

PA USTIF Auction Analysis: Data Dictionary & Lineage Report

Comprehensive Documentation of ETL Pipeline and Analytical Datasets

Technical Documentation

today

Contents

1	Executive Summary	3
2	Section 1: Source Data Inventory	4
2.1	1.1 USTIF Administrative Data	4
2.1.1	1.1.1 Actuarial Contract Data	4
2.1.2	1.1.2 Individual Claims Data	4
2.2	1.2 PA DEP External Data Sources	6
2.2.1	1.2.1 PASDA ArcGIS REST APIs	6
2.2.2	1.2.2 eMapPA External Extraction	6
2.2.3	1.2.3 eFACTS Web Scraping	6
3	Section 2: Analytical Dataset Dictionary	8
3.1	2.1 Core USTIF Datasets	8
3.1.1	2.1.1 claims_clean.csv	8
3.2	Key Linkage Field	8
3.2.1	2.1.2 contracts_clean.csv (Excluded from v2.0 Pipeline)	10
3.3	2.2 PA DEP Facility Linkage	11
3.3.1	2.2.1 facility_linkage_table.csv	11
3.4	Critical Linkage Table	11
3.5	2.3 eFACTS Compliance Data	12
3.5.1	2.3.1 efacts_facility_meta.csv	12
3.5.2	2.3.2 efacts_violations.csv	12
3.5.3	2.3.3 efacts_inspections.csv	13
3.5.4	2.3.4 efacts_permits_detail.csv	13
3.5.5	2.3.5 efacts_permits_tasks.csv	13
3.6	2.4 Remediation Data	15
3.6.1	2.4.1 efacts_remediation_summary.csv	15
3.6.2	2.4.2 efacts_remediation_substances.csv	15
3.6.3	2.4.3 efacts_remediation_milestones.csv	15
4	Section 3: Data Lineage & Transformation Logic	16
4.1	3.1 Entity Relationship Diagram	16
4.2	3.2 Key Transformations by ETL Stage	16
4.2.1	Stage 1: Raw Load (01_load_ustif_data.R)	16
4.2.2	Stage 2a: PA DEP Download (02a_padep_download.R)	16
4.2.3	Stage 2b: eFACTS Scrape (02b_efacts_scrape.R)	17
4.2.4	Stage 3: Master Dataset (03_merge_master_dataset.R)	17
4.2.5	Stage 4: Analysis Panel (04_construct_analysis_panel.R)	17
4.3	3.3 Merge Key Analysis	18
4.4	3.4 Data Quality Flags	18
4.4.1	Missingness Patterns	18
4.4.2	Known Data Limitations	18

5	Appendix A: Categorical Variable Codebooks	19
5.1	A.1 claim_status Values	19
5.2	A.2 auction_type Values	19
5.3	A.3 DEP Region Values	19
6	Appendix B: File Inventory	20
6.1	B.1 Source Files	20
6.2	B.2 Processed Files	20
6.3	B.3 External Data Files	20

1 Executive Summary

This document provides authoritative documentation for the Pennsylvania Underground Storage Tank Indemnification Fund (USTIF) auction analysis data pipeline. It serves as the canonical reference for:

- 1. **Source Data Inventory:** Raw data origins, granularity, and file structures
- 2. **Analytical Dataset Dictionary:** Complete variable definitions with lineage mapping
- 3. **Transformation Logic:** ETL operations from raw \rightarrow processed datasets

Data Architecture Version: 2.0 (PA DEP facility-centric)

Primary Data Sources:

Source	Domain	Observation Unit	Records
USTIF Administrative	Claims/Contracts	Claim \times Contract	658 contracts, 7,793 claims
PA DEP ArcGIS REST	Facility Registry	Facility	38,545 facilities
PA DEP eFACTS	Compliance/Enforcement	Inspection/Violation	Variable

2 Section 1: Source Data Inventory

2.1 1.1 USTIF Administrative Data

2.1.1 1.1.1 Actuarial Contract Data

File: Actuarial_Contract_Data_2.xlsx

Sheet: Report 1

Observation Unit: Contract (one row per contract/job)

Granularity: Contract-level; multiple contracts may exist per claim

Record Count: 658

Temporal Coverage: 1995–2024

Provider: Pennsylvania Underground Storage Tank Indemnification Fund via Third-Party Administrator (ICF)

Raw Column Schema:

Raw Column Name	Data Type	Non-Null Count	Description
Contract Jobs.Claim Number	int64	658	Foreign key to claims data
Contract Job Identifier	int64	658	Primary key for contract
Adjuster Full Name	object	658	TPA staff managing claim
CTS Site Name	object	658	Site name in Claims Tracking System
Department Name	object	658	Permit number (XX-XXXXXX format)
Contract Will Bring Site to Closure? Desc	object	658	Yes/No closure flag
Consultant Full Name	object	658	Environmental consultant name
Contract Effective Date	datetime64	658	Contract start date
Contract End Date	datetime64	600	Contract completion date
Contract Category Desc	object	658	Category classification
Bid Approval Letter to Claimant	datetime64	174	Bid approval notification date
Bid Type Desc	object	238	Auction format (SOW/Bid-to-Result)
Contract Type Desc	object	656	Contract mechanism type
Contract Base Price	float64	618	Initial contracted amount
Amount Paid to Date	float64	658	Disbursed amount (\$)
Notes	object	386	Free-text notes

2.1.2 1.1.2 Individual Claims Data

File: Actuarial_UST_Individual_Claim_Data_thru_63020_4.xlsx

Sheet: Report 1 (skip first 3 rows for header)

Observation Unit: Claim (one row per claim event)

Granularity: Claim-level; represents insurance claim for contamination incident

Record Count: 7,793

Temporal Coverage: 1988–2025

Provider: USTIF Actuarial Reports

Raw Column Schema:

Raw Column Name	Data Type	Non-Null Count	Description
Event Department Abbreviation	object	7,792	Permit number (linkage key)
Claim Number	object	7,793	Primary key for claim
Paid Loss	float64	7,793	Direct remediation payments
Incurred Loss	float64	7,793	() PaidALAE float64 2,382 Allocat Total incurred (paid + reserved)
Date of Loss Reported	datetime64	7,792	Release discovery date
Date of Claim	datetime64	7,792	USTIF claim filing date
Claim Status (as of 6/10/2025)	object	7,792	Current claim status
Date Closed	datetime64	7,045	Claim closure date
DEP Region	object	7,792	PA DEP regional office
Claimant Full Name	object	7,792	Tank owner/claimant
County	object	7,792	Pennsylvania county
Event Primary Loc. Desc	object	7,791	Location type classification
Product(s)	object	7,792	Petroleum products stored
Product Type Other Description	object	562	Other product specification

2.2 1.2 PA DEP External Data Sources

2.2.1 1.2.1 PASDA ArcGIS REST APIs

Endpoint Base: <https://mapservices.pasda.psu.edu/server/rest/services/pasda/>

Update Frequency: Monthly

Authentication: None (public API)

2.2.1.1 Active Storage Tanks (Layer 27) URL: .../DEP/MapServer/27/query

Observation Unit: Active registered facility

Records: 11,339

Key Fields: `attributes_facility_i` (permit number), `attributes_primary_fa` (eFACTS ID), `attributes_site_id`, `coordinates`

2.2.1.2 Inactive Storage Tanks (Layer 20) URL: .../DEP2/MapServer/20/query

Observation Unit: Closed/removed facility

Records: 33,765

Key Fields: Same schema as active tanks

2.2.1.3 Land Recycling Cleanup Sites (Layer 18) URL: .../DEP/MapServer/18/query

Observation Unit: Act 2 cleanup site

Records: 23,312

Key Fields: Site hierarchy (organization → client → site → facility)

2.2.2 1.2.2 eMapPA External Extraction

Endpoint Base: <https://gis.dep.pa.gov/depgisprd/rest/services/emappa/>

Provides: Alternative schema with full column names (not truncated)

Layer	Records	Notes
Active Tanks (118)	11,334	Full column names preserved
Inactive Tanks (119)	33,777	Full column names preserved
Land Recycling (26)	3,762	Act 2 program sites

2.2.3 1.2.3 eFACTS Web Scraping

Base URL: <https://www.ahs.dep.pa.gov/eFACTSWeb/>

Authentication: None (public interface)

Scraping Method: HTTP GET with HTML parsing (rvest)

Rate Limiting: 1.5s between facility pages, 0.75s between detail pages

Output Tables:

Table	Observation Unit	Est. Records
<code>efacts_facility_meta</code>	Facility	38,545
<code>efacts_tanks</code>	Sub-facility (tank)	~100,000
<code>efacts_inspections</code>	Inspection event	~200,000
<code>efacts_violations</code>	Violation × Enforcement	~50,000
<code>efacts_permits_detail</code>	Authorization	~80,000

Table	Observation Unit	Est. Records
efacts_permits_tasks	Task milestone	~150,000
efacts_remediation_summary	Incident	~20,000
efacts_remediation_substances	Substance \times Incident	~40,000
efacts_remediation_milestones	Milestone	~60,000

3 Section 2: Analytical Dataset Dictionary

3.1 2.1 Core USTIF Datasets

3.1.1 2.1.1 claims_clean.csv

Purpose: Primary claims dataset for cost outcome analysis

Observation Unit: Insurance claim

Records: 7,793

ETL Source Script: R/etl/01_load_ustif_data.R

Lineage: Actuarial_UST_Individual_Claim_Data_thru_63020_4.xlsx → cleaning → derived variables

3.2 Key Linkage Field

department (permit number in XX-XXXXX format) links to **facility_linkage_table.permit_number** with 93.7% match rate.

Variable	Type	Definition	Allowable Values	Raw Source → Transformation
department	character	Facility permit number	XX-XXXXX format	Event Department Abbreviation → <code>clean_names()</code> → <code>str_trim()</code>
claim_number	character	Unique claim identifier	Numeric string	Claim Number → <code>filter(!is.na)</code>
paid_loss	numeric	Direct remediation payments	0	Paid Loss → direct copy
paid_alae	numeric	Allocated loss adjustment expenses	0, NA	Paid ALAE → direct copy
incurred_loss	numeric	Total incurred (paid + reserve)	0	Incurred Loss → direct copy
loss_reported_date		Release discovery date	1988-01-01 to present	Date of Loss Reported → <code>as.Date()</code>
claim_date	date	USTIF claim filing date	1988-01-01 to present	Date of Claim → <code>as.Date()</code>
claim_status	character	Current claim status	Closed Eligible, Closed Withdrawn, Open Pending, Open Appealed, Open Eligible, Closed NA, Closed, NA	Claim Status (as of 6/10/2025) → <code>clean_names()</code>
closed_date	date	Claim closure date	Date or NA	Date Closed → <code>as.Date()</code>
dep_region	character	PA DEP regional office	6 regions	DEP Region → <code>str_trim()</code>
claimant_name	character	Tank owner/claimant	Free text	Claimant Full Name → direct copy
county	character	Pennsylvania county	67 PA counties	County → <code>str_to_title()</code>
location_desc	character	Location type	Private, Commercial, Local Government, etc.	Event Primary Loc. Desc → direct copy
products	character	Petroleum products	35 combinations	Product(s) → direct copy
product_other	character	Other product description	Free text, NA	Product Type Other Description → direct copy
claim_year	integer	Year of claim filing	1988–2025	DERIVED: <code>year(claim_date)</code>

Variable	Type	Definition	Allowable Values	Raw Source → Transformation
loss_year	integer	Year of loss discovery	1988–2025	DERIVED: year(loss_reported_date)
closed_year	integer	Year of closure	1988–2025, NA	DERIVED: year(closed_date)
total_paid	numeric	Total disbursement	0	DERIVED: coalesce(paid_loss, 0) + coalesce(paid_alae, 0)
is_closed	logical	Claim closed indicator	TRUE/FALSE	DERIVED: !is.na(closed_date) str_detect(claim_status, "closed")
is_open	logical	Claim open indicator	TRUE/FALSE	DERIVED: str_detect(claim_status, "open")
claim_duration_days	numeric	Days from claim to closure	0, NA	DERIVED: difftime(closed_date, claim_date, units="days")
claim_duration_years	numeric	Years from claim to closure	0, NA	DERIVED: claim_duration_days / 365.25

3.2.1 2.1.2 contracts_clean.csv (Excluded from v2.0 Pipeline)

Purpose: Contract/auction records for procurement analysis

Observation Unit: Contract/job

Records: 658

ETL Source Script: R/etl/01_load_ustif_data.R

Note: While documented, the tank construction data (`tanks_clean.csv`) is **EXCLUDED** from v2.0 analysis due to limited facility coverage (18.7% match rate). Use `facility_linkage_table.csv` instead.

Variable	Type	Definition	Allowable Values	Raw Source → Transformation
<code>claim_number</code>	character	Foreign key to claims	Numeric string	<code>Contract Jobs.Claim Number</code> → <code>as.character()</code>
<code>contract_id</code>	character	Unique contract ID	Numeric string	<code>Contract Job Identifier</code> → <code>as.character()</code>
<code>adjuster</code>	character	TPA adjuster name	16 unique names	<code>Adjuster Full Name</code> → direct copy
<code>site_name</code>	character	CTS site name	Free text	<code>CTS Site Name</code> → direct copy
<code>department</code>	character	Permit number	XX-XXXXX format	<code>Department Name</code> → direct copy
<code>brings_to_closure</code>	character	Closure flag text	Yes/No	<code>Contract Will Bring Site to Closure? Desc</code> → direct copy
<code>consultant</code>	character	Environmental consultant	79 unique firms	<code>Consultant Full Name</code> → direct copy
<code>contract_start</code>	date	Contract effective date	1995–2024	<code>Contract Effective Date</code> → <code>as.Date()</code>
<code>contract_end</code>	date	Contract completion date	Date or NA	<code>Contract End Date</code> → <code>as.Date()</code>
<code>contract_category</code>	character	Contract category	2 categories	<code>Contract Category Desc</code> → direct copy
<code>bid_approval_date</code>	date	Bid approval notification	Date or NA	<code>Bid Approval Letter to Claimant</code> → <code>as.Date()</code>
<code>bid_type</code>	character	Auction format type	Bid To Result, Scope of Work, NA	<code>Bid Type Desc</code> → direct copy
<code>contract_type</code>	character	Contract mechanism	4 types	<code>Contract Type Desc</code> → direct copy
<code>base_price</code>	numeric	Initial contract amount	0	<code>Contract Base Price</code> → direct copy
<code>amendments_total</code>	numeric	Amendment value	0	<code>Total Price of Amendments</code> → direct copy
<code>paid_to_date</code>	numeric	Disbursed amount	0	<code>Amount Paid to Date</code> → direct copy
<code>contract_year</code>	integer	Contract start year	1995–2024	DERIVED: <code>year(contract_start)</code>
<code>total_contract_value</code>	numeric	Total contract value	0	DERIVED: <code>coalesce(base_price, 0) + coalesce(amendments_total, 0)</code>
<code>is_bid_to_result</code>	logical	Bid-to-Result indicator	TRUE/FALSE	DERIVED: <code>str_detect(bid_type, "bid\\ result")</code> (case-insensitive)
<code>is_scope_of_work</code>	logical	SOW indicator	TRUE/FALSE	DERIVED: <code>str_detect(bid_type, "scope\\ sow")</code> (case-insensitive)
<code>auction_type</code>	character	Auction classification	Scope of Work, Bid-to-Result, Other/Unknown	DERIVED: Conditional on <code>is_scope_of_work</code> , <code>is_bid_to_result</code>
<code>brings_to_closure_flag</code>	logical	Closure flag	TRUE/FALSE	DERIVED: <code>str_detect(brings_to_closure, "yes\\ true\\ closure")</code>

3.3 2.2 PA DEP Facility Linkage

3.3.1 2.2.1 facility_linkage_table.csv

Purpose: Master facility crosswalk linking permit numbers to eFACTS IDs

Observation Unit: Unique registered facility

Records: 38,545

ETL Source Script: R/etl/02a_padep_download.R

Lineage: pasda_tanks_active pasda_tanks_inactive → standardize columns → deduplicate (prefer active status)

3.4 Critical Linkage Table

This is the **PRIMARY** facility reference for v2.0 architecture. Links USTIF claims (via **permit_number**) to all PA DEP subsystems (via **efacts_facility_id**).

Variable	Type	Definition	Allowable Values	Raw Source → Transformation
permit_number	character	Facility permit (PK)	XX-XXXXX format	attributes_facility_i / attributes_facility_id
efacts_facility_id	integer	eFACTS database ID (FK)	Positive integer	attributes_primary_fa / attributes_primary_facility_id
site_id	integer	PA DEP site identifier	Positive integer	attributes_site_id → direct copy
facility_name	character	Registered facility name	Free text	attributes_facility_n / attributes_facility_name
address	character	Street address	Free text	attributes_facility_a / attributes_facility_address1
city	character	City	PA municipalities	attributes_facility_c / attributes_facility_city
municipality	character	Municipality name	PA municipalities	attributes_facility_m / attributes_facility_municipality
zip	character	ZIP code	5 or 9 digit	attributes_facility_z / attributes_facility_zip
latitude	numeric	Latitude coordinate	39.7–42.3 (PA range)	attributes_latitude → direct copy
longitude	numeric	Longitude coordinate	-80.5 to -74.7 (PA range)	attributes_longitude → direct copy
owner_id	integer	Tank owner ID	Positive integer	attributes_tank_owner / attributes_tank_owner_id
owner_name	character	Tank owner name	Free text	attributes_tank_own_1 / attributes_tank_owner_name
registration_status	character	Active/Inactive flag	active, inactive	DERIVED: Source file indicator

3.5 2.3 eFACTS Compliance Data

3.5.1 2.3.1 efacts_facility_meta.csv

Purpose: Facility-level metadata from eFACTS system

Observation Unit: Facility

ETL Source Script: R/etl/02b_efacts_scrape.R

Variable	Type	Definition	Allowable Values	Scrape Source
efacts_facility_id	character	eFACTS ID (PK)	Numeric string	URL parameter
facility_id	character	Alternate facility ID	Numeric string	#ContentPlaceholder2_DetailsView → “Facility ID” row
facility_name	character	Facility name	Free text	DetailsView → “Facility Name” row
address	character	Street address	Free text	DetailsView → “Address” row
status	character	Facility status	Active, Inactive, etc.	DetailsView → “Status” row
program	character	Regulatory program	Storage Tanks, etc.	DetailsView → “Program” row
batch_import_date	datetime	Scrape timestamp	ISO datetime	Sys.time() at batch save

3.5.2 2.3.2 efacts_violations.csv

Purpose: Violation and enforcement records

Observation Unit: Violation × Enforcement action

ETL Source Script: R/etl/02b_efacts_scrape.R (State machine parser)

Variable	Type	Definition	Allowable Values	Scrape Source
inspection_id	character	Parent inspection ID	Numeric string	URL parameter from inspection link
efacts_facility_id	character	Facility ID (FK)	Numeric string	Parent facility context
violation_id	character	Violation identifier	Numeric string	table.GridViewTable[border='2'] → first column
violation_date	character	Violation date	Date string	Table row → second column
description	character	Violation description	Free text	Table row → third column
resolution	character	Resolution description	Free text	“Resolution:” row
citation	character	PA Code citation	Legal citation format	“PA Code Legal Citation:” parsing
violation_type	character	Violation classification	Category string	“Violation Type:” parsing
enforcement_id	character	Enforcement action ID	Numeric string, NA	Nested enforcement table
enf_type	character	Enforcement type	Warning, NOV, COA, etc.	“Enforcement Type:” field
penalty_assessed	character	Penalty amount assessed	Dollar amount string	“Penalty Amount Assessed:” field
penalty_collected	character	Penalty collected	Dollar amount string	“Total Amount Collected:” field
date_executed	character	Execution date	Date string	“Date Executed:” field

Variable	Type	Definition	Allowable Values	Scrape Source
penalty_final_date	character	Penalty finalization date	Date string	“Penalty Final Date:” field
total_amount_due	character	Amount due	Dollar amount string	“Total Amount Due:” field
taken_against	character	Enforcement target	Entity name	“Taken Against:” field
on_appeal	character	Appeal status	Yes/No	“On Appeal?” field
penalty_status	character	Penalty status	Status string	“Penalty Status:” field
enforcement_status	character	Enforcement status	Open/Closed	“Enforcement Status:” field
num_violations_addressed	integer	Count addressed	Integer string	“# of Violations Addressed” field
batch_import_date	datetime	Scrape timestamp	ISO datetime	Batch save time

3.5.3 2.3.3 efacts_inspections.csv

Purpose: Inspection event records

Observation Unit: Inspection

Variable	Type	Definition	Allowable Values	Scrape Source
inspection_type	character	Inspection type with ID	“Type (ID)” format	#ContentPlaceHolder2_GridView3
inspection_date	character	Inspection date	Date string	Table column
result	character	Inspection result	Pass/Fail/etc.	Table column
efacts_facility_id	character	Facility ID (FK)	Numeric string	Parent context
batch_import_date	datetime	Scrape timestamp	ISO datetime	Batch save time

3.5.4 2.3.4 efacts_permits_detail.csv

Purpose: Authorization/permit details

Observation Unit: Authorization

Variable	Type	Definition	Allowable Values	Scrape Source
authorization_id	character	Authorization ID	Numeric string	DetailsView parsing
permit_number	character	Permit number	XX-XXXXX format	DetailsView → “Permit Number”
site	character	Site name	Free text	DetailsView → “Site”
client	character	Client name	Free text	DetailsView → “Client”
authorization_type	character	Auth type	Registration, etc.	DetailsView → “Authorization Type”
application_type	character	Application type	New/Renewal/etc.	DetailsView → “Application Type”
authorization_is_for	character	Authorization purpose	Free text	DetailsView → “Authorization Is For”
date_received	character	Receipt date	Date string	DetailsView → “Date Received”
status	character	Authorization status	Active/Expired/etc.	DetailsView → “Status”
auth_id	character	Internal auth ID	Numeric string	URL parameter
efacts_facility_id	character	Facility ID (FK)	Numeric string	Parent context
batch_import_date	datetime	Scrape timestamp	ISO datetime	Batch save time

3.5.5 2.3.5 efacts_permits_tasks.csv

Purpose: Permit task milestones

Observation Unit: Task

Variable	Type	Definition	Allowable Values	Scrape Source
<code>efacts_facility_id</code>	character	Facility ID (FK)	Numeric string	Parent context
<code>auth_id</code>	character	Parent authorization ID	Numeric string	Parent auth context
<code>task_internal_id</code>	character	Composite task ID	<code>auth_id_task_index</code>	Generated
<code>task</code>	character	Task name	Free text	<code>table.StaticTable</code> → “Task” column
<code>start_date</code>	character	Task start date	Date string	Table column
<code>target_date</code>	character	Target completion date	Date string	Table column
<code>completion_date</code>	character	Actual completion date	Date string, NA	Table column
<code>batch_import_date</code>	datetime	Scrape timestamp	ISO datetime	Batch save time

3.6 2.4 Remediation Data

3.6.1 2.4.1 efacts_remediation_summary.csv

Purpose: Remediation incident summary

Observation Unit: Incident/Release

Variable	Type	Definition	Allowable Values	Scrape Source
incident_name	character	Incident identifier	Free text with ID	#ContentPlaceHolder2_GridView4
confirmed_release_date	character	Release confirmation date	Date string	Table column
type	character	Incident type	UST Release, etc.	Table column
cleanup_status	character	Cleanup status	Open, Closed, etc.	Table column
cleanup_status_date	character	Status date	Date string	Table column
efacts_facility_id	character	Facility ID (FK)	Numeric string	Parent context
batch_import_date	datetime	Scrape timestamp	ISO datetime	Batch save time

3.6.2 2.4.2 efacts_remediation_substances.csv

Purpose: Substances released per incident

Observation Unit: Substance × Incident

Variable	Type	Definition	Allowable Values	Scrape Source
incident_name	character	Parent incident name	Free text	Detail page table
confirmed_release_date	character	Release date	Date string	Table column
incident_id	character	Incident ID	Numeric string	Table column
incident_type	character	Type classification	UST Release, etc.	Table column
cleanup_status	character	Cleanup status	Status string	Table column
cleanup_status_date	character	Status date	Date string	Table column
substance_released	character	Substance name	Chemical name	#ContentPlaceHolder2_GridView1
environmental_impact	character	Impact type	Groundwater, Soil, etc.	Table column
efacts_facility_id	character	Facility ID (FK)	Numeric string	Parent context
lrpact_id	character	LRPACT ID	Numeric string	URL parameter
substance_internal_id	character	Composite ID	lrpact_id_sub_index	Generated
batch_import_date	datetime	Scrape timestamp	ISO datetime	Batch save time

3.6.3 2.4.3 efacts_remediation_milestones.csv

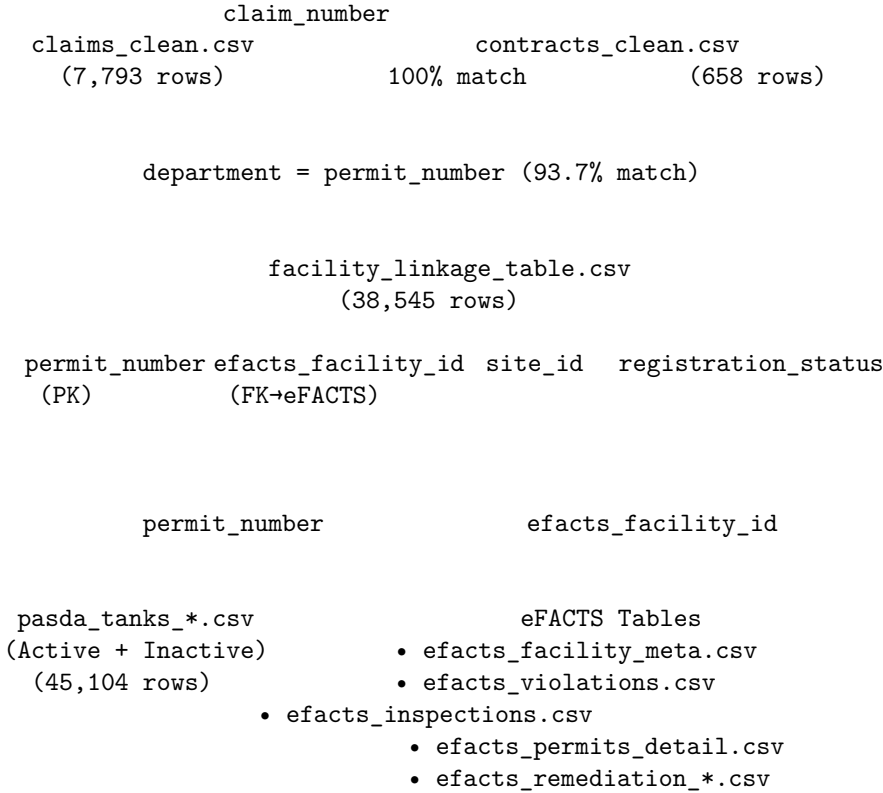
Purpose: Remediation progress milestones

Observation Unit: Milestone

Variable	Type	Definition	Allowable Values	Scrape Source
milestone_name	character	Milestone name	Free text	#ContentPlaceHolder2_GridView2
milestone_event_date	character	Event date	Date string	Table column
milestone_due_date	character	Due date	Date string	Table column
i_milestone_status	character	Milestone status	Complete, Pending, etc.	Table column
milestone_response_date	character	Response date	Date string	Table column
efacts_facility_id	character	Facility ID (FK)	Numeric string	Parent context
lrpact_id	character	Parent LRPACT ID	Numeric string	Parent context
milestone_internal_id	character	Composite ID	lrpact_id_mil_index	Generated
batch_import_date	datetime	Scrape timestamp	ISO datetime	Batch save time

4 Section 3: Data Lineage & Transformation Logic

4.1 3.1 Entity Relationship Diagram



4.2 3.2 Key Transformations by ETL Stage

4.2.1 Stage 1: Raw Load (01_load_ustif_data.R)

Operation	Input	Output	Logic
Column standardization	Raw Excel columns	snake_case names	janitor::clean_names()
Date conversion	Excel serial/datetime	R Date class	as.Date()
Derived cost metric	paid_loss, paid_alae	total_paid	coalesce(paid_loss,0) + coalesce(paid_alae,0)
Duration calculation	claim_date, closed_date	claim_duration_years	difftime(...) / 365.25
Auction type classification	bid_type	auction_type	Regex pattern matching

4.2.2 Stage 2a: PA DEP Download (02a_padep_download.R)

Operation	Input	Output	Logic
API pagination	ArcGIS REST endpoint	Full feature set	resultOffset iteration
Column standardization	attributes_* columns	Clean names	Handle PASDA truncation
Facility deduplication	Active + Inactive sets	Unique facilities	Prefer active status

Operation	Input	Output	Logic
Linkage table creation	Combined facilities	<code>facility_linkage_table</code>	<code>group_by(permit_number)</code> <code>%>% slice(1)</code>

4.2.3 Stage 2b: eFACTS Scrape (02b_efacts_scrape.R)

Operation	Input	Output	Logic
HTML parsing	eFACTS web pages	Structured tables	<code>rvest::html_table()</code>
Schema enforcement	Raw scraped data	Canonical columns	<code>enforce_schema()</code> function
Violation state machine	Nested HTML tables	Flat violation records	Row-by-row state tracking
Checkpoint/resume	Progress state	Recovery file	RDS checkpoint every 100 facilities

4.2.4 Stage 3: Master Dataset (03_merge_master_dataset.R)

Operation	Input	Output	Logic
Claims-contracts join	<code>claims, contracts</code>	Merged claims	<code>left_join(by="claim_number")</code>
Contract aggregation	Multiple contracts/claim	One row/claim	<code>group_by(claim_number)</code> <code>%>% summarise()</code>
Cost harmonization	Multiple cost fields	<code>total_cost</code>	<code>pmax(incurred_loss, total_paid)</code>
Era classification	<code>claim_year</code>	era factor	5-year bins

4.2.5 Stage 4: Analysis Panel (04_construct_analysis_panel.R)

Operation	Input	Output	Logic
Treatment definition	<code>auction_type</code>	<code>is_PFP</code>	<code>auction_type == "Bid-to-Result"</code>
IV construction	Adjuster assignments	<code>adjuster_leniency</code>	Leave-one-out mean PFP rate
Covariate imputation	Tank characteristics	Imputed values	Median imputation with missingness flags
Sample restriction	Full panel	Analysis subset	Closed claims, <code>total_paid > \$1,000</code>

4.3 3.3 Merge Key Analysis

Join	Left Key	Right Key	Match Rate	Notes
claims → contracts	claim_number	claim_number	100% (design)	Many claims have zero contracts
claims → facility_linkage	department	permit_number	93.7%	6.3% legacy/unregistered permits
facility_linkage → eFACTS	efacts_facility_id	efacts_facility_id	100%	All registered have eFACTS pages
facility_linkage → PASDA active	permit_number	attributes_facility_i	29.4%	Active subset
facility_linkage → PASDA inactive	permit_number	attributes_facility_i	87.6%	Inactive subset

4.4 3.4 Data Quality Flags

4.4.1 Missingness Patterns

Dataset	Variable	Missing %	Handling
claims_clean	paid_alae	69.4%	Coalesced to 0 in total_paid
claims_clean	closed_date	9.6%	NA retained; is_closed derived separately
contracts_clean	bid_type	63.8%	NA = non-auction contract
contracts_clean	contract_end	8.8%	Ongoing contracts
facility_linkage	latitude/longitude	<1%	NA retained

4.4.2 Known Data Limitations

1. **Tank construction data excluded:** tanks_clean.csv provides only 18.7% coverage of claims; use facility_linkage_table instead
2. **eFACTS scrape incompleteness:** Scraping may not capture all historical records
3. **Claim-facility linkage:** 6.3% of claims have legacy permit formats not in current registry
4. **Temporal gaps:** Some years have sparse contract records pre-2003

5 Appendix A: Categorical Variable Codebooks

5.1 A.1 claim_status Values

Value	Definition
Closed Eligible	Claim closed, eligible for coverage
Closed Withdrawn	Claim withdrawn by claimant
Closed NA	Closed, not applicable
Closed	Generic closed status
Open Pending	Open, awaiting review
Open Appealed	Open, under appeal
Open Eligible	Open, eligible for coverage

5.2 A.2 auction_type Values

Value	Definition	Scoring Weight
Scope of Work	Detailed specifications; cost heavily weighted	Cost > Technical
Bid-to-Result	Outcome-based; technical emphasized	Technical > Cost
Other/Unknown	Non-auction or unclassified contracts	N/A

5.3 A.3 DEP Region Values

Value	Counties Covered
PADEP Southcentral Regional Office	Adams, Bedford, Blair, Cumberland, Dauphin, Franklin, Fulton, Huntingdon, Juniata, Lancaster, Lebanon, Mifflin, Perry, York
PADEP Southwest Regional Office	Allegheny, Armstrong, Beaver, Cambria, Fayette, Greene, Indiana, Somerset, Washington, Westmoreland
PADEP Northwest Regional Office	Butler, Clarion, Crawford, Elk, Erie, Forest, Jefferson, Lawrence, McKean, Mercer, Venango, Warren
PADEP Northcentral Regional Office	Bradford, Cameron, Centre, Clearfield, Clinton, Columbia, Lycoming, Montour, Northumberland, Potter, Snyder, Sullivan, Tioga, Union
PADEP Southeast Regional Office	Bucks, Chester, Delaware, Montgomery, Philadelphia
PADEP Northeast Regional Office	Berks, Carbon, Lackawanna, Lehigh, Luzerne, Monroe, Northampton, Pike, Schuylkill, Susquehanna, Wayne, Wyoming

6 Appendix B: File Inventory

6.1 B.1 Source Files

File	Location	Size	Records
Actuarial_Contract_Data.xlsx	data/contracts/	~100KB	658
Actuarial_UST_Individuals/Claims_Data_thru_63020_4.xlsx	data/claims/	2MB	7,793

6.2 B.2 Processed Files

File	Location	Records	Primary Key
claims_clean.csv	data/processed/	7,793	claim_number
contracts_clean.csv	data/processed/	658	contract_id
facility_linkage_tables	data/external/padep/	38,545	permit_number

6.3 B.3 External Data Files

File	Source	Records
pasda_tanks_active.csv	PASDA ArcGIS	11,339
pasda_tanks_inactive.csv	PASDA ArcGIS	33,765
pasda_cleanup_sites.csv	PASDA ArcGIS	23,312
emappa_tanks_active.csv	eMapPA	11,334
emappa_tanks_inactive.csv	eMapPA	33,777
emappa_land_recycling.csv	eMapPA	3,762
efacts_facility_meta.csv	eFACTS scrape	Variable
efacts_violations.csv	eFACTS scrape	Variable
efacts_inspections.csv	eFACTS scrape	Variable
efacts_permits_detail.csv	eFACTS scrape	Variable
efacts_permits_tasks.csv	eFACTS scrape	Variable
efacts_tanks.csv	eFACTS scrape	Variable
efacts_remediation_summary.csv	eFACTS scrape	Variable
efacts_remediation_substances.csv	eFACTS scrape	Variable
efacts_remediation_milestones.csv	eFACTS scrape	Variable

Document Version: 1.0

Generated: `r Sys.Date()`

Pipeline Version: v2.0 (PA DEP facility-centric architecture)