MSIN0094 Third Assignment

Due Friday 10am, Dec 5, 2025

Candidate number: RWFD2

Word count: 1933

1. Descriptive Analytics (20 pts)

Q1 From `data_full`, generate a new variable, `final_price`, which is the actual retail price for each week (i.e., Recommended Retail Price after discounts). (**8pts**)

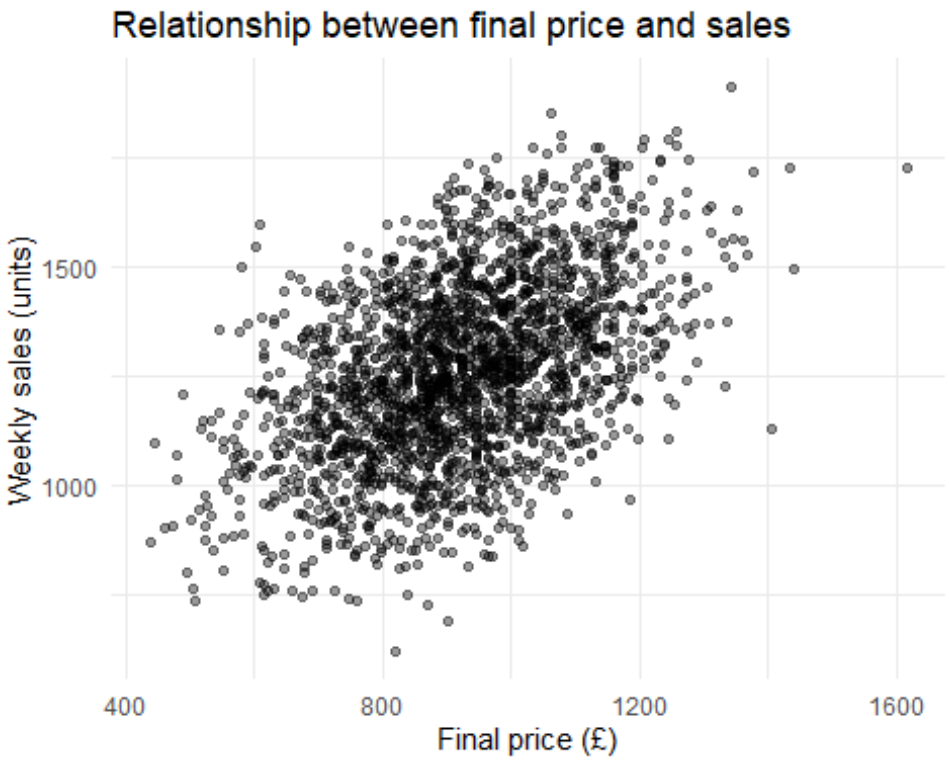
- Write your code below to generate final_price from RRP and discount. (**2pts**)

```
# write your codes below
# The discount represents the percentage reduction from the original price, so t
he customer pays only the remaining proportion of the RRP
data_full <- data_full %>%
  mutate(final_price = RRP * (1 - discount))
data_full %>%
    slice_head(n = 5)
```

```
  product_id   brand technology resolution support_HDR screensize week_id sales
1          1 Samsung       OLED      1080p           0      50-59       1  1745
2          2 Samsung       OLED         4k           1      30-39       1  1444
3          3 Samsung       OLED         4k           0      30-39       1  1183
4          4 Samsung       QLED      1080p           1        60+       1  1474
5          5 Samsung       OLED         4k           1      50-59       1  1647
   RRP discount marketing_expense cost_shifter final_price
1 1399     0.12          3794.739     618.6324     1231.12
2 1099     0.12          3794.739     571.6035      967.12
3 1049     0.18          3794.739     644.6133      860.18
4  949     0.13          3794.739     593.3050      825.63
5 1299     0.17          3794.739     646.6439     1078.17
```

- Visualise using scatter plot the relationship between final price and sales. Tips: you can use `ggplot2` and `geom_point` to create the scatter plot. Write your code below to create the scatter plot. (**2pts**)

```
# write your codes below
library(ggplot2)
ggplot(data_full, aes(x = final_price, y = sales)) +
  geom_point(alpha = 0.4) +
  labs(
    x = "Final price (£)",
    y = "Weekly sales (units)",
    title = "Relationship between final price and sales"
  ) +
  theme_minimal()
```



- Do you observe a positive or negative relationship between final price and sales? Is this relationship causal? Why or why not? (**4pts**, 150 words)

The scatter plot shows a positive relationship between final price and sales: more expensive TVs tend to sell more units in this dataset. This is the opposite of what a simple demand curve would suggest, where higher prices usually reduce quantity demanded. However, this pattern is not causal.

The relationship is likely driven by endogeneity. First, omitted variables such as product quality, screen size, technology (OLED vs LCD), and brand positioning can jointly affect both price and sales, creating a spurious positive correlation even though a price increase for the same TV would reduce demand. Second, reverse causality may be present: Amazon may charge higher prices in weeks when stronger demand is expected. Because of these confounding factors, we cannot conclude that higher prices cause higher sales. The scatter plot reflects correlation, not the true causal price effect.

Q2. Use `dplyr` to compute the average weekly dollar sales (final price * unit sales) for each brand across all weeks (i.e., the result should be 1 average per brand). Rank the brands from the highest average dollar sales to the lowest average dollar sales. (**6pts**) Which brand has the highest average weekly dollar sales? (**2pts**).

```
# write you code below
data_sales_by_brand <- data_full %>%
  mutate(dollar_sales = final_price * sales) %>%  # Creating weekly revenue vari
able
  group_by(brand) %>% # Group data by brand to compute brand-level averages
  summarise(avg_weekly_dollar_sales = mean(dollar_sales, na.rm = TRUE)) %>% # Ca
lculate the average weekly dollar sales for each brand
```

```
  ungroup() %>%
  arrange(desc(avg_weekly_dollar_sales)) # Rank brands

# please do not modify.
# print out the ranking of brands based on average weekly dollar sales
data_sales_by_brand
```

```
# A tibble: 4 × 2
  brand   avg_weekly_dollar_sales
  <chr>                     <dbl>
1 Samsung                1277231.
2 Sony                   1226644.
3 LG                     1119492.
4 Philips                1086011.
```

```
# Based on the computed rankings of average weekly dollar sales, Samsung has the highest average weekly revenue (≈ 1.28 million dollars per week), followed by Sony, LG, and Philips. This indicates that Samsung consistently generates the largest weekly sales value on Amazon among the four brands in the dataset. The result suggests that Samsung either sells higher-priced models, larger volumes, or both, leading to stronger overall weekly dollar performance relative to its competitors.
```

> ***Q3.*** In Marketing, we refer to brand equity as the additional sales a brand can obtain when everything else is equal, i.e., the causal effect of brands on sales. Does the above average sales ranking causally identify which brand has the highest brand equity? Why or why not? (**4pts**; 150 words)

The ranking of average weekly dollar sales does not causally identify which brand has the highest brand equity. The comparison reflects correlations, not the causal effect of brand names holding all other factors constant. Each brand sells a different mix of technologies (OLED vs LCD), screen sizes, resolutions, and price segments. These product characteristics strongly influence demand and are not controlled for in the simple ranking. As a result, higher sales may reflect a brand offering more premium models, more discounted products, or simply having broader product availability, rather than stronger brand equity.

Additionally, endogeneity is present. Brands may strategically adjust their prices, discounts, or product assortments based on expected demand, and Amazon's algorithmic pricing may respond to sales trends. Because these factors jointly determine both prices and sales, we cannot isolate the incremental sales attributable purely to the brand name. Therefore, the ranking cannot be interpreted as a causal measure of brand equity.

2. Marketing Mix Modeling and Endogeneity (28pts)

> ***Q4.*** Run a Marketing Mix Modeling linear regression as follows (**6pts**):

- Run the linear regression below using `fixest` package (Equation 1 hereinafter) (**2pts**).

```
# write you codes for the regression below
library(fixest)

# Run the linear regression: sales = a + b*final_price + c*marketing_expense + e
```

```
# feols() estimates an OLS model from the fixest package
ols_1 <- feols(
  sales ~ final_price + marketing_expense,
  data = data_full
)

# do not modify the code below; this is to print out the results

modelsummary(ols_1,
             stars = T,
             gof_map = c('nobs','r.squared'))
```

(Intercept)	421.271***
	(20.126)
final_price	0.618***
	(0.020)
marketing_expense	0.094***
	(0.003)
Num.Obs.	2080
R2	0.503

- p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

- Interpret the coefficients of `final_price`, including coefficients and statistical significance (**4pts**).

```
# The coefficient on final_price is 0.618 and statistically significant at the 1
% level. This means that, in this model, a £1 increase in the final retail price
is associated with an increase of about 0.62 additional units sold per week, hol
ding marketing expenditure constant. This is a positive price effect, which goes
against the usual expectation that higher prices reduce sales. The result sugges
ts that more expensive TV models in the dataset tend to sell more units, likely
because they also differ in important characteristics such as quality, screen si
ze, or technology.
# However, this estimate should not be interpreted as a causal effect. Price is
likely endogenous: Amazon adjusts prices based on expected demand, stock levels,
or competitive conditions. These factors are not controlled for in this simple r
egression. Therefore, the coefficient captures correlation rather than the true
causal impact of price on sales.
```

Q5. Based on the regression coefficients reported above, discuss the endogeneity issues with `final_price` in Equation 1. For each endogeneity cause, explain the general definitions and then give concrete examples in Amazon's context. (**12pts**)

- General definition of each endogeneity cause (**6pts**, 200 words)

Endogeneity occurs when an explanatory variable is correlated with the error term, leading to biased and inconsistent regression estimates. One key source is omitted variable bias, which

arises when an important factor that affects both the explanatory variable and the outcome is left out of the model. Because this missing factor influences both sides, the regression incorrectly attributes its effect to the included variable.

A second source is reverse causality, where the causal direction runs both ways: the explanatory variable affects the outcome, but the outcome also affects the explanatory variable. In this case, the estimated coefficient mixes both effects and cannot be interpreted causally.

A third cause is measurement error, which appears when a variable is measured with noise or imperfect accuracy. If this error is correlated with the unobserved true value, estimates become biased.

Finally, simultaneity occurs when the explanatory and dependent variables are determined at the same time by the same underlying process, making them jointly dependent.

All these situations break the independence assumption of OLS, preventing causal interpretation.

- In Amazon's context, concrete examples to illustrate each endogeneity cause (**6pts**, 200 words)

In Amazon's context, omitted variable bias arises because TVs differ in technology (OLED vs LCD), screen size, quality, brand reputation, and product age. These characteristics affect both final_price and sales. For example, premium OLED models are more expensive and also tend to sell well because of higher quality, so the regression incorrectly attributes this quality effect to price.

Reverse causality occurs when Amazon changes prices in response to expected demand. For instance, if Amazon's algorithms predict high demand for a model during a holiday week, it may increase the price. In this case, higher sales don't occur because of higher price rather, higher expected demand caused Amazon to charge a higher price.

Measurement error may arise because the reported "final_price" is based on dynamic pricing that changes within a week. If the weekly dataset records an average or approximate price, this may not match the actual prices customers paid, introducing noise that biases the estimated price effect.

Simultaneity occurs because Amazon's pricing and sales happen together in real time. Pricing algorithms update continuously based on observed purchasing behavior. When sales accelerate, the system may instantly adjust prices upward or downward. Since sales influence price at the same time price influences sales, both variables are jointly determined, creating simultaneity bias.

> ***Q6.*** If the discount each week in our dataset is randomized by Amazon each week, will Equation 1 give the causal effect of price on sales? Give your reasoning. (**6pts**; 200 words)

If Amazon truly randomised the discount each week, this would make weekly price changes more exogenous and reduce reverse causality, since prices would no longer respond to

expected demand or competitor behavior. However, this does **not** mean Equation 1 would identify the causal effect of price on sales. Final price is calculated as `RRP × (1 – discount)`, and only the discount component is randomized. RRP still reflects product characteristics such as quality, screen size, technology and brand strength—factors that strongly affect sales. Because these characteristics remain unobserved in the regression, final_price is still correlated with omitted variables, creating omitted variable bias.

Randomising discounts creates useful within-product variation, but Equation 1 mixes this with cross-sectional differences between inexpensive basic TVs and expensive premium models. To obtain a clean causal estimate, we would need a model that isolates the random price variation within the same product (for example, using product fixed effects) rather than relying on a simple OLS specification.

> ***Q7.*** From the below regression designed by another data scientist, discuss whether customers always prefer larger screens (i.e., everything else being equal, a larger screen always leads to higher sales)? (**4pts**; 150 words)

The regression suggests that larger screens generally sell more units, but it does not mean customers always prefer bigger screens. The coefficients for 40–49, 50–59, and 60+ inches are all positive and significant, showing higher sales relative to the baseline (29–39 inches) after controlling for brand, technology, resolution, HDR, price, and marketing. However, the pattern is not monotonic: the 60+ category (≈81.6) has a smaller effect than 50–59 inches (≈143.8). If customers always preferred larger screens, the coefficients would steadily increase with size, which is not the case.

This indicates that very large TVs appeal to narrower segments due to higher prices, limited living space, or installation constraints. Thus, while larger screens often perform better than smaller ones, the regression does not support the claim that “the larger the screen, the higher the sales” in all cases.

3. Instrumental Variables (20pts)

> ***Q8.*** One way to obtain causal effects of price on sales from secondary data is to use the instrumental variable method. (**12pts**)

- List two variables you you would collect as instrumental variables for `final_price`

Shipping cost shocks

```
# Changes in fuel prices, logistics disruptions, or adjustments in carrier fees
directly influence Amazon’s cost structure and therefore its pricing decisions.
When shipping costs increase, Amazon often adjusts retail prices upward to maint
ain margins.
# Relevance: Shipping costs affect the retailer’s cost of delivering products an
d therefore influence weekly pricing for TVs.
# Exclusion restriction: Short-run fluctuations in shipping costs do not directl
y influence consumer demand for a specific TV model. Customers are typically una
ware of these cost changes, and logistics shocks do not change product appeal, f
eatures, or brand perception. Thus, their only channel of influence on sales is
through price, making them a valid instrument.
```

Supplier-side cost changes

```
# Fluctuations in component prices (LCD panels, semiconductors). These shift the retailer's pricing but are unrelated to short-term consumer purchasing behaviour
# TV production relies on volatile input components such as LCD panels, semiconductors, backlights, and chips. Weekly or monthly fluctuations in these component costs alter wholesale prices charged to Amazon.
#Relevance: When supplier costs rise, wholesale prices increase, and Amazon adjusts final retail prices accordingly.
# Exclusion restriction: Consumers do not respond directly to upstream component price fluctuations. Higher panel or chip costs do not change a TV's perceived quality, screen size, or marketing appeal. They only influence demand indirectly via the resulting price adjustment.
```

- Can one use the VAT tax rate of TV products as an instrument variable for `final_price`? (**4pts**; 100 words)

Generally no: VAT directly affects the final price paid by consumers, and consumers react to this tax-inclusive price when deciding whether to buy a TV. Because VAT influences sales through the exact same channel as price, it cannot satisfy the exclusion restriction required for a valid instrument. Any change in VAT mechanically changes demand, making it impossible to separate the effect of price from the effect of the tax itself. Therefore, VAT is not a valid instrument for *final_price*.

Q9. Assume you have identified one instrument variable `cost_shifter` in `data_full`. In the code blocks below, write down the two regressions you would need to run in order to estimate the causal effects of `final_price` on `sales`, including `marketing_expense` as the only control variable (**8pts**)

- Correct first stage codes and explanation of the code (**3pts**)
- Correct second stage codes and explanation of the codes (**3pts**)

```
# show the estimation code below and describe the steps

# Stage 1: write the first-stage regression

ols_stage1 <- feols(
  # The dependent variable is final_price - the endogenous variable
  final_price ~ cost_shifter + marketing_expense,
  # Cost_shifter is the instrumental variable, marketing_expense is a control variable
  data = data_full
)

# Adding predicted price to a data_full

data_full <- data_full %>%
  mutate(predicted_price = predict(ols_stage1, newdata = data_full))

# Stage 2: write the second-stage regression
```

```r
ols_stage2 <- feols(
  # Replace final_price with the predicted, instrument-based price
  sales ~ predicted_price + marketing_expense,
  data = data_full
)

# do not modify the code below; this is to print out the results

modelsummary(list(
  "First Stage" = ols_stage1,
  "Second Stage" = ols_stage2
),
stars = T, gof_map = c('nobs','r.squared'))
```

	First Stage	Second Stage
(Intercept)	256.793***	1665.622***
	(40.790)	(63.067)
cost_shifter	1.099***	
	(0.066)	
marketing_expense	0.000	0.093***
	(0.003)	(0.003)
predicted_price		-0.734***
		(0.068)
Num.Obs.	2080	2080
R2	0.118	0.306

- p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001
- Based on the results of the two regressions, discuss the causal effect of `final_price` on `sales` (**2pts**)

```r
# In the first stage, the coefficient on cost_shifter is 1.099 and highly signif
icant, confirming that the instrument strongly predicts final_price and satisfie
s the relevance condition. Marketing_expense has no meaningful effect on price,
as expected. This first stage extracts the exogenous component of price movement
s.

# The second-stage coefficient on predicted_price is −0.734, indicating a clear
causal negative effect of price on sales6 significant at the 1% level. This mean
s that a £1 exogenous increase in price causally reduces weekly sales by about 0
.73 units, holding marketing constant. Marketing_expense remains positive and si
gnificant, showing that higher spending increases sales. Because predicted_price
reflects only exogenous price variation from the cost_shifter, it isolates how s
ales respond when price changes for reasons unrelated to demand.

# This result is much more negative than the earlier positive OLS estimate, whic
h was biased upward because premium, high-quality TVs are both expensive and sel
l more. The IV approach removes this bias and reveals the true causal effect.
```

Q10. Design the A/B/N testing (**20 pts**)

1. The natural unit of randomisation is the individual customer account. Try-on is an on-site, personalised feature, so the main decision is whether each customer sees: 1) no new feature (control), 2) Treatment A: real photo try-on, or 3) Treatment B: cartoon avatar try-on. User-level randomisation maximises sample size, balances observed and unobserved traits, and matches how decisions are made (each user chooses what to buy).

However, there are risks of spillover and crossover. Crossover may occur if the same person uses multiple accounts or devices and accidentally receives different treatments. This can be mitigated by forcing login and tying treatment to the Amazon account ID across all devices.

Spillovers can happen if customers show the feature to friends or family or talk about it, which might influence how others behave even if they were assigned to a different group. However, because the try-on feature is small and built into the normal interface, these spillover effects are likely limited. Therefore, user-level randomisation is still the best balance between practicality and statistical power.

2. In this A/B/N experiment, we divide users into three groups: Control, A, and B. Since the expected effects of the two treatments are likely similar, an equal allocation of one-third per group provides the strongest statistical power. With 100,000 users, we randomly select the required number of users for A and B using a reproducible random seed, and assign all remaining users to Control.

Treatment assignment is fixed for each user during the experiment to avoid contamination and to ensure clean measurement of treatment effects.

```
set.seed(888)

# Assume we have a dataset with 100,000 users
# data_users contains at least one identifying variable (user_id)

N <- nrow(data_users)

Error: объект 'data_users' не найден

# Defining allocation probabilities for A/B/N
p_A <- 1/3
p_B <- 1/3
p_control <- 1/3

# 1. Randomly sample users for Treatment A
A_index <- sample(1:N,
                  size = round(N * p_A),
                  replace = FALSE)

Error in 1:N: NA/NaN-аргумент

# 2. From the remaining users, randomly sample for Treatment B
remaining_after_A <- setdiff(1:N, A_index)
```

```
Error in 1:N: NA/NaN-аргумент
```

```
B_index <- sample(remaining_after_A,
                  size = round(N * p_B),
                  replace = FALSE)
```

```
Error: объект 'remaining_after_A' не найден
```

```
# 3. Remaining users go to Control
control_index <- setdiff(remaining_after_A, B_index)
```

```
Error: объект 'remaining_after_A' не найден
```

```
# 4. Assigning treatment labels
data_users <- data_users %>%
  mutate(
    treatment = case_when(
      row_number() %in% A_index ~ "A", # Real-photo try-on
      row_number() %in% B_index ~ "B", # Cartoon avatar try-on
      row_number() %in% control_index ~ "control" # No new feature
    )
  )
```

```
Error: объект 'data_users' не найден
```

3. We must calculate the minimum sample size required to detect the expected treatment effects: a £10 increase in spending for Treatment A and a £5 increase for Treatment B, given a standard deviation of £100. Using a power analysis with 80% power and a 5% significance level, we estimate the required sample size for each comparison against the control group.

```
# sample size for A vs control
ss_A <- power.t.test(
  delta     = 10,
  sd        = 100,
  sig.level = 0.05,
  power     = 0.8,
  type      = "two.sample"
)

# sample size for B vs control
ss_B <- power.t.test(
  delta     = 5,
  sd        = 100,
  sig.level = 0.05,
  power     = 0.8,
  type      = "two.sample"
)

ss_A$n; ss_B$n   # Required users per arm
```

```
[1] 1570.737
```

```
[1] 6280.064
```

The power calculations indicate that detecting the £10 lift in Treatment A requires about 1,571 users, while detecting the smaller £5 lift in Treatment B requires roughly 6,280 users. Therefore, the experiment should be based on the larger sample size to ensure adequate power for both effects.

4. We need to collect two main categories of data.

1) Randomization-check data: customer and session characteristics used to verify balance across groups, such as account ID, country, device type, Prime status, historical spending, and baseline engagement (page visits, click-throughs, add-to-cart rate).
2) Outcome data: variables required to measure treatment effects, including feature exposure, number of try-on sessions, time spent using the feature, conversion rate, order value, returns, and follow-up engagement (wishlists, recommendation clicks).

All records must include the assigned treatment label, timestamps, and whether the feature actually rendered, allowing us to assess compliance and exclude users who never received their assigned experience.

5. After collecting the data, we first perform randomization checks to verify that baseline characteristics are balanced across Control, A, and B. If any differences arise, regression controls can be accounted.

```
# Check baseline balance: Control vs A
#t.test(baseline_spend ~ treatment,
#       data = data_amazon %>%
#         filter(treatment %in% c("control", "A")))
#
# Check baseline balance: Control vs B
#t.test(baseline_spend ~ treatment,
#      data = data_amazon %>%
#         filter(treatment %in% c("control", "B")))
```

Next, we compare outcomes such as order value and conversion rate across groups.

```
# Compute average order value for each group
#data_amazon %>%
#  group_by(treatment) %>%
#  summarise(avg_order_value = mean(order_value))
```

Then test statistical significance:

```
#t.test(order_value ~ treatment,
#       data = data_amazon %>% filter(treatment %in% c("control","A")))
#t.test(order_value ~ treatment,
#       data = data_amazon %>% filter(treatment %in% c("control","B")))
```

Last step is to estimate average treatment effects using an A/B/N regression with control as the reference group:

```
#data_amazon <- data_amazon %>%
#  mutate(treat_factor = relevel(as.factor(treatment), ref = "control"))
```

```
#model <- feols(order_value ~ treat_factor, data = data_amazon)
#modelsummary(model, stars = TRUE)
```

Coefficients on A and B give the average causal lifts versus control; we compare magnitudes, confidence intervals, and implementation costs to choose the better variant.

> ***Q11.*** Finally, Tom would like to study the causal effect of Amazon rating on product sales. For instance, what is the causal effect of a 4.5-star rating on sales compared to a 4-star rating. Propose **one** natural experiment method to study this causal question. (**12pts**)

1) I would use a Regression Discontinuity Design (RDD) based on Amazon's rating rounding rule. Suppose the displayed stars are rounded to the nearest 0.5. When a product's underlying average rating moves from 4.24 to 4.26, the true quality changes very little, but the displayed rating jumps from 4.0 to 4.5 stars. Products just below and just above the 4.25 cutoff should be almost identical in quality, brand, and demand trends. The only sharp difference is the number of visible stars. Comparing sales for products on each side of this threshold gives a quasi-experimental estimate of the causal effect of a 4.5-star vs 4-star rating.

2) To implement the RDD, we need product–week panel data.

Running variable: Underlying average numerical rating (e.g. 4.23, 4.27) before rounding, distance to the cutoff 4.25.

Treatment variable: Indicator HighRating_it that equals 1 when the displayed rating is 4.5 stars (rating ≥ 4.25) and 0 when it is 4.0 stars (rating < 4.25).

Outcome variable: Weekly unit sales (revenue can be used as well) for each product.

Controls: Price, discounts, marketing expense, category dummies, stock availability, plus product ID and time variables to build fixed effects.

3) After collecting the data, we restrict the sample to products whose numerical ratings lie within a narrow window around the 4.25 cutoff. We then construct a treatment indicator equal to 1 when a product displays 4.5 stars and 0 when it displays 4.0 stars. Because products on either side of this threshold are almost identical in underlying quality, a discontinuous jump in sales can be interpreted causally.

Estimating the causal effect using the following regression:

$$Sales_{it} = \alpha + \beta HighRating_{it} + \gamma_1 price_{it} + \gamma_2 discount_{it} + \gamma_3 marketing_{it} + \mu_i + \lambda_t + \varepsilon_{it}.$$

μ_i are product fixed effects, λ_t are time fixed effects. The coefficient β measures how much sales increase when a product displays 4.5 stars instead of 4 stars. Then checking robustness by trying different bandwidths around the cutoff and alternative functional forms for the running variable.