01_data_collection

August 21, 2025

1 Project Overview: Predicting Government Contract Lapse

The goal of this project is to investigate whether we can **predict the likelihood of a government contract lapsing** (not being completed or success) using structured financial data and contract-level information.

To achieve this, the project combines structured data sources with text-based contract descriptions and evaluate machine learning approaches for predictive modeling.

1.1 Project Workflow

The project is organized across four main notebooks:

1. **01_data_preprocessing**
   - Collect and clean data from **Compustat** (financials) and **USAspending** (contracts).
   - Standardize formats and merge into a unified dataset for modeling.
2. **02_data_preprocessing**
   - Perform feature engineering and preprocessing on **structured variables** (e.g., product or service code).
   - Leave out `transaction_description` for now.
   - Train baseline models (logistic regression, random forest, XGBoost) to establish performance benchmarks.
3. **03_feature_engineering_text**
   - Process the `transaction_description` field using **LLM-based classification**.
   - Integrate these features with structured data.
   - Re-test models and compare improvements.
4. **04_model_development**
   - Select the **XGBoost classifier** as the primary model (based on performance).
   - Perform **hyperparameter optimization** using Random Search, Bayesian Optimization (TPE), and Grid Search.

- Compare and analyze results with cross-validation metrics (accuracy, ROC-AUC, PR-AUC).

1.2 Expected Outcomes

- Build a reproducible pipeline that combines structured financial/contract data with text features.

- Evaluate whether machine learning models, particularly **XGBoost**, can meaningfully predict contract lapse.

- Document the impact of adding LLM-generated features (`transaction_description`) on predictive power.

[1]:

```python
# Standard library
import glob
import os
from pathlib import Path
import logging

# Third-party
import numpy as np
import pandas as pd
import statsmodels.formula.api as smf
from rapidfuzz import fuzz, process
from joblib import Parallel, delayed
```

[2]:

```python
# Configure logging
logging.basicConfig(
    level=logging.INFO,
    format="%(asctime)s | %(levelname)s | %(message)s"
)
```

1.3 Data Sources

USAspending (Contracts):
U.S. federal contract data from **USAspending.gov**.
I downloaded the official bulk CSV files organized by yearly folders, where each folder contains multiple CSV files for that fiscal year.
For this project, I will be working with 3 years of contract data: **2022, 2023, and 2024**.

Compustat (Financials):
Firm-level financial data from **WRDS Compustat**.
This dataset provides balance sheet, income statement, and market variables that can be linked with the contract data to analyze how government contracts relate to company financials.

1.4 Inspecting Raw Data Structure

The raw data is stored as:

```
|-2022-NAICS-to-SIC-Crosswalk.xlsx
|-compustat_data.csv
|-usaspending/
       2022/
           file1.csv
           file2.csv
           ...
       2023/
           file1.csv
           file2.csv
           ...
       2024/
           file1.csv
           file2.csv
           ...
```

Loading USAspending Before loading everything, I first inspect a sample file to view **all available column names**.

This allows me to manually select the most relevant columns and then load only those, which reduces memory usage and speeds up downstream processing.

[3]:

```python
# Path to 2024 folder
data_dir = Path("usaspending/2024")

# Pick one CSV file (first one in the folder)
sample_file = next(data_dir.glob("*.csv"))
print("Sample file:", sample_file)

# Read only the header row to get column names
column_titles = pd.read_csv(sample_file, nrows=0)
list(column_titles.columns)
```

```
Sample file: usaspending/2024/FY2024_All_Contracts_Full_20250807_2.csv
```

[3]:

```
['contract_transaction_unique_key',
 'contract_award_unique_key',
 'award_id_piid',
 'modification_number',
 'transaction_number',
 'parent_award_agency_id',
 'parent_award_agency_name',
 'parent_award_id_piid',
 'parent_award_modification_number',
 'federal_action_obligation',
 'total_dollars_obligated',
 'total_outlayed_amount_for_overall_award',
 'base_and_exercised_options_value',
 'current_total_value_of_award',
 'base_and_all_options_value',
 'potential_total_value_of_award',
 'disaster_emergency_fund_codes_for_overall_award',
 'outlayed_amount_from_COVID-19_supplementals_for_overall_award',
 'obligated_amount_from_COVID-19_supplementals_for_overall_award',
 'outlayed_amount_from_IIJA_supplemental_for_overall_award',
 'obligated_amount_from_IIJA_supplemental_for_overall_award',
 'action_date',
 'action_date_fiscal_year',
 'period_of_performance_start_date',
 'period_of_performance_current_end_date',
 'period_of_performance_potential_end_date',
 'ordering_period_end_date',
 'solicitation_date',
 'awarding_agency_code',
 'awarding_agency_name',
 'awarding_sub_agency_code',
 'awarding_sub_agency_name',
 'awarding_office_code',
 'awarding_office_name',
 'funding_agency_code',
 'funding_agency_name',
 'funding_sub_agency_code',
 'funding_sub_agency_name',
 'funding_office_code',
 'funding_office_name',
 'treasury_accounts_funding_this_award',
 'federal_accounts_funding_this_award',
 'object_classes_funding_this_award',
 'program_activities_funding_this_award',
 'foreign_funding',
 'foreign_funding_description',
 'sam_exception',
 'sam_exception_description',
 'recipient_uei',
 'recipient_duns',
 'recipient_name',
 'recipient_name_raw',
 'recipient_doing_business_as_name',
 'cage_code',
 'recipient_parent_uei',
 'recipient_parent_duns',
 'recipient_parent_name',
 'recipient_parent_name_raw',
 'recipient_country_code',
 'recipient_country_name',
 'recipient_address_line_1',
 'recipient_address_line_2',
 'recipient_city_name',
 'prime_award_transaction_recipient_county_fips_code',
 'recipient_county_name',
 'prime_award_transaction_recipient_state_fips_code',
 'recipient_state_code',
 'recipient_state_name',
 'recipient_zip_4_code',
 'prime_award_transaction_recipient_cd_original',
 'prime_award_transaction_recipient_cd_current',
 'recipient_phone_number',
 'recipient_fax_number',
 'primary_place_of_performance_country_code',
 'primary_place_of_performance_country_name',
 'primary_place_of_performance_city_name',
 'prime_award_transaction_place_of_performance_county_fips_code',
 'primary_place_of_performance_county_name',
 'prime_award_transaction_place_of_performance_state_fips_code',
 'primary_place_of_performance_state_code',
 'primary_place_of_performance_state_name',
 'primary_place_of_performance_zip_4',
 'prime_award_transaction_place_of_performance_cd_original',
 'prime_award_transaction_place_of_performance_cd_current',
 'award_or_idv_flag',
 'award_type_code',
 'award_type',
 'idv_type_code',
 'idv_type',
 'multiple_or_single_award_idv_code',
 'multiple_or_single_award_idv',
 'type_of_idc_code',
 'type_of_idc',
 'type_of_contract_pricing_code',
 'type_of_contract_pricing',
 'transaction_description',
 'prime_award_base_transaction_description',
 'action_type_code',
 'action_type',
 'solicitation_identifier',
 'number_of_actions',
 'inherently_governmental_functions',
 'inherently_governmental_functions_description',
 'product_or_service_code',
 'product_or_service_code_description',
 'contract_bundling_code',
 'contract_bundling',
 'dod_claimant_program_code',
 'dod_claimant_program_description',
 'naics_code',
 'naics_description',
 'recovered_materials_sustainability_code',
 'recovered_materials_sustainability',
 'domestic_or_foreign_entity_code',
 'domestic_or_foreign_entity',
 'dod_acquisition_program_code',
 'dod_acquisition_program_description',
 'information_technology_commercial_item_category_code',
 'information_technology_commercial_item_category',
 'epa_designated_product_code',
 'epa_designated_product',
 'country_of_product_or_service_origin_code',
 'country_of_product_or_service_origin',
 'place_of_manufacture_code',
 'place_of_manufacture',
 'subcontracting_plan_code',
 'subcontracting_plan',
 'extent_competed_code',
 'extent_competed',
 'solicitation_procedures_code',
 'solicitation_procedures',
 'type_of_set_aside_code',
 'type_of_set_aside',
 'evaluated_preference_code',
 'evaluated_preference',
 'research_code',
 'research',
 'fair_opportunity_limited_sources_code',
 'fair_opportunity_limited_sources',
 'other_than_full_and_open_competition_code',
 'other_than_full_and_open_competition',
 'number_of_offers_received',
 'commercial_item_acquisition_procedures_code',
 'commercial_item_acquisition_procedures',
 'small_business_competitiveness_demonstration_program',
 'simplified_procedures_for_certain_commercial_items_code',
 'simplified_procedures_for_certain_commercial_items',
 'a76_fair_act_action_code',
 'a76_fair_act_action',
 'fed_biz_opps_code',
 'fed_biz_opps',
 'local_area_set_aside_code',
 'local_area_set_aside',
 'price_evaluation_adjustment_preference_percent_difference',
 'clinger_cohen_act_planning_code',
 'clinger_cohen_act_planning',
 'materials_supplies_articles_equipment_code',
 'materials_supplies_articles_equipment',
 'labor_standards_code',
 'labor_standards',
 'construction_wage_rate_requirements_code',
 'construction_wage_rate_requirements',
 'interagency_contracting_authority_code',
 'interagency_contracting_authority',
 'other_statutory_authority',
 'program_acronym',
 'parent_award_type_code',
 'parent_award_type',
 'parent_award_single_or_multiple_code',
 'parent_award_single_or_multiple',
 'major_program',
 'national_interest_action_code',
 'national_interest_action',
 'cost_or_pricing_data_code',
 'cost_or_pricing_data',
 'cost_accounting_standards_clause_code',
 'cost_accounting_standards_clause',
 'government_furnished_property_code',
 'government_furnished_property',
 'sea_transportation_code',
 'sea_transportation',
 'undefinitized_action_code',
 'undefinitized_action',
 'consolidated_contract_code',
 'consolidated_contract',
 'performance_based_service_acquisition_code',
 'performance_based_service_acquisition',
 'multi_year_contract_code',
 'multi_year_contract',
 'contract_financing_code',
 'contract_financing',
 'purchase_card_as_payment_method_code',
 'purchase_card_as_payment_method',
 'contingency_humanitarian_or_peacekeeping_operation_code',
 'contingency_humanitarian_or_peacekeeping_operation',
 'alaskan_native_corporation_owned_firm',
 'american_indian_owned_business',
 'indian_tribe_federally_recognized',
 'native_hawaiian_organization_owned_firm',
 'tribally_owned_firm',
 'veteran_owned_business',
 'service_disabled_veteran_owned_business',
 'woman_owned_business',
 'women_owned_small_business',
 'economically_disadvantaged_women_owned_small_business',
 'joint_venture_women_owned_small_business',
 'joint_venture_economic_disadvantaged_women_owned_small_bus',
 'minority_owned_business',
 'subcontinent_asian_asian_indian_american_owned_business',
 'asian_pacific_american_owned_business',
 'black_american_owned_business',
 'hispanic_american_owned_business',
 'native_american_owned_business',
 'other_minority_owned_business',
 'contracting_officers_determination_of_business_size',
 'contracting_officers_determination_of_business_size_code',
 'emerging_small_business',
 'community_developed_corporation_owned_firm',
 'labor_surplus_area_firm',
 'us_federal_government',
 'federally_funded_research_and_development_corp',
 'federal_agency',
 'us_state_government',
 'us_local_government',
 'city_local_government',
 'county_local_government',
 'inter_municipal_local_government',
 'local_government_owned',
 'municipality_local_government',
 'school_district_local_government',
 'township_local_government',
 'us_tribal_government',
 'foreign_government',
 'organizational_type',
 'corporate_entity_not_tax_exempt',
 'corporate_entity_tax_exempt',
 'partnership_or_limited_liability_partnership',
 'sole_proprietorship',
 'small_agricultural_cooperative',
 'international_organization',
 'us_government_entity',
 'community_development_corporation',
 'domestic_shelter',
 'educational_institution',
 'foundation',
 'hospital_flag',
 'manufacturer_of_goods',
 'veterinary_hospital',
 'hispanic_servicing_institution',
 'receives_contracts',
 'receives_financial_assistance',
 'receives_contracts_and_financial_assistance',
 'airport_authority',
 'council_of_governments',
 'housing_authorities_public_tribal',
 'interstate_entity',
 'planning_commission',
 'port_authority',
 'transit_authority',
 'subchapter_scorporation',
 'limited_liability_corporation',
 'foreign_owned',
 'for_profit_organization',
 'nonprofit_organization',
 'other_not_for_profit_organization',
 'the_ability_one_program',
 'private_university_or_college',
 'state_controlled_institution_of_higher_learning',
 '1862_land_grant_college',
 '1890_land_grant_college',
 '1994_land_grant_college',
 'minority_institution',
 'historically_black_college',
 'tribal_college',
 'alaskan_native_servicing_institution',
 'native_hawaiian_servicing_institution',
 'school_of_forestry',
 'veterinary_college',
 'dot_certified_disadvantage',
 'self_certified_small_disadvantaged_business',
 'small_disadvantaged_business',
 'c8a_program_participant',
 'historically_underutilized_business_zone_hubzone_firm',
 'sba_certified_8a_joint_venture',
 'highly_compensated_officer_1_name',
 'highly_compensated_officer_1_amount',
 'highly_compensated_officer_2_name',
 'highly_compensated_officer_2_amount',
 'highly_compensated_officer_3_name',
 'highly_compensated_officer_3_amount',
 'highly_compensated_officer_4_name',
 'highly_compensated_officer_4_amount',
 'highly_compensated_officer_5_name',
 'highly_compensated_officer_5_amount',
 'usaspending_permalink',
 'initial_report_date',
 'last_modified_date']
```

[4]:

```
# Selected columns
selected_columns = [
    # --- Identifiers ---
    'contract_transaction_unique_key',
    'contract_award_unique_key',
    'award_id_piid',

    # --- Financials ---
    'federal_action_obligation',
    'total_dollars_obligated',
    'current_total_value_of_award',
    'potential_total_value_of_award',

    # --- Dates ---
    'action_date',
    'action_date_fiscal_year',
    'period_of_performance_start_date',
    'period_of_performance_current_end_date',
    'period_of_performance_potential_end_date',

    # --- Agencies ---
    'awarding_agency_code',
    'funding_agency_code',

    # --- Recipient attributes ---
    'recipient_country_name',

    # --- Place of performance ---
    'primary_place_of_performance_country_name',

    # --- Award characteristics ---
    'award_type',
    'idv_type',
    'type_of_contract_pricing',
    'extent_competed',
    'type_of_set_aside',
    'number_of_offers_received',

    # --- Text fields for LLM parsing ---
    'transaction_description',
    # --- Categorization ---
    'product_or_service_code',
    'naics_code', 'naics_description',

    # --- High-signal risk flags ---
    'national_interest_action',
    'undefinitized_action',
    'multi_year_contract',
    'performance_based_service_acquisition',
    'contract_financing',
    'government_furnished_property',

    # --- Socio-economic ownership (collapsed set) ---
    'veteran_owned_business',
    'woman_owned_business',
    'minority_owned_business',
    'small_disadvantaged_business',

    # --- Business size ---
    'contracting_officers_determination_of_business_size',

    # --- Org type (condensed) ---
    'for_profit_organization',
    'nonprofit_organization',
    'foreign_owned',
    'us_state_government',
    'us_local_government',
    'us_tribal_government',
    'foreign_government',

    # --- Institutions ---
    'educational_institution',
    'historically_black_college',
    'tribal_college',
    'hospital_flag',

    # --- Special program ---
    'the_ability_one_program',

    # --- For Compustat merge ---
    'recipient_parent_name',
    'recipient_name',
]

print(len(selected_columns))
```

```
51
```

[5]:

```
DATA_DIR = "usaspending"
YEARS = {"2022", "2023", "2024"}

frames = []

# Loop through each subfolder inside usaspending
for year in sorted(os.listdir(DATA_DIR)):
    if year not in YEARS:
        continue

    folder_path = os.path.join(DATA_DIR, year)
    if not os.path.isdir(folder_path):
        continue

    print(f"Getting files from {year}...")

    for f in glob.glob(os.path.join(folder_path, "*.csv")):
        try:
            df = pd.read_csv(
                f, usecols=lambda c: c in selected_columns, low_memory=False
            )
            frames.append(df)
        except Exception as e:
            print(f"Error reading {f}: {e}")

df_usaspending = pd.concat(frames, ignore_index=True)
print("Final shape:", df_usaspending.shape)
```

```
Getting files from 2022…
Getting files from 2023…
Getting files from 2024…
Final shape: (20041177, 51)
```

[6]:

```
# Inspect missingness
nan_summary = (
    df_usaspending.isna().mean().sort_values(ascending=False) * 100
)
print(nan_summary)
```

```
idv_type                                              96.226180
type_of_set_aside                                     78.833993
number_of_offers_received                             67.876632
contract_financing                                    34.399018
multi_year_contract                                   30.844975
national_interest_action                              23.643996
award_type                                             3.773820
primary_place_of_performance_country_name              3.773820
current_total_value_of_award                           3.773820
```

```
period_of_performance_current_end_date                  3.773820
period_of_performance_potential_end_date                3.773820
undefinitized_action                                    1.349347
extent_competed                                         0.312856
recipient_parent_name                                   0.019495
transaction_description                                 0.006402
recipient_country_name                                  0.003817
type_of_contract_pricing                                0.003413
naics_description                                       0.003019
naics_code                                              0.003009
product_or_service_code                                 0.000529
recipient_name                                          0.000195
period_of_performance_start_date                        0.000185
contracting_officers_determination_of_business_size     0.000095
minority_owned_business                                 0.000000
hospital_flag                                           0.000000
foreign_government                                      0.000000
foreign_owned                                           0.000000
for_profit_organization                                 0.000000
nonprofit_organization                                  0.000000
us_tribal_government                                    0.000000
the_ability_one_program                                 0.000000
us_local_government                                     0.000000
historically_black_college                              0.000000
tribal_college                                          0.000000
us_state_government                                     0.000000
educational_institution                                 0.000000
contract_transaction_unique_key                         0.000000
woman_owned_business                                    0.000000
veteran_owned_business                                  0.000000
performance_based_service_acquisition                   0.000000
government_furnished_property                           0.000000
contract_award_unique_key                               0.000000
funding_agency_code                                     0.000000
awarding_agency_code                                    0.000000
action_date_fiscal_year                                 0.000000
action_date                                             0.000000
potential_total_value_of_award                          0.000000
total_dollars_obligated                                 0.000000
federal_action_obligation                               0.000000
award_id_piid                                           0.000000
small_disadvantaged_business                            0.000000
dtype: float64
```

Observation: Columns with extremely high missingness are unlikely to provide stable signals and may complicate modeling.

Based on this observation, decided remove any column with **more than 20% missing values**.

After dropping sparse columns, there are still some NaNs. Since the proportion is small, I remove all rows containing any NaNs.

[7]:

```
# Drop columns with more than 20% NaN
thresh = 0.80 * len(df_usaspending)
df_usaspending.dropna(axis=1, thresh=thresh, inplace=True)

# Drop rows with any remaining NaN
df_usaspending.dropna(axis=0, how="any", inplace=True)

df_usaspending
```

[7]:

```
                            contract_transaction_unique_key  \
0                      3600_-NONE-_36C10B22C0009_0_-NONE-_0
1                 9700_9700_SPE2DV22F78C7_0_SPE2DV17D4001_0
2                 9700_9700_SPE3SU22F3BRL_0_SPE30018DS320_0
3                 9700_9700_SPE30022F10X5_0_SPE30022D3336_0
4         7012_7001_70CTD018FR0000009_P00001_HSHQDC12D00…
…                                                       …
20041171          4732_4732_47QSCC24F0RNH_0_47QSCC23D0004_0
20041172               9700_-NONE-_SPE4A624V2460_0_-NONE-_0
20041173  1205_7529_12314421F0478_P00009_HHSN31620120009…
20041175          9700_9700_SPE3SU24F3JFB_0_SPE30020DS351_0
20041176          9700_9700_SPE30024FGL99_0_SPE30022DV015_0

                                  contract_award_unique_key  \
0                 CONT_AWD_36C10B22C0009_3600_-NONE-_-NONE-
1            CONT_AWD_SPE2DV22F78C7_9700_SPE2DV17D4001_9700
2            CONT_AWD_SPE3SU22F3BRL_9700_SPE30018DS320_9700
3            CONT_AWD_SPE30022F10X5_9700_SPE30022D3336_9700
4         CONT_AWD_70CTD018FR0000009_7012_HSHQDC12D00014…
…                                                       …
20041171     CONT_AWD_47QSCC24F0RNH_4732_47QSCC23D0004_4732
20041172          CONT_AWD_SPE4A624V2460_9700_-NONE-_-NONE-
20041173  CONT_AWD_12314421F0478_1205_HHSN316201200098W_…
20041175     CONT_AWD_SPE3SU24F3JFB_9700_SPE30020DS351_9700
20041176     CONT_AWD_SPE30024FGL99_9700_SPE30022DV015_9700

              award_id_piid  federal_action_obligation  \
0             36C10B22C0009                   20450.00
1             SPE2DV22F78C7                      38.60
2             SPE3SU22F3BRL                     701.38
3             SPE30022F10X5                   22029.85
4         70CTD018FR0000009                       0.00
…                         …                          …
20041171      47QSCC24F0RNH                     808.40
20041172      SPE4A624V2460                    1679.92
```

```
20041173      12314421F0478                      -0.01
20041175      SPE3SU24F3JFB                    1039.68
20041176      SPE30024FGL99                     405.00

          total_dollars_obligated  current_total_value_of_award  \
0                        20450.00                      20450.00
1                           38.60                         38.60
2                          701.38                        701.38
3                        22029.85                      22029.85
4                        47780.13                      47780.13
…                           …                           …
20041171                   808.40                        808.40
20041172                  1679.92                       1679.92
20041173               6339843.14                    6339843.14
20041175                  1039.68                       1039.68
20041176                   405.00                        405.00

          potential_total_value_of_award action_date  action_date_fiscal_year  \
0                               82100.00  2021-11-10                     2022
1                                  38.60  2021-11-10                     2022
2                                 701.38  2021-11-10                     2022
3                               22029.85  2021-11-10                     2022
4                               47780.13  2021-11-10                     2022
…                                  …         …                      …
20041171                          808.40  2023-11-13                     2024
20041172                         1679.92  2023-11-13                     2024
20041173                      6339843.14  2023-11-13                     2024
20041175                         1039.68  2023-11-13                     2024
20041176                          405.00  2023-11-13                     2024

         period_of_performance_start_date  … foreign_government  \
0                              2021-11-02  …                  f
1                              2021-11-10  …                  f
2                              2021-11-10  …                  f
3                              2021-11-10  …                  f
4                              2018-01-01  …                  f
…                                   …  …                …
20041171                       2023-11-13  …                  f
20041172                       2023-11-13  …                  f
20041173                       2021-07-08  …                  f
20041175                       2023-11-13  …                  f
20041176                       2023-11-13  …                  f

         educational_institution  hospital_flag  foreign_owned  \
0                              f              f              f
1                              f              f              f
2                              f              f              f
```

```
3                              f              f              f
4                              f              f              f
…                          …            …            …
20041171                       f              f              f
20041172                       f              f              f
20041173                       f              f              f
20041175                       f              f              f
20041176                       f              f              f

         for_profit_organization nonprofit_organization  \
0                              t                      f
1                              t                      f
2                              t                      f
3                              t                      f
4                              t                      f
…                          …                    …
20041171                       t                      f
20041172                       t                      f
20041173                       t                      f
20041175                       t                      f
20041176                       t                      f

         the_ability_one_program historically_black_college tribal_college  \
0                              f                          f              f
1                              f                          f              f
2                              f                          f              f
3                              f                          f              f
4                              f                          f              f
…                          …                        …            …
20041171                       f                          f              f
20041172                       f                          f              f
20041173                       f                          f              f
20041175                       f                          f              f
20041176                       f                          f              f

         small_disadvantaged_business
0                                   f
1                                   f
2                                   f
3                                   f
4                                   f
…                               …
20041171                            f
20041172                            f
20041173                            t
20041175                            f
20041176                            f
```

```
[19278683 rows x 45 columns]
```

Industry Code Alignment (NAICS → SIC) The USAspending dataset uses **NAICS** (North American Industry Classification System) codes, while Compustat uses **SIC** (Standard Industrial Classification) codes.

Since I need to merge contract data (USAspending) with firm financials (Compustat), I first convert NAICS to SIC.

This ensures both datasets use a consistent industry classification scheme.

[8]:
```python
# Load crosswalk file
df_cross = pd.read_excel('2022-NAICS-to-SIC-Crosswalk.xlsx')
df_cross
```

[8]:
```
      Input Seq 1  2022 NAICS Code  \
0               1           111110
1               2           111120
2               3           111130
3               4           111140
4               5           111150
…           …              …
2345         2346           926140
2346         2347           926150
2347         2348           927110
2348         2349           928110
2349         2350           928120

                                       2022 NAICS Title Related SIC Code  \
0                                       Soybean Farming              116
1                     Oilseed (except Soybean) Farming               119
2                             Dry Pea and Bean Farming               119
3                                         Wheat Farming              111
4                                         Corn Farming               115
…                                                 …              …
2345  Regulation of Agricultural Marketing and Commo…             9641
2346  Regulation, Licensing, and Inspection of Misce…             9651
2347                     Space Research and Technology              9661
2348                                 National Security              9711
2349                             International Affairs              9721

                      Related SIC Code Description Change to 2017 Code  \
0                                         Soybeans           No Change
1                                 Cash Grains, Nec           No Change
2                                 Cash Grains, Nec           No Change
3                                            Wheat           No Change
4                                             Corn           No Change
```

```
…                                            …                 …
2345          Regulation of Agricultural Marketing           No Change
2346  Regulation, Miscellaneous Commercial Sectors           No Change
2347                 Space Research and Technology           No Change
2348                             National Security           No Change
2349                         International Affairs           No Change

     2017 NAICS Code                                  2017 NAICS Code.1
0             111110                                    Soybean Farming
1             111120                 Oilseed (except Soybean) Farming
2             111130                         Dry Pea and Bean Farming
3             111140                                      Wheat Farming
4             111150                                     Corn Farming
…              …                                                …
2345          926140  Regulation of Agricultural Marketing and Commo…
2346          926150  Regulation, Licensing, and Inspection of Misce…
2347          927110                   Space Research and Technology
2348          928110                               National Security
2349          928120                           International Affairs

[2350 rows x 8 columns]
```

[9]:

```
# ---------- helper ----------
def build_naics_to_sic_dicts(
    df_cross,
    naics_col="2022 NAICS Code",
    sic_col="Related SIC Code",
    sic_desc_col="Related SIC Code Description"):

    """
    Build dictionaries mapping NAICS codes to SIC codes and descriptions.

    Parameters
    ----------
    df_cross : pd.DataFrame
        Crosswalk dataframe.
    naics_col : str
        Column name for NAICS codes.
    sic_col : str
        Column name for SIC codes.
    sic_desc_col : str
        Column name for SIC descriptions.

    Returns
    -------
    tuple of dict
        (map_code, map_desc)
    """

    logging.info("Building maps for %s ...", naics_col)

    # Restrict to needed cols and non-null rows
    cols = [naics_col, sic_col, sic_desc_col]
    tmp = df_cross[cols].dropna(subset=cols)
    logging.info("Rows after dropping NaN: %s", f"{len(tmp):,}")

    # Fast counts of (NAICS, SIC, DESC)
    counts = (
        tmp.value_counts(subset=cols)
        .reset_index(name="cnt")
    )
    logging.info("Unique (NAICS, SIC, Desc) combos: %s", f"{len(counts):,}")

    # Pick argmax SIC per NAICS
    counts = counts.sort_values(
        [naics_col, "cnt", sic_col],
        ascending=[True, False, True],
        kind="stable"
    )

    idx = counts.groupby(naics_col)["cnt"].idxmax()
    winners = counts.loc[idx, [naics_col, sic_col, sic_desc_col]]
    logging.info("Winners selected: %s", f"{len(winners):,}")

    # Build two small dicts
    map_code = dict(zip(winners[naics_col], winners[sic_col]))
    map_desc = dict(zip(winners[naics_col], winners[sic_desc_col]))

    logging.info("Maps built (%s NAICS mapped)", f"{len(map_code):,}")
    return map_code, map_desc


# ---------- Build map (2022 only) ----------
map22_code, map22_desc = build_naics_to_sic_dicts(df_cross)

# Ensure new cols exist
for col in ["sic4", "sic_desc"]:
    if col not in df_usaspending.columns:
        df_usaspending[col] = pd.Series(dtype="object")

# Only map FY>=2022
mask = df_usaspending["action_date_fiscal_year"].ge(2022)
```

```
df_usaspending.loc[mask, "sic4"] = df_usaspending.loc[mask, "naics_code"].
 ↪map(map22_code)
df_usaspending.loc[mask, "sic_desc"] = df_usaspending.loc[mask, "naics_code"].
 ↪map(map22_desc)

print("Mapping complete!")
```

```
2025-08-21 17:41:11,385 | INFO | Building maps for 2022 NAICS Code …
2025-08-21 17:41:11,387 | INFO | Rows after dropping NaN: 2,347
2025-08-21 17:41:11,394 | INFO | Unique (NAICS, SIC, Desc) combos: 2,153
2025-08-21 17:41:11,398 | INFO | Winners selected: 1,009
2025-08-21 17:41:11,399 | INFO | Maps built (1,009 NAICS mapped)
```

```
Mapping complete!
```

Loading Compustat (WRDS) Compustat provides standardized financial statement and market data at the firm level.
I collected this dataset from **WRDS (Wharton Research Data Services)**, which is the official access point for Compustat.

[10]:

```
df_comp = pd.read_csv('compustat_data.csv')

# Apply Standard Compustat filters
df_comp = df_comp[
    (df_comp["pddur"] == 12) &          # Annual (12 months)
    (df_comp["indfmt"] == "INDL") &    # Industrial format
    (df_comp["datafmt"] == "STD") &    # Standard format
    (df_comp["popsrc"] == "D") &       # Domestic population source
    (df_comp["consol"] == "C")         # Consolidated
].drop_duplicates()
```

```
/var/folders/m4/8tn_t7fn3n999rq0tn3djx1w0000gn/T/ipykernel_7575/3202400631.py:1:
DtypeWarning: Columns (10,12,16,17,26,30,33,34,35,920,946,947) have mixed types.
Specify dtype option on import or set low_memory=False.
  df_comp = pd.read_csv('compustat_data.csv')
```

[11]:

```
# Ensure datadate is datetime
df_comp["datadate"] = pd.to_datetime(df_comp["datadate"], errors="coerce")

# Federal fiscal year: Oct-Sep
df_comp["federal_fy"] = np.where(
    df_comp["datadate"].dt.month >= 10,
    df_comp["datadate"].dt.year + 1,    # Oct-Dec → next year
    df_comp["datadate"].dt.year         # Jan-Sep → same year
).astype(int)

# Keep only 2022-2024 federal fiscal years (same as USAspending)
df_comp = df_comp[df_comp["federal_fy"].between(2022, 2024)]
```

Merging Compustat with USAspending For this project, the goal is to align **company financials (Compustat)** with **federal contract activity (USAspending)**.
Since Compustat only contains data on **publicly listed companies**, will restrict the USAspending dataset to include only those awards that were won by **public companies**.

This ensures that every contract observation can be matched to firm-level financial data, allowing us to study how government contracts relate to company performance.

[12]:

```
# --- Align SIC codes between USAspending and Compustat ---

# Ensure Compustat's 'sich' exists
df_comp["sic4"] = (
    df_comp["sich"]
    .astype(str)                # make string
    .str.zfill(4)               # pad to 4 digits
    .str.strip()                # remove extra spaces
)

# Normalize USAspending SIC as string 4-digit
df_usaspending["sic4"] = (
    df_usaspending["sic4"]
    .astype(str)
    .str.zfill(4)
    .str.strip()
)
```

[13]:

```
# Remove bogus codes ("0nan", "nan", empty, None)
df_comp = df_comp[~df_comp["sic4"].isin(["0nan", "nan", "NaN", "None", ""])]

df_usaspending = df_usaspending[~df_usaspending["sic4"].isin(["0nan", "nan",␣
 ↪"NaN", "None", ""])]
```

[14]:

```
# Keep only sic4 values present in Compustat
valid_sic4 = df_comp["sic4"].dropna().unique()
df_usaspending = df_usaspending[df_usaspending["sic4"].isin(valid_sic4)].copy()
df_usaspending
```

[14]:

```
                             contract_transaction_unique_key  \
1                  9700_9700_SPE2DV22F78C7_0_SPE2DV17D4001_0
3                  9700_9700_SPE30022F10X5_0_SPE30022D3336_0
6                  9700_9700_SPE30022F116F_0_SPE30022D3336_0
14                 9700_9700_SPE2D622F534V_0_SPE2DE17D0006_0
16                 4732_4732_47QSCC22F0FB2_0_47QSCC21A0001_0
…                                                        …
20041164                9700_-NONE-_SPE8ES24V0062_0_-NONE-_0
20041165           9700_9700_SPE60724FADWC_0_SPE60722D0024_0
20041168  8900_4732_89503422FWA401247_P00004_47QRAD20D81…
20041169      9700_9700_W15QKN20F0189_P00014_W15QKN19D0054_1
```

```
20041173  1205_7529_12314421F0478_P00009_HHSN31620120009…

                                   contract_award_unique_key  \
1            CONT_AWD_SPE2DV22F78C7_9700_SPE2DV17D4001_9700  
3            CONT_AWD_SPE30022F10X5_9700_SPE30022D3336_9700  
6            CONT_AWD_SPE30022F116F_9700_SPE30022D3336_9700  
14           CONT_AWD_SPE2D622F534V_9700_SPE2DE17D0006_9700  
16           CONT_AWD_47QSCC22F0FB2_4732_47QSCC21A0001_4732  
…                                                     …  
20041164          CONT_AWD_SPE8ES24V0062_9700_-NONE-_-NONE-  
20041165     CONT_AWD_SPE60724FADWC_9700_SPE60722D0024_9700  
20041168  CONT_AWD_89503422FWA401247_8900_47QRAD20D8175_…  
20041169     CONT_AWD_W15QKN20F0189_9700_W15QKN19D0054_9700  
20041173  CONT_AWD_12314421F0478_1205_HHSN316201200098W_…  

              award_id_piid  federal_action_obligation  \
1             SPE2DV22F78C7                      38.60  
3             SPE30022F10X5                   22029.85  
6             SPE30022F116F                    5448.90  
14            SPE2D622F534V                      54.62  
16            47QSCC22F0FB2                    2717.00  
…                         …                    …  
20041164      SPE8ES24V0062                    4240.00  
20041165      SPE60724FADWC                    2870.99  
20041168  89503422FWA401247                   32220.00  
20041169      W15QKN20F0189                 -197225.02  
20041173      12314421F0478                      -0.01  

          total_dollars_obligated  current_total_value_of_award  \
1                           38.60                         38.60  
3                        22029.85                      22029.85  
6                         5448.90                       5448.90  
14                          54.62                         54.62  
16                        2717.00                       2717.00  
…                               …                        …  
20041164                  4240.00                       4240.00  
20041165                  2870.99                       2870.99  
20041168                332220.00                     395444.27  
20041169              22127346.92                   22127346.92  
20041173               6339843.14                    6339843.14  

          potential_total_value_of_award action_date  action_date_fiscal_year  \
1                                  38.60  2021-11-10                     2022  
3                               22029.85  2021-11-10                     2022  
6                                5448.90  2021-11-10                     2022  
14                                 54.62  2021-11-10                     2022  
16                               2717.00  2021-11-10                     2022  
```

```
…                                  …         …                      …
20041164                         4240.00  2023-11-13                     2024
20041165                         2870.99  2023-11-13                     2024
20041168                       395444.27  2023-11-13                     2024
20041169                     22127346.92  2023-11-13                     2024
20041173                      6339843.14  2023-11-13                     2024

         period_of_performance_start_date  … hospital_flag foreign_owned  \
1                              2021-11-10  …             f             f
3                              2021-11-10  …             f             f
6                              2021-11-10  …             f             f
14                             2021-11-10  …             f             f
16                             2021-11-10  …             f             f
…                                    …  …           …           …
20041164                       2023-11-13  …             f             f
20041165                       2023-11-13  …             f             f
20041168                       2022-03-02  …             f             f
20041169                       2020-02-13  …             f             t
20041173                       2021-07-08  …             f             f

          for_profit_organization  nonprofit_organization  \
1                               t                       f
3                               t                       f
6                               t                       f
14                              t                       f
16                              t                       f
…                            …                     …
20041164                        t                       f
20041165                        t                       f
20041168                        t                       f
20041169                        t                       f
20041173                        t                       f

         the_ability_one_program historically_black_college tribal_college  \
1                              f                          f              f
3                              f                          f              f
6                              f                          f              f
14                             f                          f              f
16                             f                          f              f
…                           …                        …           …
20041164                       f                          f              f
20041165                       f                          f              f
20041168                       f                          f              f
20041169                       f                          f              f
20041173                       f                          f              f

         small_disadvantaged_business    sic4  \
1                                   f  5047.0
3                                   f  2015.0
6                                   f  2015.0
14                                  f  5047.0
16                                  f  5031.0
…                               …     …
20041164                            f  2891.0
20041165                            f  2911.0
20041168                            f  8711.0
20041169                            f  7361.0
20041173                            t  7373.0

                                     sic_desc
1              Medical and Hospital Equipment
3         Poultry Slaughtering and Processing
6         Poultry Slaughtering and Processing
14             Medical and Hospital Equipment
16              Lumber, Plywood, and Millwork
…                                       …
20041164               Adhesives and Sealants
20041165                   Petroleum Refining
20041168                 Engineering Services
20041169                  Employment Agencies
20041173   Computer Integrated Systems Design

[9197874 rows x 47 columns]
```

[15]:

```python
class CompustatMerger:
    """
    Merge USAspending contracts with Compustat firm financials
    using fuzzy name matching (parent first, then recipient).
    Optimized with deduplication + vectorized similarity.
    """

    def __init__(self, threshold: int = 70, n_jobs: int = 4):
        """
        Parameters
        ----------
        threshold : int
            Fuzzy match score cutoff (0-100). Higher = stricter matches.
        n_jobs : int
            Number of parallel jobs for fuzzy matching.
        """
        self.threshold = threshold
        self.n_jobs = n_jobs

    @staticmethod
    def _normalize(series: pd.Series) -> pd.Series:
        """Normalize names to improve fuzzy matching consistency."""
        return (
            series.astype(str)
            .str.upper()
            .str.replace(r"[^\w\s]", " ", regex=True)
            .str.replace(r"\s+", " ", regex=True)
            .str.strip()
        )

    def _best_match_dict(self, names, comp_names):
        """
        Compute best fuzzy matches for a list of names against comp_names.
        Returns a dict {name -> (best_match, score)}.
        """
        if not names or not comp_names:
            return {}

        # cdist computes full similarity matrix in one call
        scores = process.cdist(names, comp_names, scorer=fuzz.token_set_ratio)
        best_idx = scores.argmax(axis=1)
        best_scores = scores[np.arange(len(names)), best_idx]

        result = {
            name: (comp_names[idx], score)
            for name, idx, score in zip(names, best_idx, best_scores)
            if score >= self.threshold
        }
        return result

    def merge(self, df_us: pd.DataFrame, df_comp: pd.DataFrame) -> pd.DataFrame:
        """
        Perform fuzzy merge of USAspending with Compustat:
        - Match on SIC + fiscal year
        - Parent name first, then recipient if parent fails
        - Multiple awards per firm get same Compustat info for that year
        """
        df_us = df_us.copy()
        df_comp = df_comp.copy()

        # Normalize names
        df_us["recipient_parent_name_norm"] = self.
↪_normalize(df_us["recipient_parent_name"])
        df_us["recipient_name_norm"] = self._normalize(df_us["recipient_name"])
        df_comp["conm_norm"] = self._normalize(df_comp["conm"])

        logging.info("Normalization of names completed")
        keep_from_comp = [
            "conm", "sic4", "federal_fy",
            "at", "sale", "revt", "ib", "lt", "ceq", "oancf",
            "cogs"
        ]

        out_rows = []

        # Pre-index Compustat by (sic, year)
        comp_index = {(sic, year): g for (sic, year), g in df_comp.
↪groupby(["sic4", "federal_fy"])}

        for i, ((sic, year), g_us) in enumerate(
            df_us.groupby(["sic4", "action_date_fiscal_year"], sort=False),␣
↪start=1
        ):
            g_comp = comp_index.get((sic, year))
            if g_comp is None:
                continue

            comp_names = g_comp["conm_norm"].dropna().unique().tolist()
            if not comp_names:
                continue

            logging.info(
                "Processing group %s: SIC=%s, Year=%s, USAspending rows=%s,␣
↪Compustat rows=%s",
                i, sic, year, len(g_us), len(g_comp)
            )

            # Deduplicate names
            unique_parents = g_us["recipient_parent_name_norm"].dropna().
↪unique().tolist()
            unique_children = g_us["recipient_name_norm"].dropna().unique().
↪tolist()

            # Fuzzy matches (dicts)
            parent_matches = self._best_match_dict(unique_parents, comp_names)
            child_matches = self._best_match_dict(unique_children, comp_names)

            # Build lookup dict: parent first, then child
            match_map = {k: v[0] for k, v in parent_matches.items()}
            for k, v in child_matches.items():
                if k not in match_map:  # only fill if parent not matched
                    match_map[k] = v[0]
            # Vectorized assignment
            g_us = g_us.copy()
            g_us["matched_name"] = (
                g_us["recipient_parent_name_norm"].map(match_map)
                .fillna(g_us["recipient_name_norm"].map(match_map))
            )

            # Drop unmatched rows
            g_us = g_us.dropna(subset=["matched_name"])
            if g_us.empty:
                continue

            # Merge once (fast, vectorized)
            cols_present = [c for c in keep_from_comp if c in g_comp.columns]
            g_comp_keep = g_comp[cols_present + ["conm_norm"]].
↪rename(columns={"conm_norm": "matched_name"})

            g_merged = g_us.merge(g_comp_keep, on="matched_name", how="left",␣
↪suffixes=("", "_comp"))
            out_rows.append(g_merged)

        if not out_rows:
            logging.warning("No matches found between USAspending and Compustat.
↪")
            return pd.DataFrame()

        df_merged = pd.concat(out_rows, ignore_index=True)

        # Clean columns
        drop_cols = [
            "recipient_parent_name", "recipient_name", "sic4_comp",␣
↪"federal_fy",
            "recipient_parent_name_norm", "recipient_name_norm", "matched_name"
        ]
        df_merged.drop(columns=[c for c in drop_cols if c in df_merged.columns],
                       errors="ignore", inplace=True)

        df_merged.rename(columns={"conm": "company_name"}, inplace=True)

        logging.info("Final merged shape: %s rows, %s columns", *df_merged.
↪shape)
        return df_merged
```

[16]:

```
merger = CompustatMerger(threshold=70)
df_merged = merger.merge(df_usaspending, df_comp)
```

```
2025-08-21 17:42:56,658 | INFO | Normalization of names completed
```

```
2025-08-21 17:43:08,195 | INFO | Processing group 1: SIC=5047.0, Year=2022,
USAspending rows=738380, Compustat rows=7
2025-08-21 17:43:09,886 | INFO | Processing group 2: SIC=2015.0, Year=2022,
USAspending rows=63969, Compustat rows=6
2025-08-21 17:43:09,969 | INFO | Processing group 3: SIC=5031.0, Year=2022,
USAspending rows=121303, Compustat rows=4
2025-08-21 17:43:10,081 | INFO | Processing group 4: SIC=4581.0, Year=2022,
USAspending rows=25125, Compustat rows=6
2025-08-21 17:43:10,147 | INFO | Processing group 5: SIC=5122.0, Year=2022,
USAspending rows=302940, Compustat rows=15
2025-08-21 17:43:10,747 | INFO | Processing group 6: SIC=3728.0, Year=2022,
USAspending rows=89001, Compustat rows=11
2025-08-21 17:43:10,892 | INFO | Processing group 7: SIC=3721.0, Year=2022,
USAspending rows=40796, Compustat rows=11
2025-08-21 17:43:11,072 | INFO | Processing group 8: SIC=2911.0, Year=2022,
USAspending rows=244983, Compustat rows=25
2025-08-21 17:43:11,449 | INFO | Processing group 9: SIC=2833.0, Year=2022,
USAspending rows=176282, Compustat rows=14
2025-08-21 17:43:11,721 | INFO | Processing group 10: SIC=3711.0, Year=2022,
USAspending rows=10476, Compustat rows=43
2025-08-21 17:43:11,769 | INFO | Processing group 11: SIC=2531.0, Year=2022,
USAspending rows=133238, Compustat rows=3
2025-08-21 17:43:11,921 | INFO | Processing group 12: SIC=7373.0, Year=2022,
USAspending rows=25812, Compustat rows=37
2025-08-21 17:43:12,042 | INFO | Processing group 13: SIC=8711.0, Year=2022,
USAspending rows=72355, Compustat rows=15
2025-08-21 17:43:12,222 | INFO | Processing group 14: SIC=3669.0, Year=2022,
USAspending rows=6833, Compustat rows=11
2025-08-21 17:43:12,248 | INFO | Processing group 15: SIC=8742.0, Year=2022,
USAspending rows=35855, Compustat rows=24
2025-08-21 17:43:12,390 | INFO | Processing group 16: SIC=7363.0, Year=2022,
USAspending rows=11805, Compustat rows=15
2025-08-21 17:43:12,420 | INFO | Processing group 17: SIC=3829.0, Year=2022,
USAspending rows=32783, Compustat rows=25
2025-08-21 17:43:12,525 | INFO | Processing group 18: SIC=4522.0, Year=2022,
USAspending rows=11274, Compustat rows=6
2025-08-21 17:43:12,544 | INFO | Processing group 19: SIC=3851.0, Year=2022,
USAspending rows=4346, Compustat rows=6
2025-08-21 17:43:12,553 | INFO | Processing group 20: SIC=3678.0, Year=2022,
USAspending rows=11249, Compustat rows=6
2025-08-21 17:43:12,572 | INFO | Processing group 21: SIC=7359.0, Year=2022,
USAspending rows=6558, Compustat rows=13
2025-08-21 17:43:12,595 | INFO | Processing group 22: SIC=3844.0, Year=2022,
USAspending rows=3015, Compustat rows=4
2025-08-21 17:43:12,601 | INFO | Processing group 23: SIC=8731.0, Year=2022,
USAspending rows=41251, Compustat rows=27
2025-08-21 17:43:12,784 | INFO | Processing group 24: SIC=6531.0, Year=2022,
USAspending rows=3299, Compustat rows=28
```

```
2025-08-21 17:43:12,803 | INFO | Processing group 25: SIC=3826.0, Year=2022,
USAspending rows=15827, Compustat rows=20
2025-08-21 17:43:12,857 | INFO | Processing group 26: SIC=3531.0, Year=2022,
USAspending rows=10513, Compustat rows=4
2025-08-21 17:43:12,872 | INFO | Processing group 27: SIC=2851.0, Year=2022,
USAspending rows=20804, Compustat rows=6
2025-08-21 17:43:12,899 | INFO | Processing group 28: SIC=4911.0, Year=2022,
USAspending rows=5739, Compustat rows=101
2025-08-21 17:43:12,970 | INFO | Processing group 29: SIC=2086.0, Year=2022,
USAspending rows=6528, Compustat rows=20
2025-08-21 17:43:12,984 | INFO | Processing group 30: SIC=5141.0, Year=2022,
USAspending rows=16989, Compustat rows=1
2025-08-21 17:43:13,001 | INFO | Processing group 31: SIC=3812.0, Year=2022,
USAspending rows=17035, Compustat rows=17
2025-08-21 17:43:13,048 | INFO | Processing group 32: SIC=2834.0, Year=2022,
USAspending rows=25211, Compustat rows=290
2025-08-21 17:43:13,195 | INFO | Processing group 33: SIC=5063.0, Year=2022,
USAspending rows=20748, Compustat rows=3
2025-08-21 17:43:13,218 | INFO | Processing group 34: SIC=3842.0, Year=2022,
USAspending rows=4589, Compustat rows=45
2025-08-21 17:43:13,246 | INFO | Processing group 35: SIC=7389.0, Year=2022,
USAspending rows=22039, Compustat rows=37
2025-08-21 17:43:13,392 | INFO | Processing group 36: SIC=8744.0, Year=2022,
USAspending rows=29655, Compustat rows=2
2025-08-21 17:43:13,443 | INFO | Processing group 37: SIC=4412.0, Year=2022,
USAspending rows=3407, Compustat rows=50
2025-08-21 17:43:13,456 | INFO | Processing group 38: SIC=4923.0, Year=2022,
USAspending rows=2378, Compustat rows=12
2025-08-21 17:43:13,464 | INFO | Processing group 39: SIC=3086.0, Year=2022,
USAspending rows=1662, Compustat rows=2
2025-08-21 17:43:13,468 | INFO | Processing group 40: SIC=2673.0, Year=2022,
USAspending rows=8797, Compustat rows=1
2025-08-21 17:43:13,478 | INFO | Processing group 41: SIC=3571.0, Year=2022,
USAspending rows=32166, Compustat rows=4
2025-08-21 17:43:13,524 | INFO | Processing group 42: SIC=2421.0, Year=2022,
USAspending rows=16558, Compustat rows=3
2025-08-21 17:43:13,554 | INFO | Processing group 43: SIC=8734.0, Year=2022,
USAspending rows=5643, Compustat rows=5
2025-08-21 17:43:13,572 | INFO | Processing group 44: SIC=5072.0, Year=2022,
USAspending rows=130925, Compustat rows=2
2025-08-21 17:43:13,710 | INFO | Processing group 45: SIC=1531.0, Year=2022,
USAspending rows=1674, Compustat rows=21
2025-08-21 17:43:13,721 | INFO | Processing group 46: SIC=8011.0, Year=2022,
USAspending rows=5116, Compustat rows=5
2025-08-21 17:43:13,736 | INFO | Processing group 47: SIC=3562.0, Year=2022,
USAspending rows=13420, Compustat rows=5
2025-08-21 17:43:13,761 | INFO | Processing group 48: SIC=3577.0, Year=2022,
USAspending rows=2454, Compustat rows=16
```

```
2025-08-21 17:43:13,773 | INFO | Processing group 49: SIC=7361.0, Year=2022,
USAspending rows=2328, Compustat rows=11
2025-08-21 17:43:13,782 | INFO | Processing group 50: SIC=2522.0, Year=2022,
USAspending rows=46475, Compustat rows=2
2025-08-21 17:43:13,862 | INFO | Processing group 51: SIC=3634.0, Year=2022,
USAspending rows=1713, Compustat rows=4
2025-08-21 17:43:13,865 | INFO | Processing group 52: SIC=4899.0, Year=2022,
USAspending rows=13111, Compustat rows=21
2025-08-21 17:43:13,895 | INFO | Processing group 53: SIC=7371.0, Year=2022,
USAspending rows=14622, Compustat rows=13
2025-08-21 17:43:13,948 | INFO | Processing group 54: SIC=3312.0, Year=2022,
USAspending rows=2400, Compustat rows=17
2025-08-21 17:43:13,957 | INFO | Processing group 55: SIC=4953.0, Year=2022,
USAspending rows=18055, Compustat rows=7
2025-08-21 17:43:13,991 | INFO | Processing group 56: SIC=8062.0, Year=2022,
USAspending rows=12852, Compustat rows=7
2025-08-21 17:43:14,012 | INFO | Processing group 57: SIC=7381.0, Year=2022,
USAspending rows=15016, Compustat rows=4
2025-08-21 17:43:14,037 | INFO | Processing group 58: SIC=4731.0, Year=2022,
USAspending rows=48576, Compustat rows=15
2025-08-21 17:43:14,099 | INFO | Processing group 59: SIC=3448.0, Year=2022,
USAspending rows=1011, Compustat rows=2
2025-08-21 17:43:14,102 | INFO | Processing group 60: SIC=3621.0, Year=2022,
USAspending rows=4397, Compustat rows=15
2025-08-21 17:43:14,122 | INFO | Processing group 61: SIC=3724.0, Year=2022,
USAspending rows=8000, Compustat rows=3
2025-08-21 17:43:14,135 | INFO | Processing group 62: SIC=8721.0, Year=2022,
USAspending rows=4012, Compustat rows=4
2025-08-21 17:43:14,141 | INFO | Processing group 63: SIC=7374.0, Year=2022,
USAspending rows=5074, Compustat rows=46
2025-08-21 17:43:14,203 | INFO | Processing group 64: SIC=8111.0, Year=2022,
USAspending rows=1879, Compustat rows=3
2025-08-21 17:43:14,208 | INFO | Processing group 65: SIC=3443.0, Year=2022,
USAspending rows=1234, Compustat rows=3
2025-08-21 17:43:14,213 | INFO | Processing group 66: SIC=3564.0, Year=2022,
USAspending rows=1955, Compustat rows=3
2025-08-21 17:43:14,218 | INFO | Processing group 67: SIC=3317.0, Year=2022,
USAspending rows=1449, Compustat rows=4
2025-08-21 17:43:14,222 | INFO | Processing group 68: SIC=3663.0, Year=2022,
USAspending rows=14333, Compustat rows=46
2025-08-21 17:43:14,321 | INFO | Processing group 69: SIC=3823.0, Year=2022,
USAspending rows=3750, Compustat rows=14
2025-08-21 17:43:14,339 | INFO | Processing group 70: SIC=2842.0, Year=2022,
USAspending rows=2079, Compustat rows=4
2025-08-21 17:43:14,343 | INFO | Processing group 71: SIC=3021.0, Year=2022,
USAspending rows=1056, Compustat rows=3
2025-08-21 17:43:14,345 | INFO | Processing group 72: SIC=7011.0, Year=2022,
USAspending rows=7273, Compustat rows=15
```

```
2025-08-21 17:43:14,370 | INFO | Processing group 73: SIC=4512.0, Year=2022,
USAspending rows=497, Compustat rows=29
2025-08-21 17:43:14,374 | INFO | Processing group 74: SIC=8741.0, Year=2022,
USAspending rows=3837, Compustat rows=7
2025-08-21 17:43:14,385 | INFO | Processing group 75: SIC=4213.0, Year=2022,
USAspending rows=382, Compustat rows=18
2025-08-21 17:43:14,389 | INFO | Processing group 76: SIC=4941.0, Year=2022,
USAspending rows=3069, Compustat rows=14
2025-08-21 17:43:14,405 | INFO | Processing group 77: SIC=8071.0, Year=2022,
USAspending rows=5288, Compustat rows=19
2025-08-21 17:43:14,425 | INFO | Processing group 78: SIC=3442.0, Year=2022,
USAspending rows=1165, Compustat rows=5
2025-08-21 17:43:14,429 | INFO | Processing group 79: SIC=3843.0, Year=2022,
USAspending rows=1568, Compustat rows=7
2025-08-21 17:43:14,434 | INFO | Processing group 80: SIC=2836.0, Year=2022,
USAspending rows=1055, Compustat rows=688
2025-08-21 17:43:14,581 | INFO | Processing group 81: SIC=2835.0, Year=2022,
USAspending rows=4252, Compustat rows=49
2025-08-21 17:43:14,602 | INFO | Processing group 82: SIC=3561.0, Year=2022,
USAspending rows=3403, Compustat rows=8
2025-08-21 17:43:14,613 | INFO | Processing group 83: SIC=2033.0, Year=2022,
USAspending rows=1180, Compustat rows=3
2025-08-21 17:43:14,616 | INFO | Processing group 84: SIC=3674.0, Year=2022,
USAspending rows=3755, Compustat rows=107
2025-08-21 17:43:14,681 | INFO | Processing group 85: SIC=6361.0, Year=2022,
USAspending rows=16, Compustat rows=5
2025-08-21 17:43:14,683 | INFO | Processing group 86: SIC=3714.0, Year=2022,
USAspending rows=1320, Compustat rows=46
2025-08-21 17:43:14,693 | INFO | Processing group 87: SIC=2891.0, Year=2022,
USAspending rows=9969, Compustat rows=3
2025-08-21 17:43:14,703 | INFO | Processing group 88: SIC=3541.0, Year=2022,
USAspending rows=1218, Compustat rows=1
2025-08-21 17:43:14,706 | INFO | Processing group 89: SIC=7822.0, Year=2022,
USAspending rows=100, Compustat rows=3
2025-08-21 17:43:14,708 | INFO | Processing group 90: SIC=1623.0, Year=2022,
USAspending rows=3789, Compustat rows=9
2025-08-21 17:43:14,725 | INFO | Processing group 91: SIC=5171.0, Year=2022,
USAspending rows=91, Compustat rows=8
2025-08-21 17:43:14,728 | INFO | Processing group 92: SIC=3672.0, Year=2022,
USAspending rows=4643, Compustat rows=10
2025-08-21 17:43:14,740 | INFO | Processing group 93: SIC=5084.0, Year=2022,
USAspending rows=4221, Compustat rows=4
2025-08-21 17:43:14,746 | INFO | Processing group 94: SIC=8093.0, Year=2022,
USAspending rows=3670, Compustat rows=5
2025-08-21 17:43:14,754 | INFO | Processing group 95: SIC=6513.0, Year=2022,
USAspending rows=397, Compustat rows=4
2025-08-21 17:43:14,757 | INFO | Processing group 96: SIC=3272.0, Year=2022,
USAspending rows=1211, Compustat rows=2
```

```
2025-08-21 17:43:14,761 | INFO | Processing group 97: SIC=8051.0, Year=2022,
USAspending rows=3593, Compustat rows=3
2025-08-21 17:43:14,771 | INFO | Processing group 98: SIC=7311.0, Year=2022,
USAspending rows=965, Compustat rows=15
2025-08-21 17:43:14,776 | INFO | Processing group 99: SIC=3661.0, Year=2022,
USAspending rows=1099, Compustat rows=14
2025-08-21 17:43:14,782 | INFO | Processing group 100: SIC=8082.0, Year=2022,
USAspending rows=339, Compustat rows=9
2025-08-21 17:43:14,785 | INFO | Processing group 101: SIC=7948.0, Year=2022,
USAspending rows=166, Compustat rows=1
2025-08-21 17:43:14,787 | INFO | Processing group 103: SIC=6411.0, Year=2022,
USAspending rows=704, Compustat rows=24
2025-08-21 17:43:14,792 | INFO | Processing group 104: SIC=2821.0, Year=2022,
USAspending rows=638, Compustat rows=13
2025-08-21 17:43:14,797 | INFO | Processing group 105: SIC=3613.0, Year=2022,
USAspending rows=4842, Compustat rows=4
2025-08-21 17:43:14,806 | INFO | Processing group 106: SIC=3341.0, Year=2022,
USAspending rows=1141, Compustat rows=1
2025-08-21 17:43:14,809 | INFO | Processing group 107: SIC=3572.0, Year=2022,
USAspending rows=1843, Compustat rows=6
2025-08-21 17:43:14,815 | INFO | Processing group 109: SIC=2452.0, Year=2022,
USAspending rows=247, Compustat rows=2
2025-08-21 17:43:14,818 | INFO | Processing group 110: SIC=5045.0, Year=2022,
USAspending rows=1211, Compustat rows=9
2025-08-21 17:43:14,825 | INFO | Processing group 111: SIC=5065.0, Year=2022,
USAspending rows=468, Compustat rows=10
2025-08-21 17:43:14,829 | INFO | Processing group 112: SIC=3651.0, Year=2022,
USAspending rows=3100, Compustat rows=16
2025-08-21 17:43:14,846 | INFO | Processing group 113: SIC=3357.0, Year=2022,
USAspending rows=1562, Compustat rows=3
2025-08-21 17:43:14,850 | INFO | Processing group 114: SIC=5093.0, Year=2022,
USAspending rows=18, Compustat rows=1
2025-08-21 17:43:14,852 | INFO | Processing group 115: SIC=3523.0, Year=2022,
USAspending rows=939, Compustat rows=12
2025-08-21 17:43:14,858 | INFO | Processing group 116: SIC=3861.0, Year=2022,
USAspending rows=7479, Compustat rows=10
2025-08-21 17:43:14,867 | INFO | Processing group 117: SIC=7323.0, Year=2022,
USAspending rows=645, Compustat rows=7
2025-08-21 17:43:14,870 | INFO | Processing group 118: SIC=2273.0, Year=2022,
USAspending rows=272, Compustat rows=3
2025-08-21 17:43:14,873 | INFO | Processing group 119: SIC=6099.0, Year=2022,
USAspending rows=524, Compustat rows=11
2025-08-21 17:43:14,876 | INFO | Processing group 120: SIC=6321.0, Year=2022,
USAspending rows=1437, Compustat rows=5
2025-08-21 17:43:14,881 | INFO | Processing group 121: SIC=6512.0, Year=2022,
USAspending rows=1417, Compustat rows=21
2025-08-21 17:43:14,896 | INFO | Processing group 122: SIC=3081.0, Year=2022,
USAspending rows=195, Compustat rows=3
```

```
2025-08-21 17:43:14,898 | INFO | Processing group 123: SIC=2844.0, Year=2022,
USAspending rows=559, Compustat rows=23
2025-08-21 17:43:14,902 | INFO | Processing group 124: SIC=7812.0, Year=2022,
USAspending rows=601, Compustat rows=7
2025-08-21 17:43:14,906 | INFO | Processing group 126: SIC=3011.0, Year=2022,
USAspending rows=1287, Compustat rows=2
2025-08-21 17:43:14,909 | INFO | Processing group 127: SIC=5812.0, Year=2022,
USAspending rows=873, Compustat rows=58
2025-08-21 17:43:14,926 | INFO | Processing group 128: SIC=3824.0, Year=2022,
USAspending rows=1883, Compustat rows=3
2025-08-21 17:43:14,931 | INFO | Processing group 129: SIC=2511.0, Year=2022,
USAspending rows=2271, Compustat rows=2
2025-08-21 17:43:14,935 | INFO | Processing group 130: SIC=3751.0, Year=2022,
USAspending rows=276, Compustat rows=7
2025-08-21 17:43:14,938 | INFO | Processing group 131: SIC=3524.0, Year=2022,
USAspending rows=437, Compustat rows=1
2025-08-21 17:43:14,940 | INFO | Processing group 132: SIC=6153.0, Year=2022,
USAspending rows=248, Compustat rows=7
2025-08-21 17:43:14,943 | INFO | Processing group 133: SIC=3715.0, Year=2022,
USAspending rows=852, Compustat rows=1
2025-08-21 17:43:14,945 | INFO | Processing group 134: SIC=3221.0, Year=2022,
USAspending rows=63, Compustat rows=1
2025-08-21 17:43:14,946 | INFO | Processing group 135: SIC=2211.0, Year=2022,
USAspending rows=1578, Compustat rows=2
2025-08-21 17:43:14,949 | INFO | Processing group 136: SIC=1382.0, Year=2022,
USAspending rows=207, Compustat rows=6
2025-08-21 17:43:14,952 | INFO | Processing group 137: SIC=5013.0, Year=2022,
USAspending rows=205, Compustat rows=2
2025-08-21 17:43:14,954 | INFO | Processing group 138: SIC=1389.0, Year=2022,
USAspending rows=588, Compustat rows=30
2025-08-21 17:43:14,962 | INFO | Processing group 139: SIC=7819.0, Year=2022,
USAspending rows=209, Compustat rows=1
2025-08-21 17:43:14,963 | INFO | Processing group 140: SIC=6163.0, Year=2022,
USAspending rows=153, Compustat rows=4
2025-08-21 17:43:14,965 | INFO | Processing group 141: SIC=8351.0, Year=2022,
USAspending rows=57, Compustat rows=1
2025-08-21 17:43:14,966 | INFO | Processing group 142: SIC=3281.0, Year=2022,
USAspending rows=131, Compustat rows=1
2025-08-21 17:43:14,968 | INFO | Processing group 143: SIC=2013.0, Year=2022,
USAspending rows=360, Compustat rows=3
2025-08-21 17:43:14,970 | INFO | Processing group 144: SIC=3533.0, Year=2022,
USAspending rows=103, Compustat rows=18
2025-08-21 17:43:14,972 | INFO | Processing group 145: SIC=6211.0, Year=2022,
USAspending rows=55, Compustat rows=40
2025-08-21 17:43:14,975 | INFO | Processing group 146: SIC=1381.0, Year=2022,
USAspending rows=14, Compustat rows=12
2025-08-21 17:43:14,977 | INFO | Processing group 147: SIC=5082.0, Year=2022,
USAspending rows=35, Compustat rows=1
```

```
2025-08-21 17:43:14,978 | INFO | Processing group 148: SIC=6331.0, Year=2022,
USAspending rows=51, Compustat rows=75
2025-08-21 17:43:14,982 | INFO | Processing group 149: SIC=5099.0, Year=2022,
USAspending rows=286, Compustat rows=1
2025-08-21 17:43:14,984 | INFO | Processing group 150: SIC=3559.0, Year=2022,
USAspending rows=59, Compustat rows=25
2025-08-21 17:43:14,987 | INFO | Processing group 151: SIC=2611.0, Year=2022,
USAspending rows=26, Compustat rows=1
2025-08-21 17:43:14,988 | INFO | Processing group 152: SIC=3411.0, Year=2022,
USAspending rows=150, Compustat rows=5
2025-08-21 17:43:14,990 | INFO | Processing group 153: SIC=5172.0, Year=2022,
USAspending rows=223, Compustat rows=11
2025-08-21 17:43:14,992 | INFO | Processing group 154: SIC=5051.0, Year=2022,
USAspending rows=58, Compustat rows=7
2025-08-21 17:43:14,995 | INFO | Processing group 155: SIC=2451.0, Year=2022,
USAspending rows=126, Compustat rows=3
2025-08-21 17:43:14,997 | INFO | Processing group 156: SIC=2085.0, Year=2022,
USAspending rows=14, Compustat rows=4
2025-08-21 17:43:14,998 | INFO | Processing group 157: SIC=2052.0, Year=2022,
USAspending rows=60, Compustat rows=1
2025-08-21 17:43:15,000 | INFO | Processing group 158: SIC=6311.0, Year=2022,
USAspending rows=127, Compustat rows=39
2025-08-21 17:43:15,004 | INFO | Processing group 159: SIC=4922.0, Year=2022,
USAspending rows=16, Compustat rows=9
2025-08-21 17:43:15,005 | INFO | Processing group 160: SIC=6141.0, Year=2022,
USAspending rows=1, Compustat rows=35
2025-08-21 17:43:15,006 | INFO | Processing group 161: SIC=3211.0, Year=2022,
USAspending rows=13, Compustat rows=1
2025-08-21 17:43:15,007 | INFO | Processing group 162: SIC=2082.0, Year=2022,
USAspending rows=27, Compustat rows=9
2025-08-21 17:43:15,009 | INFO | Processing group 163: SIC=2731.0, Year=2022,
USAspending rows=30, Compustat rows=3
2025-08-21 17:43:15,010 | INFO | Processing group 164: SIC=4011.0, Year=2022,
USAspending rows=13, Compustat rows=7
2025-08-21 17:43:15,013 | INFO | Processing group 165: SIC=6111.0, Year=2022,
USAspending rows=7, Compustat rows=4
2025-08-21 17:43:15,014 | INFO | Processing group 166: SIC=7996.0, Year=2022,
USAspending rows=8, Compustat rows=5
2025-08-21 17:43:15,015 | INFO | Processing group 167: SIC=6552.0, Year=2022,
USAspending rows=8, Compustat rows=15
2025-08-21 17:43:15,016 | INFO | Processing group 168: SIC=3652.0, Year=2022,
USAspending rows=8, Compustat rows=1
2025-08-21 17:43:15,017 | INFO | Processing group 169: SIC=7331.0, Year=2022,
USAspending rows=10, Compustat rows=1
2025-08-21 17:43:15,019 | INFO | Processing group 171: SIC=2084.0, Year=2022,
USAspending rows=7, Compustat rows=6
2025-08-21 17:43:15,020 | INFO | Processing group 172: SIC=6792.0, Year=2022,
USAspending rows=4, Compustat rows=14
```

```
2025-08-21 17:43:15,021 | INFO | Processing group 173: SIC=7841.0, Year=2022,
USAspending rows=19, Compustat rows=3
2025-08-21 17:43:15,022 | INFO | Processing group 174: SIC=1311.0, Year=2022,
USAspending rows=11, Compustat rows=141
2025-08-21 17:43:15,025 | INFO | Processing group 175: SIC=3241.0, Year=2022,
USAspending rows=7, Compustat rows=5
2025-08-21 17:43:15,026 | INFO | Processing group 176: SIC=6722.0, Year=2022,
USAspending rows=3, Compustat rows=1
2025-08-21 17:43:15,027 | INFO | Processing group 177: SIC=2024.0, Year=2022,
USAspending rows=1, Compustat rows=1
2025-08-21 17:43:15,028 | INFO | Processing group 178: SIC=2911.0, Year=2023,
USAspending rows=234625, Compustat rows=23
2025-08-21 17:43:16,098 | INFO | Processing group 179: SIC=3678.0, Year=2023,
USAspending rows=10262, Compustat rows=6
2025-08-21 17:43:16,130 | INFO | Processing group 180: SIC=2015.0, Year=2023,
USAspending rows=60507, Compustat rows=4
2025-08-21 17:43:16,229 | INFO | Processing group 181: SIC=3728.0, Year=2023,
USAspending rows=94632, Compustat rows=12
2025-08-21 17:43:16,404 | INFO | Processing group 182: SIC=8744.0, Year=2023,
USAspending rows=30408, Compustat rows=2
2025-08-21 17:43:16,482 | INFO | Processing group 183: SIC=3571.0, Year=2023,
USAspending rows=33865, Compustat rows=5
2025-08-21 17:43:16,540 | INFO | Processing group 184: SIC=8742.0, Year=2023,
USAspending rows=35966, Compustat rows=26
2025-08-21 17:43:16,683 | INFO | Processing group 185: SIC=3317.0, Year=2023,
USAspending rows=1317, Compustat rows=4
2025-08-21 17:43:16,688 | INFO | Processing group 186: SIC=7812.0, Year=2023,
USAspending rows=574, Compustat rows=9
2025-08-21 17:43:16,692 | INFO | Processing group 187: SIC=5047.0, Year=2023,
USAspending rows=719661, Compustat rows=7
2025-08-21 17:43:17,733 | INFO | Processing group 188: SIC=2833.0, Year=2023,
USAspending rows=166292, Compustat rows=12
2025-08-21 17:43:17,963 | INFO | Processing group 189: SIC=5122.0, Year=2023,
USAspending rows=353411, Compustat rows=15
2025-08-21 17:43:18,580 | INFO | Processing group 190: SIC=3721.0, Year=2023,
USAspending rows=36147, Compustat rows=13
2025-08-21 17:43:18,668 | INFO | Processing group 191: SIC=3823.0, Year=2023,
USAspending rows=3790, Compustat rows=13
2025-08-21 17:43:18,693 | INFO | Processing group 192: SIC=2673.0, Year=2023,
USAspending rows=10011, Compustat rows=1
2025-08-21 17:43:18,713 | INFO | Processing group 193: SIC=7373.0, Year=2023,
USAspending rows=27266, Compustat rows=33
2025-08-21 17:43:18,852 | INFO | Processing group 194: SIC=4731.0, Year=2023,
USAspending rows=49725, Compustat rows=16
2025-08-21 17:43:18,973 | INFO | Processing group 195: SIC=8711.0, Year=2023,
USAspending rows=71254, Compustat rows=15
2025-08-21 17:43:19,143 | INFO | Processing group 196: SIC=4522.0, Year=2023,
USAspending rows=12148, Compustat rows=8
```

```
2025-08-21 17:43:19,161 | INFO | Processing group 197: SIC=8051.0, Year=2023,
USAspending rows=3833, Compustat rows=3
2025-08-21 17:43:19,172 | INFO | Processing group 198: SIC=3021.0, Year=2023,
USAspending rows=817, Compustat rows=4
2025-08-21 17:43:19,174 | INFO | Processing group 199: SIC=4581.0, Year=2023,
USAspending rows=22853, Compustat rows=8
2025-08-21 17:43:19,233 | INFO | Processing group 200: SIC=5063.0, Year=2023,
USAspending rows=30939, Compustat rows=2
2025-08-21 17:43:19,269 | INFO | Processing group 201: SIC=8731.0, Year=2023,
USAspending rows=42964, Compustat rows=26
2025-08-21 17:43:19,453 | INFO | Processing group 202: SIC=5141.0, Year=2023,
USAspending rows=13977, Compustat rows=1
2025-08-21 17:43:19,466 | INFO | Processing group 203: SIC=3829.0, Year=2023,
USAspending rows=32408, Compustat rows=28
2025-08-21 17:43:19,573 | INFO | Processing group 204: SIC=5031.0, Year=2023,
USAspending rows=159059, Compustat rows=3
2025-08-21 17:43:19,744 | INFO | Processing group 205: SIC=3812.0, Year=2023,
USAspending rows=16674, Compustat rows=15
2025-08-21 17:43:19,793 | INFO | Processing group 206: SIC=3663.0, Year=2023,
USAspending rows=13183, Compustat rows=41
2025-08-21 17:43:19,890 | INFO | Processing group 207: SIC=4953.0, Year=2023,
USAspending rows=18973, Compustat rows=7
2025-08-21 17:43:19,937 | INFO | Processing group 208: SIC=2834.0, Year=2023,
USAspending rows=23785, Compustat rows=284
2025-08-21 17:43:20,094 | INFO | Processing group 209: SIC=2531.0, Year=2023,
USAspending rows=125665, Compustat rows=3
2025-08-21 17:43:20,243 | INFO | Processing group 210: SIC=5072.0, Year=2023,
USAspending rows=90314, Compustat rows=2
2025-08-21 17:43:20,336 | INFO | Processing group 211: SIC=8062.0, Year=2023,
USAspending rows=12991, Compustat rows=9
2025-08-21 17:43:20,358 | INFO | Processing group 212: SIC=3826.0, Year=2023,
USAspending rows=15154, Compustat rows=22
2025-08-21 17:43:20,416 | INFO | Processing group 213: SIC=4899.0, Year=2023,
USAspending rows=10016, Compustat rows=21
2025-08-21 17:43:20,448 | INFO | Processing group 214: SIC=7389.0, Year=2023,
USAspending rows=26458, Compustat rows=46
2025-08-21 17:43:20,653 | INFO | Processing group 215: SIC=3562.0, Year=2023,
USAspending rows=14572, Compustat rows=5
2025-08-21 17:43:20,680 | INFO | Processing group 216: SIC=2421.0, Year=2023,
USAspending rows=14220, Compustat rows=3
2025-08-21 17:43:20,703 | INFO | Processing group 217: SIC=2522.0, Year=2023,
USAspending rows=49280, Compustat rows=2
2025-08-21 17:43:20,766 | INFO | Processing group 218: SIC=3724.0, Year=2023,
USAspending rows=8988, Compustat rows=3
2025-08-21 17:43:20,778 | INFO | Processing group 219: SIC=3669.0, Year=2023,
USAspending rows=7801, Compustat rows=12
2025-08-21 17:43:20,802 | INFO | Processing group 220: SIC=3442.0, Year=2023,
USAspending rows=1210, Compustat rows=5
```

```
2025-08-21 17:43:20,807 | INFO | Processing group 221: SIC=8071.0, Year=2023,
USAspending rows=4720, Compustat rows=20
2025-08-21 17:43:20,827 | INFO | Processing group 222: SIC=5171.0, Year=2023,
USAspending rows=514, Compustat rows=6
2025-08-21 17:43:20,830 | INFO | Processing group 223: SIC=2851.0, Year=2023,
USAspending rows=24759, Compustat rows=4
2025-08-21 17:43:20,863 | INFO | Processing group 224: SIC=2891.0, Year=2023,
USAspending rows=19412, Compustat rows=2
2025-08-21 17:43:20,887 | INFO | Processing group 225: SIC=3312.0, Year=2023,
USAspending rows=2248, Compustat rows=16
2025-08-21 17:43:20,896 | INFO | Processing group 226: SIC=4923.0, Year=2023,
USAspending rows=2246, Compustat rows=11
2025-08-21 17:43:20,904 | INFO | Processing group 227: SIC=3711.0, Year=2023,
USAspending rows=10676, Compustat rows=48
2025-08-21 17:43:20,959 | INFO | Processing group 228: SIC=7363.0, Year=2023,
USAspending rows=11092, Compustat rows=16
2025-08-21 17:43:20,985 | INFO | Processing group 229: SIC=4941.0, Year=2023,
USAspending rows=3042, Compustat rows=14
2025-08-21 17:43:21,002 | INFO | Processing group 230: SIC=6321.0, Year=2023,
USAspending rows=1586, Compustat rows=5
2025-08-21 17:43:21,009 | INFO | Processing group 231: SIC=3443.0, Year=2023,
USAspending rows=1201, Compustat rows=4
2025-08-21 17:43:21,014 | INFO | Processing group 232: SIC=3861.0, Year=2023,
USAspending rows=8571, Compustat rows=10
2025-08-21 17:43:21,024 | INFO | Processing group 233: SIC=3086.0, Year=2023,
USAspending rows=2245, Compustat rows=2
2025-08-21 17:43:21,029 | INFO | Processing group 234: SIC=3851.0, Year=2023,
USAspending rows=4303, Compustat rows=6
2025-08-21 17:43:21,035 | INFO | Processing group 235: SIC=4412.0, Year=2023,
USAspending rows=3931, Compustat rows=50
2025-08-21 17:43:21,078 | INFO | Processing group 236: SIC=5065.0, Year=2023,
USAspending rows=376, Compustat rows=12
2025-08-21 17:43:21,102 | INFO | Processing group 237: SIC=7381.0, Year=2023,
USAspending rows=14957, Compustat rows=6
2025-08-21 17:43:21,155 | INFO | Processing group 238: SIC=7011.0, Year=2023,
USAspending rows=10505, Compustat rows=15
2025-08-21 17:43:21,195 | INFO | Processing group 239: SIC=3559.0, Year=2023,
USAspending rows=1931, Compustat rows=25
2025-08-21 17:43:21,208 | INFO | Processing group 240: SIC=5084.0, Year=2023,
USAspending rows=6777, Compustat rows=4
2025-08-21 17:43:21,220 | INFO | Processing group 241: SIC=3844.0, Year=2023,
USAspending rows=2974, Compustat rows=4
2025-08-21 17:43:21,227 | INFO | Processing group 242: SIC=3575.0, Year=2023,
USAspending rows=1993, Compustat rows=1
2025-08-21 17:43:21,232 | INFO | Processing group 243: SIC=2086.0, Year=2023,
USAspending rows=5378, Compustat rows=21
2025-08-21 17:43:21,245 | INFO | Processing group 244: SIC=2836.0, Year=2023,
USAspending rows=910, Compustat rows=688
```

```
2025-08-21 17:43:21,377 | INFO | Processing group 245: SIC=2842.0, Year=2023,
USAspending rows=2310, Compustat rows=3
2025-08-21 17:43:21,383 | INFO | Processing group 246: SIC=8734.0, Year=2023,
USAspending rows=5458, Compustat rows=6
2025-08-21 17:43:21,401 | INFO | Processing group 247: SIC=3672.0, Year=2023,
USAspending rows=4593, Compustat rows=11
2025-08-21 17:43:21,414 | INFO | Processing group 248: SIC=3634.0, Year=2023,
USAspending rows=1017, Compustat rows=5
2025-08-21 17:43:21,417 | INFO | Processing group 249: SIC=2211.0, Year=2023,
USAspending rows=1125, Compustat rows=2
2025-08-21 17:43:21,419 | INFO | Processing group 250: SIC=5051.0, Year=2023,
USAspending rows=99, Compustat rows=7
2025-08-21 17:43:21,422 | INFO | Processing group 251: SIC=7361.0, Year=2023,
USAspending rows=2074, Compustat rows=14
2025-08-21 17:43:21,432 | INFO | Processing group 252: SIC=8721.0, Year=2023,
USAspending rows=3779, Compustat rows=5
2025-08-21 17:43:21,441 | INFO | Processing group 253: SIC=1623.0, Year=2023,
USAspending rows=3707, Compustat rows=12
2025-08-21 17:43:21,462 | INFO | Processing group 254: SIC=3674.0, Year=2023,
USAspending rows=3876, Compustat rows=108
2025-08-21 17:43:21,527 | INFO | Processing group 255: SIC=8011.0, Year=2023,
USAspending rows=4407, Compustat rows=7
2025-08-21 17:43:21,541 | INFO | Processing group 256: SIC=7374.0, Year=2023,
USAspending rows=5244, Compustat rows=41
2025-08-21 17:43:21,598 | INFO | Processing group 257: SIC=7359.0, Year=2023,
USAspending rows=7214, Compustat rows=15
2025-08-21 17:43:21,625 | INFO | Processing group 258: SIC=4911.0, Year=2023,
USAspending rows=5469, Compustat rows=98
2025-08-21 17:43:21,694 | INFO | Processing group 259: SIC=3572.0, Year=2023,
USAspending rows=1817, Compustat rows=8
2025-08-21 17:43:21,700 | INFO | Processing group 260: SIC=3577.0, Year=2023,
USAspending rows=2951, Compustat rows=14
2025-08-21 17:43:21,711 | INFO | Processing group 261: SIC=3564.0, Year=2023,
USAspending rows=1767, Compustat rows=3
2025-08-21 17:43:21,716 | INFO | Processing group 262: SIC=7372.0, Year=2023,
USAspending rows=3823, Compustat rows=165
2025-08-21 17:43:21,946 | INFO | Processing group 263: SIC=6531.0, Year=2023,
USAspending rows=3722, Compustat rows=28
2025-08-21 17:43:21,965 | INFO | Processing group 264: SIC=2511.0, Year=2023,
USAspending rows=2398, Compustat rows=2
2025-08-21 17:43:21,969 | INFO | Processing group 265: SIC=3531.0, Year=2023,
USAspending rows=11122, Compustat rows=4
2025-08-21 17:43:21,984 | INFO | Processing group 266: SIC=2821.0, Year=2023,
USAspending rows=620, Compustat rows=13
2025-08-21 17:43:21,990 | INFO | Processing group 267: SIC=7311.0, Year=2023,
USAspending rows=1048, Compustat rows=15
2025-08-21 17:43:21,996 | INFO | Processing group 268: SIC=2835.0, Year=2023,
USAspending rows=3815, Compustat rows=50
```

```
2025-08-21 17:43:22,017 | INFO | Processing group 269: SIC=3842.0, Year=2023,
USAspending rows=3024, Compustat rows=46
2025-08-21 17:43:22,045 | INFO | Processing group 270: SIC=8741.0, Year=2023,
USAspending rows=3210, Compustat rows=6
2025-08-21 17:43:22,053 | INFO | Processing group 271: SIC=3613.0, Year=2023,
USAspending rows=4424, Compustat rows=4
2025-08-21 17:43:22,063 | INFO | Processing group 272: SIC=8093.0, Year=2023,
USAspending rows=3378, Compustat rows=5
2025-08-21 17:43:22,070 | INFO | Processing group 273: SIC=7371.0, Year=2023,
USAspending rows=14923, Compustat rows=18
2025-08-21 17:43:22,134 | INFO | Processing group 274: SIC=2033.0, Year=2023,
USAspending rows=1463, Compustat rows=3
2025-08-21 17:43:22,138 | INFO | Processing group 275: SIC=3541.0, Year=2023,
USAspending rows=1227, Compustat rows=1
2025-08-21 17:43:22,141 | INFO | Processing group 276: SIC=3357.0, Year=2023,
USAspending rows=1513, Compustat rows=3
2025-08-21 17:43:22,146 | INFO | Processing group 277: SIC=3561.0, Year=2023,
USAspending rows=2928, Compustat rows=7
2025-08-21 17:43:22,155 | INFO | Processing group 278: SIC=3621.0, Year=2023,
USAspending rows=3915, Compustat rows=16
2025-08-21 17:43:22,175 | INFO | Processing group 279: SIC=1382.0, Year=2023,
USAspending rows=163, Compustat rows=5
2025-08-21 17:43:22,178 | INFO | Processing group 280: SIC=3651.0, Year=2023,
USAspending rows=2554, Compustat rows=15
2025-08-21 17:43:22,193 | INFO | Processing group 281: SIC=1531.0, Year=2023,
USAspending rows=1595, Compustat rows=21
2025-08-21 17:43:22,203 | INFO | Processing group 282: SIC=5812.0, Year=2023,
USAspending rows=1127, Compustat rows=60
2025-08-21 17:43:22,222 | INFO | Processing group 283: SIC=3578.0, Year=2023,
USAspending rows=3699, Compustat rows=4
2025-08-21 17:43:22,253 | INFO | Processing group 284: SIC=3715.0, Year=2023,
USAspending rows=1121, Compustat rows=1
2025-08-21 17:43:22,260 | INFO | Processing group 285: SIC=4213.0, Year=2023,
USAspending rows=363, Compustat rows=18
2025-08-21 17:43:22,266 | INFO | Processing group 286: SIC=5013.0, Year=2023,
USAspending rows=252, Compustat rows=3
2025-08-21 17:43:22,269 | INFO | Processing group 287: SIC=3272.0, Year=2023,
USAspending rows=1105, Compustat rows=1
2025-08-21 17:43:22,272 | INFO | Processing group 288: SIC=3661.0, Year=2023,
USAspending rows=1066, Compustat rows=13
2025-08-21 17:43:22,279 | INFO | Processing group 289: SIC=6512.0, Year=2023,
USAspending rows=1317, Compustat rows=21
2025-08-21 17:43:22,294 | INFO | Processing group 290: SIC=3448.0, Year=2023,
USAspending rows=1137, Compustat rows=1
2025-08-21 17:43:22,296 | INFO | Processing group 291: SIC=2711.0, Year=2023,
USAspending rows=141, Compustat rows=5
2025-08-21 17:43:22,298 | INFO | Processing group 292: SIC=1389.0, Year=2023,
USAspending rows=561, Compustat rows=27
```

```
2025-08-21 17:43:22,305 | INFO | Processing group 293: SIC=3011.0, Year=2023,
USAspending rows=752, Compustat rows=1
2025-08-21 17:43:22,308 | INFO | Processing group 294: SIC=3081.0, Year=2023,
USAspending rows=205, Compustat rows=3
2025-08-21 17:43:22,310 | INFO | Processing group 295: SIC=3523.0, Year=2023,
USAspending rows=934, Compustat rows=12
2025-08-21 17:43:22,317 | INFO | Processing group 296: SIC=3221.0, Year=2023,
USAspending rows=64, Compustat rows=1
2025-08-21 17:43:22,318 | INFO | Processing group 297: SIC=7323.0, Year=2023,
USAspending rows=722, Compustat rows=7
2025-08-21 17:43:22,322 | INFO | Processing group 298: SIC=6513.0, Year=2023,
USAspending rows=363, Compustat rows=3
2025-08-21 17:43:22,325 | INFO | Processing group 299: SIC=6099.0, Year=2023,
USAspending rows=476, Compustat rows=11
2025-08-21 17:43:22,328 | INFO | Processing group 300: SIC=3824.0, Year=2023,
USAspending rows=1786, Compustat rows=3
2025-08-21 17:43:22,333 | INFO | Processing group 301: SIC=6792.0, Year=2023,
USAspending rows=7, Compustat rows=15
2025-08-21 17:43:22,335 | INFO | Processing group 302: SIC=8082.0, Year=2023,
USAspending rows=336, Compustat rows=8
2025-08-21 17:43:22,337 | INFO | Processing group 303: SIC=3341.0, Year=2023,
USAspending rows=1050, Compustat rows=1
2025-08-21 17:43:22,340 | INFO | Processing group 304: SIC=3411.0, Year=2023,
USAspending rows=45, Compustat rows=4
2025-08-21 17:43:22,341 | INFO | Processing group 305: SIC=4512.0, Year=2023,
USAspending rows=605, Compustat rows=28
2025-08-21 17:43:22,347 | INFO | Processing group 306: SIC=2731.0, Year=2023,
USAspending rows=225, Compustat rows=1
2025-08-21 17:43:22,349 | INFO | Processing group 307: SIC=2013.0, Year=2023,
USAspending rows=418, Compustat rows=4
2025-08-21 17:43:22,352 | INFO | Processing group 308: SIC=8111.0, Year=2023,
USAspending rows=1651, Compustat rows=3
2025-08-21 17:43:22,357 | INFO | Processing group 309: SIC=2844.0, Year=2023,
USAspending rows=624, Compustat rows=23
2025-08-21 17:43:22,360 | INFO | Processing group 310: SIC=4812.0, Year=2023,
USAspending rows=1316, Compustat rows=26
2025-08-21 17:43:22,372 | INFO | Processing group 311: SIC=4832.0, Year=2023,
USAspending rows=242, Compustat rows=13
2025-08-21 17:43:22,375 | INFO | Processing group 312: SIC=4813.0, Year=2023,
USAspending rows=2034, Compustat rows=34
2025-08-21 17:43:22,389 | INFO | Processing group 313: SIC=7819.0, Year=2023,
USAspending rows=224, Compustat rows=1
2025-08-21 17:43:22,391 | INFO | Processing group 314: SIC=3751.0, Year=2023,
USAspending rows=344, Compustat rows=11
2025-08-21 17:43:22,393 | INFO | Processing group 315: SIC=3714.0, Year=2023,
USAspending rows=1525, Compustat rows=41
2025-08-21 17:43:22,405 | INFO | Processing group 316: SIC=3843.0, Year=2023,
USAspending rows=1591, Compustat rows=7
```

```
2025-08-21 17:43:22,410 | INFO | Processing group 317: SIC=6311.0, Year=2023,
USAspending rows=133, Compustat rows=35
2025-08-21 17:43:22,412 | INFO | Processing group 319: SIC=5961.0, Year=2023,
USAspending rows=33, Compustat rows=69
2025-08-21 17:43:22,416 | INFO | Processing group 320: SIC=6411.0, Year=2023,
USAspending rows=812, Compustat rows=27
2025-08-21 17:43:22,424 | INFO | Processing group 321: SIC=2084.0, Year=2023,
USAspending rows=3, Compustat rows=5
2025-08-21 17:43:22,425 | INFO | Processing group 322: SIC=6722.0, Year=2023,
USAspending rows=4, Compustat rows=1
2025-08-21 17:43:22,426 | INFO | Processing group 323: SIC=2452.0, Year=2023,
USAspending rows=181, Compustat rows=2
2025-08-21 17:43:22,428 | INFO | Processing group 324: SIC=5045.0, Year=2023,
USAspending rows=1257, Compustat rows=11
2025-08-21 17:43:22,437 | INFO | Processing group 325: SIC=1311.0, Year=2023,
USAspending rows=16, Compustat rows=141
2025-08-21 17:43:22,439 | INFO | Processing group 326: SIC=5172.0, Year=2023,
USAspending rows=203, Compustat rows=13
2025-08-21 17:43:22,455 | INFO | Processing group 327: SIC=2082.0, Year=2023,
USAspending rows=24, Compustat rows=9
2025-08-21 17:43:22,461 | INFO | Processing group 328: SIC=2611.0, Year=2023,
USAspending rows=151, Compustat rows=1
2025-08-21 17:43:22,464 | INFO | Processing group 329: SIC=6021.0, Year=2023,
USAspending rows=207, Compustat rows=1
2025-08-21 17:43:22,469 | INFO | Processing group 330: SIC=5093.0, Year=2023,
USAspending rows=17, Compustat rows=1
2025-08-21 17:43:22,474 | INFO | Processing group 331: SIC=6163.0, Year=2023,
USAspending rows=162, Compustat rows=4
2025-08-21 17:43:22,476 | INFO | Processing group 332: SIC=6153.0, Year=2023,
USAspending rows=206, Compustat rows=9
2025-08-21 17:43:22,479 | INFO | Processing group 333: SIC=3281.0, Year=2023,
USAspending rows=132, Compustat rows=1
2025-08-21 17:43:22,482 | INFO | Processing group 334: SIC=6331.0, Year=2023,
USAspending rows=47, Compustat rows=68
2025-08-21 17:43:22,491 | INFO | Processing group 335: SIC=5099.0, Year=2023,
USAspending rows=252, Compustat rows=2
2025-08-21 17:43:22,494 | INFO | Processing group 336: SIC=2052.0, Year=2023,
USAspending rows=59, Compustat rows=1
2025-08-21 17:43:22,496 | INFO | Processing group 337: SIC=7822.0, Year=2023,
USAspending rows=107, Compustat rows=3
2025-08-21 17:43:22,497 | INFO | Processing group 338: SIC=7948.0, Year=2023,
USAspending rows=134, Compustat rows=1
2025-08-21 17:43:22,499 | INFO | Processing group 339: SIC=8351.0, Year=2023,
USAspending rows=46, Compustat rows=2
2025-08-21 17:43:22,501 | INFO | Processing group 340: SIC=2273.0, Year=2023,
USAspending rows=419, Compustat rows=3
2025-08-21 17:43:22,504 | INFO | Processing group 341: SIC=5331.0, Year=2023,
USAspending rows=71, Compustat rows=10
```

```
2025-08-21 17:43:22,505 | INFO | Processing group 342: SIC=6211.0, Year=2023,
USAspending rows=51, Compustat rows=39
2025-08-21 17:43:22,508 | INFO | Processing group 343: SIC=3524.0, Year=2023,
USAspending rows=480, Compustat rows=1
2025-08-21 17:43:22,510 | INFO | Processing group 344: SIC=2721.0, Year=2023,
USAspending rows=336, Compustat rows=4
2025-08-21 17:43:22,513 | INFO | Processing group 345: SIC=2451.0, Year=2023,
USAspending rows=111, Compustat rows=3
2025-08-21 17:43:22,516 | INFO | Processing group 346: SIC=3533.0, Year=2023,
USAspending rows=110, Compustat rows=21
2025-08-21 17:43:22,518 | INFO | Processing group 347: SIC=1381.0, Year=2023,
USAspending rows=15, Compustat rows=12
2025-08-21 17:43:22,519 | INFO | Processing group 348: SIC=4011.0, Year=2023,
USAspending rows=35, Compustat rows=5
2025-08-21 17:43:22,521 | INFO | Processing group 349: SIC=5082.0, Year=2023,
USAspending rows=64, Compustat rows=1
2025-08-21 17:43:22,522 | INFO | Processing group 350: SIC=4922.0, Year=2023,
USAspending rows=24, Compustat rows=7
2025-08-21 17:43:22,524 | INFO | Processing group 351: SIC=2085.0, Year=2023,
USAspending rows=12, Compustat rows=6
2025-08-21 17:43:22,525 | INFO | Processing group 352: SIC=7841.0, Year=2023,
USAspending rows=84, Compustat rows=2
2025-08-21 17:43:22,526 | INFO | Processing group 353: SIC=5094.0, Year=2023,
USAspending rows=28, Compustat rows=1
2025-08-21 17:43:22,527 | INFO | Processing group 354: SIC=2741.0, Year=2023,
USAspending rows=103, Compustat rows=1
2025-08-21 17:43:22,528 | INFO | Processing group 355: SIC=7996.0, Year=2023,
USAspending rows=8, Compustat rows=5
2025-08-21 17:43:22,530 | INFO | Processing group 356: SIC=3211.0, Year=2023,
USAspending rows=16, Compustat rows=1
2025-08-21 17:43:22,531 | INFO | Processing group 357: SIC=4833.0, Year=2023,
USAspending rows=78, Compustat rows=11
2025-08-21 17:43:22,533 | INFO | Processing group 358: SIC=6361.0, Year=2023,
USAspending rows=24, Compustat rows=5
2025-08-21 17:43:22,535 | INFO | Processing group 359: SIC=6141.0, Year=2023,
USAspending rows=1, Compustat rows=36
2025-08-21 17:43:22,536 | INFO | Processing group 360: SIC=7331.0, Year=2023,
USAspending rows=5, Compustat rows=1
2025-08-21 17:43:22,537 | INFO | Processing group 361: SIC=5411.0, Year=2023,
USAspending rows=27, Compustat rows=13
2025-08-21 17:43:22,538 | INFO | Processing group 362: SIC=3241.0, Year=2023,
USAspending rows=11, Compustat rows=5
2025-08-21 17:43:22,539 | INFO | Processing group 363: SIC=3652.0, Year=2023,
USAspending rows=11, Compustat rows=1
2025-08-21 17:43:22,540 | INFO | Processing group 364: SIC=6552.0, Year=2023,
USAspending rows=9, Compustat rows=15
2025-08-21 17:43:22,542 | INFO | Processing group 365: SIC=6111.0, Year=2023,
USAspending rows=7, Compustat rows=4
```

```
2025-08-21 17:43:22,543 | INFO | Processing group 366: SIC=5311.0, Year=2023,
USAspending rows=18, Compustat rows=4
2025-08-21 17:43:22,544 | INFO | Processing group 367: SIC=2024.0, Year=2023,
USAspending rows=1, Compustat rows=1
2025-08-21 17:43:22,545 | INFO | Processing group 368: SIC=1221.0, Year=2023,
USAspending rows=2, Compustat rows=1
2025-08-21 17:43:22,546 | INFO | Processing group 369: SIC=6282.0, Year=2023,
USAspending rows=4, Compustat rows=60
2025-08-21 17:43:22,548 | INFO | Processing group 370: SIC=5047.0, Year=2024,
USAspending rows=763671, Compustat rows=6
2025-08-21 17:43:24,348 | INFO | Processing group 371: SIC=2086.0, Year=2024,
USAspending rows=4668, Compustat rows=18
2025-08-21 17:43:24,362 | INFO | Processing group 372: SIC=2531.0, Year=2024,
USAspending rows=122142, Compustat rows=3
2025-08-21 17:43:24,518 | INFO | Processing group 373: SIC=2911.0, Year=2024,
USAspending rows=257388, Compustat rows=25
2025-08-21 17:43:24,955 | INFO | Processing group 374: SIC=3721.0, Year=2024,
USAspending rows=35676, Compustat rows=14
2025-08-21 17:43:25,024 | INFO | Processing group 375: SIC=5063.0, Year=2024,
USAspending rows=37089, Compustat rows=2
2025-08-21 17:43:25,068 | INFO | Processing group 376: SIC=5072.0, Year=2024,
USAspending rows=84638, Compustat rows=1
2025-08-21 17:43:25,166 | INFO | Processing group 377: SIC=5031.0, Year=2024,
USAspending rows=205994, Compustat rows=3
2025-08-21 17:43:25,344 | INFO | Processing group 378: SIC=4953.0, Year=2024,
USAspending rows=19635, Compustat rows=8
2025-08-21 17:43:25,383 | INFO | Processing group 379: SIC=2015.0, Year=2024,
USAspending rows=64030, Compustat rows=4
2025-08-21 17:43:25,475 | INFO | Processing group 380: SIC=4581.0, Year=2024,
USAspending rows=21347, Compustat rows=9
2025-08-21 17:43:25,535 | INFO | Processing group 381: SIC=2891.0, Year=2024,
USAspending rows=91507, Compustat rows=1
2025-08-21 17:43:25,650 | INFO | Processing group 382: SIC=3571.0, Year=2024,
USAspending rows=33368, Compustat rows=5
2025-08-21 17:43:25,699 | INFO | Processing group 383: SIC=3829.0, Year=2024,
USAspending rows=32404, Compustat rows=28
2025-08-21 17:43:25,815 | INFO | Processing group 384: SIC=7389.0, Year=2024,
USAspending rows=28008, Compustat rows=52
2025-08-21 17:43:26,050 | INFO | Processing group 385: SIC=2833.0, Year=2024,
USAspending rows=173275, Compustat rows=11
2025-08-21 17:43:26,250 | INFO | Processing group 386: SIC=7373.0, Year=2024,
USAspending rows=45441, Compustat rows=31
2025-08-21 17:43:26,393 | INFO | Processing group 387: SIC=5812.0, Year=2024,
USAspending rows=1313, Compustat rows=58
2025-08-21 17:43:26,412 | INFO | Processing group 388: SIC=3728.0, Year=2024,
USAspending rows=101680, Compustat rows=13
2025-08-21 17:43:26,571 | INFO | Processing group 389: SIC=4412.0, Year=2024,
USAspending rows=4656, Compustat rows=49
```

```
2025-08-21 17:43:26,582 | INFO | Processing group 390: SIC=4522.0, Year=2024,
USAspending rows=10192, Compustat rows=9
2025-08-21 17:43:26,604 | INFO | Processing group 391: SIC=4899.0, Year=2024,
USAspending rows=8736, Compustat rows=19
2025-08-21 17:43:26,633 | INFO | Processing group 392: SIC=8744.0, Year=2024,
USAspending rows=31485, Compustat rows=2
2025-08-21 17:43:26,684 | INFO | Processing group 393: SIC=8071.0, Year=2024,
USAspending rows=4914, Compustat rows=18
2025-08-21 17:43:26,704 | INFO | Processing group 394: SIC=3826.0, Year=2024,
USAspending rows=14860, Compustat rows=26
2025-08-21 17:43:26,770 | INFO | Processing group 395: SIC=3572.0, Year=2024,
USAspending rows=2068, Compustat rows=8
2025-08-21 17:43:26,776 | INFO | Processing group 396: SIC=7371.0, Year=2024,
USAspending rows=14947, Compustat rows=14
2025-08-21 17:43:26,829 | INFO | Processing group 397: SIC=5122.0, Year=2024,
USAspending rows=363868, Compustat rows=14
2025-08-21 17:43:27,416 | INFO | Processing group 398: SIC=3663.0, Year=2024,
USAspending rows=13071, Compustat rows=32
2025-08-21 17:43:27,546 | INFO | Processing group 399: SIC=3561.0, Year=2024,
USAspending rows=3190, Compustat rows=7
2025-08-21 17:43:27,557 | INFO | Processing group 400: SIC=4731.0, Year=2024,
USAspending rows=3766, Compustat rows=16
2025-08-21 17:43:27,574 | INFO | Processing group 401: SIC=6531.0, Year=2024,
USAspending rows=5243, Compustat rows=27
2025-08-21 17:43:27,597 | INFO | Processing group 402: SIC=3812.0, Year=2024,
USAspending rows=18354, Compustat rows=14
2025-08-21 17:43:27,646 | INFO | Processing group 403: SIC=2421.0, Year=2024,
USAspending rows=12727, Compustat rows=2
2025-08-21 17:43:27,669 | INFO | Processing group 404: SIC=2522.0, Year=2024,
USAspending rows=46228, Compustat rows=2
2025-08-21 17:43:27,730 | INFO | Processing group 405: SIC=2842.0, Year=2024,
USAspending rows=2566, Compustat rows=3
2025-08-21 17:43:27,736 | INFO | Processing group 406: SIC=3861.0, Year=2024,
USAspending rows=8702, Compustat rows=8
2025-08-21 17:43:27,746 | INFO | Processing group 407: SIC=3711.0, Year=2024,
USAspending rows=12802, Compustat rows=46
2025-08-21 17:43:27,804 | INFO | Processing group 408: SIC=3823.0, Year=2024,
USAspending rows=3573, Compustat rows=10
2025-08-21 17:43:27,818 | INFO | Processing group 409: SIC=7381.0, Year=2024,
USAspending rows=15718, Compustat rows=6
2025-08-21 17:43:27,848 | INFO | Processing group 410: SIC=8742.0, Year=2024,
USAspending rows=35797, Compustat rows=27
2025-08-21 17:43:28,002 | INFO | Processing group 411: SIC=8011.0, Year=2024,
USAspending rows=4633, Compustat rows=6
2025-08-21 17:43:28,014 | INFO | Processing group 412: SIC=7363.0, Year=2024,
USAspending rows=10957, Compustat rows=16
2025-08-21 17:43:28,040 | INFO | Processing group 413: SIC=2835.0, Year=2024,
USAspending rows=3827, Compustat rows=43
```

```
2025-08-21 17:43:28,059 | INFO | Processing group 414: SIC=8734.0, Year=2024,
USAspending rows=5940, Compustat rows=7
2025-08-21 17:43:28,079 | INFO | Processing group 415: SIC=3312.0, Year=2024,
USAspending rows=2043, Compustat rows=16
2025-08-21 17:43:28,086 | INFO | Processing group 416: SIC=3714.0, Year=2024,
USAspending rows=1601, Compustat rows=40
2025-08-21 17:43:28,099 | INFO | Processing group 417: SIC=7359.0, Year=2024,
USAspending rows=8337, Compustat rows=12
2025-08-21 17:43:28,123 | INFO | Processing group 418: SIC=2851.0, Year=2024,
USAspending rows=24911, Compustat rows=4
2025-08-21 17:43:28,154 | INFO | Processing group 419: SIC=4812.0, Year=2024,
USAspending rows=2810, Compustat rows=28
2025-08-21 17:43:28,169 | INFO | Processing group 420: SIC=8731.0, Year=2024,
USAspending rows=47158, Compustat rows=20
2025-08-21 17:43:28,340 | INFO | Processing group 421: SIC=1623.0, Year=2024,
USAspending rows=3735, Compustat rows=11
2025-08-21 17:43:28,357 | INFO | Processing group 422: SIC=2673.0, Year=2024,
USAspending rows=10239, Compustat rows=1
2025-08-21 17:43:28,366 | INFO | Processing group 423: SIC=8711.0, Year=2024,
USAspending rows=72337, Compustat rows=12
2025-08-21 17:43:28,528 | INFO | Processing group 424: SIC=6321.0, Year=2024,
USAspending rows=1952, Compustat rows=5
2025-08-21 17:43:28,532 | INFO | Processing group 425: SIC=4911.0, Year=2024,
USAspending rows=5087, Compustat rows=99
2025-08-21 17:43:28,602 | INFO | Processing group 426: SIC=3086.0, Year=2024,
USAspending rows=2746, Compustat rows=1
2025-08-21 17:43:28,605 | INFO | Processing group 427: SIC=3559.0, Year=2024,
USAspending rows=1326, Compustat rows=28
2025-08-21 17:43:28,619 | INFO | Processing group 428: SIC=6512.0, Year=2024,
USAspending rows=1355, Compustat rows=21
2025-08-21 17:43:28,633 | INFO | Processing group 429: SIC=7372.0, Year=2024,
USAspending rows=5295, Compustat rows=153
2025-08-21 17:43:28,902 | INFO | Processing group 430: SIC=3443.0, Year=2024,
USAspending rows=1119, Compustat rows=4
2025-08-21 17:43:28,909 | INFO | Processing group 431: SIC=3272.0, Year=2024,
USAspending rows=1058, Compustat rows=1
2025-08-21 17:43:28,912 | INFO | Processing group 432: SIC=3634.0, Year=2024,
USAspending rows=590, Compustat rows=4
2025-08-21 17:43:28,915 | INFO | Processing group 433: SIC=3562.0, Year=2024,
USAspending rows=16750, Compustat rows=5
2025-08-21 17:43:28,949 | INFO | Processing group 434: SIC=3843.0, Year=2024,
USAspending rows=1687, Compustat rows=7
2025-08-21 17:43:28,954 | INFO | Processing group 435: SIC=3531.0, Year=2024,
USAspending rows=10893, Compustat rows=4
2025-08-21 17:43:28,971 | INFO | Processing group 436: SIC=3715.0, Year=2024,
USAspending rows=1058, Compustat rows=1
2025-08-21 17:43:28,975 | INFO | Processing group 437: SIC=3851.0, Year=2024,
USAspending rows=4004, Compustat rows=6
```

```
2025-08-21 17:43:28,981 | INFO | Processing group 438: SIC=3678.0, Year=2024,
USAspending rows=10257, Compustat rows=6
2025-08-21 17:43:28,999 | INFO | Processing group 439: SIC=2834.0, Year=2024,
USAspending rows=25179, Compustat rows=238
2025-08-21 17:43:29,140 | INFO | Processing group 440: SIC=3577.0, Year=2024,
USAspending rows=2621, Compustat rows=15
2025-08-21 17:43:29,152 | INFO | Processing group 441: SIC=3613.0, Year=2024,
USAspending rows=4683, Compustat rows=4
2025-08-21 17:43:29,162 | INFO | Processing group 442: SIC=7374.0, Year=2024,
USAspending rows=5665, Compustat rows=36
2025-08-21 17:43:29,212 | INFO | Processing group 443: SIC=3317.0, Year=2024,
USAspending rows=1305, Compustat rows=4
2025-08-21 17:43:29,217 | INFO | Processing group 444: SIC=8741.0, Year=2024,
USAspending rows=3147, Compustat rows=6
2025-08-21 17:43:29,225 | INFO | Processing group 445: SIC=7011.0, Year=2024,
USAspending rows=10971, Compustat rows=15
2025-08-21 17:43:29,260 | INFO | Processing group 446: SIC=3442.0, Year=2024,
USAspending rows=1445, Compustat rows=5
2025-08-21 17:43:29,264 | INFO | Processing group 447: SIC=2511.0, Year=2024,
USAspending rows=2114, Compustat rows=2
2025-08-21 17:43:29,268 | INFO | Processing group 448: SIC=3724.0, Year=2024,
USAspending rows=9403, Compustat rows=3
2025-08-21 17:43:29,280 | INFO | Processing group 449: SIC=3541.0, Year=2024,
USAspending rows=1190, Compustat rows=1
2025-08-21 17:43:29,283 | INFO | Processing group 450: SIC=3578.0, Year=2024,
USAspending rows=4742, Compustat rows=2
2025-08-21 17:43:29,291 | INFO | Processing group 451: SIC=8093.0, Year=2024,
USAspending rows=3587, Compustat rows=5
2025-08-21 17:43:29,297 | INFO | Processing group 452: SIC=8062.0, Year=2024,
USAspending rows=14373, Compustat rows=9
2025-08-21 17:43:29,315 | INFO | Processing group 453: SIC=7361.0, Year=2024,
USAspending rows=2133, Compustat rows=15
2025-08-21 17:43:29,324 | INFO | Processing group 454: SIC=4923.0, Year=2024,
USAspending rows=2079, Compustat rows=11
2025-08-21 17:43:29,330 | INFO | Processing group 455: SIC=3669.0, Year=2024,
USAspending rows=8441, Compustat rows=11
2025-08-21 17:43:29,356 | INFO | Processing group 456: SIC=8051.0, Year=2024,
USAspending rows=3786, Compustat rows=4
2025-08-21 17:43:29,366 | INFO | Processing group 457: SIC=3824.0, Year=2024,
USAspending rows=1715, Compustat rows=2
2025-08-21 17:43:29,370 | INFO | Processing group 458: SIC=8721.0, Year=2024,
USAspending rows=3639, Compustat rows=5
2025-08-21 17:43:29,377 | INFO | Processing group 459: SIC=3621.0, Year=2024,
USAspending rows=3575, Compustat rows=15
2025-08-21 17:43:29,394 | INFO | Processing group 460: SIC=3575.0, Year=2024,
USAspending rows=2007, Compustat rows=1
2025-08-21 17:43:29,398 | INFO | Processing group 461: SIC=7323.0, Year=2024,
USAspending rows=715, Compustat rows=7
```

```
2025-08-21 17:43:29,401 | INFO | Processing group 462: SIC=3524.0, Year=2024,
USAspending rows=403, Compustat rows=1
2025-08-21 17:43:29,403 | INFO | Processing group 463: SIC=2711.0, Year=2024,
USAspending rows=173, Compustat rows=5
2025-08-21 17:43:29,404 | INFO | Processing group 464: SIC=3661.0, Year=2024,
USAspending rows=780, Compustat rows=14
2025-08-21 17:43:29,410 | INFO | Processing group 465: SIC=3674.0, Year=2024,
USAspending rows=4020, Compustat rows=104
2025-08-21 17:43:29,476 | INFO | Processing group 466: SIC=2821.0, Year=2024,
USAspending rows=591, Compustat rows=11
2025-08-21 17:43:29,481 | INFO | Processing group 467: SIC=5084.0, Year=2024,
USAspending rows=6073, Compustat rows=4
2025-08-21 17:43:29,489 | INFO | Processing group 468: SIC=5013.0, Year=2024,
USAspending rows=177, Compustat rows=2
2025-08-21 17:43:29,491 | INFO | Processing group 469: SIC=3842.0, Year=2024,
USAspending rows=3081, Compustat rows=40
2025-08-21 17:43:29,514 | INFO | Processing group 470: SIC=6311.0, Year=2024,
USAspending rows=178, Compustat rows=35
2025-08-21 17:43:29,517 | INFO | Processing group 471: SIC=3844.0, Year=2024,
USAspending rows=3013, Compustat rows=4
2025-08-21 17:43:29,521 | INFO | Processing group 472: SIC=2211.0, Year=2024,
USAspending rows=833, Compustat rows=2
2025-08-21 17:43:29,523 | INFO | Processing group 473: SIC=5051.0, Year=2024,
USAspending rows=124, Compustat rows=7
2025-08-21 17:43:29,525 | INFO | Processing group 474: SIC=5045.0, Year=2024,
USAspending rows=899, Compustat rows=10
2025-08-21 17:43:29,531 | INFO | Processing group 475: SIC=3021.0, Year=2024,
USAspending rows=814, Compustat rows=4
2025-08-21 17:43:29,533 | INFO | Processing group 476: SIC=6513.0, Year=2024,
USAspending rows=373, Compustat rows=4
2025-08-21 17:43:29,534 | INFO | Processing group 477: SIC=3448.0, Year=2024,
USAspending rows=984, Compustat rows=1
2025-08-21 17:43:29,537 | INFO | Processing group 478: SIC=4941.0, Year=2024,
USAspending rows=2558, Compustat rows=14
2025-08-21 17:43:29,551 | INFO | Processing group 479: SIC=4213.0, Year=2024,
USAspending rows=325, Compustat rows=16
2025-08-21 17:43:29,554 | INFO | Processing group 480: SIC=3651.0, Year=2024,
USAspending rows=2409, Compustat rows=12
2025-08-21 17:43:29,566 | INFO | Processing group 481: SIC=8111.0, Year=2024,
USAspending rows=1613, Compustat rows=3
2025-08-21 17:43:29,570 | INFO | Processing group 482: SIC=5331.0, Year=2024,
USAspending rows=59, Compustat rows=9
2025-08-21 17:43:29,571 | INFO | Processing group 483: SIC=7311.0, Year=2024,
USAspending rows=989, Compustat rows=16
2025-08-21 17:43:29,580 | INFO | Processing group 484: SIC=3081.0, Year=2024,
USAspending rows=172, Compustat rows=2
2025-08-21 17:43:29,581 | INFO | Processing group 485: SIC=3672.0, Year=2024,
USAspending rows=3998, Compustat rows=10
```

```
2025-08-21 17:43:29,592 | INFO | Processing group 486: SIC=4813.0, Year=2024,
USAspending rows=3522, Compustat rows=29
2025-08-21 17:43:29,608 | INFO | Processing group 487: SIC=2731.0, Year=2024,
USAspending rows=284, Compustat rows=1
2025-08-21 17:43:29,609 | INFO | Processing group 488: SIC=3751.0, Year=2024,
USAspending rows=282, Compustat rows=11
2025-08-21 17:43:29,612 | INFO | Processing group 489: SIC=3357.0, Year=2024,
USAspending rows=1561, Compustat rows=3
2025-08-21 17:43:29,616 | INFO | Processing group 490: SIC=5065.0, Year=2024,
USAspending rows=223, Compustat rows=10
2025-08-21 17:43:29,619 | INFO | Processing group 491: SIC=2836.0, Year=2024,
USAspending rows=878, Compustat rows=632
2025-08-21 17:43:29,744 | INFO | Processing group 492: SIC=2273.0, Year=2024,
USAspending rows=252, Compustat rows=3
2025-08-21 17:43:29,747 | INFO | Processing group 493: SIC=3564.0, Year=2024,
USAspending rows=1541, Compustat rows=2
2025-08-21 17:43:29,751 | INFO | Processing group 494: SIC=1389.0, Year=2024,
USAspending rows=566, Compustat rows=25
2025-08-21 17:43:29,757 | INFO | Processing group 495: SIC=5172.0, Year=2024,
USAspending rows=212, Compustat rows=13
2025-08-21 17:43:29,759 | INFO | Processing group 496: SIC=3523.0, Year=2024,
USAspending rows=709, Compustat rows=13
2025-08-21 17:43:29,765 | INFO | Processing group 497: SIC=4013.0, Year=2024,
USAspending rows=204, Compustat rows=1
2025-08-21 17:43:29,766 | INFO | Processing group 498: SIC=2033.0, Year=2024,
USAspending rows=2057, Compustat rows=3
2025-08-21 17:43:29,770 | INFO | Processing group 499: SIC=3011.0, Year=2024,
USAspending rows=748, Compustat rows=1
2025-08-21 17:43:29,773 | INFO | Processing group 500: SIC=7812.0, Year=2024,
USAspending rows=570, Compustat rows=10
2025-08-21 17:43:29,777 | INFO | Processing group 501: SIC=1531.0, Year=2024,
USAspending rows=1462, Compustat rows=22
2025-08-21 17:43:29,787 | INFO | Processing group 502: SIC=7841.0, Year=2024,
USAspending rows=28, Compustat rows=2
2025-08-21 17:43:29,788 | INFO | Processing group 503: SIC=3341.0, Year=2024,
USAspending rows=999, Compustat rows=1
2025-08-21 17:43:29,791 | INFO | Processing group 504: SIC=2611.0, Year=2024,
USAspending rows=123, Compustat rows=1
2025-08-21 17:43:29,792 | INFO | Processing group 505: SIC=5141.0, Year=2024,
USAspending rows=739, Compustat rows=1
2025-08-21 17:43:29,793 | INFO | Processing group 506: SIC=5171.0, Year=2024,
USAspending rows=666, Compustat rows=6
2025-08-21 17:43:29,797 | INFO | Processing group 508: SIC=2844.0, Year=2024,
USAspending rows=850, Compustat rows=20
2025-08-21 17:43:29,801 | INFO | Processing group 509: SIC=6411.0, Year=2024,
USAspending rows=778, Compustat rows=30
2025-08-21 17:43:29,809 | INFO | Processing group 510: SIC=2452.0, Year=2024,
USAspending rows=151, Compustat rows=2
```

```
2025-08-21 17:43:29,811 | INFO | Processing group 511: SIC=6099.0, Year=2024,
USAspending rows=589, Compustat rows=8
2025-08-21 17:43:29,814 | INFO | Processing group 512: SIC=2721.0, Year=2024,
USAspending rows=456, Compustat rows=4
2025-08-21 17:43:29,817 | INFO | Processing group 513: SIC=2013.0, Year=2024,
USAspending rows=384, Compustat rows=4
2025-08-21 17:43:29,819 | INFO | Processing group 514: SIC=1311.0, Year=2024,
USAspending rows=17, Compustat rows=125
2025-08-21 17:43:29,821 | INFO | Processing group 515: SIC=6021.0, Year=2024,
USAspending rows=166, Compustat rows=1
2025-08-21 17:43:29,823 | INFO | Processing group 516: SIC=4832.0, Year=2024,
USAspending rows=373, Compustat rows=13
2025-08-21 17:43:29,827 | INFO | Processing group 517: SIC=7822.0, Year=2024,
USAspending rows=117, Compustat rows=4
2025-08-21 17:43:29,829 | INFO | Processing group 518: SIC=6153.0, Year=2024,
USAspending rows=234, Compustat rows=10
2025-08-21 17:43:29,831 | INFO | Processing group 519: SIC=8082.0, Year=2024,
USAspending rows=352, Compustat rows=6
2025-08-21 17:43:29,834 | INFO | Processing group 520: SIC=8351.0, Year=2024,
USAspending rows=63, Compustat rows=2
2025-08-21 17:43:29,836 | INFO | Processing group 521: SIC=4512.0, Year=2024,
USAspending rows=666, Compustat rows=28
2025-08-21 17:43:29,840 | INFO | Processing group 522: SIC=2741.0, Year=2024,
USAspending rows=134, Compustat rows=1
2025-08-21 17:43:29,842 | INFO | Processing group 523: SIC=6163.0, Year=2024,
USAspending rows=120, Compustat rows=3
2025-08-21 17:43:29,843 | INFO | Processing group 524: SIC=4833.0, Year=2024,
USAspending rows=112, Compustat rows=11
2025-08-21 17:43:29,845 | INFO | Processing group 525: SIC=7819.0, Year=2024,
USAspending rows=181, Compustat rows=1
2025-08-21 17:43:29,847 | INFO | Processing group 526: SIC=6282.0, Year=2024,
USAspending rows=33, Compustat rows=55
2025-08-21 17:43:29,849 | INFO | Processing group 527: SIC=5311.0, Year=2024,
USAspending rows=36, Compustat rows=3
2025-08-21 17:43:29,850 | INFO | Processing group 528: SIC=2085.0, Year=2024,
USAspending rows=9, Compustat rows=6
2025-08-21 17:43:29,851 | INFO | Processing group 529: SIC=3281.0, Year=2024,
USAspending rows=134, Compustat rows=1
2025-08-21 17:43:29,852 | INFO | Processing group 530: SIC=5099.0, Year=2024,
USAspending rows=317, Compustat rows=2
2025-08-21 17:43:29,855 | INFO | Processing group 531: SIC=3221.0, Year=2024,
USAspending rows=89, Compustat rows=1
2025-08-21 17:43:29,857 | INFO | Processing group 532: SIC=1382.0, Year=2024,
USAspending rows=218, Compustat rows=5
2025-08-21 17:43:29,859 | INFO | Processing group 533: SIC=3533.0, Year=2024,
USAspending rows=84, Compustat rows=20
2025-08-21 17:43:29,861 | INFO | Processing group 534: SIC=7948.0, Year=2024,
USAspending rows=142, Compustat rows=1
```

```
2025-08-21 17:43:29,862 | INFO | Processing group 535: SIC=2052.0, Year=2024,
USAspending rows=58, Compustat rows=1
2025-08-21 17:43:29,864 | INFO | Processing group 536: SIC=5411.0, Year=2024,
USAspending rows=72, Compustat rows=14
2025-08-21 17:43:29,865 | INFO | Processing group 537: SIC=5082.0, Year=2024,
USAspending rows=57, Compustat rows=1
2025-08-21 17:43:29,866 | INFO | Processing group 538: SIC=6211.0, Year=2024,
USAspending rows=45, Compustat rows=39
2025-08-21 17:43:29,869 | INFO | Processing group 539: SIC=2451.0, Year=2024,
USAspending rows=99, Compustat rows=3
2025-08-21 17:43:29,871 | INFO | Processing group 540: SIC=2082.0, Year=2024,
USAspending rows=29, Compustat rows=9
2025-08-21 17:43:29,873 | INFO | Processing group 541: SIC=4011.0, Year=2024,
USAspending rows=28, Compustat rows=5
2025-08-21 17:43:29,875 | INFO | Processing group 542: SIC=1221.0, Year=2024,
USAspending rows=18, Compustat rows=1
2025-08-21 17:43:29,876 | INFO | Processing group 543: SIC=6331.0, Year=2024,
USAspending rows=52, Compustat rows=71
2025-08-21 17:43:29,879 | INFO | Processing group 544: SIC=5093.0, Year=2024,
USAspending rows=14, Compustat rows=1
2025-08-21 17:43:29,881 | INFO | Processing group 545: SIC=5961.0, Year=2024,
USAspending rows=20, Compustat rows=64
2025-08-21 17:43:29,882 | INFO | Processing group 546: SIC=4922.0, Year=2024,
USAspending rows=15, Compustat rows=8
2025-08-21 17:43:29,883 | INFO | Processing group 547: SIC=6361.0, Year=2024,
USAspending rows=29, Compustat rows=6
2025-08-21 17:43:29,885 | INFO | Processing group 548: SIC=6792.0, Year=2024,
USAspending rows=7, Compustat rows=13
2025-08-21 17:43:29,887 | INFO | Processing group 549: SIC=6552.0, Year=2024,
USAspending rows=11, Compustat rows=13
2025-08-21 17:43:29,888 | INFO | Processing group 550: SIC=3411.0, Year=2024,
USAspending rows=15, Compustat rows=4
2025-08-21 17:43:29,889 | INFO | Processing group 551: SIC=3241.0, Year=2024,
USAspending rows=15, Compustat rows=6
2025-08-21 17:43:29,890 | INFO | Processing group 552: SIC=5094.0, Year=2024,
USAspending rows=11, Compustat rows=1
2025-08-21 17:43:29,891 | INFO | Processing group 553: SIC=6111.0, Year=2024,
USAspending rows=3, Compustat rows=4
2025-08-21 17:43:29,893 | INFO | Processing group 554: SIC=7331.0, Year=2024,
USAspending rows=12, Compustat rows=1
2025-08-21 17:43:29,893 | INFO | Processing group 555: SIC=7996.0, Year=2024,
USAspending rows=3, Compustat rows=5
2025-08-21 17:43:29,894 | INFO | Processing group 556: SIC=3652.0, Year=2024,
USAspending rows=6, Compustat rows=1
2025-08-21 17:43:29,896 | INFO | Processing group 557: SIC=1381.0, Year=2024,
USAspending rows=5, Compustat rows=12
2025-08-21 17:43:29,897 | INFO | Processing group 558: SIC=2084.0, Year=2024,
USAspending rows=5, Compustat rows=4
```

```
2025-08-21 17:43:29,898 | INFO | Processing group 559: SIC=2024.0, Year=2024,
USAspending rows=2, Compustat rows=1
2025-08-21 17:43:29,899 | INFO | Processing group 560: SIC=6141.0, Year=2024,
USAspending rows=3, Compustat rows=37
2025-08-21 17:43:39,523 | INFO | Final merged shape: 1711207 rows, 55 columns
```

[17]:

```python
print(df_merged["action_date_fiscal_year"].value_counts())
```

```
action_date_fiscal_year
2022    644335
2023    559564
2024    507308
Name: count, dtype: int64
```

[18]:

```python
df_merged.to_csv('collected_data.csv', index = False)
df_merged = pd.read_csv('collected_data.csv')
df_merged
```

```
/var/folders/m4/8tn_t7fn3n999rq0tn3djx1w0000gn/T/ipykernel_7575/2167190118.py:2:
DtypeWarning: Columns (19) have mixed types. Specify dtype option on import or
set low_memory=False.
  df_merged = pd.read_csv('collected_data.csv')
```

[18]:

```
                       contract_transaction_unique_key  \
0            9700_9700_SPE2DV22F78C7_0_SPE2DV17D4001_0
1            9700_9700_SPE2D622F523K_0_SPE2DE18D0010_0
2            9700_9700_SPE2DM22FN9VG_0_SPE2DM20D9503_0
3            9700_9700_SPE2DM22FY4Z0_0_SPE2DM20D2504_0
4            9700_9700_SPE2DV22FALX8_0_SPE2DV17D6030_0
…                                                  …
1711202       7008_-NONE-_70Z08724PSTPL0002_0_-NONE-_0
1711203      1443_-NONE-_140P2123P0036_P00001_-NONE-_0
1711204      12D0_-NONE-_12FPC122P0005_P00001_-NONE-_0
1711205      9700_-NONE-_W9127N21P0149_P00002_-NONE-_0
1711206  7008_-NONE-_70Z08724PSTPL0002_P00001_-NONE-_0

                              contract_award_unique_key      award_id_piid  \
0        CONT_AWD_SPE2DV22F78C7_9700_SPE2DV17D4001_9700      SPE2DV22F78C7
1        CONT_AWD_SPE2D622F523K_9700_SPE2DE18D0010_9700      SPE2D622F523K
2        CONT_AWD_SPE2DM22FN9VG_9700_SPE2DM20D9503_9700      SPE2DM22FN9VG
3        CONT_AWD_SPE2DM22FY4Z0_9700_SPE2DM20D2504_9700      SPE2DM22FY4Z0
4        CONT_AWD_SPE2DV22FALX8_9700_SPE2DV17D6030_9700      SPE2DV22FALX8
…                                                    …                …
1711202   CONT_AWD_70Z08724PSTPL0002_7008_-NONE-_-NONE-  70Z08724PSTPL0002
1711203       CONT_AWD_140P2123P0036_1443_-NONE-_-NONE-      140P2123P0036
1711204       CONT_AWD_12FPC122P0005_12D0_-NONE-_-NONE-      12FPC122P0005
1711205       CONT_AWD_W9127N21P0149_9700_-NONE-_-NONE-      W9127N21P0149
1711206   CONT_AWD_70Z08724PSTPL0002_7008_-NONE-_-NONE-  70Z08724PSTPL0002
```

```
         federal_action_obligation  total_dollars_obligated  \
0                            38.60                    38.60   
1                           109.60                   109.60   
2                          6033.03                  6033.03   
3                            98.80                    98.80   
4                           255.75                   255.75   
…                            …                      …         
1711202                     104.00                   104.00   
1711203                       0.00                 19913.00   
1711204                    -256.00                  1988.25   
1711205                   -5000.00                  4000.00   
1711206                       0.00                   104.00   

         current_total_value_of_award  potential_total_value_of_award  \
0                               38.60                           38.60   
1                              109.60                          109.60   
2                             6033.03                         6033.03   
3                               98.80                           98.80   
4                              255.75                          255.75   
…                               …                             …         
1711202                        104.00                          104.00   
1711203                      19913.00                        19913.00   
1711204                       1988.25                         1988.25   
1711205                       4000.00                         4000.00   
1711206                        104.00                          104.00   

        action_date  action_date_fiscal_year period_of_performance_start_date  \
0        2021-11-10                     2022                       2021-11-10   
1        2021-11-10                     2022                       2021-11-10   
2        2021-11-10                     2022                       2021-11-10   
3        2021-11-10                     2022                       2021-11-10   
4        2021-11-10                     2022                       2021-11-10   
…             …                      …                               …          
1711202  2024-07-16                     2024                       2024-07-16   
1711203  2024-07-11                     2024                       2023-07-11   
1711204  2024-06-11                     2024                       2021-11-10   
1711205  2024-08-29                     2024                       2021-09-30   
1711206  2024-08-28                     2024                       2024-07-16   

         …                 company_name        at      sale      revt  \
0        …            OWENS & MINOR INC  3536.551  9785.315  9785.315   
1        …            PATTERSON COS INC  2741.630  6499.405  6499.405   
2        …            OWENS & MINOR INC  3536.551  9785.315  9785.315   
3        …            OWENS & MINOR INC  3536.551  9785.315  9785.315   
4        …            OWENS & MINOR INC  3536.551  9785.315  9785.315   
…      …                          …        …       …       …            
```

```
1711202  …  FIRST AMERICAN FINANCIAL CP  16802.800  5998.100  5998.100
1711203  …  FIRST AMERICAN FINANCIAL CP  16802.800  5998.100  5998.100
1711204  …  FIRST AMERICAN FINANCIAL CP  16802.800  5998.100  5998.100
1711205  …  FIRST AMERICAN FINANCIAL CP  16802.800  5998.100  5998.100
1711206  …  FIRST AMERICAN FINANCIAL CP  16802.800  5998.100  5998.100

              ib         lt       ceq    oancf  xrd      cogs
0        221.589   2598.050   938.501  124.177  0.0  8272.086
1        203.210   1698.995  1041.676 -980.994  NaN  5079.132
2        221.589   2598.050   938.501  124.177  0.0  8272.086
3        221.589   2598.050   938.501  124.177  0.0  8272.086
4        221.589   2598.050   938.501  124.177  0.0  8272.086
…          …        …         …       …  …        …
1711202  216.800  11940.000  4848.100  354.300  NaN  5408.100
1711203  216.800  11940.000  4848.100  354.300  NaN  5408.100
1711204  216.800  11940.000  4848.100  354.300  NaN  5408.100
1711205  216.800  11940.000  4848.100  354.300  NaN  5408.100
1711206  216.800  11940.000  4848.100  354.300  NaN  5408.100

[1711207 rows x 55 columns]
```