## Choosing a Paper

Paper A  
Paper B

Issues with A, although it seemed easy it was complicated because the source data was simulated and also the topic was too complicated to be replicated.  
We chose Paper B mostly because it was a well documented Paper with own data

## Talking about the chosen Paper

References, describir the things to be replicated

* Explain the data, the issues, how we understood
* what issues we encounter?

## Issues when coding how to correctly replicated

ie. it was difficult to do the sizing, scaling, colors, also to determine which chart we were replicating

* issues when publishing, data size, corrupt folder, issues private vs. public.
* wrong path folder, had to find what was the original path folder because although we had everything, we weren’t able to publish
* reseting and changing the project path // the Git Repo to my actual working folder.
* stucked when coding (Ctrl+C) unstuck it

## Actual replication

talk about the solutions and the issues, one project for the presentation, another for coding.

library(haven)  
library(dplyr)

Adjuntando el paquete: 'dplyr'

The following objects are masked from 'package:stats':  
  
 filter, lag

The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

library(ggplot2)  
library(janitor)

Adjuntando el paquete: 'janitor'

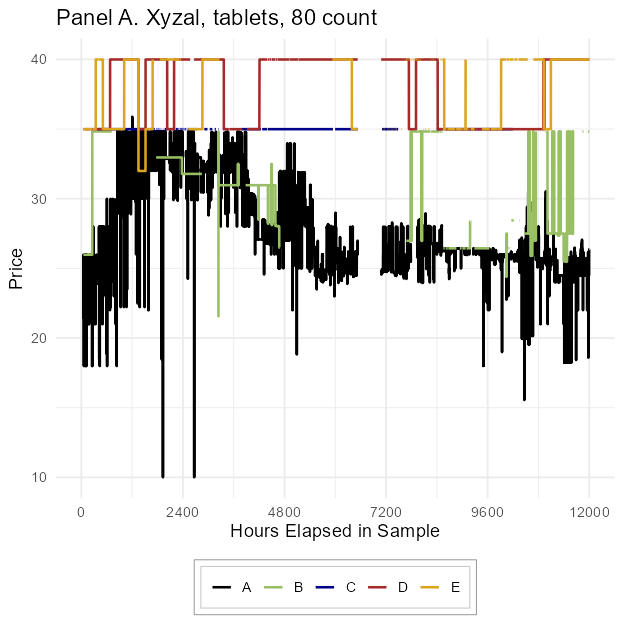
The following objects are masked from 'package:stats':  
  
 chisq.test, fisher.test

# Load and clean  
data <- read\_dta("C:/Users/Christian Casas/OneDrive - studhsf/Documents/Masters - HS Fresenius/2nd Semester/Data Science for Business/analysis/data/analysis\_data.dta") %>%  
 janitor::clean\_names()  
  
# Filter for Xyzal 80ct Tablet (non-multipack)  
data\_filtered <- data %>%  
 filter(  
 brand == "Xyzal",  
 form == "Tablet",  
 size == 80,  
 multipack == 1,  
 flag\_imputed\_price != 1  
 ) %>%  
 distinct(website, period\_id, .keep\_all = TRUE)  
  
# Get upper Y-axis limit for clean breaks, issue with the scaling since we had to stop at 12000, so it was force to stop al 12000  
x\_max <- max(ceiling(max(data\_filtered$price, na.rm = TRUE) / 200) \* 200, 12000)  
  
# Plot  
ggplot(data\_filtered, aes(x = period\_id, y = price, color = website)) +  
 geom\_line(size = 0.9) +  
 scale\_color\_manual(values = c(  
 "A" = "black",  
 "B" = "#98bf64",  
 "C" = "darkblue",  
 "D" = "brown",  
 "E" = "#DAA520"  
 )) +  
 #scale  
 scale\_x\_continuous(  
 limits = c(0, x\_max),  
 breaks = seq(0, x\_max, by = 2400)  
 ) +  
 labs(  
 title = "Panel A. Xyzal, tablets, 80 count",  
 x = "Hours Elapsed in Sample",  
 y = "Price",  
 color = "Retailer"  
 ) +  
 theme\_minimal(base\_size = 14) +  
 theme(  
 legend.position = "bottom",  
 legend.title = element\_blank(),  
 legend.background = element\_rect(fill = "white", color = "gray80", size = 0.4),  
 legend.box.background = element\_rect(color = "gray60"),  
 legend.box.margin = margin(t = 5, r = 5, b = 5, l = 5)  
 )

Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
ℹ Please use `linewidth` instead.

Warning: The `size` argument of `element\_rect()` is deprecated as of ggplot2 3.4.0.  
ℹ Please use the `linewidth` argument instead.

Warning: Removed 5714 rows containing missing values or values outside the scale range  
(`geom\_line()`).



## Replication

Here compare the paper’s graph vs. ours

Warning: Removed 5714 rows containing missing values or values outside the scale range  
(`geom\_line()`).

