Chicago Inspections & Licenses Audit

Gregory Bandy

University of Chicago - Data Engineering

Executive Summary

This project examines the relationship between the City of Chicago’s food inspection and business licensing datasets to determine whether inspection outcomes, specifically violations and failures, affect the status of business licenses. Through the creation of a relational database, data normalization, and visual analytics in Python, the link between the datasets was analyzed to identify potential gaps in regulatory feedback between the City of Chicago's agencies and departments. The analysis highlights discrepancies between inspection failures and updates to licensing status, raising questions about the integration of compliance systems. The lack of clear coordination between inspection results and licensing outcomes suggests a need for automated enforcement actions, improved data integration, and strong compliance systems.

Research Objectives

The primary objectives of this project were to:

- Investigate if inspection outcomes (e.g., violations or failed inspections) impact the licensing status of businesses.
- Assess the strength and accuracy of linkages between inspections and licenses based on license ID, business name, or location.
- Support public policy through data-driven insights into regulatory effectiveness.

To address the objectives, the following business questions were formulated:

- **Linkage validity**: Can inspections reliably be matched to licenses?
- **Enforcement feedback**: Do failed inspections lead to revoked/suspended/inactive licenses?
- **Compliance gaps**: Are businesses operating without valid licenses?

Data source: City of Chicago Open Data Portal filtered to years 2020 to 2025:

- Food Inspections Dataset
- Business Licenses Dataset

Methodology

The data engineering analysis of the inspections and business licenses dataset was conducted using a combination of Python and MySQL tools. For extraction, transformation, and loading, a data pipeline was built using Python. MySQL was then utilized for storage and query exploration & execution. The key steps included: filtering datasets to the last five years for performance and relevance, structuring data into normalized tables, matching inspections with licenses based on license ID and location, investigating mismatches and gaps in license-inspection relationships, and adding surrogate keys to support referential integrity.

Relational Database Schema

Five normalized (3NF+) relational tables were created from the two datasets. Each table below satisfies the Third Normal Form because in each, all non-key attributes depend solely on the table's primary key.

The businesses table contains the fields account_number, legal_name, dba_name, and business_id, where business_id serves as the primary key. This table is in 3NF because each non-key attribute depends solely on the primary key and not on any other non-key attributes. While account_number may appear in other tables, it is not a determinant within this table, and no transitive dependencies exist between columns.

Column	Description	Notes
business_id	Unique ID for each business	**Primary Key**
account_number	Business account reference	Can help link to licenses

dba_name	"Doing business as" name	Useful for business-facing names
legal_name	Registered legal entity name	Useful for formal/business docs

Table 1: Businesses 3rd Normal Form Table with entity descriptions.

The licenses table includes license_id, account_number, license_number, license_code, description, license_status, application_type, start_date, end_date, business_id, and location_id. license_id is the primary key. This table is in 3NF because all attributes are dependent only on license_id, and there are no transitive dependencies. While license_code and description could potentially be extracted into a separate lookup table, their presence here does not violate 3NF unless the description varies inconsistently with the license_code. These columns also do not affect the analysis, so they could be dropped from the table without changing the results or conclusions.

Column	**Description**	**Notes**
license_id	Unique identifier for each license	**Primary Key**: ensure uniqueness
license_number	External/official license number issued by the city	Could be a candidate key; possibly used in business processes
license_code	Coded value representing license type	Can be linked to description; optionally normalize to a lookup table

description	Human-readable description of the license_code	May be redundant if license_code is repeated often
license_status	Current status of the license (e.g., ACTIVE, EXPIRED)	Useful for filtering valid or expired licenses
application_type	Type of application submitted (e.g., RENEWAL, NEW)	Useful for trends or filtering
start_date	Date the license became active	Datetime format; useful for time-based analysis
end_date	Date the license expired or is set to expire	Can be NULL for active licenses; used for filtering
business_id	ID linking to the business associated with the license	Foreign key to businesses table
location_id	ID linking to the geographic location of the licensed entity	Foreign key to locations table

Table 2: Licenses 3rd Normal Form Table with entity descriptions.

The inspections table contains inspection_id, inspection_uid, inspection_date, inspection_type, result, risk_level, facility_type, and location_id. Originally, inspection_id was to be the primary key. However, inspection_uid was introduced as a surrogate key during data cleaning because it was discovered that inspection_id is not unique and consistent in the dataset. This meant that inspection_uid was a necessary surrogate key to ensure uniqueness and atomicity of the database. The table satisfies 3NF because all non-key fields describe the inspection and are directly dependent on the primary key without any transitive relationships.

Column	Description	Notes
inspection_uid	Unique surrogate identifier	**Primary Key**
inspection_id	non -unique identifier for the inspection	Same ID can be used for follow-up inspections
license_id	ID of the license being inspected	Foreign key to licenses table
inspection_date	Date the inspection took place	Required for time-series/trend analysis
inspection_type	Type/category of inspection (e.g., CANVASS, COMPLAINT)	Useful for filtering/analysis
result	Outcome of the inspection (e.g., PASS, FAIL, OUT OF BUSINESS)	Core analytical value — ties to violations
risk_level	Risk rating of the establishment	Categorical; often tied to enforcement strategy
facility_type	Type of facility inspected (e.g., Restaurant, School)	Can be standardized; useful for grouping
location_id	Geographic location of the inspected site	Foreign key to locations table

Table 3: Inspection 3rd Normal Form Table with entity descriptions.

The violations table consists of violation_id, violation_code, description, and inspection_uid. Each row represents a single violation tied to an inspection. This table is in 3NF because all attributes are fully dependent on the primary key (violation_id), and inspection_id and inspection_uid serve as foreign keys for linkage, not as determinants of other non-key attributes. There are no transitive dependencies among the fields.

Column	Description	Notes
violation_id	Surrogate unique ID	**Primary Key**
violation_code	Specific violation that occurred during inspection (text format)	Descriptive information of fault
description	Detailed information for violation	Added information in non-standard format and reporting
inspection_uid	Unique inspection id	Foreign Key from inspections

Table 4: Violations 3rd Normal Form Table with entity descriptions.

Finally, the locations table includes the following fields: address, city, state, zip, and location_id, with location_id serving as the primary key. Each row uniquely identifies a geographic location. The table is in 3NF since all location-specific fields depend solely on the primary key, and no field depends on another non-key field.

Together, these five tables form a fully normalized relational schema in 3NF. Each entity is well-scoped, and relationships are established through properly defined foreign keys, ensuring data integrity, eliminating redundancy, and efficient query performance.

EER Diagram

An Enhanced Entity-Relationship diagram was created to visualize entity relationships and highlight the relational structure. It shows normalized linkages between businesses, inspections, licenses, and violations. It is used to model the database structure and confirms:

- One business has many licenses
- One license is associated with one location

- One location can be tied to many licenses and many inspections
- One inspection belongs to one location
- One inspection has many violations

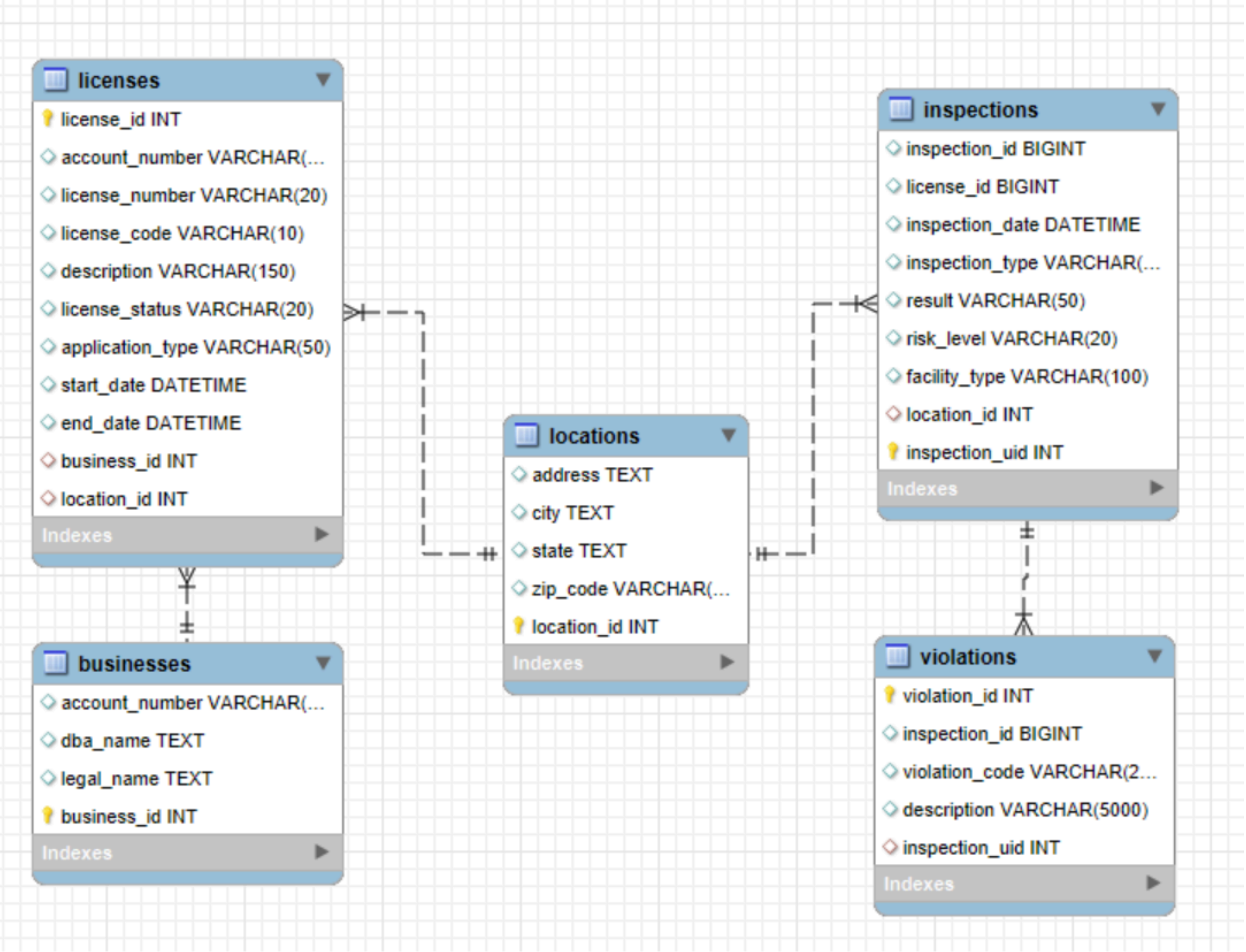


Image 1: EER Diagram of the Chicago Food Inspections and Business Licenses Relational Database.

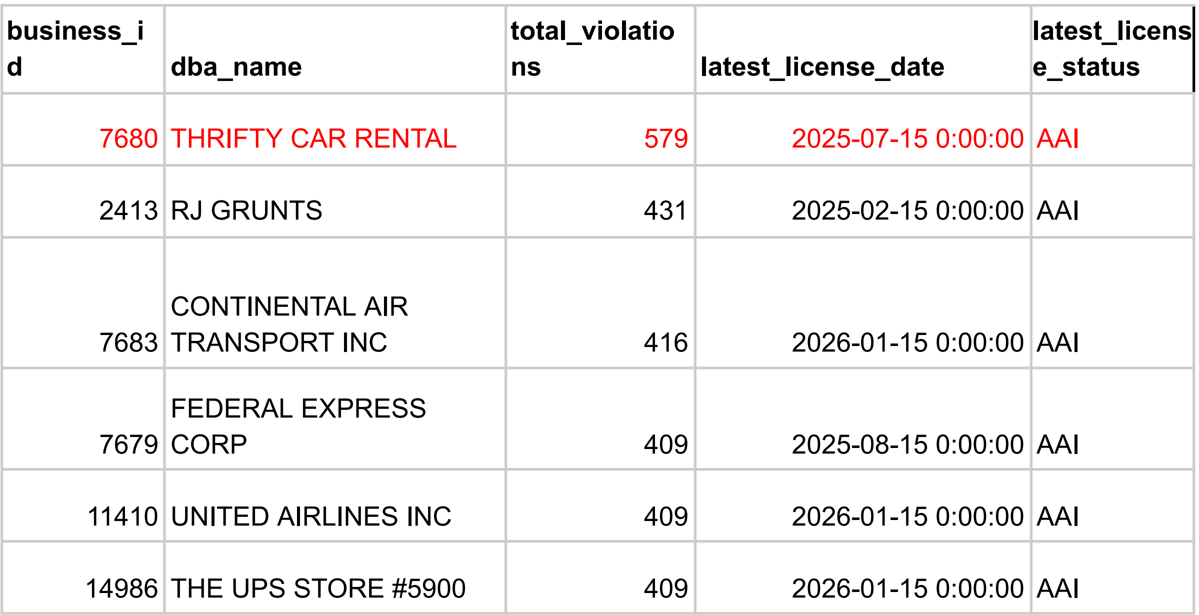
Data Analysis and Visualization

To determine the degree of linkage between inspections and license enforcement, we wrote a set of SQL queries and developed Python-based visualizations. These analyses were used to reveal whether violations had any observable effect on business licensing and to identify broader patterns of noncompliance and regulatory inconsistency.

SQL Query Insights

Top Violators:

This query identified businesses with the highest number of violations per inspection. The results showed entities such as THRIFTY CAR RENTAL and RJ GRUNTS retaining active licenses despite hundreds of violations. This directly addresses the enforcement feedback question, revealing that violations do not reliably lead to license suspension or revocation.



business_id	dba_name	total_violations	latest_license_date	latest_license_status
7680	THRIFTY CAR RENTAL	579	2025-07-15 0:00:00	AAI
2413	RJ GRUNTS	431	2025-02-15 0:00:00	AAI
7683	CONTINENTAL AIR TRANSPORT INC	416	2026-01-15 0:00:00	AAI
7679	FEDERAL EXPRESS CORP	409	2025-08-15 0:00:00	AAI
11410	UNITED AIRLINES INC	409	2026-01-15 0:00:00	AAI
14986	THE UPS STORE #5900	409	2026-01-15 0:00:00	AAI

8910	MR. QUILES MEXICAN FOOD #2	242	2023-07-15 0:00:00	AAI
20155	O'BRIENS RESTAURANT & BAR	201	2024-01-15 0:00:00	AAI
377	SHERATON CHICAGO HOTEL/TOWERS	162	2025-03-15 0:00:00	AAC
10158	JAZZ BAR	135	2022-01-15 0:00:00	AAI
11947	THE FAT SHALLOT, LLC	129	2025-05-15 0:00:00	AAI
19239	KONA ICE OF NILES	126	2023-09-15 0:00:00	AAI
12621	EL AZTECA SANCHEZ #1	121	2024-08-15 0:00:00	AAI
787	RAINBOW #643	119	2026-04-15 0:00:00	AAI
1515	LIDS #6423	119	2023-03-15 0:00:00	AAI
19382	POMPEI TACOS LLC	116	2024-09-15 0:00:00	AAI
1740	CHEESIE'S TRUCK	105	2024-12-15 0:00:00	AAI
2320	Harold's Chicken Express # 55	104	2025-11-15 0:00:00	AAI
17629	JAPAN AIRLINES	99	2022-04-15 0:00:00	AAI

Table 5: Query results of the top violators and their license status

Inspection Results vs. License Status

By grouping inspections by result and current license status, this analysis showed that the majority of failed inspections occurred under active licenses (AAI). Surprisingly, statuses such as AAC (applied) also appeared in successful inspections, suggesting unclear or inconsistent status updates. This insight reinforces that inspection outcomes are not effectively influencing license status changes. Even statuses that have been revoked (REV) have passed inspections, further emphasizing the inconsistent status update between inspections and licenses.

license_status	**inspection_result**	**total**
AAC	fail	120
AAC	pass	462
AAI	fail	2532
AAI	pass	9649
REV	fail	5
REV	pass	5

Table 6: Query result for inspection results compared with license statuses.

Businesses Operating Without Active Licenses

This query quantified how many businesses were operating without a currently active license. The result — over 15,000 businesses — indicates a major compliance gap. This addresses the compliance gaps question by demonstrating that inspections are being conducted without verifying the licensing status of establishments.

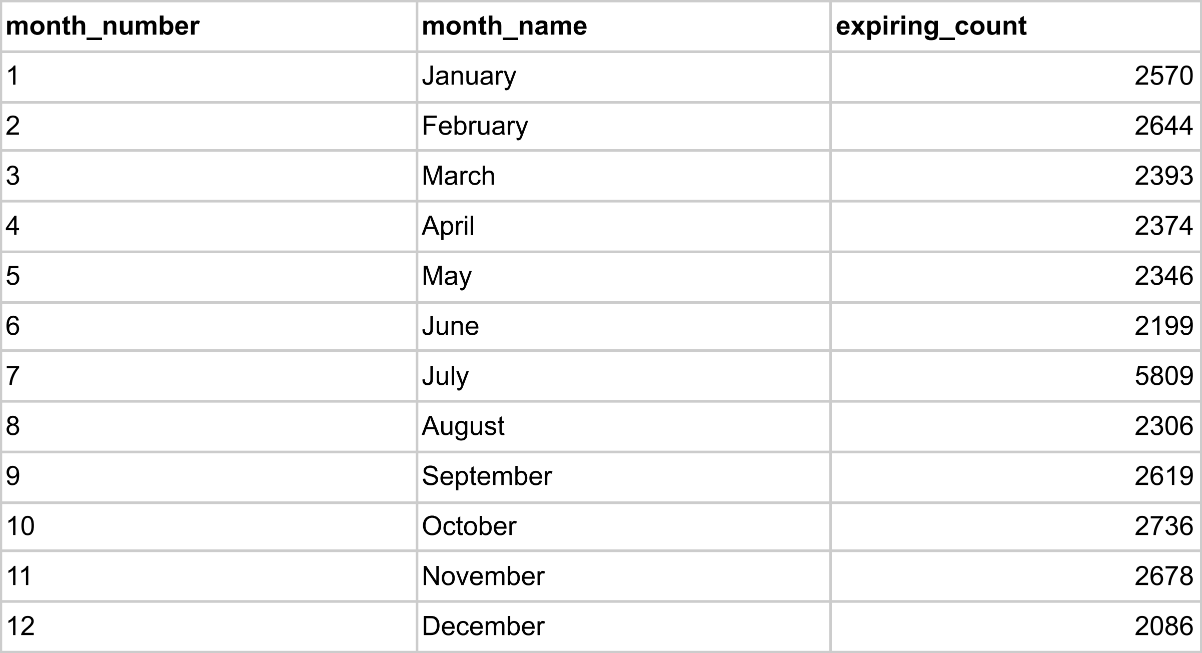
license_status	business_count
Without Active License	15972
With Active License	6161

Table 7: Query result for license statuses of businesses that are currently operating.

○

Licenses Expiring by Month (All Years)

Using date aggregation, we identified patterns in license expiration by month. A significant peak in July suggests that expiration cycles are not evenly distributed. This temporal insight could help in resource planning and aligns with the broader goal of understanding how well inspections are synchronized with licensing cycles.

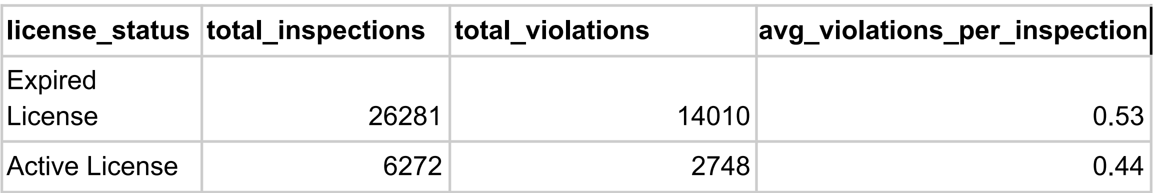


month_number	month_name	expiring_count
1	January	2570
2	February	2644
3	March	2393
4	April	2374
5	May	2346
6	June	2199
7	July	5809
8	August	2306
9	September	2619
10	October	2736
11	November	2678
12	December	2086

Table 8: Query result for licenses expiring each month aggregated over the past 5 years.

Correlation of License Expiration with Violation Rate

This analysis calculated the average number of violations per inspection for active and expired licenses. Higher violation rates among expired licenses suggest increased risk and decreased oversight. It also supports the compliance gap hypothesis, showing that expired licenses do not prevent ongoing violations.



license_status	total_inspections	total_violations	avg_violations_per_inspection
Expired License	26281	14010	0.53
Active License	6272	2748	0.44

Table 8: Query result for license expiration and violation rate.

Visualizations Developed:

Each visualization below not only complements the associated SQL findings but also helps stakeholders, such as policymakers and regulatory agencies, intuitively digest complex patterns.

Bar Chart – Top Violators and Avg Violations per Inspection

Highlights top violators by total and average violations, visualizing repeat offenders.

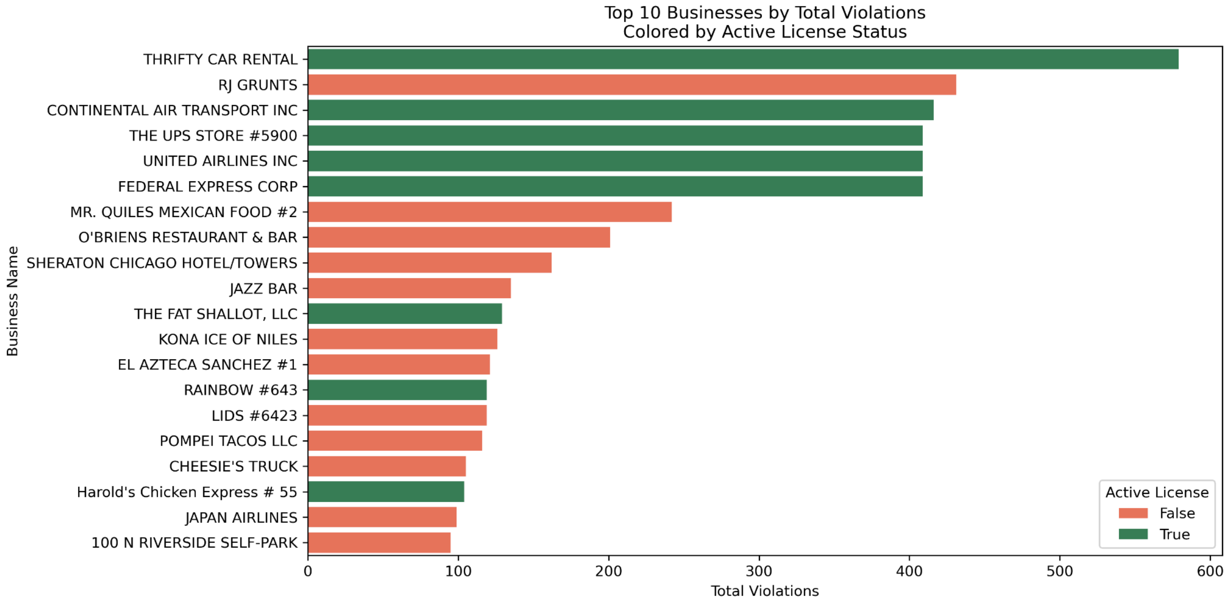
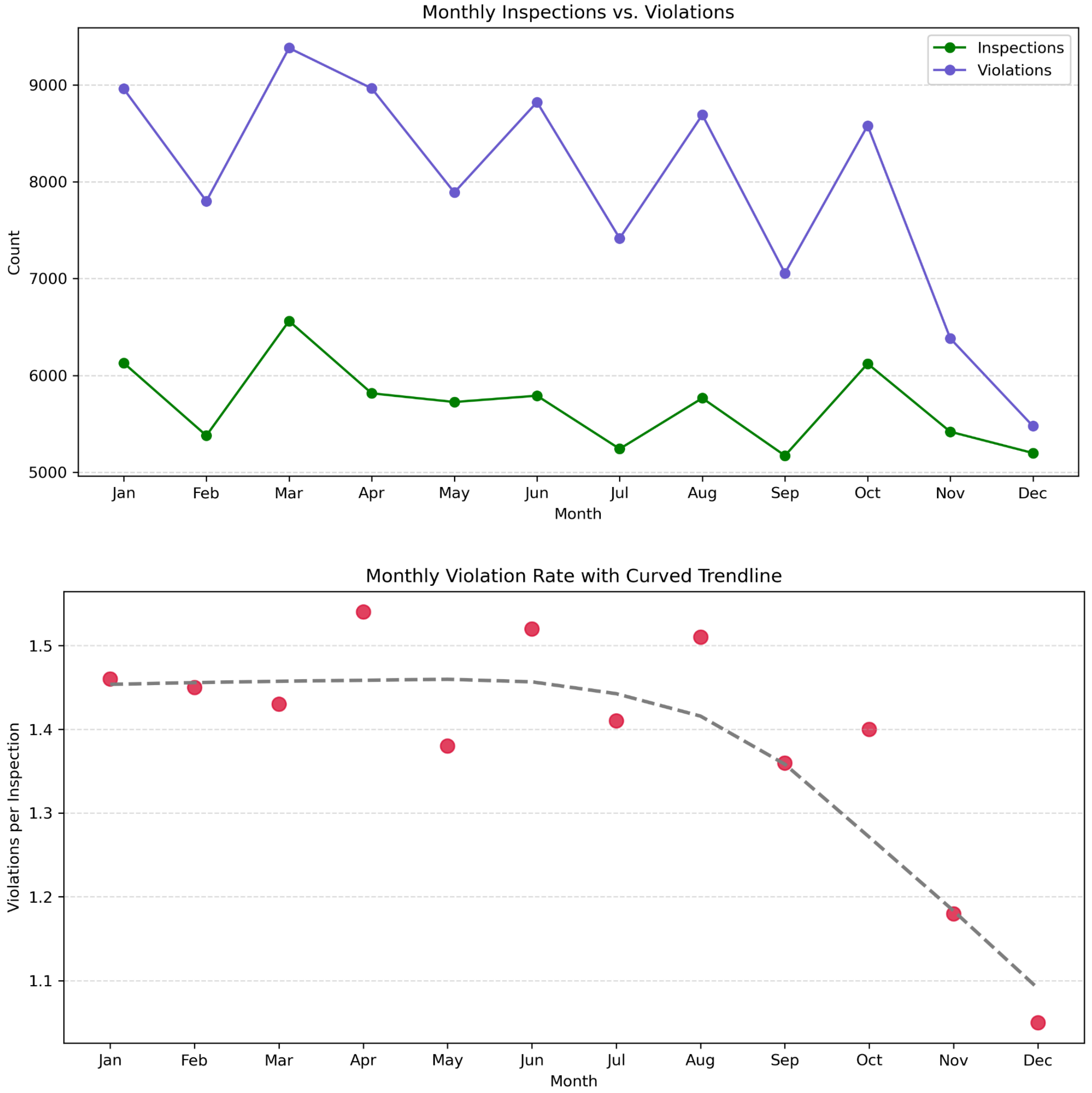


Image 2: This chart highlights the top violators based on the total and average number of violations per inspection. It visually communicates which businesses repeatedly fail inspections and emphasizes the gap between enforcement expectations and outcomes. It also illustrates which of these top violators is still operating with an active license, showing a potential compliance gap.

Line Chart – Monthly Inspections and Violation Trends with Trendline

Shows monthly inspections and violations, supporting trend analysis and bottleneck identification.


Monthly Inspections vs. Violations

Monthly Violation Rate with Curved Trendline

Image 3 & 4: These charts presents trends in monthly inspections and violations over time. This helps identify patterns, seasonality, or gaps in enforcement activities, guiding the better allocation of regulatory resources.

Heatmap – Active Business Concentration by ZIP Code

Guides policy decisions by highlighting geographic areas with the highest business activity.

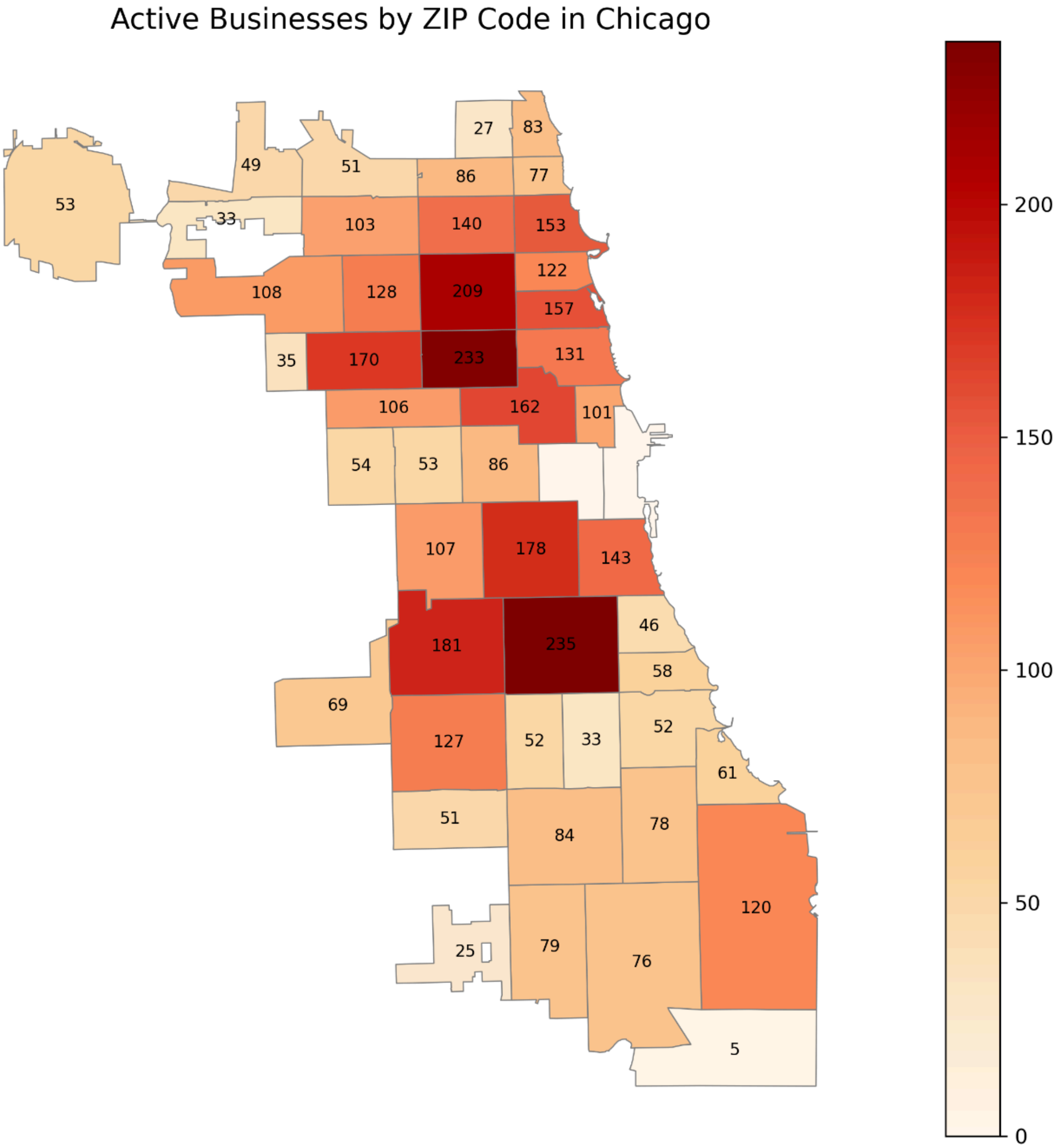


Image 5: The heatmap visualizes the density of licensed businesses across Chicago ZIP codes. It provides spatial awareness for policymakers, helping them focus attention on regions with high business concentration or recurring violations.

Line/Bar Chart – License Expiration, Inspections, and Violations by Month

Offers insight into when regulatory resources should be ramped up for renewals.

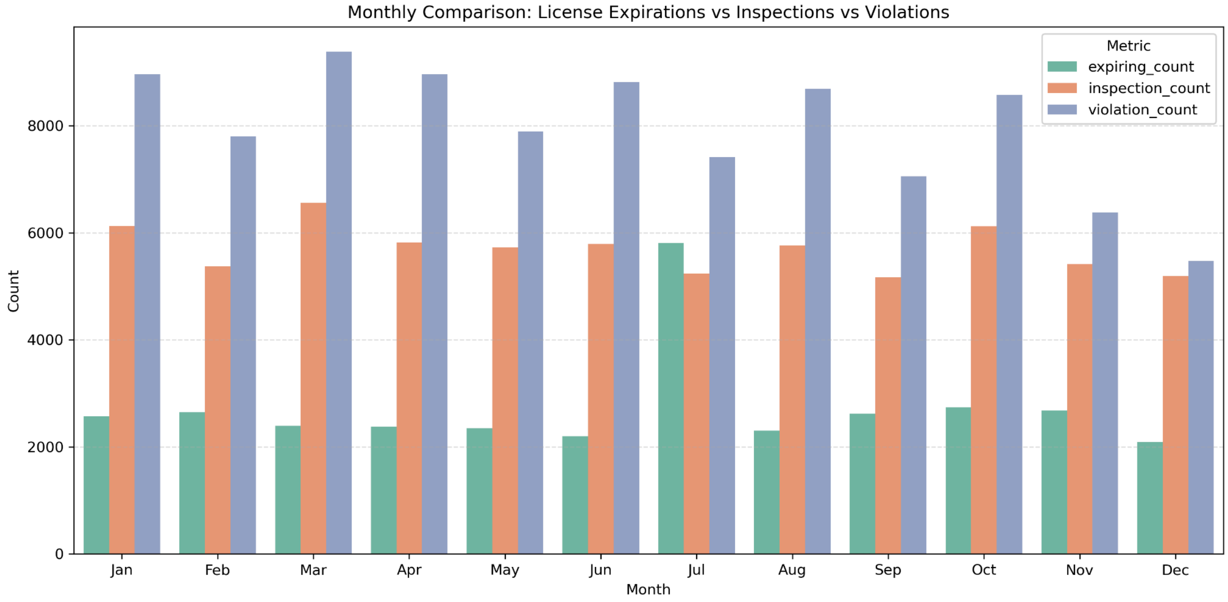
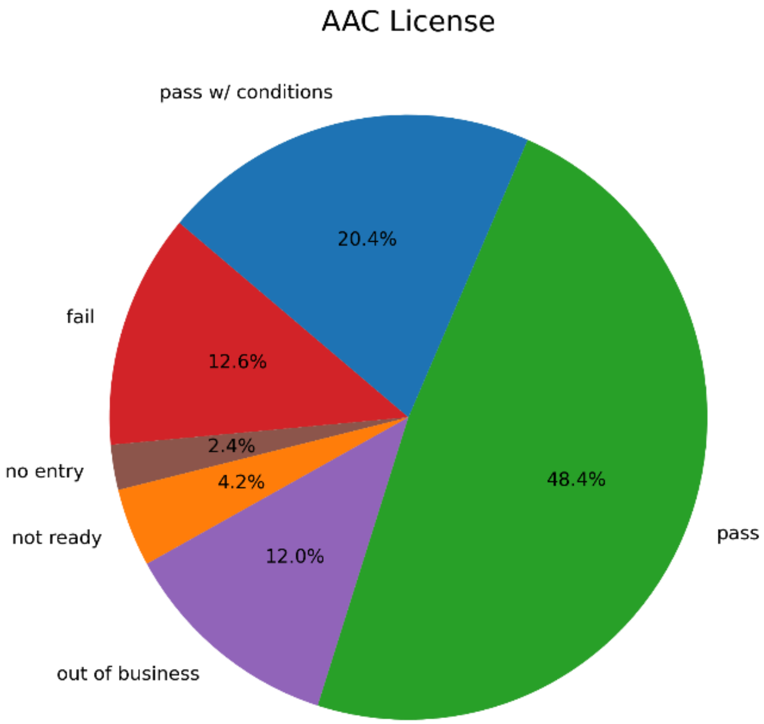
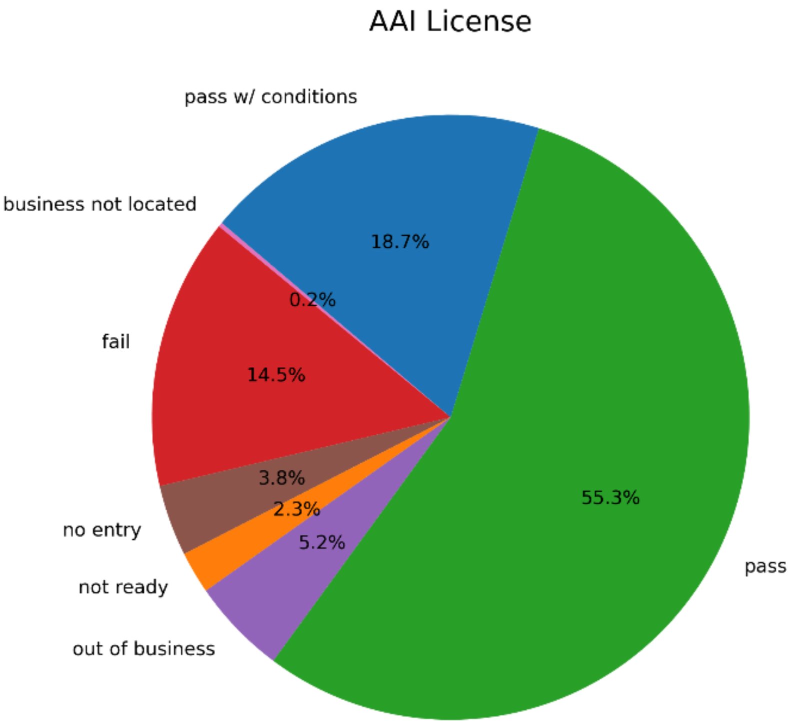


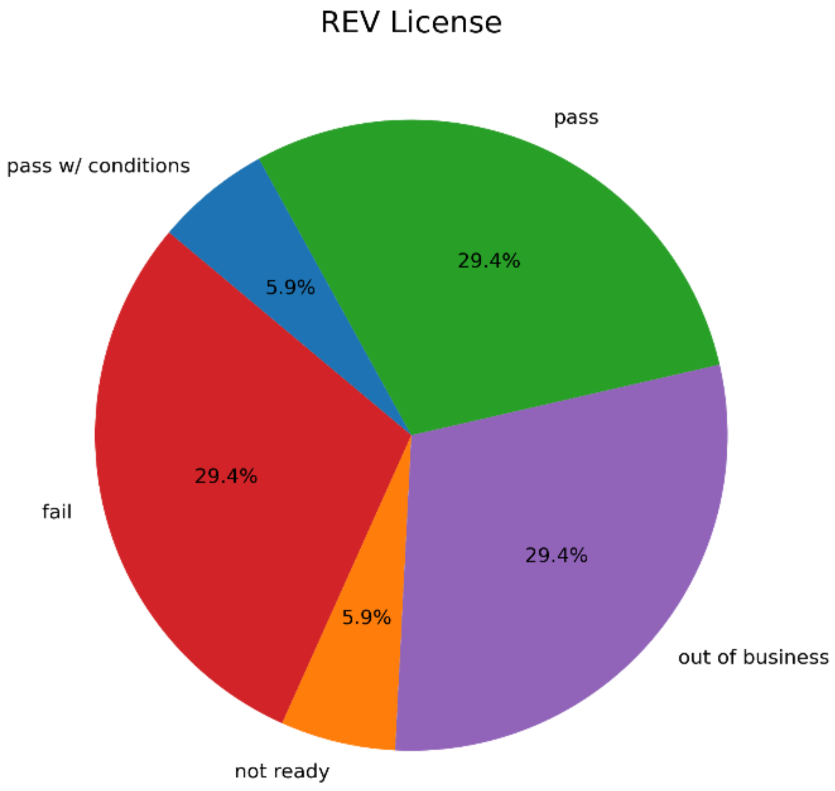
Image 6: This bar chart shows how license expirations are distributed throughout the year. By identifying expiration spikes (e.g., in July), regulatory agencies can better plan renewal outreach and inspection schedules.

Pie Charts – Inspection Results by License Status

Side-by-side comparisons of how license types relate to pass/fail results, showing systemic gaps.







These pie charts above offer a quick comparative view of inspection outcomes (pass/fail) broken down by license status. They are effective for visualizing systemic inconsistencies, such as failed inspections occurring primarily under active licenses. They also illustrate the inconsistencies in record-keeping, with businesses listed as having active licenses (AAI) falling under the category of "out of business".

Recommnedations

Building on the analytical findings, the following recommendations offer actionable steps to enhance regulatory oversight and improve public health and safety. These measures target automation, integration, and accountability, ensuring that critical signals, such as inspection failures, are not missed.

A key innovation proposed in this study is the creation of an "enforcement actions' table that records when and how enforcement measures are taken in response to inspection results. This would serve as a centralized record of city responses to violations, facilitating transparency, auditing, and policy analysis. By capturing which licenses were suspended or flagged after an inspection, the city can close the loop between inspection data and real-world regulatory consequences, improving accountability and responsiveness.

This investigation recommends that the City of Chicago consider the following steps to improve efficiency and transparency:

1. **Automate enforcement triggers**: Flag failed inspections for immediate review of your license status.
2. **Synchronize inspection and licensing systems**: Enable real-time updates across departments.
3. **Develop an enforcement actions dataset**: Capture decisions (warnings, suspensions) along with their corresponding inspection triggers.
4. **Digitize revocation workflows**: Integrate business rules into systems to automate license suspensions.
5. **Standardize key fields across datasets**: Facilitate easier joins and minimize ambiguity.

Proposed Enforcement_Actions Table Schema:

- action_id (PK)
- license_id (FK)
- inspection_id (FK)
- action_type (e.g., Warning, Suspension, Revoked)
- action_date
- trigger_reason
- manual_flag
- notes

Lessons Learned

This project provided valuable insights into the challenges of integrating regulatory datasets and revealed systemic issues in enforcement processes. It reinforced the importance of data normalization, the role of visualizations in uncovering patterns, and the need for feedback loops between inspections and license actions.

- Inspection and licensing systems are only loosely coupled; proper integration requires system-level changes.
- Data normalization and unique identifiers (UIDs) are essential for maintaining database integrity.
- SQL alone is not enough; combining data engineering and visualization enables deeper insight.
- Python visualizations are effective alternatives to Power BI for dynamic storytelling.
- Without automation, critical policy signals (e.g., failed inspections) may go unaddressed.

This project highlights the importance of aligned regulatory systems and shows how data-driven strategies can help cities enhance public health oversight and business compliance. The audit illustrates the value of combining data engineering with policy insights to drive meaningful change in the public sector. System upgrades and automation will be crucial in ensuring safer, compliant business operations in Chicago.