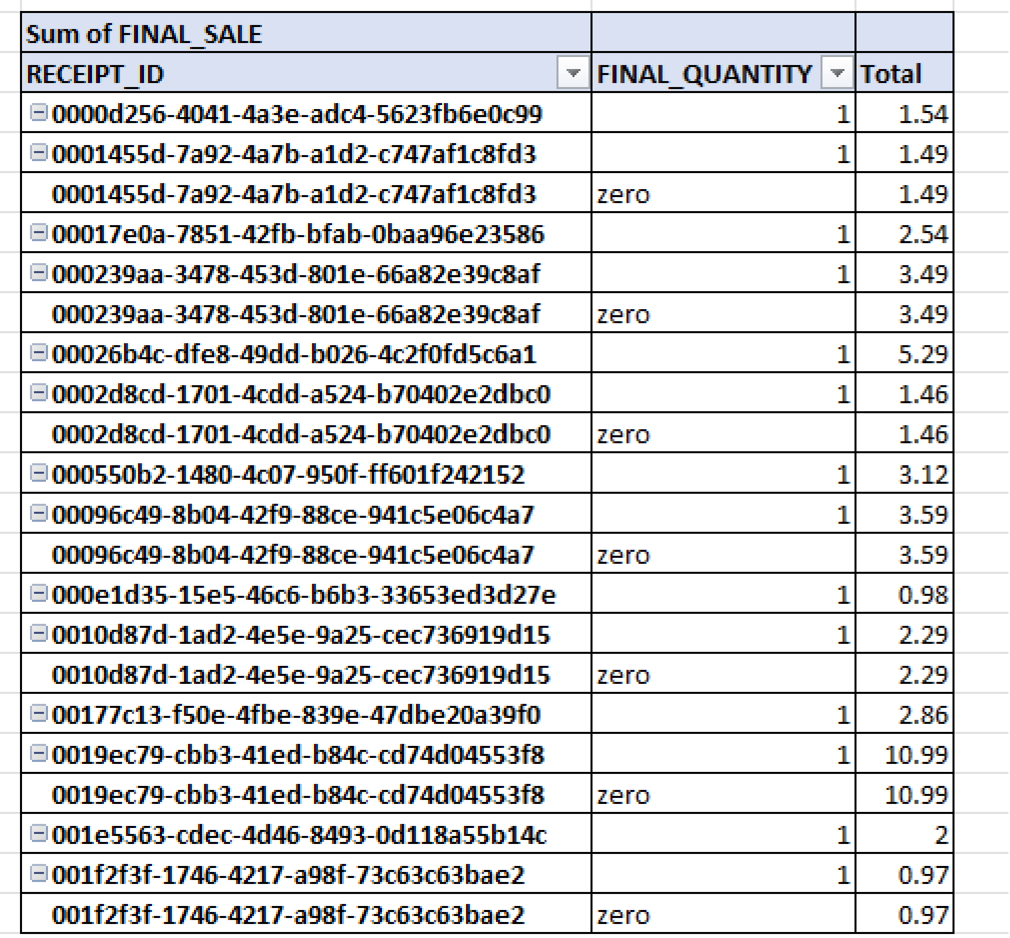
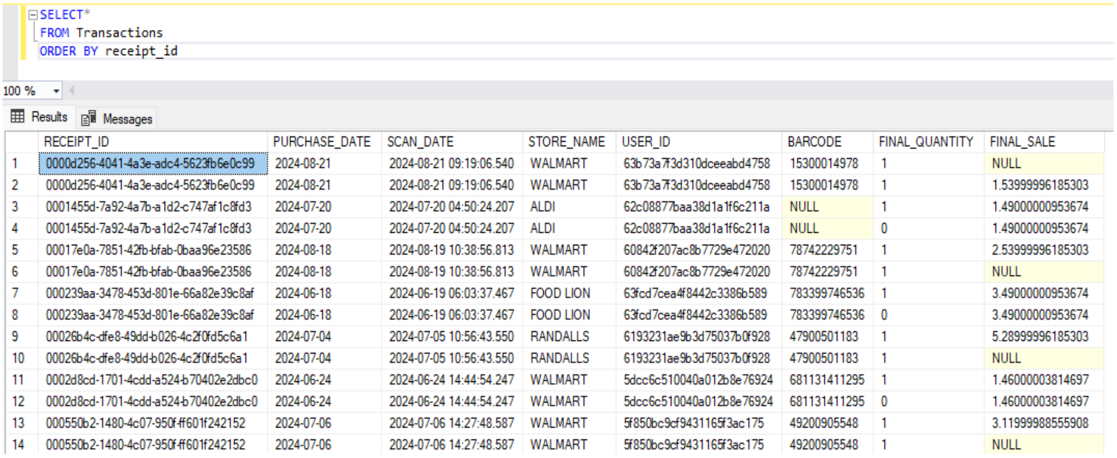
Data Quality Issues

- **Transactions**
  - There are 12,500 rows that have "zero" as a value in the final_quantity field (Example below). This should not be possible as the final_quantity field should be numeric and should be "0", not "zero". This change was made to allow for proper ingestion into SQL database and queries are reflective of this change being made.



Sum of FINAL_SALE		
RECEIPT_ID	FINAL_QUANTITY	Total
⊟0000d256-4041-4a3e-adc4-5623fb6e0c99	1	1.54
⊟0001455d-7a92-4a7b-a1d2-c747af1c8fd3	1	1.49
0001455d-7a92-4a7b-a1d2-c747af1c8fd3	zero	1.49
⊟00017e0a-7851-42fb-bfab-0baa96e23586	1	2.54
⊟000239aa-3478-453d-801e-66a82e39c8af	1	3.49
000239aa-3478-453d-801e-66a82e39c8af	zero	3.49
⊟00026b4c-dfe8-49dd-b026-4c2f0fd5c6a1	1	5.29
⊟0002d8cd-1701-4cdd-a524-b70402e2dbc0	1	1.46
0002d8cd-1701-4cdd-a524-b70402e2dbc0	zero	1.46
⊟000550b2-1480-4c07-950f-ff601f242152	1	3.12
⊟00096c49-8b04-42f9-88ce-941c5e06c4a7	1	3.59
00096c49-8b04-42f9-88ce-941c5e06c4a7	zero	3.59
⊟000e1d35-15e5-46c6-b6b3-33653ed3d27e	1	0.98
⊟0010d87d-1ad2-4e5e-9a25-cec736919d15	1	2.29
0010d87d-1ad2-4e5e-9a25-cec736919d15	zero	2.29
⊟00177c13-f50e-4fbe-839e-47dbe20a39f0	1	2.86
⊟0019ec79-cbb3-41ed-b84c-cd74d04553f8	1	10.99
0019ec79-cbb3-41ed-b84c-cd74d04553f8	zero	10.99
⊟001e5563-cdec-4d46-8493-0d118a55b14c	1	2
⊟001f2f3f-1746-4217-a98f-73c63c63bae2	1	0.97
001f2f3f-1746-4217-a98f-73c63c63bae2	zero	0.97

  - All receipts have at least 2 entries each. Both entries appear to have the same Receipt_ID, Purchase_Date, Scan_Date, Store_Name, User_ID, and Barcode but differ in the final_quantity or final_sale. Seen in the image above and below.

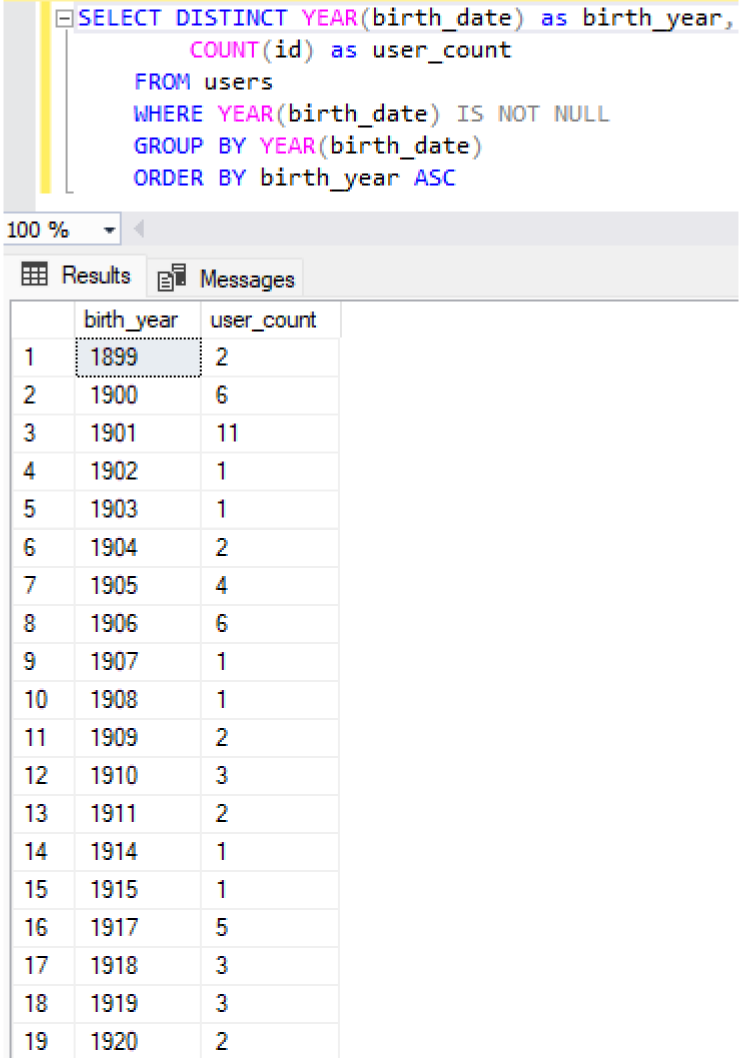


```sql
SELECT*
FROM Transactions
ORDER BY receipt_id
```

	RECEIPT_ID	PURCHASE_DATE	SCAN_DATE	STORE_NAME	USER_ID	BARCODE	FINAL_QUANTITY	FINAL_SALE
1	0000d256-4041-4a3e-adc4-5623fb6e0c99	2024-08-21	2024-08-21 09:19:06.540	WALMART	63b73a7f3d310dceeabd4758	15300014978	1	NULL
2	0000d256-4041-4a3e-adc4-5623fb6e0c99	2024-08-21	2024-08-21 09:19:06.540	WALMART	63b73a7f3d310dceeabd4758	15300014978	1	1.53999996185303
3	0001455d-7a92-4a7b-a1d2-c747af1c8fd3	2024-07-20	2024-07-20 04:50:24.207	ALDI	62c08877baa38d1a1f6c211a	NULL	1	1.49000000953674
4	0001455d-7a92-4a7b-a1d2-c747af1c8fd3	2024-07-20	2024-07-20 04:50:24.207	ALDI	62c08877baa38d1a1f6c211a	NULL	0	1.49000000953674
5	00017e0a-7851-42fb-bfab-0baa96e23586	2024-08-18	2024-08-19 10:38:56.813	WALMART	60842f207ac8b7729e472020	78742229751	1	2.53999996185303
6	00017e0a-7851-42fb-bfab-0baa96e23586	2024-08-18	2024-08-19 10:38:56.813	WALMART	60842f207ac8b7729e472020	78742229751	1	NULL
7	000239aa-3478-453d-801e-66a82e39c8af	2024-06-18	2024-06-19 06:03:37.467	FOOD LION	63fcd7cea4f8442c3386b589	783399746536	1	3.49000000953674
8	000239aa-3478-453d-801e-66a82e39c8af	2024-06-18	2024-06-19 06:03:37.467	FOOD LION	63fcd7cea4f8442c3386b589	783399746536	0	3.49000000953674
9	00026b4c-dfe8-49dd-b026-4c2f0fd5c6a1	2024-07-04	2024-07-05 10:56:43.550	RANDALLS	6193231ae9b3d75037b0f928	47900501183	1	5.28999996185303
10	00026b4c-dfe8-49dd-b026-4c2f0fd5c6a1	2024-07-04	2024-07-05 10:56:43.550	RANDALLS	6193231ae9b3d75037b0f928	47900501183	1	NULL
11	0002d8cd-1701-4cdd-a524-b70402e2dbc0	2024-06-24	2024-06-24 14:44:54.247	WALMART	5dcc6c510040a012b8e76924	681131411295	1	1.46000003814697
12	0002d8cd-1701-4cdd-a524-b70402e2dbc0	2024-06-24	2024-06-24 14:44:54.247	WALMART	5dcc6c510040a012b8e76924	681131411295	0	1.46000003814697
13	000550b2-1480-4c07-950f-ff601f242152	2024-07-06	2024-07-06 14:27:48.587	WALMART	5f850bc9cf9431165f3ac175	49200905548	1	3.11999988555908
14	000550b2-1480-4c07-950f-ff601f242152	2024-07-06	2024-07-06 14:27:48.587	WALMART	5f850bc9cf9431165f3ac175	49200905548	1	NULL

- **Users**

  - There are some users that have birth years that, while valid, are highly unlikely given that they take place 100+ years before the founding of Fetch. These are likely false birthdays and would need to be excluded from any age-based analyses.



```
SELECT DISTINCT YEAR(birth_date) as birth_year,
        COUNT(id) as user_count
    FROM users
    WHERE YEAR(birth_date) IS NOT NULL
    GROUP BY YEAR(birth_date)
    ORDER BY birth_year ASC
```

	birth_year	user_count
1	1899	2
2	1900	6
3	1901	11
4	1902	1
5	1903	1
6	1904	2
7	1905	4
8	1906	6
9	1907	1
10	1908	1
11	1909	2
12	1910	3
13	1911	2
14	1914	1
15	1915	1
16	1917	5
17	1918	3
18	1919	3
19	1920	2

  - There were 3,675 users with no birth_date, 30,508 users with no language, 4,812 users with no state, and 5,892 user with no gender listed. While these are not necessarily issues, they may limit the groupings and audiences that can be developed off of non-engagement characteristics.
  - Additionally, only 91 of the 17,694 user_ids in the Transactions table appear in the user table. Once again, this is not an issue that will cause a breakage but does limit the user-level analysis that can be done on the transaction data and implies that the User table is incomplete.

- **Products**
  - 4,025 rows are missing a barcode, leaving no way to tie them back to transactions and include them in analyses.
  - 226,472 rows have no brand or manufacturer and cannot be identified past the category level
  - 2 Listerine rows are missing their manufacturer. As there are many other Listerine entries that feature a manufacturer, this is an extremely simple fix even if needed to be done manually.
  - 547 barcodes across 11 brands have a category_1 value of “Needs Review” and need to be addressed in order to accurately capture their information during analysis

Challenging Fields to Understand

- I did not feel that any of the fields were difficult to understand. One point of note was that I could see the multiple category fields within the products table becoming confusing in conversation and discussion and could potentially benefit from differing names. Potentially something as simple as Primary Category, Sub-Category, Product Type, Product Subtype.