




Community and Clinical Data Initiative (CODI) Data Model Quality Assurance Script

v1.0 - April 8, 2025

The purpose of this tool is to enable participating community-based organizations to check their structured data extract (SDE) file against the CODI SDE specification. This tool assumes that a SDE file has been created; if not, please see the CODI SDE mapping tool which is available at <https://mitre.github.io/codi/>. Use of this tool may be limited if the health information exchange (HIE) that you will be exchanging data with has not adopted the CODI Data Model; it is recommended that you coordinate with your HIE before developing the SDE file.

This script includes a total of 18 checks. The first 6 checks validate conformance to the CODI SDE specification. Checks 7 - 18 provide output that can be assessed to determine potential data quality issues. The legend below describes symbols used throughout to indicate potential issues. All issues should be addressed prior to sending the SDE file to your HIE partner. You may expand the code to review the logic for each check.

Legend

-  - No issues found
-  - Please review
-  - Presumed error found

Run date: April 17 2025, 02:52PM

Environment variable not set – using sample data
Loading files from ./2025_1

Conformance Checks

Check 1 - Do the expected tables exist?

Only tables with presumed errors are listed. If a table is missing due to a deliberate choice (e.g., REFERRAL table is missing because your organization does not have referral information), the error may be ignored.

The input SDE file has 9 of 12 expected tables.

- ✗ - QUESTION does not exist
- ✗ - QUESTION_RESPONSE does not exist
- ✗ - REFERRAL does not exist

Check 2 - Does each table have the expected attributes in the specified format?

Only attributes with presumed errors are listed. Tables with errors in Check 1 are excluded.

- ✗ - Found attribute mismatches in ASSET_ENROLLMENT
 - Attribute ASSET_ENROLLMENT.PROGRAMID is missing
 - Attribute ASSET_ENROLLMENT.ASSET_NAME is missing
 - Attribute ASSET_ENROLLMENT.ASSET_DESCRIPTION is missing
 - Attribute ASSET_ENROLLMENT.DELIVERY_END_DATE contains invalid values: {'0/0/0000'}
- ✗ - Found attribute mismatches in ASSET_DELIVERY
 - Attribute ASSET_DELIVERY.PROGRAMID is missing
 - Attribute ASSET_DELIVERY.SDOH_INTERVENTION_PROC is missing
 - Attribute ASSET_DELIVERY.ASSET_QUANTITY is missing
- ✗ - Found attribute mismatches in PROGRAM
 - Attribute PROGRAM.PROGRAM_SETTING is missing
- ✗ - Found attribute mismatches in PROGRAM_ENROLLMENT
 - Attribute PROGRAM_ENROLLMENT.REFERRALID is missing
 - Attribute PROGRAM_ENROLLMENT.ENROLLMENT_DATE contains invalid values: {'0/0/0000'}
- ✗ - Found attribute mismatches in SESSION
 - Attribute SESSION.PROGRAM_ENROLLMENT_ID is missing

Check 3 - Are the primary keys in each table unique?

Only tables where the number of distinct keys is not equal to the total number of rows (i.e., expected distinct keys) are shown. Tables with errors in Check 1 are excluded.

- ✖ – Found duplicate keys in PRIVATE_ADDRESS_HISTORY.ADDRESSID
 - Total rows: 5886
 - Distinct Keys: 4998
- ✖ – Found duplicate keys in ASSET_ENROLLMENT.ASSET_ENROLLMENT_ID
 - Total rows: 2430
 - Distinct Keys: 1281
- ✖ – Found duplicate keys in ASSET_DELIVERY.ASSET_DELIVERY_ID
 - Total rows: 6736
 - Distinct Keys: 5697

Check 4 - Are there any values that do not match defined value sets?

This check applies to attributes with a specified value set; only attributes with presumed errors are listed. The unexpected values observed are listed for each attribute. Attributes in tables with errors in Check 1 are excluded.

- ✖ – Found unexpected value(s) in DEMOGRAPHIC
 - Attribute DEMOGRAPHIC.SEX contains: ['OT']
- ✖ – Found unexpected value(s) in ASSET_ENROLLMENT
 - Attribute ASSET_ENROLLMENT.DISPOSITION_REASON contains: ['Completed Program', 'Deceased', 'Needs Exceed Scope of Program', 'Admitted to Long Term Care (LTC) Facility', 'Moved Out of Service Area', 'No Longer Eligible', 'Refused Further Services', 'Client Withdrew from Program', 'Removed from Program', 'Unable to Contact']
- ✖ – Found unexpected value(s) in ASSET_DELIVERY
 - Attribute ASSET_DELIVERY.ASSET_UNIT contains: ['MEAL']
- ✖ – Found unexpected value(s) in PROGRAM_ENROLLMENT
 - Attribute PROGRAM_ENROLLMENT.DISPOSITION_REASON contains: ['Completed Program', 'Deceased', 'Needs Exceed Scope of Program', 'Admitted to Long Term Care (LTC) Facility', 'Moved Out of Service Area', 'No Longer Eligible', 'Refused Further Services', 'Client Withdrew from Program', 'Removed from Program', 'Unable to Contact']

Check 5 - Are any required attributes or data missing?

Only required attributes with presumed errors are listed. Optional attributes and attributes in tables with errors in Check 1 are excluded.

- ✖ – Found missing required values in ASSET_ENROLLMENT
 - Attribute ASSET_ENROLLMENT.ASSET_NAME does not exist

Check 6 - Are there any issues with referential integrity?

All tables with required foreign keys must have a matching primary key in the referenced table. Referential integrity is not checked for optional foreign keys. Tables with errors in Check 1 are excluded.

Check 6A - Are there any PATIDs contained in SDE tables that are not included in DEMOGRAPHIC?

- ✗ - Found 4212 patids in PRIVATE_ADDRESS_HISTORY that are not in DEMOGRAPHIC
- ✗ - Found 62 patids in PROGRAM_ENROLLMENT that are not in DEMOGRAPHIC
- ✗ - Found 19 patids in SESSION that are not in DEMOGRAPHIC

Check 6B - Are there any PROGRAMIDs contained in SDE tables that are not included in PROGRAM?

- ✗ - Found 576 PROGRAM IDs in PROGRAM_ENROLLMENT that are not in PROGRAM
- ✗ - Found 29 PROGRAM IDs in SESSION that are not in PROGRAM
- ✗ - Found 4 PROGRAM IDs in CURRICULUM_COMPONENT that are not in PROGRAM

Check 6C - Are there any QUESTIONNAIRE_IDs or ITEM_IDs contained in QUESTION_RESPONSE that are not included in QUESTION?

⚠ - Skipping this check because QUESTION or QUESTION_RESPONSE is not present

Data Quality Checks

Check 7 - How many rows are in each table?

The total number of rows per table should be reviewed to determine if the row counts for each table are as expected. Tables with errors in Check 1 are excluded.

⚠ - Please review closely

| | Row Count |
|-------------------------|-----------|
| DEMOGRAPHIC | 1846 |
| PRIVATE_ADDRESS_HISTORY | 5886 |
| ASSET_ENROLLMENT | 2430 |
| ASSET_DELIVERY | 6736 |
| PROGRAM | 4 |
| PROGRAM_ENROLLMENT | 664 |
| SESSION | 29 |
| CURRICULUM_COMPONENT | 4 |
| ORGANIZATION | 1 |

Check 8 - Which demographic attributes have missing values?

The total number of missing values per demographic attribute should be reviewed to determine if the level of missingness for each is as expected. Only attributes with missing values are listed. Demographic information is required to link this data to clinical data.

Reminder: Total # of rows in DEMOGRAPHIC = 1846

⚠️ - Please review closely

| Column Name | Rows missing value | Missing% |
|----------------------|--------------------|----------|
| PAT_MAIDENNAME | 1846 | 100.0 |
| PRIMARY_PHONE_TYPE | 208 | 11.3 |
| SECONDARY_PHONE_TYPE | 1648 | 89.3 |

Data Quality Checks about Programs

The following 4 checks provide output that can be assessed to determine potential data quality issues associated with programs (PROGRAMID) referenced across the PROGRAM_ENROLLMENT and SESSION tables.

Check 9 - Are the counts of unique people with program enrollments for each PROGRAM expected?

Program Enrollment Counts - Unique IDs per Program

| PROGRAMID | PATID | PROGRAM_ENROLLMENT_ID |
|-----------|-------|-----------------------|
| CON | 439 | 484 |
| CTC | 92 | 92 |
| LB | 83 | 88 |

Check 10 - Are the total counts of program enrollments and sessions for each PROGRAM expected?

Program Data Availability (PROGRAM_ENROLLMENT and SESSION tables)

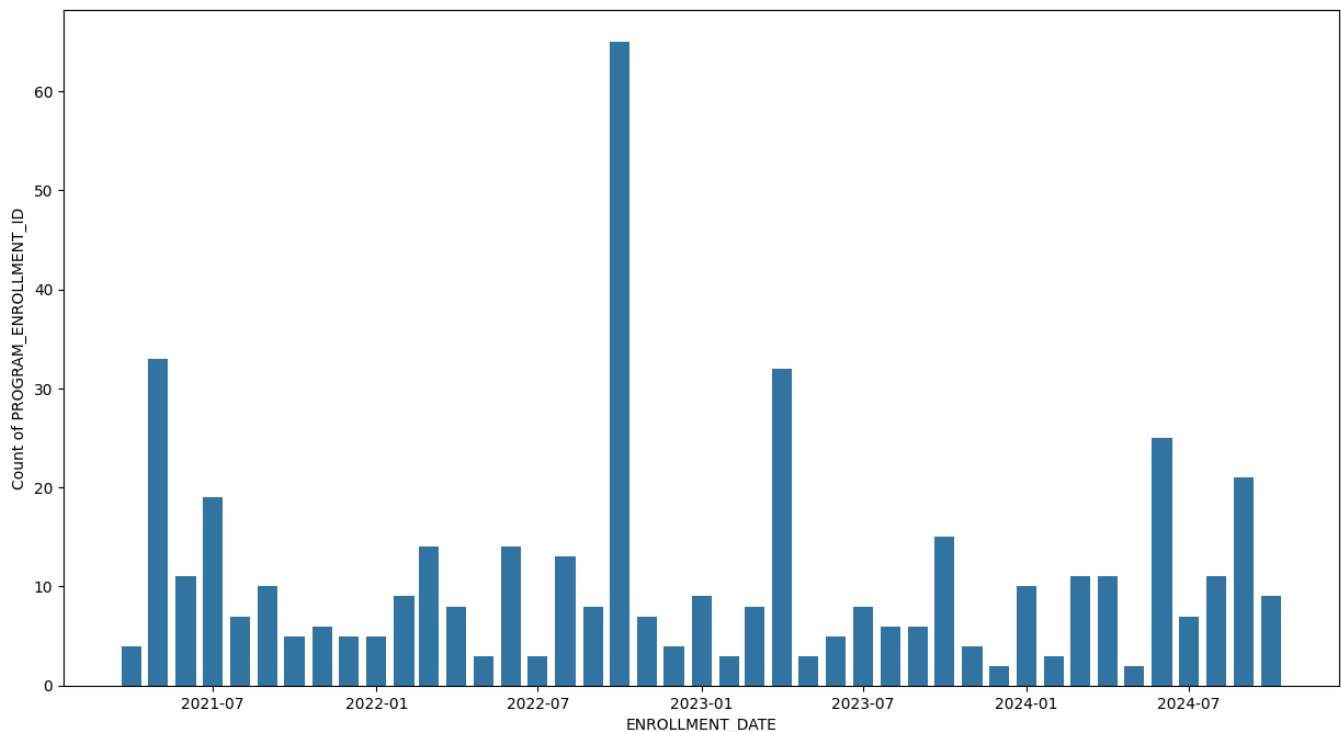
| | Total Counts | CON | CTC | LB |
|-----------------------------------|--------------|------|------|------|
| Total PATIDs with Enrollment Data | 599 | 439 | 92 | 83 |
| PATIDs with Only Enrollment Data | 574 | 439 | 67 | 83 |
| Total PATIDs with Session Data | 25 | 0 | 25 | 0 |
| PATIDs with Only Session Data | 0 | 0 | 0 | 0 |
| PATIDs with Both | 25 | 0 | 25 | 0 |
| PATIDs with Neither | 1309 | 1407 | 1816 | 1763 |

Check 11 - Are the volume of records and trends for enrollment by each PROGRAM plausible?

The following output includes a graph summarizing the total enrollments by date for each PROGRAMID. Any month-to-month enrollment changes that are greater than 25% are noted in a table titled "Large Monthly Shifts" below the graph, if detected. The enrollment trends for each program should be reviewed to determine if they are as expected.

⚠️ - Please review closely

PROGRAM_ENROLLMENT: PROGRAMID is CON

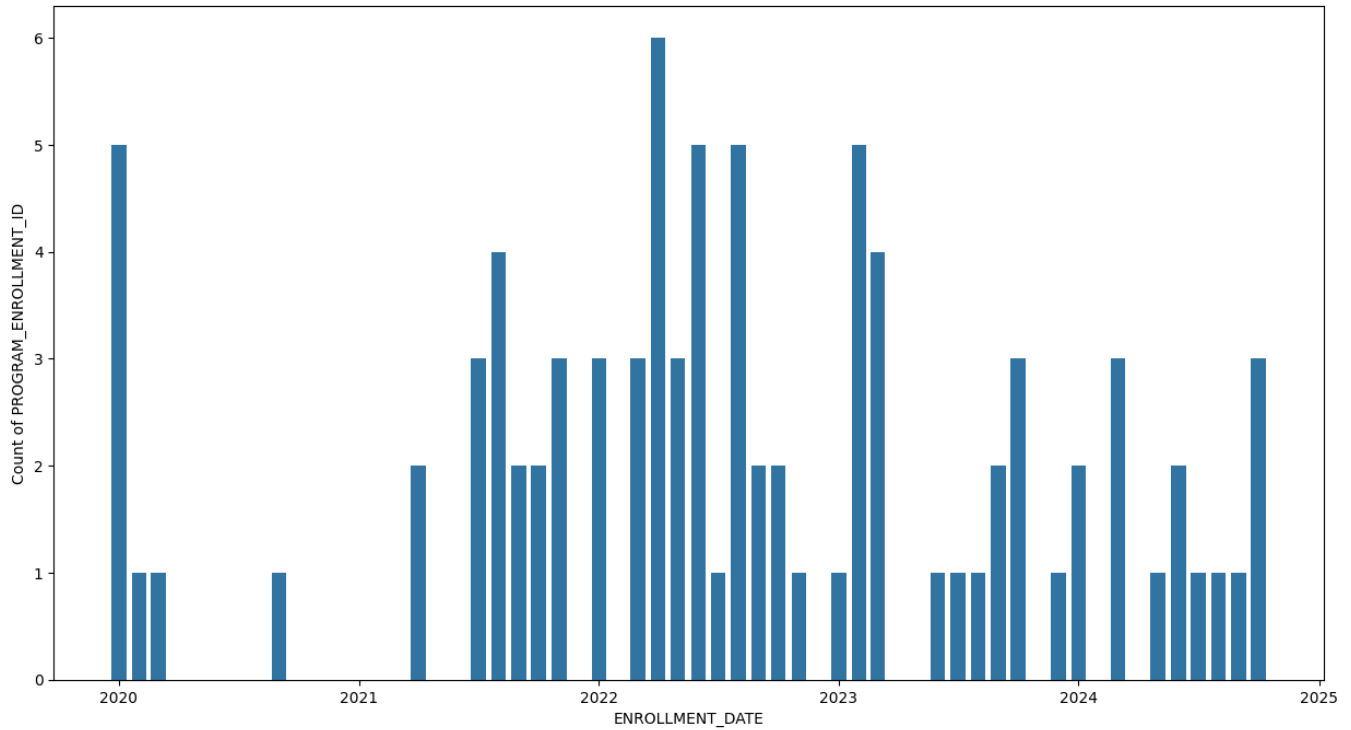


Large Monthly Shifts

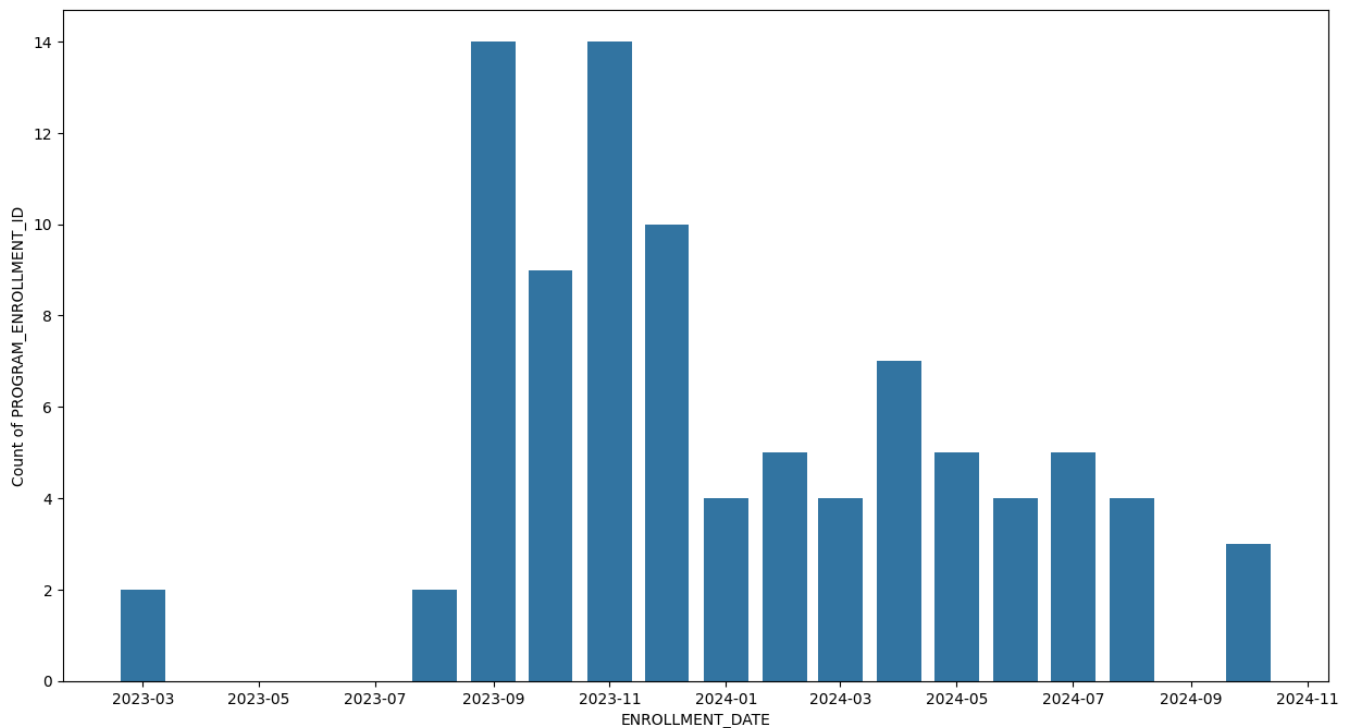
Large Monthly Shifts are changes in the value from month-to-month greater than 25%

2022-10 --> 2022-11: 65 --> 7

PROGRAM_ENROLLMENT: PROGRAMID is LB



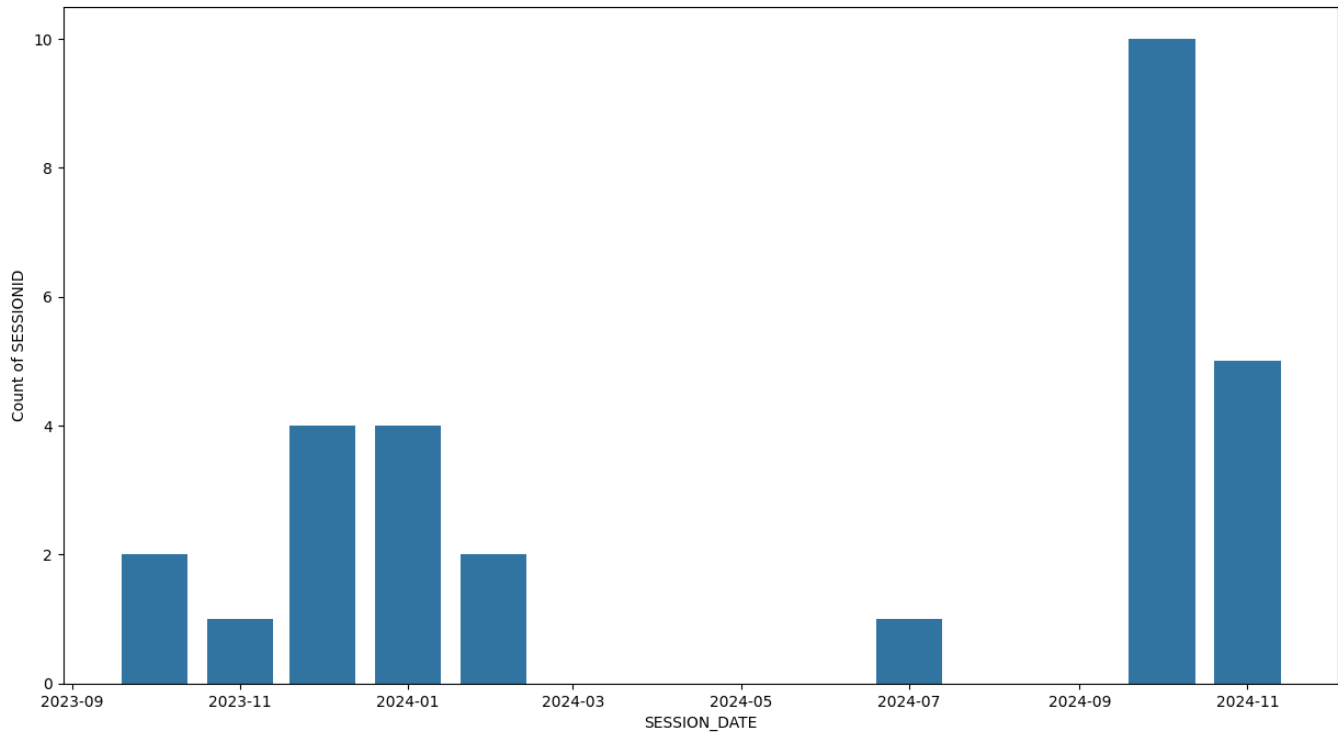
PROGRAM_ENROLLMENT: PROGRAMID is CTC



Check 12 - Are the volume of records and trends for session dates by each PROGRAM plausible?

The following output includes a graph summarizing the total participants in a session by date for each PROGRAMID. Any month-to-month session changes that are greater than 25% are noted in a table titled “Large Monthly Shifts” below the graph, if detected. The session trends for each program should be reviewed to determine if they are as expected.

SESSION: PROGRAMID is CTC



Data Quality Checks about Assets

The following 4 checks provide output that can be assessed to determine potential data quality issues associated with assets (ASSET_TYPE/ASSET_SUBTYPE) referenced across the ASSET_ENROLLMENT and ASSET_DELIVERY tables.

Check 13 - Are the counts of unique people with asset enrollments for each ASSET type/subtype expected?

Reminder: Total # of PATIDs in DEMOGRAPHIC = 1846

Asset Data Availability (ASSET_ENROLLMENT and ASSET_DELIVERY tables)

| | Total Counts | CN - FRESH | HD - SHELF STABLE | HD - FRESH | HD - FROZEN | CN - SHELF STABLE |
|-----------------------------------|-----------------|---------------|----------------------|---------------|----------------|----------------------|
| Total PATIDs with Enrollment Data | 1189 | 0 | 0 | 1187 | 1068 | 0 |
| PATIDs with Only Enrollment Data | 1048 | 0 | 0 | 1064 | 967 | 0 |
| Total PATIDs with Delivery Data | 245 | 119 | 24 | 130 | 111 | 11 |
| PATIDs with Only Delivery Data | 104 | 119 | 24 | 7 | 10 | 11 |
| PATIDs with Both | 141 | 0 | 0 | 123 | 101 | 0 |
| PATIDs with Neither | 553 | 1727 | 1822 | 652 | 768 | 1835 |

Check 14 - Are the total counts of asset enrollments and deliveries for each ASSET expected?

Asset Enrollment Counts - Unique IDs per Asset Type/Subtype

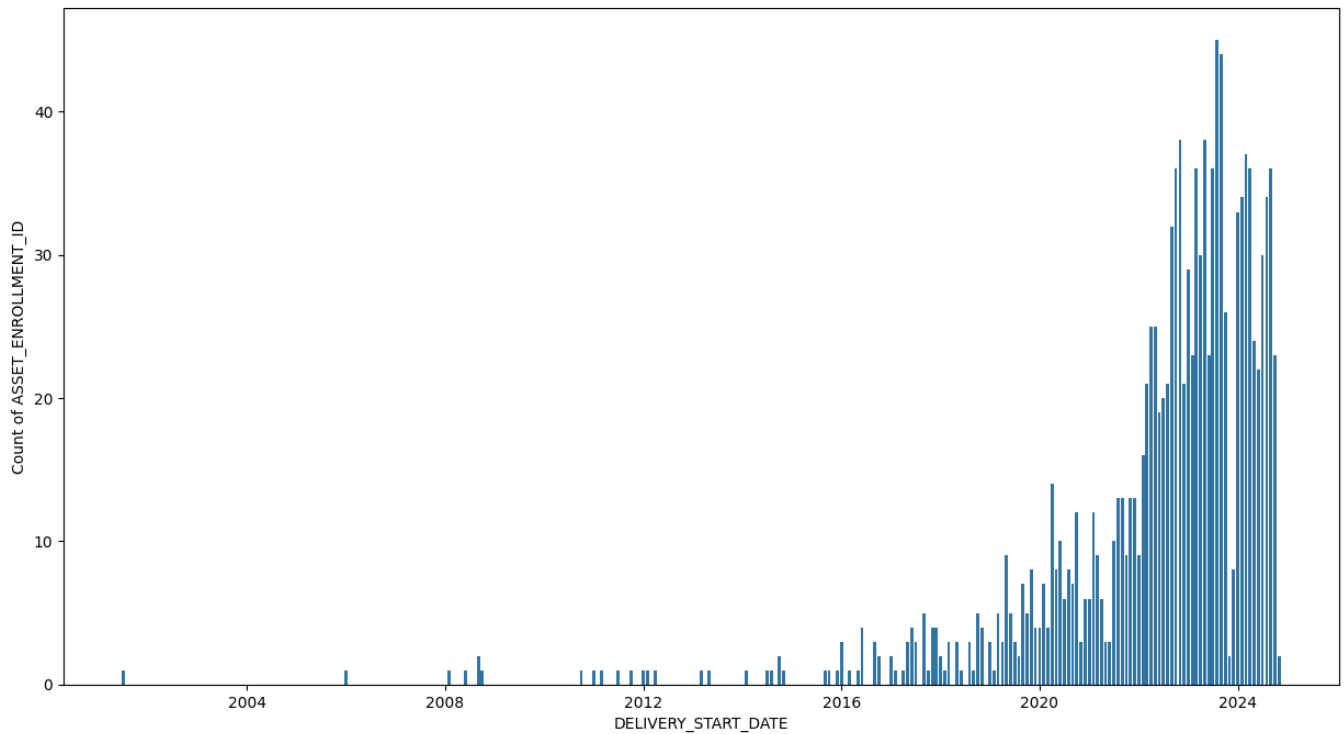
| | | PATID | ASSET_ENROLLMENT_ID |
|------------|---------------|-------|---------------------|
| ASSET_TYPE | ASSET_SUBTYPE | | |
| HD | FRESH | 1187 | 1279 |
| | FROZEN | 1068 | 1151 |

Check 15 - Are the volume of records and trends for enrollment by each ASSET type and subtype plausible?

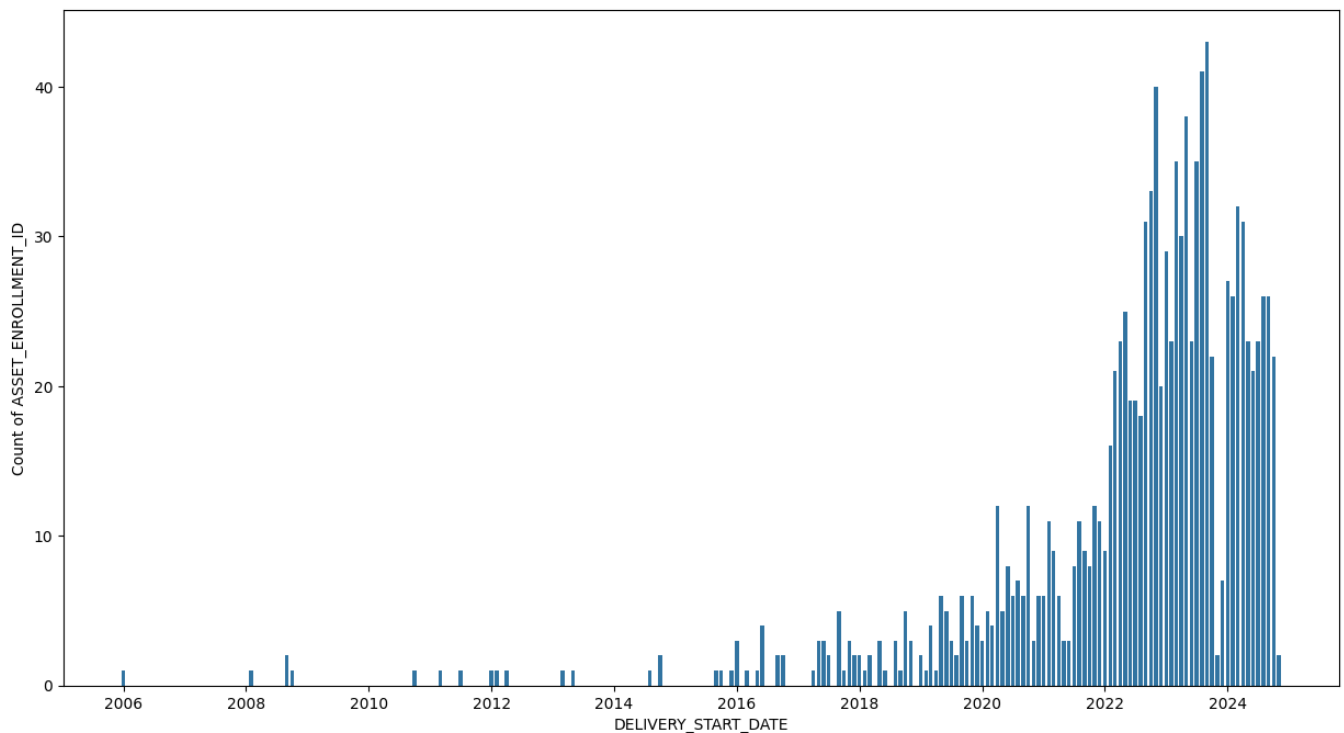
The following output includes a graph summarizing the total enrollments by date for each unique ASSET TYPE and SUBTYPE pair in the ASSET_ENROLLMENT table. Any month-to-month enrollment changes that are greater than 25% are noted in a table titled "Large Monthly Shifts" below the graph, if detected. The enrollment trends for each ASSET should be reviewed to determine if they are as expected.

 - Please review closely

ASSET_ENROLLMENT: ASSET_TYPE-ASSET_SUBTYPE is HD - FRESH



ASSET_ENROLLMENT: ASSET_TYPE-ASSET_SUBTYPE is HD - FROZEN

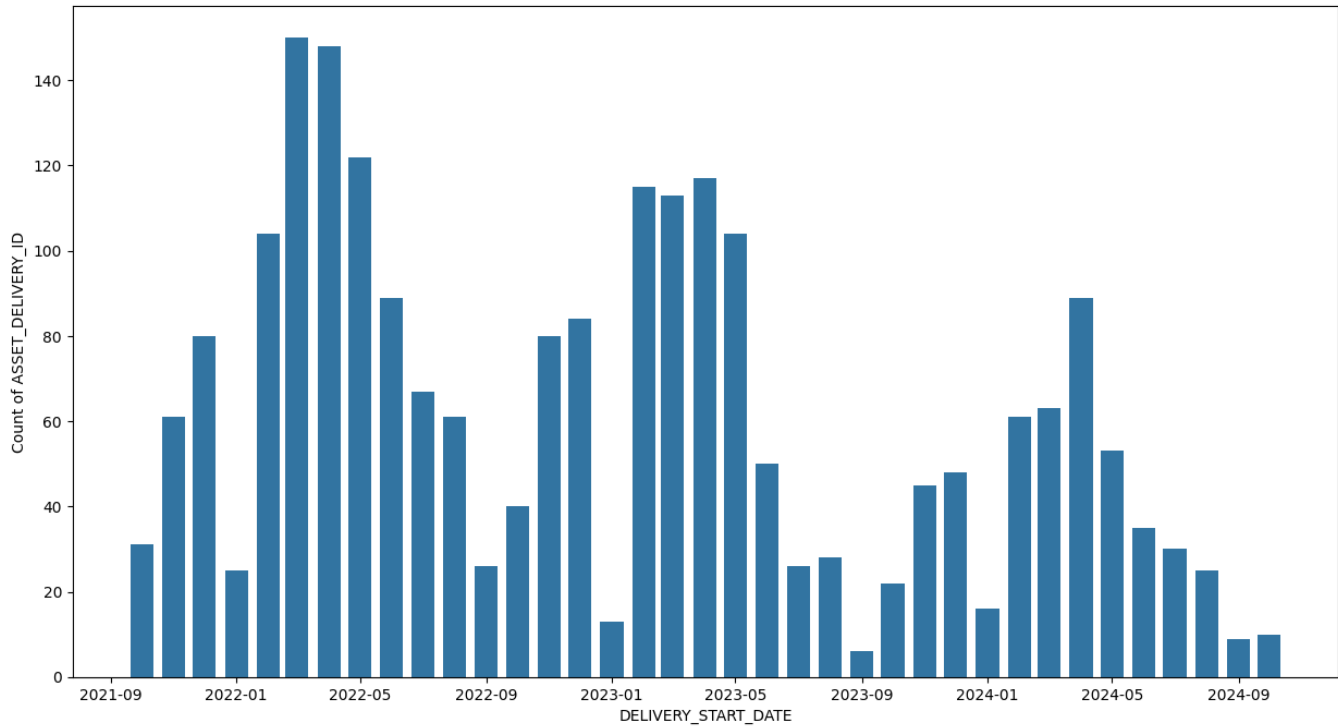


Check 16 - Are the volume of records and trends for delivery of each ASSET type and subtype plausible?

The following output includes a graph summarizing the total deliveries by date for each unique ASSET TYPE and SUBTYPE pair in the ASSET_DELIVERY table. Any month-to-month delivery changes that are greater than 25% are noted in a table titled "Large Monthly Shifts" below the graph, if detected. The delivery trends for each ASSET should be reviewed to determine if they are as expected.

⚠️ - Please review closely

ASSET_DELIVERY: ASSET_TYPE-ASSET_SUBTYPE is CN - FRESH

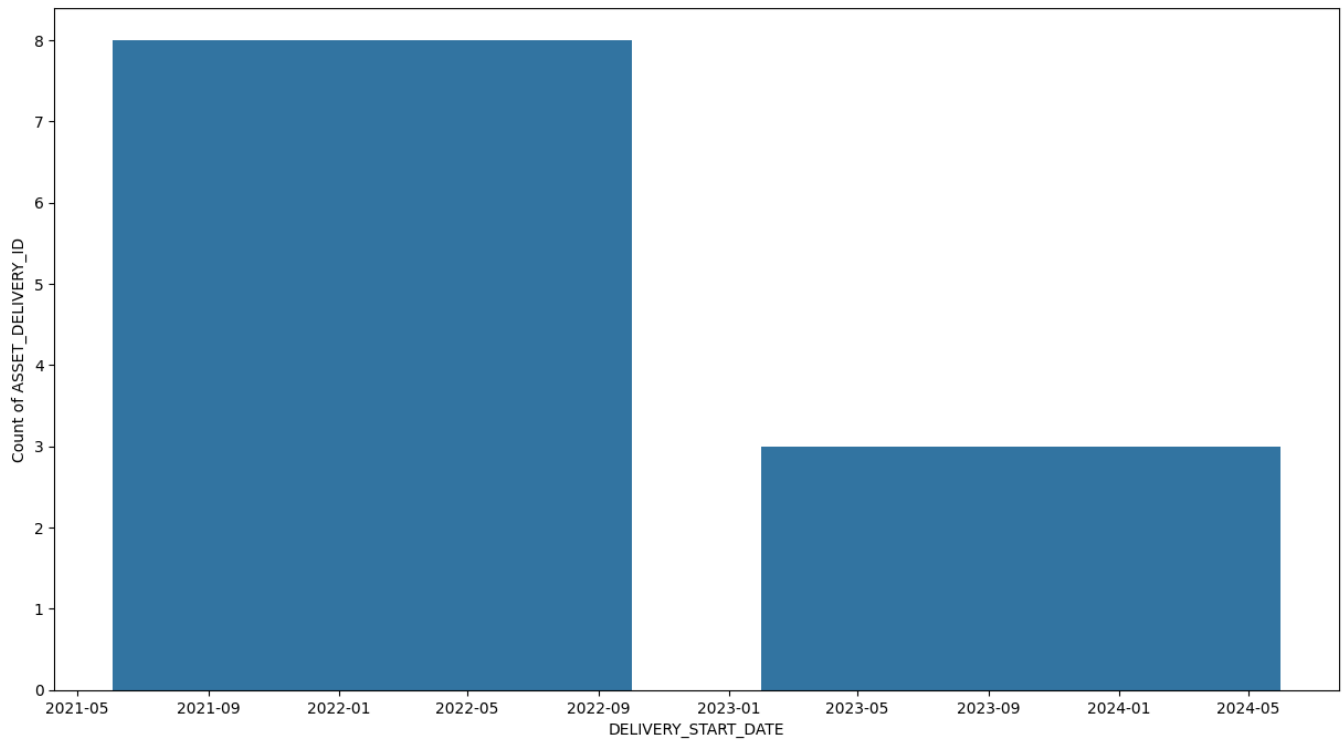


Large Monthly Shifts

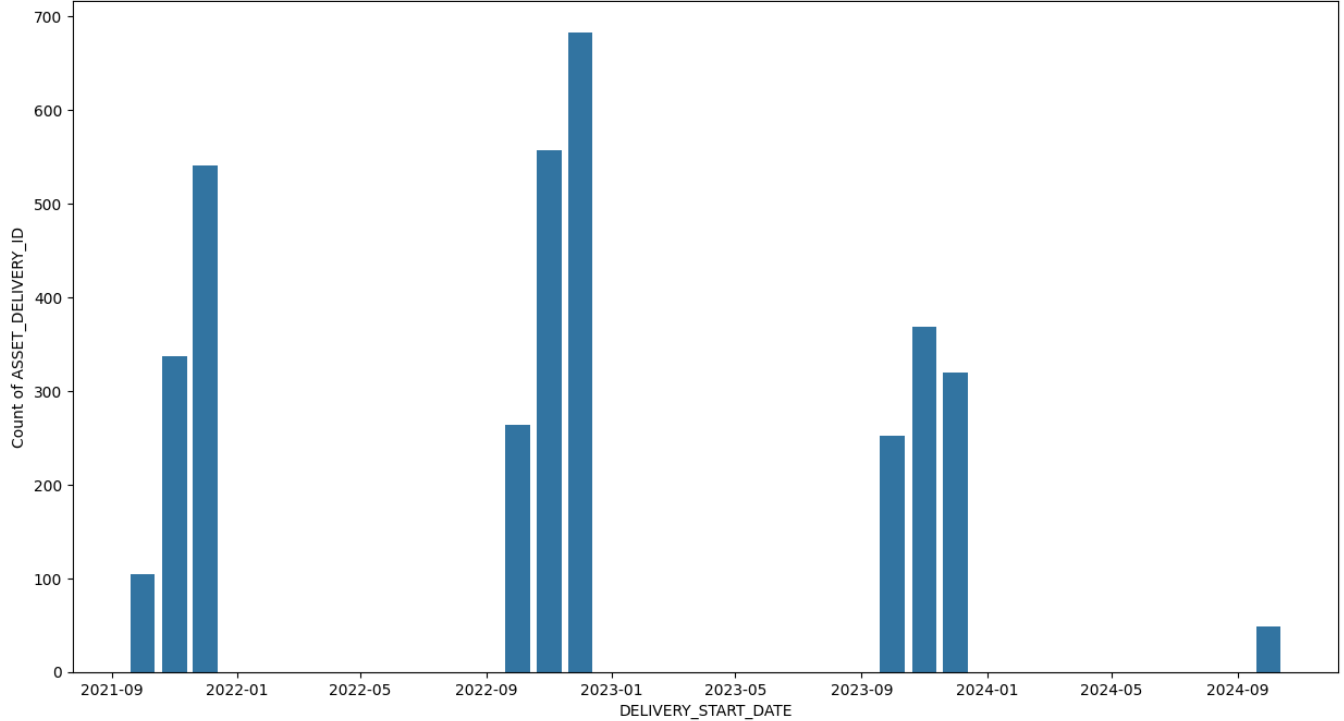
Large Monthly Shifts are changes in the value from month-to-month greater than 25%

2021-11 --> 2021-12: 61 --> 80
2021-12 --> 2022-01: 80 --> 25
2022-02 --> 2022-03: 104 --> 150
2022-05 --> 2022-06: 122 --> 89
2022-08 --> 2022-09: 61 --> 26
2022-12 --> 2023-01: 84 --> 13
2023-05 --> 2023-06: 104 --> 50
2024-03 --> 2024-04: 63 --> 89
2024-04 --> 2024-05: 89 --> 53
2024-05 --> 2024-06: 53 --> 35

ASSET_DELIVERY: ASSET_TYPE-ASSET_SUBTYPE is CN - SHELF STABLE



ASSET_DELIVERY: ASSET_TYPE-ASSET_SUBTYPE is HD - FRESH



Large Monthly Shifts

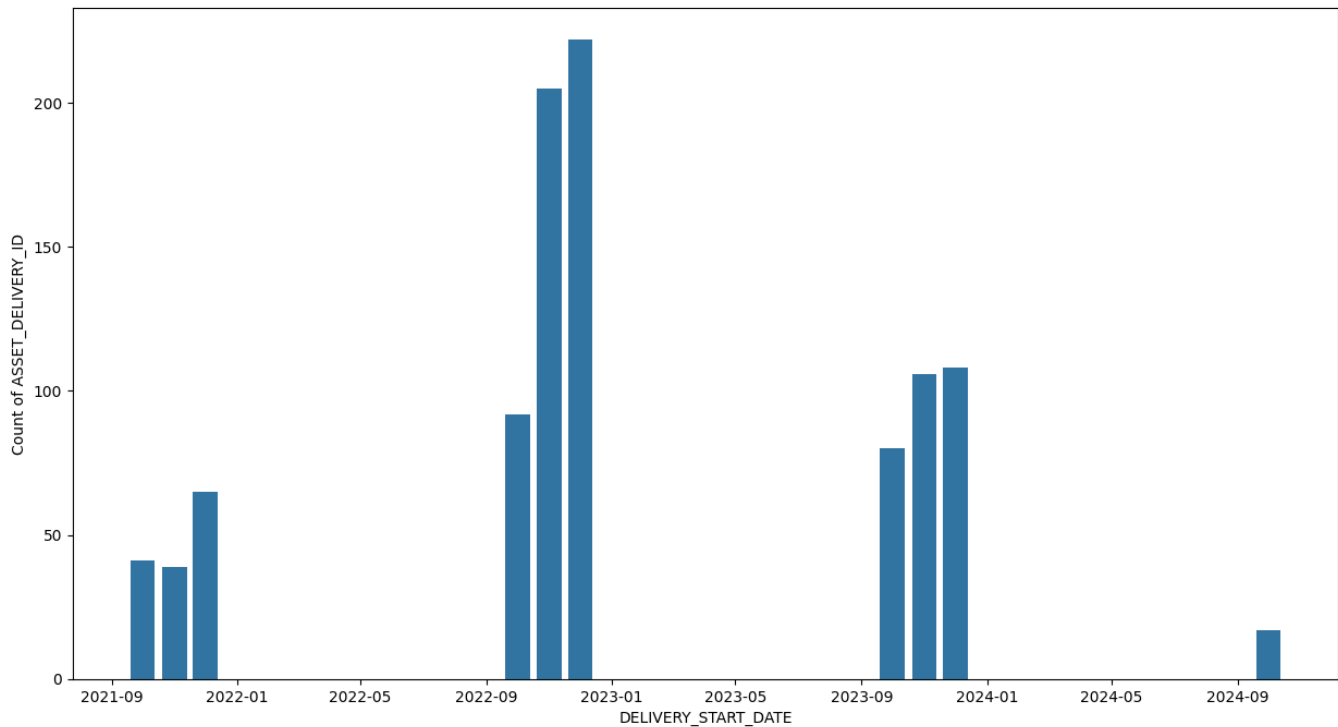
Large Monthly Shifts are changes in the value from month-to-month greater than 25%

```

2021-10 --> 2021-11: 104 --> 337
2021-11 --> 2021-12: 337 --> 541
2021-12 --> 2022-10: 541 --> 264
2022-10 --> 2022-11: 264 --> 557
2022-12 --> 2023-10: 683 --> 252
2023-10 --> 2023-11: 252 --> 369
2023-12 --> 2024-10: 320 --> 49

```

ASSET_DELIVERY: ASSET_TYPE-ASSET_SUBTYPE is HD - FROZEN



Large Monthly Shifts

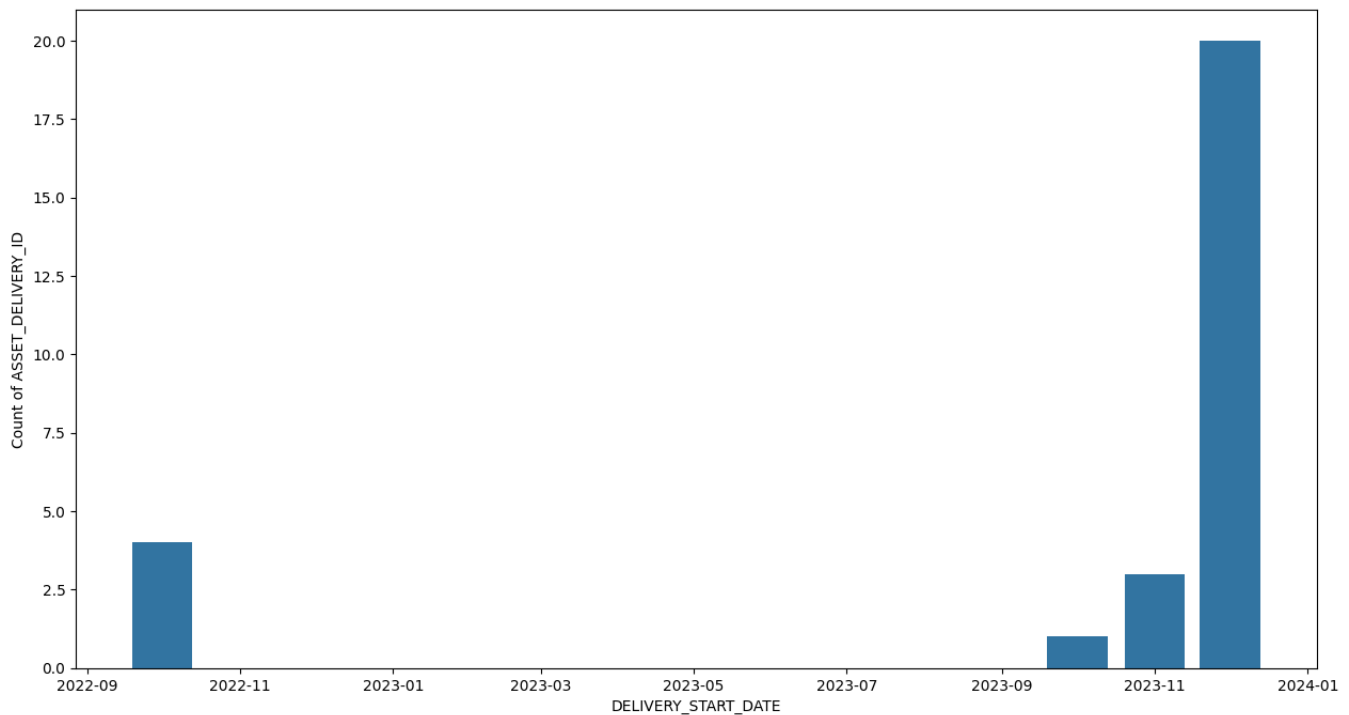
Large Monthly Shifts are changes in the value from month-to-month greater than 25%

```

2021-12 --> 2022-10: 65 --> 92
2022-10 --> 2022-11: 92 --> 205
2022-12 --> 2023-10: 222 --> 80
2023-10 --> 2023-11: 80 --> 106
2023-12 --> 2024-10: 108 --> 17

```

ASSET_DELIVERY: ASSET_TYPE-ASSET_SUBTYPE is HD - SHELF STABLE



Data Quality Checks about Referrals

Check 17 - Are the volume of records and trends of REFERRALS plausible?

The following output includes a graph summarizing the total referrals by date in the REFERRAL table. Any month-to-month referral changes that are greater than 25% are noted in a table titled "Large Monthly Shifts" below the graph, if detected. The trends should be reviewed to determine if they are as expected.

⚠️ - Please review closely

⚠️ - Skipping this check because REFERRAL does not exist

Data Quality Checks about Question Responses

Check 18 - Are the volume of records and trends of QUESTION RESPONSEs plausible?

The following output includes a graph summarizing the total question responses by date in the QUESTION_RESPONSE table. Any month-to-month changes that are greater than 25% are noted in a table titled "Large Monthly Shifts" below the graph, if detected. The trends should be reviewed to determine if they are as expected.

⚠️ - Please review closely

⚠️ - Skipping this check because QUESTION_RESPONSE does not exist