Week 3 Workbook

Odelia

3/10/2021

**Question 1: Why is this graph not printing any output?**

This is because “geom\_point” is not in the code. It is important for “geom\_point” to be in the code as it tells R to create the scatterplot. Geom defines the layout of the ggplot layer while point defines the layout to be a scatterplot. If “geom\_point” is missing from the code, R will not know that I want my data to be displayed as a plot and specifically, a scatterplot. For the plot to display successfully, after loading tidyverse, the code should be as such: ggplot(data = mtcars) + geom\_point(mapping = aes(x = mpg, y = wt, colour=factor(cyl)))

**Question 2. Using the mpg dataset, graph the relationship between city mileage and highway mileage by year manufacture.**

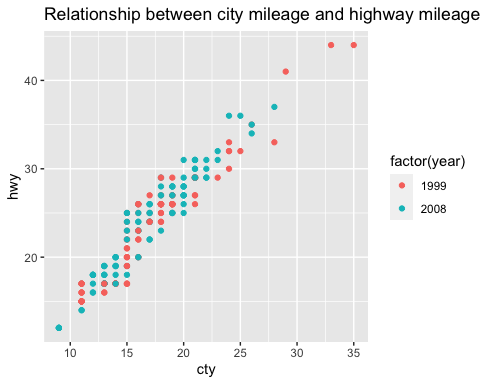
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.3 ✓ purrr 0.3.4  
## ✓ tibble 3.1.0 ✓ dplyr 1.0.5  
## ✓ tidyr 1.1.2 ✓ stringr 1.4.0  
## ✓ readr 1.4.0 ✓ forcats 0.5.1

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

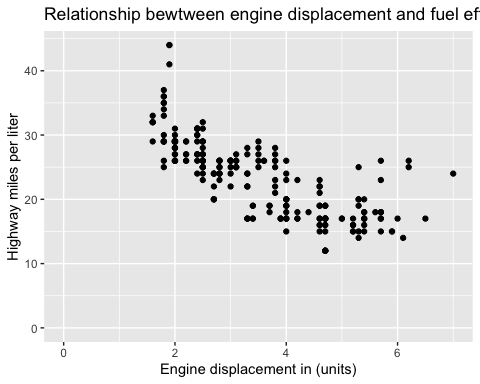
ggplot(data = mpg) + geom\_point(mapping = aes(x = cty, y = hwy, colour=factor(year))) + labs(title = "Relationship between city mileage and highway mileage by year in mpg")



**Question 3. Edit this graph so that the x axis and the y axis both start at 0**

This can be done by adding “expand\_limits(x = 0, y = 0)” to the code.

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



**Question 4: what is one benefit and one limitation for this graph above (in which the x and y values start at 0?)**

Limitation: Real world applications for internal combustion engines will always have a value greater than 0. Likewise for data that can never logically be 0, having both the x and y axes at 0 would unnecessarily make the graph’s data points more compressed together and therefore harder to read than a truncated graph.

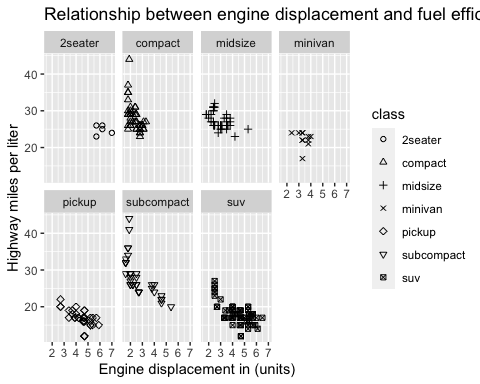
Benefit: On the other hand, truncated diagrams will distort the underlying numbers visually. A diagram that starts at 0 minimises the likelihood of people visually overestimating differences in data points.

**Question 5. Which of these two graphs do you prefer and why?**

I prefer the graph with the data points categorised by colours. Categorising data points by colours makes it easier and more comfortable for the brain to differentiate data points than if they were categorised by shapes. Especially in a dataset with a large number of data points or data points with underlying numbers that are really close or stack on top of one another, having the data points differentiated by colours would be easier to discern one data point from another, and also notice any trends in the data. It is also easier to associate the categories with their relative colours as colours attract the human eye and stimulate the brain more so than shapes. It is also empirically evident in research studies that colours have the tendency to capture better attention level, and thus, better memory than shapes.

**Question 6. add a facet to this graph for the “class” variable**

g2 <-ggplot(data = mpg) +   
 geom\_point(mapping = aes(x = displ, y = hwy, shape = class )) + scale\_shape\_manual(values = c(1:7)) +  
 labs(title = "Relationship between engine displacement and fuel efficiency in the mpg automobile dataset") +   
 xlab("Engine displacement in (units)") +   
 ylab("Highway miles per liter")   
g2 + facet\_wrap(~ class, nrow = 2)



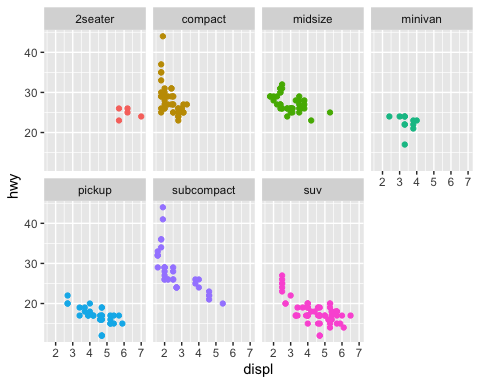
I have to add “scale\_shape\_manual(values = c(1:7))” in order for the 7th class of data points to be plotted as the shape palette can only deal with a maximum of 6 discrete values. Without it, data points for “suv” class will not be plotted or appear on the graph.

**Question 7. which graph is more informative and why?**

My opinion is that graph b is more informative. The use of facet for this data allows for a clear visualisation of the relationship between engine displacement and fuel efficiency within the individual classes of cars. It also clearly shows us the differences within each classes and also inter-class differences as well. Without facets, a single scatterplot (graph a) will be overly dense with information that overlaps one another. Graph a will allow us to plot best fit lines that tells us that the higher the engine displacement, the lower the fuel efficiency. It will still be visible to see the differences between the classes due to the colour differentiation but it will not be able to tell us information of each classes and the differences within. Hence, graph b is more informative.

**Question 8. remove the legend from the facet graph above (g4)** clea I can remove the legend from the facet graph above by adding “theme(legend.position =”none“)” to the code.

g4 <- ggplot(data = mpg) +  
 geom\_point(mapping = aes(x = displ, y = hwy, color = class)) +  
 facet\_wrap( ~ class, nrow = 2)  
g4 + theme(legend.position = "none")



**Question 9 Convert the y variable to “numeric” and graph the relationship betweeen religiosity (x-axis) and `thr\_ms`` (y-axis) in the ISSP dataset. Create new axis labels**

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 × 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()  
## .. )

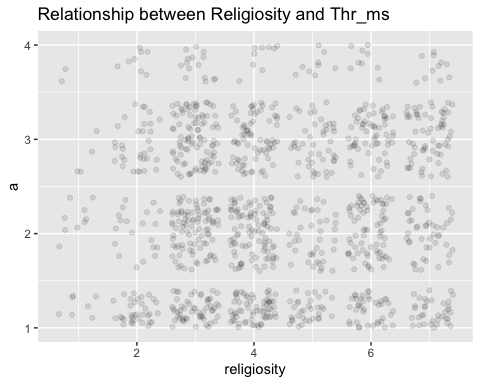
ip <- issp %>%  
 mutate(  
 id = factor(id),  
 thr\_ath = as.factor(thr\_ath),  
 thr\_bd = as.factor(thr\_bd),  
 thr\_ch = as.factor(thr\_ch),  
 thr\_hd = as.factor(thr\_hd),  
 thr\_jw = as.factor(thr\_jw),  
 thr\_ms = as.factor(thr\_ms),  
 neg\_ath = as.factor(neg\_ath),  
 neg\_bd = as.factor(neg\_bd),  
 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)  
 )  
a <- as.numeric(as.factor(ip$thr\_ms))

To convert “thr\_ms” to numeric, I added the code as such: a <- as.numeric(as.factor(ip$thr\_ms))

I then plot the relationship between religiosity and ‘thr\_ms’ as such:

ggplot(data = ip, aes(y = a, x = religiosity)) + geom\_jitter(alpha = .1) + scale\_y\_continuous(limits = c(1,4)) + labs(title = "Relationship between Religiosity and Thr\_ms")

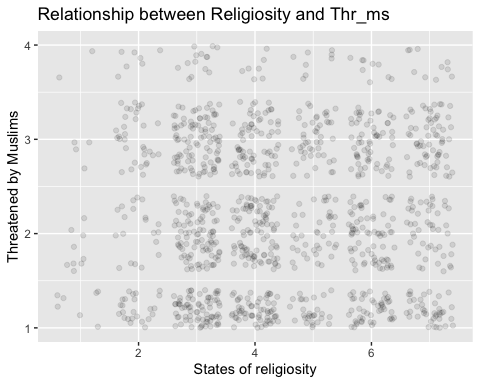
## Warning: Removed 1573 rows containing missing values (geom\_point).



To create new axis labels, I added in the following codes: xlab(“States of religiosity”) and ylab(“Threatened by Muslims”)

ggplot(data = ip, aes(y = a, x = religiosity)) + geom\_jitter(alpha = .1) + scale\_y\_continuous(limits = c(1,4)) + labs(title = "Relationship between Religiosity and Thr\_ms") + xlab("States of religiosity") + ylab("Threatened by Muslims")

## Warning: Removed 1566 rows containing missing values (geom\_point).



**Question 10. Note that I have the following graph should start from 1 and run to 4 but currently runs from 0-4. Fix the graph**

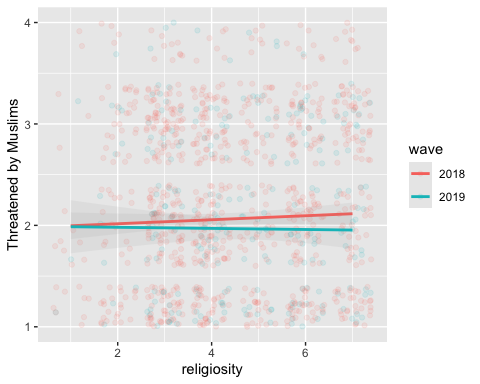
The graph can be fixed by changing the limit for the y-axis from 0 to 1: scale\_y\_continuous(limits = c(1,4))

ggplot(data = ip, aes(y = a, x = religiosity, colour = wave)) + geom\_jitter(alpha = .1) +   
 geom\_smooth(method = lm, fullrange = FALSE, alpha = 0.1) +  
 scale\_y\_continuous(limits = c(1,4)) + ylab("Threatened by Muslims")

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

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

## Warning: Removed 1586 rows containing missing values (geom\_point).



**Extra Question 11. Find one way of improving the following code and explain your answer**

I can improve the code by adding “show.values = F” and changing the x-axis text element angle from ‘20’ to ‘0’. By doing so, I am improving the code by making it an APA-formatted document suitable for reporting in scientific research. In order for a figure/graph to be suitable for reporting in an APA document, values should not be included in the graph. Values are only reported in the results writeup. I have done so by adding “show.value = F” into the code. The x-axis labels should also be parallel to the axis instead of being in an angle. Therefore, the need to change the angle to ‘0’. The result graph would be as such:

library(sjPlot)   
plot\_xtab(  
 ip$thr\_ms,  
 ip$wave,  
 show.total = F,  
 show.n = F,   
 show.values = F,  
 geom.colors = c("lightgreen", "darkred")  
 ) +  
 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 = 0))

