

CARTO

Unlocking Smarter Property Risk Assessments with Spatio-Temporal Crime Insights and CARTO

Follow @CARTO on Twitter

GeoPython 2025

Basel Switzerland

February 24-26

<https://2025.geopython.net>



Lucía García-Duarte
lgarciaduarte@cartodb.com

Data Scientist @CARTO

Agenda

Intro to cloud native CARTO

Intro to the Analytics Toolbox and Workflows

Use case: Unlocking Smarter Property Risk Assessments
with Spatio-Temporal Crime Insights and CARTO

Questions and Answers

Before we begin

All the content for this workshop including a transcript, links to all maps and code can be found in:

<https://rb.gy/geesg9>



SCAN ME

If you haven't already...

For this session, you'll need a CARTO account! If you don't have one, you can set up a free 14-day trial at app.carto.com

! There is a maximum of one CARTO account per email address. If you have previously set up a free trial with your email, we recommend using an alternative email address for this session. If you run into any issues setting up an account, please contact support@carto.com.



SCAN ME

<https://rb.gy/geesg9>



The leading platform for
Location Intelligence and
Spatial Data Science



With an end to end platform:



Technology

Managed cloud or
on-premises
platform



Data

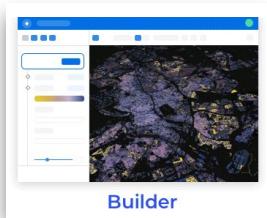
Open and
premium location
data streams



Services

Ongoing enablement
programs and custom
engagements

CARTO brings together cloud connectivity, visualization, spatial analysis and development capabilities in a unified workspace.



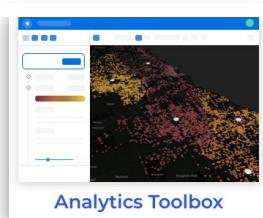
Builder



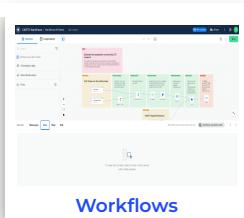
Data Explorer



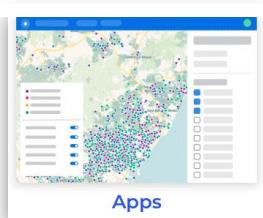
Data Observatory



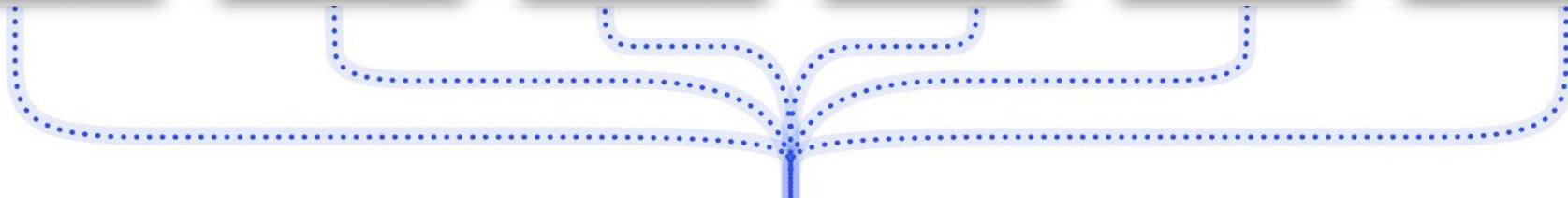
Analytics Toolbox



Workflows



Apps



C A R T O



BigQuery



snowflake*



amazon
REDSHIFT

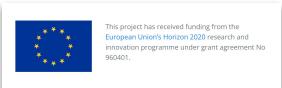


databricks

Analytics Toolbox

Overview

- Set of UDFs and Stored Procedures that unlock advanced Spatial Analytics **natively** within the data warehouses.
- Executed directly from your client, using **simple SQL** commands.
- Separated in **different levels** of abstraction, with core, advanced and domain specific functions (e.g. telco).



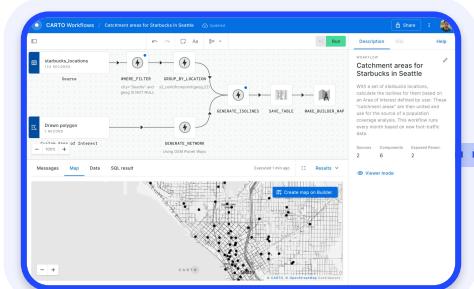
The screenshot displays the "Analytics Toolbox" interface, organized into three main horizontal sections:

- Domain-specific:** Contains a "Retail" icon and a "Coming soon" placeholder for "Logistics".
- Advanced:** Contains icons for "Tiler", "Data", "Clustering", "Random", "Routing", "Geocoding", and "Statistics".
- Core:** Contains icons for "Transformations", "Processing", "Measurements", "Placekey", "Constructors", "Accessors", "H3", "Quadkey", "S2", and "Geohash".

At the bottom of the interface, there is a footer bar featuring the BigQuery GIS logo.

Workflows

A new UI to design Spatial SQL workflows

A screenshot of the generated SQL code. The code is a complex multi-line query using PostgreSQL syntax. It includes several SELECT statements, a CREATE OR REPLACE TABLE statement, and various spatial functions like ST_GEOPOINT and ST_BUFFER. The code is divided into sections by red boxes, likely corresponding to the workflow steps shown in the first image.

User designs workflows in CARTO

Workflows are compiled into SQL

Workflows are executed on your data warehouse



CARTO — Unlock the power of spatial analysis

CARTO Platform

A Guided Tour

<https://app.carto.com/>

Workspace

- [Home](#)
- [Maps](#)
- [Workflows](#)
- [Data Explorer](#)
- [Data Observatory](#)
- [Connections](#)
- [Applications](#)
- [Developers](#)
- [Settings](#)

Welcome to CARTO

Get started by creating stunning maps and bringing your spatial data analysis with our Data Observatory, build powerful apps, and more — your location data into powerful insights.

Getting started
Bring your data and create your first map

Start with your spatial analysis

Demo

Pinpoint new store locations closest to your customers

3 weeks ago

Demo

Monitor retail store performance

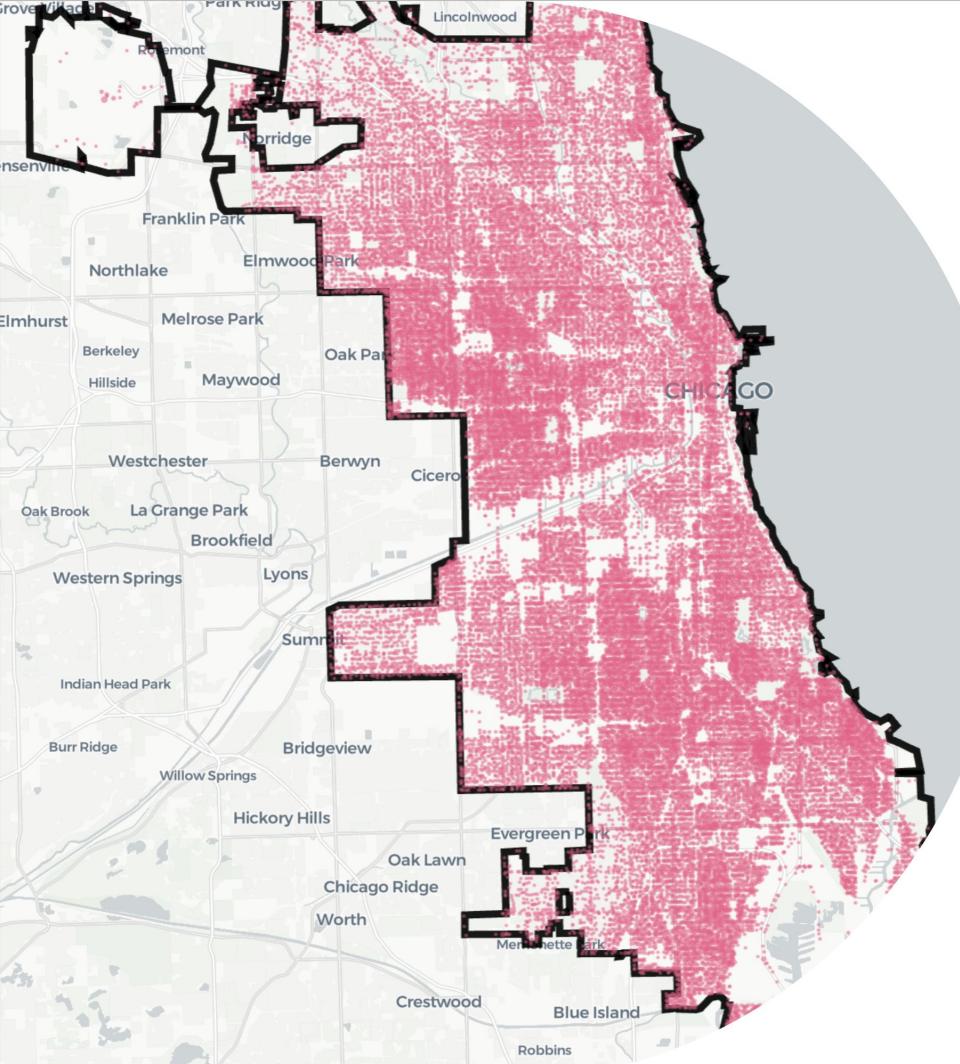
8 mth. ago

Start creating workflows

It's time for a real world example!

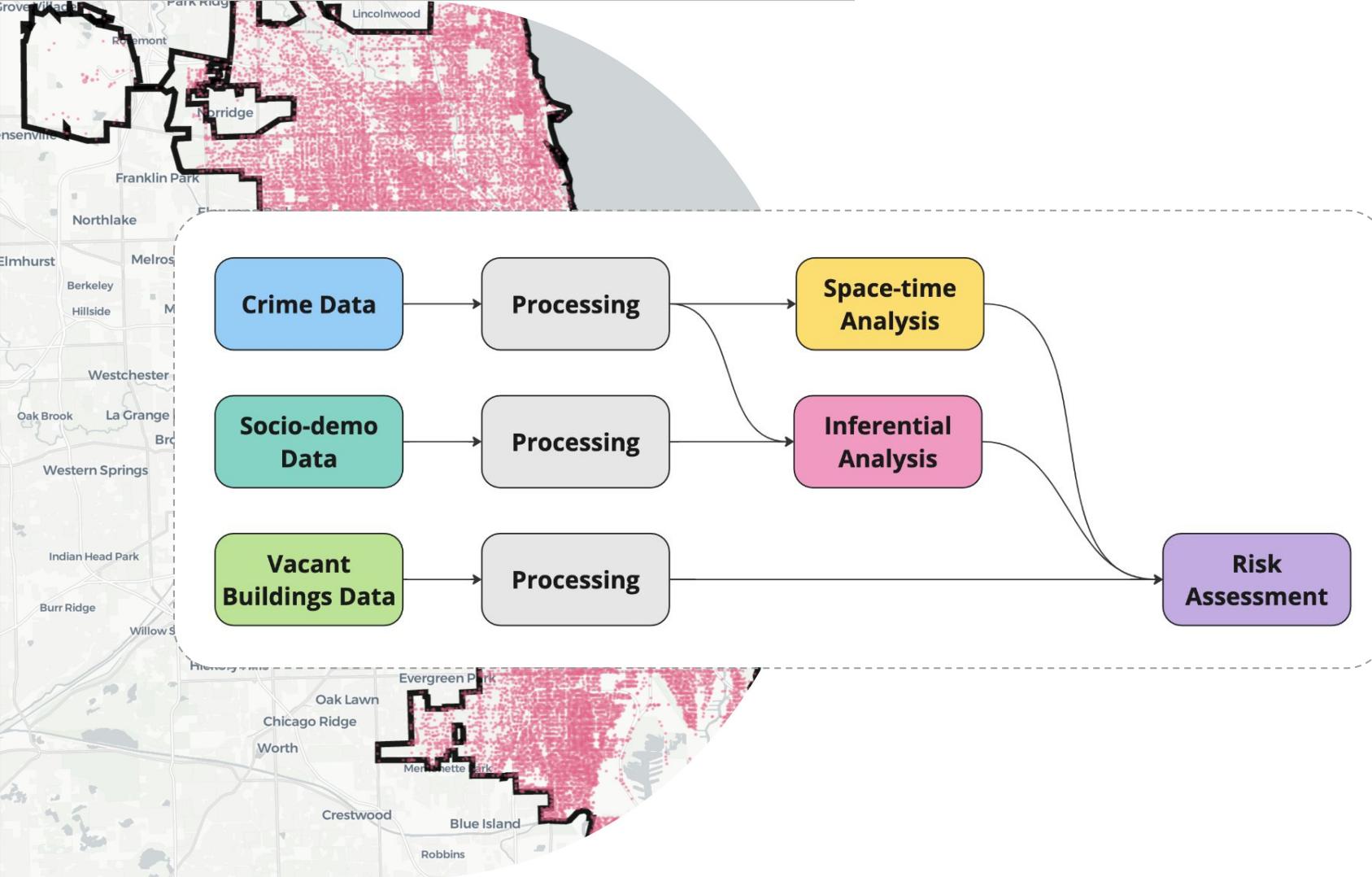
- How to take advantage of cloud native data warehouses to process, visualize, and analyze spatio-temporal big data
- How to use CARTO to analyse and model the spatial distribution and temporal evolution of real-world datasets
- How to use CARTO to create workflows and visualizations that provide decision-makers with easy access to critical insights

Unlocking Smarter Property Risk Assessments with Spatio-Temporal Crime Insights



We'll get hands-on with CARTO, showing you how to **analyze** the **patterns and changes** in violent crimes **over time and space**.

We'll learn how insurers can use the results from this analysis to **improve portfolio management** by better assessing non-financial, location-based risks related to not just crime, but also to other perils, like floods, hurricanes, hail.



Presenting the data

Violent Crime Data

- **Crime incidents** in Chicago from 2001-present
- It features **one row per report**
- The geo support is **the Census block's lon/lat coordinates**

Can be found in:

[Google BigQuery public marketplace](#)
`@bigquery-public-data.chicago_crime.crime`

◀ Product details



Chicago Crime Data

[City of Chicago](#)

Chicago Police Department crime data from 2001 to present

[VIEW DATASET](#) ↗

OVERVIEW

SAMPLES

RELATED PRODUCTS

Overview

This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. In order to protect the privacy of crime victims, addresses are shown at the block level only and specific locations are not identified.

US Census

- **American Community Survey** 5-years and 1-year estimates
- It features:
 - The **geometry** of each Census block group
 - Population, median rent, etc.

Can be found in:

[Subscribing to the ACS data in CARTO spatial catalog](#)

The screenshot shows the CARTO Spatial Data Catalog interface. At the top, there's a navigation bar with the CARTO logo and links for Platform, Solutions, Documentation, and Pricing. Below the navigation is a breadcrumb trail with a back arrow and the text "Spatial Data Catalog". The main content area has a blue header with the title "Demographics / American Community Survey" and "Sociodemographics, 2018, 5yrs - United States America (Census Block Group)". Below the header are three tabs: "Summary", "Data" (which is selected), and "Map". The "Map" tab displays a choropleth map of census block groups in a geographic area, with colors representing different demographic data. A small "Explore" button is visible in the bottom right corner of the map area. At the very bottom, a footer note states: "The American Community Survey (ACS) is an ongoing survey that pr...".

CARTO Spatial Features

- **Derived** variables including demographics, points of interest, and climatology, with **global** coverage
- It features:
 - The **H3** cell
 - The urbanity level

Can be found in:

[Subscribing to CARTO Spatial Features data in CARTO spatial catalog](#)

C A R T O Platform Solutions Documentation Pricing

← Spatial Data Catalog

Derived / CARTO
Spatial Features - United States of America
Resolution 8)

Summary Data Map

Spatial Features is a dataset curated by CARTO providing access to a

Vacant Buildings

- 311 calls for open and vacant buildings [reported to the City of Chicago](#)
- Address location, vacant or occupied status; if the building appears dangerous or hazardous...

Can be found in:

`cartobq.sdsc24_ny_workshops.CHI_boundary_3
11_vacant_buildings`



I WANT TO ▾

PROGRAMS AND INITIATIVES ▾

GOVERNMENT ▾

ABOUT ▾

Department of Buildings

Enhancing safety and quality of life for
Chicago's residents and visitors

[Buildings Home](#)

[Sign Up for E-mail Alerts](#)

Forms ▾

[Report a Problem Building](#) ▾

[Records and Data](#) ▾

Alerts ▾

[Home](#) / [Departments](#) / [Department of Buildings](#) / Vacant and Abandoned Buildings Service Requests

Vacant and Abandoned Buildings Service Requests

This data set contains 311 calls for open and vacant buildings reported to the City of Chicago between January 1, 2010 and December 2018.

Click **Menu** in the upper right-hand corner of the Data Player below to: view, print, or download this data set or access the data via API. To sort or remove columns, click **More Views**. Click the **Share** button on the left, which is just below the Menu button, to email data or post to social networks. To view the date last updated, click on the **Info** button, which is on the right just below the City seal.

Before downloading or sharing data, please read the [Terms of Use](#). For more information or assistance [Contact Us](#).

311 Service Requests - Vacant and Abandoned Buildings
Reported - Historical



Pre-processing the data

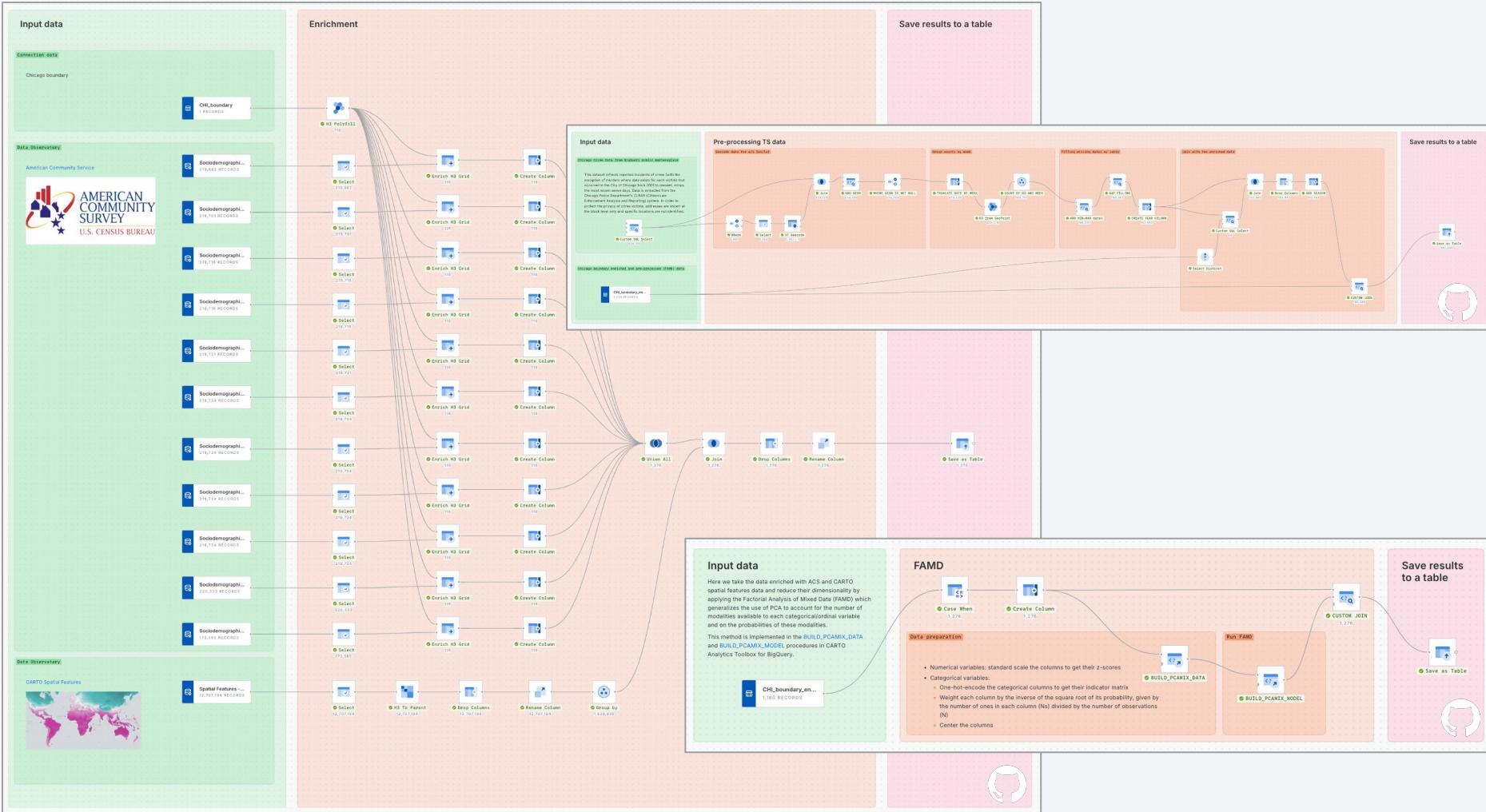
We've already...

1. Transformed all input sources into a regular grid of **Spatial Indexes** through **Enrichments**
2. Reduced the dimensionality of the enriched data to reduce model complexity (we will use this data to model crime counts) using the **Factorial Analysis of Mixed Data** algorithm

Processed data is available through BQ:

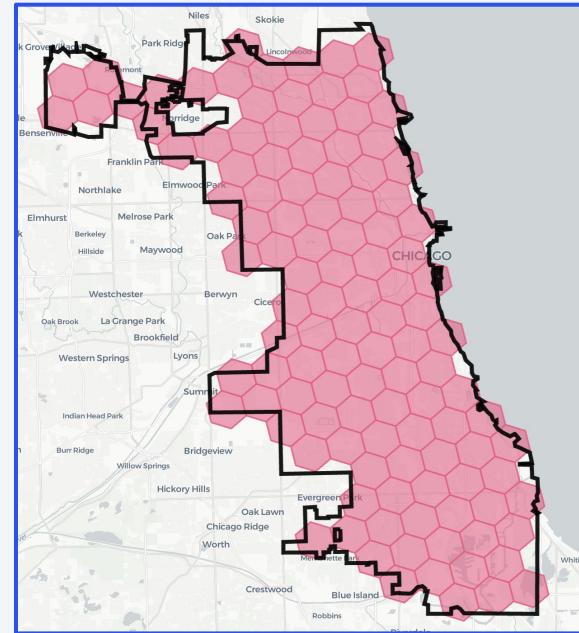
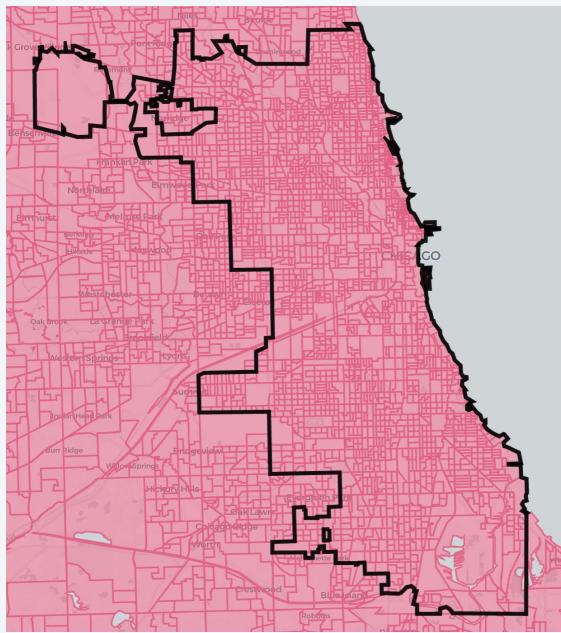
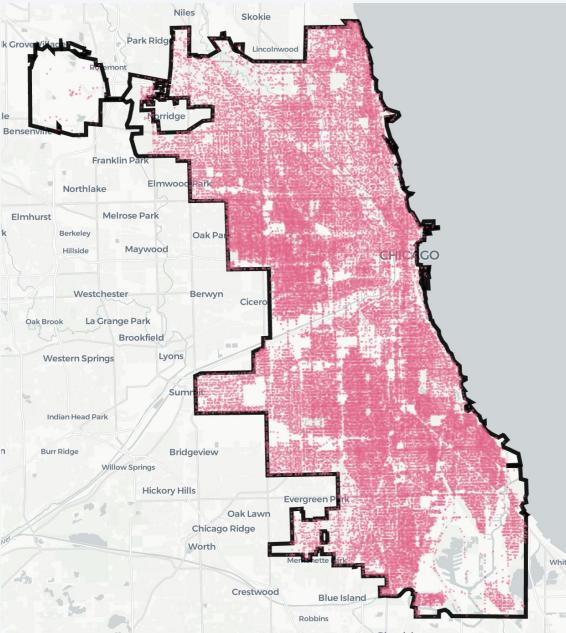
`cartobq.sdsc24_ny_workshops.CHI_boundary_enriched`

SCHEMA	DETAILS	PREVIEW	TABLE EXPLORER	PREVIEW
<input type="checkbox"/> week				DATE
<input type="checkbox"/> h3				STRING
<input type="checkbox"/> counts				INTEGER
<input type="checkbox"/> year				INTEGER
<input type="checkbox"/> month				INTEGER
<input type="checkbox"/> total_pop_sum				FLOAT
<input type="checkbox"/> median_age_avg				FLOAT
<input type="checkbox"/> median_rent_avg				FLOAT
<input type="checkbox"/> black_pop_sum				FLOAT
<input type="checkbox"/> hispanic_pop_sum				FLOAT
<input type="checkbox"/> owner_occupied_housing_units_median_value_sum				FLOAT
<input type="checkbox"/> vacant_housing_units_sum				FLOAT
<input type="checkbox"/> housing_units_sum				FLOAT
<input type="checkbox"/> families_with_young_children_sum				FLOAT
<input type="checkbox"/> urbanity_any				STRING
<input type="checkbox"/> urbanity_any_ordinal				INTEGER
<input type="checkbox"/> principal_component_1				FLOAT
<input type="checkbox"/> principal_component_2				FLOAT



We've already...

1. Transformed all input sources into a regular grid of **Spatial Indexes** through **Enrichments**

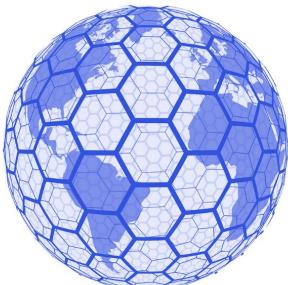


Spatial Indexes

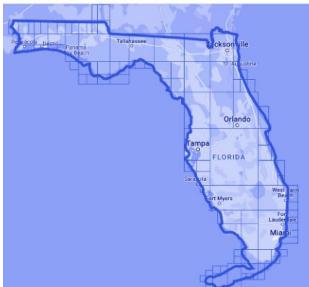
Multi-resolution, hierarchical grids that are “geolocated” by a short reference string, rather than a complex geometry:

(0,0)	(1,0)	(2,0)	(3,0)	(4,0)	(5,0)	(6,0)	(7,0)
(0,1)	(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)	(7,1)
(0,2)	(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)	(7,2)
(0,3)	(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)	(7,3)
(0,4)	(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)	(7,4)
(0,5)	(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)	(7,5)

Quadbin



H3



S2

Spatial Indexes



8a2aa84ec307fff

Geometries



```
POLYGON((-96.196141 41.125515,
-96.195606 41.125514, -96.181864
41.125507, -96.177078 41.125474,
-96.167733 41.125456, -96.160565
41.125456, -96.154682 41.125429,
-96.151094 41.125414, -96.138848
41.125395, -96.138454 41.125394,
-96.138381 41.125394, -96.137158
41.125391, -96.130043 41.125377,
-96.1301...)*
```

Spatial Indexes

Multi-resolution, hierarchical grids that are “geolocated” by a short reference string, rather than a complex geometry:

- **EFFICIENCY** → smaller to store, faster to process and more computationally efficient.
- **FLEXIBILITY** → combining datasets into one layer for ease of comparison and analysis.
- **OBJECTIVITY** → remove bias associated with traditional administrative geographies.
- **IMPACT** → visually more impactful and much easier to understand.

Spatial Indexes



8a2aa84ec307fff

Geometries

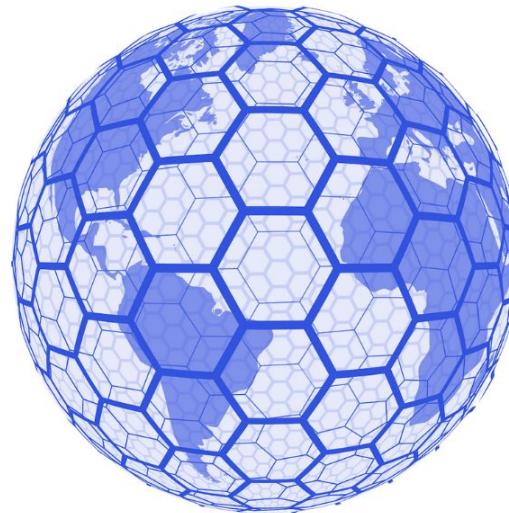


```
POLYGON((-96.196141 41.125515,  
-96.195606 41.125514, -96.181864  
41.125507, -96.177078 41.125474,  
-96.167733 41.125456, -96.160565  
41.125456, -96.154682 41.125429,  
-96.151094 41.125414, -96.138848  
41.125395, -96.138454 41.125394,  
-96.138381 41.125394, -96.137158  
41.125391, -96.130043 41.125377,  
-96.1301...)*
```

Spatial Indexes

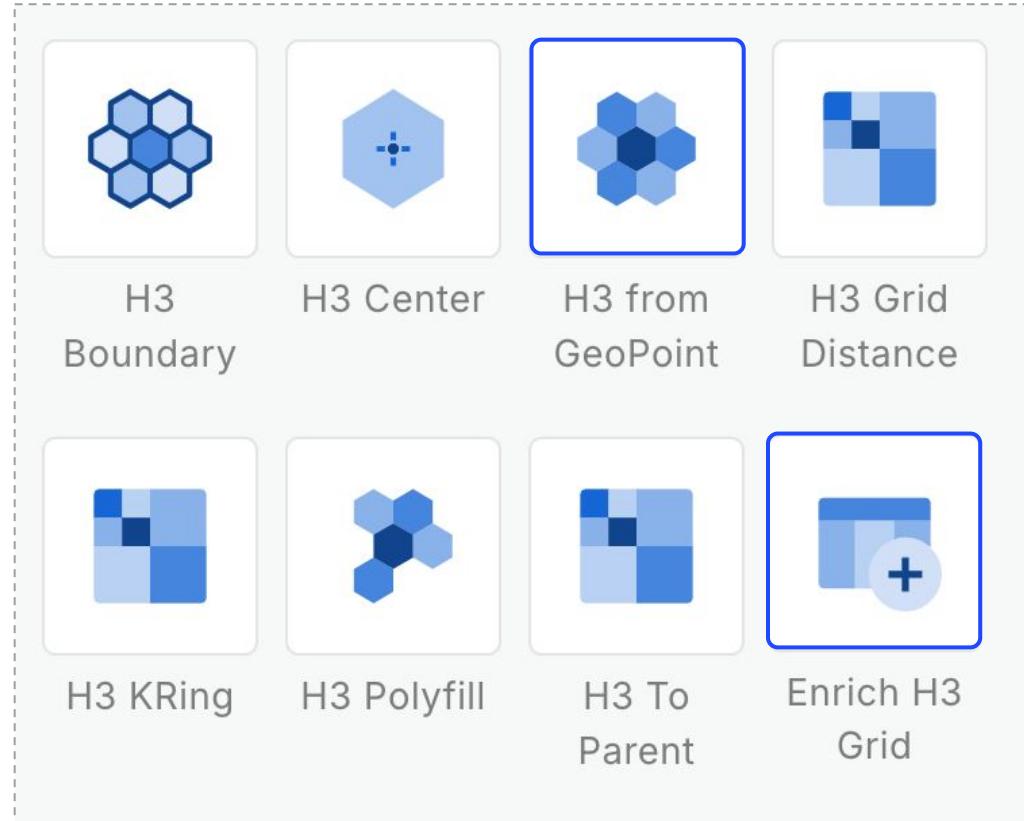
(0,0)	(1,0)	(2,0)	(3,0)	(4,0)	(5,0)	(6,0)	(7,0)
(0,1)	(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)	(7,1)
(0,2)	(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)	(7,2)
(0,3)	(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)	(7,3)
(0,4)	(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)	(7,4)
(0,5)	(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)	(7,5)

Quadbin



H3

Spatial Indexes Operations



We've already...

1. Transformed all input sources into a regular grid of Spatial Indexes through Enrichments
2. Reduced the dimensionality of the enriched data to reduce model complexity (we will use this data to model crime counts) using the **Factorial Analysis of Mixed Data** algorithm

h3_string	acs_year	number	total_pop,sum	number	median_age,avg	number	median_rent,avg	number	black_pop,sum	number	hispanic_pop,sum	number	owner_occupied_housing_units,median_value,sum	number	vacant_housing_units,sum	number	housing_units,sum	number	families_with_young_children,sum	number	urbanity,any	string	urbanity,any,ordinal	number
872759343fffff	2,020	307,795	33.3	1,010	15,025	213,888	47,903,738	5,745	99,431	23,643	rural	0												
872759343fffff	2,011	242,27	34.5	912	7,061	164,382	42,011,061	7,502	83,625	22,727	rural	0												
872759343fffff	2,019	306,248	31.1	978	14,141	209,468	42,733,316	6,187	100,315	16,793	rural	0												
872759343fffff	2,010	238,298	27.9	934	5,075	168,794	41,591,833	7,943	77,006	16,769	rural	0												
872759343fffff	2,007	272,498	26.4	877	16.99	176,076	76,078,855	8,164	89,582	12,797	rural	0												
872759343fffff	2,013	273,381	32.4	941	7,281	193,286	39,716,34	7,723	85,611	23,389	rural	0												
872759343fffff	2,014	288,164	32.1	953	0	214,027	38,723,431	4,413	96,864	22,506	rural	0												
872759343fffff	2,009	278,456	26.7	949	26,257	179,165	63,435,82	12,577	92,671	19,417	rural	0												
872759343fffff	2,008	262,569	26.2	913	15,666	159,969	76,277,437	9,267	82,08	15,004	rural	0												
872759343fffff	2,006	268,306	26.1	850	16,534	182,357	78,042,807	7,723	88,038	20,52	rural	0												
872759343fffff	2,005	268,306	26.1	850	16,534	182,357	78,042,807	7,723	88,038	20,52	rural	0												

h3_string	acs_year	number	principal_component_3	number	principal_component_2	number
872759343fffff	2,020	2,595	-1.772			
872759343fffff	2,011	2,731	-1.728			
872759343fffff	2,019	2,565	-1.934			
872759343fffff	2,010	2,517	-1.169			
872759343fffff	2,007	2,526	-2.293			
872759343fffff	2,013	2,639	-1.861			
872759343fffff	2,014	2,618	-1.879			
872759343fffff	2,009	2,461	-2.242			
872759343fffff	2,008	2,485	-2.292			
872759343fffff	2,006	2,541	-2.326			

11 socio-demo variables → 2 principal components

Factorial Analysis of Mixed data

- **Generalizes the use of PCA** to account for the number of modalities available to each categorical/ordinal variable and on the probabilities of these modalities.
- Depending on the variable type, the procedure applies the **different transformations** to the input data

BUILD_PCAMIX_DATA

`BUILD_PCAMIX_DATA(input_query, index_column, cols_num_arr, cols_cat_arr,`

Description

Prepares the input data for the [BUILD_PCAMIX_MODEL](#) procedure.

This procedure is tested against the R package [FactoMineR](#), which adopts the Factorial Analysis of Mixed Data (FAMD) method developed by [Pagés \(2004\)](#). The same method is applied here and generalizes the use of PCA to account for the number of modalities available to each categorical/ordinal variable and on the probabilities of these modalities.

Depending on the variable type, the procedure applies the following transformations to the input data:

- For the numerical variables: standard scale the columns to get their z-scores
- For the categorical variables:
 - One-hot-encode the categorical columns to get their indicator matrix
 - Weight each column by the inverse of the square root of its probability, given by the number of ones in each column (N_s) divided by the number of observations (N)
 - Center the columns

Factorial Analysis of Mixed data - Categorical

One-hot encoding

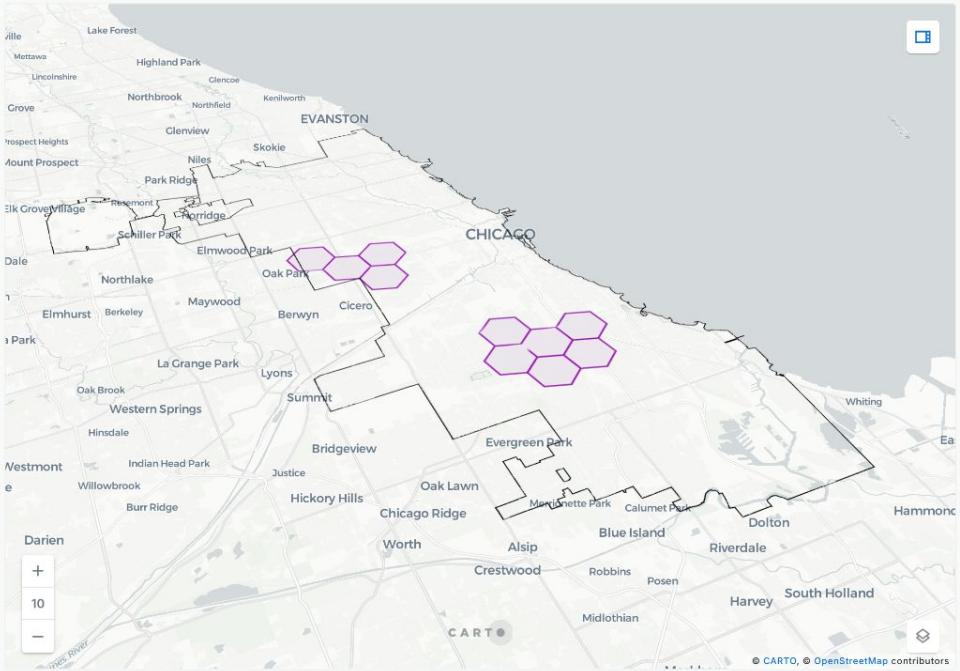
- Converts categorical variables into a binary format
- Ensures that each category is treated independently, preventing the PCA from mistakenly assuming any sort of ordinal relationship

Weighting

- Weight each OHE column by the inverse of the square root of its probability
- Ensures that less frequent categories, which may still carry meaningful information, are given more importance

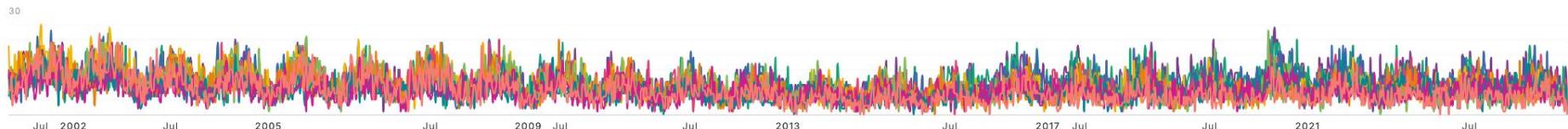
Centering

- Zero-center the data (subtracting the column mean)
- Makes sure that all features have a mean of 0 to avoid bias.



Violent crime counts (Top 10 H3 cells)

01/01/2001 - 30/12/2024



Legend (Top 10 H3 cells):

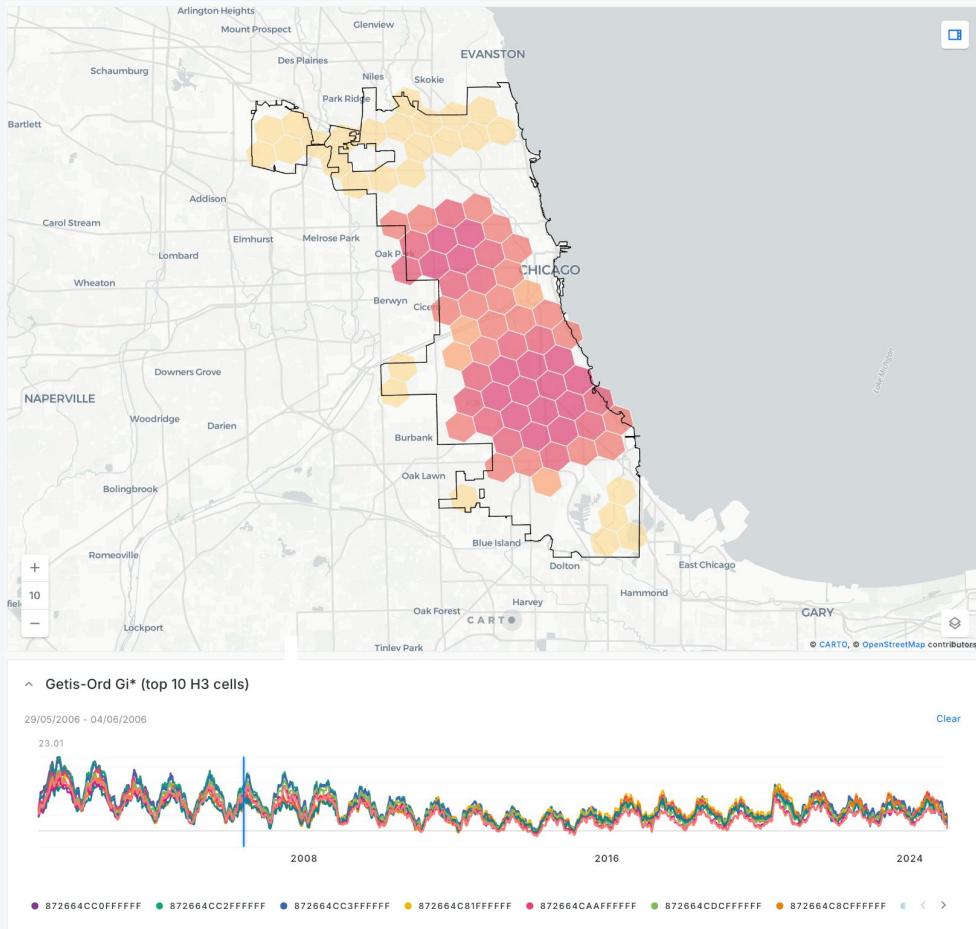
- 872664C8CFFFFF
- 872664C81FFFFFF
- 872664CAAFFFFF
- 872664CC2FFFFFF
- 872664CDCFFFFF
- 872664CD1FFFFFF
- 872664CC0FFFFFF
- 872664CC3FFFFFF
- 872664C80FFFFFF
- 872664CD5FFFFFF

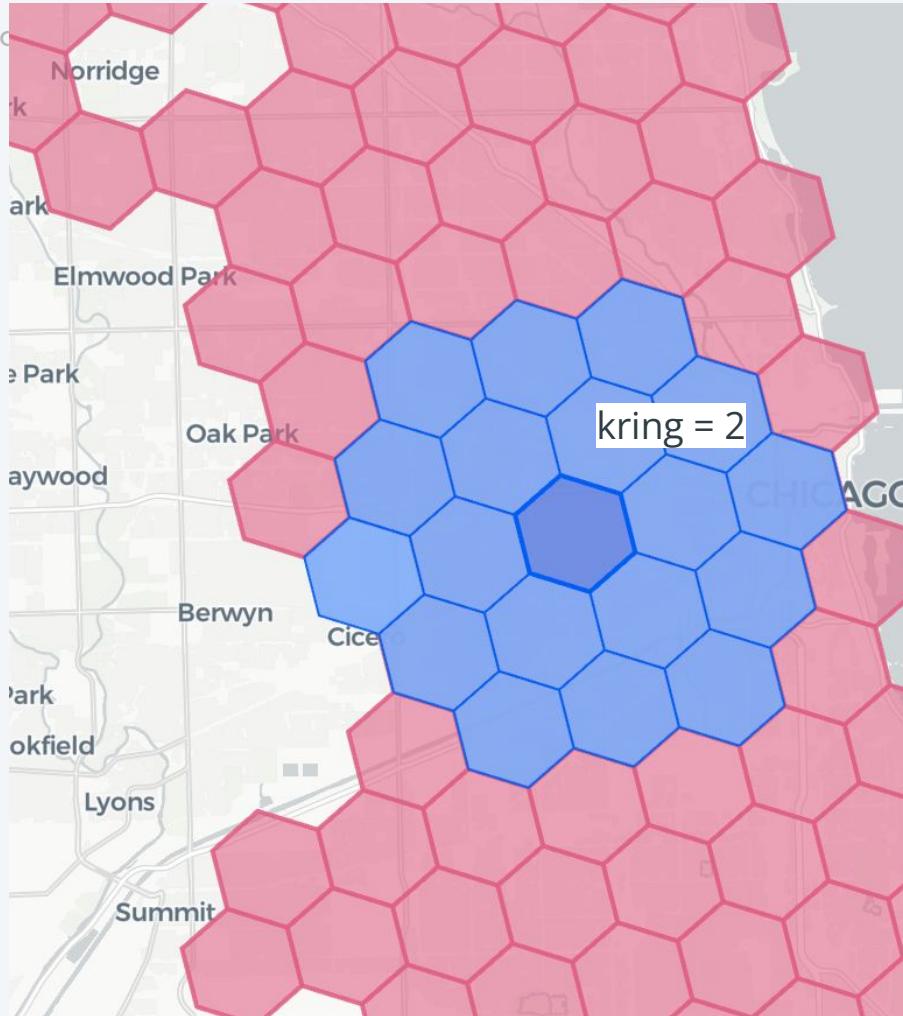
Exploratory Analysis: space-time insights

Hotspot classification

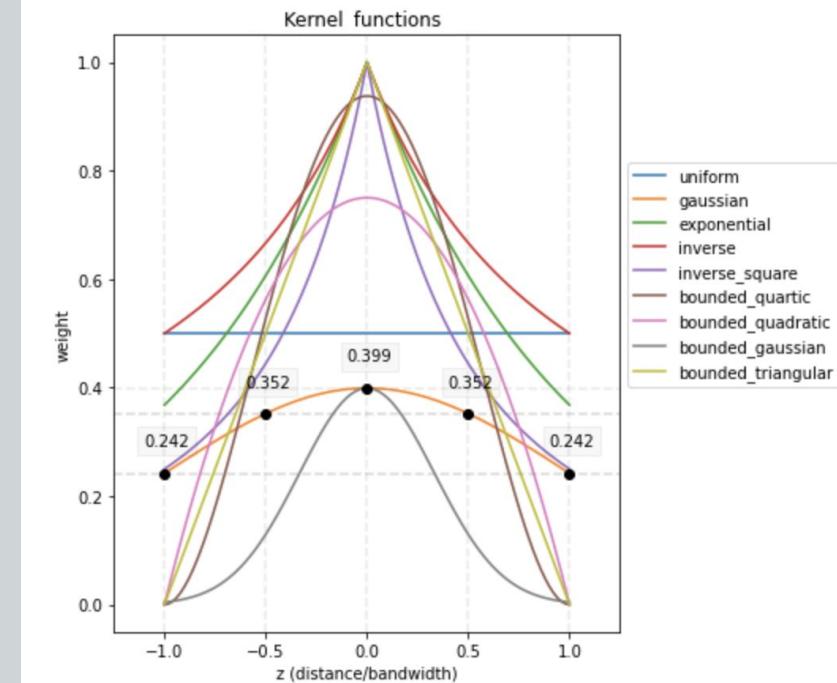
Space-time Getis-Ord

- **Identifies hot and cold spots:** high positive Z-scores indicate hot spots, while low negative Z-scores indicate cold spots.
- Allows to **compare space-time patterns:**
 - Summer vs Winter
 - 2021 vs 2022



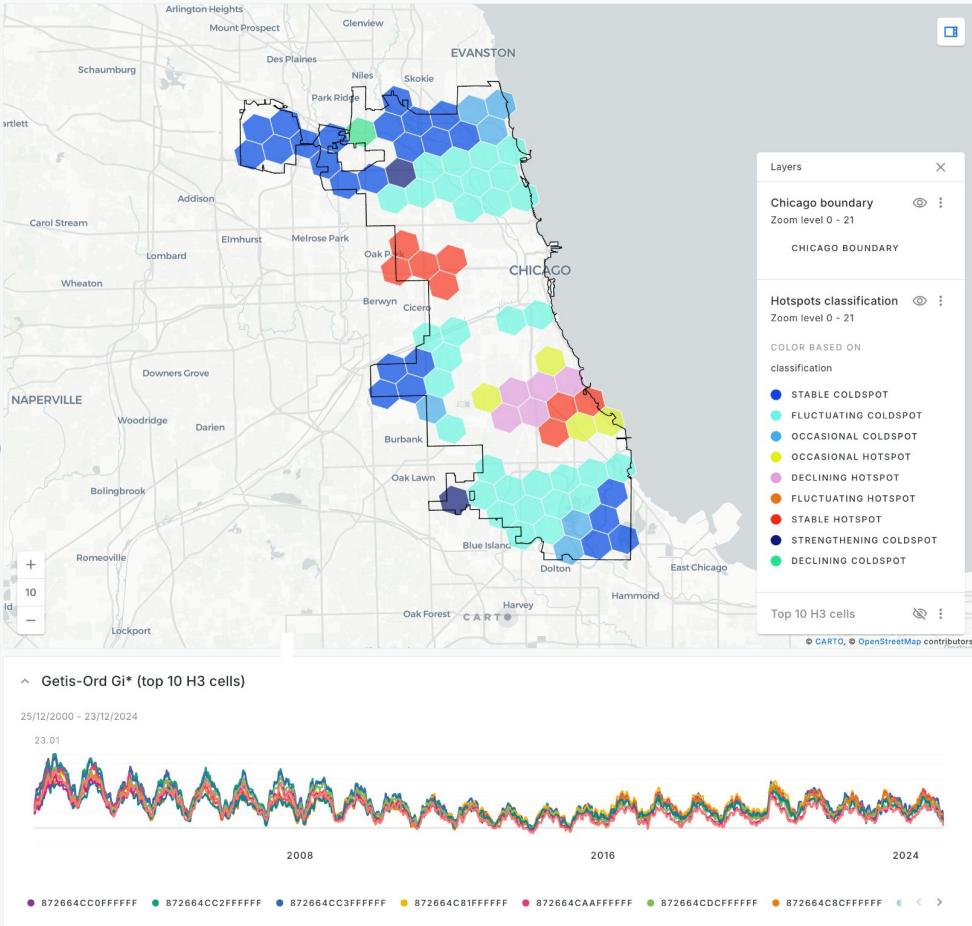


weights

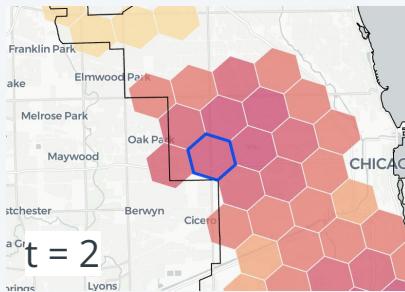
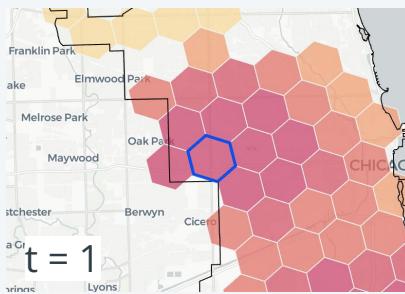
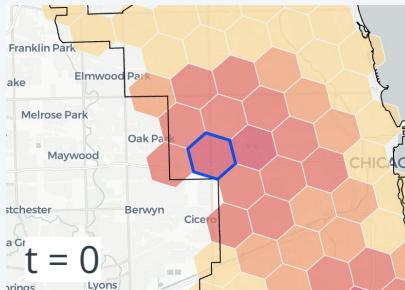


Hotspot Classification

- Serves as a digest of space-time Getis-O
- Will collapse Gi time series to a set of predefined categories
 - West Garfield Park: stable hotposts
 - Suburban villages: coldspots
 - Englewood: declining hotposts?



Category	Description
Undetected Pattern	This category applies to locations that do not exhibit any discernible patterns of hot or cold activity as defined in subsequent categories.
Incipient Hotspot	This denotes a location that has become a significant hotspot only in the latest observed time step, without any prior history of significant hotspot activity.
Sequential Hotspot	Identifies a location experiencing an unbroken series of significant hotspot activity leading up to the most recent time step, provided it had no such activity beforehand and less than 90% of all observed intervals were hotspots.
Strengthening Hotspot	A location consistently identified as a hotspot in at least 90% of time steps, including the last, where there's a statistically significant upward trend in activity intensity.
Stable Hotspot	Represents a location maintaining significant hotspot status in at least 90% of time steps without showing a clear trend in activity intensity changes over time.
Declining Hotspot	A location that has consistently been a hotspot in at least 90% of time steps, including the most recent one, but shows a statistically significant decrease in the intensity of its activity.
Occasional Hotspot	Locations that sporadically become hotspot, with less than 90% of time steps marked as significant hotspots and no instances of being a significant coldspot.
Fluctuating Hotspot	Marks a location as a significant hotspot in the latest time step that has also experienced significant coldspot phases in the past, with less than 90% of intervals as significant hotspots.
Legacy Hotspot	A location that isn't currently a hotspot but was significantly so in at least 90% of past intervals.



Input data

Chicago boundary enriched and pre-processed (FAMD) data



Hotspots analysis

Hotspot analysis identifies and measures the strength of spatio-temporal patterns by taking the user-defined neighborhood for each feature in a spatial table and calculating whether the values within that neighborhood are significantly higher or lower than across the entire table.

Find space-time hotspots

Calculate the spatio-temporal Getis-Ord Gi* (Gi*) statistic. Positive Gi* values indicate that the neighborhood values are significantly higher than across the entire table - i.e. it is a hotspot - and negative Gi* values indicate the reverse.



✓ Getis Ord Spacetime
145,348



✓ Rename Column
145,348

Classify space-time hotspots

Once we have identified hot and cold spots, we can classify them into a set of predefined categories so that the results are easier to digest.



✓ Select Distinct
9



✓ Spacetime Hotspots Classification
116

Save results to a table



✓ Save as Table
145,348



✓ Save as Table
116



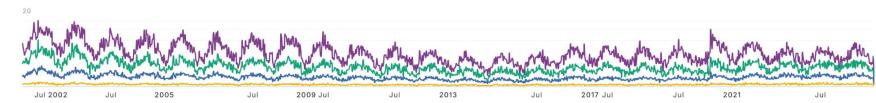
Exploratory Analysis: space-time insights

Time-series clustering

Time Series Clustering

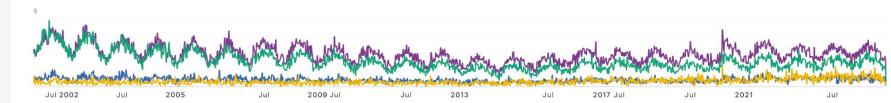
Value characteristic

- Clustering based on **step-by-step distance of the time series values**
 - K-Means with **Euclidean** distance
- The closer the **signals**, the higher the chance of being clustered together



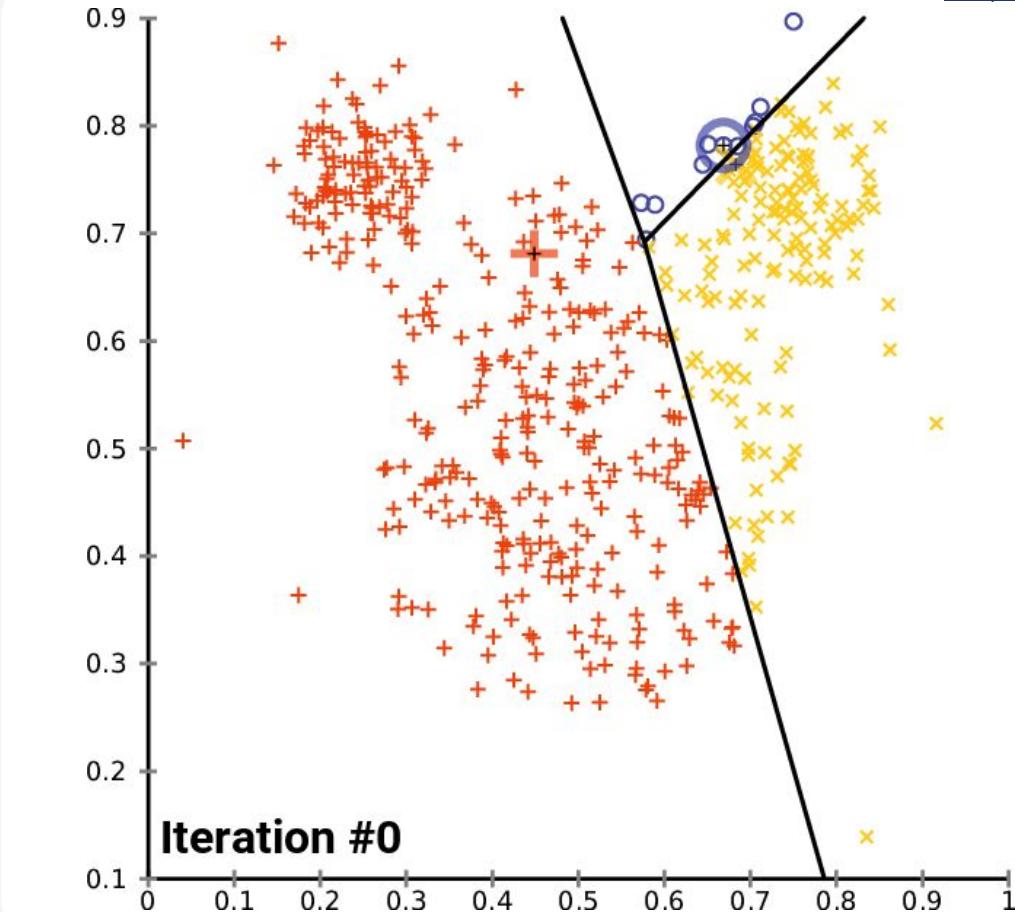
Profile characteristic

- Clustering based on the time series **dynamics** along the timespan passed
 - K-Means with **Cosine** distance
- The closer the **correlation**, the higher the chance of being clustered together

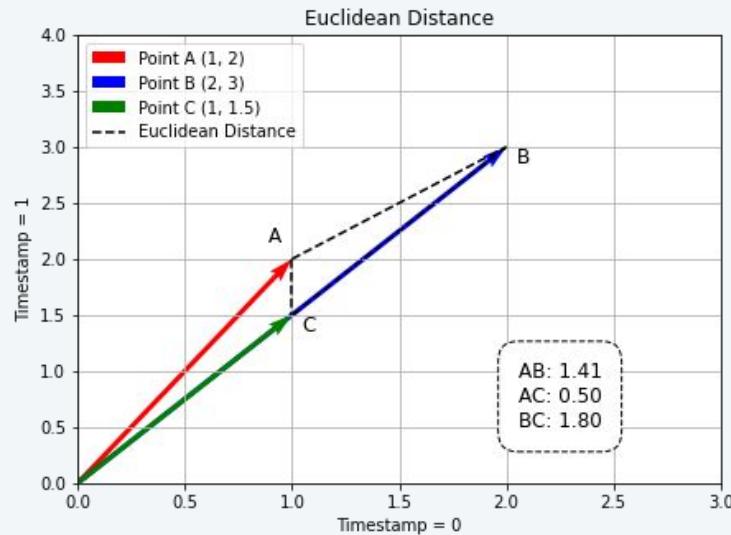


K-means clustering

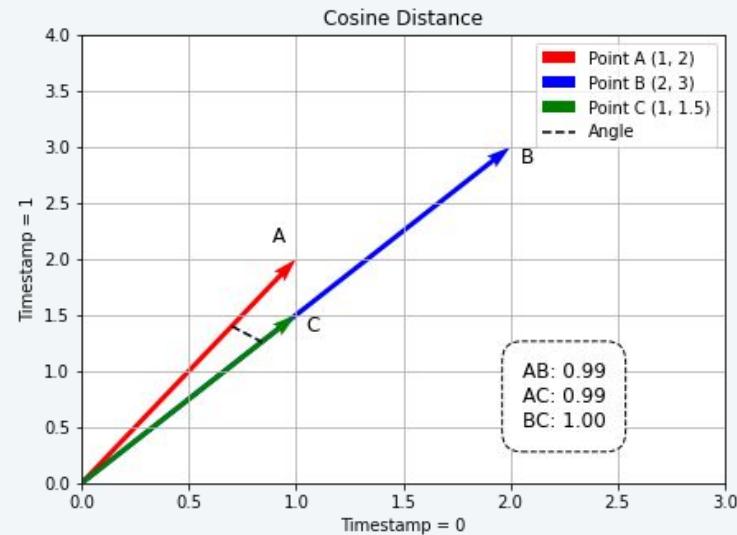
- **Unsupervised learning** algorithm that groups data into K clusters based on their **similarity**
- **Goal:** minimize the variance within each cluster, ensuring that similar points are grouped together
- It starts by randomly selecting K cluster centers (centroids), then **iteratively** assigns each data point to the nearest centroid and recalculates it until convergence



Value characteristic



Profile characteristic



Input data



Time series clustering

Time series clustering is a method used to group time series data into clusters based on similarities in their temporal patterns. It aims to find natural groupings among time series, where series within the same cluster are more similar to each other than to those in other clusters.

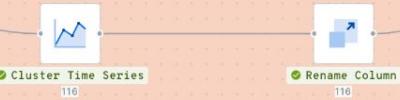
By value

The series is clustered based on the **step-by-step distance of its values**. One way to think of it is that the closer the signals, the closer the series will be understood to be and the higher the chance of being clustered together.



By profile

The series is clustered based on their **dynamics along the time span passed**. This time, the closer the correlation between two series, the higher the chance of being clustered together.

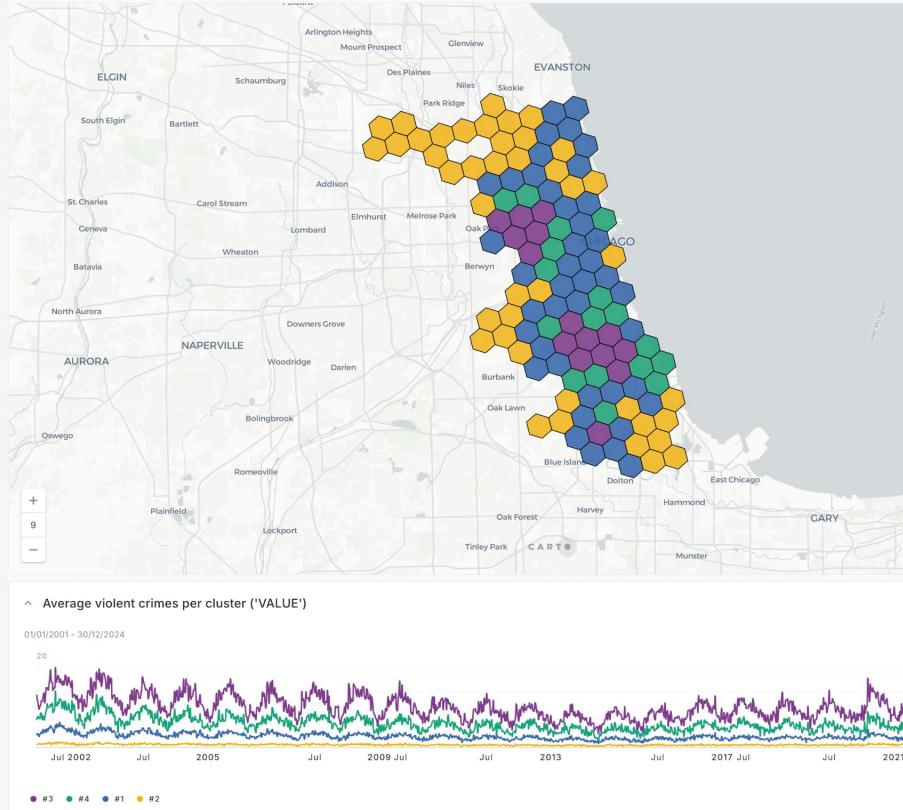


Save results to a table



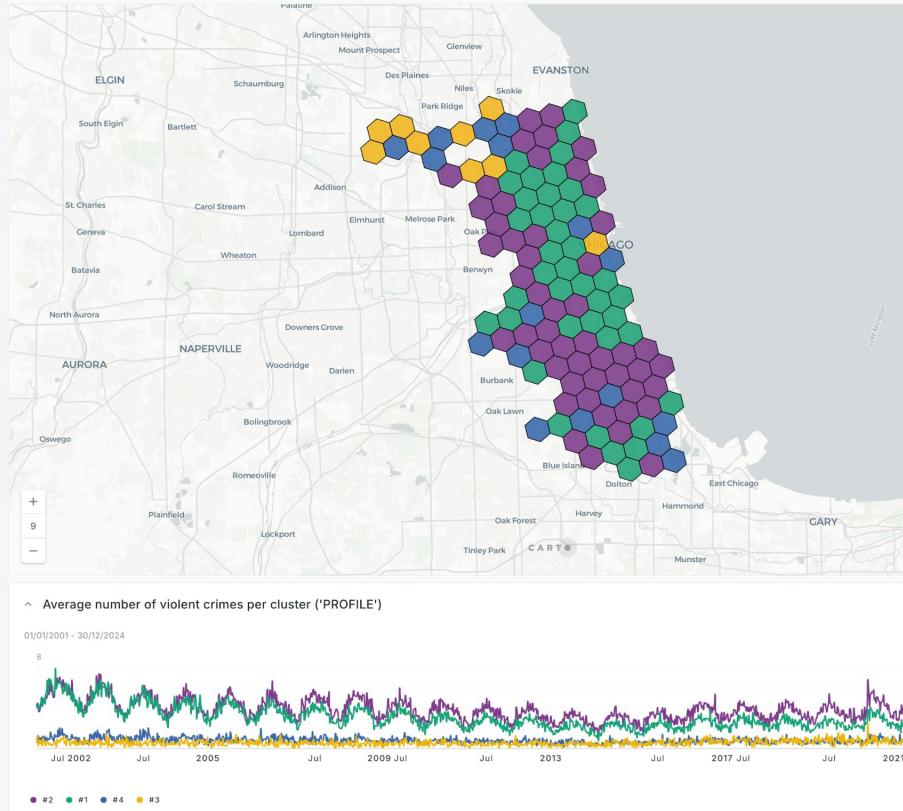
Time Series Clustering - Value

- The closer the **signals**, the higher the chance of being clustered together
 - We can clearly identify areas with very different crime levels, from group *purple* with very high-levels to group *yellow* with very low values.



Time Series Clustering - Profile

- The closer the **correlation**, the higher the chance of being clustered together
 - We can identify groups of time series with different seasonalities and trends (e.g. *purple* and *green* are characterized by a large seasonal cycle with different trends, while *yellow* and *blue* do not show any seasonal variability).



Inferential Analysis

Estimate the expected counts

Extension Packages

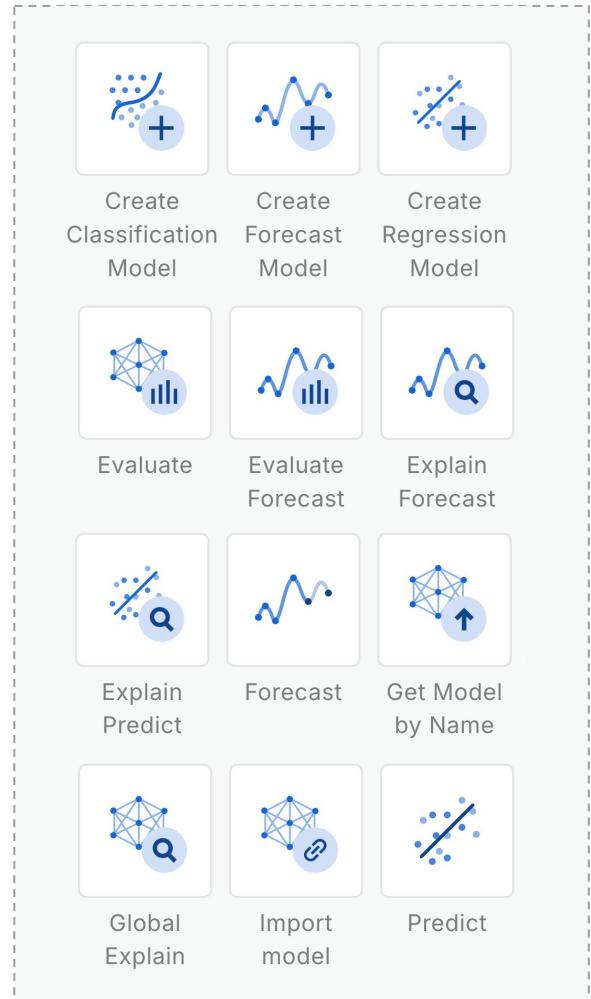
[Workflows Extension Packages](#) is a powerful tool that allows users to extend the functionality of Workflows by creating packages of **custom** components through **SQL** stored procedures, allowing users to **design and integrate components** that are **tailored to their specific needs**.

The screenshot shows the 'Extension packages' section within the CARTO Workflows interface. The top navigation bar includes 'CARTO Workflows / Untitled' and 'Updated'. Below the navigation is a search bar and three tabs: 'Explore' (selected), 'Installed', and 'Upload'. A sidebar on the left lists categories like 'Aggregations', 'CQL', 'Custom', 'Summaries', and 'Cursors'. The main area is titled 'CARTO Extension packages' and displays four cards:

- AI AND ML**
BigQuery ML for Workflows
by CARTO
Version 1.0.1
- EARTH OBSERVATION**
Google Earth Engine
by CARTO
Version 1.2.0
- ENVIRONMENT AND CLIMATE**
Google Environment APIs
by CARTO
Version 1.2.0
- TELCO**
Telco Signal Propagation
by CARTO
Version 1.1.0

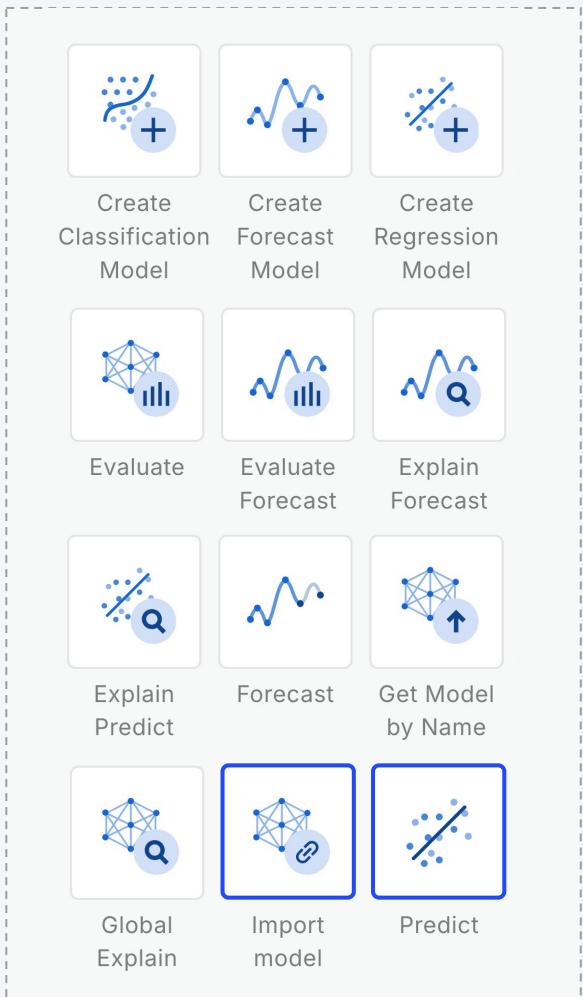
BigQuery ML for Workflows

- Allows users to exploit [BigQuery's ML capabilities](#) directly from Workflows, enabling a seamless integration of machine learning models into automated pipelines.
- With minimal coding required, users can quickly build, train, and deploy predictive models using data stored in Google BigQuery.
- All components are created on top of BigQuery ML's capabilities, with each component invoking a specific BQ ML procedure.



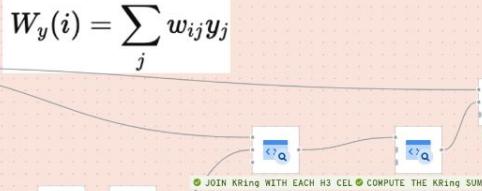
We will...

1. Extract and select features from the data
2. Train a model using Python's scikit-learn to estimate the expected crime counts per 1000 people
3. Save the model to ONNX (Open Neural Network Exchange) format
4. Import the model into BigQuery using Workflows
5. Run predictions



Input data**Create spatial lag variable**

We can add spatial-lag variables to account for the influence of neighboring or nearby regions on the variable of interest in a given location. They are derived from the idea that outcomes in one location might not be independent but influenced by outcomes in other nearby locations due to spatial interactions or spillover effects.

$$W_y(i) = \sum_j w_{ij} y_j$$
**Add temporal lags and seasonal terms**

- **Temporal lag variables** are used in time series analysis to account for the effect of past values of a variable on its current or future values. Temporal lags capture the relationship between a variable at one point in time and its previous observations, helping model delayed effects or persistence over time (a.k.a. autocorrelation)
- **Seasonal terms** can be added to model repeating seasonal behaviors. These can be represented as Fourier terms, i.e. as a periodic function by summing sine and cosine functions of different frequencies,

$$\sum_{k=1}^K \left[a_k \sin\left(\frac{2\pi k t}{T}\right) + b_k \cos\left(\frac{2\pi k t}{T}\right) \right]$$

**Save results to a table**



File Edit View Insert Cell Kernel Widgets Help
Not Trusted Python 3 (ipykernel) ○
File + % Run Cell Code

```
In [3]: import os
import pandas as pd
import numpy as np
from google.cloud import bigquery

%matplotlib inline
from matplotlib import pylab as plt
import matplotlib.dates as mdates
import seaborn as sns

import collections
```

```
In [4]: ## ****
## BIG QUERY CREDENTIALS
## ****
bq_client = bigquery.Client.from_service_account_json('my-service-account-key.json')

#!gcloud auth application-default login
#bq_client = bigquery.Client()
```

```
In [5]: bq_table = 'cartobq.sdsc24_ny_workshops.CHI_boundary_enriched_w_lags'
selected_h3 = '872664c8fffff'
```

Plotting setup

```
In [6]: import seaborn as sns
from pandas.plotting import register_matplotlib_converters
register_matplotlib_converters()
```

Input data

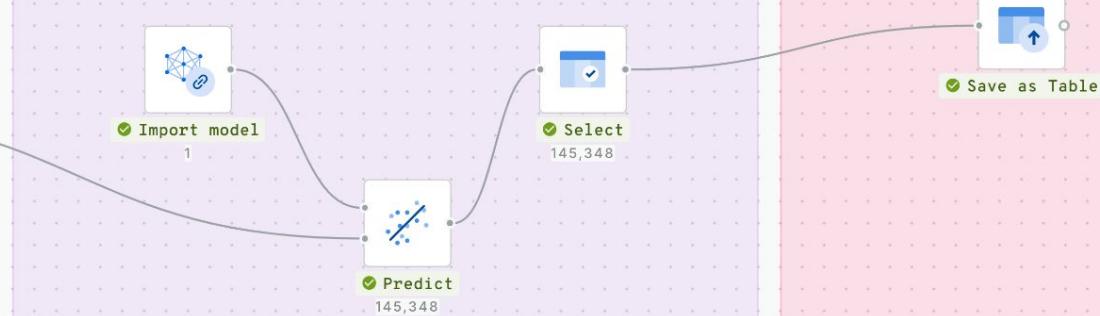
Chicago boundary enriched and pre-processed (FAMD) data:



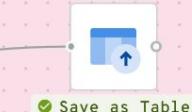
CHI_boundary_en...
145,348 RECORDS

Estimate the expected value of the series conditional on the selected covariates

To estimate the expected crime rate per 1,000 people, we can import a pre-trained model from GCS using scikit-learn. This approach is useful when training and prediction are handled in separate processes (e.g., a data scientist trains the model, while an analyst performs the forecasting) or for specific use cases such as online predictions or predictions based on different data sets.

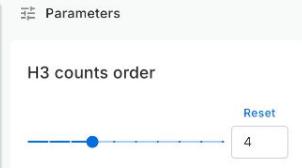
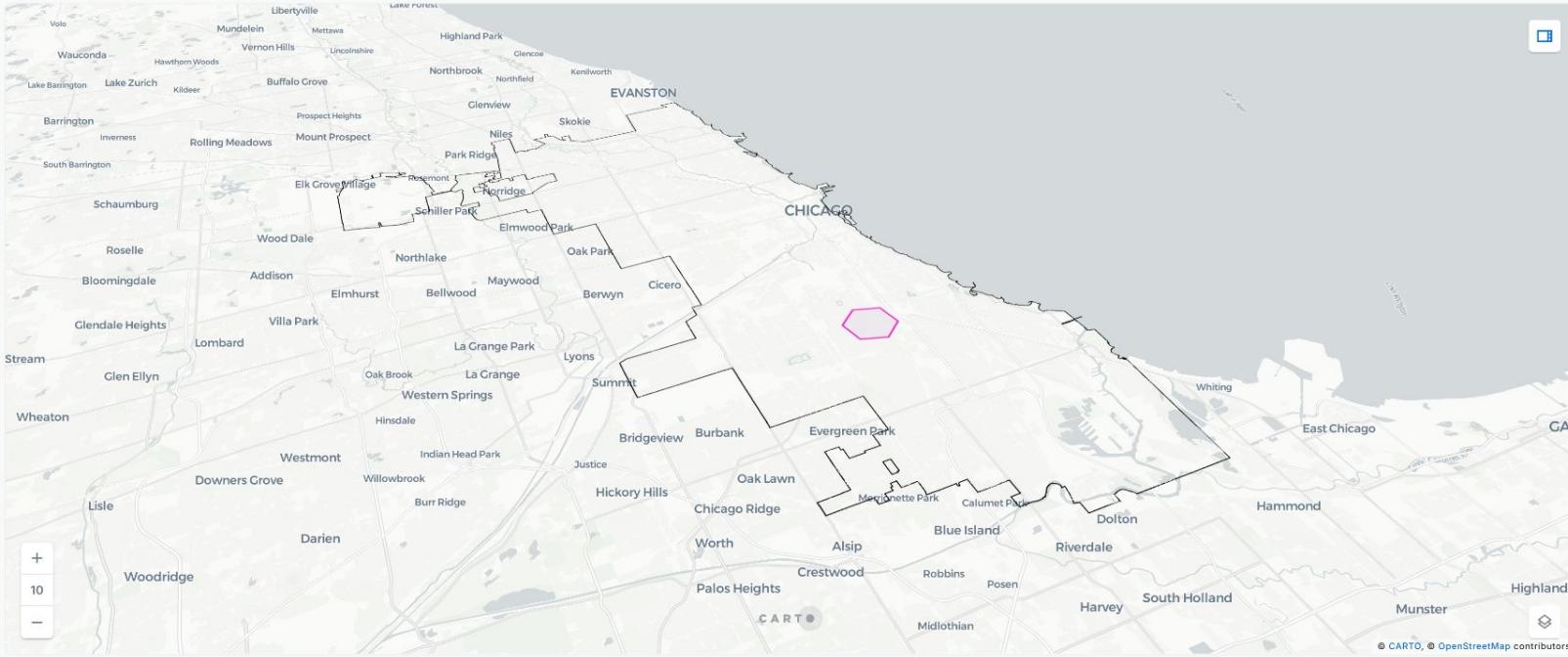


Save results to a table



Save as Table





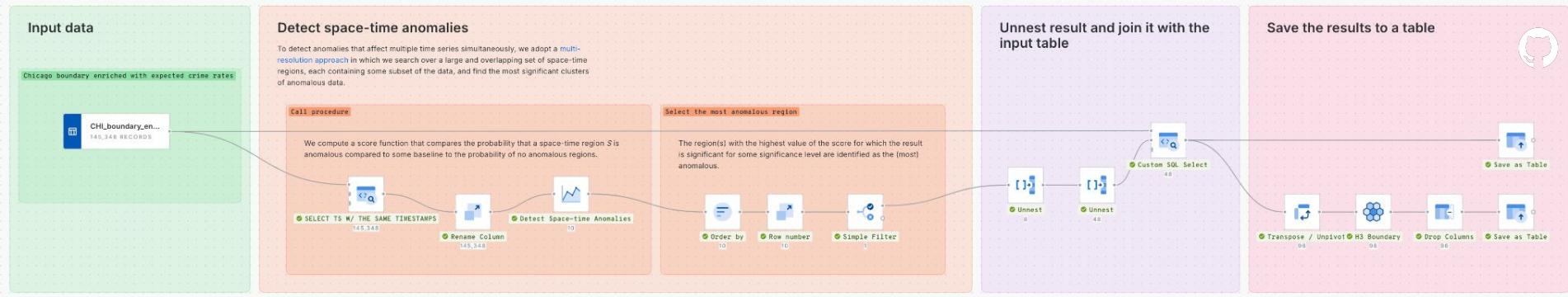
Inferential Analysis

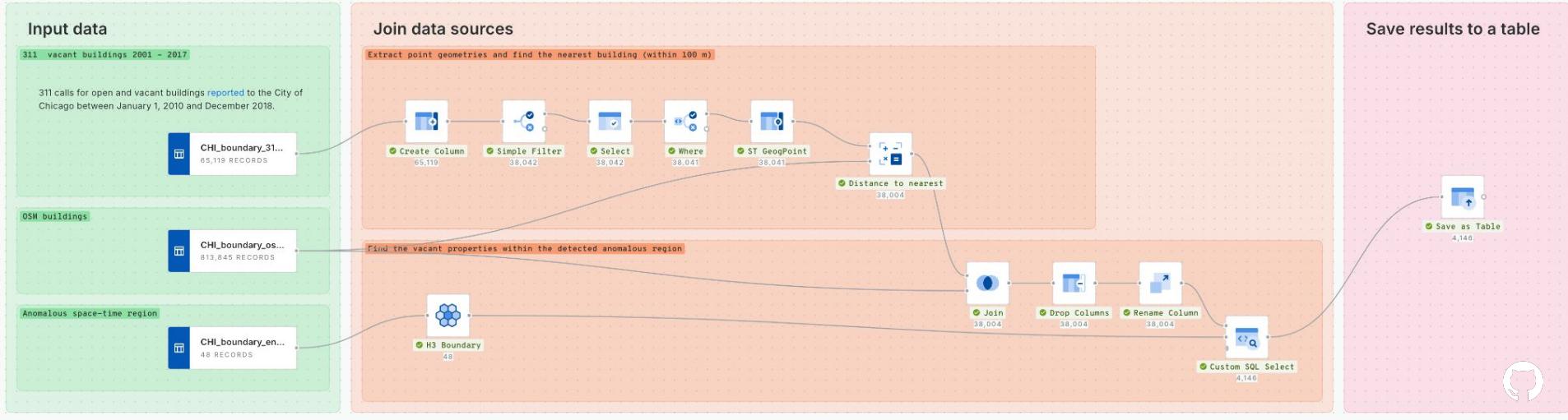
Detect emerging space-time anomalies

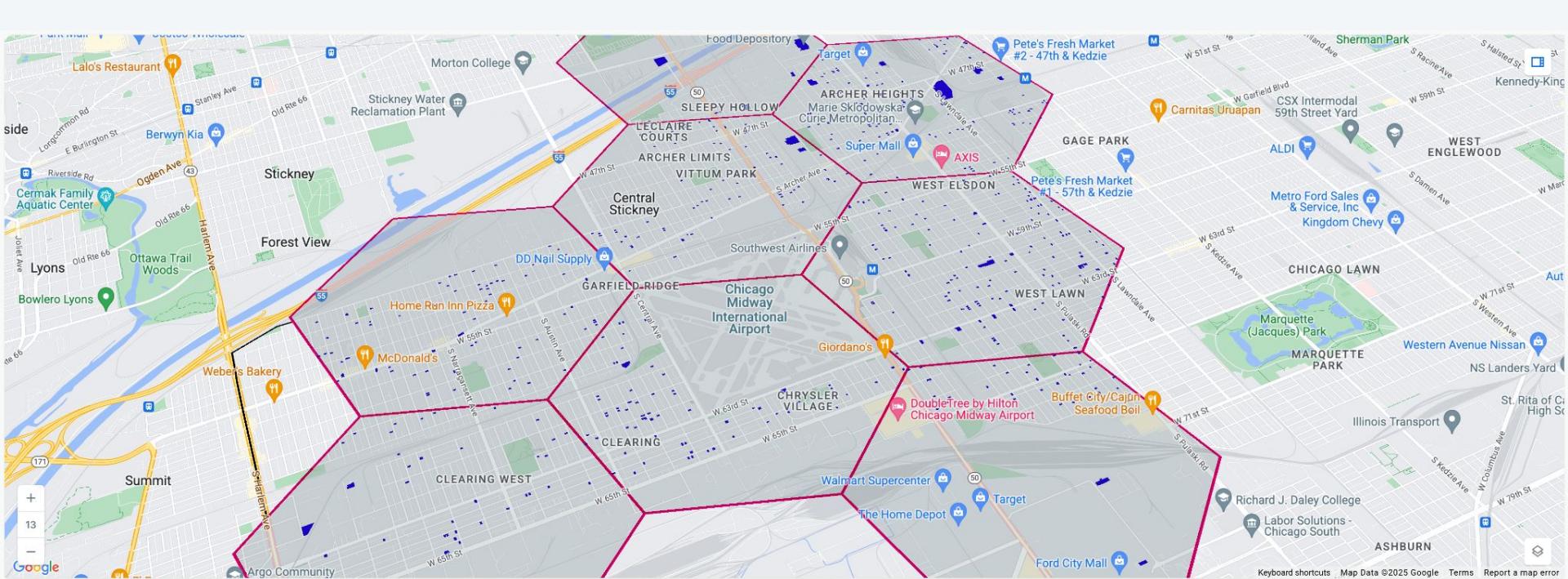
Space-time Anomaly Detection

- Based on the **generalized space-time scan statistics framework**:
 - Compute a score function that compares the probability that a space-time region is anomalous compared to some baseline to the probability of no anomalous regions.
 - The region(s) with the highest value of the score for which the result is significant for some significance level are identified as the (most) anomalous.
- Under the null hypothesis of no anomalous space-time regions, we assume **the observed values should be equal to the baseline values** given by the expected counts estimated in the previous section.









^ Observed VS. baseline counts

25/11/2024 - 30/12/2024



• COUNTS • PREDICTED_COUNTS

Closing thoughts

CARTO

Thanks!

Follow @CARTO on Twitter