CityUnmasked

Syracuse Crime and Urban Decay Analysis

Arpita Khot | Dhruvin Barot | Shreya Kulkarni

1. Problem Statement, Objective, and Intent

The Problem

Syracuse, New York is experiencing two compounding urban challenges at once. Crime is concentrated in specific corridors - persistent across years, resistant to conventional responses. At the same time, physical urban decay is spreading: buildings formally condemned as unsafe, registered vacant properties sitting unresolved for years, and code violations accumulating faster than the city can address them.

These two problems have historically been treated as separate policy domains. Housing code enforcement operated independently of public safety. Data systems were separate. Policy conversations rarely intersected. The question this project asks is whether that separation is justified or whether the two problems are geographically entangled in ways that make each harder to solve without addressing the other.

The Objective

This objective has three layers. The first is empirical: is there a measurable, statistically significant overlap between physical decay and crime in Syracuse? The second is directional: does decay precede crime, or does crime precede decay? The third is practical: if the relationship is real, can we classify each neighborhood by the type of problem it has and recommend the right intervention for each?

The Intent and Goal

The intent is not to prove a simple cause-and-effect relationship. Properties become vacant or unfit for many reasons that have nothing to do with crime like landlord economics failing, population loss, structural age, harsh upstate winters, and historical disinvestment. Any honest analysis must acknowledge this and show the places where blight exists without crime, not just where they overlap.

The goal is to produce actionable, differentiated recommendations. Different neighborhoods are experiencing different versions of the same problem. A neighborhood where crime and decay are reinforcing each other needs a simultaneous housing and public safety response. A neighborhood where blight exists without high crime needs investment and ownership reform, not increased policing. The goal is to tell those neighborhoods apart and give the city a data-driven basis for deploying the right solution in the right place.

2. Datasets and How Each Was Analyzed

2.1 Crime Data - crime_clean.csv

The crime dataset contains 25,752 incident records spanning 2023 through 2025. Each record includes the crime type, a severity score from 1 to 5, a quality-of-life flag distinguishing serious crimes from minor incidents, season, time of day, day of week, and geographic coordinates.

How it was analyzed: The dataset was analyzed for temporal patterns - which months and hours see the highest crime counts - and for composition - what share of incidents are serious vs minor. The multi-year span was used rather than a single year to reveal structural patterns that hold across time rather than anomalies from one particular year.

Key findings: Larceny, Simple Assault, and Criminal Mischief are the top crime types. 91.5% of all incidents are classified as serious crimes - this is not a quality-of-life policing problem.

2.2 Unfit Properties - Unfit_Properties.csv

This dataset contains 264 formal city violations for properties deemed unsafe or uninhabitable under New York State's Property Maintenance Code. These are the most severe end of the property condition spectrum - a city inspector has formally determined the building poses a risk. The data spans 2014 through 2025.

How it was analyzed: Violations were grouped by year to identify the growth trend. The status field (Open or Closed) was used to calculate the resolution rate. Zip code aggregation identified geographic concentration.

Key findings: 73% of all violations ever filed remain Open today. Violations grew 33 times from 2014 to 2025, with the sharpest acceleration beginning in 2021. The problem is geographically concentrated in zip codes 13204, 13205, and 13208 - Syracuse's west and south sides.

2.3 Vacant Properties - Vacant_Properties.csv

The vacant property dataset contains 1,651 registered vacant properties, of which 1,615 have usable coordinates. It includes neighborhood labels, zip codes, and a validity field indicating whether each vacancy is still active or has been resolved.

How it was analyzed: Properties were grouped by neighborhood and zip code. The active vs resolved split was calculated from the VPR_valid field. The geographic distribution was compared against the unfit properties and crime datasets to test for overlap.

Key findings: 88% of registered vacancies are still active - the highest unresolved rate across all four datasets. Brighton (244 properties) and Northside (177) lead by neighborhood. The same three zip codes that dominate unfit properties (13204, 13205, 13208) also dominate vacant properties - the first evidence of geographic co-occurrence before any cross-dataset analysis.

2.4 Code Violations - code_violations.csv

The code violations dataset contains 140,726 total records. Not all are analytically useful for this project. Administrative violations - registration failures, permit paperwork, business certifications - were excluded because they do not indicate physical decay. After filtering to physical decay violations only, 92,790 records remained.

Filtering logic: Only complaint types indicating physical property conditions were kept: Property Maintenance (Interior and Exterior), Vacant House, Overgrowth, Trash and Debris, Fire Safety, and Vacant Lot. Business certifications, rental registry failures, and permit violations were excluded.

Tiering logic: Retained violations were assigned a severity tier based on keywords in the violation text field.

Tier	Description
Tier 1 — Structural / Critical	Foundation failures, collapse risk, unfit for habitation - 10,334 violations
Tier 2 — Systems Failure	Plumbing, electrical, heating, life-safety systems - 34,427 violations
Tier 3 — Environmental Neglect	Overgrowth, trash, visible abandonment - 48,029 violations

Key findings: 92,790 violations across 10 years (2017–2026) give 108 monthly data points, far more statistical power than any other dataset for time-series analysis. The same three zip codes (13205, 13204, 13208) dominate this dataset too, confirming that four independent datasets point at the same geographic area.

3. How Each Dataset Is Visualized in the Dashboard

Tab 1 - Crime Analysis

This tab presents the crime dataset in four charts.

Top Crime Types (horizontal bar): Shows the 8 most frequent crime types ranked by total count. A horizontal bar was chosen because the crime type labels are long and readable horizontally. The ranking immediately communicates that Larceny, Simple Assault, and Criminal Mischief dominate — a mix of property and violent crime.

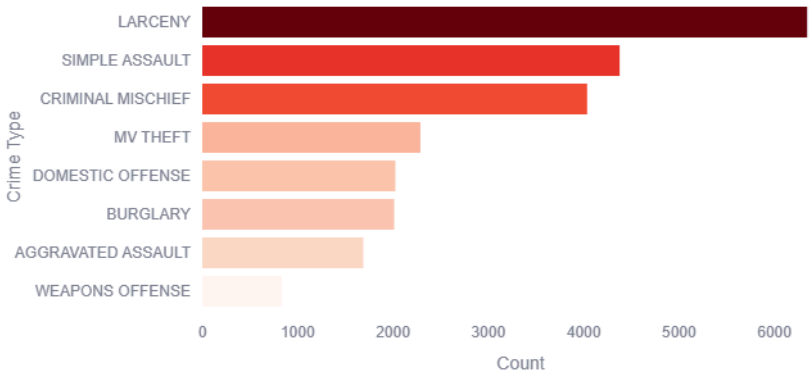

Fig 1: Top Crime Types

Serious vs Quality-of-Life (donut): Splits all incidents into serious crimes and minor quality-of-life incidents. The donut format makes the 91.5% serious share immediately visible without reading numbers. This chart establishes that the city is dealing with a serious crime problem, not a nuisance issue.

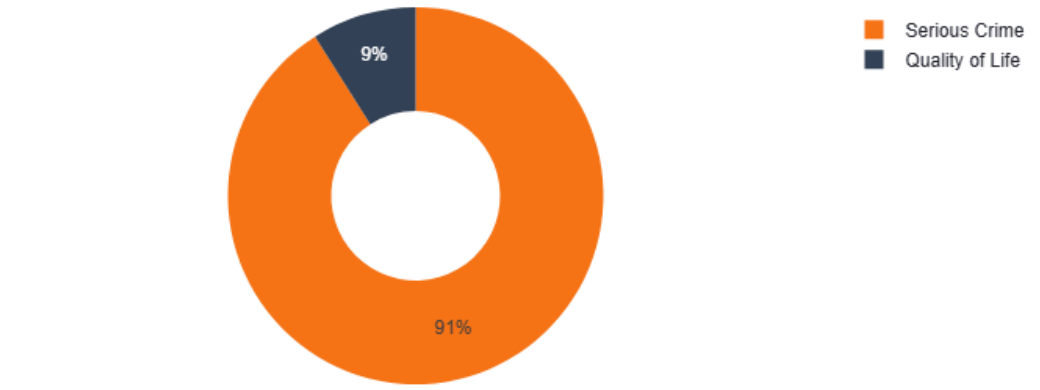

Fig 2: Serious vs. Quality-of-Life

Crime by Month (line): Aggregates incident counts by month across all years. A line chart was chosen to show the continuous seasonal rhythm - the summer peak is visually clear. The multi-year aggregation smooths out single-year anomalies and reveals the structural pattern.

Fig 3: Crime by Month

Crime by Hour of Day (bar): Shows incident counts for each hour. A bar chart was chosen because hours are discrete categories. The evening window (6pm to midnight) is immediately visible as the highest bars, which directly informs patrol deployment timing.

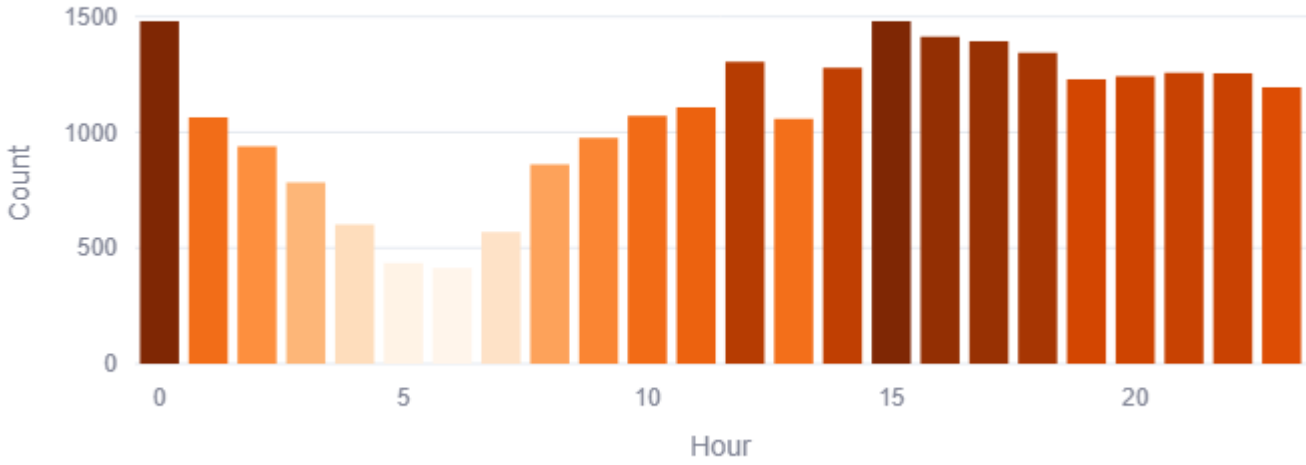

Fig 4: Crime by Hour of Day

Tab 2 - Unfit Properties

This tab presents the unfit properties dataset in four charts.

Violations Filed Per Year (bar with average line): Shows annual violation counts since 2014 with a dashed average line for reference. The growing bars crossing the average and continuing upward make the 33x acceleration visually undeniable. The post-2021 steepening is immediately apparent.

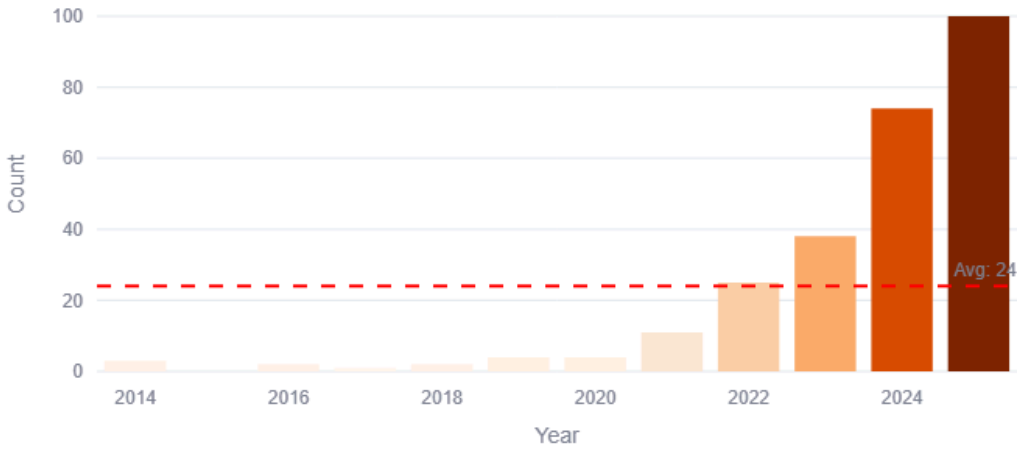

Fig 5: Violations Filed Per Year

Open vs Closed Violations (donut): The dominant red segment for Open (73%) communicates the resolution crisis in a single glance. The city is losing ground - this chart makes that concrete without requiring any numbers to be read.

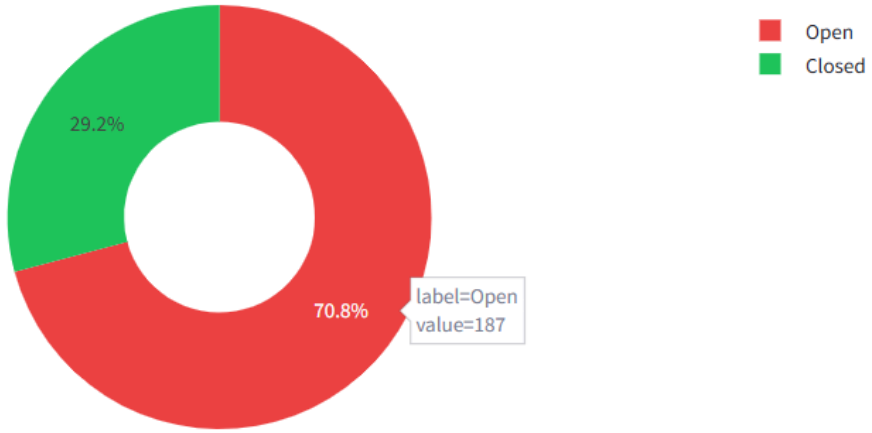

Fig 6: Open vs Closed Violations

Total Violations by Zip Code (bars): Two side-by-side bar charts show which zip codes have the most total violations and the most currently open violations. The near-identical rankings confirm that the unresolved violation problem is geographically concentrated in the same areas as the total violation problem.

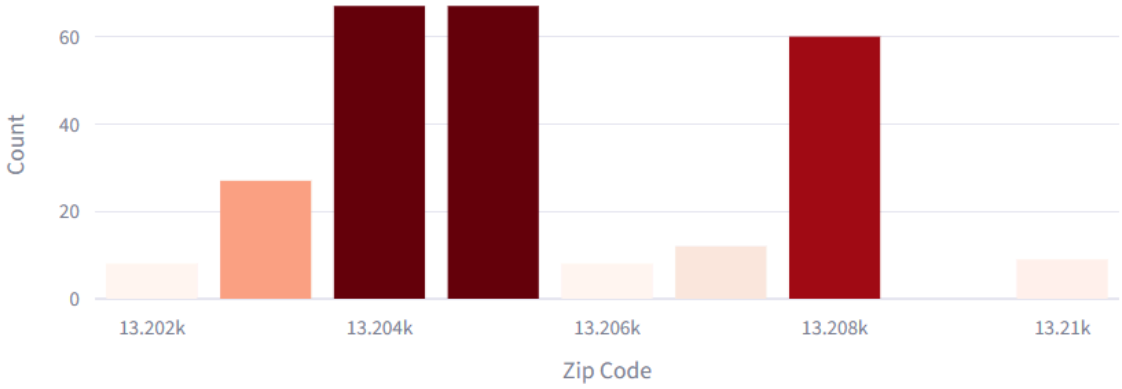

Fig 7: Total Violations by Zip Code

Tab 3 - Vacant Properties

This tab presents the vacant properties dataset in four charts using the same structure as the unfit properties tab for easy visual comparison.

Vacancies by Neighborhood (horizontal bar): Horizontal orientation used because neighborhood names are long. Brighton and Northside leading is immediately visible. The chart introduces the neighborhood-level granularity that zip codes alone cannot provide.

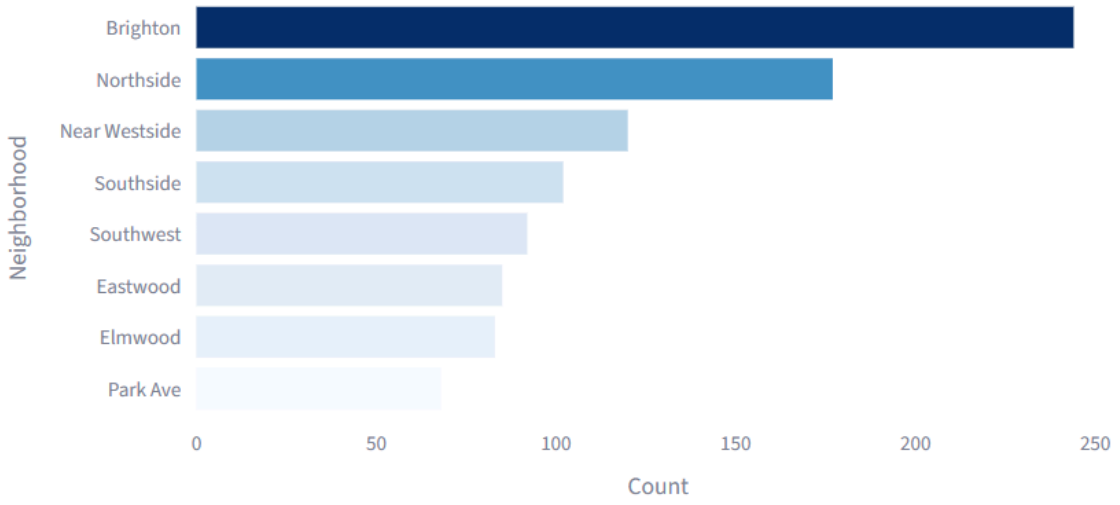

Fig 8: Vacancies by Neighborhood

Active vs Resolved (donut): The 88% active segment is larger than the comparable donut in the unfit properties tab, making the escalating unresolved rate across datasets visually comparable when both tabs are viewed in sequence.

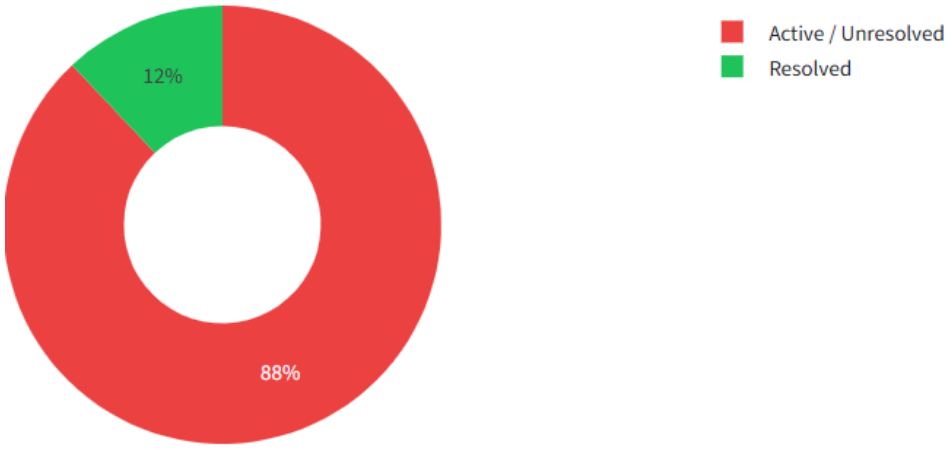

Fig 9: Active vs Resolved

Total Vacancies by Zip Code (bars): 13205's dominance (521 vacancies) is immediately visible in the leftmost bar towering over the others. The near-identical ranking between total and active confirms that these properties are accumulating, not cycling through resolution.

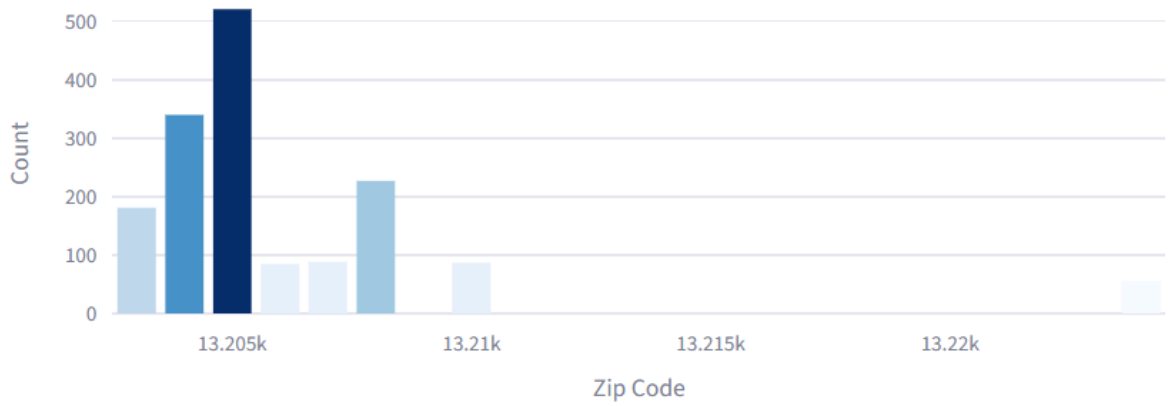

Fig 10: Total Vacancies by Zip code

Tab 5 - Code Violations

This tab presents the code violations dataset with a focus on physical severity and the causal relationship with crime.

Top Zip Codes by Violations (bar): Confirms the same geographic hotspot across a third independent dataset. 13205, 13204, 13208 dominating four separate datasets is the core geographic evidence of the project.

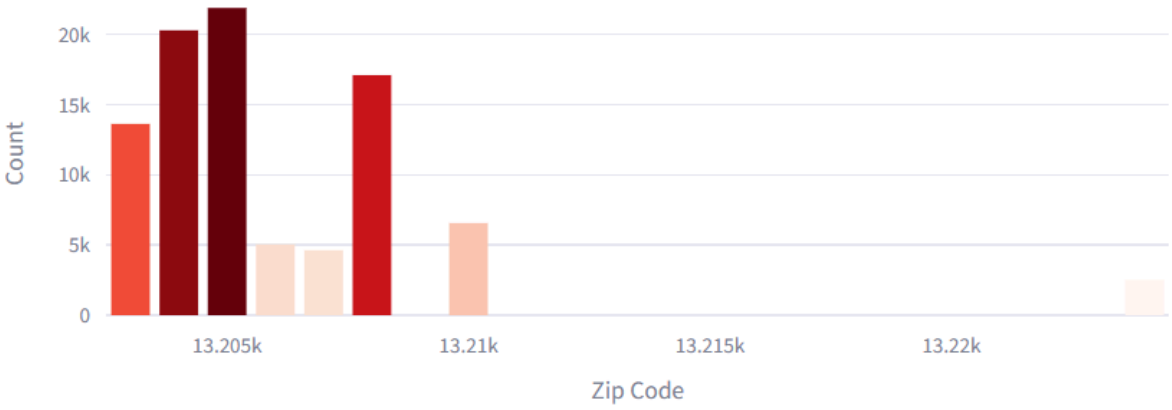

Fig 11: Top Zip Codes by Violations

Violation Mix by Tier (donut): Shows the proportion of violations at each severity tier. Environmental neglect dominates by volume, but the structural tier (Tier 1) is the most dangerous fraction. The chart communicates both the scale of visible disorder and the severity of the underlying structural problem.

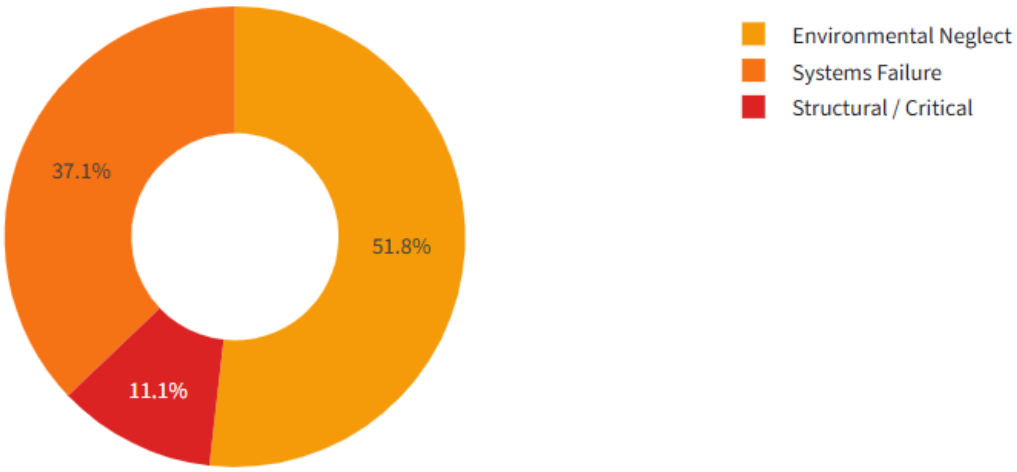

Fig 12: Violation Mix by Tier

Granger Causality P-Value Chart (grouped bar): This is the causality test visualization — explained in detail in Part 4 of this report. The chart shows p-values for each lag in both directions (violations to crime, and crime to violations). Bars below the 0.05 threshold line are statistically significant. This is the most analytically rigorous chart in the project.

Fig 13: Granger Casuality P-Value Chart

Tab 4 - Models and Algorithms Used

4.1 BallTree Haversine Spatial Join

For every crime incident, this algorithm checks whether any decay point falls within 100 meters by building a tree structure that partitions space into nested regions, allowing the search to skip large portions of the dataset rather than computing every crime-to-decay distance individually. Haversine distance is used instead of standard Euclidean distance because latitude/longitude coordinates require accounting for the earth's curvature to be accurate at street level. The output is a proximity flag per crime (near_unfit, near_vacant, near_decay) and a zone label (Near Unfit Only, Near Vacant Only, Near Both, Neither), which feed directly into the Random Forest and the Urban Decay Index.

4.2 Granger Causality Test

This statistical test asks whether past values of one time series help predict future values of another, in this case, whether monthly violation counts predict future monthly crime counts. It is run in both directions (violations to crime, and crime to violations) to avoid confirmation bias and to test whether a feedback loop exists. The code violations dataset is used rather than unfit properties because it provides 108 monthly data points (2017–2026), which gives the test genuine statistical power. Before running, each series is checked for stationarity using the Augmented Dickey-Fuller test and differenced if needed, to ensure the results reflect a genuine predictive relationship rather than a shared trend.

4.3 Weighted Composite Urban Decay Index

Each zip code is scored from 0 to 100 by normalizing three signals - crime count (weighted 40%), combined decay score of unfit plus vacant properties (35%), and percentage of unresolved violations (25%) - and summing them. Zip codes are then classified into four types using median splits: Type A (above median on both crime and decay) is the crime-blight feedback zone, Type B (high decay, low crime) is the economic abandonment zone, Type C (unfit-dominant) is the infrastructure decay zone, and Low Risk sits below median on both axes.

4.4 Logistic Regression (Crime Hotspot Prediction) The city is divided into roughly 400–500 meter grid cells, and for each year (2023, 2024, 2025) a logistic regression is trained to predict whether a grid cell will become a Q4 crime cluster based on its January–September total crime count and share of serious crimes. Predicted risk scores are averaged across all three years to identify chronic hotspots, blocks that are consistently high-risk, not just anomalous in a single year. Logistic regression was chosen over Random Forest here because the goal is a clean probability score per grid cell that can be mapped and ranked spatially, not a feature importance analysis.

Tab 5 - Urban Decay Index - How the Datasets Connect to Crime

The Urban Decay Index tab is where the project's four datasets are brought together into a single analytical framework. Each individual tab shows one dataset in isolation. This tab shows what happens when they are combined.

The Scatter Chart - Crime vs Decay by Zip Code

Each zip code is plotted as a dot, with the Urban Decay Score (unfit + vacant properties) on the horizontal axis and crime count on the vertical axis. Median lines on both axes split the chart into four quadrants. Zip codes in the top-right quadrant have both high crime and high decay these are the Type A crime-blight feedback zones, where housing conditions and public safety problems reinforce each other. Zip codes in the bottom-right quadrant have high decay but low crime the Type B economic abandonment zones, where blight exists for non-criminal reasons.

The presence of dots in this bottom-right quadrant is crucial: if blight always caused crime, it would be empty. Instead, it shows that decay and crime are distinct problems; the places where they do co-occur (top-right) are genuinely meaningful hotspots, not just artifacts of both being correlated with a third factor like poverty.

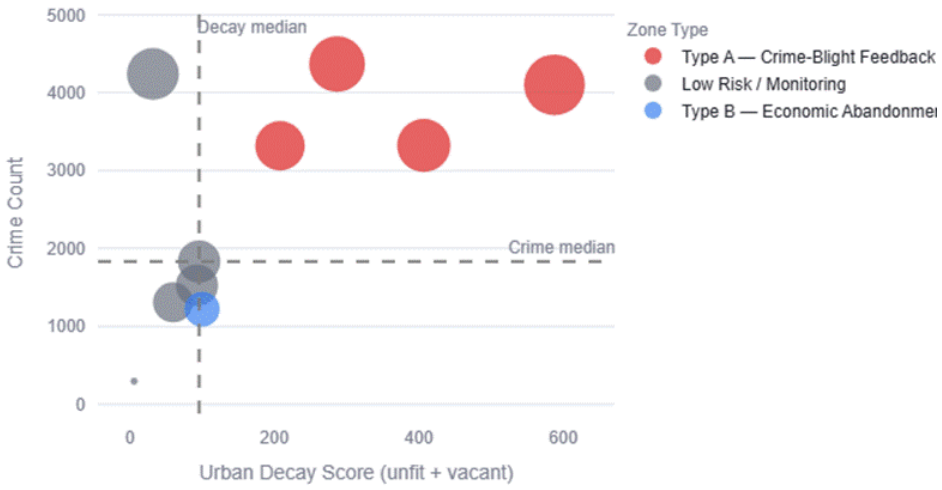

Fig 14: Crime vs Day

The Risk Score Ranking Chart

The top 10 zip codes are ranked by their composite risk score (crime 40%, decay 35%, unresolved violations 25%), with each bar color-coded by zone type. This chart is the policy delivery output. It tells the city which zip codes need attention first, and the color tells the city what kind of attention each one needs. A red bar means Type A: simultaneous housing and public safety intervention. A blue bar means Type B: investment and ownership reform, not enforcement.

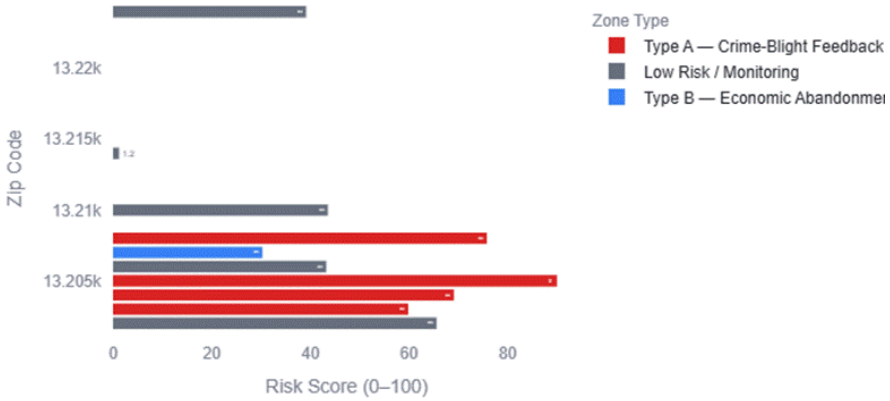

Fig 15: Zip Codes by Risk Score

Crime Types by Decay Zone

For the five most common crime types, incidents are split by which decay zone they occurred in Near Both, Near Unfit Only, Near Vacant Only, or Neither. The purpose of this chart is to test whether the decay-crime relationship is specific to serious violent crimes or applies equally across all types. If assault and robbery are disproportionately concentrated in the Near Both zone relative to their overall frequency, it means the most dangerous crimes are specifically associated with the highest-intensity decay environments. This directly informs severity-based patrol deployment decisions.

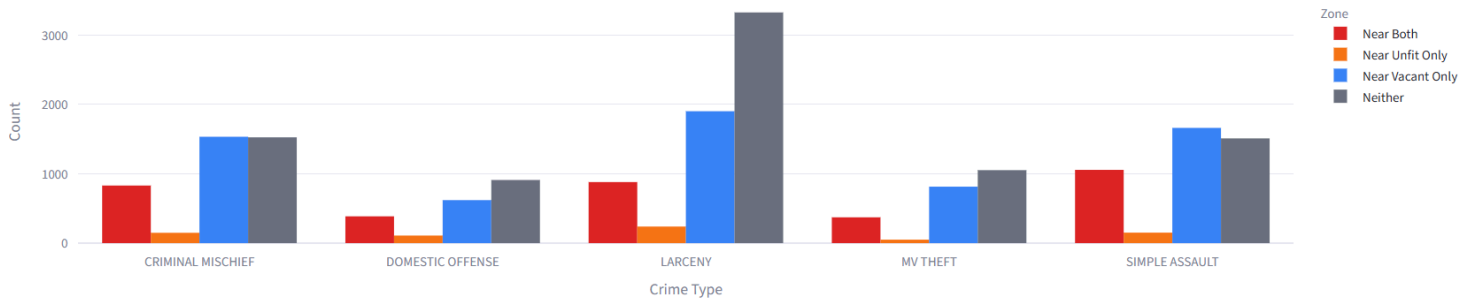

Fig 16: Crime Types by Decay Zone

The Key Finding

Zip codes 13204, 13205, and 13208 rank at the top of the risk score chart and are classified as Type A zones. These are the same three zip codes that appeared at the top of every individual dataset unfit properties, vacant properties, and code violations. Four independent datasets, each loaded and analyzed separately, all pointing at the same three zip codes is not coincidence. It is the project's strongest evidence that the problem in these areas is structural and compounding, not random. The analysis also identifies zip codes with significant vacancy and low crime economic abandonment zones where the right response is investment, not policing. This differentiation is what makes the project's recommendations defensible rather than generic.

Tab 6 - The Map Tab

Layers and Design

The Map tab places all three property datasets on a single interactive Folium map of Syracuse, with three layers that can be toggled independently. The crime layer appears as a heatmap that shows the density of incidents across all years, with brighter areas indicating more crime and providing the baseline spatial pattern. Unfit properties are shown as individual circle markers, with red markers for open, unresolved violations and gray for closed cases; each marker includes a tooltip with the address, status, and year, which is feasible and informative given there are only 264 such properties. Vacant properties are visualized as a blue–cyan heatmap rather than individual points, since 1,615 locations would clutter the map; this layer reveals vacancy density without obscuring the underlying streets and neighborhoods.

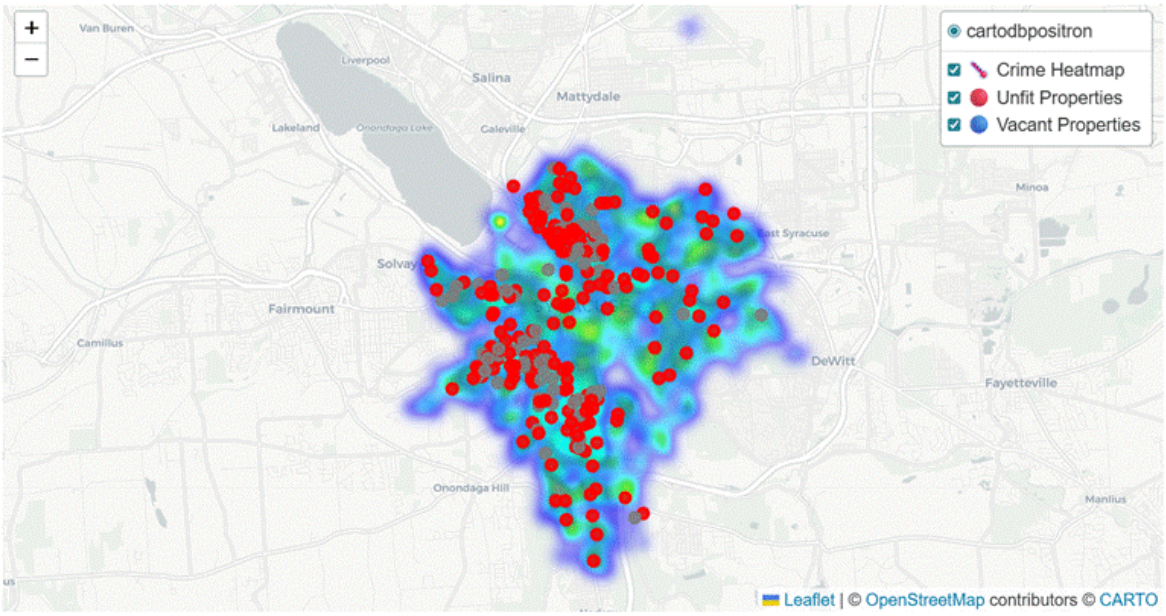

Fig 17: The Heat Map

How to Read the Map

The most important areas to interpret are the places where all three layers overlap: where the orange crime heatmap is bright, red unfit markers are present, and the blue vacant heatmap is dense on the same blocks. These intersections correspond to the Type A zones identified analytically by the Urban Decay Index, highlighting the blocks where crime and physical decay reinforce each other and where joint housing and public safety intervention is most urgent. In practice, the northwest and southwest corridors of Syracuse show the strongest three-way overlap, aligning with ZIP codes 13204, 13205, and 13208, which consistently rank as high-risk across every dataset in the project

Tab 7 - The Prediction Tab

The prediction tab contains the project's forward-looking model: a spatial-temporal hotspot predictor that identifies which city blocks are chronically at risk of becoming crime clusters in Q4 (October through December). The model uses 2023, 2024, and 2025 crime data from crime_clean.csv.

What the Model Predicts

For each approximately 400 to 500 meter grid cell in the city, the model predicts the probability that the cell will become a crime cluster in October through December, based on the crime patterns observed in January through September of the same year. This is run for each year (2023, 2024, 2025) and the predicted risk is averaged across all three years to identify cells that are persistently high-risk. A grid cell is labeled as a cluster if it receives 3 or more crimes in Q4. The threshold of 3 was chosen to distinguish genuinely dangerous blocks from blocks with isolated incidents.

The Risk Heatmap

The heatmap shows the predicted probability that each grid cell becomes a Q4 crime cluster, averaged across 2023 through 2025. Brighter colors indicate higher chronic risk. Four large labeled circles identify the neighborhoods that consistently appear in the highest-risk tier across all three years: Downtown (13202), Southside (13207), Eastside and SU area (13210), and Near Westside (13204). Blue markers identify the 10 individual grid cells with the highest average predicted risk.

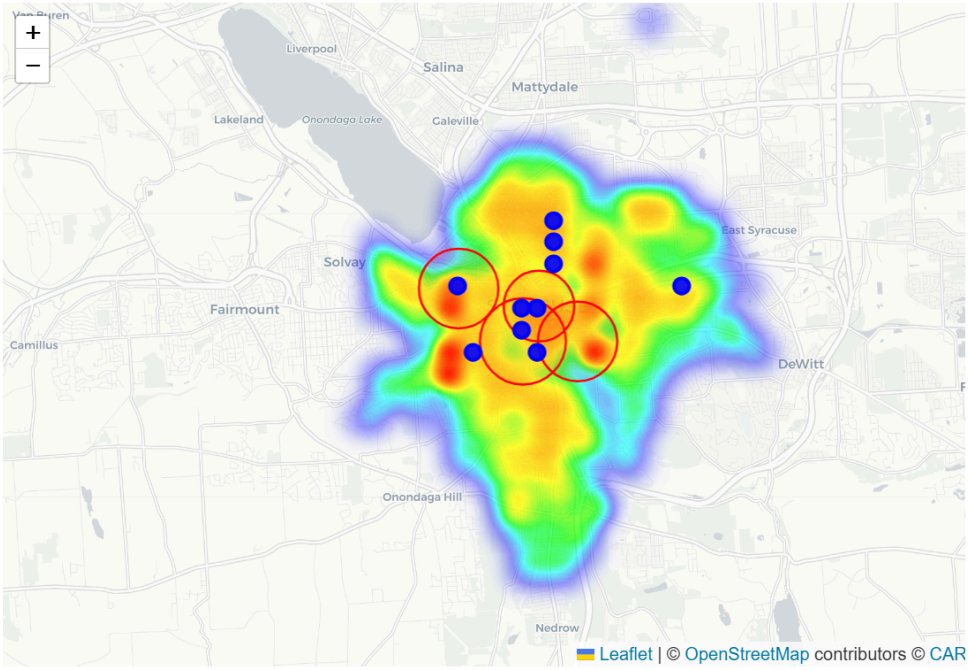

Fig 18: The Risk Map

The Top 10 Table

A table showing the 10 highest-risk grid cells ranked by **average predicted risk score**. For each grid cell, this score is computed by averaging the model's predicted probability of becoming a Q4 hotspot across the three years (2023–2025), based on features observed from January to September. Each row shows the grid cell's center coordinates, this risk score (the model's estimated probability of becoming a Q4 cluster), and the average number of crimes actually observed in Q4 across the three years. This table is the operational output of the model: a city planner or police commander can take these coordinates and identify the specific blocks that need concentrated attention **before** Q4 begins.

	Rank	Lat Center	Lon Center	Risk Score	Avg Future Crimes (Oct–Dec)
0	1	43.0475	-76.1525	1	63
1	2	43.0425	-76.1525	1	54.5
2	3	43.0625	-76.1425	1	56
3	4	43.0475	-76.1475	1	35.5
4	5	43.0525	-76.1725	1	45
5	6	43.0375	-76.1675	1	31
6	7	43.0575	-76.1425	1	29.5
7	8	43.0375	-76.1475	1	30
8	9	43.0525	-76.1025	1	38.5
9	10	43.0675	-76.1425	1	22

Fig 19: Table Predicting risk score

What the Model Suggests

The model consistently identifies four chronic hotspot areas. Downtown (13202) and Eastside (13210) are institutional and commercial centers where late-night activity, transit, and density create persistent risk, while Southside (13207) and Near Westside (13204) are residential areas where the crime - blight feedback loop from the Urban Decay Index is active. This aligns with the A/B/C zone classification: Southside and Near Westside blocks that appear in the Top 10 are Type A zones, needing simultaneous property intervention and targeted public safety, whereas Downtown and Eastside require a different strategy focused on late-night safety infrastructure, code enforcement on problem parcels, and coordination with anchor institutions.

The practical value of the model is that it shifts the city's response from reactive to proactive. Instead of deploying resources after Q4 crime clusters emerge, the city can use January - September data to anticipate where those clusters are likely to form and concentrate housing, code, and policing resources **before** the high-risk window begins.

4. Conclusion

CityUnmasked tested whether physical decay and crime in Syracuse are meaningfully connected — across four datasets and five methods, they are. Crime clusters around decay points, the Urban Decay Index and hotspot model both flag **13204, 13205, and 13208** as structurally high-risk, and forecasts show which blocks are likely to become Q4 hotspots before they form. Just as important, Type B economic abandonment zones show that blight is **not always** crime-driven. The project's value is a framework that distinguishes where the city needs joint housing + public safety action, where it needs investment without extra enforcement, and where infrastructure repair is the priority.