[redacted] Transaction Analysis

Stefanie Molin

April 19, 2017

The Problem ([redacted])

Very senior execs at [redacted] are running into situations where they are seeing purchased products in the banners. We did a test order and found that the product ID didn't fire via the conversion tag in our test order. Given that, is there a way for you to see how many sales didn't have product IDs fire in the *last 30 days*? We're now concerned that we could be showing purchased products in banners **AND** missing out on sales.

Analysis

We are going to do this in 3 steps:

1. Query Vertica for data
2. Use `dplyr` to manipulate it
3. Graph data using `ggplot2`

Query Vertica

Pull the following information by day:

- number of transactions with null internal product IDs
- total transactions

```
# QueryVertica already loaded along with user/pass
query <- "
SELECT
  day
  , COUNT(DISTINCT (CASE WHEN product_internal_id IS NULL
      THEN transaction_id ELSE NULL END)) AS null_id_sales
  , COUNT(DISTINCT transaction_id) AS total_sales
FROM
    [redacted]
WHERE
    day >= CURRENT_DATE() - 30
    AND merchant_id = 5535
GROUP BY
  day
"

data <- QueryVertica(username, query, password)
```

Manipulate Data

Use `dplyr` to prepare the data for the graph and for the conclusion; for the graph we need daily data, but for the conclusion we need an overall. Let's also add the percent of null internal ID sales to the total.

```r
library(dplyr)

# conclusion data
data_conclusion <- data %>%
  summarize(percent_null = sum(null_id_sales, na.rm = TRUE)/sum(total_sales,
                                                                na.rm = TRUE))

# graph data
data_pivot <- data %>%
  mutate(percent_null = null_id_sales/total_sales, day = as.Date(day)) %>%
  select(day, percent_null) %>%
  arrange(day)

# view graph data
library(knitr)
knitr::kable(head(data_pivot))
```

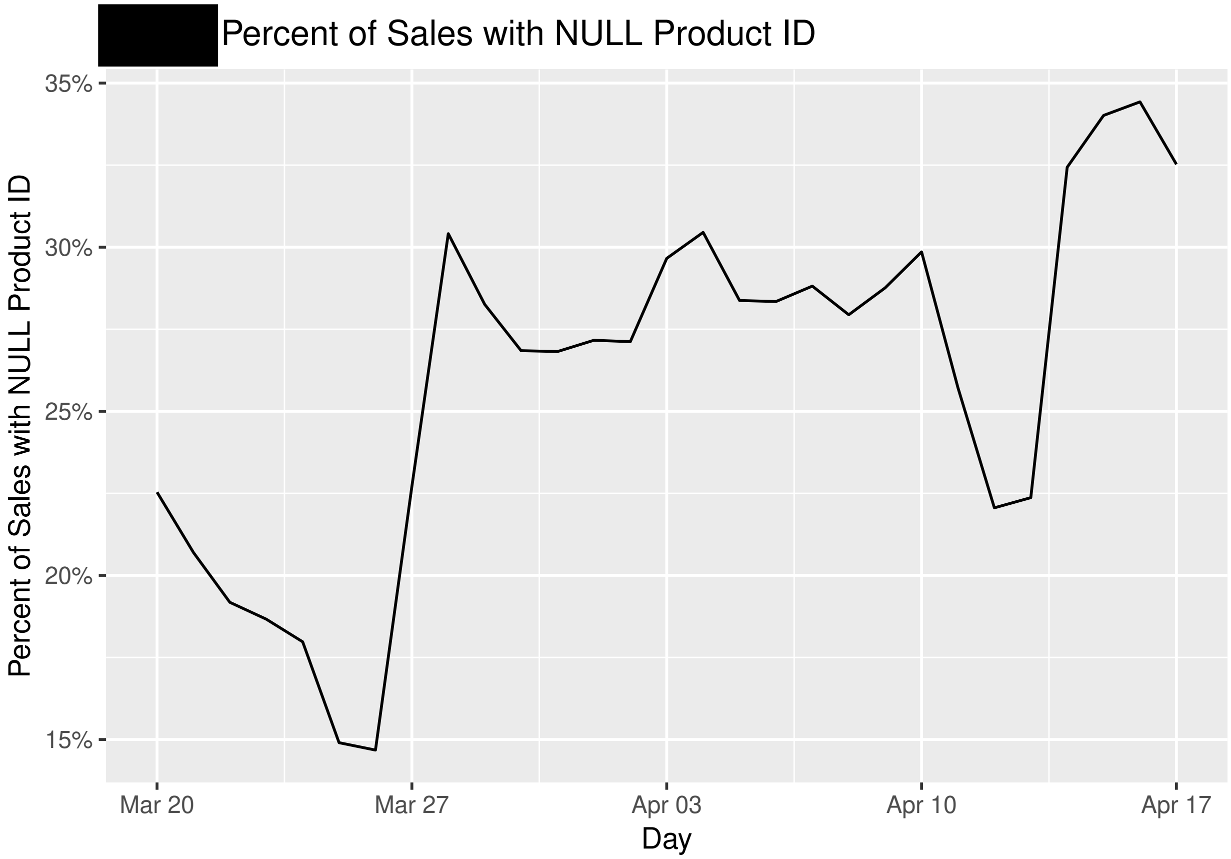
day	percent_null
2017-03-20	0.2253502
2017-03-21	0.2069759
2017-03-22	0.1917937
2017-03-23	0.1866545
2017-03-24	0.1797596
2017-03-25	0.1490247

Graph Data

With `ggplot2` create a line graph of this percentage by day.

```r
library(ggplot2)

data_pivot %>%
  ggplot(aes(x = day, y = percent_null)) +
  geom_line() +
  ggtitle("[redacted] Percent of Sales with NULL Product ID") +
  labs(x = "Day", y = "Percent of Sales with NULL Product ID") +
  scale_y_continuous(labels = scales::percent)
```



Conclusion

The way we store data on our end seems to involve a convoluted process that takes the external product ID from the client finds its internal ID in our system to store in the database; then any time we have the external ID it is from looking in our system not what the client passed. This seems to be the case because I don't see any sales in the last 30 days with an external product ID of NULL, however, I see **25.76% of sales** in the *last 30 days* that have had a NULL as the internal ID.