Data viz w/ggplot2

Part 1 - Exploratory plots

PA 434 - Federica Fusi

Before we start

Where we are

1. Dollar-sign (brief intro)

Where we are

1. Dollar-sign (brief intro)
2. Tidy data (a.k.a. cleaning your data for analysis)
    a. pivot_wider, pivot_longer
    b. managing columns
    c. recoding
    d. missing data

Where we are

1. Dollar-sign (brief intro)
2. Tidy data (a.k.a. cleaning your data for analysis)
    a. pivot_wider, pivot_longer
    b. managing columns
    c. recoding
    d. missing data
3. Extracting information with dplyr

Where we are

1. Dollar-sign (brief intro)
2. Tidy data (a.k.a. cleaning your data for analysis)
    a. pivot_wider, pivot_longer
    b. managing columns
    c. recoding
    d. missing data
3. Extracting information with dplyr
4. Data visualization
    a. Exploratory plots
    b. Explanatory plots

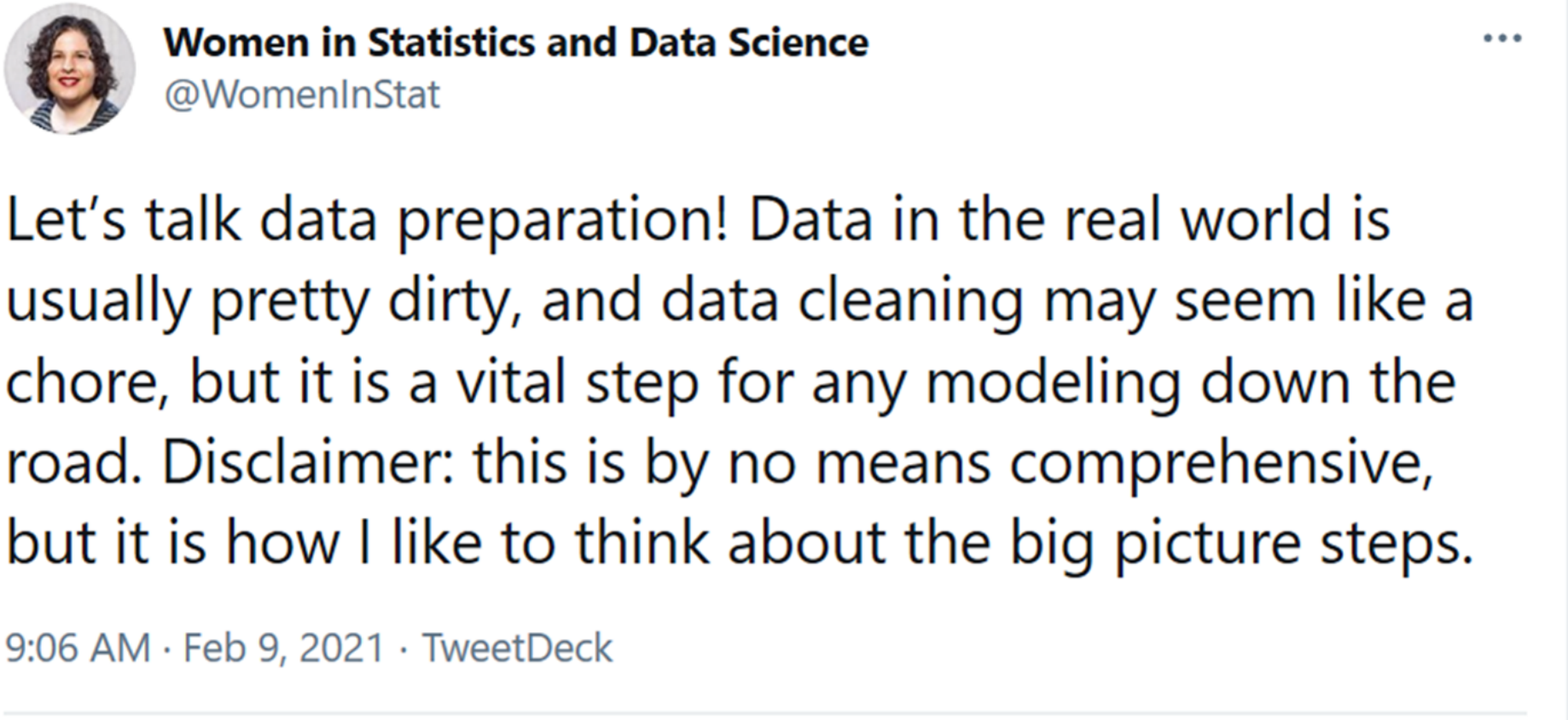
Where we are

1. Dollar-sign (brief intro)
2. Tidy data (a.k.a. cleaning your data for analysis)
    a. pivot_wider, pivot_longer
    b. managing columns
    c. recoding
    d. missing data
3. Extracting information with dplyr
4. Data visualization
    a. Exploratory plots
    b. Explanatory plots
5. In-class activity: putting everything together

A twitter thread

A nice summary of what we have done so far and why it matters!

https://twitter.com/WomenInStat/status/1359156784624242688



Women in Statistics and Data Science @WomenInStat · Feb 9

Step 2.1: Sometimes, I will try to do this in small bite-sized pieces first, like a quick pilot study to make the data more digestible. This could be isolating maybe 5 individuals with my features of interest and getting comfortable with them.

1 7

Why data cleaning

If you have a Twitter account

It can be a great professional resource to stay up-to-date.

General R accounts
ModernDive - https://twitter.com/ModernDive
RStudio - https://twitter.com/rstudiotips

Data viz
Mina Chalabi, US Guardian, Data Editor. https://twitter.com/MonaChalabi
Washington Post Graphics - https://twitter.com/PostGraphics
Wall Street Journal graphics - https://twitter.com/WSJGraphics

If you have a Twitter account

Communities

These communities share a lots of resourcs + promote work of great people in the data science and AI fields.

RLadies - https://twitter.com/RLadiesGlobal
RLadies - rotating curator - https://twitter.com/WeAreRLadies
Women in Stat - https://twitter.com/WomenInStat
Black in Data - https://twitter.com/BlkInData
Black in Ai - https://twitter.com/black_in_ai

Underrepresented community in data science
https://twitter.com/DataUmbrella
https://twitter.com/AIinclusive
https://twitter.com/DiversifyTechCo [scholarship opportunities]
https://twitter.com/codewithveni

Data viz

Data visualization

"At its core, the term ‘data visualization’ refers to any visual display of data that helps us understand the underlying data better. This can be a plot or figure of some sort or a table that summarizes the data." (Tidyverse: Skills for Data Science in R - Wright et al., 2021, p. 329)

Two main reasons to make graphs:

1. To explore the data [today's topic]

2. To communicate with others [next week's topic]

Exploratory plots

You probably haven't see many of those:

- Mostly for internal use - you (and your team) are the audience!
- Quick and dirty: draw quick information from your data in a more intuitive way
- You'll make plenty - there is no reason not to make one!
- Several uses: better understand your data, find patterns in your data, suggest modeling strategies, debug your analysis
- Not very fancy but CORRECT

Explanatory graphs

These are the fancy graphs that you see all around

- The "reader" is the audience
- Pretty, well designed graphs
- You'll make few of them, AT THE END OF YOUR ANALYSIS
- You will carefully select which graphs you need to make
- It can take quite some time to get a graph "right"

Exploratory graphs

Basic rules for an explanatory graphs

Three rules to keep in mind even when making exploratory graphs:

1. Data are displayed appropriately (i.e., the graph type matches the data type)
2. Clearly-labeled graph (e.g., axes' labels and title)
3. Axes that are not misleading

We will discuss each of these points today.

Choosing the right graph

Source

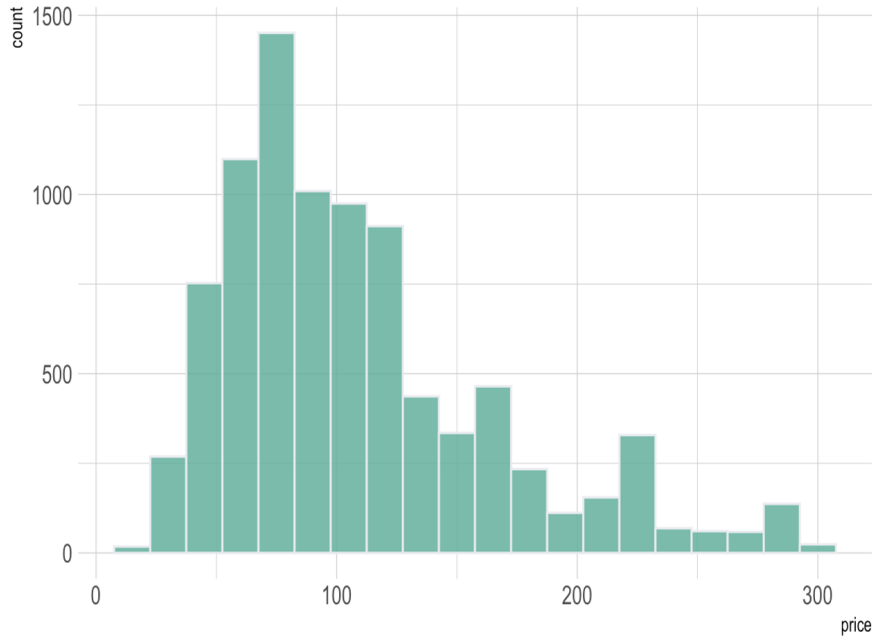
All graph pictures in this section come from this great website called **Data to Viz**, which you should add your bookmarks! There are A LOT websites on data visualization, and this is one of the most interesting to me.

- Decision tree to pick the right graph
- List of common mistakes
- R codes

There are other good websites such as depict data studio pr the R Graph Gallery.

Histograms

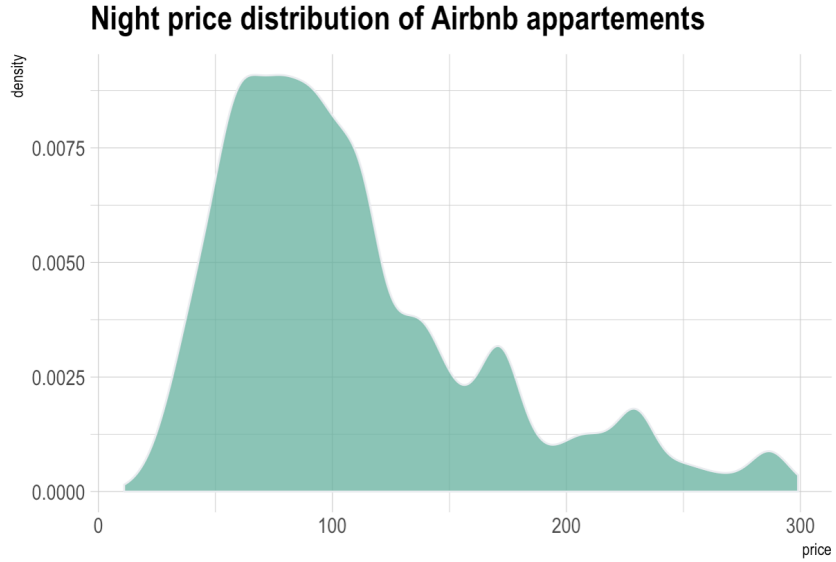
- **Variable type**: Single quantitative variables
- **Scope:** Explore the range of values that you have in your dataset - minimum, maximum, the range, where most observations are located and so on...



Histogram

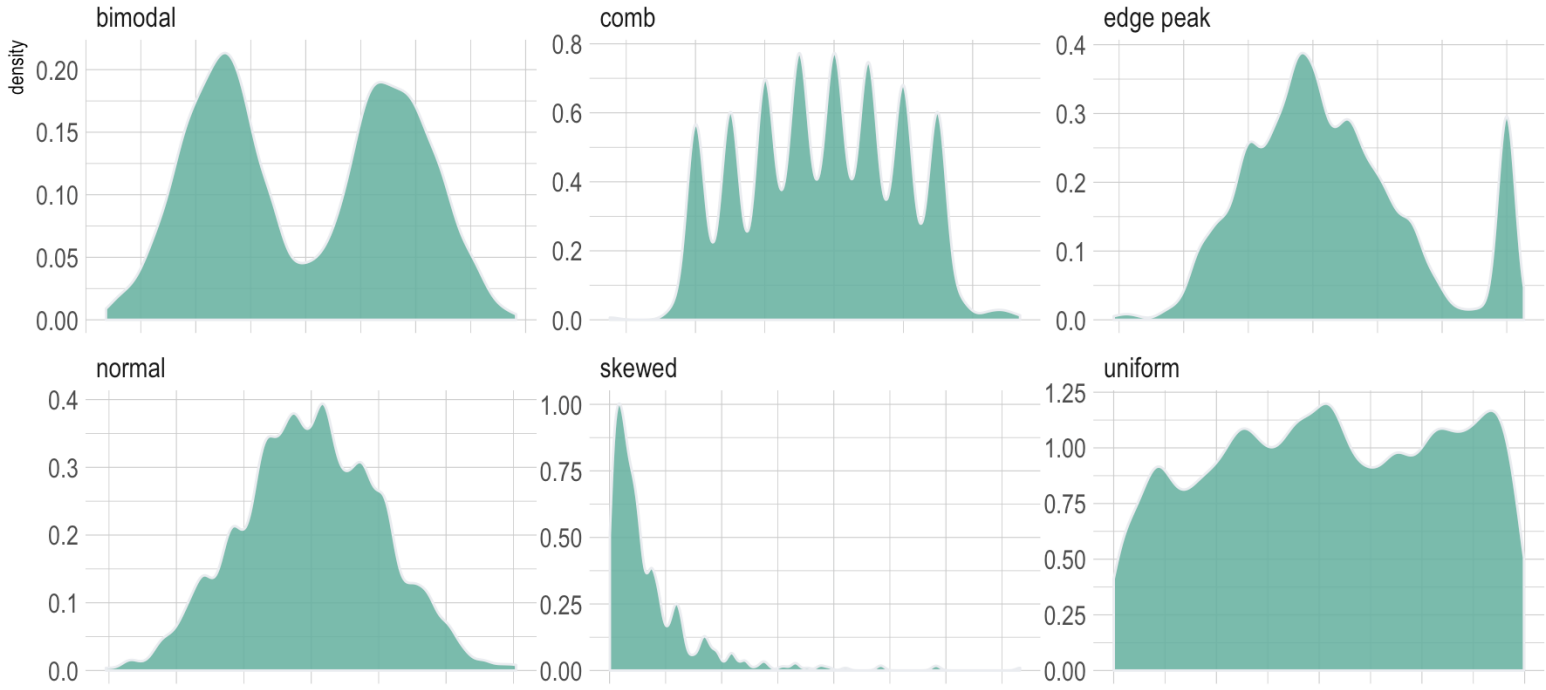
Density plots

- **Variable type:** Single quantitative variable
- **Scope:** better visualization of the distribution of a variable - less dependent on how you set your bars. For instance, it is generally used to show if your variable has a normal or skewed distribution.



Density plot

Density plot

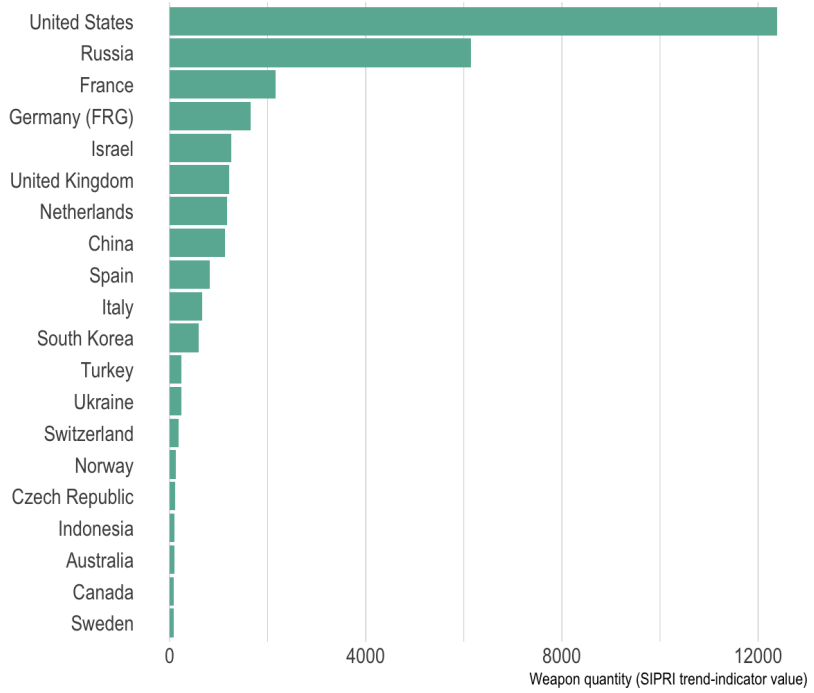


Density plot - variable distributions

If you want to know more about the Kernel density function used in density plot, you can find a nice blog post here.

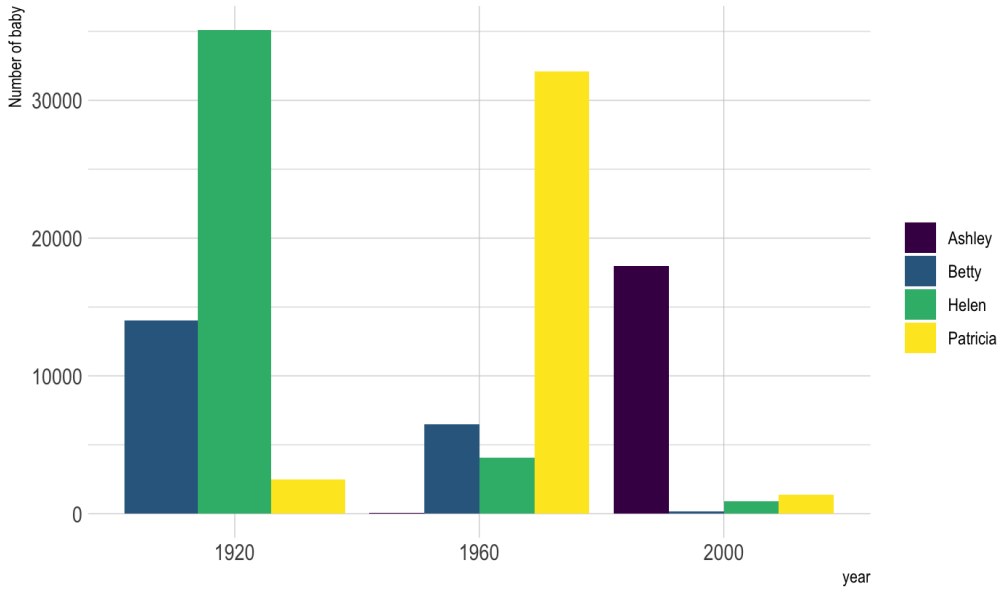
Barplot

- **Variable type:** Categorical variable that you want to break down by category
- **Scope:** Compare frequencies across categories



Grouped barplot

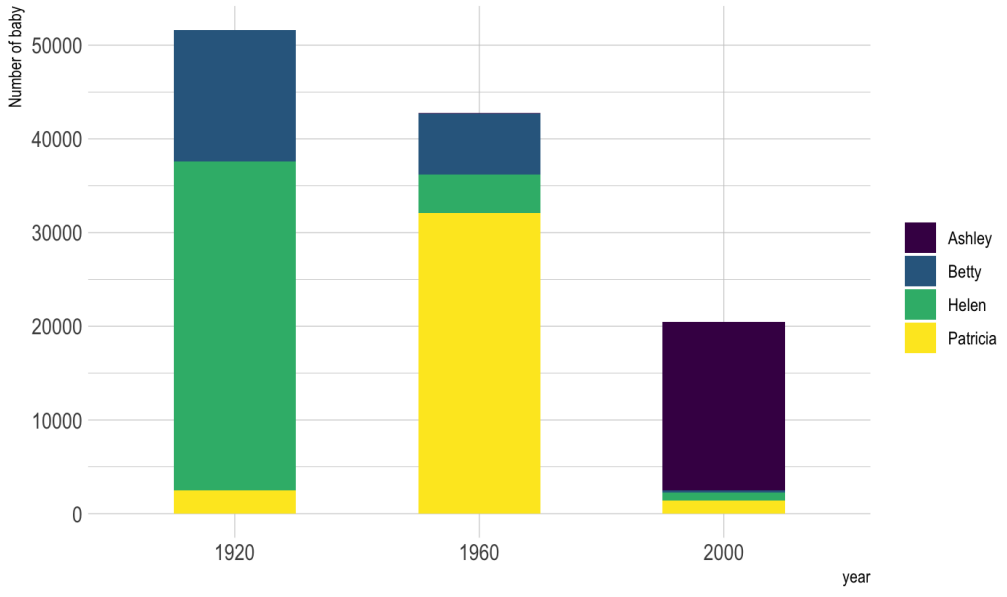
- **Variable type:** Multiple categorical variables
- **Scope:** compare across categories



Grouped barplot

Stacked barplot

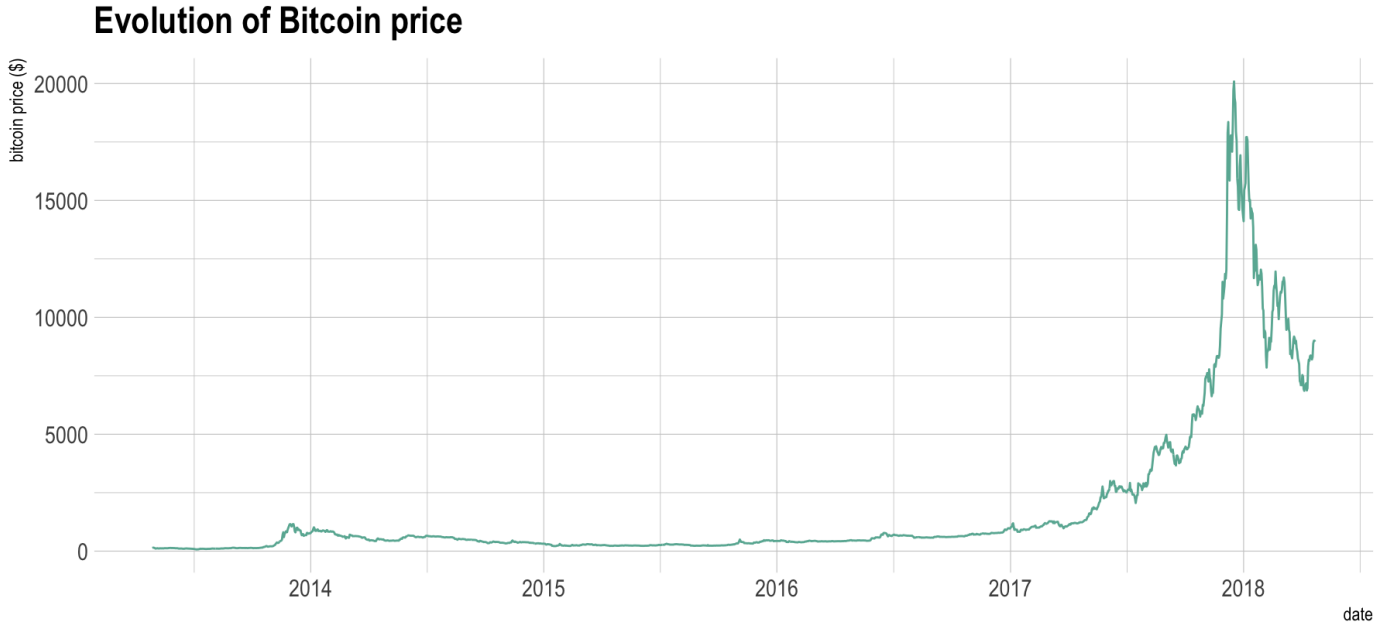
- **Variable type:** Categorical variables
- **Scope:** show the proportion of values across categorical variables



Stacked barplot

Line plots

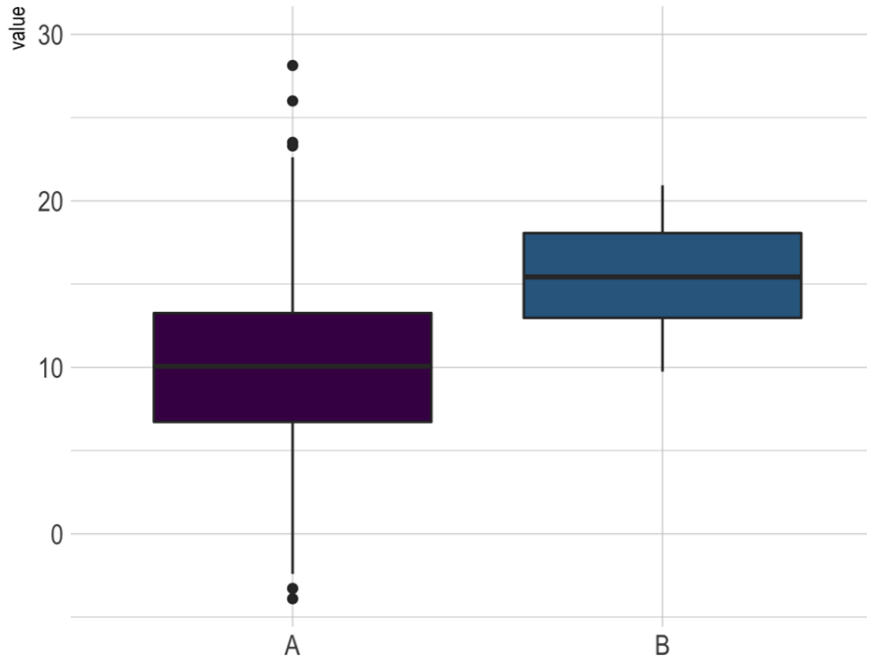
- **Variable type:** quantitative variables
- **Scope:** show the trend(s) over time

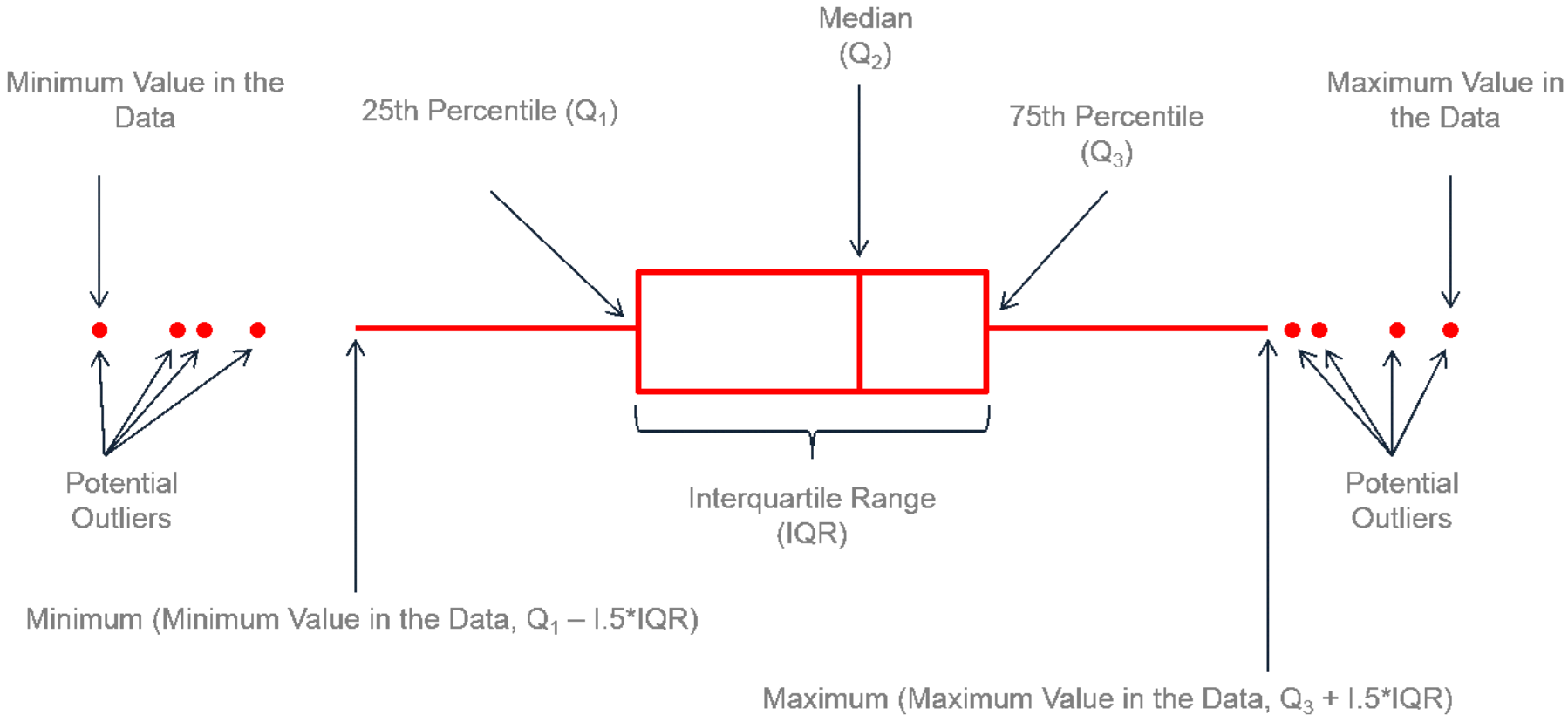


Line plot

Boxplot

- **Variable type:** quantitative variables
- **Scope:** display the summary descriptive statistics of the variable (it can be broken down by categorical variables). This is extremely valuable as exploratory graph.




Median
(Q_2)
Minimum Value in the Data
25th Percentile (Q_1)
75th Percentile
(Q_3)
Maximum Value in the Data
Potential Outliers
Interquartile Range (IQR)
Potential Outliers
Minimum (Minimum Value in the Data, Q_1 – 1.5*IQR)
Maximum (Maximum Value in the Data, Q_3 + 1.5*IQR)

Boxplot

Outliers & boxplots

Outliers are *unusual large or small values within your dataset.* They could be:

- real extreme values
- errors in the dataset (mistaken values)

Outliers & boxplots

Outliers are *unusual large or small values within your dataset.* They could be:

- real extreme values
- errors in the dataset (mistaken values)There is a common formula for outliers (which shouldn't be more important than common sense!)

UPPER: Q3 + (1.5 * INTERQUANTILE RANGE)

LOWER: Q1 - (1.5 * INTERQUANTILE RANGE)

Identiy outliers

It's hard to fathom that there are so many elderly, active gang members in Chicago who need to be tracked by police. But those aren't the only curious entries in the database.

As of this March, it also included 13 people who are supposedly 118 years old — and two others listed as 132.
Propublica

Identify outliers

- If natural variation, keep it in your dataset. You can make a note if it has a great effect on your analysis.
  - You can consider showing your analysis with and without the outlier(s)
  - There are some statistic methods that are "robust" to outliers
- If it's a mistake, fix it (if you can) or remove it

Whatever you chose, **always** document your choice.

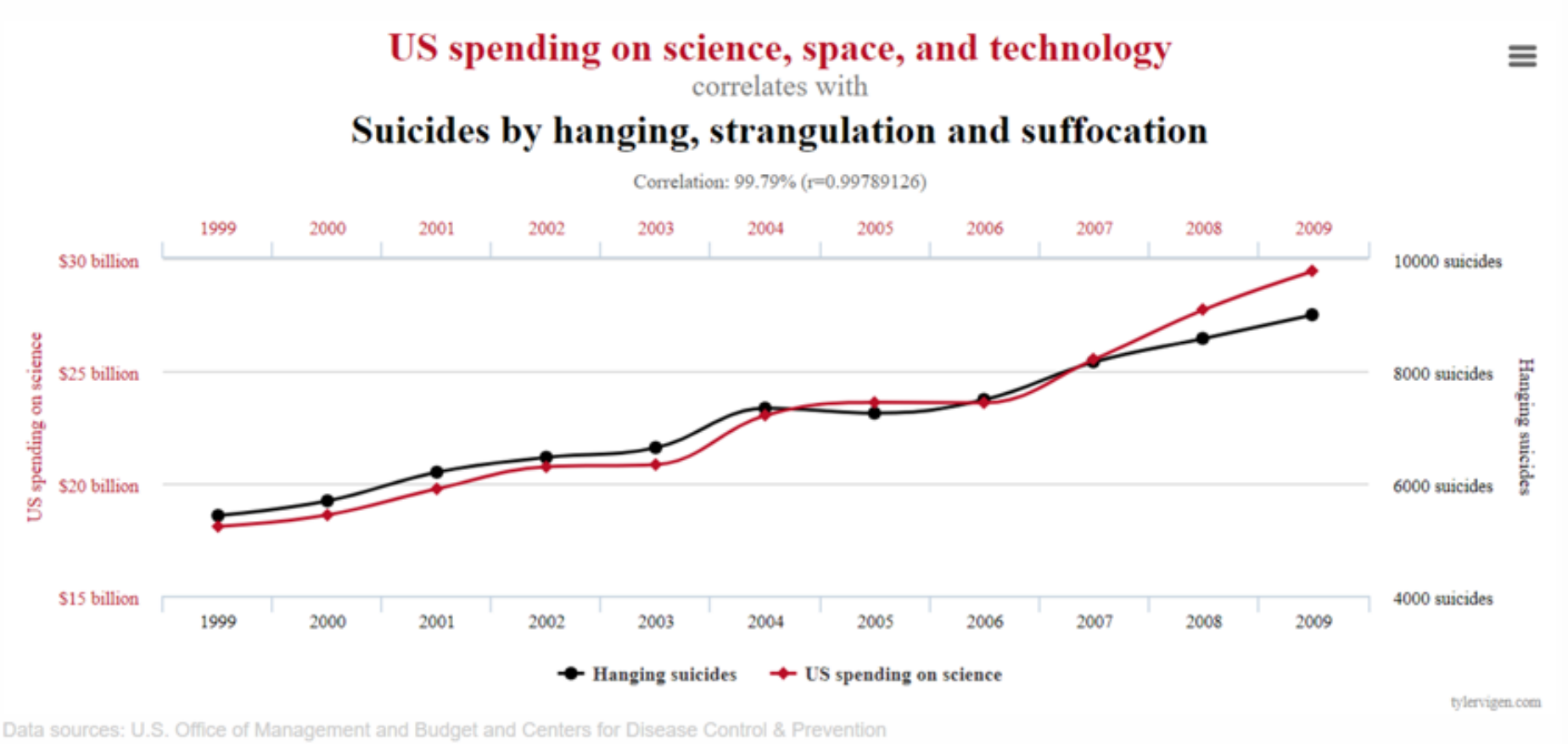
Scatterplot

- **Variable type:** continuous variable
- **Scope:** Display and investigate the relationship between two variables. It is extremely helpful as exploratory graph. *What does each point represent?*

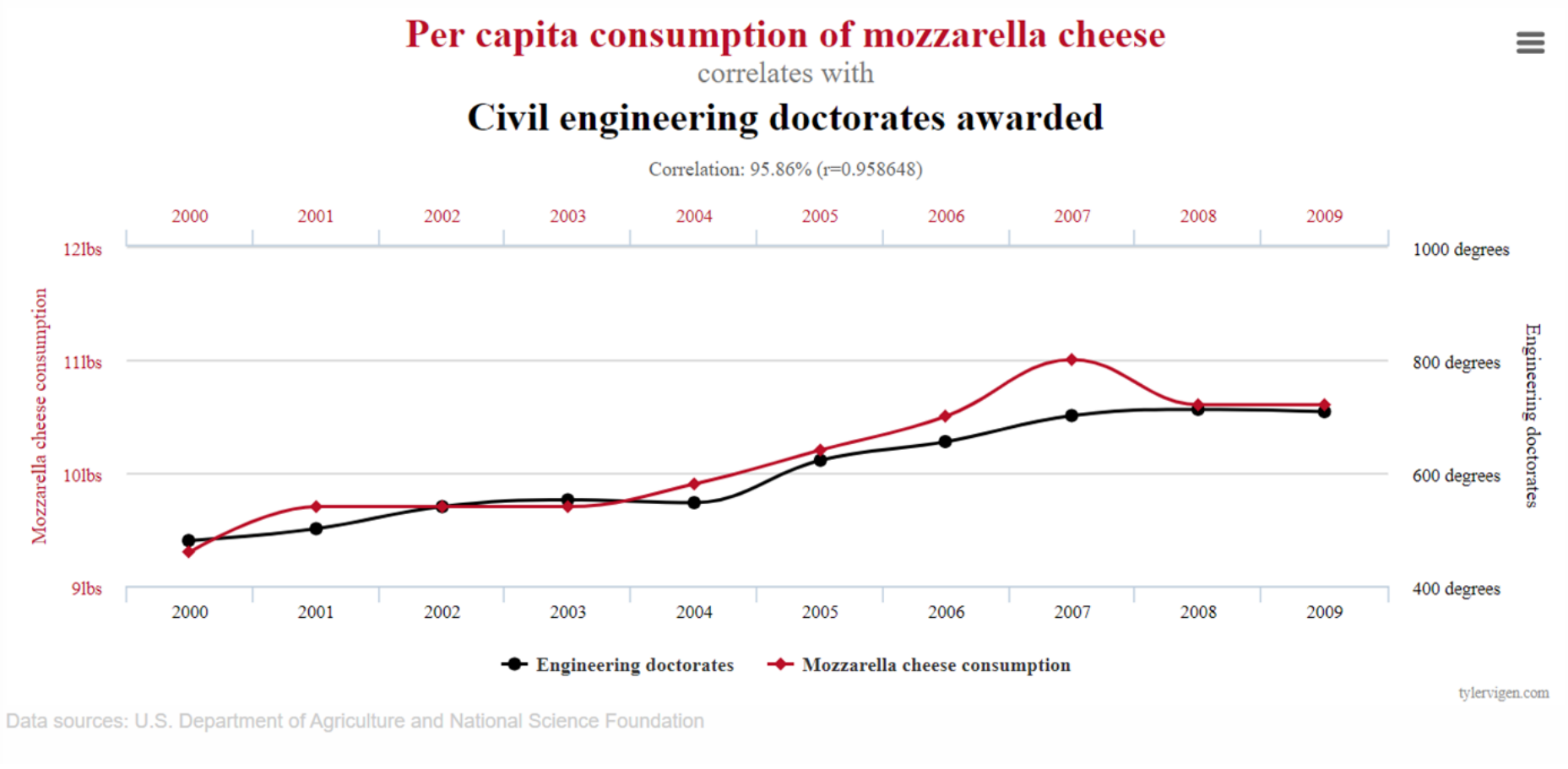
Correlations

Definition: the extent to which two variables change together at a constant change (e.g., one variable increases and the other one increases or one variable increases and the other decreases and vice versa)

It is helpful to describe relationship but it doesn't imply a cause-effect relationship. It simply means that two variables are correlated.



Spurious correlations



Spurious correlations

Correlation

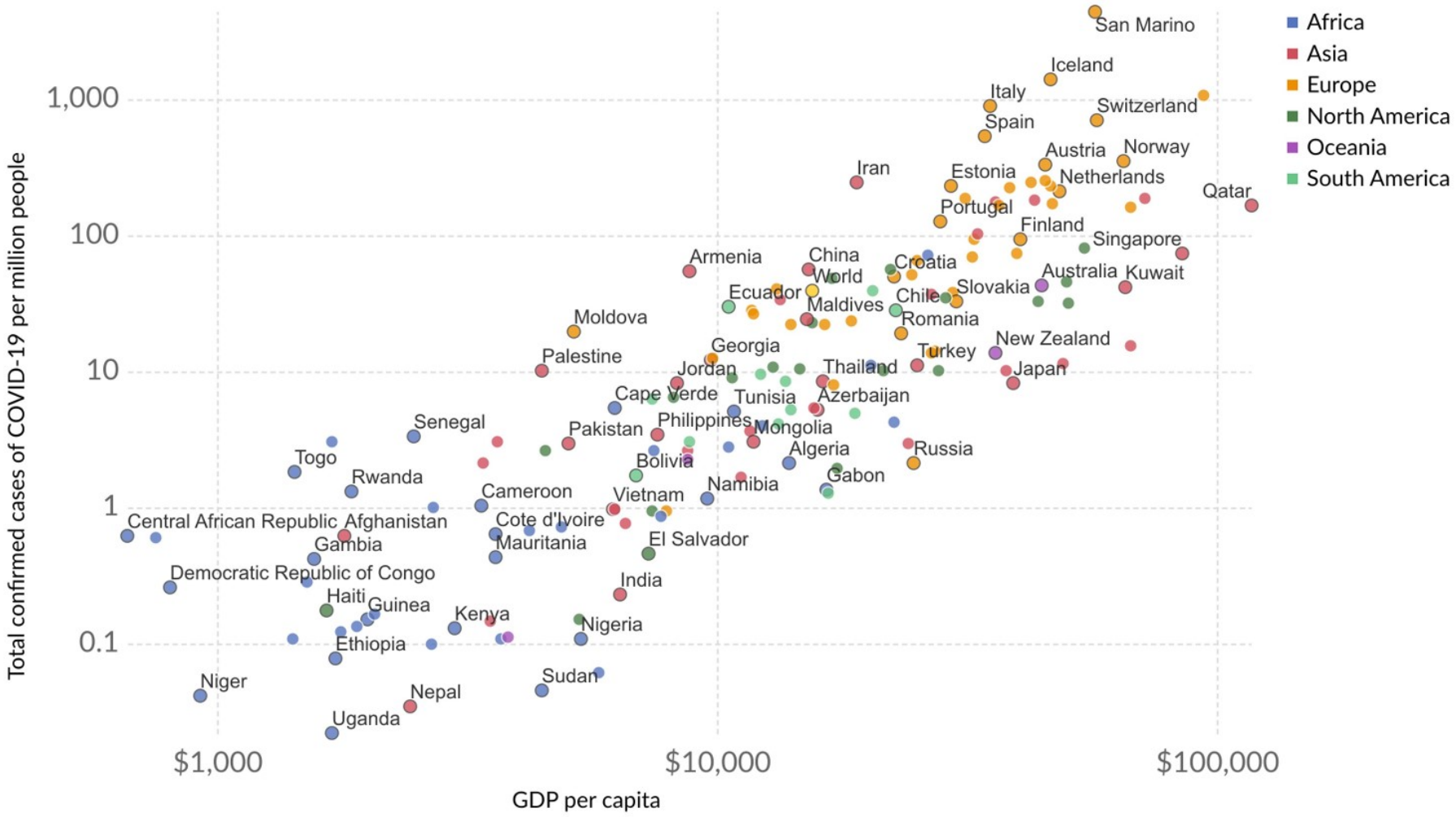
So, why should you look at correlations?

Let's look to an interesting graph (it's a year old): *What do you think about this graph? What can you learn from it?*

Total confirmed cases per million people vs. GDP per capita, Mar 22, 2020

The number of confirmed cases of COVID-19 is lower than the number of total cases. The main reason for this is limited testing.
GDP per capita is adjusted for price differences between countries (it is expressed in international dollars).

Our World in Data



Source: European CDC – Latest Situation Update Worldwide, World Bank

Correlation

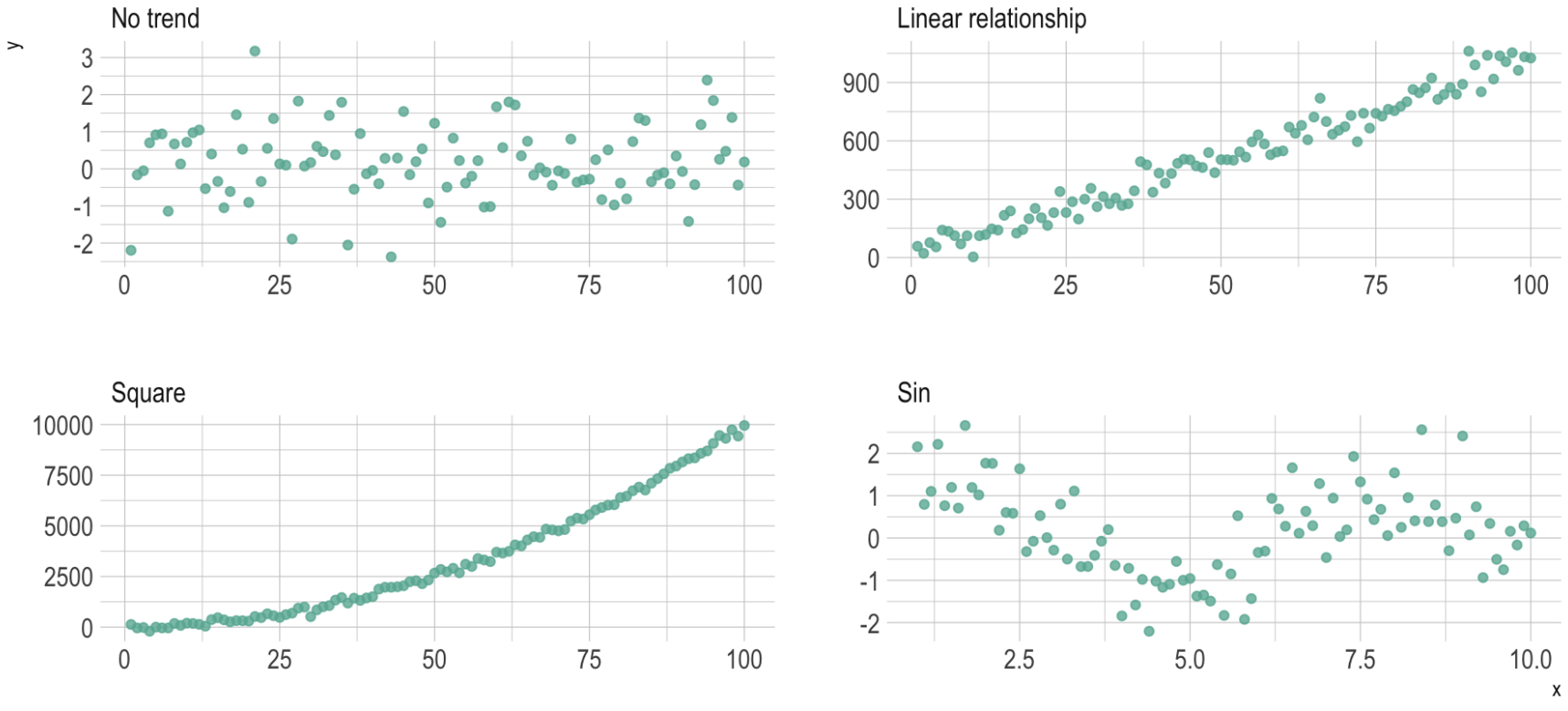
With the due cautious, a correlation can provide useful information at an initial stage of analysis - e.g., generating hypotheses or theory about social phenomenon.

The graph does not suggest causation: a higher GDP does not cause a greater number of COVID-19 cases

BUT

It does suggest a correlation that might open discussion about policies and hypotheses for research: Why do you think we observe such correlation?

- Lack of testing in poorer country
- People in richer countries are more likely to travel and therefore to get in contact with the virus OR virus might spread airports and similar crowded places
- Temperature – hypothesis that the virus diffusion is slower in warm climates
- Age: people in richer countries tend to be older



Scatterplot

A final note

- Are these all types of graphs?

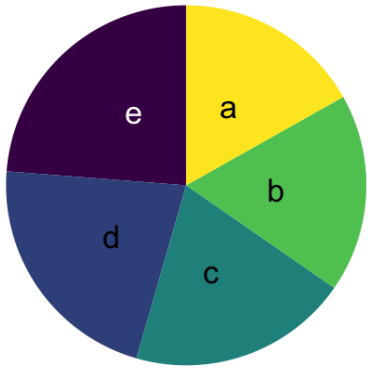
Nope, although they provide a good starting point! We will look into other types as we move forward but you are free to explore more graphs through the websites listed above (for instance, see alternatives to barplots)

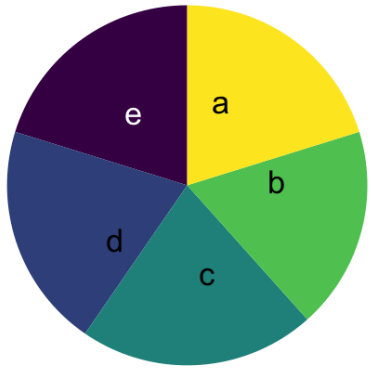
- What about pie charts?

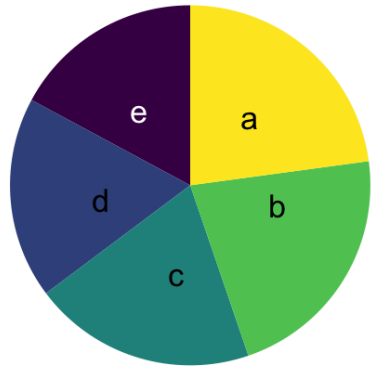
People with experience on data viz will tell you that pie charts are TO BE AVOIDED ALWAYS - It is difficult for a person to read angles and compare them. Your best alternative is to use a barplot.

piecharts vs barplots

In each graph, can you tell which is the largest group?

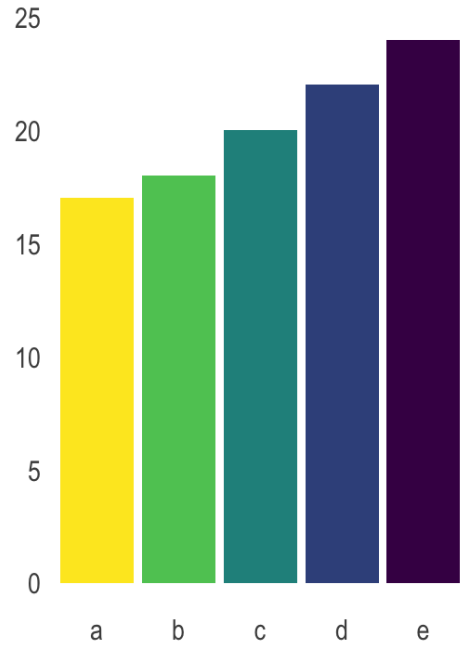


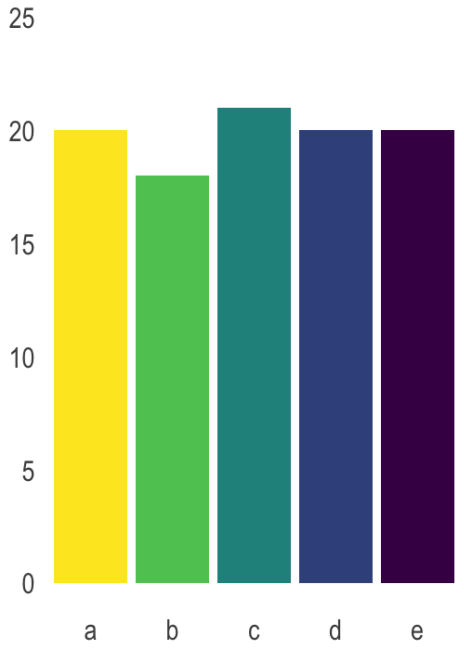


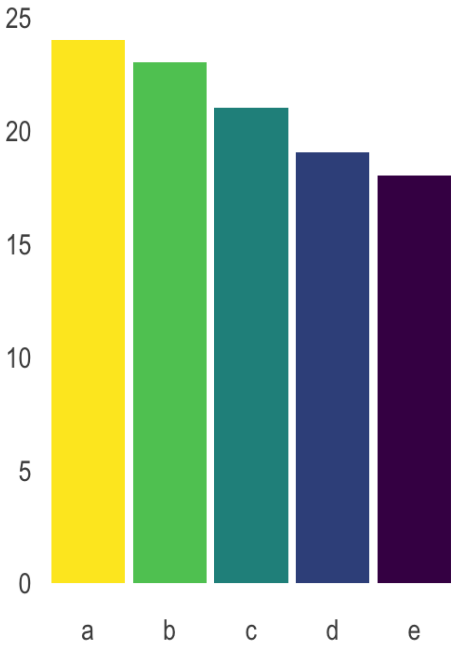


piecharts vs barplots

And now? Is it easier?







ggplot2

ggplot2

ggplot2 is the package for data visualization in R.

How it works?

- ggplot2 is built around "layers" - we are going to add layers of information on top of each others as we build our graph (e.g., information about the type of graph, information about the axes, about the title and so on...)
- Each line of the graph is separated by a + (similar to a pipe approach).

ggplot

Where to start:

Step 1 is to chose the data that you want to use

```
ggplot(data = YOUR_DATA) +  #call your data

#OR you can start with a pipe

data %>%
  ggplot()
```

Step 2 is to specify the type of graph that you want to create using **geom**

```
geom_PLOT_TYPE(mapping = aes(x = VARIABLE, y = VARIABLE))
```

geom

There are several geom functions which are pretty intuitive to remember:

- geom_point = scatterplots
- geom_histogram = histogram
- geom_bar = bar plot
- geom_boxplot = boxplot
- geom_density = density plot
- geom_dotplot = dotplot

Full list of geom_ functions can be found here

data

We are going to use the **election_turnout** dataset from the **stevedata** package.

```
#install.packages("stevedata")
library(stevedata)
```

Let's see how to do some basics graphs...

histogram

First, let's use a histogram to look at the **turnout** percentage across states

```
ggplot(data = election_turnout) +
  geom_histogram(mapping = aes(x = turnoutho))
```

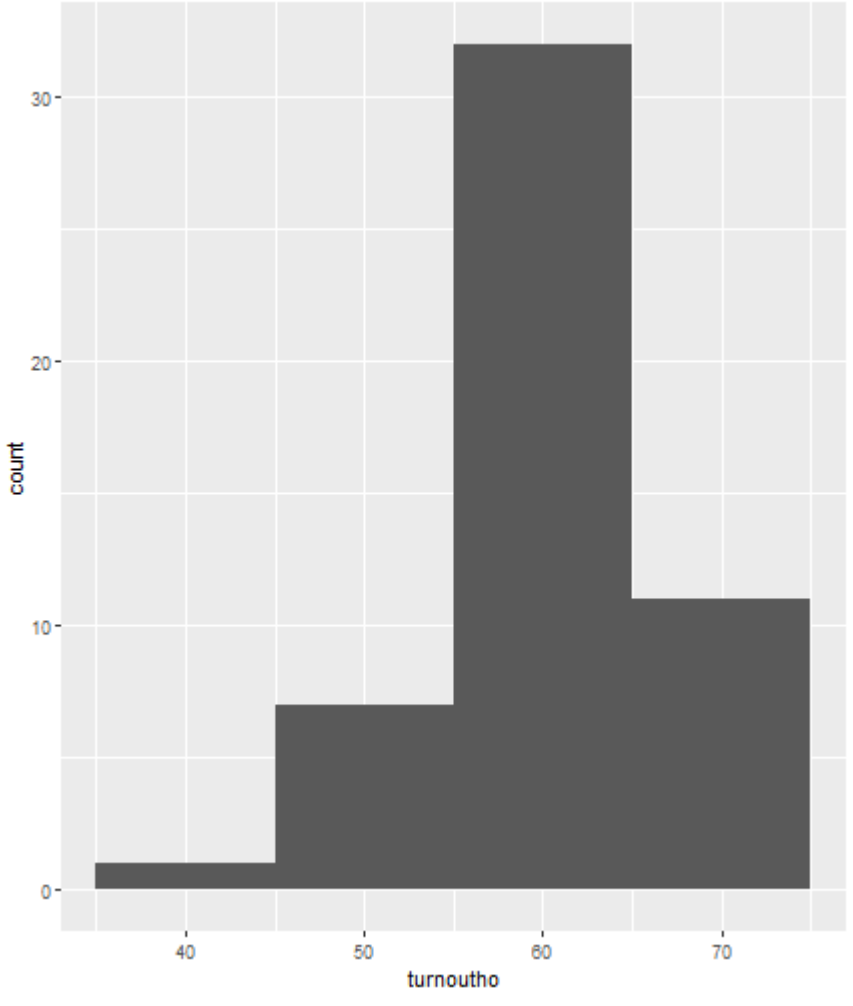
barplot bins

The width of the bins in you graph can be tricky - if you vary it, you graph representation also changes quite a bit

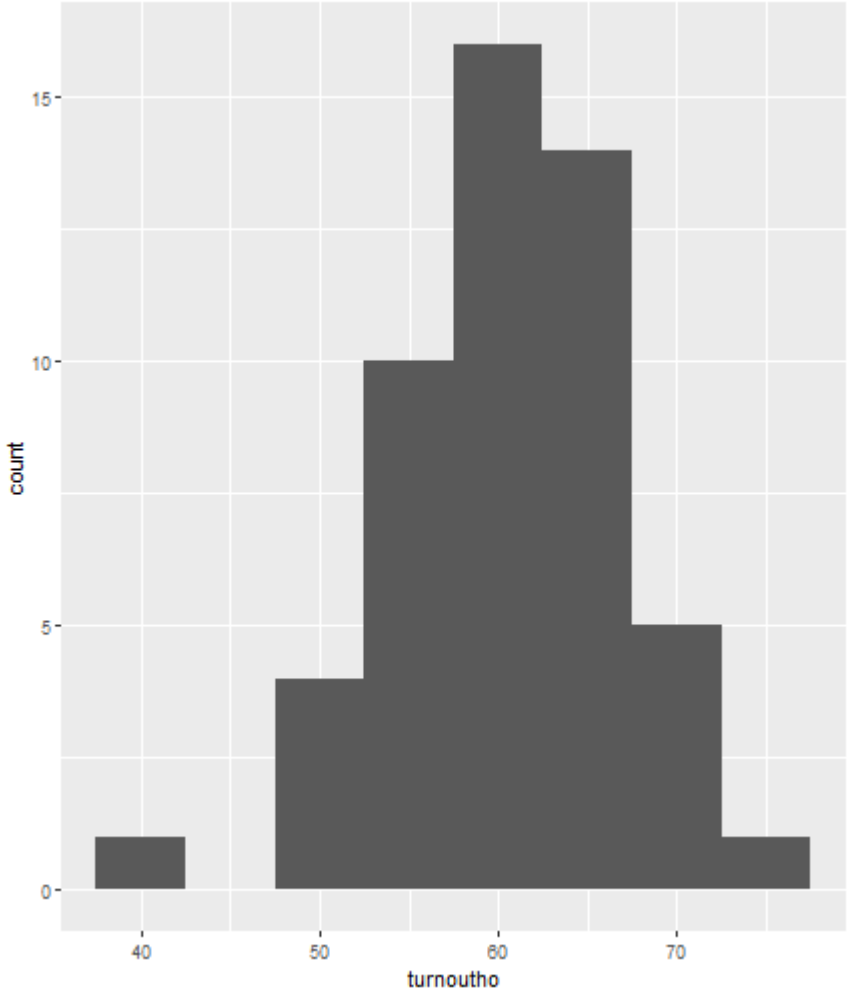
- **binwidth** defines how many observations should be included in each bin
- **bins** define how many bins you want to create.

You need to use one or the other within the aes argument.

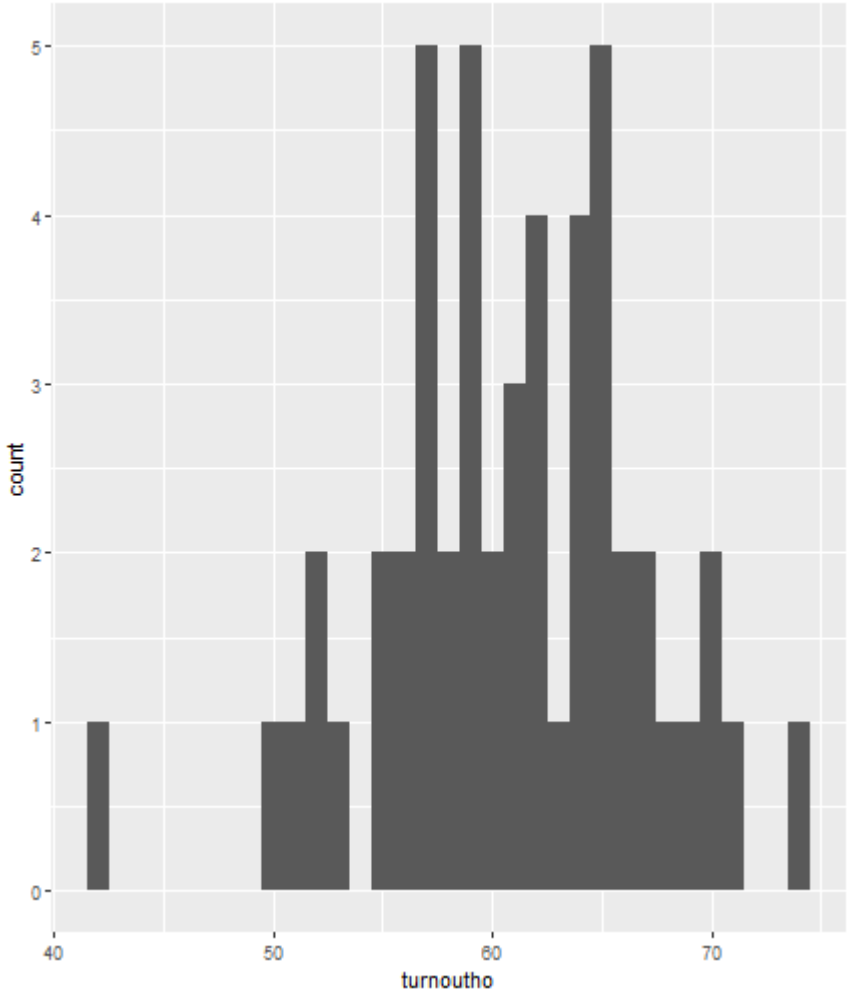
```
ggplot(data = election_turnout) +
  geom_histogram(mapping = aes(x = turnoutho),
                 binwidth = 10
                 )
```



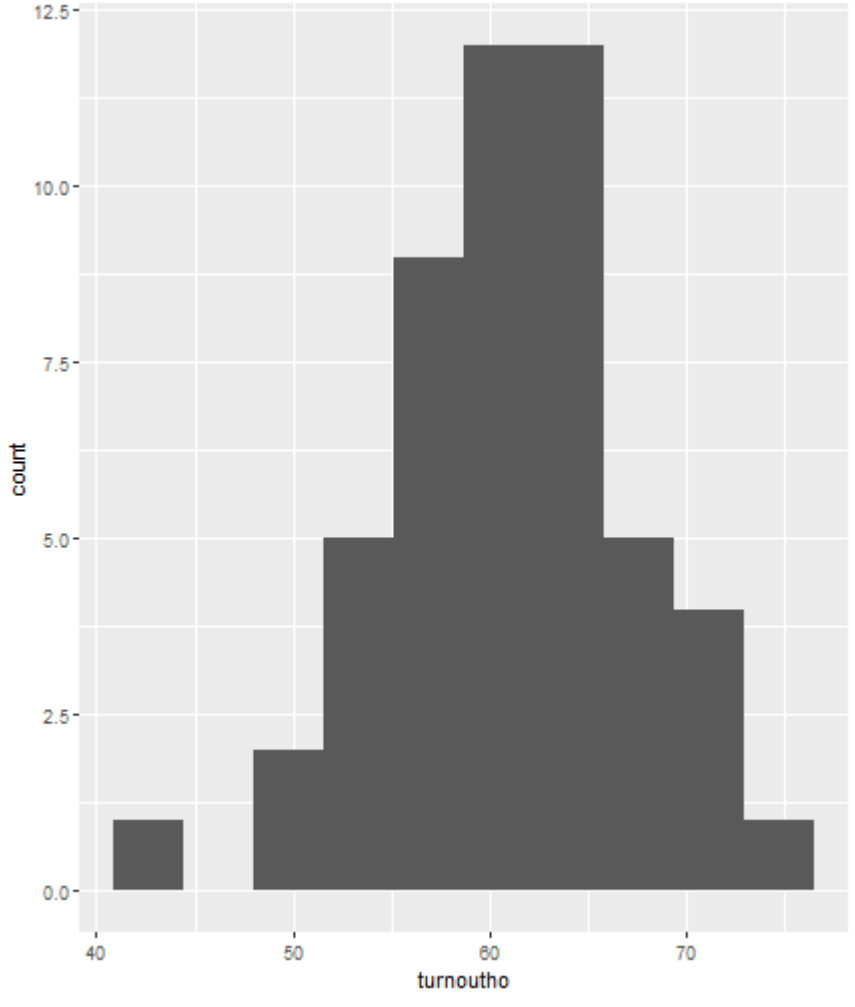
```
ggplot(data = election_turnout) +
  geom_histogram(mapping = aes(x = turnoutho),
                 binwidth = 5 #ROTATE
                 )
```



```
ggplot(data = election_turnout) +
  geom_histogram(mapping = aes(x = turnoutho),
                 binwidth = 1 #ROTATE
                 )
```



```
ggplot(data = election_turnout) +
  geom_histogram(mapping = aes(x = turnoutho),
                 bins = 10   #ROTATE
                 )
```

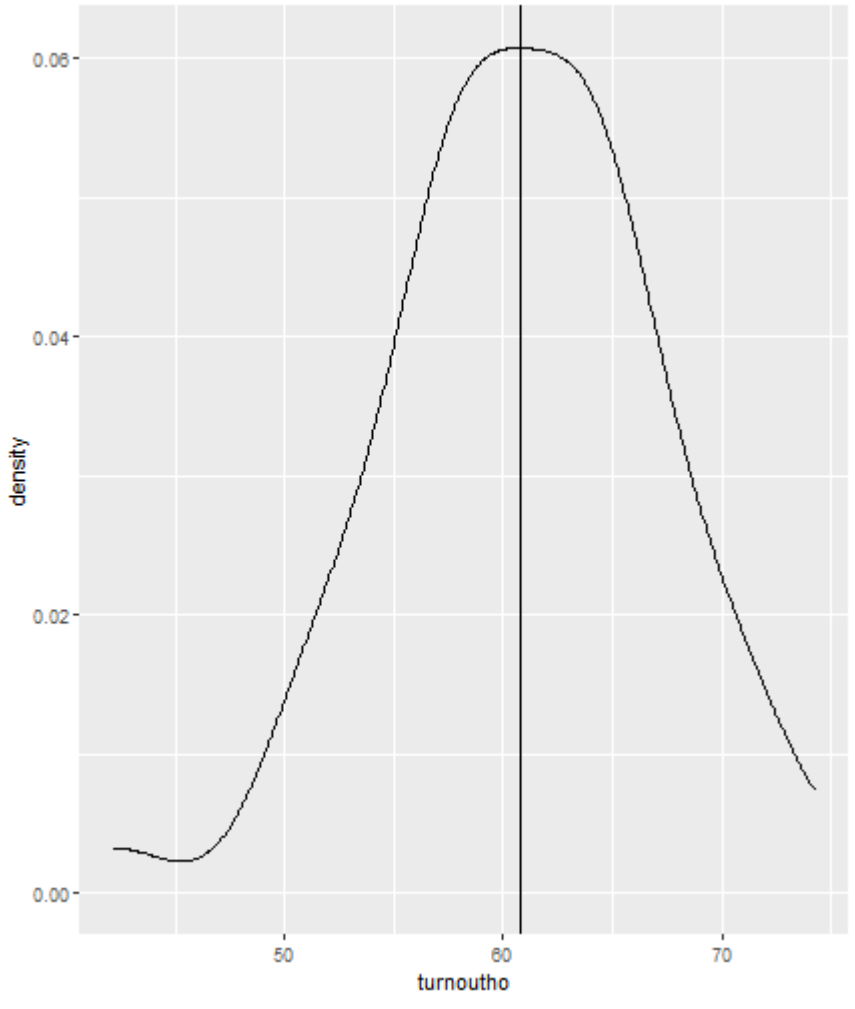


Density plot

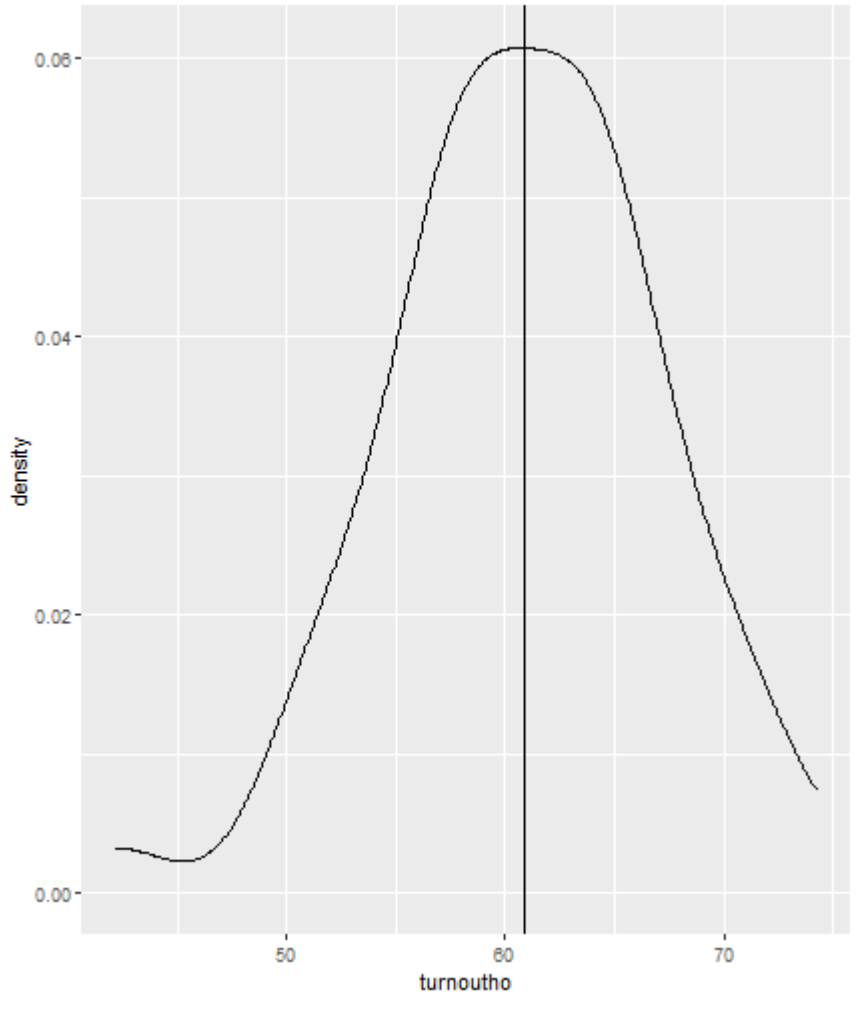
A better way to look at the distribution is by using a density plot.

```
ggplot(data = election_turnout) +
  geom_density(mapping = aes(x = turnoutho))
```

```
ggplot(data = election_turnout) +
  geom_density(mapping = aes(x = turnoutho)) +
  geom_vline(xintercept =
               mean(election_turnout$turnoutho)
             )
```



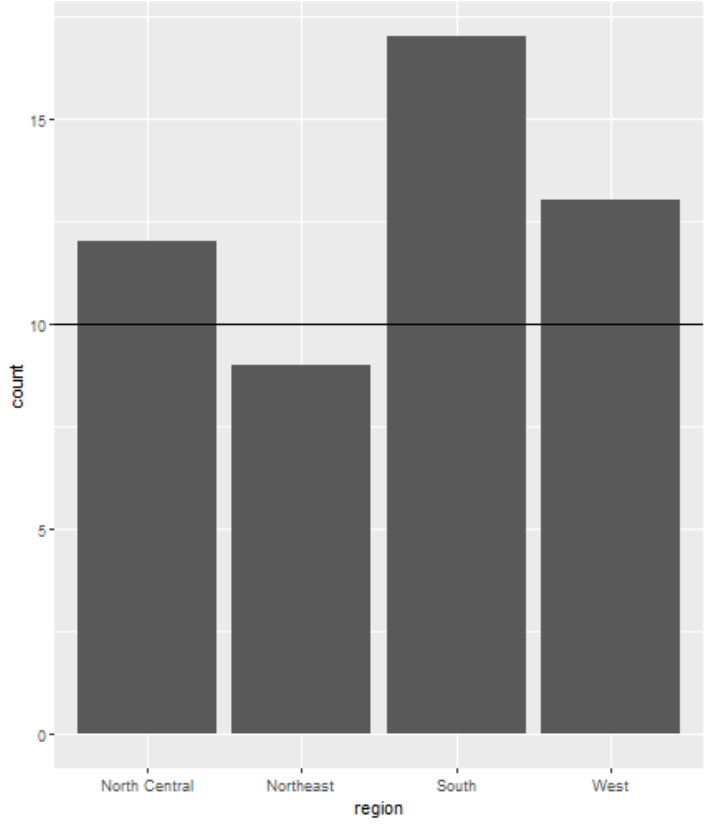
```
ggplot(data = election_turnout) +
  geom_density(mapping = aes(x = turnoutho)) +
  geom_vline(xintercept =
               median(election_turnout$turnoutho) #R
             )
```



Since mean and median are very similar, it's likely we are looking at a normal distribution

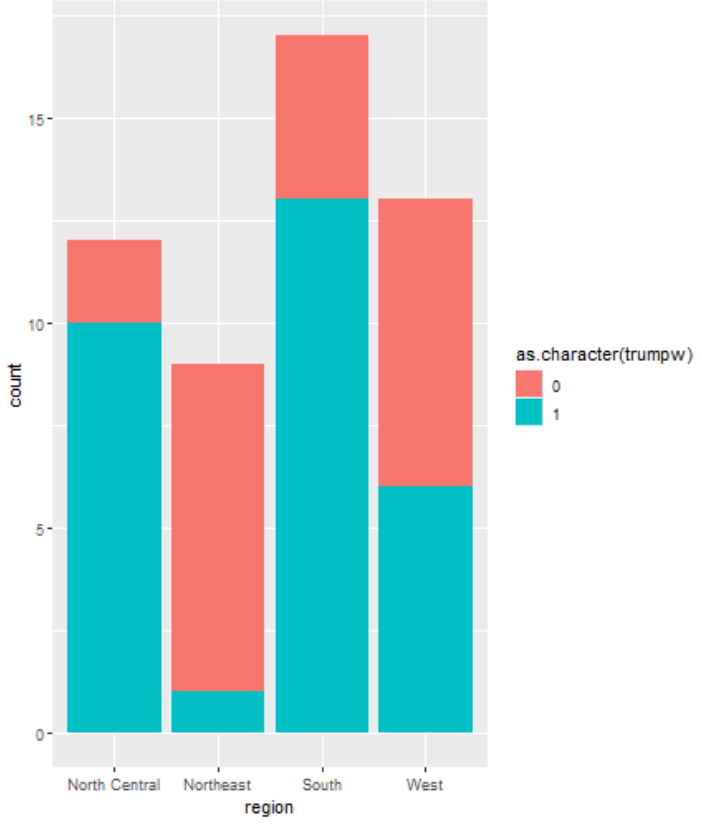
barplot

```
ggplot(data = election_turnout) +
  geom_bar(mapping = aes(x = region)) +
  geom_hline(yintercept = 10)
```



Stacked bar plot

```
ggplot(data = election_turnout) +
  geom_bar(mapping = aes(x = region, fill = as.character(trumpw)))
```

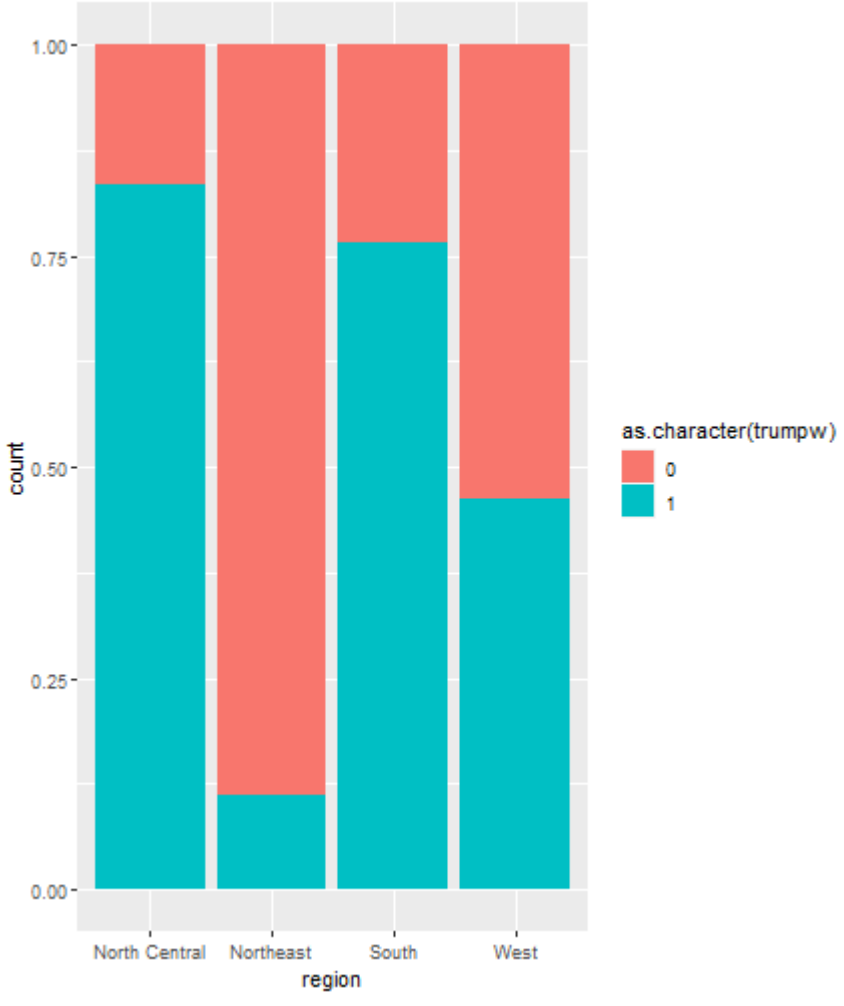


Stacked bar plot

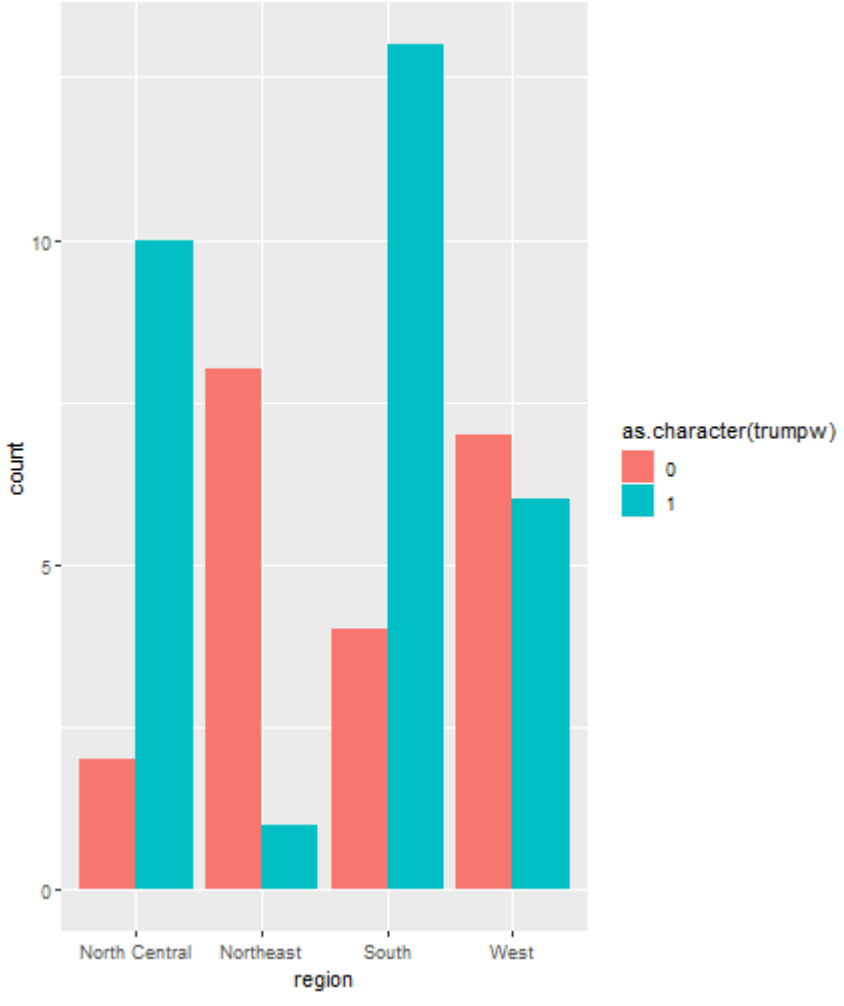
So, we use a stacked barplot to highlight a new dimension of the data (e.g., states where Trump won)

We can use the basic code of a barplot to visualize data differently...

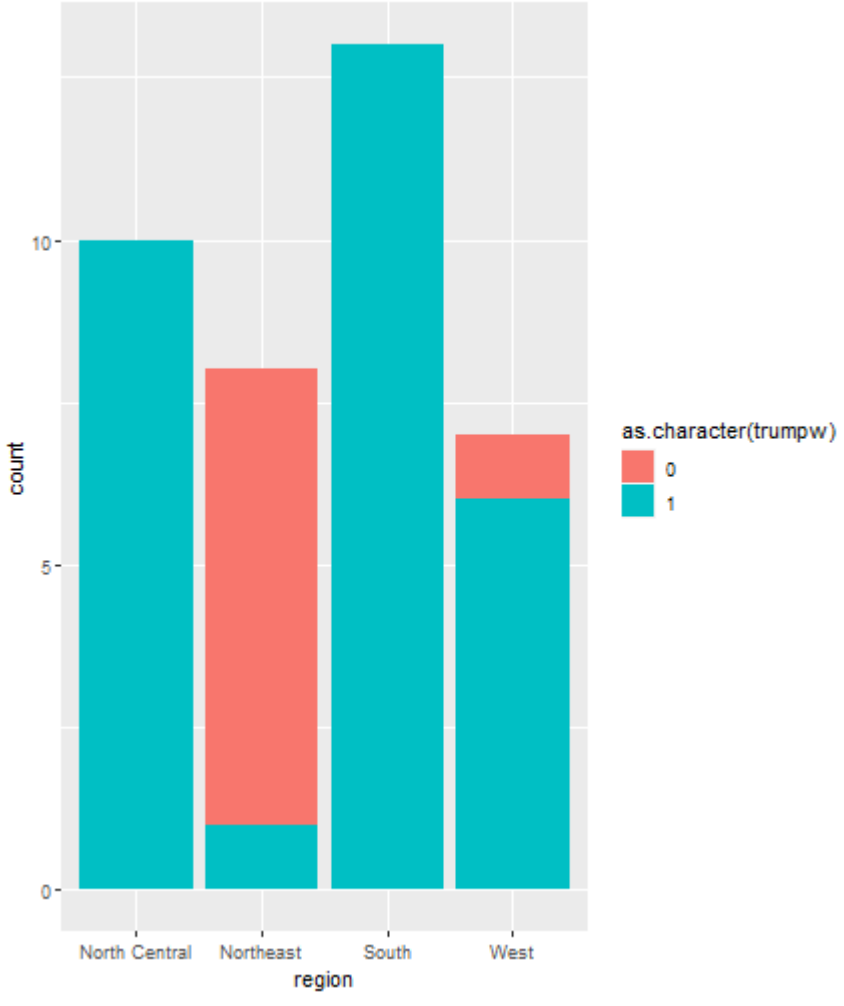
```
ggplot(data = election_turnout) +
  geom_bar(mapping = aes(x = region, fill = as.charac
           position = "fill"
           )
```



```
ggplot(data = election_turnout) +
  geom_bar(mapping = aes(x = region, fill = as.charac
           position = "dodge" #ROTATE
           )
```



```
ggplot(data = election_turnout) +
  geom_bar(mapping = aes(x = region, fill = as.charac
           position = "identity" #ROTATE
           )
```



facet

Another way to look at differences across states based off Trump's victory is by splitting the graph.

We can use *facet* to split a plot into multiple plots according to one or more categorical variables.

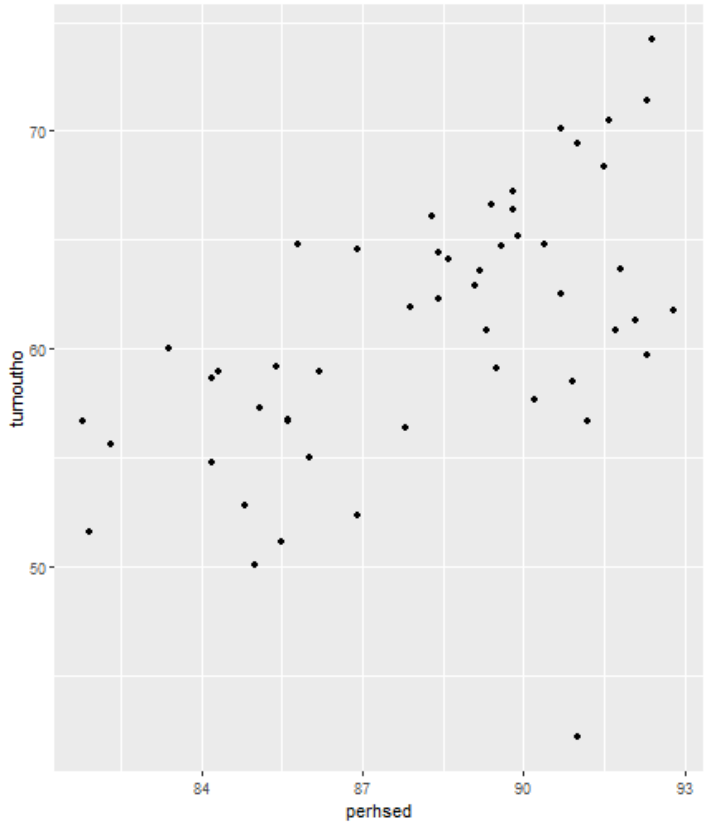
Intuition: you have a scatterplot of all your observations. You want to separate out observations for states where Trump won from

facet

```
ggplot(data = election_turnout) +
  geom_bar(mapping = aes(x = region, fill = as.character(trumpw))) +
  facet_grid(~as.character(trumpw))
```

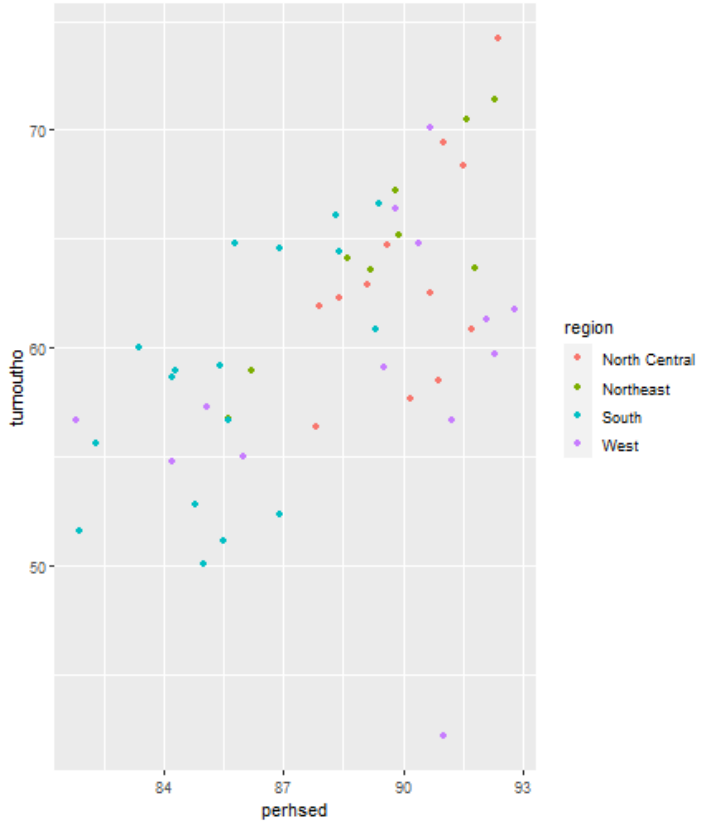
Scatterplot

```
ggplot(data = election_turnout) +
  geom_point(mapping = aes(x = perhsed, y = turnoutho))
```



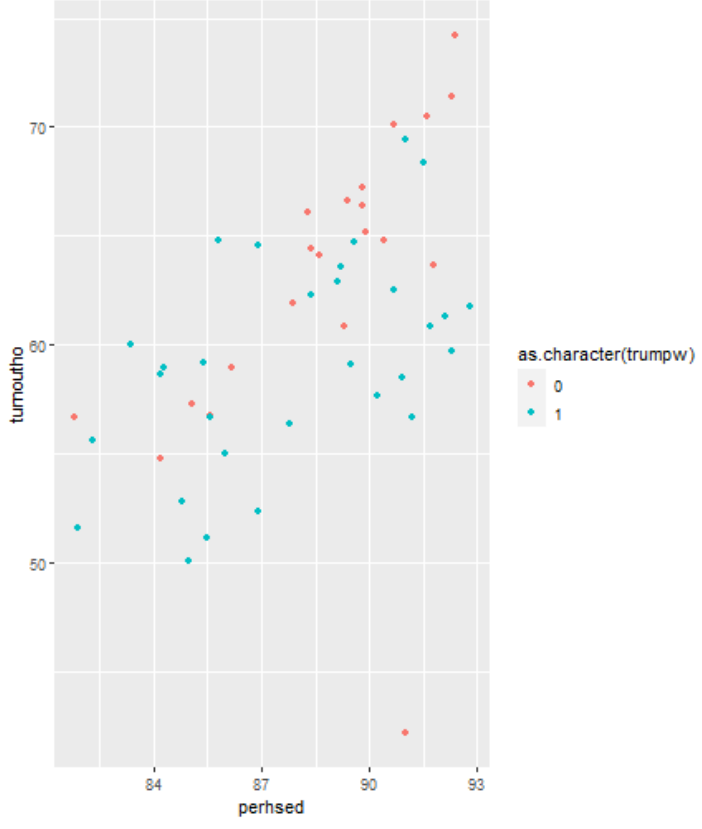
Scatterplot

```
ggplot(data = election_turnout) +
  geom_point(mapping = aes(x = perhsed, y = turnoutho, color = region))
```



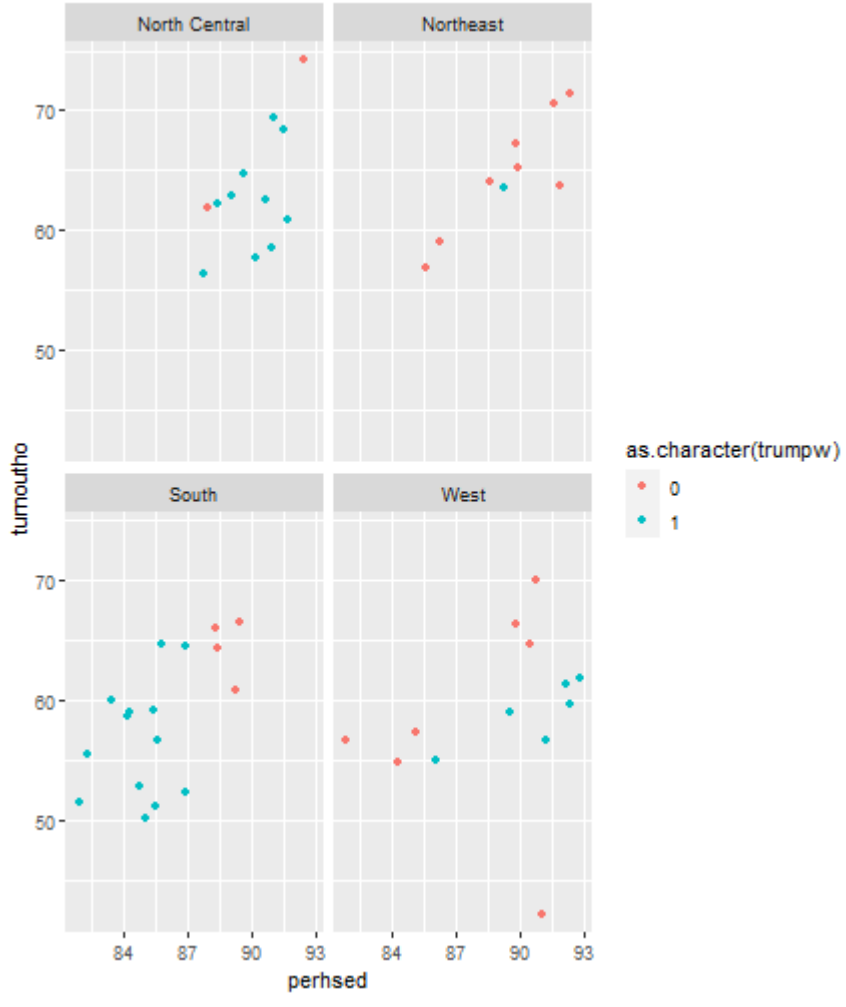
Scatterplot

```
ggplot(data = election_turnout) +
  geom_point(mapping = aes(x = perhsed, y = turnoutho, color = as.character(trumpw)))
```

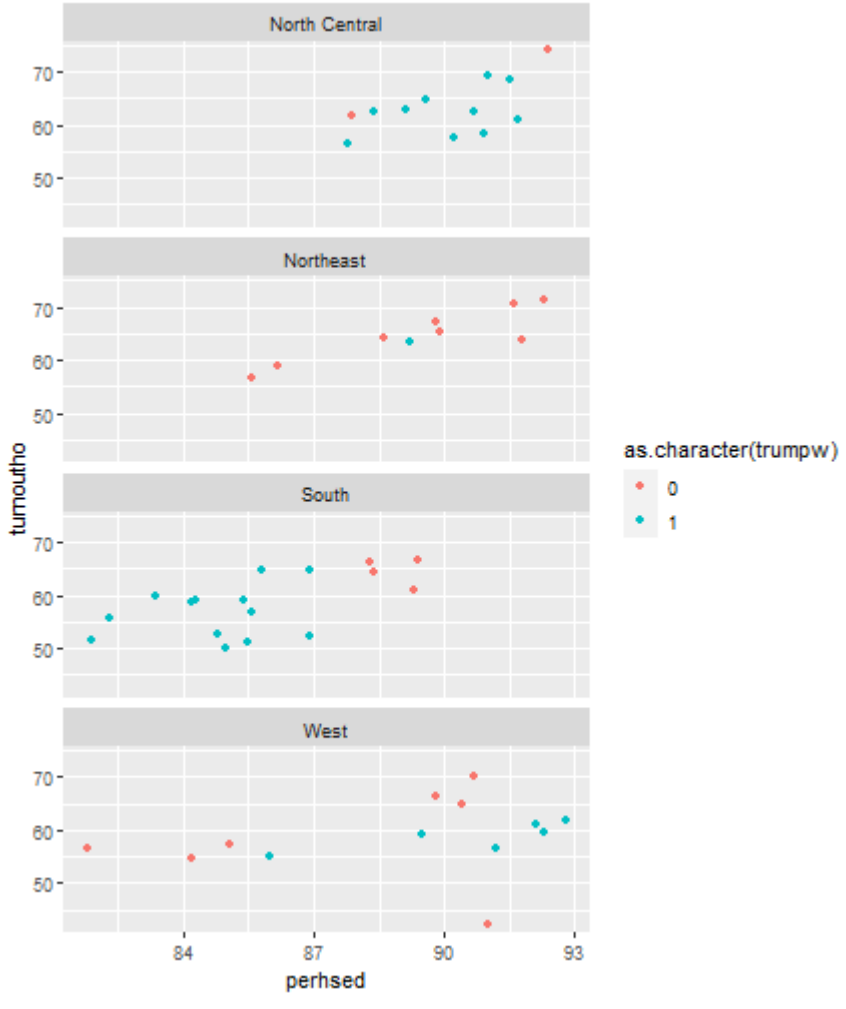


facet

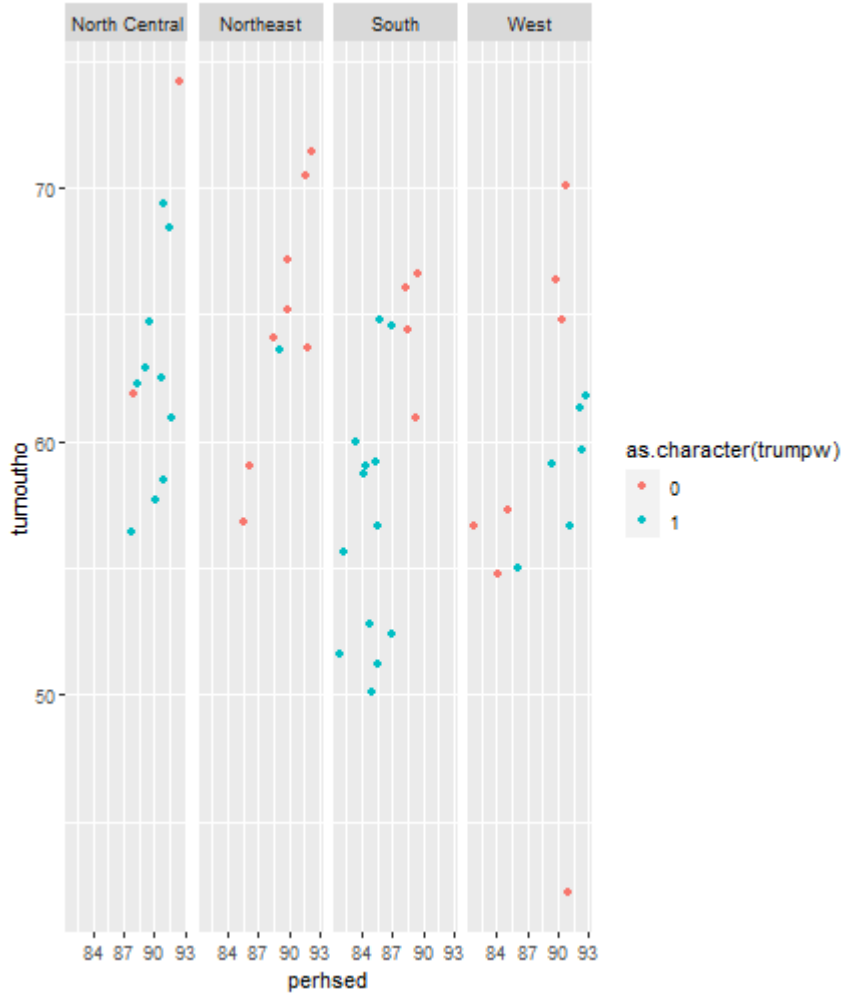
```
ggplot(data = election_turnout) +
  geom_point(mapping = aes(x = perhsed, y = turnoutho
  facet_wrap(~region)
```



```
ggplot(data = election_turnout) +
  geom_point(mapping = aes(x = perhsed, y = turnoutho
  facet_wrap(~region, ncol = 1)
```



```
ggplot(data = election_turnout) +
  geom_point(mapping = aes(x = perhsed, y = turnoutho
  facet_wrap(~region, nrow = 1)
```



Correlation

Note that you can also calculate the strenght of the correlation between the two variables, whereas:

- **-1** indicates a strong negative correlation (one variable increses, the other decreases)
- **0** indicates no correlation (the line is horizontal)
- **+1** indicates a strong positive correlation (one variable increases, the other increases)

```
cor(election_turnout$turnoutho, election_turnout$perhsed)
```

```
[1] 0.5282739
```

Correlation line

You can represent the correlation line on the graph by calculating the line parameter with a simple linear regression.

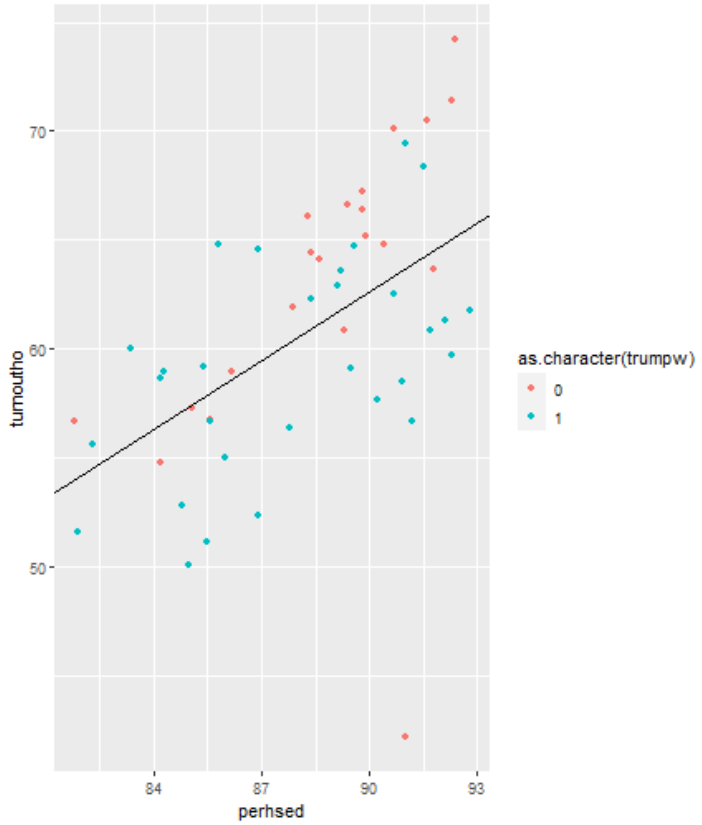
Step 1: calculate the parameters

```
reg = lm(turnoutho ~ perhsed, data = election_turnout)
```

Step 2: add them to your graph

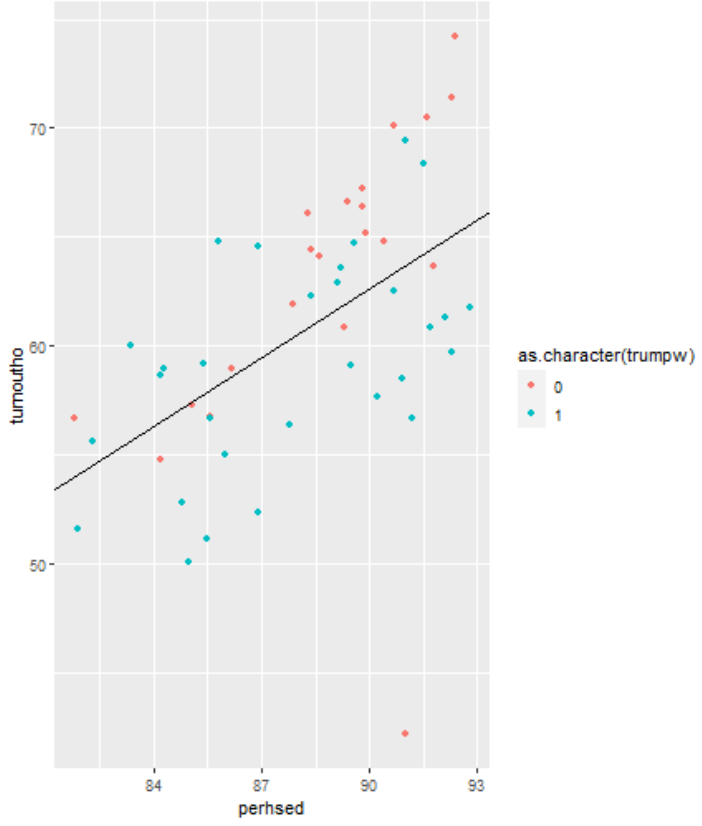
Regression line

```
ggplot(data = election_turnout) +
  geom_point(mapping = aes(x = perhsed, y = turnoutho, color = as.character(trumpw))) +
  geom_abline(intercept = -32.303, slope = 1.055)
```



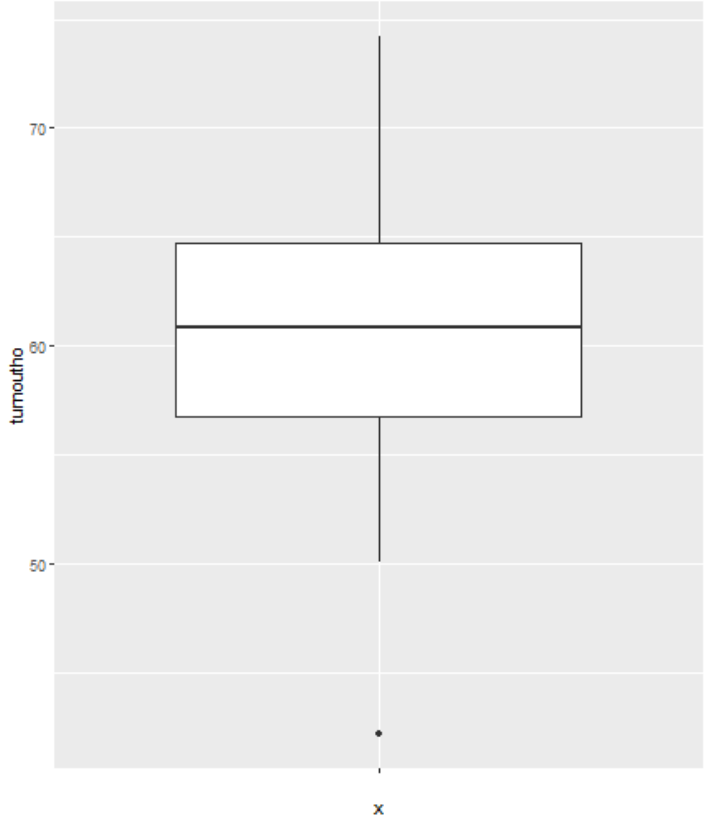
Regression line

```
ggplot(data = election_turnout) +
  geom_point(mapping = aes(x = perhsed, y = turnoutho, color = as.character(trumpw))) +
  geom_abline(intercept = reg$coefficients[1], slope = reg$coefficients[2])
```



Boxplot

```
election_turnout %>%
ggplot() +
  geom_boxplot(aes(x = "", y = turnoutho))
```



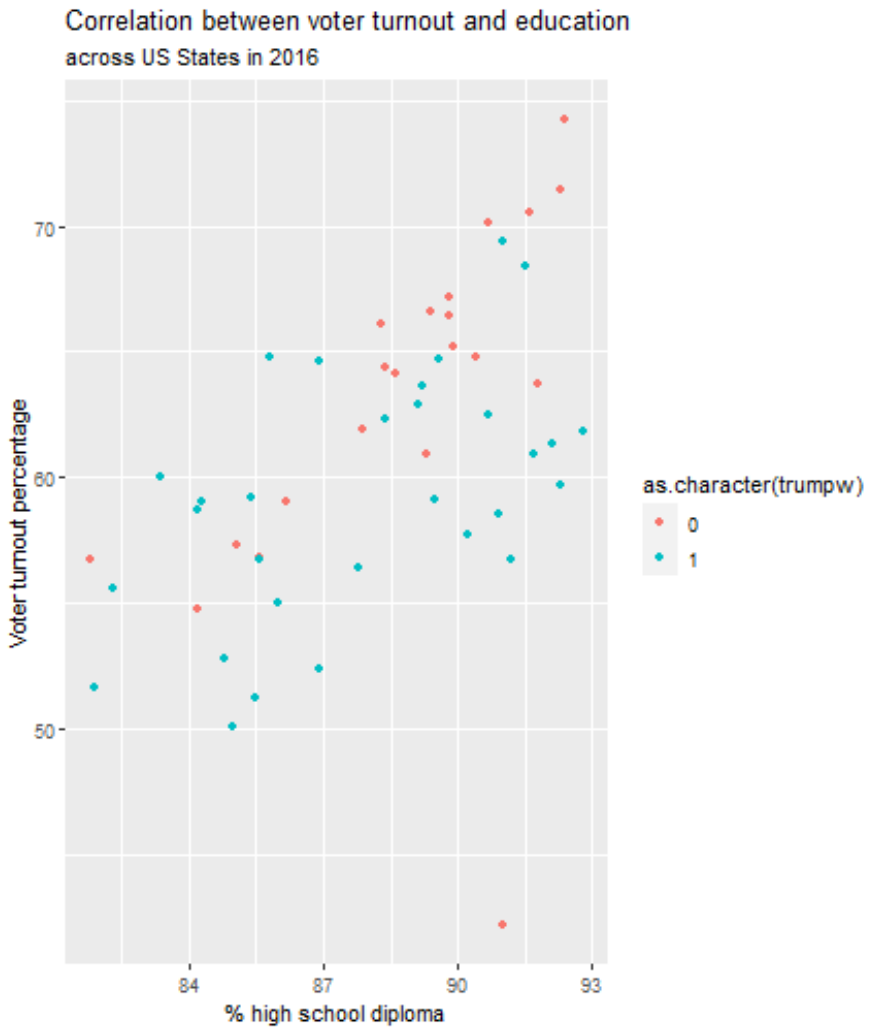
Boxplot

What would you do with that one outlier?

Label your graph

Labels and titles

```
scatterplot_label
```



Change the axis

You x and y axis need to make sense given your data (check out the summary statistics to define them). It is possible that R plot is not producing your axes correctly

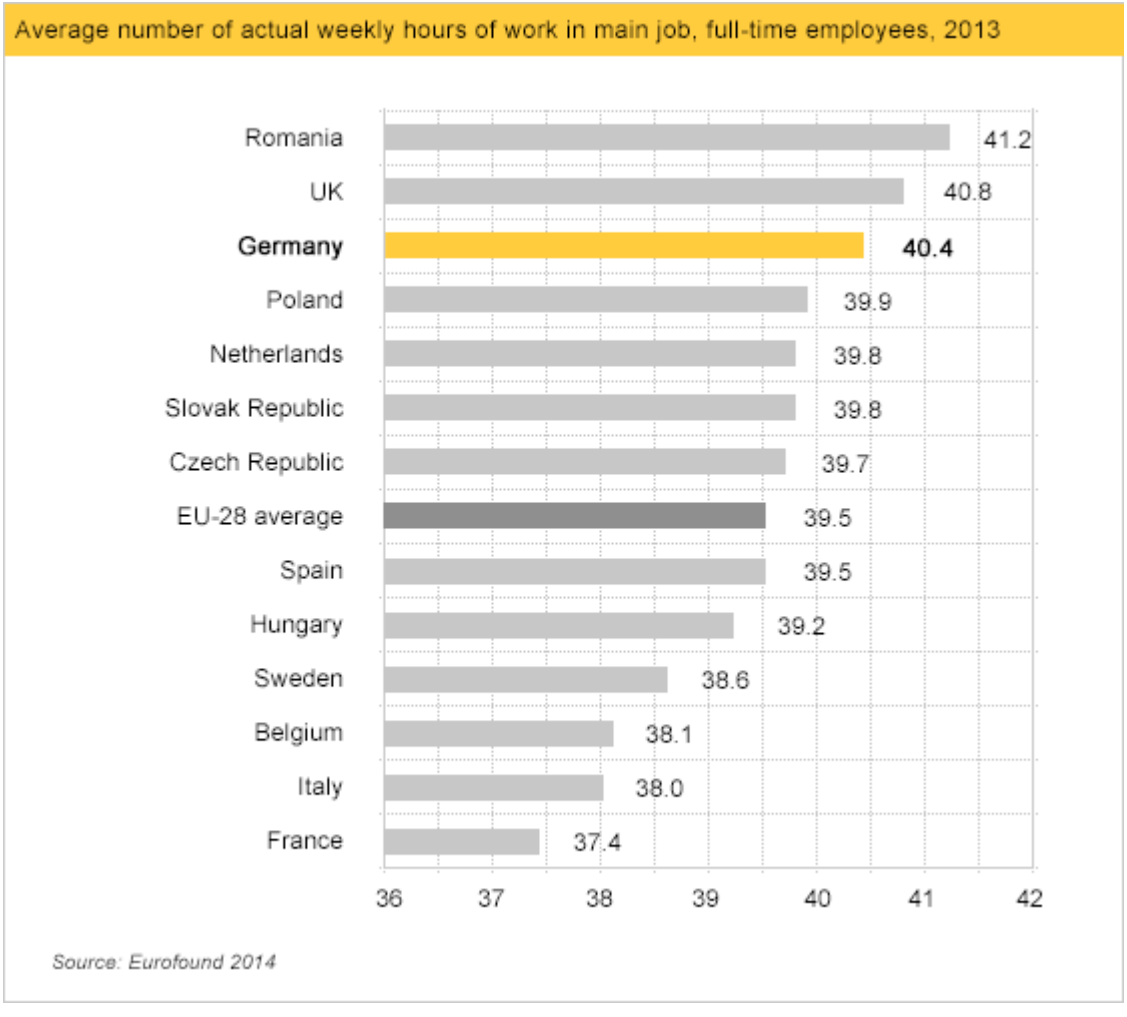
Some general rules:

Barplots:

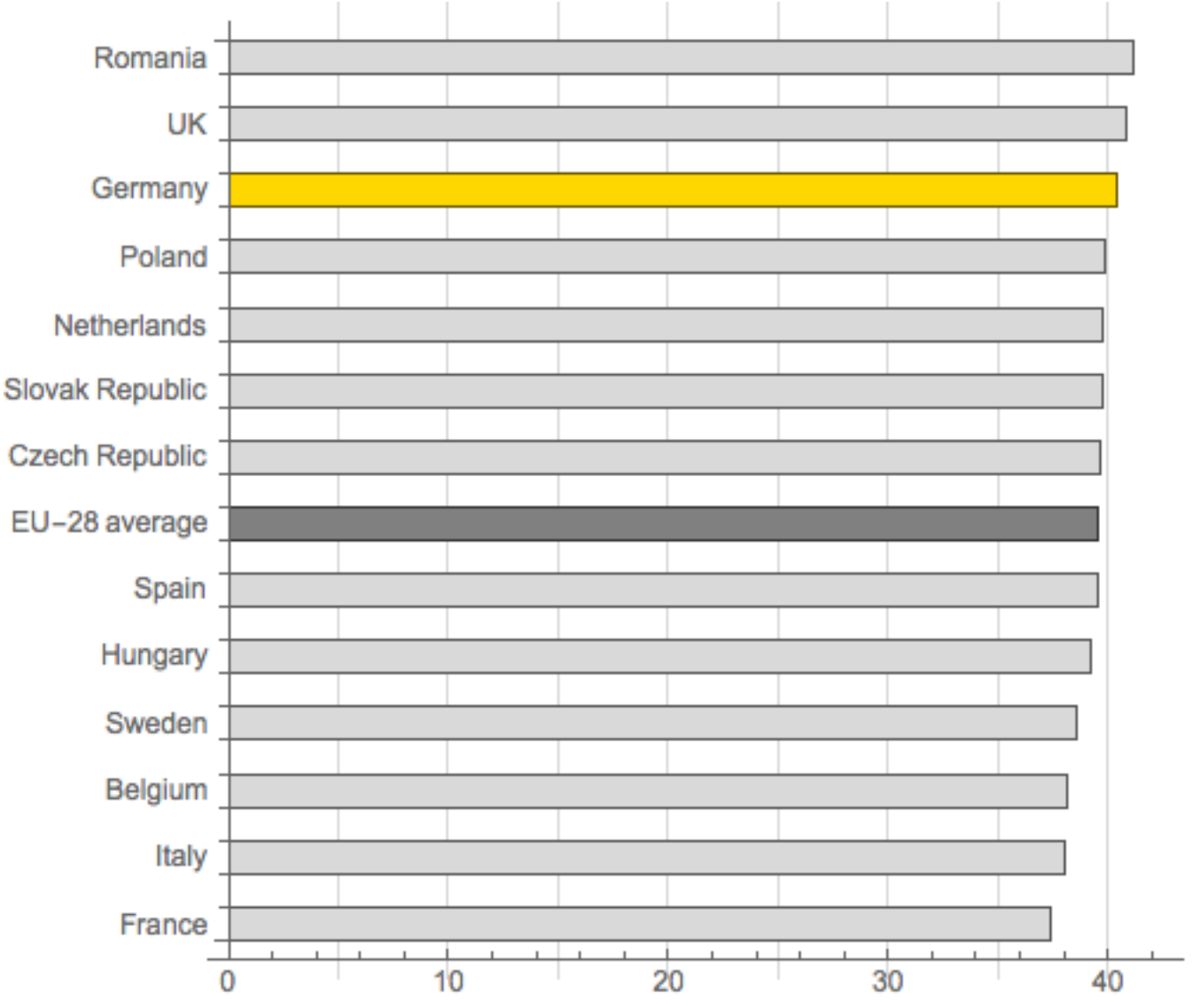
- Y-axis should start at zero to avoid distorting the visual
- The scope is to emphasize the absolute magnitude of a variable

Other graphs:

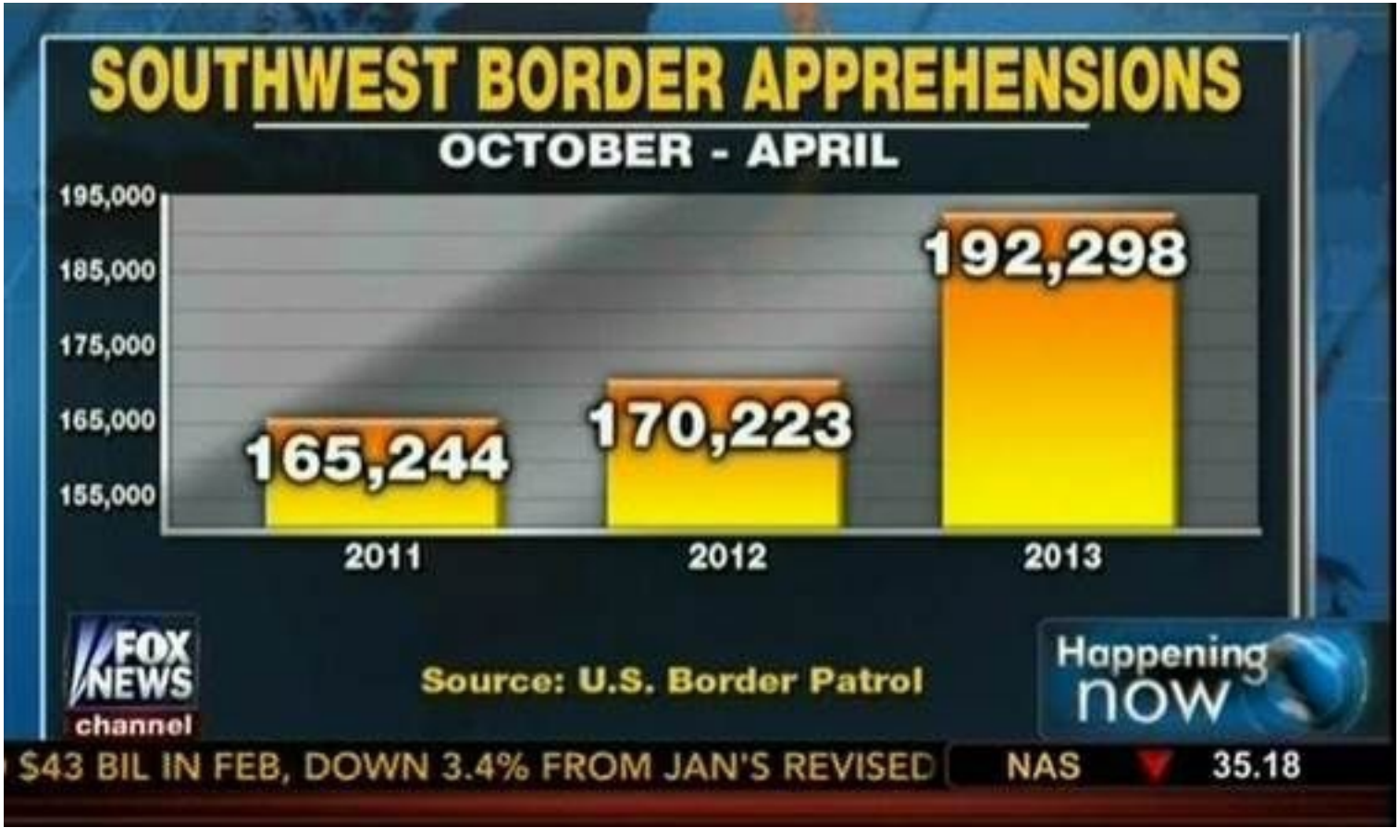
- It is not reasonable to assume that the data will assume a value of zero (e.g., stock price)
- We would not be able to observe data variation if we were to start at zero.
    - Line plot: you generally do not start at zero as the scope is to show change over time



Source: https://www.callingbullshit.org/tools/tools_misleading_axes.html



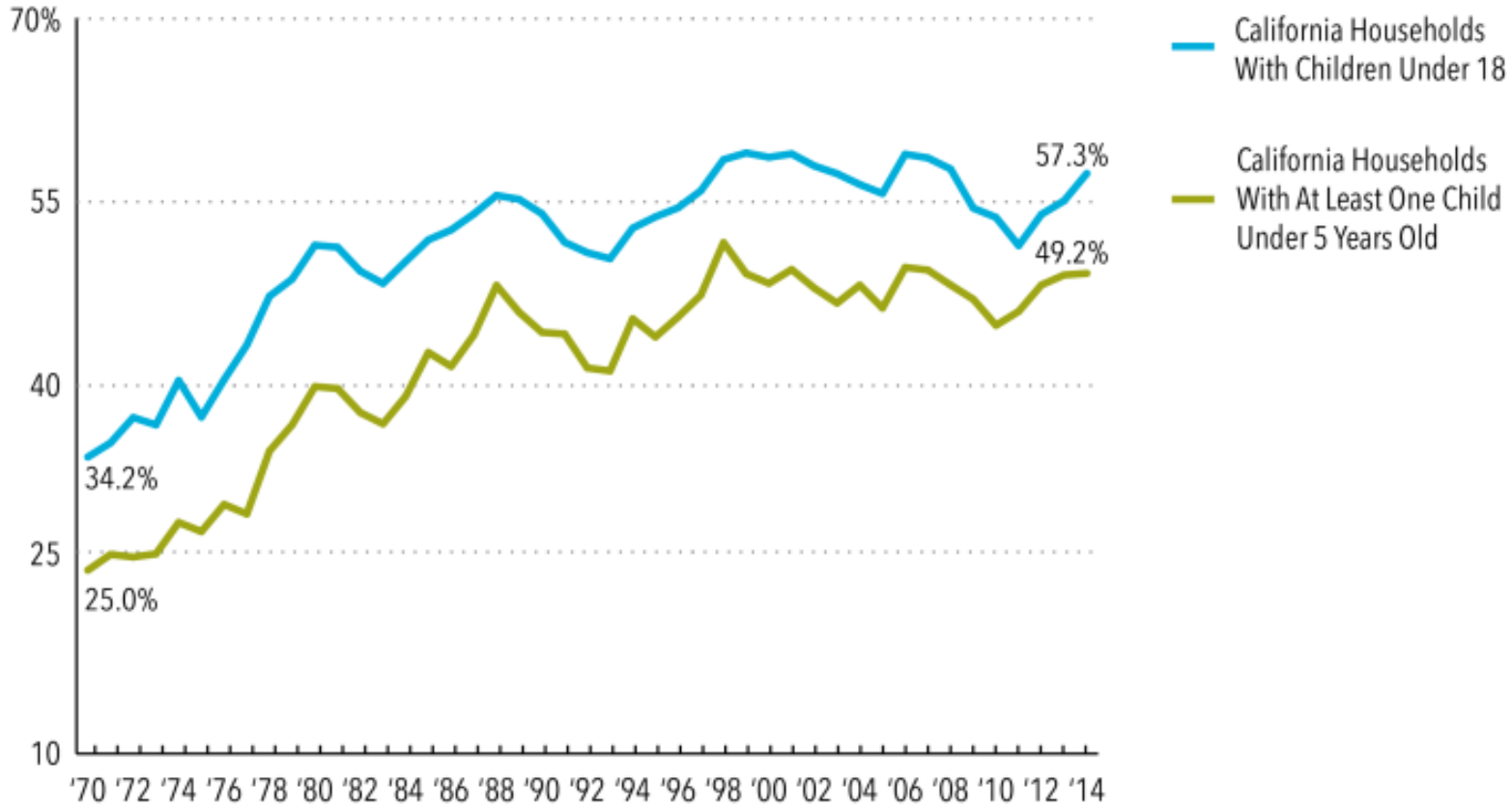
Source: https://www.callingbullshit.org/tools/tools_misleading_axes.html



Source: https://www.data-to-viz.com/caveat/cut_y_axis.html

More California Households Have All Parents Working, Making Access to Child Care an Important Priority

Percentage of California Households Where All Parents Work, 1970 to 2014

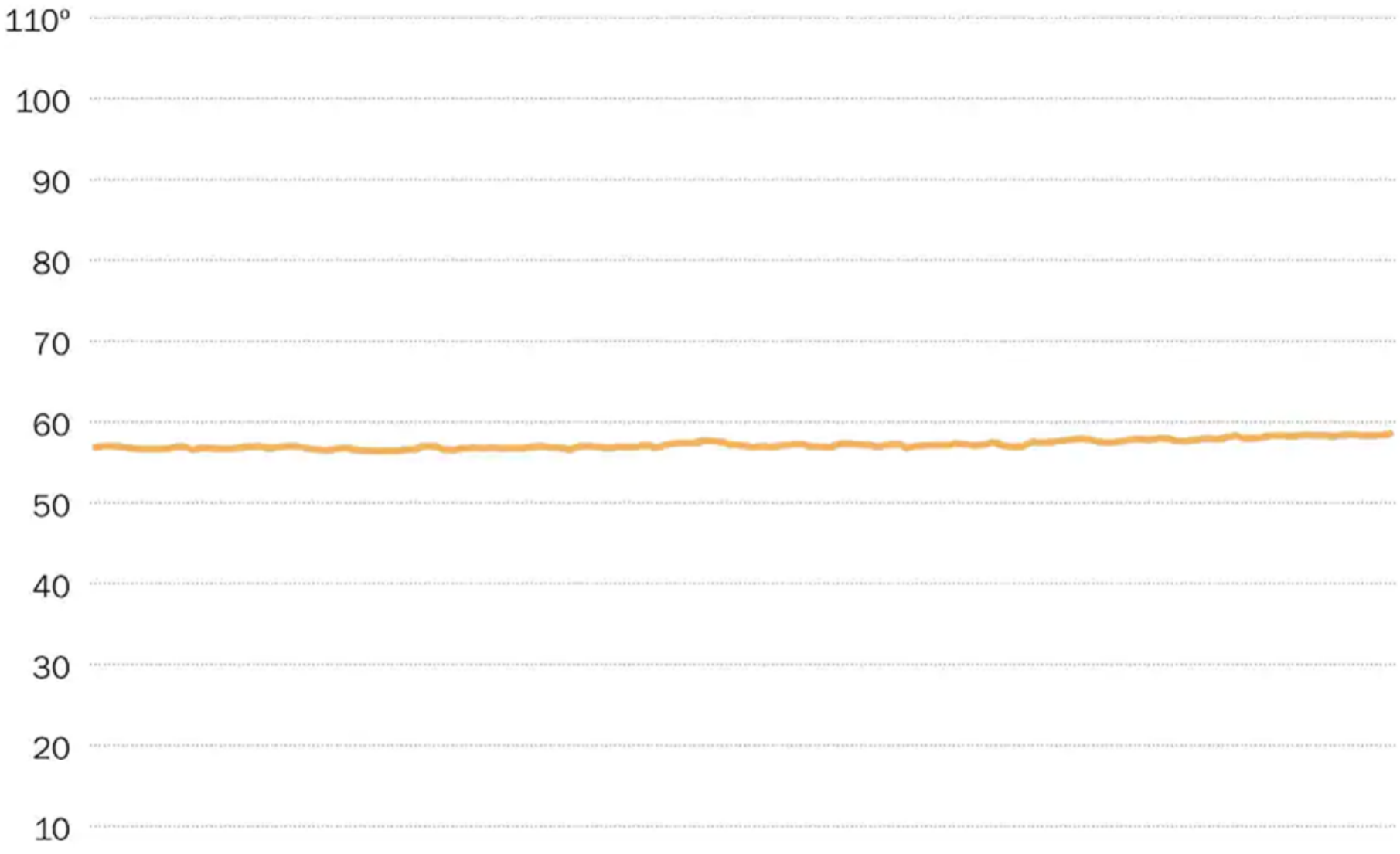


Note: A "household where all parents work" includes single-parent households and dual-earner

California Budget & Policy Center

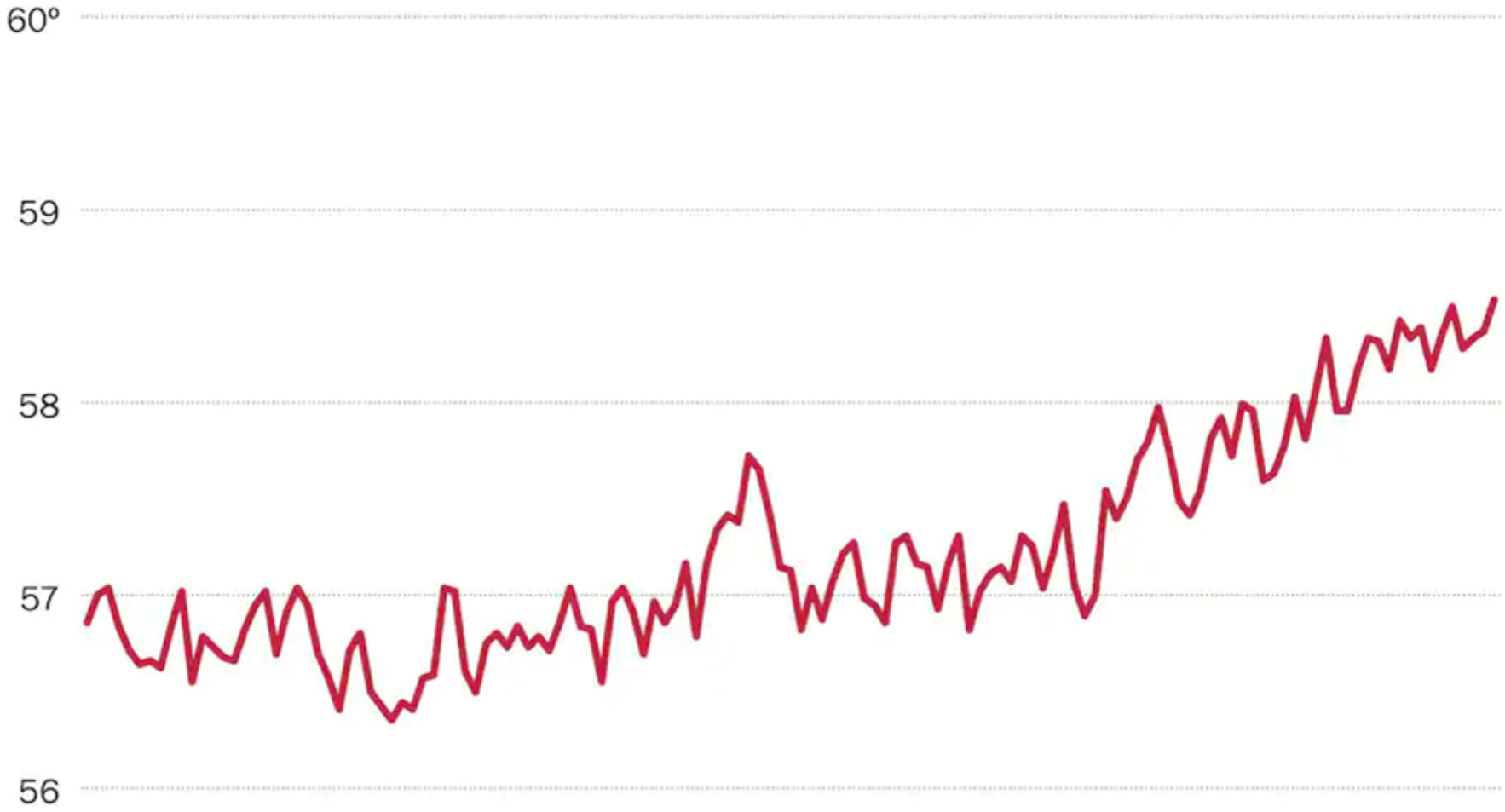
Average global temperature by year, 50x scale

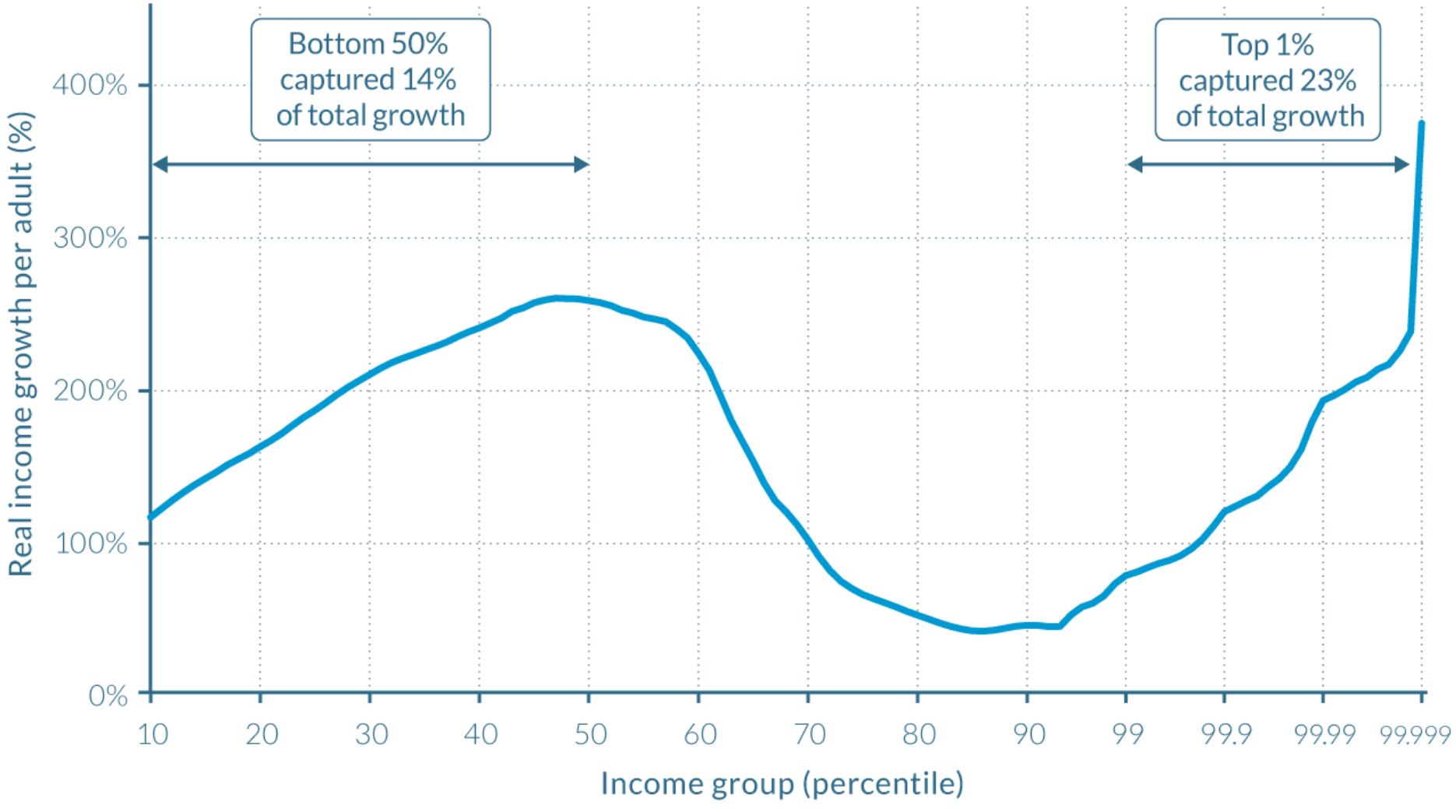
Data from NASA/GISS.



Average global temperature by year

Data from NASA/GISS.





Source: WID.world (2017). See wir2018.wid.world/methodology.html for data series and notes.

On the horizontal axis, the world population is divided into a hundred groups of equal population size and sorted in ascending order from left to right, according to each group's income level. The Top 1% group is divided into ten groups, the richest of these groups is also divided into ten groups, and the very top group is again divided into ten groups of equal population size. The vertical axis shows the total income growth of an average individual in each group between 1980 and 2016. For percentile group p99p99.1 (the poorest 10% among the world's richest 1%), growth was 77% between 1980 and 2016. The Top 1% captured 23% of total growth over this period. Income estimates account for differences in the cost of living between countries. Values are net of inflation.

Change the axis

Continuous variables

- scale_y_continous
- scale_x_continuous

Discrete variables

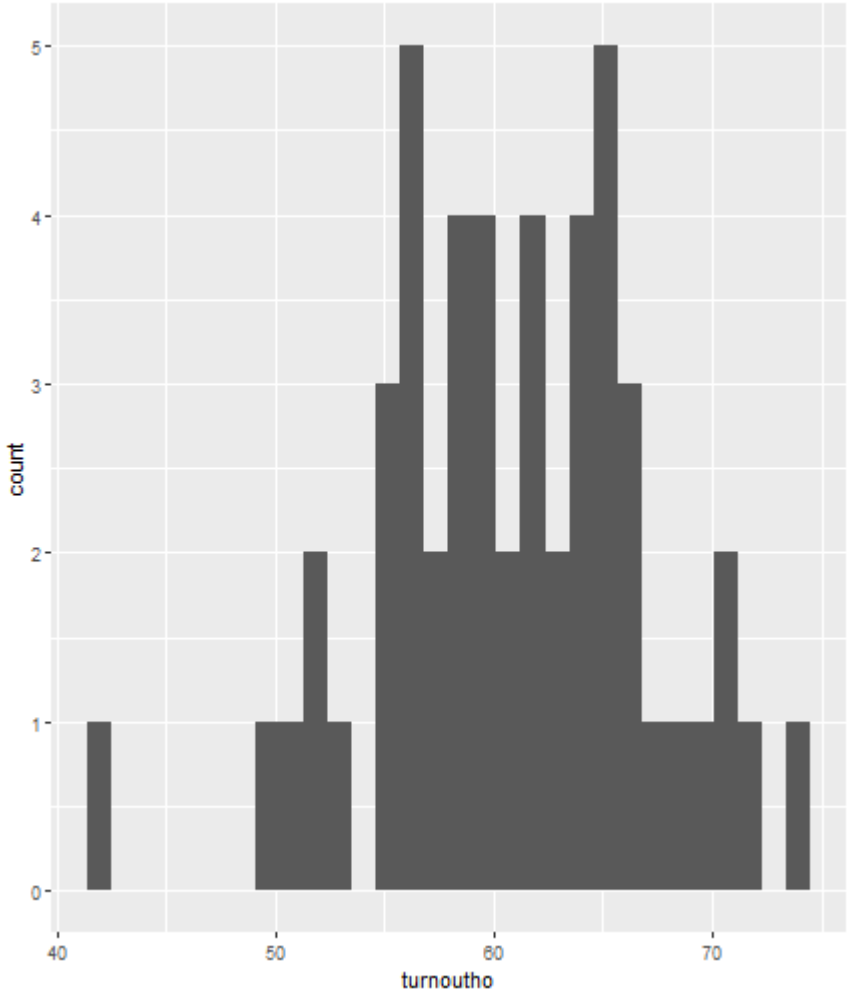
- scale_y_discrete
- scale_x_discrete

scale function

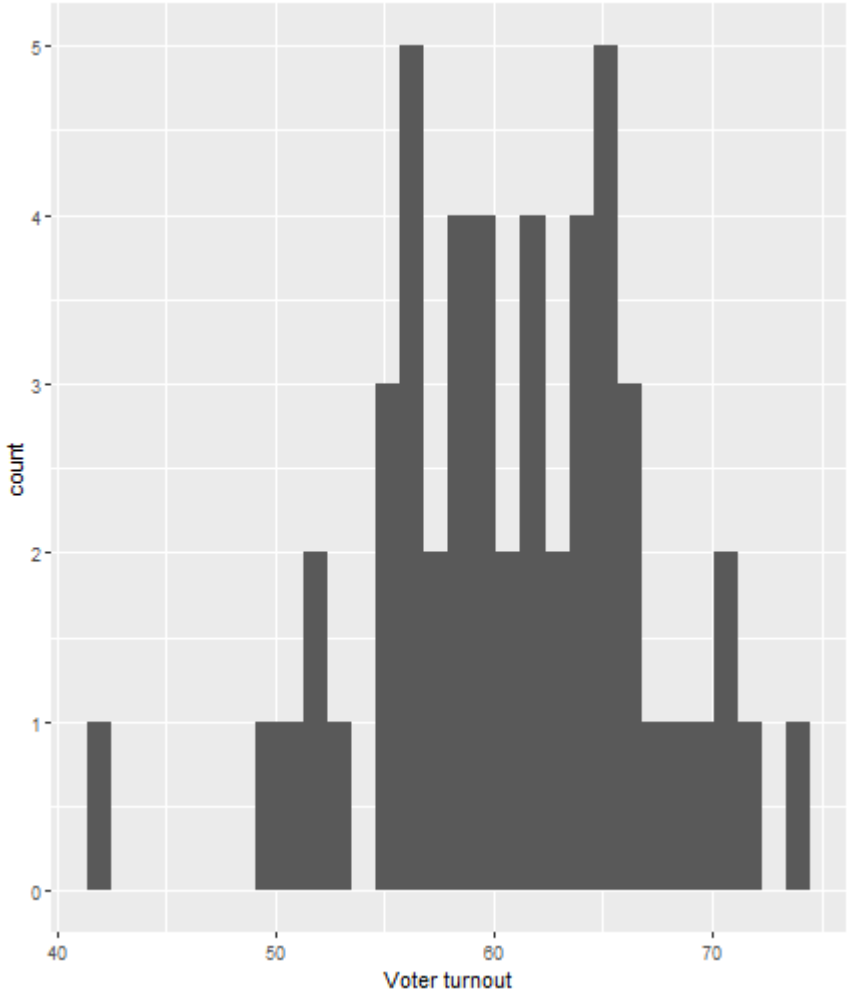
- **name:** x or y axis labels
- **breaks:** to control the breaks in axis
- **labels:** labels of axis tick marks.
- **limits:** a numeric vector specifying x or y axis limits (min, max)
- **trans:** for axis transformations. Possible values are “log2”, “log10”, …

Example

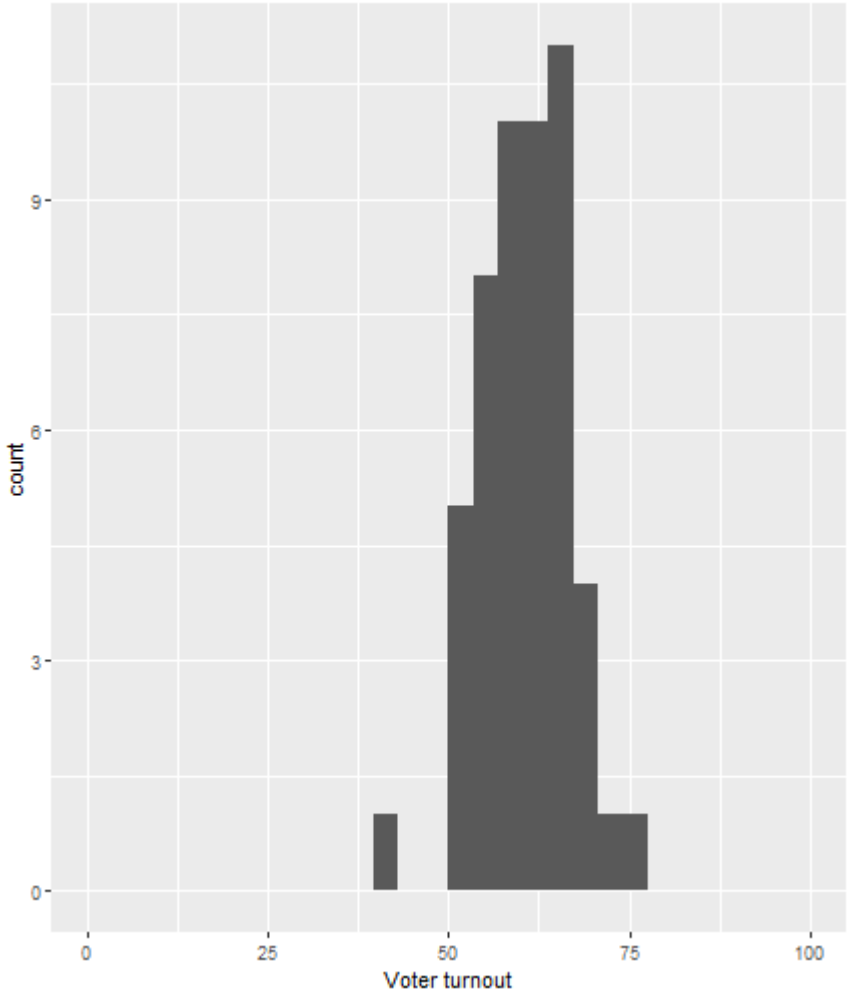
```
ggplot(data = election_turnout) +
  geom_histogram(mapping = aes(x = turnoutho)) +
  scale_x_continuous(
                       )
```



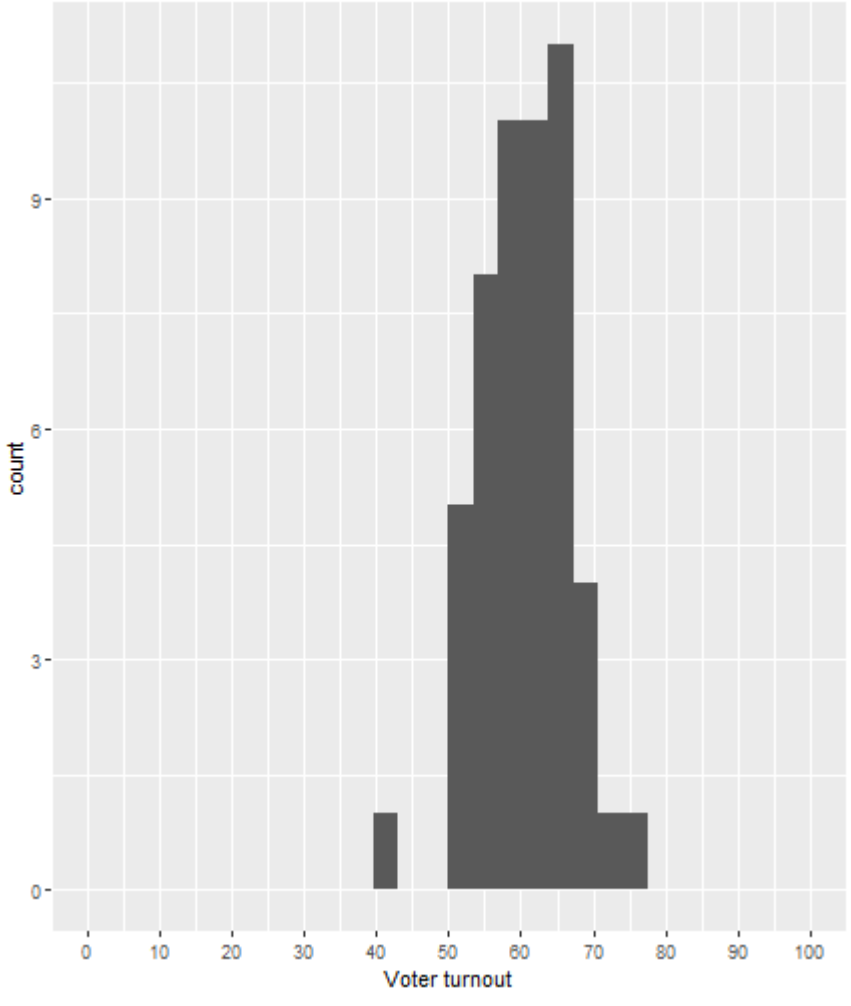
```
ggplot(data = election_turnout) +
  geom_histogram(mapping = aes(x = turnoutho)) +
  scale_x_continuous(
                       name = "Voter turnout",
                         )
```



```
ggplot(data = election_turnout) +
  geom_histogram(mapping = aes(x = turnoutho)) +
  scale_x_continuous(
                       name = "Voter turnout",
                       limits = c(0,100),
                         )
```

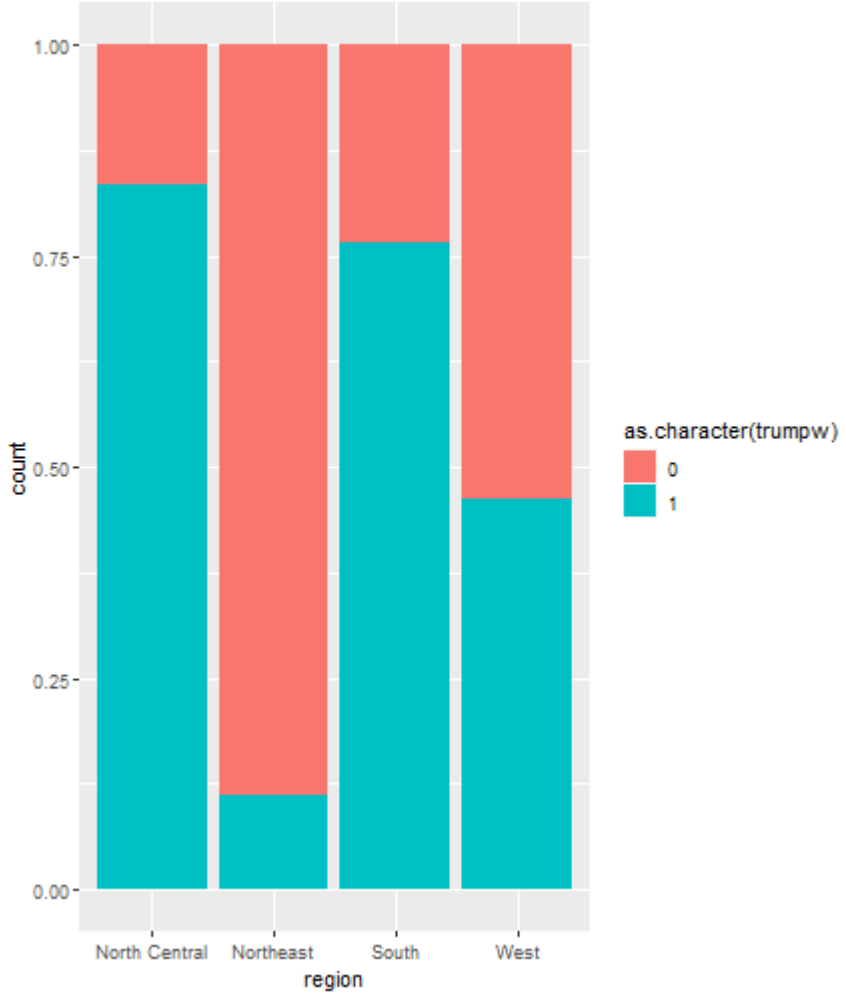


```
ggplot(data = election_turnout) +
  geom_histogram(mapping = aes(x = turnoutho)) +
  scale_x_continuous(
                      name = "Voter turnout",
                      limits = c(0,100),
                      breaks = seq(0, 100, by = 10)
                        )
```

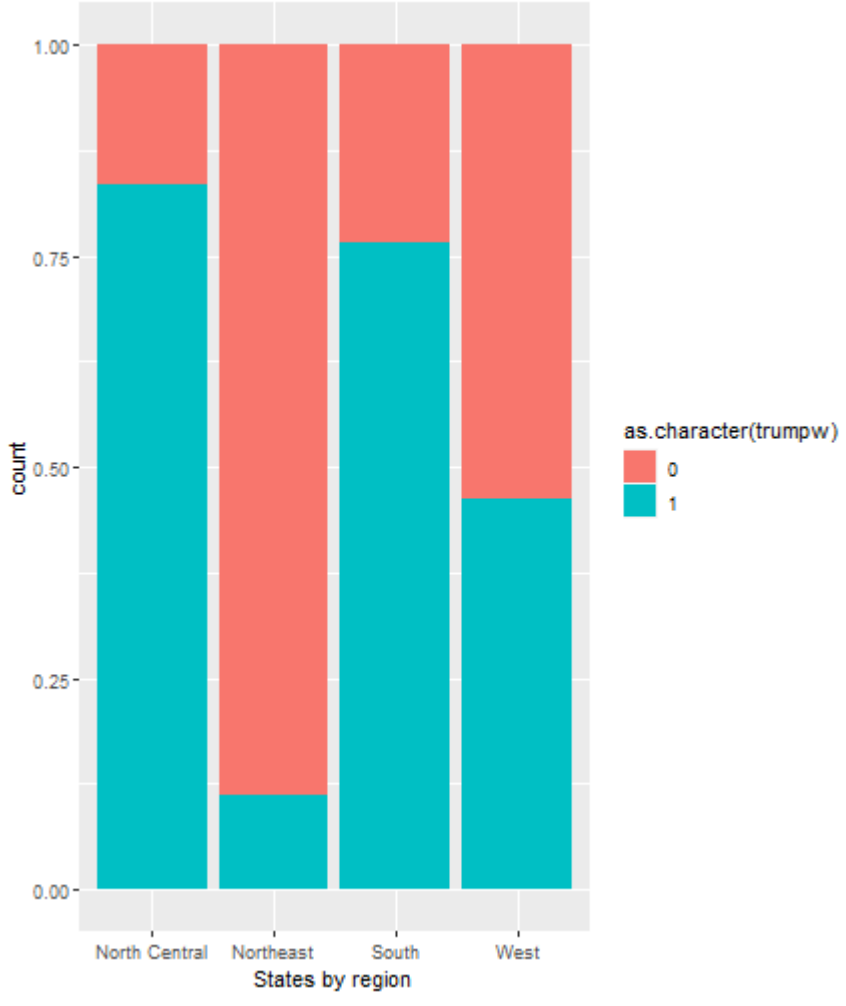


Example 2

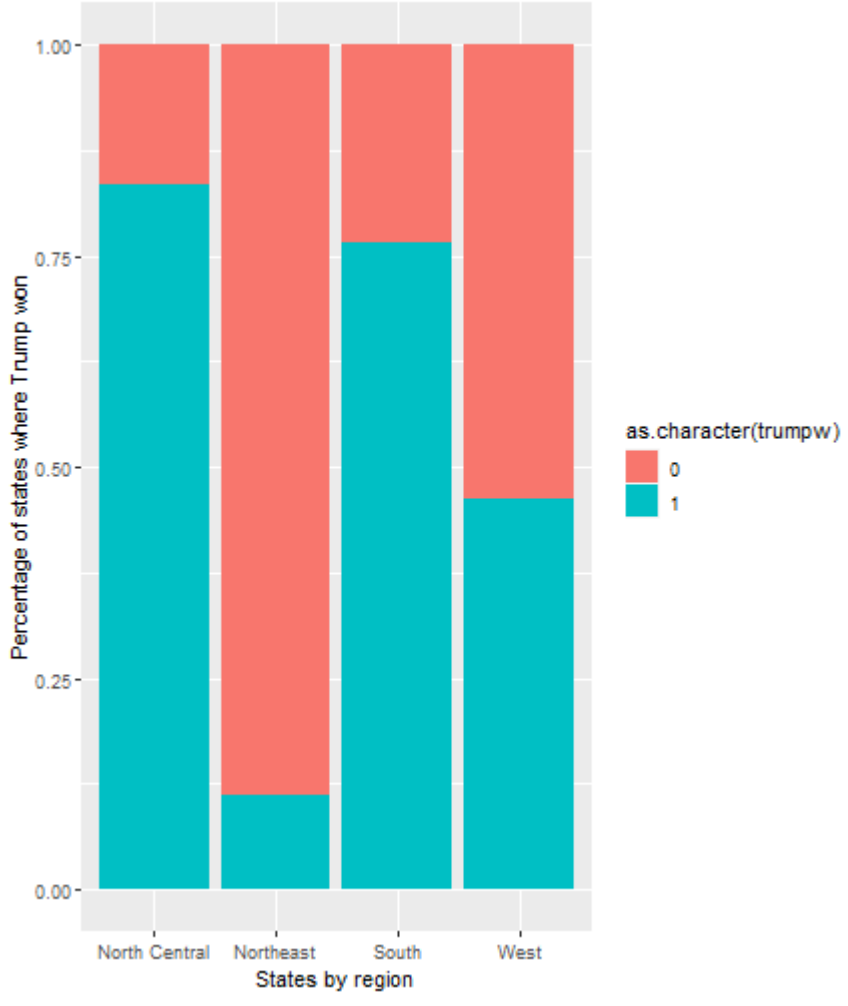
```
ggplot(data = election_turnout) +
  geom_bar(mapping = aes(x = region,
                         fill = as.character(trumpw))
                         position = "fill") +
  scale_x_discrete(
           ) +
  scale_y_continuous(
           )
```



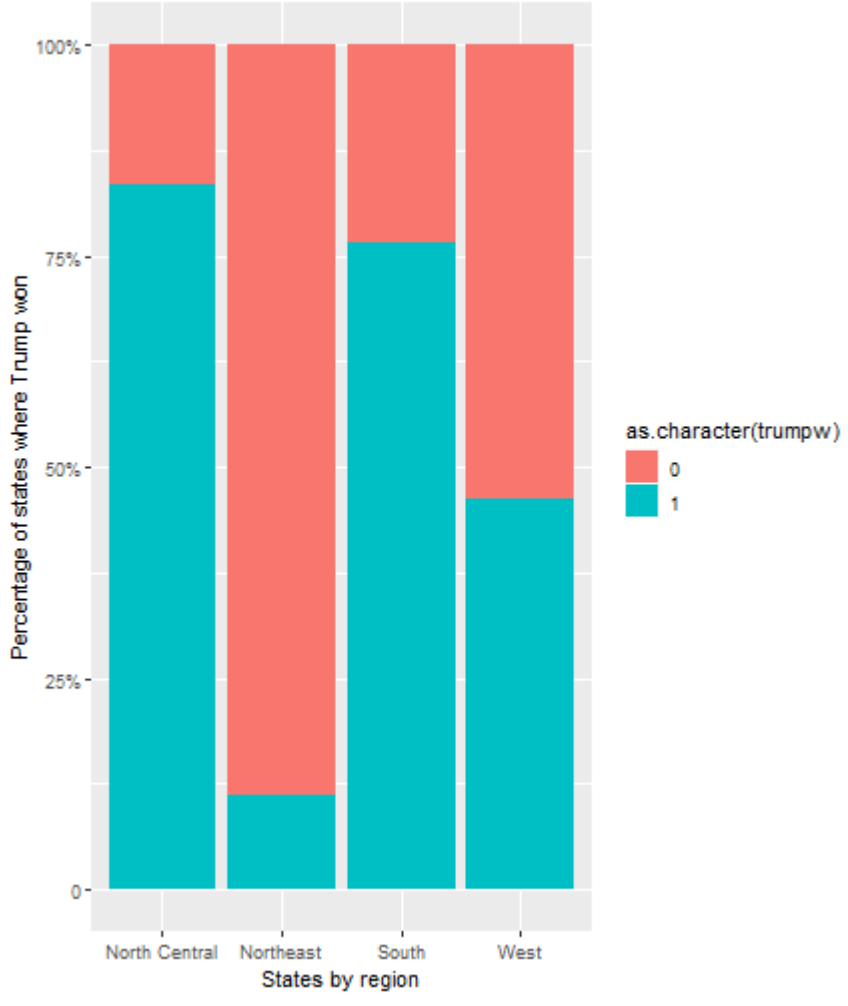
```
ggplot(data = election_turnout) +
  geom_bar(mapping = aes(x = region,
                         fill = as.character(trumpw))
                         position = "fill") +
  scale_x_discrete(
            name = "States by region"
            ) +
  scale_y_continuous(
            )
```



```
ggplot(data = election_turnout) +
  geom_bar(mapping = aes(x = region,
                         fill = as.character(trumpw))
                         position = "fill") +
  scale_x_discrete(
           name = "States by region"
           ) +
  scale_y_continuous(
           name = "Percentage of states where Trump w
           )
```

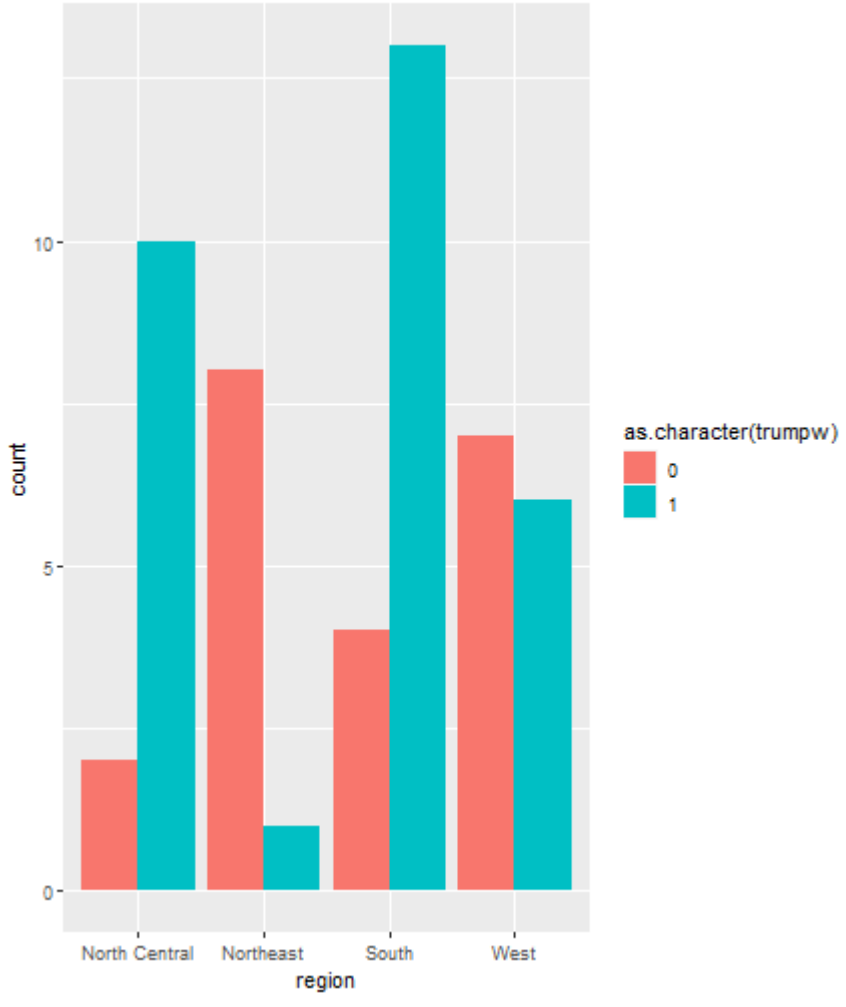


```
ggplot(data = election_turnout) +
  geom_bar(mapping = aes(x = region,
                         fill = as.character(trumpw))
                         position = "fill") +
  scale_x_discrete(
           name = "States by region"
           ) +
  scale_y_continuous(
           name = "Percentage of states where Trump v
           labels = c("0", "25%", "50%", "75%", "100%
           )
```

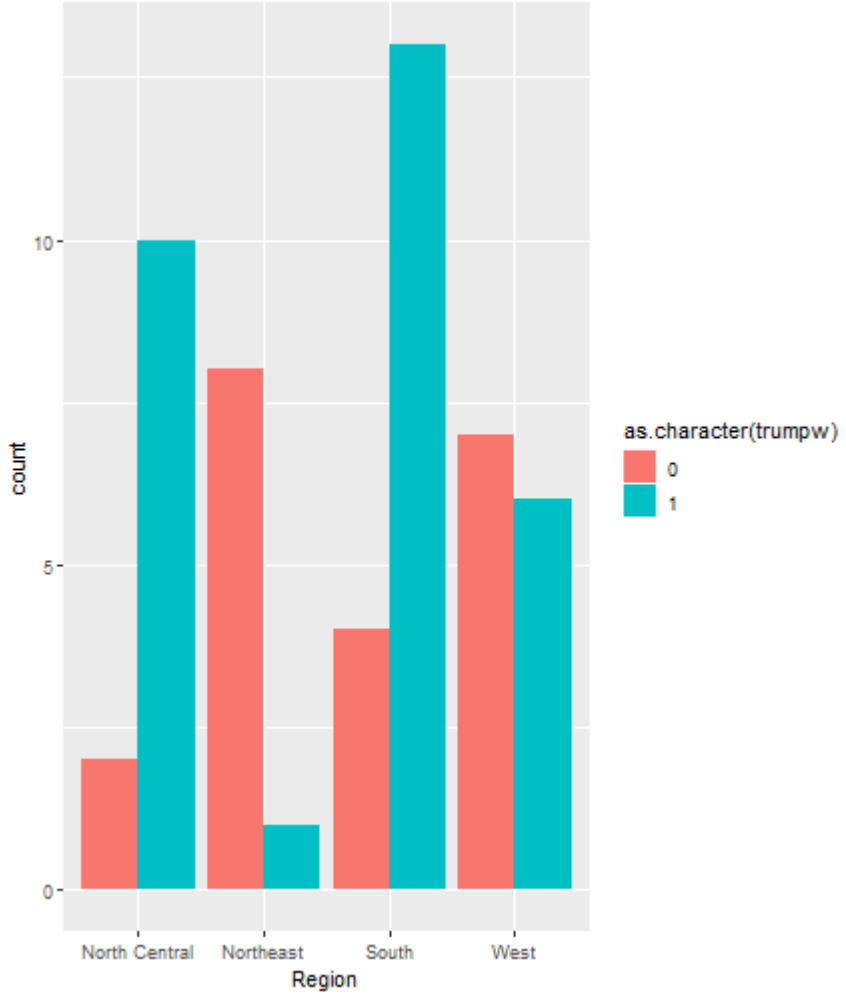


Example 3

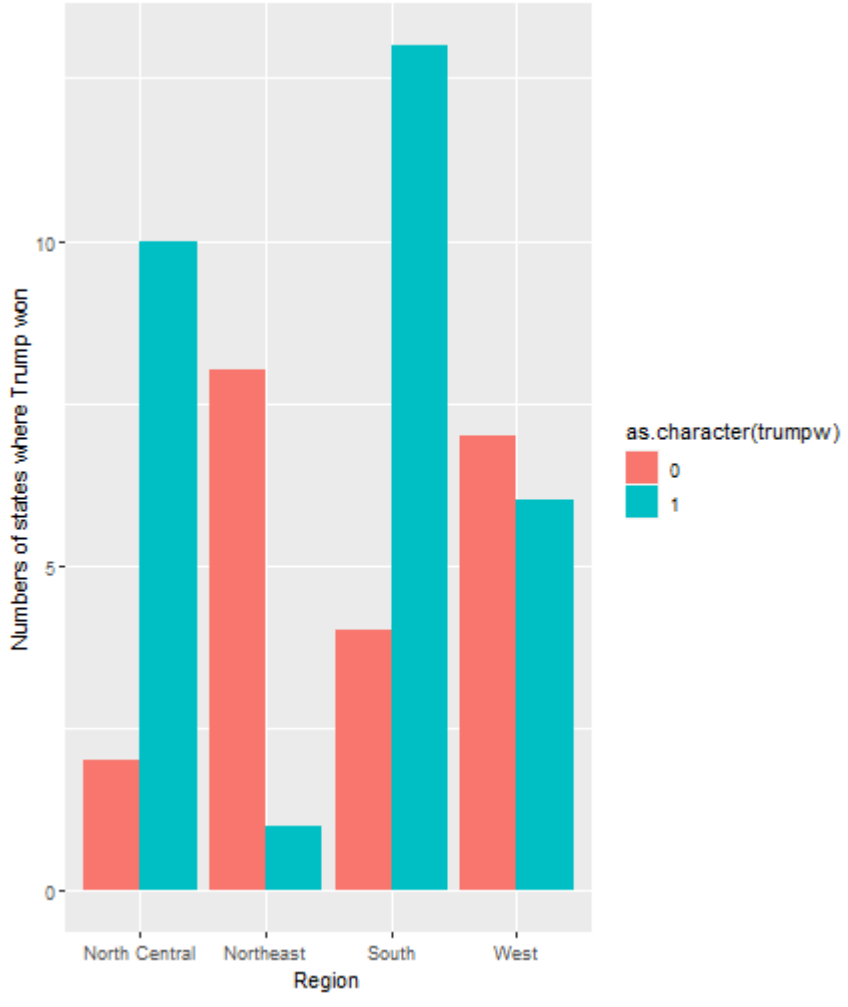
```
ggplot(data = election_turnout) +
  geom_bar(mapping = aes(x = region,
                         fill = as.character(trumpw))
                         position = "dodge") +
  scale_x_discrete(
            ) +
  scale_y_continuous(
            )
```



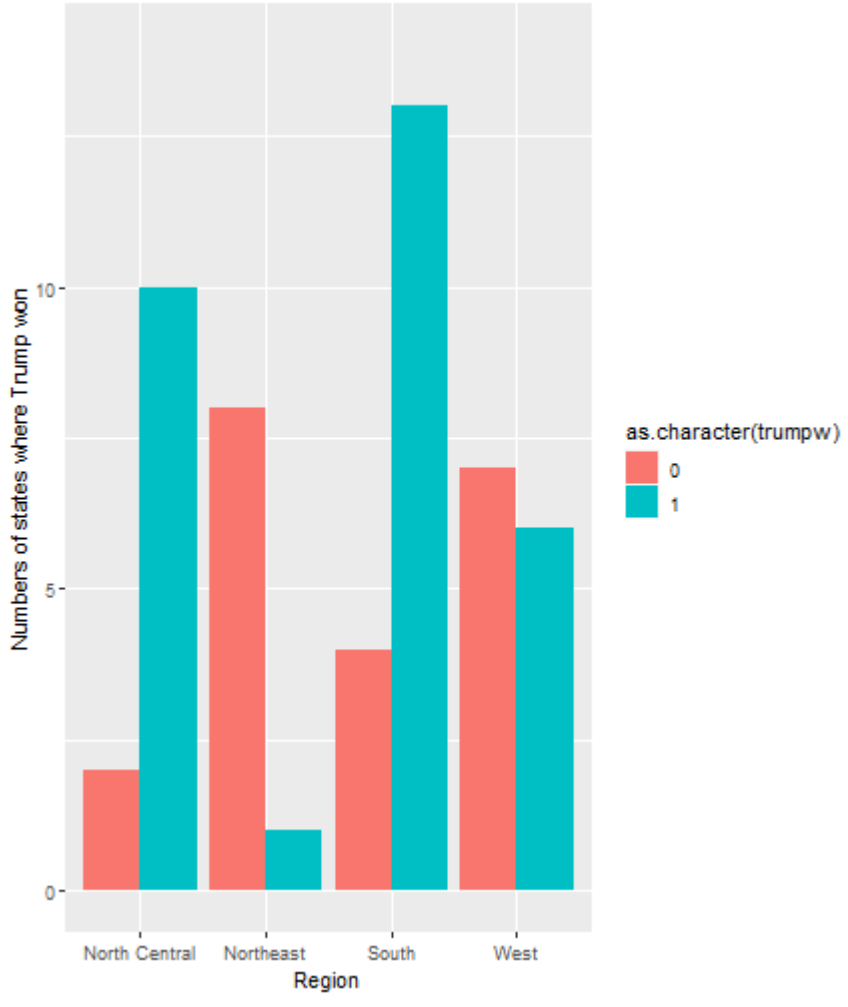
```
ggplot(data = election_turnout) +
  geom_bar(mapping = aes(x = region,
                         fill = as.character(trumpw))
                         position = "dodge") +
  scale_x_discrete(
            name = "Region"
            ) +
  scale_y_continuous(
            )
```



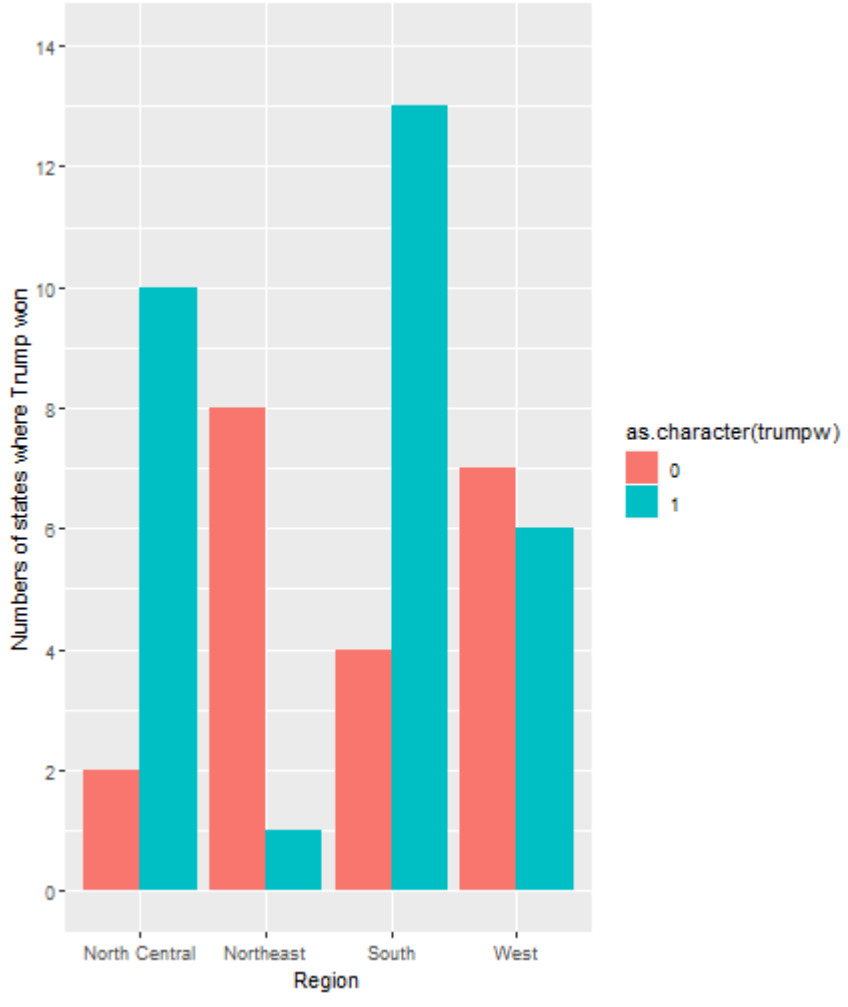
```
ggplot(data = election_turnout) +
  geom_bar(mapping = aes(x = region,
                         fill = as.character(trumpw))
                         position = "dodge") +
  scale_x_discrete(
            name = "Region"
            ) +
  scale_y_continuous(
            name = "Numbers of states where Trump wor
            )
```



```
ggplot(data = election_turnout) +
  geom_bar(mapping = aes(x = region,
                         fill = as.character(trumpw))
                         position = "dodge") +
  scale_x_discrete(
            name = "Region"
            ) +
  scale_y_continuous(
            name = "Numbers of states where Trump wor
            limits = c(0,14),
            )
```



```
ggplot(data = election_turnout) +
  geom_bar(mapping = aes(x = region,
                         fill = as.character(trumpw))
                         position = "dodge") +
  scale_x_discrete(
            name = "Region"
            ) +
  scale_y_continuous(
            name = "Numbers of states where Trump wor
            limits = c(0,14),
            breaks = c(0, 2, 4, 6, 8, 10, 12, 14),
            )
```



```
ggplot(data = election_turnout) +
  geom_bar(mapping = aes(x = region,
                         fill = as.character(trumpw))
                         position = "dodge") +
  scale_x_discrete(
            name = "Region"
            ) +
  scale_y_continuous(
            name = "Numbers of states where Trump wor
            limits = c(0,14),
            breaks = c(0, 2, 4, 6, 8, 10, 12, 14),
            labels = c("0", "2", "4", "6", "8", "10",
            )
```