Informatics 143

Information Visualization

Lecture 8

Duplication of course material for any commercial purpose without the explicit written permission of the professor is prohibited.

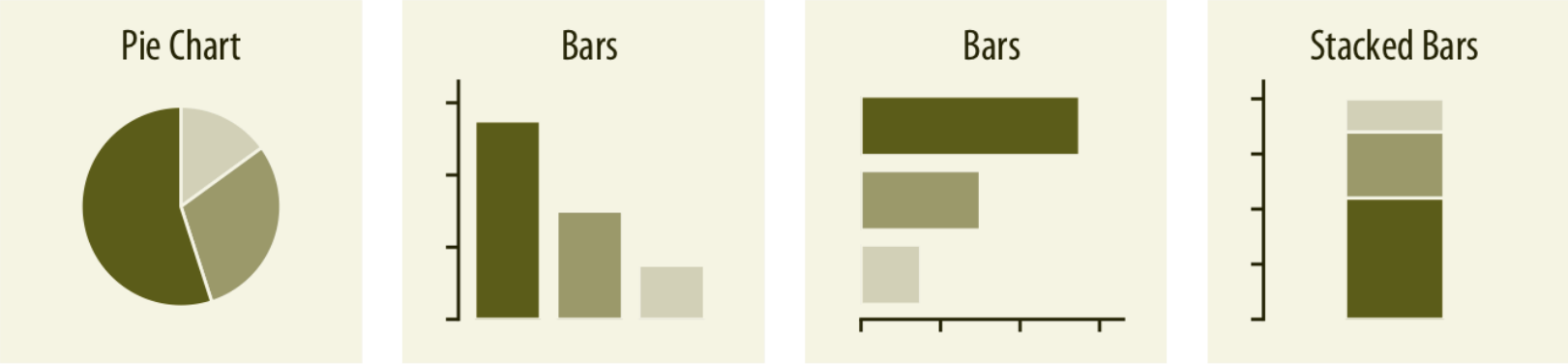
These course materials are based on books from Claus O. Wilke, Kieran Healy, Edward R. Tufte, Alberto Cairo, Colin Ware, Tamara Munzner, and others.
Powerpoint theme by Prof. André van der Hoek.

Visualization of proportions

- Useful to visualize how some groups or amount split into individual pieces that sum to a whole
  - E.g. proportions of gender in a group, votes in elections, market shares, supercomputing technologies, etc.
- Data has:
  - At least one set of values (qualitative or quantitative) *that add to a whole*
- *Some* standard geometrical mappings:
  - Pie charts
  - Bars, stacked bars and stacked densities
  - Mosaic plots
  - Treemaps
  - Parallel sets

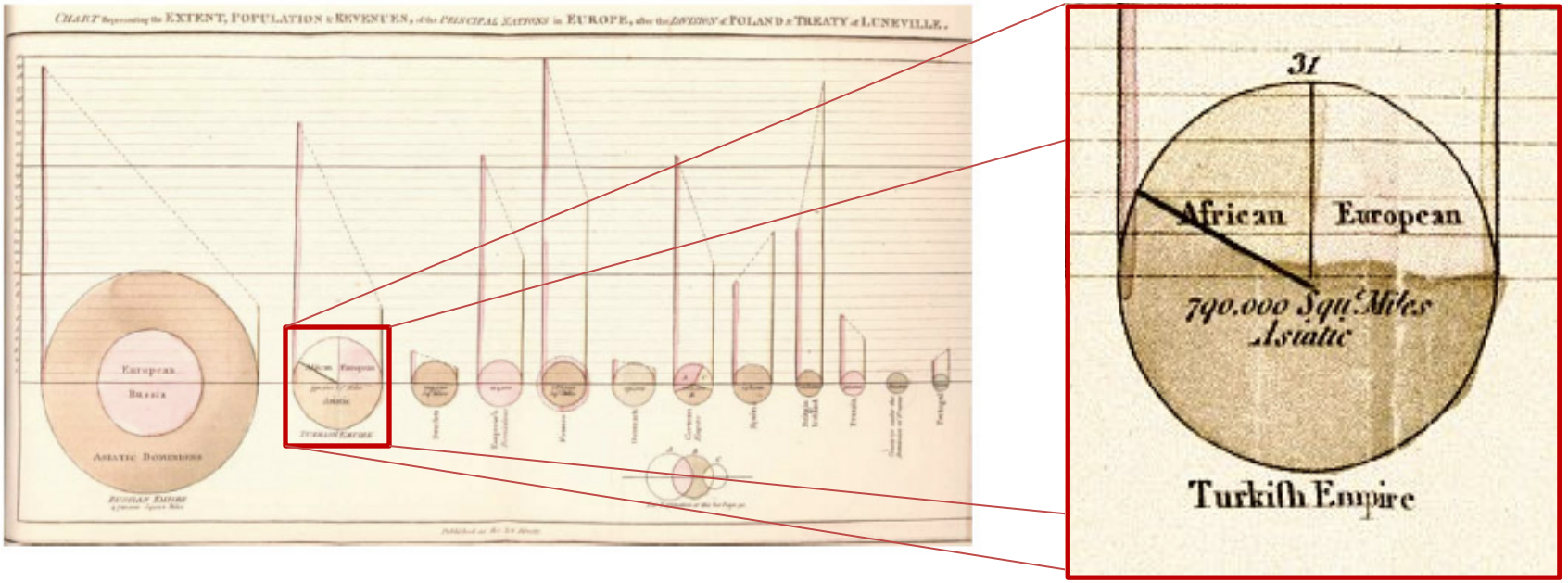
Visualization of proportions

- Individual groups/sets



Visualization of proportions

- Pie charts alternative name: circle charts
- Earliest known 1801 (William Playfair)



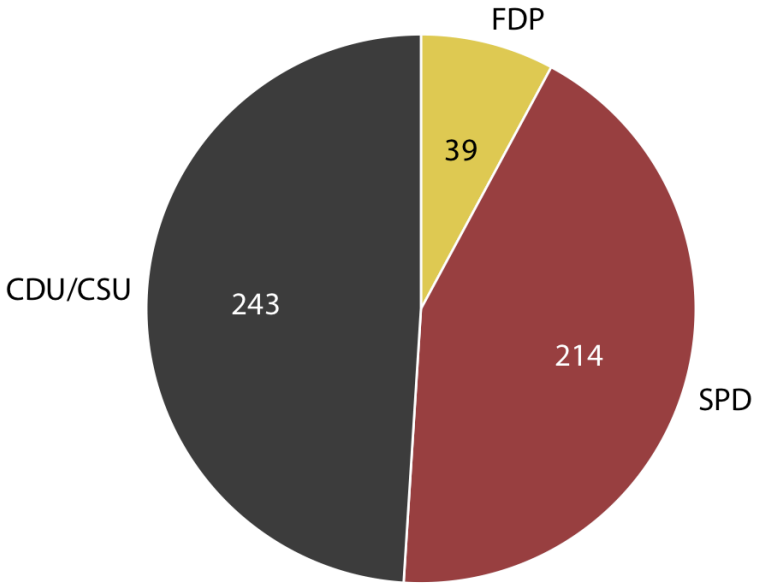
Visualization of proportions

- How to read it?
  - Suppose that you have a data set of a parliament composed by three parties. Each political party has a certain number of seats.
  - A pie chart visual representation is created when you use a circle and divide it in slices **respecting the proportions** set by the proportions of the seats in the entire parliament.

Data source: Fundamentals of Data Visualization

Visualization of proportions

- How to read it?
  - Suppose that you have a data set of a parliament composed by three parties. Each political party has a certain number of seats.
  - A pie chart visual representation is created when you use a circle and divide it in slices respecting the proportions set by the proportions of the seats in the entire parliament.



Data source: Fundamentals of Data Visualization

Visualization of proportions

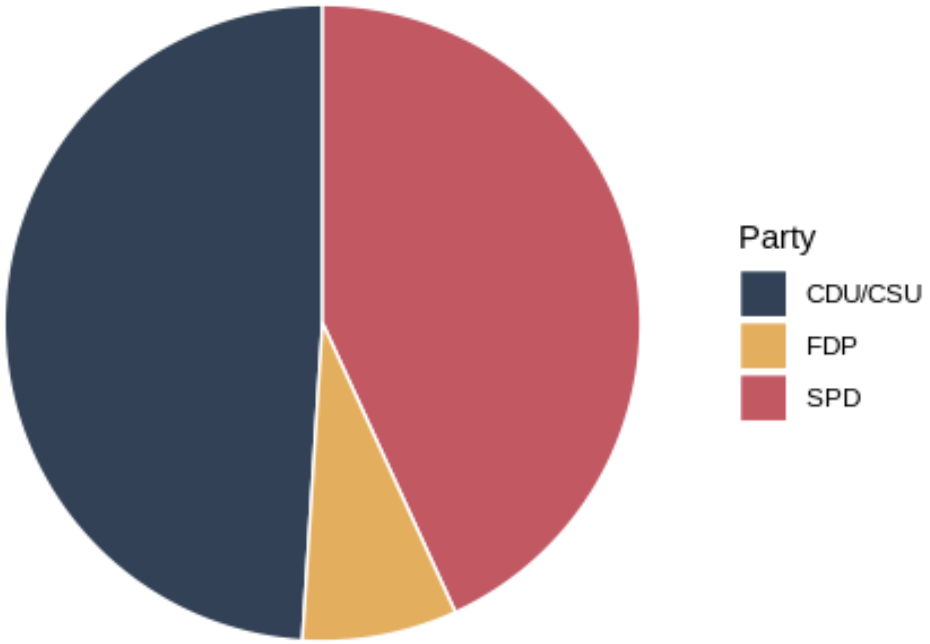
- How to build a pie chart in ggplot2?
  - **There is not a geom for it!**
  - In baseline R you can use pie(). In ggplot we can use a trick: *make a bar chart and transform it to polar coordinates.*

```
bundesdata <- data.frame(
  Party=c("CDU/CSU", "SPD", "FDP"),
  value=c(243, 214, 39)
)

ggplot(bundesdata, aes(x="", y=value, fill=Party)) +
  geom_bar(stat="identity", width=1, color="white") +
  coord_polar("y", start=0) +
  scale_fill_manual(values = c("#2e4057", "#edae49", "#d1495b")) +
  theme_void()
```

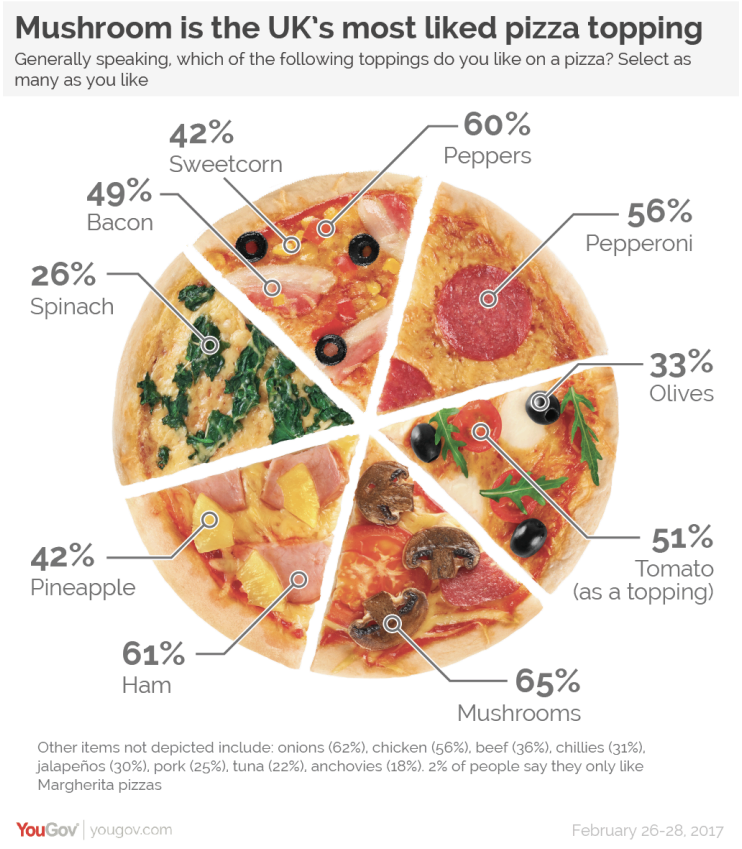
Visualization of proportions

- How to build a pie chart in ggplot2?
  - You can programatically calculate where to add labels to write the numbers of seats...



Visualization of proportions

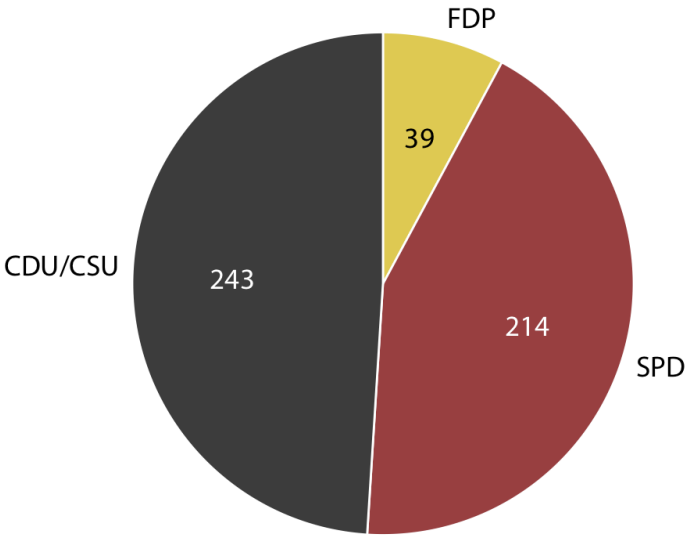
- Problems:
  - Hard for the mind to fully understand angles...
  - If talking about proportions, it should **never** go beyond 100%...
  - Do not use 3D: there is no additional information there...

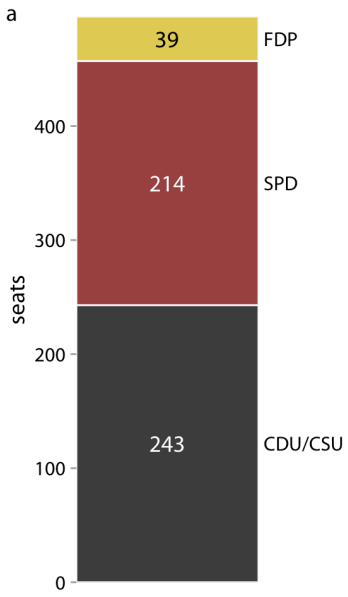


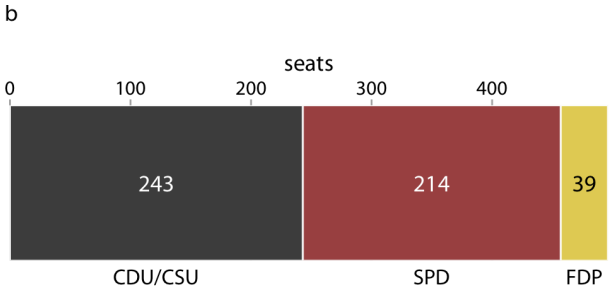
https://yougov.co.uk/topics/politics/articles-reports/2017/03/06/does-pineapple-belong-pizza

Visualization of proportions

- How to read it?
  - The area of each slice is proportional to the fraction of the total
  - **The same strategy can be used in stacked bars. *Are they better?***



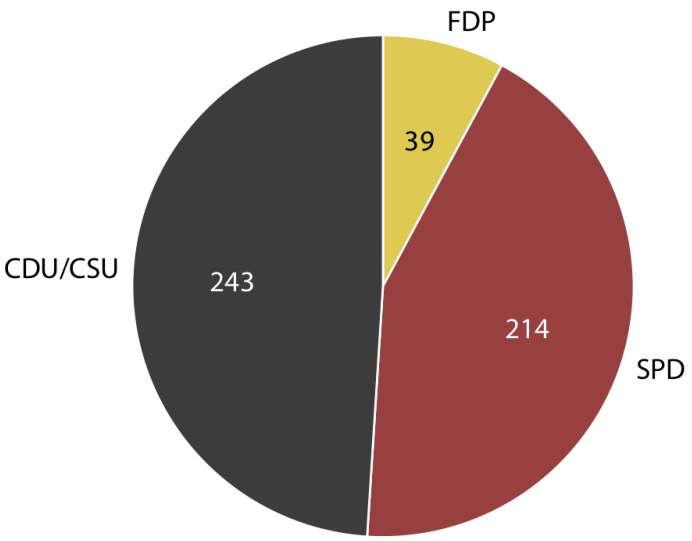


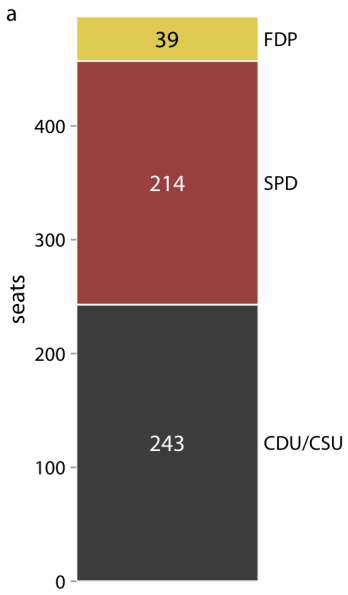


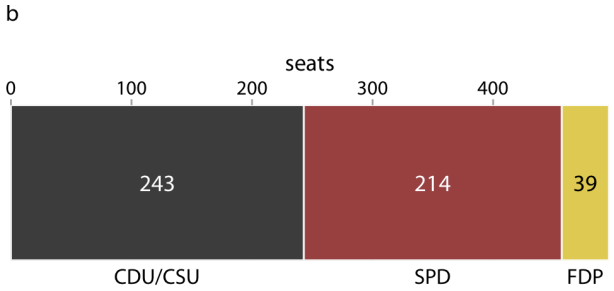
Data source: Fundamentals of Data Visualization

Visualization of proportions

- How to read it?
  - The area of each slice is proportional to the fraction of the total
  - The same strategy can be used in stacked bars. *Are they better?*







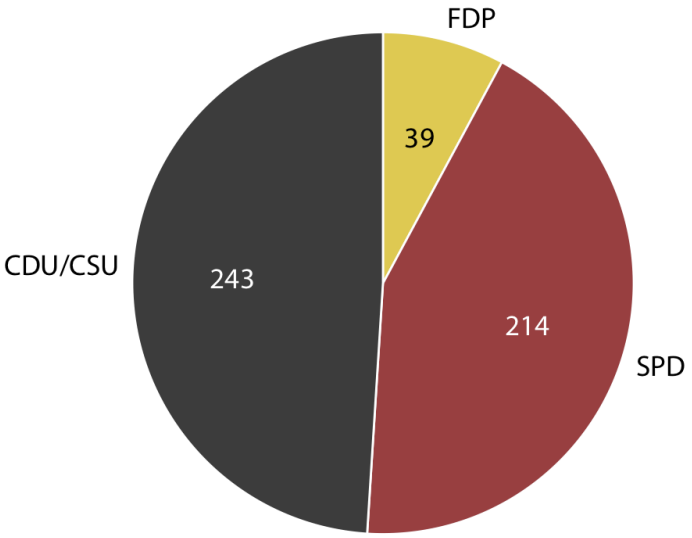
It is easier to see that SPD+FDP > CDU/CSU

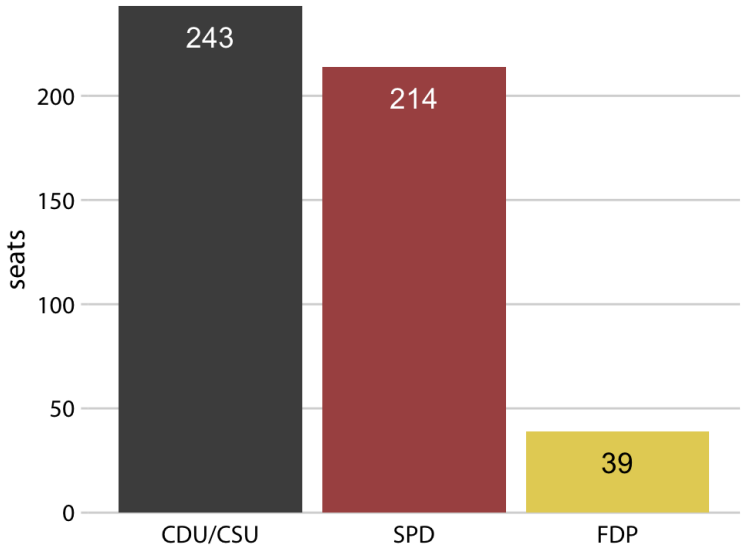
It is not obvious that SPD+FDP > CDU/CSU

Data source: Fundamentals of Data Visualization

Visualization of proportions

- How to read it?
  - The area of each slice is proportional to the fraction of the total
  - The same strategy can be used in side-by-side bars. *Are they better?*





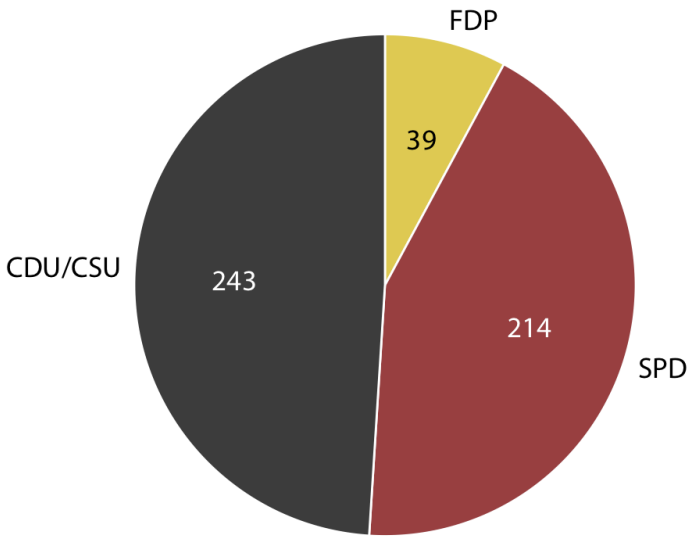
It is easier to see that SPD+FDP > CDU/CSU

It is not obvious that SPD+FDP > CDU/CSU

Data source: Fundamentals of Data Visualization

Visualization of proportions

- How to read it?



It is usually ok if you wish to visualize SMALL DATA SETS

Data source: Fundamentals of Data Visualization

Visualization of proportions

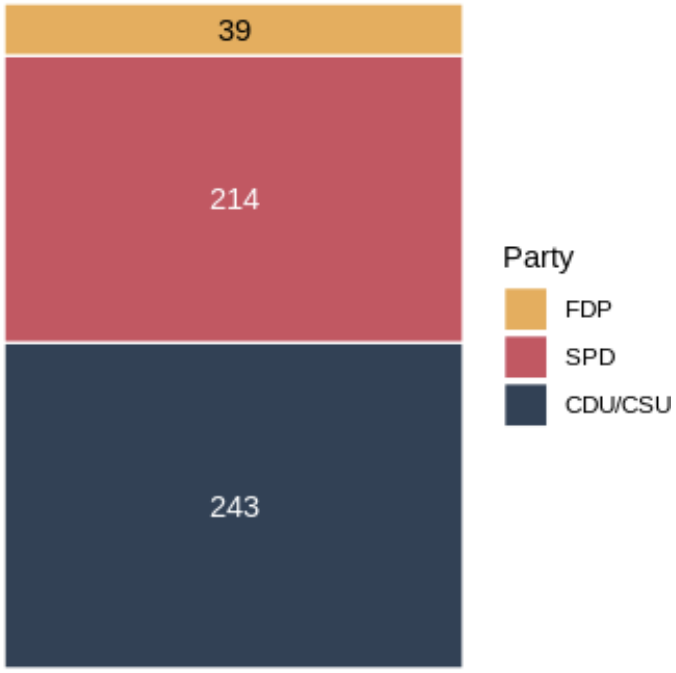
- How to build a a simple stack in ggplot2?
  - Simply use geom_col(position="stack")

```
bundesdata <- data.frame(
  Party=c("CDU/CSU", "SPD", "FDP"),
  value=c(243, 214, 39)
)

ggplot(bundesdata, aes(x = 1, y = value,
       fill = factor(Party, levels = rev(Party)))) +
  geom_col(position = "stack", color = "white") +
  geom_text(aes(x = 1.,
            y = (cumsum(data$value) - data$value/2), label = value),
            color = c("white", "white", "black")) +
  scale_fill_manual("Party",
    values = c("#edae49", "#d1495b", "#2e4057")) +
  theme_void()
```

Visualization of proportions

- How to build a a simple stack in ggplot2?
  - Simply use geom_col(position="stack")



Visualization of proportions

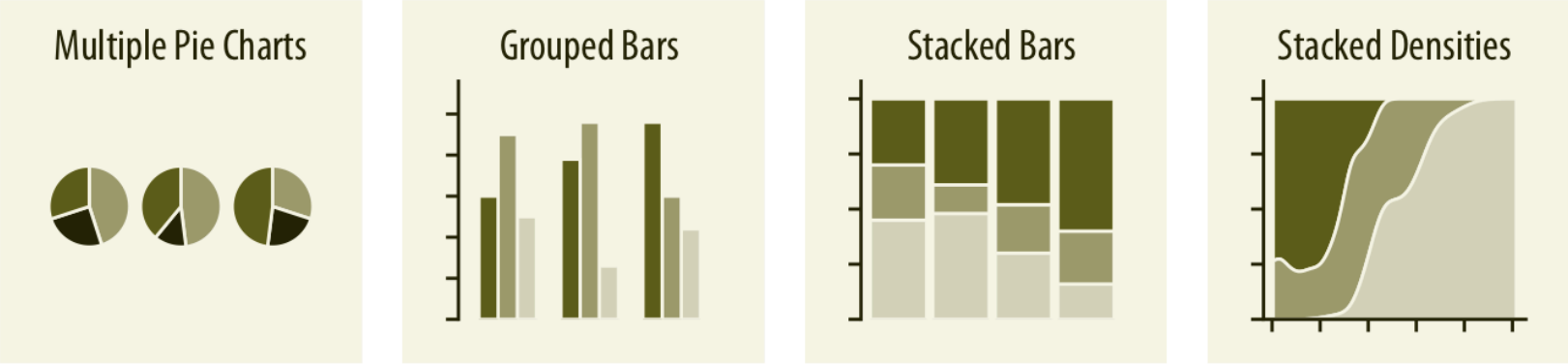
- Pie charts, stacked bars and side-by-side bars

	Pie chart	Stacked bars	Side-by-side bars
Clearly visualizes the data as proportions of a whole	✔	✔	✖
Allows easy visual comparison of the relative proportions	✖	✖	✔
Visually emphasizes simple fractions, such as 1/2, 1/3, 1/4	✔	✖	✖
Looks visually appealing even for very small datasets	✔	✖	✔
Works well when the whole is broken into many pieces	✖	✖	✔
Works well for the visualization of many sets of proportions or time series of proportions	✖	✔	✖

Data source: Fundamentals of Data Visualization

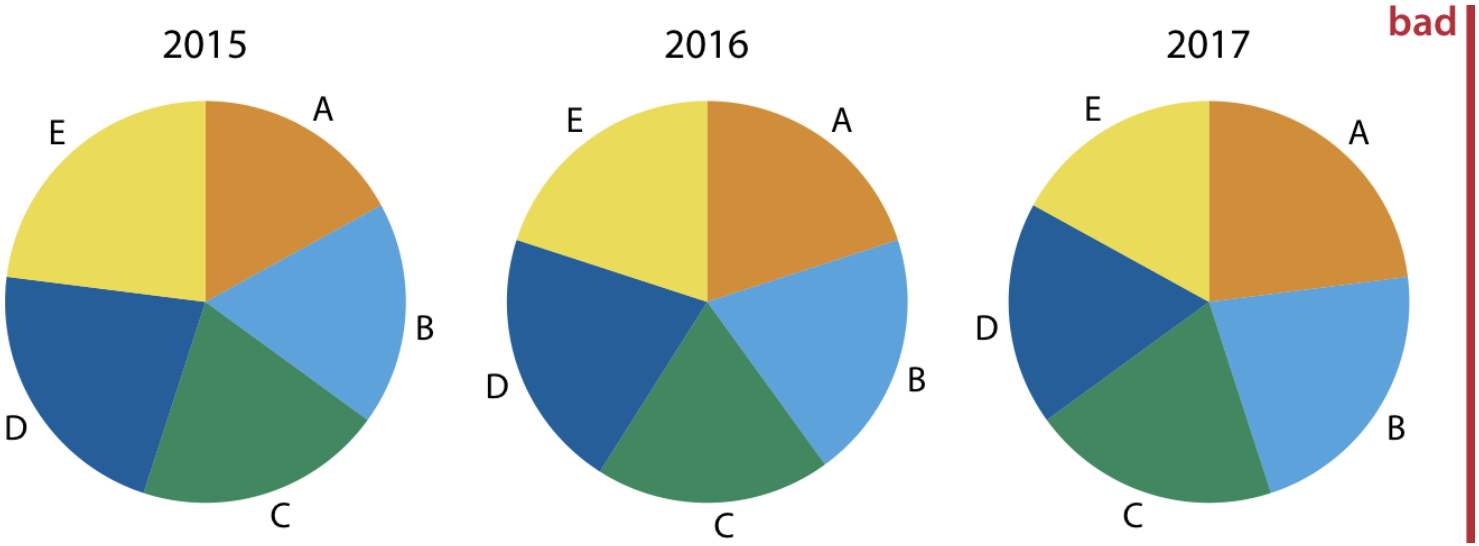
Visualization of Proportions

- Multiple groups/sets



Visualization of proportions: pie charts

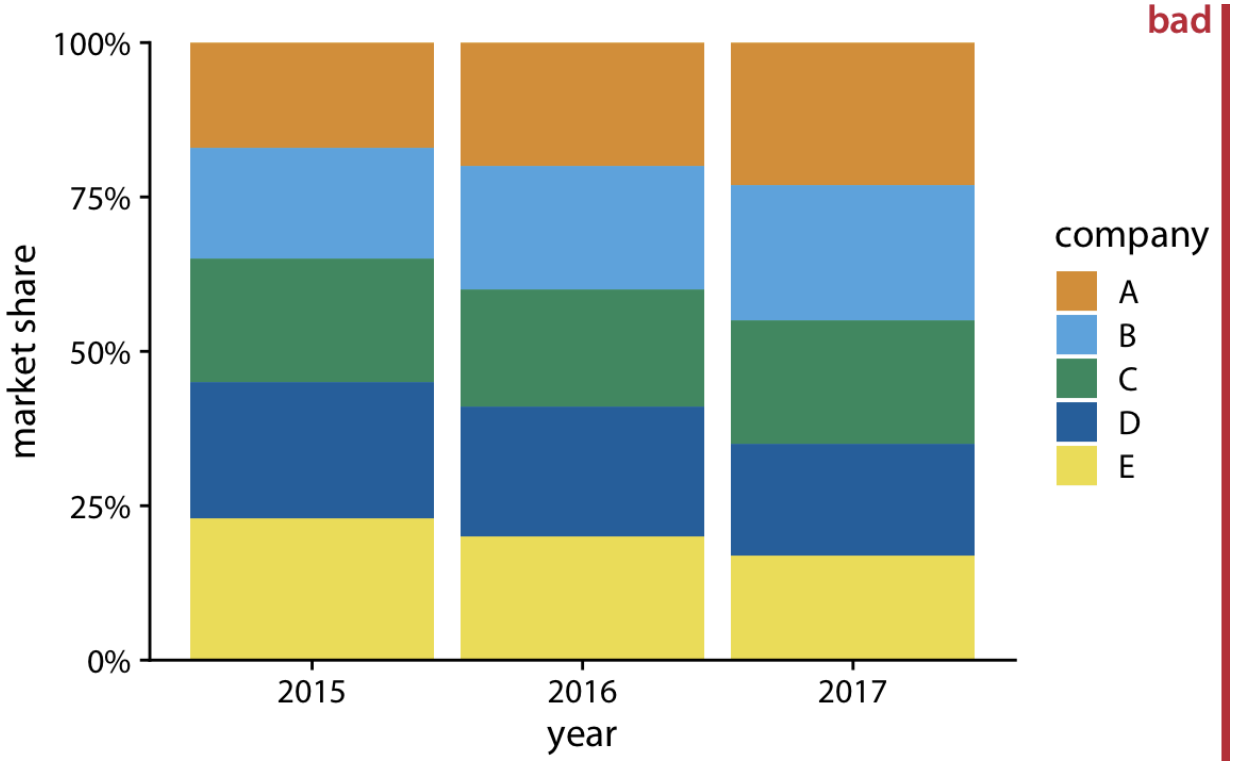
- Can you identify any evolution?



Data source: Fundamentals of Data Visualization

Visualization of proportions: stacked bars

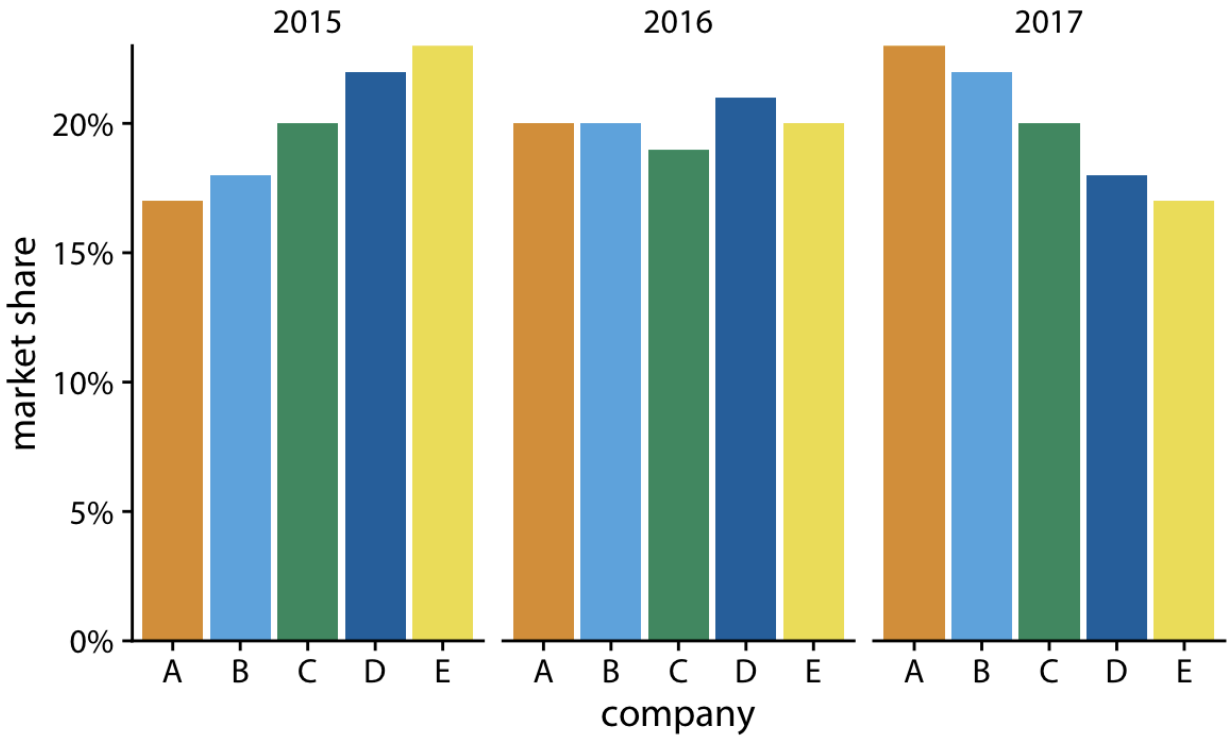
- Can you clearly see the evolution of B, C and D?
- Can you tell the relative market shares in each year?



Data source: Fundamentals of Data Visualization

Visualization of proportions : side-by-side bars

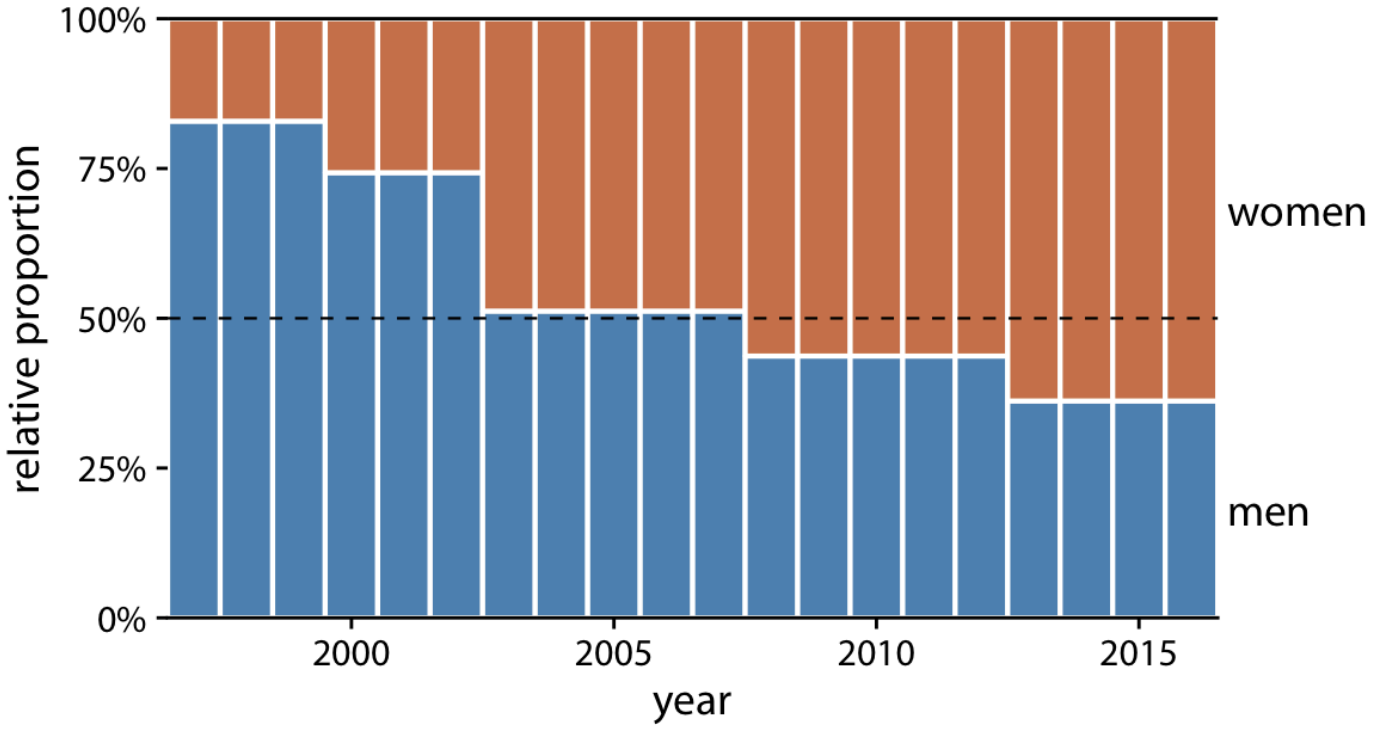
- Clearly shows that A & B increased market share
- Clearly shows that D & E reduced market share



Data source: Fundamentals of Data Visualization

Visualization of proportions

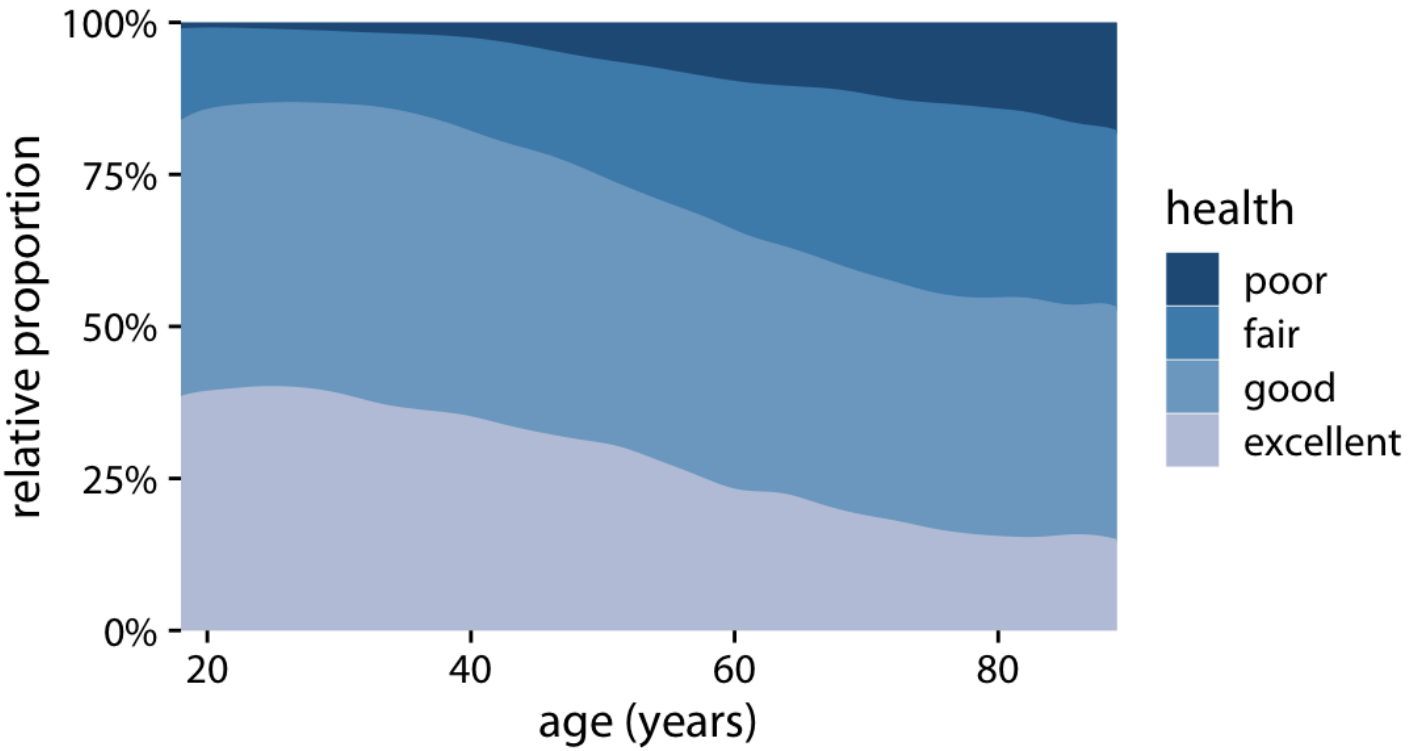
- Series of stacked bars are ok if there are only two categories in the dataset



Data source: Inter-Parliamentary Union (IPU), ipu.org.

Visualization of proportions

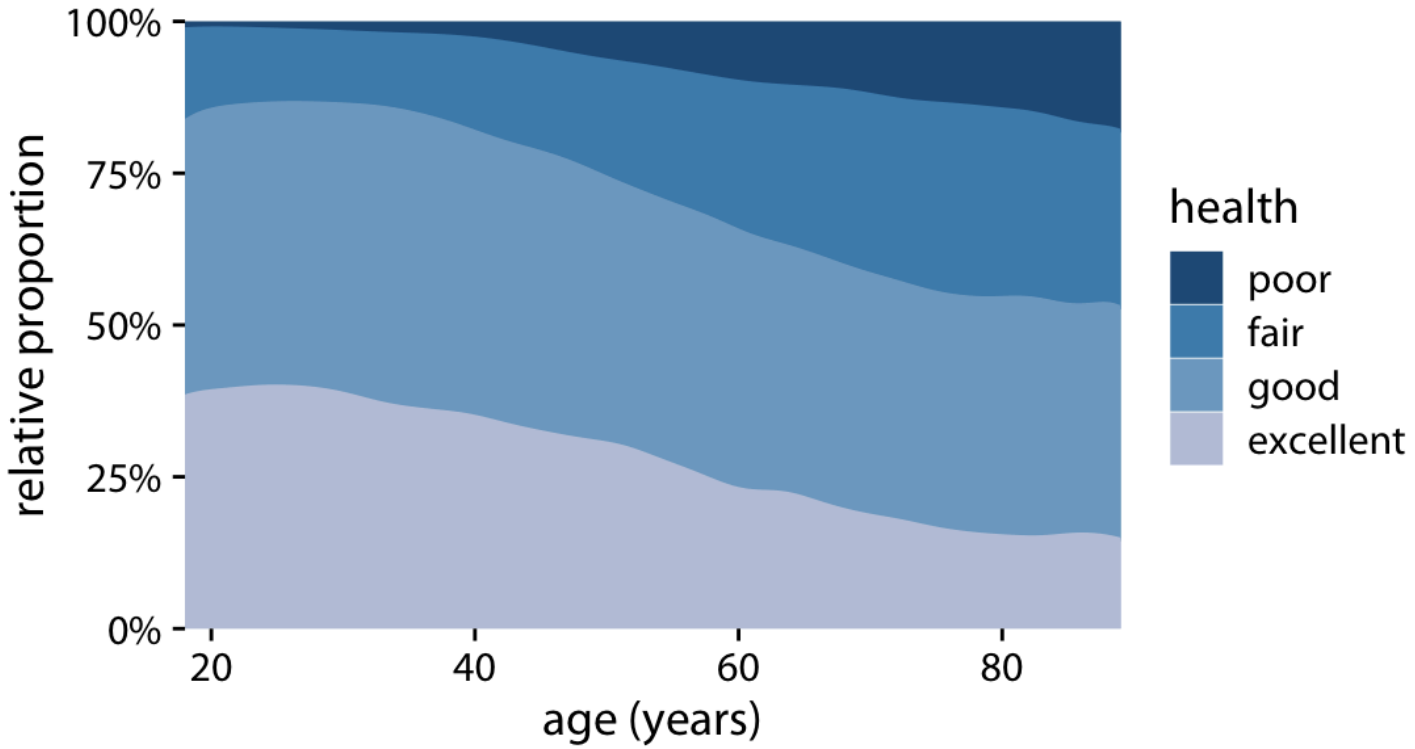
- Stacked-densities:
  - How proportions change with a continuous variable
  - Obtained from a kernel density estimation



Data source: general social survey (GSS).

Visualization of proportions

- Stacked-densities:
  - How proportions change with a continuous variable
  - Obtained from a kernel density estimation



The figure hides that there are many more young people than old people.

Data source: general social survey (GSS).

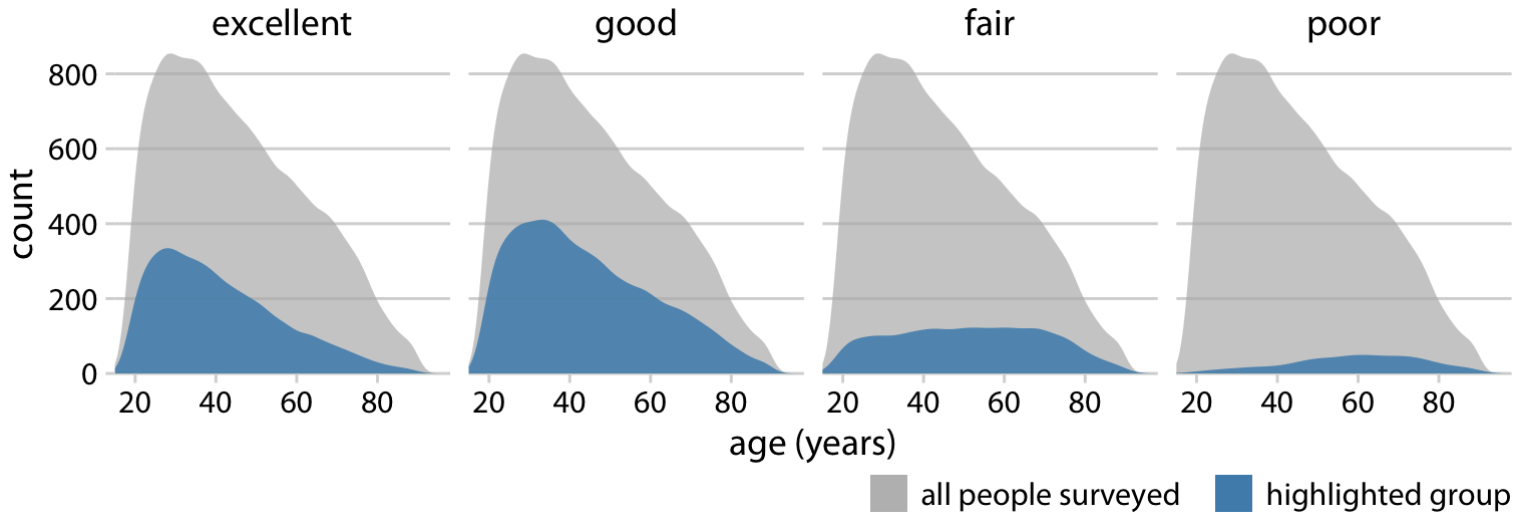
Visualization of proportions

- Multi-pannel plots using a shared element: the total

Data source: general social survey (GSS).

Visualization of proportions

- Multi-pannel plots using a shared element: the total



Only works for a small number of categories

Data source: general social survey (GSS).

Visualization of proportions

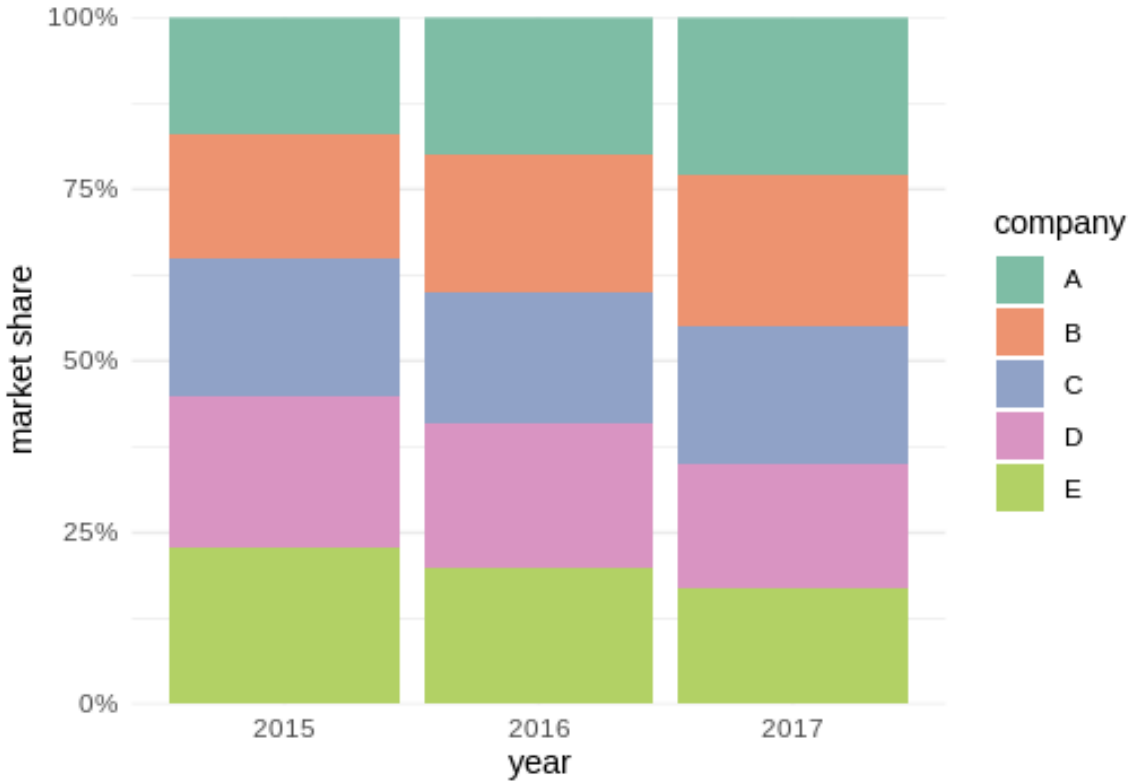
- How to build a series of stacked bars with two categories?
  - You can simply use geom_col() and set the fill aesthetics

```
marketshare <- data.frame(
                percent = c(17, 18, 20, 22, 23, 20, 20, 19, 21, 20, 23, 22, 20, 18, 17),
                company = rep(LETTERS[1:5], 3),
                year = rep(c("2015", "2016", "2017"), each = 5))

ggplot(marketshare, aes(x = year, y = percent, fill = company)) +
  geom_col(position = "stack") +
  scale_y_continuous(
    name = "Market share",
    labels = scales::percent_format(accuracy = 1, scale = 1),
    expand = c(0, 0)
  ) + xlab("Year")
  scale_fill_brewer(palette = "Set2") +
  theme_minimal()
```

Visualization of proportions

- How to build a series of stacked bars with two categories?
  - You can simply use geom_col() and set the fill aesthetics



Visualization of proportions

- How to build a side by side bars?
  - Use facet_wrap() together with the col geometry.

```
ggplot(marketshare, aes(x = company, y = percent, fill = company)) +
    geom_col() +
    facet_wrap(~year) +
    scale_y_continuous(
      name = "market share",
      labels = scales::percent_format(accuracy = 1, scale = 1),
      expand = c(0, 0)) +
    scale_fill_brewer(palette = "Set2") +
    theme_minimal()
```

Visualization of proportions

- How to build a side by side bars?
  - Use facet_wrap() together with the col geometry.

Visualization of proportions

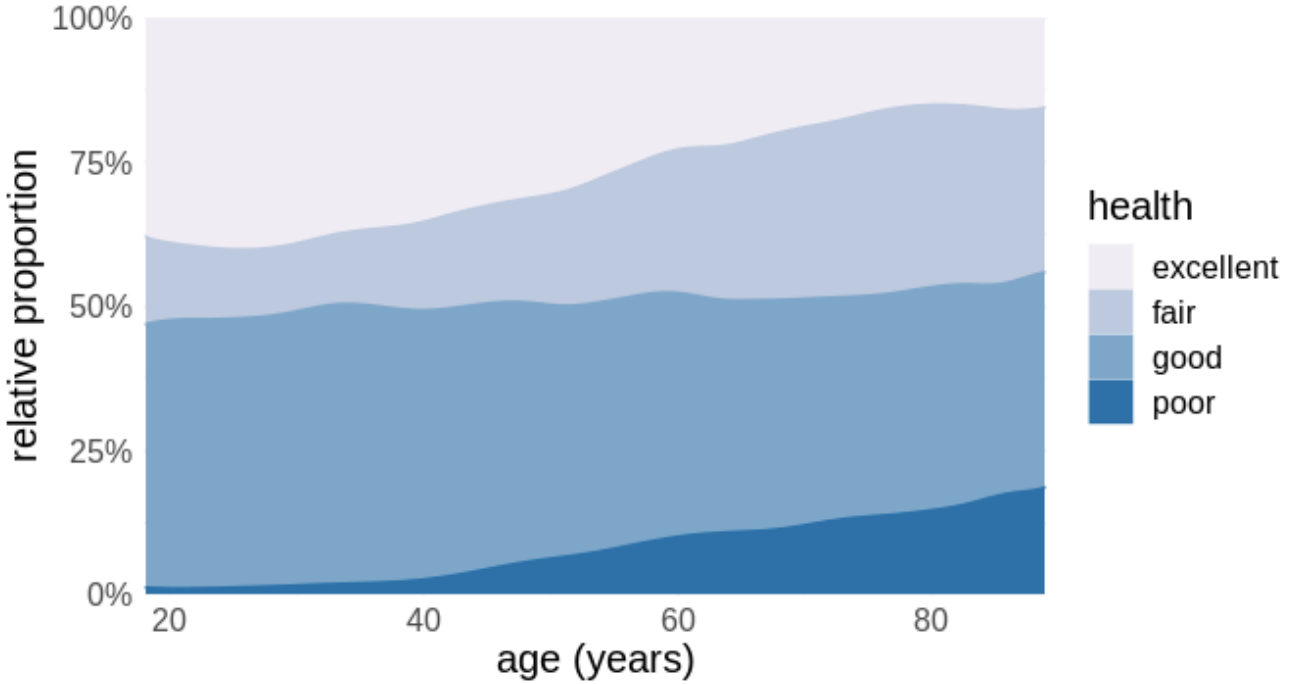
- How to build a stacked densities?
  - Just use geom_density(position = "fill") and set the fill and color mappings in the aes() call

```
happy <- na.omit(read.csv("http://www.ics.uci.edu/~algol/teaching/informatics143w2021/happy.csv"))

ggplot(happy,
       aes(x = age, y = ..count.., fill = health, color = health)) +
  geom_density(position = "fill") +
  scale_x_continuous(name = "age (years)", expand = c(0, 0)) +
  scale_y_continuous(
    expand = c(0, 0), name = "relative proportion",
    labels = scales::percent) +
  scale_fill_brewer(palette = "PuBu") +
  scale_color_brewer(palette = "PuBu") +
  theme_minimal() +
  theme(text = element_text(size=15))
```

Visualization of proportions

- How to build a stacked densities?
  - Just use geom_density(position = "fill") and set the fill and color mappings in the aes() call



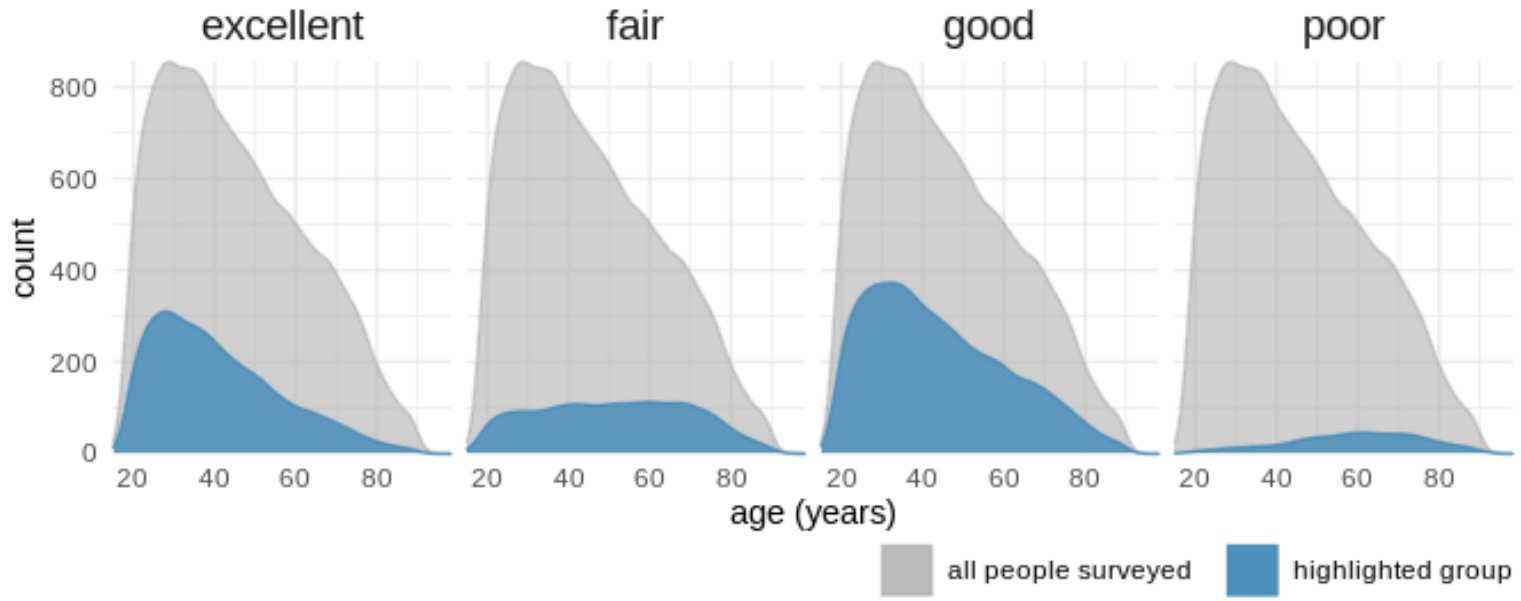
Visualization of proportions

- How to build mutiple overlapping densities?
  - Just use geom_density() twice and also facet_wrap()

```
ggplot(mutate(happy, health = health), aes(x = age, y = ..count..)) +
  geom_density(data = select(df_health, -health),
               aes(fill = "all people surveyed   ",
               color = "all people surveyed   ")) +
  geom_density(aes(fill = "highlighted group", color = "highlighted group"))+
  facet_wrap(~health, nrow = 1) +
  scale_x_continuous(name = "age (years)", limits = c(15, 98),
                     expand = c(0, 0)) +
  scale_y_continuous(name = "count", expand = c(0, 0)) +
  scale_fill_manual(values = c("#b3b3b3a0", "#2b8cbed0"), name = NULL,
                    guide = guide_legend(direction = "horizontal")) +
  scale_color_manual(values = c("#b3b3b3a0", "#2b8cbed0"), name = NULL,
                     guide = guide_legend(direction = "horizontal")) +
  coord_cartesian(clip = "off") + theme_minimal() +
  theme(axis.line.x = element_blank(),
        strip.text = element_text(size = 14,
                                  margin = margin(0, 0, 0.2, 0, "cm")),
        legend.position = "bottom",
        legend.justification = "right",
        legend.margin = margin(4.5, 0, 1.5, 0, "pt"),
        legend.spacing.x = grid::unit(4.5, "pt"),
        legend.spacing.y = grid::unit(0, "pt"),
        legend.box.spacing = grid::unit(0, "cm"))
```

Visualization of proportions

- How to build mutiple overlapping densities?
  - Just use geom_density() twice and also facet_wrap()



Visualization of multiple proportions

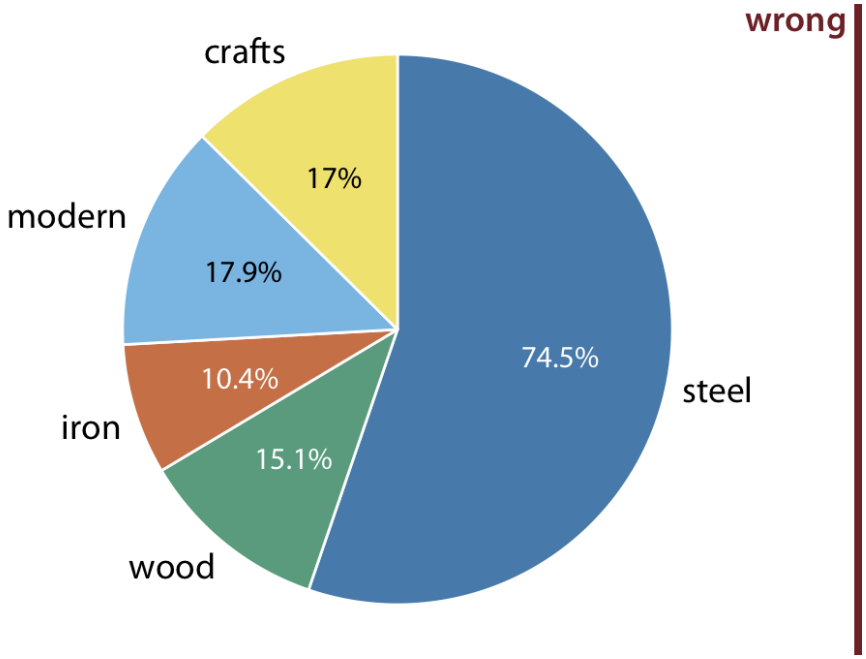
- Useful to investigate **how the proportions evolve over multiple categorical variables at once**
  - E.g. how the proportions of seats are distributed by party and gender, what is the health status of individuals by gender and maritial status, etc.

- Data has:
  - More than two set of values (qualitative or quantitative) *that add to a whole* ***and that can be split into multiple categories***

- *Some* standard geometrical mappings:
  - Mosaic plots
  - Treemaps
  - Nested pies (be careful!)
  - Parallel sets

Visualization of multiple proportions

- **What not to do...**
  - Visualize the fraction of bridges made from steel, iron, or wood **and** the fraction that are crafts or modern.

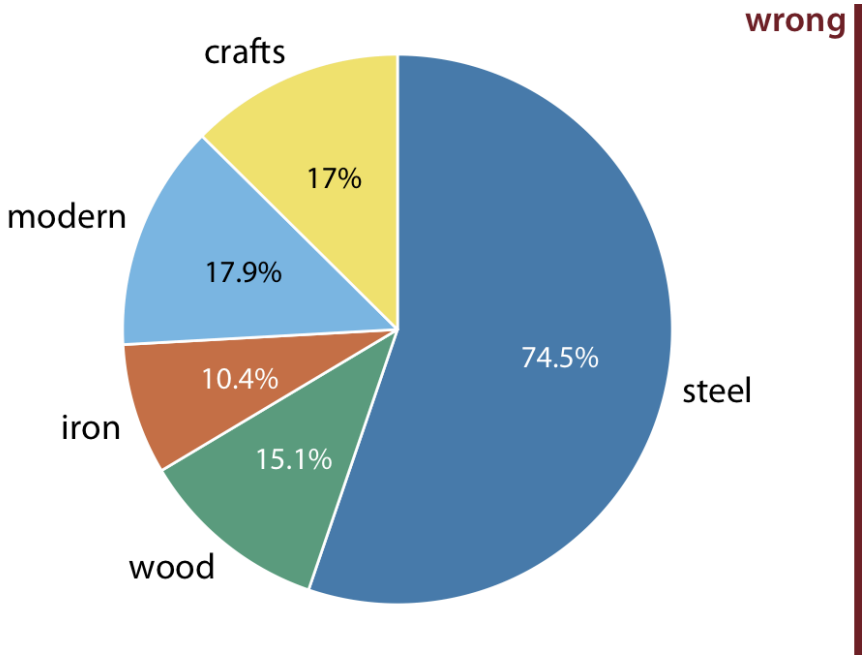


What is wrong with this figure?

Data source: Yoram Reich and Steven J. Fenves, UCI ML Repository.

Visualization of multiple proportions

- What not to do...
  - Visualize the fraction of bridges made from steel, iron, or wood **and** the fraction that are crafts or modern.

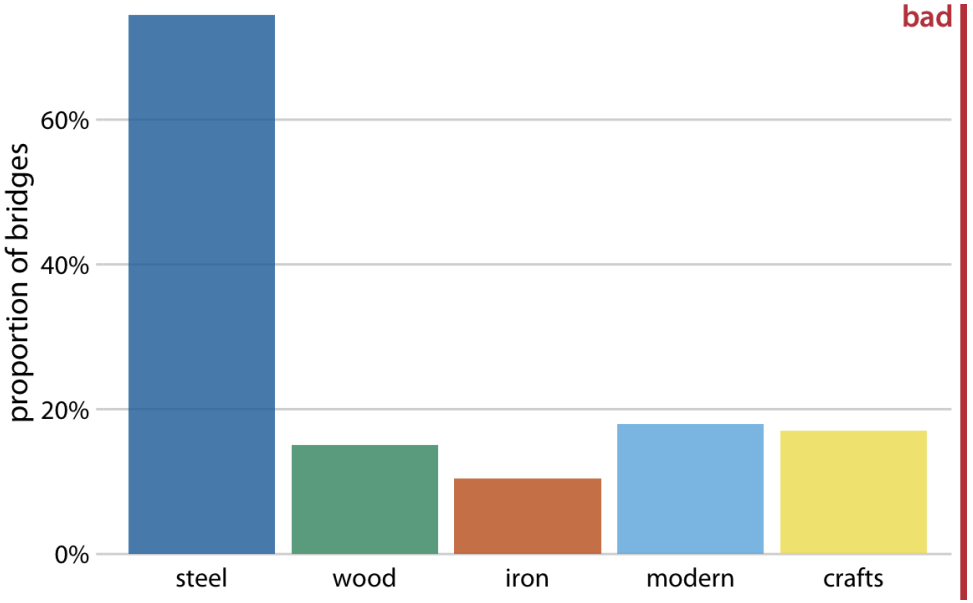


Percentages add up to more than 100%

Data source: Yoram Reich and Steven J. Fenves, UCI ML Repository.

Visualization of multiple proportions

- What not to do...
  - Visualize the fraction of bridges made from steel, iron, or wood **and** the fraction that are crafts or modern.

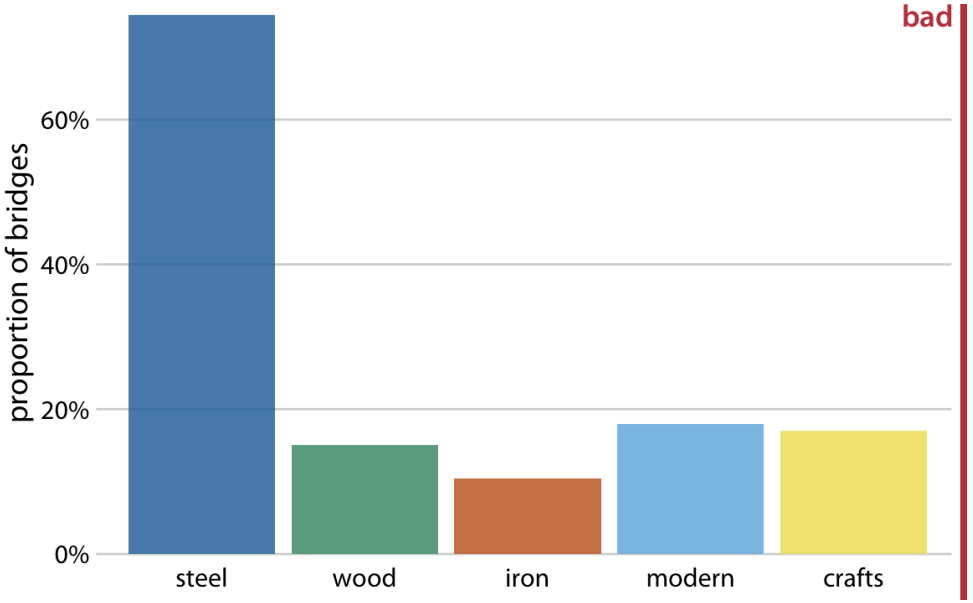


What is bad about this figure?

Data source: Yoram Reich and Steven J. Fenves, UCI ML Repository.

Visualization of multiple proportions

- What not to do...
  - Visualize the fraction of bridges made from steel, iron, or wood **and** the fraction that are crafts or modern.



It does not show that the categories overlap.

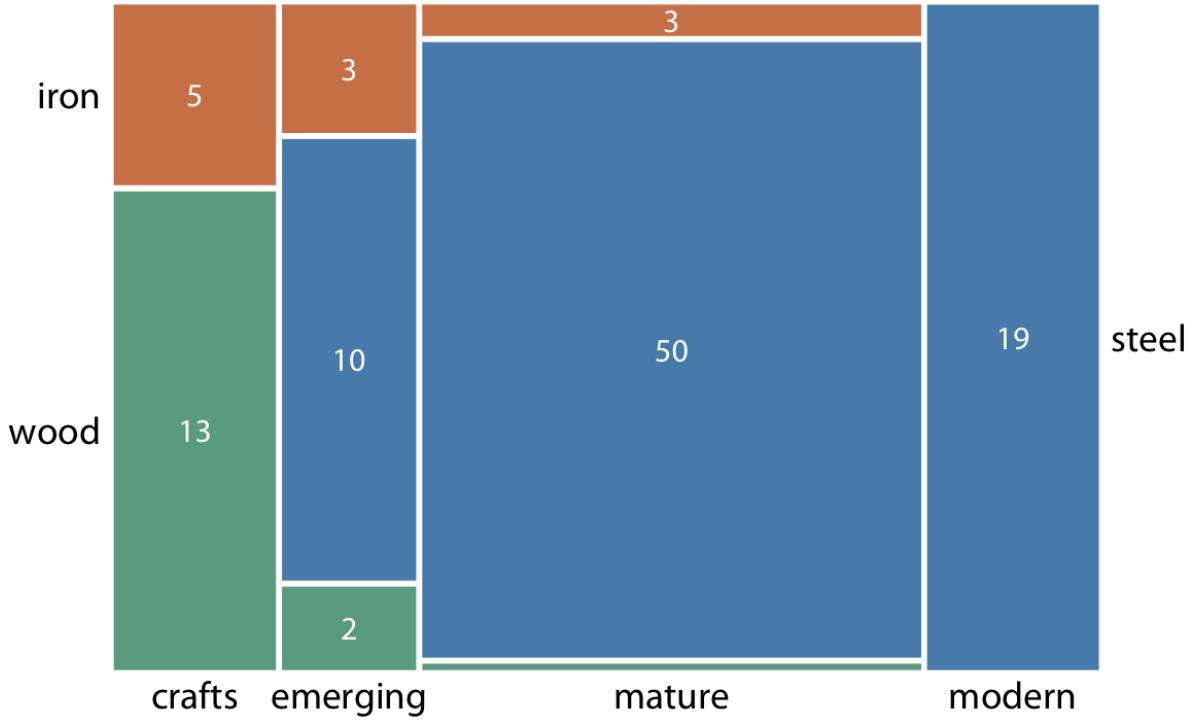
Data source: Yoram Reich and Steven J. Fenves, UCI ML Repository.

Visualization of multiple proportions

- What to do?
- Multiple groups/sets and multiple grouping variables

Visualization of multiple proportions

- Mosaic plots
  - It is not a stacked bar plot!
  - **There is variation of the sizes in the x and y axis.**
  - **Every categorical variable shown must cover all observations.**



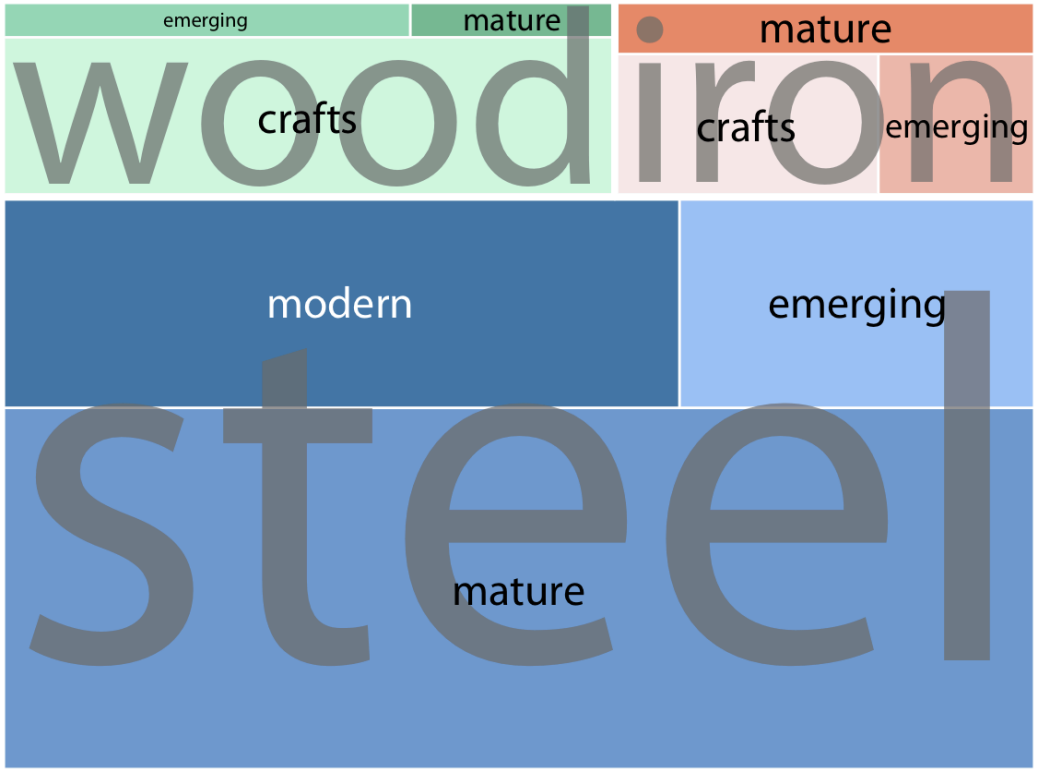
Data source: Yoram Reich and Steven J. Fenves, UCI ML Repository.

Visualization of multiple proportions

- Mosaic plots
  - *How to create the mosaic?*
  - Start by dividing the x axis into the categories following the relative proportions;
  - Within each x-category, subvidide the y-axis into the relative proportions of the objects in that category.
  - Resulting rectangle have areas that are proportional to the number of datapoints in each pair of categorical variables

Visualization of multiple proportions

- Treemaps plots
  - Similar to mosaic plots, but distinct way to split rectangles
  - Recursively nest rectangles inside each other



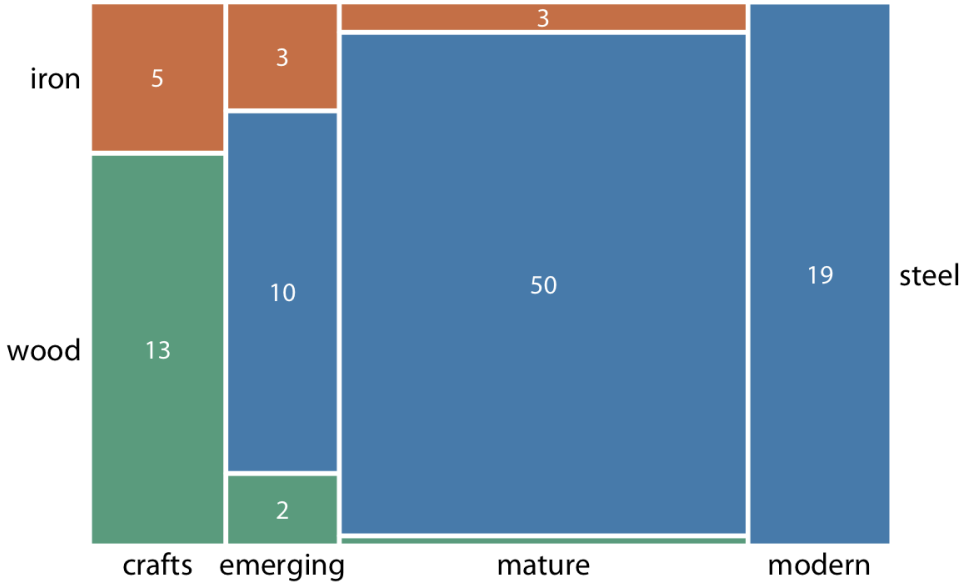
Data source: Yoram Reich and Steven J. Fenves, UCI ML Repository.

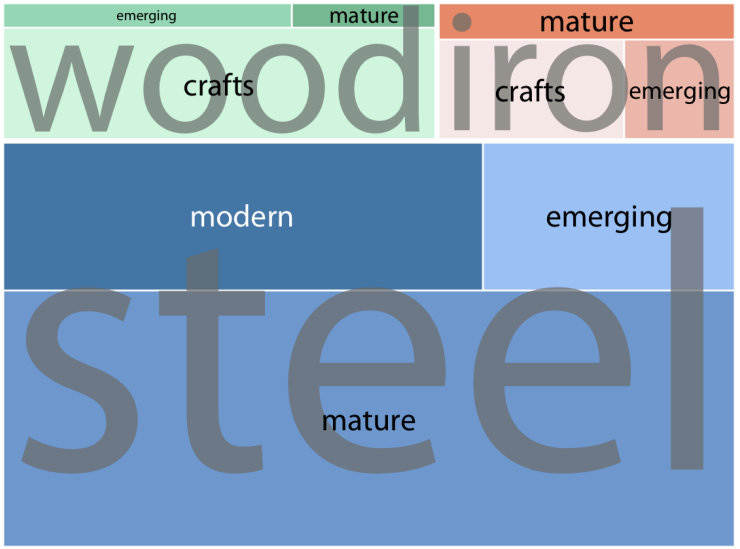
Visualization of multiple proportions

- Treemaps plots
  - *How to create the nesting?*
  - First subdivide the total area of the rectangle into proportions of one categorical variable
  - Then subdivide each rectangle following the proportions of the second categorical variable

Visualization of multiple proportions

- Mosaic and treemaps
  - Different emphasis





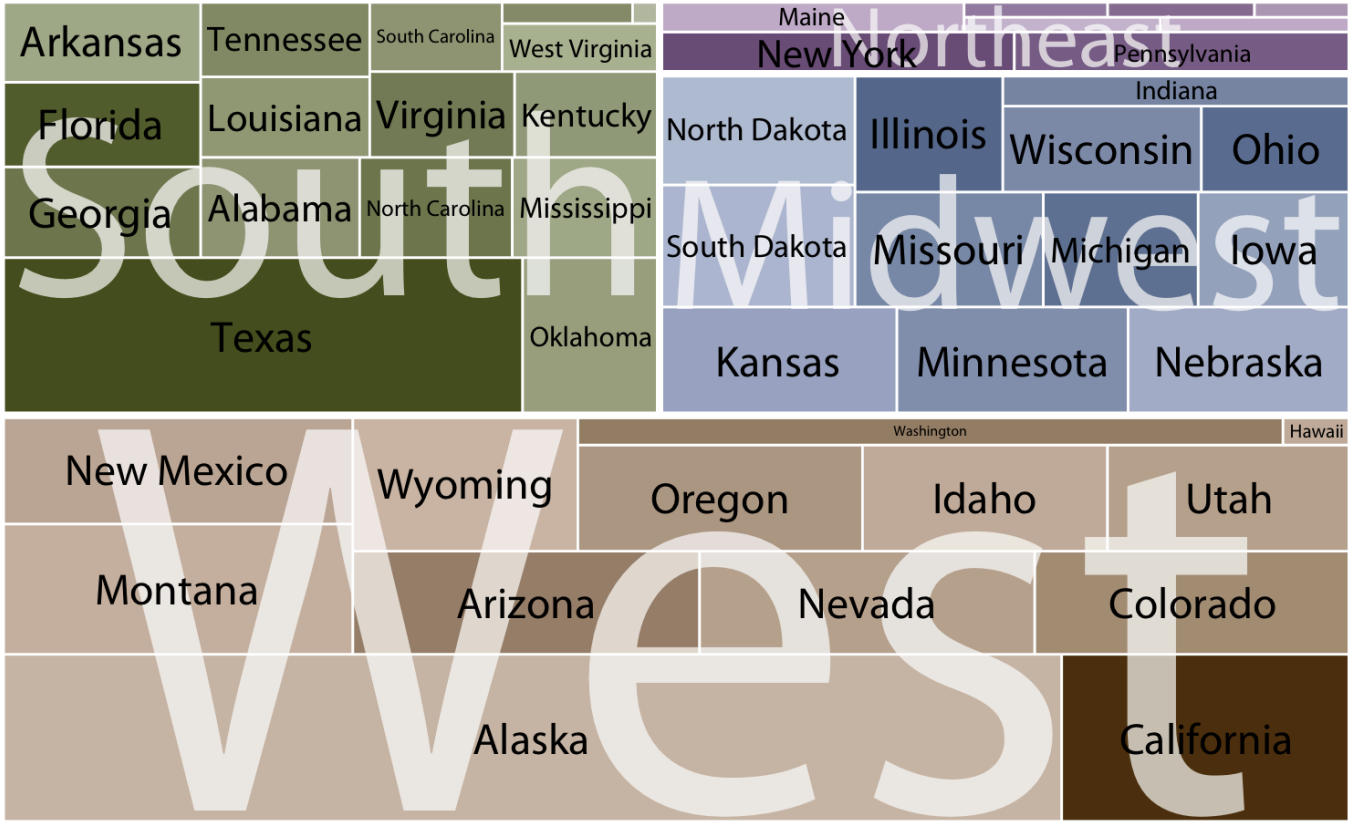
Data source: Yoram Reich and Steven J. Fenves, UCI ML Repository.

Visualization of multiple proportions

- Mosaic and treemaps
  - Mosaic plots assume that all proportions can be identified via combinations of categorical variables.
  - Tree maps allow more free combinations.

Visualization of multiple proportions

- Mosaic and treemaps
  - Mosaic plots assume that all proportions can be identified via combinations of categorical variables. Tree maps allow more free combinations.



Data source: 2010 U.S. Census

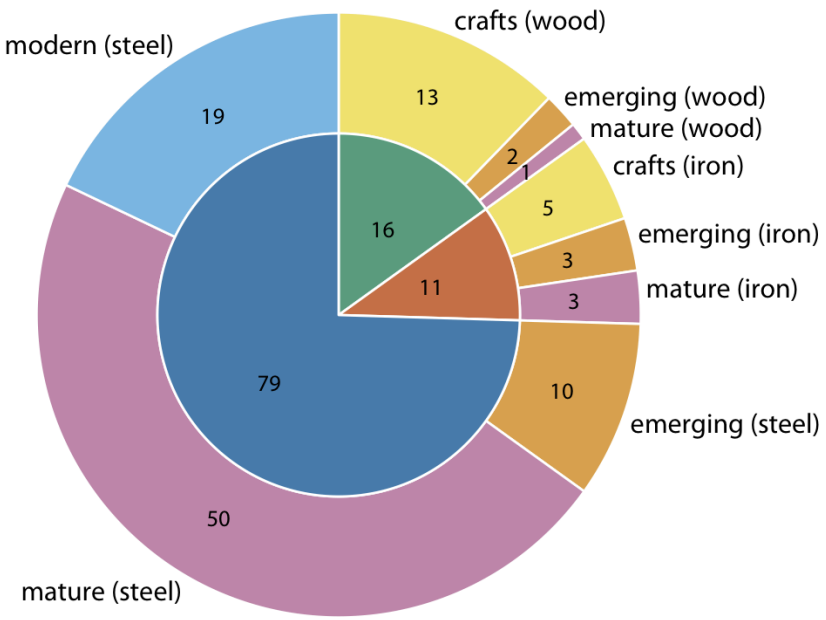
Visualization of multiple proportions

- Mosaic and treemaps
  - Different emphasis.
  - Mosaic plots assume that all proportions can be identified via combinations of categorical variables. Tree maps allow more free combinations.
  - **Both are used.**
  - **Both have similar limitations as stacked bars: difficult direct comparison due to lack of shared baselines.**
    - If you are using these techniques, showing the actual counts *can* help.

Visualization of multiple proportions

- Nested pies
  - A variant of a pie chart
  - Two possibilities : create a nested division

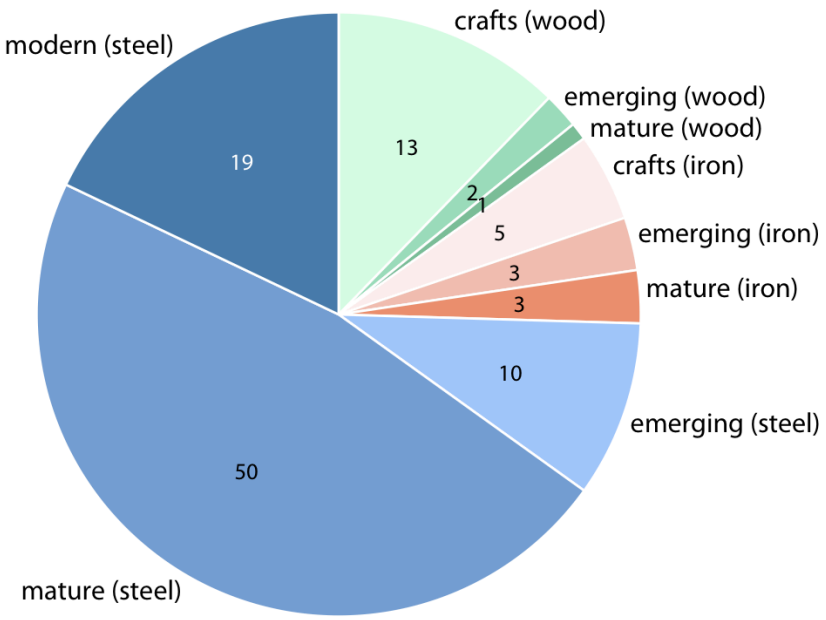
ugly



Data source: Yoram Reich and Steven J. Fenves, UCI ML Repository.

Visualization of multiple proportions

- Nested pies
  - A variant of a pie chart
  - Two possibilities : create a nested coloring



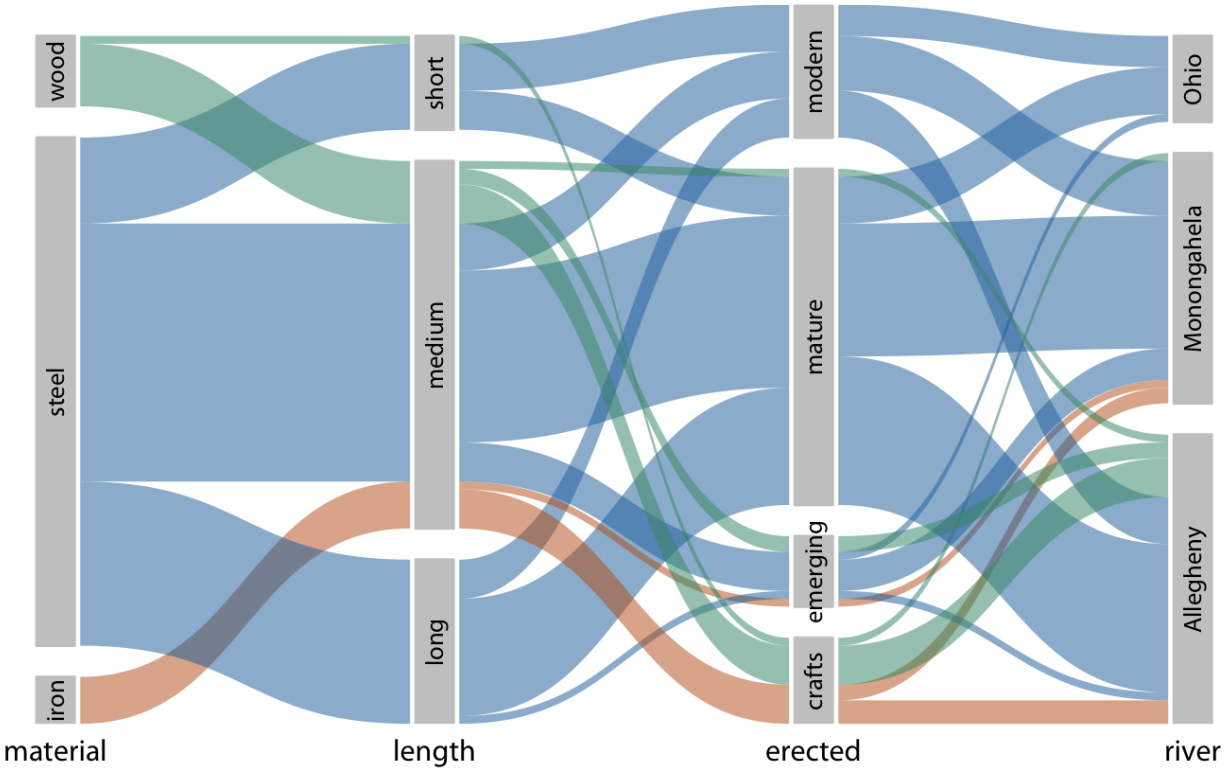
Data source: Yoram Reich and Steven J. Fenves, UCI ML Repository.

Visualization of multiple proportions

- Parallel sets
  - A mapping that works for more than two categorical variables
  - Display independent vertical bars with the relative proportions for each categorical variable
  - Join them via semi-transparent bands to indicate how the groups relate to each other.

Visualization of multiple proportions

- Parallel sets

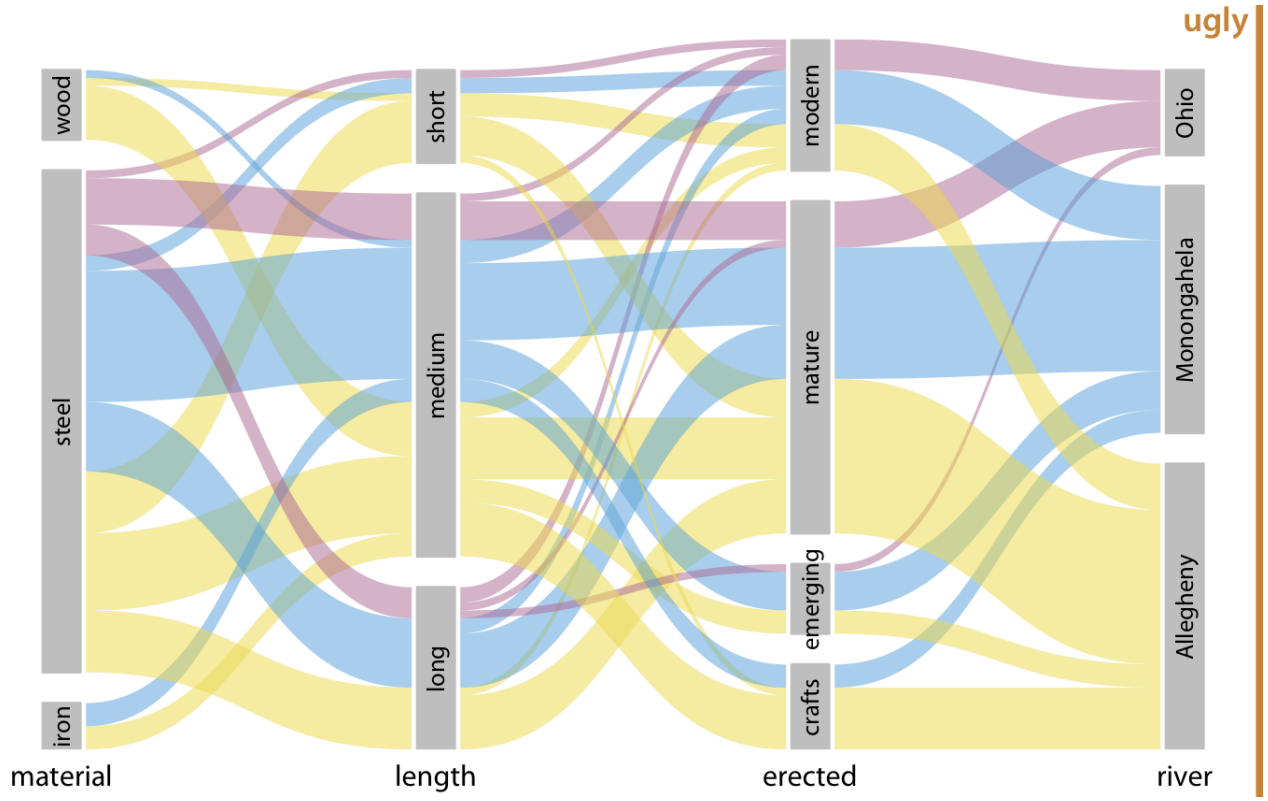


Visualization of multiple proportions

- Important to consider when using Parallel sets
  - **Usually** chose the coloring scheme **starting from the leftmost variable**
  - **Organize** the variables **to minimize the crossing** of the shaded bands

Visualization of multiple proportions

- If you ignore these hints...



Visualization of proportions

- How to build mosaics?
  - **Not very simple... But it can be done by combining bars and facets**

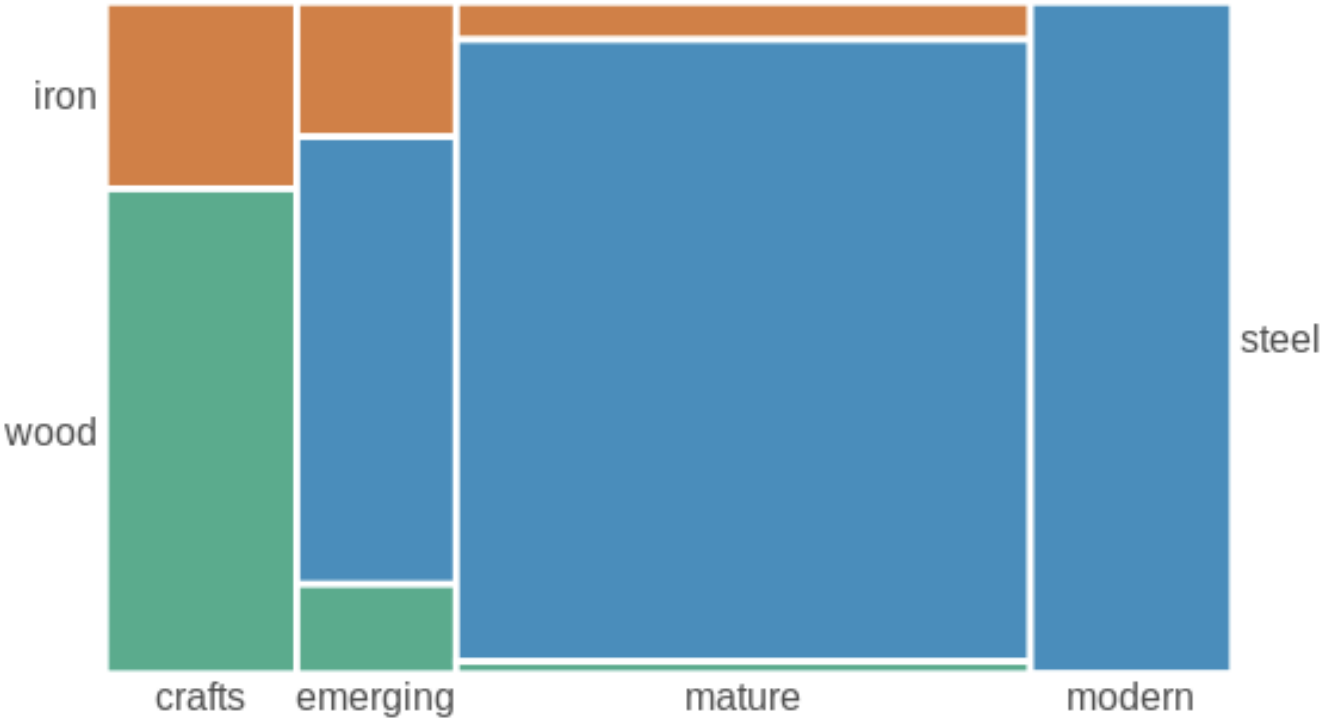
```
bridges_tidy <-
read.csv("http://www.ics.uci.edu/~algol/teaching/informatics143w2021/bridges_tidy.csv")
n_total <- sum(bridges_tidy$count)

labels_df <- group_by(bridges_tidy, erected) %>%
  filter(count != 0) %>% arrange(desc(material)) %>%
  mutate( y = (cumsum(count) - 0.5*count)/group_count, y = ifelse(erected == "mature" &
material == "wood", NA, y))


ggplot(bridges_tidy, aes(x = erected, y = count, width = group_count, fill = material) +
  geom_bar(stat = "identity", position = "fill", colour = "white", size = 1) +
  facet_grid(~erected, scales = "free_x", space = "free_x") +
  scale_y_continuous(
    name = NULL, expand = c(0, 0),
    breaks = filter(labels_df, erected == "crafts")$y,
    labels = filter(labels_df, erected == "crafts")$material,
    sec.axis = dup_axis(
      breaks = filter(labels_df, erected == "modern")$y,
      labels = filter(labels_df, erected == "modern")$material
    )) +
  scale_x_discrete(name = NULL) +
  theme_minimal() +
  theme(line = element_blank(), text = element_text(size=15),
        strip.text = element_blank(), axis.ticks.length = unit(0, "pt"),
        panel.spacing.x = unit(0, "pt"))
```

Visualization of proportions

- How to build mosaics?
  - **Not very simple... But it can be done by combining bars and facets**



Visualization of proportions

- How to build mosaics?
  - **Not very simple... But it can be done by combining bars and facets**

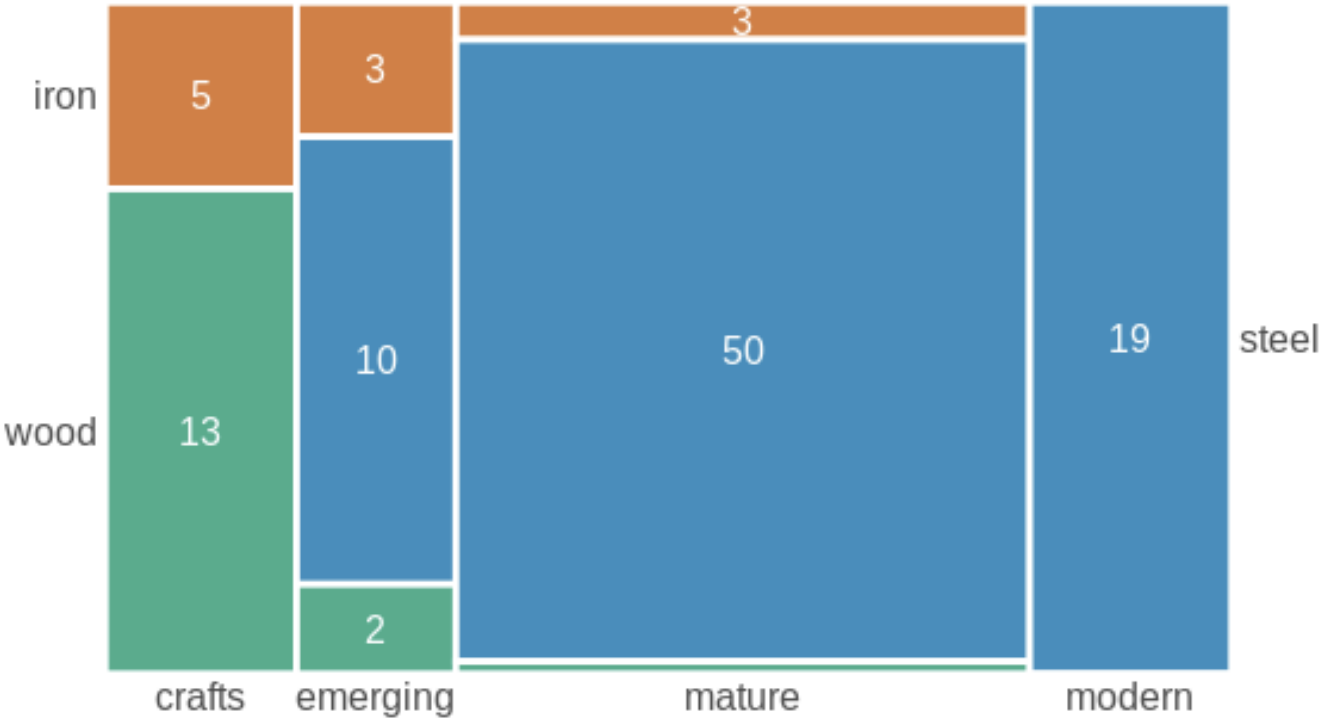
```
bridges_tidy <- read.csv("http://www.ics.uci.edu/~algol/teaching/informatics143w2021/bridges_tidy.csv")
n_total <- sum(bridges_tidy$count)

labels_df <- group_by(bridges_tidy, erected) %>%
  filter(count != 0) %>% arrange(desc(material)) %>%
  mutate( y = (cumsum(count) - 0.5*count)/group_count, y = ifelse(erected == "mature" & material == "wood", NA, y))

ggplot(bridges_tidy) +
  aes(x = erected, y = count, width = group_count, fill = material) +
  geom_bar(stat = "identity", position = "fill", colour = "white", size = 1) +
  geom_text(
    data = labels_df,
    aes(y = y, label = count, color = material),
    na.rm = TRUE,
    size = 12/.pt) +
  facet_grid(~erected, scales = "free_x", space = "free_x") +
  scale_y_continuous(
    name = NULL, expand = c(0, 0),
    breaks = filter(labels_df, erected == "crafts")$y,
    labels = filter(labels_df, erected == "crafts")$material,
    sec.axis = dup_axis(
      breaks = filter(labels_df, erected == "modern")$y,
      labels = filter(labels_df, erected == "modern")$material
    )) +
  scale_x_discrete(name = NULL) +
  scale_fill_manual(
    values = c("#D55E00D0", "#0072B2D0", "#009E73D0"),
    guide = "none") +
  scale_color_manual(
    values = c(iron = "white", wood = "white", steel = "white"),
    guide = "none") +
  coord_cartesian(clip = "off") +
  theme_minimal() +
  theme(line = element_blank(),
        strip.text = element_blank(),
        axis.ticks.length = unit(0, "pt"),
        panel.spacing.x = unit(0, "pt"))
```

Visualization of proportions

- How to build mosaics?
  - **Not very simple... But it can be done by combining bars and facets**



Visualization of proportions

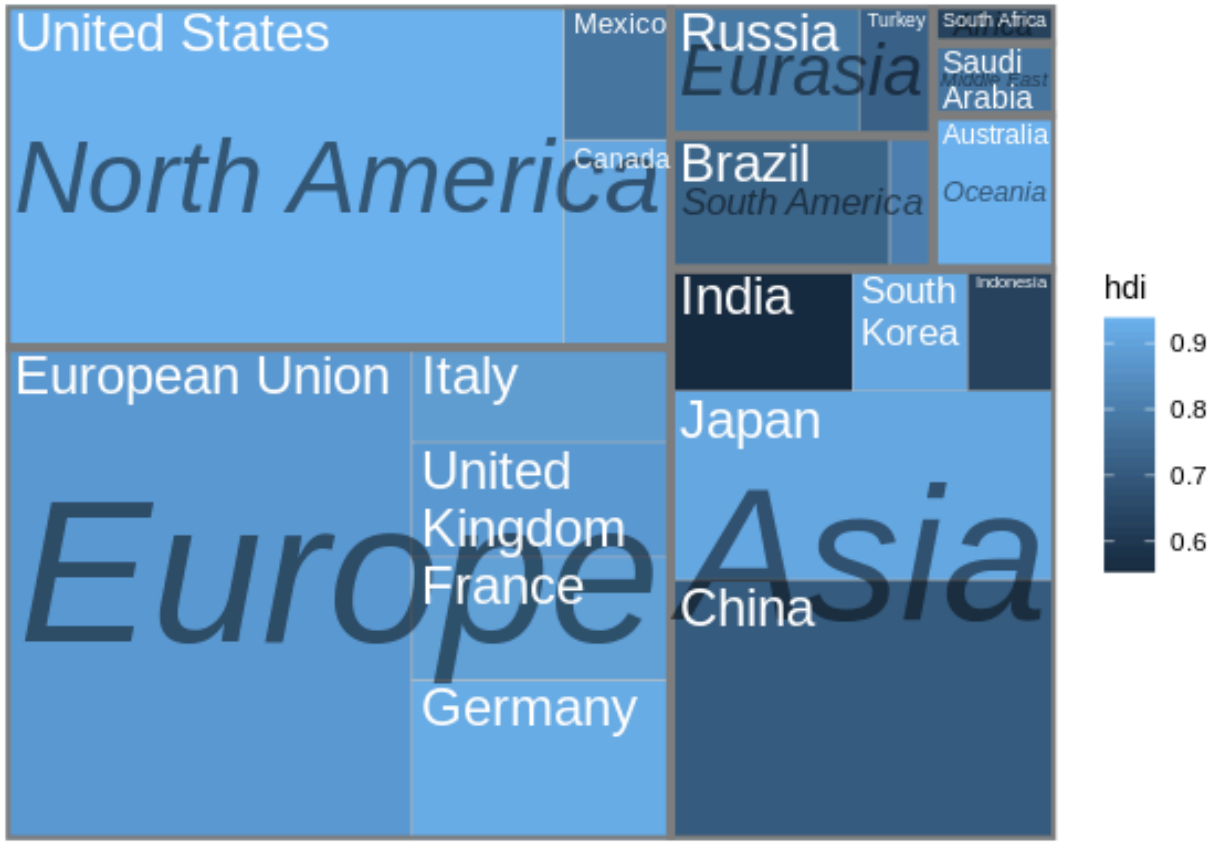
- How to build treemaps?
  - Not very simple from standard ggplot2... But it can be done using treemapif
  - Then, you can simply use the geom_treemap() call

```
devtools::install_github("wilkox/treemapify")
require(treemapify)

ggplot(G20, aes(area = gdp_mil_usd, fill = hdi,
                label = country, subgroup = region)) +
  geom_treemap() +
  geom_treemap_subgroup_border() +
  geom_treemap_subgroup_text(place = "centre", grow = T,
                             alpha = 0.5, colour = "black",
                             fontface = "italic", min.size = 0) +
  geom_treemap_text(colour = "white", place = "topleft", reflow = T)
```

Visualization of proportions

- How to build treemaps?
  - Not very simple from standard ggplot2... But it can be done using treemapif
  - Then, you can simply use the geom_treemap() call



Visualization of proportions

- How to build parallel plots?
  - Not very simple from standard ggplot2... But it can be done using ggforce
  - Then, you can simply use the geom_parallel_sets() call

```
devtools::install_github("thomasp85/ggforce")
require(ggforce)

data <- reshape2::melt(Titanic)
data <- gather_set_data(data, 1:4)

ggplot(data, aes(x, id = id, split = y, value = value)) +
  geom_parallel_sets(aes(fill = Survived), alpha = 0.3, axis.width = 0.2) +
  geom_parallel_sets_axes(axis.width = 0.2) +
  geom_parallel_sets_labels(colour = 'white') +
  theme_void()
```

Visualization of proportions

- How to build parallel plots?
  - Not very simple from standard ggplot2... But it can be done using ggforce
  - Then, you can simply use the geom_parallel_sets() call