Essential Research Toolkit for the Humanities

Week 6: Data visualization

Anna Pryslopska

May 16, 2022

Psycholinguistics and Cognitive Modeling Lab

Questions?

1. Clean `moses` using pipes.

```
moses_clean <-
  moses %>%
    na.omit() %>%
    select(ID, Item, Condition, Answer) %>%
    arrange(Item, Condition) %>%
    mutate(Answer_cleaned = fct_collapse(Answer,
           cant_answer = cant_answer,
           dont_know = dont_know,
           armstrong = armstrong,
           everest = everest,
           madrid = madrid,
           manchester = manchester,
           nobel = nobel,
           olympics = olympics,
           platypus = platypus,
           prince = prince,
           printing = printing,
           …)
```

2. Add the correct answers and rename the accuracy categories.

```
moses_answers <-
  moses_clean %>%
    merge(correct_answer, by = c("Item", "Condition"))

moses_accuracy <-
  moses_answers %>%
    mutate(Accuracy = ifelse(Answer_cleaned == Correct_Answer,
        yes = "Correct answer",
        no = ifelse(Answer_cleaned == "dont_know",
        yes = "Don't know", no = "Incorrect answer"))
```

3. Create a new column "Group".

```
moses_accuracy <-
  moses_accuracy %>%
    mutate(Group = ifelse(Condition == 1, "Moses illusion",
        ifelse(Condition == 2, "well-formed",
        ifelse(Condition == 100, "control", "distorted"))))
```

4. Summarize the results

```
moses_accuracy %>%
  group_by(Group, Accuracy) %>%
  summarise(Count = n()) %>%
  mutate(Frequency = 100*Count / sum(Count)) %>%
  select(Group, Accuracy, Frequency)
```

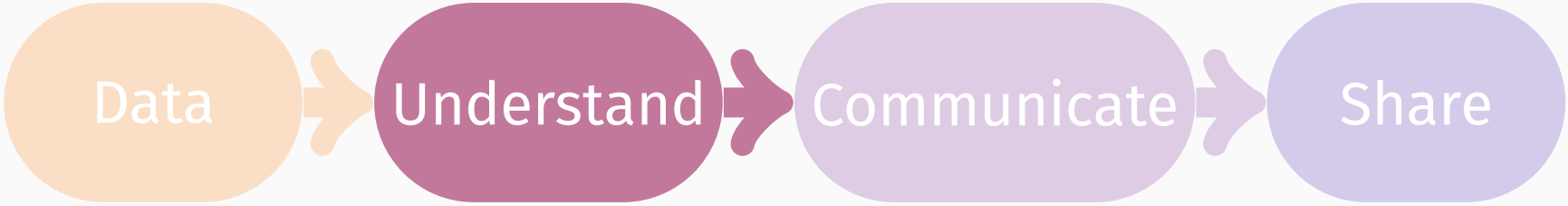
Group	Accuracy	Frequency
<chr>	<chr>	<dbl>
control	Correct answer	57.1
control	Don't know	26.9
control	Incorrect answer	16.0
distorted	Correct answer	75
distorted	Don't know	17.9
distorted	Incorrect answer	7.14
Moses illusion	Correct answer	26.4
Moses illusion	Don't know	25.7
Moses illusion	Incorrect answer	47.9
well-formed	Correct answer	75.6
well-formed	Don't know	17.8
well-formed	Incorrect answer	6.67

5. Clean and summarize `adjectives`.

```
adjectives %>%
  na.omit() %>%
  filter(ADVERB != "123" & ADVERB != "dghdhffhg" & age >=17 &
         LIST %in% 1:6 & Value %in% 1:7) %>%
  select(Value, ADJECTIVE, ADVERB) %>%
  group_by(ADVERB) %>%
  summarise(mean = mean(Value),
            sd = sd(Value),
            min = min(Value),
            max = max(Value),
            count = n()) %>%
  arrange(ADVERB)
```

	ADVERB	mean	sd	min	max	count
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<int>
1	absichtlich	3.07	1.89	1	7	3778
2	bewusst	3.37	1.96	1	7	3777
3	freiwillig	2.64	1.72	1	7	3777

Understanding → Communicating



import, clean, transform, export, visualize, model

document, create clean and beautiful reports

connect, collaborate, backup

Table of contents

1. Exporting to file
2. Data visualization
3. Plotting with R
4. Choosing colors
5. Best practices
6. Bad examples
7. Homework assignment

Exporting to file

Moving on from R

Save to **current working directory**.

```
write_csv(WHAT, "PATH TO WHERE", row.names = FALSE)
```
comma

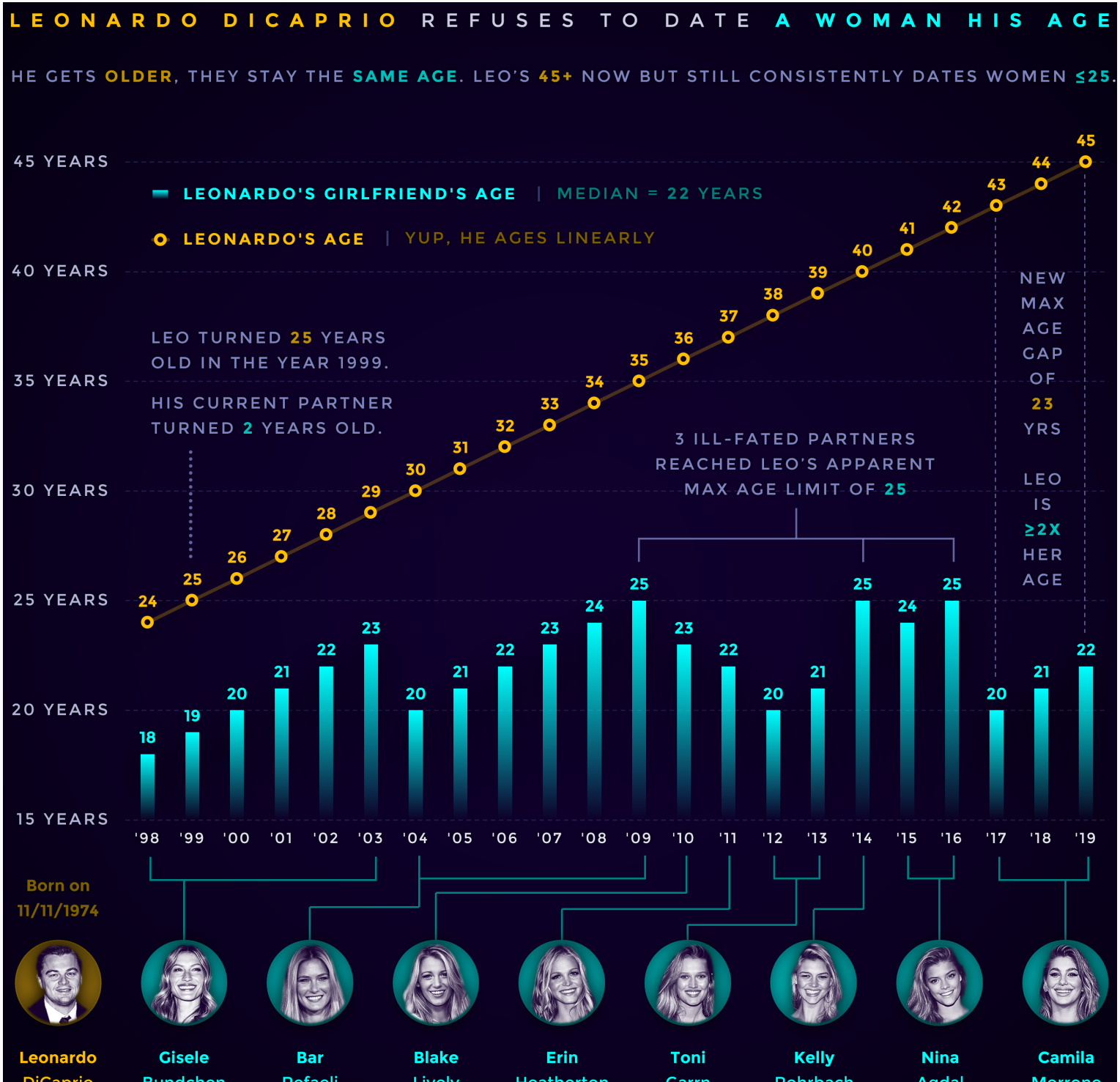
```
write_tsv()
```
tab

```
write_excel_csv()
```
CSV for Excel

```
write_delim()
```
specify how to separate

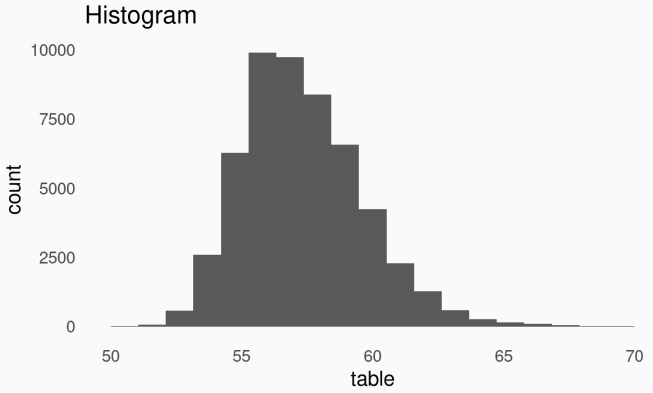
```
write_csv(moses_accuracy, "Moses accuracy.csv",
row.names = FALSE)

write_delim(moses_accuracy, "Moses accuracy.txt",
row.names = FALSE, delim = ":")
```

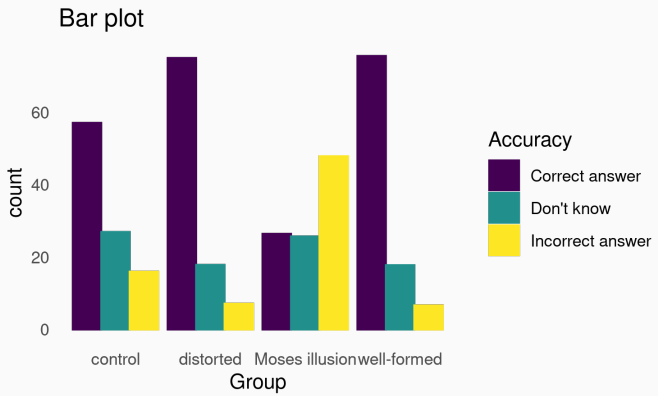
Data visualization

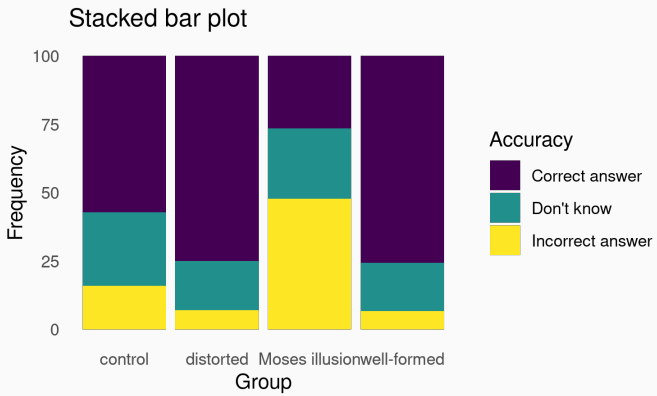

LEONARDO DICAPRIO REFUSES TO DATE A WOMAN HIS AGE
HE GETS OLDER, THEY STAY THE SAME AGE. LEO'S 45+ NOW BUT STILL CONSISTENTLY DATES WOMEN ≤25.
45 YEARS
40 YEARS
35 YEARS
30 YEARS
25 YEARS
20 YEARS
15 YEARS
LEONARDO'S GIRLFRIEND'S AGE | MEDIAN = 22 YEARS
LEONARDO'S AGE | YUP, HE AGES LINEARLY
LEO TURNED 25 YEARS OLD IN THE YEAR 1999.
HIS CURRENT PARTNER TURNED 2 YEARS OLD.
3 ILL-FATED PARTNERS REACHED LEO'S APPARENT MAX AGE LIMIT OF 25
NEW MAX AGE GAP OF 23 YRS
LEO IS ≥2X HER AGE
24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45
18 19 20 21 22 23 20 21 22 23 24 25 23 22 20 21 25 24 25 20 21 22
'98 '99 '00 '01 '02 '03 '04 '05 '06 '07 '08 '09 '10 '11 '12 '13 '14 '15 '16 '17 '18 '19
Born on 11/11/1974
Leonardo DiCaprio
Gisele Bundchen
Bar Refaeli
Blake Lively
Erin Heatherton
Toni Garrn
Kelly Rohrbach
Nina Agdal
Camila Morrone

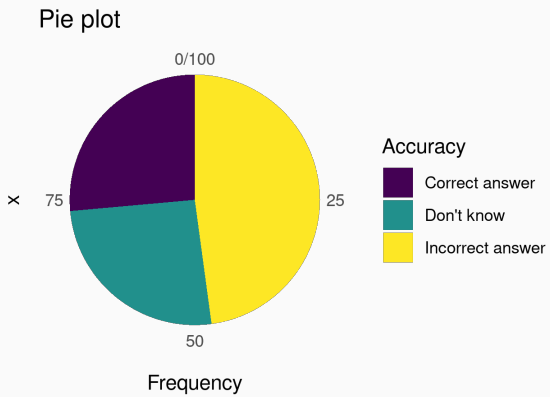
Data viz types

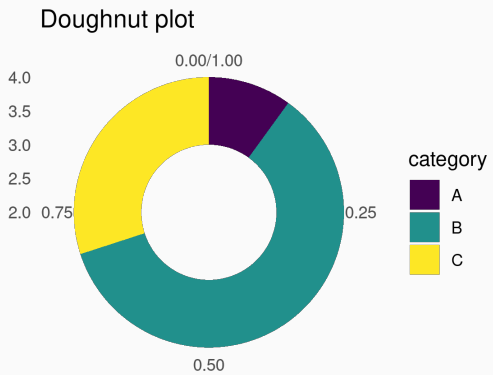
IA	Variable	Est.	SE	df	t/z	p≤	95% CI
First pass duration							
5	(intercept)	5.38	0.02	39	218.36	0.00	5.33, 5.43
5	conjunction	−0.07	0.02	892	−3.69	0.00	−0.10, −0.03
5	verb×conjunction	0.07	0.04	896	1.93	0.05	−0.00, 0.14
6	(intercept)	5.73	0.03	67	171.02	0.00	5.66, 5.79
6	conjunction	−0.10	0.02	2154	−5.91	0.00	−0.13, −0.07



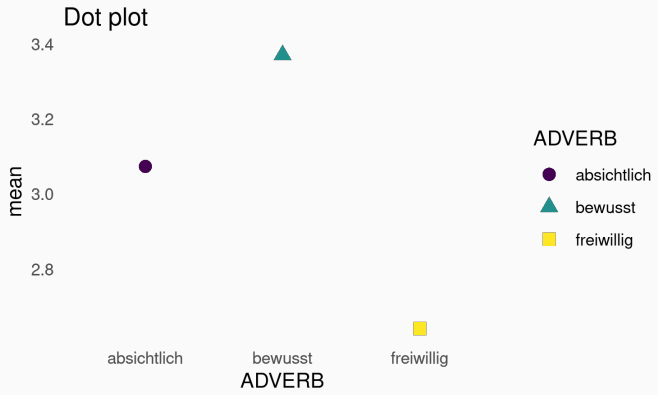


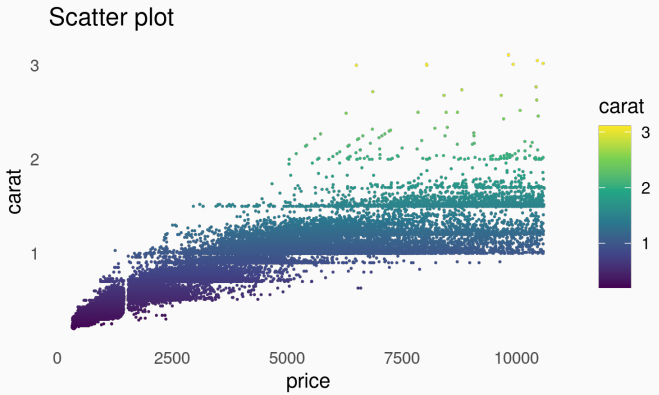


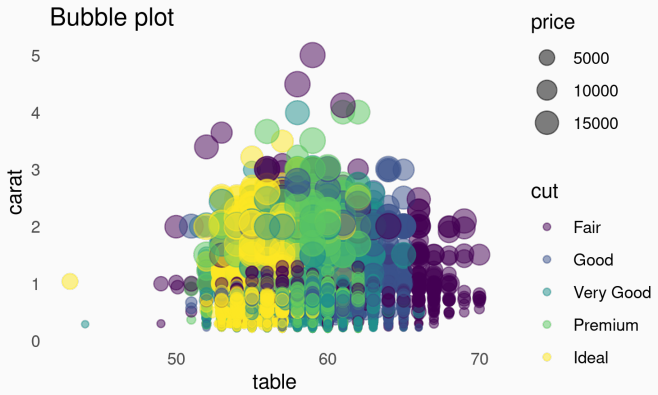


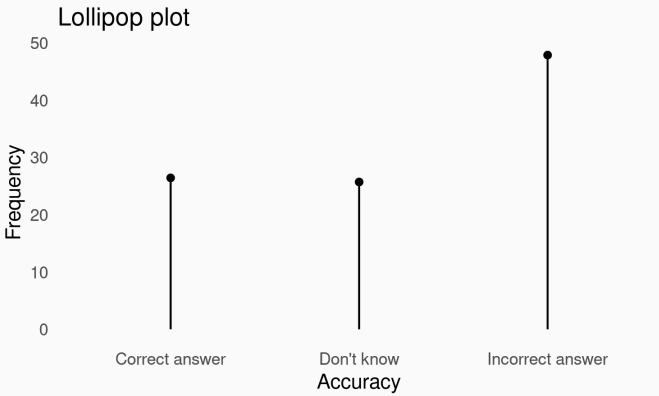


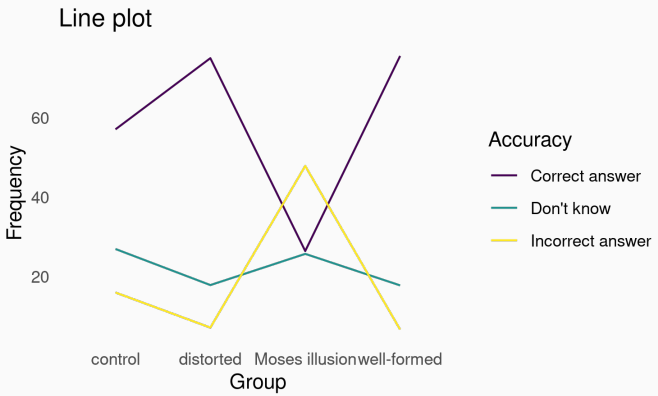
Data viz types

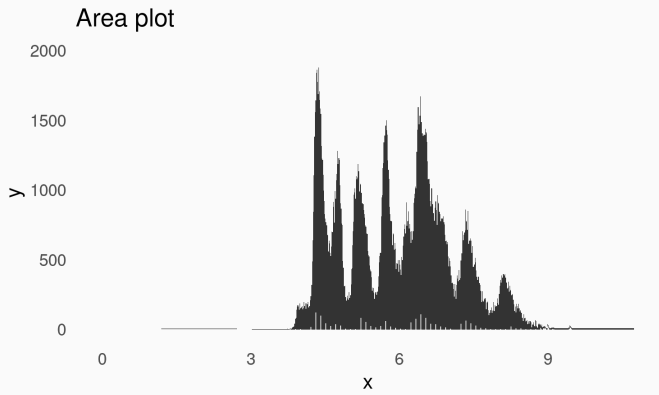




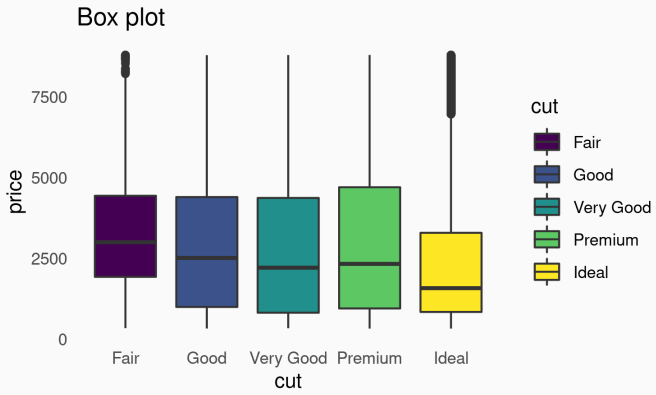


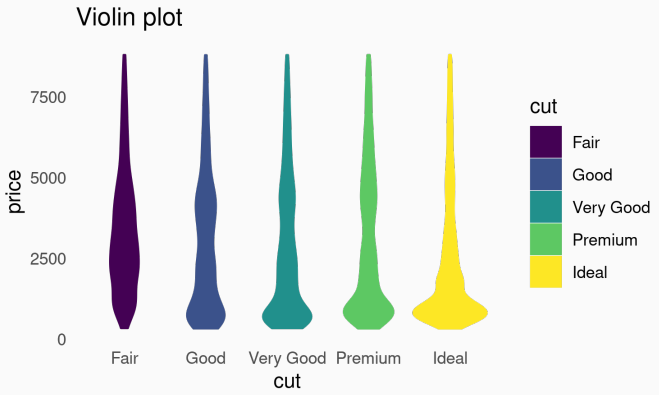


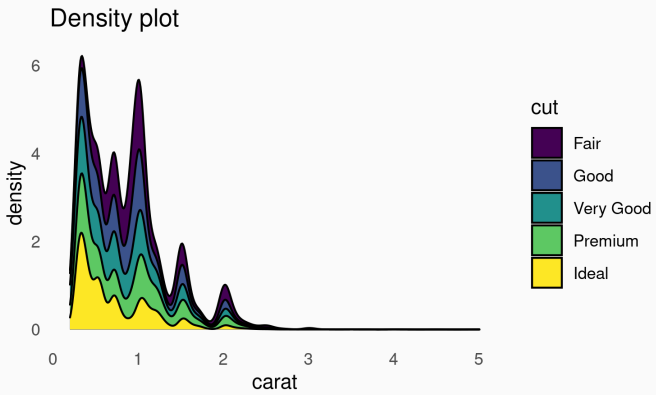


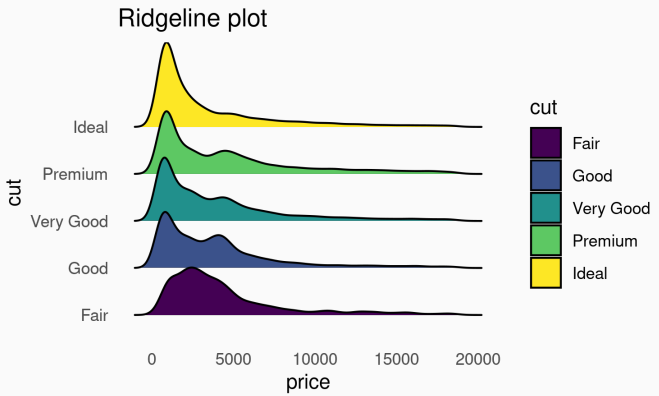


Data viz types



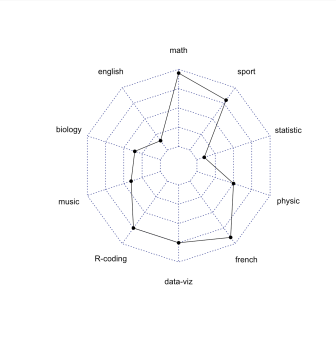




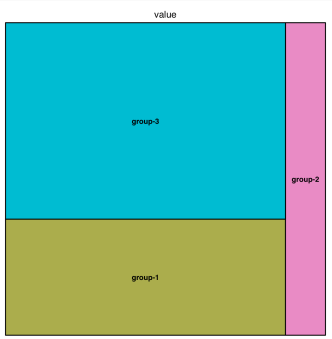


Data viz types beyond ggplot2

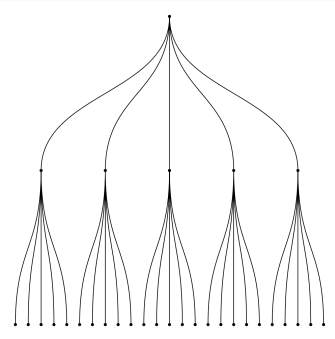
Radar



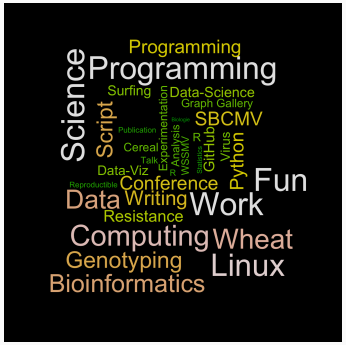
Treemap



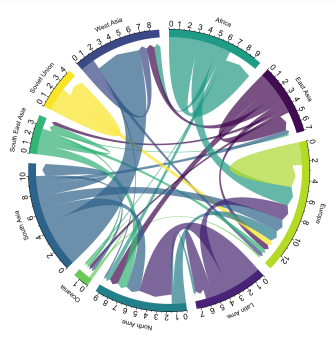
Dendrogram



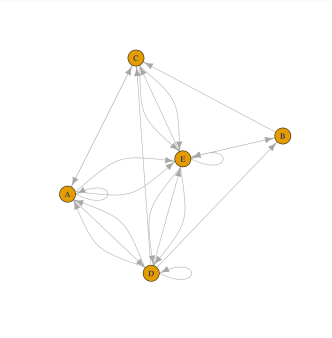
Wordcloud



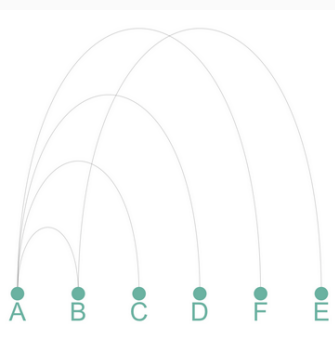
Chord



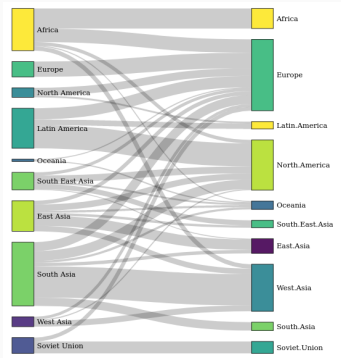
Network



Arc



Sankey



Source: `www.data-to-viz.com`

Plotting with R

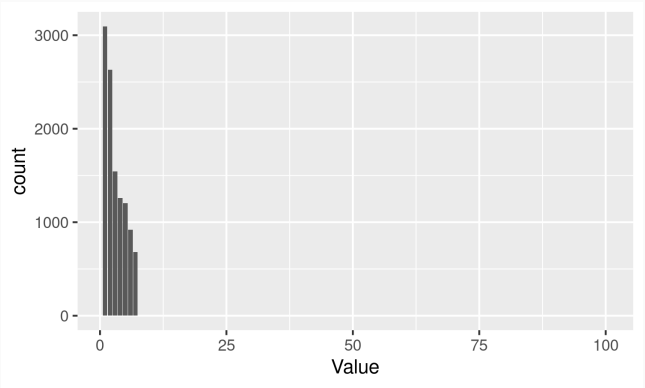
ggplot2: grammar of graphics

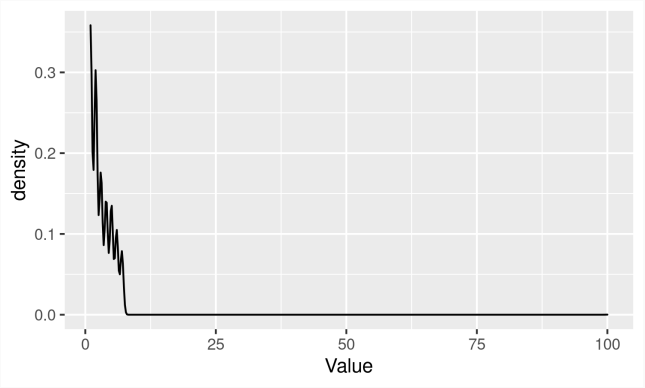
```
ggplot(data = WHERE, aes()) + creates empty plot + defaults
  geom_FUNCTION(MAPPING ARGUMENT(S)) +    adds first layer
  OTHER LAYERS +                  e.g. coordinates, facets, scales
  MORE LAYERS                     e.g. color paletts, labels, themes

ggsave("FILENAME", width= NR, height = NR)
```

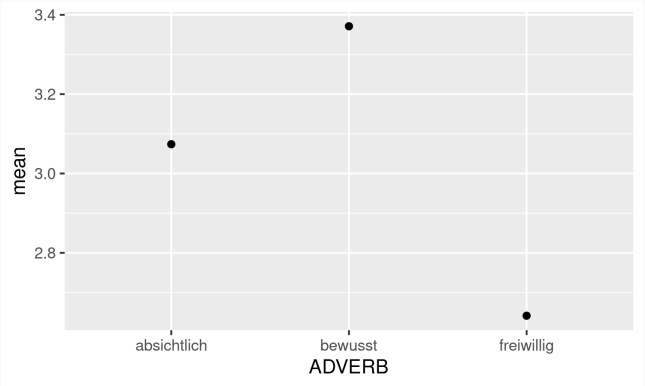
Popular geoms

Name	Usage
`geom_bar()`	Bar plots
`geom_line()`	Line plots
`geom_point()`	Dot plots
`geom_density()`	Density plots
`geom_boxplot()`	Boxplots
`geom_histogram()`	Histograms

```
ggplot(adjectives, aes(x=Value)) +
  geom_bar()
ggplot(adjectives, aes(x=Value)) +
  geom_density()
ggplot(adjectives_sum,
aes(x=ADVERB, y=mean)) +
  geom_point()
```







Aesthetic mappings

Aesthetic	Usage
x, y	values on x and y axis, respectively
fill	solid color, e.g. inside bars
colour	line colour, e.g. bar outline
size	size (e.g. of points)
shape	shape (e.g. of points)
alpha	transparency
linetype	type of line (e.g. solid, dotted, dashed)
position	*identity*, *fill*, *stack*, *jitter*, or *dodge* (syntax & options depend on geom)
stat	`"count"` for histogram, `"identity"` for existing values

```
ggplot(adjectives, aes(x=Value)) +
  geom_bar(fill="indianred1", color="#FF9999")
```

```
ggplot(adjectives, aes(x=Value)) +
  geom_density(fill="blue", color="navy")
```

```
ggplot(adjectives_sum, aes(x=ADVERB, y=mean, shape=ADVERB)) +
  geom_point(color="firebrick")
```

List of predefined color names: `https://tinyurl.com/ybf5n22w`

Plotting the Moses illusion results

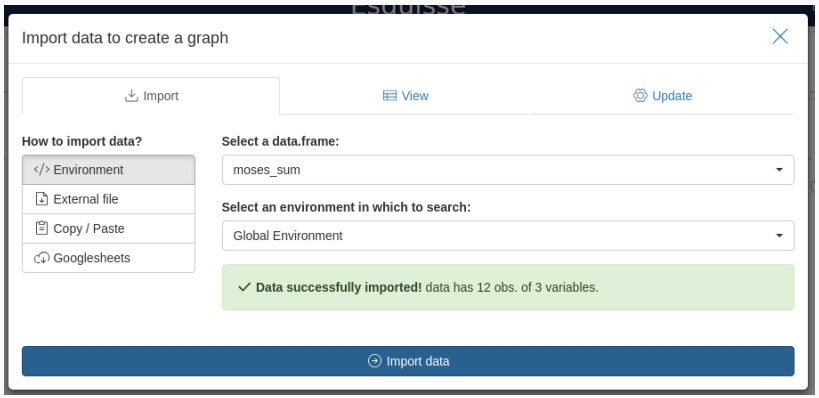
```
ggplot(moses_sum, aes(x=Accuracy, y=Frequency)) +
  geom_bar(stat="identity")

ggplot(moses_sum, aes(x=Group, y=Frequency, fill=Accuracy)) +
  geom_bar(stat="identity")

ggplot(moses_sum, aes(x=Group, y=Frequency, fill=Accuracy)) +
  geom_bar(stat="identity", position="dodge")

ggplot(moses_sum, aes(x=Group, y=Frequency, fill=Accuracy)) +
  geom_bar(stat="identity", position = "dodge") +
  scale_fill_manual(values = c("indianred1","navy","gold"))

ggplot(moses_sum, aes(x=Group, y=Frequency, fill=Accuracy)) +
  geom_bar(stat="identity", position = "dodge") +
  scale_fill_manual(values = c("indianred1","navy","god")) +
  labs(title = "Moses illusion bar plot")

ggplot(moses_sum, aes(x=Group, y=Frequency, fill=Accuracy)) +
  geom_bar(stat="identity", position = "stack") +
  scale_fill_manual(values = c("indianred1","navy","gold")) +
  labs(title = "Moses illusion bar plot")
```

Plotting the Moses illusion results

```
ggplot(moses_sum, aes(x=Group, y=Frequency, color=Accuracy)) +
  geom_point(stat="identity")

ggplot(moses_sum, aes(x=Group, y=Frequency, color=Accuracy)) +
  geom_point(stat="identity") +
  scale_color_manual(values=c("indianred1","lightseagreen","gold"))

ggplot(moses_sum, aes(x=Group, y=Frequency, shape=Accuracy)) +
  geom_point(stat="identity", size=3)

ggplot(moses_sum, aes(x=Group, y=Frequency, shape=Accuracy)) +
  geom_point(stat="identity", size=3, position=position_dodge(0.5)) +
  labs(title = "Moses illusion", subtitle = "May 2022")

ggplot(moses_sum, aes(x=Group, y=Frequency, shape=Accuracy)) +
  geom_point(stat="identity", size=3, position=position_dodge(0.5)) +
  labs(title = "Moses illusion", subtitle = "May 2022")
  theme_bw()
```

A shortcut: `esquisse`

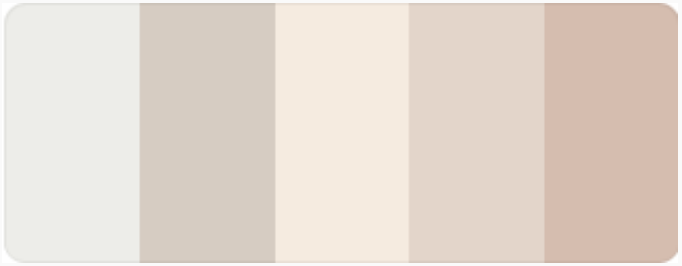
1. Install and load the package **`esquisse`**.
2. Launch the app via the function **`esquisser()`**.
3. Choose the dataset (your own or others).
4. Create, edit, and save your plots.



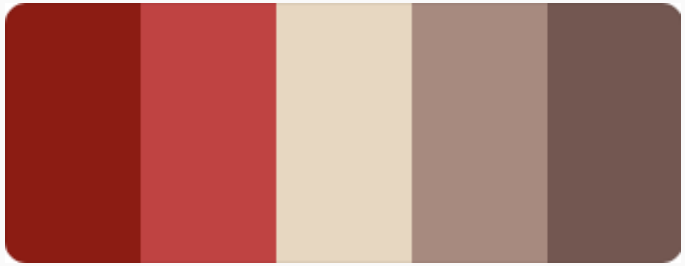
? https://tinyurl.com/sr77r3au

Choosing colors

Pick colors that are easily distinguishable



Too similar



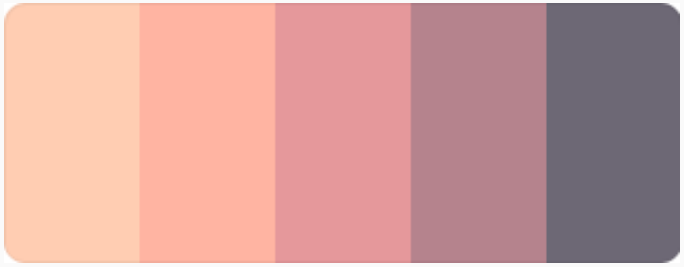
Different

Less is more

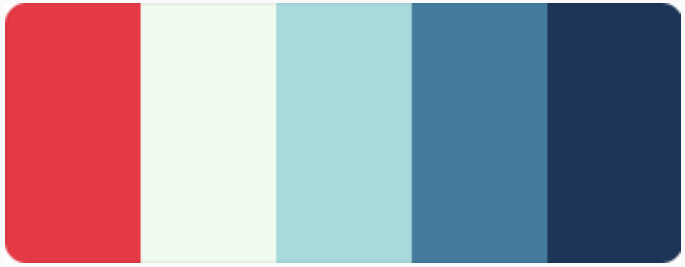
Chaotic

Peaceful

Use color to make important information stand out

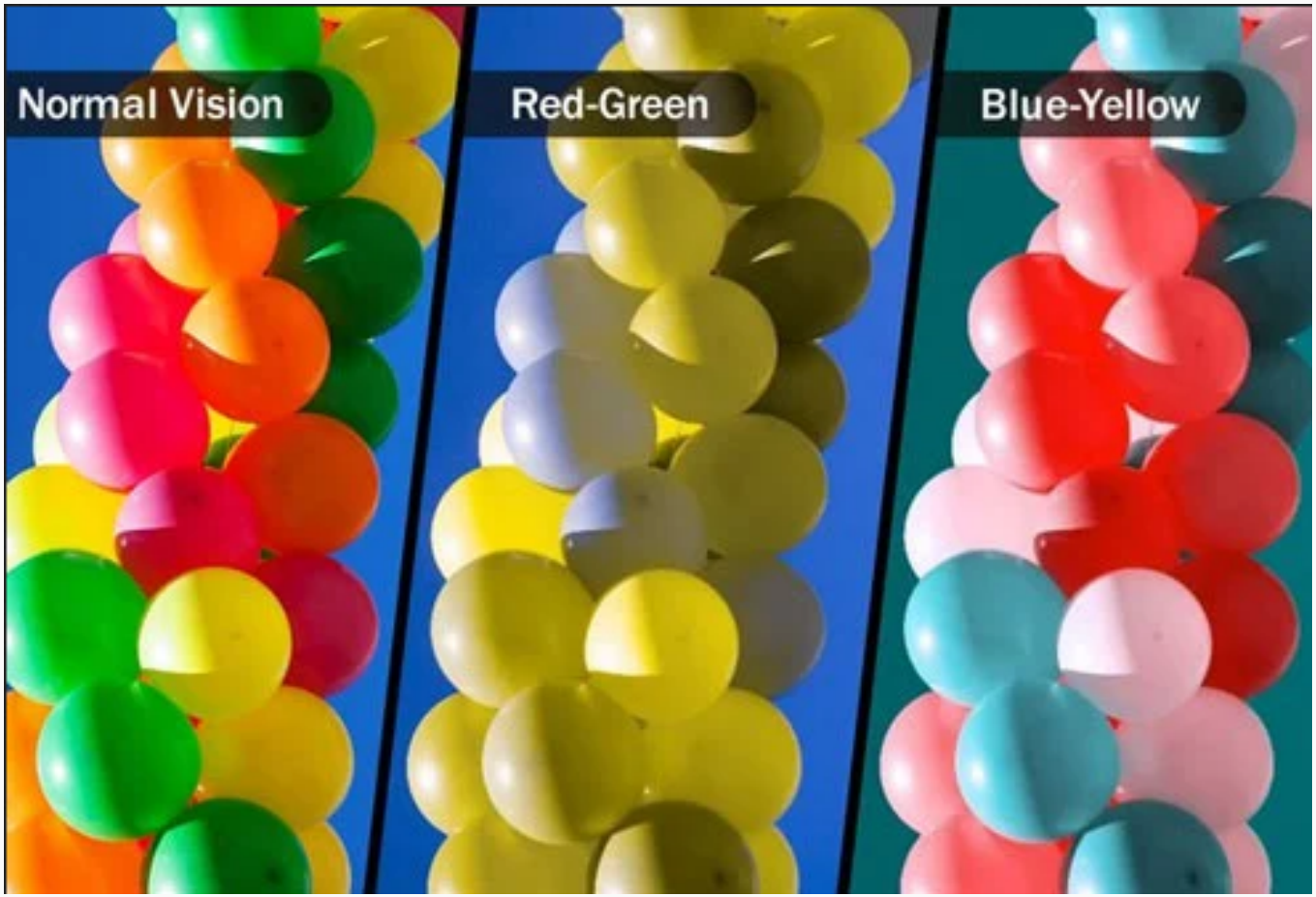


Everything is fine



Watch out!

Be inclusive

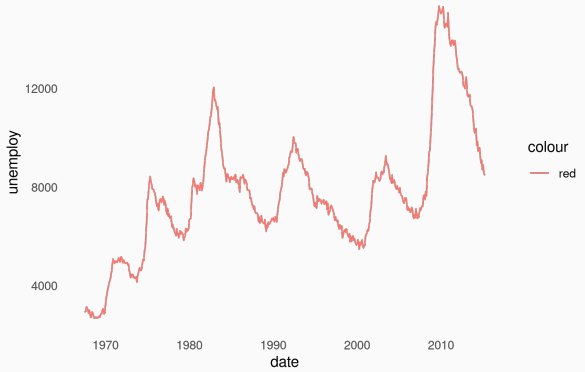


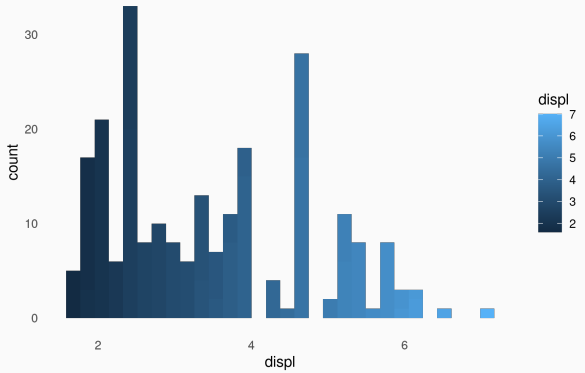
Source: `https://www.emedicinehealth.com`

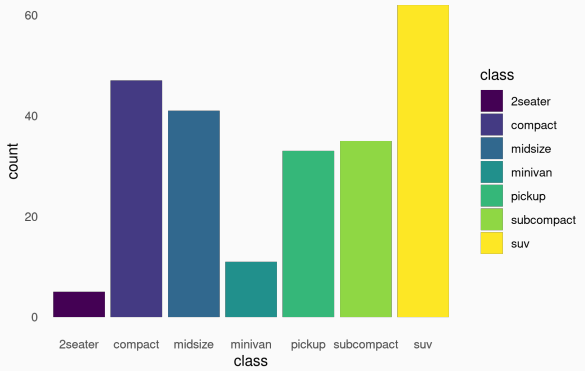
Color me skeptical

Single color for continuous data
Sequential colors for grouping
Divergent colors for comparison

You *can* procrastinate and make your own color palettes, you can use generators (e.g. `https://learnui.design/tools/data-color-picker.html`, `https://color.adobe.com`, `https://coolors.co`) or use preexisting ones (`ggplot2` or `RColorBrewer`)







ggplot2 and data viz resources

ggplot2-specific:

- http://www.cookbook-r.com
- https://ggplot2.tidyverse.org/
- https://ggplot2-book.org/
- https://r-graph-gallery.com/
- https://raw.githubusercontent.com/rstudio/cheatsheets/main/data-visualization.pdf

General data visualization

- https://blog.datawrapper.de/beautifulcolors/
- https://venngage.com/

Generative art

- https://koenderks.github.io/aRtsy/

Best practices


I love you

Bad examples

Cherrypicking

realdonaldtrump • Follow

realdonaldtrump Thank you very much, working hard!

Load more comments

riot_racer @figboot31977 I know your a Donald trump supporter. Kind of obvious

figboot31977 @riot_racer Here's a thought for ya.... You don't know crap!!! GET a LIFE and STAY the #@&# OUT OF MINE!!!

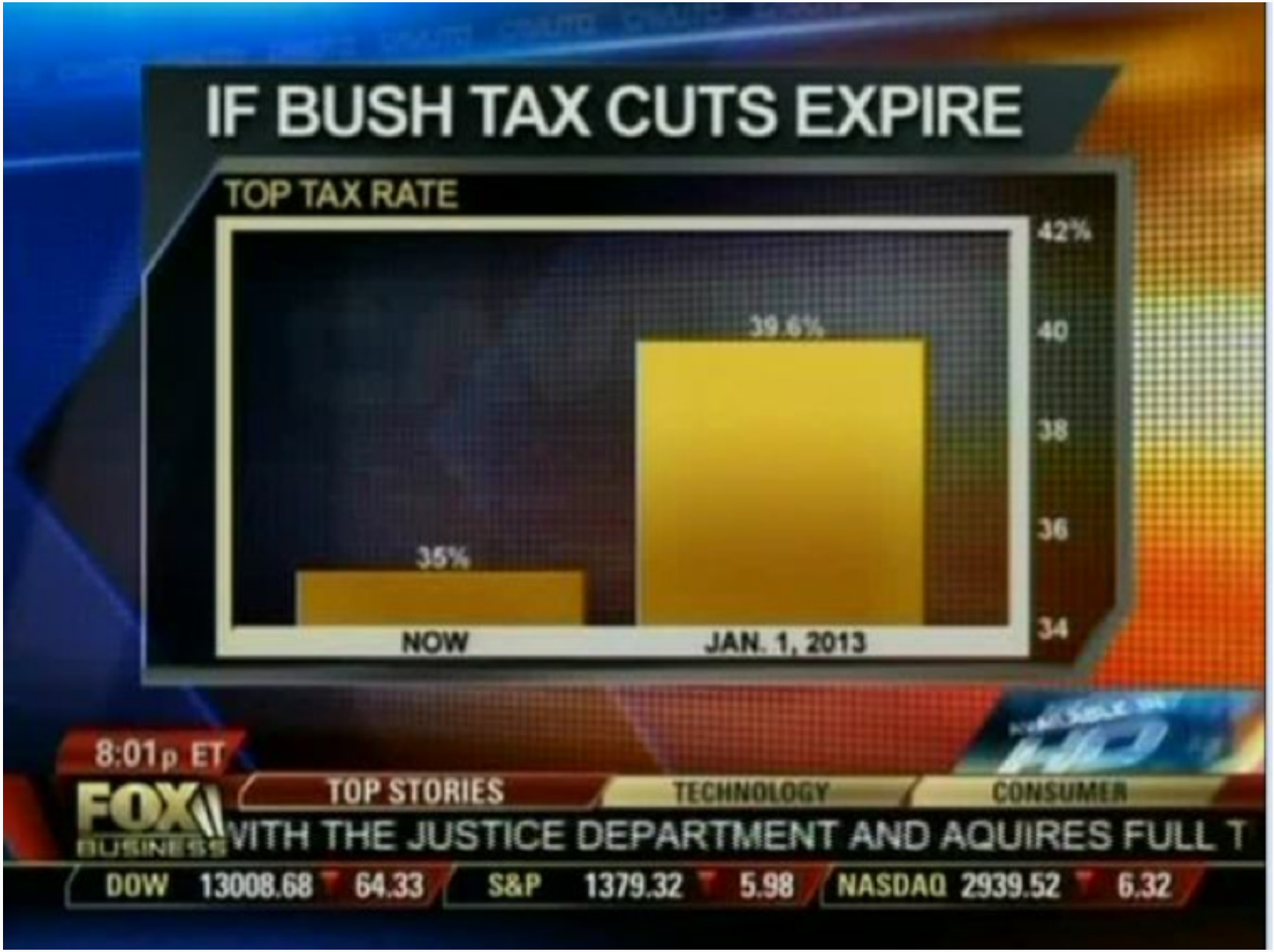
riot_racer @figboot31977 I'm not in yours. If I was then I'd have a bigger understanding of who you are. Like I said if you support

625,072 likes

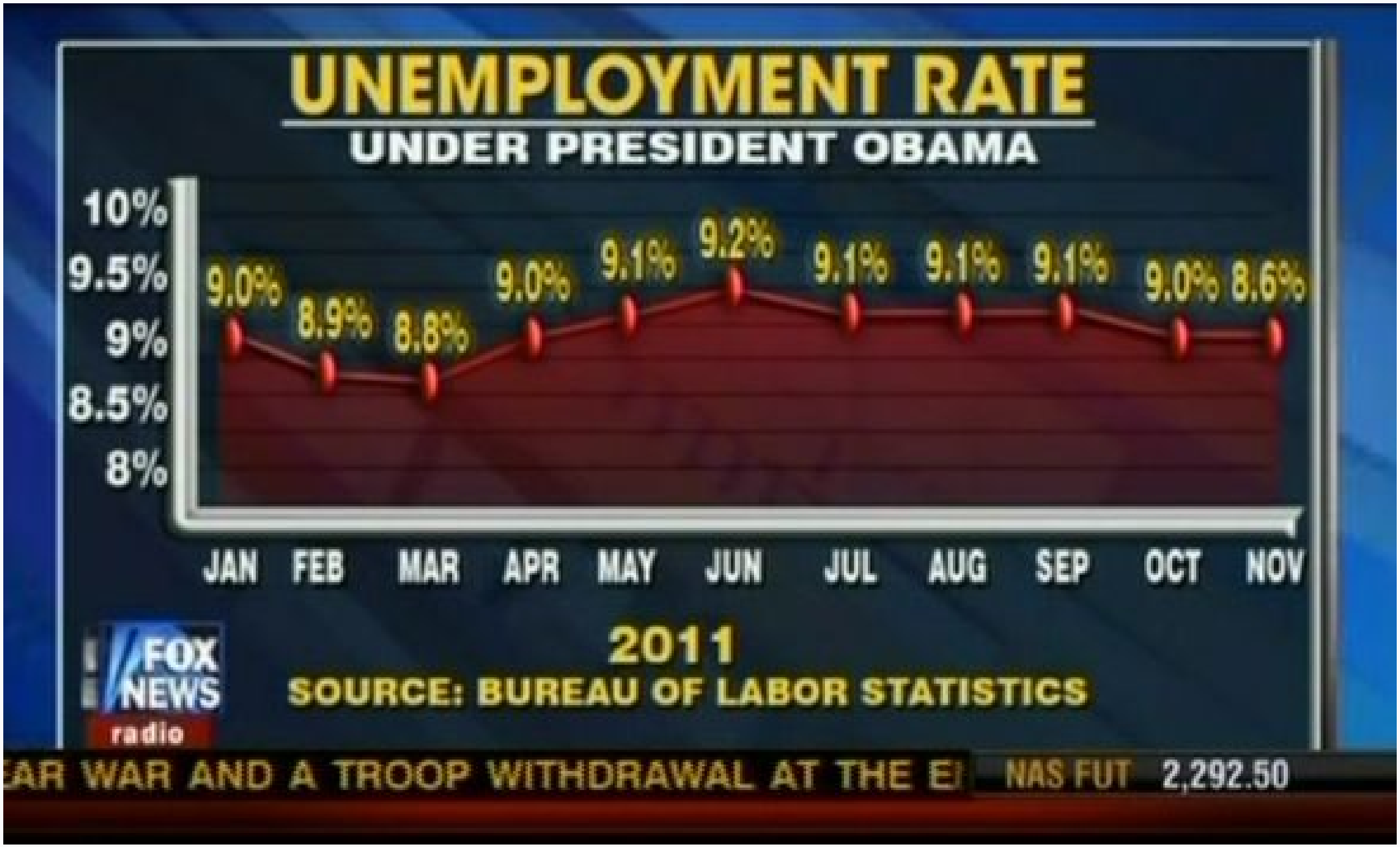
JULY 26

Log in to like or comment.

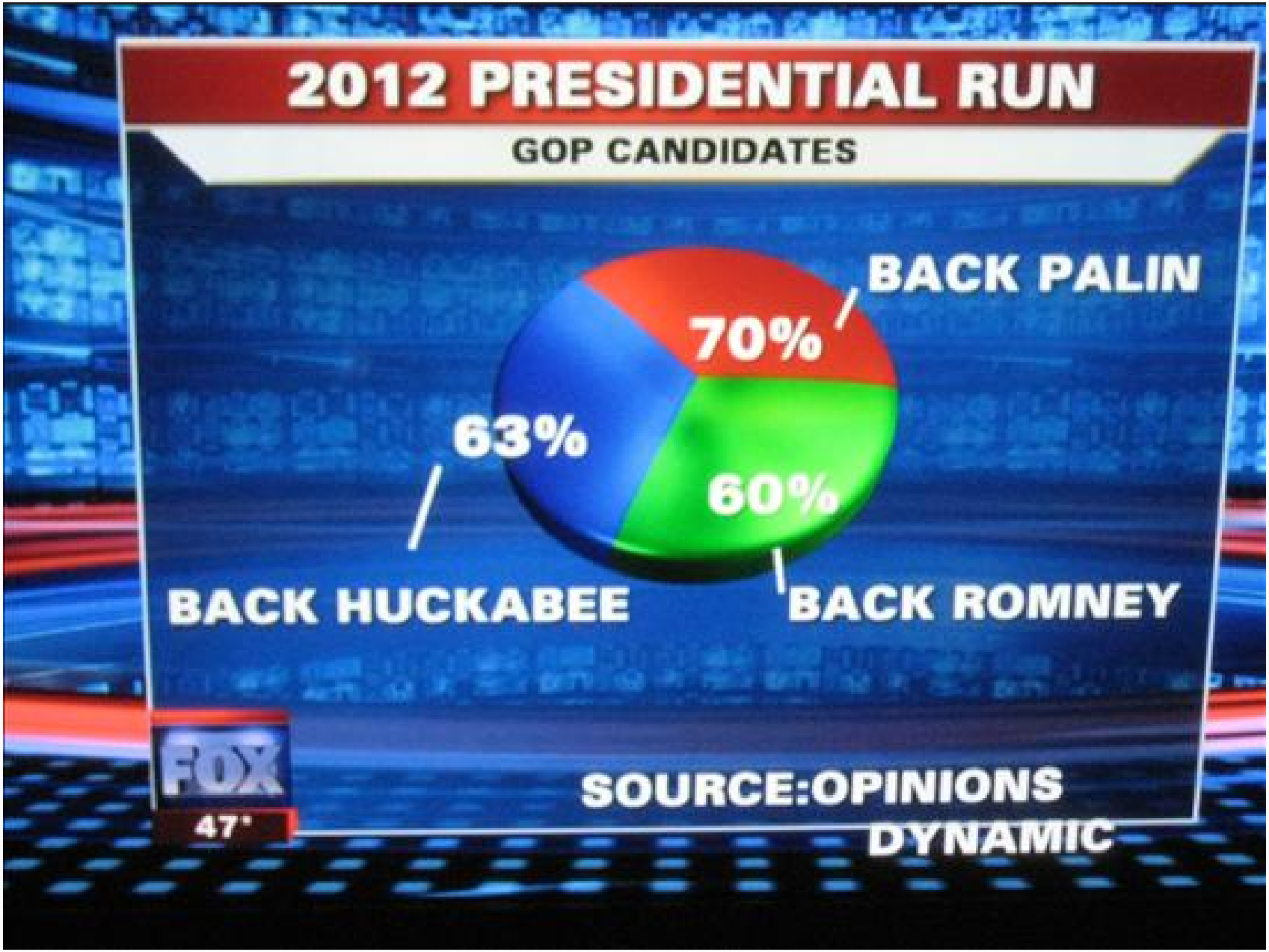
Omitting the baseline



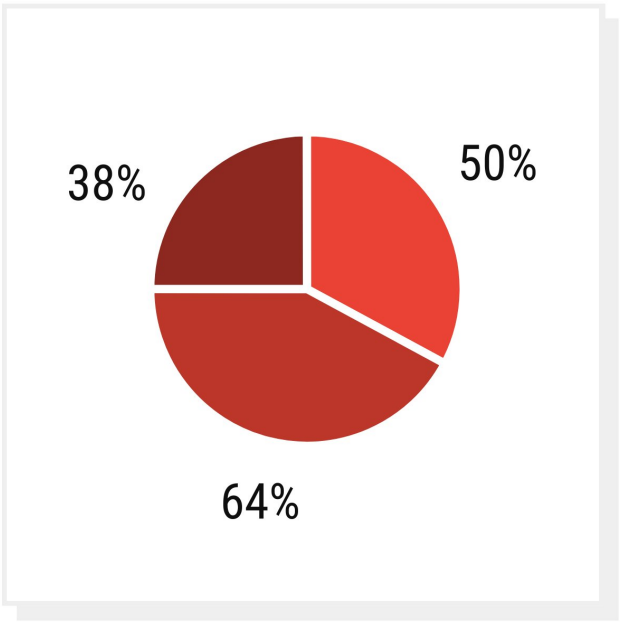
Mainpulating the y axis

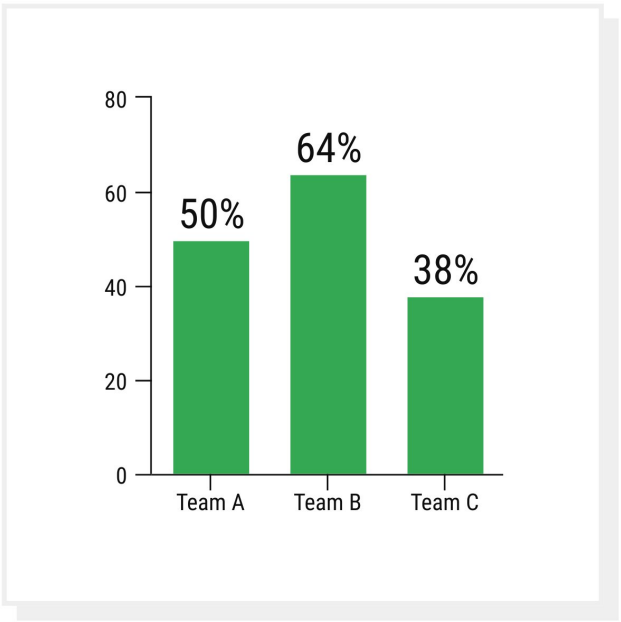


Using the wrong graph

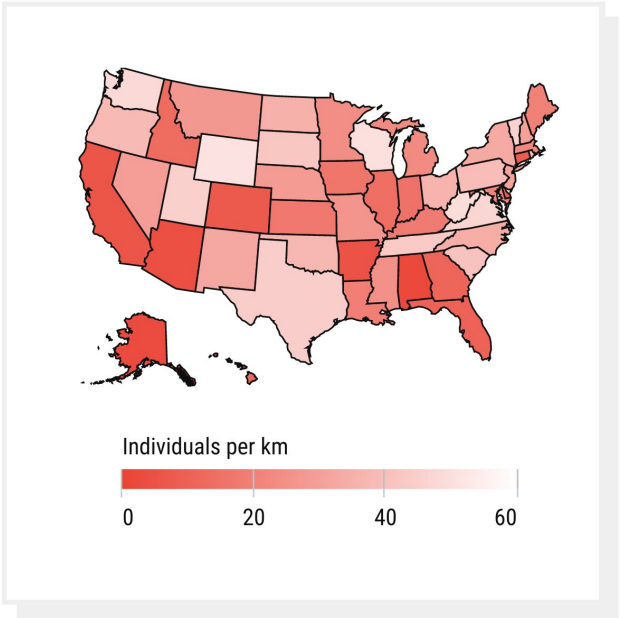


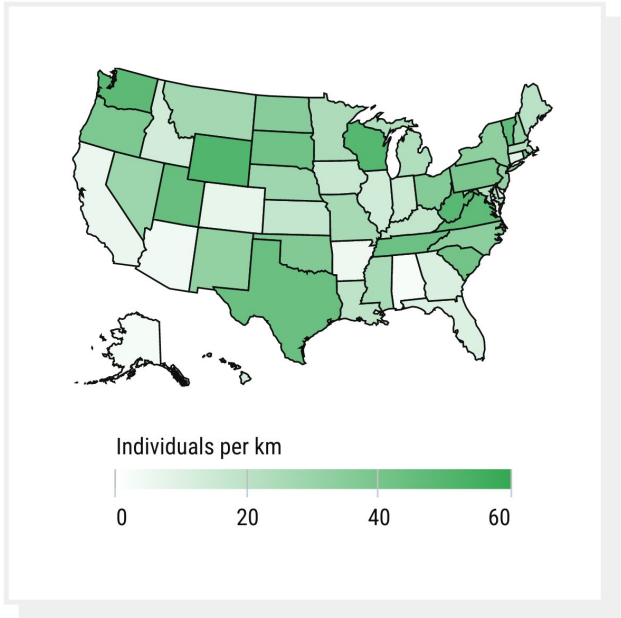
Using the wrong graph





Going against conventions





Source: `https://venngage.com`

Questions?

Homework assignment

Homework assignment due May 23

If modifying raw data, include the code to create the data frames you are using!

Create three different plot types for the `adjectives` data set (raw, cleaned, or summarized). Customize the colors and add a title.

Using the `esquisse` package, create another three plots from one of the `ggplot2` data sets (e.g. diamonds, economics, mpg). Change the legend position and add labels to your plot.

Read chapters 26-30 of R for Data Science
`https://r4ds.had.co.nz/`

Install the packages `knitr` and `rmarkdown`