Methodology: SWPA Municipal Grant Revenue Analysis

Source code

The source code for this project is publicly available on GitHub at:
https://github.com/maxachis/pa_municipal_annual_financial_report_scraper

Purpose

This document describes the methodology used to collect, process, and analyze municipal grant revenue data for 12 counties in southwestern Pennsylvania. The analysis produces a comprehensive dataset of federal, state, and local grant revenues received by individual municipalities over a nine-year period (2015--2023), enabling comparisons across municipalities, grant types, and time.

The resulting reports are designed to support regional planning and policy analysis by answering questions such as:

- How much grant revenue does each municipality receive annually, and from which level of government?
- How do aggregate grant revenues compare across municipalities of different types and sizes?
- What share of a municipality's total revenue comes from grants, and how does per-capita grant funding vary across the region?

Geographic scope and unit of analysis

Unit of analysis: Individual municipalities (boroughs, townships, and cities) as defined by the Pennsylvania Department of Community and Economic Development (DCED).

Geographic scope: 12 counties in the southwestern Pennsylvania region:

- Allegheny
- Armstrong
- Beaver
- Butler
- Cambria
- Fayette
- Greene

- Indiana
- Lawrence
- Somerset
- Washington
- Westmoreland

Temporal scope: Fiscal years 2015 through 2023, divided into two reporting periods:
- **Period 1:** 2015--2019 (5 years)
- **Period 2:** 2020--2023 (4 years)

This division reflects the structure of the original SWPA Grant Revenue dataset, which covered 2015--2019 for a subset of counties. Period 2 extends the analysis with newly collected data.

Data sources

Primary source: PA DCED Municipal Annual Financial Reports (AFRs)

Pennsylvania municipalities are required to file Annual Financial Reports with the Department of Community and Economic Development. These reports are publicly available through the DCED's Municipal Statistics portal:

- **URL:** https://apps.dced.pa.gov/munstats-public/ReportInformation2.aspx?report=mAfrForm
- **Format:** Downloadable Excel workbooks (one per municipality per year)
- **Key data location:** "Sheet3" of each workbook contains revenue line items with standardized code numbers, labels, and total amounts

Each AFR workbook includes financial line items organized by a standardized code system. The grant revenue codes used in this analysis, and their corresponding line items as they appear across all 5,970 AFR files in the dataset, are:

Federal grants (code prefix `351` -- summed to produce federal grant totals):

Code	Label
351.03	Highways and Streets
351.09	Community Development
351.00	All Other Federal Capital and Operating Grants

State grants (code prefix `354` -- summed to produce state grant totals):

Code	Label
354.03	Highways and Streets
354.09	Community Development
354.15	Recycling / Act 101
354.00	All Other State Capital and Operating Grants
355.08	Local Share Assessment/Gaming Proceeds

Local grants (code prefix `357` -- summed to produce local grant totals):

Code	Label
357.03	Highways and Streets
357.00	All Other Local Governmental Units Capital and Operating Grants

The AFR also contains code-355 line items ("State Shared Revenues and Entitlements"), which represent formula-based distributions rather than competitive or project-based grants. These are **not** included in the grant totals, with the exception of 3.55.08 :"Local Share Assessment," which represents a competitive pool of funds from gaming revenues.

Code	Label
355.01	Public Utility Realty Tax (PURTA)
355.02-355.03	Motor Vehicle Fuel Tax (Liquid Fuels Tax) and State Road Turnback
355.04	Alcoholic Beverage Licenses
355.05	General Municipal Pension System State Aid
355.07	Foreign Fire Insurance Tax Distribution
355.09	Marcellus Shale Impact Fee Distribution
355.00	All Other State Shared Revenues and Entitlements

The total revenue figure is also extracted from each report (column 10 of Sheet3). Not every municipality reports values for every code in a given year; codes with zero or no activity are sometimes omitted from the AFR entirely.

Supplementary sources

DCED Municipality Classification

Municipality type classifications (e.g., First Class Township, Borough, Third Class City) are obtained from DCED and stored as a reference dataset. These classifications are used to contextualize grant revenue patterns by municipality type.

U.S. Census Bureau Population Estimates

Population estimates are obtained from the U.S. Census Bureau API and linked to municipalities by geographic identifier (GEOID). These estimates are used to calculate per-capita grant revenue metrics. Population data is matched to the appropriate year for each reporting period.

Supplementary source: Urban/Rural Classification

Municipalities are classified as urban or rural based on the Pennsylvania DCED's 2022 classification of municipalities. This classification provides an additional dimension for analyzing grant revenue distribution patterns.

Data collection

Web scraping process

AFR workbooks are collected via automated web scraping using Playwright, a browser automation framework. The scraper navigates the DCED Municipal Statistics portal through a three-level hierarchy:

1. **County selection:** Each of the 12 target counties is selected from a dropdown menu
2. **Municipality selection:** For each county, every available municipality is selected in turn
3. **Year selection:** For each municipality, reports are requested for each year in the study period (2015--2023)

For each county/municipality/year combination, the scraper:

- Selects the appropriate dropdown values on the DCED web form
- Triggers the report display
- Waits for the report to render (with timeout handling)
- Downloads the resulting Excel workbook

- Records the scraping outcome (success, timeout, or "no AFR available") in a tracking database

The scraper includes rate-limiting to avoid overwhelming the DCED server. Due to the volume of reports (~5,400+ individual municipality-year combinations), the full scraping process takes several hours and may require multiple runs to handle intermittent timeouts.

Error handling during collection

The scraper tracks three categories of outcomes for each municipality-year combination:

- **Success:** Report downloaded and saved
- **No report available:** The DCED portal indicates no AFR was filed for that municipality-year
- **Error:** A timeout or access error occurred during download

Municipalities with no reports or errors are documented and flagged for transparency in the final output.

Data processing

Excel file parsing

Each downloaded AFR workbook is processed by extracting data from Sheet3. The extraction process:

1. Opens the workbook and reads Sheet3
2. For each row, reads the code (column 1), label (column 2), and total amount (column 10)
3. Validates each row against an expected pattern using regular expression matching
4. Classifies each code by its prefix into federal (351), state (354), or local (357) grant categories
5. Stores the code, label, category, and total amount in the database

Database storage

Processed data is stored in a normalized relational database (SQLite via SQLAlchemy) with the following core structure:

- **counties:** County identifiers and names
- **municipalities:** Municipality identifiers, names, county associations, and urban/rural classification
- **annual_reports:** One record per municipality-year filing, linking to county and municipality
- **code:** Standardized revenue codes with labels and category classifications
- **report_details:** Individual line-item totals per report per code
- **total_revenue:** Aggregate total revenue per report

Database views pre-aggregate grant totals by type:
- **federal_total_per_report:** Sum of all code-351 line items per report
- **state_total_per_report:** Sum of all code-354 line items per report
- **local_total_per_report:** Sum of all code-357 line items per report

Handling of invalid municipalities

Municipalities that are flagged as invalid (e.g., due to systematic data issues) are excluded from all report queries via a `flag_invalid_municipalities` table. This exclusion is applied consistently across all three report tabs.

Data quality and limitations

Missing and incomplete data

Not all municipalities file AFRs for every year. The extent of missing data is documented in full in the project's data notes. Key summary:
- **Period 1 (2015-2019):** 4 municipalities filed no reports at all; 12 additional municipalities have at least one missing year
- **Period 2 (2020-2023):** 14 municipalities filed no reports at all; 21 additional municipalities have at least one missing year

The municipalities missing include:

2015-2019 (4 municipalities with no reports):
1. Canoe Twp (Indiana County)
2. Portersville Boro (Butler County)
3. Eastvale Boro (Beaver County)
4. Frankfort Springs Boro (Beaver County)

2020-2023 (14 municipalities with no reports):
1. Eastvale Boro (Beaver)
2. Benson Boro (Somerset)
3. Homewood Boro (Beaver)
4. Shanksville Boro (Somerset)
5. Portersville Boro (Butler)
6. Fayette City Boro (Fayette)
7. Ashville Boro (Cambria)
8. Madison Boro (Westmoreland)
9. Fallston Boro (Beaver)
10. New Eagle Boro (Washington)
11. Wilmore Boro (Cambria)
12. South New Castle Boro (Lawrence)

13. Frankfort Springs Boro (Beaver)
14. New Alexandria Boro (Westmoreland)

Anomalous access errors

A small number of municipality-year combinations produced errors when accessed on the DCED portal, distinct from reports that are simply not available. These are tracked separately and documented within the repository.

Data reporting quality

Some municipalities report zero revenue across all categories despite having filed an AFR. For example, Allegheny County's McDonald Borough AFRs from 2015--2019 report $0 across all revenue types. These reports are included in the dataset as filed but may not reflect actual municipal finances.

Known data corrections

- McDonald Borough was initially misclassified under Westmoreland County rather than Washington County in certain records. This has been corrected.

Limitations

- **Filing compliance:** The analysis can only include data from municipalities that file AFRs. Non-filing municipalities are absent from the dataset entirely, which may introduce bias---particularly if non-filing municipalities share characteristics (e.g., smaller size, fewer resources).
- **Data entry quality:** AFR data is self-reported by municipalities. Errors in municipal data entry are passed through to this analysis.
- **Code consistency:** While the AFR code system is standardized, municipalities may not use all codes consistently. The "XX" suffix codes (e.g., 351.XX, 354.XX, 357.XX) serve as catch-all categories and may capture varying types of grants across municipalities.
- **Population estimates:** Per-capita calculations rely on Census Bureau population estimates, which carry margins of error. These margins are included in the output but not propagated through the per-capita calculations.
- **Temporal alignment:** Population estimates are point-in-time snapshots, while revenue data covers a full fiscal year. Minor misalignment between population reference dates and fiscal year boundaries is expected but unlikely to materially affect results.
- **Web scraping brittleness:** The data collection process depends on the structure of the DCED web portal. Changes to the portal may require updates to the scraper.
- **Municipal revenue focus:** This dataset reflects municipal revenue from grant sources only. It does not share information about grant dollars awarded to CDCs, businesses, authorities, or any other groups within community jurisdictions.

Data validation

The following validation checks are performed to ensure data integrity:

- **Spot-check verification:** High-value entries are manually verified against source AFR data on the DCED portal
- **Cross-reference with domain experts**: Results shared with SWPA to confirm data results are in line with expected priors.
- **Missing report confirmation:** Municipalities flagged as missing reports are manually spot-checked on the DCED portal
- **Municipality count verification:** The count of municipalities per county in the database is compared against the count listed on the DCED Annual Report web page
- **Code verification:** Revenue codes in the database are spot-checked against codes displayed in AFRs on the DCED portal
- **Population spot-checks:** Population estimates in the database are compared against Census API results

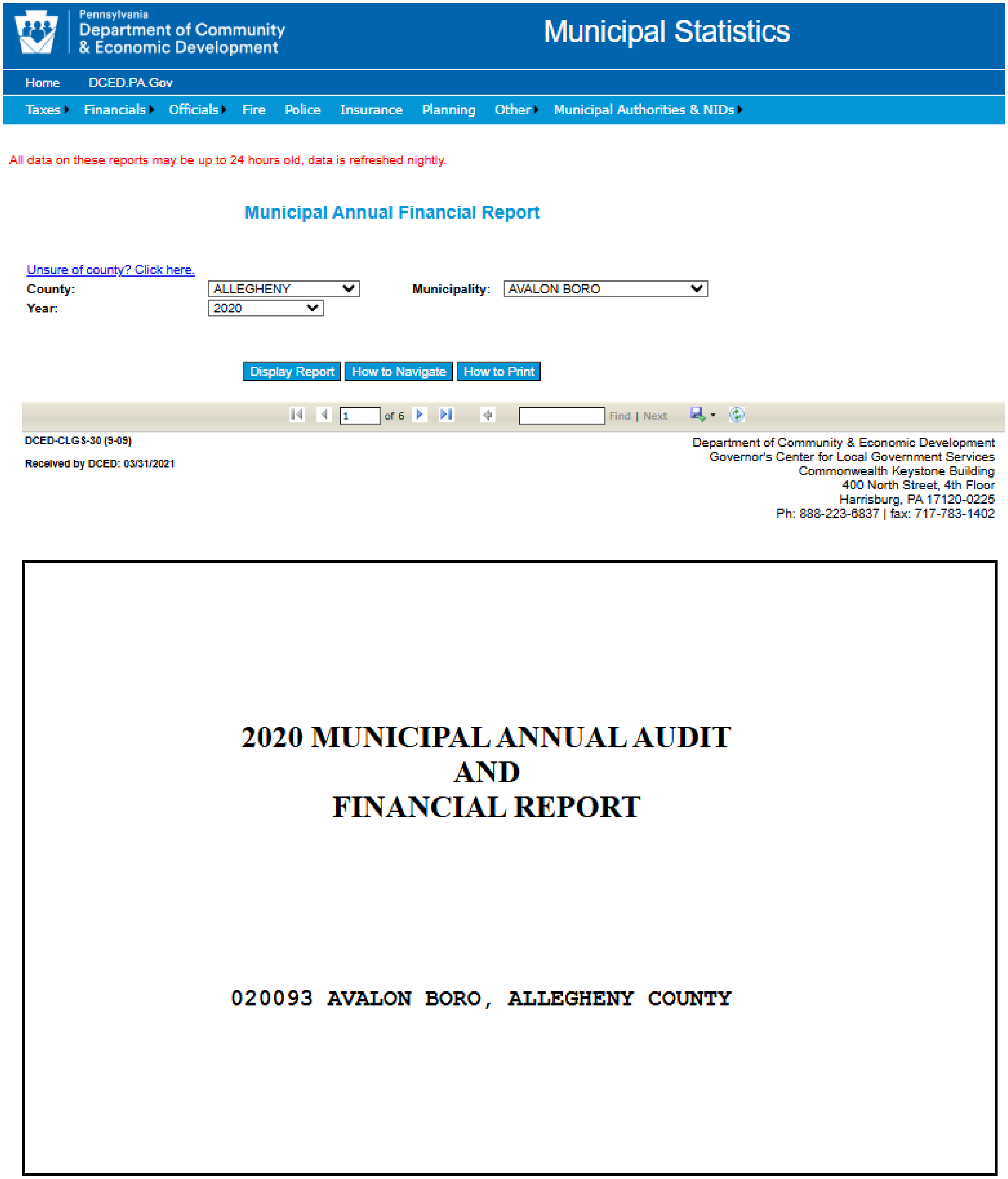
Technical implementation

Component	**Technology**
Web scraping	Playwright (browser automation)
Data processing	Python, openpyxl (Excel reading)
Database	SQLite, SQLAlchemy ORM, Alembic (migrations)
Data analysis	Polars, Pandas
Data validation	Pandera (schema validation), Pydantic (model validation)
Report generation	xlsxwriter (Excel writing)
Census data	U.S. Census Bureau API via aiohttp

Sources

- Pennsylvania Department of Community and Economic Development. "Municipal Annual Financial Report." Municipal Statistics. https://apps.dced.pa.gov/munstats-public/ReportInformation2.aspx?report=mAfrForm
- Pennsylvania Department of Community and Economic Development. Municipality classification data.

- U.S. Census Bureau. Population Estimates Program. https://www.census.gov/programs-surveys/popest.html
- Pennsylvania Department of Community and Economic Development. "List of Municipalities." Urban/rural classification, 2022.



A shot of the DCED Municipal Financial Report Page.

A repetitive process like this is ripe for automation, and that’s exactly what I did.

41		**Federal**								
42	351.03	Highways and Streets								
43	351.09	Community Development								
44	351.00	All Other Federal Capital and Operating Grants								
45	352.01	National Forest								
46	352.00	All Other Federal Shared Revenue and Entitlements								
47	353.00	Federal Payments in Lieu of Taxes								
48		**Total Federal**								
49										
50		**State**								
51	354.03	Highways and Streets	5,453							5,453
52	354.09	Community Development								
53	354.15	Recycling / Act 101								
54	354.00	All Other State Capital and Operating Grants								
55	355.01	Public Utility Realty Tax (PURTA)	920							920
56	355.02-355.03	Motor Vehicle Fuel Tax (Liquid Fuels Tax) and State Road Turnback		39,426						39,426
57	355.04	Alcoholic Beverage Licenses	200							200
58	355.05	General Municipal Pension System State Aid	11,763							11,763
59	355.07	Foreign Fire Insurance Tax Distribution	12,215							12,215
60	355.08	Local Share Assessment/Gaming Proceeds								
61	355.09	Marcellus Shale Impact Fee Distribution	178							178
62	355.00	All Other State Shared Revenues and Entitlements								
63	356.00	State Payments in Lieu of Taxes								
64		**Total State**	30,729	39,426						70,155
65										
66		**Local Government Units**								
67	357.03	Highways and Streets	8,697							8,697
68	357.00	All Other Local Governmental Units Capital and Operating Grants								
69	358.00	Local Government Unit Shared Payments for Contracted Intergovernmental Services				565,412				565,412
70	359.00	Local Governmental Units and Authorities Payments in Lieu of Taxes	216,241							216,241
71		**Total Local Government Units**	224,938			565,412				790,350

An example of report formatting, from the Aleppo Township, Allegheny County 2015 Annual Financial Report