

**Hacettepe University
Department of Industrial Engineering
Undergraduate Program
2023-2024 Fall**

**EMU 430 – Data Analytics
Week 8
November 23, 2023**

Instructor: Erdi Dasdemir

edasdemir@hacettepe.edu.tr
www.erdidasdemir.com

**Previously on
EMU430**

**Data Visualization
Principles**

**Principle:
Know When to Include 0**

**Principle:
Do Not Distort Quantities**

**Principle:
Order by a Meaningful Value**

**Principle:
Show the Data**

**Principle:
Use Common Axes to Ease
Comparisons**

**Principle:
Consider Transformations**

**Principle:
Compared Visual Cues Should be
Adjacent to Ease Comparisons**

**Principle:
Encoding a Third Variable**

**Case Study:
Vaccines**

**Principle
avoid pseudo and gratuitous 3D plots**

I drew inspiration primarily from [Dr. Rafael Irizarry's "Introduction to Data Science" Book](#) and ["Data Science" course by HarvardX on edX](#) for the slides this week.

Previously on

EMU430

World Health and Economics

Case Study

We tried to answer the following two questions:

1. Is it fair to say the world is divided into rich (Western nations) and poor (the developing world in Africa, Asia, and Latin America)?
2. Has income inequality across countries worsened during the last 40 years?



- The data set was put together by `dslabs` library, and it was created using a number of spreadsheets available from the Gapminder foundation.

```
library(dslabs)
library(tidyverse)
data(gapminder)
head(gapminder)
```

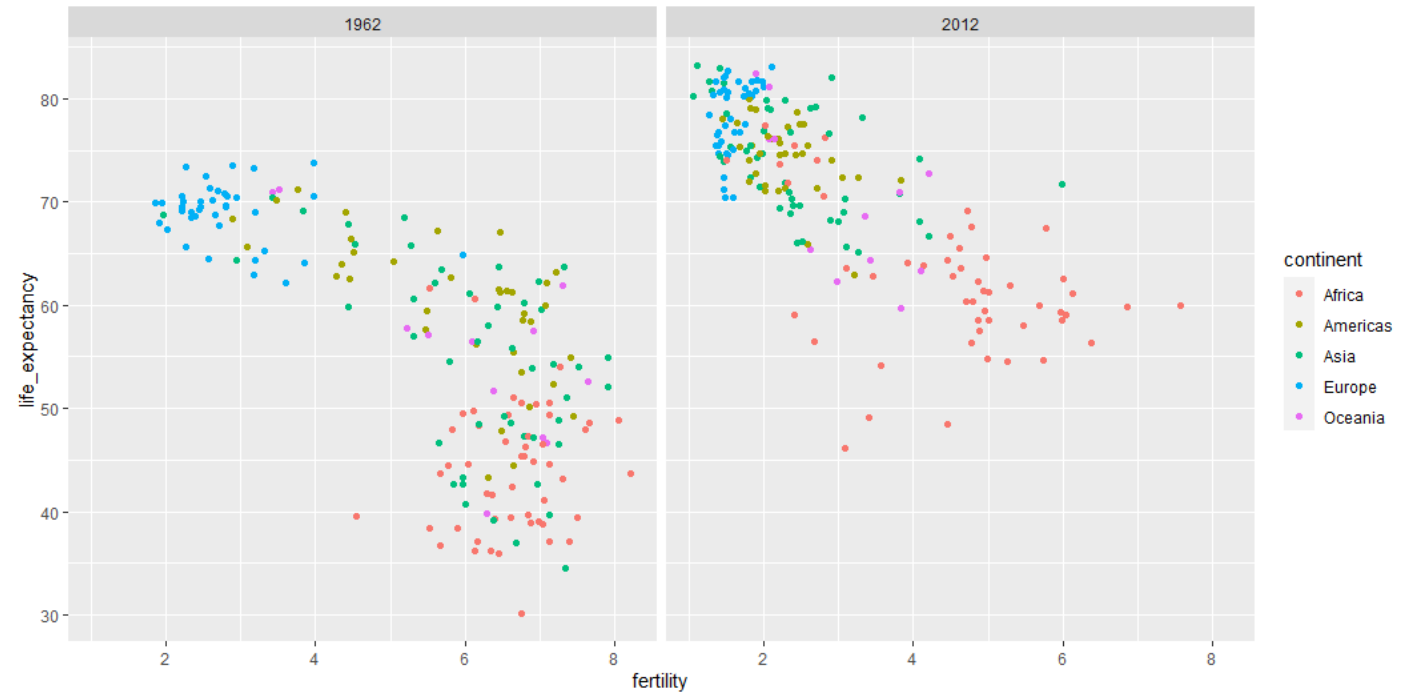
```
> head(gapminder)
```

	country	year	infant_mortality	life_expectancy	fertility	population	gdp	continent	region
1	Albania	1960	115.40	62.87	6.19	1636054	NA	Europe	Southern Europe
2	Algeria	1960	148.20	47.50	7.65	11124892	13828152297	Africa	Northern Africa
3	Angola	1960	208.00	35.98	7.32	5270844	NA	Africa	Middle Africa
4	Antigua and Barbuda	1960	NA	62.97	4.43	54681	NA	Americas	Caribbean
5	Argentina	1960	59.87	65.39	3.11	20619075	108322326649	Americas	South America
6	Armenia	1960	NA	66.86	4.55	1867396	NA	Asia	Western Asia

- We will test our knowledge regarding differences in child mortality across different countries.

Faceting: 1962 vs 2012

```
gapminder %>% filter(year %in%  
c(1962, 2012)) %>%  
ggplot(aes(fertility,  
life_expectancy, color = continent))  
+ geom_point() + facet_grid(.~year)
```



Gapminder Life Expectancy and Fertility

○ To observe this improvement over multiple years, we can add more years to our plot.

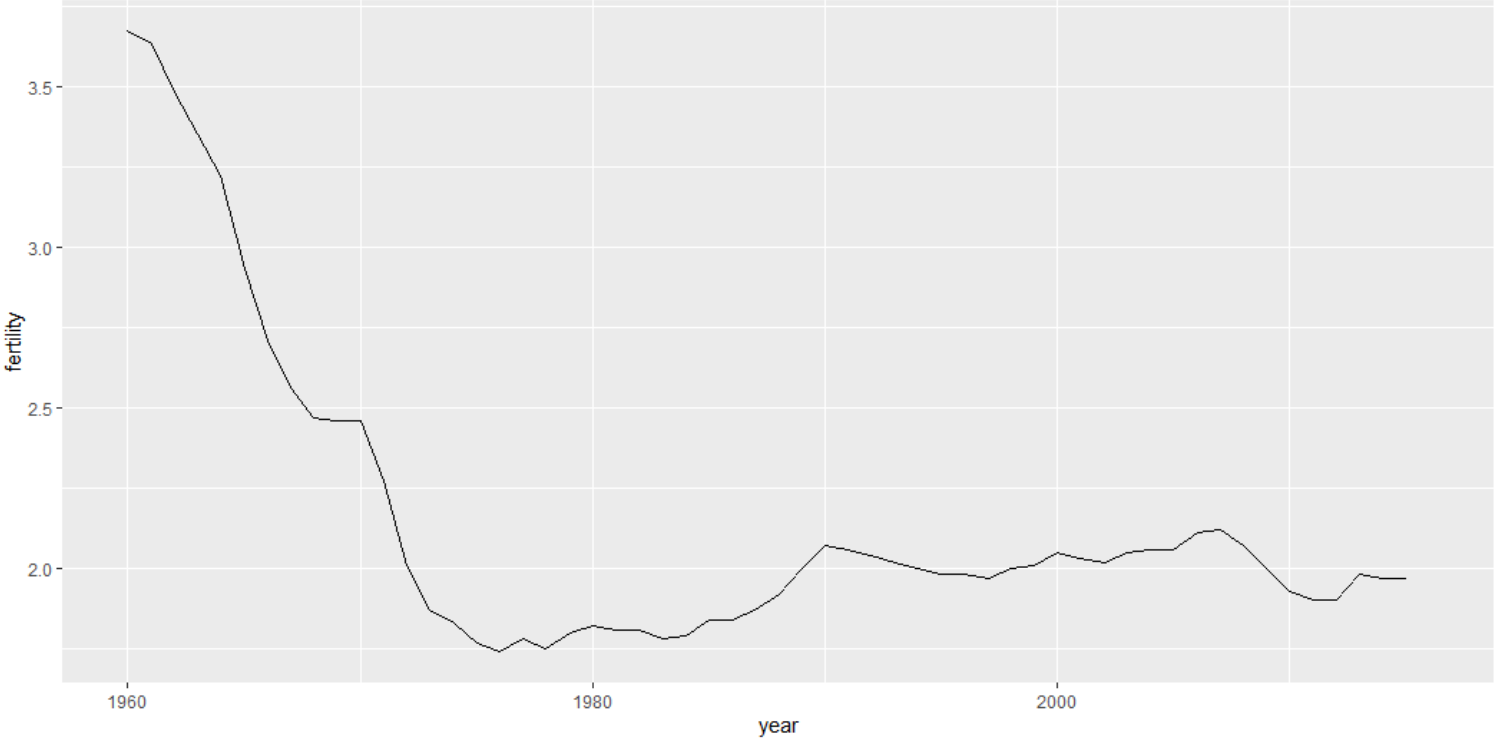
○ **facet_wrap()** function

```
gapminder %>% filter(year %in%  
c(1962, 1970, 1980, 1990, 2000,  
2012)) %>% ggplot(aes(fertility,  
life_expectancy, color = continent))  
+ geom_point() + facet_wrap(~year)
```



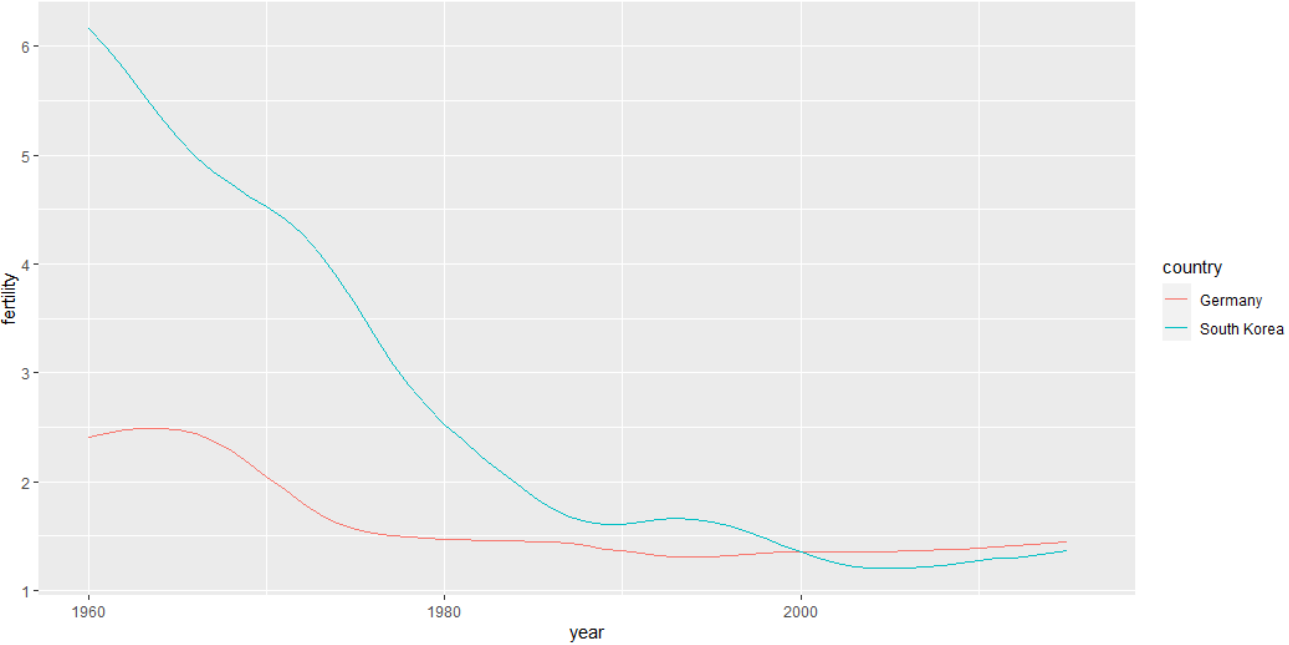
Time Series Plot

```
gapminder %>% filter(country == "United States") %>%  
ggplot(aes(x=year, y=fertility)) + geom_line()
```



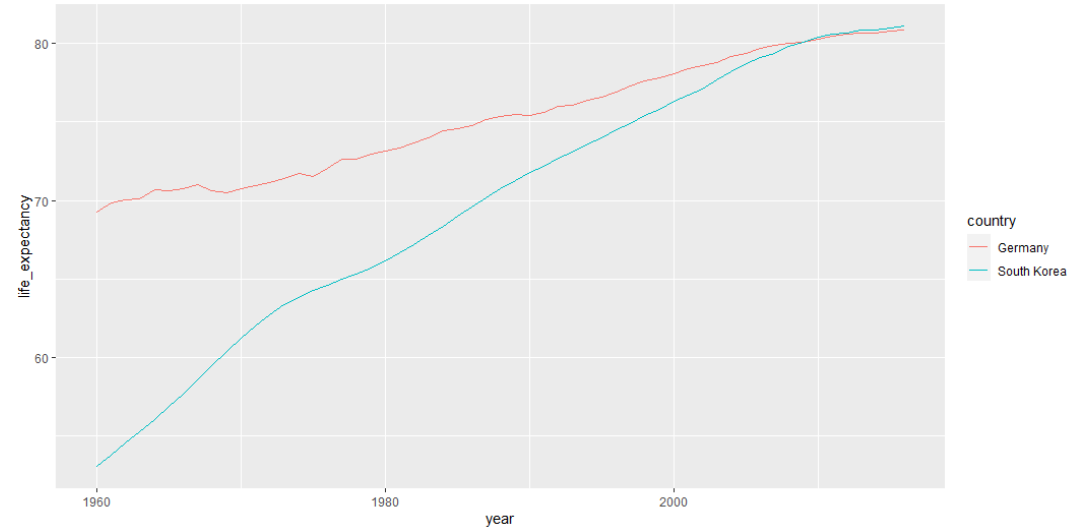
- The good thing about using colors is that ggplot automatically groups data with color mapping.

```
countries <- c("South Korea", "Germany")
gapminder %>% filter(country %in%
countries) %>% ggplot(aes(year,
fertility, color = country)) +
geom_line()
```

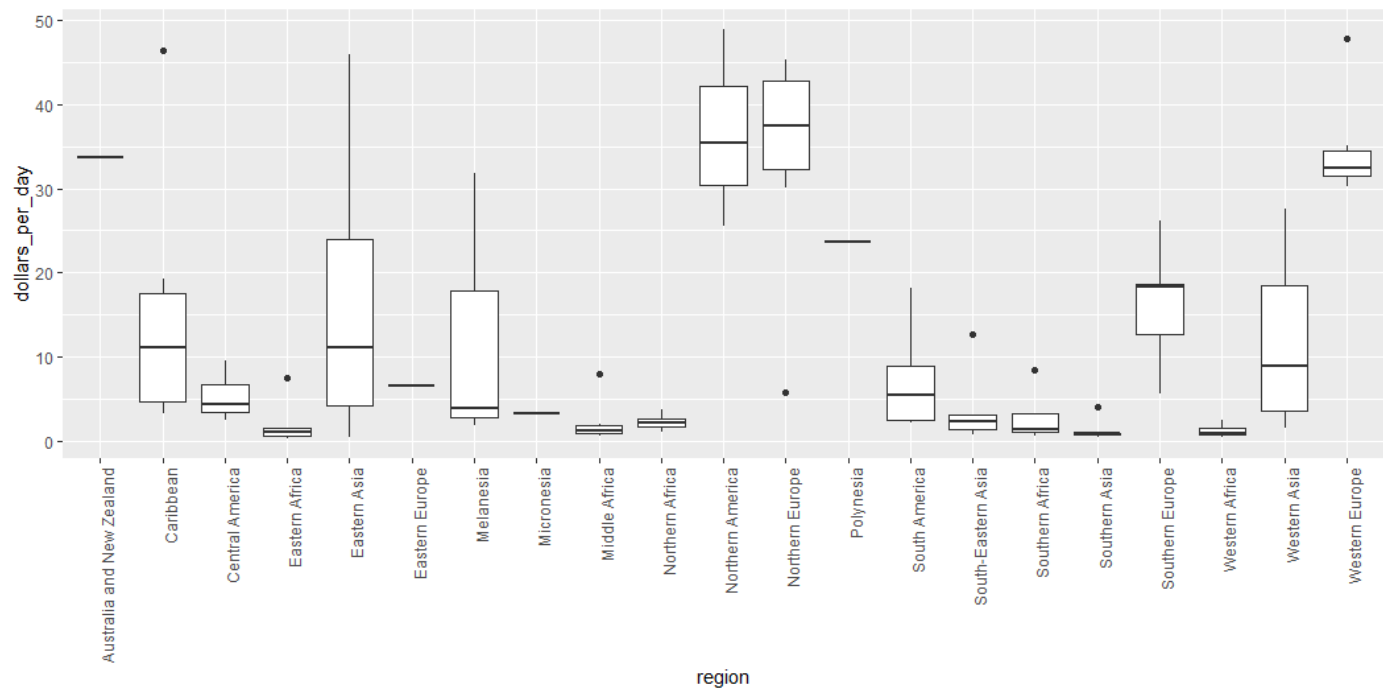


Let us look at life expectancy.

```
countries <- c("South Korea", "Germany")
gapminder %>% filter(country %in%
countries) %>% ggplot(aes(year,
life_expectancy, color = country)) +
geom_line()
```



```
past_year = 1970
p <- gapminder %>% filter(year == past_year & !is.na(gdp)) %>% ggplot(aes(region,
dollars_per_day))
p + geom_boxplot() + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



The order is alphabetically.

We can do something more meaningful.

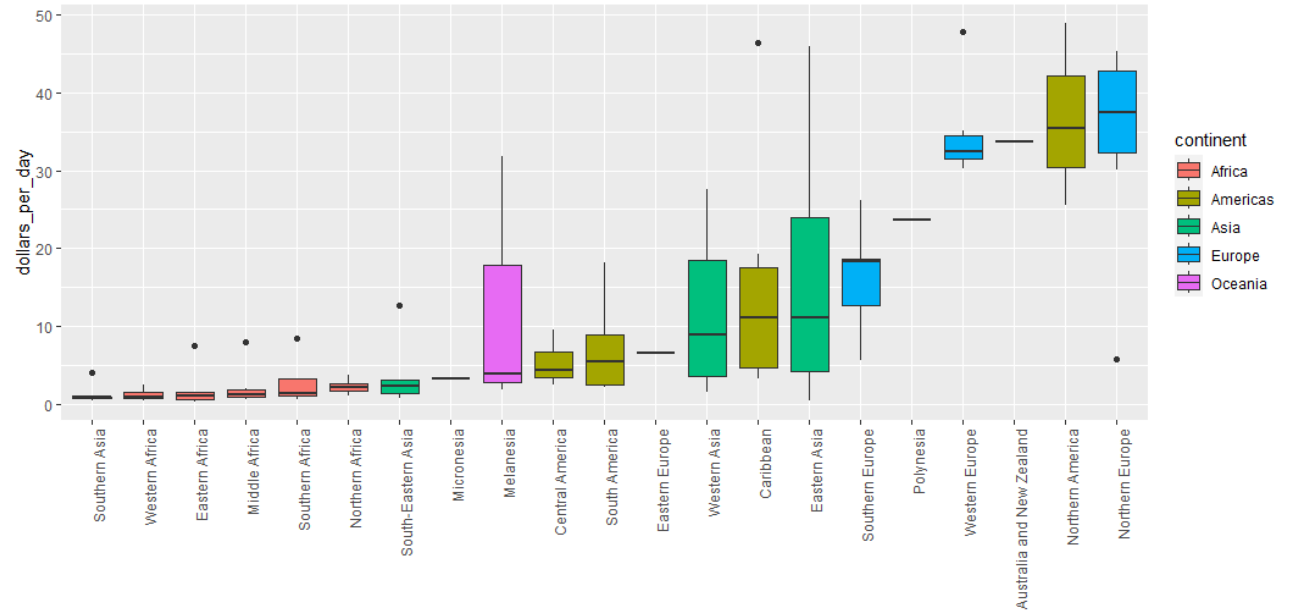
We will use `reorder()` function.

Reorder regions by their median income levels

```
past_year = 1970
```

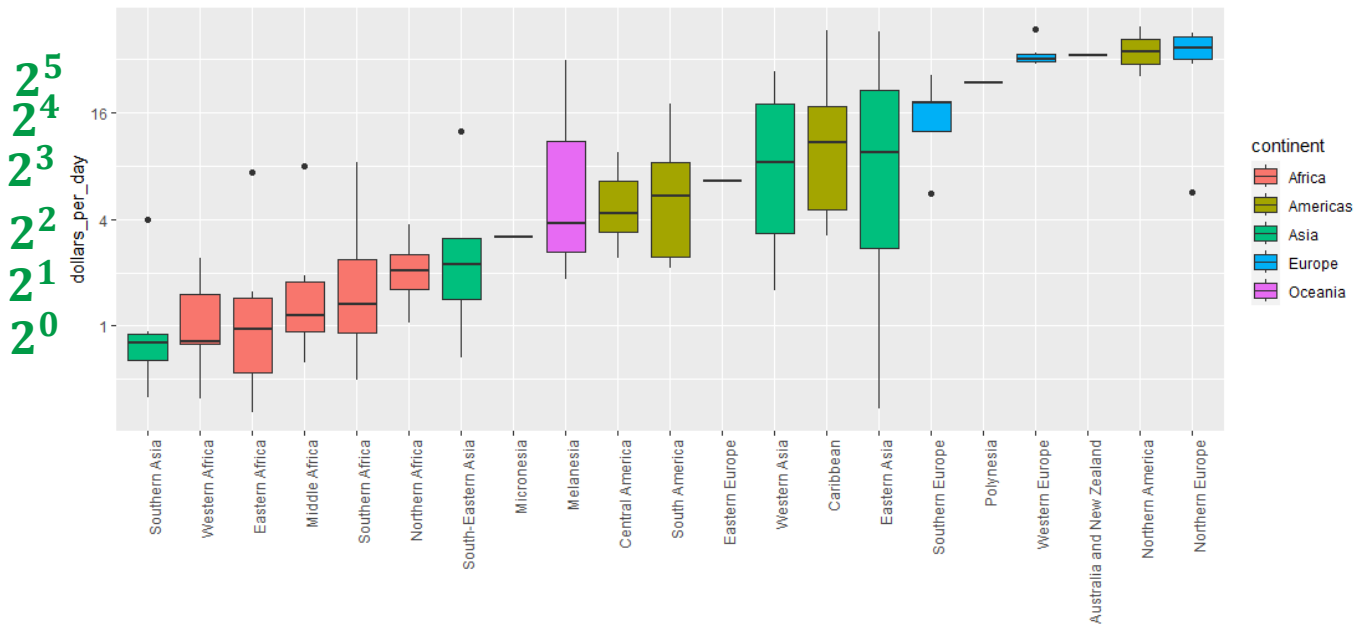
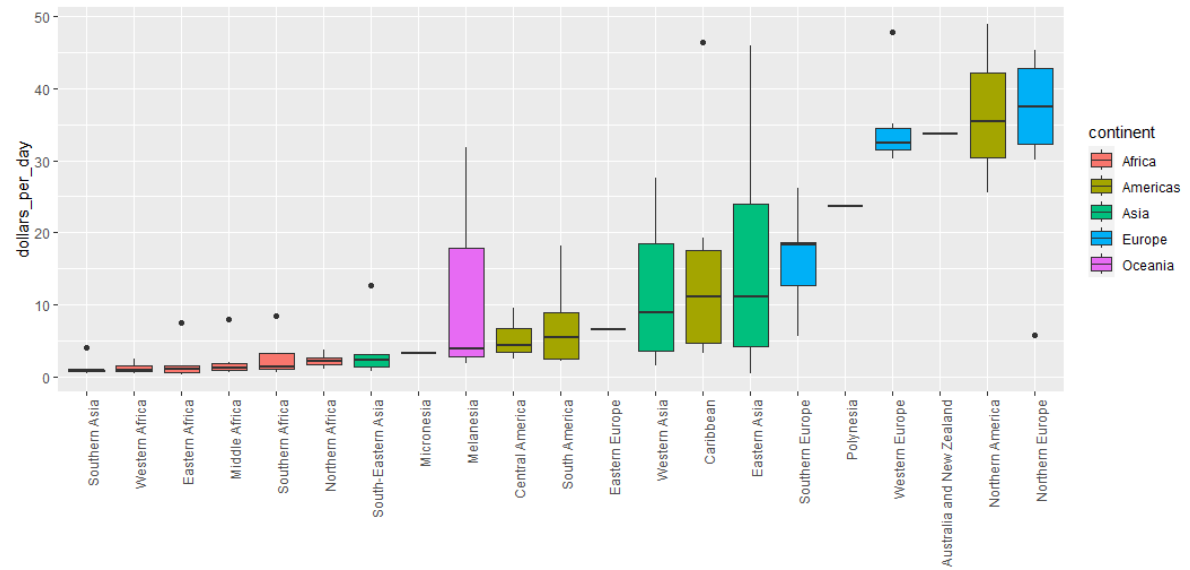
```
p <- gapminder %>% filter(year ==  
past_year & !is.na(gdp)) %>%  
mutate(region = reorder(region,  
dollars_per_day, FUN = median)) %>%  
ggplot(aes(region, dollars_per_day,  
fill = continent)) + geom_boxplot() +  
theme(axis.text.x = element_text(angle  
= 90, hjust = 1)) + xlab("")
```

p



Change scale to log scale

```
p + scale_y_continuous(trans = "log2")
```



In order to compare two years, we need to make sure that the list of countries are the same in these two years.

```
country_list_1 <- gapminder %>% filter(year == 1970 & !is.na(dollars_per_day))  
%>% .$ country
```

```
country_list_2 <- gapminder %>% filter(year == 2010 & !is.na(dollars_per_day))  
%>% .$ country
```

```
country_list <- intersect(country_list_1, country_list_2)
```

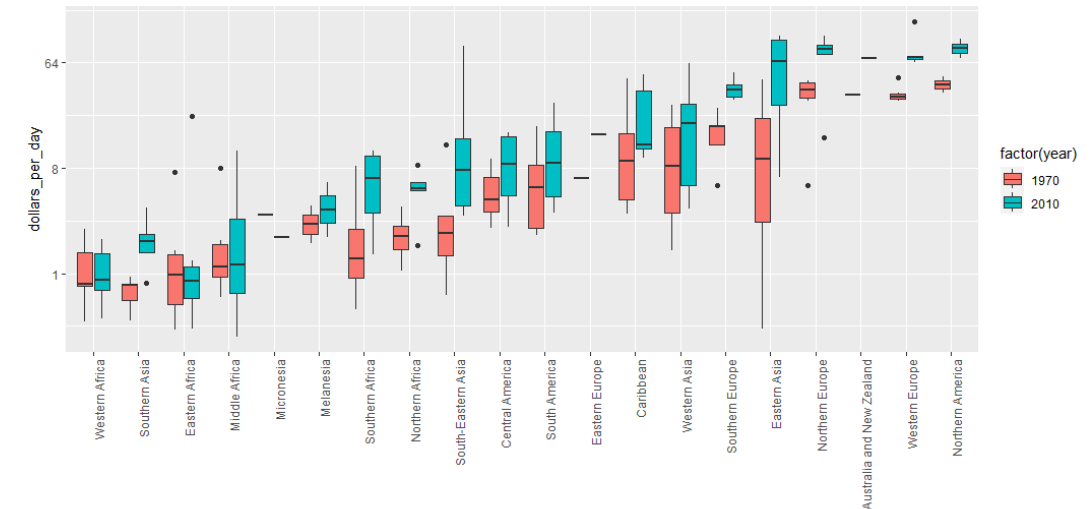

Ease comparisons

```

p <- gapminder %>% filter(year %in% c(1970,
2010) & country %in% country_list) %>%
mutate(region = reorder(region,
dollars_per_day, FUN = median)) %>% ggplot()
+ theme(axis.text.x = element_text(angle =
90, hjust = 1)) + xlab(" ") +
scale_y_continuous(trans = "log2")


p + geom_boxplot(aes(region, dollars_per_day,
fill = factor(year)))

```



Data Visualization

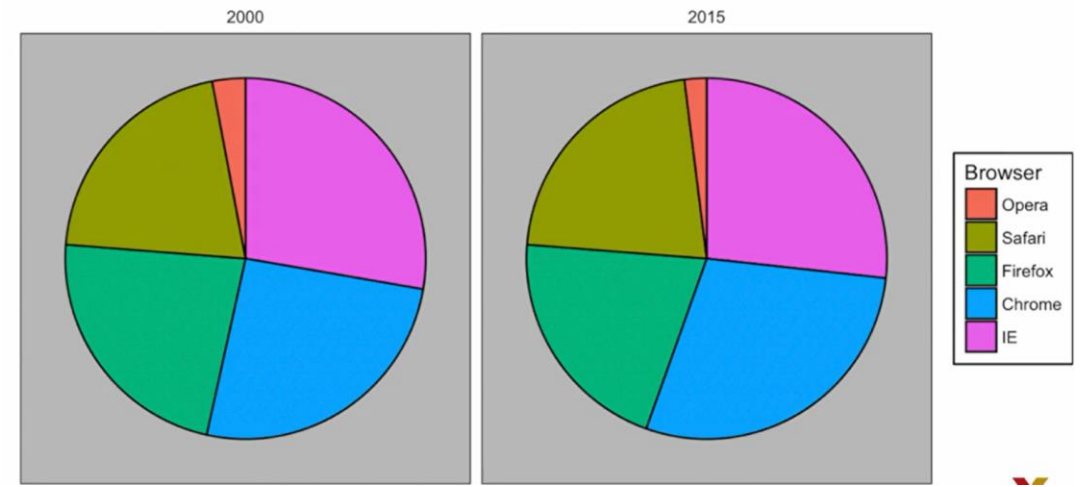
Principles

- Some general principles we can use as guidelines for effective data visualization.
- Based on a talk of [Karl Broman's Creating effective figures and tables](#) 
- Class notes from [Peter Aldhous' Introduction to Data Visualization Course](#)
- Our course book [Rafael A. Irizarry's Introduction to Data Science Book](#)

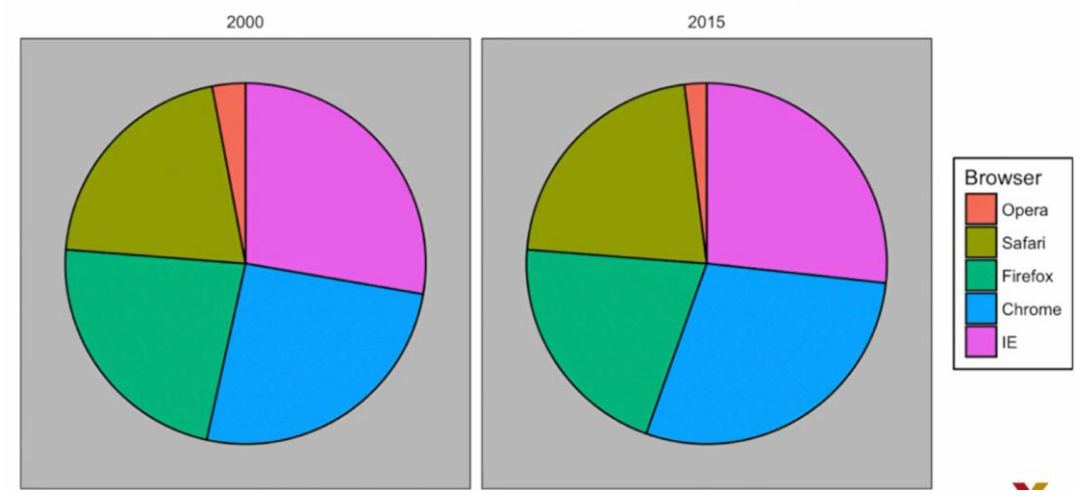
Approaches to visualize data

- position
- aligned lengths
- angles
- area
- brightness
- color hue

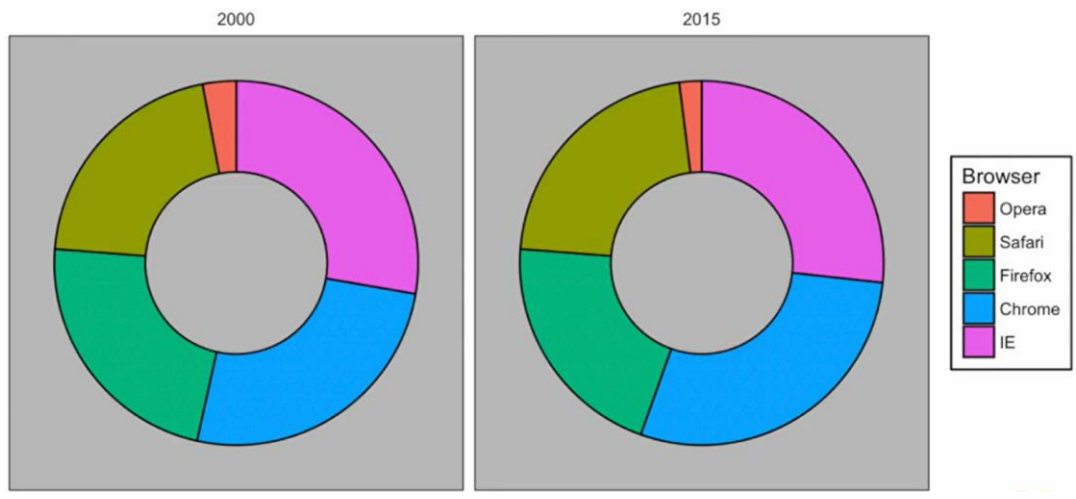
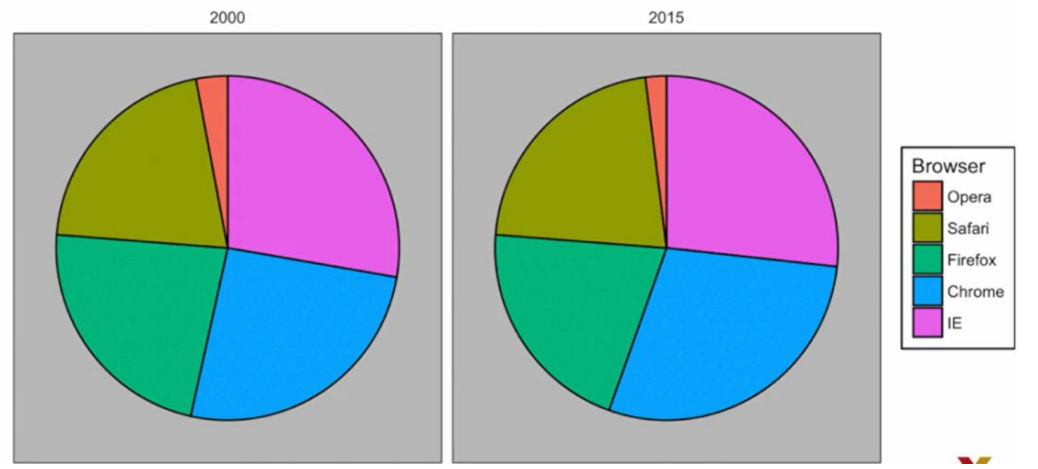
- Suppose that we want to report the results from two polls, asking what's your browser preference.
- The polls were taken in 2000, and then in 2015.
- Here, for each year, we are simply comparing five quantities and five percentages.
- A widely used graphical representation of percentages, popularized by Microsoft Excel, is the pie chart.
- Here's the pie chart for our data. There's two pie charts, one for 2000, one for 2015.



- we're representing quantities with **both areas and angles**, since each pie slice's angle and area are proportional to the quantity they represent.
- This turns out to be a suboptimal choice, since, as demonstrated by perception studies, humans are not good at precisely quantifying angles.



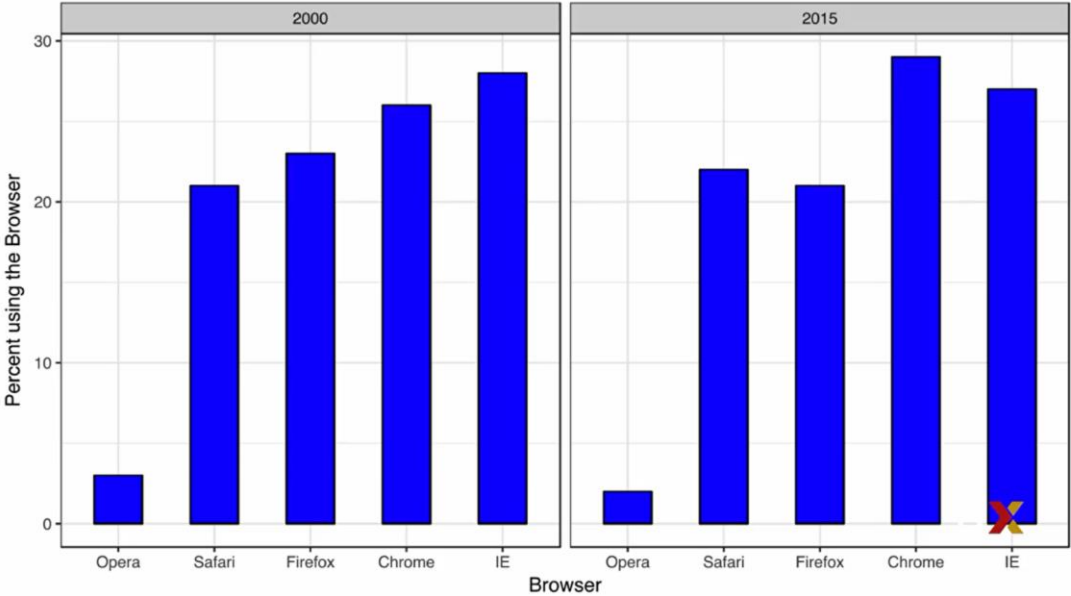
- Humans are even worse when only area is available.
- This makes the donut chart, which only uses area, even worse than the pie chart.
- Can you determine the actual percentages and rank the browser's popularity?
- Can you see how the percentages changed from 2000 to 2015?



- In this case, simply showing the numbers is not only clearer, but it would save us on print costs if making a paper version of our results.

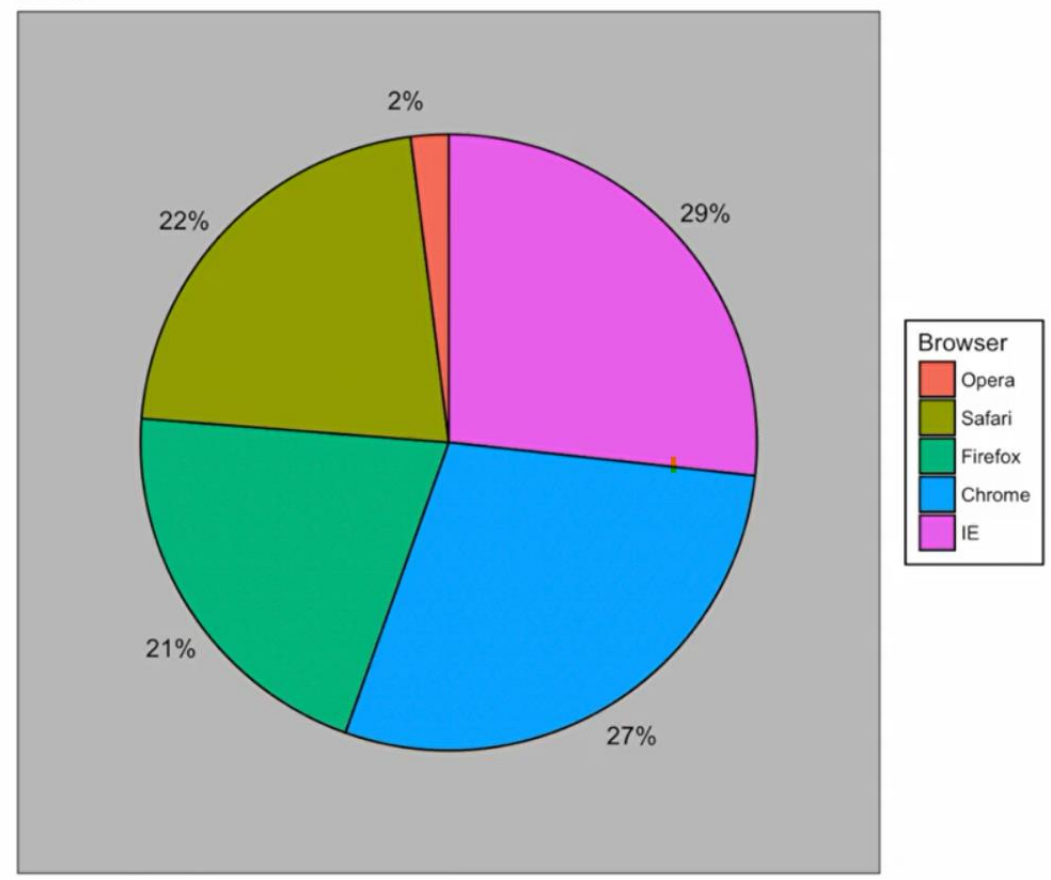
Browser	2000	2015
Opera	3	2
Safari	21	22
Firefox	23	21
Chrome	26	29
IE	28	27

- If we insist on a plot, the preferred way to plot these quantities is to use length and positions since humans are much better at judging linear measures.
- We can now determine the actual percentages by following a horizontal line to the y-axis.



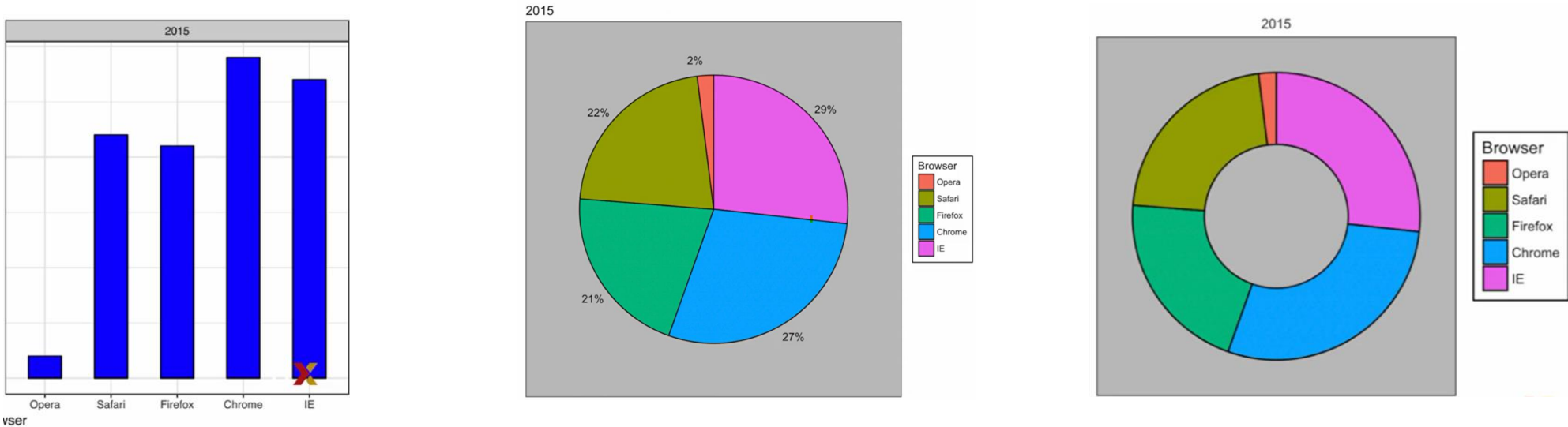
- If for some reason you need to make a pie chart, do include the percentages as numbers to avoid having to infer them from the angles or area.

2015



In summary,

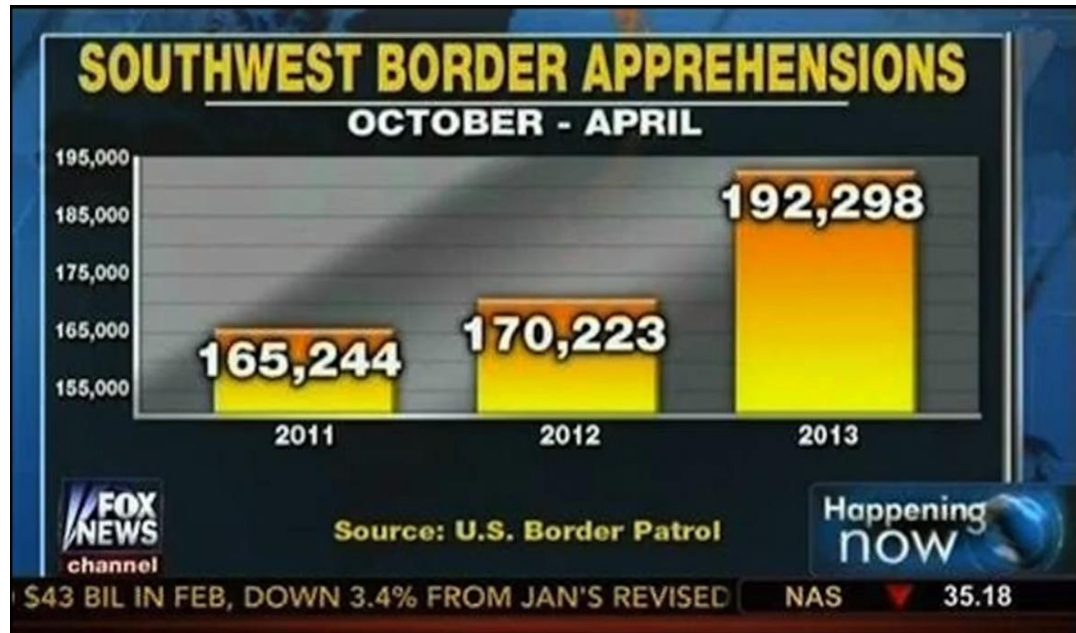
- The preferred way to display quantities in this example: position and length > angles > area.
- Brightness and color are even harder to quantify than angles and area. But, as we will see later, they are sometimes useful when more than two dimensions are being displayed.



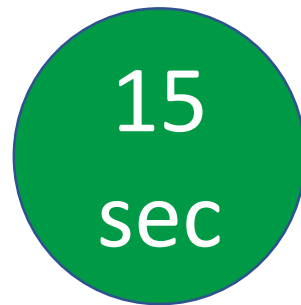
Principle:

Know When to Include 0

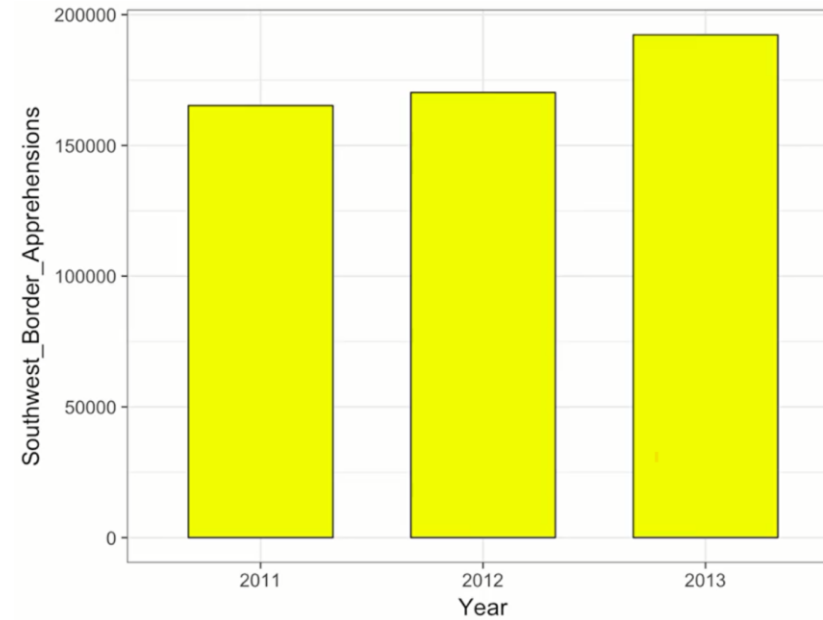
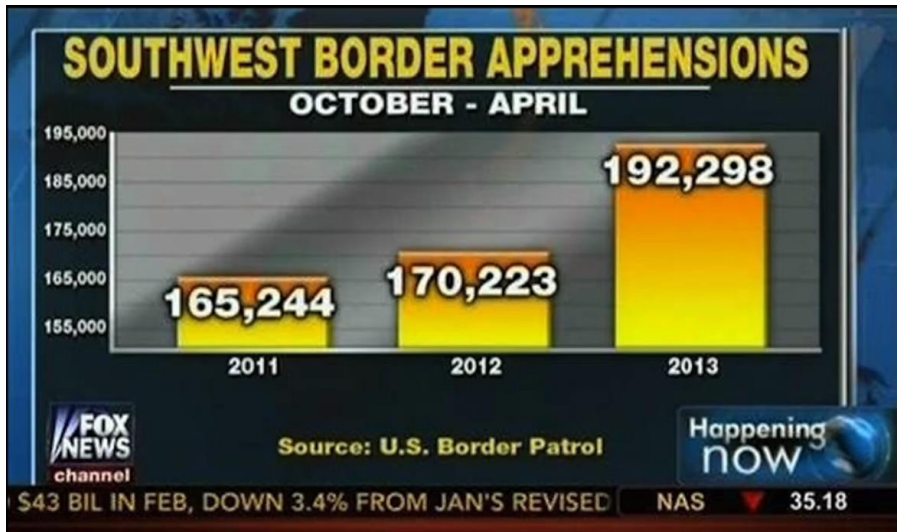
- Fox News Example:



- When using bar plots, it is dishonest not to start the bars at 0.
- This is because by using a bar plot, we are implying the length is proportional to the quantities being displayed. By avoiding 0, relatively small differences can be made to look much bigger than they actually are.
- This approach is often used by politicians or media organizations trying to exaggerate the difference

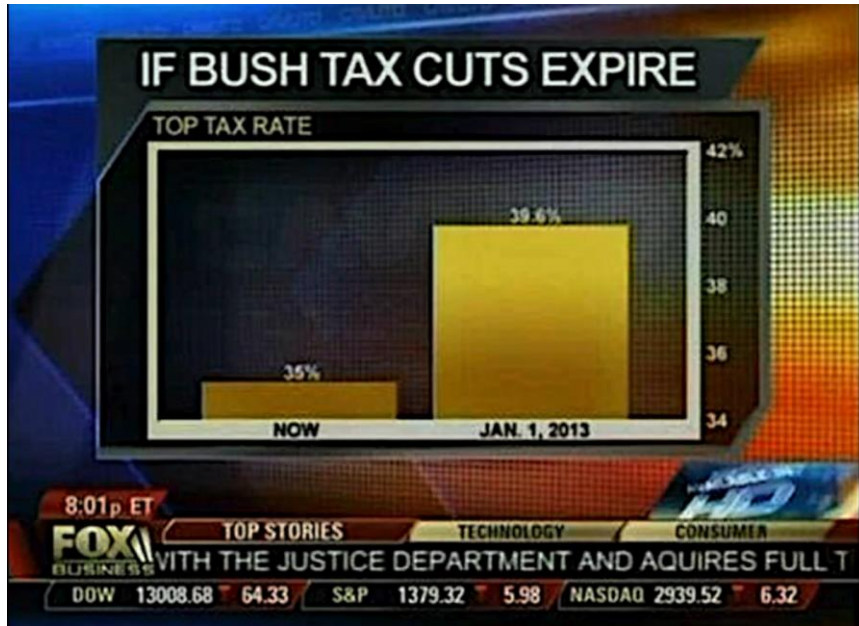


- Fox News Example:

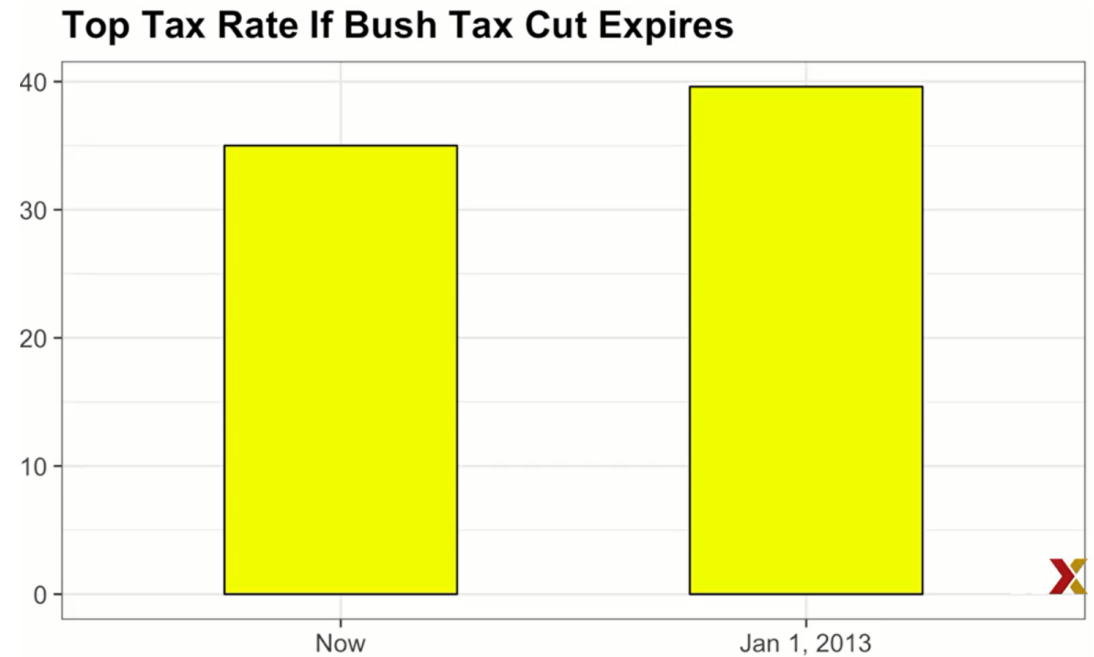


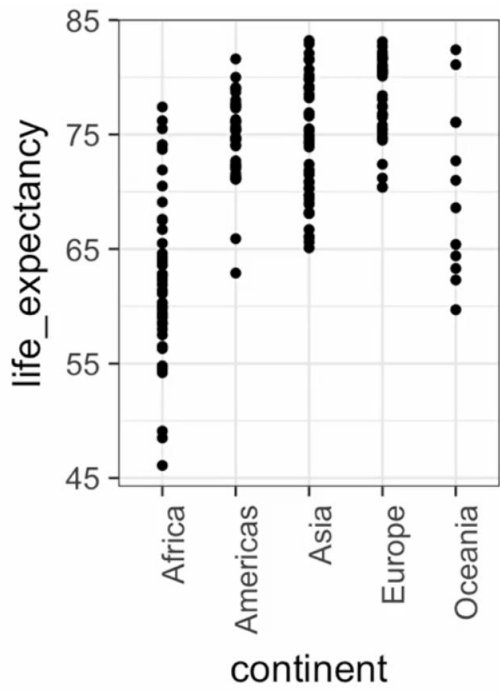
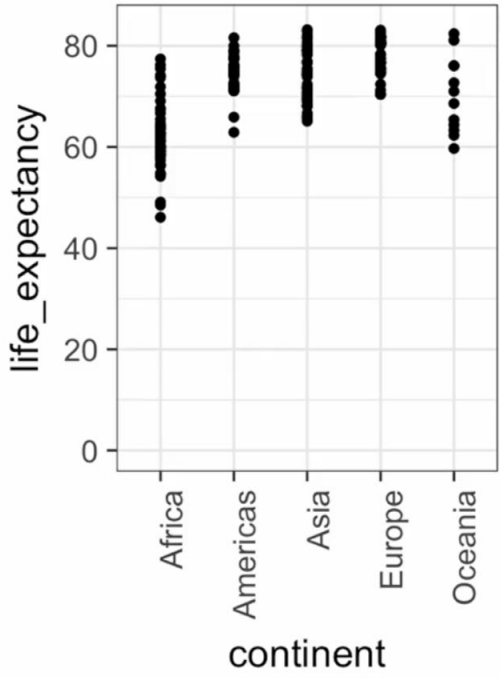
- Look how much bigger the 2013 bar looks compared to the 2011.
- They have only increased by about **16%**.
- Starting the graph at 0 illustrates this clearly. This is what it looks like if the plot includes 0.

Another Fox News Example:



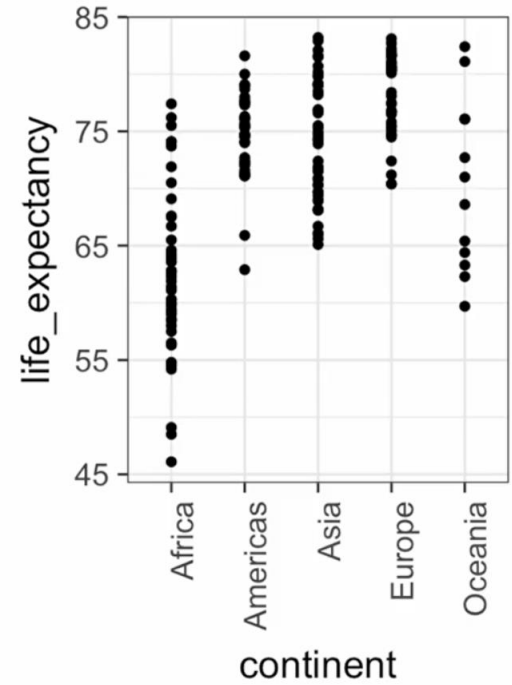
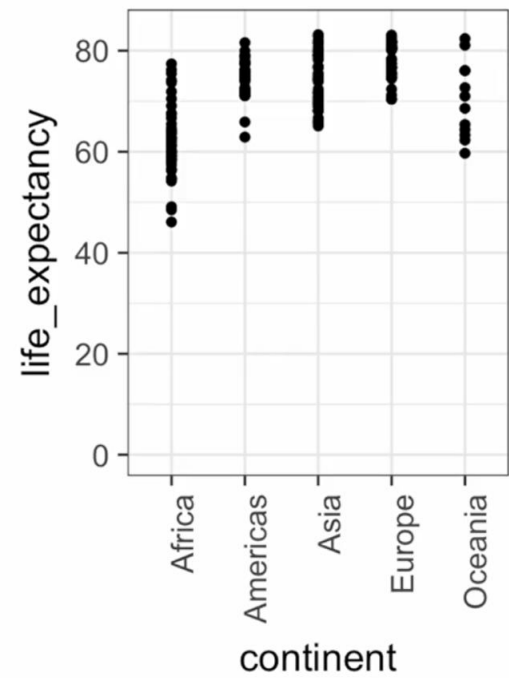
We look at the bar plots, it looks like the January 1, 2013 is about 5 times bigger than the now bar plot.





- The space between 0 and 43 adds no information and makes it harder to appreciate the between and within variability.
- When using position rather than length, then it's not necessary to include 0.
- For example, when we want to compare differences between groups relative to the variability seen within the groups.
- Here's an illustrative example showing the country average life expectancies, stratified into continents, in 2012.

- When using position rather than length, then it's not necessary to include 0.
- For example, when we want to compare differences between groups relative to the variability seen within the groups.
- Here's an illustrative example showing the country average life expectancies, stratified into continents, in 2012.



Principle:

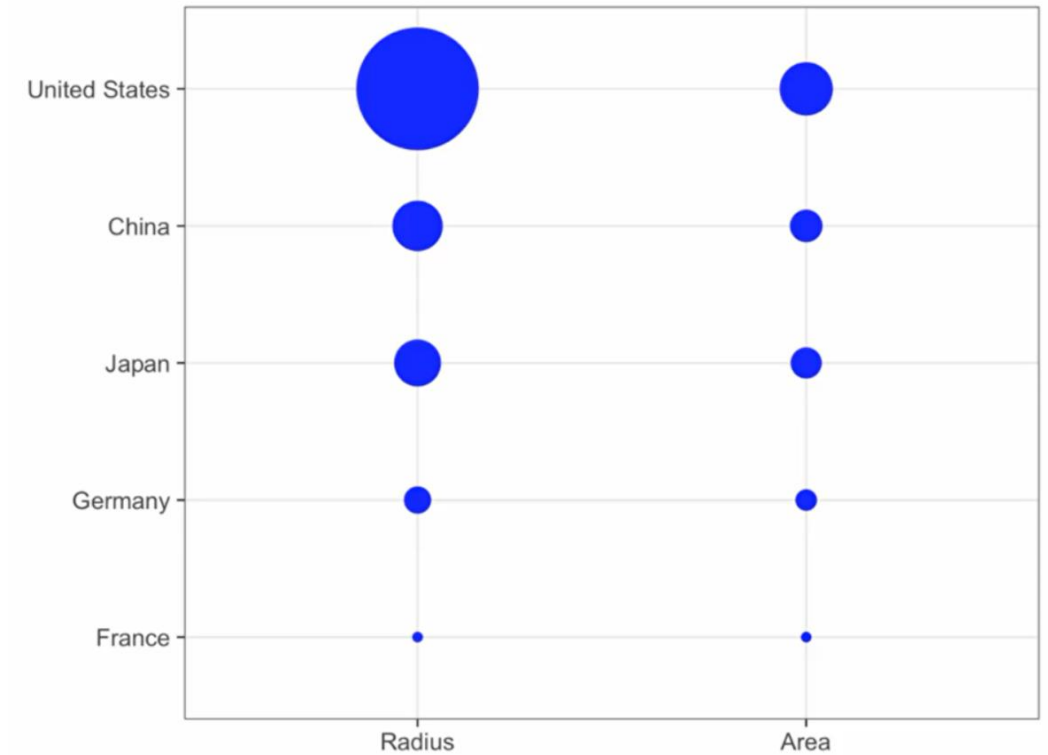
Do Not Distort Quantities

Compare the US GDP to the GDP of four competing nations.

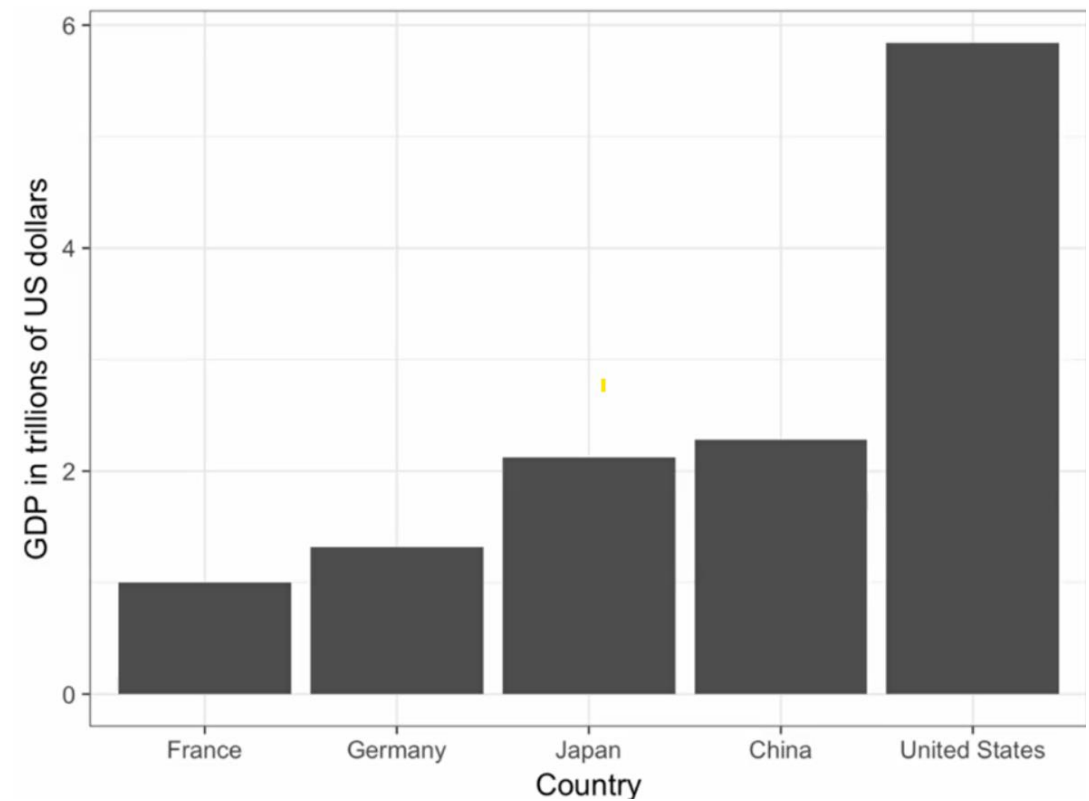


- Note that judging by the area of the circles, the US appears to have an economy over 5 times larger than China, and over 30 times larger than France.
- However, when looking at the actual numbers, one sees that this is not the case. The actual ratios are 2.6, and 5.8 times bigger than China and France respectively.

- The reason for this distortion is that the radius, rather than the area, was made to be proportional to the quantity, which implies that the proportions between the areas is squared. So 2.6 turns into 6.5, and 5.8, turns into 34.1.
- Here's a comparison of the circles we get if we make the values proportional to the radius, that's on the left, and so the area, that's on the right.



- Not surprisingly, ggplot defaults to using area rather than the radius.
- Of course, in this case, we really should not be using area at all, since we can use position and length. Here's the bar plot comparing the GDPs.



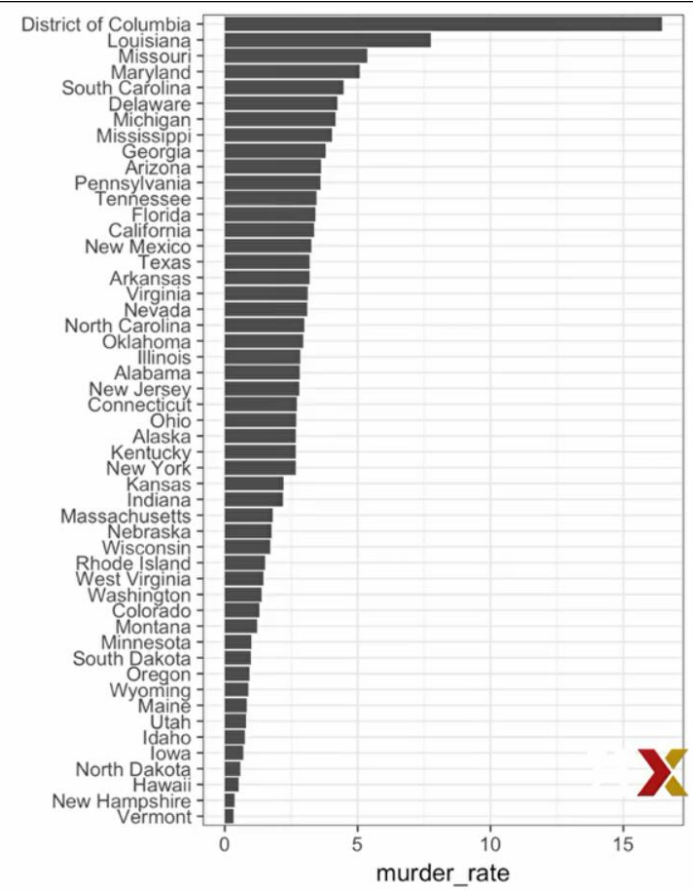
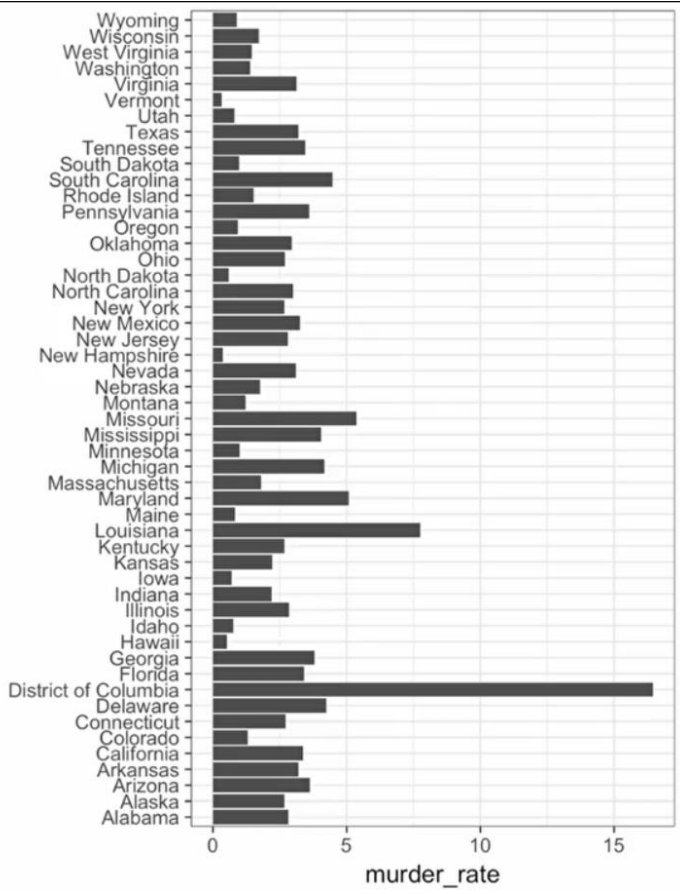
Principle:

Order by a Meaningful Value

- When one of the axes is used to show categories, as done bar plots, the default ggplot behavior is to order the categories alphabetically.
- We rarely want to use alphabetical order. It's arbitrary.
- Instead, we should order by a meaningful quantity using `reorder()` function

Principles

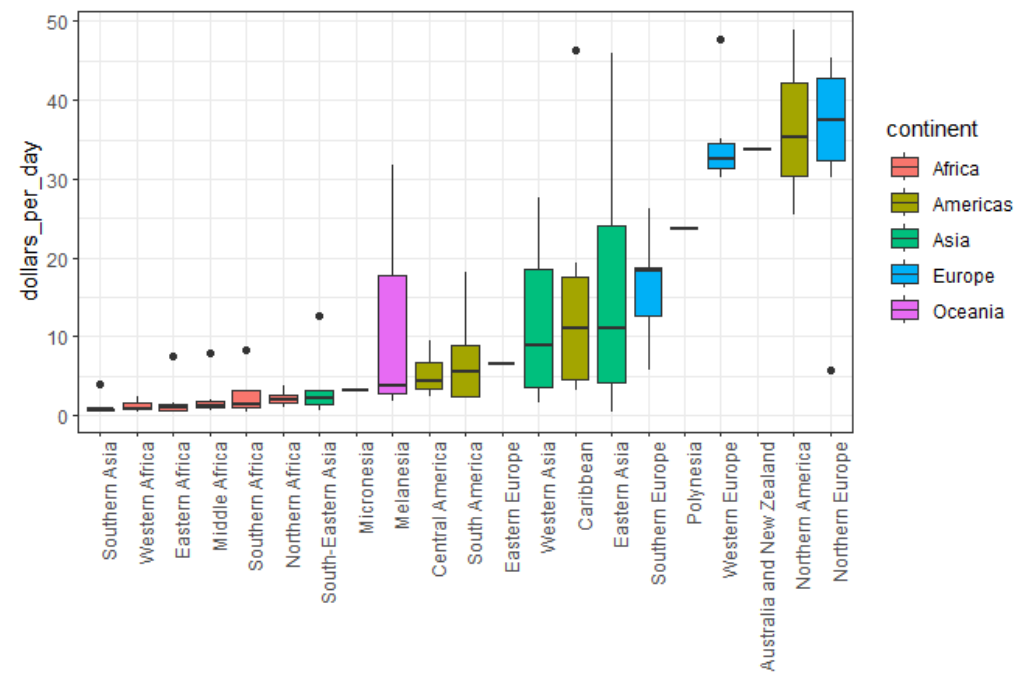
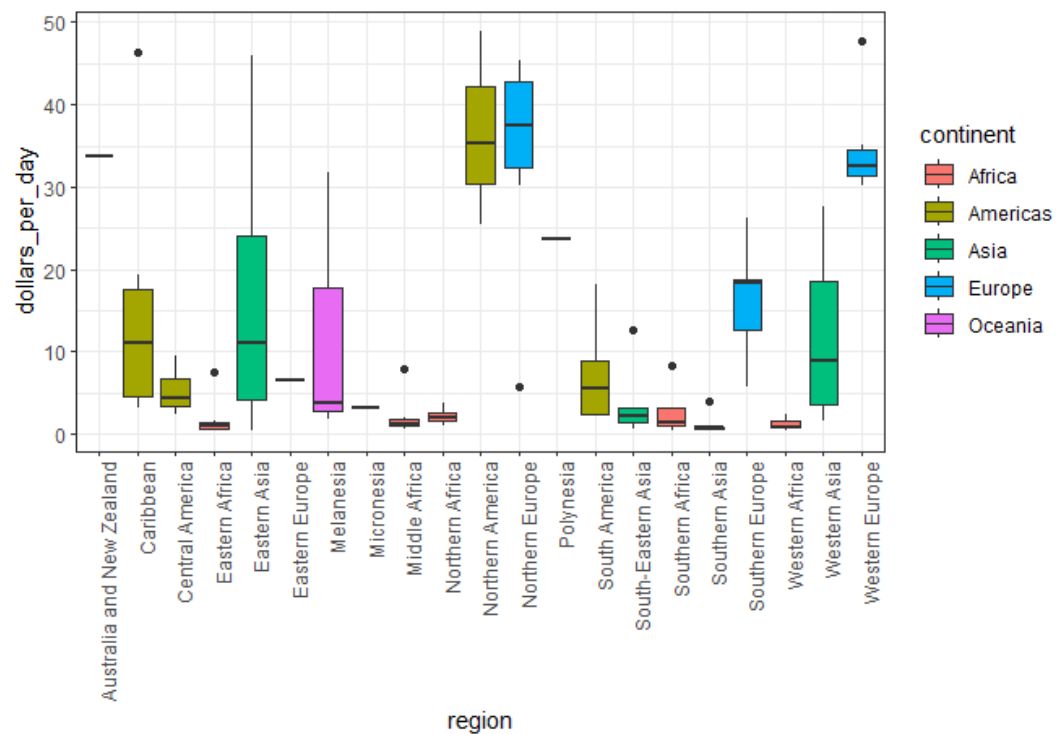
Order by a meaningful value



- To appreciate how the right order can help convey a message, suppose we want to create a plot to compare the murder rates across states.
- We're particularly interested in the most dangerous and the safest states.

Principles Order by a meaningful value

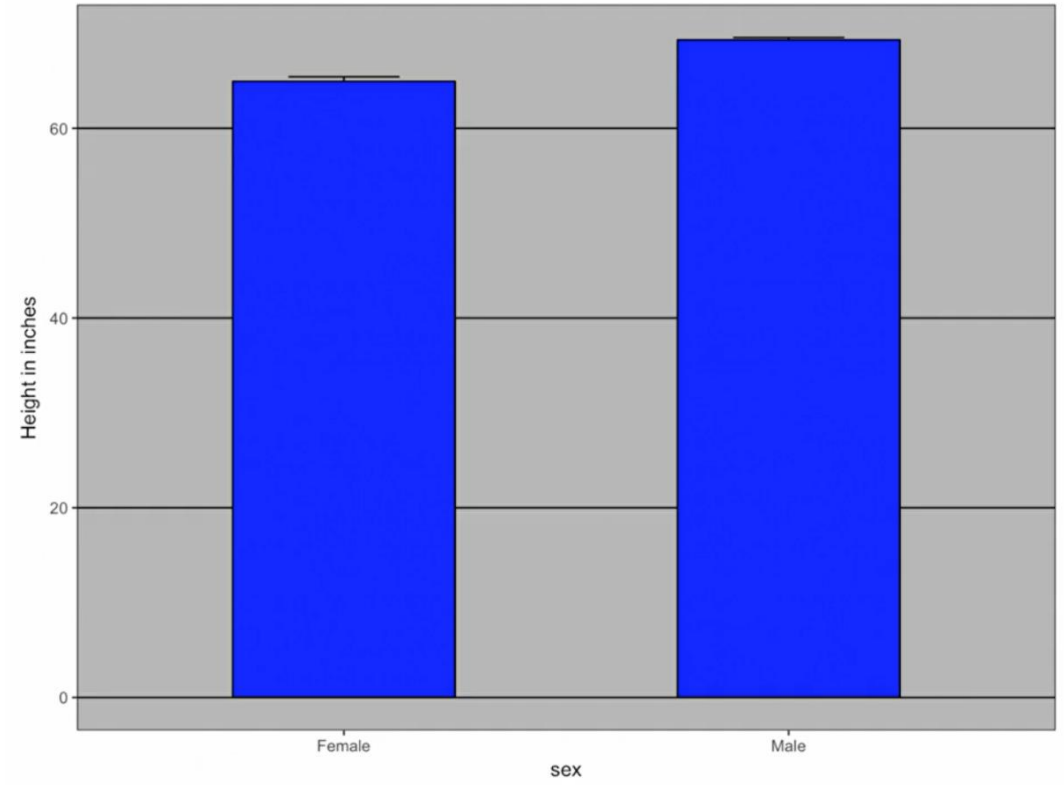
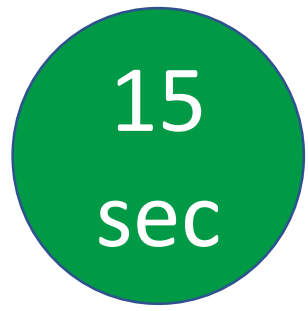
Remember our example related to income distribution across regions.



Principle:

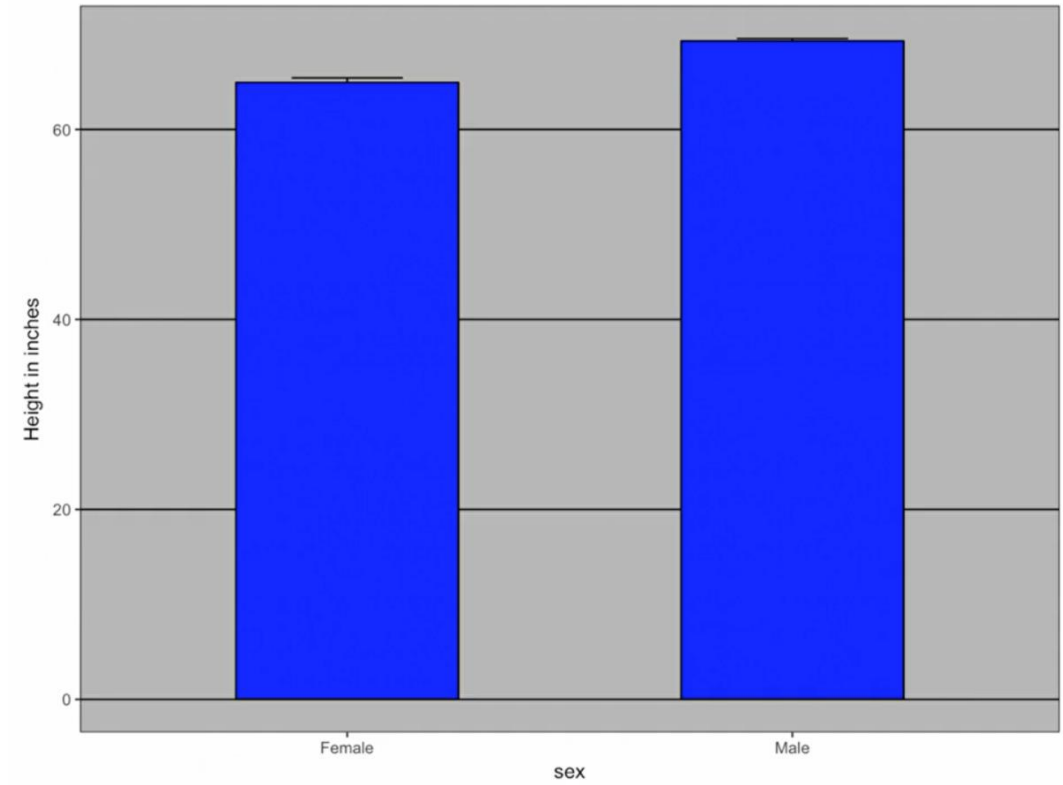
Show the Data

- We now shift our attention to displaying data with a focus on comparing groups.
- Let's assume a DM is interested in the difference in heights between males and females.
- A commonly seen plot used for comparison between groups, popularized by software such as Microsoft Excel, shows the average and the standard error.

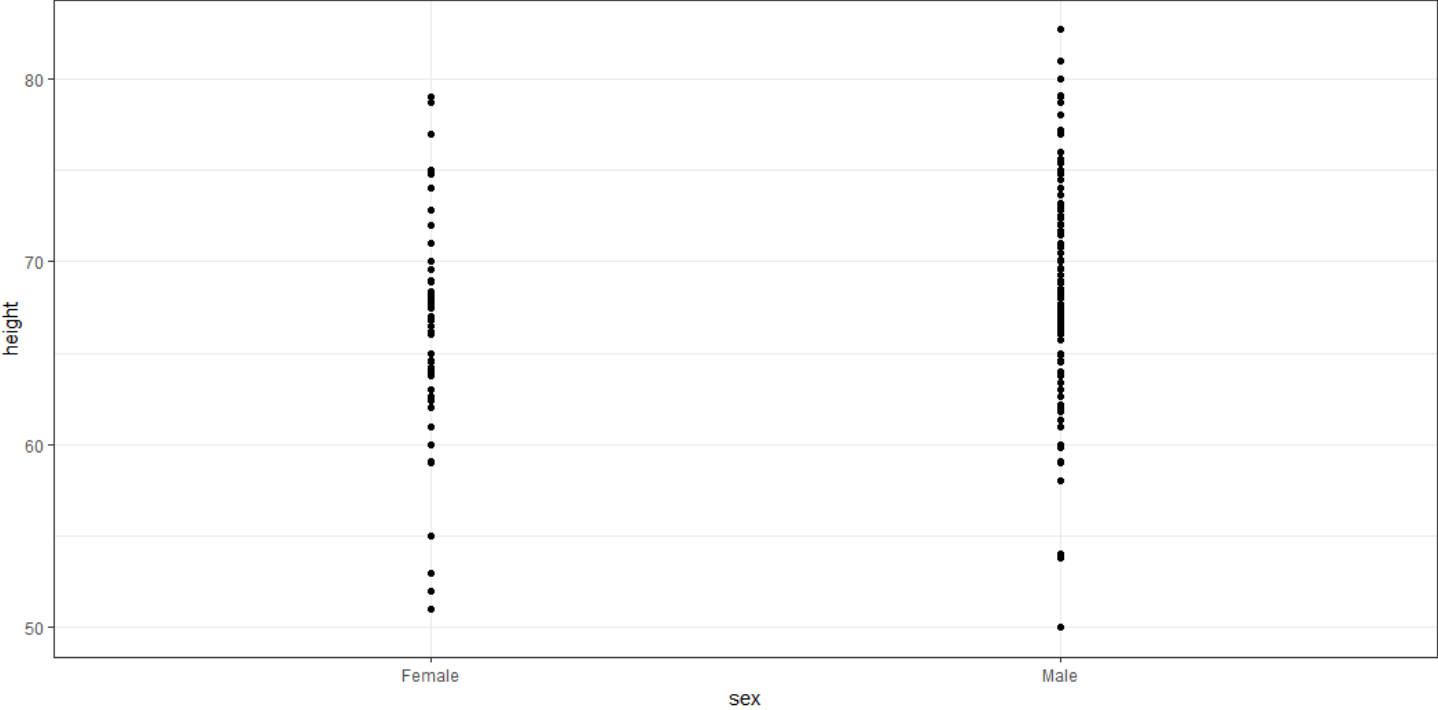


- The average of each group is represented by the top of each bar, and the antenna that we see that expands out is the average plus two standard errors.
- DM will have little information

- Note that the bars go to 0.
- Does this mean there are tiny humans measuring less than one foot?
- Are all males taller than the tallest female?
- Is there a range of heights? D
- M can't answer these questions since we have provided almost no information on the height distribution.
- This brings us back to our principle, show the data.

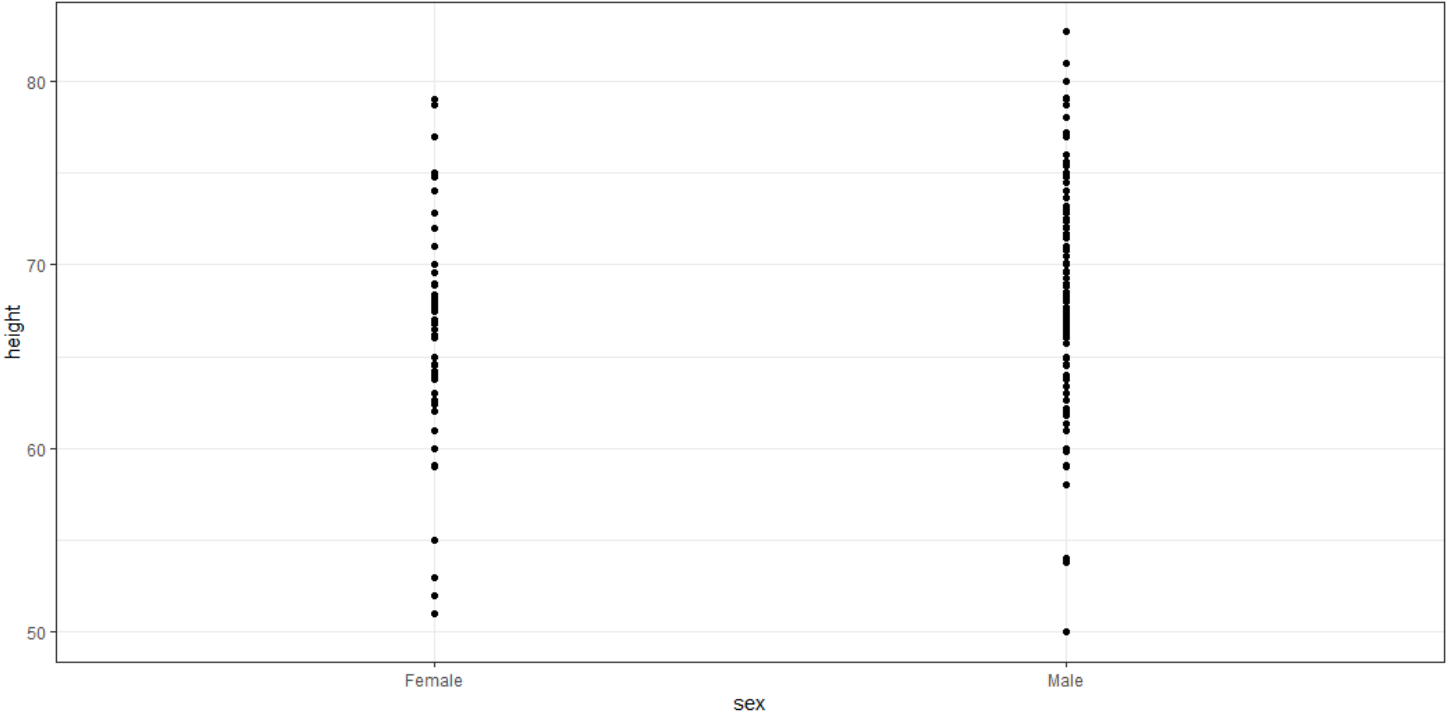
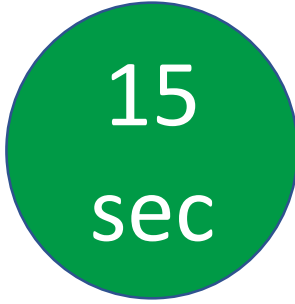


```
library(dslabs)  
library(tidyverse)  
data(heights)  
heights %>% ggplot(aes(sex, height)) + geom_point() + theme_bw()
```



- We now have a more informative plot than the bar plot by simply showing all the points.
- Just this little line of code shows you the points, the heights for females and the height for males.
- However, this plot has limitations as well.

What can be done better?



Since we can't really see all the 216 and 708 points plotted for females and males, respectively.

And many points are plotted above each other so we don't know how many there are.

jitter

Jitter is adding a small random shift to each point. In this case, adding horizontal jitter does not alter the interpretation since the height of the points doesn't change.

alpha blending

making the point somewhat transparent. Without alpha blending, the more points fall on top of each other, the darker the plot gets in that region, which also helps us get a sense of how the points are distributed.



```
heights %>% ggplot(aes(sex, height)) + geom_jitter(aes(color = sex), alpha = 0.3) + theme_bw()
```

sex
● Female
● Male



sex
● Female
● Male

Now since there are so many points, it is more effective to show distribution rather than show individual points.

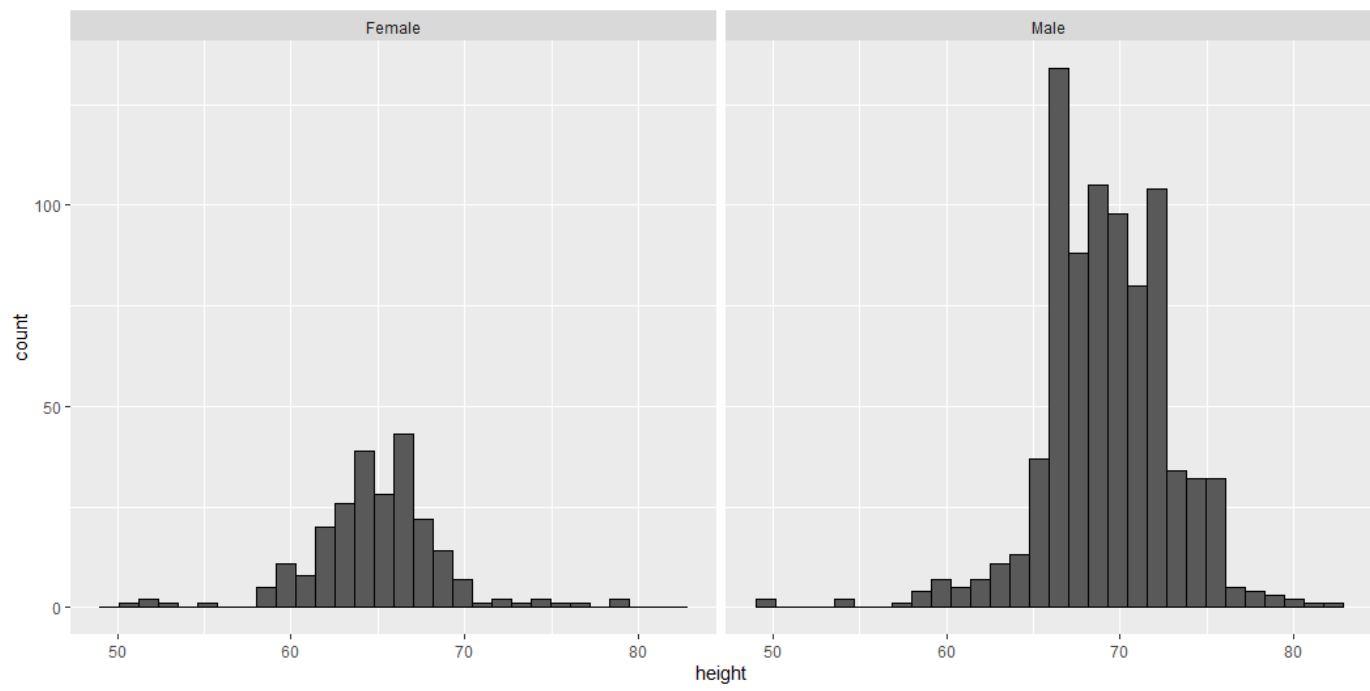
Principle:

Use Common Axes to Ease

Comparisons

Now in this case, showing all the data is not as effective as showing distributions.

```
heights %>%  
  ggplot(aes(height)) +  
  geom_histogram(colour =  
  "black") + facet_grid(.~  
  sex)
```

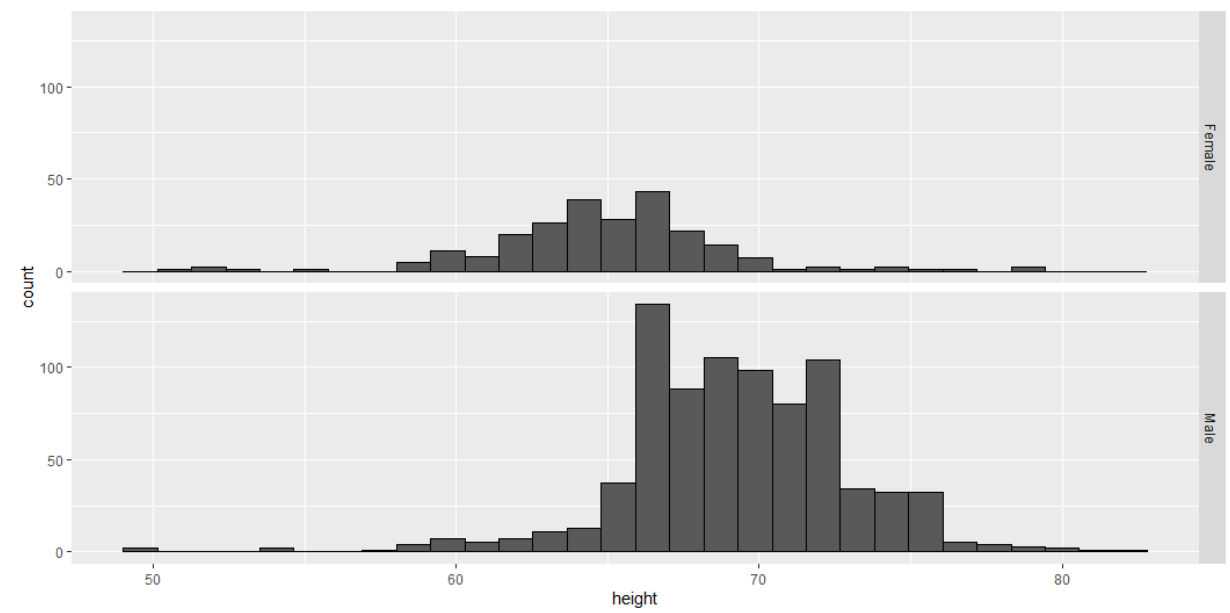


keep the axes the same

align plots

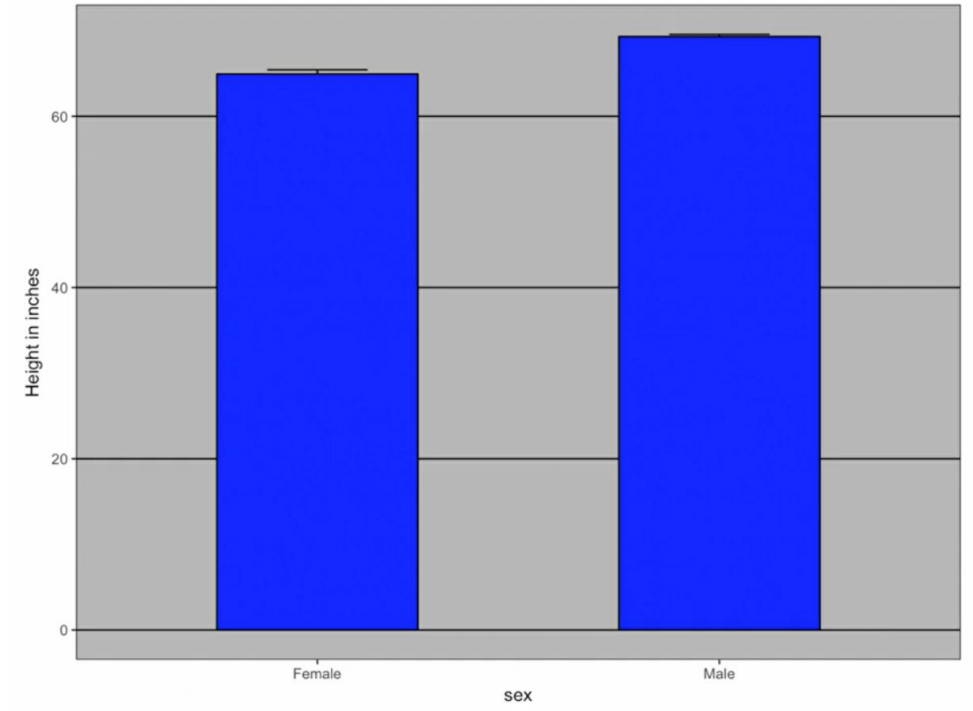
- align plots vertically to see horizontal changes, and horizontally to see vertical changes. In this example, there is a shift towards right. Putting plots vertically is much more helpful.

```
heights %>% ggplot(aes(height)) + geom_histogram(colour = "black") + facet_grid(sex ~ .)
```





SEX
● Female
● Male

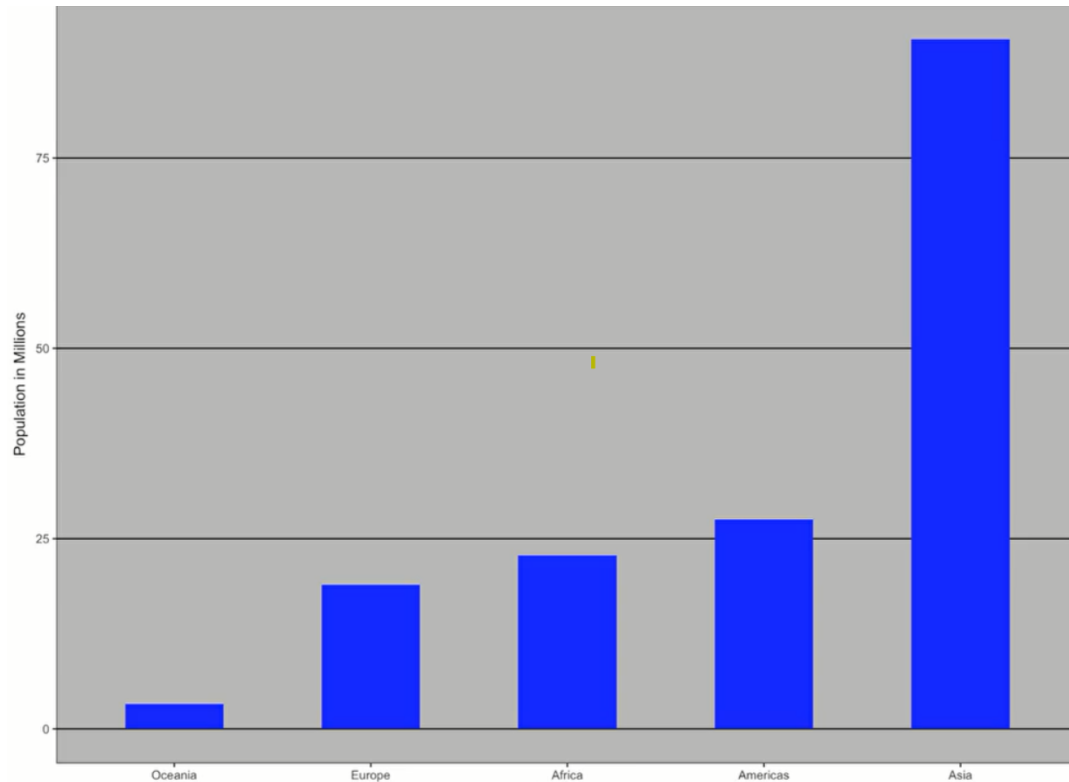
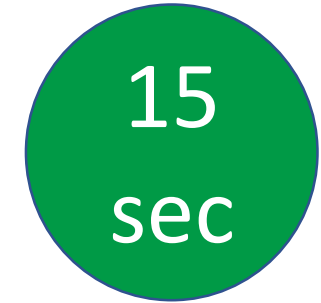


They also use common axes, which is good. Yet, they have other problems that we already discussed.

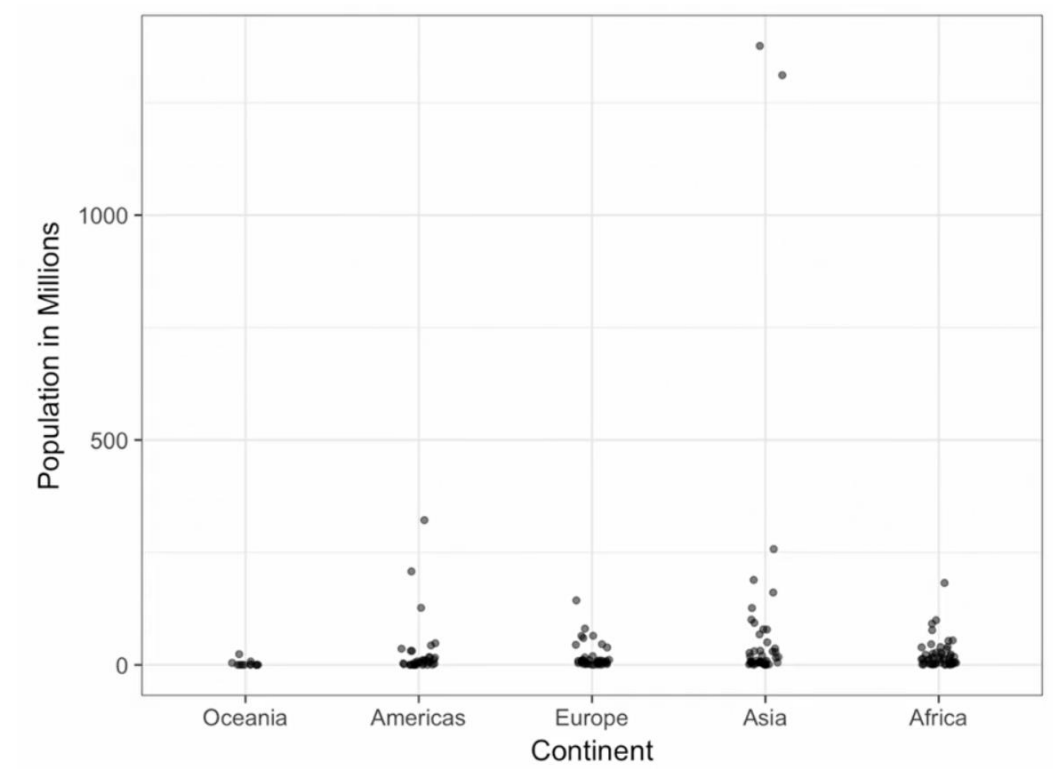
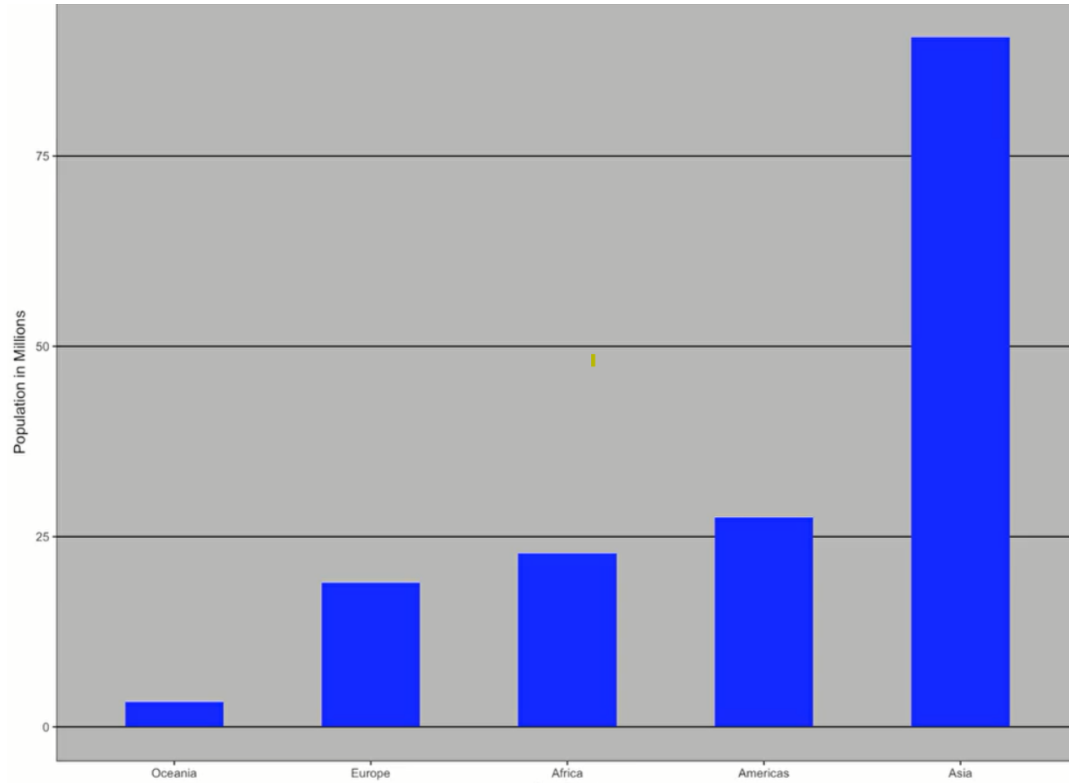
Principle:

Consider Transformations

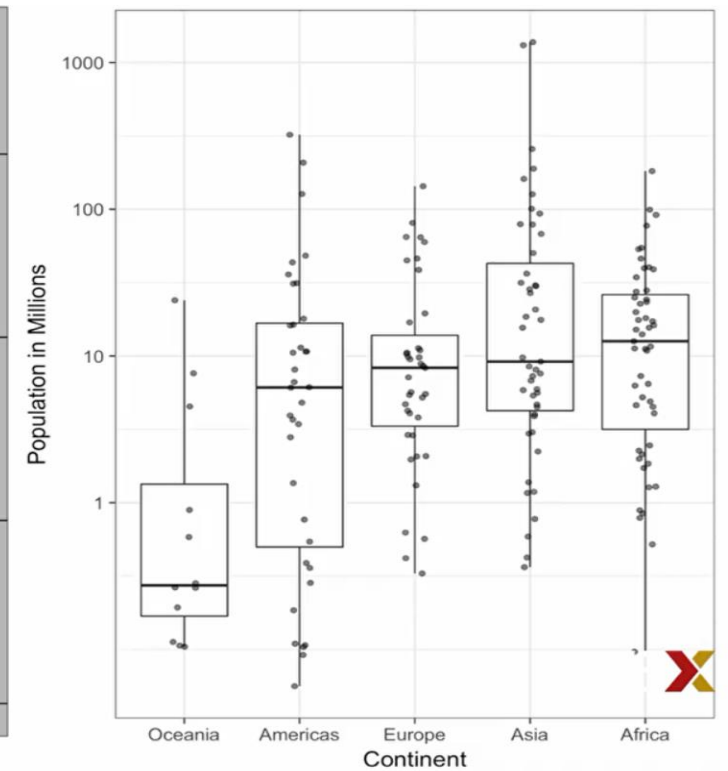
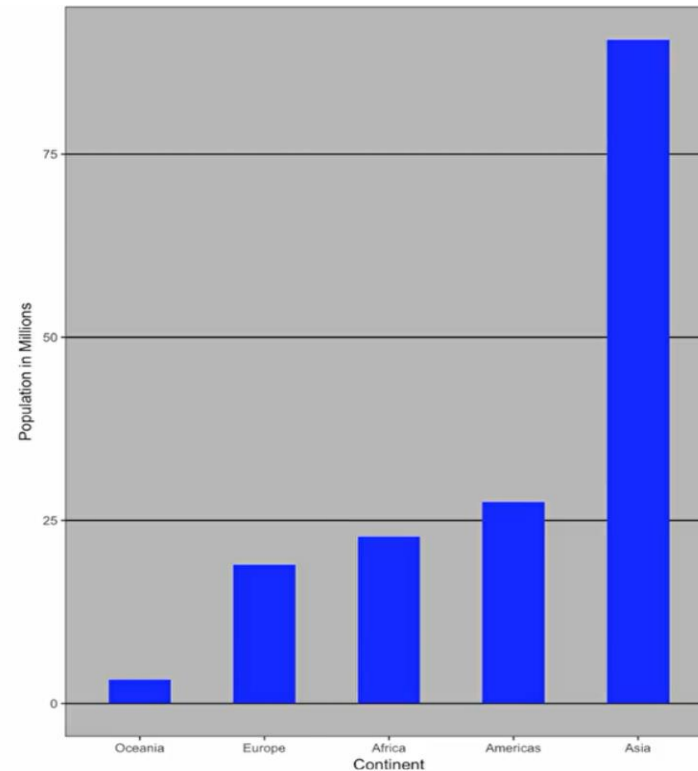
- We have motivated the use of the log transformation in cases where the changes are multiplicative.
- Population size was an example in which we found a log transformation to yield a more informative plot.
- The combination of incorrectly using bar plots, when a log transformation is merited, can be particularly distorting.



- From this plot, one would conclude that countries in Asia are much more populous than other continents. Is this a correct assumption?
- **This is due to two very large countries, which we can assume are India and China.**



- Here, using a log transformation provides a much more informative plot.
- We compare the original bar plot to a box plot using the log-scale transformation for the y-axis.

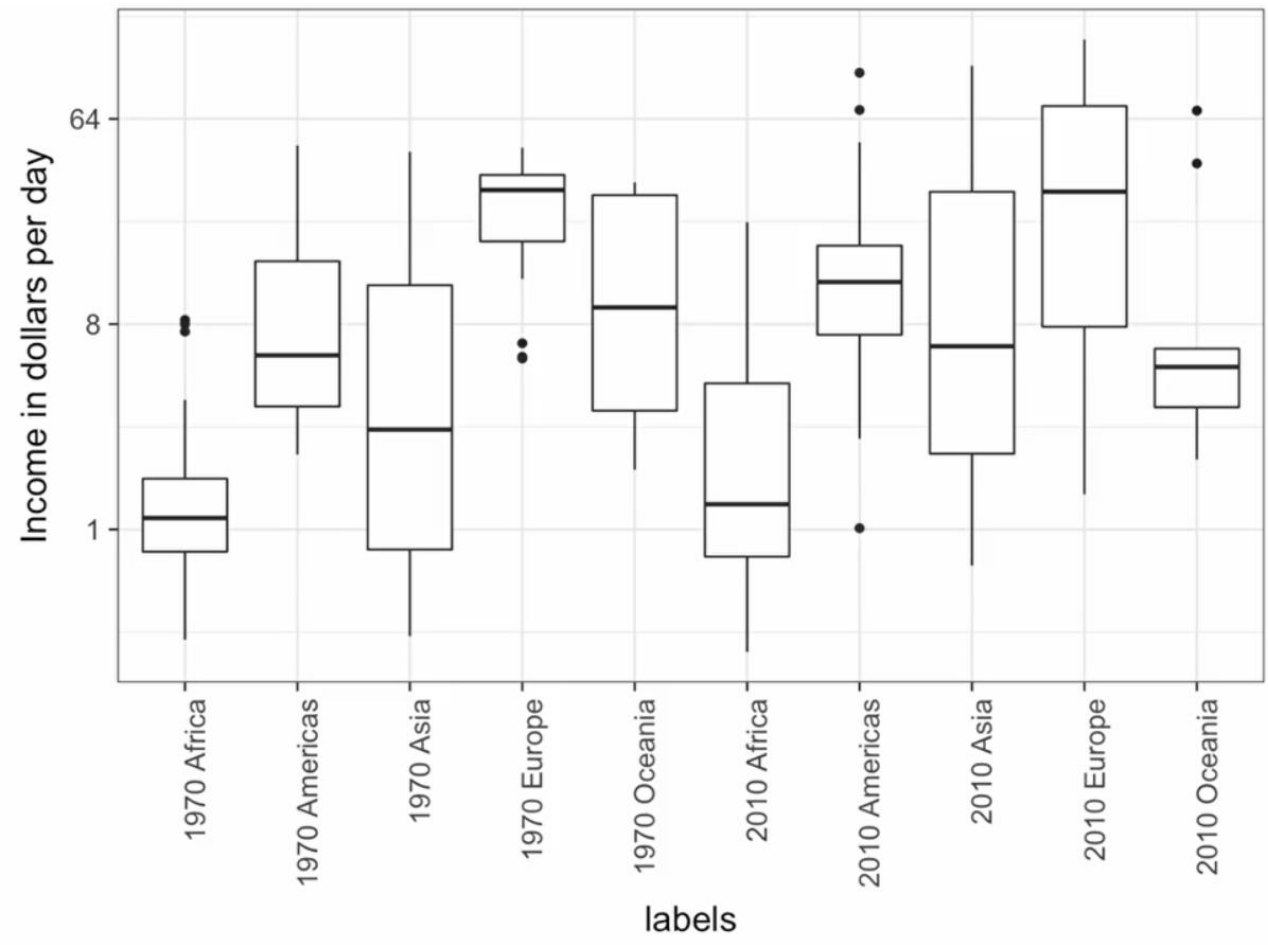
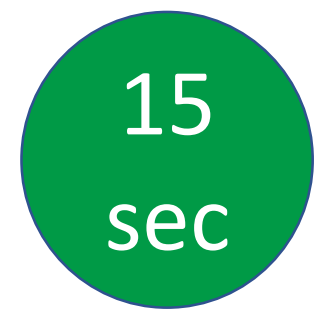


- More informative that the box plot is over the bar plot.
- For example, we see that Africa has a higher median population size than Asia.

Principle:

Compared Visual Cues Should be

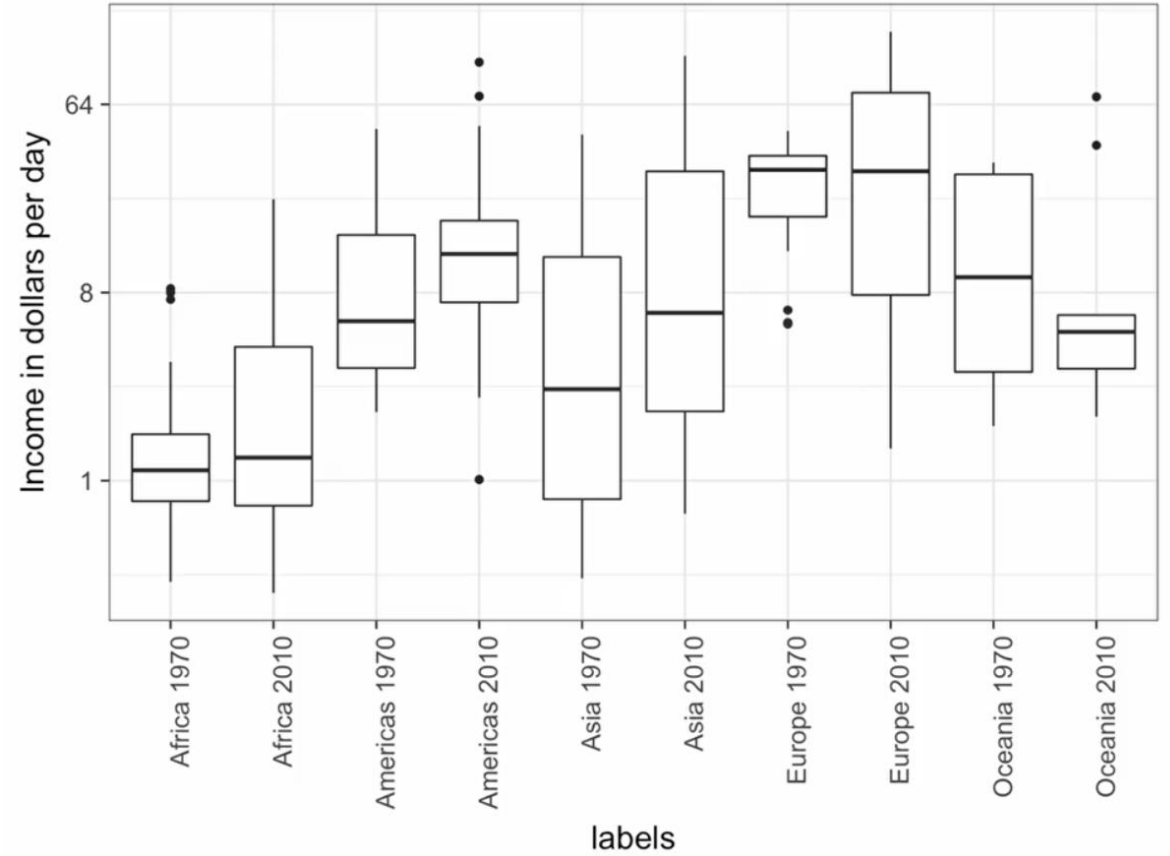
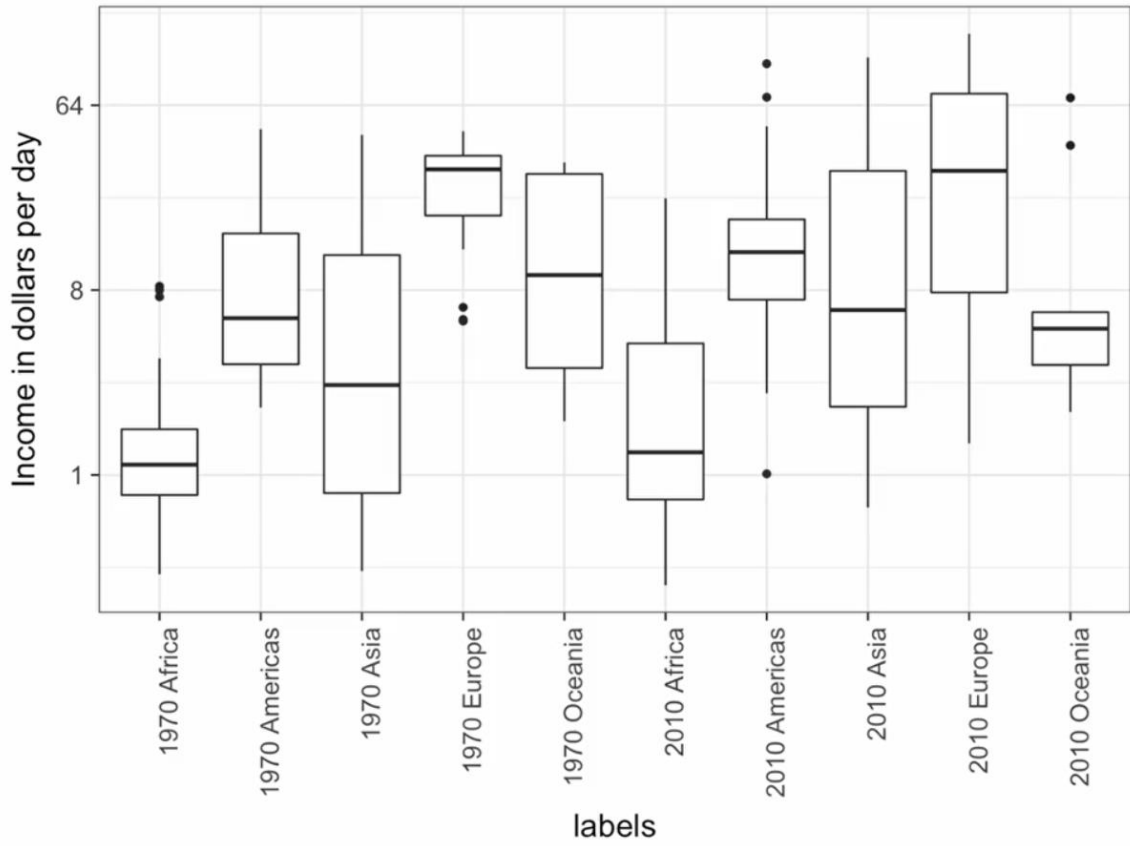
Adjacent to Ease Comparisons

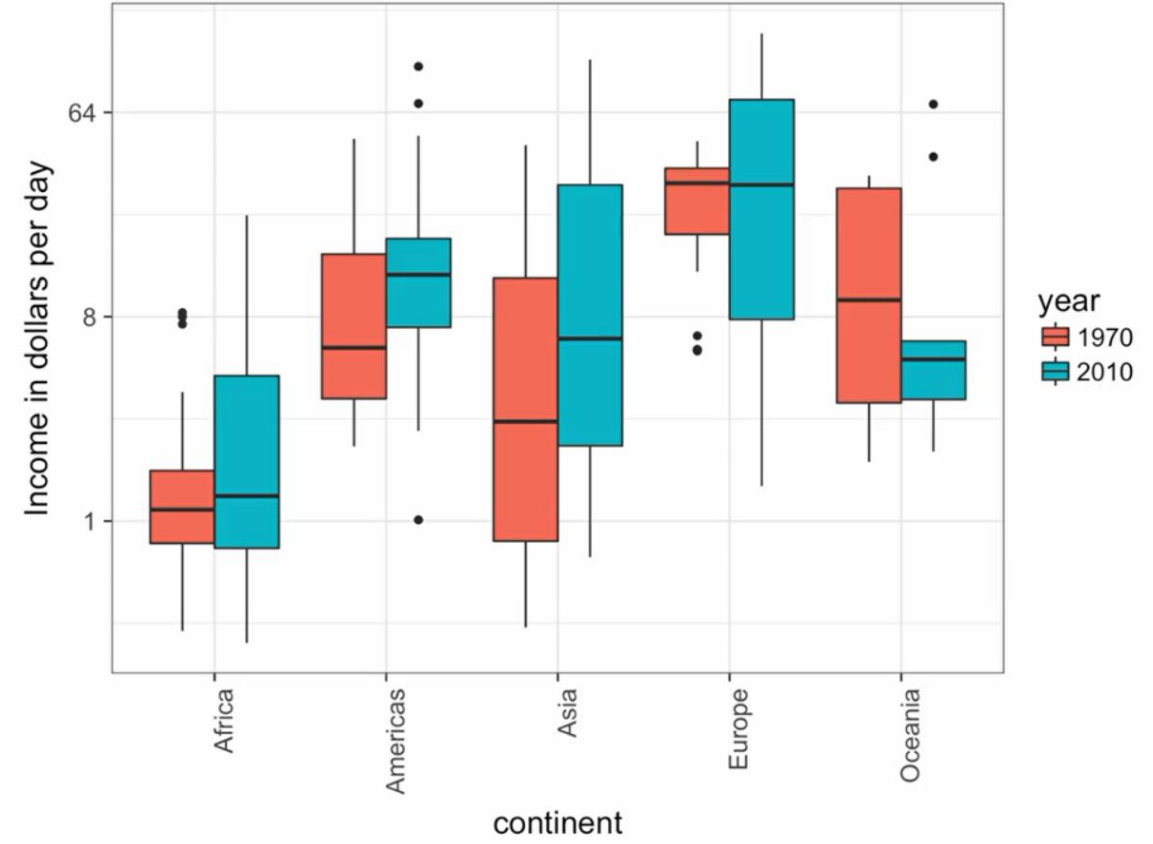
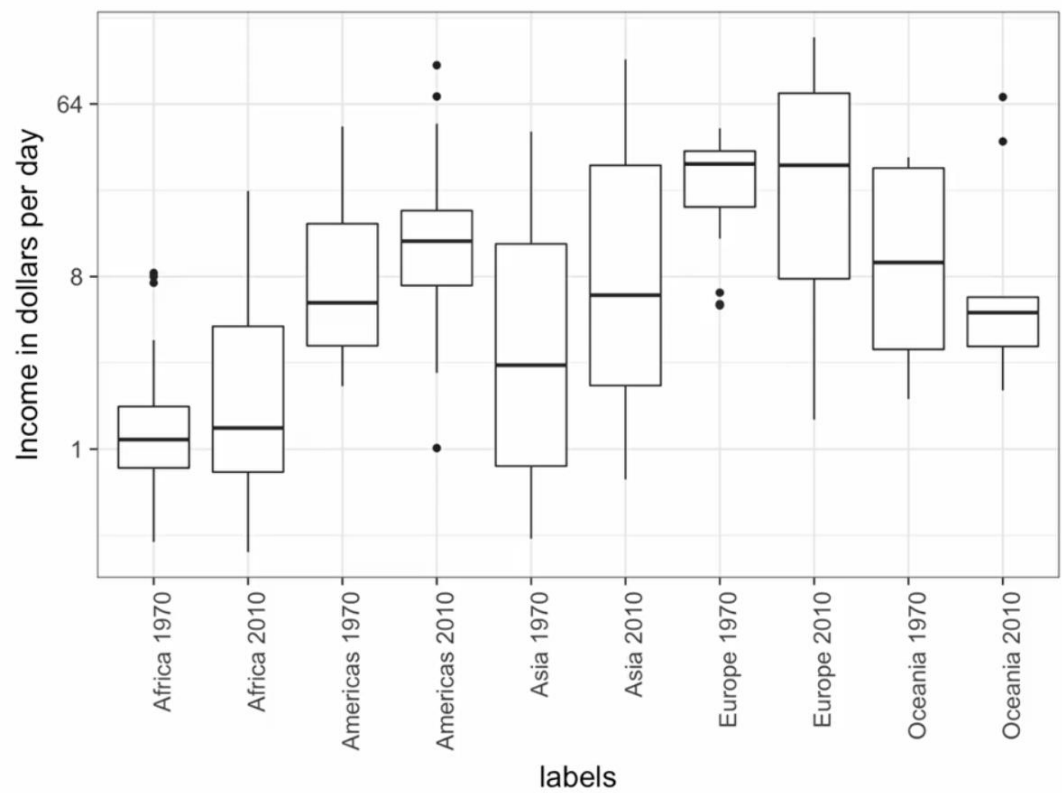


➤ When comparing income data between 1970 and 2010, across regions, we made a figure similar to this one.

➤ **What is the problem here?**

The default in ggplot is to order alphabetically. So the labels with 1970 come before the labels with 2010, making that comparison challenging.



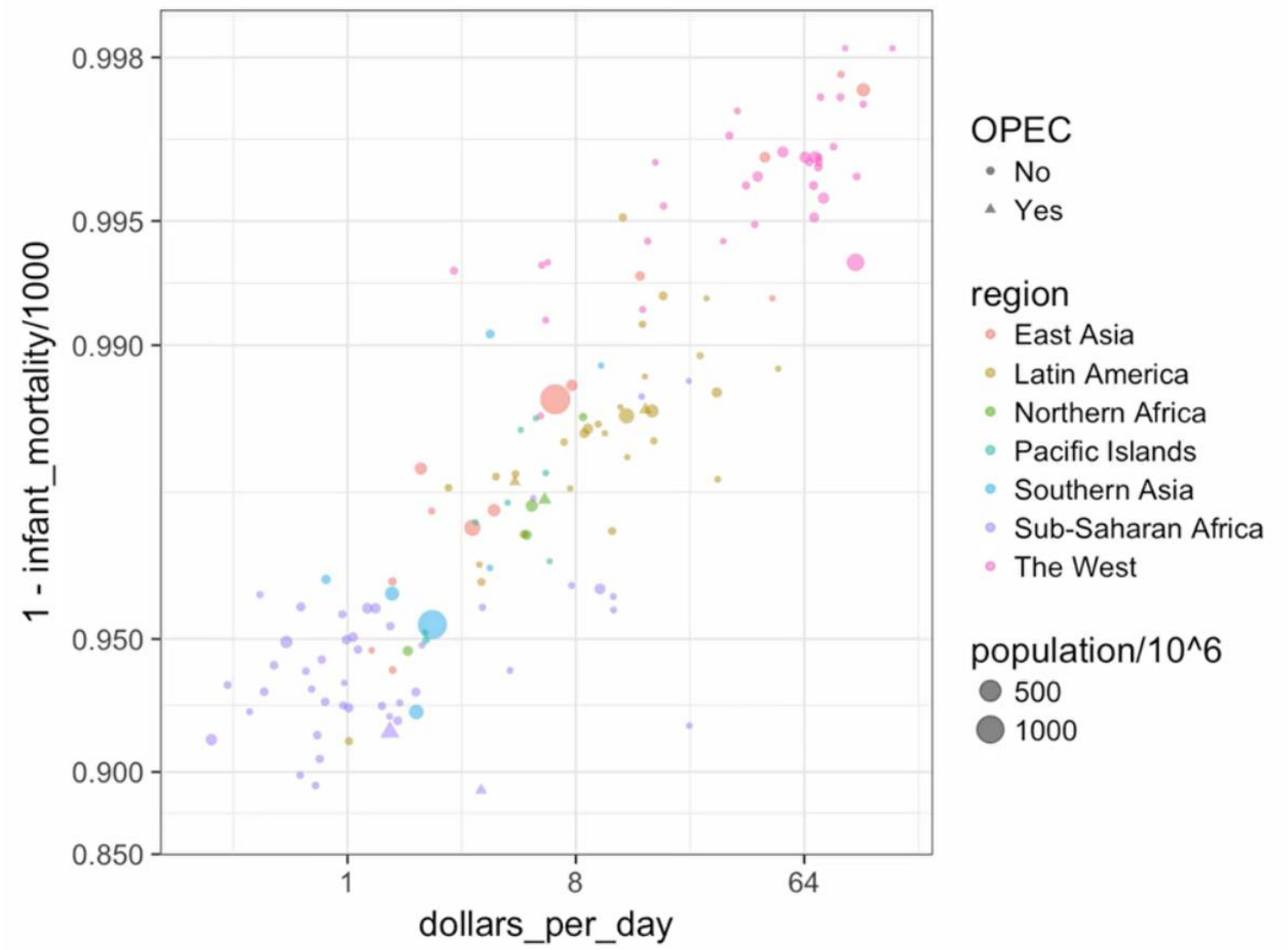


- The comparison becomes even easier if we use color to distinguish 1970 to 2010.
- Using color is another way to ease comparisons.

Principle:

Encoding a Third Variable

- We can show the relationship between infant survival rates and average income using a scatter plot.
- Here's a version of this plot where we encode **three more variables, OPEC membership, region, and population size.**



- Note that we encode categorical variables with color hue and shape.
- These shapes can be controlled with a shape argument. Here are the shapes available for use in R.



Case Study:

Vaccines

- In the 19th century, before herd immunization was achieved through vaccination programs, deaths from infectious diseases, like smallpox and polio, were common.
- However, today, despite all the scientific evidence for their importance, vaccination programs have become somewhat controversial.
- The controversy started with a paper published in 1988 and led by Andrew Wakefield claiming there was a link between the administration of the measles, mumps, and rubella MMR vaccine, and the appearance of autism and bowel disease.
- Sensationalist media have lead public to believe that vaccines were harmful.
- Effective communication of data is a strong antidote to misinformation and fear mongering.



The data used in these plots were collected, organized, and distributed by the Tycho project. They include weekly reported counts data for 7 diseases from 1928 to 2011 from all 50 states.

```
library(dslabs)
data(us_contagious_diseases)
str(us_contagious_diseases)

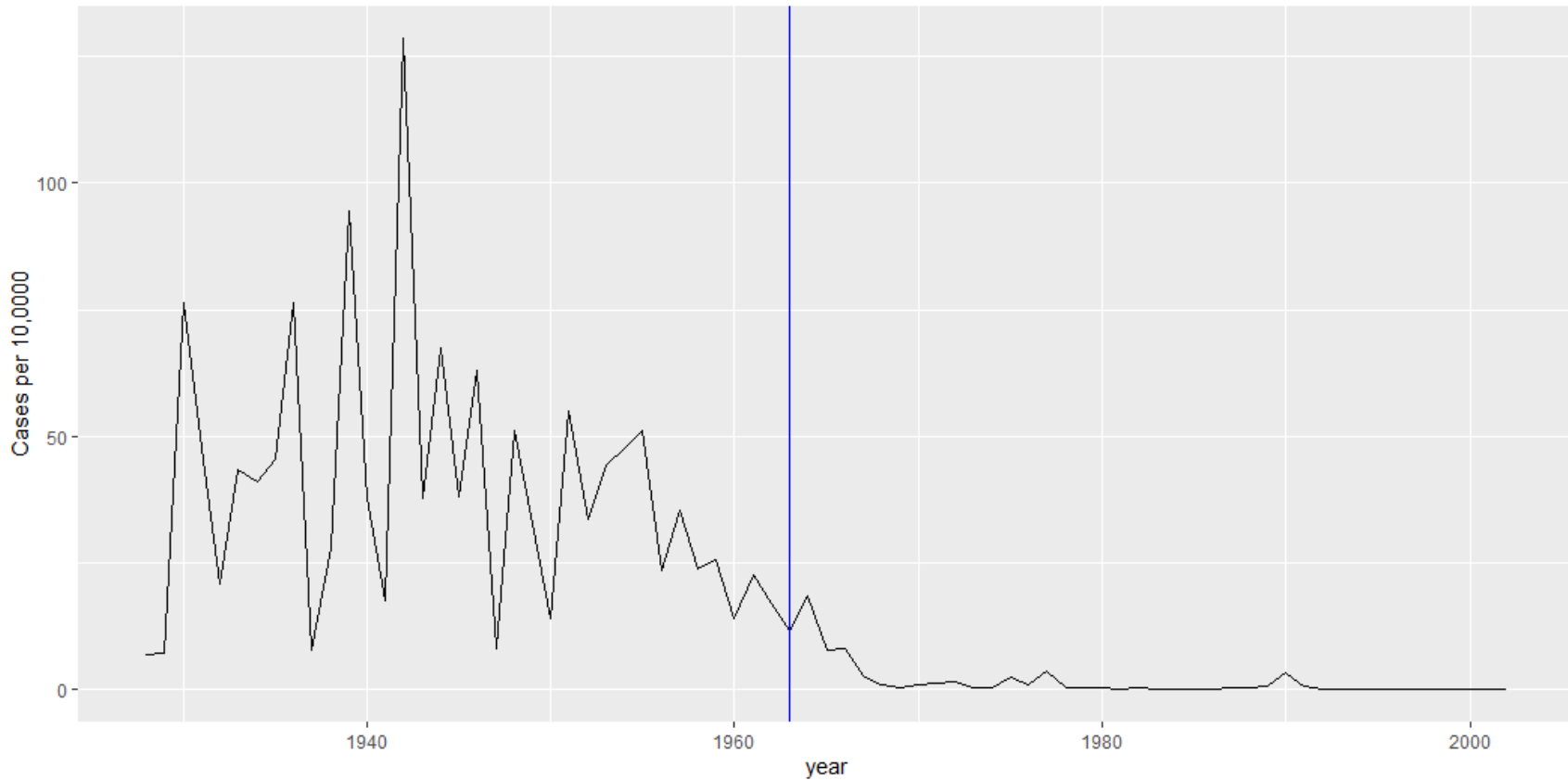
the_disease <- "Measles"

dat <- us_contagious_diseases %>% filter(!state %in% c("Hawaii", "Alaska") & disease
== the_disease) %>% mutate(rate = count / population * 10000) %>% mutate(state =
reorder(state, rate))
```

It includes a per 100,000 rate, orders states by average value of disease, and removes Alaska and Hawaii, since they only became states in the late 50s.

Let's look at the California

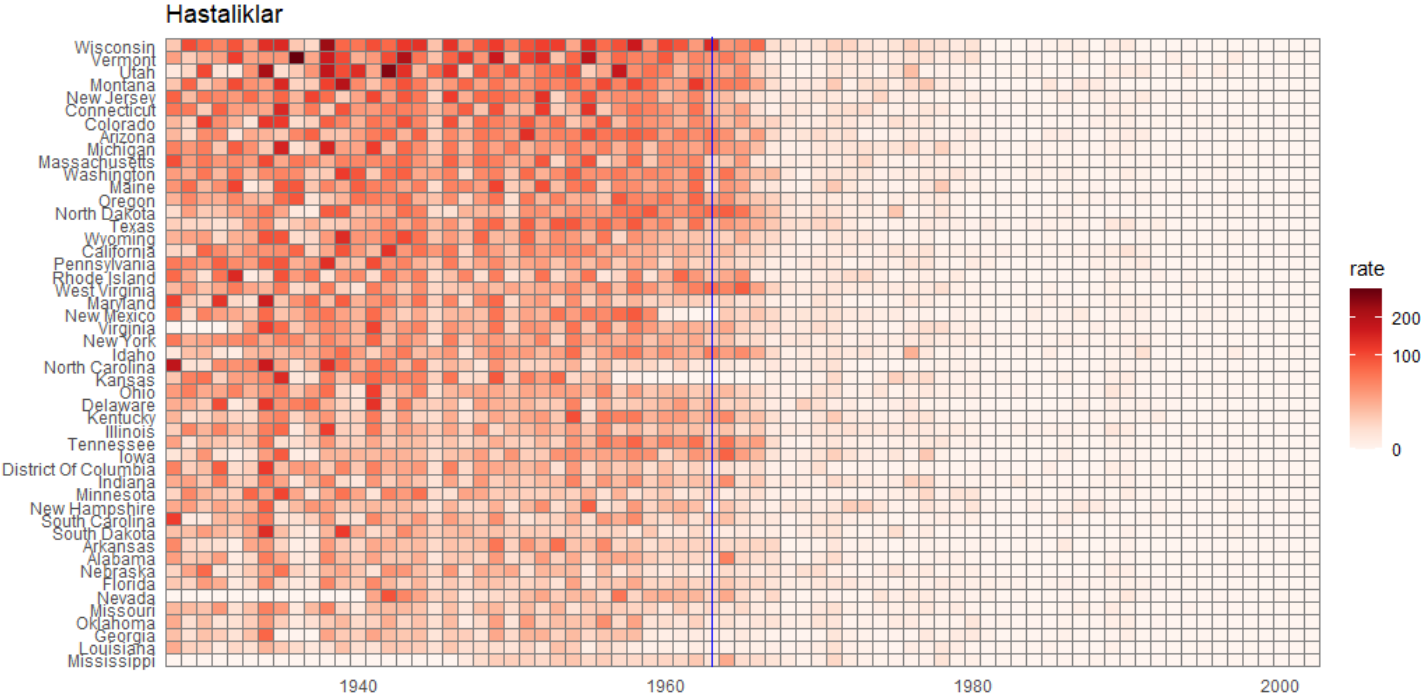
```
dat %>% filter(state == "California") %>% ggplot(aes(year, rate)) + geom_line()  
+ ylab("Cases per 10,000") + geom_vline(xintercept = 1963, col = "blue")
```



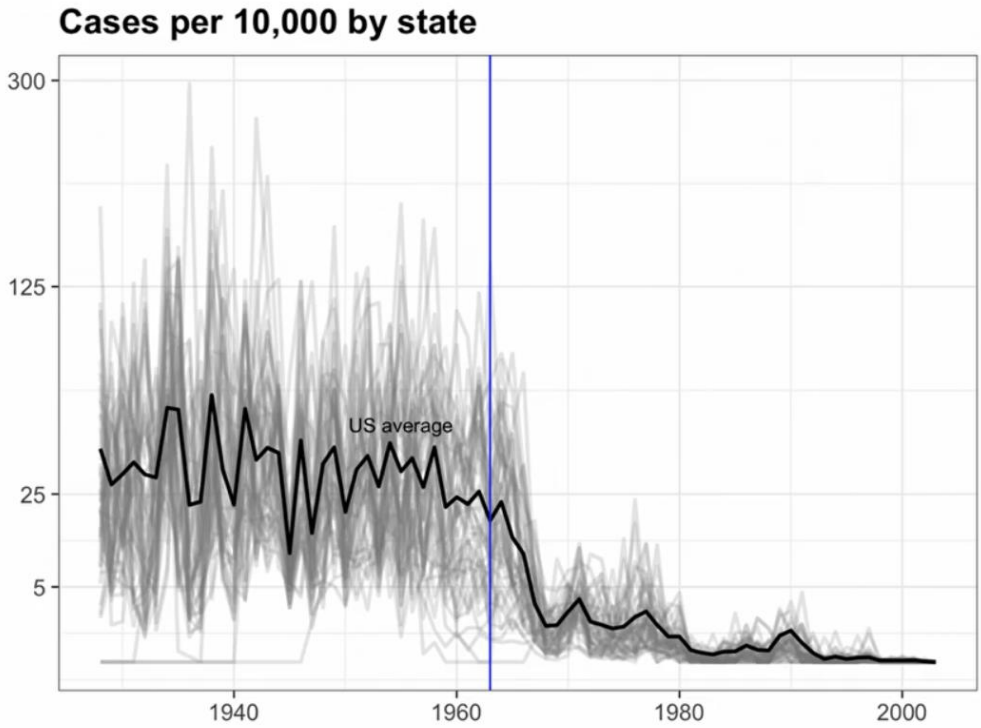
- can we show data for all states in one plot?
- We have **three variables** to show, **year, state, and rate**.

```
library(RColorBrewer)

dat %>% ggplot(aes(year, state, fill = r
+ geom_tile(color="grey50") +
scale_x_continuous(expand = c(0,0)) +
scale_fill_gradientn(colors =
brewer.pal(9,"Reds"), trans = 'sqrt') +
geom_vline(xintercept=1963, col= "blue")
theme_minimal() + theme(panel.grid =
element_blank())) + ggtitle("Hastaliklar"
xlab("") + ylab(""))
```



- One limitation of this plot is that it uses color to represent quantity, which we earlier explained makes it a bit harder to know exactly how high it is going.
- Position and length are better cues.

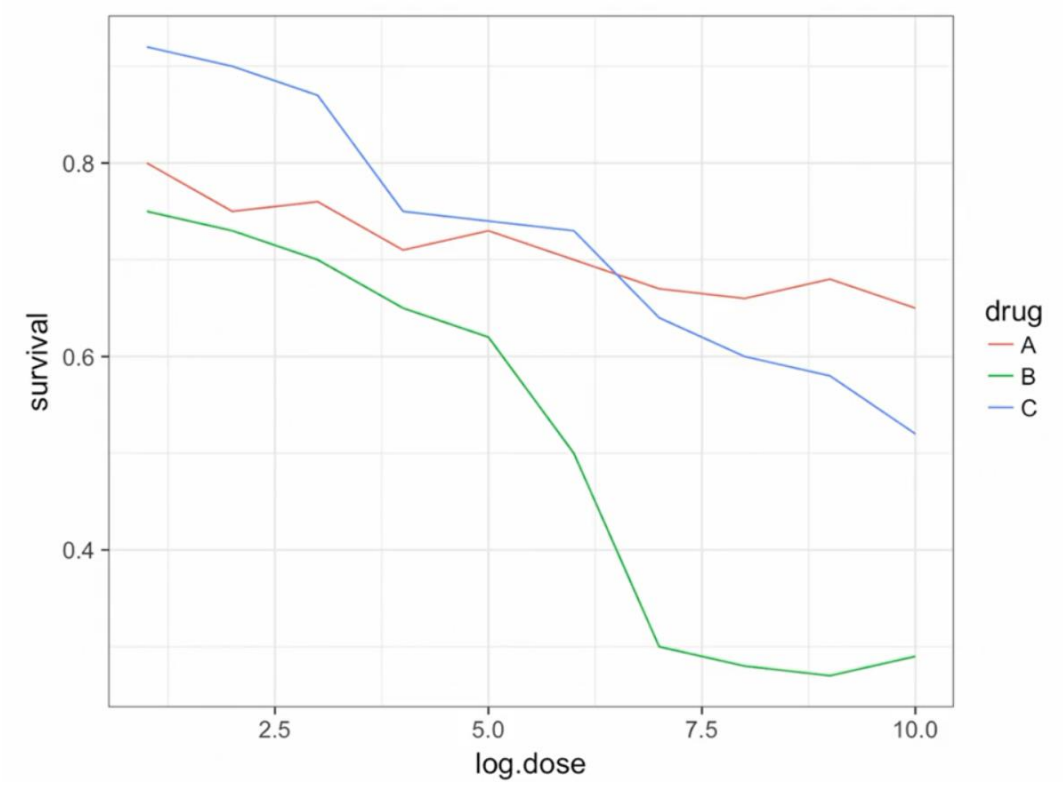
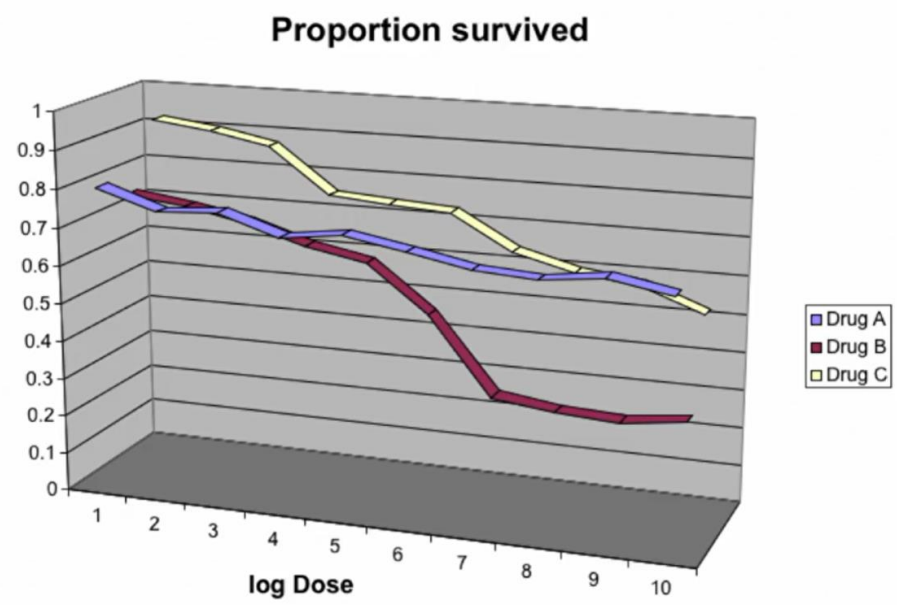


It shows very clearly how after the vaccine was introduced the rates went down across all states. It shows the same information as our previous plot, but now we can actually see what the values are.

Principle

avoid pseudo and gratuitous 3D plots

- Humans are not good at seeing in three dimensions.
- Our limitation is even worse when it's pseudo-three-dimensional, as it is when you put it on a page or a web page.



DNA Fingerprinting: A Review of the Controversy Kathryn Roeder Statistical Science Vol. 9, No. 2 (May, 1994), pp. 222-247

- Humans are not good at seeing in three dimensions.
- Our limitation is even worse when it's pseudo-three-dimensional, as it is when you put it on a page or a web page.

