CARTO

DeepFin Investor Virtual Tutorial

Spatial Data Science with CARTO

Introductions

Giulia Carella, PhD

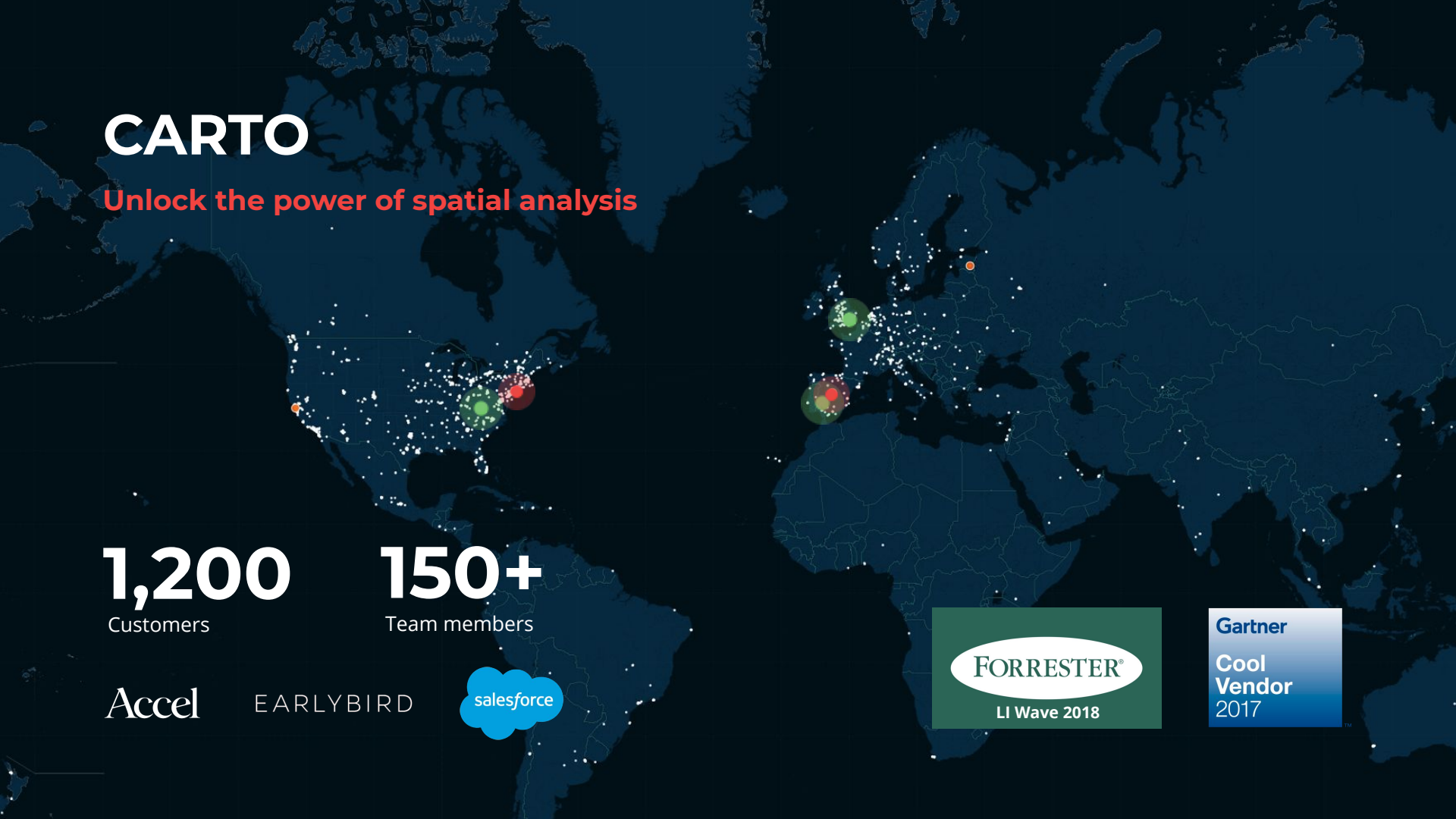
Data Scientist
at CARTO

Miguel Álvarez

Data Scientist
at CARTO

CARTO

Unlock the power of spatial analysis



1,200
Customers

150+
Team members

Accel

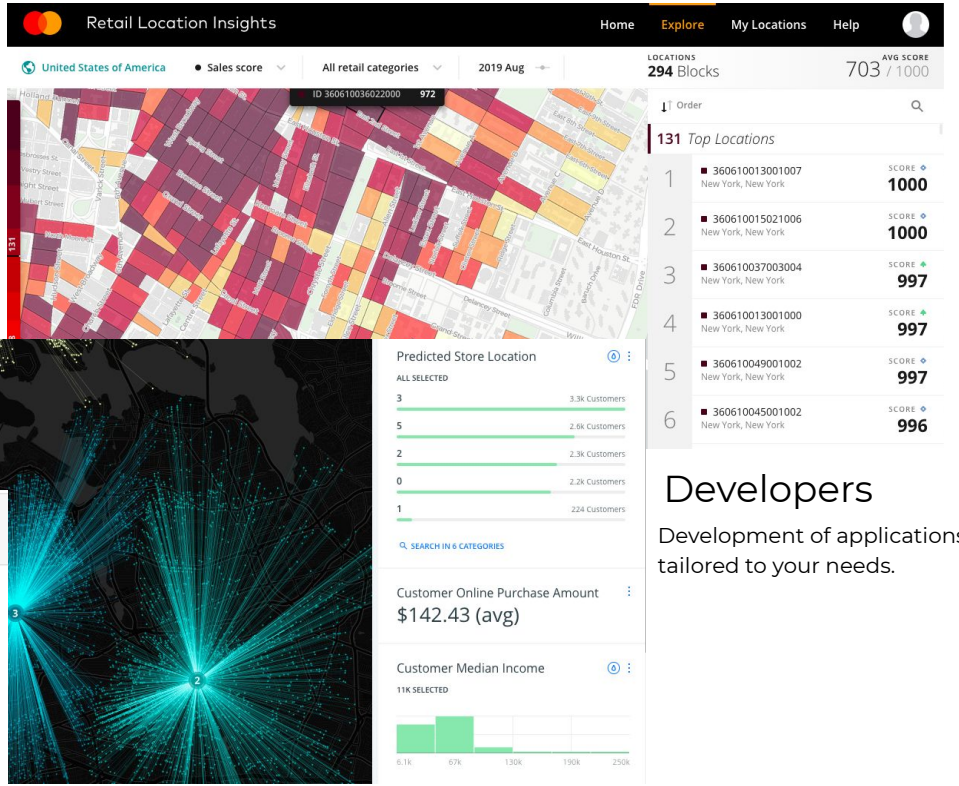
EARLYBIRD

salesforce

FORRESTER®
LI Wave 2018

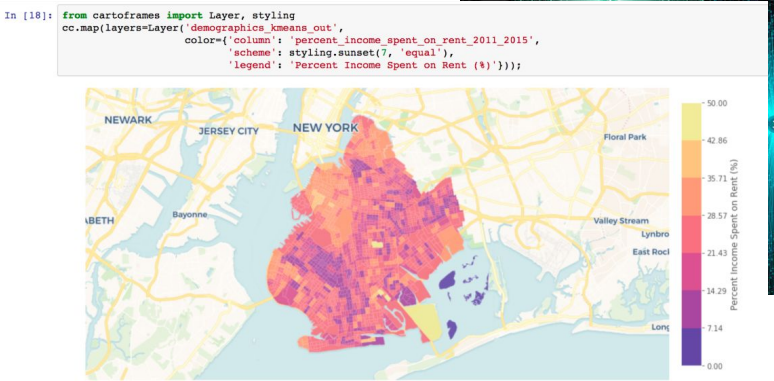
Gartner
Cool Vendor
2017

A comprehensive platform for Spatial Analysis



Developers

Development of applications tailored to your needs.



Data Scientists

Integrate data, analysis and visualization into data science flows

Data Analysts

Tools to enable business users to analyze data and create lightweight dashboards without a single line of code.

CARTO is a B2B SaaS Platform

We offer our customers end-to-end solutions supported by:

Technology

Managed cloud or on-premises platform

Data

Open and premium location data streams

Services

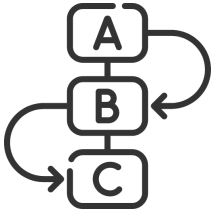
Ongoing enablement programs and custom engagements

Why CARTO for Spatial Data Science?

30% | 30% | 20% | 20%

Discovering data useful for their analysis

Evaluating and purchasing data



ETLing the data into common structures

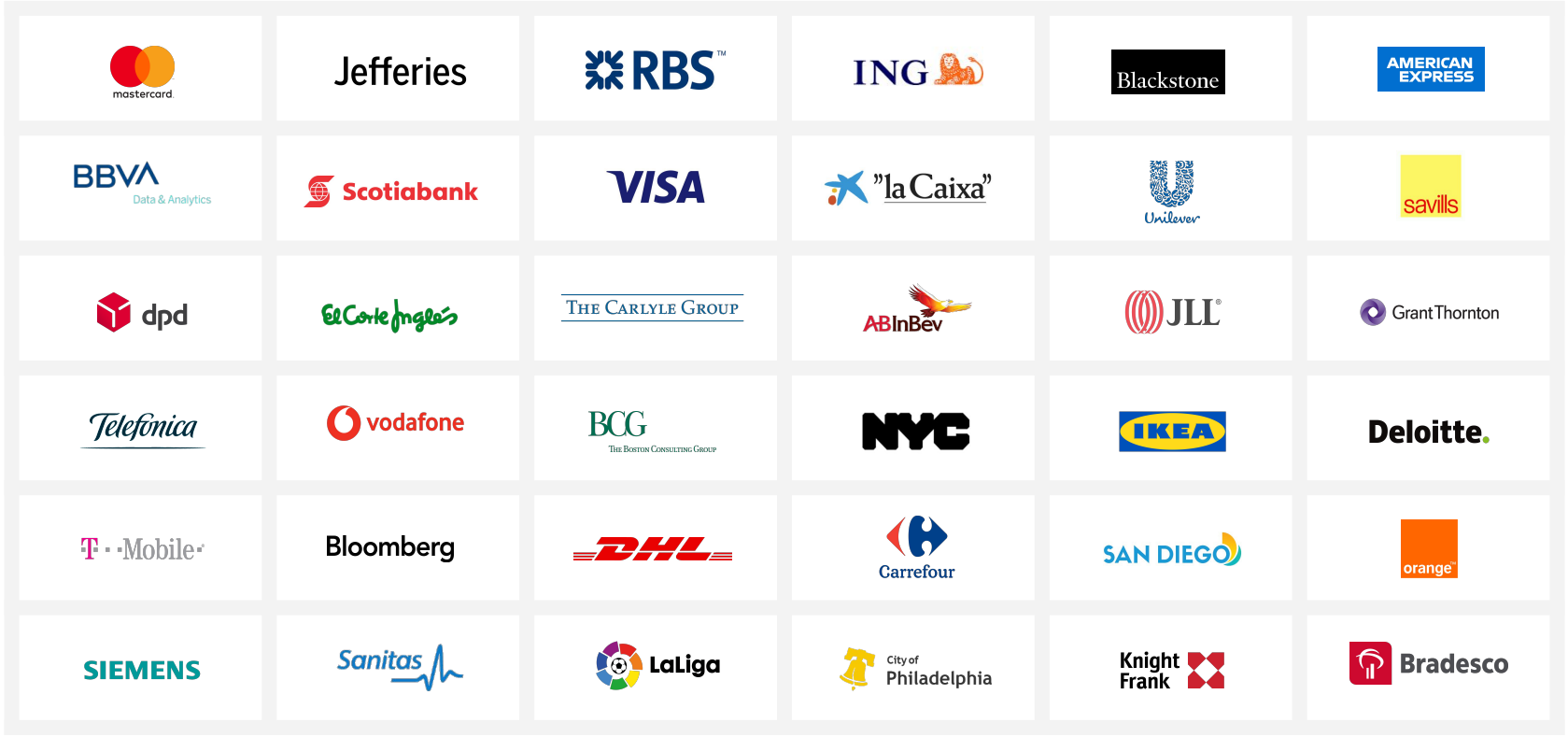
Analyzing, doing feature extraction and modeling

	CARTO	ESRI	OPEN SOURCE
Data discovery and access	✔ Data Observatory	🔶 Demographics, Landscape	🔶 POI (OSM), Demographics
Viz	✔ CARTOframes	✔	🔶 Geopandas, Folium
Geo-functions e.g. spatial joins	✔ CARTOframes	✔	🔶 Geopandas, Shapely, QGIS
LBS services e. g. geocoding, route engine, isochrones	✔ CARTOframes	✔	🔶 Openrouteservice
Easy to integrate with other DS tools Python SQL	✔	🔶	🔶

For a wide range of industries:

- Banks
- Cities & Government
- CPG
- Credit Card Providers
- Environmental & Natural Resources
- Healthcare & Pharma
- Insurance
- Investment Firms & Funds
- Logistics
- Marketing & Advertising
- Real Estate
- Retail
- Telco
- Transport
- Utilities

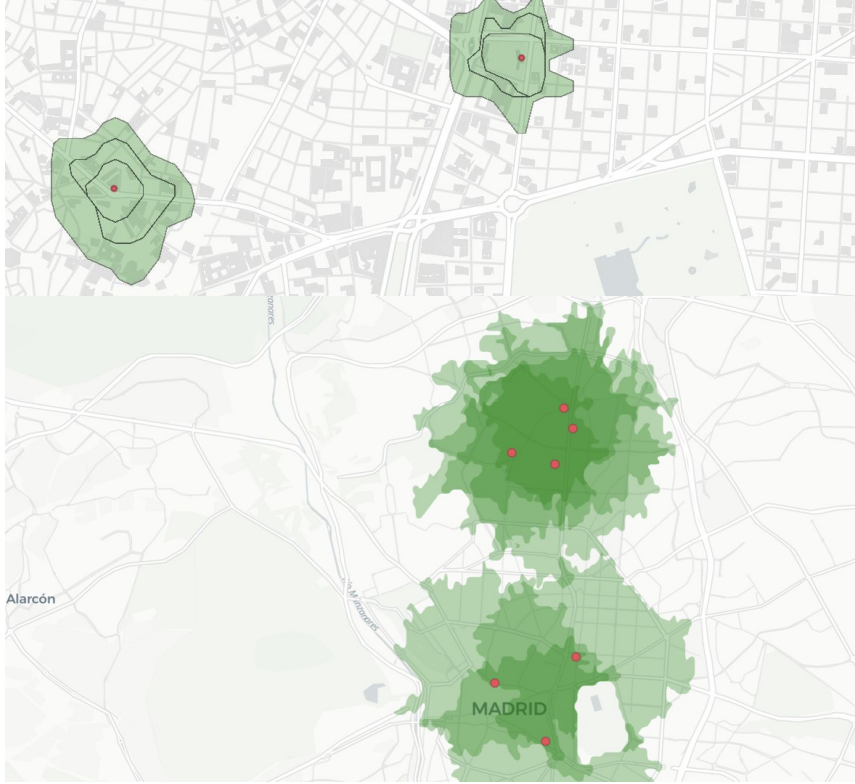
Market-leaders trust CARTO



What is Spatial Data Science?

09:10 a.m. - 09:40 a.m.

Spatial data science takes advantage of traditional GIS tools to manipulate, analyze and visualize geospatial data

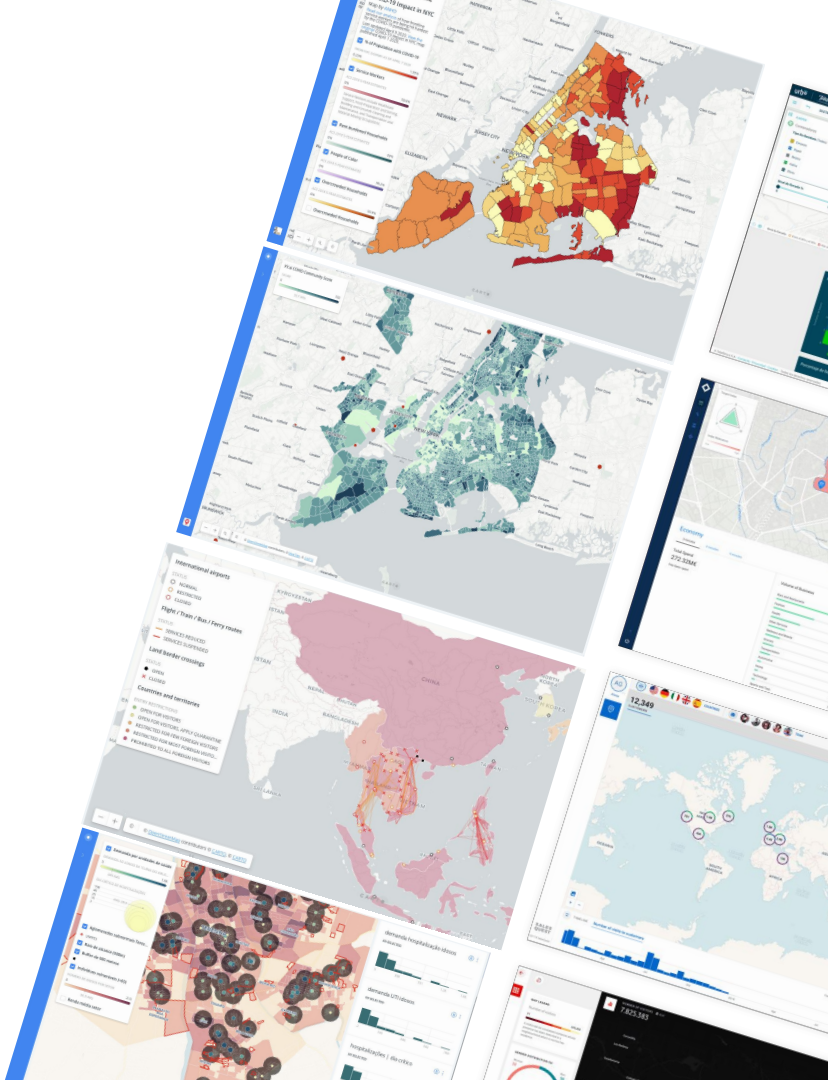


	the_geom	address
0	POINT (-3.70588 40.42049)	Gran Vía 46
1	POINT (-5.98312 37.35547)	Ebro 1

What is Spatial Data Science?

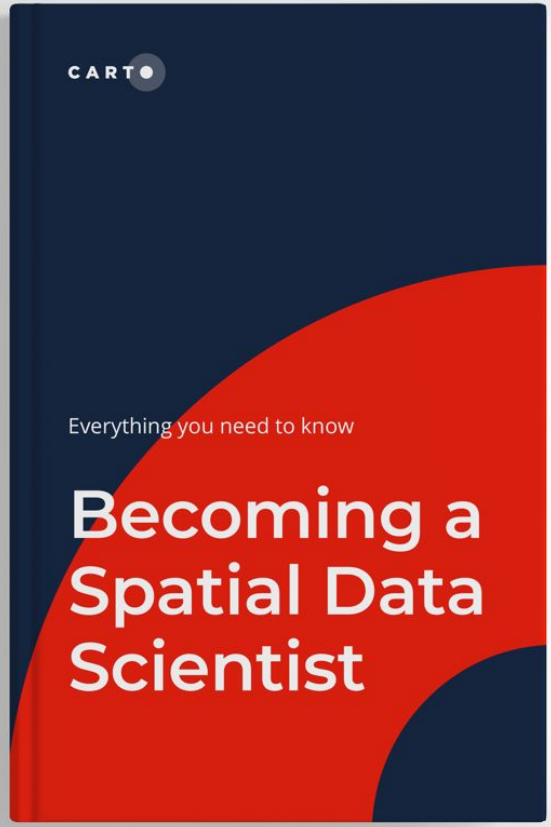
09:10 a.m. - 09:40 a.m.

"Spatial data science treats location, distance, and spatial interaction as core aspects of the data"
(Prof. Luc Anselin)



Ready to become a Spatial Expert?

https://go.carto.com/ebooks/spatial-data-science



https://github.com/CartoDB/data-science-book

Notebooks

Table of Contents

Chapter 1

- `Visualizing spatial data with CARTOframes` (static preview) - a notebook for easily visualizing your data on a map using CARTOframes.
- `Computing measures of spatial dependence` (static preview) - a notebook for exploring spatial dependence in your data and visualize the results using CARTOframes.
- `Discrete spatial models` (static preview) - a notebook with examples of spatial models for discrete processes and visualize the results using CARTOframes.
- `Continous spatial models` (static preview) - a notebook with examples of spatial models for continuous processes and visualize the results using CARTOframes.

Chapter 2

- `Agglomerative Clustering` (static preview) - a notebook demonstrating how to create spatially constrained clusters using agglomerative clustering
- `DBSCAN` (static preview) - a notebook demonstrating how to create clusters of points in geographic coordinates
- `SKATER` (static preview) - a notebook demonstrating how to create spatially constrained clusters that are homogeneous

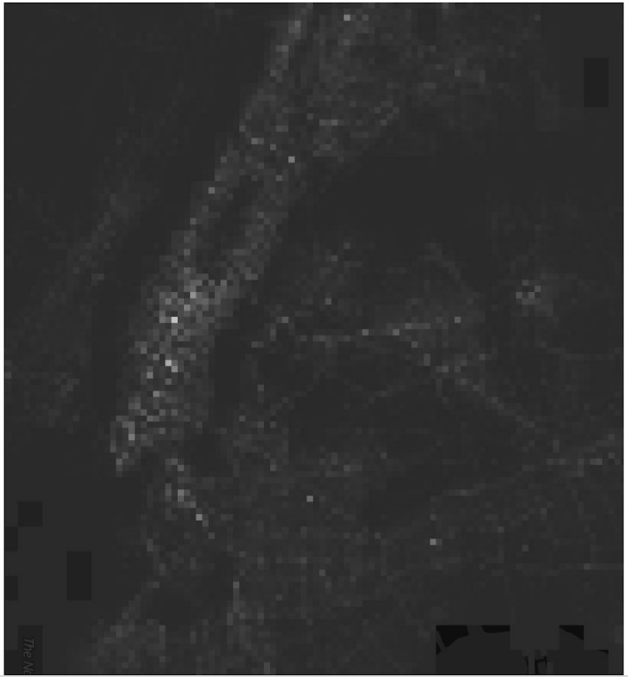
Chapter 3

- `Travelling Salesman Problem` (static preview) - a notebook demonstrating how to solve travelling salesman problem.

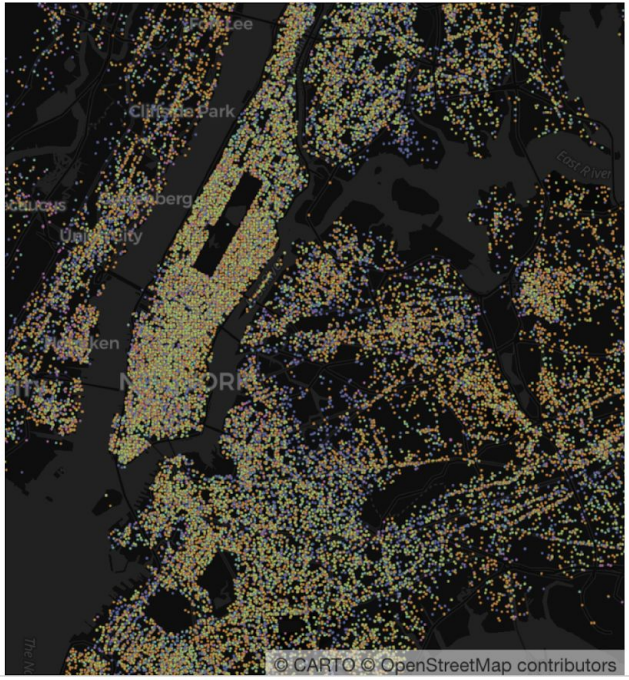
Spatial data comes in all forms and shapes

MEDIAN HOUSEHOLD INCOME

NUMBER OF VISITORS FROM GPS SOURCES



POI LOCATIONS BY CATEGORY



Why is Spatial special? Spatial Dependence

"Everything is related to everything else, but near things are more related than distant things."

(Tobler, 1970)

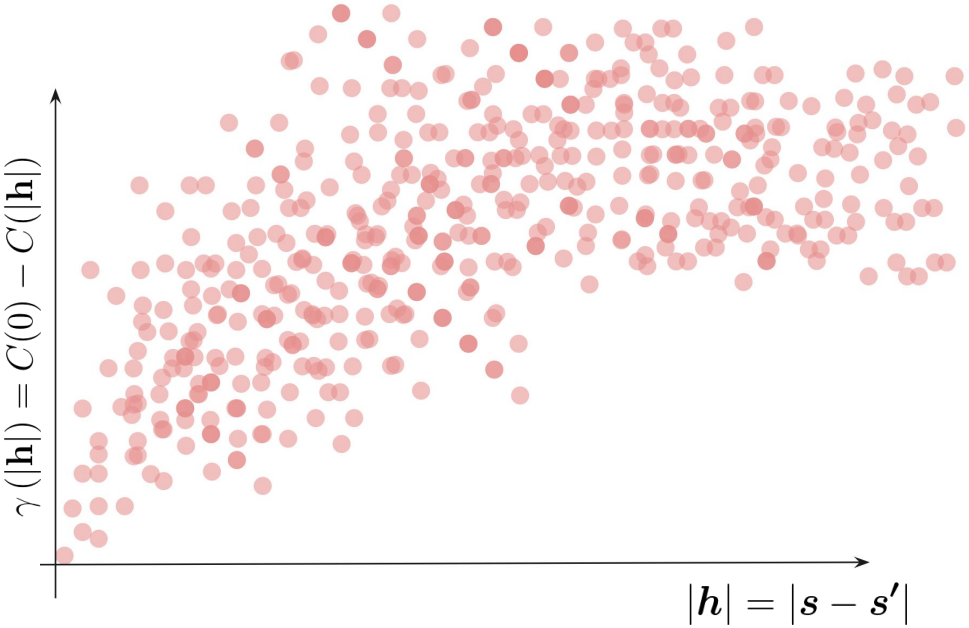
1. CONTINUOUS PROCESSES

A **Gaussian Processes (GP)** is parameterized by a mean function and covariance function

- as $|\boldsymbol{h}| = |\boldsymbol{s} - \boldsymbol{s'}|$ ↑ then $C(\cdot)\downarrow$
- $C(\cdot)$ depends on some parameters

  e.g.

$$C\left(|\mathbf{h}|\right) = \sigma^2 exp(-\frac{1}{\rho}|\mathbf{h}|)$$



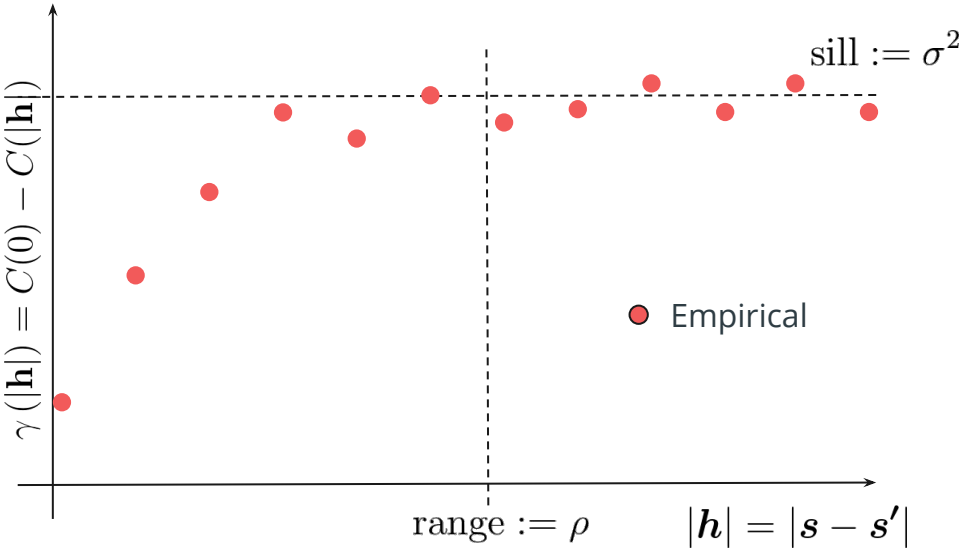
1. CONTINUOUS PROCESSES

A **Gaussian Processes (GP)** is parameterized by a mean function and covariance function

- as $|\boldsymbol{h}| = |\boldsymbol{s} - \boldsymbol{s'}| \uparrow$ then $C(\cdot)\downarrow$
- $C(\cdot)$ depends on some parameters

  e.g.

$$C\left(|\mathbf{h}|\right) = \sigma^2 exp(-\frac{1}{\rho}|\mathbf{h}|)$$



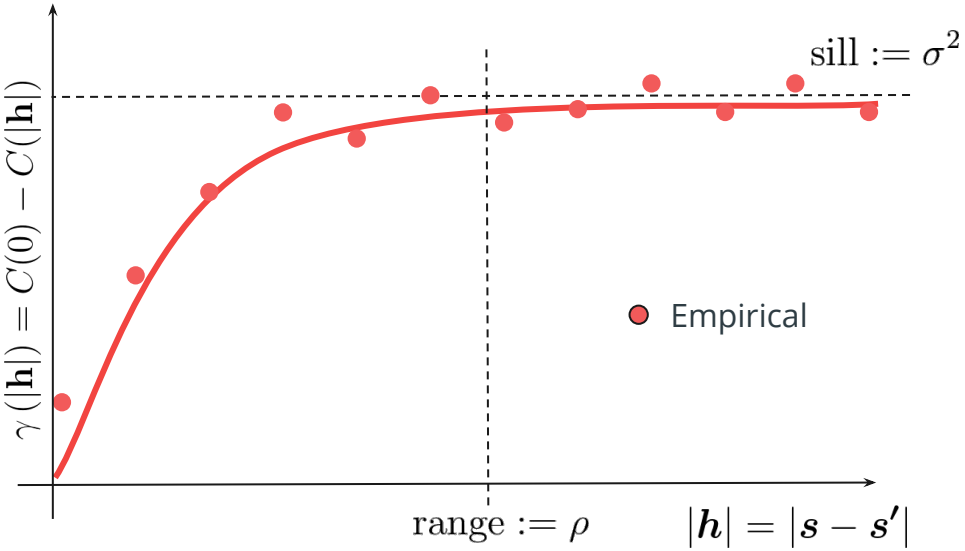
1. CONTINUOUS PROCESSES

A **Gaussian Processes (GP)** is parameterized by a mean function and covariance function

- as $|\boldsymbol{h}| = |\boldsymbol{s} - \boldsymbol{s'}|$ ↑ then ↓
- $C(\cdot)$ depends on some parameters

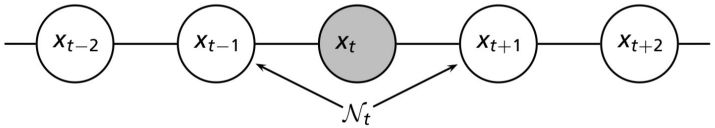
  e.g.

$$C\left(|\mathbf{h}|\right) = \sigma^2 exp(-\frac{1}{\rho}|\mathbf{h}|)$$

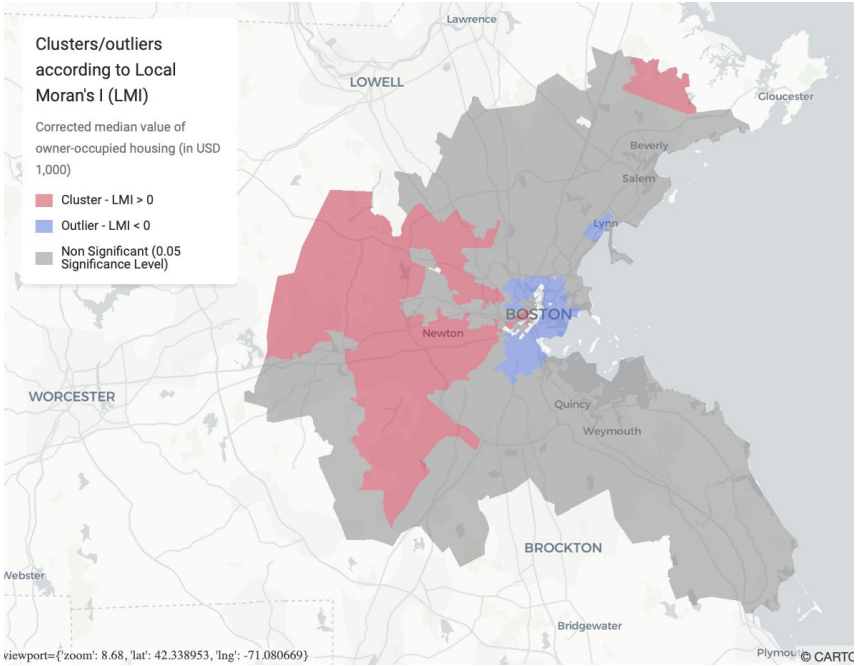


2. DISCRETE PROCESSES

- Neighborhood structures: **Gaussian Markov Random Fields (GMRF)**

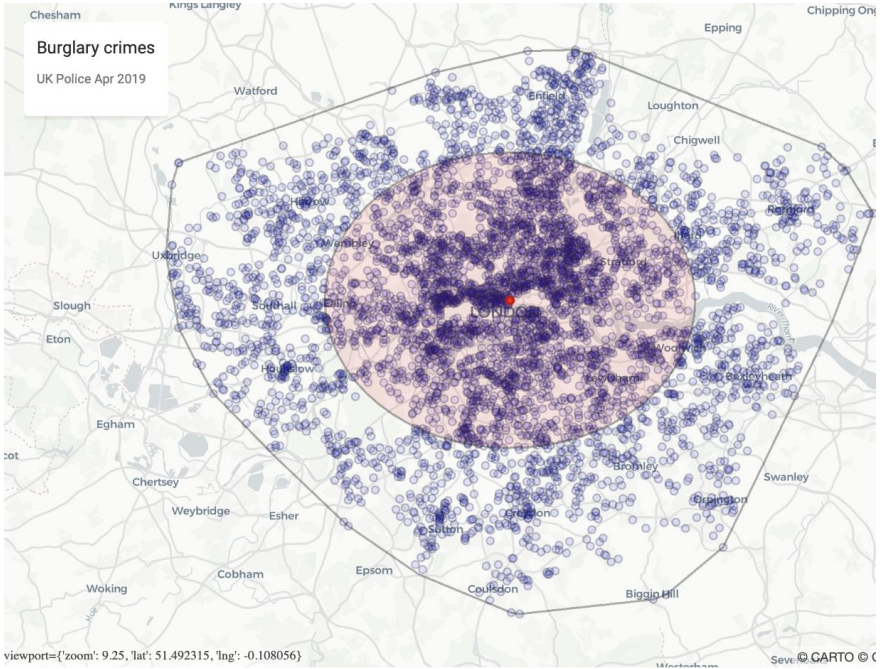


- **Autocorrelation statistics** (e.g. Moran's *I*)



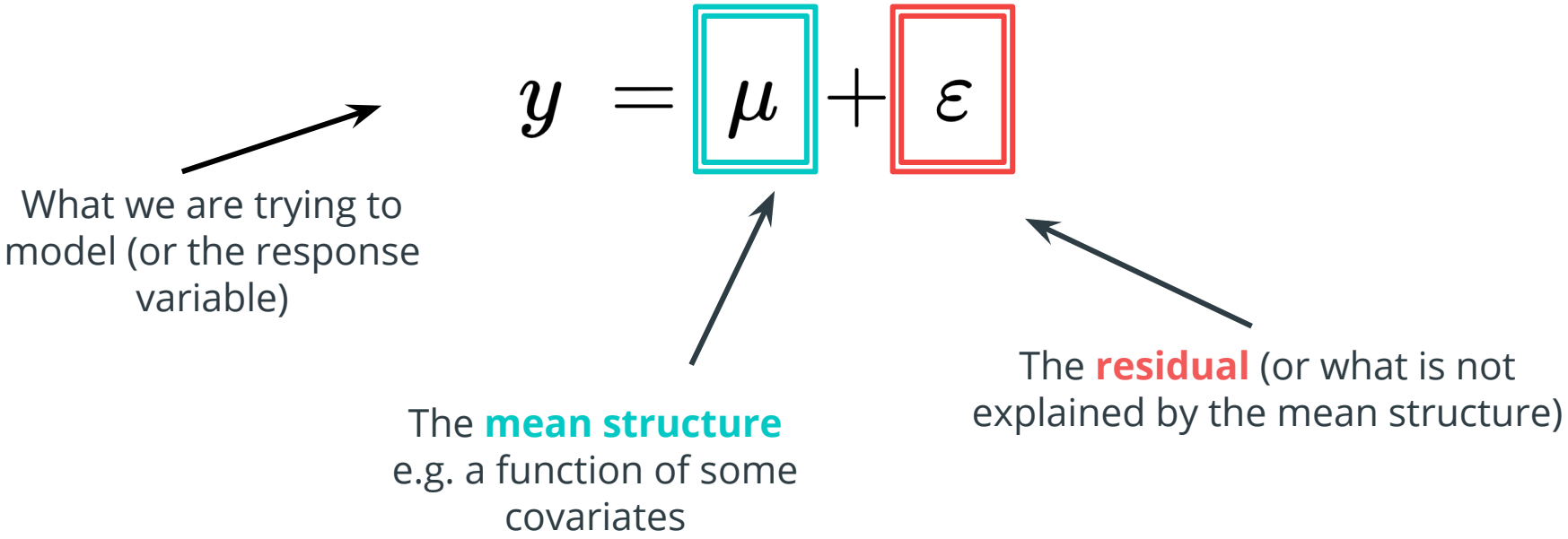
3. POINT PATTERNS

- **Complete Spatial Randomness**
- **Summary statistics**
- **Nearest neighbor analysis** (G-function, K-function, etc.)

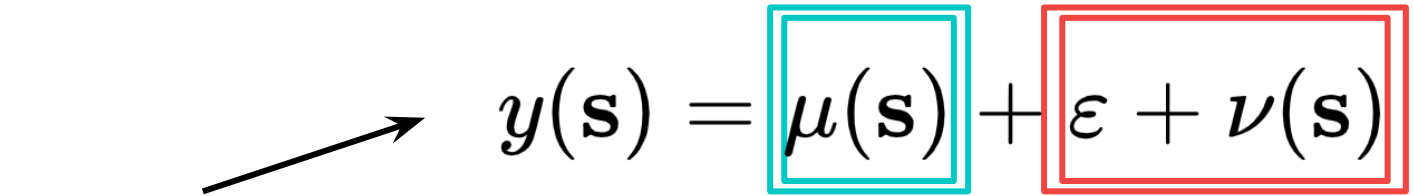


Spatial Modeling: Leveraging Location in Prediction

Spatial modelling



Spatial modelling



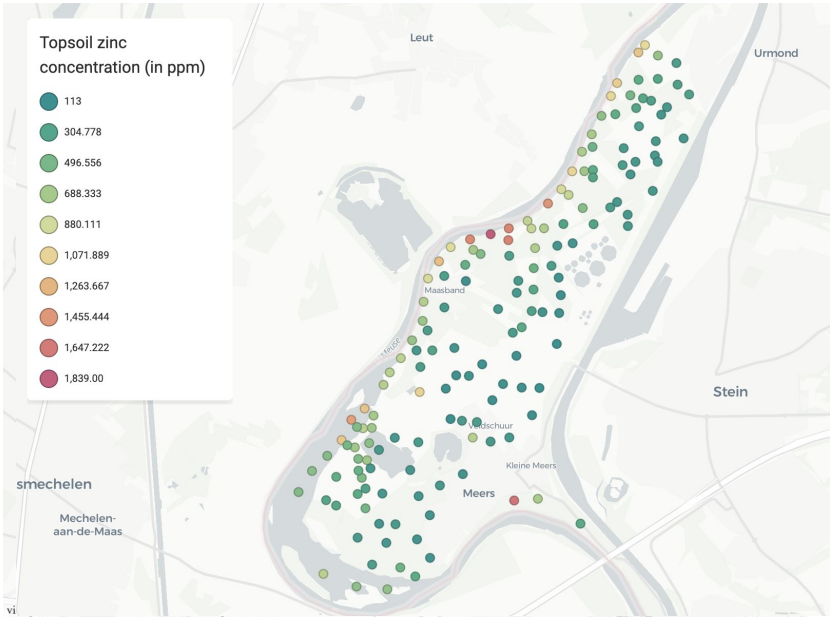
What we are trying to model (or the response variable)

The **mean structure** e.g. a function of some covariates

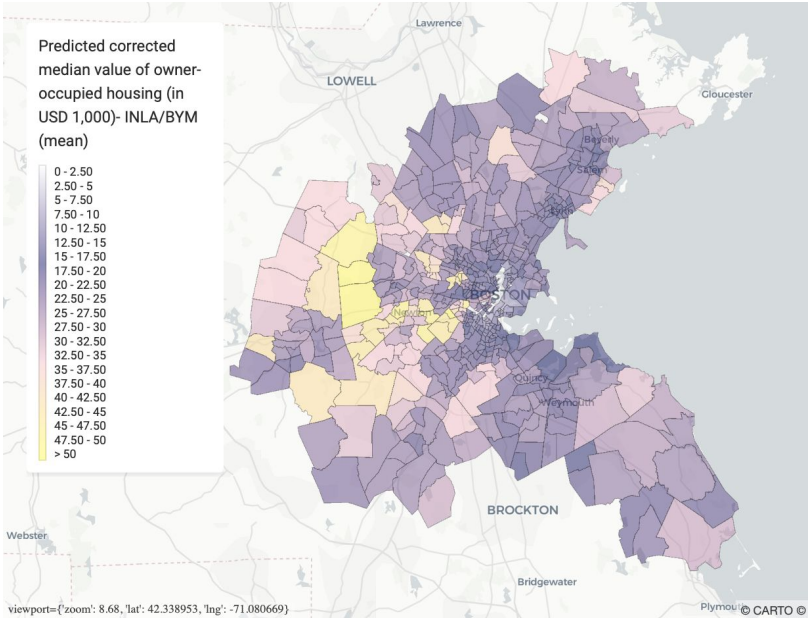
The **residual** (or what is not explained by the mean structure)

Spatial modelling

Continuous Spatial Error Models

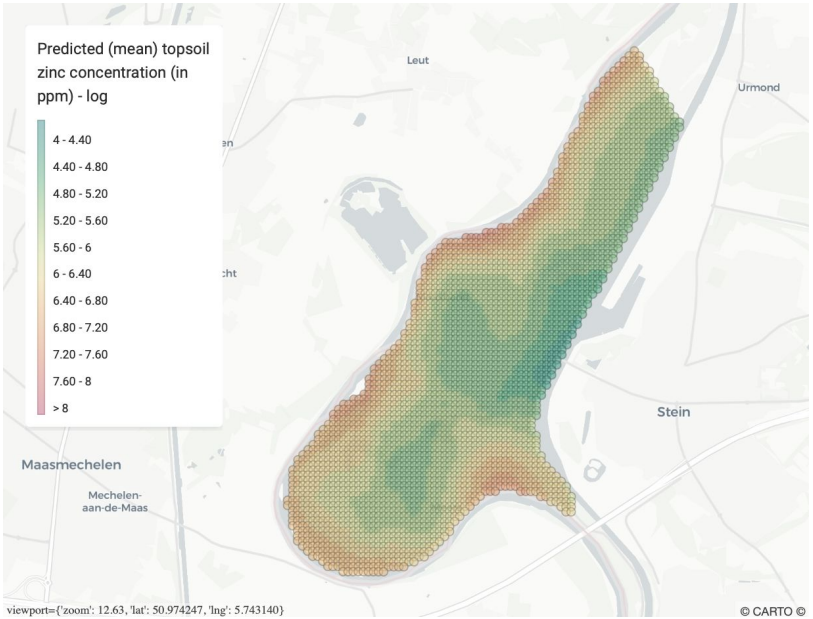


Discrete Spatial Error Models

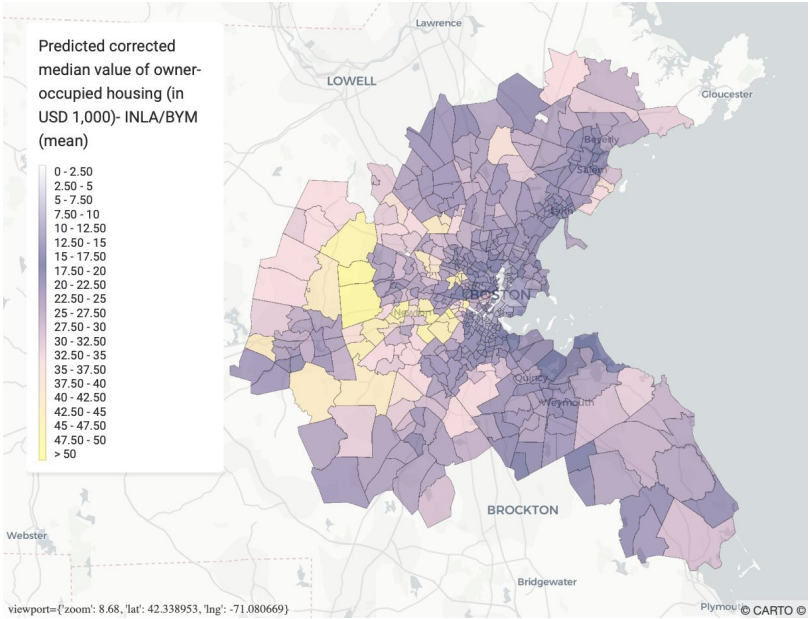


Spatial modelling

Continuous Spatial Error Models



Discrete Spatial Error Models



Spatial Clustering and Regionalization

Clustering VS Spatial Clustering

Clustering

Uses data attributes to create classes that, via those attributes, are different while staying alike within that category

- Longitude and latitude can be included as one of these attributes
- e.g. **K-means**

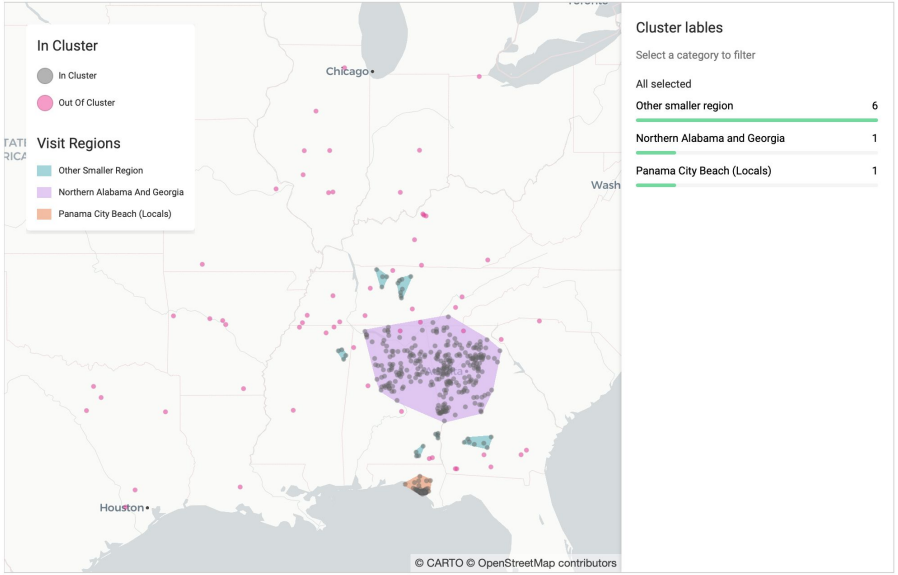
Spatial Clustering

Groups together points that are close to each other based on a distance measurement

- e.g. **DBSCAN, GENERALIZED DBSCAN**

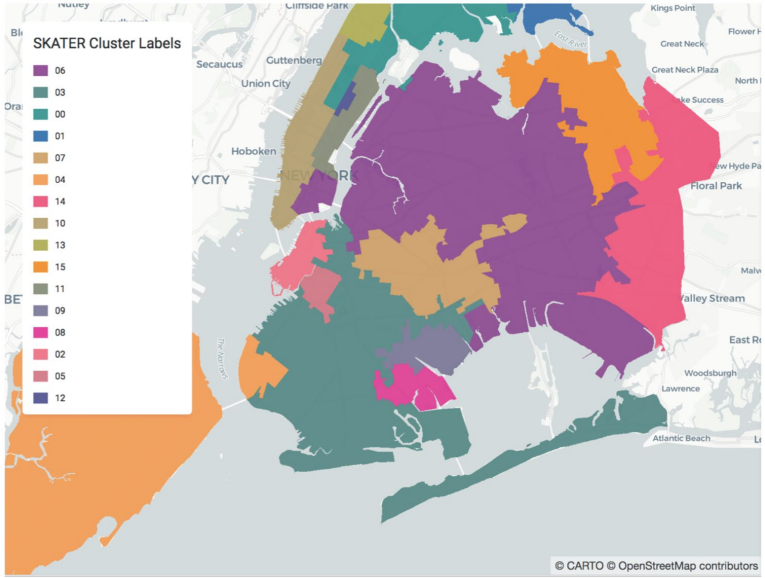
Clustering VS Regionalization

Clustering



Using DBSCAN

Regionalization



Using SKATER

Logistics Optimization with Spatial Analysis

Optimization

A typical optimization model consists of the following components:

- **Decision Variables**
  e.g. whether to open a distribution center (DC) at a specific location, whether a zip code is served by a DC, or which truck will serve one customer and when)

- **Objective Function**
  e.g. costs, service level, etc.

- **Constraints**
  e.g. physical constraints (a truck cannot transport more than its capacity), business constraints (every client should not be further than 20 miles away from the closest DC)

Exact VS approximate algorithms

Exact

Find the actual optimal solution

- Simplex Algorithm

Approximate

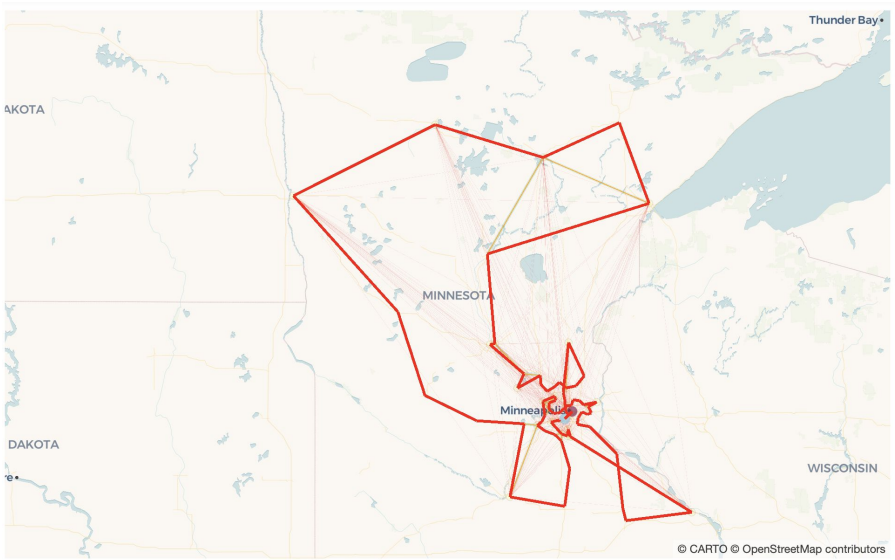
Close as possible to the optimum value in a reasonable amount of time

- Simulated Annealing
- Christofides Algorithm

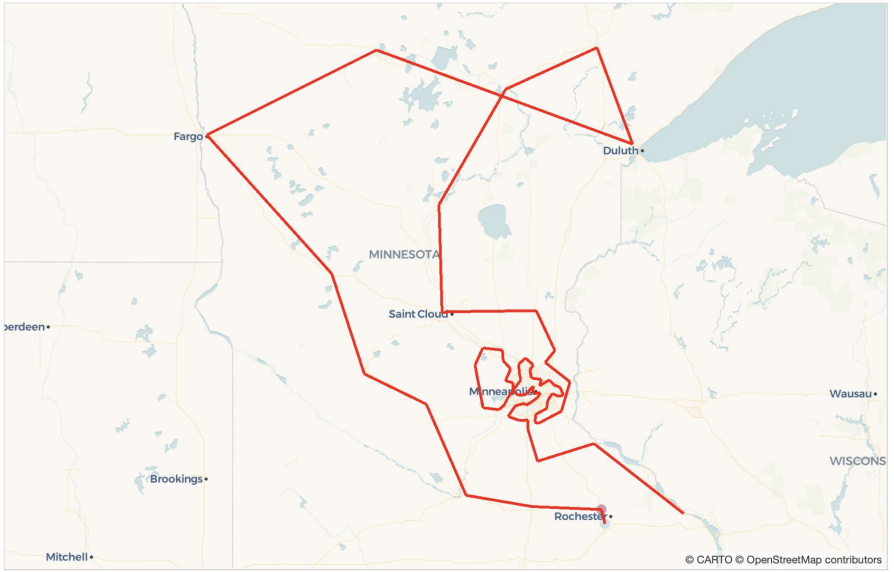
Solving the Traveling Salesman Problem

Given a list of cities and the distances between each pair of cities, what is the shortest possible route that visits each city and returns to the origin city?

Christofides Algorithm

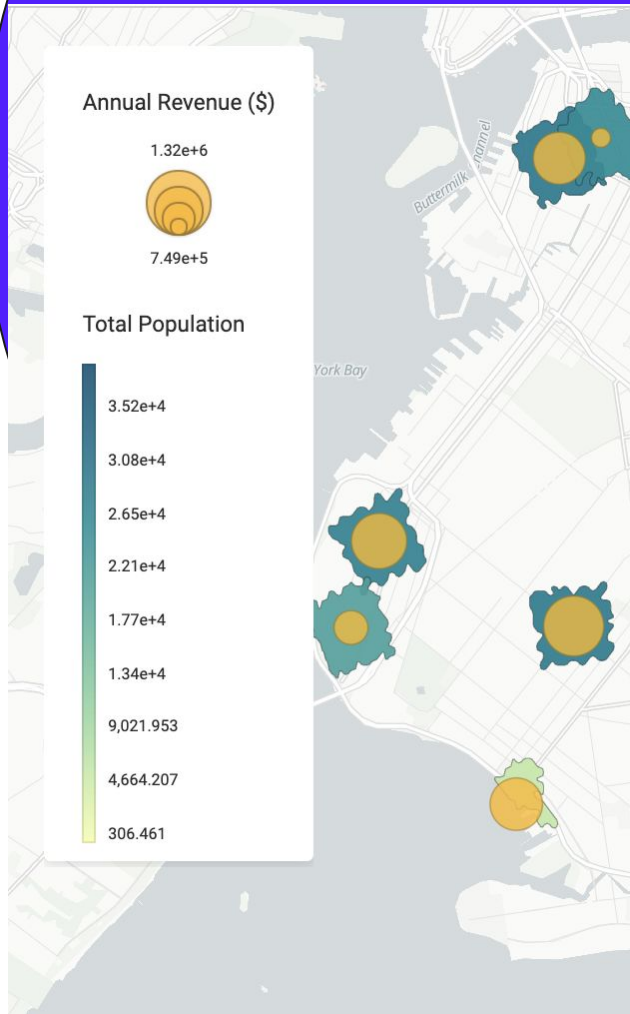


Ant Optimization



Data Science Workflows using CARTO

09:40 a.m. - 10:00 a.m



HOW IT WORKS

CARTO turns your Location Data Into Business Outcomes

Whether it's more efficient delivery routes, strategic store placements or targeted geomarketing campaigns - CARTO makes it simple in 5 key steps:

Data Ingestion

Data Enrichment

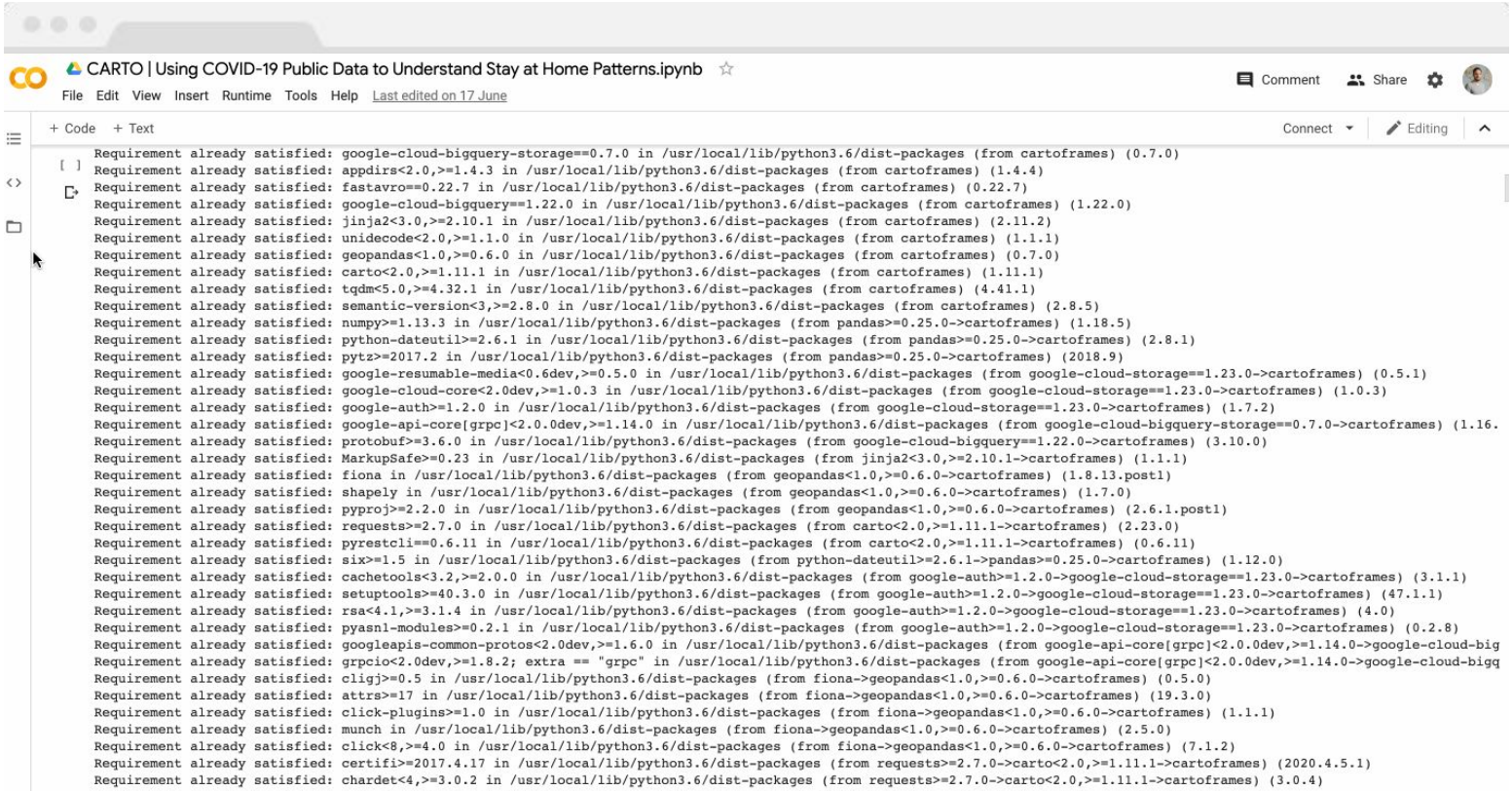
Analysis

Solutions & Visualization

Integrations

CARTOframes

A Python Library to Facilitate your Spatial Data Analysis Workflow



Why CARTOframes?

Powering end-to-end data science workflows

Explore

Clean, geocode, and visualize your data straight out of Jupyter notebooks.

Enrich

Access a wide range of datasets - all on standardized spatial aggregations to reduce your time to insight.

Analyze

Get insights from your data using our API and your own libraries, functions, and workflows.

Share

Once your analysis is done, add widgets and share your results.

1. Explore

➔ Manage your data

Load a CSV file
Load data from a CSV file

Load a JSON file
Load data from a JSON file

Load a GeoJSON file
Load data from a GeoJSON file

Load a shapefile
Load data from a shapefile

Load a CARTO table
Load data from a CARTO table

Load a CARTO SQL query
Load data from a CARTO table using a SQL Query

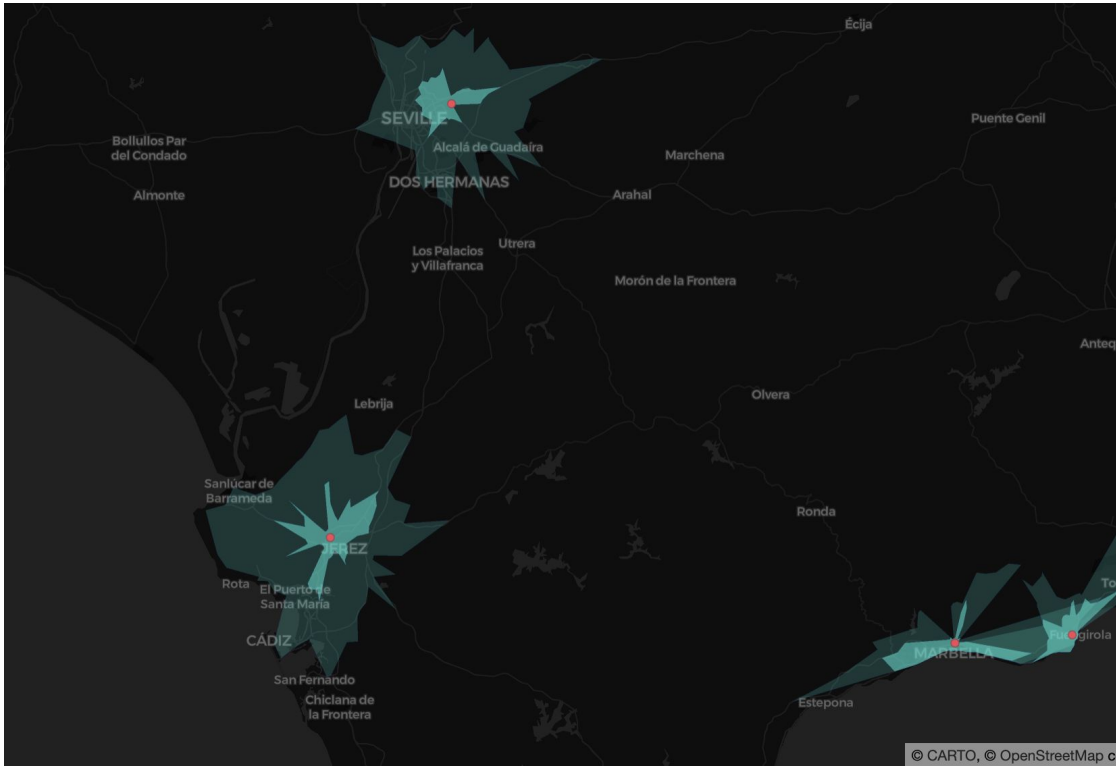
Upload to CARTO
Upload data to CARTO

Change CARTO table privacy
Change the privacy of a CARTO table

1. Explore

➔ Get your data ready

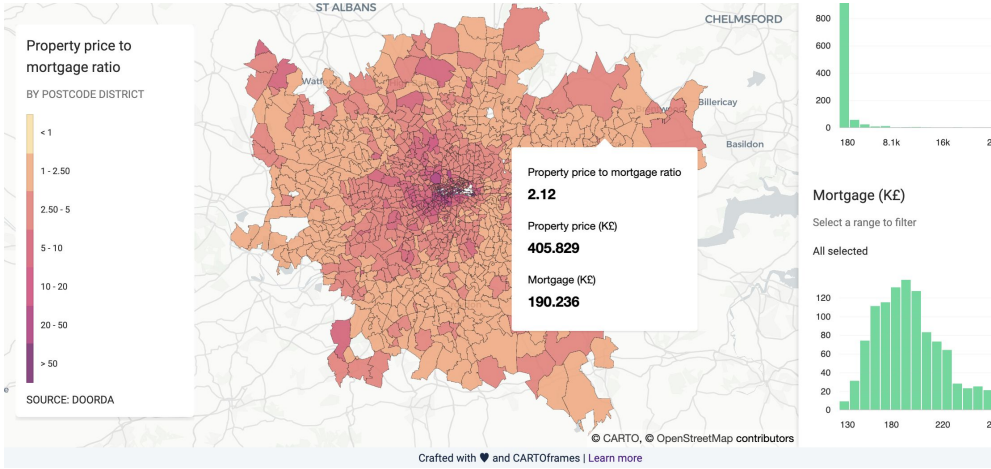
- Geocode large datasets in just one request
- Create isochrones for your points



1. Explore

➔ Visualize

- Local data and hosted datasets
- Maps with multiple layers
- Styling for numerical and categorical variables
- Custom basemaps
- Legends, pop-ups, and widgets
- Layouts



2. Enrich

➔ CARTO Data Observatory

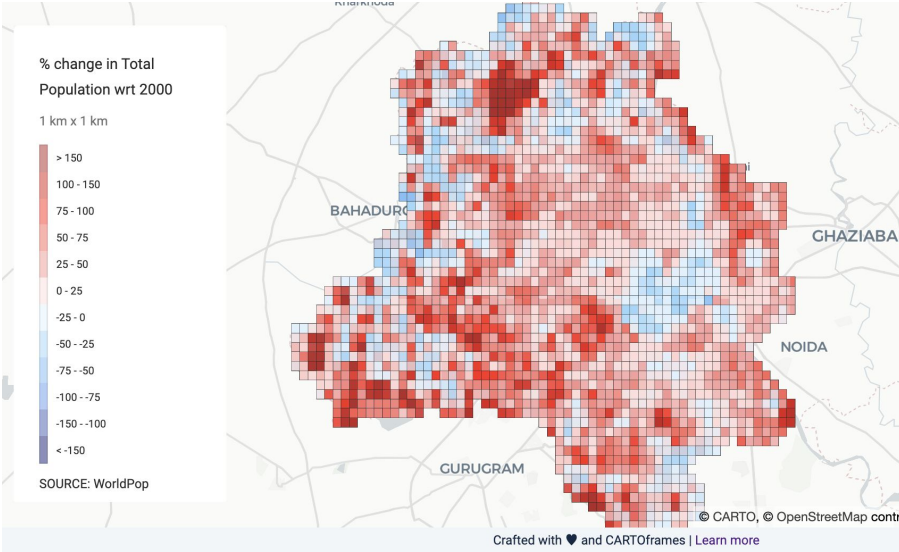
- Access to different location data streams on common geometries. Working with market-leaders, we bring together high-quality curated datasets to reduce the time to insight. https://carto.com/platform/spatial-data-catalog/

Financial	Housing
Human Mobility	POI's
Road Traffic	Environmental
Demographics	Global Boundaries
GeoSocial	COVID-19

2. Enrich

➔ Discover and enrich your data

- Direct access to our Data Observatory
- Open and premium datasets
- Discover the data you need
  - By category
  - By country
  - By geography
  - By provider
- Check stats about available datasets
- Request a dataset
- Enrich your dataframe (points or polygons)

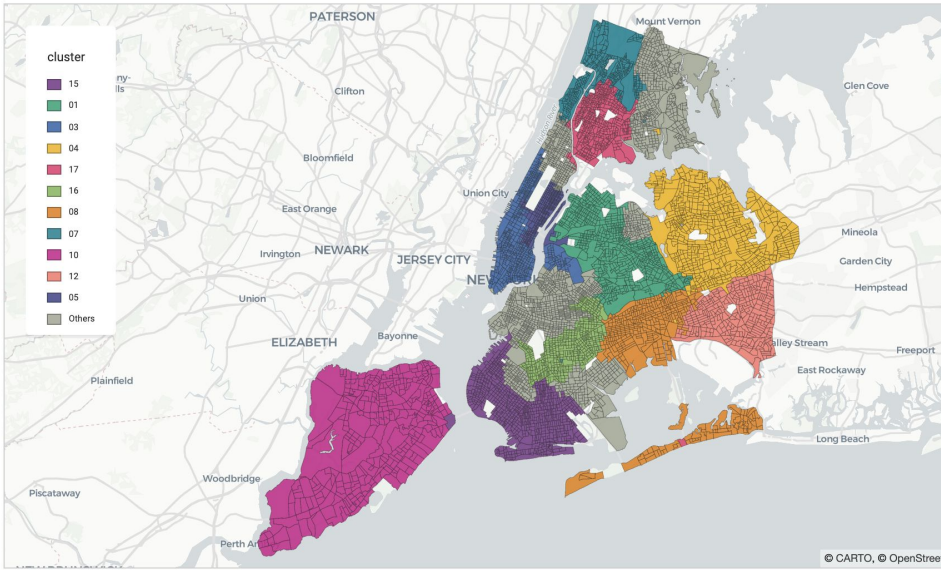


3. Analyze

We are working to help you *productize* your workflows and reduce your data preparation time, enabling you to focus on your analysis.

- ➔ **Feature engineering (soon)**
- ➔ **Spatial Clustering (soon)**
- ➔ **Data partitioning for spatial data (soon)**
- ➔ **Spatial projection (soon)**

```
[390]: Map(Layer(nyc_geoms_enriched.reset_index(),
                 style=color_category_style('cluster'),
                 popup_hover=[popup_element('geoid'),
                              popup_element('index'),
                              popup_element('cluster')]))
```



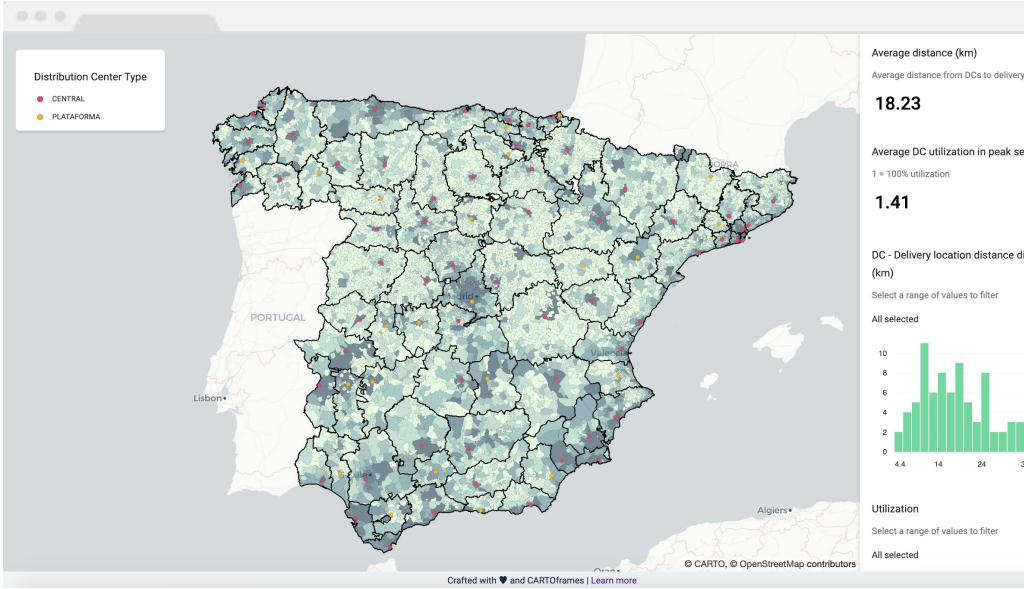
4. Share

➔ Prepare your analysis output

- Make your analysis easy to consume by others in your organization by adding widgets: histogram, category, animation control, time series.

➔ Publish

- Publish your map to CARTO and get the shareable link as response.



Break

10:00 a.m. - 10:10 a.m.

Hands-on with CARTO
Practical Spatial Data Science in Python

10:10 a.m. - 11:30 a.m.

Using CARTO tech stack, in this session we will go through a step by step demo using Jupyter notebooks, from data exploration, to external data discovery and augmentation, to model formulation and results.

- ***Site selection***: where should Starbucks open new coffee shops in Long Island, NY? In this demo we will go through a typical site selection use case, from modelling the revenues of the existing stores as a function of socioeconomic covariates, to predicting the potential revenues in new locations.

- ***Logistic spatial optimization***: where should a parcel delivery company locate their distribution and fulfilment centers? What areas should they service? In this demo we will go through a supply chain network optimization use case, from analysing past data to identify spatio-temporal patterns to building an optimization model to analyze and quantify the impact of changes in the current network.

Q&A and discussion

11:30 a.m. - 11:45 a.m.

Online

October 19th - 23rd 2020

Get your free ticket now

Spatial Data Science Conference

Thank you for listening!
Any questions?

Request a demo at CARTO.com

Giulia Carella
Data Scientist // giulia@carto.com

Miguel Álvarez
Data Scientist // marvarez@carto.com

CARTO

JPM Big Data and AI Research

11:45 a.m. - 12:00 p.m.