregion_region_wyksztalcenie_i_rok$czy_pali = as.numeric(
  palenie_a_podregion_region_wyksztalcenie_i_rok$czy_pali)

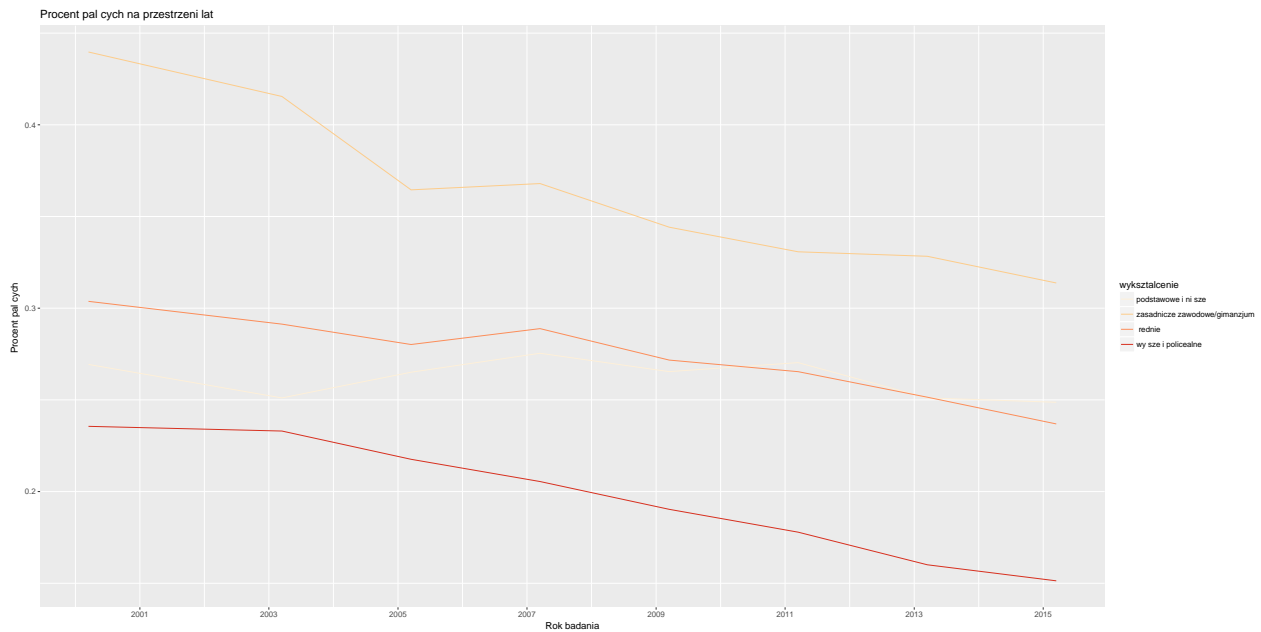
# group, count percent
palenie_a_wyksztalcenie_i_rok = palenie_a_podregion_region_wyksztalcenie_i_rok %>%
  filter(!is.na(wyksztalcenie)) %>%
  group_by(wyksztalcenie, rok) %>%
  summarise(percent = sum(czy_pali*waga)/sum(waga))

# change to datetime object
palenie_a_wyksztalcenie_i_rok$rok <- strptime(
  x = as.character(palenie_a_wyksztalcenie_i_rok$rok), format="%Y")

#First plot
ggplot(data=palenie_a_wyksztalcenie_i_rok, aes(
  x=rok, y=percent, group=wyksztalcenie, color=wyksztalcenie))+
  geom_line(stat = "identity")+
  scale_colour_brewer(type='seq', palette = 'OrRd')+
  scale_x_datetime(date_labels = "%Y", date_breaks = '2 years')+
  ggtitle('Procent palących na przestrzeni lat')+
  ylab('Procent palących')+
  xlab('Rok badania')

```

Pierwszy geom\_line:



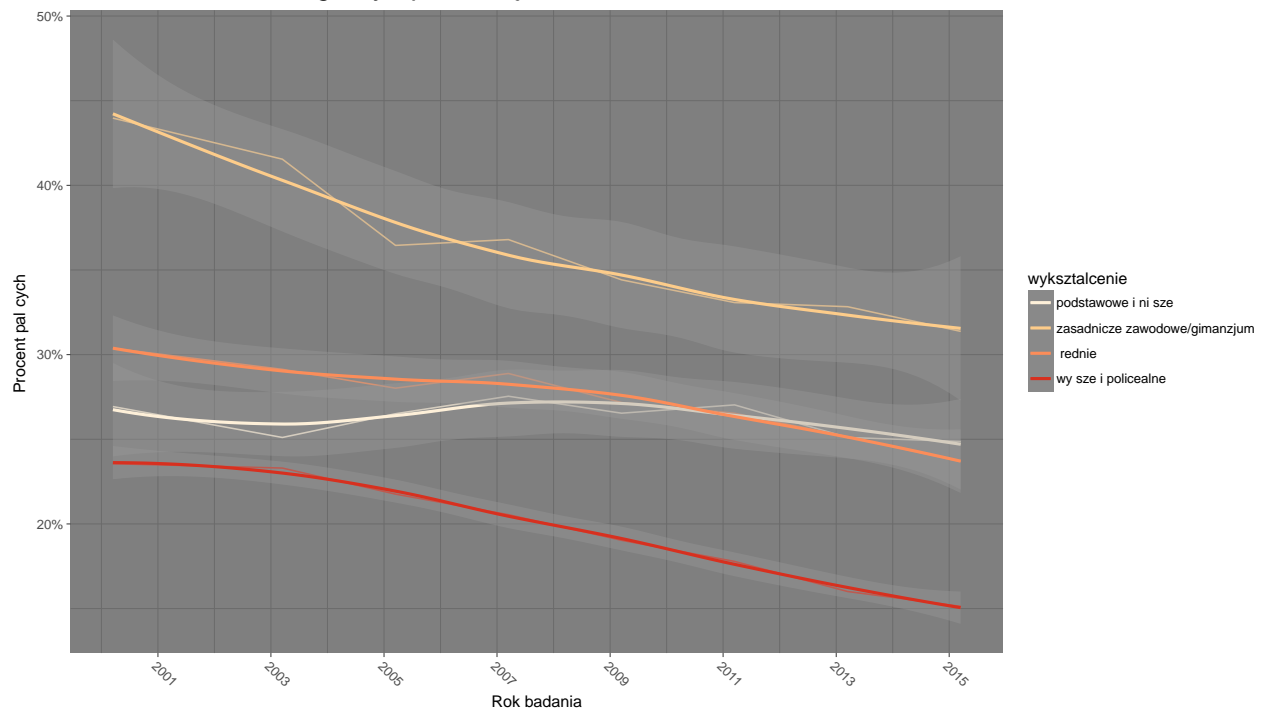
W wykresie powyżej nie pasuje nam bardzo wiele rzeczy :)

- niska czytelność
- mało czytelna skala osi Y
- chcemy czytelnika poinformować o niepewności wynikającej z trendu
- tytuł bardzo źle wygląda

```
# 2gi wykres
ggplot(data=palenie_a_wyksztalcenie_i_rok, aes(
  x=rok, y=percent, group=wyksztalcenie, color=wyksztalcenie))+
  geom_line(stat = "identity")+
  scale_colour_brewer(type='seq', palette = 'OrRd')+
  scale_x_datetime(date_labels = "%Y", date_breaks = '2 years')+
  ylab('Procent palących')+
  xlab('Rok badania')+ #####
  geom_smooth(method='loess')+
  scale_y_continuous(labels = scales::percent)+
  theme_dark()+
  ggtitle(expression(atop("Procent palących na przestrzeni lat", atop(
    italic("Dane z Diagnozy Społecznej"), "")))) +
  theme(axis.text.x = element_text(angle=-45, hjust=0, vjust=1),
        plot.title = element_text(size = 25, face = "bold", colour = "black", vjust = -1))
```

## Procent palących na przestrzeni lat

*Dane z Diagnozy Społecznej*



Na tą chwilę można dopracować:

- podpisy osi
- legenda jest mało czytelna
- tytuł

```
ggplot(data=palenie_a_wyksztalcenie_i_rok, aes(
  x=rok, y=percent, group=wyksztalcenie, color=wyksztalcenie))+
  geom_line(stat = "identity")+
  scale_colour_brewer(type='seq', palette = 'OrRd')+
  scale_x_datetime(date_labels = "%Y", date_breaks = '2 years')+
  geom_smooth(method='loess')+
  scale_y_continuous(labels = scales::percent)+
  theme_dark()+
  ggtitle(expression(atop("Procent palących na przestrzeni lat",
    atop(italic("Dane z Diagnozy Społecznej"), "")))) +
  theme(axis.text.x = element_text(angle=-45, hjust=0, vjust=1),
    plot.title = element_text(size = 25, face = "bold",
      colour = "black", vjust = -1, hjust=0.5))+#####
  theme(legend.direction = 'vertical',
    legend.key.size = unit(3, 'lines'), legend.justification = c(1, -0.015),
    legend.title = element_text(size = 18, colour = "black", vjust = 0),
    legend.text = element_text(size = 11),
    axis.title.x = element_text(size = 15, colour = 'grey35'),
    axis.title.y = element_text(size = 15, colour = 'grey35'),
    plot.background = element_rect(fill = 'grey90'),
    legend.background = element_rect(fill = 'grey90')) +
  labs(x = 'Rok badania', y = 'Procent palących', color='Wykształcenie')
```



## Procent palących na przestrzeni lat

*Dane z Diagnozy Społecznej*

