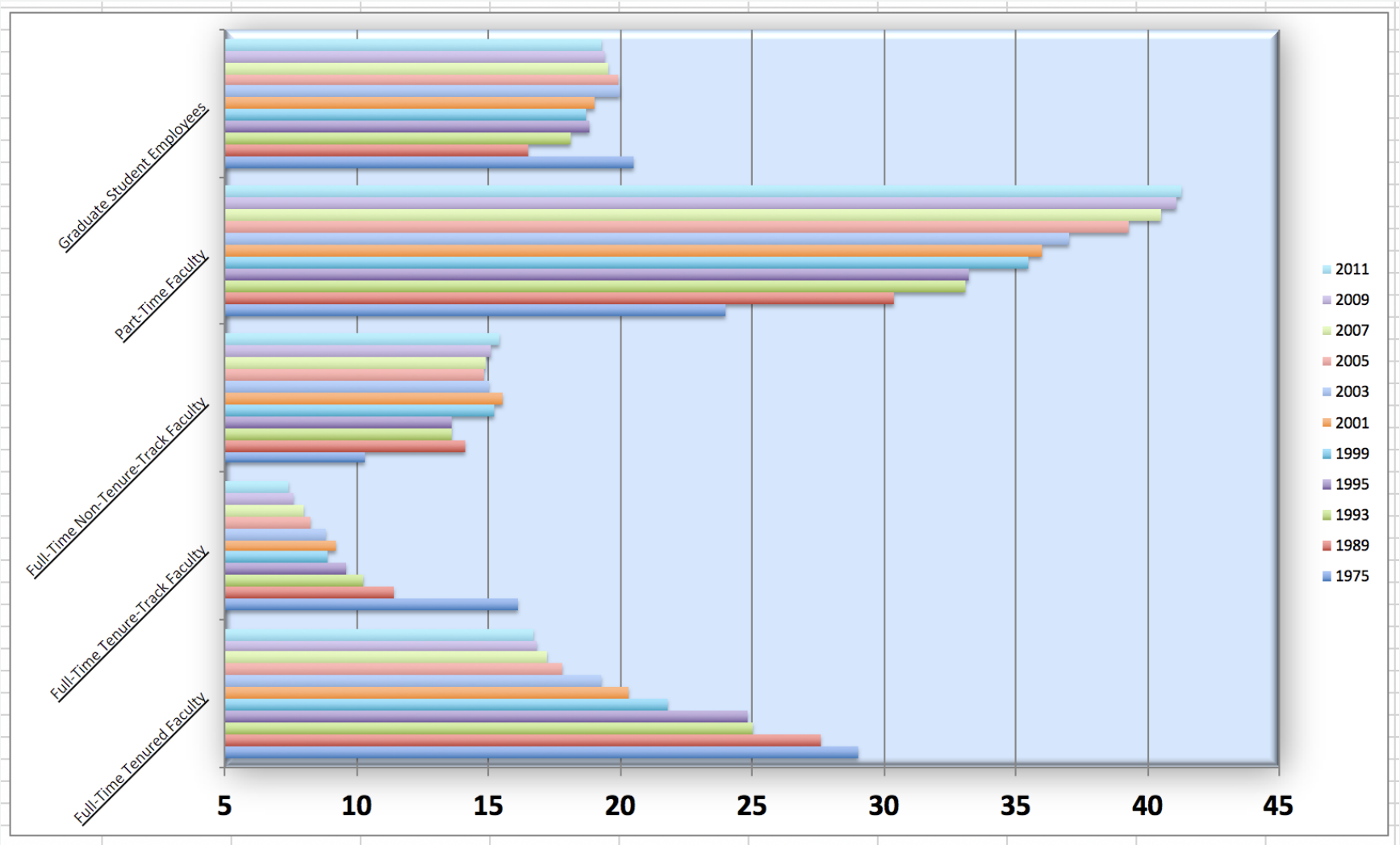
Looking at data

MACS 40700

University of Chicago

Take a sad plot, and make it better

The American Association of University Professors (AAUP) is a nonprofit membership association of faculty and other academic professionals. This report by the AAUP shows trends in instructional staff employees between 1975 and 2011, and contains an image very similar to the one given below.



Each row in this dataset represents a faculty type, and the columns are the years for which we have data. The values are percentage of hires of that type of faculty for each year.

```
staff <- read_csv("data/instructional-staff.csv")
staff
```

```
## # A tibble: 5 × 12
##   faculty_type    `1975` `1989` `1993` `1995` `1999` `2001` `2003` `2005` `2007`
##   <chr>            <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1 Full-Time Tenu…   29     27.6   25     24.8   21.8   20.3   19.3   17.8   17.2
## 2 Full-Time Tenu…   16.1   11.4   10.2    9.6    8.9    9.2    8.8    8.2    8
## 3 Full-Time Non-…   10.3   14.1   13.6   13.6   15.2   15.5   15     14.8   14.9
## 4 Part-Time Facu…   24     30.4   33.1   33.2   35.5   36     37     39.3   40.5
## 5 Graduate Stude…   20.5   16.5   18.1   18.8   18.7   19     20     19.9   19.5
## # ℹ 2 more variables: `2009` <dbl>, `2011` <dbl>
```

Recreate the visualization

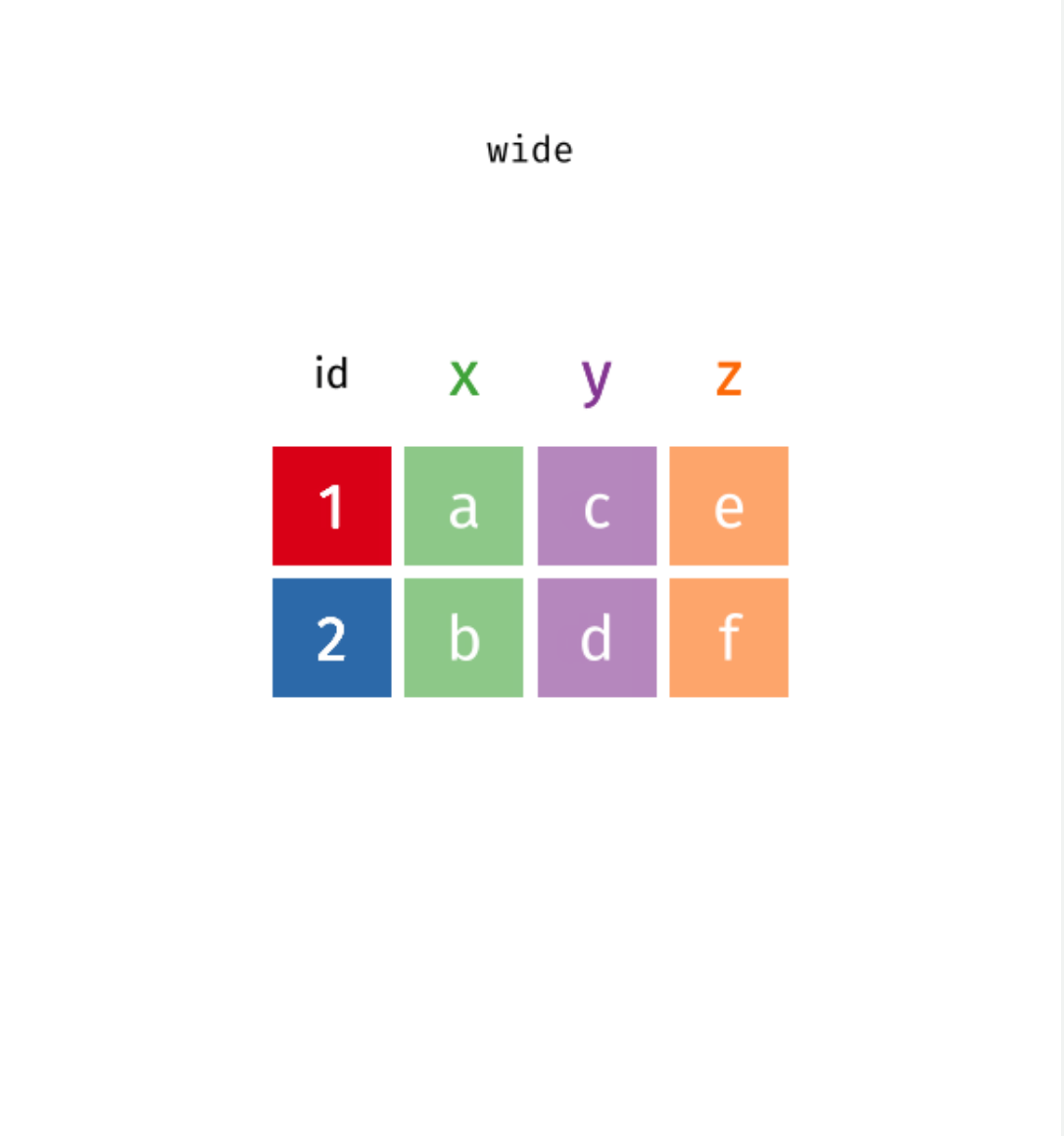
In order to recreate this visualization we need to first reshape the data to have one variable for faculty type and one variable for year. In other words, we will convert the data from the long format to wide format.

But before we do so...

If the long data will have a row for each year/faculty type combination, and there are 5 faculty types and 11 years of data, how many rows will the data have?


PIVOT!!!

`pivot_*()` function



pivot_longer()

```
pivot_longer(data, cols, names_to = "name", values_to = "value")
```

- The first argument is `data` as usual.
- The second argument, `cols`, is where you specify which columns to pivot into longer format -- in this case all columns except for the `faculty_type`
- The third argument, `names_to`, is a string specifying the name of the column to create from the data stored in the column names of data -- in this case `year`
- The fourth argument, `values_to`, is a string specifying the name of the column to create from the data stored in cell values, in this case `percentage`

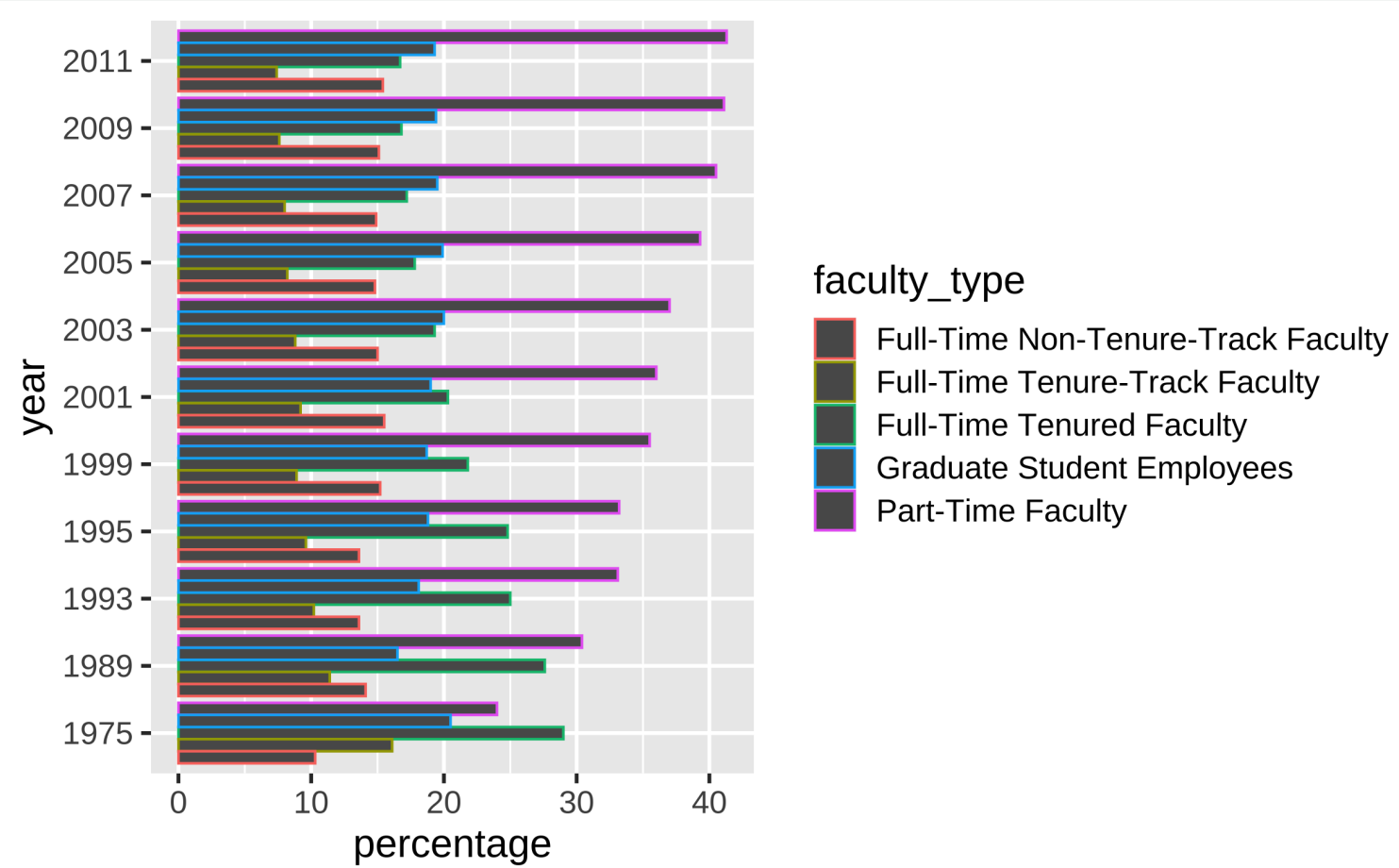
Pivot instructor data

```
library(tidyverse)

staff_long <- staff %>%
  pivot_longer(cols = -faculty_type, names_to = "year",
               values_to = "percentage") %>%
  mutate(percentage = as.numeric(percentage))

staff_long
```

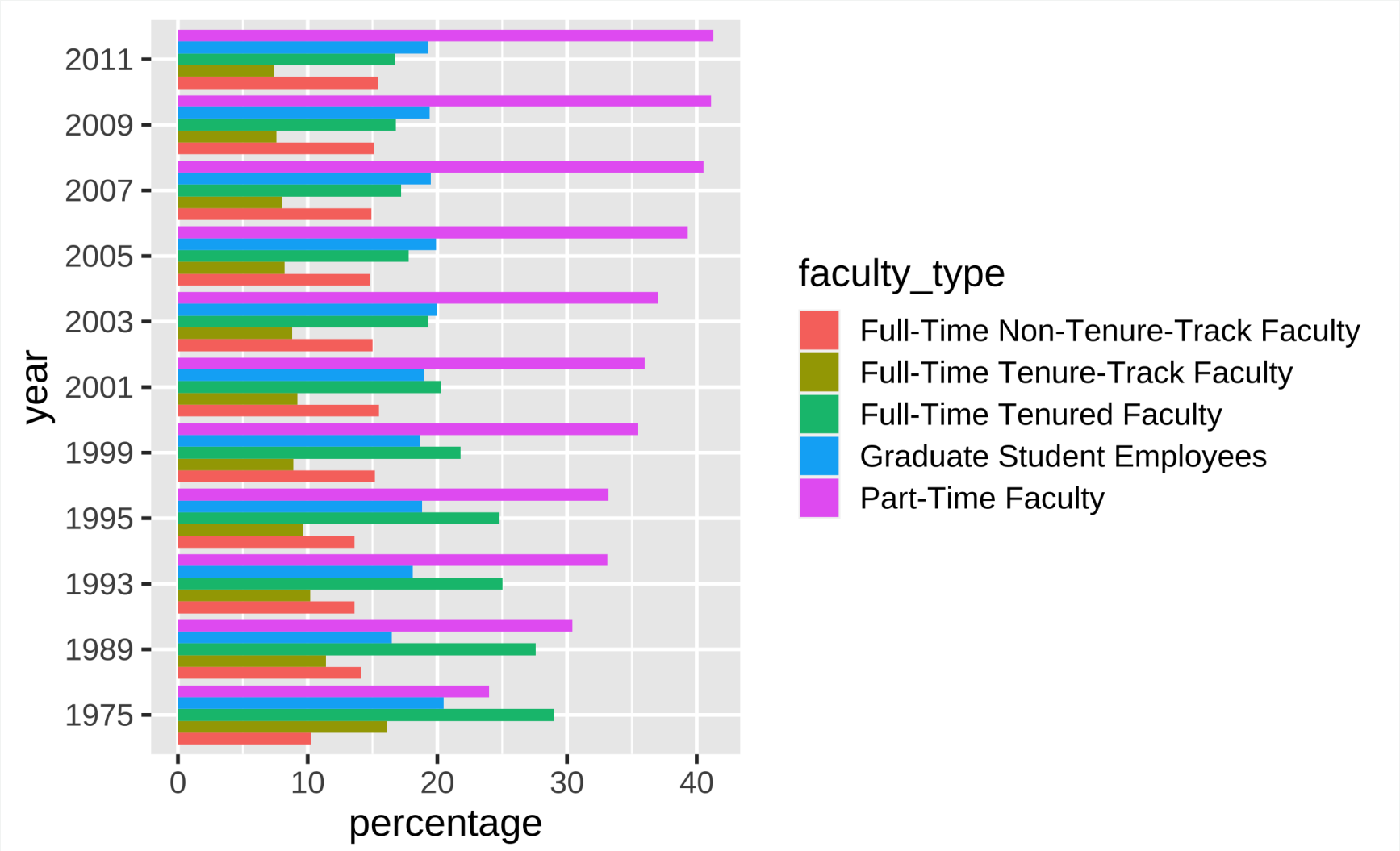
```
## # A tibble: 55 × 3
##    faculty_type              year  percentage
##    <chr>                     <chr>      <dbl>
##  1 Full-Time Tenured Faculty 1975        29
##  2 Full-Time Tenured Faculty 1989        27.6
##  3 Full-Time Tenured Faculty 1993        25
##  4 Full-Time Tenured Faculty 1995        24.8
##  5 Full-Time Tenured Faculty 1999        21.8
##  6 Full-Time Tenured Faculty 2001        20.3
##  7 Full-Time Tenured Faculty 2003        19.3
##  8 Full-Time Tenured Faculty 2005        17.8
##  9 Full-Time Tenured Faculty 2007        17.2
## 10 Full-Time Tenured Faculty 2009        16.8
## # ℹ 45 more rows
```

This doesn't look quite right, how would you fix it?

```
staff_long %>%
  ggplot(aes(x = percentage, y = year, color = faculty_type)) +
  geom_col(position = "dodge")
```

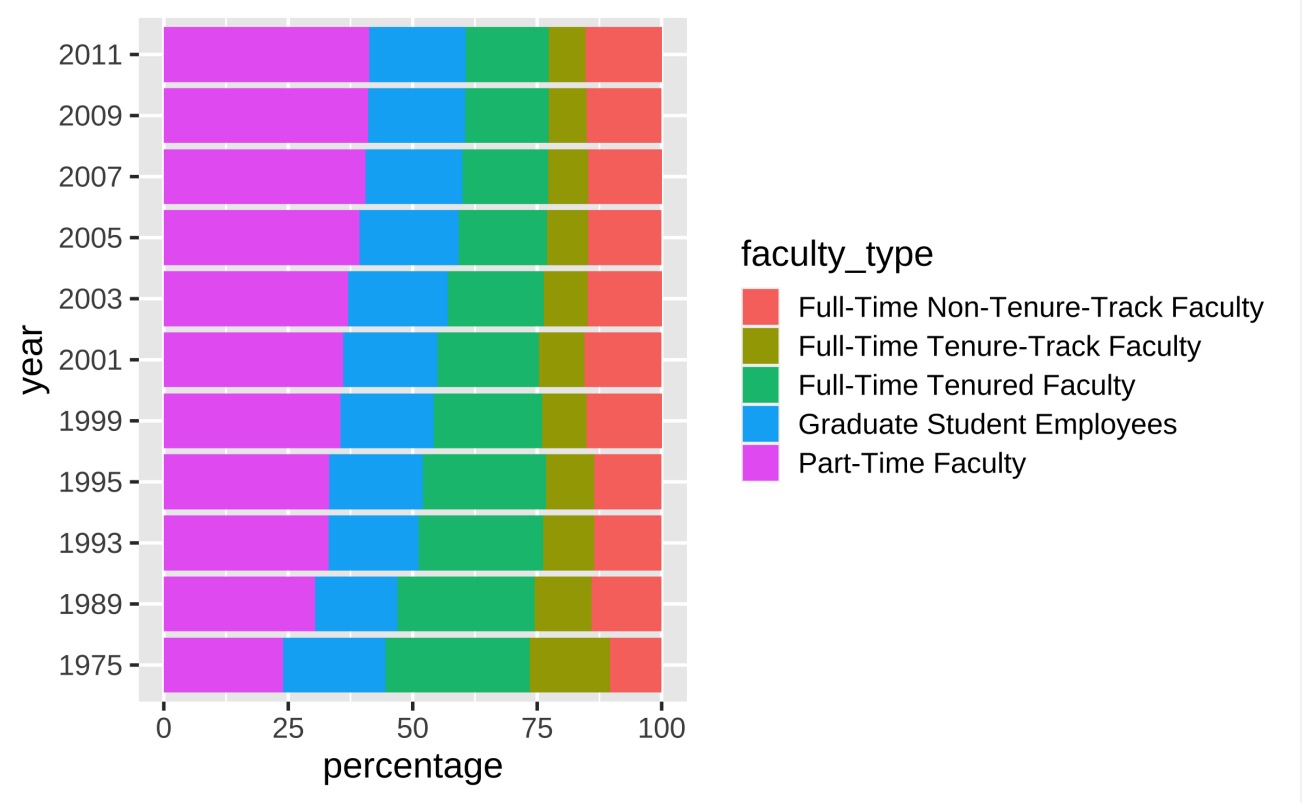


```
staff_long %>%
  ggplot(aes(x = percentage, y = year, fill = faculty_type)) +
  geom_col(position = "dodge")
```



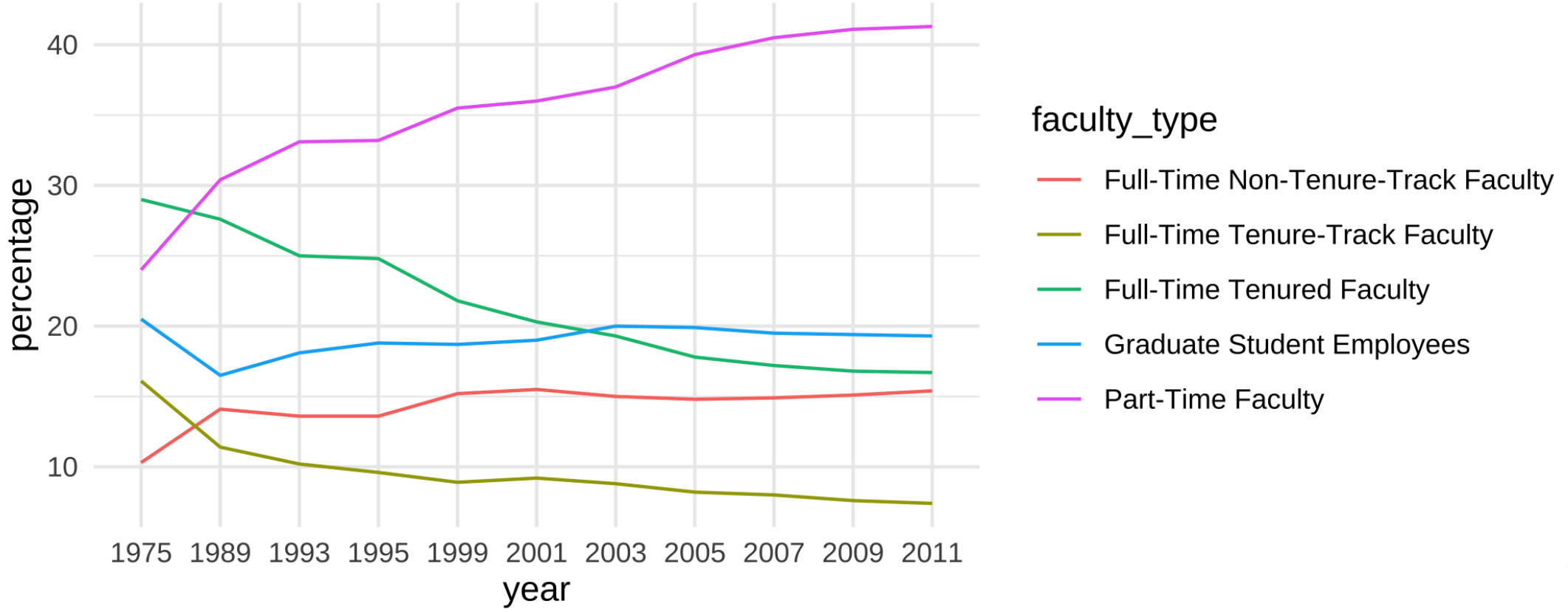
Some improvement...

```
staff_long %>%
  ggplot(aes(x = percentage, y = year, fill = faculty_type)) +
  geom_col()
```



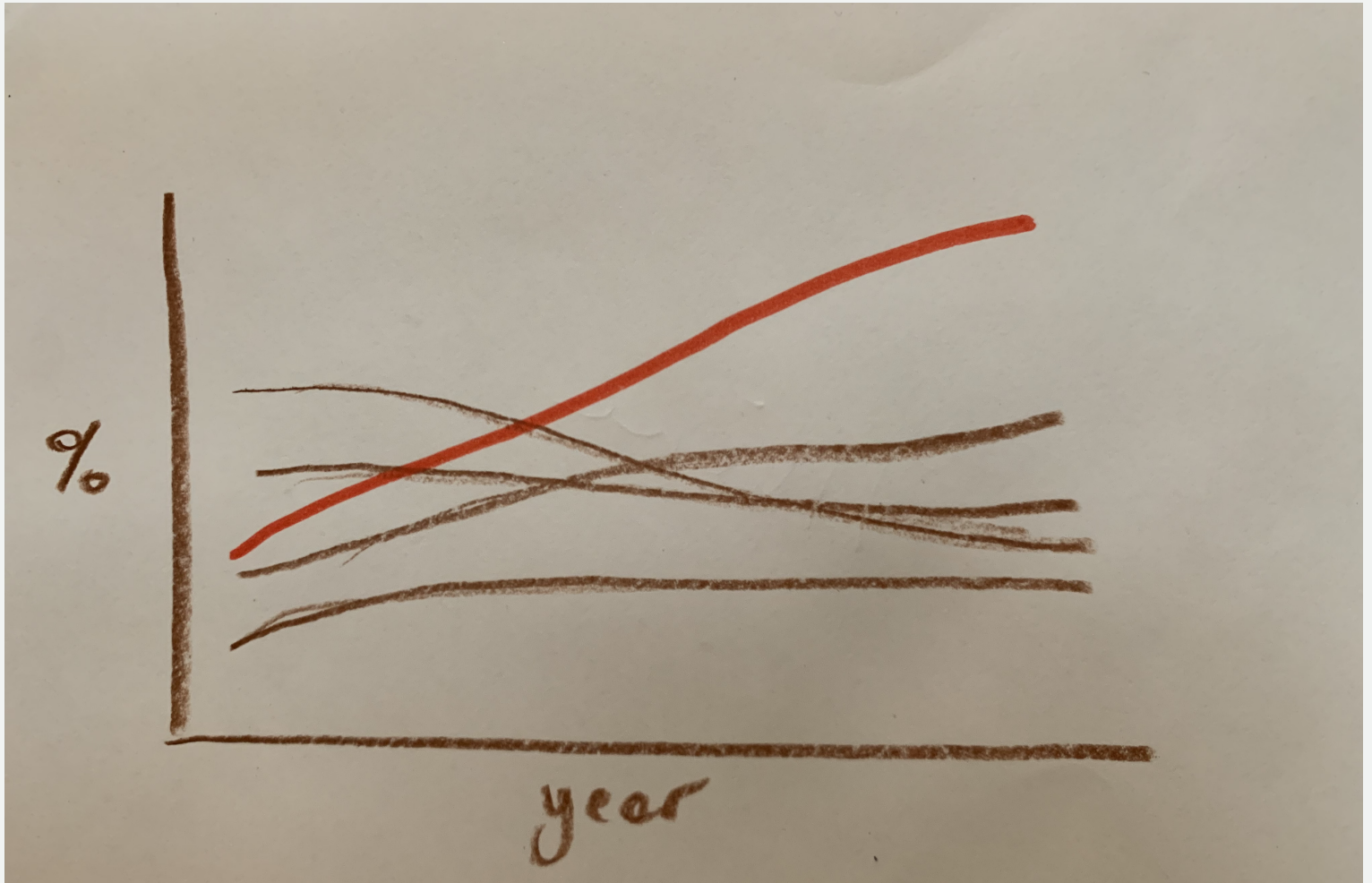
More improvement

```
staff_long %>%
  ggplot(aes(x = year, y = percentage, group = faculty_type,
             color = faculty_type)) +
  geom_line() +
  theme_minimal()
```



Goal: even more improvement!

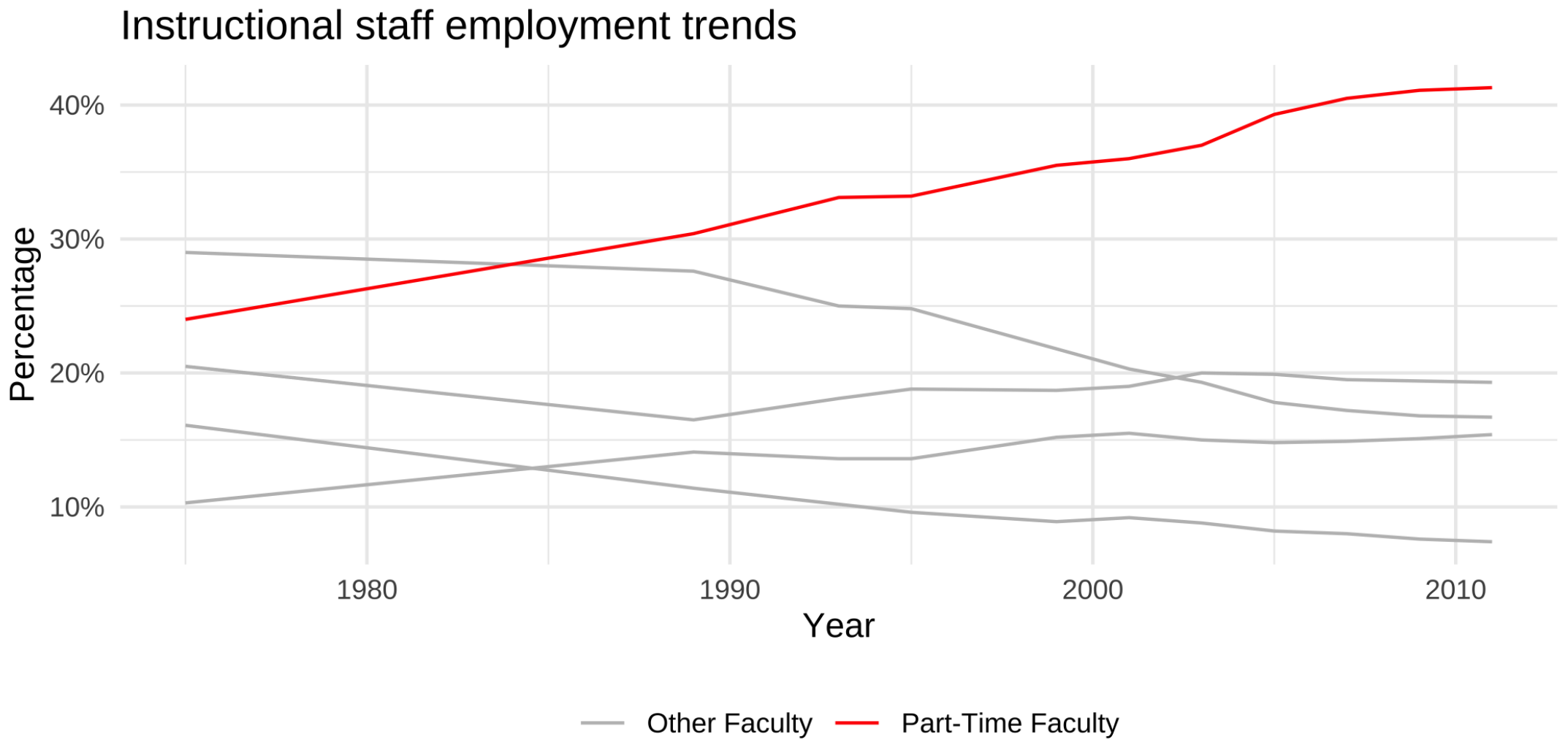
I want to achieve the following look but I have no idea how!



Asking good questions

- Describe what you want
- Describe where you are
- Create a minimal **repr**oducible **ex**ample: `reprex::reprex()`

Plot Code


Instructional staff employment trends
40%
30%
20%
10%
Percentage
1980
1990
2000
2010
Year
Other Faculty
Part-Time Faculty

Plot Code

```
library(scales)

staff_long %>%
  mutate( #<<
    part_time = if_else(faculty_type == "Part-Time Faculty", #<<
                        "Part-Time Faculty", "Other Faculty"), #<<
    year = as.numeric(year) #<<
    ) %>% #<<
  ggplot(aes(x = year, y = percentage/100, group = faculty_type,
             color = part_time)) +
  geom_line() +
  scale_color_manual(values = c("gray", "red")) + #<<
  scale_y_continuous(labels = label_percent(accuracy = 1)) + #<<
  theme_minimal() +
  labs(
    title = "Instructional staff employment trends",
    x = "Year", y = "Percentage", color = NULL
  ) +
  theme(legend.position = "bottom") #<<
```

A/B testing

Data: College education costs

- Data on four year colleges and universities in the United States (2018-19)
- Extracted from College Scorecard API
- Source: `rcis::scorecard`

rcis::scorecard

```
library(tidyverse)
# remotes::install_github("cis-ds/rcis")
library(rcis)

glimpse(scorecard)
```

```
## Rows: 1,721
## Columns: 14
## $ unitid    <dbl> 100654, 100663, 100706, 100724, 100751, 100830, 100858, 1009…
## $ name      <chr> "Alabama A & M University", "University of Alabama at Birmin…
## $ state     <chr> "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", …
## $ type      <fct> "Public", "Public", "Public", "Public", "Public", "Public", …
## $ admrate   <dbl> 0.8965, 0.8060, 0.7711, 0.9888, 0.8039, 0.9555, 0.8507, 0.60…
## $ satavg    <dbl> 959, 1245, 1300, 938, 1262, 1061, 1302, 1202, 1068, NA, 1101…
## $ cost      <dbl> 23445, 25542, 24861, 21892, 30016, 20225, 32196, 32514, 3483…
## $ netcost   <dbl> 15529, 16530, 17208, 19534, 20917, 13678, 24018, 19808, 2050…
## $ avgfacsal <dbl> 68391, 102420, 87273, 64746, 93141, 69561, 96498, 62649, 533…
## $ pctpell   <dbl> 0.7095, 0.3397, 0.2403, 0.7368, 0.1718, 0.4654, 0.1343, 0.22…
## $ comprate  <dbl> 0.2866, 0.6117, 0.5714, 0.3177, 0.7214, 0.3040, 0.7870, 0.70…
## $ firstgen  <dbl> 0.3658281, 0.3412237, 0.3101322, 0.3434343, 0.2257127, 0.381…
## $ debt      <dbl> 15250, 15085, 14000, 17500, 17671, 12000, 17500, 16000, 1425…
## $ locale    <fct> City, City, City, City, City, City, City, City, City, Suburb…
```

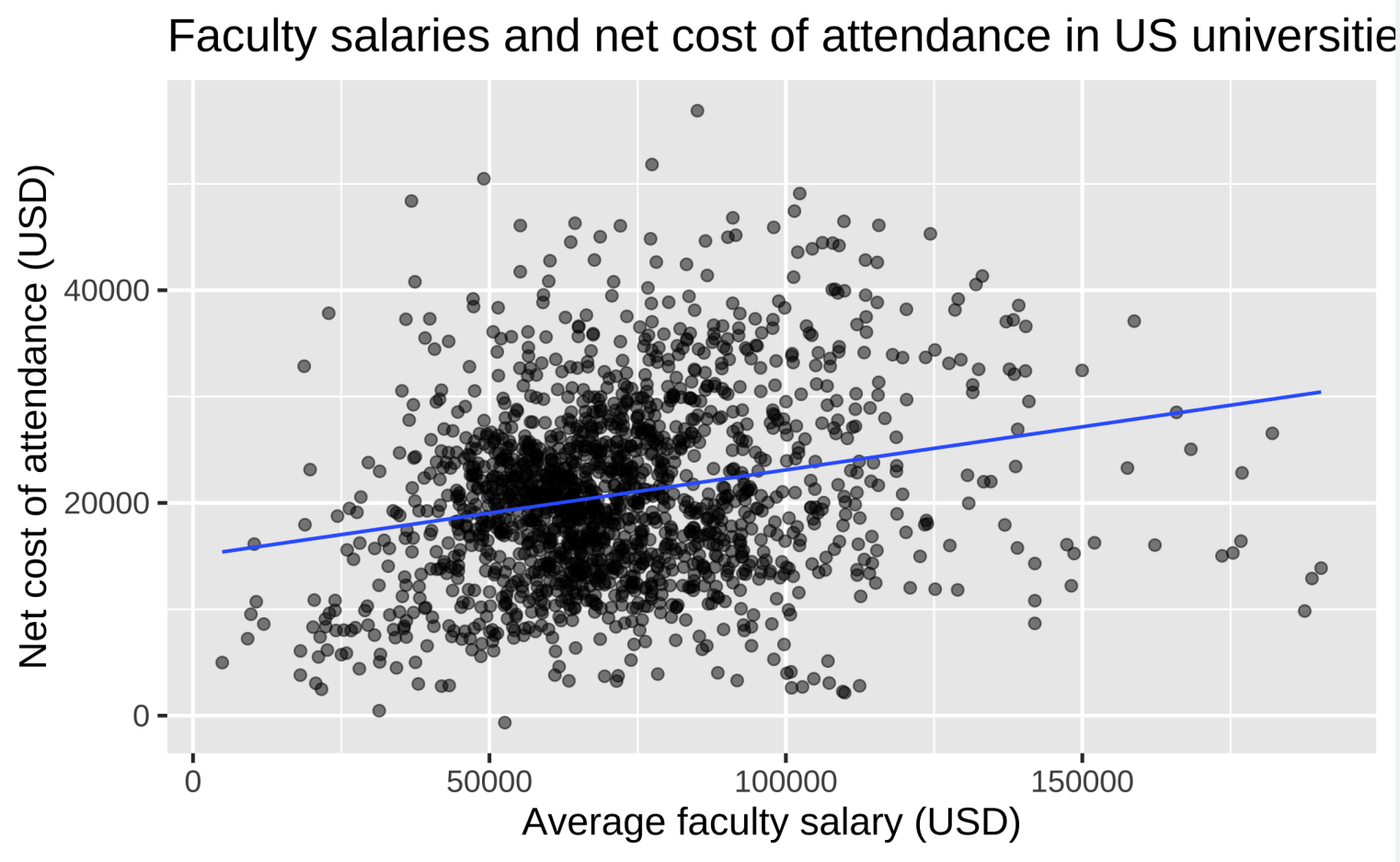
A simple visualization

Code | Plot

```
ggplot(scorecard, aes(x = avgfacsal, y = netcost)) +
  geom_point(alpha = 0.5, size = 2) +
  geom_smooth(method = "lm", se = FALSE, size = 0.7) +
  labs(
    x = "Average faculty salary (USD)",
    y = "Net cost of attendance (USD)",
    title = "Faculty salaries and net cost of attendance in US universities"
  )
```

A simple visualization

Code | Plot



New variable: pctpell_cat

```
scorecard <- scorecard %>%
  mutate(pctpell_cat = cut_interval(x = pctpell, n = 6)) %>%
  drop_na(pctpell_cat)

scorecard %>%
  select(pctpell, pctpell_cat)
```

```
## # A tibble: 1,717 × 2
##    pctpell pctpell_cat
##      <dbl> <fct>
##  1   0.710 (0.667,0.833]
##  2   0.340 (0.333,0.5]
##  3   0.240 (0.167,0.333]
##  4   0.737 (0.667,0.833]
##  5   0.172 (0.167,0.333]
##  6   0.465 (0.333,0.5]
##  7   0.134 [0,0.167]
##  8   0.226 (0.167,0.333]
##  9   0.501 (0.5,0.667]
## 10   0.697 (0.667,0.833]
## # ℹ 1,707 more rows
```

Distribution of `pctpell_cat`

```
scorecard <- scorecard %>%
  mutate(pctpell_cat = cut_interval(x = pctpell, n = 6)) %>%
  drop_na(pctpell_cat)

scorecard %>%
  count(pctpell_cat)
```

```
## # A tibble: 6 × 2
##   pctpell_cat       n
##   <fct>         <int>
## 1 [0,0.167]       169
## 2 (0.167,0.333]   630
## 3 (0.333,0.5]     609
## 4 (0.5,0.667]     205
## 5 (0.667,0.833]    73
## 6 (0.833,1]        31
```

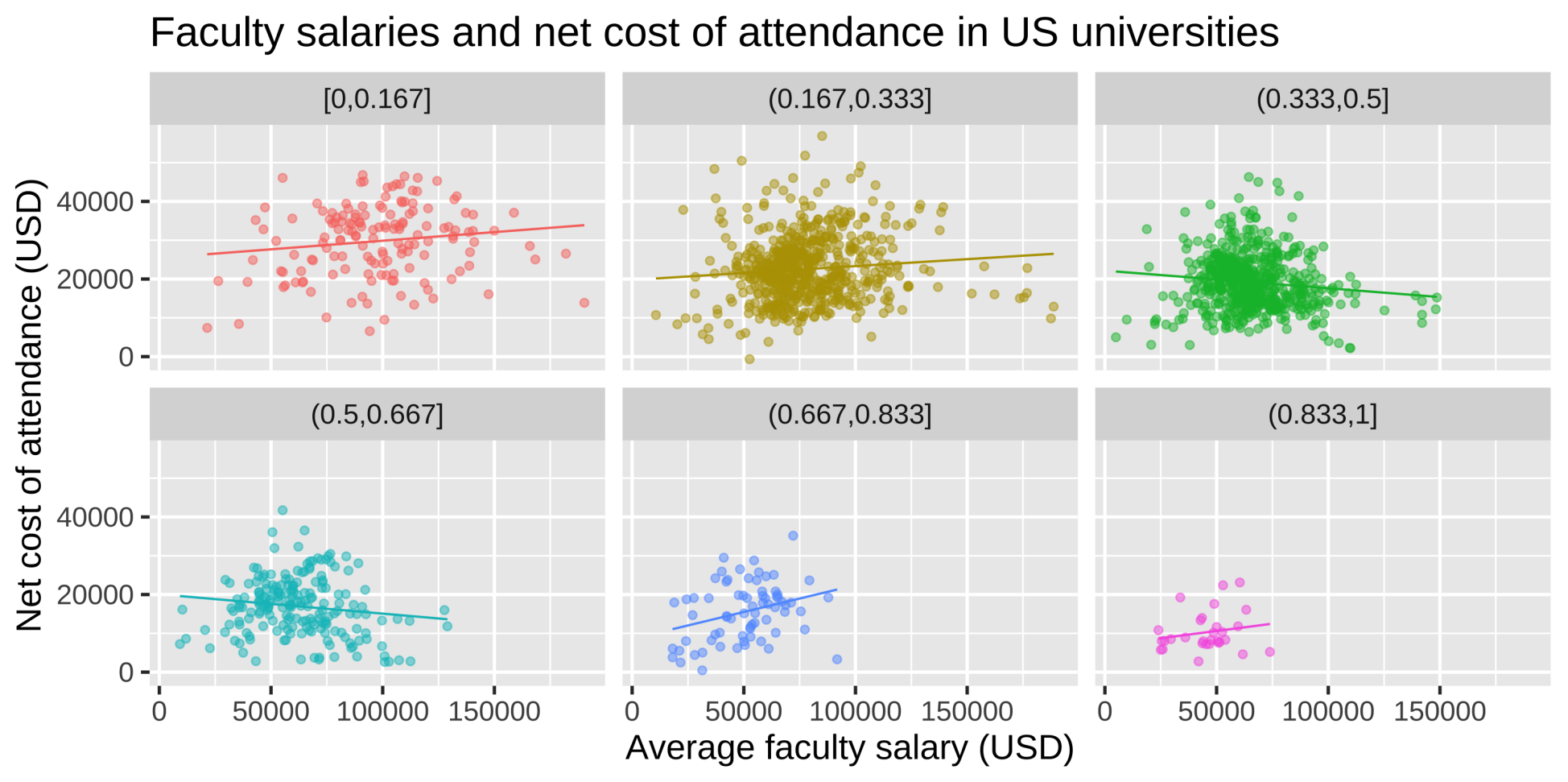
A slightly more complex visualization

Code | Plot

```
ggplot(scorecard, aes(x = avgfacsal, y = netcost, color = pctpell_cat)) +
  geom_point(alpha = 0.5, show.legend = FALSE) +
  geom_smooth(method = "lm", se = FALSE, size = 0.5, show.legend = FALSE) +
  facet_wrap(vars(pctpell_cat)) +
  labs(
    x = "Average faculty salary (USD)",
    y = "Net cost of attendance (USD)",
    color = "Percentage of Pell grant recipients",
    title = "Faculty salaries and net cost of attendance in US universities"
  )
```

A slightly more complex visualization

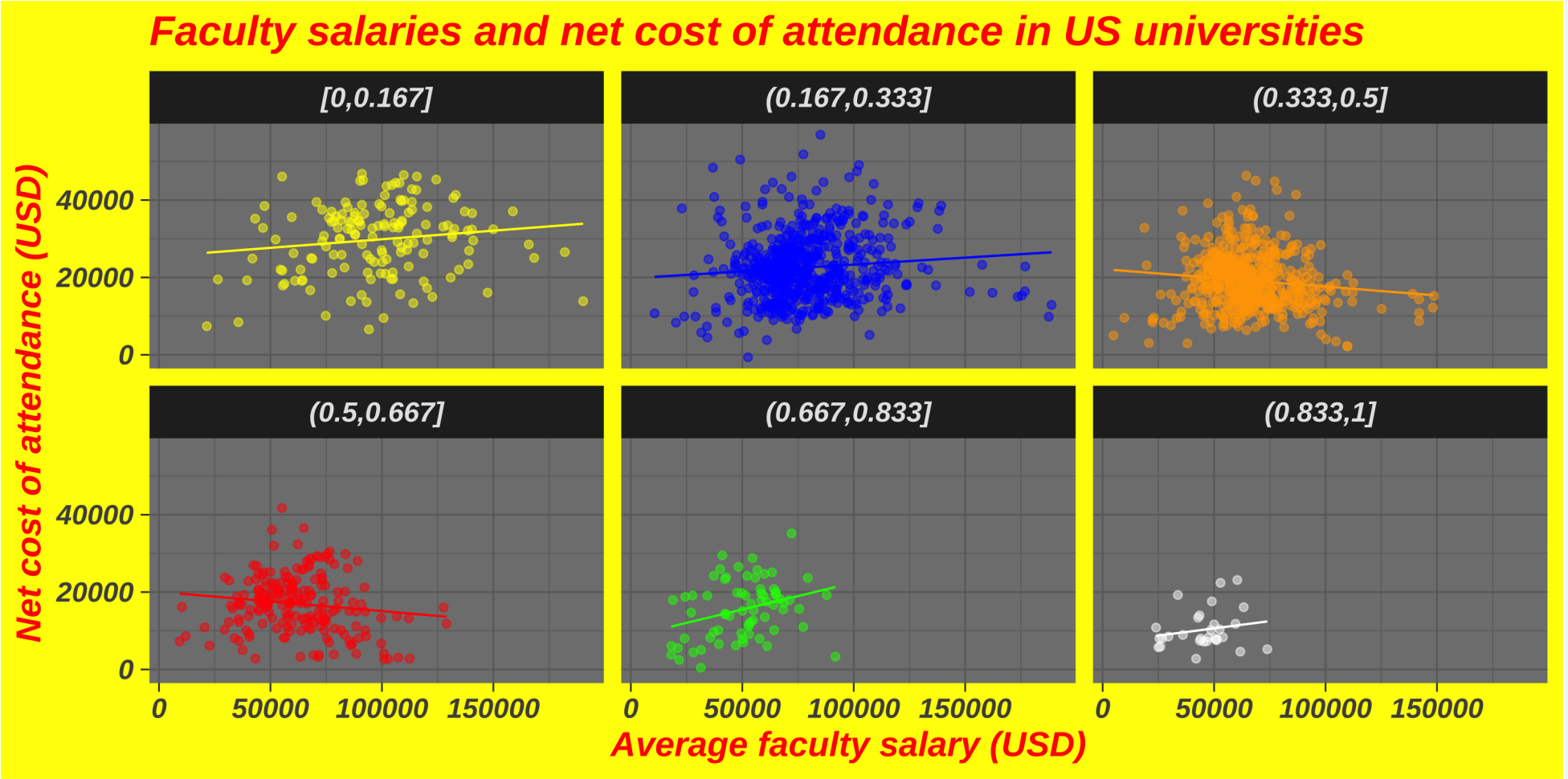
Code | Plot



In the next two slides, the same plots are created with different "cosmetic" choices. Examine the plots two given (Plot A and Plot B), and decide whcih one you prefer.

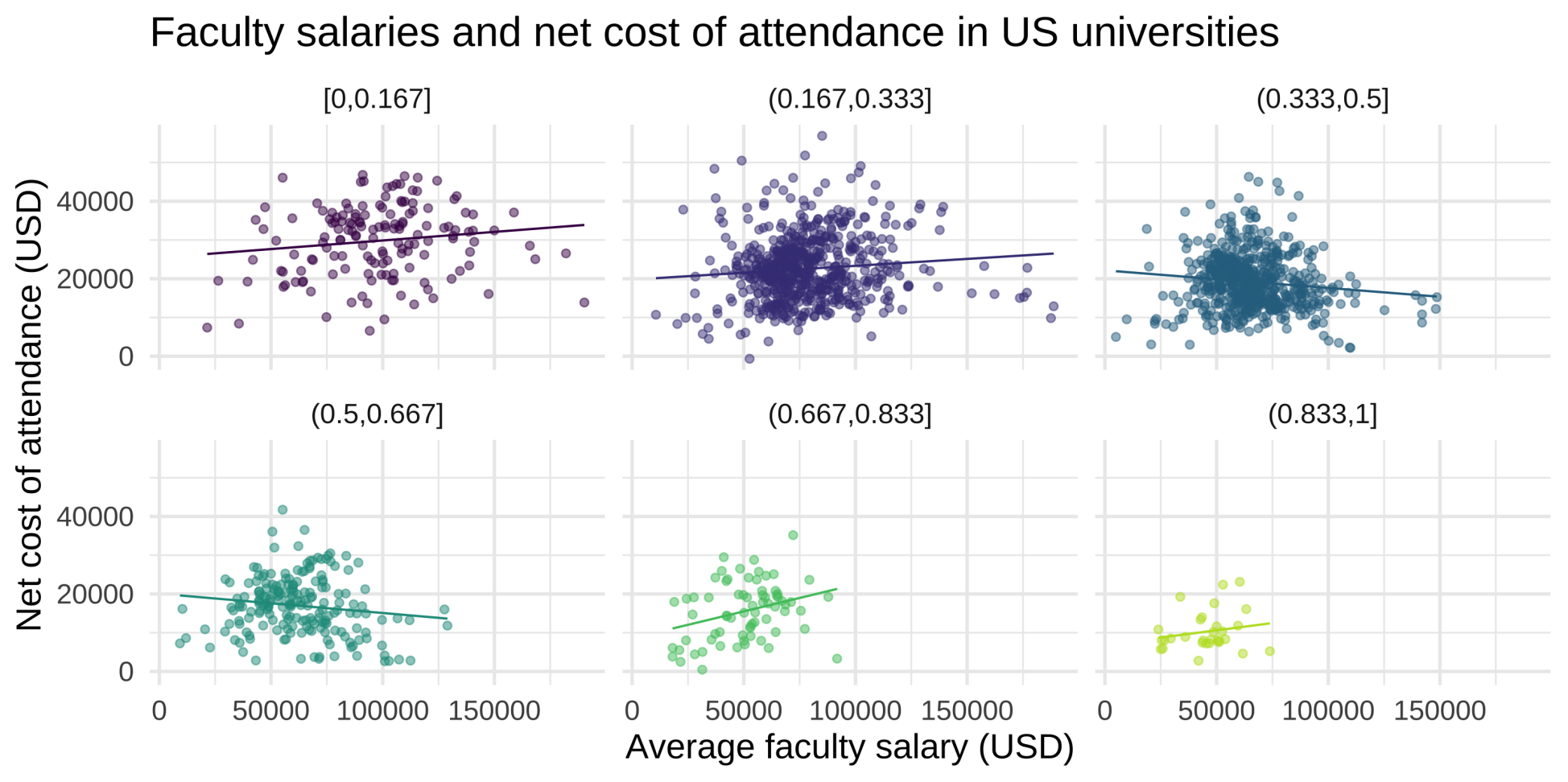
Test 1

Plot A | Plot B



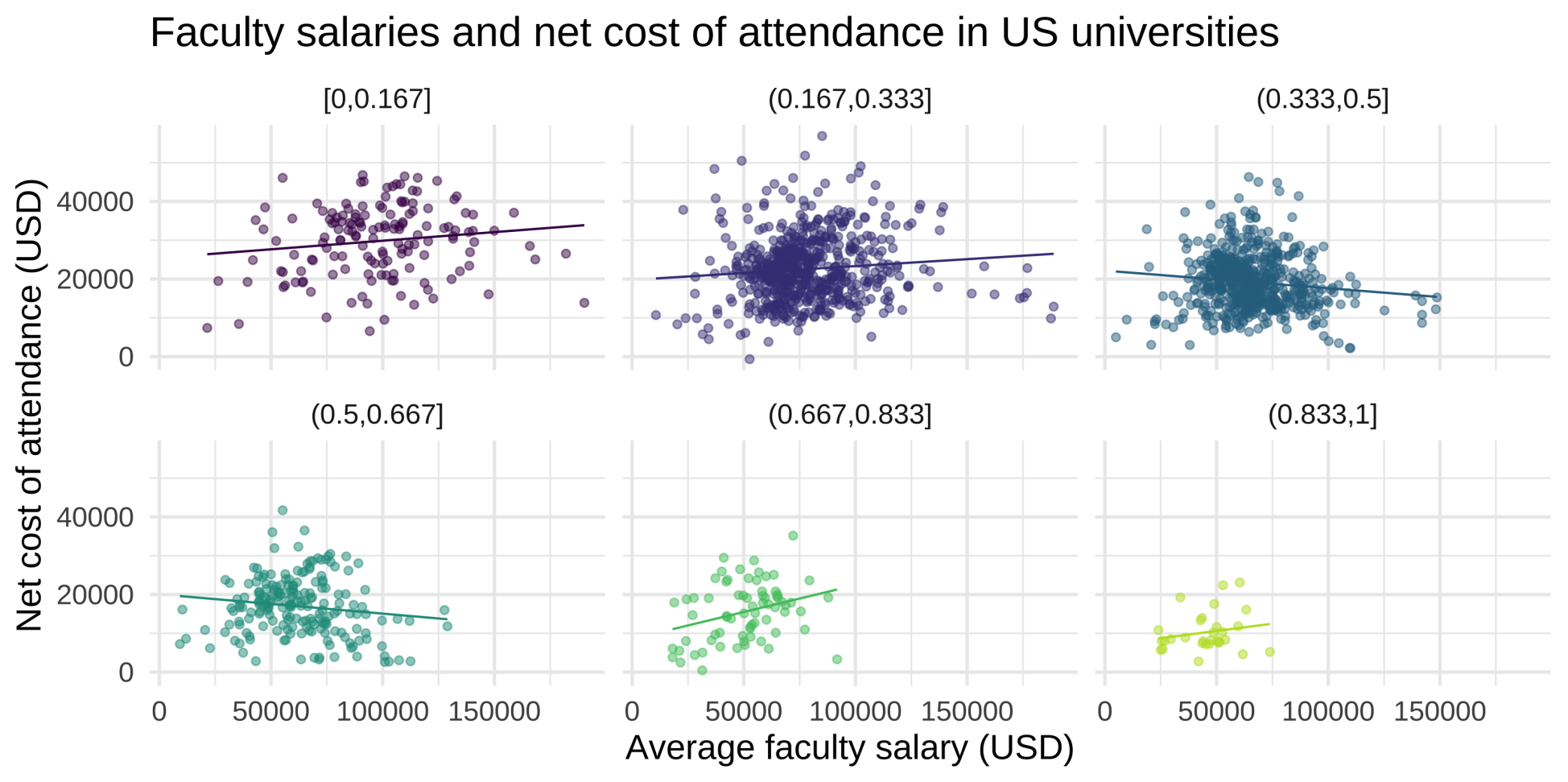
Test 1

Plot A | Plot B



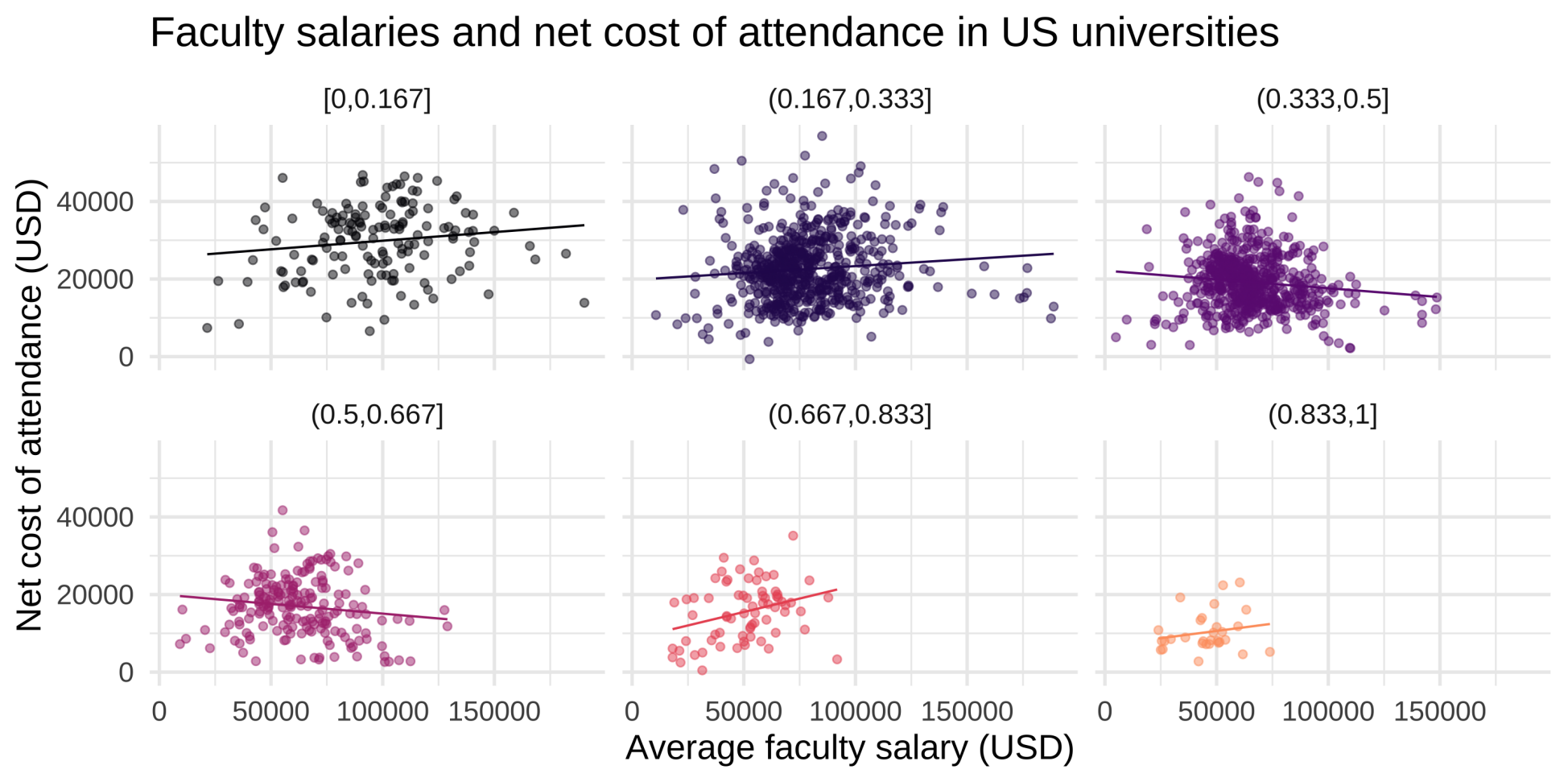
Test 2

Plot A | Plot B



Test 2

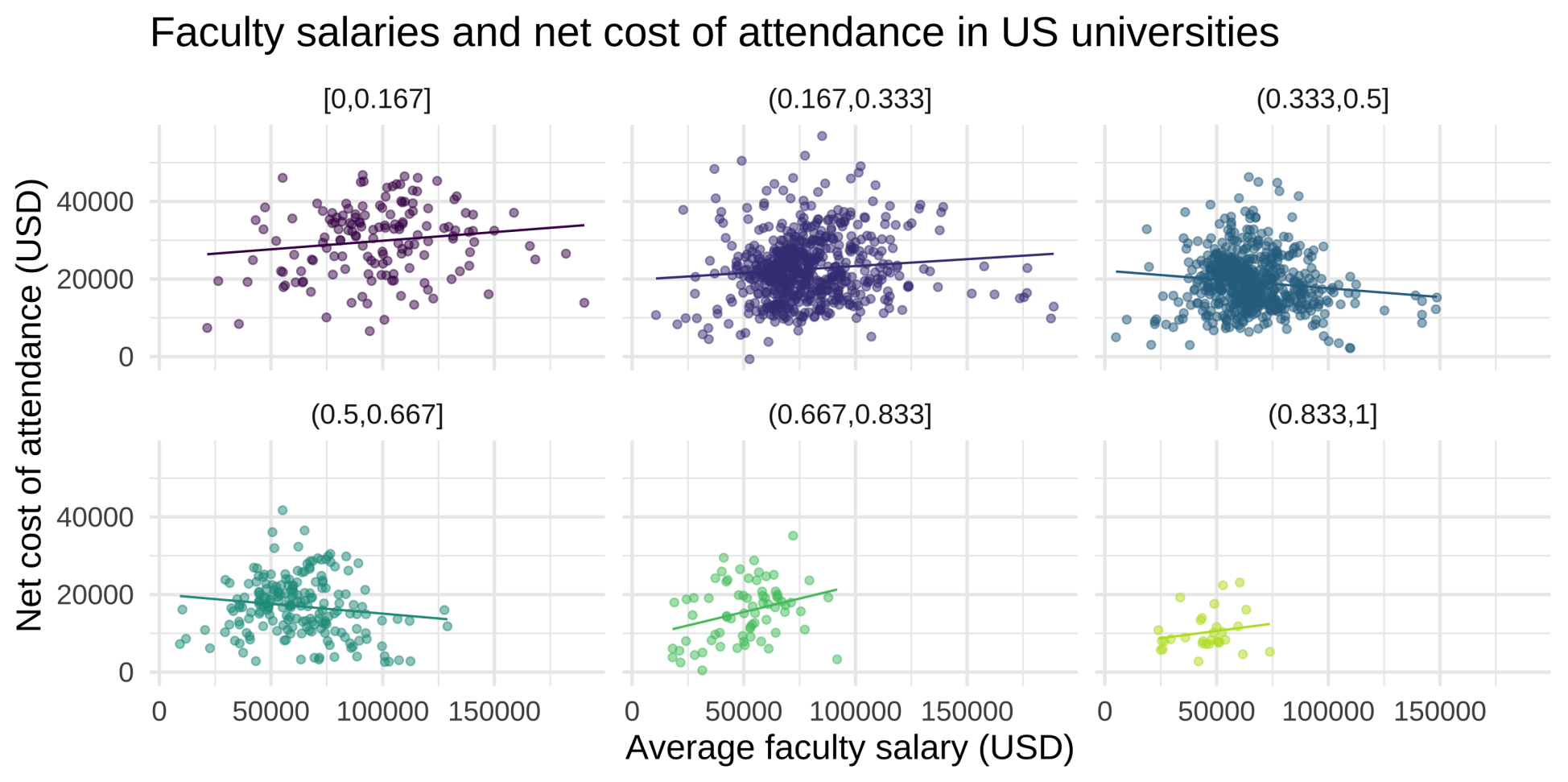
Plot A | Plot B



a deeper look at the plotting code...

Minimal theme + viridis scale, default option

Plot Code



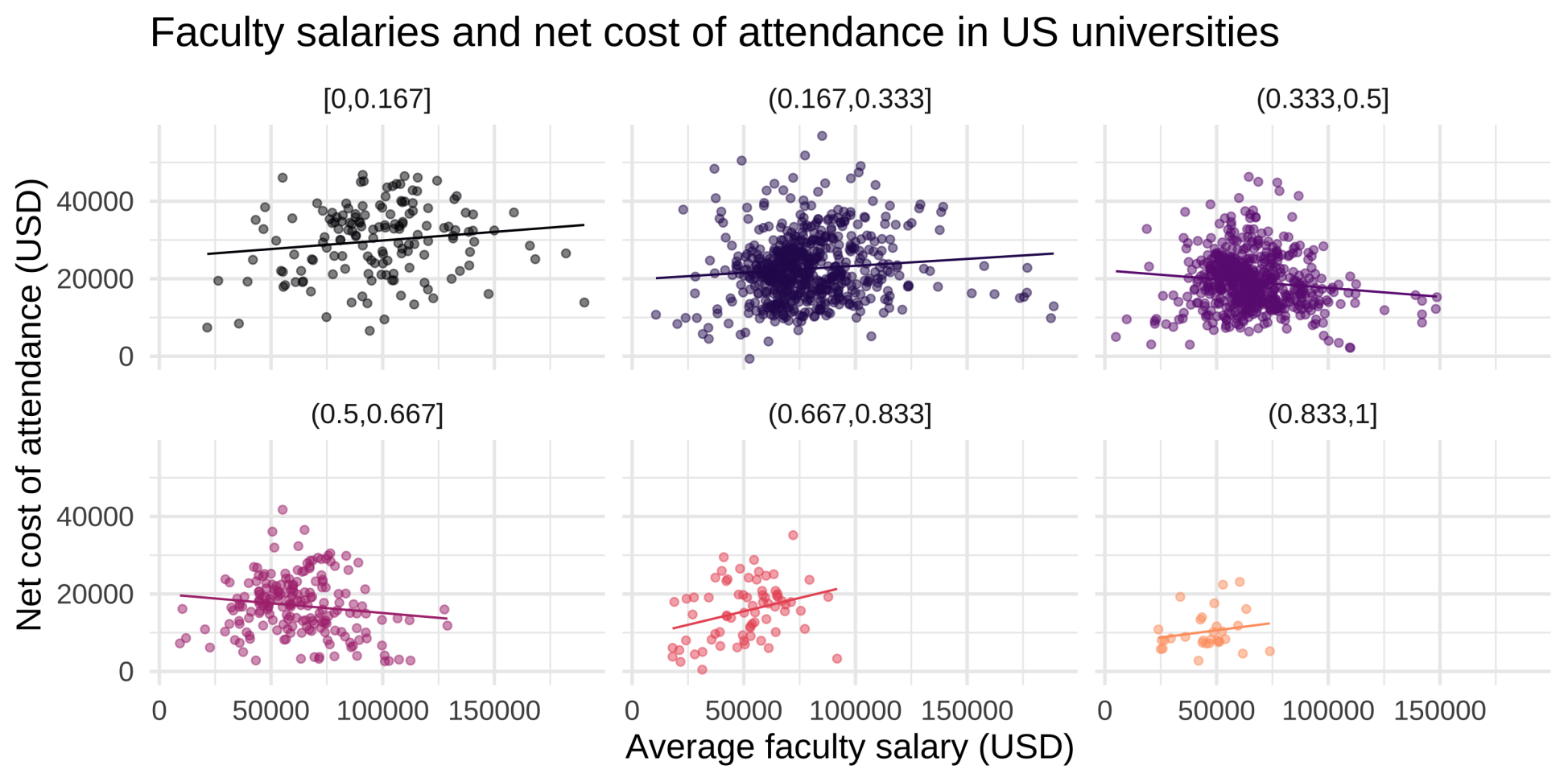
Minimal theme + viridis scale, default option

Plot | Code

```
ggplot(scorecard, aes(x = avgfacsal, y = netcost, color = pctpell_cat)) +
  geom_point(alpha = 0.5, show.legend = FALSE) +
  geom_smooth(method = "lm", se = FALSE, size = 0.5, show.legend = FALSE) +
  facet_wrap(vars(pctpell_cat)) +
  labs(
    x = "Average faculty salary (USD)",
    y = "Net cost of attendance (USD)",
    color = "Percentage of Pell grant recipients",
    title = "Faculty salaries and net cost of attendance in US universities"
  ) +
  theme_minimal(base_size = 16) +  #<<
  scale_color_viridis_d(end = 0.9)  #<<
```

Viridis scale, option A (magma)

Plot Code



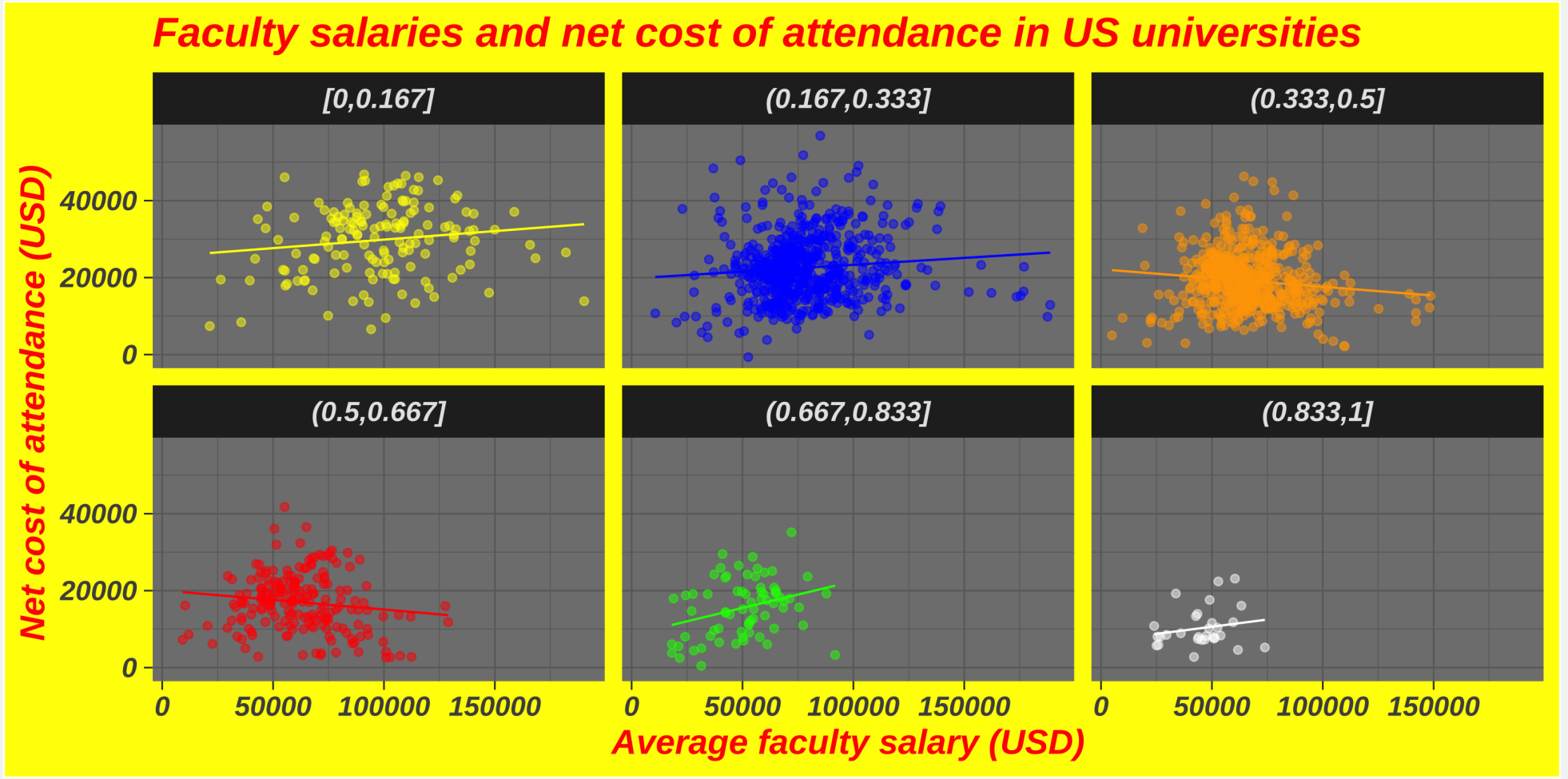
Viridis scale, option A (magma)

Plot | Code

```
ggplot(scorecard, aes(x = avgfacsal, y = netcost, color = pctpell_cat)) +
  geom_point(alpha = 0.5, show.legend = FALSE) +
  geom_smooth(method = "lm", se = FALSE, size = 0.5, show.legend = FALSE) +
  facet_wrap(vars(pctpell_cat)) +
  labs(
    x = "Average faculty salary (USD)",
    y = "Net cost of attendance (USD)",
    color = "Percentage of Pell grant recipients",
    title = "Faculty salaries and net cost of attendance in US universities"
  ) +
  theme_minimal(base_size = 16) +
  scale_color_viridis_d(end = 0.8, option = "A") #<<
```

Dark theme + further theme customization

Plot Code



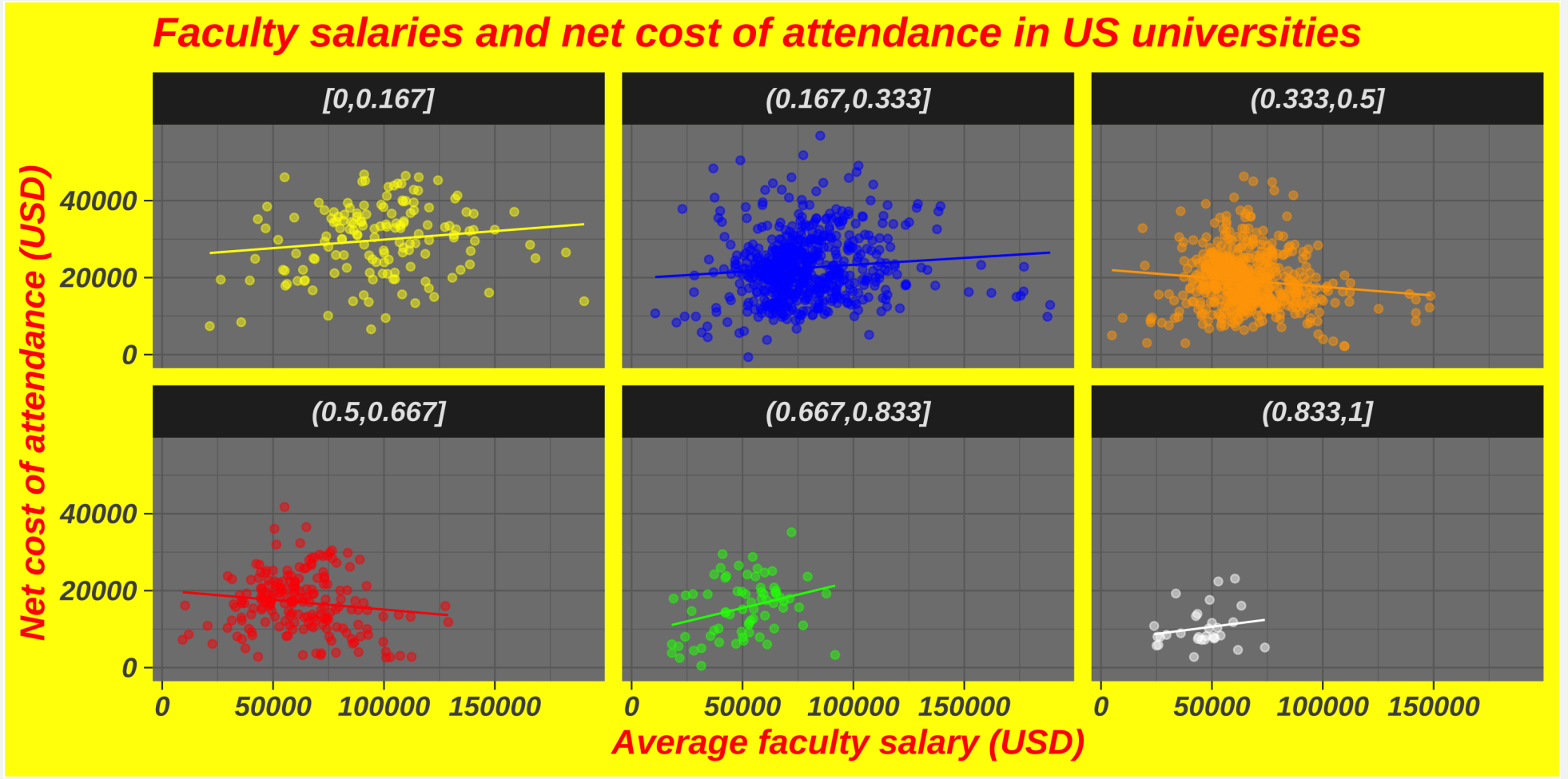
Dark theme + further theme customization

Plot | Code

```
ggplot(scorecard, aes(x = avgfacsal, y = netcost, color = pctpell_cat)) +
  geom_point(alpha = 0.5, show.legend = FALSE) +
  geom_smooth(method = "lm", se = FALSE, size = 0.5, show.legend = FALSE) +
  facet_wrap(vars(pctpell_cat)) +
  labs(
    x = "Average faculty salary (USD)",
    y = "Net cost of attendance (USD)",
    color = "Percentage of Pell grant recipients",
    title = "Faculty salaries and net cost of attendance in US universities"
  ) +
  theme_dark(base_size = 16) + #<<
  scale_color_manual(values = c("yellow", "blue", "orange", "red", "green", "white")) + #
  theme( #<<
    text = element_text(color = "red", face = "bold.italic"), #<<
    plot.background = element_rect(fill = "yellow") #<<
  ) #<<
```

What makes bad figures bad?

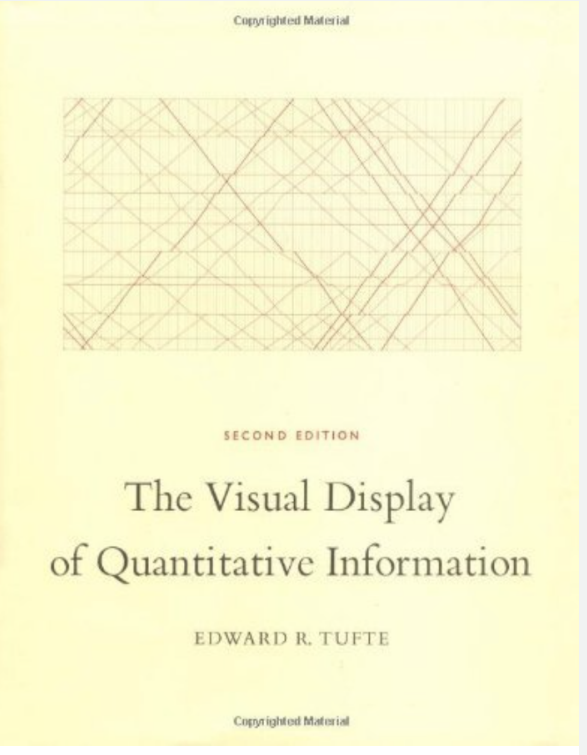
Bad taste


Faculty salaries and net cost of attendance in US universities
[0,0.167]
(0.167,0.333]
(0.333,0.5]
(0.5,0.667]
(0.667,0.833]
(0.833,1]
Net cost of attendance (USD)
40000
20000
0
0
50000
100000
150000
Average faculty salary (USD)

Data-to-ink ratio

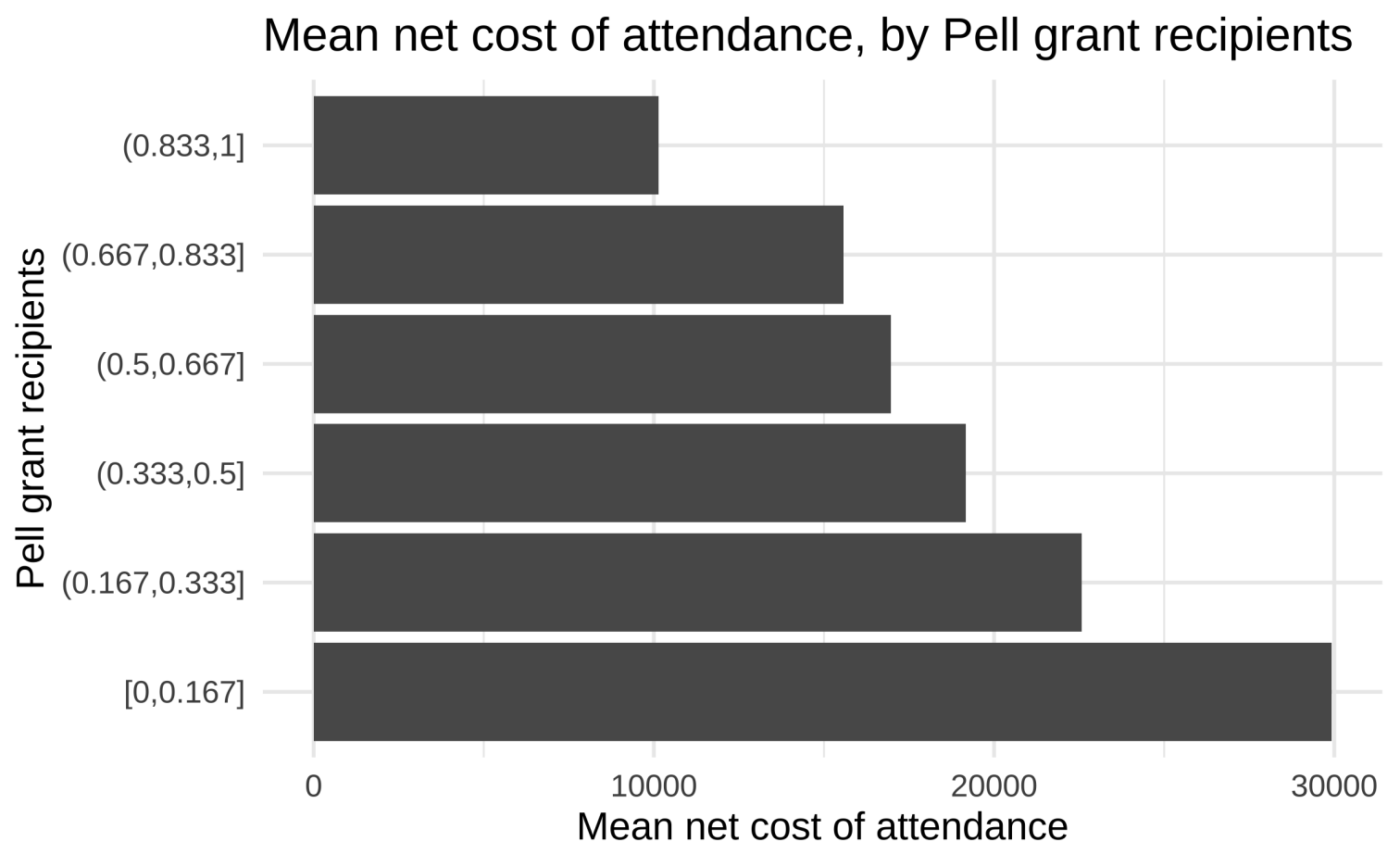
Tufte strongly recommends maximizing the **data-to-ink ratio** this in the Visual Display of Quantitative Information (Tufte, 1983).

> Graphical excellence is the well-designed presentation of interesting data—a matter of substance, of statistics, and of design … [It] consists of complex ideas communicated with clarity, precision, and efficiency. … [It] is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space … [It] is nearly always multivariate … And graphical excellence requires telling the truth about the data. (Tufte, 1983, p. 51).



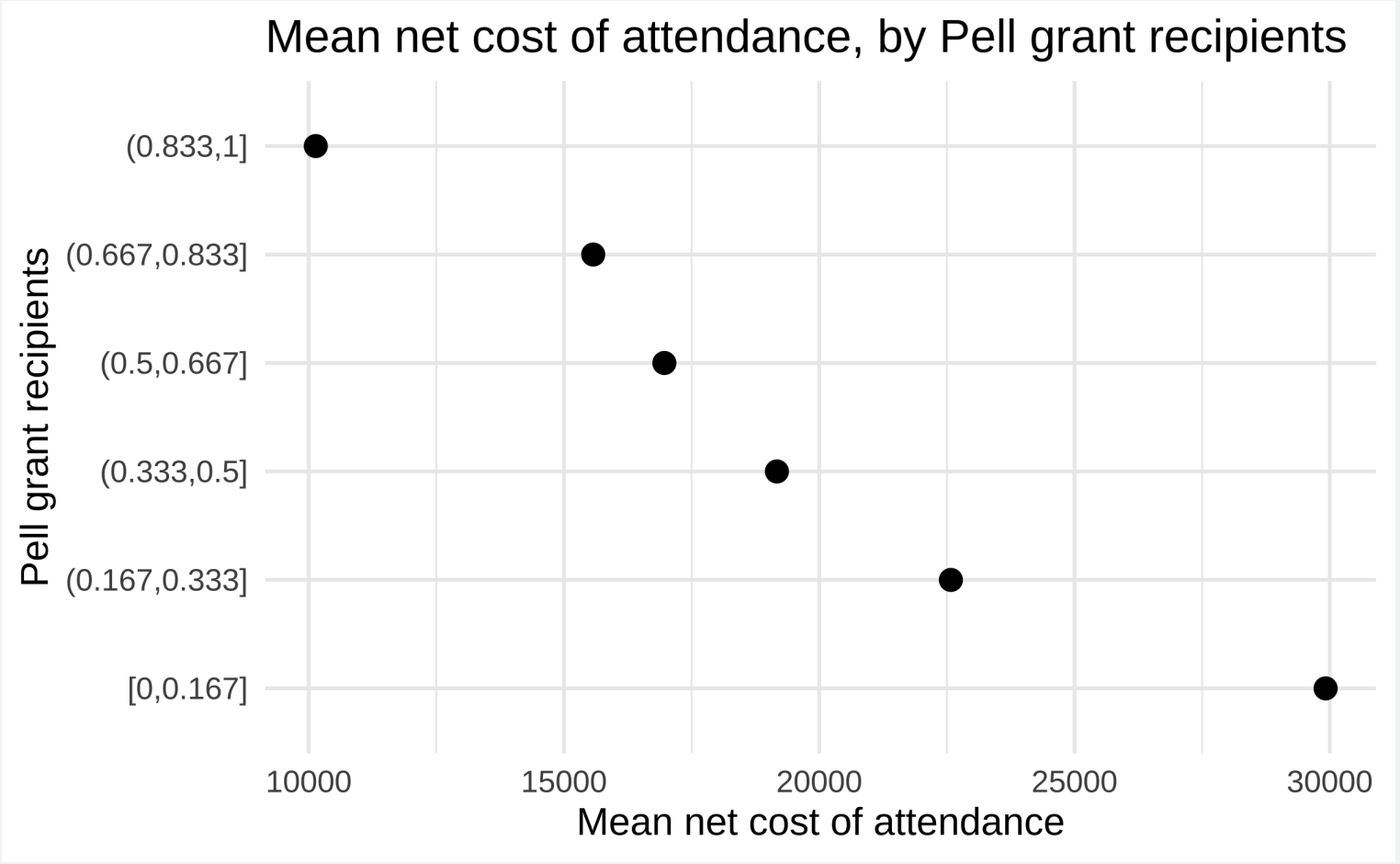
Which of the plots has higher data-to-ink ratio?

Plot A | Plot B



Which of the plots has higher data-to-ink ratio?

Plot A | Plot B



a deeper look at the plotting code...

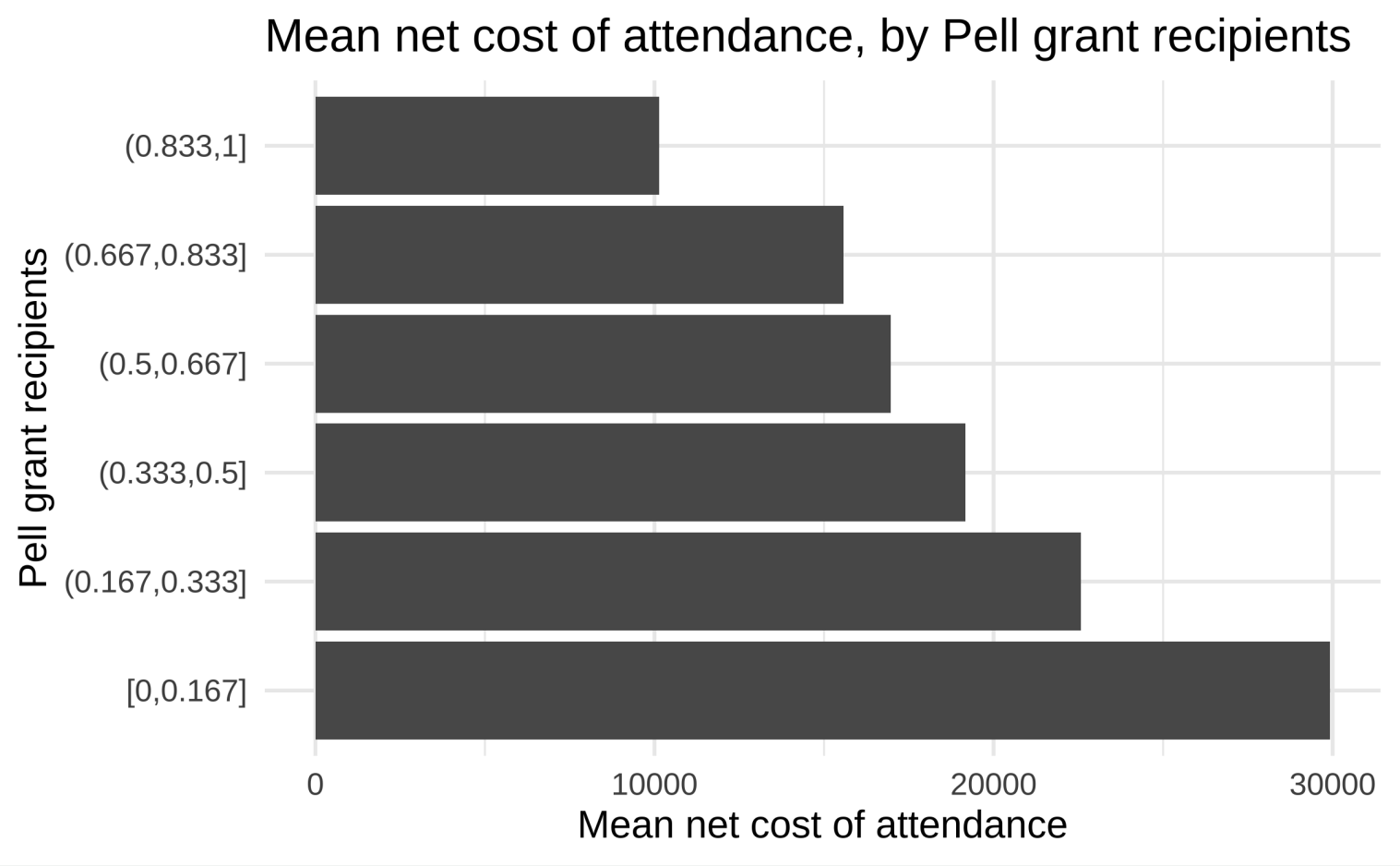
Summary statistics

```
mean_netcost_pctpell <- scorecard %>%
  group_by(pctpell_cat) %>%
  summarise(mean_netcost = mean(netcost, na.rm = TRUE))

mean_netcost_pctpell
```

```
## # A tibble: 6 × 2
##   pctpell_cat   mean_netcost
##   <fct>                <dbl>
## 1 [0,0.167]           29919.
## 2 (0.167,0.333]       22579.
## 3 (0.333,0.5]         19171.
## 4 (0.5,0.667]         16966.
## 5 (0.667,0.833]       15572.
## 6 (0.833,1]           10138.
```

Barplot

Plot Code



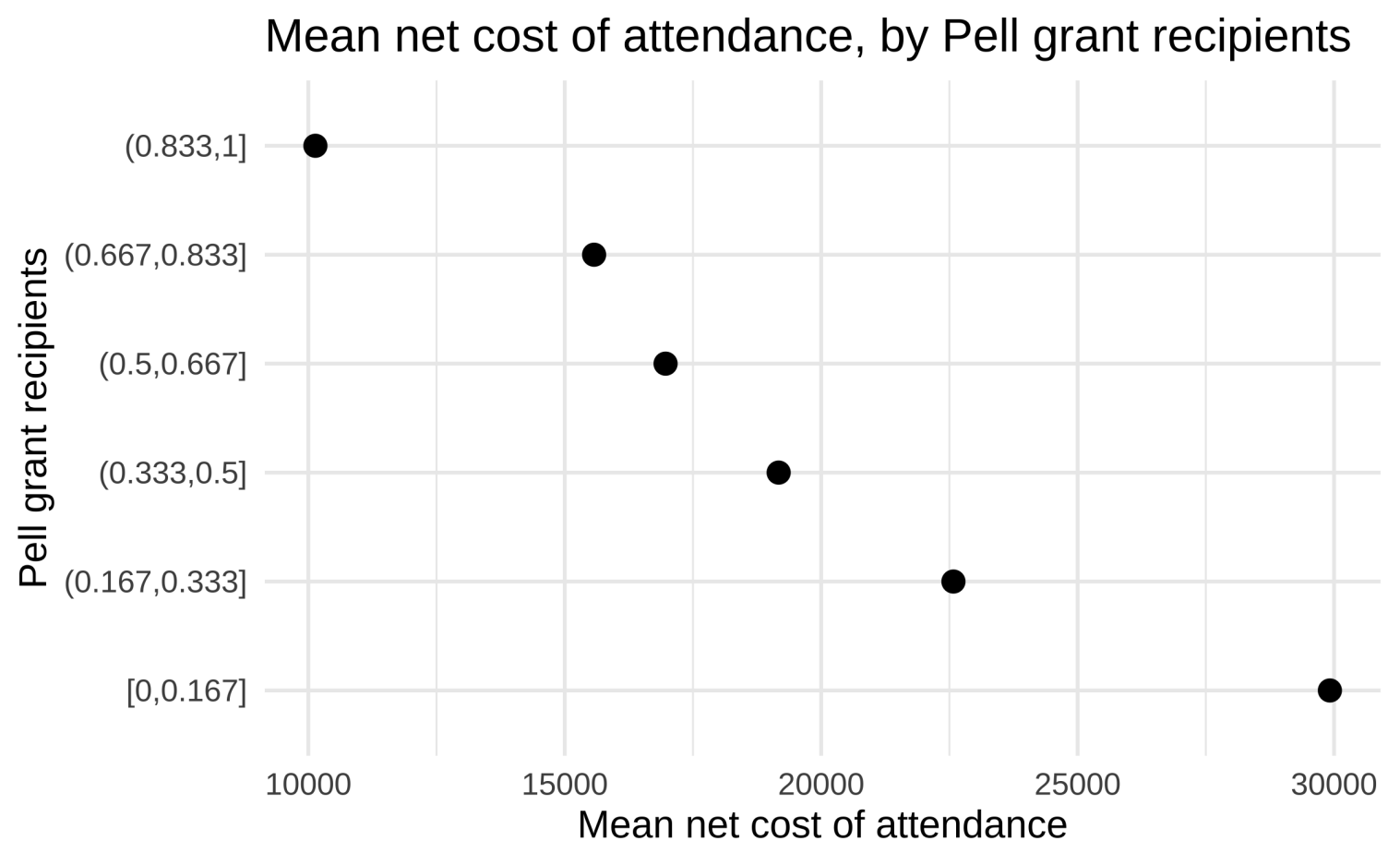
Barplot

Plot | Code

```
ggplot(mean_netcost_pctpell, aes(y = pctpell_cat, x = mean_netcost)) +
  geom_col() + #<<
  labs(
    x = "Mean net cost of attendance", y = "Pell grant recipients",
    title = "Mean net cost of attendance, by Pell grant recipients"
  ) +
  theme_minimal(base_size = 16)
```

Scatterplot

Plot Code



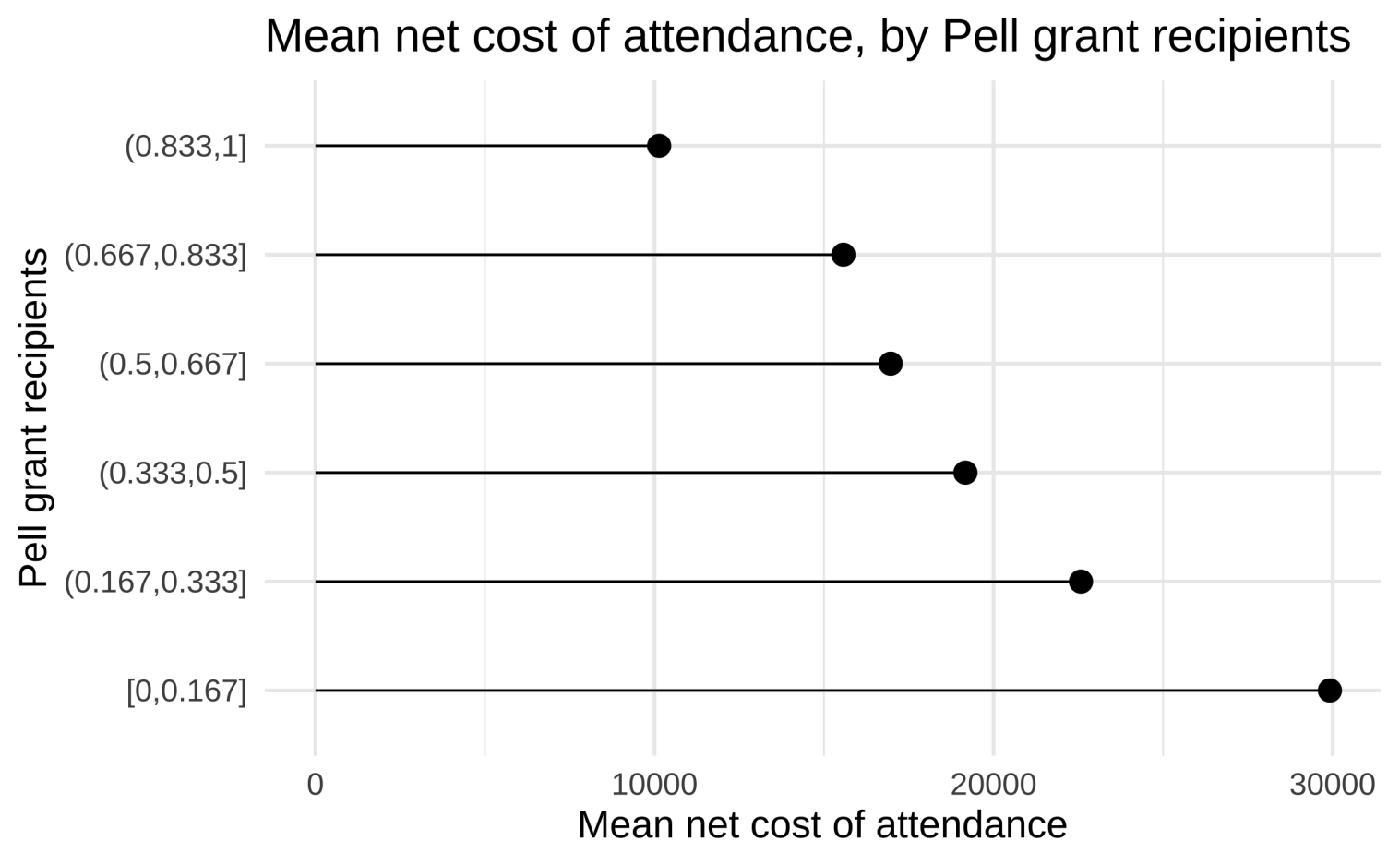
Scatterplot

Plot | Code

```
ggplot(mean_netcost_pctpell, aes(y = pctpell_cat, x = mean_netcost)) +
  geom_point(size = 4) + #<<
  labs(
    x = "Mean net cost of attendance", y = "Pell grant recipients",
    title = "Mean net cost of attendance, by Pell grant recipients"
  ) +
  theme_minimal(base_size = 16)
```

Lollipop plot -- a happy medium?

Plot Code



Lollipop plot -- a happy medium?

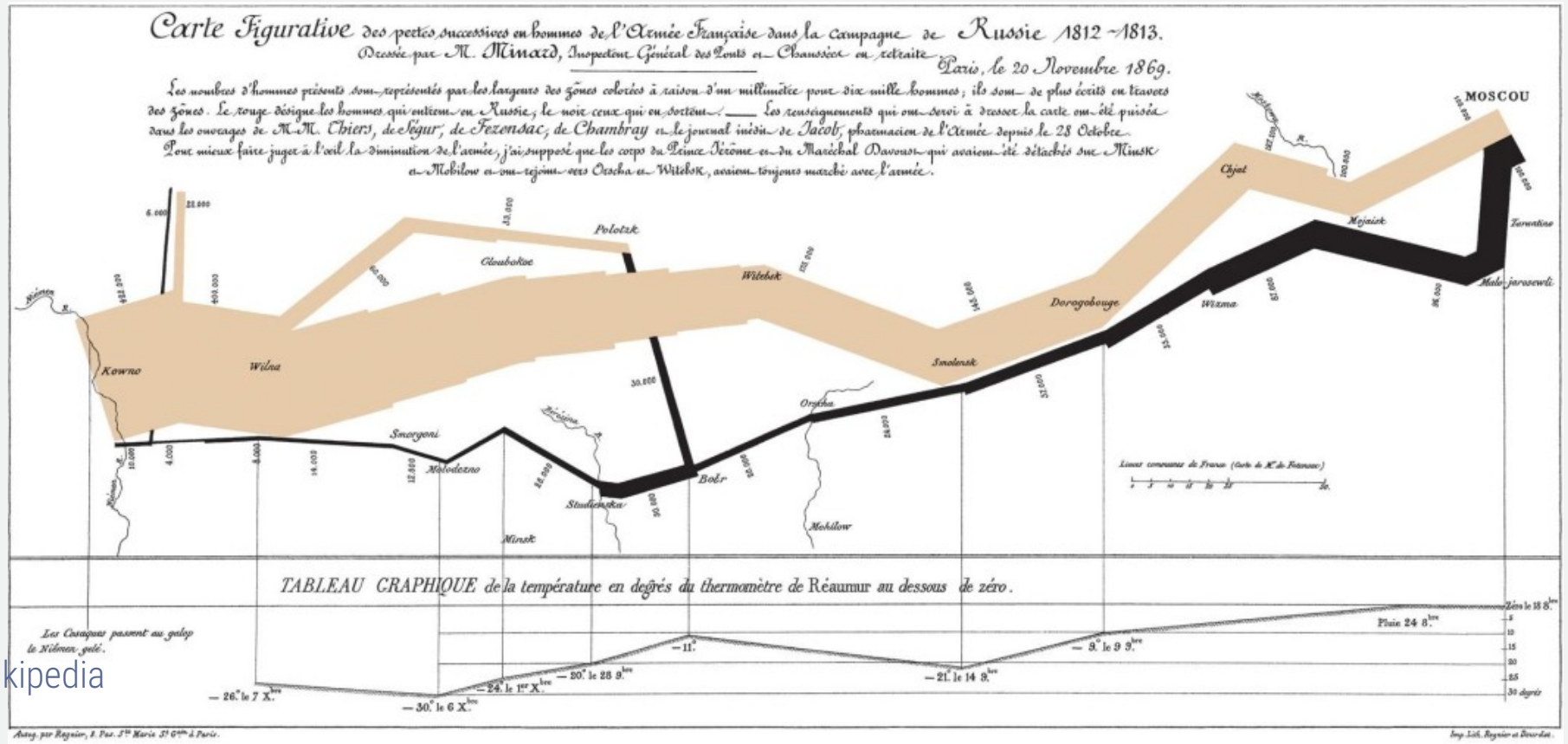
Plot | Code

```
ggplot(mean_netcost_pctpell, aes(y = pctpell_cat, x = mean_netcost)) +
  geom_point(size = 4) +
  geom_segment( #<<
    aes( #<<
      x = 0, xend = mean_netcost, #<<
      y = pctpell_cat, yend = pctpell_cat #<<
    ) #<<
  ) + #<<
  labs(
    x = "Mean net cost of attendance", y = "Pell grant recipients",
    title = "Mean net cost of attendance, by Pell grant recipients"
  ) +
  theme_minimal(base_size = 16)
```

Activity: Napoleon's retreat

05:00

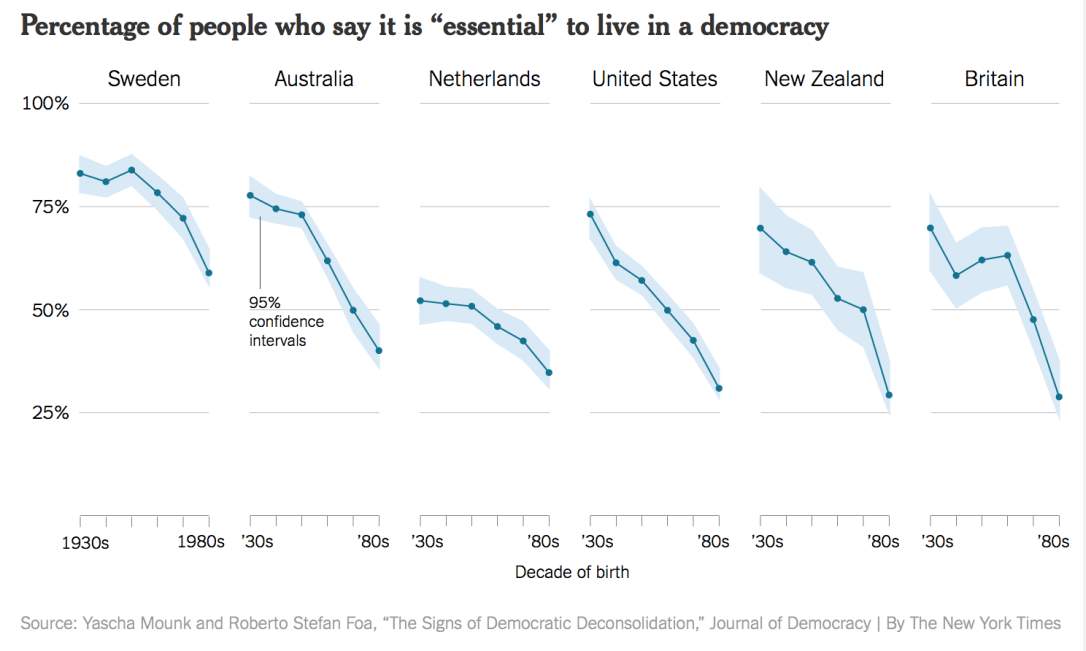
This is Minard's visualization of Napoleon's retreat. Discuss in a pair (or group) the features of the following visualization. What are the variables plotted? Which aesthetics are they mapped to?



Source: Wikipedia

Bad data

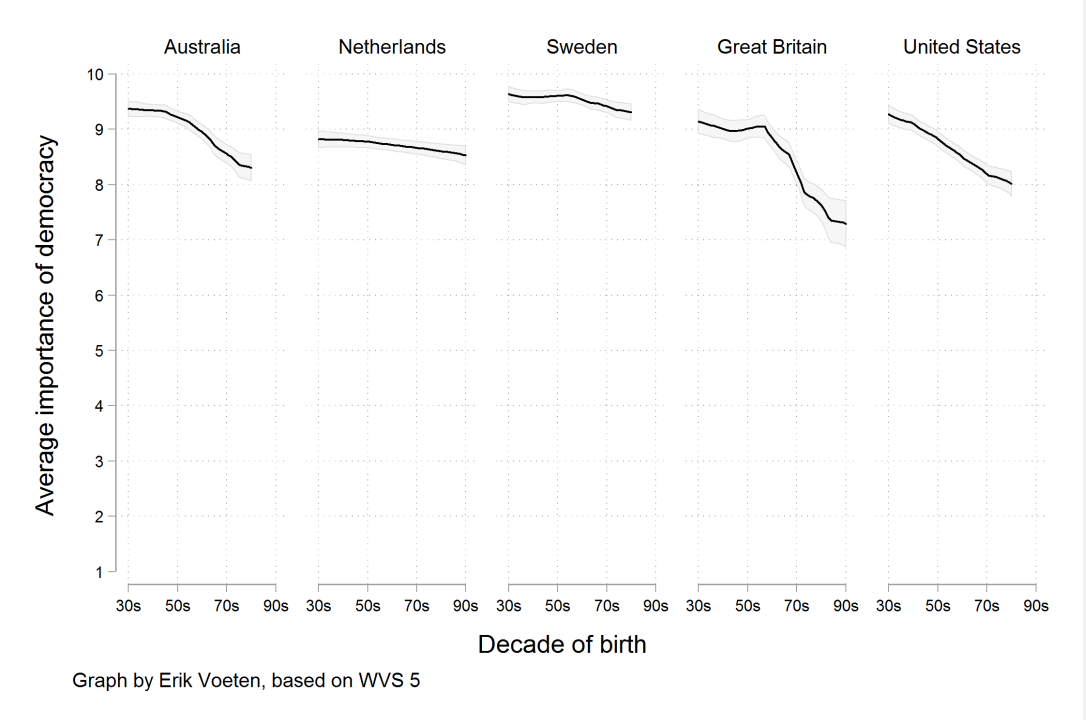
Original | Improved



Healy, Data Visualization: A practical introduction. Chapter 1. Figures 1.8 and 1.9.

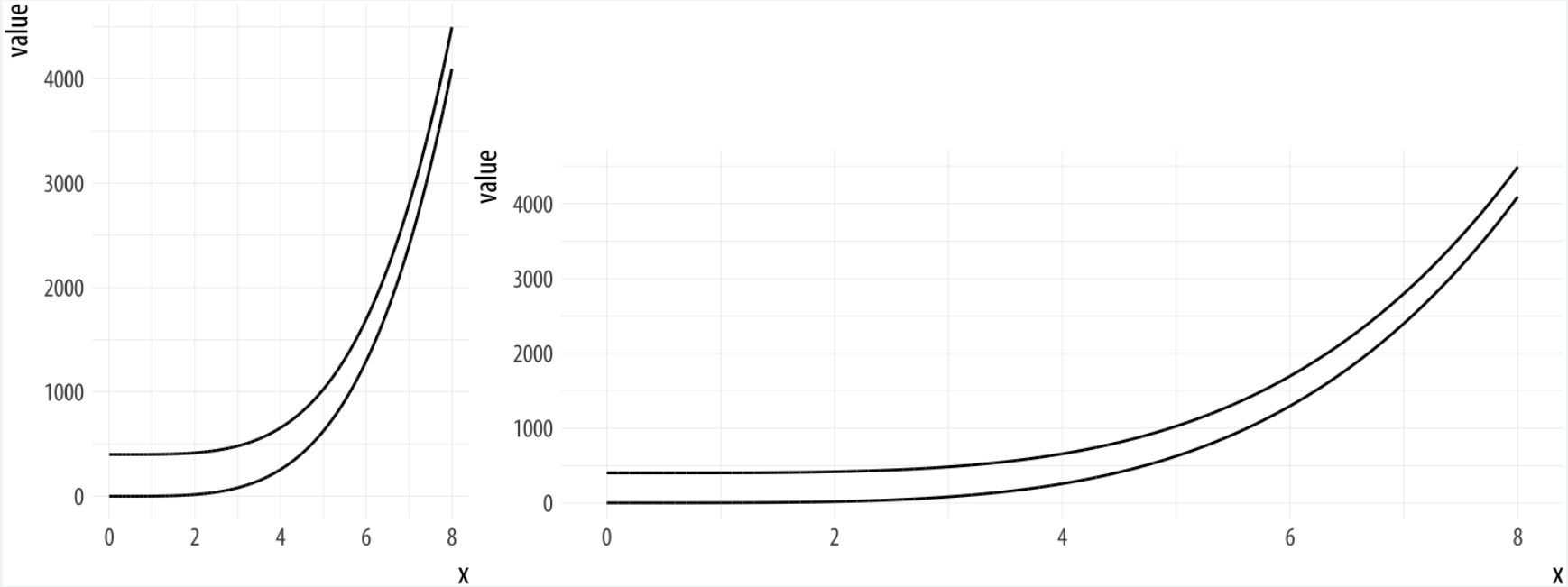
Bad data

Original | Improved



Healy, Data Visualization: A practical introduction. Chapter 1. Figures 1.8 and 1.9.

Bad perception

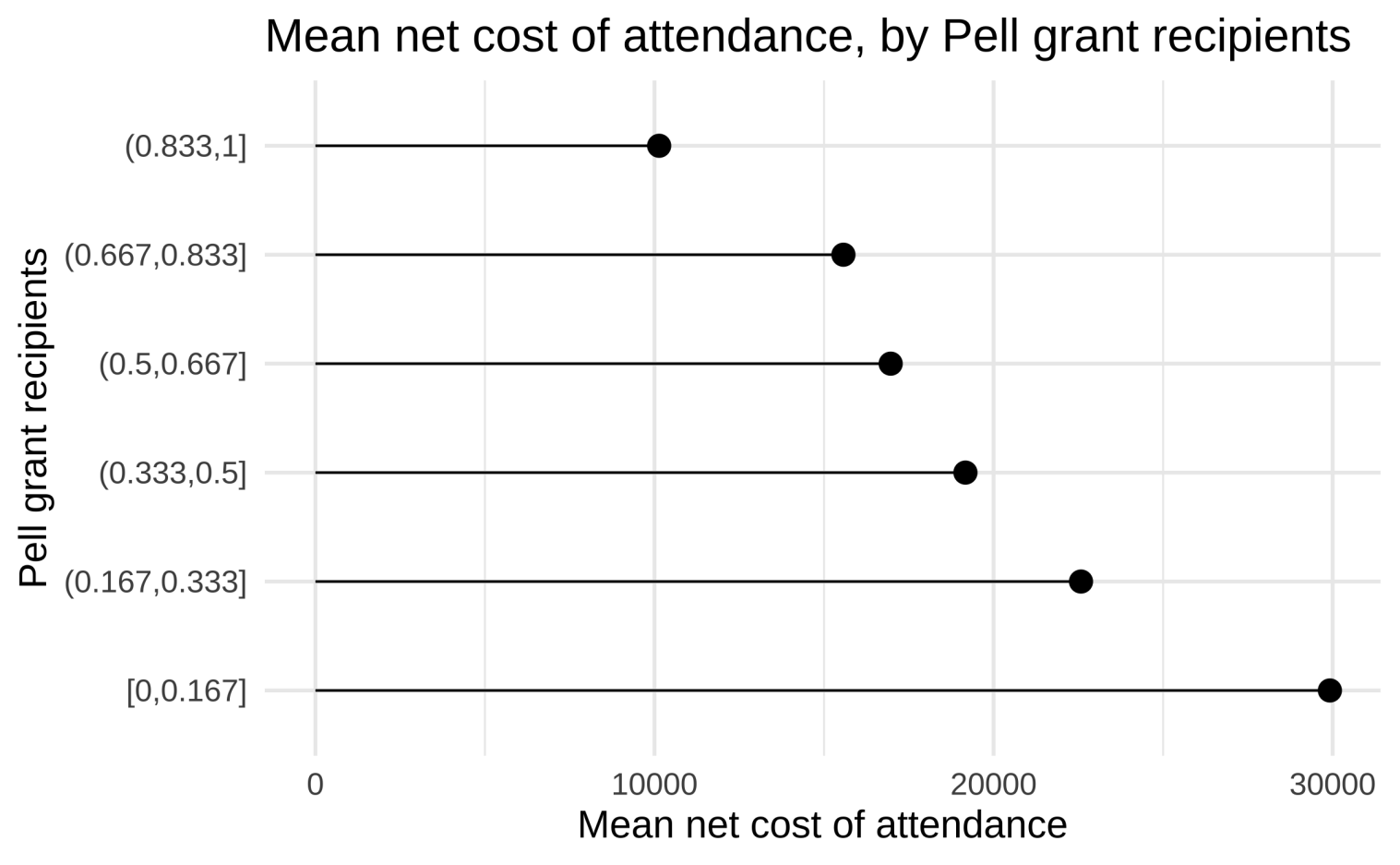


Healy, Data Visualization: A practical introduction. Chapter 1. Figure 1.12.

Aesthetic mappings in ggplot2

A second look: lollipop plot

Plot Code



A second look: lollipop plot

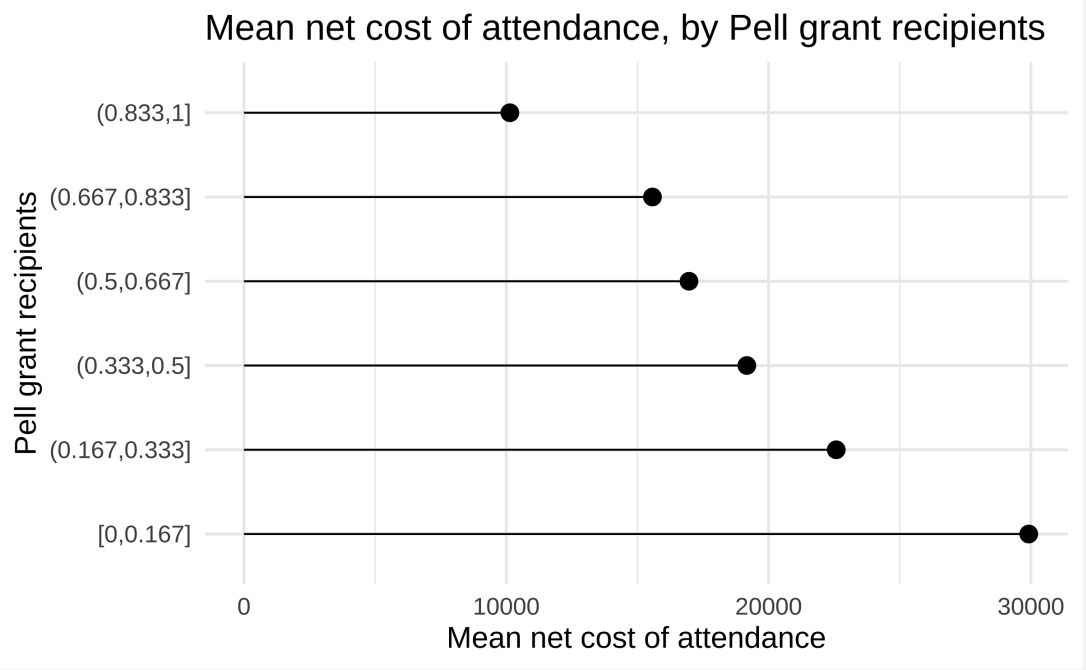
Plot | Code

```
ggplot(mean_netcost_pctpell, aes(y = pctpell_cat, x = mean_netcost)) +
  geom_point(size = 4) +
  geom_segment(
    aes(
      x = 0, xend = mean_netcost,
      y = pctpell_cat, yend = pctpell_cat
    )
  ) +
  labs(
    x = "Mean net cost of attendance", y = "Pell grant recipients",
    title = "Mean net cost of attendance, by Pell grant recipients"
  ) +
  theme_minimal(base_size = 16)
```

Activity: Spot the difference I

Can you spot the differences between the code here and the one provided in the previous slide? Are there any differences in the resulting plot? Work in a pair (or group) to answer.

Plot | Code



03:00

Activity: Spot the difference I

Can you spot the differences between the code here and the one provided in the previous slide? Are there any differences in the resulting plot? Work in a pair (or group) to answer.

Plot | Code

```
ggplot(mean_netcost_pctpell, aes(y = pctpell_cat, x = mean_netcost)) +
  geom_point(size = 4) +
  geom_segment(aes(
    xend = 0,
    yend = pctpell_cat
  )) +
  labs(
    x = "Mean net cost of attendance", y = "Pell grant recipients",
    title = "Mean net cost of attendance, by Pell grant recipients"
  ) +
  theme_minimal(base_size = 16)
```

03:00

Global vs. layer-specific aesthetics

- Aesthetic mappings can be supplied in the initial `ggplot()` call, in individual layers, or in some combination of both.
- Within each layer, you can add, override, or remove mappings.
- If you only have one layer in the plot, the way you specify aesthetics doesn't make any difference. However, the distinction is important when you start adding additional layers.

Wrap up

Think back to all the plots you saw in the lecture, without flipping back through the slides. Which plot first comes to mind? Describe it in words.