ETC1010: Data Modelling and Computing

Lecture 4: Plotting your data

Di Cook (dicook@monash.edu, @visnut)

Week 4

0

ETC1010: Data Modelling and Computing

Lecture 4: Plotting your data

Di Cook (dicook@monash.edu, @visnut)

Week 4

Overview

- Working with dates
  - Does every year have 365 days? Does every day have 24 hours? Does every minute have 60 seconds?
  - Where are you?
  - What day of the week is it? Day of the month? Week in the year?
  - Are we talking months as numbers or names?
  - How many days until we go on holidays?
- Grammar of graphics
  - Language of defining plots, that integrates with statistical thinking
  - Way to say how one plot is the same or different from another, e.g. barchart v pie chart
  - Evaluate whether one design is better than another for communication

```
ym
[1
md
[1
dm
[1
ym
[1
wd
[1
yd
[1
to
[1
to
[1
```

What's in a data?

How many ways can you write down today's date?

Write them into this window:

http://collabedit.com/x9nvf

A

```
ymd("2017-08-15")
[1] "2017-08-15"
mdy("08/15/2017")
[1] "2017-08-15"
dmy("15082017")
[1] "2017-08-15"
ymd_hms("2015:08:15 10:05:30", tz = "Australia/Melbourne")
[1] "2015-08-15 10:05:30 AEST"
wday("2017-08-15")
[1] 3
yday("2017-08-15")
[1] 227
today()
[1] "2017-08-14"
today(tz = "America/Los_Angeles")
[1] "2017-08-14"
```

Fl

Airline data

```
la
la

#

1
2
3
4
5
6
7
```

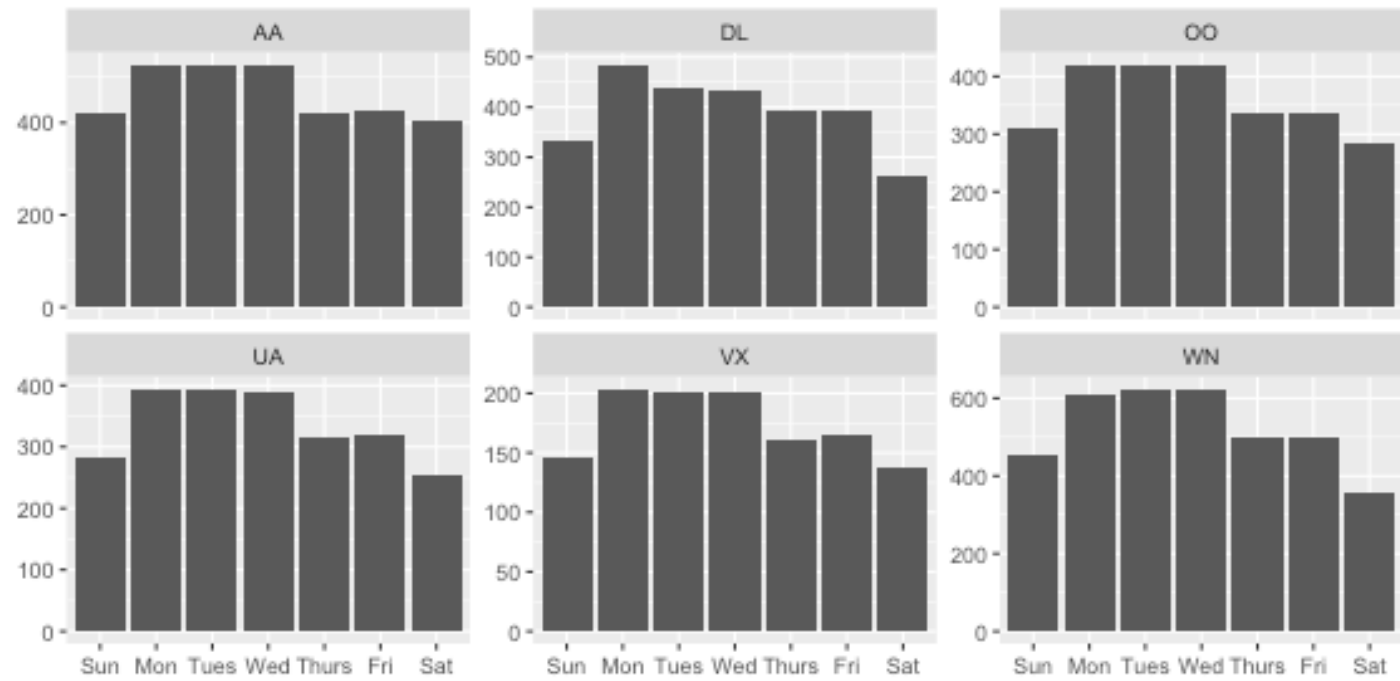
```
# A tibble: 18,166 x 44
    YEAR MONTH DAY_OF_MONTH DAY_OF_WEEK    FL_DATE UNIQUE_CARRIER
   <int> <int>        <int>       <int>     <date>          <chr>
 1  2017     5            1           1 2017-05-01             AA
 2  2017     5            2           2 2017-05-02             AA
 3  2017     5            3           3 2017-05-03             AA
 4  2017     5            4           4 2017-05-04             AA
 5  2017     5            5           5 2017-05-05             AA
 6  2017     5            6           6 2017-05-06             AA
 7  2017     5            7           7 2017-05-07             AA
 8  2017     5            8           1 2017-05-08             AA
 9  2017     5            9           2 2017-05-09             AA
10  2017     5           10           3 2017-05-10             AA
# ... with 18,156 more rows, and 38 more variables: AIRLINE_ID <int>,
#   CARRIER <chr>, TAIL_NUM <chr>, FL_NUM <int>, ORIGIN <chr>,
#   ORIGIN_CITY_NAME <chr>, ORIGIN_STATE_ABR <chr>, DEST <chr>,
#   DEST_CITY_NAME <chr>, DEST_STATE_ABR <chr>, CRS_DEP_TIME <chr>,
#   DEP_TIME <chr>, DEP_DELAY <dbl>, DEP_DELAY_NEW <dbl>, DEP_DEL15 <dbl>,
#   DEP_DELAY_GROUP <int>, TAXI_OUT <dbl>, WHEELS_OFF <chr>,
#   WHEELS_ON <chr>, TAXI_IN <dbl>, CRS_ARR_TIME <chr>, ARR_TIME <chr>,
#   ARR_DELAY <dbl>, ARR_DELAY_NEW <dbl>, ARR_DEL15 <dbl>,
#   CANCELLED <dbl>, CANCELLATION_CODE <chr>, DIVERTED <dbl>,
#   CRS_ELAPSED_TIME <dbl>, ACTUAL_ELAPSED_TIME <dbl>, AIR_TIME <dbl>,
#   DISTANCE <dbl>, CARRIER_DELAY <dbl>, WEATHER_DELAY <dbl>,
#   NAS_DELAY <dbl>, SECURITY_DELAY <dbl>, LATE_AIRCRAFT_DELAY <dbl>,
#   X44 <chr>
```

Flights per day of the week

```
lax <- lax %>% mutate(day=lubridate::wday(FL_DATE, label=TRUE))
lax %>%
  count(day)
# A tibble: 7 x 2
    day     n
  <ord> <int>
1   Sun  2267
2   Mon  3037
3  Tues  2990
4   Wed  2980
5 Thurs  2451
6   Fri  2461
7   Sat  1980
```

By carrier

```
lax %>% filter(CARRIER %in% c("WN", "AA", "DL", "OO", "UA", "VX")) %>%
  ggplot(aes(x=day)) + geom_bar() +
  facet_wrap(~CARRIER, ncol=3, scales="free_y") +
  xlab("") + ylab("")
```



Fl

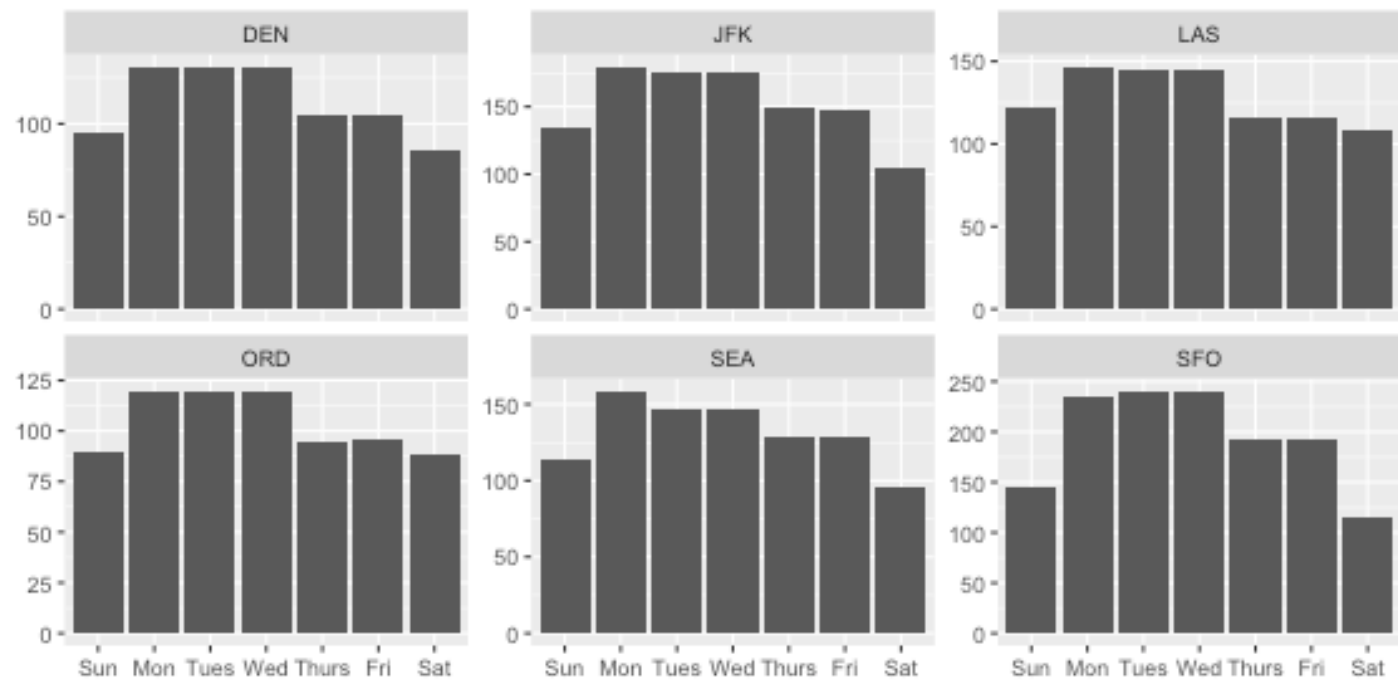
By origin

```
la
la

#

1
2
3
4
5
```

```
lax %>% filter(ORIGIN %in% c("SFO", "JFK", "SEA", "LAS", "DEN", "ORD")) %>%
  ggplot(aes(x=day)) + geom_bar() +
  facet_wrap(~ORIGIN, ncol=3, scales="free_y") +
  xlab("") + ylab("")
```



Flights by week of the year

```
lax <- lax %>% mutate(week=lubridate::week(FL_DATE))
lax %>%
  count(week)
# A tibble: 5 x 2
   week     n
  <dbl> <int>
1    18  3502
2    19  4105
3    20  4105
4    21  4135
5    22  2319
```

Yo

Ho

Your turn

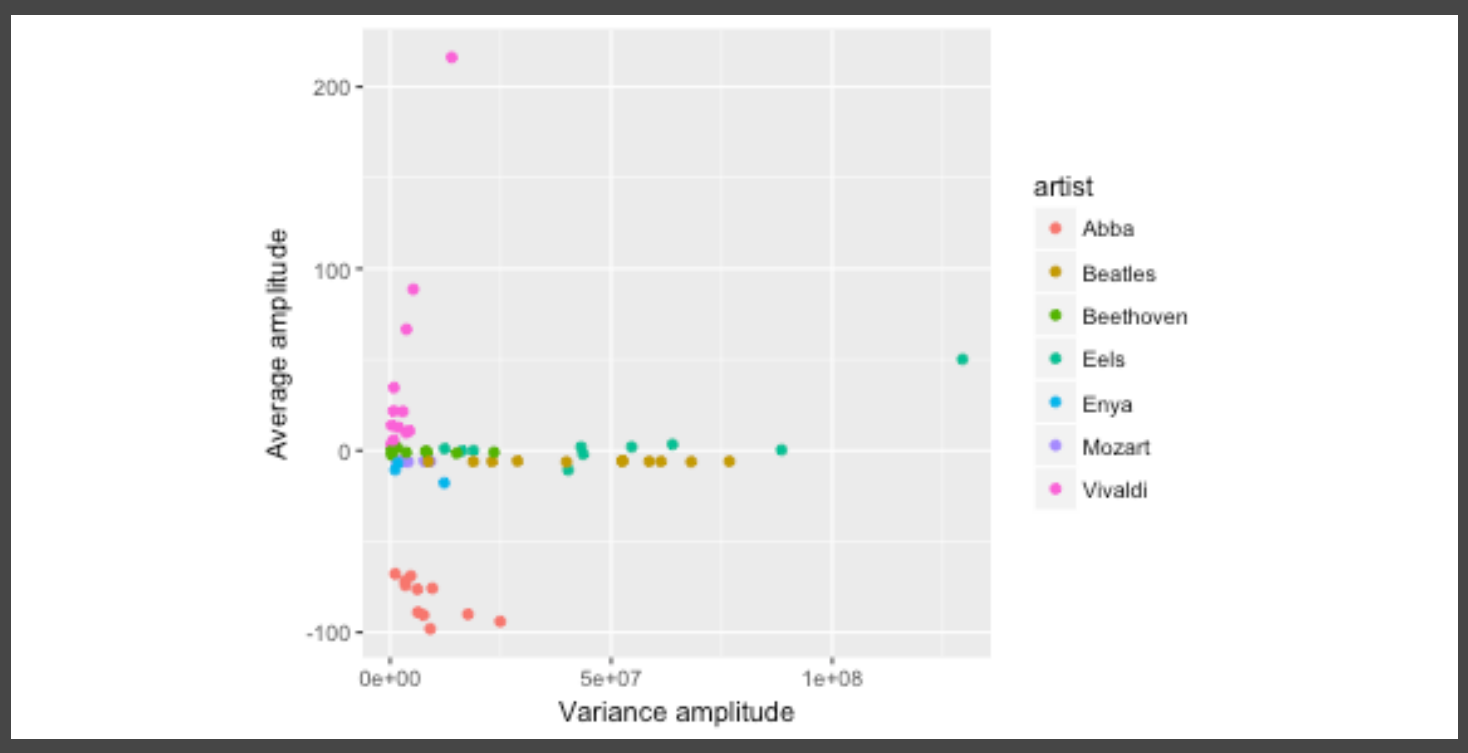
- What is a (data) plot?
- What are the three most important data plots?

Yc

Wh

Your turn

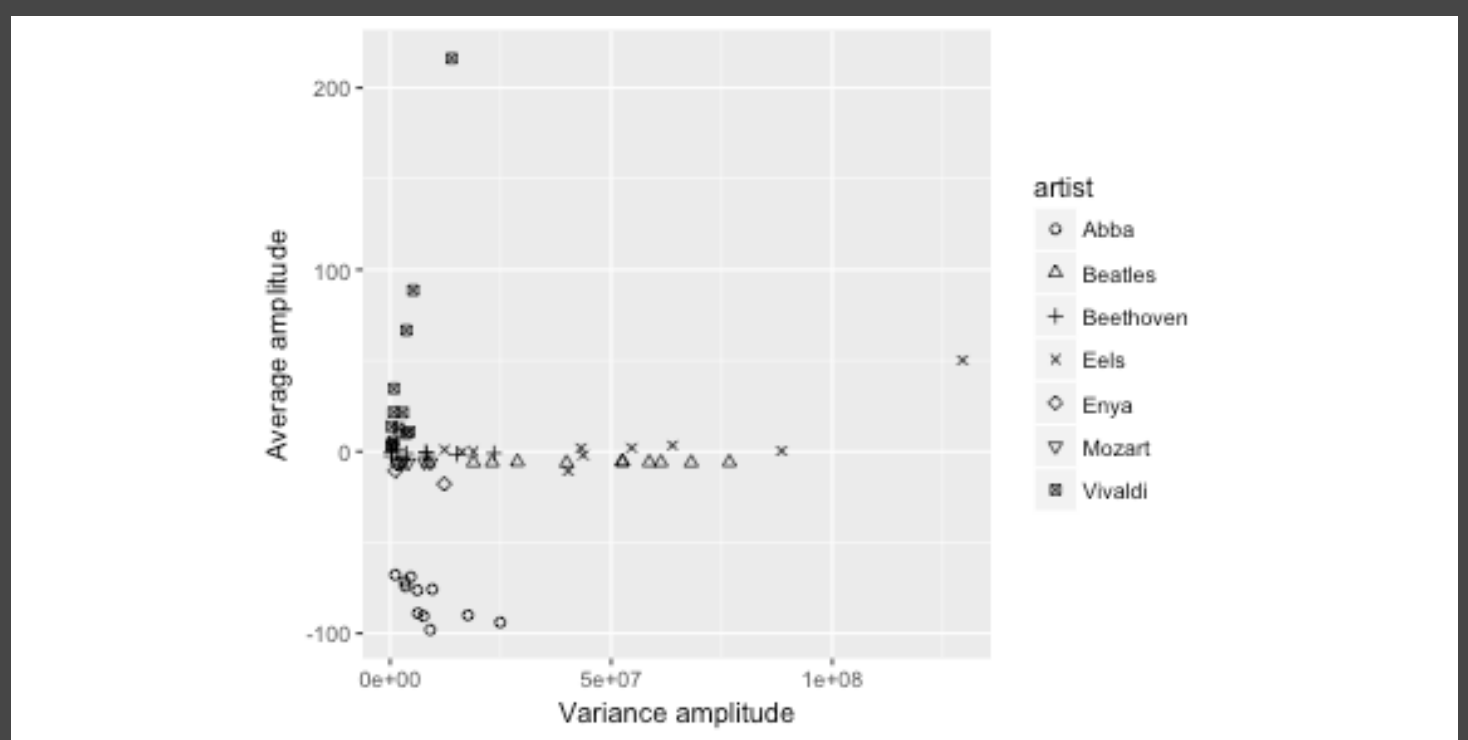
How would you describe this plot?



E

Your turn

What about this plot?



anc

Elements of a data plot

- data
- mapping of variables to graphical elements (aesthetics)
- type of plot structure to use (geom)
- transformations: log scale, ...

and ...

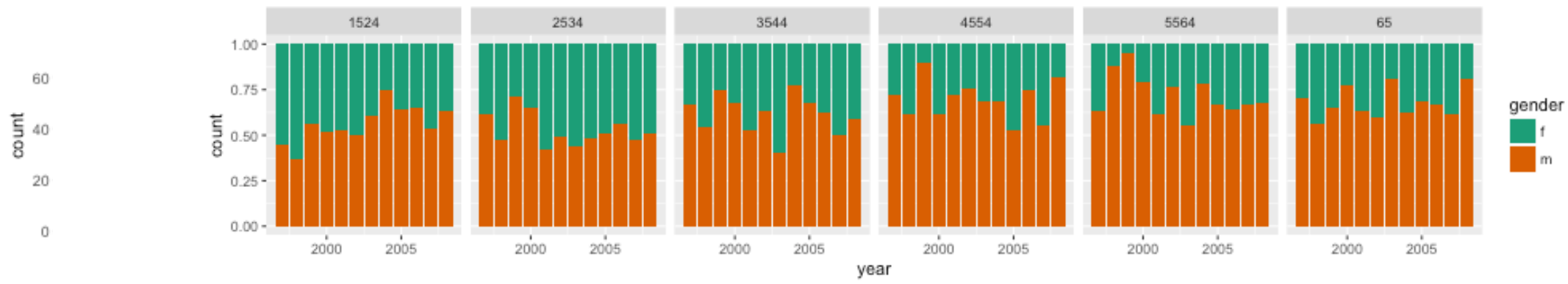
- layers: multiple geoms, multiple data sets, annotation
- facets: show subsets in different plots
- themes: modifying style

Why use a grammar of graphics?

- Remember tidy data?
- Data is organised into variables and observations.
- With a grammar, the variables are directly mapped to an element in the plot

Tuberculosis data

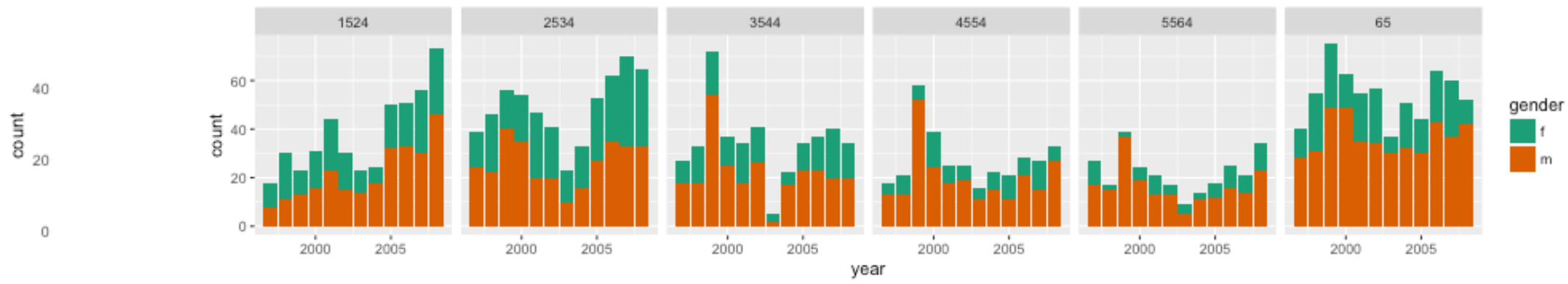
```
ggplot(tb_au, aes(x = year, y = count, fill = gender)) +
  geom_bar(stat = "identity", position = "fill") +
  facet_grid(~ age) +
  scale_fill_brewer(palette="Dark2")
```



100% charts? What so we learn?

Bar charts

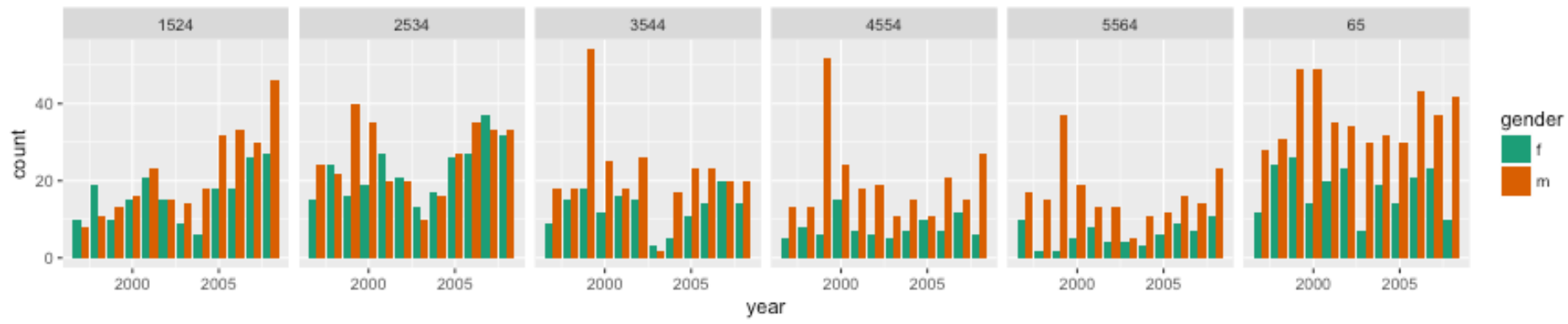
```
ggplot(tb_au, aes(x = year, y = count, fill = gender)) +
  geom_bar(stat = "identity") +
  facet_grid(~ age) +
  scale_fill_brewer(palette="Dark2")
```



What do we learn?

Side-by-side barcharts

```
ggplot(tb_au, aes(x = year, y = count, fill = gender)) +
  geom_bar(stat = "identity", position="dodge") +
  facet_grid(~ age) +
  scale_fill_brewer(palette="Dark2")
```



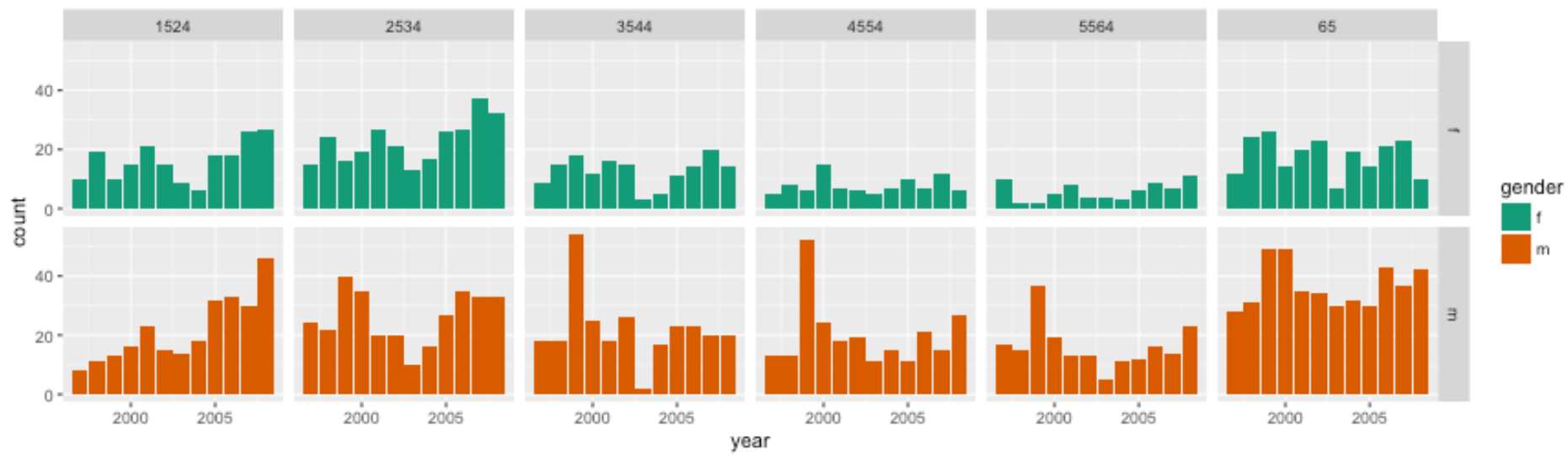
What is the focus now?

Pi

gg

Separate bar charts

```
ggplot(tb_au, aes(x = year, y = count, fill = gender)) +
  geom_bar(stat = "identity") +
  facet_grid(gender ~ age) +
  scale_fill_brewer(palette="Dark2")
```

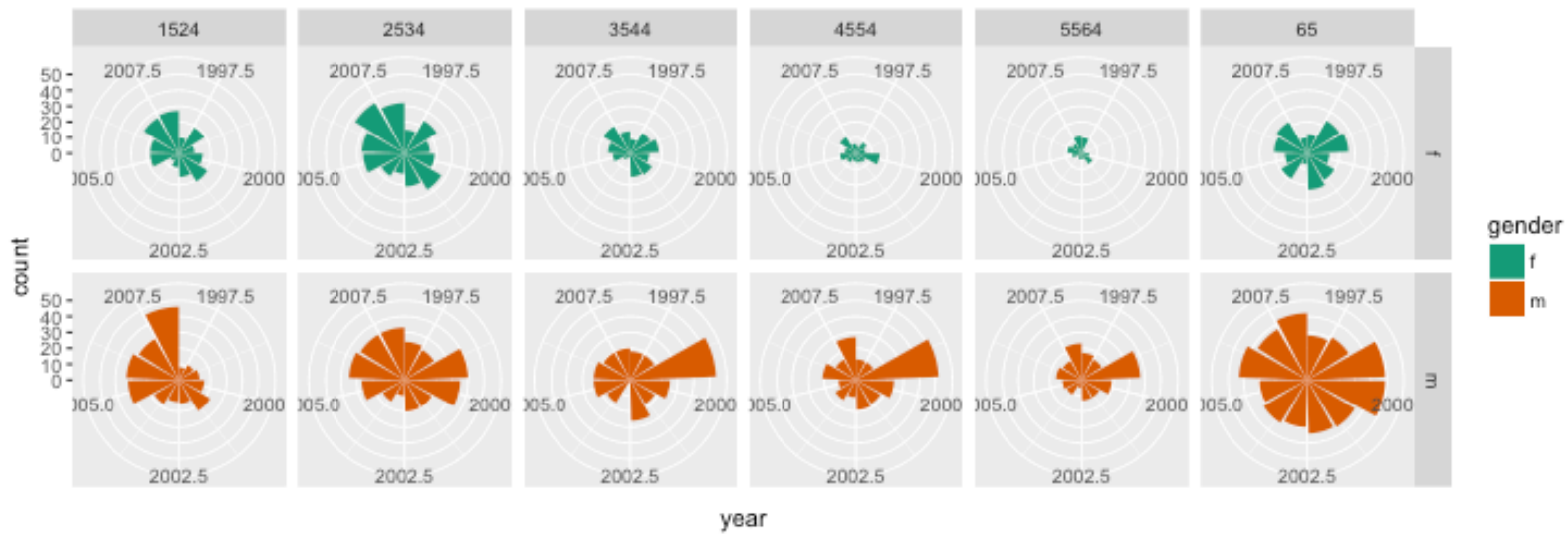


No

What is the focus now?

Pie charts?

```
ggplot(tb_au, aes(x = year, y = count, fill = gender)) +
  geom_bar(stat = "identity") +
  facet_grid(gender ~ age) +
  scale_fill_brewer(palette="Dark2") + coord_polar()
```

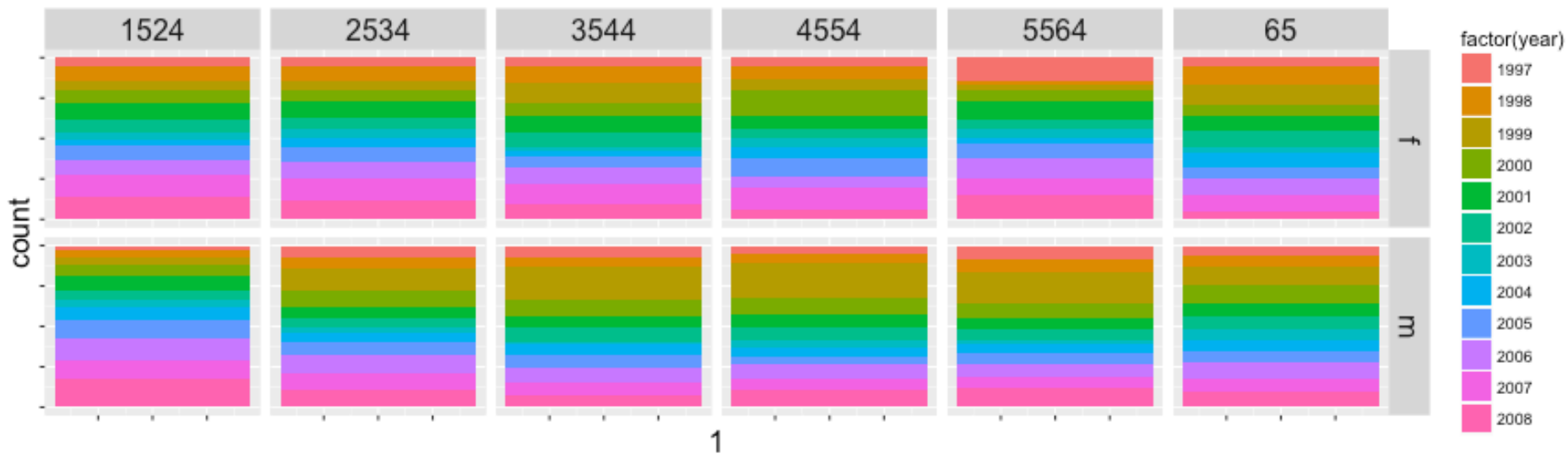


Nope! That's a rose chart.

Its

Rainbow charts?

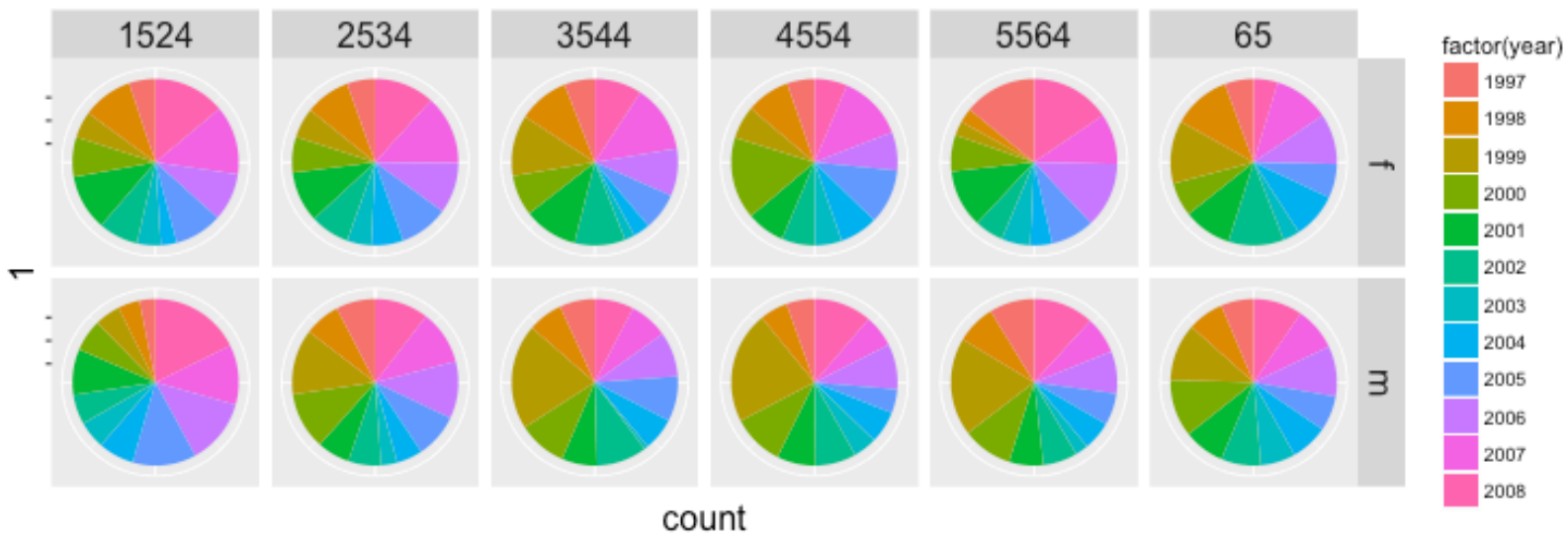
```
ggplot(tb_au, aes(x = 1, y = count, fill = factor(year))) +
  geom_bar(stat = "identity", position="fill") +
  facet_grid(gender ~ age) +
  theme(
    axis.text = element_blank(),
    strip.text = element_text(size = 16),
    axis.title = element_text(size = 16)
  )
```



Its a single stacked bar, in each facet.

Pie charts

```
ggplot(tb_au, aes(x = 1, y = count, fill = factor(year))) +
  geom_bar(stat = "identity", position="fill") +
  facet_grid(gender ~ age) +
  theme(
    axis.text = element_blank(),
    strip.text = element_text(size = 16),
    axis.title = element_text(size = 16)
  ) + coord_polar(theta="y")
```



Data - Autism

```
library(HLMdiag)
data(autism)
glimpse(autism)
Observations: 604
Variables: 7
$ childid  <int> 1, 1, 1, 1, 1, 10, 10, 10, 10, 100, 100, 100, 100, 10...
$ sicdegp  <fctr> high, high, high, high, high, low, low, low, low, hi...
$ age2     <dbl> 0, 1, 3, 7, 11, 0, 1, 7, 11, 0, 1, 3, 7, 0, 1, 7, 11,...
$ vsae     <int> 6, 7, 18, 25, 27, 9, 11, 18, 39, 15, 24, 37, 135, 8, ...
$ gender   <fctr> male, male, male, male, male, male, male, male, male...
$ race     <fctr> white, white, white, white, white, white, white, whi...
$ bestest2 <fctr> pdd, pdd, pdd, pdd, pdd, autism, autism, autism, aut...
```

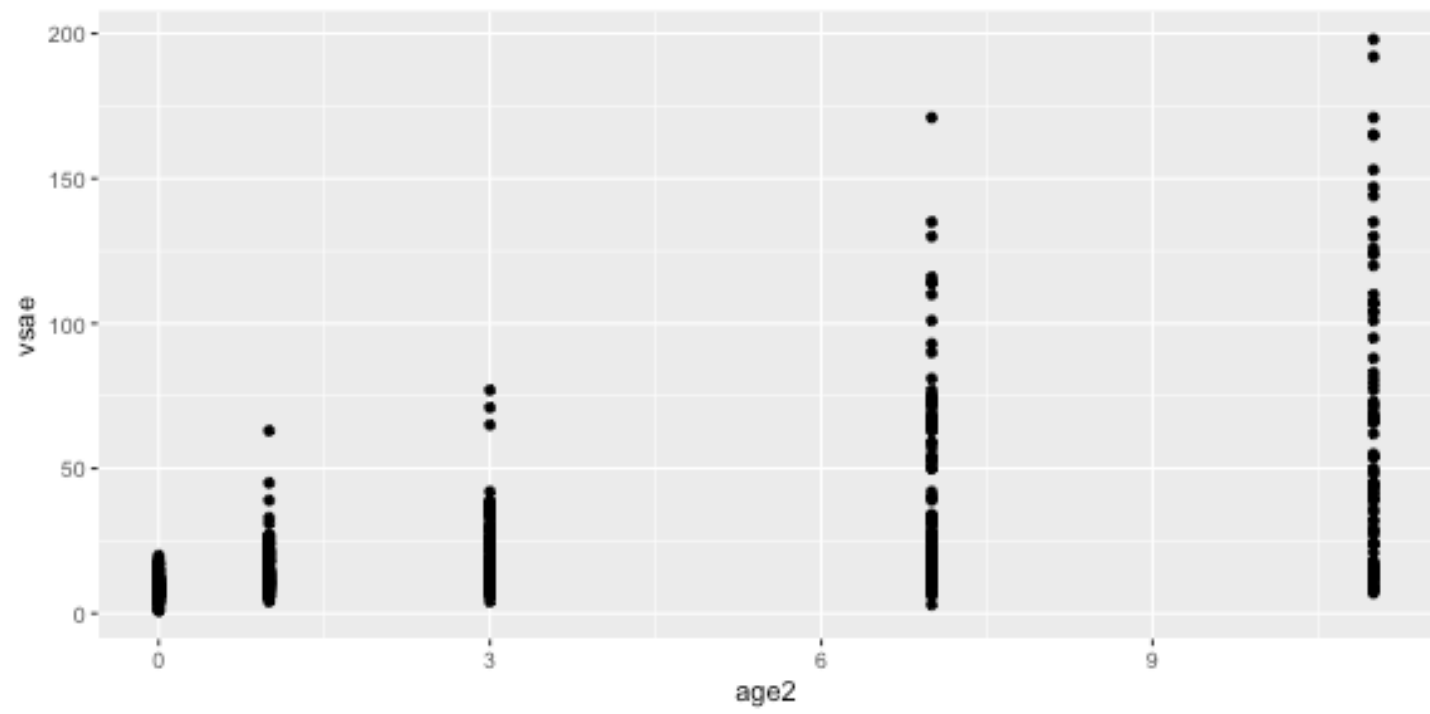
Repeated measurements (panel data, longitudinal data) for each subject. Need to examine within subject dependence, relative to between subject, and between demographic group.

Ji

gg

Plotting points

```
ggplot(autism, aes(x=age2, y=vsae)) +
  geom_point()
```

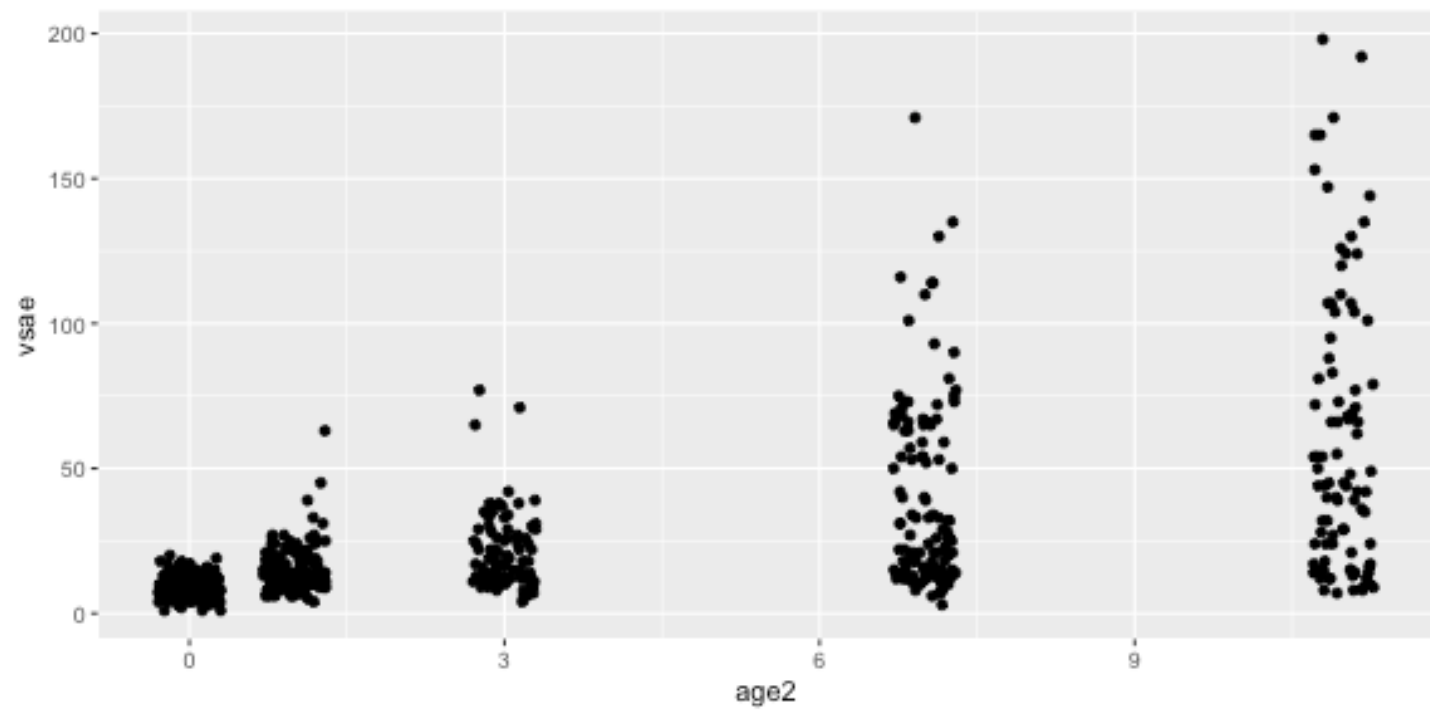


A

Jittering points

gg

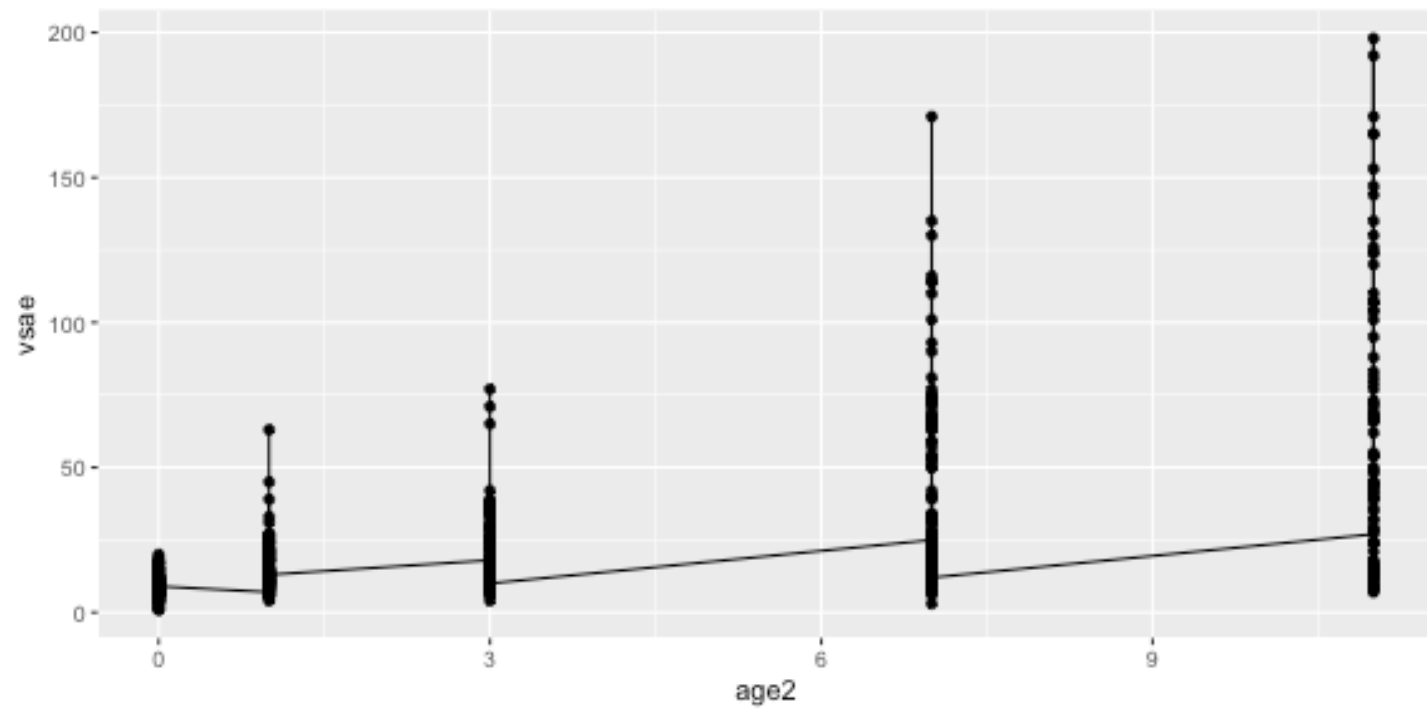
```
ggplot(autism, aes(x=age2, y=vsae)) +
  geom_jitter(width=0.3, height=0)
```



No

Adding lines

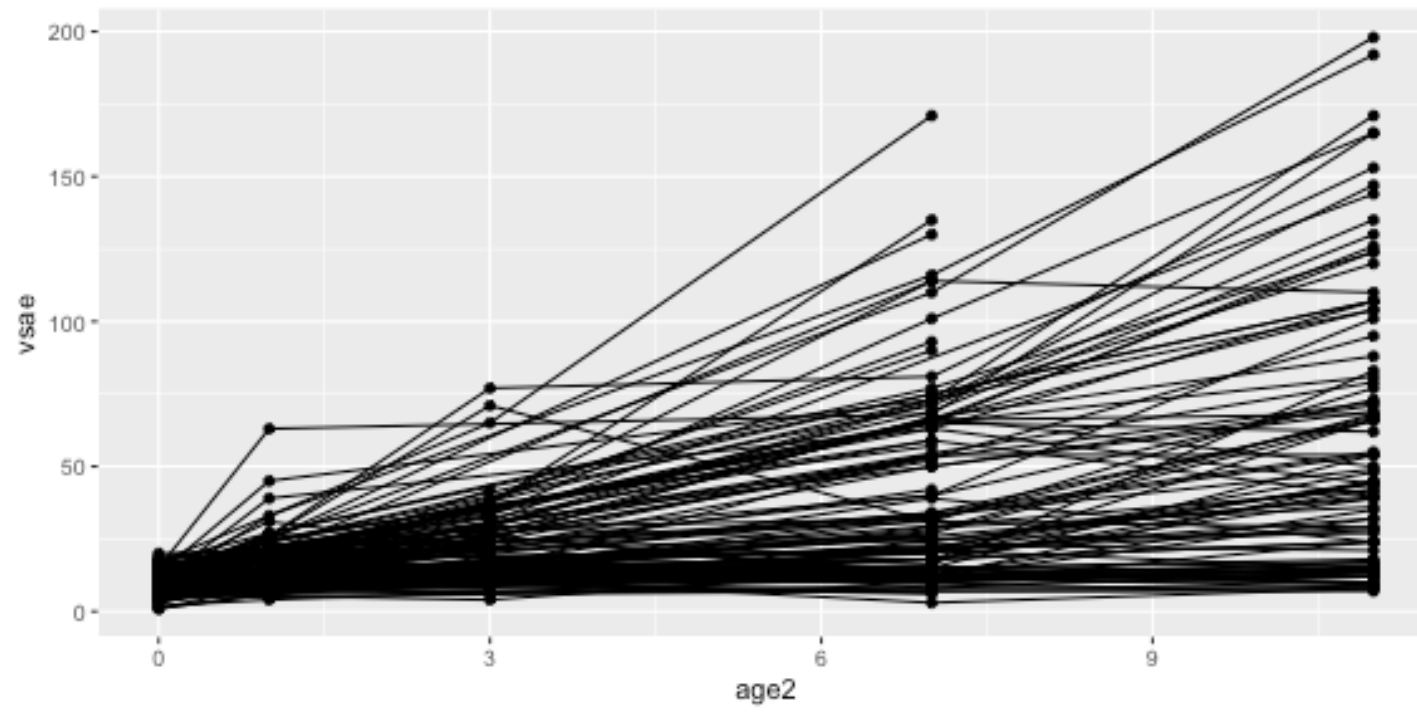
```
ggplot(autism, aes(x=age2, y=vsae)) +
  geom_point() + geom_line()
```



Not the lines we want!

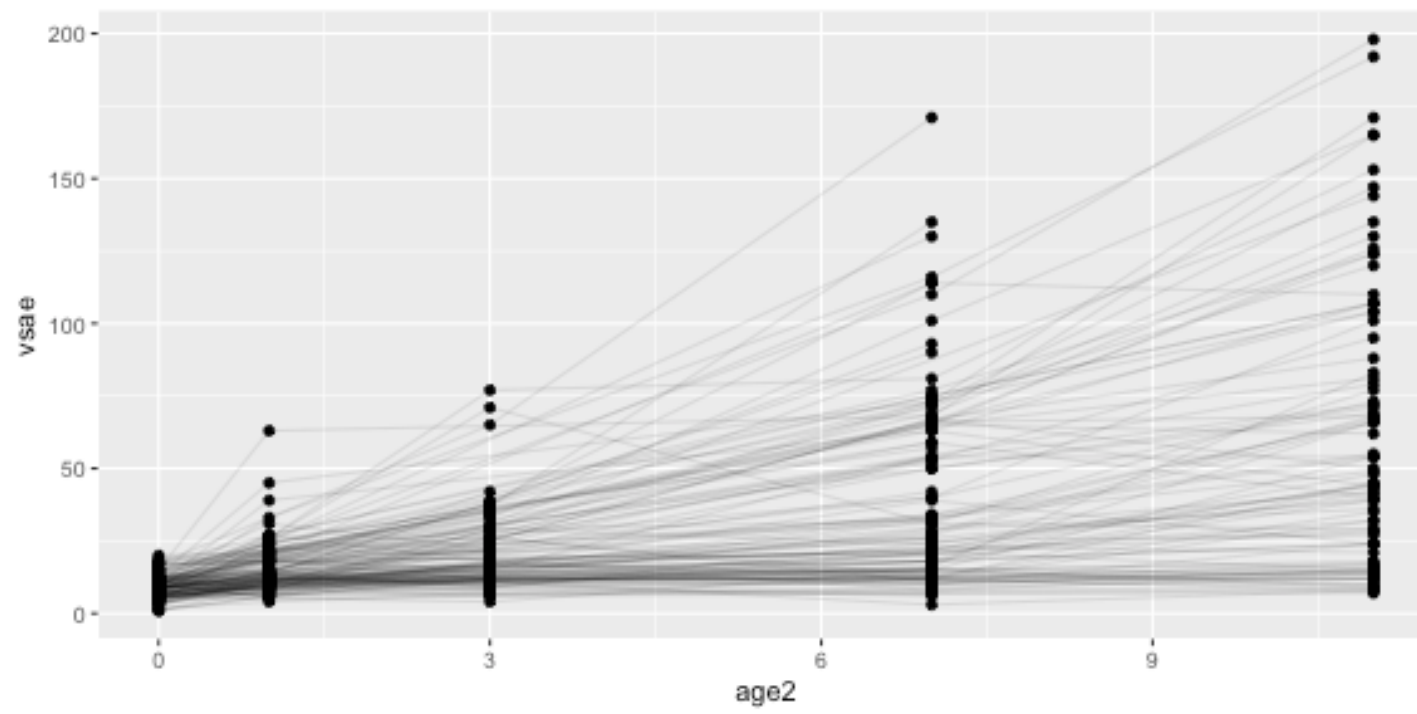
These are the lines we want

```
ggplot(autism, aes(x=age2, y=vsae, group=childid)) +
  geom_point() + geom_line()
```



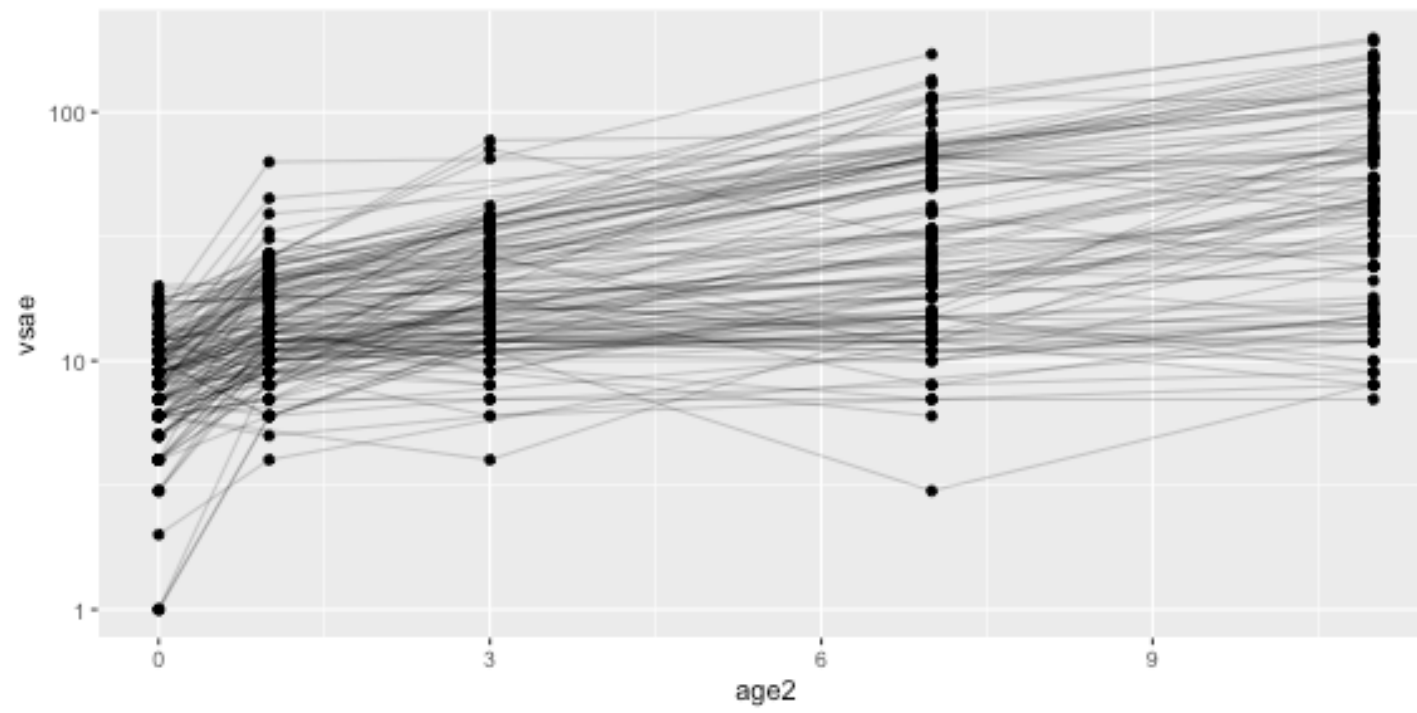
Too much ink

```
ggplot(autism, aes(x=age2, y=vsae, group=childid)) +
  geom_point() + geom_line(alpha=0.1)
```



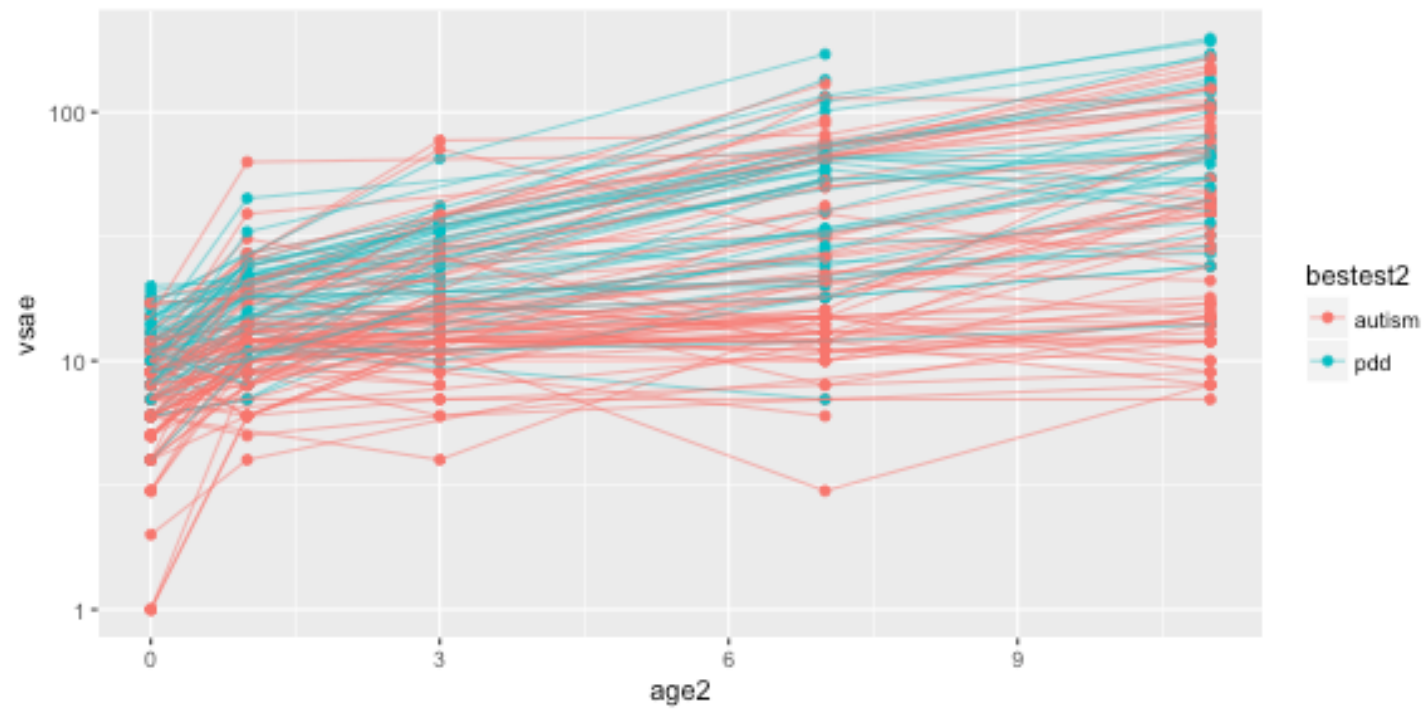
Log scale y

```
ggplot(autism, aes(x=age2, y=vsae, group=childid)) +
  geom_point() + geom_line(alpha=0.2) + scale_y_log10()
```



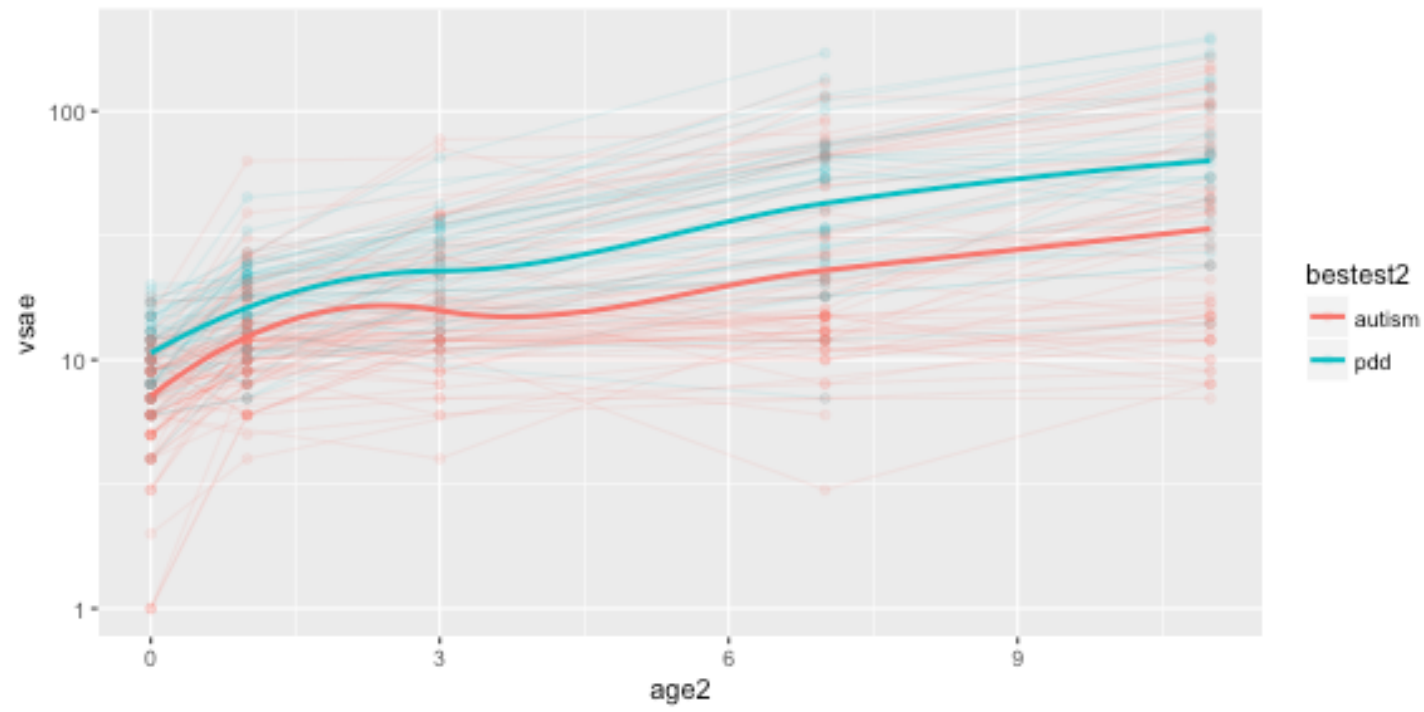
By diagnosis at age 2

```
ggplot(autism, aes(x=age2, y=vsae, group=childid, colour=bestest2)) +
  geom_point() + geom_line(alpha=0.5) + scale_y_log10()
```



Refine groups

```
ggplot(autism, aes(x=age2, y=vsae, colour=bestest2)) +
  geom_point(alpha=0.1) + geom_line(aes(group=childid), alpha=0.1) +
  geom_smooth(se=F) +
  scale_y_log10()
```



A

```
gg
```

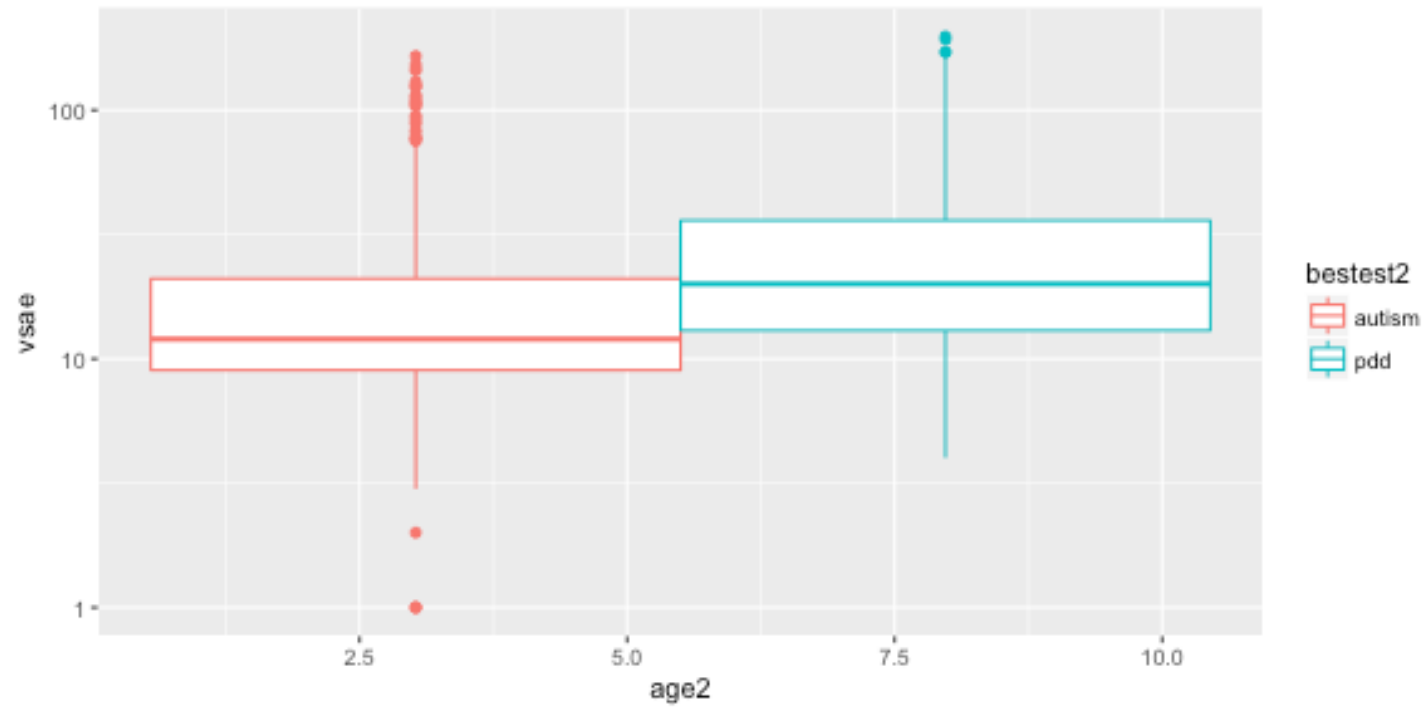
Tha

Your turn

What do we learn about autism, age, and the diagnosis at age 2?

A different look

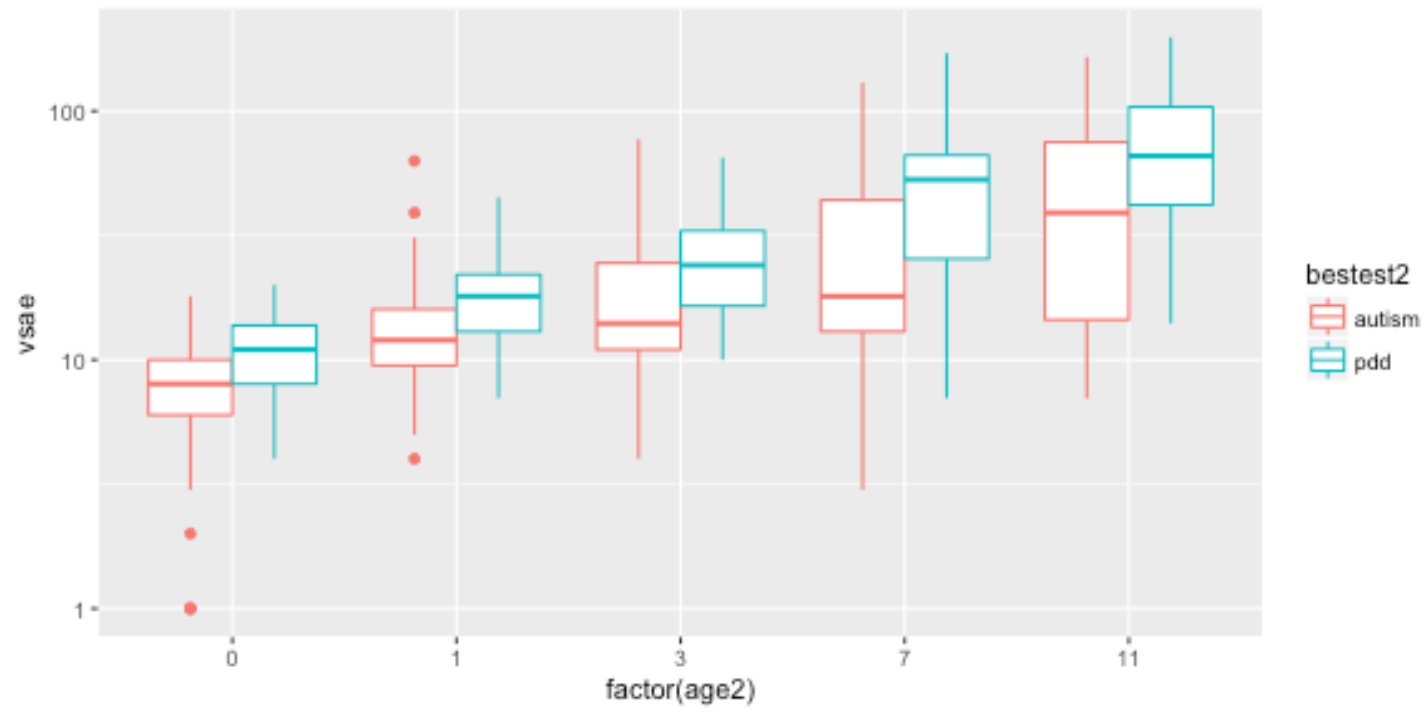
```
ggplot(autism, aes(x=age2, y=vsae, colour=bestest2)) +
  geom_boxplot() + scale_y_log10()
```



That's not what I wanted ....

For each age measured

```
ggplot(autism, aes(x=factor(age2), y=vsae, colour=bestest2)) +
  geom_boxplot() + scale_y_log10()
```



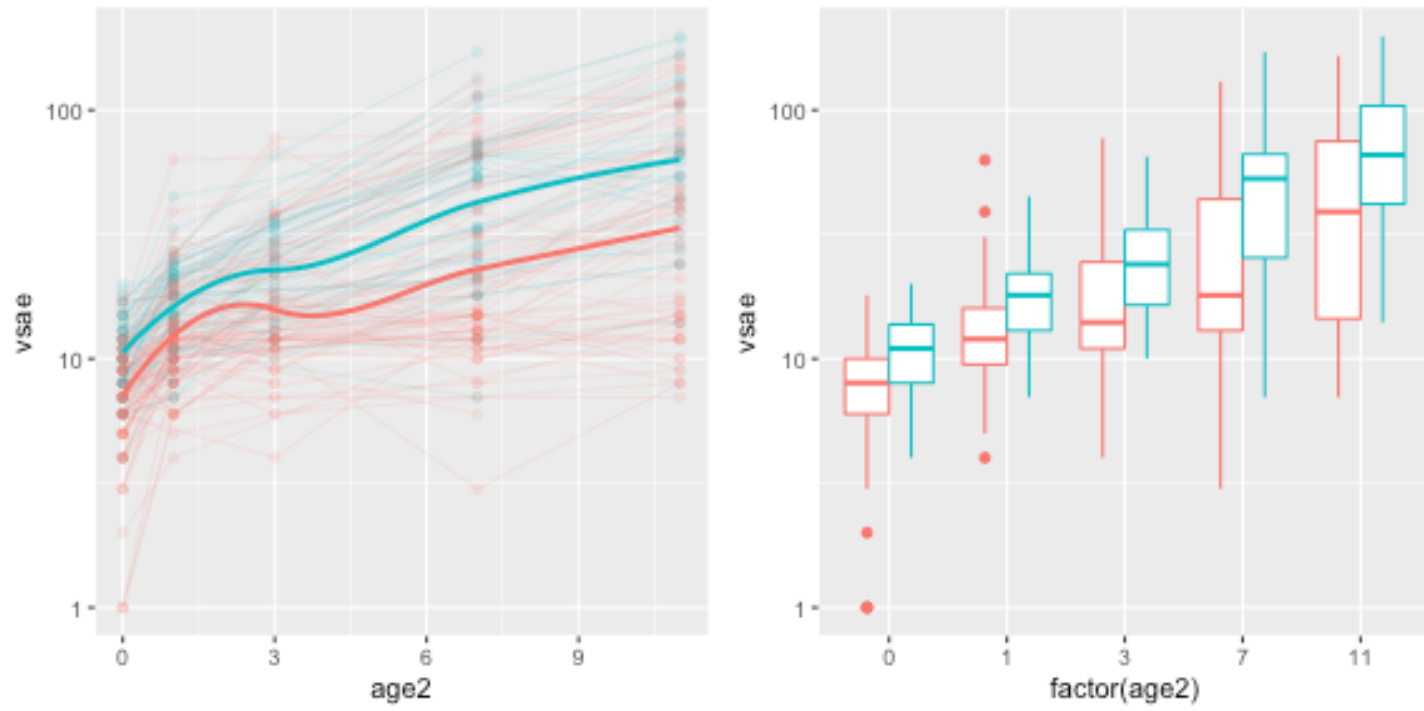
N

41%

Obs
Var
$ R
$ H
$ D
$ H
$ D
$ I
$ I
$ W
$ I
$ G
$ C
$ U
$ I
$ G
$ I
$ I
$ I
$ I
$ I
$ I
$ H
$ Have you ever smoked a cigarette in an airplane bathroom when it was against the rul

Which is better?

```
p1 <- ggplot(autism, aes(x=age2, y=vsae, colour=bestest2)) +
  geom_point(alpha=0.1) + geom_line(aes(group=childid), alpha=0.1) +
  geom_smooth(se=F) +
  scale_y_log10() + theme(legend.position="none")
p2 <- ggplot(autism, aes(x=factor(age2), y=vsae, colour=bestest2)) +
  geom_boxplot() + scale_y_log10() + theme(legend.position="none")
grid.arrange(p1, p2, ncol=2)
```



New example - Flying etiquette

41% Of Fliers Think You're Rude If You Recline Your Seat

```
Observations: 1,040
Variables: 27
$ RespondentID
$ How often do you travel by plane?
$ Do you ever recline your seat when you fly?
$ How tall are you?
$ Do you have any children under 18?
$ In a row of three seats, who should get to use the two arm rests?
$ In a row of two seats, who should get to use the middle arm rest?
$ Who should have control over the window shade?
$ Is itrude to move to an unsold seat on a plane?
$ Generally speaking, is it rude to say more than a few words tothe stranger sitting n
$ On a 6 hour flight from NYC to LA, how many times is it acceptable to get up if you'
$ Under normal circumstances, does a person who reclines their seat during a flight ha
$ Is itrude to recline your seat on a plane?
$ Given the opportunity, would you eliminate the possibility of reclining seats on pla
$ Is it rude to ask someone to switch seats with you in order to be closer to friends?
$ Is itrude to ask someone to switch seats with you in order to be closer to family?
$ Is it rude to wake a passenger up if you are trying to go to the bathroom?
$ Is itrude to wake a passenger up if you are trying to walk around?
$ In general, is itrude to bring a baby on a plane?
$ In general, is it rude to knowingly bring unruly children on a plane?
$ Have you ever used personal electronics during take off or landing in violation of a
$ Have you ever smoked a cigarette in an airplane bathroom when it was against the rul
```

S

gg

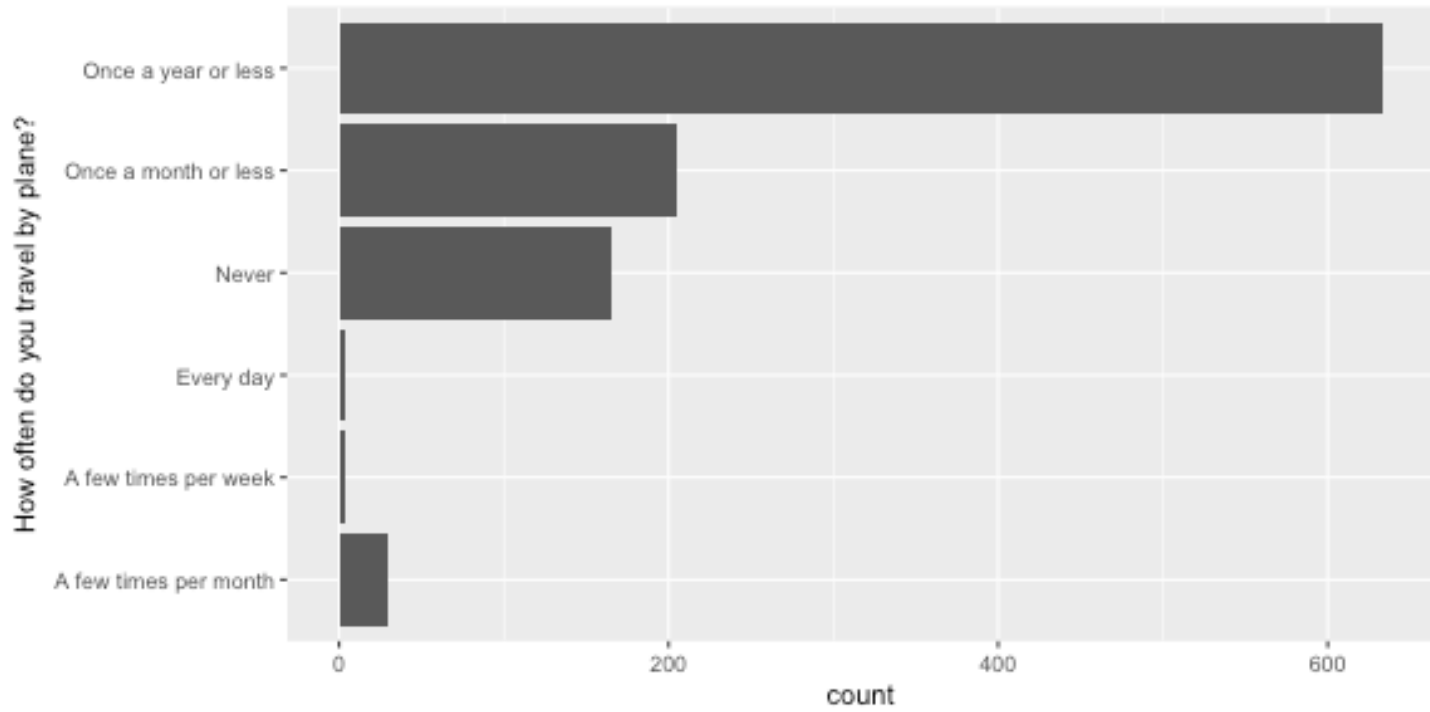
Variables

- Mix of categorical and quantitative variables.
- What mappings are appropriate?
- Area for counts of categories,
- side-by-side boxplots for mixed pair.

Cat

Support

```
ggplot(fly, aes(x=`How often do you travel by plane?`)) +
  geom_bar() + coord_flip()
```

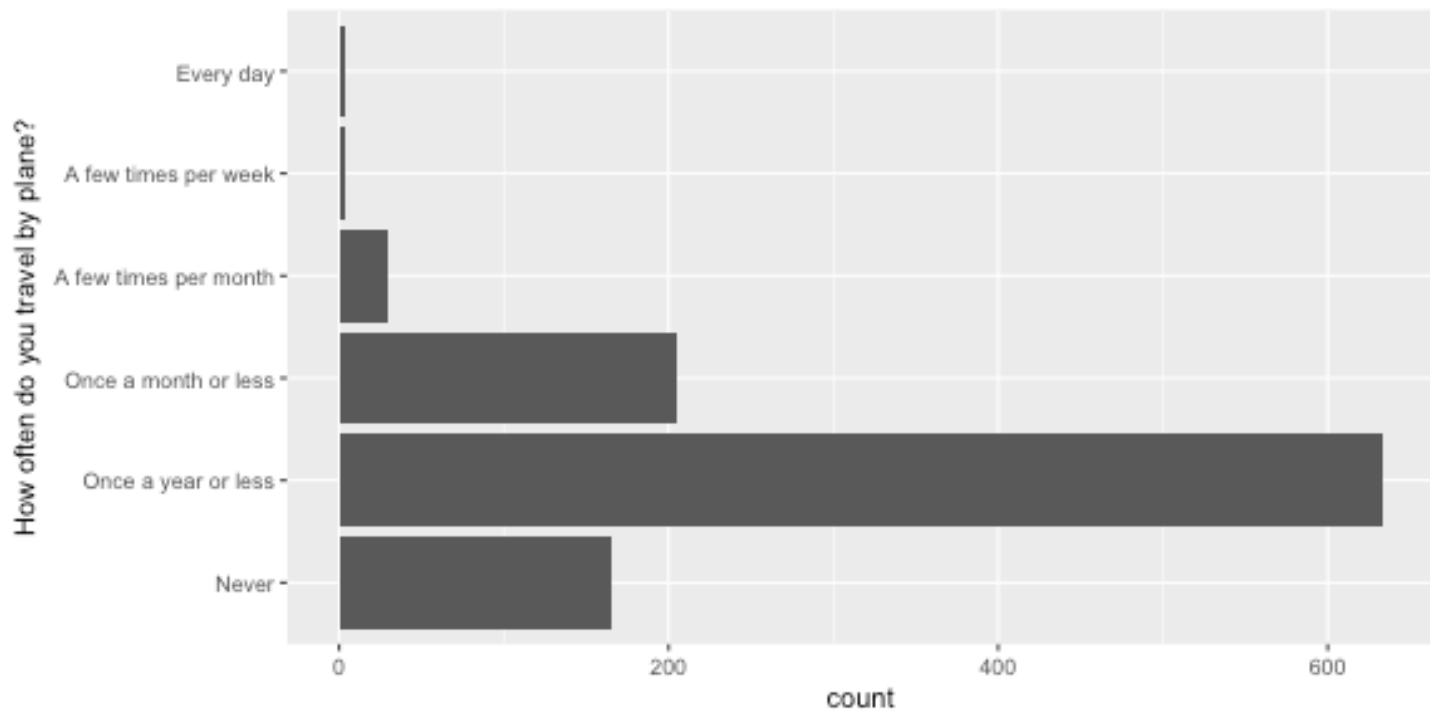


Categories are not sorted

gg

Sorted categories

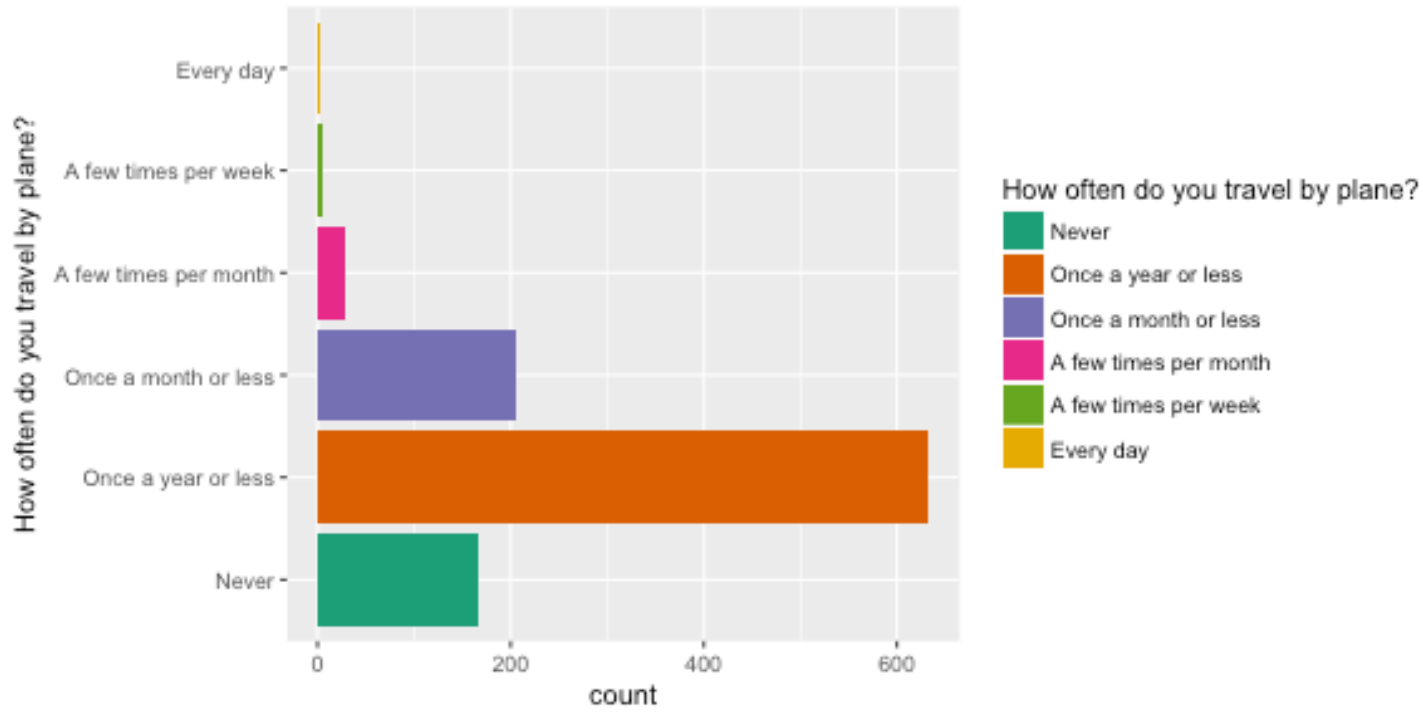
```
fly$`How often do you travel by plane?` <-
  factor(fly$`How often do you travel by plane?`, levels=c(
    "Never","Once a year or less","Once a month or less",
    "A few times per month","A few times per week","Every day"))
ggplot(fly, aes(x=`How often do you travel by plane?`)) + geom_bar() + coord_
```



Fi

```
ggplot(fly, aes(x=`How often do you travel by plane?`,
                fill=`How often do you travel by plane?`)) + geom_bar() + coo
  scale_fill_brewer(palette="Dark2")
```

```
fl
```

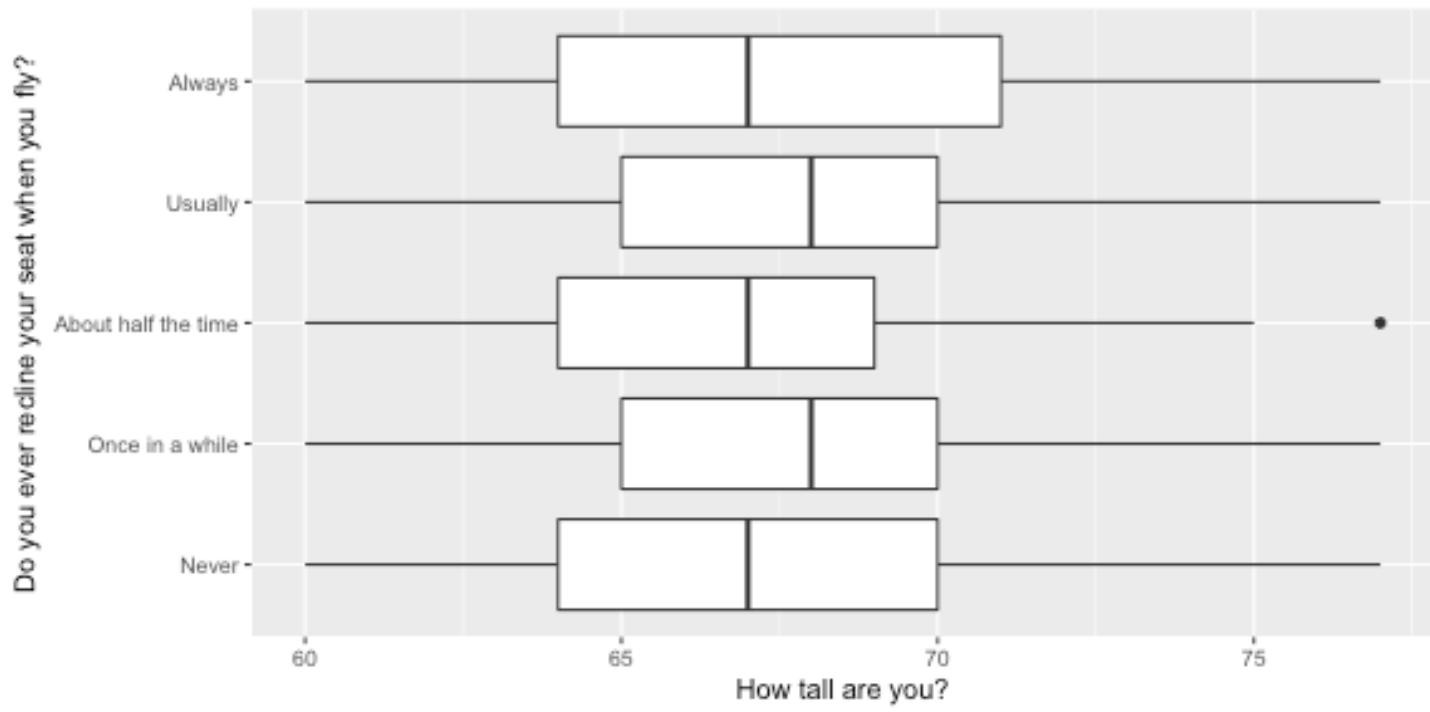


Filter data

```
fly_sub <- fly %>% filter(`How often do you travel by plane?` %in%
                            c("Once a year or less","Once a month or less"))
  filter(!is.na(`Do you ever recline your seat when you fly?`)) %>%
  filter(!is.na(Age)) %>% filter(!is.na(Gender))
```

Recline by height

```
fly_sub$`Do you ever recline your seat when you fly?` <- factor(
  fly_sub$`Do you ever recline your seat when you fly?`, levels=c(
    "Never","Once in a while","About half the time",
    "Usually","Always"))
ggplot(fly_sub, aes(y=`How tall are you?`, x=`Do you ever recline your seat w
```



Yc

Wh

Your turn

What is the difference between `colour` and `fill`?

Yc

Wh

Your turn

What is the difference between `colour` and `fill`?

- colour is for 0 or 1-dimensional elements, and

Co

Wh

the

Your turn

What is the difference between `colour` and `fill`?

- colour is for 0 or 1-dimensional elements, and
- fill is for area (2-d) geoms

Coordinate systems

What does `coord_fixed()` do? What is the difference between this and using `theme(aspect.ratio=...)`?

Coordinate systems

What does `coord_fixed()` do? What is the difference between this and using `theme(aspect.ratio=...)`?

- coord_fixed operates on the raw data values, but

43 / 62

Fa

gg

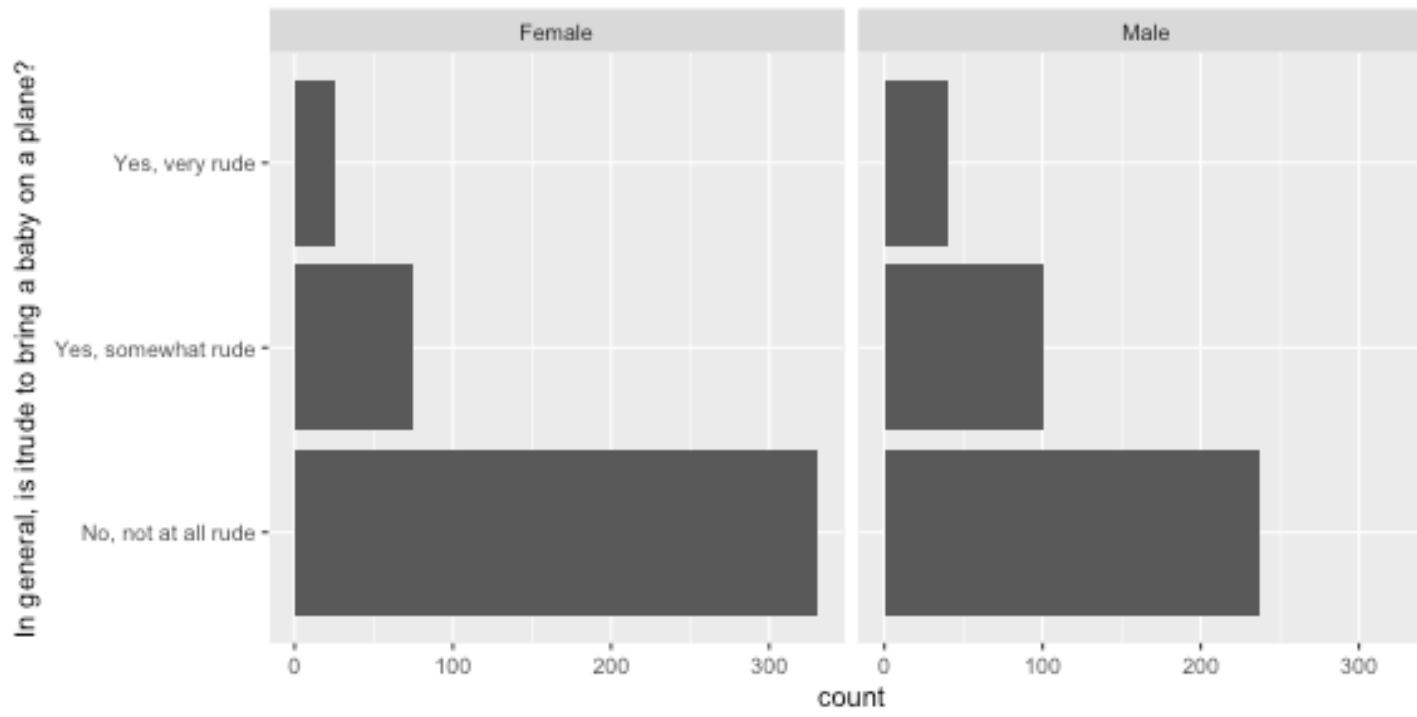
Coordinate systems

What does `coord_fixed()` do? What is the difference between this and using `theme(aspect.ratio=...)`?

- coord_fixed operates on the raw data values, but
- theme(aspect_ratio=...) works on the plot dimensions

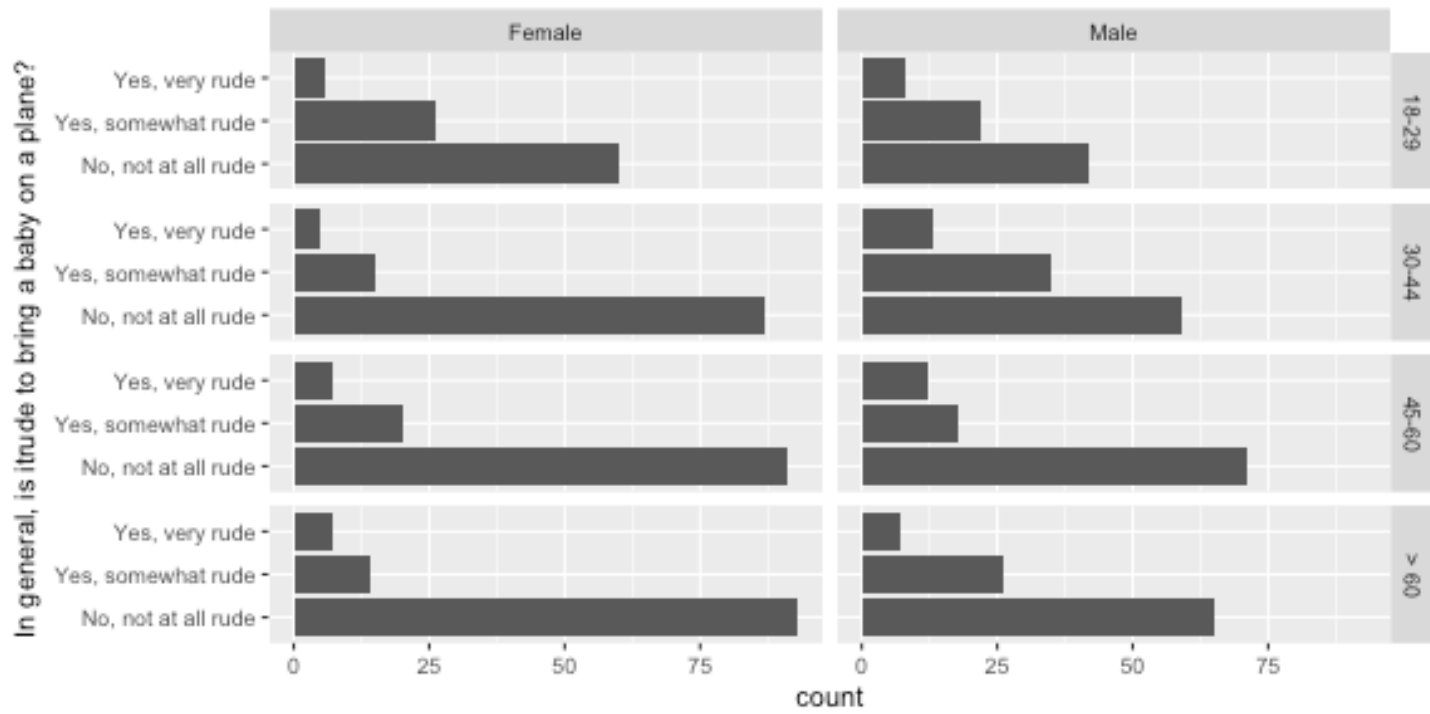
Facets

```
ggplot(fly_sub,
       aes(x=`In general, is itrude to bring a baby on a plane?`)) +
  geom_bar() + coord_flip() + facet_wrap(~Gender)
```



Facets

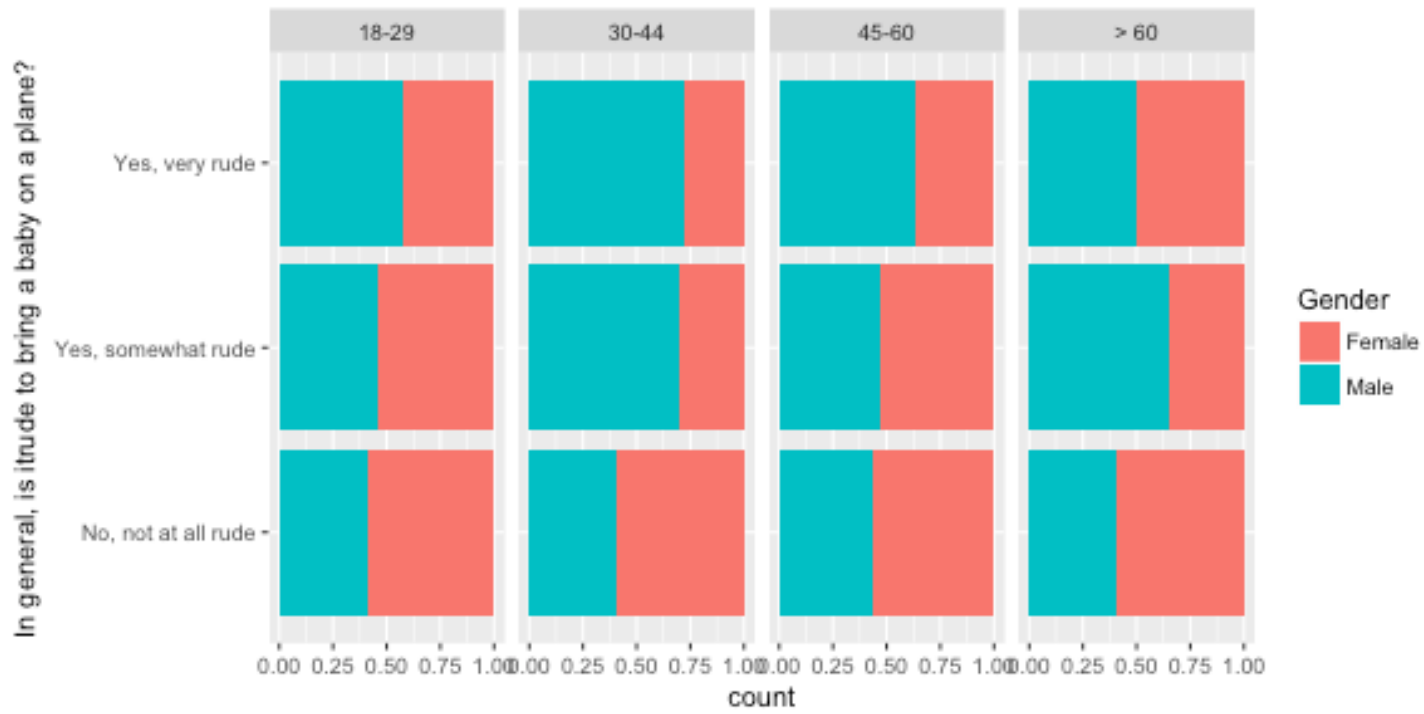
```
fly_sub$Age <- factor(fly_sub$Age, levels=c("18-29","30-44","45-60","> 60"))
ggplot(fly_sub, aes(x=`In general, is itrude to bring a baby on a plane?`)) +
  geom_bar() + coord_flip() + facet_grid(Age~Gender)
```



Color palettes - default

```
p
```

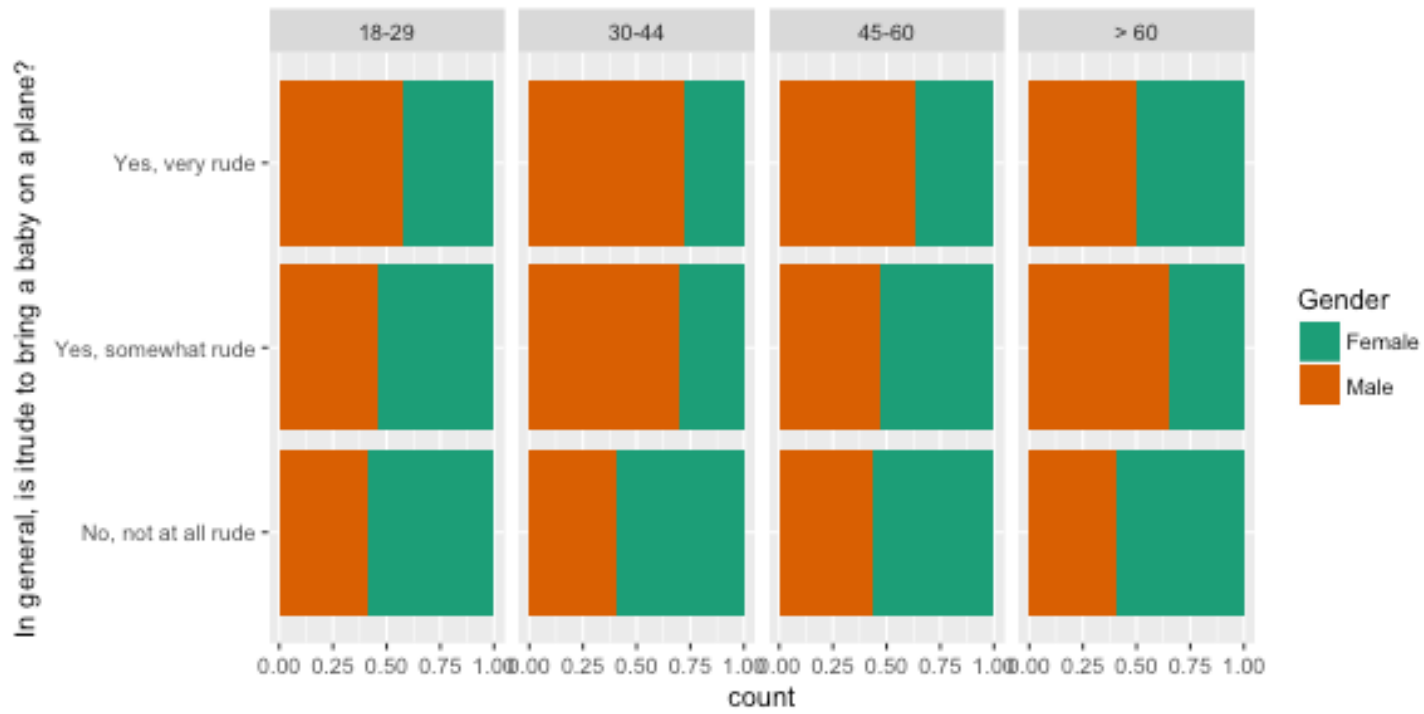
```
p <- ggplot(fly_sub, aes(x=`In general, is itrude to bring a baby on a plane?
                         fill=Gender)) +
  geom_bar(position="fill") + coord_flip() + facet_wrap(~Age, ncol=5)
p
```



What do we learn?

Color palettes - brewer

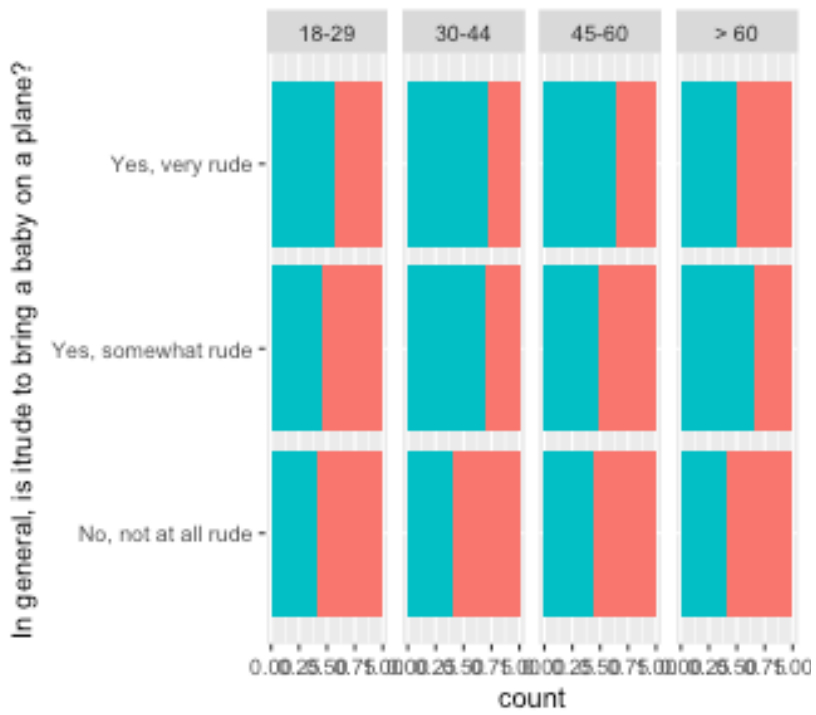
```
p + scale_fill_brewer(palette="Dark2")
```

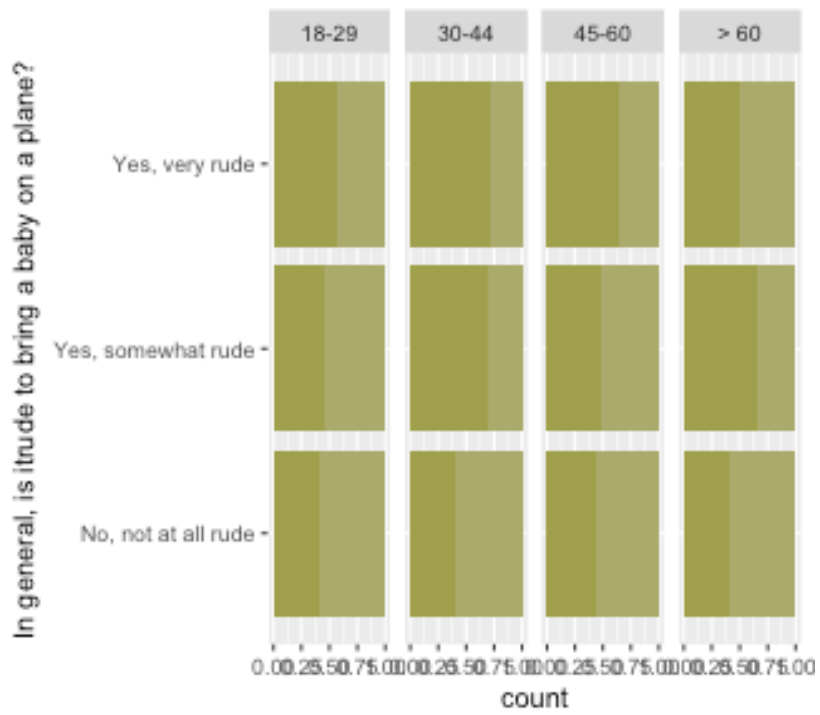


Color blind-proofing

```
library(scales)
library(dichromat)
clrs <- hue_pal()(3)
p + theme(legend.position = "none")
clrs <- dichromat(hue_pal()(3))
p + scale_fill_manual("", values=clrs) + theme(legend.position = "none")
```

P


18-29
30-44
45-60
> 60
In general, is itrude to bring a baby on a plane?
Yes, very rude
Yes, somewhat rude
No, not at all rude
0.00 0.25 0.50 0.75 1.00
count

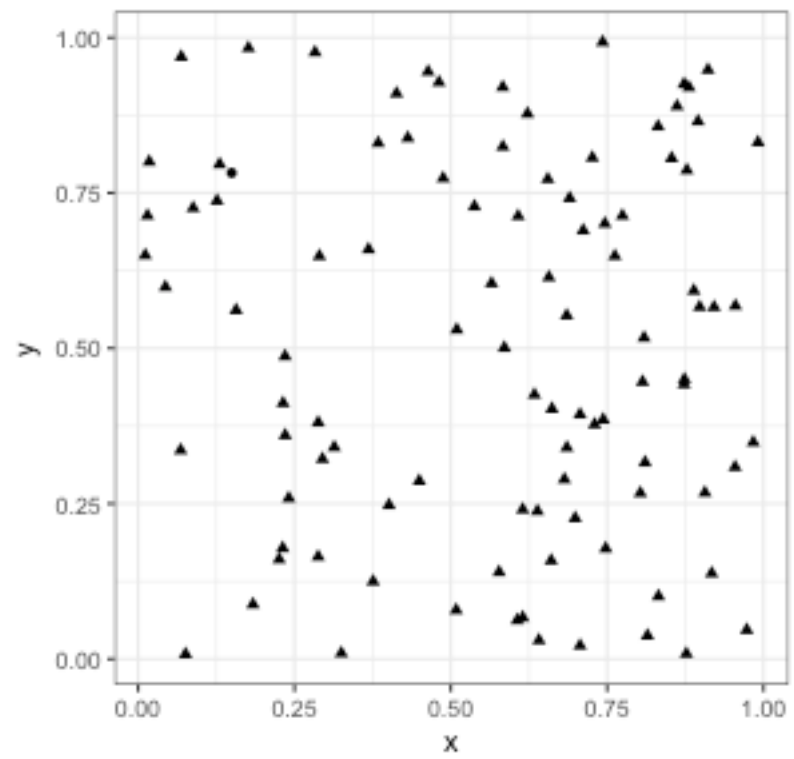

18-29
30-44
45-60
> 60
In general, is itrude to bring a baby on a plane?
Yes, very rude
Yes, somewhat rude
No, not at all rude
0.00 0.25 0.50 0.75 1.00
count

H Perceptual principles

- Hierarchy of mappings: (first) position along an axis - (last) color (Cleveland, 1984; Heer and Bostock, 2009)
- Pre-attentive: Some elements are noticed before you even realise it.
- Color: (pre-attentive) palettes - qualitative, sequential, diverging.
- Proximity: Place elements for primary comparison close together.
- Change blindness: When focus is interrupted differences may not be noticed.

Pre-attentive

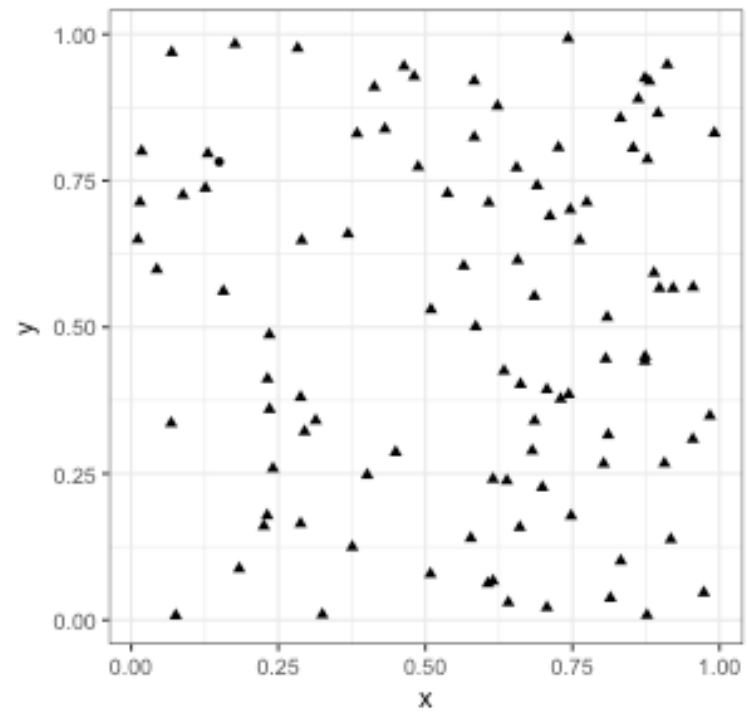
Can you find the odd one out?



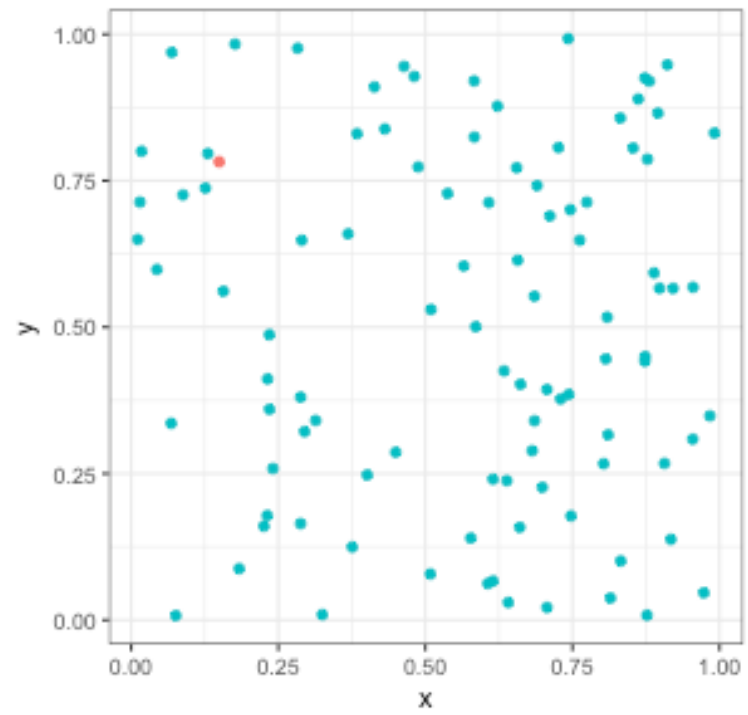
Is it

Pre-attentive

Can you find the odd one out?



Is it easier now?

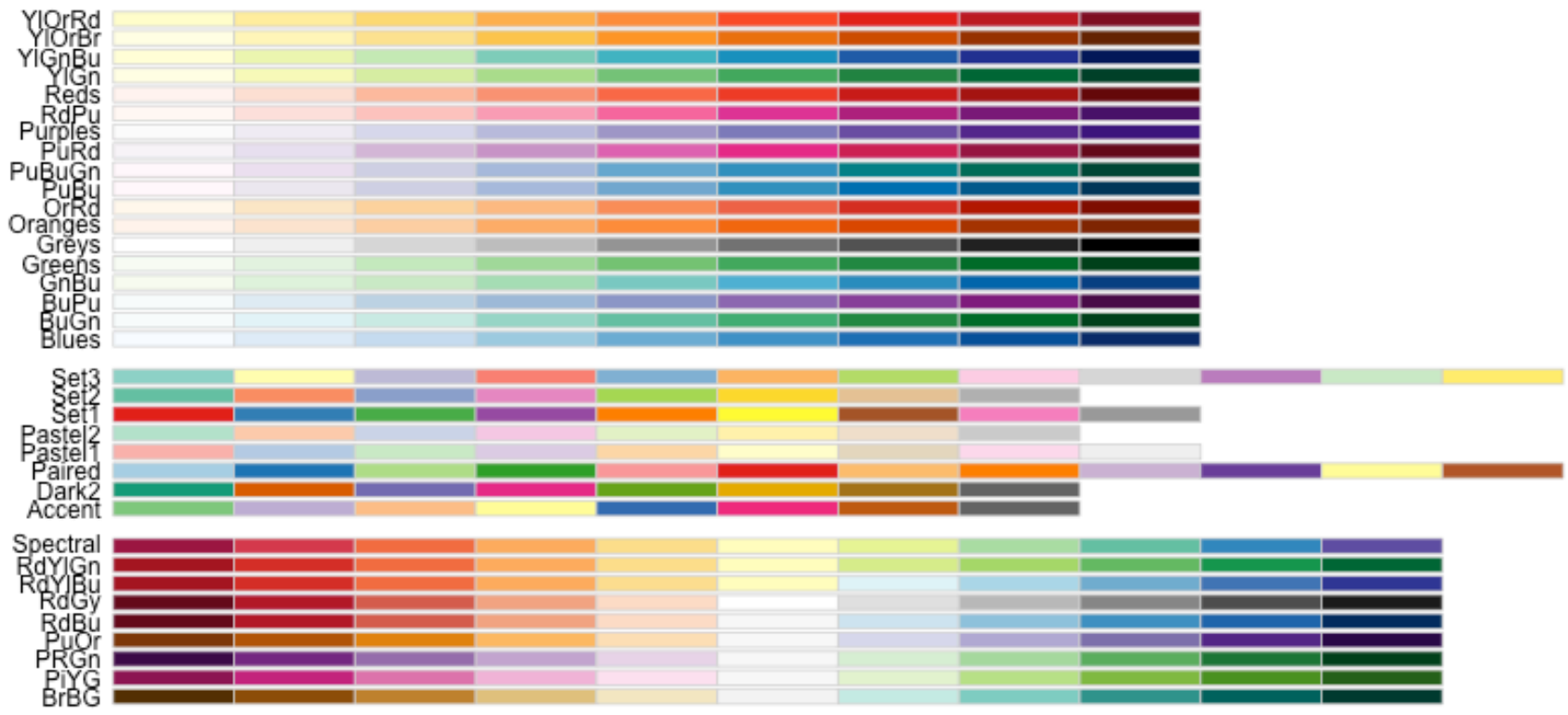


Color palettes

- Qualitative: categorical variables
- Sequential: low to high numeric values
- Diverging: negative to positive values

P

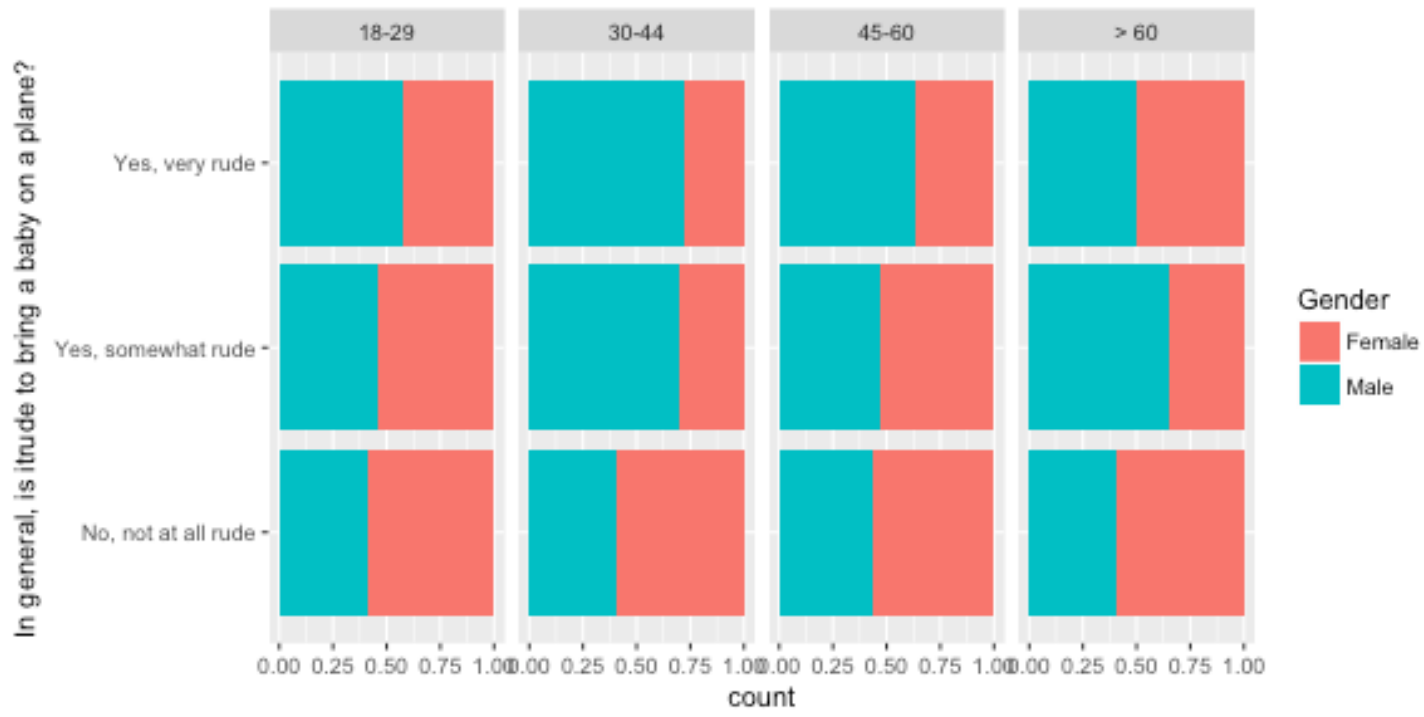
```
gg
```


YlOrRd
YlOrBr
YlGnBu
YlGn
Reds
RdPu
Purples
PuRd
PuBuGn
PuBu
OrRd
Oranges
Greys
Greens
GnBu
BuPu
BuGn
Blues
Set3
Set2
Set1
Pastel2
Pastel1
Paired
Dark2
Accent
Spectral
RdYlGn
RdYlBu
RdGy
RdBu
PuOr
PRGn
PiYG
BrBG

Wit
anc

Proximity

```
ggplot(fly_sub, aes(x=`In general, is itrude to bring a baby on a plane?`,
                    fill=Gender)) +
  geom_bar(position="fill") + coord_flip() + facet_wrap(~Age, ncol=5)
```



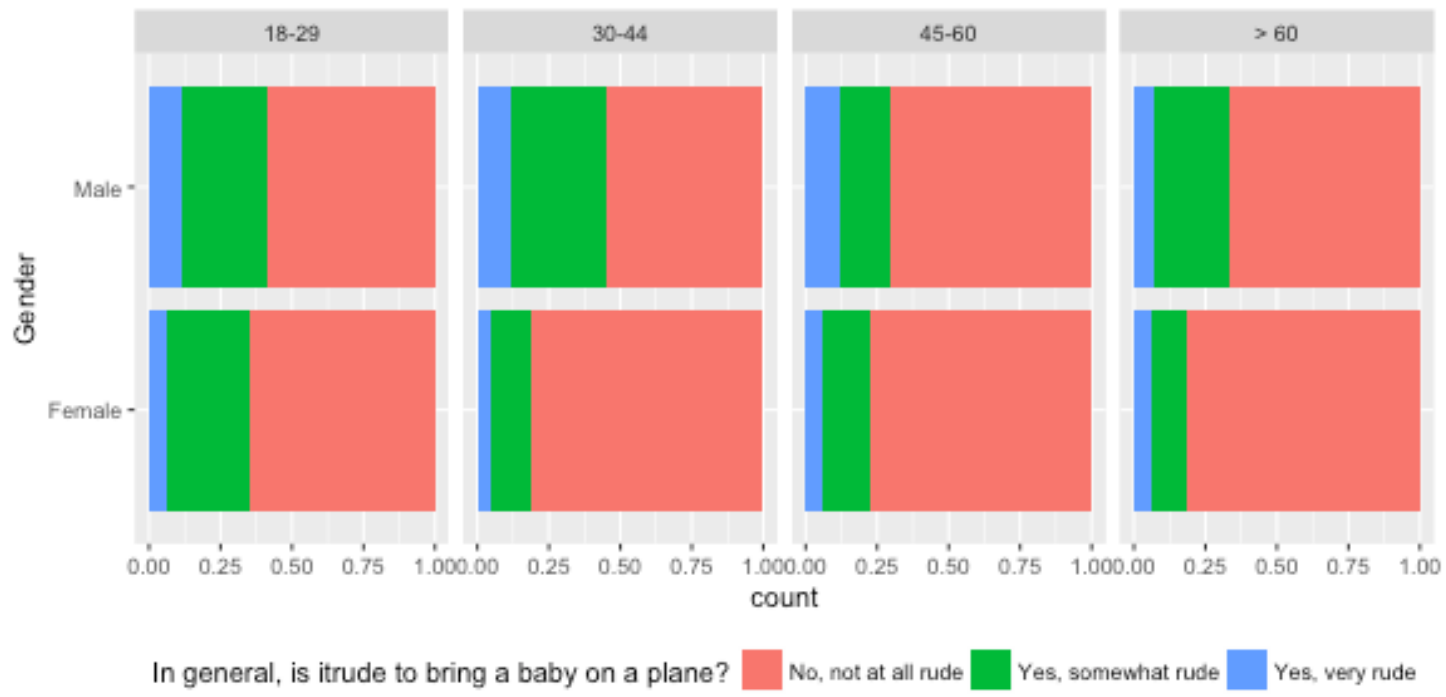
With this arrangement we can see proportion of gender within each rudeness category, and compare these across age groups. How could we arrange this differently?

A

Proximity

gg

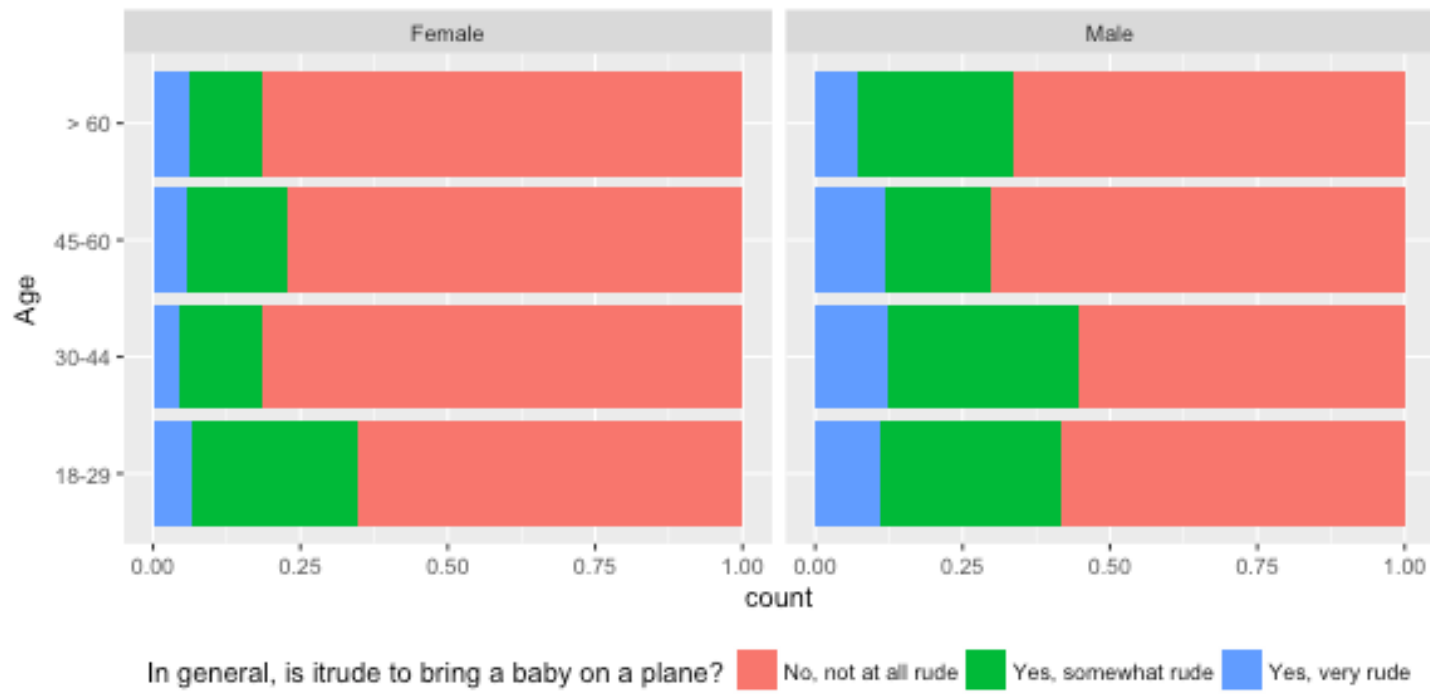
```
ggplot(fly_sub, aes(x=Gender,
   fill=`In general, is itrude to bring a baby on a plane?`)) +
  geom_bar(position="fill") + coord_flip() + facet_wrap(~Age, ncol=5) +
  theme(legend.position="bottom")
```



What is different about the comparison now?

Another arrangement

```
ggplot(fly_sub, aes(x=Age,
                    fill=`In general, is itrude to bring a baby on a plane?`)
  geom_bar(position="fill") + coord_flip() + facet_wrap(~Gender, ncol=5) +
  theme(legend.position="bottom")
```



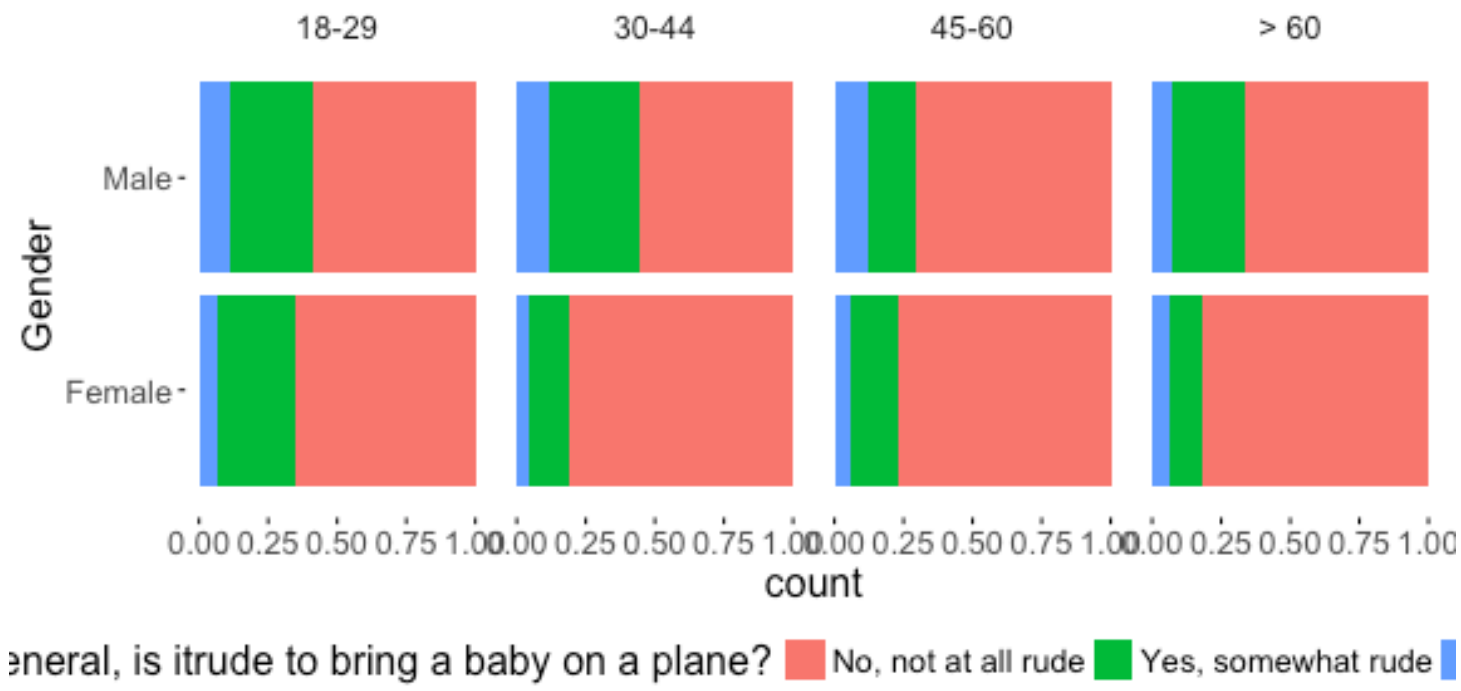
Themes

The `ggthemes` package has many different styles for the plots. Other packages such as `xkcd`, `skittles`, `wes anderson`, `beyonce`, ...

```
library(xkcd)
ggplot(fly_sub, aes(x=Gender,
                    fill=`In general, is itrude to bring a baby on a plane?`)
  geom_bar(position="fill") + coord_flip() + facet_wrap(~Age, ncol=5) +
  theme_xkcd() + theme(legend.position="bottom")
```

See the vignette for instructions on installing the xkcd font.

R


18-29
30-44
45-60
> 60
Male
Female
Gender
0.00 0.25 0.50 0.75 1.00
count
eneral, is itrude to bring a baby on a plane?
No, not at all rude
Yes, somewhat rude

Resources

- Winston Chang (2012) Cookbook for R
- Antony Unwin (2014) Graphical Data Analysis
- Naomi Robbins (2013) Creating More Effective Charts

Share and share alike