PSY 503: Foundations of Statistical Methods in Psychological Science

Statistical Models,

Probability

Suyog Chandramouli

Zoom & 311 PSH (Princeton University)

29th September, 2025

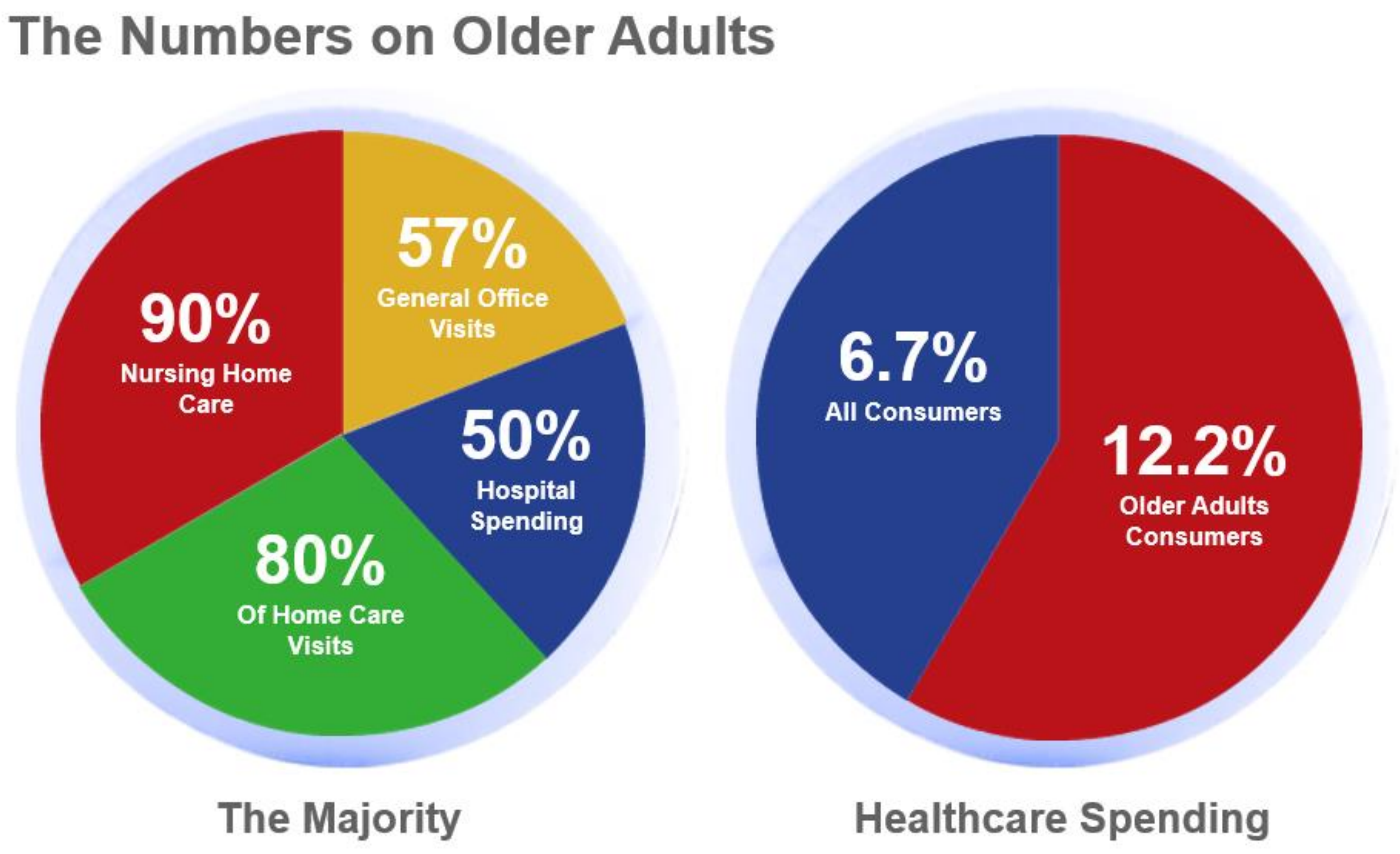
Visualization Principles

Visualization Principles

- When would you call a graph or a visualization “bad”?

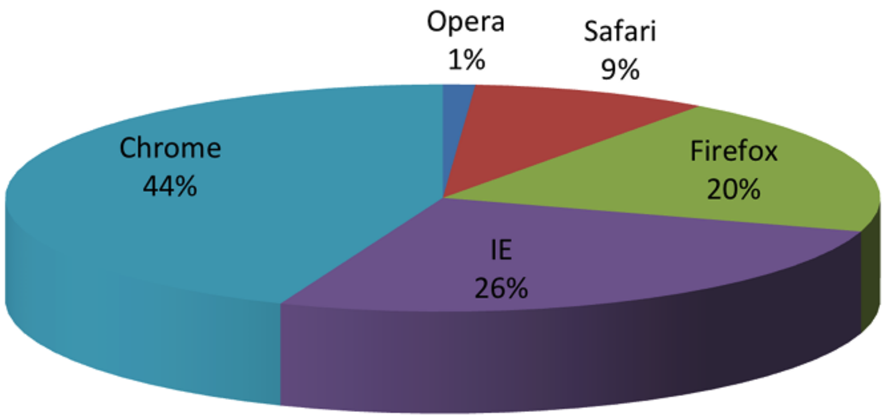
Examples of bad visualization

From: https://www.reddit.com/r/shittydataisbeautiful/



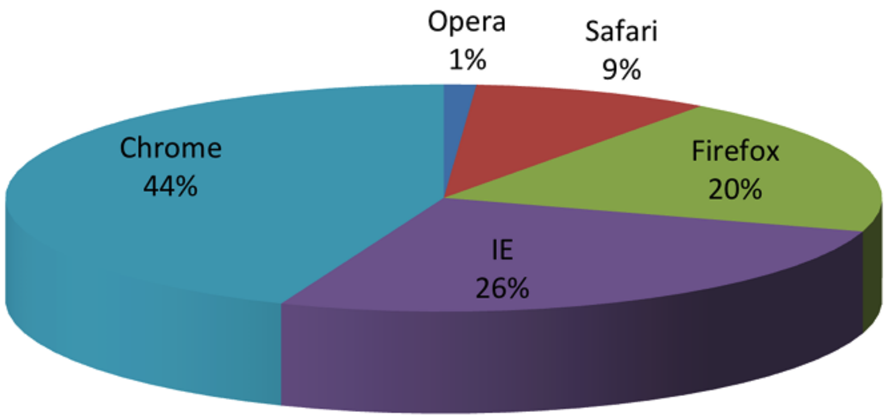
Examples of bad visualization

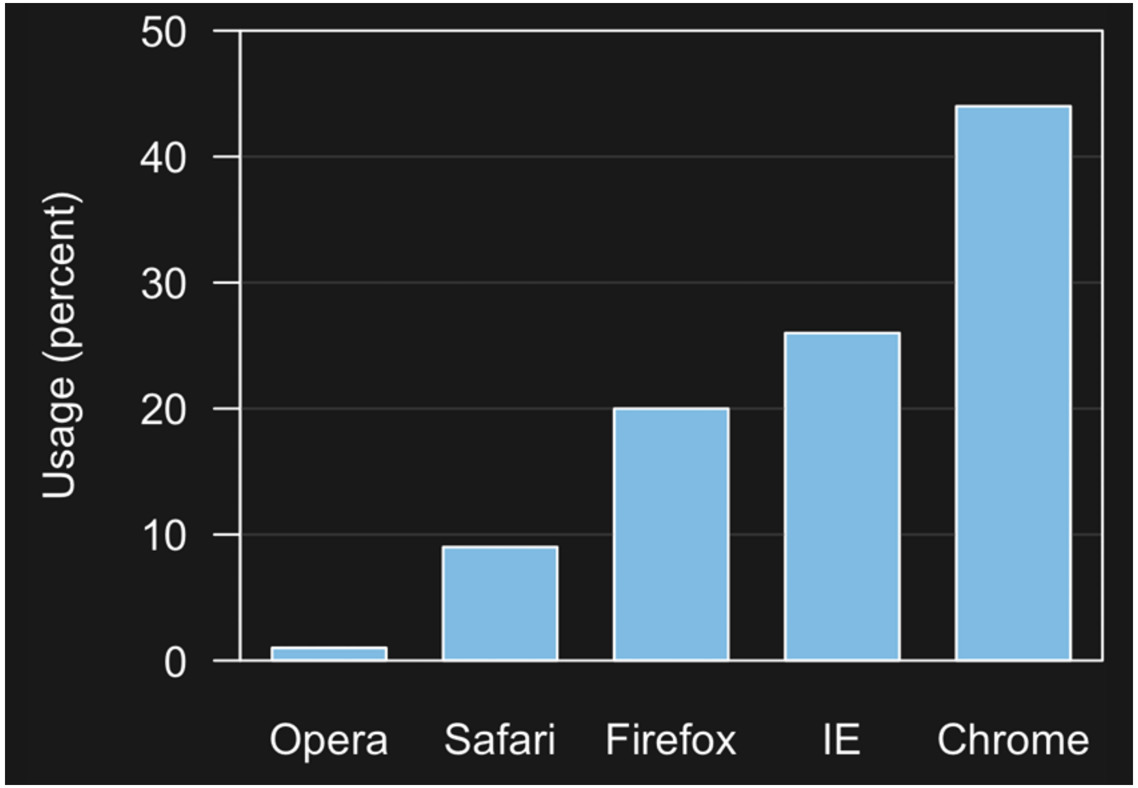
From: https://github.com/kbroman/Talk_Graphs



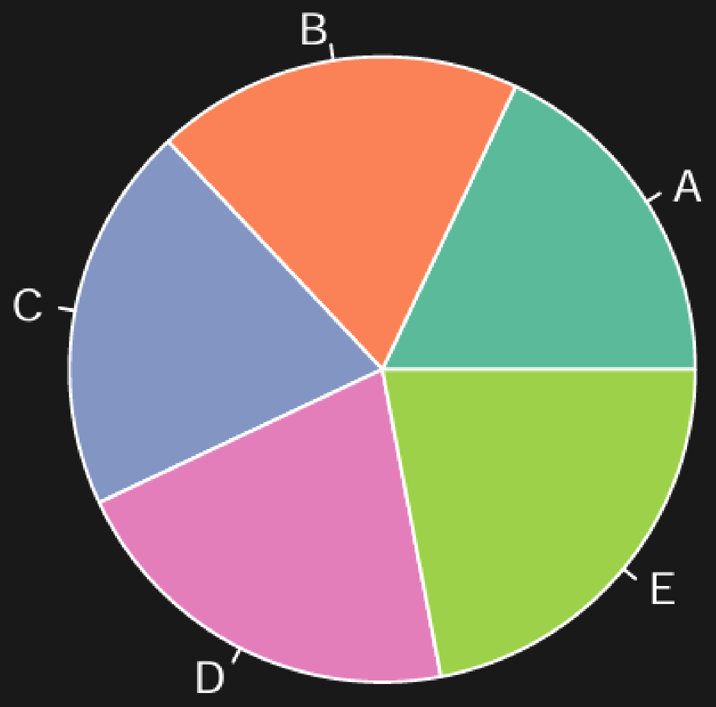
Examples of bad visualization

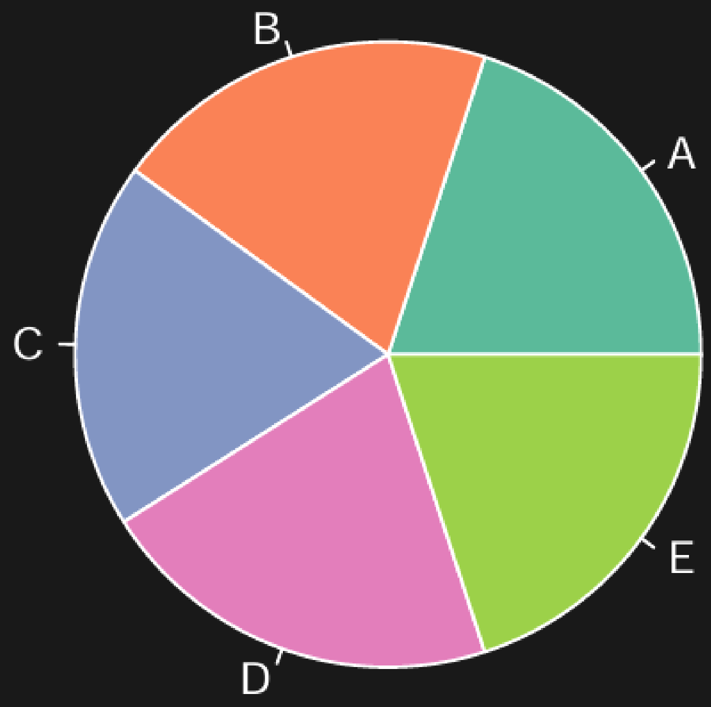
From: https://github.com/kbroman/Talk_Graphs

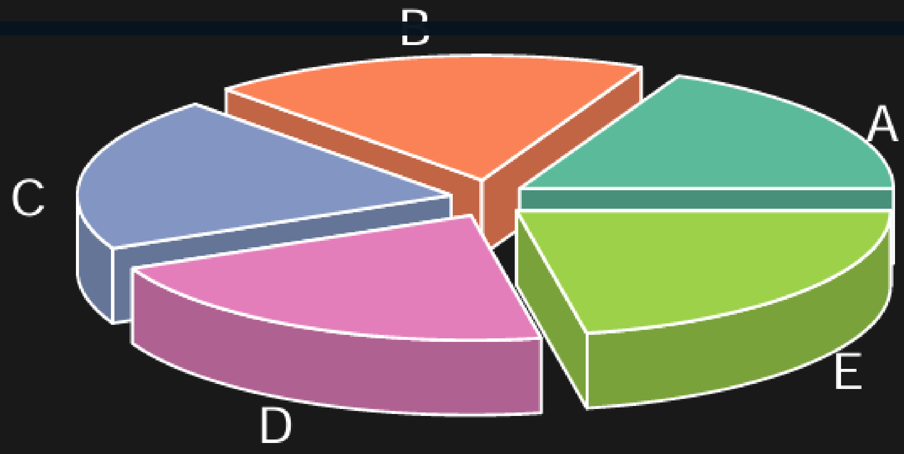


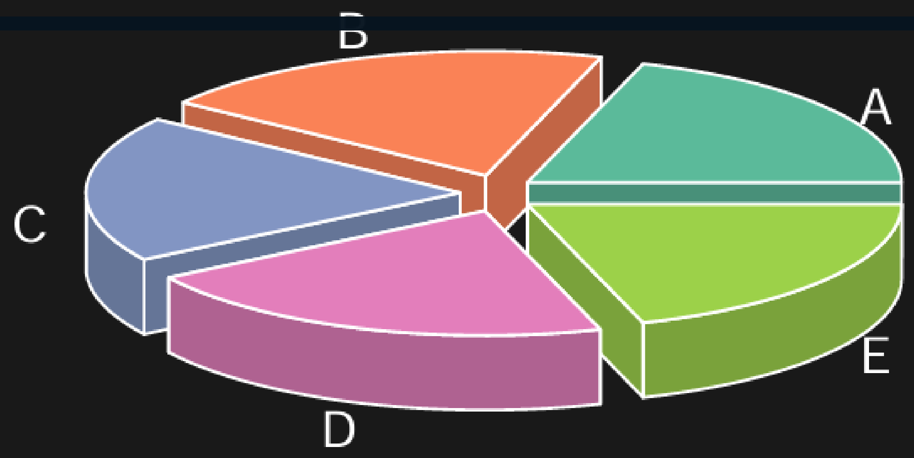


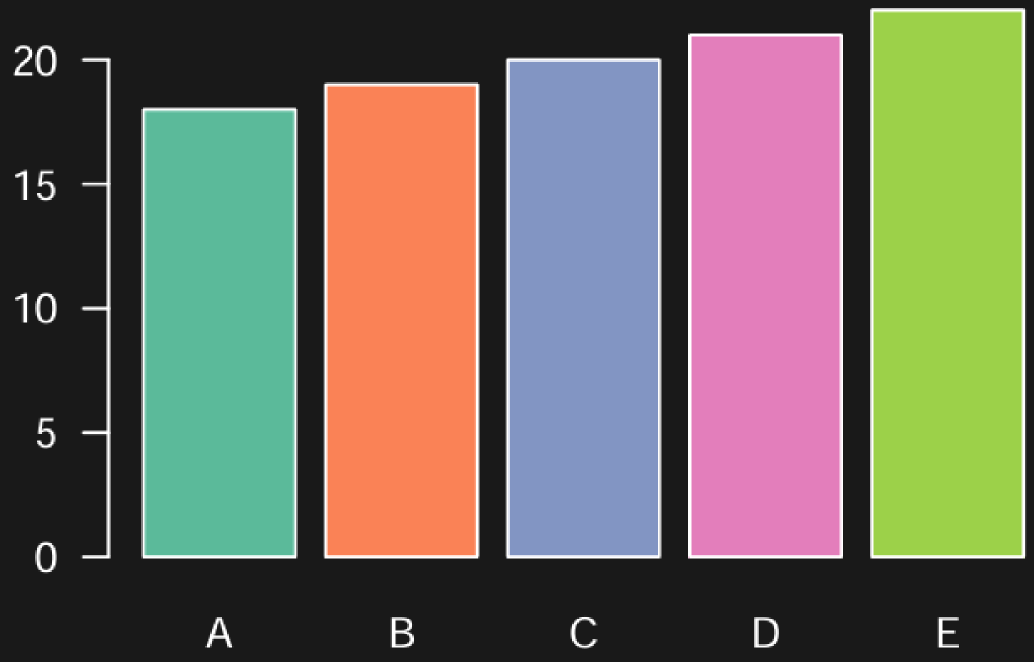
Avoid pie charts

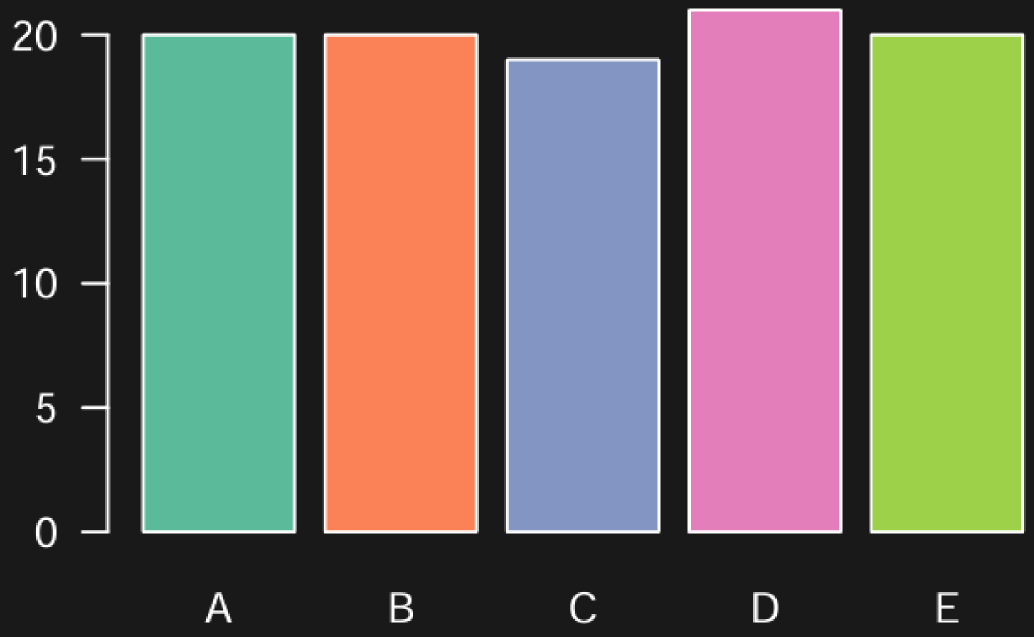


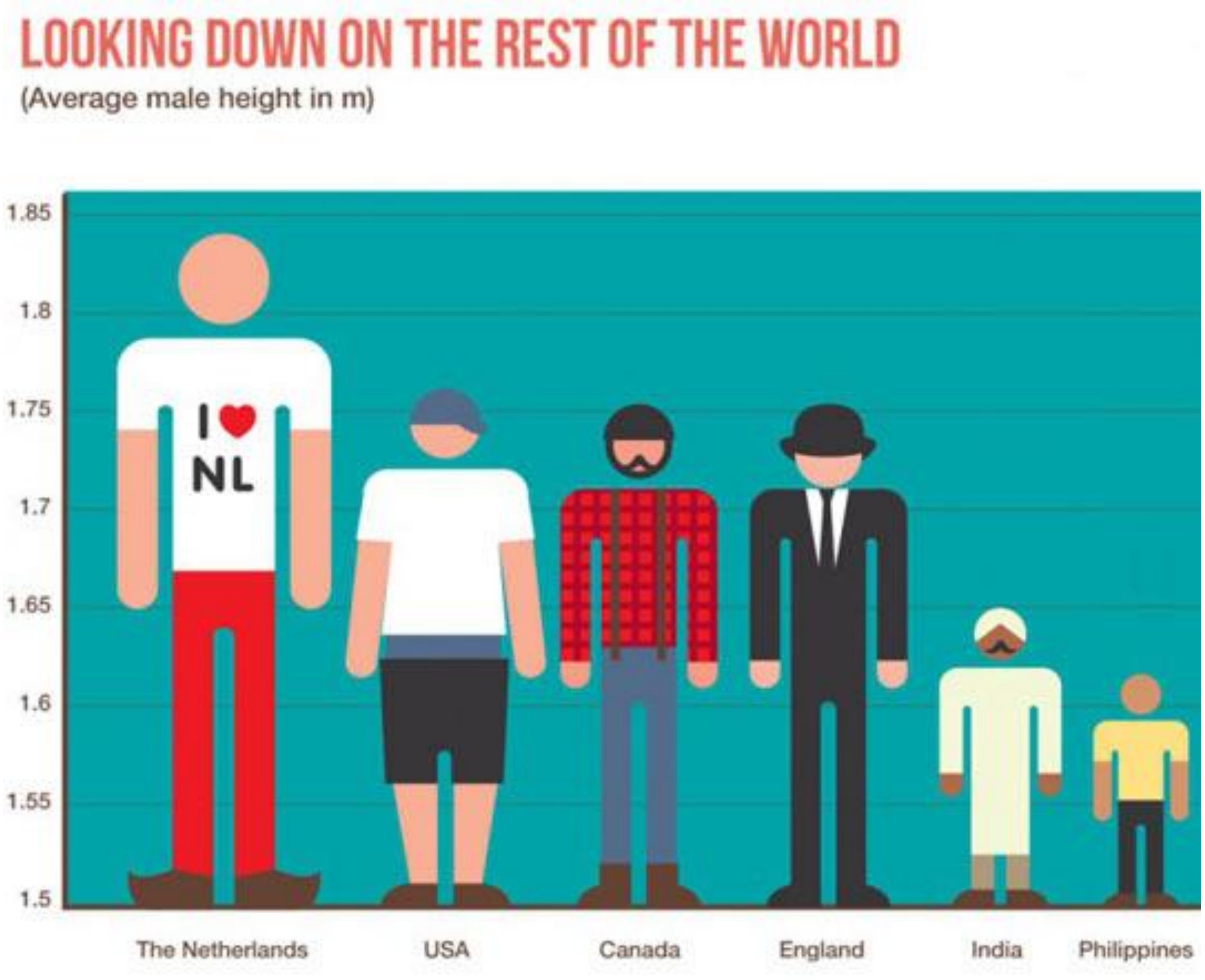



B
A
C
E
D


B
A
C
E
D


20
15
10
5
0
A
B
C
D
E


20
15
10
5
0
A
B
C
D
E



From: https://www.reddit.com/r/shittydataisbeautiful/

Heuristics for good visualizations

- Good Labeling
  - Have clear labels for:
    - Title, subtitle, axis labels (with units), legend, lines bars

Heuristics for good visualizations

- Good Labeling
  - Have clear labels for:
    - Title, subtitle, axis labels (with units), legend, lines bars

- Less is more / Keep it simple

Heuristics for good visualizations

- Good Labeling
  - Have clear labels for:
    - Title, subtitle, axis labels (with units), legend, lines bars
- Less is more / Keep it simple
  - That is optimize for Signal to Noise ratio

Heuristics for good visualizations

- Good Labeling
  - Have clear labels for:
    - Title, subtitle, axis labels (with units), legend, lines bars
- Less is more / Keep it simple
  - That is optimize for Signal to Noise ratio
  - Reduce clutter (aka chart junk)

Heuristics for good visualizations

- Good Labeling
  - Have clear labels for:
    - Title, subtitle, axis labels (with units), legend, lines bars
- Less is more / Keep it simple
  - That is optimize for Signal to Noise ratio
  - Reduce clutter (aka chart junk)
  - Avoid ornamentation (shadowing, extra illustration, etc)

Heuristics for good visualizations

- Good Labeling
  - Have clear labels for:
    - Title, subtitle, axis labels (with units), legend, lines bars

- Less is more / Keep it simple
  - That is optimize for Signal to Noise ratio
  - Reduce clutter (aka chart junk)
  - Avoid ornamentation (shadowing, extra illustration, etc)
  - Keep fonts simple (avoid uncessary **bold** or *italicization*) Don't use 3D charts. They can make it hard to discern or perceive the actual information.

Heuristics for good visualizations

- Good Labeling
  - Have clear labels for:
    - Title, subtitle, axis labels (with units), legend, lines bars

- Less is more / Keep it simple
  - That is optimize for Signal to Noise ratio
  - Reduce clutter (aka chart junk)
  - Avoid ornamentation (shadowing, extra illustration, etc)
  - Keep fonts simple (avoid uncessary **bold** or *italicization*) Don't use 3D charts. They can make it hard to discern or perceive the actual information.
  - Don't try to compare too many categories or data types in one chart

Heuristics for good visualizations

- Good Labeling
  - Have clear labels for:
    - Title, subtitle, axis labels (with units), legend, lines bars
- Less is more / Keep it simple
  - That is optimize for Signal to Noise ratio
  - Reduce clutter (aka chart junk)
  - Avoid ornamentation (shadowing, extra illustration, etc)
  - Keep fonts simple (avoid uncessary **bold** or *italicization*) Don't use 3D charts. They can make it hard to discern or perceive the actual information.
  - Don't try to compare too many categories or data types in one chart
  - Align with notation and narrative in the text, "proofread like a maniac"

Remove
to improve
(the **data-ink** ratio)

Created by Darkhorse Analytics www.darkhorseanalytics.com

From https://stat545.com/effective-graphs.html

Heuristics for good visualizations

- Less is more
    - Reduce clutter

Heuristics for good visualizations

- Less is more
  - Reduce clutter

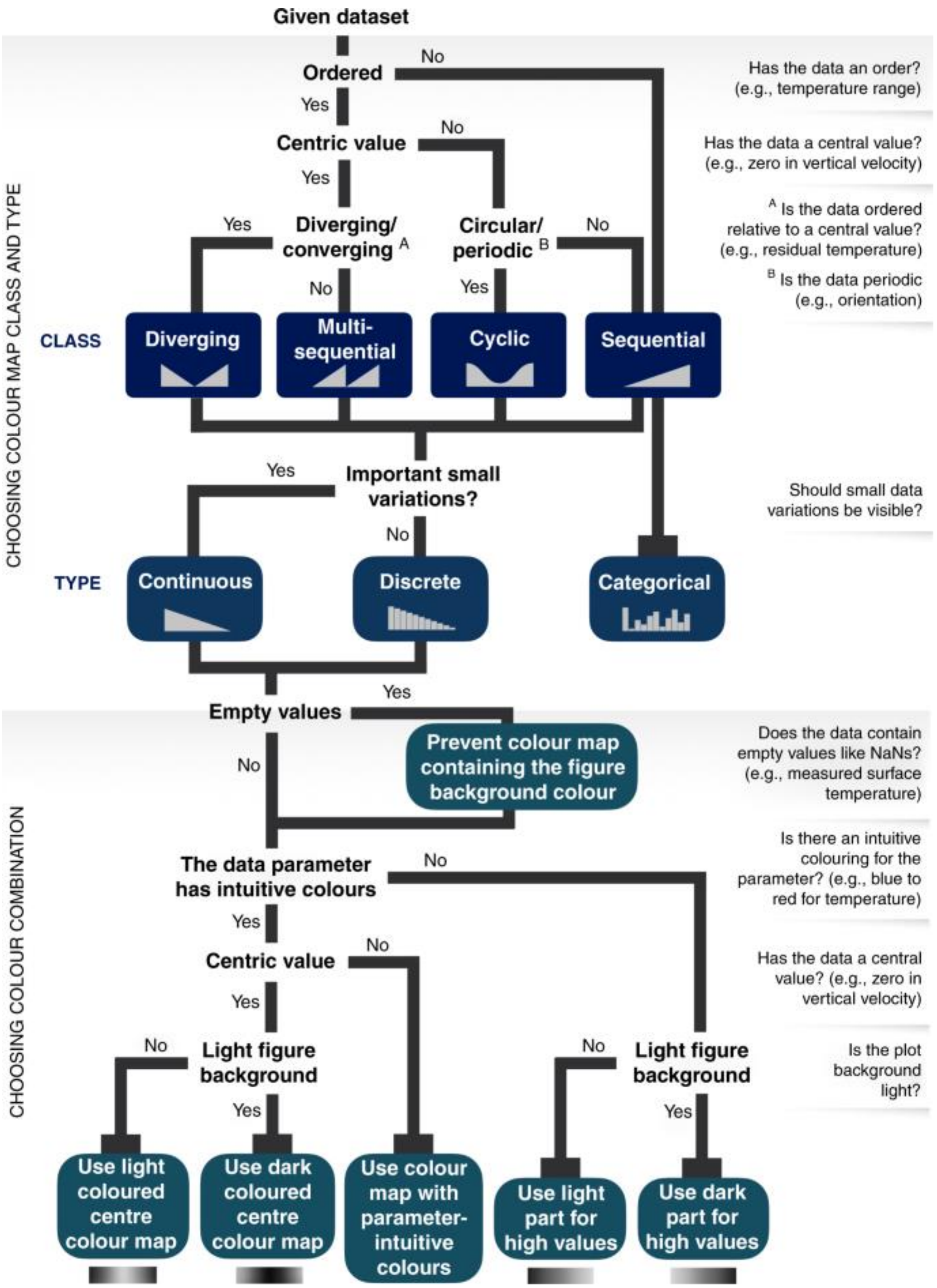
- Colors
  - Does it help in communication and add meaning?
  - Colorblind friendly?
  - Prints well in grayscale?
  - Respect intuitive coloring schemes

Heuristics for good visualizations

- Less is more
  - Reduce clutter

- Colors
  - Does it help in communication and add meaning?
  - Colorblind friendly?
  - Prints well in grayscale?
  - Respect intuitive coloring schemes

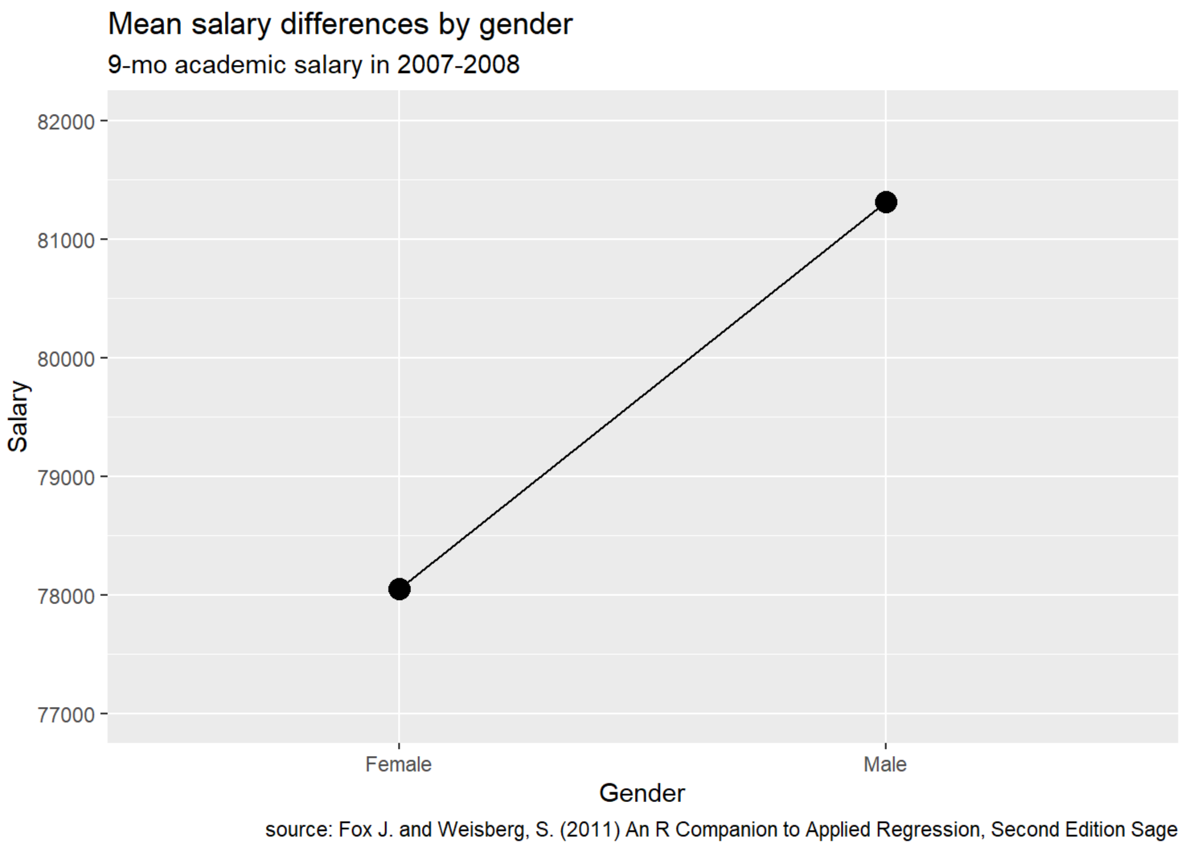
- Play around with themes/ palletes / packages
  - See: viridis package, ggthemes (e.g. scale_color_colorblind()),..



https://www.nature.com/articles/s41467-020-19160-7

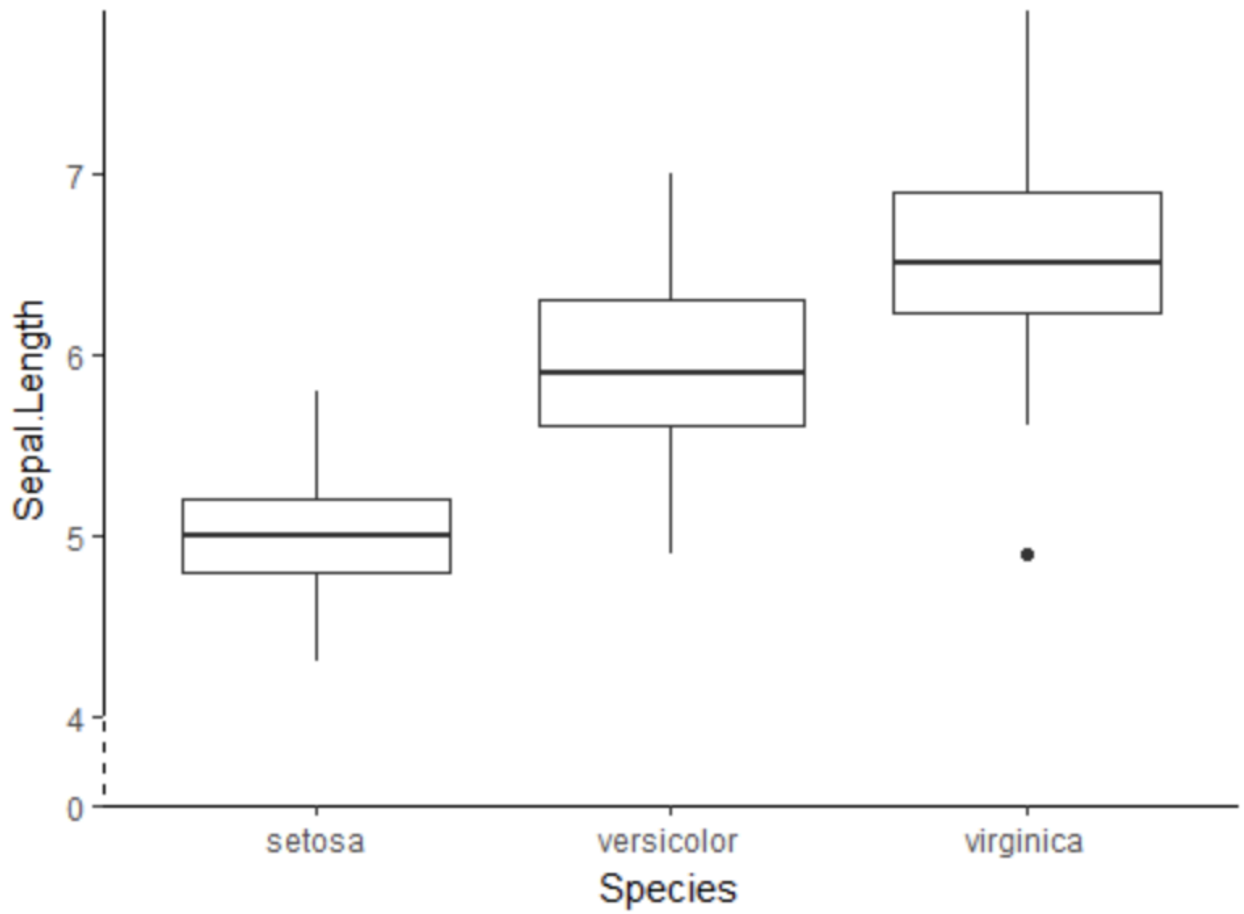
Heuristics for good visualizations

- Pay attention to axis scales
  - When possible include 0 in the scale.
  - If not, indicate this
    - Visual break / zig zags
    - Explain why this is helpful (to visualize trends more clearly, etc.)



From: https://rkabacoff.github.io/datavis/Advice.html

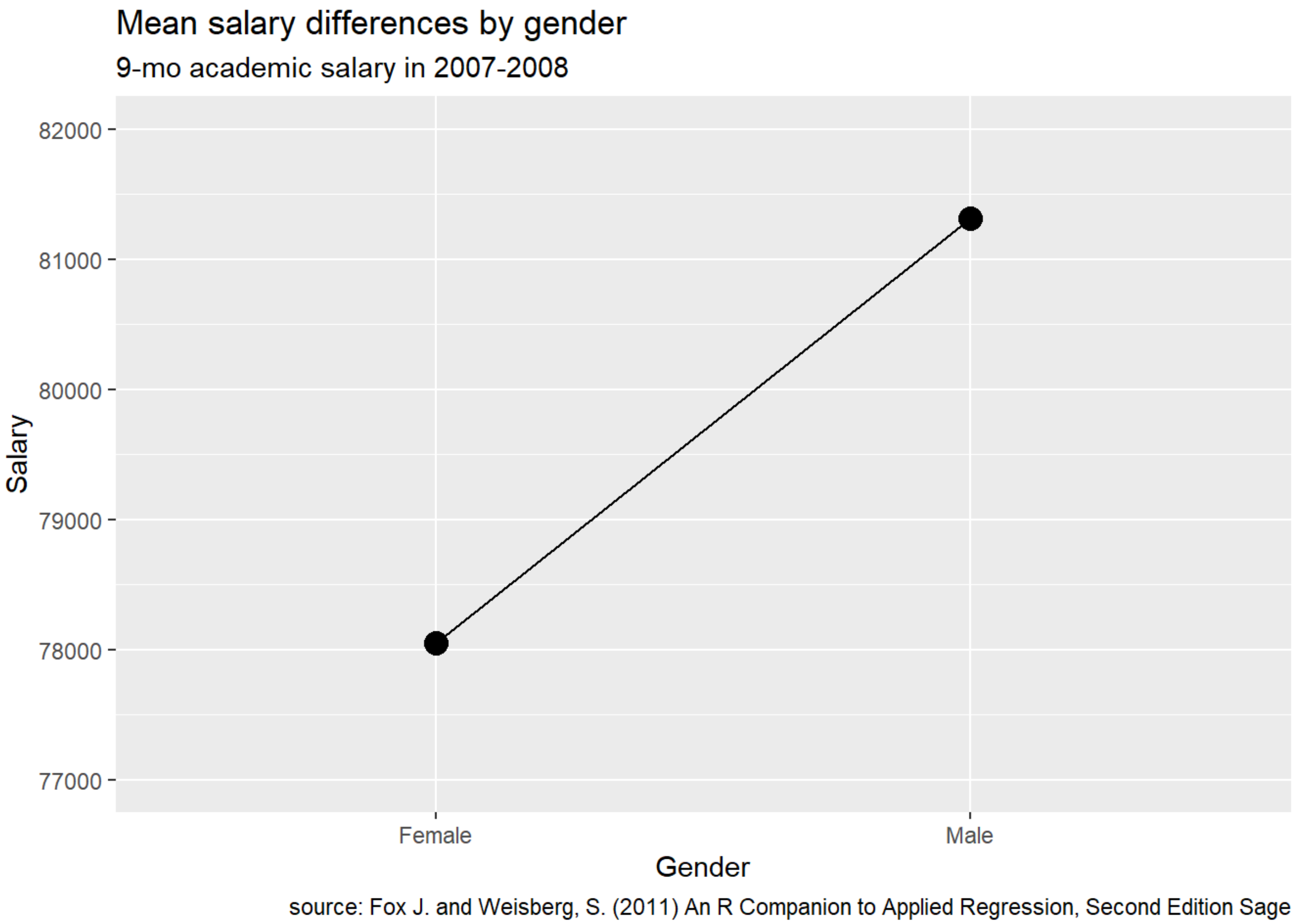
```r
ggplot(iris, aes(Species, Sepal.Length)) + geom_boxplot() +
  theme_classic() +
  scale_y_continuous(limits = c(3.5,NA), expand = c(0,0),
                     breaks = c(3.5, 4:7), labels = c(0, 4:7)) +
  theme(axis.line.y = element_blank()) +
  annotate(geom = "segment", x = -Inf, xend = -Inf, y = -Inf, yend = Inf) +
  annotate(geom = "segment", x = -Inf, xend = -Inf, y =  3.5, yend = 4,
           linetype = "dashed", color = "white")
```



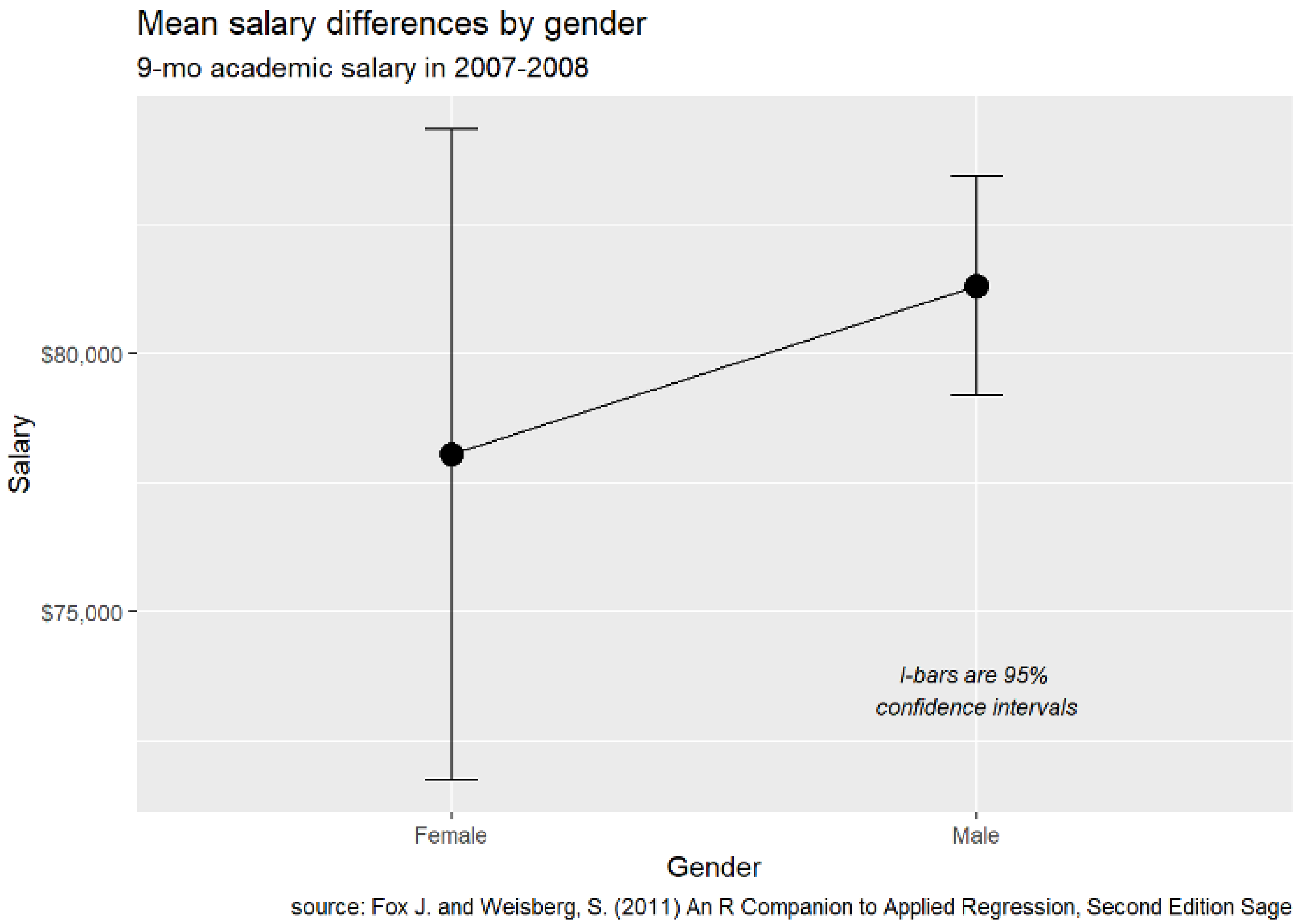
Heuristics for good visualizations

- Pay attention to axis scales
  - When possible include 0 in the scale.

- Include uncertainty/variance measures
  - Mean by itself is rarely meaningful



From: https://rkabacoff.github.io/datavis/Advice.html



From: https://rkabacoff.github.io/datavis/Advice.html

Heuristics for good visualizations

- Pay attention to axis scales
  - When possible include 0 in the scale.

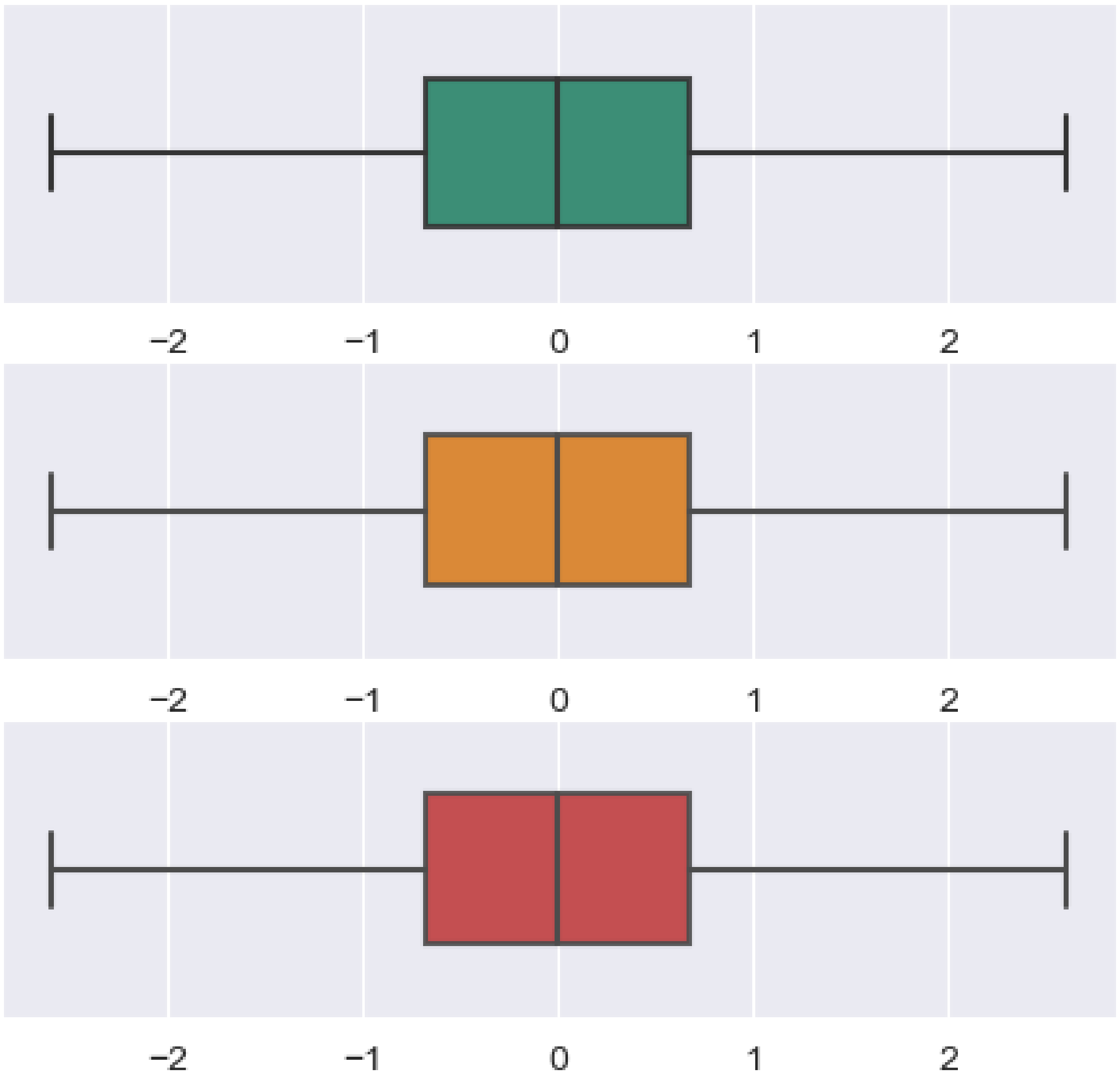
- Include uncertainty/variance measures
  - Mean by itself is rarely meaningful
  - Error bars/ Confidence intervals help

Heuristics for good visualizations

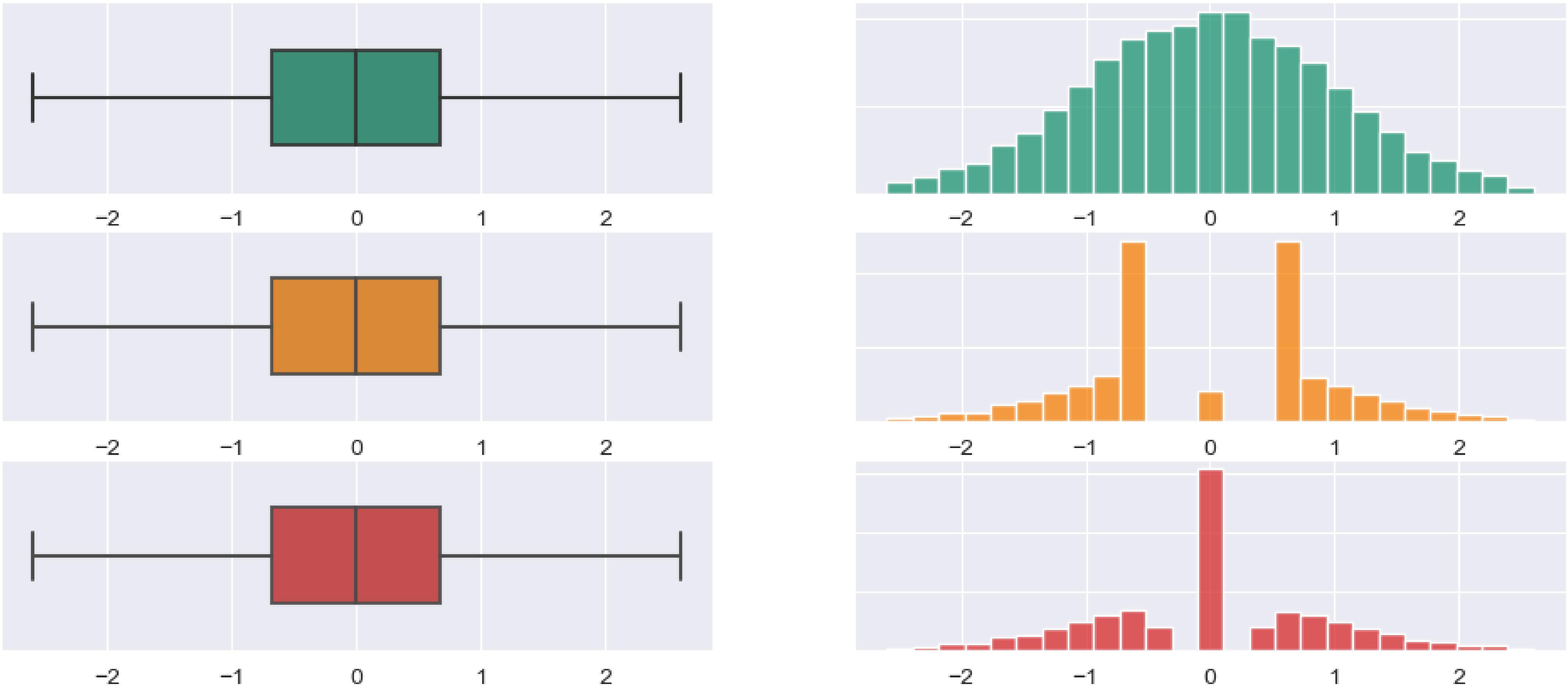
- Pay attention to axis scales
  - When possible include 0 in the scale.

- Include uncertainty/variance measures
  - Mean by itself is rarely meaningful
  - Error bars/ Confidence intervals help
  - Better still to include distributional information

Boxplot

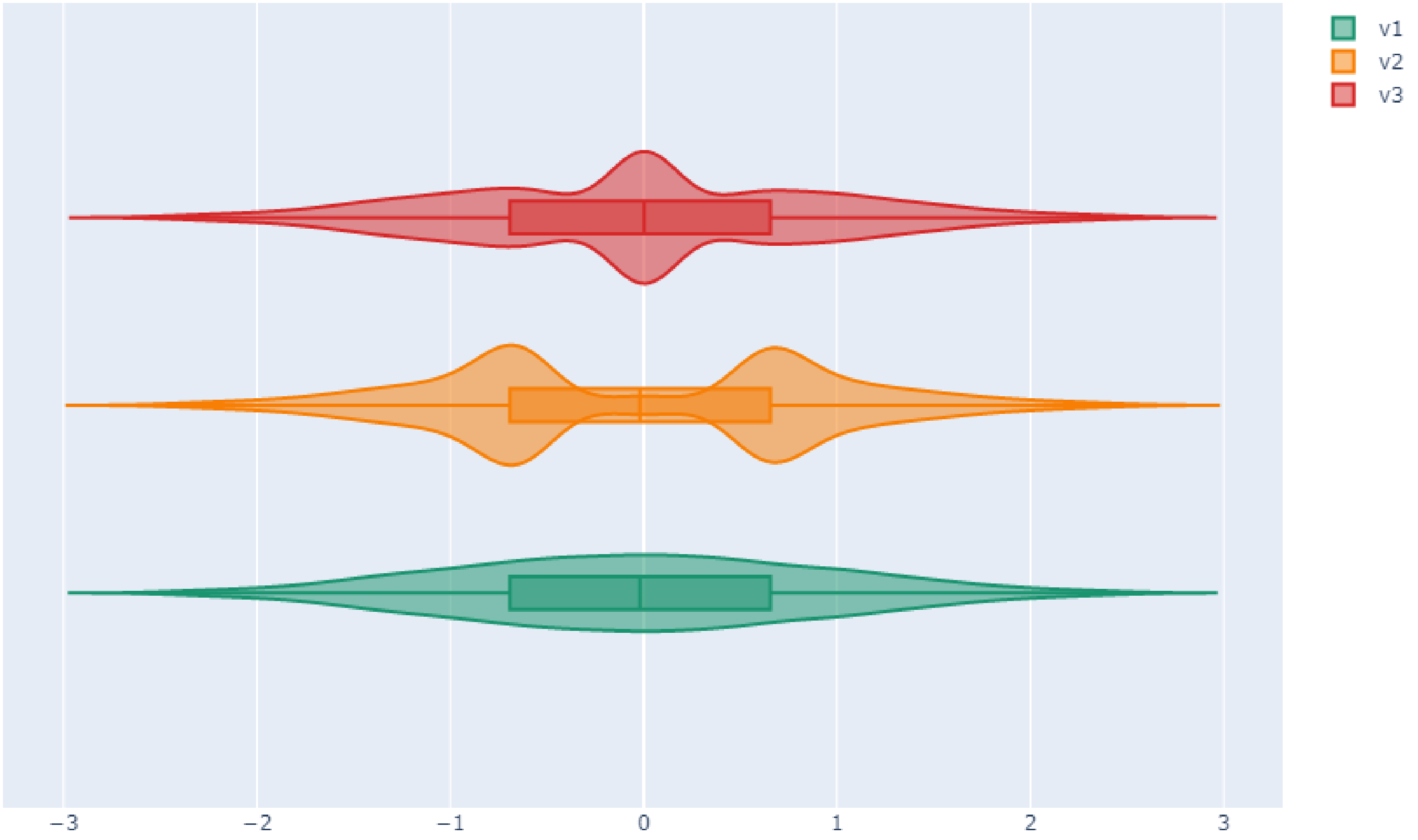


−2
−1
0
1
2
−2
−1
0
1
2
−2
−1
0
1
2

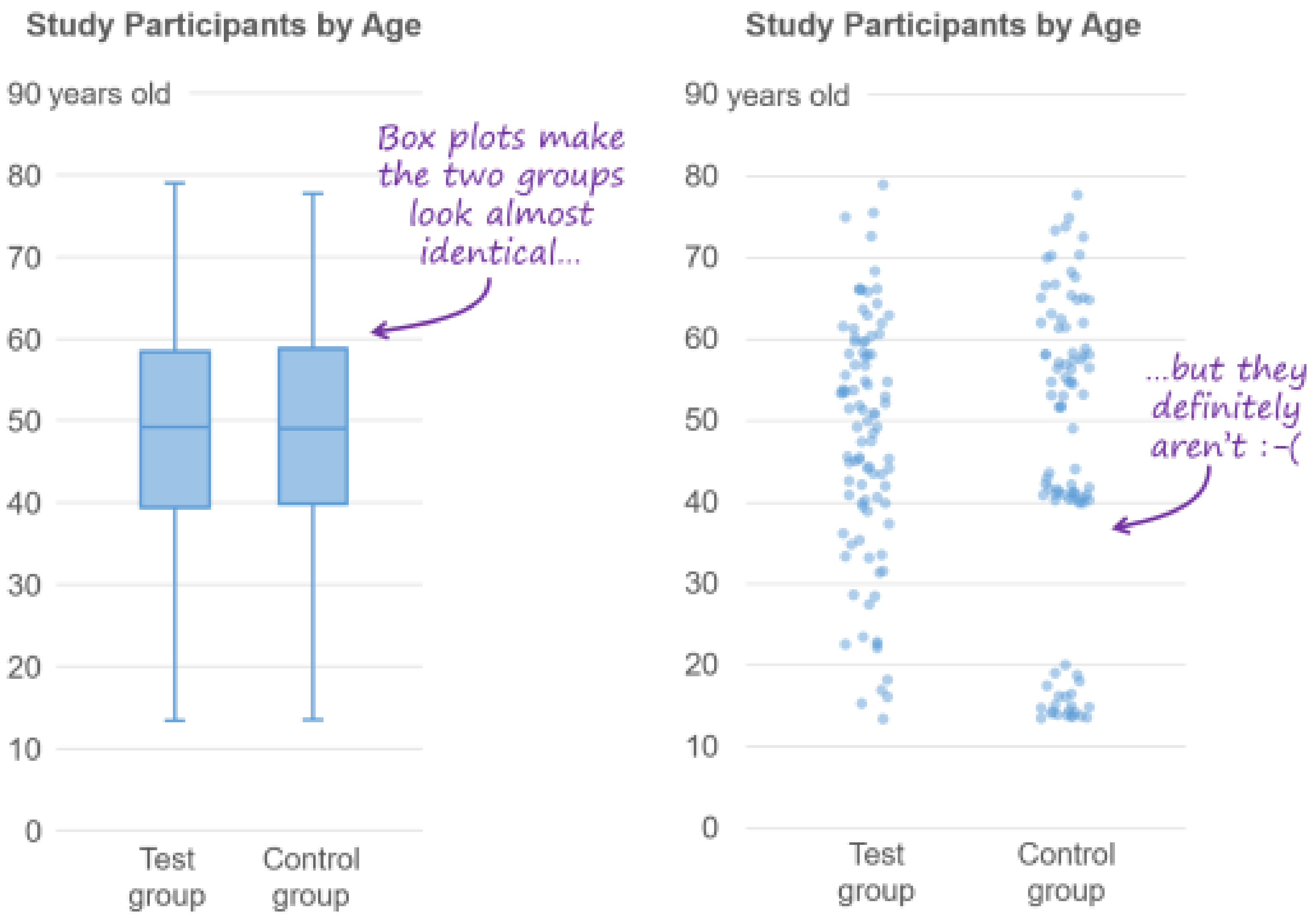


−2 −1 0 1 2
−2 −1 0 1 2
−2 −1 0 1 2
−2 −1 0 1 2
−2 −1 0 1 2
−2 −1 0 1 2

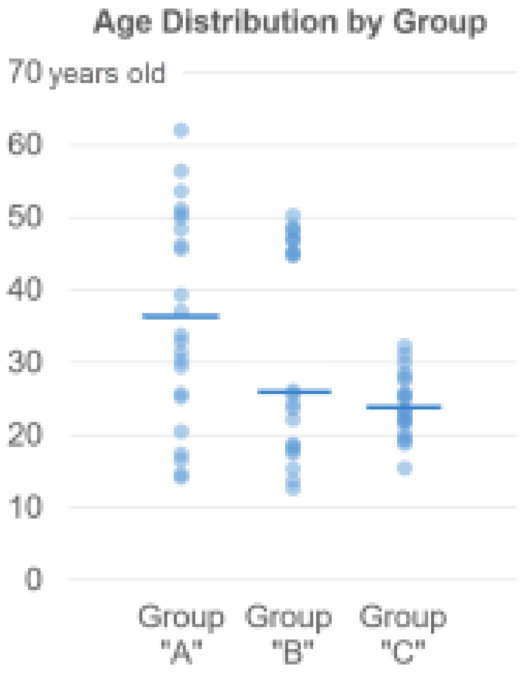
Violin plots

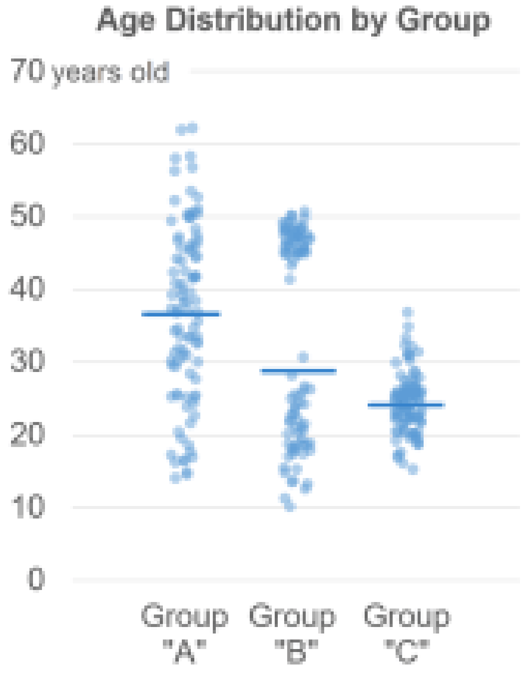


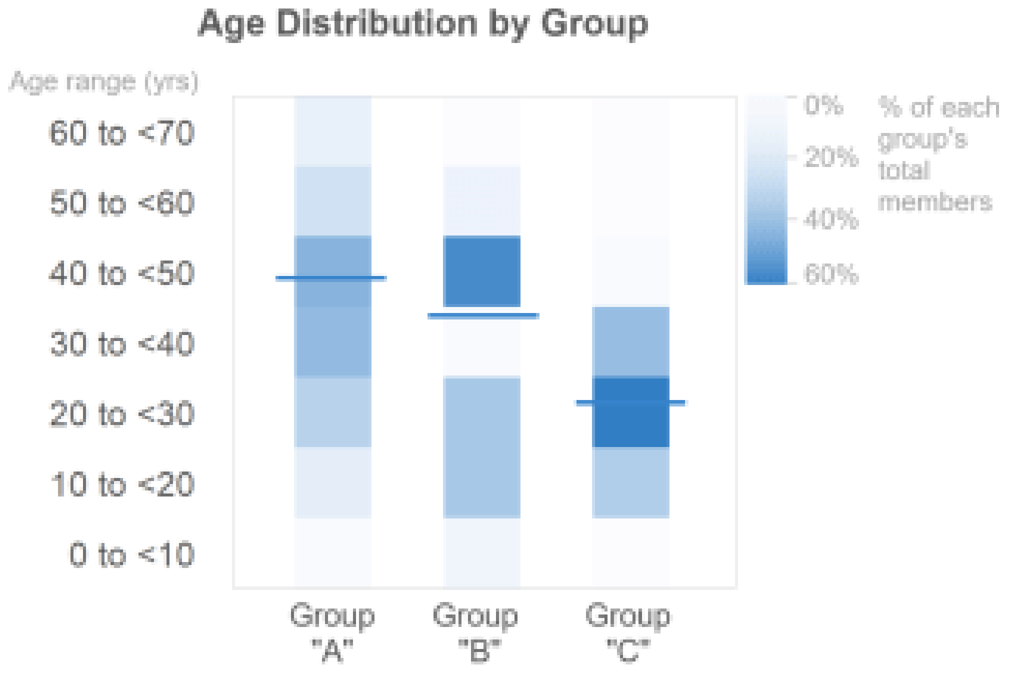
Strip plots

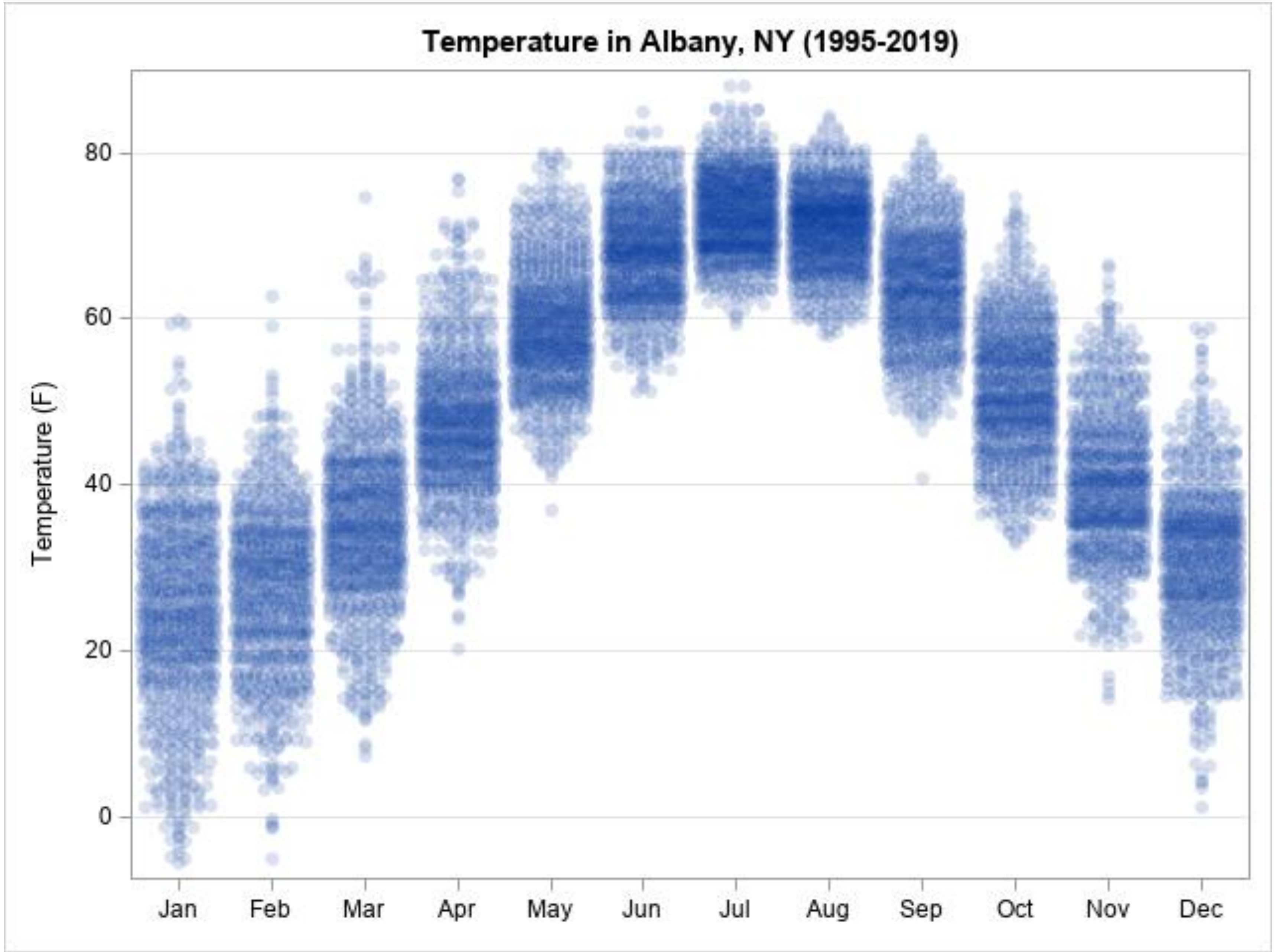


From https://nightingaledvs.com/ive-stopped-using-box-plots-should-you/


Age Distribution by Group
70 years old
60
50
40
30
20
10
0
Group "A"
Group "B"
Group "C"

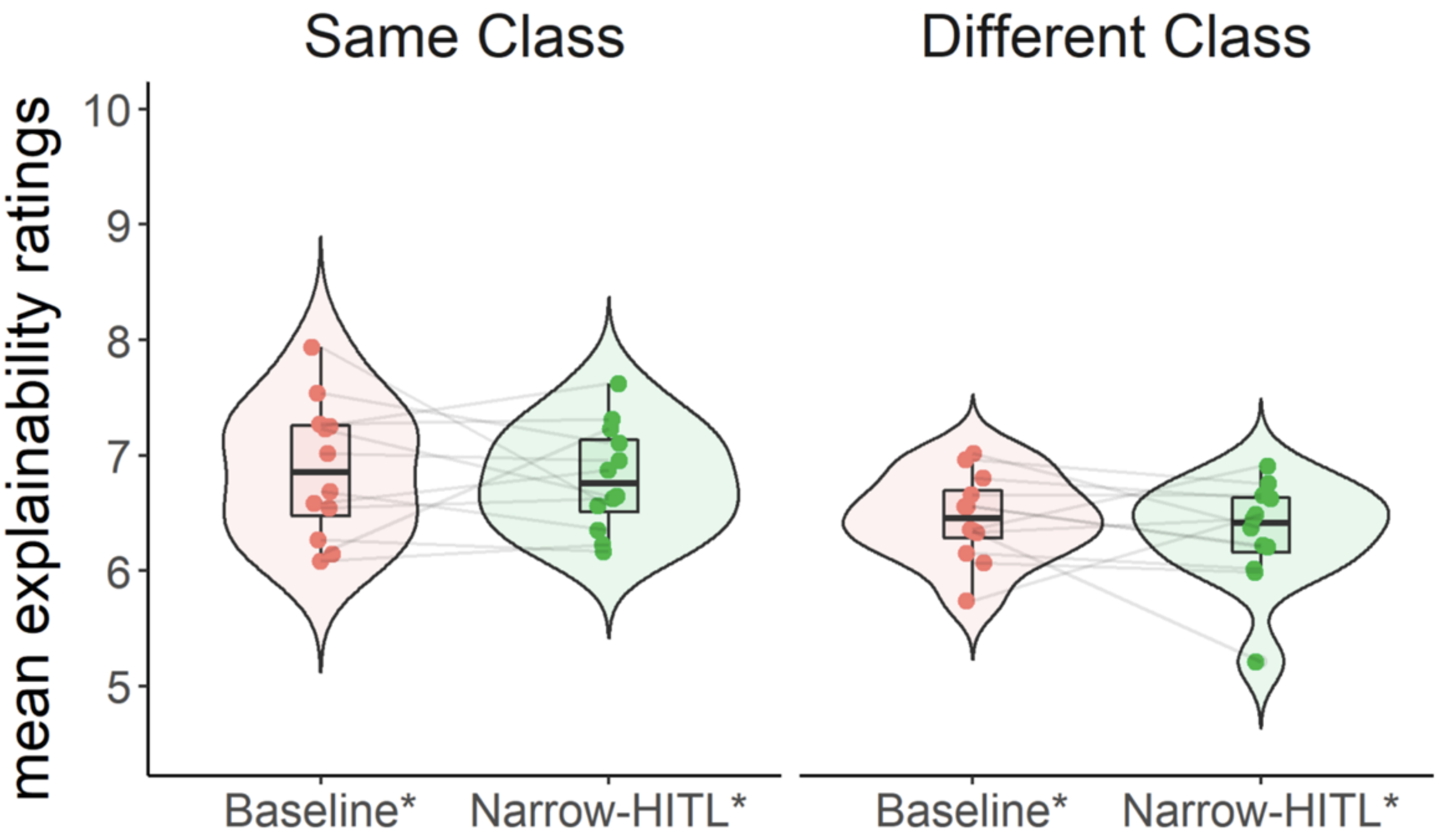

Age Distribution by Group
70 years old
60
50
40
30
20
10
0
Group "A"
Group "B"
Group "C"


Age Distribution by Group
Age range (yrs)
60 to <70
50 to <60
40 to <50
30 to <40
20 to <30
10 to <20
0 to <10
Group "A"
Group "B"
Group "C"
0%
20%
40%
60%
% of each group's total members



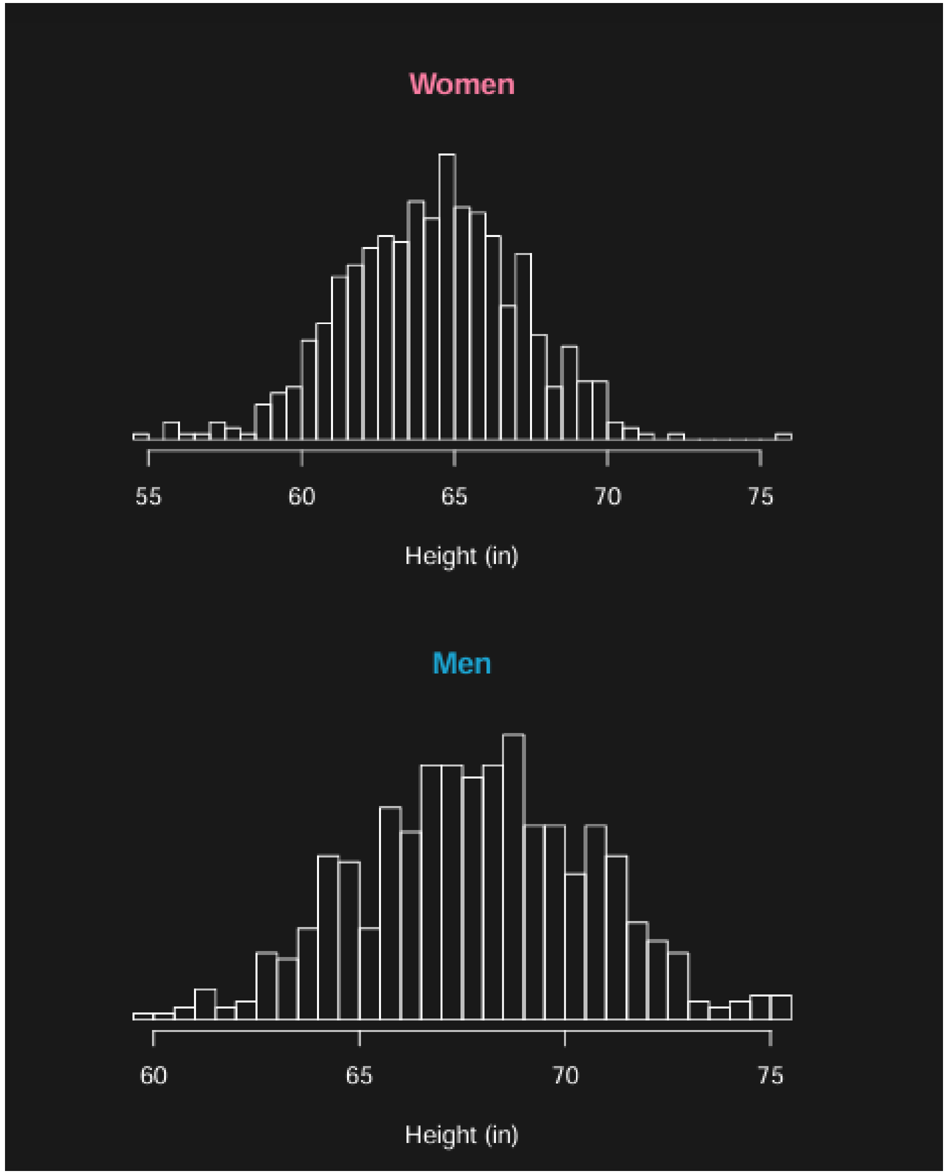
From https://blogs.sas.com/content/iml/2019/11/18/strip-plot-sas.html

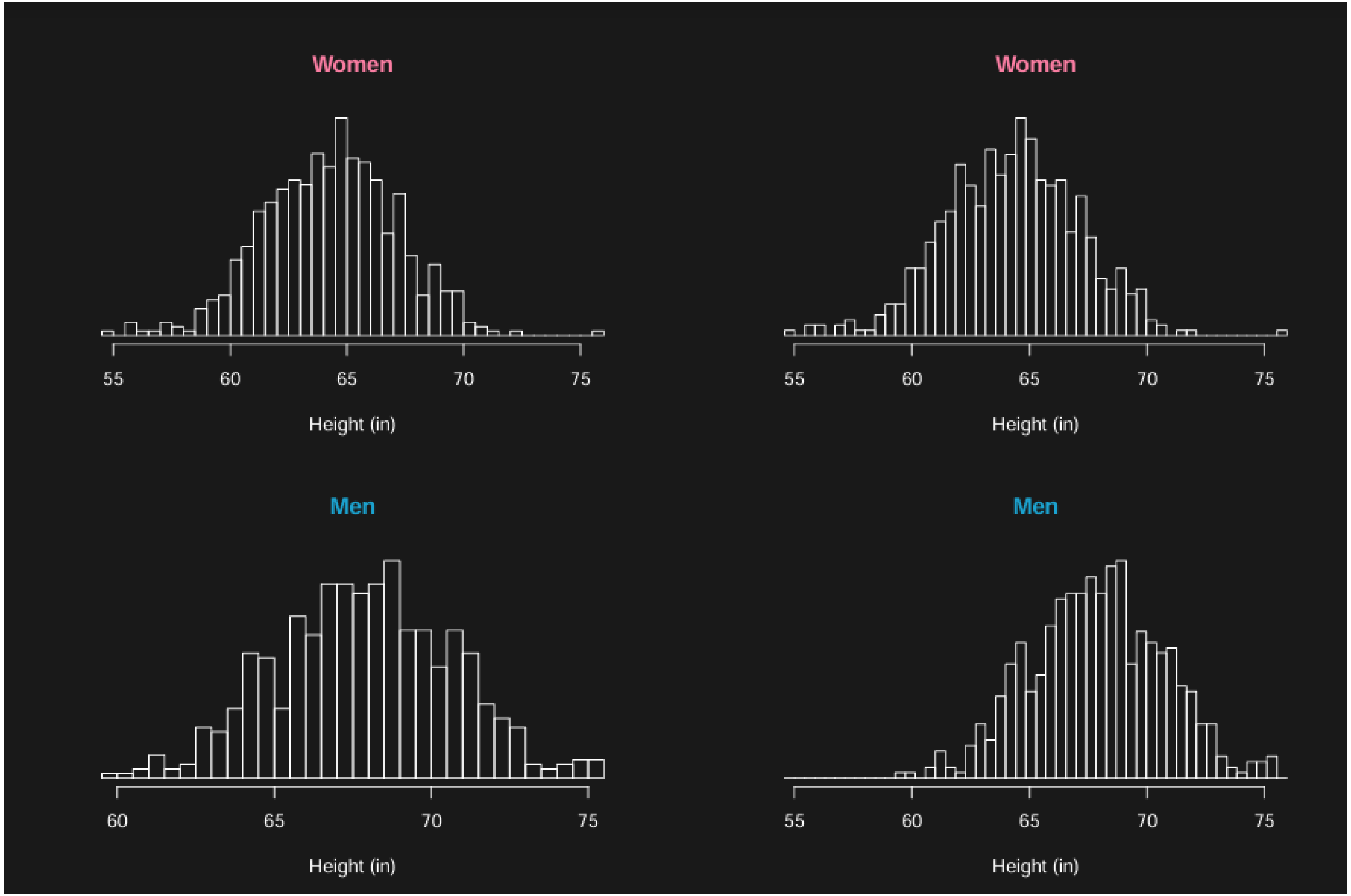
(Dot and) Violin plots



Heuristics for good visualizations

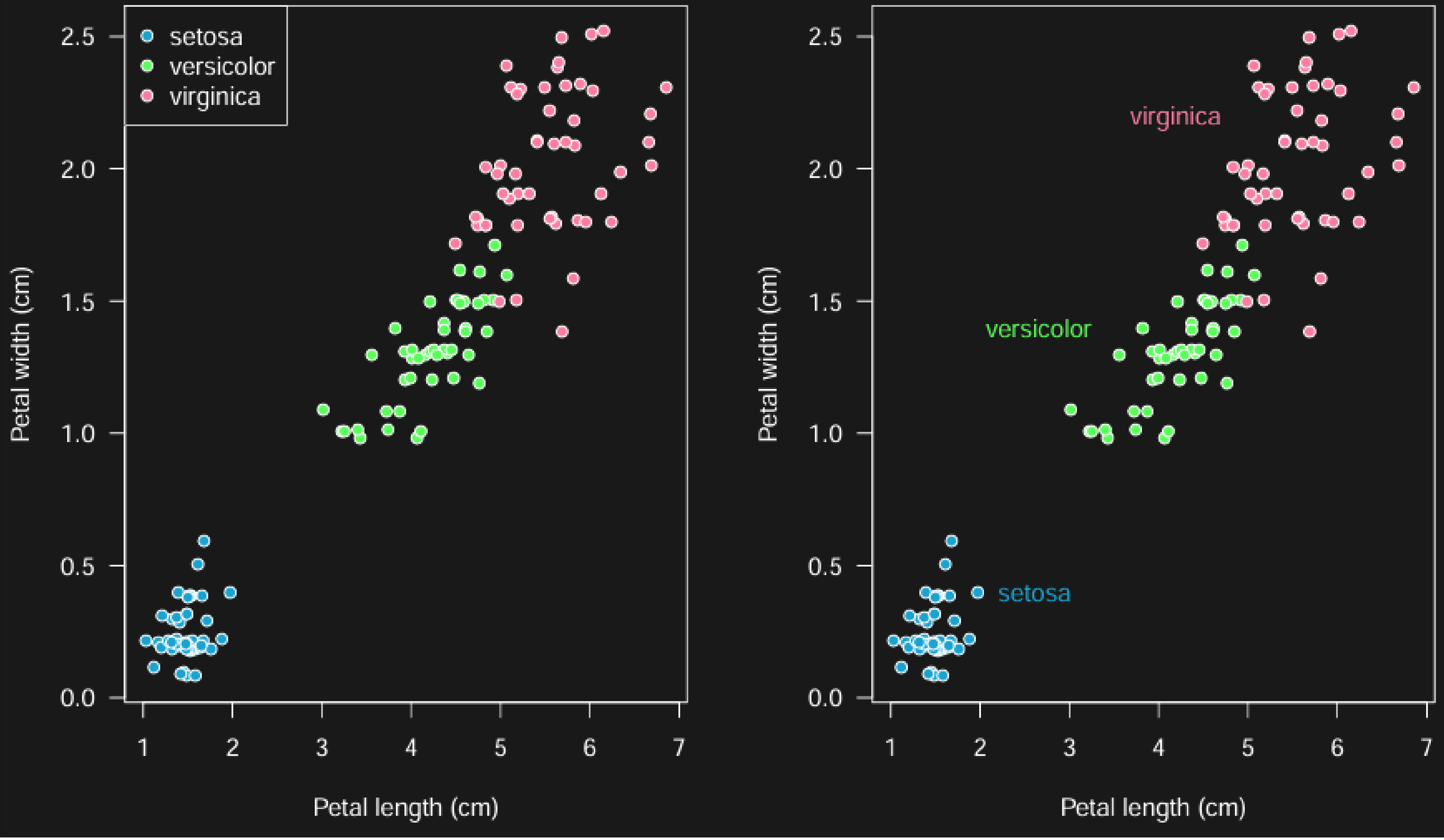
- Make it easy to compare panels / facets
  - Use common axes
  - Align them vertically


Women
55
60
65
70
75
Height (in)
Men
60
65
70
75
Height (in)



Women
55 60 65 70 75
Height (in)
Women
55 60 65 70 75
Height (in)
Men
60 65 70 75
Height (in)
Men
55 60 65 70 75
Height (in)

Use labels not legends

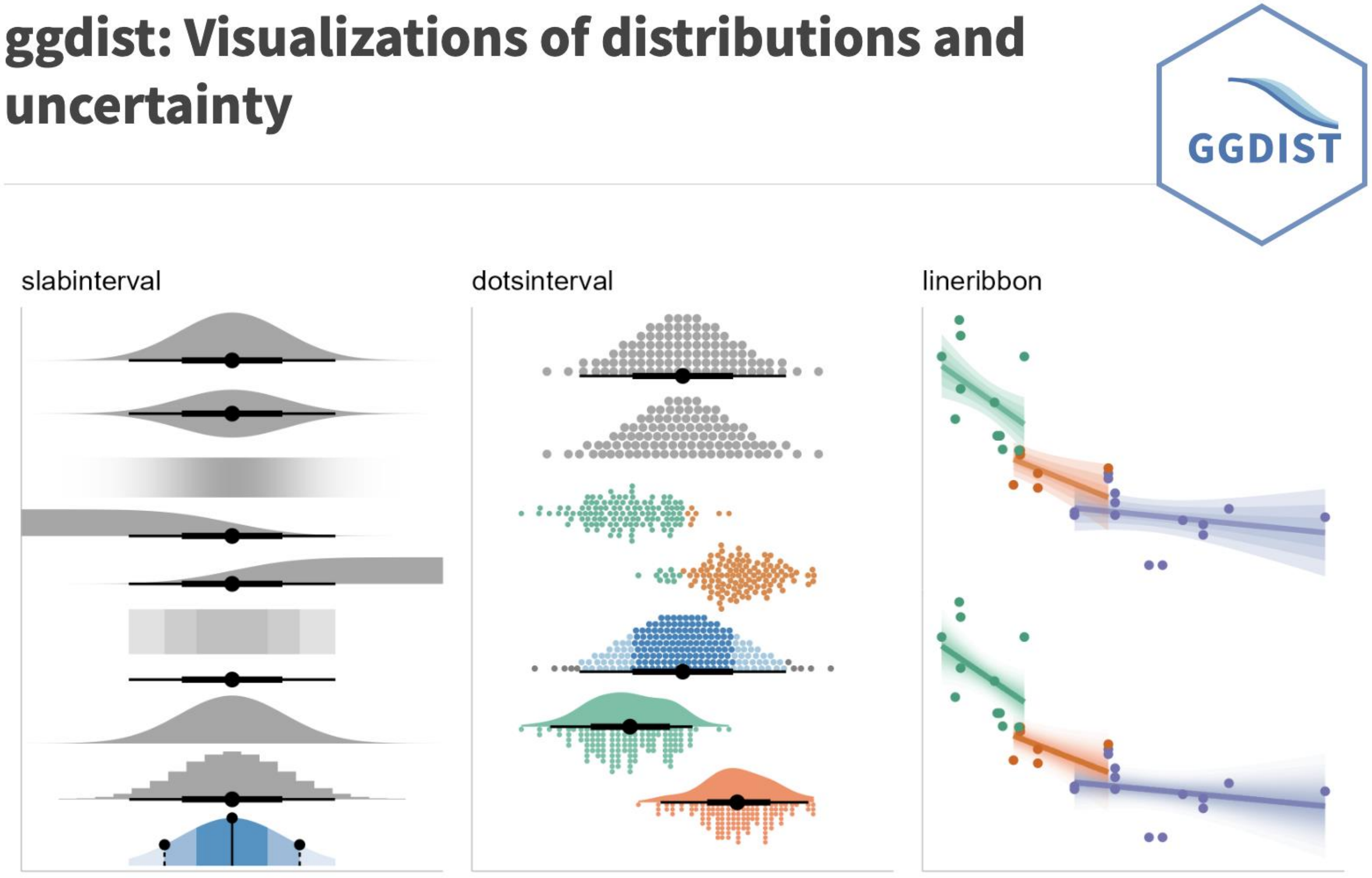


Heuristics for good visualizations (TLDR)

- **Main goal** – to communicate (with minimal reader effort) and not mislead

- Minimize clutter; keep it simple
- Use colors and aesthetics in a meaningful manner
- Ensure that axes scales aren't misleading
- Include variance measures / show distributional information
- Align things to make comparison easy

Active area of research & development



Active area of research

Designing for Interactive Exploratory Data Analysis Requires Theories of Graphical Inference

Jessica Hullman[1], Andrew Gelman[2]

[1]Northwestern University, [2]Columbia University

Jessica Hullman will be visiting Princeton in February to present in the PSY 505 seminar series.

Active area of research

Designing for Interactive Exploratory Data Analysis Requires The Graphical Inf

Jessica Hullman[1], Andrew Gelman[2]

[1]Northwestern University, [2]Columbia University

A Cognitive Interpretation of Data Analysis†

Garrett Grolemund, Hadley Wickham

Regression..

Previously…

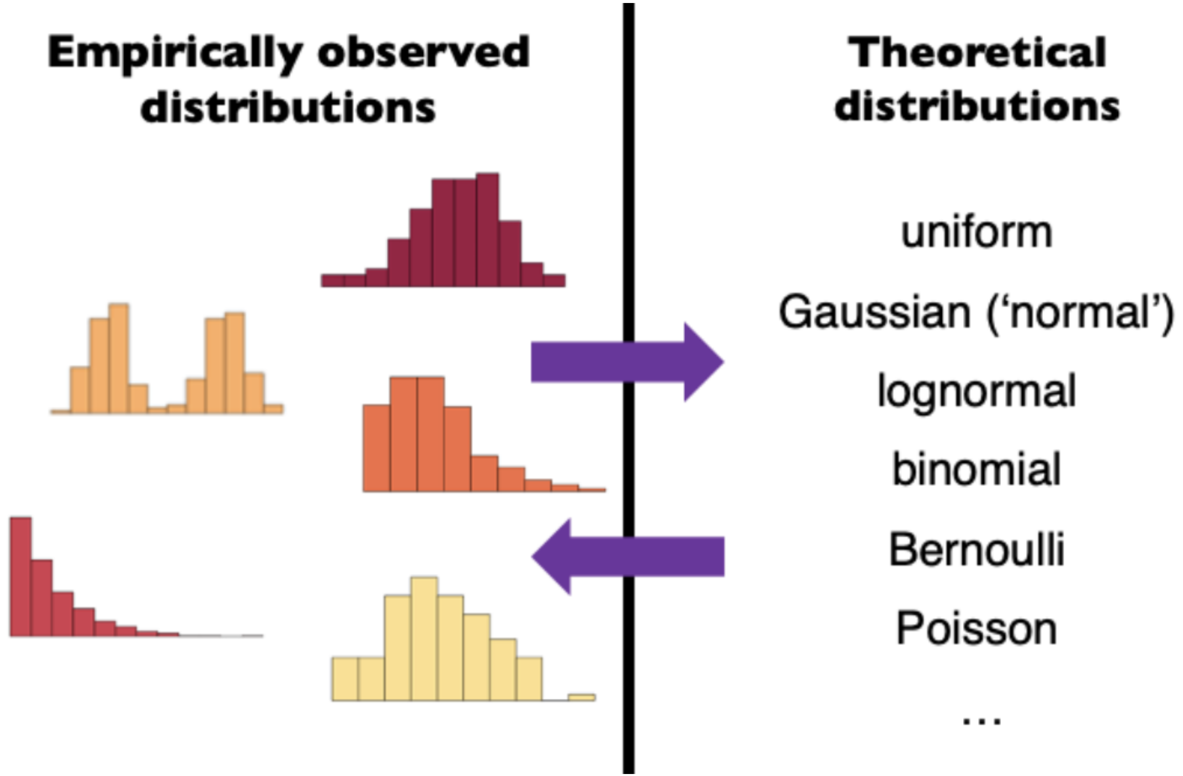
What is a model?

- Models are simplifications of things in the real world

What is a statistical modelling?

- **Statistical modeling** = “making **models** of **distributions**”

  (coming up with a plausible data generating process/ DGP)



Basic Structure of a Statistical Model

$$data = model + error$$

- Data
- Model
- Use our model to ***predict*** the value of the data for any given observation:

$$\widehat{data}_i = model_i$$

- Error (predicted – observed)

$$error_i = data_i - \widehat{data}_i$$

GLM (Generalized Linear Model)

- General mathematical framework
  - Regression all the way down
  - Highly flexible
    - Can fit qualitative (categorical) and quantitative predictors
  - Easy to interpret
  - Helps understand interrelatedness to other models
  - Easy to build to more complex models

A simple model

- Null or empty model

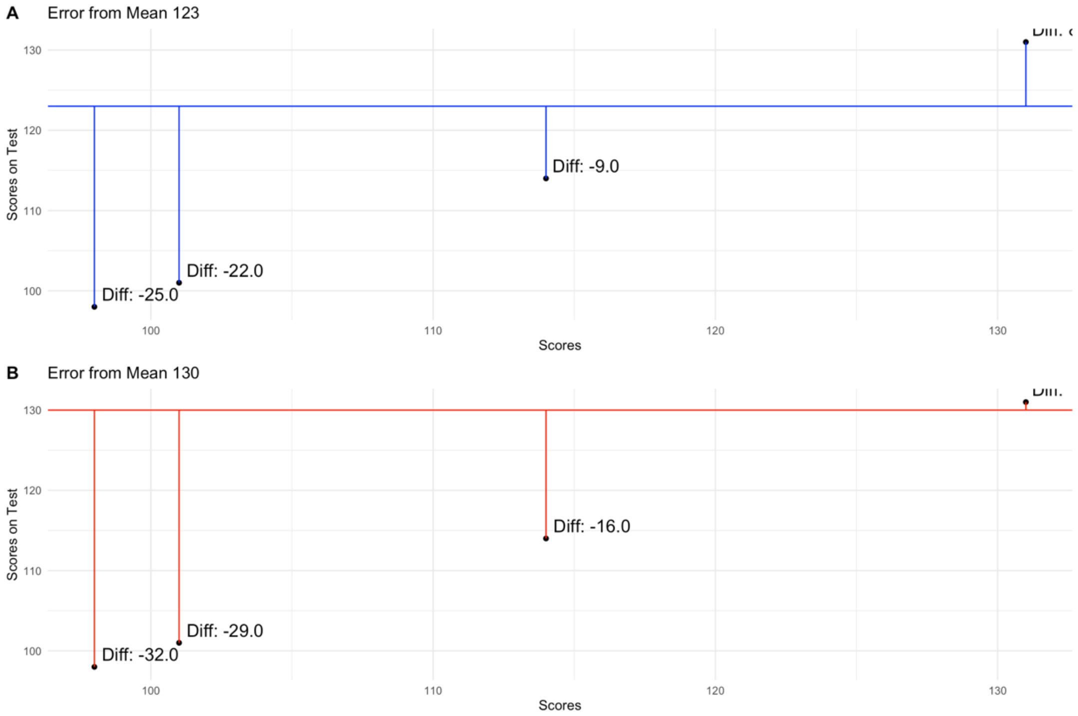
$$Y_i = \beta_0 + \epsilon$$

$$\mathrm{Y}_i = b_0 + e$$

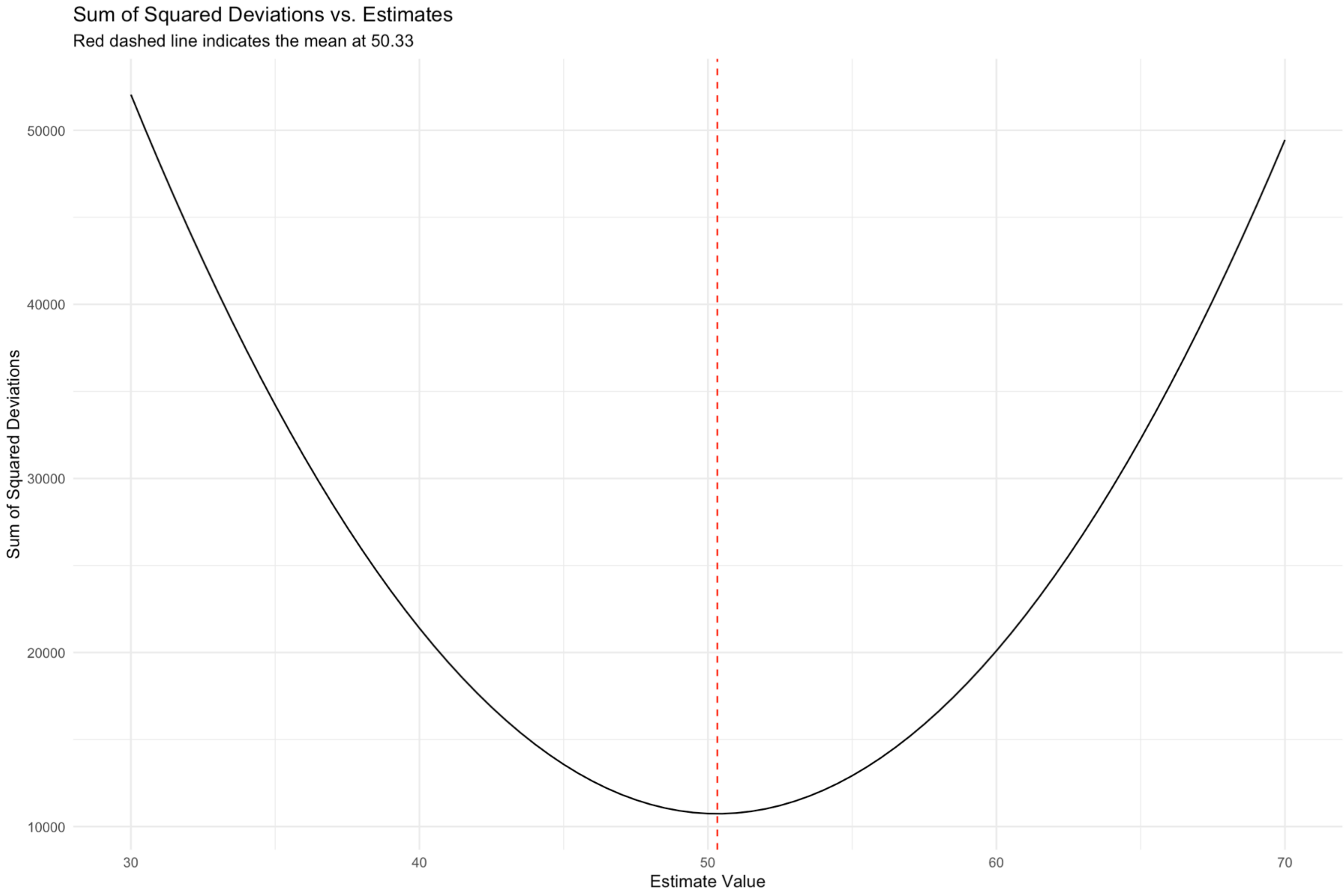
- Makes the same prediction for each observation (and we add an error sample)

Figuring out b_0

- Goal of any model is to find an “estimator” that minimizes the error
  - How we define error will determine the best estimator



SS minimized at the mean



Error Measures

- Sum of Squared Errors (SS)
  - This measures the total squared difference between observed and predicted values
  - Most commonly used in regression analysis (what we will be using)

$$SS = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

Describing error

- We should have some overall description of the accuracy of model's predictions
  - SSR
    - Standard deviation

$$s^2 = \text{MSE} = \frac{1}{n-p}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2$$

$$SD = \sqrt{MSE}$$

Fitting the empty model

We can use the `lm` function to fit the model with no predictors (Null Model / Empty model)

```
empty.model <- lm(HrsSleep2009~NULL, data =
smallNLS) empty.model
```

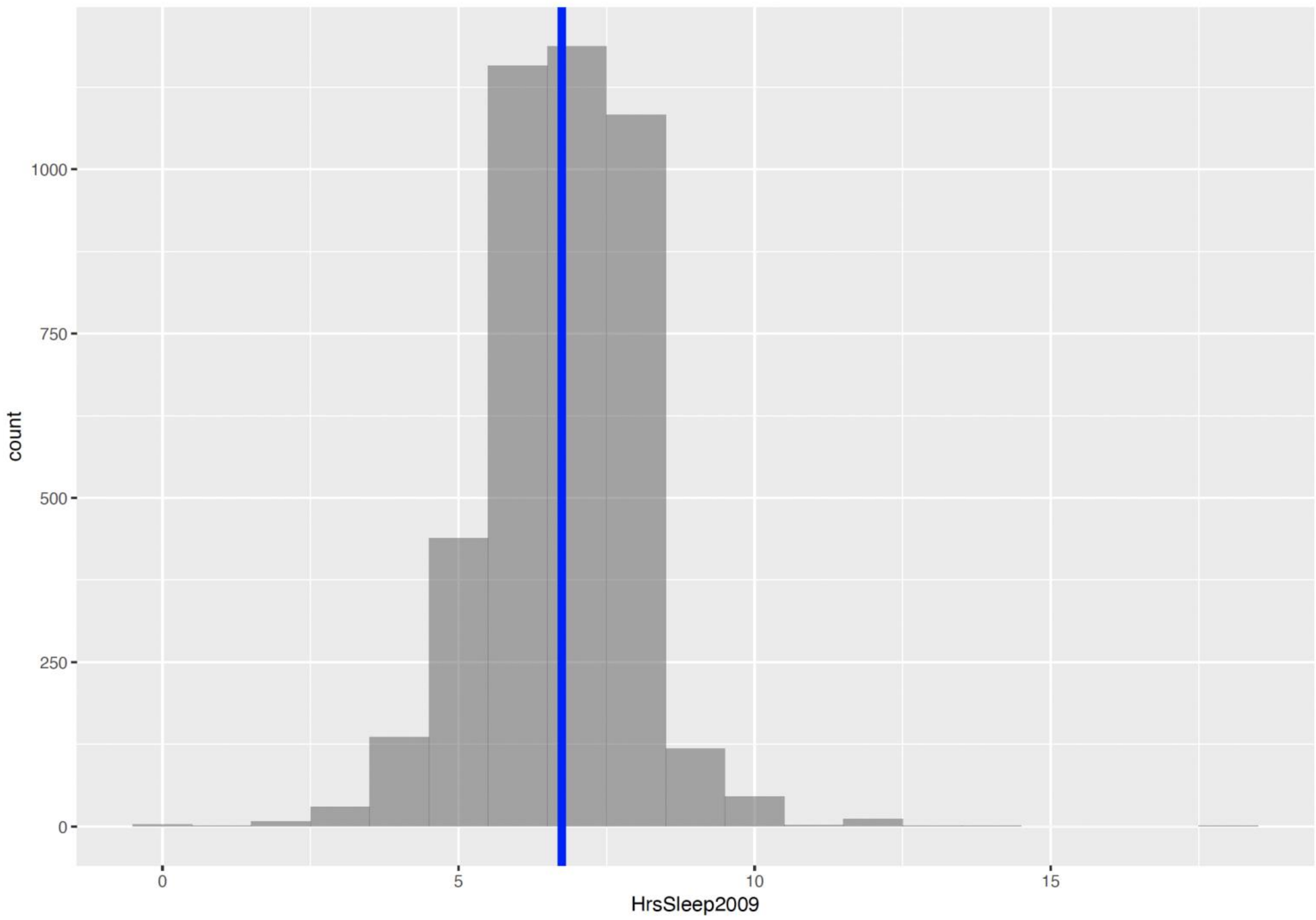
```
##
## Call:
## lm(formula = HrsSleep2009 ~ NULL, data =
smallNLS) ##
## Coefficients:
## (Intercept)
##        6.65
```

```
favstats(~HrsSleep2009, data = smallNLS)
```

```
##  min Q1 median Q3 max mean       sd  n missing
##    5  6      7  8   8 6.65 1.136708 20       0
```

Adding prediction to the plot

```
gf_histogram(~HrsSleep2009, data = NLSdata, binwidth = 1) %>%
  gf_vline(xintercept=~mean,data =SleepStats, color="blue",size=2)
```



A simple model

- Null or empty model

$$Y_i = \beta_0 + \epsilon$$

$$\epsilon = \beta_0 - Y_i$$

- Can we add more information to our model? That is, can we produce better explanations for the data, and reduce improve the error measure?

Adding explanatory variables to models

- So far, we have looked at variation, and summary measures for it
- But we haven’t really explained why it is occuring.

A simple model

- Null or empty model

$$Y_i = \beta_0 + \epsilon$$

$$\epsilon = \beta_0 - Y_i$$

- Can we add more information to our model? That is, can we produce better explanations for the data, and reduce improve the error measure?

Running example: Francis Galton’s height data

Galton’s height data

played a crucial role in the development of regression analysis.

Purpose: to understand how physical traits, specifically height, were passed from parents to children.

Data: height measurements from 928 adult children and their parents.

Methodology: For simplicity, Galton used the average height of the parents (adjusting for gender differences) as the 'parent height'.

Galton’s height data

````
```{r}
library(tidyverse)
height_data <- read.csv("galton_height.csv")
height_data
```
````

Description: df [928 × 2]

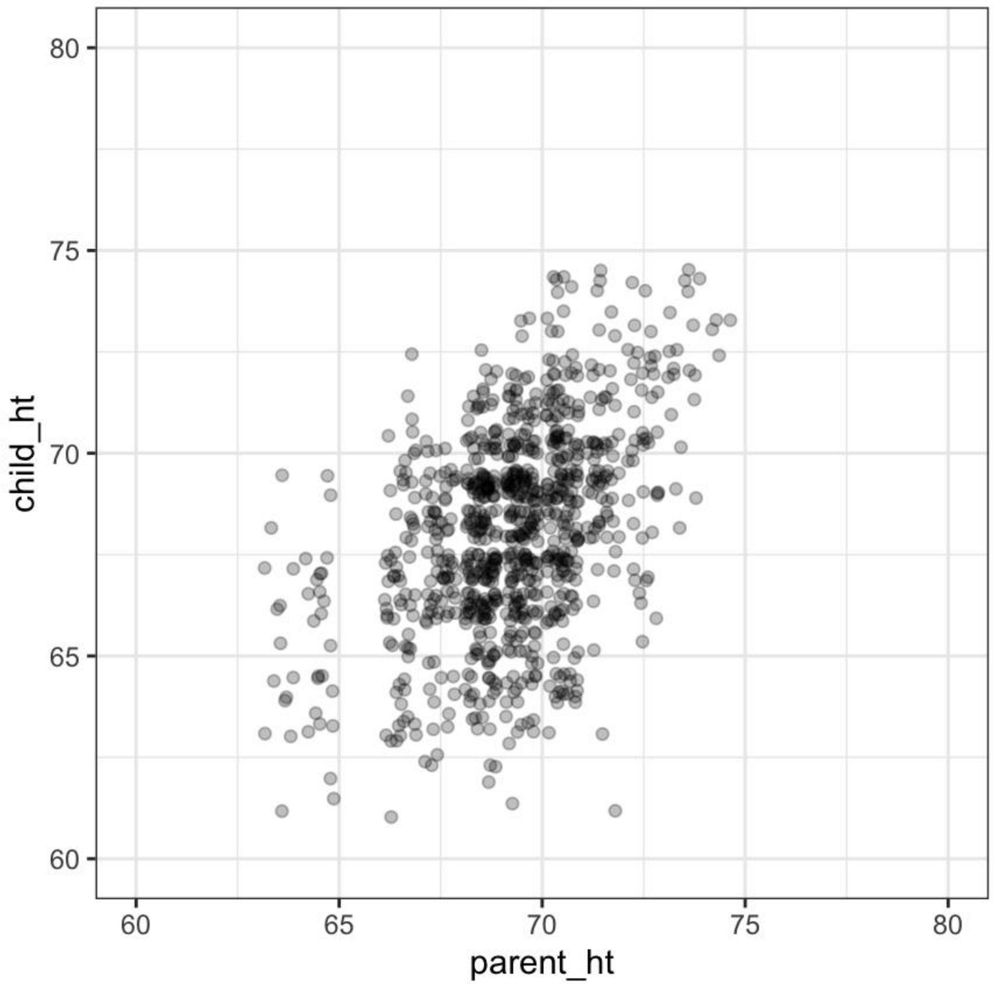
child_ht <dbl>	parent_ht <dbl>
72.2	74.5
73.2	74.5
73.2	74.5
73.2	74.5
68.2	73.5
69.2	73.5
69.2	73.5
70.2	73.5
71.2	73.5
71.2	73.5

Galton’s height data

```
height_data %>%
  ggplot(aes(y=child_ht, x= parent_ht))+
  geom_jitter(alpha = 0.25)+
  scale_x_continuous(limits = c(60,80))+
  scale_y_continuous(limits = c (60,80))+
  coord_fixed()+
  theme_bw()
```

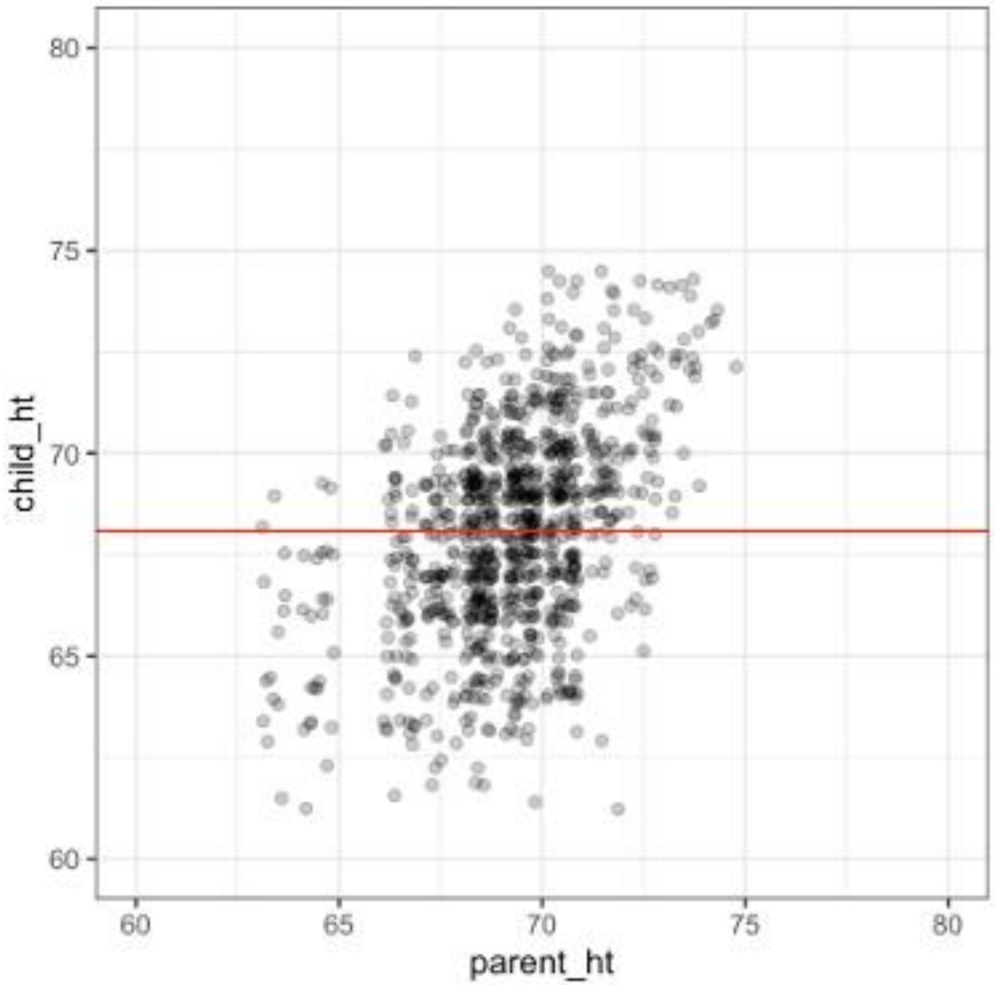
Galton's height data

```
height_data %>%
  ggplot(aes(y=child_ht, x= parent_ht))+
  geom_jitter(alpha = 0.25)+
  scale_x_continuous(limits = c(60,80))+
  scale_y_continuous(limits = c (60,80))+
  coord_fixed()+
  theme_bw()
```



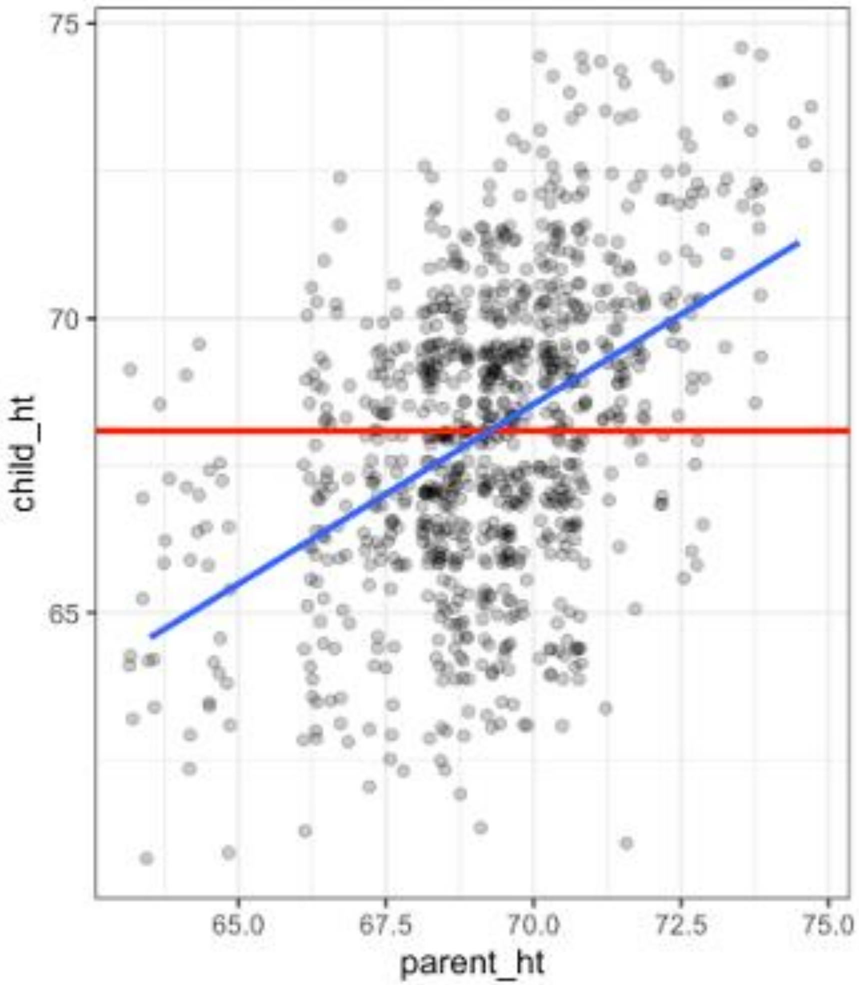
Galton's height data

```
child_mean = mean(height_data$child_ht)

height_data %>%
  ggplot(aes(y=child_ht, x= parent_ht))+
  geom_jitter(alpha = 0.25)+
  scale_x_continuous(limits = c(60,80))+
  scale_y_continuous(limits = c (60,80))+
  geom_hline(yintercept = child_mean, color="red")+
  coord_fixed()+
  theme_bw()
```



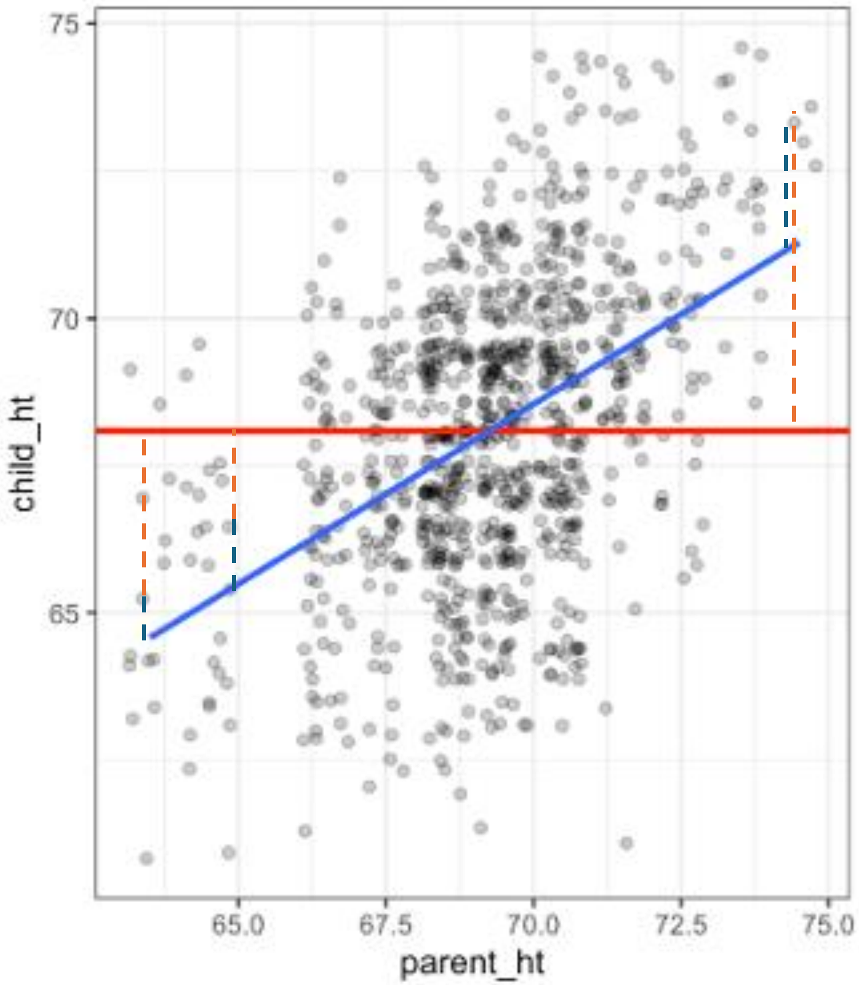
Galton's height data

```
height_data %>%
  ggplot(aes(y=child_ht, x= parent_ht))+
  geom_jitter(alpha = 0.25)+
  geom_hline(yintercept = 68.09, color="red", size = 0.9)+
  geom_smooth(method ="lm", se= FALSE) +
  coord_fixed()+
  theme_bw()
```



Galton’s height data

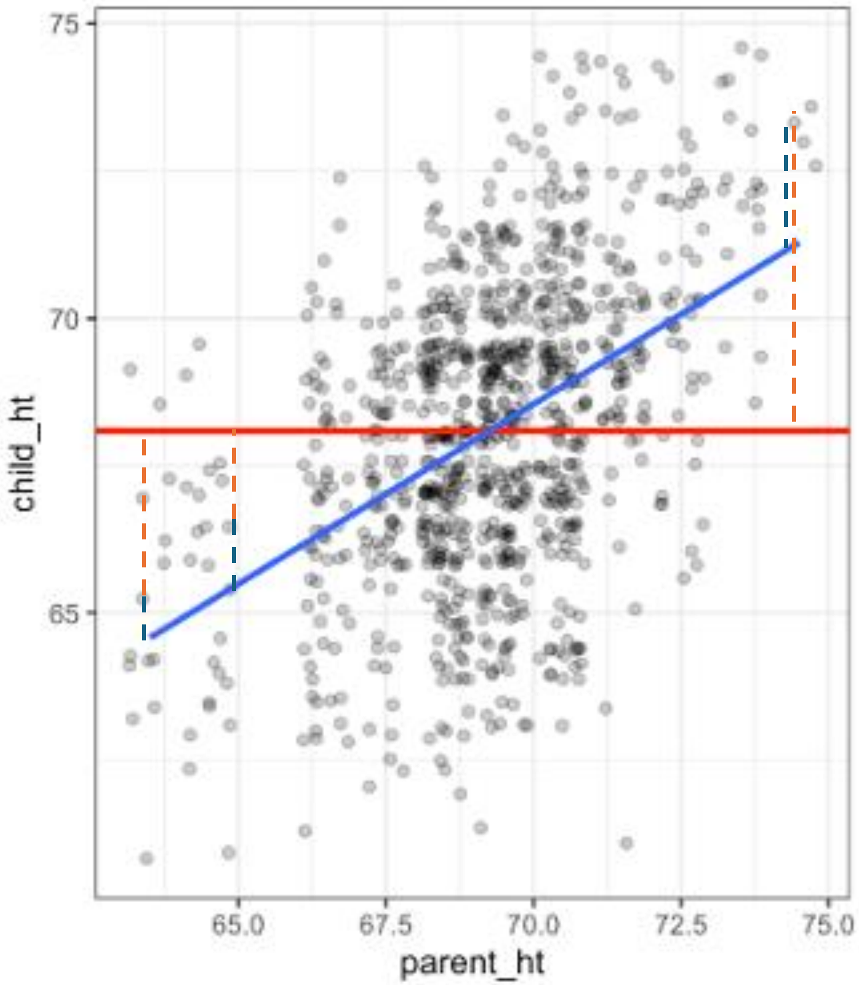
```
height_data %>%
  ggplot(aes(y=child_ht, x= parent_ht))+
  geom_jitter(alpha = 0.25)+
  geom_hline(yintercept = 68.09, color="red", size = 0.9)+
  geom_smooth(method ="lm", se= FALSE) +
  coord_fixed()+
  theme_bw()
```



Galton’s height data

```r
height_data %>%
  ggplot(aes(y=child_ht, x= parent_ht))+
  geom_jitter(alpha = 0.25)+
  geom_hline(yintercept = 68.09, color="red", size = 0.9)+
  geom_smooth(method ="lm", se= FALSE) +
  coord_fixed()+
  theme_bw()
```

$\hat{Y}_i$ = intercept + slope * parent_ht

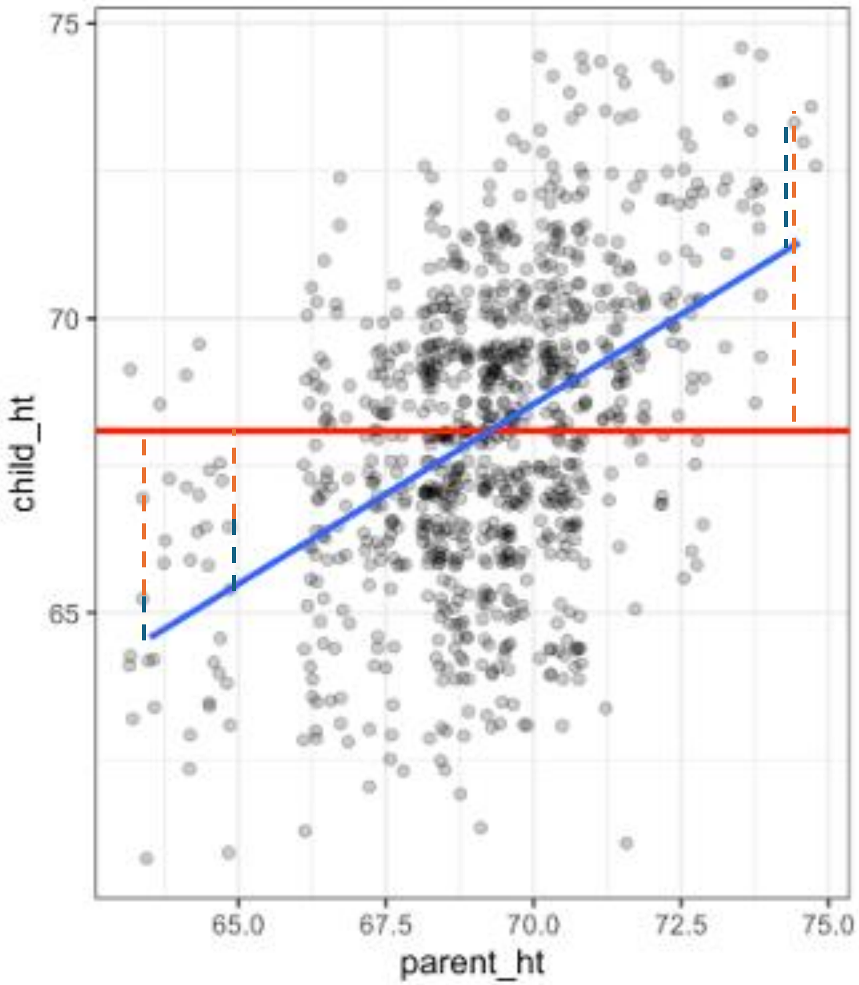


Galton's height data

```
height_data %>%
  ggplot(aes(y=child_ht, x= parent_ht))+
  geom_jitter(alpha = 0.25)+
  geom_hline(yintercept = 68.09, color="red", size = 0.9)+
  geom_smooth(method ="lm", se= FALSE) +
  coord_fixed()+
  theme_bw()
```

$\hat{Y}_i$ = intercept + slope * parent_ht

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon$$



Galton's height data

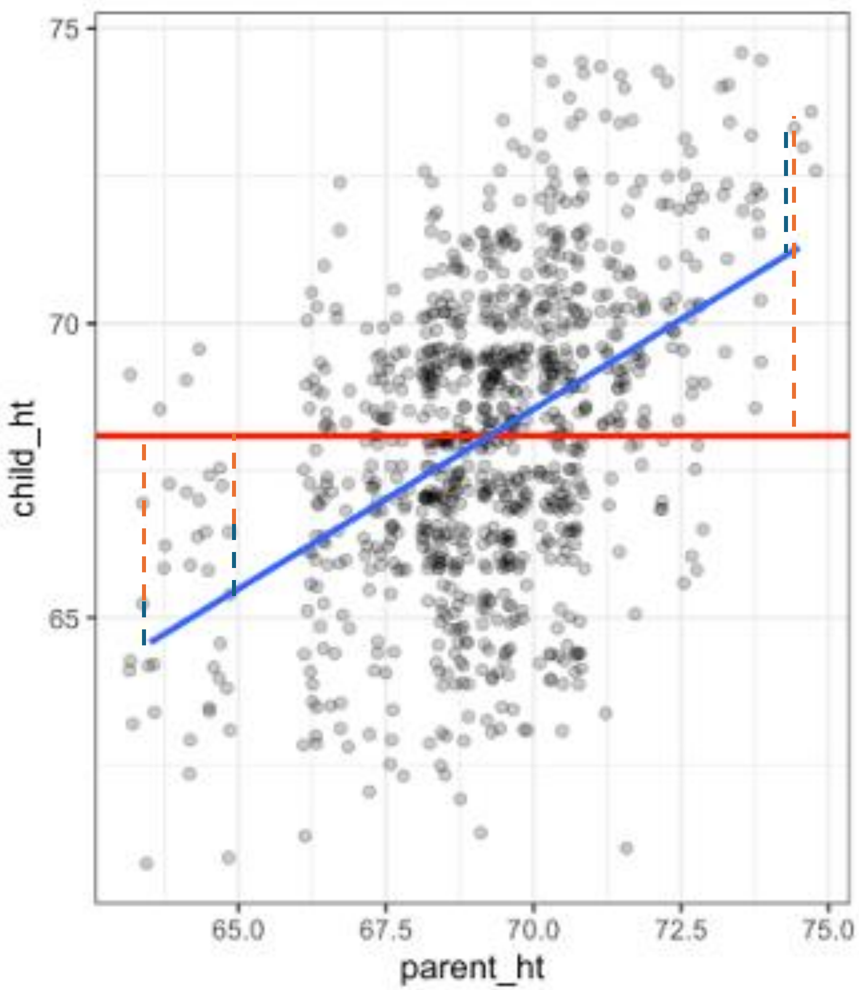
```
height_data %>%
  ggplot(aes(y=child_ht, x= parent_ht))+
  geom_jitter(alpha = 0.25)+
  geom_hline(yintercept = 68.09, color="red", size = 0.9)+
  geom_smooth(method ="lm", se= FALSE) +
  coord_fixed()+
  theme_bw()
```

$\hat{Y}_i$ = intercept + slope * parent_ht

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon$$

vs

$\hat{Y}_i = \bar{Y}_i$



Galton’s height data

```
height_data %>%
  ggplot(aes(y=child_ht, x= parent_ht))+
  geom_jitter(alpha = 0.25)+
  geom_hline(yintercept = 68.09, color="red", size = 0.9)+
  geom_smooth(method ="lm", se= FALSE) +
  coord_fixed()+
  theme_bw()
```

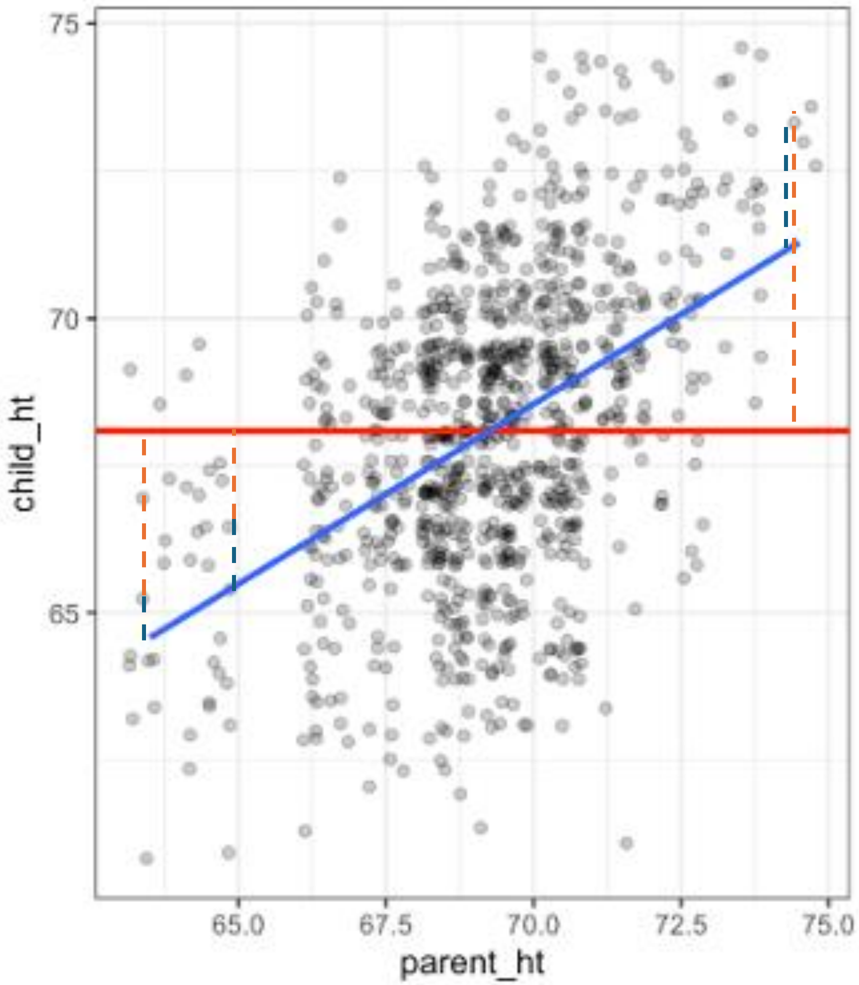
$\hat{Y}_i$ = intercept + slope * parent_ht

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon$$

vs

$\hat{Y}_i = \bar{Y}_i$

$$Y_i = \beta_0 + \epsilon$$



Galton's height data

```
height_data %>%
  ggplot(aes(y=child_ht, x= parent_ht))+
  geom_jitter(alpha = 0.25)+
  geom_hline(yintercept = 68.09, color="red", size = 0.9)+
  geom_smooth(method ="lm", se= FALSE) +
  coord_fixed()+
  theme_bw()
```

Model 1

$\hat{Y}_i$ = intercept + slope * parent_ht

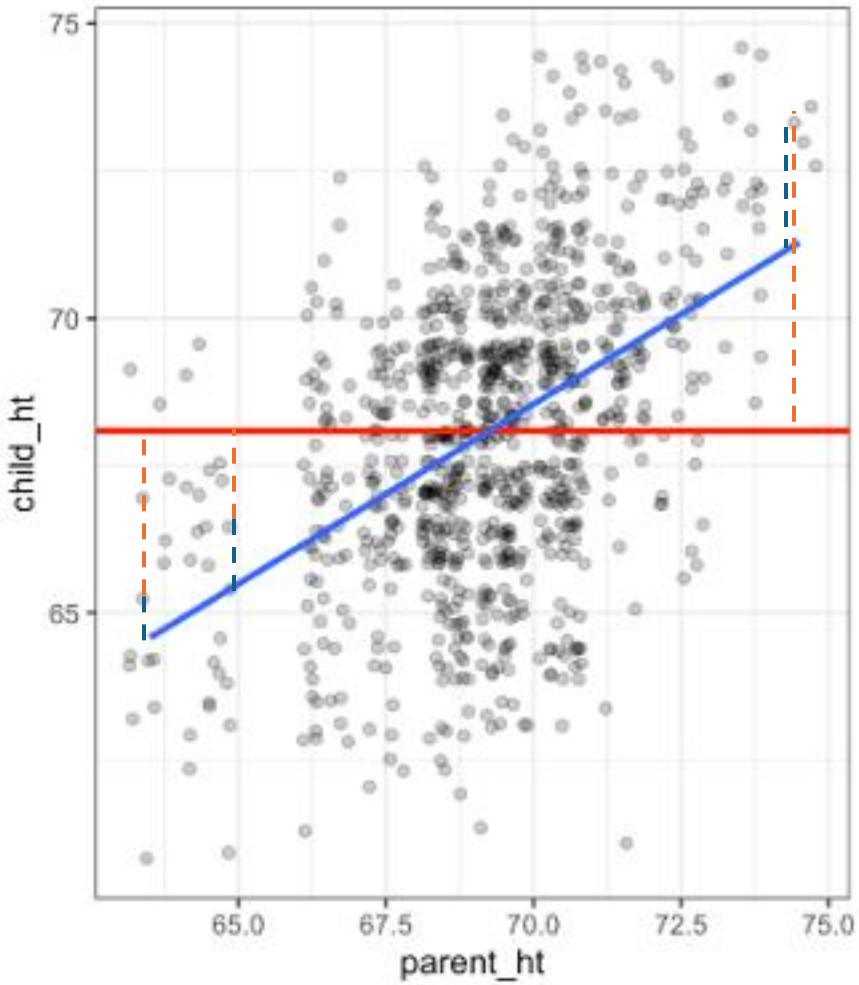
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon$$

vs

Model 2

$\hat{Y}_i = \bar{Y}_i$

$$Y_i = \beta_0 + \epsilon$$



Galton's height data

```
height_data %>%
  ggplot(aes(y=child_ht, x= parent_ht))+
  geom_jitter(alpha = 0.25)+
  geom_hline(yintercept = 68.09, color="red", size = 0.9)+
  geom_smooth(method ="lm", se= FALSE) +
  coord_fixed()+
  theme_bw()
```

Model 1

$\hat{Y}_i$ = intercept + slope * parent_ht

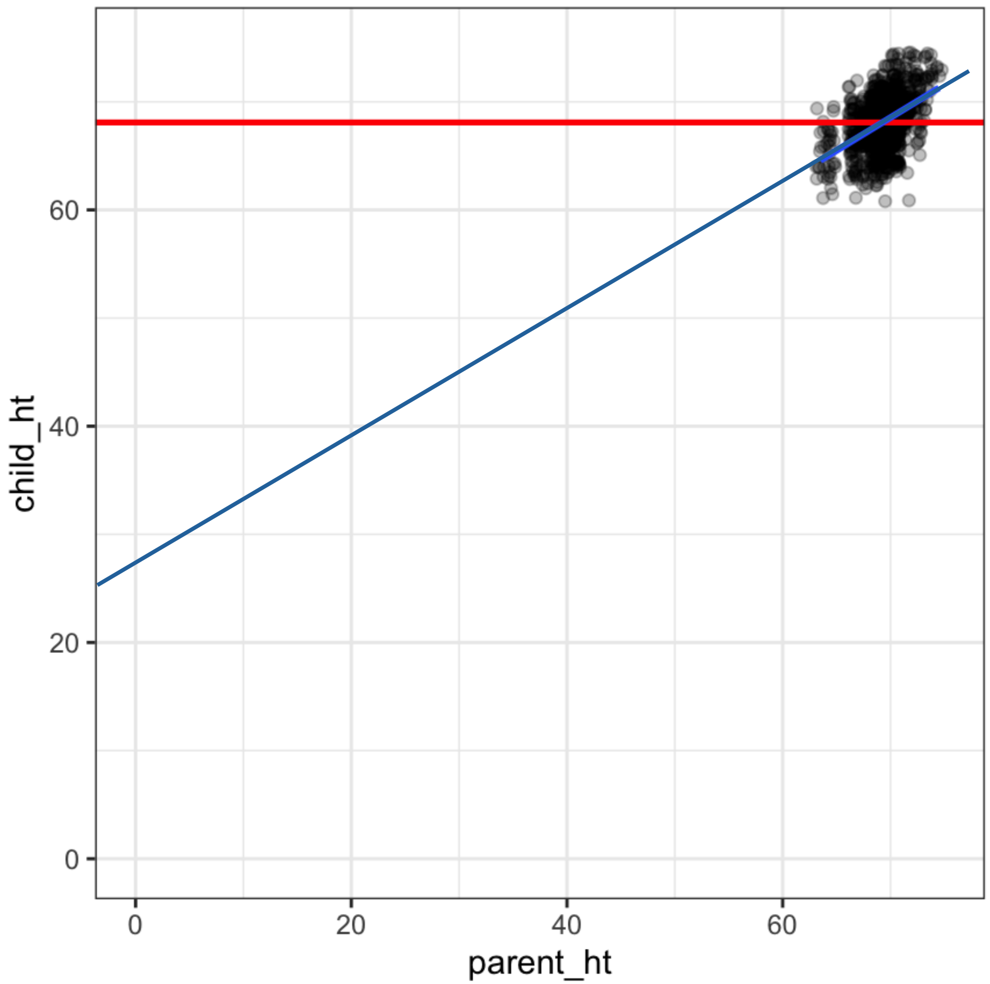
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon$$

vs

Model 2

$\hat{Y}_i = \bar{Y}_i$

$$Y_i = \beta_0 + \epsilon$$



Galton’s height data

- Linear Regression (prediction for Y changes linearly with X)

Model 1

$\hat{Y}_i$ = intercept + slope * parent_ht

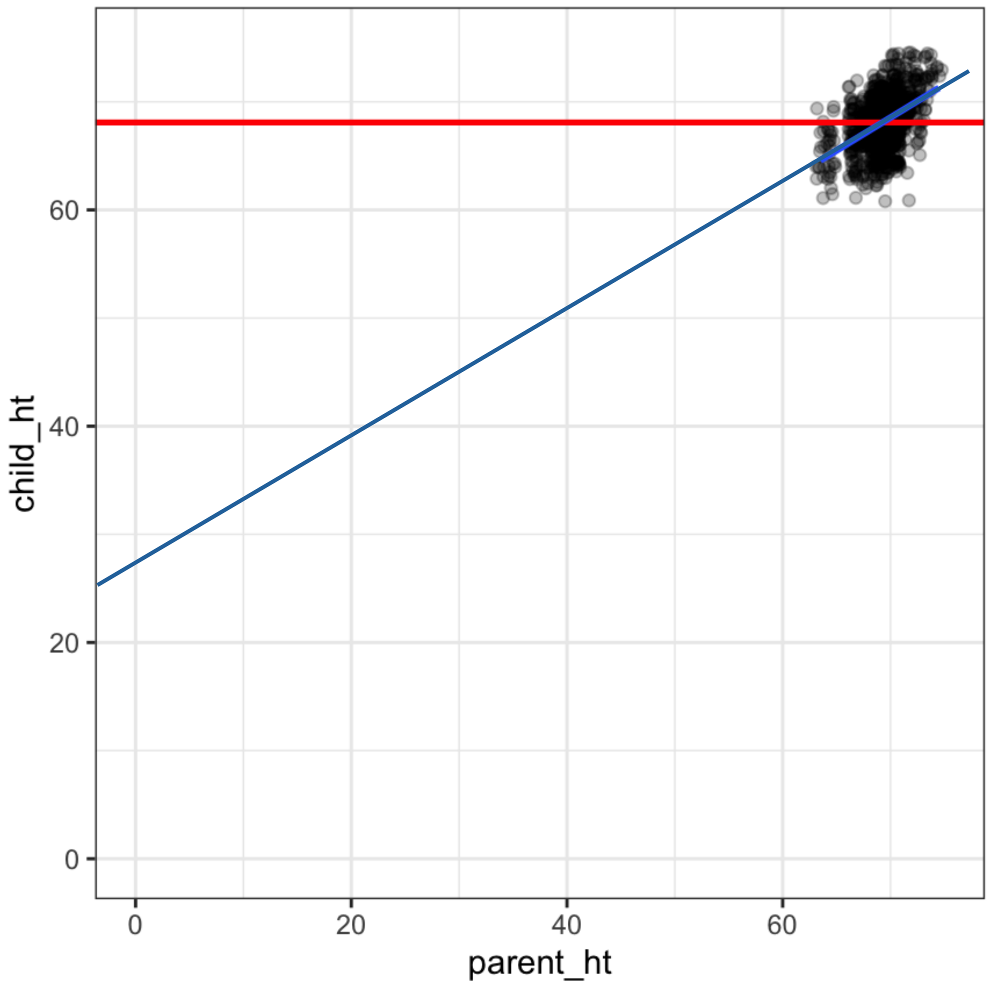
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon$$

vs

Model 2

$$\hat{Y}_i = \bar{Y}_i$$

$$Y_i = \beta_0 + \epsilon$$



Galton’s height data

- Linear Regression (prediction for Y changes linearly with X)
- Be careful about extrapolation

Model 1

$\hat{Y}_i$ = intercept + slope * parent_ht

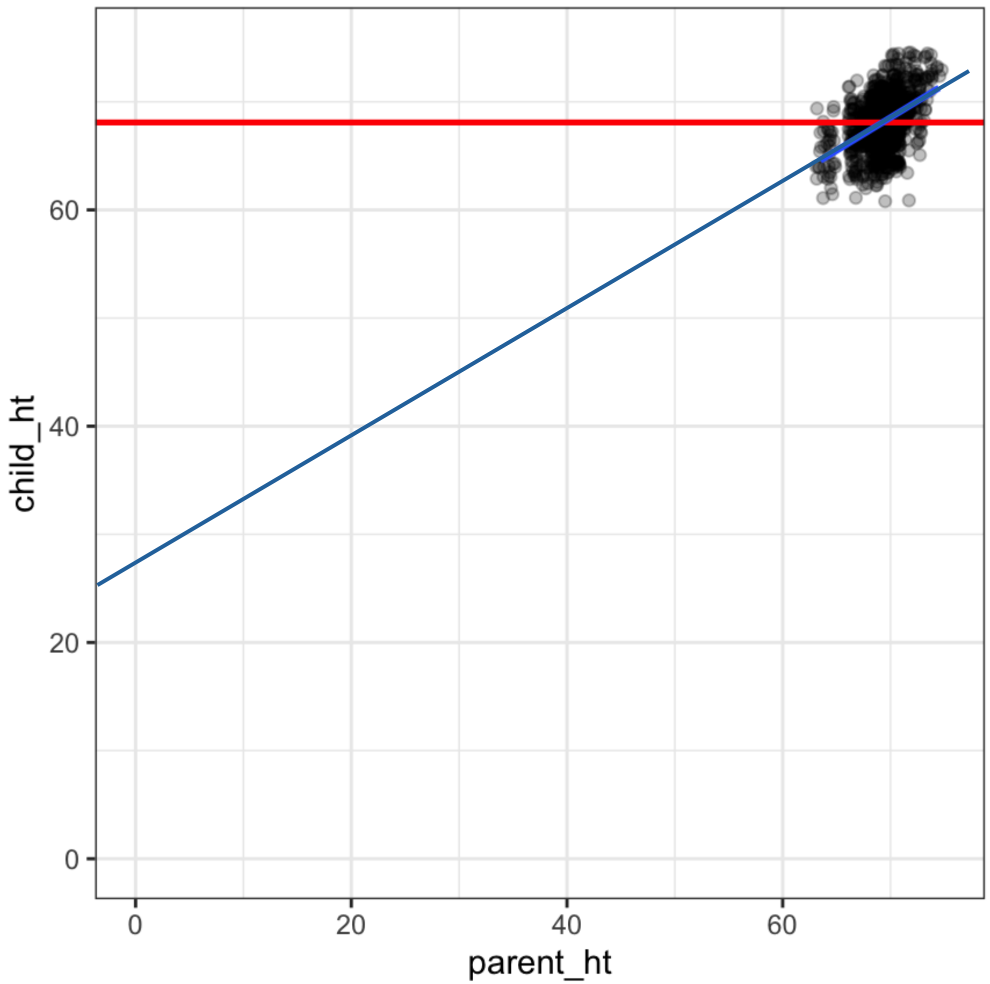
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon$$

vs

Model 2

$\hat{Y}_i = \bar{Y}_i$

$$Y_i = \beta_0 + \epsilon$$



Galton’s height data

- Linear Regression (prediction for Y changes linearly with X)
- Be careful about extrapolation
- Connection between the slope and how correlated data are?

Model 1

$\hat{Y}_i$ = intercept + slope * parent_ht

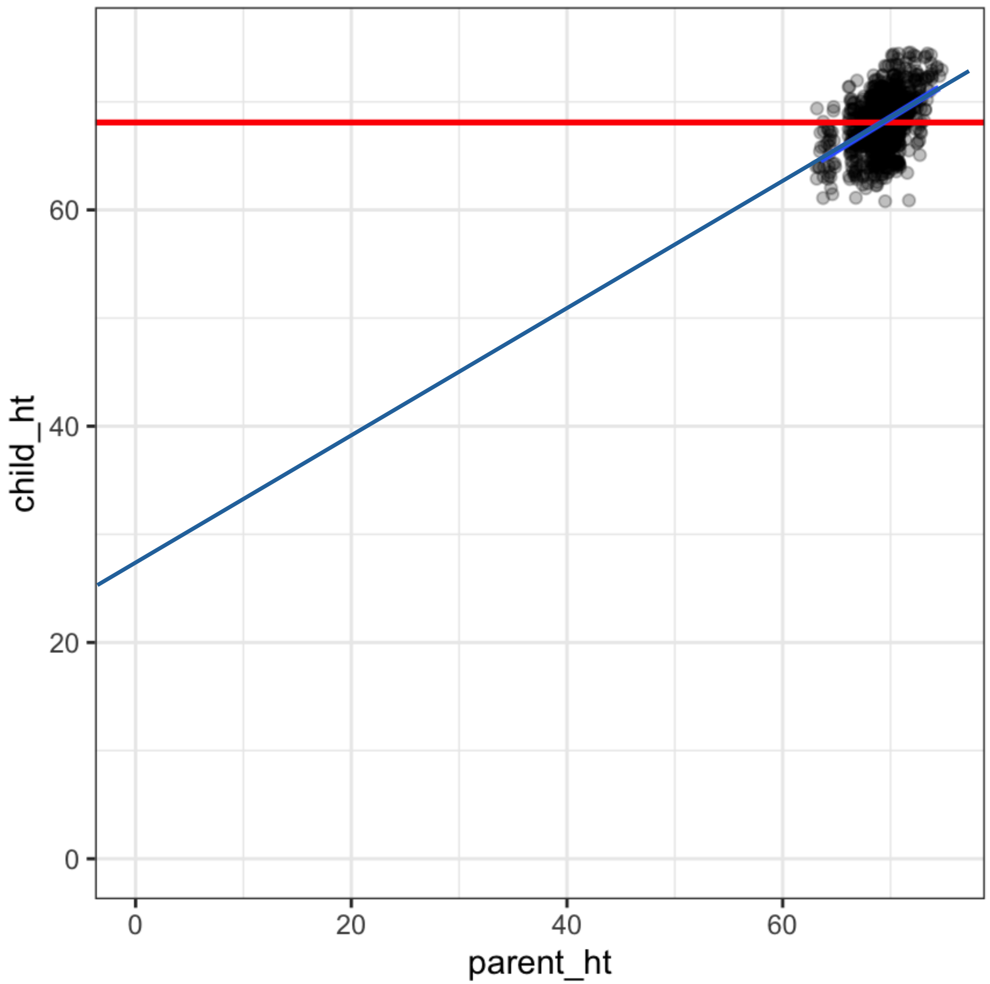
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon$$

vs

Model 2

$\hat{Y}_i = \bar{Y}_i$

$$Y_i = \beta_0 + \epsilon$$



Galton's height data

- Linear Regression (prediction for Y changes linearly with X)
- Be careful about extrapolation
- Connection between the slope and how correlated data are?
    - direction (positive vs negative slope)

Model 1

$\hat{Y}_i$ = intercept + slope * parent_ht

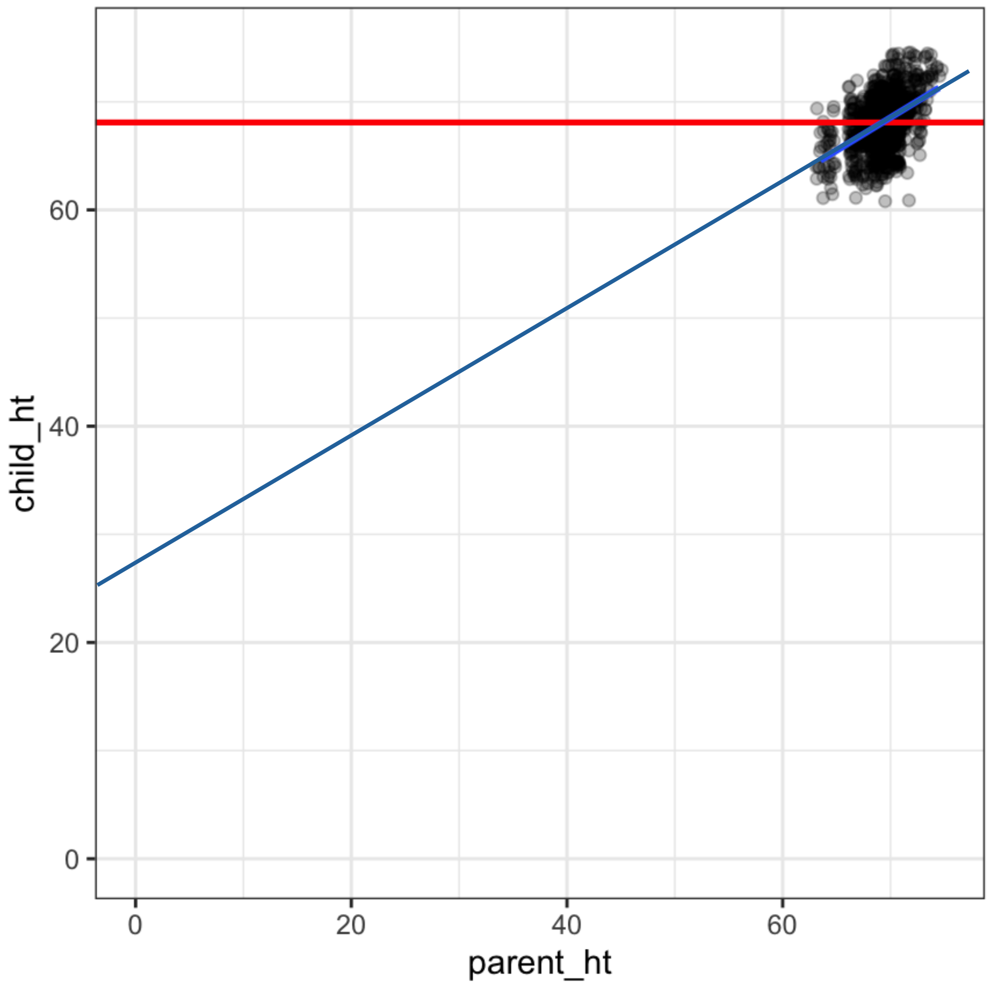
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon$$

vs

Model 2

$\hat{Y}_i = \bar{Y}_i$

$$Y_i = \beta_0 + \epsilon$$



Galton's height data

- Linear Regression (prediction for Y changes linearly with X)
- Be careful about extrapolation
- Connection between the slope and how correlated data are?
  - direction (positive vs negative slope)
  - steepness

Model 1

$\hat{Y}_i$ = intercept + slope * parent_ht

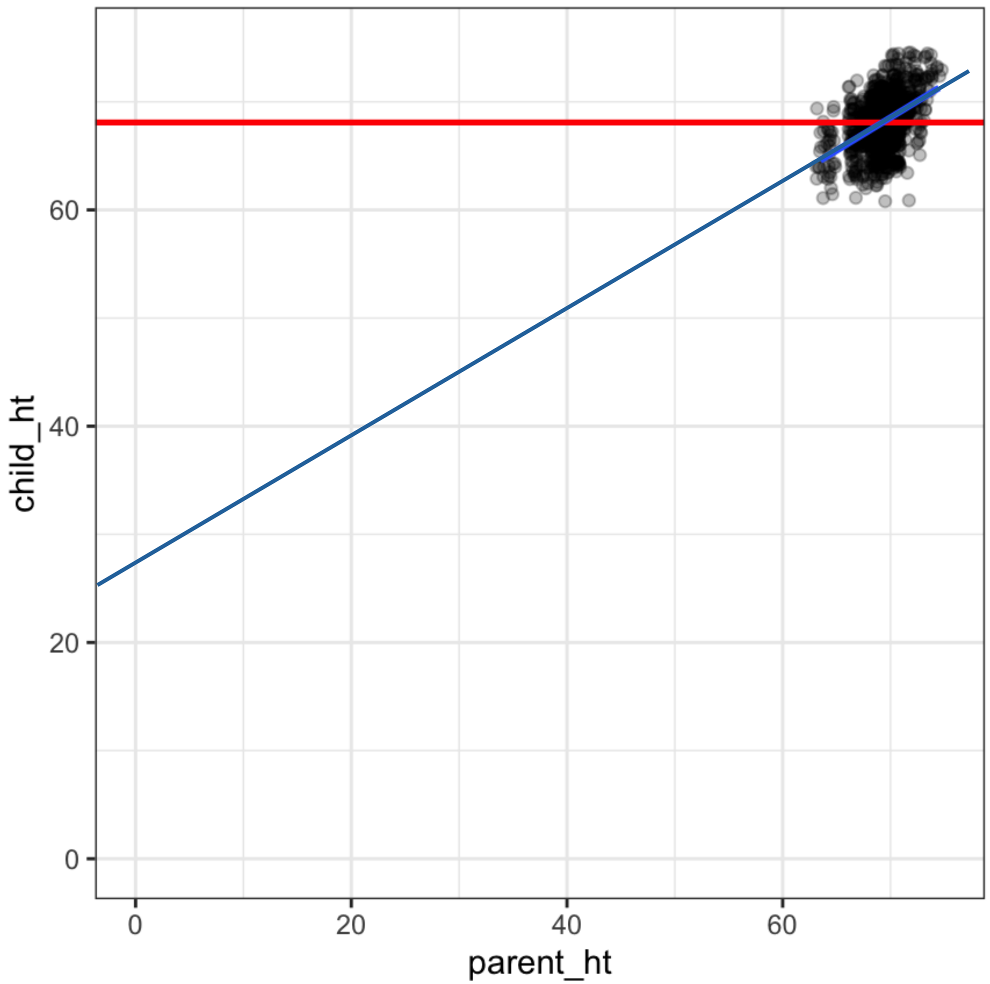
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon$$

vs

Model 2

$\hat{Y}_i = \bar{Y}_i$

$$Y_i = \beta_0 + \epsilon$$



Galton's height data

Model 1

$\hat{Y}_i$ = intercept + slope * parent_ht

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon$$

y ~ x + 1 ; y ~ x

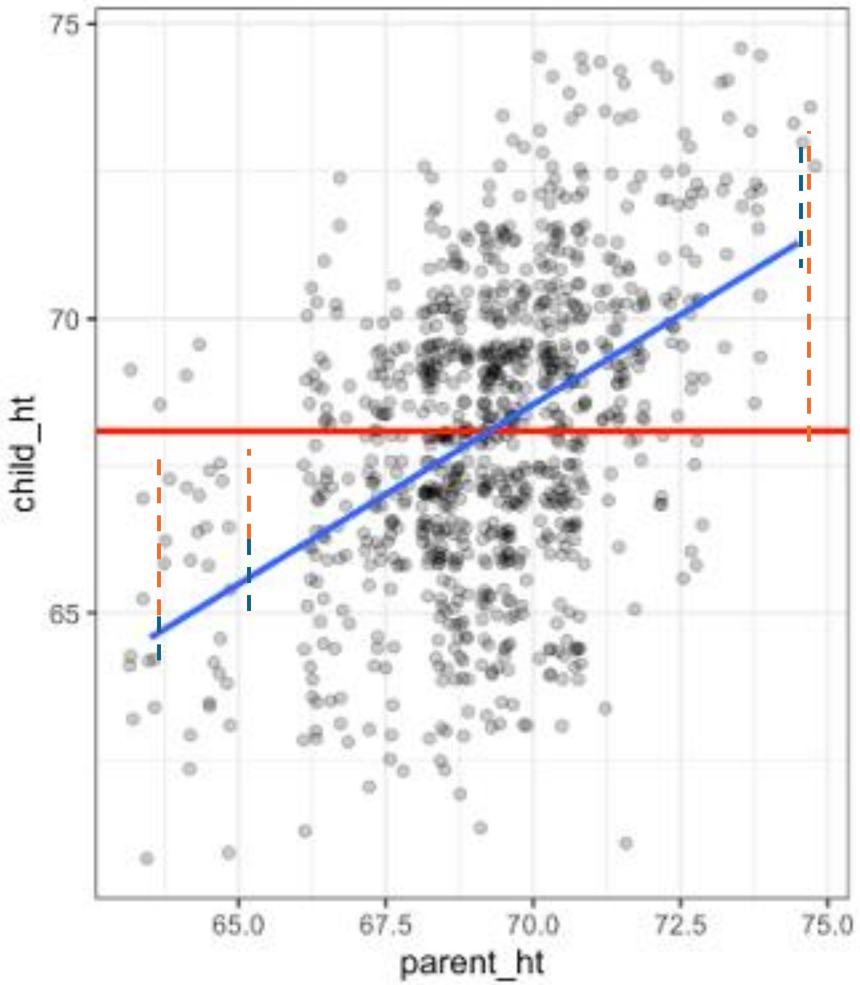
vs

Model 2

$\hat{Y}_i = \bar{Y}_i$

$$Y_i = \beta_0 + \epsilon$$

y ~ 1 ; y ~ NULL



Galton's height data

M1 = lm(child_height ~ parent_height +1, data = galton_data)

Model 1

$\hat{Y}_i$ = intercept + slope * parent_ht

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon$$

y ~ x + 1 ; y ~ x

vs

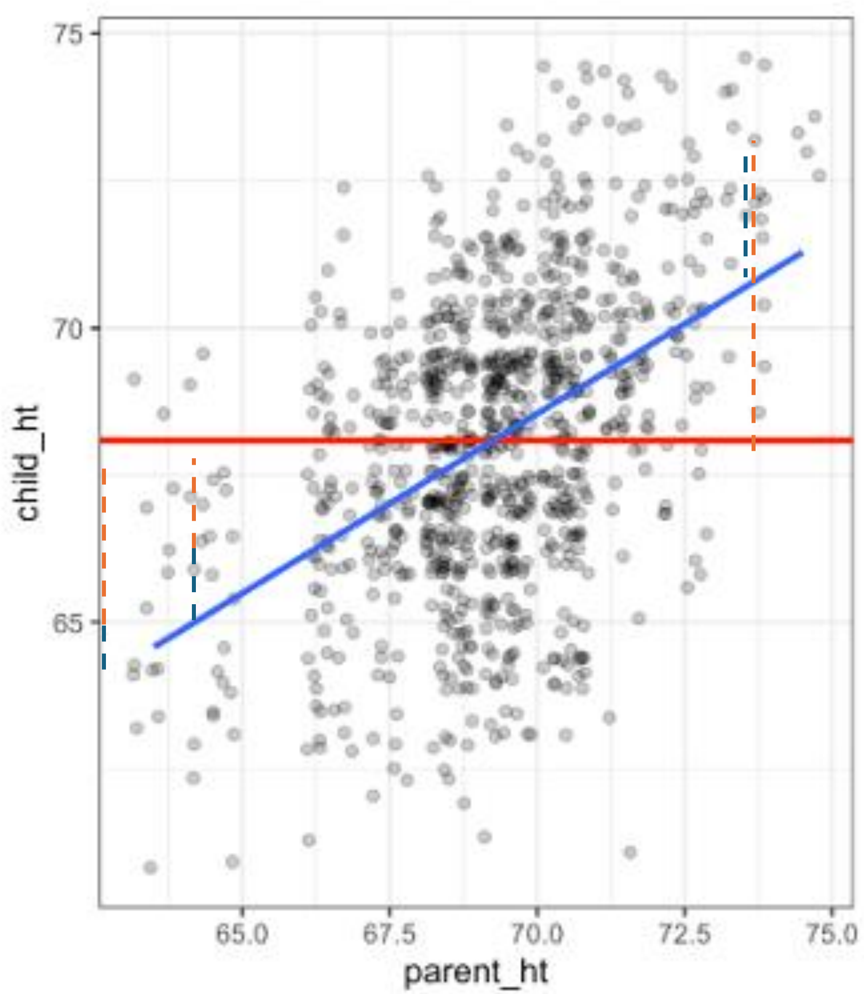
M2 = lm(child_height ~1, data = galton_data)

Model 2

$\hat{Y}_i = \bar{Y}_i$

$$Y_i = \beta_0 + \epsilon$$

y ~ 1 ; y ~ NULL



Fitting Models in R

````
```{r}
M1_SlopeAndIntercept <- lm(
  child_ht ~ parent_ht, data= height_data
)

M1_SlopeAndIntercept
```
````

```
Call:
lm(formula = child_ht ~ parent_ht, data = height_data)

Coefficients:
(Intercept)    parent_ht
    25.8486       0.6099
```

Fitting Models in R

````
```{r}
M1_SlopeAndIntercept <- lm(
  child_ht ~ parent_ht, data= height_data
)

M1_SlopeAndIntercept
```
````

```
Call:
lm(formula = child_ht ~ parent_ht, data = height_data)

Coefficients:
(Intercept)    parent_ht
    25.8486       0.6099
```

````
```{r}
M2_Mean <- lm(
  child_ht ~ 1, data= height_data
)

M2_Mean
```
````

```
Call:
lm(formula = child_ht ~ 1, data = height_data)

Coefficients:
(Intercept)
      68.09
```

Model fit summary using summary ()

```{r}
summary(M1_SlopeAndIntercept)
```

```
Call:
lm(formula = child_ht ~ parent_ht, data = height_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.2577 -1.4280  0.1323  1.5720  5.7918 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 25.84856    2.69009   9.609   <2e-16 ***
parent_ht    0.60992    0.03882  15.710   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.26 on 926 degrees of freedom
Multiple R-squared:  0.2104,	Adjusted R-squared:  0.2096 
F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

Model fit summary using summary ()

```{r}
summary(M1_SlopeAndIntercept)
```

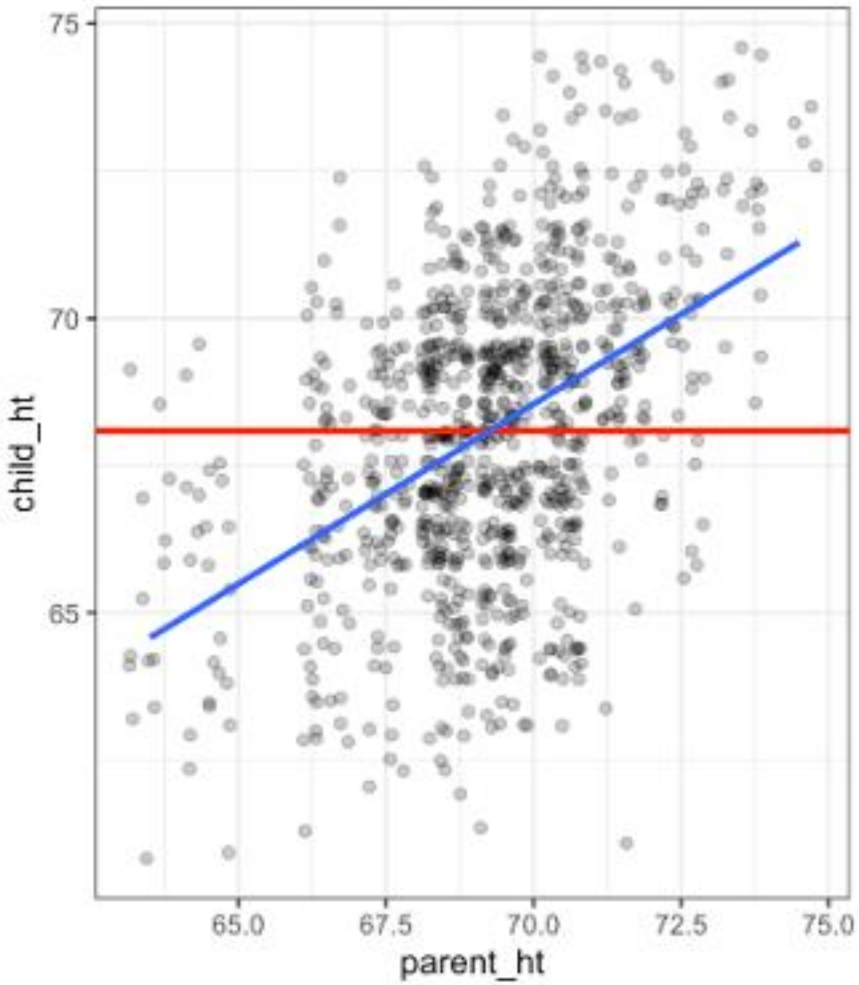
```
Call:
lm(formula = child_ht ~ parent_ht, data = height_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.2577 -1.4280  0.1323  1.5720  5.7918 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 25.84856    2.69009   9.609   <2e-16 ***
parent_ht    0.60992    0.03882  15.710   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.26 on 926 degrees of freedom
Multiple R-squared:  0.2104,	Adjusted R-squared:  0.2096 
F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

```{r}
summary(M2_Mean)
```

```
Call:
lm(formula = child_ht ~ 1, data = height_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.8933 -1.8933  0.1067  2.1067  6.1067 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 68.09332    0.08346   815.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.543 on 927 degrees of freedom
```

Making predictions in R with predict()

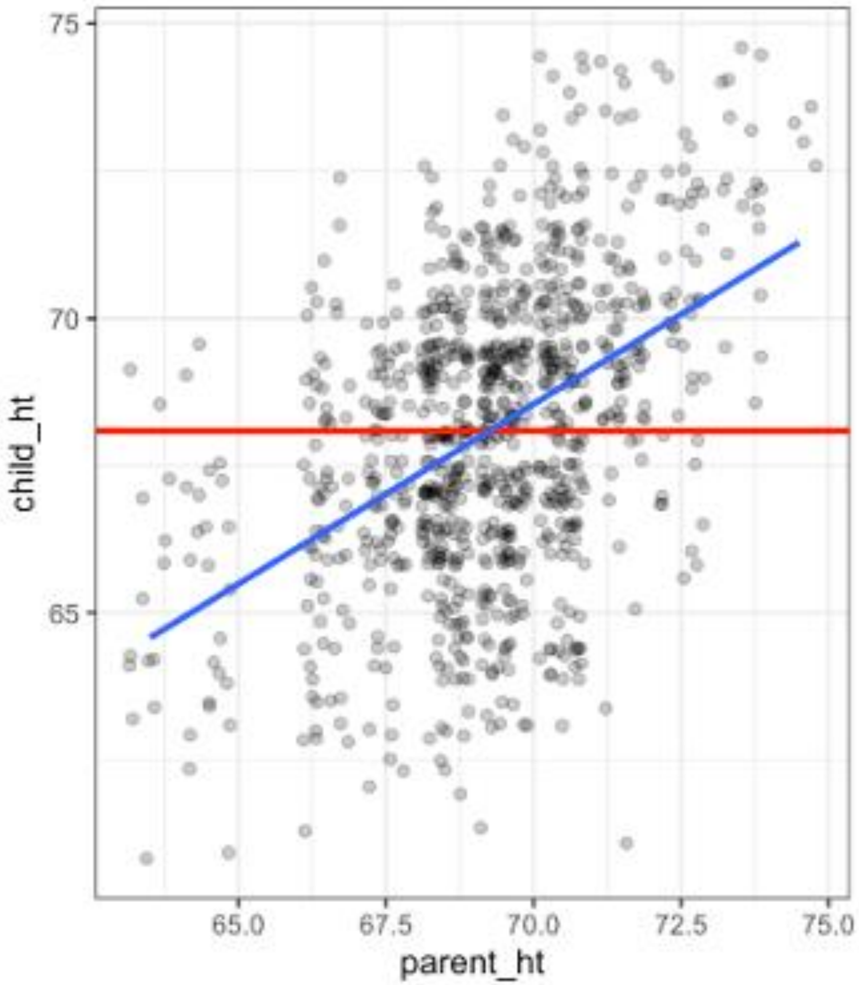
```{r}
tall_parents <- tibble(
  parent_ht = 80
)

predict(M1_SlopeAndIntercept, tall_parents)
predict (M2_Mean, tall_parents)
```



Making predictions in R with predict()

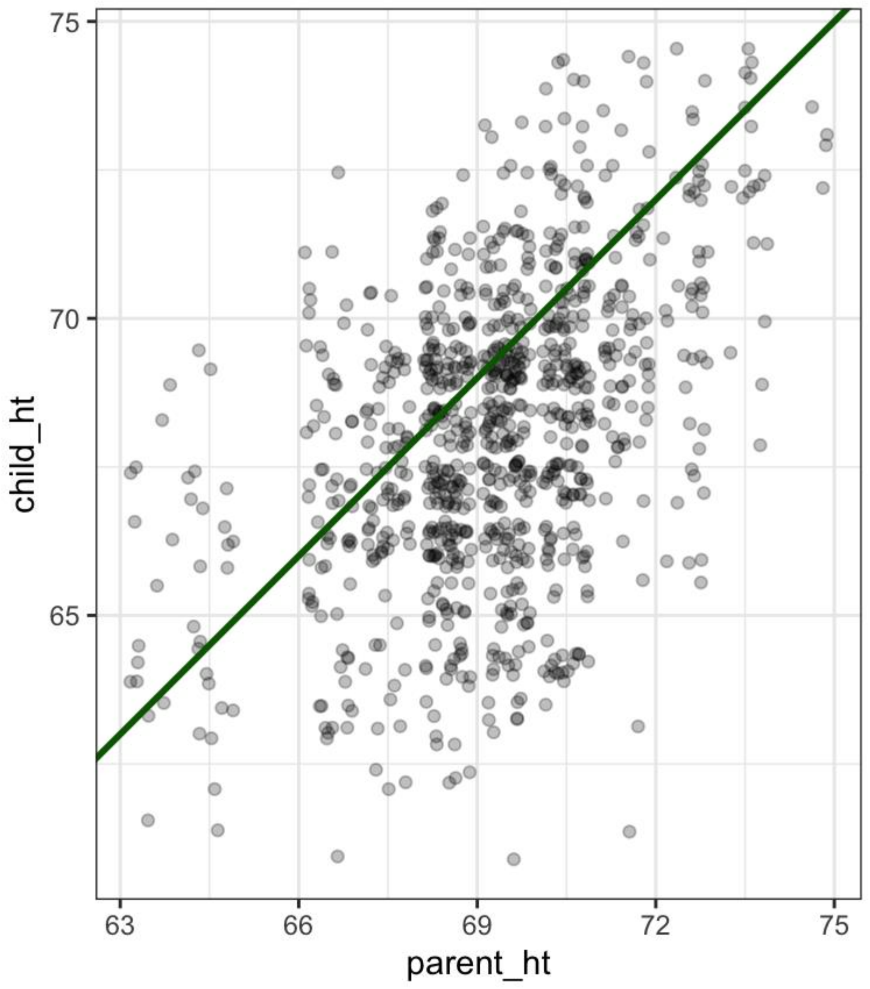
```{r}
tall_parents <- tibble(
  parent_ht = 80
)

predict(M1_SlopeAndIntercept, tall_parents)
predict (M2_Mean, tall_parents)
```

```
       1
74.64206
       1
68.09332
```



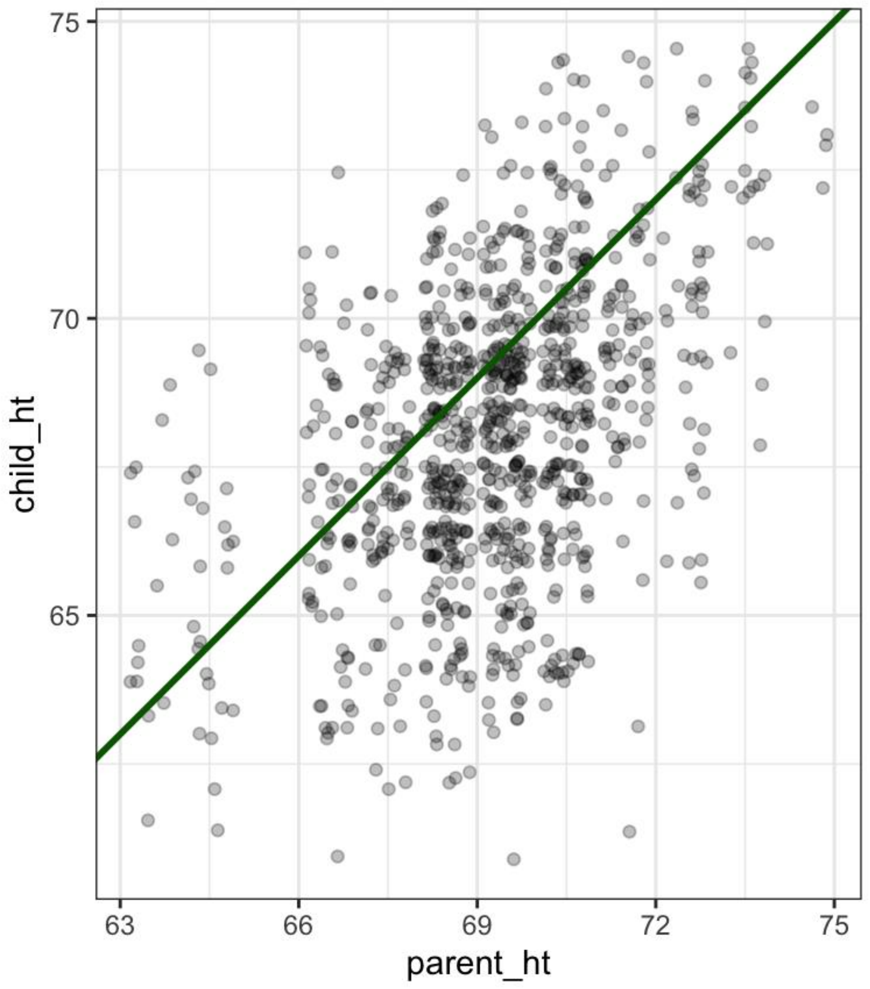
Aside: Regression to the mean

- First observed by Galton using this very dataset



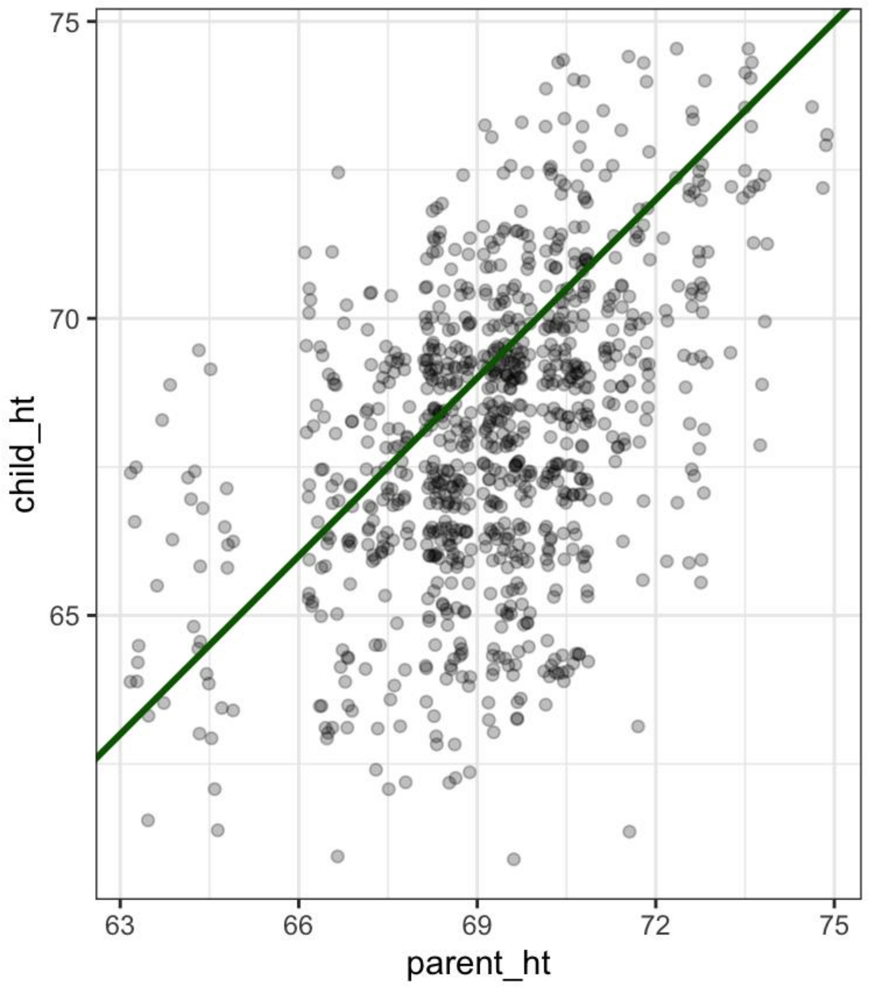
Aside: Regression to the mean

- First observed by Galton using this very dataset
- Property of data, not the model
  - tendency for extreme values in one measurement to be closer to the average in a second measurement
  - Very short parents have taller children
  - Very tall parents have shorter children



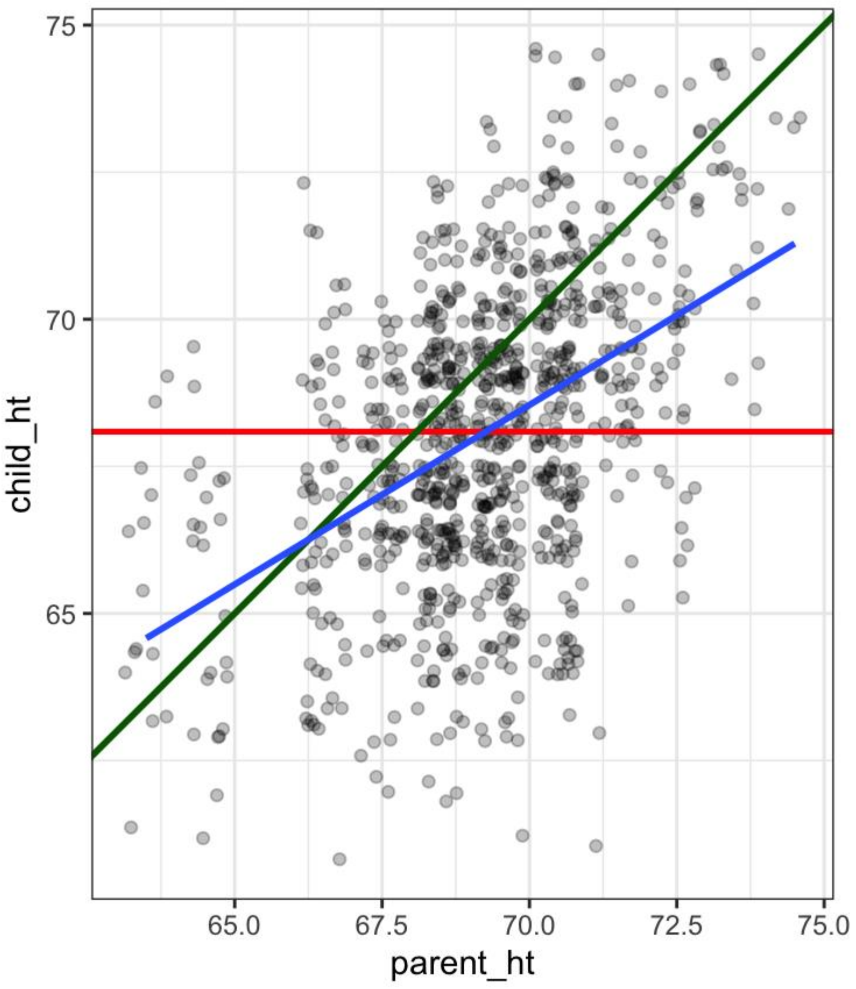
Aside: Regression to the mean

- First observed by Galton using this very dataset
- Property of data, not the model
  - tendency for extreme values in one measurement to be closer to the average in a second measurement
  - Very short parents have taller children
  - Very tall parents have shorter children
- Why?
  - parent height influenced by - Causal factors (genetic, environmental, health, socioeconomic) + natural variation + measurement error
  - child inherits only a portion of these
  - Correlation is imperfect



Aside: Regression to the mean

- First observed by Galton using this very dataset
- Property of data, not the model
  - tendency for extreme values in one measurement to be closer to the average in a second measurement
  - Very short parents have taller children
  - Very tall parents have shorter children
- Why?
  - parent height influenced by - Causal factors (genetic, environmental, health, socioeconomic) + natural variation + measurement error
  - child inherits only a portion of these
  - Correlation is imperfect



Other examples of regression to the mean?

Other examples of regression to the mean?

- Sports
  - A basketball player scores unusually high in one game but returns to their average scoring in the subsequent match

- Praise vs. Criticism
  - trainers often criticize poor performers and praise high performers.
  - worst performers do better subsequently while the top ones do worse.
  - This pattern leads to the mistaken conclusion that criticism boosts performance more than praise.

Many more examples: https://statisticsbyjim.com/basics/regression-to-the-mean/