Slow ggplot2

Evangeline Reynolds

2018-09-29

Contents

1 Introduction 5

2 Baseball, WAR, and Ethnicity 7

3 Christmas Trees 9

4 Officials' beliefs about women's representation 13

5 Maternal Leave 15

6 Traits 17

7 Salarys of Trump and Obama White House Employees 19

8 Winter Games 21

9 Brexit 25

10 Curry in London 29

11 Life Expectancy Increases 33

12 Myers Briggs 35

13 Wine 37

14 Arctic Ice 39

Chapter 1

Introduction

In late 2016 I heard Andy Kreibel (and ___) being interviewed on the Podcast “Data Stories”. He was invited onto the show because he had started the initiative #MakeoverMonday, in which he would find the data for an existing data visualization that he had come across in the media, and would recreate the visualization using Tableau, a tool that they were both expert in, but found that they weren’t practicing as regularly due to the grind of administrative jobs. The two friends shared their makeovers products with each other, but also with the world on Twitter. Soon, more people expressed interest in joining, and they two started a more organized initiative - posting the original graph and data every Sunday, so that whoever wanted to could participate in #Makeovermonday.

My first submission was late 2016, after catching wind of the exciting project via the podcast. I made a scrappy little graph about motorway casualties; sad topic, but fun graph making.

I was using base R at that time. Then in the summer of 2017 I went to a conference in Zurich, the women’s summer school for political methodology. There was a session on ggplot2. I internalized some of the basics, and decided that if I wanted to learn that (powerful - as everyone kept calling it) graphing system, then I could do it via the #MakeoverMonday weekly exercises (not that I particpated weekly). Even though most folks were using Tablaeu, the administrators didn’t seem to mind a few R and ggplot submission here and there. I got a little hooked.

Early this year Andy and Eva Murry sent a number of the participants a private message on Twitter. “We’re writing a book: #MakeoverMonday”. They were putting together a collection of a visualizations that resulted from the project, and were seeking perspectives of participants as well as permission to use some of the visualizations produced for the initiatives. Cool. I was pleased to participate. For me #MakeoverMonday allowed me to focus on the visualization task. Usually visualization comes at the end of, sometimes arguous, data cleaning — and you might already be a little spent. Having rather clean data delivered, and seeing the approaches of many other (many brilliant) data visualizers was a treat. I still need to buy my copy of the book, which contains a visualization of food prices in London as a function of how far a restaurant is from the Big Ben.

And now, using the magic of RStudio and Yihui Xie’s bookdown, I’m putting together my own little collection. Of course there is a bit of curation involved — I’m not including every plot. And, I’m revising the exact code that creates the plots in many cases, to be more consistent across plots, and also, I think, to make communicating about how the plot was built easy. This involves:

- using fewer functions (labs(title =) instead of ggtitle())
- pulling out aes() from the ggplot() function
- multiply using functions aes(x = ?) + aes(y = ?) rather than aes(x = , y =)
- using base R functions and tidyverse functions. For other packages, the :: style to call them
- write out arguments (no shortcuts) aes(x = gdppercap) not aes(gdppercap)
- order ggplot commands so that reactivity is obvious

Chapter 2

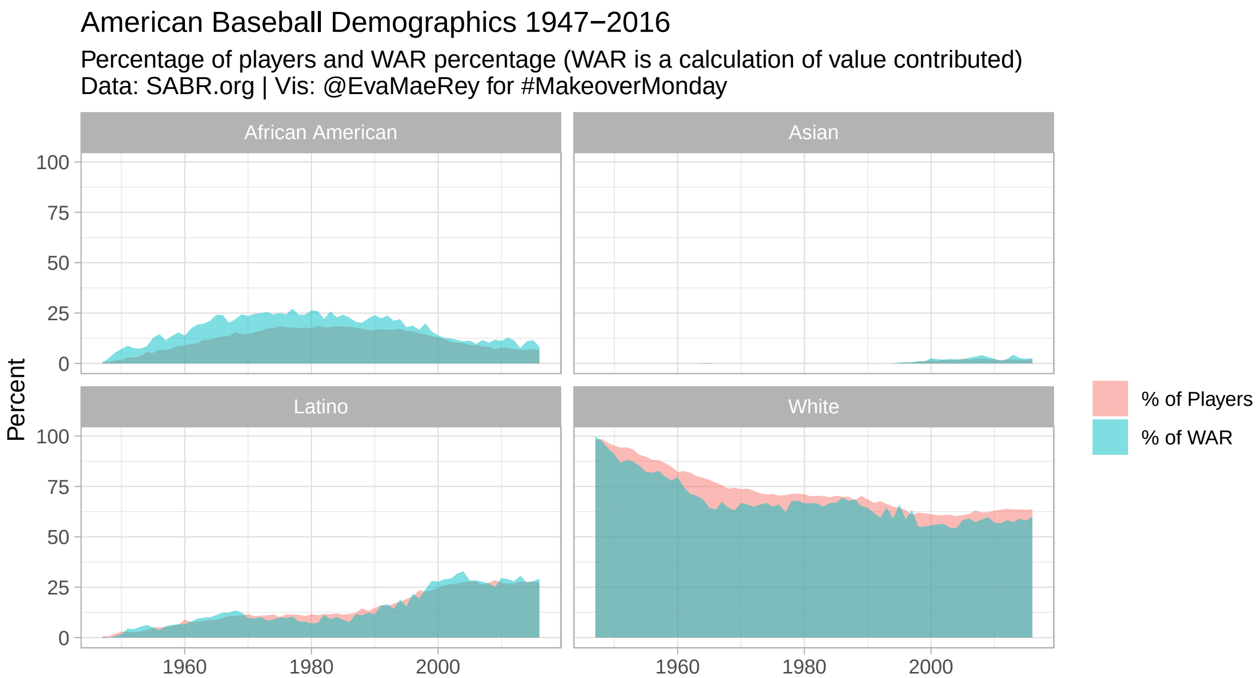
Baseball, WAR, and Ethnicity

This data visualization uses the WAR measure in baseball, a calculation based on the contributions of players. The visualizations show that new ethnicities and races started to be included in Major League baseball, the minority players that joined tended to contribute more than the expected value for players overall. For example, from 1947, when Jackie Robinson joined Major League baseball, and onward, the percent of African American players was outpaced by the percent calculated contributions (WAR) of African American players.

A random sample from the data set:

Year	Ethnicity	type	Percent
1983	African American	% of WAR	25.8
1949	African American	% of Players	1.5
1950	Latino	% of WAR	1.7
1967	Latino	% of Players	10.7
1978	Asian	% of WAR	0.0

```
ggplot(df_gather) +
  aes(x = Year) +
  aes(y = Percent) +
  aes(fill = type) +
  facet_wrap(~ Ethnicity) +
  geom_area(alpha = .5, position = "dodge") +
  labs(fill = "") +
  labs(x = "") +
  labs(title = "American Baseball Demographics 1947-2016") +
  labs(subtitle = "Percentage of players and WAR percentage (WAR is a calculation of value contributed)"
  theme_light()
```


American Baseball Demographics 1947−2016
Percentage of players and WAR percentage (WAR is a calculation of value contributed)
Data: SABR.org | Vis: @EvaMaeRey for #MakeoverMonday
African American
Asian
Latino
White
Percent
100
75
50
25
0
1960
1980
2000
% of Players
% of WAR

Chapter 3

Christmas Trees

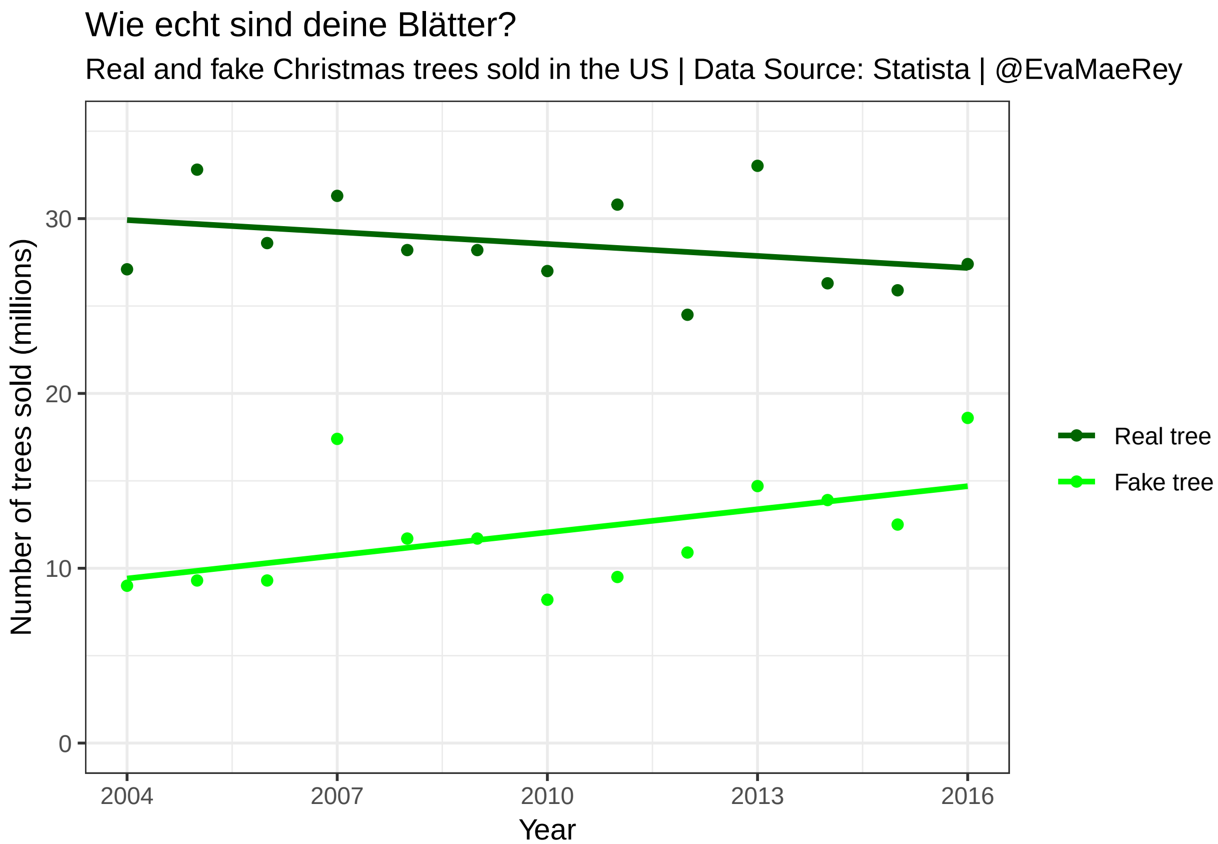
Here is a simple plot of Christmas Tree Sales in the U.S. The plot shows that artificial tree sales are on the rise, contrasting with declines in real trees. The title plays on the German Christmas Carol "O Tannenbaum", "Oh Christmas Tree" in English. "Wie echt sind deine Blätter?" means "how real are your leaves"; the original text from the carol is "Wie treu sind deine Blätter!" which means "How true your leaves are!"

I also plot the cumulative number of trees purchased of each type, artificial and real, from 2004 to 2014, comparing that to the 2016 U.S. population. Almost one real tree per person was bought over the course of 10 years!

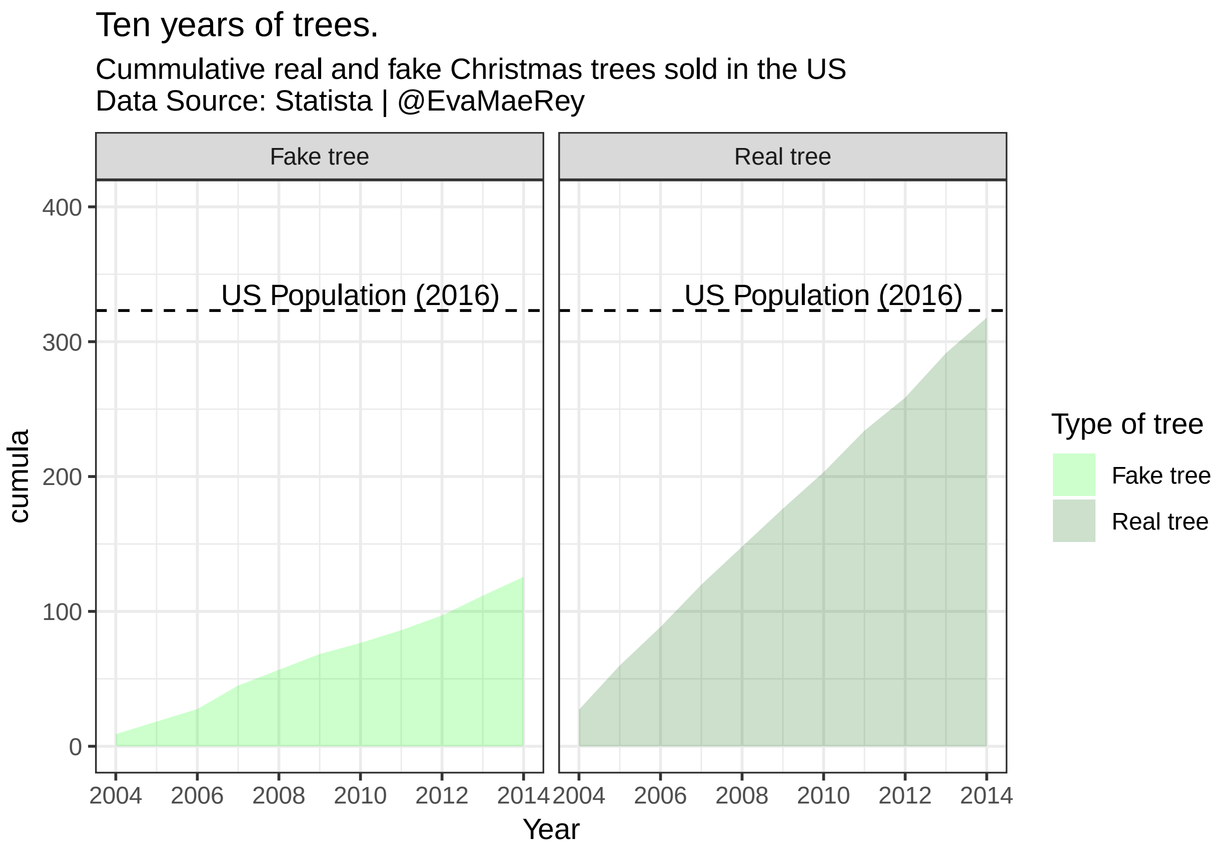
A random sample from the data set:

Year	Number of trees sold	Type of tree	Number of trees sold (millions)
2008	11700000	Fake tree	11.7
2011	9500000	Fake tree	9.5
2012	10900000	Fake tree	10.9
2015	25900000	Real tree	25.9
2006	9300000	Fake tree	9.3

```
ggplot(data = dta) +
  aes(Year) +
  aes(y = `Number of trees sold (millions)`) +
  geom_point() +
  aes(col = fct_rev(`Type of tree`)) +
  geom_smooth(method = "lm", se = F) +
  scale_color_manual(values = c("darkgreen", "green")) +
  ylim(c(0, 35)) +
  labs(col = "") +
  labs(title = "Wie echt sind deine Blätter?") +
  labs(subtitle = "Real and fake Christmas trees sold in the US | Data Source: Statista | @EvaMaeRey ")
  theme_bw()
```




```
dta <- dta %>%
  group_by(`Type of tree`) %>%
  mutate(cumula = cumsum(`Number of trees sold (millions)`))

ggplot(dta %>% filter(Year <= 2014)) +
  aes(Year) +
  aes(y = cumula) +
  aes(fill = `Type of tree`) +
  geom_hline(yintercept = 323.1, lty = 2) +
  geom_area(alpha = .2) + facet_wrap(~ `Type of tree`) +
  annotate(geom = "text", x = 2010, y = 335, label = "US Population (2016)") +
  labs(title = "Ten years of trees.") +
  labs(subtitle = "Cummulative real and fake Christmas trees sold in the US\nData Source: Statista | @E
  scale_fill_manual(values = c("green", "darkgreen")) +
  theme_bw() +
  ylim(c(0, 400))
```


Ten years of trees.
Cummulative real and fake Christmas trees sold in the US
Data Source: Statista | @EvaMaeRey
Fake tree
Real tree
US Population (2016)
US Population (2016)
400
300
200
100
0
cumula
2004
2006
2008
2010
2012
2014
Year
Type of tree
Fake tree
Real tree

Chapter 4

Officials' beliefs about women's representation

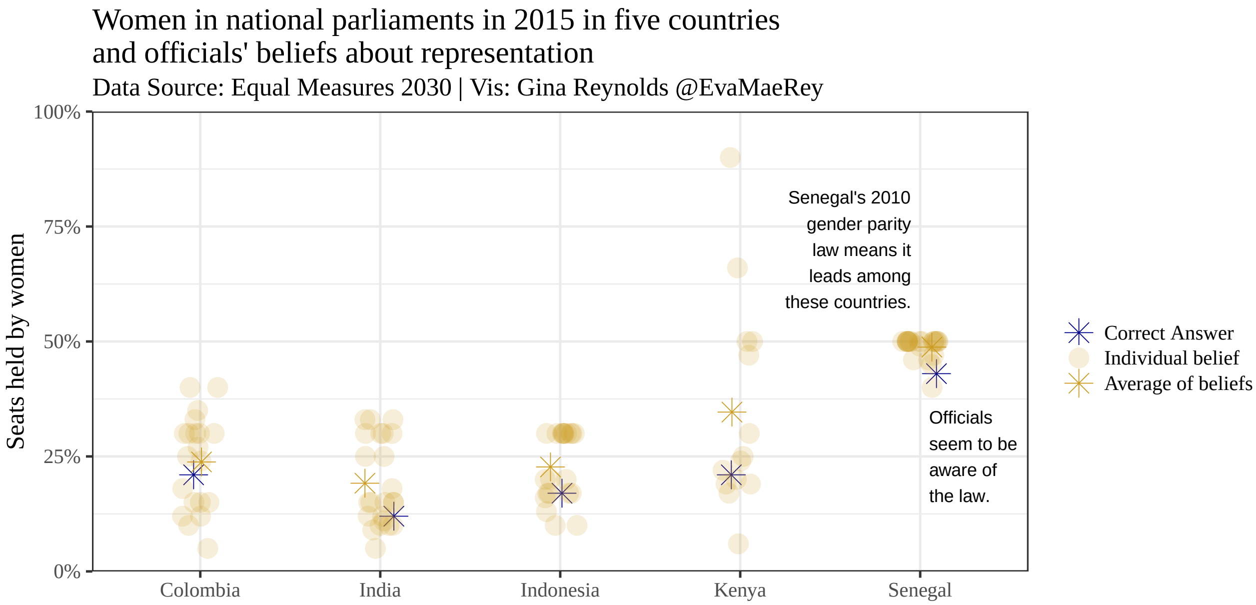
The data provided is based on a small survey of elite officials in five less developed countries. The question that arrises from the data is: How well do elites know the conditions in their countries. In general, the elites overestimate women's representation. But this is not the case in Senegal, where there are gender quotas in the Parliament. Most elites therefore estimate that the representation is about equal with men. I jitter the responses of the elites horizontally to avoid overplotting.

A random sample from the data set:

Country	Topic	value	value_type	alpha
Senegal	Share of seats held by women	0.50	Individual belief	0.3
Kenya	Share of seats held by women	0.20	Individual belief	0.3
Senegal	Share of seats held by women	0.50	Individual belief	0.3
Colombia	Share of seats held by women	0.30	Individual belief	0.3
Indonesia	Share of seats held by women	0.17	Individual belief	0.3

```
ggplot(data = df_all) +
  aes(x = Country) +
  aes(y = value) +
  aes(col = fct_inorder(value_type)) +
  aes(alpha = fct_inorder(value_type)) +
  aes(shape = fct_inorder(value_type)) +
  geom_jitter(width = .1, height = 0, size = 7) +
  geom_hline(yintercept = c(0, 100), col = "grey") +
  geom_hline(yintercept = c(50), lty = 2, col = "grey") +
  theme_bw(base_size = 20, base_family = "Times") +
  scale_y_continuous(limits = c(0, 1), expand = c(0, 0), labels = scales::percent) +
  # create just one key "" mixing aesthetics
  scale_colour_manual(name = "", values = c("darkblue", "goldenrod3", "goldenrod3")) +
  scale_alpha_manual(name = "", values = c(1, .17, 1)) +
  scale_shape_manual(name = "", values = c(8, 19, 8)) +
  annotate(geom = "text", x = 4.95, y = .70,
           label = str_wrap("Senegal's 2010 gender parity law means it leads among these countries.", 1
  annotate(geom = "text", x = 5.05, y = .250,
           label = str_wrap("Officials seem to be aware of the law.", 10), size = 5, hjust = 0) +
  labs(x = "") +
  labs(y = "Seats held by women") +
  labs(title = "Women in national parliaments in 2015 in five countries \nand officials' beliefs about
```

```
labs(subtitle = "Data Source: Equal Measures 2030 | Vis: Gina Reynolds @EvaMaeRey")
```



Chapter 5

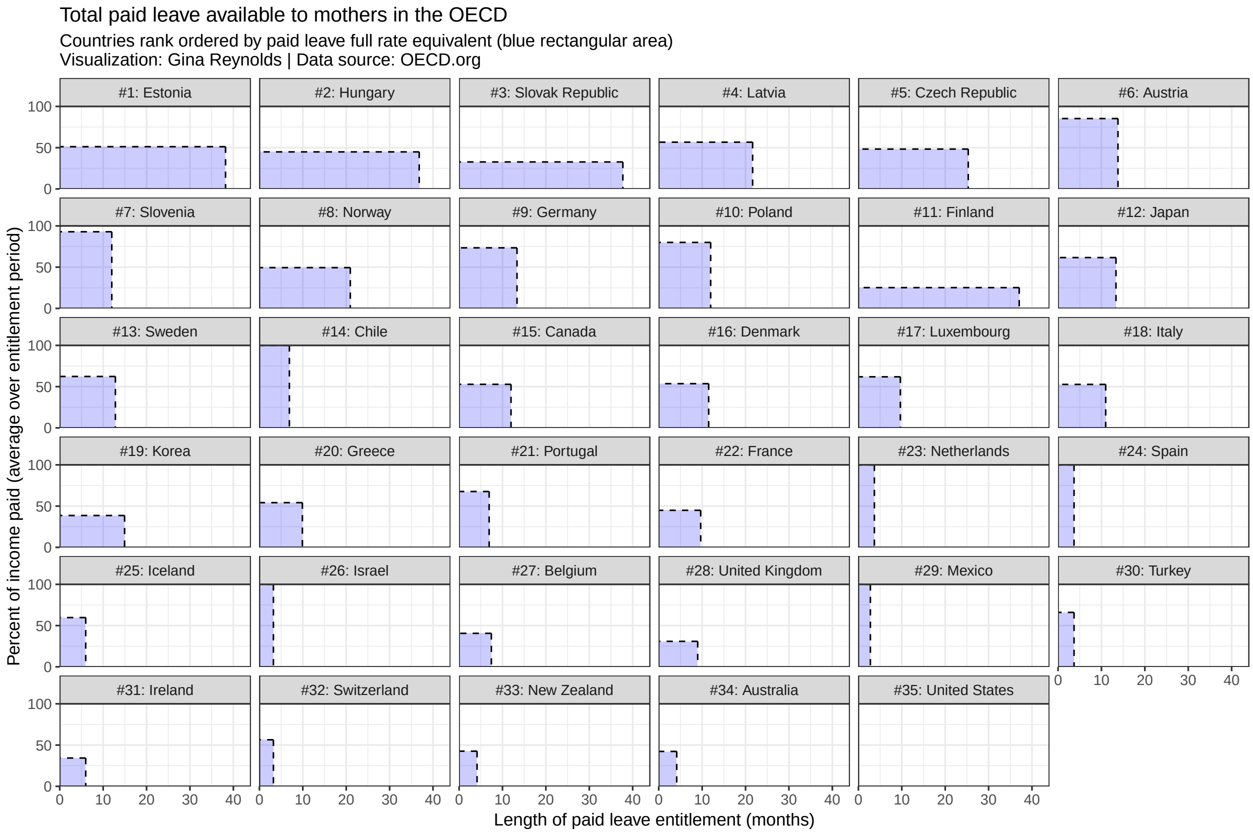
Maternal Leave

The OECD provides a comparative report on how much paid leave women are entitled to after childbirth. But leave takes different forms. In some places, the allowed leave is longer, but sometimes that means that the pay out compared to the regular salary is lower. To emphasize the different forms that law around paid leave take, I plotted the total payout available to mothers as areas of rectangles, where one side is the length of leave allowed, and the other side is the proportion of salary paid to the new mom.

A random sample from the data set:

Country	Paid maternity leave avg payment rate (%)	Paid maternity leave full rate equivalent in weeks	Paid
New Zealand	42.6	7.7	
Slovak Republic	70.0	23.8	
Slovenia	100.0	15.0	
Ireland	34.3	8.9	
Japan	67.0	9.4	

```
ggplot(df) +
  aes(x = paid_leave_months) +
  aes(y = `Total paid leave avg payment rate (%)`) +
  aes(xmin = 0) +
  aes(xmax = paid_leave_months) +
  aes(ymin = 0) +
  aes(ymax = `Total paid leave avg payment rate (%)`) +
  facet_wrap(fct_inorder(rank_name) ~ .) +
  geom_rect(fill = "blue", alpha = .2) +
  aes(yend = 0) +
  aes(xend = 0) +
  geom_segment(aes(yend = `Total paid leave avg payment rate (%)`), lty = "dashed") +
  geom_segment(aes(xend = paid_leave_months), lty = "dashed") +
  scale_y_continuous(limits = c(0, 100), expand = c(0, 0), breaks = c(0, 50, 100)) +
  scale_x_continuous(limits = c(0, 44), expand = c(0, 0)) +
  labs(x = "Length of paid leave entitlement (months)") +
  labs(y = "Percent of income paid (average over entitlement period)") +
  labs(title = "Total paid leave available to mothers in the OECD") +
  labs(subtitle = "Countries rank ordered by paid leave full rate equivalent (blue rectangular area)\nVi
  theme_bw(base_size = 12)
```


Total paid leave available to mothers in the OECD
Countries rank ordered by paid leave full rate equivalent (blue rectangular area)
Visualization: Gina Reynolds | Data source: OECD.org
#1: Estonia
#2: Hungary
#3: Slovak Republic
#4: Latvia
#5: Czech Republic
#6: Austria
#7: Slovenia
#8: Norway
#9: Germany
#10: Poland
#11: Finland
#12: Japan
#13: Sweden
#14: Chile
#15: Canada
#16: Denmark
#17: Luxembourg
#18: Italy
#19: Korea
#20: Greece
#21: Portugal
#22: France
#23: Netherlands
#24: Spain
#25: Iceland
#26: Israel
#27: Belgium
#28: United Kingdom
#29: Mexico
#30: Turkey
#31: Ireland
#32: Switzerland
#33: New Zealand
#34: Australia
#35: United States
Percent of income paid (average over entitlement period)
Length of paid leave entitlement (months)
0
50
100
10
20
30
40

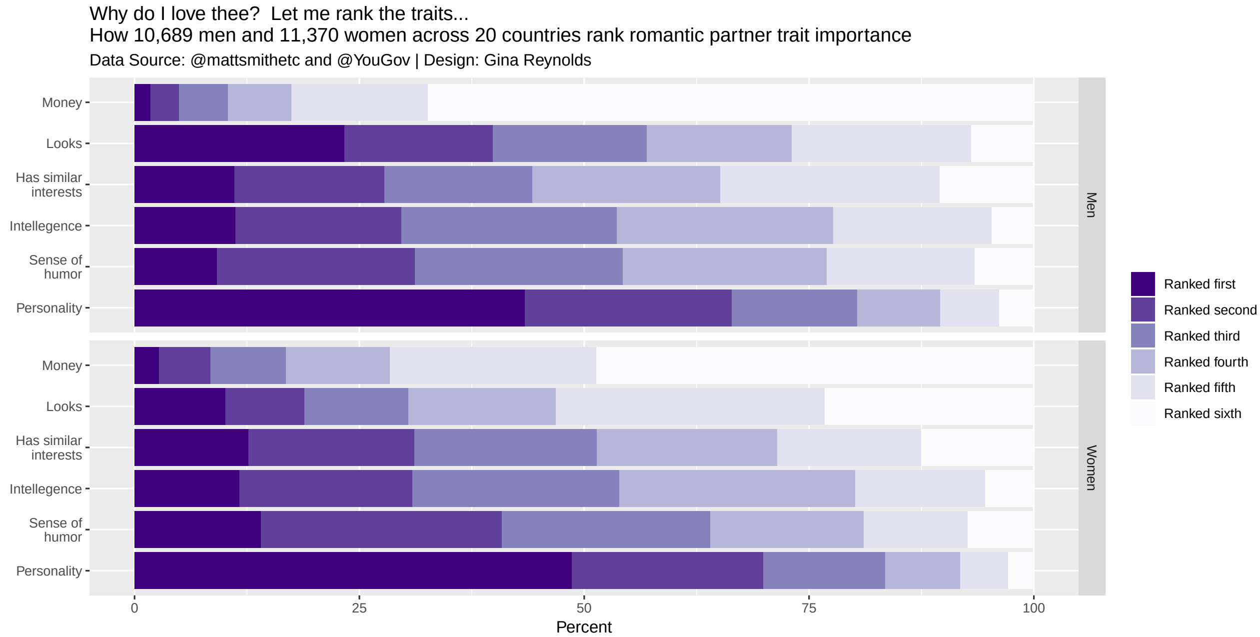
Chapter 6

Traits

A random sample from the data set:

Gender	Question_short	Rank (text)	Rank (number)	n	Percent
Men	Money	Ranked first	1	190.78	1.787746
Men	Money	Ranked sixth	6	7190.57	67.380809
Men	Looks	Ranked fifth	5	2134.35	19.967724
Men	Money	Ranked fifth	5	1617.37	15.155919
Women	Personality	Ranked fifth	5	608.01	5.331190

```
ggplot(data = world) +
  aes(x = Question_short_wrap) +
  aes(y = Percent) +
  aes(fill = `Rank (text)`) +
  facet_grid(Gender ~ .) +
  geom_col() +
  coord_flip() +
  scale_fill_manual(
    values = colorRampPalette(RColorBrewer::brewer.pal(9, "Purples"))(6)[1:6],
    guide = guide_legend(reverse = TRUE)
  ) +
  labs(fill = "") +
  xlab("") +
  labs(title = "Why do I love thee?  Let me rank the traits... \nHow 10,689 men and 11,370 women across
  labs(subtitle = "Data Source: @mattsmithetc and @YouGov | Design: Gina Reynolds")
```


Why do I love thee? Let me rank the traits...
How 10,689 men and 11,370 women across 20 countries rank romantic partner trait importance
Data Source: @mattsmithetc and @YouGov | Design: Gina Reynolds
Money
Looks
Has similar interests
Intellegence
Sense of humor
Personality
Men
Women
0
25
50
75
100
Percent
Ranked first
Ranked second
Ranked third
Ranked fourth
Ranked fifth
Ranked sixth

Chapter 7

Salarys of Trump and Obama White House Employees

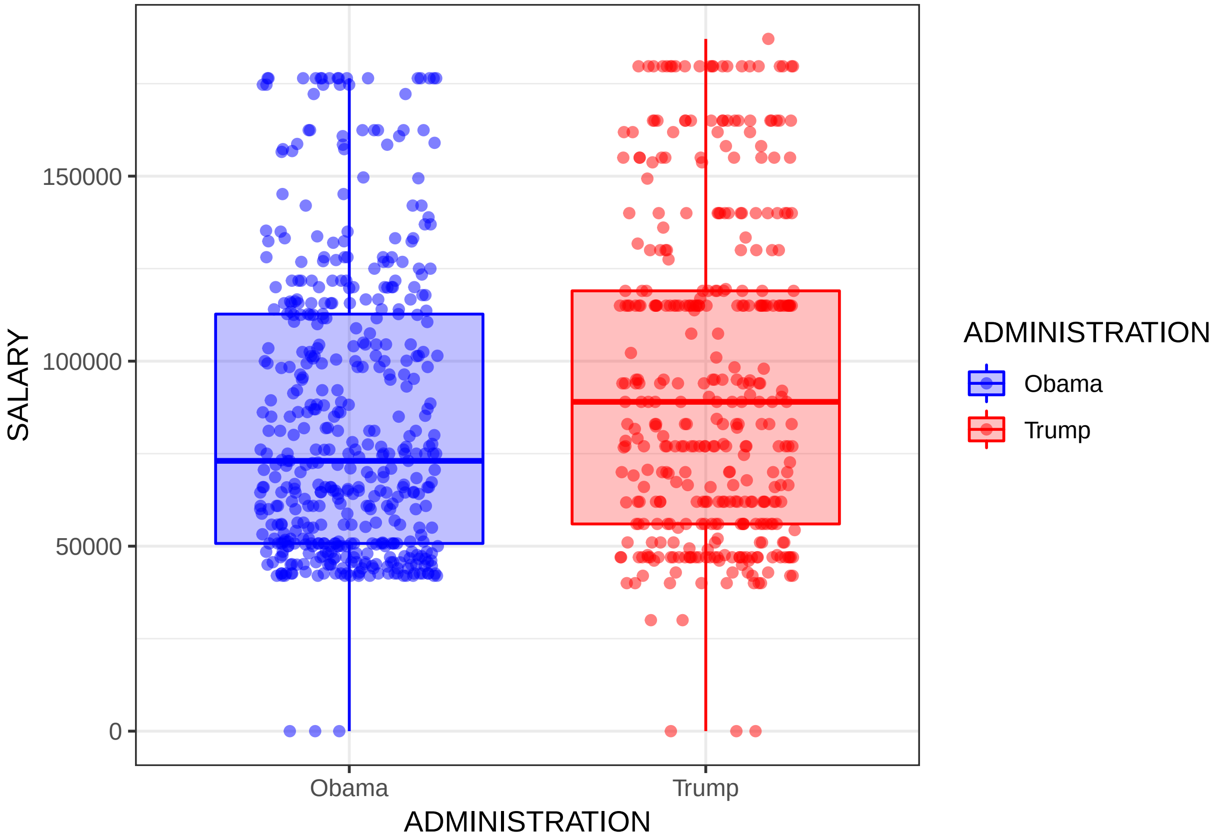
The data set, originally reported on in an NPR article, shows the difference in the distribution of salaries for the Obama and early Trump White House.

First I plot a histogram of each administration. Then I also contrast boxplots for each administration; the data points are overlayed, jittered to the widths of the boxplots. Plotly is used to make the graph interactive; mousing over will allow you to see who the point represents, their job description and exactly how much they are paid.

A random sample from the data set:

ADMINISTRATION	NAME	STATUS	SALARY	PAY BASIS	POSITION TITLE
Obama	Hoover, Zealan T.	Employee	72000	Per Annum	POLICY ADVISOR TO THE SE
Trump	Lai, Joseph G.	Employee	115000	Per Annum	SPECIAL ASSISTANT TO THE
Obama	Dickason, Christine N.	Employee	42000	Per Annum	DEPUTY ASSOCIATE DIRECT
Obama	Larimer, Becky S.	Employee	68658	Per Annum	CALLIGRAPHER
Trump	Salem, Hannah H.	Employee	83000	Per Annum	SENIOR LEAD PRESS REPRES

```
ggplot(both_data) +
  aes(x = ADMINISTRATION) +
  aes(y = SALARY) +
  geom_jitter(alpha = .5, height = 0, width = .25) +
  aes(col = ADMINISTRATION) +
  geom_boxplot(alpha = .25) +
  aes(fill = ADMINISTRATION) +
  scale_colour_manual(values = c("blue", "red")) +
  scale_fill_manual(values = c("blue", "red")) +
  theme_bw()
```


SALARY
150000
100000
50000
0
Obama
Trump
ADMINISTRATION
ADMINISTRATION
Obama
Trump

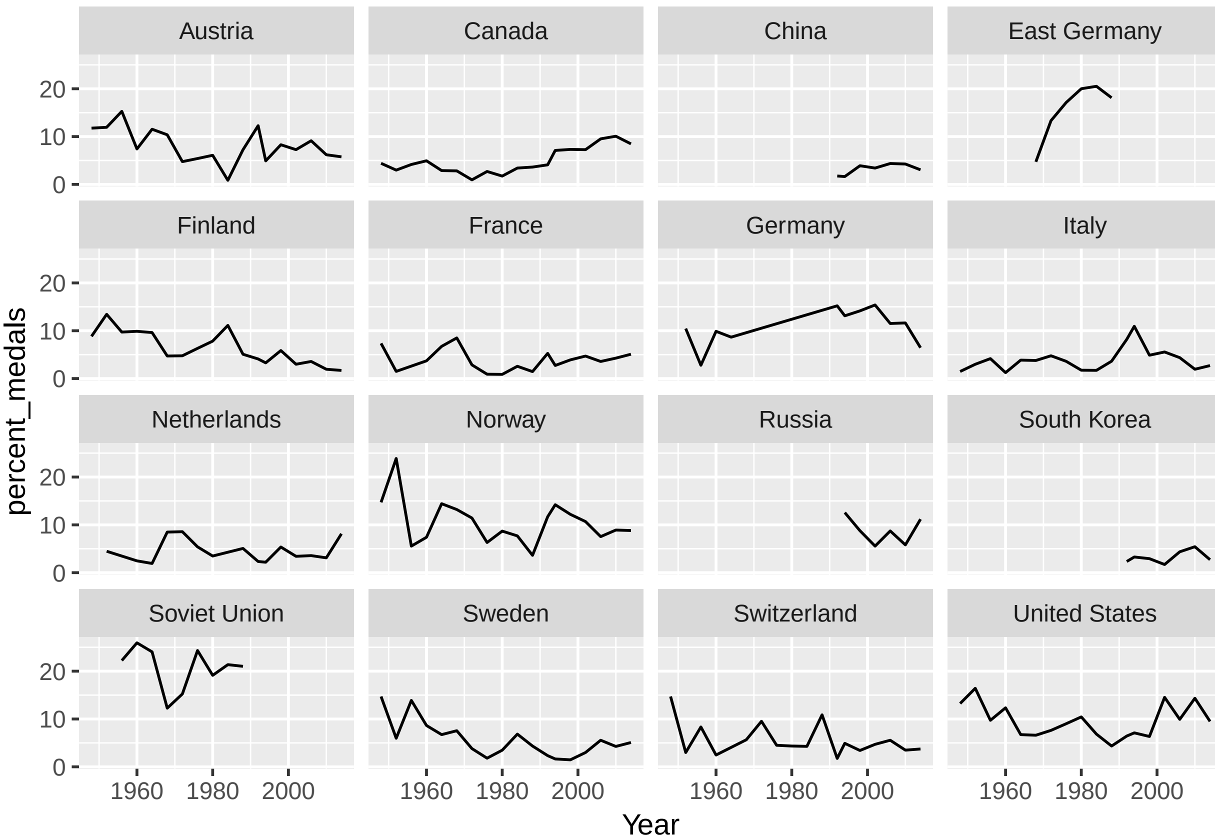
Chapter 8

Winter Games

A random sample from the data set:

Year	Sport	Event	Country	Gender	Medal Rank	Medal	Na
2006	Snowboarding	Men’s Halfpipe	United States	Men	2	silver	Da
1992	Alpine Skiing	Men’s Combined	Italy	Men	1	gold	Jos
2002	Alpine Skiing	Men’s Slalom	France	Men	2	silver	SÃ
2006	Cross-Country Skiing	Women’s 4 Ã— 5-Kilometer Relay	Germany	Women	2	silver	Ge
1956	Speedskating	Men’s 500 Meters	Soviet Union	Men	2	silver	Ra

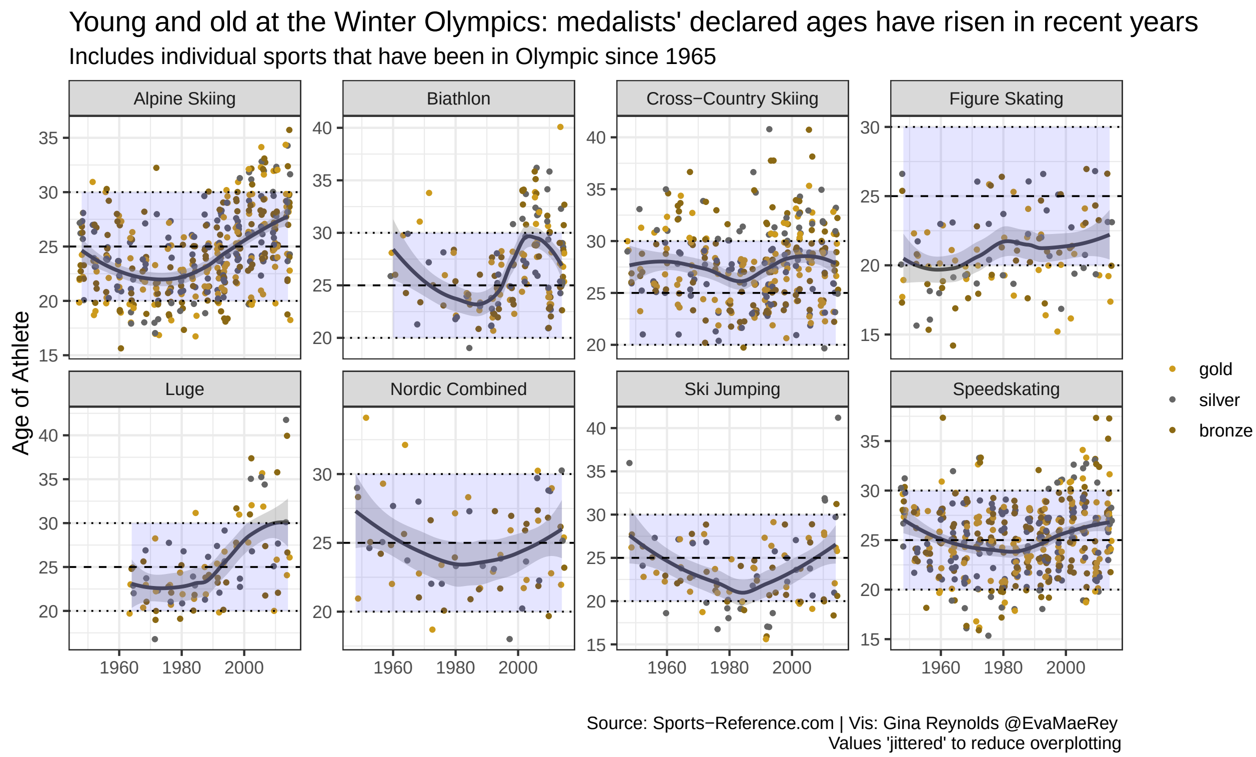
```
ggplot(data = dta) +
  aes(x = Year) +
  aes(y = percent_medals) +
  geom_line() +
  facet_wrap(~ Country)
```



A random sample from the data set:

Year	Sport	Event	Country	Gender	Medal Rank	Medal	Name of Athlete or
1994	Cross-Country Skiing	Men's 30 Kilometers	Norway	Men	2	silver	BjÃ¸rn DÃ¦hlie
1968	Cross-Country Skiing	Men's 50 Kilometers	Norway	Men	1	gold	Ole EllefsÃ¦ter
2006	Speedskating	Women's 5,000 Meters	Canada	Women	1	gold	Clara Hughes
1998	Ski Jumping	Men's Large Hill, Team	Austria	Men	3	bronze	Austria
1976	Speedskating	Men's 1,000 Meters	Norway	Men	2	silver	JÃ¸rn Didriksen

```
ggplot(dta) +
  aes(x = Year) +
  aes(y = `Age of Athlete`) +
  facet_wrap(~ Sport, scales = "free_y", nrow = 2) +
  geom_jitter(size = 1, mapping = aes(col = fct_inorder(Medal))) +
  geom_smooth(col = "grey30") +
  geom_ribbon(ymin = 20, ymax = 30, alpha = .1, fill = "blue") +
  geom_hline(yintercept = c(20, 30), lty = "dotted") +
  geom_hline(yintercept = c(25), lty = "dashed") +
  scale_color_manual(values = c("goldenrod3", "grey40", "goldenrod4"), name = "") +
  labs(x = "") +
  labs(title = "Young and old at the Winter Olympics: medalists' declared ages have risen in recent yea
  labs(subtitle = "Includes individual sports that have been in Olympic since 1965") +
  labs(caption = "Source: Sports-Reference.com | Vis: Gina Reynolds @EvaMaeRey \nValues 'jittered' to r
  theme_bw(base_size = 13)
```



Young and old at the Winter Olympics: medalists' declared ages have risen in recent years
Includes individual sports that have been in Olympic since 1965
Alpine Skiing
Biathlon
Cross−Country Skiing
Figure Skating
Luge
Nordic Combined
Ski Jumping
Speedskating
Age of Athlete
1960
1980
2000
gold
silver
bronze
Source: Sports−Reference.com | Vis: Gina Reynolds @EvaMaeRey
Values 'jittered' to reduce overplotting

Chapter 9

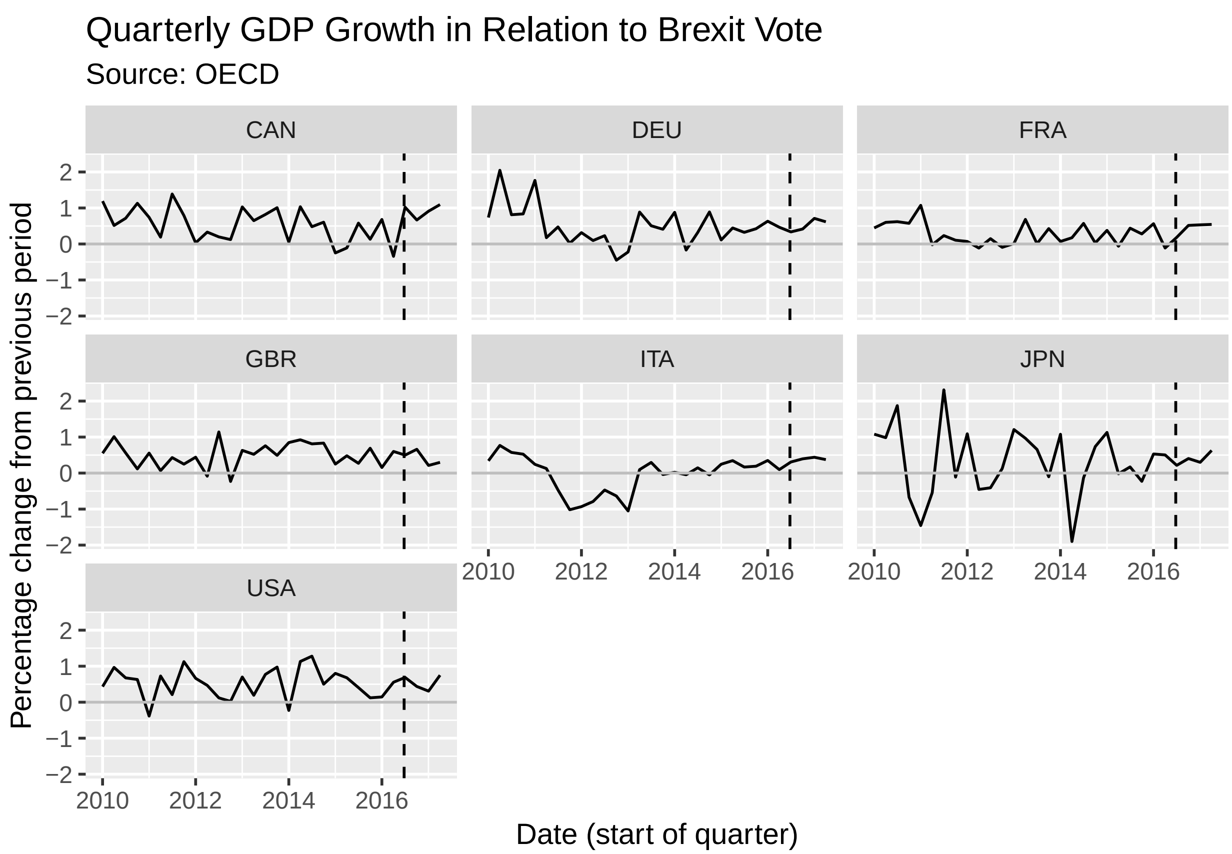
Brexit

This visualization challenge was a proposed makeover for a Financial Times visualization focusing on relative economic growth in G7 countries, with an emphasis on growth in the UK, focusing especially since Brexit. The visualization I present here is not what I created at the time of the challenge; instead it is inspired by Alan Smith a data journalist at the Financial Times, who created a really compelling visualization a couple of months after MakeoverMonday's treatment. I try to recreate his plot - which uses a ribbon to contain all G7 countries, and plot the UK's stats thereover. This declutters the graph, and makes you focus on where the UK falls among other countries, without being needlessly specific about those countries; the data story isn't about them anyway, might be Smith's thinking. My graph actually lightly traces economic growth in other countries, but deemphasizes their importance, like Smith.

A random sample from the data set:

Country	Year	Quarter	Date (start of quarter)	Percentage change from previous period	Date (start o quarter)
GBR	2013	4	2013-10-01	0.491800	2013-10-01
ITA	2012	2	2012-04-01	-0.790043	2012-04-01
USA	2015	4	2015-10-01	0.121010	2015-10-01
ITA	2013	4	2013-10-01	-0.039854	2013-10-01
GBR	2010	1	2010-01-01	0.549599	2010-01-01

```
ggplot(data = data) +
  aes(x = `Date (start of quarter)`) +
  aes(y = `Percentage change from previous period`) +
  facet_wrap(~ Country) +
  geom_line() +
  geom_hline(yintercept = 0, col = "grey") +
  geom_vline(xintercept = as.numeric(as.POSIXct("2016-06-23")), lty = "dashed") +
  labs(title = "Quarterly GDP Growth in Relation to Brexit Vote") +
  labs(subtitle = "Source: OECD")
```

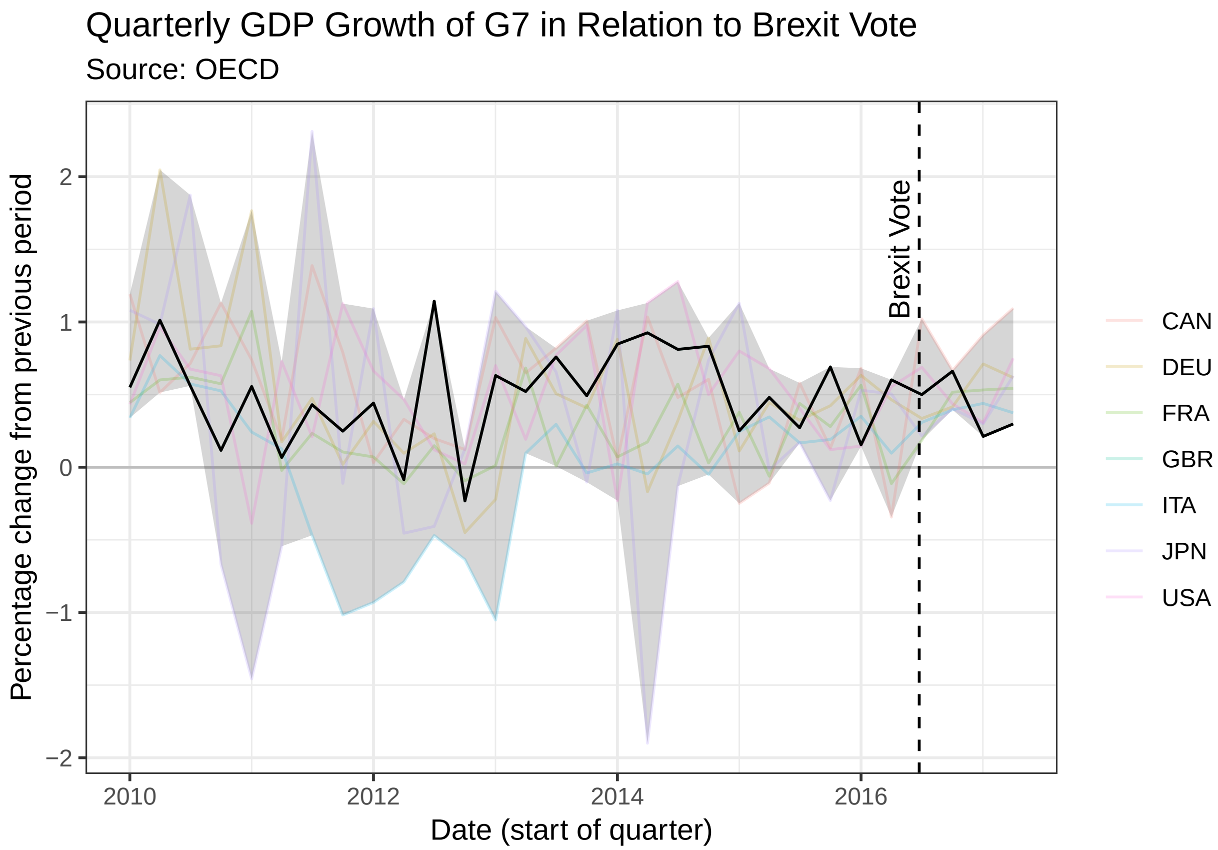


A random sample from the data set:

Country	Year	Quarter	Date (start of quarter)	Percentage change from previous period	Date (start o quarter)	
DEU	2011	3	2011-07-01	0.474351	2011-07-01	-
CAN	2013	4	2013-10-01	1.007924	2013-10-01	-
ITA	2016	3	2016-07-01	0.308248	2016-07-01	
DEU	2013	2	2013-04-01	0.886494	2013-04-01	
FRA	2016	2	2016-04-01	-0.112590	2016-04-01	-

```r
ggplot(data = data) +
  aes(x = `Date (start of quarter)`) +
  aes(y = `Percentage change from previous period`) +
  aes(ymin = min_) +
  aes(ymax = max_) +
  geom_hline(yintercept = 0, col = "grey") +
  geom_ribbon(alpha = .2) +
  geom_line(aes(col = Country), alpha = .2) +
  geom_line(data = data %>% filter(Country == "GBR"), col = "black") +
  geom_vline(xintercept = as.numeric(as.POSIXct("2016-06-23")), lty = 2) +
  annotate(
    geom = "text", x = as.POSIXct("2016-04-23"), y = 1.5,
    label = "Brexit Vote", angle = 90
  ) +
  labs(
    title = "Quarterly GDP Growth of G7 in Relation to Brexit Vote",
    subtitle = "Source: OECD",
    col = ""
  ) +
```

```r
theme_bw()
```


Quarterly GDP Growth of G7 in Relation to Brexit Vote
Source: OECD
Percentage change from previous period
2
1
0
−1
−2
2010
2012
2014
2016
Date (start of quarter)
Brexit Vote
CAN
DEU
FRA
GBR
ITA
JPN
USA

Chapter 10

Curry in London

This visualization task seemed to get at the question: Does where you eat matter. The data was the cost of identical menu items at different locations of the same restaurant, the Wetherspoon, around the UK.

First, I mapped the cost of a single menu item, the Empire Burger, across the UK. Then, I calculated the distance from Wetherspoon restaurants from the Big Ben, and plotted prices as a function of this distance – plotting only the restaurants in a 10 kilometer radius.

A random sample from the data set:

Name	Location	Latitude	Longitude	Empire State Burger	Chicken Tikka	Gammon aft
The Moon Under Water	Milton Keynes	52.04181	-0.7488899	8.75	7.49	
The Observatory	Ilkeston	52.97154	-1.3081310	8.75	6.99	
The Counting House	Congleton	53.16386	-2.2146526	8.75	7.40	
The Christopher Creeke	Bournemouth	50.72238	-1.8657451	8.75	6.40	
Bull and Stirrup Hotel	Chester	53.19436	-2.8933313	9.10	7.79	

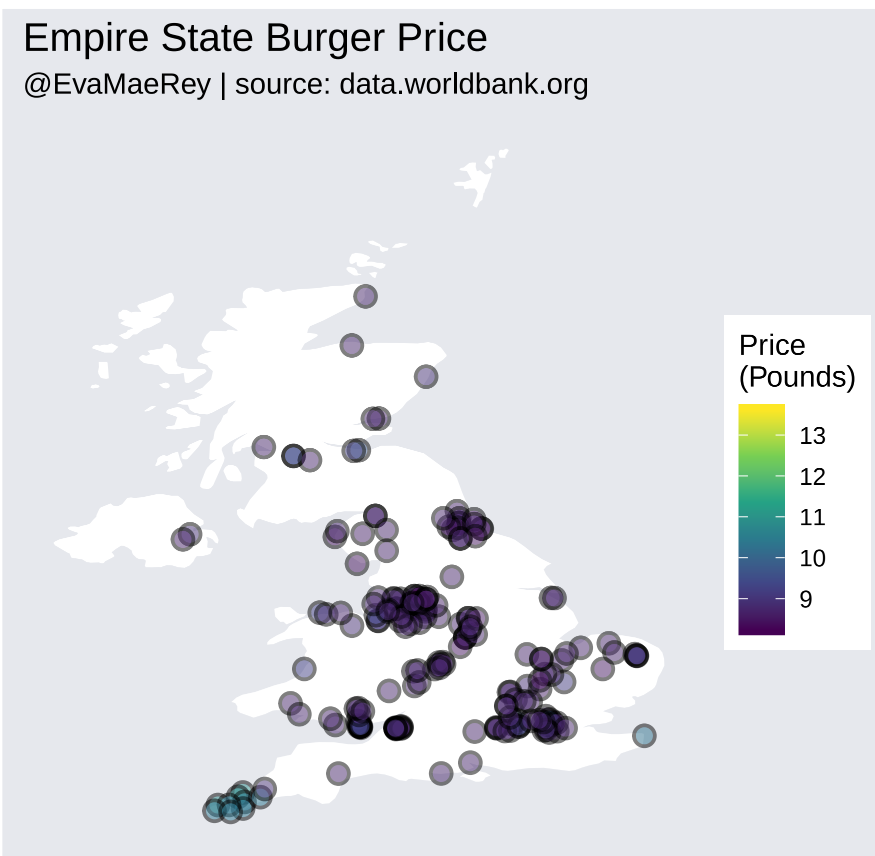
```
# Mapping data
world_map_df <- map_data("world")
```

A random sample from the data set:

	long	lat	group	order	region	subregion
71407	-72.28902	-9.629199	1124	71407	Peru	NA
89141	54.47285	42.180176	1418	89141	Turkmenistan	NA
100607	27.57383	-12.227051	1623	100607	Zambia	NA
34107	-81.81622	22.200195	457	34107	Cuba	NA
82254	138.66074	53.744774	1309	82254	Russia	NA

```
# create a blank ggplot theme
theme_opts <- theme(
  panel.grid.minor = element_blank(),
  panel.grid.major = element_blank(),
  panel.background = element_blank(),
  plot.background = element_rect(fill = "#e6e8ed"),
  panel.border = element_blank(),
  axis.line = element_blank(),
  axis.text.x = element_blank(),
  axis.text.y = element_blank(),
  axis.ticks = element_blank(),
  axis.title.x = element_blank(),
  axis.title.y = element_blank(),
```

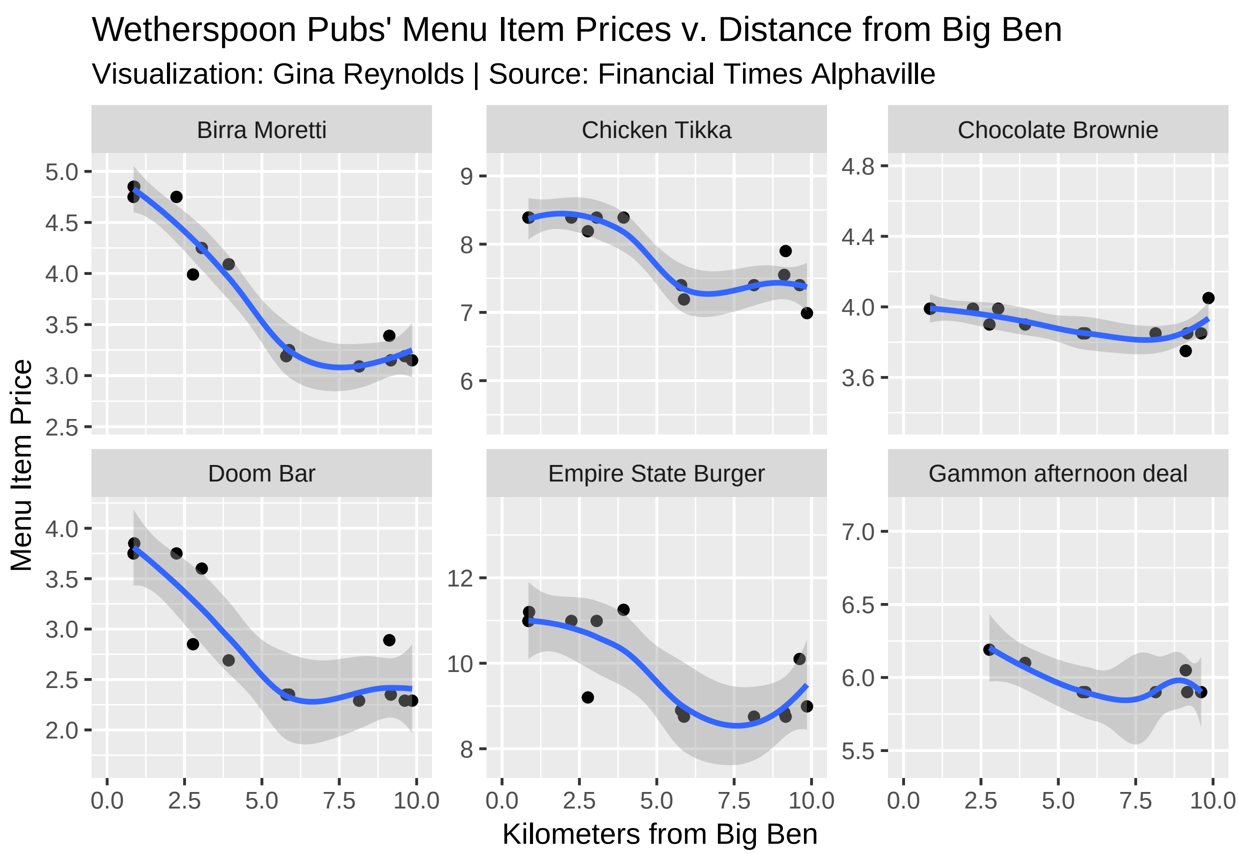
```
  plot.title = element_text(size = 15)
)


ggplot(data = world_map_df %>% filter(region == "UK")) +
  aes(x = long) +
  aes(y = lat) +
  aes(group = group) +
  geom_polygon(fill = "white") +
  coord_equal() +
  scale_fill_viridis_c(option = "viridis") +
  geom_point(data = data0,
             mapping = aes(x = Longitude, y = Latitude,
                           group = NULL, fill = `Empire State Burger`),
    colour = "black", shape = 21, stroke = 1, alpha = .5, size = 3
  ) +
  labs(fill = "Price\n(Pounds)") +
  labs(title = "Empire State Burger Price") +
  labs(subtitle = "@EvaMaeRey | source: data.worldbank.org") +
  theme_opts
```



A random sample from the data set:

Name	Location	Latitude	Longitude	Notes	Moretti as a % of a tikka	Moretti as %
The Muckle Cross	Elgin	57.64887	-3.3124440	NA	0.3770270	(
The Hatchet Inn	Newbury	51.40082	-1.3228789	NA	0.4581081	(
The Smithy Fold	Glossop	53.44258	-1.9487549	NA	0.4175676	(
The Robert Peel	Bury	53.59381	-2.2981393	NA	0.4297636	(
The Coliseum Picture Theatre	Cleethorpes	53.56090	-0.0312133	NA	0.4018445	(

```
ggplot(data = dataLong) +
  aes(x = `Kilometers from Big Ben`) +
  aes(y = `Menu Item Price`) +
  facet_wrap(~ Item, scales = "free_y") +
  geom_point() +
  geom_smooth() +
  xlim(c(0, 10)) +
  labs(title = "Wetherspoon Pubs' Menu Item Prices v. Distance from Big Ben") +
  labs(subtitle = "Visualization: Gina Reynolds | Source: Financial Times Alphaville")
```



Chapter 11

Life Expectancy Increases

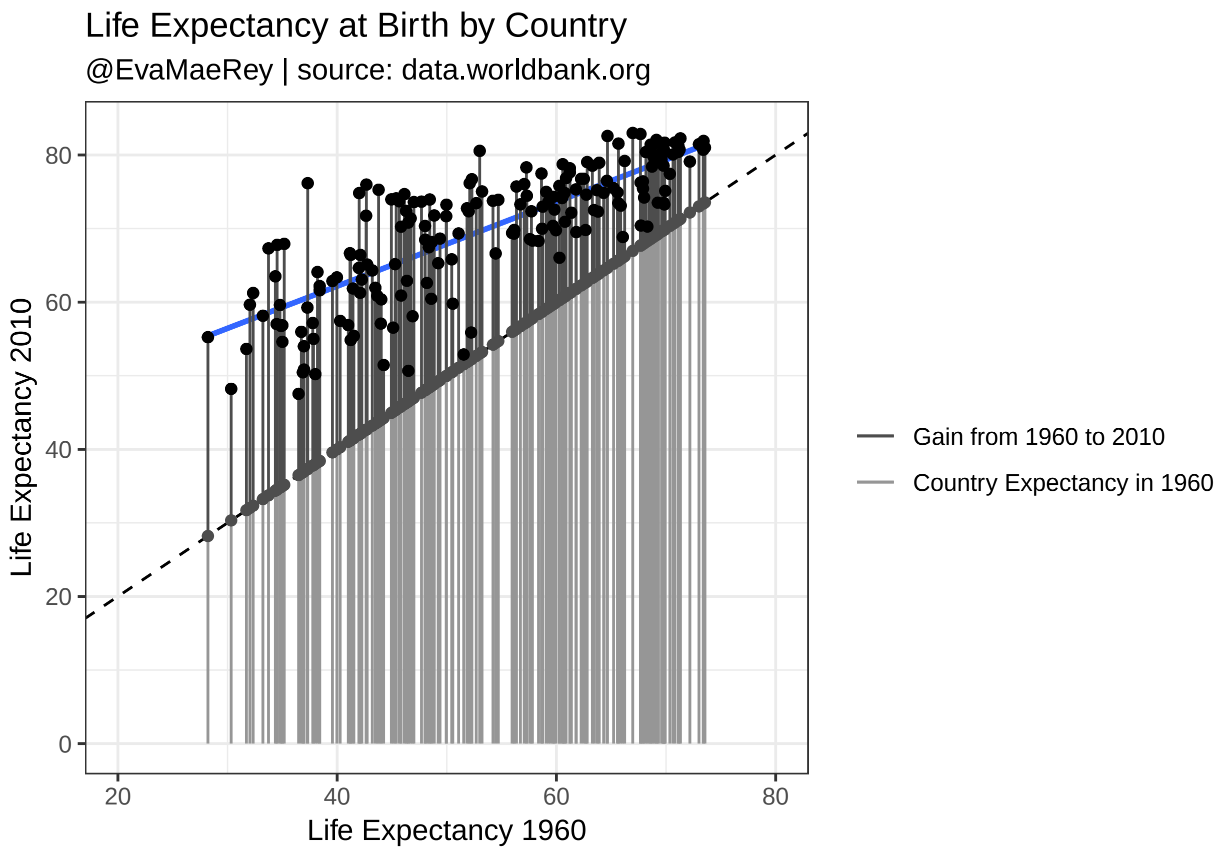
To dramatically show the increases in life expectancy by country from 1960 to 2010, I plot the variable in 1960 versus itself in 2010. The line of equivilance (a 45% angle) is used as a reference and shows the result that you would see if there where no growth. The vertical distance from this line is the increase in life expectancy. I also superimpose a linear model on top of the scatter plot. You can see that the gains are greater for countries that started off with lower life expectancies.

A random sample from the data set:

Life Expectancy 1960	Country Code	Country Name	Region	Income Group	Year
73.42317	ISL	Iceland	Europe & Central Asia	High income	2010
58.69405	UZB	Uzbekistan	Europe & Central Asia	Lower middle income	2010
62.72905	BHS	Bahamas, The	Latin America & Caribbean	High income	2010
59.73227	LKA	Sri Lanka	South Asia	Lower middle income	2010
41.01805	COD	Congo, Dem. Rep.	Sub-Saharan Africa	Low income	2010

```
ggplot(compare) +
  aes(x = `Life Expectancy 1960`) +
  aes(y = `Life Expectancy 2010`) +
  geom_point() +
  geom_smooth(se = F, method = "lm") +
  geom_abline(slope = 1, intercept = 0, lty = 2) +
  # coord_fixed() +
  aes(xend = `Life Expectancy 1960`) +
  aes(yend = `Life Expectancy 1960`) +
  geom_segment(mapping = aes(col = "Gain from 1960 to 2010")) +
  geom_segment(mapping = aes(y = 0, col = "Country Expectancy in 1960")) +
  scale_color_manual(
    breaks = c(
      "Gain from 1960 to 2010",
      "Country Expectancy in 1960"
    ),
    values = c("grey59", "grey30", "grey30")
  ) +
  geom_point(aes(y = `Life Expectancy 1960`), col = "grey30") +
  geom_point() +
  labs(subtitle = "@EvaMaeRey | source: data.worldbank.org", size = .7) +
  labs(title = "Life Expectancy at Birth by Country") +
  labs(col = "") +
  theme(legend.title = element_blank()) +
```

```
theme_bw() +
xlim(c(20, 80))
```



Chapter 12

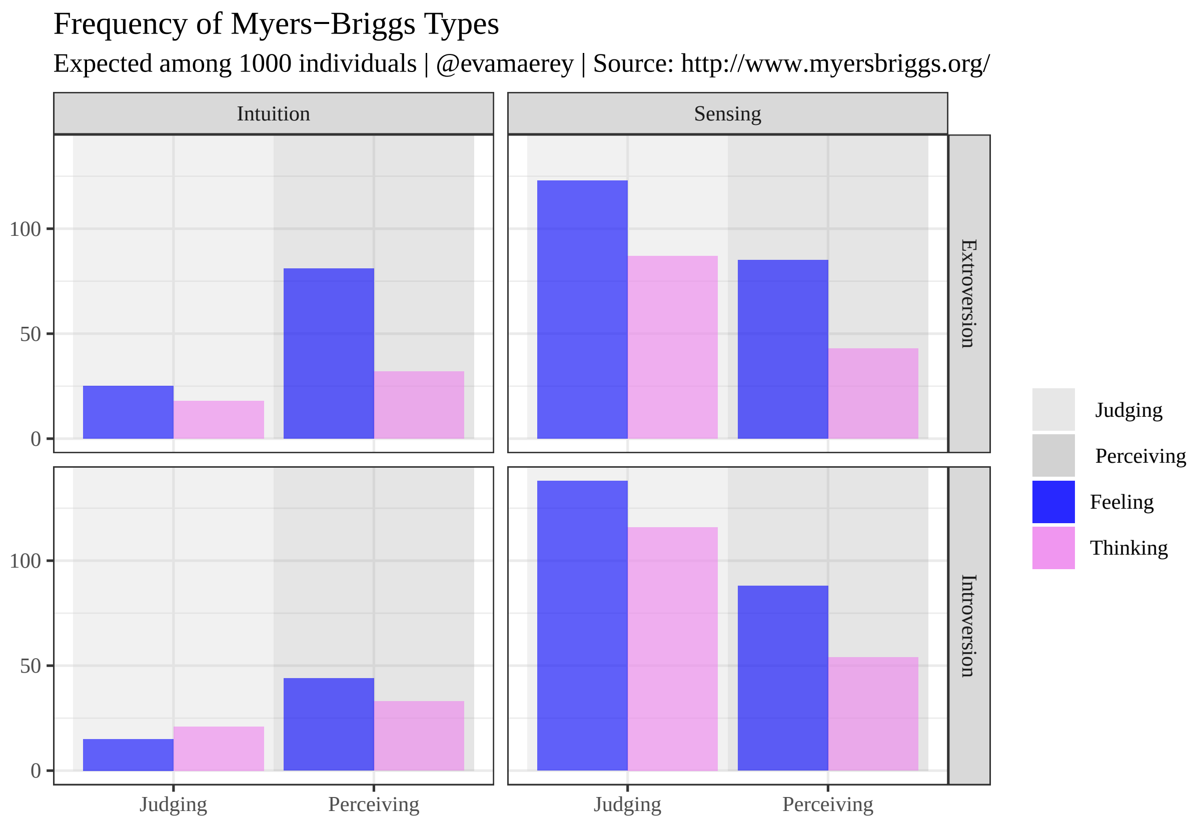
Myers Briggs

This data looks at the relationship betweeen four binary variables. The challenge is how to display that in one visualization. My first idea was to use a mosaic plot. However, I came across advice from "The Perceptual Edge", that generally advised against the use of the mosaic plot, instead favoring a kind of nested bar plot. I tried to implement that. While I do think that it is pretty, I think that it still requires a lot of the reader to interpret the graph. Perhaps more annotation could alleviate this burden.

A random sample from the data set:

(S)ensing/I(N)tuition	(T)hinking/(F)eeling	(J)udging/(P)erceiving	(E)xtroversion/(I)ntroversion	count
Sensing	Thinking	Judging	Introversion	1
Sensing	Thinking	Perceiving	Introversion	1
Sensing	Thinking	Judging	Introversion	1
Sensing	Feeling	Perceiving	Extroversion	1
Intuition	Thinking	Perceiving	Introversion	1

```
ggplot(d) +
  aes(x = `(J)udging/(P)erceiving`) +
  aes(fill = `(T)hinking/(F)eeling`) +
  facet_grid(`(E)xtroversion/(I)ntroversion` ~
               `(S)ensing/I(N)tuition`) +
  geom_rect(aes(x = NULL, y = NULL,
                xmin = mins, xmax = max,
                fill = `judging perceiving`),
                ymin = 0, ymax = 700, data = background
  ) +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = alpha(c("lightgrey", "darkgrey", "blue", "violet"), c(.3, .3, .6, .6))) +
  labs(x = "") +
  labs(y = "") +
  labs(fill = "") +
  labs(title = "Frequency of Myers-Briggs Types") +
  labs(subtitle = "Expected among 1000 individuals | @evamaerey | Source: http://www.myersbriggs.org/")
  theme_bw(base_size = 10, base_family = "Times")
```


Frequency of Myers−Briggs Types
Expected among 1000 individuals | @evamaerey | Source: http://www.myersbriggs.org/
Intuition
Sensing
Extroversion
Introversion
100
50
0
100
50
0
Judging
Perceiving
Judging
Perceiving
Judging
Perceiving
Feeling
Thinking

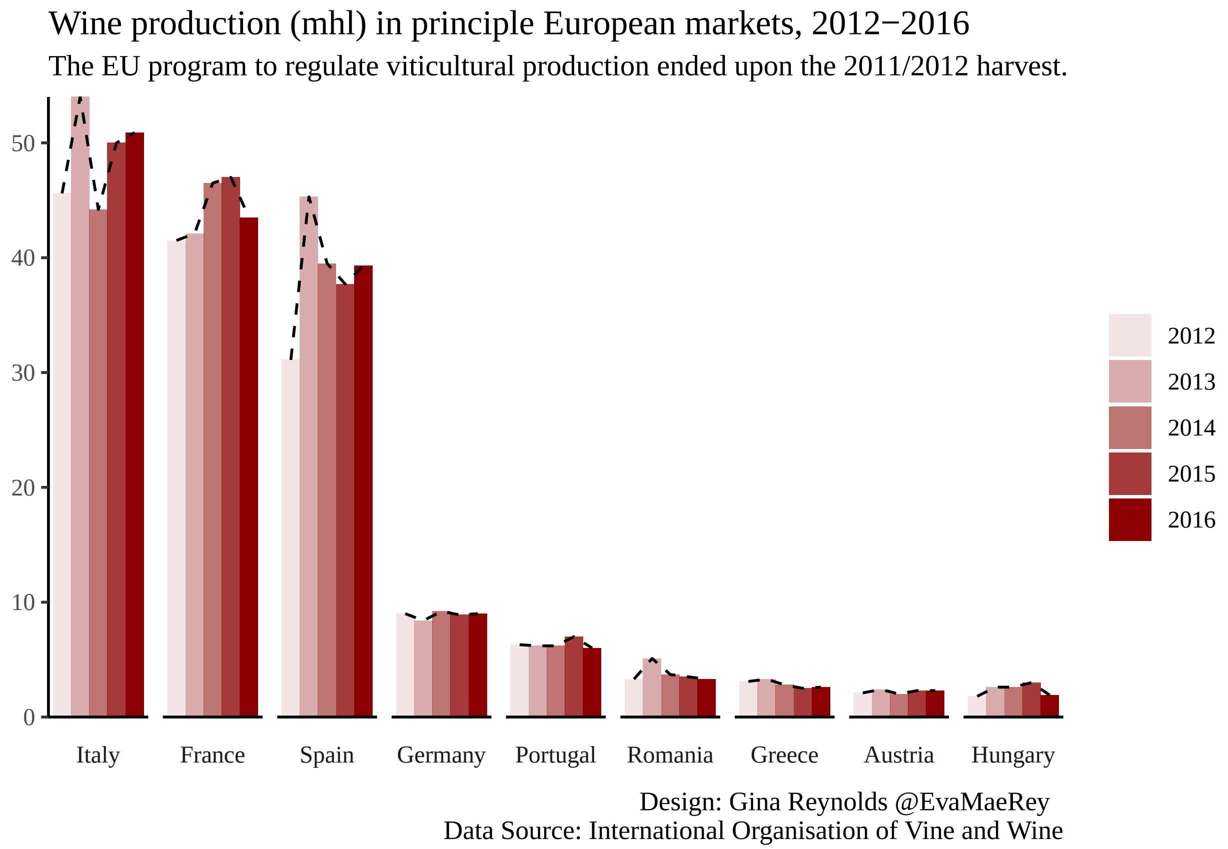
Chapter 13

Wine

Wine production in Europe may have been volitile during the years plotted because of production control policies implemented by the EU. The big three, Italy, France and Spain, particularly saw a lot of volitility early in this period.

```
df <- readxl::read_xlsx("raw_data/Wine_Production_by_country.xlsx") %>%
  filter(Country != "World total")

Europe <- c(
  "Italy", "France", "Spain",
  "Germany", "Portugal", "Romania",
  "Austria", "Greece", "Hungary"
)
```

```
  ggplot(df %>% filter(Country %in% Europe)) +
  aes(x = Year) +
  aes(y = `Wine production in mhl`) +
  facet_wrap(~ fct_inorder(Country), strip.position = "bottom", nrow = 1) +
  geom_col(aes(alpha = Year), position = "dodge", fill = "darkred", width = 1) +
  geom_line(col = "black", lty = 2) +
  scale_y_continuous(expand = c(0, 0)) +
  labs(fill = "") +
  labs(alpha = "")+
  labs(title = "Wine production (mhl) in principle European markets, 2012-2016") +
  labs(subtitle = "The EU program to regulate viticultural production ended upon the 2011/2012 harvest.
    labs(caption = "Design: Gina Reynolds @EvaMaeRey  \nData Source: International Organisation of Vine
  theme_classic(base_family = "Times") +
  theme(
    axis.title = element_blank(),
    strip.placement = "outside",
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank(),
    strip.background = element_blank(),
    plot.caption = element_text(size = 10)
  )
```


Wine production (mhl) in principle European markets, 2012−2016
The EU program to regulate viticultural production ended upon the 2011/2012 harvest.
50
40
30
20
10
0
Italy
France
Spain
Germany
Portugal
Romania
Greece
Austria
Hungary
2012
2013
2014
2015
2016
Design: Gina Reynolds @EvaMaeRey
Data Source: International Organisation of Vine and Wine

Chapter 14

Arctic Ice

This visualization shows the trend in Arctic Ice Sea Extent, data from the National Snow and Ice Data Center. If I recall correctly, the definition for coverage is the case where at least 15 percent of the sea is ice.

The visualization shows melting and freezing cycles, in accordance with the seasons — and the disconcerting trend of a general decrease in ice extent over the years.

One problem that arises is due to inconsistant number of days in each year. There is a measurement for every day, but leap years contain a extra day. Which means that plotting years over years leads to imperfect alignment. My solution was just to pretend that all the data come from a single year, 2000, and plot each of the years on that scale. The earliest year cycle and last year cycle are highlighted in white.

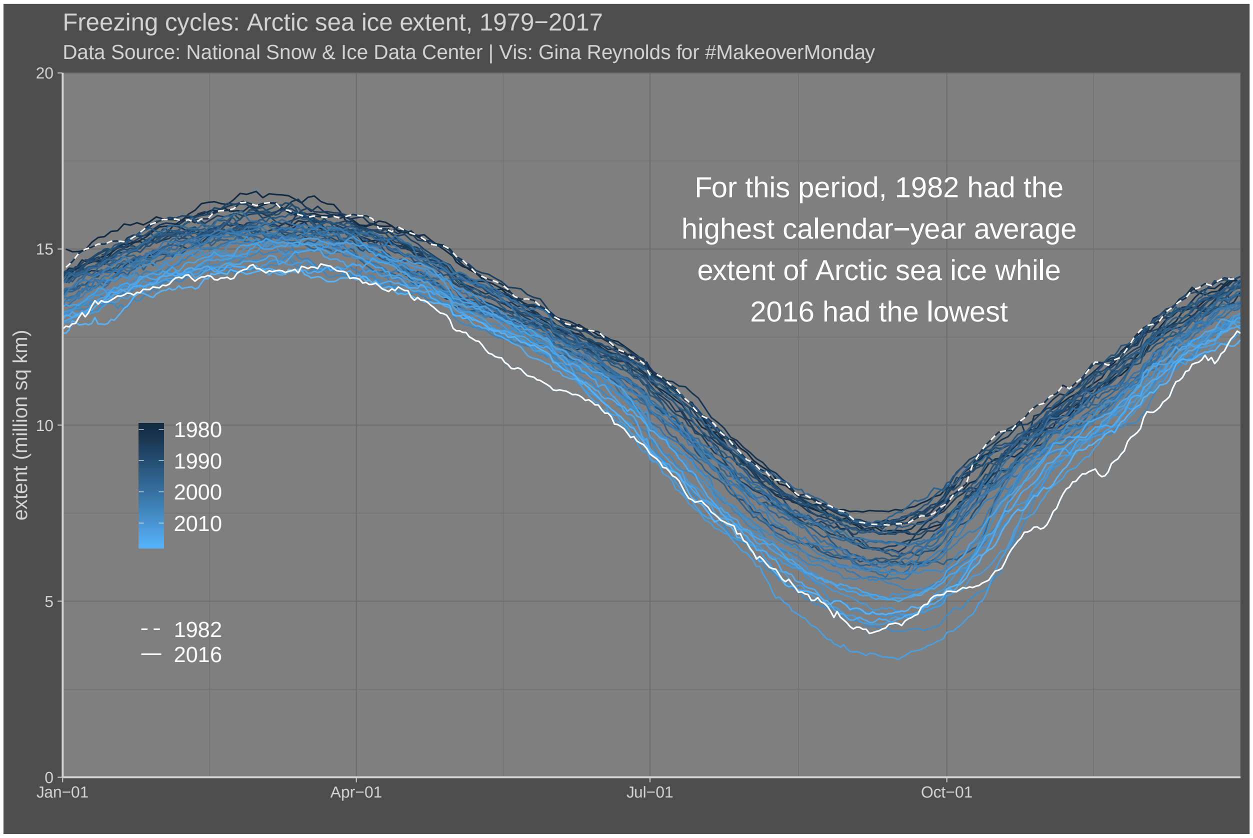
A random sample from the data set:

Date	Extent (million sq km)	year	month_day	month_day_plus	proportion_ocean_covered_in_ice	mea
1998-05-10	13.991	1998	05-10	2000-05-10	0.0388639	
1989-02-24	15.638	1989	02-24	2000-02-24	0.0434389	
1988-01-25	15.020	1988	01-25	2000-01-25	0.0417222	
2008-09-04	4.686	2008	09-04	2000-09-04	0.0130167	
1984-07-06	10.807	1984	07-06	2000-07-06	0.0300194	

year	average_coverage	num_days	average_day
1982	12.43945	182	1982-07-02 00:00:00
2016	10.15069	366	2016-07-01 12:00:00

```r
# breaks for x axis.
br <- as.numeric(lubridate::ymd(c(
  "2000-01-01", "2000-04-01",
  "2000-07-01", "2000-10-01", "2001-01-01"
)))

ggplot(df) +
  aes(x = as.numeric(month_day_plus)) +
  aes(y = `Extent (million sq km)`) +
  aes(group = year) +
  geom_line() +
  aes(col = year) +
  scale_x_continuous(
    breaks = br,
    labels = c("Jan-01", "Apr-01", "Jul-01", "Oct-01", "Jan-01"),
    expand = c(0, 0)
  ) +
```

```
  scale_y_continuous(expand = c(0, 0), limits = c(0, 20)) +
  scale_color_continuous(
    guide = guide_colourbar(reverse = TRUE),
    breaks = seq(2010, 1980, -10)
  ) +
  geom_line(aes(lty = factor(year)),
    data = df %>% filter(year == 2016 | year == 1982),
    col = "white"
  ) +
  scale_linetype_manual(
    name = "",
    values = c("dashed", "solid")
  ) +
  annotate(
    geom = "text", x = 11210, y = 15,
    label = str_wrap("For this period, 1982 had the highest calendar-year average extent of Arctic sea i
    col = "white",
    size = 7
  ) +
  labs(x = "") +
  labs(y = "extent (million sq km)") +
  labs(col = "") +
  labs(lty = "") +
  labs(title = "Freezing cycles: Arctic sea ice extent, 1979-2017") +
  labs(subtitle = "Data Source: National Snow & Ice Data Center | Vis: Gina Reynolds for #MakeoverMonday
  theme_dark(base_size = 14) +
  theme(
    legend.background = element_blank(),
    legend.position = c(0.1, .35),
    legend.text = element_text(colour = "white", size = 15),
    plot.background = element_rect(fill = "grey30"),
    plot.title = element_text(colour = "lightgrey"),
    plot.subtitle = element_text(colour = "lightgrey"),
    axis.title = element_text(colour = "lightgrey"),
    axis.line = element_line(colour = "lightgrey"),
    axis.text = element_text(colour = "lightgrey"),
    axis.ticks = element_line(colour = "lightgrey")
  )
```


Freezing cycles: Arctic sea ice extent, 1979−2017
Data Source: National Snow & Ice Data Center | Vis: Gina Reynolds for #MakeoverMonday
For this period, 1982 had the highest calendar−year average extent of Arctic sea ice while 2016 had the lowest
extent (million sq km)
20
15
10
5
0
1980
1990
2000
2010
1982
2016
Jan−01
Apr−01
Jul−01
Oct−01