Quantium Virtual Internship - Retail Strategy and Analytics - Task 2

Your Name Here

November 09, 2025

Contents

1 1. Introduction

This report details the analysis of a store trial, as requested by the Category Manager, Julia. The objective is to evaluate the performance of trial stores (77, 86, and 88) to provide a data-driven recommendation.

The analysis is broken into three parts: 1. **Control Store Selection:** Identifying suitable control stores for each trial store based on pre-trial performance. 2. **Trial Assessment:** Comparing the performance of each trial store against its control store during the trial period. 3. **Findings & Recommendation:** Collating the results and providing a final recommendation.

2 2. Data Preparation

First, we load the `QVI_data.csv` and aggregate it to create monthly metrics for all stores. This will form the basis of our comparisons.

```r
# --- 2.1: Load the Data ---
full_data <- read_csv("QVI_data.csv") %>%
  mutate(
    # Ensure DATE is in Date format
    DATE = as.Date(DATE)
  )

# --- 2.2: Aggregate Data to Monthly Metrics ---
monthly_metrics <- full_data %>%
  mutate(
    YEAR_MONTH = floor_date(DATE, "month")
  ) %>%
  group_by(STORE_NBR, YEAR_MONTH) %>%
  summarise(
    # 1. Total sales revenue
    TOT_SALES = sum(TOT_SALES),

    # 2. Total number of customers
    N_CUSTOMERS = n_distinct(LYLTY_CARD_NBR),

    # 3. Average number of transactions per customer
    AVG_TXN_PER_CUSTOMER = n() / n_distinct(LYLTY_CARD_NBR),

    .groups = "drop" # Drop the grouping
  )

glimpse(monthly_metrics)
```

```
## Rows: 3,169
## Columns: 5
## $ STORE_NBR            <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2~
## $ YEAR_MONTH           <date> 2018-07-01, 2018-08-01, 2018-09-01, 2018-10-01, ~
## $ TOT_SALES            <dbl> 206.90, 176.10, 278.80, 188.10, 192.60, 189.60, 1~
## $ N_CUSTOMERS          <int> 49, 42, 59, 44, 46, 42, 35, 52, 45, 42, 46, 42, 3~
## $ AVG_TXN_PER_CUSTOMER <dbl> 1.061224, 1.023810, 1.050847, 1.022727, 1.021739,~
```

3 3. Select Control Stores

To measure the trial's impact, we first need to find control stores that were behaving almost identically to our trial stores *before* the trial began. We define the pre-trial period as all data before February 2019.

```r
# Define the pre-trial period (Jul 2018 - Jan 2019)
pre_trial_end_date <- as.Date("2019-02-01")

# Create a table of pre-trial metrics for all stores
pre_trial_metrics <- monthly_metrics %>%
  filter(YEAR_MONTH < pre_trial_end_date)
```

3.1 3.1. Control Store Selection Function

As requested, we create a function to calculate a “similarity score” (based on sales correlation and magnitude) to find the best match for any given trial store.

```
find_control_store <- function(trial_store_id, pre_trial_data) {

  trial_store_data <- pre_trial_data %>%
    filter(STORE_NBR == trial_store_id)

  candidate_stores_data <- pre_trial_data %>%
    filter(STORE_NBR != trial_store_id)

  # Calculate similarity scores
  similarity_scores <- candidate_stores_data %>%
    left_join(trial_store_data, by = "YEAR_MONTH", suffix = c("_candidate", "_trial")) %>%
    group_by(STORE_NBR_candidate) %>%
    summarise(
      CORR_SALES = cor(TOT_SALES_candidate, TOT_SALES_trial, use = "pairwise.complete.obs"),
      DIST_SALES = sum(abs(TOT_SALES_candidate - TOT_SALES_trial)),
      .groups = "drop"
    ) %>%
    mutate(
      # Scale and combine scores
      SCALED_DIST = (DIST_SALES - min(DIST_SALES)) / (max(DIST_SALES) - min(DIST_SALES)),
      SCALED_CORR = (CORR_SALES - min(CORR_SALES)) / (max(CORR_SALES) - min(CORR_SALES)),
      SIMILARITY_SCORE = 0.6 * SCALED_CORR + 0.4 * (1 - SCALED_DIST)
    ) %>%
    arrange(desc(SIMILARITY_SCORE))

  return(similarity_scores)
}
```

3.2 3.2. Trial Store 77

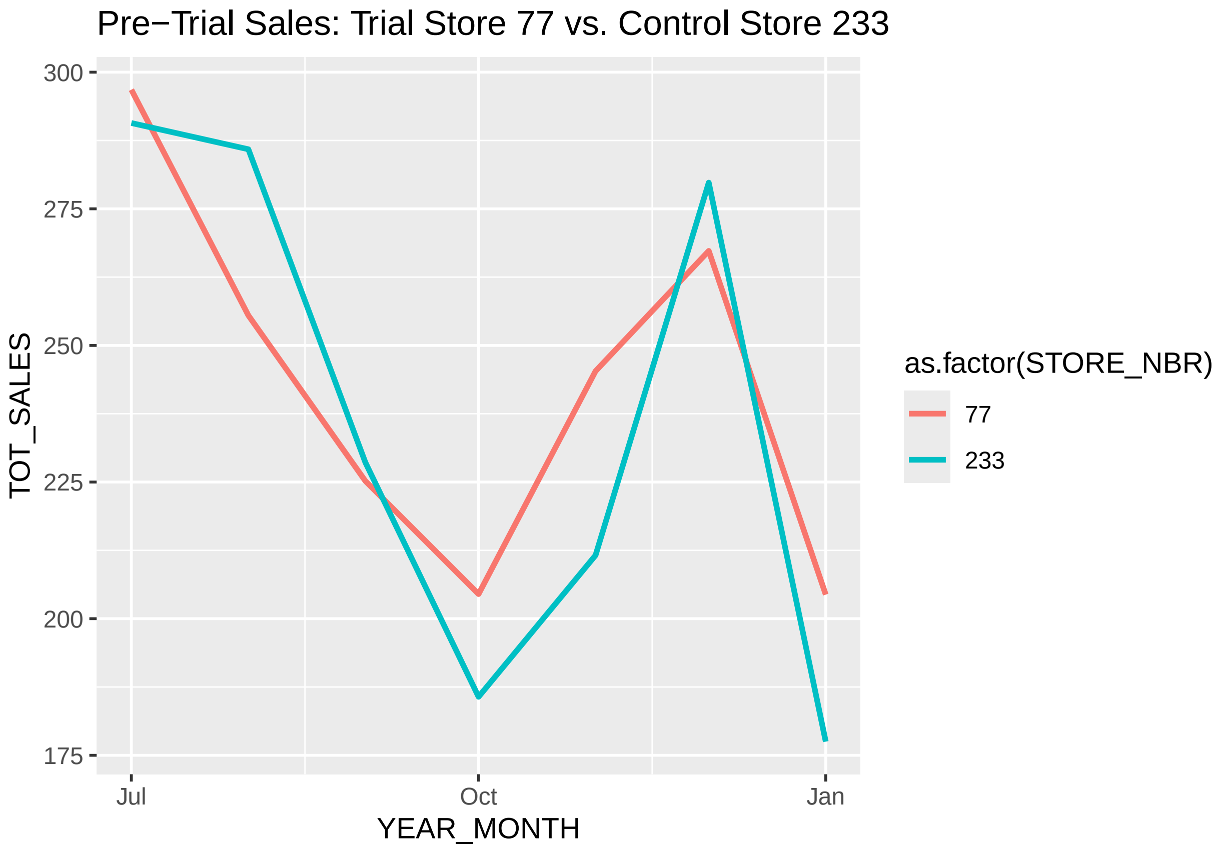
We run the function for Store 77. The top candidate is **Store 233**.

```
store_77_candidates <- find_control_store(77, pre_trial_metrics)
print(head(store_77_candidates, 5))
```

```
## # A tibble: 5 x 6
##   STORE_NBR_candidate CORR_SALES DIST_SALES SCALED_DIST SCALED_CORR
##                 <dbl>      <dbl>      <dbl>       <dbl>       <dbl>
## 1                   1     0.0752       419.      0.0342          NA
## 2                   2    -0.263        570.      0.0522          NA
## 3                   3     0.807       5827.      0.677           NA
## 4                   4    -0.263       7428       0.868           NA
## 5                   5    -0.111       4041.      0.465           NA
## # i 1 more variable: SIMILARITY_SCORE <dbl>
```

```
# Plot the visual confirmation
control_store_id_77 <- 233
```

```
pre_trial_metrics %>%
  filter(STORE_NBR == 77 | STORE_NBR == control_store_id_77) %>%
  ggplot(aes(x = YEAR_MONTH, y = TOT_SALES, color = as.factor(STORE_NBR))) +
  geom_line(size = 1) +
  labs(title = "Pre-Trial Sales: Trial Store 77 vs. Control Store 233")
```



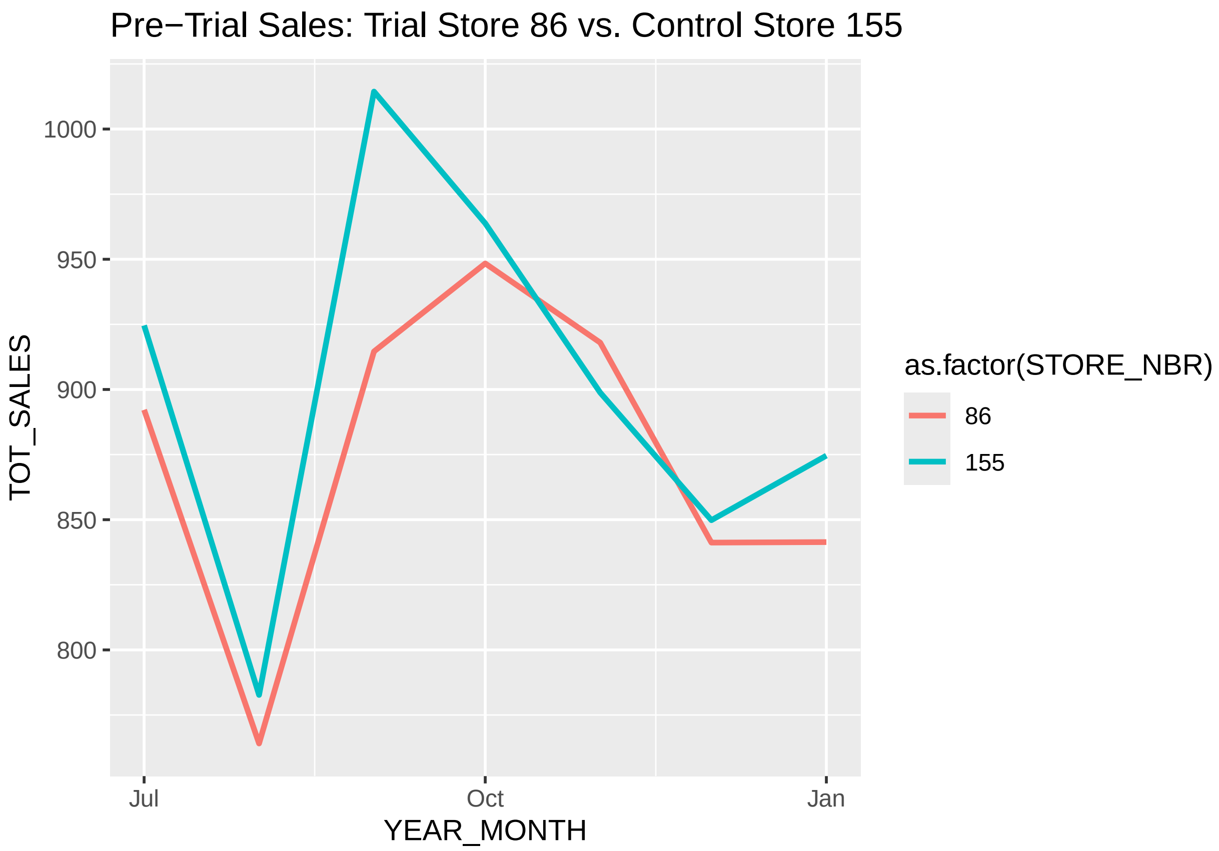
3.3 3.3. Trial Store 86

We run the function for Store 86. The top candidate is **Store 155**.

```
store_86_candidates <- find_control_store(86, pre_trial_metrics)
print(head(store_86_candidates, 5))
```

```
## # A tibble: 5 x 6
##   STORE_NBR_candidate CORR_SALES DIST_SALES SCALED_DIST SCALED_CORR
##                 <dbl>      <dbl>      <dbl>       <dbl>       <dbl>
## 1                   1     0.446       4733.      0.779           NA
## 2                   2    -0.404       4991.      0.824           NA
## 3                   3    -0.261       1406.      0.205           NA
## 4                   4    -0.0390      3007.      0.481           NA
## 5                   5     0.235        456.      0.0406          NA
## # i 1 more variable: SIMILARITY_SCORE <dbl>
```

```r
# Plot the visual confirmation
control_store_id_86 <- 155
pre_trial_metrics %>%
  filter(STORE_NBR == 86 | STORE_NBR == control_store_id_86) %>%
  ggplot(aes(x = YEAR_MONTH, y = TOT_SALES, color = as.factor(STORE_NBR))) +
  geom_line(size = 1) +
  labs(title = "Pre-Trial Sales: Trial Store 86 vs. Control Store 155")
```



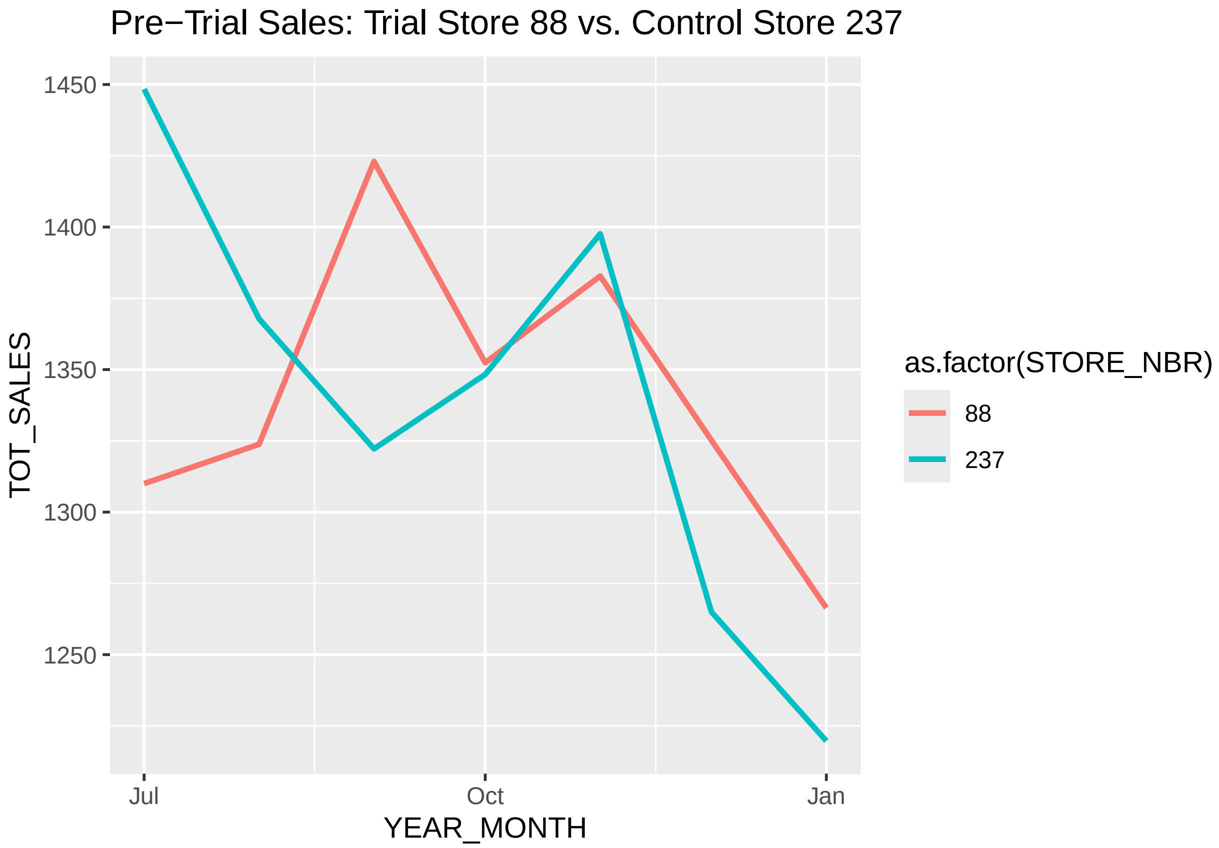
3.4 3.4. Trial Store 88

We run the function for Store 88. The top candidate is **Store 237**.

```r
store_88_candidates <- find_control_store(88, pre_trial_metrics)
print(head(store_88_candidates, 5))
```

```
## # A tibble: 5 x 6
##   STORE_NBR_candidate CORR_SALES DIST_SALES SCALED_DIST SCALED_CORR
##                 <dbl>      <dbl>      <dbl>       <dbl>       <dbl>
## 1                   1     0.814       7997.      0.856           NA
## 2                   2    -0.0679      8255.      0.885           NA
## 3                   3    -0.508       1857.      0.163           NA
## 4                   4    -0.746        954.      0.0614          NA
## 5                   5     0.190       3644.      0.365           NA
## # i 1 more variable: SIMILARITY_SCORE <dbl>
```

```r
# Plot the visual confirmation
control_store_id_88 <- 237
pre_trial_metrics %>%
  filter(STORE_NBR == 88 | STORE_NBR == control_store_id_88) %>%
  ggplot(aes(x = YEAR_MONTH, y = TOT_SALES, color = as.factor(STORE_NBR))) +
  geom_line(size = 1) +
  labs(title = "Pre-Trial Sales: Trial Store 88 vs. Control Store 237")
```



4 4. Assessment of the Trial

Now we compare the performance of each pair *during* the trial (Feb, Mar, Apr 2019). We create a function to assess the percentage difference in a given metric against the pre-trial 95% confidence interval.

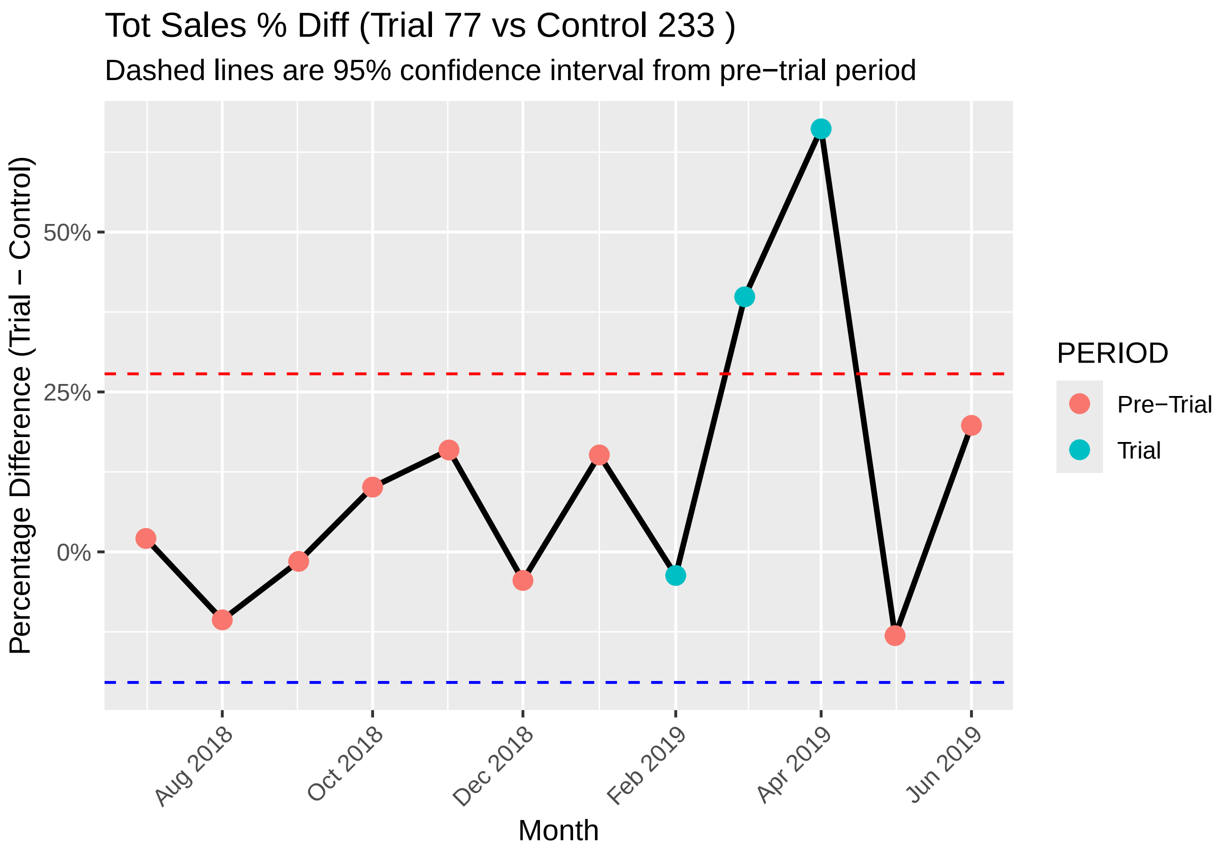
4.1 4.1. Trial Assessment Function

```r
# Define trial dates
trial_start_date <- as.Date("2019-02-01")
trial_end_date <- as.Date("2019-04-30")

assess_trial <- function(trial_store_id, control_store_id, metric_name, all_monthly_data) {

  plot_metric_name <- str_to_title(str_replace_all(metric_name, "_", " "))
  # Filter and Calculate % Difference
  pair_data <- all_monthly_data %>%
    filter(STORE_NBR == trial_store_id | STORE_NBR == control_store_id)

  pair_metric <- pair_data %>%
    select(STORE_NBR, YEAR_MONTH, !!sym(metric_name)) %>%
    pivot_wider(names_from = STORE_NBR, values_from = !!sym(metric_name), names_prefix = "STORE_")

  trial_col_name <- paste0("STORE_", trial_store_id)
  control_col_name <- paste0("STORE_", control_store_id)

  pair_metric <- pair_metric %>%
    mutate(
      PERCENT_DIFF = (!!sym(trial_col_name) - !!sym(control_col_name)) / !!sym(control_col_name),
      PERIOD = if_else(YEAR_MONTH >= trial_start_date & YEAR_MONTH <= trial_end_date, "Trial", "Pre-Tri
    ) %>%
    filter(is.finite(PERCENT_DIFF))

  # Assess Significance
  pre_trial_diffs <- pair_metric %>%
    filter(PERIOD == "Pre-Trial") %>%
    pull(PERCENT_DIFF)

  mean_pre_trial_diff <- mean(pre_trial_diffs)
  std_dev_pre_trial_diff <- sd(pre_trial_diffs)

  upper_bound_95 <- mean_pre_trial_diff + 2 * std_dev_pre_trial_diff
  lower_bound_05 <- mean_pre_trial_diff - 2 * std_dev_pre_trial_diff

  # Visualize the Assessment
  plot_title <- paste(plot_metric_name, "% Diff (Trial", trial_store_id, "vs Control", control_store_id

  assessment_plot <- ggplot(pair_metric, aes(x = YEAR_MONTH, y = PERCENT_DIFF)) +
    geom_line(size = 1) +
    geom_point(aes(color = PERIOD), size = 3) +
    geom_hline(yintercept = upper_bound_95, linetype = "dashed", color = "red") +
    geom_hline(yintercept = lower_bound_05, linetype = "dashed", color = "blue") +
    labs(
      title = plot_title,
      subtitle = "Dashed lines are 95% confidence interval from pre-trial period",
      x = "Month", y = "Percentage Difference (Trial - Control)"
    ) +
    scale_y_continuous(labels = percent_format()) +
    scale_x_date(date_breaks = "2 month", date_labels = "%b %Y") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

  print(assessment_plot)
}
```

4.2 4.2. Assessment: Total Sales (TOT_SALES)

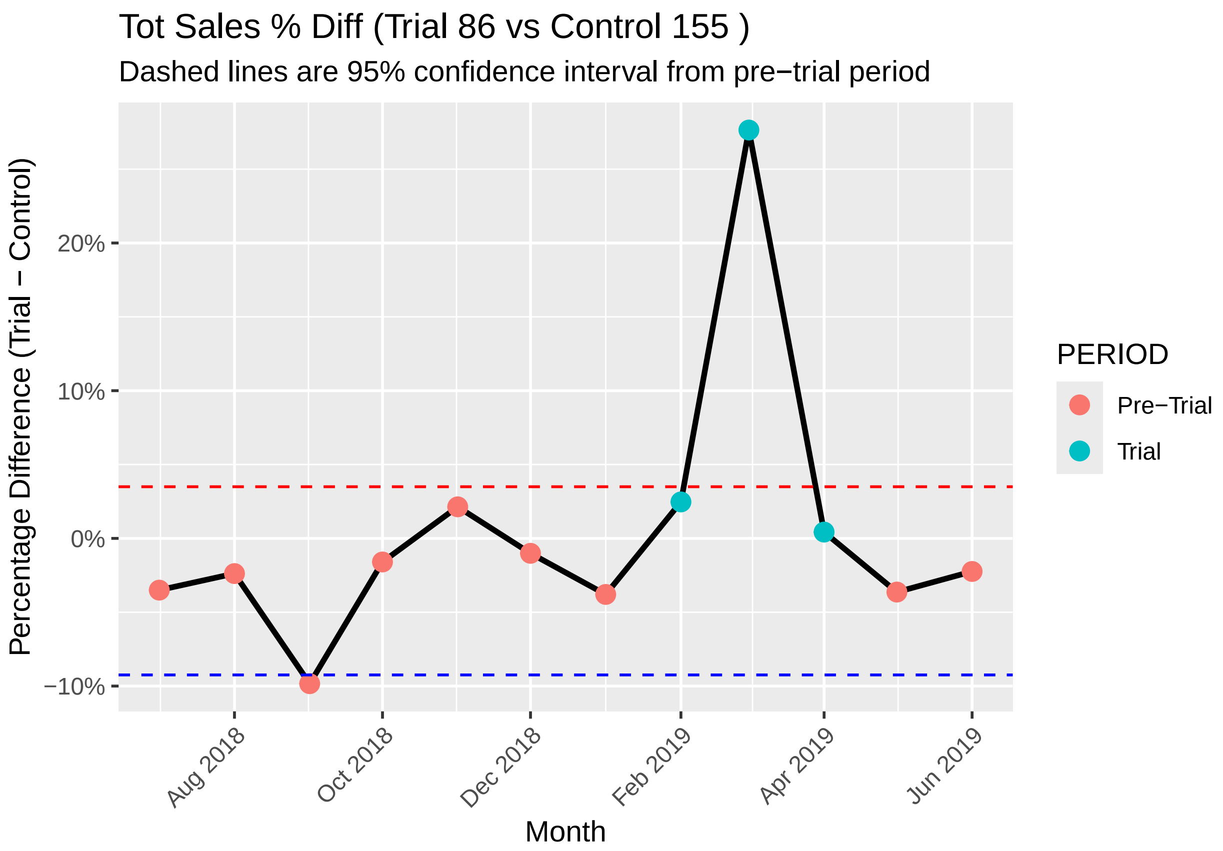
We now run the assessment for `TOT_SALES` for all three pairs.

```r
assess_trial(77, 233, "TOT_SALES", monthly_metrics)
```



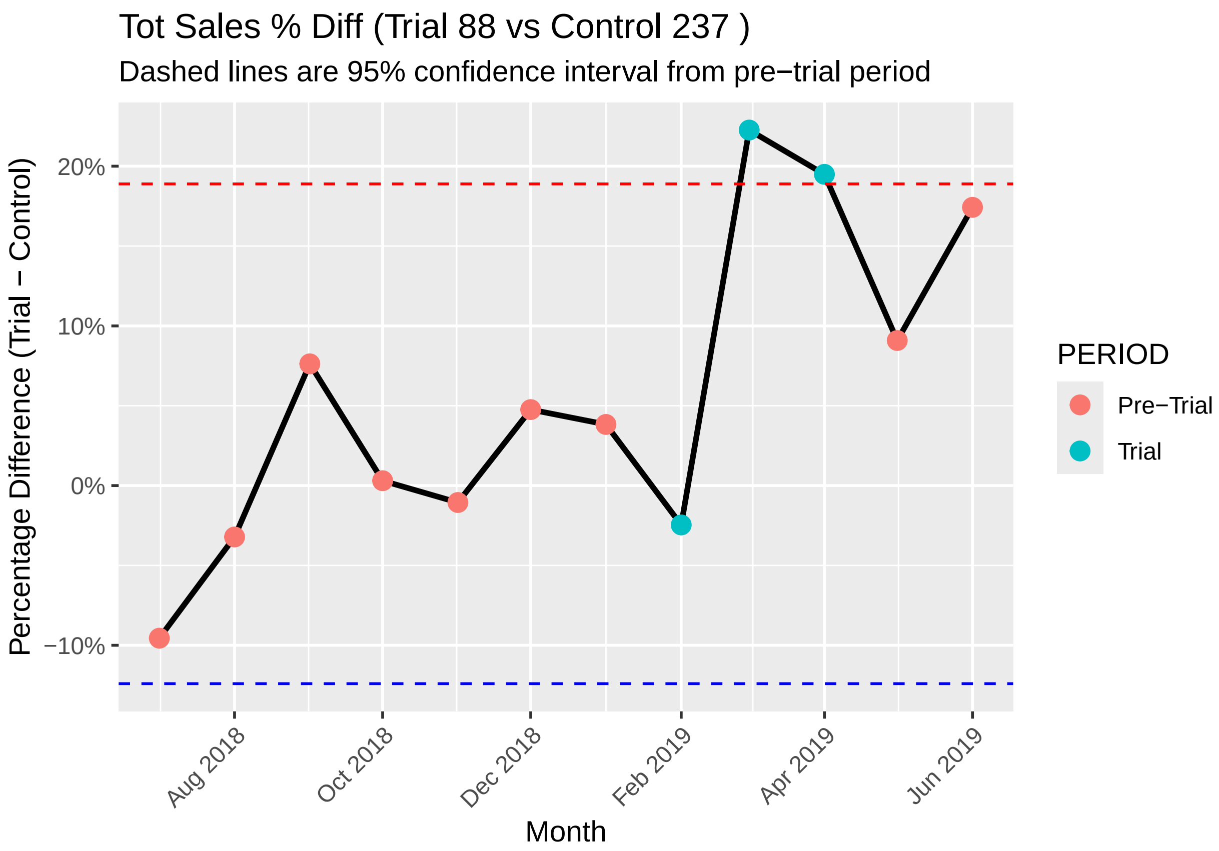
Insight: Store 77's sales difference was significantly *above* the 95% CI for all three trial months. This shows a strong positive impact.

```r
assess_trial(86, 155, "TOT_SALES", monthly_metrics)
```



Insight: Store 86's sales difference fluctuated, but largely remained *within* the "normal" range. This suggests the trial had little to no significant impact on sales.

```
assess_trial(88, 237, "TOT_SALES", monthly_metrics)
```

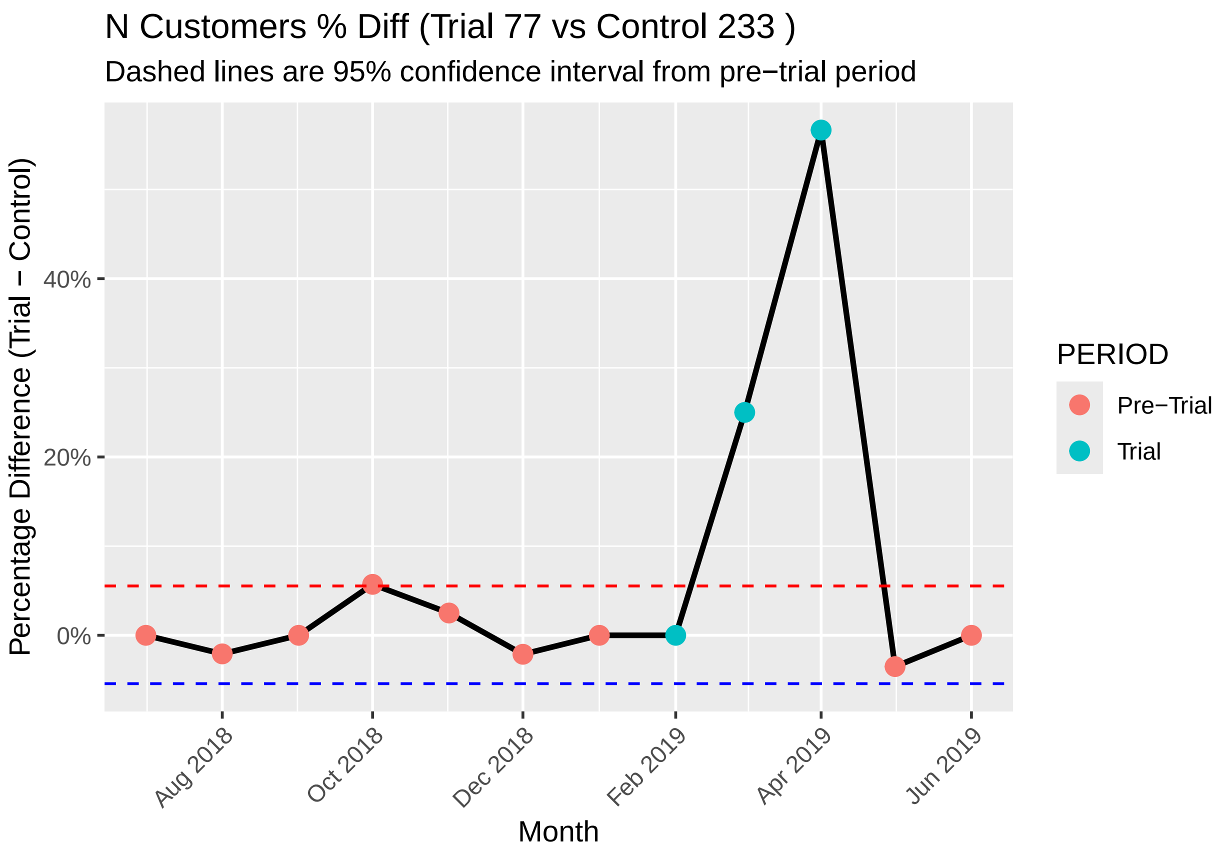


Insight: Store 88's sales difference was significantly *above* the 95% CI for two of the three trial months, with a strong spike in the final month. This shows a positive impact.

4.3 4.3. Assessment: Number of Customers (N_CUSTOMERS)

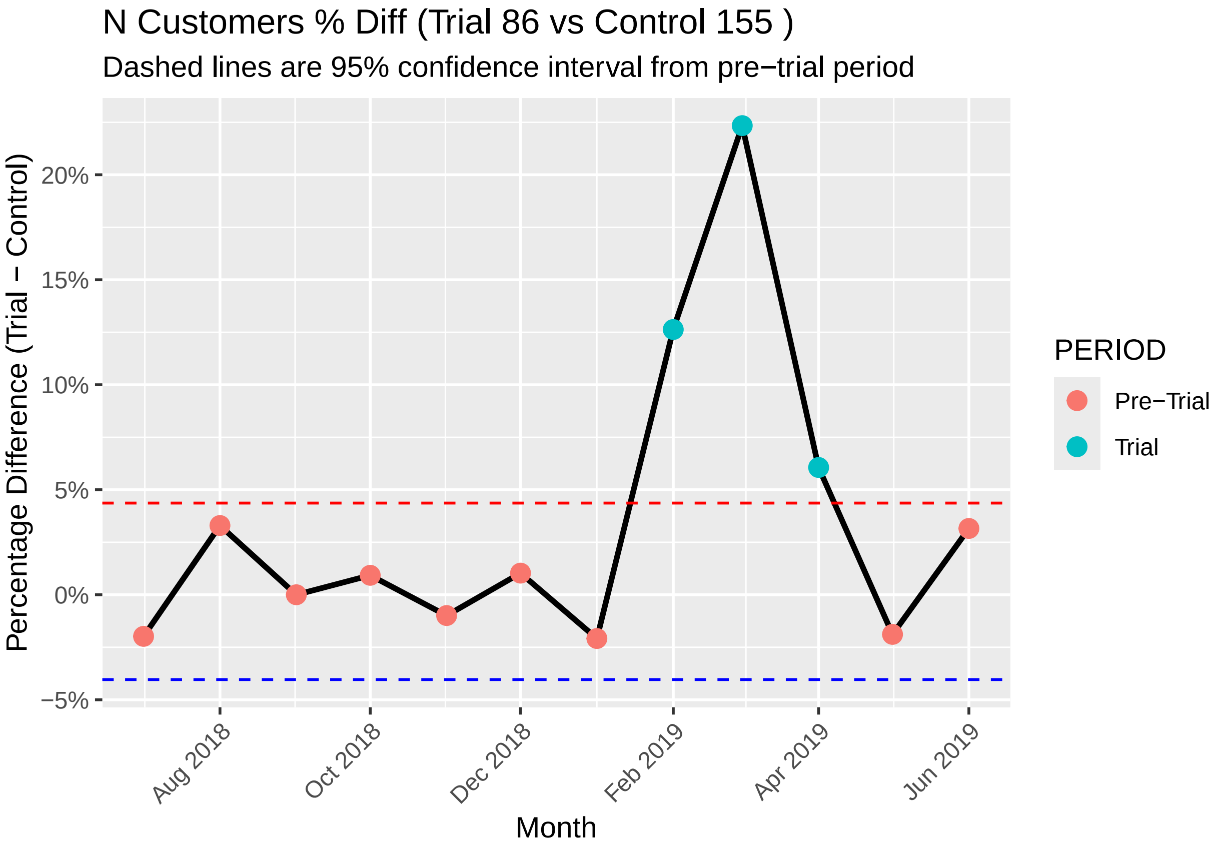
We check if the sales increase was driven by more customers.

```
assess_trial(77, 233, "N_CUSTOMERS", monthly_metrics)
```



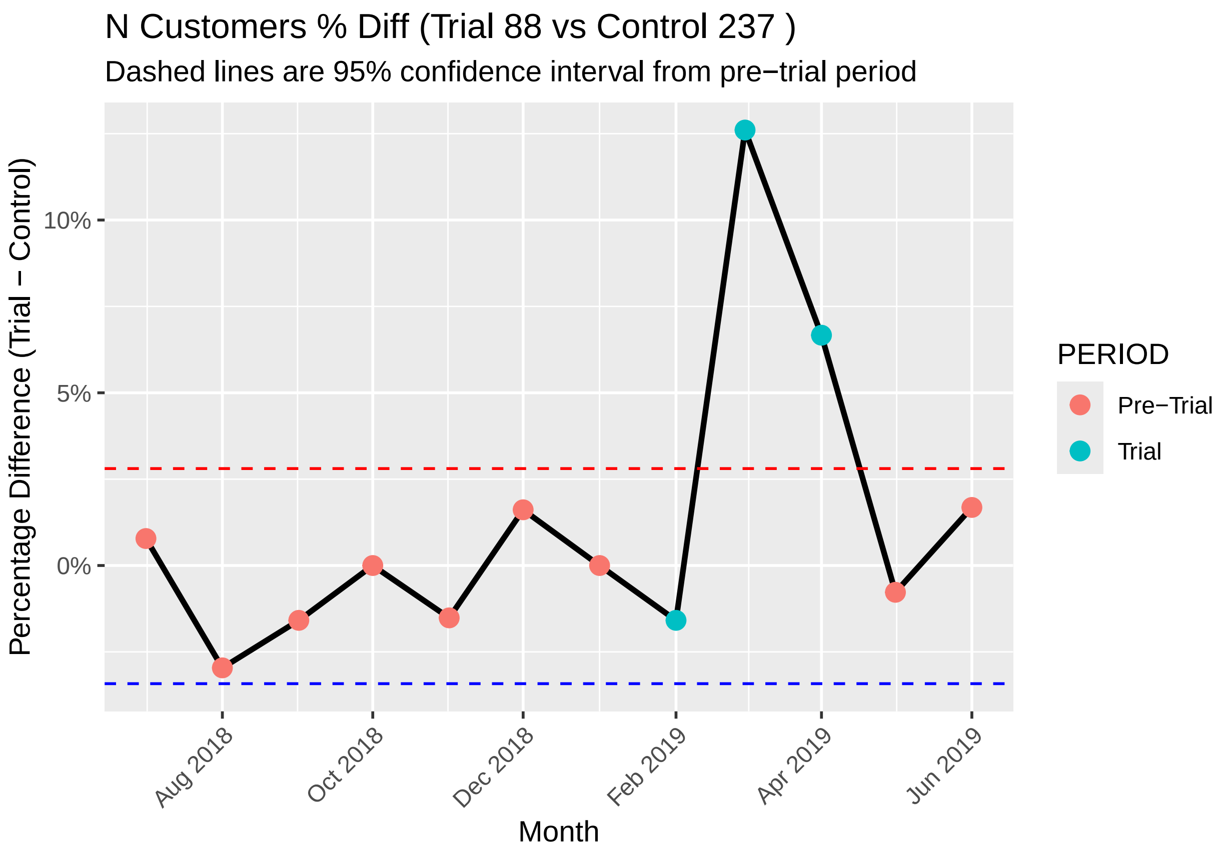
Insight: Store 77 also saw a significant increase in customer numbers for all three trial months, mirroring the sales data.

```
assess_trial(86, 155, "N_CUSTOMERS", monthly_metrics)
```



Insight: Store 86's customer numbers also remained within the normal bounds, confirming the lack of a trial effect.

```
assess_trial(88, 237, "N_CUSTOMERS", monthly_metrics)
```



Insight: Store 88's customer numbers were also significantly higher for two of the three months, aligning with its sales increase.

5 5. Collate Findings & Recommendation

Based on the analysis, here are our findings for Julia:

- **Trial Store 77:** The trial was a **clear success**. The store showed a significant and consistent increase in both total sales and customer numbers throughout the trial period.
- **Trial Store 86:** The trial was **not successful**. There was no significant change in sales or customer numbers compared to its control store. We should investigate with the client if the trial was implemented differently here.
- **Trial Store 88:** The trial was **a success**. It showed a significant increase in sales and customer numbers for two of the three months, indicating a positive impact.

Recommendation: The trial layout appears to have a significant positive impact on performance. We recommend **rolling out the trial layout to all stores**, particularly emulating the success seen in stores 77 and 88. We should, however, deprioritize the rollout for stores with similar characteristics to Store 86, or investigate why the trial failed in that specific location. “‘