# Data reviews

## /\* Receipts \*/

**select** **max**(*r*.totalSpent)

**from** Receipts *r*

-- 4721.95

*-- Thats a big number, review that receipt*

**select** \*

**from** Receipts *r*

**where** totalSpent = (**select** **max**(*r*.totalSpent) **from** Receipts *r* )

* *Turns out there are 92 records with the same amount. All occured on 2021-01-24, All the same user, but all have different bar codes, all have 620 as the Item Count*

**select** **avg**(totalSpent)

**from** Receipts *r*

*-- 927.67*

* *Quite a bit bigger than the average*

**select** **min**(*r*.totalSpent)

**from** Receipts *r*

-- 0

-- Take another look at the $0 transactions.

**select** \*

**from** Receipts *r*

**where** totalSpent = (**select** **min**(*r*.totalSpent) **from** Receipts *r* )

* *Nothing unusual here, different users, dates, items. They are all marked as FLAGGED however.*

/\* Key column counts \*/

**select** **count** (\*)

**from** Receipts *r*

*-- 2060*

**select** **count** (**distinct** *r*.ReceiptID)

**from** Receipts *r*

*-- 679*

* *Can we assume that multiple duplicate receipts are additional pages of the same receipt?*

**select** **count**(\*)

**from** Receipts *r*

**where** isnull(*r*.barcode,'') = ''

* *158 receipts with a missing bar code*

**select** **count**(*r*.barcode)

**from** Receipts *r*

**left** **join** Brands *b*

**on** *r*.barcode = *b*.barcode

**where** *b*.barcode **is** **not** **NULL**

*-- Only 64 Barcodes on the receipts file have a matching barcode in the Brands table. This might have dupes*

-- dupecheck

**select** \*

**from** Receipts *r*

**join** (**select** **distinct** barcode **from** Brands *b*) *b*

**on** *r*.barcode = *b*.barcode

* + Yes, there are 4 dupes.

-- Inconsitencies

**select** len(**rtrim**(*r*.receiptID)), **count**(\*)

**from** Receipts *r*

**group** **by** len(**rtrim**(*r*.receiptID))

Receipt ID's are all 26 bytes

**select** len(**rtrim**(*r*.barcode)), **count**(\*)

**from** Receipts *r*

**group** **by** len(**rtrim**(*r*.barcode))

**order** **by** len(**rtrim**(*r*.barcode)) **desc**

***- Barcode Inconsitencies; There are (7) different bar code lengths. Assuming these are UPC codes, since they’re goods for sale, the only valid lengths are 12 and 8.***

|  |  |
| --- | --- |
| **select len(rtrim(r.barcode)), count(\*) from Receipts r  group by len(rtrim(r.barcode)) order by len(rtrim(r.barcode)) desc** | |
|  |  |
| 13 | 4 |
| 12 | 1,465 |
| 11 | 31 |
| 10 | 82 |
| 5 | 2 |
| 4 | 317 |
| 2 | 1 |
| Blank | 158 |

Receipt Status

**select** **distinct** rewardsReceiptStatus

**from** Receipts *r*

*-- 4 types; PENDING, FLAGGED, FLAGGED, REJECTED*

**select** \*

**from** Receipts *r*

**where** isnull(rewardsReceiptStatus,'') = ''

*-- All records have a status*

-- take a look at the dates

**select** \*

**from** Receipts *r*

**where** createDate > dateScanned

**OR**

createDate > finishedDate

**or** createDate > modifyDate

*-- Nothing with a processing date earlier than the created date*

**select** \*

**from** Receipts *r*

**where** dateScanned > finishedDate

-- Nothing here

**select** \*

**from** Receipts *r*

**where** dateScanned < purchaseDate

*-- There are 13 records where the scanned date is prior to the purchase date. How can you scan an item before it's purchased?*

--Investigate this  
**select** *u*.role, *u*.userID , *r*.\*

**from** Receipts *r*

**join** Users *u*

**on** *r*.userId = *u*.userID

**where** dateScanned < purchaseDate

-- All consumer purchases, different purchasers

-- Any lag in processing?

**select**

ReceiptID,

purchaseDate,

dateScanned,

**DATEDIFF**(DAY, purchaseDate, dateScanned) **AS** *days\_difference*,

**DATEDIFF**(HOUR, purchaseDate, dateScanned) **AS** *hours\_difference*

**from** Receipts *r*

**order** **by** **DATEDIFF**(DAY, purchaseDate, dateScanned) **desc**

***-- There are several that have a time lag over 1,000 days***

***-- Duplicate rows***

***select***

***ReceiptKey,***

***ReceiptID,***

***barcode,***

***createDate,***

***dateScanned,***

***finishedDate,***

***modifyDate,***

***purchaseDate,***

***purchasedItemCount,***

***rewardsReceiptStatus,***

***totalSpent,***

***userId,***

***count(\*)***

***from***

***Receipts r***

***group by***

***ReceiptKey,***

***ReceiptID,***

***barcode,***

***createDate,***

***dateScanned,***

***finishedDate,***

***modifyDate,***

***purchaseDate,***

***purchasedItemCount,***

***rewardsReceiptStatus,***

***totalSpent,***

***userId***

***having***

***count(\*) >1***

* ***None***

## Users

**select** **count**(\*)

**from** Users *u*

-- 212 Users

**select** **count**(**distinct** userID)

**from** Users *u*

-- 212

* *No duplicate ID’s*

**select** **case** active

**when** 1 **then** 'Active'

**when** 0 **then** 'Inactive'

**end** **as** *Active*,

**count**(\*) **as** *records*

**from** Users *u*

**group** **by** active

**order** **by** active

* *211 active, 1 inactive user*

**select** **count**(\*)

**from** Users *u*

**where** isnull(createdDate,'') <> '' **and** isnull(lastLogin,'') = ''

* *40 users with no logins*

**select** *u*.signUpSource, **count**(\*)

**from** Users *u*

**group** **by** *u*.signUpSource

* *204 users signed up via Email, 3 vial Google, and 5 have no email source*

**select** *u*.state, **count**(\*)

**from** Users *u*

**group** **by** *u*.state

* *Users come from 8 states, with Wisconsin making up about 90% of them*

**select**

userID,

active,

createdDate,

lastLogin,

[role],

signUpSource,

state,

**count**(\*)

**from**

Users *u*

**group** **by**

userID,

active,

createdDate,

lastLogin,

[role],

signUpSource,

state

**having**

**count** (\*) > 1

* *No duplicate records*

## Brands

**select** **count**(\*)

**from** Brands *b*

-- 1167 records

**select** **count**(**distinct** brandID)

**from** Brands *b*

-- 1167

* No duplicate Brand ID’s

**select** **count**(\*)

**from** Brands *b*

**where** isnull(barcode,'') = ''

* *No missing barcodes*

**select** len(**rtrim**(*b*.barcode))

**from** Brands *b*

**group** **by** len(**rtrim**(*b*.barcode))

**order** **by** len(**rtrim**(*b*.barcode)) **DESC**

* *All barcodes are 12 bytes*

**select** *bc*.[ref], **count** (\*)

**from** Brands *b*

**join** BrandCPGs *bc*

**on** *b*.barcode = *bc*.barcode

**group** **by** *bc*.[ref]

* There are 1034 COGS reference records, and 147 Cpgs. Why is there a mix of Cost related references with Package goods related records? they don’t seem to belong together.
* There are 612 brands with no “top” designation.

|  |  |
| --- | --- |
| **select b.topBrand, count(\*) from Brands b  group by topBrand  order by topBrand desc** | |
| **topBrand** |  |
| 1 | 31 |
| 0 | 524 |
|  | 612 |