Workbook 03

Preston Lyons

10/03/2021

## Plot Missing Data Problem

Missing Data Code Output:

library(“tidyverse”) ggplot(data = mtcars) + aes(mpg, wt, colour=factor(cyl))

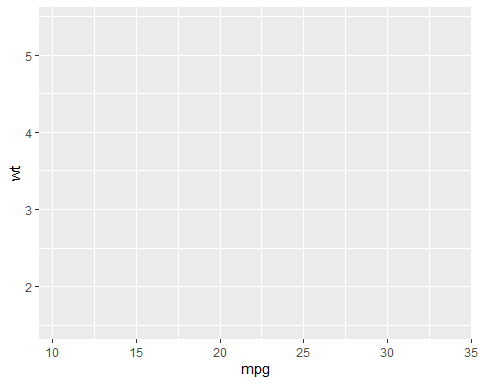
library("tidyverse")

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3 v purrr 0.3.4  
## v tibble 3.0.6 v dplyr 1.0.4  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

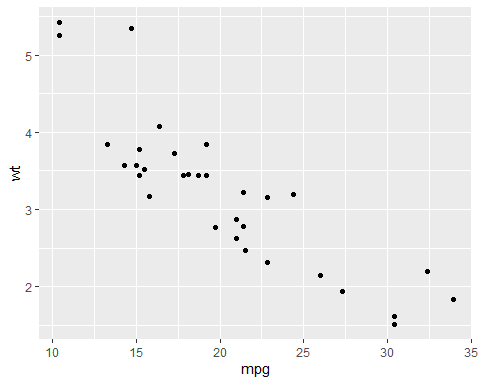
ggplot(data = mtcars) +   
 aes(mpg, wt, colour=factor(cyl))

 What appears to be missing is the “geom\_point” command preceding the aesthetic (“aes”) command, and then defining which variables go into which axes. Including that code looks like this:

ggplot(data = mtcars) + geom\_point(mapping = aes(x = mpg, y = wt))

And it produces this:

ggplot(data = mtcars) +   
geom\_point(mapping = aes(x = mpg, y = wt))

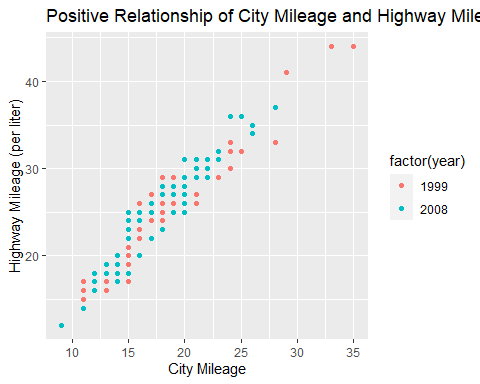


## Using the mpg dataset, graph the relationship between city milage and highway mileage by year manufacture

head(mpg)

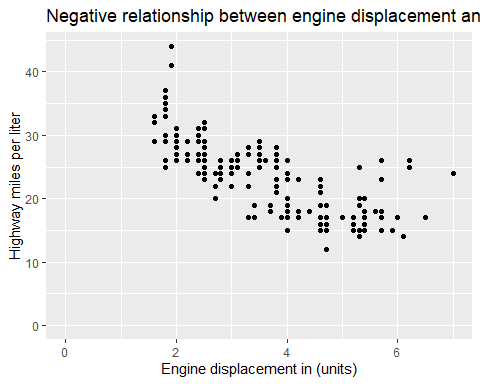
## # A tibble: 6 x 11  
## manufacturer model displ year cyl trans drv cty hwy fl class   
## <chr> <chr> <dbl> <int> <int> <chr> <chr> <int> <int> <chr> <chr>   
## 1 audi a4 1.8 1999 4 auto(l5) f 18 29 p compa~  
## 2 audi a4 1.8 1999 4 manual(m5) f 21 29 p compa~  
## 3 audi a4 2 2008 4 manual(m6) f 20 31 p compa~  
## 4 audi a4 2 2008 4 auto(av) f 21 30 p compa~  
## 5 audi a4 2.8 1999 6 auto(l5) f 16 26 p compa~  
## 6 audi a4 2.8 1999 6 manual(m5) f 18 26 p compa~

ggplot(data = mpg) +   
geom\_point(mapping = aes(x = cty, y = hwy, color=factor(year)))+  
 labs(title ="Positive Relationship of City Mileage and Highway Mileage for each Manufacture Year.") +   
 xlab("City Mileage") +   
 ylab("Highway Mileage (per liter)")



## Edit this graph so that the x axis and the y axis both start at 0

ggplot(data = mpg) +   
 geom\_point(mapping = aes(x = displ, y = hwy)) +   
 labs(title ="Negative relationship between engine displacement and fuel efficiency.") +   
 xlab("Engine displacement in (units)") +   
 ylab("Highway miles per liter") + expand\_limits(x = 0, y = 0)



## What is one benefit and one limitation for this graph above (in which the x and y values start at 0?)

Limitation: if the data doesn’t start close to zero, it’ll make what’s going on in the data (viz., patterns) harder to see in the graph

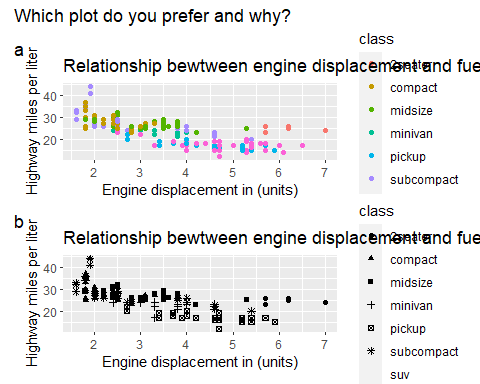
Advantage: it provides an honest look at the data, and avoids the risk of things being (or, less synically, appearing) manipulative to the reader. Marked differences found in the data can be relied on, by not starting at 0 differences may seem more pronounced, so starting at 0 avoids this problem.

## Which of these two graphs do you prefer and why?

g1 <-ggplot(data = mpg) +   
 geom\_point(mapping = aes(x = displ, y = hwy, colour = class )) +   
 labs(title = "Relationship bewtween engine displacement and fuel efficiency in the mpg automobile dataset") +   
 xlab("Engine displacement in (units)") +   
 ylab("Highway miles per liter")   
  
g2 <-ggplot(data = mpg) +   
 geom\_point(mapping = aes(x = displ, y = hwy, shape = class )) +   
 labs(title = "Relationship bewtween engine displacement and fuel efficiency in the mpg automobile dataset") +   
 xlab("Engine displacement in (units)") +   
 ylab("Highway miles per liter")  
  
library("patchwork")  
  
g1 / g2 + plot\_annotation(title = "Which plot do you prefer and why?", tag\_levels = 'a')

## Warning: The shape palette can deal with a maximum of 6 discrete values because  
## more than 6 becomes difficult to discriminate; you have 7. Consider  
## specifying shapes manually if you must have them.

## Warning: Removed 62 rows containing missing values (geom\_point).

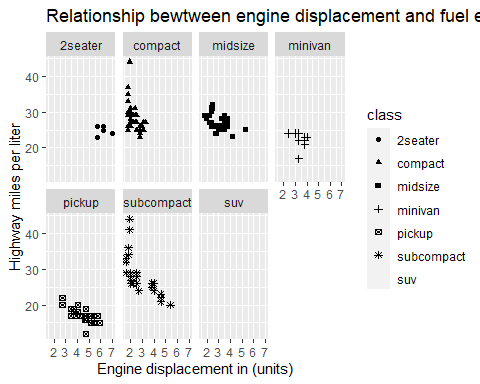
 I prefer the colourful graph (the first; g1) as it is easier to eyeball (informal inference) the graph and notice its patterns, and see where each type of vehicle is grouped.

## Add a facet to this graph for the “class” variable

ggplot(data = mpg) +   
 geom\_point(mapping = aes(x = displ, y = hwy, shape=class)) +   
 labs(title = "Relationship bewtween engine displacement and fuel efficiency in the mpg automobile dataset") +   
 xlab("Engine displacement in (units)") +   
 ylab("Highway miles per liter") +  
 facet\_wrap(~ class, nrow = 2)

## Warning: The shape palette can deal with a maximum of 6 discrete values because  
## more than 6 becomes difficult to discriminate; you have 7. Consider  
## specifying shapes manually if you must have them.

## Warning: Removed 62 rows containing missing values (geom\_point).

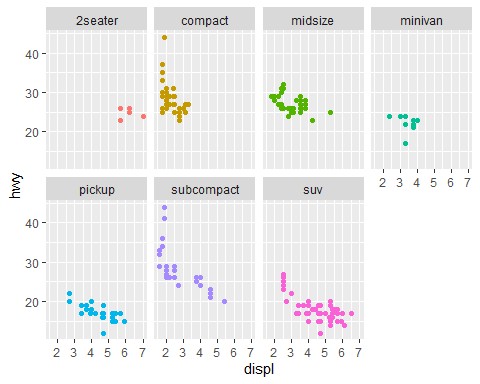


## Favourite Graph and Why

I actually prefer graph A - both have a great deal of utility, and should both be included in a publication if possible, however graph A

## Remove the legend

ggplot(data = mpg) +  
 geom\_point(mapping = aes(x = displ, y = hwy, color = class)) +  
 facet\_wrap( ~ class, nrow = 2) +  
 theme(legend.position = "none")



## Convert the y variable to “numeric” and graph the relationship betweeen religiousity (x-axis) and `thr\_mus’ (y-axis) in the ISSP dataset. Create new axis labels

# read the issp dataset for questionaire see: ISSP\_2018\_Religion\_Questionnaire\_final\_version1-2.pdf  
#   
# subset of data from the issp dataset  
issp <- readr::read\_csv2(url("https://raw.githubusercontent.com/go-bayes/psych-447/main/data/issp.csv"))

## i Using ',' as decimal and '.' as grouping mark. Use `read\_delim()` for more control.

##   
## -- Column specification --------------------------------------------------------  
## cols(  
## .default = col\_character(),  
## id = col\_double(),  
## age = col\_double(),  
## eduyears = col\_double(),  
## nzeuro = col\_double(),  
## rightwing = col\_double(),  
## thr\_ath = col\_double(),  
## wave = col\_double(),  
## religiosity = col\_double()  
## )  
## i Use `spec()` for the full column specifications.

head(issp)

## # A tibble: 6 x 21  
## id age male eduyears nzeuro rightwing neg\_ath neg\_bd neg\_ch neg\_hd   
## <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <chr> <chr> <chr> <chr>   
## 1 1 52 Not M~ 16 1 4 Neither ~ Somewh~ Somewh~ Somewh~  
## 2 1 53 Not M~ 13 1 3 Neither ~ Neithe~ Neithe~ Somewh~  
## 3 2 63 Not M~ 10 1 4 Neither ~ Neithe~ Somewh~ Neithe~  
## 4 2 64 Not M~ 12 1 1 Neither ~ Neithe~ Somewh~ Neithe~  
## 5 3 64 Male NA 1 8 Neither ~ Neithe~ Neithe~ Neithe~  
## 6 3 65 Male 16 1 7 Neither ~ Neithe~ Neithe~ Neithe~  
## # ... with 11 more variables: neg\_jw <chr>, neg\_ms <chr>, thr\_ath <dbl>,  
## # thr\_bd <chr>, thr\_ch <chr>, thr\_hd <chr>, thr\_jw <chr>, rural <chr>,  
## # thr\_ms <chr>, wave <dbl>, religiosity <dbl>

str(issp)

## spec\_tbl\_df [2,668 x 21] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ id : num [1:2668] 1 1 2 2 3 3 4 4 5 5 ...  
## $ age : num [1:2668] 52 53 63 64 64 65 29 30 41 42 ...  
## $ male : chr [1:2668] "Not Male" "Not Male" "Not Male" "Not Male" ...  
## $ eduyears : num [1:2668] 16 13 10 12 NA 16 11 13 11 14 ...  
## $ nzeuro : num [1:2668] 1 1 1 1 1 1 NA 1 1 1 ...  
## $ rightwing : num [1:2668] 4 3 4 1 8 7 8 8 6 7 ...  
## $ neg\_ath : chr [1:2668] "Neither negative nor positive" "Neither negative nor positive" "Neither negative nor positive" "Neither negative nor positive" ...  
## $ neg\_bd : chr [1:2668] "Somewhat negative" "Neither negative nor positive" "Neither negative nor positive" "Neither negative nor positive" ...  
## $ neg\_ch : chr [1:2668] "Somewhat negative" "Neither negative nor positive" "Somewhat negative" "Somewhat negative" ...  
## $ neg\_hd : chr [1:2668] "Somewhat negative" "Somewhat negative" "Neither negative nor positive" "Neither negative nor positive" ...  
## $ neg\_jw : chr [1:2668] "Somewhat negative" "Neither negative nor positive" "Neither negative nor positive" "Somewhat positive" ...  
## $ neg\_ms : chr [1:2668] "Somewhat negative" "Somewhat negative" "Somewhat negative" "Somewhat negative" ...  
## $ thr\_ath : num [1:2668] 2 2 1 1 2 2 1 1 3 NA ...  
## $ thr\_bd : chr [1:2668] "Not very threatening" "Somewhat threatening" "Not threatening at all" "Not threatening at all" ...  
## $ thr\_ch : chr [1:2668] "Not very threatening" "Somewhat threatening" "Somewhat threatening" "Somewhat threatening" ...  
## $ thr\_hd : chr [1:2668] "Not very threatening" "Somewhat threatening" "Not threatening at all" "Not threatening at all" ...  
## $ thr\_jw : chr [1:2668] "Not very threatening" "Somewhat threatening" "Not threatening at all" "Not threatening at all" ...  
## $ rural : chr [1:2668] "Not Rural" "Not Rural" "Not Rural" "Not Rural" ...  
## $ thr\_ms : chr [1:2668] "Not very threatening" "Somewhat threatening" "Somewhat threatening" "Not very threatening" ...  
## $ wave : num [1:2668] 2018 2019 2018 2019 2018 ...  
## $ religiosity: num [1:2668] 6 4 3 4 4 6 7 7 5 5 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. id = col\_double(),  
## .. age = col\_double(),  
## .. male = col\_character(),  
## .. eduyears = col\_double(),  
## .. nzeuro = col\_double(),  
## .. rightwing = col\_double(),  
## .. neg\_ath = col\_character(),  
## .. neg\_bd = col\_character(),  
## .. neg\_ch = col\_character(),  
## .. neg\_hd = col\_character(),  
## .. neg\_jw = col\_character(),  
## .. neg\_ms = col\_character(),  
## .. thr\_ath = col\_double(),  
## .. thr\_bd = col\_character(),  
## .. thr\_ch = col\_character(),  
## .. thr\_hd = col\_character(),  
## .. thr\_jw = col\_character(),  
## .. rural = col\_character(),  
## .. thr\_ms = col\_character(),  
## .. wave = col\_double(),  
## .. religiosity = col\_double()  
## .. )

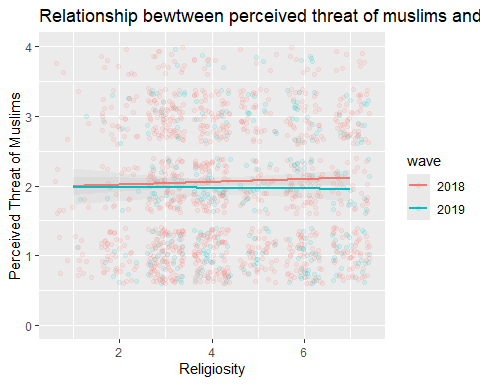
ip <- issp %>%  
 mutate(  
 id = factor(id),  
 thr\_ath = as.factor(thr\_ath),  
 thr\_bd = as.factor(thr\_bd),  
 thr\_ch = as.factor(thr\_ch),  
 thr\_hd = as.factor(thr\_hd),  
 thr\_jw = as.factor(thr\_jw),  
 thr\_ms = as.factor(thr\_ms),  
 neg\_ath = as.factor(neg\_ath),  
 neg\_bd = as.factor(neg\_bd),  
 neg\_ch = as.factor(neg\_ch),  
 neg\_hd = as.factor(neg\_hd),  
 neg\_jw = as.factor(neg\_jw),  
 neg\_ms = as.factor(neg\_ms),  
 wave = as.factor(wave),  
 nzeuro = as.factor(nzeuro),  
 eduyears = as.numeric(eduyears),  
 male = as.factor(male),  
 age = as.numeric(age),  
 rightwing = as.numeric(rightwing),  
 rural = as.factor(rural),  
 religiosity = as.numeric(religiosity)  
 )

library(ggplot2)  
ggplot(data = ip, aes(y = as.numeric(thr\_ms), x = religiosity, colour = wave)) + geom\_jitter(alpha = .1) +   
 geom\_smooth(method = lm, fullrange = FALSE, alpha = 0.1) +  
 scale\_y\_continuous(limits = c(0,4)) +  
 labs(title = "Relationship bewtween perceived threat of muslims and type of religiosity") +   
 xlab("Religiosity") +   
 ylab("Perceived Threat of Muslims")

## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 1245 rows containing non-finite values (stat\_smooth).

## Warning: Removed 1297 rows containing missing values (geom\_point).



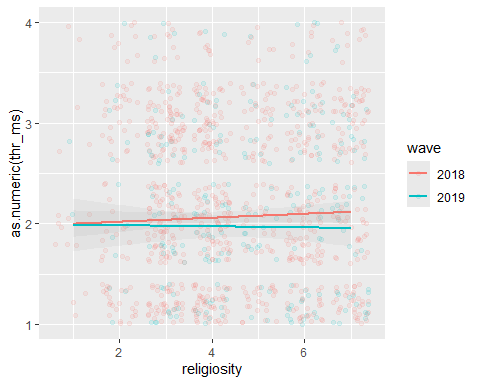
## Note that I have the following graph should start from 1 and run to 4 but currently runs from 0-4. Fix the graph

library(ggplot2)  
ggplot(data = ip, aes(y = as.numeric(thr\_ms), x = religiosity, colour = wave)) + geom\_jitter(alpha = .1) +   
 geom\_smooth(method = lm, fullrange = FALSE, alpha = 0.1) +  
 scale\_y\_continuous(limits = c(1,4))

## `geom\_smooth()` using formula 'y ~ x'

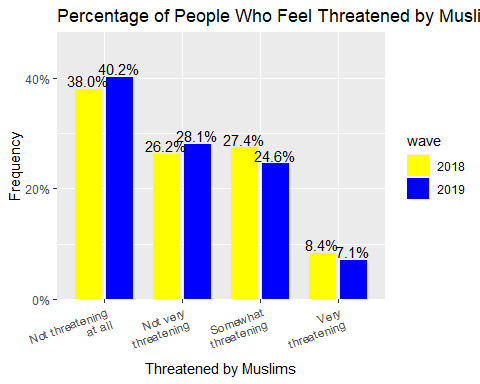
## Warning: Removed 1245 rows containing non-finite values (stat\_smooth).

## Warning: Removed 1576 rows containing missing values (geom\_point).



## Fix the Code & Discuss

library(sjPlot)  
plot\_xtab(  
 ip$thr\_ms,  
 ip$wave,  
 show.total = F,  
 show.n = F,  
 geom.colors = c("yellow", "blue")  
 ) + labs(title = "Percentage of People Who Feel Threatened by Muslims (2018 compared to 2019)") +  
 xlab("Threatened by Muslims") + ylab("Frequency") +  
 #scale\_y\_continuous(limits=c(0,7)) + #theme(plot.title = element\_text(size=09))  
 theme(axis.text.x = element\_text(angle = 20, hjust = 1))

 I’ve added a title to explain the bar chart, and changed the colours so that they were less provocative (green v red can represent good v bad) - I would like to understand how to change the x-axis from a percentage to a frequency (count) just out of interest, I assume this wasn’t done as the sample sizes are different, but if they weren’t it could be useful to know and compare tallies as well as percentages!